Vicenç Torra
Yasuo Narukawa (Eds.)

LNAI 5285

Modeling Decisions for Artificial Intelligence

5th International Conference, MDAI 2008
Sabadell, Spain, October 2008
Proceedings

Springer

Lecture Notes in Artificial Intelligence 5285

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Vicenç Torra Yasuo Narukawa (Eds.)

Modeling Decisions for Artificial Intelligence

5th International Conference, MDAI 2008
Sabadell, Spain, October 30-31, 2008
Proceedings

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial (IIIA)
Consejo Superior de Investigaciones Científicas (CSIC)
Campus UAB, 08193 Bellaterra, Catalonia, Spain
E-mail: vtorra@iiia.csic.es

Yasuo Narukawa
Toho Gakuen
3-1-10 Naka, Kunitachi, Tokyo 186-0004, Japan
E-mail: narukawa@d4.dion.ne.jp

Library of Congress Control Number: 2008935891

CR Subject Classification (1998): I.2, F.4.1, F.1, H.2.8, I.6

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-88268-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-88268-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12538812 06/3180 5 4 3 2 1 0

Preface

This volume contains papers presented at the 5th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2008), held in Sabadell, Catalonia, Spain, October 30-31. This conference followed MDAI 2004 (Barcelona, Catalonia, Spain), MDAI 2005 (Tsukuba, Japan), MDAI 2007 (Tarragona, Catalonia, Spain), and MDAI 2008 (Kitakyushu, Japan) with proceedings also published in the LNAI series (Vols. 3131, 3558, 3885, and 4617).

The aim of this conference was to provide a forum for researchers to discuss the theory and tools for modeling decisions, as well as applications that encompass decision-making processes and information-fusion techniques.

The organizers received 43 papers from 15 different countries, from Asia, Europe, and America, 19 of which are published in this volume. Each submission received at least two reviews from the Program Committee and a few external reviewers. We would like to express our gratitude to them for their work. The plenary talks presented at the conference are also included in this volume.

The conference was supported by the IIIA-CSIC, the UNESCO Chair in Data Privacy, the Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT), the Catalan Association for Artificial Intelligence (ACIA), the European Society for Fuzzy Logic and Technology (EUSFLAT), the Spanish MEC (ARES - CONSOLIDER INGENIO 2010 CSD2007-00004), and the City of Sabadell.

July 2008

Vicenç Torra
Yasuo Narukawa

Organization

General Chair

Vicenç Torra, IIIA-CSIC, Catalonia, Spain

Program Chairs

Vicenç Torra, IIIA-CSIC, Catalonia, Spain
Yasuo Narukawa, Toho Gakuen, Japan

Advisory Board

L. Godo, J. Kacprzyk, S. Miyamoto, M. Sugeno, R. R. Yager

Program Committee

G. Beliakov, U. Bodenhofer, T. Calvo, J. Domingo-Ferrer, J. Dujmovic, B. H. Far, M. Grabisch, E. Herrera-Viedma, J. Herranz, K. Hirota, M. Inuiguchi, H. Kikuchi, I. Kojadinovic, J.-L. Marichal, R. Meo, R. Mesiar, T. Murofushi, M. Ng, G. Pasi, A. Valls, Z. Xu, Y. Yoshida, G. Zhang, N. Zhong

Local Organizing Committee

I. Cano, J. Nin

Additional Referees

Y. Endo, M. Komornikova, J. Long, J. M. Mateo-Sanz, J. Nin, E. Okamoto, F. Sebé, A. Solanas, A. Stupnanova-Markova

Supporting Institutions

The UNESCO Chair in Data Privacy
The Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
The Catalan Association for Artificial Intelligence (ACIA)
The European Society for Fuzzy Logic and Technology (EUSFLAT)
The City of Sabadell
The Spanish MEC (ARES - CONSOLIDER INGENIO 2010 CSD2007-00004)

Table of Contents

Clustering and Similarity

Computational Intelligence and Optimization

Data Privacy

Toward Elucidating Language Functions in the Brain

Michio Sugeno

Faculty of Culture and Information Science, Doshisha University,
1-3 Tatara Miyakodani, Kyotanabe City, Kyoto, 610-0394 Japan

Human intelligence is characterized by the use of language rather than the brain hardware. The human brainware consists of a neural system as hardware and a language system as software. Language was created by the brain hardware, and the human brain evolved together with language over millions of years. It is, therefore, necessary to take two approaches to create the human brain: a hardware-centered approach and a software-centered approach. While the hardware-centered approach is based on computational neuroscience, it is possible to base the software-centered approach on linguistics.

With this in mind, we discuss the higher-order language functions in the brain. There are three approaches to elucidate the language functions: top-down, intermediate, and bottom-up. In the top-down approach we start from existing phenomena of language and in the bottom-up approach we start from neural processes to deal with language. The intermediate approach is something lying between the two. A major difficulty in elucidating language functions is that we lack experimental tools to directly observe detailed brain activities in dealing with language. Therefore it is desirable and rather inevitable to combine all possible approaches.

In this study we refer, as the basic theory, to Systemic Functional Linguistics (SFL) initiated by Halliday. SFL systematically describes language the system of which consists of four strata: phonology, lexicogrammar, semantics, and context. There are three metafunctions in language: ideational, interpersonal, and textual. These metafunctions penetrate the four strata. For example in the stratum of semantics, metafunctions appear as ideational, interpersonal, and textual meanings. Interpersonal meaning is ordinary meaning concerned with construing experience by seeing, hearing, thinking, and so on, interpersonal meaning concerned with enacting interpersonal relations through language, and textual meaning concerned with organizing ideational and interpersonal meanings as discourse.

In the top-down approach, we have developed a computational model of language which consists of the semiotic base describing the system of language, and the algorithms of text understanding/generation with the semiotic base. The brain is supposed to contain a neuro-computational system of language in its nature. Our ultimate goal is to identify this system. Since existing language was created by the brain and language phenomena can be fully observed, it is possible to develop a computational model of language starting from a theoretical

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 1–2, 2008.

model provided by linguistics. Based on a computational model, we shall be able to infer a neuro-computational model of language.

As for the intermediate approach, we discuss the stratified system of language in the brain by introducing some clinical evidence obtained from studies on aphasia. According to the Yamadorifs studies, the stratification of language suggested by SFL is also realized in the brain; lexicogrammar and semantics are processed in the left hemisphere while context is processed in the right hemisphere.

In the bottom-up approach, we have conducted brain experiments to analyze dynamical processes in understanding the meanings of texts with and without honorific expressions. The aim of this study is to elucidate the difference of brain activities during processing ideational meaning and interpersonal meaning, where texts with honorific expressions contain interpersonal meaning and texts without honorific expressions mainly hold ideational meaning.

Two kinds of sentences with and without honorific expressions were sequentially shown to subjects. We measured electroencephalograms and acquired event-related potentials. Then using equivalent current dipole source localization method, we analyzed the activation of the brain. It was found that the brain activities for understanding sentences with honorific expressions are different from those produced for understanding sentences without honorific expressions.

Privacy-Preserving Similarity Evaluation and Application to Remote Biometrics Authentication

Hiroaki Kikuchi[1], Kei Nagai[1], Wakaha Ogata[2], and Masakatsu Nishigaki[3]

[1] Department of Communication and Network Engineering,
School of Information and Telecommunication Engineering, Tokai university
1117 Kitakaname, Hiratsuka, Kangawa, 259-1292, Japan
Tel.: +81-463-58-1211, Fax: +81-463-50-2412,
kikn@tokai.ac.jp
[2] Graduate School of Innovation Management,
Tokyo Institute of Technology, Tokyo, Japan
wakaha@mot.titech.ac.jp
[3] Graduate School of Science and Technology, Shizuoka University
Shizuoka, Japan
nishigaki@inf.shizuoka.ac.jp

Abstract. In this paper, a new method for secure remote biometric authentication preventing the vulnerability of compromised biometrics is presented. The idea is based on a public-key cryptographical protocol, referred as *Zero-knowledge Proof*, which allows a user to prove that she has surely a valid biometric data without revealing the data. Hence, the scheme is free from the risk of disclosure of biometric data. Even if a malicious administrator has a privilege access to the private database, it is infeasible for him to learn the private template. This paper studies two well-known definitions, the *cosine correlation* and the *Euclidean distance* as similarities of given two feature vectors. Both similarities are defined with some multiplications and additions, which can be performed in privacy-preserving way because of the useful property of public-key commitment scheme, *additive homomorphism*. The estimation based on the experimental implementation shows that the private Euclidean distance scheme archives better accuracy in terms of false acceptance and rejection than the private cosine coloration scheme, but it requires about $5/2n\ell$ overhead to evaluate n-dimension feature vectors consisting of ℓ-bit integers.

1 Introduction

Biometrics identifiers are now commonly used to identify individuals in more secure and more efficient ways than the conventional password-based methods. Typically, the biometric identifiers including fingerprint, vein, iris, facial images are scanned and processed in appropriate algorithm to extract a *feature vector*, called *template*, which will be compared to newly scanned image to verify that the owner of the biometric data is legitimate or not.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 3–14, 2008.

The biometric recognition, however, are mostly made in *local* environment, e.g., a matching with the template data stored in secure smartcard (in ATM cards), or a user authentication at personal laptop PCs. The reason of limitation in local is the known vulnerabilities of *remote biometric authentication* that once a biometric template is stolen, it is stolen forever and can not be recovered. If we store our biometric data to some service provider, we immediately face risks that the server may be compromised, or a malicious administrator of the server may learn our highly private data and can disclose it.

Many researchers points out the issue in remote biometrics authentication and several attempts addressing it have been made. Ratha et al. [4] propose a "cancelable biometrics", using a morphing technique to transform a biometrical data into a randomized form, which depends on given morphing function. Jeong et al. [5] propose a changeable biometrics for face recognition using the principal component analysis (PCA) and the independent component analysis (ICA). Given two vectors chosen from PCA and ICA coefficients, they extract from an input face image the transformed vector according to a scrambling rule. When the transformed template is compromised, the scrambling rules is replaced by a new one. Juels and Sudan's "fuzzy vault scheme" [6] is an improvement upon the previous work by Juels and Wattenberg [8]. In [6], they use the polynomial reconstruction problem based on an error-collection code such as the Reed-Solomon. Clancy et al. [7] proposed a "fingerprint vault system" based on the fuzzy vault. Using multiple minutiae location sets, they use a canonical positions of minutiae, as the elements of a set. Uludag et al. [12] proposed a fuzzy vault system for fingerprint using the Lagrange interpolation and the Cyclic Redundancy Check (CRC) for testing polynomial reconstruction instead of the error-collection step.

In this paper, we present a new method for secure remote biometric authentication preventing the vulnerability of compromised biometrics. Our idea is based on a public-key cryptographical protocol, referred as *Zero-knowledge Proof*, which allows a user to prove that she has surely a valid biometric data without revealing the data. Hence, the scheme is free from the risk of disclosure of biometric data. Even if the administrator with privilege access to the private database is malicious, it is infeasible to learn the private template. Without learning the template stored at the server, he performs an evaluation of similarities between the template and the new input in privacy-preserving way.

The zero-knowledge proof is generally "expensive" in terms of communication and computation costs. The performance of schemes depends on what similarity measure is used to securely evaluated. In this paper, we study two well-known definitions, the *cosine correlation* and the *Euclidean distance* as similarities of given two feature vectors. Both similarities are defined with some multiplications and some additions, which can be performed in privacy-preserving way because of the useful property of public-key commitment scheme, *additive homomorphic*. The estimation based on the experimental implementation shows that the private Euclidean distance scheme achieves better accuracy in terms of false acceptance and rejection than the private cosine correlation scheme, but it requires

about $5/2n\ell$ overhead to evaluate n-dimension feature vectors consisting of ℓ-bit integers.

2 Preliminaries

2.1 Similarities

Let $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ be n-dimensional vectors of R^n. We consider the following two well-known similarities between $\boldsymbol{a}$ and $\boldsymbol{b}$, which will be evaluated in privacy-preserving way in later section.

Definition 1. *A Cosine Correlation is a similarity between* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *defined as*

$$c(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}|| \cdot ||\boldsymbol{b}||} = \frac{a_1 b_1 + \cdots + a_n b_n}{\sqrt{a_1^2 + \cdots + a_n^2}\sqrt{b_1^2 + \cdots + b_n^2}}$$

where $||\boldsymbol{a}||$ *is a norm of* $\boldsymbol{a}$.

Definition 2. *An Euclidean Distance,* $d(\boldsymbol{a}, \boldsymbol{b})$*, is defined as*

$$d(\boldsymbol{a}, \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}|| = \sqrt{\sum_{i}^{n} (a_i - b_i)^2}.$$

For simplification, taking the normalization of $\boldsymbol{a}$ and $\boldsymbol{b}$, we can reduce the computational cost of cosine correlation as $c(\boldsymbol{a}/||\boldsymbol{a}||, \boldsymbol{b}/||\boldsymbol{b}||) = \boldsymbol{a} \cdot \boldsymbol{b}$. Taking squared as $d(\boldsymbol{a}, \boldsymbol{b})^2$, we can omit the computation of squire root for Euclidean similarity.

2.2 Secure Commitment

A *commitment* is a cryptographical primitive to commit to a value while keeping it hidden, and to reveal the committed value later.

A function $E(m, r)$ is considered as secure commitment to message m, where r is random number, if it satisfies

1. no information reveals from $E(m, r)$, and
2. no one finds $m' \neq m$ and r' such that $E(m, r) = E(m', r')$.

Fujisaki and Okamoto proposed in [2] a probabilistic commitment scheme based on the integer factorization problem as follows.

Definition 3. *Let* N *be a composite number that no one knows the factors, and* g *and* h *be elements of* Z_N *such that* $\log_g h$ *is not known by anybody. A commitment to* m *is*

$$E(m, r) = g^m h^r \bmod N,$$

where r *is random number.*

The Fujisaki-Okamoto commitment has an *additive homomorphism*, a useful property for privacy-preserving computation, satisfying

$$E(m, r) \times E(m', r') = E(m + m', r + r') \text{ and } E(m, r)^x = E(mx, rx),$$

where the addition $m + m'$ is an ordinary arithmetic (not modular arithmetic) since we don't know the order of g and h.

2.3 Zero-Knowledge Proof of Commitment

We introduce a cryptographical protocol for proving that a committed value m lines in a specific interval $[a, b]$ without revealing m, often known as Boudot's Range Proof [9].

Definition 4. *Let F be a commitment $E(m, r)$ to message m. A* proof of knowledge of commitment *is a cryptographical protocol allowing a prover to show that committed m is in $[a, b]$ without revealing m to a verifier, denoted by*

$$PK\left\{m, r \mid F = E(m, r) \wedge m \in [a, b]\right\}$$

where r is uniformly chosen over $[-2^s N+1, 2^s N-1]$ and s is a security parameter (e.g, $s = 160$ [bit]).

The range proof takes about five times of overhead of a standard zero-knowledge proof of the committed value $PK\{m \mid F = E(m, r)\}$. Namely, it is expensive in terms of both computation and communication.

3 Private Similarity Evaluations

3.1 Overview and Assumption

In our model, Alice, a user who tries to prove her identity to server, interacting with Bob, a server who authenticate Alice based on the data Alice has registered with Bob. Assume that Alice does not fully trust Bob, who is a *honest-but-curious* party having a chance to reveal her private information. Instead of her private biometric data $\boldsymbol{x} = (x_1, \ldots, x_n)$, Alice registers the commitment of $\boldsymbol{x}$, $E(\boldsymbol{x})$, from which Bob can not learn $\boldsymbol{x}$. To authenticate her to Bob, Alice scans her fresh biometric data $\boldsymbol{y} = (y_1, \ldots, y_n)$ such that $x_i \approx y_i$ for $i = 1, \ldots, n$, and proves $\boldsymbol{x} \approx \boldsymbol{y}$ to Bob without revealing $\boldsymbol{y}$ (nor $\boldsymbol{x}$) in the zero-knowledge proof of similarities between $\boldsymbol{x}$ and $\boldsymbol{y}$.

There are many efficient protocols for proving several kinds of equalities in zero-knowledge way, and we need to prove privately that $\boldsymbol{y}$ is "close" to $\boldsymbol{x}$. It is not so hard to implement the fuzzy matching if Alice is allowed to access her tamper-proof device to recover $\boldsymbol{x}$ to be compared with new one $\boldsymbol{y}$. In the next section, we will show that the states-of-the-art cryptographical protocols allow us to evaluate similarities between any given committed vectors and to show the difference is within a range, without disclosing private biometric data to anyone. Hence, the protocol is free from the risk of private information disclosure.

3.2 Private Cosine Correlation Evaluation

We show a protocol for secure evaluation of a cosine correlation given $\boldsymbol{x}$ and $\boldsymbol{y}$ in Figure 1.

First of all, Alice needs to compute the commitment to her true private input $\boldsymbol{x}$ using random values $r_1, \ldots, r_n$ chosen uniformly over Z_N, as $E_i = E(x_i/c, r_i)$

Protocol Private-Cosine
Input: $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in Z_N^n$.

1. (Registration) Alice sends to Bob the commitment to her private input $\boldsymbol{x}$, $E_1, \ldots, E_n$, where E_i is defined $E(x_i/c, r_i) = g^{x_i/c} h^{r_i}$ with random $r_i \in Z_N$ for $i = 1, \ldots, n$, and c is the norm of $\boldsymbol{x}$, i.e., $c = ||\boldsymbol{x}||$.
2. (Authentication) Alice computes a commitment to her scanned input $\boldsymbol{y} = (y_1, \ldots, y_n)$, $G_1, \ldots, G_n$ such that $G_i = E_i^{y_i/c'}$ for $i = 1, \ldots, n$, where $c' = ||\boldsymbol{y}||$. She proves to Bob that she knows the committed input $\boldsymbol{y}$ in the zero-knowledge proof
$$PK_1 = PK\left\{ y_i/c' \mid G_i = E_i^{y_i/c'} \right\}$$
for $i = 1, \ldots, n$.
3. Bob verifies PK_1 for all i and then computes $D_C = \prod_i^n G_i$ if PK_1 is valid.
4. Alice computes a similarity $d_C = c(\boldsymbol{x}, \boldsymbol{y})$ and proves to Bob that she knows $\boldsymbol{y}$ that nearly equals to $\boldsymbol{x}$ in the zero-knowledge proof
$$PK_2 = PK\left\{ d_C, R_C \mid D_C = E(d_C, R_C) \wedge d_C \in [\tau_1, 1] \right\},$$
where $R_C = \sum_{i=1}^n r_i y_i / c'$.
5. Bob authenticates Alice if he verifies PK_2.

Fig. 1. Protocol for Cosine correlation evaluation

for $i = 1, \ldots, n$. For reducing computational cost, we use the norm $c = ||\boldsymbol{x}||$ to normalize the committed input x_i. The random values are used for making the commitment indistinguishable against Bob in a sense that he can distinguish two messages with negligible probability.

The key idea of the protocol is to evaluate the cosine correlation between template data $\boldsymbol{x}$ and an input data $\boldsymbol{y}$ without revealing private $\boldsymbol{x}$ and $\boldsymbol{y}$. The additive homomorphic property of the commitment scheme allows Bob to compute the commitment of the cosine correlation between hidden $\boldsymbol{x}$ and $\boldsymbol{y}$ at the third step as follows,

$$D_C = \prod_{i=1}^n G_{i=1}^n = \prod_{i=1}^n E(x_i/c, r_i)^{y_i/c'} = \prod_{i=1}^n E(x_i y_i / cc', r_i y_i / c')$$
$$= E(\frac{1}{||\boldsymbol{x}||\,||\boldsymbol{y}||} \sum_{i=1}^n x_i y_i, \sum_{i=1}^n r_i y_i / c') = E(c(\boldsymbol{x}, \boldsymbol{y}), R_C))$$

where R_C is a random element computed as $\sum_{i=1}^n r_i y_i / c'$. Since Alice is allowed to access the tamper-proof device to obtain random values used to commit $\boldsymbol{x}$, she is able to learn R_C, and thereby get d_C. She also needs to prove to Bob that the commitment G_i has been correctly computed as defined formula without revealing y_i in PK_1.

Protocol Private-Euclid
Input: $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in Z_N^n$.

1. (Registration) Alice sends to Bob two sequences of commitments, $E_1, \ldots, E_n$, and $\tilde{E}_1, \ldots, \tilde{E}_n$, where
$$E_i = E(x_i, r_i), \qquad \tilde{E}_i = E(x_i^2, \tilde{r}_i)$$
r_i and $\tilde{r}_i$ are random values chosen from Z_N, for $i = 1, \ldots, n$. Alice proves to Bob that she knows the corresponding input $\boldsymbol{x}$ committed by E_i, $\tilde{E}_i$ is the commitment to x_i squared in zero-knowledge proof
$$PK_3 = PK\left\{x_i, r_i \mid E_i = E(x_i, r_i) \wedge \tilde{E}_i = E_i^{x_i}\right\}.$$
Bob adds the commitments to template database if PK_3 is valid.
2. (Authentication) Alice computes three sequences of commitments with respect to her scanned input $\boldsymbol{y} = (y_1, \ldots, y_n)$,
$$F_i = E(y_i, r_i'), \quad \tilde{F}_i = E(y_i^2, \tilde{r}_i'), \quad \text{and } G_i = E_i^{y_i},$$
for $i = 1, \ldots, n$. Where, r_i' and $\tilde{r}_i'$ are random values. She proves to Bob that the commitments have been properly computed in the zero-knowledge proof
$$PK_4 = PK\left\{y_i, r_i' \mid F_i = E(y_i, r_i'), \wedge \tilde{F}_i = F_i^{y_i} \wedge G_i = E_i^{y_i}\right\}$$
for $i = 1, \ldots, n$.
3. Bob computes $D_E = \prod_{i=1}^n \tilde{E}_i \tilde{F}_i / G_i^2$ if PK_4 is valid for all i.
4. Alice computes the Euclidean distance $d_E = d(\boldsymbol{x}, \boldsymbol{y})$ and proves to Bob that she knows $\boldsymbol{y}$ that nearly equals to $\boldsymbol{x}$ in a sense of Euclidean distance, using the zero-knowledge proof
$$PK_5 = PK\left\{d_E, R_E \mid D_E = E(d_E, R_E) \wedge d_E \in [0, \tau_2]\right\},$$
where $R_E = \sum_{i=1}^n r_i x_i - 2 r_i y_i + \tilde{r}_i'$.
5. Bob authenticates Alice if he verifies PK_5.

Fig. 2. Protocol for Euclidean Distance evaluation

At the end of the protocol, using the Boudot's range proof[9] and the conjunctive proof of knowledge[11] (ZK_2), she can finally convince Bob that she has valid input $\boldsymbol{y}$ such that the similarities $d_C = c(\boldsymbol{x}, \boldsymbol{y})$ is less than pre-determined threshold τ_1, which means that Alice is surely a legitimate user.

3.3 Private Euclidean Distance Evaluation

Figure 2 shows the protocol Private-Euclid for proving Alice's private identity $\boldsymbol{y}$ is within the distance τ_2 from registered $\boldsymbol{x}$. In addition to the protocol Private-Cosine, it requires Alice to commit $\boldsymbol{x}$ squared as $\tilde{E}_i$ at the registration step. Implicitly, we use notation X for the commitment to x, and $\tilde{X}$ for the commitment to x^2 in the figure.

Table 1. Experiment Environment

item	values
fingerprint scanner	Digital Persona U. are .U4000 Digital Persona Gold SDK 2.5.0
fingerprint images resolution	50 genuine and 450 imposter images 300×300 [pixel]
image processing software platform	NIST NFIS2[3] proprietary application with Java version 1.5.0_06, Windows XP, 1.00 GHz, 512 MB

The additive homomorphic property allows us to privately evaluate the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ at side of Bob, at Step 3, as follows,

$$\begin{aligned} D_E &= \prod_{i=1}^{n} \tilde{E}_i \tilde{F}_i / G_i^2 = \prod_{i=1}^{n} E(x_i^2, r_i x_i) E(y_i^2, \tilde{r}_i') / E(2x_i y_i, 2r_i y_i) \\ &= \prod_{i=1}^{n} E(x_i^2 + y_i^2 - 2x_i y_i, r_i x_i + \tilde{r}_i' - 2r_i y_i) \\ &= E(\sum_{i=1}^{n} x_i^2 - 2x_i y_i + y_i^2, R_E) = E(||\boldsymbol{x} - \boldsymbol{y}||^2, R_E), \end{aligned}$$

letting R_E be a constant defined as $\sum_{i=1}^{n} r_i x_i - 2r_i y_i + \tilde{r}_i'$. For constructing zero-knowledge protocols PK_3, PK_4 and PK_5, we add a protocol proving that a committed number is squared number, presented in [9]. If all proofs are valid, Bob convince that Alice is a legitimate user who has registered $\boldsymbol{x}$ and hence is able to show the correctly computed commitment of $||\boldsymbol{x} - \boldsymbol{y}||^2$ less than threshold τ_2.

4 Evaluation

Most zero-knowledge protocols are designed to be secure in the cost of communicational and computational overhead, which are not often considered as significant. There are a trade-off between performance and security, e.g., reducing a probability being impersonated by half requires double amount of bits to be computed. In addition, we claim that there is one more trade-off between accuracy and performance in secure biometric authentication. The accuracy (and the performance) depends on a function for similarity to be evaluated in zero-knowledge protocol. Hence, it is not trivial to identify the optimal function of similarity for the multiple objective requirements involved each other.

4.1 Feature Vector

To compare two similarities, we performed some experiments using actual fingerprint images under the environment listed in Table 1. More than 500 live

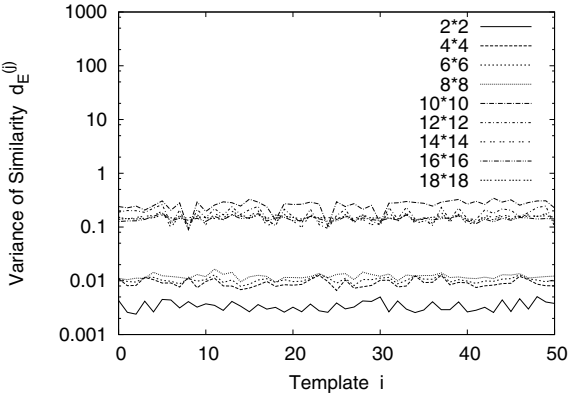


Fig. 3. Variance of Similarity d_E

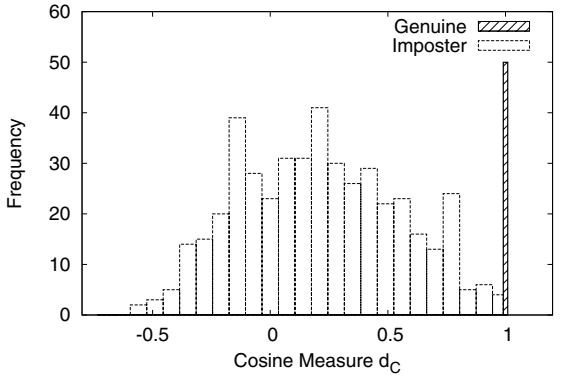


Fig. 4. Histogram of Cosine Correlations $c(\boldsymbol{a}, \boldsymbol{b})$

fingerprints are scanned and performed some sorts of image processing and extraction algorithms, which yield the feature vectors, called *ridge-valley orientation*.

The feature vector consists of a 18×18 matrix of orientations of ridges and valleys of the surface of finger, taking average for each local 16×16-pixel image. The ridge-valley orientation is quite stable against a transformation of images, thus good for the evaluation of similarities of high-dimension vectors. While, it needs to deal with empty portions of image caused by miss-scanning. To avoid some elements of feature from being zero, we take $n = L^2$ elements from the core of the 18×18 matrix. The accuracy of authentication depends on dimension n of the feature, and hence the optimal dimension is a significant issue. Figure 3 shows the variance of similarities (Euclidean distance) of two fingerprint images with respect to dimensions $n = 2 \times 2, 4 \times 4, \ldots, 18 \times 18$. From the observation of the result, we see that $n > 10 \times 10$ provides a good enough similarities to distinguish two images.

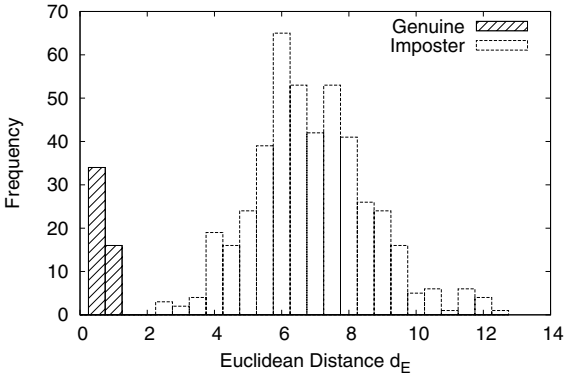


Fig. 5. Histogram of Euclidean Distances $d(\boldsymbol{a}, \boldsymbol{b})$

Figure 4 shows two distributions of cosine correlations; one between genuine and imposter images (labeled as "Imposter"), and the other one between two distinct images chosen from genuine images (as "Genuine"). The dimension of feature vector is $n = 8^2$. The genuine images are distributed within a narrow area of range, while the distribution of imposter images is broad. These distributions look almost disjoint, that is, the classification hardly ever fails.

The Euclidean distances of two feature vectors are distributed as well, shown in Figure 5. In comparison of two similarities, the distribution of genuine images is quite separate from that of imposter images in the Euclidean distance, while these are distributed closely in cosine correlations. Therefore, the accuracy of Euclidean distance is likely to be better than that of the cosine correlation.

4.2 Accuracy

We show the accuracy of authentication schemes based on the similarities in Figure 6, where overall accuracy is given as *Equal Error Rate (ERR)* of $n = 12^2$ feature vectors with respects to thresholds τ_1 and τ_2. An ERR is the rate at which both accept and reject errors are equal. Obviously, the experiment means that the Euclidean distance is superior in accuracy to the cosine correlation for all dimensions n. The result is compatible with the analysis of distributions studied in the above section.

Figure 7 show the Relative Operating Characteristic plot (ROC) for particular dimension $n = 18^2$, illustrating the change of False Rejection Rate (FRR) with respects to False Acceptance Rate (FAR). We observe that the tradeoff between these rates by varying thresholds, and the cosine correlation has higher error rate than the Euclidean distance. After all, the Euclidean distance is the better similarity measure than the cosine correlation in terms of dimensions and thresholds.

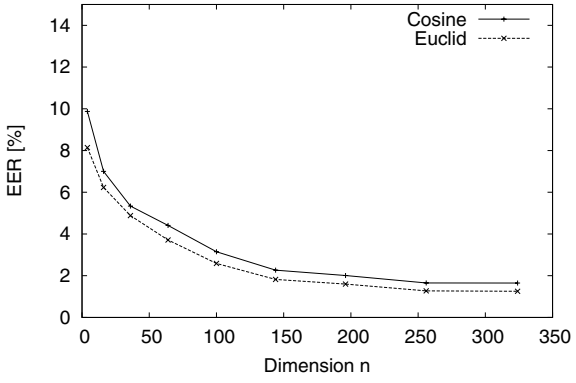


Fig. 6. Equal Error Rates (ERRs) with respect to dimension n

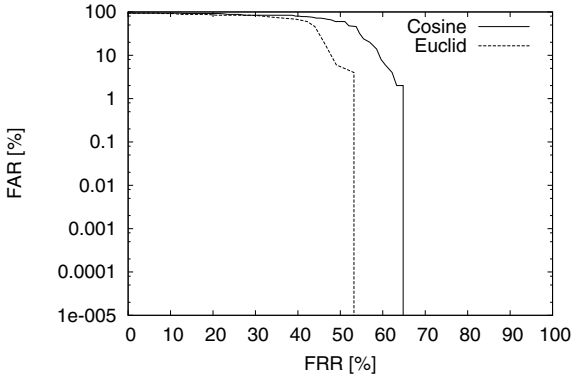


Fig. 7. Relative Operating Characteristic plot (ROC) for $n = 18^2$-dimension future vector

4.3 Performance

There are two factors for performance of protocols; a computational cost and a communication cost. The former is estimated as a number of modular exponentiations, which is the dominant factor of processing time, for each step in zero-knowledge proof. The latter is the function taking dimension n and size of modulus $\ell = |N|$, typically $\ell = 1024$ bit. We summarise the estimation of both costs in Table 2. The estimation shows that the Euclidean distance requires about double $(5/2n\ell)$ overhead to evaluate n-dimension vectors consisting of ℓ-bit integers, in theory.

In addition to the estimation from equations, we measure the processing time based on sample implementation of the protocols. Figure 8 shows the experimental result, where the constant is $\ell = |N| = 1024$, security parameter in zero-knowledge protocol is $t = 160$ bit, and dimension of feature vector ranges

Table 2. Estimation of costs for two protocol

		Private-Cosine	*Private-Euclid*
Computation	1. PK_1/PK_4	$3n$	$11n$
	3. PK_2/PK_5	19×2	19×2
	total	$3n + 38$	$11n + 38$
Communication	1. $G/G, \tilde{F}$	$n\ell$	$2n\ell$
	$PK1/PK4$	$n\ell$	$3n\ell$
	3. PK_2/PK_5	$5\ell \times 2$	$5\ell \times 2$
	total	$\ell(2n + 10)$	$\ell(5n + 10)$

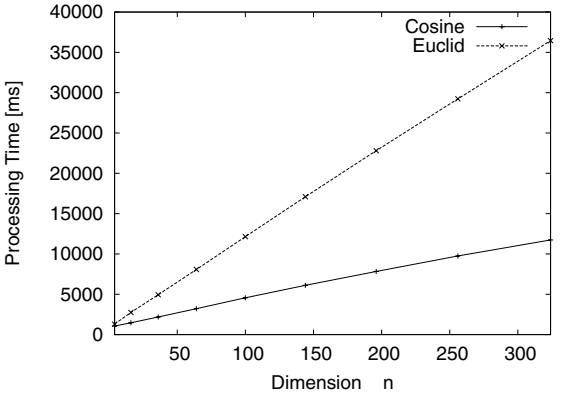


Fig. 8. Processing times for evaluating protocols with respect to dimension n

from 2×2 to 18×18. We confirm that the estimation is compatible with the experimental result. Note that there is a constant amount of time at $n = 0$, which means the overhead caused at PK_2 and PK_5. In typical setting, say $n = 8^2$, protocol Private-cosine and Private-Euclid take $3,218$ and $8,078$ [ms], respectively.

4.4 Security

The security of the proposed protocols are based on the security of the strong RSA assumption, the difficulty of the decision Diffie-Hellman problem in the random oracle model. The probability to forge the commitments in PKs can be negligible as the security parameter increases. On the other hand, the common biometric feature has a less entropy than the commitment scheme. The probability of malicious party to impersonate someone without his biometric feature is fixed at a level determined by the entropy of the feature. Hence, the zero-knowledge protocol is secure enough to apply the biometric authentication.

Our model makes an assumption of tamper-freeness of secure device that stores the template feature vector with the random values used for commitment. We consider the assumption is reasonable in practical perspective since many

secure devices are widely used in our daily life, e.g., the RFID and the smart cards. The requirement of secure device, however, is not useful from the usability point of view.

5 Conclusions

We have studied the protocols for secure similarity evaluation of vectors, Private-Cosine and Private-Euclid, based on the zero-knowledge proof of range. The Private-Cosine allows a user to convince a server that the user has a secret similar to the data stored at server in a sense of the cosine correlation, while protocol Private-Euclid uses the Euclidean distance to evaluate similarity. The latter archives better accuracy in terms of false acceptance and rejection than the former, in the cost of computational overhead. Our schemes are designed for secure remote biometric authentication that no malicious party including even server administrator can reveal private biometric data.

References

1. Chan, A., Frankel, Y., Tsiounis, Y.: Easy Come – Easy Go Divisible Cash. In: Price, W.L., Chaum, D. (eds.) EUROCRYPT 1987. LNCS, vol. 304, pp. 127–141. Springer, Heidelberg (1988)
2. Fujisaki, E., Okamoto, T.: Statistical Zero Knowledge Protocols to Prove Modular Polynomial Relations. In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 16–30. Springer, Heidelberg (1997)
3. Nist Fingerprint Image Software 2 (NFIS2), http://fingerprint.nist.gov/NFIS/
4. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal 40(3) (2001)
5. Jeong, M.Y.: Changeable Biometrics for Appearance Based Face Recognition. In: Proc. of Biometric Symposium, Biometric Consortium Conference, Baltimore (September 2006)
6. Juels, A., Sudan, M.: A Fuzzy Vault Scheme. In: Lapidoth, A., Teletar, E. (eds.) Proc. IEEE Int'l. Symp. Information Theory, p. 408 (2002)
7. Clancy, T.C., Kiyavash, N.: Secure Smartcard–Based Fingerprint Authentication. In: Proc. ACM SIGMM 2003 Multim. Biom. Met. App., pp. 45–52 (2003)
8. Juels, A., Wattenberg, M.: A Fuzzy Commitment Scheme. In: Tsudik, G. (ed.) Sixth ACM Conf. Computer and Comm. Security, pp. 28–36 (1999)
9. Boudot, F.: Efficient proofs that a committed number lies in an interval. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 431–444. Springer, Heidelberg (2000)
10. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer Science + Business Media, Heidelberg (2003)
11. Cramer, R., Damgård, I., Schoenmakers, B.: Proofs of partial knowledge and simplified design of witness hiding protocols. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 174–187. Springer, Heidelberg (1994)
12. Uludag, U., Jain, A.K.: Fuzzy Fingerprint Vault. In: Proc. Workshop: Biometrics: Challenges Arising from Theory to Practice, pp. 13–16 (2004)

Suitability Maps Based on the LSP Method

Jozo J. Dujmović[1], Guy De Tré [2], and Nico Van de Weghe [3]

[1] Department of Computer Science, San Francisco State University,
1600 Holloway Ave, San Francisco, CA 94132, U.S.A.
jozo@sfsu.edu
[2] Department of Telecommunications and Information Processing
Ghent University, St.-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
Guy.DeTre@UGent.be
[3] Department of Geography, Ghent University,
Krijgslaan 281 (S8), B-9000 Ghent, Belgium
Nico.VandeWeghe@UGent.be

Abstract. In this paper we propose the concept of logically aggregated geographic *suitability maps* (S-maps). The goal of S-maps is to provide specialized maps of the suitability degree of a selected geographic region for a specific purpose. There is a wide spectrum of purposes which include suitability for industrial development, agriculture, housing, education, recreation, etc. Our goals are to specify main concepts of S-maps development, and to identify some of the potential application areas. Our approach is based on soft computing with partial truth and graded logic functions within the framework of the LSP method.

1 Introduction

Traditional geographic maps are defined as the distribution of selected scalar indicators in the two dimensional space. Such indicators include altitude, cities, roads, airports, rivers, etc. However, each (X,Y) point has many other attributes that may be of interest for complex planning and decision making. Let the array of point attributes be $(a_1, a_2, ..., a_n)$. Such attributes may characterize *physical characteristics of terrain* (slope, altitude, material, distance from major roads, distance from green areas, distance from lakes, etc.), *available infrastructure* (supply of water, supply of electrical energy, sewage system, telecommunications, transport systems, etc.), *urban characteristics* (distance from major schools, shopping areas, entertainment, sport facilities, hospitals, the density of population, etc.), *legal status* (private property, governmental property, areas reserved for special activities), *economic development* (local industries, businesses, employability), *pollution* (air, water, noise), etc. All these attributes affect the overall suitability of a specific area for a selected type of use. In a general case the degree of suitability depends on a variety of logic conditions that evaluators specify using reasoning techniques that are typical for soft computing.

The S-map is defined as a spatial distribution of the overall degree of suitability for a specific use. Typical examples of such use are construction of industrial objects, homes, hospitals, schools, recreation areas, entertainment centers, sport facilities,

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 15–25, 2008.

shopping centers, airports, etc. In all cases decision makers are interested to evaluate and compare locations or regions from the standpoint of their suitability for a selected use. The degree of suitability E is a soft computing logical function of n attributes and we assume that its range is normalized: $E = G(a_1, a_2, ..., a_n) \in [0,1]$. The value 0 denotes an unsuitable location and the value 1 (or 100%) denotes the maximum level of suitability.

Our concept is similar to the land-use suitability maps proposed in [10]. However, the land-use suitability maps are based on outranking methods and do not support flexible logic conditions that we consider fundamental for justifiable decision making. Similar to [10] and [2], our approach is a step towards dynamic generation of specialized maps based on multicriteria decision models.

The predecessors of S-maps are composed using map algebra. Map algebra is a set based algebra for manipulating geographic data [15], or some of its generalizations (e.g. [1]) or extensions (e.g. with the temporal dimension [7], [12]). Notwithstanding that map algebra is recognized as one of the most dominant frameworks to handle GIS-based raster data [11], alternatives (e.g. [9] and [2]) have been proposed, all having their pros and contras.

While traditional maps are always produced having in mind the needs and interest of specific users, there is a clear need for specialized composite indicators that can be dynamically generated in a flexible way from geographic databases to provide information necessary for advanced public and professional decision making related to urban planning, industrial development, corporate planning, etc. In particular, there is a need for soft computing suitability maps that show suitability indicators based on flexible suitability criteria that include sophisticated logic conditions. The purpose of this paper is to propose a method for designing such maps using the LSP method for evaluating suitability. We first present the concept of S-maps and a numerical case study of their application. Then we discuss the issues of providing accurate input data.

2 Design of S-Maps

The proposed technique for creating S-maps is summarized in Fig. 1. The investigated area is divided in an orthogonal grid of square cells of size h where X,Y denote the coordinates of the center of a specific cell. Each analyzed cell is characterized by an array of n cell attributes $(a_1(X,Y), a_2(X,Y), ..., a_n(X,Y))$. The attributes are indicators that affect the ability of the analyzed cell to support some desired activity. For simplicity, the array of attributes can be denoted $(a_1, a_2, ..., a_n)$, and we assume that each attribute is a function of coordinates X,Y.

The array of attributes provides inputs for the quantitative evaluation process based on the LSP method [4,5]. After defining a complete and nonredundant list of input attributes, the next step in this process is to provide elementary attribute criteria for each component of the array of attributes. The elementary criteria are functions $g_i : R \to [0,1],\ i = 1, ..., n$. The value $e_i = g_i(a_i)$ is called the attribute (or

elementary) preference. The attribute preference denotes the degree to which the value a_i satisfies a specific requirement that reflects the selected type of evaluation.

The final step in the organization of the LSP criterion function is the development of the preference aggregation structure that logically aggregates all attribute preferences and generates the resulting overall preference that is the degree of suitability $E(X,Y) \in [0,1]$. The aggregation process can include a variety of logic conditions that are modeled using the Generalized Conjunction/Disjunction function [6] and more complex compound aggregators [5]. A classification of fundamental aggregators in the Continuous Preference Logic (CPL) is shown in Table 1.

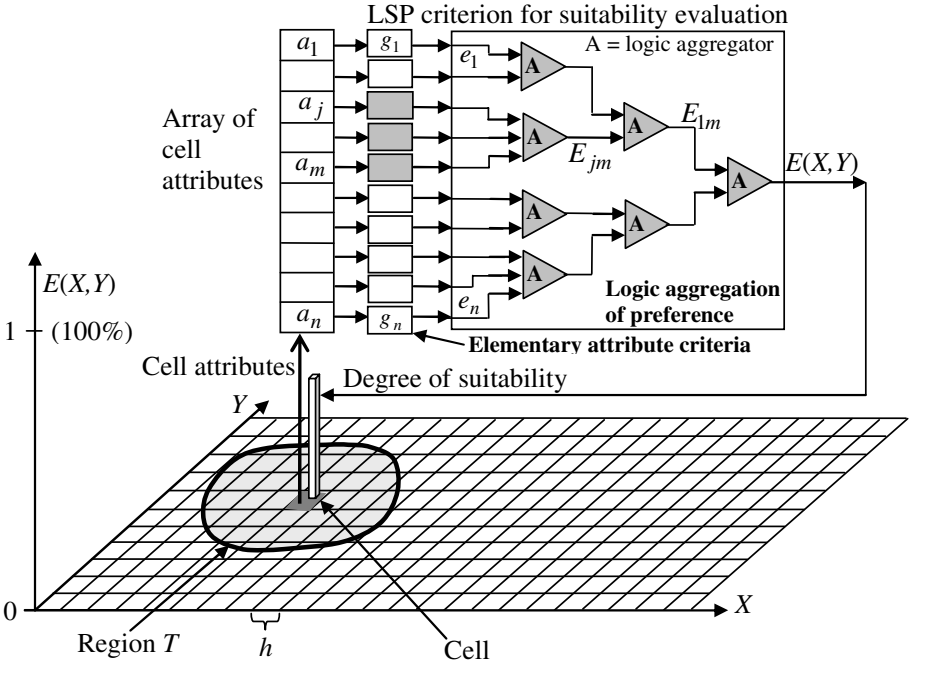


Fig. 1. The concept of S-maps

The value $E(X,Y)$ reflects the suitability of a given X,Y cell, and the distribution defined by $E(X,Y)$, $X_{\min} \le X \le X_{\max}$, $Y_{\min} \le Y \le Y_{\max}$ represents the desired S-map. The average suitability of a region T can then be computed by averaging $E(X,Y)$ over the desired region:

$$\bar{E} = \frac{\iint\limits_{T} E(X,Y)\,dX\,dY}{\iint\limits_{T} dX\,dY}$$

Table 1. Classification of fundamental CPL aggregators [6]

Aggregation Operators in Continuous Preference Logic	Basic CPL Aggregator	Full disjunction (D)	
	Generalized Conjunction/ Disjunction (GCD)	Partial disjunction	Hard partial disjunction (HPD)
			Soft partial disjunction (SPD)
		Neutral aggregator	Arithmetic mean (A)
		Partial disjunction	Soft partial conjunction (SPC)
			Hard partial conjunction (HPC)
		Full conjunction (C)	
	Compound aggregators	Simple partial absorption	Disjunctive partial absorption (DPA)
			Conjunctive partial absorption (CPA)
		Nested partial absorption	Sufficient/Desired/Optional (SDO)
			Mandatory/Desired/Optional (MDO)
		Partial equivalence, partial implication, etc.	

The mean suitability $\overline{E}$ is a useful indicator only if the distribution $E(X,Y)$ satisfies some acceptability criteria, primarily a sufficient smoothness and a low variability. For example if $E(X,Y)$ in a region *T* shows discontinuities and large variations, this can prevent some applications regardless the value of $\overline{E}$ for *T*.

It is important to note that S-maps can be dynamically generated from the database of attributes. They are flexible because the formal logic and semantic parameters that evaluate the suitability of cells can be interactively modified and adjusted by the user. By modifying the parameters the user can generate a sequence of maps that answer a variety of "what-if" questions. These answers are the primary purpose of S-maps.

3 Evaluation of Urban Expansion Suitability

The suitability for urban expansion is one of frequent and complex evaluation problems. In this section we evaluate the degree of urban expansion suitability as a case study that illustrates our method. The first step in this direction is to develop a system attribute tree that includes all attributes that will be evaluated. A simplified tree with 11 attributes is shown in Fig. 2.

Some of the attributes are considered mandatory, i.e. if they are not satisfied then the overall suitability for urban expansion is considered unacceptable and rated zero. Mandatory attributes are in Fig. 2 denoted by (+). On the other hand, there are attributes that are in our example considered nonmandatory and denoted by (-). If a nonmandatory requirement is not satisfied that will not cause rejection of the proposed location. For example, while appropriate slope and altitude are considered mandatory requirements, a good orientation of the new urban complex is considered desirable, but it is not mandatory. Similarly, good environment is highly desirable but not necessary: if other conditions are satisfied new urban areas can be built in cases where green areas and lakes are missing. Finally, the proximity to an airport is also considered nonmandatory. However, good ground transportation is considered mandatory.

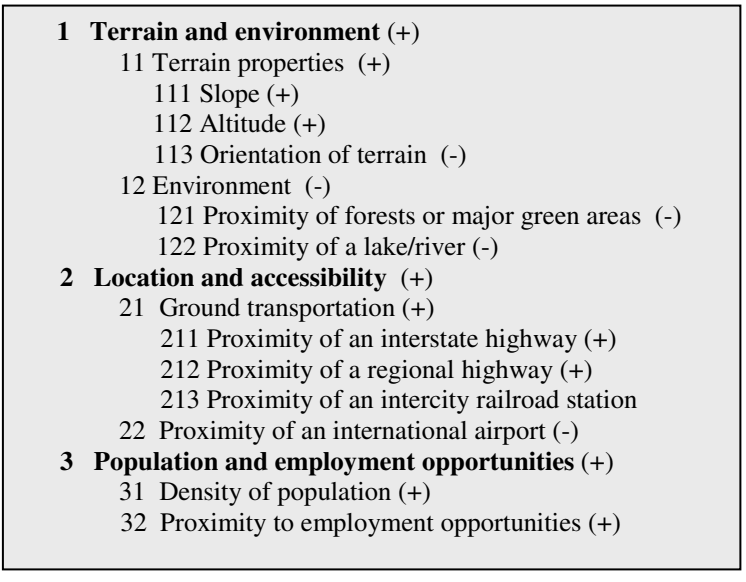


Fig. 2. System attribute tree with mandatory (+) and nonmandatory (-) components

Mandatory and nonmandatory attributes are examples of logic conditions that are present in all areas of evaluation. Additional logic conditions include the adjustable levels of simultaneity (andness) and replaceability (orness) that some groups of attributes must satisfy. Finally, all attributes are assumed to have adjustable levels of relative importance. This is the reason why it is convenient to realize the evaluation model using the LSP method.

The next step in the LSP method is to specify requirements for the above 11 attributes. The requirements are specified as functions that show the level of satisfaction with each value of the attribute. The elementary attribute criteria are shown in Fig. 3. For example, the criterion #212 specifies that it is desired that an urban complex is located in proximity of a regional highway. More precisely, the proposed elementary criterion considers that an ideal distance from the regional highway is from 100 to 200 meters. If the distance is greater than 2000 meters or less than 25 meters that is considered unacceptable.

Elementary criteria are based on piecewise linear (polygonal) approximations of functions: we define a set of justifiable breakpoints and use linear interpolation between them. This approach yields a good combination of simplicity and accuracy. For example, we expect that the distance to the railroad station is less that 15 minutes and not greater than 30 minutes; if the distance is 20 minutes we use interpolation and the resulting degree of satisfaction of this requirement will be 66.7%.

The aggregation of attribute preferences is presented in Fig. 4. Each circle has a reference number and denotes the weighted power mean aggregator $e_{out} = (W_1 e_1^r + ... + W_k e_k^r)^{1/r}$; input lines denote weights $W_1, ..., W_k$, and exponents r for aggregators A, C--, C-+, CA, and C+ are respectively 1, 0.619, -0.148, -0.72,

and -3.51 [4]. The aggregation structure includes mostly conjunctive aggregators that reflect requirements for simultaneous satisfaction of requirements. Three aggregators (identified in Fig. 4 by block numbers 11, 1, and 2) are asymmetrical partial absorptions that aggregate mandatory and desired inputs. If the desired input is 0, this causes a penalty (the average decrement of the output value) *P,* and if the desired input is 1, this causes a reward (the average increment of the output value) *R*.

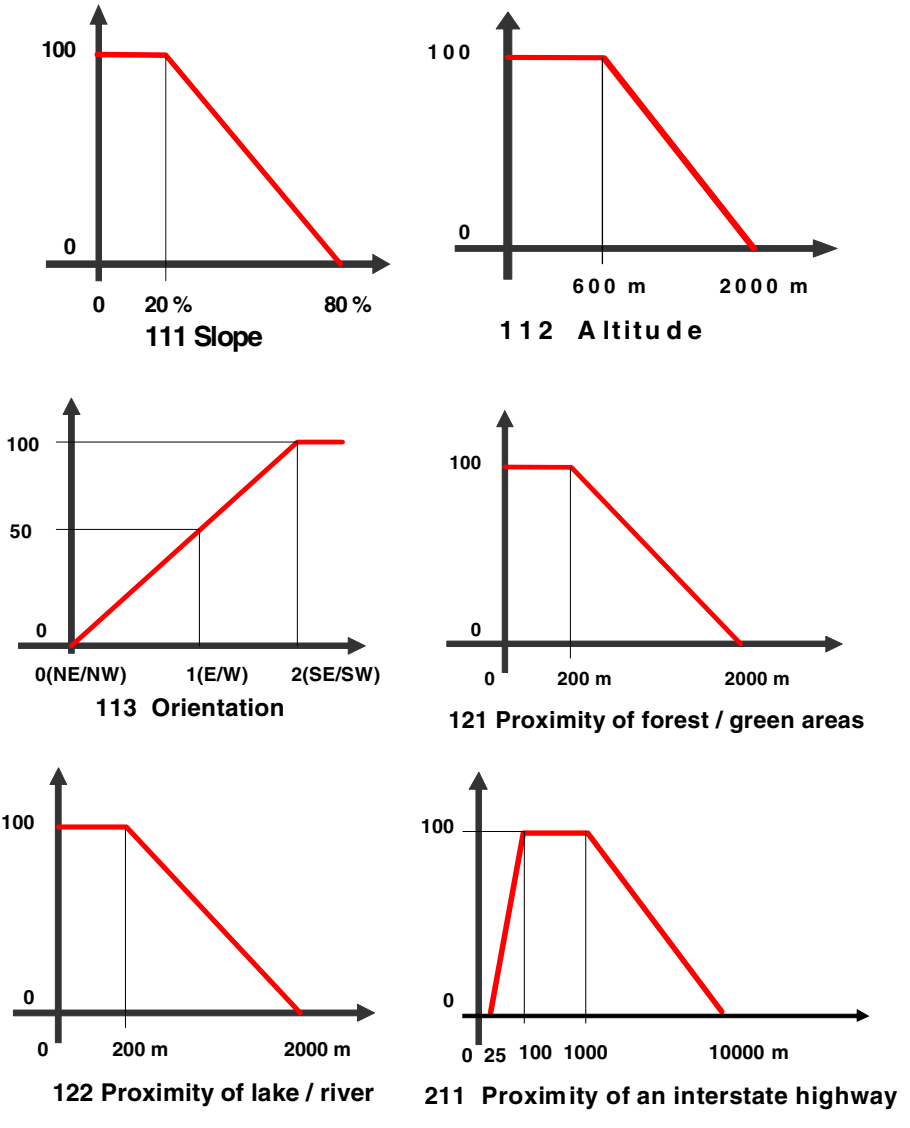
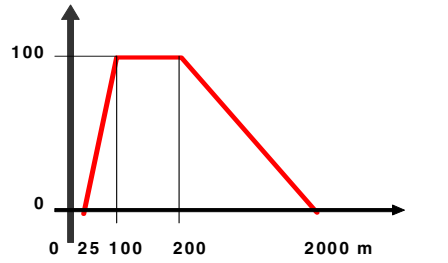
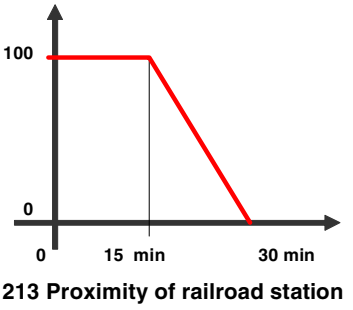
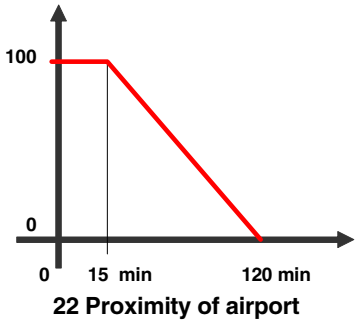


Fig. 3. Elementary attribute criteria







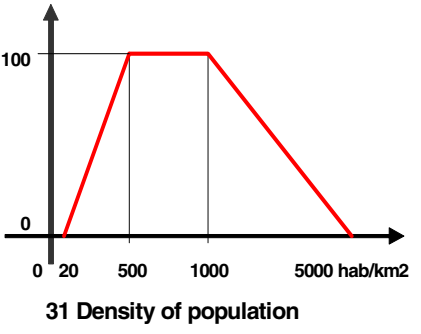


Fig. 3. *(continued)*

In the case of asymmetrical partial absorption the parameters of aggregator (the weights and the andness of partial conjunction) can be computed from the desired penalty/reward pairs. Therefore, the evaluator selects only the most appropriate levels of penalty and reward. In the case of other aggregators (generalized conjunction/disjunction [5], [6]) the evaluator selects weights that express the desired relative importance of inputs and the andness/orness that reflects the desired level of simultaneity or replaceability of inputs. Another way to determine parameters is to use software tools that compute the parameters from a training set of desired input-output pairs.

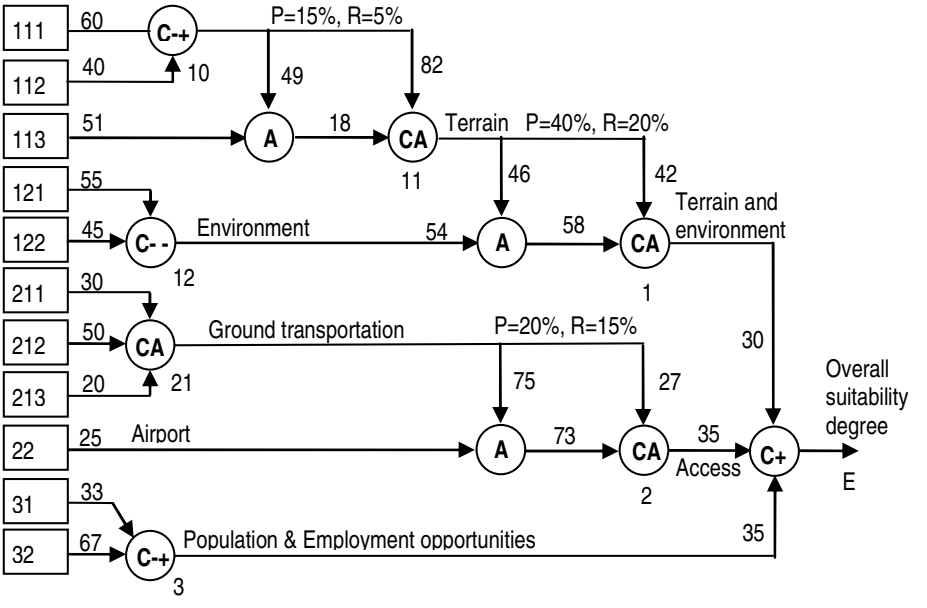


Fig. 4. The aggregation of preferences and the computation of the overall suitability

For simplicity, let us compare three locations: $L_1 = (X_1, Y_1)$, $L_2 = (X_2, Y_2)$, and $L_3 = (X_3, Y_3)$. These locations are selected as typical for three areas that are candidates for urban expansion. Their attributes are shown in Table 2. Because of differences in the available infrastructure, the cost of building in location L_2 is 30% more expensive than building in location L_1, and building in location L_3 is 20% more expensive then building in location L_1. The problem is to find which location is the most suitable for the urban expansion, and the basic results are shown in Table 3.

Table 2. Input attributes and costs for the three competitive locations

Loc	111	112	113	121	122	211	212	213	22	31	32	C
L_1	40	1200	2	3000	250	1500	50	20	100	555	30	1
L_2	18	400	1	150	500	1100	300	20	20	800	20	1.3
L_3	35	700	0	1600	700	1700	400	15	35	1500	35	1.2

Table 3. Resulting suitability degrees [%] and cost/preference indicators

Loc	10	11	12	21	1	2	3	E	E/C
L_1	62.7	65.5	26.8	48.5	51.7	42.9	86	51.75	51.75 [72.5%]
L_2	100	94.4	92.4	88.6	93.8	89.8	95.5	92.83	71.41 [100%]
L_3	81.6	69.7	42	92	60.3	89.9	77.7	72.63	60.52 [84.8%]

The most suitable location is L_2 because it satisfies almost 93% of the suitability requirements. The location L_3 is second; it satisfies 73% of the requirements. The least suitable location is L_1; it satisfies only 52% of the suitability requirements for urban expansion. If the importance of high suitability is the same as the importance of low cost then we can compare the locations using the E/C ratio, and in such a case the location L_2 is still the most convenient regardless the highest cost; the overall suitability of L_3 is approximately 15% lower than the suitability of L_2.

The presented example shows a suitability criterion that incorporates elementary attribute criteria and a number of logic conditions: adjustable andness and orness, adjustable relative importance, symmetric and asymmetric (mandatory/desired) logic conditions. Such criteria can be used for a variety of experiments with other decision requirements.

4 Data Availability and Reliability Problems

S-maps are based on assumption that accurate values are available for all attributes in the array $(a_1(X,Y),a_2(X,Y),...,a_n(X,Y))$ and for each point (X,Y) in the analyzed region. There are applications where this assumption holds, but there are also real life applications where this is not the case. Information sources might be imperfect, containing inaccurate, incomplete, or even inconsistent data. For some attributes, like distances from given map objects, we would be able to directly compute a reliable attribute value (on condition that the map is reliable). In a general case, however, we must be prepared to face situations where the necessary attribute values are not available. Several types of unavailability have been identified in [13], including *the incompleteness of data*, *the lack of sufficiently accurate data*, and *the nonexistent data*.

Handling the incompleteness of data. If an attribute value $a(X,Y)$ for a given point (X,Y) is missing, but values are available for relevant neighboring points that are close enough, then we might be able to derive an approximate value by aggregating the corresponding attribute values of these neighboring points. The problem of deriving a reliable value can thus be decomposed in three subproblems: search for relevant neighboring points, determine whether these are close enough or not, and interpolate the values.

To search for relevant neighboring points, a Triangular Irregular Network (TIN) [14] is constructed with the points for which data is available as vertices. The well-known Delaunay triangulation method [3] is used for this purpose. It is then straightforward to determine the triangle in which the point (X,Y) is located; the relevant neighboring points are the vertices of this triangle. After the triangle containing the point (X,Y) is identified the user must decide whether the accuracy of the applied interpolation method is satisfactory. This must be done in a general way and not for each point separately. In the case of unacceptable accuracy we consider data to be unknown.

In cases where we are interested in interpolation of geologic data it is convenient to use a family of nonlinear least squares estimation algorithms developed in geostatistics, primarily various forms of kriging [8,16].

Handling the lack of sufficiently accurate data. In the case of unknown data we may apply one of the following three approaches: (1) Develop a modified suitability criterion excluding unknown data, (2) Perform the analysis replacing the unknown value by a range of appropriate values or a distribution, and (3) Develop a method to modify the preference aggregation structure in an automatic way.

The case of nonexistent data. Another potential type of unavailable data is the case where data is not available because it does not exist. This means that the corresponding criterion is not applicable for the analyzed location. This is a sufficient indication that the existing suitability criterion must be redesigned.

5 Conclusions

S-maps are specialized geographic maps based on aggregating a number of attribute preferences that characterize the suitability of a geographic location for a specific use. Advantages of S-maps can be summarized as follows:

- S-maps are general and flexible in the sense that they can express the suitability of the analyzed geographic area for any specific use.
- The method of generating S-maps offers a high level of logic versatility originating from the LSP-based soft computing approach. It is easily understandable and consistent with observable properties of human reasoning in the area of evaluation.
- LSP models of suitability generate correct logic results in all points of the attribute space. The accuracy of such models cannot be reduced by unpredictable variations of attribute values. Therefore, the expected reliability of S-maps is very good.
- S-maps are dynamically generated from the database of attributes.
- Users of S-maps can experiment with various suitability criteria and dynamically investigate effects of changing their parameters.
- As versatile on-line tools, S-maps have potential of becoming an indispensable decision support means in many social, engineering, and business activities.

S-maps create various opportunities for future work. The initial efforts should be focused on improving the availability and reliability of input attribute data. There is also space for improving methods for working with incomplete and imprecise attributes. Finally, it is also necessary to develop appropriate software infrastructure that will facilitate the routine creation and experimental use of S-maps.

References

1. Camara, G., Palomo, G., de Souza, R., de Oliveira, O.: Towards a Generalized Map Algebra: Principles and Data Types, GeoInfo., pp. 66–81 (2005)
2. Chakhar, S., Mousseaua, V.: An Algebra for Multicriteria Spatial Modeling. Computers, Environment and Urban Systems 31(5), 572–596 (2007)

3. Delaunay, B.: Sur la sphère vide. Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk 7, 793–800 (1934)
4. Dujmović, J.J., Nagashima, H.: LSP Method and its Use for Evaluation of Java IDE's. International J. Approx. Reas. 41(1), 3–22 (2006)
5. Dujmović, J.J., Preference Logic, J.J.: for System Evaluation. IEEE Transactions on Fuzzy Systems 15(6), 1082–1099 (2007)
6. Dujmović, J.J.: Characteristic Forms of Generalized Conjunction/Disjunction. In: Proceedings of the IEEE World Congress on Computational Intelligence, Hong Kong (June 2008)
7. Frank, A.U.: Map Algebra Extended with Functors for Temporal Data. In: Akoka, J., Liddle, S.W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., Mayr, H.C. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 194–207. Springer, Heidelberg (2005)
8. Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University Press, New York (1997)
9. Haklay, M.: Map Calculus in GIS: A Proposal and Demonstration. International Journal of Geographical Information Science (IJGIS) 18(1), 107–125 (2004)
10. Joerin, F., Thériault, M., Musy, A.: Using GIS and outranking multicriteria analysis for land-use suitability assessment. International Journal of Geographical Information Science (IJGIS) 15(2), 153–174 (2001)
11. Longley, P., Goodchild, M., Maguire, D., Rhind, D.: Geographic Information Systems and Science. John Wiley & Sons, Ltd., Chichester (2001)
12. Mennis, J., Viger, R., Tomlin, C.: Cubic Map Algebra Functions for Spatio-Temporal Analysis. Cartography and Geographical Information Science 32(1), 17–32 (2005)
13. Motro, A.: Management of Uncertainty in Database Systems. In: Kim, W. (ed.) Modern Database Systems, The object model, Interoperability and Beyond, Addison-Wesley Publishing Company, Reading (1995)
14. Rigaux, P., Scholl, M., Voisard, A.: Spatial Databases with Applications to GIS. Morgan Kaufman Publishers, San Francisco (2002)
15. Tomlin, C.: Geographical Information Systems and Cartographic Modelling. Prentice Hall, Englewood Cliffs (1990)
16. Wackernagel, H.: Multivariate Geostatistics - An Introduction with Applications. Springer, Berlin (1995)

Aggregated Mean Ratios of an Interval Induced from Aggregation Operations

Yuji Yoshida

Faculty of Economics and Business Administration, University of Kitakyushu
4-2-1 Kitagata, Kokuraminami, Kitakyushu 802-8577, Japan
yoshida@kitakyu-u.ac.jp

Abstract. This paper deals with quasi-arithmetic means of an interval through utility functions in decision making. The mean values are discussed from the viewpoint of aggregation operators, and they are given as aggregated values of each point in the interval. We investigate the properties of the quasi-arithmetic mean and its translation invariance, and next we demonstrate the decision maker's attitude based on his utility by the quasi-arithmetic mean and the aggregated mean ratio. The dual quasi-arithmetic means are also discussed with dual aggregation operators. Finally, examples of the quasi-arithmetic means and the aggregated mean ratio for various typical utility functions are given to understand the motivation.

1 Introduction

In decision making, we often use mean values with utility functions as a criterion ([7,8,16,17]). This paper deals with means of an interval through utility functions. A mean value of an interval $[a, b]$, where a and b are real numbers, is given by the middle mean $(a+b)/2$ of the both edges a and b in classical theory. However, in decision making models like artificial intelligent, we need to estimate data subjectively. How are means of an interval evaluated under decision maker's subjective utility? In this paper, we discuss means of an interval from the viewpoint of *aggregation operators*. Kolmogorov [10] and Nagumo [12] studied aggregation operators and verified under some assumptions that the aggregated value of real numbers $x_1, x_2, \cdots, x_n (x_i \in [0,1], i = 1, 2, \cdots, n)$ is represented as

$$\xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \tag{1}$$

with a continuous strictly increasing function $f : [0,1] \mapsto [0,1]$. Starting from this equation (1), we construct means of an interval step by step in Section 2. In this paper, we consider that the mean value is given by an aggregated value of all points in the interval, and we call it a *quasi-arithmetic mean*. In Sections 3 and 4, we investigate the properties of the quasi-arithmetic mean and its translation invariance. Next, in Section 5, we introduce an *aggregated mean ratio* of the quasi-arithmetic mean by an interior ratio on the interval,

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 26–37, 2008.

and we demonstrate the correspondence among the quasi-arithmetic mean, the aggregated mean ratio and the decision maker's attitude based on his utility. The decision maker's attitudes, for example neutral, risk averse and risk loving, are characterized by the quasi-arithmetic mean and the aggregated mean ratio under necessary/sufficient conditions on utility functions. Further we investigate the movement of the aggregated mean ratio in local/global regions. In Section 6, we also consider the quasi-arithmetic means induced from dual aggregation operators, and we give their characterizations. Finally, in Section 7, we show examples of the quasi-arithmetic means and the aggregated mean ratio with various typical utility functions, and we give their relations with the classical quasi-arithmetic means.

2 Aggregation Operators

We deal with quasi-arithmetic means of an interval induced from aggregation operations from the viewpoint of subjective decision making. It is well-known that the aggregation operation can be represented with a continuous increasing function (Kolmogorov [10] and Nagumo [12]). In this paper, taking the continuous increasing function as a utility function in decision making, we discuss the decision-maker's judgment by the mean based on utility. We analyze properties of the quasi-arithmetic mean. introducing a ratio concerning the mean on the interval. First, we start from the notion of aggregation operations of several variables on $[0,1]$. Next, we construct a quasi-arithmetic means of an interval step by step. Let n be a fixed natural number, and let $\xi^n : [0,1]^n \mapsto [0,1]$ be a function. We represent it as $\xi^n(x_1, x_2, \cdots, x_n)$ for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$.

Definition 1 (n-ary aggregation operator [8]). A function $\xi^n : [0,1]^n \mapsto [0,1]$ is called an *n-ary aggregation operator* if it satisfies the following conditions (A.i) – (A.iii):

(A.i) $\xi^n(x_1, x_2, \cdots, x_n) \leq \xi^n(y_1, y_2, \cdots, y_n)$ if $x_i \leq y_i$ for all $i = 1, 2, \cdots, n$.
(A.ii) ξ^n is continuous on $[0,1]^n$.
(A.iii) $\xi^n(x_1, x_2, \cdots, x_n) = \xi^n(x_{i_1}, x_{i_2}, \cdots, x_{i_n})$ if $(i_1, i_2, \cdots, i_n) = \sigma(1, 2, \cdots, n)$, where σ is a permutation operator.

We can find another definition of n-ary aggregation in [2,3], which requires (A.1) with boundary conditions and does not require (A.ii) and (A.iii). However, in this paper we introduce the n-ary aggregation by Definition 1 to discuss quasi-arithmetic means (5) with a continuous utility function f. The conditions (A.i), (A.ii) and (A.iii) are called *monotone*, *continuous* and *neutral* respectively. For an n-ary aggregation operator ξ^n, the following properties (A.iv), (A.v) and (A.vi) are said to be *idempotent*, *strictly monotone* and *decomposable* respectively:

(A.iv) $\xi^n(x, x, \cdots, x) = x$ for all $x \in [0,1]$.
(A.v) $\xi^n(x_1, x_2, \cdots, x_n) < \xi^n(y_1, y_2, \cdots, y_n)$ whenever $x_i \leq y_i$ for all $i = 1, 2, \cdots, n$ and $x_j < y_j$ for some $j = 1, 2, \cdots, n$.

(A.vi) Put $\xi_1 := \xi^1(x_1) = x_1,\ \xi_2 := \xi^2(x_1, x_2),\ \xi_3 := \xi^3(x_1, x_2, x_3),\ \cdots, \xi_n := \xi^n(x_1, x_2, \cdots, x_n)$. Then, for $k = 1, 2, \cdots, n$ it holds that

$$\xi^n(x_1, \cdots, x_k, x_{k+1}, \cdots, x_n) = \xi^n(\xi_k, \cdots, \xi_k, x_{k+1}, \cdots, x_n).$$

The following well-known result regarding the aggregation operations is given by Kolmogorov [10] and Nagumo [12].

Lemma 1 ([10],[12]). *An n-ary aggregation operator $\xi^n : [0,1]^n \mapsto [0,1]$ satisfies (A.iv), (A.v) and (A.vi) if and only if there exists a continuous strictly increasing function $f : [0,1] \mapsto [0,1]$ such that*

$$\xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \tag{2}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$.

Definition 2 (aggregation operator [2,3]). A function $\xi : \bigcup_{n\geq 1}[0,1]^n \mapsto [0,1]$ is called an *aggregation operator* if it is given by n-ary aggregation operators ξ^n such as $\xi = \xi^n$ on $[0,1]^n$ for each $n = 1, 2, \cdots$.

In Lemma 1, we note for each $n = 1, 2, \cdots$ that when any continuous strictly increasing function $f : [0,1] \mapsto [0,1]$ is given, a function ξ^n which is described by (2) satisfies (A.i) – (A.vi) and it is an n-ary aggregation operator. Therefore, in this paper, first of all we fix a continuous strictly increasing function $f : [0,1] \mapsto [0,1]$ as the decision maker's utility, and we discuss an n-ary aggregation operator $\xi^n : [0,1]^n \mapsto [0,1]$ defined by

$$\xi(x_1, x_2, \cdots, x_n) = \xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \tag{3}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$ and $n = 1, 2, \cdots$. Then, we construct a quasi-arithmetic means of an interval under subjective decision-making from the viewpoint of an aggregation of all points in the interval. Let $[a, b]$ be a closed interval satisfying $0 \leq a < b \leq 1$. Let $\{[c_{i-1}, c_i] | i = 1, 2, \cdots, n\}$ be a partition of the interval $[a, b]$ such that $c_i := a + i(b-a)/n$ for $i = 0, 1, 2, \cdots, n$. Take a point x_i on the interval $[c_{i-1}, c_i]$ such that $x_i \in [c_{i-1}, c_i]$ for each $i = 1, 2, \cdots, n$. From (3), we define a quasi-arithmetic mean on the interval $[a, b]$ as follows

$$M([a,b]) = \lim_{n\to\infty} \xi(x_1, x_2, \cdots, x_n) = \lim_{n\to\infty} f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right). \tag{4}$$

Hence we can understand the mean (4) as the aggregated value of all points distributed uniformly on $[a, b]$ through the utility function f since $\xi(x_1, x_2, \cdots, x_n)$ is an aggregated point of reference points x_i on the small interval $[c_{i-1}, c_i]$

$(i = 1, 2, \cdots, n)$ based on the aggregation operator defined by the utility function f. By the definition of Riemann integral, we obtain

$$M([a,b]) = f^{-1}\left(\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\right) \tag{5}$$

for $[a,b](\subset [0,1])$ such that $0 \le a < b \le 1$. Hence, $M([a,b])$ represents a *mean value* given by a real number $c(\in [a,b])$ satisfying $f(c) = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x$ in the *mean value theorem*. Let $D(\subset (-\infty,\infty))$ be an interval. Extending the domain from the closed interval $[0,1]$ to D, in the next section we demonstrate the *quasi-arithmetic mean* of closed subintervals of D in the form (5) for $[a,b] \subset D$ $(a < b)$, where $f : D \mapsto (-\infty,\infty)$ is a continuous strictly increasing function for utility.

3 Quasi-arithmetic Means

Let $\mathbb{R} := (-\infty,\infty)$ be the set of all real numbers. For two bounded closed intervals $[a,b]$ and $[c,d]$, we give a partial order $\preceq$ concerning intervals as follows: $[a,b] \preceq [c,d] \iff a \le c$ and $b \le d$. Let D be a fixed interval which is not a singleton and we call it a domain. Let $\mathcal{I}(D)$ be the set of all nonempty subintervals of D and let $\mathcal{C}(D)$ be the set of all nonempty bounded closed subintervals of D. Let $f : D \mapsto \mathbb{R}$ be a continuous strictly increasing function for utility. Define a map $M : \mathcal{C}(D) \mapsto D$ by

$$M([a,b]) := \begin{cases} f^{-1}\left(\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\right) & \text{if } a < b \\ \lim_{c\downarrow a} f^{-1}\left(\frac{1}{c-a}\int_a^c f(x)\,\mathrm{d}x\right) & \text{if } a = b < \sup D \\ \lim_{c\uparrow b} f^{-1}\left(\frac{1}{b-c}\int_c^b f(x)\,\mathrm{d}x\right) & \text{if } a = b = \sup D \end{cases} \tag{6}$$

for $[a,b] \in \mathcal{C}(D)$.

Lemma 2. *A quasi-arithmetic mean $M : \mathcal{C}(D) \mapsto D$ defined by (6) has the following properties (M.i) – (M.iii):*

(M.i) *Let $[a,b] \in \mathcal{C}(D)$. Then it holds that $a \le M([a,b]) \le b$. Especially, $M([a,a]) = a$ for $a \in D$.*
(M.ii) *Let $[a,b],[c,d] \in \mathcal{C}(D)$ such that $[a,b] \preceq [c,d]$. Then $M([a,b]) \le M([c,d])$.*
(M.iii) *The map $M : \mathcal{C}(D) \mapsto D$ is continuous, i.e. it holds that*

$$\lim_{n\to\infty} M([a_n,b_n]) = M([a,b])$$

for $[a,b] \in \mathcal{C}(D)$ and $[a_n,b_n] \in \mathcal{C}(D)(n = 1,2,\cdots)$ such that $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$.

Remark 1. We note the quasi-arithmetic mean (6) also has the following properties as an evaluation of intervals.

(i) The properties (M.i) – (M.iii) in Lemma 2 are corresponding to the properties of *compensative*, *monotone* and *continuous* respectively.

(ii) In Lemma 2, we may take a continuous *strictly decreasing* function when we deal with a regret function instead of the utility function f. Then, the corresponding mean is reduced to the one with a strictly decreasing function. Actually we see the following. Let $g : D \mapsto \mathbb{R}$ be a continuous *strictly decreasing* function. Define a map $M^g : \mathcal{C}(D) \mapsto D$ by

$$M^g([a,b]) := \begin{cases} g^{-1}\left(\frac{1}{b-a}\int_a^b g(x)\,\mathrm{d}x\right) & \text{if } a < b \\ \lim_{c\downarrow a} g^{-1}\left(\frac{1}{c-a}\int_a^c g(x)\,\mathrm{d}x\right) & \text{if } a = b < \sup D \\ \lim_{c\uparrow b} g^{-1}\left(\frac{1}{b-c}\int_c^b g(x)\,\mathrm{d}x\right) & \text{if } a = b = \sup D \end{cases}$$

for $[a,b] \in \mathcal{C}(D)$. Then, setting a continuous strictly increasing function $f := -g : D \mapsto \mathbb{R}$, we obtain $g^{-1}(x) = f^{-1}(-x)$ and $M^g([a,b]) = M^f([a,b])$ or $[a,b] \in \mathcal{C}(D)$ such that $a < b$. This is the same one as (6).

Finally we can extend the quasi-arithmetic mean (6) of a closed interval to be applicable for a general interval. For intervals $(a,b], [a,b), (a,b) \in \mathcal{I}(D)$, we extend the mean M by

$$\begin{aligned} M((a,b]) &:= \lim_{c\downarrow a} M([c,b]) && \text{if } (a,b] \in \mathcal{I}(D), \\ M([a,b)) &:= \lim_{c\uparrow b} M([a,c]) && \text{if } [a,b) \in \mathcal{I}(D), \\ M((a,b)) &:= \lim_{c\downarrow a}\lim_{d\uparrow b} M([c,d]) && \text{if } (a,b) \in \mathcal{I}(D). \end{aligned}$$

Then, from the definition we have $M((a,b]) = M([a,b])$ and $M((a,b)) = M([a,b))$ if $a \neq \inf D$, and $M([a,b)) = M([a,b])$ and $M((a,b)) = M((a,b])$ if $b \neq \sup D$. We can extend the mean M regarding the ends of D in similar ways.

4 The Quasi-arithmetic Means and Translation Invariance

In this section, we discuss translation invariance of the quasi-arithmetic means. Let D be a fixed domain and let $f : D \mapsto \mathbb{R}$ be a continuous strictly increasing function for utility. Let $M : \mathcal{C}(D) \mapsto D$ be the mean given by (6).

Lemma 3. *The following conditions (a) – (c) are equivalent:*

(a) $r \cdot M([a,b]) + s = M([ra+s, rb+s])$ *for all* $[a,b] \in \mathcal{C}(D), r, s \in \mathbb{R}$ $(r > 0)$.
(b) $M([a,b]) + M([c,d]) = M([a+c, b+d])$ *for all* $[a,b], [c,d] \in \mathcal{C}(D)$.
(c) *There exists a constant* $\theta^* \in [0,1]$ *such that*

$$M([a,b]) = a + \theta^*(b-a) \quad \textit{for all } [a,b] \in \mathcal{C}(D). \tag{7}$$

Now, to represent Lemma 4(i) explicitly, we introduce the following notations. For an interval $I(\subset \mathbb{R})$ and a function $h : I \mapsto \mathbb{R}$, we define $h(I) := \{h(x) \mid x \in I\}$. If $I = [a,b] \in \mathcal{C}(D)$ and h is a continuous strictly increasing function, then

it holds that $h([a,b]) = \{h(x) \mid x \in [a,b]\} = \left[\min_{x\in[a,b]} h(x), \max_{x\in[a,b]} h(x)\right] = [h(a), h(b)]$. Let a semi-linear strictly increasing function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ by $\varphi(x) := rx + s$ $(x \in \mathbb{R})$ for $r, s \in \mathbb{R}$ such that $r > 0$. Then Lemma 4(i) implies $\varphi(M([a,b])) = M([\varphi(a), \varphi(b)]) = M(\varphi([a,b]))$ for $[a,b] \in \mathcal{C}(D)$. Therefore, using this notations, Lemma 4(i) asserts that the mean M is invariant for a semi-linear strictly increasing translation φ. Namely,

$$\varphi \circ M = M \circ \varphi,$$

where $\circ$ is the composition of maps.

Remark 2. Regarding (7) of Lemma 3, we find $\theta^* = 1/2$ in Corollary 2 of the next section. Therefore, (7) follows

$$M([a,b]) = \frac{a+b}{2} \quad \text{for all } [a,b] \in \mathcal{C}(D). \tag{8}$$

The following lemma characterizes the translation invariance properties. General translation invariance is also discussed by [11,13].

Lemma 4 (translation-invariance). *The following (i) – (iv) hold.*

(i) *Put a function $f(x) = cx + d$ on a domain $D = (-\infty, \infty)$ with constants c, d such that $c > 0$. Then,*

$$r \cdot M([a,b]) + s = M([ra + s, rb + s])$$

for $[a,b] \subset (-\infty, \infty)$, $r, s \in \mathbb{R}$ such that $r > 0$.

(ii) *Let a domain $D = [0, \infty)$. Put a function $f(x) = x^\gamma$ on D with a positive constant γ. Then,*

$$r \cdot M([a,b]) = M([ra, rb])$$

for $[a,b] \subset [0, \infty)$, $r \in \mathbb{R}$ such that $r > 0$.

(iii) *Let a domain $D = (0, \infty)$. Put a function $f(x) = \gamma \log x$ on D with a positive constant γ. Then,*

$$r \cdot M([a,b]) = M([ra, rb])$$

for $[a,b] \subset (0, \infty)$, $r \in \mathbb{R}$ such that $r > 0$.

(iv) *Let a domain $D = (-\infty, \infty)$. Put a function $f(x) = e^{\gamma x}$ on D with a positive constant γ. Then,*

$$M([a,b]) + s = M([a + s, b + s])$$

for $[a,b] \subset (-\infty, \infty)$, $s \in \mathbb{R}$.

5 Aggregated Mean Ratios of the Quasi-arithmetic Means

In this section, we introduce an aggregated mean ratios of the quasi-arithmetic mean and we discuss the correspondence among the quasi-arithmetic mean, the

aggregated mean ratio and the decision maker's attitude based on his utility. Let D be a fixed domain and let $f : D \mapsto \mathbb{R}$ be a continuous strictly increasing function for utility. Let $M : \mathcal{C}(D) \mapsto D$ be the mean given by (6). Taking the continuous strictly increasing function f as a utility function in decision making, we discuss the decision-maker's judgment by the quasi-arithmetic mean based on utility. Fix an interval $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Define an interior ratio $\theta(a, b)$ induced from a position of the quasi-arithmetic mean $M([a, b])$ on the interval $[a, b]$ by

$$\theta(a, b) := \frac{M([a, b]) - a}{b - a}. \tag{9}$$

We call it an *aggregated mean ratio* under the subjective utility f. It is trivial from (M.i) that $0 \leq \theta(a, b) \leq 1$. In this section, we investigate properties of the ratio θ and we discuss movement of the ratio $\theta(a, b)$ with respect to parameters a, b of an interval $[a, b]$ in local regions and global regions. Dujmović [4,5,6] studied a *conjunction/disjunction degree*, which is a similar type of ratio to (9) in the power case, for computer science. This paper discusses characterizations from the viewpoint of economics. Let $g : D \mapsto \mathbb{R}$ be another continuous strictly increasing function, and let $N : \mathcal{C}(D) \mapsto D$ be the mean defined by g instead of f in the way of (6):

$$N([a, b]) := \begin{cases} g^{-1}\left(\frac{1}{b-a}\int_a^b g(x)\,\mathrm{d}x\right) & \text{if } a < b \\ \lim_{c \downarrow a} g^{-1}\left(\frac{1}{c-a}\int_a^c g(x)\,\mathrm{d}x\right) & \text{if } a = b < \sup D \\ \lim_{c \uparrow b} g^{-1}\left(\frac{1}{b-c}\int_c^b g(x)\,\mathrm{d}x\right) & \text{if } a = b = \sup D \end{cases} \tag{10}$$

for $[a, b] \in \mathcal{C}(D)$. We also put the aggregated mean ratio η for the mean N:

$$\eta(a, b) := \frac{N([a, b]) - a}{b - a} \tag{11}$$

for $[a, b] \in \mathcal{C}(D)$ such that $a < b$.

Theorem 1. *Assume that f and g are C^2-class functions on D. Let $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (i) – (iii) hold.*

(i) *If f and g satisfy $f''/f' < g''/g'$ on (a, b), then $M([a, b]) < N([a, b])$ and $\theta(a, b) < \eta(a, b)$.*
(ii) *If f and g satisfy $f''/f' \leq g''/g'$ on (a, b), then $M([a, b]) \leq N([a, b])$ and $\theta(a, b) \leq \eta(a, b)$.*
(iii) *If f is semi-linear, i.e. $f(x) = rx + s$ with $r, s \in \mathbb{R}$ such that $r > 0$, then $M([a, b]) = (a + b)/2$ and $\theta(a, b) = 1/2$.*

Corollary 1. *Assume that f is a C^2-class function on D. Let $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (i) – (iv) hold.*

(i) *If f satisfies $f'' < 0$ on (a, b), then $\theta(a, b) < 1/2$.*
(ii) *If f satisfies $f'' \leq 0$ on (a, b), then $\theta(a, b) \leq 1/2$.*

(iii) *If f satisfies $f'' > 0$ on (a, b), then $\theta(a, b) > 1/2$.*
(iv) *If f satisfies $f'' \geq 0$ on (a, b), then $\theta(a, b) \geq 1/2$.*

Remark 3. In Corollary 1, $f'' = 0$ implies the decision maker's neutral attitude, $f'' < 0$ corresponds to the decision maker's risk averse attitude, and $f'' > 0$ is the decision maker's risk loving attitude. Therefore, when we may choose two functions f and g as decision maker's utilities, Theorem 1 implies that the utility f yields more risk averse results than g if $f''/f' \leq g''/g'$ on (a, b). Thus, the inequality $\theta(a, b) \leq \eta(a, b)$ means that the aggregated mean ratio $\theta(a, b)$ is more risk averse than $\eta(a, b)$. The index $-f''/f'$ is called the *Arrow-Pratt absolute risk aversion* in economics ([1,14]).

Kolesárová [9] studied relations between $M([a, b])$ and f in the power cases. The following Theorem 2 and Corollary 2 show equivalences regarding the assertion 'if - then' in Theorem 1(ii) and Corollary 1(ii).

Theorem 2. *Assume that f and g are C^2-class functions on D. Let $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (a) – (c) are equivalent.*

(a) $f''/f' \leq g''/g'$ *on* (a, b).
(b) $M([c, d]) \leq N([c, d])$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.
(c) $\theta(c, d) \leq \eta(c, d)$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.

Corollary 2. *Assume that f is a C^2-class function on D. Let $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (a) – (c) are equivalent.*

(a) $f'' \leq 0$ *on* (a, b).
(b) $M([c, d]) \leq (c + d)/2$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.
(c) $\theta(c, d) \leq 1/2$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.

In Theorem 1, we let a middle utility $h := (f + g)/2$. Let $L : \mathcal{C}(D) \mapsto D$ be the mean which is defined by h instead of f in the way of (6), and let ζ be the aggregated mean ratio for the mean L. Then, the following result implies that the estimation by the middle utility $h = (f + g)/2$ gives a middle attitude by utilities f and g in decision making.

Corollary 3. *Assume that f and g are C^2-class functions on D. Let $[a, b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (i) and (ii) hold.*

(i) *If f and g satisfy $f''/f' < g''/g'$ on (a, b), then $M([a, b]) < L([a, b]) < N([a, b])$ and $\theta(a, b) < \zeta(a, b) < \eta(a, b)$.*
(ii) *If f and g satisfy $f''/f' \leq g''/g'$ on (a, b), then $M([a, b]) \leq L([a, b]) \leq N([a, b])$ and $\theta(a, b) \leq \zeta(a, b) \leq \eta(a, b)$.*

The following theorem gives a local property of the ratio $\theta^f(a, b)$ at the neighborhood of $b = a$.

Theorem 3. *Assume that f is a C^2-class function on D. Then, it holds that*

$$\lim_{b \downarrow a} \theta(a, b) = \frac{1}{2} \quad (a \in D) \quad \text{and} \quad \lim_{a \uparrow b} \theta(a, b) = \frac{1}{2} \quad (b \in D). \tag{12}$$

In the following corollary, we find the ratio θ^* in Lemma 3(iii) must be $\theta^* = 1/2$ owing to Theorem 3.

Corollary 4. *Assume that f is a C^2-class function on D. If M satisfies*

$$r \cdot M([a,b]) + s = M([ra+s, rb+s])$$

for $[a,b] \in \mathcal{C}(D)$ and $r, s \in \mathbb{R}$ such that $r > 0$, then it holds that

$$M([a,b]) = \frac{a+b}{2} \quad \textit{for all } [a,b] \in \mathcal{C}(D).$$

Remark 4. Let $f : D \mapsto \mathbb{R}$ be a continuous strictly increasing function. We have discussed the quasi-arithmetic mean M defined by

$$M([a,b]) = f^{-1}\left(\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\right) \tag{13}$$

for $[a,b] \in \mathcal{C}(D)$ $(a < b)$. The mean value criterion $M([a,b])$ is different from the following criterion $m([a,b])$ defined by a weighted sum of the both edges a and b since $M([a,b])$ is given as the aggregated value of all points distributed uniformly on $[a,b]$:

$$m([a,b]) := \lambda a + (1-\lambda)b \tag{14}$$

for $[a,b] \in \mathcal{C}(D)$ $(a < b)$, where λ is a constant $\lambda \in [0,1]$. Actually, when we deal with the form (13), Corollary 2 shows that we cannot choose λ such that $\lambda \neq 1/2$. This paper insists that the quasi-arithmetic means must be defined by every points in the interval not only the edges of the interval.

6 Dual Quasi-arithmetic Means

We discuss quasi-arithmetic means induced from a dual aggregation operator, which is used to aggregate opinions in groups. For example, we can obtain a weak agreement against a strong agreement when we use the dual aggregation operator in group opinions instead of the original aggregation operator (Fodor and Roubens [8]).

Definition 3 (The dual aggregation [2,3]). For an n-ary aggregation operator $\xi^n : [0,1]^n \mapsto [0,1]$, *the dual aggregation operator* $\xi^d : [0,1]^n \mapsto [0,1]$ is given by

$$\xi^d(x_1, x_2, \cdots, x_n) := 1 - \xi^n(1-x_1, 1-x_2, \cdots, 1-x_n) \tag{15}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$.

We can deal with a more general dual aggregation operator ξ^d with any fixed $\kappa \in \mathbb{R}$ instead of 1 in (15). Now we introduce a quasi-arithmetic mean induced from the dual aggregation operator. Let $[a,b] \in \mathcal{C}(D)$ such that $a < b$. Let

$f : D \mapsto \mathbb{R}$ be a continuous strictly increasing function for utility. Put a quasi-arithmetic mean $M : \mathcal{C}(D) \mapsto D$ by (6):

$$M([a,b]) = f^{-1}\left(\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\right).$$

Fix any $\kappa \in \mathbb{R}$. Let a semi-linear strictly decreasing function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ by $\varphi(x) := \kappa - x$ $(x \in \mathbb{R})$. Then the mean induced from the translation φ is called *the quasi-arithmetic mean* M^d, and it is given by

$$M^d([a,b]) := \begin{cases} (f\circ\varphi)^{-1}\left(\frac{1}{b-a}\int_a^b (f\circ\varphi)(x)\,\mathrm{d}x\right) & \text{if } a < b \\ \lim_{c\downarrow a}(f\circ\varphi)^{-1}\left(\frac{1}{c-a}\int_a^c (f\circ\varphi)(x)\,\mathrm{d}x\right) & \text{if } a = b < \sup D \\ \lim_{c\uparrow b}(f\circ\varphi)^{-1}\left(\frac{1}{b-c}\int_c^b (f\circ\varphi)(x)\,\mathrm{d}x\right) & \text{if } a = b = \sup D \end{cases} \tag{16}$$

for $[a,b] \in \mathcal{C}(D)$, where $\circ$ is the composition of maps.

Lemma 5. *A dual quasi-arithmetic mean $M^d : \mathcal{C}(D) \mapsto D$ defined by (16) has the following properties (M.i) – (M.iv):*

(M.i) *Let $[a,b] \in \mathcal{C}(D)$ such that $a < b$. Then it holds that $M^d([a,b]) = \kappa - M([\kappa - b, \kappa - a])$ and*

$$\theta^d(a,b) = 1 - \theta(\kappa - b, \kappa - a),$$

where θ and θ^d are the aggregated mean ratios, which are defined by (9), corresponding to M and M^d respectively.

(M.ii) *Let $[a,b] \in \mathcal{C}(D)$. Then it holds that $a \le M^d([a,b]) \le b$. Especially, $M^d([a,a]) = a$ for $a \in D$.*

(M.iii) *Let $[a,b], [c,d] \in \mathcal{C}(D)$ satisfy$[a,b] \preceq [c,d]$. Then $M^d([a,b]) \le M^d([c,d])$.*

(M.iv) *The map $M^d : \mathcal{C}(D) \mapsto D$ is continuous, i.e. it holds that*

$$\lim_{n\to\infty} M^d([a_n,b_n]) = M^d([a,b])$$

for $[a,b] \in \mathcal{C}(D)$ and $[a_n,b_n] \in \mathcal{C}(D)(n = 1,2,\cdots)$ such that $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$.

Corollary 5. *Assume that f and g are C^2-class functions on D. Let N be the mean which is defined by g instead of f in the way of (6). Let M^d and N^d be the dual quasi-arithmetic means of M and N respectively, and let θ^d and η^d be their aggregated mean ratios respectively. Let $[a,b] \in \mathcal{C}(D)$ such that $a < b$. Then the following (i) and (ii) hold.*

(i) *If f and g satisfy $f''/f' < g''/g'$ on (a,b), then $M^d([a,b]) > N^d([a,b])$ and $\theta^d(a,b) > \eta^d(a,b)$.*

(ii) *If f and g satisfy $f''/f' \le g''/g'$ on (a,b), then $M^d([a,b]) \ge N^d([a,b])$ and $\theta^d(a,b) \ge \eta^d(a,b)$.*

7 Examples

In this section, we give examples for the previous sections. The following example shows the local property of utility functions in comparison of the decision maker's two attitudes f and g which are corresponding to the aggregated mean ratios.

Example 1. Take convex utility functions $f(x) = e^x$ and $g(x) = x^2$ on $D = (0, \infty)$. Then we have

$$\frac{f''(x)}{f'(x)} = 1 \lesseqgtr \frac{1}{x} = \frac{g''(x)}{g'(x)} \iff x \lesseqgtr 1$$

for $x \in D$. From Theorem 1, we obtain $\theta(a,b) < \eta(a,b)$ for $[a,b] \subset (0,1]$ such that $a < b$ and we also obtain $\theta(a,b) > \eta(a,b)$ for $[a,b] \subset [1,\infty)$ such that $a < b$, where $\theta(a,b)$ is the aggregated mean ratio given by $f(x)$ and $\eta(a,b)$ is the aggregated mean ratio given by $g(x)$. This shows that $f(x)$ is more risk averse than $g(x)$ in the region $(0,1)$ and that $f(x)$ is more risk loving than $g(x)$ in the region $[1,\infty)$.

Next we investigate the results in the previous sections for the typical quasi-arithmetic means.

Example 2. Take a function $f(x) = x^\gamma$ on $D = (0,\infty)$ with a constant γ such that $\gamma \neq -1, 0$. Then, for $[a,b] \subset D$ such that $a < b$, the quasi-arithmetic mean is given by the following $M^\gamma([a,b])$:

$$M^\gamma([a,b]) := \left(\frac{b^{\gamma+1} - a^{\gamma+1}}{(\gamma+1)(b-a)}\right)^{1/\gamma},$$

and Corollary 1 implies that its aggregated mean ratio satisfies

$$\theta(a,b) \lesseqgtr \frac{1}{2} \quad \text{if } \gamma \lesseqgtr 1.$$

From Theorem 3, $\lim_{b \downarrow a} \theta(a,b) = \lim_{a \uparrow b} \theta(a,b) = 1/2$ holds, and we obtain

$$\lim_{a \downarrow 0} \theta(a,b) = \lim_{b \to \infty} \theta(a,b) = \left(\frac{1}{\gamma+1}\right)^{1/\gamma}.$$

Hence, the quasi-arithmetic means for γ are as follows:

$$\begin{aligned}
M^2([a,b]) &= \left(\frac{a^2 + ab + b^2}{3}\right)^{1/2} && \text{for } \gamma = 2,\\
M^1([a,b]) &= \frac{a+b}{2} && \text{for } \gamma = 1,\\
\lim_{\gamma \to 0} M^\gamma([a,b]) &= \exp\left(\frac{b\log b - a\log a}{b-a} - 1\right) && \text{as } \gamma \to 0,
\end{aligned}$$

$$\lim_{\gamma\to -1} M^{\gamma}([a,b]) = \frac{b-a}{\log b - \log a} \quad \text{as } \gamma \to -1,$$
$$\lim_{\gamma\to \infty} M^{\gamma}([a,b]) = b \quad \text{as } \gamma \to \infty,$$
$$\lim_{\gamma\to -\infty} M^{\gamma}([a,b]) = a \quad \text{as } \gamma \to -\infty.$$

Acknowledgments

The author is grateful to anonymous referees for their useful advices.

References

1. Arrow, K.J.: Essays in the Theory of Risk-Bearing. Markham, Chicago (1971)
2. Calvo, T., Kolesárová, A., Komorníková, M., Mesiar, R.: Aggregation operators: Basic concepts, issues and properties. In: Calvo, T., et al. (eds.) Aggregation Operators: New Trends and Applications, pp. 3–104. Physica-Verlag, Springer (2002)
3. Calvo, T., Pradera, A.: Double weighted aggregation operators. Fuzzy Sets and Systems 142, 15–33 (2004)
4. Dujmović, J.J.: Weighted Conjunctive and disjunctive means and their application in system evaluation. Univ. Beograd. Publ. Elektotech. Fak. Ser. Mat. Fiz. 483, 147–158 (1974)
5. Dujmović, J.J., Larsen, H.L.: Generalized Conjunction/disjunction. International Journal of Approximate Reasoning 46, 423–446 (2007)
6. Dujmović, J.J., Nagashima, H.: LSP method and its use for evaluation of Java IDEs. International Journal of Approximate Reasoning 41, 3–22 (2006)
7. Fishburn, P.C.: Utility Theory for Decision Making. John Wiley and Sons, New York (1970)
8. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multi-Criteria Decision Support. Kluwer Academic Publishers, Dordrecht (1994)
9. Kolesárová, A.: Limit properties of quasi-arithmetic means. Fuzzy Sets and Systems 124, 65–71 (2001)
10. Kolmogoroff, A.N.: Sur la notion de la moyenne. Acad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. 12, 388–391 (1930)
11. Mesiar, R., Rückschlossová, T.: Characterization of invariant aggregation operators. Fuzzy Sets and Systems 142, 63–73 (2004)
12. Nagumo, K.: Über eine Klasse der Mittelwerte. Japanese Journal of Mathematics 6, 71–79 (1930)
13. Lázaro, J., Rückschlossová, T., Calvo, T.: Shift invariant binary aggregation operators. Fuzzy Sets and Systems 142, 51–62 (2004)
14. Pratt, J.W.: Risk Aversion in the Small and the Large. Econometrica 32, 122–136 (1964)
15. Yager, R.R.: OWA aggregation over a continuous interval argument with application to decision making. IEEE Trans. on Systems, Man, and Cybern. - Part B: Cybernetics 34, 1952–1963 (2004)
16. Yoshida, Y.: The Valuation of European Options in Uncertain Environment. European J. Oper. Res. 145, 221–229 (2003)
17. Yoshida, Y.: A discrete-time model of American put option in an uncertain environment. European J. Oper. Res. 151, 153–166 (2003)

WOWA Enhancement of the Preference Modeling in the Reference Point Method*

Włodzimierz Ogryczak

Warsaw University of Technology, Institute of Control & Computation Engineering,
00-665 Warsaw, Poland
wogrycza@ia.pw.edu.pl

Abstract. The Reference Point Method (RPM) is an interactive technique formalizing the so-called quasi-satisficing approach to multiple criteria optimization. The DM's preferences are there specified in terms of reference (target) levels for several criteria. The reference levels are further used to build the scalarizing achievement function which generates an efficient solution when optimized. Typical RPM scalarizing functions are based on the augmented min-max aggregation where the worst individual achievement minimization process is additionally regularized with the average achievement. The regularization by the average achievement is easily implementable but it may disturb the basic min-max model. We show that the OWA regularization allows one to overcome this flaw since taking into account differences among all ordered achievement values. Further, allowing to define importance weights we introduce the WOWA enhanced RPM. Both the theoretical and implementation issues of the WOWA enhanced method are analyzed. Linear Programming implementation model is developed and proven.

1 Introduction

Consider a decision problem defined as an optimization problem with m criteria (objective functions). In this paper, without loss of generality, it is assumed that all the criteria are minimized. Hence, we consider the following Multiple Criteria Optimization (MCO) problem:

$$\min \{ \ (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) \ : \ \mathbf{x} \in Q \ \} \tag{1}$$

where $\mathbf{x}$ denotes a vector of decision variables to be selected within the feasible set $Q \subset R^n$, and $\mathbf{f}(x) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ is a vector function that maps the feasible set Q into the criterion space R^m. Note that neither any specific form of the feasible set Q is assumed nor any special form of criteria $f_i(\mathbf{x})$ is required. We refer to the elements of the criterion space as outcome vectors. An outcome vector $\mathbf{y}$ is attainable if it expresses outcomes of a feasible solution, i.e. $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for some $\mathbf{x} \in Q$.

* The research was partially supported by the Polish Ministry of Science and Higher Education under the research grant N N516 4307 33.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 38–49, 2008.

Model (1) only specifies that we are interested in minimization of all objective functions f_i for $i \in I = \{1, 2, \ldots, m\}$. Thus it allows only to identify (to eliminate) obviously inefficient solutions leading to dominated outcome vectors, while still leaving the entire efficient set to look for a satisfactory compromise solution. In order to make the multiple criteria model operational for the decision support process, one needs assume some solution concept well adjusted to the DM preferences. This can be achieved with the so-called quasi-satisficing approach to multiple criteria decision problems. The best formalization of the quasi-satisficing approach to multiple criteria optimization was proposed and developed mainly by Wierzbicki [15] as the Reference Point Method (RPM). The reference point method was later extended to permit additional information from the DM and, eventually, led to efficient implementations of the so-called Aspiration/Reservation Based Decision Support (ARBDS) approach with many successful applications [2,16].

The RPM is an interactive technique. The basic concept of the interactive scheme is as follows. The DM specifies requirements in terms of reference levels, i.e., by introducing reference (target) values for several individual outcomes. The reference levels are used to build the scalarizing achievement function which generates an efficient solution when minimized. The computed efficient solution is presented to the DM as the current solution allowing comparison with previous solutions and modifications of the aspiration levels if necessary. In building the function it is assumed that the DM prefers outcomes that satisfy all the reference levels to any outcome that does not reach one or more of the reference levels.

The scalarizing achievement function can be viewed as two-stage transformation of the original outcomes. First, the strictly monotonic partial achievement functions are built to measure individual performance with respect to given reference levels. Having all the outcomes transformed into a uniform scale of individual achievements they are aggregated at the second stage to form a unique scalarization. The RPM is based on the so-called augmented (or regularized) min-max aggregation. Thus, the worst individual achievement is essentially minimized but the optimization process is additionally regularized with the term representing the average achievement. The min-max aggregation is crucial for allowing the RPM to generate all efficient solutions even for nonconvex (and particularly discrete) problems. On the other hand, the regularization is necessary to guarantee that only efficient solution are generated. The regularization by the average achievement is easily implementable but it may disturb the basic min-max model. Actually, the only consequent regularization of the min-max aggregation is the lex-min order or more practical the OWA aggregation with monotonic weights. The latter combines all the partial achievements allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, the third largest weight to the third worst achievement, and so on. The recent progress in optimization methods for ordered averages [8,11] allows one to implement the OWA RPM quite effectively. Further, following the concept of Weighted OWA [13,14] the importance weighting of several achievements may be incorporated into the RPM. Such a WOWA enhancement

of the ARBDS uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than straightforward rescaling of achievement values against those defined by the reference levels [12].

The paper is organized as follows. In the next section we formalize the scalarization achievement functions of the RPM with the detailed formulas for the ARBDS technique. In Section 3 we introduce the OWA and WOWA extensions of the RPM. We show that the WOWA enhanced RPM always generates an efficient solution to the original MCO problem complying simultaneously with the ARBDS preference model assumptions. Further, in Section 4 we develop and prove the Linear Programming implementation model for the method.

2 Scalarizations of the RPM

In the RPM method, depending on the specified reference levels, a special scalarizing achievement function is built which, when minimized, generates an efficient solution to the problem. While building the scalarizing achievement function the following properties of the preference model are assumed. First of all, each solution generated by the scalarizing function optimization must be an efficient solution of the original MCO problem. To meet this requirement the function must be strictly increasing with respect to each outcome. Second, a solution with all individual outcomes satisfying the corresponding reference levels is preferred to any solution with at least one individual outcome worse (greater) than its reference level. That means, the scalarizing achievement function minimization must enforce reaching the reference levels prior to further improving of criteria. Thus, similar to the goal programming approaches, the reference levels are treated as the targets but following the quasi-satisficing approach they are interpreted consistently with basic concepts of efficiency in the sense that the optimization is continued even when the target point has been reached already.

The generic scalarizing achievement function takes the following form [15]:

$$S(\mathbf{a}) = \max_{1 \le i \le m} \{a_i\} + \frac{\varepsilon}{m} \sum_{i=1}^{m} a_i \tag{2}$$

where ε is an arbitrary small positive number and $a_i = s_i(f_i(\mathbf{x}))$, for $i = 1, 2, \ldots, m$, are the partial achievement measuring actual performances of the individual outcomes with partial achievement functions $s_i : R \to R$ with respect to the corresponding reference levels. Let $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ represent the achievement vector. The scalarizing achievement function (2) is, essentially, defined by the worst partial (individual) achievement but additionally regularized with the sum of all partial achievements. The regularization term is introduced only to guarantee the solution efficiency in the case when the minimization of the main term (the worst partial achievement) results in a non-unique optimal solution. Due to combining two terms with arbitrarily small parameter ε, formula (2) is easily implementable and it provides a direct interpretation of the scalarizing achievement function as expressing (dis)utility.

Various functions s_i provide a wide modeling environment for measuring partial achievements [16]. The basic RPM model is based on a single vector of the reference levels, the aspiration vector $\mathbf{r}^a$ and the piecewise linear functions s_i. Real-life applications of the RPM methodology usually deal with more complex partial achievement functions defined with more than one reference point [1,16] which enriches the preference models and simplifies the interactive analysis. In particular, the ARBDS models taking advantages of two reference vectors: vector of aspiration levels $\mathbf{r}^a$ and vector of reservation levels $\mathbf{r}^r$ [2] are used, thus allowing the DM to specify requirements by introducing acceptable and required values for several outcomes. The partial achievement function s_i can be interpreted then as a measure of the DM's satisfaction with the current value of outcome of the ith criterion. It is a strictly increasing function of outcome with value $a_i = 0$ if $f_i(\mathbf{x}) = r_i^a$, and $a_i = 1$ for $f_i(\mathbf{x}) = r_i^r$. Thus the partial achievement functions map the outcomes values onto a normalized scale of the DM's satisfaction. Various functions can be built meeting those requirements [16]. The simplest for implementation is convex piece-wise linear partial achievement function introduced in the ARBDS system for the multiple criteria transshipment problems with facility location [7]:

$$a_i = s_i(f_i(\mathbf{x})) = \begin{cases} \alpha(f_i(\mathbf{x}) - r_i^a)/(r_i^r - r_i^a), & f_i(\mathbf{x}) \le r_i^a \\ (f_i(\mathbf{x}) - r_i^a)/(r_i^r - r_i^a), & r_i^a < f_i(\mathbf{x}) < r_i^r \\ \gamma(f_i(\mathbf{x}) - r_i^r)/(r_i^r - r_i^a) + 1, & f_i(\mathbf{x}) \ge r_i^r \end{cases} \tag{3}$$

where α and γ are arbitrarily defined parameters satisfying $0 < \alpha < 1 < \gamma$. Parameter α represents additional increase of the DM's satisfaction (negative dissatisfaction values) when a criterion generates outcomes better than the corresponding aspiration level. On the other hand, parameter $\gamma > 1$ represents dissatisfaction connected with outcomes worse than the reservation level.

When accepting the loss of a direct utility interpretation, one may consider more powerful lexicographic preference modeling [4,5] based on linear partial achievement $a_i = (f_i(\mathbf{x}) - r_i^a)/(r_i^r - r_i^a)$ but splitted into separate preemptive multilevel interval achievement measures: the reservation level underachievement a_i^r, the aspiration level underachievement a_i^a and the aspiration level overachievement a_i^o defined by the following formula:

$$\begin{array}{ll} a_i^r = s_i^r(f_i(\mathbf{x})) = (f_i(\mathbf{x}) - r_i^r)_+/(r_i^r - r_i^a) & \forall\ i \in I \\ a_i^a = s_i^a(f_i(\mathbf{x})) = \min\{(f_i(\mathbf{x}) - r_i^a)_+/(r_i^r - r_i^a), 1\} & \forall\ i \in I \\ a_i^o = s_i^o(f_i(\mathbf{x})) = (r_i^a - f_i(\mathbf{x}))_+/(r_i^r - r_i^a) & \forall\ i \in I \end{array} \tag{4}$$

Minimization of the scalarizing achievement function (2)–(3) is then replaced with the lexicographic optimization of the multilevel aggregations:

$$\operatorname{lex}\min_{\mathbf{x}} \{(S(\mathbf{a}^r), S(\mathbf{a}^a), S(-\mathbf{a}^o)) : \text{Eq. } (4),\ \mathbf{x} \in Q\} \tag{5}$$

Note that instead of (4), the interval achievements may be defined with the goal programming modeling techniques [6]:

$$f_i(\mathbf{x})/(r_i^r - r_i^a) + a_i^o - a_i^a - a_i^r = r_i^a,\ a_i^o \ge 0,\ 0 \le a_i^a \le 1,\ a_i^r \ge 0 \quad \forall\ i \in I \tag{6}$$

3 WOWA Extension of the RPM

The crucial properties of the RPM are related to the min-max aggregation of partial achievements while the regularization is only introduced to guarantee the aggregation monotonicity. Unfortunately, the distribution of achievements may make the min-max criterion partially passive when one specific achievement is relatively very small for all the solutions. Minimization of the worst achievement may then leave all other achievements unoptimized. Nevertheless, the selection is then made according to linear aggregation of the regularization term instead of the min-max aggregation, thus destroying the preference model of the RPM. This can be illustrated with an example of a simple discrete problem of 7 alternative feasible solutions to be selected according to 6 criteria. Table 1 presents six partial achievements for all the solutions where all the outcome values were within the corresponding intervals between the aspiration and the reservation levels. Thus the partial achievements may be viewed as a_i^a defined according to formula (4) (with $a_r^r = 0$ and $a_i^o = 0$) as well as the a_i defined according to formula (3). All the solutions are efficient. Solutions S1 to S5 reach the aspiration levels (achievement values 0.0) for four of the first five criteria while being quite far from one of them and the aspiration level for the sixth criterion as well (achievement values 0.9). Solution S6 is close to the aspiration levels (achievement values 0.2) for the first five criteria while being far only to the aspiration level for the sixth criterion (achievement values 0.9). All the solutions generate the same worst achievement value 0.9. Therefore, while using the standard augmented min-max aggregation (2) the final selection of a solution depends on the total achievement (regularization term). Actually, one of solutions S1 to S5 will be selected as better than S6.

In order to avoid inconsistencies caused by the regularization in the aggregation (2), the min-max solution may be regularized according to the ordered averaging rules [17]. This is mathematically formalized as follows. Within the space of achievement vectors we introduce map $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$ which orders the coordinates of achievements vectors in a nonincreasing order, i.e., $\Theta(a_1, \ldots, a_m) = (\theta_1(\mathbf{a}), \ldots, \theta_m(\mathbf{a}))$ iff there exists a permutation τ such that $\theta_i(\mathbf{a}) = a_{\tau(i)}$ for all i and $\theta_1(\mathbf{a}) \geq \theta_2(\mathbf{a}) \geq \ldots \geq \theta_m(\mathbf{a})$. The standard min-max aggregation depends on minimization of $\theta_1(\mathbf{a})$ and it ignores values of $\theta_i(\mathbf{a})$ for $i \geq 2$. In order to take into account all the achievement values, one needs to maximize the weighted combination of the ordered achievements thus representing the so-called Ordered Weighted Averaging (OWA) aggregation [17]. Note that the weights are then assigned to the specific positions within the ordered achievements rather than to the partial achievements themselves. With the OWA aggregation one gets the following RPM model:

$$\min_{\mathbf{x}} \{ \sum_{i=1}^{m} w_i \theta_i(\mathbf{a}) \ : \ a_i = s_i(f_i(\mathbf{x})) \ \forall \ i, \ \mathbf{x} \in Q \ \} \tag{7}$$

where $w_1 > w_2 > \ldots > w_m$ are positive and strictly decreasing weights. Actually, they should be significantly decreasing to represent regularization of

Table 1. Sample achievements with passive min-max criterion

									w	0.5	0.25	0.15	0.05	0.03	0.02	
Sol.	a_1	a_2	a_3	a_4	a_5	a_6	max	$\sum$		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	$A_{\mathbf{w}}$
S1	0.9	0.0	0.0	0.0	0.0	0.9	0.9	1.8		0.9	0.9	0.0	0.0	0.0	0.0	0.675
S2	0.0	0.9	0.0	0.0	0.0	0.9	0.9	1.8		0.9	0.9	0.0	0.0	0.0	0.0	0.675
S3	0.0	0.0	0.9	0.0	0.0	0.9	0.9	1.8		0.9	0.9	0.0	0.0	0.0	0.0	0.675
S4	0.0	0.0	0.0	0.9	0.0	0.9	0.9	1.8		0.9	0.9	0.0	0.0	0.0	0.0	0.675
S5	0.0	0.0	0.0	0.0	0.9	0.9	0.9	1.8		0.9	0.9	0.0	0.0	0.0	0.0	0.675
S6	0.2	0.2	0.2	0.2	0.2	0.9	0.9	1.9		0.9	0.2	0.2	0.2	0.2	0.2	0.550
S7	0.9	0.9	0.9	0.2	0.6	0.2	0.9	3.7		0.9	0.9	0.9	0.6	0.2	0.2	0.895

the min-max order. Note that the standard RPM model with the scalarizing achievement function (2) can be expressed as the OWA model (7) with weights $w_2 \ldots = w_m = \varepsilon/m$ and $w_1 = 1 + \varepsilon/m$ thus strictly decreasing in the case of $m = 2$. Unfortunately, for $m > 2$ it abandons the differences in weighting of the second largest achievement, the third largest one etc ($w_2 = \ldots = w_m = \varepsilon/m$). The OWA RPM model (7) allows one to differentiate all the weights by introducing decreasing series (e.g. geometric ones). One may notice that application of decreasing weights $\mathbf{w} = (0.5, 0.25, 0.15, 0.05, 0.03, 0.02)$ within the OWA RPM enables selection of solution S6 from Table 1.

Typical RPM model allows weighting of several achievements only by straightforward rescaling of the achievement values [12]. The OWA RPM model enables one to introduce importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements as defined in the so-called Weighted OWA (WOWA) aggregation [13]. Let $\mathbf{w} = (w_1, \ldots, w_m)$ be a vector of preferential (OWA) weights and let $\mathbf{p} = (p_1, \ldots, p_m)$ denote the vector of importance weights ($p_i \geq 0$ for $i = 1, 2, \ldots, m$ as well as $\sum_{i=1}^{m} p_i = 1$). The corresponding Weighted OWA aggregation of achievements $\mathbf{a} = (a_1, \ldots, a_m)$ is defined as follows:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{i=1}^{m} \omega_i \theta_i(\mathbf{a}), \qquad \omega_i = w^*(\sum_{k \leq i} p_{\tau(k)}) - w^*(\sum_{k<i} p_{\tau(k)}) \tag{8}$$

where w^* is a monotone increasing function that interpolates points $(\frac{i}{m}, \sum_{k \leq i} w_k)$ together with the point (0.0) and τ representing the ordering permutation for $\mathbf{a}$ (i.e. $a_{\tau(i)} = \theta_i(\mathbf{a})$). We focus on the linear interpolation. The WOWA may be expressed with more direct formula where preferential (OWA) weights w_i are applied to averages of the corresponding portions of ordered achievements (quantile intervals) according to the distribution defined by importance weights p_i [9,10]:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{i=1}^{m} w_i m \int_{\frac{i-1}{m}}^{\frac{i}{m}} \overline{F}_{\mathbf{a}}^{(-1)}(\xi)\, d\xi \tag{9}$$

Table 2. WOWA selection with $\mathbf{p} = (\frac{4}{12}, \frac{3}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$

w	0.5		0.25		0.15		0.05		0.03		0.02		$A_{\mathbf{w},\mathbf{p}}(\mathbf{a})$
S1	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7425
S2	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.675
S3	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5625
S4	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.45
S5	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.45
S6	0.9	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.375
S7	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.6	0.2	0.2	0.8815

Table 3. WOWA selection with $\mathbf{p} = (\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{7}{12})$

w	0.5		0.25		0.15		0.05		0.03		0.02		$A_{\mathbf{w},\mathbf{p}}(\mathbf{a})$
S1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.855
S2	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.855
S3	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.855
S4	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.855
S5	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.0	0.0	0.0	0.0	0.855
S6	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.2	0.2	0.2	0.2	0.2	0.8475
S7	0.9	0.9	0.9	0.6	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.6875

where $\overline{F}_{\mathbf{y}}^{(-1)}$ is the stepwise function $\overline{F}_{\mathbf{y}}^{(-1)}(\xi) = \theta_i(\mathbf{y})$ for $\beta_{i-1} < \xi \leq \beta_i$. It can also be mathematically formalized as follows. First, we introduce the left-continuous right tail cumulative distribution function (cdf) defined as:

$$\overline{F}_{\mathbf{y}}(d) = \sum_{i \in I} p_i \delta_i(d) \quad \text{where} \quad \delta_i(d) = \begin{cases} 1 & \text{if } y_i \geq d \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

which for any real (outcome) value d provides the measure of outcomes greater or equal to d. Next, we introduce the quantile function $\overline{F}_{\mathbf{y}}^{(-1)}$ as the right-continuous inverse of the cumulative distribution function $\overline{F}_{\mathbf{y}}$:

$$\overline{F}_{\mathbf{y}}^{(-1)}(\xi) = \sup\{\eta : \overline{F}_{\mathbf{y}}(\eta) \geq \xi\} \quad \text{for} \quad 0 < \xi \leq 1.$$

For instance applying importance weighting $\mathbf{p} = (\frac{4}{12}, \frac{3}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$ to solution achievements from Table 1 and using them together with given there OWA weights $\mathbf{w}$ one gets the WOWA aggregations from Table 2. The corresponding RPM method selects then solution S6, similarly to the case of equal importance weights. On the other hand, when increasing the importance of the last outcome achievements with $\mathbf{p} = (\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{7}{12})$ one gets the WOWA values from Table 3.

The WOWA enhanced ARBDS can be formulated as based on the following lexicographic optimization problem:

$$\operatorname{lex\,min}_{\mathbf{x}} \{(A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^r), A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^a), A_{\mathbf{w},\mathbf{p}}(-\mathbf{a}^o)) : \text{Eq. (4)}, \ \mathbf{x} \in Q\} \tag{11}$$

used to generate current solutions according to the specified preferences. We will show that problem (11) always generates an efficient solution to the original MCO problem complying simultaneously with the ARBDS preference model assumptions.

Theorem 1. *For any reference levels $r_i^a < r_i^r$, any positive weights* $\mathbf{w}$ *and* $\mathbf{p}$*, if* $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ *is an optimal solution of the problem (11), then* $\bar{\mathbf{x}}$ *is an efficient solution of the corresponding multiple criteria problem (1).*

Proof. Let $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ be an optimal solution of the problem (11) with some positive weighting vectors $\mathbf{w}$ and $\mathbf{p}$. Suppose that $\bar{x}$ is not efficient to the multiple criteria problem (1). This means, there exists a decision vector $x \in Q$ such that $f_i(\mathbf{x}) \leq f_i(\bar{\mathbf{x}})$ for all $i \in I$ and $f_{i_o}(\mathbf{x}) < f_{i_o}(\bar{\mathbf{x}})$ for some outcome index $i_o \in I$. Let us define a_i^r, a_i^a and a_i^o according to formula (4). The quadruple $(\mathbf{x}, \mathbf{a}^r, \mathbf{a}^a, \mathbf{a}^o)$ is then a feasible solution of problem (11). Moreover, $a_i^r \leq \bar{a}_i^r$, $a_i^a \leq \bar{a}_i^a$ and $a_i^o \geq \bar{a}_i^o$ for all $i \in I$ where at least one of strict inequalities $a_{i_0}^r < \bar{a}_{i_0}^r$ or $a_{i_0}^a < \bar{a}_{i_0}^a$ or $a_{i_0}^o > \bar{a}_{i_0}^o$ holds. Hence, due to strict monotonicity of the WOWA aggregation with positive weighting vectors, one gets $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^r) \leq A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r)$, $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^a) \leq A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a)$ and $A_{\mathbf{w},\mathbf{p}}(-\mathbf{a}^o) \leq A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o)$ with at least one inequality strict. The latest assertion contradicts the lexicographic optimality of $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ for problem (11), which completes the proof.

Theorem 2. *For any reference levels $r_i^a < r_i^r$, any positive weights* $\mathbf{w}$ *and* $\mathbf{p}$*, if* $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ *is an optimal solution of the problem (11), then all the reservation level underachievements $\bar{a}_i^r$ are equal 0 whenever there exists a feasible solution $\mathbf{x} \in Q$ such that $f_i(\mathbf{x}) \leq r_i^r$ for all $i \in I$.*

Proof. Let $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ be an optimal solution of the problem (11) with some positive weighting vectors $\mathbf{w}$ and $\mathbf{p}$. Suppose that $\bar{a}_{i_0}^r > 0$ for some $i_0 \in I$ and there exists a feasible solution $\mathbf{x} \in Q$ such that $f_i(\mathbf{x}) \leq r_i^r$ for all $i \in I$. Let us define a_i^r, a_i^a and a_i^o according to formula (4) and note that $a_i^r = 0$ for all $i \in I$. The quadruple $(\mathbf{x}, \mathbf{a}^r, \mathbf{a}^a, \mathbf{a}^o)$ is then a feasible solution of problem (11) and, due to positive weights, $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^r) = 0 < A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r)$ thus contradicting the lexicographic optimality of $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$.

Theorem 3. *For any reference levels $r_i^a < r_i^r$, any positive weights* $\mathbf{w}$ *and* $\mathbf{p}$*, if* $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ *is an optimal solution of the problem (11), then all the aspiration level underachievements $\bar{a}_i^a$ are equal 0 whenever there exists a feasible solution $\mathbf{x} \in Q$ such that $f_i(\mathbf{x}) \leq r_i^a$ for all $i \in I$.*

Proof. Let $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ be an optimal solution of the problem (11) with some positive weighting vectors $\mathbf{w}$ and $\mathbf{p}$. Suppose that $\bar{a}_{i_0}^a > 0$ for some $i_0 \in I$ and there exists a feasible solution $\mathbf{x} \in Q$ such that $f_i(\mathbf{x}) \leq r_i^a$ for all $i \in I$. Let us define a_i^r, a_i^a and a_i^o according to formula (4) and note that $a_i^a = a_i^r = 0$ for all $i \in I$. The quadruple $(\mathbf{x}, \mathbf{a}^r, \mathbf{a}^a, \mathbf{a}^o)$ is then a feasible solution of problem (11) and, due to positive weights, $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^a) = 0 < A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a)$ thus contradicting the lexicographic optimality of $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$.

In order to show that the WOWA ARBDS model provides us with a complete parameterization of the efficient set, we will prove in the following theorem that for each efficient solution $\bar{\mathbf{x}}$ there exist aspiration and reservation vectors for which $\bar{\mathbf{x}}$ with the corresponding values of the multilevel achievements is an optimal solution of problem (11).

Theorem 4. *If $\bar{\mathbf{x}}$ is an efficient solution of the multiple criteria problem (1), then there exist aspirations levels r_i^a such that the quadruple $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ is an optimal solution of the corresponding problem (11), for any reservation levels $r_i^r > r_i^a$ and positive weighting vectors $\mathbf{w}$ and $\mathbf{p}$.*

Proof. Let us set the aspiration levels as $r_i^a = f_i(\bar{x})$ for $i \in I$. For any reservation levels $r_i^r > r_i^a$, all the corresponding multilevel achievements defined according to formula (4) take the zero values: $\bar{\mathbf{a}}^r = 0$, $\bar{\mathbf{a}}^a = 0$ and $\bar{\mathbf{a}}^o = 0$. Suppose that for some weights the quadruple $(\bar{\mathbf{x}}, 0, 0, 0)$ is not an optimal solution of the corresponding problem (11). This means there exists a vector $\mathbf{x} \in Q$ such that $\mathbf{a}^r = 0$, $\mathbf{a}^a = 0$, $\mathbf{a}^o \geq 0$ and $A_{\mathbf{w},\mathbf{p}}(-\mathbf{a}^o) < A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o)$. Hence, $f_i(\mathbf{x}) \leq f_i(\bar{\mathbf{x}})\ \forall\ i \in I$ and $f_{i_o}(\mathbf{x}) < f_{i_o}(\bar{\mathbf{x}})$ for some index $i_o \in I$. The latest assertion contradicts the efficiency of $\bar{\mathbf{x}}$ to (1), which completes the proof.

In the proof of Theorem 4 we have used one set of preferential parameters leading to the given solution. Obviously, there are many alternative parameter settings allowing to reach this goal. For instance, one may set the reservation levels as $r_i^r = f_i(\bar{x})$ for $i \in I$ while taking any aspiration levels $r_i^a < r_i^r$.

4 Linear Programming Implementation

An important advantage of the RPM depends on its easy implementation as an expansion of the original MCO problem. Actually, even complicated partial achievement functions of the form (3) are strictly increasing and convex, thus allowing for implementation of the entire RPM model (2) by an LP expansion [7]. The same applies to the WOWA enhanced ARBDS.

Recall that formula (9) defines the WOWA value applying preferential weights w_i to importance weighted averages within quantile intervals. It may be reformulated to use the tail averages

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{k=1}^{m} w'_k m L(\mathbf{a}, \mathbf{p}, \frac{k}{m}), \qquad L(\mathbf{y}, \mathbf{p}, \xi) = \int_0^{\xi} \overline{F}_{\mathbf{y}}^{(-1)}(\alpha) d\alpha \tag{12}$$

where weights $w'_k = w_k - w_{k+1}$ for $k = 1, \ldots, m-1$ and $w'_m = w_m$ and $L(\mathbf{y}, \mathbf{p}, \xi)$ is defined by left-tail integrating of $\overline{F}_{\mathbf{y}}^{(-1)}$.

Values $L(\mathbf{a}, \mathbf{p}, \xi)$ for any $0 \leq \xi \leq 1$ can be given by optimization:

$$L(\mathbf{a}, \mathbf{p}, \xi) = \max_{s_i} \{\sum_{i=1}^{m} a_i s_i : \sum_{i=1}^{m} s_i = \xi, \quad 0 \leq s_i \leq p_i \quad \forall\ i \in I\} \tag{13}$$

Introducing dual variable t corresponding to the equation $\sum_{i=1}^{m} s_i = \xi$ and variables d_i corresponding to upper bounds on s_i one gets the following LP dual expression for $L(\mathbf{a}, \mathbf{p}, \xi)$

$$L(\mathbf{a}, \mathbf{p}, \xi) = \min_{t, d_i} \{\xi t + \sum_{i=1}^{m} p_i d_i : t + d_i \geq a_i,\ d_i \geq 0 \quad \forall\, i \in I\} \tag{14}$$

Following (12) and (14) one gets finally the following model for the WOWA enhanced ARBDS:

$$\begin{aligned}
&\operatorname{lex\,min} [\sum_{k=1}^{m} w'_k z^r_k, \sum_{k=1}^{m} w'_k z^a_k, \sum_{k=1}^{m} w'_k z^o_k] \\
&\text{s.t.}\ \ \mathbf{x} \in Q \\
&\qquad f_i(\mathbf{x})/(r^r_i - r^a_i) + a^o_i - a^a_i - a^r_i = r^a_i \qquad \forall\, i \in I \\
&\qquad a^o_i \geq 0, \quad 0 \leq a^a_i \leq 1, \quad a^r_i \geq 0 \qquad \forall\, i \in I \\
&\qquad z^r_k = k t^r_k + m \sum_{i=1}^{m} p_i d^r_{ik},\ a^r_i \leq t^r_k + d^r_{ik},\ d^r_{ik} \geq 0 \qquad \forall\, i, k \in I \\
&\qquad z^a_k = k t^a_k + m \sum_{i=1}^{m} p_i d^a_{ik},\ a^a_i \leq t^a_k + d^a_{ik},\ d^a_{ik} \geq 0 \qquad \forall\, i, k \in I \\
&\qquad z^o_k = k t^o_k + m \sum_{i=1}^{m} p_i d^o_{ik},\ -a^o_i \leq t^o_k + d^o_{ik},\ d^o_{ik} \geq 0 \qquad \forall\, i, k \in I
\end{aligned} \tag{15}$$

thus allowing for implementation as an LP expansion of the original problem. The following theorem justifies model (15) as an implementation of the WOWA ARBDS approach (11) thus preserving its preference model properties.

Theorem 5. *For any reference levels $r^a_i < r^r_i$, any positive importance weights p_i and positive strictly decreasing weights w_i, if $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ is an optimal solution of the problem (15), then it is an optimal solution of the corresponding problem (11).*

Proof. Let $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ be an optimal solution of the problem (15) with some positive weighting vectors $\mathbf{w}$ and $\mathbf{p}$. Following the WOWA formulas (12) and (14) one may notice that the problem (15) is equivalent to the following lexicographic optimization:

$$\operatorname*{lex\,min}_{\mathbf{x}} \{(A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^r), A_{\mathbf{w},\mathbf{p}}(\mathbf{a}^a), A_{\mathbf{w},\mathbf{p}}(-\mathbf{a}^o)) : \text{Eq. (6)},\ \mathbf{x} \in Q\} \tag{16}$$

Hence, if $\bar{a}^r_i$, $\bar{a}^a_i$ and $\bar{a}^o_i$ fulfill formula (4) for $\bar{\mathbf{x}}$, then the quadruple $\bar{\mathbf{x}}$ is an optimal solution of the corresponding problem (11). In order to prove that formula (4) is satisfied it is enough to show that $\bar{a}^o_i \bar{a}^a_i = 0$ and $(1 - \bar{a}^a_i)\bar{a}^r_i = 0$.

Suppose that $\bar{a}^o_{i_0} \bar{a}^a_{i_0} > 0$ for some index $i_0 \in I$. One may decrease then values of both variables $\bar{a}^o_{i_0}$ and $\bar{a}^a_{i_0}$ by the same small positive number. This means, for sufficiently small positive number δ the quadruple $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^o - \delta \mathbf{e}_{i_0}, \bar{\mathbf{a}}^a - \delta \mathbf{e}_{i_0}, \bar{\mathbf{a}}^r)$, where $\mathbf{e}_{i_0}$ denotes the unit vector corresponding to index i_0, is feasible to problem (16). Due to positive weights w_i and p_i, one gets $(A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r)$,

$A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a - \delta\mathbf{e}_{i_0}), A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o + \delta\mathbf{e}_{i_0})) <_{lex} (A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r), A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a), A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o))$ which contradicts optimality of $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ to problem (16) and thereby (15).

Further, suppose that $(1 - \bar{a}^a_{i_0})\bar{a}^r_{i_0} > 0$ for some index $i_0 \in I$. One may decrease then value of variable $\bar{a}^r_{i_0}$ and simultaneously increase $\bar{a}^a_{i_0}$ by the same small positive number. This means, for sufficiently small positive number δ the quadruple $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^-, \bar{\mathbf{a}}^a + \delta\mathbf{e}_{i_0}, \bar{\mathbf{a}}^r - \delta\mathbf{e}_{i_0})$ is feasible to problem (16). Due to positive weights w_i and p_i, one gets $(A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r - \delta\mathbf{e}_{i_0}), A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a + \delta\mathbf{e}_{i_0}), A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o)) <_{lex} (A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^r), A_{\mathbf{w},\mathbf{p}}(\bar{\mathbf{a}}^a), A_{\mathbf{w},\mathbf{p}}(-\bar{\mathbf{a}}^o))$ which contradicts optimality of $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ to problem (16) and thereby (15).

Thus $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^o, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^r)$ fulfills formula (4) and therefore it is an optimal solution of the corresponding problem (11).

Corollary 1. *For any reference levels $r^a_i < r^r_i$ any positive importance weights p_i and positive strictly decreasing weights w_i, if $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ is an optimal solution of the problem (15), then $\bar{\mathbf{x}}$ is an efficient solution of the corresponding multi-criteria problem (1).*

Corollary 2. *If $\bar{\mathbf{x}}$ is an efficient solution of the multiple criteria problem (1), then there exist aspirations levels $r^a_i = f_i(\mathbf{x})$ such that $(\bar{\mathbf{x}}, \bar{\mathbf{a}}^r, \bar{\mathbf{a}}^a, \bar{\mathbf{a}}^o)$ is an optimal solution of the corresponding problem (15), for any reservation levels $r^r_i > r^a_i$, any positive importance weights p_i and positive strictly decreasing weights w_i.*

5 Conclusions

The reference point method is a very convenient technique for interactive analysis of the multiple criteria optimization problems. It provides the DM with a tool for an open analysis of the efficient frontier. The interactive analysis is navigated with the commonly accepted control parameters expressing reference levels for the individual objective functions. The partial achievement functions quantify the DM satisfaction from the individual outcomes with respect to the given reference levels. The final scalarizing function is built as the augmented min-max aggregation of partial achievements which means that the worst individual achievement is essentially maximized but the optimization process is additionally regularized with the term representing the average achievement. The regularization by the average achievement is easily implementable but it may disturb the basic max-min aggregation. In order to avoid inconsistencies caused by the regularization, the max-min solution may be regularized with the OWA aggregation combining all the partial achievements by allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, the third largest weight to the third worst achievement, and so on. Further following the concept of the Weighted OWA [13] the importance weighting of several achievements may be incorporated into the RPM. Such a WOWA enhancement of the RPM uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than straightforward rescaling of achievement values [12].

The ordered regularizations are more complicated in implementation due to the requirement of pointwise ordering of partial achievements. However, the recent progress in optimization methods for ordered averages [8] allows one to implement the OWA RPM quite effectively by taking advantages of piecewise linear expression of the cumulated ordered achievements. Similar, model can be achieved for the WOWA enhanced ARBDS. Actually, the resulting formulation extends the original constraints and criteria with simple linear inequalities thus allowing for a quite efficient implementation.

References

1. Granat, J., Makowski, M.: ISAAP – Interactive Specification and Analysis of Aspiration-Based Preferences. Eur. J. Opnl. Res. 122, 469–485 (2000)
2. Lewandowski, A., Wierzbicki, A.P.: Aspiration Based Decision Support Systems – Theory, Software and Applications. Springer, Berlin (1989)
3. Liu, X.: Some properties of the weighted OWA operator. IEEE Trans. Systems, Man Cyber. B 368, 118–127 (2006)
4. Ogryczak, W.: Preemptive reference point method. In: Climaco, J. (ed.) Multicriteria Analysis — Proceedings of the XIth International Conference on MCDM, pp. 156–167. Springer, Berlin (1997)
5. Ogryczak, W.: On Goal Programming Formulations of the Reference Point Method. J. Opnl. Res. Soc. 52, 691–698 (2001)
6. Ogryczak, W., Lahoda, S.: Aspiration/Reservation Decision Support – A Step Beyond Goal Programming. J. MCDA 1, 101–117 (1992)
7. Ogryczak, W., Studziński, K., Zorychta, K.: DINAS: A Computer-Assisted Analysis System for Multiobjective Transshipment problems with Facility Location. Comp. Opns. Res. 19, 637–647 (1992)
8. Ogryczak, W., Śliwiński, T.: On solving linear programs with the ordered weighted averaging objective. Eur. J. Opnl. Res. 148, 80–91 (2003)
9. Ogryczak, W., Śliwiński, T.: On Optimization of the Importance Weighted OWA Aggregation of Multiple Criteria. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part I. LNCS, vol. 4705, pp. 804–817. Springer, Heidelberg (2007)
10. Ogryczak, W., Śliwiński, T.: On decision Support Under Risk by the WOWA Optimization. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 779–790. Springer, Heidelberg (2007)
11. Ogryczak, W., Tamir, A.: Minimizing the sum of the k largest functions in linear time. Inform. Proc. Let. 85, 117–122 (2003)
12. Ruiz, F., Luque, M., Cabello, J.M.: A classification of the weighting schemes in reference point procedures formultiobjective programming. J. Opnl. Res. Soc. (forthcoming)
13. Torra, V.: The weighted OWA operator. Int. J. Intell. Syst. 12, 153–166 (1997)
14. Torra, V., Narukawa, Y.: Modeling Decisions Information Fusion and Aggregation Operators. Springer, Berlin (2007)
15. Wierzbicki, A.P.: A Mathematical Basis for Satisficing Decision Making. Math. Modelling 3, 391–405 (1982)
16. Wierzbicki, A.P., Makowski, M., Wessels, J.: Model Based Decision Support Methodology with Environmental Applications. Kluwer, Dordrecht (2000)
17. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans. Systems, Man and Cyber. 18, 183–190 (1988)

Uninorms and Non-contradiction

Ana Pradera

Departamento de Ciencias de la Computación
Universidad Rey Juan Carlos
C. Tulipán s/n, 28933 Móstoles, Madrid, Spain

Abstract. This paper studies the satisfaction of the well-known Non-Contradiction Principle within the class of uninorm aggregation functions, taking into account that this principle may be interpreted in two different ways (a strong one, based on falsity, and a weaker one, relying on self-contradiction). The logical negation is represented by means of strong negation functions, and the most important classes of uninorms are examined.

Keywords: Aggregation functions. Uninorms. Strong Negations. Non-Contradiction Principle.

1 Introduction

Uninorms constitute an important and broad class of mixed aggregation functions. Since their inception in 1996 ([1]) they have attracted a significant and varied amount of research activity, ranging from theoretical developments to practical applications. Among the former one may refer to the investigations on uninorms' structure and classification (see e.g. [2–8]) or to the study of different mathematical properties, such as distributivity or modularity, that usually translate into solving functional equations (see e.g. [9–12]). On the other hand, uninorms have been and continue to be successfully applied in different fields such as expert systems ([13, 14]), fuzzy systems modeling ([15, 16]), approximate reasoning ([17–20]), fuzzy logic ([21, 22]) or fuzzy neurocomputing ([23, 24]).

Several classes and parameterized families of uninorms are nowadays available (see e.g. the recent overview given in [25]), so the question of how to choose the most suitable function for each particular application arises. Various criteria may help in making this decision, such as the requirement of some mathematical properties (e.g. continuity on some specific regions, or idempotency) or the fitting of prototypical data.

Another criterion is the behavior of the uninorm when receiving *contradictory information*: should it be tolerant, intolerant, or perhaps indifferent? In order to analyze this issue, one first needs, on one hand, to clarify what "contradictory information" means, and, on the other hand, to decide how to evaluate the function's behavior. One possible approach is the following: first, to establish that contradictory information is represented by couples $(x, N(x))$, where N is a negation function, and then to evaluate the function's behavior by checking the

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 50–61, 2008.

fulfillment of two well-known mathematical properties, the Non-Contradiction Principle (when seeking intolerant behavior) and the Excluded-Middle Principle (for tolerant behavior).

The present paper focusses on the Non-Contradiction (NC) Principle, aiming to find out which subclasses of uninorms satisfy it and subject to which conditions. Following [26], the NC Principle is understood in two different ways, one based on falsity, as in Modern Logic, and the other one based on self-contradiction, as in Ancient Logic (note that the satisfaction of this principle within the general framework of bivariate aggregation functions[1] has been studied in [27] -ancient logic interpretation- and in [28] -modern logic interpretation-).

The paper is organized as follows. Section 2 provides an overview of the most important issues regarding uninorms, strong negation functions and the NC Principle. Section 3 includes the main results of the paper. It first gives some general results on the fulfillment of the NC Principle within the class of uninorms, and then addresses the particular cases of the most important known subclasses: $\mathcal{U}_{\min}$-uninorms, representable uninoms, $]0,1[^2$-continuous uninorms and idempotent uninorms. Finally, the paper ends with some conclusions.

2 Preliminaries

Uninorms were introduced as a generalization of triangular norms and conorms (t-norms and t-conorms for short)[2]. Excluding the limiting cases of t-norms and t-conorms, uninorms are defined as follows:

Definition 1 (Uninorm). *A* uninorm *is a function* $U : [0,1]^2 \to [0,1]$ *which is associative, commutative, non-decreasing in each variable and has a neutral element* e *belonging to the open interval* $]0,1[$.

Uninorms present a conjunctive behavior when dealing with low input values (those below the neutral element e) and a disjunctive one for high values (those above e). More precisely, any uninorm U with neutral element e is associated with a t-norm T_U and a t-conorm S_U such that ([4]):

$$\forall (x,y) \in [0,e]^2, \qquad U(x,y) = e \cdot T_U\left(\frac{x}{e}, \frac{y}{e}\right),$$

$$\forall (x,y) \in [e,1]^2, \qquad U(x,y) = e + (1-e) \cdot S_U\left(\frac{x-e}{1-e}, \frac{y-e}{1-e}\right).$$

Otherwise (i.e., when receiving a mixture of low and high inputs), unimorms are averaging functions:

$$\forall (x,y) \in [0,e] \times [e,1] \cup [e,1] \times [0,e], \qquad \min(x,y) \leq U(x,y) \leq \max(x,y).$$

[1] Also known as aggregation operators: operations $A : [0,1]^2 \to [0,1]$ which are non-decreasing in each variable and that verify the boundary conditions $A(0,0) = 0$ and $A(1,1) = 1$.

[2] We shall assume that the reader is familiar with the basic concepts regarding t-norms and t-conorms, which can be found e.g. in [29] or [30].

Uninorms may be classified into two different categories ([4]): those verifying $U(0,1) = 0$, which have annihilator element $a = 0$ and are known as *conjunctive uninorms*, and those verifying $U(0,1) = 1$, which have annihilator $a = 1$ and are known as *disjunctive uninorms*. Different classes of uninorms have been identified and studied. The definition and main properties of the most important ones will be overviewed in the next section.

Recall on the other hand that so-called *strong negations* ([31]), i.e., non-increasing functions $N : [0,1] \to [0,1]$ which are involutive (that is, verify $N(N(x)) = x$ for any $x \in [0,1]$), are the most usual way for representing the logic negation. Due to their definition, strong negations are continuous and strictly decreasing functions, they satisfy the boundary conditions $N(0) = 1$ and $N(1) = 0$, and they have a unique fixed point, that we will denote x_N, verifying $0 < x_N < 1$ and $N(x_N) = x_N$. Note also that, for any $x \in [0,1]$, it is $x \leq N(x)$ if and only if $x \leq x_N$. Recall in addition that a function $N : [0,1] \to [0,1]$ is a strong negation if and only if there exists a strictly increasing bijection $\varphi : [0,1] \to [0,1]$ such that $N = \varphi^{-1} \circ (1 - Id_{[0,1]}) \circ \varphi$, i.e., $N(x) = \varphi^{-1}(1 - \varphi(x))$ for any $x \in [0,1]$. The most commonly used strong negation is the *standard negation*, obtained with $\varphi = Id_{[0,1]}$, defined as $N(x) = 1 - x$ for all $x \in [0,1]$.

Regarding the *Non-Contradiction Principle*, it is well-known that such law, in its ancient Aristotelian formulation, can be described as follows: for any statement p, the statements p and *not* p cannot be at the same time, i.e., $p \wedge \neg p$ is *impossible*. In [26] it is argued that such formulation may be interpreted in at least two different ways, depending on how the term *impossible* is understood:

- If the approach that is common in Modern Logic (ML) is adopted, the term *impossible* may be thought as *false*, and then the NC Principle may be expressed, in a structure with minimum element $\mathbf{0}$, as $p \wedge \neg p = \mathbf{0}$ for any statement p.
- Another possibility, which may be considered closer to Ancient Logic (AL), is to interpret *impossible* as *self-contradictory*, understanding that an object is self-contradictory whenever it entails its negation. In this case, the NC Principle may be written as $p \wedge \neg p \vDash \neg(p \wedge \neg p)$ for any statement p, where $\vDash$ represents an entailment relation.

In the context of aggregation functions, if the operation $\wedge$ is represented by means of a bivariate aggregation function $A : [0,1]^2 \to [0,1]$ and the logical negation is modeled by a strong negation N, the NC law can be interpreted in the two following ways:

- Modern Logic interpretation: it is said that *A satisfies NC(ML) with respect to (w.r.t.) N* if

$$\forall x \in [0,1]: \quad A(x, N(x)) = 0 \qquad \text{NC(ML)}$$

- Ancient Logic interpretation: it is said that *A satisfies NC(AL) with respect to (w.r.t.) N* if

$$\forall x \in [0,1]: \quad A(x, N(x)) \leq N(A(x, N(x))) \qquad \text{NC(AL)}$$

From now on, we will say that *A satisfies NC(ML)* (respectively, *satisfies NC(AL)*) if there exists a strong negation N such that A satisfies NC(ML) (respectively, NC(AL)) w.r.t. N.

With all the above preliminary material at hand, the main objective of the present paper can now be reformulated as follows: it consists in solving the above functional equation and functional inequality when A is taken as a uninorm, i.e., to answer the following questions:

- Which classes of uninorms do satisfy NC(ML) (respectively NC(AL))?
- With respect to what kinds of strong negations?
- Are there uninorms satisfying NC(ML) (respectively NC(AL)) w.r.t. *any* strong negation? In which classes?

3 Main Results

Let us first of all deal with the Modern Logic equation, whose solution for the case of uninorms is very simple (see also [28] where this is proved applying a more general result):

Proposition 1. *No uninorm satisfies NC(ML).*

Proof. Let U be a uninorm with neutral element $e \in]0,1[$, let N be a strong negation and let us suppose that U satisfies NC(ML) w.r.t. N. Then, choosing $x = e$, it would be $U(e, N(e)) = 0$, which is equivalent to $N(e) = 0$ or $e = 1$, and this is contradictory with the hypothesis $e \in]0,1[$. □

The case of Ancient Logic interpretation is more interesting. Let us first recall the following characterization of the NC(AL) inequality for arbitrary aggregation functions:

Proposition 2. *[27] Let A be a bivariate aggregation function on $[0,1]$ and let N be a strong negation with fixed point x_N. Then A satisfies NC(AL) w.r.t. N if and only if*

$$\forall x \in [0,1]: \quad A(x, N(x)) \leq x_N \tag{1}$$

In the case of uninorms, the range for checking the condition given in (1) can be reduced:

Proposition 3. *Let U be a uninorm with neutral element $e \in]0,1[$ and let N be a strong negation with fixed point x_N. Then U satisfies NC(AL) w.r.t. N if and only if*

$$\forall x \in [0, N(e)]: \quad U(x, N(x)) \leq x_N \tag{2}$$

Proof. We need to prove that, when dealing with uninorms, (2) implies (1), i.e., the inequality $U(x, N(x)) \leq x_N$ is true for any $x \in]N(e), 1]$. Choosing $x = N(e)$ in (2) gives $N(e) \leq x_N$, so the two following cases may be distinguished:

– If $x \in]N(e), x_N]$, it is $x \leq x_N \leq e$ and $N(x) \leq e$, and then, since any t-norm is below min,

$$U(x, N(x)) = e \cdot T_U\left(\frac{x}{e}, \frac{N(x)}{e}\right) \leq e \cdot \min\left(\frac{x}{e}, \frac{N(x)}{e}\right) = \min(x, N(x)) \leq x \leq x_N.$$

– If $x \in]x_N, 1]$, it is $N(x) \leq x_N$, i.e., $N(x) \in [0, N(e)]$ or $N(x) \in]N(e), x_N]$, so either by hypothesis or thanks to the item above it is $U(N(x), N(N(x))) = U(N(x), x) \leq x_N$, which, by commutativity, entails $U(x, N(x)) \leq x_N$.

□

Let us make explicit some intuitive consequences of the characterization given in Proposition 3:

Corollary 1.

1. *If a uninorm satisfies NC(AL), then it is necessarily a conjunctive uninorm.*
2. *If a uninorm with neutral element e satisfies NC(AL) w.r.t. a strong negation N with fixed point x_N, then $e \geq x_N$.*
3. *No uninorm satisfies NC(AL) with respect to* every *strong negation.*

Proof. The two first items are obtained taking $x = 0$ and $x = N(e)$ in (2), respectively. The last one is easily proved directly from the second item: it cannot be $e \geq x_N$ for *any* $x_N \in]0, 1[$, since $e \neq 1$. □

Therefore, uninorms satisfying NC(AL) may only be found among conjunctive ones choosing an N such that $e \geq x_N$, and only the value of the function at the couples $(x, N(x))$ where $x \leq N(e)$ (and hence $N(x) \geq e$) needs to be checked. In what follows the behavior of the main classes of conjunctive uninorms is analyzed.

3.1 The Class $\mathcal{U}_{\min}$

The class of conjunctive uninorms known as $\mathcal{U}_{\min}$ is obtained when choosing the limiting averaging function min for the region $[0, e] \times [e, 1] \cup [e, 1] \times [0, e]$. Such uninorms have been characterized as follows:

Theorem 1. *([4]) A bivariate function U on $[0, 1]$ is a conjunctive uninorm with neutral element $e \in]0, 1[$ such that the section $U(\cdot, 1)$ is continuous on $[0, e[$ if and only if there exists a t-norm T and a t-conorm S such that*

$$U(x, y) = \begin{cases} e \cdot T\left(\frac{x}{e}, \frac{y}{e}\right) & \text{if } (x, y) \in [0, e]^2, \\ e + (1 - e) \cdot S\left(\frac{x-e}{1-e}, \frac{y-e}{1-e}\right) & \text{if } (x, y) \in [e, 1]^2, \\ \min(x, y) & \text{otherwise} \end{cases}$$

The behavior of this class of uninorms with respect to the NC(AL) Principle is given in the following proposition (see also [27]):

Proposition 4. *A uninorm $U \in \mathcal{U}_{\min}$ with neutral element $e \in]0,1[$ satisfies NC(AL) w.r.t. a strong negation N with fixed point x_N if and only if $e \geq x_N$.*

Proof. If U satisfies NC(AL) w.r.t. N then it is clearly $e \geq x_N$ (see Corollary 1). Conversely, if $e \geq x_N$, it is $N(e) \leq e$, so $x \in [0, N(e)]$ implies $x \leq e$, and in such case Theorem 1 ensures $U(x, N(x)) \leq \min(x, N(x))$. Then it is $U(x, N(x)) \leq x_N$ for any $x \in [0, N(e)]$, and this means, according to Proposition 3, that U satisfies NC(AL) w.r.t. N. □

Of course, given $e \in]0,1[$ it is always possible to find a strong negation N such that $e \geq x_N$, so we have the following immediate consequence of the above characterization:

Corollary 2. *Any uninorm in $\mathcal{U}_{\min}$ satisfies NC(AL).*

Example 1. The weakest uninorm satisfying NC(AL) is the weakest uninorm, given by

$$U(x,y) = \begin{cases} 0, & \text{if } (x,y) \in [0,e[^2, \\ \max(x,y), & \text{if } (x,y) \in [e,1]^2, \\ \min(x,y) & \text{otherwise.} \end{cases}$$

Indeed, the weakest uninorm belongs to $\mathcal{U}_{\min}$ ([4], it is built by means of the weakest t-norm, the drastic product T_D, and the weakest t-conorm, max). Then, according to Proposition 4, it satisfies NC(AL) w.r.t. N as long as N is taken such that $e \geq x_N$.

Example 2. In addition to the weakest uninorm, another commonly cited example of uninorm that, according to Proposition 4, appears to satisfy NC(AL) w.r.t. any N such that $e \geq x_N$ is the one obtained from the family $\mathcal{U}_{\min}$ by choosing $T = \min$ and $S = \max$ ([1]):

$$U_c(x,y) = \begin{cases} \max(x,y) & \text{if } (x,y) \in [e,1]^2, \\ \min(x,y) & \text{otherwise.} \end{cases}$$

3.2 Representable Uninorms

Representable uninorms ([2–4]) constitute an important class of uninorms that can be built by means of univariate generating functions. Restricting ourselves to the conjunctive case:

Proposition 5. *[4] Let $u : [0,1] \to [-\infty, +\infty]$ be a strictly increasing bijection such that $u(e) = 0$ for some $e \in]0,1[$. The function $U : [0,1]^2 \to [0,1]$ given by*

$$U(x,y) = \begin{cases} u^{-1}(u(x) + u(y)), & \text{if } (x,y) \in [0,1]^2 \backslash \{(0,1),(1,0)\}, \\ 0 & \text{otherwise} \end{cases}$$

is a conjunctive uninorm with the neutral element e (known as a conjunctive representable uninorm).

The function u, which is determined up to a positive multiplicative constant, is called an additive generator of the representable uninorm U. Representable uninorms are almost continuous (i.e., continuous everywhere except at the corners $(0,1)$ and $(1,0)$), strictly increasing on $]0,1[^2$ and have an associated strong negation N_u, given by $N_u(x) = u^{-1}(-u(x))$, with fixed point $x_{N_u} = e$.

The following characterization is directly obtained from Propositions 3 and 5:

Proposition 6. *Let U be a conjunctive representable uninorm with neutral element e and additive generator u, and let N be a strong negation with fixed point x_N. Then U satisfies NC(AL) w.r.t. N if and only if*

$$\forall x \in]0, N(e)] : \quad N(x) \leq u^{-1}(u(x_N) - u(x)) \tag{3}$$

The question of whether the inequality (3) can always be satisfied is answered below:

Corollary 3. *Any conjunctive representable uninorm satisfies NC(AL).*

Proof. Choosing N such that $x_N = e$, (3) is equivalent to $N(x) \leq N_u(x)$ for all $x \in]0, N(e)]$, where N_u is the uninorm's associated strong negation. Consequently, any conjunctive representable uninorm satisfies NC(AL) at least w.r.t. its associated strong negation, as well as w.r.t. any weaker strong negation with fixed point e. □

Note also that the chosen strong negations must necessarily be weaker than the uninorm's associated negation, even if they have a different fixed point:

Proposition 7. *Let U be a conjunctive representable uninorm with neutral element e and additive generator u, and let N be a strong negation. If U satisfies NC(AL) w.r.t. N, then necessarily $N \leq N_u$.*

Proof. We have to prove that $N(x) \leq N_u(x)$ for all $x \in [0,1]$. This is obvious if $x \in \{0,1\}$; otherwise, Proposition 2 and the second item in Corollary 1 entail $U(x, N(x)) \leq e$, which is equivalent to $N(x) \leq N_u(x)$. □

If we restrict ourselves to negations having e as fixed point, then the above necessary condition becomes also sufficient:

Proposition 8. *Let U be a conjunctive representable uninorm with neutral element e and additive generator u, and let N be a strong negation with fixed point e. Then U satisfies NC(AL) w.r.t. N if and only if $N \leq N_u$.*

Proof. Taking into account the previous Proposition, we just have to prove that U satisfies NC(AL) w.r.t. N whenever it is $N \leq N_u$, but this is obvious thanks to the monotonicity of U and the fact (see the proof of Corollary 3) that U satisfies NC(AL) w.r.t. N_u. □

Example 3. An important family of parameterized conjunctive representable uninorms ([2–4]) is given, with $\lambda \in]0, +\infty[$, by

$$U_\lambda(x,y) = \begin{cases} \frac{\lambda xy}{\lambda xy+(1-x)(1-y)} & \text{if } (x,y) \in [0,1]^2 \backslash \{(0,1),(1,0)\}, \\ 0 & \text{otherwise} \end{cases}$$

U_λ has neutral element $e_\lambda = \frac{1}{1+\lambda}$ and it can be obtained by means of the additive generator $u_\lambda(x) = \log\left(\frac{\lambda x}{1-x}\right)$. According to the above results, U_λ satisfies NC(AL) w.r.t. its associated negation, N_{u_λ}, and any weaker negation with the same fixed point.

3.3 $]0,1[$-continuous Uninorms

Representable uninorms belong to a wider class of uninorms: those which are continuous on the open square $]0,1[^2$. The representation of such uninorms in the conjunctive case is recalled below:

Theorem 2. *([7], see also [8]) Let U be a conjunctive uninorm with neutral element $e \in]0,1[$ which is continuous on $]0,1[^2$. Then U can be represented as follows:*

$$U(x,y) = \begin{cases} e \cdot T\left(\frac{x}{e}, \frac{y}{e}\right) & \text{if } x,y \in [0,a], \\ u^{-1}(u(x)+u(y)) & \text{if } x,y \in]a,1[, \\ x & \text{if } x \in [0,a], y \in]a,1[\text{ or } x \in [0,c[, y = 1, \\ y & \text{if } x \in]a,1[, y \in [0,a] \text{ or } x = 1, y \in [0,c[, \\ 1 & \text{if } x \in]c,1], y = 1 \text{ or } x = 1, y \in]c,1], \\ x \text{ or } y & \text{if } x = c, y = 1 \text{ or } x = 1, y = c \end{cases}$$

where T is a continuous t-norm, $u : [a,1] \to [-\infty,+\infty]$ is a strictly increasing bijection such that $u(e) = 0$, $a \in [0,e[$, $c \in [0,a]$ and $U(c,c) = c$.

The satisfaction of the NC(AL) in the case of conjunctive $]0,1[^2$-continuous uninorms can be characterized as follows:

Proposition 9. *Let U be a conjunctive uninorm continuous on $]0,1[^2$ with neutral element $e \in]0,1[$ and let N be a strong negation with fixed point x_N. Then U satisfies NC(AL) w.r.t. N if and only if it is either $N(e) \leq a$ or else $N(x) \leq u^{-1}(u(x_N) - u(x))$ for all $x \in]a,N(e)]$, where a and u come from the representation of U given in Theorem 2.*

Proof. The proof from left to right is accomplished using Proposition 3 and the fact that $x \in]a,N(e)]$ implies, by Theorem 2, $U(x,N(x)) = u^{-1}(u(x)+u(N(x)))$. Conversely:

- If $N(e) \leq a$, Theorem 2 shows that for any $x \in [0,N(e)]$ it is $U(x,N(x)) = \min(x,N(x))$, and hence U satisfies NC(AL) w.r.t. N.
- Otherwise it is, by hypothesis, $N(x) \leq u^{-1}(u(x_N)-u(x))$ for all $x \in]a,N(e)]$, and then:
  - If $x \in]a,N(e)]$, since it is $x > a$ and $N(x) \geq e$ Theorem 2 shows that the hypothesis is equivalent to $U(x,N(x)) \leq x_N$.

- If $x \in [0, a]$, according to Theorem 2 it is either $U(x, N(x)) = \min(x, N(x))$ or $U(x, N(x)) = e \cdot T\left(\frac{x}{e}, \frac{N(x)}{e}\right)$, but both cases entail $U(x, N(x)) \leq x_N$. □

Corollary 4. *Any conjunctive uninorm continuous on* $]0, 1[^2$ *satisfies NC(AL).*

Proof. If the value a in Theorem 2 is zero, then U is a conjunctive representable uninorm and hence it satisfies NC(AL) (Corollary 3). Otherwise, according to Proposition 9 it suffices to choose a strong negation N such that $N(e) \leq a$ in order to guarantee that U satisfies NC(AL) w.r.t. N (note that $x_N \leq a$ entails this latter condition). □

3.4 Idempotent Uninorms

Idempotent uninorms, that is, those verifying $U(x, x) = x$ for all $x \in [0, 1]$, were first studied for the particular cases of left and right-continuous functions in [5], and have later on been characterized in the general case in [6]. We will concentrate on the former:

Theorem 3. *([5])*

1. *A bivariate function U on $[0, 1]$ is a left-continuous idempotent uninorm with neutral element $e \in]0, 1]$ if and only if there exists a decreasing function $g : [0, 1] \to [0, 1]$ with fixed point e satisfying*
   (i) $g(g(x)) \geq x$ for all $x \in [0, g(0)]$, and
   (ii) $g(x) = 0$ for all $x \in]g(0), 1]$
   such that U is given, for any $(x, y) \in [0, 1]^2$, by

$$U(x, y) = \begin{cases} \min(x, y) & \text{if } y \leq g(x) \text{ and } x \leq g(0), \\ \max(x, y) & \text{otherwise.} \end{cases}$$

2. *A bivariate function U on $[0, 1]$ is a right-continuous idempotent uninorm with neutral element $e \in [0, 1[$ if and only if there exists a decreasing function $g : [0, 1] \to [0, 1]$ with fixed point e satisfying*
   (i) $g(g(x)) \leq x$ for all $x \in [g(1), 1]$, and
   (ii) $g(x) = 1$ for all $x \in [0, g(1)[$
   such that U is given, for any $(x, y) \in [0, 1]^2$, by

$$U(x, y) = \begin{cases} \max(x, y) & \text{if } y \geq g(x) \text{ and } x \geq g(1), \\ \min(x, y) & \text{otherwise.} \end{cases}$$

With regards to the satisfaction of the NC(AL) law, we can first of all note the existence of at least one idempotent uninorm satisfying NC(AL): the uninorm U_c given in Example 2, which is clearly a conjunctive right-continuous idempotent uninorm that, according to Proposition 4, satisfies NC(AL) w.r.t. any N such that $e \geq x_N$. However, contrary to what happens with uninorms in $\mathcal{U}_{\min}$ and conjunctive $]0, 1[^2$-continuous uninorms, not every conjunctive idempotent uninorm satisfies NC(AL). In order to prove this, let us first note that Theorem

3 establishes that any idempotent uninorm is determined by a unary function g, usually referred to as its associated function, that may be used to characterize the classes of (conjunctive) left and right-continuous idempotent uninorms that satisfy NC(AL):

Proposition 10. *Let N be a strong negation with fixed point x_N.*

1. *A left-continuous idempotent uninorm with neutral element e and associated function g satisfies NC(AL) w.r.t. N if and only if*

$$\forall x \in [0, N(e)]: \quad N(x) \leq g(x) \tag{4}$$

2. *A right-continuous idempotent uninorm with neutral element e and associated function g satisfies NC(AL) w.r.t. N if and only if*

$$N(e) < g(1) \quad \text{or} \quad \forall x \in [g(1), N(e)]: \quad N(x) < g(x) \tag{5}$$

Proof. The results are obtained applying Proposition 3 to the characterization given in Theorem 3. □

Remark 1. Note that the conditions (4) and (5) in the above Proposition implicitly ensure the conjunctive behavior of the uninorms: indeed, it is $g(0) = 1$ for the left-continuous case (choosing $x = 0$ in (4)) and $g(1) \neq 0$ for the right-continuous one ($g(1) = 0$ would entail the false statement $g(0) > 1$ choosing $x = 0$ in (5)).

The characterization given in Proposition 10 allows for the following conclusions:

Corollary 5.

1. *Conjunctive right-continuous uninorms always satisfy NC(AL).*
2. *Not every conjunctive left-continuous uninorms satisfies NC(AL), but some do.*

Proof. First, given an arbitrary conjunctive right-continuous uninorm, it is always possible to find a strong negation N such that $N(e) < g(1)$, and then, according to Proposition 10, such uninorm satisfies NC(AL) w.r.t. N.

Now, according to Theorem 3, the function

$$g(x) = \begin{cases} 1 & \text{if } x = 0, \\ e & \text{if } 0 < x \leq e, \\ 0 & \text{otherwise} \end{cases}$$

provides a conjunctive left-continuous uninorm for which, due to the continuity of strong negations, it is impossible to find any N such that $N(x) \leq g(x)$ for all $x \in [0, N(e)]$. On the other hand, there are left-continuous idempotent uninorms satisfying NC(AL): for example, the choice of a strong negation N with fixed point $x_N = e$ as associated function directly provides a conjunctive left-continuous idempotent uninorm (called U_c^N in [5]) which obviously satisfies NC(AL) w.r.t. N. □

4 Conclusions

This paper has analyzed the behavior of uninorms when dealing with contradictory information by studying the satisfaction of the Non-Contradiction Principle under two different interpretations. The main conclusions are the following:

- No uninorm satisfies the Non-Contradiction Principle in its stronger version (NC(ML), based on falsity).
- Regarding the weaker version (NC(AL), based on self-contradiction):
  - No uninorm satisfies it w.r.t. *any* strong negation.
  - Uninorms satisfying NC(AL) w.r.t. some strong negations may be found in each of the main classes of conjunctive uninorms. Such uninorms have been characterized, in particular, in the cases of $\mathcal{U}_{\min}$-uninorms, representable uninorms, $]0,1[^2$-continuous uninorms and left and right-continuous idempotent uninorms (Propositions 4, 6, 9 and 10).
  - $\mathcal{U}_{\min}$-uninorms, conjunctive $]0,1[^2$-continuous uninorms (including representable uninorms) and conjunctive right-continuous idempotent uninorms always satisfy NC(AL) (i.e., it is always possible to find a strong negation such that the NC(AL) inequality is verified). However, even if many conjunctive left-continuous idempotent uninorm satisfy NC(AL), not every uninorm in this class satisfies it.

References

1. Yager, R., Rybalov, A.: Uninorm aggregation operators. Fuzzy Sets and Systems 80, 111–120 (1996)
2. Dombi, J.: Basic concepts for a theory of evaluation: the aggregative operator. Europ. J. Oper. Research 10, 282–293 (1982)
3. Klement, E., Mesiar, R., Pap, E.: On the relationship of associative compensatory operators to triangular norms and conorms. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems 4, 129–144 (1996)
4. Fodor, J., Yager, R., Rybalov, A.: Structure of uninorms. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems 5, 411–427 (1997)
5. De Baets, B.: Idempotent uninorms. Europ. J. Oper. Research 118, 631–642 (1999)
6. Martín, J., Mayor, G., Torrens, J.: On locally internal monotonic operations. Fuzzy Sets and Systems 137, 27–42 (2003)
7. Hu, S.K., Li, Z.F.: The structure of continuous uni-norms. Fuzzy Sets and Systems 124, 43–52 (2001)
8. Drygaś, P.: On the structure of continuous uninorms. Kybernetika 43(2), 183–196 (2007)
9. Calvo, T., De Baets, B., Fodor, J.: The functional equations of Frank and Alsina for uninorms and nullnorms. Fuzzy Sets and Systems 120, 385–394 (2001)
10. Mas, M., Mayor, G., Torrens, J.: The modularity condition for uninorms and t-operators. Fuzzy Sets and Systems 126, 207–218 (2002)
11. Mas, M., Mayor, G., Torrens, J.: The distributivity condition for uninorms and t-operators. Fuzzy Sets and Systems 128, 209–225 (2002)
12. Drewniak, J., Drygaś, P., Rak, E.: Distributivity between uninorms and nullnorms. Fuzzy Sets and Systems 159(13), 1646–1657 (2008)

13. De Baets, B., Fodor, J.: Van Melle's combining function in MYCIN is a representable uninorm: An alternative proof. Fuzzy Sets and Systems 104, 133–136 (1999)
14. Tsadiras, A., Margaritis, K.: The MYCIN certainty factor handling function as uninorm operator and its use as a threshold function in artificial neurons. Fuzzy Sets and Systems 93, 263–274 (1999)
15. Yager, R.: Uninorms in fuzzy systems modeling. Fuzzy Sets and Systems 122, 167–175 (2001)
16. Yager, R., Kreinovich, V.: Universal approximation theorem for uninorm-based fuzzy systems modeling. Fuzzy Sets and Systems 140, 331–339 (2003)
17. De Baets, B., Fodor, J.: Residual operators of uninorms. Soft Comput. 3(2), 89–100 (1999)
18. Ruiz-Aguilera, D., Torrens, J.: Residual implications and co-implications from idempotent uninorms. Kybernetika 40(1), 21–38 (2004)
19. Mas, M., Monserrat, M., Torrens, J.: Two types of implications derived from uninorms. Fuzzy Sets Syst. 158(23), 2612–2626 (2007)
20. Ruiz-Aguilera, D., Torrens, J.: Distributivity of residual implications over conjunctive and disjunctive uninorms. Fuzzy Sets and Systems 158(1), 23–37 (2007)
21. Gabbay, D.M., Metcalfe, G.: Fuzzy logics based on [0, 1]-continuous uninorms. Arch. Math. Log. 46(5-6), 425–449 (2007)
22. Marchioni, E., Montagna, F.: On triangular norms and uninorms definable in Ł $\Pi\frac{1}{2}$. Int. J. Approx. Reasoning 47(2), 179–201 (2008)
23. Pedrycz, W.: Logic-based fuzzy neurocomputing with unineurons. IEEE T. Fuzzy Systems 14(6), 860–873 (2006)
24. Pedrycz, W., Hirota, K.: Uninorm-based logic neurons as adaptive and interpretable processing constructs. Soft Comput. 11(1), 41–52 (2007)
25. Fodor, J., De Baets, B.: Uninorm basics. In: Wang, P., Ruan, D., Kerre, E. (eds.) Fuzzy Logic. A Spectrum of Theoretical & Practical Issues, pp. 49–64. Springer, Berlin (2007)
26. Trillas, E., Alsina, C., Pradera, A.: Searching for the roots of Non-Contradiction and Excluded-Middle. International Journal of General Systems 31, 499–513 (2002)
27. Pradera, A., Trillas, E.: Aggregation operators from the ancient NC and EM point of view. Kybernetika 42, 243–260 (2006)
28. Pradera, A., Trillas, E.: Aggregation, Non-Contradiction and Excluded-Middle. Mathware & Soft Computing XIII, 189–201 (2006)
29. Klement, E., Mesiar, R., Pap, E.: Triangular Norms. Kluwer, Dordrecht (2000)
30. Alsina, C., Frank, M., Schweizer, B.: Associative functions. In: Triangular Norms and Copulas. World Scientific, Singapore (2006)
31. Trillas, E.: Sobre funciones de negación en la teoría de los subconjuntos difusos. In: Stochastica, vol. III, pp. 47–59 (in Spanish) (1979); Reprinted (English version) S. Barro et al. (eds.): Advances of Fuzzy Logic, Universidad de Santiago de Compostela, pp. 31–43 (1998)

Choquet Stieltjes Integral, Losonczi's Means and OWA Operators

Vicenç Torra[1] and Yasuo Narukawa[2]

[1] IIIA-CSIC, Institut d'Investigació en Intel·ligència Artificial,
Campus UAB, 08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es
[2] Toho Gakuen,
3-1-10 Naka, Kunitachi, Tokyo, 186-0004 Japan
narukawa@d4.dion.ne.jp

Abstract. Neat OWA operators have been defined as a generalization of the OWA operators. In this paper we study these operators establishing some relationships with some other operators. In particular, we link them with the Losonczi's mean.

Keywords: Aggregation operators, Losonczi's mean, OWA operators.

1 Introduction

Aggregation operators [4,8,19] are used to combine information to obtain a datum of better quality. In recent years there is an increasing interest in these topics for their application in decision problems and artificial intelligence applications.

In this paper we focus on the Losonczi's mean, which was proposed in [9] and which generalizes Bajraktarević's mean [1,2]. Our interest is to establish some relationships between these operators and some of the ones that have been defined more recently related to the OWA operator. In particular, we will consider the OWA, and the neat OWA.

The structure of this paper is as follows. In Section 2 we review fuzzy measures and introduce a few results related to the Choquet Stieltjies integral. In Section 3, we review the aggregation operators we need later on. Then, in Section 5, we establish the relationships between these operators. Section 6 uses Losonczi's mean to introduce a Losonczi's OWA operator. The paper finishes with some conclusions.

2 Fuzzy Measures and the Choquet Stieltjes Integral

In this section, we define fuzzy measures, the Choquet integral and the Choquet Stieltjes integral, and show their basic properties.

Let X be a locally compact Hausdorff space and $\mathcal{B}$ be a class of Borel sets, that is, the smallest $\sigma-$algebra which includes the class of all closed sets. We say that $(X, \mathcal{B})$ is a measurable space.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 62–73, 2008.

Example 1. (1) The set of all real numbers R is a locally compact Hausdorff space. If $X = R$, $\mathcal{B}$ is the smallest $\sigma-$ algebra which includes the class of all closed intervals.
(2) Let $X := \{1, 2, \ldots, N\}$. X is a compact Hausdorff space with a discrete topology. Then we have $\mathcal{B} = 2^X$.

Definition 1. [17] Let $(X, \mathcal{B})$ be a measurable space. A fuzzy measure (or a non-additive measure) μ is a real valued set function, $\mu : \mathcal{B} \longrightarrow [0, 1]$ with the following properties;

(1) $\mu(\emptyset) = 0$
(2) $\mu(A) \leq \mu(B)$ whenever $A \subset B$, $A, B \in \mathcal{B}$.

We say that the triplet $(X, \mathcal{B}, \mu)$ is a fuzzy measure space if μ is a fuzzy measure.

Definition 2. Let $(X, \mathcal{B})$ be a measurable space. A function $f : X \to R$ is said to be measurable if $\{x | f(x) \geq \alpha\} \in \mathcal{B}$ for all $\alpha \in R$.

Example 2. Let f be a continuous function. Then for all $\alpha \in R$ $\{f \geq \alpha\}$ is a closed set. Therefore f is measurable.

$\mathcal{F}(X)$ denotes the class of non-negative measurable functions, that is,

$$\mathcal{F}(X) = \{f | f : X \to R^+, f : \text{measurable}\}$$

Definition 3. [5,11] Let $(X, \mathcal{B}, \mu)$ be a fuzzy measure space. *The Choquet integral* of $f \in \mathcal{F}(X)$ with respect to μ is defined by

$$(C) \int f d\mu = \int_0^\infty \mu_f(r) dr,$$

where $\mu_f(r) = \mu(\{x | f(x) \geq r\})$.

Suppose that $X = \{1, 2, \ldots, N\}$. The $i-$th order statistic $a^{(i)}$ [21] is a functional on R^N which is defined by arranging the components of $\mathbf{a} = (a_1, \cdots, a_N) \in R^N$ in the increasing order

$$a^{(1)} \leq \cdots \leq a^{(i)} \leq \cdots \leq a^{(N)}.$$

Using the $i-$th order statistics, the Choquet integral is written as

$$(C) \int \mathbf{a} d\mu = \sum_{i=1}^{N} (a^{(i)} - a^{(i-1)}) \mu(\{(i) \cdots (N)\}),$$

where we define $a^{(0)} := 0$.

Definition 4. [6] Let $f, g \in \mathcal{F}(X)$. We say that f and g are comonotonic if

$$f(x) < f(x') \Rightarrow g(x) \leq g(x')$$

for $x, x' \in X$.

We say that f and g are strongly comonotonic if

$$f(x) < f(x') \Leftrightarrow g(x) < g(x')$$

for $x, x' \in X$. $f \sim_s g$ denotes that f and g are strongly comonotonic.

Definition 5. Let I be a real-valued functional on $\mathcal{F}(X)$. We say I is *comonotonically additive* if and only if $I(f+g) = I(f) + I(g)$ for comonotonic $f, g \in \mathcal{F}(X)$.

It is well known that the Choquet integral is a comonotonically additive functional on $\mathcal{F}(X)$. Conversely a comonotonically additive functional on $\mathcal{F}(X)$ is represented as a Choquet integral [15,16].

Let $C_b(X)$ be a class of bounded continuous functions on X. Since $\sim_s$ is an equivalent relation, we can define an equivalence class $[f] \in C_b(X)/\sim_s$. Then applying Hahn-Banach's theorem we have the next theorem.

Theorem 1. *Let $(X, \mathcal{B}, \mu)$ be a fuzzy measure space. For every $f \in C_b(X)$, there exists a probability $P_{[f]}$ on $\mathcal{B}$ such that*

$$(C)\int f d\mu = \int f dP_{[f]}.$$

Now, we define the Choquet-Stieltjes integral [12].

Definition 6. *Let $(X, \mathcal{B}, \mu)$ be a fuzzy measure space and $\varphi : R^+ \to R^+$ be a non-decreasing real valued function with $\varphi(0) = 0$. Then, we can define the Lebesgue-Stieltjes measure ν_φ [14] on the real line by*

$$\nu_\varphi((a,b)) := \varphi(b-0) - \varphi(a+0),$$

where $\varphi(b+0) := \lim_{x \to b+0} \varphi(x)$ and $\varphi(a-0) := \lim_{x \to a-0} \varphi(x)$.

We define the Choquet-Stieltjes integral $CS_{\mu,\varphi}(f)$ with respect to μ, φ by

$$CS_{\mu,\varphi}(f) := \int_0^\infty \mu_f(r) d\nu_\varphi(r),$$

where $\mu_f(r) = \mu(\{x | f(x) \geq r\})$.

When we use the space $X = \{1, 2, \ldots, N\}$, the Choquet-Stieltjes integral can be rewritten, using the $i-$th order statistics, as

$$\begin{aligned} CS_{\mu,\varphi}(\mathbf{a}) &= \sum_{i=1}^{N} (\varphi(a^{(i)}) - \varphi(a^{(i-1)})) \mu(\{(i) \cdots (n)\}) \\ &= \sum_{i=1}^{n} \varphi(a^{(i)}) \{\mu(\{(i) \cdots (n)\}) - \mu(\{(i+1) \cdots (n)\})\}. \end{aligned}$$

Proposition 1. *Let $(X, \mathcal{B}, \mu)$ be a fuzzy measure space and $\varphi : R^+ \to R^+$ be a continuous and strictly increasing function with $\varphi(0) = 0$. Then the Choquet-Stieltjes integral of f with respect to μ, φ is a Choquet integral of $\varphi(f)$, that is,*

$$CS_{\mu,\varphi}(f) = (C)\int \varphi(f) d\mu.$$

3 On Some Aggregation Operators

In this section we review some aggregation operators that are later on needed in this paper. Most of the functions reviewed here can also be found in [19,20]. For the sake of simplicity, we consider aggregation operators in $[0, 1]$. Other real intervals might be also appropriate.

The review includes the Losonczi's mean. This family of means is a generalization of weighted means and of quasi-weighted means in the sense that instead of having constant weights p_i attached to inputs a_i, they have weights that are functions of the inputs. That is, the operator uses functions π_i of a_i, instead of using weights p_i.

Before establishing the Losonczi's mean, we review the weighted mean and the quasi-weighted mean to underline the similarities between the operators. We start with the definition of an aggregation operator and of a weighting vector. Note that we require aggregation operators to be idempotent although this is not always the case in the literature.

Definition 7. *Let $D \subset R^N$. An aggregation operator Ag is a function $Ag : D \to R$ with the following properties;*

(1) *(Unanimity or idempotency)*
$Ag(a, \ldots, a) = a$ *if* $(a, \ldots, a) \in D$
(2) *(Monotonicity)*
If $a_i \leq b_i$ *for all* $i = 1, \ldots, n$, $\mathbf{a} = (a_1, \ldots, a_N), \mathbf{b} = (b_1, \ldots, b_N)$ $\mathbf{a}, \mathbf{b} \in D$, *then* $Ag(\mathbf{a}) \leq Ag(\mathbf{b})$.

An aggregation operator is said to be neat when the result is invariant to any permutation of the input data. That is, if for any permutation π of $\{1, \ldots, N\}$, the following equation holds:

$$Ag(a_1, \ldots, a_N) = Ag(a_{\pi(1)}, \ldots, a_{\pi(N)}).$$

Definition 8. *A vector $\mathbf{p} = (p_1, \ldots, p_N)$ such that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$ is a weighting vector of dimension N.*

Definition 9. *Given a weighting vector $\mathbf{p} = (p_1, \ldots, p_N)$ and a function ϕ (strictly increasing with inverse ϕ^{-1}), the weighted mean WM and the quasi-weighted mean QWM are defined as follows:*

$$WM_{\mathbf{p}}(\mathbf{a}) = \sum_{i=1}^{N} p_i a_i$$

$$QWM_{\mathbf{p}}(\mathbf{a}) = \phi^{-1}\big(\sum_{i=1}^{N} p_i \phi(a_i)\big)$$

for $\mathbf{a} = (a_1, \ldots, a_N) \in R^N$.

The next proposition is obvious from Definition 9 and Proposition 1.

Proposition 2. *Let $X := \{1, 2, \ldots, N\}$ and $\mathbf{p} = (p_1, \ldots, p_N)$ be a weighting vector. We can define a probability measure P on 2^X by $P(\{i\}) := p_N$.*

$$WM_{\mathbf{p}}(\mathbf{a}) = (C) \int \mathbf{a} dP$$

$$QWM_{\mathbf{p}}(\mathbf{a}) = \phi^{-1}\big(CS_{P,\phi}(\mathbf{a})\big)$$

for $\mathbf{a} = (a_1, \ldots, a_N) \in R^N$.

As stated above, Losonczi's means corresponds to the QWM but with weighting functions $\pi_i(a_i)$ instead of weights p_i. Additionally, the weights are explicitly normalized to avoid the cumbersome requirement of functions adding one. We give now the definition taking all this into account.

Definition 10. *[9] Given functions π_i and ϕ (ϕ strictly increasing with inverse ϕ^{-1}), the Losonczi's mean is defined as follows:*

$$LM(a_1, \ldots, a_N) = \phi^{-1}\left(\frac{\sum_{i=1}^N \pi_i(a_i)\phi(a_i)}{\sum_{i=1}^N \pi_i(a_i)}\right)$$

This operator generalizes the QWM, the WM, and other means as e.g. the counter-harmonic mean $\sum a_i^p / \sum a_i^{p-1}$ (also called the BADD operator in [23]). See e.g. [3] and [19].

We have the next proposition in similar way to Proposition 2.

Proposition 3. *Let $X := \{1, 2, \ldots, N\}$, let $\pi_i : R \to R^+$ and let $\phi : R \to R$ (ϕ strictly increasing with inverse ϕ^{-1}). Then, let us define a probability measure P on 2^X by $P_{\mathbf{a}}(\{i\}) := \frac{\pi_i(a_i)}{\sum_{i=1}^N \pi_i(a_i)}$. Under these conditions, we have*

$$LM_{\pi,\phi}(\mathbf{a}) = \phi^{-1}\big(CS_{P_{\mathbf{a}},\phi}(\mathbf{a})\big)$$

for $\mathbf{a} = (a_1, \ldots, a_N) \in R^N$.

Yager introduced the Ordered Weighting Averaging operator in [22].

Definition 11. *[22] Given a weighting vector $\mathbf{w} = (w_1, \ldots, w_N)$, the Ordered Weighting Averaging operator is defined as follows:*

$$OWA_{\mathbf{w}}(\mathbf{a}) = \sum_{i=1}^N w_i a_{\sigma(i)}$$

where σ defines a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$, $\mathbf{a} = (a_1, \ldots, a_n)$.

It is obvious from this definition that the OWA is neat.

A fuzzy measure μ on $\mathcal{B}$ is said to be symmetric [10] if $\mu(A) = \mu(B)$ for $|A| = |B|$, $A, B \in \mathcal{B}$. Symmetric fuzzy measures on $\{1, \ldots, N\}$ can be represented in terms of N weights so that $\mu(A) = \sum_{i=1}^{|A|} w_i$. Using a symmetric fuzzy measure, we can represent any OWA operator as a Choquet integral.

Let $\mathbf{a}, \mathbf{b} \in R^n$ be comonotonic. Since $a_{\sigma(i)}+b_{\sigma(i)} = (a+b)_{\sigma(i)}$, $OWA_{\mathbf{w}}(\mathbf{a}+\mathbf{b}) = OWA_{\mathbf{w}}(\mathbf{a})+OWA_{\mathbf{w}}(\mathbf{b})$, that is, $OWA_{\mathbf{w}}$ is comonotonically additive. Therefore we have the next proposition.

Proposition 4. *Let $X := \{1, 2, \ldots, N\}$. For every $OWA_{\mathbf{w}}$, there exists a symmetric fuzzy measure satisfying $\mu(\{N\}) := w_1$ and $\mu(\{1, \ldots, i\}) := w_1 + \cdots + w_i$ for $i = 1, 2, \ldots, N$, such that*

$$OWA_{\mathbf{w}}(\mathbf{a}) = (C)\int \mathbf{a} d\mu$$

for $\mathbf{a} \in R_+^N$.

Since for an arbitrary $\mathbf{a}$ and $\mathbf{b}$ in R^N, it is not always true, $a_{\sigma(i)}+b_{\sigma(i)} = (a+b)_{\sigma(i)}$ for every i, $OWA_{\mathbf{w}}$ is not always additive.

4 Generalized OWA Operator

Definition 12. *Let $\mathcal{F} = (f_1, \ldots, f_N)$ where $f_i : [0,1]^N \to [0,1]$ for $i = 1, \ldots, N$ are N weighting functions such that $\sum_{i=1}^N f_i(x_1, \ldots, x_N) = 1$ for all $(x_1, \ldots, x_N) \in [0,1]^N$, then the generalized OWA (GOWA) is defined as follows:*

$$GOWA_{\mathcal{F}}(a_1, \ldots, a_N) = \sum_{i=1}^{N} w_i a_{\sigma(i)}$$

where σ defines a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$, and where w_i is defined by $w_i = f_i(a_1, \ldots, a_N)$

We have the next proposition, which is similar to Proposition 4.

Proposition 5. *Let $X := \{1, 2, \ldots, N\}$ and let $\mathcal{F} = (f_1, \ldots, f_N)$ where $f_i : [0,1]^N \to [0,1]$ for $i = 1, \ldots, N$ are N weighting functions such that $\sum_{i=1}^N f_i(x_1, \ldots, x_N) = 1$ for all $(x_1, \ldots, x_N) \in [0,1]^N$. For every $GOWA_{\mathcal{F}}$, there exists a symmetric fuzzy measure $\mu_{\mathbf{a}}$ satisfying $\mu_{\mathbf{a}}(\{N\}) := f_1(\mathbf{a})$ and $\mu_{\mathbf{a}}(\{1, \ldots, i\}) := f_1(\mathbf{a}) + \cdots + f_i(\mathbf{a})$ for $\mathbf{a} \in R^N$, $i = 1, 2, \ldots, N$, such that*

$$GOWA_{\mathcal{F}}(\mathbf{a}) = (C)\int \mathbf{a} d\mu_{\mathbf{a}}$$

for $\mathbf{a} \in R_+^N$.

Now, we introduce a generalization of Losonczi's means.

Definition 13. *Let $\mathcal{F} = (f_1, \ldots, f_N)$ where $f_i : [0,1]^N \to [0,1]$ for $i = 1, \ldots, N$ are N weighting functions such that $\sum_{i=1}^N f_i(x_1, \ldots, x_N) = 1$ for all $(x_1, \ldots, x_N) \in [0,1]^N$, then a Generalized Losonczi's mean is defined as follows:*

$$GLM_{\mathcal{F}}(a_1, \ldots, a_N) = \phi^{-1}\Big(\sum_{i=1}^{N} w_i \phi(a_i)\Big)$$

where w_i is defined by $w_i = f_i(a_1, \ldots, a_N)$

Proposition 6. *Let* $X := \{1, 2, \ldots, N\}$, *let* $\pi_i : R \to R^+$ *and* $\phi : R \to R$ *(ϕ strictly increasing with inverse ϕ^{-1}), and let* $\mathcal{F} = (f_1, \ldots, f_N)$ *where* $f_i : [0,1]^N \to [0,1]$ *for* $i = 1, \ldots, N$ *are* N *weighting functions such that* $\sum_{i=1}^N f_i(x_1, \ldots, x_N) = 1$ *for all* $(x_1, \ldots, x_N) \in [0,1]^N$. *Let us define a probability measure* P *on* 2^X *by* $P_{\mathbf{a}}(\{i\}) := \frac{\pi_i(\mathbf{a})}{\sum_{i=1}^N \pi_i(\mathbf{a})}$. *Under these conditions, we have*

$$GLM_{\mathcal{F},\phi}(\mathbf{a}) = \phi^{-1}\big(CS_{P_{\mathbf{a}},\phi}(\mathbf{a})\big)$$

for $\mathbf{a} = (a_1, \ldots, a_N) \in R^N$.

Applying Theorem 1, we obtain the proposition below.

Proposition 7. *Let* $X := \{1, 2, \ldots, N\}$ *and, let* μ *be a fuzzy measure on* 2^X. *Then, the Choquet integral with respect to* μ *is a GLM with* $\phi(x) = x$.

Let φ be a real valued function on a closed interval $[c, d]$. Then, φ is said to be *convex* if

$$\varphi(\lambda x + (1-\lambda)y) \le \lambda\varphi(x) + (1-\lambda)\varphi(y)$$

for $x, y \in [c, d]$, $0 < \lambda < 1$.

In contrast, φ is said to be *concave* if

$$\varphi(\lambda x + (1-\lambda)y) \ge \lambda\varphi(x) + (1-\lambda)\varphi(y)$$

for $x, y \in [c, d]$, $0 < \lambda < 1$.

Let μ be a fuzzy measure on $(X, \mathcal{B})$. Then, since we assume that $\mu(X) = 1$, we have the next inequalities [13].

(1) If φ is convex, then

$$(C)\int \varphi(f)d\mu \ge \varphi\left((C)\int f d\mu\right).$$

(2) If φ is concave, then

$$(C)\int \varphi(f)d\mu \le \varphi\left((C)\int f d\mu\right).$$

Applying Proposition 6, we have the next proposition.

Proposition 8. *Let* $X := \{1, 2, \ldots, N\}$, *let* $\pi_i : R \to R^+$ *and* $\phi : R \to R$ *(ϕ strictly increasing with inverse ϕ^{-1}), and let* $\mathcal{F} = (f_1, \ldots, f_N)$ *where* $f_i : [0,1]^N \to [0,1]$ *for* $i = 1, \ldots, N$ *are* N *weighting functions such that* $\sum_{i=1}^N f_i(x_1, \ldots, x_N) = 1$ *for all* $(x_1, \ldots, x_N) \in [0,1]^N$. *Then, let us define the probability measure* P *on* 2^X *by* $P_{\mathbf{a}}(\{i\}) := \frac{\pi_i(\mathbf{a})}{\sum_{i=1}^N \pi_i(\mathbf{a})}$ *for* $\mathbf{a} = (a_1, \ldots, a_N) \in R^N$. *Under these conditions, we have*

(1) *if* φ *is convex, then*

$$(C)\int \mathbf{a} dP_{\mathbf{a}} \le GLM_{\mathcal{F},\phi}(\mathbf{a}).$$

(2) *if* φ *is concave, then*

$$(C)\int \mathbf{a} dP_{\mathbf{a}} \ge GLM_{\mathcal{F},\phi}(\mathbf{a}).$$

5 On the Relationships between Operators

Proposition 9. *The GOWA operator is equivalent to a GLM with* $\phi(x) = x$.

Proof. To prove this proposition we will consider first the representation of a GOWA by a GLM with $\phi(x) = x$. Then, we will consider the reversal case. In both cases, we consider that the GLM is generated from functions f_i and that the GOWA operator is generated from functions g_i.

A GOWA operator generated by g_i can be represented by a GLM with $\phi(x) = x$ and where f_i is defined as the function g_j such that $i = \sigma(j)$. As f_i is a function of all a_i the selection of the appropriate g_j is possible within f_i. To illustrate this fact, we define this function f_i explictly:

$$f_i(a_1,\ldots,a_N) = \begin{cases} g_1(a_1,\ldots,a_N) & \text{if } a_i \le \min_{j\neq i} a_j \\ g_2(a_1,\ldots,a_N) & \text{if } \min_{j\neq i} a_j \le a_i \le \min_{j_1,j_2\neq i} \max(a_{j_1},a_{j_2}) \\ g_3(a_1,\ldots,a_N) & \text{if } \min_{j_1,j_2\neq i | j_1\neq j_2} \max(a_{j_1},a_{j_2}) \le a_i \le \\ & \qquad \le \min_{j_1,j_2,j_3\neq i} \max(a_{j_1},a_{j_2},a_{j_3}) \\ \ldots & \\ g_N(a_1,\ldots,a_N) & \text{if } \min_{j_1,\ldots,j_{N-1}\neq i | j_r\neq j_s} \max(a_{j_1},\ldots,a_{N-1}) \le a_i \end{cases}$$

A GLM with $\phi(x) = x$ and generated by functions f_i can be represented by a GOWA with g_j defined by the f_i such that $i = \sigma(j)$. Again, this function can be defined explicitly from $a_1, \ldots, a_N$. That is,

$$g_j(a_1,\ldots,a_N) = \begin{cases} f_1(a_1,\ldots,a_N) & \text{if } \min_{i_1,\ldots,i_j} \max(a_{i_1},\ldots,a_{i_j}) = a_1 \\ f_2(a_1,\ldots,a_N) & \text{if } \min_{i_1,\ldots,i_j} \max(a_{i_1},\ldots,a_{i_j}) = a_2 \\ \ldots & \\ f_N(a_1,\ldots,a_N) & \text{if } \min_{i_1,\ldots,i_j} \max(a_{i_1},\ldots,a_{i_j}) = a_N \end{cases}$$

When there exist $a_i = a_j$ for $i \neq j$, the functions g_j should be defined so that all f_i are selected.

□

Yager defined in [24] a generalized OWA. It is one of the generalizations of the OWA operator.

Definition 14. *[24] Let* $\mathcal{F} = (f_1,\ldots,f_N)$ *where* $f_i : [0,1]^N \to [0,1]$ *for* $i = 1,\ldots,N$ *are* N *weighting functions such that* $\sum_{i=1}^{N} f_i(x_1,\ldots,x_N) = 1$ *for all* $(x_1,\ldots,x_N) \in [0,1]^N$*, then Yager's generalized OWA (YGOWA) is defined as follows:*

$$YGOWA_{\mathcal{F}}(a_1,\ldots,a_N) = \sum_{i=1}^{N} w_i a_{\sigma(i)}$$

where σ *defines a permutation of* $\{1,\ldots,N\}$ *such that* $a_{\sigma(i)} \ge a_{\sigma(i+1)}$*, and where* w_i *is defined by* $w_i = f_i(a_{\sigma(1)},\ldots,a_{\sigma(N)})$

It is obvious from the definition that YGOWA operators are generalizations of GOWA and neat OWA. However not all GOWA are neat OWA. For example, using $f_i(a_1, \ldots, a_N) = a_i$ we get the GOWA $\sum(a_i a_{\sigma(i)})/\sum a_{a_i}$, that is not a neat OWA.

From Proposition 9, the next corollary follows.

Corollary 1. *The GLM generalizes the OWA, YGOWA and GOWA.*

Now we consider two particular GLM with some interesting properties.

Definition 15. *Let the Function Unanimous GLM (FUGLM) and the Dimensional Function Unanimous GLM (DFUGLM) be defined as the GLM with the following weighting functions:*

(i) *FUGLM is a GLM with* $f_i(a_1, \ldots, a_N) = f(a_i)$.
(ii) *DFUGLM is a GLM with* $f_i(a_1, \ldots, a_N) = f(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$, *where* f *is a symmetric function.*

The following properties hold for such operators.

Proposition 10. *FUGLM and DFUGLM operators are symmetric (invariant to any permutation of the input data).*

Proof. We first prove that the FUGLMs are symmetric. To do so, we give the expression of FUGLM in terms of a GLM. That is,

$$GLM_{\mathcal{F}}(a_1, \ldots, a_N) = \phi^{-1}\big(\sum_{i=1}^{N} w_i \phi(a_i)\big), \tag{1}$$

where w_i is defined by $w_i = f_i(a_1, \ldots, a_N) = f(a_i)$. Therefore, FUGLM is equivalent to

$$GLM_{\mathcal{F}}(a_1, \ldots, a_N) = \phi^{-1}\big(\sum_{i=1}^{N} f(a_i)\phi(a_i)\big)$$

Naturally, any permutation of the a_i will lead to the same terms $f(a_i)\phi(a_i)$, although in different orders.

Now, we prove that the DFUGLM is also symmetric. Taking into account that $w_i = f_i(a_1, \ldots, a_N) = f(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$ with a symmetric function f, the DFUGLM corresponds to

$$\phi^{-1}\big(\sum_{i=1}^{N} f(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)\phi(a_i)\big) \tag{2}$$

In this case, a permutation s on $\{1, \ldots, N\}$ will lead to:

$$\phi^{-1}\big(\sum_{i=1}^{N} f(a_{s(1)}, \ldots, a_{s(i-1)}, a_{s(i+1)}, \ldots, a_{s(N)})\phi(a_{s(i)})\big) \tag{3}$$

For each a_i in Equation 2, the corresponding term in Equation 3 is the $\phi(a_{s(j)})$ such that $s(j) = i$. Naturally, for both Equations 2 and 3, the elements in f are the same although in a different order. So, as the function f is symmetric, DFUGLM does not depend on the order. □

Corollary 2. *FUGLM and DFUGLM with $\phi(x) = x$ are neat OWA operators.*

6 The Losonczi's OWA

Now, for completeness, we consider Losonczi's OWA.

Definition 16. *Given functions ω_i and ϕ (ϕ monotonic increasing with inverse ϕ^{-1}), the Losonczi's OWA is defined as follows:*

$$LoOWA(a_1, \ldots, a_N) = \phi^{-1}\left(\frac{\sum_{i=1}^{N} \omega_i(a_{\sigma(i)})\phi(a_{\sigma(i)})}{\sum_{i=1}^{N} \omega_i(a_{\sigma(i)})}\right)$$

with σ defined as above.

The following result can be proven for this operator.

Proposition 11. *The LoOWA is a particular case of the GLM and it generalizes the OWA operator.*

Note that in the definition of the LoOWA, the index i refers to the elements according to their position in the ordering (the order is determined by the permutation σ). Therefore, the expression $\omega_i(a_{\sigma(i)})$ assigns the ith weight according to the value that occupies the ith position. Such value corresponds to $a_{\sigma(i)}$. When $\omega_i(a_{\sigma(i)})$ is constant, the LoOWA is a quasi-OWA operator (introduced in [7]).

The LoOWA operator can be used to model situations in which importance is given according to the order an element occupies in the input. This is a case similar to the OWA operator. In contrast to the case of the OWA, now not only the order is important but the value itself. The following two examples illustrate this situation.

- **Robot.** Let us consider a robot with 5 sensors that measure the distance to the nearest object. To avoid collisions, we are interested in giving more importance to small values than to large ones. Nevertheless, when small values are larger than a given threshold (e.g., one meter), the collision problem is not so relevant and the weight of the smallest values is diminished. This situation can be modelled with the following weighting functions $\omega_1(a) = \omega_2(a) = \omega_3(a) = \omega_4(a) = 0.2$ and

$$\omega_5(a) = \begin{cases} 1 & if\ a \leq 1 \\ 0.2 & otherwise \end{cases}$$

– **Compensation.** Let us consider the aggregation of 4 criteria so that compensation of two bad values is allowed, but only when they have passed a given threshold. This situation can be modeled defining ω_i as follows $\omega_4(x) = \omega_3(x) = 1$ if $x < 0.5$ and $\omega_4(x) = \omega_3(x) = 0.25$ otherwise. All other ω_i are defined by $\omega_i(x) = 0.25$ for all x.

In the same way that the WOWA [18] was defined as a generalization of the weighted mean and the OWA operator so that the weights of both operators are taken into account, it is straightforward to define the Losonczi's WOWA (LoWOWA) in terms of the weighting functions π_i and ω_i.

The LoWOWA will allow to incorporate in the two examples above the weights π_i used in the Losonczi'c mean. In the case of the robot, the LoWOWA would permit us to include information on the reliability of the sensors (a reliability that is a function of the data supplied by the sensor itself). In the case of the criteria, the use of LoWOWA would permit us to represent the importance of the criteria (an importance that would depend on the value assigned to the criteria itself).

7 Conclusions

In this paper we have established several relationships between some operators in the literature. We have focused in the family of means defined by Losonczi'c.

Acknowledgment

Partial support by the Generalitat de Catalunya (AGAUR, 2006BE-2 00338, 2005 SGR 00446 and 2005-SGR-00093) and by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged.

References

1. Bajraktarević, M.: Sur une equation fonctionnelle aux valeurs moyennes. Glanik Mat.-Fiz. i Astr., Zagreb 13, 243–248 (1958)
2. Bajraktarević, M.: Sur une genéralisation des moyennes quasilineaire. Publ. Inst. Math. Beograd. 3(17), 69–76 (1963)
3. Bullen, P.S., Mitrinović, D.S., Vasić, P.M.: Means and their Inequalities. D. Reidel Publishing Company (1988)
4. Calvo, T., Mayor, G., Mesiar, R. (eds.): Aggregation Operators. Physica-Verlag (2002)
5. Choquet, G.: Theory of capacities. Ann. Inst. Fourier, Grenoble 5, 131–295 (1955)
6. Dellacherie, C.: Quelques commentaires sur les prolongements de capacités. In: Séminaire de Probabilités 1969/1970, Strasbourg. Lecture Notes in Mathematics, vol. 191, pp. 77–81 (1971)
7. Fodor, J., Marichal, J.-L., Roubens, M.: Characterization of the ordered weighted averaging operators. IEEE Trans. on Fuzzy Systems 3(2), 236–240 (1995)

8. Grabisch, M., Murofushi, T., Sugeno, M. (eds.): Fuzzy Measures and Integrals: Theory and Applications. Physica-Verlag (2000)
9. Losonczi, L.: Über eine neue Klasse von Mittelwerte. Acta Sci. Math (Acta Univ. Szeged) 32, 71–78 (1971)
10. Miranda, P., Grabisch, M.: p-symmetric fuzzy measures. In: Proc. of the IPMU 2002 Conference, Annecy, France, pp. 545–552 (2002)
11. Murofushi, T., Sugeno, M.: An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. Fuzzy Sets and Systems 29, 201–227 (1989)
12. Narukawa, Y., Murofushi, T.: Choquet Stieltjes integral as a tool for decision modeling. Int. J. of Intel. Syst. 23, 115–127 (2008)
13. Narukawa, T.: Distances defined by Choquet integral. In: IEEE International Conference on Fuzzy Systems, London, CD-ROM [#1159] (2007)
14. Riesz, F., Nagy, B.: Functional analysis. Frederick Unger Publishing (1955)
15. Schmeidler, D.: Integral representation without additivity. Proceedings of the American Mathematical Society 97, 253–261 (1986)
16. Sugeno, M., Narukawa, Y., Murofushi, T.: Choquet integral and fuzzy measures on locally compact space. Fuzzy Sets and Systems 99, 205–211 (1998)
17. Sugeno, M.: Theory of fuzzy integrals and its applications, Doctoral Thesis, Tokyo Institute of Technology (1974)
18. Torra, V.: The weighted OWA operator. Int. J. of Intel. Syst. 12, 153–166 (1997)
19. Torra, V., Narukawa, Y.: Modeling decisions: information fusion and aggregation operators. Springer, Heidelberg (2007)
20. Torra, V., Narukawa, Y.: Modelització de decisions: fusió d'informació i operadors d'agregació. UAB Press (2007)
21. van der Waerden, B.L.: Mathematical statistics. Springer, Berlin (1969)
22. Yager, R.R.: On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans. on Systems, Man and Cybernetics 18, 183–190 (1988)
23. Yager, R.R., Filev, D.P.: Parameterized and-like and or-like OWA operators. Int. J. of General Systems 22, 297–316 (1994)
24. Yager, R.R.: Families of OWA operators. Fuzzy Sets and Systems 59, 125–148 (1993)

The Polytope of Fuzzy Measures and Its Adjacency Graph

Elías F. Combarro and Pedro Miranda

[1] Artificial Intelligence Center, University of Oviedo, Spain
[2] Department of Statistics and O.R., Complutense University of Madrid, Spain

Abstract. In this paper we deal with the problem of studying the structure of the polytope of fuzzy measures for finite referential sets. We prove that the diameter of the polytope of fuzzy measures is 3 for referentials of 3 elements or more. We also show that the polytope is combinatorial, whence we deduce that the adjacency graph of fuzzy measures is Hamilton connected if the cardinality of the referential set is not 2. We also give some results about the facets and edges of this polytope. Finally, we treat the corresponding results for the polytope given by the convex hull of monotone boolean functions.

Keywords: Fuzzy measures, monotone boolean functions, diameter, combinatorial polytopes.

1 Introduction and Basic Concepts

Consider a finite referential set $X = \{x_1, ..., x_n\}$ of n elements. The set X plays the role of the set of criteria in Decision Making, players in Game Theory, individuals in Welfare Theory, ... Subsets of X are denoted by capital letters $A, B, ...$ and also by $A_1, A_2, ...$ In order to avoid hard notation, for singletons $\{x_i\}$ we will usually omit braces. The set of subsets of X is denoted by $\mathcal{P}(X)$.

Definition 1. *A* **non-additive measure** *[10] or* **fuzzy measure** *[26] or* **capacity** *[5] over X is a function $\mu : \mathcal{P}(X) \to [0, 1]$ satisfying*

1. *$\mu(\emptyset) = 0$ and $\mu(X) = 1$ (boundary conditions).*
2. *If $A \subseteq B$ then $\mu(A) \leq \mu(B)$ (monotonicity).*

From a mathematical point of view, fuzzy measures constitute a generalization of probability distributions in which we remove additivity and monotonicity is imposed instead. This extension is perfectly justified in many practical situations, in which additivity is too restrictive. For example, in the field of Decision Making, models based on Probability, as those from von Neumann and Morgenstern [27] or Anscombe and Aumann [3] to cite a few, can lead to inconsistencies due to *uncertainty aversion*, as the well-known paradoxes of Ellsberg [11] or Allais [2]. However, models based on fuzzy measures [4,22] are able to handle and interpret these problems.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 74–85, 2008.

Fuzzy measures have been also successfully applied to model problems in Multicriteria Decision Making and Cooperative Games. In the former case, they allow the decision maker to introduce vetoes and favors in the model [13], as well as interactions among the different criteria [14]. In the theory of Cooperative Games, fuzzy measures represent the characteristic function of monotone games, i.e., the pay-off that each coalition can guarantee for itself; indeed, they are related to the Shapley value [23]. Other fields related to fuzzy measures are combinatorics [21], pseudo-boolean functions [16], etc. This versatility of fuzzy measures has led to a huge number of related works, both from a theoretical and from a practical point of view.

Note that we need $2^n - 2$ coefficients in order to define a fuzzy measure. We will denote the set of all fuzzy measures over X by $\mathcal{FM}(X)$. Notice that $\mathcal{FM}(X)$ is a polytope in $\mathbb{R}^{2^n-2}$ (or $\mathbb{R}^{2^n}$ if we include the coordinates for $\mu(\emptyset)$ and $\mu(X)$).

On $\mathcal{FM}(X)$ we can define a partial order given by $\mu_1 \leq \mu_2$ if and only if $\mu_1(A) \leq \mu_2(A)$, $\forall A \subseteq X$. If $\mu_1 \leq \mu_2$ or $\mu_2 \leq \mu_1$ we say that μ_1 and μ_2 are **comparable**.

A problem arising in practice is the identification of the fuzzy measure modelling a certain situation. In [6], we have dealt with the problem of identifying a fuzzy measure from sample information through genetic algorithms [12]. The cross-over operator used in the algorithm was the convex combination, possible as $\mathcal{FM}(X)$ is a polytope; this operator allows a reduction in the complexity of the algorithm. However, the use of this operator has the drawback that the search region is reduced in each iteration. Then, in order to ensure that the searched measure is inside the initial region, we need to consider the extreme points of $\mathcal{FM}(X)$ as the initial population. It has been pointed out in [20] that these extreme points are the set of $\{0,1\}$-valued measures, i.e., the set of monotone boolean functions of n variables except the constant functions 0 and 1 (that do not satisfy the boundary conditions). This extremes are also stack filters [28] and the elements of the free distributive lattice of n generators [24].

Remark that for a $\{0,1\}$-valued measure μ, there are some subsets A satisfying the following conditions:

$$\mu(A) = 1,\ \mu(B) = 1,\ \forall B \supseteq A,\ \mu(C) = 0,\ \forall C \subset A$$

We will call any subset satisfying these conditions a **minimal subset** for μ. The set of minimal subsets also forms the *qualitative Möbius transform* [15]. We will denote the families of minimal subsets by $\mathbf{C}, \mathbf{D}$, and so on. The fuzzy measure whose minimal subsets are the family $\mathbf{C}$ will be denoted by $\mu_{\mathbf{C}}$. Minimal subsets are also known as *minimal true subsets* or *minimal primes*. If we consider the lattice $(\mathcal{P}(X), \cup, \cap)$, then a minimal subset for a $\{0,1\}$-valued measure μ can be equivalently defined as a subset of X such that $\mu(A) = 1$ and whose principal filter $\mathcal{F}_A$ and principal ideal $\mathcal{I}_A$ (see [25]) satisfy

$$\mu(B) = 1,\ \forall B \in \mathcal{F}_A,\ \mu(B) = 0,\ \forall B \in \mathcal{I}_A \backslash \{A\}.$$

An example of fuzzy measure that we will use later is the one in which there is only a minimal subset $A \subseteq X$, $A \neq \emptyset$. This measure is given by

$$u_A(B) := \begin{cases} 1 \text{ if A} \subseteq \text{B} \\ 0 \text{ otherwise} \end{cases}$$

These fuzzy measures are the vertices of a special convex class of fuzzy measures called *belief functions*, that appear in the Theory of Evidence [9]. For $\emptyset$, we define the fuzzy measure $u_\emptyset$ by

$$u_\emptyset(B) := \begin{cases} 1 \text{ if B} \neq \emptyset \\ 0 \text{ if B} = \emptyset \end{cases}$$

Note that $u_\emptyset$ follows a different structure to any other u_A; indeed, its minimal subsets are $\{x_1\}, ..., \{x_n\}$ and it is not a belief function.

The set of minimal subsets of a $\{0,1\}$-valued measure determine an **antichain** (collections of sets which are pairwise uncomparable with respect to inclusion, see [1]). Then, the number of vertices of the polytope $\mathcal{FM}(X)$ is the number of different antichains on X.

The number of antichains of a set of cardinality n is known as the n-th Dedekind number, denoted D_n [8]. The first Dedekind numbers are given in Table 1 [29].

Table 1. Number of vertices of $\mathcal{FM}(X)$

n	Dedekind numbers
1	1
2	4
3	18
4	166
5	7579
6	7828352
7	2414682040996
8	56130437228687557907786

In the values given in this table we have excluded the empty antichain and the antichain which contains only the empty set, as these cases lead the constant functions 0 and 1, that are not fuzzy measures. The form of the general term of this sequence is known [17] but is, however, inefficient. Anyway, from the quantities in Table 1, it can be seen that we cannot use the vertices of $\mathcal{FM}(X)$ as initial population when n is big (and $n = 6$ is big!). Then, it is necessary to consider only a subset of the set $\{0,1\}$-valued measures; however, this set should be chosen carefully in order to cover a big part of $\mathcal{FM}(X)$ with a reduced number of vertices.

Related to this problem, we have already studied in [18] which are the *isometries* (functions maintaining distances) on $\mathcal{FM}(X)$ and the set of fuzzy measures remaining invariant for any isometric transformation.

We have also characterized whether two vertices of $\mathcal{FM}(X)$ are adjacent in [7]. More concretly, we have proved the following:

Proposition 1. *If μ_1 and μ_2 are adjacent vertices of $\mathcal{FM}(X)$, then $\mu_1 < \mu_2$ or $\mu_2 < \mu_1$.*

However, this condition is not sufficient.

Definition 2. *Let $\mathbf{C}$ and $\mathbf{D}$ be two collections of minimal sets (antichains). We say that $\mathbf{D}$ is $\mathbf{C}-$**decomposable** if there exists a partition of $\mathbf{D}$ in two non-empty subsets $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A} \not\subseteq \mathbf{C}$ and $\mathbf{B} \not\subseteq \mathbf{C}$, and if $A \in \mathbf{A}$ and $B \in \mathbf{B}$, then there exists $C \in \mathbf{C}$ such that $C \subseteq A \cup B$.*

The following can be proved:

Theorem 1. *Let $\mu_{\mathbf{D}}, \mu_{\mathbf{C}}$ be two vertices of $\mathcal{FM}(X)$ such that $\mu_{\mathbf{D}} > \mu_{\mathbf{C}}$. Then, $\mu_{\mathbf{D}}$ and $\mu_{\mathbf{C}}$ are adjacent vertices of $\mathcal{FM}(X)$ if and only if $\mathbf{D}$ is not $\mathbf{C}$-decomposable.*

Let us now turn to the convex hull of monotone Boolean functions.

Lemma 1. *The constant functions 0 and 1 are adjacent to any other monotone Boolean function.*

Thus, Theorem 1 can be extended to this polytope, thus obtaining:

Corollary 1. *Let $\mu_{\mathbf{D}}, \mu_{\mathbf{C}}$ be two vertices of the set of the convex hull of monotone Boolean functions such that $\mu_{\mathbf{D}} > \mu_{\mathbf{C}}$. Then, $\mu_{\mathbf{D}}$ and $\mu_{\mathbf{C}}$ are adjacent vertices of this polytope if and only if $\mu_{\mathbf{D}}$ is not $\mathbf{C}$-decomposable.*

In this paper, we aim to study more properties about the polytope $\mathcal{FM}(X)$. These properties could be interesting in the problem of identification of fuzzy measures. Moreover, they might shed light on the structure of $\mathcal{FM}(X)$, and could be useful in the search of families of fuzzy measures with additional properties. Besides, many of the results obtained for $\mathcal{FM}(X)$ can be extended to the convex hull of monotone boolean functions. First, we study the edges of the polytope; we show that the probability of two measures are adjacent decreases when the cardinality of X grows. We also study the facets of the polytope. We show that the diameter of $\mathcal{FM}(X)$ is 3 for $|X| > 2$. Finally, we show that the graph of $\mathcal{FM}(X)$ is Hamilton connected when $|X| \neq 2$. Corresponding results for monotone Boolean functions are also stated. We finish with the conclusions and open problems. Detailed proofs of these results can be found in [7].

2 More Results about the Adjacency

In this section, we study other properties of $\mathcal{FM}(X)$ related to adjacency. First, we show that given two monotone boolean functions, it is quite uncommon that they are adjacent.

Lemma 2. *The probability of two monotone Boolean functions of n variables taken at random being comparable (with the order relation) tends to zero when n tends to infinity.*

As an immediate consequence of this lemma and Proposition 1, we have the following result.

Corollary 2. *The probability of two monotone Boolean functions of n variables taken at random being adjacent tends to zero when n tends to infinity.*

This result also applies to the extremes of $\mathcal{FM}(X)$, since the monotone boolean functions 0 and 1 are adjacent to any other monotone Boolean function (Lemma 1). This property can be interesting in order to determine subfamilies of vertices covering a big region of $\mathcal{FM}(X)$ and with a small cardinality. For this, it makes sense to consider families of vertices that are not adjacent to each other; these families are known as *stable sets* of vertices.

Some values of the probability of two vertices of $\mathcal{FM}(X)$ being adjacent are given in the following table. In it, we are considering possible pairs of vertices **with** replacement, so that for a vertex μ, the pair (μ, μ) is possible; for two different vertices μ_1, μ_2, pairs (μ_1, μ_2) and (μ_2, μ_1) are considered once and not twice. Similar results can be obtained random choice without replacement, as Lemma 2 and Corollary 2 also hold in this situation.

Table 2. Probabilities of two vertices being adjacent for different cardinalities

n	2	3	4	5
Probability	0.5	0.45062	0.23015	0.07189

3 The Diameter of $\mathcal{FM}(X)$

Definition 3. *Given a polytope $\mathcal{F}$, the* **graph (of adjacency) of** *$\mathcal{F}$ is defined by the graph whose nodes are the vertices of $\mathcal{F}$ and whose edges join two nodes if and only if they are adjacent.*

Consider the graph of $\mathcal{FM}(X)$. Let us define the distance $d(\mu_1, \mu_2)$ between two extremes μ_1, μ_2 of $\mathcal{FM}(X)$ as the number of edges of the shortest path between them in this graph. We have shown in the previous corollary that the distance between two vertices is greater than 1 with probability tending to 1 when $|X|$ tends to infinity. Now, we will study the *diameter* of the graph, i.e. the maximum distance between two extremes.

Lemma 3. *If $|X| > 2$ and μ_1, μ_2 are two extremes of $\mathcal{FM}(X)$ such that both μ_1 and μ_2 are either adjacent to u_X or to $u_\emptyset$ (not necessarily both adjacent to the same) then $d(\mu_1, \mu_2) \leq 3$.*

Proof: It is obvious because u_X and $u_\emptyset$ are adjacent when $|X| > 2$. □

Note that for $|X| = 2$, u_X and $u_\emptyset$ are not adjacent as $\{\{x_1\}, \{x_2\}\}$ determine a $\{X\}-$decomposition (see Theorem 1). For $|X| = 1$, $u_X = u_\emptyset$.

Let us now study the distance when we consider extremes that are not adjacent to u_X nor $u_\emptyset$. We start characterizing these extremes.

Lemma 4. *If $|X| > 3$ and μ is an extreme of $\mathcal{FM}(X)$ which is neither adjacent to u_X nor to $u_\emptyset$, then there exists $x_i \in X$ such that μ can be written as*

$$\mu = \left(\bigvee_{j \neq i} u_{\{x_i, x_j\}} \right) \vee u_{X \setminus \{x_i\}}.$$

Lemma 5. *Suppose $|X| > 3$. If $\mu_i := \left(\bigvee_{j \neq i} u_{\{x_i, x_j\}} \right) \vee u_{X \setminus \{x_i\}}$, $\forall x_i \in X$, then $d(\mu_i, \mu) \leq 3$ for any μ extreme point of $\mathcal{FM}(X)$.*

Sketch of proof: It is easy to check that, for all i and j, the extremes μ_i and μ_j are both adjacent to the extreme $u_{X \setminus \{x_i\}} \vee u_{X \setminus \{x_j\}}$. Thus $d(\mu_i, \mu_j) = 2$ since μ_i and μ_j are not adjacent (neither $\mu_i > \mu_j$ nor $\mu_j > \mu_i$). Now, remark that μ_i is also adjacent to $u_{X \setminus \{x_i\}}$ which, in turn, is adjacent to u_X and to $u_\emptyset$. Hence, $d(\mu_i, u_X) = d(\mu_i, u_\emptyset) = 2$ and the result follows from the previous lemma. □

Thus, the diameter of $\mathcal{FM}(X)$ when $|X| > 3$ is at most 3. In next result we will prove that it is indeed 3.

Lemma 6. *Assume $|X| > 3$. There exist vertices in $\mathcal{FM}(X)$ whose distance is at least 3.*

Sketch of proof: It suffices to consider μ the extreme point of $\mathcal{FM}(X)$ whose minimal subsets are $\{x_1, x_2\}$ and $X \setminus \{x_1, x_2\}$ and μ' the extreme point whose minimal subsets are $\{A_i, B_i\}_{i=3}^n$ with $A_i := \{x_1, x_i\}$, $B_i := \{x_2, x_i\}$. Note that they are not comparable, whence $d(\mu, \mu') > 1$. It can be proved that it is not possible to find another extreme point μ'' being adjacent to both μ and μ'. Therefore, $d(\mu, \mu') \geq 3$. □

Joining all these results, we can state the following Theorem.

Theorem 2. *If $|X| \geq 3$, then the diameter of the graph of adjacency of the extremes of $\mathcal{FM}(X)$ is exactly 3.*

For $|X| = 3$, we can study the distance of two extreme points of $\mathcal{FM}(X)$ in Figure 1 (which has been drawn with the help of the Pigale computer program[1]). We represent each vertex by means of its minimal sets. Also, we use i instead of x_i. Thus, $\{1, 2\}, \{3\}$ stands for $(u_{x_1} \wedge u_{x_2}) \vee u_{x_3}$, and so on.

In this figure, we can define two different families of vertices: let us denote by $\mathcal{V}_1$ the nine vertices whose minimal elements are $\{\{i\}\}$, $\{\{i, j\}\}$ or $\{\{i\}, \{j\}\}$ for

[1] PIGALE: Public Implementation of a Graph Algorithm Library and Editor, H. de Fraysseix and P. Ossona de Mendez. http://pigale.sourceforge.net/

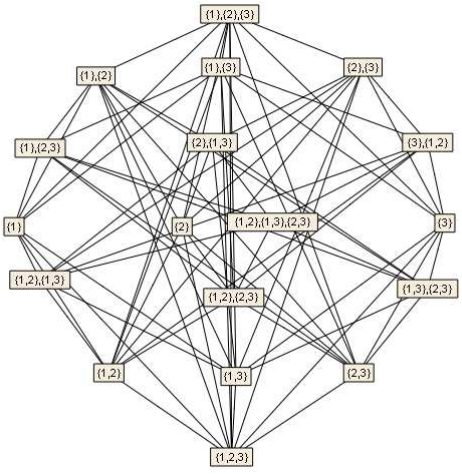


Fig. 1. Adjacency of vertices of $\mathcal{FM}(X)$ for $|X| = 3$

$i, j = 1, 2, 3$; the rest of vertices except the three whose minimal elements are $\{\{1,2,3\}\}, \{\{1\},\{2\},\{3\}\}$ and $\{\{1,2\},\{1,3\},\{2,3\}\}$ are included in $\mathcal{V}_2$. Vertices in $\mathcal{V}_1$ are simultaneously adjacent to $\{\{1,2,3\}\}$ and $\{\{1\},\{2\},\{3\}\}$, and vertices in $\mathcal{V}_2$ are adjacent to $\{\{1,2\},\{1,3\},\{2,3\}\}$. Thus, all vertices in $\mathcal{V}_2$ are at distance 2 or less among them, and all vertices in $\mathcal{V}_1$ are at distance 2 or less among them. Also, any vertex in $\mathcal{V}_1$ is adjacent to at least one vertex in $\mathcal{V}_2$ (and vice versa), so vertices in $\mathcal{V}_1$ and $\mathcal{V}_2$ are at distance 3 at most. Since $\{\{1,2\},\{1,3\},\{2,3\}\}$ is adjacent to all vertices in $\mathcal{V}_2$, it is at distance at most 2 of all vertices in $\mathcal{V}_1$. Similarly, $\{\{1,2,3\}\}$ and $\{\{1\},\{2\},\{3\}\}$ are at distance at most 2 of all vertices in $\mathcal{V}_2$. Finally, the distance between $\{\{1,2\},\{1,3\},\{2,3\}\}$ and $\{\{1,2,3\}\}$ is 3, the distance between $\{\{1,2\},\{1,3\},\{2,3\}\}$ and $\{\{1\},\{2\},\{3\}\}$ is again 3, and the distance between $\{\{1,2,3\}\}$ and $\{\{1\},\{2\},\{3\}\}$ is 1.

For $|X| = 1$, the diameter of $\mathcal{FM}(X)$ is 0, and for $|X| = 2$, it is 2. Notice that in the convex hull of all monotone boolean functions, Lemma 6 does not hold, as any measure is adjacent to the constant functions 0 and 1. In this case, the diameter is always 2, except for $|X| = 1$, whose diameter is 1 (note that there are three measures in this case: the constant 0, the constant 1 and the function whose value is 0 for $\emptyset$ and 1 for X; all are adjacent to each other).

Moreover, the following holds:

Theorem 3. *The probability that an extreme point of $\mathcal{FM}(X)$ chosen at random is adjacent to $u_\emptyset$ tends to 1 when $n = |X|$ tends to infinity.*

From this theorem, we can derive the following results:

Corollary 3. *The probability that an extreme point of $\mathcal{FM}(X)$ chosen at random is adjacent to u_X tends to 1 when $|X|$ tends to infinity.*

Corollary 4. *The probability that two extreme points μ, μ' of $\mathcal{FM}(X)$ chosen at random satisfy $d(\mu, \mu') = 2$ tends to 1 when $n = |X|$ tends to infinity.*

As a consequence, the probability that two extreme points μ, μ' of $\mathcal{FM}(X)$ chosen at random satisfy $d(\mu, \mu') = 3$, i.e. they are at maximal distance, tends to 0 when $|X|$ tends to infinity. Table 3 shows the probability of the different distances for different values of n. Distance 0 appears when $\mu = \mu'$.

Table 3. Probabilities of different distances for $\mathcal{FM}(X)$

n	$P(d(\mu,\mu')=0)$	$P(d(\mu,\mu')=1)$	$P(d(\mu,\mu')=2)$	$P(d(\mu,\mu')=3)$	D_n	Pairs at distance 1
2	0.25	0.5	0.25	0	4	8
3	0.05555	0.45062	0.44444	0.04938	18	146
4	0.00602	0.23015	0.75091	0.01292	166	6342
5	0.00013	0.07189	0.92797	$2.78 * 10^{-6}$	7579	4129670

From the last two columns of this table we can see that the number of adjacent vertices for each extreme point is not the same except for the case $|X| = 2$. Otherwise, the number pairs that are at distance one should be divisible by the number of vertices D_n, and this does not hold for $n \geq 3$.

Notice that if we embed $\mathcal{FM}(X)$ in $\mathcal{FM}(X \cup \{x_{n+1}\})$, the distance between two measures $\mu_1, \mu_2 \in \mathcal{FM}(X)$ and the distance of the corresponding embedded measures $\mu_1', \mu_2' \in \mathcal{FM}(X \cup \{x_{n+1}\})$, where μ_1' and μ_2' are defined by

$$\mu_i'(A) := \begin{cases} \mu_i(A) & \text{if } A \subseteq X \\ \mu_i(A \backslash \{x_{n+1}\}) & \text{if } x_{n+1} \in A \end{cases}$$

can be different.

On the other hand, it can be seen that the minimal subsets for μ_i and μ_i' are the same. Now, the following can be shown:

Lemma 7. *Consider μ_1 and μ_2 two monotone boolean functions such that there exists $x_i \in X$ satisfying that it does not belong to any minimal subset of μ_1 and μ_2. Then, $d(\mu_1, \mu_2) \leq 2$.*

Proof: If μ_1 and μ_2 are adjacent, then $d(\mu_1, \mu_2) = 1$ and we are done. Suppose then that they are not adjacent; then, they are adjacent to u_X as the families of minimal subsets cannot be $\{X\}$-decomposed. □

As a corollary we obtain the following result:

Corollary 5. *If we consider the polytope $\mathcal{FM}(X)$ as a sub-polytope of $\mathcal{FM}(X \cup \{x_{n+1}\})$, then the measures in $\mathcal{FM}(X)$ whose distance is 3 are at distance 2 in $\mathcal{FM}(X \cup \{x_{n+1}\})$.*

4 Some Results about the Facets of $\mathcal{FM}(X)$

Let us now study the facets of $\mathcal{FM}(X)$. More concretly, we address now the problem of obtaining the number of vertices in a facet of the polytope. The facets of a polytope are given by the points satisfying with equality a non-dummy constraint. The constraints defining $\mathcal{FM}(X)$ are

$$\mu(A) - \mu(A \setminus x_i) \geq 0, \forall \emptyset \neq A \subseteq X, x_i \in A.$$

Lemma 8. *There are not dummy constraints in* $\mathcal{FM}(X)$.

Then, given the facet defined by $\mu(A) = \mu(A \backslash x_i)$, for some $\emptyset \neq A \subseteq X$ and $x_i \in A$, the vertices in it are those satisfying either $\mu(A) = \mu(A \backslash x_i) = 0$ or $\mu(A) = \mu(A \backslash x_i) = 1$, except when A is a singleton, where the vertices are given by those measures such that $\mu(A) = 0$, and when $A = X$, where the vertices are those measures satisfying $\mu(X \backslash x_i) = 1$.

Lemma 9. *Consider* $A, B \subset X$ *such that* $|A| = |B|$. *Then, the number of vertices in any facet defined by* $\mu(A) = \mu(A \backslash x_i), x_i \in A$ *is the same as in any facet defined by* $\mu(B) = \mu(B \backslash x_j), x_j \in B$.

Lemma 10. *Consider* $A, B \subset X$ *such that* $|B| = |X| - |A| + 1$. *Then, the number of vertices in any facet defined by* $\mu(A) = \mu(A \backslash x_i), x_i \in A$ *is the same as in any facet defined by* $\mu(B) = \mu(B \backslash x_j), x_j \in B$.

For $|A| = 1$, the following can be stated:

Lemma 11. *Suppose* $A = \{x_i\}$. *The number of vertices in the corresponding facet of* $\mathcal{FM}(X)$ *is* $D_n - D_{n-1} - 1$.

Joining Lemma 10 and Lemma 11, we conclude that the number of vertices in the facet defined by $\mu(X \backslash x_i) = 1$ is $D_n - D_{n-1} - 1$.

From Lemma 9, we know that the number of vertices in a facet only depends on the cardinality of the set defining it. Let us denote $A_i := \{x_1, \ldots, x_i\}$, $i = 1, \ldots, n$ and $A_0 := \emptyset$. We will call the facet defined by $\mu(A_i) = \mu(A_{i-1})$, $i = 1, \ldots, n$, the A_i**-facet**. The number of vertices in the A_i-facet will be denoted by F_i. From Lemmas 10 and 11, we already know that $F_1 = F_n = D_n - D_{n-1} - 1$ and that $F_i = F_{n-i+1}$ for $i = 1, \ldots, n$. The following result shows further relationship between the F_i numbers and the Dedekind numbers.

Theorem 4. *For every* n *it holds*

$$F_1 + F_2 + \ldots + F_n = (n-1)D_n.$$

In Table 4 we give the exact values of the F_i numbers for some values of $|X|$. The values for $|X| = 2, 3, 4$ can be deduced from the previous results. A computer program was implemented to explicitly count the vertices in each facet and supply the values for $|X| = 5, 6$.

Table 4. Values of F_i

	F_1	F_2	F_3	F_4	F_5	F_6
$\|X\| = 2$	2	2	-	-	-	-
$\|X\| = 3$	13	10	13	-	-	-
$\|X\| = 4$	147	102	102	147	-	-
$\|X\| = 5$	7412	5739	4014	5739	7412	-
$\|X\| = 6$	7820772	7240284	4509824	4509824	7240284	7820772

5 $\mathcal{FM}(X)$ Is Combinatorial

Finally, let us study a final property of the graph of $\mathcal{FM}(X)$.

Definition 4. *[19] A convex polytope is* **combinatorial** *if it satisfies the following conditions:*

- *All its vertices are $\{0,1\}$-valued.*
- *If vertices x and y are not adjacent, then there exist two other vertices u and v such that $x + y = u + v$.*

The following can be proved:

Proposition 2. *$\mathcal{FM}(X)$ is a combinatorial polytope.*

By Lemma 1, the result also holds for the convex hull of all monotone boolean functions.

A graph is **Hamilton connected** if every pair of distinct nodes is joined by a Hamilton path. For combinatorial polyhedra, the following can be shown:

Theorem 5. *[19] Let G be the graph of a combinatorial polytope. Then G is either a hypercube or else is Hamilton connected.*

Then, the following holds:

Corollary 6. *The graph of $\mathcal{FM}(X)$ is Hamilton connected for $|X| \neq 2$.*

Proof: Remark that $\mathcal{FM}(X)$ is a hypercube for $n = 1$ and $n = 2$. For $n = 1$, the graph of $\mathcal{FM}(X)$ is trivially Hamilton connected. Moreover, it is easy to see that the graph is not Hamilton connected for $n = 2$.

If $n > 2$, it suffices from Theorem 5 to show that the graph of $\mathcal{FM}(X)$ is not a hypercube. But this holds, as the hypercube of dimension n has diameter n; from Theorem 3, the diameter of $\mathcal{FM}(X)$ is 3 if $|X| \geq 3$. Thus, if $\mathcal{FM}(X)$ is a hypercube, it must be the 3-dimensional one. On the other hand, the hypercube of dimension 3 has 8 vertices, and this is not a Dedekind number. □

For the convex hull of monotone Boolean functions, we can adapt the previous proof to conclude that the corresponding graph is Hamilton connected for any cardinality. For $|X| = 2$, note that it is not a hypercube, as there are two more functions than in the case of fuzzy measures.

6 Conclusions and Open Problems

In this paper we have studied some properties of the polytope $\mathcal{FM}(X)$. Many of the results that we have obtained apply also to the convex hull of monotone Boolean functions.

We have studied the edges and the facets of $\mathcal{FM}(X)$. We have proved that the probability of two vertices chosen at random are adjacent tends to 0 when the cardinality of X grows. For the facets, we have proved that the number of vertices in a facet depends on the cardinality of the subset defining the facet. We have also shown that there seems to be a duality relationship for the facets.

We have proved that $\mathcal{FM}(X)$ has diameter 3 when $|X| \geq 3$ and that two vertices chosen at random are at distance 2 with probability tending to 1 when $|X|$ tends to infinity.

Finally, we have shown that $\mathcal{FM}(X)$ is combinatorial, whence we have concluded that the graph of this polytope is Hamilton connected for $|X| \neq 2$.

We think that these results can shed light on the structure of $\mathcal{FM}(X)$ and the convex hull of monotone Boolean functions. Moreover, these results could be interesting in the problem of identifying a fuzzy measure. For example, if we consider the identification through genetic algorithms, we know that the set of vertices cannot be used as the initial population [6]. However, it could be interesting to study the performance of the algorithm if we consider a stable subset of vertices.

Finally, there are some problems that could be interesting to study related to $\mathcal{FM}(X)$. In this sense, it might be interesting to study more deeply the number of vertices in a facet. Another problem is to determine the number of adjacent vertices to a given extreme point of $\mathcal{FM}(X)$. For this, more research is needed.

Acknowledgements

This research has been supported in part by grant numbers MTM2007-61193 and CAM-UCM910707, and by MEC and FEDER grant TIN2007-61273. We would also like to thank an anonymous referee for interesting references.

References

1. Aigner, M.: Combinatorial Theory. Springer, Heidelberg (1979)
2. Allais, M.: Le comportement de l'homme rationnel devant le risque: critique des postulats de l'école américaine. Econometrica (21), 503–546 (1953) (in French)
3. Anscombe, F.J., Aumann, R.J.: A definition of subjective probability. The Annals of Mathematical Statistics (34), 199–205 (1963)
4. Chateauneuf, A.: Modelling attitudes towards uncertainty and risk through the use of Choquet integral. Annals of Operations Research (52), 3–20 (1994)
5. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier (5), 131–295 (1953)
6. Combarro, E.F., Miranda, P.: Identification of fuzzy measures from sample data with genetic algorithms. Computers and Operations Research 33(10), 3046–3066 (2006)

7. Combarro, E.F., Miranda, P.: On the polytope of non-additve measures. Fuzzy Sets and Systems 159(16), 2145–2162 (2008)
8. Dedekind, R.: Über Zerlegungen von Zahlen durch ihre grössten gemeinsamen Teiler. Festschrift Hoch Braunschweig Ges. Werke II, 103–148 (1897) (in German)
9. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statististics (38), 325–339 (1967)
10. Denneberg, D.: Non-additive measures and integral. Kluwer Academic, Dordrecht (1994)
11. Ellsberg, D.: Risk, ambiguity, and the Savage axioms. Quart. J. Econom. (75), 643–669 (1961)
12. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading (1989)
13. Grabisch, M.: Alternative representations of discrete fuzzy measures for decision making. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 5, 587–607 (1997)
14. Grabisch, M.: k-order additive discrete fuzzy measures and their representation. Fuzzy Sets and Systems (92), 167–189 (1997)
15. Grabisch, M.: The Möbius function on symmetric ordered structures and its application to capacities on finite sets. Discrete Mathematics 287(1-3), 17–34 (2004)
16. Hammer, P.L., Holzman, R.: On approximations of pseudo-boolean functions. Zeitschrift für Operations Research. Mathematical Methods of Operations Research (36), 3–21 (1992)
17. Kisielewicz, A.: A solution od Dedekind's problem on the number of isotone Boolean functions. J. reine angew. Math. (386), 139–144 (1988)
18. Miranda, P., Combarro, E.F.: On the structure of some families of fuzzy measures. IEEE Transactions on Fuzzy Systems 15(6), 1068–1081 (2007)
19. Naddef, D., Pulleyblank, W.R.: Hamiltonicity and Combinatorial Polyhedra. Journal of Combinatorial Theory Series B 31, 297–312 (1981)
20. Radojevic, D.: The logical representation of the discrete Choquet integral. Belgian Journal of Operations Research, Statistics and Computer Science 38(2–3), 67–89 (1998)
21. Rota, G.C.: On the foundations of combinatorial theory I. Theory of Möbius functions. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete (2), 340–368 (1964)
22. Schmeidler, D.: Integral representation without additivity. Proceedings of the American Mathematical Society 97(2), 255–261 (1986)
23. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the theory of Games. Annals of Mathematics Studies, vol. II, pp. 307–317. Princeton University Press, Princeton (1953)
24. Shmulevich, I., Selke, T.M., Coyle, E.J.: Stack Filters and Free Distributive Lattices. In: Proceeding of the 1995 IEEE Workshop on Nonlinear Signal Processing, Halkidiki, Greece, June 1995, pp. 927–930 (1995)
25. Skornjakov, L.A.: Elements of lattice theory. Adam Hilger Ltd. (1977)
26. Sugeno, M.: Theory of fuzzy integrals and its applications. Ph.D thesis, Tokyo Institute of Technology (1974)
27. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behaviour. Princeton University Press, New Jersey (1944)
28. Wendt, P., Coyle, E., Gallagher, N.J.: Stack filters. IEEE Transactions on Acoustics, Speech and Signal Processing, 898–911 (1986)
29. Wiedemann, D.: A computation of the eighth Dedekind number. Order 8, 5–6 (1991)

On Consensus Measures in Fuzzy Group Decision Making

F.J. Cabrerizo[1], S. Alonso[2], I.J. Pérez[1], and E. Herrera-Viedma[1]

[1] Dept. of Computer Science and Artificial Intelligence,
University of Granada, 18071 Granada, Spain
{cabrerizo,ijperez,viedma}@decsai.ugr.es
[2] Dept. of Software Engineering,
University of Granada, 18071 Granada, Spain
zerjioi@ugr.es

Abstract. In group decision making problems, a natural question in the consensus process is how to measure the closeness among experts' opinions in order to obtain the consensus level. To do so, different approaches have been proposed. For instance, several authors have introduced hard consensus measures varying between 0 (no consensus or partial consensus) and 1 (full consensus or complete agreement). However, consensus as a full and unanimous agreement is far from being achieved in real situations. So, in practice, a more realistic approach is to use softer consensus measures, which assess the consensus degree in a more flexible way. The aim of this paper is to identify the different existing approaches to compute soft consensus measures in fuzzy group decision making problems. Additionally, we analyze their advantages and drawbacks and study the future trends.

Keywords: group decision making, consensus process, soft consensus measures.

1 Introduction

In a classical Group Decision Making (GDM) situation there is a problem to solve, a solution set of possible alternatives, and a group of two or more experts, who express their opinions about this solution set of alternatives. These problems consist in multiple individuals interacting to reach a decision. Each expert may have unique motivations or goals and may approach the decision process from a different angle, but have a common interest in reaching eventual agreement on selecting the "best" option(s) [5,8,24]. To do this, experts have to express their preferences by means of a set of evaluations over a set of alternatives.

In a GDM problem, there are two processes to apply before obtaining a final solution [9,13,14,15,18,22,23]: *the consensus process* and *the selection process* (see Figure 1). The former consists in how to obtain the maximum degree of consensus or agreement between the set of experts on the solution set of alternatives. Normally, the consensus process is guided by a human figure called moderator [6,9,22] who is a person that does not participate in the discussion but knows

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 86–97, 2008.

the agreement in each moment of the consensus process and is in charge of supervising and addressing the consensus process toward success, i.e., to achieve the maximum possible agreement and to reduce the number of experts outside of the consensus in each new consensus round. The latter refers to how to obtain the solution set of alternatives from the opinions on the alternatives given by the experts. Clearly, it is preferable that the set of experts achieves a great agreement among their opinions before applying the selection process.

At the beginning of every GDM problem, the set of experts have diverging opinions, then, the consensus process is applied, and in each step, the degree of existing consensus among experts' opinions is measured. If the consensus degree is lower than a specified threshold, the moderator would urge experts to discuss their opinions further in an effort to bring them closer. Otherwise, the moderator would apply the selection process in order to obtain the final consensus solution to the GDM problem. In such a way, a GDM problem may be defined as a dynamic and iterative process, in which the experts, via the exchange of information and rational arguments, agree to update their opinions until they become sufficiently similar, and then, the solution alternative(s) is/are obtained. In this paper, we focus on the consensus process.

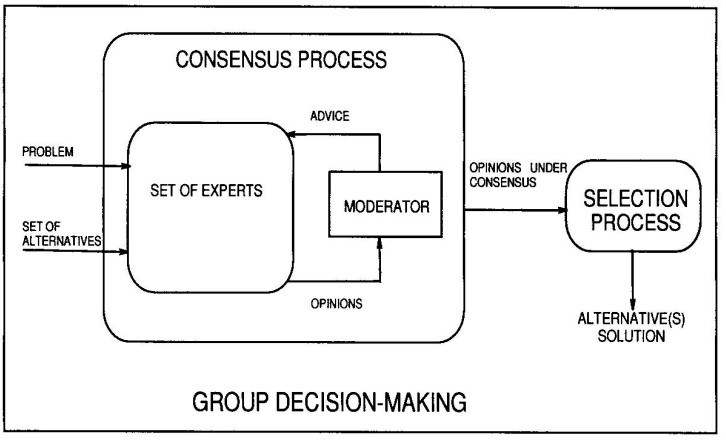


Fig. 1. Resolution process of a GDM problem

A natural question in the consensus process is how to measure the closeness among experts' opinions in order to obtain the consensus level. To do so, different approaches have been proposed. For instance, several authors have introduced *hard consensus measures* varying between 0 (no consensus or partial consensus) and 1 (full consensus o complete agreement) [2,3,26,27]. Thus, using *hard consensus measures*, in [2,3], a distance from consensus as a difference between some average preference matrix and one of several possible consensus preference matrices is determined. In [26] some measures of attitudinal similarity between individuals that is an extension of the classical Tanimoto coefficient are derived.

And, in [27], a consensus measure based on a-cuts of the respective individual fuzzy preference matrices is derived. However, consensus as a full and unanimous agreement is far from being achieved in real situations, and even if it is, in such a situation, the consensus reaching process could be unacceptably costly. So, in practice, a more realistic approach is to use *softer consensus measures* [19,20,21], which assess the consensus degree in a more flexible way, and therefore reflect the large spectrum of possible partial agreements, and guide the consensus process until widespread agreement (not always full) is achieved among experts. The soft consensus measures are based on the concept of coincidence [11], measured by means of similarity criteria defined among experts' opinions.

The aim of this paper is to identify the different existing approaches in the literature to compute soft consensus measures in fuzzy GDM problems and analyze their advantages and drawbacks. To do so, firstly, we identify three different coincidence criteria to compute soft consensus measures: *strict coincidence among preferences*, *soft coincidence among preferences* and *coincidence among solutions*. Then, we analyze their application in consensus processes of fuzzy GDM problems and study their drawbacks and advantages. Furthermore, we describe the new advanced approaches, which use the above coincidence criteria, allowing to generate recommendations to help experts change their opinions in order to obtain the highest degree of consensus possible and adapt the consensus process to increase the agreement and to reduce the number of experts' preferences that should be changed after each consensus round.

In order to do this, the paper is set up as follows. In Section 2, we present the different approaches proposed in the literature to obtain soft consensus measures in fuzzy GDM problems. In Section 3, we discuss their advantages and drawbacks. The new advanced approaches are shown in Section 4. Finally, some concluding remarks are pointed out in Section 5.

2 Approaches to Obtain Soft Consensus Measures in Fuzzy GDM Problems

In this section, we analyze different existing approaches in the literature to obtain soft consensus measures in a fuzzy GDM problem.

As aforementioned, soft consensus measures are based on the coincidence concept [11], i.e., measuring the existing coincidence among expert's opinions by means of similarity criteria. In the literature, we identify three different approaches of coincidence concept to compute soft consensus measure:

1. *Consensus models based on strict coincidence among preferences*. In this case, similarity criteria among preferences are used to compute the coincidence concept. It is assumed only two possible results: the total coincidence (value 1) or null coincidence (value 0). Some examples of this approach are the following: In [19], assuming fuzzy preference relations to represent experts' preferences, the first consensus model based on strict coincidence was defined. Given a particular alternative pair and two experts, if their preferences are equal, then they are in agreement (value 1), and otherwise they

are in disagreement (value 0). Then consensus measures are calculated across the global set of the alternatives in a hierarchical pooling process from the coincidence measured on experts' preferences and using the fuzzy majority concept represented by a linguistic quantifier [29]. In [9,10], different consensus measures based on strict coincidence were presented assuming that experts' preferences are provided by means of linguistic preference relations. Applying the strict coincidence on preferences provided by the experts for each alternative pair, the expert group is divided into subsets, one subset for each possible linguistic label used to quality the preference on the alternative pair. Then, using the cardinalities of the subsets of experts three kinds of consensus measures are defined, each one associated to the three different levels of representation of a preference relation, alternative pair, individual alternative and global relation.

2. *Consensus models based on soft coincidence among preferences.* As above, similarity criteria among preferences are used to compute the coincidence concept but, in this case, a major number of possible coincidence degrees is considered. It is assumed that the coincidence concept is a gradual concept which could be assessed with different degrees defined in the unit interval [0,1]. Some examples of this approach are the following: In [19], a first consensus model based on soft coincidence was also defined. But in this case, given a particular alternative pair and two experts, the coincidence among their preference is measured using a closeness function $s : [0, 1] \rightarrow [0, 1]$. In [20,21], some soft consensus measures defined as extensions of those shown in [19] are introduced, considering GDM problems with heterogeneous set of alternatives and heterogeneous groups of experts, respectively. In [7], an extension of soft consensus models defined in [19,20,21] is presented, which consists in the computation of consensus measures using the ordered weighted averaging (OWA) operator [28]. In [4], a soft consensus model for multi-criteria GDM problems defined in a ordinal fuzzy linguistic approach was defined. In this case, coincidence values are obtained by means of a linguistic similarity function defined directly on linguistic assessments given on the alternatives. In [11], the fuzzification of soft coincidence concept was presented. The soft coincidence is defined in each alternative pair of a linguistic preference relation as a fuzzy set defined on the set of expert pairs and characterized by closeness observed among their preferences. The closeness among preferences is established by means of ad-hoc closeness table defined among all the possible labels of linguistic term set used to represent the preferences. In [14], a soft consensus model is presented to deal with GDM problems in a multi-granular fuzzy linguistic context. As in [9,10,11], three kinds of soft consensus measures are considered. The soft coincidence among multi-granular linguistic preferences is obtained using a similarity function defined on transformation of such preferences in a basic linguistic term set. In [16], as in [9,10,11,14], a soft consensus model based on three consensus measures was proposed. In this case, experts provide their preferences by means of incomplete fuzzy preference relations assessed in [0,1] and the soft coincidence is defined using a similarity function among preferences in [0,1].

3. *Consensus models based on coincidence among on solutions.* In this case, similarity criteria among the solutions obtained from the experts' preferences are used to compute the coincidence concept and different degrees assessed in [0,1] are assumed [1,13]. Basically, we compare the positions of the alternatives between the individual solutions and the collective solution, which allows to know better the real consensus situation in each moment of the consensus process. Some examples of this approach are the following: In [13] was defined the first consensus model based on the measuring the coincidence degree between individual solutions and collective solution. In [13], it is assumed that experts represent their preferences by means of different elements of representation (relation, ordering and utilities) and then it is not possible to compare preferences. To overcome this problem authors propose to compare solutions to obtain the coincidence degrees. This means that the first step of consensus process to measure coincidence degrees is to apply a selection process to obtain a temporary collective solution and the temporary individual solutions, and measure the closeness among them. An important characteristic of this consensus model was the introduction of a recommendation system to aid experts to change their preferences in the consensus reaching process and, in such a way, to substitute the moderator's actions. In [1], a similar consensus model is presented but assuming heterogeneous GDM problems, i.e., experts with different importance degrees.

3 Discussion

In this section, some important aspects of the use of the different approaches to obtain soft consensus degrees within the decision making process are analyzed. To do so, we show the advantages and drawbacks of each one of them.

1. *Strict coincidence among preferences.* The advantage of this approach is that the computation of the consensus degrees is simple and easy because it assumed only two possible values: 1 if the opinions are equal and otherwise a value of 0. However, the drawback of this approach is that the consensus degrees obtained do not reflect the real consensus situation because it only assigns values of 1 or 0 when comparing the experts' opinions, and, for example, we obtain a consensus value 0 for two different preference situations as (very high, high) and (very high,low), when clearly in the second case the consensus value should be lower than in the first case.
2. *Soft coincidence among preferences.* The advantage of this approach is that the consensus degrees obtained are similar to the real consensus situation because they are obtained using similarity functions that assign values between 0 and 1, which are not so strict as in the above approach. The drawback of this approach is that the computation of the consensus degrees is more difficult than in the above approach because we need to define similarity criteria [14,16].
3. *Coincidence among solutions.* The advantage of this approach is that the consensus degrees are obtained comparing not the opinions or choice de-

grees but the position of the alternatives in each solution, what allows us to reflect the real consensus situation in each moment of the consensus reaching process. The drawback of this approach is that the computation of the consensus degrees is more difficult than in the above approaches because we need to define similarity criteria and it is necessary to apply a selection process before obtaining the consensus degrees.

4 New Advanced Approaches

In this section, we describe the new advanced soft consensus approaches which have been developed using the above concepts of coincidence. These approaches allow to generate recommendations to help experts change their opinions in order to obtain the highest degree of consensus possible [13,14,16] and adapt the consensus process to increase the agreement and to reduce the number of experts' preferences that should be changed after each consensus round [25].

4.1 Approaches Generating Recommendations to Help Experts

These approaches generate simple and easy rules to help experts change their opinions in order to obtain the highest degree of consensus possible. To do so, they are based on two consensus criteria, consensus degrees indicating the agreement between experts opinions and proximity measures used to find out how far the individual opinions are from the group opinion. Thus, proximity measures are used in conjunction with the consensus degrees to build a guidance advice system, which acts as a feedback mechanism that generates advice so that experts can change their opinions. Furthermore, these consensus criteria are computed at the three different levels of representation of information of a preference relation: pair of alternatives, alternative, and relation. It allows us to know the current state of consensus from different viewpoints, and therefore, to guide more correctly the consensus reaching processes. Thus, as these measures are given on three different levels for a preference relation, this measure structure will allow us to find out the consensus state of the process at different levels. For example, we will be able to identify which experts are close to the consensus solution, or in which alternatives the experts are having more trouble to reach consensus.

So, the computation of the consensus degrees assuming that experts provide their preferences by means of fuzzy preference relations, $P^h = (p^h_{ij})$, is carried out as follows. First, for each pair of experts (e_h, e_l) $(h = 1, \ldots, m-1,\ l = h+1, \ldots, m)$ a similarity matrix $SM^{hl} = (sm^{hl}_{ik})$ is defined. To do it, one of the above coincidence criteria can be used. Then, a collective similarity matrix, $SM = (sm_{ik})$, is obtained by aggregating all the similarity matrices using an aggregation function ϕ

$$sm_{ik} = \phi(sm^{hl}_{ik},\ h = 1, \ldots, m-1,\ l = h+1, \ldots, m). \tag{1}$$

Once the similarity matrices are computed, the consensus degrees are calculated at the three different levels.

1. **Level 1.** *Consensus degree on pairs of alternatives.* The consensus degree on a pair of alternatives (x_i, x_k), called cop_{ik}, is defined to measure the consensus degree amongst all the experts on that pair of alternatives. In this case, this is expressed by the element of the collective similarity matrix SM, i.e.,

$$cop_{ik} = sm_{ik}. \tag{2}$$

   The closer cop_{ik} to 1, the greater the agreement amongst all the experts on the pair of alternatives (x_i, x_k). This measure will allow the identification of those pairs of alternatives with a poor level of consensus.
2. **Level 2.** *Consensus degree on alternatives.* The consensus degree on alternative x_i, denoted ca_i, is defined to measure the consensus degree among all the experts on that alternative:

$$ca_i = \frac{\sum_{k=1;k\neq i}^{n} (cop_{ik} + cop_{ki})}{2n-2}. \tag{3}$$

   These values can be used to propose the modification of preferences associated to those alternatives with a consensus degree lower than a minimal consensus threshold γ.
3. **Level 3.** *Consensus degree on the relation.* The consensus degree on the relation, called cr is defined to measure the global consensus degree amongst all the experts' opinions. It is computed as the average of all the consensus degrees on the alternatives, i.e.,

$$cr = \frac{\sum_{i=1}^{n} ca_i}{n}. \tag{4}$$

   This is the value used to control the consensus situation.

Once consensus degrees are calculated, the proximity measures are obtained. To compute them for each expert, we need to obtain the collective preference relation, $P^c = (p^c_{ik})$, which summarizes preferences given by all the experts and is calculated by means of the aggregation of the set of individual preference relations $\{P^1, \ldots, P^m\}$ as follows

$$p^c_{ik} = \phi(p^1_{ik}, \ldots, p^m_{ik}). \tag{5}$$

with ϕ an aggregation operator.

Once P^c is obtain, we can compute the proximity measures carrying out the following two steps:

1. For each expert, e_h, a proximity matrix, $PM^h = (pm^h_{ik})$, is obtained using one of the above coincidence criteria.
2. Computation of proximity measures at three different level:
   (a) **Level 1.** *Proximity measure on pairs of alternatives.* The proximity measure of an expert e_h on a pair of alternatives (x_i, x_k) to the group's one, called pp^h_{ik}, is expressed by the element (i, k) of the proximity matrix PM^h:

$$pp^h_{ik} = pm^h_{ik}. \tag{6}$$

(b) **Level 2.** *Proximity measure on alternatives.* The proximity measure of an expert e_h on an alternative x_i to the group's one, called pa_i^h, is calculated as follows:

$$pa_i^h = \frac{\sum_{k=1,k\neq i}^{n} pp_{ik}^h}{n-1}. \tag{7}$$

(c) **Level 3.** *Proximity measure on the relation.* The proximity measure of an expert e_h on his/her unbalanced fuzzy linguistic preference relation to the group's one, called pr^h, is calculated as the average of all proximity measures on the alternatives:

$$pr^h = \frac{\sum_{i=1}^{n} pa_i^h}{n}. \tag{8}$$

The meaning of the proximity measures are the following: if they are close to 1, then they have a positive contribution for the consensus to be high, while if they are close to 0, then they have a negative contribution to the consensus. Therefore, we can use them to provide advice to the experts to change their opinions and to find out which direction that change has to follow in order to obtain the highest degree of consensus possible.

Once proximity measures are calculated, the recommendations are generated. The production of advice to achieve a solution with the highest degree of consensus possible is carried out in two steps [14]: *Identification rules* and *Direction rules*.

1. **Identification rules (IR).** We must identify the experts, alternatives and pairs of alternatives that are contributing less to reach a high degree of consensus and, therefore, should participate in the change process.

   (a) *Identification rule of experts (IR.1).* It identifies the set of experts that should receive advice on how to change some of their preference values. This set of experts, called $EXPCH$, that should change their opinions are those whose satisfaction degree on the relation is lower than the minimum consensus threshold γ. Therefore, the identification rule of experts, IR.1, is the following:

   $$EXPCH = \{e_h \mid pr^h < \gamma\} \tag{9}$$

   (b) *Identification rule of alternatives (IR.2).* It identifies the alternatives whose associated assessments should be taken into account by the above experts in the change process of their preferences. This set of alternatives is denoted as ALT. The identification rule of alternatives, IR.2, is the following:

   $$ALT = \{x_i \in X \mid ca_i < \gamma\} \tag{10}$$

   (c) *Identification rule of pairs of alternatives (IR.3).* It identifies the particular pairs of alternatives (x_i, x_k) whose respective associated assessments

p_{ik}^h the expert e_h should change. This set of pairs of alternatives is denoted as $PALT^h$. The identification rule of pairs of alternatives, IR.3, is the following:

$$PALT^h = \{(x_i, x_k) \mid x_i \in ALT \wedge e_h \in EXPCH \wedge pp_{ik}^h < \gamma\} \quad (11)$$

2. **Direction rules (DR).** We must find out the direction of the change to be recommended in each case, i.e., the direction of change to be applied to the preference assessment p_{ik}^h, with $(x_i, x_k) \in PALT^h$. To do this, we define the following four direction rules.
   (a) *DR.1.* If $p_{ik}^h > p_{ik}^c$, the expert e_h should decrease the assessment associated to the pair of alternatives (x_i, x_k), i.e., p_{ik}^h.
   (b) *DR.2.* If $p_{ik}^h < p_{ik}^c$, the expert e_h should increase the assessment associated to the pair of alternatives (x_i, x_k), i.e., p_{ik}^h.

4.2 Adaptive Approaches

These approaches are based on a refinement process of the consensus process that allows to increase the agreement and to reduce the number of experts' preferences that should be changed after each consensus round. The refinement process adapts the search for the furthest experts' preferences to the existent agreement in each round of consensus. So, when the agreement is very low (initial rounds of the consensus process), the number of changes of preferences should be bigger than when the agreement is medium or high (final rounds) (see Figure 2).

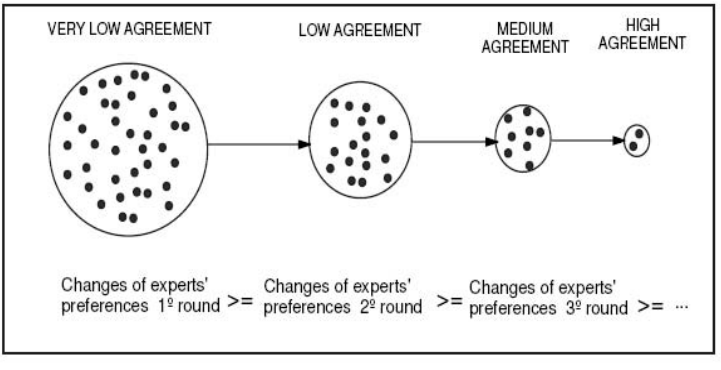


Fig. 2. Reduction of the number of changes of preferences into the consensus process

These approaches consider that in the first rounds of the consensus process, the agreement is usually very low and it seems logic that many experts' preferences should be changed. However, after several rounds, the agreement should have improved and then just the furthest experts' preferences from the collective preference should be changed. It involves that the procedure to search for the furthest experts' preferences from collective preference should be different

according to the achieved agreement in each round. Each Preference Search Procedure (PSp) should have a different behavior and should return a different set of preferences that each expert should change in order to improve the agreement in the next consensus round. In consequence of the adaptation of the consensus process to the existent agreement in each round, the number of changes of preferences suggested to experts after each consensus round will be smaller according to the favorable evolution of the level of agreement.

In this way, in the consensus process, if the agreement among experts is low, i.e, there are a lot of experts' preferences with different assessments, the number of experts which should change their preferences in order to make them closer to collective preference should be great. However, if the agreement is medium or high, it means that the majority of preferences are similar and therefore there exist a low number of experts' preferences far from the collective preference. In this case, only these experts should change them in order to improve the agreement. Keeping in mind this idea, these approaches propose distinguishing among three level of agreement: very low, low and medium consensus. Each level of consensus involves to carry out the search for the furthest preferences in a different way. So when the consensus degree cr is very low, these approaches will search for the furthest preferences on all experts, while if cr is medium, the search will be limited to the furthest experts. To do so, these approaches carries out three different PSps: PSp for very low consensus, PSp for low consensus and PSp for medium consensus. The possibility of carrying out different PSps according to the existent consensus degree in each round defines the adaptive character of our model.

5 Concluding Remarks

In this paper we have identified the different existing approaches to compute soft consensus measures in fuzzy group decision making problems and analyzed their advantages and drawbacks. Additionally, we have described the new advanced approaches allowing to generate recommendations to help experts change their opinions in order to obtain the highest degree of consensus possible and adapt the consensus process to increase the agreement and to reduce the number of experts' preferences that should be changed after each consensus round.

In the future, we think to study as to apply these consensus models in decision making problems with incomplete information and using information domains which do not allow to define similarity criteria among preferences in a direct way, as for example the unbalanced fuzzy linguistic information [12,17].

Acknowledgements

This work has been supported by the Research Projects TIN2007-61079 and SAINFOWEB-PAI00602.

References

1. Ben-Arieh, D., Chen, Z.: Linguistic-Labels Aggregation and Consensus Measure for Autocratic Decision Making using Group Recommendations. IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans 36(3), 558–568 (2006)
2. Bezdek, J., Spillman, B., Spillman, R.: Fuzzy Measures of Preferences and Consensus in Group Decision Making. In: Proc. 1977 IEEE Conf. on Decision and Control, pp. 1303–1309 (1977)
3. Bezdek, J., Spillman, B., Spillman, R.: A Fuzzy Relation Space for Group Decision Theory. Fuzzy Sets and Systems 1, 255–268 (1978)
4. Bordogna, G., Fedrizzi, M., Pasi, G.: A Linguistic Modeling of Consensus for a Fuzzy Majority in Group Decision Making. IEEE Transactions on Systems, Man and Cybernetics 27(1), 126–132 (1997)
5. Chen, S.J., Hwang, C.L.: Fuzzy Multiple Attributive Decision Making: Theory and its Applications. Springer, Berlin (1992)
6. Ephrati, E., Rosenschein, J.S.: Deriving Consensus in Multiagent Systems. Artificial Intelligence 87, 21–74 (1996)
7. Fedrizzi, M., Kacprzyk, J., Nurmi, H.: Consensus Degrees under Fuzzy Majorities and Fuzzy Preferences using OWA (Ordered Weighted Average) Operators. Control Cybernet 22, 78–86 (1993)
8. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer, Dordrecht (1994)
9. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A Model of Consensus in Group Decision Making under Linguistic Assessments. Fuzzy Sets and Systems 78, 73–87 (1996)
10. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A Rational Consensus Model in Group Decision Making using Linguistic Assessments. Fuzzy Sets and Systems 88, 31–49 (1997)
11. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: Linguistic Measures Based on Fuzzy Coincidence for Reaching Consensus in Group Decision Making. International Journal of Approximate Reasoning 16, 309–334 (1997)
12. Herrera, F., Herrera-Viedma, E., Martínez, L.: A Fuzzy Linguistic Methodology to Deal With Unbalanced Linguistic Term Sets. IEEE Transactions on Fuzzy Systems 16(2), 354–370 (2008)
13. Herrera-Viedma, E., Herrera, F., Chiclana, F.: A Consensus Model for Multiperson Decision Making with Different Preference Structures. IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans 32(3), 394–402 (2002)
14. Herrera-Viedma, E., Martínez, L., Mata, F., Chiclana, F.: A Consensus Support System Model for Group Decision-Making Problems with Multi-granular Linguistic Preference Relations. IEEE Transaction on Fuzzy Systems 13(5), 644–658 (2005)
15. Herrera-Viedma, E., Chiclana, F., Herrera, F., Alonso, S.: A Group Decision-Making Model with Incomplete Fuzzy Preference Relations Based on Additive Consistency. IEEE Transactions on Systems, Man and Cybernetics. Part B, Cybernetics 37(1), 176–189 (2007)
16. Herrera-Viedma, E., Alonso, S., Chiclana, F., Herrera, F.: A Consensus Model for Group Decision Making with Incomplete Fuzzy Preference Relations. IEEE Transactions on Fuzzy Systems 15(5), 863–877 (2007)
17. Herrera-Viedma, E., López-Herrera, A.G.: A Model of Information Retrieval System with Unbalanced Fuzzy Linguistic Information. International Journal of Intelligent Systems 22(11), 1197–1214 (2007)

18. Hwang, C.L., Lin, M.J.: Group Decision Making under Multiple Criteria: Methods and Applications. Springer, Berlin (1987)
19. Kacprzyk, J.: On Some Fuzzy Cores and Soft Consensus Measures in Group Decision Making. In: Bezdek, J. (ed.) The Analysis of Fuzzy Information, pp. 119–130. CRC Press, Boca Raton (1987)
20. Kacprzyk, J., Fedrizzi, M.: Soft Consensus Measure for Monitoring Real Consensus Reaching Processes under Fuzzy Preferences. Control Cybernet 15, 309–323 (1986)
21. Kacprzyk, J., Fedrizzi, M.: A Soft Measure of Consensus in the Setting of Partial (Fuzzy) Preferences. European Journal of Operational Research 34, 316–323 (1988)
22. Kacprzyk, J., Fedrizzi, M., Nurmi, H.: Group Decision Making and Consensus under Fuzzy Preferences and Fuzzy Majority. Fuzzy Sets and Systems 49, 21–31 (1992)
23. Kacprzyk, J., Nurmi, H., Fedrizzi, M.: Consensus under Fuzziness. Kluwer Academic Publishers, Boston (1997)
24. Marakas, G.H.: Decision Support Systems in the 21th Century, 2nd edn. Pearson Education, Inc., New Jersey (2003)
25. Mata, F.: Modelos para Sistemas de Apoyo al Consenso en Problemas de Toma de Decisión en Grupo definidos en Contextos Lingüísticos Multigranulares. Doctoral Thesis (2006)
26. Spillman, B., Bezdek, J., Spillman, R.: Coalition Analysis with Fuzzy Sets. Kybernetes 8, 203–211 (1979)
27. Spillman, B., Spillman, R., Bezdek, J.: A Fuzzy Analysis of Consensus in Small Groups. In: Wang, P.P., Chang, S.K. (eds.) Fuzzy Automata and Decision Processes, pp. 331–356. North-Holland, Amsterdam (1980)
28. Yager, R.R.: On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making. IEEE Transactions on Systems, Man, Cybernetics. Part A: Systems and Humans 18, 183–190 (1988)
29. Zadeh, L.A.: A Computational Approach to Fuzzy Quantifiers in Natural Languages. Computers and Mathematics with Applications 9, 149–184 (1983)

SBM and Bipolar Models in Data Envelopment Analysis with Interval Data

Masahiro Inuiguchi and Fumiki Mizoshita

Graduate School of Engineering Science, Osaka University,
1-3, Machikaneyama, Toyonaka, Osaka 560-8531, Japan
inuiguti@sys.es.osaka-u.ac.jp

Abstract. Under interval input-output data, 25 qualitative different efficiencies have been proposed. In this paper, SBM models for DEA with interval input-output data are investigated in order to introduce quantitative evaluation. It is shown that SBM models for 14 efficiencies are reduced to linear programming problems. Moreover, in order to evaluate decision making units in a negative way, we generalize the inverted DEA into the case of interval input-output data. Evaluation based on efficiency-inefficiency scores is investigated.

1 Introduction

The efficiency of a decision making unit (DMU) can be evaluated simply by the ratio of its output amount to its input amount. When there are many different inputs and many different outputs, it is difficult to define the total input and output amounts. In order to evaluate the efficiency of DMUs with multiple inputs and outputs, data envelopment analysis (DEA) [1] was proposed. In DEA, the efficiency of a DMU is evaluated in comparison with many DMUs having same kinds of inputs and outputs. Because of its usefulness and tractability, a lot of applications as well as methodological developments of DEA were done.

Because data are sometimes observed with a noise and/or with the inaccuracy, DEA with uncertain data is required. To this end, sensitivity analysis [2,3] was developed. This analysis usually works well in data fluctuations of only one DMU. Chance constrained models [4,13] of DEA were proposed in which input-output data are treated as random variable vectors. In this approach, we need to assume special types of probability distributions and the reduced problems for evaluating efficiency generally becomes nonlinear programming problems. The interval approach [6,10] and fuzzy set approach [7,9,11] were also proposed. In those approaches, imprecise data are represented by intervals or fuzzy numbers and the range or fuzzy set of efficiency scores are calculated. Inuiguchi and Tanino [9] proposed possible and necessity efficiencies and showed the relation with fuzzy efficiency scores. Moreover, imprecise DEA [5] was also developed in order to treat imprecise knowledge about input-output data in DEA. The model allows interval data and ordinal data, where ordinal data specifies only the order of data values but not real data values. It is shown in [5] that the efficiency evaluation problem with imprecise data is reduced to a linear programming.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 98–109, 2008.

Recently, the authors [8,12] proposed an approach to DEA with interval data. Dominance relation between DMUs are variously defined based on the combinations of four kinds of inequality relations between intervals. They proposed 25 efficiencies and showed that any efficiencies defined by logical combinations of four inequalities between intervals can be obtained by logical combinations of the 25 efficiencies. The strong-weak relations among those 25 efficiencies are shown. This implies that owing to the imprecision of data, the efficiency of DMU can be evaluated qualitatively. Moreover, they showed that the 25 efficiency test problems are solved by multi-phase simplex methods. The 25 kinds of efficiency scores are also defined but they are not consistent with their strong-weak relations.

In this paper, we extend the authors' previous approach [8,12]. First we propose efficiency scores consistent with the strong-weak relations among 25 efficiencies. In order to equip good properties, we introduce the SBM model [15]. To each of 25 efficiencies, an SBM model is formulated. We show that SBM models for 14 efficiencies can be reduced to linear programming problems.

DEA provides an optimistic evaluation so that some DMU may be positively overrated. To moderate such a positive overassessment, we introduce a negative assessment. The inverted DEA model [16] has been proposed for such a purpose. Then we extend the inverted DEA model into the case of interval input-output data. We investigate the overall evaluation combining DEA results and the inverted DEA results under interval data.

This paper is organized as follows. In next section, we briefly review the authors' approach to DEA with interval input-output data. In Section 3, we develop SBM model under interval input-output data. Moreover we investigate the inverted DEA with interval input-output data.

2 Data Envelopment Analysis with Interval Data

Data envelopment analysis (DEA) [1] is a tool to evaluate Decision Making Units (DMUs) based on the comparison among input-output data. If there is no possible activity outperforming the o-th DMU under given input-output data, the o-th DMU is regarded as efficient. In this paper, we evaluate DMUs when input-output data are given as intervals. It was shown that 25 kinds of efficiencies are obtained and that DMUs can be qualified by the 25 efficiencies [8].

We assume the i-th input data of the j-th DMU is given by interval $\mathcal{X}_{ij} = [x_{ij}^{\mathrm{L}}, x_{ij}^{\mathrm{R}}]$ and the k-th output data of the j-th DMU by interval $\mathcal{Y}_{kj} = [y_{kj}^{\mathrm{L}}, y_{kj}^{\mathrm{R}}]$. For the sake of the simplicity, we use an interval input data matrix $\mathcal{X}$ having $\mathcal{X}_{ij}$ as its (i,j)-component and an interval output matrix $\mathcal{Y}$ having $\mathcal{Y}_{kj}$ as its (k,j)-component. The interval input-output data of j-th DMU is given by $(\mathcal{X}_{\cdot j}, \mathcal{Y}_{\cdot j})$, where $\mathcal{X}_{\cdot j}$ and $\mathcal{Y}_{\cdot j}$ are j-th column of interval matrices $\mathcal{X}$ and $\mathcal{Y}$. Moreover, we use matrices $X^{\mathrm{L}} = (x_{ij}^{\mathrm{L}})$, $X^{\mathrm{R}} = (x_{ij}^{\mathrm{R}})$, $Y^{\mathrm{L}} = (y_{kj}^{\mathrm{L}})$ and $Y^{\mathrm{R}} = (y_{kj}^{\mathrm{L}})$ showing lower and upper bounds of interval matrices $\mathcal{X}$ and $\mathcal{Y}$. The j-th columns of X^{L}, X^{R}, Y^{L} and Y^{R} are denoted by $X_{\cdot j}^{\mathrm{L}}$, $X_{\cdot j}^{\mathrm{R}}$, $Y_{\cdot j}^{\mathrm{L}}$ and $Y_{\cdot j}^{\mathrm{R}}$, respectively.

In order to define efficiencies, we need to introduce dominance relations and a set of possible activities which is called a production possibility set. The following

dominance relations $\succeq^{Q}$ ($Q \in \{\Pi, \mathrm{N}, \mathrm{L}, \mathrm{R}, \mathrm{LR}, \mathrm{L|R}\}$) between two interval input-output data (Γ_1, Δ_1) and (Γ_2, Δ_2) are defined:

$$(\Gamma_1, \Delta_1) \succeq^{\Pi} (\Gamma_2, \Delta_2) \Leftrightarrow \boldsymbol{x}_1^{\mathrm{L}} \leq \boldsymbol{x}_2^{\mathrm{R}} \text{ and } \boldsymbol{y}_1^{\mathrm{L}} \geq \boldsymbol{y}_2^{\mathrm{R}}, \tag{1}$$
$$(\Gamma_1, \Delta_1) \succeq^{\mathrm{N}} (\Gamma_2, \Delta_2) \Leftrightarrow \boldsymbol{x}_1^{\mathrm{R}} \leq \boldsymbol{x}_2^{\mathrm{L}} \text{ and } \boldsymbol{y}_1^{\mathrm{R}} \geq \boldsymbol{y}_2^{\mathrm{L}}, \tag{2}$$
$$(\Gamma_1, \Delta_1) \succeq^{\mathrm{L}} (\Gamma_2, \Delta_2) \Leftrightarrow \boldsymbol{x}_1^{\mathrm{R}} \leq \boldsymbol{x}_2^{\mathrm{R}} \text{ and } \boldsymbol{y}_1^{\mathrm{L}} \geq \boldsymbol{y}_2^{\mathrm{L}}, \tag{3}$$
$$(\Gamma_1, \Delta_1) \succeq^{\mathrm{R}} (\Gamma_2, \Delta_2) \Leftrightarrow \boldsymbol{x}_1^{\mathrm{L}} \leq \boldsymbol{x}_2^{\mathrm{L}} \text{ and } \boldsymbol{y}_1^{\mathrm{R}} \geq \boldsymbol{y}_2^{\mathrm{R}}, \tag{4}$$
$$(\Gamma_1, \Delta_1) \succeq^{\mathrm{LR}} (\Gamma_2, \Delta_2) \Leftrightarrow (\Gamma_1, \Delta_1) \succeq^{\mathrm{L}} (\Gamma_2, \Delta_2) \text{ and } (\Gamma_1, \Delta_1) \succeq^{\mathrm{R}} (\Gamma_2, \Delta_2), \tag{5}$$
$$(\Gamma_1, \Delta_1) \succeq^{\mathrm{L|R}} (\Gamma_2, \Delta_2) \Leftrightarrow (\Gamma_1, \Delta_1) \succeq^{\mathrm{L}} (\Gamma_2, \Delta_2) \text{ or } (\Gamma_1, \Delta_1) \succeq^{\mathrm{R}} (\Gamma_2, \Delta_2). \tag{6}$$

Using dominance relations $\succeq^{Q}$, $Q \in \{\Pi, \mathrm{N}, \mathrm{L}, \mathrm{R}, \mathrm{LR}, \mathrm{L|R}\}$, we define 25 strong dominance relations by

$$(\Gamma_1, \Delta_1) \succ^{Q_1\text{-}Q_2} (\Gamma_2, \Delta_2) \Leftrightarrow (\Gamma_1, \Delta_1) \succeq^{Q_1} (\Gamma_2, \Delta_2) \text{ and } (\Gamma_2, \Delta_2) \not\succeq^{Q_2} (\Gamma_1, \Delta_1), \quad Q_1 \in \mathcal{Q}_1,\ Q_2 \in \mathcal{Q}_2, \tag{7}$$

where $\mathcal{Q}_1 = \{\Pi, \mathrm{N}, \mathrm{L}, \mathrm{R}, \mathrm{LR}\}$ and $\mathcal{Q}_2 = \{\Pi, \mathrm{N}, \mathrm{L}, \mathrm{R}, \mathrm{L|R}\}$. In (7), we do not consider the cases when $Q_1 = \mathrm{L|R}$ or $Q_2 = \mathrm{LR}$ because they are evaluated by the efficiencies defined by using the 25 strong dominance relations [8].

Let $\boldsymbol{e} = (1, 1, \ldots, 1)^{\mathrm{T}}$. The possible activities are uniquely defined by

$$\mathcal{P} = \{(\boldsymbol{\gamma}, \boldsymbol{\delta}) \mid (\mathcal{X}\boldsymbol{\lambda}, \mathcal{Y}\boldsymbol{\lambda}) \succeq^{\mathrm{L}} (\boldsymbol{\gamma}, \boldsymbol{\delta}), (\mathcal{X}\boldsymbol{\lambda}, \mathcal{Y}\boldsymbol{\lambda}) \succeq^{\mathrm{R}} (\boldsymbol{\gamma}, \boldsymbol{\delta}),\ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \boldsymbol{0}\}. \tag{8}$$

Then we can define 25 kinds of $(Q_1\text{-}Q_2)$-efficiencies as follows:

$$\text{the } j\text{-th DMU is } (Q_1\text{-}Q_2)\text{-efficient} \Leftrightarrow \nexists (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{P} : (\boldsymbol{\gamma}, \boldsymbol{\delta}) \succeq^{Q_1\text{-}Q_2} (\mathcal{X}_{\cdot j}, \mathcal{Y}_{\cdot j}), \tag{9}$$

where $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$.

These 25 $(Q_1\text{-}Q_2)$-efficiencies are qualitatively different. It should be noted that, owing to the uncertainty of input-output data, we can evaluate efficiencies of DMUs qualitatively. Considering the combinations of those $(Q_1\text{-}Q_2)$-efficiencies by logical connectives, we have much more efficiencies.

The strong-weak relation among 25 $(Q_1\text{-}Q_2)$-efficiencies is shown as in Figure 1. As shown in Figure 1, Π-N efficiency is the strongest. A Π-N efficient DMU stays efficient even if input-out data fluctuate in the given intervals. On the contrary, N-Π efficiency is the weakest. An N-Π efficient DMU is efficient only for a combination of input-output values in the given intervals. The others are between them but there are many kinds. Considering the definition of $(Q_1\text{-}Q_2)$-efficiency, the influence of dominance relation $\succeq^{Q_1}$ would be stronger than that of dominance relation $\succeq^{Q_2}$. In this sense, we may use only five $(Q_1\text{-}Q_2)$-efficiencies whose Q_1's are different one another in order to reduce the complexity of the analysis.

Inuiguchi and Mizoshita [8] showed that each of 25 $(Q_1\text{-}Q_2)$-efficiencies of a DMU is tested by solving a mathematical programming problem. The test problem of a $(Q_1\text{-}Q_2)$-efficiency of the o-th DMU can be formulated as the following

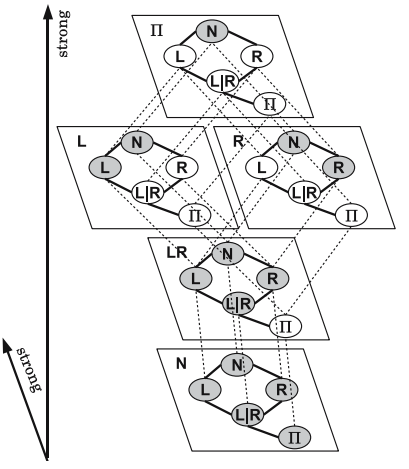


Fig. 1. Strong-weak relation among 25 efficiencies

Table 1. The correspondence between Q_i and $(Q_{i1}, Q_{i2}, Q_{i3}, Q_{i4})$, $i = 1, 2$

Q_1	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_2	Q_{21}	Q_{22}	Q_{23}	Q_{24}
Π	L	R	R	L	Π	R	L	L	R
N	R	L	L	R	N	L	R	R	L
L	L	L	R	R	L	L	L	R	R
R	R	R	L	L	R	R	R	L	L

mixed integer programming problem:

$$
\begin{aligned}
&\text{maximize } s,\\
&\text{subject to}\\
&\begin{cases}
\text{if } Q_1 \neq \mathrm{LR}\\
\quad X^{Q_{11}}\boldsymbol{\lambda} \le X_{\cdot o}^{Q_{12}},\ \ Y^{Q_{13}}\boldsymbol{\lambda} \ge Y_{\cdot o}^{Q_{14}},\\
\text{if } Q_1 = \mathrm{LR}\\
\quad X^{\mathrm{L}}\boldsymbol{\lambda} \le X_{\cdot o}^{\mathrm{L}},\ \ Y^{\mathrm{L}}\boldsymbol{\lambda} \ge Y_{\cdot o}^{\mathrm{L}},\ \ X^{\mathrm{R}}\boldsymbol{\lambda} \le X_{\cdot o}^{\mathrm{R}},\ \ Y^{\mathrm{R}}\boldsymbol{\lambda} \ge Y_{\cdot o}^{\mathrm{R}},
\end{cases}\\
&\begin{cases}
\text{if } Q_2 \neq \mathrm{L|R}\\
\displaystyle\sum_{i=1}^{m} X_{i\cdot}^{Q_{21}}\boldsymbol{\lambda} z_i^{1-} - \sum_{k=1}^{p} Y_{k\cdot}^{Q_{23}}\boldsymbol{\lambda} z_k^{1+} + s = \sum_{i=1}^{m} x_{io}^{Q_{22}} z_i^{1-} - \sum_{k=1}^{p} y_{ko}^{Q_{24}} z_k^{1+},\\
\text{if } Q_2 = \mathrm{L|R}\\
\quad \begin{cases}
\displaystyle\sum_{i=1}^{m} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_i^{1-} - \sum_{k=1}^{p} Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_k^{1+} + s \le \sum_{i=1}^{m} x_{io}^{\mathrm{L}} z_i^{1-} - \sum_{k=1}^{p} y_{ko}^{\mathrm{R}} z_k^{1+},\\
\displaystyle\sum_{i=1}^{m} X_{i\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_i^{2-} - \sum_{k=1}^{p} Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_k^{2+} + s \le \sum_{i=1}^{m} x_{io}^{\mathrm{R}} z_i^{2-} - \sum_{k=1}^{p} y_{ko}^{\mathrm{L}} z_k^{2+},
\end{cases}
\end{cases}\\
&\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda} = 1,\ \ \boldsymbol{\lambda} \ge \mathbf{0},\ \ \lambda_o = 0,\\
&z_i^{j-} \in \{0,1\},\ i = 1, 2, \ldots, m, \quad z_k^{j+} \in \{0,1\},\ k = 1, 2, \ldots, p,\ \ j = 1, 2,
\end{aligned}
\tag{10}
$$

where the correspondence between Q_i and Q_{ij}, $j = 1, 2, 3, 4$ is shown in Table 1. If Problem (10) has a feasible solution such that $s > 0$, the o-th DMU is not $(Q_1 - Q_2)$-efficient. Otherwise, the o-th DMU is $(Q_1 - Q_2)$-efficient.

In some cases, Problem (10) is reduced to a linear programming problem by the following theorem (Mizoshita and Inuiguchi [12]).

Theorem 1. *Assume one of the following assertions holds.*

1. $Q_1 \neq \mathrm{LR}$, $Q_2 \neq \mathrm{L|R}$, $X^{Q_{21}} \leq X^{Q_{11}}$, $X^{Q_{12}} \leq X^{Q_{22}}$, $Y^{Q_{13}} \leq Y^{Q_{23}}$ *and* $Y^{Q_{24}} \leq Y^{Q_{14}}$,
2. $Q_1 \neq \mathrm{LR}$, $Q_2 = \mathrm{L|R}$, $X^{Q_{12}} \leq X^{\mathrm{L}} \leq X^{Q_{11}}$ *and* $Y^{Q_{13}} \leq Y^{\mathrm{R}} \leq Y^{Q_{14}}$,
3. $Q_1 \neq \mathrm{LR}$, $Q_2 = \mathrm{L|R}$, $(X^{Q_{21}} \leq X^{\mathrm{L}} \leq X^{Q_{22}}$ *or* $X^{Q_{21}} \leq X^{\mathrm{R}} \leq X^{Q_{22}})$ *and* $(Y^{Q_{24}} \leq Y^{\mathrm{L}} \leq Y^{Q_{23}}$ *or* $Y^{Q_{24}} \leq Y^{\mathrm{R}} \leq Y^{Q_{23}})$,
4. $Q_1 = \mathrm{LR}$ *and* $Q_2 = \mathrm{L|R}$.

Then we have $\boldsymbol{z}^{1-} = \boldsymbol{e}$ *and* $\boldsymbol{z}^{1+} = \boldsymbol{e}$ *at an optimal solution of Problem (10).*

Similarly, assume one of the following assertions holds.

1. $Q_1 \neq \mathrm{LR}$, $Q_2 = \mathrm{L|R}$, $X^{Q_{12}} \leq X^{\mathrm{R}} \leq X^{Q_{11}}$ *and* $Y^{Q_{13}} \leq Y^{\mathrm{L}} \leq Y^{Q_{14}}$,
2. $Q_1 = \mathrm{LR}$ *and* $Q_2 = \mathrm{L|R}$.

Then, we have $\boldsymbol{z}^{2-} = \boldsymbol{e}$ *and* $\boldsymbol{z}^{2+} = \boldsymbol{e}$ *at an optimal solution of Problem (10).*

From Theorem 1, if Q_1-Q_2 is Π-N, N-Π, N-N, N-L, N-R, N-L|R, L-N, L-L, R-N, R-R, LR-N, LR-L, LR-R or LR-L|R, we can solve Problem (10) with fixing $\boldsymbol{z}^{1-} = \boldsymbol{e}$, $\boldsymbol{z}^{1+} = \boldsymbol{e}$, $\boldsymbol{z}^{2-} = \boldsymbol{e}$ and $\boldsymbol{z}^{2+} = \boldsymbol{e}$. In these cases, Problem (10) is reduced to a linear programming problem. In Figure 1, those efficiencies are shown by the shade. Mizoshita and Inuiguchi [12] showed that even the other 11 efficiency test problems are solved by a multi-phase phase simplex method.

In order to evaluate the proximity to $(Q_1 - Q_2)$-efficiency, Mizoshita and Inuiguchi [12] have defined a $(Q_1 - Q_2)$-efficiency score. Unfortunately, the score is counter-intuitive to the strong-weak relation among $(Q_1 - Q_2)$-efficiencies. The score of a DMU to a stronger $(Q_1 - Q_2)$-efficiency can take a larger value than that to a weaker $(Q_1 - Q_2)$-efficiency. In next section, we would like to remedy this inadequacy by introducing the slack-based measure (SBM) model.

3 SBM Model for Interval Input-Output Data

3.1 SBM Model for Crisp Input-Output Data

In CCR and BCC models of DEA, the efficiency scores of DMUs can be calculated by solving efficiency test problems. The scores depend on the choice between input-orientation and output-orientation. There are no CCR/BCC models treating input and output equally. On the contrary, the additive model of DEA treats input and output equally but no efficiency scores can be obtained. Tone [14] proposed efficiency scores by using slack variables in the additive model. Tone [15] considered desirable properties for efficiency scores and proposed the slack-based measure (SBM) model.

Let $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ be the input-output data of the i-th DMU, where $\boldsymbol{x}_i$ is the input data represented as an m-dimensional vector and $\boldsymbol{y}$ is the output data represented as a p-dimensional vector. Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be an $m \times n$ matrix of input data and $Y = (\boldsymbol{y}_i, \ldots, \boldsymbol{y}_n)$ be a $p \times n$ matrix of output data. The SBM model [15] evaluating o-th DMU's efficiency is the following linear fractional programming problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1 - \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} s_i^- / x_{io}}{1 + \dfrac{1}{p}\displaystyle\sum_{k=1}^{p} s_k^+ / y_{ro}}, \\ \text{subject to} \quad & X\boldsymbol{\lambda} + \boldsymbol{s}^- = \boldsymbol{x}_o, \\ & Y\boldsymbol{\lambda} - \boldsymbol{s}^+ = \boldsymbol{y}_o, \\ & \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda} = 1, \\ & \boldsymbol{\lambda} \geq \boldsymbol{0},\ \boldsymbol{s}^- \geq \boldsymbol{0},\ \boldsymbol{s}^+ \geq \boldsymbol{0}. \end{aligned} \tag{11}$$

The efficiency score is obtained as the optimal value to this problem. The o-th DMU is efficient if and only if the optimal value is zero. It is shown that Problem(11) is reduced to a linear programming problem.

We define a reference-set R_o with respect to the o-th DMU by $R_o = \{j \in \{1, 2, \ldots, n\} \mid \lambda_j^* > 0\}$, where λ_j^* is the j-th component of $\boldsymbol{\lambda}^*$. Then the efficiency score of SBM model has the following properties:

(P1) *Unit invariant*: The score should be invariant with respect to the units of data.
(P2) *Monotone*: The score should be monotone decreasing in each slack in input and output.
(P3) *Translation invariant*: The score should be invariant under parallel translation of the coordinate system applied.
(P4) *Reference-set dependent*: The score should be determined only by consulting the reference-set of the DMU concerned.

3.2 Extension to the Case of Interval Input-Output Data

Let us introduce SBM model into DEA with interval input-output data. We find that, in interval case, there are two comparisons between the activities of the o-th DMU and the combined activities of DMUs which are appeared in constraints with respect to Q_1 part and Q_2 part, while in the conventional case, there is only one comparison. This is caused by a fact that the dominance relation between intervals uses their lower and upper bounds. Taking care of this difference, we define the following efficiency score of the o-th DMU corresponding to the objective function of Problem (11):

$$\rho_o = \frac{1 - \dfrac{1}{2m}\displaystyle\sum_{i=1}^{m}\left(\frac{d_i^-}{x_{io}^{\mathrm{R}}} + \frac{s_i^-}{x_{io}^{\mathrm{R}}}\right)}{1 + \dfrac{1}{2p}\displaystyle\sum_{k=1}^{p}\left(\frac{d_k^+}{y_{ko}^{\mathrm{R}}} + \frac{s_k^+}{y_{ko}^{\mathrm{R}}}\right)}, \tag{12}$$

where $\boldsymbol{d}^+ = (d_1^+, \ldots, d_m^+)^{\mathrm{T}}$ and $\boldsymbol{d}^- = (d_1^-, \ldots, d_p^-)^{\mathrm{T}}$ are slack variable vectors corresponding to Q_1 part while $\boldsymbol{s}^+ = (s_1^+, \ldots, s_m^+)^{\mathrm{T}}$ and $\boldsymbol{s}^- = (s_1^-, \ldots, s_p^-)^{\mathrm{T}}$ are slack variable vectors corresponding to Q_2 part (see constraints of Problem (13) described later for more precise definitions of slack variables). Then it is conceivable to employ ρ_o in (12) as the objective function to be maximized.

In order to evaluate the efficiency, we should check the positivity of the sum $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_k^-$. If it can be positive, the o-th DMU is not $(Q_1 - Q_2)$-efficient. The maximization of ρ_o in (12) may make the sum $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_k^-$ zero even if there exists a feasible solution with $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_k^- > 0$. Then, we add a constraint $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_k^- \geq \epsilon$ where ϵ is a very small positive number.

As the result, the SBM model extended to interval input-output data is formulated as the following mixed integer programming problem:

$$
\begin{aligned}
&\text{minimize} \quad \rho_o = \frac{1 - \dfrac{1}{2m}\displaystyle\sum_{i=1}^{m}\left(\frac{d_i^-}{x_{iq}^{\mathrm{R}}} + \frac{s_i^-}{x_{iq}^{\mathrm{R}}}\right)}{1 + \dfrac{1}{2p}\displaystyle\sum_{k=1}^{p}\left(\frac{d_k^+}{y_{kq}^{\mathrm{R}}} + \frac{s_k^+}{y_{kq}^{\mathrm{R}}}\right)},\\
&\text{subject to}\\
&\quad \text{if } Q_1 \neq \mathrm{LR},\\
&\quad \begin{cases} X_{i\cdot}^{Q_{11}}\boldsymbol{\lambda} + d_i^- = x_{iq}^{Q_{12}}, \ i = 1, \ldots, m,\\ Y_{k\cdot}^{Q_{13}}\boldsymbol{\lambda} - d_k^+ = y_{kq}^{Q_{14}}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_1 = \mathrm{LR},\\
&\quad \begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\lambda} + d_i^- \leq x_{iq}^{\mathrm{L}}, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\lambda} + d_i^- \leq x_{iq}^{\mathrm{R}}, \ i = 1, \ldots, m,\\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\lambda} - d_k^+ \geq y_{kq}^{\mathrm{L}}, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\lambda} - d_k^+ \geq y_{kq}^{\mathrm{R}}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_2 \neq \mathrm{L|R},\\
&\quad \begin{cases} X_{i\cdot}^{Q_{21}}\boldsymbol{\lambda} z_i^{1-} + s_i^- = x_{iq}^{Q_{22}} z_i^{1-}, i = 1, \ldots, m,\\ Y_{k\cdot}^{Q_{23}}\boldsymbol{\lambda} z_k^{1+} - s_k^+ = y_{kq}^{Q_{24}} z_k^{1+}, k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_2 = \mathrm{L|R},\\
&\quad \begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_i^{1-} + s_i^- \leq x_{iq}^{\mathrm{L}} z_i^{1-}, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_i^{2-} + s_i^- \leq x_{iq}^{\mathrm{R}} z_i^{2-}, \ i = 1, \ldots, m,\\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_k^{2+} - s_k^+ \geq y_{kq}^{\mathrm{L}} z_k^{2+}, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_k^{1+} - s_k^+ \geq y_{kq}^{\mathrm{R}} z_k^{1+}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \sum_{i=1}^{m} s_i^+ + \sum_{k=1}^{p} s_k^- \geq \epsilon, \ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda} = 1, \ \boldsymbol{\lambda} \geq \boldsymbol{0}, \ \lambda_o = 0, \ \boldsymbol{d}^-, \ \boldsymbol{d}^+, \ \boldsymbol{s}^-, \ \boldsymbol{s}^+ \geq \boldsymbol{0},\\
&\quad z_i^{1-}, z_i^{2-}, z_k^{1+}, z_k^{2+} \in \{0, 1\}, \ i = 1, 2, \ldots, m, \ k = 1, 2, \ldots, p.
\end{aligned}
\tag{13}
$$

If there is no feasible solution, the o-th DMU is $(Q_1$-$Q_2)$-efficient. If $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_j^- = \epsilon$ holds at the obtained optimal solution, the o-th DMU is regarded as $(Q_1$-$Q_2)$-efficient (to confirm the $(Q_1$-$Q_2)$-efficiency, we may solve Problem (10)). Otherwise, the o-th DMU is not $(Q_1$-$Q_2)$-efficient and the optimal value is the score called the $(Q_1$-$Q_2)$-efficiency score. Note that s of Problem (10) is corresponding to $\sum_{i=1}^m s_i^+ + \sum_{k=1}^p s_k^-$ of Problem (13).

Problem (13) is reduced to the following problem with a linear objective function:

$$
\begin{aligned}
&\text{minimize} \quad \tau_o = t - \frac{1}{2m}\sum_{i=1}^{m}\left(\frac{D_i^-}{x_{iq}^{\mathrm{R}}} + \frac{S_i^-}{x_{iq}^{\mathrm{R}}}\right),\\
&\text{subject to}\\
&1 = t + \frac{1}{2p}\sum_{k=1}^{p}\left(\frac{D_k^+}{y_{kq}^{\mathrm{R}}} + \frac{S_k^+}{y_{kq}^{\mathrm{R}}}\right),\\
&\text{if } Q_1 \neq \mathrm{LR},\\
&\begin{cases} X_{i\cdot}^{Q_{11}}\boldsymbol{\Lambda} + D_i^- = x_{iq}^{Q_{12}}t, \ i = 1,\ldots,m,\\ Y_{k\cdot}^{Q_{13}}\boldsymbol{\Lambda} - D_k^+ = y_{kq}^{Q_{14}}t, \ k = 1,\ldots,p, \end{cases}\\
&\text{if } Q_1 = \mathrm{LR},\\
&\begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\Lambda} + D_i^- \le x_{iq}^{\mathrm{L}}t, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\Lambda} + D_i^- \le x_{iq}^{\mathrm{R}}t, \ i = 1,\ldots,m,\\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\Lambda} - D_k^+ \ge y_{kq}^{\mathrm{L}}t, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\Lambda} - D_k^+ \ge y_{kq}^{\mathrm{R}}t, \ k = 1,\ldots,p, \end{cases}\\
&\text{if } Q_2 \neq \mathrm{L|R},\\
&\begin{cases} X_{i\cdot}^{Q_{21}}\boldsymbol{\Lambda}z_i^{1-} + S_i^- = x_{iq}^{Q_{22}}z_i^{1-}t, i = 1,\ldots,m,\\ Y_{k\cdot}^{Q_{23}}\boldsymbol{\Lambda}z_k^{1+} - S_k^+ = y_{kq}^{Q_{24}}z_k^{1+}t, k = 1,\ldots,p, \end{cases}\\
&\text{if } Q_2 = \mathrm{L|R},\\
&\begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\Lambda}z_i^{1-} + S_i^- \le x_{iq}^{\mathrm{L}}z_i^{1-}t, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\Lambda}z_i^{2-} + S_i^- \le x_{iq}^{\mathrm{R}}z_i^{2-}t, \ i = 1,\ldots,m,\\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\Lambda}z_k^{1+} - S_k^+ \ge y_{kq}^{\mathrm{L}}z_k^{1+}t, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\Lambda}z_k^{2+} - S_k^+ \ge y_{kq}^{\mathrm{R}}z_k^{2+}t, \ k = 1,\ldots,p, \end{cases}\\
&\sum_{i=1}^{m} S_i^+ + \sum_{k=1}^{p} S_j^- \ge t\epsilon, \ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Lambda} = t, \ \boldsymbol{\Lambda} \ge \mathbf{0}, \ \Lambda_o = 0, \ \boldsymbol{D}^-, \ \boldsymbol{D}^+, \ \boldsymbol{S}^-, \ \boldsymbol{S}^+ \ge \mathbf{0},\\
&z_i^{1-}, z_i^{2-}, z_k^{1+}, z_k^{2+} \in \{0,1\}, \ i = 1,2,\ldots,m, \ k = 1,2,\ldots,p, \ t \ge 0.
\end{aligned}
\tag{14}
$$

Let $(\tau_o^*,\ t^*,\ \boldsymbol{\Lambda}^*,\ \boldsymbol{D}^{-*},\ \boldsymbol{D}^{+*},\ \boldsymbol{S}^{-*},\ \boldsymbol{S}^{+*},\ \boldsymbol{z}^{1-*},\ \boldsymbol{z}^{2-*},\ \boldsymbol{z}^{1+*},\ \boldsymbol{z}^{2+*})$ be an optimal solution to Problem (14). Then an optimal solution to Problem (13) is obtained as $(\rho_o^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{d}^{-*},\ \boldsymbol{d}^{+*},\ \boldsymbol{s}^{-*},\ \boldsymbol{s}^{+*},\ \boldsymbol{z}^{1-*},\ \boldsymbol{z}^{2-*},\ \boldsymbol{z}^{1+*},\ \boldsymbol{z}^{2+*})$ with definitions,

$$
\begin{aligned}
&\rho_o^* = \tau_o^*,\ \boldsymbol{\lambda}^* = \boldsymbol{\Lambda}^*/t^*,\ \boldsymbol{d}^{-*} = \boldsymbol{D}^{-*}/t^*,\ \boldsymbol{d}^{+*} = \boldsymbol{D}^{+*}/t^*,\\
&\boldsymbol{s}^{-*} = \boldsymbol{S}^{-*}/t^* \text{ and } \ \boldsymbol{s}^{+*} = \boldsymbol{S}^{+*}/t^*.
\end{aligned}
\tag{15}
$$

Theorem 1 is valid also for Problem (14). Then if Q_1-Q_2 is II-N, N-II, N-N, N-L, N-R, N-L|R, L-N, L-L, R-N, R-R, LR-N, LR-L, LR-R or LR-L|R, Problem (14) is reduced to a linear programming problem. (Q_1-Q_2)-efficiency score satisfy properties (P1), (P2), (P3) and (P4) as well as the following property:

(P5) *Consistent with strong-weak relation*: The score should be consistent with the strong-weak relation among (Q_1-Q_2) efficiencies.

4 Bipolar Evaluation Based on Interval Data

4.1 Inverted DEA

DEA gives an optimistic evaluation because it chooses the most favorable parameters for the evaluated DEA. Therefore DMUs may sometimes be positively

overrated. To moderate such a positive overassessment, we may add a negative assessment. The inverted DEA model [16] has been proposed to evaluate the inefficiency of a DMU. Then DMUs can be evaluated by the bipolar scale, efficiency vs. inefficiency. In order to introduce the bipolar assessment, we extend the inverted DEA to interval input-output data in this section.

In [16], inverted models corresponding to BCC models are formulated. In this section, we consider the inverted model corresponding to SBM model. The inverted SBM model can be formulated as

$$\begin{aligned} \text{minimize} \quad & \frac{1-\dfrac{1}{p}\displaystyle\sum_{k=1}^{p} s_k^-/y_{ro}}{1+\dfrac{1}{m}\displaystyle\sum_{i=1}^{m} s_i^+/x_{io}}, \\ \text{subject to} \quad & X\boldsymbol{\lambda}-\boldsymbol{s}^+=\boldsymbol{x}_o, \\ & Y\boldsymbol{\lambda}+\boldsymbol{s}^-=\boldsymbol{y}_o, \\ & \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda}=1, \\ & \boldsymbol{\lambda}\geq\boldsymbol{0},\ \boldsymbol{s}^-\geq\boldsymbol{0},\ \boldsymbol{s}^+\geq\boldsymbol{0}. \end{aligned} \tag{16}$$

In DEA, we search a possible activity having a larger output with a smaller input than the o-th DMU's activity and if there is no such activity, the o-th DMU is regarded as efficient. On the contrary, In the inverted DEA, we search a possible activity having a smaller output with a larger input than the o-th DMU's activity and if there is no such activity, the o-th DMU is regarded as inefficient. Therefore, the evaluation policy is totally opposite. The score obtained by the inverted SBM model can be seen as inefficiency score.

Given a pair of efficiency score θ and inefficiency score φ, the activities of DMUs can be classified into the following four categories:

High-class: DMUs with $\theta \geq \alpha$ and $\varphi < \beta$.
Commonplace: DMUs with $\theta < \alpha$ and $\varphi < \beta$.
Low-class: DMUs with $\theta < \alpha$ and $\varphi \geq \beta$.
Peculiar: DMUs with $\theta \geq \alpha$ and $\varphi \geq \beta$.

Here α and β are given thresholds.

4.2 Inverted DEA with Interval Input-Output Data

As the inverted DEA is parallel to DEA, we can easily extend the inverted DEA to the case of interval input-output data. The inverted SBM model for interval input-output data is formulated as

$$\text{minimize} \quad \eta_o = \frac{1-\dfrac{1}{2p}\displaystyle\sum_{k=1}^{p}\left(\frac{d_k^-}{y_{kq}^{\mathrm{R}}}+\frac{s_k^-}{y_{kq}^{\mathrm{R}}}\right)}{1+\dfrac{1}{2m}\displaystyle\sum_{i=1}^{m}\left(\frac{d_i^+}{x_{iq}^{\mathrm{R}}}+\frac{s_i^+}{x_{iq}^{\mathrm{R}}}\right)},$$

$$
\begin{aligned}
&\text{subject to}\\
&\quad \text{if } Q_1 \neq \text{LR},\\
&\quad \begin{cases} X_{i\cdot}^{Q_{13}}\boldsymbol{\lambda} - d_i^+ = x_{ig}^{Q_{14}}, \ i = 1, \ldots, m, \\ Y_{k\cdot}^{Q_{11}}\boldsymbol{\lambda} + d_k^- = y_{kq}^{Q_{12}}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_1 = \text{LR},\\
&\quad \begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\lambda} - d_i^+ \geq x_{iq}^{\mathrm{L}}, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\lambda} - d_i^+ \geq x_{iq}^{\mathrm{R}}, \ i = 1, \ldots, m, \\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\lambda} + d_k^- \leq y_{kq}^{\mathrm{L}}, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\lambda} + d_k^- \leq y_{kq}^{\mathrm{R}}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_2 \neq \text{L|R},\\
&\quad \begin{cases} X_{i\cdot}^{Q_{23}}\boldsymbol{\lambda} z_i^{1+} - s_i^+ = x_{ig}^{Q_{24}} z_i^{1+}, i = 1, \ldots, m, \\ Y_{k\cdot}^{Q_{21}}\boldsymbol{\lambda} z_k^{1-} + s_k^- = y_{kq}^{Q_{22}} z_k^{1-}, k = 1, \ldots, p, \end{cases}\\
&\quad \text{if } Q_2 = \text{L|R},\\
&\quad \begin{cases} X_{i\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_i^{1+} - s_i^+ \geq x_{iq}^{\mathrm{L}} z_i^{1+}, \ X_{i\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_i^{2+} - s_i^+ \geq x_{iq}^{\mathrm{R}} z_i^{2+}, \ i = 1, \ldots, m, \\ Y_{k\cdot}^{\mathrm{L}}\boldsymbol{\lambda} z_k^{2-} + s_k^- \leq y_{kq}^{\mathrm{L}} z_k^{2-}, \ Y_{k\cdot}^{\mathrm{R}}\boldsymbol{\lambda} z_k^{1-} + s_k^- \leq y_{kq}^{\mathrm{R}} z_k^{1-}, \ k = 1, \ldots, p, \end{cases}\\
&\quad \sum_{i=1}^{m} s_i^- + \sum_{k=1}^{p} s_k^+ \geq \epsilon, \ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\lambda} = 1, \ \boldsymbol{\lambda} \geq \mathbf{0}, \ \lambda_o = 0, \ \boldsymbol{d}^-, \ \boldsymbol{d}^+, \ \boldsymbol{s}^-, \ \boldsymbol{s}^+ \geq \mathbf{0},\\
&\quad z_i^{1-}, z_i^{2-}, z_k^{1+}, z_k^{2+} \in \{0, 1\}, \ i = 1, 2, \ldots, m, \ k = 1, 2, \ldots, p.
\end{aligned}
\tag{17}
$$

If no feasible solution exists, the o-th DMU is (Q_1-Q_2)-inefficient. If $\sum_{i=1}^{m} s_i^+ + \sum_{k=1}^{p} s_j^- = \epsilon$ holds at the obtained optimal solution, the o-th DMU is regarded as (Q_1-Q_2)-inefficient. Otherwise, the o-th DMU is not (Q_1-Q_2)-inefficient and the optimal value is the score called the (Q_1-Q_2)-inefficiency score. As Problem (13) is reduced to Problem (14), Problem (17) can be reduced to a simpler programming problem. Moreover, owing to Theorem 1, fourteen of them can be reduced to linear programming problems.

As we have 25 (Q_1-Q_2)-efficiencies, we have 25 (Q_1-Q_2)-inefficiencies. Combining those, we can analyze the efficiency-inefficiency of a DMU in various ways. However, this variety would be too huge. We may restrict ourselves within efficiencies and inefficiencies whose Q_1-Q_2 is Π-N, L-L, R-R, LR-L—R or N-Π.

Using (Q_1-Q_2)-efficiency and inefficiency scores, we can classify DMUs into the following five categories:

High-Class: DMUs which are (Q_1-Q_2)-efficient for some $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$ and not (Q_1-Q_2)-inefficient for all $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$. Among them, (Π-N)-efficient DMUs are the first-class.

Commonplace: DMUs which are neither (Q_1-Q_2)-efficient nor (Q_1-Q_2)-inefficient for all $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$.

Low-Class: DMUs which are not (Q_1-Q_2)-efficient for any $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$ and (Q_1-Q_2)-inefficient for some $Q_1 \in \mathcal{Q}_1$ and $Q_2 \in \mathcal{Q}_2$. Among them, (Π-N)-inefficient DMUs are the lowest.

Peculiar: DMUs which are (Π-N)-efficient and at the same time (Π-N)-inefficient.

Uncertain: DMUs which are (R-R)-efficient and at the same time (R-R)-inefficient but neither (Π-N)-efficient nor (Π-N)-inefficient.

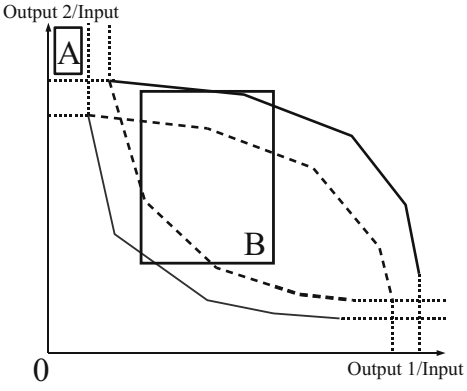


Fig. 2. Peculiar versus uncertain DMU

The difference between peculiar DMU and uncertain DMU can be illustrated in Figure 2. Figure 2 shows a case of one input and two outputs. The production possible set obtained from all DMUs except the o-th DMU are shown by four polygonal lines on (output 1/input)-(output 2/input) coordinate. If the activity of the o-th DMU is represented by box A, the o-th DMU becomes Π-N efficient since two concave polygonal lines are passing under box A. In this case, because two convex polygonal lines are passing right side of box A, the o-th DMU becomes Π-N inefficient, too. Therefore, the o-th DMU located at box A is peculiar. On the other hand, if the activity of the o-th DMU is represented by box B, the o-th DMU becomes R-R efficient since two concave polygonal lines are passing under the upper right corner point of box B. In this case, because two convex polygonal lines are passing over lower left corner point of box B, the o-th DMU becomes R-R inefficient, too. Therefore, the o-th DMU located at box B is uncertain. As shown in Figure 2, peculiarity indicates that DMU locates at the edge of production possibility set while uncertainty indicates that the input-output data of DMU is very wide.

5 Conclusions

In this paper, we have proposed SBM models with interval input-output data. By this approach, the efficiencies of DMUs can be evaluated qualitatively and quantitatively. Moreover, we have also proposed inverted DEA models with interval input-output data. With these models, bipolar evaluations of the efficiencies of DMUs are available. By the proposed approach, the robustness and possibility of the efficiency can be analyzed. The results are very different from the conventional approach using the center values of intervals.

Due to the limited space, we could not demonstrate the usefulness of the proposed approach. However, the proposed approach is applied to realworld data about activities of many Japanese banks. We observed that many kinds of efficiencies are obtained and that there exist DMUs included in different categories of the

bipolar analysis. By using interval data, a more detailed analysis with respect to the uncertainty is possible. The further applications and modifications of the proposed approach would be the future research topic.

Acknowledgments

This work has been partially supported by the Grant-in-Aid for Scientific Research (B) No.17310098.

References

1. Charnes, A., Cooper, W., Lewin, A.Y., Seiford, L.M. (eds.): Data Envelopment Analysis: Theory, Methodology and Applications. Kluwer Academic Publishers, Boston (1994)
2. Charnes, A., Neralic, L.: Sensitivity analysis of the additive model in data envelopment analysis. European Journal of Operational Research 48, 332–341 (1990)
3. Charnes, A., Rousseau, J., Semple, J.: Sensitivity and stability of efficiency classification in data envelopment analysis. Journal of Productivity Analysis 7, 5–18 (1996)
4. Cooper, W.W., Deng, H., Huang, Z., Susan, X.L.: Chance constrained programming approaches to congestion in stochastic data envelopment analysis. European Journal of Operational Research 155, 487–501 (2004)
5. Cooper, W.W., Park, K.S., Yu, G.: IDEA and AR-IDEA: Models for dealing with imprecise data in DEA. Management Science 45, 597–607 (1999)
6. Despotis, D.K., Smirlis, Y.G.: Data envelopment analysis with imprecise data. European Journal of Operational Research 140, 24–36 (2002)
7. Guo, P., Tanaka, H.: Fuzzy DEA: A perceptual evaluation method. Fuzzy Sets and Systems 119, 149–160 (2001)
8. Inuiguchi, M., Mizoshita, F.: Possibilistic data envelopment analysis with interval data: Part 1, Various efficiencies and their relations. In: Proceedings of SCIS & ISIS 2006, pp. 2293–2298 (2006)
9. Inuiguchi, M., Tanino, T.: Data envelopment analysis with fuzzy input-output data. In: Haimes, Y., Steuer, R.E. (eds.) Research and Practice in Multiple Criteria Decision Making, pp. 296–307. Springer, Berlin (2000)
10. Kao, C.: Interval efficiency measures in data envelopment analysis with imprecise data. European Journal of Operational Research 174, 1087–1099 (2006)
11. Kao, D., Liu, S.-T.: Fuzzy efficiency measures in data envelopment analysis. Fuzzy Sets and Systems 113, 427–437 (2000)
12. Mizoshita, F., Inuiguchi, M.: Possibilistic data envelopment analysis with interval data: Part 2, Efficiency tests and scores. In: Proceedings of SCIS & ISIS 2006, pp. 2299–2304 (2006)
13. Olesen, O.B., Peterson, N.C.: Chance constrained efficiency evaluation. Management Science 41, 442–457 (1995)
14. Tone, K.: An ε-free DEA and a new measure of efficiency. Journal of the Operations Research Society of Japan 36, 167–174 (1993)
15. Tone, K.: A slacks-based measure of efficiency in data envelopment analysis. European Journal of Operational Research 130, 498–509 (2001)
16. Yamada, Y., Matsui, T., Sugiyama, M.: New analysis of efficiency based on DEA. Journal of the Operations Research Society of Japan 37, 158–167 (1994) (in Japanese)

A Comparison between Two Approaches to Threat Evaluation in an Air Defense Scenario

Fredrik Johansson[1] and Göran Falkman[2]

[1] University of Skövde
fredrik.johansson@his.se
[2] University of Skövde
goran.falkman@his.se

Abstract. Threat evaluation is a high-level information fusion problem of high importance within the military domain. This task is the foundation for weapons allocation, where assignment of blue force (own) weapon systems to red force (enemy) targets is performed. In this paper, we compare two fundamentally different approaches to threat evaluation: Bayesian networks and fuzzy inference rules. We conclude that there are pros and cons with both types of approaches, and that a hybrid of the two approaches seems both promising and viable for future research.

Keywords: Bayesian networks, fuzzy inference rules, fuzzy logic, threat assessment, threat evaluation, weapons allocation.

1 Introduction

In a military environment it is often the case that decision makers in real-time have to evaluate the tactical situation and to protect defended assets against enemy targets by assigning available weapon systems to them [1]. In situations with several potential threats, a prioritizing of targets is necessary. Such an order of priority is often based on the degree of threat the targets represent to friendly defended assets. To determine which of several threats that represent the highest danger is of great importance, since errors such as prioritizing a lesser threat as a greater threat can result in engaging the wrong target, which often will have severe consequences [2]. The calculation of such threat values is often referred to as *threat evaluation*.

Threat evaluation is a part of threat analysis [3], which in an information fusion context is a central part of level 3 (impact assessment) in the well-known Joint Directories of Laboratories data fusion model [4]. Threat evaluation is the basis for *weapons allocation*, a process in which the decision-maker decides on which weapon system that should be assigned to a certain target. Research in high-level information fusion is still relatively immature [5]. As a consequence, different methods have been proposed for e.g. threat evaluation, but a systematic comparison between different approaches is lacking. Therefore, in this paper, we implement and compare two different artificial intelligence (AI) approaches to threat evaluation. The first method is the Bayesian network approach described

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 110–121, 2008.

in [6], while the other is implemented as fuzzy inference rules loosely based on a description in [7] and [8]. In a literature survey on threat evaluation, presented in [6], these methods have been identified as two fundamentally different approaches to threat evaluation.

The remainder of this paper is organized as follows. In Section 2, a description of the threat evaluation process is given. A central part of this process is a function that assigns threat values to pairs of targets and defended assets. In Section 3, the theory behind Bayesian networks is outlined, and a description of an approach to use Bayesian networks for threat value calculation is given. A similar description on the use of fuzzy logic for threat value calculation is given in Section 4. The properties of these two approaches are compared in Section 5. Finally, in Section 6 the paper is concluded, and thoughts regarding future work are presented.

2 Threat Evaluation

Consider a tactical situation where we have a set of defended assets $\mathbf{A} = \{A_1, \ldots, A_n\}$ that we are interested in to protect (e.g. friendly forces, bridges, and power plants). There is also a set of air targets $\mathbf{T} = \{T_1, \ldots, T_m\}$, which have been detected in the surveillance area. Now, the problem is to, for each target-defended asset pair (T_i, A_j) where $T_i \in \mathbf{T}$ and $A_j \in \mathbf{A}$, assign a threat value representing the degree of threat T_i poses to A_j, i.e., to define a function $f : \mathbf{T} \times \mathbf{A} \rightarrow [0, 1]$, assuming threat values between 0 (lowest possible threat value) and 1 (highest possible threat value). Based on the calculated threat values we will create a prioritized threat list, going from the most severe threat to the least. This prioritized threat list can later on be used as a basis for deciding how friendly weapon systems should be allocated to the targets.

A question then becomes how to assign threat values to target-defended asset pairs? In literature, it is often stated that a threat should be assessed as a combination of its capability and intent (cf. [9,2], and [10] (p.284)) to inflict injury or damage to defended assets. Many different capability and intent parameters for threat evaluation have been suggested throughout literature, and an overview of these are presented in [6]. Examples of parameters for threat evaluation are *target type*, *velocity*, and *time before hit* (TBH). In order to be able to calculate threat values we also need to specify the function f. The two main approaches to implement f are *rule-based* algorithms and *graphical models* [6]. We will in the following describe and compare fuzzy inference rules with Bayesian networks, where the former is an example of a rule-based algorithm, and the latter an example of a graphical model.

3 Bayesian Networks

A Bayesian network characterizes a problem domain consisting of a set of random variables $\mathbf{U} = \{X_1, ..., X_n\}$. These variables are in the Bayesian network represented as a set of corresponding nodes (vertices) $\mathbf{V}$ in an acyclic directed

graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where the set of edges $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ specifies (conditional) independence and dependence relations that hold between variables within the domain. Given the graph structure $\mathcal{G}$, a joint probability distribution P over $\mathbf{U}$ can be calculated from a set of local probability distributions associated with each node[1] X_i, using the chain rule of Bayesian networks

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|\mathbf{pa}_i), \tag{1}$$

where the set of local probability distributions are the (conditional) distributions in the product of Equation 1 (with $\mathbf{pa}_i$ we refer to an assignment of values to the parent set $\mathbf{PA}_i$ of node X_i). From this joint probability distribution we can perform *probabilistic inference*, i.e., to compute a posterior probability of interest conditional on available observations. This can mathematically be seen as given a set of observations (evidence) $\mathbf{z}$, a set of query variables $\mathbf{X}$, and a set $\mathbf{Y}$, including all variables except $\mathbf{X}$ and $\mathbf{Z}$, perform the computation of the posterior probability $P(\mathbf{X}|\mathbf{z})$. Given the full joint distribution encoded in the Bayesian network, this can in theory be computed by a brute force approach where we compute the answer by summing out the hidden (non-evidence) variables $\mathbf{Y}$:

$$P(\mathbf{X}|\mathbf{z}) = \frac{\sum_{\mathbf{Y}} P(\mathbf{X}, \mathbf{Y}, \mathbf{z})}{\sum_{\mathbf{X},\mathbf{Y}} P(\mathbf{X}, \mathbf{Y}, \mathbf{z})}. \tag{2}$$

However, this is in may cases computationally intractable, since we in this way do not make use of all the independences encoded in the network. In the work presented in this paper we have therefore used the junction tree algorithm (also known as the join tree algorithm) [11] for inference. See [12] for an excellent presentation of the algorithm.

3.1 Implementation

In the experiments, we have used the Bayesian network described in [6]. The posterior probability of interest here is $P(Threat = true|\mathbf{z})$, where $\mathbf{z}$ is a set of observations, as described above. These observations most often correspond to evidence regarding the information variables $Target\ type$, $Speed$, $Distance$, and $Time\,Before\,Hit$, however, such evidence can sometimes be soft or missing, due to limitations in sensor coverage. The structure of the Bayesian network is shown in Figure 1.

4 Fuzzy Logic

4.1 Fuzzy Sets and Membership Functions

Fuzzy logic builds upon the concept of *fuzzy sets*. Fuzzy sets can be seen as a generalization of standard crisp sets. In crisp set theory, members x of the

[1] Since the nodes in $\mathcal{G}$ are in one-to-one correspondence with the variables in $\mathbf{U}$, we use X_i to denote both variables and their corresponding nodes.

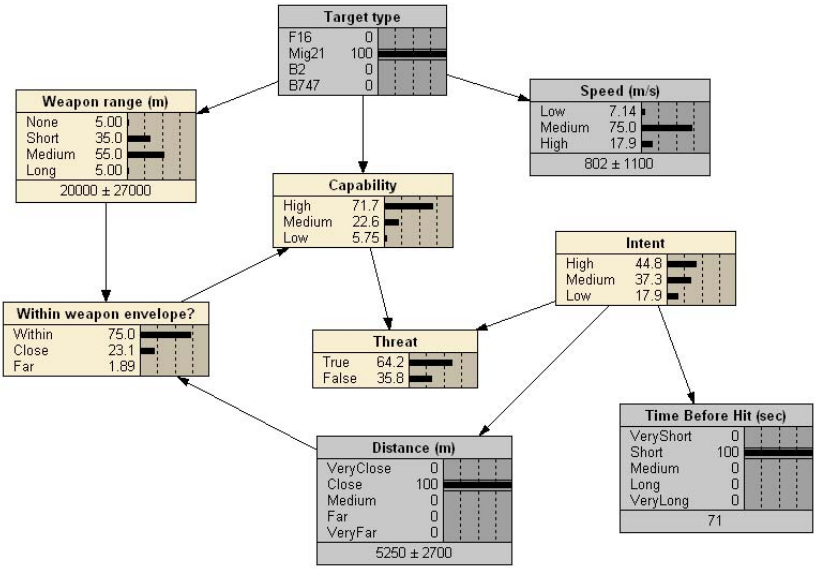


Fig. 1. Structure of the implemented Bayesian network

universal set $\mathbf{X}$ are either members or nonmembers of a set $\mathbf{A} \subseteq \mathbf{X}$. That is, there is a discrimination function μ_A that for each $x \in \mathbf{X}$ assigns a value $\mu_A(x)$, such that it is 0 if and only if $x \notin \mathbf{A}$, and 1 if and only if $x \in \mathbf{A}$. In fuzzy set theory this is generalized, so that the values assigned to x fall within a specified range, indicating the element's membership grade in the (fuzzy) set in question [13]. Larger values indicate a higher degree of membership, while lower values indicate a lower degree of membership. Hence, in fuzzy set theory there is a *membership function*

$$\mu_A : \mathbf{X} \rightarrow [0, 1]. \tag{3}$$

It must be noted that membership grades are not to be confused with probabilities. To illustrate how a fuzzy set constructed from a membership function μ_A might look like, we consider the variable *altitude*. For a given context, it might be hard to define crisp boundaries between *low*, *medium*, and *high* altitude. Instead, we can use membership functions to give more soft transitions between the possible states, as seen in Figure 2.

4.2 Logical Operations on Fuzzy Sets

Having defined the concepts of fuzzy sets, we must also be able to combine different fuzzy sets. We will here only consider two basic operations on fuzzy sets: intersection and union. In order for a function to be acceptable as an intersection operator, it must fulfill some basic requirements (cf. [14]). A function $\top$ fulfilling such requirements is known as a *t-norm*. The t-norm we will use in this paper is

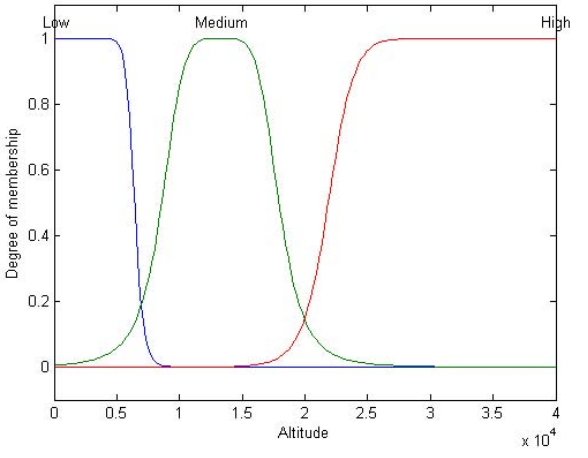


Fig. 2. Possible membership functions describing low, medium, and high altitude.

$\top_{min}$, i.e.,

$$\mu_{A\cap B}(x) = \top_{min}[\mu_A(x), \mu_B(x)]. \tag{4}$$

In the same manner, a *t-conorm* is a function $\perp$ that fulfills certain requirements for being a union operator. The t-conorm we have used in this paper is $\perp_{max}$, i.e.,

$$\mu_{A\cup B}(x) = \perp_{max}[\mu_A(x), \mu_B(x)]. \tag{5}$$

This means that for intersection of two fuzzy sets, we take the minimum, while for union, we take the maximum. According to [14], these operators are comfortable to work with both arithmetically and graphically, and are very often used.

4.3 Fuzzy Inference Rules

Fuzzy logic comprises of conditional statements such as

```
IF Speed == high AND Distance == close THEN Threat = high (1),
```

where the number within parentheses is the weight of the rule. These statements are known as fuzzy inference rules. A fuzzy inference system most often consists of several rules. As a first step in the inference process, each crisp numerical input is *fuzzified* over all qualifying fuzzy sets required by the rules, i.e., the degree to which each part of the antecedent of a rule is satisfied is determined using the specified membership functions. In next step, the fuzzy operators specified in Section 4.2, are applied to rules with more than one premise in their antecedent. In this way, all rules outputs a single value from their antecedent.

As seen in the fuzzy inference rule above, the consequent of a rule is a fuzzy set. This fuzzy set is for each rule reshaped using the output from its antecedent on a specified implication method (here we have used the *min*-operator, truncating

the fuzzy set specified by the consequent). Next, the output from all rules are combined into a single fuzzy set, using an aggregation operator (in our case the *max*-operator). Finally, the resulting aggregated fuzzy set is *defuzzified* into a single crisp value. In the experiments presented here, we have used the *centroid* calculation as a defuzzification method.

4.4 Implementation

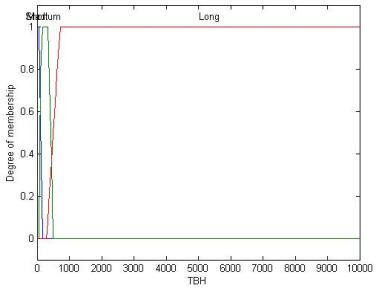
In the fuzzy inference system, we have used the same set of input variables as in the Bayesian network implementation, i.e., $Target\ type$, $Speed$, $Distance$, and $Time\ Before\ Hit$. The output of the system is the fuzzy set $Threat$. The ingoing rules in the implementation can be seen below (we have here used a more compact representation of the rules).

```
TargetType == F16 --> Threat = high (1)
TargetType == Mig21 --> Threat = medium (1)
TargetType == B2 --> Threat = high (1)
TargetType == B747 --> Threat = low (1)
TBH == short --> Threat = high (1)
TBH == medium --> Threat = medium (1)
TBH == long --> Threat = low (1)
Speed == low AND Distance == far --> Threat = low (1)
Speed == medium AND Distance == medium --> Threat = medium (1)
Speed == high AND Distance == close --> Threat = high (1)
Distance == far --> Threat = low (0.25)
Distance == medium --> Threat = medium (0.25)
Distance == close --> Threat = high (0.25)
```

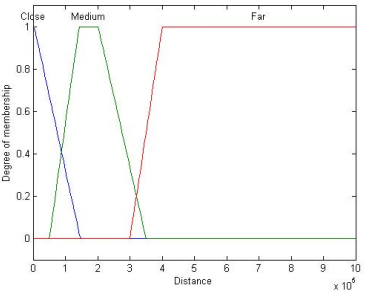
The membership functions for the input variables $Time\ Before\ Hit$, $Speed$, and $Distance$, together with the output variable $Threat$, can be seen in Figure 3. $Target\ type$ is not shown since it is implemented as an ordinary crisp set.

5 Comparison

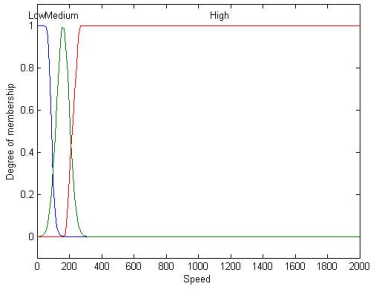
We have integrated the two methods for threat evaluation into a test bed implemented in Microsoft Visual C++ 2005. Scenarios generated in the STAGE Scenario tool are loaded into the system as XML files, and threat values are calculated in real-time for target-defended asset pairs, using the selected threat evaluation method. In order to demonstrate the approaches, we have constructed a test scenario consisting of a single defended asset and four air targets (one F-16, one B-2 bomber, and two Boeing 747). The initial heading of the targets, as well as their predefined way points, are shown in Figure 4. Speeds of the targets are close to constant, except for the F-16, which accelerates at the point of its turn against the defended asset. The reason for why we have chosen to illustrate this particular scenario is that it shows extremes with different kinds of air targets (fighters, bombers, and civilian aircrafts) with varying velocities. In



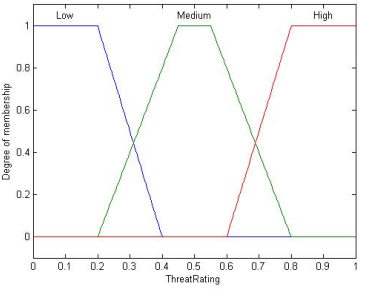
(a) Membership functions for TBH



(b) Membership functions for distance



(c) Membership functions for speed



(d) Membership functions for threat

Fig. 3. Illustration of the membership functions used in the fuzzy inference system for threat evaluation

Figure 5(a), the threat values inferred by the Bayesian network are shown, while the corresponding values inferred from the fuzzy inference system are shown in Figure 5(b) (one unit of time in the diagram corresponds to 50 updates, which is approximately ten seconds). As can be seen, the resulting ranking of the threats are quite similar, even though the threat values in general are higher in Figure 5(b). Another difference is that the threat values are more clearly separated in Figure 5(a). The output from the Bayesian network better reflects the authors opinions regarding the threat values in the scenario, but it should also be noted that the time needed to create the fuzzy inference rules was much shorter than for the development of the Bayesian network.

The output from a threat evaluation system obviously is dependent upon its parameter settings (e.g., the numbers used in the conditional probability tables in the Bayesian network, and the fuzzy sets used in the fuzzy inference system). Since there is arbitrarily many possible parameter settings, the comparison between two specific implementations is of minor interest. Rather, what is more interesting is the characteristics of the two approaches, such as whether the technique is transparent to the user or not.

As can be seen in Figure 5, the changes in threat values are smoother when using fuzzy logic than when Bayesian networks are used. Of course, the output from the Bayesian network becomes smoother the more states that are added.

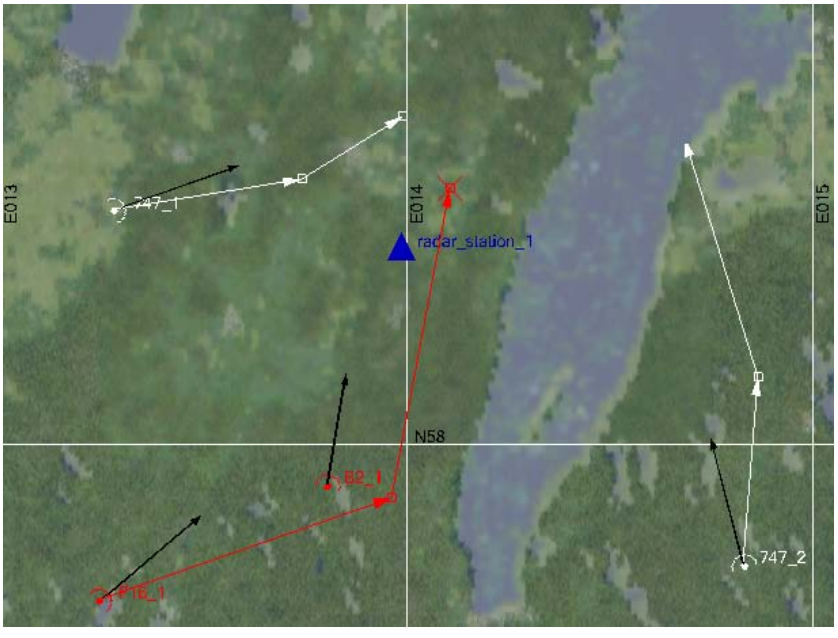
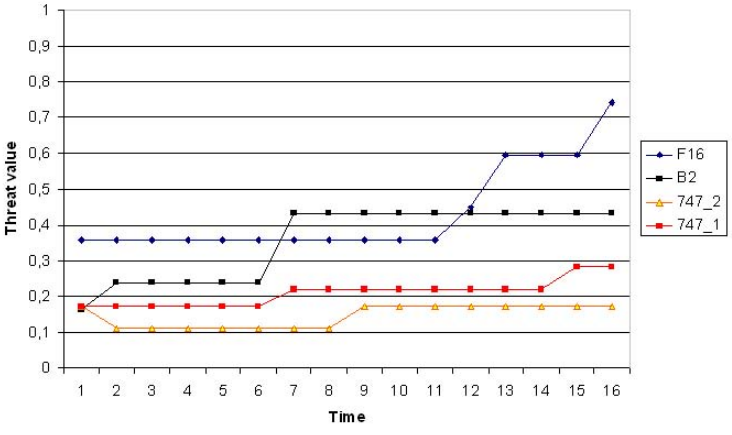


Fig. 4. Outline of the test scenario (created in STAGE Scenario)

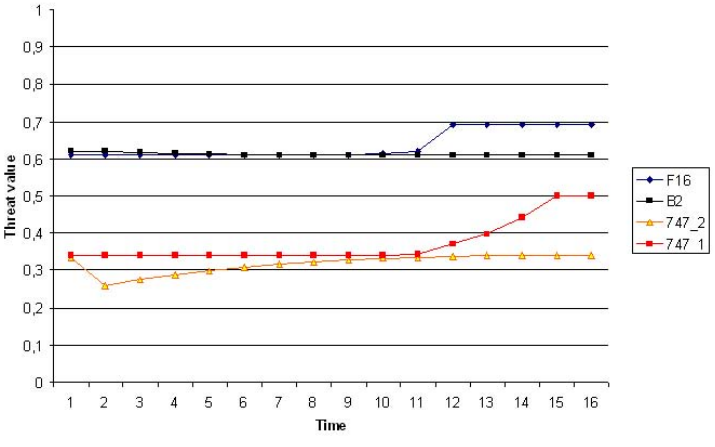
However, this comes with the cost of an increased burden in specifying more conditional probabilities. Also, the computational complexity increases with an increased number of states. This smoothness is an interesting property of fuzzy logic, compared to the hard boundaries between states when using Bayesian networks.

Another difference between the two approaches is that the inference goes in only one direction when using fuzzy inference rules (i.e., evidence regarding the input variables is entered, whereupon a crisp value for the output variable is calculated), while we in a Bayesian network can compute an arbitrary posterior probability of interest, given some evidence. As an example, we can use the Bayesian network to calculate $P(Speed = low|Distance = far)$ as well as $P(Threat = true|Distance = close, TBH = short, Speed = low)$. An even more important property of Bayesian networks is their ability to handle missing information. A fuzzy inference system is often dependent upon that we know the values of all its input variables, while this is not the case of the Bayesian network. In a hostile situation this can be very important, since we seldom have complete sensor coverage, sensors can be disturbed by countermeasures, etc.

In this paper we used the *min*- and *max*-operators as t-norm and t-conorm. However, many others have been proposed, and there is no common agreement on which that is the best. In fact, different t-norms and t-conorms seems to be appropriate for different types of problems [15]. The choice of t-norm and t-conorm will often influence the result of the fuzzy inference, hence, the choice



(a) Bayesian network



(b) Fuzzy inference rules

Fig. 5. Calculated threat value for different targets as a function of time

of the "right" t-norm and t-conorm can be crucial, and can be seen as a problem with the fuzzy logic approach. Bayesian networks does not have this problem of ad hoc solutions, since it has a sound mathematical foundation within probability theory. Nevertheless, fuzzy inference rules are very appealing in that they are easy to work with.

On the opposite, to obtain the probabilities that are required for the conditional probability tables in a Bayesian network can often be daunting [16]. This problem can be solved by learning the probabilities from data, however, this demands large amounts of data that seldom is available in this domain.

6 Conclusion and Future Work

In this paper, we have presented the problem of threat value calculation. This problem, often referred to as threat evaluation, is a high-level information fusion problem of high importance. We have presented two different approaches to threat evaluation; Bayesian networks and fuzzy inference rules. Implementations of the two approaches have been integrated into a threat evaluation system, working as a test bed for different threat evaluation techniques. The outputs from the two implemented threat evaluation approaches on a specified example scenario have been shown. The implemented Bayesian network's calculated threat values were better separated and more close to the authors opinion regarding the threat values in the scenario, however, the Bayesian network demanded more development time than the fuzzy inference system.

There are pros and cons with both approaches to threat evaluation. Strengths of Bayesian networks are that they have a sound mathematical foundation within probability theory, the ability to handle missing evidence, and that the network can be used to calculate an arbitrary probability of interest. Strengths of fuzzy inference rules are that they more quickly and easily can be created by non-professionals, and that they can produce a smoother change in output without the need for very fine discretization. Once the models have been constructed, both fuzzy inference rules and Bayesian networks are relatively transparent to the user. This is important since human users must be able to trust the system they are using. This is especially true for such critical applications as threat evaluation systems. This can be compared to black box techniques such as artificial neural networks, which are not appropriate for this kind of applications, due to their opaqueness.

Finally, an important difference between the two approaches is their ability to handle uncertain input data. In the Bayesian approach, uncertain evidence can be handled elegantly by using Pearl's method of virtual evidence (cf. [17]), while in the fuzzy logic approach, uncertain evidence can not be handled explicitly. Since target tracks that are input to the threat evaluation process almost always are imperfect to some degree, Bayesian networks are to be preferred to fuzzy logic from a theoretical uncertainty management perspective.

6.1 Future Work

We are interested in making use of the advantages of both Bayesian networks and fuzzy inference rules for future threat evaluation systems. This can be done either by creating an ensemble of the different techniques, or by incorporating one into the other, e.g., try to use fuzzy sets instead of ordinary crisp sets in the Bayesian network.

Threat evaluation is a basis for deciding on which (if any) weapon(s) to allocate to a specific target. However, to evaluate how large threat a target represents to a defended asset is not enough for making such a decision. We also need to know how valuable our defended assets are, what abilities we have to engage a specific target, the probability of successful engagement, etc. We think that

implementation of a module for making such assessments will make it easier to make better comparisons of threat evaluation techniques in the future.

Acknowledgment

This research has been supported by a grant from the Knowledge Foundation (project number: 2003/0104) to the Information Fusion research program at the University of Skövde (www.infofusion.se). The core of the BN implementation in this project is based on Netica's C++-API (Norsys Software Corp.). For the implementation of the fuzzy inference rules, we have used the Fuzzy Logic Toolbox in MATLAB. For the simulations we have used the STAGE Scenario tool from Presagis Inc.

References

1. Roux, J.N., van Vuuren, J.H.: Threat evaluation and weapon assignment decision support: A review of the state of the art. ORiON 23, 151–186 (2007)
2. Roy, J., Paradis, S., Allouche, M.: Threat evaluation for impact assessment in situation analysis systems. In: Kadar, I. (ed.) Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XI, vol. 4729, pp. 329–341 (2002)
3. Paradis, S., Benaskeur, A., Oxenham, M., Cutler, P.: Threat evaluation and weapons allocation in network-centric warfare. In: Proceedings of the 8th International Conference on Information Fusion (2005)
4. Steinberg, A., Bowman, C., White, F.: Revisions to the JDL data fusion model. In: Proceedings of the SPIE Sensor Fusion: Architectures, Algorithms, and Applications III, pp. 430–441 (1999)
5. Phister, P.W., Plonisch, I.: Data fusion "cube": A multi-dimensional perspective. In: Proceedings of the Command and Control Research and Technology Symposium (CCRTS 2002) (2002)
6. Johansson, F., Falkman, G.: A Bayesian network approach to threat evaluation with application to an air defense scenario. In: Proceedings of the 11th International Conference on Information Fusion (2008)
7. Liang, Y.: An approximate reasoning model for situation and threat assessment. In: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (2007)
8. Liang, Y.: A fuzzy knowledge based system in situation and threat assessment. Journal of Systems Science & Information 4(4), 791–802 (2006)
9. Nguyen, X.: Threat assessment in tactical airborne environments. In: Proceedings of the Fifth International Conference on Information Fusion (2002)
10. Waltz, E.L., Llinas, J.: Multisensor Data Fusion. Artech House (1990)
11. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society B50, 157–224 (1988)
12. Huang, C., Darwiche, A.: Inference in belief networks: A procedural guide. International Journal of Approximate Reasoning 15(3), 225–263 (1996)
13. Klir, G.J., Folger, T.A.: Fuzzy sets, uncertainty, and information. Prentice-Hall, Inc., Upper Saddle River (1987)

14. Kruse, R., Gebhardt, J.E., Klowon, F.: Foundations of Fuzzy Systems. John Wiley & Sons, Inc., New York (1994)
15. Kreinovich, V., Nguyen, H.T.: Which fuzzy logic is the best: Pragmatic approach (and its theoretical analysis). Fuzzy Sets and Systems 157(5), 611–614 (2006)
16. Druzdzel, M., van der Gaag, L.: Building probabilistic networks: where do the numbers come from? IEEE Transactions on Knowledge and Data Engineering 12(4), 481–486 (2000)
17. Chan, H., Darwiche, A.: On the revision of probabilistic beliefs using uncertain evidence. Artificial Intelligence 163(1), 67–90 (2005)

Fuzzy Classification Function of Standard Fuzzy *c*-Means Algorithm for Data with Tolerance Using Kernel Function

Yuchi Kanzawa[1], Yasunori Endo[2], and Sadaaki Miyamoto[2]

[1] Shibaura Institute of Technology, 3-7-5 Toyosu, Koto, Tokyo, 135-8548, Japan
kanzawa@sic.shibaura-it.ac.jp
[2] University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan

Abstract. In this paper, the fuzzy classification functions of the standard fuzzy c-means for data with tolerance using kernel functions are proposed.

First, the standard clustering algorithm for data with tolerance using kernel functions are introduced. Second, the fuzzy classification function for fuzzy *c*-means without tolerance using kernel functions is discussed as the solution of a certain optimization problem. Third, the optimization problem is shown so that the solutions are the fuzzy classification function values for the standard fuzzy *c*-means algorithms using kernel functions with respect to data with tolerance. Fourth, Karush-Kuhn-Tucker conditions of the objective function is considered, and the iterative algorithm is proposed for the optimization problem. Some numerical examples are shown.

1 Introduction

Fuzzy *c*-means (FCM) [1] is one of the well-known fuzzy clusterings and many FCM variants have been proposed after FCM. In these variants, FCM algorithm based on the concept of regularization by entropy has been proposed by one of the authors [2]. This algorithm is called regularized entropy FCM (eFCM) and is discussed not only for its usefulness but also for its mathematical relations with other techniques.

There are many cases that data have some errors in clustering. Many clustering algorithms have been proposed to classify the data with error. In these algorithms, two FCM algorithms considering the inner point in the region of data have been proposed by some of the authors [3]. The one is derived from the standard FCM and the other is from eFCM. Because these algorithms interpret the error of the data as an arbitrary point in a given region called "tolerance", these are called FCMs for data with tolerance (FCM-Ts). While FCM-T has been proposed by considering data including some errors, it can be interpreted as the method obtaining clearer classification result than the plain FCM by moving the data. In the case that some data are classified vaguely by the plain FCM, there is a chance that we can obtain the classification result closer to human judge by daring to move the data, that is, by FCM-Ts.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 122–133, 2008.

FCM-Ts are pointed out that it is difficult to classify data with nonlinear borders because FCM-Ts use squared distance between each datum and each cluster center for their dissimilarity. In order to solve this problem of FCM-Ts, two new algorithms [5] has been proposed using nonlinear transformations from the original pattern space into a higher dimensional feature space with kernel functions in Support Vector Machines (SVM) [6]. The one is called K-sFCM-T derived from sFCM-T and the other is K-eFCM-T derived from eFCM-T.

Fuzzy classification functions for the standard FCM (sFCM) and eFCM without tolerance have been proposed [2], respectively, in order to show how prototypical an arbitrary point in the data space is to a cluster by extending the membership to the whole space. Fuzzy classification function can be used to classify a brand-new data to a cluster after fuzzy clustering (sFCM or eFCM) is done for the initially given data. It is a kind of supervised classification. Fuzzy classification function can be also used to investigate the feature of the corresponding clustering algorithm since it clarify the classifying situation in whole space than only memberships for finite number of data. Such fuzzy classifications for FCM-Ts have not been proposed yet.

In this paper, we propose the fuzzy classification function for K-sFCM-T. This function values are calculated by iterative algorithms obtained from a certain optimization problem. which are derived from that fuzzy classification function of FCM without tolerance can be interpreted as an optimization problem. We also show some numerical examples of such fuzzy classification function.

The contents of this paper are the followings. In the second section, we define some notation, introduce K-sFCM-T. In the third section, we propose the fuzzy classification functions for K-sFCM-T by an iterative algorithm, which are led from the optimization problem based on that fuzzy classification function value for FCM without tolerance can be interpreted as the solution of a certain optimization problem. In the fourth section, some numerical examples are shown. In the last section, we conclude this paper.

2 Preliminaries

In this section, we define some notation and introduce the standard fuzzy c-means for data with tolerance using kernel function (K-sFCM-T). In the first subsection, we define some notations which are the data for clustering, the membership by which the each data belongs to the each cluster, the cluster centers, the tolerance for the data and the maximum tolerance for the data. In the second subsection, we introduce K-sFCM-T which is the basis of our main theme.

2.1 Notation

In this subsection, we define some notation which are the data for clustering, the membership by which the each data belongs to the each cluster, the cluster centers, the tolerance for the data and the maximum tolerance for the data.

The data set $x = \{x_i \mid x_i \in \mathrm{R}^p, i \in \{1, \ldots, N\}\}$ is given. The membership by which x_i belongs to the j-th cluster is denoted by $u_{i,j}$ $(i \in \{1, \cdots, N\},$

$j \in \{1, \cdots, C\}$) and the set of $u_{i,j}$ is denoted by $u \in \mathrm{R}^{N \times C}$ called the partition matrix. The constraint for u is

$$\sum_{j=1}^{C} u_{i,j} = 1 \quad (0 \leq u_{i,j} \leq 1).$$

A high-dimensional feature space used in SVM is denoted by H, whereas the original space R^p is called data space. H may be an infinite-dimensional metric space. Let the inner product denoted by $\langle \cdot, \cdot \rangle$. The norm of H for an element $g \in \mathrm{H}$ is given by $\|g\|_{\mathrm{H}}^2 = \langle g, g \rangle$. A transformation $\Phi : \mathrm{R}^p \to \mathrm{H}$ is employed whereby x_i is mapped into $\Phi(x_i)$. Explicit representation of $\Phi(x)$ is not usable in general but the inner product $\langle \Phi(x), \Phi(y) \rangle$ can be expressed by a kernel function

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle. \tag{1}$$

A representative kernel function is the radial basis function (RBF) kernel described as $K(x, y) = \exp(-\sigma^{-2}\|x - y\|_2^2)$ with a positive parameter σ. The cluster center set in H is denoted by $W = \{W_j \mid W_j \in \mathrm{H}, j \in \{1, \ldots, C\}$. $v = \{v_j \mid v_j \in \mathrm{R}^p, j \in \{1, \ldots, C\}\}$. The tolerance for the data x in H is denoted by $E = \{E_i \mid E_i \in \mathrm{H}, i \in \{1, \ldots, N\}\}$. The maximum tolerance is denoted by $\kappa = \{\kappa_i \mid \kappa_i \in \mathrm{R}_+, i \in \{1, \ldots, N\}\}$.

2.2 Standard FCM for Data with Tolerance Using Kernel Functions

In this subsection, we introduce K-sFCM-T [5]. This algorithm is the basis of our main theme.

K-sFCM-T is the algorithm obtained by solving the following optimization problem:

$$\underset{u,E,W}{\text{minimize}}\, J_{m,\mathrm{k,t}}(u, E, W) \quad \text{under} \quad \begin{cases} \displaystyle\sum_{j=1}^{C} u_{i,j} = 1, \\ \|E_i\|_{\mathrm{H}}^2 \leq \kappa_i^2 \quad (\kappa_i > 0), \end{cases} \tag{2}$$

where

$$J_{m,\mathrm{k,t}}(u, E, W) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j}^m \|\Phi(x_i) + E_i - W_j\|_{\mathrm{H}}^2. \tag{3}$$

The parameter m is the power satisfying $m > 1$. The optimal solution u are obtained by the following algorithm.

Algorithm 1 (K-sFCM-T)

Step 1. *Give the value of m and κ. Select a kernel function $K : \mathrm{R}^p \times \mathrm{R}^p \to \mathrm{R}$. Set the initial cluster center v.*

Step 2. *Calculate* $Y_{i,j}^{(0)}$, $Z_{j,\tilde{j}}^{(0)}$ *and* $d_{i,j}^{(0)}$ *such that*

$$Y_{i,j}^{(0)} = K(x_i, v_j^{(0)}), \quad Z_{j,\tilde{j}}^{(0)} = K(v_j^{(0)}, v_{\tilde{j}}^{(0)}), \quad d_{i,j}^{(0)} = K(x_i, x_i) - 2Y_{i,j}^{(0)} + Z_{j,j}^{(0)}. \tag{4}$$

Set t *be* 0.

Step 3. *Calculate* $u_{i,j}^{(t)}$, $U_j^{(t)}$, $\mu_i^{(t)}$, $\alpha_i^{(t+1)}$, $Y_{i,j}^{(t+1)}$, $Z_{j,\tilde{j}}^{(t+1)}$ *and* $d_{i,j}^{(t+1)}$ *such that*

$$u_{i,j}^{(t)} = 1/\sum_{k=1}^{C} \left(\frac{d_{i,j}^{(t)}}{d_{i,k}^{(t)}} \right)^{1/(m-1)}, \quad U_j^{(t)} = \sum_{i=1}^{N} {u_{i,j}^{(t)}}^m, \quad \mu_i^{(t)} = \sum_{j=1}^{C} {u_{i,j}^{(t)}}^m, \tag{5}$$

$$\begin{aligned} \alpha_i^{(t+1)} = \min\Bigg\{ & \kappa_i \Bigg({\mu_i^{(t)}}^2 K(x_i, x_i) - 2\mu_i^{(t)} \sum_{j=1}^{C} {u_{i,j}^{(t)}}^m Y_{i,j}^{(t)} \\ & + \sum_{j=1}^{C} \sum_{k=1}^{C} {u_{i,j}^{(t)}}^m {u_{i,k}^{(t)}}^m Z_{j,k}^{(t)} \Bigg)^{-1/2}, {\mu_i^{(t)}}^{-1} \Bigg\}, \end{aligned} \tag{6}$$

$$\begin{aligned} Y_{i,j}^{(t+1)} = & {U_j^{(t+1)}}^{-1} \sum_{k=1}^{N} {u_{k,j}^{(t+1)}}^m \Bigg[\left(1 - \alpha_k^{(t+1)} \mu_k^{(t+1)} \right) K(x_i, x_k) \\ & + \alpha_k^{(t+1)} \sum_{\ell=1}^{C} {u_{k,\ell}^{(t+1)}}^m Y_{i,\ell}^{(t)} \Bigg], \end{aligned} \tag{7}$$

$$\begin{aligned} Z_{j,\tilde{j}}^{(t+1)} = & {U_j^{(t+1)}}^{-1} {U_{\tilde{j}}^{(t+1)}}^{-1} \sum_{k=1}^{N} \sum_{\ell=1}^{N} {u_{k,j}^{(t+1)}}^m {u_{\ell,\tilde{j}}^{(t+1)}}^m \\ & \cdot \Bigg[\left(1 - \alpha_k^{(t+1)} \mu_k^{(t+1)} \right) \left(1 - \alpha_\ell^{(t+1)} \mu_\ell^{(t+1)} \right) K(x_k, x_\ell) \\ & + \left(1 - \alpha_k^{(t+1)} \mu_k^{(t+1)} \right) \alpha_\ell^{(t+1)} \sum_{r=1}^{C} {u_{\ell,r}^{(t+1)}}^m Y_{k,r}^{(t+1)} \\ & + \alpha_k^{(t+1)} \left(1 - \alpha_\ell^{(t+1)} \mu_\ell^{(t+1)} \right) \sum_{q=1}^{C} {u_{k,q}^{(t+1)}}^m Y_{\ell,q}^{(t+1)} \\ & + \alpha_k^{(t+1)} \alpha_\ell^{(t+1)} \sum_{q=1}^{C} \sum_{r=1}^{C} {u_{k,q}^{(t+1)}}^m {u_{\ell,r}^{(t+1)}}^m Z_{q,r}^{(t)} \Bigg], \end{aligned} \tag{8}$$

$$d_{i,j}^{(t+1)} = \left(1 - \alpha_i^{(t+1)} \mu_i^{(t+1)}\right)^2 K(x_i, x_i) + 2\left(1 - \alpha_i^{(t+1)} \mu_i^{(t+1)}\right) \cdot \sum_{k=1}^{C} \left(\alpha_i^{(t+1)} {u_{i,k}^{(t+1)}}^m - \delta_{k,j}\right) Y_{i,k}^{(t+1)} + \sum_{k=1}^{C} \sum_{\ell=1}^{C} \left(\alpha_i^{(t+1)} {u_{i,k}^{(t+1)}}^m - \delta_{k,j}\right) \left(\alpha_i^{(t+1)} {u_{i,\ell}^{(t+1)}}^m - \delta_{\ell,j}\right) Z_{k,\ell}^{(t)}. \tag{9}$$

where $\delta_{k,j}$ is the Kronecker's delta.

Step 4. *Check the stopping criterion. If the criterion is not satisfied, go back to* Step 3.

3 Fuzzy Classification Function of K-FCM-T

In this section, we propose the fuzzy classification function for K-sFCM-T (Algorithm 1).

In the first subsection, we interpret fuzzy classification function for FCM as the solution of an optimization problem. In the second subsection, we define fuzzy classification function for K-sFCM-T as a new optimization problem, and construct a new algorithm to obtain the fuzzy classification function value for K-sFCM-T by solving the optimization problem. In the third subsection, we do for K-eFCM-T.

3.1 Fuzzy Classification Function for FCM

In this subsection, we interpret fuzzy classification function for FCM as the solution of an optimization problem. This interpretation is used for considering fuzzy classification function of K-FCM-Ts.

Fuzzy classification functions [2] are available in FCMs which show how prototypical an arbitrary point in the data space is to a cluster by extending the membership $u_{i,j}$ to the whole space.

Fuzzy classification function $\tilde{u}_j(\tilde{x})$ for sFCM with respect to a brand-new datum $\tilde{x} \in \mathrm{R}^p$ is defined as

$$\tilde{u}_j(\tilde{x}) = 1 / \sum_{k=1}^{C} \left(\frac{\tilde{d}_j}{\tilde{d}_k},\right)^{\frac{1}{m-1}}, \quad \text{where} \quad \tilde{d}_j = \|\tilde{x} - v_j\|^2 \tag{10}$$

and where v_j is j-th cluster center obtained from sFCM. This fuzzy classification function value is the solution $\tilde{u}_j$ of the following optimization problem:

$$\underset{\tilde{u}}{\text{minimize}} \sum_{j=1}^{C} \tilde{u}_j^m \|\tilde{x} - v_j\|^2 \quad \text{under} \quad \sum_{j=1}^{C} \tilde{u}_j = 1 \tag{11}$$

because its Lagrange function

$$L_{m,\mathrm{FCF}}(\tilde{u}) = \sum_{j=1}^{C} \tilde{u}_j^m \|\tilde{x} - v_j\|^2 + \gamma \left(\sum_{j=1}^{C} \tilde{u}_j - 1 \right) \tag{12}$$

has KKT condition as Eq. (10), where γ is Lagrange multiplier.

Based on the above interpretation, fuzzy classification function for K-sFCM-T is proposed by solving the corresponding optimization problem with tolerance in the next subsection.

3.2 Fuzzy Classification Function for K-sFCM-T

In this subsection, we propose a fuzzy classification function for K-sFCM-T (Algorithm 1) as an iterative algorithm. First, we define an optimization problem whose component of the solution is the fuzzy classification function value for K-sFCM-T with respect to a brand-new datum. Secondly, we solve the optimization problem with KKT conditions. Finally, we construct a new algorithm from the solutions of the optimization problem.

We define the following optimization problem whose component $\tilde{u}$ of solution $(\tilde{u}, \tilde{E})$ is the fuzzy classification value for K-sFCM-T with respect to a brand-new datum $\tilde{x} \in \mathrm{R}^p$:

$$\underset{\tilde{u},\tilde{E}}{\text{minimize}}\, J_{m,\mathrm{k,t,FCF}}(\tilde{u}, \tilde{E}) \quad \text{under} \begin{cases} \displaystyle\sum_{j=1}^{C} \tilde{u}_j = 1, \\ \|\tilde{E}\|_{\mathrm{H}}^2 \le \tilde{\kappa}^2 \quad (\tilde{\kappa} > 0), \end{cases} \tag{13}$$

where

$$J_{m,\mathrm{k,t,FCF}}(\tilde{u}, \tilde{E}) = \sum_{j=1}^{C} \tilde{u}_j^m \|\Phi(\tilde{x}) + \tilde{E} - W_j\|^2 \tag{14}$$

and where W_j is j-th cluster center obtained from K-sFCM-T. $\tilde{E} \in \mathrm{H}$ is the tolerance for $\tilde{x}$ and $\tilde{\kappa} \in \mathrm{R}_+$ is the maximum tolerance for $\tilde{x}$. Its Lagrange function $L_{m,\mathrm{k,t,FCF}}$ is as below:

$$\begin{aligned} L_{m,\mathrm{k,t,FCF}}(\tilde{u}, \tilde{E}) = & \sum_{j=1}^{C} \tilde{u}_j^m \|\Phi(\tilde{x}) + \tilde{E} - W_j\|_{\mathrm{H}}^2 \\ & + \gamma \left(\sum_{j=1}^{C} \tilde{u}_j - 1 \right) + \delta \left(\|\tilde{E}\|_{\mathrm{H}}^2 - \tilde{\kappa}^2 \right), \end{aligned} \tag{15}$$

where γ and δ are Lagrange multipliers. Karush-Kuhn-Tucker conditions of $L_{m,\mathrm{k,t,FCF}}$ are below:

$$\frac{\partial L_{m,\mathrm{k,t,FCF}}}{\partial \tilde{u}_j} = 0, \quad \frac{\partial L_{m,\mathrm{k,t,FCF}}}{\partial \tilde{\varepsilon}} = 0, \quad \gamma \frac{\partial L_{m,\mathrm{k,t,FCF}}}{\partial \gamma} = 0, \quad \gamma \le 0. \tag{16}$$

The necessary condition that $\tilde{u_j}$ is the minimizing value is as below:

$$\tilde{u}_j = 1/\sum_{k=1}^{C} \left(\frac{\tilde{d}_j}{\tilde{d}_k} \right)^{\frac{1}{m-1}}, \tag{17}$$

$$\tilde{E} = -\tilde{\alpha} \left(\sum_{j=1}^{C} \tilde{u}_j^m (\Phi(\tilde{x}) - W_j) \right), \tag{18}$$

where

$$\tilde{d}_j = \|\Phi(\tilde{x}) + \tilde{E} - W_j\|_{\mathrm{H}}^2, \tag{19}$$

$$\tilde{\alpha} = \min \left\{ \tilde{\kappa} \left\| \sum_{j=1}^{C} \tilde{u}_j^m (\Phi(\tilde{x}) - W_j) \right\|^{-1}, \tilde{\mu}^{-1} \right\}, \tag{20}$$

$$\tilde{\mu} = \sum_{j=1}^{C} \tilde{u}_j^m. \tag{21}$$

Since we don't know the explicit form of Φ, we cannot calculate (17)–(21) directly. All we can know is the inner product as the value of the kernel function. Therefore, we lead other forms with the kernel function K instead of (17)–(21). Noting that (18) implies

$$\begin{aligned} \Phi(\tilde{x}) + \tilde{E} &= \Phi(\tilde{x}) - \tilde{\alpha}\tilde{\mu}\Phi(\tilde{x}) + \tilde{\alpha} \sum_{j=1}^{C} \tilde{u}_j^m W_j \\ &= (1 - \tilde{\alpha}\tilde{\mu})\Phi(\tilde{x}) + \tilde{\alpha} \sum_{j=1}^{C} \tilde{u}_j^m W_j \end{aligned} \tag{22}$$

and that the norm for H is described with inner product form, d_j and α_i are rewritten as

$$\begin{aligned} d_j &= \|\Phi(\tilde{x}) + \tilde{E} - W_j\|_{\mathrm{H}}^2 \\ &= \left\| (1 - \tilde{\alpha}\tilde{\mu})\Phi(\tilde{x}) + \tilde{\alpha} \sum_{k=1}^{C} \tilde{u}_k^m W_k - W_j \right\|_{\mathrm{H}}^2 \\ &= \left\| (1 - \tilde{\alpha}\tilde{\mu})\Phi(\tilde{x}) + \sum_{k=1}^{C} \left(\tilde{\alpha}\tilde{u}_k^m - \delta_{k,j} \right) W_k \right\|_{\mathrm{H}}^2 \\ &= (1 - \tilde{\alpha}\tilde{\mu})^2 \left\langle \Phi(\tilde{x}), \Phi(\tilde{x}) \right\rangle \\ &\quad + 2(1 - \tilde{\alpha}\tilde{\mu}) \sum_{k=1}^{C} \left(\tilde{\alpha}\tilde{u}_k^m - \delta_{k,j} \right) \left\langle \Phi(\tilde{x}), W_k \right\rangle \\ &\quad + \sum_{k=1}^{C} \sum_{\ell=1}^{C} \left(\tilde{\alpha}\tilde{u}_k^m - \delta_{k,j} \right) \left(\tilde{\alpha}\tilde{u}_\ell^m - \delta_{\ell,j} \right) \left\langle W_k, W_l \right\rangle, \end{aligned} \tag{23}$$

$$\tilde{\alpha} = \min\left\{ \tilde{\kappa}\left(\tilde{\mu}^2\langle\Phi(\tilde{x}),\Phi(\tilde{x})\rangle - 2\tilde{\mu}\sum_{j=1}^{C}\tilde{u}_j^m\langle\Phi(\tilde{x}),W_j\rangle + \sum_{j=1}^{C}\sum_{k=1}^{C}\tilde{u}_j^m\tilde{u}_k^m\langle W_j,W_k\rangle\right)^{-1/2}, \tilde{\mu}^{-1}\right\}, \tag{24}$$

where $\delta_{i,j}$ is Kronecker's delta. W_j satisfies that

$$W_j = U_j^{-1}\sum_{i=1}^{N}u_{i,j}^m\left((1-\alpha_i\mu_i)\Phi(x_i) + \alpha_i\sum_{k=1}^{C}u_{i,k}^m W_k\right) \tag{25}$$

from KKT conditions for Eq.(3)

$$W_j = U_j^{-1}\sum_{i=1}^{N}u_{i,j}^m(\Phi(x_i)+E_i), \quad E_i = -\alpha_i(\sum_{j=1}^{C}u_{i,j}^m(\Phi(x_i)-W_j)), \tag{26}$$

thus, $\langle\Phi(\tilde{x}),W_j\rangle$ satisfies

$$\langle\Phi(\tilde{x}),W_j\rangle = (U_j)^{-1}\sum_{i=1}^{N}(u_{i,j})^m\left[(1-\alpha_i\mu_i)\langle\Phi(\tilde{x}),\Phi(x_i)\rangle + \alpha_i\sum_{\ell=1}^{C}(u_{i,\ell})^m\langle\Phi(\tilde{x}),W_l'\rangle\right]. \tag{27}$$

Hence, $\langle\Phi(\tilde{x}),W_j\rangle$, denoted by y_j, can be obtained by solving the linear equation $Ay=b$, where the element $a_{j,\tilde{j}}$ of A and b_j of b are

$$a_{j,\tilde{j}} = \delta_{j,\tilde{j}} - U_j^{-1}\sum_{\ell}^{N}u_{\ell,j}^m\alpha_\ell u_{\ell,k}^m, \quad b_j = U_j^{-1}\sum_{k=1}^{N}u_{k,j}^m(1-\alpha_k\mu_k)K(x,x_k). \tag{28}$$

$\left\langle W_j, W_{\tilde{j}}\right\rangle$ has been already obtained as $Z_{j,\tilde{j}}$. It is natural that the initial tolerance $\tilde{E}$ is set to 0, which is achieved by $\tilde{\alpha}=0$. From these settings, the initial dissimilarities $\tilde{d}_j$ are given by:

$$\begin{aligned}\tilde{d}_j &= \|\Phi(\tilde{x})+\tilde{E}-W_j\|_{\mathrm{H}}^2\\ &= \|\Phi(\tilde{x})-W_j\|_{\mathrm{H}}^2\\ &= \langle\Phi(\tilde{x}),\Phi(\tilde{x})\rangle - 2\langle\Phi(\tilde{x}),W_j\rangle + \langle W_j,W_j\rangle\\ &= K(\tilde{x},\tilde{x}) - 2y_j + Z_{j,j},\end{aligned} \tag{29}$$

from which we can update $\tilde{u}$, $\tilde{\mu}$, $\tilde{\alpha}$ and $\tilde{d}$ by (17), (21), (24) and (23).

From the above discussion, we obtain the following iterative algorithm.

Algorithm 2

Step 1. *Inherit m, C, Z, K from Algorithm 1. Give the maximal tolerance value of $\tilde{\kappa}$ for $\tilde{x}$. Set $a_{j,\tilde{j}}$ and b_j such that*

$$a_{j,\tilde{j}} = \delta_{j,\tilde{j}} - U_j^{-1} \sum_{k}^{N} u_{k,j}^m \alpha_k u_{k,\tilde{j}}^m, \quad b_j = \quad U_j^{-1} \sum_{k=1}^{N} u_{k,j}^m (1 - \alpha_k \mu_k) K(x, x_k) \tag{30}$$

and solve $Ay = b$. Calculate $\tilde{d}_j$ such that

$$\tilde{d}_j = K(\tilde{x}, \tilde{x}) - 2y_j + Z_{j,j}. \tag{31}$$

Step 2. *Calculate $\tilde{u}_j$, $\tilde{\mu}_j$, $\tilde{\alpha}$ and $\tilde{d}_j$ such that*

$$\tilde{u}_j = 1 / \sum_{k=1}^{C} \left(\frac{\tilde{d}_j}{\tilde{d}_k} \right)^{\frac{1}{m-1}}, \quad \tilde{\mu} = \sum_{j=1}^{C} \tilde{u}_j^m, \tag{32}$$

$$\tilde{\alpha} = \min \left\{ \tilde{\kappa} \left(\tilde{\mu}^2 K(\tilde{x}, \tilde{x}) - 2\tilde{\mu} \sum_{j=1}^{C} \tilde{u}_j y_j + \sum_{j=1}^{C} \sum_{k=1}^{C} \tilde{u}_j \tilde{u}_k Z_{j,k} \right)^{-1/2}, \tilde{\mu}^{-1} \right\}. \tag{33}$$

$$\begin{aligned} \tilde{d}_j = & (1 - \tilde{\alpha}\tilde{\mu})^2 K(\tilde{x}, \tilde{x}) + 2(1 - \tilde{\alpha}\tilde{\mu}) \sum_{k=1}^{C} (\tilde{\alpha}\tilde{u}_k - \delta_{k,j}) y_k \\ & + \sum_{k=1}^{C} \sum_{\ell=1}^{C} (\tilde{\alpha}\tilde{u}_k - \delta_{k,j})(\tilde{\alpha}\tilde{u}_\ell - \delta_{\ell,j}) Z_{k,\ell}. \end{aligned} \tag{34}$$

Step 3. *Check the stopping criterion. If the criterion is satisfied, $\tilde{u}_j$ is the fuzzy classification function value with respect to $\tilde{x}$. Otherwise, go back to* Step 2.

4 Numerical Examples

In this section, we show some examples of fuzzy classification function by Algorithm 2. In each example, after ten trials for Algorithm 1 with different initial cluster centers are tested and the solution with the minimal objective function value is selected, Algorithm 2 is applied. For all examples, we employ RBF kernel

$$K(x, y) = \exp(-\sigma^2 \|x - y\|_2^2). \tag{35}$$

The first example is classifying the data shown in Fig. 1 into a ring shaped cluster and a ball one. We fix $\sigma^2 = 0.1$ and $m = 2$, and test three different values of

$\kappa_i \in \{0, 0.3, 0.4\}$. The cases of $\kappa_i \in \{0.0.3\}$ produce the correctly classified results shown in Fig. 2 and Fig. 3, respectively. The cases of $\kappa_i = 0.4$ cannot produce the correctly classified results shown in Fig. 4. From these figures, we can find that the larger value of κ_i, the larger membership for the cluster #1 and the larger size of the range for the cluster #1.

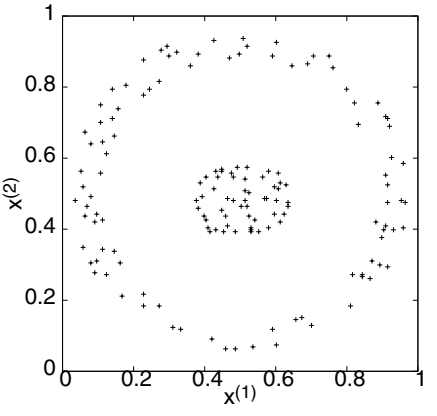


Fig. 1. Ring and Ball Data

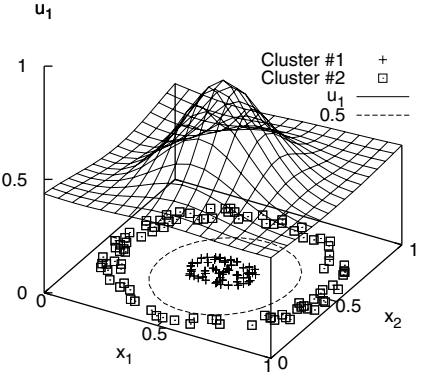


Fig. 2. Successful Classification Result of Fig. 1 by K-sFCM-T 1 with $\sigma^2 = 0.1$, $m = 2$ and $\kappa_i = 0$, and its fuzzy classification function surface by Algorithm 2

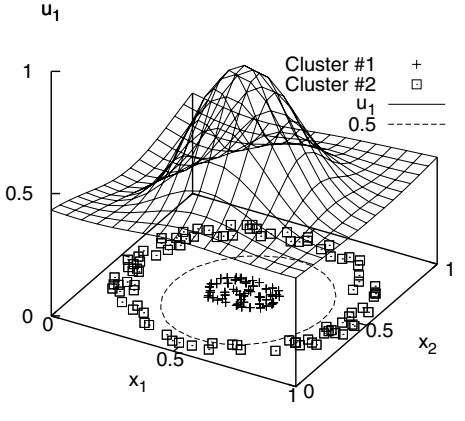


Fig. 3. Successful Classification Result of Fig. 1 by K-sFCM-T 1 with $\sigma^2 = 0.1$, $m = 2$ and $\kappa_i = 0.3$, and its fuzzy classification function surface by Algorithm 2

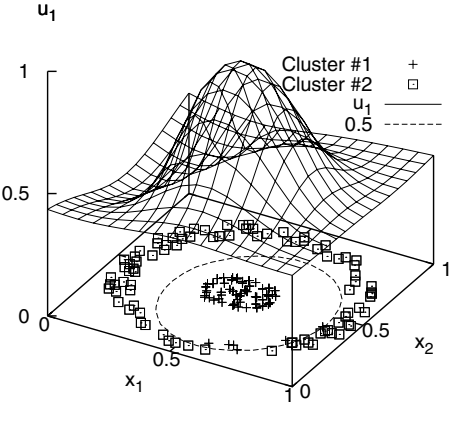


Fig. 4. Miss-Classification Result of Fig. 1 by K-sFCM-T 1 with $\sigma^2 = 0.1$, $m = 2$ and $\kappa_i = 0.4$, and its fuzzy classification function surface by Algorithm 2

The second example is classifying the data shown in Fig. 4 into two crescents shaped clusters. We fix $\sigma^2 = 0.1$ and $m = 2$ and test three different values of $\kappa_i \in \{0, 0.7, 0.8\}$. While the case of $\kappa_i = 0$ fails shown in Fig. 6, the cases of $\kappa_i = 0.7$ produce the correct result shown in Fig. 7. From these figures, we can find that the tolerance helps the incomplete nonlinearity of the introduced kernel and makes the classification border bended adequately. The case of $\kappa_i = 0.8$ cannot produce the correctly classified results since all the cluster centers correspond with each other and all the membership values are 0.5.

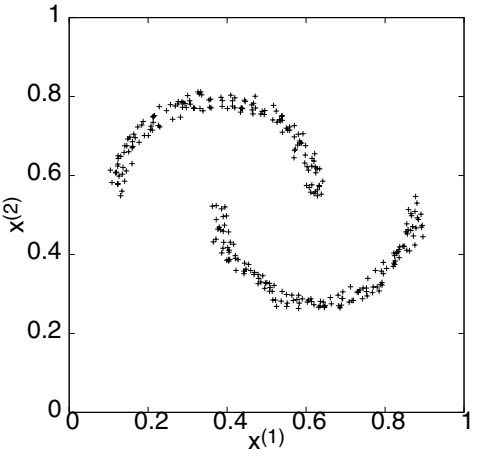


Fig. 5. Crescents Data

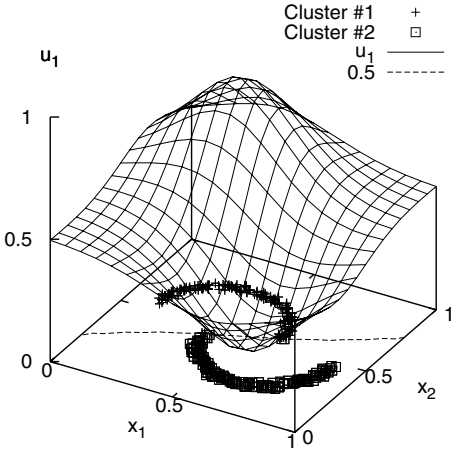


Fig. 6. Miss-classification Result of Fig. 4 by K-sFCM-T 1 with $\sigma^2 = 0.1$, $m = 2$ and $\kappa_i = 0$, and its fuzzy classification function surface by Algorithm 2

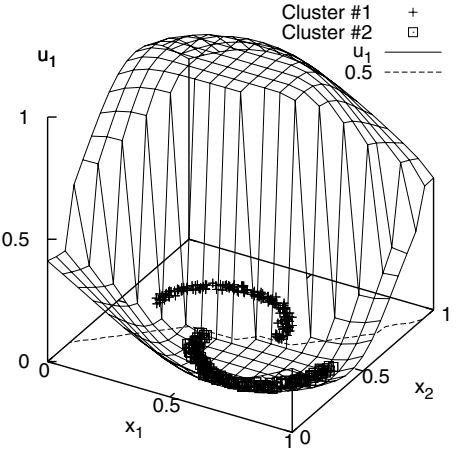


Fig. 7. Successful Classification Result of Fig. 4 by K-sFCM-T 1 with $\sigma^2 = 0.1$, $m = 2$ and $\kappa_i = 0.7$, and its fuzzy classification function surface by Algorithm 2

5 Conclusion

In this paper, we proposed the fuzzy classification function of the standard fuzzy c-means for data with tolerance using kernel functions. First, the standard clustering algorithm for data with tolerance using kernel functions was introduced. Second, the fuzzy classification function for fuzzy c-means without tolerance

using kernel functions was discussed as the solution of a certain optimization problem. Third, the optimization problem was shown so that the solutions are the fuzzy classification function values for the standard fuzzy c-means algorithm using kernel functions with respect to data with tolerance. Fourth, Karush-Kuhn-Tucker conditions of the objective function was considered, and the iterative algorithm was proposed for the optimization problem. Through some numerical examples, it was shown that how prototypical an arbitrary point in the data space is to the already obtained cluster by extending the membership to the whole space.

Note that the fuzzy classification function of K-eFCM-T can be also calculated by similar iterative algorithm, though it is omitted by the sake of pages.

As the future work, using the proposed fuzzy classification function, we will investigate fuzzy c-means algorithm using kernel functions with respect to data with tolerance. Especially, we will study how to give the best parameter κ producing adequate clustering results.

References

1. Bezdek, J.P.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Miyamoto, S., Umayahara, K.: Methods in Hard and Fuzzy Clustering. In: Liu, Z.-Q., Miyamoto, S. (eds.) Soft computing and human-centered machines. Springer, Tokyo (2000)
3. Murata, R., Endo, Y., Haruyama, H., Miyamoto, S.: On fuzzy c–means for data with tolerance. J. Advanced Computational Intelligence and Intelligent Informatics 10(5), 673–681 (2006)
4. Miyamoto, S., Suizu, D.: Fuzzy c-Means Clustering Using Kernel Functions in Support Vector Machines. J. Advanced Computational Intelligence and Intelligent Informatics 7(1), 25–30 (2003)
5. Kanzawa, Y., Endo, Y., Miyamoto, S.: Fuzzy c-Means Algorithms for Data with Tolerance using Kernel Functions (to appear)
6. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

A Similarity Measure for Sequences of Categorical Data Based on the Ordering of Common Elements

Cristina Gómez-Alonso and Aida Valls

iTAKA Research Group - Intelligent Tech. for Advanced Knowledge Acquisition
Department of Computer Science and Mathematics
Universitat Rovira i Virgili
43007 Tarragona, Catalonia, Spain
{cristina.gomez,aida.valls}@urv.cat

Abstract. Similarity measures are usually used to compare items and identify pairs or groups of similar individuals. The similarity measure strongly depends on the type of values to compare. We have faced the problem of considering that the information of the individuals is a sequence of events (i.e. sequences of web pages visited by a certain user or the personal daily schedule). Some measures for numerical sequences exist, but very few methods consider sequences of categorical data. In this paper, we present a new similarity measure for sequences of categorical labels and compare it with the previous approaches.

1 Introduction

In the last years there is an increasing interest in developing techniques to deal with sequences of data. Temporal data mining algorithms have been developed to deal with this type of data [3,6]. Understanding sequence data is becoming very important and the treatment of those sequences is expected to enable novel classes of applications in the next years [1]. For example, telecommunication companies store spatio-temporal data daily, these sequences contain detailed information about the personal or vehicular behaviour, which can allow to find interesting patterns to be used in many different applications, such as traffic control. Similarly, people surf the Internet. This is another great potential source of sequences of users'actions (e.g. web pages visited). The study of the behaviour on the Internet can also lead to interesting applications, such as intrusion detection. There are other domains that also produce temporal sequences [4]: protein sequences that describe the amino acid composition of proteins and represent the structure and function of proteins, gene information (DNA) that encode the genetic makeup, electronic health records that store the clinical history of patients, etc.

However, this type of data requires an adaptation of the classical data mining and decision making techniques applied to static data. Data are called static if all their feature values do not change with time, or change negligibly. In contrast, sequence data analysis is interested in studying the changes in the values in order to identify interesting temporal patterns.

In [8] three different approaches to deal with time series are presented: (1) to work directly with raw data, (2) to convert a raw series data into a feature vector of lower dimension and (3) to represent the sequence with a certain number of model parameters.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 134–145, 2008.

The feature-based and model-based approaches permit to apply conventional algorithms since there is no need to deal with the sequential data. However, sometimes it is not possible to build those feature vectors or models. In this work we are interested in the first approach, which requires to adapt the classical techniques in order to be able to deal with the particularities of sequential data.

In this paper we consider the problem of measuring the similarity of two time sequences of items (i.e. events). Comparing elements is a basic key point in many methods for analysing data, such as clustering techniques (which build clusters of similar objects), classification of objects into existing clusters, characterisation of prototypes, recommender systems or decision making methods (such as those based on dominance rough sets that consider dominance, indiscernibility and similarity relations [5]).

In [8] a survey of similarity/distance measures for sequential data is given. Nine measures are defined and most of them can only be applied to numerical values. In the examples of temporal sequences presented before, the items of the sequence are not numbers but categorical values (places, web pages, proteins, etc.). Although sequences of categorical values are very important nowadays, there still exist few attempts to work with them due to the inherent complexity of dealing with non-numerical values.

In this paper we present a similarity measure between two categorical sequences that is based on the comparison of the common items in the two sequences and the positions where they appear.

First in section 2, a review of other approaches to similarity measurement in time series is introduced. Section 3 presents different features that must be taken into account for working with sequences and then describes the type of sequences that we have considered. In section 4 a new similarity function is defined. Section 5 shows a case study where different similarity measures for sequences are compared. Finally, section 6 gives the conclusions and outlines the future research lines.

2 Review of Dissimilarity Measures

A dissimilarity function d on two objects i and j must satisfy the following conditions:

1. Symmetry: $d(i,j) = d(j,i)$
2. Positivity: $d(i,j) \geq 0$ for all i, j

If conditions:

3. Triangle inequality: $d(i,j) \leq d(i,k) + d(k,j)$ for all i, j, k; and
4. Reflexivity: $d(i,j) = 0$ iff $i = j$

also hold, it is called *metric* or *distance function*.

Moreover, d is a *normalized* distance function if $0 \leq d(i,j) \leq 1$ for all objects i and j.

Dissimilarity functions can be classified according to the type of value they can deal with into: numerical, categorical or mixed functions. In this section, some of the classical distance functions for static data are presented. After this, the existing distance measures for sequential data are reviewed. The cases of numerical and categorical information are presented separately.

2.1 Dissimilarity Functions for Numerical Variables

In this section, we present the most commonly used dissimilarity functions for numerical variables. Let's take two objects i and j represented by the corresponding vectors of values $i = (x_{i1}, \ldots, x_{iK})$ and $j = (x_{j1}, \ldots, x_{jK})$.

Euclidean Distance. It is the sum of the squares of the differences of the values.

$$d_2(i,j) = \sqrt[2]{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2} \tag{1}$$

City-Block or Manhattan Distance. It is the sum of the absolute differences for all the attributes of the two objects.

$$d_1(i,j) = \sum_{k=1}^{K} |x_{ik} - x_{jk}| \tag{2}$$

Minkowski Distance. It is a generic distance which is defined as the $q-th$ root of the sum of powers $q-th$ of absolute differences of the values of the two objects. Note that the *Euclidean distance* and *Manhattan distance* are particular cases for $q = 2$ and $q = 1$, respectively.

$$d_q(i,j) = \left(\sum_{k=1}^{K} |x_{ik} - x_{jk}|^q\right)^{\frac{1}{q}} \tag{3}$$

With respect to sequences of numerical values, the most common similarity measures are the following ones (in [8] these and other approaches are presented):

Short Time Series Distance. It is the sum of the squared differences of the slopes in two time series being compared.

$$d_{STS}(i,j) = \sqrt[2]{\sum_{k=1}^{K}\left(\frac{x_{j(k+1)} - x_{j(k)}}{t_{(k+1)} - t_{(k)}} - \frac{x_{i(k+1)} - x_{i(k)}}{t_{(k+1)} - t_{(k)}}\right)^2} \tag{4}$$

where t_k is the time point for data values x_{ik} and x_{jk}

Dynamic Time Warping Distance. It consists in the alignment of two series $Q = (q_1, q_2, ..., q_n)$ and $R = (r_1, r_2, ..., r_m)$ in order to minimize their difference. To this end, an $n \cdot m$ matrix is built, where the (i,j) element of the matrix contains the distance $d(q_i, r_j)$ (generally *Euclidean distance*). A *warping path* $W = w_1, w_2, ..., w_K$ is calculated, where $max(m,n) \leq K \leq m+n-1$. Then, the minimum distance between the two series is calculated as:

$$d_{DTW}(i,j) = min\left(\frac{\sum_{k=1}^{K} w_k}{K}\right) \tag{5}$$

2.2 Dissimilarity Functions for Categorical Variables

Now, the case of categorical variables for static and sequential data is outlined. For the static case, the following two distances are well-known:

Chi-Squared (χ^2) Distance. It is based on the number of objects in the dataset that have the same value than object i for the k-th variable, I_{ki}.

$$\chi^2(i,j) = \sum_{k=1}^{K} d_k(i,j) \tag{6}$$

where $d_k(i,j)$ is 0 when $x_{ik} = x_{jk}$ and $\left(\frac{1}{I_{ki}} + \frac{1}{I_{kj}}\right)$ otherwise.

Hamming Distance. It is the number of positions where the two objects are different. It is limited to cases when they have identical lengths.

$$d_H(i,j) = \sum_{k=1}^{K} d_k(i,j) \tag{7}$$

where d_k is 0 when $x_{ik} = x_{jk}$ and 1 if $x_{ik} \neq x_{jk}$.

In the case of sequences of categorical values, there are three approaches: Hamming distance (an extension of eq. 7), String metrics and Alignment-based distances.

With respect to String metrics, we have:

Edit or Levenshtein Distance. It calculates the minimum number of edit operations to transform S_1 into S_2, where an edit operation is an insertion, deletion or substitution of a single character.

Damerau-Levenshtein Distance. A modification of Levenshtein distance adding the transposition operation, which is a function that swaps two elements of a sequence.

Kullback-Liebler Divergence. It measures the difference between two probability distributions.

$$d_{KL}(i,j) = \sum_{k=1}^{K} (P_i(x|X) - P_j(x|X)) log\left(\frac{P_i(x|X)}{P_j(x|X)}\right) \tag{8}$$

where P_i denote the *conditional probability distribution* for S_i

Otherwise, *sequence alignment* comes from ADN, RNA or protein sequences studies. The main characteristic of all these cases is that elements of the sequences are characters. In this case, methods are based on the Dynamic Time Warping Distance (see eq. 5). A more detailed analysis can be found at [9,11,12].

3 Description of the Data Sequences to Compare

As it usually happens in many Artificial Intelligence techniques, the nature of the values in the data set determines the characteristics of the method that can be applied. The usual main classification distinguishes: numerical values versus categorical values. Numerical scores can be continuous, discrete or intervals, and can represent quantitative measurements, ratios or ordinal scales. Categorical values represent qualitative features

with ordered or unordered scales [7]. But, in fact, there are other data representations that can be considered as textual, spatial, image or multimedia.

As fas as temporal series are concerned, other distinctions must be done as to whether the data is uniformly or non-uniformly sampled, univariate or multivariate, and whether the sequences are of equal or unequal length [8].

In this work we want to deal with sequences of events that represent the behaviour of the user in a particular context. For example tourists visiting a city, where the sequences show the itinerary that each person has followed to visit the interesting locations in this city. A private real data set of tourists' itineraries provided by Dr. Shoval has been tested. This data set is about a city with 25 interesting places and has about 40 itineraries with lengths that range from 10 to 30 items.

Another data set we have considered is the list of sequences of visits at the Microsoft web page. The data was obtained by sampling and processing the www.microsoft.com logs. The data set records the use of 38000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site (Vroots) that he/she visited in a one week time-frame. The number of Vroots is 294, and the mean number of Vroots visits per user is 3. This data is publicly available at the UCI Machine Learning Repository [2].

These two examples of event sequences have the following common characteristics:

- Events are categorical values that belong to a finite set of linguistic labels (city locations, web pages).
- The sequences have been uniformly sampled in the sense that time slopes are not taken into account.
- The sequences are univariate, only one concept is studied.
- The lengths of the sequences of the individuals are not equal.
- Events can be repeated in the sequence (for example, a certain tourist visited the same place, Main Street, more than one time during his holidays).

To facilitate the analysis of the results, the categorical values indicating places or web pages have been substituted by simpler identifiers. An example of 14 different event sequences is given in Table 1. Each character may represent a physical place or a web page.

Table 1. Example of data sequences

Id	Sequence
1	a b
2	b c
3	a b c
4	c a b
5	d a b c
6	e d a b c
7	f e d a b c

Id	Sequence
8	c g d a b c
9	d
10	d a
11	d b
12	c b c b c b
13	b c b c b c
14	e b e b e b

4 A New Similarity Measure: Ordered-Based Sequence Similarity

The similarity measures for sequences of categorical values presented in section 2 are quite simple and are not adequate to deal with temporal event sequences like the web logs or the tourists' itineraries. To compare this type of sequences two issues are important:

1. The number common elements in the two sequences
2. The order between the common elements

The former allows us to measure if the two individuals have done the same things, that is, if they have visited the same web pages or have gone to the same places. The latter takes into account the temporal sequence of the events, that is, if two tourists have visited the Main Street after going to the City Hall or not. This second measure should also take into account if two events have taken place consecutively or not.

For example, let $T1$ and $T2$ be tourists who have visited some places of the same city: $T1 = \{a,b,c\}$ and $T2 = \{c,a,b,d\}$. Notice that, there are 3 common places and also they have visited a before b. So, we could say that they are quite similar.

In this paper we present a new approach to calculate the similarity that takes into account these two aspects. It is called *Ordering-based Sequence Similarity* (OSS) and consists, on one hand, in finding the common elements in the two sequences, and on the other hand, in comparing the positions of the elements in both sequences. The *'elements'* that are the basis of this measure can be either single events or sub-sequences of events that are considered as an indivisible groups (i.e patterns). In case of working with patterns, they must have a minimum length of two and a maximum length equal to the shortest sequence.

Definition 1. *Let i and j be two sequences of items of different lengths, $i = (x_{i,1},\ldots, x_{i,card(i)})$ and $j = (x_{j,1},\ldots,x_{j,card(j)})$. Let $L = \{l_1,...,l_n\}$ be a set of n symbols to represent all the possible elements of those sequences (L is called a language). Then, the Ordering-based Sequence Similarity (OSS) is defined as:*

$$d_{OSS}(i,j) = \frac{f(i,j)+g(i,j)}{card(i)+card(j)} \tag{9}$$

where

$$g(i,j) = card(\{x_{ik}|x_{ik} \notin j\}) + card(\{x_{jk}|x_{jk} \notin i\}) \tag{10}$$

and

$$f(i,j) = \frac{\sum_{k=1}^{n}(\sum_{p=1}^{\Delta}|i_{(l_k)}(p) - j_{(l_k)}(p)|)}{max\{card(i),card(j)\}} \tag{11}$$

where $i_{(l_k)} = \{t|i(t) = l_k\}$ and $\Delta = min(card(i_{(l_k)}),card(j_{(l_k)}))$.

This function has two parts, g is counting the number of non common elements, and f measures the similarity in the position of the elements in the sequences (the ordering). The function f is calculated in the symbols space L. So, first, each event in the sequence i is projected into L, obtaining a numerical vector fore each symbol: $i_{(l_1)}..i_{(l_n)}$. Each of these new vectors store the positions of the corresponding symbol in the sequence i.

The same is done with sequence j, obtaining $j_{(l_1)}..j_{(l_n)}$. Then the projections of the two sequences i and j into L are compared, and the difference in the positions is calculated and normalised by the maximum cardinality of the sequences i and j.

If two sequences are equal, the result of d_{OSS} is zero, because the positions are always equal ($f = 0$) and there are not uncommon elements ($g = 0$). Oppositely, if the two sequences do not share any element, then $g = card(i) + card(j)$ and $f = 0$, and d_{OSS} is equal to 1 when it is divided by $card(i) + card(j)$. The Ordering-based Sequence Similarity function always gives values between 0 and 1.

The function has the following properties:

- Symmetry: $d_{OSS}(i,j) = d_{OSS}(j,i)$
- Positivity: $d_{OSS}(i,j) \geq 0$ for all i, j
- Reflexivity: $d_{OSS}(i,j) = 0$ iff $i = j$

However, it does not fulfil the triangle inequality: $d_{OSS}(i,j) \leq d_{OSS}(i,k) + d_{OSS}(k,j)$ for all i, j, k. From these properties, it is clear that d_{OSS} is a dissimilarity but not a distance.

Proof. The proof of Symmetry, Positivity and Reflexivity is trivial by Definition 1. The Triangle Inequality does not hold. This is proven in the following counterexample.

Let A, B and C be three sequences defined by $A = \{b,c\}$, $B = \{d,a\}$ and $C = \{d,a,b,c\}$. In this case, $d_{OSS}(A,B) = 1.0$ because they do not share any item. However, $d_{OSS}(A,C) = 0.5$ because they have two elements in common. B and C also have other two common elements (and they are also in the same position), so $d_{OSS}(B,C) = 0.33$. Consequently, $d_{OSS}(A,C) + d_{OSS}(B,C) = 0.83$ which is less than $d_{OSS}(A,B)$ that is 1.0, which proofs that the triangle inequality is not fulfilled. □

As it has been pointed out, this measure can be applied to different items: single events or groups of events. Having a sequence $\{a,b,a,c,d\}$, in the first case, $i = (a,b,a,c,d)$, so x_{ij} is any individual event in the sequence. In the second case, $i = (ab,ba,ac,cd)$, so x_{ij} is any pair of consecutive items, and $i = (aba,bac,acd)$, for triplets.

The following example illustrate how d_{OSS} is calculated for single events (d_{OSS-1}) and for pairs of events (d_{OSS-2}). Let us take the two following sequences: $A = \{a,b,c,a\}$, $B = \{c,a,d,b,c,a,c\}$, with $card_A = 4$ and $card_B = 7$.

The similarity considering single items gives $d_{OSS-1}(A,B) = 0.36$. This result is obtained in the following way: symbols a, b and c are common in both sequences A and B. The projection on the symbol a are: $A_{(a)} = \{0,3\}$ and $B_{(a)} = \{1,5\}$, so $f_a(A,B) = |0-1| + |3-5| = 3$. For symbol b: $A_{(b)} = \{1\}$, $B_{(b)} = \{3\}$ and $f_b(A,B) = |1-3| = 2$. For c: $A_{(c)} = \{2\}$, $B_{(c)} = \{0,4,6\}$ and $f_c(A,B) = |2-0| = 2$. So, $f(A,B) = \frac{f_a(A,B)+f_b(A,B)+f_c(A,B)}{7} = 1$. Calculating the non common elements, we have $g(A,B) = 3$. Finally, $d_{OSS-1}(A,B) = \frac{f(A,B)+g(A,B)}{4+7} = \frac{1+3}{11} = 0.36$.

Considering the same case example with patterns of length 2, we have a greater dissimilarity, $d_{OSS-2}(A,B) = 0.629$. In this case, the sequences are $A' = \{ab,bc,ca\}$ and $B' = \{ca,ad,db,bc,ca,ac\}$ with cardinalities 3 and 6. They share 2 elements, for the bc pair we have $A_{(bc)} = \{1\}$, $B_{(bc)} = \{3\}$ and $f_{bc}(A,B) = |1-3| = 2$, while for the pair ca: $A_{(ca)} = \{2\}$, $B_{(ca)} = \{0,4\}$ and $f_c a(A,B) = |2-0| = 2$. So, $f(A,B) = \frac{f_{bc}(A,B)+f_{ca}(A,B)}{6} = 0.66$. Calculating the non common elements, we have $g(A,B) = 5$. Finally, $d_{OSS-2}(A,B) = \frac{f(A,B)+g(A,B)}{3+6} = \frac{0.66+5}{9} = 0.629$.

5 Experiments

In this section we present the results obtained with a test set of 14 registers, corresponding to the event sequences presented in Table 1. This dataset contains sequences of different lengths. Register 9 is an extreme case, with a single event in the sequence. Sequences in registers 12, 13 and 14 show the case of having repeated events.

We have tested the Ordering-based Sequence Similarity considering each event separately, OSS-1 (Table 2) and with patterns of two events, OSS-2 (Table 3).

Notice that the OSS applied to pairs of events gives higher dissimilarity values in many cases (see the number of 1's in Table 3). This is due to the fact that finding common pairs of events is much more difficult than finding common single events.

For OSS-2, the sequences in the register id=6 $\{e,d,a,b,c\}$ and id=7 $\{f,e,d,a,b,c\}$ are the most similar ones (0.19), because they share 4 common pairs in very similar positions (ed,da,ab,bc) and 1 uncommon pair (fe). The next ones are id=5 $\{d,a,b,c\}$ and id=6 that have 3 common pairs and 1 uncommon pair. And in third place we find sequences id=12 $\{c,b,c,b,c,b\}$ and id=13 $\{b,c,b,c,b,c\}$, that also share 4 common pairs and 2 uncommon ones.

On the contrary, OSS-1 considers that similarity between registers id=12 and id=13 is higher (0.08) that the one between id=6 and id=7 (0.19). This is because in the first case there are 6 common symbols (all) and in the second case there are only 5 common symbols and 1 uncommon. From our point of view, OSS-1 is able to better capture the degrees of similarity for the sequences of events than the OSS-2, since it is not so important that they happen together than in similar positions.

After analysing the OSS results, the behaviour of the OSS function has been compared to the Edit Distance, which is a measure that is quite popular for comparing sequences [4]. The Edit Distance counts the number of changes needed to be applied on

Table 2. Results of the OSS applied to individual events (OSS-1)

Id.Reg.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0.62	0.2	0.33	0.41	0.54	0.62	0.62	1	0.62	0.5	0.75	0.77	0.75
2	0.62	0	0.33	0.4	0.5	0.59	0.66	0.60	1	1	0.62	0.54	0.5	0.77
3	0.2	0.33	0	0.22	0.25	0.4	0.5	0.48	1	0.66	0.6	0.59	0.59	0.77
4	0.33	0.4	0.22	0	0.25	0.4	0.5	0.40	1	0.6	0.66	0.57	0.61	0.79
5	0.41	0.5	0.25	0.25	0	0.19	0.33	0.35	0.6	0.33	0.37	0.66	0.66	0.81
6	0.54	0.59	0.4	0.4	0.19	0	0.16	0.37	0.7	0.48	0.51	0.72	0.72	0.66
7	0.62	0.66	0.5	0.5	0.33	0.16	0	0.40	0.76	0.58	0.60	0.77	0.77	0.72
8	0.62	0.60	0.48	0.40	0.35	0.37	0.40	0	0.76	0.58	0.60	0.58	0.59	0.87
9	1	1	1	1	0.6	0.7	0.76	0.76	0	0.33	0.33	1	1	1
10	0.62	1	0.66	0.6	0.3	0.48	0.58	0.58	0.33	0	0.5	1	1	1
11	0.5	0.62	0.6	0.66	0.37	0.51	0.60	0.60	0.33	0.5	0	0.75	0.77	0.75
12	0.75	0.54	0.59	0.57	0.66	0.72	0.77	0.58	1	1	0.75	0	**0.08**	0.5
13	0.77	0.5	0.59	0.61	0.66	0.72	0.77	0.59	1	1	0.77	**0.08**	0	0.54
14	0.75	0.77	0.77	0.79	0.81	0.66	0.72	0.87	1	1	0.75	0.5	0.54	0

Table 3. Results of the OSS applied to pairs of events (OSS-2)

Id.Reg.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	0.33	0.5	0.58	0.7	0.76	0.76	-	1	1	1	1	1
2	1	0	0.5	1	0.66	0.75	0.8	0.8	-	1	1	0.7	0.66	1
3	0.33	0.5	0	0.62	0.33	0.5	0.59	0.59	-	1	1	0.71	0.74	1
4	0.5	1	0.62	0	0.6	0.70	0.77	0.77	-	1	1	1	1	1
5	0.58	0.66	0.33	0.6	0	0.25	0.4	0.4	-	0.5	1	0.77	0.8	1
6	0.7	0.75	0.5	0.70	0.25	0	**0.19**	0.39	-	0.65	1	0.82	0.84	1
7	0.76	0.8	0.59	0.77	0.4	**0.19**	0	0.4	-	0.73	1	0.86	0.88	1
8	0.76	0.8	0.59	0.77	0.4	0.39	0.4	0	-	0.73	1	0.86	0.88	1
9	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	1	1	1	1	0.5	0.65	0.73	0.73	-	0	1	1	1	1
11	1	1	1	1	1	1	1	1	-	1	0	1	1	1
12	1	0.7	0.71	1	0.77	0.82	0.86	0.86	-	1	1	0	0.28	1
13	1	0.66	0.74	1	0.8	0.84	0.88	0.88	-	1	1	0.28	0	1
14	1	1	1	1	1	1	1	1	-	1	1	1	1	0

Table 4. Results according to the Edit Distance (ED)

Id.Reg.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	0.33	0.33	0.5	0.6	0.66	0.66	1	1	0.5	0.83	0.83	0.83
2	1	0	0.33	1	0.5	0.6	0.66	0.66	1	1	1	0.66	0.66	0.83
3	0.33	0.33	0	0.66	0.25	0.4	0.5	0.5	1	1	0.66	0.66	0.66	0.83
4	0.33	1	0.66	0	0.5	0.6	0.66	0.5	1	0.66	0.66	0.66	0.66	0.83
5	0.5	0.5	0.25	0.5	0	0.2	0.33	0.33	0.75	0.5	0.5	0.66	0.66	0.83
6	0.6	0.6	0.4	0.6	0.2	0	**0.16**	0.33	0.8	0.6	0.6	0.66	0.66	0.66
7	0.66	0.66	0.5	0.66	0.33	**0.16**	0	0.33	0.83	0.66	0.66	0.83	0.66	0.83
8	0.66	0.66	0.5	0.5	0.33	0.33	0.33	0	0.83	0.66	0.66	0.66	0.66	1
9	1	1	1	1	0.75	0.8	0.83	0.83	0	0.5	0.5	1	1	1
10	1	1	1	0.66	0.5	0.6	0.66	0.66	0.5	0	0.5	1	1	1
11	0.5	1	0.66	0.66	0.5	0.6	0.66	0.66	0.5	0.5	0	0.83	0.83	0.83
12	0.83	0.66	0.66	0.66	0.66	0.66	0.83	0.66	1	1	0.83	0	0.33	0.5
13	0.83	0.66	0.66	0.66	0.66	0.66	0.66	0.66	1	1	0.83	0.33	0	0.66
14	0.83	0.83	0.83	0.83	0.83	0.66	0.83	1	1	1	0.83	0.5	0.66	0

one sequence to obtain another one. To scale the values into the unit interval, we have divided the number of changes by the length of the longest sequence. Table 4 presents the results of the normalised Edit Distance (ED) on the same data set.

The first significant difference in the results of Table 4 is the pair of registers that achieves the minimum dissimilarity in each measure. ED considers that the most similar ones are $\{e,d,a,b,c\}$ (id=6) and $\{f,e,d,a,b,c\}$ (id=7), because they are the longest sequences with a single difference, the introduction of a new symbol. OSS-1 gives the same similarity value to this pair, 0.16, but it finds another most similar pair of

Table 5. Comparison of the results

Register		OSS-1			OSS-2			ED		
		MIN		MAX	MIN		MAX	MIN		MAX
Id	Seq	Id	Value	Id	Id	Value	Id	Id	Value	Id
1	ab	3	0.2	9	3	0.33	2,10,11,12, 13,14	3,4	0.33	2,9,10
2	bc	3	0.33	9,10	3	0.5	1,4,10,11, 14	3	0.33	1,4,9,10,11
3	abc	1	0.2	9	1,5	0.33	10,11,14	5	0.25	9,10
4	cba	3	0.22	9	1	0.5	2,10,11,12, 13,14	1	0.33	2,9
5	dabc	6	0.19	14	6	0.25	11,14	6	0.2	14
6	edabc	7	0.16	12,13	7	0.19	11,14	7	0.16	9
7	fedabc	6	0.16	12,13	6	0.19	11,14	6	0.16	9,12,14
8	cgdabc	5	0.35	14	6	0.39	11,14	5,6,7	0.33	14
9	d	10,11	0.33	1,2,3,4,12, 13,14	-	-	-	10,11	0.5	1,2,3,4,12, 13,14
10	da	5,9	0.33	2,12,13,14	5	0.5	1,2,3,4,11, 12,13,14	5,9,11	0.5	1,2,3,12, 13,14
11	db	9	0.33	13	all	1.0	all	1,5,9,10	0.5	2
12	cbcbcb	13	0.08	9,10	13	0.28	1,4,10,11, 14	13	0.33	9,10
13	bcbcbc	12	0.08	9,10	12	0.28	1,4,10,11, 14	12	0.33	9,10
14	ebebeb	12	0.5	9,10	all	1.0	all	12	0.5	8,9,10

sequences: $\{c,b,c,b,c,b\}$ (id=12) and $\{b,c,b,c,b,c\}$ (id=13), which are the longest sequences that share exactly the same symbols in very similar positions. In fact, they have the same sequence $\{c,b,c,b,c\}$ but adding b before or after it.

If we consider now the first row of the matrices, the one that compares the register $\{a,b\}$ (id=1) with the rest, it can be seen that the behaviour of the Edit Distance is quite different from the one of the OSS-1. ED considers that sequence $\{a,b\}$ (id=1) is equally similar to $\{a,b,c\}$ (id=3) and $\{c,a,b\}$ (id=4). Whereas, OSS-1 considers that the former is more similar to register id=1 because both individuals have started the sequence doing the same event a, followed by b. And the difference is that the individual id=1 has stopped and the other has continued one step more. However, sequence id=4 has not started doing the same event. This difference is able to be captured if the relative ordering of the events is considered.

This difficulty to distinguish two sequences that have different items than two sequences that have the different items but in different order is the main drawback of the Edit Distance [13]. An extreme case is the result given with sequences like $\{a,a,b,b\}$ and $\{b,b,a,a\}$. ED will give a dissimilarity of 4 changes (or 1 if it is normalised). In

this case OSS-1 clearly improves ED giving a dissimilarity value of 0.125, which is more like what common sense would indicate.

To make a deeper analysis of the behaviour of the similarity matrix for clustering purposes, it is needed to identify which is the closest sequence to a given one. Table 5 shows the identifier of the register/s with minimum dissimilarity to each of the 14 case studies, for the 3 measures. Also the corresponding dissimilarity value that links those pairs of sequences is given. Finally, the sequence with maximum dissimilarity is shown.

An analysis of the table suggests that the dissimilarity function OOS-1 is more precise than ED and OOS-2 to determine the minimum and maximum values. It usually identifies unique values.

6 Conclusions and Future Work

In this paper we have proposed a new measure of dissimilarity for sequences of categorical values that considers two main criteria: which are the common and not common symbols, and which is the difference in the ordering of the common symbols in both sequences. The rationale for establishing these criteria is that the sequences to be compared contain an ordered list of events (f.i. an itinerary that indicates the places visited by a tourist), and we are interested in capturing this sequentiality of the items.

The paper shows that, for event sequences, the Ordering-based Sequence Similarity (OOS) gives better results than the Edit Distance. The results show that OSS is a proper approach to prioritize both the number of common elements as its order. It can also be easily seen that OSS also behaves better than the Hamming distance, which compares the sequences position by position and charge each mismatch 1 unit in the dissimilarity value without taking into account if the symbol appears in other nearer position. One of the future analysis to be done is a comparison with the alignment-based approaches.

We are now interested in using the Ordering-based Sequence Similarity for clustering. Building clusters is interesting in many problems, as it has been mentioned in the introduction. It can be used to learn the underlying structure of a domain. In this sense, clustering of event sequences can lead to identify groups of individuals that behave in a similar way [3]. Other use of clustering methods in which we are particularly interested in is the field of privacy preserving. Microaggregation is one of the standard tools for numerical database protection commonly in use in National Statistical Offices. In the last years, research on the protection of numerical time series has started [10], due to the increasing number of sequence data available, as argued in the introduction of this paper. The OSS similarity is a first step towards the extension of microaggregation to the case of categorical event sequences.

Acknowledgments

The authors want to specially thank the collaboration of Dr. N. Shoval. This work has been supported by the Spanish research projects E-AEGIS (TSI-2007-65406-C03) and Consolider-Ingenio 2010 ARES (CSD2007-00004).

References

1. Abul, O., Atzori, M., Bonchi, F., Giannotti, F.: Hiding sequences. In: ICDE Workshops, pp. 147–156. IEEE Computer Society, Los Alamitos (2007)
2. Asuncion, A., Newman, D.: UCI machine learning repository (2007), http://archive.ics.uci.edu/ml/
3. Dietterich, T.G.: Machine learning for sequential data: A review. In: Caelli, T., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 15–30. Springer, Heidelberg (2002)
4. Dong, G., Pei, J.: Sequence Data Mining. Advances in Database Systems, vol. 33. Springer, US (2007)
5. Figueira, J., Greco, S., Ehrgott, M.: Multiple Criteria Decision Analysis: State of the Art Surveys. ISOR & MS, vol. 78. Springer, Heidelberg (2005)
6. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco (2006)
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)
8. Liao, T.W.: Clustering of time series data–a survey. Pattern Recognition 38(11), 1857–1874 (2005)
9. Mount, D.W.: Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press (September 2004)
10. Nin, J., Torra, V.: Extending microaggregation procedures for time series protection. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Slowinski, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 899–908. Springer, Heidelberg (2006)
11. Notredame, C.: Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology 3(8), e123+ (2007)
12. Wallace, I.M., Blackshields, G., Higgins, D.G.: Multiple sequence alignments. Current Opinion in Structural Biology 15(3), 261–266 (2005)
13. Yang, J., Wang, W.: Cluseq: Efficient and effective sequence clustering. In: 19th International Conference on Data Engineering (ICDE 2003), vol. 00, p. 101 (2003)

Analytical and Numerical Evaluation of the Suppressed Fuzzy C-Means Algorithm

László Szilágyi[1,2], Sándor M. Szilágyi[1], and Zoltán Benyó[2]

[1] Sapientia - Hungarian Science University of Transylvania,
Faculty of Technical and Human Science, Târgu-Mureş, Romania
lalo@ms.sapientia.ro
[2] Budapest University of Technology and Economics, Department of Control
Engineering and Information Technology, Budapest, Hungary

Abstract. Suppressed fuzzy c-means (s-FCM) clustering was introduced in [Fan, J. L., Zhen, W. Z., Xie, W. X.: Suppressed fuzzy c-means clustering algorithm. Patt. Recogn. Lett. 24, 1607–1612 (2003)] with the intention of combining the higher speed of hard c-means (HCM) clustering with the better classification properties of fuzzy c-means (FCM) algorithm. They modified the FCM iteration to create a competition among clusters: lower degrees of memberships were diminished according to a previously set suppression rate, while the largest fuzzy membership grew by swallowing all the suppressed parts of the small ones. Suppressing the FCM algorithm was found successful in the terms of accuracy and working time, but the authors failed to answer a series of important questions. In this paper we clarify the view upon the optimality and the competitive behavior of s-FCM via analytical computations and numerical analysis.

Keywords: fuzzy c-means algorithm, suppressed fuzzy c-means algorithm, competitive clustering, alternating optimization.

1 Introduction

Fuzzy logic [19] has successfully penetrated the theory of data clustering. It took several important steps until Bezdek [5] reached the alternative optimization (AO) solution of fuzzy clustering, named fuzzy c-means algorithm (FCM), which improved the partition performance of the previously existing hard c-means clustering (HCM) by extending the membership logic.

FCM outperformed HCM in the terms of partition quality, at the cost of a slower convergence. In spite of this drawback, FCM is one of the most popular clustering algorithms not only in engineering studies, but also in a series of sciences from biology to sociology.

Several researchers have studied the convergence speed of FCM and tried to introduce modified algorithms with improved characteristics [6,12]. Another trend was to combine competitive and FCM clustering techniques. In this order, several new algorithms appeared which tried to fuzzify Kohonen's learning

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 146–157, 2008.

vector quantization (LVQ) algorithm [14,18]. Karayiannis and Bezdek [13] introduced an integrated approach to fuzzy LVQ and FCM, based on the notion of generalized mean of real numbers. These algorithms, even if their main goal wasn't to improve the convergence of accurate clustering techniques, they had a considerable contribution.

Later, Wei and Xie [16] proposed a technique called rival checked FCM to modify the fuzzy membership values given by FCM, in order to improve the performance against the clock. Their solution, which rose the highest degree of membership at the detriment of the second largest, led to unacceptable results in some cases. In order to confront this problem, Fan et al. [7] introduced the suppressed fuzzy c-means algorithm (s-FCM). The authors stated that, by prizing the highest membership and suppressing all others, the modification does not disturb the original order among clusters. They also remarked, that setting the suppression rate $\alpha = 0$ makes s-FCM and HCM identical, while $\alpha = 1$ reduces the algorithm to the conventional FCM. The s-FCM algorithm was found successful based on some numerical analysis, but unfortunately, the authors left several issues wide open:

1. They failed to prove whether s-FCM is optimal in any sense, that is, whether it minimizes any kind of objective function.
2. The extra step of s-FCM was inspired by the basis of competitive learning [7], but the authors failed to give any evidence of its competitive behavior.
3. The authors found s-FCM clustering insensitive to the fuzzyfication parameter m, based on a few experiments. However, since the competitivity of FCM is controlled using m [5,8], this cannot be the established so easily.
4. The authors failed to provide any strategy to choose the suppression rate α. This was already pointed out by Hung et al. [10,11], who formulated a criterion for α based on considerations regarding cluster validity.

Very recently, Xie et al. introduced a novel possibilistic c-means clustering [17] algorithm that interprets the fuzzy membership gap produced by s-FCM between the winner and non-winner clusters similarly to the symmetrical margin between classes provided by support vector machines [15,4] in supervised classification problems.

This paper assumes to investigate the main issues of the suppressed FCM algorithm listed above. The rest of this paper is structured as follows. Chapter 2 presents the background works that have impact on our investigations. Chapter 3 contains the analytical computations performed in order to reveal the properties of s-FCM. Chapter 4 gives a numerical analysis of the properties of s-FCM. Conclusions are given in the last chapter.

2 Background

2.1 Fuzzy C-Means

The fuzzy c-means algorithm has successful applications in a wide variety of clustering problems. The traditional FCM partitions a set of object data into a

number of c clusters based on the minimization of a quadratic objective function. The objective function to be minimized is:

$$J_{\mathrm{FCM}} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} ||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d_{ik}^2 \ , \tag{1}$$

where $\boldsymbol{x}_k$ represents the input data ($k = 1 \ldots n$), $\boldsymbol{v}_i$ represents the prototype or centroid value or representative element of cluster i ($i = 1 \ldots c$), $u_{ik} \in [0, 1]$ is the fuzzy membership function showing the degree to which datum vector $\boldsymbol{x}_k$ belongs to cluster i, $m > 1$ is the fuzzyfication parameter, and d_{ik} represents the distance between vector $\boldsymbol{x}_k$ and cluster prototype $\boldsymbol{v}_i$. According to the definition of fuzzy sets, for any input vector $\boldsymbol{x}_k$, we have

$$\sum_{i=1}^{c} u_{ik} = 1 \ . \tag{2}$$

The minimization of the objective function is reached by alternately applying the optimization of J_{FCM} over $\{u_{ik}\}$ with $\boldsymbol{v}_i$ fixed, $i = 1 \ldots c$, and the optimization of J_{FCM} over $\{\boldsymbol{v}_i\}$ with u_{ik} fixed, $i = 1 \ldots c$, $k = 1 \ldots n$ [9]. During each cycle, the optimal values are computed from the zero gradient conditions, and obtained as follows:

$$u_{ik}^{\star} = \frac{d_{ik}^{-2/(m-1)}}{\sum_{j=1}^{c} d_{jk}^{-2/(m-1)}} \quad \forall\, i = 1 \ldots c, \forall\, k = 1 \ldots n \ , \tag{3}$$

$$\boldsymbol{v}_i^{\star} = \frac{\sum_{k=1}^{n} u_{ik}^{m} \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad \forall\, i = 1 \ldots c \ . \tag{4}$$

According to the alternative optimization (AO) scheme, formulae (3) and (4) are alternately applied, until cluster prototypes stabilize.

2.2 Hard C-Means Clustering

The main difference between hard c-means and fuzzy c-means clustering is the membership logic. While in case of fuzzy clustering, the degree of membership of a vector $\boldsymbol{x}_k$ to cluster i, denoted by u_{ik}, can take any value between 0 and 1, hard clustering uses as degrees of membership only two values: 0 and 1. Although the objective function of HCM is the same as the one of FCM, using $m = 1$, the optimization formula differs from Eq. (3): for any $k = 1 \ldots n$, u_{ik} is set to 1 whenever $\boldsymbol{v}_i$ is the closest cluster prototype viewed from $\boldsymbol{x}_k$, and 0 in any different case. Ties are resolved arbitrarily. The computation of cluster prototypes is performed according to Eq. (4).

The main advantage of FCM over HCM is that it improves partition performance and reveals the classification data more reasonably [7]. However, FCM also has the well-known disadvantage of slower convergence [5].

2.3 Suppressed Fuzzy C-Means

The suppressed fuzzy c-means algorithm was introduced by Fan et al. [7], having the declared goal of improving the convergence speed of FCM, while keeping its good classification accuracy. They modified the alternative optimization scheme of FCM, by inserting an extra computational step between the application of formulae (3) and (4). This new step performs the following task: for each vector $\boldsymbol{x}_k$, after having obtained its new optimal fuzzy membership values u_{ik}, we search for the highest one $u_{w_k k}$, and declare cluster $w_k \in \{1, 2, \ldots, c\}$ the winner. The fuzzy memberships are then modified such a way, that all non-winner values are decreased via multiplying by a so-called suppression rate α, $(0 \leq \alpha \leq 1)$, and the winner membership is increased such a way, that the relation (2) is fulfilled by the modified memberships. Therefore, the extra formula of s-FCM is:

$$\mu_{w_k k} = 1 - \alpha \sum_{j \neq w_k} u_{jk} = 1 - \alpha + \alpha u_{w_k k} \quad \text{if } i = w_k \ , \tag{5}$$

$$\mu_{ik} = \alpha u_{ik} \qquad \forall i \in \{1, 2, \ldots, c\} - \{w_k\} \ , \tag{6}$$

where μ_{ik}, $i = 1 \ldots c$, $k = 1 \ldots n$, represent the fuzzy memberships obtained with the modification introduced by the s-FCM algorithm.

Fan et al. did not give a recipe for choosing a suppression rate that is optimal in any sense, or suitable for any any given purpose. They set the suppression rate to the middle of the interval ($\alpha = 0.5$), and found s-FCM insensitive to the fuzzification parameter m.

3 What Is the Suppressed Fuzzy C-Means Algorithm?

In the following subsections, we will try to give some answers to the questions previously formulated on the s-FCM algorithm. During the computations we will suppose that $0 < \alpha < 1$, to avoid divisions by zero. These cases are trivial anyway and need no investigation.

3.1 What Kind of Competition Does Suppression Introduce?

Fan et al. introduced the suppressed FCM algorithm on the basis of competitive learning [7]. This is all they said about its competitive behavior. Obviously, when the new fuzzy memberships of a vector $\boldsymbol{x}_k$ is computed in a given iteration, there is a competition among clusters, which makes the closest prototype win and all others lose.

We will start the investigation at Eq. (3). When the new degrees of membership of vector $\boldsymbol{x}_k$ are computed, everything depends on the distances d_{ik}, $i = 1 \ldots c$. If we change the scale of the metric such a way, that distances are lengthened or shortened proportionally, the obtained memberships remain the same. In other words, the ratio of two membership values, say u_{ik}/u_{jk} does not depend on the above mentioned scale.

Conversely, when memberships to non-winner clusters are proportionally suppressed via multiplying them by α, it can be interpreted as their distances from vector $\boldsymbol{x}_k$ remain constant, but as the winner cluster receives a higher degree of membership, prototype $\boldsymbol{v}_{w_k}$ is counted as it were closer to $\boldsymbol{x}_k$ than it really is. Again, in other words, when new cluster prototypes $\boldsymbol{v}_i$ are computed using the weighted averaging formula in Eq. (4), based on the suppressed fuzzy memberships, the vectors $\boldsymbol{x}_k$ whose competition the currently computed prototype has won, are taken into consideration as they were at a reduced distance $d'_{w_k k} < d_{w_k k}$, giving those vectors a higher impact than in FCM.

This reduced distance can be characterized with a learning rate $\eta = 1 - (d'_{w_k k}/d_{w_k k})$, which will be computed in the followings. We will use the notations: $\delta_{ik} = d_{ik}^{2/(1-m)}\ \forall i = 1 \dots c$, $\forall k = 1 \dots n$, and $\delta'_{w_k k} = \gamma \delta_{w_k k}$, where we expect $\gamma \geq 1$.

Using these new notations, we can rewrite (3) for both winner and non-winner clusters. For the winner cluster we have:

$$\mu_{w_k k} = \frac{\gamma \delta_{w_k k}}{\gamma \delta_{w_k k} + \sum\limits_{j=1, j \neq w_k}^{c} \delta_{jk}} = \frac{\gamma \delta_{w_k k}}{(\gamma - 1)\delta_{w_k k} + \sum\limits_{j=1}^{c} \delta_{jk}} \ , \tag{7}$$

while the non-winner clusters receive the memberships:

$$\mu_{ik} = \frac{\delta_{ik}}{\gamma \delta_{w_k k} + \sum\limits_{j=1, j \neq w_k}^{c} \delta_{jk}} = \frac{\delta_{ik}}{(\gamma - 1)\delta_{w_k k} + \sum\limits_{j=1}^{c} \delta_{jk}} \ . \tag{8}$$

Now we can compare the suppressed memberships computed in (5) with (7), and (6) with (8). From the former two we get:

$$(1 - \alpha) + \alpha \frac{\delta_{w_k k}}{\sum\limits_{j=1}^{c} \delta_{jk}} = \frac{\gamma \delta_{w_k k}}{(\gamma - 1)\delta_{w_k k} + \sum\limits_{j=1}^{c} \delta_{jk}} \ , \tag{9}$$

which implies

$$(1 - \alpha)(\gamma - 1)\delta_{w_k k} + (1 - \alpha) \sum_{j=1}^{c} \delta_{jk} + \alpha(\gamma - 1) \frac{\delta_{w_k k}^2}{\sum\limits_{j=1}^{c} \delta_{jk}} + \alpha \delta_{w_k k} = \gamma \delta_{w_k k} \ . \tag{10}$$

From here we intend to compute γ, so we go on this way:

$$\delta_{w_k k}(\gamma - 1) \left[(1 - \alpha) + \alpha \frac{\delta_{w_k k}}{\sum_{j=1}^{c} \delta_{jk}} - 1 \right] = (1 - \alpha) \left[\delta_{w_k k} - \sum_{j=1}^{c} \delta_{jk} \right] \ , \tag{11}$$

which then becomes

$$\alpha(\gamma - 1) \left[\frac{\delta_{w_k k}}{\sum_{j=1}^{c} \delta_{jk}} - 1 \right] = (1 - \alpha) \left[1 - \frac{\sum_{j=1}^{c} \delta_{jk}}{\delta_{w_k k}} \right] \ . \tag{12}$$

Taking in consideration the relation $u_{w_k k} = \delta_{w_k k} / \sum_{j=1}^{c} \delta_{jk}$, we get

$$(\gamma - 1)(u_{w_k k} - 1) = \frac{(1-\alpha)(u_{w_k k} - 1)}{\alpha u_{w_k k}} \quad . \tag{13}$$

In FCM, the degree of membership assigned to the winner cluster $u_{w_k k} = 1$ only if the input vector $\boldsymbol{x}_k$ and the cluster prototype $\boldsymbol{v}_{w_k}$ coincide. In this trivial case there is no need to compute the suppression as there is nothing to suppress. Excluding this trivial case, we may simplify the previous equation, so we get:

$$\gamma = 1 + \frac{1-\alpha}{\alpha u_{w_k k}} \quad . \tag{14}$$

On the other hand, if we start from (6) and (8) we obtain:

$$\frac{\delta_{ik}}{(\gamma - 1)\delta_{w_k k} + \sum_{j=1}^{c} \delta_{jk}} = \alpha \cdot \frac{\delta_{ik}}{\sum_{j=1}^{c} \delta_{jk}} \quad . \tag{15}$$

As δ_{ik} is never zero, this equation can be restructured as follows:

$$\sum_{j=1}^{c} \delta_{jk} = \alpha(\gamma - 1)\delta_{w_k k} + \alpha \sum_{j=1}^{c} \delta_{jk} \quad , \tag{16}$$

or

$$\gamma = 1 + \frac{(1-\alpha) \sum_{j=1}^{c} \delta_{jk}}{\alpha \delta_{w_k k}} = 1 + \frac{1-\alpha}{\alpha u_{w_k k}} \quad , \tag{17}$$

which, according to Eqs. (5) and (6), can be further transcribed as:

$$\gamma = 1 + \frac{1-\alpha}{\alpha u_{w_k k}} = \frac{\mu_{w_k k}}{\alpha u_{w_k k}} = \frac{\mu_{w_k k}}{\mu_{w_k k} - (1-\alpha)} \quad . \tag{18}$$

So we obtained the same γ value both ways. Although this is not yet the learning rate, we should discuss about the possible singularities:

- The degree of membership assigned by FCM to the winner class, $u_{w_k k}$, cannot be zero, because then all other u_{ik} values would be zero, and that contradicts the probability constraint of FCM.
- The suppression rate α can be zero, but that would reduce s-FCM to HCM, which is a trivial case with strict winner-takes-all competition.
- As the suppression rate α and the winner cluster's fuzzy membership are both in the interval $[0, 1]$, we indeed have $\gamma \geq 1$. Equality holds when $\alpha = 1$, that is, there is no suppression.

We can conclude, that Eq. (18) is valid if $0 < \alpha \leq 1$. Under these circumstances, the learning rate of the s-FCM is:

$$\eta_{\mathrm{s}} = 1 - \frac{d'_{w_k k}}{d_{w_k k}} = 1 - \gamma^{(1-m)/2} = 1 - \left(1 + \frac{1-\alpha}{\alpha u_{w_k k}}\right)^{(1-m)/2} \quad . \tag{19}$$

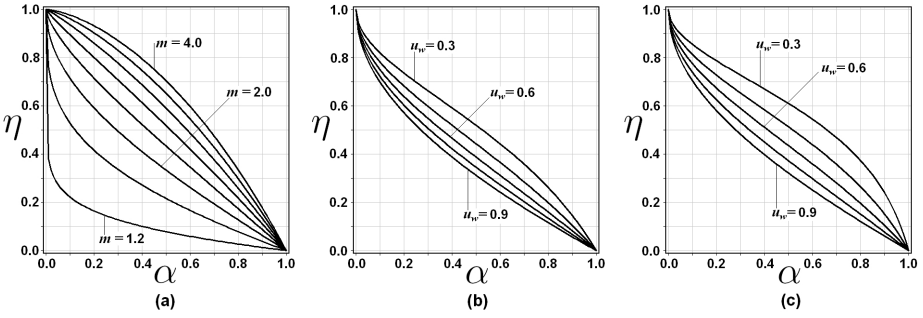


Fig. 1. Effect of sFCM's suppression rate on the learning rate: (a) with $u_w = 0.8$ and different values of m; (b) with $m = 2$ and different values of winner membership u_w (right); (c) Learning rate of Os-FCM plotted against the suppression rate, with $m = 2$ and different values of winner membership u_w

It was expected, that the fuzzyfication parameter m and the suppression rate α influence the learning rate. In addition, another factor is present, namely the fuzzy membership value of the winner cluster, $u_{w_k k}$. Some graphical representations of the learning rate vs. suppression rate are shown in Fig. 1. So far we can conclude, that s-FCM has a quasi-competitive behavior with a variable learning rate.

3.2 Is s-FCM Optimal? If so, What Does It Minimize?

Fan et al. [7] reported the fact, that s-FCM acts like HCM when the suppression rate is set to $\alpha = 0$, and to coincide with FCM when $\alpha = 1$. What happens if $0 < \alpha < 1$, remains a question.

Let us consider from the beginning, that $0 < \alpha < 1$, as the two extreme cases are trivial anyway and need no investigation. We can call s-FCM optimal if we find an objective function, whose AO minimization gives the optimization formulae:

$$\mu_{ik} = \alpha \cdot \frac{d_{ik}^{-2/(m-1)}}{\sum_{j=1}^{c} d_{jk}^{-2/(m-1)}} \quad \forall k = 1 \ldots n \quad \forall i \neq w_k \ , \tag{20}$$

$$\mu_{w_k k} = 1 - \alpha + \alpha \cdot \frac{d_{w_k k}^{-2/(m-1)}}{\sum_{j=1}^{c} d_{jk}^{-2/(m-1)}} \quad \forall k = 1 \ldots n, \; w_k = \arg\min_i \{d_{ik}\} \ , \tag{21}$$

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m \boldsymbol{x}_k}{\sum_{k=1}^{n} \mu_{ik}^m} \quad \forall i = 1 \ldots c \ . \tag{22}$$

Unfortunately, this kind of analytical function is not likely to exist. There exists at least one function, namely

$$J_{\text{not s-FCM}} = \sum_{k=1}^{n}\sum_{i=1}^{c}[\mu_{ik} - (1-\alpha)h_{ik}]^m d_{ik}^2 \ , \tag{23}$$

which satisfies the first two conditions for any $\alpha \in (0,1)$, but the generated cluster prototypes coincide with the ones of FCM. Let us now propose a novel approach for an optimal suppression and name it optimally suppressed fuzzy c-means algorithm (Os-FCM):

$$J_{\text{Os-FCM}} = \sum_{k=1}^{n}\sum_{i=1}^{c}[\alpha u_{ik}^m + (1-\alpha)h_{ik}]d_{ik}^2 \ , \tag{24}$$

where α is a parameter that is intended to mix fuzzy and hard c-means clustering, similarly to s-FCM, but creating a pure mixture of FCM and HCM. It is obvious, that there are two values of α, where s-FCM and Os-FCM coincide: 0 and 1, corresponding to HCM and FCM, respectively. In case of any other α, s-FCM and Os-FCM differ. The AO iteration formulae of Os-FCM are easy to obtain via the well-known technique of Lagrange multipliers. The update criteria we obtain for u_{ik} and h_{ik} are the same as in case of FCM and HCM algorithms, respectively. However, the update formula for cluster prototypes becomes:

$$\boldsymbol{v}_i = \frac{\sum\limits_{k=1}^{n}\left[\alpha u_{ik}^m + (1-\alpha)h_{ik}\right]\boldsymbol{x}_k}{\sum\limits_{k=1}^{n}\left[\alpha u_{ik}^m + (1-\alpha)h_{ik}\right]} \quad \forall i = 1\ldots c \ . \tag{25}$$

In the followings, we will compare the s-FCM algorithm with our newly proposed optimal clustering model, from two different points of view: (1) we will compute the quasi-competitive learning rate of Os-FCM and compare it with the one of s-FCM; (2) we will analyze the behavior of both algorithms by employing them to cluster the IRIS data [1] with several different settings.

Using the notations defined at the computation of the learning rate of s-FCM, and based on Eq. (25), we can write the following equation in these new circumstances:

$$\alpha\left(\frac{\delta_{w_k k}}{\sum_{j=1}^{c}\delta_{jk}}\right)^m + (1-\alpha) = \left(\frac{\gamma\delta_{w_k k}}{(\gamma-1)\delta_{w_k k} + \sum_{j=1}^{c}\delta_{jk}}\right)^m \ , \tag{26}$$

which represents the weighting coefficient received by a vector $\boldsymbol{x}_k$ whose competition was won by cluster prototype $\boldsymbol{v}_i$. According to Eq. (3), we can transcribe the following equation as:

$$\alpha u_{w_k k}^m + (1-\alpha) = \left(\frac{\gamma u_{w_k k}}{1 + (\gamma-1)u_{w_k k}}\right)^m \ , \tag{27}$$

which is an expression that is difficult to solve analytically. Let us ease the circumstances by setting $m = 2$, which then yields:

$$\gamma^2 u_{w_k k}^2 = [1 - \alpha + \alpha u_{w_k k}^2][1 + (\gamma - 1)u_{w_k k}]^2 \ . \tag{28}$$

This leads to the following second order equation in γ:

$$\alpha u_{w_k k}^2(1+u_{w_k k})\gamma^2 - 2u_{w_k k}[1-\alpha+\alpha u_{w_k k}^2]\gamma - (1-u_{w_k k})[1-\alpha+\alpha u_{w_k k}^2] = 0 \ , \tag{29}$$

which has only one acceptable (non-negative) solution:

$$\gamma = \frac{1 - \alpha + \alpha u_{w_k k}^2 + \sqrt{1 - \alpha + \alpha u_{w_k k}^2}}{\alpha u_{w_k k}(1 + u_{w_k k})} \ . \tag{30}$$

Let us verify the extreme values: if $\alpha \to 0$, then $\gamma \to \infty$, which is what we expected. Also, if we set $\alpha \to 1$, we get $\gamma \to \frac{u_{w_k k}^2 + u_{w_k k}}{u_{w_k k}(1+u_{w_k k})} = 1$, which means zero learning rate, corresponding to FCM. So the learning rate is given by:

$$\eta_{\mathrm{Os}} = 1 - \gamma^{\frac{1-m}{2}} = 1 - \sqrt{\frac{\alpha u_{w_k k}(1 + u_{w_k k})}{1 - \alpha + \alpha u_{w_k k}^2 + \sqrt{1 - \alpha + \alpha u_{w_k k}^2}}} \ , \tag{31}$$

its graphical representation is shown in Fig. 1(c). The curves on in this graph are quite similar but not identical with the ones shown in Fig. 1(b). This is not yet a proof for similar behavior of the two algorithms, this is what was possible to show the analytical way. The comparison of s-FCM and Os-FCM should continue with a numerical analysis.

4 Numerical Analysis

In the followings we will present some numerical analysis of the functional characteristics of the suppressed FCM algorithm. These tests are performed using the IRIS data [1], which consist of 150 labeled feature vectors of four dimensions, and the high dimensional vectors of the WINE data set [3].

A series of numerical tests targeted the clustering accuracy. It is well-known that the IRIS data cannot be perfectly classified without using the labels for supervised learning. In case of unsupervised clustering, a deterministic misclassification rate of 10% represents fine accuracy. In case of FCM, HCM, or s-FCM this accuracy is reached only if an intelligent prototype initialization scheme is applied. If the initial prototypes are not properly chosen, the algorithms might fail [2]. It is well known, that FCM is less sensitive to initialization than HCM.

We have tested the clustering accuracy and robustness, intentionally using a least smart initialization: randomly chosen input vectors, differing from each other, were set as initial cluster prototypes. The left columns in Fig. 2 show the confusion rate of s-FCM, meaning the percentage of cases when the clustering failed, vs. the suppression rate α. These results suggest, that s-FCM requires the

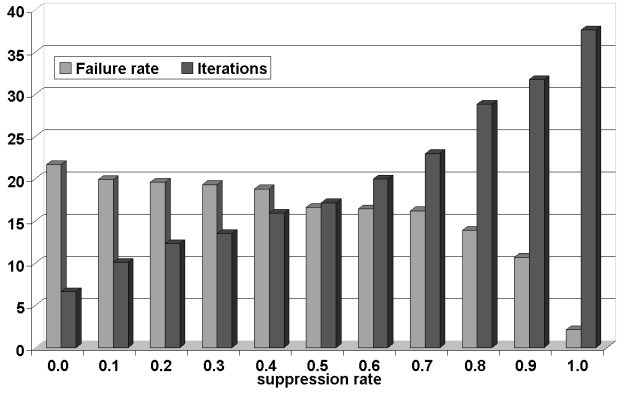


Fig. 2. Failure rate in case of random initialization, and the number necessary iterations to reach a strict convergence criterion, represented vs. the suppression rate

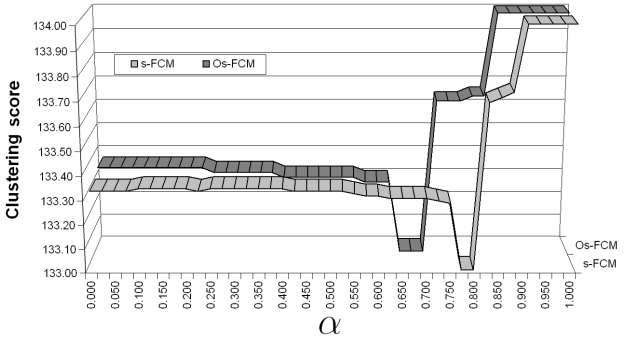


Fig. 3. Average clustering score of s-FCM and Os-FCM, plotted against α

proper initialization as much as HCM does. If the initial prototypes are chosen randomly, but choosing exactly one from each label class, then the confusion rate reduces to zero for all FCM, HCM, and s-FCM.

The right columns of in Fig. 2 show the number of necessary iterations vs. suppression rate, using the same threshold value in the stopping criterion. These values are in full accordance to those found in [7]: a suitably chosen suppression rate may reduce 2-3 times the necessary computation cycles. Taking into consideration, that an efficient implementation uses a few percent more computations to perform the fuzzy membership modification, we can conclude that suppressing the FCM algorithms is a useful idea.

Now let us compare the main behavioral parameters of the two algorithm. Figure 3 shows the clustering score (average number of correct decisions out of 150) of s-FCM and Os-FCM, averaged along several hundreds of tests performed with different initialization. Figure 4 presents the average number of iterations of s-FCM and Os-FCM, necessary to reach a given level of convergence. The data represented here are also computed from hundreds of tests.

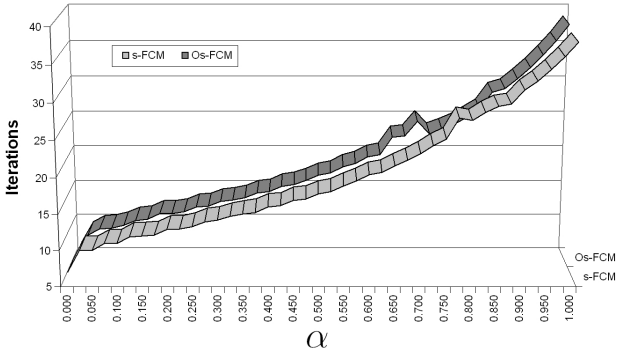


Fig. 4. Average number of necessary iterations for s-FCM and Os-FCM to reach an $\varepsilon = 10^{-8}$ convergence, plotted against α

From the shape of the two graphs it is visible, that the two methods might be related somehow. Not only the approximative values are similar for any α, but also the shape of the graphs. The first graph indicates that the result of clustering and the misclassification rate does depend on the initialization: the clustering score can be either 133 or 134 in all cases.

Tests with multidimensional data have revealed, that Os-FCM reaches the same convergence level in 4-5% less iterations, than s-FCM. Clustering accuracy of both algorithms is at the same level.

The shape of the graphs in Fig. 3 suggests that there might be a direct relation between α_s and α_{Os}. If s-FCM and Os-FCM were equivalent, then according to Eqs. (5), (6), (4) and (25), we should have $\forall u_{ik} \in [0,1], \forall m > 1$

$$\begin{cases} (\alpha_s u_{ik})^m = (\alpha_{Os} u_{ik})^m \\ (1 - \alpha_s + \alpha_s u_{ik})^m = (\alpha_{Os} u_{ik})^m + 1 - \alpha_{Os} \end{cases} . \tag{32}$$

In order to have any chance for equivalence between s-FCM(α_s) and Os-FCM (α_{Os}), this equation system should be compatible. But this isn't the case: the only case of compatibility is $m \to 1$ and $\alpha_s = \alpha_{Os}$. However, $m \to 1$ is the case of HCM, where there is nothing to suppress.

As a diagnosis of the suppressed FCM algorithm we can say: we cannot take for granted the optimality or non-optimality of s-FCM, but we can assert that it behaves very similarly to an optimal clustering model (Os-FCM).

5 Conclusions

In this paper we assumed to study the properties of the suppressed fuzzy c-means algorithm. Based on analytical computations, we have exploited the quasi-competitive behavior of s-FCM. On the other hand, using numerical analysis, we have have shown, that s-FCM succeeded to inherit the quick convergence of HCM, and the accuracy of FCM, but also has the disadvantage of being

almost as sensitive to prototype initialization as HCM is. We have proposed an optimal version of the suppressed FCM algorithm and compared the properties of the two algorithms via analytical computations and numerical tests. Although the two algorithms definitely differ, we found only slight differences between their performance: Os-FCM should be favored mostly because of its guaranteed optimality.

References

1. Anderson, E.: The IRISes of the Gaspe peninsula. Bull. Amer. IRIS Soc. 59, 2–5 (1935)
2. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proc. Symp. Discr. Alg., pp. 1027–1035 (2007)
3. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Benyó, B., Somogyi, P., Várady, P., Paláncz, B.: Classification of time series using singular values and wavelet subband analysis with ANN and SVM classifiers. J. Adv. Comput. Intell. & Intell. Inform. 10, 498–503 (2006)
5. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum, New York (1981)
6. Cannon, R.L., Dave, J.V., Bezdek, J.C.: Efficient implementation of the fuzzy c-means clustering algorithms. IEEE Trans. Patt. Anal. Machine Intell. 8, 248–255 (1986)
7. Fan, J.L., Zhen, W.Z., Xie, W.X.: Suppressed fuzzy c-means clustering algorithm. Patt. Recogn. Lett. 24, 1607–1612 (2003)
8. Hathaway, R.J., Bezdek, J.C.: Optimization of clustering by reformulation. IEEE Trans. Fuzzy Syst. 3, 241–245 (1995)
9. Hathaway, R.J., Bezdek, J.C., Hu, Y.: Generalized fuzzy c-means clustering strategies using L_p norm distances. IEEE Trans. Fuzzy Syst. 8, 576–582 (2000)
10. Hung, W.L., Yang, M.S., Chen, D.H.: Parameter selection for suppressed fuzzy c-means with an application to MRI segmentation. Patt. Recogn. Lett. 27, 424–438 (2006)
11. Hung, W.L., Chang, Y.C.: A modified fuzzy c-means algorithm for differentiation in MRI of ophtalmology. LNCS, vol. 24, pp. 340–350 (2006)
12. Kamel, M.S., Selim, S.Z.: New algorithms for solving the fuzzy clustring problem. Patt. Recogn. 27, 421–428 (1994)
13. Karayiannis, N.B., Bezdek, J.C.: An integrated approach to fuzzy learning vector quantization and fuzzy c-means clustering. IEEE Trans. Fuzzy Syst. 5, 622–628 (1997)
14. Tsao, E.C.K., Bezdek, J.C., Pal, N.R.: Fuzzy Kohonen clustering networks. Patt. Recogn. 27, 757–764 (1994)
15. Vapnik, V.: Statistical learning theory. Wiley, New York (1998)
16. Wei, L.M., Xie, W.X.: Rival checked fuzzy c-means algorithm. Acta Electr. Sin. 28, 63–66 (2000)
17. Xie, Z., Wang, S., Chung, F.L.: An enhanced possibilistic c-means clustering algorithm. Soft Comput. 12, 593–611 (2008)
18. Yair, E., Zeger, K., Gersho, A.: Competitive learning and soft competition for vector quantization design. IEEE Trans. Sign. Proc. 40, 294–309 (1992)
19. Zadeh, L.A.: Fuzzy sets. Inform. Contr. 8, 338–353 (1965)

Generalized Agglomerative Clustering with Application to Information Systems

Sadaaki Miyamoto

Department of Risk Engineering
Faculty of Systems and Information Engineering
University of Tsukuba, Ibaraki 305-8573, Japan
miyamoto@risk.tsukuba.ac.jp

Abstract. Although data clustering is relatively uninvestigated in rough set studies, there are much room for applying clustering and related techniques to this field. In this paper we focus on generalization of agglomerative clustering to information systems. A poset-valued hierarchical clustering is defined and the combination of traditional agglomerative clustering and lattice diagram of attributes in an information system is considered. Inner product spaces are available to information systems by using kernel functions in support vector machines. Different algorithms for generalized agglomerative clustering using the inner product are described. Illustrative examples are shown.

1 Introduction

Although rough sets [15,16] provide logical and analytical framework for classification, there is another area relatively unnoticed, that is, the concept of unsupervised classification [3], alias data clustering [1,4]. Although a few studies discuss clustering algorithms using rough set concepts [6,5], there are still many research possibilities of which some possible directions are shown in [14].

The present paper discusses agglomerative clustering algorithms for information systems [16]. The method of poset-valued hierarchical classification in [14] is further generalized and the use of an inner-product space is considered, which necessitates the introduction of a kernel function [17,18].

Two types of kernel functions are studied, one is based on an ordinary Gaussian kernel which has an implicit mapping into a high-dimensional Euclidean space [18], while the second has an explicit form of the mapping.

Moreover simple illustrative examples are given to grasp the concept of poset-valued clustering.

The rest of this paper is organized as follows. In Section 2, we first review agglomerative clustering that has an abstract formulation and its generalization into poset-valued hierarchical classification. Section 3 discusses inner product space for information systems that uses kernel functions for agglomerative clustering. Section 4 proposes a fuzzy subset system to obtain a family of kernels. Finally, Section 5 gives concluding remarks.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 158–166, 2008.

Throughout this paper, proofs of the propositions are omitted, as they are not difficult.

2 Generalization of Agglomerative Clustering

We first assume that the set of objects for clustering is denoted by $\mathcal{T} = \{t_1, \ldots, t_n\}$ and a generic object in $\mathcal{T}$ is also denoted by $t \in \mathcal{T}$. A dissimilarity measure $d(t, t')$ is assumed to be given between an arbitrary pair of objects $t, t' \in \mathcal{T}$; the way how a dissimilarity measure is defined will be discussed below.

A family of clusters of $\mathcal{T}$ denoted by $\mathcal{G} = \{G_1, \ldots, G_K\}$ is a partition of $\mathcal{T}$:

$$\bigcup_{i=1}^{K} G_i = \mathcal{T}, \qquad G_i \cap G_j = \emptyset \quad (i \neq j).$$

Assume that an inter-cluster distance which will be discussed below is denoted by $d(G_i, G_j)$. Furthermore, we consider a family of clusters that depends on a parameter α:

$$\mathcal{G}(\alpha) = \{G_1(\alpha), \ldots, G_K(\alpha)\}.$$

Accordingly the inter-cluster distance is $d(G_i(\alpha), G_j(\alpha))$.

2.1 Dissimilarity in an Information System

An information system is assumed to have a form of a table in which each row corresponds to an object t_k, while each column indicates an attribute a_i, $i = 1, \ldots, m$. The (k, i)-cell for an object t_k and a_i is denoted by $v_{ki} = t_k(a_i)$. The value v_{ki} can either be a nonnumeric symbol or numerical in general. We assume v_{ki} is a nonnumeric symbol, but a numerical value can be handled in a similar manner without difficulty.

For simplicity we assume that different attributes have different symbols: $v_{ki} \neq v_{lj}$ for $i \neq j$.

A simple measure of dissimilarity is assumed to be given between two symbols v_{ki} and v_{li} for an attribute a_i:

$$d(v_{ki}, v_{li}; a_i) = \begin{cases} 1 & (v_{ki} \neq v_{li}), \\ 0 & (v_{ki} = v_{li}). \end{cases} \tag{1}$$

Moreover we define

$$d(t_k, t_l; a_i) = d(t_k(a_i), t_l(a_i); a_i) \tag{2}$$

For the sake of convenience, The set of attributes is denoted by $\mathcal{A} = \{a_1, \ldots, a_m\}$. A subset of $\mathcal{A}$ is denoted by $A' = \{a_i, \ldots, a_\ell\}$, A'', and so on.

The dissimilarity (2) is naturally extended to a subset A':

$$d(t, t'; A') = \sum_{a_i \in A'} d(t(a_i), t'(a_i); a_i) \tag{3}$$

2.2 Agglomerative Clustering and Hierarchical Classification

A general procedure of agglomerative clustering is as follows [8,10].

1. Let the initial clusters be individual objects: $G_k = \{t_k\}$. Let $\mathcal{G} = \{G_1, \ldots, G_K\}$. Define inter-cluster distances as the distance between the corresponding objects: $d(G_i, G_j) = d(t_i, t_j)$. Let the number of clusters be $K = n$.
2. Merge two clusters $G' = G_p \cup G_q$ of the minimum distance:

$$(G_p, G_q) = \arg\min_{i,j} d(G_i, G_j)$$

   Add G' to $\mathcal{G}$ and delete G_p and G_q from $\mathcal{G}$. Reduce the number of clusters: $K \leftarrow K - 1$. The minimum value is stored as the level of the merge m_K:

$$m_K = d(G_p, G_q) = \min_{i,j} d(G_i, G_j).$$

3. If $K = 1$, stop, else update the distances between the merged cluster and other clusters $d(G', G_r)$, $r = 1, \ldots, K$. Go to step **2**.

There are several ways to update the distances in step **3**, and accordingly we have a number of methods of agglomerative clustering such as the single link, the complete link, and the Ward method which we will discuss here. It has been known that, in these methods, the level m_K is monotone non-decreasing:

$$m_{n-1} \leq m_{n-2} \leq \cdots \leq m_2 \leq m_1.$$

Assume $\alpha = m_K$. Then the next property is valid.

Proposition 1. *For every $\alpha \leq \alpha'$ and for each $G_i(\alpha') \in \mathcal{G}(\alpha')$ there exists $G_j(\alpha) \in \mathcal{G}(\alpha)$ such that $G_j(\alpha) \subseteq G_i(\alpha')$.*

2.3 Poset-Valued Hierarchical Classification

We generalize the last property and define a poset-valued hierarchical classification.

Definition 1. *Let P be a poset [7] of which the preorder is defined by $\preceq$. We say $\mathcal{G}(\alpha) = \{G_1(\alpha), \ldots, G_K(\alpha)\}$ $(\alpha \in P)$ is a poset-valued hierarchical classification if for every $\alpha \preceq \alpha'$ and for each $G_i(\alpha') \in \mathcal{G}(\alpha')$ there exists $G_j(\alpha) \in \mathcal{G}(\alpha)$ such that*

$$G_j(\alpha) \subseteq G_i(\alpha').$$

We write $\mathcal{G}(\alpha) \triangleright \mathcal{G}(\alpha')$ if this property holds.

Example 1. Let $P = 2^{\mathcal{A}}$: the collection of subsets of $\mathcal{A}$. P is then a lattice, i.e., a poset, by the natural inclusion of subsets.

Assume $\alpha = A'$, a subset of $\mathcal{A}$. Generate a cluster $G(\alpha) = G(A')$:

$$t, t' \in G(A') \iff d(t, t'; A') = 0$$

Table 1. An example of an information table

$\mathcal{T}$	D	E	F
t_1	a_1	b_1	c_1
t_2	a_1	b_1	c_2
t_3	a_1	b_2	c_1
t_4	a_1	b_2	c_2
t_5	a_2	b_1	c_1
t_6	a_2	b_1	c_2
t_7	a_2	b_2	c_1

It then is easy to see that the collection of $G(A')$ forms a poset-valued hierarchical classification.

Example 2. Consider the seven tuples shown in Table 1 with the schema $\mathcal{A} = (D, E, F)$. Here these three letters are attributes. The poset is $\Lambda = 2^{\mathcal{A}} = \{\emptyset, D, E, F, DE, DF, FE, DEF\}$ where the abbreviated symbol DE implies $\{D, E\}$, and so on. We have

$$\begin{aligned} \mathcal{G}(DEF) &= \{t_1, \ldots, t_7\}, \\ \mathcal{G}(DE) &= \{t_1t_2, t_3t_4, t_5t_6, t_7\} \end{aligned}$$

etc. where t_it_j is an abbreviated symbol for $\{t_i, t_j\}$.

Figure 1 shows the Hasse diagram of $\Lambda = 2^{\mathcal{A}}$ together with the partitions attached to each element of the lattice.

Example 3. Let us fix a subset $A' \in \mathcal{A}$ and perform an agglomerative clustering based on the dissimilarity $d(t, t'; A')$ using the single link method.

Let α be the pair of A' and m_K: $\alpha = (A', m_K)$, where m_K is an arbitrary level of the merge in the agglomerative algorithm. Then the all collection of α forms a poset

$$P = (\mathcal{A}, \{m_{n_1}, \ldots, m_1\}) \tag{4}$$

by the natural ordering:

$$\alpha \preceq \alpha' \iff A' \supseteq A'', \quad m_K \leq m_{K'}. \tag{5}$$

Then, it is not difficult to see that the next proposition holds.

Proposition 2. *The cluster generated at any $\alpha = (A', m_K) \in P$ using the single link forms a hierarchical classification:*

$$\alpha \preceq \alpha' \;\Rightarrow\; G(\alpha) \triangleright G(\alpha').$$

Example 4. Figure 2 shows an illustration of Example 3 in which Table 1 and the single link is used. For each node of the Hasse diagram, a dendrogram is attached. If no reversal in the dendrogram exists [8,10], then such a figure defines a poset-valued classification in the above sense.

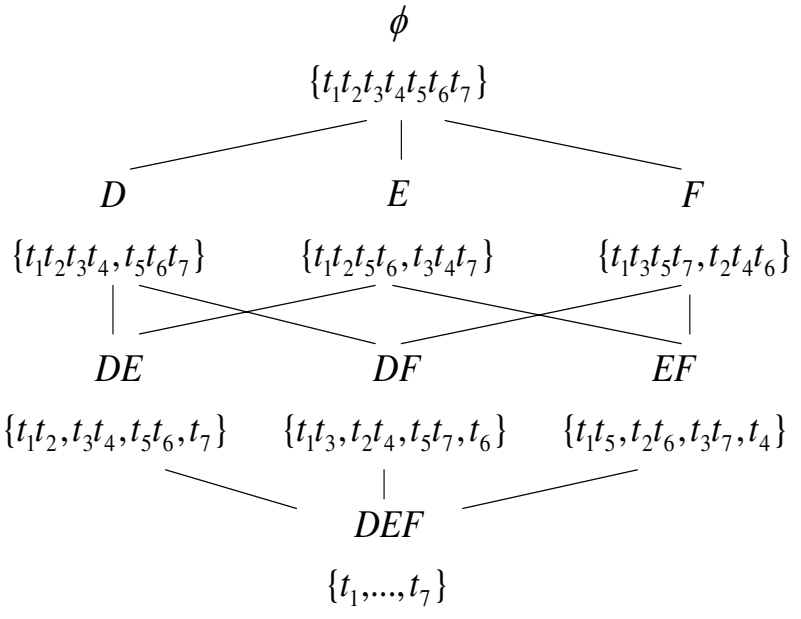


Fig. 1. An example of the poset-valued clustering

3 Inner Product Space for Information Systems

Most methods of data analysis assume that the underlying space is a Euclidean space, alias an inner product space. There are several ways to introduce a Euclidean space to information systems.

The simplest way is to map each symbol to a Euclidean space of a high-dimension: Suppose all symbols are $v_1, \ldots, v_p$. They are mapped into $\boldsymbol{R}^p$:

$$\Psi(v_1) = \frac{1}{2}(1, 0, \ldots, 0),\ \Psi(v_2) = \frac{1}{2}(0, 1, 0, \ldots, 0),\ \ldots \tag{6}$$

It is immediate to see

$$d(t, t'; a_i) = d(t(a_i), t'(a_i)) = \|\Psi(t(a_i)) - \Psi(t'(a_i))\|^2. \tag{7}$$

and hence

$$d(t, t'; A') = \sum_{a_i \in A'} \|\Psi(t(a_i)) - \Psi(t'(a_i))\|^2. \tag{8}$$

Thus for a subset A', the Cartesian product of the Euclidean space is the corresponding space. The Ward method of agglomerative clustering assumes a Euclidean space. Hence the above equality implies that the Ward method can be used for the objects in $\mathcal{T}$ for any subset A'. We thus observe that the cluster generated at $\alpha \in P$ using the Ward method forms a hierarchical classification as above.

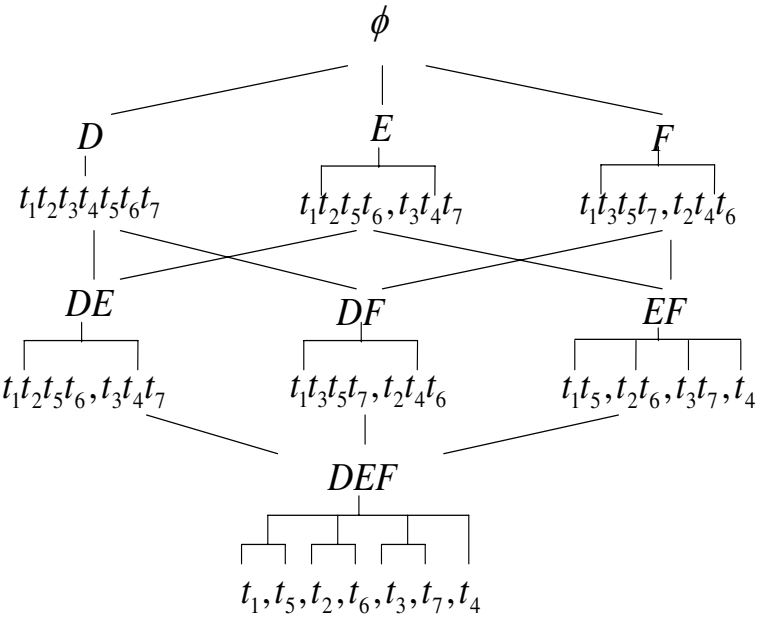


Fig. 2. An example of the poset-valued clustering with dendrograms

3.1 Kernel-Based Methods

Another approach to introduce an inner product space is to employ a kernel function [18,17] which has been used in support vector machines [18].

We have the next proposition.

Proposition 3

$$K_1(t, t'; A') = \exp(-\gamma d(t, t'; A')) \tag{9}$$

is a positive-definite kernel function, where γ is a positive constant and $d(t, t'; A')$ is given by (8).

The proof of the above proposition is given in [14]. Notice that $K_1(t, t'; A')$ is a kernel function for all $A' \in \mathcal{A}$.

The Ward Method Using Kernel Functions. It is well-known that the single link, the complete link, and the average link can be used for any kind of dissimilarity measure, while the Ward method is based on an Euclidean space.

The updating algorithm of the Ward method in Step 3 of the agglomerative clustering is as follows [10].

$$\begin{aligned} d(G', G_r) = \frac{1}{|G_p| + |G_q| + |G_r|} \Big[& (|G_p| + |G_r|) d(G_p, G_r) \\ & + (|G_q| + |G_r|) d(G_q, G_r) - |G_r| d(G_p, G_q) \Big] \end{aligned} \tag{10}$$

It is easy to see that this updating equation can be used for any kernel function, since a kernel function maps an object into a Euclidean space. Note also that the initial dissimilarity values should be defined in terms of $K(t,t')$:

$$d(t,t') = \frac{1}{2}\{K(t,t) + K(t',t') - 2K(t,t')\}.$$

Proposition 4. *The cluster generated at any $\alpha = (A', m_K) \in P$ using the Ward method with a positive definite kernel forms a hierarchical classification:*

$$\alpha \preceq \alpha' \;\Rightarrow\; G(\alpha) \triangleright G(\alpha').$$

4 A Fuzzy Subset System for a Dissimilarity

An important generalization of the dissimilarity (3) is defined by introducing a system of fuzzy subsets on $\mathcal{T}$.

Let F_ℓ $(\ell = 1, \ldots, H)$ be a system of fuzzy subsets on $\mathcal{T}$. We define a nonnegative definite kernel $K_2(t,t')$ by

$$K_2(t,t') = \sum_{\ell=1}^{H} \mu_{F_\ell}(t)\mu_{F_\ell}(t'). \tag{11}$$

It is immediate to see that the next proposition holds.

Proposition 5. *$K_2(t,t')$ is positive definite if and only if the set of vectors*

$$(\mu_{F_\ell}(t_1), \ldots, \mu_{F_\ell}(t_n)), \quad \ell = 1, \ldots, H$$

is linearly independent.

The proof is easy and omitted. We hereafter assume that the assumption on the linear independence is satisfied.

Using this kernel, we define the norm of t and the distance between t and t':

$$\|t\|^2_{K_2} = K_2(t,t) = \sum_{\ell=1}^{H} \{\mu_{F_\ell}(t)\}^2, \tag{12}$$

$$\|t - t'\|^2_{K_2} = K_2(t,t) = \sum_{\ell=1}^{H} \{\mu_{F_\ell}(t) - \mu_{F_\ell}(t')\}^2. \tag{13}$$

Accordingly, the dissimilarity using K_2 is

$$d_{K_2}(t,t') = \|t - t'\|^2_{K_2}. \tag{14}$$

We should apply this distance to an information system. For a given attribute value v_ℓ of a cell in an information system, we define $\mu_{F_\ell}(t_k)$:

$$\mu_{F_\ell}(t_k) = \begin{cases} 1 & (\exists a_i,\ t_k(a_i) = v_\ell), \\ 0 & (\text{otherwise}). \end{cases} \tag{15}$$

Actually the above F_i is not fuzzy but a crisp set.

Then the dissimilarity is defined by

$$d_{K_2}(t, t'; A') = \|t - t'\|^2_{K_2} \tag{16}$$

where attribute values v_ℓ are limited to those in A'. We have

Proposition 6. *The dissimilarity $d_{K_2}(t, t'; A')$ given by (16) is equal to the first dissimilarity (3).*

Generally the dissimilarity (16) need not be identical to the one by (3), in other words, F_i may be fuzzy and fuzziness is defined to express relatedness among different attribute values.

We remind that a kernel-based method employs a nonlinear mapping Φ from the object space into a high-dimensional inner product space. In this sense we have three mappings:

1. First method using (6) uses $\Psi(\cdot)$ as the mapping Φ.
2. Second method using
$$K_1(t, t'; A') = \exp(-\gamma d(t, t'; A'))$$
uses an implicit mapping Φ, i.e., an explicit form of Φ is unavailable.
3. Third method uses
$$\Phi(t) = (\mu_{F_1}(t), \ldots, \mu_{F_H}(t)).$$
Thus the mapping $\Phi(t)$ is explicit.

Different algorithms should be used in accordance with the explicitness of the mapping. Generally, more efficient algorithms can be used when the mapping has explicit forms, as the first and third ones. For example, $O(n^3)$ order algorithms should be used for an explicit mapping in the crisp and fuzzy c-means clustering [1,9,10], while the algorithms with explicit mappings require $O(n)$ complexity [11,12,13]. Notice also that the updating formula (10) is valid for both explicit and implicit mappings.

5 Conclusion

The concept of poset-valued agglomerative clustering has been discussed and the use of kernel functions is considered. The kernel function embeds the objects in an inner product space, whereby various methods of data analysis are applicable to information systems in addition to clustering, although such an idea is different from traditional approach in rough set studies. Hence such a method based on *continuity* should be compared with traditional *logical* approaches in order to uncover their intrinsic relations.

There seem to be a relation, e.g., in fuzzy subset systems, since a generalization of rough sets uses such a system of subsets whereby upper and lower approximations can be defined. Moreover a subset system can be compared with the concept of neighborhoods used in generalized rough sets [19] and modal logic systems [2].

Future studies will include studies of theoretical relations with above studies, and applications to real-world problems expressed as information systems.

Acknowledgment. This study has been supported by the Grant-in-Aid for Scientific Research, No.19300074, JSPS, Japan.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Chellas, B.F.: Modal Logic. Cambridge University Press, Cambridge (1980)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
4. Everitt, B.S.: Cluster Analysis, 3rd edn., Arnold, London (1993)
5. Hirano, S., Tsumoto, S.: A framework for unsupervised selection of indiscernibility threshold in rough clustering. In: Greco, S., et al. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 872–881. Springer, Heidelberg (2006)
6. Lingras, P., West, C.: Interval set clustering of web users with rough K-means. J. of Intel. Informat. Sci. 23(1), 5–16 (2004)
7. MacLane, S., Birkhoff, G.: Algebra, 2nd edn. Macmillan, Basingstoke (1979)
8. Miyamoto, S.: Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer, Dordrecht (1990)
9. Miyamoto, S., Mukaidono, M.: Fuzzy c-means as a regularization and maximum entropy approach. In: Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA 1997), Prague, Czech, June 25-30, 1997, vol. II, pp. 86–92 (1997)
10. Miyamoto, S.: Introduction to Cluster Analysis: Theory and Applications of Fuzzy Clustering, Morikita-Shuppan, Tokyo (1990) (in Japanese)
11. Miyamoto, S., Suizu, D.: Fuzzy c-means clustering using transformations into high dimensional spaces. In: Proc. of FSKD 2002: 1st International Conference on Fuzzy Systems and Knowledge Discovery, Singapore, November 18-22, 2002, pp. 656–660 (2002)
12. Miyamoto, S., Nakayama, Y.: Algorithms of hard c-means clustering using kernel functions in support vector machines. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(1), 19–24 (2003)
13. Miyamoto, S., Suizu, D.: Fuzzy c-means clustering using kernel functions in support vector machines. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(1), 25–30 (2003)
14. Miyamoto, S.: Data Clustering Algorithms for Information Systems. In: An, A., Stefanowski, J., Ramanna, S., Butz, C.J., Pedrycz, W., Wang, G. (eds.) RSFDGrC 2007. LNCS (LNAI), vol. 4482, pp. 13–24. Springer, Heidelberg (2007)
15. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
16. Pawlak, Z.: Rough Sets. Kluwer Academic Publishers, Dordrecht (1991)
17. Scholkopf, B., Burges, C.J.C., Smola, A.J. (eds.): Advances in Kernel Methods: Support Vector Learning. MIT, Cambridge (1999)
18. Vapnik, V.N.: Statistical Learning Theory. Wiley, Chichester (1998)
19. Yao, Y.Y., Wong, S.K.M., Lin, T.Y.: A review of rough set models. In: Lin, T.Y., Cercone, N. (eds.) Rough Sets and Data Mining: Analysis of Imprecise Data, pp. 47–75. Kluwer, Boston (1997)

A Comprehensive Study on Reducts in Dominance-Based Rough Set Approach

Yoshifumi Kusunoki and Masahiro Inuiguchi

Graduate School of Engineering Science, Osaka University,
1-3, Machikaneyama, Toyonaka, Osaka 560-8531, Japan
kusunoki@inulab.sys.es.osaka-u.ac.jp, inuiguti@sys.es.osaka-u.ac.jp

Abstract. In this paper, we propose new reducts in the dominance-based rough set approach. The relations with previous ones are clarified. Moreover, a comprehensive enumeration method of all kinds of reducts is proposed. We show that all kinds of reducts are enumerated based on two discernibility matrices associated with generalized decisions.

1 Introduction

Rough sets [6] have been applied to the analysis of decision tables. Because of its usefulness, rough set approaches have been used in pattern recognition, machine learning, knowledge discovery, medical informatics, decision analysis, kansei engineering and so on [3,7]. The classical rough sets are defined using the indiscernibility relation, i.e., an equivalence relation. In decision tables, the indiscernibility relation implies that all attributes are nominal. However, in the real world, we may face cases when some attribute values are ordinal. For example, consider two test scores as condition attributes and the comprehensive evaluation as the decision attribute, all attributes are ordinal. As in this example, we may sometimes suppose the monotonicity between the decision attribute and the condition attributes.

When we suppose a monotonicity between the decision attribute and the condition attributes, the classical rough set approach developed for nominal attributes is not sufficient as demonstrated by Greco et al. [2]. To overcome this insufficiency, Greco et al. [2] proposed *the dominance-based rough set approach (DRSA)*. In DRSA, upward/downward unions of decision classes instead of decision classes are approximated based on dominance relations instead of indiscernibility relations. The similar analysis to the classical rough set analysis can be performed by DRSA.

The attribute reduction is also discussed in DRSA. By the attribute reduction, superfluous attributes are removed and we find important attributes as a set of condition attributes called a *reduct*. Susmaga et al. [8] proposed reducts preserving an information measure called *a quality of sorting*. Yang et al. [9] have proposed four kinds of reducts preserving lower/upper approximations of upward/downward unions of decision classes. Inuiguchi and Yoshioka [5] have proposed several kinds of reducts preserving upper approximations, lower approximations, and/or boundary regions of upward and/or downward unions.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 167–178, 2008.

The four kinds of reducts by Yang et al. [9] are included in many kinds of reducts by Inuiguchi and Yoshioka [5]. Those are called *union-based reducts* because they do not consider the approximations of decision classes but those of upward/downward unions of decision classes. The relations among union-based reducts have been investigated in [5] but the relations between the reduct by Susmaga et al. [8] and union-based reducts have not yet.

In this paper, we define lower and upper approximations and boundary regions of decision classes in DRSA and investigate reducts preserving those approximations and boundary regions. Because we consider the approximations of decision classes, we call them *class-based reducts*. We investigate the relations among the Susmaga's reduct, union-based reducts and class-based reducts. Moreover, we propose a comprehensive method enumerating all reducts based on *the discernibility matrices* [7]. We show that all kinds of reducts can be enumerated by two discernibility matrices associated with *generalized decisions* [1].

In next section, DRSA and previous reducts are reviewed. In Section 3, approximations of decision classes in DRSA are defined and the reducts based on these approximations are proposed. Moreover, relations among many kinds of reducts are investigated. In Section 4, giving reducts based on generalized decisions, we show all kinds of reducts are enumerated by two discernibility matrices associated with general decisions. Finally, we describe conclusions in Section 5.

2 Dominance-Based Rough Set Approach and Reducts

2.1 Dominance-Based Rough Set Approach (DRSA)

In DRSA [2,8], decision tables with order relations are analyzed. A *decision table* is defined by a quadruple $\mathcal{T} = \langle U, C \cup \{d\}, V, f \rangle$, where U is a finite set of *objects* (universe), C is a finite set of *condition attributes*, $d \notin C$ is a *decision attribute*, V_q is the domain of the attribute q, $V = \bigcup_{q \in C \cup \{d\}} V_q$ is a set of *attribute values* and a total function $f : U \times C \cup \{d\} \to V$ such that $\forall x \in U, \forall q \in C \cup \{d\}, f(x,q) \in V_q$ is called an *information function*. In DRSA, we assume a total order $\succeq_d$ on V_d, a preorder $\succeq_q$ on V_q for $q \in Q \subseteq C$, where a preorder is a reflexive and transitive relation and a total order is a preorder satisfying anti-symmetry and comparability. For the sake of simplicity, we write $x \succeq_q y$ instead of $f(x,q) \succeq_q f(y,q)$ for $q \in C \cup \{d\}$. As a background knowledge, we assume a monotonicity such that $x \succeq_q y$ for all $q \in C$ implies $x \succeq_d y$.

The relation $x \succeq_q y$ is interpreted as "x is at least as good as y with respect to attribute q". For $P \subseteq C$, we define a dominance relation $\succeq_P$ by $x \succeq_P y$ if and only if $\forall q \in P$, $x \succeq_q y$, where $x \succeq_P y$ implies that x *dominates* y with respect to P. The decision attribute d partitions U into a set of decision classes $\mathcal{C} = \{Cl_1, Cl_2, ..., Cl_n\}$. For simplicity, we define $T = \{1, 2, ..., n\}$. We assume $s > t$ if and only if $\forall x \in Cl_s$, $\forall y \in Cl_t$, $x \succ_d y$.

In DRSA, the following *upward unions* $Cl_t^{\geq}$ and *downward unions* $Cl_t^{\leq}$ of decision classes are approximated by means of the dominance relation:

$$Cl_t^{\geq} = \bigcup_{k \geq t} Cl_k, \quad Cl_t^{\leq} = \bigcup_{k \leq t} Cl_k, \ t \in T \tag{1}$$

We have $Cl_1^{\geq} = Cl_n^{\leq} = U$, $Cl_t^{\geq} = U - Cl_{t-1}^{\leq}$ $(t \geq 2)$. Using dominance relation $\succeq_P$ $(P \subseteq C)$, P-dominating set and P-dominated set are respectively defined by

$$D_P^+(x) = \{y \in U \mid y \succeq_P x\}, \quad D_P^-(x) = \{y \in U \mid x \succeq_P y\}. \tag{2}$$

P-lower and *P-upper approximations* of $Cl_t^{\geq}$ are respectively defined by

$$\underline{P}(Cl_t^{\geq}) = \{x \in U \mid D_P^+(x) \subseteq Cl_t^{\geq}\}, \ \overline{P}(Cl_t^{\geq}) = \{x \in U \mid D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset\}. \tag{3}$$

Similarly, *P-lower* and *P-upper approximations* of $Cl_t^{\leq}$ are defined by

$$\underline{P}(Cl_t^{\leq}) = \{x \in U \mid D_P^-(x) \subseteq Cl_t^{\leq}\}, \ \overline{P}(Cl_t^{\leq}) = \{x \in U \mid D_P^+(x) \cap Cl_t^{\leq} \neq \emptyset\}. \tag{4}$$

The difference between the upper and lower approximations is called a *boundary region*, the boundary regions $Bn_P(Cl_t^{\geq})$ and $Bn_P(Cl_t^{\leq})$ are defined by

$$Bn_P(Cl_t^{\geq}) = \overline{P}(Cl_t^{\geq}) - \underline{P}(Cl_t^{\geq}), Bn_P(Cl_t^{\leq}) = \overline{P}(Cl_t^{\leq}) - \underline{P}(Cl_t^{\leq}). \tag{5}$$

2.2 Properties of Upper and Lower Approximations

Let $P \subseteq C$ and $t \in T$. Then the upper and lower approximations and boundary regions satisfy the following properties [2]:

$$\overline{P}(Cl_1^{\geq}) = \underline{P}(Cl_1^{\geq}) = U, \quad \overline{P}(Cl_n^{\leq}) = \underline{P}(Cl_n^{\leq}) = U, \tag{6}$$

$$\underline{P}(Cl_t^{\geq}) \subseteq Cl_t^{\geq} \subseteq \overline{P}(Cl_t^{\geq}), \quad \underline{P}(Cl_t^{\leq}) \subseteq Cl_t^{\leq} \subseteq \overline{P}(Cl_t^{\leq}), \tag{7}$$

$$\overline{P}(Cl_t^{\geq}) = U - \underline{P}(Cl_{t-1}^{\leq}), \quad \overline{P}(Cl_{t-1}^{\leq}) = U - \underline{P}(Cl_t^{\geq}), \ (t \geq 2) \tag{8}$$

$$\overline{P}(Cl_t^{\geq}) \cup \overline{P}(Cl_{t-1}^{\leq}) = U, \quad (t \geq 2) \tag{9}$$

$$Bn_P(Cl_t^{\geq}) = Bn_P(Cl_{t-1}^{\leq}), \quad (t \geq 2) \tag{10}$$

$$\overline{P}(Cl_t^{\geq}) = Bn_P(Cl_t^{\geq}) \cup Cl_t^{\geq}, \quad \overline{P}(Cl_t^{\leq}) = Bn_P(Cl_t^{\leq}) \cup Cl_t^{\leq}, \tag{11}$$

$$\underline{P}(Cl_t^{\geq}) = Cl_t^{\geq} - Bn_P(Cl_t^{\geq}), \quad \underline{P}(Cl_t^{\leq}) = Cl_t^{\leq} - Bn_P(Cl_t^{\leq}). \tag{12}$$

Let $P, Q \subseteq C$ and $s, t \in T$. Then, we have the following monotonicity properties:

$$s \geq t \Rightarrow \underline{P}(Cl_s^{\geq}) \subseteq \underline{P}(Cl_t^{\geq}), \quad s \leq t \Rightarrow \underline{P}(Cl_s^{\leq}) \subseteq \underline{P}(Cl_t^{\leq}), \tag{13}$$

$$s \geq t \Rightarrow \overline{P}(Cl_s^{\geq}) \subseteq \overline{P}(Cl_t^{\geq}), \quad s \leq t \Rightarrow \overline{P}(Cl_s^{\leq}) \subseteq \overline{P}(Cl_t^{\leq}), \tag{14}$$

$$Q \subseteq P \Rightarrow \underline{Q}(Cl_t^{\geq}) \subseteq \underline{P}(Cl_t^{\geq}), \quad Q \subseteq P \Rightarrow \underline{Q}(Cl_t^{\leq}) \subseteq \underline{P}(Cl_t^{\leq}), \tag{15}$$

$$Q \subseteq P \Rightarrow \overline{Q}(Cl_t^{\geq}) \supseteq \overline{P}(Cl_t^{\geq}), \quad Q \subseteq P \Rightarrow \overline{Q}(Cl_t^{\leq}) \supseteq \overline{P}(Cl_t^{\leq}). \tag{16}$$

2.3 Generalized Decision

The generalized decision [1] plays an important role in discernibility matrices described in Section 4. Let $P \subseteq C$ and $x \in U$. Then *P-generalized decision* $\delta_P(x)$ is defined by $\delta_P(x) = \langle l_P(x), u_P(x) \rangle$, where we define

$$l_P(x) = \min\{t \in T \mid D_P^+(x) \cap Cl_t \neq \emptyset\}, \tag{17}$$

$$u_P(x) = \max\{t \in T \mid D_P^-(x) \cap Cl_t \neq \emptyset\}. \tag{18}$$

$\delta_P(x)$ shows the interval of decision classes to which x may belong. $l_P(x)$ and $u_P(x)$ are the lower and upper bounds of the interval. $l_P(x)$ and $u_P(x)$ are monotone with respect to the inclusion relation between condition attribute sets. Namely, for $Q, P \subseteq C$ and $x \in U$, we have

$$Q \subseteq P \Rightarrow l_Q(x) \leq l_P(x),\ u_Q(x) \geq u_P(x). \tag{19}$$

Using $\delta_P(x)$, the lower and upper approximations are represented as

$$\underline{P}(Cl_t^{\geq}) = \{x \in U \mid l_P(x) \geq t\}, \quad \underline{P}(Cl_t^{\leq}) = \{x \in U \mid u_P(x) \leq t\}, \tag{20}$$

$$\overline{P}(Cl_t^{\geq}) = \{x \in U \mid u_P(x) \geq t\}, \quad \overline{P}(Cl_t^{\leq}) = \{x \in U \mid l_P(x) \leq t\}. \tag{21}$$

2.4 Previous Reducts in DRSA

Attribute reduction is one of major topics in the rough set approach. By the method, superfluous attributes are removed so that we may find important attributes as a set of attributes called a *reduct*.

In DRSA, a few approaches to attribute reduction have been already proposed. Susmaga et al. [8] proposed the reduct preserving the quality of sorting $\gamma_P(\mathcal{C})$, where for $P \subseteq C$, $\gamma_P(\mathcal{C})$ is defined by

$$\gamma_P(\mathcal{C}) = \frac{|U - \bigcup_{t \in T} Bn_P(Cl_t^{\leq})|}{|U|} = \frac{|U - \bigcup_{t \in T} Bn_P(Cl_t^{\geq})|}{|U|}. \tag{22}$$

In this paper, we call this reduct a *Q-reduct*. Yang et al. [9] proposed four kinds of reducts for an incomplete decision table with dominance relations. They are reducts preserving lower/upper approximations of upward/downward unions. Inuiguchi and Yoshioka [5] proposed several kinds of reducts and investigated their relations. They are reducts preserving only lower approximations, only upper approximations, both lower and upper approximations and boundary regions of upward/downward unions. Inuiguchi and Yoshioka showed that they are only three different kinds. Four kinds of reducts by Yang et al. [9] are same as four kinds of reducts by Inuiguchi and Yoshioka [5]. Since those reducts are based on upward and downward unions, they are called *union-based reducts* [5].

Let us show the definitions of the previously proposed reducts.

Definition 1. *(Q-reduct) A set $P \subseteq C$ is called a Q-reduct if and only if*

(Q1) $\gamma_P(\mathcal{C}) = \gamma_C(\mathcal{C})$ *and*
(Q2) $\nexists Q \subset P$ *such that* $\gamma_Q(\mathcal{C}) = \gamma_P(\mathcal{C})$.

Definition 2. *($L^{\geq}$-reduct) A set $P \subseteq C$ is called an $L^{\geq}$-reduct if and only if*

(L1$^{\geq}$) $\underline{P}(Cl_t^{\geq}) = \underline{C}(Cl_t^{\geq})$ *for all $t \in T$, and*
(L2$^{\geq}$) $\nexists Q \subset P$ *such that* $\underline{Q}(Cl_t^{\geq}) = \underline{P}(Cl_t^{\geq})$ *for all $t \in T$.*

Definition 3. *($L^{\leq}$-reduct) A set $P \subseteq C$ is called an $L^{\leq}$-reduct if and only if*

(L1$^{\leq}$) $\underline{P}(Cl_t^{\leq}) = \underline{C}(Cl_t^{\leq})$ *for all $t \in T$, and*
(L2$^{\leq}$) $\nexists Q \subset P$ *such that* $\underline{Q}(Cl_t^{\leq}) = \underline{P}(Cl_t^{\leq})$ *for all $t \in T$.*

Definition 4. *($L^\diamond$-reduct) A set $P \subseteq C$ is called an $L^\diamond$-reduct if and only if*

(L1$^\diamond$) $\underline{P}(Cl_t^{\geq}) = \underline{C}(Cl_t^{\geq})$, $\underline{P}(Cl_t^{\leq}) = \underline{C}(Cl_t^{\leq})$ *for all* $t \in T$, *and*
(L2$^\diamond$) $\nexists Q \subset P$ *such that* $\underline{Q}(Cl_t^{\geq}) = \underline{P}(Cl_t^{\geq})$, $\underline{Q}(Cl_t^{\leq}) = \underline{P}(Cl_t^{\leq})$ *for all* $t \in T$.

As shown in Inuiguchi and Yoshioka [5], we have

If P is an L$^\diamond$-reduct then P satisfies (L1$^{\geq}$) and (L1$^{\leq}$).

3 Class-Based Reducts in DRSA

3.1 Approximations of Decision Classes

In this section, we propose a few new concepts of reducts, called *class-based reducts*. Before giving the definitions, we need to define lower and upper approximations and boundary regions of decision classes. For $P \subseteq C$ and $t \in T$, lower and upper approximations of Cl_t and the boundary region of Cl_t are defined by

$$\underline{P}(Cl_t) = \underline{P}(Cl_t^{\geq}) \cap \underline{P}(Cl_t^{\leq}), \tag{23}$$
$$\overline{P}(Cl_t) = \overline{P}(Cl_t^{\geq}) \cap \overline{P}(Cl_t^{\leq}), \tag{24}$$
$$Bn_P(Cl_t) = \overline{P}(Cl_t) - \underline{P}(Cl_t). \tag{25}$$

The properties of those approximations are given in the following theorem.

Theorem 1. *For $P \subseteq C$ and $t \in T$, we have*

$$\underline{P}(Cl_t) = \{x \in U \mid l_P(x) = u_P(x) = t\}, \tag{26}$$
$$\overline{P}(Cl_t) = \{x \in U \mid l_P(x) \leq t \leq u_P(x)\}, \tag{27}$$
$$Bn_P(Cl_t) = \{x \in U \mid l_P(x) \leq t \leq u_P(x), l_P(x) < u_P(x)\}, \tag{28}$$
$$\overline{P}(Cl_t^{\geq}) = \bigcup_{k \geq t, k \in T} \overline{P}(Cl_k),\ \overline{P}(Cl_t^{\leq}) = \bigcup_{k \leq t, k \in T} \overline{P}(Cl_k), \tag{29}$$
$$\underline{P}(Cl_t) \subseteq Cl_t \subseteq \overline{P}(Cl_t), \tag{30}$$
$$\overline{P}(Cl_t) = Bn_P(Cl_t) \cup Cl_t, \tag{31}$$
$$\underline{P}(Cl_t) = Cl_t - Bn_P(Cl_t), \tag{32}$$
$$\underline{P}(Cl_t) = U - \bigcup_{k \neq t, k \in T} \overline{P}(Cl_k), \tag{33}$$
$$U - \bigcup_{k \in T} \underline{P}(Cl_k) = \bigcup_{k \in T} Bn_P(Cl_k), \tag{34}$$
$$Bn_P(Cl_t) = Bn_P(Cl_t^{\geq}) \cup Bn_P(Cl_t^{\leq}), \tag{35}$$
$$Bn_P(Cl_t) = \overline{P}(Cl_t) \cap \bigcup_{k \neq t, k \in T} \overline{P}(Cl_k). \tag{36}$$

Proof. We prove (26), (29), (33), (34) , (35) and (36) only. The others can be shown in the similar manner or straightforwardly.

First we prove (26). Suppose $x \in \{x \in U \mid l_P(x) = u_P(x) = t\}$. By the definition of $l_P(x)$, $l_P(x) = t$ implies $D_P^+(x) \cap Cl_{t-1}^{\leq} = \emptyset$. Since $Cl_t^{\geq} = U - Cl_{t-1}^{\leq}$, we have $D_P^+(x) \subseteq Cl_t^{\geq}$. Then $x \in \underline{P}(Cl_t^{\geq})$. In the same way, we obtain $u_P(x) = t$ implies $x \in \underline{P}(Cl_t^{\leq})$. Therefore, $x \in \underline{P}(Cl_t)$. On the other hand, suppose $x \in \underline{P}(Cl_t)$. Then we have $D_P^+(x) \subseteq Cl_t^{\geq}$ and $D_P^-(x) \subseteq Cl_t^{\leq}$. Because $x \in D_P^+(x)$ and $x \in D_P^-(x)$, we obtain $x \in D_P^+(x) \cap D_P^-(x) \subseteq Cl_t^{\geq} \cap Cl_t^{\leq} = Cl_t$. Moreover, $D_P^+(x) \subseteq Cl_t^{\geq}$ is equivalent to $D_P^+(x) \cap Cl_{t-1}^{\leq} = \emptyset$. Combining this with $x \in D_P^+(x)$ and $x \in Cl_t$, we obtain $l(x) = t$. Similarly, from $D_P^-(x) \subseteq Cl_t^{\leq}$, we obtain $u(x) = t$. Therefore, $x \in \{x \in U \mid l_P(x) = u_P(x) = t\}$.

We prove the second part of (29) by induction. When $t = 1$, $\overline{P}(Cl_1^{\leq}) = \overline{P}(Cl_1)$ holds. Suppose $\overline{P}(Cl_s^{\leq}) = \bigcup_{k \leq s, k \in T} \overline{P}(Cl_k)$ holds when $t = s < n$, then

$$\bigcup_{k \leq s+1} \overline{P}(Cl_k) = \overline{P}(Cl_s^{\leq}) \cup \overline{P}(Cl_{s+1}) = \overline{P}(Cl_s^{\leq}) \cup (\overline{P}(Cl_{s+1}^{\geq}) \cap \overline{P}(Cl_{s+1}^{\leq}))$$
$$= (\overline{P}(Cl_s^{\leq}) \cup \overline{P}(Cl_{s+1}^{\geq})) \cap (\overline{P}(Cl_s^{\leq}) \cup \overline{P}(Cl_{s+1}^{\leq})).$$

By (9) and (14), $\bigcup_{k \leq s+1} \overline{P}(Cl_k) = \overline{P}(Cl_{s+1}^{\leq})$ holds. Therefore, we have proved the second part of (29). The first part can be shown in the same way.

We prove (33). Applying De Morgan's law and (29) to the definition, we obtain $\underline{P}(Cl_t) = \underline{P}(Cl_t^{\leq}) \cap \underline{P}(Cl_t^{\geq}) = U - \overline{P}(Cl_{t+1}^{\geq}) \cup \overline{P}(Cl_{t-1}^{\leq}) = U - \bigcup_{k \neq t} \overline{P}(Cl_k)$.

We prove (34). Applying (9), De Morgan's law, we obtain

$$U - \bigcup_{k \in T} \underline{P}(Cl_k) = \bigcup_{k \in T} \overline{P}(Cl_k) - \bigcup_{k \in T} \underline{P}(Cl_k)$$
$$= \bigcup_{k \in T} \overline{P}(Cl_k) \cap \bigcap_{k \in T} (U - \underline{P}(Cl_k)) = \bigcup_{k \in T} (\overline{P}(Cl_k) \cap \bigcap_{l \in T} (U - \underline{P}(Cl_l))).$$

By (33), we obtain $\overline{P}(Cl_k) \subseteq U - \underline{P}(Cl_l)$ for $l \neq k$. Therefore, $U - \bigcup_{k \in T} \underline{P}(Cl_k) = \bigcup_{k \in T} (\overline{P}(Cl_k) \cap (U - \underline{P}(Cl_k))) = \bigcup_{k \in T} (\overline{P}(Cl_k) - \underline{P}(Cl_k)) = \bigcup_{k \in T} Bn_P(Cl_k)$.

We prove (35). By (20) and (21), $Bn_P(Cl_t^{\geq}) = \{x \in U \mid l_p(x) < t \leq u_P(x)\}$ and $Bn_P(Cl_t^{\leq}) = \{x \in U \mid l_p(x) \leq t < u_P(x)\}$. Then, from (28), $Bn_P(Cl_t^{\geq}) \cup Bn_P(Cl_t^{\leq}) = \{x \in U \mid l_p(x) \leq t \leq u_P(x),\ l_p(x) < u_P(x)\} = Bn_P(Cl_t)$.

Finally, we prove (36). By (33), $Bn_P(Cl_t) = \overline{P}(Cl_t) - \underline{P}(Cl_t) = \overline{P}(Cl_t) - (U - \bigcup_{k \neq t, k \in T} \overline{P}(Cl_k)) = \overline{P}(Cl_t) \cap \bigcup_{k \neq t, k \in T} \overline{P}(Cl_k)$. □

Moreover, the approximations are also monotone with respect to the inclusion relation between condition attribute sets. Let $P, Q \subseteq C$ and $t \in T$, we have

$$Q \subseteq P \Rightarrow \underline{Q}(Cl_t) \subseteq \underline{P}(Cl_t),\ \overline{Q}(Cl_t) \supseteq \overline{P}(Cl_t). \tag{37}$$

3.2 Class-Based Reducts

Now, we are ready to define new kinds of reducts. The first kind of reducts, called *L-reduct*, preserves the lower approximations of decision classes, the second kind

reduct, called *U-reduct*, preserves the upper approximations of decision classes, and the third kind of reduct, called *B-reduct*, preserves the boundary regions of decision classes. They are defined formally as follows.

Definition 5. *(L-reduct) A set $P \subseteq C$ is called an L-reduct if and only if*

(L1) $\underline{P}(Cl_t) = \underline{C}(Cl_t)$ *for all* $t \in T$, *and*
(L2) $\nexists Q \subset P$ *such that* $\underline{Q}(Cl_t) = \underline{P}(Cl_t)$ *for all* $t \in T$.

Definition 6. *(U-reduct) A set $P \subseteq C$ is called a U-reduct if and only if*

(U1) $\overline{P}(Cl_t) = \overline{C}(Cl_t)$ *for all* $t \in T$, *and*
(U2) $\nexists Q \subset P$ *such that* $\overline{Q}(Cl_t) = \overline{P}(Cl_t)$ *for all* $t \in T$.

Definition 7. *(B-reduct) A set $P \subseteq C$ is called a B-reduct if and only if*

(B1) $Bn_P(Cl_t) = Bn_C(Cl_t)$ *for all* $t \in T$, *and*
(B2) $\nexists Q \subset P$ *such that* $Bn_Q(Cl_t) = Bn_P(Cl_t)$ *for all* $t \in T$.

Those concepts are parallel to L-, U- and B-reducts [4] discussed in the setting of the classical rough sets.

From the properties of approximations, we have the following theorem which is proved easily.

Theorem 2. *We have the following assertions:*

(a) If P is a U-reduct then P satisfies (L1).
(b) P is a U-reduct if and only if P is a B-reduct.

Consequently, we have only two kinds of class-based reducts: L-reduct and U-reduct (or B-reduct). This result is parallel to the result in the classical rough sets [4].

Let us discuss relations of the proposed class-based reducts and the previous reducts introduced in Section 2.4. We have the following theorems.

Theorem 3. *P is an L-reduct if and only if P is a Q-reduct.*

Proof. By (10) and (35), we have $Bn_P(Cl_t) = Bn_P(Cl_{t-1}^{\leq}) \cup Bn_P(Cl_t^{\leq})$. It implies $\bigcup_{t\in T} Bn_P(Cl_t) = \bigcup_{t\in T} Bn_P(Cl_t^{\leq})$. From (34), the union of the lower approximations $\bigcup_{t\in T} \underline{P}(Cl_t)$ equals to the quality of sorting $\gamma_P(\mathcal{C})$. So it suffices to show $\left|\bigcup_{t\in T} \underline{P}(Cl_t)\right| = \left|\bigcup_{t\in T} \underline{C}(Cl_t)\right|$ if and only if $\forall t \in T, \underline{P}(Cl_t) = \underline{C}(Cl_t)$. From (30), $\underline{Q}(Cl_t)$ and $\underline{Q}(Cl_s)$ are disjoint for any $t, s \in T$ such that $t \neq s$ and for any $Q \subseteq C$. Then, we have $\left|\bigcup_{t\in T} \underline{Q}(Cl_t)\right| = \sum_{t\in T} |\underline{Q}(Cl_t)|$ for $Q = P, C$. From this and (37), $\left|\bigcup_{t\in T} \underline{P}(Cl_t)\right| = \left|\bigcup_{t\in T} \underline{C}(Cl_t)\right|$ if and only if $\forall t \in T, |\underline{P}(Cl_t)| = |\underline{C}(Cl_t)|$ which is equivalent to $\forall t \in T, \underline{P}(Cl_t) = \underline{C}(Cl_t)$. □

Theorem 4. *P is a U-reduct if and only if P is an $L^{\diamond}$-reduct.*

Proof. This is easily obtained from (8), (24) and (29). □

As a result, all kinds of reducts proposed in DRSA are arranged in Figure 1.

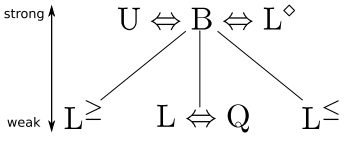


Fig. 1. Relations of reducts in DRSA

4 A Unified Approach to Discernibility Matrices

4.1 Reducts Based on Generalized Decisions

A discernibility matrix is a popular approach to enumeration of all reducts in the rough set approach. For $L^{\geq}$-, $L^{\leq}$- and $L^{\diamond}$-reducts, suitable discernibility matrices have successfully proposed in [5,9]. For Q-reduct, a similar approach to a discernibility matrix has been proposed in [8]. Therefore, all reducts in DRSA can be enumerated by those previous approaches.

In this section, we propose alternative discernibility matrices for enumerations of all reducts. In the proposed approach, calculations of lower and upper approximations of all downward and upward unions are not required but those of generalized decisions of all objects. The latter would require less computation effort than the former. Moreover, we can treat all kinds of reducts $\forall y \in U, (l_C(y) > l_C(x)$ implies $\exists q \in P, x \not\succeq_q y)$ comprehensively in the proposed approach.

We introduce some kinds of reducts based on generalized decisions.

Definition 8. *(δ-reduct) A set $P \subseteq C$ is called a δ-reduct if and only if*

($\delta 1$) $\forall x \in U, \delta_P(x) = \delta_C(x)$ *and*
($\delta 2$) $\nexists Q \subset P$ *such that* $\forall x \in U, \delta_Q(x) = \delta_P(x)$.

Definition 9. *(Lδ-reduct) A set $P \subseteq C$ is called an Lδ-reduct if and only if*

(L$\delta 1$) $\forall x \in U$, $l_C(x) = u_C(x)$ *implies* $\delta_P(x) = \delta_C(x)$ *and*
(L$\delta 2$) $\nexists Q \subset P$ *such that* $\forall x \in U$, $l_C(x) = u_C(x)$ *implies* $\delta_Q(x) = \delta_P(x)$.

Definition 10. *(l-reduct) A set $P \subseteq C$ is called an l-reduct if and only if*

($l1$) $\forall x \in U$, $l_P(x) = l_C(x)$ *and*
($l2$) $\nexists Q \subset P$ *such that* $\forall x \in U$, $l_Q(x) = l_P(x)$.

Definition 11. *(u-reduct) A set $P \subseteq C$ is called a u-reduct if and only if*

($u1$) $\forall x \in U$, $u_P(x) = u_C(x)$ *and*
($u2$) $\nexists Q \subset P$ *such that* $\forall x \in U$, $u_Q(x) = u_P(x)$.

We obtain the following theorem.

Theorem 5. *We have the following assertions:*

(a) P is a δ-reduct if and only if P is a U-reduct, i.e., an $L^{\diamond}$-reduct.
(b) P is an Lδ-reduct if and only if P is an L-reduct, i.e., a Q-reduct.
(c) P is an l-reduct if and only if P is an $L^{\geq}$-reduct.
(d) P is a u-reduct if and only if P is an $L^{\leq}$-reduct.

Proof. We prove only (b) since (a), (c) and (d) can be shown easily from (20). It suffices to prove that (Lδ1) is equivalent to (L1). From (26), (Lδ1) if and only if $\forall t \in T$, $\underline{C}(Cl_t) \subseteq \underline{P}(Cl_t))$ which is (L1). □

4.2 Discernibility Matrices

We first give an underlying theorem to define discernibility matrices.

Theorem 6. *Let $x \in U$, then we have the following equivalences:*

$$l_P(x) = l_C(x)$$
$$\text{if and only if } \ \forall y \in U, (l_C(y) < l_C(x) \text{ implies } \exists q \in P, y \not\succeq_q x), \tag{38}$$
$$u_P(x) = u_C(x)$$
$$\text{if and only if } \ \forall y \in U, (u_C(y) > u_C(x) \text{ implies } \exists q \in P, y \not\preceq_q x). \tag{39}$$

Proof. We prove only (38). The other can be shown in the same way.

First, under the supposition $\forall y \in U, (l_C(y) < l_C(x)$ implies $\exists q \in P, y \not\succeq_q x)$, we prove $l_C(x) \leq l_P(x)$ by contradiction. Assume $l_C(x) > l_P(x)$, then there exists $y \in D_P^+(x) \cap Cl_{l_P(x)}$. The fact $y \in Cl_{l_P(x)}$ and the reflexivity of $\succeq_P$ $(y \in D_P^+(y))$ implies $l_C(y) \leq l_P(x)$. From this and the assumption $l_C(x) > l_P(x)$ we have $l_C(y) < l_C(x)$. On the other hand, $y \in D_P^+(x)$ i.e., $y \succeq_P x$, i.e., $\forall q \in P, y \succeq_q x$. Facts $l_C(y) < l_C(x)$ and $y \succeq_P x$ contradict the supposition. Consequently, we have $l_C(x) \leq l_P(x)$. Moreover, from (19), $l_P(x) \leq l_C(x)$ is equivalent to $l_C(x) = l_P(x)$.

Next we prove the converse. Suppose $l_P(x) = l_C(x)$. Assume there exists $y \in U$ such that $l_C(y) < l_C(x)$ and $y \succeq_P x$. $y \succeq_P x$ implies $D_P^+(y) \subseteq D_P^+(x)$. By definition, this further implies $l_P(y) \geq l_P(x)$. Then, under the assumption, we have $l_C(y) < l_C(x) = l_P(x) \leq l_P(y)$. This contradicts (19). Then we have $l_C(y) < l_C(x)$ implies $\exists q \in P$, $y \not\succeq_q x$ for all $y \in U$. □

Now we are ready to define discernibility matrices. The l-discernibility matrix M^l and u-discernibility matrix M^u are defined as follows.

Definition 12. *The l-discernibility matrix $M^l = (m_{ij}^l)_{i,j=1,\ldots,|U|}$ is composed of (i,j)-components m_{ij}^l defined by*

$$m_{ij}^l = \begin{cases} \{q \in C \mid x_i \not\succeq_q x_j\} & \text{if } l_C(x_i) < l_C(x_j) \\ C & \text{otherwise} \end{cases} \tag{40}$$

On the other hand, the u-discernibility matrix $M^u = (m_{ij}^u)_{i,j=1,\ldots,|U|}$ is composed of (i,j)-components m_{ij}^u defined by

$$m_{ij}^u = \begin{cases} \{q \in C \mid x_i \not\preceq_q x_j\} & \text{if } u_C(x_i) > u_C(x_j) \\ C & \text{otherwise} \end{cases} \tag{41}$$

Let $\tilde{q}_i^P$ be a Boolean variable corresponding to a condition attribute q_i and a condition attribute set P defined by

$$\tilde{q}_i^P = \begin{cases} \text{true} & \text{if } q_i \in P, \\ \text{false} & \text{otherwise.} \end{cases} \tag{42}$$

Then, based on M^l and M^u, we consider the following Boolean functions $F^{\geq}$, $F^{\leq}$, F^{U} and F^{L}:

$$F^{\geq}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \bigwedge_{1 \leq i,j \leq |U|} \bigvee_{q \in m_{ij}^l} \tilde{q}^P, \tag{43}$$

$$F^{\leq}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \bigwedge_{1 \leq i,j \leq |U|} \bigvee_{q \in m_{ij}^u} \tilde{q}^P, \tag{44}$$

$$F^{\mathrm{U}}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \bigwedge_{1 \leq i,j \leq |U|} \bigvee_{q \in m_{ij}^l} \tilde{q}^P \wedge \bigwedge_{1 \leq i,j \leq |U|} \bigvee_{q \in m_{ij}^u} \tilde{q}^P, \tag{45}$$

$$F^{\mathrm{L}}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \bigwedge_{1 \leq i \leq |U|} \bigwedge_{j: l_P(x_j) = u_P(x_j)} \bigvee_{q \in m_{ij}^l} \tilde{q}^P \wedge \bigwedge_{1 \leq i \leq |U|} \bigwedge_{j: l_P(x_j) = u_P(x_j)} \bigvee_{q \in m_{ij}^u} \tilde{q}^P. \tag{46}$$

Based on Theorem 6, we have the following theorem.

Theorem 7. *We have the following equivalences:*

$$P \subseteq C \text{ satisfies } (\mathrm{L1}^{\geq}), \text{ i.e., } (l1) \text{ if and only if } F^{\geq}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \text{true}, \tag{47}$$

$$P \subseteq C \text{ satisfies } (\mathrm{L1}^{\leq}), \text{ i.e., } (u1) \text{ if and only if } F^{\leq}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \text{true}, \tag{48}$$

$$P \subseteq C \text{ satisfies } (\mathrm{U1}), \text{ i.e., } (\mathrm{L}\delta 1) \text{ if and only if } F^{\mathrm{U}}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \text{true}, \tag{49}$$

$$P \subseteq C \text{ satisfies } (\mathrm{L1}), \text{ i.e., } (\delta 1) \text{ if and only if } F^{\mathrm{L}}(\tilde{q}_1^P, ..., \tilde{q}_{|C|}^P) = \text{true}. \tag{50}$$

Proof. First three equivalences are obvious from Theorem 6 and definitions of Boolean functions. The last one needs to show that any object x such that $l_C(x) < u_C(x)$ never satisfies $l_P(x) = u_P(x)$. This is clear from (19). □

From Theorem 7, all $\mathrm{L}^{\geq}$-, $\mathrm{L}^{\leq}$-, U- and L-reducts can be obtained as all prime implicants of Boolean functions $F^{\geq}$, $F^{\leq}$, F^{U} and F^{L}, respectively.

The proposed discernibility matrices have two advantages comparing to the previous ones. One is the computational efficiency. We need to calculate neither lower approximations, upper approximations nor boundary regions but only the lower bounds $l_C(x)$ and the upper bounds $u_C(x)$. The computational complexity for the former is at least n times of that of the latter, when we do not apply (20) and (21). The other advantage is that the all Boolean functions with respect to $\mathrm{L}^{\geq}$-reduct, $\mathrm{L}^{\leq}$-reduct, U-reduct and L-reduct are obtained from only two discernibility matrices.

Table 1. A decision table

Student	Mathematics	Physics	Literature	Evaluation	l_C	u_C
S_1	good	good	good	good	good	good
S_2	good	good	medium	medium	medium	good
S_3	medium	good	medium	good	medium	good
S_4	bad	medium	good	medium	medium	medium
S_5	medium	bad	medium	bad	bad	medium
S_6	medium	bad	bad	medium	bad	medium
S_7	bad	bad	bad	bad	bad	bad

Table 2. The discernibility matrix M^l with respect to Table 1

	S_1^*	S_2	S_3	S_4^*	S_5	S_6	S_7^*
S_1	C	C	C	C	C	C	C
S_2	$\{q_3\}$	C	C	C	C	C	C
S_3	$\{q_1, q_3\}$	C	C	C	C	C	C
S_4	$\{q_1, q_2\}$	C	C	C	C	C	C
S_5	C	$\{q_1, q_2\}$	$\{q_2\}$	$\{q_2, q_3\}$	C	C	C
S_6	C	C	$\{q_2, q_3\}$	$\{q_2, q_3\}$	C	C	C
S_7	C	C	C	$\{q_2, q_3\}$	C	C	C

Table 3. The discernibility matrix M^u with respect to Table 1

	S_1^*	S_2	S_3	S_4^*	S_5	S_6	S_7^*
S_1	C	C	C	$\{q_1, q_2\}$	C	C	C
S_2	C	C	C	$\{q_1, q_2\}$	$\{q_1, q_2\}$	C	C
S_3	C	C	C	$\{q_1, q_2\}$	$\{q_2\}$	$\{q_2, q_3\}$	C
S_4	C	C	C	C	C	C	$\{q_2, q_3\}$
S_5	C	C	C	C	C	C	$\{q_1, q_3\}$
S_6	C	C	C	C	C	C	$\{q_1\}$
S_7	C	C	C	C	C	C	C

Example 1. Consider a decision table given in Table 1. This table shows student evaluation in a school. Objects are seven students, i.e., $U = \{S_1, S_2, \ldots, S_7\}$. Condition attributes are scores of mathematics (q_1), physics (q_2) and literature (q_3), while decision attribute (d) is a comprehensive evaluation. Namely, $C = \{q_1, q_2, q_3\}$. We may assume that the better scores in all subjects student takes, the better comprehensive evaluation he/she gets. The lower bounds l_C and the upper bounds u_C are shown in the rightmost two columns of Table 1.

Discernibility matrices M^l and M^u are obtained as in Table 2 and 3, respectively. The columns with asterisk shows objects belonging to one of lower approximations of decision classes. Then Boolean functions $F^{\geq}$, $F^{\leq}$, F^{U} and F^{L} are obtained as

$$F^{\geq}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = \tilde{q}_2 \wedge \tilde{q}_3 \wedge (\tilde{q}_1 \vee \tilde{q}_2) \wedge (\tilde{q}_1 \vee \tilde{q}_3) \wedge (\tilde{q}_2 \vee \tilde{q}_3) = \tilde{q}_2 \wedge \tilde{q}_3. \quad (51)$$

$$F^{\leq}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = \tilde{q}_1 \wedge \tilde{q}_2 \wedge (\tilde{q}_1 \vee \tilde{q}_2) \wedge (\tilde{q}_1 \vee \tilde{q}_3) \wedge (\tilde{q}_2 \vee \tilde{q}_3) = \tilde{q}_1 \wedge \tilde{q}_2, \quad (52)$$

$$F^{\mathrm{U}}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = F^{\geq}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) \wedge F^{\leq}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = \tilde{q}_1 \wedge \tilde{q}_2 \wedge \tilde{q}_3 \quad (53)$$

$$F^{\mathrm{L}}(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = (\tilde{q}_3 \wedge (\tilde{q}_1 \vee \tilde{q}_2) \wedge (\tilde{q}_1 \vee \tilde{q}_3) \wedge (\tilde{q}_2 \vee \tilde{q}_3)) \\ \wedge(\tilde{q}_1 \wedge (\tilde{q}_1 \vee \tilde{q}_2) \wedge (\tilde{q}_1 \vee \tilde{q}_3) \wedge (\tilde{q}_2 \vee \tilde{q}_3)) = \tilde{q}_1 \wedge \tilde{q}_3 \qquad (54)$$

Consequently, we obtain $\{q_2, q_3\}$ as a unique $\mathrm{L}^{\geq}$-reduct, $\{q_1, q_2\}$ as a unique $\mathrm{L}^{\leq}$-reduct, $C = \{q_1, q_2, q_3\}$ as a unique U-reduct and $\{q_1, q_3\}$ as a unique L-reduct. As exemplified in this example, $\mathrm{L}^{\geq}$-reduct, $\mathrm{L}^{\leq}$-reduct, U-reduct and L-reduct can be different.

5 Conclusions

In this paper, we have investigated attribute reduction in DRSA. We introduce class-based reducts and show relations with previous reducts. Moreover, we show that all kinds of reducts can be enumerated comprehensively based on two discernibility matrices associated with generalized decisions. The proposed approach to the enumeration of all reducts can be computationally efficient.

The investigations about the class-based reducts in VC-DRSA as well as VP-DRSA will be a next step of this research.

Acknowledgments

This work has been partially supported by the Grant-in-Aid for Scientific Research (B) No.17310098.

References

1. Dembczyński, K., Greco, S.: Quality of rough approximation in multi-criteria classification problems. In: Greco, S., et al. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 318–327. Springer, Heidelberg (2006)
2. Greco, S., Matarazzo, B., Slowinski, R.: Rough set theory for multicriteria decision analysis. European Journal of Operational Research 129(1), 1–47 (2001)
3. Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.): RSCTC 2006. LNCS (LNAI), vol. 4259. Springer, Heidelberg (2006)
4. Inuiguchi, M., Tsurumi, M.: Measures based on upper approximations of rough sets for analysis of attribute importance and interaction. International Journal of Innovative Computing, Information and Control 2(1), 1–12 (2006)
5. Inuiguchi, M., Yoshioka, Y.: Several reducts in dominance-based rough set approach. In: Huynh, V.-N., et al. (eds.) Interval/Probabilistic Uncertainty and Non-classical Logics, ASC46, pp. 163–175 (2008)
6. Pawlak, Z.: Rough sets. International Journal of Information and Computer Sciences 11(5), 341–356 (1982)
7. Słowiński, R. (ed.): Intelligent Decision Support. Kluwer Academic Publishers, Dordrecht (1992)
8. Susmaga, R., Słowiński, R., Greco, S., Matarazzo, B.: Generation of reducts and rules in multi-attribute and multi-criteria classification. Control and Cybernetics 29(4), 969–988 (2000)
9. Yang, X., Yang, J., Wu, C., Yu, D.: Dominance-based rough set approach and knowledge reductions in incomplete ordered information system. Information Sciences 178(4), 1219–1234 (2008)

Graph-Based Active Learning Based on Label Propagation

Jun Long, Jianping Yin, Wentao Zhao, and En Zhu

National University of Defense Technology, Changsha, Hunan 410073, China
jdragon_nudt@hotmail.com

Abstract. By only selecting the most informative instances for labeling, active learning could reduce the labeling cost when labeled instances are hard to obtain. Facing the same situation, semi-supervised learning utilize unlabeled instances to strengthen classifiers trained on labeled instances under suitable assumptions. However, the current active learning methods often ignore such effect. Combining semi-supervised learning, we propose a graph-based active learning method, which can also handle multi-class problems, in the entropy reduction framework. The proposed method trains the base classifier using a popular graph-based semi-supervised label propagation method and samples the instance with the largest expected entropy reduction for labeling. The experiments show that the proposed method outperforms the traditional sampling methods on selected datasets.

Keywords: active learning, semi-supervised learning, label propagation.

1 Introduction

In passive learning problems, a previously labeled set of instances is available for training. However, the process of labeling instances may be expensive or time-consuming in many real world applications. For example, in gene expression analysis, labeling data may require very expensive tests therefore only a small set of labeled data may be available. To reduce the labeling cost in learning problems, active learning methods, which only sample the most informative instances for labeling, were proposed.

According to the source of unlabeled instances, active learning methods can be divided into two types: pool-based and stream-based. In this paper, we focus on pool-based active learning in which the active learner can go through the entire set of unlabeled instances and select the most informative one or ones for labeling. In general, a pool-based active learning method comprises two parts: a learning engine and a sampling engine [1]. The typical process of pool-based active learning methods can be described as follows. Initially, a small training set of labeled instances and an unlabeled set are available. Then, the learning engine trains a base classifier on the original training set. After that, the sampling engine chooses the most informative instance x from the unlabeled instances and then labels x by human experts before $< x, c(x) >$ is added into the labeled set where

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 179–190, 2008.

$c(x)$ is the label of x. Then the learning engine constructs a new classifier on the updated labeled set. The whole process runs repeatedly until the evaluation index of the classifier or iteration times reaches the preset value.

Sampling criterion is the key problem in active learning. Depending on the criterion used to select instances for labeling, the current research falls under several categories: uncertainty reduction, expected-error minimization and version space reduction [2]. The uncertainty reduction approach [3] selects the instances on which the current base classifier has the least certainty. Many sampling methods apply the similar strategy [4,5]. They perform better than random sampling in most tasks, but sometimes they may select outliers. The expected-error minimization approach [6,7,8] samples the instances that minimize the future expected error rate on the test set. Such methods expect to achieve the lowest error, but they are computationally expensive and the performance heavily depends on the chosen lost function. The version space reduction approach [9,10] tries to select the instances that can reduce the volume of the version space by half. Query-by-Committee is a representative method of this approach that constructs a committee consisting of randomly selected hypotheses from the version space and selects the instances on which the disagreement within the committee is the greatest. The version space reduction approach also includes QBag [11], QBoost [11], Active DECORATE [12] and CBMPMS [13].

Most active learning methods sample instances just based on the models built on labeled instances and totally ignore the effect of unlabeled instances. However, unlabeled instances could strengthen supervised learning tasks under suitable assumptions. A variety of semi-supervised learning methods were proposed based on this idea.

Semi-supervised learning methods can be used to strengthen active learners. Some researchers have made their contributions based on such idea. McCallum and Nigam [14] presented an active learning method which constructed the base classifier on labeled and unlabeled instances using the EM algorithm for text categorization. It is efficient in text categorization, but only suitable for the generative model. Muslea [2] proposed the multi-view active learning method which selected the instances with the largest disagreement from multi-view learners. However, it requires the learning task to be a multi-view one. Zhu [8] presented a method combining active learning and semi-supervised learning using Gaussian Fields and Harmonic Functions. Belonging to the expected-error minimization approach, it tends to achieve the lowest error. However, it can only handle 2-class classification problems and its performance depends on the function used to estimate the risk.

In recent years, graph-based methods is popular in semi-supervised learning due to clear mathematical framework and strong performance with suitable models. However, few study concentrates on graph-based methods in active learning. In this paper, we propose a graph-based active learning method called GAL, which can also handle multi-class problems, in the entropy reduction framework. The proposed method trains a classifier using a popular graph-based semi-supervised label propagation method and samples the instance with the

largest expected entropy reduction for labeling. The experiments show that the proposed method obtains smaller data utilization and average deficiency than other popular active learners on selected datasets from semi-supervised learning benchmarks.

The rest of the paper is organized as follows. Section 2 provides the basic notations and related work in semi-supervised learning. Section 3 presents the entropy reduction framework for graph-based active learning. Section 4 describes the proposed graph-based active learning called GAL in detail. Section 5 shows the experimental results of the GAL method as well as other methods on selected data sets. Section 6 draws the conclusion.

2 Preliminaries

2.1 Basic Notation

The instance space X is a nonempty set containing several instances. Each instance x_i is a feature vector $< x_{i1}, x_{i2}, \cdots, x_{im} >$. Let $C = \{y_1, y_2, \cdots, y_p\}$ be the set of possible labels.

The target function c to be learned is a function $c : X \rightarrow C$ that classifies any $x \in X$ as a member of C. C has p elements. The notion $< x, c(x) >$ denotes a labeled instance and $< x, ? >$ denotes an unlabeled instance where $? \in C$. L denotes the whole set of labeled instances and U denotes the whole set of unlabeled instances.

There are l labeled instances: $< x_1, y_1 >, \cdots, < x_l, y_l >$, and u unlabeled instances:$< x_{l+1}, ? >, \cdots, < x_{l+u}, ? >$. We have $l \ll u$. The total number of instances is $n = l + u$.

Let $G =< V, E >$ be a connected graph with vertices V and edges E. Each vertex $v_i \in V$ represents an instance in $L \bigcup U$. Each edge $< v_j, v_k > \in E$ connects two vertices v_j and v_k. We have the adjacency matrix W_{nn} of G and its entry w_{ij} is the similarity between v_i and v_j. Here w_{ij} is given by Gaussian kernel of width σ:

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{1}$$

We choose a method proposed in [15] to learn the scale parameter σ.

Furthermore, we define a diagonal matrix D, in which $D_{ii} = \sum_{j=1}^{n} w_{ij}$.

Let Y_{mp} be a $m \times p$ matrix on G, and y_{ik} in Y_{mp} is the probability for instance x_i to be labeled as $y \in C$. If x_i is labeled as y_k, then $y_{ik} = 1$. And if x_i is unlabeled, then $y_{ik} = 1/p, k = 1, \cdots, p$.

2.2 Semi-supervised Learning

Similar to active learning, few labeled instances are available and a large number of unlabeled instances are provided in semi-supervised learning. Differently, semi-supervised learning aims to build better classifiers using both labeled and unlabeled instances and can not select instances for labeling.

Some assumptions should be satisfied for semi-supervised learning, including [16]: (1)*If two points x_1, x_2 are close, then so should be the corresponding outputs y_1, y_2.* (2)*If two points are in the same structure (a cluster or a manifold), then they are likely to have the same labels.* They can be called the cluster assumptions, which make sense in many real world applications. Based on these assumptions, labels of many unlabeled instances can be predicted by nearby labeled instances with high certainty. Thus the label uncertainty on all the instances can be significantly reduced.

Various semi-supervised learning methods were proposed according to different models for realizing the cluster assumptions[17]: co-training, generative models, graph-based learning, semi-supervised vector machines and so on.

Graph-based semi-supervised learning methods are most popular in recent years. In these methods, instances are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances. The cluster assumptions can be implemented in an easy way in graph-based methods. Typical graph-based semi-supervised learning methods include Mincuts [18], harmonic functions [15], label propagation [19], and manifold regularization [20].

3 A Framework of Entropy Reduction for Graph-Based Active Learning

Let $H(x_i)$ be the entropy of label probability distribution of instance x_i. Then let $H(G)$ denote the sum of $H(x)$ on all instances. Then

$$H(G) = \sum_{x_i \in L \bigcup U} \sum_{k=1}^{p} (-y_{ik} \log y_{ik}) \tag{2}$$

Thus, $H(G)$ reflects the certainty of the label probability distribution on $L \bigcup U$. Smaller $H(G)$ indicates more certain labels on G. If all instances are labeled, $H(G)$ equals to 0. Thus, minimizing $H(G)$ can be viewed as the goal of the learning tasks.

In active learning, when some instances are selected for labeling, $H(x)$ of those instances are changed from some positive value to 0. However, since obtaining labels on instances requires heavy cost in active learning, we can not label all the instances because of cost constraint.

When holding the clustering assumptions, the labels on labeled instances can be used to predict those of neighboring instances. Thus, labels propagate from labeled instances to unlabeled instances. Then, the uncertainty on labels decreases, so does $H(G)$.

Therefore, label propagation can be used to strengthen active learning under the clustering assumptions.

4 The Graph-Based Active Learning Method

Under the clustering assumptions, semi-supervised learning and active learning can be combined to construct stronger active learners using unlabeled instances.

We present a graph-based active learning method in which both learning engine and sampling engine are modified to employ semi-supervised learning. In the learning engine, a popular graph-based label propagation method is used to generate the base classifier. And in the sampling engine, the sampling criterion which tends to sample the instance with the largest expected entropy reduction is proposed.

4.1 Learning Engine in GAL

To strengthen the learning engine with unlabeled instances, we choose a popular graph-based label propagation method proposed by Zhou [16] to generate the base classifier.

The label propagation method is introduced in Algorithm 1.

Algorithm 1. the label propagation method proposed by Zhou [16]

1.Construct the matrix W_{ij} and $W_{ii} = 0$;
2.Construct the matrix $S = D^{-1/2}WD^{-1/2}$;
3.Iterate $F(t+1) = \alpha SF(t) + (1-\alpha)Y$ until convergence, where α is a parameter in (0,1);
4.Let F^* denote the limit of the sequence $F(t)$. Label each point x_i as a label $y_i = \arg\max_{j \le c} F^*_{ij}$.

F is a vectorial function $F : X \to C$ which assigns a vector F_i which denotes the predicted class probability distribution on each instances x_i. Zhou [16] proves that the sequence $\{F(t)\}$ converges to $F^* = (1-\alpha)(I - \alpha S)^{-1}Y$.

When a new instance x should be predicted by the label propagation method, we simply use the following function to calculate the label of x [21]:

$$\hat{y} = \frac{\sum_j W_X(x, x_j)\hat{y}_j}{\sum_j W_X(x, x_j)} \tag{3}$$

where W_X is the Gaussian kernel function:

$$W_X(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{4}$$

We take Eq.(3) as the base classifier in the GAL method.

The modified learning engine executes label propagation on unlabeled instances in iterations. Thus more entropy will be reduced than the method employed by the traditional learning engines.

4.2 Sampling Engine in GAL

Since the learning engine is modified to utilize label propagation on unlabeled instances, the sampling engine should be changed to cooperate with that.

Let $LP(G)$ be the label propagation operation on G. When $LP(G)$ is executed, we obtain the predicted class probability distribution $F^* = (1-\alpha)(I - \alpha S)^{-1}Y$.

Let $Label(G, v, y)$ be the operation which labels v in G as $y \in C$ and outputs the graph G with updated Y after v is labeled. When $Label(G, v_i, y_j)$ is executed, $y_{ij} \in Y_{mp}$ will become 1 and $y_{ik}, (k \neq j)$ will become 0.

Let $IG(G, v, y)$ denote the reduced entropy when v was labeled as y after label propagation. Then

$$IG(G, v, y) = H(LP(G)) - H(LP(Label(G, v, y))) \tag{5}$$

We sample the instances with the maximum expected entropy reduction. Thus, the sampling criterion is

$$ES_i = \sum_{y \in C} p(y|x_i) IG(G, v_i, y) \tag{6}$$

where $p(y|x_i)$ denotes the probability of x_i being labeled as y. According to the definition of F_i, $p(y|x_i)$ can be obtained from F_i. We sample the instances with the largest ES_i.

4.3 The Process of Graph-Based Active Learning

The process of GAL method is given in Algorithm 2.

Algorithm 2. the GAL method

Input: an initial labeled set L, an unlabeled set UL, a stopping criterion S, and an integer M which specify the number of instances sampled in each iteration.
Begin:
Construct G, W, D, Y;
repeat
1.For each instance $x_i \in UL$ compute

$$ES_i = \sum_{y \in C} p(y|x_i) Gain(G, v_i, y) \tag{7}$$

2.Select a subset A of size M from UL in which instances x_i have the largest ES_i;
3.Remove A from UL;
4.Label instances in A;
5.Add A into L;
6.Update Y;
7.Recalculate $F^* = (1-\alpha)(I - \alpha S)^{-1}Y$;
until the stopping criterion S is satisfied
End.
Output: The final F^*.

5 Experimental Results

5.1 Methodology

A series of experiments were conducted to evaluate the performance of our GAL method. Four representative active learning algorithms were tested.

- Random sampling: choosing the instance at random;
- Uncertainty sampling: choosing the instance with the largest uncertainty of prediction, as in [3];
- QBC sampling: choosing the instance that the committee members disagree with most, as in [9];
- GAL sampling(the method introduced in this paper).

The methods proposed by McCallum [14], Muslea [2] and Zhu [8] are designed for specific tasks, respectively. Thus we can not compare them in a general way.

Naive bayes was selected to be the base classifier of all other active learners. 10-fold cross-validation was used to obtain the target accuracy of the base classifier. The target accuracy is defined as the accuracy obtained by the base learning method trained on the whole dataset. All results presented were averages of ten runs. The committee size in QBC were set to 5. α was set to 0.1.

Each dataset was divided into 10 equal partitions at random and each in turn is used for testing and the remainder was used as the sampling set. Before the test started, the sampling set was divided into two parts: one is the labeled set and another is the unlabeled set. The labeled set contains only one instance selected randomly and the unlabeled set contains all the rest instances. When the test started, the active learner sampled 1 instance from the unlabeled set for labeling in each iteration. While the active learner reached the target accuracy, the test stopped.

Four datasets were chosen as the benchmarks: *g241c*, *handwritten digits*, *coil* and *secstr1500*. These datasets were from the benchmarks of *Semi-supervised learning* [21]. The reason we selected these datasets for experiments is that they can be easily obtained to compare different active learners and are widely accepted as the benchmarks for semi-supervised learning.

Some information of these datasets were given as follows:

- *g241c*: was generated to hold the cluster assumption. All instances were drawn from two unit-variance isotropic Gaussians. The label of an instance represents the Gaussian it was drawn from.
- *handwritten digits*: designed to consist of points close to a low-dimensional manifold embedded in a high-dimensional space. The instances originated from the digit '1'. Then several operations were taken on it, including translation, rotation and line thickness.
- *coil*: from the Columbia object image library, which consists images taken from different directions. 24 objects were divided into 6 classes of 4 objects each. There are 38 images in each class.

– *secstr1500*: used to predict the secondary structure of a given amino acid in a protein based on a sequence window centered around that amino acid. The dataset consisting of 83679 instances is presented to investigate how far current methods can cope with large-scale application. It is too large for PC to handle it, thus we choose a subset which includes 1500 instances from the whole dataset.

Table 1 shows the basic properties of these datasets.

Table 1. Basic properties of the datasets

Data set	Classes	Dimension	Instances	Comment
g241c	2	241	1500	artificial
digit1	2	241	1500	artificial
coil	6	241	1500	natural
secstr1500	2	315	1500	natural

Two metrics were used to compare the performance of different active learners: *data utilization* [12] and *average deficiency* [22].

Data utilization is defined as the number of sampling an active learner requires to reach the target accuracy. This metric reflects how efficiently the active learner can use the data. Smaller values of *data utilization* indicate more efficient active learning. Moreover, it is employed by many other researchers [12,11].

Average deficiency is used to evaluate how much an active learner could improve accuracy over random sampling. It is defined as:

$$Def_n(Active) = \frac{\sum_{t=1}^{n}(Acc_n(Ran) - Acc_t(Acv))}{\sum_{t=1}^{n}(Acc_n(Ran) - Acc_t(Ran))} \tag{8}$$

where n denotes the size of the whole unlabeled set, Acv denotes the active learner we want to evaluate, Ran denotes the random sampling method, and $Acc_t(Acv)$ denotes the average accuracy achieved by Acv after t sampling. Furthermore, the value of $Def_n(Acv)$ is always non-negative and smaller values in $[0, 1)$ indicate more efficient active learning [22].

5.2 Results

The *data utilization* and the *deficiency* of the different active learners were summarized in Table 2 and Table 3, respectively. In the head of Table 2, TA denotes target accuracy.

According to Table 2 and Table 3, it shows that our GAL method has a superior performance than other sampling methods on most datasets.

Table 2. Average data utilization of the different active learners

Data set	Random	Uncertain	QBC	GAL	TA
g241c	291	268	349	208	81.23%
digit1	74	55	110	40	95.56%
coil	288	351	172	144	64.42%
secstr1500	543	511	488	358	64.87%

Table 3. Average deficiency of the different active learners

Data set	Uncertain	QBC	GAL
g241c	0.7525	0.6096	0.5023
digit1	0.2436	0.1873	0.1033
coil	0.6690	0.5979	0.4587
secstr1500	0.7150	1.1634	0.5209

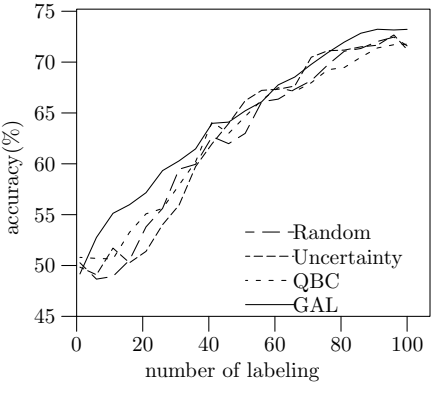


Fig. 1. learning curves on *g241c*

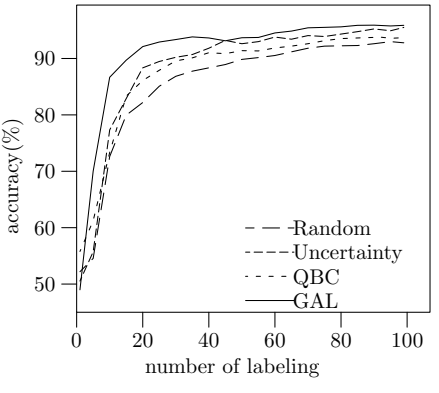


Fig. 2. learning curves on *digit1*

Figure 1-4 show the learning curves on the *g241c*, *digit1*, *coil* and *secstr1500*. In all these figures, the vertical axis shows the accuracy of the classifier and the horizontal axis shows the number of labels.

In Figure 1, all the learning curves climb substantially. The learning curve of the GAL method starts with a sharp rise and finally reaches the highest point. In Figure 2, 3 and 4, our GAL method almost outperforms the other active learners throughout the whole learning curve.

We obtain two observations in the experiments. First, our GAL method achieves significantly higher accuracy than other methods at the beginning of all tests. This indicates that the GAL method tends to sample more informative

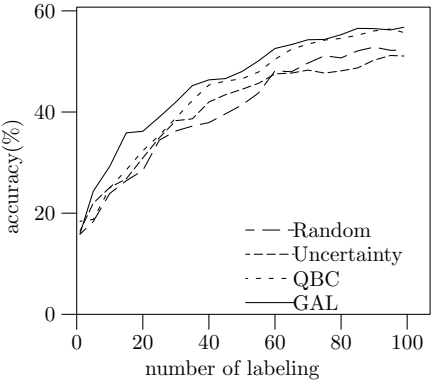


Fig. 3. learning curves on *coil*

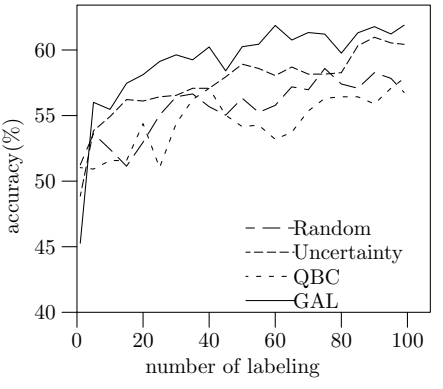


Fig. 4. learning curves on *secstr1500*

instances than other active learning methods. Second, we find that all the learning curves fluctuate during the whole cures in Figure 4. The dataset *secstr1500* is too sparse to obtain stable results. To the best of our knowledge, the current graph-based learning methods can not handle the whole dataset of *secstr* with 83679 instances efficiently. We will investigate it in our ongoing work.

6 Conclusions

In this paper, we use graph-based semi-supervised learning, which employs unlabeled instances when training, to strengthen active learning. A graph-based active learning method in an entropy reduction framework is presented. The proposed method trains a classifier using a popular graph-based semi-supervised label propagation method in the learning engine and samples the instance with the largest expected entropy reduction in the sampling engine. The experiments show that the proposed method outperforms the traditional sampling methods on selected datasets.

We make several contributions in this paper. First, an graph-based entropy reduction framework for active learning was proposed. Second, the method in which both the learning engine and the sampling engine are modified to utilize the unlabeled instances for active learning was presented. Third, the proposed graph-based active learning method can also handle multi-class learning problems.

We would like to pursue the following directions: extending the proposed method to more complex data problem and developing efficient algorithms for active learning on manifold structure.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.60603015, 60603062).

References

1. Hieu, T.N., Arnold, S.: Active learning using pre-clustering. In: Proc. 21th International Conf on Machine Learning, Banff, CA. Morgan Kaufmann, San Francisco (2004)
2. Muslea, I., Minton, S., Knoblock, C.A.: Active learning with multiple views. Journal of Artificial Intelligence Research 27, 203–233 (2006)
3. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: 17th ACM International Conference on Research and Development in Information Retrieval, pp. 3–12. Springer, Heidelberg (1994)
4. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2001)
5. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: Proc. 17th International Conf. on Machine Learning, CA, pp. 839–846. Morgan Kaufmann, San Francisco (2000)
6. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence research 4, 129–145 (1996)
7. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Proc. 18th International Conf. on Machine Learning, pp. 441–448. Morgan Kaufmann, San Francisco (2001)
8. Zhu, X., Lafferty, J.D., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data (2003)
9. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Workshop on Computational Learning Theory, San Mateo, CA, pp. 287–294. Morgan Kaufmann, San Francisco (1992)
10. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning 28, 133–168 (1997)
11. Abe, N., Mamitsuka, H.: Query learning using boosting and bagging. In: Proc. 15th International Conf. on Machine Learning, Madison, CA, pp. 1–10. Morgan Kaufmann, San Francisco (1998)
12. Melville, P., Mooney, R.J.: Diverse ensembles for active learning. In: Proc. 21th International Conf. on Machine Learning, Banff, CA, pp. 584–591. Morgan Kaufmann, San Francisco (2004)
13. Long, J., Yin, J., Zhu, E.: An active learning method based on most possible misclassification sampling using committee. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 104–113. Springer, Heidelberg (2007)
14. McCallum, A., Nigam, K.: Employing em and pool-based active learning for text classification. In: ICML, pp. 350–358 (1998)
15. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML, pp. 912–919 (2003)
16. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NIPS (2003)
17. Zhu, X.: Semi-supervised learning literature survey (2006)
18. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: ICML, pp. 19–26 (2001)

19. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation (2002)
20. Belkin, M., Niyogi, P.: Using manifold stucture for partially labeled classification. In: NIPS, pp. 929–936 (2002)
21. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)
22. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. In: ICML, pp. 19–26 (2003)

Golden Complementary Dual in Quadratic Optimization

Akifumi Kira and Seiichi Iwamoto

Faculty of Economics, Kyushu University
6-19-1 Hakozaki, Higashi, Fukuoka 812-8581, Japan
ec207004@s.kyushu-u.ac.jp, iwamoto@en.kyushu-u.ac.jp

Abstract. This paper discusses the dual of infinite-variable quadratic minimization (primal) problems from a view point of Golden ratio. We consider two pairs of primal and dual (maximization) problems. One pair yields the Golden complementary duality: (i) Both the minimum value function and the maximum value function are the identical Golden quadratic. (ii) Both the minimum point and the maximum point constitute the Golden paths. (iii) The alternate sequence of both the Golden paths constitutes another Golden path. The other yields the inverse-Golden complementary duality: (i)′ Both the minimum value function and the maximum value function are the identical inverse-Golden quadratic, (ii) and (iii).

1 Introduction

The golden section is a line segment divided into two according to the golden ratio. It is a proportion which is considered to be particularly pleasing to the eye.

The Golden ratio is one of the most beautiful numbers. The desire for optimality is inherent for humans. One minimization leads to the other maximization, which arrives at a duality. We direct our attention to both the Golden ratio and the duality. A duality of fine features is shown.

In this paper, we are concerned with dynamic optimization problems of infinitely many variables from a viewpoint of Golden duality [7,8,10]. We take two typical quadratic minimization (primal) problems with initial condition and associate each problem with a quadratic maximization (dual) problem with transversality condition. The two pairs of primal and dual problems have an interesting feature. As for the first pair, the minimum value function is Golden quadratic and the minimum point constitutes a Golden path, while so is the maximum value function and the maximum point does such another. As for the second, the minimum value function is inverse-Golden quadratic and the minimum point constitutes the same Golden path, while so is the maximum value function and the maximum point does the same such another.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 191–202, 2008.

2 Duality

A real number

$$\phi = \frac{1+\sqrt{5}}{2} \approx 1.618$$

is called *Golden number* [1,2,14]. It is the larger of the two solutions to quadratic equation (QE)

$$x^2 - x - 1 = 0. \tag{1}$$

Sometimes QE (1) is called *Fibonacci* or *Golden*. The Golden QE has two real solutions: ϕ and its *conjugate* $\overline{\phi} := 1 - \phi$. We note that

$$\phi + \overline{\phi} = 1, \qquad \phi \cdot \overline{\phi} = -1.$$

Further we have

$$\phi^{-1} = \phi - 1, \quad (\overline{\phi})^{-1} = -\phi, \quad \phi^{-1} + (\overline{\phi})^{-1} = -1, \quad \phi^{-1} \cdot (\overline{\phi})^{-1} = -1,$$
$$\phi^2 = 1 + \phi, \quad \overline{\phi}^2 = 2 - \phi, \quad \phi^2 + \overline{\phi}^2 = 3, \quad \phi^2 \cdot \overline{\phi}^2 = 1.$$

A linear function $u(x) = ax$ is called *Golden*^{n} if $a = \phi^n$, where $n = 1, -1, 2, -2, \ldots$ A quadratic function $v(x) = ax^2$ is also called *Golden*^{n} if $a = \phi^n$. In either, Golden^1 is simply called *Golden*. Golden^{-1} is also called *inverse-Golden* [9]. In this section, we consider two pairs of primal and dual problems. One pair yields a duality for the Golden quadratic function. The other pair yields a duality for the inverse-Golden quadratic function.

2.1 Golden Duality

We take an interval $[0, x]$, where $x > 0$. Let us consider the set of all divisions of the interval $[0, x]$. Each division is specified by an inner point $y \in [0, x]$, which splits the interval $[0, x]$ into two intervals $[0, y]$ and $[y, x]$. A point $(2-\phi)x$ splits the interval into two intervals $[0, (2-\phi)x]$ and $[(2-\phi)x, x]$. A point $(\phi - 1)x$ splits it into $[0, (\phi - 1)x]$ and $[(\phi - 1)x, x]$. In either case, the length constitutes the Golden ratio $(2 - \phi) : (\phi - 1) = 1 : \phi$. Thus both divisions are the *Golden section* [1,2,14].

Definition 1. *[11] A sequence $x : \{0, 1, \ldots\} \to R^1$ is called Golden if and only if either*

$$\frac{x_{t+1}}{x_t} = \phi - 1 \qquad \text{or} \qquad \frac{x_{t+1}}{x_t} = 2 - \phi.$$

Lemma 1. *[11] A Golden sequence x is either*

$$x_t = x_0(\phi - 1)^t \qquad \text{or} \qquad x_t = x_0(2 - \phi)^t.$$

We remark that

$$(\phi - 1)^t = \phi^{-t}, \qquad (2 - \phi)^t = (1 + \phi)^{-t}$$

where

$$\phi - 1 = \phi^{-1} \approx 0.618, \qquad 2 - \phi = (1+\phi)^{-1} \approx 0.382$$

Let R^∞ be the set of all sequences of real values :

$$R^\infty = \{x = (x_0, x_1, \ldots, x_n, \ldots) \,|\, x_n \in R^1 \quad n = 0, 1, \ldots\}.$$

We consider a primal problem on R^∞ [1] :

$$(\mathrm{P}_1) \qquad \begin{array}{l} \text{minimize} \quad \displaystyle\sum_{n=0}^{\infty} \left[x_n^2 + (x_n - x_{n+1})^2\right] \\ \text{subject to} \quad \text{(i)} \quad x \in R^\infty \quad \text{(ii)} \quad x_0 = c \end{array}$$

where $c \in R^1$.
A dual problem is a maximization problem of $\mu = (\mu_0,\ \mu_1, \ldots,\ \mu_n, \ldots) \in R^\infty$:

$$(\mathrm{D}_1) \qquad \begin{array}{l} \text{Maximize} \quad c^2 + 2c\mu_0 - \displaystyle\sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] \\ \text{subject to} \quad \text{(i)} \quad \mu \in R^\infty \quad \text{(ii)} \quad \lim_{n\to\infty} \mu_n = 0. \end{array}$$

We note that both problems contain a common series $\displaystyle\sum_{n=0}^{\infty} \left[y_n^2 + (y_n - y_{n+1})^2\right]$.
In either problem, we are concerned with the finite convergence case :
$\displaystyle\sum_{n=0}^{\infty} \left[y_n^2 + (y_n - y_{n+1})^2\right] < \infty$. This implies that $\lim_{n\to\infty} y_n = 0$. In Section 3, we will see that the additional transversality condition (ii) enables us to make dual of (P_1) without difficulty. Therefore, the transversality condition may be removed from the constraints.

Theorem 1. (Golden duality) (i) The primal problem (P_1) has the minimum value $m = \phi c^2$ at the point

$$\hat{x} = c\,(1,\ (2-\phi),\ \ldots,\ (2-\phi)^n, \ldots)\,.$$

(ii) The dual problem (D_1) has the maximum value $M = \phi c^2$ at the point

$$\mu^* = \phi^{-1} c\,(1,\ (2-\phi),\ \ldots,\ (2-\phi)^n, \ldots)\,.$$

We make an observation about the two optimal solutions. First, both the minimum value function and the maximum value function are the identical *Golden quadratic value function* (Golden dual).

$$m = M = \phi c^2.$$

[1] As for corresponding finite variable problems see [6], and as for their dual and others see [3,4,5,6,12].

Second, both the minimum point and the maximum point constitute one *Golden path* (Golden).

$$\begin{aligned}\hat{x} &= (x_0,\ \hat{x}_1,\ \hat{x}_2,\ \ldots,\ \hat{x}_n,\ \ldots)\\ &= (c,\ c(2-\phi),\ c(2-\phi)^2,\ \ldots,\ c(2-\phi)^n,\ \ldots)\end{aligned}$$

$$\begin{aligned}\mu^* &= (\mu_0^*,\ \mu_1^*,\ \mu_2^*,\ \ldots,\ \mu_n^*,\ \ldots)\\ &= (c\phi^{-1},\ c\phi^{-1}(2-\phi),\ c\phi^{-1}(2-\phi)^2,\ \ldots,\ c\phi^{-1}(2-\phi)^n,\ \ldots).\end{aligned}$$

Third, the alternate sequence of both the Golden paths $\breve{y}$ constitutes another *Golden path* (Golden complement).

$$\begin{aligned}\breve{y} &:= (x_0,\ \mu_0^*,\ \hat{x}_1,\ \mu_1^*,\ \hat{x}_2,\ \mu_2^*,\ \ldots,\ \hat{x}_n,\ \mu_n^*,\ \ldots)\\ &= (c,\ c\phi^{-1},\ c(2-\phi),\ c\phi^{-1}(2-\phi),\ \ldots,\ c(2-\phi)^n,\ c\phi^{-1}(2-\phi)^n,\ \ldots)\\ &= (c,\ c(\phi-1),\ c(\phi-1)^2,\ c(\phi-1)^3,\ \ldots,\ c(\phi-1)^{2n},\ c(\phi-1)^{2n+1},\ \ldots).\end{aligned}$$

How beautiful this duality is!

Thus, the duality is called *Golden complementary duality*. A proof of Theorem 1 will be given throughout the discussion in Section 3.

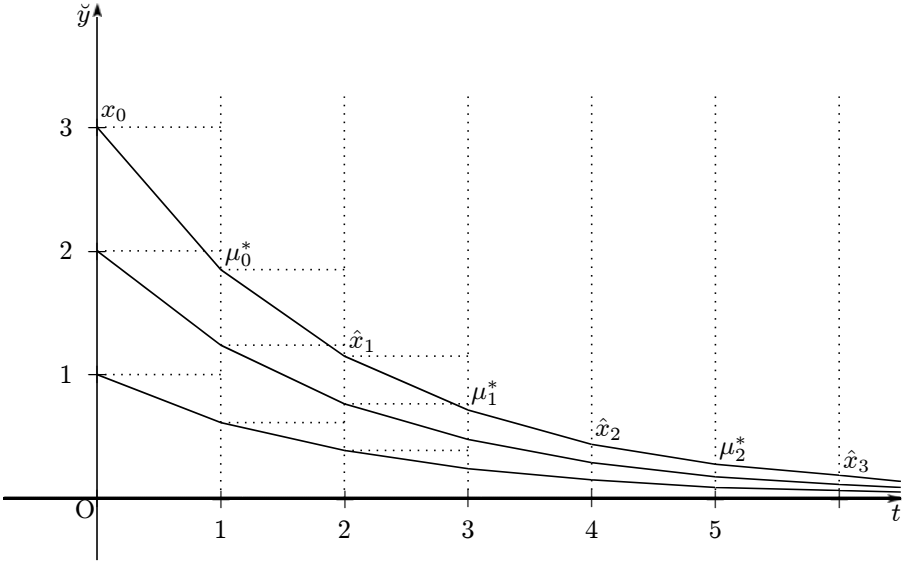


Fig. 1. Golden paths $\breve{y} = c(\phi-1)^t \quad c = 1, 2, 3$

We see that the minimum solution of (P_1) yields the maximum solution of (D_1). Let (P_1) have the minimum value $m = \phi c^2$ at the minimum point $\hat{x}$. Then we have

$$\begin{aligned}
&\text{Max}\left[c^2 + 2c\mu_0 - \sum_{n=0}^{\infty}\left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right]\right] \\
&= \underset{\mu_0}{\text{Max}}\left[c^2 + 2c\mu_0 - \min_{\{\mu_n\}_{n\geq 1}} \sum_{n=0}^{\infty}\left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right]\right] \\
&= \underset{\mu_0}{\text{Max}}\left[c^2 + 2c\mu_0 - \phi\mu_0^2\right] \\
&= \phi c^2 \qquad \text{for} \quad \mu_0 = \phi^{-1}c
\end{aligned}$$

where the minimum is attained at

$$(\hat{\mu}_1,\ \hat{\mu}_2,\ \ldots,\ \hat{\mu}_n, \ldots) = \mu_0\left(2-\phi,\ (2-\phi)^2,\ \ldots,\ (2-\phi)^n, \ldots\right).$$

Theorem 2. (Double-Golden Solution 1): $(\phi \backsim \phi^{-1})$ *The functional equation*

$$f(c) = \underset{\mu \in R^1}{\text{Max}}\left[c^2 + 2c\mu - f(\mu)\right] \qquad c \in R^1$$

has a maximum value function $f(c) = \phi c^2$ *for maximum point function* $\hat{\mu}(c) = \phi^{-1}c$.

We observe that f is Golden (and) quadratic and $\hat{\mu}$ is inverse-Golden (and) linear. This is the first double-Golden solution.

Proof. It is easily verified that f with $\hat{\mu}$ satisfies the functional equation (see also [15]). □

Corollary 1. (Double-Golden Solution 2): $(\phi \backsim \phi)$ *The functional equation*

$$f(c) = \underset{\mu \in R^1}{\text{Max}}\left[2c\mu + \mu^2 - f(\mu)\right] \qquad c \in R^1$$

has a maximum value function $f(c) = \phi c^2$ *for maximum point function* $\breve{\mu}(c) = \phi c$.

We observe that f is Golden quadratic and $\breve{\mu}$ is Golden linear, which is the second double-Golden solution.

2.2 Inverse-Golden Duality

Second we consider a primal problem

$$(P_2) \qquad \begin{aligned} &\text{minimize} \quad \sum_{n=0}^{\infty}\left[(x_n - x_{n+1})^2 + x_{n+1}^2\right] \\ &\text{subject to} \quad \text{(i)} \ \ x \in R^{\infty} \quad \text{(ii)} \ \ x_0 = c \end{aligned}$$

and a dual problem

$$(\mathrm{D}_2) \qquad \begin{array}{l} \text{Maximize } \ 2c\mu_0 - \mu_0^2 - \displaystyle\sum_{n=0}^{\infty} \left[(\mu_n - \mu_{n+1})^2 + \mu_{n+1}^2 \right] \\ \text{subject to } \ \text{(i)} \quad \mu \in R^{\infty} \quad \text{(ii)} \quad \displaystyle\lim_{n\to\infty} \mu_n = 0 \end{array}$$

where $c \in R^1$. We note that a difference between (P_1) and (P_2) is constant :

$$\sum_{n=0}^{\infty} \left[x_n^2 + (x_n - x_{n+1})^2 \right] \ = \ x_0^2 + \sum_{n=0}^{\infty} \left[(x_n - x_{n+1})^2 + x_{n+1}^2 \right].$$

The difference $x_0^2 = c^2$ is also preserved between (D_1) and (D_2). This enables us to obtain a duality in terms of inverse-Golden number $\phi^{-1} = \phi - 1$ as follows.

Theorem 3. (Inverse-Golden duality) (i) The primal problem (P_2) has the minimum value $m = \phi^{-1}c^2$ at the point

$$\hat{x} = c\,(1,\ (2-\phi),\ \ \ldots,\ (2-\phi)^n, \ldots).$$

(ii) The dual problem (D_2) has the maximum value $M = \phi^{-1}c^2$ at the point

$$\mu^* \ = \ \phi^{-1} c\,(1,\ (2-\phi),\ \ \ldots,\ (2-\phi)^n, \ldots).$$

Here we have also a Golden complementary duality :
(i) Both the minimum value function and the maximum value function are the identical *inverse-Golden quadratic* (inverse-Golden dual).

$$m \ = \ M \ = \ \phi^{-1}c^2.$$

(ii) Both the minimum point $\hat{x}$ and the maximum point μ^* constitute the Golden paths, which are the same ones in (P_1) and (D_1), respectively.
(iii) The alternate sequence of both the Golden paths constitutes another *Golden path* (Golden complement).

Further the minimum solution of (P_2) yields the maximum solution of (D_2). Let (P_2) have the minimum value $m = \phi^{-1}c^2$ at the minimum point $\hat{x}$. Then we have

$$\begin{aligned} & \mathrm{Max} \left[2c\mu_0 - \mu_0^2 - \sum_{n=0}^{\infty} \left[(\mu_n - \mu_{n+1})^2 + \mu_{n+1}^2 \right) \right] \\ = \ & \max_{\mu_0} \left[2c\mu_0 - \mu_0^2 - \min_{\{\mu_n\}_{n\geq 1}} \sum_{n=0}^{\infty} \left[(\mu_n - \mu_{n+1})^2 + \mu_{n+1}^2 \right] \right] \\ = \ & \max_{\mu_0} \left[2c\mu_0 - \mu_0^2 - \phi^{-1}\mu_0^2 \right] \\ = \ & \phi^{-1}c^2 \qquad \text{for} \quad \mu_0 \ = \ \phi^{-1}c \end{aligned}$$

where the minimum is attained at

$$(\hat{\mu}_1,\ \hat{\mu}_2,\ \ldots,\ \hat{\mu}_n, \ldots) = \mu_0 \left(2-\phi,\ (2-\phi)^2,\ \ldots,\ (2-\phi)^n, \ldots\right).$$

Theorem 4. (Double-Golden Solution 3): $(\phi^{-1} \backsim \phi^{-1})$ *The functional equation*

$$g(c) = \underset{\mu \in R^1}{\text{Max}} \left[2c\mu - \mu^2 - g(\mu)\right] \qquad c \in R^1$$

has a maximum value function $g(c) = \phi^{-1}c^2$ *for maximum point function* $\breve{\mu}(c) = \phi^{-1}c$.

Corollary 2. (Double-Golden Solution 4): $(\phi^{-1} \backsim \phi)$ *The functional equation*

$$g(c) = \underset{\mu \in R^1}{\text{Max}} \left[-c^2 + 2c\mu - g(\mu)\right] \qquad c \in R^1$$

has a maximum value function $g(c) = \phi^{-1}c^2$ *for maximum point function* $\hat{\mu}(c) = \phi c$.

3 Lagrangean Method

In this section we show how the Lagrangean method derives a maximization (dual) problem from the minimization (primal) problem.

Let us reconsider the primal problem

$$(\mathrm{P}_1) \qquad \begin{array}{ll} \text{minimize} & \displaystyle\sum_{n=0}^{\infty} \left[x_n^2 + (x_n - x_{n+1})^2\right] \\ \text{subject to} & \text{(i)}\ \ x \in R^{\infty} \quad \text{(ii)}\ \ x_0 = c. \end{array}$$

We introduce a sequence of variables $u = \{u_0,\ u_1,\ \ldots,\ u_n,\ \ldots\}$ by

$$u_n = x_{n+1} - x_n.$$

Then (P_1) is formulated into a quadratic minimization under a linear constraint

$$(\mathrm{P}_1') \qquad \begin{array}{ll} \text{minimize} & \displaystyle\sum_{n=0}^{\infty} \left(x_n^2 + u_n^2\right) \\ \text{subject to} & \text{(i)}\ \ x_{n+1} = x_n + u_n \quad n \geq 0 \quad \text{(ii)}\ \ x_0 = c. \end{array}$$

Let us now solve this problem through a Lagrangean multiplier's method. We introduce a sequence of variables $\mu = \{\mu_0,\ \mu_1,\ \ldots,\ \mu_n,\ \ldots\}$ with the property $\lim_{n\to\infty} \mu_n = 0$, which is called a *Lagrange multiplier*. Let us construct the Lagrangean

$$L(x, u, \mu) = \sum_{n=0}^{\infty} \left[x_n^2 + u_n^2 - 2\mu_n(x_{n+1} - x_n - u_n)\right].$$

Then it has the partial derivatives

$$\begin{aligned} L_{x_n} &= 2(x_n + \mu_n - \mu_{n-1}) \qquad n \geq 1 \\ L_{u_n} &= 2(u_n + \mu_n) \qquad n \geq 0 \\ L_{\mu_n} &= -2(x_{n+1} - x_n - u_n) \qquad n \geq 0. \end{aligned}$$

Here we notice that the Lagrangean is not one for a *regular* (a finite n variables with a finite m constraints) extremal problem. The problem has a countably infinite variables under a countably infinite linear constraints.

Lemma 2. *Let (x, u) be an extremum point. Then there exists a μ satisfying the condition that all the partial derivatives at point (x, u, μ) vanish:*

$$L_{x_n} = 0 \quad n \geq 1, \qquad L_{u_n} = L_{\mu_n} = 0 \quad n \geq 0. \tag{2}$$

Proof. Let $\hat{x} = (\hat{x}_n)_{n\geq 0}, \hat{u} = (\hat{u}_n)_{n\geq 0}$ be an extremum point for (P_1'). We take any large positive integer N and consider a finite-truncated conditional minimization problem of $x = (x_n)_0^N, u = (u_n)_0^N$:

$$(\mathrm{T}_N) \quad \begin{aligned} &\text{minimize} \quad \sum_{n=0}^{N} \left(x_n^2 + u_n^2\right) \\ &\text{subject to} \quad \text{(i)} \quad x_{n+1} = x_n + u_n \quad 0 \leq n \leq N-1 \\ &\qquad\qquad\quad \text{(ii)} \quad \hat{x}_{N+1} = x_N + u_N \\ &\qquad\qquad\quad \text{(iii)} \quad x_0 = c. \end{aligned}$$

This has $(2N+1)$ variables $x_1, \ldots, x_N$, $u_0, \ldots, u_N$ and $(N+1)$ linear constraints. Let us construct the Lagrangean L^N by

$$\begin{aligned} L^N(x, u, \mu) = &\sum_{n=0}^{N-1} \left[x_n^2 + u_n^2 - 2\mu_n(x_{n+1} - x_n - u_n)\right] \\ &+ x_N^2 + u_N^2 - 2\mu_N(\hat{x}_{N+1} - x_N - u_N) \quad \text{for} \quad \mu = (\mu_n)_0^N. \end{aligned}$$

Then the point $(\hat{x}_n)_1^N, (\hat{u}_n)_0^N$ is also an extremum point for the truncated problem. It satisfies the linear independent constraint qualification [13]. Therefore, Lagrange Multiplier Theorem (for a regular problem) implies that there exists a $(\mu_n^*)_0^N$ such that $(\hat{x}_n)_1^N, (\hat{u}_n)_0^N; (\mu_n^*)_0^N$ satisfies

$$L_{x_n}^N = 0 \quad 1 \leq n \leq N, \qquad L_{u_n}^N = L_{\mu_n}^N = 0 \quad 0 \leq n \leq N.$$

Thus we have

$$\begin{aligned} \hat{x}_n + \mu_n^* - \mu_{n-1}^* &= 0 \qquad 1 \leq n \leq N \\ \hat{u}_n + \mu_n^* &= 0 \qquad 0 \leq n \leq N \\ \hat{x}_{n+1} - \hat{x}_n - \hat{u}_n &= 0 \qquad 0 \leq n \leq N. \end{aligned}$$

Since N is arbitrarily large, we conclude that there exists a $(\mu_n^*)_{n\geq 0}$ such that $(\hat{x}_n)_{n\geq 1}, (\hat{u}_n)_{n\geq 0}; (\mu_n^*)_{n\geq 0}$ satisfies (2). This completes the proof. □

Then (2) is equivalent to

$$\begin{aligned} x_n &= -(\mu_n - \mu_{n-1}) \quad n \geq 1 \\ u_n &= -\mu_n \quad n \geq 0 \\ u_n &= x_{n+1} - x_n \quad n \geq 0. \end{aligned}$$

Now we solve this equivalent system in the following. Deleting u and μ, we get a system of linear equations:

$$x_{n-1} - 3x_n + x_{n+1} = 0 \quad n \geq 1, \quad x_0 = c. \tag{3}$$

Thus we have

$$u_n = x_{n+1} - x_n, \quad \mu_n = -(x_{n+1} - x_n). \tag{4}$$

We see that Eq.(3) has the solution[2]

$$x_n = c(2-\phi)^n \quad n \geq 0. \tag{5}$$

Thus we have

$$u_n = -\frac{c}{\phi}(2-\phi)^n, \quad \mu_n = \frac{c}{\phi}(2-\phi)^n. \tag{6}$$

Lemma 3. *It holds that for any finite N-stage process*

$$\begin{aligned} \sum_{n=0}^{N} \left(x_n^2 + u_n^2\right) &= \sum_{n=0}^{N} \left[x_n^2 + u_n^2 - 2\mu_n(x_{n+1} - x_n - u_n)\right] \\ &= c^2 + 2c\mu_0 - \sum_{n=1}^{N} (\mu_n - \mu_{n-1})^2 - \sum_{n=0}^{N} \mu_n^2 - 2\mu_N x_{N+1} \\ &\quad + \sum_{n=1}^{N} \left[x_n + (\mu_n - \mu_{n-1})\right]^2 + \sum_{n=0}^{N} (u_n + \mu_n)^2 . \end{aligned} \tag{7}$$

for any (x, u) satisfying (i), (ii) *and any $\mu \in R^\infty$.*

Lemma 4. *The solution (x, u) in (5), (6) is a minimum point for (P_1'). Hence, x is a minimum point for (P_1).*

Proof. We show that the (x, u) is a minimum point. Let (X, U) be any solution satisfying

$$X_{n+1} = X_n + U_n \quad n \geq 0, \quad X_0 = c.$$

Here we take the μ in (6). We may consider the set of all feasible x satisfying $\lim_{N\to\infty} \mu_N x_{N+1} = 0$ in (P_1), because of the quadratic minimization and

[2] A general solution of (3) is $x_n = A(2-\phi)^n + B(1+\phi)^n$, where $A + B = c$. The case $A = c, B = 0$ attains a minimum value for (P_1).

$\lim_{N\to\infty} \mu_N = 0$. This set in turn includes the set of all feasible x satisfying $\lim_{N\to\infty} x_N = 0$. Thus letting $N \to \infty$ in (7), we have

$$\sum_{n=0}^{\infty} \left(x_n^2 + u_n^2\right) = c^2 + 2c\mu_0 - \sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] + \sum_{n=1}^{\infty} \left[x_n + (\mu_n - \mu_{n-1})\right]^2 + \sum_{n=0}^{\infty} (u_n + \mu_n)^2 . \tag{8}$$

Similarly, we have for (X, U)

$$\sum_{n=0}^{\infty} \left(X_n^2 + U_n^2\right) = c^2 + 2c\mu_0 - \sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] + \sum_{n=1}^{\infty} \left[X_n + (\mu_n - \mu_{n-1})\right]^2 + \sum_{n=0}^{\infty} (U_n + \mu_n)^2 . \tag{9}$$

Since (x, u) satisfies

$$x_n + (\mu_n - \mu_{n-1}) = 0, \quad u_n + \mu_n = 0, \quad \lim_{n\to\infty} \mu_n = 0$$

a comparison between (8) and (9) yields

$$\sum_{n=0}^{\infty} \left(x_n^2 + u_n^2\right) \le \sum_{n=0}^{\infty} \left(X_n^2 + U_n^2\right) .$$

This completes the proof. □

From (9) we have a basic inequality as follows.

Lemma 5. *It holds that*

$$\begin{aligned} \sum_{n=0}^{\infty} \left(x_n^2 + u_n^2\right) &= c^2 + 2c\mu_0 - \sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] \\ &\quad + \sum_{n=1}^{\infty} \left[x_n + (\mu_n - \mu_{n-1})\right]^2 + \sum_{n=0}^{\infty} (u_n + \mu_n)^2 . \qquad (10) \\ &\ge c^2 + 2c\mu_0 - \sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] \end{aligned}$$

for any (x, u) *satisfying* (i), (ii) *and any* μ *satisfying* $\lim_{n\to\infty} \mu_n = 0$. *The equality holds if and only if*

$$x_n = -(\mu_n - \mu_{n-1}) \quad n \ge 1 \quad \textit{and} \quad u_n = -\mu_n \quad n \ge 0.$$

This lemma states that

$$L(\hat{x}, \hat{u} : \mu) \le L(\hat{x}, \hat{u} : \mu^*) \le L(x, u : \mu^*) \tag{11}$$

where

$$L(x,u:\mu) = \sum_{n=0}^{\infty} \left[x_n^2 + u_n^2 - 2\mu_n(x_{n+1} - x_n - u_n)\right]$$

$$\hat{x}_n = c(2-\phi)^n, \quad \hat{u}_n = -\frac{c}{\phi}(2-\phi)^n, \quad \mu_n^* = \frac{c}{\phi}(2-\phi)^n.$$

In fact, we have the equality between left-hand side and middle side:

$$L(\hat{x},\hat{u}:\mu) = L(\hat{x},\hat{u}:\mu^*) \qquad \forall \mu \ ; \ \lim_{n\to\infty} \mu_n = 0.$$

Hence we have a maximization problem for $\mu = (\mu_0, \mu_1, \ldots, \mu_n, \ldots)$ as follows:

$$(\mathrm{D}_1) \qquad \begin{array}{l} \text{Maximize } \ c^2 + 2c\mu_0 - \displaystyle\sum_{n=0}^{\infty} \left[\mu_n^2 + (\mu_n - \mu_{n+1})^2\right] \\ \text{subject to } \ \text{(i)} \ \ \mu \in R^{\infty} \ \ \text{(ii)} \ \ \lim_{n\to\infty} \mu_n = 0. \end{array}$$

Thus we have derived the desired dual problem together with the optimum solution.

Lemma 6. *The point* μ^* *with* $\mu_n^* = \dfrac{c}{\phi}(2-\phi)^n$ *attains the maximum value* $M = \phi c^2$ *for* (D_1).

4 Conclusion

We have discussed a beautiful aspect in deterministic environment where the Golden ratio has been incorporated in a complementary duality. This complementarity is differenct from the complemetary slackness in primal and dual optimization. The duality is based upon the Golden ratio. This approach has a potential in extending the duality both in dynamic decison-making and in non-deterministic environment.

References

1. Beutelspacher, A., Petri, B.: Der Goldene Schnitt 2., überarbeitete und erweiterte Auflage. Elsevier GmbH, Spectrum Akademischer Verlag, Heidelberg (1996)
2. Dunlap, R.A.: The Golden Ratio and Fibonacci Numbers. World Scientific Publishing Co. Pte. Ltd, Singapore (1977)
3. Iwamoto, S.: Inverse theorem in dynamic programming I, II, III. J. Math. Anal. Appl. 58, 113–134, 247–279, 439–448 (1977)
4. Iwamoto, S.: Dynamic programming approach to inequalities. J. Math. Anal. Appl. 58, 687–704 (1977)
5. Iwamoto, S.: Reverse function, reverse program and reverse theorem in mathematical programming. J. Math. Anal. Appl. 95, 1–19 (1983)

6. Iwamoto, S.: Theory of Dynamic Program. Kyushu Univ. Press, Fukuoka (1987) (in Japanese)
7. Iwamoto, S.: The Golden optimum solution in quadratic programming. In: Proc. of the Fourth Intl Conference on Nonlinear Analysis and Convex Analysis (NACA 2005), pp. 109–205. Yokohama Publishers, Yokohama (2007)
8. Iwamoto, S.: Golden quadruplet: optimization - inequality - identity - operator. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 14–23. Springer, Heidelberg (2007)
9. Iwamoto, S., Kira, A., Yasuda, M.: Golden duality in dynamic optimization. In: Proceedings of the Second KOSEN Workshop Mathematics, Technology and Education (MTE 2008), Ibaraki, pp. 35–47 (2008)
10. Iwamoto, S., Yasuda, M.: Dynamic programming creates the Golden ratio, too. In: Proceedings of the Workshop Decision Making under Uncertainty and Mathematical Models. Research Institute for Mathematical Sciences, Kyoto Univ., Suri Kagaku Kokyu Roku No. 1477, pp. 136–140 (2006)
11. Iwamoto, S., Yasuda, M.: Golden optimal path in discrete-time dynamic optimization processes. In: The 15-th International Conference on Difference Equations and Applications (ICDEA 2006), Kyoto (July 2006) (to appear)
12. Iwamoto, S., Tomkins, R.J., Wang, C.-L.: Some theorems on reverse inequalities. J. Math. Anal. Appl. 119, 282–299 (1986)
13. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: Neyman, J. (ed.) Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, pp. 481–492. Univ. of California Press, Berkeley (1951)
14. Walser, H.: Der Goldene Schnitt. B.G. Teubner, Leibzig (1996)
15. Walter, W.: On a functional equation of Bellman in the theory of dynamic programming. Aequations Math. 14, 435–444 (1976)

A Linear-Time Multivariate Micro-aggregation for Privacy Protection in Uniform Very Large Data Sets

Agusti Solanas and Roberto Di Pietro

UNESCO Chair in Data Privacy
Dept. Computer Engineering and Mathematics
Rovira i Virgili University
Av. Països Catalans, 26. E-43001 Tarragona. Catalonia
{agusti.solanas,roberto.dipietro}@urv.cat

Abstract. Optimally micro-aggregating a multivariate data set is known to be NP-hard, thus, heuristic approaches are used to cope with this privacy preserving problem. Unfortunately, algorithms in the literature are computationally costly, and this prevents using them on large data sets.

We propose a partitioning algorithm to micro-aggregate uniform very large data sets with cost $O(n)$. We provide the mathematical foundations proving the efficiency of our algorithm and we show that the error associated to micro-aggregation is bounded and decreases when the number of micro-aggregated records grows. The experimental results confirm the prediction of the mathematical analysis. In addition, we provide a comparison between our proposal and MDAV, a well-known micro-aggregation algorithm with cost $O(n^2)$.

1 Introduction

Information and communication technologies (ICT) foster the gathering of personal data. The old paper-based files that occupy a large amount of space are being replaced by electronic files that can be stored in tiny USB flash drives. Thus, a paradigm shift is tacking place in many of our daily activities. An exponent of this paradigm shift is the so-called *e-administration* that aims to achieve a paperless office, where all the old paper-based processes are replaced by electronic ones. The main goal of this change is to improve productivity and performance. At the same time, moving aside paper-based offices may lead to total transparency and accountability and, by extension, to better *e-governance*.

This novel way of understanding the management of data is especially reflected in very important areas such as e-commerce and health-care that must conform to strict regulations [1][5]. In addition, most countries have legislation which compels national statistical agencies to guarantee statistical confidentiality [10][11][14]. Thus, protecting individual privacy is a key issue for many institutions, namely statistical agencies, Internet companies, manufacturers, etc.

Many efforts have been devoted to develop techniques guaranteeing privacy, but there are many examples of negligence regardless. British politicians became astonished when they were told on November 20th, 2007, that two computer disks full of

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 203–214, 2008.

personal data of 25m British individuals had gone missing. The fate of the disks is unknown and the privacy of the individuals, whose personal data are lost, is in danger. This is the most recent of a series of similar unfortunate cases. In October, 2007, Her Majesty's Revenue and Customs (HMRC) lost another disk containing pension records of 15.000 people. Data on 26.5m people were stolen from the home of an employee of the Department of Veterans Affairs in America in 2006, and 658000 queries were disclosed by the AOL search engine in August of the same year. These disclosures of personal information are not new; however, due to the great advances in the Information and Communication Technologies (ICTs), it is very easy to gather large amounts of personal data, and mistakes such as the formerly explained are magnified.

Due to this dramatic increase in the amount of stored personal data, there is a real need for efficient methods to protect privacy. Micro-aggregating data is a common solution to protect the privacy of the users, whose data is stored, however, micro-aggregating very large data sets is a very costly task when current of-the-shelf methods are used.

With the aim of overcoming the limitations of the current micro-aggregation methods, we propose an efficient micro-aggregation algorithm that allows large amounts of electronic data to be micro-aggregated in linear time, *i.e.* with a cost $O(n)$. Our proposal have been designed to work with uniformly distributed data, however, it can be extended to work with other data distributions. Experiments confirm the results of the mathematical analysis and proof the usefulness of our proposal.

The rest of the paper is organised as follows. Section 2 is a summary of the main concepts of micro-aggregation and its methods. In Section 3 our proposal is explained in detail. Section 4 contains the experimental results that proof the usefulness of our method. Finally, the article concludes in Section 5.

2 Background

Personal privacy is a foundational principle stated by the Universal Declaration of Human Rights[1]. With the aim of protecting this foundational right, Statistical Disclosure Control (SDC) was proposed as a discipline that sought to transform data in such a way that they could be publicly released whilst preserving data utility and statistical confidentiality. The point was to avoid disclosure of information that could be linked to specific individual or corporate respondent entities.

It is necessary to keep the balance between the right of the individuals to privacy and the right of the society to knowledge. The solution to this problem is somewhere between two extremes: (i) **No modification**, *i.e.* maximal data utility but no privacy, and (ii) **total encryption**, *i.e.* absolute privacy but no data utility.

One of the youngest techniques proposed to keep this balance is micro-aggregation. It is an SDC technique consisting in the aggregation of individual data. It can be considered as an SDC sub-discipline devoted to the protection of individual data, also called *micro-data*. Micro-aggregation can be understood as a clustering problem with constraints on the size

[1] "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour or reputation ..." ***Universal Declaration of Human Rights.***

of the clusters. It is related to other clustering problems (*e.g.* dimension reduction or minimum squares design of clusters). However, the main difference is that micro-aggregation does not consider the number of clusters to generate or the number of dimensions to reduce, but the minimum number of records that must be grouped in the same cluster.

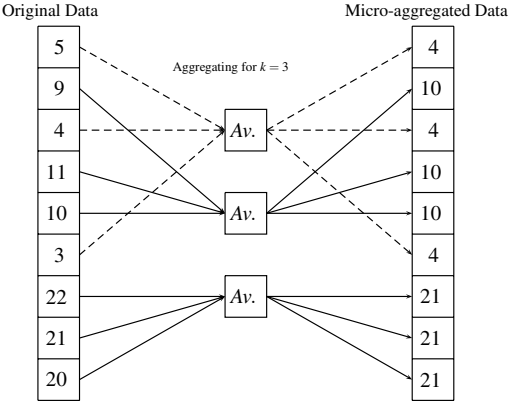


Fig. 1. Example of univariate micro-aggregation where the minimum cardinality of the groups is $k = 3$

When micro-aggregation is used, two main goals have to be kept in mind:

- *Preserving data utility.* Similar records should be micro-aggregated instead of different ones. In the example given in Figure 1, groups of three records are built and aggregated. Note that records in the same group are similar.
- *Protecting respondents' privacy.* Data have to be sufficiently modified to prevent re-identification or disclosure. In the example given in Figure 1, after micro-aggregating the original data, it is impossible to distinguish records from the same group. Thus, the probability of linking a respondent with his/her micro-data is inversely proportional to the number of aggregated records.

In order to determine the information loss produced by micro-aggregation the Sum of Squared Errors (SSE) is used (*cf.* Expression 1).

$$\mathbf{SSE} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})'(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i}) \tag{1}$$

where g is the number of groups, n_i is the number of records in the i-th group, $\mathbf{x_{ij}}$ is the j-th record in the i-th group and $\bar{\mathbf{x}}_\mathbf{i}$ is the average record of the i-th group.

This SSE measure is generally compared with the total error (SST) defined in Equation 2.

$$\mathbf{SST} = \sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})'(\mathbf{x_i} - \bar{\mathbf{x}}) \tag{2}$$

where n is the number of records in the data set, x_i is a record of the data set and $\bar{x}$ is the average record of the data set.

Given a homogeneity measure such as the SSE and a security parameter k, which determines the minimum cardinality of the groups, the micro-aggregation problem can be enunciated as follows:

> *Given a data set* **D** *with* **n** *records in a characteristic space* $\mathbb{R}^d$*, the problem consists in obtaining a k-partition*[2] $\mathcal{P}$ *of* **D***, so that the SSE of* $\mathcal{P}$ *is minimised. Once* $\mathcal{P}$ *is obtained, each record of every part of* $\mathcal{P}$ *is replaced by the average record of the part.*

The micro-aggregation problem is known to be NP-hard [9] for multivariate data sets, thus, heuristic methods must be used to solve it. There is a plethora of this kind of methods to address the multivariate micro-aggregation problem [15]. Some of them are based on building a tree structure connecting the records in the data set and partition the tree to generate a k-partition. Examples of this kind are the Minimum Spanning Tree Partitioning (MSTP) proposed in [7] and the more recently proposed μ-approximation [3]. The main limitation of the MSTP remains in its foundation, *i.e.* in the minimum spanning tree. Although it could be very useful when the data are distributed in clusters, it fails to properly adapt to the data when they are distributed in a scattered way. In addition, both algorithms have a high computational cost because they must compute the distances between all records in the data set to build the tree structure.

Instead of structuring the data in trees or graphs, an alternative way to tackle the problem is greedily building groups of similar records. Examples of this kind are the Maximum Distance (MD) method [2] and the Maximum Distance to Average Vector (MDAV) method [4,6]. A slight modification of the same construction was described in [7] under the name Centroid-Based Fixed-Size Micro-aggregation and, in [12] Solanas et al. proposed an improvement on MDAV named Variable-MDAV, V-MDAV for short. The main advantage of these methods is their simplicity but their computational complexity (*i.e.* MD has a cubic cost and the others have a quadratic cost) prevents their use with very large data sets.

In [13] genetic algorithms are used to micro-aggregate small data sets. Although this method performs very well with small amounts of data (*i.e.* with less than 100 records), it cannot be applied to large data sets. An improvement of this method that mixes MDAV and genetic algorithms was proposed in [8].

3 Our Proposal

In this section we define our proposal. Firstly, we explain its foundational ideas and a high-level algorithm. Afterwards, we provide some details of the mathematical model and, we analyse the behaviour of our method for several values of the parameters.

3.1 Rationale

Current micro-aggregation algorithms are very costly (*i.e.* at least $O(n^2)$) because they compute the similarity (*e.g.* the squared Euclidean distance) between all possible pairs

[2] A k-partition of **D** is a partition where its parts have, at least, k records of **D**.

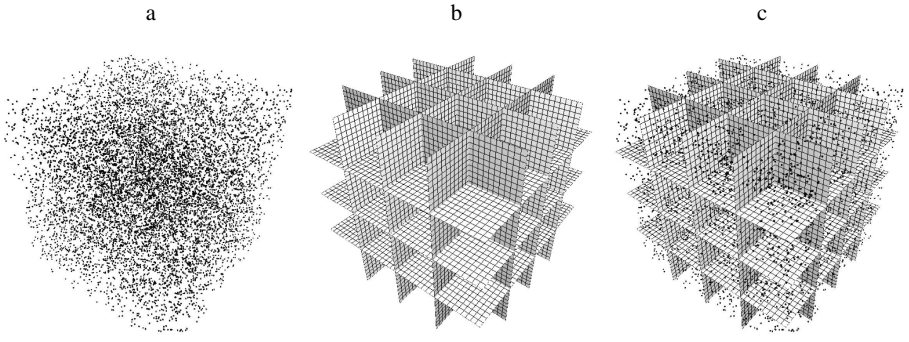


Fig. 2. Graphical example of data partitioning. a) Set of points uniformly distributed in a three dimensional space. b) Hypercubes of edge $\ell = 0.25$ generated by our algorithm. c) Wrapping of points by the generated hypercubes.

of records in the original data set. Once these similarities are known, the records are grouped.

Our proposal is absolutely different. Instead of using the similarity between records to group them, we partition the hyper-space to which the records belong. By partitioning the hyper-space we are implicitly grouping the records that are distributed inside it. By construction, the records that are in the same hypercube are very likely to be similar, thus, micro-aggregating them in the same group leads to the reduction of the SSE.

The most important contribution of our proposal is that our method dramatically reduces the time needed to micro-aggregate a given data set whilst the information loss is not significantly different from the one obtained by other costlier solutions.

Take as an example the set of records depicted in Figure 2(a). It is apparent that computing the distance between all possible pairs of records would be extremely costly. On the contrary, computing the hypercubes depicted in Figure 2(b) is easy and computationally cheap. Finally, assigning each record to a group (*i.e.* to a hypercube) is very easy thanks to the regularity of the hypercubes. Figure 2(c) depicts how the records are wrapped by their corresponding hypercubes.

3.2 Algorithm

Our method receives as input a data set **D** containing n records, and a security parameter k. Firstly, the records in **D** are normalised (step 1 in Algorithm 1) to assure that the hyper-space $\mathcal{HS}$ has a unitary edge (*i.e.* the length of the hypercube that wraps $\mathcal{HS}$ is 1). Next, the optimal length ℓ of the hypercubes is computed according to the number of records in $\mathcal{HS}$ and the privacy parameter k (step 2 in Algorithm 1) *cf.* to Section 3.5 for further details on the computation of ℓ. Once ℓ is determined, the set of hypercubes (HC) is created (step 3 in Algorithm 1). A hypercube is defined by its bounds in each dimension. For example, the bounds $\{(0,0.25),(0,0.25)\}$ define a two-dimensional hypercube (*i.e.* a square) that wraps all the points (x,y) such as $0 < x \leq 0.25$ and $0 < y \leq 0.25$. Each generated hypercube (HC_i) is identified by a unique index (i).

Algorithm 1. Efficient Hyper-Cubes-based Micro-aggregation

```
   Data: Data set D, Integer k, Integer n.
   Result: k-partition.
 1 D' ⟵Normalise(D);
 2 ℓ ⟵ ComputeHypercubeEdgeLength(n,k);
 3 HC ⟵ GenerateSetOfHypercubes(ℓ);
 4 for Each record r_i in D' do
 5 |   < HC_j, r_i > ⟵ AssingRecordToHypercube(HC,r_i);
 6 |   Card(HC_j) := Card(HC_j) + 1;
 7 end
 8 for Each Hypercube in HC do
 9 |   if Card(HC_j)< k then
10 |   |   Neigh = DetermineHypercubeNeighbour(HC_j);
11 |   |   FuseHypercubes(HC_j,Neigh);
12 |   |   HC ⟵ RemoveHypercubeFromSetOfHypercubes(HC_j);
13 |   |   Card(Neigh) := Card(Neigh)+Card(HC_j);
14 |   end
15 end
16 return set of tuples < HC_j, r_i >
```

After generating HC, each record in $\mathcal{HS}$ has to be assigned to a hypercube HC_i (step 5 in Algorithm 1). This assignment can be understood as a tuple of two indexes: the first index refers to the hypercube HC_j that contains the record and, the second index identifies the record r_i. Once a record is assigned to a hypercube, the cardinality of the hypercube (*e.g.* a counter) is increased (step 6 in Algorithm 1).

At this stage of the process, all records are already assigned to a hypercube. However, due to the fact that ℓ is an approximation (*cf.* Section 3.5), a little fraction of hypercubes can be smaller and, thus, they can contain less than k records. The aim of this algorithm is to obtain a k-partition. Hence, all hypercubes must contain, at least, k records. In order to cope with this constraint, low-cardinality hypercubes are fused with their closest neighbours (steps 9-14 in Algorithm 1). Note that, by construction, every low-cardinality hypercube has a complete hypercube (*i.e.* one having at least k records) adjacent to it, and the fusion of these hypercubes is straightforward and computationally cheap. Once all low-cardinality hypercubes have been fused, the algorithm finishes and returns the set of tuples $< HC_j, r_i >$ that represent the obtained k-partition.

3.3 Analytical Model

We assume that records in a data set **D** can be represented as points randomly scattered on a d-dimensional hyper-space $(\mathcal{HS})$.

Our proposal is based on the partition of $\mathcal{HS}$ in d-dimensional hypercubes (HCs) of edge ℓ. We indicate each of the $(1/\ell)^d$ hypercubes with a unique index (HC_i). Further, in a similar way, it is straightforward to find a bijection that binds a point in $\mathcal{HS}$ to a specific hypercube. Hence, a point q belonging to $\mathcal{HS}$ can be identified in two ways: first, using its d co-ordinates; second, via a tuple $q_{<i,j>}$, where the index i points to the

hypercube of index i (HC_i), and the index j refers to a unique index that identifies a unique point within HC_i.

Let $X_{i,j}$ be the random variable that takes on the value 1 if the point $q_{<i,j>}$ lies within HC_i, and 0 otherwise. Let C_i be the random variable that counts the number of points lying within HC_i. If we note that $E[X_{i,j}] = \ell^d$ since the points are uniformly distributed, we have that:

$$\mu_i = E[C_i] = E\left[\sum_{i=1}^{n} X_{i,j}\right] = \sum_{i=1}^{n} E[X_{i,j}] = \sum_{i=1}^{n} \ell^d = n\ell^d$$

Now, since the points in HC_i are *iid*, we have that for $0 < \varepsilon < 1$ the following equation holds:

$$Pr[|C_i - \mu_i| > \varepsilon\mu_i] \leq 2 \times exp\left(-\frac{\varepsilon^2}{3}\mu_i\right) \tag{3}$$

Note that Equation 3 provides an upper bound on the probability that a single HC_i has a number of points diverging from its mean by more (or less) than $\varepsilon\mu_i$. We are interested in identifying a condition that makes *all* of the HC_i to satisfy Equation 3. We can derive such condition as follows: let us define the event *Bad*="at least one of the hypercubes does not satisfy Equation 3", we have that this event happens with the probability given in Equation 4.

$$\begin{aligned} Pr[Bad] = \quad & Pr[|C_1 - \mu_i| > \varepsilon\mu_i \vee \ldots \vee |C_{1/\ell^d} - \mu_i| > \varepsilon\mu_i] \\ & \leq (1/\ell^d) Pr[|C_i - \mu_i| < \varepsilon\mu_i] \leq 2 \times exp\left(-\tfrac{\varepsilon^2}{3}\mu_i - \ln \ell^d\right) \end{aligned} \tag{4}$$

Note that the above equation fully characterises our model. For instance, if we set $k = \mu_i$ and $\varepsilon = 1/2$, from Equation 4 we have that:

$$Pr[Bad] \leq 2 \times exp\left(-\frac{k}{12} - \ln \ell^d\right) = 2 \times exp\left(-\frac{k}{12} - \ln \frac{k}{n}\right) = 2 \times exp\left(-\frac{k}{12} - \ln k + \ln n\right)$$

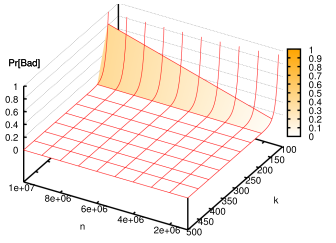
Hence, for $k = 12 \ln n$, we have:

$$Pr[Bad] \leq \frac{1}{6 \ln n}$$

that is, all the hypercubes will have a number of points that is between $k/2$ and $2k$ with probability $\frac{1}{6 \ln n}$.

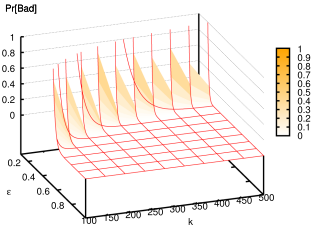
Finally, note that once set the parameters that provide the needed upper bound on the probability for the event *Bad* to happen, the corresponding value ℓ can be computed as:

$$\ell = \sqrt[d]{\frac{k}{n}} \tag{5}$$



(a) Error graph as a function of k and n for $\varepsilon = 0.5$.

(b) Error graph as a function of ε and n for $k = 160$.



(c) Error graph as a function of ε and k for $n = 1,000,000$.

Fig. 3. 3D representation of the parameters ε, k and n

3.4 Model Validation

It is worth noticing that Equation 4, whilst providing a complete characterisation for our model, is dependant on three parameters only: n, k, and ε. To assess how these parameters affect the value of Pr[*Bad*], we have varied the value k in the range $[100,\ldots,500]$ using an incremental step of 100; the value n is varied in the range $[10^6,\ldots,10^7]$, with an incremental step of 10^6, whilst ε is varied in the range $[0.1,\ldots,0.9]$ with an incremental step of 0.1. To obtain a 3D representation, we fixed one of these three parameters, whilst allowing the other two to range in the described intervals. In particular, to obtain Figure 3(a) we set $\varepsilon = 0.5$, for Figure 3(b) we chose $k = 160$, whilst $n = 10^6$ for Figure 3(c). In each of the three figures, it can be noticed that the error probability (Pr[*Bad*]) decreases exponentially fast as free parameters increase. In particular, note from Figure 3(a) ($\varepsilon = 0.5$) that it is enough to select $k > 150$ to have Pr[*Bad*] reduced to an almost negligible value. From Figure 3(b) it turns out that for any $\varepsilon > 0.4$ the error probability Pr[*Bad*] is close to zero. Finally, from Figure 3(c) ($n = 10^6$) is apparent that for any $k \geq 100$, if one chooses $\varepsilon > 0.5$, the probability Pr[*Bad*] is very close to zero.

3.5 Tuning the ℓ Value

We normalise the points in $\mathcal{HS}$, thus, $\mathcal{HS}$ has an edge equal to one. From this result, it follows that the edge of each HC_i will be between zero and one. However, note that it could be the case that $\ell \in I$, that is ℓ belongs to the set of the irrational numbers; this

situation could not be treated on a finite state machine. Hence, we add the constraint that ℓ has to be expressed as $\ell = 1/p$, where p is an integer. Therefore we introduce the value $(\overline{\ell})$ that is, an approximation of the ideal edge for the HC_i, needed due to the finiteness of the computing machines. The simplest strategy to obtain $(\overline{\ell})$, is to set $(\overline{\ell})^{-1}$ to the smallest integer larger than $(\ell)^{-1}$, that is: $(\overline{\ell})^{-1} = \lceil (\ell)^{-1} \rceil$.

Note that in this way, we have that all but a fraction ($\hat{HC}$) of the hypercubes have an edge larger than the original ℓ. Hence, for those HC_i with edge larger than ℓ, the probability that the number of points within each HC_i is less than k further decreases. However, note that there could be a a fraction of hypercubes ($\hat{HC}$) that have an edge smaller than $(\overline{\ell})$. Indeed, this fraction will always exist whenever $\ell^{-1} \neq \lceil \ell^{-1} \rceil$. Moreover, it could be possible that the resulting edge for hypercubes in $\hat{HC}$ is even smaller than ℓ. Hence, these hypercubes could host less than k points. To cope with this issue, we extend the proposed partitioning method as follows: each of the hypercube in $\hat{HC}$ (if the edge is shorter than $(\overline{\ell})$) is merged with a neighbouring hypercube that has edge $(\overline{\ell})$. It can be shown that, by construction, such a neighbour always exists and that none of such neighbours is assigned to more than one hypercube in $\hat{HC}$. Further, from the practical point of view, as intuition suggests and experimental result will confirm, the impact of this approximation is negligible.

In an extended version of this paper we plan to show that a finer, but more complex, size for ℓ exists; however due to space limitations we consider this issue out of the scope of this paper.

3.6 Dealing with Non-uniform Data Distributions

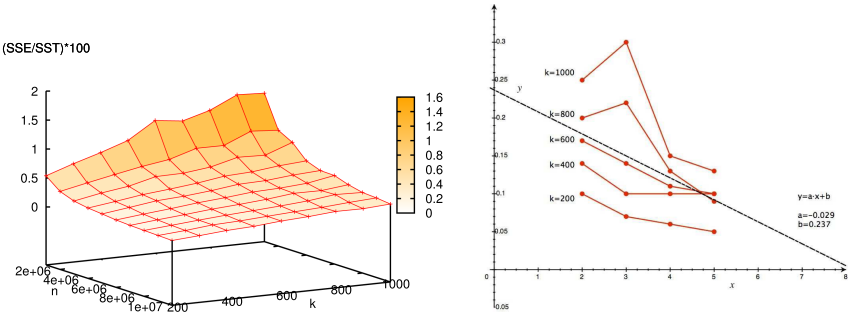
We have assumed that points in $\mathcal{HS}$ are *iid*. We are aware that, in many applications, data do not follow such a distribution pattern. However, we have detailed the mathematical foundations of the proposed method, and

we have shown that it provides very interesting results. In particular, the relative error is very low, whilst the computational cost incurred by the proposed method to partition $\mathcal{HS}$ is linear.

Furthermore, we believe that the proposed method could be used when data show a different statistical distribution, such as the Gaussian one. The key underlying idea to deal with this problem is to relax the assumption that ℓ is a constant. Indeed, we can tune ℓ to leverage the different data density in $\mathcal{HS}$ provided by a known data distribution function, so that each of the resulting HC_i still satisfy the cardinality constraint imposed by micro-aggregation. Our current investigations focus on this direction and we will tackle this problem in a future extension of this paper.

4 Experimental Results

In the following we describe the experiments that eventually show the high quality results achievable with our proposal. The experimental scenario is the following: we generated several sets of records uniformly at random in a 3-dimensional space. The size of the generated sets varied in the range $[10^6, \ldots, 10^7]$. For each of the generated sets we computed the SST (*cf.* Equation 2) and the SSE (*cf.* Equation 1) of the k-partition obtained by our method.



(a) Ratio between the SSE and SST in %.

(b) Evolution of the information loss difference between MDAV and our method

Fig. 4. Results of our method

Table 1. Brief comparison between the information loss of MDAV and our method

n	k	MDAV $\frac{SSE}{SST} \times 100$ %	Our $\frac{SSE}{SST} \times 100$ %	Diff.	n	k	MDAV $\frac{SSE}{SST} \times 100$ %	Our $\frac{SSE}{SST} \times 100$ %	Diff.
2×10^6	1000	0.75	1	0.25	4×10^6	1000	0.47	0.62	0.15
2×10^6	800	0.64	0.84	0.2	4×10^6	800	0.41	0.54	0.13
2×10^6	600	0.53	0.70	0.17	4×10^6	600	0.33	0.44	0.11
2×10^6	400	0.40	0.54	0.14	4×10^6	400	0.25	0.35	0.10
2×10^6	200	0.25	0.35	0.10	4×10^6	200	0.16	0.22	0.06
3×10^6	1000	0.57	0.87	0.3	5×10^6	1000	0.41	0.54	0.13
3×10^6	800	0.49	0.71	0.22	5×10^6	800	0.35	0.46	0.09
3×10^6	600	0.40	0.54	0.14	5×10^6	600	0.29	0.39	0.1
3×10^6	400	0.31	0.41	0.1	5×10^6	400	0.22	0.32	0.1
3×10^6	200	0.19	0.26	0.07	5×10^6	200	0.14	0.19	0.05

In Figure 4(a), on the z-axis we report the value of the ratio $(SSE/SST) * 100$, that is the percentage of the total error introduced by our micro-aggregation proposal. Note that for any value of k and n in the considered interval, the ratio does not exceed 1.6%, and that for a wide range of parameters, the same ratio is well below 0.2%. Further, the behaviour of the plot conveys another useful information: the quality of the implementation that accommodates the need of merging (few) hypercubes is highly satisfactory (*i.e.* less than 1.6% of the total error is inherited). Moreover, refining the value assigned to $(\overline{\ell})$ will further improve these results.

The fact that our algorithm has a linear cost $O(n)$ is a clear advance in the field of privacy protection by means of micro-aggregation. However, we have also compared the results of our method with a well-known $O(n^2)$ micro-aggregation algorithm *i.e.* MDAV [4,6] to proof that the dramatic computational cost reduction does not imply a significant increase in the error.

In Table 1, MDAV is compared with our method in terms of information loss. Note that we have used data sets with less than 5×10^6 records. Although our method obtains the results in few minutes, MDAV needs days to obtain them when the data sets are very large *e.g.* 10×10^6. The results in Table 1 indicate that MDAV is better than our method in terms of information loss. However the difference between both methods is not significant *i.e.* less than 0.3%. In addition, as can be clearly seen in Figure 4(b), the difference between both methods tends to decrease when the number of records increases.

These results proof that our method is useful and the only practical choice to micro-aggregate very large data sets in a reasonable time.

5 Conclusions

We have proposed an efficient multivariate micro-aggregation method that is able to micro-aggregate very large data sets in linear time $O(n)$. This is the first micro-aggregation method that achieves this goal.

We have applied our method to very large data sets of up to 10^7 records, where classical micro-aggregation algorithms cannot be applied due to their high computational cost. Thus, we have chosen values for k which are higher than the common ones. If it is necessary to obtain groups of very small cardinality, it is possible to use our method to partition the data in groups of, let's say a thousand records, and then apply a classical micro-aggregation algorithm to refine the k-partition.

In this paper we have established the foundations of a novel and efficient micro-aggregation algorithm. In the near future we plan to extend our proposal by investigating the following lines:

- Extend the algorithm to cope with non-uniform data
- Analyse the use of our algorithm as a pre-process, whose results could be used by a classical micro-aggregation algorithm to obtain a better k-partition in a reasonable time.
- Improve the SSE of our algorithm by recursively divide the hypercubes with more than $2k - 1$ records.

Acknowledgements

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organisation. This work was partly supported by the Spanish Ministry of Education through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446.

References

1. Boyens, C., Krishnan, R., Padman, R.: On privacy-preserving access to distributed heterogeneous healthcare information. In: Proceedings of the 37th Hawaii International Conference on System Sciences HICSS-37, Big Island, HI. IEEE Computer Society, Los Alamitos (2004)
2. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)

3. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Comput. Math. Appl. 55(4), 714–732 (2008)
4. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
5. HIPAA. Health insurance portability and accountability act (2004), http://www.hhs.gov/ocr/hipaa/
6. Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing, S.: μ-ARGUS version 4.0 Software and User's Manual. Statistics Netherlands, Voorburg NL (May 2005)
7. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering 17(7), 902–911 (2005)
8. Martinez-Balleste, A., Solanas, A., Domingo-Ferrer, J., Mateo-Sanz, J.M.: A genetic approach to multivariate microaggregation for database privacy. In: IEEE 23rd International Conference on Data Engineering ICDE, April 17-20, 2007, pp. 180–185 (2007)
9. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Comission for Europe 18(4), 345–354 (2001)
10. European Parliament. DIRECTIVE 2002/58/EC of the European Parliament and Council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)(2002), http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf
11. Canadian Privacy. Canadian privacy regulations (2005), http://www.media-awareness.ca/english/issues/privacy/canadian_legislation_privacy.cfm
12. Solanas, A., Martínez-Ballesté, A.: V-MDAV: Variable group size multivariate microaggregation. In: COMPSTAT 2006, Rome, pp. 917–925 (2006)
13. Solanas, A., Martinez-Balleste, A., Mateo-Sanz, J.M., Domingo-Ferrer, J.: Multivariate microaggregation based on genetic algorithms. In: 3rd International IEEE Conference on Intelligent Systems IS, pp. 65–70 (2006)
14. USPrivacy. U.S. privacy regulations (2005), http://www.media-awareness.ca/english/issues/privacy/us_legislation_privacy.cfm
15. Willenborg, L., DeWaal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)

Improving Microaggregation for Complex Record Anonymization

Jordi Pont-Tuset[1], Jordi Nin[2], Pau Medrano-Gracia[1], Josep Ll. Larriba-Pey[1], and Victor Muntés-Mulero[1]

[1]DAMA-UPC, Computer Architecture Dept.
Universitat Politècnica de Catalunya,
Campus Nord UPC, C/Jordi Girona 1-3
08034 Barcelona, Catalonia, Spain
{jpont,pmedrano,larri,vmuntes}@ac.upc.edu
[2]IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n
08193 Bellaterra, Catalonia, Spain
jnin@iiia.csic.es

Abstract. Microaggregation is one of the most commonly employed microdata protection methods. This method builds clusters of at least k original records and replaces the records in each cluster with the centroid of the cluster. Usually, when records are complex, i.e., the number of attributes of the data set is large, this data set is split into smaller blocks of attributes and microaggregation is applied to each block, successively and independently. In this way, the information loss when collapsing several values to the centroid of their group is reduced, at the cost of losing the k-anonymity property when at least two attributes of different blocks are known by the intruder.

In this work, we present a new microaggregation method called *One dimension microaggregation* ($Mic1D - \kappa$). This method gathers all the values of the data set into a single sorted vector, independently of the attribute they belong to. Then, it microaggregates all the mixed values together. Our experiments show that, using real data, our proposal obtains lower disclosure risk than previous approaches whereas the information loss is preserved.

Keywords: Microaggregation, k-anonymity, Privacy in Statistical Databases.

1 Introduction

Confidential data is usually released to third parties (e.g., politicians and researchers) for data analysis. This dissemination has to be in accordance with laws and regulations to avoid the publication of critical private information. In this situation, it is necessary to release data preserving the statistics without

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 215–226, 2008.

revealing confidential information. This is a typical problem, for instance, in official statistics institutes.

Special efforts have been made to develop a wide range of protection methods. These methods aim at guaranteeing an acceptable level of protection of the confidential data. Specific areas such as Privacy in Statistical Databases (PSD) tackle the problem of protecting confidential data in order to publicly release it, without revealing confidential information that could be linked to an specific individual or entity. Good surveys about protection methods can be found in the literature [1,3].

Recently, microaggregation has emerged as one of the most promising data protection methods. For example, the work in [7] shows that microaggregation is used by many official statistics institutes for data anonymization. The basic implementation of microaggregation [3,4,16] works as follows: given a data set with A attributes, small clusters of at least k elements (records) are built and each original record is replaced with the centroid of the cluster to which the record belongs. A certain level of privacy is ensured because k records have an identical protected value (k-*anonymity* [15,17,18]).

However, when the complexity of the records in the data set is large and, thus, the number of attributes A is large, microaggregation techniques suffer from a low statistical utility. This is so because the larger the number of attributes, the larger the distance between the original records in the data set and their corresponding centroids. Therefore, a lot of information of the original data is lost when the protected data set is released.

1 To solve this drawback, the following natural strategy is usually used: the data set is split into smaller blocks of attributes and microaggregation is independently applied to each block. In this way, the information loss is lower, at the cost of an increase of the disclosure risk. In other words, the property of k-anonymity is not ensured anymore, as we explain later in this paper.

In this work, we propose to combine a set of preprocessing steps along with the microaggregation in order to minimize the disclosure risk without losing information. We test this new method using real data sets showing that $Mic1D-\kappa$ is able to outperform previous microaggregation methods diminishing the risk of disclosure without increasing the information loss. Specifically, we compare our new method with one of the most commonly used microaggregation methods, the MDAV (Maximum Distance to Average Vector) algorithm [4], showing that $Mic1D-\kappa$ achieves lower disclosure risk than MDAV when different groups of attributes are known by an intruder.

This paper is organized as follows. In Section 2 we review some basic concepts related to protection methods in general (and microaggregation in particular). In Section 3, we present our new microaggregation method called *One dimension microaggregation*. Section 4 is devoted to compare traditional MDAV microaggregation with our new microaggregation method with real data sets; we explain the ingredients of our experiments and the obtained results. Finally, Section 5 draws some conclusions and presents some future work.

2 Preliminaries

In this section, we explain some basic concepts that will be useful for the rest of the paper. Namely, we first describe the scenario where a microdata protection method is applied to preserve the privacy of the owners of some statistical data. Then, we recall one of the most used protection methods, microaggregation, and one of its heuristic variants: MDAV. Finally, we describe a way to measure the quality of a given microaggregation method, according to the levels of privacy and statistical utility that it provides.

2.1 Statistical Data Protection

A data set X can be seen as a matrix with n rows (*records*) and A columns (*attributes*), where each row contains A attributes of an individual. The attributes in a data set can be classified according to two different categories, *identifiers* or *quasi-identifiers*, depending on their capability to identify unique individuals. Among the quasi-identifier attributes, we distinguish between *confidential* and *non-confidential* ones, depending on the kind of information they contain. Because of this, we write $X = X_{id}||X_{nc}||X_c$.

In this paper, we consider the following scenario for statistical disclosure control, which was defined in [3] to compare several protection methods.

(i) Identifier attributes in X are either removed or encrypted. Therefore reduce X to $X = X_{nc}||X_c$.
(ii) Confidential quasi-identifier attributes X_c are not modified, and so we have $X'_c = X_c$; in this way, the statistical utility of the confidential attributes is completely preserved.
(iii) A microdata protection method ρ is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data is being released. This leads to a protected data set $X'_{nc} = \rho(X_{nc})$.
(iv) The released data set is $X' = X'_{nc}||X'_c = \rho(X_{nc})||X_c$.

After applying this protection procedure, the disclosure risk caused by an intruder that, first, obtains non-confidential attributes from other sources and, then, tries to re-identify entities by using record linkage methods between these external information and X_{nc} is reduced, since X_{nc} has been obfuscated by using X'_{nc} instead.

2.2 Microaggregation

As we explained before, microaggregation builds small clusters of at least k elements of A attributes and replaces the original records by the centroid of the cluster to which the records belong.

The goal of a microaggregation method is to minimize the total Sum of the Square Error

$$SSE = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)^T (x_{ij} - \bar{x}_i),$$

where c is the total number of clusters, C_i is the i-th cluster and $\bar{x}_i$ is the centroid of C_i. The restriction is $|C_i| \geq k$, for all $i = 1, \ldots, c$.

If a microaggregation method is applied to all the A attributes of the original data set X at the same time; then, the resulting protected data set X' satisfies the property of k-anonymity [18]: each protected record can correspond to at least k original records. However, in order to increase the statistical utility of the released (protected) information, statistical agencies usually split the whole data set X into blocks of a few attributes a_i ($\sum_{i=1}^{C} a_i = A$ where C is the total number of blocks), and then apply a microaggregation method to each block, independently. In this way, k-anonymity is not preserved anymore.

In the case of univariate microaggregation ($a_i = 1$), there exist polynomial time algorithms to obtain the optimal microaggregation [8]. The main drawback of univariate microaggregation is that it provides a bad level of privacy, due to its high disclosure risk [13]. However, for the multivariate case ($a_i > 1$), the problem of finding the optimal microaggregation is NP-hard [14]. For this reason, multivariate microaggregation methods are heuristic. In this paper, we recall one of the most commonly used multivariate techniques: MDAV microaggregation.

2.3 MDAV Microagregation

The MDAV (Maximum Distance to Average Vector) algorithm [4] is a heuristic algorithm for clustering records in a data set X so that each cluster is constrained to have at least k records.

MDAV works as follows. First, MDAV computes the average record $\bar{x}$ of all records in X, then, MDAV builds two clusters. In order to build them, MDAV considers the most distant record x_r to the average record $\bar{x}$ and forms a cluster around x_r (this cluster contains x_r together with the $k-1$ closest records to x_r). When the cluster is done, all the records belonging to this cluster are removed from X. Following, the most distant record x_s from record x_r is taken and a new cluster is done around x_s, again, all the records belonging to this cluster are removed. This process is repeated until all the records are assigned to one cluster. Finally, the protected data set X' is built replacing the original records in X by the centroid of the cluster to which the record belongs.

Note that this process can be done considering all the attributes in the data set at the same time, or the data set can be split into smaller blocks of attributes and MDAV is independently applied to each block. The former option ensures the k-anonymity property with a large information loss, and the second one ensures a small information loss, but k-anonymity property is not preserved any more.

2.4 Performance Measure for Microaggregation

A microdata protection method must guarantee a certain level of privacy (low disclosure risk). At the same time, since the goal is to allow third parties to perform reliable statistical computations over the released (protected) data, the

protection method must ensure somehow that the protected data is statistically close to the original one.

Therefore, we have two inversely related aspects to measure, for each microdata protection method: the *disclosure risk* (DR), which is the risk that an intruder obtains correct links between the protected and the original data; and the *information loss* (IL) caused by the protection method. When one of them increases, the other one decreases. The two extreme cases are the following ones: (i) if the original microdata is released, then information loss is zero, but the disclosure risk is maximum; (ii) if the original microdata is encrypted and then released, the disclosure risk is (almost) zero, but the information loss is maximum.

There are different measures proposed in the literature to evaluate the quality of a data protection method. Such measures can be general (for all protection methods) or specific for a given data protection method. For instance, the goal of microaggregation is to minimize the total Sum of the Square Error SSE. Since there are no optimal solutions in polynomial time to multivariate microaggregation, and the methods used are heuristic, the actual value of SSE for a given method is a measure of its quality with regards to information loss.

Regarding privacy, microaggregation provides, by definition, some level of anonymity. If the method is applied to all the attributes (a single block), then the initial parameter k indicates the achieved anonymity: for each protected record, there are at least k possible original records which can correspond to it. However, if the original data set is split into r blocks and the microaggregation method is applied to each block independently, then the final level of anonymity obviously decreases: two records which are in the same cluster for one block of attributes may be in different clusters for other blocks, which results in two different protected records.

A possible way of computing the real level of anonymity achieved by a microaggregation method is to consider the ratio between the total number n of records and the number of protected records which are different. This gives the average size of each "global cluster" in the protected data set. We denote as k' this *real anonymity* measure:

$$k' = \frac{n}{|\{x'|x' \in X'\}|}$$

It was introduced in [11] and used in other papers like [12].

3 One Dimension Microaggregation

One dimension microaggregation (Mic1D-κ) uses the same vision of data handling based on the vectorization, sorting and partitioning of all the values in the data set presented in [9]. There are several aspects that motivate these three steps:

Vectorization. The first step is vectorization. The basic idea is to gather all the values in the data set in a single vector, independently of the attribute they belong to. This way, we are ignoring the attribute semantics and, therefore, the possible correlation between two different attributes in the data set.

Sorting. Once all the values are inserted in the unique vector, it is necessary to sort them in order to minimize the SSE value when the clusters (partitions) are done. Note that sorting the values is a way of adding noise.

Partitioning. In order to ensure a certain level of privacy (k-anonymity), we propose to split the data set in several κ-partitions and to calculate the average value for each partition. Modifying the value of κ, Mic1D-κ allow us to adjust the trade-off between information loss (SSE) and disclosure risk. Note that if the data set was not sorted, κ would not have this property.

Since the ranges of values of two different attributes could differ significantly, the sorting step may not merge all the values coming from different attributes appropriately. For this reason, after the partitioning is complete, data is normalized in each partition and it is sorted and re-partitioned again. Data normalization improves the attribute merging and therefore, it is more difficult for an intruder to re-identify an individual. Finally, data in each partition is substituted by their centroid. All the steps of Mic1D-κ are represented in Figure 1.

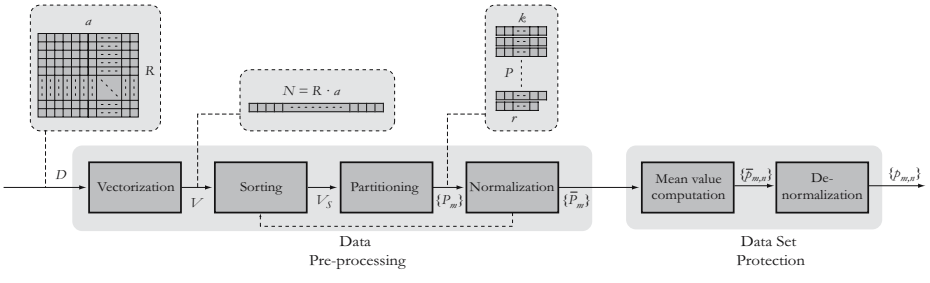


Fig. 1. Mic1D-κ schema

Formally speaking, let $\mathcal{D}$ be the original data set to be protected. We denote by R the number of records in $\mathcal{D}$. Each record consists of a numerical attributes or fields. We assume that none of the registers contain blanks. We denote by N the total number of values in $\mathcal{D}$. As a consequence, $N = R \cdot a$.

Let V be a vector of size N. Mic1D-κ treats values in the data set independently of the attribute they belong to. In other words, the concepts of record and field is ignored and the N values in the data set are placed in V. We call this process *desemantization*.

First, V is sorted increasingly. Let us denote by V_s the ordered vector of size N containing the sorted data and v_i the ith element of vector V_s, where $0 \leq i < N$.

Next, V_s is divided into smaller sub-vectors or partitions. Each sub-vector is normalized into the $[0, 1]$ interval and they are all sorted and partitioned again. We define κ, where $1 < \kappa \leq N$, as the number of values per partition. Note that if κ is not a divisor of N the last partition will contain a smaller number of values. Let P be the number of κ-partitions. We call r the number of values in the last partition where $0 \leq r < \kappa$. Therefore, $N = \kappa P + r$. If $r > 0$ the we have $P + 1$ partitions. We denote by P_m the mth partition.

Let $v_{m,n}$ be defined as the nth element of P_m:

$$\begin{cases} v_{m,n} := v_{m\kappa+n} & n = 0 \dots \kappa - 1 \quad m = 0 \dots P - 1 \\ v_{P,n} := v_{P\kappa+n} & n = 0 \dots r - 1 \end{cases}$$

For each partition P_m, the mean value of its components is computed:

$$\mu_m = \sum_{n=0}^{\kappa-1} \frac{v_{m,n}}{\kappa} \qquad m = 0 \dots P - 1 \qquad\qquad \mu_P = \sum_{n=0}^{r-1} \frac{v_{P,n}}{r}$$

where the latter expression is applied to the last partition if $r > 0$, i.e., if κ does not divide the total number of values in the data set.

The protected value $p_{m,n}$ for $v_{m,n}$ is then:

$$\begin{cases} p_{m,n} = \mu_m & n = 0 \dots \kappa - 1 \quad m = 0 \dots P - 1 \\ p_{P,n} = \mu_P & n = 0 \dots r - 1 \end{cases}$$

Finally, Mic1D-κ de-normalizes the data into the original range. The protected values are placed in the protected data set in the same place occupied by the corresponding $v_{m,n}$ in the original data set. This way, we are undoing the sorting and vectorization steps.

4 Experimental Results

We have tested Mic1D-κ and compared our results with those obtained by the MDAV algorithm, using the *Census* [19] and *Water-treatment* [10] data sets. The former was extracted using the Data Extraction System of the U.S. Census Bureau and contains 1080 records consisting of 13 numerical attributes. The latter was extracted from the UCI repository and contains 35 attributes and 380 entries. These data sets have been used in previous works [3,11] to compare different microaggregation techniques.

As shown in [12], when protecting a data set using multivariate microaggregation, the way in which the data is split to form blocks is highly relevant with regard to the degree of privacy achieved (k' value). Similarly, we have reduced both data sets to have 9 attributes, which we detail in Tables 1 and 2.

In both data sets, attributes $a1, a2$ and $a3$ are highly correlated, as well as attributes $a4, a5$ and $a6$ and attributes $a7, a8$ and $a9$. On the contrary, attributes of different blocks are non-correlated. For our experiments, when protecting data using MDAV microaggregation, we assume attributes to be split in three blocks of three attributes each. Also, we consider two situations when protecting the data sets using MDAV microaggregation: blocking correlated attributes and thus non-correlated blocks, *i.e.*, $(a1, a2, a3)$, $(a4, a5, a6)$ and $(a7, a8, a9)$; or blocking non-correlated attributes but correlated blocks, *i.e.*, $(a1, a4, a7)$, $(a2, a5, a8)$ and $(a3, a6, a9)$. Testing these two cases will let us study the impact of the choice of the attributes for the microaggregation groups, based on their correlations.

For each data set and attribute selection method, we apply MDAV microaggregation using the same parameterizations as those in previous works [11,12].

Table 1. Attribute description of the Census data set

id	Name	Description
$a1$	AGI	Adjusted gross income
$a2$	FICA	Social security retirement payroll deduction
$a3$	INTVAL	Amount of interest income
$a4$	EMCONTRB	Employer contribution for health insurance
$a5$	TAXINC	Taxable income amount
$a6$	WSALVAL	Amount: Total wage and salary
$a7$	ERNVAL	Business or farm net earnings in 19
$a8$	PEARNVAL	Total person earnings
$a9$	POTHVAL	Total other persons income

Table 2. Attribute description of the Water data set

id	Name	Description
$a1$	PH-E	Input pH to plant
$a2$	PH-P	Input pH to primary settler
$a3$	PH-D	Input pH to secondary settler
$a4$	DQO-E	Input chemical demand of oxygen to plant
$a5$	COND-P	Input conductivity to primary settler
$a6$	COND-D	Input conductivity to secondary settler
$a7$	DBO-S	Output biological demand of oxygen
$a8$	SS-S	Output suspended solids
$a9$	SED-S	Output sediments

Namely, we protect the data sets using MDAV with parameter $k = 5, 25, 50, 75, 100$ for the Census data set, and $k = 5, 10, 15, 20, 25$ for the Water-treatment data set. The selection of these values aims at covering a wide range of SSE values and, thus, studying scenarios with different *information loss* values.

For Mic1D-κ, we use $\kappa = 3000, 3200, 4000, 4400, 5000$ for the Census data set and $\kappa = 300, 500, 800, 850, 900$ for the Water-treatment data set. Note that, since Mic1D-κ *desemantizes* the data set, there is no point in considering different situations related to the correlation of the attributes and, therefore, we protect the data set just once for each parametrization. In order to make the comparison fair, we have chosen the values of κ in Mic1D-κ to obtain similar SSE values to those obtained by MDAV after protecting the data sets.

We consider that a possible intruder knows the value of three random attributes of the original data set. Different tests are performed assuming that the intruder knows different sets of three attributes. Depending on these attributes the intruder will have information coming from one or more groups. Table 3 shows all the considered possibilities.

First, we suppose that the three known attributes belong to the same MDAV microaggregated block (e.g. $(a1, a2, a3)$ in the correlated scenario or $(a1, a4, a7)$

Table 3. Different groups of variables known by the intruder

Correlated	1G	$(a1, a2, a3)$, $(a4, a5, a6)$, $(a7, a8, a9)$
	2G	$(a1, a2, a5)$, $(a1, a3, a7)$, $(a2, a3, a6)$, $(a1, a4, a5)$, $(a2, a4, a6)$ $(a5, a6, a9)$, $(a6, a7, a8)$, $(a1, a8, a9)$, $(a2, a7, a9)$
	3G	$(a1, a4, a7)$, $(a1, a5, a8)$, $(a1, a6, a9)$, $(a2, a4, a7)$, $(a2, a5, a8)$ $(a2, a6, a9)$, $(a3, a4, a7)$, $(a3, a5, a8)$, $(a3, a6, a9)$
Non-correlated	1G	$(a1, a4, a7)$, $(a2, a5, a8)$, $(a3, a6, a9)$
	2G	$(a1, a4, a5)$, $(a1, a3, a7)$, $(a4, a7, a8)$, $(a1, a2, a5)$, $(a2, a4, a8)$ $(a5, a8, a9)$, $(a3, a6, a8)$, $(a1, a6, a9)$, $(a3, a4, a9)$
	3G	$(a1, a2, a3)$, $(a1, a5, a6)$, $(a1, a8, a9)$, $(a2, a3, a4)$, $(a4, a5, a6)$ $(a4, a8, a9)$, $(a2, a3, a7)$, $(a5, a6, a7)$, $(a7, a8, a9)$

Table 4. SSE and real k' values using MDAV-k and Mic1D-κ methods assuming that different groups of variables are known by the intruder using the Census data set

		SSE	k'		
			1G	2G	3G
MDAV-k	5	64.99	5.00	1.92	1.00
	25	223.73	25.12	7.00	1.09
	50	328.31	51.43	14.66	1.41
	75	382.34	77.14	23.18	1.96
	100	428.68	108.00	35.00	3.33
Mic1D-κ	3000	32.27	8.37	9.87	5.77
	3200	89.18	11.97	13.76	8.26
	4000	129.06	20.10	22.09	13.89
	4400	310.63	23.15	26.94	17.01
	5000	738.12	72.83	76.08	55.02

Correlated attributes

		SSE	k'		
			1G	2G	3G
MDAV-k	5	58.49	5.00	1.96	1.02
	25	260.13	25.12	7.35	1.24
	50	356.47	51.43	15.86	2.05
	75	563.79	77.14	24.38	2.83
	100	721.91	108.00	36.14	4.62
Mic1D-κ	3000	32.27	5.63	8.51	8.04
	3200	89.18	8.01	11.95	11.83
	4000	129.06	13.53	19.45	19.19
	4400	310.63	16.62	23.64	22.45
	5000	738.12	59.77	67.72	67.25

Non-correlated attributes

in the non-correlated). Since the size of the three microagreggation blocks is 3, there are only three options to consider. We denote this case by 1G. Since the intruder only has access to data from one group, MDAV ensures the k-anonymity property (this is the best possible scenario for MDAV). However, note that, usually, the intruder cannot choose the attributes obtained from external sources and it might be difficult to obtain all the attributes in the same group. Second, we assume that the known attributes belong to two different MDAV microaggregated groups. There are many possible combinations of three attributes under this assumption, so nine of them were chosen randomly. We refer to this case as 2G. Finally, case 3G is defined analogously to 2G, and also

Table 5. SSE and real k' values using MDAV-k and Mic1D-κ methods assuming that different groups of variables are known by the intruder using the Water data set

		SSE	k'		
			1G	2G	3G
MDAV-k	5	28.18	5.09	1.94	1.00
	10	46.14	10.00	3.14	1.01
	15	72.03	15.20	4.42	1.01
	20	94.24	20.00	5.75	1.04
	25	114.56	25.33	7.28	1.10
Mic1D-κ	300	32.67	1.62	1.51	1.10
	500	65.89	3.25	3.39	1.76
	800	80.95	7.87	7.55	4.67
	850	132.13	9.65	10.03	6.65
	900	255.64	12.95	13.61	9.14

Correlated attributes

		SSE	k'		
			1G	2G	3G
MDAV-k	5	69.51	5.00	2.03	1.03
	10	126.21	10.00	3.55	1.16
	15	173.96	15.20	5.28	1.39
	20	259.07	20.00	7.00	1.53
	25	247.58	25.33	9.22	1.91
Mic1D-κ	300	32.67	1.11	1.35	1.35
	500	65.89	1.78	2.58	2.63
	800	80.95	4.74	7.17	6.88
	850	132.13	6.54	9.77	8.67
	900	255.64	9.07	14.52	11.71

Non-correlated attributes

nine possibilities of known attributes are considered. Note that, in both scenarios 2G and 3G, k-anonymity is not ensured by MDAV. Note also that, if the intruder had more than three attributes, it would not be possible to consider 1G. We are considering the case were the intruder only has three attributes to study a scenario were MDAV can still preserve k-anonymity.

The first column of Tables 4 and 5 presents the SSE values for all the parameterizations and situations described before. Note that the range of SSE covered by the two methods is similar, so this allows us to compare the disclosure risk of both methods fairly. For all these scenarios, we compute k' and the mean of all the k' values in each situation is presented in the second, third and fourth columns. Note that, whereas MDAV is affected by the fact that the chosen attributes are correlated or not, this effect is not noticeable using $Mic1D - \kappa$. Specifically, when the attributes in a group are not correlated, the information loss (SSE) using MDAV tends to be increased since we are trying to collapse the records in a single value, using three independent attributes or dimensions. Nevertheless, this effect can be neglected with our technique since, thanks to the data preprocessing, the whole microaggregation process is performed on a single dimension (vector of values), the semantics of attributes are ignored and the effect caused by attribute correlations is avoided.

Results show that, in general, $Mic1D - \kappa$ achieves lower disclosure risk levels (larger values of k') than those achieved by MDAV for similar information loss (SSE), especially when the attributes chosen come from different microaggregated groups (2G and 3G), which is the most common case. When the intruder has access to the three attributes coming from a single microaggregated group, MDAV presents k' values which are similar or, in some cases, even larger than those obtained by $Mic1D - \kappa$ (comparing cases with similar SSE). This is normal since MDAV preserves the k-anonymity in this case. However, in the remaining scenarios

(2G and 3G), that represent most of the cases, $Mic1D-\kappa$ achieves larger k' values than those obtained by MDAV when similar SSE values are compared.

5 Conclusions and Future Work

In this paper, we have presented a new type of microaggregation called *One Dimension microaggregation*. This microaggregation method significantly diminish the problem of attribute selection in multivariate microaggregation achieving in general a higher level of privacy than that obtained by MDAV, one of the most well-known microaggregation methods. This is specially true as, from the attributes known by the intruder, the number of these coming from different microaggregation groups of MDAV increases.

As future work, we plan to further study One Dimension microaggregation using other information loss and disclosure risk measures. We also plan to develop and implement a method for vector partitioning which considers the SSE value when the partitions are done so that we can reduce the SSE value of our method and, therefore, the information loss.

All in all, in this paper we show that microaggregation can be a very useful method for the anonymization of complex records containing a large number of attributes, when it is combined with the data preprocessing proposed in our work.

Acknowledgments

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) and by the Government of Catalunya (grant 2005-SGR-00093) is acknowledged.

The authors from DAMA-UPC want to thank Generalitat de Catalunya for its support through grant number GRE-00352 and Ministerio de Educación y Ciencia of Spain for its support through grant TIN2006-15536-C02-02. Josep L. Larriba-Pey wants to thank UPC, Generalitat de Catalunya and Ministerio de Educación y Ciencia for his I3 grant.

References

1. Adam, N.R., Wortmann, J.C.: Security-control for statistical databases: a comparative study. ACM Computing Surveys 21, 515–556 (1989)
2. Domingo-Ferrer, J., Torra, V.: Disclosure control methods and information loss for microdata. In: [6], pp. 91–110 (2001)
3. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: [6], pp. 111–133 (2001)
4. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. on Knowledge and Data Engineering 14(1), 189–201 (2002)
5. Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J.M., Sebé, F.: Efficient multivariate data-oriented microaggregation. The VLDB Journal 15, 355–369 (2006)

6. Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.): Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. Elsevier Science, Amsterdam (2001)
7. Felsö, F., Theeuwes, J., Wagner, G.: Disclosure Limitation in Use: Results of a Survey. In: [6], pp. 17–42 (2001)
8. Hansen, S., Mukherjee, S.: A Polynomial Algorithm for Optimal Univariate Microaggregation. Trans. on Knowledge and Data Engineering 15(4), 1043–1044 (2003)
9. Medrano-Gracia, P., Pont-Tuset, J., Nin, J., Muntés-Mulero, V.: Ordered Data Set Vectorization for Linear Regression on Data Privacy. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 361–372. Springer, Heidelberg (2007)
10. Murphy, P., Aha, D.W.: UCI Repository machine learning databases. University of California, Department of Information and Computer Science, Irvine (1994), http://www.ics.uci.edu/~mlearn/MLRepository.html
11. Nin, J., Herranz, J., Torra, V.: Attribute Selection in Multivariate Microaggregation. In: Post-Proc. of 11th ACM International Conference on Extending Database Technology (2008)
12. Nin, J., Herranz, J., Torra, V.: How to group attributes in multivariate microaggregation. Int. J. on Uncertainty, Fuzziness and Knowledge-Based Systems 16(1), 121–138 (2008)
13. Nin, J., Torra, V.: Analysis of the Univariate Microaggregation Disclosure Risk (submitted, 2007)
14. Oganian, A., Domingo-Ferrer, J.: On the Complexity of Optimal Microaggregation for Statistical Disclosure Control. Statistical J. United Nations Economic Commission for Europe 18(4), 345–354 (2000)
15. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI Intl. Tech. Rep. (1998)
16. Sande, G.: Exact and approximate methods for data directed microaggregation in one or more dimensions. Int. J. of Unc., Fuzz. and Knowledge Based Systems 10(5), 459–476 (2002)
17. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. of Unc., Fuzz. and Knowledge Based Systems 10(5), 571–588 (2002)
18. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. of Unc., Fuzz. and Knowledge Based Systems 10(5), 557–570 (2002)
19. U.S. Census Bureau, Data Extraction System (1990), http://www.census.gov/

A Shared Steganographic File System with Error Correction

Josep Domingo-Ferrer and Maria Bras-Amorós

Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{josep.domingo,maria.bras}@urv.cat

Abstract. Steganographic file systems are file systems where the location and even the existence of files are unknown to the users not having stored them. If the file system can be written to by several users, a user may inadvertently damage the files stored by other users. In this paper, solutions to the collision problem are proposed which rely on error-correcting codes. The storage efficiency and the privacy offered by the proposed protocols are analytically assessed.

Keywords: Information privacy and security, steganographic file systems, error-correcting codes.

1 Introduction

Steganographic file systems [1] were introduced a decade ago as file systems where the location and even the existence of files are unknown to the users not having stored them. This feature is a problem when the file system is a shared one that can be written to by several users: as a result of a (unknown) collision, a user may inadvertently damage the files stored by other users [5].

There are two simplistic approaches to deal with the collision problem:

- *Privacy reduction.* A possible solution is to reduce the risk of collisions between users by reducing the freedom of each user for placing her file in the system. However, the smaller the freedom, the smaller is privacy: the location of the files of a user becomes easier to guess by other users or by external intruders.
- *Efficiency reduction.* If the total size of user files stored in the steganographic file system is less than the size of the file system by several orders of magnitude, collisions are unlikely to occur. However, this entails a very inefficient storage use.

1.1 Contribution and Plan of This Paper

Intermediate solutions between the two simplistic approaches above are explored in this paper. We present a shared steganographic file system protocol whose aim is to offer a tradeoff between privacy and storage efficiency.

V. Torra and Y. Narukawa (Eds.): MDAI 2008, LNAI 5285, pp. 227–238, 2008.

Section 2 describes a collision-resistant file system where user files are stored as segments of straight lines in a two-dimensional space. Section 3 analyzes the effects of discretization on privacy and storage efficiency. Section 4 sketches conclusions and future work.

2 A Collision-Resistant Shared Steganographic File System in the Plane

Consider a steganographic shared file system FS whose storage space consists of a publicly readable bit matrix

$$\mathbf{M} = \{m_{ij} : 0 \leq i, j \leq N-1\}$$

of size $N \times N$, for a large positive integer N. Let the bits m_{ij} be randomly initialized (*e.g.* a bit can be set to 0 or 1 with probability 1/2). We assume that the file system FS is trusted to keep secret the locations where files are recorded.

A first naive protocol allowing a user U to store a file in $\mathbf{M}$ is given below. The protocols in this paper all use pseudorandom number generators seeded by an initial value only known to each user, so that the user can reproduce at any moment the pseudorandom values she has previously generated (*e.g.* to recover from the file system the files she has stored in it). Also, the protocols in this paper assume that the user is able to privately send information to the file system (*e.g.* by encrypting it under the file system's public key).

Protocol 1

1. *Let $F = (f_0, \cdots, f_{l-1})$ be a bit vector representing the file that U wants to store in FS. Let $\{S_i^F : i \geq 0\}$ be a pseudorandom sequence generated by U specifically for file F. For $i = 0$ to $l-1$, U computes $E = (e_0, \cdots, e_{l-1})$, where $e_i := f_i \oplus S_i^F$ and $\oplus$ is addition modulo 2.*
2. *U pseudorandomly selects a slope $a \in \mathbb{R}$, an integer intercept b and an integer offset c, the latter randomly drawn from $\{0, \cdots, N-1\}$. (In this protocol and in the remaining ones of this paper, intercepts and slopes are sampled from a uniform distribution, and slopes are selected by uniformly sampling an angle from $[-\pi/2, \pi/2)$ and then taking as slope the tangent of that angle.)*
3. *U privately sends a, b, c and E to FS.*
4. *For $i = 0$ to $l-1$, FS stores e_i in the component (x_i, y_i) of $\mathbf{M}$, where if $|a| \geq 1$ or $a = 0$*

$$x_i = (i + c) \bmod N$$

$$y_i = [a(i + c) + b] \bmod N$$

and if $|a| \in (0, 1)$

$$x_i = [(1/a)(i + c) + b] \bmod N$$

$$y_i = (i + c) \bmod N$$

with $[\cdot]$ *being the integer rounding operator. Thus,* E *is stored as a segment of a pseudorandomly chosen straight line. (If slopes* $|a| \in (0,1)$ *were not handled separately, one would come up with many points sharing the same ordinate* y_i *due to rounding, which could result in a lot of overlap at the crossing of lines with slopes under 1.) In order to prevent* E *from wrapping around itself regardless of the slope* a*, one must require* $l \leq N$*.*

To recover file F from FS, a user U must know (a, b, c, l) and the pseudorandom sequence $\{S_i^F : i \geq 0\}$. Thanks to the use of the pseudorandom sequence $\{S_i^F | i \geq 0\}$, Protocol 1 fulfills the standard requirement of steganographic file systems that stored files should be indistinguishable from the random, unused positions. As an additional precaution, FS is assumed to randomly tweak straight segments formed by unused bit positions from time to time; those fake file insertions thwart intruders from trying to infer the location of files by observing the changes in the bit matrix.

A problem when using Protocol 1 in a shared file system is that the segments corresponding to files of different users might cross each other. This is a concern only for shared file systems: if there was a single user, one could assume she can select the parameters for her files so that no crossing occurs. If a file F' is stored after a file F and the segments of both files cross each other, the bit of F' at the crossing position overrides the bit of F. With probability $1/2$, this causes an error in file F. This has two undesirable effects, which we next describe along with possible solutions:

- Errors damage the integrity of the stored files. Using error-correcting codes (ECC) to encode files before storage appears as a natural way to mitigate this problem.
- Even if errors can be corrected, their very existence may leak to a user the location of the files belonging to other users. Indeed, assume that a user U stores two files F_1 and F_2 in the file system and then keeps retrieving both files very often in order to detect any simultaneous appearance of errors in them due to their being crossed by a new file. Now, if a user U' stores a new file F' that crosses F_1 and F_2 and both crossings cause simultaneous single-bit errors in F_1 and F_2, respectively, U can infer that F' lies on the straight line connecting both erroneous bit positions. A possible way to repair this weakness would be to require that the bits of each file be randomly shuffled by the file system using a secret permutation different for each file; but this would require the help of the file system for file retrieval, which would be a disadvantage with respect to Protocol 1 above. A better option is to divide $\mathbf{M}$ into several tiles and to split a file into fragments and store each fragment in a different tile, which makes it difficult for an intruder U to locate all the fragments of a file by merely watching the crossing errors.

From now on, by a $[n, k]$ ECC we mean an error correcting code with length n and dimension k. Its transmission rate is $R = \frac{k}{n}$ (*e.g.* see [4]). If the minimum distance of the code is d, then its correction capability is $t := \lfloor \frac{d-1}{2} \rfloor$.

Protocol 2 below incorporates the above solutions to deal with crossing errors. The idea is to make tiling compatible with the storage of large files, by using several tiles to store a file: thus, only a fragment of the file is stored in a particular tile. The $N \times N$ bit matrix $\mathbf{M}$ is considered to be divided into h^2 tiles of size $N/h \times N/h$ for some integer divisor h of N. For each tile $T_{r,s}$, the file system maintains a global counter $\nu_{r,s}$ initially set to 0 that counts the number of file fragments stored in the tile.

Protocol 2

1. *Let $F = (f_0, \cdots, f_{l-1})$ be a bit vector representing the file that U wants to store in FS.*
2. *U encodes F using a binary $[n, k]$ ECC with $n \leq N/h$, to obtain an encoded file $E = (e_0, \cdots, e_{m-1})$.*
3. *U generates a pseudorandom sequence $\{S_i^E : i \geq 0\}$ specifically for file E and computes $E' = (e'_0, \cdots, e'_{m-1})$, where $e'_i := e_i \oplus S_i^E$ for $i = 0$ to $m - 1$.*
4. *U computes a pseudorandom enumeration of the tiles whose global counter is less than $t + 1$; each tile appears in the enumeration a number of times equal to the difference between $t + 1$ and its global counter (in general, repetitions of the same tile appear in different positions of the pseudorandom enumeration). Let the enumeration be $T_{r_0,s_0}, T_{r_1,s_1}, \cdots$ Each file fragment in a tile can contain up to N/h bits of E', so as many tiles from the enumeration will be taken in turn as needed to store the m bits of E'. If the number h' of necessary tiles is greater than the number of tiles in the enumeration, then exit the Protocol (there is insufficient storage to hold E'). Note that, initially, all h^2 tiles can be used,* so m can be as large as $Nh(t + 1)$ bits, which implies a maximum l as large as $\lfloor \frac{Nh(t+1)}{n} \rfloor k$; *the maximum size of storable new files will decrease as the number of tiles holding already $t + 1$ file fragments increases.*
5. *For $0 \leq j < h'$ user U pseudorandomly selects a slope $a_{r_j,s_j} \in \mathbb{R}$, an integer intercept b_{r_j,s_j} and an integer offset c_{r_j,s_j}, with the latter two randomly drawn from $\{0, \cdots, N/h - 1\}$.*
6. *U privately sends to FS the indexes of the h' chosen tiles and the slopes, intercepts and offsets chosen for each tile. U publicly sends E' to FS.*
7. *For $0 \leq j < h'$ the file system FS does:*
   (a) *Let $\nu_{r_j,s_j} = \nu_{r_j,s_j} + 1$;*
   (b) *For $i = 0$ to $N/h - 1$, store the bit $e'_{jN/h+i}$ of E in the component (x_i, y_i) of $\mathbf{M}$, where if $|a| \geq 1$ or $a = 0$*

$$x_i = r_j N/h + ((i + c_{r_j,s_j}) \bmod (N/h))$$

$$y_i = s_j N/h + ([a_{r_j,s_j}(i + c_{r_j,s_j}) + b_{r_j,s_j}] \bmod (N/h))$$

and if $|a| \in (0, 1)$

$$x_i = r_j N/h + ([(1/a_{r_j,s_j})(i + c_{r_j,s_j}) + b_{r_j,s_j}] \bmod (N/h))$$

$$y_i = s_j N/h + ((i + c_{r_j,s_j}) \bmod (N/h))$$

If slopes $|a| \in (0, 1)$ were not handled separately in the above protocol, one would come up with many points sharing the same ordinate y_i due to rounding, which could result in a lot of overlap at the crossing of lines with slopes under 1. The following lemma is a quantification of storage efficiency.

Lemma 1. *The maximum storage efficiency achievable by Protocol 2 when using a $[n, k]$ ECC is $\lfloor \frac{Nh(t+1)}{n} \rfloor \frac{k}{N^2}$. Consequently, if $\frac{Nh(t+1)}{n} \simeq \lfloor \frac{Nh(t+1)}{n} \rfloor$ then the maximum storage efficiency is approximately $(t+1)Rh/N$, where R is the transmission rate of the code.*

Proof: The optimum case is the one mentioned in Step 4 of Protocol 2: a file of size $\lfloor \frac{Nh(t+1)}{n} \rfloor k$ is stored across the tiles. By diving this file size by the total storage available N^2, we get the efficiency above. □

In the above protocol, tiles should stay large enough so that they can still be viewed as plane regions and bits can be viewed as "points" in those regions: *e.g.* if a tile consists of very few bits, rounding when computing straight segments causes a lot of crossings to occur (see Section 3.2 below for an analysis of the impact of multibit crossings on efficiency).

As to privacy, in Protocol 2 tiles only contain fragments of a file, which is thus harder to locate. Indeed, to locate a file an intruder needs to determine which tiles store fragments of the file and, within each of those tiles, where does the line storing the corresponding fragment lie. See Section 3.1 below for a discussion on the difficulty of guessing a specific line within a certain tile.

The price paid for the above advantages of Protocol 2 is that the user needs to keep more information to recover a file than in Protocol 1: h' slopes, intercepts and offsets (instead of a single slope, intercept and offset required by the previous protocols).

Example 1. Consider a shared steganographic file system with a bit matrix $\mathbf{M}$ of size $2^{20} \times 2^{20}$ (1 Terabit). Divide $\mathbf{M}$ into $2^5 \times 2^5$ tiles of $2^{15} \times 2^{15}$ bits each. Consider the primitive BCH code of length $2^{15} - 1$ over $\mathbb{F}_2$ with designed correction capability equal to $t = 10$ (i.e. designed minimum distance equal to 21). Its dimension is 32617. This means that within each tile we can encode up to 32617 bits of $t + 1 = 11$ file fragments as 11 codewords of length $2^{15} - 1$. If these codewords are inserted in the file system and they only cross each other at one bit position, it will be possible to correct at retrieval time any error in any of the 11 file fragments that is due to crossings. A total of 1024×11 file fragments with at most 32617 bits each can be inserted in $\mathbf{M}$. In this case, the maximum storage efficiency is approximately $0.000334146 \simeq 3 \cdot 10^{-4}$.

Alternatively, we can also divide $\mathbf{M}$ into $2^{10} \times 2^{10}$ tiles of $2^{10} \times 2^{10}$ bits each. Consider the primitive BCH code of length $2^{10} - 1$ over $\mathbb{F}_2$ with designed correction capability equal to $t = 10$ (i.e. designed minimum distance equal to 21). Its dimension is 923. This means that within each block we can encode 11 file fragments with at most 923 bits each as 11 codewords of length $2^{10} - 1$. If these codewords are inserted in the file system and they only cross each other at one bit position, it will be possible to correct at retrieval time any error in any

of the 11 stored file fragments that is due to crossings. A total of $2^{20} \times 11$ file fragments with at most 923 bits can be inserted in $\mathbf{M}$. In this case, the maximum storage efficiency is approximately $0.0096921 \simeq 10^{-2}$. □

3 The Effects of Discretization on Privacy and Efficiency

In Section 2, we have used the idealization that the bit matrix $\mathbf{M}$ can be regarded as a plane, where files are stored as straight lines. In fact, $\mathbf{M}$ is a grid, so files are stored as near-straight lines with discretized slopes. This has some practical consequences:

- Unlike in a tile in a continuous plane, in an $N \times N$ grid, the maximum length of a straight line as we define it can be no more than N regardless of its slope, which is consistent with the limitation on the length of the stored files in the above protocols.
- In a grid, the range of possible values for the slopes and the intercepts is finite [3], which has privacy implications: the uncertainty of an intruder about the location of the line storing a particular file is finite.
- In a grid, two discretized "straight" lines may cross each other in more than one bit position. This has implications for efficiency: there may be more than one error caused by the crossing of two files, which further limits the number of storable files in an error-free manner with respect to the continuous idealization used in Section 2.

In the next subsections, we analyze the above mentioned privacy and efficiency implications.

3.1 Discretization and Privacy

The protocols above encrypt the file by adding a pseudorandom sequence to it in order to conceal its redundancy in front of an intruder. However, the intruder could blindly (*i.e.* randomly) try to guess the line segment where a file or file fragment is stored. If she succeeded at that, she could for example tweak all bits along that line to destroy the file or file fragment; or she could attempt its decryption. Therefore, the intruder's uncertainty about the slope and the intercept of the line are measures of privacy.

For the sake of clarity, we make the following simplifications:

- We will initially assume that no tiling is used and that straight lines are stored in the entire $N \times N$ bit matrix $\mathbf{M}$. (To adapt the discussion for the case of tiling, the length N/h of the tile side must be used instead of N.)
- We will assume that the length of the file is the maximum value N. This is the worst case for privacy, because for files with maximum length, the intruder does not need to worry about the file length l and the offset c.

If $\mathbf{M}$ is an $N \times N$ grid, the intercept b is an integer value between 0 and $N-1$, where all values in the range have the same probability from the intruder's

Table 1. Discretized slopes and their probabilities in an $N \times N$ bit matrix

Index i	Slope $\hat{a}_i$	Probability $p(\hat{a}_i)$
0	$-\infty$	$(\arctan(-2(N-1)) + \pi/2)/\pi$
1	$-(N-1)$	$(\arctan(-2(N-1)/3) - \arctan(-2(N-1)))/\pi$
2	$-(N-1)/2$	$(\arctan(-2(N-1)/5) - \arctan(-2(N-1)/3))/\pi$
3	$-(N-1)/3$	$(\arctan(-2(N-1)/7) - \arctan(-2(N-1)/5))/\pi$
...	...	...
$N-1$	-1	$(\arctan(-(N-3/2)/(N-1)) - \arctan(-(N-1)/(N-3/2)))/\pi$
N	$-(N-2)/(N-1)$	$(\arctan(-(N-5/2)/(N-1)) - \arctan(-(N-3/2)/(N-1)))/\pi$
$N+1$	$-(N-3)/(N-1)$	$(\arctan(-(N-7/2)/(N-1)) - \arctan(-(N-5/2)/(N-1)))/\pi$
...	...	...
$2N-2$	0	$(\arctan(1/(2(N-1))) - \arctan(-1/(2(N-1))))/\pi$
$2N-1$	$1/(N-1)$	$(\arctan(3/(2(N-1))) - \arctan(1/(2(N-1))))/\pi$
$2N$	$2/(N-1)$	$(\arctan(5/(2(N-1))) - \arctan(3/(2(N-1))))/\pi$
...	...	...
$3N-3$	1	$(\arctan((N-1)/(N-3/2)) - \arctan((N-3/2)/(N-1)))/\pi$
$3N-2$	$(N-1)/(N-2)$	$(\arctan((N-1)/(N-5/2)) - \arctan((N-1)/(N-3/2)))/\pi$
...	...	...
$4N-3$	$N-1$	$(\arctan(2(N-1)) - \arctan(2(N-1)/3))/\pi$
$4N-4$	$+\infty$	$(\pi/2 - \arctan(2(N-1)))/\pi$

viewpoint. Thus, Shannon's entropy can be used to measure the intruder's uncertainty about the intercept as

$$H(b) = \log_2 N \tag{1}$$

The analysis for the discretized slope $\hat{a}$ is a bit more complex. We will give a lower bound for the entropy $H(\hat{a})$. A subset of the possible slopes in an $N \times N$ matrix is listed in the second column of Table 1: those $4N-3$ slopes are obtained when a straight segment starting at one corner of the $N \times N$ grid successively touches the bit positions in the opposite edges of the grid (similar to the hand of a clock touching the marks of the seconds). If these were the only possible slopes, from the intruder's point of view, the probability $p(\hat{a})$ of each discretized slope $\hat{a}$ is proportional to the size of the fraction of the angular range $[-\pi/2, \pi/2)$ such that the continuous slopes with angles in that fraction round to $\hat{a}$. For example,

$$p(+\infty) = (\pi/2 - \arctan((N-1)/(1/2)))/\pi$$

$$p(N-1) = (\arctan((N-1)/(1/2)) - \arctan((N-1)/(3/2)))/\pi$$

and so on (see third column of Table 1). In this way, the intruder's uncertainty on the slope can be lower-bounded as

$$H(\hat{a}) \geq -\sum_{i=0}^{4N-4} p(\hat{a}_i) \log_2 p(\hat{a}_i) \tag{2}$$

The following conclusions on privacy can be drawn:

- The overall privacy can be measured as the joint entropy $H(\hat{a}, b) = H(\hat{a}) + H(b)$, where additivity holds because $\hat{a}$ and b are independently selected.

- As it could be expected, both $H(b)$ and the lower bound for $H(\hat{a})$ increase with N. Since using tiles instead of the entire matrix involves replacing N with N/h, the privacy of the slope and the intercept within a tile is reduced.
- In Protocol 2, a file is stored across several tiles, each with its own slope and intercept. Assuming the worst case in which the intruder has managed to determine the h' tiles holding the file, the intruder's uncertainty on the tile intercepts and slopes can be measured by the following joint entropies

$$H(b_{r_0,s_0},\cdots,b_{r_{h'-1},s_{h'-1}}) = \sum_{i=0}^{h'-1} H(b_{r_i,s_i}) \tag{3}$$

$$H(\hat{a}_{r_0,s_0},\cdots,\hat{a}_{r_{h'-1},s_{h'-1}}) = \sum_{i=0}^{h'-1} H(\hat{a}_{r_i,s_i}) \tag{4}$$

where we have used that the slopes and intercepts are independently chosen for each tile.

3.2 Discretization and Efficiency

Two randomly chosen straight lines over the plane cross each other with probability 1 and the crossing consists of a single point. However, two discretized straight lines over an $N \times N$ grid may cross each other at more than one point. We analyze in this section the implications of this fact. For simplicity, we assume that the file system consists of a single tile (no tiling); the adaptation to several tiles is straightforward.

The first thing to note is that two discretized straight lines may have several crossings because of the modular operations. See the left-hand side of Figure 1 for an illustration. The discretized versions of $y = x$ (black dots) and $y = 5x$ (white dots) are depicted on an $N \times N$ grid, where $N = 2^4$; the number of crossings is four.

Secondly, the overlap at each crossing can consist of several bits, depending on the slopes of both lines. See the right-hand side of Figure 1. There, the discretized versions of $y = 10x/3$ and $y = 16x/5$ are depicted; there is a single crossing which involves four bits.

Analytically counting the expected number of crossings and the number of overlaps per crossing is by no means straightforward. We refer the reader to [2] for a preliminary discussion of this problem. For the sake of pragmatism, we have chosen here a simulation approach.

For $N \in \{2^i : i = 9,\cdots,21\}$ and $t \in \{1,\cdots,9\}$ we have conducted the following experiment:

1. Repeat 5000 times
   (a) Throw $t+1$ random digital straight lines of length N like the ones described in Protocol 1 into an $N \times N$ bit matrix;
   (b) Count the number $x_{N,t}$ of overlaps between the first line thrown and the subsequent t lines;

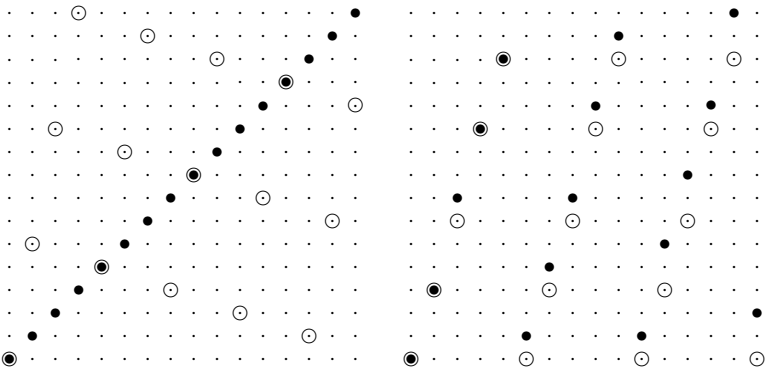


Fig. 1. Crossings of two discretized straight lines. Left, two lines with several crossings. Right, two lines with one crossing involving several bits.

2. Compute the histogram of the relative frequencies of the random variable $X_{N,t}$ modeling the number of overlaps;
3. Compute the expected value $E(X_{N,t})$ of the number of overlaps;
4. Compute the variance $\sigma^2_{X_{N,t}}$ of the number of overlaps;

For each choice of N and t, we are interested in finding a value t' such that $P(X_{N,t} \leq t') \approx 1$. We make the simplifying assumption that successive overlaps in a line occur independently. In this case, when throwing two discretized straight lines, the probability that a bit in the first line thrown is "trodden" by the second line can be estimated as $p = E(X_{N,1})/N$. When throwing $t+1$ lines, the probability that a bit in the first line is trodden by any of the subsequent t lines can be estimated as $p' = 1 - (1-p)^t$. Now, the number of bits in the first line that are trodden by any of the subsequent t lines can be modeled as a binomial random variable with N trials and success probability p'. Therefore, under the above independence assumption, $E(X_{N,t})$ can be approximated as

$$Np' \tag{5}$$

and $\sigma^2_{X_{N,t}}$ can be approximated as

$$Np'(1-p') \tag{6}$$

In Figure 2, for $N = 2^9$ and $N = 2^{12}$ and several values of t we depict $E(X_{N,t})$ and $E(X_{N,t}) + 3\sigma^2_{X_{N,t}}$, as well as their approximations resulting from Expressions (5) and (6). Figure 3 is analogous for 2^{15} and 2^{18}, respectively. Two observations are in order here:

- The experimental results obtained show that over 99% of the area in the histograms of $X_{N,t}$ for all N and t tried lies left of $E(X_{N,t}) + 3\sigma^2_{X_{N,t}}$.
- The independence-based approximations resulting from Expressions (5) and (6) overestimate the corresponding empirical magnitudes for all t tried, except $t = 1$.

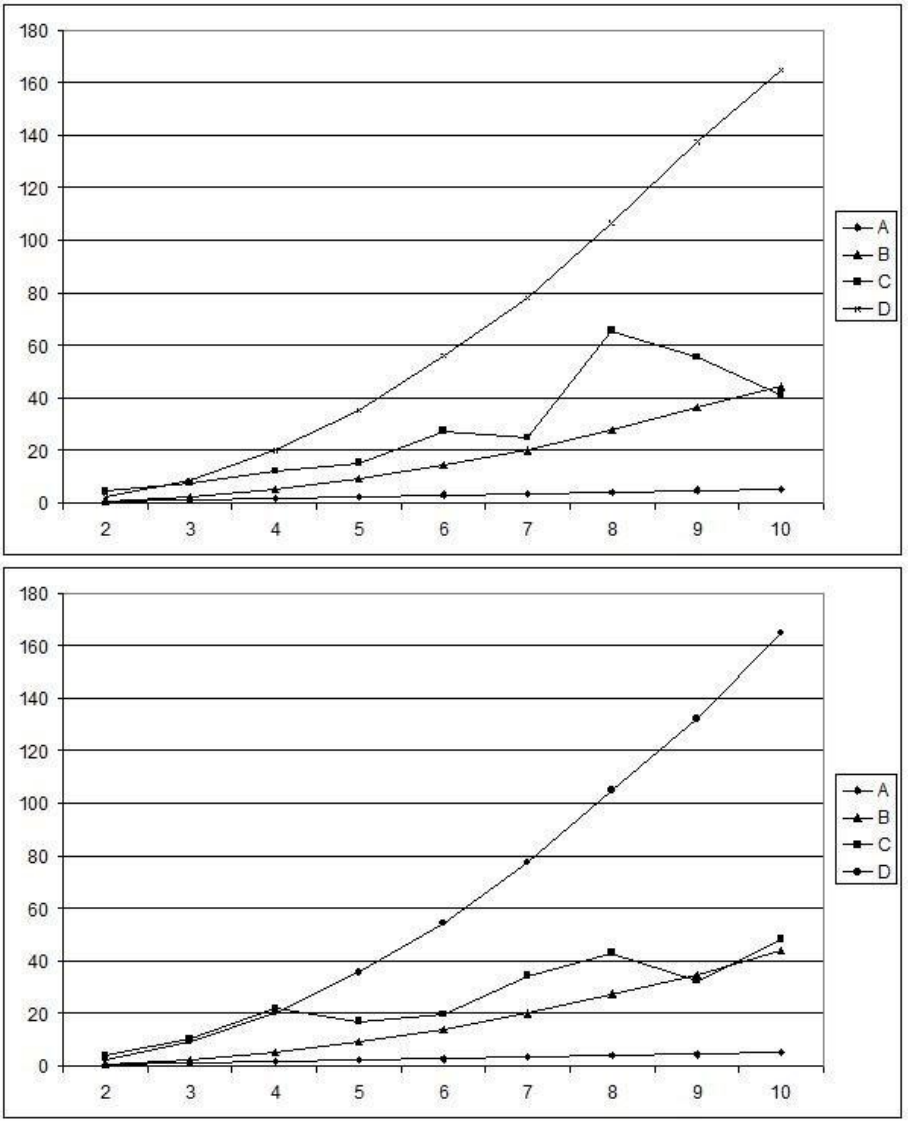


Fig. 2. Top: For $N = 2^9$ and several values of t (abscissae), A) Empirical $E(X_{N,t})$; B) Approximation Np'; C) Empirical $E(X_{N,t})+3\sigma^2_{X_{n,t}}$; D) Approximation $Np'+3Np'(1-p')$. Bottom: same for $N = 2^{12}$ and several values of t.

Therefore, to adapt protocols in Section 2 and Lemma 1 on efficiency to the real situation of discrete lines with multibit overlaps we must use a binary ECC with error-correction capability t' which, with probability almost one, is greater than the number of errors caused by overlaps in any stored file. It follows from the empirical discussion above that a suitable choice when $t > 1$ is

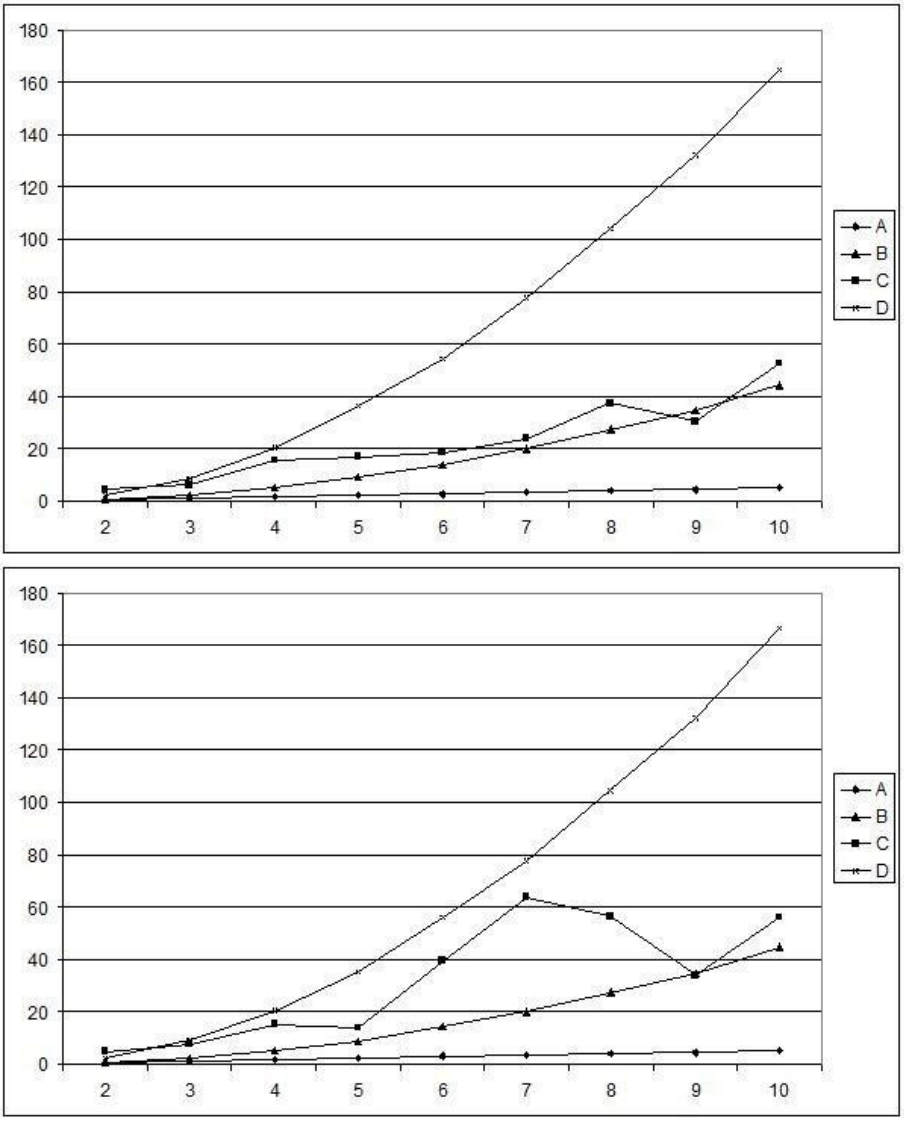


Fig. 3. Top: For $N = 2^{15}$ and several values of t (abscissae), A) Empirical $E(X_{N,t})$; B) Approximation Np'; C) Empirical $E(X_{N,t})+3\sigma^2_{X_{n,t}}$; D) Approximation $Np'+3Np'(1-p')$. Bottom: same for $N = 2^{18}$ and several values of t.

$$t' := Np' + 3Np'(1 - p') \tag{7}$$

To use Expression (7), only $E(X_{N,1})$ needs to be computed empirically. For $t = 1$, a suitable choice is

$$t' := E(X_{N,1}) + 3\sigma^2_{X_{N,1}}$$

Note that we are placing ourselves in the worst case in which every overlap causes one error (the actual probability of an overlap causing an error is 1/2). As one would expect, usually $t' > t$, which decreases storage efficiency with respect to the continuous idealization of Section 2 because a t'-correcting code will normally have a lower transmission rate than a t-correcting code.

4 Conclusion

This contribution has presented protocols for storing files in a shared steganographic file system. Their novelty is that they deal with the errors caused by successive file insertions. The privacy and the storage efficiency offered by the proposed approach have been quantified.

Future work will include finding alternative ways to store files which, without degrading privacy, are more storage-efficient than straight lines and/or can guarantee 100% correction probability.

Acknowledgments and Disclaimer

Thanks go to Markku Saarinen for providing the original motivation for this work and suggesting the use of ECC in shared steganographic file systems over the plane. We are also indebted to Francesc Sebé, Josep M. Mateo-Sanz and an anonymous reviewer for their comments. We also wish to thank Glòria Pujol for her help with the simulation work. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Anderson, R., Needham, R., Shamir, A.: The steganographic file system. In: Aucsmith, D. (ed.) Information Hiding, 2nd International Workshop, Portland, Oregon, USA (1998)
2. Bras-Amorós, M., Domingo-Ferrer, J.: On overlappings of digital straight lines. In: Jornadas de Matemática Discreta y Algorítmica, Lleida, Catalonia (2008)
3. Koplowitz, J., Lindenbaum, M., Bruckstein, A.: The number of digital straight lines on an $N \times N$ grid. IEEE Transactions on Information Theory 36(1), 192–197 (1990)
4. Roth, R.M.: Introduction to Coding Theory. Cambridge University Press, Cambridge (2006)
5. Zhou, X.: Steganographic File System. Ph.D thesis, School of Computing, National University of Singapore (2005)

Author Index