Vicenç Torra
Yasuo Narukawa
Marc Daumas (Eds.)

LNAI 6408

Modeling Decisions for Artificial Intelligence

7th International Conference, MDAI 2010
Perpignan, France, October 2010
Proceedings

Springer

Lecture Notes in Artificial Intelligence 6408

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

Subseries of Lecture Notes in Computer Science

Vicenç Torra Yasuo Narukawa
Marc Daumas (Eds.)

Modeling Decisions for Artificial Intelligence

7th International Conference, MDAI 2010
Perpignan, France, October 27-29, 2010
Proceedings

Springer

Volume Editors

Vicenç Torra
IIIA-CSIC
Campus UAB s/n, 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

Marc Daumas
Université de Perpignan
Tecnosud – Rambla de la thermodynamique
66100 Perpignan, France
E-mail: marc.daumas@univ-perp.fr

Library of Congress Control Number: 2010935676

CR Subject Classification (1998): I.2, H.3, H.4, F.1, H.2.8, J.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-16291-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16291-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

This volume contains papers presented at the 7th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2010), held in Perpignan, France, October 27–29. This conference followed MDAI 2004 (Barcelona, Catalonia, Spain), MDAI 2005 (Tsukuba, Japan), MDAI 2006 (Tarragona, Catalonia, Spain), MDAI 2007 (Kitakyushu, Japan), MDAI 2008 (Sabadell, Catalonia, Spain), and MDAI 2009 (Awaji Island, Japan) with proceedings also published in the LNAI series (Vols. 3131, 3558, 3885, 4617, 5285, and 5861).

The aim of this conference was to provide a forum for researchers to discuss 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 12 different countries, from Europe, Asia, Australia and Africa, 25 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 CNRS: Centre National de la Recherche Scientifique, the Université de Perpignan Via Domitia, the ELIAUS: Laboratoire Electronique Informatique Automatique Systèmes, IMERIR: Ecole d'Ingénierie Informatique et Robotique, 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).

August 2010

Vicenç Torra
Yasuo Narukawa
Marc Daumas

Organization

General Chair

Marc Daumas — University of Perpignan, France

Program Chairs

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

Advisory Board

Lluis Godo, Kaoru Hirota, Janusz Kacprzyk, Sadaaki Miyamoto, Michio Sugeno, Ronald R. Yager

Program Committee

Gleb Beliakov, Ulrich Bodenhofer, Tomasa Calvo, Josep Domingo-Ferrer, Jozo Dujmovic, Michel Grabisch, Enrique Herrera-Viedma, Masahiro Inuiguchi, Hiroaki Kikuchi, Ivan Kojadinovic, Xinwang Liu, Jun Long, Jean-Luc. Marichal, Rosa Meo, Radko Mesiar, Tetsuya Murai, Toshiaki Murofushi, Guillermo Navarro-Arribas, Michael Ng, Gabriella Pasi, Leszek Rutkowski, Aida Valls, Zeshui Xu, Yuji Yoshida, Gexiang Zhang, Ning Zhong.

Local Organizing Committee Chair

Adrien Toutant — University of Perpignan, France

Local Organizing Committee

Stéphane Grieu, Annick Truffert, Stéphane Thil, Michel Ventou, Patrick Vilamajó

Additional Referees

Silvia Calegari, Simon James, David Nettleton, Jordi Nin

Supporting Institutions

The French National Center for Scientific Research (CNRS)
Université de Perpignan Via Domitia
The Électronique, Informatique, Automatique & Systèmes laboratory (ELIAUS)
The Mediterranean Institute for Study and Research in Computer Science and Robotics (IMERIR)
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)

Table of Contents

Invited Papers

Regular Papers

Aggregation Operators and Decision Making

Clustering and Similarity

Computational Intelligence

Data Privacy

Relationships between Qualitative and Quantitative Scales for Aggregation Operations: The Example of Sugeno Integrals

Didier Dubois

IRIT, CNRS and Université de Toulouse
France
dubois@irit.fr

In decision applications, especially multicriteria decision- making, numerical approaches are often questionable because it is hard to elicit numerical values quantifying preference, criteria importance or uncertainty. More often than not, multicriteria decision-making methods come down to number-crunching recipes with debatable foundations. One way out of this difficulty is to adopt a qualitative approach where only maximum and minimum are used. Such methods enjoy a property of scale invariance that insures their robustness. One of the most sophisticated aggregation operation making sense on qualitative scales is Sugeno integral. It is not purely ordinal as it assumes commensurability between preference intensity and criteria importance or similarly, utility and uncertainty. However, since absolute qualitative value scales must have few levels so as to remain cognitively plausible, there are as many classes of equivalent decisions as value levels. Hence this approach suffers from a lack of discrimination power. In particular, qualitative aggregations such as Sugeno integrals cannot be strictly increasing and violate the strict Pareto property. In this talk, we report results obtained when trying to increase the discrimination power of Sugeno integrals, generalizing such refinements of the minimum and maximum as leximin and leximax. The representation of leximin and leximax by sums of numbers of different orders of magnitude (forming a super-increasing sequence) can be generalized to weighted max and min (yielding a "big-stepped" weighted average) and Sugeno integral (yielding a "big-stepped" Choquet integral). This methodology also requires qualitative monotonic set-functions to be refined by numerical set-functions, and we show they can always be belief or plausibility functions in the sense of Shafer.

References

1. Dubois, D., Fargier, H.: Making Discrete Sugeno Integrals More Discriminant. International Journal of Approximate Reasoning 50, 880–898 (2009)
2. Dubois, D., Fargier, H., Prade, H., Sabbadin, R.: A survey of qualitative decision rules under uncertainty. In: Bouyssou, D., Dubois, D., Pirlot, M., Prade, H. (eds.) Decision-making Process- Concepts and Methods, ch. 11, pp. 435–473. ISTE London & Wiley (2009)

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 1–2, 2010.

3. Dubois, D., Fargier, H.: Capacity refinements and their application to qualitative decision evaluation. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS (LNAI), vol. 5590, pp. 311–322. Springer, Heidelberg (2009)
4. Fargier, H., Sabbadin, R.: Qualitative decision under uncertainty: Back to expected utility. Artificial Intelligence 164, 245–280 (2005)
5. Prade, H., Rico, A., Serrurier, M., Raufaste, E.: Eliciting Sugeno integrals: methodology and a case study. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS (LNAI), vol. 5590, pp. 712–723. Springer, Heidelberg (2009)
6. Prade, H., Rico, A., Serrurier, M.: Elicitation of Sugeno integrals: A version space learning perpective. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 392–401. Springer, Heidelberg (2009)

User Privacy in Web Search

Josep Domingo-Ferrer

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@urv.cat

Web search engines gather a lot of information on the preferences and interests of users. They actually gather enough information to create detailed user profiles which might enable re-identification of the individuals to which those profiles correspond, *e.g.* thanks to the so-called vanity queries or to linkage of several queries known to have been submitted by the same user. In this way, a broadly used search engine like Google becomes a "big brother" in the purest Orwellian style.

In this talk, a survey will be offered of the solutions which have been proposed to preserve anonymity in web search and to fight profile creation. We will start with Private Information Retrieval (PIR) and we will highlight its lack of practicality. We will then look at some relaxations of PIR, based on standalone defense by the user or on a defense based on a peer-to-peer community in which one user submits queries by other users and viceversa.

Finally, we will sketch a new theory, called coprivacy or co-operative privacy, whose goal is to find out under which conditions the best rational option for a peer-to-peer user is to help other peers in preserving their privacy.

References

1. Aguilar-Melchor, C., Deswarte, Y.: Trustable relays for anonymous communication. Transactions on Data Privacy 2(2), 101–130 (2009)
2. AOL Search Data Scandal (August 2006), http://en.wikipedia.org/wiki/AOL_search_data_scandal
3. Beimel, A., Ishai, Y., Malkin, T.: Reducing the servers' computation in private information retrieval: Pir with preprocessing. Journal of Cryptology 17, 125–151 (2004)
4. Castellà-Roca, J., Viejo, A., Herrera-Joancomartí, J.: Preserving user's privacy in web search engines. Computer Communications 32(13-14), 1541–1551 (2009)
5. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 41–50 (1995)
6. Chor, B., Gilboa, N., Naor, M.: Private information retrieval by keywords. Technical Report TR CS0917, Department of Computer Science, Technion (1997)
7. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. Journal of the ACM 45, 965–981 (1998)

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 3–4, 2010.

8. Domingo-Ferrer, J.: Coprivacy: towards a theory of sustainable privacy. In: Domingo-Ferrer, J. (ed.) PSD 2010. LNCS, vol. 6344, pp. 258–268. Springer, Heidelberg (2010)
9. Domingo-Ferrer, J., Bras-Amorós, M., Wu, Q., Manjón, J.: User-private information retrieval based on a peer-to-peer community. Data and Knowledge Engineering 68(11), 1237–1252 (2009)
10. Domingo-Ferrer, J., Solanas, A., Castellà-Roca, J.: $h(k)$-Private information retrieval from privacy-uncooperative queryable databases. Online Information Review 33(4), 720–744 (2009)
11. Erola, A., Castellà-Roca, J., Navarro-Arribas, G., Torra, V.: Semantic microaggregation for the anonymization of query logs. In: Domingo-Ferrer, J. (ed.) PSD 2010. LNCS, vol. 6344, pp. 127–137. Springer, Heidelberg (2010)
12. Howe, D.C., Nissenbaum, H.: TrackMeNot: Resisting surveillance in web search. In: Kerr, I., Lucock, C., Steeves, V. (eds.) Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society. Oxford University Press, Oxford (2009), http://www.mrl.nyu.edu/~dhowe/trackmenot/
13. Internet World Stats, http://www.internetworldstats.com/
14. iProspect Blended Search Results Study (April 2008), http://www.iprospect.com/premiumPDFs/researchstudy_apr2008_blendedsearchresults.pdf
15. Nash, J.: Non-cooperative games. Annals of Mathematics 54, 289–295 (1951)
16. Netcraft June 2010 Web Server Survey, http://news.netcraft.com/archives/2010/06/16/june-2010-web-server-survey.html
17. Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.): Algorithmic Game Theory. Cambridge University Press, Cambridge (2007)
18. Open Directory Project, http://www.dmoz.org
19. Ogata, W., Kurosawa, K.: Oblivious keyword search. Journal of Complexity 20(2-3), 356–371 (2004)
20. Ostrovsky, R., Skeith III, W.E.: A survey of single-database PIR: techniques and applications. In: Okamoto, T., Wang, X. (eds.) PKC 2007. LNCS, vol. 4450, pp. 393–411. Springer, Heidelberg (2007)
21. Reiter, M.K., Rubin, A.D.: Crowds: anonymity for web transactions. ACM Transactions on Information Systems Security 1(1), 66–92 (1998)
22. Saint-Jean, F., Johnson, A., Boneh, D., Feigenbaum, J.: Private web search. In: Proc. of the 2007 ACM Workshop on Privacy in Electronic Society, pp. 84–90 (2007)
23. Shen, X., Tan, B., Zhai, C.X.: Privacy protection in personalized search. ACM SIGIR Forum 41(1), 4–17 (2007)
24. The Tor Project, Inc. "Tor: Overview", http://torproject.org/overview.html.en
25. Torbutton 1.2.5, https://addons.mozilla.org/ca/firefox/addon/2275/
26. Viejo, A., Castellà-Roca, J.: Using social networks to distort users' profiles generated by web search engines. Computer Networks (to appear)
27. Princeton University. WordNet: A lexical database for English, http://wordnet.princeton.edu/

A Bibliometric Index Based on Collaboration Distances

Maria Bras-Amorós[1], Josep Domingo-Ferrer[1], and Vicenç Torra[2]

[1] Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{maria.bras,josep.domingo}@urv.cat
[2] IIIA, Institut d'Investigació en Intel·ligència Artificial
CSIC, Consejo Superior de Investigaciones Científicas,
Campus UAB s/n, E-08193 Bellaterra, Catalonia
vtorra@iiia.csic.es

The h-index by Hirsch[1] has recently earned a lot of popularity in bibliometrics, being echoed in Nature and implemented in the Web of Science bibliometric database. Previous indicators were the total number of papers or the total number of citations. Following the widely accepted idea that not all papers should count equally, the h-index counts only those papers that are significant enough according to their number of citations. However, as for qualifying the significance of citations, beyond excluding self-citations by recent proposals [2,3,4,5,6], the fact that not all citations should count equally has remained unaddressed, with the exception of [7]. The h-index can be described in terms of a pool of evaluated objects (papers), a quality function on the evaluated object (citations received by each paper) and a sentencing line crossing the origin ($y = x$). When the evaluated objects are ordered by descreasing quality, then the intersection of the sentencing line with the graph of the quality function yields the index value.

Based on this abstraction, we present a new index, the c-index, in which the evaluated objects are the citations received (by a paper, an author, a research group, a journal, etc.), the quality of a citation is the collaboration distance between the authors of the cited and the citing papers when the citation appears, and the sentencing line takes a slope α between 0 and ∞. To mitigate the small world effect we suggest taking $\alpha \approx 1/4$. As a result, the new index counts only those citations which are significant enough, where significance is proportional to the collaboration distance between the cited and the citing authors.

While an h-index x means that there are x papers with at least x citations each and the rest of papers with at most x citations, a c-index x means that there are x citations (regardless of the papers to which these citations refer) at collaboration distance at least $\alpha\ x$ and the rest of citations at collaboration distance at most $\alpha\ x$.

If we want to differentiate between recurrent collaborations and occasional collaborations, a refined version of the classical distance can be defined, where the distance between two coauthors is inversely proportional to the number of joint papers between them.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 5–6, 2010.

Some of the advantages of the new c-index are:

1. It gives a solution to the problem of few but seminal contributions, which for instance means that Galois has h-index 2, and which is also especially important when evaluating journals[6].
2. It neutralizes self-citations and citations by close authors.
3. It discourages gratuitous coauthorship.
4. The new index is not linear anymore with respect to the scientific age, rewarding citations to and from novel authors and thus modernity.
5. Multiple spelling of one single reference in different citations or misspelling reference data other than authorship, which decrease the h-index, do not affect the c-index.

Together with the f-index [7], the c-index is a pioneer in the bibliometric literature in measuring the output of a scientist or a journal based at the same time on the quantity and quality of the received citations: the more distant the citing authors, the higher the quality of a citation (this notion of quality rewards contributions of broad interest).

Since any bibliometric index is referred to a particular database, it should be easy for any of the bibliometric databases to automate the computation of the c-index, just like some of them have automated the computation of classical distances (MathSciNet of the American Mathematical Society) or the computation of the h-index (Web of Science).

Being based only on citations, the c-index loses the feature of the h-index of counting how many papers among those by an author have had a decent impact. To remedy this, one might combine the h-index and the c-index by providing both of them or by mixing them (e.g. as $\sqrt{hc}$) if required for ranking purposes.

In [8] one can find an extended version of this paper with a deeper discussion on the c-index, a detailed comparison with the most recent indices, some computational hints, and some experiments.

References

1. Ball, P.: Index aims for fair ranking of scientists. Nature 436, 900 (2005)
2. Schreiber, M.: Self-citation corrections for the Hirsch index. Europhysics Letters (EPL) 78, 30002p1–30002p6 (2007)
3. Derby, B.: H-factors research metrics and self-citation. Nature Blogs (April 25, 2008)
4. Zhivotovsky, L.A., Krutovsky, K.V.: Self-citation can inflate h-index. Scientometrics 77(2), 373–375 (2008)
5. Egghe, L.: An improvement of the h-index: the g-index. ISSI Newsletter 2(1), 8–9 (2006)
6. Braun, T., Glänzel, W., Schubert, A.: A Hirsch-type index for journals. Scientometrics 69(1), 169–173 (2006)
7. Katsaros, D., Akritidis, L., Bozanis, P.: The f-index: quantifying the impact of coterminal citations on scientists' ranking. Journal of the American Society for Information Science and Technology 60(5), 1051–1056 (2009)
8. Bras-Amorós, M., Domingo-Ferrer, J., Torra, V.: A bibliometric index based on the collaboration distance between cited and citing authors (submitted 2010)

Measuring the Influence of the kth Largest Variable on Functions over the Unit Hypercube

Jean-Luc Marichal and Pierre Mathonet

Mathematics Research Unit, FSTC, University of Luxembourg,
6, rue Coudenhove-Kalergi, L-1359 Luxembourg,
Grand Duchy of Luxembourg
jean-luc.marichal@uni.lu, pierre.mathonet@uni.lu

Abstract. By considering a least squares approximation of a given square integrable function $f\colon [0,1]^n \to \mathbb{R}$ by a shifted L-statistic function (a shifted linear combination of order statistics), we define an index which measures the global influence of the kth largest variable on f. We show that this influence index has appealing properties and we interpret it as an average value of the difference quotient of f in the direction of the kth largest variable or, under certain natural conditions on f, as an average value of the derivative of f in the direction of the kth largest variable. We also discuss a few applications of this index in statistics and aggregation theory.

1 Introduction

Consider a real-valued function f of n variables $x_1, \ldots, x_n$ and suppose we want to measure a global influence degree of every variable x_i on f. A reasonable way to define such an influence degree consists in considering the coefficient of x_i in the best least squares approximation of f by affine functions of the form

$$g(x_1, \ldots, x_n) = c_0 + \sum_{i=1}^{n} c_i x_i.$$

This approach was considered in [6,10] for pseudo-Boolean functions $f\colon \{0,1\}^n \to \mathbb{R}$ and in [9] for square integrable functions $f\colon [0,1]^n \to \mathbb{R}$. It turns out that, in both cases, the influence index of x_i on f is given by an average "derivative" of f with respect to x_i.

Now, it is also natural to consider and measure a global influence degree of the smallest variable, or the largest variable, or even the kth largest variable for some $k \in \{1, \ldots, n\}$. As an application, suppose we are to choose an appropriate aggregation function $f\colon [0,1]^n \to \mathbb{R}$ to compute an average value of $[0,1]$-valued grades obtained by a student. If, for instance, we use the arithmetic mean function, we might expect that both the smallest and the largest variables are equally influent. However, if we use the geometric mean function, for which the value 0

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 7–18, 2010.

(the left endpoint of the scale) is multiplicatively absorbent, we might anticipate that the smallest variable is more influent than the largest one.

Similarly to the previous problem, to define the influence of the kth largest variable on f it is natural to consider the coefficient of $x_{(k)}$ in the best least squares approximation of f by symmetric functions of the form

$$g(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i x_{(i)},$$

where $x_{(1)}, \ldots, x_{(n)}$ are the order statistics obtained by rearranging the variables in ascending order of magnitude.

In this paper we solve this problem for square integrable functions $f\colon [0,1]^n \to \mathbb{R}$. More precisely, we completely describe the least squares approximation problem above and derive an explicit expression for the corresponding influence index (§2). We also show that this index has several natural properties, such as linearity and continuity, and we give an interpretation of it as an average value of the difference quotient of f in the direction of the kth largest variable. Under certain natural conditions on f, we also interpret the index as an average value of the derivative of f in the direction of the kth largest variable (§3). We then provide some alternative formulas for the index to possibly simplify its computation (§4) and we consider some examples including the case when f is the Lovász extension of a pseudo-Boolean function (§5). Finally, we discuss a few applications of the index (§6).

We employ the following notation throughout the paper. Let $\mathbb{I}^n$ denote the n-dimensional unit cube $[0,1]^n$. We denote by $L^2(\mathbb{I}^n)$ the class of square integrable functions $f\colon \mathbb{I}^n \to \mathbb{R}$ modulo equality almost everywhere. For any $S \subseteq [n] = \{1, \ldots, n\}$, we denote by $\mathbf{1}_S$ the characteristic vector of S in $\{0,1\}^n$ (with the particular case $\mathbf{0} = \mathbf{1}_\varnothing$).

Recall that if the $\mathbb{I}$-valued variables $x_1, \ldots, x_n$ are rearranged in ascending order of magnitude $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$, then $x_{(k)}$ is called the *kth order statistic* and the function $\mathrm{os}_k\colon \mathbb{I}^n \to \mathbb{R}$, defined as $\mathrm{os}_k(\mathbf{x}) = x_{(k)}$, is the *kth order statistic function*. As a matter of convenience, we also formally define $\mathrm{os}_0 \equiv 0$ and $\mathrm{os}_{n+1} \equiv 1$. To stress on the arity of the function, we can replace the symbols $x_{(k)}$ and os_k with $x_{k:n}$ and $\mathrm{os}_{k:n}$, respectively. For general background on order statistics, see for instance [1,4].

Finally, we use the lattice notation $\wedge$ and $\vee$ to denote the minimum and maximum functions, respectively.

2 Influence Index for the *k*th Largest Variable

An *L-statistic* function is a linear combination of the functions $\mathrm{os}_1, \ldots, \mathrm{os}_n$. A *shifted L-statistic* function is a constant plus an L-statistic function. Denote by V_L the set of shifted L-statistic functions. Clearly, V_L is spanned by the linearly independent set

$$B = \{\mathrm{os}_1, \ldots, \mathrm{os}_n, \mathrm{os}_{n+1}\} \tag{1}$$

and thus is a linear subspace of $L^2(\mathbb{I}^n)$ of dimension $n+1$. For a given function $f \in L^2(\mathbb{I}^n)$, we define the *best shifted L-statistic approximation of* f as the function $f_L \in V_L$ that minimizes the distance

$$\|f-g\|^2 = \int_{\mathbb{I}^n} (f(\mathbf{x})-g(\mathbf{x}))^2 \, d\mathbf{x}$$

among all $g \in V_L$, where $\|\cdot\|$ is the norm in $L^2(\mathbb{I}^n)$ associated with the inner product $\langle f,g\rangle = \int_{\mathbb{I}^n} f(\mathbf{x})g(\mathbf{x})\,d\mathbf{x}$. Using the general theory of Hilbert spaces, we immediately see that the solution of this approximation problem exists and is uniquely determined by the orthogonal projection of f onto V_L. This projection is given by

$$f_L = \sum_{j=1}^{n+1} a_j \operatorname{os}_j , \tag{2}$$

where the coefficients a_j (for $j \in [n+1]$) are characterized by the conditions

$$\langle f - f_L, \operatorname{os}_i\rangle = 0 \quad \text{for all} \quad i \in [n+1]. \tag{3}$$

The coefficient matrix of this sytem is the square matrix M of order $n+1$ defined by $(M)_{ij} = \langle \operatorname{os}_i, \operatorname{os}_j\rangle$ for all $i,j \in [n+1]$.

Lemma 1. *For every $i,j \in [n+1]$, we have*

$$(M)_{ij} = \frac{\min(i,j)\big(\max(i,j)+1\big)}{(n+1)(n+2)} \tag{4}$$

and

$$\frac{(M^{-1})_{ij}}{(n+1)(n+2)} = \begin{cases} 2\,, & \text{if } i=j<n+1,\\ \frac{n+1}{n+2}\,, & \text{if } i=j=n+1,\\ -1\,, & \text{if } |i-j|=1,\\ 0\,, & \text{otherwise.}\end{cases} \tag{5}$$

Recall that the central second difference operator is defined for any real sequence $(z_k)_{k\geqslant 1}$ as $\delta_k^2 z_k = z_{k+1} - 2z_k + z_{k-1}$. For every $k \in [n]$, define the function $g_k \in L^2(\mathbb{I}^n)$ as

$$g_k = -(n+1)(n+2)\,\delta_k^2 \operatorname{os}_k. \tag{6}$$

We immediately obtain the following explicit forms for the components of f_L in the basis (1).

Proposition 1. *The best shifted L-statistic approximation f_L of a function $f \in L^2(\mathbb{I}^n)$ is given by (2), where*

$$a_k = \begin{cases} \langle f, g_k\rangle\,, & \text{if } k \in [n],\\ (n+1)^2\langle f,1\rangle - (n+1)(n+2)\langle f, \operatorname{os}_n\rangle\,, & \text{if } k = n+1.\end{cases} \tag{7}$$

Now, to measure the global influence of the kth largest variable $x_{(k)}$ on an arbitrary function $f \in L^2(\mathbb{I}^n)$, we naturally define an index $I\colon L^2(\mathbb{I}^n) \times [n] \to \mathbb{R}$ as $I(f,k) = a_k$, where a_k is obtained from f by (7). We will see in the next section that this index indeed measures an influence degree.

Definition 1. *Let $I\colon L^2(\mathbb{I}^n) \times [n] \to \mathbb{R}$ be defined as $I(f,k) = \langle f, g_k \rangle$, that is*

$$I(f,k) = -(n+1)(n+2) \int_{\mathbb{I}^n} f(\mathbf{x})\, \delta_k^2\, x_{(k)}\, d\mathbf{x}. \tag{8}$$

3 Properties and Interpretations

In this section we present various properties and interpretations of the index $I(f,k)$. The first result follows immediately from Definition 1.

Proposition 2. *For every $k \in [n]$, the mapping $f \mapsto I(f,k)$ is linear and continuous.*

We now present an interpretation of $I(f,k)$ as a covariance. Considering the unit cube $\mathbb{I}^n$ as a probability space with respect to the Lebesgue measure, we see that, for any $k \in [n]$, the index $I(f,k)$ is the covariance of the random variables f and g_k. Indeed, we have $I(f,k) = E(f\, g_k) = \operatorname{cov}(f, g_k) + E(f)\, E(g_k)$, where $E(g_k) = \langle 1, g_k \rangle = I(1,k) = 0$. From the usual interpretation of the concept of covariance, we see that $I(f,k)$ is positive whenever the values of $f - E(f)$ and $g_k - E(g_k) = g_k$ have the same sign. Note that $g_k(\mathbf{x})$ is positive whenever $x_{(k)}$ is greater than $\frac{1}{2}(x_{(k+1)} + x_{(k-1)})$, which is the midpoint of the range of $x_{(k)}$ when the other order statistics are fixed at $\mathbf{x}$.

We now provide an interpretation of $I(f,k)$ as an expected value of the derivative of f in the direction of the kth largest variable (see Proposition 3).

Let S_n denote the group of permutations of $[n]$. Recall that the unit cube $\mathbb{I}^n$ can be partitioned almost everywhere into the open standard simplexes

$$\mathbb{I}^n_\pi = \{\mathbf{x} \in \mathbb{I}^n : x_{\pi(1)} < \cdots < x_{\pi(n)}\} \qquad (\pi \in S_n).$$

Definition 2. *Given $k \in [n]$, let $f\colon \cup_{\pi \in S_n} \mathbb{I}^n_\pi \to \mathbb{R}$ be a function such that the partial derivative $D_{\pi(k)} f|_{\mathbb{I}^n_\pi}$ exists for every $\pi \in S_n$. The* derivative of f in the direction (k) *is the function $D_{(k)} f\colon \cup_{\pi \in S_n} \mathbb{I}^n_\pi \to \mathbb{R}$ defined as*

$$D_{(k)} f(\mathbf{x}) = D_{\pi(k)} f(\mathbf{x}) \quad \textit{for all} \quad \mathbf{x} \in \mathbb{I}^n_\pi.$$

Now, for every $k \in [n]$, consider the function $h_k \in L^2(\mathbb{I}^n)$ defined as

$$h_k = (n+1)(n+2)(\mathrm{os}_{k+1} - \mathrm{os}_k)(\mathrm{os}_k - \mathrm{os}_{k-1}).$$

We easily see that h_k is a probability density function on $\mathbb{I}^n$. This fact can also be derived by choosing $f = \mathrm{os}_k$ in the following result.

Proposition 3. *For every $k \in [n]$ and every $f \in L^2(\mathbb{I}^n)$ such that $D_{(k)}f$ is continuous and integrable on $\cup_{\pi \in S_n} \mathbb{I}^n_\pi$, we have*

$$I(f,k) = \int_{\mathbb{I}^n} h_k(\mathbf{x})\, D_{(k)} f(\mathbf{x})\, d\mathbf{x}. \tag{9}$$

We now give an alternative interpretation of $I(f,k)$ as an expected value, which does not require the additional assumptions of Proposition 3. In this more general framework, we naturally replace the derivative with a difference quotient. To this extent, we introduce some further notation. As usual, we denote by $\mathbf{e}_i$ the ith vector of the standard basis for $\mathbb{R}^n$. For every $k \in [n]$ and every $h \in [0,1]$, we define the (k)-*difference* (or *discrete* (k)-*derivative*) operator $\Delta_{(k),h}$ over the set of real functions on $\mathbb{I}^n$ by

$$\Delta_{(k),h} f(\mathbf{x}) = f(\mathbf{x} + h\,\mathbf{e}_{\pi(k)}) - f(\mathbf{x})$$

for every $\mathbf{x} \in \mathbb{I}^n_\pi$ such that $\mathbf{x} + h\mathbf{e}_{\pi(k)} \in \mathbb{I}^n_\pi$. Thus defined, the value $\Delta_{(k),h} f(\mathbf{x})$ can be interpreted as the *marginal contribution* of $x_{(k)}$ on f at $\mathbf{x}$ with respect to the increase h. For instance, we have $\Delta_{(k),h}\, x_{(k)} = h$.

Similarly, we define the (k)-*difference quotient* operator $Q_{(k),h}$ over the set of real functions on $\mathbb{I}^n$ by $Q_{(k),h} f(\mathbf{x}) = \frac{1}{h} \Delta_{(k),h} f(\mathbf{x})$.

Theorem 1. *For every $k \in [n]$ and every $f \in L^2(\mathbb{I}^n)$, we have*

$$I(f,k) = (n+1)(n+2) \int_{\mathbb{I}^n} \int_{x_{(k)}}^{x_{(k+1)}} \Delta_{(k),y-x_{(k)}} f(\mathbf{x})\, dy\, d\mathbf{x}. \tag{10}$$

In view of Eq. (10), the index $I(f,k)$ can be interpreted (up to normalization) as a summation over all points $\mathbf{x} \in \mathbb{I}^n$ of the importance of the kth largest variable at $\mathbf{x}$, given by

$$\int_{x_{(k)}}^{x_{(k+1)}} \Delta_{(k),y-x_{(k)}} f(\mathbf{x})\, dy.$$

As an immediate consequence of Theorem 1, we have the following interpretation of the index $I(f,k)$ as an expected value of a difference quotient with respect to some distribution.

Corollary 1. *For every $k \in [n]$ and every $f \in L^2(\mathbb{I}^n)$, we have*

$$I(f,k) = \int_{\mathbb{I}^n} \int_{x_{(k)}}^{x_{(k+1)}} p_k(\mathbf{x},y)\, Q_{(k),y-x_{(k)}} f(\mathbf{x})\, dy\, d\mathbf{x},$$

where $p_k(\mathbf{x},y) = (n+1)(n+2)(y - x_{(k)})$ defines a probability density function on the set $\{(\mathbf{x},y) : \mathbf{x} \in \mathbb{I}^n,\, y \in [x_{(k)}, x_{(k+1)}]\}$.

Another important feature of the index is its invariance under the action of permutations. Recall that a permutation $\pi \in S_n$ acts on a function $f\colon \mathbb{I}^n \to \mathbb{R}$ by $\pi(f)(x_1,\ldots,x_n) = f(x_{\pi(1)},\ldots,x_{\pi(n)})$. By the change of variables theorem, we immediately see that every $\pi \in S_n$ is an isometry of $L^2(\mathbb{I}^n)$, that is, $\langle \pi(f), \pi(g) \rangle = \langle f, g \rangle$. From this fact, we derive the following result.

Proposition 4. *For every $f \in L^2(\mathbb{I}^n)$ and every $\pi \in S_n$, both functions f and $\pi(f)$ have the same best shifted L-statistic approximation f_L. Moreover, we have $\|\pi(f) - f_L\| = \|f - f_L\|$.*

With any function $f\colon \mathbb{I}^n \to \mathbb{R}$ we can associate the following symmetric function

$$\mathrm{Sym}(f) = \frac{1}{n!} \sum_{\pi \in S_n} \pi(f).$$

It follows immediately from Propositions 2 and 4 that both functions f and $\mathrm{Sym}(f)$ have the same best shifted L-statistic approximation f_L. Combining this observation with Proposition 4, we derive immediately the following corollary.

Corollary 2. *For every $k \in [n]$, every $f \in L^2(\mathbb{I}^n)$, and every $\pi \in S_n$, we have $I(f,k) = I(\pi(f),k) = I(\mathrm{Sym}(f),k)$.*

Remark 1. Corollary 2 shows that, to compute $I(f,k)$, we can replace f with $\mathrm{Sym}(f)$. For instance, if $f(\mathbf{x}) = x_i$ for some $i \in [n]$ then $\mathrm{Sym}(f) = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\sum_{i=1}^n x_{(i)}$ and hence, using Proposition 3, we obtain $I(f,k) = \frac{1}{n}$.

Given $k \in [n]$, we say that the order statistic $x_{(k)}$ is *ineffective* almost everywhere for a function $f\colon \mathbb{I}^n \to \mathbb{R}$ if $\Delta_{(k),y-x_{(k)}} f(\mathbf{x}) = 0$ for almost all $\mathbf{x} \in \cup_{\pi \in S_n} \mathbb{I}^n_\pi$ and almost all $y \in \left]x_{(k-1)}, x_{(k+1)}\right[$. For instance, given unary functions $f_1, f_2 \in L^2(\mathbb{I})$, the order statistic $x_{(1)}$ is ineffective almost everywhere for the function $f\colon \mathbb{I}^2 \to \mathbb{R}$ such that

$$f(x_1,x_2) = \begin{cases} f_1(x_1), & \text{if } x_1 > x_2, \\ f_2(x_2), & \text{if } x_1 < x_2. \end{cases}$$

The following result immediately follows from Theorem 1.

Proposition 5. *Let $k \in [n]$ and $f \in L^2(\mathbb{I}^n)$. If $x_{(k)}$ is ineffective almost everywhere for f, then $I(f,k) = 0$.*

The *dual* of a function $f\colon \mathbb{I}^n \to \mathbb{R}$ is the function $f^d\colon \mathbb{I}^n \to \mathbb{R}$ defined by $f^d(\mathbf{x}) = 1 - f(\mathbf{1}_{[n]} - \mathbf{x})$. A function $f\colon \mathbb{I}^n \to \mathbb{R}$ is said to be *self-dual* if $f^d = f$. By using the change of variables theorem, we immediately derive the following result.

Proposition 6. *For every $f \in L^2(\mathbb{I}^n)$ and every $k \in [n]$, we have $I(f^d,k) = I(f,n-k+1)$. In particular, if f is self-dual, then $I(f,k) = I(f,n-k+1)$.*

4 Alternative Expressions for the Index

The computation of the index $I(f,k)$ by means of (8) or (9) might be not very convenient due to the presence of the order statistic functions. To make those integrals either more tractable or easier to evaluate numerically, we provide in this section some alternative expressions for the index $I(f,k)$ that do not involve any order statistic.

We first derive useful formulas for the computation of the integral $\langle f, \mathrm{os}_k \rangle$ (Proposition 7). To this extent, we consider the following direct generalization of order statistic functions.

Definition 3. *For every nonempty $S = \{i_1, \ldots, i_s\} \subseteq [n]$, $s = |S|$, and every $k \in [s]$, we define the function $\mathrm{os}_{k:S} \colon \mathbb{I}^n \to \mathbb{R}$ as $\mathrm{os}_{k:S}(\mathbf{x}) = \mathrm{os}_{k:s}(x_{i_1}, \ldots, x_{i_s})$.*

To simplify the notation, we will write $x_{k:S}$ for $\mathrm{os}_{k:S}(\mathbf{x})$. Thus $x_{k:S}$ is the kth order statistic of the variables in S.

Lemma 2. *For every $s \in [n]$ and every $k \in [s]$, we have*

$$\sum_{\substack{S \subseteq [n] \\ |S|=s}} x_{k:S} = \sum_{j=k}^{n} \binom{j-1}{k-1} \binom{n-j}{s-k} x_{j:n}. \tag{11}$$

Lemma 3. *For every $k \in [n]$, we have*

$$x_{k:n} = \sum_{\substack{S \subseteq [n] \\ |S| \geqslant k}} (-1)^{|S|-k} \binom{|S|-1}{k-1} x_{|S|:S} \tag{12}$$

$$x_{k:n} = \sum_{\substack{S \subseteq [n] \\ |S| \geqslant n-k+1}} (-1)^{|S|-n+k-1} \binom{|S|-1}{n-k} x_{1:S} \tag{13}$$

Combining Lemma 3 with some classical results in measure theory, we compute several expressions of $\int_{\mathbb{I}^n} f(\mathbf{x})\, x_{(k)}\, d\mathbf{x}$:

Proposition 7. *For every function $f \in L^2(\mathbb{I}^n)$ and every $k \in [n]$, the integral $J_{k:n} = \int_{\mathbb{I}^n} f(\mathbf{x})\, x_{(k)}\, d\mathbf{x}$ is given by each of the following expressions:*

$$\int_{\mathbb{I}^n} f(\mathbf{x})\, d\mathbf{x} - \sum_{S \subseteq [n] : |S| \geqslant k} (-1)^{|S|-k} \tbinom{|S|-1}{k-1} \int_0^1 \int_{[0,y]^S} \int_{[0,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{14}$$

$$\sum_{S \subseteq [n] : |S| \geqslant n-k+1} (-1)^{|S|-n+k-1} \tbinom{|S|-1}{n-k} \int_0^1 \int_{[y,1]^S} \int_{[0,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{15}$$

$$\int_{\mathbb{I}^n} f(\mathbf{x})\, d\mathbf{x} - \sum_{S \subseteq [n] : |S| \geqslant k} \int_0^1 \int_{[0,y]^S} \int_{[y,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{16}$$

$$\sum_{S \subseteq [n] : |S| < k} \int_0^1 \int_{[0,y]^S} \int_{[y,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{17}$$

From Definition 1 and Proposition 7, we derive the following expressions for the quantity $\frac{I(f,k)}{(n+1)(n+2)}$:

$$\sum_{S \subseteq [n] : |S| \geqslant k-1} (-1)^{|S|+1-k} \tbinom{|S|+1}{k} \int_0^1 \int_{[0,y]^S} \int_{[0,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{18}$$

$$\sum_{S \subseteq [n] : |S| \geqslant n-k} (-1)^{|S|-n+k-1} \tbinom{|S|+1}{n-k+1} \int_0^1 \int_{[y,1]^S} \int_{[0,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy \tag{19}$$

$$\Big(\sum_{S \subseteq [n] : |S| = k-1} - \sum_{S \subseteq [n] : |S| = k} \Big) \int_0^1 \int_{[0,y]^S} \int_{[y,1]^{[n] \setminus S}} f(\mathbf{x})\, d\mathbf{x}\, dy. \tag{20}$$

5 Some Examples

We now apply our results to two special classes of functions, namely the multiplicative functions and the Lovász extensions of pseudo-Boolean functions. The latter class includes the so-called discrete Choquet integrals, well-known in aggregation function theory.

5.1 Multiplicative Functions

Consider the function $f(\mathbf{x}) = \prod_{i=1}^n \varphi_i(x_i)$, where $\varphi_i \in L^2(\mathbb{I})$, and set $\Phi_i(x) = \int_0^x \varphi_i(t)\,dt$ for $i = 1, \ldots, n$. By using (18), we obtain

$$\frac{I(f,k)}{(n+1)(n+2)} = \sum_{\substack{S\subseteq[n]\\ |S|\geqslant k-1}} (-1)^{|S|+1-k} \binom{|S|+1}{k} \prod_{i\in[n]\setminus S} \Phi_i(1) \int_0^1 \prod_{i\in S} \Phi_i(y)\,dy \tag{21}$$

The following result gives a concise expression for $I(f,k)$ when f is symmetric.

Proposition 8. *Let $f\colon \mathbb{I}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \prod_{i=1}^n \varphi(x_i)$, where $\varphi \in L^2(\mathbb{I})$, and let $\Phi(x) = \int_0^x \varphi(t)\,dt$. Then, for every $k \in [n]$, we have*

$$I(f,k) = \begin{cases} \Phi(1)^n \int_0^1 D_z h(z; k+1, n-k+2)|_{z=\Phi(y)/\Phi(1)}\,dy, & \text{if } \Phi(1) \neq 0,\\ (-1)^{n-k+1}(n+1)\,\frac{\Gamma(n+3)}{\Gamma(k+1)\,\Gamma(n-k+2)} \int_0^1 \Phi(y)^n\,dy, & \text{if } \Phi(1) = 0, \end{cases}$$

where $h(z;a,b) = z^{a-1}(1-z)^{b-1}/B(a,b)$ is the probability density function of the beta distribution with parameters a and b.

Example 1. Let $f\colon \mathbb{I}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \left(\prod_{i=1}^n x_i\right)^c$, where $c > -\frac{1}{2}$. For instance, the product function corresponds to $c = 1$ and the geometric mean function to $c = 1/n$. We can calculate $I(f,k)$ by using Proposition 8 with $\varphi(x) = x^c$. Using the substitution $z = y^{c+1}$ and then integrating by parts, we obtain

$$I(f,k) = c\left(\frac{1}{c+1}\right)^{n+2} \frac{\Gamma(n+3)\,\Gamma(k-1+\frac{1}{c+1})}{\Gamma(k+1)\,\Gamma(n+1+\frac{1}{c+1})} = \frac{\Gamma(k-1+\frac{1}{c+1})}{\Gamma(k+1)\,\Gamma(\frac{1}{c+1})}\, I(f,1),$$

with

$$I(f,1) = c\left(\frac{1}{c+1}\right)^{n+2} \frac{\Gamma(n+3)\,\Gamma(\frac{1}{c+1})}{\Gamma(n+1+\frac{1}{c+1})}.$$

We observe that $I(f,k) \to I(f,1)$ as $c \to -\frac{1}{2}$. Also, for $c > 0$, we have $I(f,k+1) < I(f,k)$ for every $k \in [n-1]$. As expected in this case, the smallest variables are more influent on f than the largest ones.

5.2 Lovász Extensions

Recall that an n-place (*lattice*) *term function* $p\colon \mathbb{I}^n \to \mathbb{I}$ is a combination of projections $\mathbf{x} \mapsto x_i$ ($i \in [n]$) using the fundamental lattice operations $\wedge$ and $\vee$; see [2]. For instance,

$$p(x_1, x_2, x_3) = (x_1 \wedge x_2) \vee x_3$$

is a 3-place term function. Note that, since $\mathbb{I}$ is a bounded chain, here the lattice operations $\wedge$ and $\vee$ reduce to the minimum and maximum functions, respectively.

Clearly, any shifted linear combination of n-place term functions

$$f(\mathbf{x}) = c_0 + \sum_{i=1}^{m} c_i\, p_i(\mathbf{x})$$

is a continuous function whose restriction to any standard simplex $\mathbb{I}^n_\pi$ ($\pi \in S_n$) is a shifted linear function. According to Singer [11, §2], f is then the *Lovász extension* of the pseudo-Boolean function $f|_{\{0,1\}^n}$, that is, the continuous function $f\colon \mathbb{I}^n \to \mathbb{R}$ which is defined on each standard simplex $\mathbb{I}^n_\pi$ as the unique affine function that coincides with $f|_{\{0,1\}^n}$ at the $n+1$ vertices of $\mathbb{I}^n_\pi$. Singer showed that a Lovász extension can always be written as

$$f(\mathbf{x}) = f^\pi_{n+1} + \sum_{i=1}^{n} (f^\pi_i - f^\pi_{i+1})\, x_{\pi(i)} \qquad (\mathbf{x} \in \mathbb{I}^n_\pi), \tag{22}$$

with $f^\pi_i = f(\mathbf{1}_{\{\pi(i),\ldots,\pi(n)\}}) = v_f(\{\pi(i),\ldots,\pi(n)\})$ for $i \in [n+1]$, where the set function $v_f\colon 2^{[n]} \to \mathbb{R}$ is defined as $v_f(S) = f(\mathbf{1}_S)$. In particular, $f^\pi_{n+1} = c_0 = f(\mathbf{0})$. Conversely, any continuous function $f\colon \mathbb{I}^n \to \mathbb{R}$ that reduces to an affine function on each standard simplex is a shifted linear combination of term functions:

$$f(\mathbf{x}) = \sum_{S \subseteq [n]} m_f(S)\, x_{1:S}, \tag{23}$$

where $m_f\colon 2^{[n]} \to \mathbb{R}$ is the *Möbius transform* of v_f, defined as

$$m_f(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|}\, v_f(T).$$

Indeed, expression (23) reduces to an affine function on each standard simplex and agrees with $f(\mathbf{1}_S)$ at $\mathbf{1}_S$ for every $S \subseteq [n]$. Thus the class of shifted linear combinations of n-place term functions is precisely the class of n-place Lovász extensions.

Remark 2. A nondecreasing Lovász extension $f\colon \mathbb{I}^n \to \mathbb{R}$ such that $f(\mathbf{0}) = 0$ is also called a *discrete Choquet integral*. For general background, see for instance [5].

For every nonempty $S \subseteq [n]$ and every $k \in [|S|]$, the function $\mathrm{os}_{k:S}$ is a Lovász extension and, from (12), we have

$$x_{k:S} = \sum_{\substack{T \subseteq S \\ |T| \geqslant k}} (-1)^{|T|-k} \binom{|T|-1}{k-1} x_{|T|:T}$$

The following proposition gives a concise expression for the index $I(\mathrm{os}_{j:S}, k)$. We first compute the action of the symmetrizer Sym on such functions.

Lemma 4. *For every nonempty $S \subseteq [n]$ and every $j \in [|S|]$, we have*

$$\mathrm{Sym}(\mathrm{os}_{j:S}) = \frac{1}{\binom{n}{|S|}} \sum_{\substack{T \subseteq [n] \\ |T|=|S|}} \mathrm{os}_{j:T}.$$

Proposition 9. *For every nonempty $S \subseteq [n]$, every $j \in [|S|]$, and every $k \in [n]$, we have*

$$I(\mathrm{os}_{j:S}, k) = \frac{\binom{k-1}{j-1}\binom{n-k}{|S|-j}}{\binom{n}{|S|}} \tag{24}$$

if $0 \leqslant k - j \leqslant n - |S|$, and 0, otherwise.

The following proposition gives an explicit expression for the index $I(f, k)$ when f is a Lovász extension.

Proposition 10. *If $f\colon \mathbb{I}^n \to \mathbb{R}$ is a Lovász extension, then*

$$f(\mathbf{x}) = f(\mathbf{0}) + \sum_{i=1}^{n} x_{(i)}\, D_{(i)} f(\mathbf{x}). \tag{25}$$

Moreover, for every $k \in [n]$, we have

$$I(f,k) = \overline{v}_f(n-k+1) - \overline{v}_f(n-k) = \sum_{s=1}^{n-k+1} \binom{n-k}{s-1} \overline{m}_f(s)\,, \tag{26}$$

where $\overline{v}_f(s) = \binom{n}{s}^{-1} \sum_{S \subseteq [n]:|S|=s} v_f(S)$ and $\overline{m}_f(s) = \binom{n}{s}^{-1} \sum_{S \subseteq [n]:|S|=s} m_f(S)$.

We can readily see that the shifted L-statistic functions are precisely the symmetric Lovász extensions. From this observation we derive the following result.

Proposition 11. *For any Lovász extension $f\colon \mathbb{I}^n \to \mathbb{R}$, we have $f_L = \mathrm{Sym}(f)$ and*

$$\mathrm{Sym}(f) = f(\mathbf{0}) + \sum_{i=1}^{n} I(f,i)\, \mathrm{os}_i.$$

6 Applications

We briefly discuss some applications of the influence index in aggregation theory and statistics.

6.1 Influence Index in Aggregation Theory

Several indexes (such as interaction, tolerance, and dispersion indexes) have been proposed and investigated in aggregation theory to better understand the general behavior of aggregation functions with respect to their variables; see

[5, Chap. 10]. These indexes enable one to classify the aggregation functions according to their behavioral properties. The index $I(f,k)$ can also be very informative and thus contribute to such a classification. As an example, we have computed this index for the arithmetic mean and geometric mean functions (see Remark 1 and Example 1) and we can observe for instance that the smallest variable $x_{(1)}$ has a larger influence on the latter function.

Remark 3. Noteworthy aggregation functions are the so-called conjunctive aggregation functions, that is, nondecreasing functions $f\colon \mathbb{I}^n \to \mathbb{R}$ satisfying $0 \leqslant f(\mathbf{x}) \leqslant x_{(1)}$; see [5, Chap. 3]. Although these functions are bounded from above by $x_{(1)}$, the index $I(f,k)$ need not be maximum for $k=1$. For instance, for the binary conjunctive aggregation function

$$f(x_1,x_2) = \begin{cases} 0, & \text{if } x_1 \vee x_2 < \frac{3}{4}, \\ x_1 \wedge x_2 \wedge \frac{1}{4}, & \text{otherwise,} \end{cases}$$

we have $I(f,1) = \frac{17}{128}$ and $I(f,2) = \frac{19}{64}$, and hence $I(f,1) < I(f,2)$.

In the framework of aggregation functions, it can be natural to consider and identify the functions $f \in L^2(\mathbb{I}^n)$ for which the order statistics are equally influent, that is, such that $I(f,k) = I(f,1)$ for all $k \in [n]$. As far as the Lovász extensions are concerned, we have the following result, which can be easily derived from Proposition 10 and the immediate identities

$$\overline{v}_f(s) = \sum_{t=0}^{s} \binom{s}{t} \overline{m}_f(t) \quad \text{and} \quad \overline{m}_f(s) = \sum_{t=0}^{s} (-1)^{s-t} \binom{s}{t} \overline{v}_f(t).$$

Proposition 12. *If $f\colon \mathbb{I}^n \to \mathbb{R}$ is a Lovász extension, then the following are equivalent.*

(a) We have $I(f,k) = I(f,1)$ for all $k \in [n]$.
(b) The sequence $(\overline{v}_f(s))_{s=0}^n$ is in arithmetic progression.
(c) We have $\overline{m}_f(s) = 0$ for $s = 2,\ldots,n$.

This proposition can be easily interpreted. In view of Corollary 2, Lemma 4 and Proposition 11, we see that conditions (b) and (c) above are two equivalent ways to express that the symmetric part of the Lovász extension f is the arithmetic mean, up to a multiplicative and an additive constant.

6.2 Influence Index in Statistics

It can be informative to assess the influence of every order statistic on a given statistic to measure, e.g., its behavior with respect to the extreme values. From this information we can also approximate the given statistic by a shifted L-statistic. Of course, for L-statistics (such as Winsorized means, trimmed means, linearly weighted means, quasi-ranges, Gini's mean difference; see [4, §6.3, §8.8, §9.4]), the computation of the influence indexes is immediate. However, for some other statistics such as the central moments, the indexes can be computed via (18)–(20).

Example 2. The closest shifted L-statistic to the variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is given by

$$\sigma_L^2 = \frac{1-n^2}{12n(n+3)} + \sum_{k=1}^{n} I(\sigma^2, k)\, X_{(k)},$$

with $I(\sigma^2, k) = (n+2)(2k-n-1)/(n^2(n+3))$, which can be computed from (20). We then immediately see that the smallest and largest variables are the most influent.

Acknowledgments

The authors wish to thank Samuel Nicolay for fruitful discussions. This research is supported by the internal research project F1R-MTH-PUL-09MRDO of the University of Luxembourg.

References

1. Balakrishnan, N., Rao, C.R. (eds.): Order statistics: theory & methods. Handbook of Statist, vol. 16. North-Holland, Amsterdam (1998)
2. Burris, S., Sankappanavar, H.P.: A course in universal algebra. Graduate Texts in Mathematics, vol. 78. Springer, New York (1981)
3. David, F.N., Johnson, N.L.: Statistical treatment of censored data. I. Fundamental formulae. Biometrika 41, 228–240 (1954)
4. David, H., Nagaraja, H.: Order statistics, 3rd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester (2003)
5. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation functions. Encyclopedia of Mathematics and its Applications, vol. 127. Cambridge University Press, Cambridge (2009)
6. Hammer, P., Holzman, R.: Approximations of pseudo-Boolean functions; applications to game theory. Z. Oper. Res. 36(1), 3–21 (1992)
7. Kahn, J., Kalai, G., Linial, N.: The influence of variables on Boolean functions. In: Proc. 29th Annual Symposium on Foundations of Computational Science, pp. 68–80. Computer Society Press (1988)
8. Marichal, J.-L.: The influence of variables on pseudo-Boolean functions with applications to game theory and multicriteria decision making. Discrete Appl. Math. 107(1-3), 139–164 (2000)
9. Marichal, J.-L., Mathonet, P.: Measuring the interactions among variables of functions over the unit hypercube, arXiv:0912.1547
10. Marichal, J.-L., Mathonet, P.: Weighted Banzhaf interaction index through weighted approximations of games, arXiv:1001.3052
11. Singer, I.: Extensions of functions of 0-1 variables and applications to combinatorial optimization. Numer. Funct. Anal. Optimization 7, 23–62 (1984)

Measuring the Interactions among Variables of Functions over the Unit Hypercube

Jean-Luc Marichal and Pierre Mathonet

Mathematics Research Unit, FSTC, University of Luxembourg,
6, rue Coudenhove-Kalergi, L-1359 Luxembourg,
Grand Duchy of Luxembourg
jean-luc.marichal@uni.lu, pierre.mathonet@uni.lu

Abstract. By considering a least squares approximation of a given square integrable function $f\colon [0,1]^n \to \mathbb{R}$ by a multilinear polynomial of a specified degree, we define an index which measures the overall interaction among variables of f. This definition extends the concept of Banzhaf interaction index introduced in cooperative game theory. Our approach is partly inspired from multilinear regression analysis, where interactions among the independent variables are taken into consideration. We show that this interaction index has appealing properties which naturally generalize the properties of the Banzhaf interaction index. In particular, we interpret this index as an expected value of the difference quotients of f or, under certain natural conditions on f, as an expected value of the derivatives of f. These interpretations show a strong analogy between the introduced interaction index and the overall importance index defined by Grabisch and Labreuche [7]. Finally, we discuss a few applications of the interaction index.

1 Introduction

Sophisticated mathematical models are extensively used in a variety of areas of mathematics and physics, and especially in applied fields such as engineering, life sciences, economics, finance, and many others. Here we consider the simple situation where the model aims at explaining a single dependent variable, call it y, in terms of n independent variables $x_1, \ldots, x_n$. Such a model is usually described through an equation of the form

$$y = f(x_1, \ldots, x_n),$$

where f is a real function of n variables.

Now, suppose that the function f describing the model is given and that we want to investigate its behavior through simple terms. For instance, suppose we want to measure the overall contribution (importance or influence) of each independent variable to the model. A natural approach to this problem consists in defining the overall importance of each variable as the coefficient of this variable

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 19–30, 2010.

in the least squares linear approximation of f. This approach was considered by Hammer and Holzman [11] for pseudo-Boolean functions and cooperative games $f\colon \{0,1\}^n \to \mathbb{R}$. Interestingly enough, they observed that the coefficient of each variable in the linear approximation is exactly the Banzhaf power index [2,5] of the corresponding player in the game f.

In many practical situations, the information provided by the overall importance degree of each variable may be far insufficient due to the possible interactions among the variables. Then, a more flexible approach to investigate the behavior of f consists in measuring an overall importance degree for each combination (subset) of variables. Such a concept was first introduced in [13] for Boolean functions $f\colon \{0,1\}^n \to \{0,1\}$ (see also [3,4]), then in [14] for pseudo-Boolean functions and games $f\colon \{0,1\}^n \to \mathbb{R}$ (see also [15]), and in [7] for square integrable functions $f\colon [0,1]^n \to \mathbb{R}$.

In addition to these importance indexes, we can also measure directly the interaction degree among the variables by defining an overall interaction index for each combination of variables. This concept was introduced axiomatically in [10] (see also [6]) for games $f\colon \{0,1\}^n \to \mathbb{R}$. However, it has not yet been extended to real functions defined on $[0,1]^n$. In this paper we intend to fill this gap by defining and investigating an appropriate index to measure the interaction degree among variables of a given square integrable function $f\colon [0,1]^n \to \mathbb{R}$. Our sources of inspiration to define such an index are actually threefold:

In cooperative game theory. Interaction indexes were introduced axiomatically a decade ago [10] for games $f\colon \{0,1\}^n \to \mathbb{R}$ (see also [6]). The best known interaction indexes are the Banzhaf and Shapley interaction indexes, which extend the Banzhaf and Shapley power indexes. Following Hammer and Holzman's approach [11], it was shown in [9] that the Banzhaf interaction index can be obtained from least squares approximations of the game under consideration by games whose multilinear representations are of lower degrees.

In analysis. Considering a sufficiently differentiable real function f of several variables, the *local* interaction among certain variables at a given point $\mathbf{a}$ can be obtained through the coefficients of the Taylor expansion of f at $\mathbf{a}$, that is, through the coefficients of the *local* polynomial approximation of f at $\mathbf{a}$. By contrast, if we want to define an *overall* interaction index, we naturally have to consider a *global* approximation of f by a polynomial function.

In statistics. Multilinear statistical models have been proposed to take into account the interaction among the independent variables (see for instance [1]): two-way interactions appear as the coefficients of leading terms in quadratic models, three-way interactions appear as the coefficients of leading terms in cubic models, and so forth.

On the basis of these observations, we naturally consider the least squares approximation problem of a given square integrable function $f\colon [0,1]^n \to \mathbb{R}$ by a polynomial of a given degree. As multiple occurrences in combinations of variables are not relevant, we will only consider multilinear polynomial functions. Then, given a subset $S \subseteq \{1,\ldots,n\}$, an index $\mathcal{I}(f,S)$ measuring the interaction

among the variables $\{x_i : i \in S\}$ of f is defined as the coefficient of the monomial $\prod_{i\in S} x_i$ in the best approximation of f by a multilinear polynomial of degree at most $|S|$. This definition is given and discussed in Section 2.

In Section 3 we show that this new index has many appealing properties, such as linearity, continuity, and symmetry. In particular, we show that, similarly to the Banzhaf interaction index introduced for games, the index $\mathcal{I}(f, S)$ can be interpreted in a sense as an expected value of the discrete derivative of f in the direction of S (Theorem 2) or, equivalently, as an expected value of the difference quotient of f in the direction of S (Corollary 1). Under certain natural conditions on f, the index can also be interpreted as an expected value of the derivative of f in the direction of S (Proposition 4). These latter results reveal a strong analogy between the interaction index and the overall importance index introduced by Grabisch and Labreuche [7].

In Section 4 we discuss the computation of explicit expressions of the interaction index for certain classes of functions, namely pseudo-multilinear polynomials and discrete Choquet integrals.

We employ the following notation throughout the paper. Let $\mathbb{I}^n$ denote the n-dimensional unit cube $[0,1]^n$. We denote by $F(\mathbb{I}^n)$ the class of all functions $f\colon \mathbb{I}^n \to \mathbb{R}$ and by $L^2(\mathbb{I}^n)$ the subclass of square integrable functions $f\colon \mathbb{I}^n \to \mathbb{R}$ modulo equality almost everywhere. For any $S \subseteq N = \{1, \ldots, n\}$, we denote by $\mathbf{1}_S$ the characteristic vector of S in $\{0,1\}^n$.

2 Interaction Indexes

In this section we first recall the concepts of power and interaction indexes introduced in cooperative game theory and how the Banzhaf index can be obtained from the solution of a least squares approximation problem. Then we show how this approximation problem can be extended to functions in $L^2(\mathbb{I}^n)$ and, from this extension, we introduce an interaction index for such functions.

Recall that a (*cooperative*) *game* on a finite set of players $N = \{1, \ldots, n\}$ is a set function $v\colon 2^N \to \mathbb{R}$ which assigns to each coalition S of players a real number $v(S)$ representing the *worth* of S.[1] Through the usual identification of the subsets of N with the elements of $\{0,1\}^n$, a game $v\colon 2^N \to \mathbb{R}$ can be equivalently described by a pseudo-Boolean function $f\colon \{0,1\}^n \to \mathbb{R}$. The correspondence is given by $v(S) = f(\mathbf{1}_S)$ and

$$f(\mathbf{x}) = \sum_{S\subseteq N} v(S) \prod_{i\in S} x_i \prod_{i\in N\setminus S} (1 - x_i). \tag{1}$$

Equation (1) shows that any pseudo-Boolean function $f\colon \{0,1\}^n \to \mathbb{R}$ can always be represented by a multilinear polynomial of degree at most n (see [12]), which can be further simplified into

$$f(\mathbf{x}) = \sum_{S\subseteq N} a(S) \prod_{i\in S} x_i\,, \tag{2}$$

[1] Usually, the condition $v(\varnothing) = 0$ is required for v to define a game. However, we do not need this restriction in the present paper.

where the set function $a\colon 2^N \to \mathbb{R}$, called the *Möbius transform* of v, is defined by

$$a(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|}\, v(T).$$

Let $\mathcal{G}^N$ denote the set of games on N. A *power index* [17] on N is a function $\phi\colon \mathcal{G}^N \times N \to \mathbb{R}$ that assigns to every player $i \in N$ in a game $f \in \mathcal{G}^N$ his/her prospect $\phi(f,i)$ from playing the game. An *interaction index* [10] on N is a function $I\colon \mathcal{G}^N \times 2^N \to \mathbb{R}$ that measures in a game $f \in \mathcal{G}^N$ the interaction degree among the players of a coalition $S \subseteq N$.

For instance, the *Banzhaf interaction index* [10] of a coalition $S \subseteq N$ in a game $f \in \mathcal{G}^N$ can be defined (in terms of the Möbius transformation of f) by

$$I_{\mathrm{B}}(f,S) = \sum_{T \supseteq S} \left(\frac{1}{2}\right)^{|T|-|S|} a(T), \tag{3}$$

and the *Banzhaf power index* [5] of a player $i \in N$ in a game $f \in \mathcal{G}^N$ is defined by $\phi_{\mathrm{B}}(f,i) = I_{\mathrm{B}}(f,\{i\})$.

It is noteworthy that $I_{\mathrm{B}}(f,S)$ can be interpreted as an average of the *S-difference* (or *discrete S-derivative*) $\Delta^S f$ of f. Indeed, it also writes (see [9, §2])

$$I_{\mathrm{B}}(f,S) = \frac{1}{2^n} \sum_{\mathbf{x} \in \{0,1\}^n} (\Delta^S f)(\mathbf{x}), \tag{4}$$

where $\Delta^S f$ is defined inductively by $\Delta^{\varnothing} f = f$ and $\Delta^S f = \Delta^{\{i\}} \Delta^{S\setminus\{i\}} f$ for $i \in S$, with $\Delta^{\{i\}} f(\mathbf{x}) = f(\mathbf{x} \mid x_i = 1) - f(\mathbf{x} \mid x_i = 0)$.

We now recall how the Banzhaf interaction index can be obtained from a least squares approximation problem. For $k \in \{0,\ldots,n\}$, denote by V_k the set of all multilinear polynomials $g\colon \{0,1\}^n \to \mathbb{R}$ of degree at most k, that is of the form

$$g(\mathbf{x}) = \sum_{\substack{S \subseteq N \\ |S| \leqslant k}} c(S) \prod_{i \in S} x_i\,, \tag{5}$$

where the coefficients $c(S)$ are real numbers. For a given pseudo-Boolean function $f\colon \{0,1\}^n \to \mathbb{R}$, the best kth approximation of f is the unique multilinear polynomial $f_k \in V_k$ that minimizes the distance $\sum_{\mathbf{x} \in \{0,1\}^n} (f(\mathbf{x}) - g(\mathbf{x}))^2$ among all $g \in V_k$. A closed-form expression of f_k was given in [11] for $k = 1$ and $k = 2$ and in [9] for arbitrary $k \leqslant n$. In fact, when f is given in its multilinear form (2) we obtain

$$f_k(\mathbf{x}) = \sum_{\substack{S \subseteq N \\ |S| \leqslant k}} a_k(S) \prod_{i \in S} x_i,$$

where

$$a_k(S) = a(S) + (-1)^{k-|S|} \sum_{\substack{T \supseteq S \\ |T| > k}} \binom{|T|-|S|-1}{k-|S|} \left(\frac{1}{2}\right)^{|T|-|S|} a(T).$$

It is then easy to see that

$$I_{\mathrm{B}}(f,S) = a_{|S|}(S). \tag{6}$$

Thus, $I_{\mathrm{B}}(f,S)$ is exactly the coefficient of the monomial $\prod_{i\in S} x_i$ in the best approximation of f by a multilinear polynomial of degree at most $|S|$.

Taking into account this approximation problem, we now define an interaction index for functions in $L^2(\mathbb{I}^n)$ as follows. Denote by W_k the set of all multilinear polynomials $g\colon \mathbb{I}^n \to \mathbb{R}$ of degree at most k. Clearly, these functions are also of the form (5). For a given function $f \in L^2(\mathbb{I}^n)$, we define the *best kth (multilinear) approximation of* f as the multilinear polynomial $f_k \in W_k$ that minimizes the distance

$$\int_{\mathbb{I}^n} \big(f(\mathbf{x}) - g(\mathbf{x})\big)^2 \, d\mathbf{x} \tag{7}$$

among all $g \in W_k$.

It is easy to see that W_k is a linear subspace of $L^2(\mathbb{I}^n)$ of dimension $\sum_{s=0}^{k} \binom{n}{s}$. Indeed, W_k is the linear span of the basis $B_k = \{v_S : S \subseteq N,\ |S| \leqslant k\}$, where the functions $v_S\colon \mathbb{I}^n \to \mathbb{R}$ are defined by $v_S(\mathbf{x}) = \prod_{i\in S} x_i$. Note that formula (7) also writes $\|f-g\|^2$ where $\|\cdot\|$ is the standard norm of $L^2(\mathbb{I}^n)$ associated with the inner product $\langle f,g\rangle = \int_{\mathbb{I}^n} f(\mathbf{x})g(\mathbf{x})\,d\mathbf{x}$. Therefore, using the general theory of Hilbert spaces, the solution of this approximation problem exists and is uniquely determined by the orthogonal projection of f onto W_k. This projection can be easily expressed in any orthonormal basis of W_k. But here it is very easy to see that the set $B'_k = \{w_S : S \subseteq N,\ |S| \leqslant k\}$, where $w_S\colon \mathbb{I}^n \to \mathbb{R}$ is given by

$$w_S(\mathbf{x}) = 12^{|S|/2} \prod_{i\in S} \Big(x_i - \frac{1}{2}\Big) = 12^{|S|/2} \sum_{T\subseteq S} \Big(-\frac{1}{2}\Big)^{|S|-|T|} v_T(\mathbf{x}),$$

forms such an orthonormal basis for W_k (actually, this basis can be obtained from B_k via Gram Schmidt orthogonalization).

The following immediate theorem gives the components of the best kth approximation of a function $f \in L^2(\mathbb{I}^n)$ in the bases B_k and B'_k.

Theorem 1. *For every $k \in \{0,\ldots,n\}$, the best kth approximation of $f \in L^2(\mathbb{I}^n)$ is the function*

$$f_k = \sum_{\substack{T\subseteq N\\ |T|\leqslant k}} \langle f, w_T\rangle\, w_T \;=\; \sum_{\substack{S\subseteq N\\ |S|\leqslant k}} a_k(S)\, v_S\,, \tag{8}$$

where

$$a_k(S) = \sum_{\substack{T\supseteq S\\ |T|\leqslant k}} \Big(-\frac{1}{2}\Big)^{|T|-|S|} 12^{|T|/2}\, \langle f, w_T\rangle. \tag{9}$$

By analogy with (6), to measure the interaction degree among variables of an arbitrary function $f \in L^2(\mathbb{I}^n)$, we naturally define an index $\mathcal{I}\colon L^2(\mathbb{I}^n)\times 2^N \to \mathbb{R}$ as $\mathcal{I}(f,S) = a_{|S|}(S)$, where $a_{|S|}(S)$ is obtained from f by (9). We will see in the next section that this index indeed measures an importance degree when $|S| = 1$ and an interaction degree when $|S| \geqslant 2$.

Definition 1. *Let* $\mathcal{I}\colon L^2(\mathbb{I}^n) \times 2^N \to \mathbb{R}$ *be defined as* $\mathcal{I}(f,S) = 12^{|S|/2}\langle f, w_S\rangle$, *that is,*

$$\mathcal{I}(f,S) = 12^{|S|} \int_{\mathbb{I}^n} f(\mathbf{x}) \prod_{i\in S} \left(x_i - \frac{1}{2}\right) d\mathbf{x}. \tag{10}$$

Thus we have defined an interaction index from an approximation (projection) problem. Conversely, this index characterizes this approximation problem. Indeed, as the following result shows, the best kth approximation of $f \in L^2(\mathbb{I}^n)$ is the unique function of W_k that preserves the interaction index for all the s-subsets such that $s \leqslant k$. The discrete analogue of this result was established in [9] for the Banzhaf interaction index (3).

Proposition 1. *A function* $f_k \in W_k$ *is the best* k*th approximation of* $f \in L^2(\mathbb{I}^n)$ *if and only if* $\mathcal{I}(f,S) = \mathcal{I}(f_k,S)$ *for all* $S \subseteq N$ *such that* $|S| \leqslant k$.

3 Properties and Interpretations

Most of the interaction indexes defined for games, including the Banzhaf interaction index, share a set of fundamental properties such as linearity, symmetry, and k-monotonicity (see [6]). Many of them can also be expressed as expected values of the discrete derivatives (differences) of their arguments (see for instance (4)). In this section we show that the index $\mathcal{I}$ fulfills direct generalizations of these properties to the framework of functions of $L^2(\mathbb{I}^n)$. In particular, we show that $\mathcal{I}(f,S)$ can be interpreted as an expected value of the difference quotient of f in the direction of S or, under certain natural conditions on f, as an expected value of the derivative of f in the direction of S.

The first result follows from the very definition of the index.

Proposition 2. *For every* $S \subseteq N$, *the mapping* $f \mapsto \mathcal{I}(f,S)$ *is linear and continuous.*

Recall that if π is a permutation on N, then, for every function $f \in F(\mathbb{I}^n)$, the permutation π acts on f by $\pi(f)(x_1,\ldots,x_n) = f(x_{\pi(1)},\ldots,x_{\pi(n)})$. The following result is then an easy consequence of the change of variables theorem.

Proposition 3. *The index* $\mathcal{I}$ *is symmetric. That is, for every permutation* π *on* N, *every* $f \in L^2(\mathbb{I}^n)$, *and every* $S \subseteq N$, *we have* $\mathcal{I}(\pi(f),\pi(S)) = \mathcal{I}(f,S)$.

We now provide an interpretation of $\mathcal{I}(f,S)$ as an expected value of the S-derivative $D^S f$ of f. The proof immediately follows from repeated integrations by parts of (10) and thus is omitted.

For $S \subseteq N$, denote by h_S the probability density function of independent beta distributions on $\mathbb{I}^n$ with parameters $\alpha = \beta = 2$, that is, $h_S(\mathbf{x}) = 6^{|S|}\prod_{i\in S} x_i (1-x_i)$.

Proposition 4. *For every* $S \subseteq N$ *and every* $f \in L^2(\mathbb{I}^n)$ *such that* $D^T f$ *is continuous and integrable on* $]0,1[^n$ *for all* $T \subseteq S$, *we have*

$$\mathcal{I}(f,S) = \int_{\mathbb{I}^n} h_S(\mathbf{x})\, D^S f(\mathbf{x})\, d\mathbf{x}. \tag{11}$$

Remark 1. (a) Formulas (4) and (11) show a strong analogy between the indexes I_{B} and $\mathcal{I}$. Indeed, $I_{\mathrm{B}}(f,S)$ is the expected value of the S-difference of f with respect to the discrete uniform distribution whereas $\mathcal{I}(f,S)$ is the expected value of the S-derivative of f with respect to a beta distribution. We will see in Theorem 2 a similar interpretation of $\mathcal{I}(f,S)$ which does not require all the assumptions of Proposition 4.

(b) Propositions 1 and 4 reveal an analogy between least squares approximations and Taylor expansion formula. Indeed, while the k-degree Taylor expansion of f at a given point $\mathbf{a}$ can be seen as the unique polynomial of degree at most k whose derivatives at $\mathbf{a}$ coincide with the derivatives of f at the same point, the best kth approximation of f is the unique multilinear polynomial of degree at most k that agrees with f in all average S-derivatives for $|S| \leqslant k$.

We now give an alternative interpretation of $\mathcal{I}(f,S)$ as an expected value, which does not require the additional assumptions of Proposition 4. In this more general framework, we naturally replace the derivative with a difference quotient. To this extent, we introduce some further notation. As usual, we denote by $\mathbf{e}_i$ the ith vector of the standard basis for $\mathbb{R}^n$. For every $S \subseteq N$ and every $\mathbf{h} \in \mathbb{I}^n$, we define the *$S$-shift* operator $E^S_{\mathbf{h}}$ on $F(\mathbb{I}^n)$ by

$$E^S_{\mathbf{h}} f(\mathbf{x}) = f\left(\mathbf{x} + \sum_{j \in S} h_j \mathbf{e}_j\right)$$

for every $\mathbf{x} \in \mathbb{I}^n$ such that $\mathbf{x} + \mathbf{h} \in \mathbb{I}^n$.

We also define the *S-difference* (or *discrete S-derivative*) operator $\Delta^S_{\mathbf{h}}$ on $F(\mathbb{I}^n)$ inductively by $\Delta^{\varnothing}_{\mathbf{h}} f = f$ and $\Delta^S_{\mathbf{h}} f = \Delta^{\{i\}}_{\mathbf{h}} \Delta^{S\setminus\{i\}}_{\mathbf{h}} f$ for $i \in S$, with $\Delta^{\{i\}}_{\mathbf{h}} f(\mathbf{x}) = E^{\{i\}}_{\mathbf{h}} f(\mathbf{x}) - f(\mathbf{x})$. Similarly, we define the *S-difference quotient* operator $Q^S_{\mathbf{h}}$ on $F(\mathbb{I}^n)$ by $Q^{\varnothing}_{\mathbf{h}} f = f$ and $Q^S_{\mathbf{h}} f = Q^{\{i\}}_{\mathbf{h}} Q^{S\setminus\{i\}}_{\mathbf{h}} f$ for $i \in S$, with $Q^{\{i\}}_{\mathbf{h}} f(\mathbf{x}) = \frac{1}{h_i} \Delta^{\{i\}}_{\mathbf{h}} f(\mathbf{x})$.

The next straightforward lemma provides a direct link between the difference operators and the shift operators. It actually shows that, for every fixed $\mathbf{h} \in \mathbb{I}^n$, the map $S \mapsto \Delta^S_{\mathbf{h}}$ is nothing other than the Möbius transform of the map $S \mapsto E^S_{\mathbf{h}}$.

Lemma 1. *For every $f \in F(\mathbb{I}^n)$ and every $S \subseteq N$, we have*

$$\Delta^S_{\mathbf{h}} f(\mathbf{x}) = \sum_{T \subseteq S} (-1)^{|S|-|T|} E^T_{\mathbf{h}} f(\mathbf{x}). \tag{12}$$

Let us interpret the S-difference operator through a simple example. For $n = 3$ and $S = \{1,2\}$, we have

$$\Delta^S_{\mathbf{h}} f(\mathbf{x}) = f(x_1+h_1, x_2+h_2, x_3) - f(x_1+h_1, x_2, x_3) - f(x_1, x_2+h_2, x_3) + f(x_1, x_2, x_3).$$

In complete analogy with the discrete concept of marginal interaction among players in a coalition $S \subseteq N$ (see [9, §2]), the value $\Delta^S_{\mathbf{h}} f(\mathbf{x})$ can be interpreted

as the *marginal interaction* among variables x_i ($i \in S$) at $\mathbf{x}$ with respect to the increases h_i for $i \in S$.

Setting $\mathbf{h} = \mathbf{y} - \mathbf{x}$ in the example above, we obtain

$$\Delta^S_{\mathbf{y}-\mathbf{x}} f(\mathbf{x}) = f(y_1, y_2, x_3) - f(y_1, x_2, x_3) - f(x_1, y_2, x_3) + f(x_1, x_2, x_3).$$

If $x_i \leqslant y_i$ for every $i \in S$, then $\Delta^S_{\mathbf{y}-\mathbf{x}} f(\mathbf{x})$ is naturally called the *f-volume* of the box $\prod_{i\in S}[x_i, y_i]$. The following straightforward lemma shows that, when $f = v_S$, $\Delta^S_{\mathbf{y}-\mathbf{x}} f(\mathbf{x})$ is exactly the volume of the box $\prod_{i\in S}[x_i, y_i]$.

Lemma 2. *For every $S \subseteq N$, we have $\Delta^S_{\mathbf{y}-\mathbf{x}} v_S(\mathbf{x}) = \prod_{i\in S}(y_i - x_i)$.*

In the remaining part of this paper, the notation $\mathbf{y}_S \in [\mathbf{x}_S, \mathbf{1}]$ means that $y_i \in [x_i, 1]$ for every $i \in S$.

Theorem 2. *For every $f \in L^2(\mathbb{I}^n)$ and every $S \subseteq N$, we have*

$$\mathcal{I}(f, S) = \frac{1}{\mu(S)} \int_{\mathbf{x}\in\mathbb{I}^n} \int_{\mathbf{y}_S\in[\mathbf{x}_S,\mathbf{1}]} \Delta^S_{\mathbf{y}-\mathbf{x}} f(\mathbf{x})\, d\mathbf{y}_S\, d\mathbf{x}, \tag{13}$$

where

$$\mu(S) = \int_{\mathbf{x}\in\mathbb{I}^n} \int_{\mathbf{y}_S\in[\mathbf{x}_S,\mathbf{1}]} \Delta^S_{\mathbf{y}-\mathbf{x}} v_S(\mathbf{x})\, d\mathbf{y}_S\, d\mathbf{x} = 6^{-|S|}.$$

Remark 2. (a) By Lemma 2, we see that $\mathcal{I}(f, S)$ can be interpreted as the average f-volume of the box $\prod_{i\in S}[x_i, y_i]$ divided by its average volume, when $\mathbf{x}$ and $\mathbf{y}_S$ are chosen at random with the uniform distribution.

(b) As already mentioned in Remark 1(a), Theorem 2 appears as a natural generalization of formula (4) (similarly to Proposition 4) in the sense that the marginal interaction $\Delta^S_{\mathbf{h}} f(\mathbf{x})$ at $\mathbf{x}$ is averaged over the whole domain $\mathbb{I}^n$ (instead of its vertices).

(c) We note an analogy between formula (13) and the importance index defined by Grabisch and Labreuche in [7, Theorem 1]. Indeed, up to the normalization constant, this importance index is obtained by replacing in formula (13) the operator $\Delta^S_{\mathbf{y}-\mathbf{x}}$ by $E^S_{\mathbf{y}-\mathbf{x}} - I$. Moreover, when S is a singleton, both operators coincide and so do the normalization constants.

As an immediate consequence of Theorem 2, we have the following interpretation of the index $\mathcal{I}$ as an expected value of the difference quotients of its argument with respect to some probability distribution.

Corollary 1. *For every $f \in L^2(\mathbb{I}^n)$ and every $S \subseteq N$, we have*

$$\mathcal{I}(f, S) = \int_{\mathbf{x}\in\mathbb{I}^n} \int_{\mathbf{y}_S\in[\mathbf{x}_S,\mathbf{1}]} p_S(\mathbf{x}, \mathbf{y}_S)\, Q^S_{\mathbf{y}-\mathbf{x}} f(\mathbf{x})\, d\mathbf{y}_S\, d\mathbf{x},$$

where the function $p_S(\mathbf{x}, \mathbf{y}_S) = 6^{|S|} \prod_{i\in S}(y_i - x_i)$ defines a probability density function on the set $\{(\mathbf{x}, \mathbf{y}_S) : \mathbf{x} \in \mathbb{I}^n, \mathbf{y}_S \in [\mathbf{x}_S, \mathbf{1}]\}$.

Let us now analyze the behavior of the interaction index $\mathcal{I}$ on some special classes of functions. The following properties generalize in a very natural way to our setting the behavior of the Banzhaf interaction index I_B with respect to the presence of null players and dummy coalitions.

Recall that a null player in a game (or a set function) $v \in \mathcal{G}^N$ is a player $i \in N$ such that $v(T \cup \{i\}) = v(T)$ for every $T \subseteq N \setminus \{i\}$. Equivalently, the corresponding pseudo-Boolean function $f\colon \{0,1\}^n \to \mathbb{R}$, given by (1), is independent of x_i. The notion of null player for games is then naturally extended through the notion of ineffective variables for functions in $F(\mathbb{I}^n)$ as follows. A variable x_i $(i \in N)$ is said to be *ineffective* for a function f in $F(\mathbb{I}^n)$ if $f(\mathbf{x}) = E^{\{i\}}_{-\mathbf{x}} f(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{I}^n$, or equivalently, if $\Delta^{\{i\}}_{\mathbf{y}-\mathbf{x}} f(\mathbf{x}) = 0$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$.

Define $I_f = \{i \in N : x_i \text{ ineffective for } f\}$. From either (10) or (13), we immediately derive the following result, which states that any combination of variables containing at least one ineffective variable for a function $f \in L^2(\mathbb{I}^n)$ has necessarily a zero interaction.

Proposition 5. *For every $f \in L^2(\mathbb{I}^n)$ and every $S \subseteq N$ such that $S \cap I_f \neq \varnothing$, we have $\mathcal{I}(f, S) = 0$.*

We say that a coalition $S \subseteq N$ is *dummy* in a game (or a set function) $v \in \mathcal{G}^N$ if $v(R \cup T) = v(R) + v(T) - v(\varnothing)$ for every $R \subseteq S$ and every $T \subseteq N \setminus S$. This means that $\{S, N \setminus S\}$ forms a partition of N such that, for every coalition $K \subseteq N$, the relative worth $v(K) - v(\varnothing)$ is the sum of the relative worths of its intersections with S and $N \setminus S$. It follows that a coalition S and its complement $N \setminus S$ are simultaneously dummy in any game $v \in \mathcal{G}^N$.

We propose the following extension of this concept.

Definition 2. *We say that a subset $S \subseteq N$ is* dummy *for a function $f \in F(\mathbb{I}^n)$ if $f(\mathbf{x}) = E^S_{-\mathbf{x}} f(\mathbf{x}) + E^{N \setminus S}_{-\mathbf{x}} f(\mathbf{x}) - f(\mathbf{0})$ for every $\mathbf{x} \in \mathbb{I}^n$.*

The following proposition gives an immediate interpretation of this definition.

Proposition 6. *A subset $S \subseteq N$ is* dummy *for a function $f \in F(\mathbb{I}^n)$ if and only if there exist functions $f_S, f_{N \setminus S} \in F(\mathbb{I}^n)$ such that $I_{f_S} \supseteq N \setminus S$, $I_{f_{N \setminus S}} \supseteq S$ and $f = f_S + f_{N \setminus S}$.*

The following result expresses the natural idea that interaction index for subsets that are properly partitioned by a dummy subset must be zero. It is an immediate consequence of Propositions 2, 5, and 6.

Proposition 7. *For every $f \in L^2(\mathbb{I}^n)$, every nonempty subset $S \subseteq N$ that is dummy for f, and every subset $K \subseteq N$ such that $K \cap S \neq \varnothing$ and $K \setminus S \neq \varnothing$, we have $\mathcal{I}(f, K) = 0$.*

4 Applications

We now calculate explicit expressions of the interaction index for two classes of functions, namely pseudo-multilinear polynomials and discrete Choquet integrals.

4.1 Pseudo-multilinear polynomials

As a first application, we derive an explicit expression of the index $\mathcal{I}$ for the class of pseudo-multilinear polynomials, that is, the class of multilinear polynomials with transformed variables.

Definition 3. *We say that a function $f \in L^2(\mathbb{I}^n)$ is a* pseudo-multilinear polynomial *if there exists a multilinear polynomial $g \in F(\mathbb{R}^n)$ and n unary functions $\varphi_1, \ldots, \varphi_n \in L^2(\mathbb{I})$ such that $f(\mathbf{x}) = g(\varphi_1(x_1), \ldots, \varphi_n(x_n))$ for every $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{I}^n$.*

Using expression (5) of multilinear polynomials, we immediately see that any pseudo-multilinear polynomial $f \in L^2(\mathbb{I}^n)$ can be written in the form

$$f(\mathbf{x}) = \sum_{T \subseteq N} a(T) \prod_{i \in T} \varphi_i(x_i).$$

The following result yields an explicit expression of the interaction index for this function in terms of the interaction indexes for the unary functions $\varphi_1, \ldots, \varphi_n$.

Proposition 8. *For every pseudo-multilinear polynomial $f \in L^2(\mathbb{I}^n)$ and every $S \subseteq N$, we have*

$$\mathcal{I}(f, S) = \sum_{T \supseteq S} a(T) \prod_{i \in T \setminus S} \mathcal{I}(\varphi_i, \varnothing) \prod_{i \in S} \mathcal{I}(\varphi_i, \{i\}).$$

Remark 3. Proposition 8 can actually be easily extended to functions of the form

$$f(\mathbf{x}) = \sum_{T \subseteq N} a(T) \prod_{i \in T} \varphi_i^T(x_i),$$

where $\varphi_i^T \in L^2(\mathbb{I})$ for $i = 1, \ldots, n$ and $T \subseteq N$.

An interesting subclass of pseudo-multilinear polynomials is the class of multiplicative functions, that is, functions of the form $f(\mathbf{x}) = \prod_{i=1}^n \varphi_i(x_i)$, where $\varphi_1, \ldots, \varphi_n \in L^2(\mathbb{I})$. For every multiplicative function $f \in L^2(\mathbb{I}^n)$ and every $S \subseteq N$, assuming $\mathcal{I}(f, \varnothing) \neq 0$, the ratio $\mathcal{I}(f, S)/\mathcal{I}(f, \varnothing)$ is also multiplicative in the sense that

$$\frac{\mathcal{I}(f, S)}{\mathcal{I}(f, \varnothing)} = \prod_{i \in S} \frac{\mathcal{I}(\varphi_i, \{i\})}{\mathcal{I}(\varphi_i, \varnothing)}. \tag{14}$$

4.2 The Discrete Choquet Integrals

A *discrete Choquet integral* is a function $f \in F(\mathbb{I}^n)$ of the form

$$f(\mathbf{x}) = \sum_{T \subseteq N} a(T) \min_{i \in T} x_i, \tag{15}$$

where the set function $a\colon 2^N \to \mathbb{R}$ is nondecreasing with respect to set inclusion and such that $a(\varnothing) = 0$ and $\sum_{S\subseteq N} a(S) = 1$.[2] These functions are mainly used in aggregation function theory and decision making. For general background, see for instance [8, Section 5.4].

The following proposition yields an explicit expression of the interaction index for the class of discrete Choquet integrals. We first consider a lemma and recall that the *beta function* is defined, for any integers $p, q > 0$, by

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt = \frac{(p-1)!(q-1)!}{(p+q-1)!}\,.$$

Proposition 9. *If $f \in F(\mathbb{I}^n)$ is of the form (15), then we have*

$$\mathcal{I}(f,S) = 6^{|S|} \sum_{T\supseteq S} a(T)\, B(|S|+1, |T|+1).$$

Remark 4. The map $a \mapsto \mathcal{I}(f,S) = 6^{|S|}\sum_{T\supseteq S} a(T)\, B(|S|+1,|T|+1)$ defines an interaction index, in the sense of [6], that is not a probabilistic index (see [6, Section 3.3]). However, if we normalize this interaction index (with respect to $|S|$) to get a probabilistic index, we actually divide $\mathcal{I}(f,S)$ by $6^{|S|}B(|S|+1,|S|+1)$ and retrieve the index I_M defined in [16].

Acknowledgments

The authors wish to thank Michel Beine, Miguel Couceiro, Paul Gérard, and Samuel Nicolay for fruitful discussions. This research is supported by the internal research project F1R-MTH-PUL-09MRDO of the University of Luxembourg.

References

1. Aiken, L.S., West, S.G.: Multiple Regression: Testing and Interpreting Interactions. Sage Publications, Newbury Park (1991)
2. Banzhaf, J.F.: Weighted voting doesn't work: A mathematical analysis. Rutgers Law Review 19, 317–343 (1965)
3. Ben-Or, M., Linial, N.: Collective coin flipping. In: Randomness and Computation, pp. 91–115. Academic Press, New York (1990); Earlier version: Collective coin flipping, robust voting games and minima of Banzhaf values. In: Proc. 26th IEEE Symposium on the Foundation of Computer Sciences, Portland, pp. 408–416 (1985)
4. Bourgain, J., Kahn, J., Kalai, G., Katznelson, Y., Linial, N.: The influence of variables in product spaces. Isr. J. Math. 77(1-2), 55–64 (1992)
5. Dubey, P., Shapley, L.S.: Mathematical properties of the Banzhaf power index. Math. Oper. Res. 4, 99–131 (1979)
6. Fujimoto, K., Kojadinovic, I., Marichal, J.-L.: Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. Games Econom. Behav. 55(1), 72–99 (2006)

[2] Whether the conditions on the set function a are assumed or not, the function given in (15) is also called the *Lovász extension* of the pseudo-Boolean function $f|_{\{0,1\}^n}$.

7. Grabisch, M., Labreuche, C.: How to improve acts: An alternative representation of the importance of criteria in MCDM. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 9(2), 145–157 (2001)
8. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation functions. Encyclopedia of Mathematics and its Applications, vol. 127. Cambridge University Press, Cambridge (2009)
9. Grabisch, M., Marichal, J.-L., Roubens, M.: Equivalent representations of set functions. Math. Oper. Res. 25(2), 157–178 (2000)
10. Grabisch, M., Roubens, M.: An axiomatic approach to the concept of interaction among players in cooperative games. Int. J. Game Theory 28(4), 547–565 (1999)
11. Hammer, P., Holzman, R.: Approximations of pseudo-Boolean functions; applications to game theory. Z. Oper. Res. 36(1), 3–21 (1992)
12. Hammer, P., Rudeanu, S.: Boolean methods in operations research and related areas. Springer, Heidelberg (1968)
13. Kahn, J., Kalai, G., Linial, N.: The influence of variables on Boolean functions. In: Proc. 29th Annual Symposium on Foundations of Computational Science, pp. 68–80. Computer Society Press (1988)
14. Marichal, J.-L.: The influence of variables on pseudo-Boolean functions with applications to game theory and multicriteria decision making. Discrete Appl. Math. 107(1-3), 139–164 (2000)
15. Marichal, J.-L., Kojadinovic, I., Fujimoto, K.: Axiomatic characterizations of generalized values. Discrete Applied Mathematics 155(1), 26–43 (2007)
16. Marichal, J.-L., Mathonet, P.: Approximations of Lovász extensions and their induced interaction index. Discrete Appl. Math. 156(1), 11–24 (2008)
17. Shapley, L.: A value for n-person games. In: Contributions to the Theory of Games II. Annals of Mathematics Studies, vol. 28, pp. 307–317. Princeton University Press, Princeton (1953)

Weighted Quasi-arithmetic Means and Conditional Expectations

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. In this paper, the weighted quasi-arithmetic means are discussed from the viewpoint of utility functions and background risks in economics, and they are represented by weighting functions and conditional expectations. Using these representations, an index for background risks in stochastic environments is derived through the weighted quasi-arithmetic means. The first-order stochastic dominance and the risk premium are demonstrated using the weighted quasi-arithmetic means and the aggregated mean ratios, and they are characterized by the background risk index. Finally, examples of the weighted quasi-arithmetic mean and the aggregated mean ratio for various typical utility functions are given.

1 Introduction

Weighted quasi-arithmetic means are important tools in the subjective estimation of data in decision making such as management, artificial intelligence and so on ([3,4,5]), and it is also strongly related to utility functions and background risks in economics ([6]). This paper analyzes quasi-arithmetic means of an interval through utility functions and weighting functions. Yoshida [12,13] has studied weighted quasi-arithmetic means of an interval by weighted aggregation operations from the viewpoint of subjective decision making where Kolmogorov [9] and Nagumo [10] studied the aggregation operators and Aczél [1] developed the theory regarding weighted aggregation. In this paper, we take a continuous strictly increasing function $f : [a, b] \mapsto (-\infty, \infty)$ as a decision maker's utility function, and we put a continuous function $w : [a, b] \mapsto (0, \infty)$ as a weighting function. Then we define a *weighted quasi-arithmetic mean* on a closed interval $[a, b]$ with the utility f in the background risk w by

$$f^{-1}\left(\int_a^b f(x)w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x\right).$$

Hence, it represents a *mean value* given by a real number $c(\in [a, b])$ satisfying

$$f(c)\int_a^b w(x)\,\mathrm{d}x = \int_a^b f(x)w(x)\,\mathrm{d}x$$

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 31–42, 2010.

in the *mean value theorem*. This paper discusses the weighted quasi-arithmetic means from the viewpoint of utility functions and background risks in economics. Representing the weighted quasi-arithmetic means by conditional expectations, we derive an index for risks in stochastic environments, and we also discuss the first-order stochastic dominance and the risk premium using the weighted quasi-arithmetic means and the aggregated mean ratios.

In Section 2, we give definitions of *the weighted quasi-arithmetic mean* and an *aggregated mean ratio* of the weighted quasi-arithmetic mean by an interior ratio on the interval, and we demonstrate the relation among the weighted quasi-arithmetic mean, the aggregated mean ratio and the decision maker's preference/attitude based on his utility. In economics, the decision maker's attitudes, for example neutral, risk averse and risk loving, are characterized to Arrow-Pratt index of the utility function([2,11,7,8]). In Section 3, this paper characterizes the weighted quasi-arithmetic means and the mean ratios by not only utility functions but also weighing functions as an index for risks in stochastic environments. Next we investigate the properties of the weighted quasi-arithmetic means and the aggregated mean ratios regarding combinations of utility functions and weighting functions. Representing the weighted quasi-arithmetic means by conditional expectations, we investigate the relation between the index for background risks and the risk premium in economics. We also discuss the first-order stochastic dominance through the weighted quasi-arithmetic means. Finally, in Section 4, we show a lot of examples of the weighted quasi-arithmetic means and the aggregated mean ratios with various typical utility functions, and we demonstrate their relations with the classical quasi-arithmetic means.

2 Weighted Quasi-arithmetic Means and Their Properties

In this section, we introduce weighted quasi-arithmetic means and aggregated mean ratios regarding with utility functions and weighting functions, and we discuss sufficient conditions on utility functions and weighting functions to characterize the decision maker's attitude based on the quasi-arithmetic mean and the aggregated mean ratio. Let D be a fixed interval which is not a singleton and we call it a domain. Let $\mathcal{C}(D)$ be the set of all nonempty bounded closed subintervals of D and let $\mathcal{C}(D)_< := \{[a,b] \in \mathcal{C}(D) | a < b\}$. Let $f : D \mapsto (-\infty, \infty)$ be a continuous strictly increasing function for utility, and let $w : D \mapsto (0, \infty)$ be a continuous function for weighting. For a closed interval $[a,b] \in \mathcal{C}(D)_<$, a mapping $M_w^f : \mathcal{C}(D) \mapsto D$ given by

$$M_w^f([a,b]) := f^{-1}\left(\int_a^b f(x)w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x \right) \tag{1}$$

is called the *weighted quasi-arithmetic mean* with a specified weighting w. Next for a closed interval $[a,b] \in \mathcal{C}(D)_<$ we define an interior ratio $\theta_w^f(a,b)$ from a position of the weighted quasi-arithmetic mean $M_w^f([a,b])$ on the interval $[a,b]$ by

$$\theta_w^f(a,b) := \frac{M_w^f([a,b]) - a}{b-a}. \tag{2}$$

Dujmović [3,4,5] studied a *conjunction/disjunction degree*, which is a similar type of ratio in the power case, for computer science. This paper discusses their characterizations from the viewpoint of economics by conditional expectations. Now we let $g : D \mapsto (-\infty, \infty)$ be another continuous strictly increasing function for utility. Let $M_w^g : \mathcal{C}(D) \mapsto D$ be the weighted quasi-arithmetic mean defined by g instead of f in the way of (1) and we put the aggregated mean ratio θ_w^g for M_w^g. Then we obtain the following results.

Lemma 1 ([13]). *Let f and g be C^2-class utility functions on D. Let $[a, b] \in \mathcal{C}(D)_<$. Then the following (a) – (c) are equivalent.*

(a) $f''/f' \leq g''/g'$ *on* (a, b).
(b) $M_w^f([c, d]) \leq M_w^g([c, d])$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.
(c) $\theta_w^f(c, d) \leq \theta_w^g(c, d)$ *for all* $[c, d]$ *satisfying* $[c, d] \subset [a, b]$ *and* $c < d$.

When we may choose two utility functions f and g as decision maker's utilities, Lemma 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 $\theta_w^f(a, b) \leq \theta_w^g(a, b)$ implies that the aggregated mean ratio $\theta_w^f(a, b)$ is more risk averse than $\theta_w^g(a, b)$. The function $-f''/f'$ is called the *Arrow-Pratt index* and it implies the degree of absolute risk aversion in economics ([2,11]).

3 Weighted Quasi-arithmetic Means and Background Risks

In this paper, we focus on weighting functions w as risk factors of stochastic environments in the weighted quasi-arithmetic mean (1) and we characterize it in relation to the conditional expectation. Let D be a fixed domain and let $f : D \mapsto (-\infty, \infty)$ be a fixed continuous strictly increasing function for utility. The following theorem implies the properties of the weighted quasi-arithmetic mean M_w^f and the ratio θ_w^f concerning weighting w.

Theorem 1. *Let $w : D \mapsto (0, \infty)$ and $v : D \mapsto (0, \infty)$ be C^1-class weighting functions. Let $[a, b] \in \mathcal{C}(D)_<$. Then the following (i) and (ii) hold.*

(i) *If w and v satisfy $w'/w < v'/v$ on (a, b), it holds that $M_w^f([a, b]) < M_v^f([a, b])$ and $\theta_w^f([a, b]) < \theta_v^f([a, b])$.*
(ii) *If w and v satisfy $w'/w \leq v'/v$ on (a, b), it holds that $M_w^f([a, b]) \leq M_v^f([a, b])$ and $\theta_w^f([a, b]) \leq \theta_v^f([a, b])$.*

In Theorem 1, we note that $w'/w \leq v'/v$ on (a, b) is a sufficient condition so that the weighting w yields lower estimation than the weighting v. Further, the following Theorem 2 shows an equivalence regarding the assertion 'if - then' in Theorem 1(ii).

Theorem 2. *Let $w : D \mapsto (0, \infty)$ and $v : D \mapsto (0, \infty)$ be C^1-class weighting functions. Let $[a, b] \in \mathcal{C}(D)_<$. Then the following (a) – (c) are equivalent.*

(a) $w'/w \leq v'/v$ *on* (a,b).
(b) $M_w^f([c,d]) \leq M_v^f([c,d])$ *for all* $[c,d]$ *satisfying* $[c,d] \subset [a,b]$ *and* $c < d$.
(c) $\theta_w^f(c,d) \leq \theta_v^f(c,d)$ *for all* $[c,d]$ *satisfying* $[c,d] \subset [a,b]$ *and* $c < d$.

In the following proposition, (i) implies that the estimation by a utility $h = (f+g)/2$ gives a *middle attitude* by the both utilities f and g and (ii) shows that a weighting function $u = (w+v)/2$ gives a *middle-level risk* of the both risks w and v in stochastic environments.

Proposition 1. *Let* $[a,b] \in \mathcal{C}(D)_<$. *Then the following (i) and (ii) holds.*

(i) *Let* f *and* g *be* C^2*-class utility functions on* D. *Let* $h := (f+g)/2$. *If* f *and* g *satisfy* $f''/f' \leq g''/g'$ *on* (a,b), *then* $M_w^f([a,b]) \leq M_w^h([a,b]) \leq M_w^g([a,b])$ *and* $\theta_w^f(a,b) \leq \theta_w^h(a,b) \leq \theta_w^g(a,b)$.
(ii) *Let* $w : D \mapsto (0,\infty)$ *and* $v : D \mapsto (0,\infty)$ *be* C^1*-class weighting functions. Let* $u := (w+v)/2$. *If* w *and* v *satisfy* $w'/w \leq v'/v$ *on* (a,b), *then* $M_w^f([a,b]) \leq M_u^f([a,b]) \leq M_v^f([a,b])$ *and* $\theta_w^f(a,b) \leq \theta_u^f(a,b) \leq \theta_v^f(a,b)$.

The Arrow-Pratt index $-f''/f'$ implies the degree of absolute risk aversion. On the other hand, the index $-w'/w$, which is introduced in this paper, is related to the *background risks* of stochastic environments in economics ([8]). In the rest of this section, using the representation of conditional expectations, we investigate the relation between the index $-w'/w$ and the background risks. Let (Ω, P) be a probability space, where P is a non-atomic probability measure on Ω.

Definition 1. For random variables X and Y on Ω, it is said that the random variable X is *dominated by* the random variable Y in the sense of *the first-order stochastic dominance* if

$$P(X < x) \geq P(Y < x) \text{ for any real number } x. \tag{3}$$

Then the following result is well-known for the first-order stochastic dominance in economics (Arrow [2], Gollier [7], Eeckhoudt et al. [8]).

Proposition 2. *Let* X *and* Y *be random variables on* Ω. *Then, the random variable* X *is dominated by the random variable* Y *in the sense of the first-order stochastic dominance if and only if it holds that*

$$E(f(X)) \leq E(f(Y)) \tag{4}$$

for any increasing utility function $f : (-\infty,\infty) \mapsto (-\infty,\infty)$ *satisfying a tail condition* $\lim_{x\to\pm\infty} f(x)(P(X<x) - P(Y<x)) = 0$.

The first-order stochastic dominance (3) means that the stochastic environment X is risky than the stochastic environment Y, and it shows in (4) that all decision makers estimate the stochastic environment X lower than the stochastic environment Y. Then the decision makers prefer the stochastic environment Y to the stochastic environment X with their any increasing utility functions f.

Let X be a real random variable on Ω with a C^1-class density function w on $(-\infty, \infty)$. Since the conditional expectation of the utility $f(X)$ is

$$E(f(X) \mid a < X < b) = \frac{E(f(X)1_{\{a<X<b\}})}{P(a < X < b)} = \frac{\int_a^b f(x)w(x)\,\mathrm{d}x}{\int_a^b w(x)\,\mathrm{d}x}, \tag{5}$$

it holds that

$$M_w^f([a,b]) = f^{-1}\left(\frac{\int_a^b f(x)w(x)\,\mathrm{d}x}{\int_a^b w(x)\,\mathrm{d}x}\right) = f^{-1}(E(f(X) \mid a < X < b)) \tag{6}$$

for real numbers $a, b\,(a < b)$, where $1_{\{\cdot\}}$ implies the characteristic function of a set. From Theorem 2 and (6), we obtain the following result together with Proposition 2.

Corollary 1. *Let X and Y be random variables on Ω which have C^1-class density functions w and v on $(-\infty, \infty)$ respectively. If*

$$\frac{w'}{w} \leq \frac{v'}{v} \quad \text{on } (-\infty, \infty), \tag{7}$$

then the random variable X is dominated by the random variable Y in the sense of the first-order stochastic dominance.

From this corollary, (7) is a sufficient condition for the first-order stochastic dominance (3) where the stochastic environment X is risky than the stochastic environment Y. Hence we find that (7) is useful to estimate the risk-level of stochastic environments and it is easy to check in actual problems (Example 3). In this paper, we call $-w'/w$ the *background risk index*. We note that the first-order stochastic dominance (3) is a risk criterion in a global area $D = (-\infty, \infty)$ for stochastic environments and it is represented by integrals in (4), however the background risk index $-w'/w$ can measure risks even in local areas since it is represented by differentials.

Next we discuss risk premiums regarding risk averse in financial management ([7,8]). Let $z \in D$, which implies an *initial wealth*, and let $[a, b] \in \mathcal{C}(D_z)_<$, where $D_z := \{x - z \mid x \in D\}$. Let X be a random variable on Ω, which implies a *stochastic environment with some risk*. A decision maker with a utility f is called *risk averse on* (a, b) if

$$E(f(z + X) \mid a < X < b) \leq f(E(z + X \mid a < X < b)). \tag{8}$$

A sufficient condition for the risk averse is that the utility function f is concave. Let w be a density function on D for the random variable X. Hence, in the following (9), a real number $\pi_w^f(a, b)$ is called *the risk premium on* (a, b) ([7,8]) if it satisfies

$$E(f(z + X) \mid a < X < b) = f(z - \pi_w^f(a, b)). \tag{9}$$

Eq.(9) means that the decision maker accepts the risk arising from the random variable X by paying the risk premium $\pi_w^f(a, b)$.

Theorem 3. *Let f be a continuous strictly increasing utility function on D. Let X be a random variable on Ω which has a C^1-class density function w on D. The risk premium in (9) is given by*

$$\pi_w^f(a,b) = -M_w^h([a,b]), \tag{10}$$

where $h(x) := f(z+x)$ for $x \in (a-z, b-z)$.

Then we obtain the following two theorems. Theorem 4 is from Lemma 1 and Theorem 3, and it gives the relation between the Arrow-Pratt index and the risk premium. On the other hand, Theorem 4 is from Theorems 2 and 3, and it gives the relation between the background risk index and the risk premium.

Theorem 4. *Let an initial wealth $z \in D$ and let $[a,b] \in \mathcal{C}(D_z)_<$. Let f and g be continuous strictly increasing utility functions on D. Let X be random variable on Ω which has a C^1-class density function w. Then the following (a) and (b) are equivalent.*

(a) $f''/f' \le g''/g'$ *on* $(z+a, z+b)$.
(b) $\pi_w^f(c,d) \ge \pi_w^g(c,d)$ *for all* $[c,d]$ *satisfying* $[c,d] \subset [a,b]$ *and* $c < d$.

Theorem 5. *Let f be a continuous strictly increasing utility function on D. Let X and Y be random variables on Ω which have C^1-class density functions w and v respectively. Then the following (a) and (b) are equivalent.*

(a) $w'/w \le v'/v$ *on* (a,b).
(b) $\pi_w^f(c,d) \ge \pi_v^f(c,d)$ *for all* $[c,d]$ *satisfying* $[c,d] \subset [a,b]$ *and* $c < d$.

4 Examples

In this section, we give examples for weighted quasi-arithmetic means $M_w^f([a,b])$ and the aggregated mean ratio $\theta_w^f(a,b)$ which are presented in the previous sections. First we investigate examples of weighting functions w, and next we discuss examples of utility functions f.

Example 1. We deal with a utility function $f(x) = x$ for $x \in (-\infty, \infty)$. Then $f''(x)/f'(x) = 0$. For a closed interval $[a,b] \in \mathcal{C}(D)_<$, we define the *neutral weighted mean* $N_w([a,b])$ and its aggregated mean ratio $\nu_w(a,b)$ by

$$N_w([a,b]) := \int_a^b x\ w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x \tag{11}$$

and

$$\nu_w(a,b) := \frac{N_w([a,b]) - a}{b-a} = \int_a^b (x-a)w(x)\,\mathrm{d}x \Big/ \int_a^b (b-a)w(x)\,\mathrm{d}x. \tag{12}$$

(i) Take a weighting function $w(x) = x^\alpha$ on $D = (0, \infty)$ with a constant α such that $\alpha \neq -2$ and $\alpha \neq -1$. Then $w'(x)/w(x) = \alpha/x$. Let $[a, b] \subset D = (0, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{(\alpha + 1)(b^{\alpha+2} - a^{\alpha+2})}{(\alpha + 2)(b^{\alpha+1} - a^{\alpha+1})}.$$

Further, it holds that $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$ ([13, Theorem 5.9]) and $\lim_{a \downarrow 0} \nu_w(a, b) = \lim_{b \to \infty} \nu_w(a, b) = (\alpha + 1)/(\alpha + 2)$. Weighted quasi-arithmetic means $M_w^f([a, b])$ for other utility functions f are given by Table 1.

Table 1. Weighted quasi-arithmetic means for utility functions f ($w(x) = x^\alpha$)

f	f''/f'	$M_w^f([a, b])$
$rx + s$ $(r > 0)$	0	$\dfrac{(\alpha + 1)(b^{\alpha+2} - a^{\alpha+2})}{(\alpha + 2)(b^{\alpha+1} - a^{\alpha+1})}$
x^r $(r \neq 0)$	$\dfrac{r - 1}{x}$	$\left(\dfrac{(\alpha + 1)(b^{r+\alpha+1} - a^{r+\alpha+1})}{(r + \alpha + 1)(b^{\alpha+1} - a^{\alpha+1})}\right)^{1/r}$
$r \log x$ $(r > 0)$	$-\dfrac{1}{x}$	$\exp\left(\dfrac{b^{\alpha+1} \log b - a^{\alpha+1} \log a}{b^{\alpha+1} - a^{\alpha+1}} - \dfrac{1}{\alpha + 1}\right)$
e^{sx} $(s \neq 0)$	s	$\dfrac{1}{s} \log\left(\dfrac{(\alpha + 1)(\Gamma(\alpha + 1, -sb) - \Gamma(\alpha + 1, -sa))}{s^{\alpha+1}(b^{\alpha+1} - a^{\alpha+1})}\right)$

Here in Table 1 we put

$$\Gamma(\alpha + 1, c) := \int_c^\infty t^\alpha e^{-t} dt$$

for real numbers c.

(ii) Take a weighting function $w(x) = x^{-2}$ on $D = (0, \infty)$ with $\alpha = -2$. Then $w'(x)/w(x) = -2/x$. Let $[a, b] \subset D = (0, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{ab(\log b - \log a)}{b - a}.$$

Further, it holds that $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$ ([13, Theorem 5.9]) and $\lim_{a \downarrow 0} \nu_w(a, b) = \lim_{b \to \infty} \nu_w(a, b) = 0$.

(iii) Take a weighting function $w(x) = x^{-1}$ on $D = (0, \infty)$ with $\alpha = -1$. Then $w'(x)/w(x) = -1/x$. Let $[a, b] \subset D = (0, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{b - a}{\log b - \log a}.$$

Further, it holds that $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$ ([13, Theorem 5.9]) and $\lim_{a \downarrow 0} \nu_w(a, b) = \lim_{b \to \infty} \nu_w(a, b) = 0$.

(iv) Take a weighting function $w(x) = c_0 + c_1 x + c_2 x^2$ on $D = (0, \infty)$ with positive constants $c_0, c_1.c_2$. Then

$$\frac{w'(x)}{w(x)} = \frac{c_1 + 2c_2 x}{c_0 + c_1 x + c_2 x^2}.$$

Let $[a, b] \subset D = (0, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{\frac{1}{2}c_0(b^2 - a^2) + \frac{1}{3}c_1(b^3 - a^3) + \frac{1}{4}c_2(b^4 - a^4)}{c_0(b - a) + \frac{1}{2}c_1(b^2 - a^2) + \frac{1}{3}c_2(b^3 - a^3)}.$$

Further, it holds that $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$,

$$\lim_{a \downarrow 0} \nu_w(a, b) = \frac{6c_0 + 4c_1 b + 3c_2 b^2}{12c_0 + 6c_1 b + 4c_2 b^2} \quad \text{and} \quad \lim_{b \to \infty} \nu_w(a, b) = \frac{3}{4}.$$

(v) Take a weighting function $w(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$ on $D = (0, \infty)$ with positive constants $c_0, c_1.c_2, \cdots, c_n$. Then

$$\frac{w'(x)}{w(x)} = \frac{\sum_{k=0}^{n-1}(k+1)c_{k+1}x^k}{\sum_{k=0}^{n} c_k x^k}.$$

Let $[a, b] \subset D = (0, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{\sum_{k=0}^{n} \frac{1}{k+2} c_k (b^{k+2} - a^{k+2})}{\sum_{k=0}^{n} \frac{1}{k+1} c_k (b^{k+1} - a^{k+1})}.$$

Further, it holds that $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$,

$$\lim_{a \downarrow 0} \nu_w(a, b) = \frac{\sum_{k=0}^{n} \frac{1}{k+2} c_k b^{k+2}}{\sum_{k=0}^{n} \frac{1}{k+1} c_k b^{k+1}} \quad \text{and} \quad \lim_{b \to \infty} \nu_w(a, b) = \frac{n+1}{n+2}.$$

(vi) Take a weighting function $w(x) = e^{-\beta x}$ on $D = (-\infty, \infty)$ with a non-zero constant β. Then $w'(x)/w(x) = -\beta$. Let $[a, b] \subset D = (-\infty, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{e^{-\beta b}(\beta b + 1) - e^{-\beta a}(\beta a + 1)}{\beta(e^{-\beta b} - e^{-\beta a})}.$$

Further, $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$ and $\lim_{a \to -\infty} \nu_w(a, b) = \lim_{b \to \infty} \nu_w(a, b) = 1$.

(vii) Take a weighting function $w(x) = e^{-\gamma^2 x^2}$ on $D = (-\infty, \infty)$ with a positive constant γ. Then $w'(x)/w(x) = -2\gamma^2 x$. Let $[a, b] \subset D = (-\infty, \infty)$ such that $a < b$. Then, we have

$$N_w([a, b]) = \frac{e^{-\gamma^2 a^2} - e^{-\gamma^2 b^2}}{2\gamma \int_{\gamma a}^{\gamma b} e^{-x^2}\, dx}.$$

Further, $\lim_{b \downarrow a} \nu_w(a, b) = \lim_{a \uparrow b} \nu_w(a, b) = 1/2$, $\lim_{a \to -\infty} \nu_w(a, b) = 1$ and $\lim_{b \to \infty} \nu_w(a, b) = 0$.

Table 2. Neutral weighted means for weighting functions w $(f(x) = x)$

w	w'/w	$N_w([a, b])$
x^α $(\alpha \neq -2, -1)$	$\dfrac{\alpha}{x}$	$\dfrac{(\alpha+1)(b^{\alpha+2} - a^{\alpha+2})}{(\alpha+2)(b^{\alpha+1} - a^{\alpha+1})}$
$\sum_{k=0}^n c_k x^k$ $(c_0, c_1, \cdots, c_n > 0)$	$\dfrac{\sum_{k=0}^{n-1}(k+1)c_{k+1}x^k}{\sum_{k=0}^n c_k x^k}$	$\dfrac{\sum_{k=0}^n \frac{1}{k+2} c_k (b^{k+2} - a^{k+2})}{\sum_{k=0}^n \frac{1}{k+1} c_k (b^{k+1} - a^{k+1})}$
$e^{-\beta x}$ $(\beta \neq 0)$	$-\beta$	$\dfrac{e^{-\beta b}(\beta b + 1) - e^{-\beta a}(\beta a + 1)}{\beta(e^{-\beta b} - e^{-\beta a})}$
$e^{-\gamma^2 x^2}$ $(\gamma > 0)$	$-2\gamma^2 x$	$\dfrac{e^{-\gamma^2 a^2} - e^{-\gamma^2 b^2}}{2\gamma \int_{\gamma a}^{\gamma b} e^{-x^2}\, \mathrm{d}x}$

Some results in Example 1 are listed in Table 2. Next we show the relation between the weighted quasi-arithmetic means and the typical means.

Example 2. Let the domain $D = (0, \infty)$. Take a function $f(x) = x^r$ and $w(x) = x^\alpha$ on D with constants r, α satisfying $r \neq 0$. Then $f''(x)/f'(x) = (r-1)/x$ and $w'(x)/w(x) = \alpha/x$. Hence we can deal with not only $r > 0$ for increasing function $f = x^r$ but also $r < 0$ for decreasing function $f(x) = x^r$ ([13, Remark 3.2(1)]). Then, for $[a, b] \subset D$ such that $a < b$, the weighted quasi-arithmetic mean is given by the following $M_{(\alpha)}^{(r)}([a, b]) := M_w^f([a, b])$:

$$M_{(\alpha)}^{(r)}([a, b]) = \left(\frac{(\alpha+1)(b^{r+\alpha+1} - a^{r+\alpha+1})}{(r+\alpha+1)(b^{\alpha+1} - a^{\alpha+1})} \right)^{1/r}$$

if $r \neq 0, \alpha \neq -1, r+\alpha \neq -1$. The limiting values regrading r and α are

$$\lim_{\alpha \to -r-1} M_{(\alpha)}^{(r)}([a,b]) = ab \left(\frac{r(\log b - \log a)}{b^r - a^r} \right)^{1/r} \quad \text{if } r \neq 0,$$

$$\lim_{\alpha \to -1} M_{(\alpha)}^{(r)}([a,b]) = \left(\frac{r(\log b - \log a)}{b^r - a^r} \right)^{-1/r} \quad \text{if } r \neq 0,$$

$$\lim_{r \to 0} M_{(\alpha)}^{(r)}([a,b]) = \exp\left(\frac{b^{\alpha+1} \log b - a^{\alpha+1} \log a}{b^{\alpha+1} - a^{\alpha+1}} - \frac{1}{\alpha+1} \right) \quad \text{if } \alpha \neq -1,$$

$$\lim_{\alpha \to -1} \lim_{r \to 0} M_{(\alpha)}^{(r)}([a,b]) = \sqrt{ab},$$

$$\lim_{r \to -\infty} M_{(\alpha)}^{(r)}([a,b]) = a,$$

$$\lim_{r \to \infty} M_{(\alpha)}^{(r)}([a,b]) = b.$$

Finally we show the relation between the weighted quasi-arithmetic means and their application to economics.

Example 3. We give an example for Corollary 1 by normal distributions on stochastic environments. Let random variables X and Y have normal distributions on Ω with density functions w and v respectively as follows. Let μ_X and μ_X be the means and let σ_X and σ_Y be the standard deviations for w and v respectively, i.e.,

$$w(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\,\sigma_X^2} \right) \text{ and } v(x) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(x-\mu_Y)^2}{2\,\sigma_Y^2} \right)$$

for real numbers x. Then we have

$$\begin{aligned}
&\frac{w'(x)}{w(x)} \leq \frac{v'(x)}{v(x)} \\
&\iff -\frac{x-\mu_X}{\sigma_X^2} \leq -\frac{x-\mu_Y}{\sigma_Y^2} \\
&\iff \sigma_X^2 \mu_Y - \sigma_Y^2 \mu_X \geq (\sigma_X^2 - \sigma_Y^2) x \\
&\iff \begin{cases} x \geq \frac{\sigma_X^2 \mu_Y - \sigma_Y^2 \mu_X}{\sigma_X^2 - \sigma_Y^2} & \text{if } \sigma_X < \sigma_Y \\ x \leq \frac{\sigma_X^2 \mu_Y - \sigma_Y^2 \mu_X}{\sigma_X^2 - \sigma_Y^2} & \text{if } \sigma_X > \sigma_Y \\ \text{all } x \in (-\infty, \infty) & \text{if } \sigma_X = \sigma_Y \text{ and } \mu_X \leq \mu_Y \\ \text{no } x & \text{if } \sigma_X = \sigma_Y \text{ and } \mu_X > \mu_Y. \end{cases}
\end{aligned}$$

Define a domain D by

$$D := \begin{cases} \left(\frac{\sigma_X^2 \mu_Y - \sigma_Y^2 \mu_X}{\sigma_X^2 - \sigma_Y^2}, \infty \right) & \text{if } \sigma_X < \sigma_Y \\ \left(-\infty, \frac{\sigma_X^2 \mu_Y - \sigma_Y^2 \mu_X}{\sigma_X^2 - \sigma_Y^2} \right) & \text{if } \sigma_X > \sigma_Y \\ (-\infty, \infty) & \text{if } \sigma_X = \sigma_Y \text{ and } \mu_X \leq \mu_Y \\ \emptyset & \text{if } \sigma_X = \sigma_Y \text{ and } \mu_X > \mu_Y. \end{cases}$$

From Theorems 2 and 5, we get $M_w^f([a,b]) \leq M_v^f([a,b])$ and $\pi_w^f(a,b) = -M_w^f([a,b]) \geq -M_v^f([a,b]) = \pi_v^f(a,b)$ for subintervals $[a,b] \subset D$. By Theorem 3, we can also calculate the risk premium $\pi_w^f(a,b)$ for a classical utility function $f(x) = 1 - e^{-x}$ as follows:

$$\pi_w^f(a,b) = \frac{\mathrm{Erf}\left(\frac{a-\mu_X}{\sqrt{2}\sigma_X}\right) - \mathrm{Erf}\left(\frac{b-\mu_X}{\sqrt{2}\sigma_X}\right)}{2(-a+b+e^{a+z}-e^{b+z})} + \frac{e^{z+\mu_X+\frac{\sigma_X^2}{2}}\left(\mathrm{Erf}\left(\frac{-a+\mu_X+\sigma_X^2}{\sqrt{2}\sigma_X}\right) - \mathrm{Erf}\left(\frac{-b+\mu_X+\sigma_X^2}{\sqrt{2}\sigma_X}\right)\right)}{2(-a+b+e^{a+z}-e^{b+z})},$$

where

$$\mathrm{Erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

for real numbers x. Further, by Corollary 1, if $\sigma_X = \sigma_Y$ and $\mu_X \leq \mu_Y$, all decision makers prefers the stochastic environment Y to the stochastic environment X for his any increasing utility f, i.e. it holds that $E(f(X)) \leq E(f(Y))$ for any increasing utility function f, which is equivalent that X is dominated by Y in the sense of the first-order stochastic dominance (Proposition 1).

5 Conclusions

We have analyzed the weighted quasi-arithmetic means with utility functions and weighting for random factors in stochastic environments. The background risk index is first introduced through weighting functions as an index of risk-levels for stochastic environments, and its relations to the first-order stochastic dominance and the risk premium are demonstrated with conditional expectations. We have investigated a lot of examples of the weighted quasi-arithmetic mean and the aggregated mean ratio for various typical utility functions. The stochastic dominance is a risk criterion in a global area for stochastic environments however using the background risk index $-w'/w$ we can analyze risks even in local areas. The background risk index $-w'/w$ will be useful and easy to calculate in actual problems.

References

1. Aczél, J.: On weighted mean values. Bulletin of the American Math. Society 54, 392–400 (1948)
2. Arrow, K.J.: Essays in the Theory of Risk-Bearing, Markham, Chicago (1971)
3. 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)
4. Dujmović, J.J., Larsen, H.L.: Generalized Conjunction/disjunction. International Journal of Approximate Reasoning 46, 423–446 (2007)

5. Dujmović, J.J., Nagashima, H.: LSP method and its use for evaluation of Java IDEs. International Journal of Approximate Reasoning 41, 3–22 (2006)
6. Fishburn, P.C.: Utility Theory for Decision Making. John Wiley and Sons, New York (1970)
7. Gollier, G.: The Economics of Risk and Time. MIT Publishers, Cambridge (2001)
8. Eeckhoudt, L., Gollier, G., Schkesinger, H.: Economic and Financial Decisions under Risk. Princeton University Press, Princeton (2005)
9. Kolmogoroff, A.N.: Sur la notion de la moyenne. Acad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. 12, 388–391 (1930)
10. Nagumo, K.: Über eine Klasse der Mittelwerte. Japanese Journal of Mathematics 6, 71–79 (1930)
11. Pratt, J.W.: Risk Aversion in the Small and the Large. Econometrica 32, 122–136 (1964)
12. Yoshida, Y.: Aggregated mean ratios of an interval induced from aggregation operations. In: Torra, V., Narukawa, Y. (eds.) MDAI 2008. LNCS (LNAI), vol. 5285, pp. 26–37. Springer, Heidelberg (2008)
13. Yoshida, Y.: Quasi-arithmetic means and ratios of an interval induced from weighted aggregation operations. Soft Computing 14, 473–485 (2010)

Modelling Group Decision Making Problems in Changeable Conditions

Ignacio J. Pérez[1], Sergio Alonso[2],
Francisco J. Cabrerizo[3], and Enrique Herrera-Viedma[1]

[1] Dept. of Computer Science and Artificial Intelligence, University of Granada, Spain
ijperez@decsai.ugr.es, viedma@decsai.ugr.es
[2] Dept. of Software Engineering, University of Granada, Spain
zerjioi@ugr.es,
[3] Dept. of Software Engineering and Computer Systems, Distance Learning University of Spain (UNED), Madrid, Spain
cabrerizo@issi.uned.es

Abstract. The aim of this paper is to present a new group decision making model with two important characteristics: i) we apply mobile technologies in the decision process and ii) the set of alternatives is not constant through time. We implement a prototype of a mobile decision support system based on changeable sets of alternatives. Using their mobile devices (as mobile phones or PDAs), experts can provide/receive information in anywhere and anytime. The prototype also incorporates a new system to manage the alternatives and thus, to give more realism to decision processes allowing to manage changeable set of alternatives, focussing the discussion in a subset of them that changes in each stage of the process.

Keywords: Group decision making, mobile internet, dynamic environment.

1 Introduction

Group Decision Making (GDM) arises from many real world situations [1, 2]. As a result, the study of decision making is necessary and important not only in Decision Theory but also in areas such as Management Science, Operations Research, Politics, Social Psychology, and so on. In such problems, there are a set of alternatives to solve a problem and a group of experts trying to achieve a common solution. To do this, experts have to express their preferences by means of a set of evaluations over the set of alternatives.

Nowadays, we are realizing many significant advances in the way human interact with technology. The spread of e-services and wireless or mobile devices has increased accessibility to data and, in turn, influenced the way in which users make decisions while they are on the move. Users can make real-time decisions based on the most up-to-date data accessed via wireless devices, such as portable computers, mobile phones, and personal digital assistants (PDAs), which are usually carried all the time and allows to make decisions anytime and anywhere. Thus, the adoption of the latest mobil technologies extends opportunities and allows to carry out consensus processes where previously could not be correctly addressed. Such adoption is based on the assumption that if the communications are improved the decisions will be upgraded, because

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 43–54, 2010.

the discussion could be focussed on the problem with less time wasted on unimportant issues [3, 4].

Usually, resolution methods for GDM problems are static, that is, it is assumed that the number of alternatives and experts acting in the GDM problem remains fixed throughout the decision making process. However, in real decision situations we find dynamic GDM problems in which the number of alternatives and/or experts could vary during the decision making process. Sometimes, where the decision process is slow or it takes a long time, the set of feasible alternatives is dynamic because their availability or feasibility could change through the decision making time. For example, in e-commerce decision frameworks, where the alternatives are the items that could be bought, it is possible that the availability of some of these items changes while experts are discussing and making the decision, even, new good items might become available. In this paper, we assume GDM problems with changeable set of alternatives.

The aim of this paper is to present a prototype of Decision Support System (DSS) to deal automatically with dynamic GDM problems assuming different preference representations and based on mobile technologies. We present a tool to control the possible changes of alternatives that could appear through the decision making process. At every stage of the decision process, the users (i) will be informed with updated data about the current stage of the decision process, (ii) will receive recommendations to help them to change their preferences, and (iii) will be able to send their updated preferences at any moment, thus improving the user participation in the GDM process. . In order to build a flexible framework and give a high degree of freedom to represent the preferences, experts are allowed to provide their preferences in any of the following four ways: (i) as a preference ordering of the alternatives, (ii) as an utility function, (iii) as a fuzzy preference relation, or (iv) as a multiplicative preference relation.

To do so, the paper is set out as follows: Some considerations about GDM problems and mobile technologies are presented in Section 2. Section 3 deals with the prototype which implements such mobile DSS. Finally, in Section 4 we point out our conclusions.

2 Preliminaries

In this section we present the classical GDM model and the advantages of using mobile technology in GDM problems.

2.1 Group Decision Making Models

In a GDM problem we have a finite set of feasible alternatives, $X = \{x_1, x_2, \ldots, x_n\}$, $(n \geq 2)$, to be ranked from best to worst using the information given by a set of experts, $E = \{e_1, e_2, \ldots, e_m\}$, $(m \geq 2)$.

Usual resolution methods for GDM problems include two different processes [5, 6] (see Figure 1):

1. *Consensus process:* Clearly, in any decision process, it is preferable that the experts reach a high degree of consensus on the solution set of alternatives. Thus, this process refers to how to obtain the maximum degree of consensus or agreement between the set of experts on the solution alternatives.

2. *Selection process:* This process consists in how to obtain the solution set of alternatives from the opinions on the alternatives given by the experts.

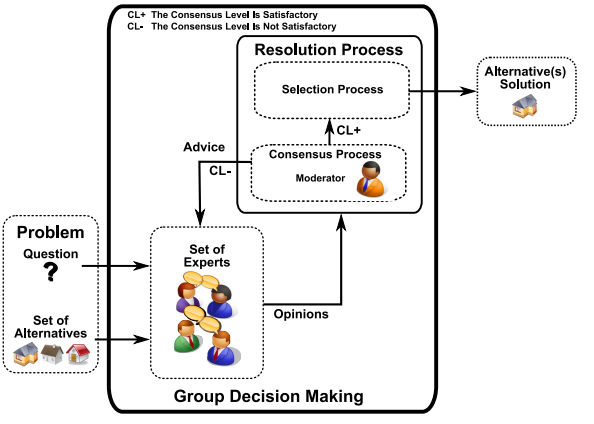


Fig. 1. Resolution process of a GDM

Usually, resolution methods for GDM problems are static, that is, it is assumed that the number of alternatives and experts acting in the GDM problem remains fixed throughout the decision making process. However, in real decision situations we find dynamic GDM problems in which the number of alternatives and/or experts could vary during the decision making process. In this paper, we assume GDM problems with changeable sets of alternatives.

On the other hand, as each expert, $e_k \in E$, has his own ideas, attitudes, motivations and personality, it is quite natural to think that different experts could express their preferences in a different way. This fact has led some authors [7, 8, 9, 10, 11, 12] to assume that experts' preferences over the set of alternatives may be represented in different ways. Amongst these, the most frequently used in decision making theory are:

- *Preference orderings of alternatives:* $O^k = \{o^k(1), ..., o^k(n)\}$, where $o^k(\cdot)$ is a permutation function over the index set, $\{1, ..., n\}$, for the expert, e_k, defining an ordered vector of alternatives, from best to worst.
- *Utility functions:* $U^k = \{u_1^k, ..., u_n^k\}$, $u_i^k \in [0, 1]$, where u_i^k represents the utility evaluation given by the expert e_k to x_i.
- *Fuzzy preference relations:* $P^k \subset X\text{x}X$, with a membership function, $\mu_{P^k} : X\text{x}X \rightarrow [0, 1]$, where $\mu_{P^k}(x_i, x_j) = p_{ij}^k$ denotes the preference degree of x_i over x_j.
- *Multiplicative preference relations:* $A^k \subset X\text{x}X$, where the intensity of preference, a_{ij}^k, is measured using a ratio scale, particularly the 1/9 to 9 scale;

2.2 Mobile Technologies Usage in GDM Problems

During the last decade, organizations have moved from face-to-face group environments to virtual group environments using communication technology. More and more

workers use mobile devices to coordinate and share information with other people. The main objective is that the members of the group could work in an ideal way where they are, having all the necessary information to take the right decisions [3, 4, 13, 14].

To support the new generation of decision makers and to add real-time process in the GDM problem field, many authors have proposed to develop decision support systems based on mobile technologies [15, 16]. Similarly, we propose to incorporate mobile technologies in a DSS obtaining a Mobile DSS (MDSS). Using such a technology should enable a user to maximize the advantages and minimize the drawbacks of DSSs.

The need of a face-to-face meeting disappears with the use of this model, being the own computer system who acts as moderator. Experts can communicate with the system directly using their mobile device from any place in the world and at any time. Hereby, a continuous information flow among the system and each member of the group is produced, which can help to reach the consensus between the experts on a faster way and to obtain better decisions.

In addition, MDSS can help to reduce the time constraint in the decision process. Thus, the time saved by using the MDSS can be used to do an exhaustive analysis of the problem and obtain a better problem definition. This time also could be used to identify more feasible alternative solutions to the problem, and thus, the evaluation of a large set of alternatives would increase the possibility of finding a better solution. The MDSS helps to the resolution of GDM problems providing a propitious environment for the communication, increasing the satisfaction of the user and, in this way, improving the final decisions.

3 A Mobile DSS Based on Changeable Sets of Alternatives

In this section we describe the mobile DSS that incorporates a tool to manage GDM problems in which the set of alternatives could change throughout the decision process. It allows to develop GDM processes at anytime and anywhere, and simulate with more accuracy level the real processes of human decision making which are developed in dynamic environments as the Web, financial investment, health, etc. In what follows we explain the "client/server" architecture of the mobile DSS and the communication and work flow that summarizes the functions of the DSS.

3.1 Server Side

We assume that before to start the GDM process, the moderator selects the feasible set of experts and alternatives and establishes the remaining parameters of the problem. Thus, the structure of the proposed Mobile DSS server is composed of the following five processes: (i) uniformization process, (ii) selection process, (iii) consensus process, (iv) dynamic choice process of alternatives, and (v) feedback process (Figure 2).

Uniformization Process: To give a higher degree of freedom to the system, we assume that experts can present their preferences using any of the preference representations presented in section 2.1. Therefore, it is necessary to make the information uniform before applying the consensus and selection processes. As in [8] we propose to

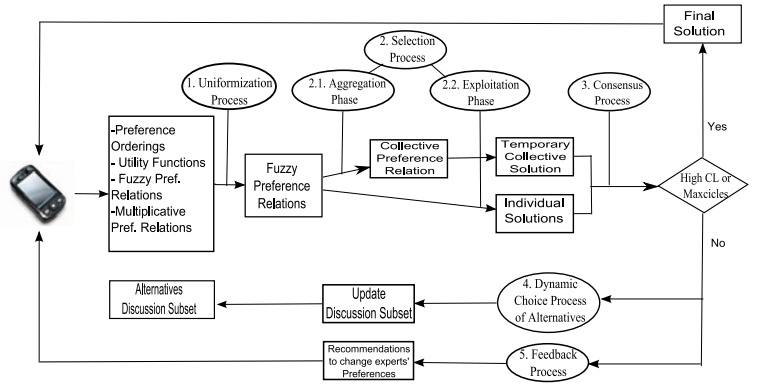


Fig. 2. Structure of the DSS server with multiple preference representation structures

use fuzzy preference relations as the base element to uniform experts' preferences and the following transformation functions are used [8]: $f^1\left(o_i^k, o_j^k\right) = \frac{1}{2}\left(1 + \frac{o_j^k - o_i^k}{n-1}\right)$, $f^2\left(u_i^k, u_j^k\right) = \frac{\left(u_i^k\right)^2}{\left(u_i^k\right)^2 + \left(u_j^k\right)^2}$, $f^3\left(a_{ij}^k\right) = \frac{1}{2}\left(1 + \log_9 a_{ij}^k\right)$.

Selection Process: Once the information is made uniform, we have a set of m individual fuzzy preference relations and then we apply a selection process which has two phases [2, 17]: (i) *aggregation* and (ii) *exploitation*.

- *Aggregation phase:*
  This phase defines a collective preference relation, $P^c = \left(p_{ij}^c\right)$, obtained by means of the aggregation of all individual fuzzy preference relations $\left\{P^1, P^2, \ldots, P^m\right\}$. It indicates the global preference between every pair of alternatives according to the majority of experts' opinions. For example, the aggregation could be carried out by means of an OWA operator [18, 19].
- *Exploitation phase:*
  This phase transforms the global information about the alternatives into a global ranking of them, from which the set of solution alternatives is obtained. The global ranking is obtained applying two choice degrees of alternatives to the collective fuzzy preference relation [20]: the *quantifier guided dominance degree* (QGDD) and the *quantifier guided non dominance degree* (QGNDD).
  Finally, the solution X_{sol} is obtained by applying these two choice degrees, and thus, selecting the alternatives with maximum choice degrees.

Consensus Process: In our mobile DSS, we use a consensus model for GDM problems with different preference representations as it was done in [21]. This model presents the following main characteristics:

- It is based on two soft consensus criteria: global consensus measure on the set of alternatives X, symbolized as C_X, and the proximity measures of each expert e_i on X, called P_X^i.

– Both consensus criteria are defined by comparing the individual solutions with the collective solution using as comparison criterion the positions of the alternatives in each solution.

Initially, in this consensus model we consider that in any nontrivial GDM problem the experts disagree in their opinions so that consensus has to be viewed as an iterated process. This means that agreement is obtained only after some rounds of consultation. In each round, the DSS calculates both the consensus and the proximity measures. The consensus measures evaluate the agreement existing among experts and the proximity measures are used in the feedback mechanism to support the group discussion phase of the consensus process.

Dynamic Choice Process of Alternatives: In real world we find many dynamic decision frameworks: health, financial investment, military operations, Web. In such cases, due to different factors the set of solution alternatives could vary throughout the decision process.

Classical GDM models are defined within static frameworks. In order to make the decision making process more realistic, we provide a new tool to deal with dynamic alternatives in decision making. In such a way, we can solve dynamic decision problems in which, at every stage of the process, the discussion could be centered on different alternatives.

To do so, we define a method which allows us to remove and insert new alternatives into the discussion process. Firstly, the system identifies those worst alternatives that might be removed and the new alternatives to include in the set. This new alternatives can be obtained from a set of new alternatives appeared at a time or from the supply set of alternatives that includes all the alternatives that we had at the beginning of the process but that were not included in the discussion subset because the limitation of this due to specific parameters of the problem. It is worth noting that we assume that alternatives are independent and the inclusion or elimination of one alternative can not change the ranking of other pairs of alternatives.

Thus, the method has two different phases: (1) Remove old bad alternatives and and (2) Insert new good alternatives.

1. The first phase manages situations in which some alternatives of the discussion subset are not available at the moment due to some dynamic external factors or because the experts have evaluated them poorly and they have a low dominance degree ($QGDD$). Therefore, the system checks the availability and the $QGDD$ of each alternative in the current discussion subset. If some alternative is not available or has a $QGDD$ lower than a threshold ($minQGDD$), the system looks for a new good alternative in the new alternatives subset. If this subset is empty, the system uses the supply subset of alternatives provided by the expert at the beginning of the decision process and that were not taken into account then because of the impossibility to compare all the alternatives at the same time. Then, the system asks for the experts' opinions about the replacement and acts according to them (see Figure 3).
2. The second case manages the opposite situation, that is, when some new alternatives have emerged. Basically, the system checks if some new good alternatives

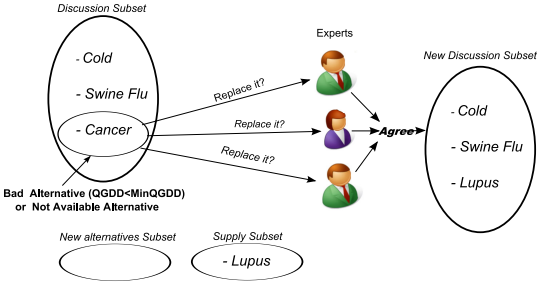


Fig. 3. Dynamic choice process of alternatives: Case 1

have appeared in the new alternatives subset due to some dynamic external factors. If this is the case, the system has to identify the worst alternatives of the current discussion subset. To do this, the system uses the dominance degree $QGDD$ of all alternatives again to choose the worst alternatives. Then, the system asks for the experts' opinions about the replacement and acts according to them (see Figure 4).

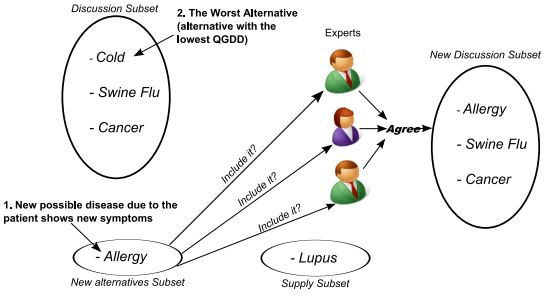


Fig. 4. Dynamic choice process of alternatives: Case 2

Feedback Process: To guide the change of the experts' opinions, the DSS simulates a group discussion session in which a feedback mechanism is applied to quickly obtain a high level of consensus. This mechanism is able to substitute the moderator's actions in the consensus reaching process. The main problem is how to find a way of making individual positions converge and, therefore, how to support the experts in obtaining and agreeing with a particular solution.

When the consensus measure C_X has not reached the required consensus level (CL) and the number of rounds has not reached a maximum number of iterations (MAXCYCLE), defined prior to the beginning of the decision process, the experts' opinions must be modified. As aforementioned, we are using the proximity measures to build a feedback mechanism so that experts can change their opinions and narrow their positions.

3.2 Client Side

For the implementation of the DSS we have chosen a thin client model. This model depends primarily on the central server for the processing activities. This prototype is designed to operate on mobile devices with Internet connection.

The client software has to show to the experts the next eight interfaces:

- Connection: The device must be connected to the network to send/receive information to the server.
- Authentication: The device will ask for a user and password data to access the system.
- Problem description: When a decision process is started, the device shows to the experts a brief description of the problem and the discussion subset of alternatives (see Figure 5 a).
- Selection of preference representations (see Figure 5 b).

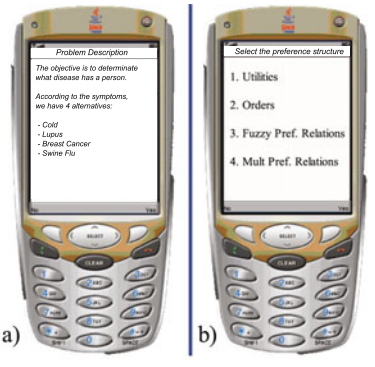


Fig. 5. Problem description and selection of preference representations

- Insertion of preferences: The device will have four different interfaces, one for each different format of preference representation (see Figure 6).
- Change of alternatives: When a bad or not available alternative deserves to be removed from the discussion subset, or a new alternative deserves be inserted in the discussion subset, using the new management process of alternatives, the experts can assess if they want to update the discussion subset by changing these alternatives (see Figure 7).
- Feedback: When opinions should be modified, the device shows to the experts the recommendations and lets them send their new preferences (see Figure 8 a).
- Output: At the end of the decision process, the device will show to the experts the set of solution alternatives as an ordered set of alternatives marking the most relevant ones (see Figure 8 b). The system shows an additional scoring (QGDD) of each alternative when the problem needs more than one of them to be solved. Moreover, if the minimum consensus level is not reached and temporary solution becomes final solution because the maximum number of feedback cycles has been reached, the system notes this situation and it shows the current consensus level on the screen.

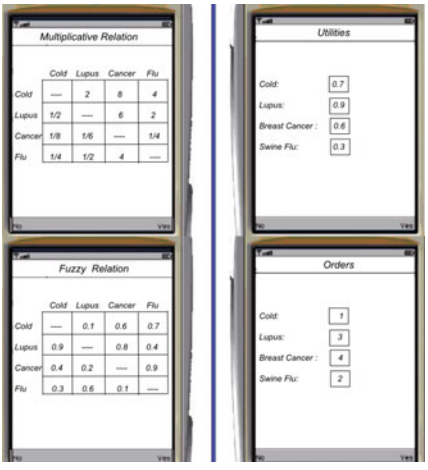


Fig. 6. Insertion of preferences

Fig. 7. Change of alternatives question

On the technical side of the development of the client part of the DSS, it is worth noting that the client application complies with the MIDP 2.0 specifications [22], and that the J2ME Wireless Toolkit 2.2 [23] provided by SUN was used in the development phase. This wireless toolkit is a set of tools that provide J2ME developers with some emulation environments, documentation, and examples to develop MIDP-compliant applications. The application was later tested with a toy example using a JAVA-enabled mobile phone on a GSM network using a GPRS-enabled SIM card. If the discussion subset of alternatives is large and can not be displayed on the screen, Java interface provides scrolling tools that allows displaying bigger interfaces on small screens. The MIDP application is packaged inside a JAVA archive (JAR) file, which contains the applications classes and resource files. This JAR file is the one that actually is downloaded to the physical device (mobile phone) along with the JAVA application descriptor file when an expert wants to use the MDSS.

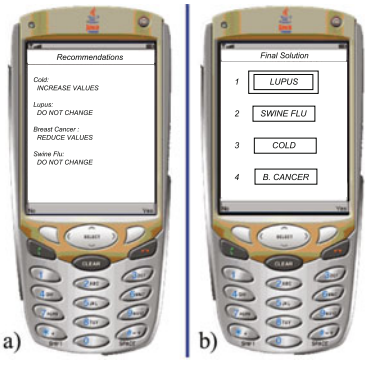


Fig. 8. Recommendations and Final Solution

3.3 Communication and Work Flow

The DSS has to carry out the following functions, also represented in figure 9. In the diagram we can see all the functions of the system, the form in which they are connected together with the database, and the order in which each of them is executed.

0. Database initialization
1. Verify the user messages and store the main information
2. Make the experts' preferences uniform
3. Computation of the set of solution alternatives
4. Computation of the consensus measures
5. Control the consensus state
6. Control the change of alternatives

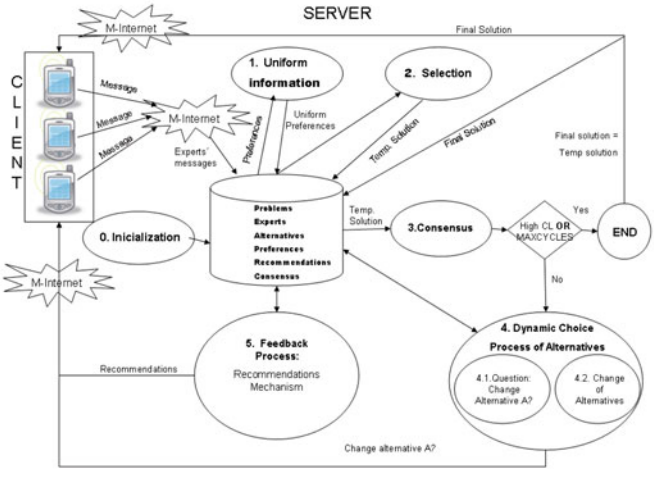


Fig. 9. Functions Scheme of the System

7. Generate the recommendations
8. Go to step 1

4 Conclusions

We have presented a prototype of mobile DSS for GDM problems based on dynamic decision environments which uses the advantages of mobile Internet technologies to improve the user satisfaction with the decision process and develop decision processes at anytime and anywhere. The system allows to model dynamic decision environments because it incorporates a new tool to manage the changes of alternatives in the set of solution alternatives through decision process. We have used mobile phones as the device used by the experts to send their preferences but the structure of the prototype is designed to use any other mobile device as PDAs.

Acknowledgements

This paper has been developed with the financing of FEDER funds in FUZZYLING project (TIN2007-61079), PETRI project (PET2007-0460), project of Ministry of Public Works (90/07) and Excellence Andalusian Project (TIC5299).

References

[1] Kacprzyk, J., Fedrizzi, M.: Multiperson decision making models using fuzzy sets and possibility theory. Kluwer Academic Publishers, Dordrecht (1990)
[2] Roubens, M.: Fuzzy sets and decision analysis. Fuzzy Sets and Systems 90(2), 199–206 (1997)
[3] Katz, J.: Handbook of Mobile Communication Studies. MIT Press, Cambridge (2008)
[4] Schiller, J.: Mobile Communications, 2nd edn. Addison-Wesley, Reading (2003)
[5] Cabrerizo, F., Alonso, S., Herrera-Viedma, E.: A consensus model for group decision making problems with unbalanced fuzzy linguistic information. International Journal of Information Technology & Decision Making 8(1), 109–131 (2009)
[6] Herrera, F., Herrera-Viedma, E., Verdegay, J.: A model of consensus in group decision making under linguistic assessments. Fuzzy Sets and Systems 78(1), 73–87 (1996)
[7] Buyukozkan, G., Feyzioglu, O., Ruan, D.: Fuzzy group decision-making to multiple preference formats in quality function deployment. Computers In Industry 58(5), 392–402 (2007)
[8] Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. Fuzzy Sets and Systems 97(1), 33–48 (1998)
[9] Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating multiplicative preference relations in a multiplicative decision making model based on fuzzy preference relations. Fuzzy Sets and Systems 122(2), 277–291 (2001)
[10] Ma, J., Fan, Z., Jiang, Y.: An optimization approach to multiperson decision making based on different formats of preference information. IEEE Transactions on Systems, Man and Cybernetics part A-Systems and Humans 36(5), 876–889 (2006)
[11] Xu, Z.S.: Multiple-attribute group decision making with different formats of preference information on attributes. IEEE Transactions on Systems, Man and Cybernetics part B-Cybernetics 37(6), 1500–1511 (2007)

[12] Xu, Z.S., Chen, J.: Magdm linear-programming models with distinct uncertain preference structures. IEEE Transactions on Systems, Man and Cybernetics part B-Cybernetics 38(5), 1356–1370 (2008)
[13] Imielinski, T., Badrinath, B.: Mobile wireless computing: challenges in data management. Communications of the ACM 37(10), 18–28 (1994)
[14] Wen, W., Chen, Y., Pao, H.: A mobile knowledge management decision support system for automatically conducting an electronic business. Knowledge-Based Systems 21(7) (2008)
[15] Eren, A., Subasi, A., Coskun, O.: A decision support system for telemedicine through the mobile telecommunications platform. Journal of Medical Systems 32(1) (2008)
[16] Ricci, F., Nguyen, Q.: Acquiring and revising preferences in a critique-based mobile recommender system. IEEE Intelligent Systems 22(3) (2007)
[17] Triantaphyllou, E.: Multi-criteria decision making methods: a comparative study. Kluwer Academic Publishers, Dordrecht (2000)
[18] Yager, R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics 18(1), 183–190 (1988)
[19] Yager, R.: Weighted maximum entropy owa aggregation with applications to decision making under risk. IEEE Transactions on Systems, Man and Cybernetics part A-Systems and Humans 39(3), 555–564 (2009)
[20] Herrera, F., Herrera-Viedma, E., Verdegay, J.: A sequential selection process in group decision making with a linguistic assessment approach. Information Sciences 85(4), 223–239 (1995)
[21] 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)
[22] Midp 2.0 specifications
[23] Java wireless toolkit

Individual Opinions-Based Judgment Aggregation Procedures

Farah Benamara[1], Souhila Kaci[2], and Gabriella Pigozzi[3]

[1] IRIT-CNRS, Toulouse, France
benamara@irit.fr
[2] Université Lille-Nord de France, Artois, France
kaci@cril.univ-artois.fr
[3] University of Luxembourg, Luxembourg
gabriella.pigozzi@uni.lu

Abstract. Judgment aggregation is a recent formal discipline that studies how to aggregate individual judgments on logically connected propositions to form collective decisions on the same propositions. Despite the apparent simplicity of the problem, the aggregation of individual judgments can result in an inconsistent outcome. This seriously troubles this research field. Expert panels, legal courts, boards, and councils are only some examples of group decision situations that confront themselves with such aggregation problems. So far, the existing framework and procedures considered in the literature are idealized. Our goal is to enrich standard judgment aggregation by allowing the individuals to agree or disagree on the decision rule. Moreover, the group members have the possibility to abstain or express neutral judgments. This provides a more realistic framework and, at the same time, consents the definition of an aggregation procedure that escapes the inconsistent group outcome.

1 Introduction

Judgment aggregation is a recent formal discipline that studies how to aggregate individual judgments to form collective decisions. Examples are expert panels, legal courts, boards, and councils [7]. This field has recently attracted attention in multi-agent systems and artificial intelligence, in particular due to the relations with belief merging [13], for example for the combination of opinions of equally reliable individuals.

Judgment aggregation problems consider a group of people stating their views (in the binary form of 1 or 0) on some logically interconnected propositions. As an example, consider the board of a research funding agency whose members have to decide whether to support a research project (*conclusion* D) on the basis of three criteria : originality (P), quality (Q), and applicability (R), that is the decision rule can be expressed as $(P \wedge Q \wedge R) \leftrightarrow D$. As we will see, problems arise because seemingly reasonable aggregation procedures lead to paradoxical outcomes.

Clearly, the problems investigated in this new field are relevant and common to many situations. Nevertheless, the procedures considered so far in the literature are idealized. To provide a more realistic framework and to provide an

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 55–66, 2010.

aggregation procedure that does not run into inconsistent social outcomes are the goals of the paper. More specifically, we propose to extend standard judgment aggregation to take into account two main considerations.

First, we introduce the notion of *judgment status*. It is not realistic to expect that group members state their judgments on each proposition, or that they always have a clear position on each proposition. Our model allows members to express classical binary judgment, a neutral judgment or to abstain on some or all propositions that the individuals consider as irrelevant. In our example of the project funding, suppose that the applicability criterion (R) has been introduced only recently following some new regulation that impose all research funding agency to be evaluated on the basis of likeness to attract the interest of private funding. Suppose also that some members dissent with the criterion R because they believe that this will damage pure theoretical projects to the benefit of pure applied ones. Some members can believe that R is completely irrelevant and thus abstain to give a judgment on R. On the other hand, some other members can believe that the criterion R is relevant but they prefer to be neutral because they are not able to assess its applicability, or because they are indifferent to its value. It is worth noticing that abstention and neutral judgments denote distinct positions. The difference will be clarified later in the paper.

Second, we introduce the notion of *acceptance of the decision rule*. Our framework allows the group members to state whether or not they agree on the rule governing the decision. In our example, some members may disagree on the way the propositions are logically connected whereas some other members can accept the decision rule even if they believe that some propositions are irrelevant.

We present a flexible judgment aggregation approach, in which individuals can express 0/1 judgments as well as being neutral or abstain on some propositions, and can participate in the group decision procedure while disagreeing on the imposed decision rule. What kind of aggregation procedure does such a new framework advocate for? We suggest that the group decision procedure should be responsive of the group's opinion about the decision rule. If the majority (or a pre-fixed quota) of the individuals accepts the decision rule, the aggregation procedure will be based on the criteria of the rule. If the group rejects the rule, the only issue the group can confidently express its opinion about is the conclusion D.

The remainder of this paper is organized as follows. After necessary background on the problem of judgment aggregation, we present our general framework. We then introduce the formal representation, the aggregation procedure and show that our framework offers a solution to the judgment aggregation dilemma. Lastly, we compare our approach to some related work and then we conclude.

2 Judgment Aggregation

In the original problem of judgment aggregation [5,6], a court has to make a decision on whether a person is liable of breaching a contract (proposition D). The judges have to reach a verdict following the legal doctrine. This states that

a person is liable if and only if there was a contract (P) and there was a conduct constituting breach of such a contract (Q). The legal doctrine can be formally expressed by the rule $(P \wedge Q) \leftrightarrow D$. Each member of the court expresses her judgment on the propositions P, Q and D such that the rule $(P \wedge Q) \leftrightarrow D$ is satisfied.

Suppose now that the three members of the court make their judgments according to Table 1.

Table 1. Doctrinal paradox. Premises: P = There was a contract, Q = There was conduct constituting breach of such a contract. Conclusion: $D = (P \wedge Q)$ = There was a breach of contract.

	P	Q	$D = (P \wedge Q)$
Judge A	1	0	0
Judge B	0	1	0
Judge C	1	1	1
Majority	1	1	0

Each judge expresses a consistent judgment, i.e., she says yes to D if and only if she says yes to both P and Q. However, *proposition-wise* majority voting (consisting in the separate aggregation of the votes for each proposition P, Q and D via majority rule) results in a majority for P and Q and yet a majority for $\neg D$. This is an inconsistent collective result, in the sense that $\{P, Q, \neg D, (P \wedge Q) \leftrightarrow D\}$ is inconsistent in propositional logic. The paradox lies in the fact that majority voting can lead a group of rational agents to endorse an irrational collective judgment, i.e., to have a majority believing that the defendant should be left free while *another* majority deems there are reasons to sentence her. The literature on judgment aggregation refers to such problem as the *doctrinal paradox*. Clearly, the relevance of such aggregation problems goes beyond the specific court example and affects all collective decisions on logically interconnected propositions.

The first two ways to avoid the inconsistency that have been suggested are the *premise-based procedure* (PBP) and the *conclusion-based procedure* (CBP) [12]. According to the PBP, each member casts her judgment on each premise. The conclusion is then inferred from the judgment of the majority of the group on the premises using the rule $(P \wedge Q) \leftrightarrow D$. In the example above, the PBP would declare the defendant liable of breaching the contract.

According to the CBP, the members decide privately on P and Q and only express their opinions on D publicly. The judgment of the group is then inferred from applying the majority rule to the agent judgments on the conclusion. The defendant will be declared liable if and only if a majority of the judges actually believes that she is liable, and no reasons for the court decision could be supplied. In the example, contrary to the PBP, the application of the CBP would free the defendant.

3 General Framework

Let us first introduce some terminology from standard judgment aggregation. A set of agents $N = \{1, 2, \ldots, k\}$, with $k \geq 3$, has to make judgments on logically interconnected propositions of a language $\mathcal{L}$. The set of propositions on which the judgments have to be made is called *agenda* (denoted by $\mathcal{A}$), and this is divided between premises and conclusion. A (individual or collective) *judgment set* is the set of propositions believed by the agents or the group. An k-tuple $(J_1, J_2, \ldots, J_k)$ of agents judgment sets is called *profile*. A *judgment aggregation rule* F assigns a collective judgment set J to each profile $(J_1, J_2, \ldots, J_k)$ of agents judgment sets. A judgment set J is *consistent* if it is a consistent set in $\mathcal{L}$, and is *complete* if, for any $P \in \mathcal{L}$, $P \in J$ or $\neg P \in J$ but not both.

In our framework a judgment aggregation has the form:

$$C \leftrightarrow D, \tag{1}$$

where C is a general propositional formula built on literals representing criteria P_i and D is a literal representing the final decision (or conclusion). In the following (1) is referred to as the *decision rule*. We assume that C is neither a tautology nor a contradiction, this ensures that the criteria are logically independent (as in the standard framework). Moreover, if one of the criteria is the negation of a propositional variable (e.g. $\neg Q$), the members express their judgments on criteria as given, i.e., on P, $\neg Q$, etc. Considering decision rules where C is a conjunction is not a limitation of our approach. Notably, our analysis extends to other truth-functional combinations of literals as well, e.g. the disjunctive decision rule $(P \vee Q \vee R) \leftrightarrow D$, because this rule is equivalent to $(\neg P \wedge \neg Q \wedge \neg R) \leftrightarrow \neg D$.

Let us now formalize the extensions we intend to give to standard judgments aggregation, namely: *judgment status* and *acceptance of the decision rule*.

3.1 Judgment Status

We distinguish four possible judgments: classical binary judgment 1 (for) or 0 (against), neutral judgment and abstention. As classical binary judgment is already at work, we only detail abstention and neutral judgments.

Neutral judgments. Neutrality captures those situations in which members do not have a clear position on a specific issue, do not feel competent, or simply prefer not to take position on that matter. We represent a neutral judgment by a question mark "?". For example, given $(P \wedge Q) \leftrightarrow D$, if a member believes P to be true but does not know about Q, then her judgment set will be $\{(1, ?, 0)\}$. A group member may express a neutral judgment w.r.t. some or all criteria, and on the conclusion as well.

Abstention. We represent an abstention by a cross "X". The difference between abstention and a neutral judgment is that a member abstains on a criterion

P_i when she deems that criterion *irrelevant* for the decision D. Consequently, abstention on criteria are ignored in the decision process. However, a member cannot abstain on the decision D. When a member takes part into a collective decision process, she is expected to express her judgment on D.

3.2 Acceptance/Rejection of the Decision Rule

Each member j says whether she accepts ($Accept_j = 1$) or rejects ($Accept_j = 0$) the decision rule.

Acceptance of the decision rule. This means that either, for a member j, the criteria $P_1, \cdots, P_n$ are the all and only relevant ones to make a judgment on D, or that (1) contains some irrelevant criteria together with all relevant ones. In the first case, j will give 1/0 or neutral judgments on each criterion. In the second case, she will abstain on the criteria that she deems to be irrelevant and will give a judgment only on the relevant criteria. Of course, D is computed only on the basis of the criteria on which she expressed a judgment, i.e., she did not abstain. It is worth observing that the original legal paradox of judgment aggregation is an instance of this case, where all group members (the judges) have to endorse the legal code, or behave as if this is the case.

Rejection of the decision rule. There are two cases under which a member can reject the decision rule. The first is when she believes that criteria P_i are not adequate, i.e., some criteria are missing. In this case, j fixes the value of D according also to the missing criteria. The intuition is that, if a member wants to have her saying in a decision process, but considers the adopted rule unable to capture the relevant criteria for the decision, she must be able to express her judgment on the conclusion while making explicit that she deems the rule to be not appropriate.

The second situation is when j agrees on the criteria P_i in the decision rule, but disagrees on the way these criteria are logically connected. For e.g. the rule is $(P \wedge Q) \leftrightarrow D$ and according to j the decision rule should instead be $(P \vee Q) \leftrightarrow D$. Member j will therefore assign 0/1 or neutral judgments on the criteria while deciding on the value of D according to $(P \vee Q) \leftrightarrow D$.

It is important to notice that members may express judgments on the premises even when they reject the rule. The reason is that in case they are the only one to reject the rule (or in any case there is not majority rejecting the rule), the decision procedure will be PBP (see Section 4) in which case they can have their saying on the premises.

Example 1. Consider our running example of the board of a research funding agency whose members have to decide which research project to support on the basis of three criteria: originality (P), quality (Q), and applicability (R). Suppose that the decision rule is $(P \wedge Q \wedge R) \leftrightarrow D$. The five members state their judgments on P, Q and R as in Table 2.

Table 2. Acceptance of the general rule and individual judgments

	Acceptance	P	Q	R	D
M_1	1	0	0	1	0
M_2	1	1	X	1	1
M_3	1	?	0	0	0
M_4	0	1	X	1	?
M_5	0	0	1	1	1

The first three members agree with the decision rule $(P \wedge Q \wedge R) \leftrightarrow D$ since for them $Accept_j = 1$ (for $j = 1, 2, 3$). M_1 thinks that the criteria P, Q, and R are the all and only relevant attributes for funding a project whereas M_2 thinks that the criterion Q is irrelevant and decides then to abstain to give any judgement on Q. The decision of M_2 is derived on the basis of P and R only. The third member also agrees on the decision rule, but unlike the first two, she is neutral on P. Since for M_3, Q and R false and she accepts the rule, D is also false. Finally, M_4 and M_5 reject the decision rule. M_4 thinks that Q is irrelevant and that there are missing criteria. So she abstains to give any judgement on Q and gives neutral judgement on D according to P, R and the missing criteria. M_5 does not accept the rule for other reasons: she thinks that the criteria are relevant but not correctly linked. Indeed she expresses her opinions on all the propositions but she gives her judgement to D following the rule $(P \vee Q \vee R) \leftrightarrow D$.

4 Representation and Aggregation Procedure

We represent a judgment expressed by a member j by the following tuple

$$J_j = (Accept_j, P_{1j}, \cdots, P_{nj}, D_j),$$

where $P_{ij} \in \{0, 1, ?, X\}$ and $D_j \in \{0, 1, ?\}$.

D_j is either derived following the decision rule or fixed by the group member depending on whether she accepts the general rule or not.

Given a set of judgments $\{J_1, \cdots, J_k\}$, the collective decision is represented as follows:

$$D = (Accept_{agg}, P_{agg_1}, \cdots, P_{agg_n}, D_{agg}),$$

such that:

- $Accept_{agg}$ is the majority (or any other quota rule) of $Accept_1, \cdots, Accept_k$. If there are as many members accepting the rule as members rejecting it, we fix $Accept_{agg} = 0$. In social choice theory, tie-breaking rules are usually random. Since in our approach, all members assign a value to the conclusion, it is preferable to break the tie in favour of $Accept_{agg} = 0$ (so turning to CBP) than by a random choice (this will be detailed later in this section).
- P_{agg_i} is the majority of $P_{i1}, \cdots, P_{ik}$ following proposition-wise majority voting. Abstentions on P_i are ignored when computing P_{agg_i} since those criteria

are irrelevant. Neutral judgments simply follow the majority. In case of indecision, i.e., a tie between the number of $P_{ij} = 1$ and $P_{ij} = 0$, we put $P_{agg_i} =?$.

– D_{agg} is computed by PBP or CBP. The procedure is fixed according to $Accept_{agg}$ as follows :
  - if $Accept_{agg} = 0$ then we use CBP and D_{agg} is computed on the basis of $D_1, \cdots, D_k$. This is intuitively meaningful since $Accept_{agg} = 0$ means that the group members thought that the decision rule was not the right one for that decision, so the only reasonable thing they can say is the final conclusion. D_{agg} is calculated by simple majority voting. Neutral judgments simply follow the majority. In case of indecision, we have the following subcases:
    * if indecision is not allowed, then we do not accept neutrality on D. We then propose to have either a 'pessimistic' ($D_{agg} = 0$) or an 'optimistic' ($D_{agg} = 1$) solution on the conclusion. Such a choice is publicly stated at the begining of the decision process and is fixed by the same person or organization that fixed the decision rule. It is reasonable to expect that the way the conclusion is decided in case of indecision depends on the context: In a legal context, for example, it is preferred to release a culprit rather than condemn an innocent. On the other hand, if we must hire a person, it is reasonable to opt for the optimistic solution, i.e., the indecision is interpreted as a positive decision.
    * if indecision is allowed at the begining of the decision process then we put $D_{agg} =?$
  - if $Accept_{agg} = 1$ then PBP is used and D_{agg} is derived by the collective judgments on the premises following the decision rule. If the aggregation of $P_{agg_1}, \cdots, P_{agg_n}$ results in ? following the decision rule and indecision is not allowed then we will have $D_{agg} = 1$ in case of an optimistic reasoning or $D_{agg} = 0$ in case of an pessimistic reasoning.
  - In case there are as many members who accept the rule as individuals who reject the rule, CBP is used. The reason is that those who reject the rule derive the value of D_{agg} using also the missing criteria or what they think are the correct logical relations among criteria. In both cases, using PBP and deriving D_{agg} by the given decision rule would not reflect their opinions. If CBP returns a tie between $D = 0$ and $D = 1$, this is handled in the same way as in the case where $Accept_{agg} = 0$.

We now illustrate the procedure with our running example.

Example 2. Table 3 gives the judgments expressed by five members of our funding board. Only M_1 rejects the rule because she believes that some criteria are missing. She fixes the value of D according also to the missing criteria, this is why $D = 0$ despite the fact that we have $P = 1$ and $Q = 1$. Since $Accept_{agg} = 1$ we use premise-based procedure. We get $D_{agg} = 1$ following the decision rule.

Table 3. Example of judgment aggregation with acceptance of the general rule

	Acceptance	P	Q	R	D
M_1	0	1	1	?	0
M_2	1	0	1	1	0
M_3	1	X	1	1	1
M_4	1	1	1	1	1
M_5	1	1	0	1	0
collective decision	1	1	1	1	1

Example 3. Let us now consider individuals who have to make a collective decision using the rule $(P \wedge Q \wedge R) \leftrightarrow D$, with a majority thinking that it is not appropriate, i.e., $Accept_{agg} = 0$. Suppose that their judgments are as in Table 4.

Table 4. Example of judgment aggregation with rejection of the general rule

	Acceptance	P	Q	R	D
M_1	0	1	0	0	1
M_2	1	1	1	1	1
M_3	0	0	0	1	1
M_4	0	X	0	X	0
M_5	1	1	0	1	0
collective decision	0	1	0	1	1

A majority of members do not accept the decision rule. As we have seen, when $Accept_{agg} = 0$ it means that the group members believe that the most important criteria for the decision are missing or that criteria are relevant but not well connected. Therefore, they express their judgments on the criteria in the rule, but their decision on the conclusion D takes into account what they believe are the missing attributes or the right decision rule. For example, M_3 states that $D = 1$ despite the fact that $P = 0$ and $Q = 0$ because she thinks that the rule is $(P \vee Q \vee R) \leftrightarrow D$. In this situation, the group will conclude $D_{agg} = 1$ following a CBP.

Please note that letting the majority deciding on the group acceptance of the decision rule is just one possibility. Nothing forbids to fix a different quota, such as unanimity or a quota of 2/3 of the agents in order to accept or reject the decision rule at the group level. According to our framework, the original doctrinal paradox would be solved by PBP. In the court example, all judges have to give judgments according to the legal doctrine. Hence, PBP would be enforced. This would be in line with what advocated by some legal theorists, that is in a legal verdict reasons are more important than the final decision as these will form the legal *corpus* for future verdicts.

5 A Solution to the Dilemma

In this section we compare our approach to standard judgment aggregation. In particular, our approach can be seen not only as a more realistic and flexible framework for judgment aggregation problems, but also as an escape route from the paradoxes that trouble judgment aggregation. In order to illustrate why this is the case, we will state the first impossibility theorem [8], recall why PBP and CBP are considered escape routes from the dilemma and, finally, show that our approach is an alternative solution.

The first impossibility theorem of judgment aggregation[1] showed that there exists no aggregation function F that satisfies the following few desirable conditions:

Universal Domain: The domain of F is the set of all profiles of consistent and complete judgment sets.

Anonymity: For any profiles $(J_1, \ldots, J_k)$, $(J'_1, \ldots, J'_k)$ in the domain that are permutations of each other, $F(J_1, \ldots, J_k) = F(J'_1, \ldots, J'_k)$. Intuitively, this means that all agents have equal weight.

Systematicity: For any $P, Q \in \mathcal{A}$ and any profiles $(J_1, \ldots, J_k)$, $(J'_1, \ldots, J'_k)$ in the domain, if $\forall j \in N, P \in J_j \leftrightarrow Q \in J'_j$, then $P \in F(J_1, \ldots, J_k) \leftrightarrow Q \in F(J'_1, \ldots J'_k)$. This condition ensures that the collective judgment on each proposition depends only on the agent judgments on that proposition, and that the aggregation rule is the same across all propositions. Systematicity is clearly a very strong condition. In subsequent impossibility results, systematicity has been weakened to the independence of irrelevant alternatives:

Independence of Irrelevant Alternatives (IIA): For any $P \in \mathcal{A}$ and any profiles $(J_1, \ldots, J_k)$, $(J'_1, \ldots, J'_k)$ in the domain, if $\forall j \in N, P \in J_j \leftrightarrow P \in J'_j$, then $P \in F(J_1, \ldots, J_k) \leftrightarrow P \in F(J'_1, \ldots J'_k)$. In other words, IIA is systematicity without the neutrality condition, requiring that all propositions are equally treated.

It should be now clear why PBP and CBP are escape routes to the dilemmas of judgment aggregation. PBP is a procedure that relaxes the independence of irrelevant alternatives condition: The individuals are requested to express their judgments only on the premises and the collective value of the conclusion is derived by the aggregated values on the premises following the decision rule. On the other hand, CBP can never generate a paradoxical outcome as the conclusion is a literal.

However, as attractive as these procedures can appear, they leave a major open problem: When proposition-wise majority voting collapses into an inconsistent group outcome, PBP and CBP give opposite solutions, as we have seen in the court example. The question is then how we can decide between these two solutions.

[1] For other impossibility theorems that strengthened and expanded the original formulation, see [9].

Our approach is an attempt to provide an answer to this question. By extending the standard judgment aggregation framework and allowing the group members to accept or reject the given decision rule, we introduce a way to decide between PBP and CBP. When the majority (or any other pre-fixed quota of the voters) agrees with the decision rule, the individual judgments are aggregated by PBP. On the other hand, when the group does not agree with the rule, the only opinion the group can provide is about the final decision, so the aggregation procedure will turn to CBP.

Since our approach can always decide whether the aggregation function is PBP or CBP, we provide an escape from the paradoxes that plague standard judgment aggregation.

6 Related Works

In this section we refer to works that proposed to relax some of the assumptions made in the classical judgment aggregation framework. However, our model is the first that combines all these different aspects and introduces new ones.

Results in judgment aggregation usually assume complete judgment sets both at the individual and collective level. Gärdenfors [4] was the first to criticize such assumption as being too strong and unrealistic. He allows voters to abstain from expressing judgments on some propositions in the agenda. He proves that, if the judgment sets may not be complete (but logically closed and consistent), then every aggregation function that is IIA and Paretian[2], must be oligarchic[3]. Gärdenfors' framework requires the agenda to have a very rich logical structure (with an infinite number of issues). More recently, Dokow and Holzman [3] extended Gärdenfors' result and consider finite agendas. Again, impossibility results are obtained. Hence, relaxing the completeness assumption does not avoid the impossibility results.

Nevertheless, allowing the voters to not express their judgments on some of the issues in the agenda provides a more realistic model of judgment aggregation, which is the aim of our paper. In order to avoid confusion, we must observe that we distinguish abstaining from being neutral with respect to an issue in the agenda. Abstentions in Gärdenfors and Dokow and Holzman' works correspond to what we call "neutral judgments". In our model, a voter abstains on a criterion when she deems that this criterion is irrelevant. In this case, she does not state her judgments on a criterion.

In a recent paper, Miller [10] considers judgment aggregation problems in which members have different views on how the premises are connected to

[2] A *Paretian* aggregation function is such that, if all the members in the group adopt the same position on a certain issue, this position will be adopted at the collective level as well.

[3] An aggregation function is *oligarchic* if, for every issue in the agenda, the group adopts a position 0 (resp. 1) if and only if all the members of a subset of the group (the oligarchy) adopt position 0 (resp. 1) on that issue. Clearly, when there is only one member in the oligarchy, it corresponds to dictatorship.

the conclusion. This means that there is no imposed decision rule but, given a set of premises and a conclusion, each member expresses her judgments on the propositions in the agenda as well as providing the decision rule she has used. Miller's framework allows members to use decision rules in which only some of the premises appear. However, group members are requested to express judgments also on criteria that they deem irrelevant for the final decision. The question addressed is whether, once the members have voted following their own decision rules, it is possible to determine a group decision rule that represents how the group see the logical relations between premises and conclusion. Unless the unanimity rule is used, the answer is negative.

Another related work is [11], which considers judgment aggregation situations in which there is a gap between necessary and sufficient conditions to justify a certain decision on the conclusion. An example is the reviewing process for the publication of a paper. Suppose that the criteria for recommending publication of a manuscript (D) are correctness of the results (P) and originality of the ideas (Q). We may agree that P is necessary for recommending publication ($\neg P \rightarrow \neg D$), and also that $P \wedge Q \rightarrow D$. However, the gap between necessary and sufficient conditions is the situation in which we judge the results contained in the submission to be correct but the ideas not original: $P \wedge \neg Q$ is consistent with both acceptance and rejection of a paper. Members may have different views on such gaps. The question posed in [11] is how to justify such individual discrepancies at the group level. Possibilities results are explored, in which majority voting on the conclusion is combined with no veto power on the premises.

Despite the similarities of the above contributions with our approach, the key feature of our proposal is to present a normative procedure: the aggregation rule is PBP or CBP depending on the members view on the given decision rule. The group will always be able to take a decision and, when most of them consider the decision rule to be appropriate, the group will also be able to provide reasons for that decision.

Another way to make the aggregation procedure reactive to the individual opinions about the decision rule is to allow the group members to assign weights to the criteria, as in [2]. The way the group decision is derived (PBP or CBP) depends on whether the final weights are above or below a fixed threshold. The problem of fixing a threshold is common to other frameworks that use similar quantitative approach (see for example, the work by (Dietrich and List 2005) using quota rules). Even if these approaches also provides a more realistic framework to judgment aggregation, problems have to be solved such as : where do the weights come from and how/who should fix the threshold.

7 Conclusion and Future Work

We extended standard judgment aggregation procedure in order to take into account the judgement status and the acceptance of the decision rule. Our framework allows group members to express 0/1 judgments as well as being neutral or abstain on some propositions. In addition, it allows individuals to state whether

or not they agree on the rule governing the decision. The aggregation procedure we propose is more reactive to individual opinions since we use a flexible decision rule (conclusion-based or premises-based) according to the acceptance/rejection of the rule by the group members. Our approach is more realistic and flexible compared to standard judgment aggregation procedure. In addition, it provides an escape route from the paradoxes that trouble judgment aggregation.

This work can be extended in different directions, among which we plan to investigate the relationship between acceptance of the decision rule in our framework and works on coalitions [14]. More precisely, we intend to study how group members can form coalitions and manipulate their judgments in order to drive the decision process in a particular direction. In addition, we plan to investigate the relationship with opinion aggregation in order to go beyond binary judgments [1]. A fuzzy approach would also allow to express the degree of confidence in the decision rule.

References

1. Ben-Arieh, D., Chen, Z.: Linguistic group decision-making: opinion aggregation and measures of consensus. Fuzzy Optimization and Decision Making 5(4), 371–386 (2007)
2. Benamara, F., Kaci, S., Pigozzi, G.: Judgment aggregation with rule confidence scores. In: 12th International Workshop on Non-Monotonic Reasoning (NMR 2008), pp. 2–9 (2008)
3. Dokow, E., Holzman, R.: Aggregation of binary evaluations with abstentions. Journal of Economic Theory 145(2), 544–561 (2010)
4. Gärdenfors, P.: A representation theorem for voting with logical consequences. Economics and Philosophy 22, 181–190 (2006)
5. Kornhauser, L.A., Sager, L.G.: Unpacking the court. Yale Law Journal 96, 82–117 (1986)
6. Kornhauser, L.A., Sager, L.G.: The one and the many: Adjudication in collegial courts. California Law Review 81, 1–51 (1993)
7. List, C.: Judgment aggregation - a bibliography on the discursive dilemma, the doctrinal paradox and decisions on multiple propositions (2007), http://personal.lse.ac.uk/LIST/doctrinalparadox.htm
8. List, C., Pettit, P.: Aggregating sets of judgments: An impossibility result. Economics and Philosophy 18, 89–110 (2002)
9. List, C., Puppe, C.: Judgment aggregation: A survey. In: Pattanaik, P., Anand, P., Puppe, C. (eds.) Oxford Handbook of Rational and Social Choice. Oxford University Press, Oxford (2009)
10. Miller, M.: Judgment aggregation and subjective decision-making. Economics and Philosophy 24, 205–231 (2008)
11. Nehring, K., Puppe, C.: Justifiable group choice. Journal of Economic Theory 145(2), 583–602 (2010)
12. Pettit, P.: Deliberative democracy and the discursive dilemma. Philosophical Issues 11, 268–299 (2001)
13. Pigozzi, G.: Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. Synthese 152(2), 285–298 (2006)
14. Shehory, O., Kraus, S.: Methods for task allocation via agent coalition formation. Artificial Intelligence 101(1-2), 165–200 (1998)

Aggregation of Bounded Fuzzy Natural Number-Valued Multisets

Jaume Casasnovas and J. Vicente Riera

University of Balearic Islands, Palma de Mallorca E 07122, Spain
jaume.casasnovas@uib.es, jvicente.riera@uib.es

Abstract. Multisets (also called bags) are like-structures where an element can appear more than once. Recently, several generalizations of this concept have been studied. In this article we deal with a new extension of this concept, the bounded fuzzy natural number-valued multisets. On this kind of bags, a bounded distributive lattice structure is presented and a partial order is defined. Moreover, we study operations of aggregations (t-norms and t-conorms) and we provide two methods for their construction.

1 Introduction

Multisets (also called bags in the literature [25]) are like-structures where an element can appear more than once. Formally, a multiset over a set of types X is a mapping M defined from X to the set $\mathbb{N} = \{1, 2, \cdots, \}$ of natural numbers. A survey of the mathematics of multisets, including their axiomatic foundation, can be found in [2]. The multisets have been studied by several researchers in computer science from different points of view. For example, their applications to data analysis and decision making [19], their applications to flexible querying [22] or the monograph on multiset processing [13].

According to the interpretation of a multiset $M : X \to \mathbb{N}$, it describes a set or *universe*, Ω, which consists of $M(x)$ "exact" copies of each type $x \in X$. Specifically, for each $x \in X$, $M(x)$ is the account of elements or cardinal of the subset $\Omega_x \subset \Omega$. The number $M(x)$ is usually called the *multiplicity* of x in the multiset M. One of the most natural and simple example is the multiset of prime factors of a natural number n. Thus, the number 504 has the factorization $504 = 2^3 \cdot 3^2 \cdot 7^1$ which gives the multiset $\{2, 2, 2, 3, 3, 7\}$.

Notice that all properties (inclusion, equality, etc.) and operations (addition, union, intersection, etc.) between multisets stem from similar properties and operations of the set of natural numbers. So, a deep study of the valuation set of multisets over a universe X allow us to obtain new properties. And a change of this valuation set allows us to get new extensions.

In [3], the authors introduced a more general definition of "extended multiset" as mappings $M : X \to L$, where L is a finite or infinite chain of natural numbers, or, even, it can be $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$, with the usual operations and order. This definition allows to extend several aggregation operators defined in L, such as t-norms or t-conorms, to multisets.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 67–78, 2010.

Another natural generalization of this interpretation of multisets leads to the notions *Real-Valued Bags and bag relations* [20] or *multisets with fuzzy values* [16,20] over a set of *types* X. Such a multiset describes for each $x \in X$, a set Ω_x consisting this time of "possibly inexact" copies of x with a degree of similarity valued in [0,1]. In this way, in [18] an immediate generalization of crisp multisets using fuzzy numbers instead of natural numbers is proposed. So, provided a suitable definition of fuzzy number (triangular, trapezoidal, Gauss-shaped, etc [15]), it is possible to consider fuzzy Number-Valued multisets defined over X.

Analogously to the crisp case and in order to define a "multiplicity" or *fuzzy multiplicity* of each type for a *fuzzy multiset* over X, we need to associate to each $x \in X$ the *cardinality* of the fuzzy set Ω_x. The problem of "counting" fuzzy sets has generated a lot of literature since Zadeh's first definition of the cardinality of fuzzy sets [14,15]. In particular, the scalar cardinalities of fuzzy sets, which associate to each fuzzy set a positive real number, have been studied from the axiomatic point of view [12] with the aim of capturing different ways of counting additive aspects of fuzzy sets like the cardinalities of supports, of levels, of cores, etc. In a similar way, the fuzzy cardinalities of fuzzy sets [11,14], which associate to any fuzzy set a fuzzy natural number, have also been studied from the axiomatic point of view.

Taking into account that the fuzzy cardinality of a fuzzy set is a fuzzy natural number, i.e., a discrete fuzzy number whose support is a subset of consecutive natural numbers, in [10] the authors defined *Fuzzy Natural Number-Valued multiset* as mappings $M : X \to FNN$ where FNN is the set of fuzzy natural numbers. On this type of multisets, monoidal and lattice structures were studied.

On the other hand, in [3] the authors deal with multisets whose multiplicities are possibly bounded due to circumstances of the framework where they are defined. As a consequence, in this paper we propose a new extension of the concept of multiset, the bounded fuzzy natural number-valued multisets. On this new set of multisets we define a structure of bounded distributive lattice. And, on this bounded partially ordered set we define triangular norms and conorms. Moreover, we propose two methods to get t-norms and t-conorms. The first method uses t-norms(t-conorms) on $\mathcal{A}_1^L = \{u \in FNN \mid supp(u) \subseteq L = \{0, 1, \cdots, m\}\}$. And the second one uses divisible t-norms(t-conorms) on the finite chain $L = \{0, 1, \cdots, m\}$ of natural numbers.

2 Preliminaries

2.1 Multisets

Let X be a crisp set. A (*crisp*) *multiset* over X is a mapping $M : X \to \mathbb{N}$, where $\mathbb{N}$ stands for the set of natural numbers including the 0. A multiset M over X is *finite* if its *support*

$$supp(M) = \{x \in X | M(x) > 0\}$$

is a finite subset of X. We shall denote the sets of all multisets over a set X by $MS(X)$, and by $\perp$ the *null multiset*, defined by $\perp(x) = 0$ for each $x \in X$.

For every $A, B \in MS(X)$, their *sum* [21] $A + B$ is the multiset defined pointwise by

$$(A + B)(x) = A(x) + B(x), \quad x \in X.$$

Let us mention here that it has been argued that this sum +, also called *additive union*, is the right notion of union of multisets. According to the interpretation of multisets as sets of copies of types explained in the introduction, this sum corresponds to the disjoint union of sets, as it interprets that all copies of each x in the set represented by A are different from all copies of it in the set represented by B. This additive sum has quite different properties from the ordinary union of sets. For instance, the collection of submultisets of a given multiset is not closed under this operation and consequently no sensible notion of complement within this collection exists.

For every $A, B \in MS(X)$, their *join* $A \vee B$ and *meet* $A \wedge B$ are respectively the multisets over X defined pointwise by $(A \vee B)(x) = \max(A(x), B(x))$ and $(A \wedge B)(x) = \min(A(x), B(x))$, $x \in X$. If A and B are finite, then $A + B$, $A \vee B$ and $A \wedge B$ are also finite. A partial order $\leq$ on $MS(X)$ is defined by $A \leq B$ if and only if $A(x) \leq B(x)$ for every $x \in X$. If $A \leq B$, then their *difference* $B - A$ is the multiset defined pointwise by

$$(B - A)(x) = B(x) - A(x).$$

2.2 Triangular Norms and Conorms on Partially Ordered Sets

Let $(P; \leq)$ be a non-trivial bounded partially ordered set (poset) with "e" and "m" as minimum and maximum elements respectively.

Definition 2.1. [1] *A* triangular norm *(briefly t-norm) on P is a binary operation $T : P \times P \to P$ such that for all $x, y, z \in P$ the following axioms are satisfied:*

1. $T(x, y) = T(y, x)$ *(commutativity)*
2. $T(T(x, y), z) = T(x, T(y, z))$ *(associativity)*
3. $T(x, y) \leq T(x', y')$ *whenever* $x \leq x', y \leq y'$ *(monotonicity)*
4. $T(x, m) = x$ *(boundary condition)*

Definition 2.2. *A* triangular conorm *(t-conorm for short) on P is a binary operation $S : P \times P \to P$ which, for all $x, y, z \in P$ satisfies (1), (2), (3) and (4′): $S(x, e) = x$, as boundary condition.*

2.3 Triangular Norms and Conorms on Discrete Settings

Let L be the totally ordered set $L = \{0, 1, \ldots, m\} \subset \mathbb{N}$. A t-norm(t-conorm) defined on L will be called a discrete t-norm(t-conorm).

Definition 2.3. [17] *A t-norm(t-conorm) $T(S) : L \times L \to L$ is said to be* smooth *if it satisfies $T(S)(x+1, y) - T(S)(x, y) \leq 1$ and $T(S)(x, y+1) - T(S)(x, y) \leq 1$.*

Definition 2.4. [17] *A t-norm(t-conorm) $T : L \times L \to L$ is said to be* divisible *if it satisfies: For all $x, y \in L$ with $x \leq y$, there is $z \in L$ such that $x = T(y, z)(y = S(x, z))$.*

2.4 Discrete Fuzzy Numbers

By a fuzzy subset of the set of real numbers, we mean a function $u : \mathbb{R} \to [0,1]$. For each fuzzy subset u, let $u^\alpha = \{x \in \mathbb{R} : u(x) \geq \alpha\}$ for any $\alpha \in (0,1]$ be its α-level set (or α-cut). By $supp(u)$, we mean the support of u, i.e. the set $\{x \in \mathbb{R} : u(x) > 0\}$. By u^0, we mean the closure of $supp(u)$.

Definition 2.5. *[23] A fuzzy subset u of the set of real numbers $\mathbb{R}$ with membership mapping $u : \mathbb{R} \to [0,1]$ is called* discrete fuzzy number *if its support is finite, i.e., there are $x_1, ..., x_n \in \mathbb{R}$ with $x_1 < x_2 < ... < x_n$ such that $supp(u) = \{x_1, ..., x_n\}$, and there are natural numbers s, t with $1 \leq s \leq t \leq n$ such that:*

1. *$u(x_i)$=1 for any natural number i with $s \leq i \leq t$ (*core*)*
2. *$u(x_i) \leq u(x_j)$ for each natural number i, j with $1 \leq i \leq j \leq s$*
3. *$u(x_i) \geq u(x_j)$ for each natural number i, j with $t \leq i \leq j \leq n$*

Remark 2.1. If the fuzzy subset u is a discrete fuzzy number then the support of u coincides with its closure, i.e. $supp(u) = u^0$.

From now on, the notation DFN stands for the set of discrete fuzzy numbers.

The operations addition, maximum and minimum between discrete fuzzy numbers defined through Extension principle [15] can yield fuzzy subsets that do not satisfy the conditions to be discrete fuzzy numbers [4,24]. In [4,5,6,24], this drawback is studied and a new method to define these operations is proposed. So, the next result holds [24]:

Theorem 2.1. *Let $u, v \in DFN$, the fuzzy subset denoted by $u \underset{W}{\oplus} v$, such that it has as r-cuts the sets $[u \underset{W}{\oplus} v]^r = \{x \in supp(u) + supp(v) : \min([u]^r + [v]^r) \leq x \leq \max([u]^r + [v]^r)\}$ for each $r \in [0,1]$ where $\min([u]^r + [v]^r) = \min\{x : x \in [u]^r + [v]^r\}$, $\max([u]^r + [v]^r) = \max\{x : x \in [u]^r + [v]^r\}$ and $(u \underset{W}{\oplus} v)(x) = \sup\{r \in [0,1]$ such that $x \in [u \underset{W}{\oplus} v]^r\}$ is a discrete fuzzy number.*

On the other hand, in [6], the following result is obtained:

Proposition 2.1. *For each $u, v \in DFN$, there exist two unique discrete fuzzy numbers, which we will denote by $MIN_w(u,v)$ and $MAX_w(u,v)$, such that they have the sets $MIN_w(u,v)^\alpha$ and $MAX_w(u,v)^\alpha$ as α-cuts respectively, where*

$$MIN_w(u,v)^\alpha = \{z \in supp(u) \bigwedge supp(v) | \min(x_1^\alpha, y_1^\alpha) \leq z \leq \min(x_p^\alpha, y_k^\alpha)\}$$

$$MAX_w(u,v)^\alpha = \{z \in supp(u) \bigvee supp(v) | \max(x_1^\alpha, y_1^\alpha) \leq z \leq \max(x_p^\alpha, y_k^\alpha)\}$$

for each $\alpha \in [0,1]$, being $u^\alpha = \{x_1^\alpha, \cdots, x_p^\alpha\}$, $v^\alpha = \{y_1^\alpha, \cdots, y_k^\alpha\}$ the α-cuts of u and v respectively. And, $supp(u) \bigwedge supp(v) = \{z = min(x,y) | x \in supp(u), y \in supp(v)\}$ and $supp(u) \bigvee supp(v) = \{z = max(x,y) | x \in supp(u), y \in supp(v)\}$

3 Operations on Fuzzy Natural Numbers

From now on, the notation fnn stands for a fuzzy natural number (i.e. discrete fuzzy numbers whose support only includes consecutive natural numbers) and FNN stands for the set of fuzzy natural numbers.

3.1 Addition of Fuzzy Natural Numbers

It is well known [15] that, in the case of continuous fuzzy numbers the addition obtained by extending the usual addition of real numbers through the extension principle is associative and commutative. But the fuzzy natural numbers are not continuous on $\mathbb{R}$.

In [5], the authors proved that in the case in which the discrete fuzzy numbers have as support an arithmetic sequence or a subset of consecutive natural numbers it is possible to use the Zadeh's extension principle to obtain its addition. Moreover, we know [24] the next result:

Proposition 3.1. *Let us consider $u, v \in DFN$. If $u \oplus v \in DFN$ where $u \oplus v$ denotes the addition of u and v using the Zadeh's extension principle, then $u \oplus v$ and $u \underset{W}{\oplus} v$ are identical, where $u \underset{W}{\oplus} v$ is the discrete fuzzy number obtained from u and v according to Theorem 2.1.*

Remark 3.1. A consequence of the previous proposition is that if we prove a property for the operation $\underset{W}{\oplus}$ in the set of fuzzy natural numbers, we will obtain the same property for the operation $\oplus$ in this set.

Theorem 3.1. *[10] The set FNN of the fuzzy natural numbers is a commutative monoid with the Zadeh's addition as a monoidal operation.*

3.2 Maximum and Minimum of Fuzzy Natural Numbers

With respect to the maximum and the minimum of two fuzzy natural numbers, the authors have proved in [6] the following proposition:

Proposition 3.2. *[6] Let u, v be two fuzzy natural numbers. Then $MAX(u, v)$, defined through the extension principle, coincides with $MAX_w(u, v)$. So, if $u, v \in FNN$, $MAX(u, v)$ is a fuzzy natural number and $MAX(u, v) \in FNN$. Analogously, $MIN(u, v)$, defined through the extension principle, coincides with the fnn $MIN_w(u, v)$. So, if $u, v \in FNN$, then $MIN(u, v)$ is a fuzzy natural number and $MIN(u, v) \in FNN$.*

But we have studied in [7] the associativity, commutativity, idempotence, absorption and distributivity for the operations MIN_w and MAX_w between discrete fuzzy numbers in general and between fuzzy natural numbers in particular and we obtained the following proposition:

Proposition 3.3. *[7] The set of discrete fuzzy numbers whose support is a sequence of consecutive natural numbers (FNN,MIN_w,MAX_w) is a distributive lattice.*

If we gather the previous Propositions 3.2 and 3.3, then we obtain the following consequence:

Proposition 3.4. *[7] The set of discrete fuzzy numbers whose support is a sequence of consecutive natural numbers (FNN,MIN,MAX) is a distributive lattice.*

With the aim of studying the monotony for the addition of two fuzzy natural numbers, we need a definition of order:

Definition 3.1. *[7] From the operations MIN_w and MAX_w, we can define a partial order on FNN on the following way:*
$u \preceq v$ if and only if $MIN_w(u,v) = u$, or equivalently, $u \preceq v$ if and only if $MAX_w(u,v) = v$ for any $u, v \in FNN$. Equivalently, we can also define the partial ordering in terms of α-cuts:
$u \preceq v$ if and only if $\min(u^\alpha, v^\alpha) = u^\alpha$
$u \preceq v$ if and only if $\max(u^\alpha, v^\alpha) = v^\alpha$

Proposition 3.5. *[10] Let $u, v, w, t \in FNN$. If $u \preceq v$ and $w \preceq t$ where $\preceq$ denotes the partial order in FNN defined in Definition 3.1 then $u \oplus w \preceq v \oplus t$, where $\oplus$ denotes the Zadeh's addition.*

3.3 Discrete Fuzzy Numbers Obtained by Extending Discrete t-norms(t-conorms) Defined on a Finite Chain

Let us consider a discrete t-norm(t-conorm) $T(S)$ on the finite chain $L = \{0, 1, \cdots, m\} \subset \mathbb{N}$. If X and Y are subsets of L, then the subset $\{T(x,y) | x \in \mathsf{X}, y \in \mathsf{Y}\} \subseteq L$ will be denoted by $T(\mathsf{X}, \mathsf{Y})$. Analogously, $S(\mathsf{X}, \mathsf{Y}) = \{S(x,y) | x \in \mathsf{X}, y \in \mathsf{Y}\}$.

So, if we consider the α-cut sets, $u^\alpha = \{x_1^\alpha, ..., x_p^\alpha\}$, $v^\alpha = \{y_1^\alpha, ..., y_k^\alpha\}$, for u and v respectively then $T(u^\alpha, v^\alpha) = \{T(x,y) | x \in u^\alpha, y \in v^\alpha\}$ and $S(u^\alpha, v^\alpha) = \{S(x,y) | x \in u^\alpha, y \in v^\alpha\}$ for each $\alpha \in [0,1]$, where u^0 and v^0 denote $supp(u)$ and $supp(v)$ respectively.

Definition 3.2. *[8] For each $\alpha \in [0,1]$, let us consider the sets*

$$C^\alpha = \{z \in T(supp(u), supp(v)) | \min T(u^\alpha, v^\alpha) \leq z \leq \max T(u^\alpha, v^\alpha)\} \text{ and}$$

$$D^\alpha = \{z \in S(supp(u), supp(v)) | \min S(u^\alpha, v^\alpha) \leq z \leq \max S(u^\alpha, v^\alpha)\}$$

Theorem 3.2. *[8] There exists a unique discrete fuzzy number that will be denoted by $\mathcal{T}(u,v)(\mathcal{S}(u,v))$ such that $\mathcal{T}(u,v)^\alpha = C^\alpha (\mathcal{S}(u,v)^\alpha = D^\alpha)$ for each $\alpha \in [0,1]$ and $\mathcal{T}(u,v)(z) = sup\{\alpha \in [0,1] : z \in C^\alpha\} (\mathcal{S}(u,v)(z) = sup\{\alpha \in [0,1] : z \in D^\alpha\})$*

From now on the set $\mathcal{A}_1^L = \{u \in FNN \mid supp(u) \subseteq L = \{0, 1, \cdots, m\}\}$ will be called the set of bounded fuzzy natural numbers and each element of this set will be called a bounded fuzzy natural number (in short bfnn). In [9] the authors showed the next result

Theorem 3.3. *The triplet $(\mathcal{A}_1^L, MIN_w, MAX_w)$ is a bounded distributive lattice and the fnn $\widehat{0}$ and $\widehat{m}$, defined by $\widehat{0}(i) = 1$ if $i = 0$ and $\widehat{0}(i) = 0$, otherwise, $\widehat{m}(i) = 1$ if $i = m$ and $\widehat{m}(i) = 0$, otherwise, are the lower and the upper bound, respectively.*

On the other hand, it is well known [1] that it is possible to generalize the concept of t-norm (t-conorm) using any bounded partially ordered set instead of the unit interval. Using this idea and Theorem 3.3 we can build t-norms and t-conorms on the bounded distributive lattice $\mathcal{A}_1^L$.

Theorem 3.4. *[9] Let $T(S)$ be a divisible t-norm(t-conorm) on L and let*

$$\begin{aligned} \mathcal{T}(\mathcal{S}) : \mathcal{A}_1^L \times \mathcal{A}_1^L &\to \mathcal{A}_1^L \\ (A, B) &\longmapsto \mathcal{T}(u, v)(\mathcal{S}(u, v)) \end{aligned}$$

be the binary operation (which will be called the extension of the t-norm(t-conorm) $T(S)$ to $\mathcal{A}_1^L$), where $\mathcal{T}(u, v)(\mathcal{S}(u, v))$ are defined according to Theorem 3.2. Then, $\mathcal{T}(\mathcal{S})$ is a t-norm(t-conorm) on the bounded set $\mathcal{A}_1^L$.

4 Operations on Fuzzy Natural Number-Valued Multisets

4.1 FNN-Valued Multisets[10]

Definition 4.1. *A Fuzzy Natural Number-valued multiset defined over an universe X is a mapping $M : X \to FNN$ i.e. for all $x \in X$, $M(x)$ is a fuzzy natural number.*

Remark 4.1. We will denote the set of Fuzzy Natural Number-valued multisets defined over an universe X by $FNNM(X)$. Finally, the abbreviation fnnm will denote a Fuzzy Natural Number-valued multiset.

The properties of the addition of fuzzy natural numbers studied in the previous Section 3, will allow us to define the addition of fuzzy natural number-valued multisets and to study the monoidal structure of this set.

Definition 4.2. *Let $A, B : X \to FNN$ be two Fuzzy Natural Number-valued multisets. The* sum *of A and B will be the Fuzzy Natural Number-valued Multiset pointwise defined for all $x \in X$ by*

$$(A + B)(x) = A(x) \oplus B(x)$$

where the fnn $A(x) \oplus B(x)$ is obtained following the Zadeh's extension principle or equivalently using the method considered in Theorem 2.1.

Proposition 4.1. *The set $FNNM(X)$ of the fuzzy natural number-valued multisets over X is a commutative monoid with the addition as a monoidal operation.*

Analogously to the addition, the properties of the maximum and minimum of fnn studied in the previous section will allow us to define the maximum and minimum of fuzzy natural number-valued multisets and to study the order and the lattice structure of this set.

Definition 4.3. *Let $A, B : X \to FNN$ be two Fuzzy Natural Number-valued Multisets. The* join *and the* meet *of A and B will be the Fuzzy Natural Number-valued Multiset, pointwise defined for all $x \in X$ as*

$$(A \vee B)(x) = MAX\{A(x), B(x)\} \text{ and } (A \wedge B)(x) = MIN\{A(x), B(x)\}$$

respectively, where the fnn $MAX\{A(x), B(x)\}$ and $MIN\{A(x), B(x)\}$ are obtained according to the method presented in Proposition 2.1.

Proposition 4.2. *As long as, for all $x \in X$, $A(x) \in FNN$ and $B(x) \in FNN$, then $MAX\{A(x), B(x)\}$ and $MIN\{A(x), B(x)\}$ can be obtained by means of the extension principle.*

Proposition 4.3. *Let $A, B : X \to FNN$ be two Fuzzy Natural Number-valued Multisets. The binary relationship:*
$A \leq B$ if and only if $A \vee B = B$ and/or $A \wedge B = A$ i.e. $MAX\{A(x), B(x)\} = B(x), \forall x \in X$ (or $MIN\{A(x), B(x)\} = A(x), \forall x \in X$) is a partial order on the set $FNNM(X)$.

Proposition 4.4. *The set $FNNM(X)$ of the fuzzy natural number-valued multisets over X is a lattice with the partial order defined in Proposition 4.3 and the meet and join operations proposed in Definition 4.3.*

Proposition 4.5. *Let $A, B, C, D \in FNNM(X)$. If $A \leq B$ and $C \leq D$ where $\leq$ denotes the partial order in $FNNM(X)$ defined in Proposition 4.3 then $A + C \leq B + D$, where $+$ denotes the addition considered in Definition 4.2.*

5 Bounded Fuzzy Natural Numbers-Valued Multisets

Let us consider the finite chain $L = \{0, 1, \cdots, m\}$ of natural numbers and the set $\mathcal{A}_1^L = \{A \in FNN \mid supp(A) \subseteq L\}$ of bounded fuzzy natural numbers.

Definition 5.1. *Let X be a finite set or univers. A bounded fuzzy natural number-valued multiset is a function*

$$\begin{aligned} M : X &\longrightarrow \mathcal{A}_1^L \\ x &\longmapsto M(x) \end{aligned}$$

where $M(x)$ is a bounded fuzzy natural number. Usually, the function $\mathcal{M}(\cdot)$ is called count or multiplicity of M.

Remark 5.1. We will denote the set of Bounded Fuzzy Natural Number-valued multisets defined over an universe X by $BFNNM(X)$. Finally, the abbreviation bfnnm will denote a Bounded Fuzzy Natural Number-valued multiset.

The lattice structure on the set of fuzzy natural numbers considered in the previous section, will allow us to define similar algebraic structures and lattice operations (meet and join) on the set $BFNNM(X)$.

5.1 Distributive Bounded Sublattices of FNNM(X)

According to Proposition 4.4, we know that $FNNM(X)$ constitutes a partially ordered set which is a lattice. Now, using this fact, we want to see that the set $BFNNM(X)$ is a bounded distributive sublattice of the lattice $FNNM(X)$.

Definition 5.2. *Let $A, B : X \to BFNN$ be two Bounded Fuzzy Natural Number-valued Multisets. The* join *and the* meet *of A and B will be the Bounded Fuzzy Natural Number-valued Multiset, pointwise defined for all $x \in X$ as*

$$(A \vee B)(x) = MAX\{A(x), B(x)\} \text{ and } (A \wedge B)(x) = MIN\{A(x), B(x)\}$$

respectively, where the bfnn $MAX\{A(x), B(x)\}$ and $MIN\{A(x), B(x)\}$ are obtained according to the method presented in Proposition 2.1.

Remark 5.2. It is straightforward to see that if $A(x), B(x) \in \mathcal{A}_1^L$ for all $x \in X$ then the fnn $(A \vee B)(x) = MAX\{A(x), B(x)\}$ and $(A \wedge B)(x) = MIN\{A(x), B(x)\}$ belong to $BFNN$. So, the above operations $(A \vee B)$ and $(A \wedge B)$ are well defined. Moreover from proposition 3.2, $MAX\{A(x), B(x)\}$ and $MIN\{A(x), B(x)\}$ can be obtained by means of the extension principle as well.

Similarly to Proposition 4.3, it is possible to build a partial order on the set $BFNNM(X)$ using the operations join and meet considered in Definition 5.2:

Proposition 5.1. *Let $A, B : X \to BFNN$ be two Bounded Fuzzy Natural Number-valued Multisets. The binary relationship:*
$A \leq B$ if and only if $A \vee B = B$ and/or $A \wedge B = A$ i.e. $MAX\{A(x), B(x)\} = B(x), \forall x \in X$ (or $MIN\{A(x), B(x)\} = A(x), \forall x \in X$) is a partial order on the set $BFNNM(X)$.

Now, we will use this partial order to show that the set $BFNNM(X)$ has a structure of bounded distributive lattice.

Proposition 5.2. *The set $BFNNM(X)$ of the bounded fuzzy natural number-valued multisets over X is a bounded distributive lattice with the partial order defined in Proposition 5.1 and the meet and join operations proposed in Definition 5.2.*

Proof. The distributive lattice structure follows because $(\mathcal{A}_1^L, MIN, MAX)$ is a bounded distributive lattice (see Proposition 3.3). Moreover it is easy to see that the bfnnm M_0 such that $M_0(x) = \widehat{0}$ for all $x \in X$ (being $\widehat{0}$ the minimum of

lattice of bounded fuzzy natural numbers $(\mathcal{A}_1^L, MIN, MAX))$ is the minimum of $BFNNM(X)$. And, M_m such that $M_m(x) = \widehat{m}$ for all $x \in X$ (being $\widehat{m}$ the maximum of the lattice of bounded fuzzy natural numbers $(\mathcal{A}_1^L, MIN, MAX))$ is the maximum of $BFNNM(X)$.

5.2 Triangular Norms and Triangular Conorms on $BFNNM(X)$

As we have discussed in the previous Section 3.3, we know [1] that it is possible to consider t-norms(t-conorms) on any bounded partially ordered set. For this reason, we can define t-norms(t-conorms) on the bounded distributive lattice $(BFNNM(X), MIN, MAX, M_0, M_m)$.

Definition 5.3. *A t-norm(t-conorm)* $\mathbb{T}(\mathbb{S})$ *on the bounded partially ordered set* $BFNNM(X)$ *is a function*

$$\begin{array}{rl} \mathbb{T}(\mathbb{S}) : BFNNM(X) \times BFNNM(X) \to & BFNNM(X) \\ (A, B) \longmapsto & \mathbb{T}(A, B)(\mathbb{S}(A, B)) \end{array}$$

such that fulfills the following properties:

i) Commutativity: For all $M, N \in BFNNM(X)$

$$\mathbb{T}(A, B) = \mathbb{T}(A, B) \text{ and } \mathbb{S}(A, B) = \mathbb{S}(A, B)$$

ii) Monotonicity: For $A \leq B$, $C \leq D$

$$\mathbb{T}(A, C) \leq \mathbb{T}(B, D) \text{ and } \mathbb{S}(A, C) \leq \mathbb{S}(B, D)$$

iii) Associativity: For all $A, B, C \in BFNNM(X)$

$$\mathbb{T}(\mathbb{T}(A, B), C) = \mathbb{T}(A, \mathbb{T}(B, C)) \text{ and } \mathbb{S}(\mathbb{S}(A, B), C) = \mathbb{S}(A, \mathbb{S}(B, C))$$

iv) Boundary condition: For all $A \in BFNNM(X)$

$$\mathbb{T}(A, M_m) = A \text{ and } \mathbb{S}(A, M_0) = A$$

In the next proposition we will see that it is possible to construct a t-norm on the bounded distributive lattice $BFNNM(X)$ from a t-norm defined on the bounded distributive lattice $\mathcal{A}_1^L$.

Proposition 5.3. *For each t-norm* $\mathcal{T}$ *defined on* $\mathcal{A}_1^L$ *it is possible to build a t-norm* $\mathbb{T}$ *on the bounded distributive lattice* $BFNNM(X)$ *on the following way:* $\mathbb{T}(A, B)$ *is the bfnnm such that for each* $x \in X$

$$\mathbb{T}(A, B)(x) = \mathcal{T}(A(x), B(x))$$

Proof. It is straightforward because $\mathcal{T}$ is a t-norm.

Analogously,

Proposition 5.4. *For each t-conorm $\mathcal{S}$ defined on $\mathcal{A}_1^L$ it is possible to give a t-norm $\$$ on the bounded distributive lattice $BFNNM(X)$ on the following way: $\$(A, B)$ is the bfnnm such that for each $x \in X$*

$$\$(A, B)(x) = \mathcal{S}(A(x), B(x))$$

Proof. It is straightforward because $\mathcal{S}$ is a t-conorm.

From Theorem 3.4, we know that if $T(S)$ are divisible t-norm(t-conorm) on $L = \{0, \cdots, m\}$ it is possible to construct a t-norm(t-conorm) on the bounded distributive lattice of bounded fuzzy natural numbers $\mathcal{A}_1^L$. Using this fact we will see that for each divisible t-norm(t-conorm) on L it is possible to obtain a t-norm(t-conorm) on bounded partially ordered set $BFNNM(X)$.

Proposition 5.5. *For each divisible t-norm T defined on the finite chain L it is possible to build a t-norm $\mathbb{T}$ on the bounded distributive lattice $BFNNM(X)$.*

Proof. From Theorem 3.4 for each divisible t-norm T on L it is possible to obtain a t-norm $\mathcal{T}$ on $\mathcal{A}_1^L$. Now from Proposition 5.3 the proof is straightforward.

Similarly,

Proposition 5.6. *For each divisible t-conorm S defined on the finite chain L it is possible to build a t-conorm $\$$ on the bounded distributive lattice $BFNNM(X)$.*

6 Conclusion

We have introduced a possible extension of the concept of multiset, the bounded fuzzy natural number-valued multisets. On these bags, a bounded distributive lattice structure is presented and triangular operations have been defined.

Future studies aim to investigate the properties of these triangular operations and their application to build negation function and implication function on this bounded lattice.

Acknowledgments. We would like to express our thanks to anonymous reviewers who have contributed to improve this article. This work has been partially supported by the MTM2009-10962 project grant.

References

1. De Baets, B., Mesiar, R.: Triangular norms on product lattices. Fuzzy Sets and Systems 104, 61–75 (1999)
2. Blizard, W.D.: The development of multiset theory. Modern Logic, 319–352 (1991)
3. Casasnovas, J., Mayor, G.: Discrete t-norms and operations on extended multisets. Fuzzy sets and Systems 159, 1165–1177 (2008)
4. Casasnovas, J., Riera, J.V.: On the addition of discrete fuzzy numbers. WSEAS Transactions on Mathematics, 549–554 (2006).

5. Casasnovas, J., Riera, J.V.: Discrete fuzzy numbers defined on a subset of natural numbers. Theoretical Advances and Applications of Fuzzy Logic and Soft Computing: Advances in Soft Computing 42, 573–582 (2007)
6. Casasnovas, J., Riera, J.V.: Maximum and minimum of discrete fuzzy numbers. Frontiers in Artificial Intelligence and Applications: Artificial Intelligence Research and Development 163, 273–280 (2007)
7. Casasnovas, J., Riera, J.V.: Lattice properties of discrete fuzzy numbers under extended min and max. In: Proceedings IFSA-EUSFLAT, pp. 647–652 (2009)
8. Casasnovas, J., Riera, J.V.: Extension of discrete t-norms and t-conorms to discrete fuzzy numbers. In: Proceedings of the Fifth International Summer School on Aggregation Operators (AGOP 2009), pp. 77–82 (2009)
9. Casasnovas, J., Riera, J.V.: Triangular norms and conorms on the set of discrete fuzzy numbers. In: IPMU (accepted 2010)
10. Casasnovas, J., Riera, J.V.: Aggregation and arithmetic operations on fuzzy natural number-valued multisets. In: Proceedings ESTYLF 2010, pp. 309–314 (2010)
11. Casasnovas, J., Torrens, J.: An Axiomatic Approach to the fuzzy cardinality of finite fuzzy sets. Fuzzy Sets and Systems 133, 193–209 (2003)
12. Casasnovas, J., Torrens, J.: Scalar cardinalities of finite fuzzy sets for t-norms and t-conorms. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11, 599–615 (2003)
13. Calude, C.S., Păun, G., et al. (eds.): Multiset Processing, Mathematical, Computer Science and Molecular Computing Points of View. LNCS, vol. 2235. Springer, Heidelberg (2001)
14. Dubois, D.: A new definition of the fuzzy cardinality of finite sets preserving the classical additivity property. Bull. Stud. Ecxch. Fuzziness Appl. BUSEFAL (5), 11–12 (1981)
15. Klir, G., Bo, Y.: Fuzzy sets and fuzzy logic. Theory and Applications. Prentice Hall, Englewood Cliffs (1995)
16. Li, B.: Fuzzy Bags and Applications. Fuzzy Sets and Systems 34, 61–71 (1990)
17. Mayor, G., Torrens, J.: Triangular norms on discrete settings. In: Klement, E.P., Mesiar, R. (eds.) Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms, pp. 189–230. Elsevier, Amsterdam (2005)
18. Miyamoto, S.: Generalizations of Multisets and Rough Approximations. International Journal of Intel. Systems 19, 639–652 (2004)
19. Miyamoto, S.: Generalized Bags, Bag Relations and Applications to Data Analysis and Decision Making. In: Torra, V., Narukawa, Y., Inuiguchi, M. (eds.) MDAI 2009. LNCS (LNAI), vol. 5861, pp. 37–54. Springer, Heidelberg (2009)
20. Miyamoto, S.: Operations for Real-Valued Bags and Bag Relations. In: Proceedings IFSA-EUSFLAT 2009, pp. 612–617 (2009)
21. Syropulos, A.: Mathematics of Multisets. In: Calude, C.S., Pun, G., Rozenberg, G., Salomaa, A. (eds.) Multiset Processing. LNCS, vol. 2235, pp. 154–160. Springer, Heidelberg (2001)
22. Rocacher, D.: On fuzzy bags and their application to flexible quering. Fuzzy Sets and Systems 140, 93–110 (2003)
23. Voxman, W.: Canonical representations of discrete fuzzy numbers. Fuzzy Sets and Systems 54, 457–466 (2001)
24. Wang, G., Wu, C., Zhao, C.: Representation and Operations of discrete fuzzy numbers. Southeast Asian Bulletin of Mathematics 28, 1003–1010 (2005)
25. Yager, R.: On the theory of bags. Inter. J. General Systems 13, 23–37 (1986)

Sugeno Utility Functions I: Axiomatizations

Miguel Couceiro[1] and Tamás Waldhauser[1,2]

[1] University of Luxembourg
6, rue Richard Coudenhove-Kalergi, L-1359 Luxembourg
miguel.couceiro@uni.lu
[2] Bolyai Institute, University of Szeged
Aradi vértanúk tere 1, H-6720 Szeged, Hungary
twaldha@math.u-szeged.hu

Abstract. In this paper we consider a multicriteria aggregation model where local utility functions of different sorts are aggregated using Sugeno integrals, and which we refer to as Sugeno utility functions. We propose a general approach to study such functions via the notion of pseudo-Sugeno integral (or, equivalently, pseudo-polynomial function), which naturally generalizes that of Sugeno integral, and provide several axiomatizations for this class of functions.

Keywords: Pseudo-Sugeno integral, pseudo-polynomial function, local utility function, overall utility function, Sugeno utility function, axiomatization.

1 Introduction

The importance of aggregation functions is made apparent by their wide use, not only in pure mathematics (e.g., in the theory of functional equations, measure and integration theory), but also in several applied fields such as operations research, computer and information sciences, economics and social sciences, as well as in other experimental areas of physics and natural sciences. For general background, see [1,14] and for a recent reference, see [13].

In many applications, the values to be aggregated are first to be transformed by mappings $\varphi_i \colon X_i \to Y$, $i = 1, \ldots, n$, so that the transformed values (which are usually real numbers) can be aggregated in a meaningful way by a function $M \colon Y^n \to Y$. The resulting composed function $U \colon X_1 \times \cdots \times X_n \to Y$ is then defined by

$$U(x_1, \ldots, x_n) = M(\varphi_1(x_1), \ldots, \varphi_n(x_n)). \tag{1}$$

Such an aggregation model is used for instance in multicriteria decision making where the criteria are not commensurate. Here each φ_i is a local utility function, i.e., order-preserving mapping, and the resulting function U is referred to as an overall utility function (also called global preference function). For general background see [2].

In this paper, we consider this aggregation model in a purely ordinal decision setting, where Y and each X_i are bounded chains L and L_i, respectively, and

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 79–90, 2010.

where $M\colon L^n \to L$ is a Sugeno integral (see [10,19,20]) or, more generally, a lattice polynomial function. We refer to the resulting compositions as pseudo-Sugeno integrals and pseudo-polynomial functions, respectively. The particular case when each L_i is the same chain L', and each φ_i is the same mapping $\varphi\colon L' \to L$, was studied in [8] where the corresponding compositions $U = M \circ \varphi$ were called quasi-Sugeno integrals and quasi-polynomial functions. Such mappings were characterized as solutions of certain functional equations and in terms of necessary and sufficient conditions which have natural interpretations in decision making and aggregation theory.

Here, we take a similar approach and study pseudo-Sugeno integrals from an axiomatic point of view, and seek necessary and sufficient conditions for a given function to be factorizable as a composition of a Sugeno integral with unary maps. The importance of such an axiomatization is attested by the fact that this framework subsumes the Sugeno utility model. Since overall utility functions (1) where M is a Sugeno integral, coincide exactly with order-preserving pseudo-Sugeno integrals (see Sect. 5 in the companion paper [9]), we are particular interested in the case when the inner mappings φ_i are local utility functions.

The paper is organized as follows. In Sect. 2 we recall the basic definitions and terminology, as well as the necessary results concerning polynomial functions (and, in particular, Sugeno integrals) used in the sequel. In Sect. 3, we focus on pseudo-Sugeno integrals as a tool to study certain overall utility functions. We introduce the notion of pseudo-polynomial function in Subsect. 3.1 and show that, even though seemingly more general, it can be equivalently defined in terms of Sugeno integrals. An axiomatization of this class of generalized polynomial functions is given in Subsect. 3.2. Sugeno utility functions are introduced in Subsect. 3.3, as certain order-preserving pseudo-Sugeno integrals, and then characterized in Subsect. 3.4 by means of necessary and sufficient conditions which extend well-known properties in aggregation function theory. Within this general setting for studying Sugeno utility functions, it is natural to consider the inverse problem which asks for factorizations of a Sugeno utility function as a composition of a Sugeno integral with local utility functions. This question is addressed in Sect. 4, and left as an open problem to be considered in the companion paper [9] submitted to this same volume.

2 Lattice Polynomial Functions and Sugeno Integrals

2.1 Preliminaries

Throughout this paper, let L denote an arbitrary bounded chain endowed with lattice operations $\wedge$ and $\vee$, and with least and greatest elements 0_L and 1_L, respectively, where the subscripts may be omitted when the underlying lattice is clear from the context. A subset S of a chain L is said to be *convex* if for every $a, b \in S$ and every $c \in L$ such that $a \leq c \leq b$, we have $c \in S$. For any subset $S \subseteq L$, we denote by $\mathrm{cl}(S)$ the convex hull of S, that is, the smallest convex subset of L containing S. For instance, if $a, b \in L$ such that $a \leq b$, then $\mathrm{cl}(\{a, b\}) = [a, b] = \{c \in L : a \leq c \leq b\}$.

For an integer $n \geq 1$, we set $[n] = \{1, \ldots, n\}$. Let σ be a permutation on $[n]$. The *standard simplex* of L^n associated with σ is the subset $L^n_\sigma \subseteq L^n$ defined by

$$L^n_\sigma = \{\boldsymbol{x} \in L^n \colon x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}\}.$$

Given arbitrary bounded chains L_i, $i \in [n]$, their Cartesian product $\prod_{i\in[n]} L_i$ constitutes a bounded distributive lattice by defining

$$\boldsymbol{a} \wedge \boldsymbol{b} = (a_1 \wedge b_1, \ldots, a_n \wedge b_n), \quad \text{and} \quad \boldsymbol{a} \vee \boldsymbol{b} = (a_1 \vee b_1, \ldots, a_n \vee b_n).$$

For $k = 1, \ldots, n$ and $c \in L_k$, we use $\boldsymbol{x}^c_k$ to denote the vector whose ith component is c, if $i = k$, and x_i, otherwise.

In the case when $L_i = L$, for all $i \in [n]$, we also use the following notation. For $c \in L$ and $\boldsymbol{x} \in L^n$, let $\boldsymbol{x} \wedge c = (x_1 \wedge c, \ldots, x_n \wedge c)$ and $\boldsymbol{x} \vee c = (x_1 \vee c, \ldots, x_n \vee c)$, and denote by $[\boldsymbol{x}]_c$ the n-tuple whose ith component is 0, if $x_i \leq c$, and x_i, otherwise, and by $[\boldsymbol{x}]^c$ the n-tuple whose ith component is 1, if $x_i \geq c$, and x_i, otherwise.

Let $f \colon \prod_{i\in[n]} L_i \to L$ be a function. The *range* of f is given by $\mathrm{ran}(f) = \{f(\boldsymbol{x}) : \boldsymbol{x} \in \prod_{i\in[n]} L_i\}$. Also, f is said to be *order-preserving* if, for every $\boldsymbol{a}, \boldsymbol{b} \in \prod_{i\in[n]} L_i$ such that $\boldsymbol{a} \leq \boldsymbol{b}$, we have $f(\boldsymbol{a}) \leq f(\boldsymbol{b})$. A well-known example of an order-preserving function is the *median* function $\mathrm{med} \colon L^3 \to L$ given by $\mathrm{med}(x_1, x_2, x_3) = (x_1 \wedge x_2) \vee (x_1 \wedge x_3) \vee (x_2 \wedge x_3)$. Given a vector $\boldsymbol{x} \in L^m$, $m \geq 1$, set $\langle \boldsymbol{x} \rangle_f = \mathrm{med}(f(\boldsymbol{0}), \boldsymbol{x}, f(\boldsymbol{1}))$.

2.2 Basic Background on Polynomial Functions and Sugeno Integrals

In this subsection we recall some well-known results concerning polynomial functions that will be needed hereinafter. For further background, we refer the reader to [4,5,6,7,11,12,18].

Recall that a (*lattice*) *polynomial function* on L is any map $p \colon L^n \to L$ which can be obtained as a composition of the lattice operations $\wedge$ and $\vee$, the projections $\boldsymbol{x} \mapsto x_i$ and the constant functions $\boldsymbol{x} \mapsto c$, $c \in L$.

Fact 1. *Every polynomial function $p : L^n \to L$ is order-preserving and range-idempotent, that is, $p(c, \ldots, c) = c$, for every $c \in \mathrm{ran}(p)$.*

Polynomial functions are known to generalize certain prominent fuzzy integrals, namely, the so-called (*discrete*) *Sugeno integrals*. Indeed, as observed in [17], Sugeno integrals coincide exactly with those polynomial functions $q : L^n \to L$ which are *idempotent*, that is, satisfy $q(c, \ldots, c) = c$, for every $c \in L$, and in particular satisfy $\mathrm{ran}(q) = L$. We shall take this as our working definition of the Sugeno integral; for the original definition (as an integral with respect to a fuzzy measure) see, e.g., [13,19,20].

As shown by Goodstein [11], polynomial functions over bounded distributive lattices (in particular, over bounded chains) have very neat normal form representations. For $I \subseteq [n]$, let $\boldsymbol{e}_I$ be the *characteristic vector* of I, i.e., the n-tuple in L^n whose i-th component is 1 if $i \in I$, and 0 otherwise.

Proposition 2 (Goodstein [11]). *A function $p\colon L^n \to L$ is a polynomial function if and only if*

$$p(x_1,\dots,x_n) = \bigvee_{I\subseteq[n]} \big(p(\boldsymbol{e}_I) \wedge \bigwedge_{i\in I} x_i\big). \tag{2}$$

Furthermore, the function given by (2) is a Sugeno integral if and only if $p(\boldsymbol{e}_\emptyset) = 0$ and $p(\boldsymbol{e}_{[n]}) = 1$.

Remark 3. Observe that, by Proposition 2, every polynomial function $p\colon L^n \to L$ is uniquely determined by its restriction to $\{0,1\}^n$. Also, since every lattice polynomial function is order-preserving, we have that the coefficients in (2) are monotone increasing, i.e., $p(\boldsymbol{e}_I) \le p(\boldsymbol{e}_J)$ whenever $I \subseteq J$. Moreover, a function $f\colon \{0,1\}^n \to L$ can be extended to a polynomial function over L if and only if it is order-preserving.

Remark 4. It follows from Goodstein's theorem that every unary polynomial function is of the form

$$p(x) = s \vee (x \wedge t) = \operatorname{med}(s,x,t) = \begin{cases} s, & \text{if } x < s,\\ x, & \text{if } x \in [s,t],\\ t, & \text{if } t < x, \end{cases} \tag{3}$$

where $s = p(0), t = p(1)$. In other words, $p(x)$ is a truncated identity. Figure 1 shows the graph of this function in the case when L is the real unit interval $[0,1]$.

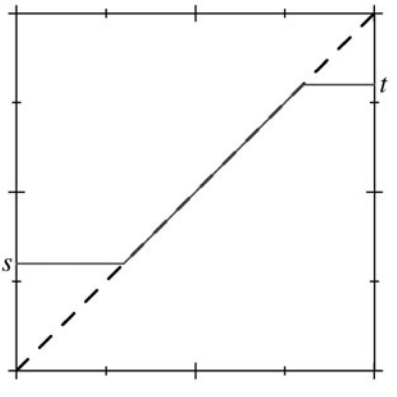


Fig. 1. A typical unary polynomial function

It is noteworthy that every polynomial function p as in (2) can be represented by $p = \langle q \rangle_p$ where q is the Sugeno integral given by

$$q(x_1,\dots,x_n) = \bigvee_{\emptyset\subsetneq I\subsetneq[n]} \big(p(\boldsymbol{e}_I) \wedge \bigwedge_{i\in I} x_i\big) \vee \bigwedge_{i\in[n]} x_i.$$

2.3 Characterizations of Polynomial Functions

The following results reassemble the various characterizations of polynomial functions obtained in [5]. For further background see, e.g., [6,7,13].

Theorem 5. *Let $p\colon L^n \to L$ be a function on an arbitrary bounded chain L. The following conditions are equivalent:*

(i) p is a polynomial function.
(ii) p is median decomposable, that is, for every $\boldsymbol{x} \in L^n$,

$$p(\boldsymbol{x}) = \operatorname{med}\left(p(\boldsymbol{x}_k^0), x_k, p(\boldsymbol{x}_k^1)\right) \qquad (k = 1, \ldots, n).$$

(iii) p is order-preserving, and $\operatorname{cl}(\operatorname{ran}(p))$-min and $\operatorname{cl}(\operatorname{ran}(p))$-max homogeneous, that is, for every $\boldsymbol{x} \in L^n$ and every $c \in \operatorname{cl}(\operatorname{ran}(p))$,

$$p(\boldsymbol{x} \wedge c) = p(\boldsymbol{x}) \wedge c \quad \text{and} \quad p(\boldsymbol{x} \vee c) = p(\boldsymbol{x}) \vee c, \quad \text{resp.}$$

(iv) p is order-preserving, range-idempotent, and horizontally minitive and maxitive, that is, for every $\boldsymbol{x} \in L^n$ and every $c \in L$,

$$p(\boldsymbol{x}) = p(\boldsymbol{x} \vee c) \wedge p([\boldsymbol{x}]^c) \quad \text{and} \quad p(\boldsymbol{x}) = p(\boldsymbol{x} \wedge c) \vee p([\boldsymbol{x}]_c), \quad \text{resp.}$$

Remark 6. Note that, by the equivalence $(i) \Leftrightarrow (iii)$, for every polynomial function $p\colon L^n \to L$, $p(\boldsymbol{x}) = \langle p(\boldsymbol{x})\rangle_p = p(\langle \boldsymbol{x}\rangle_p)$. Moreover, for every function $f\colon L^m \to L$ and every Sugeno integral $q\colon L^n \to L$, we have $\langle q(\boldsymbol{x})\rangle_f = q(\langle \boldsymbol{x}\rangle_f)$.

Theorem 5 is a refinement of the Main Theorem in [5] stated for functions over bounded distributive lattices. As shown in [7], in the case when L is a chain, Theorem 5 can be strengthened since the conditions need to be verified only on vectors of a certain prescribed type. Moreover, further characterizations are available and given in terms of conditions of somewhat different flavor, as the following theorem illustrates [7].

Theorem 7. *A function $p\colon L^n \to L$ is a polynomial function if and only if it is range-idempotent, and comonotonic minitive and maxitive, that is, for every permutation σ on $[n]$, and every $\boldsymbol{x}, \boldsymbol{x}' \in L^n_\sigma$,*

$$p(\boldsymbol{x} \wedge \boldsymbol{x}') = p(\boldsymbol{x}) \wedge p(\boldsymbol{x}') \quad \text{and} \quad p(\boldsymbol{x} \vee \boldsymbol{x}') = p(\boldsymbol{x}) \vee p(\boldsymbol{x}'), \quad \text{resp.}$$

3 Pseudo-Sugeno Integrals and Sugeno Utility Functions

In this section we study certain prominent function classes in the realm of multicriteria decision making. More precisely, we investigate overall utility functions $U\colon \prod_{i\in[n]} L_i \to L$ which can obtained by aggregating various local utility functions (i.e., order-preserving mappings) $\varphi_i\colon L_i \to L$, $i \in [n]$, using Sugeno integrals.

To this extent, in Subsect. 3.1 we introduce the wider class of pseudo-polynomial functions, and we present their axiomatization in Subsect. 3.2.

As we will see, pseudo-polynomial functions can be equivalently defined in terms of Sugeno integrals, and thus they model certain processes within multicriteria decision making. This is observed in Subsect. 3.3 where the notion of a Sugeno utility function $U\colon \prod_{i\in[n]} L_i \to L$ associated with given local utility functions $\varphi_i\colon L_i \to L$, $i \in [n]$, is discussed. Using the axiomatization of pseudo-polynomial functions, in Subsect. 3.4 we establish several characterizations of Sugeno utility functions based on Sugeno integrals given in terms of necessary and sufficient conditions which naturally extend those presented in Subsect. 2.3.

3.1 Pseudo-Sugeno Integrals and Pseudo-Polynomial Functions

Let L and $L_1, \dots, L_n$ be bounded chains. In the sequel, we shall denote the top and bottom elements of $L_1, \dots, L_n$ and L by 1 and 0, respectively. This convention will not give rise to ambiguities. We shall say that a mapping $\varphi_i\colon L_i \to L$, $i \in [n]$, satisfies the *boundary conditions* if for every $x \in L_i$,

$$\varphi(0) \le \varphi(x) \le \varphi(1) \quad \text{or} \quad \varphi(1) \le \varphi(x) \le \varphi(0).$$

Observe that if φ is order-preserving, then it satisfies the boundary conditions.

A function $f\colon \prod_{i\in[n]} L_i \to L$ is a *pseudo-polynomial function* if there is a polynomial function $p\colon L^n \to L$ and there are unary functions $\varphi_i\colon L_i \to L$, $i \in [n]$, satisfying the boundary conditions, such that

$$f(\boldsymbol{x}) = p(\varphi_1(x_1), \dots, \varphi_n(x_n)). \tag{4}$$

If p is a Sugeno integral, then we say that f is a *pseudo-Sugeno integral*. As the following result asserts, the notions of pseudo-polynomial function and pseudo-Sugeno integral turn out to be equivalent.

Proposition 8. *A function $f\colon \prod_{i\in[n]} L_i \to L$ is a pseudo-polynomial function if and only if it is a pseudo-Sugeno integral.*

Proof. Clearly, every pseudo-Sugeno integral is a pseudo-polynomial function. Conversely, if $f\colon \prod_{i\in[n]} L_i \to L$ of the form $f = p(\varphi_1(x_1), \dots, \varphi_n(x_n))$ for a lattice polynomial p, then by setting $\phi_i = \langle \varphi_i \rangle_p$ and taking q as a Sugeno integral such that $p = \langle q \rangle_p$, we have

$$\begin{aligned} f(\boldsymbol{x}) = \langle q(\varphi_1(x_1), \dots, \varphi_n(x_n)) \rangle_p &= q(\langle \varphi_1(x_1) \rangle_p, \dots, \langle \varphi_n(x_n) \rangle_p) \\ &= q(\phi_1(x_1), \dots, \phi_n(x_n)), \end{aligned}$$

and thus f is a pseudo-Sugeno integral. □

Remark 9. Clearly, $f(\boldsymbol{x}_k^0) \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_k^1)$ or $f(\boldsymbol{x}_k^1) \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_k^0)$ depending on whether $\varphi_k(0) \le \varphi_k(x) \le \varphi_k(1)$ or $\varphi_k(1) \le \varphi_k(x) \le \varphi_k(0)$, respectively.

3.2 A Characterization of Pseudo-Sugeno Integrals

Throughout this subsection, we assume that the unary maps $\varphi_i\colon L_i \to L$ considered, satisfy the boundary conditions.

We say that $f\colon \prod_{i\in[n]} L_i \to L$ is *pseudo-median decomposable* if for each $k \in [n]$ there is a unary function $\varphi_k\colon L_k \to L$ such that

$$f(\boldsymbol{x}) = \operatorname{med}\left(f(\boldsymbol{x}_k^0), \varphi_k(x_k), f(\boldsymbol{x}_k^1)\right) \tag{5}$$

for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$. Note that if f is pseudo-median decomposable w.r.t. unary functions $\varphi_i\colon L_i \to L$, $i \in [n]$, then for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and $k \in [n]$, we have $f(\boldsymbol{x}_k^0) \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_k^1)$ or $f(\boldsymbol{x}_k^1) \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_k^0)$.

Theorem 10. *Let $f\colon \prod_{i\in[n]} L_i \to L$ be a function. Then f is a pseudo-Sugeno integral if and only if f is pseudo-median decomposable.*

Proof. First we show that the condition is necessary. Suppose that $f\colon \prod_{i\in[n]} L_i \to L$ is of the form $f(\boldsymbol{x}) = q(\varphi_1(x_1), \ldots, \varphi_n(x_n))$ for some Sugeno integral q and unary functions φ_k satisfying the boundary conditions. Without loss of generality, assume that $k = 1$. So let us fix the values of $x_2, \ldots, x_n$, and let us consider the unary polynomial function $u(y) = q(y, \varphi_2(x_2), \ldots, \varphi_n(x_n))$.

Setting $a = \varphi_1(0), b = \varphi_1(1), y_1 = \varphi_1(x_1)$, the equality to prove takes the form $u(y_1) = \operatorname{med}(u(a), y_1, u(b))$. This becomes clear if we take into account that u is of the form (3), and by the boundary conditions either $a \le y_1 \le b$ or $b \le y_1 \le a$ (see also Fig. 1).

To verify that the condition is sufficient, just observe that applying (5) repeatedly to each variable of f we can straightforwardly obtain a representation of f as $f(\boldsymbol{x}) = p(\varphi_1(x_1), \ldots, \varphi_n(x_n))$ for some polynomial function p. Thus, f is a pseudo-polynomial function and, by Proposition 8, it is a pseudo-Sugeno integral. □

3.3 Motivation: Overall Utility Functions

Despite the theoretical interest, the motivation for the study of pseudo-Sugeno integrals (or, equivalently, pseudo-polynomial functions) is deeply rooted in multicriteria decision making. Let $\varphi_i\colon L_i \to L$, $i \in [n]$, be local utility functions (i.e., order-preserving mappings) having a common range $\mathcal{R} \subseteq L$, and let $M\colon L^n \to L$ be an aggregation function. The *overall utility function* associated with φ_i, $i \in [n]$, and M is the mapping $U\colon \prod_{i\in[n]} L_i \to L$ defined by

$$U(\boldsymbol{x}) = M(\varphi_1(x_1), \ldots, \varphi_n(x_n)). \tag{6}$$

For background on overall utility functions, see e.g. [2].

Thus, pseudo-Sugeno integrals subsume those overall utility functions (6) where the aggregation function M is a Sugeno integral. In the sequel we shall refer to a mapping $f\colon \prod_{i\in[n]} L_i \to L$ for which there are local utility functions

φ_i, $i \in [n]$, and a Sugeno integral (or, equivalently, a polynomial function) q, such that

$$f(\boldsymbol{x}) = q(\varphi_1(x_1), \ldots, \varphi_n(x_n)), \tag{7}$$

as a *Sugeno utility function*. As it will become clear in [9], these Sugeno utility functions coincide exactly with those pseudo-Sugeno integrals (or equivalently, pseudo-polynomial functions) which are order-preserving. Also, by taking $L_1 = \cdots = L_n = L$ and $\varphi_1 = \cdots = \varphi_n = \varphi$, it follows that Sugeno utility functions subsume the notions of quasi-Sugeno integral and quasi-polynomial function in the terminology of [8].

Remark 11. Note that the condition that $\varphi_i \colon L_i \to L$, $i \in [n]$ have a common range $\mathcal{R}$ is not really restrictive, since each φ_i can be extended to a local utility function $\varphi_i' \colon L_i' \to L$, where $L_i \subseteq L_i'$, in such a way that each φ_i', $i \in [n]$, has the same range $\mathcal{R} \subseteq L$. In fact, if $\mathcal{R}_i$ is the range of φ_i, for each $i \in [n]$, then $\mathcal{R}$ can be chosen as the interval

$$\mathrm{cl}(\bigcup_{i\in[n]} \mathcal{R}_i) = [\bigwedge_{i\in[n]} \varphi_i(0), \bigvee_{i\in[n]} \varphi_i(1)].$$

In this way, if $f' \colon \prod_{i\in[n]} L_i' \to L$ is such that $f'(\boldsymbol{x}) = q(\varphi_1'(x_1), \ldots, \varphi_n'(x_n))$, then the restriction of f' to $\prod_{i\in[n]} L_i$ is of the form $f(\boldsymbol{x}) = q(\varphi_1(x_1), \ldots, \varphi_n(x_n))$.

3.4 Characterizations of Sugeno Utility Functions

In view of the remark above, in this subsection we will assume that the local utility functions $\varphi_i \colon L_i \to L$, $i \in [n]$, considered have the same range $\mathcal{R} \subseteq L$. Since local utility functions satisfy the boundary conditions, from Theorem 10 we get the following characterization of Sugeno utility functions.

Corollary 12. *A function $f \colon \prod_{i\in[n]} L_i \to L$ is a Sugeno utility function if and only if it is pseudo-median decomposable w.r.t. local utility functions.*

We will provide further axiomatizations of Sugeno utility functions extending those of polynomial functions given in Subsect. 2.3 as well as those of quasi-polynomial functions given in [8]. For the sake of simplicity, given $\varphi_i \colon L_i \to L$, $i \in [n]$, we make use of the shorthand notation $\overline{\varphi}(\boldsymbol{x}) = (\varphi_1(x_1), \ldots, \varphi_n(x_n))$ and $\overline{\varphi}^{-1}(c) = \{\boldsymbol{d} : \overline{\varphi}(\boldsymbol{d}) = c\}$, for every $c \in \mathcal{R}$.

We say that a function $f \colon \prod_{i\in[n]} L_i \to L$ is *pseudo-max homogeneous* (resp. *pseudo-min homogeneous*) if there are local utility functions $\varphi_i \colon L_i \to L$, $i \in [n]$, such that for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and $c \in \mathcal{R}$,

$$f(\boldsymbol{x} \vee \boldsymbol{d}) = f(\boldsymbol{x}) \vee c \quad (\text{resp. } f(\boldsymbol{x} \wedge \boldsymbol{d}) = f(\boldsymbol{x}) \wedge c), \quad \text{whenever } \boldsymbol{d} \in \overline{\varphi}^{-1}(c). \tag{8}$$

Fact 13. *Let $f \colon \prod_{i\in[n]} L_i \to L$ be a function, and let $\varphi_i \colon L_i \to L$, $i \in [n]$, be local utility functions. If f is pseudo-min homogeneous and pseudo-max homogeneous w.r.t. $\varphi_1, \ldots, \varphi_n$, then it satisfies the condition*

$$\text{for every } c \in \mathcal{R} \text{ and } \boldsymbol{d} \in \overline{\varphi}^{-1}(c),\ f(\boldsymbol{d}) = c. \tag{9}$$

Lemma 14. *If $f(x_1,\dots,x_n) = q(\varphi(x_1),\dots,\varphi_n(x_n))$ for some Sugeno integral $q\colon L^n \to L$ and local utility functions $\varphi_1,\dots,\varphi_n$, then f is pseudo-min homogeneous and pseudo-max homogeneous w.r.t. $\varphi_1,\dots,\varphi_n$.*

Proof. Let $\mathcal{R}$ be the common range of $\varphi_1,\dots,\varphi_n$, let $c \in \mathcal{R}$ and $\boldsymbol{d} \in \overline{\varphi}^{-1}(c)$. By Theorem 5 and the fact that each φ_k is order-preserving, we have

$$\begin{aligned} f(\boldsymbol{x} \vee \boldsymbol{d}) &= q(\overline{\varphi}(\boldsymbol{x} \vee \boldsymbol{d})) = q(\overline{\varphi}(\boldsymbol{x}) \vee \overline{\varphi}(\boldsymbol{d})) \\ &= q(\overline{\varphi}(\boldsymbol{x}) \vee c) = q(\overline{\varphi}(\boldsymbol{x})) \vee c = f(\boldsymbol{x}) \vee c, \end{aligned}$$

and hence, f is pseudo-max homogeneous. The dual statement follows similarly. □

For $\boldsymbol{x}, \boldsymbol{d} \in \prod_{i\in[n]} L_i$, let $[\boldsymbol{x}]_{\boldsymbol{d}}$ be the n-tuple whose ith component is 0_{L_i}, if $x_i \le d_i$, and x_i, otherwise, and dually let $[\boldsymbol{x}]^{\boldsymbol{d}}$ be the n-tuple whose ith component is 1_{L_i}, if $x_i \ge d_i$, and x_i, otherwise. We say that $f\colon \prod_{i\in[n]} L_i \to L$ is *pseudo-horizontally maxitive* (resp. *pseudo-horizontally minitive*) if there are local utility functions $\varphi_i\colon L_i \to L$, $i \in [n]$, such that for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and $c \in \mathcal{R}$, if $\boldsymbol{d} \in \overline{\varphi}^{-1}(c)$, then

$$f(\boldsymbol{x}) = f(\boldsymbol{x} \wedge \boldsymbol{d}) \vee f([\boldsymbol{x}]_{\boldsymbol{d}}) \quad (\text{resp. } f(\boldsymbol{x}) = f(\boldsymbol{x} \vee \boldsymbol{d}) \wedge f([\boldsymbol{x}]^{\boldsymbol{d}})). \tag{10}$$

Lemma 15. *If $f\colon \prod_{i\in[n]} L_i \to L$ is order-preserving, pseudo-horizontally minitive (resp. pseudo-horizontally maxitive) and satisfies (9), then it is pseudo-min homogeneous (resp. pseudo-max homogeneous).*

Proof. If $f\colon \prod_{i\in[n]} L_i \to L$ is order-preserving, pseudo-horizontally minitive and satisfies (9) w.r.t. $\varphi_1,\dots,\varphi_n$, then for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$, $c \in \mathcal{R}$, $\boldsymbol{d} \in \overline{\varphi}^{-1}(c)$

$$\begin{aligned} f(\boldsymbol{x}) \wedge c = f(\boldsymbol{x}) \wedge f(\boldsymbol{d}) \;&\ge\; f(\boldsymbol{x} \wedge \boldsymbol{d}) \;=\; f((\boldsymbol{x} \wedge \boldsymbol{d}) \vee \boldsymbol{d}) \wedge f([\boldsymbol{x} \wedge \boldsymbol{d}]^{\boldsymbol{d}}) \\ = f(\boldsymbol{d}) \wedge f([\boldsymbol{x}]^{\boldsymbol{d}}) \;&\ge\; f(\boldsymbol{d}) \wedge f(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \wedge c. \end{aligned}$$

Hence f is pseudo-min homogeneous w.r.t. $\varphi_1,\dots,\varphi_n$. The dual statement can be proved similarly. □

Lemma 16. *Suppose that $f\colon \prod_{i\in[n]} L_i \to L$ is order-preserving and pseudo-min homogeneous (resp. pseudo-max homogeneous), and satisfies (9). Then f is pseudo-max homogeneous (resp. pseudo-min homogeneous) if and only if it is pseudo-horizontally maxitive (resp. pseudo-horizontally minitive).*

Proof. Suppose that $f\colon \prod_{i\in[n]} L_i \to L$ is order-preserving and pseudo-min homogeneous and satisfies (9) w.r.t. $\varphi_1,\dots,\varphi_n$. Assume first that f is pseudo-max homogeneous w.r.t. $\varphi_1,\dots,\varphi_n$. For every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and $\boldsymbol{d} \in \overline{\varphi}^{-1}(c)$, where $c \in \mathcal{R}$, we have

$$\begin{aligned} f(\boldsymbol{x} \wedge \boldsymbol{d}) \vee f([\boldsymbol{x}]_{\boldsymbol{d}}) &= \big(f(\boldsymbol{x}) \wedge c\big) \vee f([\boldsymbol{x}]_{\boldsymbol{d}}) = \big(f(\boldsymbol{x}) \vee f([\boldsymbol{x}]_{\boldsymbol{d}})\big) \wedge \big(c \vee f([\boldsymbol{x}]_{\boldsymbol{d}})\big) \\ &= f(\boldsymbol{x}) \wedge f(\boldsymbol{d} \vee [\boldsymbol{x}]_{\boldsymbol{d}}) = f(\boldsymbol{x}), \end{aligned}$$

and hence f is pseudo-horizontally maxitive w.r.t. $\varphi_1,\dots,\varphi_n$.

Conversely, if f is pseudo-horizontally maxitive w.r.t. $\varphi_1, \ldots, \varphi_n$, then by Lemma 15 f is pseudo-max homogeneous w.r.t. $\varphi_1, \ldots, \varphi_n$. The dual statement can be proved similarly. □

Lemma 17. *If $f\colon \prod_{i\in[n]} L_i \to L$ is order-preserving, pseudo-min homogeneous and pseudo-horizontally maxitive, then it is pseudo-median decomposable w.r.t. local utility functions.*

Proof. Let $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and let $k \in [n]$. If f is pseudo-horizontally maxitive, say w.r.t. $\varphi_1, \ldots, \varphi_n$, then $f(\boldsymbol{x}) = f(\boldsymbol{x} \wedge \boldsymbol{d}) \vee f([\boldsymbol{x}]_{\boldsymbol{d}})$, where the kth component of $\boldsymbol{d} \in \overline{\varphi}^{-1}(\varphi_k(x_k))$ is x_k. Now if f is pseudo-min homogeneous, then $f(\boldsymbol{x} \wedge \boldsymbol{d}) = f(\boldsymbol{x}^1_k \wedge \boldsymbol{d}) = f(\boldsymbol{x}^1_k) \wedge \varphi_k(x_k)$, and by the definition of $[\boldsymbol{x}]_{\boldsymbol{d}}$, we have $f([\boldsymbol{x}]_{\boldsymbol{d}}) \leq f(\boldsymbol{x}^0_k)$. Thus,

$$\begin{aligned} f(\boldsymbol{x}) &= \operatorname{med}\big(f(\boldsymbol{x}^0_k), f(\boldsymbol{x}), f(\boldsymbol{x}^1_k)\big) = \big(f(\boldsymbol{x}^0_k) \vee f(\boldsymbol{x})\big) \wedge f(\boldsymbol{x}^1_k) \\ &= \big(f(\boldsymbol{x}^0_k) \vee (f(\boldsymbol{x}^1_k) \wedge \varphi_k(x_k))\big) \wedge f(\boldsymbol{x}^1_k) = f(\boldsymbol{x}^0_k) \vee \big(f(\boldsymbol{x}^1_k) \wedge \varphi_k(x_k)\big) \\ &= \operatorname{med}\big(f(\boldsymbol{x}^0_k), \varphi_k(x_k), f(\boldsymbol{x}^1_k)\big). \end{aligned}$$

Since this holds for every $\boldsymbol{x} \in \prod_{i\in[n]} L_i$ and $k \in [n]$, f is pseudo-median decomposable. □

We can also extend the comonotonic properties as follows. We say that a function $f\colon \prod_{i\in[n]} L_i \to L$ is *pseudo-comonotonic minitive* (resp. *pseudo-comonotonic maxitive*) if there are local utility functions $\varphi_i\colon L_i \to L$, $i \in [n]$, such that for every permutation σ on $[n]$, and every $\boldsymbol{x}, \boldsymbol{x}'$ such that $\overline{\varphi}(\boldsymbol{x}), \overline{\varphi}(\boldsymbol{x}') \in L^n_\sigma$,

$$f(\boldsymbol{x} \wedge \boldsymbol{x}') = f(\boldsymbol{x}) \wedge f(\boldsymbol{x}') \quad (\text{resp. } f(\boldsymbol{x} \vee \boldsymbol{x}') = f(\boldsymbol{x}) \vee f(\boldsymbol{x}')).$$

The following fact is straightforward.

Fact 18. *Every Sugeno utility function of the form (7) is pseudo-comonotonic minitive and maxitive. Moreover, if a function is pseudo-comonotonic minitive (resp. pseudo-comonotonic maxitive) and satisfies (9), then it is pseudo-min homogeneous (resp. pseudo-max homogeneous).*

Let $\mathbf{P}$ be the set comprising the properties of pseudo-min homogeneity, pseudo-horizontal minitivity and pseudo-comonotic minitivity, and let $\mathbf{P}^d$ be the set comprising the corresponding dual properties. The following result generalizes the various characterizations of polynomial functions given in Subsect. 2.3.

Theorem 19. *Let $f\colon \prod_{i\in[n]} L_i \to L$ be an order-preserving function. The following assertions are equivalent:*

(i) f is a Sugeno utility function.
(ii) f is pseudo-median decomposable w.r.t. local utility functions.
(iii) f is $P_1 \in \mathbf{P}$ and $P_2 \in \mathbf{P}^d$, and satisfies (9).

Proof. By Corollary 12, we have $(i) \Leftrightarrow (ii)$. By Lemma 14, we also have that if (i) holds, then f is pseudo-min homogeneous and pseudo-max homogeneous. Furthermore, by Fact 18 and Lemmas 15, 16 and 17, we have that any two formulations of (iii) are equivalent. By Lemma 17, $(iii) \Rightarrow (ii)$. □

Remark 20. By Fact 13, if P_1 and P_2 are the pseudo-homogeneity properties, then (9) becomes redundant in (iii). Similarly, by Lemma 17, Corollary 12, and $(i) \Rightarrow (iii)$ of Theorem 19, if P_1 is pseudo-min homogeneity (pseudo-horizontal minitivity) property, and P_2 is pseudo-horizontal maxitivity (pseudo-max homogeneity) property, then (9) is redundant in (iii).

4 Concluding Remarks

Theorem 19 provides necessary and sufficient conditions for an order-preserving function $f\colon \prod_{i\in[n]} L_i \to L$ to be a Sugeno utility function, that is, to be factorized into a composition

$$f(x_1,\ldots,x_n) = q(\varphi_1(x_1),\ldots,\varphi_n(x_n)), \tag{11}$$

where $\varphi_i\colon L_i \to L$, $i \in [n]$, are local utility functions and q is a Sugeno integral. However, knowing that f is a Sugeno utility function, no clues are given on how to derive such a factorization (11). Thus, we are left with the following problem:

Problem. Given a Sugeno utility function $f\colon \prod_{i\in[n]} L_i \to L$, construct local utility functions $\varphi_k\colon L_k \to L$, $k \in [n]$, and a Sugeno integral q such that f fulfills (11).

This problem is considered and solved in the companion paper [9] also submitted to MDAI2010.

Acknowledgments. We would like to thank Jean-Luc Marichal for useful discussions and for bringing this topic to our attention. The second named author acknowledges that the present project is supported by the National Research Fund, Luxembourg, and cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND), and supported by the Hungarian National Foundation for Scientific Research under grant no. K77409.

References

1. Beliakov, G., Pradera, A., Calvo, T.: Aggregation Functions: A Guide for Practitioners. Studies in Fuzziness and Soft Computing, vol. 221. Springer, Berlin (2007)
2. Bouyssou, D., Dubois, D., Prade, H., Pirlot, M. (eds.): Decision-Making Process - Concepts and Methods. ISTE/John Wiley (2009)
3. Bouyssou, D., Marchant, T., Pirlot, M.: A Conjoint Measurement Approach to the Discrete Sugeno Integral. In: The Mathematics of Preference, Choice and Order, pp. 85–109. Springer, Berlin (2009)

4. Burris, S., Sankappanavar, H.P.: A Course in Universal Algebra. Graduate Texts in Mathematics, vol. 78. Springer, New York (1981)
5. Couceiro, M., Marichal, J.-L.: Polynomial Functions over Bounded Distributive Lattices. Journal of Multiple-Valued Logic and Soft Computing (to appear), http://arxiv.org/abs/0901.4888
6. Couceiro, M., Marichal, J.-L.: Characterizations of Discrete Sugeno Integrals as Polynomial Functions over Distributive Lattices. Fuzzy Sets and Systems 161(5), 694–707 (2010)
7. Couceiro, M., Marichal, J.-L.: Representations and Characterizations of Polynomial Functions on Chains. Journal of Multiple-Valued Logic and Soft Computing 16(1-2), 65–86 (2010)
8. Couceiro, M., Marichal, J.-L.: Axiomatizations of Quasi-Polynomial Functions on Bounded Chains. Aequationes Mathematicae 396(1), 195–213 (2009)
9. Couceiro, M., Waldhauser, T.: Sugeno Utility Functions II: Factorizations. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) MDAI 2010. LNCS (LNAI), vol. 6408, pp. 91–103. Springer, Heidelberg (2010)
10. Dubois, D., Marichal, J.-L., Prade, H., Roubens, M., Sabbadin, R.: The Use of the Discrete Sugeno Integral in Decision-Making: a Survey. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 9(5), 539–561 (2001)
11. Goodstein, R.L.: The Solution of Equations in a Lattice. Proc. Roy. Soc. Edinburgh Sect. A 67, 231–242 (1965/1967)
12. Grätzer, G.: General Lattice Theory. Birkhäuser, Berlin (2003)
13. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation Functions. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge (2009)
14. Grabisch, M., Murofushi, T., Sugeno, M. (eds.): Fuzzy Measures and Integrals - Theory and Applications. Studies in Fuzziness and Soft Computing, vol. 40. Physica-Verlag, Heidelberg (2000)
15. Kuczma, M., Choczewski, B., Ger, R.: Iterative Functional Equations. Cambridge University Press, Cambridge (1990)
16. Marichal, J.-L.: An Axiomatic Approach of the Discrete Sugeno Integral as a Tool to Aggregate Interacting Criteria in a Qualitative Framework. IEEE Trans. Fuzzy Syst. 9(1), 164–172 (2001)
17. Marichal, J.-L.: Weighted Lattice Polynomials. Discrete Mathematics 309(4), 814–820 (2009)
18. Rudeanu, S.: Lattice Functions and Equations. Springer Series in Discrete Mathematics and Theoretical Computer Science. Springer, London (2001)
19. Sugeno, M.: Theory of Fuzzy Integrals and its Applications. PhD thesis, Tokyo Institute of Technology, Tokyo (1974)
20. Sugeno, M.: Fuzzy Measures and Fuzzy Integrals–a Survey. In: Gupta, M.M., Saridis, G.N., Gaines, B.R. (eds.) Fuzzy Automata and Decision Processes, pp. 89–102. North-Holland, New York (1977)

Sugeno Utility Functions II: Factorizations

Miguel Couceiro[1] and Tamás Waldhauser[1,2]

[1] University of Luxembourg
6, rue Richard Coudenhove-Kalergi, L-1359 Luxembourg
miguel.couceiro@uni.lu
[2] Bolyai Institute, University of Szeged
Aradi vértanúk tere 1, H-6720 Szeged, Hungary
twaldha@math.u-szeged.hu

Abstract. In this paper we address and solve the problem posed in the companion paper [3] of factorizing an overall utility function as a composition $q(\varphi_1(x_1), \ldots, \varphi_n(x_n))$ of a Sugeno integral q with local utility functions φ_i, if such a factorization exists.

Keywords: Sugeno integral, local utility function, overall utility function, Sugeno utility function, factorization.

1 Introduction

In the companion paper [3], we considered a multicriteria aggregation model where local utility functions (i.e., order-preserving mappings) $\varphi_i \colon L_i \to L$, $i = 1, \ldots, n$, are aggregated using a (discrete) Sugeno integral $q \colon L^n \to L$, thus giving rise to an overall utility function $f \colon L_1 \times \cdots \times L_n \to L$ defined by

$$f(x_1, \ldots, x_n) = q(\varphi_1(x_1), \ldots, \varphi_n(x_n)). \tag{1}$$

Such functions were called Sugeno utility functions in [3]. More general classes of functions were also considered, where the inner functions φ_i are not necessarily order-preserving, and where the outer function q is either a Sugeno integral or a (lattice) polynomial function. The resulting functions were referred to as pseudo-Sugeno integrals and pseudo-polynomial functions, respectively.

This aggregation model is deeply rooted in multicriteria decision making, where the variables x_i represent different properties of the alternatives (e.g., price, speed, safety, comfort level of a car), and the overall utility function (also called global preference function) assigns a score to the alternatives that helps the decision maker to choose the best one (e.g., to choose the car to buy). A similar situation is that of subjective evaluation (see [1]): f outputs the overall rating of a certain product by customers, and the variables x_i represent the various properties of that product. The way in which these properties influence the overall rating can give information about the attitude of the customers. A factorization of the (empirically) given overall utility function f in the form (1) can be used for such an analysis; this is our main motivation for addressing this problem.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 91–103, 2010.

In [3] we established necessary and sufficient conditions which guarantee the existence of factorizations of functions $f\colon L_1 \times \cdots \times L_n \to L$ as compositions (1). However, no hint was given on how to obtain such factorizations. In this paper, we address and solve this problem by providing a canonical construction of such a Sugeno integral q and utility functions φ_i so that their composition satisfies $f = q(\varphi_1, \ldots, \varphi_n)$.

The paper is organized as follows. In Sect. 2 we recall the background on lattice polynomial functions, Sugeno integrals and Sugeno utility functions needed throughout the paper (for further background and references see the companion paper [3]). In Sect. 3 we present a method to construct a factorization of a Sugeno utility function f, which is illustrated in Subsect. 3.3 by means of a concrete example. Finally, in Sect. 4 we prove the correctness of the procedure by showing that the Sugeno integral and the local utility functions constructed in Sect. 3 indeed give a factorization of f.

2 Preliminaries

2.1 Lattice Polynomials and Sugeno Integrals

Let L be a chain endowed with the lattice operations $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Clearly, L is a distributive lattice. A chain L is *complete* if every nonempty subset S of L has a greatest lower bound (infimum) denoted by $\bigwedge S$, and a least upper bound (supremum) denoted by $\bigvee S$. A chain is *bounded*, if it has least and greatest elements, usually denoted by 0_L and 1_L, respectively, or simply by 0 and 1, when there is no risk of ambiguity. Observe that if L is complete, then it is bounded. In most applications the chains considered are either closed real intervals or finite chains, and these are all complete. Hence throughout the paper, $L_1, \ldots, L_n$ and L will always denote complete chains.

An n-ary (*lattice*) *polynomial function* on L is a function $p\colon L^n \to L$ that can be built from projections $(x_1, \ldots, x_n) \mapsto x_i$ and constants by a finite number of applications of the lattice operations $\wedge, \vee$ (for a recent reference, see [2]). The notion of polynomial functions subsumes certain important fuzzy integrals, namely, Sugeno integrals. As it was observed in [8,9], (*discrete*) *Sugeno integrals* can be defined as certain polynomial functions, namely, those polynomial functions $q\colon L^n \to L$ satisfying $q(a, \ldots, a) = a$ for all $a \in L$. We will work with this definition of the Sugeno integral; for the original definition (as an integral with respect to a fuzzy measure) see, e.g., [7,10,11].

Polynomial functions have a neat disjunctive normal form representation, as shown by the following theorem of Goodstein [5]. Let $[n] = \{1, \ldots, n\}$, and for $I \subseteq [n]$ let $\mathbf{e}_I \in L^n$ be the characteristic vector of I, i.e., the vector whose ith component is 1 if $i \in I$ and 0 if $i \notin I$.

Theorem 1. *A function $p\colon L^n \to L$ is a polynomial function if and only if*

$$p(x_1, \ldots, x_n) = \bigvee_{I \subseteq [n]} \big(p(\mathbf{e}_I) \wedge \bigwedge_{i \in I} x_i\big).$$

Such a function is a Sugeno integral iff $p(\mathbf{e}_\emptyset) = 0$ and $p(\mathbf{e}_{[n]}) = 1$.

In the sequel we will make use of the following property of polynomial functions [2].

Proposition 2. *For every polynomial function* $p\colon L^n \to L$ *and* $k \in [n]$ *we have*

$$p\big(x_1, \ldots, x_{k-1}, p\,(x_1, \ldots, x_n)\,, x_{k+1}, \ldots, x_n\big) = p\,(x_1, \ldots, x_n)\,.$$

An important polynomial function is the *median function* $\mathrm{med}\colon L^3 \to L$ defined by $\mathrm{med}\,(x, y, z) = (x \wedge y) \vee (x \wedge z) \vee (y \wedge z)$. If a, b, c are pairwise different, then $\mathrm{med}\,(a, b, c)$ is the middle one of these three elements (w.r.t. the ordering of L), while if there is a repetition among a, b, c, then $\mathrm{med}\,(a, b, c)$ equals this repeated value.

2.2 Sugeno Utility Functions

By a *Sugeno utility function* we mean a function $f\colon L_1 \times \cdots \times L_n \to L$ of the form

$$f\,(x_1, \ldots, x_n) = q\,(\varphi_1\,(x_1)\,, \ldots, \varphi_n\,(x_n))\,, \tag{2}$$

where $q\colon L^n \to L$ is a Sugeno integral, and each $\varphi_i\colon L_i \to L$ is an order-preserving function, so-called *local utility function*. Such functions can model various situations where one needs to aggregate several inputs into a single output in a meaningful way. The local utility functions φ_i map the various inputs x_i (which are measured on possibly different scales L_i) into a single scale L, and then the aggregation function q, in this case a Sugeno integral, combines them into a single value. For general background see [1,4,6,7].

A function $\varphi_i\colon L_i \to L$ satisfies the *boundary conditions* if $\varphi_i\,(x_i)$ lies between $\varphi_i\,(0_{L_i})$ and $\varphi_i\,(1_{L_i})$ for all $x_i \in L_i$. *Pseudo-Sugeno integrals* were defined in [3] as functions f of the form (2), where q is a Sugeno integral, and the inner functions φ_i satisfy the boundary conditions. Order-preserving functions clearly satisfy the boundary conditions, hence the class of pseudo-Sugeno integrals subsumes the class of Sugeno utility functions. We will see in Sect. 5 that Sugeno utility functions coincide with order-preserving pseudo-Sugeno integrals.

A fundamental tool in our study of Sugeno utility functions is the following *pseudo-median decomposition* formula. This result was stated and proved in [3] in a stronger form, where the pseudo-median decomposition formula was shown to characterize the wider class of pseudo-Sugeno integrals.

Theorem 3. *If* $f\colon L_1 \times \cdots \times L_n \to L$ *is a Sugeno utility function as in* (2)*, then for all* $k \in [n]$ *and* $\mathbf{x} \in L_1 \times \cdots \times L_n$ *we have*

$$f\,(\mathbf{x}) = \mathrm{med}\left(f\left(\mathbf{x}_k^0\right), \varphi_k\,(x_k)\,, f\left(\mathbf{x}_k^1\right)\right), \tag{3}$$

where $\mathbf{x}_k^0$ *(resp.* $\mathbf{x}_k^1$*) is the vector obtained from* $\mathbf{x}$ *by replacing its* k*th component by* 0_{L_k} *(resp.* 1_{L_k}*).*

Let us give a rough idea of how we can use the pseudo-median decomposition to extract from the global function f information about the local utility functions

φ_k. The key observation is that if $f\left(\mathbf{x}_k^0\right) < f(\mathbf{x}) < f\left(\mathbf{x}_k^1\right)$, then (3) implies that $\varphi_k(x_k) = f(\mathbf{x})$. So let us imagine that we fix all but the kth component of $\mathbf{x}$, and we continuously increase x_k from 0 to 1 in L_k. Let a (resp. b) be the first (resp. last) value of x_k where $f(\mathbf{x}) > f\left(\mathbf{x}_k^0\right)$ (resp. $f(\mathbf{x}) < f\left(\mathbf{x}_k^1\right)$). Then $f(\mathbf{x})$, viewed as a unary function of x_k, consists of three pieces: it is constant $s = f\left(\mathbf{x}_k^0\right)$ from 0 to a, coincides with φ_k from a to b, and is constant $t = f\left(\mathbf{x}_k^1\right)$ from b to 1 (see Fig. 1, where L_k and L are chosen to be the unit interval $[0,1] \subseteq \mathbb{R}$). Thus we can see some part of φ_k through the "window" $[a,b]$. Fixing the components of $\mathbf{x}$ (other than x_k) to some other values, we may open other windows, which may expose other parts of φ_k. If we could find sufficiently many windows, then we could recover φ_k but, unfortunately, this is not always the case. (In fact, as we shall see in the example of Subsect. 3.3, the local utility functions are not always uniquely determined by f.) In Sect. 3 we will develop this idea to find a candidate for φ_k, and we will show in Sect. 4 that the candidate that we construct is indeed appropriate in the sense that it can be used in factorizing a given Sugeno utility function.

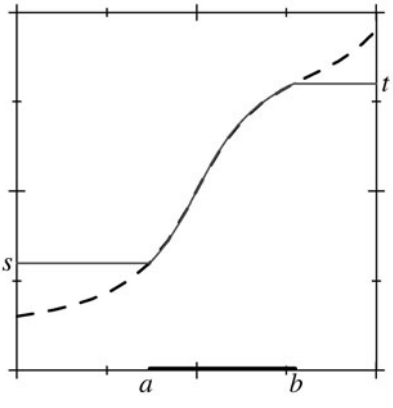


Fig. 1. The graph of φ_k as seen through a window

3 The Construction

Throughout this section let $f\colon L_1 \times \cdots \times L_n \to L$ be a Sugeno utiliy function. Knowing that f can be factorized as in (2), we will show how to construct in a canonical way a possibly different Sugeno integral q^f and local utility functions φ_i^f such that $f = q^f\left(\varphi_1^f, \ldots, \varphi_n^f\right)$. It is important that q^f and φ_i^f can be computed only from f, without having any information about q and φ_i (which are assumed to exist). In the first subsection we construct the Sugeno integral q^f, and then we describe the procedure to find suitable local utility functions φ_i^f. The latter is substantially more involved, therefore we conclude this section with a concrete example, and we defer the proof of correctness of the procedure to Sect. 4.

We will assume in the sequel that f depends on all of its variables. If this is not the case, e.g., f does not depend on its first variable, then there is a Sugeno utility function $g\colon L_2 \times \cdots \times L_n \to L$ such that $f(x_1, \ldots, x_n) = g(x_2, \ldots, x_n)$. In this case one could consider the function g instead of f, and find a factorization for this function. (If g still has inessential variables, then we can eliminate them in a similar way.)

3.1 Constructing the Sugeno Integral

The following result, which is essentially a generalization of Theorem 1, provides an appropriate Sugeno integral in order to factorize a Sugeno utility function.

Theorem 4. *If $f(x_1, \ldots, x_n) = q(\varphi_1(x_1), \ldots, \varphi_n(x_n))$ is a Sugeno utility function, then $f(x_1, \ldots, x_n) = q^f(\varphi_1(x_1), \ldots, \varphi_n(x_n))$, where $q^f\colon L^n \to L$ is the polynomial function given by*

$$q^f(y_1, \ldots, y_n) = \bigvee_{I \subseteq [n]} \left(f(\mathbf{e}_I) \wedge \bigwedge_{i \in I} y_i \right).$$

Proof. We need to prove that the following identity holds:

$$f(x_1, \ldots, x_n) = \bigvee_{I \subseteq [n]} \left(f(\mathbf{e}_I) \wedge \bigwedge_{i \in I} \varphi_i(x_i) \right). \tag{4}$$

We apply induction on n. If $n = 1$, then the right hand side of (4) takes the form $f(0) \vee (f(1) \wedge \varphi_1(x_1)) = \operatorname{med}(f(0), \varphi_1(x_1), f(1))$, which equals $f(x_1)$ by (3). Now suppose that the statement of the theorem is true for all Sugeno utility functions in $n-1$ variables. Applying the pseudo-median decomposition to f with $k = n$ we obtain

$$\begin{aligned} f(x_1, \ldots, x_n) &= \operatorname{med}(f_0(x_1, \ldots, x_{n-1}), \varphi_n(x_n), f_1(x_1, \ldots, x_{n-1})) \\ &= f_0(x_1, \ldots, x_{n-1}) \vee (f_1(x_1, \ldots, x_{n-1}) \wedge \varphi_n(x_n)), \end{aligned} \tag{5}$$

where f_0 and f_1 are the $(n-1)$-ary Sugeno utility functions defined by

$$\begin{aligned} f_0(x_1, \ldots, x_{n-1}) &= f(x_1, \ldots, x_{n-1}, 0), \\ f_1(x_1, \ldots, x_{n-1}) &= f(x_1, \ldots, x_{n-1}, 1). \end{aligned}$$

Let us apply the induction hypothesis for these functions:

$$f_0(x_1, \ldots, x_{n-1}) = \bigvee_{I \subseteq [n-1]} \left(f_0(\mathbf{e}_I) \wedge \bigwedge_{i \in I} \varphi_i(x_i) \right) = \bigvee_{I \subseteq [n-1]} \left(f(\mathbf{e}_I) \wedge \bigwedge_{i \in I} \varphi_i(x_i) \right),$$

$$f_1(x_1, \ldots, x_{n-1}) = \bigvee_{I \subseteq [n-1]} \left(f_1(\mathbf{e}_I) \wedge \bigwedge_{i \in I} \varphi_i(x_i) \right) = \bigvee_{I \subseteq [n-1]} \left(f(\mathbf{e}_{I \cup \{n\}}) \wedge \bigwedge_{i \in I} \varphi_i(x_i) \right).$$

Substituting back into (5) and using distributivity we obtain the desired equality (4). □

The polynomial q^f given in the above theorem is a Sugeno integral if and only if $f(0, \ldots, 0) = 0$ and $f(1, \ldots, 1) = 1$. It is natural to assume that the latter holds, since otherwise the parts of L that lie outside the interval $[f(0, \ldots, 0), f(1, \ldots, 1)]$ are "useless"; we may remove them without changing anything in the problem.

3.2 Constructing the Local Utility Functions

We only present the construction of φ_1^f; the other local utility functions can be constructed similarly. For any $x_1 \in L_1$ we partition $L_2 \times \cdots \times L_n$ into the following four disjoint sets:

$$\begin{aligned}
\mathcal{W}_{x_1} &= \{(x_2,\dots,x_n) : f(0,x_2,\dots,x_n) < f(x_1,x_2,\dots,x_n) < f(1,x_2,\dots,x_n)\},\\
\mathcal{L}_{x_1} &= \{(x_2,\dots,x_n) : f(0,x_2,\dots,x_n) < f(x_1,x_2,\dots,x_n) = f(1,x_2,\dots,x_n)\},\\
\mathcal{U}_{x_1} &= \{(x_2,\dots,x_n) : f(0,x_2,\dots,x_n) = f(x_1,x_2,\dots,x_n) < f(1,x_2,\dots,x_n)\},\\
\mathcal{E}_{x_1} &= \{(x_2,\dots,x_n) : f(0,x_2,\dots,x_n) = f(x_1,x_2,\dots,x_n) = f(1,x_2,\dots,x_n)\}.
\end{aligned}$$

Observe that $\mathcal{E}_{x_1}$ bears no information on x_1; we only introduce it for notational convenience.

From the pseudo-median decomposition formula (3) we know that

$$f(x_1,x_2,\dots,x_n) = \operatorname{med}(f(0,x_2,\dots,x_n), \varphi_1(x_1), f(1,x_2,\dots,x_n)).$$

Examining this formula, we immediately get the following implications for all $x_1 \in L_1$ and $(x_2,\dots,x_n) \in L_2 \times \cdots \times L_n$:

$$\begin{aligned}
(x_2,\dots,x_n) \in \mathcal{W}_{x_1} &\implies \varphi_1(x_1) = f(x_1,x_2,\dots,x_n),\\
(x_2,\dots,x_n) \in \mathcal{L}_{x_1} &\implies \varphi_1(x_1) \geq f(x_1,x_2,\dots,x_n),\\
(x_2,\dots,x_n) \in \mathcal{U}_{x_1} &\implies \varphi_1(x_1) \leq f(x_1,x_2,\dots,x_n).
\end{aligned}$$

Thus, if $\mathcal{W}_{x_1}$ is not empty, then we can see $\varphi_1(x_1)$ through a window, and we can determine its exact value. Furthermore, $\mathcal{L}_{x_1}$ and $\mathcal{U}_{x_1}$ provide lower and upper bounds, respectively, whenever they are not empty. We introduce the following notation for these values:

$$\varphi_1(x_1) = w_{x_1} = f(x_1,x_2,\dots,x_n) \text{ if } (x_2,\dots,x_n) \in \mathcal{W}_{x_1}, \tag{6}$$

$$\varphi_1(x_1) \geq l_{x_1} = \bigvee_{(x_2,\dots,x_n)\in\mathcal{L}_{x_1}} f(x_1,x_2,\dots,x_n) \text{ if } \mathcal{L}_{x_1} \neq \emptyset, \tag{7}$$

$$\varphi_1(x_1) \leq u_{x_1} = \bigwedge_{(x_2,\dots,x_n)\in\mathcal{U}_{x_1}} f(x_1,x_2,\dots,x_n) \text{ if } \mathcal{U}_{x_1} \neq \emptyset. \tag{8}$$

If any of the sets $\mathcal{W}_{x_1}, \mathcal{L}_{x_1}, \mathcal{U}_{x_1}$ is empty, then the corresponding values $w_{x_1}, l_{x_1}, u_{x_1}$ are undefined.

Now we are able to define a function $\varphi_1^f : L_1 \to L$ that will serve as a replacement of φ_1:

(W) if $\mathcal{W}_{x_1} \neq \emptyset$ then let $\varphi_1^f(x_1) = w_{x_1}$;
(L) if $\mathcal{W}_{x_1} = \emptyset, \mathcal{L}_{x_1} \neq \emptyset, \mathcal{U}_{x_1} = \emptyset$ then let $\varphi_1^f(x_1) = l_{x_1}$;
(U) if $\mathcal{W}_{x_1} = \emptyset, \mathcal{L}_{x_1} = \emptyset, \mathcal{U}_{x_1} \neq \emptyset$ then let $\varphi_1^f(x_1) = u_{x_1}$;
(LU) if $\mathcal{W}_{x_1} = \emptyset, \mathcal{L}_{x_1} \neq \emptyset, \mathcal{U}_{x_1} \neq \emptyset$ then let $\varphi_1^f(x_1) = l_{x_1}$.

It is important to note that φ_1^f is computed only from f, without reference to φ_1. Let us also observe that the four cases above cover all possibilities since $\mathcal{W}_{x_1} = \mathcal{U}_{x_1} = \mathcal{L}_{x_1} = \emptyset$ is ruled out by the assumption that f depends on its first variable. In the case (LU) we could have chosen any element from the interval $[l_{x_1}, u_{x_1}]$ (see Remark 6); we chose l_{x_1} just to make the construction canonical. We will also prove in Lemma 8 that φ_1^f is indeed a good candidate in the sense that $f = q\big(\varphi_1^f, \varphi_2, \ldots, \varphi_n\big)$.

3.3 An Example

Let us illustrate our construction with a concrete (albeit fictitious) example. Customers evaluate hotels along three criteria, namely quality of services, price, and whether the hotel has a good location. Service is evaluated on a four-level scale L_1: `*<**<***<****`, price is evaluated on a three-level scale L_2: `-<0<+` (where "`-`" means expensive, thus less desirable, and "`+`" means cheap, thus more desirable), and the third scale is L_3: `n(o)<y(es)`. In addition, each hotel receives an overall rating on the scale $L : 1 < \cdots < 8$, which gives the overall utility function $f\colon L_1 \times L_2 \times L_3 \to L$ (see Table 1(a)). We will find a factorization of this function, and we will analyse its structure in order to draw conclusions about the nature of the "human aggregation" that the customers (unconsciously) perform when forming their opinions about hotels. First we apply Theorem 4 to find the underlying Sugeno integral:

$$q^f(y_1, y_2, y_3) = 1 \vee (2 \wedge y_1) \vee (2 \wedge y_2) \vee (3 \wedge y_3) \\ \vee (2 \wedge y_1 \wedge y_2) \vee (8 \wedge y_1 \wedge y_3) \vee (6 \wedge y_2 \wedge y_3) \vee (8 \wedge y_1 \wedge y_2 \wedge y_3)\,.$$

Since 1 (resp. 8) is the least (resp. greatest) element of L, this polynomial function q^f is indeed a Sugeno integral. We can simplify q^f by cancelling those terms which are absorbed by some other terms in the disjunction:

$$q^f(y_1, y_2, y_3) = (2 \wedge y_1) \vee (2 \wedge y_2) \vee (3 \wedge y_3) \vee (y_1 \wedge y_3) \vee (6 \wedge y_2 \wedge y_3)\,.$$

We will be able to perform further simplifications after constructing the local utility functions. Table 1(b) shows the partitions of $L_2 \times L_3$ corresponding to the four possible elements $x_1 \in L_1$. The numbers in parentheses are the values of $f(x_1, x_2, x_3)$ (recall that we do not compute any values for the sets $\mathcal{E}_{x_1}$); these are used to compute the numbers $l_{x_1}, w_{x_1}, u_{x_1}$ shown in Table 1(c). This table contains these data for all $x_2 \in L_2$ and $x_3 \in L_3$ as well, together with the values of $\varphi_1^f(x_1), \varphi_2^f(x_2), \varphi_3^f(x_3)$.

Now that we know that the greatest value of φ_2^f is 6, we can simplify the Sugeno integral q^f by replacing $6 \wedge y_2 \wedge y_3$ with $y_2 \wedge y_3$, and "factoring out" $y_1 \vee y_2$:

$$(3 \wedge y_3) \vee ((y_1 \vee y_2) \wedge (2 \vee y_3)) = \operatorname{med}(3 \wedge y_3, y_1 \vee y_2, 2 \vee y_3)\,.$$

Note that this polynomial function is different from q^f, but it gives the same overall utility function f. This example shows that the Sugeno integral is not

Table 1. The hotel example

(a) The overall utility function

service	price	location	f
*	-	n	1
**	-	n	2
***	-	n	2
****	-	n	2
*	0	n	2
**	0	n	2
***	0	n	2
****	0	n	2
*	+	n	2
**	+	n	2
***	+	n	2
****	+	n	2
*	-	y	3
**	-	y	3
***	-	y	7
****	-	y	8
*	0	y	5
**	0	y	5
***	0	y	7
****	0	y	8
*	+	y	6
**	+	y	6
***	+	y	7
****	+	y	8

(b) The partitions of $L_2 \times L_3$

	*	**	***	****
(-,n)	$\mathcal{U}_{*}\,(1)$	$\mathcal{L}_{**}\,(2)$	$\mathcal{L}_{***}\,(2)$	$\mathcal{L}_{****}\,(2)$
(0,n)	$\mathcal{E}_{*}$	$\mathcal{E}_{**}$	$\mathcal{E}_{***}$	$\mathcal{E}_{****}$
(+,n)	$\mathcal{E}_{*}$	$\mathcal{E}_{**}$	$\mathcal{E}_{***}$	$\mathcal{E}_{****}$
(-,y)	$\mathcal{U}_{*}\,(3)$	$\mathcal{U}_{**}\,(3)$	$\mathcal{W}_{***}\,(7)$	$\mathcal{L}_{****}\,(8)$
(0,y)	$\mathcal{U}_{*}\,(5)$	$\mathcal{U}_{**}\,(5)$	$\mathcal{W}_{***}\,(7)$	$\mathcal{L}_{****}\,(8)$
(+,y)	$\mathcal{U}_{*}\,(6)$	$\mathcal{U}_{**}\,(6)$	$\mathcal{W}_{***}\,(7)$	$\mathcal{L}_{****}\,(8)$

(c) The local utility functions

	l	w	u	φ_1^f
*			1	1
**	2		3	2
***	2	7		7
****	8			8

	l	w	u	φ_2^f
-			1	1
0		5		5
+	6			6

	l	w	u	φ_3^f
n			1	1
y	8			8

uniquely determined by f, and neither are the local utility functions (e.g., we could have chosen $\varphi_1^f(**) = 3$ according to Remark 6).

To better understand the behaviour of f, let us separate two cases upon the location of the hotel:

$$f(x_1,x_2,x_3) = \operatorname{med}\big(3 \wedge \varphi_3^f(x_3), \varphi_1^f(x_1) \vee \varphi_2^f(x_2), 2 \vee \varphi_3^f(x_3)\big) \tag{9}$$
$$= \begin{cases} \varphi_1^f(x_1) \vee \varphi_2^f(x_2) \vee 3, & \text{if } x_3 = \mathtt{y}, \\ \big(\varphi_1^f(x_1) \vee \varphi_2^f(x_2)\big) \wedge 2, & \text{if } x_3 = \mathtt{n}. \end{cases}$$

We can see from (9) that once x_3 is fixed, what matters is the higher one of $\varphi_1^f(x_1)$ and $\varphi_2^f(x_2)$. Thus, instead of aiming at an average level in both, a better strategy would be to maximize one of them. Moreover, φ_1^f either outputs

very low or very high scores, whereas φ_2^f is almost maximized once the price is not very bad. Hence it seems more reasonable to focus on service rather than on price. The third variable can radically change the final outcome, but little can be done to improve the location of the hotel.

4 Proof of Correctness

In this section we show that the construction described in the previous section indeed provides a factorization of the Sugeno utility function f. First we prove that the functions φ_i^f are local utility functions, i.e., order-preserving functions. As before, we only consider the case $i = 1$; the other cases can be treated in an analogous way.

Theorem 5. *For any Sugeno utility function f, the function φ_1^f defined by the rules* (W),(L),(U),(LU) *in Sect. 3 is order-preserving.*

Proof. We fix $a \leq b \in L_1$ and show that $\varphi_1^f(a) \leq \varphi_1^f(b)$. First let us assume that $\mathcal{W}_a \neq \emptyset$, and let us fix an arbitrary $(x_2, \ldots, x_n) \in \mathcal{W}_a$. Then $\varphi_1^f(a) = w_a = f(a, x_2, \ldots, x_n)$, and since f is order-preserving, by the definition of $\mathcal{W}_a$ we get

$$f(0, x_2, \ldots, x_n) < f(a, x_2, \ldots, x_n) \leq f(b, x_2, \ldots, x_n) \leq f(1, x_2, \ldots, x_n).$$

If $f(b, x_2, \ldots, x_n) < f(1, x_2, \ldots, x_n)$ then $(x_2, \ldots, x_n) \in \mathcal{W}_b$, hence, by (6), $\varphi_1^f(b) = w_b = f(b, x_2, \ldots, x_n)$. If $f(b, x_2, \ldots, x_n) = f(1, x_2, \ldots, x_n)$, then $(x_2, \ldots, x_n) \in \mathcal{L}_b$, therefore $\varphi_1^f(b) \geq l_b \geq f(b, x_2, \ldots, x_n)$ by (7). In both cases we obtain that

$$\varphi_1^f(a) = w_a = f(a, x_2, \ldots, x_n) \leq f(b, x_2, \ldots, x_n) \leq \varphi_1^f(b),$$

since f is order-preserving.

The case $\mathcal{W}_b \neq \emptyset$ can be dealt with similarly. So let us consider the remaining case $\mathcal{W}_a = \mathcal{W}_b = \emptyset$. Then

$$\mathcal{L}_a \cup \mathcal{U}_a = L_2 \times \cdots \times L_n \setminus \mathcal{E}_a = L_2 \times \cdots \times L_n \setminus \mathcal{E}_b = \mathcal{L}_b \cup \mathcal{U}_b.$$

Futhermore, from $a \leq b$ we can conclude that $\mathcal{L}_a \subseteq \mathcal{L}_b$ and $\mathcal{U}_a \supseteq \mathcal{U}_b$ by making use of the fact that f is order-preserving. This implies that either $\mathcal{L}_a \subset \mathcal{L}_b$ and $\mathcal{U}_a \supset \mathcal{U}_b$, or $\mathcal{L}_a = \mathcal{L}_b$ and $\mathcal{U}_a = \mathcal{U}_b$. In the first case, choosing an arbitrary $(x_2, \ldots, x_n) \in \mathcal{L}_b \setminus \mathcal{L}_a = \mathcal{U}_a \setminus \mathcal{U}_b$ we obtain the desired inequality with the help of (7) and (8): $\varphi_1^f(a) \leq u_a \leq f(a, x_2, \ldots, x_n) \leq f(b, x_2, \ldots, x_n) \leq l_b \leq \varphi_1^f(b)$.

In the second case, we claim that $f(a, x_2, \ldots, x_n) = f(b, x_2, \ldots, x_n)$ for all $(x_2, \ldots, x_n) \in L_2 \times \cdots \times L_n$. This is clear if $(x_2, \ldots, x_n) \in \mathcal{E}_a = \mathcal{E}_b$. If $(x_2, \ldots, x_n) \in \mathcal{L}_a = \mathcal{L}_b$, then

$$f(a, x_2, \ldots, x_n) = f(1, x_2, \ldots, x_n) = f(b, x_2, \ldots, x_n).$$

If $(x_2, \ldots, x_n) \in \mathcal{U}_a = \mathcal{U}_b$, then

$$f(a, x_2, \ldots, x_n) = f(0, x_2, \ldots, x_n) = f(b, x_2, \ldots, x_n).$$

Thus, when determining l_a and l_b according to (7), we have to compute the join of exactly the same elements, hence $l_a = l_b$ (if they are defined). Similarly, we have $u_a = u_b$ whenever they are defined. Therefore $\varphi_1^f(a)$ and $\varphi_1^f(b)$ coincide, no matter which rule (L),(U) or (LU) was used to compute their values. □

Remark 6. We can see from the proof of the above theorem that (LU) could be relaxed: $\varphi_1^f(x_1)$ could be chosen to be any element of $[l_{x_1}, u_{x_1}]$ with the convention that whenever we encounter the same interval $[l_{x_1}, u_{x_1}]$ for different values of x_1, we always choose the same element of this interval. This guarantees that φ_1^f will be order-preserving. All of the proofs below work with this relaxed rule as well, since they rely only on the fact that $\varphi_1^f(x_1) \in [l_{x_1}, u_{x_1}]$ whenever $\varphi_1^f(x_1)$ is determined by rule (LU).

Next we prove that the function φ_1^f can be used in the factorization of the Sugeno utility function f. Let us recall that, since f is a Sugeno utility function, $f(x_1, x_2, \ldots, x_n) = q(\varphi_1(x_1), \varphi_2(x_2), \ldots, \varphi_n(x_n))$ for some Sugeno integral q and local utility functions φ_i. Let us denote $f'(x_1, x_2, \ldots, x_n) = q\big(\varphi_1^f(x_1), \varphi_2(x_2), \ldots, \varphi_n(x_n)\big)$. Observe that f' is also a Sugeno utility function.

Lemma 7. *For all* $(x_2, \ldots, x_n) \in L_2 \times \cdots \times L_n$ *we have*

$$f'(0, x_2, \ldots, x_n) = f(0, x_2, \ldots, x_n),$$
$$f'(1, x_2, \ldots, x_n) = f(1, x_2, \ldots, x_n).$$

Proof. We prove the first equality; the proof of the second equality is similar. Let us observe first that $\mathcal{W}_0 = \mathcal{L}_0 = \emptyset$, and $\mathcal{U}_0 \neq \emptyset$, since otherwise we had $L_2 \times \cdots \times L_n = \mathcal{E}_0$, contradicting our assumption that f depends on its first variable. Thus $\varphi_1^f(0)$ is determined by the rule (U), and $\varphi_1^f(0) = u_0 \geq \varphi_1(0)$ according to (8). Since q is order-preserving, this immediately implies that $f'(0, x_2, \ldots, x_n) \geq f(0, x_2, \ldots, x_n)$. For the other inequality we treat the two cases $(x_2, \ldots, x_n) \in \mathcal{U}_0$ and $(x_2, \ldots, x_n) \in \mathcal{E}_0$ separately.

If $(x_2, \ldots, x_n) \in \mathcal{U}_0$, then $f(0, x_2, \ldots, x_n)$ is one of the elements whose meet gives u_0 in (8), therefore $u_0 \leq f(0, x_2, \ldots, x_n)$. Thus we have

$$\begin{aligned} f'(0, x_2, \ldots, x_n) &= q\big(\varphi_1^f(0), \varphi_2(x_2), \ldots, \varphi_n(x_n)\big) \\ &= q(u_0, \varphi_2(x_2), \ldots, \varphi_n(x_n)) \\ &\leq q\big(f(0, x_2, \ldots, x_n), \varphi_2(x_2), \ldots, \varphi_n(x_n)\big) \\ &= q\big(q(\varphi_1(0), \varphi_2(x_2), \ldots, \varphi_n(x_n)), \varphi_2(x_2), \ldots, \varphi_n(x_n)\big). \end{aligned}$$

By Proposition 2, the right hand side equals

$$q(\varphi_1(0), \varphi_2(x_2), \ldots, \varphi_n(x_n)) = f(0, x_2, \ldots, x_n).$$

Hence we can conclude that $f'(0, x_2, \ldots, x_n) \leq f(0, x_2, \ldots, x_n)$.

Now let us assume that $(x_2,\dots,x_n)\in\mathcal{E}_0$. We have observed at the beginning of the proof that $\varphi_1^f(0)=u_0\geq\varphi_1(0)$. In a similar manner one can see that $\varphi_1^f(1)=l_1\leq\varphi_1(1)$, and therefore $\varphi_1^f(0)\leq\varphi_1^f(1)\leq\varphi_1(1)$ since φ_1^f is order-preserving. This allows us to make the following estimate:

$$\begin{aligned}f'(0,x_2,\dots,x_n)&=q\big(\varphi_1^f(0),\varphi_2(x_2),\dots,\varphi_n(x_n)\big)\\&\leq q\left(\varphi_1(1),\varphi_2(x_2),\dots,\varphi_n(x_n)\right)\\&=f(1,x_2,\dots,x_n).\end{aligned}$$

However, $f(1,x_2,\dots,x_n)=f(0,x_2,\dots,x_n)$ as $(x_2,\dots,x_n)\in\mathcal{E}_0$, so we can again conclude that $f'(0,x_2,\dots,x_n)\leq f(0,x_2,\dots,x_n)$. □

Lemma 8. *For all* $(x_1,\dots,x_n)\in L_1\times\cdots\times L_n$ *we have* $f'(x_1,\dots,x_n)=f(x_1,\dots,x_n)$.

Proof. Using the pseudo-median decomposition and Lemma 7 we obtain the following expression for f':

$$\begin{aligned}f'(x_1,x_2,\dots,x_n)&=\operatorname{med}\big(f'(0,x_2,\dots,x_n),\varphi_1^f(x_1),f'(1,x_2,\dots,x_n)\big)\\&=\operatorname{med}\big(f(0,x_2,\dots,x_n),\varphi_1^f(x_1),f(1,x_2,\dots,x_n)\big).\end{aligned}$$

Thus it suffices to show that

$$\operatorname{med}\big(f(0,x_2,\dots,x_n),\varphi_1^f(x_1),f(1,x_2,\dots,x_n)\big)=f(x_1,x_2,\dots,x_n). \tag{10}$$

We separate four cases with respect to the partition of $L_2\times\cdots\times L_n$.

If $(x_2,\dots,x_n)\in\mathcal{W}_{x_1}$, then $\varphi_1^f(x_1)=w_{x_1}=\varphi_1(x_1)$ by (6), hence (10) is nothing else but the pseudo-median decomposition of f.

If $(x_2,\dots,x_n)\in\mathcal{L}_{x_1}$, then $\varphi_1^f(x_1)\geq l_{x_1}$ no matter which one of the rules (W),(L),(U),(LU) was used to define $\varphi_1^f(x_1)$. Then by (7) and by the definition of $\mathcal{L}_{x_1}$ we get

$$\varphi_1^f(x_1)\geq l_{x_1}\geq f(x_1,x_2,\dots,x_n)=f(1,x_2,\dots,x_n).$$

Therefore, the left hand side of (10) equals $f(1,x_2,\dots,x_n)$, and right hand side has the same value, since $(x_2,\dots,x_n)\in\mathcal{L}_{x_1}$.

The case $(x_2,\dots,x_n)\in\mathcal{U}_{x_1}$ follows similarly. Finally, if $(x_2,\dots,x_n)\in\mathcal{E}_{x_1}$, then the left hand side of (10) is $f(0,x_2,\dots,x_n)=f(1,x_2,\dots,x_n)$ independently of the value of $\varphi_1^f(x_1)$, and right hand side has the same value, since $(x_2,\dots,x_n)\in\mathcal{E}_{x_1}$. □

Now we are ready to prove the main result of this paper.

Theorem 9. *For any Sugeno utility function f, the Sugeno integral q^f and the local utility functions φ_i^f defined in Sect. 3 give a factorization of f:*

$$f(x_1,\dots,x_n) = q^f\big(\varphi_1^f(x_1),\dots,\varphi_n^f(x_n)\big).$$

Proof. Lemma 8 shows that $f(x_1,\dots,x_n) = q\big(\varphi_1^f(x_1),\varphi_2(x_2),\dots,\varphi_n(x_n)\big)$. In a similar way one can show that φ_2 can be replaced by φ_2^f: $f(x_1,\dots,x_n) = q\big(\varphi_1^f(x_1),\varphi_2^f(x_2),\dots,\varphi_n(x_n)\big)$. By recursive reasoning, we can replace the local utility functions one by one, and we get

$$f(x_1,x_2,\dots,x_n) = q\big(\varphi_1^f(x_1),\varphi_2^f(x_2),\dots,\varphi_n^f(x_n)\big).$$

Now applying Theorem 4 to this latter factorization of f we obtain $f(x_1,\dots,x_n) = q^f\big(\varphi_1^f(x_1),\varphi_2^f(x_2),\dots,\varphi_n^f(x_n)\big)$. □

5 Concluding Remarks

We have given a method to factorize any Sugeno utility function f into a composition $f = q^f\big(\varphi_1^f,\dots,\varphi_n^f\big)$ of a Sugeno integral q^f with local utility functions φ_i^f. Such a factorization can be applied to analyse the behaviour of f, which can be useful in many problems in decision making. However, in many situations, we do not know whether our overall utility function f is a Sugeno utility function. This can be decided by making use of the various characterizations given in [3]. Alternatively, one can apply the construction of Sect. 3 directly to f. If at some point the construction fails (e.g, there are several values for w_{x_1} or $l_{x_1} > w_{x_1}$, etc.), then f does not have such a factorization. If the construction works, then we obtain a function $q^f\big(\varphi_1^f(x_1),\varphi_2^f(x_2),\dots,\varphi_n^f(x_n)\big)$. If this function coincides with f, then we have obtained the desired factorization of f, otherwise f is not a Sugeno utility function.

As we have mentioned, the pseudo-median decomposition formula (3) is valid for the wider class of pseudo-Sugeno integrals [3]. Let us observe that in our proofs we never made use of the fact that the functions φ_i are order-preserving, only the order-preservation of f (and the pseudo-median decomposition) was used. Thus Theorems 5 and 9 hold in this more general setting, and this implies that a pseudo-Sugeno integral is order-preserving if and only if it is a Sugeno utility function.

Acknowledgments. We would like to thank Jean-Luc Marichal for useful discussions and for bringing this topic to our attention. The second named author acknowledges that the present project is supported by the National Research Fund, Luxembourg, and cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND), and supported by the Hungarian National Foundation for Scientific Research under grant no. K77409.

References

1. Bouyssou, D., Dubois, D., Prade, H., Pirlot, M. (eds.): Decision-Making Process - Concepts and Methods. ISTE/John Wiley (2009)
2. Couceiro, M., Marichal, J.-L.: Polynomial functions over bounded distributive lattices. Journal of Multiple-Valued Logic and Soft Computing (to appear), http://arxiv.org/abs/0901.4888
3. Couceiro, M., Waldhauser, T.: Sugeno utility functions I: Axiomatizations. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) MDAI 2010. LNCS (LNAI), vol. 6408, pp. 79–90. Springer, Heidelberg (2010)
4. Dubois, D., Marichal, J.-L., Prade, H., Roubens, M., Sabbadin, R.: The Use of the Discrete Sugeno Integral in Decision-Making: a Survey. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 9(5), 539–561 (2001)
5. Goodstein, R.L.: The Solution of Equations in a Lattice. Proc. Roy. Soc. Edinburgh Sect. A 67, 231–242 (1965/1967)
6. Grabisch, M.: The application of fuzzy integrals in multicriteria decision making. Eur. J. Oper. Res. 89(3), 445–456 (1996)
7. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation Functions. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge (2009)
8. Marichal, J.-L.: On Sugeno integral as an aggregation function. Fuzzy Sets and Systems 114, 347–365 (2000)
9. Marichal, J.-L.: Weighted lattice polynomials. Discrete Mathematics 309(4), 814–820 (2009)
10. Sugeno, M.: Theory of Fuzzy Integrals and its Applications. PhD thesis, Tokyo Institute of Technology, Tokyo (1974)
11. Sugeno, M.: Fuzzy Measures and Fuzzy Integrals–a Survey. In: Gupta, M.M., Saridis, G.N., Gaines, B.R. (eds.) Fuzzy Automata and Decision Processes, pp. 89–102. North-Holland, New York (1977)

Managing Information Fusion with Formal Concept Analysis

Zainab Assaghir[1], Mehdi Kaytoue[1], Amedeo Napoli[1], and Henri Prade[2]

[1] Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
Campus Scientifique, B.P. 70239 – 54500 Vandœuvre-lès-Nancy – France
[2] Institut de Recherche en Informatique de Toulouse (IRIT)
118 Route de Narbonne – 31062 Toulouse – France
assaghiz@loria.fr, kaytouem@loria.fr, napoli@loria.fr, prade@irit.fr

Abstract. The main problem addressed in this paper is the merging of numerical information provided by several sources (databases, experts...). Merging pieces of information into an interpretable and useful format is a tricky task even when an information fusion method is chosen. Fusion results may not be in suitable form for being used in decision analysis. This is generally due to the fact that information sources are heterogeneous and provide inconsistent information, which may lead to imprecise results. In this paper, we propose the use of Formal Concept Analysis and more specifically pattern structures for organizing the results of fusion methods. This allows us to associate any subset of sources with its information fusion result. Once a fusion operator is chosen, a concept lattice is built. With examples throughout this paper, we show that this concept lattice gives an interesting classification of fusion results. When the fusion global result is too imprecise, the method enables the users to identify what maximal subset of sources that would support a more precise and useful result. Instead of providing a unique fusion result, the method yields a structured view of partial results labelled by subsets of sources. Finally, an experiment on a real-world application has been carried out for decision aid in agricultural practices.

1 Introduction

In this paper, we present a method for managing information fusion based on Formal Concept Analysis (FCA) when information is numerical. The problem of information fusion is encountered in various fields of application, e.g sensor fusion, multiple source interrogation systems. Information fusion consists of merging, or exploiting conjointly, several sources of information for answering questions of interest and make proper decisions [1]. A fusion operator is an operation summarizing all information given by sources into an interpretable information, for example the interval intersection for numerical information.

Several fusion operators were proposed for combining uncertain information [2, 3, 4, 5, 6, 7] and no universal method is available [2]. Dubois and Prade [2] overviewed how fuzzy set theory can address the information fusion problem

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 104–115, 2010.

Table 1. Information dataset given by sources

	m_1	m_2
g_1	$[1, 5]$	$[1, 9]$
g_2	$[2, 3]$	$[1, 3]$
g_3	$[4, 7]$	$[6, 7]$
g_4	$[6, 10]$	$[8, 9]$

Table 2. A formal context

	m_1	m_2
g_1	×	
g_2	×	
g_3	×	×
g_4		×

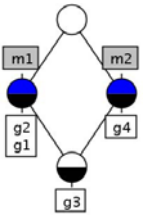


Fig. 1. Concept lattice raised from Table 2

and proposed several fusion operators for numerical information. More recently, a fusion operator based on the notion of Maximal Consistent Subset (MCS) has been proposed for finding a global point of view when no meta-knowledge is available about sources (reliability, conflict) [8,9]. These works apply the fusion operator on the set of all sources and consider the resulting information. Other approaches define their proper fusion operator in a lattice structure to combine symbolic information [6,7].

In this work, we use FCA to study all subsets of sources and their information fusion results. The main ability of FCA is to produce formal concepts corresponding to maximal sets of sources associated with a fused information. The concepts are ordered and form a structure called concept lattice. We show that this lattice contains the information fusion result considering all sources proposed by [2,8,9]. Moreover, the lattice is meaningful for organizing information fusion results of different subsets of sources and allows more flexibility for the user. Moreover, the lattice keeps a track of the origin of the information such as presented in [3] for the fusion of symbolic information.

This work can be used in many applications where it is necessary to find a suitable value summarizing several values coming from multiple sources. Here, we use an experiment in agronomy for decision helping in agricultural practices.

The paper is organized as follows. Section 2 presents and illustrates the basics of fusion operators. Section 3 introduces the preliminaries on FCA and its generalization for handling numerical data. Then, Section 4 shows how FCA is well suited for organizing different information fusion results. Section 5 describes a real-world experiment: a concept lattice embedding fusion results is interpreted for making decisions about agricultural practices.

2 Basics of Numerical Information Fusion Operators

According to previous works, there are three kinds of behaviors for the fusion operators: conjunctive, disjunctive and trade-off operators [1,2,4].

Before introducing these operators, we introduce the following notations: n is the number of sources. $\mathbb{I}_m$ is the set of all values given for the variable m. $f_m : \mathbb{I}_m \to \mathbb{R}$ denotes a fusion operator returning the fusion result for variable m.

The *conjunctive operator* is the counterpart to a set intersection. The imprecision and the uncertainty in the information associated with the result of a

conjunction is less than the imprecision or the uncertainty of each source alone. A conjunctive operator makes the assumption that all the sources are reliable, and usually results in a precise information. If there is some conflict in the information (i.e. at least one source is not fully reliable), then the result of the conjunction can be insufficiently reliable, or even empty. The conjunctive operator for a variable m is defined by $f_m(\mathbb{I}_m) = \bigcap_{i=1,\ldots,n} I_i$, e.g., in Table 1, $f_{m_1}(I_1, \ldots, I_4) = \emptyset$ represents the intersection of intervals of m_1 with $I_1 = [1, 5], I_2 = [2, 3], I_3 = [4, 7]$ and $I_4 = [6, 10]$.

The *disjunctive operator* is the counterpart to a set union. The uncertainty (or the imprecision) resulting from a disjunction is higher than the uncertainty (or the imprecision) of all sources together. A disjunctive operator makes the assumption that at least one source is reliable. The result of a disjunctive operator can be considered as reliable, but is also often (too) weakly informative. The disjunctive operator for the variable m, is defined by $f_m(\mathbb{I}) = \bigcup_{i=1,\ldots,n} I_i$, e.g. , $f_{m_1}(I_1, \ldots, I_4) = [1, 10]$ that represents the union of the intervals of m_1.

The *trade-off operators* lie between conjunctive and disjunctive behaviors, and are typically used when sources are partly conflicting. They try to achieve a good balance between informativeness and reliability [2]. The fusion based on MCS is an example of trade-off operators.

Maximal consistent subset fusion method. When no information is available about sources, like conflict between sources, or reliability of sources, a reasonable fusion method should take into account the information provided by all sources. At the same time, it should try to keep a maximum of informativeness. The notion of MCS is a natural way to achieve these two goals.

The idea of MCS goes back to Rescher and Manor [10]. This notion is currently used in the fusion of logical formulas [5] but also of numerical data [8,9]. Given a set of n intervals $\mathbb{I} = \{I_1, I_2, \ldots, I_n\}$, a subset $K \subseteq \mathbb{I}$ is *consistent* if $\bigcap_{i=1}^{|K|} K_i \neq \emptyset$ with $K_i \in K$ and *maximal* if it does not exist a proper super-set $K' \supseteq K$ that is also consistent. In Table 1, the set $K_1 = \{I_1, I_2\}$ is a MCS of the set $\mathbb{I}_{m_1}$, since $I_1 \cap I_2 \neq \emptyset$ and is maximal w.r.t. intersection property.

The fusion operator of n sources based on MCS consists in applying a disjunctive operator on their MCS. For example, the MCS fusion result for m_1 in Table 1 is $f_{m_1}(I_1, \ldots, I_4) = [2, 3] \cup [4, 5] \cup [6, 7]$, as illustrated in Figure 2. The MCS notion appears as a natural way to conciliate the two objectives of gaining information and of remaining in agreement with all sources in information fusion problem. Generally, finding MCS is a problem having exponential complexity [11]. Dubois et al. [8] introduce a linear algorithm to compute the MCS of n intervals.

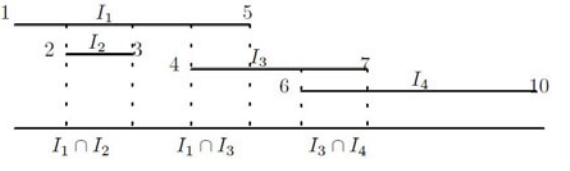


Fig. 2. MCS computed from Table 1 for the variable m_1

Properties of fusion operators. Generally, all fusion operators are commutative and idempotent. The conjunctive and disjunctive operators are associative but not the trade-off fusion operators (more details in [9]). If the final result of the fusion is not convex, it is always possible to take its convex hull (loosing some information in the process but gaining computational tractability). Conjunctive fusion result is convex but this is not the case for the others operators in general.

In conclusion, for merging numerical information, a common fusion operator has to be used. This is specially important in case of heterogeneous sources. Fusion operators are often based on assumptions or on meta-knowledge available about the sources (reliability, conflict) and the domain. Sometimes, it happens that the fusion result is not directly useful for decision. For example, in [12] the fused information must be convex, and the convexification of MCS leads to an imprecise result. Here, we propose to identify and characterize interesting subsets of sources, providing more useful fused information. Accordingly, we show how a fusion operator can be embedded in the framework of Formal Concept Analysis (FCA) to build a concept lattice yielding a structured view of partial results labelled by subsets of sources, instead of providing a unique fusion result.

3 Formal Concept Analysis

3.1 Basics

Formal concept analysis (FCA) [13] starts with a formal context (G, M, I) where G denotes a set of objects, M a set of attributes, and $I \subseteq G \times M$ a binary relation between G and M[1]. The statement $(g, m) \in I$ is interpreted as "the object g has attribute m". An example of formal context is given by Table 2 where a table entry contains a cross ($\times$) iff the object in row has the attribute in column, e.g. g_1 has the attribute m_1, i.e. $(g_1, m_1) \in I$. The two operators $(\cdot)'$ define a Galois connection between the powersets $(2^G, \subseteq)$ and $(2^M, \subseteq)$, with $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\} \qquad B' = \{g \in G \mid \forall m \in B : gIm\}$$

For $A \subseteq G$, $B \subseteq M$, a pair (A, B), such that $A' = B$ and $B' = A$, is called a *(formal) concept*, e.g. $(\{g_1, g_2, g_3\}, \{m_1\})$. In (A, B), the set A is called the *extent* and the set B the *intent* of the concept (A, B). Concepts are partially ordered by $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ ($\Leftrightarrow B_2 \subseteq B_1$), e.g. the concept $(\{g_3\}, \{m_1, m_2\})$ is a sub-concept of $(\{g_1, g_2, g_3\}, \{m_1\})$. With respect to this partial order, the set of all formal concepts forms a complete lattice called the *concept lattice* of the formal context (G, M, I). Figure 1 shows the concept lattice[2] associated with the context in Table 2. On the diagram, each node denotes a concept while a line denotes an order relation between two concepts. Due to *reduced labeling*, the extent of a concept is composed of all objects lying in the extents of its sub-concepts. Dually, the intent of a concept is composed of all attributes in the

[1] In this paper, we similarly use the terms object and information source on one hand, and variable and attribute on the other hand.

[2] The lattice diagram is designed with ConExp, `http://conexp.sourceforge.net/`.

intents of its super-concepts. The top concept ($\top$) is the highest and the bottom concept ($\bot$) is the lowest in the lattice.

The concept lattice provides a classification of objects in a domain. It entails both notions of maximality and generalization/specialization: a concept corresponds to a maximal set of objects (extent) sharing a common maximal set of attributes (intent) ; the generalization/specialization is given by the partial ordering of concepts.

However, real-world data like in biology, agronomy, etc., are not binary, but rather consist in complex data composed of numbers, graphs, etc. The data are classically processed with FCA after a data transformation, called *conceptual scaling*, e.g. discretization. Transformations generally imply an important loss of information and arbitrary choices, which must be avoided in the context of information fusion. For example, an object has the attribute m_1 (resp. m_2) in the binary Table 2 iff its values for this attribute are less than 7 (resp. greater than 5) in the numerical Table 1. With other choices, we may obtain another table, and hence another concept lattice with a different interpretation. Therefore, handling numerical data for information fusion purposes with FCA is not straightforward.

3.2 Pattern Structures for Complex Data

Instead of transforming data, one may directly work on the original data. For that purpose, a *pattern structure* is defined as a generalization of a formal context to complex data [14]. It still maps objects to their descriptions, the latter being partially ordered. When working with classical FCA, the object descriptions are sets of attributes, and are partially ordered by set inclusion, w.r.t. set intersection: let $P, Q \subseteq M$ two attributes sets, then $P \subseteq Q \Leftrightarrow P \cap Q = P$, and $(M, \subseteq)$, also written $(M, \cap)$, is a partially ordered set of object descriptions. The set intersection $\cap$ behaves as a meet operator, denoted by $\sqcap$, in a semi-lattice: it is *idempotent*, *commutative*, and *associative*. Therefore, a pattern structure naturally entails a Galois connection between the powerset of objects $(2^G, \subseteq)$ and a meet-semi-lattice of descriptions denoted by $(D, \sqcap)$.

Formally, let G be a set of objects, let $(D, \sqcap)$ be a meet-semi-lattice of potential object descriptions and let $\delta : G \longrightarrow D$ be a mapping. Then $(G, (D, \sqcap), \delta)$ is called a *pattern structure*. Elements of D are called *patterns* and are ordered by the subsumption relation $\sqsubseteq$: given $c, d \in D$ one has $c \sqsubseteq d \Longleftrightarrow c \sqcap d = c$. A pattern structure $(G, (D, \sqcap), \delta)$ gives rise to the following derivation operators $(\cdot)^{\square}$, given $A \subseteq G$ and $d \in (D, \sqcap)$:

$$A^{\square} = \prod_{g \in A} \delta(g) \qquad\qquad d^{\square} = \{g \in G | d \sqsubseteq \delta(g)\}$$

These operators form a Galois connection between $(2^G, \subseteq)$ and $(D, \sqsubseteq)$. *(Pattern) concepts* of $(G, (D, \sqcap), \delta)$ are pairs of the form (A, d), $A \subseteq G$, $d \in (D, \sqcap)$, such that $A^{\square} = d$ and $A = d^{\square}$. For a pattern concept (A, d), d is called a *pattern intent* and is a common description of all objects in A, called *pattern extent*. When partially ordered by $(A_1, d_1) \leq (A_2, d_2) \Leftrightarrow A_1 \subseteq A_2$ $(\Leftrightarrow d_2 \sqsubseteq d_1)$, the set of all concepts forms a complete lattice called a *(pattern) concept lattice*.

Pattern structures allow to consider complex data in full compliance with the FCA formalism. It requires to define a meet operator on object descriptions, inducing their partial order. In fact, as for scaling in classical FCA, the choice of an operator depends on expert knowledge, and to which extent will the resulting concept lattice be used. Several attempts were done to define such operators, on sets of graphs [14], numerical data [15], logical formulas [16], etc. In the following, we discuss how a fusion operator can be seen as a meet operator.

4 Organizing Information Fusion Results with FCA

We show here that FCA provides a suitable framework for organizing sources and their information fusion results, allowing more flexibility for the users of fusion results.

Definition (Information fusion space). *An information fusion space D_m is composed of the information available for a variable m and all their possible fusion results, w.r.t a fusion operator f_m.*

For example, with the variable m_1 in Table 1 and f_m as the interval intersection, $D_m = \{[1,5],[4,7],[6,10],[2,3],[4,5],[6,7],\emptyset\}$.

4.1 Formalizing a Fusion Operator as a Meet Operator

Let us consider a single variable $m \in M$, its fusion space D_m corresponding to a chosen fusion operator f_m. When f_m is idempotent, commutative and associative, (D_m, f_m) is a meet-semi-lattice, since f_m behaves as a meet operator. This is the case for any conjunctive or disjunctive fusion operator, and we have $c \sqcap d = f_m(c,d), \forall c,d \in D_m$, meaning that the meet of two elements of D_m corresponds to their fusion.

For example, let us consider the numerical variable m_1 in Table 1, and the conjunctive fusion operator f_{m_1} that corresponds to the interval intersection $\cap$. Figure 3 shows the meet-semi-lattice (D_{m_1}, f_{m_1}). The interval labelling a node is the meet of all intervals labelling its ascending nodes, i.e. the resulting information fusion w.r.t f_{m_1} of the sources given the intervals labelling its ascending nodes. In the example, $f_{m_1}([4,7],[6,10]) = [6,7]$ is the fusion of objects g_3 and g_4 for the variable m_1, and $f_{m_1}([2,3],[1,5]) = [2,3]$ for objects g_1 and g_2. Therefore, we have partially ordered the fusion space D_{m_1} with $c \cap d = c \Leftrightarrow c \subseteq d, \forall c,d \in D_{m_1}$. This order is a particular instance of the pattern subsumption relation defined in pattern structures. It means, in this example, that an interval is subsumed by any larger one, e.g. $[2,3] \sqsubseteq [1,5]$ since $[2,3] \subseteq [1,5]$. For example, we have $[2,3] \sqcap [1,5] = [2,3] \Leftrightarrow [2,3] \sqsubseteq [1,5]$ in terms of semi-lattice, corresponding to $[2,3] \cap [1,5] = [2,3] \Leftrightarrow [2,3] \subseteq [1,5]$ in interval inclusion terms. Note that a disjunctive fusion operator is handled similarly.

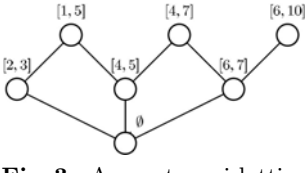


Fig. 3. A meet-semi-lattice of intervals

4.2 Building a Concept Lattice from Information Sources

Given G a set of sources, $m \in M$ a single variable, (D_m, f_m) the meet-semi-lattice of fusion results, and δ a mapping that gives to any object its information for the variable m, then $(G, (D_m, f_m), \delta)$ is a pattern structure. On the example, we have $(G, (D_{m_1}, f_{m_1}), \delta)$. (D_{m_1}, f_{m_1}) is described in the previous subsection. Descriptions of sources g_1 and g_2 are respectively $\delta(g_1) = [1, 5]$ and $\delta(g_2) = [2, 3]$. Then, the general Galois connection can be used to compute and order concepts:

$$\begin{aligned} \{g_1, g_2\}^{\square} &= [1,5] \sqcap [2,3] \\ &= f_{m_1}([1,5],[2,3]) \\ &= [2,3] \end{aligned} \qquad \begin{aligned} [2,3]^{\square} &= \{g \in G \mid [2,3] \sqsubseteq \delta(g)\} \\ &= \{g \in G \mid [2,3] \subseteq \delta(g)\} \\ &= \{g_1, g_2\}. \end{aligned}$$

Since $\{g_1, g_2\} = [2, 3]$ and $[2, 3]^{\square} = \{g_1, g_2\}$, the pair $(\{g_1, g_2\}, [2, 3])$ is a concept. Efficient FCA algorithms can extract the set of all formal concepts and order them within a concept lattice [17]. They can be easily adapted to compute in pattern structures [14,15]. The lattice of our example is given in Figure 4.

4.3 Concept Lattice Interpretation

A concept (A, d) of $(G, (D_{m_1}, f_{m_1}), \delta)$, is interesting from many points of view, as illustrated with the concept $(\{g_1, g_2\}, [2, 3])$.

- Its intent d provides the fusion resulting from objects in A, e.g. $[2, 3]$ is the conjunctive fusion f_{m_1} of the information from sources g_1 and g_2.
- No other object can be added to A without changing d, e.g. $\{g_1, g_2\}$ is the maximal set of sources whose conjunctive information fusion is $[2, 3]$.
- The extent A keeps the track of the origin of the information, e.g. it is known that the new information $[2, 3]$ comes from the information of g_1 and g_2.

The resulting concept lattice provides a suitable classification of information sources and their information fusion results. In Figure 4, a concept extent is read with reduced labelling. However, for sake of readability, intents are given for each concept (not reduced). For example, the node labelled with $[6, 7]$ represents the concept $(\{g_3, g_4\}, [6, 7])$. Due to concept ordering, a concept provides the fusion result of a subset of the extent of its super-concepts (generalization/specialization). Then, the navigation in the lattice gives interesting insights into the fusion results. This allows more flexibility for decision making. For example, in related works, only the fusion of information of all objects is considered which corresponds to the most general concept ($\top$) in the lattice. This result does not always allow to make a decision, e.g. an empty intersection in our example. Then it is interesting to observe subsets of objects, by navigating in the lattice.

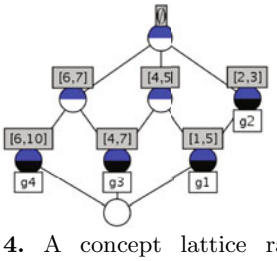


Fig. 4. A concept lattice raised from Table 1 for the variable m_1

4.4 Lattice Based on Maximal Consistent Subsets

The fusion operator f_m based on the notion of MCS is idempotent and commutative, but not associative. For example in Table 1, $f_{m_1}(f_{m_1}([1,5],[2,3]),[4,7]) = [2,3] \cup [4,7]$ and $f_{m_1}(f_{m_1}([1,5],[4,7]),[2,3]) = [2,3] \cup [4,5]$. Then, the fusion operator cannot be directly used as a meet operator to build a concept lattice.

However, since this operator returns the union of all MCS, we can firstly compute all MCS for a given variable, denoted by the set K and then use the disjunctive operator on the MCS as a meet operator to define a meet-semi-lattice $(K, \cup)$. Formally, we consider $(\mathcal{O}, (K, \cup), \delta)$ as a pattern structure where $\mathcal{O}$ is a multi-set of sources, each element is set of sources of one MCS $k \in K$, i.e. $\delta(o) \in K, \forall o \in \mathcal{O}$. For example, the MCS of intervals for m_1 are $[2,3]$, $[4,5]$ and $[6,7]$ given respectively by $\{g_1, g_2\}$, $\{g_1, g_3\}$ and $\{g_3, g_4\}$. Then, $\mathcal{O}$ represents the multi-set $\{\{g_1, g_2\}, \{g_1, g_3\}, \{g_3, g_4\}\}$ with $\delta(\{g_1, g_2\}) = [2,3]$ (meaning that the interval of values $[2,3]$ is related to the sources g_1 and g_2), $\delta(\{g_1, g_3\}) = [4,5]$ and $\delta(\{g_3, g_4\}) = [6,7]$. Then, we use an interval union as a meet operator. The resulting concept lattice is given in Figure 5. A concept extent is read with reduced labelling. A concept intent is given here for each concept. For example, in Figure 5, the right concept in the second line is $(\{\{g_1, g_2\}, \{g_1, g_3\}\}, [2,3] \cup [4,5])$ giving the values of m_1 w.r.t. the sources $\{g_1, g_2\}$ and $\{g_1, g_3\}$. Moreover, these values represent the MCS fusion result of the subset $\{g_1, g_2, g_3\}$. The concept $\top$ corresponds to the union of all MCS that is the MCS fusion result of all sources.

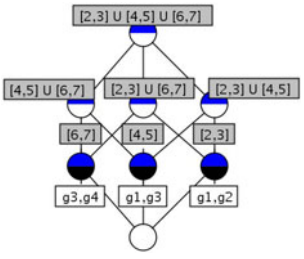


Fig. 5. Concept lattice with MCS

The method used here to obtain the lattice based on MCS does not consider all subsets of objects with their MCS fusion results. This is due to the non-associativity of the MCS fusion operator. Thus, the concept lattice does not contain all subsets of G with their MCS fusion results since the interval union is used on the MCS of data and not directly on the data given by sources. Nevertheless, the concept lattice helps us to keep the origin of the information and gives more flexibility for the users in the choice of a maximal consistent subset of sources in many application fields.

4.5 Embedding Several Variables in the Concept Lattice

Sources can provide values for different variables. For example, Table 1 involves objects described by vectors of intervals, where each dimension, i.e. column, corresponds to a unique variable, e.g. the description of the object g_1 is denoted by $\delta(g_1) = \langle [1,5], [1,9] \rangle$. It can be interesting to compute the fusion information for all variables simultaneously.

To formalize a pattern structure in this case, one defines a meet operator, i.e. fusion operator in our settings, for each dimension, or variable. Assuming that there is a canonical order on vector dimensions, the meet of two vectors is defined

as the meet on each dimension. This induces a partial order of object descriptions [15]. Thus, we consider the pattern structure $(G, (D, \sqcap), \delta)$, where G is a set of sources, $(D, \sqcap)$ is a meet-semi-lattice of vectors, and each vector dimension is provided with the fusion operator f_m corresponding to the variable m.

Going back to Table 1, descriptions of objects g_1 and g_2 are respectively the vectors $\langle [1,5],[1,9]\rangle$ and $\langle [2,3],[1,3]\rangle$. When the fusion operator for both dimension is the interval intersection, the meet of these two vectors is $\langle [1,5],[1,9]\rangle \sqcap \langle [2,3],[1,3]\rangle = \langle [2,3],[1,3]\rangle$. The subsumption relation for vectors is defined similarly: $\langle [2,3],[1,3]\rangle \sqsubseteq \langle [1,5],[1,9]\rangle$ as $[2,3] \subseteq [1,5]$ and $[1,3] \subseteq [1,9]$. Then, the general Galois connection can be used to compute and order concepts:

$$\begin{array}{ll} \{g_1,g_2\}^{\square} = \langle [1,5],[1,9]\rangle \sqcap \langle [2,3],[1,3]\rangle & \langle [2,3],[1,3]\rangle^{\square} = \{g \in G \mid \langle [2,3],[1,3]\rangle \sqsubseteq \delta(g)\} \\ \quad = \langle [2,3],[1,3]\rangle & \quad = \{g_1,g_2\} \end{array}$$

In this way, a concept represents a set of sources and their fusion w.r.t. all variables, such as no other source can be added without changing the fusion result for any variable. The variables can be either symbolic or numerical since a fusion operator is chosen for each variable.

When the fusion operator is based on MCS, we follow the pre-processing introduced above for each variable (see Section 4.4). Then, we consider the set of all MCS for all variables. Thus, we consider the pattern structure $(\mathcal{O}, (K, \sqcap), \delta)$, where $\mathcal{O}$ is the set of subsets of sources providing the MCS for all variables, $(K, \sqcap)$ is a meet-semi-lattice of vectors. Each subset in $\mathcal{O}$ is described for each dimension by a maximal interval of values if the subset represents a MCS for the corresponding dimension, otherwise the dimension description is empty. In the example, recalling that an object denotes a set of sources giving a MCS, the description of the object $\{g_1, g_2\}$ is $\delta(\{g_1, g_2\}) = \langle [2,3],[1,3]\rangle$ where $[2,3]$ and $[1,3]$ are respectively a MCS for m_1 and m_2. By contrast, the description of the object $\{g_3, g_4\}$ is $\delta(\{g_3, g_4\}) = \langle [6,7], \emptyset\rangle$ since the subset $\{g_3, g_4\}$ does not represent a MCS for the variable m_2.

This framework on fusion operators has been used on real-world data as explained in the next section.

5 A Real-World Application in Agronomy

Data and problem settings. Agronomists compute indicators for evaluating the impact of agricultural practices on the environment. Questions such as the following are of importance: what are the consequences of the application of a pesticide given its characteristic, the period of application, and the characteristics of the field? The risk level for a pesticide to reach groundwater is computed by the indicator I_{gro} in [18]. Agronomists try to make a diagnosis w.r.t. the value of I_{gro}. A value below 7 indicates that the farmer has to change its practices (pesticide, soil, date, etc.). By contrast, a value above 7 indicates that the practices of the farmer are environmental friendly [19]. Pesticide characteristics depend on the chemical characteristics of the product while pesticide period application and field characteristics depend on domain knowledge [19]. This

knowledge lies in information sources among which books, databases, and expert knowledge in agronomy. Then values for some characteristics vary w.r.t. sources. Here, we are interested in the use of pesticide *sulcotrione* and its influence on the groundwater. Sulcotrione is a herbicide marketed since 1993. It is used to control a wide range of grasses weeds in maize crops. Sulcotrione is generally weakly absorbed by soils [20]. Three characteristics of *Sulcotrione* are needed to compute the indicator I_{gro}, namely $DT50$, koc, and ADI (more details on these characteristics can be found in [18], and are not crucial for the understanding of this paper). Table 3 (simplified data) gives the values of the characteristics $DT50$ and koc according to 9 different information sources. The symbol "?" represents the case when the information source does not give data for the characteristic. The value of ADI for the *sulcotrione* is 0.00005. Agronomists look to find a suitable value for each characteristic to be considered for computing the I_{gro} indicator, hence facing an information fusion problem.

Table 3. Characteristics of *Sulcotrione*

	DT50 day	koc L/kg
BUS	[2,74]	?
PM11	[15,72]	?
PM12	?	[44,940]
PM13	?	[44,940]
INRA	?	[1.08,8.98]
Com98	[2,6]	[17,160]
AGXf	[2,6]	[1.08,160]
AGXl	[15,74]	[1.08,160]

Lattice construction and interpretation. To combine the different pieces of information, a common fusion operator has to be defined. In this application, (1) the sources are heterogeneous (2) no *a priori* knowledge about sources and characteristics is available. Therefore, an appropriate fusion operator is the MCS fusion operator. The MCS for the variable $DT50$ are K_1 and K_2, resp. K_3 and K_4 for koc (see Table 4). Table 5 results from the pre-processing of Table 3, detailed in Section 4.4. The resulting concept lattice is given in Figure 6 with 16 concepts. A concept extent is read with reduced labelling. A concept intent is not given in vectorial form for sake of readability: it is read from the intents of sub-concepts, for example, the intent of the concept C_1 is $\{(DT50, [15, 72]), (koc, [44, 160])\}$. But, if two sub-concepts intents give different values for a same attribute, then the union of values is considered. For example, the intent of the concept C_2 is $\{(DT50, [2, 6] \cup [15, 72]), (koc, [44, 160])\}$ and its sub-concepts intents are $\{(DT50, [2, 6])\}$, $\{(DT50, [15, 72])\}$ and $\{(koc, [44, 160])\}$. Moreover, each concept intent in the lattice represents the MCS fusion result of the subset of sources in the extent. The highest concept in the lattice corresponds to the MCS fusion result of all sources for all characteristics. For example, the "most right-down" concept is $(\{K_1\}, \{(DT50, [2, 6])\})$ where $[2, 6]$ is the MCS fusion result of the subset $K_1 = \{BUS, Com98, AGXf\}$ and its "most right" superconcept is $(\{K_1, K_2\}, \{(DT50, [2, 6] \cup [15, 72])\}$ where $[2, 6] \cup [15, 72]$ is the fusion result of the set $K_1 \cup K_2 = \{BUS, PM11, AGXl, Com98, AGXf\}$.

Results and discussion. The computing of a lower and higher bound for the indicator and the consequences of the results on agronomic practices and pollution are detailed and discussed in [12], but will not be detailed here as this is not necessary. It is required to consider the convex hull of the fusion result for computing the indicator. The concept lattice allows the users of I_{gro} and experts to give several diagnosis for the farmer. For example, let us consider the concept

Table 4. Label of all MCS

K_1	$\{BUS, Com98, AGXf\}$
K_2	$\{BUS, PM11, AGXl\}$
K_3	$\{INRA, Com98, AGXf, AGXl\}$
K_4	$\{PM12, PM13, Com98, AGXf, AGXl\}$

Table 5. Table 3 pre-processed

	DT50 (days)	koc (L/kg)
K_1	[2,6]	$\emptyset$
K_2	[15,72]	$\emptyset$
K_3	$\emptyset$	[1.08,8.98]
K_4	$\emptyset$	[44,160]

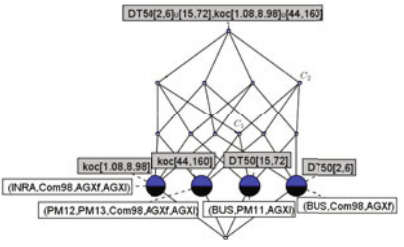


Fig. 6. Concept lattice built from Table 5

$\top$ that represents the fusion result of all sources for all characteristics. Then, $DT50$ and koc lie respectively in $[2, 72]$ and $[1.08, 160]$. With these values, the computed value for I_{gro} is $[4, 10]$. This interval is not useful since all values in $[4, 10]$ are neither smaller than 7 nor greater than 7 and the expert cannot make a decision on the practices of the farmer.

Now the indicator I_{gro} can be also computed choosing either intervals of values in higher or lower level concepts. For instance, if we consider the values of $DT50$ in $[2, 6]$, koc in $[44, 160]$ then we obtain the interval $[9.97, 10]$ for I_{gro} and the practices of the farmer are environmental friendly since the I_{gro} value is greater than 7. However, if $DT50 = [15, 72]$ and $koc = [1.08, 8.98]$, the resulting interval for I_{gro} is $[4.32, 4.32]$ indicating that the farmer must change its practices since values of I_{gro} are smaller than 7. Anyhow, we obtain, with these concepts, precise results of I_{gro}, which is not the case with the fusion global result when using the most general concept. The concept lattice allows to identify what maximal subsets of sources support the most precise results. A further step is to consider these precise results in a decision process.

6 Conclusion

In this paper, we claim that Formal Concept Analysis has the capability of supporting a decision making process in the presence of information fusion problems, even when information are complex, e.g. numbers, thanks to the formalism of pattern structures. A real-world experiment in agronomy showed that when a fusion result does not allow to make a decision, the concept lattice helps the expert by considering an ordered hierarchy of concepts, given the fusion from different maximal sets of sources. Some fusion operators can directly be used to build a concept lattice, e.g. conjunctive and disjunctive operators. To deal with the operator based on maximal coherent subsets (MCS), we proposed to transform the data since MCS is not an associative operator, and the resulting concept lattice entails fusion results of interest.

We have considered the case when information are represented by fuzzy intervals and possibility distributions in [21]. As perspective, it is interesting to study how other fusion operators can be embedded in a concept lattice, as well as meta-information on sources (when available).

References

1. Bloch, I., Hunter, A., Ayoun, A., Benferhat, S., Besnard, P., Cholvy, L., Cooke, R., Dubois, D., Fargier, H.: Fusion: general concepts and characteristics. International Journal of Intelligent Systems 16, 1107–1134 (2001)
2. Dubois, D., Prade, H.: Possibility theory and data fusion in poorly informed environments. Control Eng. Practice 2, 811–823 (1994)
3. Dubois, D., Lang, J., Prade, H.: Dealing with multi-source information in possibilistic logic. In: European Conference on Artificial Intelligence, pp. 38–42 (1992)
4. Dubois, D., Prade, H.: Possibility theory in information fusion. In: Data fusion and Perception, pp. 53–76 (2001)
5. Benferhat, S., Dubois, D., Prade, H.: Reasoning in inconsistent stratified knowledge bases. In: Int. Symp. on Multiple-Valued Logic, pp. 184–189 (1996)
6. Chaudron, L., Maille, N.: Le modèle des cubes: représentation algébrique des conjonctions de propriétés. In: Reconnaissance des Formes et Intelligence Artificielle, RFIA (2000) (in French)
7. Phan-Luong, V.: A framework for integrating information sources under lattice structure. Information Fusion 9(2), 278–292 (2008)
8. Dubois, D., Fargier, H., Prade, H.: Multiple source information fusion: a practical inconsistency tolerant approach. In: IPMU, pp. 1047–1054 (2000)
9. Destercke, S., Dubois, D., Chojnacki, E.: Possibilistic information fusion using maximal coherent subsets. IEEE Transactions on Fuzzy Systems 17, 79–92 (2009)
10. Rescher, N., Manor, R.: On inference from inconsistent premisses. Theory and Decision 1, 179–217 (1970)
11. Malouf, R.: Maximal consistent subsets. Comput. Linguist. 33(2), 153–160 (2007)
12. Assaghir, Z., Girardin, P., Napoli, A.: Fuzzy logic approach to represent and propagate imprecision in agri-environmental indicator assessment. In: IFSA/EUSFLAT Conf., pp. 707–712 (2009)
13. Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Heidelberg (1999)
14. Ganter, B., Kuznetsov, S.O.: Pattern Structures and Their Projections. In: Delugach, H.S., Stumme, G. (eds.) ICCS 2001. LNCS (LNAI), vol. 2120, pp. 129–142. Springer, Heidelberg (2001)
15. Kaytoue, M., Duplessis, S., Kuznetsov, S.O., Napoli, A.: Two FCA-Based Methods for Mining Gene Expression Data. In: Formal Concept Analysis. LNCS, vol. 5548. Springer, Heidelberg (2009)
16. Chaudron, L., Maille, N.: Generalized formal concept analysis. In: Ganter, B., Mineau, G.W. (eds.) ICCS 2000. LNCS, vol. 1867, pp. 357–370. Springer, Heidelberg (2000)
17. Kuznetsov, S.O., Obiedkov, S.A.: Comparing Performance of Algorithms for Generating Concept Lattices. J. Exp. Theor. Artif. Intell. 14, 189–216 (2002)
18. Van der Werf, H., Zimmer, C.: An indicator of pesticide environmental impact based on a fuzzy expert system. Chemosphere 36(10), 2225–2249 (1998)
19. Bockstaller, C., Girardin, P., Van Der Werf, H.: Use of agro-ecological indicators for the evaluation of farming systems. European Journal of Agronomy 7(1-3) (1997)
20. Baer, U., Calvet, R.: Fate of Soil Applied Herbicides: Experimental Data and Prediction of Dissipation Kinetics. J. Environ. Qual. 28(6), 1765–1777 (1999)
21. Assaghir, Z., Kaytoue, M., Prade, H.: A possibility theory-oriented discussion of conceptual pattern structures. In: Int. Conf. on Scalable Uncertainty Management (SUM). LNCS. Springer, Heidelberg (2010)

Indefinite Kernel Fuzzy *c*-Means Clustering Algorithms

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, Japan

Abstract. This paper proposes two types of kernel fuzzy *c*-means algorithms with an indefinite kernel. Both algorithms are based on the fact that the relational fuzzy *c*-means algorithm is a special case of the kernel fuzzy *c*-means algorithm. The first proposed algorithm adaptively updated the indefinite kernel matrix such that the dissimilarity between each datum and each cluster center in the feature space is non-negative, instead of subtracting the minimal eigenvalue of the given kernel matrix as its preprocess. This derivation follows the manner in which the non-Euclidean relational fuzzy *c*-means algorithm is derived from the original relational fuzzy *c*-means one. The second proposed method produces the memberships by solving the optimization problem in which the constraint of non-negative memberships is added to the one of K-sFCM. This derivation follows the manner in which the non-Euclidean fuzzy relational clustering algorithm is derived from the original relational fuzzy *c*-means one. Through a numerical example, the proposed algorithms are discussed.

Keywords: Indefinite Kernel, Kernel Fuzzy *c*-Means, Non-Euclidean Relational Fuzzy *c*-Means, Non-Euclidean Fuzzy Relational Clustering.

1 Introduction

Fuzzy *c*-means (FCM) [1] is a well-known fuzzy clustering method that is derived from hard *c*-means (HCM), also called *k*-means. Among the many FCM variants proposed thus far, one is the FCM algorithm based on the concept of regularization by entropy [2]. This algorithm is called entropy regularized FCM (eFCM) and is discussed not only because of its usefulness but also because of its mathematical relationships with other techniques. We call the FCM proposed in [1] standard FCM (sFCM) in order to distinguish it from eFCM.

In order to cluster data with nonlinear borders, three algorithms [3], [4] have been proposed using nonlinear transformation from the original pattern space into a higher-dimensional feature space with kernel functions in Support Vector Machine (SVM) [5]. These algorithms are called K-HCM, K-sFCM, and K-eFCM, and they are derived from HCM, sFCM, and eFCM, respectively. For simplicity, we generally call them K-CM.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 116–128, 2010.

An explicit mapping is generally unknown for kernel data analysis but their inner product should be known. However, an explicit mapping has been introduced by one of the authors and the appearance of K-CM in a higher-dimensional space has been described via kernel principal component analysis using the explicit mapping [6], [7].

A kernel matrix must be positive-definite for K-CM. However, an irresponsibly introduced kernel matrix is not always positive-definite. In particular, K-sFCM with an indefinite kernel matrix has a risk in that memberships cannot be calculated after the dissimilarity between a datum and a cluster center is updated to be negative, whereas K-HCM and K-eFCM can be continued even if it is negative although the meaningfulness of the obtained result is unknown. Although indefinite kernel matrices can be transformed to positive-definite ones by subtracting the minimal eigenvalue from their diagonal components, this incurs some computational costs.

While the above mentioned clustering methods assume that the data are given as the points in Euclidean space, there are the case that only the dissimilarity or similarity between each datum, called relational data, is given. sFCM has been developed for such relational data into relational fuzzy c-means (RFCM) [8]; however, RFCM has a drawback in that it cannot be applied to non-Euclidean relational data. In order to overcome this drawback, non-Euclidean RFCM (NERFCM) has been proposed [9], in which RFCM is executed by adaptively updating relational data. On the other hand, fuzzy analysis (FANNY) [10], a relational clustering method for non-Euclidean relational data, has been extended to more general fuzzifier parameters into non-Euclidean fuzzy relational clustering (NEFRC) [11]. Both FANNY and NEFRC are obtained by solving the optimization in which the constraint of non-negative memberships is added to that of RFCM.

In this paper, two types of algorithms are proposed for K-sFCM with an indefinite kernel matrix. Both algorithms are based on the fact that RFCM is a special case of K-sFCM. The first proposed algorithm is derived in a manner similar to how NERFCM is derived from RFCM, and the second one is derived in a manner similar to how NEFRC is derived from RFCM.

The remainder of this paper is organized as follows. In the second section, we introduce K-CM, RFCM, NERFCM, and NEFRC. In the third section, we show that RFCM is a special case of K-sFCM; this serves as the basis for our two proposed algorithms. In the fourth section, we propose two types of K-sFCM with an indefinite kernel matrix. In the fifth section, we show some numerical examples. In the last section, we conclude this paper.

2 Preliminaries

In this section, we introduce K-CM, RFCM, NERFCM, and NEFRC. K-CM, that is, K-HCM, K-sFCM, and K-eFCM, is the basis for our proposed methods. RFCM and NERFCM are introduced because the manner in which RFCM is modified into NERFCM is applied in our first proposed method to modify K-sFCM. NEFRC is introduced because we follow this derivation for our second proposed method.

2.1 K-CM

For a given data set $X = \{x_i \mid i \in \{1, \ldots, N\}\}$, K-CM assumes that the kernel matrix $K \in \mathbb{R}^{N \times N}$ is given. Let $\mathbb{H}$ be a higher-dimensional feature space, $\Phi : X \to \mathbb{H}$ be a map from data set X to the feature space $\mathbb{H}$, $W = \{W_j \in \mathbb{H} \mid j \in \{1, \cdots, C\}\}$ be a set of cluster centers in the feature space, and $u_{i,j}$ ($i \in \{1, \cdots, N\}, j \in \{1, \cdots, C\}$) be the membership by which x_i belongs to the j-th cluster. The set of $u_{i,j}$ is denoted by $u \in \mathbb{R}^{N \times C}$, and this is called the partition matrix.

K-CM is obtained by solving the following optimization problem:

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2 \tag{1}$$

$$\text{subject to } \sum_{j=1}^{C} u_{i,j} = 1 \tag{2}$$

with $u_{i,j} \in \{0, 1\}$ for K-HCM,

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j}^m \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2 \tag{3}$$

$$\tag{4}$$

subject to Eq. (2) for K-sFCM, and

$$\underset{u,W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2 + \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \log(u_{i,j}) \tag{5}$$

subject to Eq. (2) for K-eFCM. Generally, Φ cannot be given explicitly, and a kernel function $\mathcal{K} : x \times x \to \mathbb{R}$ is assumed to be given; this function describes the inner product value on the feature space by pairs of elements in the data set

$$\mathcal{K}(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle. \tag{6}$$

However, it can be interpreted that Φ can be given explicitly by letting $\mathbb{H} = \mathbb{R}^N$, $\Phi(x_i) = e_i$, the ℓ-th element of which is $\delta_{i,\ell}$ of Kronecker's delta, and by introducing $K \in \mathbb{R}^{N \times N}$ such that

$$K_{i,j} = \langle \Phi(x_i), \Phi(x_j) \rangle. \tag{7}$$

From this discussion, K-CM is given by the following algorithm.

Algorithm 1 (K-CM)

Step 1. *Fix $m > 1$ for K-sFCM and $\lambda > 0$ for K-eFCM. Assume a kernel matrix $K \in \mathbb{R}^{N \times N}$ and an initial partition matrix u.*

Step 2. *Update cluster centers as*

$$W_j = (u_{1,j}, \cdots, u_{N,j})^{\mathsf{T}} / \sum_{i=1}^{N} u_{i,j} \tag{8}$$

for K-HCM and K-eFCM, and

$$W_j = \left(u_{1,j}^m, \cdots, u_{N,j}^m\right)^{\mathsf{T}} / \sum_{i=1}^{N} u_{i,j}^m \tag{9}$$

for K-sFCM.

Step 3. *Update dissimilarity between each data and each cluster center as*

$$d_{i,j} = (e_i - W_j)^{\mathsf{T}} K (e_i - W_j). \tag{10}$$

Step 4. *Update membership as*

$$u_{i,j} = \begin{cases} 1 \ (j = \arg \quad \min_k d_{i,k}) \\ 0 \ (\mathit{Otherwise}) \end{cases} \tag{11}$$

for K-HCM,

$$u_{i,j} = 1 / \sum_{k=1}^{C} \left(\frac{d_{i,j}}{d_{i,k}} \right)^{1/(m-1)} \tag{12}$$

for K-sFCM, and

$$u_{i,j} = \frac{\exp(-\lambda d_{i,j})}{\sum_{k=1}^{C} \exp(-\lambda d_{i,k})} \tag{13}$$

for K-eFCM.

Step 5. *If (u, d, W) converge, terminate this algorithm. Otherwise, return to* Step 2.

2.2 RFCM, NERFCM, and NEFRC

RFCM, NERFCM, and NEFRC assume that the dissimilarity data matrix $R \in \mathbb{R}^{N \times N}$ is given, in which $R_{i,j}$ is the dissimilarity between the datum x_i and the datum x_k. RFCM is obtained by solving the following optimization problem:

$$\underset{u}{\text{minimize}} \sum_{j=1}^{C} \frac{\sum_{i=1}^{N} \sum_{k=1}^{N} u_{i,j}^m u_{k,j}^m R_{i,k}}{2 \sum_{t=1}^{N} u_{t,j}^m} \tag{14}$$

subject to Eq. (2). RFCM is given by the following algorithm.

Algorithm 2 (RFCM)

Step 1. *Fix $m > 1$ and assume an initial partition matrix u.*

Step 2. *Update $v_j \in \mathbb{R}^N$ as*

$$v_j = \left(u_{1,j}^m, \cdots, u_{N,j}^m \right)^{\mathsf{T}} / \sum_{i=1}^{N} u_{i,j}^m. \tag{15}$$

Step 3. *Update $d_{i,j}$ as*

$$d_{i,j} = (Rv_j)_i - v_j^{\mathsf{T}} R v_j / 2. \tag{16}$$

Step 4. *Update membership as*

$$u_{i,j} = 1 / \sum_{k=1}^{C} \left(\frac{d_{i,j}}{d_{i,k}} \right)^{1/(m-1)}. \tag{17}$$

Step 5. *If stopping criterion is satisfied, terminate this algorithm. Otherwise, return to* Step 2.

We say that a matrix $R \in \mathbb{R}^{N \times N}$ is *Euclidean* if there exists a set of points $\{y_1, \cdots, y_N\} \in \mathbb{R}^{N-1}$ such that $R_{i,j} = \|y_i - y_j\|_2^2$, and we say that R is *non-Euclidean* if no such set of points exists. A given R is not always *Euclidean*, and in this case, RFCM has a risk in that the membership cannot be calculated after $d_{i,j}$ is updated to be negative.

In order to overcome this drawback, the following revision of R, called β-spread, has been considered:

$$R_\beta = R + \beta(\mathbf{1} - E), \tag{18}$$

where β is a scalar; $\mathbf{1}$, an $N \times N$ matrix all elements of which are 1; and E, the N-dimensional unit matrix. RFCM with β-spread is given by the following algorithm of NERFCM.

Algorithm 3 (NERFCM)

Step 1. *Fix $m > 1$ and assume an initial partition matrix u. Set $\beta = 0$.*
Step 2. *Execute* Step 2 *in Algorithm 2.*
Step 3. *Update $d_{i,j}$ as*

$$d_{i,j} = (R_\beta v_j)_i - v_j^{\mathsf{T}} R_\beta v_j / 2. \tag{19}$$

Step 4. *If $d_{i,j} < 0$, update $\Delta\beta, d_{i,j}$, and β as*

$$\Delta\beta = \max\{-2d_{i,j}/\|e_i - v_j\|^2\}, \tag{20}$$

$$d_{i,j} \leftarrow d_{i,j} + \Delta\beta/2\|e_i - v_j\|^2, \tag{21}$$

$$\beta \leftarrow \beta + \Delta\beta. \tag{22}$$

Step 5. *Execute* Step 4 *in Algorithm 2.*
Step 6. *If a stopping criterion is satisfied, terminate this algorithm. Otherwise, return to* Step 2.

Modification (20)–(22) of the original RFCM algorithm calculates a reasonable underestimate of the minimal shift required to transform the current R_β into a Euclidean matrix; this shift is then implemented by updating the current $d_{i,j}$ and β.

In order to overcome the drawback of RFCM, another modification, non-Euclidean fuzzy relational clustering (NEFRC), has been proposed by solving the optimization problem in which the constraint of non-negative membership

$$u_{i,j} \geq 0 \tag{23}$$

is added to that of RFCM, that is, Eq. (14), (2). By solving this optimization problem, NEFRC is given by the following algorithm.

Algorithm 4 (NEFRC)

Step 1. *Fix $m > 1$ and assume an initial partition matrix u. Set $\beta = 0$.*
Step 2. *Calculate $a_{i,j} \in \mathbb{R}$ as*

$$a_{i,j} = \frac{m u_{i,j}^{m-2} \sum_{k=1}^N u_{k,j}^m R_{i,k}}{\sum_{k=1}^N u_{k,j}^m} - \frac{m u_{i,j}^{m-2} \sum_{\ell=1}^N \sum_{k=1}^N u_{k,j}^m u_{\ell,j}^m R_{k,\ell}}{2\left(\sum_{k=1}^N u_{k,j}^m\right)^2}. \tag{24}$$

Step 3. *Define the sets J^- and J^+ as*

$$J^- = \left\{ j \mid \frac{1/a_{i,j}}{\sum_{k=1}^C 1/a_{i,k}} < 0 \right\}, \tag{25}$$

$$J^+ = \left\{ j \mid \frac{1/a_{i,j}}{\sum_{k=1}^C 1/a_{i,k}} \geq 0 \right\}. \tag{26}$$

Step 4. *Calculate membership as*

$$u_{i,j} = \begin{cases} 0 & (j \in J^-) \\ \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}} & (j \in J^+) \end{cases} \tag{27}$$

Step 5. *If a stopping criterion is satisfied, terminate this algorithm. Otherwise, return to* Step 2.

3 RFCM Is a Case of K-sFCM

RFCM is a special case of K-sFCM. This is shown as follows. First, the update equation of v_j (15) in RFCM is described as

$$v_j = \frac{\sum_{i=1}^{N} u_{i,j}^m e_i}{\sum_{i=1}^{N} u_{i,j}^m}; \tag{28}$$

this corresponds to (9) in K-sFCM. Second, the update equation of $d_{i,j}$ (16) in RFCM is described as

$$d_{i,j} = (e_i - v_j)^\mathsf{T} \left(-\frac{1}{2} PRP \right) (e_i - v_j), \tag{29}$$

where

$$P = E - \mathbf{1}/N, \tag{30}$$

$$E : N\text{-dimensional unit matrix}, \tag{31}$$

$$\mathbf{1} : N\text{-dimensional matrix with all elements of 1}; \tag{32}$$

this corresponds to (10) in K-sFCM with the kernel matrix $K = -\frac{1}{2}PRP$. Last, the update equation of $u_{i,j}$ (17) in RFCM corresponds to (12) in K-sFCM. From the above discussions, we find that RFCM corresponds to K-sFCM with the kernel matrix $K = -\frac{1}{2}PRP$. This relationship can apply to RHCM and K-HCM, and entropy regularized RFCM and K-eFCM.

4 Indefinite K-sFCM

The kernel matrix must be positive-definite for K-CM. However, an irresponsibly introduced kernel matrix is not always positive-definite. K-sFCM with an indefinite kernel matrix has a risk in that the memberships cannot be calculated after the dissimilarity between a datum and a cluster center is updated to be negative. For example, if $C = 2$, $m > 2$, $d_{i,1} = -2$, and $d_{i,2} = 1$, then $(d_{i,1}/d_{i,2})^{1/(m-1)}$ and $(d_{i,2}/d_{i,1})^{1/(m-1)}$ are no longer real valued, and the algorithm cannot be continued. On the other hand, K-HCM and K-eFCM can be continued even if it is negative, although the meaningfulness of the obtained result is unknown. Although indefinite kernel matrices can be transformed to positive-definite ones by subtracting the minimal eigenvalue from their diagonal components, this incurs needs some computational costs.

In this section, we propose two algorithms for K-sFCM with an indefinite kernel matrix (IK-sFCM). The first one executes K-sFCM by updating the kernel matrix adaptively such that the dissimilarity between each datum and each

cluster center is non-negative; this is equivalent to the manner in which RFCM is modified into NERFCM. The membership in the second one is obtained by solving the optimization problem such that the constraint of non-negative memberships is added to the relaxed problem of K-sFCM; this is equivalent to the manner in which FANNY and NEFRC are obtained.

4.1 Indefinite K-sFCM by Revising Kernel Matrix

In this subsection, we propose an algorithm for K-sFCM with an indefinite kernel matrix (IK-sFCM), in which K-sFCM is executed by updating the kernel matrix adaptively such that the dissimilarity between each datum and each cluster center is non-negative. This derivation is equivalent to the manner in which RFCM is modified into NERFCM.

The proposed algorithm is described as follows.

Algorithm 5 (IK-sFCM by Revising Kernel Matrix)

STEP 1. *Fix $m > 1$ for K-sFCM. Assume a kernel matrix $K \in \mathbb{R}^{N\times N}$ and an initial partition matrix u. Set $\beta = 0$ and $K_0 = K$.*
STEP 2. *Execute* STEP 2 *in Algorithm 1.*
STEP 3. *Update $d_{i,j}$ as*

$$d_{i,j} = (e_i - W_j)^{\mathsf{T}} K_\beta (e_i - W_j). \tag{33}$$

STEP 4. *If $d_{i,j} < 0$, update $\Delta\beta, d_{i,j}$, and β as*

$$\Delta\beta = \max\{-d_{i,j}/\|e_i - W_j\|_2^2\}, \tag{34}$$

$$d_{i,j} \leftarrow d_{i,j} + \Delta\beta\|e_i - W_j\|^2, \tag{35}$$

$$\beta \leftarrow \beta + \Delta\beta, \tag{36}$$

$$K_\beta \leftarrow K_\beta + \Delta\beta E. \tag{37}$$

STEP 5. *Execute* STEP 4 *in Algorithm 1.*
STEP 6. *If a stopping criterion is satisfied, terminate this algorithm. Otherwise, return to* STEP 2.

We can see that K-sFCM (Algorithm 1) and this proposed algorithm are identical except for Eqs. (34)–(37) that apply whenever some negative $d_{i,j}$ is encountered. $d_{i,j}$ corresponds to the squared Euclidean distance between $\sqrt{K_\beta}\Phi(x_i)$ and $\sqrt{K_\beta}W_j$ if K_β is positive-definite. It follows that a negative value of $d_{i,j}$ implies that the feature space $\mathbb{H}$ is no longer a Euclidean space, indicating that the current value of β should be incremented by some $\Delta\beta > 0$ so that the K-sFCM iteration can be continued for the new shift value $\beta + \Delta\beta$. We consider that the definition of the increment $\Delta\beta$ in (34) is reasonable in that it provides a meaningful lower bound for the minimal increment $\Delta\beta$ required to make the new K_β positive-definite only for $(e_i - W_j)$. To see this, we rewrite the formula for $\Delta\beta$ in (34) as

$$\Delta\beta = \max\{-((e_i - W_j)^{\mathsf{T}} K_\beta (e_i - W_j))/((e_i - W_j)^{\mathsf{T}}(e_i - W_j))\}, \tag{38}$$

in which we see that $\Delta\beta$ is the maximum of CN Rayleigh quotients involving $-K_\beta$. $\Delta\beta$ defined as the maximum of the Rayleigh quotients will provide a useful underestimate of the largest eigenvalue of $-K_\beta$, that is, the smallest eigenvalue

of K_β. Underestimation is important as we want to avoid excessively adding β from the diagonal elements of K_β, because this could adversely affect both the computational complexity of the algorithm and the interpretability of clustering outputs.

Although this discussion provides some justification for Eq. (34), it is still necessary to verify that the updated $d_{i,j}$ in Eq. (35) are non-negative and correspond to $d_{i,j}$ for the newly updated β in Eq. (36). The updated $d_{i,j}$ is non-negative iff the right-hand side of Eq. (35) satisfies

$$d_{i,j} + \Delta\beta \|e_i - W_j\|_2^2 \geq 0, \tag{39}$$

that is,

$$\Delta\beta \geq -d_{i,j}/\|e_i - W_j\|_2^2. \tag{40}$$

Therefore, the non-negativity of $d_{i,j}$ follows from Eq. (34) and (40). Finally, we verify that $d_{i,j}$ in Eq. (35) is consistent with Eq. (33) for the shift $\beta + \Delta\beta$ in Eq. (36). Letting $d_{i,j}(\gamma)$ denote $d_{i,j}$ in Eq. (33) corresponding to K_γ, we can show that

$$d_{i,j}(\beta + \Delta\beta) = d_{i,j}(\beta) + \Delta\beta \|e_i - W_j\|_2^2. \tag{41}$$

In summary, modifications (34)–(37) of the original K-sFCM calculate a reasonable underestimate of the minimal shift required to transform the current K_β into a positive-definite one only for $(e_i - W_j)$; this shift is then implemented by updating the current $d_{i,j}$ and β.

4.2 Indefinite K-sFCM by Non-negative Constraint of Membership

In this subsection, we propose another algorithm for K-sFCM with an indefinite kernel matrix (IK-sFCM). In this algorithm, the membership is obtained by solving the optimization problem in which the constraint of negative memberships is added to the relaxed one of K-sFCM. This derivation is equivalent to the manner in which FANNY and NEFRC are obtained. We note that we cannot use this derivation for K-HCM and K-eFCM because the obtained algorithms with the constraint of non-negative memberships are the same as the respective original ones.

First, we describe the proposed algorithm and then, we discuss its derivation. The proposed algorithm is described as follows.

Algorithm 6 (IK-sFCM with Non-negative Membership Constraint)

Step 1. *Fix $m > 1$ and assume a kernel matrix $K \in \mathbb{R}^{N\times N}$ and an initial partition matrix u.*

Step 2. *Execute* Step 2 *in Algorithm 7.*

Step 3. *Execute* Step 3 *in Algorithm 7.*

Step 4. *Calculate $a_{i,j} \in \mathbb{R}$ as*

$$a_{i,j} = m d_{i,j} u_{i,j}^{m-2}. \tag{42}$$

Step 5. *Define the sets J^- and J^+ as*

$$J^- = \left\{ j \mid \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}} \leq 0 \right\}, \tag{43}$$

$$J^+ = \left\{ j \mid \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}} > 0 \right\}. \tag{44}$$

Step 6. *Update the membership* $u_{i,j}$ *as*

$$u_{i,j} = \begin{cases} 0 & \text{for} \quad j \in J^-, \\ \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}} & \text{for} \quad j \in J^+. \end{cases} \tag{45}$$

Step 7. *If a stopping criterion is satisfied, terminate this algorithm. Otherwise, return to* Step 2.

We can see that K-sFCM (Algorithm 1) and this proposed algorithm are identical except for equations for updating the memberships, Eqs. (12) in K-sFCM and Eqs. (42)–(45) in Algorithm 6. Eqs. (42)–(45) are obtained by solving the following optimization problem:

$$\underset{u}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}^{m}(\Phi(x_i) - W_j)^{\mathsf{T}} K(\Phi(x_i) - W_j) \quad \text{subject to } u_{i,j} \geq 0 \tag{46}$$

and Eq. (2). If K is positive-definite and the non-negative condition of u is deleted, the above optimization problem corresponds to that for K-sFCM. The Lagrange function L for this optimization problem is given as

$$\begin{aligned} L(u) =& \sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}^{m}(\Phi(x_i) - W_j)^{\mathsf{T}} K(\Phi(x_i) - W_j) \\ & - \sum_{i=1}^{N} \gamma_i \left(\sum_{j=1}^{C} u_{i,j} - 1 \right) - \sum_{i=1}^{N}\sum_{j=1}^{C} \psi_{i,j} u_{i,j}, \end{aligned} \tag{47}$$

where γ and ψ are Karush-Kuhn-Tucker vectors. The optimal condition is described as

$$\frac{\partial L}{\partial u_{i,j}} = 0, \tag{48}$$

$$\sum_{j=1}^{C} u_{i,j} = 1, \tag{49}$$

$$\psi_{i,j} \geq 0, \tag{50}$$

$$\psi_{i,j} u_{i,j} = 0. \tag{51}$$

The optimal $u_{i,j}$ is obtained by fixing $i \in \{1, \cdots, N\}$ as follows. By introducing $a_{i,j} \in \mathbb{R}$ as

$$a_{i,j} = m u_{i,j}^{m-2} (\Phi(x_i) - W_j)^{\mathsf{T}} K(\Phi(x_i) - W_j), \tag{52}$$

condition (48) is given as

$$u_{i,j} a_{i,j} - \gamma_i - \psi_{i,j} = 0, \tag{53}$$

from which we have $u_{i,j}$ as

$$u_{i,j} = \frac{\gamma_i}{a_{i,j}} + \frac{\psi_{i,j}}{a_{i,j}}. \tag{54}$$

Condition (49) determines γ_i, and $u_{i,j}$ is given as

$$u_{i,j} = \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}} - \frac{\sum_{k=1}^{C} \psi_{i,k}/a_{i,k}}{a_{i,j} \sum_{k=1}^{C} 1/a_{i,k}} + \frac{\psi_{i,j}}{a_{i,j}}. \tag{55}$$

If $\psi_{i,j} = 0$ for all $j \in \{1, \cdots, C\}$, we have

$$u_{i,j} = \frac{1/a_{i,j}}{\sum_{k=1}^{C} 1/a_{i,k}}. \tag{56}$$

If $\psi_{i,j} > 0$ for at least one $j \in \{1, \cdots, C\}$, we have $u_{i,j} = 0$ from condition (51). Let the two sets J^- and J^+ be defined as

$$J^- = \{j \mid u_{i,j} = 0\}, \tag{57}$$

$$J^+ = \{j \mid u_{i,j} > 0\}. \tag{58}$$

$\psi_{i,j}$ for $j \in J^-$ is given as

$$\psi_{i,j} = \frac{\psi_{i,j} \sum_{k \in J^-} / a_{i,k} - 1}{\sum_{k \in J^-} 1/a_{i,k} + \sum_{k \in J^+} 1/a_{i,k}}, \tag{59}$$

and we solve for $\psi_{i,j}$ to obtain

$$\psi_{i,j} = -\frac{1}{\sum_{k \in J^-} 1/a_{i,k}} \text{ for } j \in J^-. \tag{60}$$

Using this result, $u_{i,j}$ for J^+ is obtained as

$$u_{i,j} = \frac{1/a_{i,j}}{\sum_{k \in J^+}^{C} 1/a_{i,k}}. \tag{61}$$

The above discussion is summarized into Eqs. (42)–(45) in Algorithm 6.

On the other hand, the updating equation of W_j is not derived from any optimization problem but only follows Eq. (9) in Algorithm 1. Even if we consider the following optimization problem for W_j as

$$\underset{W}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j}^m (\Phi(x_i) - W_j)^\mathsf{T} K (\Phi(x_i) - W_j), \tag{62}$$

no optimal solutions for W_j exist because K is not positive-definite and this objective function is not convex. Therefore, we consider that the proposed algorithm lacks for some theoretical background, and this must be investigated in our future work.

We can see an analogy between Algorithm 6 and Algorithm 4 in that Eqs. (43)–(45) and Eqs. (25)–(27) are equivalent and that the three decomposed equations of Eq. (24) with two intermediate variables $\tilde{W}_j$ and $\tilde{d}_{i,j}$ as

$$a_{i,j} = m \tilde{d}_{i,j} u_{i,j}^{m-2}, \tag{63}$$

$$\tilde{d}_{i,j} = (e_i - \tilde{W}_j)^\mathsf{T} (\frac{1}{2} PRP)(e_i - \tilde{W}_j), \tag{64}$$

$$\tilde{W}_j = \frac{\sum_{k=1}^{N} u_{i,j}^m e_k}{\sum_{k=1}^{N} u_{i,j}^m}, \tag{65}$$

correspond to Eq. (42), (10), and (9), respectively,

5 Numerical Example

In this section, we compare our two proposed algorithms with K-sFCM to divide the simple data set, constructed by 11 elements in the two dimensional Euclidean space, shown in Fig. 1 into two clusters. We can expect a correct clustering to identify of the left five points as one cluster, the right five points as another cluster, and the other point in the midst as in-between.

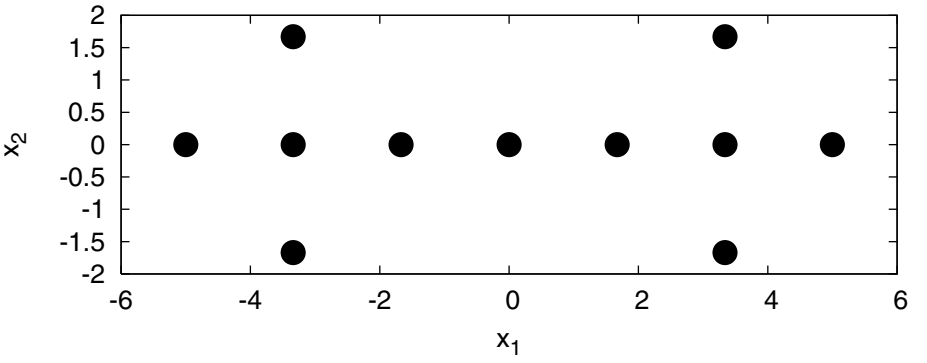


Fig. 1. Data

We applied our two proposed algorithm and K-sFCM to this problem with the fuzzifier parameter $m = 2$ using the following three kernel matrices

$$K^{(\ell)} = -\frac{1}{2} P R^{(\ell)} P + \alpha_\ell E \quad (\ell \in \{1, 2, 3\}), \tag{66}$$

where

$$R^{(1)}_{i,\tilde{i}} = \|x_i - x_{\tilde{i}}\|_2^2, \tag{67}$$

$$R^{(2)}_{i,\tilde{i}} = R^{(3)} = \|x_i - x_{\tilde{i}}\|_1^2, \tag{68}$$

$$\alpha_1 = \alpha_2 = 0, \tag{69}$$

$$\alpha_3 = 48. \tag{70}$$

The maximal and minimal eigenvalues of $K^{(\ell)}$ are shown in Table. 1, from which we can find that $K^{(1)}$ is positive-semidefinite, $K^{(2)}$ is indefinite, $K^{(3)}$ is positive-semidefinite by subtracting the minimal eigenvalue from the diagonal components of $K^{(2)}$.

Table 1. Maximal and minimal eigenvalues, and definiteness of $K^{(\ell)}$

	$K^{(1)}$	$K^{(2)}$	$K^{(3)}$
$\lambda_{\max}$	212	278	326
$\lambda_{\min}$	0	-48	0
definiteness	positive-semidefinite	indefinite	positive-semidefinite

In the result with each algorithm and with each kernel matrix from the initial partition matrix

$$u = \begin{pmatrix} 0.75\ 0.75\ 0.75\ 0.75\ 0.75\ 0.25\ 0.25\ 0.25\ 0.25\ 0.25\ 0.25 \\ 0.25\ 0.25\ 0.25\ 0.25\ 0.25\ 0.75\ 0.75\ 0.75\ 0.75\ 0.75\ 0.75 \end{pmatrix}, \tag{71}$$

the membership value $u_{1,1}$ for each case is described in Table 2. From these results, we find the following things:

1. Although K-sFCM can produce results for positive-semidefinite kernel matrices, $K^{(1)}$ and $K^{(3)}$, it cannot be continued for indefinite one, $K^{(2)}$. This exemplifies our motivation of this paper.

2. Even if the given kernel matrix $K^{(2)}$ is indefinite, K-sFCM can produce a result by adjusting $K^{(2)}$ into $K^{(3)}$, subtracting the minimal eigenvalue -48 from the diagonal components of $K^{(2)}$. However, such the obtained memberships are fuzzier. This shows the defect of adjusting a given indefinite kernel matrix to a positive-semidefinite one by subtracting the minimal eigenvalue from the diagonal components of the given kernel matrix.
3. All three algorithms produce the same results for positive-semidefinite kernel matrices, $K^{(1)}$ and $K^{(3)}$. This exemplifies that all three algorithms are identical for positive-semidefinite kernel matrix.
4. Comparing the result of K-sFCM with $K^{(3)}$ and the ones of our first proposed algorithm (Algorithm 5) with $K^{(2)}$, our first proposed algorithm produce clearer memberships than K-sFCM by adjusting the given kernel matrix $K^{(2)}$ into $K^{(3)}$. This exemplifies that, although both algorithms adjust a given kernel matrix, our first proposed algorithm suppresses the adjusting at the minimum for the dissimilarity between each data and each cluster center to be non-negative.
5. Comparing our two proposed algorithms (Algorithm 5 and Algorithm 6) with $K^{(2)}$, both algorithms produce the same memberships. It is not unknown whether this observation can be proved for other data set and other indefinite kernel, which should be investigated as our future work.

Table 2. Obtained membership $u_{1,1}$ by our two proposed algorithms and K-sFCM

	$u_{1,1}$		
	K-sFCM (Algorithm 1)	Algorithm 5	Algorithm 6
$K^{(1)}$	0.93	0.93	0.93
$K^{(2)}$	—	0.90	0.90
$K^{(3)}$	0.75	0.75	0.75

6 Conclusion

In this paper, we proposed two types of kernel fuzzy c-means clustering algorithms with an indefinite kernel matrix. Both algorithms are based on the fact that RFCM is a special case of K-sFCM. Following the manner in which RFCM is modified to NERFCM, the first proposed algorithm executes K-sFCM by adaptively updating the kernel matrix such that the dissimilarity between each datum and each cluster center is non-negative. The membership in the second one is obtained by solving the optimization in which the constraint of non-negative memberships is added to the relaxed one of K-sFCM; this follows the optimization problems of FANNY and NEFRC.

In our future work, (1) we intend to investigate the proposed algorithms with K-HCM and K-eFCM for an indefinite kernel matrix because the meaningfulness of the obtained result is unknown despite the fact that they can be continued even if it is negative; (2) we will experiment the performance og the proposed methods with bigger and more complicated data; and (3) we will apply the proposed methods to semi-supervised fuzzy clustering methods [12].

References

1. Bezdek, J.C.: 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. Miyamoto, S., Nakayama, Y.: Algorithms of Hard c-Means Clustering Using Kernel Functions in Support Vector Machines. JACIII 7(1), 19–24 (2003)
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. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
6. Miyamoto, S., Kawasaki, Y., Sawazaki, K.: An Explicit Mapping for Kernel Data Analysis and Application to Text Analysis. In: Proc. IFSA-EUSFLAT 2009, pp. 618–623 (2009)
7. Miyamoto, S., Sawazaki, K.: An Explicit Mapping for Kernel Data Analysis and Application to c-Means Clustering. In: Proc. NOLTA 2009, pp. 556–559 (2009)
8. Hathaway, R.J., Davenport, J.W., Bezdek, J.C.: Relational Duals of the c-means Clustering Algorithms. Pattern Recognition 22(2), 205–212 (1989)
9. Hathaway, R.J., Bezdek, J.C.: NERF C-means: Non-Euclidean Relational Fuzzy Clustering. Pattern Recognition 27, 429–437 (1994)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
11. Davé, R.N., Sen, S.: Robust Fuzzy Clustering of Relational Data. IEEE Trans. on Fuzzy Systems 10(6), 713–727 (2002)
12. Kanzawa, Y., Endo, Y., Miyamoto, S.: Some Pairwise Constrained Semi-Supervised Fuzzy c-Means Clustering. LNCS (LNAI), vol. 5681, pp. 268–281. Springer, Heidelberg (2009)

Algorithms in Sequential Fuzzy Regression Models Based on Least Absolute Deviations

Hengjin Tang[1] and Sadaaki Miyamoto[2]

[1] Graduate School of Systems and Information Engineering
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
tang@soft.risk.tsukuba.ac.jp
[2] Department of Risk Engineering, Faculty of Systems and Information Engineering
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
miyamoto@risk.tsukuba.ac.jp

Abstract. The method of fuzzy c-regression models is known to be useful in real applications, but there are two drawbacks. First, the results have a strong dependency on the predefined number of clusters. Second, the method of least squares is frequently sensitive to outliers or noises. To avoid these drawbacks, we apply a method of sequentially extracting one cluster at a time using noise-detecting method to fuzzy c-regression models which enables an automatic determination of clusters. Moreover regression models are based on least absolute deviations (FCRMLAD) which are known to be robust to noises. We show the effectiveness of the proposed method by using numerical examples.

Keywords: sequential clustering, fuzzy clustering, fuzzy c-regression models, least absolute deviations.

1 Introduction

Two major data analysis problems are known to be data classification and regression. The former has the output of categories whereas the latter has continuous output. There is another class of problems of data clustering, which is related to classification problems in the sense that the result is classes although it does not assume any output variable.

An interesting application of clustering and regression is the combination of the both, which is known to be c-regression problem, that is, the method should output c regression models of which each model has a cluster of data. A best-known method for this is fuzzy c-regression models (FCRM) [7] which is a variation of fuzzy c-means (FCM) [1], [9].

A drawback in this method is that we have to specify the number of clusters beforehand and the result strongly depends on that number. In the case of data clustering, one of the authors [3], [5] proposed an algorithm of sequential extraction of clusters using a noise-detecting method based on the idea of Davé and Krishnapuram [2].

In this paper we focus on c-regression models and apply this idea. Moreover it has been noted that regression models based on the least absolute deviation

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 129–139, 2010.

(LAD) are more robust to noises than the ordinary least square method. We hence propose two algorithms for LAD c-regression models. One uses the linear programming, whereas the other is far more efficient. Instead, the second algorithm is for scalar-valued independent variable alone, while the first can be used for vector-valued independent variables.

The proposed method of sequential fuzzy c-regression models based on least absolute deviations is called SFCRMLAD for simplicity. To show effectiveness of the proposed method, we compare it with fuzzy c-regression models based on least squares (called FCRMLS), fuzzy c-regression models based on least absolute deviations (called FCRMLAD), and sequential fuzzy c-regression models based on least squares (SFCRMLS) by an illustrative example.

2 Fuzzy c-Regression Models

Fuzzy c-Regression Models (FCRM), which is to obtain clusters and corresponding regression models, have been proposed by Hathaway and Bezdek [7]. We assume data set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ in which $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{R}^p$ are data of the independent variable $\boldsymbol{x}$ and $y_1, \ldots, y_n \in \boldsymbol{R}$ are those of the dependent variable y.

Our aim is to determine the c regression models:

$$y = f_i(\boldsymbol{x}; \boldsymbol{\beta}_i) + e_i, \quad i = 1, \ldots, c \tag{1}$$

and assume the regression models to be linear:

$$f_i(\boldsymbol{x}; \boldsymbol{\beta}_i) = \sum_{j=1}^{p} \beta_{ij} x_j + \beta_{i,p+1} \tag{2}$$

for simplicity. We put

$$\boldsymbol{z}_k = (\boldsymbol{x}_k, 1)^T = (x_{k1}, \ldots, x_{kp}, 1)^T \tag{3}$$

$$\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{i,p+1}) \tag{4}$$

in order to simplify the derivation.

In this paper, we use two different criteria: least squares (LS) and least absolute deviations (LAD). We call FCRM based on LS and LAD for Fuzzy c-Regression Models Based on Least Squares (FCRMLS) and Fuzzy c-Regression Models Based on Least Absolute Deviations (FCRMLAD), respectively.

2.1 Fuzzy c-Regression Models Based on Least Squares

Fuzzy c-Regression Models Based on Least Squares (FCRMLS) use the next dissimilarity between y_k and $f_i(\boldsymbol{x}_k; \boldsymbol{\beta}_i)$

$$D_{ki} = (y_k - f_i(\boldsymbol{x}_k; \boldsymbol{\beta}_i))^2 \tag{5}$$

and consider the next objective function:

$$J_{FCRMLS}(U, B) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ki})^m D_{ki} \tag{6}$$

where u_{ki} is the membership parameter which represents the belongingness of $(\boldsymbol{x}_k, y_k)$ against regression model i, m is the fuzzify parameter, U is a membership matrix and $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_c)$ is a regression parameter.

The algorithm of FCRM is following:

FCRMLS Algorithm
FCRMLS1: Set the initial value U
FCRMLS2: Repeat calculation B and U as solutions of alternative optimization of (6) until convergence.
End of FCRMLS

The initial values of U are randomly generated using uniform distribution on a unit interval. An alternative optimization means that one of U and B is fixed and the objective function is minimized with respect to the other variable, which is used throughout various fuzzy clustering algorithms.

The optimal solutions for U and B are as follows:

$$u_{ki} = \frac{(1/D_{ki})^{\frac{1}{m-1}}}{\sum_{j=1}^{c}(1/D_{kj})^{\frac{1}{m-1}}}, \tag{7}$$

$$\boldsymbol{\beta}_i = (\sum_{k=1}^{n}(u_{ki})^m \boldsymbol{z}_k \boldsymbol{z}_k^T)^{-1}(\sum_{k=1}^{n}(u_{ki})^m y_k \boldsymbol{z}_k). \tag{8}$$

2.2 Fuzzy c-Regression Models Based on Least Absolute Deviations for Vector-Valued Independent Variables

The method based on LAD requires more computation than LS but is known to have robustness [4], [6], [10]. Notice that when we compare the LS to the squared Euclidian distance, the LAD can be compared to the L_1 metric (Manhattan distance).

Fuzzy c-Regression Models Based on Least Absolute Deviations (FCRMLAD) uses the next dissimilarity between y_k and $f_i(\boldsymbol{x}; \boldsymbol{\beta}_i)$:

$$D_{ki} = |y_k - f_i(\boldsymbol{x}_k; \boldsymbol{\beta}_i)| \tag{9}$$

and consider the next objective function:

$$J_{FCRMLAD}(U, B) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m D_{ki}. \tag{10}$$

The optimal solution for U is same as that of FCRMLS. In the case of FCRMLAD, we have to solve a linear programming problem to obtain the optimal B. Since there is no guarantee that β_{ij} is non-negative, we put

$$\beta_{ij} = \beta_{ij}^{+} - \beta_{ij}^{-} \tag{11}$$

where β_{ij}^{+} and β_{ij}^{-} $(i = 1, \ldots, c, \quad j = 1, \ldots, p+1)$ are non-negative variables.

Thus the minimization of $J_{FCRMLAD}$ is equivalent to the following linear programming problem:

$$\begin{aligned}
&\min \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ki})^m r_{ki} \\
&y_k - \sum_{j=1}^{p+1}(\beta_{ij+} - \beta_{ij-})z_{kj} \leq r_{ki} \\
&y_k - \sum_{j=1}^{p+1}(\beta_{ij+} - \beta_{ij-})z_{kj} \geq -r_{ki} \\
&r_{ki},\ \beta_{ij+},\ \beta_{ij-} \geq 0
\end{aligned}$$

where the variables are r_{ki}, β_{ij}^{+} and β_{ij}^{-} $(i = 1, \ldots, c, \quad k = 1, \ldots, n, \quad j = 1, \ldots, p+1)$.

2.3 Fuzzy *c*-Regression Models Based on Least Absolute Deviations for Scalar-Valued Independent Variables

If the independent variable x is scalar-valued, the optimal solution of B can be solved by a more efficient algorithm based on an old idea of Boscovich (see, e.g., [6]).

Since each regression equation is independent, we optimize them separately and consider the next objective function:

$$\min \sum_{k=1}^{n}(u_{ki})^m \mid y_k - \beta_{i1}x_k - \beta_{i2} \mid \tag{12}$$

We assume

$$\sum_{k=1}^{n}(u_{ki})^m(y_k - \beta_{i1}x_k - \beta_{i2}) = 0 \tag{13}$$

which means the sum of the residuals between the data set and the estimated regression equation is zero. We put

$$\bar{x}_i = \frac{\sum_{k=1}^{n}(u_{ki})^m x_k}{\sum_{k=1}^{n}(u_{ki})^m} \tag{14}$$

$$\bar{y}_i = \frac{\sum_{k=1}^{n}(u_{ki})^m y_k}{\sum_{k=1}^{n}(u_{ki})^m} \tag{15}$$

and substitute (14) and (15) into (12) and obtain

$$\beta_{i2} = \bar{y}_i - \beta_{i1}\bar{x}_i. \tag{16}$$

We substitute (16) into (12) and obtain

$$\min \sum_{k=1}^{n}(u_{ki})^m \mid x_k - \bar{x}_i \mid\mid \frac{y_k - \bar{y}_i}{x_k - \bar{x}_i} - \beta_{i1} \mid \tag{17}$$

We put

$$w_{ki} = (u_{ki})^m \mid x_k - \bar{x}_i \mid, \tag{18}$$

$$\alpha_{ki} = \frac{y_k - \bar{y}_i}{x_k - \bar{x}_i} \tag{19}$$

and substitute (18) and (19) into (17) and obtain

$$\min \sum_{k=1}^{n} w_{ki} \mid \alpha_{ki} - \beta_{i1} \mid . \tag{20}$$

Although (20) is not differentiable on $\boldsymbol{R}$, we extend the derivative of (20) on α_{ki}:

$$dF(\beta_{i1}) = \sum_{k=1}^{n} w_{ki} \mathrm{sgn}(\alpha_{ki} - \beta_{i1}) \tag{21}$$

where

$$\mathrm{sgn}(w) = \begin{cases} 1 & (w > 0) \\ -1 & (w < 0) \end{cases} \tag{22}$$

Thus, $dF(\beta_{i1})$ is a step function which is right continuous and monotone non-decreasing. So, the minimization element for (20) is one of α_{ki} at which $dF(\beta_{i1})$ changes its sign. More precisely, α_{ri} is the optimal solution of (20) if and only if $dF(\beta_{i1}) < 0$ for $\beta_{i1} < \alpha_{ri}$ and $dF(\beta_{i1}) \geq 0$ for $\beta_{i1} \geq \alpha_{ri}$.

The optimal solution for β_{i1} is calculated as follows:

$$\beta_{i1} = \alpha_{ri} = \frac{y_r - \bar{y}_i}{x_r - \bar{x}_i}. \tag{23}$$

We thus have the next algorithm to derive the optimal $\boldsymbol{\beta}_i$:

Algorithm: Optimization of $\boldsymbol{\beta}_i$ in a single regression model

Step1: Calculate

$$\bar{x}_i = \frac{\sum_{k=1}^{n} (u_{ki})^m x_k}{\sum_{k=1}^{n} (u_{ki})^m}, \tag{24}$$

$$\bar{y}_i = \frac{\sum_{k=1}^{n} (u_{ki})^m y_k}{\sum_{k=1}^{n} (u_{ki})^m}, \tag{25}$$

$$w_{ki} = (u_{ki})^m \mid x_k - \bar{x}_i \mid \tag{26}$$

$$\alpha_{ki} = \frac{y_k - \bar{y}_i}{x_k - \bar{x}_i}. \tag{27}$$

Step2: Rearrange α_{ki} in ascending order and store them in $\alpha_{ki}^{'}$.

Step3: Rearrange w_{ki} corresponding to the order of $\alpha_{ki}^{'}$ and store them in q_{ki}.

Step4: Calculate

$$S = -\frac{1}{2}\sum_{k=1}^{n} q_{ki}$$

$r = 0$

While$(S < 0)\{$

$\quad r = r + 1$

$\quad S = S + q_{ri}$

$\}$

$$\beta_{i1} = \alpha_{ri}^{'} \tag{28}$$

$$\beta_{i2} = \bar{y}_i - \beta_{i1}\bar{x}_i \tag{29}$$

3 Sequential Fuzzy Clustering

One of the authors has proposed different algorithms for sequential extraction of clusters [3], [5], that is, one cluster is extracted at a time and another cluster will be found from the rest of data, and the extraction continues until no sufficient data exist. Among these methods, we use an algorithm based on noise clustering [8].

Sequential fuzzy clustering (SFC) uses the next objective function:

$$J_{SFC} = \sum_{k=1}^{n}(u_{k1})^m D_{k1} + \sum_{k=1}^{n}(u_{k0})^m \delta. \tag{30}$$

Note that there are only two clusters: u_{k1} is the membership to the extracted cluster 1 and u_{k0} is the membership to the noise cluster 0; $\delta > 0$ is a parameter which represents every object has a constant dissimilarity δ from the noise cluster.

This algorithm applies a variation of noise clustering [8], [2] to extract regression models sequentially.

The optimal solution of U is calculated as follows:

$$u_{k1} = \frac{(1/D_{k1})^{\frac{1}{m-1}}}{(1/D_{k1})^{\frac{1}{m-1}} + (1/\delta)^{\frac{1}{m-1}}} \tag{31}$$

$$u_{k0} = \frac{(1/\delta)^{\frac{1}{m-1}}}{(1/D_{k1})^{\frac{1}{m-1}} + (1/\delta)^{\frac{1}{m-1}}} \tag{32}$$

and the optimal solution B for regression models is calculated as same as that in FCRM.

We put X as a data set which we aim to analyze and $C^{(t)}$ is the number t cluster which is extracted by SFC. Algorithm of SFC is following:

SFC Algorithm
SFC1: Set the initial elements of data set $X^{(0)} = X$, $t = 0$, the initial value U and B.
SFC2: Repeat alternate optimization for (30) until convergence.
SFC3: Extract cluster $C^{(t+1)}$ that belongs to the elements with $u_{k1} > 0.5$.
SFC4: Let $X^{(t+1)} = X^{(t)} - C^{(t+1)}$. If $X^{(t+1)}$ does not have sufficient elements to extract one more cluster, stop; otherwise go to **SFC2**.
End of SFC

We apply SFC to FCRMLS and FCRMLAD, and call them Sequential Fuzzy c-Regression Models Based on Least Squares (SFCRMLS) and Sequential Fuzzy c-Regression Models Based on Least Absolute Deviations (SFCRMLAD), respectively.

4 Numerical Examples

In this section, we show numerical examples of clustering for an artificial data set and a real data set called GDP data[*1]. The former is the example of data set which contain many noises, while the latter is the example of data set which contain few noises.

Figures 1 and 2 show the results using FCRMLS and FCRMLAD where two clusters are assumed. Figures 3 and 4 show the sequentially extracted clusters using SFCRMLS and figure 5 describes overall results of SFCRMLS. Figures 6 and 7 show the sequentially extracted clusters using SFCRMLAD and figure 8 describes overall results of SFCRMLAD. For noisy data, our proposed algorithm SFCRMLAD seems to have more robustness than FCRM and SFCRM, and its results are almost the same as those of FCRMLAD. Note moreover that SFCRMLAD does not need the predefined number of clusters.

Figures 9-12 show the results using FCRMLS, FCRMLAD, SFCRMLS and SFCRMLAD, respectively. For non-noisy data, SFCRMLAD seems to have almost same performance as FCRMLS, FCRMLAD and SFCRMLS.

To summarize, our proposed approach SFCRMLAD can handle both of noisy-data and non-noisy data without predefining the number of clusters. This characteristic is very important because we can have an appropriate clustering result without knowing the degree of noise in data set and its appropriate number of clusters.

[*1] This data set which is unpublished shows relation between GDP and energy consumption in Asian countries.

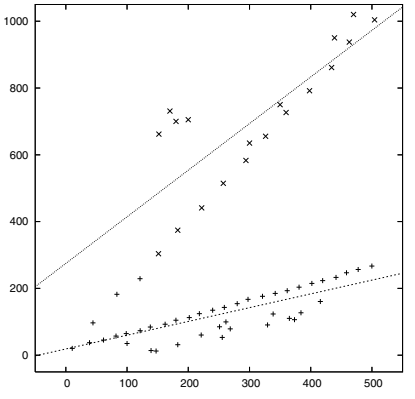


Fig. 1. Two regression models using fuzzy c-regression models based on least squares (FCRMLS), where two clusters were assumed

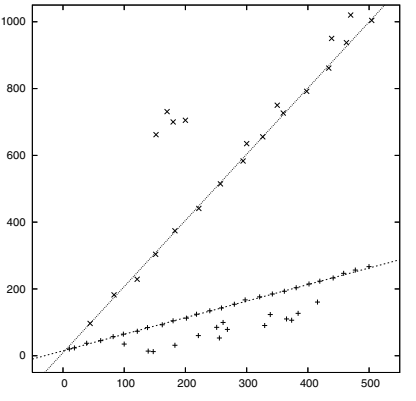


Fig. 2. Two regression models using fuzzy c-regression models based on least absolute deviations (FCRMLAD), where two clusters were assumed

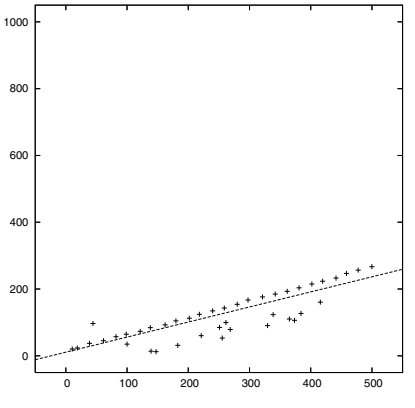


Fig. 3. First extracted cluster of SFCRMLS, $\delta = 10000$

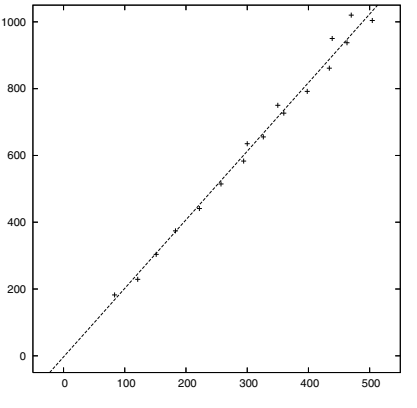


Fig. 4. Second extracted cluster of SFCRMLS, $\delta = 10000$

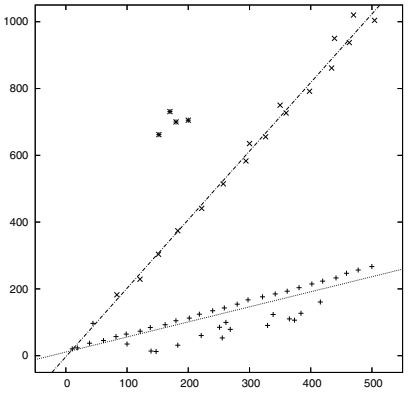


Fig. 5. Overall results of sequential extraction of clusters using SFCRLS $\delta = 10000$ where $+$, $\times$, and $*$ mean first cluster, second cluster and noise cluster, respectively

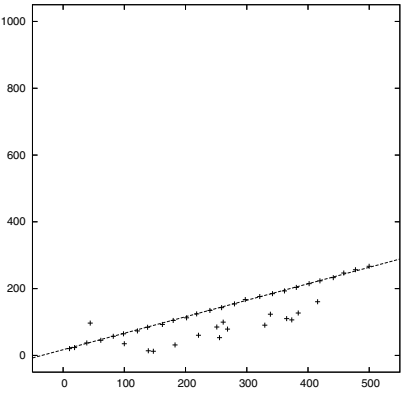


Fig. 6. First extracted cluster of SFCRMLAD, $\delta = 100$

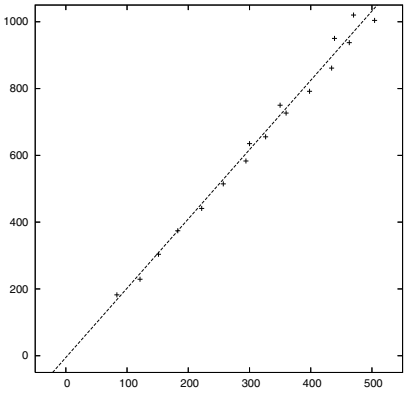


Fig. 7. Second extracted cluster of SFCRMLAD, $\delta = 100$

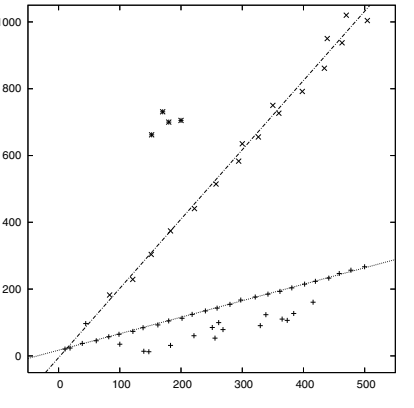


Fig. 8. Overall results of sequential extraction of clusters using SFCRMLAD $\delta = 100$ where $+$, $\times$, and $*$ mean first cluster, second cluster and noise cluster, respectively

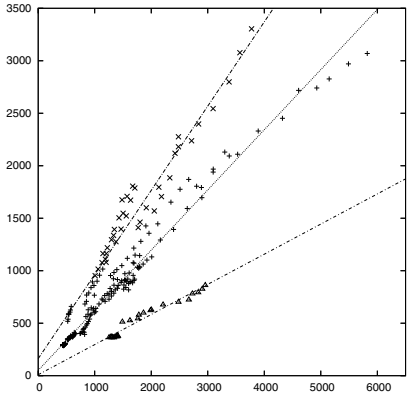


Fig. 9. Three regression models for GDP data using fuzzy c-regression models based on least squares (FCRMLS), where three clusters were assumed

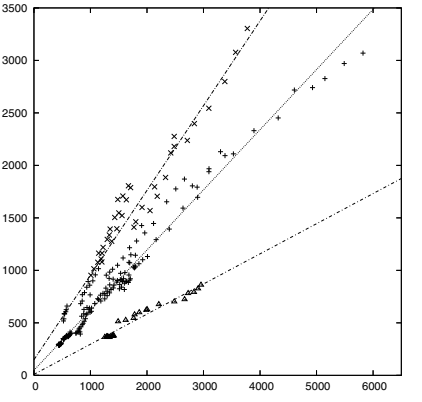


Fig. 10. Three regression models for GDP data using fuzzy c-regression models based on least absolute deviations (FCRMLAD), where three clusters were assumed

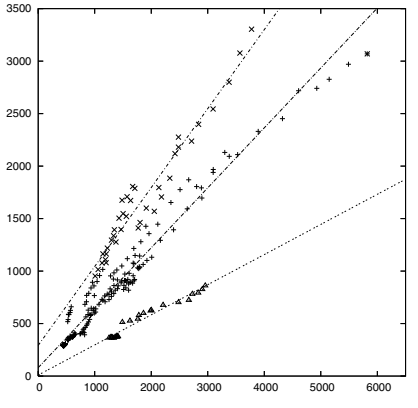


Fig. 11. Overall results of sequential extraction of clusters for GDP data using SFCRMLS $\delta = 90000$ where $+$, $\times$, $\triangle$ and $*$ mean first cluster, second cluster, third cluster and noise cluster, respectively

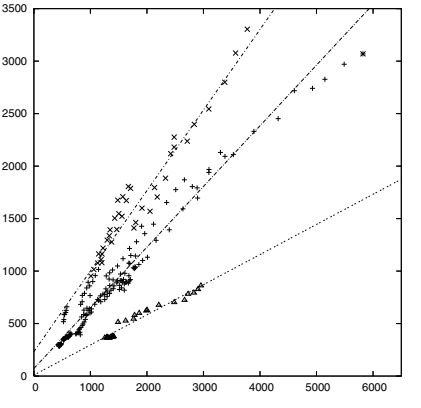


Fig. 12. Overall results of sequential extraction of clusters for GDP data using SFCRMLAD $\delta = 300$ where $+$, $\times$, $\triangle$ and $*$ mean first cluster, second cluster, third cluster and noise cluster, respectively

5 Conclusion

We have studied algorithms of sequential fuzzy clustering which relates to noise clustering [8] and we proposed algorithms for regression models based on least absolute deviations. We moreover showed their usefulness by numerical examples.

We emphasize sequential algorithms shouldn't be overlooked because they have the strong advantage of the automatic determination of the number of clusters. Moreover, since there are many variations of fuzzy clustering, we can apply sequential algorithms to those variations in clustering and that would bring us further development of clustering algorithms.

Real world problems sometimes have many data with many dimensions and complex structures. As a future work, we will apply our algorithm against the data which has more dimensions and clusters.

Acknowledgement

We thank Professor Yoji Uchiyama, University of Tsukuba for his providing GDP data.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Davé, R.N., Krishnapuram, R.: Robust clustering methods: a unified view. IEEE Trans. on Fuzzy Systems 5(2), 270–293 (1997)
3. Miyamoto, S., Arai, K.: Different Sequential Clustering Algorithms and Sequential Regression Models. In: FUZZ-IEEE 2009, Korea (August 20-24, 2009)
4. Bloomfield, P., Steiger, W.L.: Least Absolute Deviations: Theory, Applications and Algorithms. Birkhauser, Basel (1983)
5. Miyamoto, S., Kuroda, Y., Arai, K.: Algorithms for Sequential Extraction of Clusters by Possibilistic Method and Comparison with Mountain Clustering. Journal of Advanced Computational Intelligence and Intelligent Informatics 12(5) (2008)
6. Eisenhart, C.: Boscovich and the Combination of Observations. In: Whyte, L.L. (ed.) Studies of his Life and Work. Allen and Unwin (1961)
7. Hathaway, R.J., Bezdek, J.C.: Switching regression models and fuzzy clustering. IEEE Trans. on Fuzzy Systems 1, 195–204 (1993)
8. Davé, R.N.: Characterization and detection of noize in clustering. Pattern Recognition Letters 12, 657–664 (1991)
9. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering. Springer, Berlin (2008)
10. Jajuga, K.: L_1-Norm Based Fuzzy Clustering. Fuzzy Sets and Systems 39, 43–50 (1991)

A Generalized Approach to the Suppressed Fuzzy *c*-Means Algorithm

László Szilágyi[1,2], Sándor M. Szilágyi[1], and Csilla Kiss[1,3]

[1] Sapientia - Hungarian Science University of Transylvania,
Faculty of Technical and Human Science, Tîrgu-Mureş, Romania
lalo@ms.sapientia.ro
[2] Budapest University of Technology and Economics, Department of Control
Engineering and Information Technology, Budapest, Hungary
[3] Babeş-Bolyai University of Cluj-Napoca, Romania
Faculty of Mathematics and Computer Science

Abstract. Suppressed fuzzy *c*-means (s-FCM) clustering was introduced with the intention of combining the higher convergence speed of hard *c*-means (HCM) clustering with the finer partition quality of fuzzy *c*-means (FCM) algorithm. Suppression modifies the FCM iteration by creating a competition among clusters: lower degrees of memberships are reduced via multiplication with a previously set constant suppression rate, while the largest fuzzy membership grows by swallowing all the suppressed parts of the small ones. Suppressing the FCM algorithm was found successful in terms of accuracy and working time. In this paper we introduce some generalized formulations of the suppression rule, leading to an infinite number of new clustering algorithms. Based on a large amount of numerical tests performed in multidimensional environment, some generalized forms of suppression proved to give more accurate partitions than FCM and s-FCM.

Keywords: fuzzy *c*-means algorithm, suppressed fuzzy *c*-means algorithm, competitive clustering, context sensitive suppression rules.

1 Introduction

One of the first important applications of fuzzy logic [13] was the introduction of fuzzy partitions [8] in classification theory. After several more steps of evolution (e.g. Dunn [3]), Bezdek [2] 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.

Several researches have been elaborated to improve the convergence speed of FCM and to introduce modified algorithms with improved characteristics [7,10,11,12]. One of the recent such approaches was the suppressed fuzzy *c*-means algorithms (s-FCM) proposed by Fan et al. [4], having the main goal to

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 140–151, 2010.

make the convergence quicker without significantly losing from the quality of the partition. The authors introduced an extra computational step in each iteration of the FCM algorithm, aimed to strengthen the competition among clusters according to a suppression rate $\alpha \in [0, 1]$. They found that s-FCM successfully accomplished its main goals, and is insensitive to the fuzzy exponent m, but they failed to provide any evidence of the competition that stands behind s-FCM.

In a previous work [9] we have provided a detailed characterization of the competitive behavior of s-FCM, by the means of a newly introduced quasi-learning rate (QLR). In this paper, we propose some extensions to the theory of suppressed FCM by providing some generalized suppression rules based on the QLR. The rest of this paper is structured as follows. Section 2 presents the background works standing at the basis of our investigations. Section 3 introduces several types of generalized suppression rules, and provides analytical details on their characteristics. Section 4 produces a numerical analysis of the generalized suppression rules. Conclusions are given in the last section.

2 Preliminaries

2.1 Fuzzy and Hard c-Means

The conventional FCM partitions a set of object data into a number of c clusters based on the minimization of a quadratic objective function, formulated as:

$$J_{\text{FCM}} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} ||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d_{ik}^2 \ , \tag{1}$$

where $\boldsymbol{x}_k$ represents the input data ($k = 1 \dots n$), $\boldsymbol{v}_i$ represents the prototype or centroid value or representative element of cluster i ($i = 1 \dots c$), $u_{ik} \in [0, 1]$ is the fuzzy membership function showing the degree to which input vector $\boldsymbol{x}_k$ belongs to cluster i, $m > 1$ is the fuzzyfication parameter or fuzzy exponent, and $d_{ik} = ||\boldsymbol{x}_k - \boldsymbol{v}_i||$. According to the definition of fuzzy sets, for any input vector $\boldsymbol{x}_k$, we have $\sum_{i=1}^{c} u_{ik} = 1$. The minimization of the objective function is reached by alternately applying the optimization of J_{FCM} over $\{u_{ik}\}$ with $\boldsymbol{v}_i$ fixed, and the optimization of J_{FCM} over $\{\boldsymbol{v}_i\}$ with u_{ik} fixed, [5]. During each cycle, the optimal values are computed from the zero gradient conditions, and obtained as follows:

$$u_{ik}^{\star} = \frac{d_{ik}^{-2/(m-1)}}{\sum_{j=1}^{c} d_{jk}^{-2/(m-1)}} \quad \forall\, i = 1 \dots c, \forall\, k = 1 \dots n \ , \tag{2}$$

$$\boldsymbol{v}_i^{\star} = \frac{\sum_{k=1}^{n} u_{ik}^{m} \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad \forall\, i = 1 \dots c \ . \tag{3}$$

According to the alternative optimization (AO) scheme, formulae (2) and (3) are alternately applied, until cluster prototypes stabilize.

HCM is the extreme case of FCM, when $m \to 1$. In this case the partition logic is restricted to two values: 0 and 1. In every iteration, input vectors are

assigned to one cluster each, whose prototype is situated closest from the input vector. Ties are resolved arbitrarily.

Using fuzzy memberships instead of the hard ones leads to finer partitions, which unfortunately are obtained in far more iterations [2,4].

2.2 Suppressed Fuzzy *c*-Means

The suppressed fuzzy c-means algorithm was introduced by Fan et al. [4], having the declared goal of improving the convergence speed of FCM, while keeping its good classification accuracy. The proposed suppression is performed by an extra step inserted in the optimization algorithm between the application of formulae (2) and (3). After having computed the new fuzzy membership functions of vector $\boldsymbol{x}_k$ with respect to all clusters, that is the u_{ik} values ($i = 1 \dots c$), the cluster to which the vector belongs most is declared winner, and the degrees of membership are adjusted according to the following formula:

$$\mu_{ik} = \begin{cases} 1 - \alpha + \alpha u_{ik} & \text{if } i = w_k \\ \alpha u_{ik} & \text{if } i \neq w_k \end{cases} , \tag{4}$$

where w_k stands for the index of the cluster that won the competition for input vector $\boldsymbol{x}_k$, and α ($0 \leq \alpha \leq 1$) represents the so-called suppression rate. These modified memberships, which also satisfy the probabilistic constraint, will be then used instead of the u_{ik} values at the computation of the new cluster prototypes.

It is obvious, that the two extreme values of α reproduce already known algorithms. In this order, $\alpha = 1$ produces no suppression and thus we have the FCM algorithm, while $\alpha = 0$ reduces all non-winner degrees of membership to zero, causing a binary partition specific to the HCM algorithm.

Fan et al. [4] proposed using constant suppression rates in order to obtain quick convergence. They did not give a recipe for choosing a suppression rate that is optimal in any sense, or suitable for any given purpose. The proposed suppression schemes were found successful in speeding up the FCM algorithm, and keeping its accuracy in certain studied cases.

Later, Hung et al. [6] introduced some time varying suppression techniques inspired by optimal control strategies. Their approach proved successful in finding details on ophtalmological MRI images.

In a previous work [9] focusing on identifying the behavior of s-FCM, we have established a mathematical relation describing the effect of suppression. We found that the proportional suppression of non-winner memberships is mathematically equal to a virtual shortening of the distance between the winner cluster's prototype and the given input vector. This phenomenon is depicted in Fig. 1. We have characterized this phenomenon with a quasi learning rate (QLR) denoted by η_s, defined as in the case of conventional competitive algorithms:

$$\eta_s = 1 - \frac{d'_{w_k k}}{d_{w_k k}} . \tag{5}$$

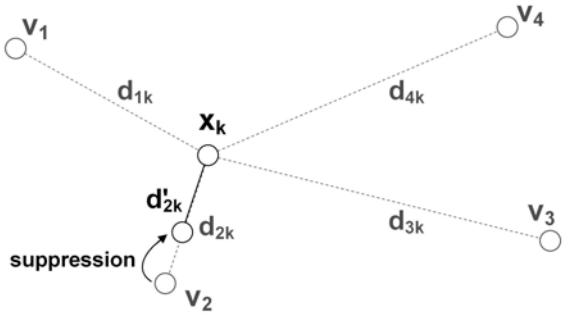


Fig. 1. The effect of suppression: cluster 2 is the winner here, so $w_k = 2$. The virtually reduced distance provides an increased membership degree to the winner cluster, while all non-winner memberships will be proportionally suppressed.

Starting from Eqs. (2), (4), and (5), we have computed the formula of the QLR and obtained the followings:

$$\eta_{\mathrm{s}} = \begin{cases} 1 - \left(1 + \frac{1-\alpha}{\alpha u_{w_k k}}\right)^{\frac{1-m}{2}} & \text{if } 0 < \alpha \leq 1 \\ 1 & \text{if } \alpha = 0 \end{cases}, \tag{6}$$

where $u_{w_k k}$ is the fuzzy membership of vector $\boldsymbol{x}_k$ with respect to its winner cluster, α is the suppression rate, and m is the fuzzy exponent. In the singularity case $\alpha = 0$, we have a winner-takes-all competition indicated by $\eta_{\mathrm{s}} = 1$, while the winner fuzzy membership cannot be zero ($u_{w_k k} \geq 1/c$).

Figure 2 exhibits some constrained plots of the QLR function, showing the behavior of the suppression operation.

3 Methods

The suppressed FCM algorithm, as proposed by Fan et al. [4], works with a constant suppression rate. Varying the suppression rate could possibly occur two different ways:

1. *Time variant suppression* means to apply a suppression rate that varies from iteration to iteration, as a function of the iteration count. This suppression rule was applied by Hung et al. in [6].
2. *Context sensitive* or *data sensitive suppression* means to define a time invariant rule of suppression, which provides a dedicated suppression rate α_k to each input vector $\boldsymbol{x}_k$ in each iteration.

Considering the fact that the most important goal of the suppression is to achieve a quicker convergence without losing the fine partitioning quality of FCM, it is not advisable to change the suppression rule in every iteration. This is why in the followings, we will exploit the possibilities of the second generalization way, namely we will define some specific suppression rules and will apply them

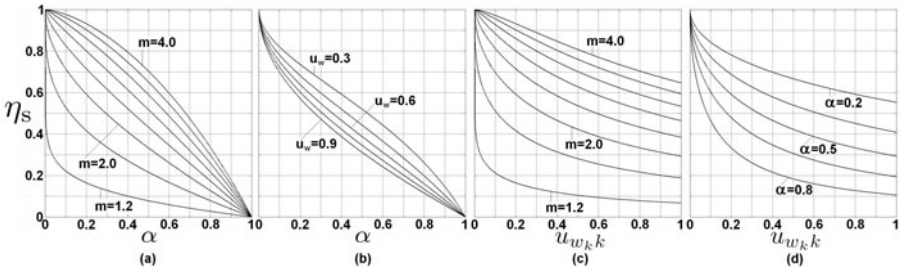


Fig. 2. Graphs of the learning rate in various circumstances: (a) with constant $u_{w_k k} = 0.8$ and different values of m, plotted against α; (b) with constant $m = 2$ and different values of winner membership $u_{w_k k}$, plotted against α; (c) with constant $\alpha = 0.5$ and different values of m, plotted against $u_{w_k k}$; (d) with constant $m = 2$ and different values of α, plotted against $u_{w_k k}$

until the convergence is achieved. All algorithms that will be proposed in the followings, will be called generalized suppressed FCM (gs-FCM), but there will be several types of them.

3.1 Proposed Suppression Rules

Constant learning rate. Let us define the first suppression rule such a way, that the QLR value is constant: for any input vector $\boldsymbol{x}_k$ in any iteration, suppression rate α_k is chosen so that $\eta_s = \theta$, where $\theta \in [0, 1]$ is a predefined constant. Obviously $\theta = 0$ implies no suppression, making the algorithm equivalent with FCM, while $\theta = 1$ brings us back to the HCM algorithm.

For any other value of θ, taking $m > 1$ and $u_{w_k k} > 0$ for granted, we may apply the first row of Eq. (6), and we obtain

$$1 - \left(1 + \frac{1-\alpha_k}{\alpha_k u_{w_k k}}\right)^{\frac{1-m}{2}} = \theta \quad \Leftrightarrow \quad \frac{1-\alpha_k}{\alpha_k u_{w_k k}} = (1-\theta)^{\frac{2}{1-m}} - 1 \ , \tag{7}$$

which then yields

$$\alpha_k = \left[1 - u_{w_k k} + u_{w_k k}(1-\theta)^{\frac{2}{1-m}}\right]^{-1} \ . \tag{8}$$

Figure 3 represents the variation of the suppression rate α_k against the winner fuzzy membership $u_{w_k k}$, under various constraints. On the left side, curves stand for different values of the parameter θ at constant fuzzy exponent $m = 2$. On the right side, plots represent α_k vs. the winner fuzzy membership for various fuzzy exponents m at constant learning rate $\theta = 0.5$. In the followings we will refer to this suppression rule and the derived algorithms as θ-type gs-FCM.

Learning rate defined as a function of the winner fuzzy membership. In this section we will suppose, that the QLR varies according to a function

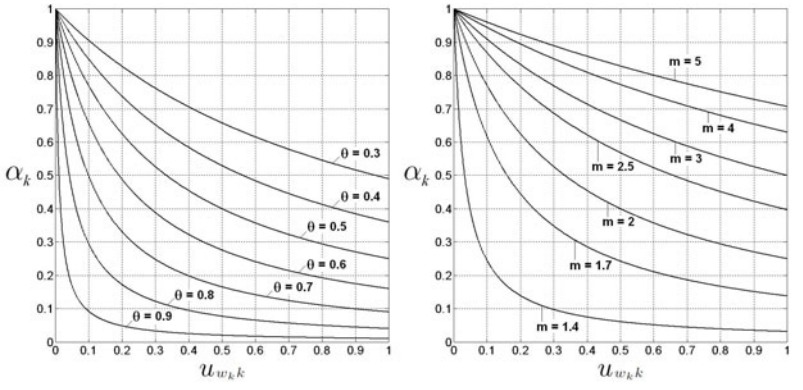


Fig. 3. Graphical representation of the α_k suppression rate vs. the value of the winner fuzzy membership: at constant fuzzy exponent $m = 2$ (left), and constant learning rate $\theta = 0.5$ (right)

of the winner membership $u_{w_k k}$: $\eta_s = f(u_{w_k k})$, where $f : [0,1] \to [0,1]$ is a continuous function. This implies:

$$1 - \left(1 + \frac{1-\alpha_k}{\alpha_k u_{w_k k}}\right)^{\frac{1-m}{2}} = f(u_{w_k k}) \Leftrightarrow \frac{1-\alpha_k}{\alpha_k u_{w_k k}} = (1 - f(u_{w_k k}))^{\frac{2}{1-m}} - 1 \ , \quad (9)$$

leading to

$$\alpha_k = \left[1 - u_{w_k k} + u_{w_k k}(1 - f(u_{w_k k}))^{\frac{2}{1-m}}\right]^{-1} \ . \quad (10)$$

For example, a learning rate that linearly decreases with the winner fuzzy membership, can be defined by $f(u_{w_k k}) = 1 - \rho u_{w_k k}$ with $0 \leq \rho \leq 1$, which gives the suppression rate

$$\alpha_k = \left[1 - u_{w_k k} + \rho^{\frac{2}{1-m}} u_{w_k k}^{\frac{3-m}{1-m}}\right]^{-1} \ . \quad (11)$$

Figure 4 exhibits the behavior of the above described example. On the left side, the variation of the quasi learning rate against the winner fuzzy membership is shown, for three different values of the parameter ρ. This is the definition of the ρ-type generalized suppression rule, which is invariant of the fuzzy exponent m. On the right side, the variation of the obtained suppression rate α_k against the winner fuzzy membership $u_{w_k k}$ is displayed, under various constraints.

In order to prove the flexibility of the generalized s-FCM algorithm, let us introduce some suppression rules derived from further $\eta_s(u_{w_k k})$ functions. Figure 5 shows the characteristics of two such algorithms: on each graph, the dotted line represents the definition function, while the continuous lines indicate the resulting suppression rules. In section 4, we will refer to these algorithms as special algorithm 1 (left) and special algorithm 2 (right).

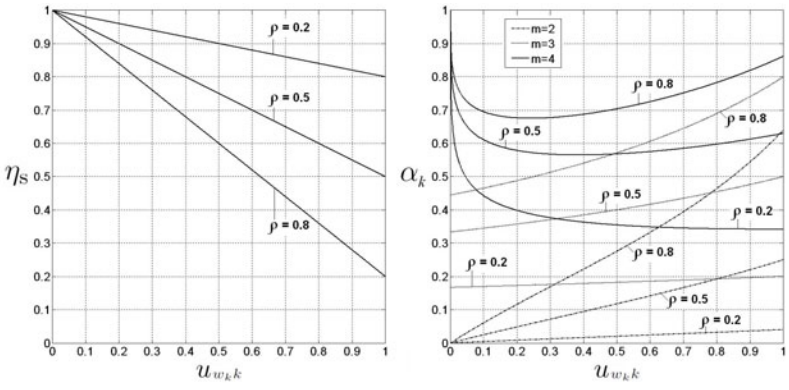


Fig. 4. Learning rate defined as a function ($\eta_s = 1 - \rho u_{w_k k}$) of the winner fuzzy membership: plot of learning rate vs. winner fuzzy membership (left), and the resulting suppression rates vs. winner fuzzy membership, captured in various circumstances (right)

Direct formula between $\mu_{w_k k}$ and $u_{w_k k}$. According to this approach, we may formulate a direct dependence rule between the winner fuzzy membership before and after suppression. For example, let us increase the winner fuzzy membership with the relativistic speed addition formula

$$\mu_{w_k k} = \frac{u_{w_k k} + \tau}{1 + u_{w_k k}\tau} \quad , \tag{12}$$

where $0 \leq \tau \leq 1$. Extreme value $\tau = 0$ makes no suppression giving the conventional FCM algorithm, while $\tau = 1$ obviously yields the HCM. In any other case, according to Eq. (4), we will have

$$1 - \alpha_k + \alpha_k u_{w_k k} = \frac{u_{w_k k} + \tau}{1 + u_{w_k k}\tau} \quad \Leftrightarrow \quad \alpha_k (1 - u_{w_k k}) = 1 - \frac{u_{w_k k} + \tau}{1 + u_{w_k k}} \quad . \tag{13}$$

In case of $u_{w_k k} = 1$ the suppression rate is irrelevant, so we get

$$\alpha_k = \frac{1 + u_{w_k k}\tau - u_{w_k k} - \tau}{(1 - u_{w_k k})(1 + u_{w_k k})} = \frac{1 - \tau}{1 + u_{w_k k}\tau} \quad . \tag{14}$$

In the followings, we will refer to the above formula as the τ-type suppression rule. Another suppression rule can be derived from the equation $\mu_{w_k k} = u^{\sigma}_{w_k k}$ with $\sigma \in [0, 1]$. Parameters τ and σ have similar effects on suppression at the boundaries of the interval: zero value means no suppression (FCM), while $\tau = 1$ or $\sigma = 1$ leads to HCM.

Figure 6 shows the characteristics of two different suppression rules based on direct formula between $\mu_{w_k k}$ and $u_{w_k k}$. The image on the left side contains two sets of curves representing plots of the learning rate against the winner fuzzy membership, in case of various choices of σ or τ, while the right side image displays plots of the obtained suppression rates against the winner fuzzy membership, under the same circumstances.

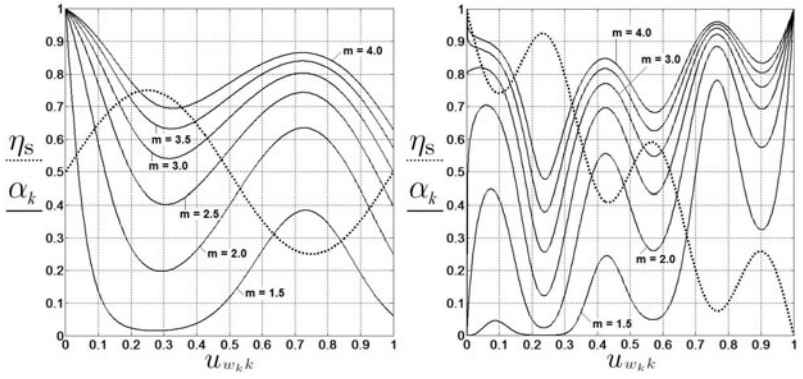


Fig. 5. Learning rate defined as functions of the winner fuzzy membership, demonstrated on two odd shaped functions: $\eta_s = (2 + \sin(2\pi u_{w_k k}))/4$ (left), and $\eta_s = (1 - u_{w_k k}) - \frac{1}{6}\sin(6\pi u_{w_k k})$ (right). On both sides, the dotted line represents η_s vs. $u_{w_k k}$, while the continuous graphs display α_k vs. $u_{w_k k}$.

3.2 Algorithm

Let us summarize the steps of the generalized s-FCM algorithm:

1. Initialize cluster prototypes with randomly chosen vectors from the input dataset, and set the value of the fuzzy exponent m.
2. Choose suppression rule and set the value of the parameters, if any (e.g. θ, ρ, σ, or τ).
3. Compute fuzzy membership with the conventional formula of FCM, Eq. (2).
4. For each input vector $\boldsymbol{x_k}$, find the winner cluster, set w_k equal to the index of the winner cluster, and compute the suppression rate α_k according to the suppression rule, with Eq. (8), (10), or (14).
5. For each input vector $\boldsymbol{x_k}$, compute suppressed fuzzy memberships with the conventional suppression formula 4, using the suppression rate α_k.
6. Update cluster prototypes using the suppressed fuzzy memberships, as in the original suppressed FCM algorithm.
7. Repeat steps 3–6 until the norm of the variation of the cluster prototypes reduces under a predefined constant ε.

4 Numerical Analysis

In the followings, we will present some numerical analysis of the functional characteristics of the conventional suppressed FCM algorithm and its generalizations. These tests were performed in multidimensional environment, using the WINE dataset [1], which consists of 178 labeled feature vectors of 13 dimensions, divided into three clusters of different sizes.

A series of numerical tests targeted the clustering accuracy. Each tested algorithm (s-FCM, gs-FCM of types θ, ρ, τ) with each parameter setting (m varied

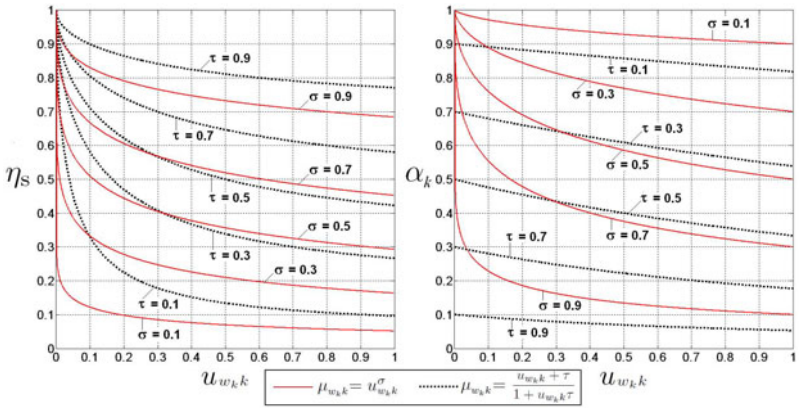


Fig. 6. Direct formula between the winner membership before and after suppression – two different suppression rules: plot of learning rate vs. winner fuzzy membership (left), and the resulting suppression rates vs. winner fuzzy membership, captured in various circumstances (right)

from 1.2 to 3.2, suppression parameter varying from 0 to 1 with step of 0.01) was performed 200 times, using previously selected random initializations of cluster prototypes. The labels of the input vectors were used as ground truth. The number of correct decisions was averaged along the 200 tests, separately for each algorithm. The obtained accuracy results are displayed in Figs. 7 and 8.

Figure 7 exhibits the average number of correct decisions produced by each algorithm, plotted against the value of the suppression parameter α, θ, ρ, or τ, at six different values of the fuzzy exponent m. In order to emphasize the correlation among the curves, two of the plots were reversed (plotted against $1 - \theta$ and $1 - \tau$, respectively). This figure reveals that gs-FCM can produce better accuracy than s-FCM, and both gs-FCM and s-FCM can perform better than the conventional FCM algorithm, when using suitably adjusted parameters.

Figure 8 displays averaged accuracy and speed results of the above mentioned four algorithms at selected values of their parameters, and the two special algorithms presented in Fig. 5, all plotted against the fuzzy exponent m. This figure suggests that the generalized suppression schemes can improve the accuracy of clustering, while the convergence is definitely quicker than in case of the conventional FCM. Comparing the performances of gs-FCM variants one can remark that more accurate solutions usually demand a few iterations more than less accurate ones. Although it seems strange, the special algorithm 2 can produce excellent partitions when using with high valued fuzzy exponent.

Table 1 makes a comparison of the proposed algorithms, showing the best averaged misclassification rates obtained by these algorithms along the variation interval of their suppression parameter. It is visible, that for any value of the fuzzy exponent, the suppression rule of type θ performed better than all others.

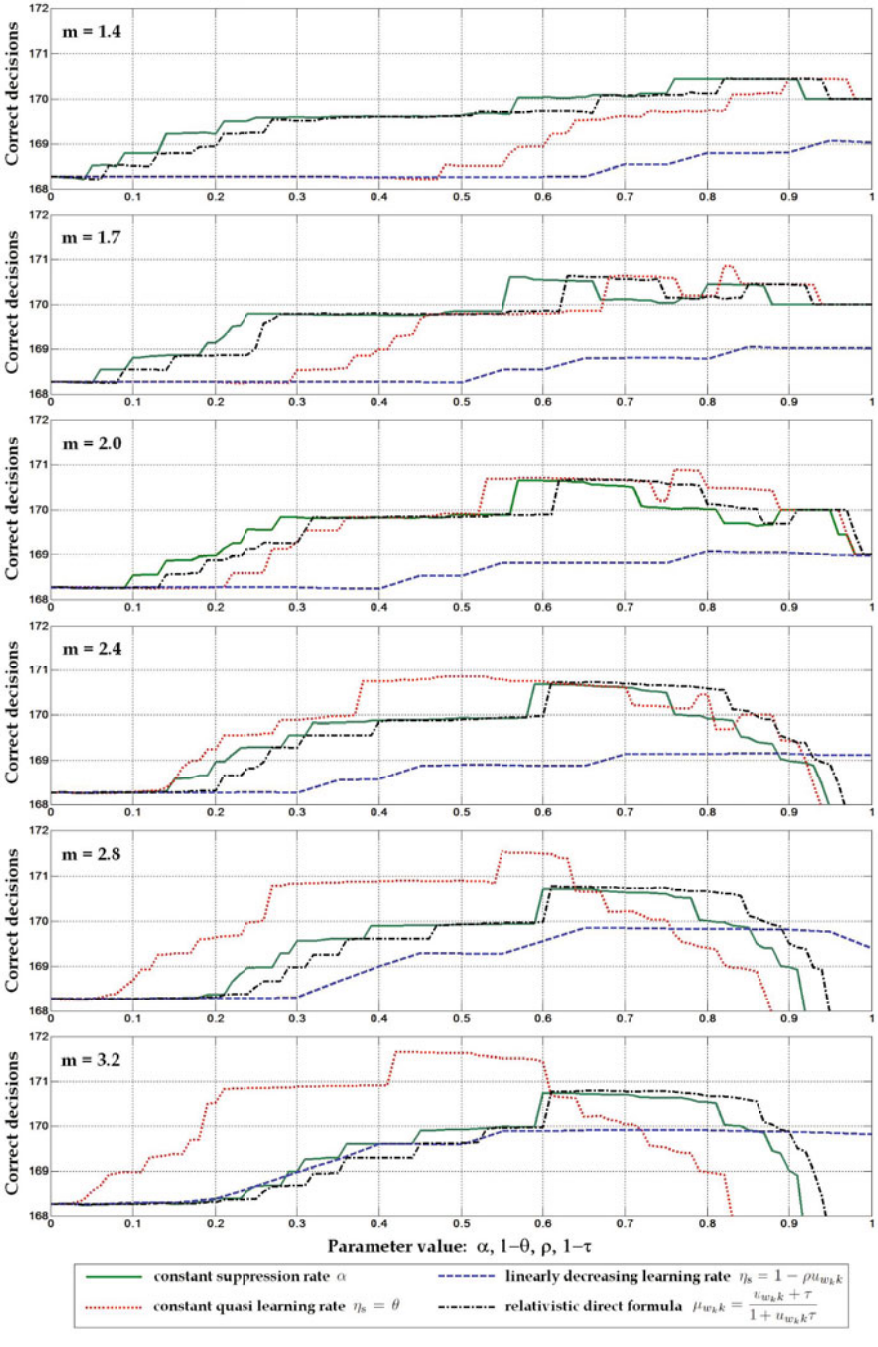


Fig. 7. The accuracy of the conventional s-FCM, compared with the generalized suppressions of type θ, ρ, and τ, at various levels of the fuzzy exponent m

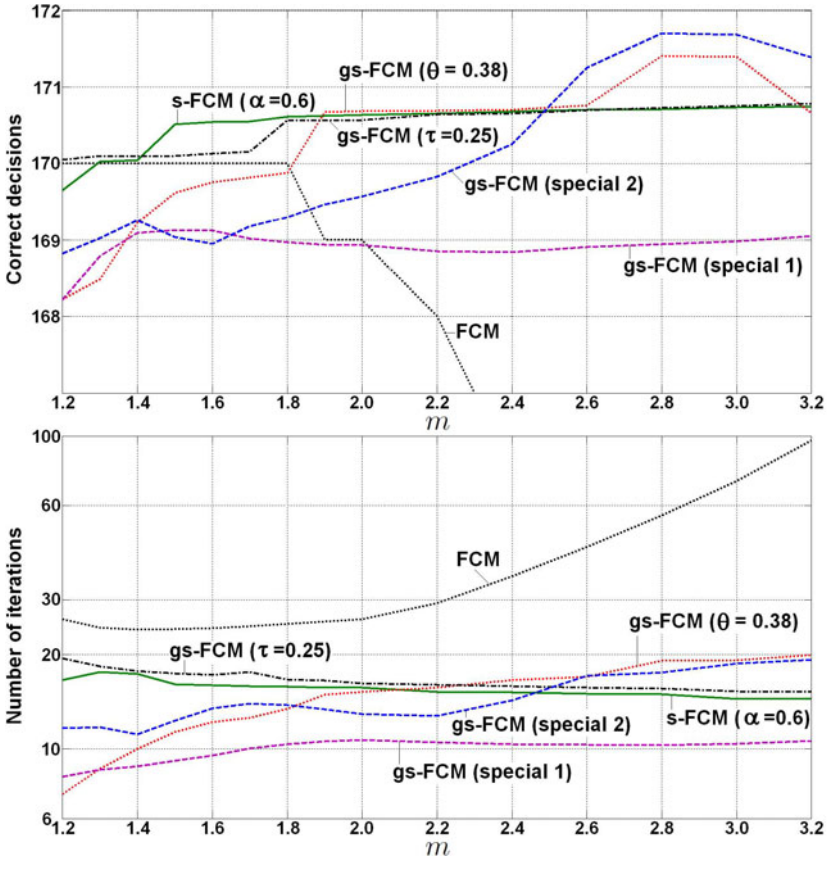


Fig. 8. Comparative evaluation of six clustering algorithms: number of correct decisions out of 178, plotted against fuzzy exponent m (top), number of necessary iterations plotted against fuzzy exponent m (bottom)

Table 1. Lowest misclassification rates of each algorithm at various fuzzy exponents

Exponent	Fuzzy	Suppressed	Generalized suppressed FCM		
m	c-means [2]	FCM [4]	θ-type	ρ-type	τ-type
1.2	4.4944 %	4.2388 %	4.2388 %	5.1826 %	4.2416 %
1.4	4.4944 %	4.2416 %	4.2360 %	5.0169 %	4.2388 %
1.6	4.4944 %	4.1826 %	4.0169 %	5.0225 %	4.1545 %
1.8	4.4944 %	4.1404 %	3.9972 %	5.0337 %	4.1320 %
2.0	5.0562 %	4.1348 %	3.9916 %	5.0169 %	4.1208 %
2.2	5.6180 %	4.1236 %	3.9775 %	4.9944 %	4.0927 %
2.4	6.7416 %	4.1124 %	4.0056 %	4.9691 %	4.0843 %
2.6	7.8652 %	4.0927 %	4.0000 %	4.7612 %	4.0702 %
2.8	8.4270 %	4.0955 %	3.6208 %	4.5843 %	4.0646 %
3.0	9.5506 %	4.0843 %	3.5758 %	4.5646 %	4.0534 %
3.2	10.674 %	4.0758 %	3.5702 %	4.5449 %	4.0449 %

5 Conclusions

In this paper we have proposed several different generalization schemes for the suppressed FCM algorithm, from which an infinite number of new clustering algorithms can be derived. Numerical tests have revealed the superiority of certain generalized forms over FCM and s-FCM. Further works will aim at testing several more suppression functions on a wider scale of test data sets, to reveal further properties of the generalized suppressed fuzzy c-means algorithm.

References

1. 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
2. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum, New York (1981)
3. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J. Cybern. 3, 32–57 (1974)
4. Fan, J.L., Zhen, W.Z., Xie, W.X.: Suppressed fuzzy c-means clustering algorithm. Patt. Recogn. Lett. 24, 1607–1612 (2003)
5. 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)
6. 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)
7. Kamel, M.S., Selim, S.Z.: New algorithms for solving the fuzzy clustring problem. Patt. Recogn. 27, 421–428 (1994)
8. Ruspini, E.H.: A new approach to clustering. Inform. Contr. 16, 22–32 (1969)
9. Szilágyi, L., Szilágyi, S.M., Benyó, Z.: Analytical and numerical evaluation of the suppressed fuzzy c-means algorithm: a study on the competition in c-means clustering models. Soft. Comput. 14, 495–505 (2010)
10. Tsao, E.C.K., Bezdek, J.C., Pal, N.R.: Fuzzy Kohonen clustering networks. Patt. Recogn. 27, 757–764 (1994)
11. Xie, Z., Wang, S., Chung, F.L.: An enhanced possibilistic c-means clustering algorithm. Soft Comput. 12, 593–611 (2008)
12. Yair, E., Zeger, K., Gersho, A.: Competitive learning and soft competition for vector quantization design. IEEE Trans. Sign. Proc. 40, 294–309 (1992)
13. Zadeh, L.A.: Fuzzy sets. Inform. Contr. 8, 338–353 (1965)

Semi-supervised Agglomerative Hierarchical Clustering Using Clusterwise Tolerance Based Pairwise Constraints

Yukihiro Hamasuna[1,2], Yasunori Endo[1], and Sadaaki Miyamoto[1]

[1] Department of Risk Engineering, Faculty of Systems and Information Engineering, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki, 305-8573, Japan
{endo,miyamoto}@risk.tsukuba.ac.jp
[2] Research Fellow of the Japan Society for the Promotion of Science
yhama@soft.risk.tsukuba.ac.jp

Abstract. Recently, semi-supervised clustering has been remarked and discussed in many researches. In semi-supervised clustering, pairwise constraints, that is, must-link and cannot-link are frequently used in order to improve clustering results by using prior knowledges or informations. In this paper, we will propose a clusterwise tolerance based pairwise constraint. In addition, we will propose semi-supervised agglomerative hierarchical clustering algorithms with centroid method based on it. Moreover, we will show the effectiveness of proposed method through numerical examples.

Keywords: semi-supervised clustering, agglomerative hierarchical clustering, centroid method, clusterwise tolerance, pairwise constraints.

1 Introduction

The aim of data mining methods is to discover important properties or knowledges from a large quantity of data. Recently, semi-supervised learning has also been remarked and discussed in many researches [1]. In the field of clustering [2,3], pairwise constraints are frequently used in order to improve clustering results by using prior knowledges or prior informations [4,5]. Also, pairwise constraints problems are considered by using probabilistic model [6], fuzzy clustering model [7]. In addition, soft constraints which are introduced as penalty terms to the objective function are another way [9,10]. In case of the methods with soft constraints, pairwise constraints are not always satisfied. These hard and soft constraints are frequently considered in semi-supervised learning methods.

In recent years, semi-supervised clustering which are based on k-means and fuzzy c-means clustering heve been widely discussed [4,7,9,10]. Also, semi-supervised clustering methods which are based on agglomerative hierarchical clustering (AHC) are discussed [11,12,13]. In these methods, pairwise constraints reffered to must-link and cannot-link are used as a prior or background knowledges about which objects should be in the same or different cluster [4]. However,

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 152–162, 2010.

because of the squared Euclidean-norm which is used as dissimilarity, it is difficult to introduce pairwise constraints in the Euclidean space. In Constrained Complete–Link(CCL) proposed by Klein [12], cannot-link constraint is handled as $d(G, G') = +\infty$. This means that a point is at the infinity, which generally breaks the Euclidean space. In order to avoid such problems, the significant methods with kernel function have been proposed [7,10]. In these methods with kernel function, pairwise constraints are considered not input space but high-dimensional feature space.

By the way, the concept of clusterwise tolerance has been proposed in order to handle different sizes or shapes of clusters [14]. This clusterwise tolerance is based on the concept of tolerance [15]. The squared Euclidean-norm is rewritten as the dissimilarity between data with clusterwise tolerance vector and cluster center. By using the concept of clusterwise tolerance, we can handle different sizes or shapes of clusters in the Euclidean space. From that sense, we propose clusterwise tolerance based pairwise constraints in order to introduce pairwise constraints into the Euclidean space in natural way. In addition, we propose semi-supervised agglomerative hierarchical clustering method based on it.

The contents of this paper are the followings. In the second section, we introduce some symbols, agglomerative hierarchical clustering algorithm (AHC) and pairwise constraints. In the third section, we propose clusterwise tolerance based pairwise constraints. In the forth section, we propose semi-supervised agglomerative hierarchical clustering using clusterwise tolerance based pairwise constraints. In the fifth section, we show the effectiveness of proposed method through numerical examples. In the last section, we conclude this paper.

2 Preparation

First, a set of data or objects to be clustered is given. A data set is denoted by $X = \{x_1, \ldots, x_n\}$ in which x_k, $(k = 1, \ldots, n)$ is an object. In most cases, $x_1, \ldots, x_n$ are vectors of real p-dimensional space $\Re^p$, that is, an object $x_k \in \Re^p$. Generally, a hard cluster is denoted by G_i is a subset of X. A set of clusters is denoted as follows:

$$\mathcal{G} = \{G_1, G_2, \ldots, G_C\},$$

where the clusters are disjoint and their union is a set of data as follows:

$$\bigcup_{i=1}^{C} G_i = X, \quad G_i \cap G_j = \emptyset \quad (i \neq j).$$

2.1 Agglomerative Hierarchical Clustering

In agglomerative hierarchical clustering (AHC), the dissimilarity denoted by $d(G, G')$ $(G, G' \in \mathcal{G})$ is used for measuring nearness between two clusters.

First, we describe a general algorithm of AHC [16,17].

Algorithm 1. AHC

AHC 1 Assume that initial clusters are given by

$$\mathcal{G} = \left\{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{N_0}\right\}.$$

Set $C = N_0$. (C is the number of clusters and N_0 is the initial number of clusters)

$$G_i = \hat{G}_i (i = 1, \ldots, C).$$

Calculate $d(G, G')$ for all pairs $G, G' \in \mathcal{G}$.

AHC2 Search the pair of minimum dissimilarity:

$$(G_p, G_q) = \arg \min_{G, G' \in \mathcal{G}} d(G, G').$$

Merge: $G_r = G_p \cup G_q$.

Add G_r to $\mathcal{G}$ and delete G_p, G_q from $\mathcal{G}$.

$C := C - 1$.

If $C = 1$ then stop and output the dendrogram. Otherwise, go to **AHC 3**.

AHC 3 Update dissimilarity $d(G_r, G'')$ for all $G'' \in \mathcal{G}$.

Go to **AHC 2**.

End AHC.

2.2 Centroid Method

In AHC procedure, there are five methods for updating dissimilarity, that is, single linkage, complete linkage, average linkage, centroid method, and ward method. Especially, centroid method and ward method are based on the Euclidean space.

In this paper, we focus centroid method described below. First, we note two definitions of centroid method, that is, the centroid of cluster and the dissimilarity between two clusters.

Let the centroid of a cluster G be

$$M(G) = \frac{1}{|G|} \sum_{x_k \in G} x_k, \tag{1}$$

and let the squared Euclidean-norm used as dissimilarity be

$$d(G, G') = \|M(G) - M(G')\|^2. \tag{2}$$

2.3 Pairwise Constraints

Typical examples of pairwise constraints are must-link and cannot-link [4]. These constraints are considered as a prior or background knowledges about which objects should be in the same or different cluster. A set $ML = \{(x_i, x_j)\} \subset X \times X$ consists of must-link pairs so that x_i and x_j should be in the same cluster, while another set $CL = \{(x_k, x_l)\} \subset X \times X$ consists of cannot-link pairs so that x_k

and x_l should be in different cluster. Obviously, *ML* and *CL* are assumed to be symmetric, that is, if $(x_i, x_j) \in ML$ then $(x_j, x_i) \in ML$, and if $(x_k, x_l) \in CL$ then $(x_l, x_k) \in CL$.

In many studies, these pairwise constraints are considered as hard constraints. This means that pairwise constraints *ML* and *CL* are always satisfied in clustering procedures and results. Many semi-supervised clustering methods based on such hard constrains have been proposed in order to improve clustering results by using background knowledges or prior informations of data sets [4,5,7].

3 Clusterwise Tolerance Based Pairwise Constraints

3.1 Clusterwise Tolerance

Each object has the tolerance κ_k which means the upper bound of clusterwise tolerance vectors. A tolerance κ_k means the admissible range of each clusterwise tolerance vector. A set of clusterwise tolerance vector is defined as $\Delta = \{\delta_{11}, \ldots, \delta_{kl}, \ldots, \delta_{nn}\}$ in which δ_{kk} is a clusterwise tolerance vector. $\delta_{11}, \ldots, \delta_{nn}$ are vectors of p-dimensional real space $\Re^p$. A clusterwise tolerance vector is the vector within the range of tolerance.

If $(x_i, x_j) \in ML$, δ_{ij} and δ_{ji} are calculated to be near each other, while $(x_k, x_l) \in CL$, δ_{kl} and δ_{lk} are calculated to be distant each other.

A constraint for clusterwise tolerance vector is as follows:

$$\|\delta_{kl}\|^2 \leq (\kappa_k)^2 \quad (\kappa_k \geq 0), {}^{\forall}k,\ l. \tag{3}$$

Figure 1 shows a clusterwise tolerance in $\Re^2$.

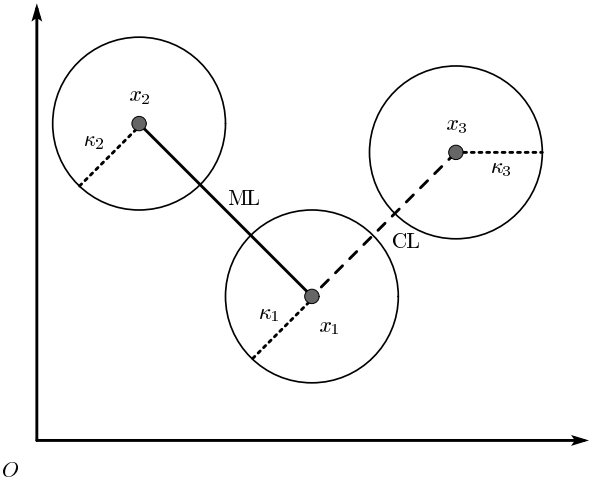


Fig. 1. An illustrative example of the concept of clusterwise tolerance

In this example, $(x_1, x_2) \in ML$ and $(x_1, x_3) \in CL$. Also, each object has tolerance. Therefore, the dissimilarity by centroid method are calculated as follows:

$$\begin{aligned} d(x_1, x_2) &= (\|x_1 - x_2\| - \kappa_1 - \kappa_2)^2, \\ d(x_1, x_3) &= (\|x_1 - x_3\| + \kappa_1 + \kappa_3)^2, \\ d(x_2, x_3) &= \|x_2 - x_3\|^2. \end{aligned}$$

3.2 Clusterwise Tolerance Based Pairwise Constraints

First, a set of must or cannot-linked objects are defined. A set $ML(G; x)$ consists of must-linked objects which in cluster G with an object x, while $CL(G; x)$ consists of cannot-linked objects which in cluster G with an object x.

$$ML(G; x) = \{\xi \mid \xi \in G, (\xi, x) \in ML\}, \tag{4}$$

$$CL(G; x) = \{\xi \mid \xi \in G, (\xi, x) \in CL\}. \tag{5}$$

In adition, $ML(G; G')$ is defined as an union of sets $ML(G; x)$, while $CL(G; G')$ is defined as an union of sets $CL(G; x)$ as follows:

$$ML(G; G') = \bigcup_{x \in G'} ML(G; x), \tag{6}$$

$$CL(G; G') = \bigcup_{x \in G'} CL(G; x). \tag{7}$$

A concept of clusterwise tolerance based pairwise constraints uses these sets in order to calculate the clusterwise tolerance which is defined between clusters.

Here, we propose clusterwise tolerance based pairwise constraints. A value of $K(G; G')$ is the sum of tolerance κ_k which in a set of must or cannot-linked objects.

$$K(G; G') = \sum_{x_k \in ML_{(G;G')}} \kappa_k - \sum_{x_l \in CL_{(G;G')}} \kappa_l. \tag{8}$$

If $K(G; G') > 0$, G is considered must-linked cluster with G', while $K(G; G') < 0$, G is considered cannot-linked cluster with G'. The upper bound of clusterwise tolerance is defined as $|K(G; G')|$. Obviously, it is depended on the value of κ_k whether G is must or cannot-linked with G'. Therefore, $K(G; G')$ and $K(G'; G)$ are asymmetric.

Next, we show an illustrative example of clusterwise tolerance based pairwise constraints. Figure 2 is a simple example of proposed method.

In this example, $(x_1, x_4), (x_3, x_5) \in ML$ and $(x_2, x_6) \in CL$. Therefore, $ML(G; G')$, $CL(G; G')$ and $K(G; G')$ are as follows:

$$\begin{aligned} ML(G; G') &= \{x_1, x_3\}, \\ CL(G; G') &= \{x_2\}, \\ K(G; G') &= \kappa_1 + \kappa_3 - \kappa_2. \end{aligned}$$

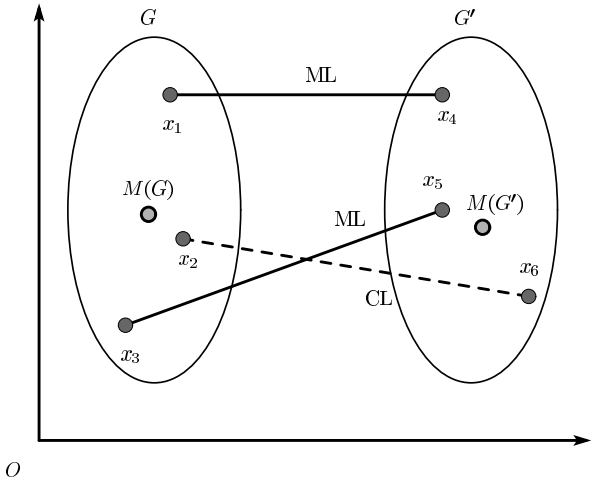


Fig. 2. An illustrative example of clusterwise tolerance based pairwise constraints

Also, $ML\,(G';G)$, $CL\,(G';G)$ and $K(G';G)$ are as follows:

$$\begin{aligned} ML\,(G';G) &= \{x_4, x_5\}\,, \\ CL\,(G';G) &= \{x_6\}\,, \\ K\,(G';G) &= \kappa_4 + \kappa_5 - \kappa_6. \end{aligned}$$

It assumed that each κ_k is the same value, $K(G;G')$ and $K(G';G)$ are both positive. This means that G and G' are must-linked clusters each other.

4 Semi-supervised Agglomerative Hierarchical Clustering Using Clusterwise Tolerance Based Pairwise Constraints

In this section, we propose semi-supervised AHC with centroid method using clusterwise tolerance based pairwise constraints (AHC–CTP). In proposed method, the centroid of each cluster is calculuted as the same procedure (1), while the dissimilarity between two clusters are as follows:

$$d\,(G,G') = \begin{cases} \left(\|M\,(G) - M\,(G')\,\| - K\,(G;G') - K\,(G';G)\right)^2 \\ \qquad (\|M\,(G) - M\,(G')\,\| > K\,(G;G') + K\,(G';G))\,, \\ 0 \qquad (\text{otherwise})\,. \end{cases} \tag{9}$$

Next, we describe an algorithm of AHC–CTP.

Algorithm 2. AHC–CTP

AHC–CTP 1 Assume that initial clusters are given by

$$\mathcal{G} = \left\{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{N_0}\right\}.$$

Set $C = N_0$. (C is the number of clusters and N_0 is the initial number of clusters)

$$G_i = \hat{G}_i (i = 1, \ldots, C).$$

Set ML, CL, κ_k and $M(G)$.
Calculate $K(G; G')$ using (8) for all pairs $G, G' \in \mathcal{G}$.
Calulate $d(G, G')$ using (9) for all pairs $G, G' \in \mathcal{G}$.

AHC–CTP 2 Search the pair of minimum dissimilarity:

$$(G_p, G_q) = \arg \min_{G, G' \in \mathcal{G}} d(G, G').$$

Merge: $G_r = G_p \cup G_q$.
Add G_r to $\mathcal{G}$ and delete G_p, G_q from $\mathcal{G}$.
$C := C - 1$.
If $C = 1$ then stop and output the dendrogram. Otherwise, go to **AHC–CTP 3**.

AHC–CTP 3 Update $ML(G_r; G'')$, $CL(G_r; G'')$, $K(G_r; G'')$, $M(G_r)$ and $d(G_r, G'')$ for all $G'' \in \mathcal{G}$.
Go to **AHC–CTP 2**.

End AHC–CTP.

5 Numerical Examples

In this section, we show numerical examples with a simple artificial data set. This data set consists of nine objects allocated in two dimensional pattern space described in Table 1. This data set should be classified into three clusters.

First, we show classification result of conventional centroid method. Fig. 3 is an illustrative example of conventional centroid method. In these classification results, '$\circ$', '$\triangle$', and '$\times$' mean each cluster.

$$G_1 = \{x_1, x_3, x_4, x_5, x_6\},$$
$$G_2 = \{x_7, x_8, x_9\},$$
$$G_3 = \{x_2\}.$$

Table 1. Data set $\{x_k \mid x_k \in \Re^p, k = 1 \sim 9\}$

k	(x_{k1}, x_{k2})
1	(2.00,3.00)
2	(2.00,7.00)
3	(2.50,5.00)
4	(4.00,5.00)
5	(5.00,5.00)
6	(6.00,5.00)
7	(7.50,5.00)
8	(8.00,3.00)
9	(8.00,7.00)

Second, we show classification result of proposed method with ML described as follows:

$$ML = \{(x_1, x_3), (x_2, x_3), (x_7, x_8), (x_7, x_9)\}.$$

Here, each object which in ML has the tolerance $\kappa_k = 1.0$. Fig. 4 is an illustrative example of these conditions. The classification result of proposed method with ML is as follows:

$$\begin{aligned} G_1 &= \{x_1, x_2, x_3\}, \\ G_2 &= \{x_4, x_5, x_6\}, \\ G_3 &= \{x_7, x_8, x_9\}. \end{aligned}$$

Third, we show classification result of proposed method with CL described as follows:

$$CL = \{(x_3, x_4), (x_6, x_7)\}.$$

Here, each object which in CL has the tolerance $\kappa_k = 1.0$. Fig. 5 is an illustrative example of this case. The classification result of proposed method with CL is as follows:

$$\begin{aligned} G_1 &= \{x_1, x_2, x_3\}, \\ G_2 &= \{x_4, x_5, x_6\}, \\ G_3 &= \{x_7, x_8, x_9\}. \end{aligned}$$

In addition, we show classification result of proposed method with ML and CL described as follows:

$$\begin{aligned} ML &= \{(x_2, x_3), (x_6, x_7)\}, \\ CL &= \{(x_1, x_8), (x_4, x_5)\}. \end{aligned}$$

Here, each object which in ML or CL has the tolerance $\kappa_k = 1.0$. Fig. 6 is an illustrative example of this case. The classification result of proposed method with ML and CL is as follows:

$$\begin{aligned} G_1 &= \{x_1, x_2, x_3, x_4\}, \\ G_2 &= \{x_5, x_6, x_7, x_8\}, \\ G_3 &= \{x_9\}. \end{aligned}$$

From these results, the effectiveness of proposed method is shown. In AHC procedure, each object is merged one by one. In our proposed method, the pairwise constraints are handled by using clusterwise tolerance based ones without breaking the Euclidean space. Moreover, the difference between proposed method and conventional one is verified. As a result, merging process which is obtained in the form of dendrogram is different from conventional centroid method. These kinds of properties are quite different from other semi-supervised clustering methods which are based on pairwise constraints.

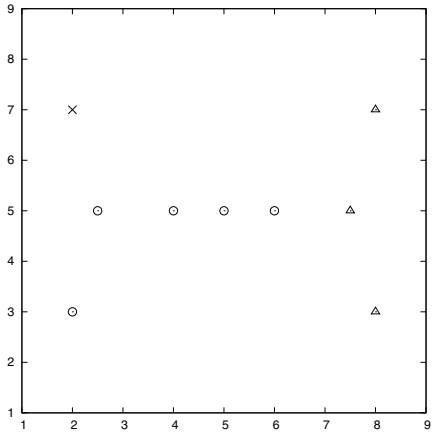


Fig. 3. Classification result of conventional centroid method

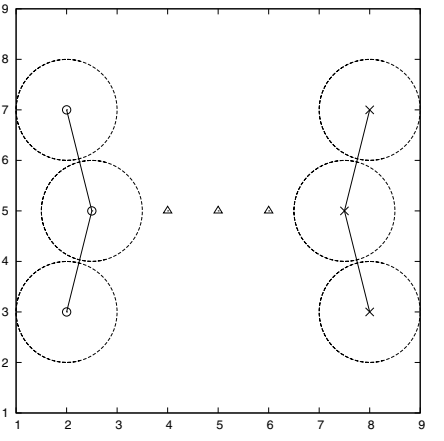


Fig. 4. Classification result with $\kappa_k = 1.0$ and $ML = \{(x_1, x_3), (x_2, x_3), (x_7, x_8), (x_7, x_9)\}$

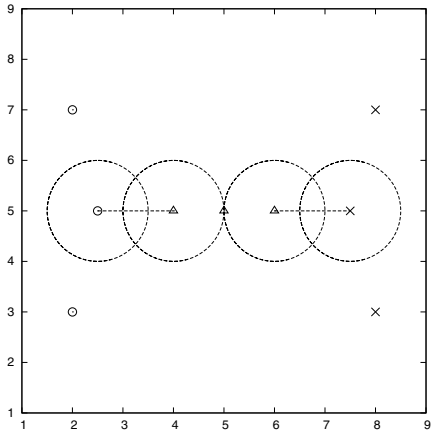


Fig. 5. Classification result with $\kappa_k = 1.0$ and $CL = \{(x_3, x_4), (x_6, x_7)\}$

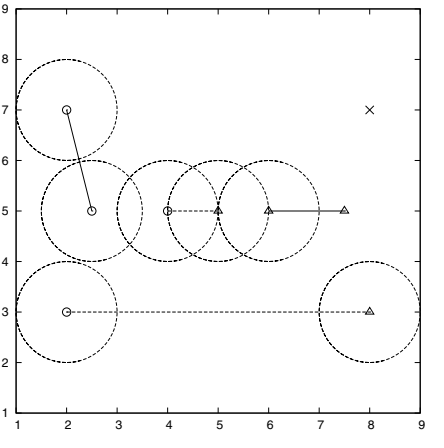


Fig. 6. Classification result with $\kappa_k = 1.0$ and $ML = \{(x_2, x_3), (x_6, x_7)\}$, $CL = \{(x_1, x_8), (x_4, x_5)\}$

6 Conclusions

In this paper, we proposed semi-supervised agglomerative hierarchical clustering using clusterwise tolerance based pairwise constraints (AHC–CTP). The proposed method can handle the pairwise constraints without breaking the Euclidean space by using the concept of clusterwise tolerance. Moreover, we showed the effectiveness of proposed method through numerical examples. The proposed

method is quite different from other semi-supervised clustering methods which are based on pairwise constraints.

In future works, we will show numerical examples with various kinds of data sets. Next, we will compare the proposed method with conventional AHC methods from the viewpoint of merging process in dendrogram. Moreover, we will consider the way to apply the clusterwise tolerance based pairwise constraints to non-hierarchical clustering methods.

Acknowledgments

This study is partly supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and the Grant-in-Aid for Scientific Research (C) (Project No.21500212) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
3. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering. Springer, Heidelberg (2008)
4. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained k-means clustering with background knowledge. In: Proc. of the 18th International Conference on Machine Learning (ICML 2001), pp. 577–584 (2001)
5. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. In: Proc. of the SIAM International Conference on Data Mining (SDM 2004), pp. 333–344 (2004)
6. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 59–68 (2004)
7. Miyamoto, S., Yamazaki, M., Terami, A.: On semi-supervised clustering with pairwise constraints. In: Proc. of The 7th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2009), pp. 245–254 (2009) (CD-ROM)
8. Endo, Y., Hamasuna, Y., Yamashiro, M., Miyamoto, S.: On semi-supervised fuzzy c-means clustering. In: Proc. of 2009 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2009), pp. 1119–1124 (2009)
9. Yan, B., Domeniconi, C.: An adaptive kernel method for semi-supervised clustering. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, pp. 521–532. Springer, Heidelberg (2006)
10. Kulis, B., Basu, S., Dhillon, I., Mooney, R.: Semi-supervised graph clustering: a kernel approach. Machine Learning 74(1), 1–22 (2009)
11. Talavera, L., Béjar, J.: Integrating declarative knowledge in hierarchical clustering tasks. In: Hand, D.J., Kok, J.N., Berthold, M.R. (eds.) IDA 1999. LNCS, vol. 1642, pp. 211–222. Springer, Heidelberg (1999)

12. Klein, D., Kamvar, S., Manning, C.: From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering. In: Proc. of the 19th International Conference on Machine Learning (ICML 2002), pp. 307–314 (2002)
13. Davidson, I., Ravi, S.S.: Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: Proc. of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (KDD 2005), pp. 59–70 (2005)
14. Hamasuna, Y., Endo, Y., Miyamoto, S.: On Tolerant Fuzzy c-Means. Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII) 13(4), 421–427 (2009)
15. Endo, Y., Murata, R., Haruyama, H., Miyamoto, S.: Fuzzy c-Means for Data with Tolerance. In: Proc. of International Symposium on Nonlinear Theory and Its Applications (Nolta 2005), pp. 345–348 (2005)
16. Miyamoto, S.: Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer, Dordrecht (1990)
17. Miyamoto, S.: Introduction to Cluster Analysis: Theory and Applications of Fuzzy Clustering, Morikita-Shuppan, Tokyo (1999) (in Japanese)

Gallbladder Segmentation in 2-D Ultrasound Images Using Deformable Contour Methods

Marcin Ciecholewski

Institute of Computer Science, Jagiellonian University,
ul. Łojasiewicza 6, 30-348 Kraków, Poland
marcin.ciecholewski@ii.uj.edu.pl

Abstract. Segmenting the gallbladder from an ultrasonography (US) image allows background elements which are immaterial in the diagnostic process to be eliminated. In this project, several active contour models were used to extract the shape of the gallbladder, both for cases free of lesions, and for those showing specific disease units, namely: lithiasis, polyps, anatomical changes, such as folds or turns of the gallbladder. First, the histogram normalization transformation was executed allowing the contrast of US images to be improved. The approximate edge of the gallbladder was found by applying one of the active contour models like the motion equation, a center-point model or a balloon model. An operation of adding up areas delimited by the determined contours was also executed to more exactly approximate the shape of the gallbladder in US images. Then, the fragment of the image located outside the gallbladder contour was eliminated from the image. The tests conducted have shown that for the 220 US images of the gallbladder, the area error rate (AER) amounted to 16.4%.

1 Introduction

Computer-assisted methods aimed at facilitating the extraction of organ shapes from medical images and helping to diagnose disease entities are currently rapidly developed. However, for some important organs like the gallbladder there are no ready, practical solutions to help physicians in their work.

The job of extracting the gallbladder structure from US images is a difficult process because images have uneven backgrounds, as shown in Fig. 1. In addition, there is a large variety of gallbladder shapes in US images due to individual traits of patients, among other reasons. US images can also present such disease units as lithiasis, polyps, changes of the organ shape like folds, turns and others which hinder extracting the contour.

In general, literature includes many publications about extracting shapes of organs from US images. One group of algorithms are these that detect edges in the image [1, 2]. Edges are usually located in areas with a high gradient value on the image, where the values of the grey level clearly change, e.g. from black to white. Edge algorithms yield inexact results when detecting an edge that is dotted and unclear. They are also computationally complex and leave noise which

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 163–174, 2010.

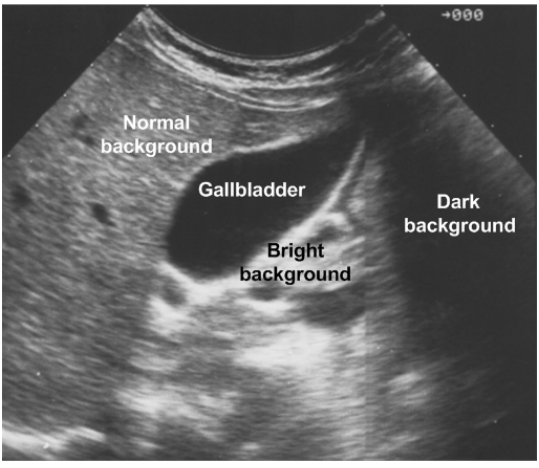


Fig. 1. An example US image of the gallbladder

needs to be eliminated later. Another solution is offered by algorithms based on textures. Richard and Keen [11] have developed an algorithm designed for detecting edges in US images using the classification of pixels corresponding to specific characteristics of textures. Although the algorithm is fully automatic, the authors note that it is computationally complex. The computational complexity of methods based on texture analysis is usually equal to $O(n^4) : W \times H \times r^2$ where: W is the image width, H is its height, and r denotes the lenght of the ROI side.

Algorithms based on deformable models like 2D AAM (the active appearance model) and the active contour (ACM) yield very exact results with relatively low calculation [13, 15]. They are usually semi-automatic methods where the initial contour or the average shape model is initiated by the user. AAM models contain information about the average shape of an object, e.g. the lumbar section of the spine on a digital x-ray image [13] and data describing the most characteristic modifications of this shape observed in the training set. The form of the model may be modified by algorithms which try to fit it to the actual shape while not allowing unnatural deformations to appear. The active contour is a mathematical model of a deformable curve located within a two-dimensional environment of an external field created by the local characteristics of the image. The fitting of the model to the shape of the object is an iterative process just as in the case of AAM. Active contour models have been used for US images to determine the shape of such organs as: the carotid artery [7] and the liver [6]. However, they have not yet been used to support the US diagnostics of the gallbladder. In this publication, the following active contour models have been used to determine the approximate area of the gallbladder: the motion equation, the center-point model and the balloon model. To more exactly approximate the shape of the gallbladder in US images, an operation of adding areas delimited by the determined contours was also executed. The research was conducted on

220 cases from different patients, including US images without lesions and ones showing lesions like: lithiasis, polyps and changes in the shape of the organ, such as folds or turns of the gallbladder. This article has the following structure. Section 2 presents methods of determining the contour of the gallbladder in US images. Section 3 describes the method for segmenting the gallbladder shape in US images. The following section discusses the experiments conducted. The last section contains a summary and sets out directions of future research.

2 Determining the Gallbladder Contour in US Images

This section presents a method of determining the approximate contour of the gallbladder in US images. First, the histogram normalization transformation is executed in order to improve the contrast of US images. The next action is to determine the approximate contour of the gallbladder by applying one of the active contour like the motion equation, a center-point model and a balloon model. To more precisely approximate the shape of the gallbladder, the adding up of areas for the determined contours can be used.

2.1 Histogram Normalization Transformation

The histogram is a one-dimensional statistical function obtained by counting the number of pixels corresponding to specific grey levels. The histogram normalization transformation makes it possible to improve the contrast of images if the values of image brightness do not cover the entire range of possible values.

Let $g : M^2 \rightarrow Z$ be a grey level US image containing the structure of the gallbladder and $(x, y) \in [0, M-1] \times [0, M-1]$ define the coordinates of the pixel. Then: $g(x, y) \in Z$. The Z set defines integers from the $[0, 2^B-1]$ interval, whereas B is the number of bits chosen to represent a single pixel. Assuming that a single pixel is represented by one byte of memory, we get $Z = \{g : g(x, y) \in [0, 255]\}$. The histogram $h(k) : Z \rightarrow Z$ is defined as the following set:

$$h(k) = \{(x, y) : g(x, y) = k\} \tag{1}$$

where $k \in [0, 255]$ is the value of the grey level.

Let $g_{HN} : M^2 \rightarrow Z$ be the histogram normalization transformation. It is represented by the following relationship:

$$g_{HN}(x, y) = LUT(k) = \frac{255}{g_{max} - g_{min}} \cdot (k - g_{min}) \quad if \ g_{min} < k < 255 \tag{2}$$

Where LUT(look-up table) is the adjustment table allowing grey levels of the input image to be changed according to the values stored in that table. We assume that $LUT(k) = 0$ if $k \leq g_{min}$ and $LUT(k) = 255$ if $k \geq 255$. The g_{max} variable is the maximum value of grey levels on the image, whereas g_{min} is the minimum value.

Details of these operations are shown in Fig. 2. Figures 2(a) and 2(b)) contain the image of the gallbladder structure g and its histogram $h(k)$. Figures 2(c) and 2(d) show the g_{HN} image after the histogram normalization transformation and the streched histogram graph.

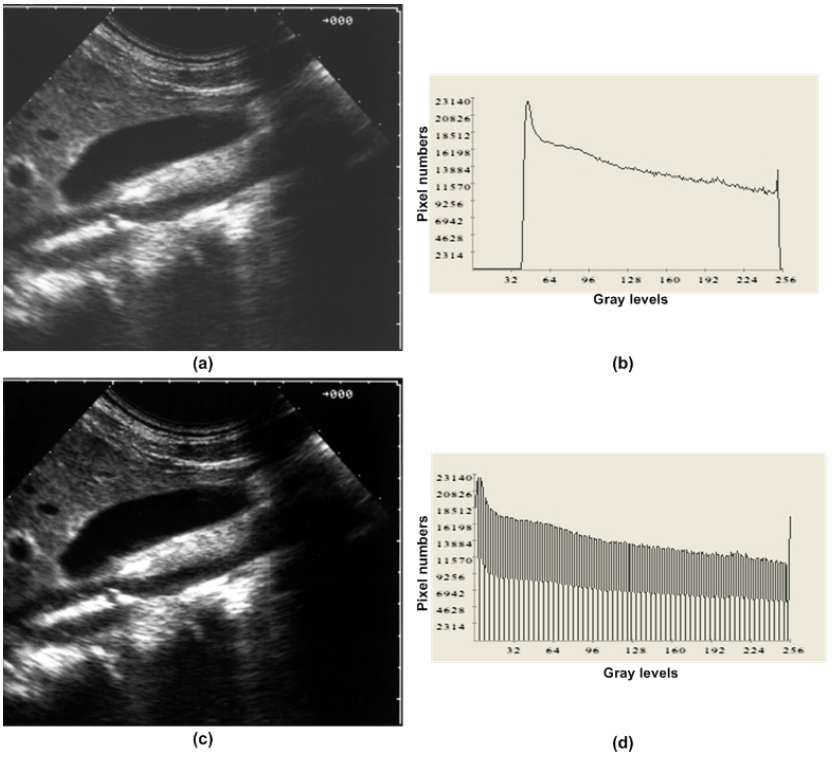


Fig. 2. Transformations in a US image of the gallbladder. (a) A US image containing the structure of the gallbladder. (b) The histogram $h(k)$. (c) The image after the histogram normalization transformation. (d) The histogram graph after the normalization transformation.

2.2 Active Contour Method

An active contour is a mathematical model of a deformable curve made of an abstract, flexible material which reacts to deformations like rubber and springy wire at the same time [8]. In a 2D image analysis context, an active contour is a flat curve which can change its shape dynamically and fit itself to image elements such as edges or borders. The concept of contour shape formation for matching image edges is explained in Fig. 3. The objective of contour movements is to find the best fit, in terms of some cost function, as a trade-off between the contour curvature and the boundary of the image object under analysis. In [8] the potential energy function of the active contour has been proposed to play the role of this cost function. The energy function is given by the following integral equation:

$$E_S = \int_0^{S_{m-1}} [E_i(v(s)) + E_e(v(s)) + E_p(v(s))]ds \tag{3}$$

where the parametric equation $v(s) = (x(s), y(s))$ defines the position of the curve, E_i represents the internal potential energy of the contour, E_e is the energy which models external constraints imposed onto the contour shape, and E_p represents component energies derived from image features, e.g. the image brightness distribution. The notation of the energy function in the discrete format is more convenient in the computer implementation of deformable models:

$$E_S = \sum_{s=0}^{S_{m-1}} [E_i(v(s)) + E_e(v(s)) + E_p(v(s))] \tag{4}$$

In this case, the energy equation is interpreted as the total of component energies of all nodal points. The symbol s symbol is the index identifying the nodal point.

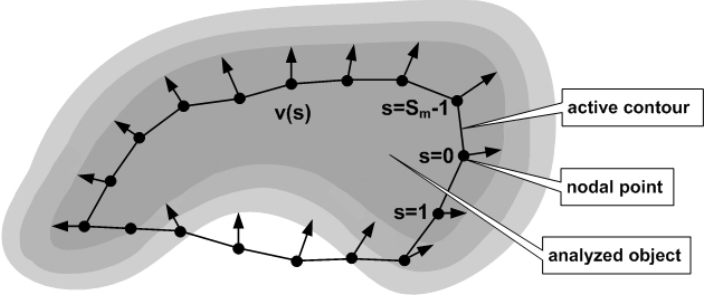


Fig. 3. Building a model of the active contour method. Arrows represent the directions in which nodal points move towards the edge of the analyzed object.

2.3 Motion Equation Model

In this project, the motion equation model proposed in article [9] has been used. This model is treated here as a flexible object of a specific mass moving within an environment of a defined viscosity. Energy E_S is minimized by changing it into the kinetic energy of moving masses of nodal points, subsequently lost as a result of moving within a viscous environment. To model the shifts of individual nodal points, a motion equation of the following form is used:

$$m\frac{\delta^2 v(s,t)}{\delta t^2} + l\frac{\delta v(s,t)}{\delta t} = F(s,t) \tag{5}$$

$$F(s) = -\nabla E_S(s) \tag{6}$$

where $v(s,t)$ is the vector of the nodal point coordinates, m is the mass assigned to every node of the graph, l is the viscosity coefficient of the environment, and F is the vector representing all forces acting on the nodes of the structure. The force F for a single nodal point can be determined as the negated value of the gradient of energy E_S calculated in the image (6). The use of the motion

equation (5) to describe contour dynamics makes it possible to quickly determine the contour balance state and does not require determining the total minimum value of energy E_S shown by equation (4). In the computer implementation, equation (5) is presented in the discrete form of:

$$m[v(s,t) - 2v(s,t-1) + v(s,t-2)] + l[v(s,t) - v(s,t-1)] = F(s,t-1) \quad (7)$$

After determining the location of the nodal point at the moment t, we obtain a formula allowing the location of nodal point at the time t to be calculated iteratively based on the values of forces F and their location in the previous two iterations. We obtain:

$$v(s,t) = \frac{F(s,t-1) + m(2v(s,t-1) - v(s,t-2)) + lv(s,t-1)}{m+l} \quad (8)$$

The numerical convergence and stability of equation (8) depends on the values of parameters m and l, as well as on the way in which force F has been defined. In the case of deformable models, the value of this force depends on many factors, including the features of the analyzed image. The energy minimization method coupled with the motion equation makes it possible to subsequently, in individual iterations, change the location of individual nodal points or of all points at the same time. In the first case, the order of node location modification can be random or defined. If the location of all nodes is modified in the same iteration, equation (8) can be written in the matrix form. The locations of nodes in iteration t are determined based on the values calculated in the previous iteration. The iterative equations have the following form:

$$x_t = \frac{Ax_{t-1} + f_x(x_{t-1}, y_{t-1}) + m(2x_{t-1} - x_{t-2}) + lx_{t-1}}{m+l} \quad (9)$$

$$y_t = \frac{Ay_{t-1} + f_y(x_{t-1}, y_{t-1}) + m(2y_{t-1} - y_{t-2}) + ly_{t-1}}{m+l} \quad (10)$$

In the case of the active contour, matrix A is a pentadiagonal one. For other models, it is a sparse matrix in which the number of elements per row is constant. The number of multiplication and addition operations increases linearly along with the increasing number of nodal points, and not with the square of their number. This is why this method is convenient for models with a large number of nodal points.

2.4 Center-Point Model and Balloon Model

In the case of objects that are oval in shape, the active contour model can be simplified by restricting the freedom of movement of individual nodal points [15], Fig. 4(a). The modification of the original model consists in individual nodal points moving only along half lines with beginnings in the center point. These half lines are distributed radially at equal angular intervals and their number is equal to the number of nodal points. A model in which the movement of nodes

is restricted in the above manner is called a center-point model. In this model, the location of every nodal point is defined by its s index which unambiguously determines the half line slope angle $\alpha(s) = 2\pi/S_m$ and by the $r(s)$ coordinate which is the distance of the nodal point from the center point (the beginning of the half line). The equation allowing the coordinates in the polar system to be replaced with coordinates in the Cartesian system has the following form:

$$v[s] = [x(s), y(s)] = [x_0 + r(s)cos2\pi/S_m \cdot s, y_0 + r(s)sin2\pi/S_m \cdot s] \qquad (11)$$

where (x_0, y_0) are the coordinates of the center point.

In the case of the balloon model [5], the contour curve is treated as the skin of the balloon on the inner side of which the compressed gas which fills the balloon exerts a pressure. The force vector is perpendicular to the tangent to the curve of the contour and oriented to the outside of the area delineated by this curve as shown in Fig. 4(b).

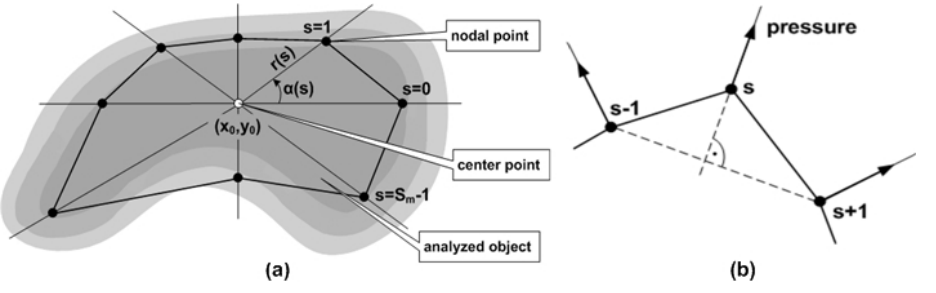


Fig. 4. Active contour models: (a) A model with a center point with the coordinates (x_0, y_0) (b) A balloon model.

2.5 Adding Up Areas Determined Using Active Contour Models

The software being developed to facilitate determining the shape of the gallbladder in US images adds up areas identified using active contour models such as the motion equation, the balloon model and the center-point model. This is aimed at facilitating a more exact approximation of the gallbladder shape in US images. Adding up areas in an image is an elementary operation defined in mathematical morphology [12].

Let us assume that the set $GB \subset M^2$ determines the area of the gallbladder approximated by the total of $k \geq 1$ of sets obtained by using active contour models, i.e.:

$$GB = GB_1 \cup GB_2 \cup \cdots \cup GB_k \qquad (12)$$

During the experiments conducted it turned out that by adding up areas, the shape of the gallbladder can be determined more precisely, particularly if anatomical anomalies like gallbladder folds or turns occur.

Fig. 5 presents the extraction of the shape of a folded gallbladder using the sum of three sets $GB = GB_1 \cup GB_2 \cup GB_3$ representing areas delineated with

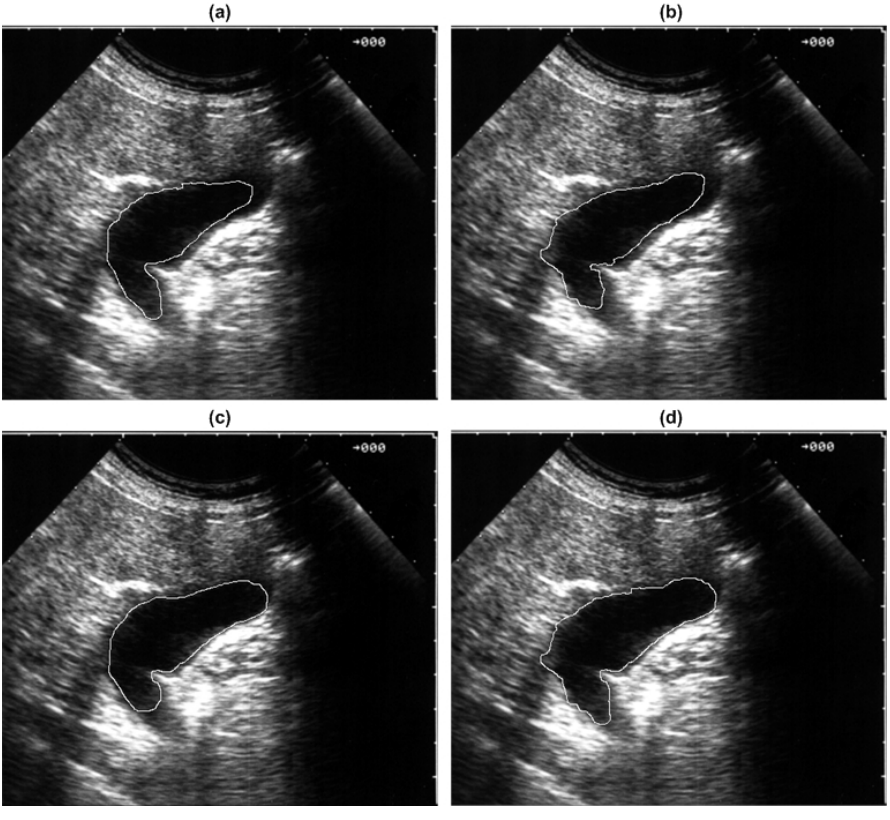


Fig. 5. Extracting the shape of the gallbladder in US images using active contour methods. (a) The gallbladder edge marked using the motion equation model. (b) The gallbladder contour obtained using the center-point model.(c) The gallbladder edge marked after using the balloon model. (d) The edge of the gallbladder shape determined by adding up the three areas determined under (a), (b) and (c).

the contour calculated using: the motion equation model - Fig. 5(a), the center-point model - Fig. 5(b) and the balloon model - Fig. 5(c). As a result of adding up these three areas, a better approximation of the gallbladder shape was obtained, as shown in Fig. 5(d).

3 Gallbladder Segmentation in a US Image

The proposed method of segmenting the gallbladder shape in a US image makes use of the calculated values of coordinates identifying the gallbladder contour determined using one of the active contour models presented in sections 2.3 and 2.4 or using the method of area adding up presented in section 2.5. In order to extract the organ from the image, we have defined two areas identifying image

fragments: GB - the area inside the gallbladder contour and BG - the area constituting the image background. Under these assumptions, the segmentation is executed in such a way that in the US image showing the gallbladder and defined by the mapping $g_{HN} : M^2 \rightarrow Z$ after the histogram normalization (2), its fragment is replaced with the BG area in which all pixels are set to white in colour. We obtain:

$$g' = \begin{cases} g_{HN} & \text{if } (x, y) \in GB \\ 255 \text{ (white)} & \text{if } (x, y) \in BG \end{cases} \tag{13}$$

Figure 6 presents the gallbladder segmentation in a sample US image. Fig. 6(a) shows an image with the gallbladder contour marked. Fig 6(b) and 6(c) contain the US image with the segmented shape of the gallbladder and its histogram $h(k)$.

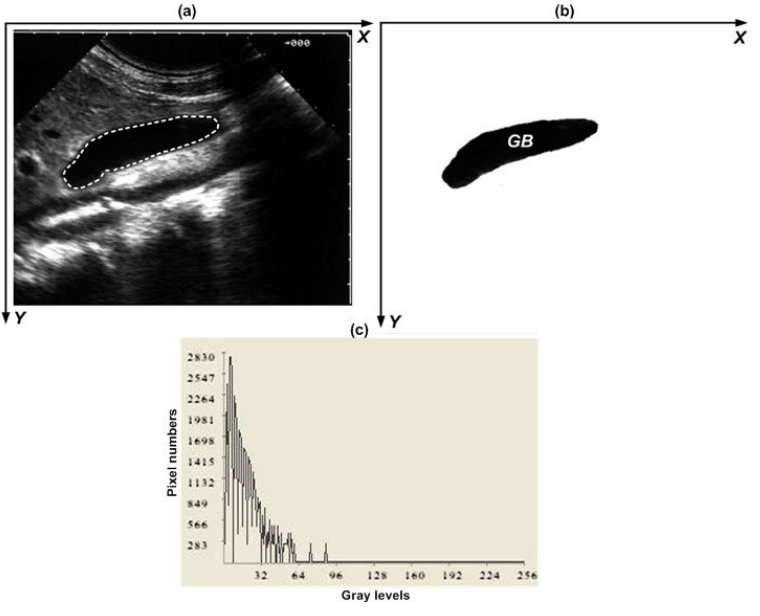


Fig. 6. The gallbladder segmentation in a US image. (a) An image with the gallbladder contour marked. (b) The segmented shape of the gallbladder. (c) Histogram of the image with the segmented shape of the gallbladder.

Figures 7(a) and 7(c) show images with the gallbladder contour marked. Figures 7(b) and 7(d) contain the US images with the segmented shape of the gallbladder. Figures 7(a) and 7(b) show images with with lithiasis, while figures 7(c) and 7(d) a fold of the gallbladder.

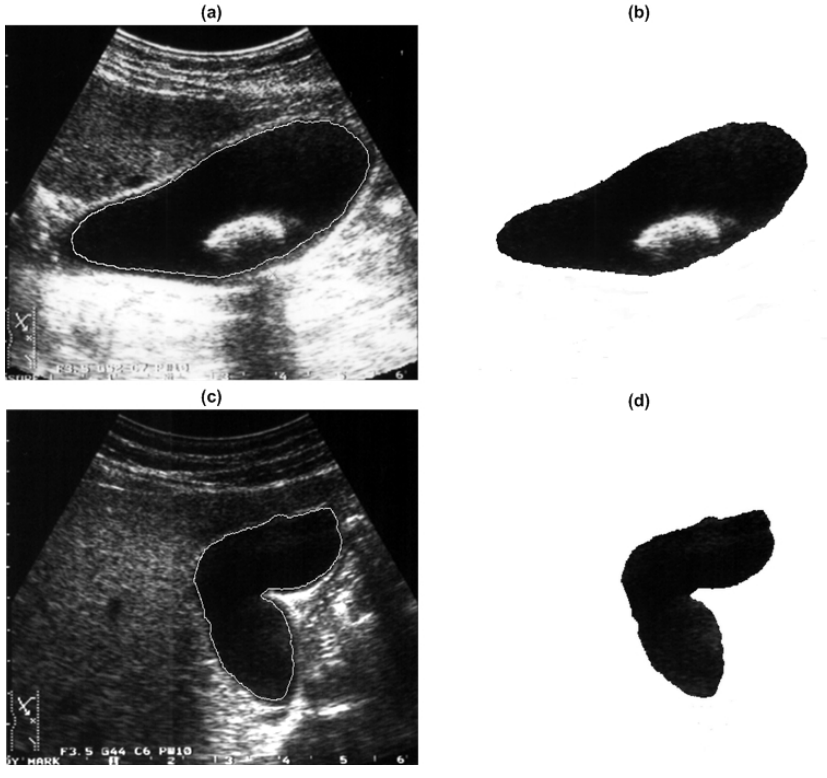


Fig. 7. The gallbladder segmentation in US images using active contour methods. (a), (b) An image with visible cholecystolithiasis. (c), (d) A gallbladder fold.

4 Completed Experiments and Selected Research Results

In order to estimate the precision of models used to determine the approximate contour of the gallbladder, the area error rate (AER) was used. The material from the Department of Image Diagnostics of the Regional Specialist Hospital in Gdańsk, Poland, was used.

4.1 Area Error Rate

The area error rate AER is an estimated value which allows a percentage change in the difference between occupied areas of an image to be compared. The difference in areas is calculated between the area extracted using the active contour model and that extracted manually (MSR). Let $Lv_{accon} \subset Z^2$ be the fragment of the image obtained using the active contour method and $Lv_{manual} \subset Z^2$ signify the image fragment extracted manually. Let $UR = Lv_{accon} \cup Lv_{manual}$ and $IR = Lv_{accon} \cap Lv_{manual}$. The AER is defined as follows:

$$AER = \frac{a_{UR} - a_{IR}}{a_{MSR}} \times 100\% \tag{14}$$

It was assumed that a_{UR} is the number of pixels within the area UR, while a_{IR} signifies the number of pixels within the area IR, and a_{MSR} is the number of pixels in the manually extracted area MSR. The MSR area of the gallbladder surface was determined by a radiologist and the AER for the 220 US images analyzed amounted to 16.4%. Table 1 presents the results of experiments for particular disease units in relation to the number of cases.

Table 1. Test results for 220 US images of the gallbladder

Patients	No. of images	AER
No lesions	100	11%
Lithiasis	70	21%
Polyp	20	18.7%
Fold/Turn	30	15.2%
Total	220	16.4%

5 Summary and Further Research Directions

This article presents a method of segmenting the shape of the gallbladder from US images developed for a computer system supporting the early diagnostics of gallbladder lesions. First, the histogram normalization transformation was executed allowing the contrast of US images to be improved. The approximate edge of the gallbladder is determined by applying one of the active contour models like the motion equation, a center-point model or a balloon model. To more precisely approximate the shape of the gallbladder, the adding up of the determined contours can be used. The fragment of the image located outside the gallbladder contour is eliminated from the image. The active contour method using the applied models coupled with the area adding up operation yielded quite precise results for both healthy organs and those showing specific disease units, namely: lithiasis, polyps, folds and turns of the gallbladder. For the 220 US images analyzed, the area error rate amounted to 16.4% Further research will be aimed at reducing the AER for images showing lesions such as lithiasis and polyps, if they are located close to the gallbladder edge. In addition, there is a need to make it possible to more precisely approximate tapering ends and the approximation of corners in the gallbladder contours determined.

Acknowledgements

This research was financed with state budget funds for science for 2009-2012 as research project of the Ministry of Science and Higher Education: N N519 406837.

References

1. Aarnink, R.G., Pathak, S.D., de la Rosette, J.J., Debruyne, F.M., Kim, Y., et al.: Edge detection in prostatic ultrasound images using integrated edge maps. Ultrasonics 36, 635–642 (1998)
2. Bodzioch, S.: Information reduction in digital image and its influence on the improvement of recognition process. Automatics, Semi-annual Journal of the AGH University of Science an Technology 8(2), 137–150 (2004)
3. Ciecholewski, M., Dębski, K.: Automatic Segmentation of the Liver in CT Images Using a Model of Approximate Contour. In: Levi, A., Savaş, E., Yenigün, H., Balcısoy, S., Saygın, Y. (eds.) ISCIS 2006. LNCS, vol. 4263, pp. 75–84. Springer, Heidelberg (2006)
4. Ciecholewski, M., Ogiela, M.: Automatic Segmentation of Single and Multiple Neoplastic Hepatic Lesions in CT Images. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2007. LNCS, vol. 4528, pp. 63–71. Springer, Heidelberg (2007)
5. Cohen, L.D., Cohen, I.: Finite-Element Methods for Active Contour Models and Balloons for 2-D and 3-D Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1131–1147 (1993)
6. Cvancarova, M., Albregtsen, T.F., Brabrand, K., Samset, E.: Segmentation of ultrasound images of liver tumors applying snake algorithms and GVF. Computer Methods and Programs in Biomedicine 84(2-3), 86–98 (2006)
7. Hamou, A.K., Osman, S., El-Sakka, M.R.: Carotid Ultrasound Segmentation Using DP Active Contours. In: Kamel, M.S., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 961–971. Springer, Heidelberg (2007)
8. Kass, M., Witkin, A., Terauzopoulos, D.: Snakes: Active Contour Models. International Journal of Computer Vision 1(4), 321–331 (1988)
9. Leymarie, F., Levine, M.D.: Simulating the Grassfire Transform using an Active Contour Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(1), 56–75 (1992)
10. Neuenschwander, W., Fua, P., Kuebler, O.: From Ziplock Snakes to Velcro Surfaces, Automatic Extraction of Man Made Objects from Aerial and Space Images, Monte Verita, pp. 105–114. Birkhaeuser, Basel (1995)
11. Richard, W.D., Keen, C.G.: Automated texture-based segmentation of ultrasound images of the prostate. Comput. Med. Imaging Graph. 20(3), 131–140 (1996)
12. Ritter, G.X., Wilson, J.N.: Computer Vision Algorithms in Image Algebra. CRC Press, Boca Raton (2000)
13. Roberts, M.G., Cootes, T.F., Adams, J.E.: Automatic segmentation of lumbar vertebrae on digitised radiographs using linked active appearance models. Proc. Medical Image Understanding and Analysis 2, 120–124 (2006)
14. Schilling, R.J., Harris, S.L.: Applied numerical methods for engineers. Brooks/Cole Publishing Com., Pacific Grove (2000)
15. Szczypiński, P., Strumiłło, P.: Application of an Active Contour Model for Extraction of Fuzzy and Broken Image Edges. Machine Graphics & Vision 5(4), 579–594 (1996)

Pattern Mining on Stars with FP-Growth

Andreia Silva and Cláudia Antunes

Instituto Superior Técnico
Av Rovisco Pais 1
1049-001 Lisboa
{andreia.silva,claudia.antunes}@ist.utl.pt

Abstract. Most existing data mining (DM) approaches look for patterns in a single table. Multi-relational DM approaches, on the other hand, look for patterns that involve multiple tables. In recent years, the most common DM techniques have been extended to the multi-relational case, but there are few dedicated to star schemas. These schemas are composed of a central fact table, linking a set of dimension tables, and joining all the tables before mining may not be a feasible solution. This work proposes a method for frequent pattern mining in a star schema based on FP-Growth. It does not materialize the entire join between the tables. Instead, it constructs an FP-Tree for each dimension and then combines them to form a super FP-Tree, that will serve as input to FP-Growth.

1 Introduction

While most existing data mining approaches look for patterns in a single data table, multi-relational data mining (MRDM) approaches look for patterns that involve multiple tables (relations) from a relational database or data warehouse. In recent years, the most common types of patterns and approaches considered in data mining have been extended to the multi-relational case and MRDM now encompasses multi-relational (MR) association rule discovery, MR decision trees and MR distance-based methods, among others. MRDM approaches have been successfully applied to a number of problems in a variety of areas [4].

From those works, just a few are dedicated to frequent itemset mining on star schemas [2,8,10].

This work aims to find frequent patterns in a set of tables of a data warehouse, following a star schema, without materializing the join of its tables.

A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process [7]. In terms of data modeling, a data warehouse consists of one or several dimensional models that are composed of a central fact table and a set of surrounding dimension tables, each corresponding to one of the dimensions of the fact table. The most used dimensional model is the star schema, which consists of multiple dimension tables that are associated by foreign keys to a central fact table.

At first glance, it may seem easy to join the tables of a star schema, and then do the mining process on the joined result [8]. However, when multiple tables

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 175–186, 2010.

are joined, the resulting table will be much larger and the mining process more expensive and time consuming. There are two major problems: First, in large applications, often the join of all related tables cannot be realistically computed because of the distributed nature of data, large dimension tables and the many-to-many relationship blow up. Second, even if the join can be computed, the multifold increase in both size and dimensionality presents a huge overhead to the already expensive pattern mining process:

(1) the number of columns will be close to the sum of the number of columns in the individual tables.
(2) If the join result is stored on disk, the I/O cost will increase significantly for multiple scanning steps;
(3) For mining frequent itemsets of small sizes, a large portion of the I/O cost is wasted on reading the full records containing irrelevant dimensions;
(4) Each tuple in a dimension table will be read multiple times in one scan of the joined result. The number of times that a tuple appears in the fact table is the number of times the whole tuple will be read in the joined result.

One of the great potential benefits of MRDM is the ability to automate the mining process to a significant extent. Fulfilling this potential requires solving the significant efficiency problems that arise when attempting to do data mining directly from a relational database, instead of a single pre-extracted flat file[3].

The proposed algorithm is an adaptation of FP-Growth [5] to mine a star schema. The main idea is to adapt the construction of the FP-Tree, so that FP-Growth can run.

Like FP-Growth, it scans each table only twice: first to count the support of each item, and second to construct the FP-Tree. It is divided in 3 stages:

1. **Support Counting:** The fact table is scanned to count the support of each foreign key.
2. **Local Mining:** An FP-Tree is constructed for each dimension table (DimFP-Tree), with a slight modification of the original FP-Tree, taking into account the support calculated in the previous step.
3. **Global Mining:** The FP-Trees of each dimension are combined to form a Super FP-Tree, according to each fact and an established order of dimensions. This Super FP-Tree is then mined with FP-Growth, without a change, giving all the frequent patterns.

Several orders for the dimensions were studied and are presented and compared in this work.

The rest of this paper is organized as follows. Section 2 presents the related work on MRDM on star schemas. The proposed algorithm is described on section 3. Section 4 gives some experimental results and section 5 presents the conclusions.

2 Related Work

The work related to multi-relational pattern mining on star schemas is increasing. Experiments showed that the approach of “mining before join” outperforms the

approach of "join before mining" even when the latter adopts known to be fastest single-table mining algorithms[8].

Jensen and Soparkar (2000) presented an Apriori based algorithm [2], that first generates frequent itemsets in each single table using a slightly modified version of Apriori[1], and then looks for frequent itemsets whose items belong to distinct tables via a multi-dimensional count array. It does not construct the whole joined table but processes each row as the row is formed, thus storage cost for the joined table is avoided. However, the number of candidates generated explodes as the number of dimensions, attributes and values increase.

Ng et al.(2002) proposed an efficient algorithm without actually performing the join operation [8]. They perform local mining on each dimension table, and then "bind" two dimension tables at each iteration, i.e. mine all frequent itemsets with items from two different tables without joining them. After binding, those two tables are virtually combined into one, which will be "binded" to the next dimension table. They use vertical data format, and a prefix tree to compress the fact table and to speed up the calculation of support.

Xu and Xie (2006) proposed MultiClose [10], which discover frequent closed itemsets without materializing join tables. It first converts the dimension tables to vertical data format, and then mines each of them with a closed algorithm. After local mining, frequent closed itemsets are stored in two-level hash table result trees, and the frequent closed itemsets across two tables are discovered by traversing those result trees.

Several multi-relational methods have been developed by the Inductive Logic Programming community over the recent years, but they are usually not scalable with respect to the number of relations and attributes in the database. Therefore they are inefficient for databases with complex schemas. Another drawback of the ILP approaches is that they need the data in the form of prolog tables.

There are other algorithms for finding multi-relational frequent itemsets, however they just consider one common attribute at a time. They would have to run as much times as the number of dimensions, since there is no attribute common to all the tables. Instead, in a star schema, the fact table has one attribute in common with each dimension, and the dimensions have no common attribute between them. The patterns discovered by those algorithms will not reflect the relationships between dimensions.

The proposed algorithm also mines star schemas without computing the entire join nor materializing join tables. After constructing an FP-tree for each dimension, the corresponding tables are discarded. The trees already take into account the minimum frequency and incorporate transaction ids. The super FP-tree that represents the whole star is then constructed, by aggregating the dimension trees all together, based on the fact table. Finally, this tree in mined using the known pattern growth method, FP-Growth[5].

3 Mining Stars

Consider a relational database modeled as a star schema.

There are multiple dimension tables, which we will denote as A, B, C, ..., each containing only one primary key, denoted by transaction id (tid), some other attributes and no foreign keys. In fact, what we want is to discover potentially useful relationships among the attributes, other than primary keys. The set of values for an attribute is called the domain of the attribute.

In order to simplify our discussion we assume the fact table, denoted as FT, only contains the tids from dimension tables as foreign keys (tid_A, tid_B, tid_C, ...). If it contains some fields other than primary keys, we can place them into an extra dimension table and insert a new foreign key corresponding to it into the fact table. They are considered facts or measures.

As example, lets consider a dataset of the real movies database used in the experiments, only with three dimensions: Award (A), Studio (B) and Movie (C). Table 1 presents a conceptual representation of the dimension tables, where a_i, b_i and c_i denote the tid of dimension tables A, B and C, respectively, and x_i, y_i and z_i denote each possible value of A, B and C. Table 2a shows the fact table.

This example will be used to show how the proposed algorithm works, with a minimum support equals to 40% of the database. The sample of the database has

Table 1. Dimension Tables

(a) Table A

tid_A	Itemsets	Support
a_1	x_1	3
a_2	$x_2x_3x_4$	1
a_3	$x_5x_6x_7$	5
a_4	$x_8x_6x_7$	1

(b) Table B

tid_B	Itemsets	Support
b_1	y_1	2
b_2	$y_2y_3y_4$	1
b_3	$y_5y_3y_6$	1
b_4	$y_7y_3y_6$	2
b_5	y_8y_3	1
b_6	$y_9y_3y_6$	2
b_7	$y_{10}y_{11}y_6$	1

(c) Table C

tid_C	Itemsets	Support
c_1	z_1z_2	1
c_2	z_3z_4	1
c_3	z_5z_4	1
c_4	z_6z_4	1
c_5	z_7	1
c_6	z_8z_2	1
c_7	z_9z_4	1
c_8	$z_{10}z_{11}$	1
c_9	$z_{12}z_4$	1
c_{10}	$z_{13}z_4$	1

Table 2. Fact table and the frequent itemsets corresponding to each tid

(a) Fact Table

tid_A	tid_B	tid_C
a_3	b_2	c_1
a_3	b_4	c_2
a_3	b_6	c_3
a_3	b_4	c_4
a_1	b_1	c_5
a_1	b_5	c_6
a_3	b_7	c_7
a_1	b_1	c_8
a_2	b_3	c_9
a_4	b_6	c_{10}

(b) Denormalized Fact

$Itemsets_A$	$Itemsets_B$	$Itemsets_C$
$x_6x_7x_5$	y_3	–
$x_6x_7x_5$	y_3y_6	z_4
$x_6x_7x_5$	y_3y_6	z_4
$x_6x_7x_5$	y_3y_6	z_4
–	–	–
–	y_3	–
$x_6x_7x_5$	y_6	z_4
–	–	–
–	y_3y_6	z_4
x_6x_7	y_3y_6	z_4

10 transactions, therefore 40% of it corresponds to 4 transactions. This means that an itemset is frequent if its support is no less than 4 transactions.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of distinct literals, called items. A subset of items is denoted as an itemset. A transaction $T = (t_{id}, X)$ is a tuple where t_{id} is a transaction-id and X is an itemset in I. Each table, in a relational database D, is a set of transactions. The *support* (or occurrence frequency) of an itemset It, is the number of transactions containing It in the database. It is frequent if its support is no less than a predefined minimum support threshold, σ.

In a database modeled as a star schema, where there are several tables, we have to be more specific: the *local support* of an itemset It, with items belonging to a table A (A.localSup(It)), is the number of occurrences of It in A. For example, on table 1a, A.localSup(x_1) = 1 and A.localSup(x_6) = 2. The *global support* (or just *support*) of an itemset It is the number of transactions of the fact table containing all the *tid*s that contain It, as in equation 1:

$$globalSup(It) = \sum_{tid}^{tid(It)} FT.localSup(tid) \tag{1}$$

Following the example above, globalSup(x_1) = FT.localSup(a_1) = 3 and globalSup(x_6) = FT.localSup(a_3) +FT.localSup(a_4) = 5 + 1 = 6.

3.1 The Algorithm

Star FP-Growth mines multiple relations for frequent patterns in a database following a star schema. The result is the same as mining the joined table, but without materializing it.

It is based on FP-Growth [5], and the main idea is to construct a Super FP-Tree, combining the FP-Trees of each dimension, so that the original FP-Growth can run and find multi relational patterns. Like FP-Growth, it scans each table only twice: first to count the support of each item and second to construct the FP-Tree.

The overall steps are:

Step 1: Support Counting: The fact table is scanned to count the support of each *tid* of each dimension. In the example, the *tid* support is shown in the third column of each dimension table (Table 1).

Step 2: Local Mining: An FP-Tree is constructed for each dimension table (DimFP-Tree), with a slight modification of the original FP-Tree, taking into account the support calculated in the previous step.

Step 3: Global Mining:

Step 3.1: Construct the Super FP-Tree: The DimFP-Trees of each dimension are combined to form a Super FP-Tree, according to each fact and an established order among dimensions.

Step 3.2: Mining the Super FP-Tree: Run FP-Growth [5], without a change, with the Super FP-Tree and the minimum support threshold. The result of this step is a list of all patterns, not just those relating to one dimension, but also those which relate the various dimensions.

Constructing the DimFP-Tree. A DimFP-Tree is very similar to an FP-Tree[5]. There are two major differences between the construction of a DimFP-Tree and an FP-Tree:

First, the support used here is the global support of each item, i.e. we consider the occurrences of an item in all database, not only in the item's table. Therefore, a node does not start with the support equals to one, but with support = support(T), with T the *tid* of the transaction that originated the node. It also is not incremented by only one, but by support(T).

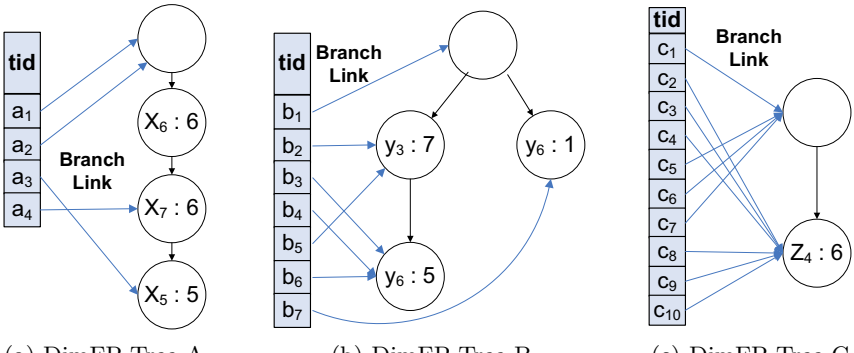
For example, b_4 has a support = 2, which means that that transaction occurs two times in the database. Adding two times the same transaction with support = 1 is the same as adding it one time with support = 2. Starting a node with support = support(T) and incrementing by support(T) avoids being repeatedly inserting the same transaction.

Second, instead of the header table, the DimFP-Tree has other structure, the branch table, that keeps track of the path correspondent to each *tid*. It stores the last node of that path for each *tid*. This structure will help the global mining. If we want to know which frequent items belong to a *tid*, we follow the link in that table to find the last node, and then we just have to climb through its parents till we reach the root node. The items of the nodes in the path we took are the frequent items of that transaction.

The result is the same as if we have the table with the transactions and their frequent sorted items. However, the size of the tree is usually much smaller than its original database, therefore, not having that table materialized in memory usually saves a lot of space and avoids duplicates [6]. If we keep the table instead of the tree, and if two transactions have the same frequent items, those items will be repeated two times. With the tree, there is just one path corresponding to the two transactions. Further, shared parts can also be merged using the tree. Therefore, in a larger scale, the more transactions there are, the greater the difference.

Lets consider the construction of the DimFP-Tree (figure 1b) of table B:

Step 2 starts with a first scan to the table B to calculate the support of each item. Only y_3 and y_6 are frequent, i.e. have a support no less than four transactions ($sup(y_3) = 7$ and $sup(y_6) = 6$). Therefore, only the itemsets containing just y_3



(a) DimFP-Tree A (b) DimFP-Tree B (c) DimFP-Tree C

Fig. 1. DimFP-Trees for each dimension table

and/or y_6 would be frequent, according to the anti-monotone property [1]. For each transaction $b_1, b_2, \ldots, b_7$, frequent items are selected and sorted according to the support descending order, and then inserted in the tree. At the same time, the branch table is constructed, linking each *tid* to the respective node in the tree. b_1, for example, does not have any frequent item, therefore its branch link links to the root node. b_3 corresponds to the first path of the tree, therefore its branch link points to the last node in that path. The other transactions follow the same reasoning. Figure 1 shows the results for all dimensions.

Constructing the Super FP-Tree. The Super FP-Tree is just like an FP-Tree, since it will serve as input to FP-Growth. The construction is very similar to the construction of an FP-Tree [5]. Despite this, there are three differences:

First, it is not necessary the first scan to any table to calculate the supports. They are already calculated and stored in each dimension tree.

Second, a fact is a set of *tid*s, therefore the denormalization of each fact is necessary before ordering the items or inserting them in the tree. A denormalized fact is an itemset with the items corresponding to its *tid*s. Through the branch table of each DimFP-Tree we can get the path corresponding to the itemset of each *tid*. Furthermore, according to the anti-monotone property, if an itemset is not frequent, no other itemset containing it will be. Thus, we only check the frequent items in the transactions of each *tid*, ensuring that the final tree has only the frequent items of each dimension. Table 2b, one can see the result of denormalizing each fact of our example.

And third, the ordering of items in a transaction does not have to be the frequency descending order. As verified in the improvement of FP-Growth proposed in [6], an FP-tree based on frequency descending ordering may not always be minimal.

The support descending ordering enhances the compactness of the FP-tree structure. However, this does not mean that the tree so constructed always achieves the maximal compactness. With the knowledge of particular data characteristics, it is sometimes possible to achieve even better compression.

There are two related and important properties of the FP-tree that can be derived from its construction process [6].

On one hand, given a transaction database DB, and without considering the root, the size of an FP-tree is bounded by $\sum_{T \in DB} |freq(T)|$.

The height of the tree is bounded by $max_{T \in DB}\{|freq(T)|\}$, where $freq(T)$ gives the frequent items of transaction T.

This means that the number of nodes of an FP-Tree (size) is, at most, the number of frequent items in all the transactions, and the number of levels (height) is, at most, the maximal number of frequent items in a transaction.

On the other hand, given a transaction database DB, the number of paths in an FP-tree is bounded by $|DB|$, i.e., it is, at most, the number of transactions in the database, if each transaction contributes to one different path of the FP-tree, with the length equal to the number of frequent items in that transaction.

With those lemmas in mind, several orders among dimensions were studied and compared in terms of the properties defined above. The three most relevant are the following:

1. Support descending order of items

This ordering does not have into account any order of dimensions. After denormalizing the fact, the transaction may have items from multiple dimensions. Sorting them in a support descending order may result in an itemset with items from multiple dimensions intermixed. This is the order used in the original FP-Growth. So, the tree resulting from applying this ordering is the same as the tree resulting from joining the tables in one and applying directly FP-Growth to it (but in this case, the joining is not materialized).

2. Support descending order of dimensions

As the support descending ordering enhances the compactness of the FP-tree structure, it is a promising order for the dimensions. Dimensions with higher support are more likely to be shared and thus arranged closer to the top of the FP-tree.

Since each dimension can have multiple items with different supports, we consider that the dimension support corresponds to the support of its least frequent item. Dimensions with the same support are ordered alphabetically, and the items of a dimension are also ordered in a support descending order.

Note that, with this ordering, items from multiple dimensions are not intermixed. Items from dimensions with higher support will always appear before those from dimensions with lower support.

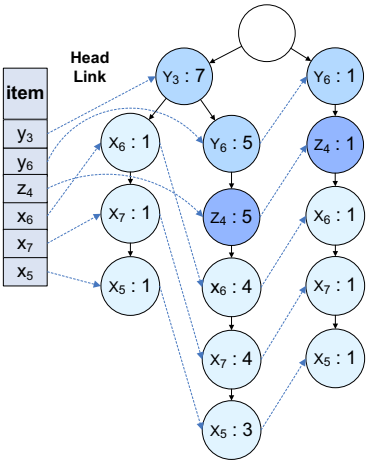


Fig. 2. Super FP-Tree with a support descending order of dimensions

In our example, the lowest support in dimensions is five in dimension A, and six in B and C. The support descending order of these dimensions is $B \rightarrow C \rightarrow A$ $((y_i s) \rightarrow (z_i s) \rightarrow (x_i s))$. Figure 2 presents the resulting Super FP-Tree. The branch table shows the items according to this ordering.

3. Path ascending order of dimensions

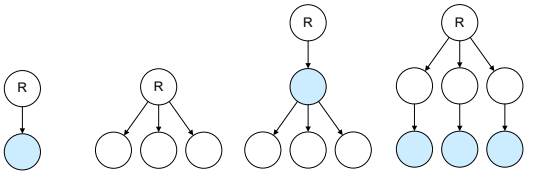
If we look at the number of paths of a tree, $|paths|$ (or just P), we can state that, when joining 2 or more trees of different $|paths|$ (i.e. adding one tree to every leaf of the other), the order in which the trees are joint influences the size of the resulting tree. Note that the $|paths|$ of the joined tree is always the same and equals to $\prod_{t \in TS} |P(t)|$, where TS is the set of the trees we want to join. The number of paths of a tree is the same as the number of leafs.

Its size (without the root) is given by

$$\sum_{i=1}^{|TS|} \left(\prod_{j=1}^{i-1} P(t_j) \right) \times size(t_i) \tag{2}$$

where $P(t)$ gives the number of paths in the tree t and TS is the set of trees in the order they are joint. The explanation is the following: a tree is inserted in each leaf of the tree immediately above, which in turn, was also inserted in each leaf of the preceding tree.

For example, imagine we have a tree A with 1 path, and another, B, with 3 paths (Figure 3a and 3b). If we join A with B, a copy of B is inserted in each leaf of tree A. Therefore, the resulting tree will have 4 nodes (without the root), as shown in figure 3c. Joining B with A will result in a tree with more nodes, 6 (figure 3d). This is because, the more leafs the tree above had, more copies of the tree below are needed.



(a) Tree A (b) Tree B (c) A with B (d) B with A

Fig. 3. Joining two trees

As we stated, joining trees in a path ascending order, will result in the smallest tree. However, joining two DimFP-Trees is not that linear. We may have to insert one tree not just in the leafs of the other, but also in a middle node, if that node corresponds to the last node of one transaction. Expressions above are also valid for this case, if we consider that $P(t)$ gives the number of nodes that correspond to the last node of a transaction.

Summing, this ordering consists in joining DimFP-Trees in a path ascending order. Dimensions with the same number of paths are ordered alphabetically, and the items of one dimension are ordered in a support descending order, like the other orderings.

These orderings have been applied to the movies database and the results are presented in section 4. Note that the set of frequent items is independent of the order applied. The result is the same for every orderings.

4 Experimental Results

The Super FP-Tree is the main structure of this algorithm. This is the tree that will be mined with FP-Growth, and this is the tree that holds the patterns we want to find. Therefore, the Super FP-Tree is the central object of these experiments.

Our goal is to analyze the impact of different orderings on the performance of our algorithm. In order to do that, the three orderings for the construction of the Super FP-Tree are compared, varying the minimum support threshold and the time spent in each step is analyzed.

The dataset is a real movies database [9]. The real database has six dimensions, with different numbers of records, from about 20 to 11000, and with a fact table with about 11000 transactions.

Among the data, we encounter a description of the directors, producers and awards received for each film and time information about them.

To achieve reliable results, the data was split into five equal datasets. The tests were applied to each dataset and we considered the average of each local result. Therefore, we analyze about 2000 facts at each time.

When comparing the size of the Super FP-Tree (figure 4), i.e. the number of nodes, the tree that is more compact is the one resulting from applying some order to the dimensions. In terms of compression, the support descending and path ascending orders of dimensions are very similar, and better than the support descending order of items. In this experiments, this ordering gave always the tree with more nodes. This difference happens because assigning an ordering for the dimensions, taking into account their characteristics and the properties described above, increase the number of shared nodes, and therefore, the compactness of the Super FP-Tree.

Note that the support descending order of items gives the same results as constructing an FP-Tree from the flat table (resulting from the join of the star), therefore, it serves as a reference to the other orderings.

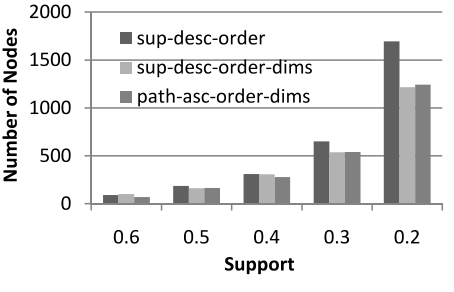


Fig. 4. Average size of the Super FP-Tree

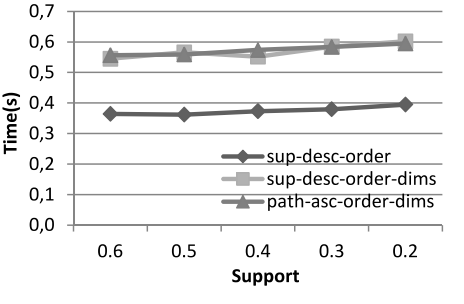


Fig. 5. Average time on Super FP-Tree construction

In terms of the number of paths in the Super FP-Tree, the resulting trees are very similar. Although the support descending order of items gives the less compact tree, it gives a tree with slightly less paths than the other orders.

The average time spent in the mining process was also studied. On average, 98% of the time is spent in the construction of the DimFP-Trees (step 2), and counting the global support (step 1) only takes 0,20% of the time. The difference between the application of the three orderings is mostly seen in step 3. As can be seen in figure 5, the support descending order of items takes less time constructing the Super FP-Tree than the other orders. With the support descending and the path ascending orders of dimensions, each transaction of the fact table has to be ordered according to that ordering before it can be inserted in the tree, yielding the previous results.

Even though the time for the construction of the Super FP-Tree is smaller for the support descending order of items, the FP-Growth will take longer to

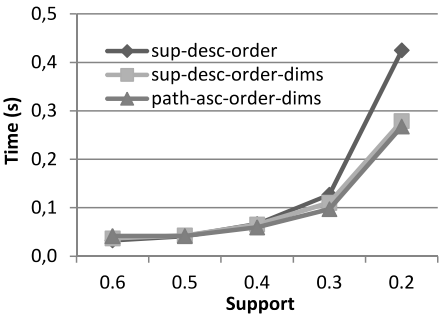


Fig. 6. Average time of FP-Growth

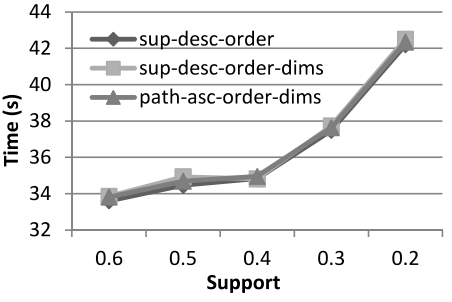


Fig. 7. Average total time on Star FP-Growth

execute (figure 6), due to resulting tree's size (figure 4 shows its size is bigger). Therefore, in the end, the total time needed for running Star FP-Growth (figure 7) is very similar for all orderings.

The size of the trees, as well as the time spent, depend not just on the size and characteristics of the data, but also on what the user wants. As minimum support decreases, the number of patterns increases, and therefore the tree and time will also increase. However, the memory needed to keep the trees will generally be much less than the memory needed to keep the tables.

The computer used to run the experiments was an Intel Xeon E5310 1.60GHz (Quad Core), with 2GB of RAM. The operating system used was GNU/Linux amd64 and the algorithm was implemented using the Java Programming language (Java Virtual Machine version 1.6.0_02). The tables were maintained in memory, as well as all the trees. However, the dimension tables were freed before Global Mining, as well as the fact table before running FP-Growth.

5 Conclusions

Star FP-Growth is a simple algorithm for mining patterns in a star schema. It does not perform the join of the tables, making use of the star properties. Building a tree for each dimension taking into account the global support of items allows us to discard the respective table and to keep only the frequent items. The main purpose is to prepare the FP-Tree that represents the data, combining the dimension trees, so that it can serve as input to FP-Growth.

Three orderings for dimensions were analyzed and the results state that applying a support descending or a path ascending order for the dimensions achieve better compression than the usual support descending order of items. The time spent in the mining process was very similar for both orderings, but it actually depends on the size and characteristics of the data, and on what we want: minimum support, performance, memory.

Using a pattern growth method and the FP-Tree gives us an important benefit: the size of an FP-tree is bounded by the size of its corresponding database

because each transaction will contribute at most one path to the FP-tree, with the length equal to the number of frequent items in that transaction. Since there are often a lot of sharing of frequent items among transactions, the size of the tree is usually much smaller than its original database. Unlike the Apriori-like method which may generate an exponential number of candidates in the worst case, under no circumstances, may an FP-tree with an exponential number of nodes be generated.

If the tree cannot be maintained in main memory, several techniques can be used, whether representing and storing the tree in hard disk, or partitioning the database into a set of projected databases, and then for each projected database, constructing and mining its corresponding FP-tree [6].

The proposed algorithm can also be generalized to be applied to a snowflake structure, where there is a star structure with a fact table FT, but a dimension table can be replaced by another fact table FT', which is connected to a set of other dimension tables. We can consider mining across dimension tables related by FT' first. Then consider the resulting Super FP-Tree as a derived DimFP-Tree and continue processing the star structure with FT. This means that mining a snowflake starts from their "leaves".

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB 1994, Proceedings of 20th International Conference on Very Large Data Bases, pp. 487–499 (September 1994)
2. Crestana-Jensen, V., Soparkar, N.: Frequent itemset counting across multiple tables. In: Terano, T., Chen, A.L.P. (eds.) PAKDD 2000. LNCS, vol. 1805, pp. 49–61. Springer, Heidelberg (2000)
3. Domingos, P.: Prospects and challenges for multi-relational data mining. SIGKDD Explor. Newsl. 5(1), 80–83 (2003)
4. Džeroski, S.: Multi-relational data mining: an introduction. SIGKDD Explor. Newsl. 5(1), 1–16 (2003)
5. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: SIGMOD 2000: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 1–12. ACM, New York (2000)
6. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)
7. Inmon, W.H.: Building the data warehouse, 2nd edn. John Wiley & Sons, Inc., New York (1996)
8. Ng, E.K.K., Fu, A.W.-C., Wang, K.: Mining association rules from stars. In: Proceedings of the 2002 IEEE International Conference on Data Mining, pp. 322–329 (2002)
9. Wiederhold, G.: Movies database documentation (1989)
10. Xu, L.-J., Xie, K.-L.: A novel algorithm for frequent itemset mining in data warehouses. Journal of Zhejiang University - Science A 7(2), 216–224 (2006)

A Computational Intelligence Based Framework for One-Subsequence-Ahead Forecasting of Nonstationary Time Series

Vasile Georgescu

Department of Mathematical Economics, University of Craiova, A.I.Cuza str. 13, 200585 Craiova, Romania
vgeo@central.ucv.ro

Abstract. This paper proposes a mix of noise filtering, fuzzy clustering, neural mapping and predictive techniques for one-subsequence-ahead forecasting of nonstationary time series. Optionally, we may start with de-noising the time series by wavelet decomposition. A non-overlapping subsequence time series clustering procedure with a sliding window is next addressed, by using a lower-bound of the Dynamic Time Warping distance as a dissimilarity measure, when applying the Fuzzy C-Means algorithm. Afterwards, the subsequence time series transition function is learned by neural mapping, consisting of deriving, for each subsequence time series, the degrees to which it belongs to the c cluster prototypes, when the $p \cdot c$ membership degrees of the previous p subsequences are presented as inputs to the neural network. Finally, this transition function is applied to forecasting one-subsequence-ahead time series, as a weighted mean of the c cluster prototypes to which it belongs, and the S&P 500 data are used for testing.

Keywords: Computational intelligence, Subsequence time series fuzzy clustering, Neural mapping, One-subsequence-ahead forecasting of time series.

1 Introduction

The prediction of financial markets is a very complex task, because the financial time series are inherently noisy, non-stationary, and deterministically chaotic (i.e., short-term random but long-term deterministic). In principle, stock trading can be profitable if the direction of price movement can be predicted consistently. However, due to the "near-random-walk" behavior of stock prices, many experimental works show little evidence of predictability when out-of-sample forecasts are considered.

In order to ameliorate the stock market forecasting accuracy, numerous computational intelligence based techniques have been proposed previously. Among them, feedforward and recurrent neural networks (NNs) gained increasing popularity. Hybridizations of NNs and genetic algorithms (GAs) have also been proposed in an attempt to avoid the local convergence of the gradient descent algorithms and thus to accurately predict the stock price index and the direction of its change. They, however, did not bear outstanding prediction accuracy partly because of the tremendous noise and non-stationary characteristics in stock market data.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 187–194, 2010.

On the other hand, the presence of short-term randomness suggests that larger profits can be consistently generated if long-term movements in the stock are accurately predicted rather than short-term movements. Unfortunately, most of the proposed models focused on the accurate forecasting of the levels (i.e. value) of the underlying stock index (e.g., the next day's closing price forecast). Actually, the absolute value of a stock price is usually not as interesting as the shape of up and down movements. As an alternative to *one-value-ahead forecast*, the approach in this paper proposes a novel *one-subsequence-ahead forecasting technique*, which focuses on the predictability of the direction of stock index movement. Our approach also differs from other studies that consider the sign of movements and thus convert the prediction problem into a classification task, which can be carried out with classification tools, such as Support Vector Machines, random forest, logit models and so on.

The proposed framework consists of four stages: the preprocessing stage; the subsequence time series fuzzy clustering stage; the neural mapping based learning stage of the subsequence time series fuzzy transition function; the one-subsequence-ahead time series forecasting stage.

2 Time Series Preprocessing

This stage consists of de-noising data by wavelet decomposition and some other transformations that rely heavily on the selection of a distance measure for clustering.

The Discrete Wavelet Transform (DWT, [8]) uses scaled and shifted versions of a mother wavelet function, usually with compact support, to form either an orthonormal basis (Haar wavelet, Daubechies) or a bi-orthonormal basis (Symlets, Coiflets). Wavelets allow cutting up data into different frequency components (called approximations and details), and then studying each component with a resolution matched to its scale. They can help de-noise inherently noisy data such as financial time series through wavelet shrinkage and thresholding methods, developed by David Donoho ([2]). The idea is to set to zero all wavelet coefficients corresponding to details in the data set that are less than a particular threshold. These coefficients are used in an inverse wavelet transformation to reconstruct the data set. An important advantage is that the de-noising is carried out without smoothing out the sharp structures and thus can help to increase both the clustering accuracy and predictive performance.

Care has to be taken in choosing suitable transformations such that the time series distance measure chosen in the clustering stage is meaningful to the application. Normalization of data is common practice when using Fuzzy C-Means, which means applying scaling and vertical translation to the time series as a whole. Moreover, as we already mentioned, the absolute value of a stock price is not as interesting as the shape of up and down movements. Thus, for allowing stock prices comparisons subsequence by subsequence, a local translation is also necessary, in such a way to have each subsequence starting from zero. A subset of 8192 daily closing prices drawn from the S&P 500 stock index data and used for training, as well as the normalized and de-noised data are shown in Fig. 1, were a level 5 decomposition with Sym8 wavelets and a fixed form soft thresholding were used.

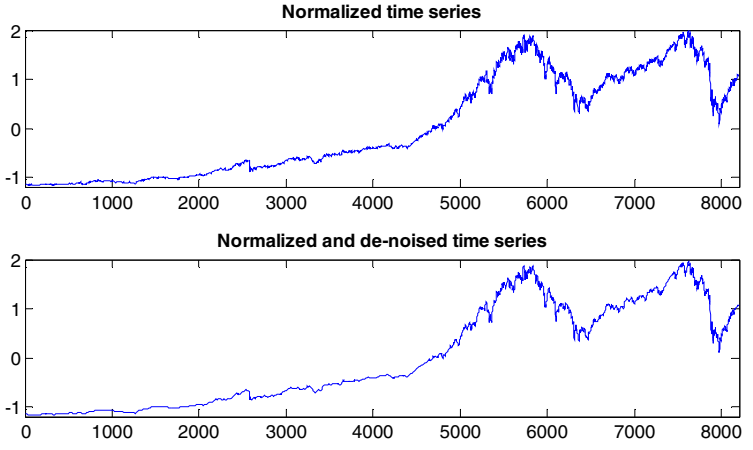


Fig. 1. A normalized and de-noised data subset, drawn from the S&P 500 stock index data

3 Subsequence Time Series Fuzzy Clustering

The idea in subsequence time series (STS) clustering is as follows. Just a single long time series is given at the start of the clustering process, from which we extract short series with a sliding window. The resulting set of subsequences are then clustered, such that each time series is allowed to belong to each cluster to a certain degree, because of the fuzzy nature of the fuzzy c-means algorithm we use. The window width and the time delay between consecutive windows are two key choices. The window width depends on the application; it could be some larger time unit (e.g., 32 days for time series sampled as daily S&P 500 stock index, in our application). Overlapping or non-overlapping windows can be used. If the delay is equal to the window width, the problem is essentially converted to non-overlapping subsequence time series clustering. We will follow this approach, being motivated by the Keogh's criticism presented in [6], where using overlapping windows has been shown to produce meaningless results, due to a surprising anomaly: cluster centers obtained using STS clustering closely resemble "sine waves", irrespective of the nature of original time series itself, being caused by the superposition of slightly shifted subsequences.

Using larger time delays for placing the windows does not really solve the problem as long as there is some overlap. Also, the less overlap, the more problematic the choice of the offsets becomes.

Since clustering relies strongly on a good choice of the dissimilarity measure, this leads to adopting an appropriate distance, depending on the very nature of the subsequence time series.

Let $S = y_m, \ldots, y_{m+w-1}$ be a subsequence with length w of time series $Y = y_1, \ldots, y_n$, where $1 \le m \le n - w + 1$. Subsequences will be represented as vectors in a w-dimensional vector space. For relatively short time series, shape-based distances, such as L_p norms, are commonly used to compare their overall appearance. The Euclidean distance (L_2) is the most widely used shape-based

distance. Other L_p norms can be used as well, such as Manhattan (L_1) and Maximum (L_∞), putting different emphasis on large deviations.

There are several pitfalls when using an L_p distance on time series: it does not allow for different *baselines* in the time sequences; it is very sensitive to *phase shifts* in time; it does not allow for *acceleration and deceleration* along the time axis (time warping). Another problem with L_p distances of time series is when scaling and translation of the amplitudes or the time axis are considered, or when outliers and noisy regions are present.

A number of non-metric distance measures have been defined to overcome some of these problems. Small distortions of the time axis are commonly addressed with non-uniform time warping, more precisely with Dynamic Time Warping (DTW, [1], [5]). The DTW distance is an extensively used technique in speech recognition and allows warping of the time axes (acceleration–deceleration of signals along the time dimension) in order to align the shapes of the two time series better. The two series can also be of different lengths. The optimal alignment is found by calculating the shortest warping path in the matrix of distances between all pairs of time points under several constraints (boundary conditions, continuity, monotonicity).

The warping path is also constrained in a global sense by limiting how far it may stray from the diagonal. The subset of the matrix that the warping path is allowed to visit is called the warping window. The two most common constraints in the literature are the Sakoe-Chiba band and the Itakura parallelogram. We can view a global or local constraint as constraining the indices of the warping path $w_k = (i, j)_k$, such that $j - r \le i \le j + r$, where r is a term defining the allowed range of warping, for a given point in a sequence. In the case of the Sakoe-Chiba band (see Fig. 2), r is independent of i; for the Itakura parallelogram, r is a function of i.

DTW is a much more robust distance measure for time series than L_2, allowing similar shapes to match even if they are out of phase in the time axis. Unfortunately, however, DWT is calculated using dynamic programming with time complexity $O(n^2)$. Recent approaches focus more on approximating the DTW distance by bounding it from below. For example, a novel, linear time (i.e., with complexity reduced to $O(n)$), lower bound of the DTW distance, was proposed in [7]. The

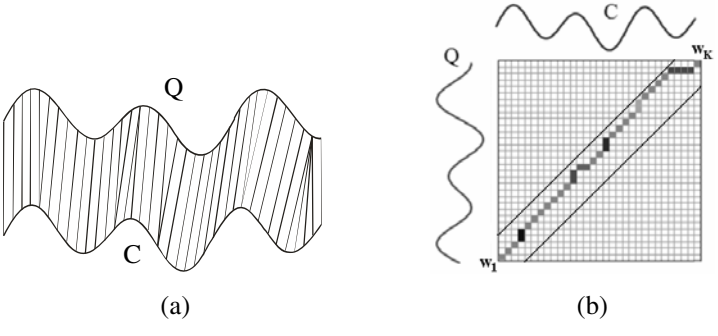


Fig. 2. (a) Aligning two time sequences using DTW. (b) Optimal warping path with the Sakoe-Chiba band as global constraints.

intuition behind the approach is the construction of a special "envelope" around the query. It can be shown that the Euclidean distance between a potential match and the nearest orthogonal point on the envelope lower bounds the DTW distance. To index this representation, an approximate bounding envelope is created.

Let $Q = \{q_1, \ldots, q_n\}$ and $C = \{c_1, \ldots, , c_m\}$ be two subsequences and $w_k = (i, j)_k$ be the warping path, such that $j - r \le i \le j + r$, where r is a term defining the range of warping for a given point in a sequence. The term r can be used to define two new sequences, L and U, where $L_i = \min(q_{i-r} : q_{i+r})$, $U_i = \max(q_{i-r} : q_{i+r})$, with L and U standing for *Lower* and *Upper*, respectively. An obvious but important property of L and U is the following: $\forall i, U_i \ge q_i \ge L_i$. Given L and U, a lower bounding measure for DTW can now be defined (see Fig. 3):

$$LB\text{-}Keogh(Q, C) = \sqrt{\begin{cases} \sum_{i=1}^{n} (c_i - U_i)^2 & if \quad c_i > U_i \\ \sum_{i=1}^{n} (c_i - U_i)^2 & if \quad c_i < U_i \\ 0 & otherwise \end{cases}}. \tag{1}$$

We are now going to generalize the fuzzy c-means algorithm to subsequence time series clustering. In this particular context, the entities to be clustered, denoted by x_k, and the cluster prototypes (centroids), denoted by v_i, are both set-defined objects, i.e. subsequence time series. The centroids are computed as weighted means, where the weights, denoted by u_{ik}, are the fuzzy membership degrees to which each subsequence belongs to a cluster. Both the *DTW* and *LB-Keogh* distances outperform L_2 and thus qualify better to be used with the fuzzy c-means algorithm. However, the *LB-Keogh*'s lower bound of DTW distance has been preferred, due to its linear time complexity. Fig. 4 plots the cluster centroids (prototypes) and the subsequence time series grouped around each centroid.

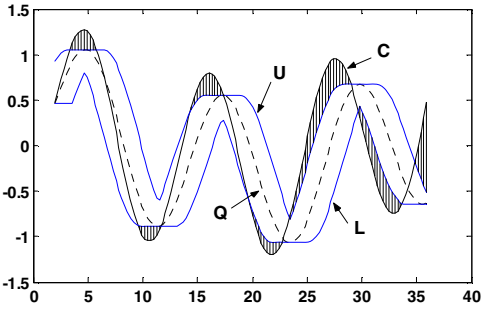


Fig. 3. The lower bounding function LB-Keogh(Q,C). The original sequence Q is enclosed in the bounding envelope of U and L.

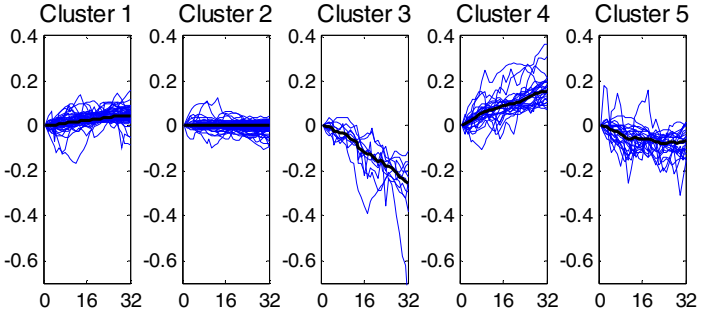


Fig. 4. The cluster centroids and the subsequence time series grouped around each centroid

4 Estimation of the Fuzzy Transition Function between Clusters by Neural Mapping

At this stage, a fuzzy transition function between clusters must be learned, which is a nonlinear vector function mapping a number of p c-dimensional membership degree vectors $\mu(STS_{t-j+1})$, $j=1,\ldots,p$, into a c-dimensional membership degree vector $\mu(STS_{t+1})$, i.e., $\mu(STS_{t+1}) = f\left(\mu(STS_t),\ldots,\mu(STS_{t-p+1})\right)$, where STS_{t+1} is the subsequence time series to be predicted. In our experiment, $p=2$ and $c=5$.

Neural networks are well known for their capability to be universal approximators (i.e., to estimate almost any computable function on a compact set arbitrarily closely, provided that enough experimental data are available). Actually, we use a *multilayer perceptron* network with two layers: one hidden layer with the tan-sigmoid transfer function and one output layer with the log-sigmoid transfer function (the latter allows constraining the output of the network between 0 and 1). The dimensions of input and

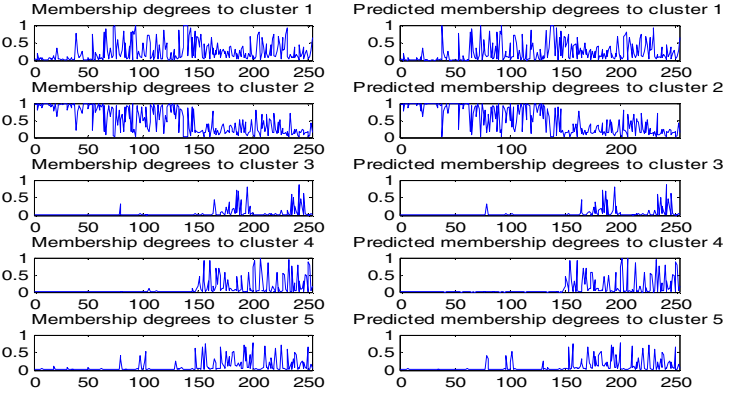


Fig. 5. Accurate neural mapping: actual and predicted membership degrees to which each of the 256 subsequence time series belongs to one of the 5 clusters

output spaces are $p \cdot c$ and c, respectively. This neural architecture is known to have the capability of approximating any nonlinear function with a finite number of discontinuities arbitrarily well, given sufficient neurons in the hidden layer.

5 One-Subsequence-Ahead Forecasting of Time Series

The one-subsequence-ahead forecast can then be obtained as a weighted mean of the c cluster prototypes (v_i), each one representing a subsequence time series:

$$STS_{t+1} = \sum_{i=1}^{c} v_i \cdot \mu\left(STS_{t+1}\right). \tag{2}$$

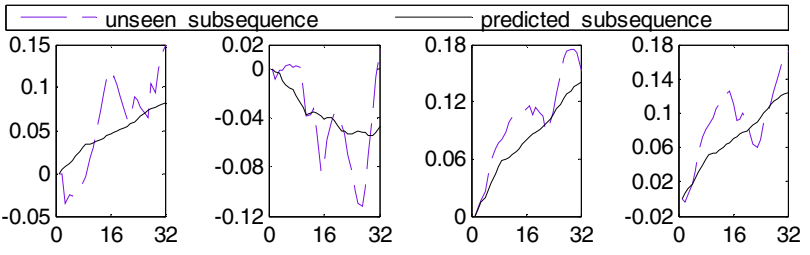


Fig. 6. One-subsequence-ahead forecasting of 4 unseen subsequences (i.e., not included in the training dataset). This covers 4*32=128 daily stock price index value forecasts.

Each forecast is a 32-length subsequence, obtained as follows: the membership degrees of the previous p subsequences to each of the c clusters are first computed; the membership degrees to which the next subsequence belongs to each cluster are then found by neural mapping; these membership degrees are finally used to compute the weighted mean of the c cluster prototypes (v_i), representing the forecast.

6 Conclusion

Predicting price levels is an intriguing, challenging, and admittedly risky endeavor. Technical analysis uses trend following strategies to forecast future price movements and to infer trading decision rules, based on the assertion that price changes have inertia. Although experimental works show little evidence of predictability, many traders consider accuracy rates of about (or greater than) 55% to be consistently profitable. However, one-value-ahead forecasting of price levels is not as useful as the shape of long-term up and down movements, due to their inherent short-term randomness. The approach in this paper proposed a novel one-subsequence-ahead forecasting framework, based on a mix of computational intelligence techniques that allow the prediction of stock index movements in a more robust way, focusing on predicting one price subsequence rather than one price level at a time.

References

1. Berndt, D.J., Clifford, J.: Finding patterns in time series: A dynamic programming approach. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 229–248. AAAI Press, Menlo Park (1996)
2. Donoho, D.: Nonlinear Wavelet Methods for Recovery of Signals, Densities and Spectra from Indirect and Noisy Data. In: Daubechies, I. (ed.) Different Perspectives on Wavelets, Proceeding of Symposia in Applied Mathematics, vol. 47, pp. 173–205. Amer. Math. Soc., Providence (1993)
3. Georgescu, V.: Generalizations of Fuzzy C-Means Algorithm to Granular Feature Spaces, based on Underlying Fuzzy Metrics: Issues and Related Works. In: 13th IFSA World Congress and 6th Conference of EUSFLAT, Lisbon, Portugal, pp. 1791–1796 (2009)
4. Georgescu, V.: A Time Series Knowledge Mining Framework Exploiting the Synergy between Subsequence Clustering and Predictive Markovian Models. Fuzzy Economic Review XIV(1), 41–66 (2009)
5. Keogh, E., Pazzani, M.J.: Scaling up dynamic time warping to massive datasets. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 1–11. Springer, Heidelberg (1999)
6. Keogh, E., Lin, J., Truppel, W.: Clustering of time series subsequences is meaningless: implications for previous and future research. In: 3rd IEEE International Conference on Data Mining, pp. 115–122 (2003)
7. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems 7, 358–386 (2005)
8. Mallat, S.G., Peyré, G.: A Wavelet Tour of Signal Processing: The Sparse Way, 3rd edn. Academic Press, London (2009)

Non-hierarchical Clustering of Decision Tables toward Rough Set-Based Group Decision Aid

Masahiro Inuiguchi, Ryuta Enomoto, and Yoshifumi Kusunoki

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
inuiguti@sys.es.osaka-u.ac.jp

Abstract. In order to analyze the distribution of mind-sets (collections of evaluations) in a group, a hierarchical clustering of decision tables has been examined. By the method, we know clusters of mind-set but the clusters are not always optimal in some criterion. In this paper, we develop non-hierarchical clustering techniques for decision tables. In order to treat positive and negative evaluations to a common profile, we use a vector of rough membership values to represent individual opinion to a profile. Using rough membership values, we develop a K-means method as well as fuzzy c-means methods for clustering decision tables. We examined the proposed methods in clustering real world decision tables.

Keywords: Decision table, rough membership function, clustering, K-means, fuzzy c-means.

1 Introduction

The rough set approaches [8] to data mining and knowledge discovery are popular and applied to various fields such as expert systems, decision analysis, medical informatics, civil engineering, and so on. They treat the inconsistency among data reasonably and induce decision rules as well as important attributes. In this paper, we take a step toward the application of rough sets to group decision support. So far, rough set approaches have been applied mainly to single decision tables with homogeneous evaluations of objects. However, as is often encountered in the real world, individual evaluations are diversified. Then it is not always a good idea to put all individual opinions in a single decision table.

Some researchers [1,5] have proposed to treat the mind-set of each individual as a decision table, i.e., a collection of evaluations by a person, and investigated the analysis of multiple decision tables. A group of researchers [1,6] in Kansei engineering initiated this approach. They proposed merging individual rules to obtain rules supported by more designers in the setting of product design development. The individual design rules are induced from each decision table by the rough set approach and thus, it may require a big computational effort. Inuiguchi and Miyajima [5] and Inuiguchi [2] have been proposed to define rough sets under multiple decision tables in order to reduce the computational load. Once rough sets are defined under multiple decision tables, we may extend rough

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 195–206, 2010.

set approaches to cases with multiple decision tables. From this point of view, Inuiguchi [3] has proposed rule induction method under multiple decision tables. Yamamoto and Inuiguchi [9] has extended this approach to cases when condition and decision attributes are ordinal and monotonous.

Considering the variety of individual opinions and mind-sets, Inuiguchi and Furudono [4] proposed clustering decision tables so that each cluster is composed of decision tables representing similar mind-sets. By such a clustering method, we may roughly figure out the distribution of mind-sets in a population. Inuiguchi and Furudono's approach is based on an agglomerative hierarchical clustering algorithm (AHC algorithm, for short) so that the result is not always optimal in the sense of a certain evaluation function under a selected number of clusters. Moreover the adopted similarity measure does not reflect the differences between individual opinions of cluster members and cluster opinions.

In this paper, we develop a non-hierarchical clustering algorithm for multiple decision tables. More precisely, we investigate K-means clustering and fuzzy c-means clustering for multiple decision tables. We develop those clustering methods for multiple decision tables. We examine the performances of those methods by numerical experiments using real world data.

In next section, we briefly introduce the rough set theory and the previous clustering approach. In Section 3, we develop the non-hierarchical clustering algorithms. In Section 4, we describe the numerical experiments and show the results. The advantages of the proposed approaches are emphasized.

2 Rough Set Theory and the Previous Approach

2.1 Decision Tables and Rough Sets

Rough sets are often applied to decision tables. A decision table is composed of a set of objects U, a set of condition attributes C and a decision attribute d. A decision table is denoted by $(U, C \cup \{d\})$. We regard each attribute $a \in C \cup \{d\}$ as a function from U to V_a, where V_a is the set of attribute values a takes. An example of a decision table is given in Table 1. In Table 1, we have $U = \{u_i,\ i = 1, 2, \ldots, 10\}$, $C = \{\text{Design}, \text{Function}, \text{Size}\}$ and $d = \text{Dec.}$ (Decision).

Given a decision table $(U, C \cup \{d\})$, we define a condition attribute pattern which we call a profile $Inf_C(u)$ of an object $u \in U$ by

$$Inf_C(u) = \bigcup_{a \in C} \{\langle a, a(u) \rangle\}, \tag{1}$$

where $a(u)$ shows the attribute value of u with respect to attribute $a \in C \cup \{d\}$. The set W_C^U of all profiles in the given decision table is defined by

$$W_C^U = \{Inf_C(u) : u \in U\}. \tag{2}$$

Let V_d be the set of decision attribute values. An opinion can be seen as an evaluation of a profile so that a pair (w, v_d), $w \in W_C^U$, $v_d \in V_d$ is regarded as

Table 1. An example of decision table

object	Design	Function	Size	Dec.
u_1	classic	simple	compact	accept
u_2	classic	multiple	compact	accept
u_3	classic	multiple	normal	reject
u_4	modern	simple	compact	reject
u_5	modern	simple	normal	reject
u_6	classic	multiple	compact	accept
u_7	modern	multiple	normal	reject
u_8	classic	simple	compact	accept
u_9	classic	multiple	normal	accept
u_{10}	modern	multiple	normal	reject

Table 2. A profile-based decision table

profile	Design	Function	Size	σ
w_1	classic	simple	compact	(2,0)
w_2	classic	multiple	compact	(2,0)
w_3	classic	multiple	normal	(1,1)
w_4	modern	simple	compact	(0,1)
w_5	modern	simple	normal	(0,1)
w_6	modern	multiple	normal	(0,2)

an opinion in this paper. Then frequency function σ_C and rough membership function μ_C are defined as follows for an opinion (w, v_d),

$$\sigma_C(w, v_d) = |Inf_C^{-1}(w) \cap d^{-1}(v_d)|, \tag{3}$$

$$\mu_C(w, v_d) = \frac{|Inf_C^{-1}(w) \cap d^{-1}(v_d)|}{|Inf_C^{-1}(w)|}, \tag{4}$$

where Inf_C^{-1} and d^{-1} are inverse images of Inf_C and d, respectively, i.e., $Inf_C^{-1}(w) = \{u \in U : Inf_C(u) = w\}$ and $d^{-1}(v_d) = \{u \in U : d(u) = v_d\}$. $\sigma_C(w, v_d)$ shows the number of objects whose profiles are w and whose decision attribute values are v_d. $\mu_C(w, v_d)$ shows the ratio of objects which take decision attribute value v_d to all objects whose profiles are w. Given $\sigma_C(w, v_d)$ for every $v_d \in V_d$, we obtain $\mu_C(w, v_d)$ as

$$\mu_C(w, v_d) = \frac{\sigma_C(w, v_d)}{\sum_{v_d \in V_d} \sigma_C(w, v_d)}. \tag{5}$$

However $\sigma_C(w, v_d)$ cannot be obtained from $\mu_C(w, v_d)$ for every $v_d \in V_d$. We can rewrite a decision table described by profiles $w \in W_C^U$ and frequencies $\{\sigma_C(w, v_d) : v_d \in V_d\}$. For example, the decision table shown in Table 1 can be rewritten as a table shown in Table 2. In Table 2, each entry in column 'σ' shows a vector $(\sigma_C(w_j, \text{accept}), \sigma_C(w_j, \text{reject}))$. In rough set analysis, the order of objects appearing in a decision table does not affect the results of the analysis. Then having a decision table described by profiles $w \in W_C^U$ as in Table 2 is equivalent to having a usual decision table as in Table 1. From this fact, we assume that decision tables are given by using profiles in what follows.

The lower and upper approximations composing a rough set with respect to $v_d \in V_d$ are defined as sets of profiles instead of objects by;

$$\underline{C}(v_d) = \{w_i \in W_C^U : \mu_C(w_i, v_d) = 1\}, \tag{6}$$

$$\overline{C}(v_d) = \{w_i \in W_C^U : \mu_C(w_i, v_d) > 0\}. \tag{7}$$

The relations of $\underline{C}(v_d)$ and $\overline{C}(v_d)$ with usual lower and upper approximations $C_*(\hat{v_d})$ and $C^*(\hat{v_d})$ are given as

$$C_*(\hat{v_d}) = Inf_C^{-1}(\underline{C}(d(\hat{v_d}))) = Inf_C^{-1}(\underline{C}(v_d)), \tag{8}$$
$$C^*(\hat{v_d}) = Inf_C^{-1}(\overline{C}(d(\hat{v_d}))) = Inf_C^{-1}(\overline{C}(v_d)), \tag{9}$$

where $\hat{v_d}$ is a set of objects taking decision attribute value $v_d \in V_d$, i.e., $\hat{v_d} = \{u \in U \mid d(u) = v_d\}$. A rough set of $\hat{v_d}$ is often defined by a pair $(C_*(\hat{v_d}), C^*(\hat{v_d}))$. In this paper, a pair $(\underline{C}(v_d), \overline{C}(v_d))$ is called a rough set with respect to v_d.

2.2 Agglomerative Hierarchical Clustering of Decision Tables

In this paper, we assume that p decision tables T_i, $i = 1, 2, \ldots, p$ are given. Those decision tables have common condition attributes C and a common decision attribute d. However, objects can be different among those decision tables. Accordingly, we assume that U_i is the object set of decision table T_i so that the profile set of decision table T_i is written by $W_C^{U_i}$. For the sake of simplicity, we define $W_C = \bigcup W_C^{U_i}$.

Given a similarity between decision tables, we may apply AHC algorithms. Then we describe the similarity proposed by Inuiguchi and Furudono [4]. The pairs of elements of lower approximations and the corresponding decision attribute value can be considered the essential parts of decision makers' opinions. They proposed a similarity using lower approximations.

First, a representative of a cluster of decision tables is defined. Let $\boldsymbol{T} = \{T_1, \ldots, T_p\}$ be a cluster of decision tables. The relative frequency of an opinion (w, v_d) can be defined by

$$\tau^{\boldsymbol{T}}(w, v_d) = \frac{\left|\{T \in \boldsymbol{T} : w \in \underline{C}^T(v_d)\}\right|}{|\boldsymbol{T}|}, \tag{10}$$

where $|Z|$ shows the cardinality of a set Z and $\underline{C}^T(v_d)$ is the lower approximation of v_d under a decision table T. Then the representative of cluster $\boldsymbol{T}$ is defined as a distribution of relative frequencies over profiles.

Given a decision table $T_k \in \boldsymbol{T}$, the reflection degree of T_k in the cluster $\boldsymbol{T}$ can be defined as a weighted average of relative frequencies $\tau^{\boldsymbol{T}}(w, v_d)$ of profiles (w, v_d) such that $w \in \underline{C}^{T_k}(v_d)$. The weights are determined by frequencies of profiles appeared in T_k. Namely, the reflection degree of T_k, $R(T_k)$ is defined by

$$R_q(T_k) = \frac{\displaystyle\sum_{v_d \in V_d} \sum_{w \in \underline{C}^{T_k}(v_d)} \left| Inf_{T_k,C}^{-1}(w) \cap d_{T_k}^{-1}(v_d) \right| \tau^{\boldsymbol{T}}(w, v_d)^q}{\displaystyle\sum_{v_d \in V_d} \sum_{w \in \underline{C}^{T_k}(v_d)} \left| Inf_{T_k,C}^{-1}(w) \cap d_{T_k}^{-1}(v_d) \right|}, \tag{11}$$

where $Inf_{T_k,C}^{-1}(w)$ is a set of objects which have profile w in decision table T_k. $d_{T_k}^{-1}(v_d)$ is a set of objects whose decision attribute values take v_d in decision

table T_k. q is a constant parameter introduced to control the significance of high degree concurrence. However, $q = 1$ seems appropriate as far as Inuiguchi and Furudono's experiments [4]. We assume that $q = 1$ in what follows.

Then the average of $R_q(T)$ over all $T \in \boldsymbol{T}$ can be regarded as the strength forming a cluster $\boldsymbol{T}$. Namely, it can be defined as

$$F_q(\boldsymbol{T}) = \frac{\sum_{T \in \boldsymbol{T}} R_q(T)}{|\boldsymbol{T}|}. \tag{12}$$

The higher $F_q(\boldsymbol{T})$ is, the easier cluster $\boldsymbol{T}$ forms. Then the similarity among decision tables in a cluster $\boldsymbol{T}$ can be defined by $F_q(\boldsymbol{T})$.

For the sake of utilizing this idea in the AHC algorithm, we can define the similarity between two clusters $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ by

$$S(\boldsymbol{T}_1, \boldsymbol{T}_2) = F_q(\boldsymbol{T}_1 \cup \boldsymbol{T}_2). \tag{13}$$

Using this similarity between two clusters of decision tables, we can classify decision tables by the following AHC algorithm:

Step 1. Define $q = 1$. Let $cl = p$ and $\boldsymbol{T}_i = \{T_i\}$, $i = 1, 2, \ldots, n$, where p is the number of given decision tables. Calculate $S_q(\boldsymbol{T}_i, \boldsymbol{T}_j)$ for all i and j such that $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$.

Step 2. Combine two clusters having the highest similarity. Namely, let

$$S_q(\boldsymbol{T}_s, \boldsymbol{T}_r) = \max_{i,j:i \neq j} S_q(\boldsymbol{T}_i, \boldsymbol{T}_j)$$

and add $\boldsymbol{T}' = \boldsymbol{T}_s \cup \boldsymbol{T}_r$ and erase $\boldsymbol{T}_s$ and $\boldsymbol{T}_r$. Update $cl = cl - 1$. If $cl = 1$ then output the dendrogram and terminate the algorithm.

Step 3. Calculate $S_q(\boldsymbol{T}', \boldsymbol{T}_j)$, for all $j \in \{1, 2, \ldots, p\}$ such that $j \neq, r$. Repeat the procedure until the termination at Step 2.

3 Non-hierarchical Clustering of Decision Tables

3.1 The Dissimilarity Measure

In order to apply K-means/fuzzy c-means clustering algorithms to p decision tables, the definitions of cluster center (representative) and some similarity/dissimilarity measure between a decision table and a cluster center are requested. This is because the clustering algorithms usually optimize the sum of similarity/dissimilarity measures between decision tables and corresponding cluster centers. In order to utilize the similarity measure defined by (13) with singletons $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, we should define the cluster center as a decision table. In some way, it is possible but the resulting optimization problem becomes intractable due to the discreteness and complexity of the cluster center.

A similarity/dissimilarity measure by which the optimization problem becomes tractable is preferable. In order to define such a similarity/dissimilarity

measure, we identify a decision table T_i with a totally ordered set of rough membership values (a vector of rough membership values),

$$\boldsymbol{\mu}^{T_i} = \left\{ \mu_C^{T_i}(w, v_d) : w \in W_C,\ v_d \in V_d \right\}, \tag{14}$$

where $\mu_C^{T_i}(w, v_d)$ is the rough membership value of profiles $w \in W_C^{T_i}$ at decision attribute value v_d under decision table T_i. When opinion (w, v_d) does not appear in T_i, we define $\mu_C^{T_i}(w, v_d) = 0$. Note that if profile w is missing in T_i, we have $\sum_{v_d \in V_d} \mu_C^{T_i}(w, v_d) = 0$. The order is defined by profile w and decision attribute value v_d.

The center of q-th cluster can be defined also by a totally ordered set of variables $\tau_q(w, v_d)$ in the unit intervals $[0, 1]$, i.e.,

$$\boldsymbol{\tau}_q = \{\tau_q(w, v_d) : w \in W_C,\ v_d \in V_d\}. \tag{15}$$

As the result, we may define a dissimilarity measure, more specifically, a distance between decision table T_i and the clustering center $\boldsymbol{\tau}_q$ by

$$D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q) = \sum_{w \in W_C} \sum_{v_d \in V_d} \left(\mu_C^{T_i}(w, v_d) - \tau_q(w, v_d) \right)^2. \tag{16}$$

The adoption of rough membership value brings not only in the tractability of the optimization problems but also in the consideration of opinions with $\mu_C^{T_i}(w, v_d) < 1$.

3.2 K-Means Clustering

Let us divide p decision tables into k clusters. Using rough membership values $\boldsymbol{\mu}^{T_i}$ and cluster centers $\boldsymbol{\tau}_q$, we may use the following objective function to be minimized for K-means clustering:

$$J_{\mathrm{km}}(U, V) = \sum_{i=1}^{p} \sum_{q=1}^{k} u_{iq} D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q), \tag{17}$$

where $U = \{u_{iq} : i = 1, 2, \ldots, p,\ q = 1, 2, \ldots, k\}$ and $V = \{\boldsymbol{\tau}_q : q = 1, 2, \ldots, k\}$. We impose the following constraints on U:

$$\mathcal{U}_{\mathrm{km}} = \left\{ \sum_{q=1}^{k} u_{iq} = 1,\ u_{iq} \in \{0, 1\}, i = 1, 2, \ldots, p,\ q = 1, 2, \ldots, k \right\}. \tag{18}$$

As is in the literature [7], this clustering problem is solved by the alternative optimization technique and the following algorithm is obtained:

Step 1. Generate initial values $\bar{U} \in \mathcal{U}_{\mathrm{km}}$ for U.
Step 2. Calculate $\bar{V} = \arg\min_V J_{\mathrm{km}}(\bar{U}, V)$.
Step 3. Calculate $\bar{U} = \arg\min_{U \in \mathcal{U}_{\mathrm{km}}} J_{\mathrm{km}}(U, \bar{V})$.

Step 4. If $\bar{U}$ or $\bar{V}$ is convergent. stop; else go to Step 2.

Optimal solutions at Steps 2 and 3 are respectively obtained as

$$\bar{\tau}_q(w, v_d) = \sum_{i=1}^{p} \bar{u}_{iq} \mu_C^{T_i}(w, v_d), \quad \bar{u}_{iq} = \begin{cases} 1, \text{ if } q = \arg\min_l D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_l), \\ 0, \text{ otherwise.} \end{cases} \tag{19}$$

There are several convergence criteria [7] but we adopt the following rule in this paper. Let $\hat{U} = \{\hat{u}_{iq} : i = 1, 2, \ldots, p,\ q = 1, 2, \ldots, k\}$ be the value of U in the previous iteration and ε be a sufficiently small positive constant value. If $\max_{i,q} |\bar{u}_{iq} - \hat{u}_{iq}| < \varepsilon$, we consider that $\bar{U} = \{\bar{u}_{iq} : i = 1, 2, \ldots, p,\ q = 1, 2, \ldots, k\}$ is convergent.

3.3 Fuzzy c-Means Clustering

In fuzzy c-means clustering, a fuzzy membership of a decision table to a cluster is allowed. The objective function to be minimized is defined by

$$J_{\text{fcm}}(U, V) = \sum_{i=1}^{p} \sum_{q=1}^{c} (u_{iq})^m D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q), \tag{20}$$

where p decision tables are divided into c clusters and $m > 1$. Because of the fuzziness in membership u_{iq}, the constraint on U is relaxed to

$$\mathcal{U}_{\text{fcm}} = \left\{ \sum_{q=1}^{k} u_{iq} = 1,\ u_{iq} \in [0, 1], i = 1, 2, \ldots, p,\ q = 1, 2, \ldots, k \right\}. \tag{21}$$

The algorithm is the same as K-means clustering algorithm with replacements $\mathcal{U}_{\text{km}}$ with $\mathcal{U}_{\text{fcm}}$ and J_{km} with J_{fcm}.

Optimal solutions at Steps 2 and 3 are respectively obtained as

$$\tau_q(w, v_d) = \frac{\sum_{i=1}^{p} (\bar{u}_{iq})^m \mu_C^{T_i}(w, v_d)}{\sum_{i=1}^{p} (\bar{u}_{iq})^m}, \quad u_{iq} = \left[\sum_{r=1}^{c} \left(\frac{D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q)}{D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_r)} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{22}$$

Note that when $m = 1$, J_{fcm} degenerates to J_{km} and the clustering methods become equivalent although their constraints are different. Then $m > 1$ would be interesting in the fuzzy c-means clustering.

3.4 Fuzzy c-Means Clustering with Entropy Regularization

Viewing the objective function J_{fcm} with $m > 1$ as a kind of regularization, fuzzy c-means clustering with entropy regularization [7] has been proposed adopting an entropy function as a regularization function.

$$J_{\text{efc}}(U, V) = \sum_{i=1}^{p} \sum_{q=1}^{c} (u_{iq}) D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q) + \lambda^{-1} \sum_{i=1}^{p} \sum_{q=1}^{c} u_{iq} \log u_{iq}, \tag{23}$$

where $\lambda > 0$. The constraints on u_{iq} are the same as usual fuzzy c-means clustering, i.e., U_{fcm}.

Even in this case, the algorithm is the same as K-means clustering algorithm with replacements $\mathcal{U}_{\mathrm{km}}$ with $\mathcal{U}_{\mathrm{efc}}$ and J_{km} with J_{fcm}.

Optimal solutions at Steps 2 and 3 are respectively obtained as

$$\tau_q(w, v_d) = \sum_{i=1}^{p} \bar{u}_{iq} \mu_C^{T_i}(w, v_d), \quad u_{iq} = \frac{\mathrm{e}^{-\lambda D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_q)}}{\sum_{r=1}^{c} \mathrm{e}^{-\lambda D(\boldsymbol{\mu}^{T_i}, \boldsymbol{\tau}_r)}}. \tag{24}$$

3.5 A Modification

In the clustering results of the previously described methods, we can observe that the sum of $\bar{\tau}_r(w, v_d)$, $v_d \in V_d$ is not always one. This can be seen easily by a case that $\mu_C^{T_i}(w, v_d) = 0$ for $v_d \in V_d$, $i = 1, \ldots, p-1$ and $\mu_C^{T_p}(w, v_d) = 1/|V_d|$ for $v_d \in V_d$. From (19), (22) and (24), we know that $\bar{\tau}_r(w, v_d) = 0$ if $\bar{u}_{pr} = 0$ and $\bar{\tau}_r(w, v_d) \leq 1/|V_d|$ otherwise, and that if $\bar{u}_{pr} > 0$ for $r = 1, 2, \ldots, c$, we have $\bar{\tau}_r(w, v_d) < 1/|V_d|$ for some $r \in [1, c]$. On the other hand, for $w \in \bigcap_{i=1}^{p} W_C^{U_i}$, from (19), (22) and (24), we know $\sum_{v_d \in V_d} \bar{\tau}_r(w, v_d) = 1$.

The fact $\sum_{v_d \in V_d} \bar{\tau}_r(w, v_d) < 1$ may debase the quality of clustering results because absent profiles are improperly reflected to clustering centers $\boldsymbol{\tau}_q$, $q = 1, 2, \ldots, p$.

Two modifications are conceivable for the modification of this debasement: one is to introduce the estimated value to absent profiles as done by Inuiguchi and Miyajima [5] and Yamamoto and Inuiguchi [9] and the other is to discard the decision tables missing w on calculation of $\tau_q(w, v_d)$, $q = 1, 2, \ldots, p$, $v_d \in V_d$. In this paper, we take the latter approach. The modification is very simple. Namely, we replace W_C with $W_C^{U_i}$ in (16).

4 Examinations by Real World Data

4.1 Data Sets

To examine the performances of the proposed approaches in comparison with the previous AHC approach, we execute numerical experiments. We use two real data sets which we collected through questionnaires. Data-set 1 concerns the student preference among Japanese companies to be employed. We collected data from 18 university students belonging to laboratories in systems engineering field. Twenty-one companies were selected and each student evaluated 12 companies randomly chosen from 21 companies. The questionnaire includes company's business activity (conservative/innovative), internationality (high/low), job offer (specialist/regular), contribution to society (high/low) and employment wish (high/medium/low).

Data-set 2 concerns preferences among simple pictorial figures. We collected data from 21 people. Each examinee evaluates his/her fondness (like/dislike) of

36 pictorial figures. The 36 pictorial figures are all combinations of 4 shapes, 3 colors and 3 patterns.

Due to the page restriction, we describe the part of results in the experiments with Data-set 1.

4.2 Experiments

We applied k-means, fuzzy c-means, entropy based fuzzy c-means, modified fuzzy c-means and the previous AHC method to Data-set 1. Using 10 different random seeds, we prepared 10 initializations for those methods. For the number of clusters, we consider three cases $p = 3, 4, 5$. The parameters m and λ are varied appropriately: m for fuzzy c-means is varied from 1.1 to 1.5 with step size 0.1, λ for entropy based fuzzy c-means from 0.9 to 1.3 with step size 0.1 and m for modified fuzzy c-means varied from 1.2 to 2.0 with step size 0.2.

We calculate 10 criteria for evaluations of clustering performances. Among them, in this paper, we describe three criteria. Adopting the idea of variable precision rough set model [10], we define the opinions (w, v_d) of decision table T_i by $\mu_C^{T_i}(w, v_d) \geq \alpha$ and opinions (w, v_d) in a cluster G_q by $\tau_q(w, v_d) \geq \beta$ with predetermined values $\alpha, \beta \in [0.5, 1]$. For the sake of simplicity, we use the following profile sets:

$$\underline{C}_\alpha^{T_i}(v_d) = \{w \in V_C^{U_i} \mid \mu_C^{T_i}(w, v_d) \geq \alpha\}, \quad \underline{C}_\beta^{G_q}(v_d) = \{w \in V_C^{U_i} \mid \tau_q(w, v_d) \geq \beta\}. \tag{25}$$

$\underline{C}_\alpha^{T_i}(v_d)$ corresponds to the set of positive members with respect to v_d in the variable precision rough set model. We define positive members of cluster G_q by

$$T_i \in G_q \overset{\text{def}}{\Leftrightarrow} u_{iq} \geq 0.5, \tag{26}$$

where, $T_i \in G_q$ stands for T_i is a positive member of G_q. Note that we may have decision tables T_i which are not positive members of any clusters and decision tables T_i which are positive members of two clusters.

Assuming that we cluster p decision tables to c clusters G_q, $q = 1, 2, \ldots, c$, the following three criteria are considered:

Concurrence of opinions between a cluster and its members: A concurrence magnitude of opinions between G_q and $T_i \in G_q$ can be defined by

$$Cnc(G_q, T_i) = \sum_{v_d \in V_d} \sum_{w \in \underline{C}_\alpha^{T_i}(v_d) \cap \underline{C}_\beta^{G_q}(v_d)} \left| Inf_{T_i,C}^{-1}(w) \cap d_{T_i}^{-1} \right|. \tag{27}$$

Then the total concurrence magnitude of the clustering results can be defined by

$$TCnc = \sum_{q=1}^{c} \sum_{T_i \in G_q} Cnc(G_q, T_i). \tag{28}$$

Conflict of opinions between a cluster and its members: A conflict magnitude of opinions between G_q and $T_i \in G_q$ can be defined by

$$Cnf(G_q, T_i) = \sum_{v_d{}^1 \in V_d} \sum_{\substack{v_d{}^2 \in V_d \\ v_d{}^2 \neq v_d{}^1}} \left| \underline{C}_\alpha^{T_i}(v_d^1) \cap \underline{C}_\beta^{G_q}(v_d^2) \right|. \tag{29}$$

Then the total conflict magnitude of the clustering results can be defined by

$$TCnf = \sum_{q=1}^{c} \sum_{T_i \in G_q} Cnf(G_q, T_i). \tag{30}$$

Ratio of between-cluster difference to within-cluster difference: The difference between clusters G_q and G_r is defined by

$$BCl(G_q, G_r) = \sum_{w \in V_C} \sum_{v_d \in V_d} (\tau_q(w, v_d) - \tau_r(w, v_d))^2. \tag{31}$$

The total between-cluster difference is defined by

$$TBCl = \sum_{q=1}^{c-1} \sum_{r=q+1}^{c} BCl(G_q, G_r). \tag{32}$$

On the other hand, the within-cluster difference of a cluster G_q can be defined by

$$WCl(G_q) = \frac{1}{|G_q|} \sum_{T_i \bar{\in} G_q} \sum_{w \in V_C} \sum_{v_d \in V_d} (\mu_C^{T_i}(w, v_d) - \tau_q(w, v_d))^2. \tag{33}$$

Then the total within-cluster difference is defined by

$$TWCl = \sum_{q=1}^{c} WCl(G_q) \tag{34}$$

Finally, the ratio of between-cluster difference to within-cluster difference is defined by

$$RBW = \frac{TBCl}{TWCl} \tag{35}$$

Considering the page restriction, in Table 3, we show only the results in the case of $p = 4$ when the previous AHC method seems to perform the best classification. In Table 3, KM, FCM(m), AHC, EFC(λ) and MFC(m) stand for K-means, fuzzy c-means with parameter m, the conventional AHC, entropy-based fuzzy c-means with parameter λ and modified fuzzy c-means with parameter m. Entry '*ave*±*dev*' of Table 3 shows the average value and the standard deviation of 10 executions.

As shown in Table 3, by selecting a suitable parameter, fuzzy c-means clustering including entropy-based and modified ones can produce a better clustering

Table 3. Results of Numerical Experiments ($p = 4 : \alpha = \beta = 0.5$)

	KM	FCM(1.1)	FCM(1.2)	FCM(1.3)	FCM(1.4)	FCM(1.5)
TCnc	39.3±5.83	45.3±5.59	43.9±4.01	41.3±3.41	24.0±0	0±0
TCnf	6.00±2.05	5.1±2.21	4.60±1.56	3.50±1.02	1±0	0±0
RBW	8.76±2.75	17.17±6.30	19.81±5.54	12.90±2.73	8.30±0	0±0
	AHC	EFC(0.9)	EFC(1)	EFC(1.1)	EFC(1.2)	EFC(1.3)
TCnc	30	35±7.35	37.7±5.31	40.6±4.27	41.4±3.32	41.6±3.41
TCnf	2	2.60±0.49	2.90±1.30	3.20±1.17	3.00±1.18	3.10±1.04
RBW	31.44	10.63±5.31	11.49±3.16	13.75±3.27	15.83±3.50	16.54±3.60
	MFC(1.2)	MFC(1.4)	MFC(1.6)	MFC(1.8)	MFC(2.0)	
TCnc	86.9±2.21	88.9±1.37	85.6±2.94	80.6±0.92	66.8±3.82	
TCnf	7.90±1.92	7.60±1.50	6.70±0.64	4.70±0.46	2.60±0.80	
RBW	72.25±13.59	77.10±9.50	77.45±9.49	75.11±10.69	68.85±6.87	

results in the concurrence of opinions between a cluster and its members as well as in the conflict of opinions between a cluster and its members than K-means clustering. However the conflict of opinions between a cluster and its members is not improved in the usual fuzzy c-means and in the entropy-based fuzzy c-means from that of the conventional AHC.

The improvement by the modified fuzzy c-means is remarkable in the concurrence of opinions as well as in the ratio of between-cluster difference to within-cluster difference. In the usual fuzzy c-means and the entropy-based fuzzy c-means, cluster opinions are less than the modified fuzzy c-means because τ_q tends to be small for profiles absent in some decision tables.

From the stability of the obtained results can be observed by counting how many times the same results are obtained in 10 executions with different initializations. The conventional AHC and FCM(1.4) were stable so that they produced same results in 10 executions. EFC(1.1) and MFC(1.8) were rather stable because they produced same results 8 times and 6 times. However, K-means method was not stable at all because it produced 10 different results.

From the quality of classification and the stability, fuzzy c-means approach, especially the modified one would be useful in clustering decision tables.

5 Concluding Remarks

We have proposed K-means clustering and fuzzy c-means clustering of decision tables. The results of the numerical experiments show that the modified fuzzy c-means clustering methods are advantageous in the concurrence of opinions between a cluster and its members as well as in the ratio of between-cluster difference to within-cluster difference.

The adopted treatment of missing profiles in this paper is to discard decision tables with no evaluation of the profile. The alternative approach, i.e., estimating the decision attribute values of missing profiles by the existing evaluations

in decision tables is one of our future topics. The proposed dissimilarity measure reflects neither the number of objects supporting opinions, the certainty on opinions nor the similarity between decision attribute values. The considerations of other dissimilarity measures and the examinations by real data are also future research topics. Moreover, we should tactfully use hierarchical and non-hierarchical clustering methods to a set of decision tables. As is known in the literature, the former is useful when the suitable number of clusters is unknown, while the latter is useful when it is known. The dissimilarity measures can be shared in both methods. The relations between those methods have not considerably investigated yet. The mutual developments of both methods would be helpful to discover the useful/convincing clustering approaches. Finally, the evaluations of the clustering results of decision tables are one of most difficult and important issues. They are useful in the selection of clustering results and should reflect the aim/application of the analysis. Some unique evaluations for clustering decision tables would be desired.

Acknowledgments. The first author acknowledges that this work was partially supported by the Grant-in-Aid for Scientific Research (B) No.19300074.

References

1. Enomoto, Y., Harada, T., Inoue, T., Mori, N.: Analysis of Choice for Audio Products Using Annexation Reduct System. The Bull. of Jpn. Soc. for the Sci. of Des. 49(5), 11–20 (2003) (in Japanese)
2. Inuiguchi, M.: A Multi-Agent Rough Set Model toward Group Decision Analysis. Kansei Eng. Int. 6(3), 33–40 (2006)
3. Inuiguchi, M.: Three Approaches to Rule Induction from Multiple Decision Tables. In: Proc. of the 12th Czech-Jpn Semin. on Data Anal. and Dec. Mak. under Uncertain., pp. 41–50 (2009)
4. Inuiguchi, M., Furudono, T.: Clustering Analysis of Individual Opinions Given by Decision Tables. In: Proc. of the 10th Czech-Jpn Semin. on Data Anal. and Dec. Mak. under Uncertain., pp. 41–53 (2007)
5. Inuiguchi, M., Miyajima, T.: Variable Precision Rough Set Approach to Multiple Decision Tables. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 304–313. Springer, Heidelberg (2005)
6. Itou, K., Enomoto, Y., Harada, T.: Influence of Annexation Order to Plural Annexation Condition Parts of Decision Rules. In: Proc. of 19th Fuzzy Syst. Symp., pp. 529–532 (2003) (in Japanese)
7. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications. Springer, Heidelberg (2008)
8. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Boston (1991)
9. Yamamoto, S., Inuiguchi, M.: A Variable Precision Dominance-based Rough Set Approach to Multiple Decision Tables. In: Inuiguchi, M., et al. (eds.) Proc. of MDAI 2009, pp. 153–164 (2009), CD-ROM
10. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. 46(1), 39–59 (1993)

Revisiting Natural Actor-Critics with Value Function Approximation

Matthieu Geist and Olivier Pietquin

IMS Research Group, Supélec, Metz, France

Abstract. Actor-critics architectures have become popular during the last decade in the field of reinforcement learning because of the introduction of the policy gradient with function approximation theorem. It allows combining rationally actor-critic architectures with value function approximation and therefore addressing large-scale problems. Recent researches led to the replacement of policy gradient by a natural policy gradient, improving the efficiency of the corresponding algorithms. However, a common drawback of these approaches is that they require the manipulation of the so-called advantage function which does not satisfy any Bellman equation. Consequently, derivation of actor-critic algorithms is not straightforward. In this paper, we re-derive theorems in a way that allows reasoning directly with the state-action value function (or Q-function) and thus relying on the Bellman equation again. Consequently, new forms of critics can easily be integrated in the actor-critic framework.

1 Introduction

Reinforcement learning (RL) is generally considered as the machine learning answer to the optimal control problem. In this paradigm, an agent learns to control optimally a dynamic system through interactions. At each time step i, the dynamic system is in a given state s_i and receives from the agent a command (or action) a_i. According to its own dynamics, the system transits to a new state s_{i+1}, and a reward r_i is given to the agent. The objective is to learn a control policy which maximizes the expected cumulative discounted reward.

Actor-critics approaches were among the first to be proposed for handling the RL problem [1]. In this setting, two structures are maintained, one for the actor (the control organ) and one for the critic (the value function which models the expected cumulative reward to be maximized). One advantage of such an approach is that it does not require knowledge about the system dynamics to learn an optimal policy. However, the introduction of the state-action value (or Q-) function [2] led to a focus of research community in pure critics methods, for which the control policy is derived from the Q-function and has no longer a specific representation. Actually, in contrast with value function, state-action value function allows deriving a greedy policy without knowing system dynamics, and function approximation (which is a way to handle large problems) is easier to combine with pure critics approaches. Pure critic algorithms therefore aim at learning this Q-function. However, actor-critics have numerous advantages over pure critics: a separate representation is maintained for the policy (in which we are

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 207–218, 2010.

ultimately interested), they somehow implicitly solve a problem known as dilemma between exploration and exploitation, they handle well large action spaces (which is not the case of pure critics, as some maxima over actions have always to be computed), and above all errors in the Q-function estimation can lead to bad derived policies.

A major march for actor-critics is the policy gradient with function approximation theorem [3,4]. This result allows combining actor-critics with value function approximation, which was a major lack of the field. Another important improvement is the natural policy gradient [5] which replaces the gradient ascent over policy parameters by a natural gradient ascent improving consequently the efficiency of resulting algorithms. These results share the drawback that they lead to work with the advantage function which does not satisfy a Bellman equation. Consequently, derivation of practical algorithms is not straightforward, as it requires estimating the advantage function which is unnatural in RL. In this paper, we reformulate (and re-prove) the theorems so as to work directly with the state-action value function (policy gradient with state-action value function approximation in Sec. 2 and natural policy gradient with approximation in Sec. 3). The position of this contribution compared to previous works is discussed in Sec. 4. This allows a very straightforward derivation of new actor-critic algorithms, some of them being proposed here (Sec. 5) and briefly experimented (Sec. 6). All results are given for the discounted cumulative reward case, however they can be easily extended to the average reward case.

2 Policy Gradient

The system to be controlled is standardly modeled as a Markov decision process (MDP) $\{S, A, P, R, \gamma\}$ [6]: S is the (finite) state space, A the (finite) action space, $P \in \mathcal{P}(S)^{S\times A}$ the set of transition probabilities, $R \in \mathbb{R}^{S\times A\times S}$ the deterministic reward function and γ the forgetting factor. Actions are selected according to a stochastic policy $\pi \in \mathcal{P}(A)^S$ which has to be optimized. The criteria to be maximized is the expected discounted cumulative reward starting in state s_0 and then following the policy π:

$$\rho(\pi) = E[\sum_{i=0}^{\infty} \gamma^i r_i | s_0, \pi] \tag{1}$$

State-action and state value functions are respectively defined as:

$$Q^\pi(s,a) = E[\sum_{i=0}^{\infty} \gamma^i r_i | s_0 = s, a_0 = a, \pi] \text{ and } V^\pi(s) = E_{a|s,\pi}[Q^\pi(s,a)] \tag{2}$$

Both state-action and state value functions satisfy a Bellman equation:

$$Q^\pi(s,a) = \sum_{s'\in S} p(s'|s,a)(r(s,a,s') + \gamma \sum_{a'\in A} \pi(a'|s')Q^\pi(s',a')) \tag{3}$$

$$V^\pi(s) = \sum_{s',a} p(s'|s,a)\pi(a|s)(r(s,a,s') + \gamma V^\pi(s')) \tag{4}$$

The advantage function is defined as their difference: $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. The advantage function does not satisfy any Bellman equation and it is obvious that $E_{a|s,\pi}[A^\pi(s,a)] = 0$. Let also d^π be the discounted weighting of states encountered starting at s_0 and then following π: $d^\pi(s) = \sum_{i=0}^{\infty} \gamma^i p(s_i = s|s_0, \pi)$.

Let the policy π_ω be parameterized by a parameter vector ω (and differentiable with respect to its parameters). We also assume that it is never zero ($\forall s, a, \omega$, $\pi_\omega(a|s) > 0$). The policy gradient approach consists in correcting parameters by following a gradient ascent:

$$\omega_i = \omega_{i-1} + \alpha_i \nabla_\omega \rho(\pi_{\omega_{i-1}}) \tag{5}$$

An important result is the policy gradient theorem (see [3] for a proof).

Theorem 1 (Policy gradient). *The gradient of the performance metric respectively to the policy parameters is:*

$$\nabla_\omega \rho(\pi_\omega) = \sum_{s \in S} d^\pi(s) \sum_{a \in A} Q^{\pi_\omega}(s,a) \nabla_\omega \pi_\omega(a|s) \tag{6}$$

An interesting thing would be to replace the true state-action value function $Q^{\pi_\omega}(s,a)$ by some approximation $\hat{Q}(s,a)$. Actually, it is possible thanks to the policy gradient with function approximation theorem [3,4], if the approximation is "good enough" and if the representations for the actor and the critic are "compatible". The proposed approach differs from previous ones on this last point.

Definition 1 (Semi-compatible approximation). *Let the approximation of the state-action value function be parameterized by two parameter vectors θ and ξ such that:*

$$\hat{Q}_{\theta,\xi}(s,a) = f_\theta(s,a) + g_\xi(s) \tag{7}$$

This approximation is said to be semi-compatible if its state-action part f_θ is compatible with the policy parameterization in the sense that:

$$\nabla_\omega \ln \pi_\omega(a|s) = \nabla_\theta f_\theta(s,a) \tag{8}$$

This definition differs from previous works in the sense that it adds a state-dependent parameterization g_ξ to the approximation, whereas in previous approaches the Q-function is approximated by only f_θ. However, $E_{a|s,\pi_\omega}[f_\theta(s,a)] = E_{a|s,\pi_\omega}[\theta^T \nabla_\omega \ln \pi_\omega(a|s)] = \theta^T \nabla_\omega E_{a|s,\pi_\omega}[1] = 0$, thus f_θ is an approximation of the advantage function rather than of the state-action value function. On the other hand, $E_{a|s,\pi_\omega}[\hat{Q}_{\theta,\xi}(s,a)] = g_\xi(s)$, thus g_ξ is an approximation of the value function. We propose to rederive all classic results with this new parameterization. The interest of this approach is that it simplifies the design of the critic, as it implies to work directly with the state-action value function.

Theorem 2 (Policy gradient with state-action value function approximation). *Let $\hat{Q}_{\theta,\xi}$ be semi-compatible as defined before. Moreover, assume that it is a good approximation in the sense that it is a local optimum of the square error between the true state-action value function Q^{π_ω} and its approximation*[1] *(in other words, the distance*

[1] As $\sum_{s \in S} d^\pi(s) = \frac{1}{1-\gamma}$, d^π is not really a distribution and the notation $E_{s|d^\pi}[h(s)] = \sum_{s \in S} d^\pi(s) h(s)$ is slightly abusive.

between the true state-action value function and its estimate should be small):

$$\begin{aligned} &\nabla_{\theta,\xi} E_{s,a|d^{\pi_\omega},\pi_\omega}[(Q^{\pi_\omega}(s,a) - \hat{Q}_{\theta,\xi}(s,a))^2] = 0 \\ \Leftrightarrow\ &E_{s,a|d^{\pi_\omega},\pi_\omega}[(Q^{\pi_\omega}(s,a) - \hat{Q}_{\theta,\xi}(s,a))\nabla_{\theta,\xi}\hat{Q}_{\theta,\xi}(s,a)] = 0 \end{aligned} \tag{9}$$

Then the policy gradient satisfies:

$$\nabla_\omega \rho(\pi_\omega) = \sum_{s\in S} d^\pi(s) \sum_{a\in A} \hat{Q}_{\theta,\xi}(s,a)\nabla_\omega \pi_\omega(a|s) \tag{10}$$

Proof. The gradient of $\hat{Q}_{\theta,\xi}$ is: $\nabla_{\theta,\xi}\hat{Q}_{\theta,\xi}(s,a) = [\nabla_\theta^T f_\theta(s,a), \nabla_\xi^T g_\xi(s)]^T$. Let also $\Delta Q(s,a)$ be defined as: $\Delta Q(s,a) = Q^{\pi_\omega}(s,a) - \hat{Q}_{\theta,\xi}(s,a)$. Part of condition (9) corresponding to parameters θ can thus be extracted and expanded thanks to the semi-compatibility condition (8):

$$\begin{aligned} 0 &= E_{s,a|d^{\pi_\omega},\pi_\omega}[\Delta Q(s,a)\nabla_\theta f_\theta(s,a)] \\ &= E_{s,a|d^{\pi_\omega},\pi_\omega}[\Delta Q(s,a)\nabla_\omega \ln \pi_\omega(a|s)] \\ &= \sum_{s\in S} d^{\pi_\omega}(s) \sum_{a\in A} \pi_\omega(a|s)\Delta Q(s,a)\nabla_\omega \ln \pi_\omega(a|s) \end{aligned}$$

However $\pi_\omega(a|s)\nabla_\omega \ln \pi_\omega(a|s) = \nabla_\omega \pi_\omega(a|s)$ thus:

$$\sum_{s\in S} d^{\pi_\omega}(s) \sum_{a\in A} \Delta Q(s,a)\nabla_\omega \pi_\omega(a|s) = 0 \tag{11}$$

Substracting Eq. (11) to Eq. (6) gives the result:

$$\begin{aligned} \nabla_\omega \rho(\omega) &= \sum_{s\in S} d^{\pi_\omega}(s) \sum_{a\in A} Q^{\pi_\omega}(s,a)\nabla_\omega \pi_\omega(a|s) - 0 \\ &= E_{s|d^{\pi_\omega}}[\sum_{a\in A} (Q^{\pi_\omega}(s,a) - \Delta Q(s,a))\nabla_\omega \pi_\omega(a|s)] \\ &= \sum_{s\in S} d^{\pi_\omega}(s) \sum_{a\in A} \hat{Q}_{\theta,\xi}(s,a)\nabla_\omega \pi_\omega(a|s) \end{aligned} \tag{12}$$

□

Notice that this results still holds by replacing $\hat{Q}_{\theta,\xi}(s,a)$ by $f_\theta(s,a)+b(s)$ where $b(s)$ is any baseline only depending on states, see [3]. Moreover, the minimum variance baseline for the state-action value function estimator is the value function $V^{\pi_\omega}(s)$ itself [7]. Recall that $g_\xi(s)$ is in fact an estimate of this value function, so using $\hat{Q}_{\theta,\xi}$ lets envision a low variance estimate.

Thanks to this result, new actor-critics algorithms can be naturally derived. The actor is the policy π_ω parameterized by ω, corrected by a gradient ascent using for example a sampled version of (10), and the critic is the approximated state-action value function $\hat{Q}_{\theta,\xi}$ parameterized by $[\theta^T, \xi^T]$ satisfying the semi-compatibility condition (8) and for which parameters are learnt such that condition (9) is satisfied.

However, another important progress for actor-critics is to correct policy parameters according to a natural gradient ascent, and we examine this point before proposing some practical algorithms.

3 Natural Policy Gradient

The idea of natural policy gradient is to correct the policy representation according to a natural gradient ascent rather than a gradient ascent. The natural gradient $\tilde{\nabla}$ is the gradient pre-multiplied by the inverse of the Fisher information matrix [8]:

$$\tilde{\nabla}\rho(\pi_\omega) = G^{-1}(\omega)\nabla\rho(\pi_\omega) \tag{13}$$

with the Fisher information matrix being equal to (see [5]):

$$G(\omega) = E_{s,a|d^{\pi_\omega},\pi_\omega}[\nabla_\omega \ln \pi_\omega(a|s)\nabla_\omega^T \ln \pi_\omega(a|s)] \tag{14}$$

The natural policy gradient was first introduced in [9] from a pure actor perspective. It was then used in [5] from an actor-critic perspective. They show an important result: the natural gradient is actually the advantage function parameter vector under the compatible approximation assumption. We show here that this result still holds for the proposed extended parameterization $\hat{Q}_{\theta,\xi}$.

Theorem 3 (Natural policy gradient with state-action value function approximation). *Let $\hat{Q}_{\theta,\xi}$ be semi-compatible as defined before and satisfying condition (9). Then the natural policy gradient satisfies:*

$$\tilde{\nabla}_\omega\rho(\pi_\omega) = \theta \tag{15}$$

Proof. Under these assumptions, theorem 2 applies:

$$\begin{aligned}\nabla_\omega\rho(\omega) &= \sum_{s\in S} d^{\pi_\omega}(s)\sum_{a\in A}\hat{Q}_{\theta,\xi}(s,a)\nabla_\omega\pi_\omega(a|s) \\ &= \sum_{s\in S} d^{\pi_\omega}(s)\sum_{a\in A}(f_\theta(s,a)+g_\xi(s))\nabla_\omega\pi_\omega(a|s)\end{aligned} \tag{16}$$

As the term g_ξ does not depends on actions, it disappears from the above equation: $\sum_a g_\xi(s)\nabla_\omega\pi_\omega(a|s) = g_\xi(s)\nabla_\omega\sum_a \pi_\omega(a|s) = g_\xi(s)\nabla_\omega 1 = 0$. As $\nabla_\omega\pi_\omega(a|s) = \pi_\omega(a|s)\nabla_\omega \ln \pi_\omega(a|s)$ and as $f_\theta(s,a) = \theta^T\nabla_\omega \ln \pi_\omega(a|s)$ (semi-compatibility condition), Eq. (16) leads to:

$$\begin{aligned}\nabla_\omega\rho(\omega) &= \sum_{s\in S} d^{\pi_\omega}(s)\sum_{a\in A} f_\theta(s,a)\nabla_\omega\pi_\omega(a|s) \\ &= E_{s,a|d^{\pi_\omega},\pi_\omega}[\theta^T\nabla_\omega \ln \pi_\omega(a|s)\nabla_\omega \ln \pi_\omega(a|s)] \\ &= E_{s,a|d^{\pi_\omega},\pi_\omega}[\nabla_\omega \ln \pi_\omega(a|s)\nabla_\omega^T \ln \pi_\omega(a|s)]\theta\end{aligned} \tag{17}$$

Recall the definition (14) of $G(\omega)$, this leads to: $\nabla_\omega\rho(\omega) = G(\omega)\theta$. This last equation and the natural gradient definition (13) lead directly to the result:

$$\tilde{\nabla}\rho(\omega) = G^{-1}(\omega)\nabla\rho(\omega) = G^{-1}(\omega)G(\omega)\theta = \theta \tag{18}$$

□

Thanks to this result, other actor-critic algorithms can be derived. The principle is the same as previously, but the gradient ascent is replaced by a natural gradient ascent, which is actually straightforward thanks to the above result.

4 Position to Previous Works

In previous works [3,5,7], theorems are derived using a parameterization f_θ of the advantage function. In order to obtain practically a critic, the advantage function (which does not satisfy a Bellman equation) has to be estimated. This is done either by adding a value component to the advantage function [5] or by using a TD error (linked to a value estimate) as the target for the advantage function. In this paper, we consider directly a parameterization of the Q-function and rederive theorems consequently. From a technical point of view, the new proofs heavily rely on the fact that the policy gradient is invariant to a state-dependent bias, which is a long known result [3]. Consequently, we do not really propose new theoretical insights on actor-critic architectures. However, from a practical point of view, our approach allows working directly with the state-action value function, which makes easier critics derivations. Moreover, notice that resulting algorithms are different from previously published ones, as shown in the next section.

5 Deriving New Actor-Critic Algorithms

Given these results, deriving practical actor-critic algorithms is quite direct. The actor is updated according to the natural-gradient, and the only remaining choice is the critic learner. We propose three critics here, the first one being based on TD with function approximation [6] and on a two-timescale approach [7], the two other ones being based on a Kalman-based Temporal Differences framework [10].

5.1 TD-NAC

The first algorithm, which we call TD-NAC (TD-based Natural Actor-Critic), is based on the classical TD with function approximation. A semi-compatible parameterization $\hat{Q}_{\theta,\xi}$ is adopted, and the critic is updated as follows:

$$\begin{pmatrix}\theta_i \\ \xi_i\end{pmatrix} = \begin{pmatrix}\theta_{i-1} \\ \xi_{i-1}\end{pmatrix} + \alpha_i \delta_i \nabla_{\theta,\xi} \hat{Q}_{\theta_{i-1},\xi_{i-1}}(s_i, a_i) \tag{19}$$

where α_i is the learning rate and δ_i the temporal difference error:

$$\delta_i = r_i + \gamma \hat{Q}_{\theta_{i-1},\xi_{i-1}}(s_{i+1}, a_{i+1}) - \hat{Q}_{\theta_{i-1},\xi_{i-1}}(s_i, a_i) \tag{20}$$

Deriving the critic update rule is thus a very direct application of the TD algorithm, much more direct than starting from the advantage function. A remaining problem is to ensure condition (9). We follow [7] and use two different timescales for the actor and the critic. The actor is updated using another learning rate β_i such that:

$$\sum_i \alpha_i = \sum_i \beta_i = \infty, \ \sum_i \alpha_i^2 = \sum_i \beta_i^2 < \infty, \ \lim_{i\to\infty} \frac{\beta_i}{\alpha_i} = 0 \tag{21}$$

The idea behind this is that in order to ensure condition (9), the actor should remain stationary from the critic point of view. Conditions (21) ensure that the critic converges

faster. The TD-NAC algorithm can be summarized as follows, the temporal difference error δ_i being defined in Eq. (20):

$$\theta_i = \theta_{i-1} + \alpha_i \delta_i \nabla_\omega \ln \pi_{\omega_{i-1}}(s_i|a_i) \tag{22}$$
$$\xi_i = \xi_{i-1} + \alpha_i \delta_i \nabla_\xi g_{\xi_{i-1}(s_i)} \tag{23}$$
$$\omega_i = \omega_{i-1} + \beta_i \theta_i \tag{24}$$

This algorithm is very close[2] to algorithm 3 in [7], which was actually first proposed in [11][3] (who call it NTD for Natural policy gradient using the Temporal Differences). Their principle is to estimate the value function (ξ_i parameters) and to use the associated temporal difference error as a target for the advantage function, which is actually a form of bootstrapping. This is less direct than the proposed approach which considers directly the Q-function and is therefore less dependent to the learning algorithm used to estimate it. The critic update for these algorithms is as follows:

$$\delta'_i = r_i + g_{\xi_{i-1}}(s_{i+1}) - g_{\xi_{i-1}}(s_i) \tag{25}$$
$$\theta_i = \theta_{i-1} + \alpha_i \nabla_\theta (f_{\theta_{i-1}}(s_i, a_i))(\delta'_i - f_{\theta_{i-1}}(s_i, a_i)) \tag{26}$$
$$\xi_i = \xi_{i-1} + \alpha_i \nabla_\xi (g_{\xi_{i-1}}(s_i))\delta'_i \tag{27}$$

Recall that $g_{\xi_{i-1}}$ (resp. $f_{\theta_{i-1}}$) is an approximation of the value (resp. advantage) function. Therefore, the difference between TD-NAC and NTD is that $E_{a_{i+1}|s_{i+1}}[\delta_i]$ is used instead of δ_i for the θ update, and that $E_{a_i|s_i}[[E_{a_{i+1}|s_{i+1}}[\delta_i]]$ is used instead of δ_i for the ξ update. Roughly speaking, the value function estimate is sometimes used instead of state-action value function estimate. These slight variations about the TD error are not new in reinforcement learning, see for example [12] and references therein.

5.2 KNAC

TD-NAC is based on a first-order critic. Using a second-order critic should speed up learning, as such algorithms are more sample-efficient. Actually, [5] introduced a natural actor-critic based on the Least-Squares Temporal Differences (LSTD) algorithm of [13]. However, in order to satisfy condition (9), this actor-critic algorithm is usually considered in a batch setting. Contrary to TD-NAC for which policy is improved at each time-step, the policy is maintained, its state-action value function is evaluated using obtained trajectories, and then the policy is improved. The advantage of using a second-order algorithm is thus somehow lost, as the policy cannot be improved after each interaction.

Kalman Temporal Differences (KTD) is another second-order algorithm, introduced in [10]. It has some interesting aspects, such as nonlinear parameterization handling. However, the feature we are interested in is its ability to handle non-stationarities. In an (online) actor-critic, the policy is updated after each interaction and is therefore not stationary. Consequently, the associated state-action value function is non-stationary

[2] In [7] algorithms are derived in the average reward case. However, extension to the discounted cumulative reward case is quite direct, and what we say is based on this extension.

[3] This algorithm is really derived in the discounted reward case.

too. To handle this problem (linked to condition (9)), the two-timescale approach is used in TD-NAC (the actor is stationary from the critic point of view). We argue that using a critic which tracks the state-action value function rather than converging to it is another manner to handle this problem, and thus to satisfy condition (9). Actually, [14] show theoretically that this condition is at least satisfied for a deterministic MDP and a stationary policy, and they show empirically that it still holds for non-stationary policies. This idea of using an adaptive critic also appears in [15], where a recursive form of LSTD integrating a forgetting factor (less general than the evolution model of KTD to be presented) is considered[4].

A semi-compatible parameterization $\hat{Q}_{\theta,\xi}$ is adopted. KTD algorithm is derived from a so-called *state-space* formulation[5], which in our case is given by:

$$\begin{cases} \begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} = \begin{pmatrix} \theta_{i-1} \\ \xi_{i-1} \end{pmatrix} + v_i \\ r_i = \hat{Q}_{\theta_i,\xi_i}(s_i, a_i) - \gamma \hat{Q}_{\theta_i,\xi_i}(s_{i+1}, a_{i+1}) + n_i \end{cases} \tag{28}$$

The first equation is the evolution equation, it specifies that parameters (which are modeled as random variables) evolve according to a random walk driven by the evolution noise v_i (to be chosen by the practitioner). It allows handling non-stationarity and avoiding local minima. The second equation is the observation equation which links the reward to the estimated state-action value function through a sampled Bellman equation. The observation noise n_i (also to be chosen) is an inductive bias which arises from the fact that the true state-action value function does not necessarily exists in the hypothesis space spanned by parameters.

The critic practical update is obtained directly from state-space model (28) and using the KTD-SARSA algorithm described in [10]. It is not difficult, but it takes room, so it is not fully described here. It should be sufficient to know that the critic is updated according to:

$$\begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} = \begin{pmatrix} \theta_{i-1} \\ \xi_{i-1} \end{pmatrix} + K_i(r_i - \hat{r}_i) \tag{29}$$

where K_i is the Kalman gain, which computation is detailed in the aforementioned paper, and where $\hat{r}_i$ is the prediction of the reward according to past estimates of the parameters and using the observation equation. Actually, $r_i - \hat{r}_i$ is a temporal difference error which takes into account the statistical nature of parameters in this model. The actor is updated according to the natural gradient ascent (24), θ_i being estimated by KTD. Notice that here there is only one β_i learning rate (there is no learning rate for KTD). We call the resulting natural actor-critic algorithm KNAC (Kalman-based Natural Actor-Critic), which can be summarized:

$$\begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} = \begin{pmatrix} \theta_{i-1} \\ \xi_{i-1} \end{pmatrix} + K_i(r_i - \hat{r}_i) \tag{30}$$

$$\omega_i = \omega_{i-1} + \beta_i \theta_i \tag{31}$$

[4] However, quite surprisingly, they also consider eligibility traces which induce a memory effect and therefore harm the non-stationary handling ability.

[5] The name state-space comes from the Kalman filtering literature and should no be confused with the state space of the MDP.

Recall that the main difference between NTD and TD-NAC is the replacement of the state-action value function by the value function in the temporal difference error. This can be easily adapted to KNAC: the term $\hat{Q}_{\xi_i}(s_{i+1}, a_{i+1})$ can be replaced by $g_{\xi_i}(s_{i+1})$ in the observation equation of state-space model (28):

$$\begin{cases} \begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} = \begin{pmatrix} \theta_{i-1} \\ \xi_{i-1} \end{pmatrix} + v_i \\ r_i = f_{\theta_i}(s_i, a_i) + g_{\xi_i}(s_i) - \gamma g_{\xi_i}(s_{i+1}) + n_i \end{cases} \tag{32}$$

We call this alternative algorithm aKNAC for averaged KNAC.

6 Experimental Results

In this section, we compare the three proposed new actor-critics to the NTD algorithm of [11] (which we recall to be equivalent to what would have been algorithm 3 of [7] in the discounted cumulative reward case). The benchmark on which these algorithms are compared is the inverted pendulum as described for example in [16].

This task requires balancing a pendulum of unknown length and mass at the upright position by applying forces to the cart it is attached to. Three actions are allowed: left force (-1), right force (+1), or no force (0). The associated state space consists in vertical angle φ and angular velocity $\dot{\varphi}$ of the pendulum. Deterministic transitions are computed according to physical dynamics of the system, and depends on current action a: $\ddot{\varphi} = \frac{g\sin(\varphi) - \beta ml\dot{\varphi}^2\sin(2\varphi)/2 - 50\beta\cos(\varphi)a}{4l/3 - \beta ml\cos^2(\varphi)}$ where g is the gravity constant, m and l the mass and the length of the pendulum, M the mass of the cart, and $\beta = \frac{1}{m+M}$. The reward is the cosine of the angular position, that is $r_i = \cos(\varphi_i)$, and the episode ends when $|\varphi_i| \geq \frac{\pi}{2}$. The discount factor γ is set to 0.95.

The policy is parameterized according to a Gibbs distribution. Let p be the size of the ω parameter vector (and thus of θ) and q the size of the ξ parameter vector. Let $\phi(s, a) = (\phi_i(s, a))_{1 \leq i \leq p}$ be a linear feature vector. The parameterized policy is given by:

$$\pi_\omega(a|s) = \frac{\exp\left(\phi(s, a)^T \omega\right)}{\sum_{b \in A} \exp\left(\phi(s, b)^T \omega\right)} \tag{33}$$

The semi-compatibility condition is therefore:

$$\nabla_\theta f_\theta(s, a) = \nabla_\omega \ln(\pi_\omega(a|s)) = \phi(s, a) - \sum_{b \in A} \pi_\omega(b|s)\phi(s, b) \tag{34}$$

Consequently, the f_θ function is given by:

$$f_\theta(s, a) = (\phi(s, a) - \sum_{b \in A} \pi_\omega(b|s)\phi(s, b))^T \theta \tag{35}$$

The parameterization is composed of a constant term and a set of 9 equispaced Gaussian kernels (centered in $\{-\frac{\pi}{4}, 0, \frac{\pi}{4}\} \times \{-1, 0, 1\}$ and with a standard deviation of 1) for each action. Thus there is a set of $p = 30$ basis functions. A parameterization has also

to be chosen for the g_ξ part of the estimated state-action value function. Let $\psi(s) = (\psi_i(s))_{1 \leq i \leq q}$ be a feature vector, we choose a linear parameterization:

$$g_\xi(s) = \psi(s)^T \xi \tag{36}$$

The parameterization is also composed of a constant term and a set of 9 equispaced Gaussian kernels (centered in $\{-\frac{\pi}{4}, 0, \frac{\pi}{4}\} \times \{-1, 0, 1\}$ and with a standard deviation of 1). Thus there is a set of $q = 10$ basis functions.

Some parameters have to be chosen for all algorithms. The ones we provide here allow obtaining good results. They are probably not optimal (better results could certainly have been obtained by testing more systematically all parameters), however orders of magnitude are valid. For all algorithms, the initial parameter vector $[\theta_0^T, \xi_0^T]$ is set to zero. For NTD and TD-NAC the critic learning rate is set to $\alpha_i = \alpha_0 \frac{\alpha_c}{\alpha_c + i^{\frac{2}{3}}}$ with $\alpha_0 = 10^{-2}$ and $\alpha_c = 10^4$. For all algorithms the actor learning rate is set to $\beta_i = \beta_0 \frac{\beta_c}{\beta_c + i}$ with $\beta_0 = 10^{-3}$ and $\beta_c = 10^4$. These learning rates satisfy the two-timescale condition (21) for NTD and TD-NAC. KNAC and aKNAC are based on KTD. This critic is a second-order algorithm for which parameters are modelled as random variables. A prior variance over theses parameters P_0 has to be chosen. Here we set $P_0 = I$ with I the identity matrix. Evolution and observation noises have also to be chosen. They are centered by assumption, and KTD only needs their second-order moments. In this experiment, the variance of the observation noise is set to $P_{n_i} = 10^{-1}$. An adaptive evolution noise is chosen, and its variance is set to $P_{v_i} = \eta P_{i-1}$ where $\eta = 10^{-5}$ is a forgetting factor and P_{i-1} is the estimate of parameters variance at time $i-1$ which is computed in the KTD algorithm. See [14] for a discussion on the choice of these parameters.

Algorithms are compared on their ability to learn the optimal policy. At the beginning of each episode the pendulum is initialized in a position close to the equilibrium $(\varphi, \dot{\varphi}) = (0, 0)$. The performance is measured as the number of steps the pendulum is maintained is the admissible zone (otherwise speaking, the length of the episode) in function of the number of episodes. A maximum of 3000 timesteps is allowed (the optimal policy would lead to an infinite episode). Consequently, the higher the curve is, the better the control is (and the better the corresponding algorithm too). Results presented on Fig. 1 are averaged over 100 independent trials.

Results of NTD and TD-NAC are not significantly different. Both algorithms learn the optimal policy in about 1600 episodes. However TD-NAC critic is simpler to derive (or at least more direct). KNAC learns the optimal policy faster, in about 600 episodes. This was to be expected: NTD and TD-NAC are based on a first-order critic, which is less sample-efficient than KTD which is a second-order critic. The aKNAC algorithm is even more efficient than KNAC, it learns the optimal policy in about only 200 episodes. We explain this by the fact that the aKNAC critic takes into account the expectation over action (by replacing $\hat{Q}_{\theta,\xi}(s', a')$ by $g_\xi(s') = E_{a'|s'}[\hat{Q}_{\theta,\xi}(s', a')]$ in the observation equation). This provides a better state-action value function estimation, see Bellman Eq. (3). These results show empirically the validity of the proposed alternative actor-critic theorems, as well as the interest of using a second-order critic handling non-stationarities to speed up learning in an online setting.

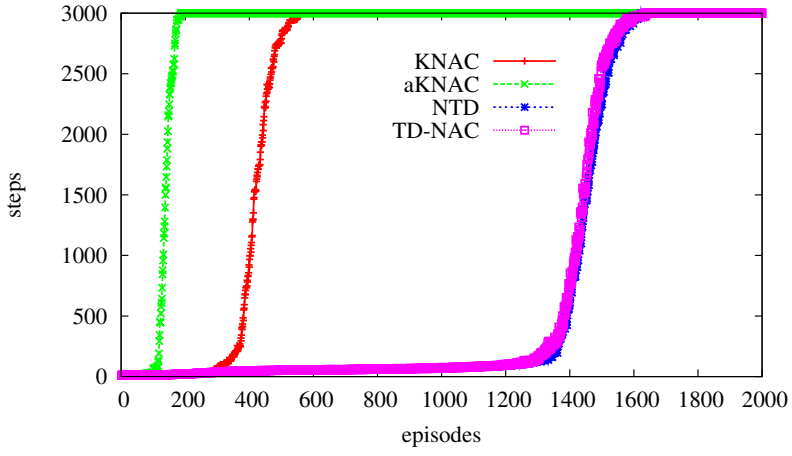


Fig. 1. Comparison of algorithms on the inverted pendulum task

7 Conclusion

In this paper we have presented alternative results concerning policy gradient with function approximation and natural policy gradient. They allow working directly with the state-action value function rather than with the advantage function. If these results do not change fundamentally the recent actor-critic theory, they allow deriving new critics in an easier way, as the state-action value function satisfies a Bellman equation, contrary to the advantage function. We have also illustrated the ease of critic design by introducing three new actor-critic algorithms, TD-NAC, KNAC and aKNAC. The first one is derived using TD with function approximation and a two-timescale approach, and it is close to the NTD algorithm [11] and to algorithm 3 of [7] as discussed before. KNAC and aKNAC are derived using the KTD framework of [10] which provides second-order function approximators able to handle non-stationarities. All these algorithms have been compared on the classic inverted pendulum task.

Actor-critic with function approximation is a very interesting paradigm. However, a number of questions remain open. An important problem is to design a critic which satisfies condition (9). So far, most of critics were designed in a batch setting [5] or using a two-timescale approach [7]. In this paper, we have proposed to use a critic which handles non-stationarities in order to satisfy this condition. Interesting perspectives would be to provide some guarantees for the proposed approach (that is choosing a provably appropriate evolution noise) and to discover new ways to ensure this condition. Another important point is the (semi-) compatibility condition and its implications. Actually, choosing a parameterization for the state-action value function or the policy is a difficult and problem-dependent choice itself. This condition renders this choice even more difficult. An interesting perspective would be to propose some feature selection framework (that is learning the structure of the representation in addition to its parameters) for such actor-critic algorithms.

Acknowledgements. The authors thank the European Community (FP7/2007-2013, grant agreement 216594, CLASSiC project : www.classic-project.org) and the Région Lorraine for financial support.

References

1. Barto, A.G., Sutton, R.S., Anderson, C.W.: Neuronlike adaptive elements that can solve difficult learning control problems, pp. 535–549 (1988)
2. Watkins, C.: Learning from Delayed Rewards. PhD thesis, Cambridge University, Cambridge, England (1989)
3. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation. In: Advances in Neural Information Processing Systems (NIPS 12), pp. 1057–1063 (2000)
4. Konda, V.R., Tsitsiklis, J.N.: Actor-Critic Algorithms. In: Advances in Neural Information Processing Systems, NIPS 12 (2000)
5. Peters, J., Vijayakumar, S., Schaal, S.: Reinforcement Learning for Humanoid Robotics. In: Third IEEE-RAS International Conference on Humanoid Robots, Humanoids 2003 (2003)
6. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. In: Adaptive Computation and Machine Learning, 3rd edn. The MIT Press, Cambridge (1998)
7. Bhatnagar, S., Sutton, R.S., Ghavamzadeh, M., Lee, M.: Incremental Natural Actor-Critic Algorithms. In: Advances in Neural Information Processing Systems (NIPS 21), Vancouver, Canada (2007)
8. Amari, S.I.: Natural gradient works efficiently in learning. Neural Computation 10, 251–276 (1998)
9. Kakade, S.: A Natural Policy Gradient. In: Advances in Neural Information Processing Systems (NIPS 14), pp. 1531–1538 (2002)
10. Geist, M., Pietquin, O., Fricout, G.: Kalman Temporal Differences: the deterministic case. In: Proceedings of the IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL 2009), Nashville, TN, USA (2009)
11. Morimura, T., Uchibe, E., Doya, K.: Utilizing the Natural Gradient in Temporal Difference Reinforcement Learning with Eligibility Traces. In: 2nd Internatinal Symposium on Information Geometry and its Applications, Tokyo, Japan, pp. 256–263 (2005)
12. Wiering, M., van Hasselt, H.: The QV Family Compared to Other Reinforcement Learning Algorithms. In: IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL 2009), Nashville, TN, USA (2009)
13. Bradtke, S.J., Barto, A.G.: Linear Least-Squares algorithms for temporal difference learning. Machine Learning 22, 33–57 (1996)
14. Geist, M., Pietquin, O., Fricout, G.: Tracking in reinforcement learning. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009. LNCS, vol. 5863, pp. 502–511. Springer, Heidelberg (2009)
15. Park, J., Kim, J., Kang, D.: An RLS-Based Natural Actor-Critic Algorithm for Locomotion of a Two-Linked Robot Arm. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 65–72. Springer, Heidelberg (2005)
16. Lagoudakis, M.G., Parr, R.: Least-Squares Policy Iteration. Journal of Machine Learning Research 4, 1107–1149 (2003)

A Cost-Continuity Model for Web Search

David F. Nettleton and Joan Codina

Pompeu Fabra University, Tanger, 122-140, 08018 Barcelona, Spain
david.nettleton@upf.edu, joan.codina@upf.edu

Abstract. In this paper we present and empirically evaluate a 'continuity-cost model' for Internet query sessions made by users. We study the relation of different 'cost factors' for a user query session, with the continuity of the user in that query session, and the order of the query in the query session. We define cost indicators from the available query log data, which are to be studied in relation to continuity and to the order/number of the query (1st, 2nd, 3rd, ..). One of our hypotheses is that cost related factors will reflect the step by step nature of the query session process. We use descriptive statistics together with rule induction to identify the most relevant factors and observable trends, and produce three classifier data models, one for each 'query number', using the 'continuity flag' as classifier label. Using the cost factors, we identify trends relating continuity/query number to user behavior, and we can use that information, for example, to make decisions about caching and query recommendation.

Keywords: Web-mining, web query-sessions, data analysis and modeling.

1 Introduction

Individual user behavior when searching for information in Internet may at first sight seem a chaotic activity, especially when we try to analyse high volume search engine logs looking for trends and patterns. In the literature there are many authors who have tried analysing web search data, deriving descriptive factors from the basic information, and a smaller number of authors who have proposed 'behavior models' based on empirical observations and user studies, such as Fox [1], Hassan[2], Baeza-Yates[3], Nettleton[4] and Ntoulas[5].

In [1], Fox carried out a user study of 146 users and 2560 query sessions, in which the users proportioned feedback about the grade of satisfaction of their query session. Bayesian modeling and 'gene analysis' techniques were used for modeling using factors such as clickthrough, the time spent on the search result page, and how the user exited from the session. In [2], Hassan conducted experiments to show that models using user behavior are more predictive of goal success than those using document relevance. Their data was derived from a commercial search engine query log, with a total of 2172 query sessions, and they defined "sequence models" using time distributions, related to 'cumulative gain', with a Markov model representation. Their findings proposed that for user behavior, sequence and time distributions are more accurate than static models or predictions based on document relevance.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 219–230, 2010.

In [3], the user's path was traced through web site links, relating the user behavior to the connectivity of each site visited, and different schemes are evaluated for modeling user behavior, including Markov Chains. Sugiyama [6] evaluated constructing user profiles from past browsing behavior of the user, which required identified users and one day's browsing data. Lee et al [7] developed an approach for the automatic detection of user 'goals' in web search, using a reduced set of 30 pre-selected queries from which ambiguous queries had been eliminated.

User behavior can be modeled in different ways. For example, the Markovian model [2,3] describes and represents transitions between frequent states in a query session, in the form of a finite state model. Silverstein[8] made the observation that Web users in general tend to formulate short queries, select few pages and an important proportion of them refine their initial query in order to retrieve more relevant documents.

Craswell [9] proposed a "position-bias" score to model the way users choose search engine results, which is similar to the 'SRC' factor we have defined (see Section 3). Teevan [10] has adopted a more classic IR approach, such as that based on statistics of re-finding and relevance feedback. An alternative approach is the "query clarity" feature proposed by Cronen-Townsend et al [11]. This model has a greater complexity, being based on the query language IR model which requires access to statistics of the content of each document. In [11], a method was presented for predicting query performance by computing the relative entropy between a query language model and the corresponding collection language model. The resulting 'clarity score' measures the coherence of the language usage in documents whose models are likely to generate the query. They propose that 'clarity scores' measure the ambiguity of a query with respect to a collection of documents.

'Continuity-cost model'. Our analogy in the current work to model user search behavior is to consider that the user approaches information search and selection as a series of succesive steps, which we propose has a broad analogy to the Nash's 'centipede game' [12,13]. The 'centipede game' represents a sequence of steps by a person, each with a given 'cost' and perceived 'benefit'. After each step of the game, the person can decide to continue or to quit, and it is assumed that this decision is influenced by the persons' perception of the cost incurred and benefit achieved. In the web search context, and with the data we have available (typical large volume web log), we clearly have 'cost' data (time spent, clicks made, …), but we do not have 'benefit' data, the latter being much more difficult to know or to quantify. In the present study, we define an approximation to 'benefit', in which we detect patterns where the user seems to be showing interest in specific documents (results).

The rest of the paper is organized as follows: in Section 1.1 we outline the motivations for the study; in Section 2 we describe the design of the cost-continuity model, its application to web search and how query sessions are represented; in Section 3 we describe the basic available data and derived factors; in Section 4 we describe the data analysis of user continuity for query sessions using descriptive statistics and create a classifier model using rule induction; finally, some conclusions are presented in Section 5.

1.1 Motivations for the Study

We will now comment two aspects which can be applied by the approach and which are the practical motivation of the work: **(i)** Optimization of the use of the Cache and **(ii)** User support and query recommendation.

Cache: If we can anticipate with a reasonable accuracy that the user will continue in the same query session, then we can use this information to make decisions about retaining the results of the previous query in the cache. The saving on selective caching with respect to caching everything "just in case" is very significant, in terms of computational cost and memory storage cost, for the millions of queries being constantly launched in large search engines. Therefore we can directly equate a 'continuity' flag to the flag which tells us to keep the results of a query in the cache. We can define two types of cost and saving: (i) the cost of keeping the results in cache and the saving obtained by not keeping them in cache - this is a memory use consideration; (ii) the cost of having to re-execute the query to generate all the results and the saving of not having to re-execute the query – this is a computational consideration. Other authors, such as Teevan [10], have adopted different approaches, such as that based on statistics of re-finding, or caching on a per-query basis.

User Support and Query Recommendation: If we can detect trends with a vision of the query session as a whole, then we can help the user by giving them 'meta information' to guide their query session, rather than hints simply on a per-query basis. This implies an understanding of the user intent which improves on the typical spelling corrections or 'did you mean' hints. The system would indicate how the user can improve his search strategy, with contextual suggestions. As with the case of Caching, a more simplistic approach is to offer query recommendations based only on the current query. We propose that the query session approach is more complex but provides a significantly more complete picture of what the user is doing as a sequence of events.

2 Design of the Cost-Continuity Model

In this Section we explain the approach and design of the 'Cost-continuity model'.

We propose that the steps in the query session have a continuity which depends on the cost incurred by the user, and the perceived information gain (distribution pattern) of the user. For example, one possible behavior (distribution) may be identified which indicates that the user is 'homing in on', or spending more time reading, specific information which is relevant to his/her query. The user may discontinue a search sequence for two main reasons: (i) the user has found the information he was looking for, or is at least satisfied with the information found; (ii) the user has not found the information he was looking for, and the cost incurred has exceeded that which the user was prepared to spend on the given search. We could also postulate a third option (iii) when the user doesn't know how to reformulate their query to make progress. In the present work we have considered this option as included in option (ii), given that the user may perceive a potential high cost for obtaining the information. Also, an interpretation of "think-time" has been factored into the cost formula (See Section 3).

2.1 Application of the 'Cost-Continuity Model' to Web Search

In the case of a query session, we assume that there is a payoff in terms of an increasing cumulative information gain which reaches a maximum value after N queries, and an increasing cumulative cost in terms of time and effort. In a query

session, the user makes one or more queries Q_i, each with its distribution b_i in terms of information gain and cost c_i in terms of user effort, the latter of which we can measure in terms of elapsed time, number of clicks, and other available data. In the case of a session in which the user progressively refines his query, we could say that, after each query and/or after each click on a result, the user will evaluate if it is worthwhile to spend more time searching and clicking, or if with the information he already has, he has satisfied his information need and will therefore leave the "game".

With reference to Fig. 1, each successive query (Q_1, Q_2 **…**) has an option to continue (C) or terminate (T). If the option to terminate is taken after making query Q_i, (vertical leg of the sequence), then the benefit will be b_i and the cost c_i. We recall that the benefit will intuitively represent the users interest (for which we have used a "distribution" factor, see Section 3) and the cost will represent the time and effort made by the user. We note that, in the two player centipede game [12,13], the values at each leg 'T' (Fig. 1), correspond to the payoff for player 1 and player 2, respectively. In the 'Continuity –cost model' with one player, the values correspond to the distribution and the cost, respectively. The user has the option to continue clicking on results for the current query, or can postulate a new query, or can quit the session.

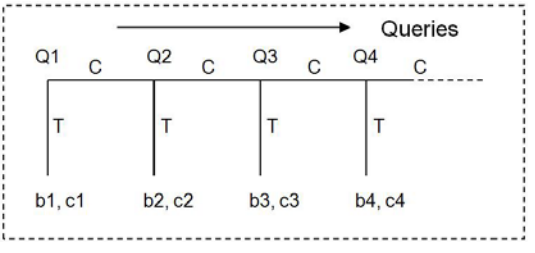


Fig. 1. 'Cost model' representation of a query session, where Q1, Q2 …, are queries, C indicates that the user continues, T indicates that he terminates, b1=benefit for Q1 and c1=cost for Q1

2.2 Representation of Query Sessions

One difficulty for studying real query sessions is the high proportion of queries which seem chaotic in nature. This is due to different reasons: non-expert users, users who have inefficient or ineffective search methods, poor style in formulating queries, lack of knowledge of what they are searching for, a high number of spelling and typographical errors etc.

If we limit our study by selecting query sessions which appear to follow a logical sequence of queries to find some specific information, we still have complex situation in which we must identify trends and anticipate what a user will do. If we are dealing with anonymous users in a high volume search engine query log, we have to wait at least until they (seem to have) completed a query and studied the returned documents, in order to make any assessment about what they will do next (for example, launch another query). A query session may consist of just one query, or two queries, or three or more, although the majority (a frequency of 95% was found for TodoCL [14]) do not consist of more than three queries. Also, for the first three queries, the ratio of

users who quit to those who make a new query (in the same query session) is 75% and 25%, respectively. We propose dividing the possible options of what a user may do into six categories. Thus, after posing a first query (Q_1) the user can quit, or continue. If the user continues he will pose a second query (Q_2), from which he can continue or quit, and if the user continues after the second query, he will pose a third query (Q_3) from which he can continue or quit. This subdivision enables us to study the data for each 'query level', and partitioned by those who quit and those who continue. In 'well formed' queries, we assume the user has kept following the same informational 'trail' and has not deviated to another theme or changed the basic objective of what s/he is looking for.

3 Basic Available Data and Derived Factors

Basic Available Data Attributes: For each query and results browsing in a query session, we have the following data: (i) Number of clicks made by the user; (ii) hold time for each document (result) clicked on, from which we derived the average hold time and the total elapsed time for a query; (iii) the ranking of each of the documents clicked in the results list of the search engine, from which we can derived the average ranking and the sum of the rankings; (iv) number of terms in the query; (v) number of documents retrieved by the query. We also have the total number of documents indexed by the search engine, which of course is a constant for all the calculations, and is used in the derived factor C_1 (see below).

3.1 Derived Cost Factors

The following factors have been developed as an approximation to the effort spent for a given query.

Query Formulation Cost 1 (*C1*). Let D_T be the total number of documents indexed by the search engine, D_R the number of documents retrieved by the given query. Then we define the cost factor as:

$$C1 = \frac{\log\left(\frac{D_T}{D_R+1}\right)}{\log(D_T)} = 1 - \frac{\log(D_R+1)}{\log(D_T)} \tag{1}$$

which serves as a kind of IDF (Inverse Document Frequency) and gives us an idea of the 'goodness' of the terms chosen. This is because the smaller the ratio, the fewer results are returned and the more specific the query is. This formula gives an approximation for the time/effort spent by the user to think of which terms to use. We assume that the query terms are well formed (correct spelling and use of valid characters). Modern search engines usually filter user queries for spelling and correct usage.

As an alternative to the query formulation cost feature proposed based on an anonymous web log, a user study could be conducted to establish cost features. A second alternative would be to try to estimate the average time that a given user spends looking at search results, and then subtract it from the "hold time" to get some estimate of query reformulation time.

Query Formulation Cost 2 (NUMTERMS): The number of terms in the query, which is a data value derived directly from the search engine log.

Effort Spent Locating Documents in Results (SRC). We define SRC as the sum of the rankings of the documents clicked by the user. This gives an approximation for the effort the user spends locating the documents which s/he finds interesting.

$$SRC = \sum_{i \in Clicked\ Docs} ranking_i \tag{2}$$

That is, SRC is equal to the sum of the rankings of the clicked documents: if the user clicks documents ranked 1, 5 and 8, then the SRC will be equal to 1+5+8=14.

Our definition of SRC is similar to "position-bias" scoring models, such as that proposed by Craswell [9]. An alternative approach would be the "query clarity" feature proposed by Cronen-Townsend et al [11]. However, this model has a greater complexity, being based on the query language model which requires access to statistics of the content of each document.

Elapsed Time. We define ELAPSED_TIME as the sum of hold times for the documents selected for a given query. This is a data value derived directly from the search engine log.

Other Factors. We define the following factor separately from the cost factors, given that it is proposed as a sort of 'benefit' or 'interest measure'.

Browsing Time Distribution (SDISTRIB). This is a factor defined in terms of the distribution of the hold times. HOLDTIME *H* represents the number of seconds a user maintained open a given document selected from the results list. It is also implicitly related to NUMCLICKS, the number of clicks made by the user on results documents for the corresponding query. Thus, for one query of a query session, we make the following definition:

$$\text{Distribution} \sqrt{\frac{\sum (x-\bar{x})^2}{N}} \cong \sqrt{\sum HT^2} \tag{3}$$

This provides information about the "time distribution" of the query, one of the key factors included in the models defined in [2,3].

*Ratios. C*alculated for the second and third queries for the following variables: **Elapsed_Time, Av_HoldTime, NumTerms, C1, SRC, SDistrib, NumClicks, Docs_Retrieved, Av_Rank.** Thus, for example, **Ratio_Elapsed_Time** represents the elapsed time of the current query divided by the elapsed time of the previous query.

Accumulators. Calculated for the second and third queries of the following variables: **Elapsed_Time, Av_HoldTime, NumTerms, C1, SRC, SDistrib, NumClicks, Docs_Retrieved, Av_Rank.** Thus, for example, **Sum_Docs_Retrieved** represents the number of documents retrieved by the current query plus the number of documents retrieved by all previous queries, in the current query session.

Choice of Factors: We initially applied a Principal Components and Chi-Square analysis to the basic data variables, the candidate derived factors and the continuity flag for each query level. Variables were then selected based on the best correlation values. With respect to the definitions, the SRC, C1 and SDISTRIB factors were partially inspired from the literature: SRC is similar to the "position-bias" scoring

model of Craswell [9], C1 is a pseudo IDF value, see Teevan [10] and SDISTRIB is a similar idea to that presented in [2,3].

Example Data Values for a Query Session: With reference to Fig. 2, we see a real query session of three queries, taken from the query log data used for the study. We observe that the user progressively refines the query, by adding more specific query terms. Above each query we see the values for the number of clicks and the number of documents retrieved, and below each query we see the calculated values for the four cost factors and the distribution factor, as described in Section 3.

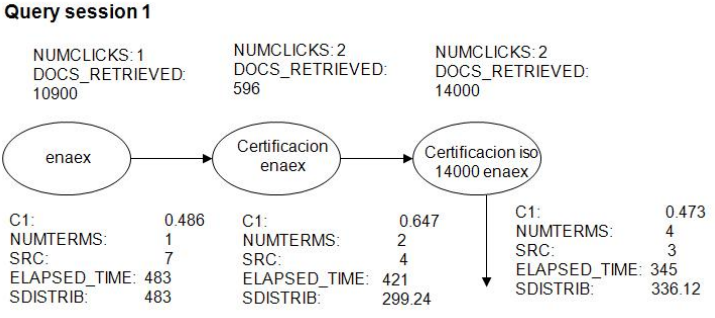


Fig. 2. Example of a State diagram for a Query Session

4 Data Analysis of User Continuity for Query Sessions

In this Section we show results of the data analysis and data modeling using the IM4Data (IBM Intelligent Miner for Data V6.1.1) Data Mining tool [15].

4.1 Preprocessing and Data Sampling

The original data has one line per click, with only the time stamp and the rank of the clicked document. Therefore we calculated the hold time of a document using a standard criteria in web search log analysis, which is to consider it as the time to the next click. We initially select 10,000 candidate clicks from a web query log of 252,000, whose hold time is not zero (unique clicks) nor excessively large (there are characteristics including very long hold times, in general greater than 900 seconds, which may indicate the user is inactive before the screen window which contains the document). In the literature studied on web search logs, a value of approx. 900 seconds is generally used as a limit for valid hold times [3,8]. It is important to note that if there is no click on a new document for the current query results, which occurs more often when there are few clicks per query, then the value of the last "hold time" is taken as time to next click (if there is a new query) or the average of the hold times of the current query (if there is more than one) or zero if there has been no document selections. This introduces an approximation, but we have observed empirically that in general it gives a good estimation of the true hold time, and statistically is much better than assigning hold time equal to zero.

4.2 Data Analysis - Descriptive Statistics

We have used the IM4Data [15] Bi-variate Statistics option using the Chi-Squared statistic. This allows us to rank the variables by the chi-square statistic relative to the partition variable (continuity) and visualize and compare the trends for each variable in one partition with the corresponding variable in the other partition. In Figs. 3 and 4, the resulting graphics for each attribute are in descending order (from left to right) of the Chi-Squared statistic, which gives the importance relative to the partition variable. Numeric attributes are represented by a histogram of their distribution. In all the histograms, the grey filled rectangles indicate the frequency for the whole dataset, and the other rectangles (which terminate above or below the grey rectangles, represent the frequency for the given partition.

Descriptive Statistics: Query 1. We will now make some observations about the most significant tendencies shown by variables in the dataset corresponding to the first query of a query session. With reference to Fig. 3, we observe that for users who choose to continue (right hand figure, partition=11) the attribute NUMCLICKS has a higher frequency for fewer clicks (distribution for the given partition shown by unfilled rectangles with grey borders). This means that users who continue have made, in general, fewer clicks on results.

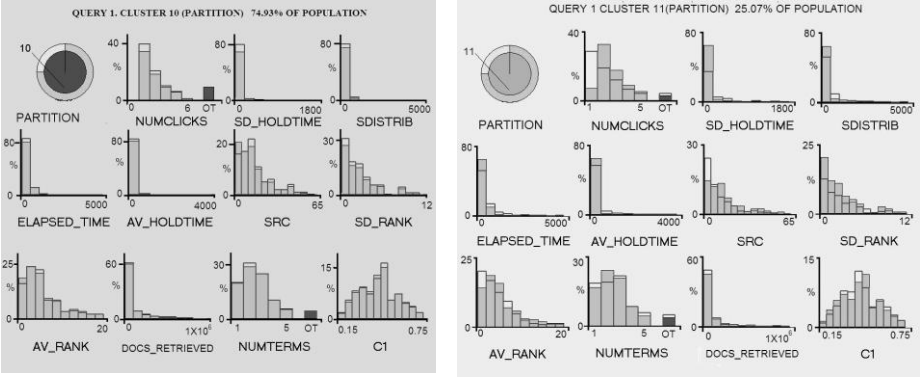


Fig. 3. Distributions of attributes of dataset 'Query 1', for users who quit (left) and users who continue (right); attributes ordered by Chi-Square

Also, if we observe the attribute SD_HOLDTIME, there is a lower frequency for small standard deviations, which means that hold times tend to be more different for users who go on to make a second query. This same tendency is shown by the derived factor SDISTRIB, which, we recall, is derived from the hold times of the documents. In the case of ELAPSED_TIME, we observe that users who continue spend slightly more time looking at the documents. This slight tendency is also shown by AV_HOLDTIME. In the case of SRC, a stronger tendency is visible, in which a higher frequency is shown for small values. This indicates that users who continue tend to choose results which are closer to the top of the ranking. Finally, SD_RANK tends to have lower frequencies for lower standard deviations of RANK, which means that the selected results tend to be more dispersed for users who go on to make a second query.

4.3 Classification Using Tree Induction

We have used the IM4Data tree induction algorithm to produce an induced tree of the input attributes with the classifier label as the continuity flag, for each query number. For the sake of brevity, we only show details of the decision tree for Query 1.

Induced Classification Tree: Query 1. With reference to Fig. 4, we see the pruned tree (to level 6) induced by IM4Data on the Query 1 dataset, including the details of the decision nodes and classification nodes. We observe that attributes 'NUMCLICKS' and 'SD_HOLDTIME' have been used in the upper part of the tree (NUMCLICKS < 1.5, SD_HOLDTIME < 19.92). Thus, they represent the most general and discriminatory factors to classify 'PARTITION', that is the users who continue (PARTITION=11) and users who do not continue (PARTITION=10). We note that lower down in the tree the attributes 'SDISTRIB', 'SRC' and 'AV_RANK' have been used, which implies they are used for more specific cases. The triangular nodes indicate that a sub-tree exists below the level of the pruning limit, and the corresponding classification data is given. An example rule derived from the decision tree of Fig. 4 is:

```
IF NUMCLICKS < 1.5   THEN 11 (continue): 100%, 51
```

which is interpreted as: if the number of clicks is less than 1.5 then there are 51 cases (with 100% accuracy) in which the user continues. A second example rule derived from the decision tree of Fig. 5 is:

```
IF NUMCLICKS BETWEEN 1.5 AND 2.5 AND HOLDTIME < 19.92
THEN 11 (continue): 100%, 26
```

Which is interpreted as: if the number of clicks is between 1.5 and 2.5 and the hold time is less than 19.92 then there are 26 cases (with 100% accuracy) in which the user continues. The relationship between NUMCLICKS, SD_HOLDTIME and AV_HOLDTIME could be as a consequence that these measures express a similar

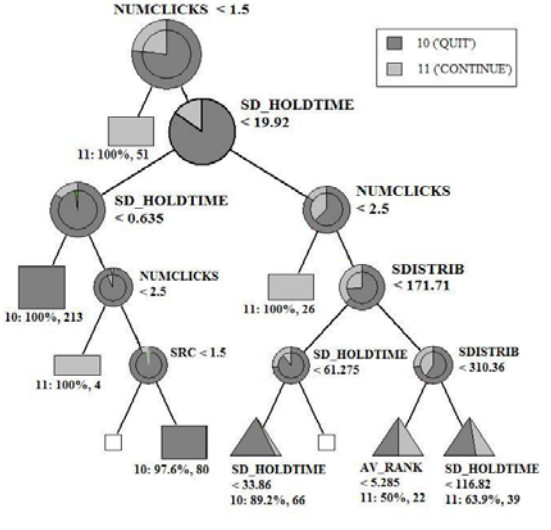


Fig. 4. Pruned Classification Tree: dataset 'Query 1'

kind of information. One option would be normalize and aggregate them to reduce the final number of measures, which may give a greater balance to the classification tree.

Classification Precision. With reference to Table 1, we present the test results (test folds) for the tree induction model built from the Query 1, Query 2 and Query 3 datasets. We observe that the model for Query 1 is the strongest (overall precision over 5 folds is 92.59%). The model for Query 2 has a precision of 78.57%, and the model for Query 3 has an overall precision of 76.78. For the Query 1 model, the low percentage of false positives and false negatives over the five folds indicates that we have a 'robust' model. Also the classification accuracy for the minority class ('continue') is high, at 81.3%. We observe that the results for Queries Q2 and Q3, indicate that it is progressively more difficult to model the data, for successive queries. With reference to Q2 and Q3, the precision for the 'continue' class drops to 65.6% for Q2 and 58.5% for Q3, with false negative rates of 32.4% and 41.9%, respectively.

Table 1. 'Query 1, 2 and 3 ': test precision for 5x2 fold cross validation

		Q1		Q2		Q3	
			MP*		MP		MP
fold1	stop†	97.5, 6.54	93.07	89.1, 12.8	82.27	81.9, 16.1	79.55
	continue††	94.9, 3.48		86.4, 11.7		60.0, 45.0	
fold2	stop	94.9, 3.48	93.44	86.4, 11.7	80.08	78.7, 10.8	71.25
	continue	88.8, 15.9		62.5, 31.5		58.3, 40.9	
fold3	stop	93.3, 6.22	90.44	85.3, 12.4	80.03	87.1, 15.8	71.25
	continue	82.1, 19.0		66.7, 27.6		54.0, 37.2	
fold4	stop	100.0, 7.81	93.68	77.8, 14.0	74.64	73.5, 16.8	79.26
	continue	75. 2, 0.0		66.3, 35.1		62.8, 43.0	
fold5	stop	96.2, 6.05	92.40	78.1, 11.8	76.09	75.6, 15.9	76.43
	continue	80.4, 13.0		70.3, 35.0		57.6, 44.2	
Geometric mean for folds	stop	96.3, 5.82	92.59	83.2, 11.8	78.57	79.2, 14.9	76.78
	continue	81.3, 3.18		65.6, 32.4		58.5, 41.9	

*MP=Model Precision, †stop{%Rate: True Positive, False Positive}, ††continue{%Rate: True Negative, False Negative}.

Note that, in Table 1, with reference to True/False Positive/Negative, a Positive result corresponds to 'stop' and a Negative result corresponds to 'continue'. Therefore, a True Positive corresponds to 'stop=true' and a True Negative corresponds to 'continue' ≡'not stop'='true'. The nature of query sessions indicates that it is progressively more difficult to identify trends in user behavior beyond the first query of the query session. Also, the majority of users (79%) are classifiable by the Query 1 model. We note that the best results for Query models 1 and 2 were obtained by using the inputs chosen by a Principal Components analysis. On the other hand, the best inputs for Query model 3 were found by using Chi-Square. Finally, we

note that, although we gave all inputs (basic, derived factors, summations and ratios) to Query models 2 and 3, the best results were given by the ratios and summations, and many of the derived cost factors were used preferentially to the basic data attributes. In the case of the Query 1 model, we used only the basic data and factors, given, of course, that the ratios and summations do not exist for the first query.

5 Conclusions

In this paper we have presented and evaluated a novel conceptual model for defining individual user behavior during query sessions for Internet search. We have defined some novel cost/distribution calculations which we have used, together with the basic query log data, as descriptive variables of the query sessions. The partition category is a flag which indicates if the user will make a new query ('continue') or finish the query session ('quit'). We have used descriptive statistical techniques and rule induction to look for trends in the data, using real query sessions taken from the 'TodoCL' query log. The work has enabled us to identify different trends for each query level (1st, 2nd and 3rd query), and define a strong model for the first Query (overall precision of 92.6% with low false positive and false negative rates), relating the continuity flag to the input attributes, which covers the majority (79%) of the queries sent to the search engine.

References

[1] Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM Trans. Inf. Syst. 23(2), 147–168 (2005)

[2] Hassan, A., Jones, R., Klinkner, K.L.: Beyond DCG: user behavior as a predictor of a successful search. In: Proc. 3rd ACM Int. Conf. WSDM 2010, New York, pp. 221–230 (2010)

[3] Baeza-Yates, R., Hurtado, C., Mendoza, M., Dupret, G.: Modeling user search behavior. In: Proc. 3rd Latin Am. Web Congress 2005, Buenos Aires, pp. 242–251 (October 2005)

[4] Nettleton, D.F., Baeza-Yates, R.: Web retrieval: Techniques for the aggregation and selection of queries and answers. Int. Journal of Intelligent Systems 23(12), 1223–1234 (2008)

[5] Ntoulas, A., Cho, J., Olston, C.: What's new on the web? The evolution of the web from a search engine perspective. In: Proc. 13th Int. WWW Conf., New York, US (May 2004)

[6] Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: Proc. 13th Int. WWW Conf. (May 2004)

[7] Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: Proc. 14th Int. World Wide Web Conference, Chiba, Japan (May 2005)

[8] Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a Very Large Web Search Engine Query Log. ACM SIGIR Forum 33(1), 6–12 (1999)

[9] Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An Experimental Comparison of Click Position-Bias Models. In: Proc. Int. Conf. on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, pp. 87–94 (2008)

[10] Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing Search via Automated Analysis of Interests and Activities. In: Proc. 28th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Salvador, Brazil, pp. 449–456 (2005)
[11] Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting Query Performance. In: Proc. 25th Int. ACM SIGIR Conf. on R+D in Information Retrieval, Finland, pp. 299–306 (2002)
[12] McKelvey, R.D., Palfrey, T.R.: An Experimental Study of the Centipede Game. Econometrica 60, 803–836 (1992)
[13] Nash, J.: Equilibrium points in n-person games. Proceedings of the National Academy of the USA 36(1), 48–49 (1950)
[14] TodoCL. Chilean Internet Search Engine (2007), `http://www.todocl.com`
[15] Im4Data, Using the Intelligent Miner for Data V8 Rel. 1. IBM Redbooks, SH12-6394-00 (2002)

An Enhanced Framework of Subjective Logic for Semantic Document Analysis

Sukanya Manna, B. Sumudu. U. Mendis, and Tom Gedeon

School of Computer Science, The Australian National University, Canberra, ACT 0200, Australia
{sukanya.manna,sumudu.mendis,tom.gedeon}@anu.edu.au

Abstract. Unlike propositional logic which works on truth or falsity of statements, human judgements are subjective in nature having certain degree of uncertainty. Two different people will analyse and interpret a document in two different ways based on their background and current focus. In this paper we present an enhanced framework of subjective logic for automated single document analysis where each sentence in the document represents a proposition, and 'opinions' are constructed about this proposition to focus the degree of uncertainty associated with it. The 'opinion' about a sentence determines the significance of that sentence in a document. The input arguments are built automatically from a document in the form of evidence; then they are analyzed based on subjective logic parameters. Two different approaches are described here. The first utilises "bag of words" concept. However, this approach tends to miss the underlying semantic meanings of the context, so we further enhanced it into the latter approach which incorporates semantic information of the context, by extending the basic definitions of subjective logic.

1 Introduction

Subjective logic [1] is a logic which operates on subjective beliefs about the world, and uses the term *opinion* to denote the representation of a subjective belief. An 'opinion' can be interpreted as a probability measure containing secondary uncertainty, and as such subjective logic can be seen as an extension of both probability calculus and binary logic. It is suitable for modeling and analysing situations involving uncertainty and incomplete knowledge [1], [2].

Jøsang et al. [2] claims that, subjective logic is mainly designed to apply and interpret different real world problems in artificial intelligence reliability analysis [3], authentication [4], and legal reasoning [5] where evidence is gathered from multiple sources with manual intervention like the case of open systems. Subjective logic also seems very suitable for reasoning about intrusion attacks because on the one hand an attack can be considered to be a crisp event, i.e. an attack either takes place or not, while on the other beliefs about intrusion can have varying degrees of certainty [6]. By analogy we can infer that any kind of decision making process, which works on crisp event but has uncertainty associated with its judgement or consequence can be dealt with subjective logic.

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 231–242, 2010.

In a document computing area, the picture is quite different; where only source of information is the document itself. This represents more of a closed system where the information source is restricted to a particular origin; which is a document in this case. When analysing single documents using subjective logic, the sets of arguments are generated automatically as evidence from the information available in it. It is mainly done by exploiting the structure and semantics of the text being considered.

When a document is read by a human, they analyse it by identifying the main idea of the source text and filtering what is essential in the information conveyed by the text. This step further involves differentiating complementary or superfluous information according to the intended purposes of the writers, with respect to what they aim at the readers to grasp. In [7], the authors have pointed out that the context of a given piece of text is interpreted and understood by a different person in a different fashion. Thus we see that human understanding and reasoning is *subjective* in nature unlike propositional logic which deals with either truth or falsity of a statement. Furthermore information provided by different persons can be either linguistically or factually different, with a prevalent degree of impreciseness and uncertainty.

In this paper, our main aim is to formulate an enhanced model based on subjective logic to analyse documents in a way which more similar to human judgements capturing uncertainty. Each sentence of a document represents specific facts about the document; we consider them to be propositions and define 'opinions' about these propositions. Thus we present a framework for automatically determining opinions about a sentence, using subjective logic because of its property of 'uncertain probability' measure. We portray two different concepts; 'bag of words' and further enhancement of the model with semantic information from the document; as 'bag of words' tend to lose the semantic binding of the context.

2 Representing Uncertain Probabilities: Subjective Logic (SL) Basics

In subjective logic, first order measure of evidence are expressed as belief mass distribution functions over frame of discernment. All these belief measure representations in subjective logic, which are called '*opinions*', also contain a base rate parameter which express the a priori belief in the absence of evidence. Philosophically, 'opinions' are quantitative representations of evidence as perceived by humans or by other intelligent agents [8]. This portraits a scenario which is an open system where evidence are gathered from different sources.

A frame of discernment Θ contains the set of possible states. It is assumed that the system cannot be in more than one elementary state at the same time. However, if an elementary state is assumed to be true then all the superstate can be considered true as well. In fact Θ is by definition always true because it contains a true state.

The elementary states in the frame of discernment Θ will be called atomic states because they do not contain any substates. The powerset of Θ, denoted

by 2^{Θ}, contains atomic states, and all possible combinations of atomic states, including Θ. A frame of discernment can be finite or infinite, in which cases the corresponding powerset is also finite or infinite.

An observer assigns a belief mass to various states based on its strength of belief that the state (or one of its substates) is true. We have directly taken the basic definitions from the original paper [2] which we have used to build up evidence from a document in our study.

Definition 1 (Belief Mass Assignment). *Let Θ be a frame of discernment. If with each substate $x \in 2^{\Theta}$ a number $m_{\Theta}(x)$ is associated such that:*
1. $m_{\Theta}(x) \geq 0$
2. $m_{\Theta}(\emptyset) = 0$
3. $\sum_{x \in 2^{\Theta}} m_{\Theta}(x) = 1$
then m_{Θ} is called a belief mass assignment in Θ, or BMA for short. For each substate $x \in 2^{\Theta}$, the number $m_{\Theta}(x)$ is called the belief mass of x.

Definition 2 (Belief Function). *Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the belief function corresponding with m_{Θ} is the function $b: 2^{\Theta} \rightarrow [0,1]$ defined by:*

$$b(x) = \sum_{y \subseteq x} m_{\Theta}(y), \quad x, y \in 2^{\Theta} \tag{1}$$

Definition 3 (Disbelief Function). *Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the disbelief function corresponding with m_{Θ} is the function $d: 2^{\Theta} \rightarrow [0,1]$ defined by:*

$$d(x) = \sum_{y \cap x = \emptyset} m_{\Theta}(y), \quad x, y \in 2^{\Theta}. \tag{2}$$

Definition 4 (Uncertainty Function). *Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the uncertainty function corresponding with m_{Θ} is the function $u: 2^{\Theta} [0,1]$ defined by:*

$$u(x) = \sum_{\substack{y \cap x \neq \emptyset \\ y \not\subseteq x}} m_{\Theta}(y), \quad x, y \in 2^{\Theta}. \tag{3}$$

From Josang's concept, we can get the **Belief Function Additivity** which is expressed as:

$$b(x) + d(x) + u(x) = 1, \quad x \in 2^{\Theta},\ x \neq \emptyset. \tag{4}$$

Definition 5 (Relative Atomicity). *Let Θ be a frame of discernment and let $x, y \in 2^{\Theta}$. Then for any given $y \neq \emptyset$ the relative atomicity of x to y is the function $a: 2^{\Theta} \rightarrow [0,1]$ defined by:*

$$a(x/y) = \frac{|x \cap y|}{|y|}, \quad x, y \in 2^{\Theta},\ y \neq \emptyset. \tag{5}$$

It can be observed that $x \cap y = \emptyset \Rightarrow = 0$ and that $y \subseteq x \Rightarrow a(x/y) = 1$. In all other cases relative atomicity will be a value between 0 and 1. The relative atomicity of an atomic state to its frame of discernment, denoted by $a(x/\Theta)$, can simply be written as $a(x)$. If nothing else is specified, the relative atomicity of a state then refers to the frame of discernment.

Definition 6 (Probability Expectation). *Let Θ be a frame of discernment with BMA m_Θ then the probability expectation function corresponding with m_Θ is the function $E: 2^\Theta \rightarrow [0,1]$ defined by:*

$$E(x) = \sum_y m_\Theta(y) a(x/y), \quad x, y \in 2^\Theta. \tag{6}$$

Definition 7 (Opinion). *Let Θ be a binary frame of discernment with 2 atomic states x and $\neg x$, and let m_Θ be a BMA on Θ where $b(x)$, $d(x)$, $u(x)$, and $a(x)$ represent the belief, disbelief, uncertainty and relative atomicity functions on x in 2^Θ respectively. Then the opinion about x, denoted by w_x is the tuple defined by:*

$$w(x) \equiv (b(x), d(x), u(x), a(x)). \tag{7}$$

For compactness and simplicity of notation we will in the following denote belief, disbelief, uncertainty and relative atomicity functions as b_x, d_x, u_x and a_x respectively.

Definition 8 (Ordering of Opinions). *Let ω_x and ω_y , be two opinions. They can be ordered according to the following criteria by priority:*
1. The opinion with the greatest probability expectation is the greatest opinion.
2. The opinion with the least uncertainty is the greatest opinion.
3. The opinion with the least relative atomicity is the greatest opinion.

3 Subjective Logic in Document Analysis

How can we define evidence in a document related to its overall meaning[1]? This is what we are building here automatically. We consider words, phrases or co-occurrence of words, semantic associations, or a sentence itself to be evidence present in a document. Now, based on this, our basic motivation is to formulate 'opinion' about a proposition, which is a sentence in this case. Stronger the opinions about a sentence, more is its significance in the document. These opinions are measured by probability expectation of a sentence as defined in (6). Greater the probability expectation, more significant is the sentence.

3.1 Representation of a Document

Assumptions. We propose the following framework for the practical application of subjective logic in a document computing context.

[1] From here, we simply write 'evidence' to express that the "evidence in a document related to its overall meaning".

1. All the words or terms (removing the stop words) in a document are atomic. However, some sentences can have single word.
2. The sentences are unique, i.e., each of them occur only once in a given document.

A document consists of sentences. In this paper, a sentence is considered to be a set of words. In a document, sentences are separated by stop marks (".", "!", "?"). Terms (stop words excluded) are extracted and the frequencies (i.e. number of occurrences) of the words in each sentence are calculated.

Let us now define the notations which we will be using in the paper. Θ is the frame of discernment. We represent a document as a collection of words, which is

$$\Theta = D_w = \{w_1, w_2, ..., w_n\} \tag{8}$$

where, D_w is a document consisting of words. w_1, $w_2 ... w_n$ and $|D_w| = n$. Now,

$$\rho(\Theta) = \{\{w_1\}, \{w_2\}, ..., \{w_1, w_2, w_3, ..., w_n\}\} \equiv 2^{\Theta} \tag{9}$$

$$|\rho(\Theta)| = 2^n \tag{10}$$

Since a document is a collection of sentences, it can also be represented as

$$D_s = \{s_1, s_2, ..., s_t\} \tag{11}$$

where t is a finite integer and each s_i is an element of $\rho(\Theta)$. Each sentence is comprised of words, which belong to the whole word collection of the document D_w. We thus represent each sentence by,

$$S_l = \{w_i w_k ... w_r\} \in \Theta \tag{12}$$

where, $1 \leq i,\ k,\ r \leq n$ and $S_l \in \rho(\Theta)$.

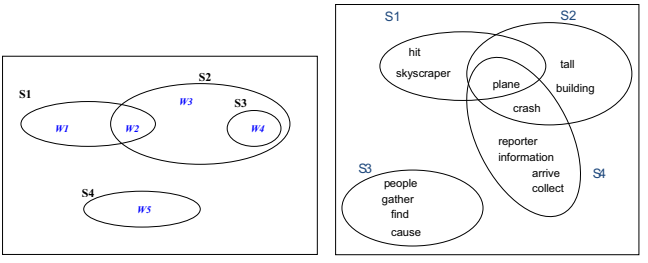
3.2 Example of Documents

Eg:1- A generic example. In fig.1, we illustrate a generic document D with four sentences $D_s = \{s_1, s_2, s_3, s_4\}$ and a list of unique words $D_w = \{w_1, w_2, w_3, w_4, w_5\}$. Atomic events are the single words w_1 to w_5 and the non atomic events are the sentences from s_1 to s_4; but in this case s_3 and s_4 are atomic. Each sentence is composed of both atomic and non atomic events. These are used as evidence for subjective logic formulation in this study.

Eg:2- A specific example. Here is another sample document which consists of four different real sentences, $D_s = \{s_1, s_2, s_3, s_4\}$.

1. A plane hits a skyscraper.
2. A plane crashed into a tall building.
3. People gathered to find out the cause.
4. Reporters arrived to collect information about the crash.

We will refer to this example in the following sections for explaining our representations of subjective logic.



(a) Generic example (b) Real example

Fig. 1. Representation of 'bag of words' form of sentences in a document

3.3 Modeling 'Opinions' about a Sentence in a Document

In this section we present the formulation of 'opinion' about a sentence in a document using subjective logic explained in sec.2. Now, let us explain step-wise computation of opinion based on sec.2 equations for the examples considered. Examples 1 and 2 as shown in subsec.3.2 are of same kind expect the fact that eg.2 represents real words in place of symbols of eg.1. So, in this section, we illustrate the computation for any one of these, i.e, eg.1.

BMA calculation: BMA is explained in def.1. Now, for a document, we calculate BMA for each event by,

$$m(x) = \frac{F(x)}{Z}, \tag{13}$$

where $F(x) = \sum_{k=1}^{t} f_{x_k}$, where t is the total number of sentences in the document, $x \in 2^{\Theta}$, and f_{x_k} is the frequency of occurrence of event x in sentence k. In other words, it is the total frequency of that event in all the sentences (or the whole document).

$$Z = \sum_{\forall x \neq \Phi} F(x), \quad x \in 2^{\Theta} \tag{14}$$

Z is the total frequency of the all the existing events (whose frequency is non zero). In the given example 1, we have 7 valid states and their corresponding frequencies in the document are: $F(\{\{w_1\}) = 1,\ F(\{w_2\}) = 2,\ F(\{w_3\}) = 1,\ F(\{w_4\}) = 2,\ F(\{w_5\}) = 1, F(\{w_1, w_2\}) = 1,\ F(\{w_2, w_3, w_4\}) = 1\}$. Therefore, $Z = 9$ in this case. Using (13), we calculate BMA for each of the states (or events) in the given example shown in fig.1. So, for eg.1, we have $m(w_1) = \frac{1}{9},\ m(w_2) = \frac{2}{9},\ m(s_1) = m(w_1, w_2) = \frac{1}{9}$ ... $m(s_4) = m(w_5) = \frac{1}{9}$

Figure 1(b) is the diagrammatic representation of example 2 of subsec.3.2. The words shown in the diagram are processed by stemming and stop words removed. This is a 'bag of words' representation of the document. Here, the number of atomic states (or events) are 14 and non-atomic states are 4. Now total frequency for all of these 18 states is 21 (which means $Z = 21$) (calculated exactly in the same way as the generic example). Now, using (13), we get the BMA for each of these states respectively; provided the frequency of each non stop words in

each sentence is 1 as per example 2; such as, $m(hit) = \frac{1}{21}$, $m(skyscraper) = \frac{1}{21}$, $m(plane) = \frac{2}{21}$, $m(crash) = \frac{2}{21}$...and so on.

Belief, Disbelief, and Uncertainty. Using definitions from sec. 2, we use equations (1), (2), and (3) to calculate the belief, disbelief and uncertainty of a sentence respectively. We illustrate the computation using eg.1's s_1 by,
$b(s_1) = m(w_1) + m(w_2) + m(w_1, w_2) = \frac{4}{9}$
$d(s_1) = m(w_3) + m(w_4) + m(w_5) = \frac{4}{9}$
$u(s_1) = 1 - (b(s_1) + d(s_1)) = \frac{1}{9}$; using (4). For eg.2, we calculate these parameters in the same way as shown for eg.1.

Calculation of relative atomicity, probability expectation and 'opinion' about a sentence. Here inorder to calculate probability expectation, we first need to find relative atomicities. Again, using equations (5), (6), and (7) of sec. 2, we compute relative atomicity for sentence s_1 of eg.1 as:
$a(s_1/w_1) = \frac{|s_1 \cap w_1|}{|w_1|} = \frac{1}{1} = 1$
$a(s_1/w_2) = \frac{|s_1 \cap w_2|}{|w_2|} = \frac{1}{1} = 1$
... $a(s_1/w_5) = a(s_1/s_4) = \frac{|s_1 \cap w_5|}{|w_5|} = \frac{0}{1} = 0$
Likewise, we calculate the atomicity for other sentences. So, the probability expectation is then obtained by, $E(s_1) = m(w_1)a(s_1/w_1) + m(w_2)a(s_1/w_2) + m(\{w_1, w_2\})a(s_1/\{w_1, w_2\}) + ... + m(w_5)a(s_1/w_5)$ Thus $E(s_1) = \frac{13}{27} = 0.48$. Thus opinion ($\omega_{s_1}$ or $\omega(s_1)$)about a sentence s_1 can be expressed using these four parameters by (7) as, $\omega(s_1) = (0.44,\ 0.44,\ 0.11,\ 4.33)$. Likewise, we compute the parameters in the same way for eg.2.

4 Extension of Subjective Logic with Semantic Information of a Document

In this section, we extend basic subjective logic model explained in the previous sec.3 where we have already shown, how to define 'opinion' about a sentence in a document considering words, phrases and sentences to be atomic or composite events as different sources of evidence. But we used 'bag of words' for formulating this measure, which is a superficial approach according to information retrieval context. Only root form of words are used for frequency measure where the underlying semantic relations between events are ignored. Hence, here we use semantic similarity as a measure to find relatedness of concepts of sentences whose 'opinions' are desired.

4.1 Why Do We Need Semantic Information?

What we write or say are very context sensitive. A same word can be linguistically expressed differently in different contexts; at the same time, different words can linguistically express same thing at a particular context. If we look at our example 2 of sec.3,

sentence 1: "A plane hits a skyscraper."
sentence 2: "A plane crashed into a tall building."
Anyone can easily infer that both the sentences are similar in their context though different words are used to express it. Similarly, if we look at other two sentences,
sentence 3: "People gathered to find out the cause."
sentence 4: "Reporters arrived to collect information about the crash."
the inference will again be same.

In fig.2, we illustrate 4 sentences with overlap only if the words in them are exactly same with same parts of speech (POS) tag. The dotted lines show which words are most similar in their meanings in sentences. In sentence 1 and 2, phrases 'hits a skyscraper' is similar as 'crashed into a tall building' or the word pairs like 'hits' and 'crash', 'people' and 'reporters', 'gathered' and 'arrived' etc have great similarity in their meanings. Index terms are not enough to find this kind of analogies as they look for only exact matches between words, which in this case failed to find any kind of relations among the sentences of eg.2. We thus extend and redefine subjective logic belief measures by incorporating semantic information about word, phrase, and sentence similarities from the document. To accomplish this, we used WordNet [9] as a lexical dictionary to gather semantic information about each word of sentences; thus making the whole decision making process context sensitive.

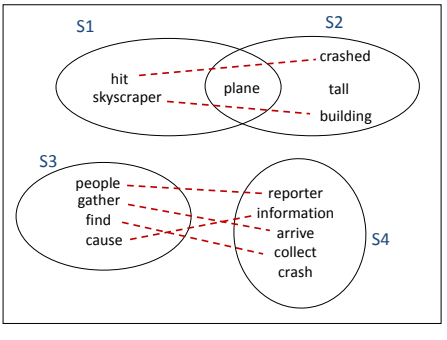


Fig. 2. An example of a document with semantic overlap

4.2 Measure of Semantic Similarity

Two words are contextually similar, if they share similar senses. To perform this automatically, we require WordNet [9], an online lexicon database, to compute this measure. Each word can have one or more synsets based on different senses of their existence also in different parts of speech like noun, verb, adjective, and adverb. Same word in different parts of speech convey different meaning to the context in which they are used. In both sentences 2 and 4, the word 'crash' occurs but in two different parts of speech; *verb* for former and *noun* for latter; obviously imparting different sense to the context. So, considering only root form

of any word, misses out the semantic meaning of it. To overcome this problem of 'bag of words' concept, we introduce a similarity measure α.

Definition 9 (Semantic Similarity). *Let Θ be a frame of discernment, and let $x, y \in 2^{\Theta}$. Then for any x and y, semantic similarity is the function $\alpha : 2^{\Theta} \rightarrow [0, 1]$ defined by*

$$\alpha(x, y) = SimScore(x, y) \quad x, y \in 2^{\Theta}. \tag{15}$$

where $SimScore(x, y)$ is a function which determines the semantic similarity measure between x and y provided the elements of x and y are in the same parts of speech. This can be any kind of similarity score like gloss overlap [10], path based measures [11], [12], edge based measures or sentence similarity measure [13]. We use a threshold κ to define the degree of similarity. Thus we can say,
$\alpha = 1$, x and y are identical
$\kappa \leq \alpha < 1$, x and y are similar
$0 \leq \alpha < \kappa$, x and y are dissimilar, where $\kappa \in [0, 1]$.
Generally $\kappa = 0.5$ is taken as a standard value for similarity scores [14].

4.3 Enhanced Belief Measures Using Semantic Information

In this section, we present an extension of subjective logic formulation for document analysis using semantic information. The equations are redefined using the similarity score α as shown in (15).

Computation of BMA: We compute belief mass assignment in the same way as shown in (13). The only difference is in the frequency calculation of atomic states; where we consider parts of speech of the words as well instead of only the root forms. For example, in eg.2, the word 'crash' is in two different parts of speech (POS) in s_2 and s_4, so these belong to two separate atomic events. Likewise, 'plane' being in same POS (noun) for both sentences 1 and 2 will have a total frequency count of 2 for that state.

Now, for example 2, there are 18 different states existing, and the frequency of each state can be represented as: $F(\{plane^{noun}\}) = 2$, $F(\{hit^{verb}\}) = 1$, $F(\{crash^{noun}\}) = 1$, $F(\{crash^{verb}\}) = 1$, $F(\{building^{noun}\}) = 1$, ..., $F(\{plane^{noun}, hit^{verb}, skyscraper^{noun}\}) = F(s_1) = 1$. Thus we get $Z = 19$ for this case. We compute BMA by (13) using these values computed, $m(plane^{noun}) = \frac{2}{19}$, $m(building^{noun}) = \frac{1}{19}$ and likewise for other events.

Similarity scores for example 2: For different belief measures, we need to use similarity score between two events. Let us assign similarity scores for each word pair belonging to same parts of speech (using example 2). Suppose, sentence 1 be the proposition we considered. So some of the similarity scores which are necessary for finding opinion about s_1 can be:
$\alpha(plane^{noun}_{s_1}, plane^{noun}_{s_2}) = 1$, $\alpha(plane^{noun}_{s_1}, building^{noun}_{s_2}) = 0.1$,
$\alpha(hit^{verb}_{s_1}, crash^{verb}_{s_2}) = 0.7$, $\alpha(skyscraper^{noun}_{s_1}, plane^{noun}_{s_2}) = 0.08$,
$\alpha(skyscraper^{noun}_{s_1}, building^{noun}_{s_2}) = 0.85$, $\alpha(plane^{noun}_{s_1}, people^{noun}_{s_3}) = 0.03$,

$\alpha(plane^{noun}_{s_1}, cause^{noun}_{s_3}) = 0.01$, ...
Likewise we compute α for all other word pairs. For this analysis, we also require similarity between composite events which can be computed using hierarchical document signature [13]. Using this method, word-sentence similarity and sentence-sentence similarity can be computed. Now, let us present some similarities of composite events for s_1,
$\alpha(s_1, s_2) = 0.5,\ \alpha(s_1, plane^{noun}_{s_2}) = 0.8,\ \alpha(s_1, crash^{verb}_{s_2}) = 0.6,\ \alpha(s_1, tall^{adj}_{s_2}) = 0.01,\ ...,\ \alpha(s_1, people^{noun}_{s_3}) = 0.02,\ \alpha(s_1, find^{verb}_{s_3}) = 0.01,\ \alpha(s_1, gather^{verb}_{s_3}) = 0.01,\ ..., \alpha(s_1, reporter^{noun}_{s_4}) = 0.01,\ \alpha(s_1, s_4) = 0.2$. The values of α shown here are solely based on intuitions and general understanding of semantics of the text considered.

Definition 10 (Semantic Belief Function). *Let Θ be a frame of discernment, m_Θ be a BMA and α be semantic similarity on Θ respectively. Then the belief function corresponding with m_Θ and alpha is the function $b^s : 2^\Theta \rightarrow [0, 1]$ defined by:*

$$b^s(x) = \sum_{\forall y | \alpha(x,y) \leq 1} m_\Theta(y), \quad x, y \in 2^\Theta,\ y \subseteq x \tag{16}$$

Thus, as per the similarity values provided, belief of sentence 1 is computed as,

$$\begin{aligned} b^s(s_1) &= m(plane^{noun}) \times \alpha(s_1, plane^{noun}) + m(hit^{verb}) \times \alpha(s_1, hit^{verb}) \\ &\quad + m(skyscraper^{noun}) \times \alpha(s_1, skyscraper^{noun}) + m(s_1) \times \alpha(s_1, s_1) \\ &= \frac{2}{19} \times 0.8 + \frac{1}{19} \times 0.5 + \frac{1}{19} \times 0.4 + \frac{1}{19} \times 0.8 \end{aligned}$$

Definition 11 (Semantic Disbelief Function). *Let Θ be a frame of discernment, m_Θ be a BMA and α be semantic similarity on Θ respectively. Then the disbelief function corresponding with m_Θ and α is the function $d^s : 2^\Theta \rightarrow [0, 1]$ defined by:*

$$d^s(x) = \sum_{\forall y | \alpha(x,y) < \kappa} \alpha(x, y) m_\Theta(y), \quad x, y \in 2^\Theta. \tag{17}$$

Now for disbelief calculation, we look for $0 \leq \alpha < \kappa$. Here, $\alpha(s_1, people^{noun}_{s_3}) = 0.02$, $\alpha(s_1, gather^{verb}_{s_3}) = 0.01$, $\alpha(s_1, reporter^{noun}_{s_4}) = 0.01$, are all less than $\kappa = 0.5$, s_1 do not have significant semantic overlap with sentences s_3 and s_4. So, they are part of disbelief. Thus,

$$\begin{aligned} d^s(s_1) &= \alpha(s_1, people^{noun}_{s_3}) m(people^{noun}_{s_3}) + \\ &\quad \alpha(s_1, gather^{verb}_{s_3}) m(gather^{verb}_{s_3}) + \\ &\quad ... + \alpha(s_1, reporter^{noun}_{s_4}) m(reporter^{noun}_{s_4}) + \\ &\quad ... + \alpha(s_1, s_3) m(s_3) + ... \\ &= (0.02 \times \frac{1}{19}) + (0.01 \times \frac{1}{19}) + (0.01 \times \frac{1}{19}) + ... \end{aligned}$$

Definition 12 (Semantic Uncertainty Function). *Let Θ be a frame of discernment, m_Θ be a BMA and α be semantic similarity on Θ respectively. Then the disbelief function corresponding with m_Θ and α is the function $u^s : 2^\Theta \rightarrow [0,1]$ defined by:*

$$u^s(x) = \sum_{1>\forall y|\alpha(x,y)\geq\kappa} \alpha(x,y)m_\Theta(y), \quad x,y \in 2^\Theta. \tag{18}$$

In case of uncertainty calculation, we consider $1 > \alpha \geq \kappa$, where $\kappa = 0.5$. Here, $\alpha(s_1, s_2) = 0.5$, $\alpha(s_1, plane_{s_2}^{noun}) = 0.8$, $\alpha(s_1, crash_{s_2}^{verb}) = 0.6$,..., have $\alpha \geq 0.5$; so these implies that s_1 has substantial overlap with s_2. Thus,

$$\begin{aligned} u^s(s_1) = {} & \alpha(s_1, plane_{s_2}^{noun})m(plane_{s_2}^{noun}) + \\ & \alpha(s_1, crash_{s_2}^{verb})m(crash_{s_2}^{verb}) + ... \\ & \alpha(s_1, s_2)m(s_2) \\ = {} & (0.8 \times \frac{2}{19}) + (0.6 \times \frac{1}{19}) + ... + (0.5 \times \frac{1}{19}) \end{aligned}$$

In this situation, the *Semantic Belief Function* will no longer hold strict additivity like (4) and is thus expressed as:

$$b^s(x) + d^s(x) + u^s(x) \leq 1, \quad x \in 2^\Theta,\ x \neq \emptyset. \tag{19}$$

Definition 13 (Semantic Relative Atomicity). *Let Θ be a frame of discernment, let $x, y \in 2^\Theta$, and let $\alpha(x,y)$ be semantic similarity of x and y. Then for any given $y \neq \emptyset$ the relative atomicity of x to y is the function $a : 2^\Theta \rightarrow [0,1]$ defined by:*

$$a^s(x/y) = \frac{\sum_{j=1}^{|y|} \bigvee_{i=1}^{|x|} \alpha(x_i, y_j)}{|y|}, \quad x, y \in 2^\Theta,\ x_i \in x,\ y_j \in y. \tag{20}$$

where x_i and y_i are atomic elements of x and y respectively. So, according to fig.2, $a^s(s_1/s_2) = \frac{1.0+0.7+0.85}{4}$, where $\alpha(plane_{s_1}^{noun}, plane_{s_2}^{noun}) = 1$, $\alpha(hit_{s_1}^{verb}, crash_{s_2}^{verb}) = 0.7$, and $\alpha(skyscraper_{s_1}^{noun}, building_{s_2}^{noun}) = 0.85$ (assuming α values based on meanings) respectively; but this is not the case when 'bag of words' are considered.

The **probability expectation** and **opinion** will remain same as (6) and (7) except the fact that the parameters will be replaced by the extended parameters based on semantic analysis, and hence represented as,

$$E^s(x) = \sum_y m_\Theta(y)a^s(x/y), \quad x, y \in 2^\Theta. \tag{21}$$

$$w^s(x) \equiv (b^s(x),\ d^s(x),\ u^s(x),\ a^s(x)). \tag{22}$$

Now, using the parameters like belief, disbelief, uncertainty, relative atomicity and BMA computed for s_1 we can get probability expectation (21) and opinion (22).

5 Conclusion

In this paper, we presented an enhanced framework of subjective logic for document analysis. Two different aspects of the model are shown. The former is simple computation of the original subjective logic [2] model using 'bag of words'. For the latter, we redefined all the definitions based on the semantic relatedness of concepts encountered in sentences and have shown how this approach is more significant for document analysis. As a future work we tend to determine the similarity threshold κ automatically by using some optimization algorithms.

References

1. Jøsang, A.: Artificial reasoning with subjective logic. In: Proceedings of the Second Australian Workshop on Commonsense Reasoning, vol. 48 (1997), Perth:[sn]
2. Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
3. Jøsang, A.: Reliability Analysis with Uncertain Probabilities. In: Proceedings of the 4th International Conference on Probabilistic Safety Assessment and Management (PSAM4). Springer, Heidelberg (1998)
4. Jøsang, A.: An algebra for assessing trust in certification chains. In: Proceedings of the Network and Distributed Systems Security Symposium (NDSS 1999). The Internet Society, San Diego (1999) (Citeseer)
5. Jøsang, A., Bondi, V.: Legal reasoning with subjective logic. Artificial Intelligence and Law 8(4), 289–315 (2000)
6. Svensson, H., Jøsang, A.: Correlation of Intrusion Alarms with Subjective Logic
7. Pardo, T., Rino, L., Nunes, M.: Extractive summarization: how to identify the gist of a text. In: The Proceedings of the 1st International Information Technology Symposium–I2TS, Citeseer, pp. 1–6 (2002)
8. Jøsang, A.: Belief Calculus. ArXiv Computer Science e-prints (June 2006)
9. Miller, G.: WordNet: a lexical database for English. Communications of the ACM 38(11), 41 (1995)
10. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation, pp. 24–26. ACM, New York (1986)
11. Lin, D.: Using syntactic dependency as local context to resolve word sense ambiguity. In: Annual Meeting-Association For Computational Linguistics, vol. 35, pp. 64–71. Association For Computational Linguistics (1997)
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: International Joint Conference on Artificial Intelligence, Citeseer, vol. 14, pp. 448–453 (1995)
13. Manna, S., Gedeon, T.: Semantic Hierarchical Document Signature In Determining Sentence Similarity. In: Proceedings of the 19th International Conference on Fuzzy Systems (accepted 2010)
14. Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 305–316. Springer, Heidelberg (2008)

Ontology-Based Anonymization of Categorical Values

Sergio Martínez, David Sánchez, and Aida Valls

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Avda. Països Catalans, 26, 43007 Tarragona, Spain
{sergio.martinezl, aida.valls, david.sanchez}@urv.cat

Abstract. The analysis of sensible data requires a proper anonymization of values in order to preserve the privacy of individuals. Information loss should be minimized during the masking process in order to enable a proper exploitation of data. Even though several masking methods have been designed for numerical data, very few of them deal with categorical (textual) information. In this case, the quality of the anonymized dataset is closely related to the preservation of semantics, a dimension which is commonly neglected of shallowly considered in related words. In this paper, a new masking method for unbounded categorical attributes is proposed. It relies on the knowledge modeled in ontologies in order to semantically interpret the input data and perform data transformations aiming to minimize the loss of semantic content. On the contrary to exhaustive methods based on simple hierarchical structures, our approach relies on a set of heuristics in order to guide and optimize the masking process, ensuring its scalability when dealing with big and heterogenous datasets and wide ontologies. The evaluation performed over real textual data suggests that our method is able to produce anonymized datasets which significantly preserve data semantics in comparison to apporaches based on data distribution metrics.

Keywords: Ontologies, Data analysis, Privacy-preserving data-mining, K-anonymity, Semantic similarity.

1 Introduction

Statistical agencies are an important source of information for intelligent data analysis and decision making. Those agencies collect responses of a set of individuals for which privacy must be guaranteed. So, before distributing the data, a masking method should be used in order to anonymize the data file and minimize the re-identification risk. The privacy level associated to masked data is typically related to the fulfilment of the *k-anonymity* property [16]. This property establishes that each anonymized record in a data set (i.e. a set of attribute values associated to an individual) has to be indistinguishable with at least *k*-1 other records within the same dataset, according to its individual attribute values.

However, in order to preserve the utility of the values (i.e. to make the anonymized data as useful as possible from the analysis and data mining point of view), it is important that the anonymization method minimizes the information loss that is

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 243–254, 2010.

inherent to the masking process. This is measured by means of a quality metric. Up to this moment, most of the attention has been paid to numerical data or bounded categorical attributes. The goal of the masking methods for numerical data was to maintain the statistical characteristics of the dataset [4]. For categorical attributes, which represent a discrete enumeration of modalities (i.e. bounded vocabulary), quality metrics are focused on maintaining the probability distribution of the values in the masked file. This has been criticized by several authors [19] as value distribution does not capture important dimensions of data utility. In fact, as categorical attributes typically represent concepts, their utility should be associated to the preservation of their inherent semantics. Omitting those semantics during the anonymization process can hamper the application of data analysis or decision making processes on those data, since the conclusions obtained can be significantly different from those obtained from the original data file.

In any case, with the success of the Information Society, textual data have grown both in size and importance. Those values can be obtained with traditional questionnaires where the user can answer with a short sentence or a noun phrase, such as "*Main hobby*" or "*Most preferred type of food*". This kind of attributes has a potentially unbounded set of values that represent a concept with a concrete semantic. Those attributes are more challenging than those corresponding to a limited set of modalities. In order to properly interpret and compare them, the similarities between their meaning, at a conceptual level, should be taken into consideration (e.g. for hobbies, *trekking* is more similar to *jogging* than to *dancing*).

Due to the ambiguity of human languages and the complexity and knowledge modelling, very few masking methods have considered the semantics of attribute values in some degree. In fact, many approaches [1, 15, 16] completely ignore this issue, dealing with textual data in a naïve way, proposing arbitrary suppressions or substitutions aimed to fulfil *k*-anonymy and preserve the distribution of the input data, but neglecting the importance of the meaning of the data. As it will be discussed in Section 2, even though there exist approaches exploiting knowledge structures during the anonymization, they consider semantics in a very shallow and ad-hoc manner and tackle the anonymization in an exhaustive manner, hampering their scalability and applicability as a general-purpose solution.

In order to overcome those limitations, we propose a new method of local anonymization for unbounded categorical attributes, which exploits ontologies [2] as knowledge background to support the anonymization process from a semantic point of view. Ontologies offer a formal, explicit and machine readable structuring of a set of concepts by means of a semantic network where multiple hierarchies are defined and semantic relations are explicitly modelled as links between concepts [6]. Thanks to initiatives such as the Semantic Web [3], many ontologies have been created in the last years, bringing the development of general purpose knowledge sources (such as WordNet [5] for English words), as well as specific domain terminologies (e.g. medical sources such as UMLS -Unified Medical Language System-).

Due to the large size of general purpose ontologies (with respect to ad-hoc knowledge structured exploited in previous approaches [1, 7, 12, 15, 16]), our algorithm tackles the anonymization in an heuristic fashion, providing better scalability with respect to the size of the ontology and the input data than related works based on exhaustive search.

The rest of the paper is organized as follows. Section 2 reviews methods for privacy protection of categorical data that take into account some kind of semantic information. Section 3 introduces classical metrics aimed to measure data quality and present other ways of semantically measuring the information loss by exploiting ontologies. In section 4, the proposed anonymization method is detailed. Section 5 is devoted to evaluate our method by applying it to real data obtained from a survey at the National Park "*Delta del Ebre*" in Catalonia, Spain. The final section contains the conclusions and future work.

2 Related Work

In the previous knowledge-based masking methods, the set of values of a categorical attribute are represented by means of Value Generalization Hierarchies (VGHs) [1, 7, 12, 15, 16]. In those cases, ad-hoc manually constructed tree-like structures are defined according to input data, where categorical labels represent leafs of the hierarchy and they are recursively subsumed by common generalizations. The masking process consists on substituting the original values by a more general one, obtained from the hierarchical structure. This generalization process decreases the number of distinct tuples and, in consequence, increases the level of *k*-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. Typically, the selection is made according to a quality metric that measures the information loss derived from the value substitution.

More in detail, in [11, 15, 16] authors propose a hierarchical scheme in which all values of an attribute are generalized to the same level of the VGH. The number of valid generalizations for an attribute is the height of the VGH for that attribute. The concrete generalization is selected by generating all the possible ones for each value and selecting the combination that provides the closest generalizations in all cases fulfilling the desired level of *k*-anonymity. In this case, the level of generalization is used as a measure of information loss.

Iyengar [8] presented a more flexible scheme which also uses a VGH, where each value of an attribute can be generalized to a different level of the hierarchy. This scheme allows a much larger space of possible generalizations. Again, for all values, all the possible generalizations fulfilling the *k*-anonymity are generated. Then, a genetic algorithm finds the optimization of a set of information loss metrics.

T. Li and N. Li [12] propose three generalization schemes. First, the Set Partitioning Scheme (SPS) represents an unsupervised approach in which each partition of the attribute domain represents a generalization. This supposes the most flexible generalization scheme but the size of the solution space grows enormously, meanwhile the benefits of a semantically coherent VGH are not exploited. The Guided Set Partitioning Scheme (GSPS) uses a VGH to restrict the partitions of the attribute domain and exploits the height of the lowest common ancestor of two values as a metric of semantic distance. Finally, the Guided Oriented Partition Scheme (GOPS) adds ordering restrictions to the generalized groups of values to restrict even more the possible generalizations. In all three cases, all the possible generalizations allowed by the proposed scheme are constructed, selecting the one that minimizes the information loss (evaluated by means of the discernibility metric [1]).

He and Naughton [7] propose a partitioning algorithm in which generalizations are created in a Top-Down fashion and the best one, according to quality metric (Normalized Certainty Penalty [17]), is recursively refined. Xu et al. [19] proposes a Utility-based generalization algorithm. The method supports defining different "utility" functions for each attribute, according to the importance of each attribute.

All the approaches relying on a VGH present a series of drawbacks. On one hand, VGHs are manually constructed from the attribute value set of the input data. So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If input data values change, the VGH should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in [12] a dozen of different values per attribute are considered in average) this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with hundreds or thousands of individual answers). On the other hand, the fact that VGHs are constructed from input data (which represents a limited sample of the underlying domain of knowledge), produces ad-hoc and small hierarchies with a much reduced taxonomical detail. It is common to observe VGHs with three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels [5]. From a semantic point of view, VGHs offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. As a result, the space for valid generalizations that a VGH offers would be much smaller than when exploiting an ontology. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss due to generalizations. As stated above, some authors try to overcome this problem by making arbitrary generalizations, but this introduces a considerable computational burden and lacks of a proper semantic background. Moreover, the quality of the result would depend on the structure of the VGH that, due to its limited scope, offers a partial and biased view of the domain.

From the point of view of semantic understanding of the input data, in order to overcome the limitations of the presented methods, one may consider their application over a wide and detailed general ontology like WordNet. WordNet [5] is a freely available lexical database that describes and organizes more than 100,000 general English concepts, which are semantically structured in an ontological fashion. WordNet contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets), each expressing a distinct concept (i.e. a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (subclass-of), meronymy (part-of), etc. The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far the most common relation, representing more than an 80% of all the modeled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies) and up to 29 different senses (for the "line" word). Considering those dimensions, the size of the generalization space would be several orders of magnitude bigger than when using ad-hoc VGHs. However, as most of the presented approaches make generalizations in an exhaustive fashion, the generalization space is exponentially large according to the depth of the hierarchy, the branching factor and the values to evaluate. So, those approaches are computationally too expensive and hardly applicable in such a big ontology like WordNet.

In order to be able to exploit the semantic background provided by big ontologies like WordNet, we present a non-exhaustive heuristic value substitution which, bounding the search space according to the input data values and based on the theory of *semantic similarity* (see Section 3), is able to scale well in such a big ontology while minimizing the loss of semantics.

3 Quality Metrics

As stated above, the goal of an anonymization method is finding a transformation of the original data, which satisfies *k-anonymity* while minimizing the information loss and, in consequence, maximizing the utility of the resulting data.

In the literature, various metrics have been proposed and exploited [1, 7, 8, 11, 12, 19] to measure the quality of anonymized data. Classical metrics, such as Dicernability Metric (DM) [1], evaluate the distribution of *n* records (corresponding to *n* individuals) into *g* groups of identical values, generated after the anonymization process. Concretely, (DM) assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalization (1). A uniform distribution of values in equally sized groups (with respect to the original data) is the goal.

$$DM = \sum_{i=1}^{n} |g_i|^2 \tag{1}$$

However, metrics based on data distribution do not capture how *semantically similar* the anonymized set is with respect to the original data. As stated in the introduction, preservation of semantics when dealing with textual attributes is crucial in order to be able to interpret and exploit anonymized data. In fact, this aspect is, from the utility point of view, more important than the distribution of the anonymized dataset when aiming to describe or understand a record by means of its attributes.

In order to minimize the loss of semantics between original and anonymized datasets, we propose relying on the theory of *semantic similarity* [9]. Semantic similarity measures the taxonomical alikeness between words based on the semantic evidences extracted from one or several knowledge sources. Ontologies like WordNet offer wide and detailed views of knowledge domains and, in consequence, represent an ideal source from which computing semantic similarity [9]. As stated in the introduction, ontologies offer a graph model in which semantic interrelations are modeled as links between concepts. As a result, semantic similarity can be estimated as a function of the taxonomic inter-link distance.

In an is-a hierarchy, the simplest way to estimate the distance between two concepts c_1 and c_2 is by calculating the shortest *Path Length* (i.e. the minimum number of links) connecting these concepts (2) [14].

$$dis_{pL}(c_1, c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \tag{2}$$

However, this measure omits the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar that those belonging to a lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure [18] also takes into account the depth of the concepts in the hierarchy (3).

$$sim_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}, \tag{3}$$

where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

Based on the same principles Leacock and Chodorow [10] also proposed a measure that considers both the shortest path between two concepts (in fact, the number of nodes N_p from c_1 to c_2) and the depth D of the taxonomy in which they occur in a non-linear fashion (4).

$$sim_{l\&c}(c_1, c_2) = -\log(N_p / 2D) \tag{4}$$

Those measures will be exploited by our approach in order minimize the loss of semantics during the substitution of sensible values.

4 Ontology-Based Anonymization of Categorical Data

Exhaustive generalization methods are too expensive to be applicable over wide ontologies like WordNet. Moreover, the fact that values to anonymize correspond to leafs of the VGH implies that values are only substituted by more general ones (which unnecessarily imposes constraints on the space of valid generalizations).

Our approach, which aims to provide local anonymization of attribute values, tackles the problem in a different manner. Thanks to the wide coverage of WordNet, one would be able to map sensible values to ontological nodes which do not necessary represent leafs of a hierarchy. As a result, semantically related concepts can be retrieved going through the hierarchy/ies to which the value belongs. Moreover ontological hierarchies are designed in a much general and fine grained fashion than ad-hoc VGHs, according to the agreement of domain knowledge experts and the input data. Those facts open the possibility of substituting sensible values by a much wider and knowledge-coherent set of semantically similar elements, including taxonomical subsumers (as done in generalization methods) but also with hierarchical siblings (with the same taxonomical depth) or specializations (located in a lower level). In fact, in many situations, a specialization may be more similar that a subsumer because, as stated in section 3, concepts belonging to lower levels of a hierarchy have less differentiated meanings due to their concreteness. As a result, the value change would result in less information loss and a higher preservation of data utility from a semantic point of view.

In order to ensure that value substitutions lead to the fulfillment of the desired degree of privacy, we should substitute each sensible value for another one that increases the level of k-anonymity. This implies that either value pairs are substituted for a new one which is "near" to both of them, or that one value is changed for another one already existing in the data set; in both cases, the goal is to make both values indistinguishable. It is important to note that, in all cases, the loss of semantic content would be equivalent: if all values of the dataset are semantically far, so are their related nodes, resulting in an inevitable high loss of semantics either by changing them for the nearest node to both of them or by substituting one for the

other. As the first option would lead to an enormous set of possible substitutions according to all the semantically related concepts available in the ontology for each sensible value, we opted for the second strategy. As a result, the space of valid substitutions is bounded to the number of *different* values available in the dataset.

The most appropriate value to which a non anonymous one should be substituted is the one that minimizes the semantic distance with respect to the original. So, semantic similarity metrics introduced in section 3 (which explore and quantify the distance of ontological nodes in the semantic network) can be used to select the substitution and minimize the loss of semantic content. As a result of a value replacement, the number of different values is decreased and the *k*-anonymity is increased. The process is repeated until the whole dataset fulfills the desired *k*-anonymity level.

As we are dealing with values represented by text labels, it is also necessary to morphologically process them in order to detect different lexicalizations of the same concept (e.g. singular/plural forms). We apply a stemming algorithm to detect conceptually equivalent values in the dataset.

Notice that the order in which the values to be replaced are selected may affect the anonymization. The generation of the optimum result implies generating all possible substitution iterations for all sensible values and picking the order that maximizes the quality of the result set. As unbounded textual attributes may usually correspond to a high number of different answers, many of them being unique, the amount of values not fulfilling the *k*-anonymity would be high. Consequently, as the cost of generating all the possible combinations is O(n!), it is computationally too expensive. In order to ensure the scalability of our approach, we implemented several heuristics that aim to select, at each step, the substitution that would likely maximize the quality of the result.

The first heuristic consists on selecting the value with the lowest number of repetitions in the original set (i.e. the more identifiable). The motivation is that those values would require a higher number of substitutions in order to fulfill the desired *k*-anonymity level. In case of a tie (e.g. several unique values, which would be very common with free text attributes), the algorithm selects the value for which its best substitution (according to the quality metric) leads to the minimum semantic information loss (according to the same quality metric), aiming to maximize the quality of the result dataset. Finally, if several replacements imply the same information loss (which would be quite rare), the algorithm selects the value for which the *k*-anonymity level resulting from that change is lower. Again, values which are more difficult to anonymize are prioritized, as they require more substitutions.

Formally, the algorithm has the following inputs: *D*, a set of *n* categorical values for a single attribute (i.e. an unbounded list of textual noun phrases, each one referring to an ontological concept) and the desired level of *k*-anonymity. The algorithm outputs the anonymized version of *D*.

```
1   Ontology-based local anonymization (D, k)
2      D' := stem (D)
3      D' := rank by number of repetitions (D')
4      v := first value (D')
5      while (number of repetitions (v, D') < k) do
6         V := values with the same number of repetitions (v, D')
7         V_max := set of values with the maximum similarity (D', V)
8         v' := value with minimum resulting k-anonymity (D', V_max)
```

```
9          D' := replace all occurrences of the value in the set (v', D')
10         D' := rank by number of repetitions (D')
11         v := first value (D')
12      end while
13   end
```

First, all words of the attribute dataset are stemmed, so that, two words are considered equal if their morphological roots are identical (line #2). The set is ascending ranked according to the number of value repetitions; then, the first value (v) is the register with the lowest k-anonymity (line #4). It checks if the corresponding value fulfils the k-anonymity according to the number of repetitions (line #5). If k-anonymity is fulfilled, the entire set will be anonymized. Otherwise, the value should be replaced. The algorithm selects all the values with the same minimum number of repetitions (line #6) and finds another value in the dataset which results in the maximum semantic similarity according to a given semantic metric (from those introduced in section 3) (line #7). If several substitutions are equally optimum, the value whose replacement results in the lowest k-anonymity level (i.e. repetitions) is selected (line #8). Finally, all the occurrences in the dataset for that value are substituted (line #9) and the dataset is reordered. The process finishes when no more replacements are needed, because the dataset is k-anonymous.

The most computationally expensive function corresponds to the calculation of the semantic similarity between value pairs, executed p^2 times in the line #7, being p the number of different labels in the attribute ($p \leq n$, being n the total number of attribute values). In the worst case, when the main loop (line #5) ends, this calculation is executed $p^2 \cdot p = p^3$ times. However, as the total set of different values are known a priori and do not change during the masking process (unlike generalization methods), it is possible to pre-calculate and store the similarities between all of them. This avoids repeating similarity measuring calculus for already evaluated value pairs. In this manner, the calculation of the similarity measure is executed a priori only p^2 times. It is important to note that the computational cost of our algorithm uniquely depends on the number of different labels, unlike the related works that depend on the total size of the dataset and on the depth and branching factor of the hierarchy (which represent an exponentially large generalization space).

5 Evaluation

We have evaluated the proposed method by applying it to a dataset consisting on textual answers to the question "*What has been the main reason to visit Delta del Ebre?"* retrieved from polls made by "*Observatori de la Fundació d'Estudis Turístics Costa Daurada*" at the Catalan National Park "*Delta del Ebre*". Examples of common answers are: nature, relaxation, fauna, culture, second residence, etc.

The dataset consists on a set of textual and unbounded answers regarding user preferences expressed by means of a noun phrase (with one or several words). As answers are open, the disclosure risk is high and, therefore, individuals are easily identifiable. The dataset is composed by 975 individual registers, with 221 different responses, being 84 of them unique. Note that this sample represents a much wider and heterogeneous test bed than those reported in related works [12], which are focused on bounded categorical values.

As those answers correspond to general and widely used concepts (i.e. sports, beach, nature, etc.) all of them have been found in WordNet 2.1 (that it is used to calculate semantic similarities), corresponding to one or several synsets. The Porter Stemming Algorithm [13] was used to extract the morphological root of words and to detect semantically equivalent answers. WordNet queries for concepts also implement stemming in order to map each concept label to different lexicalizations.

We evaluated our approach from two points of view. First, we measured the contribution of the designed heuristics in guiding the substitution process towards minimizing the information loss form a semantic point of view (as detailed in section 4). We used Wu and Palmer, Leacock and Chodorow and Path Length measures (see section 3) as quality metrics.

As baseline, we implemented a naïve substitution method that consists on replacing each sensible value by a random one from the same dataset. Following the same basic algorithm presented in section 4, each random change would increase the level of k-anonymity; the process ends when all values are anonymized. Values are ordered alphabetically, in order to avoid depending on the initial order of data. The results obtained for the random substitution are the average of 5 executions.

We compared our heuristic approach against the random substitution for different levels of *k*. To evaluate the quality of the masked dataset from a semantic point of view, we measured how semantically similar the replaced values are, in average, with respect to the original ones. We computed the averaged difference of semantics between original and anonymized sets using the Wu and Palmer's (Fig. 1) and Path Length (Fig. 2) measures.

Analyzing the figures, we can observe that our approach is able to improve the random substitution by a considerable margin. This indicates the usefulness and necessity of a heuristic substitution aimed to minimize the semantic content loss of the original dataset. This is even more noticeable for a high *k* level. Evaluating the semantic distance in function of the desired level of *k*-anonymity, one can observe a linear tendency with a very smooth growth. This is very convenient and shows that our approach performs well regardless the desired level of anonymization. Regarding the different semantic similarity measures, they provide very similar and highly correlated results. This is coherent, as all of them are based on the same ontological features (i.e. absolute path length and/or the taxonomical depth) and, even though

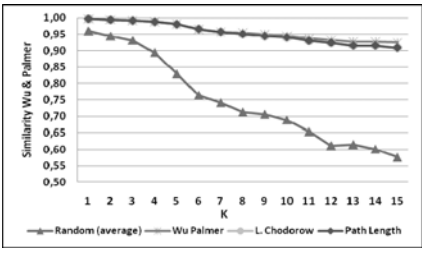


Fig. 1. Semantic similarity of the anonymized dataset

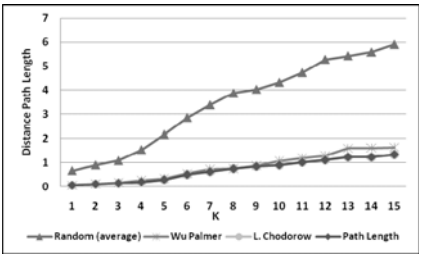


Fig. 2. Distance Path Length of the anonymized dataset

similarity values are different, the relative ranking of words is very similar. In fact, Path length and Leacock and Chorodow measures gave identical results as the later is equivalent to the former but normalized to a constant factor (i.e. the ontology depth).

On the other hand, in order to show the importance of a semantically focused anonymization, we simulated the effect that a more traditional making schema, aimed to optimize the distribution of the masked dataset (as stated at the beginning of section 3), will represent on the resulting dataset. This has been done by using the Discernability metric (eq. 1) in our algorithm instead a semantic similarity measure as a quality metric. Both approaches (semantic, based on Wu and Palmer's measure, and distributional, based on Discernability metric) have been compared by evaluating the semantic loss of the anonymized dataset (for different levels of *k*). Again, this loss is computed as the semantic similarity with respect to the original data by means of the Wu and Palmer's measure (see Fig. 3).

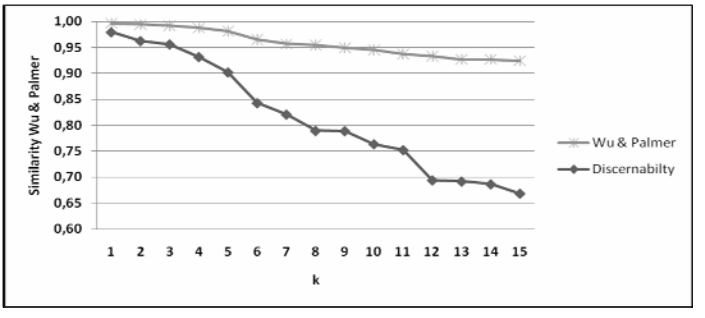


Fig. 3. Semantic similarity for our method with respect to a distributional metric

The figure shows that the optimization of dataset distribution and the preservation of information semantics are not correlated. In fact, there exists a very noticeable semantic loss in the resulting dataset for *k* values above 5. As stated in the introduction, the utility of textual information is highly dependent on its semantics. One can see that classical approaches focused on providing uniform groups of masked values may significantly modify dataset's meaning, hampering their exploitation.

From a temporal perspective, executing our method over a 2.4 GHz Intel Core processor with 4 GB RAM, the runtime of the anonymization process ranged from 0.7 to 1.3 seconds (according to the desired level of *k*-anonymity) as shown in Fig. 4. The pre-calculus of the semantic similarities between all value pairs of the dataset lasted 2.24 minutes. One can easily see how, as stated in section 4, similarity computation represents the most computationally expensive function, and how the minimization of the number of calculus results in a very noticeable optimization of runtime. Runtimes are also much lower than those reported by related works (several hours [12, 19]) based on generalization schemas and very limited VGHs and bounded categorical data (3-4 levels of depth and an average of a dozen of values [12]). This shows the scalability of our method when applied with large and heterogeneous textual data and big and wide ontologies like WordNet.

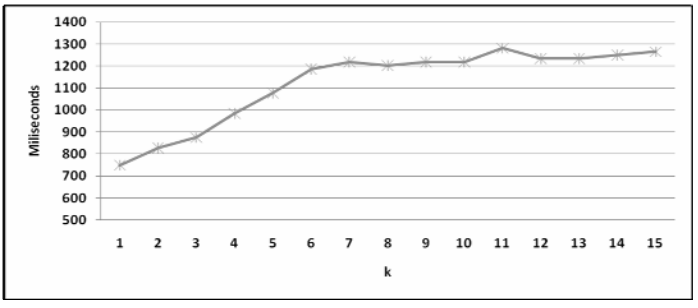


Fig. 4. Anonymization process runtime according to the level of k-anonymity

6 Conclusions

Categorical anonymization aims avoiding disclosure by fulfilling a desired level of *k*-anonymity and, at the same time, maximizing of data utility in order to properly exploit them. Previous approaches neglected or very shallowly considered the semantic content of textual data.

This paper proposes a local masking method for unbounded categorical data based on the exploitation of wide and general ontologies aimed to preserve the semantics of the dataset. Special care has been put in ensuring the scalability of the method when dealing with large and heterogeneous datasets (which are very common when involving text attributes) and big ontologies like WordNet. By enabling the exploitation of those already available ontologies we avoid the necessity of constructing ad-hoc hierarchies according to data labels like VGH-based schemas, which supposes a serious cost and limits the method's applicability.

As future lines of research, we plan to extend our method to global anonymization of complete registers, where different attributes should be masked simultaneously. We will also compare our results with those obtained when using an ad-hoc VGH instead of WordNet [20].

Acknowledgements

Thanks are given to "Observatori de la Funcació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for providing data. This work is supported the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02). Sergio Martínez Lluís is supported by the Universitat Rovira i Virgili predoctoral research grant.

References

1. Bayardo, R.J., Agrawal, R.: Data privacy through optimal *k*-anonymization. In: Proceedings of the 21st International Conference on Data Engineering (ICDE), pp. 217–228 (2005)
2. Cimiano, P.: Ontology Learning and Population from Text. In: Algorithms, Evaluation and Applications. Springer, Heidelberg (2006)

3. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, S.J.: A Search and Metadata Engine for the Semantic Web. In: Proc. 13th ACM Conference on Information and Knowledge Management, pp. 652–659. ACM Press, New York (2004)
4. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms. Adv. in Database Systems, vol. 34, pp. 53–80. Springer, New York (2008)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
6. Guarino, N.: Formal Ontology in Information Systems. In: Guarino, N. (ed.) 1st Int. Conf. on Formal Ontology in Information Systems, pp. 3–15. IOS Press, Trento (1998)
7. He, Y., Naughton, J.F.: Anonymization of Set-Valued Data via Top-Down, Local Generalization. In: 35th Int. Conf. VLDB, Lyon, France, vol. 2, pp. 934–945 (2009)
8. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), pp. 279–288 (2002)
9. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. Int. Conf. on Research in Computational Linguistics, Japan, pp. 19–33 (1997)
10. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum (ed.) WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
11. Lefevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity. In: Proceedings on the 22nd Int. Conf. on Data Engineering, ICDE, p. 25 (2006)
12. Li, T., Li, N.: Towards optimal *k*-anonymization. Data & Knowledge Engineering 65, 22–39 (2008)
13. Porter: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
14. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 9(1), 17–30 (1989)
15. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: *k*--anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
16. Sweeney, L.: *k*-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
17. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. In: Proc. of VLDB (2008)
18. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proc. 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, pp. 133–138 (1994)
19. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Wai-Chee Fu, A.: Utility-Based Anonymization for Privacy Preservation with Less Information Loss. In: ACM SIGKDD Explorations Newsletter, vol. 8 I.2, pp. 21–30 (2006)
20. Martínez, S., Sánchez, D., Valls, A., Batet, M.: The role of ontologies in the anonymization of textual variables. In: Proceedings of the 13th International Conference of the Catalan Association of Artificial Intelligence (2010) (in Press)

Rational Privacy Disclosure in Social Networks

Josep Domingo-Ferrer

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@urv.cat

Abstract. Social networking web sites or social networks for short (SNs) have become an important web service with a broad range of applications. In an SN, a user publishes and shares information and services. We propose a utility function to measure the rational benefit derived by a user from her participation in an SN, in terms of information acquired vs information provided. We show that independently and selfishly maximizing this utility leads users to "free-riding", *i.e.* getting information about other users and offering no information about themselves. This results in SN shutdown (no functionality). We then propose protocols to achieve a correlated equilibrium between users, in which they coordinate their disclosures in view of jointly maximizing their utilities. The proposed protocol can be used to assist an SN user in making rational decisions regarding which of her attributes she reveals to other users.

Keywords: Social networks, Data privacy, Game theory.

1 Introduction

Social networking web sites or social networks for short (SNs) have become an important web service with a broad range of applications: collaborative work, collaborative service rating, resource sharing, friend search, etc. Facebook, MySpace, Xing, LinkedIn, etc., are well-known examples. In an SN, a user publishes and shares information and services.

There are two types of privacy in SNs:

– *Relationship privacy.* In some SNs, the user can specify how much it trusts other users, by assigning them a trust level. It is also possible to establish several types of relationships among users (like "colleague of", "friend of", etc.). The trust level and the relationship type are used to decide whether access is granted to resources and services being offered. The availability of information on relationships (trust level, relationship type) has increased with the advent of the Semantic Web and raises privacy concerns: knowing who is trusted by whom and to what extent discloses a lot about the users thoughts and feelings; in fact, knowing relationships discloses the social network topology and this can allow re-identification of users even if

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 255–265, 2010.

they pretend to stay anonymous [5]. Relationship privacy is about allowing the normal operation of the SN while allowing users to preserve their relationships and trust levels as private as possible (see [2,3]).

– *Content privacy.* This type of privacy applies to *all* SNs and *is* the subject of this paper. The information content a user publishes clearly affects her privacy. Recently, a privacy risk score [4] has been proposed for the user to evaluate the privacy risk caused by the publication of a certain information. Let the information attributes published by the users in an SN be labeled from 1 to n. Then the privacy score risk of user j is

$$PR(j) = \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j,k) \tag{1}$$

where $V(i,j,k)$ is the visibility of user j's value for attribute i to users who are k links away from j and β_{ik} is the privacy sensitivity of attribute i (how embarrassing it is for a user to reveal attribute i to people k links away). The visibility $V(i,j,k) = 1$ if user j makes her attribute j visible to those users k links away from j; it is zero otherwise. An interesting special case is the dichotomous case, in which an attribute is either kept hidden or published for everyone; in the dicotomous case, $V(i,j) = 1$ means that user j publishes her attribute i and $V(i,j) = 0$ means user j keeps her attribute i secret. The dichotomous privacy score is

$$PR_2(j) = \sum_{i=1}^{n} \beta_i V(i,j) \tag{2}$$

Regarding the above privacy risk score, note that the greater it is for a user, the lower is the privacy preservation utility for that user. On the other hand, $PR(j)$ is a monotonically increasing function of the *sensitivity* of the user's attributes and the *visibility* these attributes get. Also, as noted in [4], the sensitivity β_{ik} is monotonically increasing with k, that is, if $k < k'$ then $\beta_{ik} \leq \beta_{ik'}$.

1.1 Contribution and Plan of This Paper

The aim of this paper is to provide protocols to assist a user in an SN in making rational decisions regarding which of her attributes she reveals to other users.

In Section 2, we define the utility a user derives from participating in an SN as the functionality the user gets divided by privacy risk score the user incurs. By functionality, we mean what the user can see about other users in the SN (we do not mean performance or similar issues). In terms of that privacy-functionality utility, we show in Section 3 that, if users independently choose their disclosure strategies, the dominant strategy (and hence the Nash equilibrium) is for SN users to "free-ride", *i.e.* to try to learn as much as possible from other users and disclose nothing about themselves, which leads to shutting down the SN. This zero-utility outcome can be improved for all users if they coordinate their strategies. In Section 4, we propose protocols to assist users in achieving correlated

equilibrium, that is, to help them to jointly maximize their utilities by revealing their attributes to each other in a correlated way. Simulation results are given in Section 5. Finally, conclusions and future research directions are summarized in Section 6.

2 A Privacy-Functionality Score

As mentioned above, our definition of user utility can be roughly summarized as the amount of information the user can see about other users in the SN divided by the amount of information the user shows about herself. This "rational" utility does not probably explain the attitude of the typical Facebook user, who tends to tell her friends a lot about herself, without caring much what she gets in return. Our definition of utility is more adapted to social networks for professionals, like Xing or LinkedIn: in those networks, employers and job applicants tend to disclose their information in a more targeted and cautious way.

We quantify the above idea of the utility a user j derives from participating in an SN by using the following privacy-functionality score

$$PRF(j) = \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j',k) I(j,j',k)}{1+PR(j)}$$

$$= \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j',k) I(j,j',k)}{1+\sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j,k)} \quad (3)$$

where $I(j,j',k)$ is 1 if j and j' are k links away from each other, and it is 0 otherwise.

Note that:

- $PRF(j)$ decreases as the privacy score $PR(j)$ in its denominator increases, that is, as user j discloses more of her privacy.
- $PRF(j)$ increases as its numerator increases; this numerator adds up the components of privacy scores of users $j' \neq j$ due to those users disclosing attribute values to j.

The dichomotous version of the privacy-functionality score is simply:

$$PRF_2(j) = \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,j')}{1+PR(j)}$$

$$= \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,j')}{1+\sum_{i=1}^{n} \beta_i V(i,j)} \quad (4)$$

3 The SN Functionality-Privacy Game with Independent Strategies

If we regard $PRF(j)$ as a game-theoretic utility function [6], the higher $PRF(j)$, the higher the utility for user j. Let us first deal with the dichotomous case, for simplicity.

The set of possible strategies S_j available to user j are the numbers from 0 to $2^n - 1$. In the binary expression of a strategy $s_j \in S_j$, a 1 in position $i \in \{0, \cdots, n-1\}$ means that, under s_j, j publishes attribute $i+1$ ($V(i+1, j) = 1$), whereas a 0 means that, under s_j, j keeps attribute $i+1$ secret ($V(i+1, j) = 0$).

Now, consider a strategy vector $s = (s_1, \cdots, s_N)$ formed by the strategies *independently and selfishly* chosen by all users. When user j chooses s_j, denote by s_{-j} the $N-1$ dimensional vector of the strategies chosen by the other users. If we use $PRF_2(j)$ to quantify the utility $u_j(s)$ incurred by user j, we have

$$u_j(s) = \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \beta_i V(i, j')}{1 + \sum_{i=1}^{n} \beta_i V(i, j)} \tag{5}$$

where the values $V(i, j)$, $i = 1, \cdots, n$, are those specified by the binary expansion of s_j.

It turns out that the strategy vector all zeros, *i.e.* $s^0 = (0, 0, \cdots, 0)$ is dominant, because, for any user j and each alternate strategy vector s', we have

$$u_j(s_j^0, s'_{-j}) \geq u_j(s'_j, s'_{-j}) \tag{6}$$

Disclosing no information is better for user j than disclosing some information, assuming that each user chooses her strategy independently. Since a dominant strategy is also a Nash equilibrium [6], the strategy vector s^0 is also a Nash equilibrium; this can be checked directly, because Inequality (6) implies

$$u_j(s_j^0, s_{-j}^0) \geq u_j(s'_j, s_{-j}^0)$$

Thus, it turns out that rational and independent choice of strategies leads to no user offering any information on the SN, which results in the SN being shut down. It is easy to show that this also holds in the general case ($k > 1$).

A similar pessimistic result is known for the P2P file sharing game, in which the system goal is to leverage the upload bandwidth of the downloading peers: the dominant strategy is for all peers to attempt "free-riding", that is, to refuse to upload [1], which causes the system to shut down.

Example 1. The simplest version of the above game is one with two users having each one attribute, which they may decide to keep hidden (a strategy denoted by H, which implies visibility 0 for the attribute) or publish (a strategy denoted by P, which implies visibility 1). Assuming a sensitivity $\beta = 1$ for that attribute and using Expression (5), the user utilities for each possible strategy vector are as follows:

$$u_1(H, H) = 0; u_1(H, P) = 1; u_1(P, H) = 0; u_1(P, P) = 1/2$$

$$u_2(H, H) = 0; u_2(H, P) = 0; u_2(P, H) = 1; u_1(P, P) = 1/2$$

This simple game can be expressed in matrix form:

User 2 / User 1	H	P
H	0 0	0 1
P	1 0	1/2 1/2

The above matrix corresponds to the Prisoner's Dilemma [6], perhaps the best-known and best-studied game. Consistently with our argument for the general case, it turns out that (H, H) is a dominant strategy, because:

$$u_1(H,P) = 1 \geq u_1(P,P) = 1/2; u_1(H,H) = 0 \geq u_1(P,H) = 0$$

$$u_2(P,H) = 1 \geq u_1(P,P) = 1/2; u_2(H,H) = 0 \geq u_2(H,P) = 0$$

The second and fourth equations above guarantee that (H, H) is a Nash equilibrium (in fact, the only one). The Prisoner's Dilemma with $N > 2$ users is known as the Pollution Game [6] and corresponds to the dichotomous SN game considered above.

4 The SN Functionality-Privacy Game with Correlated Strategies

The outcome of independent rational behavior by users, provided by Nash equilibria and dominant strategies, can be inferior to a centrally designed outcome. This is clearly seen in Example 1: the strategy (P, P) would give more utility than (H, H) to *both* users. However, usually no trusted third-party accepted by all users is available to enforce correlated strategies; in that situation, the problem is how User 1 (resp. User 2) can guess whether User 2 (resp. User 1) will choose P.

Using a solution based on cryptographic protocols for bitwise fair exchange of secrets would be an option, but it seems impractical in current social networks, as it would require a cryptographic infrastructure, unavailable in most SNs.

A more practical solution to this problem may be based on direct reciprocity (*i.e.* tit-for-tat) or reputation, two approaches largely used in the context of P2P file-sharing systems. We describe below two correlated equilibrium protocols based on tit-for-tat and reputation, respectively. They are intended as "assistants" to the human user of the SN in deciding whether to disclose an attribute to another user; however, the ultimate decision belongs to the human, who may quit and renounce to reach the equilibrium. In particular, in both protocols, User 1, as the initiator, first takes the risk of not being corresponded by User 2. However, the "loss" of User 1 will be limited to those attributes she disclosed in the last iteration. User 1 will not disclose to User 2 her remaining, more sensitive attributes.

In both protocols, we introduce user-dependent attribute sensitivities. For example, whereas some people boast their religion (and even dress according to it), for other people this is a very sensitive attribute. Also, when a user evaluates the sensitivity of the attributes received from the other user, the evaluating user is forced to use her own sensitivity scale, because she cannot be assumed to know the evaluated user's sensitivity scale. The real sensitivity scale of an individual is normally highly confidential; for that same reason, if someone discloses her sensitivity scale, there is no guarantee that it is her real scale.

4.1 Adaptation of the Dichotomous Game to Tit-for-Tat

In the protocol below, β_{ij} denotes sensitivity of attribute i according to User j's sensitivity scale. We assume that disclosing an attribute means making it visible to the other user in the protocol (not to all users): therefore, we write $V(i, j, j')$ to denote the visibility of attribute i granted by User j to User j'. Initially, all visibilities are assumed to be zero.

Protocol 1 (Tit-for-tat correlated equilibrium)
User 1 does:

1. *Set* $Quit := 0$.
2. *While* $Quit = 0$ *do:*
   (a) *If User 1 has already disclosed all her attributes* ($V(i, 1, 2) = 1$ *for all* i) *then set* $Quit := 0$.
   (b) *Disclose to User 2 the attribute* i^* *such that*

$$i^* = \arg \min_{i:V(i,1,2)=0} \beta_{i,1}$$

   that is, the least sensitive attribute among those not yet disclosed. Disclosure implies setting $V(i^*, 1, 2) := 1$.
   (c) *Request User 2 to disclose to User 1 the same attribute* i^* *disclosed by User 1 to User 2.*
   (d) *If User 1 does not receive User 2's value for the same attribute* i^*, *then set* $Quit := 1$.

While simple, Protocol 1 has the shortcoming of requiring that the ordering of attributes by sensitivity be the same for User 1 and User 2. Indeed, after User 1 discloses her least sensitive attribute i^*, she expects User 2 to disclose exactly that same attribute i^*. This will only happen if User 2 also considers i^* as her least sensitive undisclosed attribute. In case User 1 does not get i^*, she will consider there is no reciprocity and she will quit the protocol; thus, the protocol lacks robustness. One could change the protocol so that the exchange of attributes is groupwise (several attributes exchanged at a time). However, the issue arises as to which is a reasonable group size: *e.g.* disclosing all attributes in a single iteration is very robust but it is quite risky for User 1, who makes the first move. The reputation of User 2 is a way to decide on the group size. This the idea of the protocol in the next section.

4.2 Adaptation of the Non-dichotomous Game to Reputation

We adapt the non-dichotomous game as follows:

- The parameter $k \in \{1, \cdots, \ell\}$ will be used as an *intimacy level* rather than as a link distance; the greater k, the lower is intimacy. When a User j first interacts with another User j', User j admits User j' in the lowest intimacy level $k \in \ell$ (that of first-time acquaintances). Subsequent interactions may result in User j' being admitted by User j into higher intimacy levels (with smaller k).
- Attribute sensitivities β_{ijk} will now depend on the specific attribute i, the sensitivity scale of User j and the intimacy level k.
- Each User j assigns to each other User j' a reputation $v_{jj'}$ defined as the maximum sensitivity of the attributes User j is willing to show to User j'. Note that reputation is different from intimacy level: a user is probably less intimate with her psychotherapist than with an office colleague, but she surely assigns a greater reputation to her psychotherapist.
- The visibility $V(i, j, j', k)$ denotes whether attribute i is *first* made visible by User j to User j' at intimacy level k. That is, $V(i, j, j', k) = 1$ means that k is the greatest value (the lowest intimacy level) for which attribute i is made visible by User j to User j'; on the other hand, $V(i, j, j', k) = 0$ may mean that either the attribute is not visible to j' at level k or that it is visible and was first made visible by j to j' for some $k' > k$.

In this way, the utility for User j becomes

$$PRF(j) = \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ijk} V(i, j', j, k)}{1 + \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ijk} V(i, j, j', k)} \tag{7}$$

The above definition of visibility ensures that disclosure of attribute i by User j' to User j is counted in $PRF(j)$ only for one intimacy level, the lowest one (that is, the greatest k) at which attribute i is disclosed by User j' to User j.

Note that, in Expression (7), the sensitivities in the numerator are those corresponding to User j, because User j does not know the sensitivity scale of the other users j'.

We next specify a protocol for correlated equilibrium in a game with two users, each of which have values for n attributes numbered from 1 to n, with the sensitivity of user j's attribute i vs level k users being β_{ijk}. Reputation v_{12} in the protocol below is taken in the same range as attribute sensitivities. All visibilities are initially zero.

Protocol 2 (Reputation-based correlated equilibrium)
User 1 does:

1. *If available, use a previous value v_{12} for the reputation of User 2. If none is available, initialize v_{12} with a prior estimate (this guess may be human-assisted, if the human behind User 1 wishes it). Set $k := \ell$ and Quit $:= 0$.*
2. *While $k \geq 1$ and Quit $= 0$ do:*

(a) *Disclose to User 2, that is, set* $V(i,1,2,k) := 1$*, all attributes* i *such that* $V(i,1,2,k) = 0$ *and* $\beta_{i,1,k} \leq v_{12}$*; if there are no disclosable attributes, set Quit* $:= 1$.
(b) *Request User 2 to disclose to User 1 the same attributes disclosed by User 1 to User 2.*
(c) *If User 1 does not receive User 2's values for the same attributes User 1 disclosed, then*
  – *Set Quit* $:= 1$.
  Else
  – *Call* UPDATE(v_{12}).
  – *Set* $k := k - 1$.

Protocol 2 is more robust than Protocol 1 because in the former the users exchange several attributes at a time, not just one, so that some differences in the attribute sensitivity ordering can be tolerated. In Protocol 2, the procedure UPDATE(v_{12}) is used by User 1 to decide whether:

– The reputation v_{12} is kept unaltered in the next iteration;
– The reputation v_{12} is increased to the maximum between v_{12} and the maximum sensitivity of the attributes received from User 2, according to User 1's own sensitivity scale;
– The reputation v_{12} is decreased due to the content of the attributes disclosed by User 2 in the current iteration (*e.g.* if User 2 reveals that she has been in jail, this may be a very sensitive attribute, but probably it will cause her reputation vs User 1 to decrease). In particular, decreasing v_{12} to 0 is a way for User 1 to quit Protocol 2.

Clearly, updating someone's reputation is a procedure that is likely to need the intervention of the human behind User 1, as it involves subjective judgment. Specifically, User 1 (or rather the human behind her) might decide *not* to increase the reputation of User 2 to the the maximum sensitivity of the attributes received from User 2 if User 1 does not wish to correspond to the overtures of User 2. However, even if the reputation stays the same or decreases, User 2 might learn new attributes from User 1 in the next intimacy level $k-1$, because of the monotonicity of the sensitivities, *i.e.* because $\beta_{ijk'} \leq \beta_{ijk}$ for all $k' < k$. If User 1 wants to make sure that User 2 will not learn any further attribute, User 1 should set

$$v_{12} < \min_i \beta_{i,1,k-1}$$

which will cause the protocol to be quit in the next iteration.

In Protocol 2, User 2, when requested at Step 2b for the first time, acts in slave mode but proceeds much like User 1:

– User 2 assigns an initial reputation v_{21} to User 1 (maybe in a human-assisted way).
– User 2 uses the same value of k as User 1 in every iteration (starting with $k = \ell$).
– User 2 updates v_{21} (similarly to the way described above for User 1 vs v_{12}).
– User 2 decides whether she can disclose the attributes i requested by User in Step 2b above (by setting $V(i,2,1,k) := 1$) by checking whether $\beta_{i,2,k} \leq v_{21}$.

5 Simulation Results

In this section, we report on experimental results. We simulated Protocol 2 ($N = 2$ users). We took the number of attributes to be $n = 10$ and the number of intimacy levels to be $\ell = 16$. We did the following 1000 times:

- Generate attribute sensitivities $\beta_{ij\ell}$, for $i = 1, \cdots, 10$ and $j = 1, 2$ by randomly and uniformly drawing from $[0, 1]$.
- For $k = \ell - 1$ down to 1 generate β_{ijk}, for $i = 1, \cdots, 10$ and $j = 1, 2$ by randomly and uniformly drawing from $[0, \beta_{i,j,k+1}]$.
- Run Protocol 2 for the previous attribute sensitivities. Initial reputations are randomly and uniformly drawn from $[0, 1]$. The human-made decision in UPDATE(v) about the other user's reputation was simulated as follows:
  - Leave v unaltered with probability 0.45.
  - Increase v with probability 0.45 to the maximum between v and the maximum sensitivity of the attributes received from the other user (according to the decision-maker's sensitivity scale).
  - Decrease v with probability 0.1 to a value uniformly and randomly drawn from $[0, v]$.

The average number of iterations performed in one run of Protocol 2 was 10.99, that is, the protocol was quit on average after 11 iterations, out of the maximum 16 iterations.

Figure 1 shows the growth of User 1 and User 2's utilities as Protocol 2 progresses. Utilities are measured with Expression (7). It can be seen that utilities start growing for both users (improving the utility for both users is the purpose of correlated equilibrium) and they stabilize after the first four iterations.

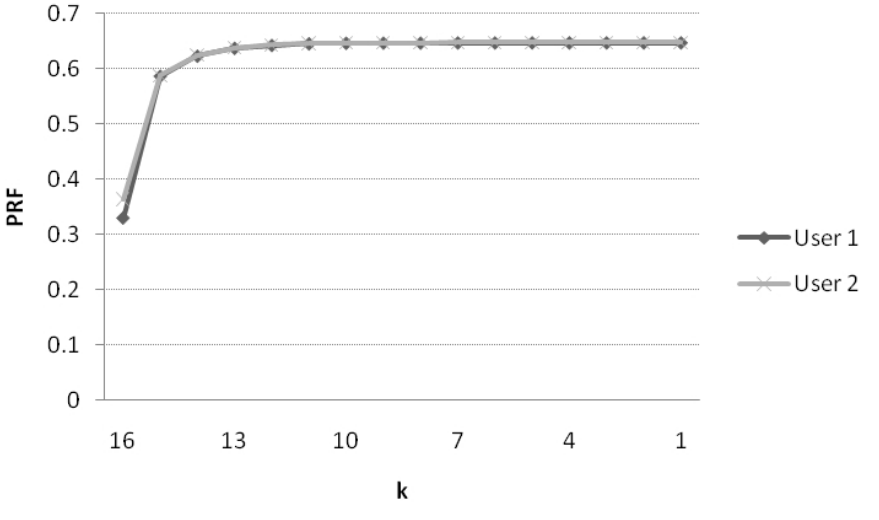


Fig. 1. Evolution of the user utilities measured with Expression (7)

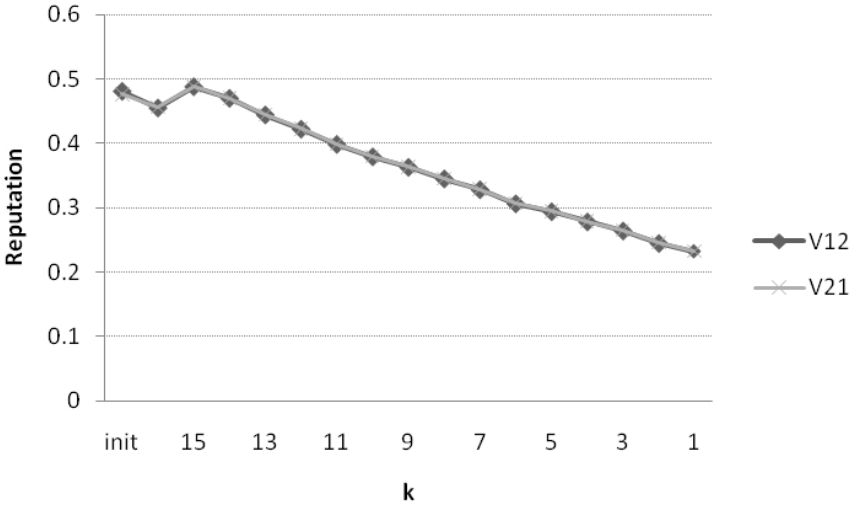


Fig. 2. Evolution of the user reputations

Figure 2 shows the evolution of the reputations User 1 and User 2 assign to each other. Users start with an initial reputation, then this reputation increases because they reveal some highly sensitive attributes; in the next higher intimacy levels, attribute sensitivities are lower because of monotonicity, hence reputation cannot grow due to new high-sensitivity disclosures; it can only decrease, according to the simulated UPDATE(v).

6 Conclusions

We have characterized the utility a user derives from an SN as the information she learns on other users divided by the information she discloses on herself. In terms of this utility, we have shown that, if a user must choose a disclosure strategy without knowing the strategies of other users, her best option is to reveal nothing, which renders the SN useless and provides zero utility to all users. However, better outcomes are possible if users coordinate their disclosure strategies, that is, if they attempt to achieve a correlated equilibrium. We have provided protocols to pursue such an equilibrium by assisting the user in rationally deciding which of her attributes she reveals. Empirical results show that our second protocol results in a utility increase for the two users participating in it.

Future research will include:

– Extending our protocols from pairwise correlated equilibria (taking the users of the SN two by two) to groupwise correlated equilibria (simultaneously correlating strategies of $N > 2$ users);

– Investigating other utility functions which could reasonably model the disclosure attitude of users of SNs for personal contact like Facebook;
– Incorporating concepts such as security and user authentication, which are quite challenging due to the very nature of social networks.

Acknowledgments and Disclaimer

Thanks go to Úrsula González-Nicolás for carrying out the simulations. This work was partly funded 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 through grant 2009 SGR 1135. The author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia. He holds the UNESCO Chair in Data Privacy, but the views expressed in this paper are his own and do not commit UNESCO.

References

1. Babaioff, M., Chuang, J., Feldman, M.: Incentives in peer-to-peer systems. In: Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.) Algorithmic Game Theory, pp. 593–611. Cambridge University Press, Cambridge (2007)
2. Carminati, B., Ferrari, E., Perego, A.: Private relationships in social networks. In: Proceedings of ICDE 2007 Third International Workshop on Privacy Data Management, pp. 163–171. IEEE Computer Society, Los Alamitos (2007)
3. Domingo-Ferrer, J., Viejo, A., Sebé, F., González-Nicolás, Ú.: Privacy homomorphisms for social networks with private relationships. Computer Networks 52, 3007–3016 (2008)
4. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. In: Proc. of ICDM 2009-The 9th IEEE International Conference on Data Mining, pp. 288–297 (2009)
5. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 30th Symposium on Security and Privacy, Oakland, California, pp. 173–187 (2009)
6. Tardos, É., Vazirani, V.V.: Basic solution concepts and computational issues. In: Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.) Algorithmic Game Theory, pp. 3–28. Cambridge University Press, Cambridge (2007)

Towards Semantic Microaggregation of Categorical Data for Confidential Documents

Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra

IIIA, Institut d'Investigació en Intel·ligència Artificial
- CSIC, Consejo Superior de Investigaciones Científicas,
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)
{vtorra,guille,dabril}@iiia.csic.es

Abstract. In the data privacy context, specifically, in statistical disclosure control techniques, microaggregation is a well-known microdata protection method, ensuring the confidentiality of each individual. In this paper, we propose a new approach of microaggregation to deal with semantic sets of categorical data, like text documents. This method relies on the WordNet framework that provides complete semantic relationship taxonomy between words. Therefore, this extension aims ensure the confidentiality of text documents, but at the same time, it should preserve the general meaning. We apply some measures to evaluate the quality of the protection method relying on information loss.

1 Introduction

It is not uncommon to find situations where we need to provide information regarding the contents of a set of confidential documents. Documents such as research project proposals, research papers submitted for publication, confidential law suites, medical records, confidential reports (by companies of law enforcement agencies), etc. cannot be publicly revealed, but being able to provide some information about them might be very useful. The information disclosed should provide enough accuracy to allow generic tasks such as the classification of documents into categories of topics, but preserving their confidentiality.

A simple approach in these situations is to provide, for each document, a vector of keywords or terms. These terms can be manually specified, or most commonly, automatically generated. For instance, by providing the N most frequent terms for each document. The terms, are then used by classification, clustering, or generic information retrieval algorithms. The problem with this approach is to ensure that the vector of terms from the document do not reveal confidential information. A relatively large vector provides more accuracy but at the same time might reveal too much information regarding the contents of the document.

In this paper we propose a novel approach to deal with this problem. Given a set of confidential documents we provide a vector of terms for each document, which ensures that a certain degree of confidentiality is preserved. The degree of confidentiality or privacy is measured in terms of k-anonymity with respect to the whole set of documents. That is, in the resulting set of document vectors, there

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 266–276, 2010.

will be k indistinguishable vectors. To do so we rely in semantic generalizations through the use of a semantic microaggregation approach for categorical data, which makes use of WordNet [5].

2 Plan of the Paper and Preliminaries

Given a set of confidential documents and a vector of terms that represent them, we use a semantic microaggregation approach to ensure k-anonymity among the documents.

To evaluate our approach, we rely on relatively large document vectors (50 and 100 terms) automatically generated from the documents. The documents are a set of papers published in the proceedings of the conference *Modeling Decisions for Artificial Intelligence* (MDAI), in the years 2007 [13], 2008 [14], and 2009 [15], which sum up to ≈ 50. The terms are selected based on their frequency. Note that how the terms are generated or selected does not have much influence in our solution, we have choose this technique because it is the most common approach. In Section 2.1 we overview the generation of document vectors.

Given the document vectors, we apply a semantic microaggregation to produce the protected vectors. Microaggregation is a well known technique used in statistical disclosure control and we extend it here to use semantic information. Section 2.2 reviews microaggregation. Our proposal on semantic microaggregation is introduced in Section 3. Section 4 provides some results from our proposal, and Section 5 concludes the paper.

2.1 Document Vectors

We have a set of m confidential documents D, and each document is represented by a document vector, which contains the most relevant terms of the document. The relevance of the selected terms is determined by their frequency. The documents are automatically parsed and tokenized following [4], then we eliminate common English stop-words, words with less than tree letters, and words which are not in WordNet. The resulting set of terms are used to calculate the document vectors.

By considering only the words included in WordNet we are eliminating some words, which can result in a loss of information. These words are normally common names or very specific terms used in specific research fields. It is important to remark that in this work we are using WordNet as a generic ontology for the English language. When the application domain is known, other domain-specific ontologies can be used such as the UMLS (Unified Medical Language System) [11] for biomedical data.

Each document is represented by a vector d, which contains the N most frequent terms t with their associated frequency weight w. The weight $\omega_{i,j}$ is computed as:

$$\omega_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the term t_i in document d_j, and $\sum_k n_{k,j}$ is the number of occurrences of all the terms in document d_j.

The document vector d_j for the jth document is sorted by frequencies. Formally,

$$d_j = ((t_{\sigma(1),j}, \omega_{\sigma(1),j}), (t_{\sigma(2),j}, \omega_{\sigma(2),j}), \ldots, (t_{\sigma(N),j}, \omega_{\sigma(N),j}) \quad (2)$$

where σ is a permutation such that $\omega_{\sigma(i),j} \geq \omega_{\sigma(i+1),j}$ for all $i = 1, .., N-1$.

The vector of relevant terms based on their frequency provides a good approximation to the contents of the document. Moreover it can easily be used for classification of documents in categories or topics by automated algorithms.

2.2 Microaggregation

Microaggregation is a statistical disclosure control technique, which provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

Privacy is ensured because all clusters have at least a predefined number of elements, and therefore, there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data. Thus, k can be seen as the privacy level provided by the microaggregation.

Microaggregation was originally [1] defined for numerical attributes, but later extended to other domains. E.g., to categorical data in [9] (see also [3]), and in constrained domains in [10].

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least k records.
- **Aggregation.** For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. Let v_i be the representative of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$\begin{array}{ll} \text{Minimize} & SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij} (d(x_j, v_i))^2 \\ \text{Subject to} & \sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\ & 2k \geq \sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\ & u_{ij} \in \{0, 1\} \end{array}$$

Algorithm 1. MDAV

Data: X: original data set, k: integer
Result: X': protected data set

```
1  begin
2    while (|X| ≥ 3 * k) do
3      Compute average record x̄ of all records in X;
4      Consider the most distant record x_r to the average record x̄;
5      Form a cluster around x_r. The cluster contains x_r together with the
       k − 1 closest records to x_r;
6      Remove these records from data set X;
7      Find the most distant record x_s from record x_r;
8      Form a cluster around x_s. The cluster contains x_s together with the
       k − 1 closest records to x_s;
9      Remove these records from data set X;
10   if (|X| >= 2 * k) then
11     Compute the average record x̄ of all records in X;
12     Consider the most distant record x_r to the average record x̄;
13     Form a cluster around x_r. The cluster contains x_r together with the
       k − 1 closest records to x_r;
14     Remove these records from data set X;
15   Form a cluster with the remaining records;
16 end
```

For numerical data it is usual to require that $d(x, v)$ is the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors, and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i is defined as the arithmetic mean of the records in the cluster. I.e., $v_i = \sum_{j=1}^{n} u_{ij} x_i / \sum_{j=1}^{n} u_{ij}$. As the solution of this problem is NP-Hard [6] when we consider more than one variable at a time (multivariate microaggregation), heuristic methods have been developed.

MDAV [2] (Maximum Distance to Average Vector) is one of such existing algorithms. It is explained in detail in Algorithm 1, when applied to a data set X with n records and A attributes. The implementation of MDAV for categorical data is given in [3].

Note that when all variables are considered at once, microaggregation is a way to implement k-anonymity [7,8].

3 Semantic Microaggregation

In order to provide a semantic microaggregation of the document vectors, we use a semantic approach both for the partition of the data into clusters, and for the aggregation of vectors to compute the centroids of the clusters. The partition is based on a semantic distance on vectors, which relies in the Wu-Palmer similarity [12] to compute distances between terms using WordNet. On

the other hand, the aggregation is computed by generalizing terms in a WordNet taxonomy. The following sections provide a description of our approach.

WordNet structures nouns, verbs, adjectives, and verbs, into sets of cognitive synonyms called *synsets* which express concrete concepts. These synsets are interlinked by several conceptual-semantic and lexical relations.

3.1 Term Distance

We rely on the Wu-Palmer measure, which provides a similarity function for two given *synsets*, defined as:

$$sim_{wup}(s_1, s_2) = \frac{2\ depth(lcs(s_1, s_2))}{depth(s_1) + depth(s_2)} \tag{3}$$

where $lcs(s_1, s_2)$ denotes the least common subsumer (most specific ancestor node) of the two synsets s_1 and s_2 in a WordNet taxonomy, and $depth(s)$ is the length of the path from s to the root of the taxonomy. Given that multiple inheritance is allowed in WordNet taxonomies, there might be more than one candidate for $lcs(s_1, s_2)$, in this case the deepest one in the taxonomy is chosen. Note that this similarity ranges from 1 (equal synsets) to 0 (actually never reaches 0, which is only assigned to non-comparable synsets).

We can easily convert sim_{wup} into a distance function as:

$$dst_{wup}(s_1, s_2) = 1 - sim_{wup}(s_1, s_2) \tag{4}$$

Since a term (or word) can belong to more than one synset in WordNet that is, it can have more than one conceptual meaning, we opt to determine the distance between terms as the minimum distance between all their subsets. If we denote as $syns(t)$ the set of synsets that contain the term t, we define our distance dst_t on two terms t_1, t_2 as:

$$dst_t(t_1, t_2) = \min\{dst_{wup}(s_i, s_j) \mid (s_i, s_j) \in syns(t_1) \times syns(t_2)\} \tag{5}$$

As an example, Table 1 show the term distance dst_t between the terms *computer* and *butterfly*. For each term we find their sysnsets and we take the minimum dst_{wup} between each pair of sysnset. Synsets are denotes with a name followed of a letter denoting whether the synset is a noun (n) or verb (v), and an ID. As shown, we have that $dst_t('computer', 'butterfly') = 0.4286$.

Table 1. Example of term distance calculation

butterfly \ computer	⟨computer.n.01⟩	⟨calculator.n.01⟩
⟨butterfly.n.01⟩	0.6190	**0.4286**
⟨butterfly.n.02⟩	0.9048	0.8888
⟨butterfly.v.01⟩	-	-
⟨butterfly.v.02⟩	-	-
⟨chat_up.v.01⟩	-	-

3.2 Document Vector Distance

Once we have defined the semantic distance between terms, we use it to define the semantic distance between two document vectors. In order to do the partition term of the microaggregation process.

We define the document distance as the mean of the minimum semantic distances between the terms of both documents. More precisely, we take the minimum distance between the first term of the first document and all terms of the second one. This process is then repeated for all terms of the first document. Then, the mean of these minimum distances is returned. This distance is defined as:

$$dst_{doc}(d_1, d_2) = \frac{1}{len(d_1)} \sum_{t_i \in d_1} \min_{t_j \in d_2} dst_t(t_i, t_j) \tag{6}$$

where $len(d_1)$ is the length of the first document.

As an example, we consider two simple document vectors d_1 and d_2 with four terms each one with their respectively frequencies.

$$d_1 = (('butterfly', 6), ('performance', 4), ('pen', 2), ('dog', 2))$$
$$d_2 = (('computer', 8), ('cat', 6), ('approach', 4), ('beetle', 2))$$

Table 2 shows all the distances between the terms of the different vectors and emphasizes the minimum ones. Note that the dashes in this table denotes the impossible relations, because when a term of the first document is assigned to other of the second document, this one is removed from the list of possible terms of the second document and then the other distances will be performed with the remaining terms.

Finally, we calculate the mean with these minimum distances to compute the semantic distance between both documents.

$$dst_{doc}(d_1, d_2) = \tfrac{1}{4}(0.130 + 0.250 + 0.143 + 0.333) = 0.214$$

Table 2. Distances between terms of two documents

$t_j \in d_2$ \ $t_i \in d_1$	'butterfly'	'performance'	'pen'	'dog'
'computer'	0.428	0.600	**0.333**	-
'cat'	0.454	0.444	0.333	**0.143**
'approach'	0.238	**0.250**	-	-
'beetle'	**0.130**	-	-	-

3.3 Document Vector Aggregation

The second operation of the microaggregation is the aggregation, which computes a new vector, that represents the cluster representative or centroid. In this case, we need to form this centroid taking into account the semantic meaning of the different elements of the vectors.

The semantic aggregation process for two different document vectors is defined as the aggregation function $\mathbb{C}$:

$$\mathbb{C}(d_1, d_2) = \bigcup_{t_i \in d_1} \{lch(t_i, \arg\min_{t_j \in d_2} dst_t(t_i, t_j)), \alpha(t_i, \arg\min_{t_j \in d_2} dst_t(t_i, t_j))\} \quad (7)$$

where $lch(t_i, t_j)$ (lowest common hypernym) denotes the lowest term in the WordNet hierarchy, which both terms, t_i and t_j, have in common, and $\alpha(t_i, t_j)$ is the mean frequency of both terms. This ensures the preservation of the frequencies in the microaggregated data. Note that the term distance is used again to find the semantically closer relation between the terms of both documents, in order to generalize the meaning of each pair in one term using the function lch.

As we said, this definition of $\mathbb{C}$ only accepts two documents vectors. But, it can be generalized easily. For clusters with more than two elements, the process will iterate aggregating the centroid, obtained with the two first documents with the following vector, and so on for all the vectors of the cluster.

The following example illustrates an aggregation process between two document vectors. We use the same simple vectors d_1 and d_2 used in the previous section. In Table 3 we can see the lowest term in common in the hierarchy by the relation of the two terms that have the minimum distance between them, and also, it shows the mean of both frequencies. The resulting centroid for this cluster is the following:

$$\mathbb{C}(d_1, d_2) = (('insect', 4), ('action', 4), ('instrumentality', 5), ('carnivore', 4))$$

Table 3. Hypernyms between terms with minimum distance of two documents

d_2 \ d_1	('butterfly', 6)	('performance', 4)	('pen', 2)	('dog', 2)
('computer', 8)	-	-	('instrumentality', 5)	-
('cat', 6)	-	-	-	('carnivore', 4)
('approach', 4)	-	('action', 4)	-	-
('beetle', 2)	('insect', 4)	-	-	-

3.4 Ilustrative Example

In order to understand better the semantic microaggreation process explained above, we give a toy example, using an original small dataset integrated by four documents as input of the process.

Table 4 (top) shows firstly the original file, integrated by four documents with three terms and their respective term frequency for each of them. The table also shows the protected output file obtained after the microaggregation process with a k value of 2. As you can see the output file has four documents as the original file, but it only has two different records. The first document centroid represents the set of documents that talk about computers parts, and the second one join

the two original documents that talk about different animals. Therefore, we can say that with the protected file we can deduce the general topics of the documents, but we cannot know the specific topics of the original dataset.

Table 4. Example of semantic microaggregation. Original and its respective protected dataset.

Original Data File
(('keyboard', 0.3), ('laptop', 0.4), ('software', 0.3))
(('horse', 0.7), ('dog', 0.2), ('cat', 0.1))
(('hardware', 0.3), ('screen', 0.3), ('computer', 0.4)))
(('lion', 0.5), ('monkey', 0.3), ('tiger', 0.2))
Protected Data File
(('abstraction', 0.3), ('computer', 0.4), ('instrumentality', 0.3))
(('big_cat', 0.3), ('carnivore', 0.2), ('placental', 0.5))
(('abstraction', 0.3), ('computer', 0.4), ('instrumentality', 0.3))
(('big_cat', 0.3), ('carnivore', 0.2), ('placental', 0.5))

4 Evaluation

In order to evaluate the semantic microaggregation described we have used 50 published papers during the last three years in the *Modeling Decisions for Artificial Intelligence*(MDAI) conference. We have created two different data sets from these 50 documents. One with a set of document vectors with the 50 more frequent terms, and another with the 100 most frequent terms. To simplify, we call them respectively *f50x50* and *f100x50*.

As stated in Section 2.1 we only consider the words included in WordNet, which result in some minor loss of information. In this concrete case we lose some common names (for example from the bibliography of each paper), and some very specific terms. More precisely, if we consider the set of words from *f50x50 f100x50* with words included in WordNet and without them, the average similarity measured by the Jaccard similarity between both sets is 0.769557, and 0.771153 respectively[1]. Again we recall that this works is just an illustrative experiment that could be improved by considering domain-specific ontologies.

Both files have been protected with different values of the parameter k in the range from 2 to 10, and then, compared them with different evaluation measures. We have not computed values of k greater than 10 due to the limited size or the test dataset, and to the fact that as we will see, with $k = 10$ we already have a high degree of information loss.

The first measure, SSE, is the sum of squares to measure homogeneity in clustering and is defined as

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (dst_{doc}(x_{ij}, \bar{x}_i))^2 \tag{8}$$

[1] The Jaccard similarity coefficient measures the similarity between two sets A and B as $\frac{|A \cap B|}{|A \cup B|}$.

Table 5. Evaluation values of both data sets according to k

k	Data Set	SSE	SSA	SST	L
2	f50x50	4.938	30.929	35.867	13.766
	f100x50	4.936	37.119	42.055	11.736
3	f50x50	11.407	21.390	32.797	34.780
	f100x50	12.049	29.733	41.782	28.838
4	f50x50	15.693	21.556	37.249	42.131
	f100x50	16.647	22.759	39.406	42.245
5	f50x50	20.404	11.890	32.294	63.181
	f100x50	21.070	19.157	40.227	52.377
6	f50x50	23.072	17.372	40.444	57.046
	f100x50	24.516	18.336	42.852	57.212
7	f50x50	25.109	11.332	36.441	68.903
	f100x50	26.712	18.981	45.693	58.560
8	f50x50	27.034	8.986	36.0194	75.053
	f100x50	27.662	16.101	43.763	63.209
9	f50x50	28.529	10.085	38.614	73.883
	f100x50	30.107	11.657	41.764	72.088
10	f50x50	31.670	5.680	37.350	84.793
	f100x50	31.455	10.857	42.312	74.341

where g is the number of groups and n_i the number of individuals in the ith group. Naturally, ($n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$). In the same way x_{ij} is the jth record in the ith group and $\bar{x}_i$ denotes the average data vector over the ith group. The lower SSE, the higher the within-group homogeneity.

The SSA measure is a measure to evaluate homogeneity between-groups.

$$SSA = \sum_{i=1}^{g} n_i (dst_{doc}(\bar{x}_i, \bar{x}))^2 \tag{9}$$

where $\bar{x}$ is the average vector over the whole set of n individuals. The higher SSA, the lower the between-groups homogeneity.

The SST measure is the total sum of squares ($SST = SSA + SSE$), or, equivalently,

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (dst_{doc}(x_{ij}, \bar{x}))^2 \tag{10}$$

The last measure is the normalized information loss and is defined as

$$L = \frac{SSE}{SST} \times 100 \tag{11}$$

The optimal k-partition is defined by the one that minimizes the SSE measure (i.e., maximizes the within-group homogeneity) and maximizes the SSA measure

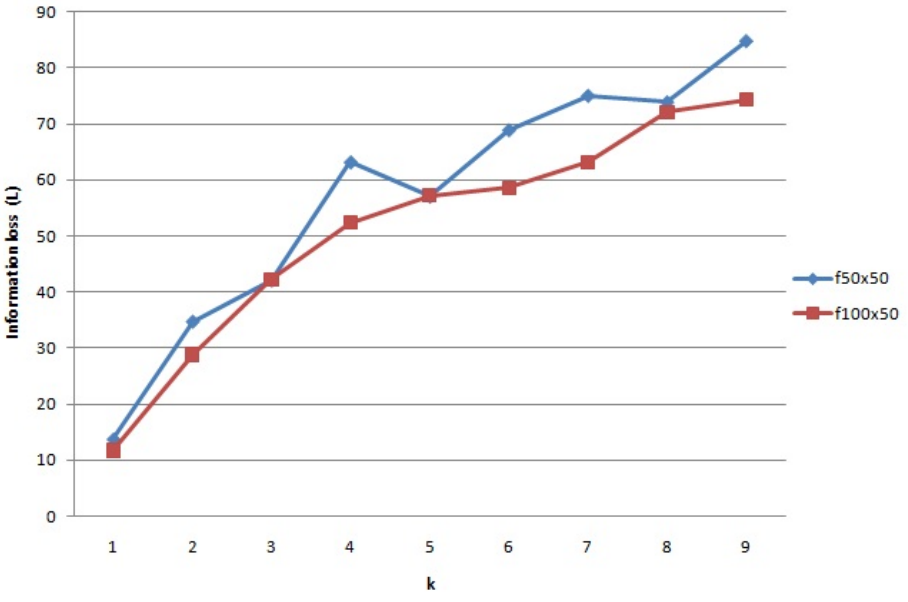


Fig. 1. Plot of information loss (L) vs. privacy level k

(i.e, minimizes the between-group homogeneity). Note that the higher within-group homogeneity, the lower the information loss.

Table 5 shows the evaluation values defining how optimal is the k-partition for each one of these protected files. As expected, the SSE values increases as k increases. It means that within-group homogeneity decreases when the number of documents per cluster increases.

On the contrary, SSA values decrease when k decrease. This is reasonable because when k grows, there are less centroids and homogeneity between clusters decreases.

Finally, we focus on the information loss. As expected, when k increases, the information loss also increases. Moreover, we can appreciate that the dataset with 50 terms, f50x50, results into a higher information loss than the dataset with 100 terms. You can see it clearly in Figure 1.

After the analysis, we can say that the best parameter is the k with values between 3 and 5, because they are the ones with a lower information loss value. At this point, we do not consider 2 as an acceptable value for k, because in this case the protection level is too weak for ensuring data confidentiality.

5 Conclusions

In this paper, we have introduced an extension of microaggregation for text documents, taking into account the semantic meaning of terms. Thanks to WordNet framework, we have modified the partition and the aggregation parts of MDAV algorithm adapting their functionality with semantic relationships. Furthermore, we have presented some results analyzing this semantic microaggregation.

As a future work, we will study information loss extensively when applying different information retrieval tools. We will also consider the use of frequencies of the terms when computing the semantic distances. Moreover, we will research in disclosure risk measures techniques for this microaggregation extension.

Acknowledgments

Partial support by the Spanish MICINN (projects TSI2007-65406-C03-02, ARES- CONSOLIDER INGENIO 2010 CSD2007-00004) is acknowledged. G. Navarro-Arribas enjoys a Juan de la Cierva grant (JCI-2008-3162) from the Spanish MICINN.

References

1. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, pp. 195–204 (1993)
2. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
3. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
4. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Computational Linguistics 32(4), 485–525 (2006)
5. Miller, G.: WordNet - About Us. WordNet. Princeton University, Princeton (2010), http://wordnet.princeton.edu
6. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commision for Europe 18(4), 345–353 (2001)
7. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
8. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
9. Torra, V.: Microaggregation for categorical variables: A median based approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
10. Torra, V.: Constrained microaggregation: Adding constraints for data editing. Transactions on Data Privacy 1(2), 86–104 (2008)
11. U.S. National Library of Medicine, National Institutes of Health, Unified Medical Language System (UMLS) (2010), http://www.nlm.nih.gov/research/umls/
12. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics, Morristown (1994)
13. Torra, V., Narukawa, Y., Yoshida, Y. (eds.): MDAI 2007. LNCS (LNAI), vol. 4617. Springer, Heidelberg (2007)
14. Torra, V., Narukawa, Y. (eds.): MDAI 2008. LNCS (LNAI), vol. 5285. Springer, Heidelberg (2008)
15. Torra, V., Narukawa, Y., Inuiguchi, M. (eds.): MDAI 2009. LNCS, vol. 5861. Springer, Heidelberg (2009)

Using Classification Methods to Evaluate Attribute Disclosure Risk

Jordi Nin[1], Javier Herranz[2], and Vicenç Torra[3]

[1] CNRS ; LAAS ; 7 avenue du Colonel Roche, 31077 Toulouse Cedex 4, France
[2] Dept. Matemàtica Aplicada IV, Universitat Politècnica de Catalunya,
C. Jordi Girona 1-3, Mòdul C-3, 08034 Barcelona (Spain)
[3] IIIA, Artificial Intelligence Research Institute, CSIC, Spanish National Research Council
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)
jnin@laas.fr,jherranz@ma4.upc.edu,vtorra@iiia.csic.es

Abstract. Statistical Disclosure Control protection methods perturb the non-confidential attributes of an original dataset and publish the perturbed results along with the values of confidential attributes. Traditionally, such a method is considered to achieve a good privacy level if attackers who try to link an original record with its perturbed counterpart have a low success probability. Another opinion is lately gaining popularity: the protection methods should resist not only record re-identification attacks, but also attacks that try to guess the true value of some confidential attribute of some original record(s). This is known as attribute disclosure risk.

In this paper we propose a quite simple strategy to estimate the attribute disclosure risk suffered by a protection method: using a classifier, constructed from the protected (public) dataset, to predict the attribute values of some original record. After defining this approach in detail, we describe some experiments that show the power and danger of the approach: very popular protection methods suffer from very high attribute disclosure risk values.

Keywords: Attribute Disclosure Control, Classification, Privacy-Preserving Data Perturbation.

1 Introduction

There are many real situations where confidential individual data is published by statistical agencies, to be used by decision makers, politicians, researchers, etc. This dissemination should ensure, however, that the privacy of individuals is protected in some way, to be in accordance with current laws and regulations. A very popular way to achieve a certain level of privacy in this scenario is the application of *perturbative protection methods* to the data, before making it public. The research community that studies such protection methods is known as *Statistical Disclosure Control* (SDC) [1]. The general SDC scenario is as follows. An entity has a dataset X containing some confidential information. The entity releases a protected version X' of the dataset to the public, so that external parties can use this data for analysis. Besides protecting the privacy of the individuals, the main goal is that the protection method preserves as much as possible

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 277–286, 2010.

the statistical utility of the original data. Of course, the values of privacy and statistical utility are inversely related.

There are several ways to measure this trade-off. Maybe the most simple and intuitive one is the *score* [4] metric. It just measures the average between two quantities: one of them analyzes the information loss produced by the application of the protection method, and the other one evaluates the risk that an intruder can obtain any information that breaks the privacy of the individual, after the protected dataset has been released.

Information loss (IL) measures the statistical utility of the protected dataset, comparing its usefulness with respect to the one of the original one. Different approaches are used to calculate the information loss. Originally, in [4] the authors calculate the average divergence of some statistical values when they are computed on both the original and the protected datasets. A probabilistic variation of these measures (PIL) was presented in [12] to ensure that the information loss value is always within the interval [0,1].

In the computation of the risk component of the score (called *disclosure risk*), one of the considered values is the risk of re-identification (also known as *entity disclosure risk*): an intruder wants to link an original record with the corresponding protected record in the released dataset, using some record linkage protocol [19]. The entity disclosure risk is then the percentage of correct links that are found by the record linkage protocol. For example, in any perturbation method that ensures k-*anonymity* [18], the entity disclosure risk is upper bounded by $1/k$, where k is the minimum number of indistinguishable records in the protected dataset.

Recently, many researches have also considered the problem of estimating the *attribute disclosure risk* for confidential attributes [11,20]. Traditionally, perturbative protection methods are only applied to the non-confidential attributes of the datasets, whereas the original confidential attributes remain unchanged. Maybe an intruder is not able to successfully link an original record with its protected record (for example due to k-anonymity). But he may still be able to predict some values of some confidential attributes for this original record with very high probability. This probability is what we call the *attribute disclosure risk*. Perturbation methods which suffer from high attribute disclosure risk should be considered as bad ones, even if their entity disclosure risk values are low. Up to now, most of the works analyzing the attribute disclosure risk are focused on anonymization methods ensuring k-anonymity. This has led to the definition of concepts such as p-sensitive k-anonymity [20] and l-diversity k-anonymity [11].

In this paper, we propose a way to estimate the attribute disclosure risk of any SDC perturbation method (enjoying k-anonymity or not), by means of classification techniques. Roughly speaking, the idea is to use the protected dataset X' as the training dataset of a classifier having as output (class label) a confidential attribute at. The original dataset X is then used as the testing dataset. The percentage of original records that are correctly classified (in other words, the *accuracy* of the classifier) will be considered as an estimation (more specifically, a lower bound) of the attribute disclosure risk, for the corresponding SDC perturbation method and attribute at. After detailing this new approach to estimate the attribute disclosure risk, we will describe a set of experiments that we have run, considering different classifiers and SDC perturbation methods. Our

experiments show that the attribute disclosure risk is sometimes very high, even for configurations where entity disclosure risk is close to negligible.

The rest of the paper is organized as follows. Firstly, in Section 2 we introduce some basic concepts about classification techniques and anonymization methods. Then, in Section 3 we detail our approach of using classification methods to estimate the attribute disclosure risk of a SDC perturbation protocol. The experiments that we have performed are described in Section 4, where we also highlight some of the most relevant obtained results. Finally, Section 5 concludes the paper and contains some lines for possible future work.

2 Preliminaries

In this section we briefly describe the basic concepts on classification and SDC anonymization methods needed to understand the rest of the paper.

2.1 Classification Introduction

The task of a classifier is to learn from specific examples of instances, each one represented by a set of attribute values and *labeled* by class values, a general mapping from the attribute space to classes that allows to classify or predict the class values for future instances. The performance of a classifier is measured as its ability to produce correct labels on unseen data. Since one cannot measure the performance of a classifier on future data, this performance is evaluated by studying the behavior of the classifier on a testing labeled dataset. The most simple and common measure for classifier performance is the percentage of records in the testing dataset that are correctly classified. This percentage is called the *accuracy* (ACC) of the classifier.

While many classifiers have been defined in the literature, none is universally better than the others in terms of their predictive power. The choice of the classifier depends therefore on the characteristics of the data and on the requirements of the classification task and the model built (computational cost, stability, interpretability etc.) The main classifiers used in data mining practice are Decision Trees, Naive Bayes, k-Nearest Neighbor (k-NN), and the Support Vector Machine (SVM).

The Decision Tree classifier [17] builds tests of single attribute values that lead to subsets of instances with a highly predictable class label. Decision trees are highly popular due to their interpretability.

The SVM classifier [21] lifts the classification task from its original data space to a much more high-dimensional feature space, and then learns a linear classifier in that space using the so called *kernel trick* that performs the computation in the data space. SVM often produces higher performance than other classifiers, but suffers from lack of interpretability.

The Naive Bayes classifier [5] predicts a class by combining, in a simple manner, prior probabilities of a class value as determined by values of each individual attribute. Naive Bayes is highly efficient to learn and to apply.

The k-nearest neighbor classifier [2] determines the class of an instance by choosing the most common class of its k closest neighbors. The method is often chosen due to understandability of its underlying principle by the users.

2.2 Anonymization Methods

In our experiments we consider different SDC perturbative methods. They are described below using the following generic notation: X is the original dataset, with n rows (or records) and m columns (or attributes). Therefore, x_{ij} represents the value of record i for attribute j.

Additive and Multiplicative Noise. are perhaps the simplest and most intuitive data perturbation methods. In additive noise [6], each value x_{ij} of the original dataset X is replaced with $x'_{ij} = x_{ij} + \epsilon$, where ϵ is the noise. The simplest approach is that ϵ is a normally distributed error drawn from a random variable $\epsilon \sim N(0, \sigma^2_\epsilon)$, and that the variance of ϵ is proportional to that of the original attributes. In multiplicative noise [10,7] each original value x_{ij} is replaced with $x'_{ij} = x_{ij} \cdot \epsilon$, where the noise ϵ follows a specific distribution which depends on the original values for attribute j.

Rank Swapping. [3] with parameter p and with respect to an attribute j can be defined as follows. Firstly, the records of X are sorted in increasing order of the values x_{ij} of the attribute j. To simplify notation, let us assume that the records are already sorted, that is $x_{ij} \leq x_{\ell j}$ for all $1 \leq i < \ell \leq n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen from the limited range $i < \ell \leq i + p$. When rank swapping is applied to a dataset, the algorithm explained above is run for each attribute to be protected, in a sequential way. The parameter p is used to control the swap range. Normally, p is defined as a percentage of the total number of records in X. Therefore, when p increases, the difference between x_{ij} and $x_{\ell j}$ may increase accordingly. This fact increases privacy, but of course the differences between the original and the protected dataset are higher, thereby decreasing the statistical utility of the data. As noted in [15], the fact that each value is swapped with a value in a fixed, closed rank makes this basic rank swapping method more prone to re-identification attacks, decreasing privacy protection offered by this method. To mitigate this drawback, a variant of rank swapping is proposed in [15], where some values (with a small but still non-negligible probability) are swapped with values out of the theoretical rank. In this paper we use rank swapping p-distribution: It defines the swap interval using a normal probability distribution defined by $\mu = \sigma = 0.5 \cdot p$.

Microaggregation. is one of the most common methods used to obtain k-anonymity for numerical data: groups of k close records are identified and substituted by their centroid. In this way, an original record is protected against disclosure risk in the sense that k protected records have exactly the same probability to correspond to that original record. To achieve minimum information loss, the goal is to find an optimal microaggregation that minimizes the sum of distances between original records and protected records (centroids). Since the optimal solution to this problem is NP-hard [16] (for the general multivariate case), many effective heuristic algorithms have been proposed to provide good quality results.

Among these methods, we can list the Centroid-based fixed-size (CBFS) algorithm [8]. It works as follows. Firstly, the average record $\bar{x}$ of all records in X is computed. The most distant record x_r to the average record $\bar{x}$ is considered, and a cluster

around x_r is formed, containing x_r together with the $k-1$ closest records to x_r. All records belonging to this cluster are removed from X. Among the remaining records, the most distant record to $\bar{x}$ is considered, a cluster is formed, etc. The process is repeated until all the records are assigned to one cluster. Finally, the protected dataset X' is built by replacing each original record in X with the centroid of the cluster to which the record belongs.

In the last years, some researchers have pointed out that k-anonymity may not be enough to ensure privacy. The notions of p-sensitivity [20], l-diversity [11] and t-closeness [9] have been proposed to address this weakness of k-anonymity. The goal is to ensure that the distribution of the confidential values in each of the final clusters satisfies some properties (a minimum number of different values, a minimum entropy value, a distribution very close to the distribution of the confidential values in the entire original dataset X, etc.). In our experiments, we have tuned CBFS in several ways to provide either p-sensitivity or l-diversity. These modifications, of course, lead to a decrease of the statistical utility, with respect to standard CBFS.

3 Estimating Attribute Disclosure Risk through Classification

Attributes in a dataset $X = X_{nc}||X_c$ can be divided into non-confidential and confidential attributes, depending on the kind of information they contain. Since the most interesting statistical information is usually contained in confidential attributes, X_c a typical approach when implementing a SDC perturbation method in practice is to keep these attributes unchanged and to apply the perturbation method ρ only to non-confidential attributes: $X'_{nc} = \rho(X_{nc})$. Therefore, the protected dataset that is released to the public is $X' = X'_{nc}||X_c$.

Once a protected dataset X' is published, different kinds of attacks can be mounted by intruders. In the most extreme case, an intruder is assumed to know all the original non-confidential attributes corresponding to some record $x \in X$, for example if he has obtained this information from another dataset. The goal of the intruder is to obtain the values of the confidential attributes for this record. A way of obtaining this information is by correctly linking x with the corresponding protected record $x' \in X'$. The success probability of this re-identification attacks are what we call entity disclosure risk.

As we have pointed out in the Introduction, a SDC protection method should not be considered secure only because it leads to low values for the entity disclosure risk. Maybe it is difficult for the intruder to link an original record with its corresponding protected record, but this attacker has a way to predict (some) values of the confidential attributes of the original records with high probability. Given a perturbed dataset X', a confidential attribute at^* of the original dataset X, and the value of all the non-confidential attributes of some original record $x \in X$, we define the *attribute disclosure risk* for at^* as the probability that an intruder obtains the correct value of attribute at^* for record x. We will denote this value as $\mathrm{ADR}(\mathsf{at}^*, x)$.

The most naive approach that an intruder can follow to obtain the correct value of attribute at^* for record x is to look for the most dominant value of attribute at^* in X', and to claim that this will be the value for record x. In other words, the percentage of records that have the dominant value in attribute at^* is already a lower bound for

$\text{ADR}(\mathsf{at}^*, x)$. The goal when designing a good re-identification method is to obtain more realistic estimations for the attribute disclosure risk suffered by the perturbation.

3.1 The Proposed Approach

For example, a less naive intruder could take the original record $x \in X$, look for the protected record $x'_c \in X'$ which is the closest to x and output the value of attribute at^* for x' as the candidate. Maybe the record x'_c is not the protected record corresponding to x, but even in this case there are some chances that the value of the attribute at^* is the same. More generally, an intruder could look for the k protected records in X' which are the closest to x, and take the most dominant value among the values of attribute at^* for these records as the candidate for the value of attribute at^* for x. Note that this last strategy is very related to what the k-nearest neighbor classifier does. Our idea is to generalize this approach even further, by considering any possible classifier. The resulting strategy that an intruder could follow to obtain the value of attribute at^* for x is as follows.

1. Taking the dataset X' as input and the attribute at^* as the label class, construct a classifier $\mathcal{C}$ for this class.
2. Apply the classifier $\mathcal{C}$ to the (future) instance x to predict the class value for x.
3. Output the obtained prediction as the candidate for the value of attribute at^* for x.

To estimate the attribute disclosure risk of attribute at^* offered by a perturbation method, the implementer (who knows the entire original dataset X) can run this routine for each original record $x \in X$ and count the percentage of records for which the true value of attribute at^* is found with this strategy.

A good way to estimate the ADR is to follow the standard procedure of k-fold cross validation, adapted to the situation that we are considering: the original dataset X is randomly partitioned into k subsamples, a single subsample is retained as the validation data (non-protected) for testing the ADR, and the remaining $k - 1$ subsamples are protected. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation.10-fold cross-validation is commonly used.

We define the *attribute disclosure risk induced by classifier $\mathcal{C}$ on attribute at^**, $\text{ADR}_{\mathcal{C}}(\mathsf{at}^*)$, as the probability that the strategy described above leads to a correct guess of the value of at^* for a given original record.

4 Experimental Analysis

In this section we explain the experiments we have carried out to test the values of the attribute disclosure risk that can be estimated using our approach.

4.1 Description of Datasets, Perturbation Methods and Classifiers

Regarding the datasets X, we have selected two datasets from the UCI repository [14] and one dataset extracted from the U. S. Census Bureau [13] using the Data Extraction

System (DES). These three data sets have the following properties: (i) the attributes are numerical; (ii) there is an attribute with a few (and quite uniformly distributed) possible values, which will be chosen as the class confidential attribute, at^*. The description of these datasets can be found in Table 1. For the class confidential attribute at^*, we have noted the percentage of records in the dataset that belong to each class; in particular, let us recall that the percentage of the dominant value already gives the naive lower bound for ADR(at^*).

Table 1. Datasets description

	Abalone	Vehicle	Census
Records	4177	846	13518
Attributes	9	19	13
Classes	M(36%)	opel(25%)	L (20%)
	F(31%)	saab(25%)	VL (20%)
	I(33%)	bus(25%)	M (20%)
		van(24%)	H (20%)
			VH (20%)

Finally, we have considered the SDC protection methods described in Section 2.2. We have applied many different parameterizations for each of the tested SDC methods, from few protection to a lot of protection. For additive noise addition, we have used the following values for the variance modification, $\alpha \in \{1, 2, 3, 5, 10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 750, 1000\}$. For multiplicative noise addition, we have used $\alpha \in \{2.5, 5, 7.5, 10, 12.5, 15, 20, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$. For rank swapping p-distribution, we have considered the values $p \in \{1, 3, 5, 12, 25, 35, 50\}$. Recall that p determines the average of the normal distribution which defines the length of the interval where swapping is done. For k-microaggregation, we have used CBFS with different values for the parameter $k \in \{5, 10, 15, 20, 25, 50, 75, 100, 125, 150\}$. Regarding our versions of CBFS which ensure p-sensitivity or l-diversity, we have used the same ten values for parameter k, and then values $p \in \{2, 3\}$ and $l \in \{2, 3\}$ for the parameters related to sensitivity and diversity of the chosen class attribute. Note that the values of p, l make sense only when they are smaller than the number of possible values for the class attribute (which is 3,4 or 5 in the considered datasets).

Regarding the classifiers, we have used the implementations available in WEKA [22], for the four classifiers: Decision Trees (DT), Naive Bayes (NB), k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM), which are the most popular and effective classifiers used in everyday data mining practice. The classifier method k-NN depends on a parameter k. For each dataset, we have done one experiment with k-NN for different values of k, between 2 and 20, without protecting the training dataset this time, and we have selected the value of k which gives the best results.

4.2 Presentation of the Results

Table 2 contains the results obtained with some of the considered parameterizations of the SDC perturbation methods. We have included in these tables three different but

significative parameterizations for each SDC method, reflecting weak protection, medium protection and high protection levels. For each parameterization, we first include for completeness the probabilistic information loss (PIL), so to provide a measure on the data utility offered by each method. Then, we give the entity disclosure risk (EDR) value obtained through a standard distance-based record linkage approach: for each original record x, the protected record $x'_c \in X'$ which is the closest to x is output as the candidate for being the protected version of x. The EDR is the percentage of original records which are correctly linked by following this approach. Finally, we have followed the strategy described in Section 3.1 to compute the attribute disclosure risk $\text{ADR}_{\mathcal{C}}(\mathsf{at}^*)$ induced by classifier $\mathcal{C}$ on attribute at^*, for the classifiers $\mathcal{C}$ =DT,NB,SVM,kNN. For simplicity, we have denoted the maximum of these values $\text{ADR}_{\mathcal{C}}(\mathsf{at}^*)$ simply as ADR, and we have included in the table only this maximum value, which is the most accurate lower bound for the real attribute disclosure risk, in each case.

Table 2. Results obtained with the three datasets and different SDC perturbation methods

	Abalone			Census			Vehicle		
	PIL	EDR	ADR	PIL	EDR	ADR	PIL	EDR	ADR
Noise, $\alpha = 3$	7.90%	64.8%	54.49%	10.65%	79.34%	92.02%	21.16%	90.65%	73.51%
Noise, $\alpha = 10$	24.65%	24.53%	54.37%	27.47%	41.17%	90.17%	30.82%	81.89%	73.16%
Noise, $\alpha = 100$	73.94%	0.00%	53.20%	77.59%	0.12%	73.53%	80.15%	2.58%	63.82%
MultNoise, $\alpha = 5$	13.50%	68.20%	54.44%	14.85%	81.62%	91.77%	23.55%	89.41%	73.00%
MultNoise, $\alpha = 10$	24.81%	24.75%	54.32%	28.00%	41.30%	89.97%	30.62%	82.07%	73.00%
MultNoise, $\alpha = 100$	74.29%	0.00%	53.27%	77.18%	0.10%	73.95%	80.26%	2.24%	62.00%
RS p-dist, $p = 2$	22.12%	38.45%	54.37%	27.82%	66.58%	91.85%	21.14%	79.93%	73.29%
RS p-dist, $p = 10$	29.00%	0.00%	54.35%	38.68%	0.00%	83.20%	34.44%	3.21%	66.77%
RS p-dist, $p = 50$	39.96%	0.00%	53.20%	40.35%	0.00%	50.61%	47.49%	0.00%	47.52%
CBFS, $k = 5$	39.05%	4.61%	54.56%	44.01%	8.38%	91.11%	53.25%	8.31%	71.75%
CBFS, $k = 25$	58.08%	0.57%	54.01%	59.48%	1.10%	90.01%	69.82%	1.18%	64.43%
CBFS, $k = 100$	63.55%	0.03%	54.10%	67.14%	0.00%	86.23%	76.39%	0.00%	42.44%
CBFS 2-sen, $k = 25$	58.08%	0.55%	54.13%	77.99%	0.85%	89.69%	70.02%	1.14%	65.26%
CBFS 3-sen, $k = 25$	73.00%	0.00%	45.00%	86.58%	0.29%	77.92%	70.48%	1.12%	64.43%
CBFS 2-div, $k = 25$	61.55%	0.40%	54.37%	88.00%	0.10%	76.62%	70.83%	1.07%	63.48%
CBFS 3-div, $k = 25$	86.00%	0.00%	40.00%	92.73%	0.00%	53.58%	74.14%	1.01%	58.53%

4.3 Discussion of the Results

On the one hand, observing Table 2, we can see that the information loss (PIL) value obviously increases with the protection parameter; on the other hand, the entity disclosure risk (EDR) value always decreases when the protection is higher. Regarding the values of the attribute disclosure risk (ADR), we notice that they remain quite high, even when the EDR values significantly decrease. See for example the Census dataset protected using additive noise with $\alpha = 100$. In this configuration the EDR value is equal to 0.12% (almost no disclosure risk) while the ADR value is equal to 73.53%. This means that one intruder is able to predict (infer) the confidential information of one individual with a probability around 75%. Note that the naive approach of selecting the majority class in this Census dataset would lead to an attribute disclosure risk value around 20%.

In addition, one could have a clear expectation that ADR values should decrease as the level of protection applied to a dataset increases. But this is not the case with many of the considered combinations of dataset / protection method. See for instance the results with the Abalone dataset protected using additive (or multiplicative) noise, rank swapping or standard CBFS. In such combinations the ADR values are more or less constantly over 50%, whilst the random expected value (majority class) would be around 36%. Then in this case, it is clear the intruder is able to discover some knowledge about the data owners independently of the method (and configuration) used to protect the dataset. The same fact applies in the CBFS 2-sensitive and CBFS 2-diversity configurations.

Only when we consider 3-sensitivity or 3-diversity (the maximum possible) we reach ADR values close to the one obtained with the naive majority approach. The main drawback of these two configurations is their high PIL value, *i*.e. their statistical utility is quite low. In the case of datasets having more than three possible confidential values, the ADR values obtained by our approach for these two configurations are still very high. See for instance the values obtained with the Census dataset (there ares five possible confidential values for the class at^*), with the CBFS 3-diversity method. In this case, the ADR is equal to 58.53%, almost three times higher than value that would be obtained through the majority class strategy.

Summing up, the obtained (and sometimes startling) results illustrate that the SDC protection methods we test should be revisited in order to be less prone to attribute re-identification attacks.

5 Conclusions

We have proposed in this work the use of classifiers on the protected released dataset to measure the attribute disclosure risk (ADR) suffered by SDC perturbation methods. We have provided a set of formal definitions for the resulting estimations of the attribute disclosure risk, as well as a set of experiments showing that this approach leads to quite high values of the ADR for many popular SDC protection methods.

As future work we would like to work on a generic post-processing technique, which can be applied to any protected dataset X' with the goal of reducing the attribute disclosure risk suffered by X'.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged. Javier Herranz enjoys a *Ramón y Cajal* grant, partially funded by the European Social Fund (ESF), from Spanish MICINN Ministry. Jordi Nin is partially supported by the European Community through the 7th Framework Programme Marie Curie Intra-European fellowship, contract No 235226.

References

1. Adam, N.R., Worthmann, J.C.: Security-control methods for statistical databases: a comparative study. ACM Computing Surveys 21(4), 515–556 (1989)
2. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)
3. Dalenius, T., Reiss, S.: Data-swapping: a technique for disclosure control. Journal of Statistical Planning and Inference 6, 73–85 (1982)
4. Domingo-Ferrer, J., Torra, V.: Disclosure control methods and information loss for microdata. In: Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 91–110 (2001)
5. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29(2), 103–130 (1997)
6. Kim, J.: A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the ASA Section on Survey Research Methodology, pp. 303–308 (1986)
7. Kim, J., Winkler, W.E.: Multiplicative noise for masking continuous data. Research report series (statistics 2003-01), U. S. Bureau of the Census (2003)
8. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering 17(7), 902–911 (2005)
9. Li, N., Li, T.: t-closeness: Privacy beyond k-anonymity and -diversity. In: Proc. of IEEE Int. Conf. on Data Engineering (2007)
10. Liu, K., Kargupta, H., Ryan, J.: Random projection based multiplicative data perturbation for privacy preserving data mining. IEEE Transactions on Knowledge and Data Engineering 18(1), 92–106 (2006)
11. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: IEEE Int. Conf. on Data Engineering (2006)
12. Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F.: Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery 11(2), 181–193 (2005)
13. U.S. Census Bureau. Data extraction system (2009), http://www.census.gov/
14. Murphy, P., Aha, D.: UCI Repository machine learning databases. University of California, Department of Information and Computer Science, Irvine (1994)
15. Nin, J., Herranz, J., Torra, V.: Rethinking rank swapping to decrease disclosure risk. Data and Knowledge Engineering 64(1), 346–364 (2008)
16. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal United Nations Economic Commission for Europe 18(4), 345–354 (2000)
17. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
18. Samatari, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI Intl. Tech. Rep. (1998)
19. Torra, V., Nin, J.: Record linkage for database integration using fuzzy integrals. Int. Journal of Intelligent Systems (IJIS) 23(6), 715–734 (2008)
20. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: IEEE Int. Conf. on Data Engineering Workshops (2006)
21. Vapnik, V.: The support vector method. In: Int. Conference on Artificial Neural Networks, pp. 263–271 (1997)
22. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)

A Misleading Attack against Semi-supervised Learning for Intrusion Detection

Fangzhou Zhu, Jun Long, Wentao Zhao, and Zhiping Cai

Huazhong University of Science and Technology, Wuhan, 430074, China
National University of Defense Technology, Changsha, Hunan 410073, China

Abstract. Machine learning has became a popular method for intrusion detection due to self-adaption for changing situation. Limited to lack of high quality labeled instances, some researchers focused on semi-supervised learning to utilize unlabeled instances enhancing classification. But involving the unlabeled instances into learning process also introduces vulnerability: attackers can generate fake unlabeled instances to mislead the final classifier so that a few intrusions can not be detected. We show how attackers can influence the semi-supervised classifier by constructing unlabeled instances in this paper. And a possible defence method which based on active learning is proposed. Experiments show that the misleading attack can reduce the accuracy of the semi-supervised learning method and the presented defense method against the misleading attack can obtain higher accuracy than the original semi-supervised learner under the proposed attack.

Keywords: semi-supervised learning, intrusion detection, active learning.

1 Introduction

Recently, machine learning has been applied to many real world problems. Especially, intrusion detection, which monitors network packets to detect whether malicious behavior happens, begins to utilize machine learning techniques [1]. However, as high quality history data requires heavy labor of experts or expensive monitoring process, it is hard to collect a large number of labeled instances for training. Thus, some researchers focus on using semi-supervised learning methods to aid classification by unlabeled instances, which is easier to collect than labeled instances [2,3].

Unfortunately, attackers may actively disturb the learning process to mislead the intrusion detection system using learning methods. For example, Newsome et al. [4] found a correlated outlier attack against a Bayes-based learning method. The attacker could add some features, which can be found in normal instances, to malicious instances. Thus the classifier trained on such dataset tends to misclassify normal instances to malicious instances.

Currently, such attacks against learning methods are mainly aiming at supervised learning methods. When semi-supervised learning methods are introduced

V. Torra, Y. Narukawa, and M. Daumas (Eds.): MDAI 2010, LNAI 6408, pp. 287–298, 2010.

to intrusion detection systems, such attacks can not work because the great amount of unlabeled instances could provide distribution information to correct the misled classifier.

Nevertheless, since the unlabeled instances can be easily collected, the attackers are more likely to pollute them. We present a novel attack method to defeat the semi-supervised learning for intrusion detection and propose a defense technique in this paper. This work is to remind researchers that semi-supervised learning may be very dangerous because the unlabeled instances can be easily polluted by attackers.

The rest of the paper can be described as follows: the related work is introduced in section 2; and we present a semi-supervised learning framework for intrusion detection in section 3; then we show a misleading attack against the semi-supervised learning method in section 4; a possible defense method is proposed in section 5; and we show the experimental results in section 6; finally, we draw the conclusions in section 7.

2 Related Work

2.1 Machine Learning for Intrusion Detection

Wenke Lee et al. [1] utilized data mining method to find features relevant to intrusions and proposed an anomaly filter to block such intrusions. From that time, machine learning became a hot direction in intrusion detection.

The current intrusion detection techniques include misuse and anomaly detection.

- *Misuse detection.* Attack behaviors are explicitly defined and all events matching these specification are classified as intrusions.
- *Anomaly detection.* A model of normal events is build and all events deviating the normal models are predicted as intrusions.

There are only few researches focusing on machine learning for misuse detection, such as the methods proposed by C. Kruegel et al. [5] and Dae-Ki Kang et al. [6].

Currently, anomaly detection is the major application of machine learning techniques in the area of intrusion detection. Related work includes K-Nearest Neighbor Classifier[7], Application-Layer intrusion detection[8], instance-based approaches[9], clustering methods[10], probabilistic learning methods[11] and so on.

2.2 Semi-supervised Learning

Semi-supervised learning aims to build better classifiers using both labeled and unlabeled instances in the situation that few labeled instances are available and a large number of unlabeled instances can be easily collected.

Some assumptions should be satisfied for semi-supervised learning, including [12]: (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 clustering 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.

A large number of semi-supervised learning methods were proposed in recent years. They can be sumerized into the following categories: self training[13], generative models[14], low density separation[15], and graph-based methods[16,17].

2.3 Attacks against Machine Learning for Intrusion Detection

According to the methods used to attack machine learning systems, typical attacks against machine learning process can be summarized into two categories [18]: *Causative attacks* and *Exploratory attacks*.

- *Causative attacks*: The attackers pollute the training instances to mislead the trained classifier. Such attacks include the red herring attack[4], the correlated outlier attack[4], the allergy attack [19] and so on.
- *Exploratory attacks*: The attackers do not alter the training instances but probe the generated classier to find the classification boundary. Thus the instances can be misclassified by the classifier will be known by the attackers. Typical attacks include the polymorphic blending attack [20], the reverse engineering attack [21], the mimicry attack against "stide" [22] and so on.

Causative attacks are more dangerous than *Exploratory attacks* but need very strong assumptions. The attackers need to change the labeled instances for launching *Causative attacks*. But in the intrusion detection environment, the labeled instances are verified by human experts and are carefully protected. Thus, they are not easy to be polluted.

3 A Semi-supervised Learning Framework for Intrusion Detection

The task of intrusion detection is to monitor network packets and classify them as normal or malicious according to the features of network packets.

In this section, we construct a framework of semi-supervised learning for intrusion detection.

3.1 Preliminaries

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 $Y = \{y_1, y_2, \cdots, y_p\}$ be the set of possible labels. We just consider *normal* and *malicious* in this paper, thus there are only $0, 1 \in Y$. 0 denotes *normal* and 1 denotes *malicious*.

The target function f to be learned is a function $f : X \to Y$ that classifies any $x \in X$ as a member of Y. Y has p elements. The notion $< x, f(x) >$ denotes a labeled instance and $< x, ? >$ denotes an unlabeled instance where $? \in Y$. 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:$< x_{l+1}, ? >, \cdots, < x_{l+u}, ? >$. Usually, we have $l \ll u$. The total number of instances is $n = l + u$.

A typical supervised machine learning process can be divided into 2 stages: the training stage and the testing stage.

- In the training stage, the system collects lots of labeled instances L and then trains a classifier h, which is a function mapping $x \in X$ to $y \in Y$, on L.
- In the testing stage, the new coming unlabeled instance is submitted to the trained classifier h and then h returns $h(x)$ to the system.

When using semi-supervised learning method, the classifier will be trained on labeled instances and unlabeled instances.

3.2 General Learning Process

The general learning process for intrusion detection can be described as follows.

First, the system monitors the network packets and transforms each packet to an instance. Then after a long time of monitoring, there are lots of collected instances. Human experts can analyze those instances and label some of them. Thus, there are labeled instances and unlabeled instances in the system. Since labeled instances require the labor of human experts and then it could be tedious and time consuming, there are often too few labeled instances compared with unlabeled instances.

After that, the system utilizes machine learning methods to train a classifier on historical labeled instances and classifies the new coming instance using the generated classifier.

3.3 The Naive Bayes Classifier

We use Naive Bayes as our basic classifier for intrusion detection.

The classifier f can be defined as follows:

$$f(x) = \arg\max_{y_j \in Y} P(y_j | x_1, x_2, \ldots, x_m) \tag{1}$$

$$= \arg\max_{y_j \in Y} \frac{P(x_1, x_2, \ldots, x_m | y_j) P(y_j)}{P(x_1, x_2, \ldots, x_m)} \tag{2}$$

$$= \arg\max_{y_j \in Y} P(x_1, x_2, \ldots, x_m | y_j) P(y_j) x \tag{3}$$

$$= \arg\max_{y_j \in Y} P(y_j) \prod_i P(x_i | y_j) \tag{4}$$

$P(x_i|y_j)$ and $P(y_j)$ can be calculated on the labeled instances L.

3.4 Self Training for Intrusion Detection

According to the clustering assumption, there should be only few instances near the classification boundary, thus the distribution of unlabeled instances can be used to determine the classification boundary.

For simplification, we choose the self training method, which is a typical semi-supervised learning method, to train the classifier. Algorithm 1 shows the process.

Algorithm 1. the Self-training-based Semi-supervised Learning Algorithm

repeat
1.Train the Naive Bayes classifier f on L;
2.Predict $f(x)$ for each x in U;
3.Select m instances with the lowest entropy from U as A and Add them with their predicted label $f(x)$ to L;
4.Delete A from U;
until $U = \varnothing$

4 A Misleading Attack

The self training method can enhance classification by using unlabeled instances because the distribution of unlabeled instances provide information of the classification boundary. But in real environments, the classifier can not distinguish whether an unlabeled instance is fake or true because the attacker can send any network packet that the monitoring sensor simply transforms it into an instance and submit it to the learning process. Thus, the attacker can generate a large number of fake unlabeled instances to change the distribution of instances and then to influence the classification boundary when the learning process involve unlabeled instances.

In this section, we propose a method to generate fake unlabeled instances to attack the semi-supervised learning method presented in section 3. The attack just misleads the classifier to recognize *malicious packets* as *normal packets*.

4.1 Instance Template Selection

First, we select some unlabeled instances which have the highest uncertainty in U and are classified as *normal* by the classifier. In this situation, it implies such instances are *normal* and near the classification boundary. These instances will be the templates of new generated instances.

The uncertainty of an instance can be defined as the class distribution entropy of that instance.

$$C(x) = \sum_{y \in Y} (-P(y|x) log P(y|x)) \tag{5}$$

Formula 5 shows the definition of uncertainty where x denotes an instance and Y denotes the class label set.

4.2 Misleading Instance Generation

After the instance templates are selected, we randomly select a few attributes of those instances and disturb the value of those attributes a little. Several instances will be constructed on each instance template.

After misleading instances are generated, we should evaluate the instance to ensure it still belongs to *normal* and its uncertainty is still high.

The evaluation function is defined in formula 5. And we specify a threshold ε. If the uncertainty of the generated instance is below ε, then it will be abandoned.

The algorithm can be described in Algorithm 2.

Algorithm 2. Instance Generation Algorithm

Input: an instance templates set *TES*, a number t to specify how many instances can be generated on each instance template, the certainty threshold ε and a set G of generated instances.
Begin:
repeat
1.For each instance $x_i \in TES$,
repeat
1.1 Select a subset of attributes ;
1.2 Add some change to the value of the selected attributes;
1.3 Evaluate the certainty of the new generated instance and delete the instance with the uncertainty below ε.
1.4 Add the changed instance into G.
until t new generated instances were added into G
until $x_i \in TES$ are all iterated

The algorithm can guarantee these new generated instances are classified as *normal* and have high uncertainty. Thus the number of *normal* instances near the classification boundary is much larger than normal situations and then the classifier tends to recognize the instances near the classification boundary as *normal*. Then if the attacker launches an intrusion whose corresponding instance is near the classification boundary, it could be misclassified.

5 Possible Defense

The reason why the attack can mislead the classifier lies on the fact that the unlabeled instances can be generated without any verification and their labels can only be predicted by the current classifier. Thus, the attackers can easily change the density of distribution in the neighborhood of the classification boundary. This could be dangerous because we determine the classification boundary according to density of distribution in semi-supervised learning.

Nevertheless, if we can obtain the labels of such generated instances, the true classification boundary can be determined.

Therefore, we can utilize active learning to defense the misleading attack. In active learning, the system can query the label of an instance and receive

a "oracle" telling the true label of the instance. The main advantage of active learning compared to other learning methods is that it can choose the most informative instances for labeling, thus the labeling cost could be saved. Then we can only query the labels of a few uncertain instances to determine the "real" classification boundary.

The key problem of active learning is how to evaluate the informative instance for labeling. A typical way is to use the class distribution entropy for evaluation. Therefore, we present the semi-supervised learning algorithm with defense in Algorithm 3.

Algorithm 3. the Self-training-based Semi-supervised Learning Algorithm with Defense

repeat
1.Train the Naive Bayes classifier f from L;
2.Predict each x in U;
3.Select n instances with the highest entropy from U and query the true labels of them.
4.Add these instances with their true labels into L and delete them from U.
5.Train the Naive Bayes classifier f from L again;
6.Select m instances with the lowest entropy from U as A and Add them with their predicted label $f(x)$ to L;
7.Delete A from U;
until $U = \varnothing$

In Algorithm 3, any instances with high uncertainty are sampled for labeling. These instances are near the classification boundary, thus the misleading attack can not change the classification boundary very much.

6 Experiments Results

6.1 Methodology

We conducted a series of experiments to test the methods proposed in this paper. The tested methods can be described as follows.

- NaiveBayes classifier for intrusion detection (NB): The NaiveBayes method is used to train the intrusion detection classifier.
- Self-training-based NaiveBayes classifier for intrusion detection (SSLNB): The NaiveBayes method is used to train the intrusion detection classifier and unlabeled instances were selected to enhance the classification based on the self training method shown in Algorithm 1;
- Self-training-based NaiveBayes classifier under the misleading attack (SSLNBMA): Lots of fake instances are generated to mislead the self-training-based NaiveBayes classifier;

– The active learning method defending the misleading attack (ALSSL): We use the active learning method to query the labels of some uncertain instances and train the classifier on them with the labeled instances so that the true classification boundary can be determined.

We select the datasets from the 1999 KDD intrusion detection contests as our test dataset. The datasets were provided by MIT Lincoln Lab. They were gathered from a local-area network simulating a typical military network enviroment with a wide variety of intrusions over a period of 9 weeks. The datasets can be download form the UCI KDD repository [23].

In the datasets, each instance has 41 features and 1 class feature which includes 22 attack types and the normal event. For simplification, we transformed the dataset into a 2-class dataset with "normal" and "malicious" as the class labels. The whole intrusion detection dataset is quite large and we selected 5000 instances from the original dataset randomly as our benchmark dataset.

When testing each method we listed above, the dataset was randomly divided into three part: the unlabeled set *US*, the labeled set *LS* and the testing set *TS*. *TS* contains 800 instances. When the experiment begins, *LS* contains only 1 instance selected randomly and *US* contains the rest. Then we selected 1 instance in *US* randomly and move it into *LS* in each iteration. Thus we trained the basic classifier on *LS* and recorded the accuracy of the classifier on *TS*.

We change the classifier in the experiments. In NB, the classifier is the Naive-Bayes classifier. In SSLNB, the classifier is the selftraining-based NaiveBayes classifier described in Algorithm 1 and in each iteration, we prelabel 200 instances. In SSLNBMA, we generated fake instances based on Algorithm 2 and ε is set to 0.292. In ALSSL, we query 2 instances for labeling in each iteration.

6.2 Results

Figure 1-3 show the results of these experiments. Each recorded data is the average of 20 runs.

In Figure 1, we show the learning curve of the NB method comparing with that of the SSLNB method. The vertical axis shows the accuracy of the classifier and the horizontal axis shows the number of labeled instances.

From Figure 1, we can see all the learning curves climb rapidly at the beginning and then continue to rise slowly. At last, both curves tend to reach the accuracy about 97.5%. Maybe in other areas, the accuracy about 97.5% is good enough. But in intrusion detection, a classifier with the accuracy about 97.5% could cause heavy damage. A classifier for intrusion detection could face thousands of network packets a day, the accuracy about 97.5% means lots of them can be misclassified. Even if there is only one misclassified event which can get the root permission of the targeted system, the intrusion detection classifier fails. Thus, the accuracy about 97.5% does not mean that the classifier is sufficient to protect the system. But, we do not focus on accuracy in this paper. What we care about is whether the semi-supervised learning can raise the accuracy compared with supervised learning and whether the misleading attack can reduce the accuracy.

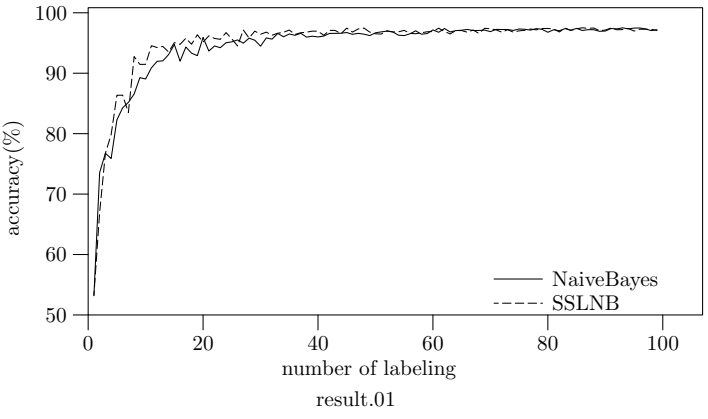


Fig. 1. Learning curves of NB and SSLNB

In figure 1, the SSLNB method has a superior performance than the NB method. It means that we can just label a small number of instances and then the classifier can quickly raise the accuracy by using lots of unlabeled instances with no cost.

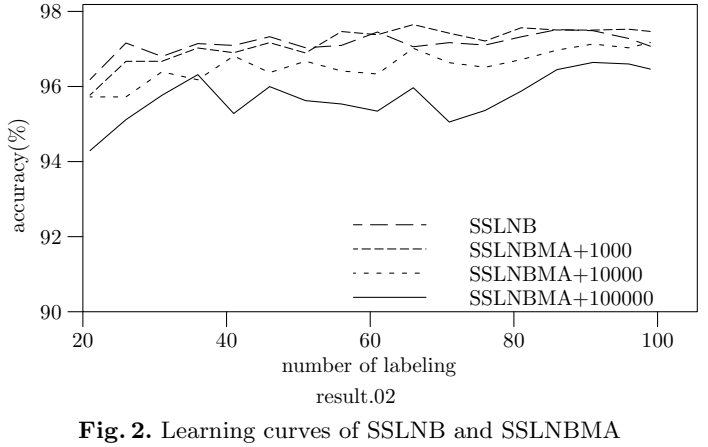


Fig. 2. Learning curves of SSLNB and SSLNBMA

Figure 2 shows the learning curves of the SSLNB method under the misleading attack with different number of fake instances generated. In the experiment, we change the number of generated instances. From figure 2, we can found the attack can cause no obvious decline in accuracy when there are 1000 generated instances. But when we generated 10000 fake instances, the accuracy of the SSLNB method under the misleading attack is a little lower than that of the SSLNB under no attacks. Moreover, when 100000 fake instances were generated, the accuracy of the SSLNB method under the misleading attack has a very

obvious decline than that of other learning curves. In real environments, it is more dangerous because the attacker can easily construct millions of such fake instances in very short time.

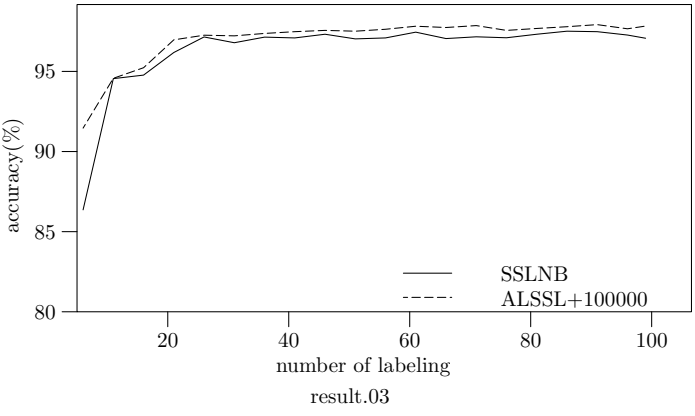


Fig. 3. Learning curves of SSLNB and ALSSL

Figure 3 exhibits the learning curves of the SSLNB under the misleading attack when we utilize the active learning method. It shows that the misleading attack can not decrease the accuracy when active learning is used for defense. The only cost of the defense is that the active learning need to ask for the real label of the generated instances. This requires complicated analysis and labor of human experts. But for security, the labeling cost is worthy.

7 Conclusions

Machine learning have been used for intrusion detection to construct more adaptive classifiers. Currently, semi-supervised learning begin to play important role in this area due to limited high quality labeled instances. But it will fail if the attacker exploit the vulnerability of semi-supervised learning.

In this paper, we build a general semi-supervised learning framework for intrusion detection based on self training. And then we propose a misleading attack method, in which a large number of fake instances were generated, aiming at the self training method. Such attack can reduce the accuracy of the build semi-supervised learner. We also presented a possible defense method, which utilize active learning, to handle the misleading attack.

We would like to pursue the following directions: build the semi-supervised learning framework based on support vector machines and then provide more formal analysis of the attack and defense techniques related to semi-supervised learning.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.60603015, 60603062).

References

1. Lee, W., Stolfo, S.J.: A framework for constructing features and models for intrusion detection systems. ACM Trans. Inf. Syst. Secur. 3(4), 227–261 (2000)
2. Mao, C.H., Lee, H.M., Parikh, D., Chen, T., Huang, S.Y.: Semi-supervised cotraining and active learning based approach for multi-view intrusion detection. In: SAC 2009: Proceedings of the 2009 ACM Symposium on Applied Computing, pp. 2042–2048. ACM, New York (2009)
3. Lane, T.: A decision-theoretic, semi-supervised model for intrusion detection (2004)
4. Newsome, J., Karp, B., Song, D.X.: Paragraph: Thwarting signature learning by training maliciously. In: Zamboni, D., Krügel, C. (eds.) RAID 2006. LNCS, vol. 4219, pp. 81–105. Springer, Heidelberg (2006)
5. Krügel, C., Toth, T.: Using decision trees to improve signature-based intrusion detection. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 173–191. Springer, Heidelberg (2003)
6. Kang, D.K., Fuller, D., Honavar, V.: Learning classifiers for misuse detection using a bag of system calls representation. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (eds.) ISI 2005. LNCS, vol. 3495, pp. 511–516. Springer, Heidelberg (2005)
7. Liao, Y.: Machine learning in intrusion detection. PhD thesis, Davis, CA, USA (2005)
8. Rieck, K.: Machine Learning for Application-Layer Intrusion Detection. PhD thesis, Berlin, Germany (2009)
9. Liao, Y., Vemuri, V.R.: Use of k-nearest neighbor classifier for intrusion detection. Computers & Security 21(5), 439–448 (2002)
10. Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: SDM (2003)
11. Mahoney, M.V., Chan, P.K.: Learning nonstationary models of normal network traffic for detecting novel attacks. In: KDD, pp. 376–385 (2002)
12. Zhu, X., Goldberg, A.B.: Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (2009)
13. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196 (1995)
14. Mccallum, A.K., Mitchell, T.: Text classification from labeled and unlabeled documents using em. In: Machine Learning, pp. 103–134 (2000)
15. Joachims, T.: Transductive inference for text classification using support vector machines, pp. 200–209. Morgan Kaufmann, San Francisco (1999)
16. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation (2002)
17. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434 (2006)

18. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., Tygar, J.D.: Can machine learning be secure? In: ASIACCS, pp. 16–25 (2006)
19. Chung, S.P., Mok, A.K.: Collaborative intrusion prevention. In: WETICE, pp. 395–400 (2007)
20. Fogla, P., Lee, W.: Evading network anomaly detection systems: formal reasoning and practical techniques. In: ACM Conference on Computer and Communications Security, pp. 59–68 (2006)
21. Lowd, D., Meek, C.: Adversarial learning. In: KDD, pp. 641–647 (2005)
22. Tan, K.M.C., Killourhy, K.S., Maxion, R.A.: Undermining an anomaly-based intrusion detection system using common exploits. In: Wespi, A., Vigna, G., Deri, L. (eds.) RAID 2002. LNCS, vol. 2516, pp. 54–73. Springer, Heidelberg (2002)
23. Archive, T.U.K.: Kdd cup 1999 data (October 1999)

Author Index