

Sébastien Destercke
Thierry Denoeux
Fabio Cuzzolin
Arnaud Martin (Eds.)

LNAI 11069

Belief Functions: Theory and Applications

5th International Conference, BELIEF 2018
Compiègne, France, September 17–21, 2018
Proceedings

 Springer

Lecture Notes in Artificial Intelligence

11069

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/1244>

Sébastien Destercke · Thierry Denoeux
Fabio Cuzzolin · Arnaud Martin (Eds.)

Belief Functions: Theory and Applications

5th International Conference, BELIEF 2018
Compiègne, France, September 17–21, 2018
Proceedings

Editors

Sébastien Destercke 
University of Technology of Compiègne
Compiègne Cedex
France

Thierry Denoeux
UMR CNRS 7253 Heudiasyc
Université de Technologie de Compiègne
Compiègne Cedex
France

Fabio Cuzzolin
Department of Computing
and Communication
Oxford Brookes University
Oxford
UK

Arnaud Martin
Université de Rennes 1
Lannion Cedex
France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-99382-9 ISBN 978-3-319-99383-6 (eBook)
<https://doi.org/10.1007/978-3-319-99383-6>

Library of Congress Control Number: 2018951243

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The theory of belief functions, also known as evidence theory or Dempster–Shafer theory, was first introduced by Arthur P. Dempster in the context of statistical inference, and was later developed by Glenn Shafer as a general framework for modeling epistemic uncertainty. These early contributions have been the starting points of many important developments, including the transferable belief model and the theory of hints. The theory of belief functions is now well established as a general framework for reasoning with uncertainty, and has well understood connections with other frameworks such as probability, possibility, and imprecise probability theories.

The series of biennial International Conferences on Belief Functions (BELIEF) is dedicated to the confrontation of ideas, the reporting of recent achievements, and the presentation of the wide range of applications of this theory. This conference series was launched in Brest, France, in 2010. It subsequently took place in Compiègne (2012), Oxford (2014) and Prague (2016). In 2018, the conference was held again in Compiègne, during September 17–21. The reason for such a quick come-back was to seize the opportunity to have a joint event with the 9th International Conference on Soft Methods in Probability and Statistics (SMPS). Such a joint meeting promotes interactions and discussions between different communities working on different aspects of uncertainty theories.

This volume contains the proceedings of the 5th International Conference on Belief Functions. The joint event collected 61 accepted submissions, each reviewed by at least two reviewers. Thirty-three of these are included in the present volume. Original contributions were solicited on theoretical aspects (including, for example, statistical inference, mathematical foundations, continuous belief functions) as well as on applications in various areas including classification, statistics, data fusion, network analysis, and intelligent vehicles. The resulting proceedings were easily produced through the use of EasyChair.

We would like to thank all the persons who made this volume and this conference possible: all contributing authors, organizers, Program Committee members who helped to build such an attractive program. We are especially grateful to our three invited speakers, Thomas Augustin (*Ludwig-Maximilians-Universität München*) for his talk “Belief Functions and Valid Statistical Inference,” Scott Ferson (*University of Liverpool*) for his talk “Non-Laplacian Uncertainty: Practical Consequences of an Ugly Paradigm Shift About How We Handle not Knowing,” and Ryan Martin (*North Carolina State University*) for his talk “Belief Functions and Valid Statistical Inference.” We would like to thank all our generous sponsors: Elsevier and the *International Journal of Approximate Reasoning*, the Laboratory of Excellence MS2T, the Heudisyc laboratory, the International Society of Information Fusion (ISIF), the Compiègne

University of Technology, and the city of Compiègne. Furthermore, we would like to thank the editors of the Springer series *Lecture Notes in Computer Science*, and Springer for their dedication to the production of this volume.

June 2018

Sébastien Destercke
Thierry Denoeux
Fabio Cuzzolin
Arnaud Martin

Organization

Program Committee

Alessandro Antonucci	IDSIA, Switzerland
Thomas Augustin	University of Munich (LMU), Munich
Giulianella Coletti	University of Perugia, Italy
Olivier Colot	Université Lille 1, France
Ana Colubi	University of Oviedo, Spain
Frank Coolen	Durham University, UK
Inés Couso	University of Oviedo, Spain
Fabio Cuzzolin	Oxford Brookes University, UK
Fabio D'Andreagiovanni	Université de Technologie de Compiègne, UMR CNRS Heudiasyc, France
Pierpaolo D'Urso	Sapienza University of Rome, Italy
Bernard De Baets	Ghent University, Belgium
Thierry Denoëux	Université de Technologie de Compiègne, UMR CNRS Heudiasyc, France
Sébastien Destercke	Université de Technologie de Compiègne, UMR CNRS Heudiasyc, France
Jean Dezert	Onera, France
Didier Dubois	Université de Paul Sabatier, Toulouse, UMR IRIT, France
Fabrizio Durante	Università del Salento, Lecce, Italy
Zied Elouedi	Institut Supérieur de Gestion de Tunis, Tunisia
Ramasso Emmanuel	Ecole Nationale Supérieure de Mécanique et des Microtechniques, FEMTO-ST, France
Maria Brigida Ferraro	Sapienza University of Rome, Italy
Maria Angeles Gil Alvarez	University of Oviedo, Spain
Lluis Godo	Artificial Intelligence Research Institute, IIIA - CSIC, Spain
Gil González-Rodríguez	University of Oviedo, Spain
Michel Grabisch	Université Paris I, France
Przemyslaw Grzegorzewski	Polish Academy of Sciences, Systems Research Institute, Poland
Olgierd Hryniewicz	Polish Academy of Sciences, Systems Research Institute, Poland
Radim Jirousek	University of Economics, Czech Republic
Anne-Laure Jousselme	NATO Centre for Maritime Research and Experimentation (CMRE), Italy
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Vaclav Kratochvíl	UTIA, Czech Republic

Rudolf Kruse	University of Magdeburg, Germany
Eric Lefevre	LGI2A Université d'Artois, France
Liping Liu	University of Akron, USA
María Asunción Lubiano	University of Oviedo, Spain
Arnaud Martin	Université de Rennes 1/IRISA, France
Ronald W. J. Meester	Vrije Universiteit Amsterdam, The Netherlands
David Mercier	Université d'Artois, France
Radko Mesiar	Slovak University of Technology Bratislava, Slovakia
Rombaut Michele	Gipsa-lab, France
Daniel Milan	Institute of Computer Science, The Czech Academy of Sciences, Czech Republic
Enrique Miranda	University of Oviedo, Spain
Ignacio Montes	Carlos III University of Madrid, Spain
Susana Montes	University of Oviedo, Spain
Serafin Moral	University of Granada, Spain
Frédéric Pichon	Université d'Artois, France
Benjamin Quost	Université de Technologie de Compiègne, UMR CNRS Heudiasyc, France
Ana Belén Ramos Guajardo	University of Oviedo, Spain
Johan Schubert	Swedish Defence Research Agency, Sweden
Ferson Scott	University of Liverpool, Institute for Risk and Uncertainty, UK
Prakash P. Shenoy	University of Kansas School of Business, USA
Beatriz Sinova	University of Oviedo, Spain
Martin Stepnicka	IRAFM, University of Ostrava, Czech Republic
Barbara Vantaggi	Sapienza University of Rome, Italy
Jirina Vejnárova	Institute of Information Theory and Automation of the AS, Czech Republic
Paolo Vicig	University of Trieste, Italy
Liu Zhunga	Northwestern Polytechnical University, China

Contents

An Evidential Collaborative Filtering Approach Based on Items Contents Clustering	1
<i>Raoua Abdelkhalek, Imen Boukhris, and Zied Elouedi</i>	
The Belief Functions Theory for Sensors Localization in Indoor Wireless Networks	10
<i>Daniel Alshamaa, Farah Mourad-Cehade, and Paul Honeine</i>	
On Evidential Clustering with Partial Supervision	14
<i>Violaine Antoine, Kévin Gravouil, and Nicolas Labroche</i>	
Exploiting Domain-Experts Knowledge Within an Evidential Process for Case Base Maintenance	22
<i>Safa Ben Ayed, Zied Elouedi, and Eric Lefevre</i>	
The Kantorovich Problem and Wasserstein Metric in the Theory of Belief Functions	31
<i>Andrey G. Bronevich and Igor N. Rozenberg</i>	
Generalised Max Entropy Classifiers	39
<i>Fabio Cuzzolin</i>	
General Geometry of Belief Function Combination	48
<i>Fabio Cuzzolin</i>	
Logistic Regression Revisited: Belief Function Analysis	57
<i>Thierry Denoeux</i>	
From Relations Between Sets to Relations Between Belief Functions.	65
<i>Sébastien Destercke, Frédéric Pichon, and John Klein</i>	
Application of Belief Functions to Levee Assessment	73
<i>Théo Dezert, Yannick Fargier, Sérgio Palma Lopes, and Philippe Côte</i>	
Prejudiced Information Fusion Using Belief Functions.	77
<i>Didier Dubois, Francis Faux, and Henri Prade</i>	
A Heuristic Approach for the Robust Flight Level Assignment Problem.	86
<i>Akli Fundo, Dritan Nace, and Chenghao Wang</i>	
Study of Distributed Data Fusion Using Dempster's Rule and Cautious Operator	95
<i>Romain Guyard and Véronique Cherfaoui</i>	

Uncertainty-Aware Parzen-Rosenblatt Classifier for Multiattribute Data	103
<i>Ali Hamache, Mohamed El Yazid Boudaren, Houdaifa Boukersoul, Islam Debicha, Hamza Sadouk, Rezki Zibani, Ahmed Habbouchi, and Omar Merouani</i>	
Birnbaum's Importance Measure Extended for Non-coherent Systems	112
<i>Ayyoub Imakhlaf and Mohamed Sallak</i>	
Evidential Independence Maximization on Twitter Network	121
<i>Siwar Jendoubi, Mouna Chebbah, and Arnaud Martin</i>	
An Evidential k -nearest Neighbors Combination Rule for Tree Species Recognition	129
<i>Siwar Jendoubi, Didier Coquin, and Reda Boukezzoula</i>	
A Compact Belief Rule-Based Classification System with Evidential Clustering	137
<i>Lianmeng Jiao, Xiaojiao Geng, and Quan Pan</i>	
A Decomposable Entropy of Belief Functions in the Dempster-Shafer Theory	146
<i>Radim Jiroušek and Prakash P. Shenoy</i>	
An Evidential K -Nearest Neighbor Classifier Based on Contextual Discounting and Likelihood Maximization	155
<i>Orakanya Kanjanatarakul, Siwarat Kuson, and Thierry Denoeux</i>	
Measuring Market Performance with Stochastic Demand: Price of Anarchy and Price of Uncertainty	163
<i>Costis Melolidakis, Stefanos Leonardos, and Constandina Koki</i>	
On the Conflict Measures Agreed with the Combining Rules	172
<i>Alexander Lepskiy</i>	
Linear Belief Functions for Data Analytics.	181
<i>Liping Liu</i>	
Outer Approximations of Coherent Lower Probabilities Using Belief Functions	190
<i>Ignacio Montes, Enrique Miranda, and Paolo Viciò</i>	
An Ordered Family of Consistency Measures of Belief Functions	199
<i>Nadia Ben Abdallah, Anne-Laure Jousselme, and Frédéric Pichon</i>	
Active Evidential Calibration of Binary SVM Classifiers	208
<i>Sébastien Ramel, Frédéric Pichon, and François Delmotte</i>	
Decision Making: A Beliefs, Preferences and Constraints Model	217
<i>Aouatef Rouahi, Kais Ben Salah, and Khaled Ghédira</i>	

Belief and Plausibility Functions on the Space of Scalar Products
and Applications 226
Juan J. Salamanca

E2CM: An Evolutionary Version of Evidential C-Means
Clustering Algorithm 234
*Zhi-gang Su, Hong-yu Zhou, Pei-hong Wang, Gang Zhao,
and Ming Zhao*

Contrasting Two Laws of Large Numbers from Possibility Theory
and Imprecise Probability 243
Pedro Terán and Elisa Pis Vigil

Improved Performance of EK-NNClus by Selecting
Appropriate Parameter 252
Qian Wang and Zhi-gang Su

An Empirical Study to Determine the Optimal k in Ek-NNclus Method 260
Yiru Zhang, Tassadit Bouadi, and Arnaud Martin

Evidential Community Detection Based on Density Peaks 269
Kuang Zhou, Quan Pan, and Arnaud Martin

Author Index 279



An Evidential Collaborative Filtering Approach Based on Items Contents Clustering

Raoua Abdelkhalek^(✉), Imen Boukhris, and Zied Elouedi

LARODEC, Institut Supérieur de Gestion de Tunis,
Université de Tunis, Tunisia

abdelkhalek_raoua@live.fr, imen.boukhris@hotmail.com, zied.elouedi@gmx.fr

Abstract. Recommender Systems (RSs) have emerged as powerful tools to provide the users with personalized recommendations and to guide them in their decision making process. Among the various recommendation approaches, Collaborative Filtering (CF) is considered as one of the most popular techniques in RSs. CF techniques are categorized into model-based and memory-based. Model-based approaches consist in learning a model from past ratings to perform predictions while memory-based ones predict ratings by selecting the most similar users (user-based) or the most similar items (item-based). In both types, recommendations are fully based on users' past ratings. However, aside from users' ratings, exploiting additional information such as items' features would enhance the accuracy of the provided predictions. Another crucial challenge in the RSs area would be to handle uncertainty arising throughout the prediction process. That is why, in this paper, we propose an item-based Collaborative Filtering under the belief function theory that not only takes advantages of both model- and memory-based CF approaches but also integrates items' contents in the recommendation process.

Keywords: Recommender systems · Collaborative filtering
Model-based · Memory-based · Belief function theory
Uncertainty · Items contents

1 Introduction

Recommender Systems (RSs) [1] are considered as an efficient tool to cope with the information overload problem. Such systems generally try to predict the users' future ratings on unseen items and provide personalized recommendations accordingly. In the research area of RSs, Collaborative Filtering (CF) approaches [2] are considered among the most popular strategies commonly adopted in this field. According to how they process the rating matrix, CF systems can be divided into model-based and memory-based categories. Memory-based CF, also referred to as neighborhood-based, compute the similarities between users (user-based) or items (item-based) and then select the most similar ones for recommendations. Commonly, Pearson and Cosine correlation coefficients are the most

widely used similarity measures in the neighborhood-based CF approaches [3]. In contrast to memory-based systems, which rely on the users' ratings directly in the prediction, model-based approaches exploit these ratings to build a model which is used to predict ratings. In traditional CF systems, the final predictions represent the user's preference for a given item as a rating score (i.e., a hard rating). The predicted value indicates whether the concerned item would interest the active user or not. Nonetheless, most of existing recommendation techniques have not considered the important issue of uncertainty which reigns in real-world problems. Such uncertainty needs to be appropriately represented and processed so as to improve quality and reliability of RSs [4]. The belief function theory (BFT) [5,6] is considered among the most used theories for reasoning under uncertainty [7]. In our paper, we embrace this theory to quantify and represent the uncertainty in the recommendation process. Furthermore, CF strategies rely only on the available ratings given by the users. However, various additional information stretching beyond the rating matrix can generally be available such as items' features (i.e., contents). Naturally, the more sources of information about the given items are exploited, the more effective the performance of recommendations will be. Hence, we propose an evidential CF approach that deals with uncertainty in both clusters assignment and final predictions while making use of the items' contents aside from their corresponding ratings.

This paper is organized as follows: Sect. 2 gives the necessary background of the belief function framework. In Sect. 3, we provide some related work of the Collaborative Filtering recommender. Section 4 describes our proposed approach. Section 5 depicts the experimental results. Finally, Sect. 6 concludes the paper and reports some potential future works.

2 Background on the Belief Function Theory

Let Θ be the frame of discernment representing the set of n elementary events such that: $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$. It contains hypotheses concerning the given problem. The power set of Θ , denoted by 2^Θ , is the set of all possible subsets of Θ . A basic belief assignment (*bba*) expresses the belief committed to each element of 2^Θ . It corresponds to the mapping function $m : 2^\Theta \rightarrow [0, 1]$ such that $\sum_{E \subseteq \Theta} m(E) = 1$ where $m(E)$ represents the basic belief mass (*bbm*) stating the part of belief exactly committed to the event E .

When an event $E \subseteq \Theta$ has $m(E) > 0$, it is called a focal element. A discounting mechanism can be adopted to account for reliability of the independent sources such that:

$$m^\alpha(E) = (1 - \alpha) \cdot m(E), \forall E \subset \Theta; m^\alpha(\Theta) = \alpha + (1 - \alpha) \cdot m(\Theta)$$

where $\alpha \in [0, 1]$ is the discounting factor.

The fusion of two *bba*'s m_1 and m_2 derived from two reliable and independent sources of evidence can be performed using Dempster's rule of combination. It is defined as follows, where the empty set \emptyset is the unique set having no elements.

$$(m_1 \oplus m_2)(E) = k \cdot \sum_{F, G \subseteq \Theta: F \cap G = E} m_1(F) \cdot m_2(G)$$

$$\text{where } (m_1 \oplus m_2)(\emptyset) = 0 \text{ and } k^{-1} = 1 - \sum_{F, G \subseteq \Theta: F \cap G = \emptyset} m_1(F) \cdot m_2(G).$$

To make decisions, beliefs can be transformed into a pignistic probability $BetP(E)$ computed as follows:

$BetP(E) = \sum_{F \subseteq \Theta} \frac{|E \cap F|}{|F|} \frac{m(F)}{(1 - m(\emptyset))}$ for all $E \subseteq \Theta$. The hypothesis having the highest value of $BetP(E)$ is then selected.

In order to process uncertain data, a panoply of machine learning techniques has been proposed under this theory, such as the Evidential K-Nearest Neighbors [8] which allows a credal classification of the objects and the Evidential c-means (ECM) [9] which allows the objects to belong to more than only one cluster, which is referred to as credal partition.

3 Related Work on Collaborative Filtering

Much research has been recently devoted to the development of CF approaches aiming to enhance the accuracy and the performance of the recommendations. Clustering-based methods are among the widely used techniques in model-based CF. In these approaches, a cluster model is created based on the available ratings and predictions are then made based on these clusters. For instance, a clustering based CF approach has been proposed in [10] to group the users in different clusters based on their ratings and predictions have been performed accordingly. A graph cut-based clustering approach has been proposed in [11] to facilitate the formation of similar user groups. In our work, we consider only item-based CF where items are clustered into groups rather than users. While CF techniques rely basically on users' ratings to provide recommendations, RSs research directions are now emerging to exploit, not only the items' ratings, but also additional information that goes beyond the rating matrix such as items' contents. In [12], the correlation between movies genres is computed and traditional user-based CF is then used to predict ratings. In [13], authors have developed a TV program RS where they combine items' contents with users' preferences and the matrix factorization technique has been applied. On the other hand, uncertainty can arise in many different ways when dealing with RSs. Consequently, it is fundamental to take it into account. Recent works have emphasized the benefits of the incorporation of such uncertainty using the belief function theory. Indeed, authors in [14, 15] have extended traditional item-based CF under the belief function theory where the items' ratings have been represented by belief functions and combined to provide the final predictions. A clustering based CF approach has also been proposed under this theory where ECM has been involved to cluster items based only on their ratings. Predictions are then computed as average values of the similar items' ratings [16].

4 ECF-IC: Evidential CF Based on Items Contents

In our proposed recommendation approach, we tend to exploit the intuition of both model-based and neighborhood-based approaches while making use, not only of the items' ratings but also of their corresponding features. In order to deal with uncertainty, the belief function theory is adopted and several tools and machine learning techniques under this theory come into play. The whole process is illustrated in Fig. 1. Two major steps characterize the new approach namely the model building step where a model is learned based on items features and the evidential prediction step where the rating matrix is explored and predictions are performed based on the K-similar items.

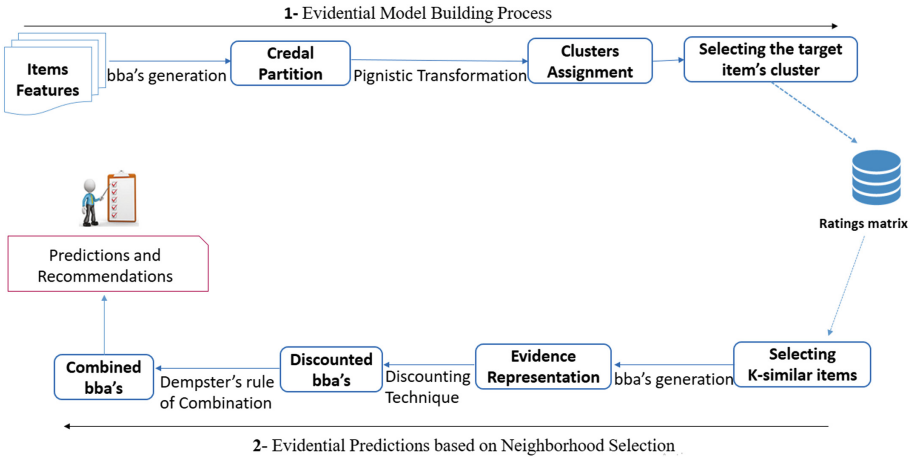


Fig. 1. Evidential CF approach based on items contents clustering

4.1 Model Building

In the first step, items' features are exploited in order to generate soft clusters among the items using the Evidential c-means technique. We define the frame of discernment $\Omega_1 = \{c_1, c_2, \dots, c_M\}$ where M corresponds to the number of clusters c . By exploiting the items contents, we aim in this phase to generate a credal partition of the items. Thus, each given item in the system can belong to any subsets of Ω_1 . For this purpose, we involve the Evidential c-means (ECM) since this efficient soft clustering technique allows to allocate, for each item in the rating matrix, a mass of belief not only to single clusters, but also to any subsets of Ω_1 . Before performing the evidential clustering process, we normalize the items' features to be considered on the same scale as proposed in [17]. For i^{th} attribute and k^{th} value of a given item A , we obtain the normalized value as follows: $NV_{A_{ik}} = (A_{ik} - A_{i,min}) / (A_{i,max} - A_{i,min})$. Once all the items features are normalized, the cluster centers, commonly referred to as prototypes, are

randomly initialized. The Euclidean distance between each item and the non empty subsets of Ω_1 is then computed and the credal partition is derived. More details about the credal partition process and parameters can be found in [9]. Finally, we compute the pignistic probability induced by each *bba*. Each item is assigned to its corresponding cluster based on the derived pignistic probabilities. Once the items clustering is performed, the items belonging to the same cluster as the target item are selected to be used in the second phase. Note that the clustering process performed before the neighborhood selection is justified by its ability to improve the scalability performance of neighborhood-based CF. Thus, items similarities are computed between the target item and the items belonging to the same cluster rather than the whole items in the system.

4.2 Predictions and Recommendations

In this phase, we define $\Omega_2 = \{\omega_1, \omega_2, \dots, \omega_n\}$ where n is the number of the possible ratings ω and $\omega_1 < \omega_2 < \dots < \omega_n$. In each cluster, the distance between the target item and the other items is computed as follows:

$$d(a, b) = \frac{\sqrt{\sum_{u \in (u_a \cap u_b)} (\omega_{u,a} - \omega_{u,b})^2}}{|u_a \cap u_b|}$$

$\omega_{u,a}$ and $\omega_{u,b}$ are the ratings of the user u for the target item a and the item b . u_a and u_b are the users who rated both items a and b . Accordingly, the K-similar items are extracted. Each similar item involves a particular hypothesis about the predicted rating. Hence, we generate a *bba* over each rating provided by the selected neighbor as well as the whole frame of discernment Ω_2 [8].

$$m_{a,b}(\{\omega_i\}) = \alpha_0 \exp^{-(\gamma_{\omega_i}^2 \times d(a,b)^2)}; m_{a,b}(\Omega_2) = 1 - \alpha_0 \exp^{-(\gamma_{\omega_i}^2 \times d(a,b)^2)}$$

Following [8], α_0 is initialized to the value 0.95 and γ_{ω_i} is computed as the inverse of the mean distance between each couple of items sharing the same ratings. We integrate the discounting technique [5] to quantify the reliability of each similar item where we define the discounting factor β as: $\beta = d(a,b)/\max(d)$. $\max(d)$ is the maximum value of the computed distances. We assume that the more similar the item is, the more reliable its evidence is. The discounted *bba*'s are then obtained such as:

$$m_{a,b}^\beta(\{\omega_i\}) = (1 - \beta) \cdot m_{a,b}(\{\omega_i\}); m_{a,b}^\beta(\Omega_2) = \beta + (1 - \beta) \cdot m_{a,b}(\Omega_2)$$

Once the different *bba*'s provided by the K-Nearest Neighbors are generated, they can be combined using Dempster's rule of combination. Inspired by [8], the following equations can be applied:

$$\begin{aligned} \forall \omega_i \in \{\omega_1, \dots, \omega_N\} \quad m^\beta(\{\omega_i\}) &= \frac{1}{Z} (1 - \prod_{i \in S_K} (1 - \alpha_{\omega_i})) \cdot \prod_{\omega_j \neq \omega_i} \prod_{i \in S_K} (1 - \alpha_{\omega_j}) \\ m^\beta(\Omega_2) &= \frac{1}{Z} \prod_{i=1}^N (1 - \prod_{i \in S_K} (1 - \alpha_{\omega_i})) \end{aligned}$$

where S_K is the set containing the K -nearest neighbors of the target item over the user-item matrix. N is the number of the ratings provided by the similar items, α_{ω_i} is the belief committed to the rating ω_i , α_{ω_j} is the belief committed to the rating $\omega_j \neq \omega_i$, Z is a normalized factor defined by:

$$Z = \sum_{i=1}^N (1 - \prod_{i \in S_K} (1 - \alpha_{\omega_i}) \prod_{\omega_j \neq \omega_i} \prod_{i \in S_K} (1 - \alpha_{\omega_j}) + \prod_{i=1}^N (\prod_{i \in S_K} (1 - \alpha_{\omega_j})))$$

5 Experimental Analysis

In our experiments, we use the well-known MovieLens¹ data set in order to evaluate our proposal. Such data set contains 1682 movies rated by 943 users. These ratings are integer scores between 1 (dislike) and 5 (like). We follow the methodology in [18] which consists in ranking the movies rated by the 943 users according to the number of the total ratings such as: $Nb_{user}(movie_1) \geq Nb_{user}(movie_2) \geq \dots \geq Nb_{user}(movie_{1682})$ where $Nb_{user}(movie_i)$ is the number of users who rated the $movie_i$. We extract 10 subsets by progressively increasing the number of the missing rates. Thus, since few ratings provided for the total number of items are available, each subset will contain a specific number of ratings leading to different degrees of sparsity.

Evaluation Measures

We rely on two evaluation metrics: The *Mean Absolute Error* (MAE) defined as: $MAE = \frac{1}{\|\hat{R}_{u,i}\|} \sum_{u,i} |\hat{R}_{u,i} - R_{u,i}|$ and the precision measure defined as follows: $Precision = \frac{IR}{IR+UR}$. Note that $R_{u,i}$ is the real rating for the user u on the item i and $\hat{R}_{u,i}$ is the predicted value. $\|\hat{R}_{u,i}\|$ is the total number of the predicted ratings. IR indicates that an interesting item has been correctly recommended while UR indicates that an uninteresting item has been incorrectly recommended. The lower the MAE is, the more accurate the predictions are while the highest precision indicates a better recommendation quality.

Experimental Results

Based on the 10 extracted subsets, we run our experiments while switching each time the number of clusters c . For each experiment, we use the values $c = 2$, $c = 3$, $c = 4$ and $c = 5$. Then, the MAE and the precision results are computed for each value. We set $\alpha_0 = 0.95$, as in [8], for the evidential prediction process. For the obtained clusters, we set the number of the K -nearest neighbors K to the value $\|c\| - 1$ since most of the best results were achieved with this value. Note that $\|c\|$ corresponds to the number of items in the cluster c . Our proposed approach, that we denote by ECF-IC, is compared against five traditional item-based CF systems: The evidential model-based CF (ECL) [16], the two evidential memory-based CF namely, the evidential item-based CF (EV) [14] and the discounting-based item-based CF (DE) [15]. Besides, we run the

¹ <http://movielens.org>.

traditional Pearson item-based CF (P) and Cosine item-based CF (C) [3]. The obtained results are depicted in Table 1. The combination of both memory- and model-based strategies leads to better results compared to the single evidential memory- and model-based approaches. Furthermore, the integration of items contents in the recommendation process shows a great improvement in the performance of the CF recommender compared to ECL which relies only users' ratings. Overall, our approach achieves better results in term of MAE with a value of 0.788 compared to 0.925 and 0.914 for both Pearson and Cosine CF, 0.809 for EV, 0.789 for DE and 0.793 for ECL. When it comes to the precision measure, the average value of the new approach (0.757) outperforms EV (0.733), DE (0.743), ECL (0.75) as well as Pearson and Cosine approaches (0.706).

Table 1. Overall MAE and Precision

Measures	Subsets	Sparsity	EV	C	P	DE	ECL	ECF-IC
MAE	S_1	53%	0.751	0.824	0.839	0.711	0.749	0.787
Precision			0.79	0.778	0.774	0.774	0.792	0.806
MAE	S_2	56.83%	0.84	0.87	0.936	0.802	0.8	0.81
Precision			0.76	0.739	0.737	0.748	0.74	0.736
MAE	S_3	59.8%	0.761	0.825	0.863	0.836	0.747	0.78
Precision			0.77	0.749	0.752	0.711	0.785	0.753
MAE	S_4	62.7%	0.763	0.876	0.905	0.743	0.793	0.748
Precision			0.763	0.745	0.746	0.775	0.782	0.830
MAE	S_5	68.72%	0.831	1	0.990	0.802	0.845	0.763
Precision			0.741	0.69	0.707	0.787	0.752	0.811
MAE	S_6	72.5%	0.851	0.917	0.976	0.843	0.8	0.785
Precision			0.735	0.733	0.732	0.74	0.813	0.78
MAE	S_7	75%	0.744	0.877	0.943	0.736	0.733	0.84
Precision			0.78	0.745	0.752	0.783	0.805	0.829
MAE	S_8	80.8%	0.718	0.848	0.927	0.723	0.762	0.745
Precision			0.778	0.718	0.729	0.821	0.755	0.733
MAE	S_9	87.4%	0.840	0.978	0.958	0.839	0.873	0.754
Precision			0.707	0.654	0.665	0.74	0.73	0.797
MAE	S_{10}	95.9%	0.991	1.13	0.913	0.978	0.83	0.87
Precision			0.513	0.509	0.463	0.431	0.55	0.5
Overall MAE			0.809	0.914	0.925	0.789	0.793	0.788
Overall Precision			0.733	0.706	0.706	0.743	0.75	0.757

6 Conclusion

In this paper, we have proposed a new evidential CF approach that incorporates items contents in the prediction process while combining both model-based and memory-based strategies. Based on the items contents, the key idea is to learn a model and to perform predictions accordingly whilst handling the uncertainty that occurs in the different steps of the recommendation process. As future work, we intend to better exploit the credal partition in the model building where all the *bba*'s of the different clusters will be considered rather than the most significant one. We intend also to perform more experiments using other real-world data sets. It would also be interesting to propose a hybrid CF approach where both items and users clustering are performed.

References

1. Ricci, F., Rokach, L., Shapira, B.: Recommender systems: introduction and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) *Recommender Systems Handbook*, pp. 1–34. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7637-6_1
2. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. In: *Advances in Artificial Intelligence*, pp. 1–19. Hindawi Publishing Corporation (2009)
3. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: *International Conference on World Wide Web*, pp. 285–295. ACM (2001)
4. Jung, S.Y., Hong, J.-H., Kim, T.-S.: A statistical model for user preference. *IEEE Trans. Knowl. Data Eng.* **17**(6), 834–843 (2005)
5. Dempster, A.P.: A generalization of Bayesian inference. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **30**, 205–247 (1968)
6. Shafer, G.: *A Mathematical Theory of Evidence*, vol. 1. Princeton University Press, Princeton (1976)
7. Cuzzolin, F.: On the orthogonal projection of a belief function. In: Mellouli, K. (ed.) *ECSQARU 2007. LNCS (LNAI)*, vol. 4724, pp. 356–367. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75256-1_33
8. Denoeux, T.: A K-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**, 804–813 (1995)
9. Masson, M.H., Denoeux, T.: ECM: an evidential version of the fuzzy c-means algorithm. *Pattern Recogn.* **41**, 1384–1397 (2008)
10. Zhang, J., Lin, Y., Lin, M., Liu, J.: An effective collaborative filtering algorithm based on user preference clustering. *Appl. Intell.* **45**, 230–240 (2016)
11. Bellogin, A., Parapar, J.: Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In: *ACM Conference on Recommender Systems*, pp. 213–216. ACM (2012)
12. Hwang, T.G., Park, C.S., Hong, J.H., Kim, S.K.: An algorithm for movie classification and recommendation using genre correlation. *Multimedia Tools Appl.* **75**(20), 12843–12858 (2016)

13. Barragns-Martnez, A.B., Costa-Montenegro, E., Burguillo, J.C., Rey-Lpez, M., Mikic-Fonte, F.A., Peleteiro, A.: A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. *Inf. Sci.* **180**(22), 4290–4311 (2010)
14. Abdelkhalek, R., Boukhris, I., Elouedi, Z.: Evidential item-based collaborative filtering. In: Lehner, F., Fteimi, N. (eds.) *KSEM 2016. LNCS (LNAI)*, vol. 9983, pp. 628–639. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47650-6_49
15. Abdelkhalek, R., Boukhris, I., Elouedi, Z.: Assessing items reliability for collaborative filtering within the belief function framework. In: Jallouli, R., Zaiane, O.R., Bach Tobji, M.A., Srarfi Tabbane, R., Nijholt, A. (eds.) *ICDEc 2017. LNBIP*, vol. 290, pp. 208–217. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62737-3_18
16. Abdelkhalek, R., Boukhris, I., Elouedi, Z.: A clustering approach for collaborative filtering under the belief function framework. In: Antonucci, A., Cholvy, L., Papini, O. (eds.) *ECSQARU 2017. LNCS (LNAI)*, vol. 10369, pp. 169–178. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61581-3_16
17. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington (2016)
18. Su, X., Khoshgoftaar, T.M.: Collaborative filtering for multi-class data using Bayesian networks. *Int. J. Artif. Intell. Tools* **17**, 71–85 (2008)



The Belief Functions Theory for Sensors Localization in Indoor Wireless Networks

Daniel Alshamaa^{1(✉)}, Farah Mourad-Chehade¹, and Paul Honeine²

¹ Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France

{daniel.alshamaa, farah.chehade}@utt.fr

² LITIS lab, Université de Rouen Normandie, Rouen, France

paul.honeine@univ-rouen.fr

Abstract. This paper investigates the usage of the belief functions theory to localize sensors in indoor environments. The problem is tackled as a zoning localization where the objective is to determine the zone where the mobile sensor resides at any instant. The proposed approach uses the belief functions theory to define an evidence framework, for estimating the most probable sensor's zone. Real experiments demonstrate the effectiveness of this approach as compared to other localization methods.

1 Introduction

Localization is an essential issue in wireless sensor networks to process the information retrieved by sensor nodes. This paper proposes a zoning-based localization technique that makes use of the belief functions theory (BFT) to combine evidence revealed at each sensor. The proposed approach is constituted of two phases. In an offline phase, received signal strength indicators (RSSIs) received from neighboring WiFi Access Points (APs) are collected in each zone and a fingerprints database is built. The kernel density estimation is then used to represent the measurements and set mass functions over the zones. In the same manner, mass functions are also constructed over supersets of zones, by concatenating zones data. In an online phase, the collected RSSIs of a mobile sensor are used in the belief functions framework to determine its zone. Since APs are not completely reliable, their associated masses are discounted according to their error rate. Afterwards, the fusion of all evidence is carried by combining masses using the conjunctive rule of combination. Finally, the pignistic transformation is applied to assign evidence to singleton sets that are the original zones. The zone having the highest evidence is then selected. Experiments on real data illustrate the performance of the belief functions framework for localization of sensors against other localization techniques.

2 Belief Functions Localization Method

2.1 Problem Formulation

The localization problem is tackled in the following manner. Let N_Z be the number of zones of the targeted area, denoted by Z_j , $j \in \{1, 2, \dots, N_Z\}$.

© Springer Nature Switzerland AG 2018

S. Destercke et al. (Eds.): BELIEF 2018, LNAI 11069, pp. 10–13, 2018.

https://doi.org/10.1007/978-3-319-99383-6_2

Let N_{AP} be the number of detected APs, denoted by AP_k , $k \in \{1, 2, \dots, N_{AP}\}$. Let $\rho_{j,k,r}$, $r \in \{1, \dots, \ell_j\}$, be the set of ℓ_j measurements collected in an offline phase in the zone Z_j with respect to AP_k . Let ρ_t be the vector of N_{AP} RSSI measurements collected by the mobile sensor at the instant t from all the APs. The aim of the proposed algorithm is to determine the zone $\hat{Z}_{j,t}$ having the highest evidence, $\hat{Z}_{j,t} = \arg \max_{Z_j} \mathcal{W}_t(Z_j)$, such that $\mathcal{W}_t(Z_j)$ represents the evidence in having the mobile sensor of observation ρ_t residing in the zone Z_j at instant t .

2.2 Mass Assignment

In the offline phase, the kernel density estimation (KDE) is proposed to model the distribution of the collected measurements $\rho_{j,k,r}$, $r \in \{1, \dots, \ell_j\}$, of each zone j according to each AP AP_k . The density estimate $Q_{KDE,Z_j,k}(\cdot)$ is calculated as,

$$Q_{KDE,Z_j,k}(\cdot) = \frac{1}{\ell_j \times h} \sum_{r=1}^{\ell_j} \mathcal{K} \left(\frac{\cdot - \rho_{j,k,r}}{h} \right), \quad (1)$$

where $\mathcal{K}(u)$ is a Gaussian kernel, and h its bandwidth,

$$\mathcal{K}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \quad (2)$$

A practical approach to determine h is to maximize the pseudo-likelihood leave-one-out cross validation,

$$ML(h) = \ell_j^{-1} \sum_{r=1}^{\ell_j} \log \left[\sum_{r' \neq r} \mathcal{K} \left(\frac{\rho_{j,k,r'} - \rho_{j,k,r}}{h} \right) \right] - \log[(\ell_j - 1)h]. \quad (3)$$

The computations are conducted in the same manner for all the supersets of the zones. Let A be a superset of zones. Then, the RSSIs related to all zones of A are considered to construct the kernel density estimate related to A , denoted $Q_{KDE,A,k}(\cdot)$ as in Eq. (1). In the online phase, once a new measurement $\rho_t = (\rho_{t,1}, \dots, \rho_{t,N_{AP}})$ is carried for localization, the kernel density estimates obtained in the offline phase is used with the belief functions theory to determine the zone of the sensor. Let \mathcal{Z} be the set of all possible zones Z_j , $j \in \{1, \dots, N_Z\}$, and let $2^{\mathcal{Z}}$ be the set of all supersets of \mathcal{Z} , i.e., $2^{\mathcal{Z}} = \{\{Z_1\}, \dots, \mathcal{Z}\}$. The mass function (MF) $m_{AP_k,t} : 2^{\mathcal{Z}} \rightarrow [0, 1]$, defined according to AP_k is calculated as follows [1],

$$m_{AP_k,t}(A) = Q_{KDE,A,k}(\rho_{t,k}). \quad (4)$$

2.3 Discounting Operation

The detected APs are not completely reliable. Indeed, each AP could yield an erroneous interpretation of evidence for some observations. In order to correct this, one can discount the MFs of Eq. (4) by taking into account the error rate

of the AP. The discounted MF $\alpha m_{AP_k,t}$ related to AP_k having an error rate α_k is deduced from $m_{AP_k,t}$ as follows [2],

$$\alpha m_{AP_k,t}(A) = \begin{cases} (1 - \alpha_k)m_{AP_k,t}(A), & \text{if } A \in 2^{\mathcal{Z}}, A \neq \mathcal{Z}; \\ \alpha_k + (1 - \alpha_k)m_{AP_k,t}(A), & \text{if } A = \mathcal{Z}. \end{cases} \quad (5)$$

By doing this, the amounts of evidence given to the subsets of \mathcal{Z} are reduced, and the remaining evidence is given to the whole set \mathcal{Z} . The source AP_k is assumed not reliable if, according to an observation $\rho_{k,\cdot}$ being truly in A , it associates more evidence to any set other than A , that is, the mass associated to A is less than the mass of another subset of $2^{\mathcal{Z}}$. Let $\epsilon_k(A)$ be the error rate related to the set A with respect to AP_k . Then,

$$\epsilon_k(A) = \int_{\mathbb{D}_{k,A}} Q_{KDE,A,k}(\rho) d\rho, \quad (6)$$

such that $\mathbb{D}_{k,A}$ is the domain of error of set A according to AP_k , defined as,

$$\mathbb{D}_{k,A} = \{\rho \mid Q_{KDE,A,k}(\rho) \leq \max_{A' \in 2^{\mathcal{Z}}, A' \neq A} (Q_{KDE,A',k}(\rho))\}. \quad (7)$$

The error rate α_k of AP_k is then the average error of all subsets according to this AP, namely

$$\alpha_k = \frac{\sum_{A \in 2^{\mathcal{Z}}} \epsilon_k(A)}{2^{|\mathcal{Z}|} - 1}. \quad (8)$$

2.4 Evidence Fusion

The evidence is then combined by aggregating the information coming from all the detected APs [3]. The mass functions can then be combined using the conjunctive rule of combination as follows,

$$m_{\cap,t}(A) = \sum_{\substack{A^{(k)} \in 2^{\mathcal{Z}} \\ \cap_k A^{(k)} = A}} \prod_{k=1}^{N_{AP}} \alpha m_{AP_k,t}(A^{(k)}), \quad (9)$$

$\forall A \in 2^{\mathcal{Z}}$, with $A^{(k)}$ is the subset A with respect to the Access Point AP_k .

2.5 Decision

An adequate notion of the BFT to attribute masses to singleton sets $A \in 2^{\mathcal{Z}}$ is the pignistic level [4]. It is defined as follows,

$$BetP_t(A) = \sum_{A \subseteq A'} \frac{m_{\cap,t}(A')}{|A'|}, \quad (10)$$

The zone $\hat{Z}_{j,t}$ having the highest evidence at instant t is then selected,

$$\hat{Z}_{j,t} = \arg \max_{Z_j} BetP_t(\{Z_j\}), j \in \{1, \dots, N_Z\}. \quad (11)$$

3 Experiments

Real experiments are conducted in the Living Lab of the University of Technology of Troyes, France. The considered floor of approximated area of 500 m² is partitioned into 19 zones, where 12 AP networks could be detected. A Set of 50 measurements is taken in each zone, of which 30 are randomly used to construct the databases, and the others are kept for test. The proposed approach is compared to other localization techniques such as weighted k -nearest neighbors algorithm (WKNN) presented in [5] and a Multinomial logistic regression (MLR) presented in [6]. The proposed method achieves an accuracy of 85.26% outperforming the WKNN with 83.82% and the MLR with 82.94%.

4 Conclusion and Future Work

This paper presented a belief functions framework for localization of sensors in indoor wireless networks. The kernel density estimation was used to set mass functions, and the belief functions theory combined evidence to determine the sensor's zone. Experiments on real data prove the effectiveness of the approach as compared to other localization techniques. Future work will focus on using the mobility as another source of information.

Acknowledgment. The authors would like to thank the European Regional Development Fund and Grand Est region in France for funding this work.

References

1. Alshamaa, D., Mourad-Chehade, F., Honeine, P.: A hierarchical classification method using belief functions. *Sig. Process.* **148**, 68–77 (2018)
2. Mercier, D., Lefèvre, É., Delmotte, F.: Belief functions contextual discounting and canonical decompositions. *Int. J. Approximate Reasoning* **53**(2), 146–158 (2012)
3. Fu, C., Yang, S.: The conjunctive combination of interval-valued belief structures from dependent sources. *Int. J. Approximate Reasoning* **53**(5), 769–785 (2012)
4. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. I. *J. Approximate Reasoning* **9**(1), 1–35 (1993)
5. Koyuncu, H., Yang, S.H.: A 2D positioning system using WSNs in indoor environment. *Int. J. Electr. Comput. Sci. IJECS-IJENS* **11**(3), 70–77 (2011)
6. Liu, D., Li, T., Liang, D.: Incorporating logistic regression to decision-theoretic rough sets for classifications. *Int. J. Approximate Reasoning* **55**(1), 197–210 (2014)



On Evidential Clustering with Partial Supervision

Violaine Antoine¹(✉), Kévin Gravouil^{1,2}, and Nicolas Labroche³

¹ Clermont Auvergne University, UMR 6158, LIMOS,
63000 Clermont-Ferrand, France
`violaine.antoine@uca.fr`

² Clermont Auvergne University, INRA, MEDIS, LMGE,
63000 Clermont-Ferrand, France
`kevin.gravouil@uca.fr`

³ University of Tours, LIFAT, EA 6300, Blois, France
`nicolas.labroche@univ-tours.fr`

Abstract. This paper introduces a new semi-supervised evidential clustering algorithm. It considers label constraints and exploits the evidence theory to create a credal partition coherent with the background knowledge. The main characteristics of the new method is its ability to express the uncertainties of partial prior information by assigning each constrained object to a set of labels. It enriches previous existing algorithm that allows the preservation of the uncertainty in the constraint by adding the possibility to favor crisp decision following the inherent structure of the dataset. The advantages of the proposed approach are illustrated using both a synthetic dataset and a real genomics dataset.

Keywords: Evidential clustering · Partial labels
Semi-supervised clustering · Belief function

1 Introduction

Evidential clustering algorithms, such as ECM [1], rely on the theoretical foundation of belief functions and evidence theory [2] and allow to express many types of uncertainty about the assignment of an object to a cluster. It enables to handle crisp single cluster assignment, as well as cluster membership degrees, total ignorance and outliers detection. The credal partition, which is formed with the assignments of all the objects, generalizes other soft partitions such as fuzzy, possibilistic or rough partitions [3].

Clustering is a complex unsupervised task that often requires additional assumptions to determine relevant solutions. The performances of a clustering algorithm can be highly improved by using background knowledge [4]. To this end, several semi-supervised evidential clustering approaches have been proposed [5–7]. In [7], the SECM-pl algorithm integrates prior information in the form of labeled data instances. The particularity of SECM-pl is its ability to

handle partial knowledge, which corresponds to the uncertainty about the assignment of an object to several classes. This partial knowledge is controlled by the algorithm in such a way that the uncertainty can be preserved.

In this paper, we propose an approach that generalizes SECM-pl, which maintains a high flexibility on the constraints, by favoring a decision making on the constraints. The paper is organized as follows: Sect. 2 recalls the basics concerning the evidence theory and its application in clustering. Section 3 details the novel SECM algorithm and focuses on how labels constraints are expressed and incorporated in ECM. Section 4 presents experimental settings and results. Finally a discussion and future work are presented in Sect. 5.

2 Preliminaries

2.1 Belief Functions

The evidence theory (or belief functions theory) [2, 8] is a mathematical framework that enables to reflect the state of partial and unreliable knowledge. Let $\Omega = \{\omega_1, \dots, \omega_c\}$ be the frame of discernment where ω_i is the true state of the system which will be defined below. The mass function $m : 2^\Omega \rightarrow [0, 1]$, also called basic belief assignment (bba), measures the degree of belief that ω_i belongs to a subset $A \subseteq \Omega$. It satisfies $\sum_{A \subseteq \Omega} m(A) = 1$. Any subset A such that $m(A) > 0$ is named a focal set of m . Given a mass function m , the plausibility function $pl : 2^\Omega \rightarrow [0, 1]$ is defined by:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega. \quad (1)$$

The quantity $pl(A)$ corresponds to the maximal degree of belief that could be given to A . To make a decision, a mass function can be transformed into a pignistic probability distribution *BetP* [8].

2.2 Evidential C-Means

Evidential clustering algorithms generate for each object $\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n \in \mathbb{R}^p$ a mass function \mathbf{m}_i on the set $\Omega = \{\omega_1, \dots, \omega_c\}$ denoting the clusters. The collection $\mathbf{M} = (\mathbf{m}_1, \dots, \mathbf{m}_n)$ forms the credal partition and allows to represent the uncertainties and imprecisions regarding the class membership of each object. ECM [1] is the credibilistic version of Fuzzy C-Means [9]. It considers for each subset $A_j \subseteq \Omega$ a representation of the subset with a prototype vector \mathbf{v}_j in \mathbb{R}^p . The objective function is:

$$J_{ECM}(\mathbf{M}, \mathbf{V}) = \sum_{i=1}^n \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^n \rho^2 m_{i\emptyset}^\beta, \quad (2)$$

where \mathbf{V} is the collection of prototypes, $m_{ij} = m_i(A_j)$ corresponds to the bba of the object \mathbf{x}_i for the subset A_j , $m_{i\emptyset}$ denotes the mass of \mathbf{x}_i allocated to the

empty set and d_{ij}^2 represents the squared Euclidean distance between \mathbf{x}_i and the prototype \mathbf{v}_j . The last term of the objective function enables to handle the empty set which can be interpreted as a cluster for outliers. The ρ parameter is a fixed coefficient representing the distance between any object and the empty set. The two parameters α and $\beta > 1$ are introduced to penalize the degree of belief assigned to subsets with a high cardinality and to control the fuzziness of the partition. The objective function is subject to

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ik} + m_{i\emptyset} = 1; \quad m_{ij} \geq 0 \quad \forall i = \{1, \dots, n\}, \forall j/A_j \subseteq \Omega. \quad (3)$$

2.3 SECM-pl

The main idea of the algorithm [7] is to add a penalty term in the objective function of ECM, in order to take into account a set of already labeled objects. Any mass function which partially or fully respects a constraint on a label ω_k has a high plausibility $pl(\omega_k)$ given to the label. Similarly, an object constrained in several classes, i.e. on the set $A_j \subset \Omega$ is respected with mass functions given a high plausibility $pl(A_j)$. Thus, the following penalty term has been proposed:

$$J_S = \sum_{i=1}^n \sum_{A_j \subset \Omega, A_j \neq \emptyset} b_{ij}(1 - Pl_i(A_j)), \quad (4)$$

where $b_{ij} = 1$ if \mathbf{x}_i is constrained on A_j and 0 otherwise.

3 New ECM Algorithm with Partial Supervision

3.1 Modeling the Constraints

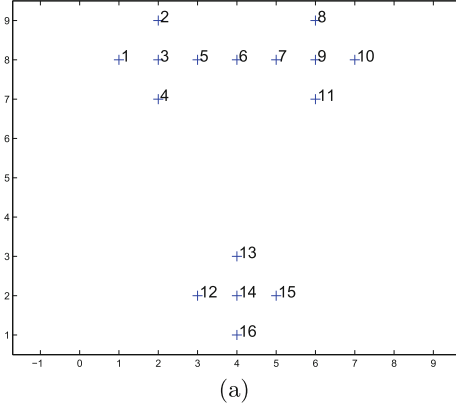
Let us consider a set of partially labeled constraints, i.e. a collection of objects \mathbf{x}_i such that $\mathbf{x}_i \in A_j, \forall A_j \neq \emptyset$. If A_j is a singleton, then the object i belongs to a class with certainty. Otherwise, \mathbf{x}_i belongs to a class listed in A_j without knowing which one more precisely. Notice that $\mathbf{x}_i \in \Omega$ corresponds to complete ignorance concerning the class of the object i . Degrees of belief containing the set of clusters A_j or a part of it should be favored as well as mass functions of subsets with a low cardinality. Thus, we define the measure $1 \geq T_{ij} \geq 0$ by the following formula:

$$T_{ij} = T_i(A_j) = \sum_{A_j \cap A_l \neq \emptyset} \frac{|A_j \cap A_l|^{\frac{r}{2}}}{|A_l|^r} m_{il}, \quad \forall i \in \{1 \dots n\}, A_j \subseteq \Omega, \quad (5)$$

where $r \geq 0$ controls a degree of penalization of the subsets. The coefficient $|A_l|^r$ is used to penalize subsets with a high cardinality and $|A_j \cap A_l|^{\frac{r}{2}}$ allows to concentrate efforts on subsets containing mostly elements of A_j . Notice that when $r = 0$, T_{ij} corresponds to the plausibility that the object \mathbf{x}_i belongs to A_j . For the rest of the paper, we set $r = 1$.

3.2 Illustration

The behavior of the new measure T_{ij} is illustrated with the DiamondK3 dataset presented Fig. 1(a). This dataset is composed of 15 objects that should be separated into 3 groups. As it can be observed, points 13 to 16 are well isolated, whereas objects 1 to 11 seem to correspond to two natural clusters connected by the object 6. Let us suppose that some partial knowledge is available: e.g. object 6 is in the cluster ω_1 and object 13 belongs either to ω_1 or to ω_3 , but not to ω_2 . Thus, we obtain the two following constraints: $\mathbf{x}_6 \in \{\omega_1\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_3\}$.



$m_i(A_j)$	\emptyset	0 0 0 0 0
	ω_1	1 0 0 0 0
	ω_2	0 0 0 0 1
	$\{\omega_1, \omega_2\}$	0 0 1 0 0
	ω_3	0 0 0 0 0
	$\{\omega_1, \omega_3\}$	0 1 0 0 0
	$\{\omega_2, \omega_3\}$	0 0 0 0 0
Ω	0 0 0 1 0	
$T_i(A_j)$	ω_1	1 $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{3}$ 0
	$\{\omega_1, \omega_3\}$	1 $\frac{\sqrt{2}}{2}$ $\frac{1}{2}$ $\frac{1}{3}$ 0
$Pl(A_j)$	ω_1	1 1 1 1 0
	$\{\omega_1, \omega_3\}$	1 1 1 1 0

(b)

Fig. 1. DiamondK3 dataset (a) and illustration of the proposed penalty term $T_i(A_j)$ when considering several possible mass functions and compared to penalty term based on plausibility $Pl(A_j)$ for previous SECM-pl [7] (b).

Figure 1(b) presents in each column a set of possible mass functions for an object \mathbf{x}_i coming from the DiamondK3 dataset. First, let us consider that $\mathbf{x}_i = \mathbf{x}_6$ and let us assume that $m_6(\omega_1) = 1$ as shown in the first column of Fig. 1(b). Thus, the constraint is respected and it can be observed that $T_6(\omega_1) = 1$. Inversely, if $m_6(\omega_2) = 1$ as presented in the last column of Fig. 1(b), then the constraint is totally neglected and $T_6(\omega_1) = 0$. Other columns illustrate partial respect of the constraint, since the bba is allocated to subsets containing the label ω_1 . The larger the cardinality of the subset, the lower the value of T_{ij} .

Let us assume that $\mathbf{x}_i = \mathbf{x}_{13}$ and let us focus on the value obtained by $T_i(\{\omega_1, \omega_3\})$ for the set of possible mass functions. As it can be observed, $T_{ij} = 0$ when no focal sets contain ω_1 and/or ω_3 . Conversely, if there exists a degree of belief not null on a subset including at least one of the classes included in the constraint, then $T_{ij} > 0$. As previously, the larger the cardinality of the subset, the lower the value of T_{ij} . For the same amount of subsets, for example columns 2 and 3 in Fig. 1(b), a higher value is given to subsets containing the most of classes in the constraint, i.e. $\{\omega_1, \omega_3\}$. This is a significant difference with the plausibility measure for which all subsets intersecting with the constraints contribute equally to the final value.

3.3 Objective Function and Optimization

Based on the mass function m_i of an object i , we can quantify the degree to which a partial constraint $\mathbf{x}_i \in A_j$ is respected by computing T_{ij} in Eq. (5). $T_{ij} = 1$ when the belief is given to a cluster in A_j and is 0 when the belief is assigned to none of the clusters included in A_j , i.e. when the constraint is not respected. If we consider now that the bbas have to be found, a natural requirement is to obtain a value of T_{ij} as high as possible if there exists a constraint such that $\mathbf{x}_i \in A_j$. This goal is achieved by minimizing the following objective function:

$$J_{SECM}(M, V) = (1 - \gamma) \frac{1}{2^{cn}} J_{ECM}(M, V) + \gamma \frac{1}{s} \sum_{i=1}^n \sum_{A_j \subset \Omega, A_j \neq \emptyset} b_{ij} (1 - T_{ij}), \quad (6)$$

such that constraints (3) are respected, s corresponds to the number of constraints, and $b_{ij} = 1$ if $\mathbf{x}_i \in A_j$, i.e if the object i is constrained with A_j and 0 otherwise.

The coefficient γ controls the tradeoff between the objective function of ECM and the constraints. Notice that if $r = 0$ for the computation of T_{ij} , then J_{SECM} is identical to the objective function proposed in [7]. Such setting allows the penalty term to give equal importance to any subset intersecting with the constraints, whereas $r > 0$ favors subsets with low cardinality. As ECM, the credal partitioning is carried out through an iterative optimization of the objective function, with the update of the mass functions and the prototypes. If β is set to 2, then the problem becomes quadratic with linear constraints and can be resolved with classical methods, for instance [10].

4 Experimentations

4.1 Toy Example

To illustrate the behavior of the SECM algorithm, we used the DiamondK3 dataset. First, an execution of ECM is performed with $\alpha = 1$, $\beta = 2$, $\rho^2 = 10^3$ and the final mass functions for the most representative subsets varying with the objects number are presented Fig. 2(a). It can be seen that ECM identifies the 3 clusters by assigning the belief to the 3 singletons. The object 6, which is located between the cluster ω_1 and ω_2 , is ambiguous as it can belong to either ω_1 or ω_2 . Thus, ECM assigns for \mathbf{x}_6 a high mass for the subset $\{\omega_1, \omega_2\}$.

Let us consider now that the following set of constraints are available: $\mathbf{x}_5 \in \{\omega_1\}$, $\mathbf{x}_6 \in \{\omega_2\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_2\}$. The SECM algorithm is executed with $\gamma = 0.5$ and the credal partition obtained is presented Fig. 2(b). As it can be observed, constraints are well respected. The object 6, previously ambiguous with the ECM algorithm, is now assigned with certainty to ω_2 . Similarly, the object 5 had with ECM its belief divided into $\{\omega_1, \omega_2\}$ and ω_1 , whereas now all its belief is given to $\{\omega_1\}$. Finally, the mass function $m_{13}(\omega_3)$ for the object 13, which is already high with ECM, has increased with SECM. It shows that SECM is able to constrained \mathbf{x}_{13} more specifically on ω_3 following the inherent structure of the dataset.

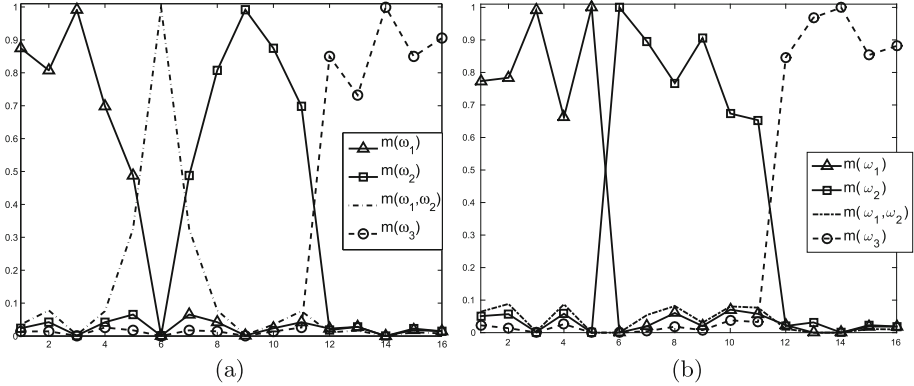


Fig. 2. Credal partitions obtained for DiamondK3 with (a) ECM and (b) SECM such that $\mathbf{x}_6 \in \{\omega_1\}$ and $\mathbf{x}_5 \in \{\omega_2\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_3\}$.

4.2 Genomics Application

Dataset: Dozens of thousands microorganism’s genomes are available in public databases. We selected three known genomes from the RefSeq database [11], namely *Clostridium acetobutylicum*, *Bacillus cereus* and *Brachyspira hyodysenteriae*, to simulate a small microbial community. DNA sequences were extracted from these genomes then embedded in numerical vectors using normalized tetranucleotide frequencies with a CONCOCT-inspired approach [12]. The final dataset, called tetragen, is composed of 22 attributes and 1188 objects corresponding to DNA sequences. Classes, i.e. the genomes *B. hyodysenteriae*, *C. acetobutylicum* and *B. cereus* contain respectively 288, 383 and 517 instances. In order to obtain the tetragen dataset, the largest DNA sequences were divided in several objects. We took benefit of this process to create label constraints: we assigned two DNA sequences composed of 13 and 21 objects in the subsets $\{B. cereus\}$ and $\{B. cereus, B. hyodysenteriae\}$ respectively. As a consequence, we obtained a dataset composed of 2.9% of constrained objects. Figure 3 presents the class and prior information used for the tetragen dataset.

Experimental Protocol: For both ECM and SECM, we performed 10 executions with random initialization of the centroids and kept the credal partition giving the minimum value for the objective function. To synthesize the information provided by the partitions, we transformed them into hard credal partitions by assigning each object to the subset of classes with the highest mass. Figures 4(a) and (b) illustrates the obtained results. As it can be observed, constraints helped SECM to impact the boundary of ω_3 .

In order to compare the methods, partitions obtained with ECM and SECM were transformed into hard partitions by selecting the cluster with the maximal pignistic probability. Then, their agreement with the real partition were measured with the Adjusted Rand Index (ARI) [13] and the Normalized Mutual Information (NMI). Both of them provide a 1 value when the partitions totally

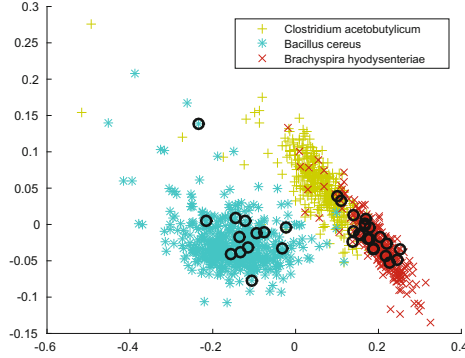


Fig. 3. Real classes (color) and constrained objects (encircled) for the tetragen data set. (Color figure online)

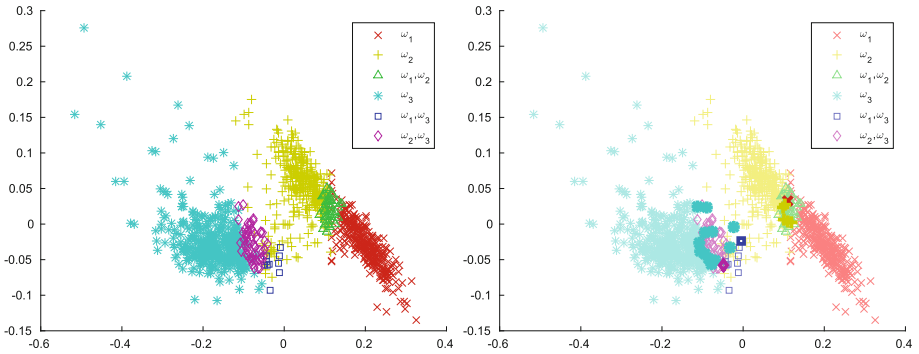


Fig. 4. Hard credal partition obtained with (a) ECM and (b) SECM for tetragen. Colors are lightened in (b) for objects for which the assignment has not changed between the two algorithms. (Color figure online)

match. With ECM, we obtained $\text{ARI}=0.75$ and $\text{NMI}=0.71$ whereas SECM gives an $\text{ARI}=0.78$ and a $\text{NMI}=0.73$. It shows that a few number of constrained objects, even partially labeled, can lead our clustering algorithm to a better result than ECM.

5 Conclusion

In this paper, a new semi-supervised clustering algorithm called SECM is proposed. It generalizes previous approach [7] based on partial label constraints. The new penalty term can be parameterized to favor either any credal partition for which constraints are still plausible or only credal partitions for which constrained objects have belief on subsets with low cardinalities. A proof of concept is provided and shows the benefits of the new algorithm. Finally, a real test

is performed on genomics data set and shows the necessity of such expressive approaches in real use case.

In the future, extensive tests on real and synthetic datasets should be conducted in order to show the influence of the parameter r and to compare various semi-supervised clustering algorithms. The genomics use case should also be developed as it offers a relevant testbed for partial user knowledge integration. A further work is to scale SECM for larger datasets, in order to apply the algorithm in a real genomics application.

References

1. Masson, M.H., Denœux, T.: ECM: an evidential version of the fuzzy c-means algorithm. *Patt. Recogn.* **41**(4), 1384–1397 (2008)
2. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
3. Denœux, T., Kanjanatarakul, O.: Beyond fuzzy, possibilistic and rough: an investigation of belief functions in clustering. In: Ferraro, M.B., et al. (eds.) *Soft Methods for Data Science*. AISC, vol. 456, pp. 157–164. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-42972-4_20
4. Basu, S., Davidson, I., Wagstaff, K.: *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. CRC Press, Boca Raton (2008)
5. Antoine, V., Quost, B., Masson, M.H., Denœux, T.: CECM: constrained evidential c-means algorithm. *Comput. Stat. Data Anal.* **56**, 894–914 (2012)
6. Antoine, V., Quost, B., Masson, M.H., Denœux, T.: Evidential clustering with instance-level constraints for proximity data. *Soft Comput.* **18**(7), 1321–1335 (2014)
7. Antoine, V., Labroche, N., Vu, V.V.: Evidential seed-based semi-supervised clustering. In: *Soft Computing and Intelligent Systems (SCIS)*, Kitakyushu, Japan, pp. 706–711. IEEE, December 2014
8. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**, 191–234 (1994)
9. Bezdek, J.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York (1981)
10. Ye, Y., Tse, E.: An extension of Karmarkar’s projective algorithm for convex quadratic programming. *Math. Program.* **44**(1), 157–179 (1989)
11. Pruitt, K., Tatusova, T., Maglott, D.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**(Suppl. 1), D61–D65 (2006)
12. Alneberg, J., et al.: Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**(11), 1144 (2014)
13. Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985)



Exploiting Domain-Experts Knowledge Within an Evidential Process for Case Base Maintenance

Safa Ben Ayed^{1,2(✉)}, Zied Elouedi¹, and Eric Lefevre²

¹ LARODEC, Institut Supérieur de Gestion de Tunis,
Université de Tunis, Tunis, Tunisia
zied.elouedi@gmx.fr

² LGI2A, Univ. Artois, EA 3926, 62400 Béthune, France
safa.ben.ayed@hotmail.fr, eric.lefevre@univ-artois.fr

Abstract. Case Base Maintenance (CBM) presents one of the key factors success for Case Based Reasoning (CBR) systems. Thence, several CBM policies are proposed to improve their problem-solving performance and competence. However, to the best of our knowledge, all of them are not able to make use of prior knowledge which can be offered by domain experts, especially that CBR is widely applied in real-life domains. For instance, given symptoms of two different cases in medicine area, the doctor can affirm that these two cases should never follow the same treatment, or conversely. This kind of prior knowledge is presented in form of *Cannot-Link* and *Must-link* constraints. In addition, most of them cannot manage uncertainty in cases during CBM. To overcome this shortcoming, we propose, in this paper, a CBM policy that handles constraints to exploit experts' knowledge during case base learning along with managing uncertainty using the belief function theory. This new CBM approach consists mainly in noisy and redundant cases deletion.

1 Introduction

Case Based Reasoning is a methodology for reasoning through adapting previous experiences to solve new problems. Each success solving operation will be retained for future learning, where an incremental aspect characterizes the case bases evolution [1]. As CBR systems are widely applied within real-life domains, and as they are designed to work over long time frames, the Case Base Maintenance (CBM) becomes a fundamental task to guarantee their success. In fact, CBM has been defined as the field that cares on implementing policies that aim to reach a particular set of performance objectives through revising the content and the organization of case bases [2]. Indeed, we note a great interest within current research that addresses issues for growing case bases. For instance, CBM policies may be divided into two strategies, even to the optimization strategy where the deletion is done after optimizing a given evaluation criterion, or to the partition strategy which allows to treat a set of small case bases independently. In the latter strategy, uncertainty about the membership of cases to the

different classes (clusters) have also been handled [3,4]. However, these CBM policies are not offering the possibility to exploit background knowledge which can be provided by an expert of domain in which the CBR system is deployed. Therefore, we aim, in this paper, to propose a new CBM approach based on an evidential clustering to manage uncertainty about the membership of cases. Moreover, this approach handles extra-information for cases clustering presented in the form of two types of constraints [5]: *Must-link* constraints which specify that two cases have the same solution and *Cannot-link* constraints which specify that two solutions cannot belong to the same cluster. To do, we used then the Constrained Evidential C-Means algorithm (CECM) [6]. The remainder of this paper is organized as follows. Section 2 reviews briefly some CBM approaches based on clustering techniques. Section 3 describes the used constrained evidential clustering technique called CECM. Our new CBM approach will be detailed in Sect. 4. Throughout Sect. 5, we discuss experimental settings, the pairwise constraints generation, testing strategy, and results.

2 Clustering-Based CBM Policies

Intuitively, when addressing the problem of maintaining a large case base, its decomposition into a number of related closely cases groups appears to be a good solution for their maintenance. Indeed, clustering techniques have been well applied within CBR since the notions of neighborhood and distances between cases are well presented. Actually, there are several works in this way. However, during the rest of this Section, two of them which handle uncertainty regarding the membership of cases to different clusters will be reviewed. The first one is called SCBM noting “Soft case base maintenance method based on competence model” which groups cases within the frame of fuzzy sets theory [7]. Then, it tries to detect the right case types to be removed without decreasing the competence of the CBR system. The second policy is named ECTD for “Evidential Clustering and case Types Detection for case base maintenance” which is more able to manage uncertainty using the belief function theory [8,9]. First, ECTD applies ECM [10] algorithm to group cases and obtain the credal partition of cases along with the different clusters centers. Then, it reasons on the way of detecting four types of cases in order to be able at the end to eliminate noisiness and redundancy. However, techniques used inside these methods do not allow to make use of the background knowledge that helps to guide to the best solution. For this paper, we consider prior knowledge in form of Must-link and Cannot-link constraints. To do, we apply on the case base a constrained evidential clustering technique as presented in the following Section.

3 Constrained Evidential Clustering Technique: CECM

When dealing with clustering-based CBM policies, it is gainful to express prior knowledge in form of instance level constraints as indicated in the Introduction. In what follows, we will present CECM through its constraints expression and work standard.

3.1 Constraints Expression by CECM

Let two objects \mathbf{o}_i and \mathbf{o}_j and their associated mass functions m_i and m_j . The mass function $m_{i \times j}$ regarding their joint class membership may be calculated in the Cartesian product $\Omega^2 = \Omega \times \Omega$, as the combination between m_i and m_j [11] such that:

$$m_{i \times j}(A \times B) = m_i(A) m_j(B), \quad A, B \subseteq \Omega, A \neq \emptyset, B \neq \emptyset \quad (1a)$$

$$m_{i \times j}(\emptyset) = m_i(\emptyset) + m_j(\emptyset) - m_j(\emptyset) m_j(\emptyset) \quad (1b)$$

Let the subset $\theta = \{(\omega_1, \omega_1), (\omega_2, \omega_2), \dots, (\omega_c, \omega_c)\}$ in Ω^2 (where c is the number of classes) presents the event “*The pair of objects \mathbf{o}_i and \mathbf{o}_j belong to the same class*”. Therefore, after calculating the plausibility $pl_{i \times j}$ from $m_{i \times j}$, the value $pl_{i \times j}(\theta) = 0$ corresponds to a Cannot-link constraint (\mathcal{C}) between \mathbf{o}_i and \mathbf{o}_j and the value $pl_{i \times j}(\bar{\theta}) = 0$ corresponds to a Must-link constraint (\mathcal{M}) between \mathbf{o}_i and \mathbf{o}_j .

3.2 Objective Function and Optimization of CECM

First of all, let mention that CECM [6] is a variant of ECM [10] algorithm (noisiness is assigned to the empty set partition). The principle of both of them during the evidential clustering is to minimize an objective function in order to maximize distances between objects belonging to different classes and minimizing those belonging to the same one. The objective function for ECM algorithm is defined such that:

$$J_{ECM}(M, V) = \frac{1}{2^c n} \left[\sum_{i=1}^n \sum_{A_k \neq \emptyset} |A_k|^\alpha m_{ik}^\beta d_{ik}^2 + \sum_{i=1}^n \rho^2 m_{i\emptyset}^\beta \right] \quad (2)$$

subject to:

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, \dots, n \quad (3)$$

where M represents the credal partition of n objects to c clusters, V presents 2^c clusters centers, d_{ij} represents a given distance between \mathbf{o}_i and \mathbf{o}_j , ρ and β are two parameters to treat noisy objects, and the coefficient α controls the penalization of degree's allocation to subsets with high cardinality.

CECM algorithm shares the same standard of ECM with an additional requirement that $pl_{i \times j}(\theta)$ (respectively $pl_{i \times j}(\bar{\theta})$) should be as low as possible if $(\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{C}$ (respectively $(\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{M}$). Consequently, its objective function to be minimized is defined such that:

$$J_{CECM}(M, V) = (1 - \xi) J_{ECM}(M, V) + \xi J_{CONST} \quad (4)$$

where the parameter ξ controls the balance between constraints and geometrical model, and J_{CONST} , which indicates \mathcal{C} and \mathcal{M} violating cost, is defined such that:

$$J_{CONST} = \frac{1}{|\mathcal{M}| + |\mathcal{C}|} \left[\sum_{(\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{M}} pl_{i \times j}(\bar{\theta}) + \sum_{(\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{C}} pl_{i \times j}(\theta) \right] \quad (5)$$

To minimize Eq. 4, an alternate optimization scheme has been proposed in [6] aiming to fix the partition matrix M and the centroid matrix V . Furthermore, CECM with adaptive metric (Mahalanobis distance) is proposed to support arbitrary shapes of clusters. More details of optimization will be found on [6].

4 Maintaining Case Bases Through Constrained Evidential Clustering and Case Types Detection (CECTD)

In this Section, we present the different steps of our CBM approach. To build our case base maintainer, our method applies the constrained evidential clustering analysis, detects cases that should be eliminated from the case base, and performs the maintenance.

4.1 Case Bases Clustering with Background Knowledge

First, we perform on case bases the CECM constrained evidential clustering as presented in Sect. 3, where each object is considered as a case and its class presents the solution part of that case. The background knowledge is presented as case-level constraints. Actually, CECM algorithm manages uncertainty by offering clusters centers along with the credal partition which provides the belief degree of cases membership to the different partitions. These two outputs are the source of case types detection strategy.

4.2 Case Types Detection

Several works on the CBM field divide cases into different types according to their role towards to whole case base or their competence for other problems resolution. In this paper, we classify cases into four types [3, 4] such that:

- *Noisy cases*: They present a distortion of values and cannot be correctly classified in any one of clusters.
- *Similar cases*: They present a number of cases which are so close that they are considered as redundant.
- *Isolated cases*: They are dissimilar and situated in clusters borders.
- *Internal cases*: They present the center of each group of similar case.

Detect Noisy Cases. Since CECM algorithm allocates a high belief's degree to the empty set for noisy cases, we propose, as in [4], to detect them such that:

$$\mathbf{x}_i \in NC \text{ iff } m_i(\emptyset) > \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_i(A_j) \quad (6)$$

where \mathbf{x}_i presents one case and NC represents the set of all the Noisy cases.

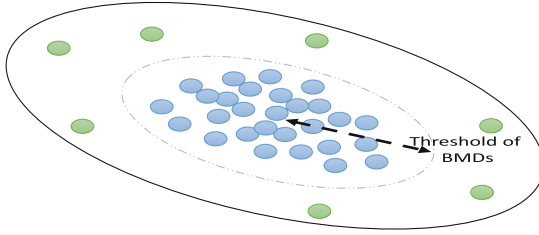


Fig. 1. Distinguish between similar and isolated cases within a cluster using a threshold

Distinguish Between Similar and Isolated Cases. Let c clusters are obtained after cases clustering step. Logically, the majority of cases are situated in the core of each cluster (Similar cases). However, we find some cases which are isolated and far somehow to the cluster’s center (Isolated cases). To distinguish between these two types, we compare cluster-case distance to a given threshold (Th_k) which has been defined as the mean of all cases distances to a given cluster’s center (see Fig. 1). To calculate the distance between a case and cluster’s center, we chose to use the following Belief Mahalanobis Distance (BMD) [4]:

$$BMD(\mathbf{x}_i, \mathbf{v}_k) = \sqrt{(\mathbf{x}_i - \mathbf{v}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mathbf{v}_k)} \quad (7)$$

where \mathbf{v}_k is the k^{th} cluster’s center generated by CECM, and Σ_k presents the *Belief Covariance Matrix* which has been presented in [6] as follows:

$$\Sigma_k = \sum_{i=1}^n \sum_{A_j \ni w_k, A_j \subseteq \Omega} m_{ij}^2 |A_j|^{\alpha-1} (\mathbf{x}_i - \bar{\mathbf{v}}_j)(\mathbf{x}_i - \bar{\mathbf{v}}_j)^T \quad (8)$$

where k is the cluster’s number with $k = 1, \dots, c$, m_{ij} and $\bar{\mathbf{v}}_j$ are respectively the credal partition and their prototypes defined by CECM.

Ultimately, we distinguish between Similar and Isolated cases such that:

$$\mathbf{x}_i \in \begin{cases} SC_k & \text{if } \exists k / BMD(\mathbf{x}_i, \mathbf{v}_k) < Th_k \\ IsC & \text{Otherwise} \end{cases} \quad (9)$$

where SC_k is the set of similar cases, IsC is the set of Isolated ones and the threshold Th_k is defined such that:

$$Th_k = \frac{\sum_{\mathbf{x}_i \notin NC} BMD(\mathbf{x}_i, \mathbf{v}_k)}{\#TotalCases - \#NoisyCases} \quad (10)$$

Flag Internal Cases. From each group of Similar cases, we have to flag an internal case as a representative for covering all of them. Hence, we choose to detect this case as the closest one to each cluster’s center using BMD. Hence, they can be formally defined such that:

$$\mathbf{x}_i \in InC \text{ iff } \exists k; \neg \exists \mathbf{x}_j / BMD(\mathbf{x}_j, \mathbf{v}_k) < BMD(\mathbf{x}_i, \mathbf{v}_k) \quad (11)$$

where \mathbf{x}_i and \mathbf{x}_j are two cases, and InC represents the set of Internal cases.

4.3 Case Base Maintenance

While maintenance, we aim to remove cases that are dispensable or distorting the problem-solving process. Through this idea, we remove cases detected as Similar in order to eliminate redundancy and improve performance, as well as Noisy cases so as to improve the competence of CBR systems in problem resolution.

5 Experimental Study Using Artificial Constraints

During this Section, we aim to differently generate the pairwise *Must-link* and *Cannot-link* constraints, as well as to validate our new CBM method benefit.

5.1 Experimental Setting

Our new CBM approach has been developed using R-3.3.2 and it is tested on a number of numeric case bases from UCI Repository which are described in Table 1 by their references, number of attributes, size, number of classes and their classes distribution. While developing, default values are taken for the CECM parameters, and the number of clusters and classes were equally taken. Besides, we used CECM with adaptive metric to consider arbitrary clusters' shape.

Table 1. UCI data sets used in our experimental study

Case base	Reference	Attributes	Instances	Classes	Class distribution
Sonar	SN	60	208	2	97/111
Ionosphere	IO	34	351	2	226/125
Heberman	HB	3	306	2	225/81
Seeds	SD	7	210	3	70/70/70
Mammographic	MM	6	961	2	516/445
Banknote authentication	BA	5	1372	2	762/610

5.2 Pairwise Constraints Generation

The aim of this subsection is to implement two different ways for artificially-generating constraints in conjunction with experiments applied on our method. The idea consists in randomly picking two cases. If they are classified with high degree of certainty ($m_i(A) > 0.5$ with A is a singleton partition), we generate a constraint through their solution (If they have the same solution, we create a Must-link constraint, otherwise we generate a Cannot-Link constraint). Therefore, we perform the following two ways:

- Batch constraints generation (CECTD_{bat}): Apply ECM algorithm (CECM without constraint), generate a number of constraints equal to 10% of the case base size. Then, apply our CECTD method.

- Alternate constraints generation (CECTD_{alt}): Within the first step of our method, we alternate between running CECM and generating randomly one constraint having high degree of certainty, until reaching 10% of constraints.

5.3 Maintenance Testing Strategy

To measure the effectiveness of our maintaining method, we track the following testing strategy. Each case base is divided into Training set (T_r) and Test set (T_s), and we apply our maintaining method on T_r to obtain a modified Training set (T'_r). Then, we compute three evaluation criteria as follows:

1. Classify T_s from T'_r using 1-Nearest Neighbor algorithm. Therefore, the classification accuracy to measure the performance is calculated such that:

$$PCC(\%) = \frac{\# \text{ correct classifications on } T_s}{\text{size of } T_s} \times 100$$

2. Measure the Retrieval Time (RT) as the time spent to classify all cases' instances in T_r using 1-NN.
3. Calculate the storage size as the data Retention Rate (RR) of T_r comparing to T'_r as follows:

$$RR (\%) = \frac{\text{size of } T'_r}{\text{size of } T_r} \times 100$$

The final estimation of each evaluation criterion is obtained by averaging ten trials values using 10-Folds cross validation technique.

5.4 Experimental Results

According to the evaluation criteria mentioned above, we compare our method with its two different ways to generate constraints (CECTD_{bat} and CECTD_{alt}) to the Initial case base (ICBR) as well as to ECTD method [4]. Results are therefore shown in Tables 2 and 3. Obviously, we tolerate some degradation in accuracy after maintenance at the aim of accelerating cases retrieving task and improving CBR systems performance. Nevertheless, Table 2 shows some improvements in accuracy especially with the alternate version of our approach. For instance, it moves from 80.78% to 82.10% after applying CECTD_{alt}. In parallel, Table 3 presents, in term of cases retention rate and retrieval time, how our approach can notably boost CBR systems. Herein, we note that we were able to reduce more than half of all case bases. For example, ‘‘Heberman’’ dataset were reduced by CECTD_{alt} until almost quarter. Moreover, even with using 1-NN for classification, we clearly note the improvement of retrieval time values particularly comparing to the Initial non-maintained case base, where all of them move from about 0.1 s to about 0.001 s.

Table 2. Accuracy evaluation (%)

Case bases	ICBR	ECTD	CECTD _{bat}	CECTD _{alt}
SN	80.78	68.31	79.78	82.10
IO	85.47	79.45	85.00	84.90
HB	72.88	67.23	70.85	72.88
SD	90.00	83.16	88.70	90.18
MM	79.81	72.13	80.01	79.92
BA	99.12	86.40	88.97	95.14

Table 3. Data Retention Rate (%) and Retrieval Time (s) evaluation

CB	ICBR		ECTD		CECTD _{bat}		CECTD _{alt}	
	RR	RT	RR	RT	RR	RT	RR	RT
SN	100	0.1003	48.98	0.0021	48.50	0.0026	46.51	0.0020
IO	100	0.0094	37.04	0.0017	35.36	0.0017	33.89	0.0015
HB	100	0.0993	29.72	0.0027	34.52	0.0021	28.14	0.0019
SD	100	0.0911	44.13	0.0023	45.77	0.0018	43.98	0.0016
MM	100	0.0852	26.23	0.0014	39.57	0.0016	40.02	0.0022
BA	100	0.1033	31.82	0.0026	44.54	0.0036	39.15	0.0027

6 Conclusion

Aiming at the performance and learning capability issues that the growing scale of CBR case bases brings, a new CBM approach based on a constrained evidential clustering technique has been developed, in this paper, using two ways for constraints generation with managing uncertainty. Better results are offered, during experiments, when generating constraints one by one alternatively with running CECM.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. In: AI Communications, pp. 39–59. IOS Press (1994)
2. Wilson, D.C., Leake, D.B.: Maintaining case-based reasoners: dimensions and directions. In: Computational Intelligence, pp. 196–213 (2001)
3. Smiti, A., Elouedi, Z.: SCBM: soft case base maintenance method based on competence model. J. Comput. Sci. (2017). <https://doi.org/10.1016/j.jocs.2017.09.013>
4. Ben Ayed, S., Elouedi, Z., Lefevre, E.: ECTD: evidential clustering and case types detection for case base maintenance. In: International Conference on Computer Systems and Applications, pp. 1462–1469. IEEE (2017)

5. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584 (2001)
6. Antoine, V., Quost, B., Masson, M., Denoeux, T.: CECM: constrained evidential c-means algorithm. *Comput. Stat. Data Anal.* **56**, 894–914 (2012)
7. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
8. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
9. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
10. Masson, M.H., Denoeux, T.: ECM: an evidential version of the fuzzy c-means algorithm. *Patt. Recogn.* **41**, 1384–1397 (2008)
11. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**, 191–234 (1994)



The Kantorovich Problem and Wasserstein Metric in the Theory of Belief Functions

Andrey G. Bronevich¹(✉) and Igor N. Rozenberg²

¹ National Research University Higher School of Economics,
Myasnitskaya 20, 101000 Moscow, Russia

brone@mail.ru

² JSC “Research and Design Institute for Information Technology,
Signalling and Telecommunications on Railway Transport”,

Orlikov per. 5, Building 1, 107996 Moscow, Russia

I.Rozenberg@gismps.ru

Abstract. The aim of this paper is to show that the Kantorovich problem, well known in models of economics and very intensively studied in probability theory in recent years, can be viewed as the basis of some constructions in the theory of belief functions. We demonstrate this by analyzing specialization relation for finitely defined belief functions and belief functions defined on reals. In addition, for such belief functions we consider the Wasserstein metric and study its connections to disjunctions of belief functions.

Keywords: Random sets · Belief functions · Specialization
Kantorovich problem · Wasserstein metric

1 Introduction

Theory of belief functions has been successfully applied in many fields such as decision-making [1], data analysis [2–4], image processing [5–7], but the unified background of this theory are random sets [8]. Certainly, in practice the simplest random sets are used, such as finite random sets [9] or random sets on reals [10], because in other cases the underlying problems seem to be can be solved only theoretically. In the paper, we give the main notions from the theory of belief functions based on random sets and show how introduced constructions can be viewed through the Kantorovich problem [11, 12], especially we study the specialization relation and the Wasserstein metric on belief functions. This metric seems to be introduced very naturally because it can be considered as an extension of a metric defined on usual sets.

The paper has the following structure. In Sect. 2 we describe the Kantorovich problem and connected with this problem the Wasserstein metric defined on probability measures. In Sect. 3 we give basic constructions from the theory of

belief functions using random sets. In Sect. 4 we introduce the inclusion of random sets known as specialization relation for belief functions. We show how this relation can be viewed through the Kantorovich problem, and we characterize the inclusion of random sets through lower and upper subsets of the corresponding partially ordered set. In Sect. 5 we introduce the Wasserstein metric on random sets and study its connections to disjunctions of belief functions. We finish the paper with conclusions and lighten problems for future research.

2 The Kantorovich Problem

The Kantorovich problem [11, 12] is well known in economics and it can be formally formulated as follows. Let (X, \mathfrak{A}, μ) and (Y, \mathfrak{B}, ν) be probability spaces, and let $\mathcal{M}(\mu, \nu)$ be the set of all probability measures on $(X \times Y, \mathfrak{A} \otimes \mathfrak{B})$ with marginals μ and ν on X and Y respectively. Then the Kantorovich problem consists in finding a probability measure $\hat{\sigma} \in \mathcal{M}(\mu, \nu)$ providing the infimum of the functional

$$K(\mu, \nu, c) = \inf_{\sigma \in \mathcal{M}(\mu, \nu)} \int_{X \times Y} c(x, y) d\sigma$$

for the cost function $c : X \times Y \rightarrow [0, +\infty)$. If X and Y are finite sets, i.e. $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$, then we can define probability measures μ, ν, σ by probabilities $\mu_i = \mu(\{x_i\})$, $\nu_j = \nu(\{y_j\})$, $\sigma_{ij} = \sigma(\{(x_i, y_j)\})$ and the Kantorovich problem is simplified to the linear programming problem:

$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) \sigma_{ij} \rightarrow \min, \\ & \left\{ \begin{array}{l} \sum_{i=1}^n \sigma_{ij} = \nu_j, \quad j = 1, \dots, m, \\ \sum_{j=1}^m \sigma_{ij} = \mu_i, \quad i = 1, \dots, n, \\ \sigma_{ij} \geq 0, \quad i = 1, \dots, n, \quad j = 1, \dots, m. \end{array} \right. \end{aligned}$$

In economics, the Kantorovich problem consists in the following. Assume that X is the set of factories, producing the same goods and Y is the set of storages. The factory x_i produces amount of goods ν_i that should be kept in storages y_j with volumes ν_j , and $\sum_{i=1}^n \mu_i = \sum_{j=1}^m \nu_j = 1$, and $c(x_i, y_j)$ gives as the transportation cost of the unit of goods from the factory x_i to the storage y_j . Then the optimal values σ_{ij} give us the optimal transportation plan for produced goods.

Assume that $X = Y$ and $d(x, y) = c(x, y)$ is a metric in X , then $d_W(\mu, \nu) = K(\mu, \nu, d)$ is the Wasserstein metric [13, 14] on probability measures. Obviously, many constructions in the theory of belief functions are linked with the Kantorovich problem (see, for example, [15–17]), and the aim of our paper is to lighten these constructions and to extend them for belief functions on reals.

3 Theory of Belief Functions: Main Notions and Constructions

Let $(\Omega, \mathfrak{A}, \mu)$ be a probability space and (X, \mathfrak{B}) be a measurable space and $\mathcal{A} \subseteq \mathfrak{B}$ be a collection of subsets in X . Then every mapping $\Xi : \Omega \rightarrow \mathcal{A}$ is called a random set if for every $B \in \mathfrak{B}$ the set $\{\omega \in \Omega | \Xi(\omega) \subseteq B\}$ is in \mathfrak{A} . Obviously, the mapping Ξ induces the algebra \mathfrak{C} on \mathcal{A} that consists of sets $\{B \in \mathcal{A} | \Xi^{-1}(B) \in \mathfrak{A}\}$ for any $A \in \mathfrak{A}$, and the probability measure P on \mathfrak{C} is defined by $P(C) = \mu(A)$ if $C = \{B \in \mathcal{A} | \Xi^{-1}(B) \in \mathfrak{A}\}$. In the theory of belief functions the probability measure P is called the basic probability assignment and the set function

$$Bel(B) = \mu(\{y \in Y | \Xi(y) \subseteq B\})$$

is called a belief function. Since $\{\omega \in \Omega | \Xi(\omega) \cap B \neq \emptyset\} = \Omega \setminus \{\omega \in \Omega | \Xi(\omega) \subseteq \bar{B}\}$, the set $\{\omega \in \Omega | \Xi(\omega) \cap B \neq \emptyset\}$, $B \in \mathfrak{B}$ is also measurable, and we can introduce the set function

$$Pl(B) = \mu(\{y \in Y | \Xi(y) \cap B \neq \emptyset\}),$$

called the plausibility function. Obviously, they satisfy the dual relation:

$$Pl(B) = 1 - Bel(\bar{B}), \quad B \in \mathfrak{B}.$$

We will illustrate these notions by two notable examples.

Example 1. If \mathcal{A} is finite (this is the case when the set X is also finite, \mathfrak{B} is the powerset of X and $\mathcal{A} = \mathfrak{B}$), then we can define the basic probability assignment by the set function $m : \mathcal{A} \rightarrow [0, 1]$ assuming that $m(A) = P(\{A\})$, $A \in \mathcal{A}$. In addition,

$$Bel(B) = \sum_{A \in \mathcal{A} | A \subseteq B} m(A), \quad Pl(B) = \sum_{A \in \mathcal{A} | A \cap B \neq \emptyset} m(A) \text{ for any } B \in \mathfrak{B}.$$

Example 2. Assume that $X = \mathbb{R}$, \mathfrak{B} is the σ -algebra of Borel measurable subsets in \mathbb{R} , and $\mathcal{A} = \{[a, b] | a \leq b, a, b \in \mathbb{R}\}$. Consider a probability space $(\Omega, \mathfrak{A}, \mu)$, in which $\Omega = \mathbb{R}^2$ and \mathfrak{A} consists of Borel measurable subsets of \mathbb{R}^2 and $\mu(\{(x, y) \in \mathbb{R}^2 | x + y \leq 0\}) = 0$. Then we can define a random set $\Xi : \Omega \rightarrow \mathcal{A}$ assuming that $\Xi(x, y) = [-x, y]$. We see that for $B = [a, b]$

$$\{\omega \in \Omega | \Xi(\omega) \subseteq B\} = \{(x, y) \in \mathbb{R}^2 | x \leq -a, y \leq b, x + y \geq 0\}.$$

This set is depicted on Fig. 1. If $F : \mathbb{R}^2 \rightarrow [0, 1]$ is the cumulative distribution function for μ , i.e. $F(x, y) = \mu((-\infty, x] \times (-\infty, y])$, then $Bel([a, b]) = F(-a, b)$. Smets considers in [10] continuous belief functions when μ defines a continuous random variable on \mathbb{R}^2 . Obviously, if set B is represented as a union of mutually disjoint intervals A_i , i.e. $B = \bigcup_{i=1}^n A_i$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then $Bel(B) = \sum_{i=1}^n Bel(A_i)$.

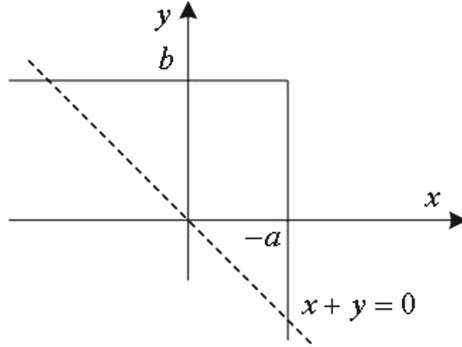


Fig. 1. The set $\{\omega \in \Omega | \Xi(\omega) \subseteq B\}$

4 Inclusion of Random Sets

At first, we will introduce the specialization relation [18] that has the same role as inclusion for usual sets. Let Ξ_1 and Ξ_2 be random sets, then formally $\Xi_1 \subseteq \Xi_2$ iff there is a joint probability distribution P of Ξ_1 and Ξ_2 such that $P(\Xi_1 \subseteq \Xi_2) = 1$. If $\Xi_1 \subseteq \Xi_2$, then Ξ_1 is called a specialization of Ξ_2 . We will illustrate this notion on previous examples of random sets.

Example 3. Let us use assumptions and notations from Example 2 and random sets Ξ_1 and Ξ_2 are given by their basic probability assignments $m_1 : \mathcal{A} \rightarrow [0, 1]$ and $m_2 : \mathcal{A} \rightarrow [0, 1]$. Then $\Xi_1 \subseteq \Xi_2$ if there is their joint probability assignment $m : \mathcal{A} \times \mathcal{A} \rightarrow [0, 1]$, such that $m(A, B) = 0$ if $A \not\subseteq B$ and

$$\begin{cases} \sum_{B \in \mathcal{A}} m(A, B) = m_1(A), \\ \sum_{A \in \mathcal{A}} m(A, B) = m_2(B). \end{cases} \tag{1}$$

We can check whether $\Xi_1 \subseteq \Xi_2$ solving the Kantorovich optimization problem w.r.t. m given (1).

Example 4. Let us use assumptions and notations from Example 2. Consider random sets Ξ_μ and Ξ_ν generated by probability measures μ and ν on the algebra \mathfrak{A} and the mapping $\Xi(x, y) = [-x, y]$. In this case $\Xi(\mathbf{x}) \subseteq \Xi(\mathbf{y})$ for $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ if $\mathbf{x} \leq \mathbf{y}$ (we use here the notation: $\mathbf{x} \leq \mathbf{y}$ if $x_1 \leq y_1$ and $x_2 \leq y_2$). Then $\Xi_\mu \subseteq \Xi_\nu$ iff there is a probability measure γ on $\mathfrak{A} \otimes \mathfrak{A}$ with corresponding marginals μ and ν such that $\gamma(\{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^2 \times \mathbb{R}^2 | \mathbf{x} \leq \mathbf{y}\}) = 1$.

Now we will give the characterization of the specialization relation in terms of upper subsets [19] of partially ordered sets. Consider first the case from Example 1. Then the algebra $\mathfrak{B} = 2^X$ can be viewed as a partially ordered set w.r.t. inclusion of sets. A subset $\mathfrak{f} \subseteq 2^X$ is called an upper subset (semi-filter) in algebra 2^X if $A \in \mathfrak{f}$ and $A \subseteq B$ implies $B \in \mathfrak{f}$.

Proposition 1. ¹Let us consider the case from Example 3. Then $\Xi_1 \subseteq \Xi_2$ iff $\sum_{A \in 2^X \setminus \mathfrak{f}} m_1(A) + \sum_{A \in \mathfrak{f}} m_2(A) \geq 1$ for any upper set \mathfrak{f} in 2^X .

Remark 1. The inequality from Proposition 1 can be rewritten in the form

$$\sum_{A \in \mathfrak{f}} m_2(A) \geq \sum_{A \in \mathfrak{f}} m_1(A).$$

Clearly, the set $\mathfrak{f}_B = \{A \in 2^X | A \cap B \neq \emptyset\}$ is a upper set in 2^X and we see that such upper sets define plausibility functions $Pl_i(B) = \sum_{A \in \mathfrak{f}_B} m_i(A)$, $i = 1, 2$, on 2^X . Thus, $\Xi_1 \subseteq \Xi_2$ implies $Pl_1(B) \leq Pl_2(B)$ for all $B \in 2^X$.

Remark 2. We can equivalently reformulate Proposition 1 through the notion of lower subset (semi-ideal) [19] of a partially ordered set. We will call the subset $\mathfrak{g} \subseteq 2^X$ a lower set in 2^X if $B \in \mathfrak{g}$ and $A \subseteq B$ implies $A \in \mathfrak{g}$. Obviously, if \mathfrak{f} is an upper set in 2^X , then $2^X \setminus \mathfrak{f}$ is a lower set in 2^X and vice versa. Thus, $\Xi_1 \subseteq \Xi_2$ iff $\sum_{A \in \mathfrak{g}} m_2(A) \leq \sum_{A \in \mathfrak{g}} m_1(A)$ for every lower set \mathfrak{g} in 2^X . For example, if we consider lower sets \mathfrak{f}_B from Remark 1, then we can define a lower set $\mathfrak{g}_B = 2^X \setminus \mathfrak{f}_B = \{A \in 2^X | A \subseteq B\}$. Then $Bel_i(B) = \sum_{A \in \mathfrak{g}_B} m_i(A)$, $i = 1, 2$, on 2^X . Thus, $\Xi_1 \subseteq \Xi_2$ implies $Bel_1(B) \geq Bel_2(B)$ for all $B \in 2^X$.

Next example shows that the inequalities $Pl_1(B) \leq Pl_2(B)$ for all $B \in 2^X$ (or $Bel_1(B) \geq Bel_2(B)$ for all $B \in 2^X$) do not provide the inclusion of random sets.

Example 5. Let $X = \{x_1, x_2, x_3\}$ and let basic probability assignments m_1 and m_2 be defined by $m_1(\{x_i\}) = 1/6$, $i = 1, 2, 3$; $m_1(X) = 1/2$; $m_2(\{x_1, x_2\}) = m_2(\{x_1, x_3\}) = m_2(\{x_2, x_3\}) = 1/3$. Then $Pl_1(B) \leq Pl_2(B)$ for all $B \in 2^X$, but $\Xi_1 \not\subseteq \Xi_2$. For proving last statement it is sufficient to consider the upper set $\mathfrak{f} = \{X\}$ and to notice that $m_1(X) > m_2(X)$.

Remark 3. If we define random sets as in Example 2, then we should consider the partially ordered set \mathbb{R}^2 w.r.t. \leq . In this case a subset $\mathfrak{f} \subseteq \mathbb{R}^2$ is called an upper set in \mathbb{R}^2 if $\mathbf{x} \in \mathfrak{f}$ and $\mathbf{x} \leq \mathbf{y}$ implies $\mathbf{y} \in \mathfrak{f}$.

Proposition 2. Let random sets Ξ_μ and Ξ_ν be defined as in Example 4 and $\Xi_\mu \subseteq \Xi_\nu$. Then $\mu(\mathfrak{f}) \leq \nu(\mathfrak{f})$ for any upper set $\mathfrak{f} \in \mathfrak{A}$.

Proof. Assume that $\Xi_\mu \subseteq \Xi_\nu$ and a measure γ is defined like in Example 4. Then

$$\mu(\mathfrak{f}) = \gamma(\{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^2 \times \mathbb{R}^2 | \mathbf{u} \in \mathfrak{f}, \mathbf{u} \leq \mathbf{v}\}).$$

Because $\mathbf{u} \in \mathfrak{f}$, $\mathbf{u} \leq \mathbf{v}$ implies $\mathbf{v} \in \mathfrak{f}$, we infer that

$$\mu(\mathfrak{f}) \leq \gamma(\{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^2 \times \mathbb{R}^2 | \mathbf{v} \in \mathfrak{f}\}) = \nu(\mathfrak{f}).$$

¹ The proof of this proposition is based on Ford-Fulkerson Theorem for the network flow problem and it is omitted because of the required format of the paper.

Remark 4. We can reformulate the statement from Proposition 2 using the notion of lower set in \mathbb{R}^2 . We will call the subset $\mathfrak{g} \subseteq \mathbb{R}^2$ a lower set in \mathbb{R}^2 if $\mathbf{x} \in \mathfrak{g}$, $\mathbf{x} \geq \mathbf{y}$ implies $\mathbf{y} \in \mathfrak{g}$. It is easy to check that if \mathfrak{f} is an upper set in \mathbb{R}^2 , then $\mathbb{R}^2 \setminus \mathfrak{f}$ is a lower set in \mathbb{R}^2 . Thus, $\Xi_\mu \subseteq \Xi_\nu$ implies that $\mu(\mathfrak{g}) \geq \nu(\mathfrak{g})$ for any lower set $\mathfrak{g} \in \mathfrak{A}$. As an example of lower set can be viewed the set $\mathfrak{g}_\mathbf{x} = \{\mathbf{y} \in \mathbb{R}^2 | \mathbf{y} \leq \mathbf{x}\}$. Obviously, using $\mathfrak{g}_\mathbf{x}$ we can compute cumulative distribution functions F_μ and F_ν by $F_\mu(\mathbf{x}) = \mu(\mathfrak{g}_\mathbf{x})$ and $F_\nu(\mathbf{x}) = \nu(\mathfrak{g}_\mathbf{x})$. Thus, $\Xi_\mu \subseteq \Xi_\nu$ implies the following inequalities $F_\mu(\mathbf{x}) \geq F_\nu(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^2$. Next example shows that these inequalities do not imply the inclusion $\Xi_\mu \subseteq \Xi_\nu$.

Example 6. Assume that probability measures μ and ν are finitely defined and

$$\begin{aligned} \mu(\{(1, 1)\}) &= 0.3, \mu(\{(1, 2)\}) = 0.1, \mu(\{(2, 1)\}) = 0.1, \mu(\{(2, 2)\}) = 0.5; \\ \nu(\{(1, 1)\}) &= 0.2, \nu(\{(1, 2)\}) = 0.2, \nu(\{(2, 1)\}) = 0.2, \nu(\{(2, 2)\}) = 0.4. \end{aligned}$$

Then $F_\mu(\mathbf{x}) \geq F_\nu(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^2$. However, $\Xi_\mu \not\subseteq \Xi_\nu$, because $\mu(\mathfrak{f}_\mathbf{y}) \geq \nu(\mathfrak{f}_\mathbf{y})$, where $\mathfrak{f}_\mathbf{y} = \{\mathbf{x} \in \mathbb{R}^2 | \mathbf{x} \geq \mathbf{y}\}$ and $\mathbf{y} = (2, 2)$.

5 The Wasserstein Metric on Random Sets

The Wasserstein metric allows us to extend the distance defined on the sets to the distance defined on random sets. For instance, if we consider random sets defined in Example 1, then the possible distance between sets $A, B \in 2^X$ is $d(A, B) = |(A \setminus B) \cup (B \setminus A)|$, i.e. the cardinality of symmetrical difference between sets A and B . Then to compute the distance between random sets we should solve the Kantorovich problem

$$\begin{aligned} d(\Xi_1, \Xi_2) &= \min \sum_{A \in 2^X} \sum_{B \in 2^X} m(A, B) d(A, B) \\ &\begin{cases} \sum_{B \in 2^X} m(A, B) = m_1(A), A \in 2^X, \\ \sum_{A \in 2^X} m(A, B) = m_2(B), B \in 2^X. \end{cases} \end{aligned} \quad (2)$$

Now we will find the connection between the introduced Wasserstein distance and disjunctions of random sets. We call the random set Ξ_3 the disjunction of random sets Ξ_1 and Ξ_2 if there is the joint probability assignment m satisfying (2), and the basic probability assignment m_3 for Ξ_3 is computed by

$$m_3(C) = \sum_{A, B \in 2^X | A \cup B = C} m(A, B).$$

We will define the cardinality of random set Ξ_i with the basic probability assignment m_i by $|\Xi_i| = \sum_{A \in 2^X} m_i(A) |A|$.

Proposition 3. *Let a random set Ξ_3 be the disjunction of Ξ_1 and Ξ_2 with the smallest cardinality. Then $d(\Xi_1, \Xi_2) = 2|\Xi_3| - |\Xi_1| - |\Xi_2|$.*

Proof. The truth of the proposition follows from the equality $d(A, B) = 2|A \cup B| - |A| - |B|$, which is valid for usual sets $A, B \in 2^X$.

Let us consider how introduced constructions for random sets from Example 1 can be defined for random sets from Example 2. In this case for measuring cardinality of a segment $[a, b]$ we can use the Lebesgue measure of this segment defined by $V([a, b]) = b - a$. Then the cardinality of a random set Ξ_μ can be evaluated by the integral $V(\Xi_\mu) = \int_{\mathbb{R}^2} v(\mathbf{x})dF_\mu(\mathbf{x})$, where $v(x, y) = x + y$. The use of usual union of segments for defining the disjunction of random sets is not well suited for our problem, because the union of segments is not the segment in general. Thus, we define the disjunction of segments $[a_1, b_1]$ and $[a_2, b_2]$ by $[\min\{a_1, a_2\}, \max\{b_1, b_2\}]$. If we depict such segments in \mathbb{R}^2 like in Example 2, then $\mathbf{z} = (z_1, z_2)$ is the disjunction of segments $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ if $(z_1, z_2) = (\max\{x_1, y_1\}, \max\{x_2, y_2\})$ ($\mathbf{z} = \mathbf{x} \vee \mathbf{y}$ for short). A random set Ξ_η is called the disjunction of random sets Ξ_μ and Ξ_ν if there is a joint probability distribution γ on $\mathfrak{A} \otimes \mathfrak{A}$ with marginals μ and ν , and the random set Ξ_η can be obtained from γ by the mapping $f : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$, where $f(\mathbf{x}, \mathbf{y}) = \mathbf{x} \vee \mathbf{y}$.

Proposition 4. *Let $\mathfrak{X} = \{\Xi_\eta\}$ be the set of all possible disjunctions of random sets Ξ_μ and Ξ_ν . Then the functional $d(\Xi_\mu, \Xi_\nu) = 2 \inf \{V(\Xi_\eta) | \Xi_\eta \in \mathfrak{X}\} - V(\Xi_\mu) - V(\Xi_\nu)$ is the Wasserstein metric on random sets.*

Proof. Assume that Ξ_η is obtained from γ by the mapping $f : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$, where $f(\mathbf{x}, \mathbf{y}) = \mathbf{x} \vee \mathbf{y}$. Then

$$2V(\Xi_\eta) - V(\Xi_\mu) - V(\Xi_\nu) = \int_{\mathbb{R}^2 \times \mathbb{R}^2} (2v(\mathbf{x} \vee \mathbf{y}) - v(\mathbf{x}) - v(\mathbf{y}))dF_\gamma(\mathbf{x}, \mathbf{y}).$$

We see that $2v(\mathbf{x} \vee \mathbf{y}) - v(\mathbf{x}) - v(\mathbf{y}) = 2(\max\{x_1, y_1\} + \max\{x_2, y_2\}) - x_1 - y_1 - x_2 - y_2 = |x_1 - y_1| + |x_2 - y_2|$, where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$, i.e. $d(\mathbf{x}, \mathbf{y}) = 2v(\mathbf{x} \vee \mathbf{y}) - v(\mathbf{x}) - v(\mathbf{y})$ is a metric on \mathbb{R}^2 . This implies the proposition.

6 Conclusion

We show that the Kantorovich problem appears naturally in many constructions of the theory of belief functions. However, this problem seems to be tractable, when we can describe random sets by discrete probability distributions and it can be represented as a linear programming problem. We are certain that the Wasserstein metric on belief functions can be used in many applications, for example, for measuring conflict in weather forecasts as shown in [16].

Acknowledgment. This work has been supported by the grant 18-01-00877 of RFBR (Russian Foundation for Basic Research).

References

1. Smets, P.: Decision making in a context where uncertainty is represented by belief functions. In: Srivastava, R.P., Mock, T.J. (eds.) *Belief Functions in Business Decisions*, pp. 17–61. Springer, Heidelberg (2002). https://doi.org/10.1007/978-3-7908-1798-0_2
2. Denoeux, T.: A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. SMC* **25**(5), 804–813 (1995)
3. Bronevich, A.G., Lepskiy, A.E., Penikas, H.I.: Coherence analysis of financial analysts' recommendations in the framework of evidence theory. *CEUR-Workshop* **1687**, 12–23 (2016)
4. Kutynina, E., Lepskiy, A.: Aggregation of forecasts and recommendations of financial analysts in the framework of evidence theory. In: Kacprzyk, J., Szmidt, E., Zadrożny, S., Atanassov, K.T., Krawczak, M. (eds.) *IWIFSGN/EUSFLAT - 2017. AISC*, vol. 642, pp. 370–381. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-66824-6_33
5. Bloch, I.: Defining belief functions using mathematical morphology - application to image fusion under imprecision. *Int. J. Approximate Reasoning* **48**, 437–465 (2008)
6. Lin, T.-C.: Partition belief median filter based on Dempster-Shafer theory in image processing. *Patt. Recogn.* **41**, 139–151 (2008)
7. Lin, T.-C.: Decision-based filter based on SVM and evidence theory for image noise removal. *Neural Comput. Appl.* **21**, 685–793 (2012)
8. Molchanov, I.: *Theory of Random Sets*. Springer, London (2005)
9. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
10. Smets, P.: Belief functions on real numbers. *Int. J. Approximate Reasoning* **40**(3), 181–223 (2005)
11. Kantorovich, L.V.: On mass moving. *Dokl. Akad. Nauk SSSR* **37**(7–8), 227–229 (1942)
12. Bogachev, V.I., Kolesnikov, A.V.: The Monge-Kantorovich problem: achievements, connections, and perspectives. *Russ. Math. Surv.* **67**(5), 1–110 (2012)
13. Givens, C.R., Shortt, R.M.: A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* **31**(2), 231–240 (1984)
14. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth Mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40**(2), 99–121 (2000)
15. Bronevich, A.G., Rozenberg, I.N.: The choice of generalized Dempster-Shafer rules for aggregating belief functions. *Int. J. Approximate Reasoning* **56**, 122–136 (2015)
16. Bronevich, A.G., Spiridenkova, N.S.: Measuring uncertainty for interval belief structures and its application for analyzing weather forecasts. In: Kacprzyk, J., Szmidt, E., Zadrożny, S., Atanassov, K., Krawczak, M. (eds.) *Advances in Fuzzy Logic and Technology 2017. Advances in Intelligent Systems and Computing*, vol. 641, pp. 273–285. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-66830-7_25
17. Han, D., Dezert, J., Yang, Y.: New distance measures of evidence based on belief intervals. In: Cuzzolin, F. (ed.) *BELIEF 2014. LNCS (LNAI)*, vol. 8764, pp. 432–441. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11191-9_47
18. Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *Int. J. Gen. Syst.* **12**(3), 193–226 (1986)
19. Davey, B.A., Priestley, H.A.: *Introduction to Lattices and Order*. Cambridge University Press, Cambridge (2002)



Generalised Max Entropy Classifiers

Fabio Cuzzolin^(✉)

Oxford Brookes University, Oxford, UK
fabio.cuzzolin@brookes.ac.uk

Abstract. In this paper we propose a generalised maximum-entropy classification framework, in which the empirical expectation of the feature functions is bounded by the lower and upper expectations associated with the lower and upper probabilities associated with a belief measure. This generalised setting permits a more cautious appreciation of the information content of a training set. We analytically derive the Karush-Kuhn-Tucker conditions for the generalised max-entropy classifier in the case in which a Shannon-like entropy is adopted.

Keywords: Classification · Max entropy · Constrained optimisation

1 Introduction

The emergence of new challenging real-world applications has exposed serious issues with current approaches to model adaptation in machine learning. Existing theory and algorithms focus on fitting the available training data, but cannot provide worst-case guarantees in mission-critical applications. Vapnik's statistical learning theory is useless for model selection, as the bounds on generalisation errors it predicts are too wide to be useful, and rely on the assumption that training and testing data come from the same (unknown) distribution. The crucial question is: what exactly can one infer from a training set?

Max entropy classifiers [19] provide a significant example, due to their simplicity and widespread application. There, the entropy of the sought joint (or conditional) probability distribution of data and class is maximised, following the *maximum entropy principle* that the least informative distribution which matches the available evidence should be chosen. Having picked a set of *feature functions*, selected to efficiently encode the training information, the joint distribution is subject to the constraint that their empirical expectation equals that associated with the max entropy distribution. The assumptions that (i) training and test data come from the same probability distribution, and that (ii) the empirical expectation of the training data is correct, and the model expectation should match it, are rather strong, and work against generalisation power.

A way around this issue is to adopt as models convex sets of probability distributions, rather than standard probability measures. Random sets, in particular, are mathematically equivalent to a special class of credal sets induced by probability mass assignments on the power set of the sample space.

When random sets are defined on finite domain, they are often called *belief functions* [20]. One can then envisage a robust theory of learning based on generalising traditional statistical learning theory in order to allow for test data to be sampled from a *different* probability distribution than the training data, under the weaker assumption that both belong to the same random set.

In this paper we make a step in that direction by generalising the max entropy classification framework. We take the view that a training set does not provide, in general, sufficient information to precisely estimate the joint probability distribution of class and data. We assume instead that a belief measure can be estimated, providing lower and upper bounds on the joint probability of data and class. As in the classical case, an appropriate measure of entropy for belief measures is maximised. In opposition to the classical case, however, the empirical expectation of the chosen feature functions is assumed to be *compatible* with lower and upper bounds associated with the sought belief measure. This leads to a constrained optimisation problem with inequality constraints, rather than equality ones, which needs to be solved by looking at the Karush-Kuhn-Tucker (KKT) conditions. Due to the concavity of the objective function and the convexity of the constraints, KKT conditions are both necessary and sufficient.

Related Work. A significant amount of work has been conducted in the past on machine learning approaches based on belief theory. Most efforts were directed at developing clustering tools, including evidential clustering [4], evidential and belief C-means [15]. Ensemble classification [23], in particular, has been extensively studied. Concerning classification, Denoeux [5] proposed in a seminal work a k-nearest neighbor classifier based on belief theory. Relevantly to this paper, interesting work has been conducted to generalise the framework of decision trees to situations in which uncertainty is encoded by belief functions, mainly by Elouedi and co-authors [7], and Vannoorenberghe and Denoeux [22].

Paper Outline. After reviewing in Sect. 2 max-entropy classification, we recall in Sect. 3 the necessary notions of belief theory. In Sect. 4 the possible generalisations of Shannon’s entropy to the case of belief measures are reviewed. In Sect. 5 the generalised max-entropy problem is formulated, together with the associated Kush-Karun-Tucker conditions. It is shown that for several generalised measures of entropy the KKT conditions are necessary and sufficient for the optimised of generalised max-entropy (Sect. 5.1). In Sect. 5.2 we derive the analytical expression of the system of KKT conditions for the case of a Shannon-like entropy for belief measures. Section 6 concludes the paper.

2 Max-entropy Classifiers

The objective of *maximum entropy classifiers* is to maximise the Shannon entropy of the conditional classification distribution $p(C_k|x)$, where $x \in X$ is the observable and $C_k \in \mathcal{C} = \{C_1, \dots, C_K\}$ is the associated class.

Given a training set in which each observation is attached a class, namely: $\mathcal{D} = \{(x_i, y_i), i = 1, \dots, N | x_i \in X, y_i \in \mathcal{C}\}$, a set M of *feature maps* is designed, $\phi(x, C_k) = [\phi_1(x, C_k), \dots, \phi_M(x, C_k)]'$ whose values depend on both the object

observed and its class. Each feature map $\phi_m : X \times \mathcal{C} \rightarrow \mathbb{R}$ is then a random variable whose expectation is: $E[\phi_m] = \sum_{x,k} p(x, C_k) \phi_m(x, C_k)$. In opposition, the *empirical* expectation of ϕ_m is: $\hat{E}[\phi_m] = \sum_{x,k} \hat{p}(x, C_k) \phi_m(x, C_k)$, where \hat{p} is a histogram constructed by counting occurrences of the pair (x, C_k) in the training set: $\hat{p}(x, C_k) = \frac{1}{N} \sum_{(x_i, y_i) \in \mathcal{D}} \delta(x_i = x \wedge y_i = C_k)$. The theoretical expectation $E[\phi_m]$ can be approximated by decomposing $p(x, C_k) = p(x)p(C_k|x)$ via Bayes' rule, and approximating the (unknown) prior of the observations $p(x)$ with the empirical prior \hat{p} , i.e., the histogram of observed values in the training set: $\tilde{E}[\phi_m] = \sum_{x,k} \hat{p}(x)p(C_k|x)\phi_m(x, C_k)$.

Definition 1. *Given a training set $\mathcal{D} = \{(x_i, y_i), i = 1, \dots, N | x_i \in X, y_i \in \mathcal{C}\}$ related to problem of classifying $x \in X$ as belonging to one of the classes $\mathcal{C} = \{C_1, \dots, C_K\}$, the max entropy classifier is the conditional probability $p^*(C_k|x)$ such that: $p^*(C_k|x) \doteq \arg \max_{p(C_k|x)} H_s(P)$, where H_s is the traditional Shannon entropy, subject to: $\tilde{E}_p[\phi_m] = \hat{E}[\phi_m] \forall m = 1, \dots, M$.*

The constraint requires the classifier to be consistent with the empirical frequencies of the features in the training set, while seeking the least informative probability distribution that does so. The solution of the maximum entropy classification problem (Definition 1) is the so-called *log-linear model*: $p^*(C_k|x) = \frac{1}{Z_\lambda(x)} e^{\sum_m \lambda_m \phi_m(x, C_k)}$, where $\lambda = [\lambda_1, \dots, \lambda_M]'$ are the Lagrange multipliers associated with the linear constraints $\tilde{E}_p[\phi_m] = \hat{E}[\phi_m]$, and $Z_\lambda(x)$ is a normalisation factor. The related classification function is: $y(x) = \arg \max_k \sum_m \lambda_m \phi_m(x, C_k)$, i.e., x is assigned the class which maximises the linear combination of the feature functions with coefficients λ .

3 Belief Functions

Definition 2. *A basic probability assignment (BPA) [1] over a discrete set Θ is a function $m : 2^\Theta \rightarrow [0, 1]$ defined on $2^\Theta = \{A \subseteq \Theta\}$ such that: $m(\emptyset) = 0$, $\sum_{A \subseteq \Theta} m(A) = 1$. The belief function (BF) associated with a BPA $m : 2^\Theta \rightarrow [0, 1]$ is the set function $Bel : 2^\Theta \rightarrow [0, 1]$ defined as: $Bel(A) = \sum_{B \subseteq A} m(B)$.*

The elements of the power set 2^Θ associated with non-zero values of m are called the *focal elements* of m . For each subset ('event') $A \subseteq \Theta$ the quantity $Bel(A)$ is called the *degree of belief* that the outcome lies in A , and represents the total belief committed to a set of outcomes A by the available evidence m . Dually, the *upper probability* of A : $Pl(A) \doteq 1 - Bel(\bar{A})$, $\bar{A} = \Theta \setminus A$, expresses the 'plausibility' of a proposition A or, in other words, the amount of evidence *not against* A [3]. The *plausibility function* $Pl : 2^\Theta \rightarrow [0, 1]$ thus conveys the same information as Bel , and can be expressed as: $Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \geq Bel(A)$.

Belief functions are mathematically equivalent to a special class of credal sets (convex sets of probability measures), as each BF Bel is associated with the set $\mathcal{P}[Bel] = \{P : P(A) \geq Bel(A)\}$ of probabilities dominating it. Its centre of mass is the *pignistic function* $BetP[Bel](x) = \sum_{A \ni x} m(A)/|A|$, $x \in \Theta$.

Given a function $f : \Theta \rightarrow \mathbb{R}$, the *lower expectation* and *upper expectation* of f w.r.t. Bel are, respectively: $E_{Bel^*}[f] \doteq \inf_{P \in \mathcal{P}[Bel]} E_P[f] = \sum_{A \subseteq \Theta} m(A) \inf_{x \in A} f(x)$, $E_{Bel}^*[f] \doteq \sup_{P \in \mathcal{P}[Bel]} E_P[f] = \sum_{A \subseteq \Theta} m(A) \sup_{x \in A} f(x)$.

4 Measures of Generalised Entropy

The issue of how to assess the level of uncertainty associated with a belief function [10] is not trivial, as authors such as Yager and Klir argued that there are several facets to uncertainty, such as *conflict* (or discord, dissonance) and *non-specificity* (also called vagueness, ambiguity or imprecision).

Some measures are directly inspired by Shannon's entropy of probability measures: $H_s[p] = -\sum_{x \in \Theta} p(x) \log p(x)$. While Nguyen's measure is a direct generalisation in which probability values are replaced by mass values [17]: $H_n[m] = -\sum_{A \in \mathcal{F}} m(A) \log m(A)$, where \mathcal{F} is the list of focal elements of m , in Yager's entropy [24] probabilities are (partly) replaced by plausibilities: $H_y[m] = -\sum_{A \in \mathcal{F}} m(A) \log Pl(A)$. Hohle's *measure of confusion* [9] is the dual measure: $H_o[m] = -\sum_{A \in \mathcal{F}} m(A) \log Bel(A)$. All such measures only capture the 'conflict' portion of uncertainty. Other measures are designed to capture the *specificity* of belief measures, i.e., the degree of concentration of the mass assigned to focal elements. A first such measure was due to Klir, Dubois & Prade [6]: $H_d[m] = \sum_{A \in \mathcal{F}} m(A) \log |A|$, and can be considered as a generalization of Hartley's entropy ($H = \log(|\Theta|)$) to belief functions. A more sophisticated proposal by Pal [18]: $H_a[m] = \sum_{A \in \mathcal{F}} m(A)/|A|$, assesses the dispersion of the evidence and is linked to the pignistic transform. A final proposal based on the commonality function $Q(A) = \sum_{B \supseteq A} m(B)$ is due to Smets: $H_t = \sum_{A \in \mathcal{F}} \log(\frac{1}{Q(A)})$.

Composite measures, such as Lamata and Moral's $H_l[m] = H_y[m] + H_d[m]$ [14], as designed to capture both entropy and specificity. Klir & Ramer [13] proposed a 'global uncertainty measure' defined as: $H_k[m] = D[m] + H_d[m]$, where: $D(m) = -\sum_{A \in \mathcal{F}} m(A) \log[\sum_{B \in \mathcal{F}} m(B) \frac{|A \cap B|}{|B|}]$. Pal et al. [18] argued that none of these composite measures is really satisfactory, as they do not admit a unique maximum and there is no sounding rationale for simply adding conflict and non-specificity measures together.

In the credal interpretation of belief functions, Harmanec and Klir's *aggregated uncertainty* (AU) [8] is defined as the maximal Shannon entropy of all the probabilities consistent with the given BF: $H_h[m] = \max_{P \in \mathcal{P}[Bel]} \{H_s[P]\}$. $H_h[m]$ is the minimal measure meeting a set of rationality requirements which include: symmetry, continuity, expansibility, subadditivity, additivity, monotonicity, normalisation. Similarly, Maeda and Ichihashi [16] proposed a composite measure $H_i[m] = H_h[m] + H_d[m]$ whose first component consists of the maximum entropy of the set of probability distributions consistent with m , and whose second part is the generalized Hartley entropy. As both H_h and H_i have high computational complexity, Jousselme et al. [11] proposed an *ambiguity measure* (AM), as the classical entropy of the pignistic function: $H_j[m] = H_s[BetP[m]]$.

Jirousek and Shenoy [10] analysed all these proposal in 2016, assessing them versus a number of significant properties, concluding that only the

Maeda-Ichihashi proposal meets all these properties. The issue remains still unsettled. In the following we will adopt a straightforward generalisation of Shannon’s entropy, and a few selected proposals based on their concavity property.

5 Generalised Max-entropy Problem

Technically, in order to generalise the max-entropy optimisation problem (Definition 1) to the case of belief functions, we need to: (i) choose an appropriate measure of entropy for belief function as the objective function; (ii) revisit the constraints that the (theoretical) expectations of the feature maps are equal to the empirical ones computed over the training set.

As for (ii), it is sensible to require that the empirical expectation of the feature functions is bracketed by the lower and upper expectations associated with the sought belief function $Bel : 2^{X \times \mathcal{C}} \rightarrow [0, 1]$. In this paper we only make use of the 2-monotonicity of belief functions, and write:

$$\sum_{(x, C_k)} Bel(x, C_k) \phi_m(x, C_k) \leq \hat{E}[\phi_m] \leq \sum_{(x, C_k)} Pl(x, C_k) \phi_m(x, C_k) \quad (1)$$

$\forall m = 1, \dots, M$, as we only consider probability intervals on singleton elements $(x, C_k) \in X \times \mathcal{C}$. Fully fledged lower and upper expectations (cfr. Sect. 3), which express the full monotonicity of BFs, will be considered in future work.

Going even further, should constraints of the form (1) be enforced on all possible subsets $A \subset X \times \mathcal{C}$, rather than just singleton pairs (x, C_k) ? This goes back to the question of what information does a training set actually carry. More general constraints would require extending the domain of feature functions to set values – we will investigate this idea in the near future as well.

5.1 Formulation and Karush-Kuhn-Tucker (KKT) Conditions

In the same classification setting of Sect. 2, the *maximum belief entropy classifier* is the joint belief measure $Bel^*(x, C_k) : 2^{X \times \mathcal{C}} \rightarrow [0, 1]$ which solves the following optimisation problem: $Bel^*(x, C_k) \doteq \arg \max_{Bel(x, C_k)} H(Bel)$ subject to the inequality constraints (1), where H is an appropriate measure of entropy for belief measures. As the above optimisation problem involves inequality constraints (1), as opposed to the equality constraints of traditional max entropy classifiers, we need to analyse the Karush-Kuhn-Tucker (KKT) [12] necessary conditions for a belief function Bel to be an optimal solution to the problem.

Definition 3. *Suppose that the objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ of a nonlinear optimisation problem $\arg \max_x f(x)$ subject to: $g_i(x) \leq 0$ $i = 1, \dots, m$, $h_j(x) = 0$ $j = 1, \dots, l$ are continuously differentiable at a point x^* . If x^* is a local optimum, under appropriate regularity conditions then there exist constants μ_i , ($i = 1, \dots, m$) and*

λ_j ($j = 1, \dots, l$), called KKT multipliers, such that the following conditions hold:

1. Stationarity: $\nabla f(x^*) = \sum_{i=1}^m \mu_i \nabla g_i(x^*) + \sum_{j=1}^l \lambda_j \nabla h_j(x^*)$;
2. Primal feasibility: $g_i(x^*) \leq 0 \forall i = 1, \dots, m$, and $h_j(x^*) = 0, \forall j = 1, \dots, l$;
3. Dual feasibility: $\mu_i \geq 0$ for all $i = 1, \dots, m$;
4. Complementary slackness: $\mu_i g_i(x^*) = 0$ for all $i = 1, \dots, m$.

Crucially, the KKT conditions are also sufficient whenever the objective function f is concave, the inequality constraints g_i are continuously differentiable convex functions, and the equality constraints h_j are affine¹.

Theorem 1. *If either $H_t, H_n, H_d, H_s[Bel]$ or $H_s[Pl]$ is adopted as measure of entropy, the generalised max entropy optimisation problem has concave objective function and convex constraints. Therefore, the KKT conditions are sufficient for the optimality of its solution(s).*

Concavity of the entropy objective function. It is well known that Shannon's entropy is a concave function of probability distributions, represented as vectors of probability values². Furthermore: any linear combination of concave functions is concave; a monotonic and concave function of a concave function is still concave; the logarithm is a concave function.

As shown by Smets [21], the transformations which map mass vectors to vectors of belief (and commonality) values are linear, as they can be expressed in the form of matrices. In particular, $\mathbf{bel} = BfrM\mathbf{m}$, where $BfrM$ is a matrix whose (A, B) entry is: $BfrM(A, B) = 1$ if $B \subseteq A$, 0 otherwise, and \mathbf{bel}, \mathbf{m} are vectors collecting the belief (mass) values of all events $A \subseteq \Theta$. The same can be said of the mapping $\mathbf{q} = QfrM\mathbf{m}$ between a mass vector and the associated commonality vector. As a consequence, belief, plausibility and commonality are all linear (and therefore concave) functions of a mass vector.

Using this matrix representation, it is easy to conclude that several of the entropies defined in Sect. 4 are indeed concave. In particular, Smets' specificity measure $H_t = \sum_A \log(\frac{1}{Q(A)})$ is concave, as a linear combination of concave functions. Nguyen's entropy $H_n = -\sum_A m(A) \log(m(A)) = H_s[m]$ is also concave, as the Shannon's entropy of a mass assignment. Dubois and Prade's measure $H_d = \sum_A m(A) \log(|A|)$ is also concave with respect to m , as a linear combination of mass values. Direct applications of Shannon's entropy function to Bel and Pl : $H_{Bel}[m] = H_s[Bel] = \sum_{A \subseteq \Theta} Bel(A) \log(\frac{1}{Bel(A)})$, $H_{Pl}[m] = H_s[Pl] = \sum_{A \subseteq \Theta} Pl(A) \log(\frac{1}{Pl(A)})$ are also trivially concave, due to the concavity of the entropy function and to the linearity of the mapping from m to Bel, Pl . Drawing conclusions on the other measures is less immediate, as they involve products of concave functions (which are not, in general, guaranteed to be concave).

Convexity of the Interval Expectation Constraints. As for the constraints (1) of the generalised max entropy problem, we first note that (1) can be decomposed into

¹ More general sufficient conditions can be given in terms of *invecity* [2] requirements.

² <http://projecteuclid.org/euclid.lnms/1215465631>.

the following pair of constraints: $g_m^1(m) \doteq \sum_{x,k} Bel(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m] \leq 0$, $g_m^2(m) = \sum_{x,k} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)] \leq 0$ for all $m = 1, \dots, M$. The first inequality constraint is a linear combination of linear functions of the sought mass assignment $m^* : 2^{X \times C} \rightarrow [0, 1]$ (since Bel^* results from applying a matrix transformation to m^*). As $\mathbf{pl} = 1 - \mathbf{J}bel = 1 - \mathbf{J}BfrMm$, constraint g_m^2 is also a linear combination of mass values. Hence, as linear function, constraints g_m^1 and g_m^2 are both concave and convex.

5.2 Belief Max-entropy Classifier for Shannon's Entropy

For the Shannon-like entropy Condition 1. (stationarity), applied to the sought optimal BF $Bel^* : 2^{X \times C} \rightarrow [0, 1]$, reads as: $\nabla H_{Bel}(Bel^*) = \sum_{m=1}^M \mu_m^1 \nabla g_m^1(Bel^*) + \mu_m^2 \nabla g_m^2(Bel^*)$. The components of ∇H_{Bel} are the partial derivatives of the entropy with respect to the mass values $m(\bar{B})$, for all $\bar{B} \subseteq \Theta$. They read as:

$$\frac{\partial H_{Bel}}{\partial m(\bar{B})} = \frac{\partial}{\partial m(\bar{B})} \sum_{A \supseteq \bar{B}} \left[- \left(\sum_{B \subseteq A} m(B) \right) \log \left(\sum_{B \subseteq A} m(B) \right) \right] = - \sum_{A \supseteq \bar{B}} [1 + \log Bel(A)].$$

As for $\nabla g_m^1(Bel^*)$ we have: $\frac{\partial g_m^1}{\partial m(\bar{B})} = \frac{\partial}{\partial m(\bar{B})} \sum_{(x, C_k) \in \Theta} Bel(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m] = \frac{\partial}{\partial m(\bar{B})} \sum_{(x, C_k) \in \Theta} m(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m]$ which is equal to $\phi_m(x, C_k)$ for $\bar{B} = \{(x, C_k)\}$, 0 otherwise³. As for the second set of constraints: $\frac{\partial g_m^2}{\partial m(\bar{B})} = \frac{\partial}{\partial m(\bar{B})} \sum_{(x, C_k) \in \Theta} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)]$ which, recalling that $Pl(x, C_k) = \sum_{B \cap \{(x, C_k)\} \neq \emptyset} m(B)$, becomes equal to $= - \sum_{(x, C_k) \in \bar{B}} \phi_m(x, C_k)$.

Assembling all our results, the KKT stationarity conditions for the generalised, belief-theoretical maximum entropy problem amount to, for all $\bar{B} \subset X \times C$:

$$\begin{cases} - \sum_{A \supseteq \bar{B}} [1 + \log Bel(A)] = \sum_{m=1}^M \phi_m(\bar{x}, \bar{C}_k)[\mu_m^1 - \mu_m^2], |\bar{B} = \{(\bar{x}, \bar{C}_k)\}| = 1, \\ - \sum_{A \supseteq \bar{B}} [1 + \log Bel(A)] = \sum_{m=1}^M \mu_m^2 \sum_{(x, C_k) \in \bar{B}} \phi_m(x, C_k), |\bar{B}| > 1. \end{cases} \quad (2)$$

The other conditions are, $\forall m = 1, \dots, M$, (1) (primal feasibility), $\mu_m^1, \mu_m^2 \geq 0$ (dual feasibility), and complementary slackness: $\mu_m^1 \sum_{(x, C_k) \in \Theta} Bel(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m] = 0$, $\mu_m^2 \sum_{(x, C_k) \in \Theta} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)] = 0$.

6 Conclusions

In this paper we proposed a generalisation of the max entropy classifier entropy in which the assumptions that test and training data are sampled by a same probability distribution, and that the empirical expectation of the feature functions is 'correct' are relaxed in the formalism of belief theory. We also studied

³ If we could define feature functions over non singletons subsets $A \subseteq \Theta$, this would simply generalise to $\phi(\bar{B})$ for all $\bar{B} \subseteq \Theta$.

the conditions under which the associated KKT conditions are necessary and sufficient for the optimality of the solution. Much work remains: (i) providing analytical model expressions, similar to log-linear models, for the Shannon-like and other major entropy measures for belief functions; (ii) analysing the case in which the full lower and upper expectations are plugged in; (iii) comparing the resulting classifiers; (iv) analysing a formulation based on the least commitment principle, rather than max entropy, for the objective function to optimise; finally, (v) relaxing the constraint that feature functions be defined on singleton pairs (x, C_k) , in a further generalisation of this important framework.

References

1. Augustin, T.: Modeling weak information with generalized basic probability assignments. In: Bock, H.H., Polasek, W. (eds.) *Data Analysis and Information Systems*, pp. 101–113. Springer, Heidelberg (1996). https://doi.org/10.1007/978-3-642-80098-6_9
2. Ben-Israel, A., et al.: What is invexity? *J. Austral. Math. Soc. Ser. B* **28**, 1–9 (1986)
3. Cuzzolin, F.: Three alternative combinatorial formulations of the theory of evidence. *Intell. Data Anal.* **14**(4), 439–464 (2010)
4. Denoeux, T., Masson, M.-H.: EVCLUS: evidential clustering of proximity data. *IEEE Trans. Syst. Man Cybern. B* **34**(1), 95–109 (2004)
5. Denceux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (1995)
6. Dubois, D., Prade, H.: Properties of measures of information in evidence and possibility theories. *Fuzzy Sets Syst.* **100**, 35–49 (1999)
7. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. *Int. J. Approximate Reasoning* **28**(2–3), 91–124 (2001)
8. Harmanec, D., Klir, G.J.: Measuring total uncertainty in Dempster-Shafer theory: a novel approach. *Int. J. Gen. Syst.* **22**(4), 405–419 (1994)
9. Hohle, U.: Entropy with respect to plausibility measures. In: *Proceedings of the 12th IEEE Symposium on Multiple-Valued Logic*, pp. 167–169 (1982)
10. Jirousek, R., Shenoy, P.P.: Entropy of belief functions in the Dempster-Shafer theory: a new perspective. In: *Proceedings of BELIEF*, pp. 3–13 (2016)
11. Jousselme, A.L.: Measuring ambiguity in the evidence theory. *IEEE Trans. Syst. Man Cybern. A* **36**(5), 890–903 (2006)
12. Karush, W.: Minima of functions of several variables with inequalities as side constraints. M.Sc. dissertation, Department of Mathematics, University of Chicago (1939)
13. Klir, G.J.: Measures of uncertainty in the Dempster-Shafer theory of evidence. In: *Advances in the Dempster-Shafer theory of evidence*, pp. 35–49 (1994)
14. Lamata, M.T., Moral, S.: Measures of entropy in the theory of evidence. *Int. J. Gen. Syst.* **14**(4), 297–305 (1988)
15. Liu, Z.: Belief c-means: an extension of fuzzy c-means algorithm in belief functions framework. *Pattern Recogn. Lett.* **33**(3), 291–300 (2012)
16. Maeda, Y., Ichihashi, H.: An uncertainty measure with monotonicity under the random set inclusion. *Int. J. Gen. Syst.* **21**(4), 379–392 (1993)
17. Nguyen, H.: On entropy of random sets and possibility distributions. In: *The Analysis of Fuzzy Information*, pp. 145–156 (1985)

18. Pal, N.R., Bezdek, J.C., Hemasinha, R.: Uncertainty measures for evidential reasoning II: a new measure of total uncertainty. *Int. J. Approximate Reasoning* **8**, 1–16 (1993)
19. Pietra, S.D.: Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(4), 380–393 (1997)
20. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
21. Smets, P.: The application of the matrix calculus to belief functions. *Int. J. Approximate Reasoning* **31**(1–2), 1–30 (2002)
22. Vannoorenberghe, P., et al.: Handling uncertain labels in multiclass problems using belief decision trees. In: *Proceedings of IPMU (2002)*
23. Xu, L.: Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Syst. Man Cybern.* **22**(3), 418–435 (1992)
24. Yager, R.R.: Entropy and specificity in a mathematical theory of evidence. *Int. J. Gen. Syst.* **9**, 249–260 (1983)



General Geometry of Belief Function Combination

Fabio Cuzzolin^(✉)

Oxford Brookes University, Oxford, UK
fabio.cuzzolin@brookes.ac.uk

Abstract. In this paper we build on previous work on the geometry of Dempster's rule to investigate the geometric behaviour of various other combination rules, including Yager's, Dubois', and disjunctive combination, starting from the case of binary frames of discernment. Believability measures for unnormalised belief functions are also considered. A research programme to complete this analysis is outlined.

Keywords: Geometry · Yager's and Dubois' combination
Conjunctive and disjunctive combination
Unnormalised belief functions

1 Introduction

In the geometric approach to uncertainty and belief function theory [3], belief measures are represented as points of a convex space, termed *belief space* \mathcal{B} [2]. In a series of papers, in particular, this author studied the behaviour of Dempster's rule of combination in this geometric setting [1]. An earlier analysis of Dempster's rule on binary domains can be found in [6].

In this work, we start to extend this geometric analysis to several other major combination operators, including Yager's [10] and Dubois' rules, but also the disjunctive operator [8]. The final objective of the research programme is a comparative geometric analysis of combination rules, which would eventually allow us to describe the 'cone' of possible future belief states under stronger or weaker assumptions on reliability and independence of sources, associated with conjunctive and disjunctive combination. The bulk of the analysis focusses on standard, normalised belief functions – towards the end, however, we also consider unnormalised belief functions [9] and provide some preliminary results.

We start by giving a general definition of *conditional subspace* (cfr. [3], Chap. 8), as the set of possible future states under a given combination rule.

Definition 1. *Given a belief function (BF) $Bel \in \mathcal{B}$ we call conditional subspace $\langle Bel \rangle_{\odot}$ the set of all \odot combinations of Bel with any other BF Bel' defined on the same frame, where \odot is an arbitrary combination rule, assuming their combination exists: $\langle Bel \rangle_{\odot} \doteq \{ Bel \odot Bel', Bel' \in \mathcal{B} \text{ s.t. } \exists (Bel \odot Bel') \}$.*

Our analysis will be conducted on binary spaces, and used to formulate conjectures on the case of general frames of discernment. We will first recall the necessary notions of the geometric approach to belief theory in Sect. 2. We will consider Yager's and Dubois' rules in Sect. 3, disjunctive combination in Sect. 4, to cover the behaviour of unnormalised BFs in Sect. 5. We will draw some verdicts and outline future work in our Conclusions.

2 Belief Functions and Their Geometry

Belief functions. A basic probability assignment (BPA) [7] over a discrete set (frame) Θ is a function $m : 2^\Theta \rightarrow [0, 1]$ defined on $2^\Theta = \{A \subseteq \Theta\}$ such that: $m(\emptyset) = 0$, $\sum_{A \subseteq \Theta} m(A) = 1$. The *belief function* (BF) associated with a BPA $m : 2^\Theta \rightarrow [0, 1]$ is the function $Bel : 2^\Theta \rightarrow [0, 1]$ defined as: $Bel(A) = \sum_{B \subseteq A} m(B)$. The elements of the power set 2^Θ associated with non-zero values of m are called the *focal elements* of m . For each subset ('event') $A \subseteq \Theta$ the quantity $Bel(A)$ is called the *degree of belief* that the outcome lies in A . Dempster's combination $Bel_1 \oplus Bel_2$ of two belief functions on Θ is the unique BF there with as focal elements all the non-empty intersections of focal elements of Bel_1 and Bel_2 , and basic probability assignment: $m_\oplus(A) = \frac{m_\cap(A)}{1 - m_\cap(\emptyset)}$, where $m_\cap(A) = \sum_{B \cap C = A} m_1(B)m_2(C)$ and m_i is the BPA of the input BF Bel_i .

Belief space. Given a frame of discernment Θ , a BF Bel is specified by its $N - 2$ belief values $\{Bel(A), \emptyset \subsetneq A \subsetneq \Theta\}$, $N \doteq 2^{|\Theta|}$, and can then be represented as a point of \mathbb{R}^{N-2} . The *belief space* [1, 2] associated with Θ is the set of points \mathcal{B} of \mathbb{R}^{N-2} which correspond to proper belief functions. It can be proven that the belief space \mathcal{B} is the convex closure Cl of all the vectors associated with categorical BFs Bel_A (such that $m(A) = 1$): $\mathcal{B} = Cl(Bel_A, \emptyset \subsetneq A \subseteq \Theta) = \{\sum_{\emptyset \subsetneq A \subseteq \Theta} \alpha_A Bel_A, \alpha_A \geq 0 \forall A, \sum_A \alpha_A = 1\}$, an $(N - 2)$ -dimensional *simplex*.

Geometry of Dempster's rule. In [3] we proved that the conditional subspace $\langle Bel \rangle$ under Dempster's combination is $\langle Bel \rangle = Cl\{Bel \oplus Bel_A, A \subseteq \mathcal{C}_{Bel}\}$, where \mathcal{C}_{Bel} is the union of the focal elements of Bel (see Fig. 1, in light blue, for the binary case $\Theta_2 = \{x, y\}$). Dempster's combination of a BF Bel with another BF Bel' with mass m' describes, for $m'(y) \in \mathbb{R}^1$, a straight line in the belief space, except the point with coordinates: $F_x(Bel) = [1, -\frac{m(\Theta_2)}{m(x)}]'$,² which coincides with the limit of $Bel \oplus Bel'$ for $m'(y) \rightarrow \pm\infty$. This is true for every value of $m'(x) \in [0, 1]$. Indeed, all the collections of Dempster's sums $Bel \oplus Bel'$ with $m'(x) = k = const$ have a common intersection at the point $F_x(Bel)$, which is located outside the belief space. In the same way, this holds for the sets $\{Bel \oplus Bel' : m'(y) = l = const\}$, which each form a distinct line passing through a twin point: $F_y(Bel) = [-\frac{m(\Theta)}{m(y)}, 1]'$.

We call $F_x(Bel), F_y(Bel)$ the *foci* of the conditional subspace $\langle Bel \rangle$.

Dempster's rule thus admits an elegant geometric construction in the belief space, illustrated, for the binary case, in Fig. 1.

¹ For Dempster's rule can be extended to pseudo belief functions.

² We write $m(x)$ instead of $m(\{x\})$, Bel_x rather than $Bel_{\{x\}}$ to simplify the notation.

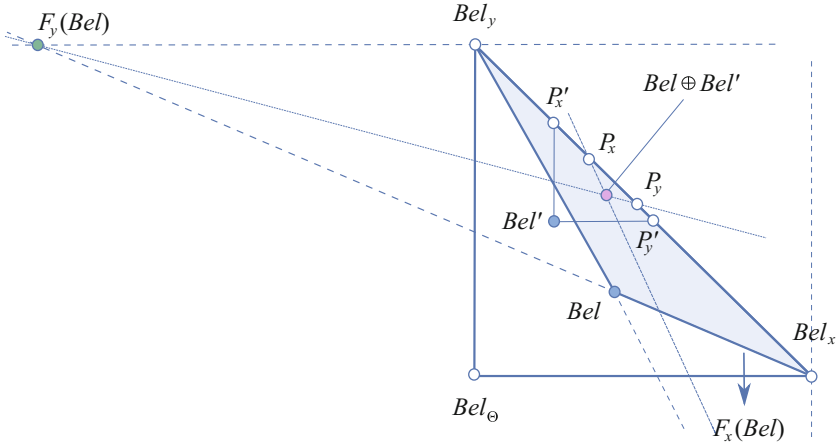


Fig. 1. Graphical construction of Dempster’s combination in the binary belief space.

Algorithm 1. Dempster’s rule: geometric construction in \mathcal{B}_2 .

- 1: **procedure** GEODEMPSTER2(Bel, Bel')
 - 2: compute the foci $F_x(Bel), F_y(Bel)$ of the conditional subspace $\langle Bel \rangle$;
 - 3: project Bel' onto \mathcal{P} along the orthogonal directions, obtaining P'_x and P'_y ;
 - 4: combine Bel with P'_x and P'_y (a much simpler operation) to get P_x and P_y ;
 - 5: draw the lines $\overline{P_x F_x(Bel)}$ and $\overline{P_y F_y(Bel)}$: their intersection is the desired orthogonal sum $Bel \oplus Bel'$.
 - 6: **end procedure**
-

These notions can be naturally extended to finite frames with an arbitrary number $|\Theta|$ of elements ([3], Chap. 8).

3 Geometry of Yager’s and Dubois’ Rules

Yager’s and Dubois’ rules. Yager’s rule [10] is based on the view that conflict is generated by non-reliable information sources. In response, the conflicting mass (here denoted by $m_\cap(\emptyset)$) is re-assigned to the whole frame of discernment Θ :

$$m_\otimes(A) = \begin{cases} m_\cap(A) & \emptyset \neq A \subsetneq \Theta \\ m_\cap(\Theta) + m_\cap(\emptyset) & A = \Theta. \end{cases} \quad (1)$$

The combination operator proposed by Dubois and Prade [5] comes from applying the *minimum specificity* principle to the cases in which the focal elements B, C of two input BF’s do not intersect, and assigns their product mass to $B \cup C$:

$$m_D(A) = m_\cap(A) + \sum_{B \cup C = A, B \cap C = \emptyset} m_1(B)m_2(C). \quad (2)$$

Analysis on binary frames. On binary frames, $\Theta = \{x, y\}$ Yager's rule (1) and Dubois' rule (2) coincide, as the only conflicting focal elements are $\{x\}$ and $\{y\}$, whose union is Θ itself:

$$\begin{aligned} m_{\otimes}(x) &= m_1(x)(1 - m_2(y)) + m_1(\Theta)m_2(x), \\ m_{\otimes}(y) &= m_1(y)(1 - m_2(x)) + m_1(\Theta)m_2(y), \\ m_{\otimes}(\Theta) &= m_1(x)m_2(y) + m_1(y)m_2(x) + m_1(\Theta)m_2(\Theta). \end{aligned} \tag{3}$$

Using (3) we can easily show that:

$$\begin{aligned} Bel_{\otimes}Bel_x &= [m(x) + m(\Theta), 0, m(y)]'; & Bel_{\otimes}Bel_y &= [0, m(y) + m(\Theta), m(x)]'; \\ Bel_{\otimes}Bel_{\Theta} &= Bel = [m(x), m(y), m(\Theta)], \end{aligned} \tag{4}$$

once adopting the vector notation $Bel = [Bel(x), Bel(y), Bel(\Theta)]'$.

The conditional subspace $\langle Bel \rangle_{\otimes}$ (Fig. 2(left)) is thus the convex closure of the points (4): $\langle Bel \rangle_{\otimes} = Cl(Bel, Bel_{\otimes}Bel_x, Bel_{\otimes}Bel_y)$.

Comparing (3) with (4), it is easy to see that

$$Bel_1 \otimes Bel_2 = m_2(x)(Bel_1 \otimes Bel_x) + m_2(y)(Bel_1 \otimes Bel_y) + m_2(\Theta)(Bel_1 \otimes Bel_{\Theta}),$$

i.e., the simplicial coordinates of Bel_2 in the binary belief space \mathcal{B}_2 and of the Yager combination $Bel_1 \otimes Bel_2$ in the conditional subspace $\langle Bel_1 \rangle_{\otimes}$ coincide.

We can then conjecture the following.

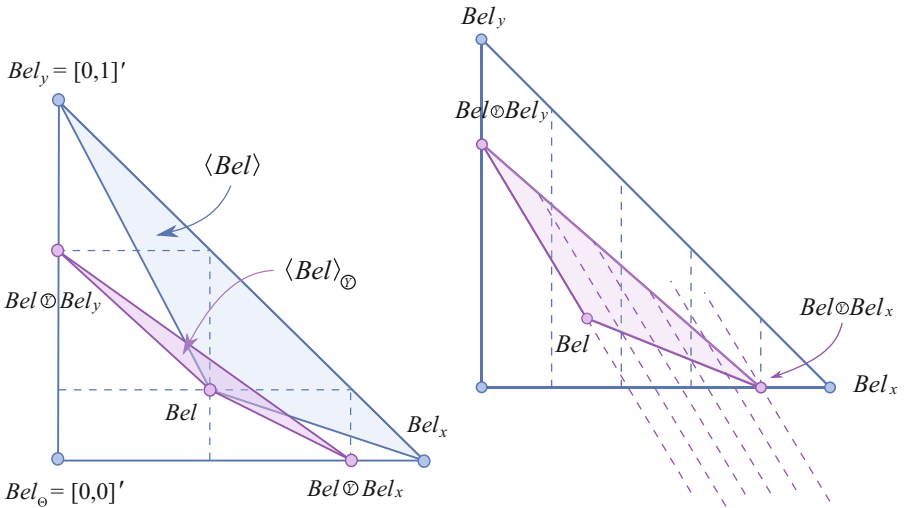


Fig. 2. (Left) Conditional subspace $\langle Bel \rangle_{\otimes}$ for Yager's (and Dubois') combination rule on a binary frame $\Theta = \{x, y\}$. Dempster's $\langle Bel \rangle$ is also shown for comparison. (Right) In Yager's combination, the images of constant mass loci (dashed blue segments) do not converge to a focus, but form parallel lines (dashed purple, cfr. Fig. 1). (Color figure online)

Conjecture 1. Yager combination and *affine combination* commute. Namely:

$$Bel \circledast \left(\sum_i \alpha_i Bel_i \right) = \sum_i \alpha_i Bel \circledast Bel_i, \quad \alpha_i \in \mathbb{R} \forall i, \sum_i \alpha_i = 1.$$

As commutativity is the basis for the geometric analysis of Dempster's rule [1], this opens the way for a similar geometric construction for Yager's rule. However, as shown in Fig. 2 (right), images of constant mass loci under Yager's rule are parallel, and there are no foci. From (3) it follows that:

$$\lim_{m_2(y) \rightarrow -\infty} \frac{m_{\circledast}(y)}{m_{\circledast}(x)} = \frac{m_1(y)(1 - m_2(x)) + m_1(\Theta)m_2(y)}{m_1(x)(1 - m_2(y)) + m_1(\Theta)m_2(x)} = -\frac{m_1(\Theta)}{m_1(x)},$$

and similarly for the loci with $m_2(y) = \text{const}$.

Nevertheless, as we will rigorously prove in upcoming work, Yager's combination also admits a geometric construction based on intersecting linear spaces which are images of constant mass loci.

4 Geometry of Disjunctive Combination

Disjunctive combination [8] is the natural, cautious dual of Dempster's combination. The operator follows from the assumption that the consensus between two sources of evidence is best represented by the union of the supported hypotheses, rather than by their intersection. An algebraic analysis of disjunctive combination on binary frames, in the form of 'Dempster semigroups', is due to Daniel [4]. Combination results are there visualised in a way similar to that presented here, although the focus is not on the geometry.

Conditional subspace. By definition: $m_{\circledcirc}(x) = m_1(x)m_2(x)$, $m_{\circledcirc}(y) = m_1(y)m_2(y)$, $m_{\circledcirc}(\Theta) = 1 - m_1(x)m_2(x) - m_1(y)m_2(y)$. Hence, in the usual vector notation:

$$\begin{aligned} Bel \circledcirc Bel_x &= [m(x), 0, 1 - m(x)]'; & Bel \circledcirc Bel_y &= [0, m(y), 1 - m(y)]'; \\ & & Bel \circledcirc Bel_{\Theta} &= Bel_{\Theta}. \end{aligned} \quad (5)$$

The conditional subspace $\langle Bel \rangle_{\circledcirc}$ is thus the convex closure of the points (5):

$$\langle Bel \rangle_{\circledcirc} = Cl(Bel, Bel \circledcirc Bel_x, Bel \circledcirc Bel_y)$$

(see Fig. 3). As in Yager's case: $Bel \circledcirc [\alpha Bel' + (1 - \alpha) Bel''] = [m(x)(\alpha m'(x) + (1 - \alpha)m''(x)), m(y)(\alpha m'(y) + (1 - \alpha)m''(y))]'$ $= \alpha Bel \circledcirc Bel' + (1 - \alpha) Bel \circledcirc Bel''$, i.e., \circledcirc commutes with *affine combination*, at least in the binary case.

Pointwise behaviour. As in Yager's case, for disjunctive combination images of constant mass loci are parallel to each other. Actually, they are parallel to the corresponding constant mass loci and the coordinate axes (observe in Fig. 3 (left) the locus $m'(x) = m''(x) = 1/3$ and its image in the conditional subspace $\langle Bel \rangle_{\circledcirc}$, with coordinate $1/3m(x)$). We can prove the following.

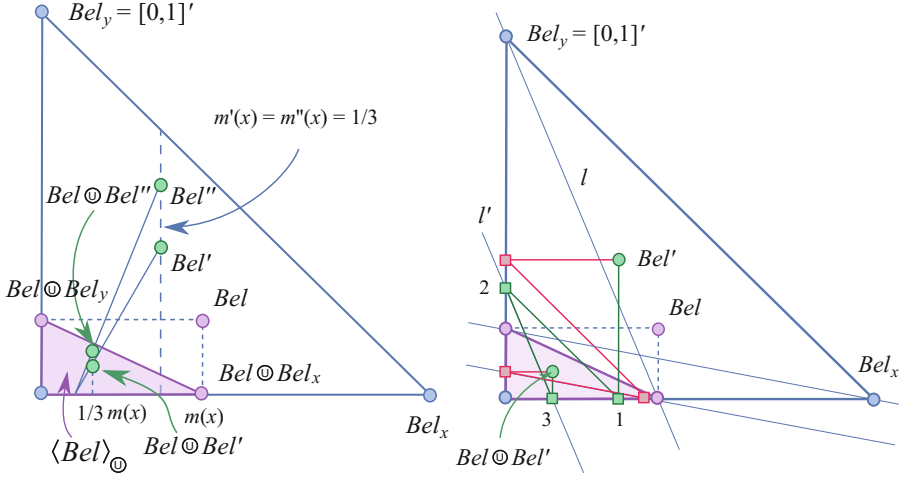


Fig. 3. (Left) Conditional subspace $\langle Bel \rangle_{\odot}$ for disjunctive combination on a binary frame. (Right) Geometric construction for the disjunctive combination of two belief functions Bel, Bel' on a binary frame.

Theorem 1. In the binary case $\Theta = \{x, y\}$, all the lines joining Bel' and $Bel_{\odot}Bel'$ for any $Bel' \in \mathcal{B}$ intersect at the point:

$$\overline{m(x)} = m'(x) \frac{m(x) - m(y)}{1 - m(y)}, \quad \overline{m(y)} = 0. \tag{6}$$

Proof. Recalling the equation of the line joining two points (χ_1, v_1) and (χ_2, v_2) of \mathbb{R}^2 , with coordinates (χ, v) : $(v - v_1) = \frac{v_2 - v_1}{\chi_2 - \chi_1}(\chi - \chi_1)$, we can identify the line joining Bel' and $Bel_{\odot}Bel'$ as:

$$(v - m'(y)) = \frac{m(y)m'(y) - m'(y)}{m(x)m'(x) - m'(x)}(\chi - m'(x)).$$

Its intersection with $v = 0$ is the point (6), which does not depend on $m'(y)$ (i.e., on the vertical location of Bel' on the constant mass loci).

A geometric construction for the disjunctive combination $Bel_{\odot}Bel'$ of two BFs in \mathcal{B}_2 is provided by simple trigonometric arguments (Fig. 3 (right)):

1. starting from Bel' , find its orthogonal projection onto the horizontal axis, with coordinate $m'(x)$ (point 1);
2. draw the line with slope 45° passing through such projection, and intersect it with the vertical axis, at coordinate $v = m'(x)$ (point 2);
3. finally, take the line l passing through Bel_y and the orthogonal projection of Bel onto the horizontal axis, and draw a parallel one l' through point 2 – its intersection with the horizontal axis (point 3) is the x coordinate $m(x)m'(x)$ of the desired combination.

A similar construction (in magenta) allows us to locate the y coordinate of the combination (as shown in Fig. 3 (right)).

5 Combination of Unnormalised Belief Functions

In the case of unnormalised belief functions (those for which $m(\emptyset) \geq 0$, UBFs [9]), Dempster's rule is replaced by *conjunctive combination*: $m_{\odot}(A) \doteq m_{\cap}(A)$. Disjunctive combination itself needs to be reassessed for UBFs as well.

In the unnormalised case, a distinction exists between the *belief* measure $Bel(A) \doteq \sum_{\emptyset \neq B \subseteq A} m(B)$ and the *believability* (in Smets' terminology) measure of an event A , denoted by: $b(A) \doteq \sum_{\emptyset \subseteq B \subseteq A} m(B)$. Here we analyse the geometric behavior of the latter, in which case \emptyset is not treated as an exception: the case of belief measures is left to future work. As $b(\Theta) = 1$, as usual, we neglect the related coordinate and represent believability functions as points of a Cartesian space of dimension $|2^{\Theta}| - 1$ (as \emptyset cannot be ignored anymore).

Conjunctive combination on the binary frame. In the case of a binary frame, the conjunctive combination of two belief functions Bel_1 and Bel_2 yields:

$$\begin{aligned} m_{\odot}(\emptyset) &= m_1(\emptyset) + m_2(\emptyset) - m_1(\emptyset)m_2(\emptyset) + m_1(x)m_2(y) + m_1(y)m_2(x), \\ m_{\odot}(x) &= m_1(x)(m_2(x) + m_2(\Theta)) + m_1(\Theta)m_2(x), \\ m_{\odot}(y) &= m_1(y)(m_2(y) + m_2(\Theta)) + m_1(\Theta)m_2(y), \\ m_{\odot}(\Theta) &= m_1(\Theta)m_2(\Theta). \end{aligned} \tag{7}$$

Conditional subspace for conjunctive combination. The global behaviour of \odot in the binary (unnormalised) case can then be understood in terms of its conditional subspace, this time in \mathbb{R}^3 . We have, after denoting $b = [b(\emptyset), b(x), b(y)]'$:

$$\begin{aligned} b_{\odot}b_{\emptyset} &= b_{\emptyset} = [1, 1, 1]'; \\ b_{\odot}b_x &= (m(\emptyset) + m(y))b_{\emptyset} + (m(x) + m(\Theta))b_x \\ &= [m(\emptyset) + m(y), 1, m(\emptyset) + m(y)]' = b(y)b_{\emptyset} + (1 - b(y))b_x; \\ b_{\odot}b_y &= (m(\emptyset) + m(x))b_{\emptyset} + (m(y) + m(\Theta))b_y \\ &= [m(\emptyset) + m(x), m(\emptyset) + m(x), 1]' = b(x)b_{\emptyset} + (1 - b(x))b_y; \\ b_{\odot}b_{\Theta} &= b, \end{aligned} \tag{8}$$

as $b_x = [0, 1, 0]'$, $b_y = [0, 0, 1]'$, $b_{\emptyset} = [1, 1, 1]'$ and $b_{\Theta} = [0, 0, 0]'$. From (8), we can note that the vertex $b_{\odot}b_x$ belongs to the line joining b_{\emptyset} and b_x , with affine coordinate given by the believability assigned by b to the other outcome y . Similarly, the vertex $b_{\odot}b_y$ belongs to the line joining b_{\emptyset} and b_y , with coordinate given by the believability assigned by b to outcome x (see Fig. 4).

Conditional subspace for disjunctive combination. As for the disjunctive combination, it is easy to see that in the unnormalised case we get: $b_{\odot}b_{\Theta} = b_{\Theta}$, $b_{\odot}b_x = b(x)b_x + (1 - b(x))b_{\Theta}$, $b_{\odot}b_{\emptyset} = b$, $b_{\odot}b_y = b(y)b_y + (1 - b(y))b_{\Theta}$, so that the conditional subspace is as in Fig. 4. Note that, in the unnormalised case, there is a unit element to \odot , namely b_{\emptyset} . We can observe a clear symmetry between the subspaces induced by disjunctive and conjunctive combination.

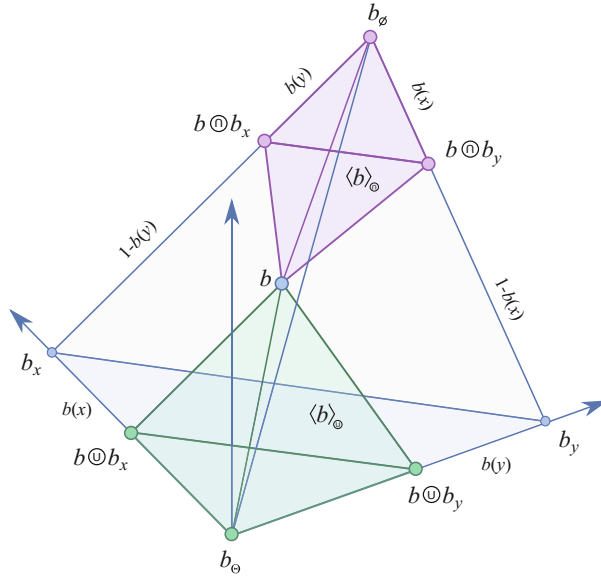


Fig. 4. Conditional subspaces induced by \odot and \ominus in a binary frame, for the case of unnormalised belief functions.

6 Conclusions

A number of questions remain open after this preliminary geometric analysis of other combination rules on binary spaces, and its extension to the case of unnormalised belief functions. In particular, the general pointwise geometric behaviour of disjunctive combination, in both the normalised and unnormalised case, needs to be understood. The question of whether disjunctive combination commutes with affine combination in general belief spaces remains open. A dual query concerns the conjunctive rule, as the alter ego of Dempster’s rule in the unnormalised case. The general pointwise geometric behaviour of conjunctive and disjunctive combinations in the unnormalised case, as well as the complete description of their conditional subspaces, will also be subject of future work. The bold and cautious rules, which are also inherently defined for unnormalised belief functions, will also be analysed.

References

1. Cuzzolin, F.: Geometry of Dempster’s rule of combination. *IEEE Trans. Syst. Man Cybern. Part B* **34**(2), 961–977 (2004)
2. Cuzzolin, F.: A geometric approach to the theory of evidence. *IEEE Trans. Syst. Man Cybern. Part C* **38**(4), 522–534 (2008)
3. Cuzzolin, F.: *The Geometry of Uncertainty*. Springer, New York (2018). <https://doi.org/10.1007/978-1-4615-0813-7>

4. Daniel, M.: Algebraic structures related to the combination of belief functions. *Scientiae Mathematicae Japonicae* **60**(2), 501–511 (2004)
5. Dubois, D., Prade, H.: Representation and combination of uncertainty with belief functions and possibility measures. *Comput. Intell.* **4**(3), 244–264 (1988)
6. Hajek, P., Valdes, J.J.: Generalized algebraic foundations of uncertainty processing in rule-based expert systems (dempsteroids). *Comput. Artif. Intell.* **10**(1), 29–42 (1991)
7. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
8. Smets, P.: Belief functions : the disjunctive rule of combination and the generalized Bayesian theorem. *Int. J. Approx. Reason.* **9**, 1–35 (1993)
9. Smets, P.: The nature of the unnormalized beliefs encountered in the transferable belief model. In: *Proceedings of UAI 1992*, pp. 292–297 (1992)
10. Yager, R.R.: On the Dempster-Shafer framework and new combination rules. *Inf. Sci.* **41**(2), 93–137 (1987)



Logistic Regression Revisited: Belief Function Analysis

Thierry Denoeux^(✉)

Université de Technologie de Compiègne, CNRS,
UMR 7253 Heudiasyc, Compiègne, France
tdenoeux@utc.fr

Abstract. We show that the weighted sum and softmax operations performed in logistic regression classifiers can be interpreted in terms of evidence aggregation using Dempster's rule of combination. From that perspective, the output probabilities from such classifiers can be seen as normalized plausibilities, for some mass functions that can be laid bare. This finding suggests that the theory of belief functions is a more general framework for classifier construction than is usually considered.

Keywords: Evidence theory · Dempster-Shafer theory
Classification · Machine learning

1 Introduction

In the last twenty years, the Dempster-Shafer (DS) theory of belief functions has been increasingly applied to classification. One direction of research is classifier fusion: classifier outputs are expressed as belief functions and combined by Dempster's rule or any other rule (see, e.g., [1, 7, 8]). Another approach is to design *evidential classifiers*, which can be defined as classifiers built from basic principles of DS theory. Typically, an evidential classifier has the structure depicted in Fig. 1: when presented by a feature vector x , the system computes k mass functions m_1, \dots, m_k defined on the set Θ of classes, based on a learning set. These mass functions are then combined using Dempster's rule, or any other rule. The first evidential classifier was the evidential k -nearest neighbor classifier [3], in which mass functions m_j are constructed from the k nearest neighbor of x , and combined by Dempster's rule. In the evidential neural network classifier [5], a similar principle is applied, but mass functions are constructed based on the distances to prototypes, and the whole system is trained to minimize an error function.

In this paper, we show that not only these particular distance-based classifiers, but also a broad class of widely-used classifiers, including logistic regression and its nonlinear extensions, can be seen as evidential classifiers. This finding leads us to the conclusion that DS theory is a much more general framework for classifier construction than was initially believed.

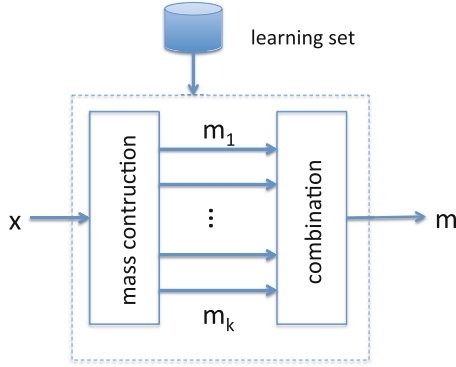


Fig. 1. Basic structure of an evidential classifier.

The rest of the paper is organized as follows. Some background definitions will first be recalled in Sect. 2. A general model of feature-based evidence will be described in Sect. 3, where we will show that the normalized plausibility function, after combining the evidence of J features, is identical to the output of logistic regression. The recovery of the full mass function will then be addressed, and a simple example will be given in Sect. 4. Section 5 will conclude the paper.

2 Background

In this section, we first recall some basic notions and definitions needed in the rest of the paper. The notion of weight of evidence will first be recalled in Sect. 2.1, and some notations for logistic regression will be introduced in Sect. 2.2.

2.1 Weights of Evidence

Let us consider a simple mass function m on a frame Θ , such that

$$m(A) = s, \quad m(\Theta) = 1 - s,$$

where s is a degree of support in $[0, 1]$. Typically, such a mass function represents some elementary piece of evidence supporting hypothesis A . Shafer [9, p. 77] defines the *weight* of this evidence as $w = -\ln(1 - s)$. Conversely, we thus have $s = 1 - \exp(-w)$. The rationale for this definition is that weights of evidence are additive: if m_1 and m_2 are two simple mass functions focussed on the same subset A , with weights w_1 and w_2 , then the orthogonal sum $m_1 \oplus m_2$ corresponds to the weight $w_1 + w_2$. If we denote a simple mass function with focal set A and weight w by A^w , we thus have $A^{w_1} \oplus A^{w_2} = A^{w_1 + w_2}$. It follows that any separable mass function can be written as $m = \bigoplus_{\emptyset \neq A \subseteq \Theta} A^{w_A}$, where w_A is the weight of evidence pointing to A . We note that, in [6], following [10], we used the term “weight” for $-\ln w$. As we will see, the additivity property is central in our analysis: we thus stick to Shafer’s terminology and notation in this paper.

2.2 Logistic Regression

Consider a multi-category¹ classification problem with J -dimensional feature vector $x = (x_1, \dots, x_J)$ and class variable $Y \in \Theta = \{\theta_1, \dots, \theta_K\}$ with $K > 2$. In the logistic regression model, we assume the logarithms of the posterior class probabilities $\mathbb{P}(Y = \theta_k|x)$ to be affine functions of x , i.e.,

$$\ln \mathbb{P}(Y = \theta_k|x) = \sum_{j=1}^J \beta_{jk} x_j + \beta_{0k} + \gamma, \quad \forall k \in \llbracket 1, K \rrbracket, \quad (1)$$

where β_{jk} , $j = 0, \dots, J$ are parameters and γ is a constant. Using the equation $\sum_{k=1}^K \mathbb{P}(Y = \theta_k|x) = 1$, we easily get the following expressions for the posterior probabilities,

$$\mathbb{P}(Y = \theta_k|x) = \frac{\exp\left(\sum_{j=1}^J \beta_{jk} x_j + \beta_{0k}\right)}{\sum_{l=1}^K \exp\left(\sum_{j=1}^J \beta_{jl} x_j + \beta_{0l}\right)}. \quad (2)$$

This transformation from arbitrary real quantities (1) to probabilities is sometimes referred to as the *softmax transformation*. Parameters β_{jk} are usually estimated by maximizing the conditional likelihood. In feedforward neural networks with a softmax output layer, a similar approach is used, with variables x_j defined as the outputs of the last hidden layer of neurons. These variables are themselves defined as complex nonlinear functions of the input variables, which are optimized together with the decision layer weights β_{jk} . Logistic regression is functionally equivalent to a feedforward neural network with no hidden layer.

3 Model

We consider a multi-category classification problem as described in Sect. 2.2. We assume that each feature x_j provides some evidence about the class variable Y . For each θ_k , the evidence of feature x_j points either to the singleton $\{\theta_k\}$ or to its complement $\overline{\{\theta_k\}}$, depending on the sign of

$$w_{jk} = \beta_{jk} x_j + \alpha_{jk}, \quad (3)$$

where $(\beta_{jk}, \alpha_{jk})$, $k = 1, \dots, K$, $j = 1, \dots, J$ are parameters. The *weights of evidence* for $\{\theta_k\}$ and $\overline{\{\theta_k\}}$ are, respectively,

$$w_{jk}^+ = (w_{jk})_+ \quad \text{and} \quad w_{jk}^- = (w_{jk})_-, \quad (4)$$

where $(\cdot)_+$ and $(\cdot)_-$ denote, respectively, the positive and the negative parts. For each feature x_j and each class θ_k , we thus have two simple mass functions

$$m_{jk}^+ = \{\theta_k\}^{w_{jk}^+} \quad \text{and} \quad m_{jk}^- = \overline{\{\theta_k\}}^{w_{jk}^-}. \quad (5)$$

¹ The case of binary classification with $K = 2$ classes requires a separate treatment. Due to space constraints, we focus on the multi-category case in this paper.

Assuming these mass functions to be independent, they can be combined by Dempster's rule. Let

$$m_k^+ = \bigoplus_{j=1}^J m_{jk}^+ = \{\theta_k\}^{w_k^+} \quad \text{and} \quad m_k^- = \bigoplus_{j=1}^J m_{jk}^- = \overline{\{\theta_k\}}^{w_k^-}$$

where

$$w_k^+ = \sum_{j=1}^J w_{jk}^+ \quad \text{and} \quad w_k^- = \sum_{j=1}^J w_{jk}^-. \quad (6)$$

The contour functions pl_k^+ and pl_k^- associated, respectively, with m_k^+ and m_k^- are

$$pl_k^+(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_k, \\ \exp(-w_k^+) & \text{otherwise,} \end{cases}$$

and

$$pl_k^-(\theta) = \begin{cases} \exp(-w_k^-) & \text{if } \theta = \theta_k, \\ 1 & \text{otherwise.} \end{cases}$$

Now, let

$$m^+ = \bigoplus_{k=1}^K m_k^+ \quad \text{and} \quad m^- = \bigoplus_{k=1}^K m_k^-,$$

and let pl^+ and pl^- be the corresponding contour functions. We have

$$\begin{aligned} pl^+(\theta_k) &\propto \prod_{l=1}^K pl_l^+(\theta_k) = \exp\left(-\sum_{l \neq k} w_l^+\right) = \exp\left(-\sum_{l=1}^K w_l^+\right) \exp(w_k^+) \\ &\propto \exp(w_k^+), \end{aligned}$$

and

$$pl^-(\theta_k) \propto \prod_{l=1}^K pl_l^-(\theta_k) = \exp(-w_k^-).$$

Finally, let $m = m^+ \oplus m^-$ and let pl be the corresponding contour function. We have

$$\begin{aligned} pl(\theta_k) &\propto pl^+(\theta_k)pl^-(\theta_k) \propto \exp(w_k^+ - w_k^-) \\ &\propto \exp\left(\sum_{j=1}^J w_{jk}\right) = \exp\left(\sum_{j=1}^J \beta_{jk}x_j + \sum_{j=1}^J \alpha_{jk}\right). \end{aligned}$$

Let p be the probability mass function induced from m by the plausibility-probability transformation [2], and let

$$\beta_{0k} = \sum_{j=1}^J \alpha_{jk}. \quad (7)$$

We have

$$p(\theta_k) = \frac{\exp\left(\sum_{j=1}^J \beta_{jk}x_j + \beta_{0k}\right)}{\sum_{l=1}^K \exp\left(\sum_{j=1}^J \beta_{jl}x_j + \beta_{0l}\right)}, \quad (8)$$

which is equivalent to (2). We thus have proved that the output probabilities computed by a logistic regression classifier can be seen as the normalized plausibilities obtained after combining elementary mass functions (5) by Dempster’s rule: these classifiers are, thus, evidential classifiers as defined in Sect. 1.

4 Recovering the Mass Function

Having shown that the output probabilities of logistic regression classifiers are normalized plausibilities, it is interesting to recover the underlying output mass function, defined as

$$m = \bigoplus_{k=1}^K \left(\{\theta_k\}^{w_k^+} \oplus \overline{\{\theta_k\}}^{w_k^-} \right). \quad (9)$$

Its complete expression can be derived (after some tedious calculation), but it cannot be given here for lack of space.

There is, however, a difficulty related to the identifiability of the weights w_k^+ and w_k^- . First, parameters β_{jk} are not themselves identifiable, because adding any constant vector \mathbf{c} to each vector $\beta_k = (\beta_{0k}, \dots, \beta_{Jk})$ produces the same normalized plausibilities (8). Secondly, for given β_{0k} , any α_{jk} verifying (7) will yield the same probabilities (8). This problem is addressed in the next section.

4.1 Identification

To identifying the underlying output mass function, we propose to apply the Least Commitment Principle, by searching for the mass function m^* of the form (9) verifying (8) and such that the sum of the squared weights of evidence is minimum. More precisely, let $\{(x_i, y_i)\}_{i=1}^n$ be a learning set, let $\widehat{\beta}_{jk}$ be the maximum likelihood estimates of the weights β_{jk} , and let $\boldsymbol{\alpha}$ denote the vector of parameters α_{jk} . Any $\beta_{jk}^* = \widehat{\beta}_{jk} + c_j$ will verify (8). The parameter values β_{jk}^* and α_{jk}^* minimizing the sum of the squared weights of evidence can thus be found by solving the following minimization problem

$$\min f(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{i=1}^n \sum_{j=1}^J \sum_{k=1}^K \left[(\widehat{\beta}_{jk} + c_j)x_{ij} + \alpha_{jk} \right]^2 \quad (10)$$

subject to

$$\sum_{j=1}^J \alpha_{jk} = \widehat{\beta}_{0k} + c_0, \quad \forall k \in \llbracket 1, K \rrbracket. \quad (11)$$

In (10), x_{ij} denotes the value of feature j for learning vector x_i . Developing the square in (10), we get

$$f(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{j,k} (\widehat{\beta}_{jk} + c_j)^2 \left(\sum_i x_{ij}^2 \right) + n \sum_{j,k} \alpha_{jk}^2 + 2 \sum_{j,k} (\widehat{\beta}_{jk} + c_j) \alpha_{jk} \sum_i x_{ij}. \quad (12)$$

Assuming that the input variables x_j have been centered, we have $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = s_j^2$, where s_j^2 is the empirical variance of feature x_j . Equation (12) then simplifies to

$$f(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{j,k} s_j^2 (\widehat{\beta}_{jk} + c_j)^2 + n \sum_{j,k} \alpha_{jk}^2. \quad (13)$$

Due to constraint (11), for any c_0 , the second term in the right-hand side of (13) is minimized for $\alpha_{jk} = \frac{1}{J}(\widehat{\beta}_{0k} + c_0)$, for all $j \in \llbracket 1, J \rrbracket$ and $k \in \llbracket 1, K \rrbracket$. Hence, the problem becomes

$$\min_{\mathbf{c}} f(\mathbf{c}) = \sum_{j=1}^J s_j^2 \left\{ \sum_{k=1}^K (\widehat{\beta}_{jk} + c_j)^2 \right\} + \frac{n}{J} \sum_{k=1}^K (\widehat{\beta}_{0k} + c_0)^2.$$

Each of the $J + 1$ terms in this sum can be minimized separately. The solution can easily be found to be

$$c_j^* = -\frac{1}{K} \sum_{k=1}^K \widehat{\beta}_{jk}, \quad \forall j \in \llbracket 0, J \rrbracket$$

The optimum coefficients are, thus,

$$\beta_{jk}^* = \widehat{\beta}_{jk} - \frac{1}{K} \sum_{l=1}^K \widehat{\beta}_{jl}, \quad \forall j \in \llbracket 0, J \rrbracket, \forall k \in \llbracket 1, K \rrbracket$$

and

$$\alpha_{jk}^* = \beta_{0k}^* / J, \quad \forall j \in \llbracket 1, J \rrbracket, \forall k \in \llbracket 1, K \rrbracket. \quad (14)$$

To get the least committed mass function m^* with minimum sum of squared weights of evidence and verifying (8), we thus need to center the rows of the $(J + 1) \times K$ matrix $B = (\beta_{jk}^*)$, set α_{jk}^* according to (14), and compute the weights of evidence w_k^- and w_k^+ from (3), (4) and (6).

4.2 Example

As a simple example, let us consider simulated data with $J = 1$ feature, $K = 3$ classes, and Gaussian conditional distributions $X|\theta_k \sim \mathcal{N}(\mu_k, 1)$, with $\mu_1 = -1$, $\mu_2 = 0$ and $\mu_3 = 1$. We randomly generated 10,000 from each of the three conditional distributions, we standardized the data and we trained a logistic regression classifier on these data. Decisions are usually based on the posterior

class probabilities $\mathbb{P}(\theta_k|x)$ displayed in Fig. 2(a). Figure 3 shows the underlying masses, computed as explained in Sect. 4.1. As we can see, masses are assigned to subsets of classes in regions where these classes overlap, as could be expected. Figure 2(b) shows the contour functions $pl(\theta_k|x)$ vs x . Interestingly, the graphs of these functions have quite different shapes, as compared to those of the posterior probabilities shown in Fig. 2(a). Whereas decisions with probabilistic classifiers are classically based on minimum expected loss, seeing logistic regression classifiers as evidential classifiers opens the possibility to experiment with other rules such as minimum lower or upper expected loss [4] or interval dominance [11].

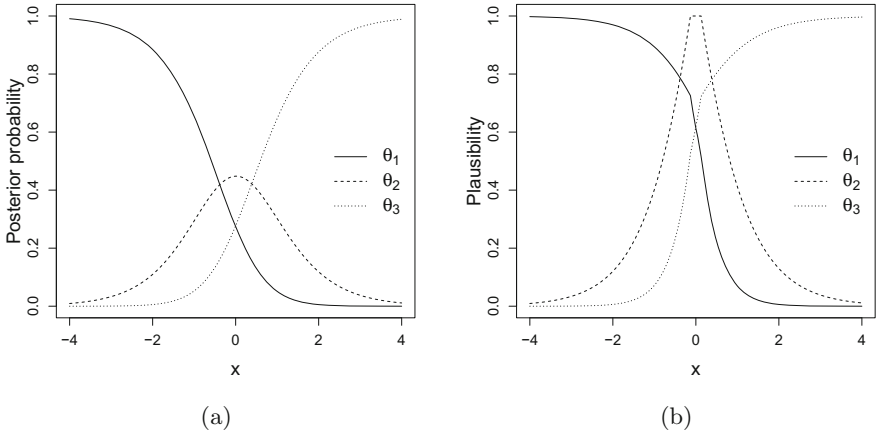


Fig. 2. Posterior class probabilities $\mathbb{P}(\theta_k|x)$ (a) and contour functions $pl(\theta_k|x)$ for the logistic regression example.

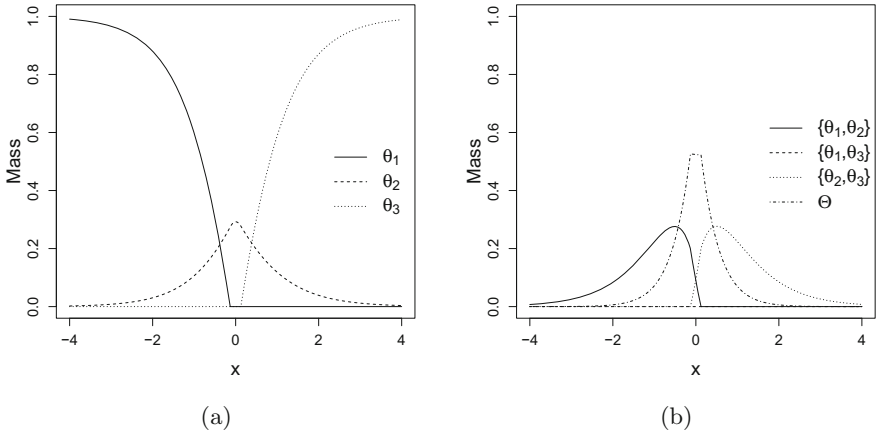


Fig. 3. Masses on singletons (a) and compound hypotheses (b) vs. x for the logistic regression example.

5 Conclusions

We have shown that logistic regression classifiers and also, as a consequence, generalized linear classifiers such as feedforward neural network classifiers, which essentially perform logistic regression in the output layer, can be seen as pooling evidence using Dempster's rule of combination. This finding may have important implications, as it opens the way to a DS analysis of many widely used classifiers, beyond the particular distance-based classifiers introduced in [3, 5]. In future work, we will deepen this analysis by exploring the consequences of viewing neural network classifiers as evidential classifiers, in terms of decision strategies, classifier fusion, and handling missing or uncertain inputs, among other research directions.

References

1. Bi, Y., Guan, J., Bell, D.: The combination of multiple classifiers using an evidential reasoning approach. *Artif. Intell.* **172**(15), 1731–1751 (2008)
2. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. *Int. J. Approximate Reasoning* **41**(3), 314–330 (2006)
3. Denœux, T.: A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(05), 804–813 (1995)
4. Denœux, T.: Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recogn.* **30**(7), 1095–1107 (1997)
5. Denœux, T.: A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern. A* **30**(2), 131–150 (2000)
6. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artif. Intell.* **172**, 234–264 (2008)
7. Quost, B., Masson, M.-H., Denœux, T.: Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *Int. J. Approximate Reasoning* **52**(3), 353–374 (2011)
8. Rogova, G.: Combining the results of several neural network classifiers. *Neural Netw.* **7**(5), 777–781 (1994)
9. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
10. Smets, P.: The canonical decomposition of a weighted belief. In: *International Joint Conference on Artificial Intelligence*, pp. 1896–1901. Morgan Kaufman, San Mateo (1995)
11. Troffaes, M.C.: Decision making under uncertainty using imprecise probabilities. *Int. J. Approximate Reasoning* **45**(1), 17–29 (2007)



From Relations Between Sets to Relations Between Belief Functions

Sébastien Destercke¹(✉), Frédéric Pichon², and John Klein³

¹ UMR CNRS 7253 Heudiasyc, Sorbonne Universités, Université de technologie de Compiègne CS 60319, 60203 Compiègne cedex, France

`sebastien.destercke@hds.utc.fr`

² Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), 62400 Béthune, France

`frederic.pichon@univ-artois.fr`

³ Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRISTAL - Centre de Recherche en Informatique Signal et Automatique de Lille, 59000 Lille, France

`john.klein@univ-lille1.fr`

Abstract. In uncertainty theories, a common problem is to define how we can extend relations between sets (e.g., inclusion, ranking, consistency, . . .) to corresponding notions between uncertainty representations. Such definitions can then be used to perform the same operations as those that are done for sets: measuring information content, ordering alternatives or checking consistency, to name a few. In this paper, we propose a general way to extend set relations to belief functions, using constrained stochastic matrices to identify those belief functions in relation. We then study some properties of our proposal, as well as its relations with existing works focusing on peculiar relations.

Keywords: Set relations · Belief functions · Specificity · Ranking Consistency

1 Introduction

One can define many relations between two (or more) subsets A, B of some finite set X , i.e. between elements of some boolean algebra $(2^X, \cap, \cup, .C)$. Such relations can check whether the sets are consistent ($A \cap B \neq \emptyset$); whether one set is more informative than another, or imply one another ($A \subseteq B$); when the space on which they are defined is ordered, whether one set is “higher” than another ($A \prec B$); etc. These relations can then be related to practical problems such as restoring consistency or ranking alternatives.

To address the same questions in those uncertainty theories that formally generalize set theory (based, e.g., on possibility distributions, belief functions or sets of probabilities), it is desirable to carry over relations between sets to uncertainty representations. Given the higher expressiveness of such theories, the problem is ill-posed in the sense that there is not a unique way to do so.

We can cite as a typical example the notion of inclusion between belief functions, that has many definitions [3]. Yet, such works usually focus on extending one particular relation in meaningful ways.

In this paper, we propose a simple way to extend any set relation to an equivalent relation between belief functions, in the sense that the relation is exactly recovered when considering categorical belief functions (i.e., belief functions reduced to one focal element). Basically, we require that for a pair of belief functions to be in relation, there must exist at least one stochastic matrix such that one of these belief functions is obtained as the dot product of the matrix with the other belief function. Additionally, the matrix is constrained to have non-null entries on pairs of focal sets satisfying the relation to extend.

We develop and study the properties of our proposal in Sect. 2, in which we also include necessary reminders. We then show in Sect. 3 how this proposal is linked to previously proposed relations between belief functions, as well as to other related results. We will focus, in particular, on the notions of information specificity, of consistency, and of ranking. Finally, we formalize in Sect. 4 how we can say whether a relation is preserved through functional mapping of a variable to another one, and provide some results about the inclusion and ranking cases.

2 Main Proposal

2.1 Definitions

A belief function on a finite space $X = \{x_1, \dots, x_K\}$ is in one-to-one correspondence with a mass function $m_i : 2^X \rightarrow [0, 1]$ that satisfies $\sum_{A \subseteq X} m_i(A) = 1$. From such a mass function, the belief and plausibility of an event A respectively read

$$Bel_i(A) = \sum_{E \subseteq A} m_i(E) \text{ and } Pl_i(A) = \sum_{E \cap A \neq \emptyset} m_i(E).$$

If $m_i(\emptyset) = 0$, they can be interpreted as bounds of the probability $P(A)$ of A , inducing the probability set $\mathcal{P}_i = \{P : Bel_i(A) \leq P(A) \leq Pl_i(A), \forall A \subseteq X\}$. We denote by \mathcal{B}^X the set of all belief functions on X . A particularly interesting subclass of belief functions for this study will be the one of categorical ones. A categorical mass function, denoted m_B , is such that $m_B(B) = 1$.

Let us now consider a relation such that for any ordered pair $(A, B) \subseteq X^2$, we will denote by \mathbf{ARB} the truth of the relation between A and B (\mathbf{R} is thus a binary relation on 2^X , or equivalently a subset of $2^X \times 2^X$). We then propose the following simple definition to extend this relation to belief functions, i.e. into a relation on \mathcal{B}^X :

Definition 1. *Given two mass functions m_1, m_2 and a subset relation \mathbf{R} , we say that $m_1 \tilde{\mathbf{R}} m_2$ iff there is a (left)¹ stochastic matrix $S_{\mathbf{R}}$ such that $\forall A, B \subseteq X$*

$$m_1(A) = \sum_{B \subseteq X} S_{\mathbf{R}}(A, B) m_2(B) \tag{1}$$

$$\text{with } S(A, B) > 0 \wedge m_2(B) > 0 \implies \mathbf{ARB}. \tag{2}$$

¹ We use left-stochasticity only throughout the paper.

It is easily checked that $\tilde{\mathbf{R}}$ is a generalisation of \mathbf{R} in the sense that

$$m_A \tilde{\mathbf{R}} m_B \Leftrightarrow \mathbf{A} \mathbf{R} \mathbf{B}, \quad \forall A, B \subseteq X. \quad (3)$$

Indeed, if $\mathbf{A} \mathbf{R} \mathbf{B}$, we can choose $S_{\mathbf{R}}(E, F) = m_A(E)$ and this matrix matches the conditions of Definition 1, hence $m_A \tilde{\mathbf{R}} m_B$. Also, there is only one relation $\tilde{\mathbf{R}}$ on belief functions spanned by Definition 1 from a given set relation \mathbf{R} . Suppose two such belief function relations exist. If a matrix matching the conditions of Definition 1 was found for the first one then the same matrix also works for the other and the relations are equivalent. Similarly, if $\tilde{\mathbf{R}}$ is spanned by Definition 1 from a given set relation \mathbf{R} then it cannot be spanned by other set relations in the same way. This is an immediate consequence of (3). Consequently, we will use the same notation for a relation \mathbf{R} on the subset or belief function side in the remainder of the paper.

Definition 1 is inspired from previous works on specificity of belief functions [3, 4, 6], as well as on recent proposals dealing with belief function ordering [5]. As these works dealt with directional, or rather asymmetric relations, Definition 1 is naturally asymmetric. However, Proposition 1 shows that it has a somehow symmetric counterpart.

Proposition 1. *Consider two mass functions m_1, m_2 and a belief function relation \mathbf{R} . Then the two following conditions are equivalent:*

1. *there is a stochastic matrix $S_{\mathbf{R}}(A, B)$ such that*

$$m_1(A) = \sum_{B \subseteq X} S_{\mathbf{R}}(A, B) m_2(B),$$

$$\text{with } S_{\mathbf{R}}(A, B) > 0 \wedge m_2(B) > 0 \implies \mathbf{A} \mathbf{R} \mathbf{B}.$$

2. *there is a joint mass function m_{12} on $2^X \times 2^X$ such that*

$$m_{12}(A, B) > 0 \implies \mathbf{A} \mathbf{R} \mathbf{B}, \quad (4)$$

$$m_1(A) = \sum_B m_{12}(A, B), \quad (5)$$

$$m_2(B) = \sum_A m_{12}(A, B). \quad (6)$$

Proof (Sketch). 1. \implies 2: from a matrix $S_{\mathbf{R}}(A, B)$, we can deduce a joint $m_{12}(A, B) = m_2(B) S_{\mathbf{R}}(A, B)$ for any A, B which satisfies 2.

2. \implies 1: from a joint $m_{12}(A, B)$ satisfying (4)–(6), define $S_{\mathbf{R}}(A, B) = m_{12}(A, B) / m_2(B)$ if $m_2(B) > 0$, and with arbitrary values making it (left)-stochastic if $m_2(B) = 0$. This matrix satisfies 1.

This proposition shows, in particular, that any stochastic matrix $S_{\mathbf{R}}$ can be associated to a unique joint mass function m_{12} , and vice-versa. Also note that, using a transformation similar to the one used in the proof, we can build a stochastic matrix $S'_{\mathbf{R}}$ such that $S'_{\mathbf{R}}(B, A) = m_{12}(A, B) / m_1(A)$ if $m_1(A) > 0$, and be with arbitrary values else. S' is such that $m_2(B) = \sum_{A \subseteq X} S'_{\mathbf{R}}(B, A) m_1(A)$ with $S'_{\mathbf{R}}(A, B) > 0$ and $m_1(A) > 0$ implying $\mathbf{B} \mathbf{R} \mathbf{A}$, but not necessarily $\mathbf{A} \mathbf{R} \mathbf{B}$.

2.2 Relation Properties Preservation

We may now wonder how much of the initial relation \mathbf{R} properties between sets do still exist when extended in this way to belief functions. We will now provide a series of results for common properties, either by providing proofs or counter-examples. We will keep the proposition/proof format, to provide a uniform presentation.

Proposition 2 (Preserved symmetry). *If \mathbf{R} is symmetric on sets, it is so on belief functions:*

$$m_1 \mathbf{R} m_2 \equiv m_2 \mathbf{R} m_1, \forall m_1, m_2$$

Proof (Sketch). If \mathbf{R} is symmetric, then $S'_{\mathbf{R}}(A, B)$ as defined above is such that $S'_{\mathbf{R}}(A, B) > 0$ and $m_1(A) > 0$ implies $A \mathbf{R} B$.

Proposition 3 (Unpreserved antisymmetry). *If \mathbf{R} is antisymmetric on sets, it is not necessarily so on belief functions, as we may have*

$$m_1 \mathbf{R} m_2 \wedge m_2 \mathbf{R} m_1 \text{ and } m_2 \neq m_1$$

Proof. Consider two mass functions that are positives only on subsets A, B, C and such that

$$\begin{aligned} m_1(A) &= 0.3, m_1(B) = 0.5, m_1(C) = 0.2, \\ m_2(A) &= 0.4, m_2(B) = 0.3, m_2(C) = 0.3, \end{aligned}$$

and the antisymmetric relation \mathbf{R} on A, B, C summarised by the matrix

$$\begin{array}{c} \begin{array}{ccc} & A & B & C \\ A & \left[\begin{array}{cc} A \mathbf{R} A & A \mathbf{R} B \\ B \mathbf{R} B & B \mathbf{R} C \end{array} \right] \\ B & & & \\ C & \left[\begin{array}{cc} C \mathbf{R} A & C \mathbf{R} C \end{array} \right] \end{array} \end{array}$$

We can then build two different joint mass functions such that $m_1 \mathbf{R} m_2$ and $m_2 \mathbf{R} m_1$. □

Proposition 4 (Unpreserved asymmetry). *If \mathbf{R} is asymmetric on sets, it is not necessarily so on belief functions, as we may have*

$$m_1 \mathbf{R} m_2 \text{ and } m_2 \mathbf{R} m_1$$

Proof. Simply consider two mass functions m_1, m_2 that are positive only on subsets A, B, C, D, E and such that

$$\begin{aligned} m_1(A) &= 0.2, m_1(B) = 0.3, m_1(C) = 0.2, m_1(D) = 0.1, m_1(E) = 0.2, \\ m_2(A) &= 0.2, m_2(B) = 0.1, m_2(C) = 0.3, m_2(D) = 0.3, m_2(E) = 0.1 \end{aligned}$$

as well as the asymmetric relation \mathbf{R} on those subsets summarised by the matrix

$$\begin{matrix} & A & B & C & D & E \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \left[\begin{array}{ccccc} & & & ARC & ARD \\ BRA & & & & BRE \\ & CRB & & CRD & \\ & DRB & & & DRE \\ ERA & & ERC & & \end{array} \right] \end{matrix}$$

We can then build two different joint mass functions such that $m_1\mathbf{R}m_2$ and $m_2\mathbf{R}m_1$. □

Proposition 5 (Preserved reflexivity). *If \mathbf{R} is reflexive on sets, it is so on belief functions:*

$$\forall m, \text{ we have } m\mathbf{R}m$$

Proof (sketch). Just consider the joint mass function $m_{12}(A, A) = m(A)$ if $m_1 = m_2 = m$.

Proposition 6 (Unpreserved irreflexivity). *If \mathbf{R} is irreflexive on sets, it is not necessarily so on belief functions, as we may have $m\mathbf{R}m$ for some $m \in \mathcal{B}$.*

Proof. Consider the following mass function

$$m(A_1) = 0.5, m(A_2) = 0.5$$

and the relation \mathbf{R} summarised in the following matrix

$$\begin{matrix} & A_1 & A_2 \\ \begin{matrix} A_1 \\ A_2 \end{matrix} & \left[\begin{array}{cc} & A_2 \\ A_1\mathbf{R}A_2 \\ A_2\mathbf{R}A_1 & \end{array} \right] \end{matrix}$$

which is irreflexive. However, the joint $m(A_1, A_2) = m(A_2, A_1) = 0.5$ shows that we have $m\mathbf{R}m$, hence \mathbf{R} may not be irreflexive for belief functions. □

Proposition 7 (Preserved transitivity). *If \mathbf{R} is transitive on sets, it is so on belief functions:*

$$m_1\mathbf{R}m_2 \wedge m_2\mathbf{R}m_3 \implies m_1\mathbf{R}m_3$$

Proof (sketch). Consider two stochastic matrices $S_{\mathbf{R}_{12}}$ and $S_{\mathbf{R}_{23}}$ satisfying Definition 1, we can show that their product gives a matrix $S_{\mathbf{R}_{13}}$ satisfying Definition 1.

Proposition 8 (Unpreserved completeness). *If \mathbf{R} is complete (or total) on sets, it is not necessarily so on belief functions: for any two m_1, m_2 we may have neither $m_1\mathbf{R}m_2$ nor $m_2\mathbf{R}m_1$.*

Proof. Consider the following mass functions

$$m_1(A_1) = 0.6, m_1(A_2) = 0.4; \quad m_2(B_1) = m_2(B_2) = 0.5$$

and the relation \mathbf{R} summarised in the following matrix

$$\begin{matrix} & B_1 & B_2 \\ A_1 & [A_1\mathbf{R}B_1 & B_2\mathbf{R}A_1] \\ A_2 & [B_1\mathbf{R}A_2 & A_2\mathbf{R}B_2] \end{matrix}$$

It is clear that any joint mass function respecting conditions (5)–(6) must give a non-null mass to both (A_1, B_1) and (A_1, B_2) , hence we have neither $m_1\mathbf{R}m_2$ nor $m_2\mathbf{R}m_1$. \square

Table 1 summarises our obtained results. Note that some properties unper- served in general can nevertheless be preserved in peculiar cases (e.g., antisym- metry of inclusion is preserved, as specialisation is antisymmetric). This opens the way to various further questions (i.e., what happens when considering poset structures).

Table 1. Summary of properties preservation

\mathbf{R} on 2^X is	$\rightarrow \mathbf{R}$ on \mathcal{B}^X is	\mathbf{R} on 2^X is	$\rightarrow \mathbf{R}$ on \mathcal{B}^X is
Symmetric	Yes	Irreflexive	No
Antisymmetric	No	Transitive	Yes
Asymmetric	No	Complete	No
Reflexive	Yes		

3 Related Works

3.1 Inclusion and Consistency

In the case where the relations are either inclusion or consistency, then we retrieve well-known results of the literature:

- in the case of inclusion we have $A\mathbf{R}B$ iff $A \subseteq B$, our definition is essentially the same as specialisation [3], since checking $m_2(B) > 0$ is unnecessary in this case.
- in the case of consistency, we have $A\mathbf{R}B$ iff $A \cap B \neq \emptyset$, and one can see that $m_1\mathbf{R}m_2$ iff there is a joint mass affecting positive mass to pairs of sets having a non-empty intersection. This is equivalent to require $\mathcal{P}_1 \cap \mathcal{P}_2 \neq \emptyset$, with \mathcal{P}_i the probability set induced by m_i [1].

3.2 Rankings

When the space X is ordered (with $x_i \leq x_{i+1}$) and possibly infinite, it makes sense to consider relations of the kind “higher than” in order to compare sets.

There are many ways to rank two sets A, B , such as:

- Single-bound dominance, that can be declined itself into four notions:
 - loose dominance: $\mathbf{AR}_{\leq LD} B$ if $\min A \leq \max B$
 - lower bound: $\mathbf{AR}_{\leq LB} B$ if $\min A \leq \min B$
 - upper bound: $\mathbf{AR}_{\leq UP} B$ if $\max A \leq \max B$
 - strict dominance: $\mathbf{AR}_{\leq SD} B$ if $\max A \leq \min B$
- Pairwise-bound or lattice dominance: $\mathbf{AR}_{\leq PD} B$ if $\min A \leq \min B$ and $\max A \leq \max B$, whose extension to belief functions studied in [5] correspond to our proposal.

Extensions of this kind are connected to the extensions of stochastic dominance explored in [2].

4 Preservation by Functional Mapping

In this section, we investigate how we can check whether a relation is preserved by a functional mapping, in the univariate case (multivariate case easily follows). Such mappings are indeed used in lots of applications involving uncertainty propagation (e.g., multi-criteria decision making, risk analysis, ...).

Let f be some function with domain X and codomain Y , i.e., $f : X \rightarrow Y$. The image $f(A)$ of $A \subseteq X$ under f is the subset $f(A) = \{f(x) : x \in A\} \subseteq Y$. More generally, the image $f(m)$ of some mass function $m \in \mathcal{B}^X$ under f is the mass function $f(m) \in \mathcal{B}^Y$ defined as

$$f(m)(B) = \sum_{f(A)=B} m(A) \text{ for all } B \subseteq Y. \tag{7}$$

Definition 1. Let $f : X \rightarrow Y$. Let \mathbf{R}^X and \mathbf{R}^Y be relations on 2^X and 2^Y , respectively. The pair $(\mathbf{R}^X, \mathbf{R}^Y)$ of relations \mathbf{R}^X and \mathbf{R}^Y is said to be compatible with respect to f (f -compatible for short) if, for all $A, B \subseteq X$,

$$\mathbf{AR}^X B \Rightarrow f(A)\mathbf{R}^Y f(B).$$

Example 1. Let \mathbf{R}_{\subseteq}^X be the relation corresponding to inclusion on X , i.e., $\mathbf{AR}_{\subseteq}^X B$ iff $A \subseteq B$, $A, B \subseteq X$. Similarly, let \mathbf{R}_{\subseteq}^Y denote inclusion on Y . Since for any function f and any $A, B \subseteq X$ such that $A \subseteq B$ it holds that $f(A) \subseteq f(B)$, the pair $(\mathbf{R}_{\subseteq}^X, \mathbf{R}_{\subseteq}^Y)$ is f -compatible for any f . Similarly, the pair $(\mathbf{R}_{\subseteq}^X, \mathbf{R}_{\subseteq}^Y)$ is f -compatible for any f .

Now, let X and Y be two ordered spaces and let $\mathbf{R}_{\leq PD}^X$ and $\mathbf{R}_{\leq PD}^Y$ be the relations corresponding to pairwise-bound dominance on X and on Y , respectively. Then, the f -compatibility of pair $(\mathbf{R}_{\leq PD}^X, \mathbf{R}_{\leq PD}^Y)$ depends on f . In particular, if f is monotonically non-decreasing, we have $f(A) \leq_{PD} f(B)$ for all $A, B \subseteq X$ such that $A \leq_{PD} B$, and thus the pair $(\mathbf{R}_{\leq PD}^X, \mathbf{R}_{\leq PD}^Y)$ is f -compatible for such f . However, this pair is not f -compatible when f is monotonically non-increasing since in general we have in this case $A \leq_{PD} B \not\Rightarrow f(A) \leq_{PD} f(B)$.

Proposition 9 (Preserved compatibility). *If $(\mathbf{R}^X, \mathbf{R}^Y)$ is f -compatible, it is so on belief functions:*

$$m_1 \mathbf{R}^X m_2 \Rightarrow f(m_1) \mathbf{R}^Y f(m_2). \quad (8)$$

Proof (sketch). Consider a joint mass m_{12} on X satisfying (4)–(6) for \mathbf{R}^X . Then if $\mathbf{R}^X, \mathbf{R}^Y$ is f -compatible, mapping m_{12} through f results in a joint mass showing that $f(m_1) \mathbf{R}^Y f(m_2)$.

We note that Proposition 9 and its straightforward extension to multivariate functions (not presented here due to lack of space) generalize results in [4, 5] concerning inclusion and ranking.

5 Conclusion

In this paper, a very general way to extend a binary relation on sets to a binary relation on belief functions is introduced. Several results are provided to assess which properties of the relation are preserved through this mechanism. Our proposal is also connected to more specific generalisation of binary relations, such as the notion of specialisation. Consequently, our results are also a generalisation of pre-existing ones for specific relations.





We believe that the original ideas presented in this paper shall reach out a large audience of belief function practitioners wishing to address multi-criteria decision making, reliability analysis or optimisation problems, in which some relations such as ranking or information loss relations can play a significant role.

References

1. Chateauneuf, A.: Combination of compatible belief functions and relation of specificity. In: Fedrizzi, M., Kacprzyk, J., Yager, R.R. (eds.) *Advances in the Dempster-Shafer Theory of Evidence*, pp. 97–114. Wiley, New York (1994)
2. Denœux, T.: Extending stochastic ordering to belief functions on the real line. *Inf. Sci.* **179**(9), 1362–1376 (2009)
3. Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *Int. J. Gen. Syst.* **12**(3), 193–226 (1986)
4. Dubois, D., Prade, H.: Random sets and fuzzy interval analysis. *Fuzzy Sets Syst.* **42**, 87–101 (1991)
5. Helal, N., Pichon, F., Porumbel, D., Mercier, D., Lefèvre, E.: The capacitated vehicle routing problem with evidential demands. *Int. J. Approx. Reason.* **95**, 124–151 (2018)
6. Yager, R.R.: The entailment principle for Dempster-Shafer granules. *Int. J. Intell. Syst.* **1**(4), 247–262 (1986)



Application of Belief Functions to Levee Assessment

Théo Dezert^{1,2} , Yannick Fargier^{2,3} , Sérgio Palma Lopes¹ ,
and Philippe Côte¹ 

¹ IFSTTAR, GERS, GeoEND, 44344 Bouguenais, France
{theo.dezert,sergio.lopes,philippe.cote}@ifsttar.fr

² Cerema Direction territoriale Normandie-Centre, 41000 Blois, France

³ Univ Lyon, IFSTTAR, GERS, RRO, 69675 Bron, France
yannick.fargier@ifsttar.fr

Abstract. We propose the use of Smets and PCR5 rules to merge artificial geophysical and geotechnical data, as part of fluvial levee assessment. It highlights the ability to characterize the presence of interfaces and a geological anomaly.

Keywords: Levee assessment · Geophysics · Geotechnical testing
Belief functions · Data fusion

1 Introduction

Fluvial levees are manmade structures built for flood protection. They are considered as hazardous structures that can fail and lead to disastrous consequences such as human or material loss and economic disasters. There are globally acknowledged methodologies for levee assessment that include geophysical and geotechnical investigation methods [1]. Geophysical methods are non-intrusive and provide physical information on large volumes of subsoil with high output and potentially significant related uncertainties. These associated uncertainties are notably due to the indirect and integrating aspects of the methods and to the resolution of inverse problems. Geotechnical investigation methods are intrusive and provide more punctual and more accurate information. An important issue of assessment of levees is to be able to combine geophysical and geotechnical data taking into consideration their respective associated uncertainties, imprecisions and spatial distributions. In this work, we suggest the use of Belief Functions (BFs) and combination rules to merge artificial geophysical (electrical resistivities) and geotechnical (cone bearing) data to display their ability to discriminate three sets of soils. We assume that the reader is familiar with the BFs introduced by Shafer in [2]. The use of BFs requires: (1) to select a common frame of discernment (FoD) of the considered problem, (2) to determine the masses of belief or Basic Belief Assignments (BBAs) from available data (geophysical and geotechnical) and (3) to choose a rule of combination.

Supported by the Pays de la Loire Region.

2 FoD and BBAs Construction

For the addressed levee problematic, we consider three classes of distinct soils θ_1 , θ_2 and θ_3 . Because the FoD, Θ , must consist of a set of exhaustive and exclusive hypotheses, we will be using a fourth class θ_4 to cover the physical characteristics not included in the three first sets. The FoD is common to both information sources. We use $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. The construction of the BBAs for each data source consists in assigning each data type to Θ .

BBA Construction From Geophysical Data: since the electrical resistivity (ER) tomography method is one of the most employed, we propose the use of ER as geophysical data. We consider two soil layers: an upper conductive layer ($10 \Omega.m$) standing for clays [3] and a subjacent and more resistive one ($10^2 \Omega.m$) standing for silts starting at 10.4 m depth. A very resistive anomaly ($10^3 \Omega.m$) standing for a sandy lens of about 10.5 m high and 21.25 m wide, is finally positioned between these two first media. We then associate ER classes to specific soils (split into ranges of ER) to Θ : $\theta_1 = [5, 20]$, $\theta_2 = [50, 2 \cdot 10^2]$, $\theta_3 = [5 \cdot 10^2, 2 \cdot 10^3]$ and $\theta_4 = [1, 5 \cup]20, 50 \cup]2 \cdot 10^2, 5 \cdot 10^2 \cup]2 \cdot 10^3, 10^4]$. We use Res2Dmod free software [4] to simulate noised data acquisition from a chosen resistivity model (Fig. 1a) and then use the Res2Dinv software [5] to obtain the inverted ER section as one would get from the processing of survey data (Fig. 1b). The distinction between clays and silts can easily be made while the discrimination of the anomaly is not obvious. We finally use the Res2dinv discretization grid for the BBA $m_1(\cdot)$ corresponding to each event of 2^Θ . The values of the masses are set using the Wasserstein distances between an inverted ER value \pm its uncertainty issued from Res2dinv and the interval corresponding to each event, so as each cell of the grid gets a normalized BBA.

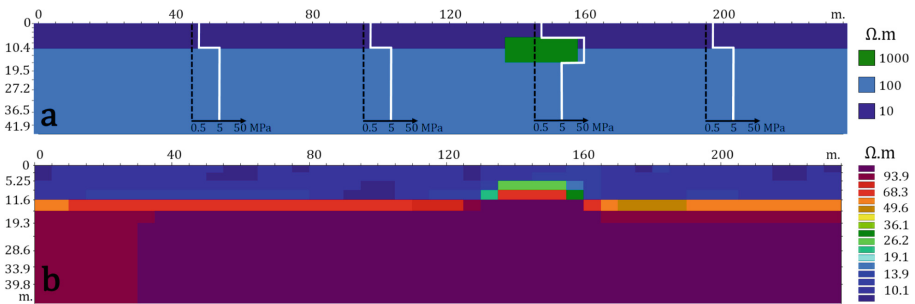


Fig. 1. 2D section of subsoil displaying (a) true ER with boreholes position in dashed line and associated cone bearing values in white and (b) inverted ER.

BBA Construction From Geotechnical Data: as geotechnical data, we use artificial cone bearing values (expressed in MPa). These information could be obtained from a cone penetrometer test investigation campaign. We simulate a

data acquisition from 4 boreholes with an interspacing of 50 m (as recommended in [6]), drilled to 40 m depth with an acquisition every meter (Fig. 1a). One of the boreholes is positioned so that it goes through the resistive anomaly. We consider the following assignment of intervals of cone bearing values to Θ : $\theta_1 = [0.3, 0.7]$, $\theta_2 = [3, 7]$, $\theta_3 = [30, 70]$ and $\theta_4 = [10^{-2}, 0.3[\cup]0.7, 3[\cup]7, 30[\cup]70, 10^2]$ that can be associated to specific soils [7], such as clays for low values, silty soils for intermediate values and sands for higher ones. We assume a belief mass equal to 1 in the borehole and impose a lateral decrease of the trust in the data. The geotechnical grid depends on the boreholes distance and on the acquisition rate. Thus, for each cell, a second BBA $m_2(\cdot)$ is fixed, entering in the fusion process.

3 BBAs Combination and Preliminary Results

We propose a fusion mesh containing all the meshes from the geotechnical and geophysical grids in order to avoid the unnecessary data alteration due to interpolations. The merging process is carried out on two meshes of same dimension. The data fusion consists in combining $m_1(\cdot)$ and $m_2(\cdot)$ assigned to each cell of the grid. Many rules of BBA combination have been proposed. Here we present only two of them: Smets' rule [8] and the Proportional Conflict Redistribution rule no. 5 (PCR5) [9] allowing the redistribution of all partial conflicts proportionately to the masses involved in them. We use PCR5 since we combine only two sources of evidence thus PCR6 is equivalent to PCR5 rule [9] in this case. Smets' rule (conjunctive rule under an open-world assumption) allows the quantification of the classical conflict level represented by (Eq. 1):

$$m_{12}(\emptyset) = \sum_{X_1, X_2 \subseteq \Theta | X_1 \cap X_2 = \emptyset} m_1(X_1)m_2(X_2) \quad (1)$$

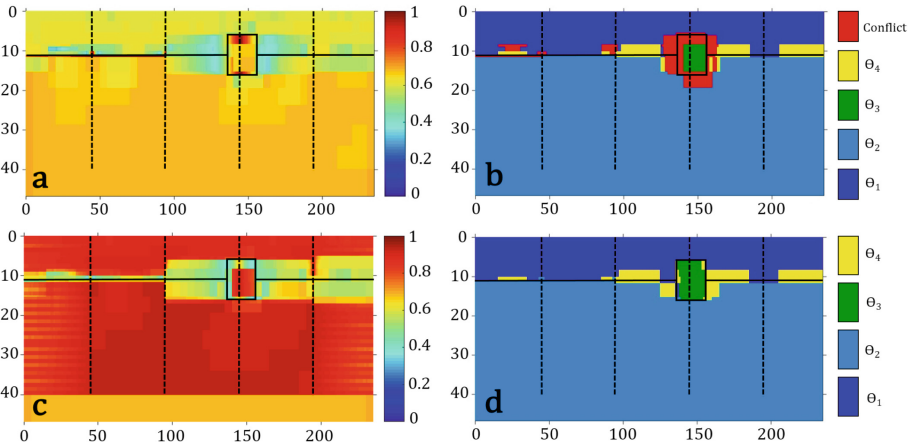


Fig. 2. Data fusion with Smets' combination rule (a, b) and with PCR5 (c, d). (a) and (c) represent the BBAs associated to the events with the highest mass, presented in (b) and (d) respectively. The black lines stand for the interfaces fixed in the ER model (Fig. 1a) while the dashed lines stand for the boreholes position.

Thanks to it, we are able to point out the conflictual zones around the horizontal interfaces and the resistive anomaly (Fig. 2b). The fusion, following PCR5 (closed world assumption)[9] (Fig. 2d) is very close to the true model we imposed (Fig. 1a), giving a clearer view of the interface and of the vertical and horizontal extension of the resistive anomaly compared to the image given by the inverted ER (Fig. 1b). As a decision-making support, we choose to represent the events having the highest belief masses (Fig. 2b and d) and their related degrees of belief (Fig. 2a and c).

4 Conclusion

The use of BFs for investigation of levees is promising. It is able to highlight the presence of an interface between two media much more precisely than the geophysical method alone. Furthermore, it enables the reliable estimation of the complete extension of an anomaly with high ER and cone bearing values. Without normalization, Smets' combination rule easily spotlights the conflicting zones. Such information could be precious during an investigation campaign, indicating areas where survey has to be reinforced. In future work, we will focus on parametric studies to choose the best decreasing functions for the lateral propagation of the geotechnical information. Finally, we will test our algorithm using real data acquired on a scale model and on a levee.

References

1. Fauchard, C., Mériaux, P.: Geophysical and geotechnical methods for diagnosing flood protection dikes: guide for implementation and interpretation. Quae (2007)
2. Shafer, G.: A Mathematical Theory of Evidence, Princeton University Press, Princeton (1976)
3. Palacky, G., West, G.F.: Electromagnetic methods in applied geophysics. Resistivity Characteristics of Geologic Targets, pp. 52–129 (1987)
4. Loke, M.H.: RES2DMOD ver. 3.01: Rapid 2D resistivity forward modelling using the finite difference and finite-element methods. Software manual (2002)
5. Loke, M.H., Barker, R.D.: Rapid least-squares inversion of apparent resistivity pseudosections by a quasi-Newton method. *Geophys. Prospect.* **44**(1), 131–152 (1996)
6. Phoon, K.K., Kulhawy, F.H.: Characterization of geotechnical variability. *Can. Geotech. J.* **36**(4), 612–624 (1999)
7. Robertson, P.K.: Soil classification using the cone penetration test. *Can. Geotech. J.* **27**(1), 151–158 (1990)
8. Smets, P.: The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(5), 447–458 (1990)
9. Smarandache, F., Dezert, J.: Advances and applications of DSMT for information fusion-*Collected works - ARP*, volume 3 (2009)



Prejudiced Information Fusion Using Belief Functions

Didier Dubois^(✉), Francis Faux, and Henri Prade

Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse,
CNRS, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France
{dubois,faux,prade}@irit.fr

Abstract. G. Shafer views belief functions as the result of the fusion of elementary partially reliable testimonies from different sources. But any belief function cannot be seen as the combination of simple support functions representing such testimonies. Indeed the result of such a combination only yields a special kind of belief functions called separable. In 1995, Ph. Smets has indicated that any belief function can be seen as the combination of so-called *generalized* simple support functions. We propose a new interpretation of this result in terms of a pair of separable belief functions, one of them modelling testimonies while the other represents the idea of prejudice. The role of the latter is to weaken the weights of the focal sets of the former separable belief function. This bipolar view accounts for a form of resistance to accept the information supplied by the sources, which differs from the discounting of sources.

1 Introduction

G. Shafer [1] has presented his theory of belief functions essentially as an approach to the fusion of unreliable elementary testimonies, each being represented by simple support functions. However many belief functions prove to be not separable, i.e., not the orthogonal sum of simple support functions. Ph. Smets [2] tries to remedy this difficulty by generalizing simple support functions, showing that any belief function is the conjunctive combination of such generalized elementary belief functions (where some masses can be negative). Using a retraction operation, he shows that any belief function can be decomposed into two separable belief functions. One represents the fusion of elementary testimonies (expressing confidence), and the other (expressing doubt) plays the role of a moderator that can annihilate, via retraction, some information supplied by the former, possibly resulting in ignorance. This pair of belief functions is called “Latent Belief Structure” by Smets.

In this paper, we present a bipolar belief function model which pushes the notion of “Latent Belief Structure” further. In a belief function, the doubt component is assumed to reflect a cognitive bias interpreted as a prejudice, pertaining to the information supplied by the confidence component. This cognitive bias leads to weaken the strength attached to the combination of some elementary

testimonies appearing in the confidence part, thus expressing a lack of trust in the information obtained by merging these testimonies.

The organization of the rest paper is as follows. In Sect. 2, some necessary background on belief functions is introduced. In Sect. 3, we propose new results about the decomposition of belief functions, providing new insights in the weight function introduced by Smets [2], as well as conditions for separability in a simple case. Section 4 presents a generalized setting for the merging of elementary testimonies in the presence of prejudices, focusing on the process of belief attenuation by means of the retraction operation. This framework is illustrated on the Linda example [3], highlighting the difference between belief retraction and source discounting.

2 Separable Belief Functions

In Shafer evidence theory, the uncertainty concerning an agent's state of belief on a finite set of possible situations, called the frame of discernment Ω is represented by a basic belief assignment (BBA) or mass function m defined as a mapping $m : 2^\Omega$ to $[0, 1]$ verifying $\sum_{A \subseteq \Omega} m(A) = 1$. Each subset $A \subseteq \Omega$ such as $m(A) > 0$ is called a *focal set* of m . A BBA m is called *normal* if \emptyset is not a focal set (subnormal otherwise), *vacuous* if Ω is the only focal element, *non-dogmatic* if Ω is a focal set, *categorical* if m has only one focal set different from Ω .

An elementary testimony T with strength $1-x$ in favor of a non-contradictory proposition $A \in 2^\Omega$ is represented by a simple BBA (SBBA) $m : 2^\Omega \rightarrow [0, 1]$ such that $m(A) = 1-x$, for $A \neq \Omega$ and $m(\Omega) = x$, with $x \in [0, 1]$ and is denoted by $m = A^x$. The value x , we call *diffidence weight*, evaluates the lack of reliability of the testimony (or the source of information). A vacuous BBA can thus be denoted by A^1 for any $A \subset \Omega$, and a categorical BBA $A \neq \Omega$ can be denoted by A^0 .

A belief function $Bel(A)$ is a non-additive set function which represents the total quantity of belief in the subset A of Ω and is defined by $Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$. A BBA m can be equivalently represented by its associated plausibility and commonality functions respectively defined for all $A \subseteq \Omega$ by $Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) = 1 - Bel(\bar{A})$ and $Q(A) = \sum_{B \supseteq A} m(B)$.

The conjunctive combination of BBA's m_j derived from k distinct sources, denoted by m_{\odot} is expressed by $m_{\odot}(A) = \sum_{A_1 \cap \dots \cap A_k = A} \left(\prod_{j=1}^k m_j(A_j) \right)$. Note that m_{\odot} is not always normal. Dempster's rule, denoted by \oplus , is a normalized version of the conjunctive combination rule and is defined such that: $m_{\oplus}(\emptyset) = 0$ and $m_{\oplus}(A) = K \cdot m_{\odot}$ for $A \neq \emptyset$. The normalization factor K is of the form $(1 - c(m_1, \dots, m_k))^{-1}$ where $c(m_1, \dots, m_k) = \sum_{A_1 \cap \dots \cap A_k = \emptyset} \left(\prod_{j=1}^k m_j(A_j) \right) < 1$ represents the amount of conflict between the sources. These two combination rules are commutative, associative, and generally used to combine BBAs from distinct sources. The Dempster rule is simply expressed using the commonality functions as: $Q_1 \oplus \dots \oplus Q_k = K \cdot Q_1 \cdot Q_2 \cdots Q_k$.

In Shafer’s view [1], a separable BBA is the result of Dempster’s rule of combination of simple BBAs: $m = \bigoplus_{\emptyset \neq A \subset \Omega} A^{w(A)}$, $w(A) \in [0, 1]$, $\forall A \subset \Omega, A \neq \emptyset$. We call the mapping $w : 2^\Omega \setminus \{\Omega\} \rightarrow (0, 1]$ a *diffidence function*. If the BBA is non-dogmatic ($m(\Omega) > 0$), this representation is unique, and $w(A) > 0, \forall A \subset \Omega$. Denœux [4] has extended this concept to the conjunctive combination of subnormal BBA’s $\bigodot_{\emptyset \neq A \subset \Omega} A^{w(A)}$, $w(A) \in [0, 1] \forall A \subset \Omega$.

Shafer [1] [Th. 7.2 p.143] shows that if *Bel* is a separable belief function, and *A* and *B* are two of its focal sets such as $A \cap B \neq \emptyset$, then $A \cap B$ is a focal set of *Bel*. The condition $A \cap B \neq \emptyset$ can be dropped if we allow for sub-normalized belief functions. But the converse is not true. This necessary condition clearly indicates that not all belief functions are separable. To overcome this difficulty, Smets [2] generalized the concept of simple support function, considering A^x such that $x \in (0, +\infty)$. Smets has shown that any non dogmatic BBA can be decomposed into the conjunctive combination of generalized BBA’s: $m = \bigodot_{\emptyset \neq A \subset \Omega} A^{w(A)}$, extending the range of diffidence functions w to $(0, +\infty)$. For every $A \subset \Omega$, the weights $w(A)$ are obtained from the commonality function of m as: $w(A) = \prod_{B \supseteq A} Q(B)^{(-1)^{|B|-|A|+1}} = \frac{\prod_{C \cap A = \emptyset, |C| \text{ odd}} Q(A \cup C)}{\prod_{C \cap A = \emptyset, |C| \text{ even}} Q(A \cup C)}$.

3 The Bipolar Decomposition of a Belief Function

We can write the decomposition of a non-dogmatic belief function as $m = (\bigodot_{A \in \mathcal{C}} A^{w^+(A)}) \otimes (\bigodot_{B \in \mathcal{D}} B^{w^-(B)})$, where

- w^+ and w^- are standard diffidence functions in $(0, 1)$ defined from the original one w associated to m , such that: $w^+(A) = \min(1, w(A))$, and $w^-(A) = \min(1, 1/w(A)), \forall A \subset \Omega$.
- \mathcal{C} and $\mathcal{D} \subseteq 2^\Omega, w(A) < 1$ if $A \in \mathcal{C}$ and $w(B) > 1$ if $B \in \mathcal{D}$.
- \otimes defined by $m_1 \otimes m_2 = (\bigodot_{\emptyset \neq A \subset \Omega} A^{w_1(A)}) \bigodot (\bigodot_{\emptyset \neq B \subset \Omega} B^{w_2(B)})$ is the *retraction operation*, also obtained by the division of commonality functions: $Q_1 \otimes Q_2(X) = \frac{Q_1(X)}{Q_2(X)}, \forall X \subseteq \Omega$, called *decombination* [2] or *removal* [5]. Ginsberg [6] and Kramosil [7] have exploited this division rule.
- Factors of the form $A^{w(A)}$ represent testimonies in favor of *A* if $w(A) < 1$, and will be called *prejudices* against believing *A* if $w(A) > 1$.

A belief function is separable if and only if $w(A) \leq 1, \forall A \subset \Omega$ in the above decomposition. In that case, the set of focal sets of m contains Ω and is closed under conjunction [1]. So a separable belief function will be of the unique form: $m = \bigodot_{A \in \mathcal{C}} A^{w^+(A)}$.

A mass function m can thus be decomposed in a unique irredundant way as a pair (m^+, m^-) , of separable belief functions induced by BBAs m^+ and m^- , such that $m = m^+ \otimes m^-$. The confidence component denoted by m^+ is a BBA obtained from the merging of SBBAs, with focal sets in \mathcal{C} , and the diffidence component denoted by m^- is a BBA obtained likewise, with focal sets in \mathcal{D} . By construction, $\mathcal{C} \cap \mathcal{D} = \emptyset$. The pair (m^+, m^-) of separable BBAs is called a latent belief structure [2] more recently studied in [4,8,9]. The existence of positive

and negative information is generally coined under the term *bipolarity* [10], an idea applied to latent belief structures in [11]. A general study of the canonical conjunctive decomposition of a belief function was realised by Ke et al. [12] and Pichon [13], albeit without focusing on its possible meaning.

In the following we are interested in retrieving the mass function m from its diffidence function w via the commonality function rather by the conjunctive combination. First, note that the expression $\prod_{B \supseteq A} Q(B)^{(-1)^{|B|-|A|+1}}$ makes sense for $A = \Omega$, and we get $w(\Omega) = 1/Q(\Omega)$. So function w can be extended to the whole of 2^S , even if only sets $A \subset \Omega$ appear in the decomposition formula. In previous studies, $w(\Omega)$ remained undefined. Of course, $w(\Omega) > 1$ but this will be also the case for the diffidence weights of other subsets for non-separable belief functions.

Noticing that $m(A) = \sum_{A \subseteq B} (-1)^{|B|-|A|} Q(B)$, and moreover $\log w(A) = \sum_{A \subseteq B} (-1)^{|B|-|A|+1} \log Q(B)$, it is clear that m is to Q what $-\log w$ is to $\log Q$. Since $Q(A) = \sum_{A \subseteq B} m(B)$, we have $\log Q(A) = \sum_{A \subseteq B} \log(1/w(B)) = \log \prod_{A \subseteq B} \frac{1}{w(B)}$. Hence,

$$Q(A) = \frac{1}{\prod_{A \subseteq B} w(B)} \quad (1)$$

Note that in (1), the weight $w(\Omega)$ appears explicitly in all the expressions of $Q(A)$ for all subsets A . Hence we can retrieve the BBA m , from the diffidence function w computed from it, directly as $m(E) = \sum_{E \subseteq A} (-1)^{|B|-|A|} (\frac{1}{\prod_{A \subseteq B} w(B)})$. In particular we can have the following result:

Proposition 1. *A diffidence function computed from m via (1) is such that $\prod_{A \subseteq \Omega} w(A) = 1$.*

Proof. We know that commonalities satisfy $Q(\emptyset) = 1$. Using (1) yields $Q(\emptyset) = \frac{1}{\prod_{\emptyset \subseteq B} w(B)} = 1$. So, $\prod_{A \subseteq \Omega} w(A) = 1$.

It gives a general definition of a diffidence function as a mapping $w : 2^\Omega \rightarrow (0, +\infty)$, such that $\prod_{A \subseteq \Omega} w(A) = 1$ and $w(\Omega) \geq 1$. Note that the mass function m_w derived from any function w defined in this way is not always positive. Indeed, suppose that $w(A) = \lambda < 1$, $w(B) = \mu > 1$, $w(C) = 1$, $C \neq \Omega$ otherwise (so $w(\Omega) = 1/\lambda\mu > 1$). By means of the conjunctive rule, one gets the BBA: $m(A \cap B) = (1 - \lambda)(1 - \mu)$, $m(A) = \lambda(1 - \mu)$, $m(B) = (1 - \lambda)\mu$, $m(\Omega) = \lambda\mu$. It is clear that $m(A \cap B)$, $m(A)$ are negative, in general. So the mapping $m \mapsto w$ is injective, but it is not surjective. Namely, given a diffidence function w such that $\prod_{A \subseteq \Omega} w(A) = 1$ and $w(\Omega) \geq 1$, Q_w obtained by (1) is a decreasing set-function that ranges on $[0, 1]$, but decreasingness is not sufficient to ensure that masses obtained from function Q_w are all positive, i.e., Q_w is not always a commonality function. On the other hand, diffidence functions such that $w(A) \leq 1, \forall A \subset \Omega$ are in one to one correspondence with BBAs of separable belief functions.

Example: Two Overlapping Focal Sets on a 4-Element Frame. Let $\Omega = \{a, b, c, d\}$. We denote $\{a\}$ by a , $\{a, b\}$ by ab , etc. Consider m with $m(ab) = \beta; m(ac) = \gamma; m(a) = \alpha$ with $\alpha + \beta + \gamma < 1$, (hence $m(\Omega) = 1 - (\alpha + \beta + \gamma)$). Note that $Q(a) = 1, Q(ab) = 1 - \alpha - \gamma, Q(ac) = 1 - \alpha - \beta, Q(B) = 1 - \alpha - \beta - \gamma$ for other non-empty sets B .

We can decompose m as a combination $m = \{ab\}^{w(ab)} \odot \{ac\}^{w(ac)} \odot \{a\}^{w(a)}$. Its diffidence function is given in Table 1.

Table 1. Decomposition with focal sets: ab, ac, a and Ω

A	m	w	Inverse solution
a	α	$w(a) = \frac{(1-\alpha-\gamma)(1-\alpha-\beta)}{1-\alpha-\beta-\gamma}$	$1 - (w(ab) + w(ac) - w(ab)w(ac))w(a)$
ab	β	$w(ab) = \frac{1-\alpha-\beta-\gamma}{1-\alpha-\gamma}$	$(1 - w(ab))w(ac)w(a)$
ac	γ	$w(ac) = \frac{1-\alpha-\beta-\gamma}{1-\alpha-\beta}$	$(1 - w(ac))w(ab)w(a)$
$abcd$	$1 - \alpha - \beta - \gamma$	$w(\Omega) = \frac{1}{1-\alpha-\beta-\gamma}$	$w(ac)w(ab)w(a)$
<i>other subsets</i>	0	1	0

It is a separable belief function if the diffidence weights are ≤ 1 . Note that $w(ab) = \frac{1-\alpha-\beta-\gamma}{1-\alpha-\gamma} < 1$ and $w(bc) = \frac{1-\alpha-\beta-\gamma}{1-\alpha-\beta} < 1$ but it is not always the case for $w(a) = \frac{(1-\alpha-\gamma)(1-\alpha-\beta)}{1-\alpha-\beta-\gamma}$. The condition of separability of the belief function m is $\alpha^2 + \alpha(-1 + \beta + \gamma) + \beta\gamma \leq 0$. Fixing β, γ , this condition is of the form $\alpha_1 \leq \alpha \leq \alpha_2$, with $\alpha_1 = \frac{(1-\beta-\gamma) - \sqrt{(-1+\beta+\gamma)^2 - 4\beta\gamma}}{2}$ and $\alpha_2 = \frac{(1-\beta-\gamma) + \sqrt{(-1+\beta+\gamma)^2 - 4\beta\gamma}}{2}$, provided that $(1 - \beta - \gamma)^2 \geq 4\beta\gamma$. The latter condition is valid only if β and γ are small enough, that is if $\sqrt{\beta} + \sqrt{\gamma} \leq 1$. Besides note that $0 \leq \alpha_1 \leq \alpha_2 \leq 1 - \beta - \gamma$.

It is of interest to consider the special case when $\beta = \gamma$. It is easy to verify that $\alpha_1 = \frac{1-2\beta-\sqrt{1-4\beta}}{2}$ and $\alpha_2 = \frac{1-2\beta+\sqrt{1-4\beta}}{2}$. We must have $\beta \leq 0.25$ otherwise the belief function cannot be decomposable (α_1 and α_2 are not defined). For $\beta = 0.25$ we have that $\alpha_1 = \alpha_2 = 0.25$. See the graph of the functions giving α_1 and α_2 in terms of β on Fig. 1. It indicates the zone of non-separability under the line $1 - 2\beta$ and on the right-hand side of the curve for α_1 and α_2 .

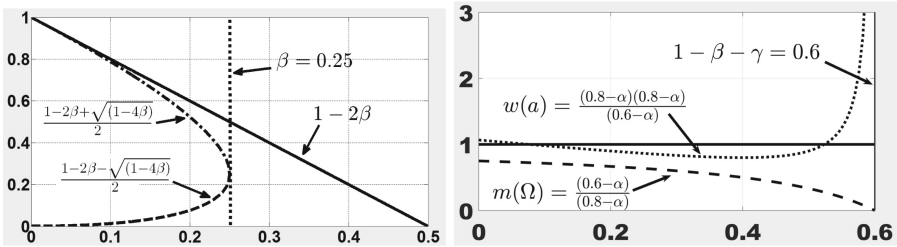


Fig. 1. Left: $m(a)$ in terms of $\beta = \gamma$ if $w(a) = 1$. Right: diffidence weights in terms of α

It may sound strange that there are two separability thresholds α_1 and α_2 . Actually, it means that, fixing β and γ , there are still two possibilities for choosing $m(a)$ such that m is the conjunctive combination of two SBBA's m_b and m_c respectively focused on ab and ac . Let $\lambda = m_b(ab)$ and $\mu = m_c(ac)$. By definition, we have $\beta = \lambda(1 - \mu)$ and $\gamma = (1 - \lambda)\mu$. Suppose without loss of generality that $\lambda\mu \leq (1 - \lambda)(1 - \mu)$. There are two possible choices for m_b and m_c :

- $m_b^1(ab) = \lambda$ and $m_c^1(ac) = \mu$. Then $\alpha_1 = \lambda\mu$, where a is weakly supported.
- $m_b^2(ab) = 1 - \mu$ and $m_c^2(ac) = 1 - \lambda$. Then $\alpha_2 = (1 - \lambda)(1 - \mu)$, where a is strongly supported ($\lambda\mu$ is small). Note that $m_b^1(ab) = 1 - m_c^2(ac)$.

When m defined by parameters α, β, γ is separable, we get $w(a) = 1$, which leads to the condition $(\alpha + \beta)(\alpha + \gamma) = \alpha$. Hence $w(ab) = 1 - \alpha - \beta$ and $w(ac) = 1 - \alpha - \gamma$. So we can define $m_b(ab) = \alpha + \beta$, $m_b(\Omega) = 1 - \alpha - \beta$; and $m_c(ac) = \alpha + \gamma$, $m_c(\Omega) = 1 - \alpha - \gamma$. Choosing $\alpha = \alpha_1$ or α_2 leads to respective pairs of SBBA's (m_b^1, m_c^1) and (m_b^2, m_c^2) . We can check that indeed these pairs are related by the condition $m_b^1(ab) = 1 - m_c^2(ac)$, that is, $\alpha_1 + \beta + \alpha_2 + \gamma = 1$.

Finally, we can study the variation of the diffidence weights when α ranges from 0 to its maximum $1 - \beta - \gamma$. Note that $w(a)$ is the mass of Ω for the SBBA m_a focusing on a , when considering the decomposition of the BBA m . The less $w(a)$ the stronger is the testimony pointing to a , the testimony is not present if $w(a) = 1$, and it becomes a prejudice against a when $w(a) > 1$. It can be checked (see Fig. 1 right) that:

- For $\alpha = 0$, we get $w(a) = \frac{(1-\beta)(1-\gamma)}{1-\beta-\gamma} > 1$.
- $w(a)$ decreases with α until a value $\underline{\alpha} = 1 - \beta - \gamma - \sqrt{\beta\gamma}$ where the derivative vanishes. The minimal value of $w(a)$ is $\frac{(\beta+\sqrt{\beta\gamma})(\gamma+\sqrt{\beta\gamma})}{\sqrt{\beta\gamma}}$ and it is less than 1 only if $\sqrt{\beta} + \sqrt{\gamma} \leq 1$, as seen earlier. When α_1 and α_2 exist, $\alpha_1 \leq \underline{\alpha} \leq \alpha_2$, and they coincide if and only if $\sqrt{\beta} + \sqrt{\gamma} = 1$.
- $w(a)$ increases with $\alpha \geq \underline{\alpha}$ and $\lim_{\alpha \rightarrow 1-\beta-\gamma} w(a) = +\infty$.

Looking at the right part of Fig. 1, we note that when $\alpha = 0$, $w(a) > 1$ and testimonies in favor of ab and ac are weak; so the prejudice against a is strong enough to erase the focal set a from m . When $w(a)$ reaches its minimal value, the prejudice in favor of a is maximal. When α is close to its maximum value $1 - \beta - \gamma$, testimonies in favor of ab and ac are less and less challenged since their diffidence weights get close to 0, while the prejudice against a rapidly increases to infinity. At the limit, we get a dogmatic belief function with $m(a) = 1 - \beta - \gamma$ and the prejudice no longer compensates the elementary testimonies in favor of ab and ac .

4 Prejudiced Information Fusion

A generalized SBBA focused on a subset E with diffidence weight x represents the idea that “one has some reason to believe that the actual world is in E (and nothing more)” when x is small ($x < 1$), whereas, when $x > 1$, it expresses

the idea that “one has some reason not to believe that the actual world is in E ” [2], what we called *prejudice*. Note that the latter does not mean that we have a reason to believe the complement \bar{E} of E (which would mean assigning a weight $x < 1$ to \bar{E}). In this section, we try to provide an interpretation of non-separable belief functions in terms of merging elementary testimonies with prejudices that weaken the result of the former merging. The idea is that the agent possessing a prejudice of strength $y > 1$ against believing E is ready to doubt about the truth of E whenever receiving a testimony claiming that E is true. More generally, the combination $A^x \odot B^y$ of a simple BBA A^x , $x < 1$ with a simple prejudice B^y , $y > 1$ yields the diffidence function $w(\cdot)$ such that:

$$\text{if } B \neq A, w(E) = \begin{cases} x & \text{if } E = A \\ y & \text{if } E = B \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

and $w(A) = xy$ if $A = B$. So, it is a belief function if and only if $A = B$ and $xy < 1$. It is equivalent to *erode* the testimony A^x with another testimony $A^{1/y}$ using retraction. In particular, $A^x \odot A^{1/x}$ yields total ignorance. However, erosion cannot alter A^x by retracting $B \neq A$.

We can compare the erosion with discounting an SSB A^x : the discounting procedure reduces the mass $1 - x$ bearing on A with a factor $\delta \in [0, 1]$ and yields $m_\delta(A) = A^{(1-\delta)+\delta x}$, which is equal to A^{xy} provided that $0 < \delta = \frac{1-xy}{1-x} \leq 1$ since $y > 1$, that is $1 < y < 1/x$.

More generally we can retract a focal set B from a separable mass function m . Consider $m = \odot_{i=1}^k A_i^{w_i}$ and its combination with a prejudice B^x , $x > 1$. Focal sets of m are of the form $E_I = \cap_{i \in I} A_i$, $I \subseteq \{1, \dots, k\}$ with masses $m(E_I) = \prod_{i \in I} (1 - w_i) \prod_{i \notin I} w_i$ (where we allow that some E_I 's may be identical). Combining this mass function with E_J^x yields a mass function m' such that $m'(E_J) = xm(E_J) + (1 - x)(\sum_{I \subset J} m(E_I)) = xm(E_J) + (1 - x)Bel(E_J)$ (where $E_\emptyset = \Omega$). So E_I is erased from the focal sets of m by E_J^x if and only if $x = \frac{Bel(E_J)}{Bel(E_J) - m(E_J)} = \frac{\sum_{I \subset J} m(E_I)}{\sum_{I \subset J} m(E_I)} = 1 / (1 - \prod_{i \in J} (1 - w_i))$, which is clearly more than 1. Note that we can erode a single focal set via retraction, while discounting affects all focal sets to the same extent. Similarly, it can be checked that if $J \subset I$, $m'(E_J) = (\prod_{i \notin I} w_i) \prod_{i \in I \setminus J} (1 - w_i) (1 - x + x \prod_{i \in J} (1 - w_i)) = 0$ if and only if $x = 1 / (1 - \prod_{i \in J} (1 - w_i))$ again, while if $J \not\subset I$, $m'(E_I) = xm(E_I)$, which is provably less than 1. In other words, retracting the focal set E_J erases all focal sets $E_I \subset E_J$ as well, namely all combinations between the merging of information from sources indexed in J , with information from other sources.

So we can consider that any belief function comes from merging unreliable elementary testimonies, with prejudices that weaken the weights pertaining to the conjunctions of information items coming from sources. It is indeed natural to consider that information we receive from the outside is challenged by our prior information taking the form of stereotypes, or prejudices that one is often unaware of. The receiver is reluctant to consider the result of such conjunction valid. For instance, consider a variant of the Linda problem [3]. In this case,

the bank teller Linda, depicted as a philanthropist, is found by participants to a psychological experiment, more likely to be a philanthropist bank teller than a bank teller, because the former looks more “representative” or typical of persons who might fit the description of Linda. Here we consider the case when we receive two testimonies, namely one (B^v) claiming that Linda is a banker and another one A^w that she is a philanthropist. The fusion process leads us to allocate a belief degree $(1 - v)(1 - w)$ to the fact that she is a philanthropist bank teller. However, a prejudiced individual would hardly believe that a bank teller can be philanthropist, and would like to erode, possibly erase, this belief by combining the result of the fusion with the generalized SSB $(A \cap B)^u$ with $1 < u \leq 1/(v + w - vw)$, which leads to a belief degree equal to $1 - (u + v - vw)u$, that is all the lesser as the prejudice is strong.

5 Conclusion

This paper revisits the decomposition of a belief function into a combination of generalized simple support functions proposed by Smets [2] showing that it can be viewed as the merging of uncertain testimonies and of prejudices against the results of their partial conjunctions. We have laid bare new formal properties of the diffidence function w and shown how to reconstruct the BBA m from it via Moebius-like transforms. Our results strengthen the approach to belief function based on the merging of pieces of evidence, as opposed to the approach based on upper and lower probability. Future research can be a study of the information ordering based on diffidence functions, introduced by Denœux [4], on which our results can shed more light.

References

1. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
2. Smets, P.: The canonical decomposition of a weighted belief. In: Proceedings 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, vol. 2, pp. 1896–1901, 20–25 August 1995
3. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* **90**, 293–315 (1983)
4. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artif. Intell.* **172**(2), 234–264 (2008)
5. Shenoy, P.P.: Conditional independence in valuation-based systems. *Int. J. Approx. Reason.* **10**(3), 203–234 (1994)
6. Ginsberg, M.L.: Non-monotonic reasoning using Dempster’s rule. In: Proceedings of National Conference on Artificial Intelligence, Austin, TX, pp. 126–129, 6–10 August 1984
7. Kramosil, I.: Probabilistic Analysis of Belief Functions. Kluwer, New York (2001)
8. Pichon, F., Denœux, T.: On Latent belief structures. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 368–380. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75256-1_34

9. Schubert, J.: Clustering decomposed belief functions using generalized weights of conflict. *Int. J. Approx. Reason.* **48**(2), 466–480 (2008)
10. Dubois, D., Prade, H.: An introduction to bipolar representations of information and preference. *Int. J. Intell. Syst.* **23**(8), 866–877 (2008)
11. Dubois, D., Prade, H., Smets, P.: “Not impossible” vs. “guaranteed possible” in fusion and revision. In: Benferhat, S., Besnard, P. (eds.) ECSQARU 2001. LNCS (LNAI), vol. 2143, pp. 522–531. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44652-4_46
12. Ke, X., Ma, L., Wang, Y.: Some notes on canonical decomposition and separability of a belief function. In: Cuzzolin, F. (ed.) BELIEF 2014. LNCS (LNAI), vol. 8764, pp. 153–160. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11191-9_17
13. Pichon, F.: Canonical decomposition of belief functions based on Teugels representation of the multivariate Bernoulli distribution. *Inf. Sci.* **428**, 76–104 (2018)



A Heuristic Approach for the Robust Flight Level Assignment Problem

Akli Fundo¹, Dritan Nace²(✉), and Chenghao Wang²

¹ University Polytechnic of Tirana, Sheshi Nene Tereza, Tirana, Albania

² Sorbonne université, Université de technologie de Compiègne, CNRS, UMR 7253, Heudiasyc - CS 60 319 - 60 203, Compiègne Cedex, France
nace@utc.fr

Abstract. The paper studies the flight level assignment (FLA) problem and its robust variant. Our goal is reducing the total cost (and more specifically the flight delay) induced by airspace congestion through an appropriated FLA taking account of uncertainties such as weather condition, flight velocity, flight departure time, etc. Among these uncertainties, we assume that the flight departure time, which follows a Mixture Gaussian Distribution, is certainly one of the main uncertainty factors worthy to deal with. The deterministic FLA problem is formulated through an Integer Linear Programming (ILP) model, which becomes trickier when the uncertainty aspect is considered. The FLA problem is strongly NP-hard and solving it exactly is out of reach even for moderate realistic instances. Hence, we propose an approximated optimization approach to solve the robust FLA problem. The main idea is to decompose the problem by levels and solving it separately while handling the connecting constraints between levels. Numerical results illustrate our findings.

Keywords: Robust optimization · Flight level assignment
Linear programming · Hoeffding's inequalities
Monte-Carlo Simulation

1 Introduction

With the high increasing demand for commercial flights each year, the Air Traffic Management (ATM) is becoming more and more complex and less efficient in reducing air traffic congestion. With respect to air traffic congestion, two main types can be identified corresponding to areas of airspace: terminal congestion (around airports) and en-route congestion (between airports). We are interested in reducing the en-route congestion and its induced cost while taking into account uncertainties. The paper studies the flight level assignment (FLA) problem and its robust variant. Our goal is reducing the total cost (and more specifically the flight delay) induced by airspace congestion through an appropriated FLA taking account of uncertainties such as weather condition, flight velocity, flight departure time, etc. We have shown in [7] that the FLA problem is NP-hard even for instances with only three altitude levels. It may also

be shown easily that for a single altitude level the problem of maximizing the number of flights accommodated to this level is NP-hard by reduction to the maximum independent set problem. This work is in continuation of [7, 11].

This paper is organized as follows. After this introduction, in Sect. 2 we report a short discussion on related works and position the problem with respect to ATM. In Sect. 3 we present our approach and discuss in detail the FLA problem for a single level. In Sect. 4, we firstly report a discussion on conflict probability estimation based on the flight departure time and its induced cost, then computational results illustrate our findings. Finally, a conclusion is given in Sect. 5.

2 Context of the Work

Related Works. Optimization problems in ATM have been widely studied these last decades. We focus on some works related to a certain extent to the flight level assignment problem. Let us firstly refer to [6], Cook et al. have shown that how uncertainty affects the ATM system is the key element to a proper model and control it and improve its performance. The source of uncertainties varies from aircraft velocity and weather condition to flight departure time, etc. Based on uncertain predicted trajectories, Irvine presented in [10] a more simplified geometrical calculation of conflict probabilities. Babak et al. in [1] studied on the stochastic methods of conflict situation detection and conflict probability evaluation. A more recent study which accounts for the effects of wind uncertainties was presented in [8].

For a conflict resolution by rerouting, let us firstly cite Bertsimas and Stock [3] who show how to optimally control aircraft by rerouting, delaying, or adjusting the speeds of the aircraft in the ATC (Air Traffic Control) system to avoid airspace regions with reduced capacities due to weather conditions. In [4], Bertsimas et al. proposed a new ILP model for large-scale instances which covered all the phases of each flight and solved it for an optimal combination of flow management actions, including rerouting decisions. Constans et al. have proposed minimizing potential conflict quantity by dynamically imposing feasible modifications on the speeds of the aircraft in [5]. In [7, 11] we have already presented some work on FLA problem. This paper uses a similar mathematical model and extends it for the case of aircraft departure time following Mixture Gaussian distribution. New numerical results are also reported.

Problem Description. In real air traffic management, the airspace is regulated by a certain number of rules, one of them is the “Semicircular Rule”. According to this rule, an aircraft is not assigned consecutively one by one altitude levels but two by two at least, as in the European airspace, and even four by four in the USA since the European airspace is more restrictive than the American one. The air traffic controllers classify the aircraft depending on the angle of motion of the trajectories. So, they divide the set of aircraft under consideration in two groups: the ones flying with an angle of motion between $-\pi$ and 0 radians (for instance) and the rest of aircraft (flying in the opposite direction). The aircraft

in the first group are requested to fly only in the odd altitude levels and the ones in the second group are requested to fly only in the even altitude levels, even if they must change their altitude due to other conditions during the flight. In order to provide more safety to the airspace, the air traffic controllers follow these guidelines to reduce the number of conflict situations. So, if a conflict situation takes place, at least one aircraft is requested to follow some maneuver as heading angle or velocity changes, or in some situation climbing or descending to the following altitude level in which it is allowed to fly according to the Semicircular Rule. In our work we assume that during the planning phase, there will be a fixed level assigned to each aircraft and the aircraft is supposed to stay to this level for or the entire enroute flight period. We assume that for each aircraft there is a most preferred flight level which is decided mostly by the type of the aircraft and fuel consumption considerations. There are also some other alternative eligible immediate upper or lower levels, which allow to deal with congestion, whereas involving an additional cost. This paper deals with the problem of assigning a set of flights with given flight paths to different levels such that potential costs of conflict over all flights are minimized. We explore a stochastic version under a robust optimization framework. Some numerical results based on a test instance are also reported.

3 Mathematical Model and Solution Approach

We start with the mathematical formulation of the robust FLA problem with probability constraints. We assume that each constraint has to be feasible with some probability $1 - \epsilon$.

Notation:

- L gives the set of the flight levels l and F^l groups all flights allowed to fly to level l .
- x_i^l is a binary variable that takes value 1 when the aircraft i flies on level l and 0 otherwise;
- b_i^l gives the profit associated with flight i when assigned at level l ;
- P_i^l gives the admissible cost for a given flight i at level l ;
- S_i^l gives the set of flights j having a potential conflict with flight i when they fly in the same level l ;
- p_{ij} is the induced cost associated with aircraft i when resolving a potential conflict with aircraft j ;
- M_i is a large number.

Assuming separate probability conditions, the mathematical formulation of the probabilistic FLA problem follows:

$$\max \quad \sum_{i \in F^l, l \in L} b_i x_i \quad (1a)$$

$$s.t. \quad Pr\left(\sum_{j \in S_i^l} p_{ij} x_j + M_i x_i \leq M_i + P_i\right) \geq 1 - \epsilon, \forall i \in F^l, l \in L \quad (1b)$$

$$\sum_{l \in L_i} x_i^l = 1, \forall i \in F, \quad (1c)$$

$$x_i^l \in \{0, 1\}, \forall i \in F, l \in L_i. \quad (1d)$$

Probability constraints (1b) ensure for each aircraft that the sum of experienced costs/delays will not exceed some given admissible cost with a high probability $1 - \epsilon$. Constraints (1c) assigns each aircraft to some of its eligible levels, while the objective function looks for a solution that assigns the aircraft to the most preferred flight levels possible. Following the Bertsimas and Sim work [2], we can deduce the robust variant of the above problem by converting the probability constraints through some deterministic ones. This yields some ILP problem, which is at least as difficult as the conventional deterministic problem. All this justifies heading to approximated methods to deal with it. The main idea behind the proposed approach is to decompose the problem by altitude levels and deal with each of them separately. We handle the connections between levels through a greedy algorithm described at the end of the Section. We report below a detailed study of the problem associated to a single flight level called RP^l .

3.1 The Problem Associated with a Single Flight Level (RP^l)

Similarly to above, the mathematical formulation associated with the probabilistic FLA restricted to level l follows:

$$\begin{aligned} \max \quad & \sum_{i \in F^l} b_i x_i \\ s.t. \quad & Pr\left(\sum_{j \in S_i^l} p_{ij} x_j \leq M_i(1 - x_i) + P_i\right) \geq 1 - \epsilon, \forall i \in F^l \\ & x_i \in \{0, 1\}, \forall i \in F^l \end{aligned} \quad (2)$$

where for sake of simplicity we use b_i, x_i, P_i instead of b_i^l, x_i^l, P_i^l . Note also that p_{ij} stands here for a random variable.

The above program is known to be a very difficult one. One way to tackle it is to use the Bertsimas and Sim model [2] which is used under some mild probability conditions not applied in our problem. Hence, to solve the above problem we have opted to use the model introduced in [12]. Intuitively, we introduce a parameter vector $\gamma \in [0, 1]^{|F^l|}$ which allows tuning the robustness of the solution

in a convenient way. Applying this idea, we obtain the following model denoted below $RP_{l\gamma}$:

$$\begin{aligned} \max \quad & \sum_{i \in F^l} b_i x_i \\ \text{s.t.} \quad & M_i x_i + \min \left\{ \sum_{j \in S_i^l} \bar{p}_{ij} x_j, \gamma_i \cdot \sum_{j \in S_i^l} \bar{p}_{ij} \right\} \leq M_i + P_i, \forall i \in F^l \\ & x_i \in \{0, 1\}, \forall i \in F^l. \end{aligned} \quad (3)$$

where \bar{p}_{ij} gives the maximal value that can be attained by p_{ij} . The above formulation can be simplified a lot. Let us focus on the robust constraint i . Either we consider the worst case (maximum conflict induced costs), or we have a constraint: $M_i x_i + \gamma_i \cdot \sum_{j \in S_i^l} \bar{p}_{ij} \leq M_i + P_i$. In this latter case, two sub-cases occur: when $\gamma_i \cdot \sum_{j \in S_i^l} \bar{p}_{ij} > P_i$, then $x_i = 0$; when $\gamma_i \cdot \sum_{j \in S_i^l} \bar{p}_{ij} \leq P_i$, we have a dummy constraint which can be ignored.

These three cases are in fact summarized in the two following ones:

- either flight i has total conflict costs less than the admissible cost and no constraint is necessary to model this situation;
- or flight i is associated with maximal conflict costs, that is constraint $M_i x_i + \sum_{j \in S_i^l} \bar{p}_{ij} x_j \leq M_i + P_i$ represents this situation.

Hence, the analysis of the above robust model leads to a new one, which is very simple. Indeed, for a given value of γ_i we know in advance if the constraint corresponding to flight i is necessary to be put in the model or not. Let denote with $I_c \subseteq F^l$ a subset of concerned flights with respect to a given vector γ . In this way, instead of vector γ we use the subset I_c as a parameter enabling to tune robustness. We denote the corresponding problem by $RP^l(I_c)$.

$$\begin{aligned} \max \quad & \sum_{i \in F^l} b_i x_i \\ \text{s.t.} \quad & M_i x_i + \sum_{j \in S_i^l} \bar{p}_{ij} x_j \leq M_i + P_i, \forall i \in I_c \\ & x_i \in \{0, 1\}, \forall i \in F^l \end{aligned} \quad (4)$$

With respect to vector γ considered, the size of the above LP varies between a few constraints (for small values of γ_i) and all constraints (for $\gamma_i = 1, \forall i$).

In the heuristic we use several parameters as $p_{ij}, \bar{p}_{ij}, P_i$ for which we have developed specific estimation methods not presented here because of lack of space. The main idea behind the Algorithm 1 is to build the solution by taking into account only the most restrictive constraints while the other flights are set by default to their most preferred flight level. The feasibility of the obtained solution is checked and if necessary new constraints are added in the ILP. Hence, an important aspect studied in this work is the estimation of solution's feasibility probability as presented below.

Algorithm 1. A heuristic approach for RP^l

procedure SOLVERPL

Set $I_c \leftarrow \emptyset$;

Select a few (say 5) number of flights i maximizing the $\sum_{j \in S_i^l} \bar{p}_{ij} - P_i$ value;

Set $I_c \leftarrow I_c \cup \{i\}$; Solve $RP^l(I_c)$; Let x^* be the initial solution found;

while True do
if feasibility probability of $x^* \geq 1 - \epsilon$ for all concerned flights **then**

An approximate robust solution is found ;Stop.

else

Select flight i such that $Pr(\sum_{j \in S_i^l} p_{ij}x_j^* \leq P_i) \geq 1 - \epsilon$ is the most violated;

Set $I_c \leftarrow I_c \cup \{i\}$; Solve $RP^l(I_c)$; Let x^* be the optimal solution found.

3.2 Solution Feasibility Estimation

Note first that the main uncertain parameter that we have considered is the departure time. Hereby we assume that the flight departure time follows a 4-component Mixture Gaussian Distribution proposed in [14].

We discuss now the methods that estimate the feasibility of solution assuming separated constraints for each flight: $Pr(\sum_{j \in S_i^l} p_{ij}x_j + M_i x_i \leq M_i + P_i) \geq 1 - \epsilon$. As $Pr(\sum_{j \in S_i^l} p_{ij}x_j + M_i x_i \leq M_i + P_i) \geq Pr(\sum_{j \in S_i^l} p_{ij}x_j \leq P_i)$, we restrict ourselves in ensuring that $Pr(\sum_{j \in S_i^l} p_{ij}x_j \leq P_i) \geq 1 - \epsilon$ for all $x_i = 1$.

Conservative Robust Method: we consider first the Soyster model [13], which looks for a solution robust to the worst case. This gives: $\sum_{j \in S_i^l} \bar{p}_{ij}x_j \leq P_i$ for all $x_i = 1$, which is equivalent to $Pr(\sum_{j \in S_i^l} p_{ij}x_j \leq P_i) = 1$.

Probability Bound method: We apply the Hoeffding's Inequality [9], which gives:

$$\begin{aligned} Pr\left(\sum_{j \in S_i^l} p_{ij}x_j \geq P_i\right) &= Pr\left(\sum_{j \in S_i^l} p_{ij}x_j - E\left[\sum_{j \in S_i^l} p_{ij}x_j\right] \geq P_i - E\left[\sum_{j \in S_i^l} p_{ij}x_j\right]\right) \\ &\leq \exp\left(-2\left(P_i - \sum_{j \in S_i^l} E[p_{ij}]x_j\right)^2 / \left(\sum_{j \in S_i^l} \bar{p}_{ij}^2 x_j\right)\right) = \epsilon_i \end{aligned} \quad (5)$$

However, when $P_i \leq \sum_{j \in S_i^l} E[p_{ij}]x_j$, by definition of Hoeffding's Inequality, the above formula can't be applied, we thus set the probability $Pr(\sum_{j \in S_i^l} p_{ij}x_j \leq P_i)$ as zero. In case that P_i is bigger than the sum of all upper bounds of random variables, then the probability is surely 1. Thus, we obtain a piece-wise probability function as follows:

$$Pr\left(\sum_{j \in S_i^l} p_{ij}x_j \leq P_i\right) = \begin{cases} 0, & \text{if } P_i \leq \sum_{j \in S_i^l} E[p_{ij}]x_j \\ 1, & \text{if } \sum_{j \in S_i^l} \bar{p}_{ij}x_j \leq P_i \\ 1 - \epsilon_i, & \text{otherwise} \end{cases} \quad (6)$$

Sampling Method: The last method tested is based on Monte-Carlo Simulation. We have randomly generated a large number of scenarios where for each

flight the departure time is generated following the above mentioned Mixture Gaussian distribution.

3.3 Putting All the Pieces Together

We describe now a heuristic approach for the Robust FLA problem, that is deciding flight level assignment robust to uncertainties that can affect flights, essentially due to fluctuation on departure time. The main idea behind the Algorithm is to decompose the problem by altitude levels and deal with each of them separately (as described above), while handling the connections between levels.

Algorithm 2. ApproxRobustFLA

Step 0:

Order levels in L following decreasing order of loads (estimated by the number of concerned flights in their most preferred level)

Step 1:

Proceed with flight level assignment separately for each level (following the order set in Step 0); solve problem RP^l involving all flights with the most preferred flight level l and other unassigned flights in F^l ; fix the level for flights assigned in the obtained solution.

Step 2:

if All flights are assigned or the maximal number of iterations is exceeded, **then** Stop;
else increase admissible cost for each unassigned flight and go to **Step 1**;

4 Implementation and Numerical Results

The code is realized with C++ with Cplex 12 under Ubuntu 16.04 LTS-64 bits, i7-7820 HQ CPU @2.90GHz, 16G RAM. The test data corresponds to French air traffic of August 12th, 1999. Table 1 presents the characteristics of test data.

Table 1. Test instance

Network	Number of flights	Used airports	Used WayPoints
NET_FR	1273	134	715

In Table 2, P_i is bounded in $[0, 30]$, calculated by $P_i = \text{duration_of_flight}_i * \text{coPi}$ (where CoPi indicates the percentage of flight time allowed for conflict resolution), and the maximal number of iterations in Algorithm 2 is set to 10. **eps** stands for the infeasibility tolerance of solution, **#CL** indicates the number of flight changes from their most preferred flight level to a feasible one, **#UF**

Table 2. Numerical results

param			RobustDet			Hoeffding				Monte-Carlo				
Ins	coPi	eps	#CL	#UF	#CM	#ElaTi	#CL	#UF	#CM	#ElaTi	#CL	#UF	#CM	#ElaTi
B	0.05	0.05	276	13	13	24.18	260	11	13	20.89	165	4	11	2,267.90
		0.10				27.75	230	12	12	12.62	155	3		1,874.15
		0.15				26.11	218	10	12	18.60	147	3		1,627.20
		0.20				26.06	193	9	12	20.31	149	4		1,586.39
		0.25				25.69	206	8	12	11.08	142	3		1,560.20
	0.10	0.05	128	3	11	8.83	118	2	11	7.93	96	1	11	1,428.87
		0.10				9.88	114	2		8.11	67	0		808.32
		0.15				8.98	114	1		9.15	53	0		578.40
		0.20				8.27	115	1		8.27	40	0		348.89
		0.25				8.86	113	1		8.55	39	0		368.84
	0.15	0.05	72	0	11	2.55	62	0	11	3.13	22	0	11	193.63
		0.10				1.98	42			1.77	19			180.00
		0.15				1.68	39			1.75	15			156.91
		0.20				1.86	37			1.41	11			97.80
		0.25				1.62	35			1.76	10			95.71
I	0.05	0.05	272	15	14	29.91	261	9	16	25.70	178	2	14	2,217.22
		0.10				30.76	241	13	14	18.51	173	3		2,624.45
		0.15				31.64	233	4	14	12.96	157	2		2,055.12
		0.20				30.49	215	9	14	12.05	156	1		2,680.39
		0.25				31.53	211	8	14	12.63	156	1		2,363.03
	0.10	0.05	137	2	14	10.79	126	2	14	9.03	69	0	14	994.31
		0.10				9.95	126	1		8.99	52			771.11
		0.15				9.94	123	0		9.38	49			673.50
		0.20				10.00	111	0		6.70	46			577.38
		0.25				9.46	116	1		9.55	42			468.15
	0.15	0.05	74	1	14	9.82	53	0	14	1.75	25	0	14	296.24
		0.10				10.55	52			2.38	21			246.30
		0.15				9.72	43			1.55	17			256.52
		0.20				10.20	43			1.72	14			205.04
		0.25				10.50	39			1.48	12			192.00

denotes the number of unassigned flights, **#CM** specifies the maximal number of potential conflict occurring for a flight in the given feasible solution, **ElaTi** gives the elapsed time on seconds to get a robust feasible solution. We have tested two types of instances: the B (basic) instances are these reported in Table 1 and I (incremented) instances give the basic instances incremented with 15% additional flights among the existing ones but scheduled 5 hours later. The obtained results show clearly that Monte-Carlo estimation method gives more satisfactory results for all scenarios.

5 Conclusion

In our work, we deal with robust FLA problem assuming the flight departure time as the main source of uncertainty [14]. We experiment several methods

showing that the sampling method (Monte-Carlo Simulation) gives an accurate solution when the distribution of flight-induced cost is hard to analyze, however, the biggest inconvenience is that this method is expensive on computation time. Therefore, an analytical approximate method will be in focus of our future work.

References

1. Babak, V., Kharchenko, V., Vasylyev, V.: Methods of conflict probability estimation and decision making for air traffic management. *Aviation* **10**(1), 3–9 (2006)
2. Bertsimas, D., Sim, M.: The price of robustness. *Oper. Res.* **52**(1), 35–53 (2004)
3. Bertsimas, D., Patterson, S.S.: The traffic flow management rerouting problem in air traffic control: a dynamic network flow approach. *Transp. Sci.* **34**(3), 239–255 (2000)
4. Bertsimas, D., Lulli, G., Odoni, A.: An integer optimization approach to large-scale air traffic flow management. *Oper. Res.* **59**(1), 211–227 (2011)
5. Constans, S., Fontaine, B., Fondacci, R.: Minimizing potential conflict quantity with speed control. In: *Proceedings of the 4th Eurocontrol Innovative Research Workshop And Exhibition*, pp. 265–274, December 2005
6. Cook, A.: Applying complexity science to air traffic management. *J. Air Transp. Manage.* **42**, 149–158 (2015)
7. Fundo, A., Nace, D., Savourey, D., Gjata, F.: The robust flight level assignment problem. In: *PGMO Days*, 8–9 November, Saclay, France (2016)
8. Hernández, E., Valenzuela, A., Rivas, D.: Probabilistic aircraft conflict detection considering ensemble weather forecast. In: *6th SESAR Innovation Days*, The Netherlands (2016)
9. Hoeffding, W.: Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **58**(301), 13–30 (1963)
10. Irvine, R.: A geometrical approach to conflict probability estimation. *Air Traffic Control Q.* **10**(2), 85–113 (2002)
11. Klopfenstein, O., Nace, D.: The robust flight level assignment problem. In: *ICRAT 2008*, 3rd International Conference on Research in Air Transportation (2008)
12. Klopfenstein, O.: Tractable algorithms for chance-constrained combinatorial problems. *RAIRO Oper. Res.* **43**(2), 157–187 (2009)
13. Soyster, A.L.: Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper. Res.* **21**(5), 1154–1157 (1973)
14. Tu, Y., Ball, M.O., Jank, W.S.: Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *J. Am. Stat. Assoc.* **103**(481), 112–125 (2008)



Study of Distributed Data Fusion Using Dempster's Rule and Cautious Operator

Romain Guyard^(✉) and Véronique Cherfaoui

Sorbonne Université, Université de Technologie de Compiègne,
CNRS Heudiasyc UMR, 7253 Compiègne, France
romain.guyard@hds.utc.fr

Abstract. This paper presents new algorithms to process data exchanged in vehicular networks. In previous works, a distributed data fusion method using belief functions to model uncertainties has been proposed for smart cars network. Since the origin of data coming from other cars is unknown, this algorithm uses the idempotent cautious operator in order to prevent data incest. This operator has been proved to be efficient in the case of transient errors and ensures the fusion convergence. However, since the cautious operator is idempotent, the quantity of concordant sources does not change the result of fusion. Thus we propose several schemes adding Dempster's rule in order to improve the fusion when we can ensure that data come from independent sources. We introduce three new combinations layout of Dempster's rule and cautious operator and we compare them using real data coming from experiments involving several communicating cars in the context of the COMOSEF project.

1 Introduction

Most smart car perception system are based on data coming from embedded sensors. However, despite technological improvements, their capabilities are limited. Thus, we propose to use information coming from other vehicles thanks to wireless network. Two methods are possible. The first one, known as cloud computing, uses a central server that gathers data from all the vehicles, then computes interesting results and sends them back. Even if this method is currently privileged, it has drawbacks. For instance, it introduces latency due to network exchanges that can be a problem in a context of high topological dynamics. Moreover, private data from users are sent to a third party and that can be an invasion of privacy. The other method, studied in this paper consists of peer-to-peer communication between vehicles (VANETs) [1]. By exchanging data with the vehicle neighbors and with neighbors of neighbors it is possible to benefit of other's knowledge to complete our local sensors capacities. Trusting other cars may not always be possible because their data can be wrong, intentionally or not. Thus, the algorithm we will use must work even if some neighbors send false data. Some research shows the possibility of hybrid VANETs and cloud computing [2].

Vehicle networks are different from regular networks because the topology can change at any time. A WIFI norm called 802.11p has been developed for this kind of dynamics. In order to converge to a common value, a robust and dynamic algorithm for distributed data fusion has been proposed in [3]. Each node of the network takes its local value and fuses it using the cautious operator [4] with data that neighbors sent to it. Then, it sends the result to all other cars which do the same thing. Thus, a same value is fused at each hop and then we have to be careful to not artificially increase the confidence due to cycle fusion. This algorithm has limitations. Firstly the idempotent cautious operator gives the same result if there is one car or multiple cars with the same local value. Secondly it doesn't fuse data over time. This paper compare alternatives of the fusion process by combining Dempster's rule with the cautious operator in order to increase the importance of the result.

2 Distributed Data Fusion Algorithms

2.1 Belief Combination

Dempster's rule shown in Eq. 1 is the most used fusion operator in the belief function framework.

$$\begin{cases} (m_1 \oplus m_2)(\emptyset) = 0 \\ A \neq \emptyset & (m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B)m_2(C) \\ \text{where } K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \end{cases} \quad (1)$$

With Dempster's rule, sources must be independent. In the car to car communication context, the independence is not guaranteed, so the fusion operator have to be idempotent. To fulfill this requirement, the cautious operator has been proposed in [4]. This operator is applied on weight using the operator minimum as shown in Eq. 2. Let be a variable ω , taking values in a finite set Ω called frame of discernment, 2^Ω the set of subsets of Ω and $A \subseteq \Omega$ a set. Let $m(A)$ be the mass function and $w(A)$ the weigh function.

$$(w_1 \otimes w_2)(A) = \min(w_1, w_2) \quad (2)$$

In order to maximize the truthfulness of the fusion, Dempster's rule should be used in priority if the sources are independent. If they are dependent, the cautious operator must be used.

2.2 Original Algorithm

This paper is based on the smart cars data fusion algorithm proposed in [3]. Its objective is to detect events using observations originated from multiple cars. The knowledge about an event of the fleet of vehicles converges to a common value (self-stabilization) using car to car communications. Algorithm 1 (INv[u] and OUTv are mass functions), illustrated in Fig. 1 (\otimes represents the cautious

operator), shows that the combination is done using the cautious operator. The following section will discuss possible improvement of this operator. Every loop, every data represented by masses functions received from neighbor is fused with the local BBA and the result is broadcasted. The $r()$ function is a discounting operation that reduces the importance of old and distant data and has been discussed in [5]. In [6], the authors present a different algorithm that does not need to use the cautious operator. In this method, instead of sending fused data, local data is broadcasted. Since every local data is associated with the source vehicle, there is no data incest, thus Dempster’s rule can be used. However, this require a message to be sent for every observation therefore it can be network intensive. Moreover every node knows personal data which can be a privacy issue. Future work should compare the two approaches.

Algorithm 1. Distributed data fusion algorithm

- 1: Upon the arrival of a new message:
 - 2: receive(*dist.mass*) from node u
 - 3: $INv[u] \leftarrow dist.mass$
 - 4: Upon the expiration of the timer of the node v
 - 5: $OUTv \leftarrow compute_local_confidence()$
 - 6: **for each** u in INv **do**
 - 7: $OUTv \leftarrow OUTv \otimes r(INv[u])$
 - 8: **end for**
 - 9: send($OUTv$) in the neighborhood
 - 10: Remove old messages in INv
 - 11: Restart the timer
-

2.3 Modification of the Distributed Algorithm

We propose in this paper to study different scheme of distributed fusion. The idea is to use Dempster’s rule that increase the knowledge when data come from two independent sources. The cautious operator should be used otherwise.

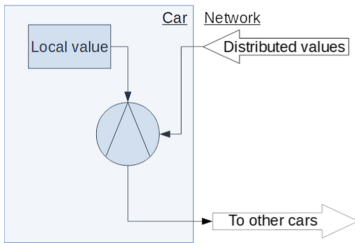


Fig. 1. Cautious operator only

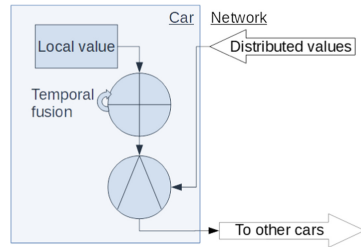


Fig. 2. Dempster’s rule before sending

Indeed, local value is always independent of values coming from the network. It is then possible to combine the with Dempster’s rule the local value and the

result of the cautious operator applied to all values sent by other cars (distributed values). Figure 2 (\oplus represents Dempster’s rule) shows the fusion diagram of this proposition. As proposed in [7] and commonly done in dynamic data fusion, we can add a temporal fusion of the local value with Dempster’s rule. This operation is done before the network data fusion as shown in Fig. 3. Finally, we propose a fourth fusion diagram in Fig. 4 that assumes that nodes send both distributed and local values. Dempster’s rule can be used to combine neighbor local values since independent local data are fused only once.

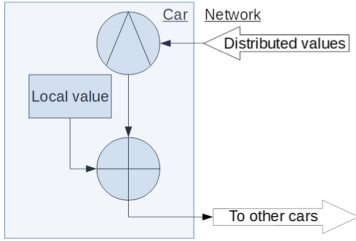


Fig. 3. Local temporal Dempster loop

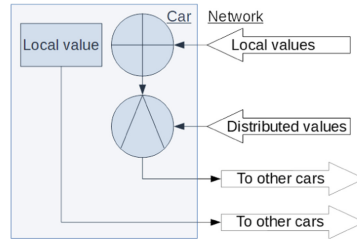


Fig. 4. By sending local values to neighbors

3 Comparison of Fusion Scheme with Experimental Data

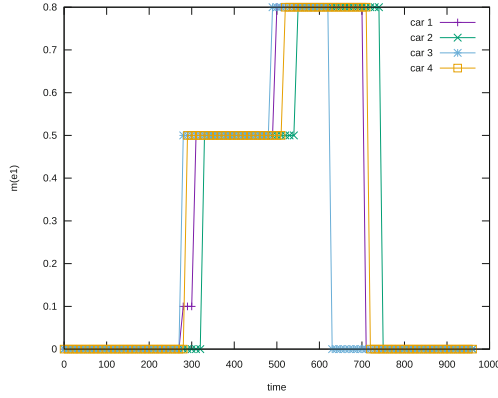
3.1 Dataset of Comosef Project

The European project CoMoSeF (Cooperative Mobility Service for the Future) has been launched in July 2012 and ended in 2016 [8]. The goal of this project is the creation of services and devices for cooperative application in transports. An experimentation has been done in order to demonstrate the efficiency of the algorithm in the French test site [9]. Heavy rains has been simulated by 10 vehicles in 3 different platoons by using wipers at given positions. The algorithm has generated alerts that has been broadcast to other cars but also to RSU (Road Side Unit). Live alerts generated by the distributed data fusion has been shown on a website.

The original data of the Comosef experimentation has not been recorded. Few days after an equivalent experimentation has been done with 4 cars. There is 4 levels of wiper speed: off, alternate, moderate and fast. Only one car is able to be in alternate mode. In this experimentation, cars follow each other and start their wiper at the same place. The speed increases until the fast mode is reached. The wipers are then stopped without going through intermediate levels. The frame of discernment is $\Omega = \{r, c, s\}$ with r representing the event “heavy rain”, c the event “Cloudy” and s the event “Sunny”. Masses are computed from the wiper speed using Table 1. Figure 5 shows for each vehicle the mass on event “heavy rain” with local data coming directly from wipers. The following of this paper will use those data in order to show the importance of how data fusion operators are combined.

Table 1. BBA for Comosef data

Wiper speed	\emptyset	$\{r\}$	$\{c\}$	$\{r, c\}$	$\{s\}$	$\{r, s\}$	$\{c, s\}$	Ω
Off	0	0	0	0	0.8	0	0	0.2
Alternate	0	0.1	0	0.5	0	0	0	0.4
Moderate	0	0.5	0	0.2	0	0	0	0.3
Fast	0	0.8	0	0	0	0	0	0.2

**Fig. 5.** Local data $m(r)$ of the four cars during the Comosef experimentation

3.2 Comparison of Fusion Scheme

In this section we compare the different combinations of operators using Comosef experimental data.

Comparison in the Original Scenario. Figure 6 shows the result of the fusion using the 4 variants of the algorithm. The top left graph represents the fusion using only the cautious operator. It shows only few changes with the local data. It can be observed that the fusion warns cars of rain before it happens. Even if the same measurement is done multiple time by all the cars, the result of the fusion stays low. The top right and the bottom left graphs show the fusion with the Dempster's rule respectively added before and after the cautious operator. They have similar behavior in this case. The Dempster's rule enables the fusion to rise to almost 1 when all cars agree on the same value. The last graph is even closer to reach 1, even if the wipers are not at full speed.

Comparison with Altered Scenarios. As previously seen, data from the Comosef experiment are almost perfect and every cars communicate without issue. In this part we simulate some changes that could occur in other scenarios. Figure 7 shows the 4 variants of the algorithm supposing that car #1 has

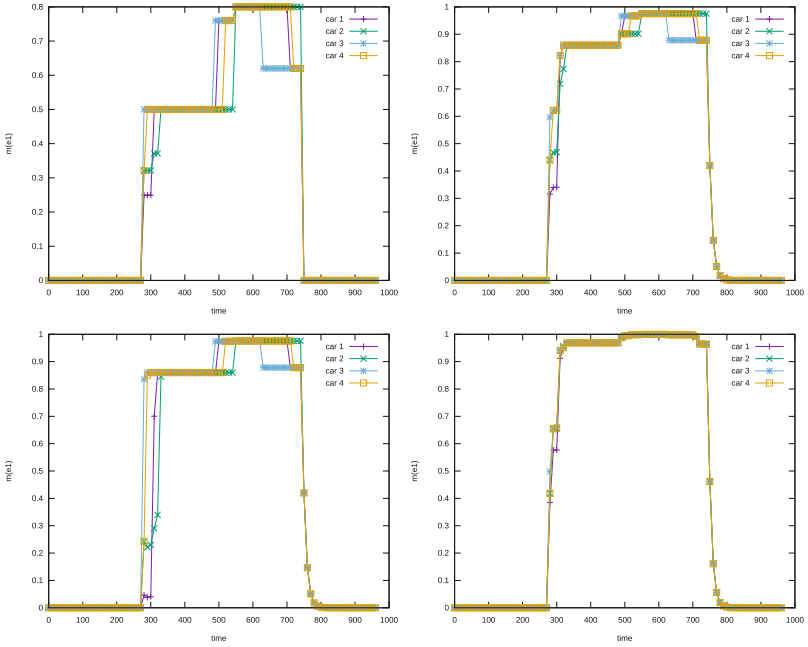


Fig. 6. Comoset data fusion using different combination operators

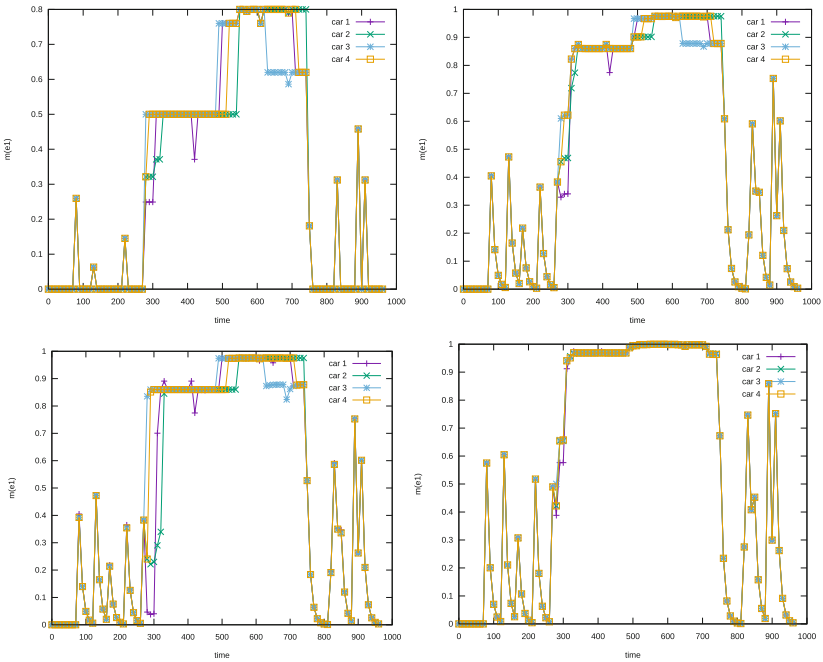


Fig. 7. Comoset data fusion with error injected

5% chance of getting wrong data from the CAN. We can see that the original algorithm is less impacted by transient errors. This behavior comes from the smoothing of Dempster’s rule. Errors are only present when cars do not have any information. When the wipers are on, errors are negligible.

In order to compare the fusion on different scenarios, we have created a metric called *fusion error*. The fusion error is the mean of the differences between the real value and the pignistic probabilities of fused data computed at each time divided by the number of cars. The ground truth is 1 if a wiper is turned on. Table 2 shows the fusion errors of the four fusion algorithms in four scenarios. In the original scenario, doing a fusion is better than not doing any fusion. As previously seen adding Dempster’s rule at one hop is the best fusion in this case. In the case of error injected in one car, the difference is lower but still present. When all cars have errors, there is almost no benefit to perform fusion. Finally, we can observe there is no changes in the case the WiFi allows communication only with the vehicles just before and just after the car.

Table 2. Fusion errors: (Fusion 1 = cautious only, Fusion 2.1 = with Dempster after cautious, Fusion 2.2 = with local Dempster loop and Fusion 3 = Fusion with Dempster on neighbor local values)

	No fusion	Fusion 1	Fusion 2.1	Fusion 2.2	Fusion 3
Original scenario	0.381	0.365	0.300	0.298	0.275
Errors car #1	0.383	0.374	0.326	0.325	0.309
Errors all cars	0.387	0.395	0.368	0.375	0.364
Low WiFi range	0.382	0.365	0.300	0.297	0.276

4 Conclusion

In this paper we have studied different scheme to fuse data in a distributed way. Three algorithms have been presented compared on experimental data. The algorithm given the best result uses Dempster’s rule with the data that are guaranteed coming from independent sources and cautious operator else. But this data fusion architecture requires to send more data that could be considerate as private. Future work should focus on testing these algorithms with more cars and in more realistic scenarios.

Acknowledgments. This work was carried out in the framework of the challenge DAPAD (Distributed and Augmented vehicle Perception to support Autonomous Driving) funded by the Labex MS2T, supported by the French Government, through the program “Investments for the future” managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). The author would like to thank Bertrand Ducourthial for his work on the Comosef project.

References

1. Da Cunha, F.D., Boukerche, A., Villas, L., Viana, A.C., Loureiro, A.A.: Data communication in VANETs: a survey, challenges and applications. Ph.D. thesis, March 2014
2. Hussain, R., Son, J., Eun, H., Kim, S., Oh, H.: Rethinking vehicular communications: merging vanet with cloud computing. In: 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 606–609, IEEE (2012)
3. Ducourthial, B., Cherfaoui, V., Denoeux, T.: Self-stabilizing distributed data fusion. In: Richa, A.W., Scheideler, C. (eds.) SSS 2012. LNCS, vol. 7596, pp. 148–162. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33536-5_15
4. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artif. Intell.* **172**, 234–264 (2008)
5. Guyard, R., Cherfaoui, V.: Study of discounting methods applied to canonical decomposition of belief functions. In: 21st International Conference on Information Fusion, Cambridge, United Kingdom (Great Britain), July 2018
6. Farah, M., Mercier, D., Delmotte, F., Lefèvre, É.: Methods using belief functions to manage imperfect information concerning events on the road in VANETs. *Transp. Res. Part C Emerg. Technol.* **67**, 299–320 (2016)
7. El Zoghby, N.: Distributed data fusion in VANETs. Ph.D. thesis, Université de Technologie de Compiègne, February 2014
8. <https://www.celticplus.eu/project-comosef/>
9. Ducourthial, B., Cherfaoui, V.: Experiments with self-stabilizing distributed data fusion. In: 2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS), pp. 289–296, September 2016



Uncertainty-Aware Parzen-Rosenblatt Classifier for Multiattribute Data

Ali Hamache, Mohamed El Yazid Boudaren^(✉), Houdaifa Boukersoul,
Islam Debicha, Hamza Sadouk, Rezki Zibani, Ahmed Habbouchi,
and Omar Merouani

Ecole Militaire Polytechnique, PO Box 17, 16111 Bordj El Bahri, Algiers, Algeria
boudaren@gmail.com

Abstract. Dempster-Shafer theory has proven to be one of the most powerful tools for data fusion and reasoning under uncertainty. Despite the huge number of frameworks proposed in this area, determining the basic probability assignment remains an open issue. To address this problem, this paper proposes a novel Dempster-Shafer scheme based on Parzen-Rosenblatt windowing for multi-attribute data classification. More explicitly, training data are used to construct approximate distributions for each hypothesis, and per each data attribute, using Parzen-Rosenblatt window density estimation. Such distributions are then used at the classification stage, to generate mass functions and reach a consensus decision using the pignistic transform. To validate the proposed scheme, experiments are carried out on some pattern classification benchmarks. The results obtained show the interest of the proposed approach with respect to some recent state-of-the-art methods.

Keywords: Classification · Dempster-Shafer theory · Multimodal data

1 Introduction

Data fusion improves decision making quality when heterogeneous sources of data are available, mainly by exploiting redundancy and complementariness among sources. One among the most flexible mathematical tools, the Dempster-Shafer theory (DST) [21,22] generalizes the Bayesian theory by (i) allowing each source to incorporate information in different levels of detail, which allows for uncertainty handling (unsure, imprecise, unreliable or missing information); and, (ii) offering a powerful mechanism for consensus decision making. Such theory has then been extensively applied in many fields [3,9,11,13]. In spite of this popularity, the crucial step which consists in defining the basic probability assignments is still an open problem. Most approaches determine mass values heuristically from data characteristics based on some measures like fuzzy membership degrees [30], distance to cluster centers [2] or probability densities [14].

The contribution of this paper falls under this latter category and relies on the use of DST for both mass functions construction and decision making.

Probabilistic frameworks for mass generation benefit from the rich literature of conventional probabilistic classifiers. Most of such approaches [1, 14, 20] represent the information related to each data attribute through probability density functions (PDFs), typically Gaussian. Such densities are then converted into beliefs which can be fused later to reach a collaborative decision. Compound hypotheses may be assigned masses by subtracting mass values related to involved individual hypotheses [14, 20] or through the mixture of distributions associated to such hypotheses [1]. Let us point out that Gaussian densities have been commonly considered for their easiness for most applications. When this assumption does not hold, however, the decision making performance may be significantly altered. More elaborated approaches overcome this limitation by converting data features to an equivalent normal space [28].

In this paper, we propose to cope with this drawback more efficiently by constructing PDFs that fit better original data histograms rather than projecting them into to a new Gaussian-like space. More explicitly, a kernel-smoothing estimation [4, 26] is applied to training data to infer an approximate PDF for each exclusive hypothesis, and per each data attribute. Hence, such PDFs can be of any form. In particular, they may be nonGaussian. At the classification stage, a given datum is assigned a set of masses generated, in some way, from the above mentioned densities. Multi-attribute masses are then fused through Dempster’s rule to reap a consensus mass. The classification decision is then inferred using some rules like the “maximum of plausibility” or the “Pignistic transform” [24]. We will see that doing so, one may improve classification accuracy. Let us point out that attributes are assumed independent here. Such assumption is usually set especially when attributes correspond to experts’ opinions. For dependent sources problem in DST, the reader may refer to [23].

The remainder of this paper is organized as follows: Sect. 2 recalls the basics of DST and Parzen-Rosenblatt window density estimation. Section 3 describes the proposed approach and explains its different steps. Experiments conducted on some universal datasets are presented and discussed in Sect. 4. Finally, concluding remarks and future improvements are given in Sect. 5.

2 Preliminaries

In this section, we briefly recall some basic notions of Dempster-Shafer theory and Parzen Rosenblatt window density estimation.

2.1 Dempster-Shafer Theory

Let $\Omega = \{\omega_1, \dots, \omega_K\}$, and let $\mathcal{P}(\Omega) = \{A_1, \dots, A_Q\}$ be its power set, with $Q = 2^K$. A function M defined from $\mathcal{P}(\Omega)$ to $[0, 1]$ is called a “basic belief assignment” (*bba*) if $M(\emptyset) = 0$ and $\sum_{A \in \mathcal{P}(\Omega)} M(A) = 1$. A *bba* M defines then a “plausibility” function Pl from $\mathcal{P}(\Omega)$ to $[0, 1]$ by $Pl(A) = \sum_{A \cap B \neq \emptyset} M(B)$,

and a “credibility” function Cr from $\mathcal{P}(\Omega)$ to $[0, 1]$ by $Cr(A) = \sum_{B \subset A} M(B)$. Also, both aforementioned functions are linked by $Pl(A) + Cr(A^c) = 1$. Furthermore, a probability function p can be considered as a particular case for which $Pl = Cr = p$.

When two *bbas* M_1 and M_2 describe two pieces of evidence, we can fuse them using the so called “Dempster-Shafer fusion” (DS fusion), which gives $M = M_1 \oplus M_2$ defined by:

$$M(A) = (M_1 \oplus M_2)(A) \propto \sum_{B_1 \cap B_2 = A} M_1(B_1)M_2(B_2) \tag{1}$$

Finally, an evidential *bba* M can be transformed into a probabilistic one using Smets method, according to which each mass of belief $M(A)$ is equally distributed among all elements of A , leading to the so called “pignistic probability”, Bet , given by:

$$Bet(\omega_i) = \sum_{\omega_i \in A \subseteq \Omega} \frac{M(A)}{|A|} \tag{2}$$

where $|A|$ is the number of elements of Ω in A .

2.2 Parzen-Rosenblatt Density Estimation

In statistics, Parzen-Rosenblatt window method [18,19], also termed kernel density estimation, is a fundamental data smoothing where inferences about the population are made, based on a finite data sample. It can be perceived as a non-parametric technique aiming to construct the PDF f , of an unknown shape, associated to a random variable X . Let (x_1, x_2, \dots, x_N) be a sample of realizations of such a random variable. The problem is then to estimate f values at several points of interest. Kernel smoothing is then a generalization of histogram smoothing in which a window, of some predefined form, centred at each point is used to estimate the density value at that point. For this purpose, the following estimator is used: $\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x_i}{h}\right)$, where $K(\cdot)$ is the kernel - a non-negative zero-mean function that integrates to one - and $h > 0$ is a smoothing parameter called the “bandwidth” or “kernel width”. Also, a wide range of kernel functions can be used.

3 Parzen-Rosenblatt Dempster-Shafer Classifier

In this section, we describe the theoretical fundament of the proposed classification approach, which will be called Parzen-Rosenblatt Dempster-Shafer classifier (PR-DS). For this purpose, let us assume we have a sample of N pre-labeled multiattribute data (Z_1, \dots, Z_N) where each datum $Z_n = (X_n, Y_n)$ with $X_n \in \Omega = \{\omega_1, \dots, \omega_K\}$ being the label, and $Y_n = (Y_n^1, \dots, Y_n^P) \in \mathbb{R}^P$ being the P -attribute observation. The problem is then to estimate the label of any new observation $Y_{n'}$ that is optimal with respect to some criterion.

In what follows, we first describe the training process conducted on the pre-labeled data sample (Z_1, \dots, Z_N) . Then, we show how our classifier assigns a new observation $Y_{n'}$ to one of the K labels.

3.1 Training

Let us consider the above pre-labeled multiattribute data (Z_1, \dots, Z_N) . According to our PR-DS scheme, training consists in estimating for each class $\omega_k \in \Omega$ and for each attribute p ($1 < p < P$), the Parzen-Rosenblatt density \hat{f}_k^p as described in Sect. 2. For further weighting sake, 5-fold cross-validation classification is achieved based on each attribute (taken alone) using the above Parzen-Rosenblatt PDFs according to maximum likelihood. F-measure is then evaluated per each attribute p and per each hypothesis $A_q \in \mathcal{P}(\Omega)$. To this end, the $K \times K$ confusion matrix is converted to a 2×2 one including only A_q and $\Omega \setminus A_q$. Then, the F-measure value obtained is itself the weighting factor α_q^p .

3.2 Supervised Classification

For a given new observation $Y_{n'}$, partial report about the identity of $X_{n'}$ can be made at each individual attribute level through a mass function M^p , on $P(\Omega)$, generated based on the Parzen-Rosenblatt PDF estimated at the training stage. Such reports are then combined to reap a consensus report M . Final decision is then be deduced through the Pignistic transform applied to M . In the following, we describe our approach step by step.

Step 1: Generation of Mass Functions. To define the mass associated to attribute p , let us consider the rank function δ_p defined from $\{1, \dots, K\}$ to Ω such as $\delta_p(k)$ is the k -ranked element of Ω in terms of \hat{f}^p , i.e. $\hat{f}_{\delta_p(1)}^p(Y_{n'}^p) \leq \hat{f}_{\delta_p(2)}^p(Y_{n'}^p) \leq \dots \leq \hat{f}_{\delta_p(K)}^p(Y_{n'}^p)$. Then, \mathcal{M}^p is derived as follows:

$$\begin{cases} \mathcal{M}^p(\Omega) \propto K \hat{f}_{\delta_p(1)}^p(Y_{n'}^p) \\ \mathcal{M}^p(\{\omega_{\delta_p(k)}, \dots, \omega_{\delta_p(K)}\}) \propto (K - k + 1) [\hat{f}_{\delta_p(k)}^p(Y_{n'}^p) - \hat{f}_{\delta_p(k-1)}^p(Y_{n'}^p)], \text{ for } k > 1 \end{cases} \quad (3)$$

Step 2: Weighting of Mass Functions. To adjust the definitive mass associated to attribute p , we apply a weakening based on the weight α_q^p associated to each hypothesis $A_q \in P(\Omega)$ as follows:

$$\begin{cases} M^p(A_q) = \alpha_q^p \mathcal{M}^p(A_q), \text{ for } A_q \subsetneq \Omega \\ M^p(\Omega) = 1 - \sum_{A_q \subsetneq \Omega} M^p(A_q) \end{cases} \quad (4)$$

Step 3: Combination of Mass Functions. Mass functions associated to different attributes are then combined into one collaborative mass $M = \bigoplus_{p=1}^P M^p$:

$$M(B) \propto \sum_{\bigcap_{p=1}^P B_p=B} \left[\prod_{p=1}^P M^p(B_p) \right], \text{ for } B, B_p \in \mathcal{P}(\Omega) \quad (5)$$

Step 4: Decision Making. Based on M , the final decision is then taken according to the Pignistic transform:

$$\hat{X}_{n'} = \arg \max_{\omega_k} \sum_{A \ni \omega_k} \frac{M(A)}{|A|} \quad (6)$$

Remark 1: If for some datum Z_n , observation at attribute p : Y_n^p is missing, this situation is handled by defining the mass function M^p as follows:

$$\begin{cases} M^p(A) = 0; \text{ for } A \in \mathcal{P}(\Omega) - \{\Omega\} \\ M^p(\Omega) = 1 \end{cases} \quad (7)$$

Remark 2: In addition to missing data handling, our classifier benefits from all other advantages of Dempster-Shafer theory and can handle situations where information are unreliable or uncertain.

Remark 3: Mass generation through (3) is conceived in such a manner that the Pignistic transform applied to the generated mass be proportional to the original Parzen-Rosenblatt density values. Formula (3) can thus be perceived as an intuitive reverse-Pignistic transform.

3.3 Unsupervised Classification

Our classifier can also be applied in the unsupervised context where no labeled data are available for training. To this end, training is applied to the whole data Y considering an initial coarse classification which is then updated iteratively until an end criterion is reached. More explicitly, the unsupervised classification runs as follows:

1. Perform a clustering \hat{X}^0 of Y (using K -means for instance);
2. $i \leftarrow 0$;
3. Derive Parzen-Rosenblatt PDFs from (Y, \hat{X}^i) ;
4. Infer \hat{X}^{i+1} using steps 1, 3 and 4 of the supervised context;
5. if end criterion is not reached: $i \leftarrow i + 1$ and go to 3;

4 Experiments

In this section, we assess the performance of the proposed PR-DS approach with respect to eleven state-of-the-art approaches: naive Bayes classifier (NBC) [12],

Table 1. Description of experimental datasets

Dataset	Items	Classes	Attributes	Missing data
Iris	150	3	4	No
Heart	270	2	13	No
Wine	178	3	13	No
Australian	69	2	14	Yes
Hepatitis	155	2	19	Yes
Sonar	208	2	60	No

linear Bayes normal classifier (LDC) [16], K nearest neighbors classifier (K-NNC) [7], nearest mean classifier (NMC) [25], quadratic discriminant classifier (QDC), support vector machine (SVM, [6]), random forests (RF [5]) and four DS-based approaches: normal distribution-based classifier (NDBC) [28], K-nearest neighbor D-S theory (KNN-DST) [8], evidential calibration (EC) [29] and weighted fuzzy Dempster-Shafer framework (WFDSF) [17]. To perform our comparative analysis, we consider six datasets from the universal UCI machine learning repository [15]: Iris, Heart, Wine, Australian, Hepatitis and Sonar. The characteristics of these datasets are provided in Table 1.

Classification performance of each method will be assessed in terms of overall accuracy. More explicitly, a five-fold cross validation is applied to each method per each dataset. The same process is repeated 100 times. The average of such runs is then used for comparison. To produce results associated to state-of-the-art methods, we acknowledge the use of Waikato Environment for Knowledge Analysis (WEKA) [27] and Matlab Pattern Recognition toolbox (PRTools) [10]. As for the four DS-based approaches, we have adopted results reported in [17].

Training process of our proposed PR-DS classifier is conducted considering a set of kernel functions. Then, the kernel exhibiting the best performance is selected per each dataset. Also, the unsupervised version of our PR-DS (denoted U-PR-DS) classifier has been compared to two clustering methods: K-means (KM) and Fuzzy C-means (FCM). The end criterion considered here is the convergence of an objective function Z defined as in Fuzzy C-means with fuzzy membership degrees replaced by Pignistic probabilities. The results obtained are illustrated in Table 2.

Table 2. Classification rates of different classifiers on multimodal benchmarks

Dataset	NBC	LDC	K-NNC	NMC	QDC	KNN-DST	NDBC	EC	WFDSF	SVM	RF	PR-DS	KM	FCM	U-PR-DS
Iris	95.48	97.94	96.16	92.34	97.39	95.33	94.00	94.67	96.00	97.65	94.72	96.23	89.33	89.33	89.33
Heart	84.13	83.47	65.65	64.42	81.99	76.30	82.59	83.70	85.56	83.07	82.47	80.42	57.57	58.24	60.60
Wine	97.29	98.65	72.39	72.39	98.91	93.84	96.63	97.17	98.32	95.65	97.70	97.31	70.22	68.53	92.70
Australian	77.19	85.97	68.64	64.93	79.72	78.41	80.01	80.60	85.20	85.29	86.95	82.25	55.94	56.08	55.94
Hepatitis	83.54	84.96	78.60	64.25	82.21	80.57	79.40	79.88	83.85	84.50	84.83	86.58	74.19	69.67	79.35
Sonar	68.19	74.11	80.99	66.28	75.21	79.81	72.57	68.26	77.02	77.48	83.17	76.61	55.28	55.28	55.77
Means	84.30	87.52	77.07	70.77	85.90	84.04	84.20	84.04	87.65	87.27	88.31	86.57	67.09	66.19	72.28

Overall, the results provided by our classifier are competitive compared to the state-of-the-art ones. Among evidential approaches, our PR-DS classifier is only outperformed by WFDSF which is based on the combination of more than one DS-based approach [17].

On the other hand, our PR-DS classifier yields the best performance for dataset Hepatitis. In the unsupervised context, the PR-DS performs significantly better than the well-known K-means and Fuzzy *c*-means.

To show how important is step 2 of our PR-DS classifier in the supervised context, let us consider the dataset Australian. During training, we have noticed that attribute 14 is rather misleading. Indeed, applying our PR-DS scheme without source weighting of step 2 yields 68.29% of accuracy. Doing the same while ignoring attribute 14 leads to an accuracy of 83.52%. Applying the weakening of step 2 reduces the impact of attribute 14 leading to an accuracy of 82.25%.

5 Conclusion

In this paper, we introduced a novel approach for multiattribute data classification. The proposed approach is based on Parzen-Rosenblatt density estimation for exclusive and compound hypotheses in accordance with Dempster-Shafer theory. Final classification decisions are then inferred via Pignistic probabilities. The novelty of our classifier with respect to other ones using similar architectures relies in (i) considering more flexible likelihood densities which allows to consider non-Gaussian distributions; and (ii) adopting a new mass generation scheme from such densities. Our approach has been assessed against state-of-the-art methods through experiments achieved on standard multimodal data benchmarks. An interesting future direction would be to exploit the unsupervised version of our classifier within evidential Markov models to improve their performance. Another interesting extension would be to consider other combination rules than Dempster's one and further investigate other weighting mechanisms.

References

1. Bendjebbour, A., Delignon, Y., Fouque, L., Samson, V., Pieczynski, W.: Multi-sensor image segmentation using Dempster-Shafer fusion in Markov fields context. *IEEE Trans. Geosci. Remote Sens.* **39**(8), 1789–1798 (2001)
2. Bloch, I.: Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recogn. Lett.* **17**(8), 905–919 (1996)
3. Boudaren, M.E.Y., An, L., Pieczynski, W.: Dempster-Shafer fusion of evidential pairwise Markov fields. *Int. J. Approximate Reasoning* **74**, 13–29 (2016)
4. Bowman, A.W., Azzalini, A.: *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations*, vol. 18. OUP, Oxford (1997)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 27 (2011)

7. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967)
8. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (1995)
9. Denœux, T.: 40 years of Dempster-Shafer theory. *Int. J. Approximate Reasoning* **79**, 1–6 (2016)
10. Duin, R., Juszczak, P., Paclik, P., Pekalska, E., De Ridder, D., Tax, D., Verzakov, S.: A matlab toolbox for pattern recognition. *PRTools Version 3*, 109–111 (2000)
11. Guo, H., Shi, W., Deng, Y.: Evaluating sensor reliability in classification problems based on evidence theory. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **36**(5), 970–981 (2006)
12. Hu, B.G.: What are the differences between Bayesian classifiers and mutual-information classifiers? *IEEE Trans. Neural Netw. Learn. Syst.* **25**(2), 249–264 (2014)
13. Jones, R.W., Lowe, A., Harrison, M.J.: A framework for intelligent medical diagnosis using the theory of evidence. *Knowl. Based Syst.* **15**(1), 77–84 (2002)
14. Le Hegarat-Masclé, S., Bloch, I., Vidal-Madjar, D.: Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Trans. Geosci. Remote Sens.* **35**(4), 1018–1031 (1997)
15. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
16. Liu, C., Wechsler, H.: Robust coding schemes for indexing and retrieval from large face databases. *IEEE Trans. Image Process.* **9**(1), 132–137 (2000)
17. Liu, Y.T., Pal, N.R., Marathe, A.R., Lin, C.T.: Weighted fuzzy Dempster-Shafer framework for multimodal information integration. *IEEE Trans. Fuzzy Syst.* **26**(1), 338–352 (2018)
18. Parzen, E.: On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**(3), 1065–1076 (1962)
19. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27**(3), 832–837 (1956)
20. Salzenstein, F., Boudraa, A.O.: Unsupervised multisensor data fusion approach. In: *Sixth International Symposium on Signal Processing and its Applications*, vol. 1, pp. 152–155. IEEE (2001)
21. Shafer, G.: *A Mathematical Theory of Evidence*, vol. 1. Princeton University Press, Princeton (1976)
22. Shafer, G.: A mathematical theory of evidence turns 40. *Int. J. Approximate Reasoning* **79**, 7–25 (2016)
23. Shafer, G.: The problem of dependent evidence. *Int. J. Approximate Reasoning* **79**, 41–44 (2016)
24. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**(2), 191–234 (1994)
25. Veenman, C.J., Reinders, M.J.: The nearest subclass classifier: a compromise between the nearest mean and nearest neighbor classifier. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(9), 1417–1429 (2005)
26. Wand, M.P., Jones, M.C.: *Kernel Smoothing*. CRC Press, London (1994)
27. Wi, H., Eibe, F.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman (2011)
28. Xu, P., Deng, Y., Su, X., Mahadevan, S.: A new method to determine basic probability assignment from training data. *Knowl. Based Syst.* **46**, 69–80 (2013)

29. Xu, P., Davoine, F., Zha, H., Denoeux, T.: Evidential calibration of binary SVM classifiers. *Int. J. Approximate Reasoning* **72**, 55–70 (2016)
30. Zhu, Y.M., Bentabet, L., Dupuis, O., Babot, D., Rombaut, M.: Automatic determination of mass functions in Dempster-Shafer theory using fuzzy c-means and spatial neighborhood information for image segmentation. *Opt. Eng.* **41**(4), 760–770 (2002)



Birnbaum's Importance Measure Extended for Non-coherent Systems

Ayyoub Imakhlaf^(✉) and Mohamed Sallak

Sorbonne universités, Université de technologie de Compiègne,
Heudiasyc UMR 7253, CS 60 319, 60 203, Compiègne cedex, France
ayyoub.imakhlaf@hds.utc.fr, mohamed.sallak@uni-heidelberg.de

Abstract. We introduce an extended Birnbaum component importance measure considering epistemic and aleatory uncertainty adapted to non-coherent systems. The belief function theory is proposed as a framework for taking into account both types of uncertainty. The objective is to rank components according to their importance in system working. This importance measure was introduced for coherent systems; however, the increasing complexity of modern systems introduces the case of non-coherent systems. This is why we should consider these kinds of systems. In this work, we propose a method to compute the importance measure of the components of non-coherent systems in the framework of belief functions theory.

Keywords: Birnbaum importance measure · Reliability analysis
Non-coherent systems · Belief function theory

1 Introduction

The Birnbaum importance measure was developed by Birnbaum (1968) [Bir68]. It was introduced for coherent systems where its computation is relatively straightforward. It could be interpreted as the rate at which the system reliability function increases as the reliability of the component increases. It could also be interpreted as the difference between the conditional probability that the system works knowing the component works, and the conditional probability that the systems works knowing that the component fails. However, increasing complexity of modern systems introduces the case of non-coherent systems. These systems are defined as systems that not satisfying at least one of the coherency condition: the monotony of the structure function of the system and the relevancy of its components. Therefore, in this kind of systems the failure state of a component could be as important as its working state. Andrews [And00] demonstrated that in the case of multi-tasking systems the non-occurrence could be important for the occurrence of the top event. In [CCR08], the authors listed several non-coherent systems. Then, it became relevant to consider these systems and needs to then adequately. Thus, it is required to extend Birnbaum's

measure for non-coherent systems, all the while maintaining a straightforward computation.

The first extension of Birnbaum importance measure was proposed in 1983 [Jac83]. However, Andrews and Beeson [AB03] showed, on one hand, that this extension ranks components incorrectly. Andrews and Beeson retake the example proposed by Jackson in [Jac83] and showed that the most important component is not the one obtained by the method of Jackson. On the other hand, they propose a new extension in which they consider separately the contribution of component working state to the system working state, and the contribution of component failing state to the system working state. The authors also noted that, due to the consensus terms, this extension leads to several results because of reliability functions could be syntactically different even if it is algebraically equivalent. Recently, Aliee *et al.* [ABGT17] defined the criticality indicator variable of component C_i , and showed that, if the consensus terms are explicitly included in the Boolean structure function, then the criticality indicator variable of a component C_i is equal to the Andrews's extension. They showed also that it is equal to Birnbaum importance measure when the system is coherent.

In this paper, the importance measure proposed in [ABGT17] is extended to take into account aleatory and epistemic uncertainty in the framework of belief functions theory. The rest of this paper is organized as follows: Sect. 2 introduces the theory of belief function. Section 3 reviews definitions and basic concept in reliability assessment. Section 4 presents an algorithm to compute components' importance measure under epistemic uncertainty for non-coherent systems. Finally, Sect. 5 concludes the paper.

2 Belief Functions Theory

The publication of the work of Dempster [Dem68] on upper and lower probabilities as well as that of Shafer [SB79] on the theory of evidence describe what is commonly called Dempster-Shafer theory. Afterwards, Smets [Sme92] reinterpreted Shafer's work and introduced belief functions theory.

This theory is a framework that enables the experts to represent and manipulate epistemic and aleatory uncertainties. It is a generalization of the probability theory as it assigns probability to subsets instead of singletons.

2.1 Belief, Plausibility, and Mass Function

A finite set Ω of mutually exclusive elements is called a frame of discernment. A subset $A \in \mathcal{P}(\Omega)$ is called a proposition, where $\mathcal{P}(\Omega)$ is the power set of Ω . The mass function, denoted m , is defined as a mapping from $\mathcal{P}(\Omega)$ in $[0, 1]$. It assigns a mass value between 0 and 1 to each proposition A of $\mathcal{P}(\Omega)$, such that

$$\sum_{A \in \mathcal{P}(\Omega)} m(A) = 1 \quad (1)$$

The mass of a subset A is interpreted as the degree of belief assigned to the hypothesis that the truth lies in the proposition A . Such an assignment implies total ignorance about the belief over all subsets of A . Every subset A such that $m(A) > 0$ is called a focal set. The two important measures of uncertainty provided by belief function theory are called belief function, and the plausibility function. They are defined respectively by:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$$

The interval $[Bel(A), Pl(A)]$ represents the uncertainty about the state of A .

2.2 Combination, Marginalization, and Vacuous Extension

In order to perform inference operations, mass functions representing different pieces of evidence need to be combined. Combination rules are an important part of belief functions theory. Several types of combination rules within the framework of belief functions [SF02]. In this section, we present the conjunctive combination rule which is the rule used in this work.

The conjunctive combination rule, denoted \odot , allows one to combine masses that are defined over the same frame of discernment and are induced by distinct bodies of evidence. The new mass obtained reflects a conjunctive combination of the underlying evidence. The conjunctive rule corresponds to an AND operation.

More formally, let m_1 and m_2 be two mass functions defined on the same frame of discernment, and induced by distinct pieces of evidence. The mass function $m_{1\odot 2} = m_1 \odot m_2$ obtained using the conjunctive rule \odot is defined as follows:

$$m_{1\odot 2}(A) = \sum_{B \cap C = A} m_1(B)m_2(C), \quad \forall A \subseteq \Omega \tag{2}$$

The marginalization to Ω_i of a mass $m^{\Omega_i \times \Omega_j}$, defined on the Cartesian product $\Omega_i \times \Omega_j$, is defined as follows, $\forall A \subseteq \Omega_i$:

$$m^{\Omega_i \times \Omega_j \downarrow \Omega_j}(A) = \sum_{\substack{B \subseteq \Omega_i \times \Omega_j, \\ Proj(B \downarrow \Omega_i) = A}} m^{\Omega_i \times \Omega_j}(B) \tag{3}$$

where $Proj(B \downarrow \Omega_i) = \{a \in \Omega_i \mid \exists b \in \Omega_j, (a, b) \in B\}$.

The inverse operation is called vacuous extension. It is done from Ω_i to $\Omega_i \times \Omega_j$ as follow:

$$m^{\Omega_i \uparrow \Omega_i \times \Omega_j}(A) = \begin{cases} m^{\Omega_i}(A) & \text{if } A = B \times \Omega_j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

3 Reliability Assessment

In a system whose working state depends on the working state of its components, some of these components contribute to system working state more than others. Birnbaum was the first who introduce a quantitative definition of this concept of importance [Bir68].

3.1 Basic Concept

For a system S with n components we assume that each component C_i can be in one and only one of two states: a working state or a failing state. As the state of a component C_i is aleatory, hence the state of each component C_i is described by binary random variable X_i that take values in $\Omega_i = \{0, 1\}$. X_i is a Bernoulli distributed random variable with success probability r_i . $r_i = P(X_i = 1)$ is called the reliability of the component C_i .

Additionally, we define the product space Θ as the Cartesian product of components' frames of discernments, that is, $\Theta = \times_{i=1}^n \Omega_i$. There are 2^n possible realizations of Θ that are called elementary event and denoted by X . They correspond to every possible combination of the states taken by components. In a similar way, the system S can be in one and only one of two states: a working state 1_S or a failing state 0_S . The state of the system is described by a binary random variable X_S that take values in $\Omega_S = \{0, 1\}$. We further define the reliability function of the system, defined from $[0, 1]^n$ to $[0, 1]$, as $R_S(r) = P(X_S = 1)$.

Definition 1. The state of the system S is directly related to the states of its components C_i . The relation between the state of the system X_S and the elementary event X is given by a binary function φ defined from $\{0, 1\}^n$ to $0, 1$. This function is called the structure function of the system:

$$X_S = \varphi(X) = \begin{cases} 1 & \text{if } S \text{ works} \\ 0 & \text{if } S \text{ fails} \end{cases}$$

Definition 2. A system is called monotone if its structure function φ is increasing i.e.

$$\forall i \in \{1, \dots, n\}, \forall X \in \Theta, \varphi(1, X_{-i}) \geq \varphi(0, X_{-i})$$

On another word, a system is monotone if its state cannot be improved when a component fails. It means that if a system does not work it will not work when a component fails.

Definition 3. A component C_i is called relevant if

$$\exists X \in \Theta, \varphi(1, X_{-i}) \neq \varphi(0, X_{-i})$$

on other words, a component is relevant if there is at least one configuration in which the state of the system is different according to the state of the component C_i .

Definition 4. A system is coherent if it is monotone and all its components are relevant. A system is non-coherent if it does not satisfy at least one of the coherency conditions.

Example. Let us consider an anomaly detection system. The system consists of a component C_1 that is supposed to activate a subsystem A when it works. However, when C_1 fails the system initiate an emergency procedure consisting on a subsystem B . We consider that the system works if the subsystem A and the component C_1 work or if C_1 fails and the subsystem B . According to the previous explanation the Boolean structure function is given by the following equation:

$$\varphi(X_1, X_A, X_B) = (X_1 \wedge X_A) \vee (\overline{X_1} \wedge X_B)$$

The example we propose is just an hypothetical one, a realistic examples is proposed in [ZM87]. This system is not small (19 components) this is why we do not use it in this paper.

Birnbaum Importance Measure. The quantitative definition introduced by Birnbaum can just be applied to coherent and not repairable system. In order to numerically rank the contribution components, the Birnbaum importance measure quantifies the contribution of each component between 0 and 1 : 0 signifies the lowest level of importance.

The Birnbaum importance measure of a component C_i is calculated from the reliability function of the system R_S as follow:

$$B_i(r_{-i}) = \frac{\partial R_S}{\partial r_i}(r) \quad (5)$$

$B_i(r_{-i})$ can be interpreted as the rate at which the system reliability increases as the reliability of component C_i increases.

Birnbaum Importance Measure Extended for Non-coherent System.

Andrews and Beeson [AB03] generalized the Birnbaum importance measure to both coherent and non-coherent systems by considering the contribution of component working state and component failing state separately. Recently, Aliee *et al.* [ABGT17] introduced a Boolean expression to represent the notion that the component C_i is critical. It is called the criticality indicator variable of component C_i , it is noted by Ψ_i^{ABGT} , and it is defined as follow:

$$\Psi_i^{ABGT}(X_{-i}) = [\varphi(1, X_{-i}) \wedge \overline{\varphi(0, X_{-i})}] \vee [\overline{\varphi(1, X_{-i})} \wedge \varphi(0, X_{-i})] \quad (6)$$

where $X_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.

The expression of Ψ_i^{ABGT} quantify the criticality of a component C_i either when it is in a failing state or in a working state i.e. The authors showed, on one hand, that if the component C_i is failure-critical the first term in (6) is equal to 1 and the second to 0. On the other hand, if C_i is repair-critical, then the first term in (6) is equal to 0 and the second to 1.

Therefore, the authors defined the importance measure I_i^{ABGT} as the probability that the criticality indicator variable of component C_i is equal to 1, i.e.

$$I_i^{ABGT}(r_{-i}) = P(\Psi_i^{ABGT}(X_{-i}) = 1) \tag{7}$$

where $r_{-i} = (r_1, \dots, r_{i-1}, r_{i+1}, \dots, r_n)$.

In other words, I_i^{ABGT} is the probability that the component C_i is failure-critical or repair critical.

They showed that I^{ABGT} is equal to the Birnbaum importance measure for coherent systems. They also showed that is equal to Andrews and Beeson's extension.

Example. let us retake the example given previously, and suppose that the reliability of C_1 is $r_1 = 0.99$, the reliability of A is $r_A = 0.95$, and the reliability of B is $r_B = 0.97$. Then,

$$\begin{aligned} I_1^{ABGT}(r_{-1}) &= P((X_A \wedge \bar{X}_B) \vee (\bar{X}_A \wedge X_B) = 1) = r_A + r_B - 2r_A r_B = 0,077 \\ I_A^{ABGT}(r_{-A}) &= P(X_1 = 1) = 0,99 \\ I_B^{ABGT}(r_{-B}) &= P(\bar{X}_1 = 1) = 1 - r_1 = 0,01 \end{aligned}$$

Hence, according to this result A is the most important one, which is not surprising because C_1 is highly reliable.

3.2 Experts Assessment

To compute the importance of a component C_i , we have to focus, on one hand, on the other components C_j ($j = 1, \dots, i - 1, i + 1, \dots, n$) that have an impact on the working state of S and, on the other hand, on the criticality indicator variable of C_i . Then, we need to know the reliability of each component to compute I_i^{ABGT} . For some components, we can only ask experts to give us their reliability, according to their experiences. Particularly if these components are too expensive or it is impractical to be observed them directly. In this case, experts should take into account uncertainty and the sensitivity of their assessment.

The belief functions theory helps experts by giving them the opportunity to quantify more adequately their uncertainties. In the context of this work, when we ask experts to assess components reliability, they express their beliefs by defining a mass function m_i for each component C_i . Hence, knowing that $\forall i = 1, \dots, n, m_i(\emptyset) = 0$, the expert has to assess $m_i(1)$ (respectively $m_i(0) =$ and $m_i(\Omega_i)$) which represents the degree of belief on the occurrence of C_i (respectively the degree of belief on the non-occurrence of C_i , and the ignorance about the state C_i).

Equation (1) allows experts to give only their beliefs about two focal sets among three, the belief about the third set is deduced from the two others. Therefore, they are free to choose one of the following combinations:

1. $m_i(1)$ and $m_i(0)$.
2. $m_i(1)$ and $m_i(\Omega_i)$.
3. $m_i(0)$ and $m_i(\Omega_i)$.

We can ask several experts to assess the mass function of the same component. The paper [SF02] presents different combination rules to aggregate the masses given by experts into a single one. This point is beyond scope of this paper.

On the other hand, we need to construct a mass function for Ψ_i^{ABGT} , denoted by $m_{\Psi_i^{ABGT}}$. As this structure is perfectly known, then the mass is categorical, i.e. it has only one focal set. This focal set is given by:

$$A = \{(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n, X_S) \in \Theta \times \Omega_S, \Psi_i^{ABGT}(X_{-i}) = X_S\} \quad (8)$$

where X_i is equal to 1 if C_i works, and it is equal to 0 if it fails.

4 Components' Importance Under Epistemic Uncertainty

Given several pieces of evidence, we have to capitalize on these pieces of evidence to construct our belief on the working state of S . Aguirre *et al.* proposed an algorithm to compute the reliability of a system using components mass functions and the configuration mass in [ASS15]. The purpose is to assert an uncertain measure about the reliability of the system given by a mass function m_S defined from Ω_S in $[0, 1]$. This algorithm is used to compute the importance measure of a component using (6). The fundamental steps are given in the following and summarizing in Fig. 1:

1. The mass function m_j of a component C_j has to be extended to the same space in which $m_{\Psi_i^{ABGT}}$ is defined (c.f. Eq. (4)).

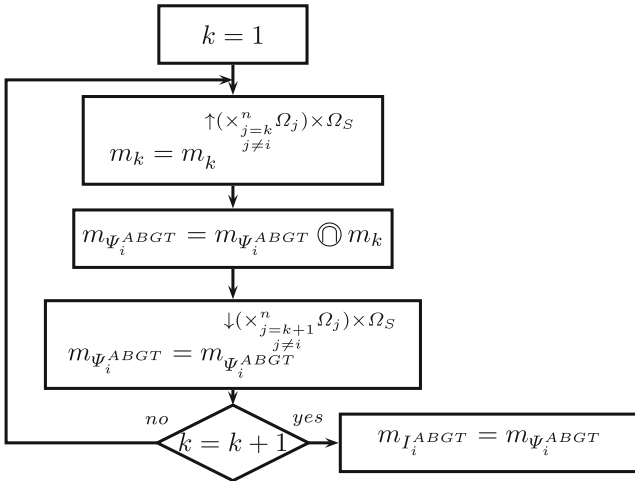


Fig. 1. Algorithm to compute $m_{I_i^{ABGT}}$

2. The extended mass function m_j of a component C_j has to be combined with $m_{\Psi_i^{ABGT}}$.
3. The mass obtained previously is marginalized to $\Omega_1 \times \dots \times \Omega_{i-1} \times \Omega_{i+1} \times \dots \times \Omega_n \times \Omega_S \setminus \Omega_j$ (c.f. Eq. (3)).
4. The first, the second, and the third steps are repeated, using the mass function obtained in the third step instead of $m_{\Psi_i^{ABGT}}$, until the obtaining of a mass function defined on Ω_S . This mass function represents the mass $m_{I_i^{ABGT}}$.

This algorithm is justified by the axioms of local computations [SS08]. These axioms indicate, *inter alia*, that the mass combination can be done in any order. However, this algorithm is impractical, because the number of operations grows exponentially according to the size of the system. Thus, the authors showed that the reliability of a coherent system can be easily obtained using belief and plausibility functions considered, respectively, as components reliability lower and upper bounds. This is due to the monotony of the structure function in the case of coherent system. However, in the case of non-coherent systems the brute approach, described previously, should be used.

5 Conclusion

In this work, we have recalled basic concept of reliability analysis in the framework of belief functions theory. This paper focuses on the concept of importance measures, especially those adapted to non-coherent systems. We first studied importance measures based on the criticality indicator variable. Then, we have tried to extended them to make them compatible with belief functions theory in order to take into account aleatory and epistemic uncertainty. For this purpose, we have proposed an algorithm to compute the importance measure of components of non-coherent systems. This paper is a preliminary work, it needs to be more formalized. On the other hand, the method proposed is considered to be a brute force approach. Thus, it has to be optimized to make it useful for reliability researchers.

References

- [AB03] Andrews, J.D., Beeson, S.: Birnbaum's measure of component importance for noncoherent systems. *IEEE Trans. Reliab.* **52**(2), 213–219 (2003)
- [ABGT17] Aliee, H., Borgonovo, E., Glaß, M., Teich, J.: On the boolean extension of the birnbaum importance to non-coherent systems. *Reliab. Eng. Syst. Saf.* **160**, 191–200 (2017)
- [And00] Andrews, J.D.: To not or not to not!! In: *Proceedings of the 18th International System Safety Conference, USA* (2000)
- [ASS15] Aguirre, M.F., Sallak, M., Schon, W.: An efficient method for reliability analysis of systems under epistemic uncertainty using belief function theory. *IEEE Trans. Reliab.* **64**(3), 893–909 (2015)
- [Bir68] Birnbaum, Z.W.: On the importance of different components in a multicomponent system. Technical report, Washington Univ Seattle Lab of Statistical Research (1968)

- [CCR08] Contini, S., Cojazzi, G.G.M., Renda, G.: On the use of non-coherent fault trees in safety and security studies. *Reliab. Eng. Syst. Saf.* **93**(12), 1886–1895 (2008)
- [Dem68] Dempster, A.P.: Upper and lower probabilities generated by a random closed interval. *Ann. Math. Stat.* **39**(3), 957–966 (1968)
- [Jac83] Jackson, P.S.: On the s-importance of elements and prime implicants of non-coherent systems. *IEEE Trans. Reliab.* **32**(1), 21–25 (1983)
- [SB79] Shafer, G., Breipohl, A.M.: Reliability described by belief functions. In: *Annual Reliability and Maintainability Symposium*, Washington, D.C., pp. 23–27 (1979)
- [SF02] Sentz, K., Ferson, S.: Combination of evidence in Dempster-Shafer theory. SAND 2002-0835 Technical report. Sandia National Laboratories, USA (2002)
- [Sme92] Smets, P.: The transferable belief model and random sets. *Int. J. Intell. Syst.* **7**(1), 37–46 (1992)
- [SS08] Shenoy, P.P., Shafer, G.: Axioms for probability and belief-function propagation. In: Yager, R.R., Liu, L. (eds.) *Classic Works of the Dempster-Shafer Theory of Belief Functions*, vol. 219, pp. 499–528. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-44792-4_20
- [ZM87] Zhang, Q., Mei, Q.: Reliability analysis for a real non-coherent system. *IEEE Trans. Reliab.* **36**(4), 436–439 (1987)



Evidential Independence Maximization on Twitter Network

Siwar Jendoubi^{1,2(✉)}, Mouna Chebbah³, and Arnaud Martin⁴

¹ LISTIC, University Savoie Mont Blanc, 74944 Annecy-le-Vieux, France

siwar.jendoubi@univ-smb.fr

² LARODEC, University of Tunis, ISG Tunis, 2000 Le Bardo, Tunisia

³ LARODEC, Univ. Manouba, ESEN, Manouba, Tunisia

mouna.chebbah@esen.tn

⁴ Univ Rennes 1, CNRS, IRISA, Lannion, France

Arnaud.Martin@univ-rennes1.fr

Abstract. Detecting independent users in online social networks is an interesting research issue. In fact, independent users cannot generally be influenced, they are independent in their choices and decisions. Independent users may attract other users and make them adopt their point of view. A user is qualified as independent when his/her point of view does not depend on others ideas. Thus, the behavior of such a user is independent from other behaviors. Detecting independent users is interesting because a part of them can be influencers. Independent users that are not influencers can be directly targeted as they cannot be influenced. In this paper, we present an evidential independence maximization approach for Twitter users. The proposed approach is based on three metrics reflecting users behaviors. We propose a useful approach for detecting influencers. Indeed, we consider the independence as a characteristic of influencers even if not all independent users are influencers. The proposed approach is experimented on real data crawled from Twitter.

Keywords: Independence measure · Independence maximization
Theory of belief functions · Twitter social network · Influence

1 Introduction

Nowadays, most of web users are connected over *online social networks (OSN)* like Facebook, Twitter, LinkedIn, *etc.* OSN Users are different and may have distinguishable characteristics. Some of them are active and others are passive. Some of them are dependent on others, thus their choices, points of views and ideas depend on others. Other users are independent and impose their own choices and points of view. These users are independent from others and may be influencing them. Therefore, in this paper, we assume that the independence is a characteristic of influence users in the network. However, we cannot consider all independent users as influencers.

Independent users are more active and they attract others with their activities on OSN. Detecting these users is an interesting task for many companies to promote their business over OSNs. A part of independent users is influencers. Independent users that are not influencers can be directly targeted as they cannot be influenced. OSN provided a wide spread platform to promote new products and services by several companies. To summarize, companies propagate their new products through influencers and may also target independent users who are not targeted otherwise.

Previous researches were already interested in measuring the independence of users in OSN. However, the independence was never studied from the influence point of view. Kudelka *et al.* [4] proposed to quantify the dependence between vertices of an OSN considered as a network in the aim of community detection. Chehibi *et al.* [1] proposed a dependence measure for Twitter. Their proposed approach is detailed in this paper from an independence point of view. Indeed, our independence maximization approach uses their independence measure.

Twitter limits the access to its data, thus we cannot obtain all information about all users. Therefore, we propose an approximate estimation using the *theory of belief functions* [2, 6]. It models uncertainty, imprecision, incompleteness, total and partial ignorance. Besides the theory of belief functions provides a mathematical framework for combination [2, 7]. Recently, the theory of belief functions was used to estimate the influence on Twitter. In fact, Jendoubi *et al.* [3] introduce an evidential influence measure for Twitter. Their measure fuses three Twitter metrics to quantify the user's influence: *followers*, *mentions*, *retweets*.

In this paper, we study the independence in OSN from the influence point of view. Then, we consider the influence of independent users in OSN. In fact, the notions of independence and influence were never studied together in the literature. Then, we propose an evidential independence maximization model for Twitter users. The aim is to detect the most independent users that may be influencers. Indeed, we consider the independence useful to characterize influence users. In addition, this hypothesis validates the independence measure proposed in [1] with regards to the influence maximization. Furthermore, we study the independence of Twitter users through a set of experiments.

The sequel of the paper is organized as follows: We detail the approach of estimating users independence in Sect. 2. Then, we detail the independence maximization model in Sect. 3. Finally, before concluding in Sect. 5, we detail experimental results on real data collected from Twitter in Sect. 4.

2 Independence on Twitter

Twitter is an OSN that allows its users to connect to each others through an explicit relation, *i.e.* *follow* and/or through many implicit relations, *i.e.* a *retweet*, a *mention* or a *citation*.

In this paper, we propose to study the users behavior through the implicit relations. Thus, a retweet is an information tweeted by a user from the tweets of

another user connected with him. The number of retweets reflects the amount that a user adopts opinions of others. A mention is a message directly sent to another specific user to communicate with him. Finally, a citation is the fact that a user cites other users in their tweets.

Thus, retweets, mentions and citations reflect amounts of adoption of others ideas by a specific user. In this paper, we propose to estimate degrees of independence between Twitter users. A user of Twitter u is independent from another user v when information provided by u are not affected by the information produced by v . When a user u is independent from v , the number of times that u retweets, mentions and cites v is quite small.

Therefore, we propose to estimate the independence degrees of Twitter users based on their numbers of follows, retweets, mentions and citations.

A user u of Twitter is independent from another user v if u is following v and u does not frequently retweet tweets of v or/and, u does not frequently mention v in his tweets.

Figure 1 summarizes the approach of user’s independence estimation on Twitter proposed in [1]. The approach is in three steps:

- Step 1. Weights estimation: w define a weight for each implicit relation: retweet, mention and citation. Thus, we define 3 weights (w_r, w_m, w_c), such that w_r is the weight of retweets, w_m is the weight of mentions and w_c is the weight of citations.
- Step 2. Mass functions estimation: a mass function is estimated from each weight. Each mass function reflects the degree of belief on the users independence from the 3 (incomplete) collected information. We define 3 mass functions (m_r, m_m, m_c), such that m_r, m_m and m_c reflect the degree of belief on the users independence knowing the weight of retweets, mentions and citations.
- Step 3. Independence degree estimation: mass functions m_r, m_m and m_c are combined in order to deduce independence degrees by considering the 3 aspects of retweets, mentions and citations.

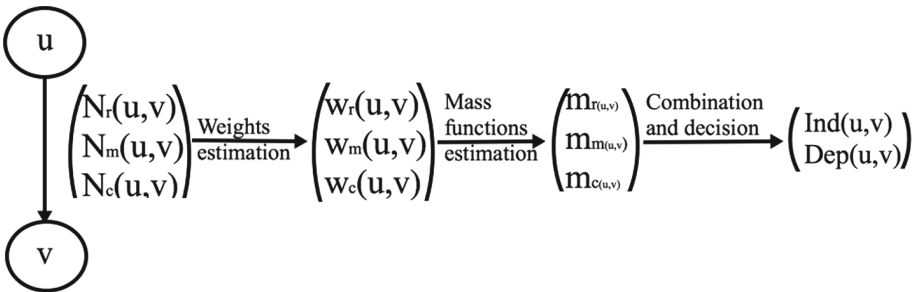


Fig. 1. Independence estimation

2.1 Step 1: Weights Estimation

Let $G = (V, E)$ be an OSN such that V is the set of nodes, E is the set of links, $u \in V$ is a follower of $v \in V$ on Twitter. A user u following a user v can retweet, mention or/and cite v . The number of retweets, mentions or/and citations may indicate the independence or the dependence of u on v . Thus, a vector of weights (w_r, w_m, w_c) is assigned to each link (u, v) as shown in Fig. 1.

The weights w_r , w_m and w_c of the link $(u, v) \in E$ are computed as follows:

$$w_i(u, v) = \frac{N_i(u, v)}{NT_i(u)} \quad (1)$$

such that $i = \{r, m, c\}$. Thus:

1. The retweet weight $w_r(u, v)$, is the number of times that u has retweeted v 's tweets ($N_r(u, v)$) proportioned by the total number of u 's retweet ($NT_r(u)$).
2. The mention weight $w_m(u, v)$, is the number of times that u has mentioned v ($N_m(u, v)$) proportioned by the total number of mentions of u ($NT_m(u)$).
3. The citation weight $w_c(u, v)$, is the number of times that u has cited v ($N_c(u, v)$) proportioned by the total number of u 's citations ($NT_c(u)$).

2.2 Step 2: Mass Functions Estimation

Weights computed in step 1 may induce to some degree of belief on the users independence. Thus, a mass function is built from each weight. Let $\mathcal{I} = \{D, I\}$ be the frame of discernment of the independence where D is the hypothesis that users are dependent and I is the hypothesis that users are independent. Mass functions are estimated as follows:

$$\begin{cases} m_{i(u,v)}^{\mathcal{I}}(\{D\}) = \alpha_{i_u} \times w_i(u, v) \\ m_{i(u,v)}^{\mathcal{I}}(\{I\}) = \alpha_{i_u} \times (1 - w_i(u, v)) \\ m_{i(u,v)}^{\mathcal{I}}(\{D, I\}) = 1 - \alpha_{i_u} \end{cases} \quad (2)$$

such that $i = \{r, m, c\}$. Thus:

1. The mass function $m_{r(u,v)}^{\mathcal{I}}$ is deduced from the retweet weight $w_r(u, v)$. Note that $\alpha_{r_u} = \frac{NT_r(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets T_u . The estimation of the mass function $m_{r(u,v)}^{\mathcal{I}}$ is more reliable when the number of retweets is big enough in comparison with the total number of tweets.
2. The mass function $m_{m(u,v)}^{\mathcal{I}}$ is deduced from the mention weight $w_m(u, v)$ and $\alpha_{m_u} = \frac{NT_m(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets quoted by u with respect to the total number of tweets of u .
3. The mass function $m_{c(u,v)}^{\mathcal{I}}$ is deduced from the citation weight $w_c(u, v)$ and where $\alpha_{c_u} = \frac{NT_c(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets of u mentioning v with respect to the total number of tweets of u .

2.3 Step 3: Independence Degree Estimation

Mass functions $m_{r(u,v)}^{\mathcal{I}}$, $m_{m(u,v)}^{\mathcal{I}}(D)$ and $m_{c(u,v)}^{\mathcal{I}}$ are combined with Dempster's rule of combination as follows:

$$m_{(u,v)}^{\mathcal{I}} = m_{r(u,v)}^{\mathcal{I}} \oplus m_{m(u,v)}^{\mathcal{I}} \oplus m_{c(u,v)}^{\mathcal{I}} \quad (3)$$

Finally, degrees of independence $Ind(u, v)$ and dependence $Dep(u, v)$ corresponds to pignistic probabilities computed from the combined mass function $m_{(u,v)}^{\mathcal{I}}$ such that:

$$\begin{cases} Dep(u, v) = BetP(D) \\ Ind(u, v) = BetP(I) \end{cases} \quad (4)$$

The independence degree $Ind(u, v)$ is non-negative, it is either positive or null. It lies in the interval $[0, 1]$. When $Ind(u, v) = 1$, u is totally independent from v ; $Ind(u, v) = 0$ implies that u is totally dependent of v . Decision is made according to the maximum of pignistic probabilities. If $Dep(u, v) \geq Ind(u, v)$ then u is dependent on v , in the opposite case, if $Ind(u, v) > Dep(u, v)$, u is independent from v .

3 Independence Maximization

The independence measure can be considered as an influence measure. In fact, social influencers are characterized by their independence from the other users. Then, we propose to validate the proposed independence measure by using it to detect influencers, we call this task independence maximization. The maximization of the user's independence in this paper is similar to the problem of influence maximization presented in [3]. In fact, we can maximize the independence through a maximization model that was defined for the influence, we just need to replace the influence measure with an independence measure that has the same mathematical properties which are the monotonicity and the submodularity.

To maximize the independence in the network, we define the amount of independence of a set of nodes, S , on the network. It is the total independence given to S from all users in the network. First, we estimate the independence of S to a user v as follows:

$$Ind(S, v) = \begin{cases} 1 & \text{if } v \in S \\ \sum_{u \in S} \sum_{x \in IN(v) \cup v} Ind(u, x) \times Ind(x, v) & \text{otherwise} \end{cases} \quad (5)$$

where $Ind(v, v) = 1$ and $IN(v)$ is the set of in-neighbors of v , *i.e.* the set of nodes linked to v through a directed link having v as destination. Next, we define the independence spread function that estimates the amount of independence of S on the network as follows:

$$\sigma(S) = \sum_{v \in V} Ind(S, v) \quad (6)$$

We are looking for S on the network that maximizes $\sigma(S)$, *i.e.* $\operatorname{argmax}_S \sigma(S)$.

The independence maximization is an NP-Hard problem. Besides, the function $\sigma(S)$, is monotone and sub-modular. Then, a greedy-based solution can provide a good approximation of the optimal independence users set S . In this case, the cost effective lazy-forward algorithm (CELF) [5] is adapted to maximize the independence. Furthermore, it is a two pass maximization algorithm that is about 700 times faster than the greedy algorithm.

4 Experiments

In our experiments, we crawled the Twitter network using the streaming API on 20/01/2018. We obtained 54960 users, 686542 implicit relations between them (retweet, mention and citation) and 352420 tweets. Next, we used the independent maximization model introduced in the previous section to detect influencers in the collected network. We fixed the number of the detected nodes (size of S) to 100.

We study the independent maximization model according to for criteria of the detected nodes which are the number of accumulated mentions $\#Mention$, the number of accumulated retweets $\#Retweet$, the number of accumulated tweets $\#Tweet$ and the number of accumulated citations $\#Citation$. Indeed, these criteria are considered as quality indicators of detected nodes. Then, higher their values are, better the quality of detected nodes is.

Figure 2 presents the obtained results according to the four fixed criteria, *i.e.* $\#Mention$, $\#Retweet$, $\#Tweet$ and $\#Citation$ receptively. According to Fig. 2, the detected users (horizontal axis) have a good quality especially in terms of $\#Mention$, $\#Retweet$ and $\#Tweet$. In fact, the detected users have more than 4500 accumulated mentions, about 1200 accumulated retweets and more than 1800 accumulated tweets. These observations mean that the detected users are active in the network in terms of tweets. Also, their content is frequently propagated (retweeted). Besides, they are frequently mentioned in others tweets. Whereas, we notice a less important number of citations of the detected users. In fact, we have 18 accumulated citation which is relatively small compared to $\#Mention$, $\#Retweet$ and $\#Tweet$. We think that this is a result of weakness of the proportion of the citations in the data.

These observations confirm the assumption introduced in the previous section, then we can deduce that influencers are characterized by their independence from the other users in the network. In fact, the detected users using the proposed independence maximization model have a good quality according to the chosen criteria which confirms that they are influencers in the network.

In this paper, the main purpose is to validate the independence measure through detecting influencers. In fact, the independence is one important characteristic of influencers. The experiments presented in this section confirm this fact. Indeed, the detected users have a good quality according to the chosen criteria. However, the independence itself is not sufficient as an influence measure

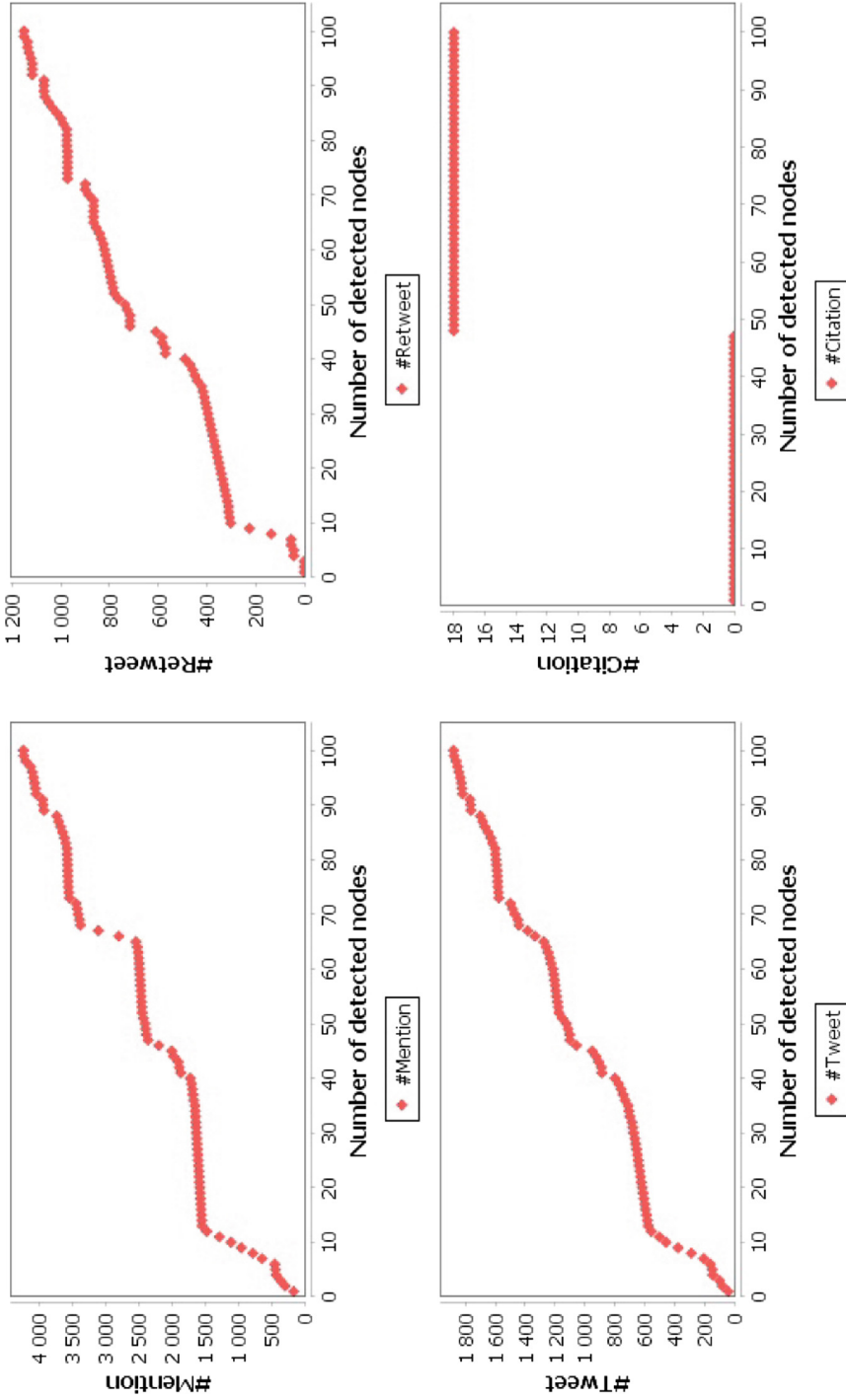


Fig. 2. Detected users using the proposed independence measure according to #Mention, #Retweet, #Tweet and #Citation

and we can obtain better results by fusing it with other influence behaviors in the network like the user's position for example.

5 Conclusion

In this paper, we study the independence of Twitter users proposed in [1] from the influence point of view. Furthermore, we propose an independence maximization model that can be useful to detect influencers. In fact, a common property of social influencers is their independence from the other users in the network. Then, we use an independence measure to estimate the user's influence and to detect a set of influencers that maximizes the global independence in the network. Next, we experiment the proposed solution on real world data collected from Twitter and we study the quality of selected users according to their #Mention, #Retweet, #Tweet and #Citation.

In future works, we will study in a more refined way the notions of influence, dependence and independence to compare them. Besides, we will search to define an influence measure that fuses the user's independence with other influence behaviors like the user's activities and position.

References

1. Chehibi, M., Chebbah, M., Martin, A.: Independence of sources in social networks. In: Medina, J., et al. (eds.) IPMU 2018. CCIS, vol. 853, pp. 418–428. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91473-2_36
2. Dempster, A.P.: Upper and lower probabilities induced by a multiple valued mapping. *Annals Math. Stat.* **38**(2), 325–339 (1967)
3. Jendoubi, S., Martin, A., Liétard, L., Ben Hadji, H., Ben Yaghlane, B.: Two evidential data based models for influence maximization in Twitter. *Knowl. Based Syst.* **121**, 58–70 (2017)
4. Kudelka, M., Drázdilová, P., Ochodkova, E., Slaninová, K., Horak, Z.: Local community detection and visualization: experiment based on student data. In: Kudělka, M., Pokorný, J., Snášel, V., Abraham, A. (eds.) *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011)*, Prague, Czech Republic, August 2011, pp. 291–303. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-31603-6_25
5. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: *Proceedings of KDD 2007*, pp. 420–429, August 2007
6. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
7. Smets, P.: The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(5), 447–458 (1990)



An Evidential k -nearest Neighbors Combination Rule for Tree Species Recognition

Siwar Jendoubi^{1,2}(✉), Didier Coquin¹, and Reda Boukezzoula¹

¹ LISTIC, University Savoie Mont Blanc, 74944 Annecy-le-Vieux, France
{siwar.jendoubi,didier.coquin,reda.boukezzoula}@univ-smb.fr

² LARODEC, University of Tunis, ISG Tunis, 2000 Le Bardo, Tunisia

Abstract. The task of tree species recognition is to recognize the tree species using photos of their leaves and barks. In this paper, we propose an evidential k -nearest neighbors (k -NN) combination rule. The proposed rule is adapted to classification problems where we have a large number of classes with an intra-class variability and an inter-class similarity like the problem of tree species recognition. Finally, we compare the performance of the proposed solution to the evidential k -NN.

Keywords: Tree species recognition · Belief functions theory · k -NN

1 Introduction

The tree species recognition¹ searches to identify the tree species through photos of leaves and barks taken with a smartphone. The automation of this task is very useful for non botanist users who want to learn more about trees. The idea is to help a user and to teach him how to recognize tree species. Trees recognition is a challenging classification problem.

The k -nearest neighbors (k -NN) classification is a fundamental and simple technique. In fact, all we need to use it is a training set that contains a representative labeled data sample of all possible classes in a given problem and a distance metric. Then, to classify a new data point x , k -NN computes the point distance with all points in the training set and selects the k -nearest neighbors. Finally, k -NN chooses the class of x according to the majority vote principle. The k -NN classifier is sensitive to the value of its main parameter k .

Denœux [6] introduced a k -NN decision rule based on the theory of belief functions [9]. The advantage of this rule is that it considers the distances from the nearest neighbors in the decision step (classification step) which leads to more accurate results. In fact, *Denœux*'s solution combines the evidence from the nearest neighbors through the framework of the theory of belief functions.

¹ This work has been supported by the French National Agency for Research with the reference ANR-15-CES38-0004 (ReVeRIES project).

The resulting decision rule is more robust to the conflicting information, *i.e.* when the object to be classified is close to different classes, and information sparsity, *i.e.* when the object to be classified is far from all patterns in the training set.

The evidential k -NN (Ek -NN) is accurate in many existing classification problems. However, when we have a large number of classes, this accuracy may decrease. This problem becomes serious when we have an intra-class variability and an inter-class similarity like the problem of tree species recognition. This fact increases the conflict between species and leads to miss classification. This limitation was addressed in the literature. In fact, [8, 11] proposed variants of the Ek -NN based on the Parametric Conjunctive t-Rules and an hybrid Dempster-Yager Rule respectively.

The main contributions of this paper are the following: first, we propose an evidential k -NN rule that is more adapted for the large number of classes problem. Besides, the proposed solution deals with the conflict between species and reduces it. In fact, we propose to use a modified version of the large number of source combination algorithm introduced by Zhou *et al.* [13]. Second, we prove the performance of the proposed solution through a set of experiments on the trees recognition problem. Then, we show that the proposed solution is more accurate than the Ek -NN.

This paper is organized as follows: Sect. 2 is dedicated to present the tree species recognition problem. Section 3 details the Ek -NN trees recognition system. Section 4 introduces the combination rule for large number of classes recognition. Section 5 presents some experiments. Finally, Sect. 6 concludes the paper.

2 Tree Species Recognition

The trees species recognition is the problem of identifying trees from their leaves, barks, flowers, *etc.* In this work, we are interested in recognizing trees from their leaves and barks. In the nature, recognizing trees is not an easy task and it needs a botanist. In these last years, many researches were conducted to automate this task. However, this is not a simple task. Indeed, in the nature there is a large variety of species. Besides, it is very common to find similarities between different species and a variability of trees in the same species. Let take the example of leaves in Fig. 1, the leaves (a) and (b) look different, but they belong to the holly species. Besides, the leaves (b) and (c) are very similar, but they belong to two different species (holly and oak).

In the literature, many solutions was proposed for this problem [1, 3, 7]. Besides, many smartphone applications are now available like Pl@ntNet and Folia. Pl@ntNet [7] gives good recognition rates. However, it needs to be connected to Internet. Moreover, it provides the results without any explanation. Whereas, we want to provide the user with information to help him get in the world of botany. In the other hand, we have Folia [3], this application does not need Internet connection. In fact, it recognizes the tree through a limited number of attributes extracted from leaves photos then it can be run on a smartphone.

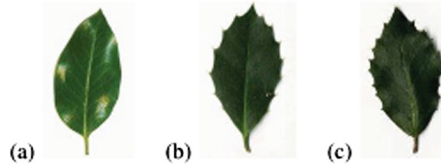


Fig. 1. Intra-species variability and interspecies similarity: (a) and (b) belong to the holly species and (c) belongs to the oak species

Besides, this application explains to the user its results. However, the current version of Folia recognizes only photos of leaves.

Ben Ameur et al. [1] introduced a model-based evidential solution that recognizes trees from leaves and barks. Their solution classifies the leaf and bark photos of the tree using the random forest classifier. Then, they use the inverse pignistic [12] operator to transform the classification results to a consonant mass distribution. Next, they fuse the consonant masses of leaf and bark to obtain a mass distribution that is used to recognize the tree. In this paper, we search to improve the results of [1]. Furthermore, we want to avoid the random forest classification step that is consuming in execution time.

In the next section, we detail the evidential k -NN species recognition system.

3 Evidential k -nearest Neighbors for Trees Recognition

To recognize tree species from leaves and barks, we follow the botanists strategy. In fact, they identify the different morphological characteristics of leaves (apex, shape, *etc*) and barks (color, gabor, *etc*). Then, we extract from each characteristic a vector of attributes to characterize it. From each leaf photo, three vectors are extracted. The first one characterizes the apex and the base of the leaf, the second represents the margin and the last characterizes the polygonal model. From each bark photo, four attributes vectors are extracted which characterize respectively: (1) the color hue H of the HSV space, (2) the texture (gabor) space, (3 & 4) vertical and horizontal orientation of the bark texture. *Bertrand et al.* [2] detail these characteristics. We consider each characteristic as a source of information.

We define a classification system as described in Fig. 2. This architecture is useful to provide the user with classification results according to each characteristic separately and with an explanation of the results. Next, to classify a new leaf/bark photo, we extract the same characteristics. Then, we apply an evidential k -NN (Ek -NN) on each characteristics as presented in Fig. 2. The output of each Ek -NN is a BBA distribution defined on the frame of all possible species. Next, we combine the obtained BBAs to make a decision according to leaves, barks and combined leaves and barks.

The Ek -NN [6] starts like the probabilistic k -NN by estimating the distance between an unclassified object x and all the elements in the training set. Next, it takes the k neighbors having the least distances to x . At this step, the

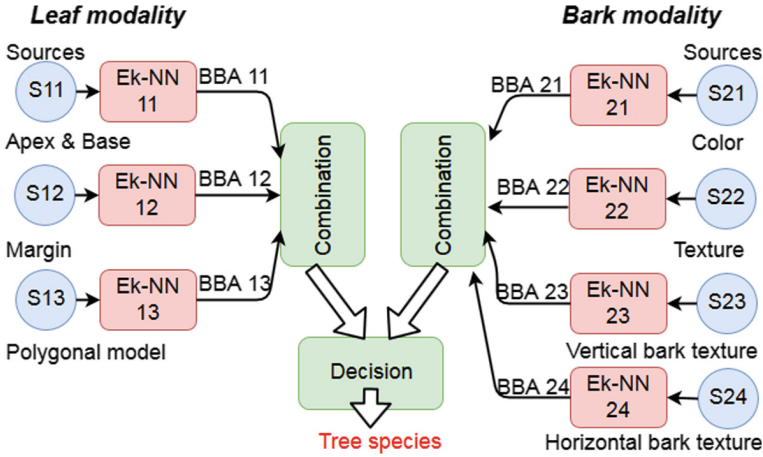


Fig. 2. Evidential k -NN based species recognition

probabilistic k -NN chooses the class of x according to the majority vote principle. Whereas, Ek -NN estimates a mass distribution for each nearest neighbor, i , using its distance to x . Let us define the frame of discernment $\Omega = \{c_1, c_2, \dots, c_n\}$. The Ek -NN estimates the mass m_i^Ω for each nearest neighbor i as follows:

$$m_i^\Omega(c_i) = \alpha_0 \Phi_i(d_i) \tag{1}$$

$$m_i^\Omega(\Omega) = 1 - m_i^\Omega(c_i) \tag{2}$$

such that c_i is the class of the neighbor i , d_i is the distance between x and i and Φ_i is a decreasing function that may be defined as $\Phi_i(d_i) = e^{-\gamma_i d_i^\beta}$ where $\gamma_i > 0$ and $\beta \in [0, 1]$. Then, the Ek -NN combines the obtained mass functions from all the neighbors in order to obtain a decision mass distribution. The combination is done, generally, using the Dempster’s rule [5]. Finally, the decision may be taken using the maximum pignistic.

The inter-species similarities and the intra-species variability make harder the recognition task for the Ek -NN. These two facts lead to an imprecise and uncertain environment for the Ek -NN. Besides, they generates an important conflict between the species. Given the example presented in Fig. 3, we have a leaf and we want to identify its species. The Ek -NN selects the k nearest neighbors to the given leaf. In Fig. 3 we have many species that are near to the unknown leaf. If we take a small k value ($k = 3$ for example), the true class (C6 in this case) will not appear in the nearest neighbors set (C1, C4, C5). Then, we need to choose a big value of k in order to increase the probability that the true class appears. In our experiments we fixed k to 20 according to an Akaike’s information criterion (AIC) [4].

After selecting the nearest neighbors, the Ek -NN estimates a simple BBA for each selected neighbor. Besides, the distances between the unknown leaf and its nearest neighbors are small and almost equal. Then we will obtain k BBA with a

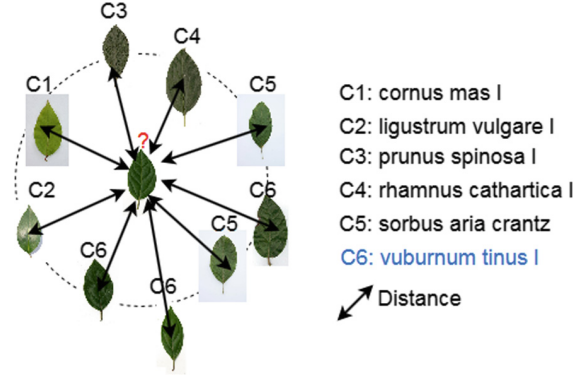


Fig. 3. Inter-species similarities and intra-species variabilities effects on the selected nearest neighbors

high amount of belief on the neighbor’s class (C_i). Combining these BBAs leads to a high conflict value (mass of the empty set $m(\emptyset)$) that tends to 1. This fact leads to a miss classification even when we use the Dempster’s rule.

4 Proposed Evidential k -nearest Neighbors Combination Rule

To better distribute the conflict and to resolve the combination problem introduced in the previous section, we propose to combine the BBAs estimated from the nearest neighbors using the *large number of sources combination algorithm (LNS)* proposed in [13]. LNS takes as input a set of simple masses, *i.e.* a mass with two focal elements among them Ω . Then, the LNS input is the set of mass functions estimated from neighbors. In the first step, LNS clusters the masses into θ clusters according to their focal element $A \neq \Omega$. Next, it combines each cluster masses using a combination rule ([13] used the *conjunctive rule of combination (CRC)* [10]).

After combining the cluster masses, we obtain θ masses, one by cluster. The next step is the reliability-based discounting step. Then, the clusters are seen as sources of information and the main purpose is to consider the reliability of each cluster through the following hypothesis: *the larger the number of masses in the cluster θ_j is, the more reliable the cluster θ_j is.* Then, we propose to estimate the reliability coefficient of the j^{th} cluster as $\varepsilon_j = \frac{\text{Card}(\theta_j)}{k}$. Next, each cluster mass is discounted using its reliability coefficient. The last step of the LNS algorithm is to combine all discounted cluster masses using a combination rule ([13] used the CRC).

The advantage of the LNS algorithm is that it offers the possibility to use two different combination rules as it combines masses in two levels, *i.e.* intra and inter-clusters. Then, we propose the *disjunctive rule of combination (DRC)*

[10] to combine intra-cluster masses and the CRC to combine the inter clusters masses. In the next section, we present a set of experiments on the tree species recognition problem and we show the contribution of the DRC in the LNS algorithm.

5 Experiments

To evaluate the proposed solution, we use a dataset from the ImageClef challenge. The dataset contains photos of trees from mainland France. Those photos are taken in the wild by non-professionals. Then, the photos in this dataset are similar to those a user may take in the nature. The dataset contains 2572 leaves photos and 895 barks photos for training and 820 couple of leaves and barks photos for testing. In the dataset, we have 72 tree species to recognize. In the experiments, we fixed the k parameter to 20 according to an AIC criterion, $\alpha_0 = 0.944$, $\gamma_i = 0.76$ and $\beta = 3.6$. We note that the classifier returns the ten most likely tree species. Then we compare LNS evidential k -NN rule (LNS Ek -NN) with the Ek -NN according to the accuracy that the good species is among the first ten returned species.

In a first experiment, we compare the proposed LNS Ek -NN with the Ek -NN using the Dempster’s rule to combine the k masses. We used the conjunctive combination rule to combine the classification results of leaves characteristics and those of barks, and we used the disjunctive combination rule to combine the results of leaves and barks. Figure 4 presents the classification results of the two experimented classifiers according to leaves characteristics, barks characteristics and combined leaves and barks. According to Fig. 4, the Ek -NN is not efficient as it gives low classification rates. Whereas, the proposed solution gives good classification rates. In fact, we have got an accuracy equals to 49.87% for the first recognized species and 89.14% for ten species from the combined leaves and barks classifiers.

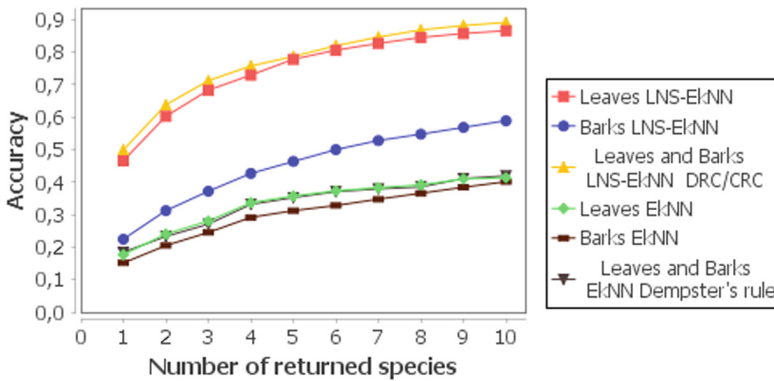


Fig. 4. Comparison between the proposed LNS evidential k -NN and the *Denœux*’s evidential k -NN according to leaves, barks and combined leaves and barks accuracy

In a second experiment, we study the impact of the combination rule on the accuracy of the combined leaves and barks classifiers. Results of this experiment are shown in Fig. 5. Then, we tested the Ek -NN with the Dempster’s rule, CRC and DRC. We notice that the used combination rule has an important impact on the accuracy of this classifier. In fact, the CRC and the DRC succeeded to improve the accuracy of the Ek -NN.

We said in the previous section that the LNS algorithm allows the use of two different combination rules to combine intra and inter-clusters masses. Then in Fig. 5, we present the accuracy using the DRC to combine intra-cluster masses and the CRC to combine the inter clusters masses compared to the accuracy when we use the CRC for both of them as proposed by [13]. According to this experiment, the DRC has a positive impact on the accuracy of the proposed LNS evidential k -NN classifier. Then, we have 49.87% for the first species recognized with the DRC (red curve) against 45.48% when we use only the CRC (blue curve). Besides, we have 89.14% for ten species with the DRC (red curve) against 88.56% when we use only the CRC (blue curve). Furthermore, the effect of the DRC is more important for the first eight detected species. In fact, when we use the DRC combination rule to combine intra-clusters masses, the mass value on the global ignorance, *i.e.* Ω , in the resulting distribution is more important than its value when we use the CRC, and this fact, allowed the classifier to more consider the uncertainty and to improve the results as shown in Fig. 5.

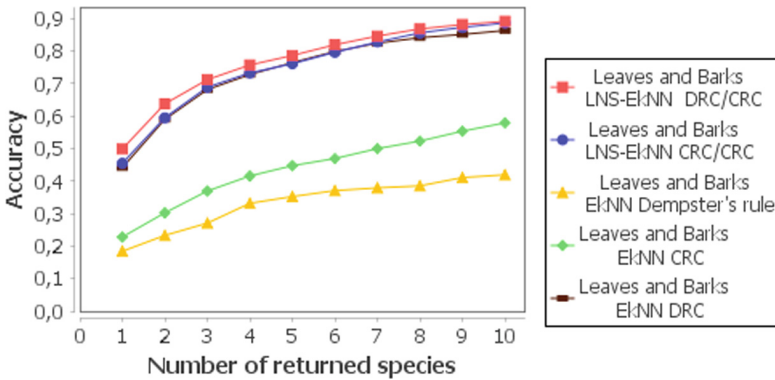


Fig. 5. Impact of the combination rule on the classifier accuracy: fusion of leaves and barks classifiers

From the experiments, we can conclude that the LNS combination algorithm is adapted to combine the estimated masses of the evidential k -NN. In fact, as we see above the proposed LNS evidential k -NN classifier is more accurate in recognizing tree species than the Ek -NN.

6 Conclusion

To sum up, in this paper we focus on the problem of tree species recognition which is a challenging task. We propose the LNS evidential k -nearest neighbors classifier. This solution is more adapted for classification when we have a large number of classes problem. In fact, the LNS combination algorithm is permanent to combine evidence from large number of sources [13] and to deal with the conflict which is the case of the problem in this paper.

In the future works, we will search to improve the achieved results in order to give more accurate recognition to the end user. Then, we will search to optimize the Ek -NN parameters according to [14]. Another good solution may be to use an adapted classifier for each extracted characteristic from leaves and barks.

References

1. Ben Ameer, R., Coquin, D., Valet, L.: Influence of the basic belief assignments construction on the behavior of a fusion system for tree species recognition. In: Proceedings of 20th International Conference on Information Fusion, Xi'an, China, IEEE FUSION, July 2017
2. Bertrand, S., Ameer, R.B., Cerutti, G., Coquin, D., Valet, L., Tougne, L.: Bark and leaf fusion systems to improve automatic tree species recognition. *Ecol. Inf.* **46**, 57–73 (2018)
3. Cerutti, G., Tougne, L., Mille, J., Vacavant, A., Coquin, D.: Understanding leaves in natural images—a model-based approach for tree species identification. *Comput. Vis. Image Underst.* **117**(10), 1482–1501 (2013)
4. Cetin, M.C., Erar, A.: Variable selection with akaike information criteria: a comparative study. *Hacettepe J. Math. Stat.* **31**, 89–97 (2002)
5. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
6. Dencœux, T.: A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **25**(5), 804–813 (1995)
7. Goëau, H., et al.: Pl@ntnet mobile app. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 423–424 (2013)
8. Lian, C., Ruan, S., Dencœux, T.: An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recognit.* **48**(7), 2318–2327 (2015)
9. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Belmont (1976)
10. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *Int. J. Approximate Reasoning* **9**, 1–35 (1993)
11. Su, Z.G., Denœux, T., Hao, Y.S., Zhao, M.: Evidential k -NN classification with enhanced performance via optimizing a class of parametric conjunctive t -rules. *Knowl. Based Syst.* **142**, 7–16 (2018)
12. Sudano, J.J.: Inverse pignistic probability transforms. In: Proceedings of FUSION, pp. 763–768 (2002)
13. Zhou, K., Martin, A., Pan, Q.: Evidence combination for a large number of sources. In: Proceedings of 20th International Conference on Information Fusion, Xi'an, China, IEEE FUSION, July 2017
14. Zouhal, L.M., Dencœux, T.: An evidence-theoretic k -NN rule with parameter optimization. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **28**(2), 263–271 (1998)



A Compact Belief Rule-Based Classification System with Evidential Clustering

Lianmeng Jiao, Xiaojiao Geng^(✉), and Quan Pan

School of Automation, Northwestern Polytechnical University,
Xi'an 710072, People's Republic of China
{jiaolianmeng, quanpan}@nwpu.edu.cn, xiaojiaogeng@mail.nwpu.edu.cn

Abstract. In this paper, a rule learning method based on the evidential C-means clustering is proposed to efficiently design a compact belief rule-based classification system. In this method, the evidential C-means algorithm is first used to obtain credal partitions of the training set. The clustering process operates in a supervised way by means of weighted product-space clustering with the goals of obtaining both good inter-cluster separability and inner-cluster pureness. Then the antecedent part of a belief rule is defined by projecting each multi-dimensional credal partition onto each feature. The consequent class and the weight of each belief rule are identified by combing those training patterns belonging to each hard credal partition within the framework of belief functions. An experiment based on several real data sets was carried out to show the effectiveness of the proposed method.

1 Introduction

Pattern classification is an active field in machine learning and artificial intelligence. Its main purpose is to assign the objects, represented by attribute (or feature) vectors to predefined group of classes. In the past five decades, a variety of classification techniques, such as support vector machines (SVM), neural networks (NN), naive Bayes (NB), K-nearest neighbors (K-NN), rule-based classification (RBC), decision trees (DT), have been proposed [1]. Among these methods, RBC not only has its own advantage in classification result interpreting, but also can be easily enhanced and complemented by adding new rules from experts based on their domain knowledge. One of the most representative RBC methods is the fuzzy rule-based classification system (FRBCS) [4, 6], which is developed incorporating fuzzy sets. The FRBCS is widely employed due to its capability of building a linguistic model interpretable to users. It has been successfully applied to many real-world classification tasks where model interpretability is important, including, but not limited to, terrain classification [12], intrusion detection [10], fault classification [13], target recognition [14], and disease diagnosis [2].

In [8], we have extended the FRBCS within the framework of Dempster-Shafer theory or belief function theory [5, 11], and developed a belief rule-based classification system (BRBCS) to address imprecise or incomplete information in complex classification problems. Compared with the fuzzy rule, the consequent part of the belief rule is in a belief distribution form, which is more informative to characterize different kinds of uncertain information existing in the training set. In addition, in the reasoning process, the class label of a query pattern is decided by combining the consequent parts of all the activated belief rules, which can reduce the risk of misclassification in noisy conditions. In many situations, this method is found experimentally to yield better accuracy and robustness than FRBCS using the same information.

Rule learning is the most important issue in developing the BRBCS. In [8], a heuristic belief rule base (BRB) learning method was developed by defining belief rules based on individuals of the training patterns, and the resulting BRB can provide an accurate mapping between the feature space and the class space. However, with this method, a higher number of data generally induces the BRB with larger size. This may lead to a large rule base for big data set, which affects the interpretability of the classification model. Motivated by the above consideration, in this paper, a compact belief rule-based classification system (CBRBCS) is developed for a better trade-off between accuracy and interpretability. We propose to learn a compact BRB based on partitions of the training set realized with clustering techniques. The evidential C-mean (ECM) algorithm [9], which extended the fuzzy C-mean algorithm within the framework of belief functions, is used for its capability to address imprecise and partial information existed in observed data. The clustering process operates in a supervised way by means of weighted product-space clustering in order to take into account the class labels. As belief rules are constructed based on credal partitions of the training set, this method can reduce the number of generated rules greatly.

The rest of the paper is organized as follows. In Sect. 2, the basics of the belief rule-based classification system and the evidential C-mean algorithm are reviewed. The compact BRB learning with evidential clustering is developed in Sect. 3 and then several benchmark data sets are used to evaluate the performance of the proposed method in Sect. 4. At last, Sect. 5 concludes the paper.

2 Background

2.1 Belief Rule-Based Classification System (BRBCS)

A belief rule-based classification system is composed of two main conceptual components, the belief rule base (BRB) and the belief reasoning method (BRM). The BRB establishes a mapping between the space of pattern features and the space of consequent classes, and the BRM provides a mechanism to classify a query pattern based on the BRB [8].

For an M -class (denoted as $\mathcal{C} = \{c_1, c_2, \dots, c_M\}$) classification problem with P features, the BRB consists of a collection of belief rules defined as follows:

$$R^j : \text{If } x_1 \text{ is } A_1^j \text{ and } x_2 \text{ is } A_2^j \text{ and } \dots \text{ and } x_P \text{ is } A_P^j, \\ \text{then class is } \mathbf{C}^j = \left\{ (c_1, \beta_1^j), \dots, (c_M, \beta_M^j) \right\}, \\ \text{with rule weight } \theta^j, \quad j = 1, 2, \dots,$$

where x_1, x_2, \dots, x_P represent the antecedent features and $\mathbf{A}^j = (A_1^j, A_2^j, \dots, A_P^j)$ is the antecedent part of the belief rule R^j with each A_p^j belonging to fuzzy partitions $\{A_{p,1}, A_{p,2}, \dots, A_{p,n_p}\}$ associated with p -th feature, $p = 1, \dots, P$. β_k^j is the belief degree that input data $\mathbf{x} = (x_1, x_2, \dots, x_P)$ belongs to c_k , $k = 1, \dots, M$. In the belief structure, the consequence may be incomplete, i.e., $\sum_{k=1}^M \beta_k^j \leq 1$, and the left belief $1 - \sum_{k=1}^M \beta_k^j$ denotes the degree of global ignorance about the consequence. The rule weight θ^j with $0 \leq \theta^j \leq 1$, characterizes the certainty grade of the belief rule R^j .

The BRB can be learned from training data or derived from expert knowledge [7]. In [8], we developed a heuristic BRB learning method within the framework of belief functions. To generate the BRB, this method uses the following steps.

Step1: *Partition of the feature space.*

The fuzzy grid-based method is used to divide the P -dimensional feature space into $\prod_{p=1}^P n_p$ fuzzy regions.

Step2: *Generation of the consequent class for each fuzzy region.*

Each training pattern is assigned to the fuzzy region with the greatest matching degree, and the class labels of training patterns assigned to the same fuzzy region are combined to get the consequent class.

Step3: *Generation of the rule weights.*

The rule weights are determined by two measures called confidence and support jointly.

Once the BRB is generated, the BRM is used to classify a query pattern by combining the consequent parts of all the activated belief rules (refer to [8] for details of this reasoning method).

2.2 Evidential C-Means (ECM)

In [9], the evidential C-means (ECM) algorithm was proposed to derive credal partitions from object data. The class membership of an object \mathbf{x}_i is represented by a mass function m_i over a given frame of discernment $\Omega = \{\omega_1, \omega_2, \dots, \omega_C\}$. This representation is able to model all situations ranging from complete ignorance to full certainty concerning the class of the object.

The credal partitions of N observed data $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^P$ are then defined as the N -tuple $M = (m_1, m_2, \dots, m_N)$. For each object \mathbf{x}_i , the quantities $m_{ij} = m_i(A_j)(A_j \subseteq \Omega, A_j \neq \emptyset)$ are determined in such a way that the mass of belief m_{ij} is low (high) when the distance d_{ij} between object \mathbf{x}_i and focal set A_j

is high (low). The distance between object \mathbf{x}_i and focal set A_j is calculated by $d_{ij} = \|\mathbf{x}_i - \bar{\mathbf{v}}_j\|$, where $\bar{\mathbf{v}}_j$ is the barycenter of the centers associated to the classes composing A_j . Denoting \mathbf{v}_k the center of the single cluster ω_k , the barycenter $\bar{\mathbf{v}}_j$ is calculated as

$$\bar{\mathbf{v}}_j = \frac{1}{|A_j|} \sum_{k=1}^C s_{kj} \mathbf{v}_k \text{ with } s_{kj} = \begin{cases} 1, & \text{if } \omega_k \in A_j \\ 0, & \text{otherwise} \end{cases}. \quad (1)$$

Finally, the objective function used to derive the credal partition matrix M of size $(2^C \times N)$ and the cluster center matrix V of size $(C \times P)$ given by

$$J_{\text{ECM}}(M, V) = \sum_{i=1}^N \sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^N \delta^2 m_{i\emptyset}^\beta, \quad (2)$$

subject to

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1, \quad \forall i = 1, \dots, N, \quad (3)$$

where $\beta > 1$ is a weighting exponent that controls the fuzziness of the partition, $\delta > 0$ controls the amount of data considered as outliers and $m_{i\emptyset}$ denotes $m_i(\emptyset)$, the amount of evidence that the class of object \mathbf{x}_i does not lie in Ω . The weighting coefficient $|A_j|^\alpha$ was introduced to penalize the subsets in Ω of high cardinality and the exponent $\alpha \geq 0$ allows to control the degree of penalization. The objective function is minimized using an iterative algorithm, which alternatively optimizes the credal partition matrix M and the cluster center matrix V .

3 Compact BRB Learning with ECM

As reviewed in Sect. 2.1, in the traditional BRB learning method, belief rules are defined based on individuals of the training patterns. This may lead to a large rule base for big data set, which affects the interpretability of the classification model. In this section, we propose to learn a compact BRB based on a partition of the training set realized with clustering techniques. The ECM algorithm is used here in order to incorporate the additional degrees of freedom and information obtained from the derived credal partition, in the belief rule-based classification system.

3.1 Credal Partition of the Feature Space

In typical classification problems, a set of N labeled patterns $\mathcal{T} = \{(\mathbf{x}_1, c^{(1)}), (\mathbf{x}_2, c^{(2)}), \dots, (\mathbf{x}_N, c^{(N)})\}$ with input vectors $\mathbf{x}_i \in \mathbb{R}^P$ and class labels $c^{(i)} \in \{c_1, c_2, \dots, c_M\}$ are available, and the problem is to classify a query pattern \mathbf{y} based on the training set \mathcal{T} . Different from unsupervised clustering problems

which only consider the inter-cluster separability, a good partition of labeled patterns should also take into account the inner-cluster pureness. For this purpose, we cluster the N labeled patterns in the following weighted product space

$$\mathbf{z} = (\mathbf{x} \times Wc), \quad (4)$$

where $W \geq 0$ controls the weight of class labels in clustering process. If $W = 0$, it just reduces to the unsupervised clustering, and as $W \rightarrow \infty$, the resulting clusters are divided only based on the class labels. A suggested choice of W for balancing the effects of feature values and class values is

$$W = \sqrt{\frac{\sum_{p=1}^P \sigma_p^2}{\sigma_c^2}}, \quad (5)$$

where σ_p^2 is the variance of p -th feature values, $p = 1, 2, \dots, P$, and σ_c^2 is the variance of class values.

With given weight W and the number of clusters C , the ECM clustering algorithm is used for the training set \mathcal{T} to discover credal partitions of the feature space. Noting that in ECM the training patterns assigned to empty set are considered as outliers, which are adverse to classification, we only construct belief rules based on $2^C - 1$ non-empty subsets of partitions obtained from the clustering algorithm. From the obtained credal partition matrix M , whose ij -th element $m_{ij} \rightarrow [0, 1]$ is the membership degree of the data \mathbf{x}_i in partition j , it is possible to extract the fuzzy sets in the antecedent parts of the belief rules.

One-dimensional antecedent fuzzy sets A_p^j are obtained from the multidimensional credal partition M by point wise projection [3] onto the space of the antecedent features x_p , $p = 1, 2, \dots, P$:

$$\mu_{A_p^j}(x_{ip}) = \text{proj}_p(m_{ij}). \quad (6)$$

With the above point-wise defined membership, a continuous membership function $\mu_{A_p^j}(x)$ for fuzzy sets A_p^j can be approximated. Several types of functions such as triangular, trapezoidal or Gaussian, can be used. In this work we choose the Gaussian membership function of the form

$$\mu_{A_p^j}(x) = f(x; \bar{v}_{jp}, \sigma_{jp}) = e^{\left(-\frac{(x - \bar{v}_{jp})^2}{2\sigma_{jp}^2}\right)}, \quad (7)$$

where \bar{v}_{jp} is the mean value calculated as Eq. (1), and σ_{jp} is the standard variance to be estimated.

In this way, for each credal partition j , $j = 1, 2, \dots, 2^C - 1$, a series of fuzzy sets $A_1^j, A_2^j, \dots, A_P^j$ can be defined on the antecedent features with Gaussian membership functions, which constitute the antecedent part of belief rule R^j .

3.2 Generation of the Consequent Class

Based on the credal partition matrix M , the training set \mathcal{T} can be divided into 2^C groups by assigning each pattern to the partition with highest mass:

$$\mathcal{T}^j = \{(\mathbf{x}_i, c^{(i)}) | m_{ij} = \max_k m_{ik}, i = 1, \dots, N\}, \quad j = 1, 2, \dots, 2^C. \quad (8)$$

The training subsets \mathcal{T}^j for $j = 1, 2, \dots, 2^C$ define a hard credal partition [9] of the training set \mathcal{T} . The subset \mathcal{T}^{2^C} , which contains the outliers, is discarded. In the following, we will derive the consequent class of belief rule R^j by combining the class information of patterns in the reminder subset $\mathcal{T}^j, j = 1, 2, \dots, 2^C - 1$.

First, for any pattern $\mathbf{x}_i \in \mathcal{T}^j$, we calculate the matching degree with antecedent part of belief rule R^j using the geometric mean operator as

$$\mu_{\mathbf{A}^j}(\mathbf{x}_i) = \sqrt[P]{\prod_{p=1}^P \mu_{A_p^j}(x_{ip})}, \quad (9)$$

where $\mu_{A_p^j}$ is the membership function of the fuzzy set A_p^j defined in Eq. (7).

Then, assume the class label of pattern \mathbf{x}_i is c_k , which takes value in class set \mathcal{C} . This can be regarded as a piece of evidence that increases the belief of the consequent class belongs to c_k . However, this piece of evidence does not by itself provide full certainty. In belief function theory, this can be expressed by saying that only some part of the belief (measured by the matching degree $\mu_{\mathbf{A}^j}(\mathbf{x}_i)$) is committed to c_k . Because $\text{Class}(\mathbf{x}_i) = c_k$ does not point to any other particular class, the rest of the belief should be assigned to the frame of discernment \mathcal{C} representing global ignorance. Therefore, this item of evidence can be represented by a mass function $m^j(\cdot|\mathbf{x}_i)$ verifying:

$$\begin{cases} m^j(\{c_k\}|\mathbf{x}_i) = \mu_{\mathbf{A}^j}(\mathbf{x}_i) \\ m^j(\mathcal{C}|\mathbf{x}_i) = 1 - \mu_{\mathbf{A}^j}(\mathbf{x}_i) \\ m^j(A|\mathbf{x}_i) = 0, \quad \forall A \in 2^{\mathcal{C}} \setminus \{\mathcal{C}, \{c_k\}\} \end{cases}. \quad (10)$$

Finally, the mass functions derived from all of the patterns in \mathcal{T}^j are combined to obtain the consequent class of belief rule R^j . As the items of evidence from different labeled patterns are collected independently, the Dempster's rule [5] is used in this work to synthesizing the final consequent class membership as

$$m^j = \bigoplus_{\mathbf{x}_i \in \mathcal{T}^j} m^j(\cdot|\mathbf{x}_i). \quad (11)$$

Noting that all the sources of evidence have only one focal set except the global set \mathcal{C} , the computation of Dempster's rule is quite efficient. The belief degrees of the consequent class of rule R^j are then obtained as $\beta_k^j = m^j(\{c_k\}), k = 1, 2, \dots, M$.

3.3 Generation of the Rule Weights

As in [8], the rule weights can be derived based on two concepts called *confidence* and *support*, which are often used for evaluating association rules in data mining fields. The confidence is a measure of the validity of one rule, which is defined as

$$c(R^j) = 1 - \overline{K^j}, \quad (12)$$

where $0 \leq \overline{K}^j \leq 1$ is the average conflict factor, which measures the conflict among those pieces of evidence used for building the consequent class of rule R^j :

$$\overline{K}^j = \begin{cases} 0, & \text{if } |\mathcal{T}^j| = 1, \\ \frac{1}{|\mathcal{T}^j|(|\mathcal{T}^j|-1)} \sum_{\substack{\mathbf{x}_p, \mathbf{x}_q \in \mathcal{T}^j; \\ c(p) \neq c(q)}} \mu_{\mathbf{A}^j}(\mathbf{x}_p) \mu_{\mathbf{A}^j}(\mathbf{x}_q), & \text{otherwise.} \end{cases} \quad (13)$$

with $|\mathcal{T}^j|$ denoting the number of training patterns in j -th hard credal partition.

On the other hand, the support indicates the grade of the coverage by one rule, which is defined as the ratio of the number of covered patterns to the total pattern number:

$$s(R^j) = \frac{|\mathcal{T}^j|}{N}. \quad (14)$$

Based on the above two measures, the rule weights are finally derived as

$$\theta^j = \frac{c(R^j)s(R^j)}{\max_j \{c(R^j)s(R^j), j = 1, \dots, 2^C - 1\}}, \quad j = 1, 2, \dots, 2^C - 1. \quad (15)$$

4 Experiment

In this experiment, four well-known benchmark data sets from UCI Repository of Machine Learning Databases (<http://archive.ics.uci.edu>) are used to evaluate the performance of the proposed compact belief rule-based classification system (CBRBCS). The main characteristics of the four data sets are summarized in Table 1.

Table 1. Statistics of the benchmark data sets used in the experiment.

Data set	# of instances	# of features	# of classes
Diabetes	768	8	2
Letter	20,000	16	26
Segment	2,310	19	7
Vehicle	846	18	4

To develop the experiments, we consider the *B-Fold Cross-Validation* (B-CV) model. Each data set is divided into B blocks, with $B - 1$ blocks as a training set and the remaining block as a test set. Therefore, each block is used exactly once as a test set. We use the 5-CV here, i.e., five random partitions of the original data set, with four of them (80%) as the training set and the remainder (20%) as the test set. For each data set, we consider the average results of the five partitions.

The performance of the proposed classifier is compared with the traditional BRBCS as well as several classical classification methods including K-NN, C4.5, and naive Bayes (NBayes). For BRBCS the typical five partitions for each feature are used. For CBRBCS, the default values of open parameters in ECM are used. The number of clusters C is optimized by minimizing the validity index of credal partitions defined in [9]. Besides, in order to reduce the complexity, we constrain the focal sets to be either Ω , or to be composed of at most two classes. Table 2 shows the classification error rates of different methods as well as the numbers of generated rules for BRBCS and CBRBCS (the numbers after virgule). It can be seen that the performance of the two belief rule-based classifiers, i.e., BRBCS and CBRBCS, is comparable with the classical methods. Compared with the traditional BRBCS, the proposed CBRBCS obtains a better trade-off between accuracy and interpretability (little accuracy is sacrificed with much smaller size of rules).

Table 2. Classification performance of CBRBCS in comparison with other methods.

Data set	K-NN	C4.5	NBayes	BRBCS	CBRBCS
Diabetes	0.324(5)	0.270(4)	0.262(3)	0.218(1)/248	0.254(2)/37
Letter	0.068(3)	0.132(4)	0.529(5)	0.058(1)/3696	0.063(2)/121
Segment	0.077(2)	0.040(1)	0.265(5)	0.115(3)/827	0.130(4)/79
Vehicle	0.275(2)	0.266(1)	0.558(5)	0.278(3)/633	0.296(4)/46

5 Conclusions

In this paper, a compact belief rule-based classification system with evidential C-mean clustering has been proposed to overcome the limitations of the traditional BRBCS in large data set conditions. Instead of defining belief rules for individuals of the training patterns, belief rules are constructed based on credal partitions of the training set. This method can discover the underling data structure, which can be successfully translated into belief rules. The experiment based on benchmark data sets have shown that the proposed classifier is competitive compared with the classical methods and can obtain a better trade-off between accuracy and interpretability than the traditional one.

Acknowledgments. This work is partially supported by China Natural Science Foundation (Nos. 61790552 and 61672431).

References

1. Aggarwal, C.C.: *Data Classification: Algorithm and Applications*. Chapman & Hall, Boca Raton (2014)
2. Akbarzadeh-Totonchi, M.R., Moshtagh-Khorasani, M.: A hierarchical fuzzy rule-based approach to aphasia diagnosis. *J. Biomed. Inform.* **40**, 465–475 (2007)
3. Almeida, R.J., Denœux, T., Kaymak, U.: Constructing rule-based models using the belief functions framework. In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R. (eds.) *IPMU 2012. CCIS*, vol. 299, pp. 554–563. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-31718-7-57>
4. Chi, Z., Yan, H., Pham, T.: *Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition*. World Scientific, Singapore (1996)
5. Dempster, A.: Upper and lower probabilities induced by multivalued mapping. *Ann. Math. Statist.* **38**, 325–339 (1967)
6. Ishibuchi, H., Nozaki, K., Tanaka, H.: Distributed representation of fuzzy rules and its application to pattern classification. *Fuzzy Sets Syst.* **52**, 21–32 (1992)
7. Jiao, L., Denœux, T., Pan, Q.: A hybrid belief rule-based classification system based on uncertain training data and expert knowledge. *IEEE Trans. Syst. Man Cybern. Syst.* **46**(12), 1711–1723 (2016)
8. Jiao, L., Pan, Q., Denœux, T., Liang, Y., Feng, X.: Belief rule-based classification system: extension of FRBCS in belief functions framework. *Inform. Sci.* **309**(1), 26–49 (2015)
9. Masson, M.H., Denœux, T.: Clustering interval-valued data using belief functions. *Pattern Recogn. Lett.* **25**, 163–171 (2004)
10. Samantaray, S.R.: Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. *Appl. Soft Comput.* **13**, 928–938 (2013)
11. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
12. Stavrakoudis, D.G., Galidaki, G.N., Gitas, I.Z., Theocharis, J.B.: A genetic fuzzy-rule-based classifier for land cover classification from hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **50**, 130–148 (2012)
13. Tsang, C., Kwong, S., Wang, H.: A systematic fuzzy rule based approach for fault classification in transmission lines. *Pattern Recogn.* **40**, 2373–2391 (2007)
14. Wu, H., Mendel, J.: Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers. *IEEE Trans. Fuzzy Syst.* **15**, 56–72 (2007)



A Decomposable Entropy of Belief Functions in the Dempster-Shafer Theory

Radim Jiroušek¹ and Prakash P. Shenoy²(✉)

¹ Faculty of Management, University of Economics,
and The Czech Academy of Sciences, Institute of Information Theory
and Automation, Jindřichův Hradec and Prague, Czech Republic

radim@utia.cas.cz

² School of Business, University of Kansas, Lawrence, KS 66045, USA

pshenoy@ku.edu

Abstract. We define entropy of belief functions in the Dempster-Shafer (D-S) theory that satisfies a compound distributions property that is analogous to the property that characterizes Shannon's definitions of entropy and conditional entropy for discrete probability distributions. None of the existing definitions of entropy for belief functions in the D-S theory satisfy such a compound distributions property. We describe some important properties of our definition.

1 Introduction

The main goal of this paper is to define entropy of belief functions in the Dempster-Shafer's theory [2, 4] that satisfies a compound distributions property analogous to the one that characterizes Shannon's definitions of entropy and conditional entropy for discrete probability distributions [6]. If $P_{X,Y}$ is a probability mass function (PMF) of (X, Y) , and it is decomposed into PMF P_X for X , and conditional probability table $P_{Y|X}$ so that $P_{X,Y} = P_X \otimes P_{Y|X}$, then Shannon's definitions of entropy and conditional entropy satisfy $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X})$. Here, \otimes denotes probabilistic combination, which is point-wise multiplication followed by normalization.

In this paper, we provide definitions of entropy and conditional entropy of belief functions, so that if $m_{X,Y}$ is a basic probability assignment (BPA) for (X, Y) that is constructed from a BPA m_X for X , and a conditional BPA $m_{Y|X}$ for Y given X , such that $m_{X,Y} = m_X \oplus m_{Y|X}$, where \oplus is Dempster's combination rule, then our definitions satisfy $H(m_{X,Y}) = H(m_X) + H(m_{Y|X})$. This is the main contribution of this paper. Our definitions of entropy and conditional entropy have several nice properties similar to corresponding properties of Shannon's entropy. Here, we do not delve into philosophical discussions about what entropy means. Our exposition focusses exclusively on mathematical properties of entropy.

2 Shannon's Definition of Entropy

In this section, we briefly review Shannon's definition of entropy of PMFs of discrete random variables, and its properties. Most of the material in this section is taken from [6].

Definition 1. Suppose P_X is a PMF of discrete variable X . The *entropy* of P_X , denoted by $H_s(P_X)$, is defined as

$$H_s(P_X) = - \sum_{x \in \Omega_X} P_X(x) \log_2(P_X(x)). \quad (1)$$

Suppose $P_{X,Y}$ is a joint PMF of (X, Y) . Then, the *joint* entropy of $P_{X,Y}$ is as in Eq. (1), i.e.,

$$H_s(P_{X,Y}) = - \sum_{(x,y) \in \Omega_{X,Y}} P_{X,Y}(x,y) \log_2(P_{X,Y}(x,y)).$$

Suppose $P_{X,Y}$ is a PMF of (X, Y) with P_X as its marginal PMF for X . Suppose we observe $X = a$ for some $a \in \Omega_X$ such that $P_X(a) > 0$. This observation is represented by the PMF $P_{X=a}$ for X such that $P_{X=a}(a) = 1$. Let $P_{Y|a} = (P_{X,Y} \otimes P_{X=a})^{\downarrow Y}$ denote the posterior PMF of Y , where \otimes denotes pointwise multiplication followed by normalization, the combination rule in probability theory. The *posterior* entropy of $P_{Y|a}$ is as in Eq. (1), i.e., $H_s(P_{Y|a}) = - \sum_{y \in \Omega_Y} P_{Y|a}(y) \log_2(P_{Y|a}(y))$.

Shannon [6] derives the expression for entropy of P_X axiomatically using four axioms as follows:

1. Axiom 1 (*Existence*): $H(P_X)$ exists.
2. Axiom 2 (*Continuity*): $H(P_X)$ should be a continuous function of P_X .
3. Axiom 3 (*Monotonicity*): If we have an equally likely PMF, then $H(P_X)$ should be a monotonically increasing function of $|\Omega_X|$.
4. Axiom 4 (*Compound distributions*): If a PMF is factored into two PMFs, then its entropy should be the sum of entropies of its factors, e.g., $P_{X,Y}(x,y) = P_X(x) P_{Y|x}(y)$, then $H(P_{X,Y}) = H(P_X) + \sum_{x \in \Omega_X} P_X(x) H(P_{Y|x})$.

Shannon [6] proves that the only function H_s that satisfies Axioms 1–4 is of the form $H_s(P_X) = -K \sum_{x \in \Omega_X} P_X(x) \log(P_X(x))$, where K is a constant that depends on the choice of units of measurement.

Let $P_{Y|X} : \Omega_{X,Y} \rightarrow [0, 1]$ be a function such that $P_{Y|X}(x,y) = P_{Y|x}(y)$ for all $(x,y) \in \Omega_{X,Y}$. As $P_{Y|x}(y)$ is only defined for $x \in \Omega_X$ such that $P_X(x) > 0$, we will assume that $P_{Y|X}$ is only defined for $x \in \Omega_X$ such that $P_X(x) > 0$. $P_{Y|X}$ is not a PMF, but can be considered as a collection of PMFs, and it is called a conditional probability table (CPT) in the Bayesian network literature. If we combine P_X and $P_{Y|X}$, we obtain $P_{X,Y}$, i.e., $P_{X,Y} = P_X \otimes P_{Y|X}$.

Definition 2. Suppose $P_{Y|X}$ is a CPT of Y given X for all $x \in \Omega_X$ such that $P_X(x) > 0$. Then the *conditional* entropy of $P_{Y|X}$ is defined as

$$H_s(P_{Y|X}) = \sum_{x \in \Omega_X} P_X(x) H_s(P_{Y|x}). \tag{2}$$

It follows from Axiom 4 that

$$H_s(P_{X,Y}) = H_s(P_X \otimes P_{Y|X}) = H_s(P_X) + H_s(P_{Y|X}). \tag{3}$$

3 Basic Definitions of the D-S Belief Functions Theory

In this section we review the basic definitions in the D-S belief functions theory, including functional representations of uncertain knowledge, and operations for making inferences from such knowledge.

Belief functions can be represented in four different ways: basic probability assignments (BPAs), belief functions, plausibility functions, and commonality functions. Here, we focus only on BPAs and commonality functions.

BPAs. Suppose X is a random variable with state space Ω_X . Let 2^{Ω_X} denote the set of all *non-empty* subsets of Ω_X . A BPA m for X is a function $m : 2^{\Omega_X} \rightarrow [0, 1]$ such that

$$\sum_{\mathbf{a} \in 2^{\Omega_X}} m(\mathbf{a}) = 1. \tag{4}$$

The non-empty subsets $\mathbf{a} \in 2^{\Omega_X}$ such that $m(\mathbf{a}) > 0$ are called *focal* elements of m . We say m is *consonant* if the focal elements of m are nested, i.e., if $\mathbf{a}_1 \subset \dots \subset \mathbf{a}_m$, where $\{\mathbf{a}_1, \dots, \mathbf{a}_r\}$ denotes the set of all focal elements of m . We say m is *quasi-consonant* if the intersection of all focal elements of m is non-empty. A BPA that is consonant is also quasi-consonant, but not vice-versa. Thus, a BPA with focal elements $\{x_1, x_2\}$ and $\{x_1, x_3\}$ is quasi-consonant, but not consonant. If all focal elements of m are singleton subsets of Ω_X , then we say m is *Bayesian*. In this case, m is equivalent to the PMF P for X such that $P(x) = m(\{x\})$ for each $x \in \Omega_X$. If Ω_X is a focal element, then we say m is *non-dogmatic*, and *dogmatic* otherwise. Thus, a Bayesian BPA is dogmatic.

Commonality Functions. The information in a BPA m can also be represented by a corresponding commonality function Q_m that is defined as follows.

$$Q_m(\mathbf{a}) = \sum_{\mathbf{b} \in 2^{\Omega_X} : \mathbf{b} \supseteq \mathbf{a}} m(\mathbf{b}) \tag{5}$$

for all $\mathbf{a} \in 2^{\Omega_X}$. Q_m is a non-increasing function in the sense that if $\mathbf{b} \subseteq \mathbf{a}$, then $Q_m(\mathbf{b}) \geq Q_m(\mathbf{a})$. Finally, Q_m is a normalized function in the sense that

$$\sum_{\mathbf{a} \in 2^{\Omega_X}} (-1)^{|\mathbf{a}|+1} Q_m(\mathbf{a}) = \sum_{\mathbf{b} \in 2^{\Omega_X}} m(\mathbf{b}) = 1. \tag{6}$$

Thus, any non-increasing, non-negative function that satisfies Eq. (6) qualifies as a commonality function.

Next, we describe two main operations for making inferences.

Dempster's Combination Rule. In the D-S theory, we can combine two BPAs m_1 and m_2 representing distinct pieces of evidence by Dempster's rule [2] and obtain the BPA $m_1 \oplus m_2$, which represents the combined evidence.

Let \mathcal{X} denote a finite set of variables. The state space of \mathcal{X} is $\times_{X \in \mathcal{X}} \Omega_X$. Thus, if $\mathcal{X} = \{X, Y\}$ then the state space of $\{X, Y\}$ is $\Omega_X \times \Omega_Y$.

Projection of states simply means dropping extra coordinates; for example, if (x, y) is a state of (X, Y) , then the projection of (x, y) to X , denoted by $(x, y)^{\downarrow X}$, is simply x , which is a state of X .

Projection of subsets of states is achieved by projecting every state in the subset. Suppose $\mathbf{b} \in 2^{\Omega_{X,Y}}$. Then $\mathbf{b}^{\downarrow X} = \{x \in \Omega_X : (x, y) \in \mathbf{b}\}$. Notice that $\mathbf{b}^{\downarrow X} \in 2^{\Omega_X}$.

Vacuous extension of a subset of states of \mathcal{X}_1 to a subset of states of \mathcal{X}_2 , where $\mathcal{X}_2 \supseteq \mathcal{X}_1$, is a cylinder set extension, i.e., if $\mathbf{a} \in 2^{\mathcal{X}_1}$, then $\mathbf{a}^{\uparrow \mathcal{X}_2} = \mathbf{a} \times \Omega_{\mathcal{X}_2 \setminus \mathcal{X}_1}$. Thus, if $\mathbf{a} \in 2^{\Omega_X}$, then $\mathbf{a}^{\uparrow \{X,Y\}} = \mathbf{a} \times \Omega_Y$.

Suppose m_X is a BPA for X , and \mathcal{X} is such that $X \in \mathcal{X}$. Then the vacuous extension of m to \mathcal{X} , denoted by $m_X^{\uparrow \mathcal{X}}$, is the BPA for \mathcal{X} such that $m_X^{\uparrow \mathcal{X}}(\mathbf{a}^{\uparrow \mathcal{X}}) = m_X(\mathbf{a})$, for all $\mathbf{a} \in 2^{\Omega_X}$, i.e., all focal elements of $m_X^{\uparrow \mathcal{X}}$ are vacuous extensions of focal elements of m_X to \mathcal{X} , and they have the same corresponding values.

We will define Dempster's rule in terms of commonality functions [4]. Suppose m_1 and m_2 are BPAs for \mathcal{X}_1 and \mathcal{X}_2 , respectively. Suppose $Q_{m_1^{\uparrow \mathcal{X}}}$ and $Q_{m_2^{\uparrow \mathcal{X}}}$ are commonality functions corresponding to BPAs $m_1^{\uparrow \mathcal{X}}$ and $m_2^{\uparrow \mathcal{X}}$, respectively, where $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. The commonality function $Q_{m_1 \oplus m_2}$ corresponding to BPA $m_1 \oplus m_2$ is

$$Q_{m_1 \oplus m_2}(\mathbf{a}) = K^{-1} Q_{m_1^{\uparrow \mathcal{X}}}(\mathbf{a}) Q_{m_2^{\uparrow \mathcal{X}}}(\mathbf{a}), \tag{7}$$

for all $\mathbf{a} \in 2^{\Omega_{\mathcal{X}}}$, where the normalization constant K is

$$K = \sum_{\mathbf{a} \in 2^{\Omega_{\mathcal{X}}}} (-1)^{|\mathbf{a}|+1} Q_{m_1^{\uparrow \mathcal{X}}}(\mathbf{a}) Q_{m_2^{\uparrow \mathcal{X}}}(\mathbf{a}). \tag{8}$$

The definition of Dempster's rule assumes that the normalization constant K is non-zero. If $K = 0$, then the two BPAs m_1 and m_2 are said to be in *total conflict* and cannot be combined. If $K = 1$, we say m_1 and m_2 are *non-conflicting*.

Marginalization. Marginalization in D-S theory is addition of values of BPAs. Suppose m is a BPA for \mathcal{X} . Then, the marginal of m for \mathcal{X}_1 , where $\mathcal{X}_1 \subseteq \mathcal{X}$, denoted by $m^{\downarrow \mathcal{X}_1}$, is a BPA for \mathcal{X}_1 such that for each $\mathbf{a} \in 2^{\Omega_{\mathcal{X}_1}}$,

$$m^{\downarrow \mathcal{X}_1}(\mathbf{a}) = \sum_{\mathbf{b} \in 2^{\Omega_{\mathcal{X}}} : \mathbf{b}^{\downarrow \mathcal{X}_1} = \mathbf{a}} m(\mathbf{b}). \tag{9}$$

Conditional belief functions. Consider a BPA m_X for X such that $m_X(\{x\}) > 0$. Suppose that there is a BPA for Y expressing our belief about Y if we know

that $X = x$, and denote it by $m_{Y|x}$. Notice that $m_{Y|x} : 2^{\Omega_Y} \rightarrow [0, 1]$ is such that $\sum_{\mathbf{a} \in 2^{\Omega_Y}} m_{Y|x}(\mathbf{a}) = 1$. We can embed this BPA for Y into a conditional BPA for (X, Y) , which is denoted by $m_{x,Y}$, such that the following four conditions hold. First, $m_{x,Y}$ tells us nothing about X , i.e., $m_{x,Y}^{\downarrow X}(\Omega_X) = 1$. Second, $m_{x,Y}$ tells us nothing about Y , i.e., $m_{x,Y}^{\downarrow Y}(\Omega_Y) = 1$. Third, if we combine $m_{x,Y}$ with the BPA $m_{X=x}$ for X such $m_{X=x}(\{x\}) = 1$ using Dempster's rule, and marginalize the result to Y we obtain $m_{Y|x}$, i.e., $(m_{x,Y} \oplus m_{X=x})^{\downarrow Y} = m_{Y|x}$. Fourth, if we combine $m_{x,Y}$ with the BPA $m_{X=\bar{x}}$ for X such $m_{X=\bar{x}}(\{\bar{x}\}) = 1$ using Dempster's rule, and marginalize the result to Y we obtain the vacuous BPA for Y , i.e., $(m_{x,Y} \oplus m_{X=\bar{x}})^{\downarrow Y}(\Omega_Y) = 1$. One way to obtain such an embedding is suggested by Smets [7] (see also [5]), called *conditional embedding*, and it consists of taking each focal element $\mathbf{b} \in 2^{\Omega_Y}$ of $m_{Y|x}$, and converting it to a corresponding focal element of $m_{x,Y}$ (with the same mass) as follows: $(\{x\} \times \mathbf{b}) \cup ((\Omega_X \setminus \{x\}) \times \Omega_Y)$. It is easy to confirm that this method of embedding satisfies all four conditions mentioned above.

This completes our brief review of the D-S belief function theory. For further details, the reader is referred to [4].

4 A Decomposable Entropy for the D-S Theory

In this section, we provide a new definition of entropy of belief functions in the D-S theory, and describe its properties. This new definition is designed to satisfy a compound distributions property analogous to the compound distribution property that characterizes Shannon's entropy of PMFs.

Definition 3. Suppose m_X is a BPA for X with state space Ω_X , and suppose Q_{m_X} denotes the commonality function corresponding to m_X . Then the entropy of m_X , denoted by $H(m_X)$, is defined as follows:

$$H(m_X) = \sum_{\mathbf{a} \in 2^{\Omega_X}} (-1)^{|\mathbf{a}|} Q_{m_X}(\mathbf{a}) \log_2(Q_{m_X}(\mathbf{a})). \quad (10)$$

If $Q_{m_X}(\mathbf{a}) = 0$, we will follow the convention that $Q_{m_X}(\mathbf{a}) \log_2(Q_{m_X}(\mathbf{a})) = 0$ as $\lim_{\theta \rightarrow 0^+} \theta \log_2(\theta) = 0$.

This is a new definition of entropy that has not been proposed earlier in the literature. The closest definition is due to Smets [8], where $H(m)$ is defined as

$$H(m) = - \sum_{\mathbf{a} \in 2^{\Omega_X}} \log_2(Q_m(\mathbf{a})),$$

but only for non-dogmatic BPAs m . Our definition holds for all BPAs. Also, our sum is an alternating weighted sum, whose sign depends on the cardinality of non-empty subset \mathbf{a} .

Suppose $m_{X,Y}$ is a joint BPA for (X, Y) . Then the *joint* entropy of $m_{X,Y}$ is as in Eq. (10), i.e.,

$$H(m_{X,Y}) = \sum_{\mathbf{a} \in 2^{\Omega_{X,Y}}} (-1)^{|\mathbf{a}|} Q_{m_{X,Y}}(\mathbf{a}) \log_2(Q_{m_{X,Y}}(\mathbf{a})).$$

Suppose $m_{X,Y}$ is a BPA for (X, Y) with m_X as its marginal BPA for X . Suppose we observe $X = a$ for some $a \in \Omega_X$ such that $m_X(\{a\}) > 0$. This observation is represented by the BPA $m_{X=a}$ such that $m_{X=a}(\{a\}) = 1$. Let $m_{Y|a} = (m_{X,Y} \oplus m_{X=a})^{\downarrow Y}$ denote the posterior BPA for Y , and its posterior entropy is as in Eq. (10), i.e., $H(m_{Y|a}) = \sum_{\mathbf{a} \in 2^{\Omega_Y}} (-1)^{|\mathbf{a}|} Q_{m_{Y|a}}(\mathbf{a}) \log_2(Q_{m_{Y|a}}(\mathbf{a}))$.

The following theorem says vacuous extension of a BPA does not change its entropy.¹

Theorem 1. *If m is a BPA for X with $\Omega_X = \{x, \bar{x}\}$, and m' is a vacuous extension of m to (X, Y) , where $\Omega_Y = \{y, \bar{y}\}$, then $H(m') = H(m)$.*

Definition 4. Suppose m_X is a BPA for X such that $m_X(x) > 0$. Let $m_{x,Y}$ denote a BPA for (X, Y) representing a conditional BPA of Y given $X = x$. We define entropy of conditional BPA $m_{x,Y}$ as follows:

$$H(m_{x,Y}) = \sum_{\mathbf{a} \in 2^{\Omega_{X,Y}}} (-1)^{|\mathbf{a}|} Q_{m_X^{\uparrow\{x,Y\}}}(\mathbf{a}) Q_{m_{x,Y}}(\mathbf{a}) \log_2(Q_{m_{x,Y}}(\mathbf{a})). \tag{11}$$

The definition in Eq. (11) is analogous to Eq. (2) for the probabilistic case. We have the following result about conditional entropy.

Theorem 2. *Suppose m_X is a BPA for X such that $\Omega_X = \{x, \bar{x}\}$ and $m_X(\{x\}) > 0$. Suppose Y is such that $\Omega_Y = \{y, \bar{y}\}$, and $m_{Y|x}$ is a BPA for Y given $X = x$. Let $m_{x,Y}$ denote a conditional BPA for (X, Y) obtained from $m_{Y|x}$ by conditional embedding. Then,*

$$H(m_{x,Y}) = m_X(\{x\})H(m_{Y|x}). \tag{12}$$

If $\Omega_X = \{x, \bar{x}\}$ and assuming $m_X(\bar{x}) > 0$, it follows from Eq. (11) that

$$H(m_{\bar{x},Y}) = \sum_{\mathbf{a} \in 2^{\Omega_{X,Y}}} (-1)^{|\mathbf{a}|} Q_{m_X^{\uparrow\{x,Y\}}}(\mathbf{a}) Q_{m_{\bar{x},Y}}(\mathbf{a}) \log_2(Q_{m_{\bar{x},Y}}(\mathbf{a})).$$

Also, from Theorem 2, it follows that:

$$H(m_{\bar{x},Y}) = m_X(\{\bar{x}\})H(m_{Y|\bar{x}}).$$

As the contexts in $m_{x,Y}$ and $m_{\bar{x},Y}$ are disjoint, and the beliefs of the contexts are described by the same BPA m_X such that $m_X(x) > 0$ and $m_X(\bar{x}) > 0$, we have the following result.

¹ For lack of space, proofs of all theorems and properties are omitted, and can be found in a working paper that can be downloaded from <http://pshenoy.faculty.ku.edu/Papers/WP334.pdf>.

Theorem 3. *Suppose X and Y are such that $\Omega_X = \{x, \bar{x}\}$, and $\Omega_Y = \{y, \bar{y}\}$. Suppose that we have non-vacuous conditional BPAs $m_{Y|x}$, and $m_{Y|\bar{x}}$ for Y such that $m_X(\{x\}) > 0$, $m_X(\{\bar{x}\}) > 0$, and after conditional embedding, these are represented by conditional BPAs $m_{x,Y}$ and $m_{\bar{x},Y}$ for (X, Y) . Then,*

$$H(m_{Y|X}) = H(m_{x,Y} \oplus m_{\bar{x},Y}) = H(m_{x,Y}) + H(m_{\bar{x},Y}). \quad (13)$$

Notice that the result in Eq. (13) is analogous of the definition of conditional entropy in Eq. (2) in the probabilistic case.

Next, we state the main result of this paper.

Theorem 4. *Suppose X and Y are such that $\Omega_X = \{x, \bar{x}\}$, and $\Omega_Y = \{y, \bar{y}\}$. Suppose m_X is a BPA for X such that $m_X > 0$ and $m_X(\bar{x}) > 0$, and $m_{Y|X} = m_{x,Y} \oplus m_{\bar{x},Y}$ is a conditional BPA for Y given X . Let $m_{X,Y} = m_X \oplus m_{Y|X}$. Then,*

$$H(m_{X,Y}) = H(m_X) + H(m_{Y|X}). \quad (14)$$

Next, we show that a probability model for (X, Y) can be replicated exactly in the DS theory. Furthermore, our definition of entropy for all BPAs will coincide with Shannon's entropy of the corresponding probabilistic function.

Theorem 5. *Suppose X and Y are such that $\Omega_X = \{x, \bar{x}\}$, and $\Omega_Y = \{y, \bar{y}\}$. Suppose P_X is a PMF for X such that $P_X(x) > 0$, and $P_X(\bar{x}) > 0$, and $P_{Y|X}$ is a CPT for Y given X , i.e., $P_{Y|X}(x, y) = P_{Y|\bar{x}}(y)$, where $P_{Y|x}$ is the conditional PMF for Y given $X = x$ for all $(x, y) \in \Omega_{X,Y}$. Let $P_{X,Y} = P_X \otimes P_{Y|X}$. Let m_X denote the Bayesian BPA corresponding to P_X , let $m_{Y|x}$ and $m_{Y|\bar{x}}$ denote the Bayesian BPAs for Y corresponding to PMFs $P_{Y|x}$ and $P_{Y|\bar{x}}$ for Y . Let $m_{x,Y}$ and $m_{\bar{x},Y}$ denote the conditional BPAs for (X, Y) obtained by conditional embedding of $m_{Y|x}$ and $m_{Y|\bar{x}}$. Let $m_{Y|X} = m_{x,Y} \oplus m_{\bar{x},Y}$, and let $m_{X,Y} = m_X \oplus m_{Y|X}$. Then, $m_{X,Y}$ is a Bayesian BPA for (X, Y) corresponding to PMF $P_{X,Y}$,*

$$H(m_{X,Y}) = H_s(P_{X,Y}), \quad (15)$$

$$H(m_X) = H_s(P_X), \text{ and} \quad (16)$$

$$H(m_{Y|X}) = H_s(P_{Y|X}). \quad (17)$$

Notice that $m_{x,Y}$, $m_{\bar{x},Y}$, and $m_{Y|X}$, are not Bayesian BPAs.

5 Other Properties

Some further properties of our definition in Eq. (10) are as follows.

Non-negativity. Suppose m is a BPA for X and suppose $|\Omega_X| = 2$. Then, $H(m) \geq 0$. For $|\Omega_X| > 2$, $H(m)$ does *not* satisfy the non-negativity property.

Example 1. Consider a BPA m for X with $\Omega_X = \{a, b, c\}$ such that $m(\{a, b\}) = m(\{a, c\}) = m(\{b, c\}) = \frac{1}{3}$. Then Q_m is as follows: $Q_m(\{a\}) = Q_m(\{b\}) = Q_m(\{c\}) = \frac{2}{3}$, $Q_m(\{a, b\}) = Q_m(\{a, c\}) = Q_m(\{b, c\}) = \frac{1}{3}$, and $Q_m(\{a, b, c\}) = 0$. Then, $H(m) = -3 \cdot \frac{2}{3} \log_2(\frac{2}{3}) + 3 \cdot \frac{1}{3} \log_2(\frac{1}{3}) = \log_2(\frac{3}{4}) \approx -0.415$. \square

We conjecture that $H(m) \geq \log_2(\frac{n}{2(n-1)})$, where $n = |\Omega_X|$. This is based on a BPA where each of $\binom{n}{2}$ doubleton subsets has a mass of $1/\binom{n}{2}$. If the conjecture is true, $H(m)$ would be on the scale from $[\log_2(\frac{n}{2(n-1)}), \log_2(n)]$. $\lim_{n \rightarrow \infty} \log_2(\frac{n}{2(n-1)}) = -1$. Lack of non-negativity is not a serious drawback. Shannon’s definition of entropy for continuous random variables characterized by probability density functions can be negative [6].

Quasi-consonant. Suppose m is a BPA for X . If m is quasi-consonant, then $H(m) = 0$. As consonant BPAs are also quasi-consonant, $H(m) = 0$ for consonant BPAs. This property suggests that $H(m)$ is a measure of “dissonance” in m .

Maximum entropy. Suppose m is a BPA for X with state space Ω_X . Then, $H(m) \leq \log_2(|\Omega_X|)$, with equality if and only if $m = m_u$, where m_u is the Bayesian equiprobable BPA for X . This is similar to the corresponding property of Shannon’s definition for PMFs.

6 Summary and Conclusion

The most important property of our definition of entropy is the compound distributions property. Such a property is not satisfied by any of the past definitions of entropy reviewed in [3], nor by the definition proposed there. The compound distributions property is fundamental to Shannon’s definition of entropy as it constitutes the main property that characterizes Shannon’s definition.

We should also note that the compound distributions property only applies to belief functions that are constructed from marginals and conditional belief functions. Given an arbitrary joint belief function, it is not always possible to factor it into marginals and conditionals that produce the given joint. Thus, our new definition is of particular interest for the class of joint belief functions that do factor into marginals and conditionals. In particular, it applies to graphical belief functions that are constructed from directed acyclic graphs models, also known as Bayesian networks, but whose potentials are described by belief functions [1].

This is work in progress. Although we have stated Theorems 1–5 for the case where X and Y are binary, we believe these theorems hold more generally, and are currently in the process of proving them.

References

1. Almond, R.G.: Graphical Belief Modeling. Chapman & Hall, London (1995)
2. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. **38**(2), 325–339 (1967)

3. Jiroušek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. *Int. J. Approx. Reason.* **92**(1), 49–65 (2018)
4. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
5. Shafer, G.: Belief functions and parametric models. *J. R. Stat. Soc. Ser. B* **44**(3), 322–352 (1982)
6. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(379–423), 623–656 (1948)
7. Smets, P.: *Un modele mathematico-statistique simulant le processus du diagnostic medical*. Ph.D. thesis, Free University of Brussels (1978)
8. Smets, P.: Information content of an evidence. *Int. J. Man Mach Stud.* **19**, 33–43 (1983)



An Evidential K -Nearest Neighbor Classifier Based on Contextual Discounting and Likelihood Maximization

Orakanya Kanjanatarakul¹, Siwarat Kuson², and Thierry Denoeux³(✉)

¹ Faculty of Management Sciences, Chiang Mai Rajabhat University,
Chiang Mai, Thailand
orakanyaa@gmail.com

² Faculty of Economics, Maejo University, Chiang Mai, Thailand
ksiwarat@gmail.com

³ Université de Technologie de Compiègne, CNRS,
UMR 7253 Heudiasyc, Compiègne, France
tdenoeux@utc.fr

Abstract. The evidential K nearest neighbor classifier is based on discounting evidence from learning instances in a neighborhood of the pattern to be classified. To adapt the method to partially supervised data, we propose to replace the classical discounting operation by contextual discounting, a more complex operation based on as many discount rates as classes. The parameters of the method are tuned by maximizing the evidential likelihood, an extended notion of likelihood based on uncertain data. The resulting classifier is shown to outperform alternative methods in partially supervised learning tasks.

Keywords: Belief functions · Dempster-Shafer theory · Classification
Machine learning · Partially supervised learning · Soft labels

1 Introduction

Since its introduction in [2], the evidential K -nearest neighbor (EKNN) classifier has been used extensively and several variants have been developed (see, e.g., [5–8, 14] for some applications and recent developments). The EKNN classifier is based on the following simple ideas: (1) each neighbor of the pattern x to be classified is considered as a piece of evidence about the class of x , represented by a mass function; (2) each mass function is discounted based on its distance to x ; and (3) the discounted mass functions induced by the K nearest neighbors of x are combined by Dempster's rule.

In [2], the parameters used to define the discount rate as a function of distance were fixed heuristically, and the method was shown to outperform other

This research was supported by the Center of Excellence in Econometrics at Chiang Mai University.

K -nearest neighbor rules. In [15], the authors showed that the performances of the method could be further improved by learning the parameters through minimizing the mean squares error (MSE) between pignistic probabilities and class indicator variables. In [4], the EKNN rule was extended to the case where the class label of training patterns is only partially known, and described by a possibility distribution. However, the learning procedure defined in [15] cannot be straightforwardly extended to the partially labeled setting because (1) the discount rate defined in the procedure depends on the class of the neighboring pattern, and (2) combining arbitrary mass functions and computing pignistic probabilities has exponential complexity in the worst case.

In this paper, we revisit the EKNN classifier by exploiting some recent developments in the theory of belief functions: (1) The discounting operation is replaced by *contextual discounting* [9], allowing us to define one discount rate parameter per class even in the partially labeled case; and (2) instead of the MSE and pignistic probabilities, we propose to use the *conditional evidential likelihood* criterion [3, 11], which allows us to account for partial class labels in a natural way, and can be computed in linear time as a function of the number of classes.

The rest of this paper is organized as follows. The EKNN classifier and classical discounting operation are first recalled in Sect. 2. The Contextual-Discounting Evidential K -NN (CD-EKNN) classifier is then introduced in Sect. 3, and experimental results are reported in Sect. 4. Section 5 concludes the paper.

2 Background

In this section, we provide a reminder of the main notions needed in the rest of the paper. The EKNN classifier will first be recalled in Sect. 2.1, and the contextual discounting operation will be presented in Sect. 2.2.

2.1 Evidential K -NN Classifier

Consider a classification problem with c classes in $\Omega = \{\omega_1, \dots, \omega_c\}$, and a learning set $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$ of n examples (x_i, y_i) , where x_i is a p -dimensional feature vector describing example i , and $y_i \in \Omega$ is the class of that example. Let x be new pattern to be classified, and $\mathcal{N}_K(x)$ the set of its K nearest neighbors in \mathcal{L} , according to some distance d (usually, the Euclidean distance when the p features are numerical). In [2, 15], it was assumed that each neighbor $x_j \in \mathcal{N}_K(x)$ induces a simple mass function \hat{m}_j defined as

$$\hat{m}_j(\{\omega_k\}) = \beta_k(d_j)y_{jk}, \quad k = 1, \dots, c \quad (1a)$$

$$\hat{m}_j(\Omega_k) = 1 - \beta_k(d_j), \quad (1b)$$

where $y_{jk} = 1$ if $y_j = \omega_k$ and $y_{jk} = 0$ otherwise, $d_j = d(x, x_j)$ and β_k is a decreasing function, usually taken as $\beta_k = \alpha \exp(-\gamma_k d_j^2)$, where α is a coefficient in $[0, 1]$ and the γ_k 's are strictly positive scale parameters. By pooling mass

functions \widehat{m}_j induced by the K nearest neighbors of x using Dempster's rule, we get the combined mass function \widehat{m} , which summarizes the evidence about the class of x based on its K nearest neighbors.

In [15], it was proposed to leave parameter α fixed and to learn parameter vector $\gamma = (\gamma_1, \dots, \gamma_c)$ by minimizing the following error function,

$$C(\gamma) = \sum_{i=1}^n \sum_{k=1}^c (\widehat{Betp}_i(\omega_k) - y_{ik})^2, \quad (2)$$

where \widehat{Betp}_i is the pignistic probability distribution computed from mass function \widehat{m}_i obtained from the K nearest neighbors of x_i . Because this classifier is based on c learnable parameters γ_k , $k = 1, \dots, c$, it will be later referred to as the γ_k -EKNN classifier.

The idea of applying the EKNN procedure to partially labeled data $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^n$, where m_i is an arbitrary mass function that represents partial knowledge about the class of example x_i was already suggested in [2] and explored further in [4]. Indeed, mass function \widehat{m}_j in (1) can be seen as the discounted version of the certain mass function $m_j(\{y_j\}) = 1$, with discount rate $1 - \beta_k(d_j)$ if $y_j = \{\omega_k\}$. The same discounting notion can be applied whatever the form of m_j , but the discount rate can no longer depend on y_j when it is unknown. Consequently, the extension is not straightforward. Also, the combination by Dempster's rule and the calculation of the pignistic probabilities in (2) have exponential complexities for arbitrary mass functions m_i , which makes the method less attractive unless c is very small. These issues will be addressed in Sect. 3, based on the notion of contextual discounting recalled hereafter.

2.2 Contextual Discounting

Let m be a mass function on $\Omega = \{\omega_1, \dots, \omega_c\}$ and β a coefficient in $[0, 1]$. The *discounting* operation [12] with discount rate $1 - \beta$ transforms m into the following mass function:

$${}^\alpha m = \beta m + (1 - \beta)m_\?, \quad (3)$$

where $m_\?$ is the vacuous mass function defined by $m_\?(\Omega) = 1$. This operation can be justified as follows [13]. Assume that m is provided by a source that may be reliable (R) or not ($\neg R$). If the source is reliable, we adopt its opinion as ours, i.e., we set $m(\cdot|R) = m$. If it is not reliable, then it leaves us in a state of total ignorance, i.e., $m(\cdot|\neg R) = m_\?$. Furthermore, assume that we have the following mass function on $\mathcal{R} = \{R, \neg R\}$: $m_{\mathcal{R}}(\{R\}) = \beta$ and $m_{\mathcal{R}}(\mathcal{R}) = 1 - \beta$, i.e., our degree of belief that the source is reliable is equal to β . Then, combining the two mass functions $m(\cdot|R)$ (after deconditioning) and $m_{\mathcal{R}}$ yields precisely ${}^\alpha m$ in (3), after marginalizing on Ω .

In [9], the authors generalized the discounting operation using the notion of *contextual discounting*. In the corresponding refined model, $m(\cdot|R)$ and $m(\cdot|\neg R)$ are defined as before, but our beliefs about the reliability of the source are now

defined given each state in Ω , i.e., we have c conditional mass functions defined by $m_{\mathcal{R}}(\{R\}|\omega_k) = \beta_k$ and $m_{\mathcal{R}}(\mathcal{R}|\omega_k) = 1 - \beta_k$, for $k = 1, \dots, c$. Combining $m(\cdot|R)$ with mass functions $m_{\mathcal{R}}(\cdot|\omega_k)$ after deconditioning yields the following discounted mass function,

$$\beta m(A) = \sum_{B \subseteq A} m(B) \left(\prod_{\omega_k \in A \setminus B} (1 - \beta_k) \prod_{\omega_l \in \bar{A}} \beta_l \right) \quad (4)$$

for all $A \subseteq \Omega$, where $\beta = (\beta_1, \dots, \beta_c)$, and a product of terms is equal to 1 if the index set is empty. The associated contour function is

$$\beta pl(\omega_k) = 1 - \beta_k + \beta_k pl(\omega_k), \quad k = 1, \dots, c, \quad (5)$$

where pl is the contour function corresponding to m .

3 Contextual-Discounting Evidential K -NN Classifier

An alternative to the γ_k -EKNN classifier based on contextual discounting will first be defined in Sect. 3.1, and learning the parameters in this model will be addressed in Sect. 3.2.

3.1 Extending the EKNN Classifier to Partially Labelled Data

As the EKNN classifier is based on discounting, it can be readily generalized using contextual discounting. More precisely, let us assume that we have a partially labeled learning set $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^n$. (The fully supervised case is recovered when all mass functions m_i are certain). Let x be a pattern to be classified, and x_j one of its K nearest neighbors. In [4], it was proposed to generalize (1) by discounting each neighbor mass function m_j with discount rate $1 - \beta(d_j) = 1 - \alpha \exp(-\gamma d_j^2)$. We then have two learnable parameters: coefficient α and a single scale parameter γ . This rule will later be referred to as the (α, γ) -EKNN classifier.

In this paper, we propose to use contextual discounting (4) instead of classical discounting. The resulting rule, called *Contextual Discounting Evidential K -nearest neighbor* (CD-EKNN) is based on c coefficients $\beta_k(d_j)$ defined by

$$\beta_k(d_j) = \alpha \exp(-\gamma_k d_j^2), \quad k = 1, \dots, c, \quad (6)$$

and there are $c + 1$ learnable parameters $\alpha \in [0, 1]$ and $\gamma_k \geq 0$, $k = 1, \dots, c$.

Whereas the discounted mass function \hat{m}_j may have a complicated expression, its contour function can be obtained from (5) as

$$\hat{pl}_j(\omega_k) = 1 - \beta_k(d_j) + \beta_k(d_j) pl_j(\omega_k), \quad k = 1, \dots, c, \quad (7)$$

and the combined contour function after pooling the evidence of the K nearest neighbors is

$$\hat{pl}(\omega_k) \propto \prod_{x_j \in \mathcal{N}_K(x)} [1 - \beta_k(d_j) + \beta_k(d_j) pl_j(\omega_k)], \quad k = 1, \dots, c. \quad (8)$$

We note that \widehat{pl} can be computed, up to a multiplicative constant, in time proportional to the number K of neighbors and the number of c of classes. The contour function is all we need to make decisions and, as we will see in the next section, to train the classifier by maximizing the evidential likelihood criterion.

3.2 Learning

To learn the parameters $\theta = (\alpha, \gamma_1, \dots, \gamma_c)$ of the CD-EKNN classifier defined in Sect. 3.1, we propose to maximize the *evidential likelihood* function introduced in [3]. Before, we introduce the evidential likelihood for this model, let us recall the expression of the “classical likelihood” in the case of fully supervised data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$. Let \widehat{pl}_i the contour function computed for instance i based on its K nearest neighbors using (8), and let \widehat{p}_i be the probability distribution obtained from \widehat{pl}_i by normalization. The conditional likelihood (given feature vectors x_1, \dots, x_n) after observing the true class labels y_1, \dots, y_n is

$$L_c(\theta) = \prod_{i=1}^n \prod_{k=1}^c \widehat{p}_i(\omega_k)^{y_{ik}}. \quad (9)$$

In the partially supervised learning case, the learning set is of the form $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^n$, where m_i is a mass function that represents our partial knowledge of the class of x_i . An extension of the likelihood function for such uncertain data was introduced and justified in [3]. Basically, the term $\prod_{k=1}^c \widehat{p}_i(\omega_k)^{y_{ik}}$ in (9) is replaced by the expected plausibility $\sum_{k=1}^c \widehat{p}_i(\omega_k)pl_i(\omega_k)$. The *evidential likelihood* is then defined as

$$L_e(\theta) = \prod_{i=1}^n \sum_{k=1}^c \widehat{p}_i(\omega_k)pl_i(\omega_k), \quad (10)$$

We note that the evidential likelihood (10) boils down to the classical likelihood (9) when all mass functions m_i are certain, i.e., when $pl_i(\omega_k) = y_{ik}$ for all i and k . The evidential log-likelihood $\log L_e(\theta)$ can be maximized using an iterative optimization procedure such as Newton’s method.

4 Numerical Experiments

In this section, we present some results with one simulated and two real datasets, in which label uncertainty was simulated by corrupting labels with noise and representing uncertainty using suitable mass functions. The simulated data were generated from $c = 2$ Gaussian distributions with densities $\mathcal{N}(\mu_k, \sigma_k^2 I)$, where $\mu_1 = (0, 0)^T$, $\mu_2 = (1, 0)^T$, $\sigma_1^2 = 0.1I$, $\sigma_2^2 = 2I$, and I is the identity matrix. Each simulated dataset had 100 vectors from each class. The real data were the Ionosphere data ($n = 351$ instances, $p = 34$ features and $c = 2$ classes) and the Sonar data ($n = 204$, $p = 60$, $c = 2$), both from the UCI Machine Learning Repository¹.

¹ Available at <http://archive.ics.uci.edu/ml>.

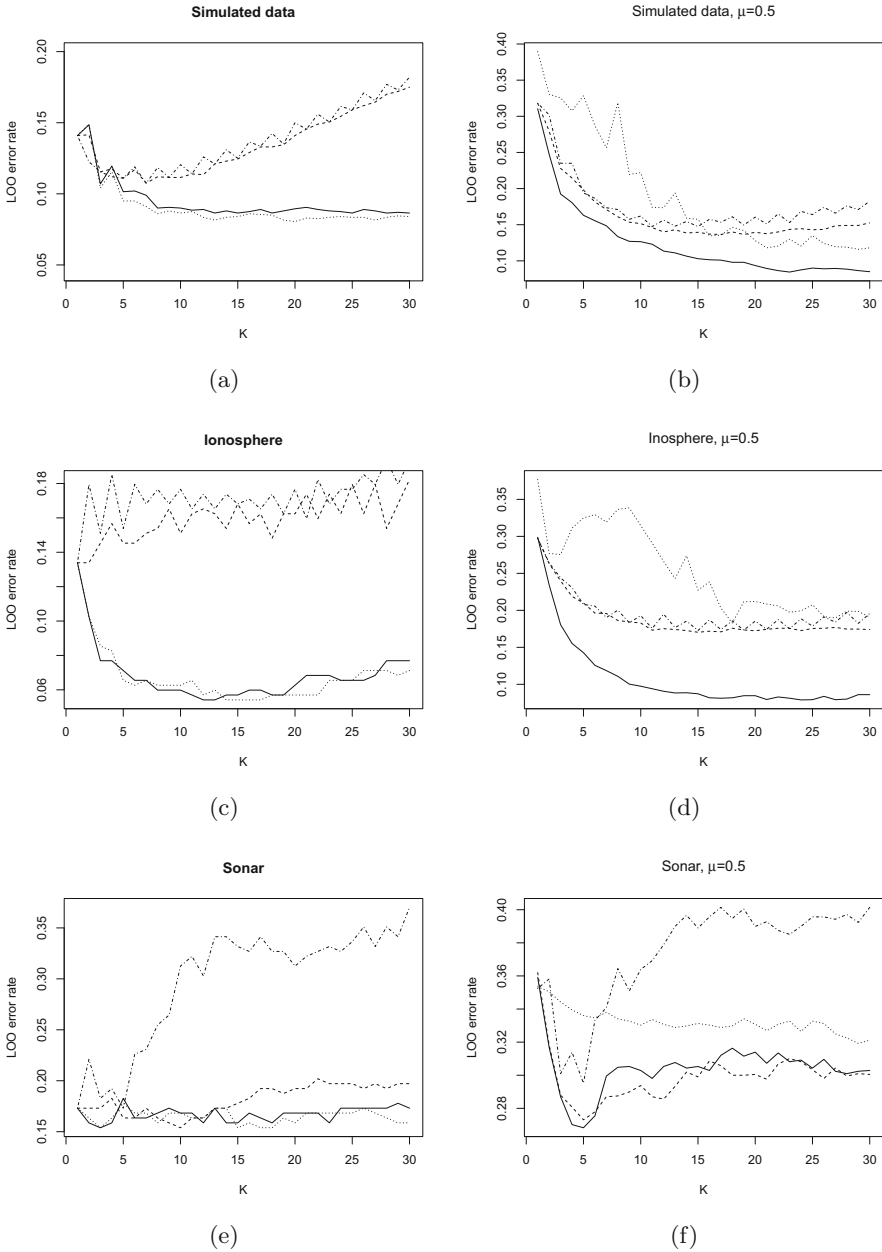


Fig. 1. Leave-one-out error rates vs. number K of neighbors for fully supervised (a, c, e) and partially supervised (b, d, f) datasets. The methods are: the CD-EKNN classifier (solid lines), the (α, γ) -EKNN classifier (dashed lines), the original γ_k -EKNN classifier (dotted lines) and the voting K -NN rule (dash-dotted lines).

Figure 1 shows the leave-one-out error rates as functions of the number K of neighbors, in two learning situations: with true class labels (Figs. 1(a), (c) and (e)), and with uncertain (soft) class labels (Figs. 1(b), (d) and (f)). To generate the uncertain labels m_i , we proceeded as in [1, 11]: for each instance i , a number p_i was generated from a beta distribution with mean $\mu = 0.5$ and variance 0.04. Then, with probability p_i , the class label y_i of instance i was replaced by y'_i picked randomly from Ω . Otherwise, we set $y'_i = y_i$. Contour function pl_i was then defined as $pl_i(\{y'_i\}) = 1$ and $pl_i(\{\omega\}) = p_i$ for all $\omega \neq y'_i$. This procedure guarantees that the soft label pl_i is all the more uncertain that the label with maximum plausibility has the more chance of being incorrect.

For each dataset and each learning situation, we considered four classifiers: (1) the (α, γ) -EKNN rule based on classical discounting and criterion (10); (2) the CD-EKNN rule with c scale parameters $\gamma_1, \dots, \gamma_c$ trained with criterion (10); (3) the original γ_k -EKNN rule recalled in Sect. 2.1, trained with criterion (2); and (4) the voting K -NN rule. As the γ_k -EKNN and voting K -NN classifiers can only handle fully supervised data with certain labels, we used the noisy labels y'_i with these classifiers, instead of the soft labels m_i .

As can be seen from Figs. 1(a), (c) and (e), the original γ_k -EKNN and CD-EKNN rules have similar performances in the fully supervised case, and they perform better than the (α, γ) -EKNN rule. On the simulated data, the (α, γ) -EKNN rule does not even outperform the voting K -NN rule (Fig. 1(a)), whereas it performs much better on the Sonar data (Fig. 1(e)).

When applied to data with soft labels, the CD-EKNN classifier clearly has the best performances. In contrast, the γ_k -EKNN and voting K -NN classifiers, which use noisy labels, perform poorly. This result confirms similar findings reported in [1, 3, 11] for parametric classifiers. The CD-EKNN classifier also performs better than the (α, γ) -EKNN rule, except on the Sonar data, for which they achieve similar error rates.

5 Conclusions

The EKNN classifier introduced in [2] and perfected in [15] has proved very efficient for fully supervised classification. Because it applies different discount rates to neighbors from different classes, the method cannot be readily extended to the partially supervised learning situation, in which we only have uncertain information about the class of learning instances. Also, it is not clear how the MSE criterion used in [15] could be generalized in the case of partially labeled data. In this paper, we have proposed a solution to this problem by replacing classical discounting with contextual discounting introduced in [9]. The underlying idea is that the reliability of the information from different neighbors depends on the class of the pattern to be classified. We also replaced the MSE by the conditional likelihood, which has already been generalized to uncertain data in [3]. The resulting CD-EKNN classifier was shown to perform very well with partially supervised data, while performing as well as the original EKNN classifier with fully supervised data.

In contrast with the original EKNN classifier, which assigns masses only to singletons and the whole frame of discernment, the CD-EKNN classifier generates more general mass functions, as a result of applying the contextual discounting operation. In future work, it will be interesting to study how masses assigned to various subsets of classes can be interpreted, and to find out if this richer information can be exploited for, e.g., classifier combination. Beyond discounting, other contextual mass correction mechanisms such as introduced in [10] could also be investigated.

References

1. Côme, E., Oukhellou, L., Dencœur, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. *Pattern Recogn.* **42**(3), 334–348 (2009)
2. Dencœur, T.: A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(05), 804–813 (1995)
3. Dencœur, T.: Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. Knowl. Data Eng.* **25**(1), 119–130 (2013)
4. Dencœur, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets Syst.* **122**(3), 47–62 (2001)
5. Guettari, N., Capelle-Laizé, A.S., Carré, P.: Blind image steganalysis based on evidential k -nearest neighbors. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 2742–2746, September 2016
6. Lian, C., Ruan, S., Dencœur, T.: An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recogn.* **48**, 2318–2327 (2015)
7. Lian, C., Ruan, S., Dencœur, T.: Dissimilarity metric learning in the belief function framework. *IEEE Trans. Fuzzy Syst.* **24**(6), 1555–1564 (2016)
8. Liu, Z.-G., Pan, Q., Dezert, J.: A new belief-based K -nearest neighbor classification method. *Pattern Recogn.* **46**(3), 834–844 (2013)
9. Mercier, D., Quost, B., Dencœur, T.: Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Inf. Fusion* **9**(2), 246–258 (2008)
10. Pichon, F., Mercier, D., Lefèvre, E., Delmotte, F.: Proposition and learning of some belief function contextual correction mechanisms. *Int. J. Approx. Reason.* **72**, 4–42 (2016)
11. Quost, B., Dencœur, T., Li, S.: Parametric classification with soft labels using the evidential em algorithm: linear discriminant analysis versus logistic regression. *Adv. Data Anal. Classif.* **11**(4), 659–690 (2017)
12. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
13. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**, 191–243 (1994)
14. Su, Z.-G., Denoeux, T., Hao, Y.-S., Zhao, M.: Evidential K -NN classification with enhanced performance via optimizing a class of parametric conjunctive t -rules. *Knowl.-Based Syst.* **142**, 7–16 (2018)
15. Zouhal, L.M., Dencœur, T.: An evidence-theoretic k -NN rule with parameter optimization. *IEEE Trans. Syst. Man Cybern. C* **28**(2), 263–271 (1998)



Measuring Market Performance with Stochastic Demand: Price of Anarchy and Price of Uncertainty

Costis Melolidakis¹, Stefanos Leonardos^{1(✉)}, and Constandina Koki²

¹ National and Kapodistrian University of Athens, 157 84 Athens, Greece
{cmelol,sleonardos}@math.uoa.gr

² Athens University of Economics and Business, 104 34 Athens, Greece
kokiconst@aueb.gr

Abstract. Globally operating suppliers face the rising challenge of wholesale pricing under scarce data about retail demand, in contrast to better informed, locally operating retailers. At the same time, as local businesses proliferate, markets congest and retail competition increases. To capture these strategic considerations, we employ the classic Cournot model and extend it to a two-stage supply chain with an upstream supplier who operates under demand uncertainty and multiple downstream retailers who compete over quantity. The supplier's belief about retail demand is modeled via a continuous probability distribution function F . If F has the *decreasing generalized mean residual life* property, then the supplier's optimal pricing policy exists and is the unique fixed point of the *mean residual life* function. We evaluate the realized *Price of Uncertainty* and show that there exist demand levels for which market performs better when the supplier prices under demand uncertainty. In general, performance worsens for lower values of realized demand. We examine the effects of increasing competition on supply chain efficiency via the realized *Price of Anarchy* and complement our findings with numerical results.

Keywords: Nash equilibrium · Generalized mean residual life
Continuous beliefs · Price of Uncertainty · Price of Anarchy

1 Introduction

The increasingly present trend of geographically distributed markets affects supply chain performance in unexpected ways. Internationally operating suppliers procure retailers via internet platforms or intricate networks with information latency. Consumer data that is easily accessible to the retailers due to their proximity to the market, may often not be available to their international suppliers. Concurrently, and aided by new technologies, local retail businesses are sprouting at a rapid pace. These trends give rise to new information and competition structures between downstream members (retailers) and their upstream contemporaries (suppliers) in modern supply chains.

The questions that rise in this changing environment, mainly concern the issues of market efficiency. How does the market perform when the supplier prices without knowing the retailers willingness-to-pay for his product? Do competing retailers have incentives to reveal private information to the supplier that they may have about retail demand? To capture these considerations and study this emerging phenomenon, in [7,8], we employ the classic Cournot model of competition and extend it to the following two-stage game: in the first-stage (acting as a Stackelberg leader), a revenue-maximizing supplier sets the wholesale price of a product under incomplete information about market demand. Demand or equivalently, the supplier's belief about it, is modeled via a continuous probability distribution. In the second-stage, the competing retailers observe wholesale price and realized market demand and engage in a classic Cournot competition. Retail price is determined by an affine inverse demand function.

Classic models, see e.g., [4,5,9,10], study market efficiency when demand is realized after the strategic decisions of all supply chain members – wholesale pricing and retailers' orders. In contrast, performance of markets in which uncertainty is resolved at an intermediate stage, has not been yet properly understood.

Contributions – Outline: Based on the equilibrium analysis in [7], the present paper aims to fill this gap by following the methodology of [5]. To measure the effects of demand uncertainty and second-stage competition on market performance and efficiency, we modify the tools of Price of Anarchy, as defined in [11] and Price of Uncertainty, c.f. [1], to account for *realized values* of demand. In Sect. 2, we provide the model description and in Sect. 3, the existing results from [7] on which the current analysis is based. Our findings, both analytical and numerical are presented in Sect. 4 and summarized in Sect. 5.

2 The Model

An upstream supplier (or manufacturer) produces a single homogeneous good at constant marginal cost, normalized to 0, and sells it to a set of $N = \{1, 2, \dots, n\}$ downstream retailers. The supplier has ample quantity to cover any possible demand and his only decision variable is the wholesale price r . The retailers observe r and the market demand α and choose simultaneously and independently their order-quantities $q_i(r | \alpha), i \in N$. They face no uncertainty about the demand and the quantity that they order from the supplier is equal to the quantity that they sell to the market (in equilibrium). The retail price is determined by an affine demand function $p = (\alpha - q(r))^+$, where α is the *demand parameter* or *demand level* and $q(r) := \sum_{i=1}^n q_i(r)$. Contrary to the retailers, we assume that at the point of his decision, the supplier has incomplete information about the actual market demand.

Game-Theoretic Formulation: This supply chain can be represented as a two-stage game, in which the supplier acts in the first and the retailers in the

second stage. A strategy for the supplier is a price $r \geq 0$ and a strategy for retailer i is a function $q_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, which specifies the quantity that retailer i will order for any possible cost r . Payoffs are determined via the strategy profile $(r, \mathbf{q}(r))$, where $\mathbf{q}(r) = (q_i(r))_{i=1}^n$. Given cost r , the profit function $\pi_i(\mathbf{q}(r) | r)$ or simply $\pi_i(\mathbf{q} | r)$, of retailer $i \in N$, is $\pi_i(\mathbf{q} | r) = q_i(\alpha - q)^+ - rq_i$. For a given value of α , the supplier's profit function, π_s is $\pi_s(r | \alpha) = rq(r)$ for $0 \leq r < \alpha$, where $q(r)$ depends on α via $\pi_i(\mathbf{q} | r)$.

Continuous Beliefs: To model the supplier's uncertainty about retail demand, we assume that after his pricing decision, but prior to the order-decisions of the retailers, a value for α is realized from a continuous distribution F , with finite mean $\mathbb{E}\alpha < +\infty$ and nonnegative values, i.e. $F(0) = 0$. Equivalently, F can be thought of as the supplier's belief about the demand parameter and, hence, about the retailers' willingness-to-pay his price. We will use the notation $\bar{F} := 1 - F$ for the survival function and $\alpha_L := \sup\{r \geq 0 : F(r) = 0\} \geq 0$, $\alpha_H := \inf\{r \geq 0 : F(r) = 1\} \leq +\infty$ for the support of F respectively. The instance $\alpha_L = \alpha_H$ corresponds to the reference case of deterministic demand. In any other case, i.e., for $\alpha_L < \alpha_H$, the supplier's payoff function π_s becomes stochastic: $\pi_s(r) = \mathbb{E}\pi_s(r | \alpha)$. All of the above are assumed to be common knowledge among the participants in the market (the supplier and the retailers).

3 Existing Results

We consider only subgame perfect equilibria, i.e. strategy profiles $(r, \mathbf{q}(r))$ such that $\mathbf{q}(r)$ is an equilibrium in the second stage and $q_i(r)$ is a best response against any r . The equilibrium behavior of this market has been analyzed in [7]. In the reference case of deterministic demand, i.e., for $\alpha_L = \alpha_H$, each retailer $i = 1, 2, \dots, n$ orders quantity $q_i^*(r | \alpha) = \frac{1}{n+1}(\alpha - r)^+$. Hence, the supplier's payoff on the equilibrium path becomes $\pi_s(r | \alpha) = rq^*(r | \alpha) = \frac{n}{n+1}r(\alpha - r)^+$, for $0 \leq r$. Maximization of π_s with respect to r yields that the complete information two-stage game has a unique subgame perfect Nash equilibrium, under which the supplier sells with optimal price $r^*(\alpha) = \frac{1}{2}\alpha$ and each of the retailers orders quantity $q_i^*(r) = \frac{1}{n+1}(\alpha - r)^+$, $i = 1, 2, \dots, n$. To proceed with the equilibrium representation in the stochastic case, we first introduce some notation.

Generalized Mean Residual Life: Let $\alpha \sim F$ be a nonnegative random variable with finite expectation $\mathbb{E}\alpha < +\infty$. The *mean residual life (MRL)* function $m(r)$ of α is defined as

$$m(r) := \mathbb{E}(\alpha - r | \alpha > r) = \frac{1}{\bar{F}(r)} \int_r^\infty \bar{F}(u) du, \quad \text{for } r < \alpha_H$$

and $m(r) := 0$, otherwise, see, e.g., [3]. In [7], we introduced the *generalized mean residual life (GMRL)* function $\ell(r)$, defined as $\ell(r) := \frac{m(r)}{r}$, for $0 < r < \alpha_H$,

in analogy to the *generalized failure rate (GFR)* function $g(r) := rh(r)$, where $h(r) := f(r)/\bar{F}(r)$ denotes the hazard rate of F and the *increasing generalized failure rate (IGFR)* unimodality condition, defined in [5] and studied in [2, 6]. If $\ell(r)$ is *decreasing*, then F has the *(DGMRL) property*. The relationship between the (IGFR) and (DGMRL) classes of random variables is studied in [7].

Market Equilibrium: Using this terminology, we can express the supplier’s optimal pricing strategy in terms of the MRL function and formulate sufficient conditions on the demand distribution, under which a subgame perfect equilibrium exists and is unique.

Theorem 1 ([7]). *Assume that the supplier’s belief about the unknown, non-negative demand parameter, α , is represented by a continuous distribution F , with support inbetween α_L and α_H with $0 \leq \alpha_L < \alpha_H \leq \infty$.*

(a) *If an optimal price r^* for the supplier exists, then r^* satisfies the fixed point equation*

$$r^* = m(r^*) \tag{1}$$

(b) *If F is strictly DGMRL and $\mathbb{E}\alpha^2$ is finite, then in equilibrium, the optimal price r^* of the supplier exists and is the unique solution of (1).*

4 Supply Chain Efficiency

To study the degree in which demand uncertainty affects the realized market profits, we fix a realized demand level α and compare the individual realized profits of the supplier and each retailer between the scenario in which the supplier prices before demand realization and the scenario in which the supplier prices after demand realization. For clarity, the results are summarized in Table 1.

Table 1. Wholesale price in equilibrium and realized profits when the supplier prices under demand uncertainty (left column) and under deterministic demand (right column).

	Upstream demand for supplier	
	Uncertain $\alpha \sim F$	Deterministic α
Equilibrium wholesale price	$r^* = m_F(r^*)$	$\alpha/2$
	Realized profits in equilibrium	
Supplier	$\Pi_s^U = \frac{n}{n+1} (\alpha - r^*)^+$	$\Pi_s^D = \frac{n}{n+1} (\alpha/2)^2$
Retailer i	$\Pi_i^U = \frac{1}{(n+1)^2} ((\alpha - r^*)^+)^2$	$\Pi_i^D = \frac{n}{(n+1)^2} (\alpha/2)^2$
Aggregate	$\Pi_{\text{Agg}}^U = \Pi_s^U + \sum_{i=1}^n \Pi_i^U$	$\Pi_{\text{Agg}}^D = \Pi_s^D + \sum_{i=1}^n \Pi_i^D$

4.1 Price of Uncertainty

By Table 1, for each retailer, we have that $\frac{1}{(n+1)^2} \left((\alpha - r^*)^+ \right)^2 \geq \frac{1}{(n+1)^2} \left(\frac{\alpha}{2} \right)^2$ for all values of $\alpha \geq 2r^*$. This implies that for larger values of the realized demand, the retailers are better off if the supplier prices under demand uncertainty. In contrast, the supplier is never better off when he prices under demand uncertainty, as is intuitively expected. Indeed $\frac{n}{n+1} r^* (\alpha - r^*)^+ \leq \frac{n}{n+1} (\alpha/2)^2$ for all values of α , with equality if and only if $\alpha = 2r^*$. The ratio of the supplier's realized profit in the scenario with demand uncertainty to the scenario without demand uncertainty is equal to $4 \cdot \frac{r^*}{\alpha} \left(1 - \frac{r^*}{\alpha} \right)$ and hence it has the shape shown in Fig. 1, independently of the underlying demand distribution.

Similar findings are obtained when we compare the market's aggregate realized profits (supplier and retailers) between these two scenarios. This is accomplished via the ratio of aggregate realized market profits under stochastic demand to the aggregate realized market profits under deterministic demand, which we term the realized *Price of Uncertainty* (PoU), motivated by a similar notion that is studied in [1]. Specifically,

$$\text{PoU} := \sup_{F \in \mathcal{G}} \sup_{\alpha} \left\{ \frac{\Pi_{\text{Agg}}^U}{\Pi_{\text{Agg}}^D} \right\} = \sup_{F \in \mathcal{G}} \sup_{\alpha} \left\{ \frac{\Pi_s^U + \sum_{i=1}^n \Pi_i^U}{\Pi_s^D + \sum_{i=1}^n \Pi_i^D} \right\}$$

in which we restrict attention to the class \mathcal{G} of nonnegative DGMRL random variables to retain equilibrium uniqueness. Intuitively, one expects the system to perform worse under demand uncertainty which translates to PoU being bounded above by 1. However, this is not the case as the next Theorem states.

Theorem 2. *The realized PoU of the stochastic market is given by $\text{PoU} = 1 + \mathcal{O}(n^{-2})$, independently of the underlying demand distribution. The upper bound is attained for realized demand $\alpha^* = \frac{n}{n-1} \cdot 2r^*$.*

Proof. By Table 1, a direct substitution yields that the inner ratio is equal to

$$\frac{\Pi_s^U + \sum_{i=1}^n \Pi_i^U}{\Pi_s^D + \sum_{i=1}^n \Pi_i^D} = \frac{\frac{n}{n+1} r^* (\alpha - r^*)^+ + n \left(\frac{1}{n+1} (\alpha - r^*)^+ \right)^2}{\frac{n}{n+1} \left(\frac{\alpha}{2} \right)^2 + \frac{n}{(n+1)^2} \left(\frac{\alpha}{2} \right)^2}$$

Hence, $\text{PoU} = \sup_{F \in \mathcal{G}} \sup_{\alpha} \left\{ \frac{4}{(n+2)\alpha^2} (\alpha - r^*)^+ (\alpha + nr^*) \right\}$. For realized demand $\alpha < r^*$, there is a stockout and the market operates worst under demand uncertainty. However, for realized demand values $\alpha > r^*$, the aggregate market profits of the supplier and the retailers may be larger if the supplier prices under

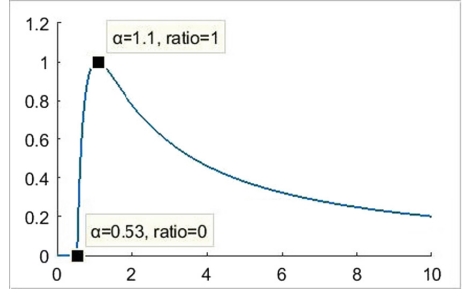


Fig. 1. Ratio of the supplier's realized profits with and without demand uncertainty for $\alpha \sim \text{Weibull}(1, 2)$.

demand uncertainty. To see this, we take the partial derivative of the previous ratio with respect to α

$$\frac{\partial}{\partial \alpha} \left(\frac{4}{(n+2)\alpha^2} (\alpha - r^*)^+ (\alpha + nr^*) \right) = \frac{4r^*}{(n+2)\alpha^3} (2nr^* - \alpha(n-1))$$

which shows that the ratio is increasing on $[r^*, \frac{2n}{n-1}r^*)$, and decreasing thereafter. The ratio is maximized for $\alpha = \frac{2n}{n-1}r^*$, yielding a value of $1 + \frac{1}{n^2+2n}$, which does not depend on the underlying distribution F and which is larger than 1 for any number n of competing second-stage retailers. \square

The values for which the ratio exceeds 1, depend on n . Specifically, for $n \geq 3$, we have that $\frac{4}{(n+2)\alpha^2} (\alpha - r^*)^+ (\alpha + nr^*) \geq 1$ for values of α in $[2r^*, \frac{2n}{n-2}r^*]$. In this case, the upper bound decreases to $2r^*$ as $n \rightarrow \infty$. For $n = 2$, the upper bound is equal to infinity, i.e., the range of α for which the ratio exceeds 1 is equal to $[2r^*, +\infty)$. In all cases, the lower bound is independent of n . Finally, by taking the partial derivative with respect to n , we find that the PoU is nondecreasing in n for realized values of α in $[r^*, 2r^*]$ and decreasing in n thereafter, again independently of the underlying demand distribution. This is illustrated in Fig. 2.

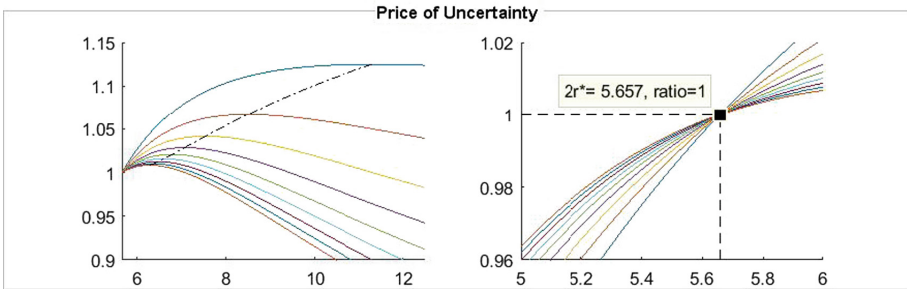


Fig. 2. Ratio of the aggregate market realized profits with and without demand uncertainty for $n = 2$ to $n = 10$ with $\alpha \sim \text{Gamma}(2, 2)$. The dashed curve in the left panel passes through the points $\alpha^* = \frac{n}{n-1} \cdot 2r^*$ on which the PoU is attained for each n . The curves are decreasing in n , i.e. the highest curve corresponds to $n = 2$ and the lowest to $n = 10$. The right panel shows the behavior of the curves in a neighborhood of their intersection point, $2r^* \approx 5.657$. Prior to the intersection, the ratio is increasing in n , whereas after the intersection the ratio is decreasing in n .

4.2 Price of Anarchy

As a benchmark, we will first determine the equilibrium behavior and performance of an integrated supply chain. The integrated firms' decision variable is now the retail price r , and hence its expected profit π_{Int} is given by $\pi_{\text{Int}}(r) = r\mathbb{E}(\alpha - r)^+ = rm(r)\bar{F}(r)$. By the same argument as in the proof of Theorem 1, π_{Int} is maximized at $r^* = m(r^*)$. In particular, the equilibrium price of both the

integrated and non-integrated supplier is the same. Hence, the integrated firm's realized profit in equilibrium is equal to $\Pi_{\text{Int}}^U(r^* | \alpha) = r^* (\alpha - r^*)^+$ (Table 2).

In a similar fashion to [11], we define the realized *Price of Anarchy (PoA)* of the system as the worst-case ratio of the realized profit of the centralized supply chain, Π_{Int}^U , to the realized aggregate profit of the decentralized supply chain, $\Pi_{\text{Dec}}^U := \Pi_s^U + \sum_{i=1}^n \Pi_i^U$. Again, to retain equilibrium uniqueness, we restrict attention to the class \mathcal{G} of nonnegative DGMRL random variables. If the realized demand α is less than r^* , then both the centralized and decentralized chains make 0 profits. Hence, we define the PoA as: $\text{PoA} := \sup_{F \in \mathcal{G}} \sup_{\alpha > r^*} \left\{ \frac{\Pi_{\text{Int}}^U}{\Pi_{\text{Dec}}^U} \right\}$. We then have

Table 2. Realized profits for the integrated firm under the two scenarios. The equilibrium wholesale prices remain the same as in the decentralized market, cf. Table 1.

	Realized profits in equilibrium	
	Uncertain demand $\alpha \sim F$	Deterministic demand α
Integrated firm	$\Pi_{\text{Int}}^U = r^* (\alpha - r^*)^+$	$\Pi_{\text{Int}}^D = (\alpha/2)^2$

Theorem 3. *The realized PoA of the stochastic market is given by $\text{PoA} = 1 + 1/n$ independently of the underlying demand distribution. The upper bound is asymptotically attained for $\alpha \searrow r^*$.*

Proof. By a direct substitution in the definition of PoA, the inner term equals $\frac{(n+1)^2}{n} \cdot \left(n + \frac{\alpha}{r^*}\right)^{-1}$. Since $\left(n + \frac{\alpha}{r^*}\right)^{-1}$ decreases in the ratio α/r^* , the inner sup is attained asymptotically for $\alpha \searrow r^*$. Hence,

$$\text{PoA} = \sup_{F \in \mathcal{G}} \left\{ \frac{(n+1)^2}{n} \cdot (n+1)^{-1} \right\} = 1 + \frac{1}{n} \quad (2)$$

Theorem 3 implies that the market becomes less efficient in the worst-case scenario, i.e., for a realized demand $\alpha \searrow r^*$, as the number of downstream retailers increases. In general, as can be directly inferred by partial differentiation with respect to n , for realized values of $\alpha < 2r^*$, the inner term of the sup expression in (2) is decreasing in n . For realized values of $\alpha \geq 2r^*$, the

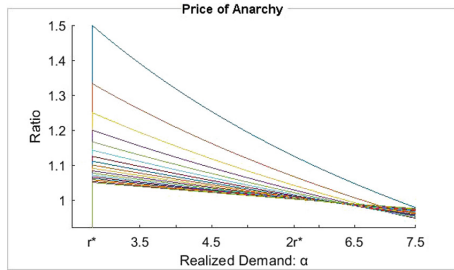


Fig. 3. Ratio of the integrated firm's to the decentralized market's aggregate profits for $n = 2, \dots, 20$ with $\alpha \sim \text{Gamma}(2, 2)$. For each n the realized PoA is attained as $\alpha \searrow r^* \approx 2.83$. For $n(\alpha - 2r^*) \leq \alpha$, the curves are nonincreasing in n , which results in the non-linearity (with respect to n) for values of α in $[2r^*, 3r^*]$.

ratio is increasing in n when $n \geq \alpha/(\alpha - 2r^*)$ and decreasing in n otherwise. These findings are shown graphically in Fig. 3.

Finally, a similar calculation yields that the PoA of the deterministic market is equal to $1 + \mathcal{O}(n^{-2})$. The realized demand α simplifies in the inner ratio and hence this upper bound is constant and independent of the demand level. Notably, the PoA in the deterministic market is equal to PoU in the stochastic market, cf. Theorem 2.

5 Conclusions

The present study complements the findings of [7, 8], by focusing to the effects of demand uncertainty on market efficiency. Based on the realized market profits, we measured the effects of uncertainty via the realized Price of Uncertainty. Counterintuitively, there exist demand levels for which the retailers' and the market's aggregate profits are higher when the supplier prices under demand uncertainty. This is achieved in expense of the supplier's welfare who is always better off under deterministic demand. The realized Price of Anarchy revealed that for any demand level, the integrated chain performs better – in terms of efficiency – as the number of competing retailers increases. Upper bounds of inefficiency are attained for lower values of realized demand. Despite these intuitions, the present analysis is limited in extent. Price differentiation and mechanisms that will incentivize retailers to honestly reveal their private information about demand, constitute promising lines of ongoing research.

References

1. Balcan, M.-F., Blum, A., Mansour, Y.: The price of uncertainty. *ACM Trans. Econ. Comput.* **1**(3), 15:1–15:29 (2013). <https://doi.org/10.1145/2509413.2509415>
2. Banciu, M., Mirchandani, P.: Technical note - new results concerning probability distributions with increasing generalized failure rates. *Oper. Res.* **61**(4), 925–931 (2013). <https://doi.org/10.1287/opre.2013.1198>
3. Belzunce, F., Martinez-Riquelme, C., Mulero J.: Univariate stochastic orders. In: *An Introduction to Stochastic Orders*, Chap. 2. Academic Press (2016). <https://doi.org/10.1016/B978-0-12-803768-3.00002-8>
4. Herweg, F.: The expectation-based loss-averse newsvendor. *Econ. Lett.* **120**(3), 429–432 (2013). <https://doi.org/10.1016/j.econlet.2013.05.035>
5. Lariviere, M., Porteus, E.: Selling to the newsvendor: an analysis of price-only contracts. *Manuf. Serv. Oper. Manag.* **3**(4), 293–305 (2001). <https://doi.org/10.1287/msom.3.4.293.9971>
6. Lariviere, M.: A note on probability distributions with increasing generalized failure rates. *Oper. Res.* **54**(3), 602–604 (2006). https://doi.org/10.1007/978-1-4615-4949-9_8
7. Leonardos, S., Melolidakis, C.: Selling to cournot oligopolists: pricing under uncertainty & generalized mean residual life (2017). <https://arxiv.org/abs/1709.09618>
8. Leonardos, S., Melolidakis, C.: Comparative Statics via Stochastic Orderings in a Two-Echelon Market with Upstream Demand Uncertainty. *AIRO Springer Series* (2018, forthcoming). <https://arxiv.org/abs/1803.03451>

9. Lu, Y., Simchi-Levi, D.: On the unimodality of the profit function of the pricing newsvendor. *Prod. Oper. Manag.* **22**, 615–625 (2013). <https://doi.org/10.1111/j.1937-5956.2012.01419.x>
10. Mandal, P., Kaul, R., Jain, T.: Stocking and pricing decisions under endogenous demand and reference point effects. *Eur. J. Oper. Res.* **264**(1), 181–199 (2018). <https://doi.org/10.1016/j.ejor.2017.05.053>
11. Perakis, G., Roels, G.: The price of anarchy in supply chains: quantifying the efficiency of price-only contracts. *Manag. Sci.* **53**(8), 1249–1268 (2007). <https://doi.org/10.1287/mnsc.1060.0656>



On the Conflict Measures Agreed with the Combining Rules

Alexander Lepskiy^(✉) 

Higher School of Economics, 20 Myasnitskaya Ulitsa, Moscow 101000, Russia
alex.lepskiy@gmail.com

Abstract. The conflict measures induced by the conjunctive and disjunctive combining rules are studied in this paper in the framework of evidence theory. The coherence of conflict measures with combining rules is introduced and studied. In addition, the structure of conjunctive and disjunctive conflict measures is studied in the paper. In particular, it is shown that the metric and entropy components can be distinguished in such measures. Moreover, these components are changed differently after combining of the bodies of evidence.

Keywords: Conflict measure · Evidence theory · Combining rule

1 Introduction

Various factors must be considered when deciding about using of combining rules in the framework of evidence theory [3, 15]. The value of a conflict measure between bodies of evidence is the important characteristic when deciding about expediency of use a particular rule. In the recent years, the study of a conflict measures has been increasingly developing into an independent research area. Axiomatics and various approaches to the evaluation of the conflict between the bodies of evidence (external conflict) were considered in [1, 4, 8, 12, 13]. The notion of internal conflict of evidence studied in [2, 10, 11, 14]. But we are considering only the external conflict in this paper. The choice of a specific measure for estimation of a conflict depends on a solvable problem. For example, if we estimate conflict between bodies of evidence with the aim of decision making about combining of evidence, then the conflict measure must be agreed in some sense with the combining rule. So Dempster's rule of combination agrees naturally with a conjunctive conflict measure. The conditions of agreement for the other combining rules are not obvious. In the given paper we study the conflict measures that are induced by conjunctive and disjunctive combining rules. The link of consistency conditions with axioms of conflict measure is studied.

In addition, the structure of conjunctive and disjunctive conflict measures studied in this paper too. In particular, we showed that it is possible to allocate the metric and entropic components in such measures. Moreover, these components are changed in different ways when the bodies of evidence are aggregated.

The main aim of this paper consists in the study of some factors (the choice of a conflict measure, the consistency conditions with combining rules, the entropy of evidence, etc.) that should be considered when we make a decision on combining of bodies of evidence.

The structure of the remainder of the paper is as follows. First, in Sect. 2, we shall recall the basic concepts of evidence theory. Axioms of a conflict measure will be discussed in Sect. 3. The conflict measures that are induced by conjunctive and disjunctive combining rules are considered in Sect. 4. The notion of coherence of conflict measures and combining rules is introduced in Sect. 5. In Sect. 6, we showed that the metric and entropic components can be allocated in the conjunctive and disjunctive conflict measures. The change of metric and entropic components after combination bodies of evidence is discussed in Sect. 7. Finally, some conclusions are presented in Sect. 8.

2 Basic Definitions and Notations of Evidence Theory

We shall recall the basic concepts of evidence theory [3, 15]. Let X be a finite set and 2^X be a powerset of X . The mass function $m : 2^X \rightarrow [0, 1]$ is considered and $\sum_{A \subseteq X} m(A) = 1$. The value $m(A)$ characterizes the relative part of evidence that the actual alternative from X belongs to set $A \in 2^X$. The subset $A \in 2^X$ is called a focal element, if $m(A) > 0$. Let $\mathcal{A} = \{A_i\}$ be a set of all focal elements of evidence. The pair $F = (\mathcal{A}, m)$ is called a body of evidence. Let $\mathcal{F}(X)$ be a set of all body of evidence on X .

If $\mathcal{A} = \{A\}$, then $F_A = (\mathcal{A}, m) = (A, 1)$ is called a categorical body of evidence. In particular F_X is called a vacuous body of evidence. If $F_j = (\mathcal{A}_j, m_j) \in \mathcal{F}(X)$, $0 \leq \alpha_j \leq 1$, $j = 1, \dots, n$ and $\sum_{j=1}^n \alpha_j = 1$, then $F = (\mathcal{A}, m) \in \mathcal{F}(X)$, where $\mathcal{A} = \bigcup_{j=1}^n \mathcal{A}_j$, $m(A) = \sum_{j=1}^n \alpha_j m_j(A)$. In this case, we will write $F = \sum_{j=1}^n \alpha_j F_j$. In particular, any body of evidence $F = (\mathcal{A}, m)$ can be represented as $F = \sum_{A \in \mathcal{A}} m(A) F_A$.

Let we have two bodies of evidence $F_1 = (\mathcal{A}_1, m_1)$ and $F_2 = (\mathcal{A}_2, m_2)$ which represent two information sources. The different combining rules R are considered in evidence theory: $R : \mathcal{F}(X) \times \mathcal{F}(X) \rightarrow \mathcal{F}(X)$. For example, the non-normalized conjunctive rule $D_0(F_1, F_2)$ is considered

$$m^{D_0}(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad A \in 2^X.$$

The value $K^D(F_1, F_2) = m^{D_0}(\emptyset)$ characterizes the amount of conflict between two sources of information (but not only, see [12]) described by the bodies of evidence F_1 and F_2 . We call the value $K^D(F_1, F_2) = m^{D_0}(\emptyset)$ the conjunctive conflict measure. If $K^D \neq 1$, then the classical Dempster rule for combining of two evidence can be defined: $m^D(A) = \frac{1}{1-K^D} m^{D_0}(A)$, $A \neq \emptyset$, $m^D(\emptyset) = 0$. The conflict management in conjunctive rule was discussed in [9].

Dubois and Prade's disjunctive consensus rule is a dual rule to Dempster's rule in some sense. This rule is defined by a formula [6]:

$$m^{DP}(A) = \sum_{B \cup C = A} m_1(B) m_2(C), \quad A \in 2^X.$$

In [7] a mixed conjunctive and disjunctive rule was discussed.

The negation (or complement) $\bar{F} = (\bar{\mathcal{A}}, \bar{m})$ of a body of evidence $F = (\mathcal{A}, m)$ is defined as $\bar{\mathcal{A}} = \{\bar{A} : A \in \mathcal{A}\}$ and $\bar{m}(A) = m(\bar{A}) \forall A \in \bar{\mathcal{A}}$, where \bar{A} denotes the complement of A [5]. Note that if we have $F = \sum_{A \in \mathcal{A}} m(A)F_A$, then we have $\bar{F} = \sum_{A \in \mathcal{A}} m(A)F_{\bar{A}}$. The duality relation is true for the non-normalized conjunctive rule and disjunctive consensus rule by analogy with De Morgan's law [5]:

$$\overline{D_0(F_1, F_2)} = DP(\bar{F}_1, \bar{F}_2). \quad (1)$$

We shall consider also the dual body of evidence $F^{(-)} = (\bar{\mathcal{A}}, m^{(-)})$ with respect to body of evidence $F = (\mathcal{A}, m)$, where $m^{(-)}(\bar{A}) = \frac{1}{N-1}(1 - m(A)) \forall A \in \mathcal{A}$, $N = |\mathcal{A}| > 1$.

3 Axioms of Conflict Measures

In general, it is desirable that the conflict measure $K(F_1, F_2)$ between bodies of evidence satisfies the following conditions (axioms) [1, 4, 13]:

- A1: $0 \leq K(F_1, F_2) \leq 1$ for all $F_1, F_2 \in \mathcal{F}(X)$ (non-negativity and normalization);
- A2: $K(F_1, F_2) = K(F_2, F_1)$ for all $F_1, F_2 \in \mathcal{F}(X)$ (symmetry);
- A3: $K(F, F) = 0$ for all $F \in \mathcal{F}(X)$ (nilpotency);
- A4: $K(F', F) \geq K(F'', F)$, if $F' = (\mathcal{A}', m)$, $F'' = (\mathcal{A}'', m)$, where $\mathcal{A}' = \{A'_i\}$, $\mathcal{A}'' = \{A''_i\}$ and $A'_i \subseteq A''_i$ for all i and $F \in \mathcal{F}(X)$ (antimonotonicity with respect to imprecision of evidence);
- A5: $K(F_X, F) = 0$ for all $F \in \mathcal{F}(X)$ (ignorance is bliss [4]);
- A6: $K(F_A, F_B) = 1$, if $A \cap B = \emptyset$.

Furthermore, if we assume that the empty set can be a focal element (the value $m(\emptyset)$ can be interpreted as the degree of confidence in the fact that the true alternative $x \notin X$), then we assume that the axioms A3 and A5 are satisfied for all $F \in \mathcal{F}(X) \setminus \{F_\emptyset\}$ and we will also consider the following axiom:

- A7: $K(F_\emptyset, F) = 1$ for all $F \in \mathcal{F}(X) \setminus \{F_\emptyset\}$.

The other axioms for conflict measures are also considered (see, e.g., [4]). We note that some axioms (for example, A4 and A6) are consistent with the conjunctive combining rule (see Sect. 5).

4 Conflict Measures Induced by Conjunctive and Disjunctive Combining Rules

Let us assume that the information from the two sources is described by means of two bodies evidence $F_1 = (\mathcal{A}_1, m_1)$ $F_2 = (\mathcal{A}_2, m_2)$. Then $K^D(F_1, F_2)$ can be considered as a conflict measure induced by conjunctive rule. This measure satisfies the axioms A1, A2, A4–A7.

Various conflict measures induced by the disjunctive consensus rule can be considered. These measures can satisfy certain axioms of the conflict measure.

Below we consider the following conflict measures induced by a disjunctive rule (we will call them disjunctive conflict measures):

$$K_1^{DP}(F_1, F_2) = \sum_{B \cup C = X} m_1(B)m_2(C), \quad K_2^{DP}(F_1, F_2) = 1 - K_1^{DP}(F_1, F_2).$$

Note that the measure $K_1^{DP}(F_1, F_2)$ satisfies only axioms A1, A2 and A7 (and the particular case of condition A6: $K(F_A, F_{\bar{A}}) = 1$). But the measure $K_2^{DP}(F_1, F_2)$ satisfies axioms A1, A2, A4, A5. The following relationship between conjunctive and disjunctive conflict measures is true. This relationship reflects the duality relation (1).

Proposition 1. $K_1^{DP}(F_1, F_2) = K^D(\bar{F}_1, \bar{F}_2)$.

The simple relations are true for the conjunctive and disjunctive conflict measures on the disjoint belief structure.

Proposition 2. *If $F_1 = (\mathcal{A}, m_1)$ and $F_2 = (\mathcal{A}, m_2)$, where $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$ ($A' \neq A''$), then: (1) $K_1^{DP}(F_1, F_2^{(-)}) = \frac{1}{N-1}K^D(F_1, F_2)$; (2) $K_2^{DP}(F_1, \bar{F}_2) = K^D(F_1, F_2)$.*

Proposition 3. *If $F_1 = (\mathcal{A}_1, m_1)$, $F_2 = (\mathcal{A}_2, m_2)$ and $\mathcal{A}_1 = \bar{\mathcal{A}}_2$, then $K^D(F_1, F_2) = 1$ implies $K_1^{DP}(F_1, F_2) = 1$.*

5 The Coherence of Conflict Measures and Combining Rules

The value of a conflict measure is an important factor for decision making about using of combining rules for aggregation of information from a few sources. In this case, the conflict measure serves as a priori characteristic of the applicability of the combining rule. The great value of a conflict measure means that we should not do the aggregation of these bodies of evidence. It is clear that the choice of a combining rule and a conflict measure must be coordinated to a certain degree in such problems. Let us consider the following matching conditions.

Definition 1. *A combining rule R and a conflict measure K are called positively agreed if $K(F_1, F_2) \leq K(R(F_1, F_2), F_i)$, $i = 1, 2$ for all $F_1, F_2 \in \mathcal{F}(X)$. The pair R and K is called negatively agreed if the opposite inequality holds.*

The positive (negative) coherence means that the value of a conflict measure between the resulting body of evidence and any operand will not decrease (will not increase) with respect to the value of a conflict measure between operands after application of combining rule.

Proposition 4.

(1) *A conjunctive (non-normalized) combining rule D_0 and a conflict measure K^D are positively agreed;*

- (2) a disjunctive combining rule DP and a conflict measure K_1^{DP} are positively agreed;
- (3) a disjunctive combining rule DP and a conflict measure K_2^{DP} are negatively agreed.

It is easy to see that the coherence of the conflict measure with the combining rule makes some axioms dependent or contradictory. For example,

- (1) if a conflict measure is positively agreed with a conjunctive combining rule or a disjunctive combining rule, then axiom A5 follows from axiom A3 because $K(F_X, F) \leq K(D(F_X, F), F) = K(F, F) = 0$ and $K(F_X, F) \leq K(DP(F_X, F), F_X) = K(F_X, F_X) = 0$;
- (2) if a conflict measure is positively agreed with a conjunctive combining rule, then axiom A6 implies that $K(F_\emptyset, F_A) = 1$ for all $A \neq \emptyset$ (particular case of axiom A7) because $1 = K(F_A, F_{\bar{A}}) \leq K(D(F_A, F_{\bar{A}}), F_A) = K(F_\emptyset, F_A)$;
- (3) if a conflict measure is positively agreed with a disjunctive combining rule, then A5 and A6 axioms are contradictory as well as A3 and A7 axioms because $1 = K(F_A, F_{\bar{A}}) \leq K(DP(F_A, F_{\bar{A}}), F_A) = K(F_X, F_A) = 0$ and $1 = K(F_\emptyset, F) \leq K(DP(F_\emptyset, F), F) = K(F, F) = 0$;
- (4) if a conflict measure is negatively agreed with a conjunctive combining rule, then axiom A5 implies axiom A3 because $0 = K(F_X, F) \geq K(D(F_X, F), F) = K(F, F)$;
- (5) if a conflict measure is negatively agreed with a conjunctive combining rule, then axiom A7 implies that $K(F_A, F_{\bar{A}}) = 1$ for all $A \neq \emptyset$ (particular case of axiom A6) because $K(F_A, F_{\bar{A}}) \geq K(D(F_A, F_{\bar{A}}), F_A) = K(F_\emptyset, F_A) = 1$.

Thus, if we are talking about the desired conditions of conflict measure, then we must take into consideration the problem being solved, the used combining rule and, consequently, the type of their coherence.

6 Metric and Entropic Components of a Conflict Measure

When we take a decision on combining of bodies of evidence we pay attention not only on the value of a conflict measure. Consider the following example.

Example 1. Let us assume that there are three candidates $X = \{x_1, x_2, x_3\}$ for a certain position. Three experts expressed their preference to these candidates as three bodies of evidence $F_1 = \frac{1}{3}F_{\{x_1\}} + \frac{1}{3}F_{\{x_2\}} + \frac{1}{3}F_{\{x_3\}}$, $F_2 = \frac{1}{3}F_{\{x_1, x_2\}} + \frac{2}{3}F_{\{x_3\}}$, $F_3 = \frac{7}{8}F_{\{x_2\}} + \frac{1}{8}F_{\{x_2, x_3\}}$. The conjunctive conflict measures are equal $K^D(F_1, F_2) = 5/9$, $K^D(F_1, F_3) = 5/8$ and $K^D(F_2, F_3) = 7/12$, i.e. $K^D(F_1, F_2) < K^D(F_2, F_3) < K^D(F_1, F_3)$. We will choose for combining a couple of bodies of evidence F_1 and F_2 with the lowest measure of conflict. We get the new body of evidence after combining by Dempster's rule: $D(F_1, F_2) = \frac{1}{4}F_{\{x_1\}} + \frac{1}{4}F_{\{x_2\}} + \frac{1}{2}F_{\{x_3\}}$. The preference is given to a third candidate in this case. At the same time, the evidence F_1 is irrelevant because first expert did not give preference to any of the candidates. If we

find a combination of the second and third bodies of evidence, then we get $D(F_2, F_3) = \frac{4}{5}F_{\{x_2\}} + \frac{1}{5}F_{\{x_3\}}$, i.e. the preference is given to a second candidate in this case. The situation is similar when we use a disjunctive conflict measure and a disjunctive rule: $K_1^{DP}(F_1, F_2) = \frac{1}{9} > K_1^{DP}(F_1, F_3) = K_1^{DP}(F_2, F_3) = \frac{1}{24}$; $DP(F_1, F_3) = \frac{7}{24}F_{\{x_2\}} + \frac{7}{24}F_{\{x_1, x_2\}} + \frac{9}{24}F_{\{x_2, x_3\}} + \frac{1}{24}F_{\{x_1, x_2, x_3\}}$; $DP(F_2, F_3) = \frac{7}{24}F_{\{x_1, x_2\}} + \frac{2}{3}F_{\{x_2, x_3\}} + \frac{1}{24}F_{\{x_1, x_2, x_3\}}$. We obtain approximately equal values of the mass function for three focal elements after combining the F_1 and F_3 . On the contrary, the combination of the second and third sources gives us that the preferred candidate is in the pair $\{x_2, x_3\}$. This example can be explained by the fact that the first body of evidence has a uniform probability distribution. It has high Shannon entropy and it is better not to use for combining. However, the entropic and metric components can be isolated in the conflict measure.

Conjunctive conflict measure for two bodies of evidence $F_1 = (\mathcal{A}_1, m_1)$ and $F_2 = (\mathcal{A}_2, m_2)$ can be rewritten as follows

$$\begin{aligned}
 K^D(F_1, F_2) &= \sum_{B \in \mathcal{A}_1, C \in \mathcal{A}_2, B \cap C = \emptyset} m_1(B)m_2(C) = 1 - \sum_{B \cap C \neq \emptyset} m_1(B)m_2(C) \\
 &= \frac{1}{2} \left(2 - 2 \sum_{B, C} q_{B, C} m_1(B)m_2(C) \right) - \sum_{B, C} (t_{B, C} - q_{B, C}) m_1(B)m_2(C), \quad (2)
 \end{aligned}$$

where $Q = (q_{B, C})$ is a symmetric positive definite matrix which satisfies the conditions: (1) $q_{B, C} \in [0, 1] \forall B, C \in 2^X$; (2) $q_{B, C} = 0$, if $B \cap C = \emptyset$; (3) $q_{B, B} = 1 \forall B \in 2^X$; $T = (t_{B, C})$, $t_{B, C} = \begin{cases} 1, & B \cap C \neq \emptyset, \\ 0, & B \cap C = \emptyset. \end{cases}$ Let $R = (r_{B, C})$, $r_{B, C} = t_{B, C} - q_{B, C}$. For example, Jaccard index, $q_{B, C} = \frac{|B \cap C|}{|B \cup C|}$, $\forall B, C \neq \emptyset$ is an example of coefficients $q_{B, C}$. We will consider a scalar product $(\mathbf{x}, \mathbf{y})_Q := \mathbf{x}^T Q \mathbf{y}$ and corresponding norm $\|\mathbf{x}\|_Q := \sqrt{\mathbf{x}^T Q \mathbf{x}}$ in the real vector space $A_{2|X|_1}$. In particular, if $q_{B, C}$ is Jaccard index, then $d_J(F_1, F_2) = \frac{1}{\sqrt{2}} \|\mathbf{m}_1 - \mathbf{m}_2\|_Q$ is Jousselme distance [8] that is widely used in evidence theory.

Let $S_X = \left\{ \mathbf{t} = (t_k)_{k=1}^{2^{|X|}-1} : t_k \in [0, 1] \forall k, \sum_k t_k = 1 \right\}$ is a simplex. We consider a functional $E_Q : S_X \rightarrow [0, 1]$,

$$E_Q(F) = E_Q(\mathbf{m}) = 1 - \|\mathbf{m}\|_Q^2 = \sum_B m(B) \left(1 - \sum_C q_{B, C} m(C) \right), \quad F = (\mathcal{A}, m).$$

This functional is close to an entropy functional in some of its properties: $\mathbf{t}^{(\max)} = \arg \max_{S_X} E_Q(\mathbf{t}) = \arg \min_{S_X} \|\mathbf{t}\|_Q^2$, $\mathbf{t}^{(\min)} = \arg \min E_Q(\mathbf{t})$, if $\exists j : t_j^{(\min)} = 1$ and $t_k^{(\min)} = 0 \forall k \neq j$ (categorical evidence), $E_Q(\mathbf{t}^{(\min)}) = 0$. The next Proposition follows from (2).

Proposition 5. *We have for the conjunctive conflict measure*

$$K^D(F_1, F_2) = \frac{1}{2} \left(E_Q(\mathbf{m}_1) + E_Q(\mathbf{m}_2) + \|\mathbf{m}_1 - \mathbf{m}_2\|_Q^2 \right) - \sum_{B,C} r_{B,C} m_1(B) m_2(C). \quad (3)$$

The formula (3) shows that the conjunctive conflict measure can be represented as a sum of average value of entropy-type functionals of bodies of evidence, the distance between bodies of evidence and a last summand that characterizes the interaction of weakly intersecting focal elements.

Corollary 1. *If $F_1 = (\mathcal{A}, m_1)$ and $F_2 = (\mathcal{A}, m_2)$, where $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$, then*

$$K^D(F_1, F_2) = \frac{1}{2} (E_I(\mathbf{m}_1) + E_I(\mathbf{m}_2)) + \frac{1}{2} \|\mathbf{m}_1 - \mathbf{m}_2\|_I^2, \quad (4)$$

where I is the identity matrix and $\|\mathbf{x}\|_I := \sqrt{\mathbf{x}^T \mathbf{x}}$ is the Euclidean norm.

Note that functional $E_I(\mathbf{t}) = \sum_B t(B)(1 - t(B))$ is defined on the simplex S_X and satisfies the conditions: $\mathbf{t}^{(\max)} = \arg \max E_I(\mathbf{t})$, if $t_k^{(\max)} = \frac{1}{2^{|X|-1}} \forall k$ (uniform distribution), $E_I(\mathbf{t}^{(\max)}) = 1 - \frac{1}{2^{|X|-1}}$; $\mathbf{t}^{(\min)} = \arg \min E_I(\mathbf{t})$, if $\exists j : t_j^{(\min)} = 1$ and $t_k^{(\min)} = 0 \forall k \neq j$ (categorical evidence), $E_I(\mathbf{t}^{(\min)}) = 0$. In addition, we have $E_I(\mathbf{t}) \leq S(\mathbf{t}) := -\sum_k t_k \log_2 t_k$ (Shannon entropy). Thus, the conjunctive conflict measure is equal in this case the average value of the entropy-type functionals and the square of the distance between the mass functions of the two bodies of evidence.

Note that the conjunctive conflict measure satisfies the triangle inequality on the disjoint belief structures.

Proposition 6. *If $F_i = (\mathcal{A}, m_i)$, $i = 1, 2, 3$, where $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$, then $K^D(F_1, F_3) \leq K^D(F_1, F_2) + K^D(F_2, F_3)$.*

Proposition 2 implies that we have the following representation for a disjunctive conflict measure and a special case of belief structures $F_1 = (\mathcal{A}, m_1)$ and $F_2 = (\mathcal{A}, m_2)$, where $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$, $N = |\mathcal{A}| > 1$:

$$K_1^{DP}(F_1, F_2^{(-)}) = \frac{1}{2(N-1)} (E_I(\mathbf{m}_1) + E_I(\mathbf{m}_2)) + \frac{1}{2(N-1)} \|\mathbf{m}_1 - \mathbf{m}_2\|_I^2,$$

$$K_2^{DP}(F_1, \bar{F}_2) = K^D(F_1, F_2) = \frac{1}{2} (E_I(\mathbf{m}_1) + E_I(\mathbf{m}_2)) + \frac{1}{2} \|\mathbf{m}_1 - \mathbf{m}_2\|_I^2.$$

7 Changing of Metric and Entropic Components of a Conflict Measure after Combining

By definition, we have that a conflict measure is not decreased after combining of bodies of evidence in the case of positive compatibility. On the other hand, metric and entropic components can be isolated in the conjunctive conflict measure. We have the question about changing of these parts when bodies of evidence are combined.

Proposition 7. *If $F_1 = (\mathcal{A}, m_1)$ and $F_2 = (\mathcal{A}, m_2)$, where $\emptyset \notin \mathcal{A}$ and $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$, then the metric component of a conjunctive conflict measure does not decrease after application of a conjunctive rule.*

The entropic component of a conjunctive conflict measure can be increased or decreased after application of a conjunctive rule.

Proposition 8. *If $F_1 = (\mathcal{A}, m_1)$ and $F_2 = (\mathcal{A}, m_2)$, where $\emptyset \notin \mathcal{A}$ and $A' \cap A'' = \emptyset \forall A', A'' \in \mathcal{A}$, then $E_I(DP(F_1, \bar{F}_2)) \geq E_I(\bar{F}_2)$.*

By other words, the value of entropy-type functional does not decrease after combining of bodies of evidence F_1 and \bar{F}_2 with the help of disjunctive rule with respect to value of entropy-type functional of body \bar{F}_2 .

The metric component of a disjunctive conflict measure can be increased or decreased after application of a disjunctive rule.

8 Conclusions

Conflict measures induced by the conjunctive and disjunctive combining rules were studied in this paper. In particular, some of the consistency conditions between the combining rules and conflict measures were discussed. The relationship of consistency conditions and the axiomatic of a conflict measure is shown.

In addition, it is shown that the metric and entropic components can be isolated into the conjunctive conflict measures. It is shown that the entropic component of evidence is an important characteristic (together with the value of a conflict measure) in decisions about the choice of the bodies of evidence for combining. It is shown in some special cases (disjoint belief structures) that the metric component of conjunctive conflict measure is not decreased after applying of conjunctive combining rule. In addition, it is shown that the value of entropic component is not decreased after combining of bodies of evidence with the help of disjunctive rules.

Acknowledgments. The study has been funded by the Russian Academic Excellence Project ‘5-100’. This work was also partially supported by the grant 18-01-00877 of RFBR (Russian Foundation for Basic Research).

References

1. Bronevich, A., Lepskiy, A., Penikas, H.: The application of conflict measure to estimating incoherence of analyst’s forecasts about the cost of shares of Russian companies. *Procedia Comput. Sci.* **55**, 1113–1122 (2015)
2. Daniel, M.: Conflicts within and between belief functions. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) *IPMU 2010. LNCS (LNAI)*, vol. 6178, pp. 696–705. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14049-5_71
3. Dempster, A.P.: Upper and lower probabilities induced by multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)

4. Destercke, S., Burger, T.: Toward an axiomatic definition of conflict between belief functions. *IEEE Trans. Cybern.* **43**(2), 585–596 (2013)
5. Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *Int. J. Gen. Syst.* **12**(3), 193–226 (1986)
6. Dubois, D., Prade, H.: On the combination of evidence in various mathematical frameworks. In: Flamm, J., Luisi, T. (eds.) *Reliability Data Collection and Analysis*, pp. 213–241. Kluwer Acad. Publ., Dordrecht (1992)
7. Florea, M.C., Jousselme, A.-L., Bossé, É., Grenier, D.: Robust combination rules for evidence theory. *Inf. Fusion* **10**(2), 183–197 (2009)
8. Jousselme, A.-L., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. *Inf. Fusion* **2**, 91–101 (2001)
9. Lefèvre, E., Elouedi, Z.: How to preserve the conflict as an alarm in the combination of belief functions? *Decis. Support Syst.* **56**, 326–333 (2013)
10. Lepskiy, A.: On internal conflict as an external conflict of a decomposition of evidence. In: Vejnárová, J., Kratochvíl, V. (eds.) *BELIEF 2016. LNCS (LNAI)*, vol. 9861, pp. 25–34. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45559-4_3
11. Lepskiy, A.: Decomposition of evidence and internal conflict. *Procedia Comput. Sci.* **122**, 186–193 (2017)
12. Liu, W.: Analysing the degree of conflict among belief functions. *Artif. Intell.* **170**, 909–924 (2006)
13. Martin, A.: About conflict in the theory of belief functions. In: Denoeux, T., Masson, M.H. (eds.) *Belief Functions: Theory and Applications. Advances in Intelligent and Soft Computing*, vol. 164, pp. 161–168. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29461-7_19
14. Schubert, J.: The internal conflict of a belief function. In: Denoeux, T., Masson, M.H. (eds.) *Belief Functions: Theory and Applications. Advances in Intelligent and Soft Computing*, vol. 164, pp. 169–177. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29461-7_20
15. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton (1976)



Linear Belief Functions for Data Analytics

Liping Liu^(✉)

The University of Akron, Akron, USA

liu@acm.org

Abstract. This paper studies the application of linear belief functions to both classic and Bayesian statistic data analysis. In particular, it explores how to combine direct observations with/without distributional assumptions as linear belief functions for estimating population mean, how to combine system equations and measurement equations with direct observations in time series models and Bayesian linear regressions. It illustrates the use of Linear Model Operating Systems (LMOS).

1 Introduction

Linear belief functions are an extension of the Dempster-Shafer theory to the case where the variables of interest are Gaussian [3]. The notion manifests a wide range of linear models such as linear regressions, linear equations, and ignorance, as well as marginal and conditional multivariate normal distributions of linear combinations of variables, which can all be uniformly represented as matrices and combined as the addition of the matrices via the Dempster's rule [4, 5].

This paper studies the application of linear belief functions to the estimation of population means, time series analysis, and linear regressions. Although these analyses can be done without belief functions, this study is important in three perspectives. First, it illustrates the applicability of belief functions to these problems. Considering that evidence involved in the analyses, including equations, distributional assumptions, and observations are all special cases of belief functions, uniformly expressed as matrices, and combined as matrix additions, the belief function approach is more elegant. Second, in Bayesian analysis, ignorance will have to be expressed using improper priors, the belief function approach is free from such a burden. Third, the big data analytics aims at using population data to predict the behavior of individuals, but often we cannot wholly load a data set into computer memory for analysis. Since the combination of belief functions is commutative and associative, the belief function approach has the potential to meet the challenge by slicing the population data.

2 Combining Direct Observations

Normal distribution $X \sim N(\mu, \Sigma)$ is represented in the matrix form of a linear belief function as $M(X) = [\mu \ \Sigma]^T$ or fully swept form $M(\vec{X}) = [\mu \Sigma^{-1} \ -\Sigma^{-1}]^T$.

As special cases, $M(X) = [x \ 0]^T$ represents certainty $X = x$ with 0 being zero covariance matrix, and $M(\vec{X}) = [0 \ 0]^T$ full ignorance on X . Assume we made n independent direct observations on X : $X = x_1, X = x_2, \dots, X = x_n$. They can be combined as belief functions. First, we represent each observation as $M_i(X) = [x_i \ 0]^T, i = 1, 2, \dots, n$. Then, we fully sweep each matrix using imaginary extreme numbers $e = 1/0$ [6]: $M_i(\vec{X}) = [x_i e \ -Ie]^T$. The combination is simply the sum of their fully swept matrices [5], and so the combination of the observations is $M(\vec{X}) = [(\sum_{i=1}^n x_i)e \ -nIe]^T$. Doing a full reverse sweeping (or unsweep) on $M(\vec{X})$ results in $M(X) = [\frac{\sum_{i=1}^n x_i}{n} \ 0]^T$.

Theorem 1. *The combination of independent direct observations on the same variables is the sample mean of the observations.*

Next assume each observation x_i has an error term $\epsilon_i \sim N(0, \Sigma_i)$. Then we can represent observation i as $X = x_i + \epsilon_i, \epsilon_i \sim N(0, \Sigma_i)$, both of which are linear belief functions: the first is a linear equation of X and ϵ_i with constant x_i , and the second is a normal distribution.

$$M_{1i}(X, \vec{\epsilon}_i) = \begin{bmatrix} x_i & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, M_{2i}(\epsilon_i) = \begin{bmatrix} 0 \\ \Sigma_i \end{bmatrix}$$

Note $M_{1i}(X, \vec{\epsilon})$ is $M_{1i}(X, \epsilon_i)$ being partially swept on ϵ_i . In general, the following defines how to sweep $M(X, Y)$ on X into $M(\vec{X}, Y)$:

$$\begin{bmatrix} \mu_X & \mu_Y \\ \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \implies \begin{bmatrix} \mu_X(\Sigma_{XX})^{-1} & \mu_Y - \mu_X(\Sigma_{XX})^{-1}\Sigma_{XY} \\ -(\Sigma_{XX})^{-1} & (\Sigma_{XX})^{-1}\Sigma_{XY} \\ \Sigma_{YX}(\Sigma_{XX})^{-1} & \Sigma_{YY} - \Sigma_{YX}(\Sigma_{XX})^{-1}\Sigma_{XY} \end{bmatrix}$$

$M(\vec{X}, Y)$ contains linear regression $Y = a + bX + \epsilon$ with coefficient $b = (\Sigma_{XX})^{-1}\Sigma_{XY}$ and intercept $a = \mu_Y - \mu_X(\Sigma_{XX})^{-1}\Sigma_{XY}$. Also, $\mu_X(\Sigma_{XX})^{-1}$ and $-(\Sigma_{XX})^{-1}$ constitute $M(\vec{X})$, and $\epsilon \sim N(0, \Sigma_{YY} - \Sigma_{YX}(\Sigma_{XX})^{-1}\Sigma_{XY})$. As a special case, linear equation $Y = a + bX$ has ignorance on X and no white noise, and so in $M(\vec{X}, Y)$, the last terms vanish.

To combine belief functions M_{1i} and M_{2i} , we need to sweep M_{2i} from ϵ_i since ϵ_i is a common variable between M_{1i} and M_{2i} [5]. $M_{2i}(\vec{\epsilon}_i) = [0 \ -(\Sigma_i)^{-1}]^T$ is then added to $M_{1i}(X, \vec{\epsilon}_i)$ to obtain the combination as

$$M_i(X, \vec{\epsilon}_i) = \begin{bmatrix} x_i & 0 \\ 0 & 1 \\ 1 & -(\Sigma_i)^{-1} \end{bmatrix}$$

Doing a reverse sweeping on $M_i(X, \vec{\epsilon}_i)$ from ϵ_i produces

$$M_i(X, \epsilon_i) = \begin{bmatrix} x_i & 0 \\ \Sigma_i & 1 \times \Sigma_i \\ \Sigma_i \times 1 & \Sigma_i \end{bmatrix}$$

Marginalizing $M_i(X, \epsilon_i)$ to X , or removing variable ϵ_i , we have $M_i(X) = [x_i \ \Sigma_i]^T$, which is a compact representation of observation i : $X \sim N(x_i, \Sigma_i)$. Because combination is commutative and associative, we can just fully sweep each $M_i(X)$: $M_i(\vec{X}) = [x_i(\Sigma_i)^{-1} \ -(\Sigma_i)^{-1}]^T$ and sum the swept matrices.

$$M(\vec{X}) = \sum_{i=1}^n M_i(\vec{X}) = \begin{bmatrix} \sum_{i=1}^n x_i(\Sigma_i)^{-1} \\ -\sum_{i=1}^n (\Sigma_i)^{-1} \end{bmatrix}$$

Doing a reverse sweeping obtains

$$M(X) = \begin{bmatrix} \sum_{i=1}^n x_i(\Sigma_i)^{-1} \ [\sum_{i=1}^n (\Sigma_i)^{-1}]^{-1} \\ [\sum_{i=1}^n (\Sigma_i)^{-1}]^{-1} \end{bmatrix}$$

Theorem 2. Assume x_1, x_2, \dots, x_n are n independent observations on random vector X with white noise error term $\epsilon_i \sim N(0, \Sigma_i), i = 1, 2, \dots, n$. Then the combined linear belief function of X has mean value $\sum_{i=1}^n x_i(\Sigma_i)^{-1} [\sum_{i=1}^n (\Sigma_i)^{-1}]^{-1}$ and error $\epsilon \sim N(0, [\sum_{i=1}^n (\Sigma_i)^{-1}]^{-1})$.

Corollary 1. Assume x_1, x_2, \dots, x_n are n independent observations on variable X with errors ϵ_i , which has standard deviation $\sigma_i, i = 1, 2, \dots, n$. Then the combined linear belief function of X has mean value

$$\frac{\frac{x_1}{(\sigma_1)^2} + \frac{x_2}{(\sigma_2)^2} + \dots + \frac{x_n}{(\sigma_n)^2}}{\frac{1}{(\sigma_1)^2} + \frac{1}{(\sigma_2)^2} + \dots + \frac{1}{(\sigma_n)^2}}$$

and variance $\frac{1}{\frac{1}{(\sigma_1)^2} + \frac{1}{(\sigma_2)^2} + \dots + \frac{1}{(\sigma_n)^2}} < \min(\sigma_1, \sigma_2, \dots, \sigma_n)$.

Corollary 2. Assume x_1, x_2, \dots, x_n are n independent observations on variable X with identical error ϵ with standard deviation σ . Then the combined linear belief function of X is $N(\bar{x}, \frac{\sigma^2}{n})$.

Corollary 2 is a familiar result; the combined linear belief function of X is identical to the Bayesian posterior distribution of population mean μ with a known standard deviation σ and an improper uniform prior (improper because the prior is not a probability distribution):

$$p(\mu|x_1, x_2, \dots, x_n) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} e^{-\frac{1}{2}(\frac{\mu-\bar{x}}{\sigma/\sqrt{n}})^2}$$

One may also verify that Corollary 1 is identical to the Bayesian posterior distribution of μ with a improper uniform prior $p(\mu)$ and known but different standard deviations $\sigma_1, \sigma_2, \dots, \sigma_n$ by carrying out some tedious operations according to Bayes rule:

$$p(\mu|x_1, x_2, \dots, x_n) = \frac{p(\mu) \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma_i})^2}}{\int p(\mu) \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma_i})^2} d\mu}$$

According to the corollary, the combined mean for X is the inverse-variance weighted average of the original observations. This result makes an intuitive sense; the larger the error, the less reliable the observation. The formula discounts each observation by the inverse variance of its error term. Therefore, only when all the measurements have independent, identical error terms, e.g., obtained by using the same stable measurement device, the combined mean is the arithmetic average of the observations.

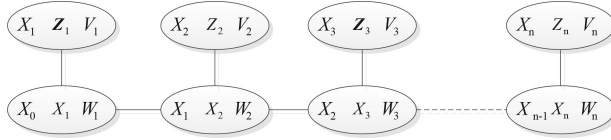
3 Dynamic Linear Models

Dynamic linear models are a general class of non-stationary time series models. A simple case is an autoregressive model, where the current value of a variable is predicted against its previous values, e.g., stock price p_t at time t is predicted by the prices of the three prior closings: $p_t = \beta_0 + \beta_1 p_{t-1} + \beta_2 p_{t-2} + \beta_3 p_{t-3} + \epsilon_t$, where ϵ_t is the associated error term. More sophisticated models include terms to model trends, seasonality, covariates, and autoregressive components. Kalman filters are a special case [1]. Let X_t be the state of the system at time t , u_t be the control input, and w_t the noise. Then the dynamics of the system may be described by the linear equation $X_{t+1} = A_t X_t + B_t u_t + w_t$ with A_t and B_t being the coefficient matrices, and $t = 0, 1, 2, \dots$. The state of the system may be observed with errors at each time. Let Z_t be the observed result of the system at time t and V_t be the measurement error. Then the measurements may be expressed by another linear equation $Z_t = C_t X_t + V_t$ with C_t being the coefficient matrix. Assume all the noises and errors are independent. The problem is to estimate X_{t+1} by using the observations $X_0, Z_1, Z_2, \dots, Z_t$.

Dynamic linear models can be more complex than Kalman filters can handle. Just like stock prices, X_{t+1} may depend not only on X_t and u_t but also earlier states and controls; e.g., economic policies u_t often take many years to be effective in affecting the state of economy X_t . Similarly, measurement Z_t may depend not only on X_t but also earlier measurements or states. For example, the measurement of temperature outside the combustion chamber of a rocket (Z_t) depends on both the current and previous internal temperatures (X_t, X_{t-1}) of the chamber due to the lag of heat diffusion: $Z_t = C_t X_t + C_{t-1} X_{t-1} + V_t$.

Regardless how complex general dynamic linear models are, the dynamic system and measurement equations are linear equations, and noises and measurement errors are Gaussian. Observations, linear equations, and Gaussian distributions are all special cases of linear belief functions, and they can be dynamically combined to propagate existing evidence to predict the current state or estimate model parameters. Here I will use a numerical example via LMOS to illustrate the idea. Suppose we want to monitor a voltage between two points, which is a constant subject to a random noise: $X_t = X_{t-1} + W_t$ with $X_0 = 120$, $W_t \sim N(0, 10^{-5})$, and $t = 1, 2, \dots$. The measurements Z_t are subject to errors V_t : $Z_t = X_t + V_t$ with $V_t \sim N(0, 0.01)$. The problem is to estimate X_{t+1} based on Z_1, Z_2, \dots, Z_t . Each linear belief function may be graphically represented by

a hyperedge containing the variables of the function. Using triangulation and maximum cardinality search, we assemble the variables into a join tree [4, 8]:



Attached to each node in the join-tree is a matrix, which may be the combination of several original models or may be vacuous if there exists no knowledge bearing on the variables. Here at node $\{X_0, X_1, W_1\}$ is the combination of linear model $X_1 = X_0 + W_1$, $W_1 \sim N(0, 0.00001)$, and $X_0 = 12$ and represented by Matrix (a). At each leaf node $\{X_i, Z_i, V_i\}$ is combination of three models: $Z_i = X_i + V_i$, $V_i \sim N(0, 0.01)$, and a direct observation on Z_i . At each node $\{X_i, X_{i+1}, W_{i+1}\}$ is the combination of linear equation $X_{i+1} = X_i + W_{i+1}$ and $W_{i+1} \sim N(0, 10^{-5})$. To illustrate the calculation, let us assume direct observations $Z_1 = 11.09$ and $Z_2 = 12.05$ and combine the corresponding belief functions for $\{X_1, Z_1, V_1\}$, $\{X_1, X_2, W_2\}$, and $\{X_2, Z_2, V_2\}$ as shown in Matrix (b–d).

	X1	X0	W1
► Mean	12.00	12.00	0.000
X1	0.005000	0.000	0.005000
X0	0.000	0.000	0.000
W1	0.005000	0.000	0.005000

(a)

	Z1	X1	V1
► Mean	11.09	11.09	0.000
Z1	0.000	0.000	0.000
X1	0.000	0.01000	-0.01000
V1	0.000	-0.01000	0.01000

(b)

	X2	(X1)	W2
► Mean	0.000	0.000	0.000
X2	0.005000	1.000	0.005000
(X1)	1.000	0.000	0.000
W2	0.005000	0.000	0.005000

(c)

	Z2	X2	V2
► Mean	12.05	12.05	0.000
Z2	0.000	0.000	0.000
X2	0.000	0.01000	-0.01000
V2	0.000	-0.01000	0.01000

(d)

	X1	X0	W1
► Mean	11.70	12.00	-0.3033
X1	0.003333	0.000	0.003333
X0	0.000	0.000	0.000
W1	0.003333	0.000	0.003333

(e)

	X2	X1	W2
► Mean	11.86	11.76	0.09636
X2	0.004545	0.001818	0.002727
X1	0.001818	0.002727	-0.0009091
W2	0.002727	-0.0009091	0.003636

(f)

Since this is a real time dynamic system, propagation must be in chronological order, i.e., using the information in the past to update the current system status. Propagation starts from a leaf node, and each node waits to send its message to a next neighbor until it has received messages from all of its other neighbors. The message to be sent is the combination of the received messages and the matrix stored at the node. Since knowledge on variables not contained in the receiving node is irrelevant to the receiver, we marginalize the combination to the intersection between the sender and the receiver. $\{X_1, Z_1, V_1\}$ is the leaf node to start propagation. Its message to $\{X_0, X_1, W_1\}$ is the marginal of Matrix (b) to X_1 : $M_{1 \rightarrow 2}(X_1) = (11.09, 0.01)^T$, which is combined with Matrix (a) at $\{X_0, X_1, W_1\}$ to produce Matrix (e).

The message to be sent from $\{X_0, X_1, W_1\}$ to $\{X_1, X_2, W_2\}$ is $M_{2 \rightarrow 3}(X_1) = (11.70, 0.0033)^T$, which is the marginalization of Matrix (e), and the message

from $\{X_2, Z_2, V_2\}$ to $\{X_1, X_2, W_2\}$ is $M_{4 \rightarrow 3}(X_2) = (12.05, 0.01)^T$, which is the marginalization of Matrix (d). These messages are combined with Matrix (c) to produce Matrix (f), and so $X_2 \sim N(11.86, 0.0045)$, which is then passed on to Node $\{X_2, X_3, W_3\}$ to continue the propagation.

4 Bayesian Linear Regression

Bayesian linear regression is essentially the combination of two equations: $Y = BA + \epsilon$ and $Y = y$ along with the assumption of a normal distribution $\epsilon \sim N(0, \Sigma)$ [2, 7]. Here Y is a row vector of n response variables, B is a $1 \times m$ vector representing unknown coefficients, A is a $m \times n$ matrix made of the observations of the independent variables, ϵ is a row vector of n noise variables, y is a vector of direct observations on Y . Assuming Σ is known, then we have:

$$M_1(Y, \vec{A}, \vec{\epsilon}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A^T & I \\ A & 0 & 0 \\ I & 0 & 0 \end{bmatrix}, M_2(Y) = \begin{bmatrix} y \\ 0 \end{bmatrix}, M_3(\epsilon) = \begin{bmatrix} 0 \\ \Sigma \end{bmatrix}$$

First we combine $M_1(Y, \vec{A}, \vec{\epsilon})$ with $M_3(\epsilon)$. Since ϵ is a common variable, we need to sweep the later into $M_3(\vec{\epsilon}) = [0 \ -\Sigma^{-1}]^T$ and then add it with the former per the new matrix addition rule [5]:

$$M_{12}(Y, \vec{B}, \vec{\epsilon}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A^T & I \\ A & 0 & 0 \\ I & 0 & -\Sigma^{-1} \end{bmatrix}$$

Doing a reverse sweeping on $M_{12}(Y, \vec{B}, \vec{\epsilon})$ from ϵ produces

$$M_{12}(Y, \vec{B}, \epsilon) = \begin{bmatrix} 0 & 0 & 0 \\ \Sigma & A^T & \Sigma \\ A & 0 & 0 \\ \Sigma & 0 & \Sigma \end{bmatrix}$$

Now Y is common between $M_{12}(Y, \vec{B}, \epsilon)$ and $M_2(Y)$, so we weep each from Y :

$$M_{12}(\vec{Y}, \vec{B}, \epsilon) = \begin{bmatrix} 0 & 0 & 0 \\ -\Sigma^{-1} & \Sigma^{-1}A^T & I \\ A\Sigma^{-1} & -A\Sigma^{-1}A^T & -A \\ I & -A^T & 0 \end{bmatrix}, M_2(\vec{Y}) = \begin{bmatrix} ye \\ -eI \end{bmatrix}$$

which are then added together to obtain

$$M_{123}(\vec{Y}, \vec{B}, \epsilon) = \begin{bmatrix} ye & 0 & 0 \\ -\Sigma^{-1} - eI & \Sigma^{-1}A^T & I \\ A\Sigma^{-1} & -A\Sigma^{-1}A^T & -A \\ I & -A^T & 0 \end{bmatrix}$$

Next we reversely sweep $M(\vec{Y}, \vec{B}, \epsilon)$ from Y and get rid of the extreme numbers involving e because $(\Sigma^{-1} + eI)^{-1} = \Sigma(\frac{1}{e}I + \Sigma)^{-1}\frac{1}{e} = 0$, and $e(\Sigma^{-1} + eI)^{-1} = I$:

$$M(Y, \vec{B}, \epsilon) = \begin{bmatrix} y & y\Sigma^{-1}A^T & y \\ 0 & 0 & 0 \\ 0 & -A\Sigma^{-1}A^T & -A \\ 0 & -A^T & 0 \end{bmatrix}$$

Finally, doing a reverse sweeping from B will produce the final belief function:

$$M(Y, B, \epsilon) = \begin{bmatrix} y (y\Sigma^{-1}A^T)(A\Sigma^{-1}A^T)^{-1} y - (y\Sigma^{-1}A^T)(A\Sigma^{-1}A^T + A^{-1})^{-1} A \\ 0 & 0 & 0 \\ 0 & (A\Sigma^{-1}A^T)^{-1} & -[(A\Sigma^{-1}A^T)^{-1}A \\ 0 & -A^T(A\Sigma^{-1}A^T)^{-1} & A^T(A\Sigma^{-1}A^T)^{-1}A \end{bmatrix}$$

which implies that the unknown coefficients B has an estimated posterior distribution $N[(y\Sigma^{-1}A^T)(A\Sigma^{-1}A^T)^{-1}, (A\Sigma^{-1}A^T)^{-1}]$. In the following, I will use a synthetic sample of six observations on two independent variables and one response variable Y to illustrate the result.

X_1	X_2	Y	LMOS Free Format	LMOS Free Format	LMOS Matrix
1.2	2.4	2.3	1.2a + 2.4b + c + e1 = y	y = 2.3	e1 ~ N(0,4)
1.1	2.5	2.6	1.1a + 2.5b + c + e2 = y	y = 2.6	e2 ~ N(0,4)
1.9	3.2	3.2	1.9a + 3.2b + c + e3 = y	y = 3.2	e3 ~ N(0,4)
2.5	4.8	4.5	2.5a + 4.8b + c + e4 = y	y = 4.5	e4 ~ N(0,4)
1.2	2.3	1.9	1.2a + 2.3b + c + e5 = y	y = 1.9	e5 ~ N(0,4)
1.3	2.9	2.8	1.3a + 2.9b + c + e6 = y	y = 2.8	e26 ~ N(0,4)

Assume the linear regression model to be estimated is $Y = aX_1 + bX_2 + c + \epsilon$ with ϵ being assumed to be a white noise with known error: $\epsilon \sim N(0, 4)$. There are a few alternative methods to specify the models in LMOS. We can represent each sample observation (x_{i1}, x_{i2}, y_i) by two linear equations $Y_i = ax_{i1} + bx_{i2} + c + \epsilon_i$, $Y_i = y_i$ and a normal distribution $\epsilon_i \sim N(0, 4)$. Then plugging in the sample data, we will create 6 identical distributions $e_i \sim N(0, 4)$ and 12 linear equations. If one is familiar with the matrix representation of linear regression models, he can further simplify the data entry by combining each linear equation with distributional assumption directly. For example, $1.2a + 2.4b + c + e_1 = 2.3$ (or $c = 2.3 - 1.2a - 2.4b - e_1$) and $e_1 \sim N(0, 4)$ can be represented directly as

$$M(c, \vec{a}, \vec{b}, \vec{e}_1) = \begin{bmatrix} 0 & 0 & 2.3 & 0 \\ 0 & 0 & -1.2 & 0 \\ 0 & 0 & -2.4 & 0 \\ -1.2 & -2.4 & 0 & -1 \\ 0 & 0 & -1 & -4^{-1} \end{bmatrix}$$

To combine the models, we should avoid combining a joint distribution to a linear equation directly. So we first combine the twelve linear equations into a joint linear system model and then combine the joint linear system with the

distribution of one random variable at a time, and after the combination, conduct a reverse sweeping of the result from the random variable. Then we combine the result with the next marginal distribution, and so on. The final result, after removing response variables, is shown below:

	e6	b	a	c	e1	e2	e3	e4	e5
Mean	0.0072656	1.0163	-0.11985	0.0013816	0.0034112	0.18980	0.17430	-0.079811	-0.29496
e6	1.5793	-2.9441	5.3833	-0.039614	0.64556	1.4783	-0.76745	0.71317	0.35114
b	-2.9441	11.248	-18.242	-5.9616	0.85585	-2.0932	4.6265	-2.4257	1.9807
a	5.3833	-18.242	32.192	5.6682	-0.51840	4.5250	-8.4595	1.4122	-2.3426
c	-0.039614	-5.9616	5.6682	9.9597	-2.4536	-1.2906	-1.6520	4.4856	-3.0498
e1	0.64556	0.85585	-0.51840	-2.4536	1.0216	0.88422	0.69984	-0.35848	1.1072
e2	1.4783	-2.0932	4.5250	-1.2906	0.88422	1.5460	-0.60875	0.025309	0.67490
e3	-0.76745	4.6265	-8.4595	-1.6520	0.69984	-0.60875	2.9203	0.59359	1.1625
e4	0.71317	-2.4257	1.4122	4.4856	-0.35848	0.025309	0.59359	3.6275	-0.60105
e5	0.35114	1.9807	-2.3426	-3.0498	1.1072	0.67490	1.1625	-0.60105	1.3053

Therefore, the regression coefficients have an estimated posterior distribution

$$(a, b, c) \sim N\left[\begin{pmatrix} -0.11985 \\ 1.0163 \\ 0.0013816 \end{pmatrix}, \begin{pmatrix} 32.192 & -18.242 & 5.6682 \\ -18.242 & 11.248 & -5.9616 \\ 5.6682 & -5.9616 & 9.9597 \end{pmatrix}\right]$$

and the estimated residuals are $\hat{\epsilon}_1 = 0.0034112$, $\hat{\epsilon}_2 = 0.18980$, etc.

5 Conclusion

This paper studies the application of linear belief functions to statistic analysis. When combining multiple independent measurements without errors over the same variables as linear belief functions, the result is the sample mean of the observations. When errors are assumed to be white noises, the result is the Bayesian posterior distribution of the population mean (without imposing improper priors). When combining the measurements over the variables with the linear regression equation that links the variables, the result is the Bayesian posterior distribution of regression coefficients. Again, there is no need to assume an improper prior over the coefficients as in Bayesian linear regression.

This paper also introduced Linear Model Operating System (LMOS, developed and made available to the public at lmos.org by the author) as the tool to represent and combine linear models. LMOS allows one to enter distributions as matrices and direct measurements and/or linear equations as free-entry equations. The paper carried out two extended examples. Using a numerical example of monitoring voltages between two points, the paper shows how to perform the Kalman filter estimation as the combination of dynamic system equations, measurement equations, and observations over a Markov tree. Using synthetic data of measurements, the paper shows how to obtain posterior distributions of population mean and regression coefficients.

References

1. Dempster, A.P.: Construction and local computation aspects of network belief functions. In: Oliver, R.M., Smith, J.Q. (eds.) *Influence Diagrams, Belief Nets, and Decision Analysis*, pp. 121–141. Wiley, Chichester (1989)
2. Dempster, A.P.: Normal belief functions and the Kalman filter. In: Saleh, A.K.M.E. (ed.) *Data Analysis from Statistical Foundations*, pp. 65–84. Nova Science Publishers, Hauppauge (2001)
3. Liu, L.: A theory of Gaussian belief functions. *Int. J. Approx. Reason.* **14**, 95–126 (1996)
4. Liu, L.: Local computation of Gaussian belief functions. *Int. J. Approx. Reason.* **22**, 217–248 (1999)
5. Liu, L.: A new matrix addition rule for combining linear belief functions. In: Vejnarová, J., Kratochvíl, V. (eds.) *BELIEF 2016. LNCS (LNAI)*, vol. 9861, pp. 14–24. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45559-4_2
6. Liu, L.: Imaginary numbers for combining linear equation models via dempster's rule. *Int. J. Approx. Reason.* **55**, 294–310 (2014)
7. Monney, P.A.: Analyzing linear regression models with hints and Dempster-Shafer theory. *Int. J. Intell. Syst.* **18**, 5–29 (2003)
8. Shafer, G., Shenoy, P.P.: Probability propagation. *Ann. Math. Artif. Intell.* **2**, 327–352 (1990)



Outer Approximations of Coherent Lower Probabilities Using Belief Functions

Ignacio Montes¹, Enrique Miranda^{1(✉)}, and Paolo Vicig²

¹ Department of Statistics and O.R., University of Oviedo, Oviedo, Spain
{imontes,mirandaenrique}@uniovi.es

² DEAMS, University of Trieste, Trieste, Italy
paolo.vicig@deams.units.it

Abstract. We investigate the problem of outer approximating a coherent lower probability with a more tractable model. In particular, in this work we focus on the outer approximations made by belief functions. We show that they can be obtained by solving a linear programming problem. In addition, we consider the subfamily of necessity measures, and show that in that case we can determine all the undominated outer approximations in a simple manner.

1 Introduction

Coherent lower probabilities are one of the most prominent models within imprecise probability theory [1]. They can be given a behavioural interpretation in terms of acceptable betting rates, thus extending Bruno de Finetti's work on subjective probability theory; at the same time, they are also equivalent to convex sets of probability measures (*credal sets*), meaning that they can be regarded as an epistemic model of imprecise information.

In spite of this, coherent lower probabilities also have a number of drawbacks that hinder their use in the practice. For instance, their associated credal sets do not possess a straightforward representation in terms of extreme points; and their extension to lower *previsions* of gambles is not unique in general. For these reasons, it becomes interesting to approximate a coherent lower probability by a more tractable model. In a previous contribution [2], we did so by means of 2-monotone lower probabilities, that overcome some of the issues mentioned above: there is a simple procedure to determine the number of extreme points of their associated credal sets [3], and they can be uniquely extended to gambles by means of the Choquet integral [4].

Although our previous results are promising, the use of 2-monotone capacities is not without issues; the most important one, in our view, is the lack of a compelling interpretation of 2-monotonicity. This has led us to study the approximation of coherent lower probabilities by means of *completely* monotone lower probabilities, or belief functions. They have a number of advantages: first, they have a clear interpretation from Shafer's Evidence Theory [5]; they can be equivalently represented by means of multi-valued mappings [6]; and still they

are sufficiently general to include as particular cases many interesting models from imprecise probability theory, such as probability boxes [7] or possibility measures [8].

The rest of the contribution is organized as follows: after giving some preliminary concepts in Sect. 2, in Sect. 3 we deal with the problem of outer approximating a coherent lower probability. We recall our results for 2-monotone lower probabilities in Sect. 3.1, investigate the problem for belief functions in Sect. 3.2 and consider the particular case of possibility measures in Sect. 3.3. Some additional comments are given in Sect. 4. Due to space limitations, several results, comments as well as proofs have been omitted.

2 Preliminaries

Let $\mathcal{X} = \{x_1, \dots, x_n\}$ denote a finite universe with cardinality n . A *lower probability* on $\mathcal{P}(\mathcal{X})$ is a function $\underline{P} : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$. Under an epistemic interpretation, $\underline{P}(A)$ may be understood as a lower bound for the unknown probability $P_0(A)$ of the event A . In that case, the available information about the probability measure P_0 is given by the *credal set* associated with \underline{P} :

$$\mathcal{M}(\underline{P}) = \{P \text{ probability measure} \mid P(A) \geq \underline{P}(A) \forall A \subseteq \mathcal{X}\}.$$

The minimum requirement on \underline{P} we shall consider in this paper is that the bounds it provides for every event can be attained by some probability in $\mathcal{M}(\underline{P})$.

Definition 1. [1] A lower probability \underline{P} on $\mathcal{P}(\mathcal{X})$ is called *coherent* when its credal set $\mathcal{M}(\underline{P})$ is non-empty and $\underline{P}(A) = \min_{P \in \mathcal{M}(\underline{P})} P(A)$ for every $A \subseteq \mathcal{X}$.

The conjugate of a lower probability \underline{P} , denoted by \overline{P} , is called *upper probability* and it is given by $\overline{P}(A) = 1 - \underline{P}(A^c)$ for every $A \subseteq \mathcal{X}$. $\overline{P}(A)$ can be interpreted as an upper bound for the unknown probability of A . When \underline{P} is coherent, \overline{P} can also be computed by $\overline{P}(A) = \max\{P(A) \mid P \in \mathcal{M}(\underline{P})\}$ for every $A \subseteq \mathcal{X}$.

One very interesting property that a coherent lower probability may satisfy is that of k -monotonicity.

Definition 2. [4] A lower probability $\underline{P} : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$ is *k -monotone* if for every $p \leq k$, and for every $A_1, \dots, A_p \subseteq \mathcal{X}$ it holds that:

$$\underline{P}\left(\bigcup_{i=1}^p A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1, \dots, p\}} (-1)^{|I|+1} \underline{P}\left(\bigcap_{i \in I} A_i\right).$$

In particular, 2-monotone lower probabilities possess a number of interesting properties: for instance, the extreme points of their associated credal set can be easily determined using the permutations of the possibility space [3]; moreover, they have a unique extension as an expectation operator that preserves 2-monotonicity: their Choquet integral [9].

If \underline{P} is k -monotone for every k , it is called *completely monotone*. It corresponds to a *belief function* within evidence theory, and we shall denote it Bel in

this paper. The conjugate upper probability of a belief function is called *plausibility function* and we shall denote it Pl . A belief function can be equivalently expressed in terms of its *Möbius inverse*, which is given by [5]:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B) \quad \forall A \subseteq \mathcal{X}.$$

This function m satisfies $\sum_{A \subseteq \mathcal{X}} m(A) = 1$ and $m(A) \in [0, 1]$ for every $A \subseteq \mathcal{X}$. Conversely, m determines the belief function by:

$$Bel(A) = \sum_{B \subseteq A} m(B).$$

Given the Möbius inverse m , those events A with strictly positive mass, $m(A) > 0$, are called *focal events*.

A particular case of plausibility functions are the possibility measures. They are connected to the theory of fuzzy sets.

Definition 3. [10] A *possibility measure* $\Pi : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$ is a function satisfying $\Pi(\emptyset) = 0$, $\Pi(\mathcal{X}) = 1$ and $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}$ for every $A, B \subseteq \mathcal{X}$.

A possibility measure is an instance of plausibility function, while its conjugate *necessity* measure is a belief function. They correspond to the particular case when the focal events are nested by set inclusion, meaning that for every two focal events E_1, E_2 , either $E_1 \subseteq E_2$ or $E_2 \subseteq E_1$.

Notation: We shall denote by $\mathcal{C}_2, \mathcal{C}_\infty$ and \mathcal{C}_Π the classes of 2-monotone lower probabilities, belief functions and possibility measures on $\mathcal{P}(\mathcal{X})$, respectively.

3 Outer Approximations of Coherent Lower Probabilities

In a recent paper [2] we investigated how to approximate a coherent lower probability \underline{P} by a 2-monotone lower probability \underline{Q} that at the same time (a) does not introduce new information; (b) is as close as possible to the original model. In this way, if \mathcal{C} denotes a class of coherent lower probabilities, we said that $\underline{Q} \in \mathcal{C}$ is an *outer approximation* of \underline{P} in \mathcal{C} if $\underline{Q} \leq \underline{P}$, and it is called *undominated* if there is no $\underline{Q}' \in \mathcal{C}$ such that $\underline{Q} \leq \underline{Q}' \leq \underline{P}$.

3.1 Outer Approximations in \mathcal{C}_2

One important issue is that of determining how close the outer approximation is to the original model. In [2], in addition to discussing other possibilities, we proposed to use the distance put forward by Baroni and Vicig in [11], given by

$$d(\underline{P}, \underline{Q}) := \sum_{E \subseteq \mathcal{X}} (\underline{P}(E) - \underline{Q}(E)). \tag{1}$$

If we interpret $\underline{P}(E) - \underline{Q}(E)$ as the additional imprecision introduced on E when replacing $\underline{P}(E)$ with $\underline{Q}(E)$, then $d(\underline{P}, \underline{Q})$ can be understood as the total imprecision added by the outer approximation \underline{Q} .

In [2], we obtained undominated outer approximations in \mathcal{C}_2 by using a linear programming problem and minimizing the distance (1). Next proposition summarizes some of our results.

Proposition 1. [2] *Let \underline{P} be a coherent lower probability, and let $\mathcal{C}'_2(\underline{P})$ denote the class of undominated outer approximations of \underline{P} in \mathcal{C}_2 .*

1. $\mathcal{C}'_2(\underline{P})$ is non-empty, and may have infinite cardinality.
2. $\underline{Q}(\{x\}) = \underline{P}(\{x\})$ for every $x \in \mathcal{X}$ and every $\underline{Q} \in \mathcal{C}'_2(\underline{P})$.
3. $\underline{P}(A) = \max_{\underline{Q} \in \mathcal{C}'_2(\underline{P})} \underline{Q}(A)$ for every $A \subseteq \mathcal{X}$.

3.2 Outer Approximations in \mathcal{C}_∞

In this section, we outer approximate a coherent lower probability by means of a belief function. Similarly to our work in [2], we propose to obtain outer approximations that minimize the distance (1) between the initial lower probability \underline{P} and the belief function: $d(\underline{P}, Bel) = \sum_{E \subseteq \mathcal{X}} (\underline{P}(E) - Bel(E))$. In terms of the Möbius inverse, this can be equivalently expressed as:

$$d(\underline{P}, Bel) = \sum_{E \subseteq \mathcal{X}} \left(\underline{P}(E) - \sum_{B \subseteq E} m(B) \right). \quad (2)$$

Let $\mathcal{C}'_\infty(\underline{P})$ denote the class of undominated outer approximations of \underline{P} in \mathcal{C}_∞ .

Proposition 2. *Let $\underline{P} : \mathcal{P}(X) \rightarrow [0, 1]$ be a coherent lower probability, and consider the problem of minimizing (2) where m is subject to the following constraints:*

$$\sum_{B \subseteq \mathcal{X}} m(B) = 1, \quad m(B) \geq 0 \quad \forall B \subseteq \mathcal{X}. \quad (\text{LP-bel.1})$$

$$\sum_{B \subseteq E} m(B) \leq \underline{P}(E) \quad \forall E \subseteq \mathcal{X}. \quad (\text{LP-bel.2})$$

1. *The feasible region of this linear programming problem is non-empty.*
2. *Any optimal solution of the linear programming problem belongs to $\mathcal{C}'_\infty(\underline{P})$.*
3. *If for a fixed event A we add the constraint*

$$\sum_{B \subseteq A} m(B) = \underline{P}(A), \quad (\text{LP-bel.3A})$$

then the feasible region of the new linear programming problem is non-empty, any optimal solution Bel belongs to $\mathcal{C}'_\infty(\underline{P})$ and satisfies $Bel(A) = \underline{P}(A)$.

4. If $\mathcal{C}'_\infty(\underline{P})$ denotes the union, for every $A \subseteq X$, of the sets of belief functions that minimize (2) subject to (LP-bel.1)–(LP-bel.3A), then for any event E it holds that $\underline{P}(E) = \max_{Q \in \mathcal{C}'_\infty(\underline{P})} Q(E)$.

This result parallels much of our work in [2]: it tells us that we can obtain undominated outer approximations by means of linear programming, and that we can guarantee the equality $Bel(A) = \underline{P}(A)$ for a fixed event A just by adding the constraint (LP-bel.3A). Some detailed comments about the complexity associated with solving the linear programming problem (LP-bel.1)–(LP-bel.2) in Property 2 can be found in [12].

The main difference with Property 1 is that undominated outer approximations in $\mathcal{C}'_\infty(\underline{P})$ may not agree with \underline{P} on singletons, and also they may not determine the same order on \mathcal{X} . Since belief functions are in particular 2-monotone, any outer approximation in \mathcal{C}_∞ is also an outer approximation in \mathcal{C}_2 . However, we do not have the inclusion $\mathcal{C}'_\infty(\underline{P}) \subseteq \mathcal{C}'_2(\underline{P})$: an undominated outer approximation in \mathcal{C}_∞ may be dominated in \mathcal{C}_2 , as we shall see in Example 1.

3.3 Outer Approximations in \mathcal{C}_Π

We focus now on the subfamily of belief functions given by necessity measures. Taking conjugacy into account, a necessity measure N^* outer approximates a coherent lower probability \underline{P} if and only if its conjugate possibility measure Π^* outer approximates the conjugate upper probability \overline{P} of \underline{P} , in the sense that $\overline{P}(A) \leq \Pi^*(A)$ for every $A \subseteq \mathcal{X}$. Since possibility measures appear more frequently in the literature than necessity measures, we shall formulate the problem in this equivalent manner.

Let $\mathcal{C}'_\Pi(\overline{P})$ denote the class of possibility measures Π^* that outer approximate \overline{P} and are *non-dominating* in $\mathcal{C}_\Pi(\overline{P})$, meaning that there is no other Π' in $\mathcal{C}_\Pi(\overline{P})$ such that $\overline{P} \leq \Pi' \leq \Pi^*$. Our next result characterizes this class.

Proposition 3. *Let $\overline{P} : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$ be a coherent upper probability satisfying $\overline{P}(\{x_i\}) > 0$ for any $x_i \in \mathcal{X}$. For any permutation σ of $\{1, \dots, n\}$, define $\Pi_\sigma : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$ by:*

$$\begin{aligned} \Pi_\sigma(\{x_{\sigma(1)}\}) &= \overline{P}(\{x_{\sigma(1)}\}) \text{ and} \\ \Pi_\sigma(\{x_{\sigma(i)}\}) &= \max_{A \in \mathcal{A}_{\sigma(i)}} \overline{P}(A \cup \{x_{\sigma(i)}\}), \text{ where for every } i > 1: \\ \mathcal{A}_{\sigma(i)} &= \left\{ A \subseteq \{x_{\sigma(1)}, \dots, x_{\sigma(i-1)}\} \mid \overline{P}(A \cup \{x_{\sigma(i)}\}) > \max_{x \in A} \Pi_\sigma(\{x\}) \right\}, \end{aligned}$$

and let $\Pi_\sigma(A) = \max_{x \in A} \Pi_\sigma(\{x\})$ for every other $A \subseteq \mathcal{X}$. Then:

1. $\mathcal{C}'_\Pi(\overline{P}) = \{\Pi_\sigma : \sigma \in S_n\}$, where S_n is the set of permutations of $\{1, \dots, n\}$.
2. For every event $A \subseteq \mathcal{X}$, $\overline{P}(A) = \min_{\sigma \in S_n} \Pi_\sigma(A)$.

This result provides us with a simple constructive method for obtaining the undominated outer approximations of \overline{P} in \mathcal{C}_Π . We also deduce that there are at

most $n!$ different undominated outer approximations. It is not difficult to show that this bound is tight.

In this result, we are assuming that $\bar{P}(\{x_i\}) > 0$ for every $x_i \in \mathcal{X}$. This assumption is not restrictive: if we consider the set $\mathcal{X}^* = \{x \in \mathcal{X} \mid \bar{P}(\{x\}) > 0\}$, that is bound to be non-empty due to the coherence of \bar{P} , there exists a one-to-one correspondence between the credal sets $\mathcal{M}_1 := \{P : P(A) \leq \bar{P}(A) \forall A \subseteq \mathcal{X}\}$ and $\mathcal{M}_2 := \{P : P(A) \leq \bar{P}(A) \forall A \subseteq \mathcal{X}^*\}$, because any $P \in \mathcal{M}_1$ satisfies $P(\mathcal{X} \setminus \mathcal{X}^*) = 0$. As a consequence, any non-dominating outer approximation Π^* of the restriction of \bar{P} to $\mathcal{P}(\mathcal{X}^*)$ can be extended to a non-dominating outer approximation Π' of \bar{P} , simply by making $\Pi'(\{x\}) = \Pi^*(\{x\})$ if $x \in \mathcal{X}^*$, $\Pi'(\{x\}) = 0$ if $x \in \mathcal{X} \setminus \mathcal{X}^*$ and $\Pi'(A) = \max_{x \in A} \Pi'(\{x\}) \forall A \subseteq \mathcal{X}$.

Remark 1. A somewhat related procedure to that in Property 3 was considered by Dubois and Prade in [13] and [14, Sect. 3.3] with the name of *Optimal Mass Allocation Procedure*; they used it to deal with the problem of outer approximating belief functions by means of possibility measures. In their formulation, given a permutation σ , they consider the nested family of events $E_j^\sigma = \{x_{\sigma(1)}, \dots, x_{\sigma(j)}\}$ for $j = 1, \dots, n$. If A_1, \dots, A_k are the focal events of the initial belief function to be outer approximated, for every $i = 1, \dots, k$ they define the value $f_\sigma(i) = \min\{j \mid A_i \subseteq E_j^\sigma\}$, and from it they define the mass of E_j^σ by:

$$m^\sigma(E_j^\sigma) = \sum_{i: f_\sigma(i)=j} m(A_i), \quad \forall j = 1, \dots, n.$$

It holds that $m^\sigma(E_1^\sigma) + \dots + m^\sigma(E_n^\sigma) = 1$ and $E_1^\sigma \subseteq \dots \subseteq E_n^\sigma$, so m^σ defines a possibility measure by means of the formula $\Pi(A) = \sum_{E_j^\sigma \cap A \neq \emptyset} m^\sigma(E_j^\sigma)$. Although this possibility measure does not coincide with the one we have denoted Π_σ in Property 3, in the end both procedures give rise to all elements in $\mathcal{C}'_\Pi(\bar{P})$. Note, nevertheless, that the procedure in [14] may, unlike ours, also produce dominating outer approximations. \blacklozenge

Although Property 3 provides a procedure for determining non-dominating outer approximations in \mathcal{C}_Π , we should be aware that the non-dominating outer approximations in \mathcal{C}_Π may be conjugate to necessity measures that are dominated in \mathcal{C}_∞ , as our next example shows:

Example 1. Let us consider a four-element space \mathcal{X} and the lower probability \underline{P} given in Table 1. To see that it is coherent, note that it is the lower envelope of the probabilities $(0.1, 0, 0.4, 0.5)$, $(0.4, 0.1, 0.2, 0.3)$ and $(0.3, 0.3, 0, 0.4)$. If we minimize Eq. (2) with constraints (LP-bel.1)–(LP-bel.2), we obtain the optimal solutions Bel_1 and Bel_2 as well as their convex combinations. If we add the additional constraint (LP-bel.3A) with $A = \{x_3, x_4\}$, we obtain a linear programming problem with infinite solutions; one of them is Bel_3 . Table 1 also gives an undominated 2-monotone lower probability \underline{Q} that outer approximates \underline{P} . It holds that Bel_2 is dominated by \underline{Q} , whence we see that Bel_2 is an undominated outer approximation of \underline{P} in \mathcal{C}_∞ , but not in \mathcal{C}_2 .

Let us now apply the procedure in Property 3 to obtain the possibility measure associated with the permutation $\sigma_1 = (1, 2, 3, 4)$. First of all, we define $\Pi_{\sigma_1}(\{x_1\}) = \overline{P}(\{x_1\}) = 0.4$. Then:

$$\mathcal{A}_2 = \{A \subseteq \{x_1\} \mid \overline{P}(A \cup \{x_2\}) > \max_{x \in A} \Pi_{\sigma_1}(\{x\})\} = \{\emptyset, \{x_1\}\}, \text{ and}$$

$$\Pi_{\sigma_1}(\{x_2\}) = \max\{\overline{P}(\emptyset \cup \{x_2\}), \overline{P}(\{x_1\} \cup \{x_2\})\} = \overline{P}(\{x_1, x_2\}) = 0.6.$$

Iterating the procedure,

$$\mathcal{A}_3 = \{A \subseteq \{x_1, x_2\} \mid \overline{P}(A \cup \{x_3\}) > \max_{x \in A} \Pi_{\sigma_1}(\{x\})\} = \{\emptyset, \{x_1\}, \{x_1, x_2\}\},$$

whence $\Pi_{\sigma_1}(\{x_3\}) = \overline{P}(\{x_1, x_2, x_3\}) = 0.7$, and finally, $\Pi_{\sigma_1}(\{x_4\}) = 1$. The associated possibility measure is depicted in Table 1. Its conjugate necessity measure N_{σ_1} is dominated by Bel_3 . ◆

Table 1. Coherent lower probability from Example 1 and its outer approximations.

A	$\underline{P}(A)$	$\overline{P}(A)$	\underline{Q}	Bel_1	Bel_2	Bel_3	Π_{σ_1}	N_{σ_1}
$\{x_1\}$	0.1	0.4	0.1	0.1	0.1	0.1	0.4	0
$\{x_2\}$	0	0.3	0	0	0	0	0.6	0
$\{x_3\}$	0	0.4	0	0	0	0	0.7	0
$\{x_4\}$	0.3	0.5	0.3	0.3	0.3	0.3	1	0.3
$\{x_1, x_2\}$	0.1	0.6	0.1	0.1	0.1	0.1	0.6	0
$\{x_1, x_3\}$	0.3	0.6	0.3	0.2	0.3	0.1	0.7	0
$\{x_1, x_4\}$	0.6	0.7	0.5	0.6	0.5	0.6	1	0.3
$\{x_2, x_3\}$	0.3	0.4	0.2	0.3	0.2	0.2	0.7	0
$\{x_2, x_4\}$	0.4	0.7	0.4	0.3	0.4	0.3	1	0.3
$\{x_3, x_4\}$	0.4	0.9	0.4	0.3	0.3	0.4	1	0.4
$\{x_1, x_2, x_3\}$	0.5	0.7	0.5	0.5	0.5	0.4	0.7	0
$\{x_1, x_2, x_4\}$	0.6	1	0.6	0.6	0.6	0.6	1	0.3
$\{x_1, x_3, x_4\}$	0.7	1	0.7	0.7	0.7	0.7	1	0.4
$\{x_2, x_3, x_4\}$	0.6	0.9	0.6	0.6	0.6	0.6	1	0.6
\mathcal{X}	1	1	1	1	1	1	1	1

This example also shows that the non-dominating outer approximations in \mathcal{C}_Π do not preserve the order between the events, in the sense that $\overline{P}(A) = \overline{P}(B) \not\Rightarrow \Pi(A) = \Pi(B)$ and $\overline{P}(A) < \overline{P}(B) \not\Rightarrow \Pi(A) \leq \Pi(B)$. To see this, it suffices to compare \overline{P} and Π_{σ_1} on singletons. A procedure for defining non-dominating outer approximations in \mathcal{C}_Π that preserve the ordered preferences between the events can be found in [11, Sect. 6.3].

4 Conclusions

In this paper, we have investigated the problem of outer approximating a coherent lower probability by means of belief functions. We have focused on those belief functions that are at the same time as close as possible to the initial model, while not adding new information, and we have shown that we can obtain these by means of a linear programming problem, and that they allow us to retrieve the initial coherent lower probability.

In the particular case of possibility measures we have provided a constructive procedure for obtaining the non-dominating outer approximations, proving thus that their number is upper bounded by $n!$. Our procedure is related to the optimal mass allocation procedure of Dubois and Prade.

As future lines of research, we would like to consider other particular families of belief functions, such as probability boxes, and to look at the representation in terms of multi-valued mappings. In addition, we would like to investigate how to elicit an outer approximation among all of the possible ones.

Acknowledgements. We acknowledge the financial support by project TIN2014-59543-P.

References

1. Walley, P.: *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London (1991)
2. Montes, I., Miranda, E., Vicig, P.: 2-monotone outer approximations of coherent lower probabilities. *Int. J. Approximate Reasoning* **101**, 181–205 (2018)
3. Shapley, L.S.: Cores of convex games. *Int. J. Game Theor.* **1**, 11–26 (1971)
4. Choquet, G.: Theory of capacities. *Annales de l'Institut Fourier* **5**, 131–295 (1953–1954)
5. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
6. Nguyen, H.T.: On random sets and belief functions. *J. Math. Anal. Appl.* **65**(3), 531–542 (1978)
7. Troffaes, M.C.M., Destercke, S.: Probability boxes on totally preordered spaces for multivariate modelling. *Int. J. Approximate Reasoning* **52**(6), 767–791 (2011)
8. Dubois, D., Prade, H.: Possibility theory: qualitative and quantitative aspects. In: Smets, P. (ed.) *Handbook on Defeasible Reasoning and Uncertainty Management Systems*. Volume 1: Quantified Representation of Uncertainty and Imprecision, pp. 169–226. Kluwer Academic Publishers, Dordrecht (1998)
9. de Cooman, G., Troffaes, M.C.M., Miranda, E.: n -Monotone exact functionals. *J. Math. Anal. Appl.* **347**, 143–156 (2008)
10. Dubois, D., Prade, H.: *Possibility Theory*. Plenum Press, New York (1988)
11. Baroni, P., Vicig, P.: An uncertainty interchange format with imprecise probabilities. *Int. J. Approximate Reasoning* **40**, 147–180 (2005)
12. Quaeghebeur, E.: Completely monotone outer approximations of lower probabilities on finite possibility spaces. In: Li, S., Wang, X., Okazaki, Y., Kawabe, J., Murofushi, T., Guan, L. (eds.) *Nonlinear Mathematics for Uncertainty and its Applications*. *Advances in Intelligent and Soft Computing*, vol. 100. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22833-9_20

13. Dubois, D., Prade, H.: Fuzzy sets and statistical data. *Eur. J. Oper. Res.* **25**(3), 345–356 (1986)
14. Dubois, D., Prade, H.: Consonant approximations of belief functions. *Int. J. Approximate Reasoning* **4**(5–6), 419–449 (1990)



An Ordered Family of Consistency Measures of Belief Functions

Nadia Ben Abdallah¹(✉), Anne-Laure Josselme¹, and Frédéric Pichon²

¹ NATO STO Centre for Maritime Research and Experimentation,
Viale San Bartolomeo, 400, La Spezia, Italy
nadia.benabdallah@cmre.nato.int

² Université d'Artois, EA 3926, Laboratoire de Génie
Informatique et d'Automatique de l'Artois (LGI2A), 62400 Béthune, France

Abstract. We propose a family of measures of consistency (and induced conflict) derived from a definition of consistent belief functions introduced previously. Besides satisfying the desired properties of monotonicity, boundedness, and extreme values, the novel family encompasses the existing probabilistic and logical consistency measures which are shown to correspond to two extremes of the family (lower sharp bound and upper asymptotic limit respectively). We illustrate the definitions and measures of consistency within an example of vessel destination estimation with inconsistent sources.

1 Introduction

In maritime security, the measurement of inconsistency may reveal maritime anomalies such as vessels deviating from normalcy (*e.g.*, “off-route vessels”, “too fast vessels”) and those possibly spoofing the Automatic Identification System (AIS) signal (by, *e.g.*, changing their actual type, concealing their current position, hiding their actual destination) to hide suspect behaviour [1,2]. Having a sound and proper measurement of inconsistency or, equivalently, consistency is thus of paramount importance for such intelligent systems.

Theoretical research on (in)consistency was pioneered by the artificial intelligence community working on knowledge bases over logical languages. Classical logic is explosive, *i.e.*, everything is a consequence of an inconsistency, so solving inconsistent knowledge bases is a major challenge. A variety of approaches have been proposed in the literature. Hunter and Konieczny [3] introduced the minimal inconsistent sets, while some other authors [4,5] proposed to attach probabilities or degrees of beliefs to propositions rather than truth values.

The consistency notion plays also a central role in the belief function setting [6,7] as it is directly related to the way conflict between pieces of evidence may be defined: as the inconsistency yielded by their conjunctive combination [8]. The two notions of inconsistency and conflict have been subject to studies whose starting point was often the logical interpretation of belief functions. Cuzzolin [9] provided a definition of consistent belief functions as a counterpart of consistent

knowledge bases [10]. Destercke and Burger [8] proposed an axiomatic approach to conflict which extends the properties of conflict between sets. Recently, Pichon et al. [11] revisited and extended some of Destercke and Burger’s results. In particular, they proposed a novel family of consistency definitions that encompasses the so-called probabilistic and logical definitions proposed in [8].

In this paper, we pursue this work and propose a new parametrised family of consistency (and their associated conflict) measures, study their properties and illustrate their use and interest on a vessel destination estimation problem. It is organized as follows. Necessary concepts of belief function theory as well as classical consistency and conflict notions and measures are first recalled. In Sect. 3, a parameterised family of consistency (and their associated conflict) measures is unveiled and its special cases and properties are discussed. We conclude and sketch the steps for future work in Sect. 4.

2 Background

In this section, we provide a brief reminder of necessary concepts on belief functions and recall the existing consistency definitions and measures in this setting.

2.1 Uncertainty Representation with Belief Functions

Preliminaries. Let the belief about the actual value of an uncertain variable \mathbf{x} defined over a frame \mathcal{X} be represented by a mass function which is a mapping $m : 2^{\mathcal{X}} \rightarrow [0, 1]$ such that $\sum_{A \subseteq \mathcal{X}} m(A) = 1$. \mathcal{M} denotes the set of mass functions defined over \mathcal{X} . The set of focal sets of m is denoted $\mathcal{F}(m)$ and its cardinality is denoted $\mathfrak{F} := |\mathcal{F}(m)|$. We allow $m(\emptyset)$, the mass associated to the empty set, to be strictly positive, which captures the fact that the true value of \mathbf{x} may be outside the frame of discernment.

Example 1. We denote by $\mathcal{X} = \{\text{Imperia, Savona, Genova, La Spezia, Livorno}\} = \{d_1, d_2, d_3, d_4, d_5\}$ the set of possible destinations of a vessel. A cleaning-matching algorithm that “cleans” the AIS reported destination by formatting it in the standard format and matches it to a standard database (the World Port Index) of port names is applied. The algorithm identifies “SAVONA” as the closest name in the World Port Index with a confidence degree of 0.8, and identifies “SAVOONGA” (Alaska region) as a possible match, with a confidence of 0.2. This can be encoded by the following mass function: $m_1(d_2) = 0.8; m_1(\emptyset) = 0.2$.

Information encoded in a mass function m can be equivalently represented by different set-measures among which the plausibility Pl and the commonality q measures defined for every $A \in 2^{\mathcal{X}}$ by:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B); \quad q(A) = \sum_{B \supseteq A} m(B). \quad (1)$$

The contour function pl is defined over \mathcal{X} by:

$$pl(x) = Pl(\{x\}) = q(\{x\}). \quad (2)$$

When two mass functions provided by two sources inform on the same entity, their combination can be performed in several ways. In particular, if the sources are independent, the combination is performed using the conjunctive operator:

$$m_{1\odot 2}(A) = \sum_{B \cap C = A} m_1(B)m_2(C) \quad \forall A \subseteq \mathcal{X},$$

The commonality function satisfies:

$$q_{1\odot 2}(A) = q_1(A)q_2(A) \quad \forall A \subseteq \mathcal{X}. \tag{3}$$

Conflict and Consistency. Different definitions of (total) consistency have been proposed in the belief function literature, among which the so-called probabilistic and logical definitions [8]:

Definition 1 (Logical consistency [8]). A mass function m is logically consistent iff $\bigcap_{A \in \mathcal{F}} A \neq \emptyset$.

Definition 2 (Probabilistic consistency [8]). A mass function m is probabilistically consistent iff $m(\emptyset) = 0$.

Two desirable properties of consistency measures for a mass function m are also provided in [8]:

1. Property 1 (Bounded): A measure of consistency should be bounded.
2. Property 2 (Extreme consistent values): A measure of consistency should reach its maximal value if and only if m is totally consistent (according to the considered definition of total consistency), and its minimal value if and only if m is totally inconsistent, i.e., $m(\emptyset) = 1$.

Two consistency measures satisfying these properties for, respectively, Definitions 1 and 2 of total consistency, are [8]:

$$\phi_\pi(m) = \max_{x \in \mathcal{X}} pl(x); \quad \phi_m(m) = 1 - m(\emptyset). \tag{4}$$

Other measures have been proposed in the literature, such as Yager’s [12]:

$$\phi_Y(m) = \sum_{A \cap B \neq \emptyset} m(A)m(B). \tag{5}$$

A conflict measure $\kappa_x : \mathcal{M} \times \mathcal{M} \rightarrow [0, 1]$ between two mass functions can be defined from a consistency measure by:

$$\kappa_x(m_1, m_2) := 1 - \phi_x(m_{1\odot 2}), \tag{6}$$

where $x \in \{m, Y, \pi\}$. These three conflict measures have been proved in [8, 11] to satisfy a set of desirable axioms for a conflict measure proposed in [8].

2.2 *N*-consistency Definition and Measures

Probabilistic and logical consistency definitions require, respectively, each of the focal sets, and the intersection of all focal sets, to be non-empty. In-between properties of the focal sets may also be useful to capture as illustrated by the following example.

Example 2. We consider two other sources that inform about the destination of the vessel: a Track-to-Route algorithm (S_2) which associates a vessel to pre-computed maritime routes based on its kinematics features, and a Vessel Traffic Service (VTS) operator (S_3) who is monitoring the marine traffic for some port authority, and who based on experience provides a subjective assessment of the vessels destination. S_2 and S_3 provide the following assessments: $m_2(\{d_3\}) = 0.2$; $m_2(\{d_1, d_2, d_3\}) = 0.6$; $m_2(\{d_1, d_2\}) = 0.2$, and $m_3(\{d_4\}) = 0.1$; $m_3(\{d_3\}) = 0.1$; $m_3(\{d_1, d_2\}) = 0.8$.

Both mass functions are equally consistent according to the probabilistic and logical measures as we can see that they satisfy: $\phi_m(m_2) = \phi_m(m_3) = 1$ and $\phi_\pi(m_2) = \phi_\pi(m_3) = 0.8$. We can therefore not compare the two assessments in terms of internal consistency using these measures. However, if we refine the analysis and consider for instance the pairwise intersection of the focal sets, m_2 appears “less inconsistent” than m_3 since its focal sets are “more” (pairwise) intersecting. This suggests the definition of other measures of internal consistency that can capture refined notions of consistency based on the degree of intersection of the focal sets.

To this aim, we proposed recently in [11] a family of definitions of consistency:

Definition 3 (*N*-consistency [11]). *A mass function m is said to be consistent of order N (*N*-consistent for short), with $1 \leq N \leq \mathfrak{F}$, iff its focal sets are *N*-wise consistent, i.e., if $\forall \mathcal{F}_N \subseteq \mathcal{F}$ s.t. $|\mathcal{F}_N| = N$, we have:*

$$\bigcap_{A \in \mathcal{F}_N} A \neq \emptyset.$$

In addition, we proposed in [11] an associated family of consistency measures ϕ_N from \mathcal{M} to $[0, 1]$ defined for any $m \in \mathcal{M}$ and $1 \leq N \leq \mathfrak{F}$ by:

$$\phi_N(m) = 1 - m^{(N)}(\emptyset), \tag{7}$$

where $m^{(N)}$ denotes the result of the conjunctive combination of m with itself N times ($m^{(1)} = m$).

The family satisfies the following properties [11]:

- For every $N \in [1, \mathfrak{F}]$, ϕ_N satisfies Properties 1 and 2 in the case where total consistency is understood according to the *N*-consistency definition.
- Probabilistic and Yager consistency (definition and measure) coincide, respectively, with 1-consistency and 2-consistency.
- m is logically consistent iff it is \mathfrak{F} -consistent.

– The family is monotonic in N for any given mass function m :

$$\phi_1(m) = \phi_m(m) \geq \phi_2(m) \geq \dots \geq \phi_{\mathfrak{F}}(m).$$

Example 3. Going back to the previous example, we have:

$\phi_1(m_2) = 1 > \phi_2(m_2) = 0.92 > \phi_3(m_2) = \phi_{\mathfrak{F}}(m_2) = 0.88 > \phi_{\pi}(m_2) = 0.8$; and $\phi_1(m_3) = 1 > \phi_{\pi}(m_3) = 0.8 > \phi_2(m_3) = 0.66 > \phi_3(m_3) = \phi_{\mathfrak{F}}(m_3) = 0.514$. Using these measures, it becomes possible to compare m_2 and m_3 in terms of internal consistency: in particular, the intuition that m_2 appears less inconsistent than m_3 when considering pairwise consistency is captured by measure ϕ_2 .

It appears that the measure ϕ_{π} does not belong to and can not be ordered within the ϕ_N family. In particular, it is not possible to compare the two measures that capture the same notion of logical consistency: ϕ_{π} and $\phi_{\mathfrak{F}}$. In the following, we show that such a comparison becomes possible through a simple transformation of the ϕ_N family.

3 Monotonically Ordered Consistency Measures

In the following, we first propose a new family of consistency measures derived from the ϕ_N family (Sect. 3.1), and then show that the probabilistic ϕ_m and logical ϕ_{π} consistency measures belong to the family and can be ordered within it (Sect. 3.2).

3.1 A New Family of Consistency Measures

Definition 4. Let m be a mass function of \mathcal{M} . The ψ_N measure of N -consistency of m , for $N \in \mathbb{N}_{>0}$, is the measure $\psi_N : \mathcal{M} \rightarrow [0, 1]$ defined for any $m \in \mathcal{M}$ by:

$$\psi_N(m) := (1 - m^{(N)}(\emptyset))^{\frac{1}{N}}. \tag{8}$$

Note that the new proposed family is simply the N -th root of the probabilistic consistency of the family of mass functions $m^{(N)}$, $N \in \mathbb{N}_{>0}$, since:

$$\psi_N(m) = (\psi_1(m^{(N)}))^{\frac{1}{N}}. \tag{9}$$

It encompasses the probabilistic consistency measure which is retrieved when $N = 1$ and $\psi_1(m) = \phi_m(m)$. However, contrary to the ϕ_N family, the 2-consistency measure ψ_2 does not coincide anymore with Yager’s ϕ_Y . We however have that $\psi_2 = \sqrt{\phi_Y}$.

Proposition 1. ψ_N measures satisfy Properties 1 and 2 for the N -consistency definition.

Proof Sketch. This stems from the result that the measure ϕ_m , or equivalently ψ_1 , satisfies both properties and the relation (9) between the measures ψ_N and ψ_1 .

Although it is obvious that the family ϕ_N is monotonic in N for all $m \in \mathcal{M}$, the equivalent result for the family ψ_N still holds but is less trivial, as stated by the proposition below and following proof.

Proposition 2. *For all $m \in \mathcal{M}$ and $N \geq 1$:*

$$\psi_N(m) \geq \psi_{N+1}(m).$$

Proof Sketch. For any two mass functions m_1 and m_2 in \mathcal{M} we have:

$$m_{1 \odot_2}(\emptyset) \geq 1 - (1 - m_1(\emptyset))(1 - m_2(\emptyset)).$$

The right-hand value is reached when the non-empty focal sets of both mass functions intersect, in which case there is no creation of empty focal sets in the combination, and $m_{1 \odot_2}(\emptyset)$ is solely due to the propagation of the masses of the empty sets of both mass functions. When $m_1 = m_2$, it is easy to prove by recursion and using the previous inequality that the following relation holds between $m(\emptyset)$ and $m^{(N)}(\emptyset)$:

$$m^{(N)}(\emptyset) \geq 1 - (1 - m(\emptyset))^N, \text{ which is equivalent to } \phi_1(m) \geq (\phi_N(m))^{\frac{1}{N}}.$$

When $m_1 = m$ and $m_2 = m^{(N)}$, the first inequality yields: $m_{1 \odot_2}(\emptyset) = m^{(N+1)}(\emptyset) \geq 1 - (1 - m^{(N)}(\emptyset))(1 - m(\emptyset))$. Since $m(\emptyset)$ and $m^{(N)}(\emptyset)$ are related by the recursive relation, we can deduce that:
 $m^{(N+1)}(\emptyset) \geq 1 - (1 - m^{(N)}(\emptyset))(1 - m^{(N)}(\emptyset))^{\frac{1}{N}}$, i.e., $\psi_{N+1}(m) \leq \psi_N(m)$.

In the following, we study the relation between the existing and the new measures.

3.2 Relation with the Existing Measures

We are interested in studying the relation between the logical consistency measure ϕ_π and the proposed family, in particular $\psi_{\mathfrak{F}}$ since, we recall, $\psi_{\mathfrak{F}}$ captures the same definition of total consistency as ϕ_π .

We start by reporting a result on the relation between the first term of the family, i.e., the probabilistic consistency measure, and the logical one ϕ_π .

Lemma 1. *Every mass function $m \in \mathcal{M}$ with \mathfrak{F} focal sets satisfies:*

$$\psi_1(m) \geq \phi_\pi(m) \geq \frac{\psi_1(m)}{\mathfrak{F}^*},$$

where \mathfrak{F}^* denotes the number of non-empty focal sets of m .

Proof Sketch. The left-hand side of the inequality is a known result [8]. The right-hand part stems from observing that:

$$\phi_\pi(m) \geq \max_{A \in \mathcal{F}} (m(A)) \geq \frac{1 - m(\emptyset)}{\mathfrak{F}^*}.$$

Actually, the result in Lemma 1 holds between measures ϕ_π and ψ_N for all $N \geq 1$:

Proposition 3. *For every $m \in \mathcal{M}$ with \mathfrak{F} focal sets, and for every $N \geq 1$:*

$$\psi_N(m) \geq \phi_\pi(m) \geq \frac{\psi_N(m)}{(\mathfrak{F}_N^*)^{\frac{1}{N}}},$$

where \mathfrak{F}_N^* denotes the number of non-empty focal sets of $m^{(N)}$. Also, the series $\psi_N(m)$ converges asymptotically to $\phi_\pi(m)$:

$$\lim_{N \rightarrow \infty} \psi_N(m) = \phi_\pi(m).$$

Proof Sketch. The inequalities stem from Lemma 1 applied to the mass function $m^{(N)}$, together with Eq. (9) and the relation: $\phi_\pi(m^{(N)}) = (\phi_\pi(m))^N$ which stems from Eqs. (3) and (2). For the second part of the proposition, $\psi_N(m)$ is a decreasing bounded series, so it converges. Since the number of focal sets of $m^{(N)}$ stops increasing after \mathfrak{F} auto-combinations of m , then $\lim_{N \rightarrow \infty} (\mathfrak{F}_N^*)$ is a constant and $\lim_{N \rightarrow \infty} (\mathfrak{F}_N^*)^{\frac{1}{N}} = 1$.

By combining Propositions 2 and 3, it appears that the logical consistency measure corresponds to the upper asymptotic limit of the ψ_N family:

Proposition 4. *For every mass function m defined over \mathcal{X} with \mathfrak{F} focal sets:*

$$\phi_m(m) = \psi_1(m) \geq \psi_2(m) \geq \dots \geq \psi_{\mathfrak{F}}(m) \geq \phi_\pi(m) = \lim_{N \rightarrow \infty} \psi_N(m).$$

and for every pair m_1 and m_2 :

$$\kappa_m(m_1, m_2) \leq \kappa_2(m_1, m_2) \leq \dots \leq \kappa_{\mathfrak{F}_{12}}(m_1, m_2) \leq \kappa_\pi(m_1, m_2) = \lim_{N \rightarrow \infty} \kappa_N(m_1, m_2).$$

where $\kappa_N(m_1, m_2) = 1 - \psi_N(m_1 \odot_2 m_2)$ and \mathfrak{F}_{12} the number of focal sets of $m_1 \odot_2 m_2$.

The analysis of the internal consistency of the belief function resulting from the conjunctive combination of the belief functions issued by some sources, *i.e.*, of their conflict, can be used in several ways to improve the estimation confidence on the fusion output. This can be done by discounting or discarding the most conflicting sources [13], or re-questioning those that are inconsistent with a certain reference source. A deep analysis of the conflict is therefore necessary as illustrated hereafter.

Example 4. To estimate the destination of the vessel, both the Track-to-Route and VTS operator rely on some extra contextual information (S_4) encoded by a mass function m_4 . The mass functions m_2 and m_3 are actually the results of the conjunctive combination of some mass functions m_{2b} and m_{3b} encoding the specific sources knowledge and m_4 : $m_2 = m_{2b} \odot_4 m_4$; $m_3 = m_{3b} \odot_4 m_4$. We are interested in determining which of the sources is more in conflict with the contextual knowledge which is highly reliable and trusted. The conflict values are, using

$\kappa_N(m_{2b}, m_4) = 1 - \psi_N(m_2)$ and $\kappa_N(m_{3b}, m_4) = 1 - \psi_N(m_3)$:
 $\kappa_m(m_{2b}, m_4) = 0$; $\kappa_2(m_{2b}, m_4) = 0.041$; $\kappa_3(m_{2b}, m_4) = 0.042$; $\kappa_\pi(m_{2b}, m_4) = 0.2$;
 $\kappa_m(m_{3b}, m_4) = 0$; $\kappa_2(m_{3b}, m_4) = 0.187$; $\kappa_3(m_{3b}, m_4) = 0.198$; $\kappa_\pi(m_{3b}, m_4) = 0.2$.
 The probabilistic and logical conflict measures do not allow one to identify which of S_2 and S_3 is more in conflict with S_4 , while the in-between conflict measures do, and suggest that S_3 is more conflicting with S_4 than S_2 .

4 Conclusions and Future Work

In this paper, we proposed a parametrised family of consistency measures and illustrated its properties and interest with an example of multi-source vessel destination estimation problem. The family satisfies desired consistency measures properties such as boundedness and extreme values, and is monotonic. In addition, it subsumes the probabilistic and logical measures as, respectively, the lower sharp bound and the upper asymptotic limit.

In a future work, we will investigate the geometric interpretation of the proposed family of measures as well as the partial order induced by the vector $(\psi_1, \dots, \psi_{\mathfrak{F}})$ on the mass functions space. It will also be interesting to know whether the new conflict measures introduced in this paper satisfy the conflict axioms of [8]. Other open questions such as the choice of the “best” measure (or level of consistency) will also be addressed considering theoretical and practical aspects such as the computational cost or some user’s expectations about the measures semantics.

Acknowledgements. This research was supported by NATO Allied Command Transformation (ACT).

References

1. Pallotta, G., Jousselme, A.L.: Data-driven detection and context-based classification of maritime anomalies. In: Proceedings of FUSION, pp. 1152–1159 (2015)
2. Ray, C., Gallen, R., Iphar, C., Napoli, A., Bouju, A.: DeAIS project: detection of AIS spoofing and resulting risks. In: Proceedings of OCEANS, pp. 1–6 (2015)
3. Hunter, A., Konieczny, S.: Measuring inconsistency through minimal inconsistent sets. In: Proceedings of KR, pp. 358–366 (2008)
4. Knight, K.: Measuring inconsistency. *J. Philos. Log.* **31**(1), 77–98 (2002)
5. Ma, J., Liu, W., Miller, P.: A characteristic function approach to inconsistency measures for knowledge bases. In: Hüllermeier, E., Link, S., Fober, T., Seeger, B. (eds.) SUM 2012. LNCS (LNAI), vol. 7520, pp. 473–485. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33362-0_36
6. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
7. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**(2), 191–234 (1994)
8. Destercke, S., Burger, T.: Toward an axiomatic definition of conflict between belief functions. *IEEE Trans. Cybern.* **43**(2), 585–596 (2013)

9. Cuzzolin, F.: Consistent transformations of belief functions. arXiv preprint [arXiv:1407.8151](https://arxiv.org/abs/1407.8151) (2014)
10. Saffiotti, A.: A belief-function logic. In: Proceedings of AAAI, pp. 642–647 (1992)
11. Pichon, F., Jousselme, A.L., Ben Abdallah, N.: Several shades of conflict. Under revision (2018)
12. Yager, R.R.: On considerations of credibility of evidence. *Int. J. Approx. Reason.* **7**(1/2), 45–72 (1992)
13. Pichon, F., Destercke, S., Burger, T.: A consistency-specificity trade-off to select source behavior in information fusion. *IEEE Trans. Cybern.* **45**(4), 598–609 (2015)



Active Evidential Calibration of Binary SVM Classifiers

Sébastien Ramel^(✉), Frédéric Pichon, and François Delmotte

Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), 62400 Béthune, France
{sebastien.ramel, frederic.pichon, francois.delmotte}@univ-artois.fr

Abstract. Evidential calibration methods of binary classifiers improve upon probabilistic calibration methods by representing explicitly the calibration uncertainty due to the amount of training (labelled) data. This justified yet undesirable uncertainty can be reduced by adding training data, which are in general costly. Hence the need for strategies that, given a pool of unlabelled data, will point to interesting data to be labelled, *i.e.*, to data inducing a drop in uncertainty greater than a random selection. Two such strategies are considered in this paper and applied to an ensemble of binary SVM classifiers on some classical binary classification datasets. Experimental results show the interest of the approach.

Keywords: Belief functions · Evidential calibration · Active learning

1 Introduction

Probabilistic calibration methods, such as isotonic and logistic (Platt scaling) regressions, allow to learn from training data how to transform classifier outputs into probabilities that an instance belongs to each of the classes [1]. They are useful for the many applications where it is important to provide such probabilities rather than mere crisp decisions and where the available classifiers output scores, such as SVMs, or inaccurate probabilities, such as Naive Bayes [1, 2]. Besides, they have been mainly designed so far for binary classification.

A limitation of these methods is that they do not take into account the uncertainty due to the amount of training data in their probability estimates and, in particular, the less training data, the more uncertain the probability estimates [3]. To address this issue, the calibration problem has been considered recently in the framework of belief function theory, yielding so-called evidential calibration methods (see [3] for the calibration of a single binary classifier and [4] for the calibration of an ensemble of binary classifiers). These latter methods are able to represent explicitly the uncertainty due to the amount of training data, which is important in critical application domains and also leads to better classification performance than probabilistic calibration methods as shown in [3, 4].

While it is important to represent the aforementioned uncertainty, it is even better if this uncertainty is as small as possible. In order to reduce it, one needs to bring in additional training (labelled) data, which may be costly and hence must be done in an efficient manner, *i.e.*, such that for any given number of added labelled data the uncertainty is reduced as much as possible. It is a similar problem to that of active learning [5], except that the primary focus is not on improving accuracy but rather on reducing uncertainty, and it is the problem tackled in this paper.

Specifically, we consider the following setting: we assume an initial set of labelled data from which some classifiers can be evidentially calibrated, and then we consider that it is possible to ask iteratively an oracle to label some data from a pool of data with missing labels. We study two strategies to decide which instances from the pool should be given to the oracle. These strategies are in the spirit of the so-called uncertainty sampling strategy framework from active learning [5], where instances in the pool are ordered according to how much the current classifier is the most unsure about.

This paper is organized as follows. Section 2 recalls the necessary background on the evidential calibration of binary classifiers. Then, Sect. 3 presents two strategies for the active evidential calibration of such classifiers and reports experimental results when these strategies are applied to binary SVM classifiers. Finally, Sect. 4 concludes the paper.

2 Evidential Calibration

Evidential calibration of binary classifiers, as introduced in [3] for the case of a single classifier and further developed in [4] for an ensemble of classifiers, relies on some recent results by Kanjanatarakul *et al.* [6, 7] concerning the prediction of a Bernoulli random variable, which are recalled in the next section.

We will assume that the reader has some basic knowledge of the theory of belief functions (a reminder can be found in [7]).

2.1 Prediction of a Bernoulli Random Variable

Kanjanatarakul *et al.* [6, 7] proposed a general approach which, given some knowledge about some parameter θ obtained by observing a realization x of some random quantity X with distribution $f_\theta(x)$ and represented by a belief function Bel_x^θ ¹, makes it possible to make statements in the form of a belief function $Bel_x^{\mathbb{Y}}$ about some random quantity $Y \in \mathbb{Y}$ whose conditional distribution $g_{x,\theta}(y)$ given $X = x$ depends on θ .

¹ Bel_x^θ must be *induced by a source* [7]. It may be obtained by a number of evidential methods to statistical inference, and in particular the likelihood-based evidential method [8] in which case Bel_x^θ is the consonant belief function whose contour function is the normalized likelihood function given the observed data x .

In particular, if Y is a binary random variable ($\mathbb{Y} = \{0, 1\}$) with associated Bernoulli distribution $\mathcal{B}(\theta)$, $\theta \in [0, 1]$, and if Bel_x^\ominus is a consonant belief function whose associated contour function pl_x^\ominus is unimodal and continuous, we have [6]:

$$Bel_x^{\mathbb{Y}}(\{1\}) = \hat{\theta} - \int_0^{\hat{\theta}} pl_x^\ominus(u) du, \quad Pl_x^{\mathbb{Y}}(\{1\}) = \hat{\theta} + \int_{\hat{\theta}}^1 pl_x^\ominus(u) du, \quad (1)$$

where $\hat{\theta}$ maximizes pl_x^\ominus . The degree of belief $Bel_x^{\mathbb{Y}}(\{1\})$ represents the amount of evidence strictly supporting $Y = 1$ while the plausibility $Pl_x^{\mathbb{Y}}(\{1\}) = 1 - Bel_x^{\mathbb{Y}}(\{0\})$ is the amount of evidence not contradicting it. Besides, the difference $Pl_x^{\mathbb{Y}}(\{1\}) - Bel_x^{\mathbb{Y}}(\{1\})$, which is equal to the mass $m_x^{\mathbb{Y}}(\{0, 1\})$ assigned to the ignorance, is merely the area under the contour function pl_x^\ominus and the size of this area tends to 0 if, *e.g.*, X follows a binomial distribution with parameters n and θ , Bel_x^\ominus is obtained using the likelihood-based method and n tends to infinity [6].

2.2 Evidential Calibration Methods

Let $\mathcal{C} = \{(s_1, y_1), \dots, (s_n, y_n)\}$ be some training data in a binary classification problem, where $s_i \in \mathbb{S}$ for some domain \mathbb{S} is the output provided by a pre-trained classifier for the i -th training sample with label $y_i \in \{0, 1\}$. For a test sample of output $s \in \mathbb{S}$ and unknown label $y \in \{0, 1\}$, any evidential calibration method proposed in [3] returns two values: the belief $Bel_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$ and plausibility $Pl_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$ that $y = 1$. These methods obtain these two values by seeing the label y of the test sample as the realization of a random variable Y with a Bernoulli distribution $\mathcal{B}(\theta)$ given knowledge about θ represented by some consonant belief function $Bel_{\mathcal{C},s}^\ominus$ with contour function $pl_{\mathcal{C},s}^\ominus$ depending on \mathcal{C} and s , and by applying then to Y the prediction approach recalled in Sect. 2.1.

The only difference between the evidential calibration methods in [3] is thus the way $pl_{\mathcal{C},s}^\ominus$ is defined. There are indeed several ways to define $pl_{\mathcal{C},s}^\ominus$: it depends on which probabilistic calibration method is extended and on which evidential approach to statistical inference is used (see [3, Sect. 4] for details). In this paper, we focus on the evidential calibration methods where $pl_{\mathcal{C},s}^\ominus$ is obtained using the likelihood-based evidential approach to statistical inference, as Xu *et al.* [3] showed that this is the approach presenting overall the best performances.

More precisely, let us consider two cases: $\mathbb{S} = \{0, 1\}$ and $\mathbb{S} = \mathbb{R}$. The case $\mathbb{S} = \{0, 1\}$ corresponds to a classifier returning binary outputs and it will allow us to investigate in Sect. 3 the behaviours of our active evidential calibration strategies in a simple setting. The case $\mathbb{S} = \mathbb{R}$ corresponds to a classifier returning scores, such as a SVM classifier, and it will allow us to recall shortly and progressively the arguably most involved and best evidential calibration method considered so far to deal with an ensemble of classifiers – the behaviours of our active strategies with respect to this latter calibration scheme of an ensemble of classifiers will also be investigated in Sect. 3.

The case $\mathbb{S} = \{0, 1\}$ can be handled using the likelihood-based evidential extension of the binning calibration method [3], in which case we have²:

$$pl_{\mathcal{C},s}^{\Theta}(\theta) = \frac{\theta^{k_s}(1-\theta)^{n_s-k_s}}{\hat{\theta}_s^{k_s}(1-\hat{\theta}_s)^{n_s-k_s}}, \quad \forall s \in \mathbb{S}, \quad (2)$$

with $k_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s, y_i = 1\}|$, $n_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s\}|$ and $\hat{\theta}_s = k_s/n_s$.

The case $\mathbb{S} = \mathbb{R}$ can be handled using the likelihood-based evidential extension of the logistic regression [3], in which case $pl_{\mathcal{C},s}^{\Theta}$ is defined as:

$$pl_{\mathcal{C},s}^{\Theta}(\theta) = \sup_{\sigma_1 \in \mathbb{R}} pl_{\mathcal{C}}^{\Sigma}(\ln(\theta^{-1} - 1) - \sigma_1 s, \sigma_1), \quad \forall s \in \mathbb{S}, \quad (3)$$

with $pl_{\mathcal{C}}^{\Sigma}(\sigma) = \frac{L(\sigma)}{L(\hat{\sigma})}$, $\forall \sigma = (\sigma_0, \sigma_1) \in \Sigma = \mathbb{R}^2$, where $L(\sigma) = \prod_{i=1}^n p_i^{t_i} (1-p_i)^{1-t_i}$, with $p_i = \frac{1}{1+\exp(\sigma_0+\sigma_1 s_i)}$ and $t_i = \frac{N_1+1}{N_1+2}$ if $y_i = 1$, $t_i = \frac{1}{N_0+2}$ if $y_i = 0$, with $N_j = |\{(s_i, y_i) \in \mathcal{C} | y_i = j\}|$, and $\hat{\sigma}$ maximizing L .

Of particular interest is that using $pl_{\mathcal{C},s}^{\Theta}$ defined by (2) or by (3) in Eq. (1), $Pl_{\mathcal{C},s}^{\mathbb{Y}}(\{1\}) - Bel_{\mathcal{C},s}^{\mathbb{Y}}(\{1\}) = m_{\mathcal{C},s}^{\mathbb{Y}}(\{0, 1\})$ decreases as n increases [3]. In other words, $m_{\mathcal{C},s}^{\mathbb{Y}}(\{0, 1\})$ reflects the amount of training data, and in particular the less training data there are, the more ignorance or uncertainty there is.

Let us now consider a somewhat more complex problem, where we have an ensemble of m classifiers such that given a test sample of unknown label $y \in \{0, 1\}$, we obtain a vector of outputs $\mathbf{s} = (s^1, \dots, s^m) \in \mathbb{R}^m$ with s^j the output of the j -th classifier. In order to be able to interpret \mathbf{s} with respect to y , a solution proposed in [4] consists in calibrating *jointly* the classifiers. A joint calibration proceeds similarly as the calibration of a single classifier: the label y is seen as the realization of a random variable with a Bernoulli distribution $\mathcal{B}(\theta)$ and a belief function $Bel_{\mathcal{C},\mathbf{s}}^{\mathbb{Y}}$ is derived using the prediction approach (1) from knowledge on θ represented by a contour function $pl_{\mathcal{C},\mathbf{s}}^{\Theta}$ depending on \mathbf{s} and a training set $\mathcal{C} = \{(\mathbf{s}_1, y_1), \dots, (\mathbf{s}_n, y_n)\}$ where \mathbf{s}_i is the output vector provided by the m classifiers for the i -th training sample with label $y_i \in \{0, 1\}$. More specifically, Minary *et al.* [4] proposed an evidential joint calibration corresponding to the likelihood-based evidential extension of the multiple logistic regression, which is a generalization of the evidential logistic regression recalled above and, in particular, the definition of $pl_{\mathcal{C},\mathbf{s}}^{\Theta}$ derived in [4] is a straightforward multivariate generalization of (3) (due to lack of space, we refer the reader to [4] for the detailed definition of $pl_{\mathcal{C},\mathbf{s}}^{\Theta}$).

3 Active Evidential Calibration

As we have seen, evidential calibration methods return for a test sample with classifier output s a degree of belief $Bel_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$ and a plausibility $Pl_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$

² Equation (2) corresponds to a degenerate binning approach with only two bins. It can be derived rigorously without referring to the evidential binning calibration, by following a similar reasoning to the one used in [3] to obtain this latter calibration.

representing, respectively, the amount of evidence strictly supporting that the label y of the sample is 1 and the amount of evidence not contradicting it. Hence, the greater the interval $[Bel_{\mathcal{C},s}^{\mathbb{Y}}(\{1\}), Pl_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})]$, the more uncertain one is about the actual support that should be given to $y = 1$. It is thus clear that while it is important that uncertainty induced by the training data be represented, this uncertainty should be small enough otherwise no useful conclusion about y may be drawn, that is, the calibrated classifier is not useful.

In order to reduce the uncertainty, one needs to add some training (labelled) data. It is generally possible and relatively easy to obtain some unlabelled data but, depending on the domain, labelling it may be costly. Besides, it may be the case that not all training data are equivalent with respect to the drop in uncertainty that they induce. Hence, it seems useful to devise some strategies that, given a pool of unlabelled data, will point to interesting data to be labelled, that is, to data that will induce a drop in uncertainty greater than selecting at random data in the pool. We refer to such strategies as active evidential calibration strategies, or active strategies for short, in opposition to the passive strategy, which is the selection at random. We propose two such strategies in Sect. 3.1, which we then test on a single classifier and on an ensemble of classifiers in Sects. 3.2 and 3.3, respectively.

3.1 Active Strategies

In pool-based active learning [5], an active learner asks queries in the form of unlabelled instances (taken from the pool) to be labeled by an oracle, and the labeled instances are then moved to the learning set, with the aim that classification accuracy will improve faster than with a random selection strategy. Several query strategy frameworks have been proposed [5]. In particular, *uncertainty sampling* for a classifier with probabilistic outputs selects the unlabelled pool instance for which the classifier output has the greatest (Shannon) entropy.

Since our aim is to reduce the uncertainty represented by the quantity $m_{\mathcal{C},s}^{\mathbb{Y}}(\{0,1\})$ for any given test instance of score $s \in \mathbb{S}$, a natural query strategy is to select from a pool $\mathcal{P} = \{s_1^{\mathcal{P}}, \dots, s_p^{\mathcal{P}}\}$ of unlabelled instances with classifier outputs $s_k^{\mathcal{P}}$, $k = 1, \dots, p$, the instance $s^* \in \mathcal{P}$ that has the greatest uncertainty $m_{\mathcal{C},s^*}^{\mathbb{Y}}(\{0,1\})$. We note that an uncertainty measure for a mass function $m^{\mathbb{Y}}$ is the generalized Hartley measure [9], which evaluates its nonspecificity and is defined as $GH(m^{\mathbb{Y}}) := \sum_{A \subseteq \mathbb{Y}} m^{\mathbb{Y}}(A) \log_2 |A|$; if $\mathbb{Y} = \{0,1\}$, we have $GH(m^{\mathbb{Y}}) = m^{\mathbb{Y}}(\{0,1\})$. Hence, this strategy is similar to that of uncertainty sampling in active learning, except that it uses another uncertainty measure (the generalized Hartley measure instead of the Shannon entropy), and may thus be called Hartley Sampling (HS). It selects the instance $s_{HS}^* \in \mathcal{P}$ such that

$$s_{HS}^* = \arg \max_{s^{\mathcal{P}} \in \mathcal{P}} GH(m_{\mathcal{C},s^{\mathcal{P}}}^{\mathbb{Y}}). \quad (4)$$

In addition to the HS strategy, we consider for comparison purposes another query strategy, which is closer to uncertainty sampling of active learning: this

second strategy, called Pignistic Sampling (PS), selects the instance $s_{PS}^* \in \mathcal{P}$ whose associated pignistic probability distribution [10] denoted $BetP(m_{\mathcal{C}, s_{PS}^*}^{\mathbb{Y}})$ has the greatest (Shannon) entropy:

$$s_{PS}^* = \arg \max_{s^{\mathcal{P}} \in \mathcal{P}} H(BetP(m_{\mathcal{C}, s^{\mathcal{P}}}^{\mathbb{Y}})), \quad (5)$$

with $H(P)$ the Shannon entropy of probability distribution P . Note that since uncertainty sampling is designed to improve accuracy, one might expect that PS will improve accuracy, but it is not clear whether it will improve uncertainty.

Let us remark that the generalized Hartley measure and the Shannon entropy of the pignistic transformation have previously shown their interest in improving classification accuracy in the context of active classification [11].

3.2 Active Evidential Calibration of a Classifier with Binary Outputs

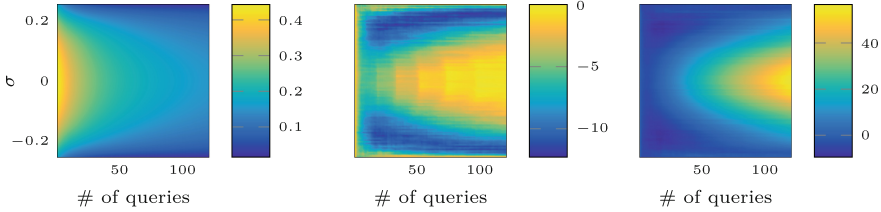
The active strategies described in the previous section are first tested with respect to a single classifier with binary outputs, *i.e.*, $\mathbb{S} = \{0, 1\}$, in which case the classifier is calibrated using (2). The test is conducted using simulated data.

Specifically, let $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, denote a given bivariate Bernoulli distribution for the pair (S, Y) of binary random variables S and Y , where S represents the classifier output and Y the true class. Such a distribution is completely characterized by the marginal probabilities $P(S = 1)$ and $P(Y = 1)$ and the covariance σ between S and Y [12].

In our experiment, we chose $P(S = 1) = P(Y = 1) = 0.5$ and considered all possible joint distributions $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, having those marginals: these are all the distributions that are obtained by choosing $\sigma \in [-0.25, 0.25]$, which is the range of possible values for σ given these marginals.

We drew randomly 10^6 samples in each of these joint distributions. We used a 1000-fold cross-validation procedure over these samples: the samples are randomly split into 1000 folds. Each fold (which contains 1000 samples) is in turn considered as the test set, and the other folds are combined to obtain a dataset which is randomly split into two parts: the first part composed of 10 instances is used as initial training data set \mathcal{C} for the evidential calibration of the classifier, and the second part composed of the remaining instances acts as the pool. The maximal number of queries for each query strategy (HS, PS and Random Sampling (RS)) was set to 120. For each fold used as test set and for each query strategy, we computed the average of the uncertainty, *i.e.*, ignorance with respect to the label after calibration, of the test instances as the number of queries increases. Finally, we averaged these latter averages over the 1000 test folds.

Figure 1 shows the performances in terms of uncertainty reduction achieved by the active strategies HS and PS with respect to the passive one (RS) used as reference. HS performs globally better than RS (up to 12% better) – it becomes equivalent to RS when σ gets closer to 0 and the number of queries increases, as



(a) Average uncertainty of RS (b) Uncertainty change over 1000-fold cross validation. (in %) from RS to HS. (c) Uncertainty change over 1000-fold cross validation. (in %) from RS to PS.

Fig. 1. Comparison of active strategies for a classifier with binary outputs.

well as when σ gets closer to -0.25 and 0.25 , which are all extreme dependence situations between S and Y . PS is beneficial with respect to RS for roughly the same zones as HS, albeit to a slightly lesser extent, but clearly detrimental (up to 55% worse) as the number of queries increases and as we get closer to $\sigma = 0$. Let us note that similar figures are obtained when other marginal probabilities $P(S = 1)$ and $P(Y = 1)$ are used (the figures are then somewhat distorted versions of the ones presented here).

3.3 Active Evidential Joint Calibration of Binary SVM Classifiers

The active strategies are now tested with respect to an ensemble of 3 SVM classifiers (trained with the LIBSVM library), which are jointly calibrated using the evidential multiple logistic regression described in Sect. 2.2. We used 6 binary classification datasets from the UCI repository: Australian, Heart, Ionosphere, Sonar, WDBC, Diabetes. Each dataset was randomly partitioned into 6 subsets: 3 subsets of 20 instances each to train each SVM, one subset of 100 instances to act as test set (except for Sonar, for which we used only 50 test samples due to its relatively small size), one subset of 10 instances to train the initial joint calibration of the classifiers, one subset containing the remaining instances and acting as the pool. Over the test set, we computed the average uncertainty of the strategies RS, HS and PS, as well as their Brier score (mean squared error), which is a standard performance (accuracy-like) measure for probabilistic calibration methods [1, 2] (to compute this score, we transformed the belief functions yielded by the evidential calibration into probability distributions using the pignistic transformation). We limited the number of queries to 20. The whole process was repeated for 100 rounds of random partitioning, and the obtained results were averaged over the rounds and then over the 6 datasets. These averages are presented in Fig. 2. As in the previous experiment, HS is better than PS to improve uncertainty, and this time PS is always better uncertainty-wise than RS. In addition, HS is better with respect to the Brier score than PS, which in turn improves upon RS. Overall, this experiment indicates that both strategies HS and PS may improve the uncertainty as well as the Brier score in comparison to RS, and that HS may be a better choice than PS.

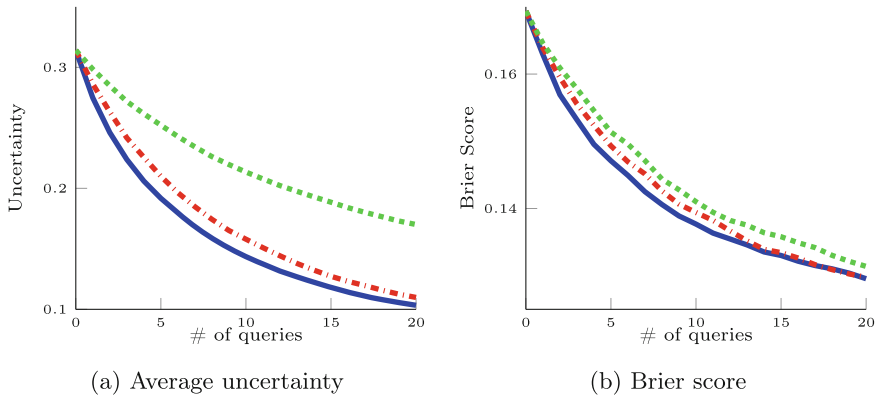


Fig. 2. Comparison of strategies HS (solid blue), PS (dash-dot red) and RS (dotted green) for an ensemble of SVM classifiers. (Color figure online)

4 Conclusions

In this paper, the benefits of two active strategies with respect to reducing the uncertainty (and also improving the performance) of the evidential calibration of binary classifiers were investigated. Preliminary experiments showed that while the Pignistic sampling strategy may be beneficial, it may be surpassed by Hartley sampling. Future works include conducting more extensive experiments (with other classifiers, datasets, calibration methods, training sets and pool sizes) to refine these conclusions, finding theoretical explanations for them in the spirit of those existing in active learning [5] and applying the approach to a driver state detection system whose calibration data are costly.

Acknowledgements. This work is funded in part by the ELSAT2020 project, which is co-financed by the European Union with the European Regional Development Fund, the French state and the Hauts de France Region Council.

References

1. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of ICML, pp. 625–632 (2005)
2. Zhong, W., Kwok, J.T.: Accurate probability calibration for multiple classifiers. In: Proceedings of IJCAI, pp. 1939–1945 (2013)
3. Xu, P., Davoine, F., Zha, H., Denooux, T.: Evidential calibration of binary SVM classifiers. *Int. J. Approx. Reason.* **72**, 55–70 (2016)
4. Minary, P., Pichon, F., Mercier, D., Lefevre, E., Droit, B.: Evidential joint calibration of binary SVM classifiers using logistic regression. In: Proceedings of SUM, pp. 405–411 (2017)
5. Settles, B.: Active learning literature survey. Computer Sciences Technical report 1648, University of Wisconsin-Madison (2009)

6. Kanjanatarakul, O., Sriboonchitta, S., Dencœux, T.: Forecasting using belief functions: an application to marketing econometrics. *Int. J. Approx. Reason.* **55**(5), 1113–1128 (2014)
7. Kanjanatarakul, O., Dencœux, T., Sriboonchitta, S.: Prediction of future observations using belief functions: a likelihood-based approach. *Int. J. Approx. Reason.* **72**, 71–94 (2016)
8. Dencœux, T.: Likelihood-based belief function: Justification and some extensions to low-quality data. *Int. J. Approx. Reason.* **55**(7), 1535–1547 (2014)
9. Klir, G.J.: *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley, Hoboken (2005)
10. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**(2), 191–234 (1994)
11. Reineking, T.: Active classification using belief functions and information gain maximization. *Int. J. Approx. Reason.* **72**, 43–54 (2016)
12. Teugels, J.L.: Some representation of the multivariate Bernoulli and binomial distributions. *J. Multivar. Anal.* **32**, 256–268 (1990)



Decision Making: A Beliefs, Preferences and Constraints Model

Aouatef Rouahi^{1(✉)}, Kais Ben Salah^{2(✉)}, and Khaled Ghédira^{3(✉)}

¹ ISG of Tunis, Tunis University, Tunis, Tunisia
rouahi.aouatef@hotmail.fr

² FCIT, University of Jeddah, Jeddah, Saudi Arabia
k.bensalah@uj.edu.sa

³ Central University, Tunis, Tunisia
khaled.ghedira@universitecentrale.tn

Abstract. Beliefs, preferences and constraints occur together in many real world problems. However, reasoning with such intertwinement is rather unexplored in the AI literature. In this paper, we introduce a model whereby agents seek for decisions that satisfy their preferences based on their beliefs subject to certain constraints by extending the soft constraints framework to the belief function theory. Constraint-based solving machinery are then adapted for solving such kind of problems. A specific branch and bound algorithm is introduced.

Keywords: Beliefs · Preferences · Constraints
Belief function theory · Soft constraints · Uncertainty

1 Introduction

Motivation. In our everyday tasks, decision-making stems from the interplay of beliefs, preferences and constraints. In spite of its importance, such intertwinement is rather unexplored in the AI literature. In this paper, we propose a BPC¹ model whereby agents seek for decisions that satisfy their preferences based on their beliefs subject to certain constraints. Under the assumptions that the agents act with full external and internal information, beliefs are often ignored or confounded with preferences. Far from being primitive, agents' preferences should depend on their beliefs about the properties and the outcomes of the alternatives especially in those situations in which an agent may only have partial information about alternatives, i.e., ill-defined alternatives. In addition, even if information is available, it can be ambiguous, contradictory or excessive. Thus, belief modeling separately from preferences is needed, so an agent can express his hesitation. Accordingly, the notion of belief-based preference is introduced. Once preferences are fixed, decisions can be inferred given the constraints that determine which alternatives are feasible. The belief function

¹ BPC stands for Beliefs, Preferences and Constraints.

theory [4, 10, 11] offers a sound mathematical basis that faithfully recognizes all our belief states. Moreover, by considering its Transferable Belief Model (TBM) interpretation [11], we introduce a two-level preference perspective: the belief base and the preference derived from it. Soft constraints [3], equipped with a powerful solving machinery, provide an interesting way to model and reason with quantitative preference relations and constraints. Our purpose in this paper is to bring into sharper focus the interesting interplay between beliefs, preferences and constraints by introducing an extended soft constraint formalism to the belief function theory.

Related Work. While intertwined preferences and constraints is thoroughly studied in the AI literature, beliefs and preferences are rarely considered except some studies of the logic of preference [7, 8] investigating preference dynamics under belief change. To the best of our knowledge, in the constraint satisfaction field, this work provides the first connection between the belief function theory and soft constraints machinery. Nevertheless, a variety of proposals has been introduced to extend soft constraints framework to deal with imperfect preferences without referring to the imperfection origin. The work in [5] considers incomplete soft constraint problems where some of the preferences may be allowed to be missing as long as it is feasible to find an optimal solution, otherwise, the agent will be required to add some preferences. In this work, incompleteness is interpreted as temporary inability or unwillingness to provide preferences over some alternatives. In our approach, we consider incompleteness as a decisive undesirability to compare some alternatives with regard to the available evidence, thus, we do not require the agent to supply further information. Another proposal considers preference intervals [6] to model imprecision in preference intensity. In our work, we assume that the preference intensities are, precisely, stated. However, we permit ties in the preference list. Other work that addresses uncertainty in soft constraints using the possibility theory is shown in [9] where some alternatives may be ill-defined, i.e., one cannot decide their values. In our work, we, thoroughly, address uncertainty in the case where the alternatives may be ill-defined so the agent cannot express his preferences in the form of “yes/no” but he could reply “I somewhat prefer this alternative”, i.e., preference intensity, or he may hesitate to express his hesitation by replying “I am not sure”, or he may simply say “I do not know” to express his ignorance.

The remainder of this paper is organized as follows: Sect. 2 reviews some preliminaries. We present the BPC model and its basic components in Sect. 3. In Sect. 4, reasoning with preferences and constraints to construct solutions are discussed and a specific branch and bound algorithm is introduced. Conclusions and further researches are drawn in Sect. 5.

2 Preliminaries

2.1 Belief Function Theory

The belief function theory was first initiated by [4] and then extended by [10]. Several interpretations have been introduced such as the well known TBM

established by [11]. Let Θ be a frame of discernment representing a finite set of elementary alternatives. A basic belief assignment (bba) m is the mapping from elements of the power set 2^Θ to $[0, 1]$ such that $\sum_{\theta \in 2^\Theta} m(\theta) = 1$.

The basic belief mass (bbm) $m(\theta)$, assigned to some subset θ of Θ , is a positive finite amount of support that is derived from the available pieces of evidence and exactly given to the set θ and not to any specific subset of θ by lack of evidence. While constraining $m(\emptyset) = 0$ corresponds to a closed-world assumption [10], allowing $m(\emptyset) \geq 0$ corresponds to an open world assumption [11] where the empty set bbm is considered as the internal conflict within an individual bba [1].

Given a set of alternatives Θ and a given bba m , we want to establish an ordering over Θ based on m . Many decision-making criteria have been developed in the literature. We are interested in decision based on maximum of pignistic probability (*BetP*) that offers a compromise between pessimistic and optimistic strategies, where higher probability degree indicates more preferred alternative. Hence, The bba m is reformed to a subjective probability measure *BetP* as follows: $BetP(A) = \frac{1}{1-m(\emptyset)} \sum_{\theta \subseteq A} \frac{|A \cap \theta| \cdot m(\theta)}{|\theta|}; \forall A \in \Theta$.

2.2 Soft Constraints

Soft constraints framework, namely Semiring-based CSP (SCSP)[3] is a generic framework to quantitative preferences covering many specific others. Given a c-semiring $S = \langle A, +, \times, 0, 1 \rangle$, a finite set D , and an ordered set of variables V , a soft constraint is a pair $\langle def, con \rangle$ where $con \subseteq V$ and $def : D^{|con|} \rightarrow A$. The associated degree from A with each alternative indicates to which extent it is preferred. The operations $(+)$ and (\times) are respectively used for comparing and combining preference degrees in order to select the best solution. We refer the reader to [2] for details.

3 Beliefs, Preferences and Constraints Model

A BPC model \wp is a tuple $(X, D, B - Pref, Cons)$, involving a finite set of variables X , its associated finite domains D and a finite set of belief-based preferences $B - Pref$. A belief-based preference $b - pref$ is a belief-soft constraint defined by the tuple (S, A, B, R) , where, $S \subseteq X$ is the scope of the preference delimiting the set of variables That $b - pref$ involves, $A \subseteq D^{|S|}$ is the set of the alternatives on which the preference relation is established, B is the belief base on A and R is the derived preference relation from B .

Example 1. Peter is buying an evening outfit (Dress-Shoes-Bag) for his wife Alice on an e-commerce shop for their first wedding anniversary. The shop provides tailored recommendations to their costumers based on their belief-based preferences and constraints. We have $X = \{Dr(dress), S(shoes), B(bag)\}$ with $D(Dr) = \{D_b(\text{black dress}) \$1000; D_r(\text{red dress}) \$650\}$, $D(S) = \{S_w(\text{white shoes}) \$300; S_r(\text{red shoes}) \$185\}$, $D(B) = \{B_b(\text{black bag}) \$100; B_r(\text{red bag}) \$60\}$.

Each item is associated with its equivalent price. As the outfit is for his wife, Peter has some beliefs about his preferences. Due to budget constraints, Peter cannot afford more than \$1250 for the outfit.

3.1 Belief Modeling

Given the scope of the preference S and the related set of alternatives A , the agent’s beliefs B over A are modeled in terms of a partial order \succeq induced by the bba m on $\{a_i \cup \emptyset\}$ to $[0,1]$, such that, $\bigcup_{(i=1)}^k a_i = A: \succeq = \{(\theta_1, \theta_2) | m(\theta_1) \geq m(\theta_2)\}$. The instance $\theta_1 \succeq \theta_2$ stands for “the betterness of θ_1 is at least as supported as the betterness of θ_2 ”, \succeq giving the evidence held by the agent is reflexive, transitive and antisymmetric as its associated strict component \triangleright (“is strictly supported to”) is irreflexive, transitive and asymmetric, its indifference component \equiv (“is as supported as”) is reflexive, symmetric and composed of (θ, θ) pairs only, and its associated incomparability relation \bowtie (“is incomparable to”) is irreflexive, not transitive and symmetric. In *Example 1.*, The belief bases of Peter is shown in Table 1.

Table 1. The belief bases induced from *Example 1.*

	$b - pref_1$	$b - pref_2$	$b - pref_3$
S	$\{Dr\}$	$\{Dr, S\}$	$\{S, B\}$
A	$\{D_b, D_r\}$	$\{(D_b, S_w), (D_b, S_r), (D_r, S_w), (D_r, S_r)\}$	$\{(S_w, B_b), (S_w, B_r), (S_r, B_b), (S_r, B_r)\}$
B	$D_b:0.7$ $D_r:0.3$	$(D_r, S_w):0.6$ $(D_b, S_w), (D_b, S_r):0.4$ $(D_r, S_r):0$	$(S_r, B_b):0.4$ $(S_w, B_b), (S_w, B_r):0.3$ $(S_r, B_r):0.2$ $\emptyset:0.1$
\succeq instances	$D_b \triangleright S_r$	$(D_r, S_w) \bowtie (D_b, S_w)^a$	$(S_r, B_b) \equiv (S_r, B_b)$

^a (D_r, S_w) is incomparable to (D_b, S_w) because we do not know the exact associated bba to (D_b, S_w) as it is tied with (D_b, S_r) by lack of evidence.

By means of our two-level preference approach, we have been able to capture all the epistemic states of the agent towards his preferences: full knowledge (e.g., well-informed agent); partial ignorance (e.g., $b - pref_2$ in Table 1 where some alternatives are tied); total ignorance if the only supported element is A; hesitation where the agent may want to express his beliefs about his preferences with some degree of hesitation (e.g., the bba associated with the empty set in $b - pref_3$) instead of expressing his total ignorance; null support where the agent has no evidence to believe that an alternative can be somehow good or bad.

3.2 Preference Deriving

As the agent belief base is outlined, his preference relations are derived as a total preorder \succeq^B induced by the *BetP* measures over A , the set of alternatives: $\succeq^B = \{(a_1, a_2) | (BetP(a_1) \geq BetP(a_2))\}$.

The relation \succeq^B is reflexive, complete and transitive. Given the belief-based preference relation \succeq^B and two alternatives $a_1, a_2 \in A$, we distinguish between two relations over a_1 and a_2 :

- a_1 is strictly preferred to a_2 , denoted by $a_1 \succ^B a_2$, when $a_1 \succeq^B a_2$ holds but $a_2 \succeq^B a_1$ does not. \succ^B is irreflexive, transitive and asymmetric.
- a_1 is indifferent to a_2 denoted by $a_1 \approx^B a_2$, when both $a_1 \succeq^B a_2$ and $a_2 \succeq^B a_1$ hold. \approx^B is reflexive, transitive and symmetric.

The derived preference relations from belief bases in *Example 1* are shown in Table 2.

Table 2. The derived preference induced from *Example 1*.

	$b - pref_1$	$b - pref_2$	$b - pref_3$
R	$D_b:0.7$ $D_r:0.3$	$(D_r, S_w):0.6$ $(D_b, S_w):0.2$ $(D_b, S_r):0.2$ $(D_r, S_r):0$	$(S_r, B_b):0.44$ $(S_w, B_b):0.17$ $(S_w, B_r):0.17$ $(S_r, B_r):0.22$
\succeq^B instances	$D_b \succ^B D_r$	$(D_r, S_w) \succ^B (D_r, S_r)$	$(S_w, B_b) \approx^B (S_w, B_r)$

3.3 Constraints Modeling

Constraints represent limitations that winnow the set of alternatives we can opt for in a given situation. In *Example 1*, Peter has one constraint $c1(\sum_{i=1..3} p_i \leq \$1250)$, where p_1, p_2 and p_3 are respectively the prices of the dress, the shoes and the bag. Once preferences and constraints are given, decisions are determinative.

4 Reasoning with Preferences and Constraints

Let S be a set of variables, we will use the notation ω_S to denote an outcome resulting from assigning a value to each variable in S from its equivalent domain. We will say that an outcome is complete iff it is defined on X , otherwise it is said to be partial. Consider a $b - pref_i$ defined on the set of variables S_i , $\delta(i, \omega_{S_i}) = BetP_i(\omega_{S_i})$ will denote the satisfaction degree of $b - pref_i$ by some outcome $\omega_{S_i} \in A_i$. $b - pref_i$ is said to be satisfied by ω_{S_i} , noted $\omega_{S_i} \models b - pref_i$, iff $\delta(i, \omega_{S_i}) > 0$. Solving a BPC problem consists in finding a complete outcome ω_X^* , if it exists, that satisfies all the constraints in *Cons* and is optimal with respect to the preferences in $B - Pref$.

4.1 Operations on Preferences

Given a $b - pref_i$ defined on S_i and some outcome ω_S such that $S_i \subseteq S \subseteq X$, then $\delta(i, \omega_S) = \delta(i, \omega \downarrow_{S_i}^S)$ ² will be the local satisfaction degree of $b - pref_i$ by ω_S . Hence, The global degree of joint satisfaction of the set of n belief-based preferences $B - Pref$ defined on the set of variables X by a given complete outcome ω_X is obtained by combining the local satisfaction degrees as follows: $\delta(B - Pref, \omega_X) = \otimes \delta(i, \omega_X) = \otimes \delta(i, \omega \downarrow_{S_i}^X), \forall i = 1..n$.

Different combination operators \otimes can be used that reflect various attitudes towards preferences satisfaction such as Min, Max, Product and Average combinations. Due to limited space, we will only adopt the product combination approach that offers more discriminating ordering than the Min and Max combinations and does not tolerate the falsification of any preference at variance with the average combination.

Given a $b - pref_i$ defined on S_i and some outcome ω_S such that $S \subseteq S_i \subseteq X$, then the estimated satisfaction degree of $b - pref_i$ by ω_S will be $\delta^e(i, \omega_S) = Max\{\delta(i, \omega_{S_i}) | \omega_{S_i} \in \Omega_{\omega \uparrow_{S_i}^S}\}$ ³.

4.2 Constructing Solutions

Given a BPC problem $\wp(X, D, B - Pref, Cons)$, every feasible complete outcome with respect to $Cons$ that jointly satisfies $B - Pref$ to a global satisfaction degree greater than 0 (whatever the used approach), is considered as a solution: $\omega_X \in S(P) \Leftrightarrow \omega_X \models Cons \wedge \delta(B - Pref, \omega_X) > 0$.

The global satisfaction degree induces a total preorder over the set of feasible outcomes, so that the best outcome will be the one that, maximally, satisfies B-Pref: $\omega_X^* = argmax_{\omega_X \in S(P)} \delta(B - Pref, \omega_X)$.

4.3 PDBB Algorithm

Commonly, when solving such constrained optimization problems, Branch and Bound (BB) algorithm is the most widely used. It incrementally builds, by assigning a variable with a value selected from its domain, outcomes prospected to be solutions, where it early on aborts every partial outcome that cannot be extended to construct a better solution than the one found so far using some upper (**B**) and lower (**b**) bounds. At each level, instead of assigning one variable with a value from its domain, we propose to assign multiple variables with values from the preference relation covering those variables, hence, the Preference-Directed BB (PDBB: see Algorithm 1). We illustrate, in Fig. 1, the

² $\omega \downarrow_{S_i}^S = \omega_{S_i} = \{(v_{i1}, \dots, v_{im}) | v_{ik} = v_j \text{ if } x_{ik} = x_j\}$ such that $S = \{x_1, \dots, x_l\}$ and $S_i = \{x_{i1}, \dots, x_{im}\}$.

³ $\Omega_{\omega \uparrow_{S_i}^S}$ is the set of outcomes resulted from the extension of ω_S from S to S_i .

PDBB execution to solve the problem described in *Example 1*. in Sect. 3. In Fig. 1, the outcomes having a red (X) are discarded because they are unfeasible, however, the outcomes with green (X) are aborted because they cannot lead to a better solution. For the *Example 1*, the best outfit for Peter's wife is { Dress: black; Shoes: red; Bag: red}.

Algorithm 1. PDBB Algorithm

```

input :  $(X, R_c, Cons, S, \omega_S, B, b)$ 
/*  $R_c = \{R_i | \bigcup_{i=1}^m S_i = X\}$  is minimal;  $S = \emptyset; \omega_S = \emptyset; B = 0; b = 1$  */
output:  $(\omega_X^*, B)$ 

while  $R_c$  is not empty do
  select and remove a relation  $R_i \in R_c$  ;
   $S \leftarrow S \cup S_i$ ;
  while  $A_i$  is not empty do
    select and remove best  $a \in A_i$ ;
     $\omega_S \leftarrow \omega_S \cup a$ ;
    if  $\omega_S \models Cons$  then
      Compute a lower bound b for  $\omega_S$ ;
      /*  $b = \delta(B - pref_a, \omega_S) \otimes \delta^e(B - pref_{\bar{a}}, \omega_S)$  such that B-Pref $_a$ 
         is the set of preferences activated by the current
         assignment and B-Pref $_{\bar{a}}$  is the rest of B-Pref that are
         not yet implied. */
      if  $b > B$  then
        if  $S = X$  then
           $B \leftarrow b$ ;
           $\omega_X^* \leftarrow \omega_S$ ;
          Print  $(\omega_X^*, B)$ ;
          if  $B = 1$  then
            return "Finished";
        else
          PDBB( $X, R_c, Cons, S, \omega_S, B, b$ );
      else
        PDBB( $X, R_c, Cons, S, \omega_S - a, B, b$ );
    else
      PDBB( $X, R_c, Cons, S, \omega_S - a, B, b$ );
  else
    PDBB( $X, R_c, Cons, S, \omega_S - a, B, b$ );

```

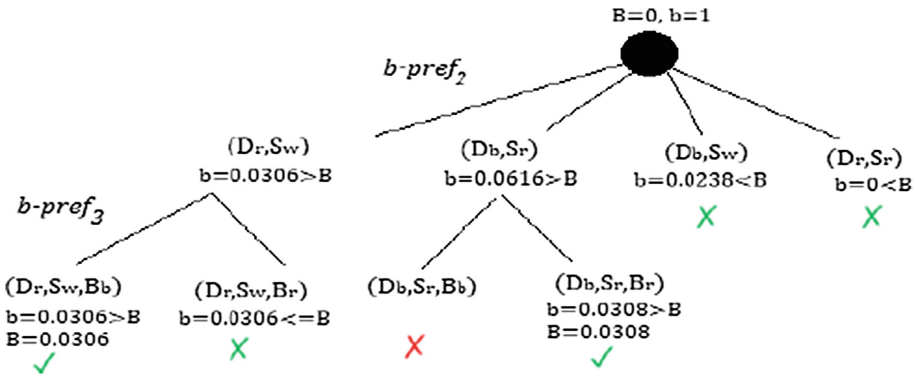


Fig. 1. PDBB for the problem in *Example 1*.

5 Conclusion and Further Work

We have introduced a decision making model whereby agents seek for decisions that satisfy their preferences based on their beliefs subject to certain constraints. Due to space limit, we could not report the experimental results, however it was proven that the PDBB is less costly than the classical BB w.r.t the number of visited nodes. The outlined PDBB search could be improved by introducing heuristics for the order of checking the preferences. Further research targets exploiting the expressiveness offered by the belief-based preferences model in order to enlarge the scope of the issues that can be tackled such as prioritized preferences, preference change using the belief revision process. We can also address the bipolar preferences exploiting the negative and positive belief notions. We also intend to introduce the weak preference relation using thresholds. Finally, we plan to explore how our model can be employed in decision support applications like recommender systems and combinatorial auctions.

References

1. Ayoun, A., Smets, P.: Data association in multi-target detection using the transferable belief model. *Int. J. Intell. Syst.* **16**(10), 1167–1182 (2001)
2. Barták, R.: Modelling soft constraints: a survey. *Neural Netw. World* **12**, 421–431 (2002)
3. Bistarelli, S., Montanari, U., Rossi, F.: Constraint solving over semirings. In: *Proceedings of the IJCAI95*. Morgan Kaufman (1995)
4. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**(2), 325–339 (1967)
5. Gelain, M., Pini, M.S., Rossi, F., Venable, K.B.: Dealing with incomplete preferences in soft constraint problems. In: Bessière, C. (ed.) *CP 2007*. LNCS, vol. 4741, pp. 286–300. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74970-7_22

6. Gelain, M., Pini, M.S., Rossi, F., Venable, K.B., Wilson, N.: Interval-valued soft constraint problems. *Ann. Math. Artif. Intell.* **58**(3–4), 261–298 (2010)
7. Lang, J., van der Torre, L.: Preference change triggered by belief change: a principled approach. In: Bonanno, G., Löwe, B., van der Hoek, W. (eds.) *LOFT 2008*. LNCS (LNAI), vol. 6006, pp. 86–111. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15164-4_5
8. Liu, F.: Changing for the better: preference dynamics and agent diversity. Ph.D. dissertation, University of Amsterdam (2008)
9. Pini, M.S., Rossi, F.: Uncertainty in soft constraint problems. In: van Beek, P. (ed.) *CP 2005*. LNCS, vol. 3709, pp. 865–865. Springer, Heidelberg (2005). https://doi.org/10.1007/11564751_103
10. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ (1976)
11. Smets, P., Kennes, R.: The transferable belief model. *Artif. Intell.* **66**, 191–234 (1994)



Belief and Plausibility Functions on the Space of Scalar Products and Applications

Juan J. Salamanca^(✉)

Escuela Politécnica de Ingeniería, Departamento de Estadística e I.O. y D.M.,
Universidad de Oviedo, 33071 Gijón, Spain
salamancajuan@uniovi.es

Abstract. We study the problem of a vector space over \mathbb{R} whose Euclidean scalar product is unknown. From piece of evidence given by certain experts, we are able to build suitable belief and plausibility functions on the space of scalar products. We pay special attention to study contradictions and degrees of conflict. As a possible application, for a random variable of the vector space, we are able to get the related belief and plausibility functions for its variance.

1 Introduction

There exist real situations where the problem is to determine a scalar product which a vector space is endowed with. The first example is quite evident: the calibration of several cameras (here calibration means obtaining the measure of our Euclidean space in a fixed units -as meters-) (see, for instance, [10]). The second example comes from the fact that any ellipse can be written as the ball of unitary radius for an appropriate scalar product; regard that a wide variety of statistical approaches have been considered in the literature for finding a suitable ellipse (see [5] for instance and references therein). The third example comes from investments. In this setting, some risks are modeled with several variables, where the variances are interpreted as uncertainty (see [13] for instance). Regarding such variables as (the coordinates of) a multivariate one, the scalar product of the joint sample space determines the weights of such variables (see Sect. 4). The last example is more theoretical. Recall that a copula determines the joint probability distribution of two random variables in terms of their marginals. Moreover, this copula summarizes the possible dependence between the variables. When no precise information is available, a natural approach consists on getting suitable bounds and/or considering stochastic orders (see, for instance, [8, 9] or [14] and references therein). We also refer to [2, 6] for related questions (besides [3, 11, 12], which are the main references).

For these geometric problems, we desire to get an appropriate approach from the Dempster-Shafer theory. Then, our main objective is to develop that theory in this direction.

Our starting framework here is a finite-dimensional vector space \mathbb{V} endowed with an unknown scalar product. However, we assume that there exists a set of experts which provide some partial information about the scalar product (see next section).

We need an extra object which can serve to join evidence. That object is a canonical, natural, geometrical order in our space of discernment.

By means of this order, we are able to consider a very wide class of evidence, enlarging the class of information that experts can provide. This order will split our problem into two main questions: (i) how big at most the real scalar product is? and (ii) how small at least the real scalar product is?

In this setting, we find the focal sets. The problem is its complexity. We are able to solve it in the following sense: it is possible to find unique scalar products which bound from above and from below that focal sets (see Theorem 1 and compare with [1], for instance). We interpret them as the global consensual measurements for any geometrical or statistical element.

This work is organized as follows: in Sect. 2, we set the finite-dimensional space of discernment and the main mathematical tools related to the order. Section 3 is aimed to consider all distinguished points as the Dempster-Shafer theory provides. Finally, in Sect. 4, we get an important application for the study of the variance of a random variable.

2 A Finite-Dimensional Space of Discernment Endowed with an Order

Let \mathbb{V} be an n -dimensional vector space over \mathbb{R} . Recall that an Euclidean scalar product is a non-degenerate positive-definite bilinear form $g : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$. Though we can consider the space of Euclidean scalar products, we will take instead a bigger space. We define

$$G := \{g : \text{bilinear form, } g(x, x) \geq 0, \forall x \in \mathbb{V}\}.$$

That is, we allow to have null directions. This fact is motivated mainly by the following reason: one expert may have a lack of information about measures along a direction. Therefore, no extra charge of mathematics is needed to get a suitable framework (see next section).

Clearly, G is a convex subset of the space of all bilinear forms of \mathbb{V} , $L(\mathbb{V}, \mathbb{V})$. Recall that $L(\mathbb{V}, \mathbb{V})$ has also a structure of a vector space over \mathbb{R} . Latter, we introduce a canonical order in G : we will say that g_0 is bigger than g_1 , written $g_0 \geq g_1$, if $g_0(x, x) \geq g_1(x, x)$ for any $x \in \mathbb{V}$. At the same time, we will say that g_1 is smaller than g_0 . The interpretation is clear: a measurement from g_0 provides a value as least as it provides from g_1 . Particularly, if X is a random variable of \mathbb{V} , its variance is bigger (or equal) in (\mathbb{V}, g_0) than in (\mathbb{V}, g_1) . This order will play a fundamental role in our problem.

Without a doubt, the order \geq is reflexive, antisymmetric and transitive. However, it is not complete. To illustrate it, consider $\{e_1, e_2\}$ a basis of a vector

space \mathbb{V} and take g_0 and g_1 defined by $g_0(e_1, e_1) = 1, g_1(e_1, e_1) = 2, g_0(e_1, e_2) = g_1(e_1, e_2) = 0, g_0(e_2, e_2) = 2, g_1(e_2, e_2) = 1$. Then, neither $g_0 \geq g_1$ nor $g_0 \leq g_1$.

Assume that we are given a collection $\{g_i\}_{i=1}^n, g_i \in G$. We could consider the (non-linear) functionals $\psi^+_{\{g_1, \dots, g_n\}}(x, x) := \max_{1 \leq i \leq n} g_i(x, x)$ and $\psi^-_{\{g_1, \dots, g_n\}}(x, x) := \min_{1 \leq i \leq n} g_i(x, x)$. The great advantage of consider these functionals is to get more precise values for the different statistical parameters (see Sect. 4). As disadvantages, none of them is a scalar product, and computations will be difficult. Consequently, we desire to find a scalar product bigger (or smaller) than the g_i with certain property of minimality (resp. maximality). For these purposes, we introduce the following sets

$$G^+(g_1, \dots, g_n) := \{g \in G : g \geq g_i, \forall i \in 1, \dots, n\},$$

$$G^-(g_1, \dots, g_n) := \{g \in G : g \leq g_i, \forall i \in 1, \dots, n\}.$$

That is, G^+ (resp. G^-) contains all the scalar products which are bigger (resp. smaller) than any g_i . Note that G^+ is nonempty: $\sum_i g_i \in G^+$. The set G^- is also non-empty: the null scalar product 0_g (defined by $0_g(x, x) = 0$, for any vector x) belongs to G^- . The main properties of these sets appear in the following result,

Lemma 1. *Let $\{g_i\}_{i=1}^n$ be a finite collection of elements of G . Then,*

- (i) $G^+(g_1, \dots, g_n)$ and $G^-(g_1, \dots, g_n)$ are convex sets.
- (ii) $G^+(g_1, \dots, g_n) = G$ if and only if $g_1 = \dots = g_n = 0_g$.
- (ii') $G^-(g_1, \dots, g_n) = 0_g$ if and only if there exists a basis $\{e_j\}$ of \mathbb{V} such that $\min_{1 \leq i \leq n} g_i(e_j, e_j) = 0$ for any j .
- (iii) $g \in G^+(g_1, \dots, g_n)$ if and only if $g(x, x) \geq \psi^+_{\{g_1, \dots, g_n\}}(x, x)$ for any vector x .
- (iii') $g \in G^-(g_1, \dots, g_n)$ if and only if $g(x, x) \leq \psi^-_{\{g_1, \dots, g_n\}}(x, x)$ for any vector x .
- (iv) $G^+(g_1, \dots, g_n) \subseteq G^+(g_1, \dots, g_{n-1})$; and $G^+(g_1, \dots, g_n) = G^+(g_1, \dots, g_{n-1})$ if and only if $g_n(x, x) \leq \psi^+_{\{g_1, \dots, g_{n-1}\}}(x, x)$ for any vector x . In particular, $G^+(g_1, \dots, g_n) = G^+(g_1, \dots, g_n, 0_g) = G^+(g_1, \dots, g_n, g_n)$.
- (iv') $G^-(g_1, \dots, g_n) \subseteq G^-(g_1, \dots, g_{n-1})$; and $G^-(g_1, \dots, g_n) = G^-(g_1, \dots, g_{n-1})$ if and only if $g_n(x, x) \geq \psi^-_{\{g_1, \dots, g_{n-1}\}}(x, x)$ for any vector x . In particular, $G^-(g_1, \dots, g_n) = G^-(g_1, \dots, g_n, \sum_{i=1}^n g_i) = G^-(g_1, \dots, g_n, g_n)$.
- (v) $G^+(g_1, \dots, g_n) = G^+(g_1, \dots, g_l) \cap G^+(g_{l+1}, \dots, g_n), 1 \leq l < n$.
- (v') $G^-(g_1, \dots, g_n) = G^-(g_1, \dots, g_l) \cap G^-(g_{l+1}, \dots, g_n), 1 \leq l < n$.
- (vi) For any permutation σ of $\{1, \dots, n\}$, $G^+(g_1, \dots, g_n) = G^+(g_{\sigma(1)}, \dots, g_{\sigma(n)})$ (idem for G^-).
- (vii) $G^+(g_1, \dots, g_n) \cap G^-(g_1, \dots, g_n) = \emptyset$ unless $g_1 = \dots = g_n$, in which case $G^+(g_1, \dots, g_n) \cap G^-(g_1, \dots, g_n) = \{g_1\}$.

We will say that g is an optimum scalar product in G^+ if for any scalar product g' such that $g' \leq g, g' \neq g$, it holds $g' \notin G^+$. Analogously, we will say that g is an optimum operator in G^- if for any scalar product g' such that $g' \geq g, g' \neq g$, it holds: $g' \notin G^-$. An optimum scalar product of G^+ will provide a minimum of the upper bounds of a dispersive parameter; equivalently,

an optimum scalar product of G^- will provide a maximum of lowers bounds of a dispersive parameter.

It is clear that uniqueness is of a great interest. The following result guarantees this fact,

Theorem 1. *There exists a unique optimum scalar product in $G^+(g_1, \dots, g_n)$; there exists a unique optimum scalar product in $G^-(g_1, \dots, g_n)$.*

Proof (Sketch for first statement). Assume that there are two optimum scalar products, g_a, g_b in $G^+(g_1, \dots, g_n)$. By assumptions, neither $g \leq g'$ nor $g \geq g'$. Endow \mathbb{V} with g_a as scalar product. Then there exists a g_a -orthonormal basis where the Gram's matrix of g_b is diagonal (g_b is the identity matrix). Comparing eigenvalues one arrives to a contradiction due to the fact that g_a and g_b are optimum scalar products.

Then, given g_1, \dots, g_n , we denote by g^+ and g^- the optimum scalar product in G^+ and G^- , respectively.

3 Setting Mass, Belief and Plausibility Functions

We need to establish what kind of information an expert shall provide. It seems quite restrictive assuming that a expert gives a set of G where the real scalar product can be. Note that the power set of the set of scalar products is too big to be considered. Moreover, we would find serious problems when two different experts provide disjointed sets. To avoid these problems, we split our problem into two propositions: **(i)** how big at most the scalar product is? **(ii)** how small at least the scalar product is? Then, we structure the original problem into the knowledge of *(i)* and *(ii)*. Observe that this framework is appropriate to simplify complex information that an expert can give. In particular, reasoning as in the proof of Theorem 1, it can be proved: if an expert gives a set of G as possible scalar products, a maximal (and a minimal) scalar product can always be found (that is, we can simplify that information into two scalar products; hence, we find no loss of generality but precision in this framework).

Definition. A mass function for *(i)* (or for *(ii)*) is a finite collection of elements of G endowed with a mass; that is, $\{(g_1, m_1), \dots, (g_n, m_n)\}$, $g_i \in G$, $m_i \in (0, 1]$, with $\sum_{i=1}^n m_i = 1$. We will denote a mass function for *(i)* by \mathcal{M}^+ , for *(ii)* by \mathcal{M}^- , and jointly by $(\mathcal{M}^+, \mathcal{M}^-)$.

The interpretation is clear: for *(i)*, an expert shows a mass m_i to be the real scalar product smaller than g_i . A similar interpretation is provided for *(ii)*. We will do computations for how big (resp. small) at most (resp. at least) is a dispersive parameter taking into account uniquely the mass function(s) related to *(i)* (resp. *(ii)*). Equivalently, we can present our conclusions in terms of intervals.

The set $G^+(g_i)$, $g_i \in \mathcal{M}^+$ can be thought as a focal set¹ -as named in Dempster-Shafer theory. On the other hand, if the scalar products of \mathcal{M}^+ (resp. \mathcal{M}^-) are totally ordered (i.e., we can write in any case $g_1 \leq g_2 \leq \dots g_n$), we would say that \mathcal{M}^+ (resp. \mathcal{M}^-) is consonant. This property simplify our study cases.

Modeling Exact Information. An expert asserts that the real scalar product is g if and only if her/his mass functions are $\mathcal{M}^+ = \{(g, 1)\}$ and $\mathcal{M}^- = \{(g, 1)\}$.

Modeling Lack of Information. If an expert has no information about (i), he/she can provide $\mathcal{M}^+ = \{(0_g, 1)\}$. Similarly, if an expert has no information about (ii), he/she can provide the algebraic sum of the scalar products of his/her partners; that is, $\mathcal{M}^+ = \{(\sum_n g_n, 1)\}$, where any g_n belongs to a mass function for (ii) of a partner. Note that Lemma 1.(iv) – (iv') assures that this lack of information does not modify the computation of a scalar product which all experts can agree on.

For (i), if an expert only knows how big the scalar product is when restricted onto a closed subspace L , let us say $g|_L$, via the inclusion map of L into \mathbb{V} , he/she can establish a scalar product of \mathbb{V} : $g|_L + \sum_n g_n$, where any g_n belongs to a mass function for (i) of a partner. For (ii), a similar argument can be applied. Again, the reason for these considerations is to provide suitable scalar products on \mathbb{V} which do not modify the information where there is a lack. Note that this procedure can be applied even in the case of partial lack of information. That is, if a lack appears only on an assignment with mass m , that information should be considered as a scalar product making use of the previous procedure, while the other assignments remain.

Contradictions. An expert can provide contradictory information. Let us illustrate it. Consider that $\{e_1\}$ is a basis of \mathbb{V} . An expert could establish $g_0 : g_0(e_1, e_1) = 1$ with mass 1 for (i), and $g_1 : g_1(e_1, e_1) = 2$ for (ii). That is, when asked how big at most the real scalar product is, he/she says a smaller quantity than when asked how small it is. Observe that this kind of contradiction cannot appear if we only take a framework for (i) or (ii) disjointedly.

To clarify the different kind of information that an expert can provide, we propose the following classification: a mass function $(\mathcal{M}^+, \mathcal{M}^-)$ is said to be *clear* if the optimum for G^+ is smaller than the optimum for G^- ; is said to be *contradictory* when the optimum for G^+ is greater than the optimum for G^- ; otherwise, is said to be *unclear*. The terminology is obvious: when considering a random variable of \mathbb{V} , if the pair of mass functions is clear, we will obtain a maximum value for a dispersive parameter which is bigger than its minimum; if it is contradictory, it happens the opposite; and if it is unclear, any case could occur.

Projecting. Let $\{e_1, \dots, e_m\}$ be a basis of a vector space \mathbb{V} . Eventually, we desire to study the projection onto a closed subspace L , let us say the linear subspace generated by $\{e_1, \dots, e_l\}$, $l < m$. If the original information is given

¹ Analogously for $G^-(g_j)$.

by a scalar product g of \mathbb{V} , we can consider its restriction onto L , $g|_L$. Then, we can operate intrinsically on L . Observe that projecting onto a closed subspace represents a marginalization.

Finally, we are able to introduce a suitable Belief and Plausibility function,

Belief Function. Given a mass function for (i) , $\mathcal{M}^+ = \{(g_1, m_1), \dots, (g_n, m_n)\}$, the total degree of evidence for $A \equiv$ ‘the real scalar product is g as maximum’ is the total amount of evidence which implies A ,

$$\text{Bel}^+(A) := \sum_{i:g_i \leq g} m_i.$$

Analogously, given a mass function for (ii) , \mathcal{M}^- , we define $\text{Bel}^-(B)$, $B \equiv$ ‘the real scalar product is g as minimum’.

We have: $\text{Bel}^+(0_g) = 0$ (except in the pathological case $g_i = 0_g$ for any i); $\text{Bel}^-(0_g) = 1$; $\text{Bel}^+(g^+) = 1$, $\text{Bel}^-(g^-) = 1$.

Plausibility Function. Given a mass function for (i) , \mathcal{M}^+ , the total degree of evidence which does not contradict $A \equiv$ ‘the real scalar product g as maximum’ is

$$\text{Pl}^+(g) := \sum_{g_i: g_i \not\leq g} m_i.$$

Analogously we define $\text{Pl}^-(g)$ for (ii) .

We have: $\text{Pl}^+(0_g) = 0$ (unless in the pathological case $g_i = 0_g$ for any i); $\text{Pl}^-(0_g) = 1$; $\text{Pl}^+(g^+) = 1$, $\text{Pl}^-(g^-) = 1$; $\text{Bel}^+(g) \leq \text{Pl}^+(g)$; $\text{Bel}^-(g) \leq \text{Pl}^-(g)$.

The interpretations are quite straightforward. For instance, $\text{Pl}^-(0_g) = 1$ is equivalent to: *the real scalar product g satisfies $g(x, x) \geq 0$ for any vector x with degree of belief 1.*

Observe that the belief and plausibility functions defined above can be extended to functions taking values on the Borel σ -algebra of G (recall that G is finite-dimensional). However, we prefer to pay attention to it concisely in another future work.

To close comments on these definitions, note that the negation of the proposition $A \equiv$ ‘the real scalar product is big at most g ’ does not lie uniquely in \mathcal{M}^+ , but also in \mathcal{M}^- . For this reason, we refrain from mixing Bel^+ , Bel^- , Pl^+ and Pl^- -like the classical formula $\text{Pl}(A) = 1 - \text{Bel}(\bar{A})$.

Combination Rule (see [7, 16], for instance). Now, we assume there exist r experts with mass functions $(\mathcal{M}_i^+, \mathcal{M}_i^-)$, $1 \leq i \leq r$. Initially, we put a weight for each expert.

Observe that, although any expert is clear in his/her mass functions, it may exist a degree of conflict when all the information is mixed. We will say that there exists no degree of conflict at all when $g^+ \geq g^-$. This fact implies always, directly, the existence of an interval I of \mathbb{R} with the following property: all experts coincide in the fact that the real value of a dispersive parameter belongs to I , whatever the random variable is.

Now, let us quantify the degree of conflict for two experts A and B with equal weight,

$$2K = \sum_{g_i \in M_a^+; g_j \in M_b^-; g_i \leq g_j} m_i^a(g_i) m_j^a(g_j) + \sum_{g_i \in M_a^-; g_j \in M_b^+; g_i \geq g_j} m_i^a(g_i) m_j^a(g_j).$$

That is, it represents the total amount of belief that must be rejected if we desire an agreement among A and B . The previous formula can be naturally extended to the case of several experts endowed with certain weights.

Assuming that the weights are fixed, from the information provided by r experts with mass functions $\{(\mathcal{M}_i^+, \mathcal{M}_i^-)\}_{i=1}^r$, we can build a global mass function $(\mathcal{M}_T^+, \mathcal{M}_T^-)$. Summing up, $g \in \mathcal{M}_T^+$ with mass m if there exist $l \geq 1$ experts with weights α_i such that each of them assigns $m_i (> 0)$ to g for (i) and $m = \sum_{i=1}^l \alpha_i m_i$. Analogously for \mathcal{M}_T^- . At the end, we get an unique global mass function, reducing the problem of several experts to only one.

4 An Application to Study Variance

Let X be a random variable of a vector space \mathbb{V} . Then, its expected value -when it exists-, $\mathbb{E}[X]$ does not depend on the scalar product of \mathbb{V} . However, this fact is not longer true for any dispersive parameter. In this setting, let \mathbb{V}^* be the dual vector space of \mathbb{V} . Recall the symmetric covariance-variance bilinear form: $C[X] : \mathbb{V}^* \times \mathbb{V}^* \rightarrow \mathbb{R}$, $C[X](x^*, y^*) = \mathbb{E}[(x^*(X)) \cdot (y^*(X))] - (\mathbb{E}[x^*(X)]) \cdot (\mathbb{E}[y^*(X)])$. Only by means of a scalar product g , we can obtain its metric trace², that is, its variance. The variance of X is then $\text{Var}[X] = \text{tr}_g C[X]$. In a practical context, this computation can be done in terms of matrices. Although we restrict ourselves to the variance, this procedure can be applied to other parameters.

Following literature (see, for instance, [4, 15]), we will understand as solution for our original problem (i) , a Belief function which assigns, to a non-negative real x , the total amount of evidence for $A \equiv$ ‘the variance is at most x ’. That is, given a finite set of experts with global mass functions \mathcal{M}_T^+ ,

$$\mathcal{BEL}(\text{Var}[X] \leq x) := \sum_i m_i(g_i) \quad |i : g_i \in \mathcal{M}_T^+, \text{tr}_{g_i} C[X] \leq x.$$

With obvious changes, we can write $\mathcal{BEL}^+(\text{Var}[X] < x)$ and $\mathcal{BEL}(\text{Var}[X] \geq x)$. The question now is how to write $\mathcal{BEL}(x_0 \leq \text{Var}[X] \leq x_1)$ assuming now that we have $(\mathcal{M}_T^-, \mathcal{M}_T^+)$. By an equity axiom, we get,

$$\mathcal{BEL}(x_0 \leq \text{Var}[X] \leq x_1) = 0.5[\mathcal{BEL}(\text{Var}[X] \leq x_1) + \mathcal{BEL}(\text{Var}[X] \geq x_0)].$$

For the case of plausibility, longer formulas can be written.

² A non-negative scalar which is invariant under change of basis.

Making use of Lemma 1: for a random variable X of a vector space \mathbb{V} and a finite set of experts,

$$\mathcal{BEL}(x_0 \leq \text{Var}[X] \leq x_1) \leq \mathcal{BEL}(x_2 \leq \text{Var}[X] \leq x_3)$$

where $x_2 \leq x_0 \leq x_1 \leq x_3$ and $x_1 \leq x_3$.

To finish this work, let us observe that, when all the experts accord g to be the real scalar product, then $\mathcal{BEL}(x_0 \leq \text{Var}[X] \leq x_1) = 1$ if $\text{tr}_g C[X] \in [x_0, x_1]$ and 0 otherwise.

References

1. Bauer, M.: Approximation algorithms and decision making in the Dempster-Shafer theory of evidence - an empirical study. *Int. J. Approximate Reasoning* **17**, 217–232 (1997)
2. Cuzzolin, F.: Two new Bayesian approximations of belief functions based on convex geometry. *IEEE Trans. SMC Part B*. **37**, 993–1008 (2007)
3. Dempster, A.P.: Upper and lower probabilities generated by a random closed interval. *Ann. Math. Stat.* **3**, 957–966 (1968)
4. Dymova, L., Sevastjanov, P.: The operations on interval-valued intuitionistic fuzzy values in the framework of Dempster-Shafer theory. *Inf. Sci.* **360**, 256–272 (2016)
5. Friendly, M., Monette, G., Fox, J.: Elliptical insights: understanding statistical methods through elliptical Geometry. *Stat. Sci.* **28**, 1–39 (2013)
6. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: comprehensive survey and generalizations. *Int. J. Approximate Reasoning*. **5**, 118–145 (2012)
7. Martin, A.: Reliability and combination rule in the theory of belief functions. In: 12th International Conference on Information Fusion, Seattle, WA, USA (2009)
8. Montes, I., Miranda, E., Pelessoni, R., Vicig, P.: Sklar’s theorem in an imprecise setting. *Fuzzy Sets Syst.* **278C**, 48–66 (2015)
9. Nelsen, R.B., Quesada, J.J., Rodríguez-Lallena, J.A., Úbeda, M.: Best-possible bounds on sets of bivariate distribution functions. *J. Multivariate Anal.* **90**, 348–358 (2004)
10. Tsai, R.Y., Lenz, R.K.: A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE Trans. Robot. Autom.* **5**, 345–358 (1989)
11. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
12. Shafer, G.: Allocations of probability. *Ann. Probab.* **5**, 827–839 (1979)
13. Shaked, M., Shanthikumar, G.: *Stochastic Orders*. Springer Series in Statistics. Springer, New York (2007)
14. Yager, R.R.: Joint cumulative distribution functions for Dempster-Shafer belief structures using copulas. *Fuzzy Optimization Decis. Making* **4**, 393–414 (2013)
15. Wu, J., Zhao, T., Li, S., Own, C.M.: Belief interval of Dempster-Shafer theory for line-of-sight identification in indoor positioning applications. *Sensors (Basel)* **17**, 1–18 (2017)
16. Zadeh, L.A.: A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI Mag.* **7**, 85–90 (1986)



E2CM: An Evolutionary Version of Evidential C-Means Clustering Algorithm

Zhi-gang Su¹(✉), Hong-yu Zhou¹, Pei-hong Wang¹, Gang Zhao¹,
and Ming Zhao²

¹ School of Energy and Environment,
Southeast University, Nanjing, Jiangsu, China
zhigangsu@seu.edu.cn

² Research Institute of Yunnan Power Grid Co. Ltd.,
Kunming, Yunnan, China

Abstract. This paper aims to propose an Evolutionary version of Evidential C-Mean (E2CM) clustering method based on a Variable string length Artificial Bee Colony (VABC) algorithm. In the E2CM, the centers of clusters are encoded in form of a population of strings with variable length to search optimal number of clusters as well as locations of centers based on the VABC, by minimizing objective function *non-specificity*, in which the assignment of objects to the population of cluster centers are performed by the ECM. One significant merit of the E2CM is that it can automatically create a credal partition without requiring the number of clusters as a priority. A numerical example is used to intuitively verify our conclusions.

Keywords: Dempster-Shafer theory · Belief functions
Evidential clustering · Swarm intelligent algorithm

1 Introduction

Evidential clustering describes uncertainty in the membership of objects to clusters using a Dempster-Shafer mass functions [11]. Roughly speaking, a mass function can be seen as a collection of sets with corresponding masses. A collection of such mass functions for n objects is called a credal partition. In recent decade, evidential clustering shows its powerful ability to reveal data structure and attracts more and more attentions in artificial intelligent societies [2].

In [5], an evidential clustering algorithm called EVCLUS was first proposed to deal with partition of relational data. After this, Masson and Denoeux proposed an evidential version of Fuzzy C-Mean clustering algorithm, called Evidential

This work is supported in part by the National Natural Science Foundation of China (51676034), and by the Key Project of Yunnan Power Grid Co. Ltd. (YNYJ2016043).

C-Means (ECM), for attribute data in [9] and relational data in [10]. In practice, there usually exist must-link and/or cannot-link pairwise constraints among objects. To solve this problem, a constrained version of ECM (CECM) was proposed consequently in [1]. Notice that in either the ECM or CECM, the strategy to derive the barycenters sometimes leads to uninformative composite clusters. To overcome this drawback, Liu et al. proposed respectively a Belief and Credal C-Means clustering algorithm [7, 8] by redefining the distance between an object and prototype of a meta-cluster.

However, all the aforementioned evidential clustering requires the number of clusters as a priority. Furthermore, the locations of centers are found by minimizing a cost function based on an alternate optimization algorithm. In a more recent work [3], Denoeux and Kanjanatarakul proposed a new evidential clustering method, called EK-NNclus, based on the Evidential k -nearest neighbor (EK-NN) classification rule [4]. The EK-NNclus can automatically determine the number of clusters without requiring any priori. Nevertheless, in contrast to other evidential clusterings, EK-NNclus creates a credal partition consisting of only singletons and the whole frame of discernment. Therefore, it is interesting to propose a new evidential clustering that can create a credal partition with more or full focal sets and without requiring the number of clusters as priori.

Motivating from above considerations, this paper aims to propose an Evolutionary version of ECM algorithm (E2CM) based on an evolutionary algorithm rather than an alternate optimization algorithm, i.e., the Variable string length Artificial Bee Colony (VABC) algorithm proposed in our previous work [12]. It will be interestingly seen that the proposed E2CM algorithm can automatically create a credal partition consisting of full focal sets without requiring as prior the number of clusters.

The rest of the paper is organized as follows. In the Sect. 2, the ECM algorithm is briefly recalled. Section 3 introduces the E2CM algorithm. The performance of E2CM is briefly validated by an numerical example in Sect. 4. The last section concludes the paper.

2 Background: Evidential C-Mean Clustering

We suggest that readers are familiar with some basic concepts of theory of belief functions [11] by considering the length of paper.

For a given dataset $X = \{x_1, x_2, \dots, x_n\}' \in R^{n \times p}$ and a class of clusters $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$, deriving a credal partition $\mathcal{M} = \{m_1, m_2, \dots, m_n\}' \in R^{n \times 2^c}$ from X implies determining the quantities $m_{i_j}^\Omega = m_i^\Omega(A_j), A_j \subseteq \Omega$ for each object x_i in such a way that $m_{i_j}^\Omega$ is high (respectively, low) when the distance d_{i_j} between x_i and the focal set A_j is small (respectively, large). Suppose that each cluster ω_k is represented by prototype $v_k \in R^p$. The barycenter for the nonempty composite cluster (i.e., focal set) A_j can be defined as

$$\bar{v}_j = |A_j|^{-1} \sum_{k=1}^c s_{kj} v_k, \quad (1)$$

where $s_{kj} = 1$ if $\omega_k \in A_j$, $s_{kj} = 0$ otherwise; $|\cdot|$ denotes the cardinality of a set or a string/sequence. Hence, the Euclidean distance d_{ij} can be calculated by $d_{ij}^2 = \|x_i - \bar{v}_j\|^2$. ECM proposes to create the credal partition \mathcal{M} and the cluster center matrix $\mathcal{V} = \{v_1, v_2, \dots, v_c\}' \in R^{c \times p}$ by minimizing

$$J_{ECM}(\mathcal{M}, \mathcal{V}) = \sum_{i=1}^n \sum_{j:\emptyset \neq A_j \subseteq \Omega} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^n \delta^2 m_{i\emptyset}, \tag{2}$$

$$s.t. \quad \sum_{j:\emptyset \neq A_j \subseteq \Omega} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, 2, \dots, n, \tag{3}$$

where $m_{i\emptyset}$ denotes the mass allocated to the empty set for object x_i ; the weighting coefficient α aims at penalizing focal sets with high cardinality; the exponent β controls the fuzziness of the partition, and distance δ^2 controls the number of objects that assumed to be outliers.

With fixed \mathcal{V} , the masses in \mathcal{M} can be obtained for $i = 1, 2, \dots, n, j : \emptyset \neq A_j \subseteq \Omega$ as

$$\begin{cases} m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}}, \\ m_{i\emptyset} = 1 - \sum_{A_k \neq \emptyset} m_{ik}. \end{cases} \tag{4}$$

Once the credal partition is fixed, the cluster centers matrix \mathcal{V} can be updated by solving the following equality

$$H\mathcal{V} = B, \tag{5}$$

with two matrices $B_{c \times p}$ and $H_{c \times c}$ defined by $B_{lq} = \sum_{i=1}^n x_{iq} \sum_{\omega_l \in A_j} |A_j|^{\alpha-1} m_{ij}^\beta$, $H_{lk} = \sum_{i=1}^n \sum_{\omega_l, \omega_k \subseteq A_j} |A_j|^{\alpha-2} m_{ij}^\beta, l, k = 1, 2, \dots, c, q = 1, 2, \dots, p$.

Finally, a validity index is needed to determine a suitable number of clusters. The *non-specificity* is such a popular choice defined by

$$N(c, \mathcal{M}) = \frac{1}{n \log_2(c)} \times \sum_{i=1}^n \left[\sum_{j:\emptyset \neq A_j \subseteq \Omega} m_i(A_j) \log_2 |A_j| + m_{i\emptyset} \log_2(c) \right]. \tag{6}$$

3 Evolutionary Evidential C-Means Clustering

3.1 Motivations and the Basic Idea

The VABC is a generalized variant of the Artificial Bee Colony algorithm (ABC) [6] with variable length of genotypes. In the VABC, each solution to the problem under consideration is called a food source and represented by a real-valued string with variable length. This variable length of representation allows to find number of and locations of genotypes simultaneously. As the ABC, VABC algorithm classifies the foraging artificial bees into three groups, namely, employed bees, onlookers and scouts. A bee that is currently exploiting a food source is called an employed bee. A bee waiting in the hive for making decision to choose a food source is named as an onlooker. A bee carrying out a random search for a new

food source is called a scout. The fitness of a solution corresponds to the nectar amount of this food source, and therefore the highest fitness induces optimal solution(s). The VABC is an iterative process, starting with a population of randomly generated solutions or food sources, and repeats following four steps:

1. Send the employed bees onto the food sources and then measure their nectar amounts (i.e., the fitness).
2. Select the food sources by the onlookers after sharing the information of employed bees and determine the nectar amount of the food sources.
3. Determine the scout bees and send them onto the possible food sources.
4. Perform mutation operations on food sources to guarantee convergence.

With above interpretations in mind, the ECM can be viewed as an optimization problem aiming to find a suitable number and locations of genotypes as well as assignments of data objects to these genotypes, by minimizing non-specificity instead (rather than (2)). This motivates us to solve such optimization problem via the VABC algorithm, and thus to propose an evolutionary version of ECM.

The basic idea of E2CM is interpreted as follows. A population of strings (i.e., food sources) with variable length are randomly initialized to represent possible number of clusters and locations of centers among data objects. By taking a p -dimensional clustering task as an example, a string $S_i = \{v_1^{i'}, v_2^{i'}, \dots, v_c^{i'}\} \in R^{1 \times cp}$ indicates number c of clusters locating respectively at centers $v_k^i \in R^p, k = 1, 2, \dots, c$. For each string in the population, a credal partition can then be created by the ECM algorithm (i.e., according to the (4)). Therefore, a set of credal partitions can be derived. Each credal partition is evaluated according to the non-specificity (6). The strings minimizing the non-specificity will be considered as the optimal solutions in the current iteration, and all strings will forage and mutate to search the better objective in the coming iterations. The E2CM will be terminated when meeting some terminations, and outputs the optimal string(s) including optimal number of clusters and locations of centers as well as corresponding credal partition.

3.2 Realization of the E2CM

Given an upper bound for the number of clusters, i.e., C_{max} , a population of N strings/food sources, representing the possible number of clusters and locations of centers, are initialized according to

$$S_{ij} = \underline{S}_j + rand[0, 1] \times (\bar{S}_j - \underline{S}_j), i = 1, 2, \dots, N, j \in \{1, 2, \dots, L_i\}, \quad (7)$$

where S_{ij} is the j th dimension of string S_i , and \underline{S}_j and \bar{S}_j are respectively the lower and upper bounds of the j th attribute of S_i ; $rand[0, 1]$ denotes a uniform random number in the range $[0, 1]$; L_i , the length of string S_i , is defined as $L_i := p \times [\text{round}(rand[0, 1] \times (C_{max} - 2)) + 2]$ with a function $\text{round}(\cdot)$ rounding a number to its nearest integer.

Note that each food source S_i can be decoded into number L_i/p of cluster centers $\mathcal{V}_i = \{v_1^i, v_2^i, \dots, v_{L_i/p}^i\}'$. Therefore, the j th dimension of S_i is exactly

the r th dimensional space of objects x_i , where r is the remainder of division $\frac{j}{p}$ if remainder exists, $r := p$ otherwise. We have $\underline{S}_j := \min_{1 \leq i \leq n} \{x_{ij}\}$ and $\bar{S}_j := \max_{1 \leq i \leq n} \{x_{ij}\}$. Furthermore, for a given \mathcal{V}_i , a credal partition \mathcal{M}_i can be derived according to (4). Then, the nectar amount of food source S_i can be evaluated by

$$fit_i = 1/Objective(S_i), \tag{8}$$

where $Objective(S_i) := N(L_i/p, \mathcal{M}_i)$ is the objective function of E2CM.

On the employed bee stage, the employed bees foraging on food sources S_i are first to search candidate new food source positions from precious ones in order to achieve more nectar amounts according to

$$S_{ij}^{new} = S_{ij} + \emptyset_{ij}(S_{ij} - S_{kq}), \tag{9}$$

where $k \in \{1, 2, \dots, N\}$ and $k \neq i$; \emptyset_{ij} is a uniformly random number in $[-1, 1]$; j and q denote resp. the j th and q th dimension of S_i and S_k . If $L_k \geq L_i$, integer j is randomly generated in $[1, L_i]$ and then is assigned to q , i.e., $q = j$; otherwise, it can first randomly generate the integer q in $[1, L_q]$ and then set $j = q$.

On the onlooker stage, by sharing information (e.g., fitness of food sources) with employed bees, onlookers then begin to explore new food sources according to (9) in the neighborhood of food sources selected. Whether a food source can be selected or not depends upon the following probability

$$p_i = fit_i / \sum_{i=1}^n fit_i. \tag{10}$$

Obviously, the higher fitness fit_i is, the more probability the food source S_i can be selected by an onlooker.

On the scout stage, the food source whose nectar is abandoned by the employed and/or onlooker bees is replaced with a new food source by the scouts. This is simulated by randomly producing a food source position according to (7) and replacing it with the abandoned one. In addition, if a food source position cannot be improved further through a predetermined number of cycles denoted by *limit*, then that food source is assumed to be abandoned.

Finally, to guarantee convergency of the length of strings (i.e., food sources), some mutation operations are defined as follows:

- If the length of a string is equal to that of the string holding the best fitness, not any mutation operation will be required.
- If the length of a string is longer than that of the string holding the best fitness, a sequence of successive p dimensions are randomly selected and removed from the string; otherwise, one object in X is randomly selected and added at the end of the string.

The probability to perform mutations on a string is defined as

$$p_i^m = \begin{cases} k_1 \frac{fit_{max} - fit_i}{fit_{max} - fit_{avg}}, & \text{if } fit_i > fit_{avg}, \\ k_2, & \text{otherwise,} \end{cases} \tag{11}$$

where k_1 and k_2 are positive constants and both are set to 0.5; fit_{max} and fit_{avg} are respectively the maximum and average fitness values of the population of all food sources. Note that, the higher p_i^m is (e.g., than a given threshold P_0), more opportunity the mutation operation will be performed on S_i .

According to above interpretations, the E2CM is summarized as follows.

Algorithm 1. E2CM clustering algorithm

Input: $N, C_{max}, P_0, \alpha, \beta, \delta, limit$, termination threshold $iter_{max}$, and data objects $\{x_i \in R^p, i = 1, 2, \dots, n\}$.

Output: centers in form of S^{best} and associated credal partition \mathcal{M} .

Encode a population of possible cluster centers S_i according to (7);

$iterations = 0, limit_i = 0$;

while $iterations < iter_{max}$ **do**

for $i=1$ to N **do** % **Employed bee stage**

 Decode food sources S_i and calculate \mathcal{M}_i according to (4);

 Evaluate fitness fit_i for S_i according to (8);

 Update food sources S_i by (9) and reevaluate their fit_i^{new} ;

If $fit_i^{new} > fit_i$, **then** $limit_i \leftarrow 0$, **else** $limit_i \leftarrow limit_i + 1$;

for $i=1$ to N **do** % **Onlooker bee stage**

 Calculate probabilities p_i for the updated S_i using (10);

if $rand[0, 1] < p_i$ **then**

 Explore new ones from S_i by (9) and reevaluate

fit_i^{new} ;

If $fit_i^{new} > fit_i$, **then** $limit_i \leftarrow 0$, **else** $limit_i \leftarrow limit_i + 1$;

for $i=1$ to N **do** % **Scout bee Stage**

if $limit_i > limit$ **then**

 Initialize S_i by (7) and reset $limit_i = 0$;

for $i=1$ to N **do** % **Mutation stage**

 Calculate probabilities p_i^m for S_i by (11);

if $p_i^m > P_0$ **then**

 Perform mutation operations on S_i ;

 Record the best string/solution as S^{best} achieving the highest fitness;

$iterations = iterations + 1$;

Using notations in Algorithm 1, the time complexity of E2CM is analyzed as follows. At each iteration, the time complexity on the employed bee stage is smaller than $O(2^{C_{max}}N + N \log N)$, and is about $O(N \log N)$ on the onlooker stage the scout stage and mutation operations consumes computation $O(NC_{max})$ and $O(N)$, respectively. Therefore, the total time complexity is about $O(iter_{max} \cdot N \cdot (2^{C_{max}} + \log N))$.

4 An Numerical Example

In this section, just an numerical example is used to intuitively verify the performance of E2CM due to the restriction on the length of paper.

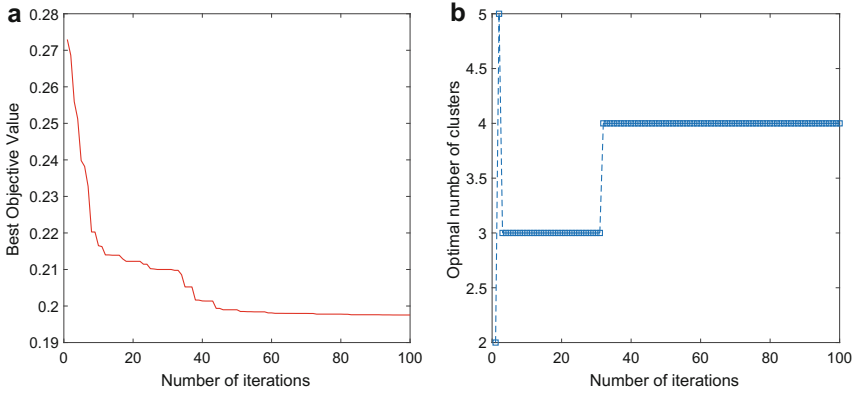


Fig. 1. The objective function (a) and number of clusters (b) corresponding to the best string(s)/solution(s) in each iteration via the E2CM for the four-class dataset

The numerical example considers the *four-class* dataset [9], consisting of equal size 100 data objects in each class. Given $iter_{max} = 100, N = 100, C_{max} = 6, P_0 = 0.4, limit = 50, \alpha = 2, \beta = 2$ and $\delta^2 = 20$. Some experimental results in one study case are presented in Figs. 1 and 2.

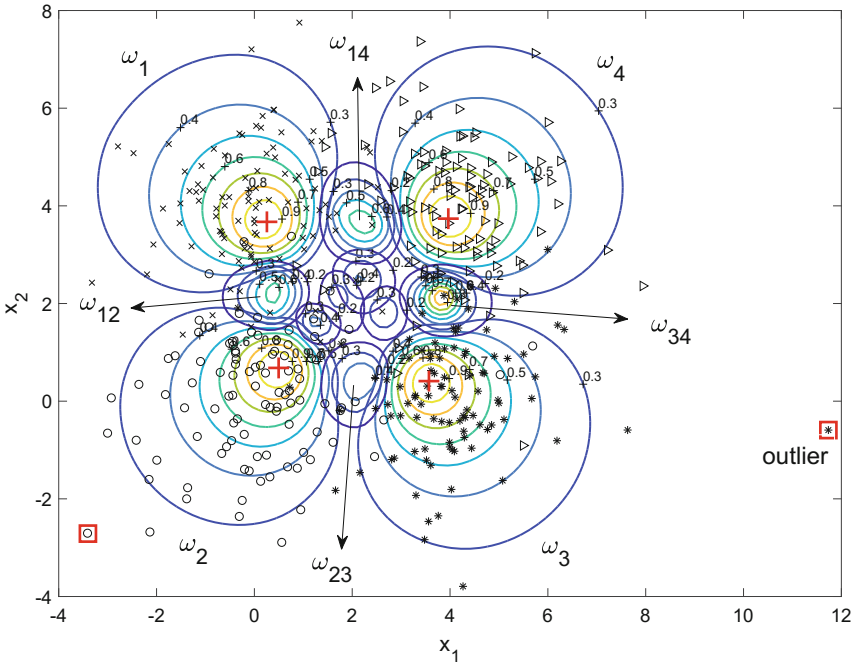


Fig. 2. Contours of credal partition for the four-class dataset via the E2CM where symbol “+” indicates locations of cluster centers and ω_{jk} means $\{\omega_j, \omega_k\}$

From Fig. 1a, we can see that the E2CM converges to the minimal non-specificity, and the number of clusters finally evolves to the optimal one, i.e., the four, as shown in Fig. 1b. This verifies that the optimal number of clusters as well as locations of centers can be found by the E2CM without knowing any priori on the cluster centers. Furthermore, E2CM reveals the data structure in a more meaningful way than hard and fuzzy partitions, as shown in Fig. 2

5 Conclusion

This paper proposes an Evolutionary version of Evidential C-Mean (E2CM) based on a Variable string length Artificial Bee Colony (VABC) algorithm. The E2CM algorithm can simultaneously find optimal number of clusters as well as locations of centers without requiring the number of clusters as priori. Furthermore, as the ECM, the E2CM can also derive a credal partition with ability to reveal data structure in a more meaningful way than classic partitional clusterings. A numerical example is used to show the performance of E2CM.

Some practical issues as well as more simulations did not be presented and will be discussed in near future. Furthermore, there are some further works. The first one is to extend the E2CM to deal with instance constraints. The second one is to consider more than one objective functions by the E2CM so as to achieve much more appropriate performance.

References

1. Antoine, V., Quost, B., Masson, M.-H., Denoeux, T.: CECM: constrained evidential C-Means algorithm. *Comput. Stat. Data Anal.* **56**(4), 894–914 (2012)
2. Denoeux, T., Kanjanatarakul, O.: Evidential clustering: a review. In: 5th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM), Da Nang, 2 December 2016, pp. 24–35 (2016)
3. Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S.: EK-NNclus: a clustering procedure based on the evidential K-nearest neighbor rule. *Knowl. Based Syst.* **88**, 57–69 (2015)
4. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (1995)
5. Denoeux, T., Masson, M.-H.: EVCLUS: evidential clustering of proximity data. *IEEE Trans. Syst. Man Cybern. Part B* **34**, 95–109 (2004)
6. Karaboga, D., Basturk, B.: On the performance of Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **8**(1), 687–697 (2008)
7. Liu, Z.-G., Dezert, J., Mercier, G., Pan, Q.: Belief C-Means: an extension of fuzzy C-Means algorithm in belief functions framework. *Pattern Recogn. Lett.* **33**(3), 291–300 (2012)
8. Liu, Z.-G., Pan, Q., Dezert, J., Mercier, G.: Credal C-Means clustering method based on belief functions. *Knowl. Based Syst.* **74**(1), 119–132 (2015)
9. Masson, M.-H., Denoeux, T.: ECM: an evidential version of the fuzzy C-Means algorithm. *Pattern Recogn.* **41**(4), 1384–1397 (2008)
10. Masson, M.-H., Denoeux, T.: RECM: relational evidential C-Means algorithm. *Pattern Recogn. Lett.* **30**(11), 1015–1026 (2009)

11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Su, Z.-G., Wang, P.-H., Shen, J.: Automatic fuzzy partition approach using Variable string length Artificial Bee Colony (VABC) algorithm. Appl. Soft Comput. **12**(11), 3421–3441 (2012)



Contrasting Two Laws of Large Numbers from Possibility Theory and Imprecise Probability

Pedro Terán^(✉) and Elisa Pis Vigil

Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Oviedo, Spain
teranpedro@uniovi.es, epis1@alumno.uned.es

Abstract. The law of large numbers for coherent lower previsions (specifically, Choquet integrals against belief measures) can be applied to possibility measures, yielding that sample averages are asymptotically confined in a compact interval. This interval differs from the one appearing in the law of large numbers from possibility theory. In order to understand this phenomenon, we undertake an in-depth study of the compatibility of the assumptions in those results. It turns out that, although there is no incompatibility between their conclusions, their assumptions can only be simultaneously satisfied if the possibility distributions of the variables are 0–1 valued.

1 The Problem

This contribution is part of a systematic analysis of the relationships between the laws of large numbers in different uncertainty frameworks (plausibility/belief measures, upper/lower probabilities, upper/lower previsions, sublinear expectations) in the particular case that they are applied to possibility measures. The main part of that analysis is [13].

In [11, Theorem 2.6], the first author obtained the following law of large numbers for possibilistic variables.

Theorem 1. *Let X be a bounded variable in a possibility space $(\Omega, \mathcal{A}, \Pi)$ such that the possibility distribution π_X of X is upper semicontinuous. Let $\{X_n\}_n$ be a sequence of variables such that*

- (i) X_n are product related,
- (ii) X_n are identically distributed as X .

Then, for any fixed $\varepsilon > 0$,

$$N \left(\mathcal{M}[X] - \varepsilon < n^{-1} \sum_{i=1}^n X_i < \mathbb{M}[X] + \varepsilon \right) \rightarrow 1.$$

The first author's research in this paper was partially funded by Asturias's *Consejería de Economía y Empleo* (FC-15-GRUPIN14-101) and by Spain's *Ministerio de Economía y Competitividad* (MTM2015–63971–P).

Here \mathcal{A} is a σ -algebra, $\Pi : \mathcal{A} \rightarrow [0, 1]$ a possibility measure with N its dual necessity measure, π_X is given by $\pi_X(x) = \Pi(X = x)$, and $\mathcal{M}[X], \mathbb{M}[X]$ are the infimum and supremum, respectively, of the 1-cut of π_X . The requirement that X_n are *product related* means

$$\Pi(X_1 = x_1, \dots, X_n = x_n) = \Pi(X_1 = x_1) \dots \Pi(X_n = x_n)$$

for all $n \in \mathbb{N}$ and $x_i \in \mathbb{R}$. Indeed, Theorem 1 in its original presentation also considers the more general situation that the product is generalized to a continuous Archimedean triangular norm. This law of large numbers is in line with previous results in the literature of possibility measures [5, 6, 8, 10]. However, it must be compared to the law of large numbers of De Cooman and Miranda [2, Theorem 2] developed in the context of coherent lower previsions as a generalization of the law of large numbers for belief measures [7] (see also [9]) in which a similar limit interval appears but involves Choquet integrals instead. For more information on Choquet integrals and lower previsions, the reader is referred to Denneberg and Walley’s books, respectively [3, 14].

Indeed, by observing that the Choquet integral against a necessity measure is a coherent lower prevision, and also rewriting their result in a way closer to Theorem 1 for ease of comparison, we obtain the following.

Theorem 2 (De Cooman and Miranda). *Let X be a bounded variable in a possibility space $(\Omega, \mathcal{P}(\Omega), \Pi)$. Let $\{X_n\}_n$ be a sequence of variables such that*

- (i') (X_1, \dots, X_n) is forward factorizing for the Choquet integral E_N , for each $n \in \mathbb{N}$,
- (ii') X_n are uniformly bounded and such that $E_N[X_n] = E_N[X]$ and $E_\Pi[X_n] = E_\Pi[X]$ for all $n \in \mathbb{N}$.

Then, for any fixed $\varepsilon > 0$,

$$N \left(E_N[X] - \varepsilon < n^{-1} \sum_{i=1}^n X_i < E_\Pi[X] + \varepsilon \right) \rightarrow 1.$$

We emphasize that this version is weaker than [2, Theorem 2] in several respects, but will be better suited to our purpose. In it, E_N and E_Π denote the Choquet integrals with respect to N and Π , respectively. Condition (ii') is obviously satisfied if X_n are identically distributed as X , i.e. if (ii) holds. The restriction to the σ -algebra of parts $\mathcal{P}(\Omega)$ is inessential. Therefore, the substantial difference in the assumptions is condition (i') of *forward factorization*, namely the property that

$$E_N[g(X_1, \dots, X_{n-1})(h(X_n) - E_N[h(X_n)])] \geq 0 \tag{1}$$

for all $n \in \mathbb{N}$ and bounded functions $g : \mathbb{R}^{n-1} \rightarrow [0, \infty)$ and $h : \mathbb{R} \rightarrow \mathbb{R}$.

Condition (i') is rather different from condition (i), and any relationship between them is not obviously visible. In this communication, we aim at clarifying their relationships or lack thereof, in view of the fact that different conclusions

appear in Theorems 1 and 2. Both results claim that the averages $n^{-1} \sum_{i=1}^n X_i$ tend to be asymptotically confined inside a compact interval, but each yields a different interval: $[\mathcal{M}[X], \mathbb{M}[X]]$ in Theorem 1 and $[E_N[X], E_{\Pi}[X]]$ in Theorem 2.

There are three important remarks to be made. Firstly, both intervals have a special significance in fuzzy and possibility theory, making their study particularly relevant. Indeed, if the possibility distribution π_X is a fuzzy interval then $[\mathcal{M}[X], \mathbb{M}[X]]$ is its *core* and $[E_N[X], E_{\Pi}[X]]$ is its *mean value* in the sense of Dubois and Prade [4] (as follows immediately from the fact that $E_N[X]$ and $E_{\Pi}[X]$ are the infimum and the supremum of all expectations of X against probability measures dominated by Π , see [1, Lemma A.2] or [12, Proposition 3.5]).

Secondly, there is no incompatibility between both conclusions, since it is always the case that $[\mathcal{M}[X], \mathbb{M}[X]] \subset [E_N[X], E_{\Pi}[X]]$. Therefore the task of contrasting (i) and (i') is not a trivial one.

And thirdly, it may happen that $[E_N[X], E_{\Pi}[X]]$ is significantly larger than $[\mathcal{M}[X], \mathbb{M}[X]]$, which is reduced to a point if there is a unique point $x \in \mathbb{R}$ such that $\Pi(X = x) = 1$.

We will proceed by analyzing a specific type of function depending on two events, which eventually leads to 625 systems of equations and inequations, at least one of which must be satisfied if (i,i') hold simultaneously. Patient work reduces those systems to 14, which are finally shown to have solutions only under restrictive conditions, yielding the result stated in the abstract (see Corollary 5 below).

2 Forward Factorization and Product Relatedness

In this section we will prove that conditions (i) and (i') are compatible only under very special circumstances. To that end it is enough to consider the situation of a couple of variables X, Y instead of a whole sequence.

Our first result shows that certain functions of X and Y must have Choquet integrals of opposite signs under those conditions. Below, I_A and I_B denote the indicator functions of events A, B . We also use the notation \vee for the maximum.

Proposition 3. *Let X, Y be bounded variables in a possibility space $(\Omega, \mathcal{P}(\Omega), \Pi)$. Then, for any $A, B \subset \Omega$,*

(a) *If (X, Y) is forward factorizing for the Choquet integral E_N , then*

$$E_N[I_A(X)(I_B(Y) - N(Y \in B))] \geq 0.$$

(b) *If X and Y are product related, then*

$$E_N[I_A(X)(I_B(Y) - N(Y \in B))] \leq 0.$$

Proof. Part (a) follows directly from (1), taking $g = I_A$ and $h = I_B$ and observing

$$E_N[I_B(Y)] = E_N[I_{\{Y \in B\}}] = N(Y \in B).$$

As regards part (b), set

$$\kappa = E_N[I_A(X)(I_B(Y) - N(Y \in B))]$$

and the variable

$$Z = \begin{cases} 1, & X \in A, Y \in B \\ N(Y \in B), & X \notin A \\ 0, & X \in A, Y \notin B. \end{cases}$$

Now we work towards expressing κ in terms of possibilities:

$$\begin{aligned} \kappa &= E_N[I_A(X)I_B(Y) + N(Y \in B)I_{A^c}(X) - N(Y \in B)] \\ &= E_N[Z] - N(Y \in B) \\ &= N(Y \in B)N(\{X \in A, Y \in B\} \cup \{X \notin A\}) \\ &\quad + (1 - N(Y \in B))N(X \in A, Y \in B) - N(Y \in B) \\ &= N(Y \in B)N(\{X \in A, Y \in B^c\}^c) \\ &\quad + (1 - N(Y \in B))N(X \in A, Y \in B) - N(Y \in B) \\ &= -N(Y \in B)(1 - N(\{X \in A, Y \in B^c\}^c)) \\ &\quad + (1 - N(Y \in B))N(X \in A, Y \in B) \\ &= -(1 - \Pi(Y \in B^c))\Pi(X \in A, Y \in B^c) \\ &\quad + \Pi(Y \in B^c)(1 - \Pi(\{X \in A, Y \in B\}^c)). \end{aligned}$$

Let a, b, c, d be the possibilities of the events involved as summarized in the following table:

Π	$\ Y \in B \quad Y \in B^c \ $		
$X \in A$	a	b	$a \vee b$
$X \in A^c$	c	d	$c \vee d$
	$a \vee c$	$b \vee d$	

With that notation,

$$\kappa = -(1 - (b \vee d)) \cdot b + (b \vee d)(1 - (b \vee c \vee d)).$$

Since X and Y are product related,

$$\begin{aligned} b &= \Pi(X \in A, Y \in B^c) = \sup_{x \in A, y \in B^c} \Pi(X = x, Y = y) \\ &= \sup_{x \in A, y \in B^c} \pi_X(x)\pi_Y(y) = \Pi(X \in A)\Pi(Y \in B^c) = (a \vee b)(b \vee d), \end{aligned}$$

whence

$$\begin{aligned} \kappa &= -(1 - (b \vee d)) \cdot (a \vee b) \cdot (b \vee d) + (b \vee d)(1 - (b \vee c \vee d)) \\ &= (b \vee d)[1 - (b \vee c \vee d) - (1 - (b \vee d))(a \vee b)]. \end{aligned} \tag{2}$$

Observing

$$1 = \Pi(\Omega) = a \vee b \vee c \vee d,$$

there are two possibilities:

CASE 1. If $a = 1$, then

$$\kappa = (b \vee d)[(b \vee d) - (b \vee c \vee d)] \leq 0.$$

CASE 2. If $b \vee c \vee d = 1$, then

$$\kappa = -(b \vee d)(1 - (b \vee d))(a \vee b) \leq 0.$$

Hence $\kappa \leq 0$ and the proof is complete. □

It is clear from Proposition 3 that forward factorization and product relatedness can occur simultaneously only if, in the notation of its proof, $\kappa = 0$. That has definite consequences for the possible distributions of X and Y , as our main result shows that at least one of them must be uniform, i.e. there is a set A such that $\pi_X = I_A$ or $\pi_Y = I_A$.

Theorem 4. *Let X and Y be bounded variables in a possibility space $(\Omega, \mathcal{P}(\Omega), \Pi)$. Conditions*

- (I) (X, Y) is forward factorizing for the Choquet integral E_N
- (II) X and Y are product related

cannot be simultaneously met unless at least one of the variables is uniform.

Proof. Let $A, B \subseteq \mathbb{R}$. By Proposition 3, if both (I) and (II) hold then it must be

$$E_N[I_A(X)(I_B(Y) - N(Y \in B))] = 0$$

and therefore for a, b, c, d in the notation of (2) we have

$$(b \vee d)[1 - (b \vee c \vee d) - (1 - (b \vee d))(a \vee b)] = 0, \tag{3}$$

whence

$$b \vee d = 0 \text{ (i.e. } b = d = 0)$$

or

$$1 - (b \vee c \vee d) = (1 - (b \vee d))(a \vee b).$$

Since $a \vee b \vee c \vee d = 1$, there are 3 possibilities for the latter equation:

- . If $a = 1$, it becomes $1 - (b \vee c \vee d) = 1 - (b \vee d)$ i.e. $c \leq b \vee d$.
- . If $b = 1$ or $d = 1$, then it always holds.
- . If $c = 1$, it becomes $0 = (1 - (b \vee d))(a \vee b)$, whence $b \vee d = 1$ or $a = b = 0$.

The solution $c = 1, b \vee d = 1$ is already included in either case $b = 1$ or $d = 1$, whence (3) is rewritten as

$$\begin{aligned}
 b = d = 0 \quad \text{or} \quad a = 1, c \leq b \vee d \quad \text{or} \quad b = 1 \\
 \text{or} \quad c = 1, a = b = 0 \quad \text{or} \quad d = 1.
 \end{aligned}
 \tag{4}$$

The same reasoning applies to the pairs of events (A^c, B) , (A, B^c) and (A^c, B^c) , from which the analogous conditions

$$\begin{aligned}
 d = b = 0 \quad \text{or} \quad c = 1, a \leq d \vee b \quad \text{or} \quad d = 1 \\
 \text{or} \quad a = 1, c = d = 0 \quad \text{or} \quad b = 1;
 \end{aligned}
 \tag{5}$$

$$\begin{aligned}
 a = c = 0 \quad \text{or} \quad b = 1, d \leq a \vee c \quad \text{or} \quad a = 1 \\
 \text{or} \quad d = 1, b = a = 0 \quad \text{or} \quad c = 1;
 \end{aligned}
 \tag{6}$$

$$\begin{aligned}
 c = a = 0 \quad \text{or} \quad d = 1, b \leq c \vee a \quad \text{or} \quad c = 1 \\
 \text{or} \quad b = 1, d = c = 0 \quad \text{or} \quad a = 1.
 \end{aligned}
 \tag{7}$$

are derived. Thus conditions (4) through (7) might simultaneously be satisfied in $5^4 = 625$ different ways. Since a, b, c, d come from a possibility measure we have the restrictions

$$0 \leq a \leq 1, \quad 0 \leq b \leq 1, \quad 0 \leq c \leq 1, \quad 0 \leq d \leq 1, \quad a \vee b \vee c \vee d = 1 \tag{8}$$

as well as, by the product relatedness,

$$\begin{aligned}
 a &= (a \vee b)(a \vee c), & b &= (a \vee b)(b \vee d), \\
 c &= (a \vee c)(c \vee d), & d &= (b \vee d)(c \vee d).
 \end{aligned}
 \tag{9}$$

The task of finding a, b, c, d is thus tantamount to solving these 625 systems of 13 to 17 equations and inequations in 4 unknowns.

We start by combining restrictions (4) through (7), adding one at a time and always using (8) to simplify the obtained conditions if convenient (thus, for example, $a = b = 0$ would replace $a \vee b = 0$).

Conditions (4) and (5) can be satisfied in 25 ways, of which the following 6 contain all others:

1. $b = d = 0$
2. $a = c = 1, b \vee d = 1$
3. $c = d = 0, a = 1$
4. $b = 1$
5. $a = b = 0, c = 1$
6. $d = 1$

Merging these with (6), conditions (4) through (6) can be satisfied in 30 ways, of which the following 14 contain all others:

1. $b = d = 0, a = 1$
2. $b = d = 0, c = 1$
3. $a = b = 1$
4. $a = c = 1, b \vee d = 1$
5. $c = d = 0, a = 1$
6. $a = c = 0, b = 1$
7. $b = 1, d \leq a \vee c$
8. $b = c = 1$
9. $a = b = 0, c = 1$
10. $a = c = 0, d = 1$
11. $b = d = 1, a \vee c = 1$
12. $a = d = 1$
13. $a = b = 0, d = 1$
14. $c = d = 1$

Merging these with (7), conditions (4) through (7) can be satisfied in 70 ways, of which the following 14 contain all others:

1. $b = d = 0, a = 1$
2. $b = d = 0, c = 1$
3. $a = b = 1$
4. $a = c = 1, b \vee d = 1$
5. $c = d = 0, a = 1$
6. $a = c = 0, b = 1$
7. $b = d = 1, a \vee c = 1$
8. $b = c = 1$
9. $c = d = 0, b = 1$
10. $a = b = 0, c = 1$
11. $a = c = 0, d = 1$
12. $a = d = 1$
13. $a = b = 0, d = 1$
14. $c = d = 1$

With a direct inspection of (9) in each of the fourteen cases, after eliminating redundancies and imposing (8) on the range of the variables we finally arrive at the following ten families of solutions:

1. $a = 0, b = 0, c = 1, d \in [0, 1]$.
2. $a = 0, b = 0, c \in [0, 1], d = 1$.
3. $a = 0, b = 1, c = 0, d \in [0, 1]$.
4. $a = 0, b \in [0, 1], c = 0, d = 1$.
5. $a = 1, b = 0, c \in [0, 1], d = 1$.
6. $a \in [0, 1], b = 0, c = 1, d = 0$.
7. $a = 1, b \in [0, 1], c = 0, d = 0$.
8. $a \in [0, 1], b = 1, c = 0, d = 0$.

- 9. $a = 1, b = 1, c = d \in [0, 1]$.
- 10. $a = b \in [0, 1], c = 1, d = 1$.

Reasoning by contradiction, assume now that X and Y were both not uniform. By definition, there would exist $x, y \in \mathbb{R}$ such that

$$p := \Pi(X = x) \in (0, 1), \quad q := \Pi(Y = y) \in (0, 1).$$

Taking $A = \{x\}$ and $B = \{y\}$ above, using (8) and (9) we obtain the table

Π	$Y \in B$	$Y \in B^c$
$X \in A$	pq	p
$X \in A^c$	q	1
	q	1

representing a solution which nonetheless is not in any of the ten families above, a contradiction. Therefore, indeed X or Y must be uniform. □

As a consequence, for sequences of variables we obtain the following corollary.

Corollary 5. *Let $\{X_n\}_n$ be a sequence of identically distributed variables in a possibility space $(\Omega, \mathcal{P}(\Omega), \Pi)$. If both forward factorization and product relatedness, i.e. conditions (i) and (i'), hold, then the X_n must have uniform possibility distributions.*

3 Discussion

It is interesting that the conditions studied here are barely compatible, in the sense that a sequence of identically distributed variables satisfying both must have distributions giving possibility 0 or 1 to every event. Thus Theorems 1 and 2 are complementary as regards those assumptions.

The original laws of large numbers from which they have been simplified are also complementary in that both have content not covered by the other. The law from Possibility Theory covers the situation that the marginals of the X_n are linked by a triangular norm more general than the product, whereas the one from Imprecise Probability is of course applicable beyond possibility measures and also shows that the speed of the convergence is exponential.

It would be tempting to conclude that this ‘almost incompatibility’ is the explanation of the fact that both laws exhibit different limit intervals, specially since $[\mathcal{M}[X], \mathbb{M}[X]] = [E_N[X], E_\Pi[X]]$ when both conditions apply (as follows from [13]).

However, it must be emphasized that such a conclusion is not warranted, i.e. it is unclear whether the larger interval in Theorem 2 is actually optimal under condition (i') when applied to possibility measures.

References

1. Castaldo, A., Maccheroni, F., Marinacci, M.: Random correspondences as bundles of random variables. *Sankhya Indian J. Stat.* **66**, 409–427 (2004)
2. de Cooman, G., Miranda, E.: Weak and strong laws of large numbers for coherent lower previsions. *J. Stat. Plan. Inference* **138**, 2409–2432 (2008)
3. Denneberg, D.: *Non-additive Measure and Integral*. Kluwer, Dordrecht (1994)
4. Dubois, D., Prade, H.: The mean value of a fuzzy number. *Fuzzy Sets Syst.* **24**, 279–300 (1987)
5. Fullér, R.: A law of large numbers for fuzzy numbers. *Fuzzy Sets Syst.* **45**, 299–303 (1992)
6. Hong, D.H., Kim, Y.M.: A law of large numbers for fuzzy numbers in a Banach space. *Fuzzy Sets Syst.* **77**, 349–354 (1996)
7. Maccheroni, F., Marinacci, M.: A strong law of large numbers for capacities. *Ann. Probab.* **33**, 1171–1178 (2005)
8. Puhalskii, A.: *Large Deviations and Idempotent Probability*. Chapman & Hall/CRC, Boca Raton (2001)
9. Rébillé, Y.: Laws of large numbers for continuous belief measures on compact spaces. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **17**, 685–704 (2009)
10. Terán, P.: On convergence in necessity and its laws of large numbers. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O. (eds.) *Soft Methods for Handling Variability and Imprecision*. ASC, vol. 48, pp. 289–296. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85027-4_35
11. Terán, P.: Law of large numbers for the possibilistic mean value. *Fuzzy Sets Syst.* **245**, 116–124 (2014)
12. Terán, P.: Laws of large numbers without additivity. *Trans. Am. Math. Soc.* **366**, 5431–5451 (2014)
13. Terán, P.: A unified approach to the laws of large numbers for possibility measures in the context of more general uncertainty theories (Submitted for publication)
14. Walley, P.: *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London (1991)



Improved Performance of EK-NNClus by Selecting Appropriate Parameter

Qian Wang^{1(✉)} and Zhi-gang Su²

¹ School of Energy and Power, Jiangsu University of Science and Technology,
Zhenjiang 212003, China
wangqian16@just.edu.cn

² School of Energy and Environment, Southeast University,
Nanjing 210096, China

Abstract. EK-NNclus is an evidential clustering method based on the evidential K -nearest neighbors classification rule. Its one significant merit is that it does not require any priori on the number of clusters. However, the EK-NNclus suffers from the influence of number K . In other words, the performance of EK-NNclus is sensitive to K : if the number K is too small, the natural cluster may be split into two or more clusters; otherwise, two or more natural clusters may be merged into one cluster. In this paper, we indicated that tuning the parameters (such as α in the discounting function) can take full advantage of the distances between the object and its nearest neighbors, which can prevent natural clusters from being merged. Some numerical experiments were conducted and the experimental results suggested that the performance of EK-NNclus can be improved if appropriate α is selected.

Keywords: EK-NNClus · Evidence theory · Clustering performance

1 Introduction

Clustering algorithm has been widely used in all kinds of fields, such as image processing [1, 2], process control [3] and fault diagnosis [4, 5]. Fuzzy c -means (FCM) [6], proposed by Dunn, is one of the most popular clustering algorithm so far. However, FCM has two disadvantages: Firstly, its robustness is poor; secondly, it cannot well capture the imprecise information in the objects since it is based on the probabilistic framework. To overcome these disadvantages, Masson and Denoeux [7] proposed a new version of fuzzy c -means algorithm, named evidential c -means algorithm (ECM). ECM is based on the evidential theory, which is good at modelling both uncertainty and imprecision. In ECM, a mass belief, the counterpart in FCM, is not only allocating to a single cluster, but also to a meta cluster (containing two or more single clusters). In addition, the robustness will be improved since the noise can be distinguished by a null cluster.

In all those clustering methods mentioned above, the cluster number, which is difficult to be determined, should be fixed in advance. To encompass the problem, Denoeux [8] proposed a novel clustering algorithm named evidential K -nearest neighbors clustering method (EK-NNclus). It is a decision-directed approach to clustering, the classifier, based on the evidential K -nearest neighbors rule, is used to label the objects. After each object is labelled, the process will be repeated until no changes take place in the labels. With the help of Hopfield neural network [9,10], Denoeux proved that the clustering algorithm could converge to a fixed point. After convergence, the mass belief of both single and meta cluster could be obtained. Therefore, it is a robust clustering method. However, the number of nearest neighbors K should be fixed in EK-NNclus. If the K is too small, a natural cluster will be split into two or more clusters; if the K is too large, two or more natural clusters will be merged into one cluster. When K is relatively large, the parameter α in the discounting function could prevent natural clusters from being merged if it is appropriately selected.

Therefore, K and α are two key parameters in the EK-NNclus algorithm. The performance of EK-NNclus algorithm will be improved if these parameters are appropriately selected. The objective of our paper is to improve the clustering performance by selecting the suitable parameters. The rest of the paper is organized as follows. In Sect. 2, EK-NNclus algorithm will be recalled. Parameter selection and experiments are conducted in Sect. 3, concluded remarks are presented in Sect. 4.

2 Background

2.1 Belief Function Theory

Let $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ be the frame of discernment, a collectively exhaustive and mutually exclusive set of c hypotheses or propositions. The mass function m is defined on the power set $2^\Omega = \{A : A \subseteq \Omega\} \rightarrow [0, 1]$, it is said to be basic belief function (BBA) if the following equation satisfies: $\sum_{A \subseteq \Omega} m(A) = 1$. A BBA is normal if $m(\emptyset) = 0$ otherwise it is subnormal. Any subset A of Ω such that $m(A) > 0$ is called a focal element. From the BBA, an evidential function called plausibility function can be defined in Eq. (1). Another function $pl : \Omega \rightarrow [0, 1]$ such that $pl(\omega_i) = Pl(\{\omega_i\})(\omega_i \in \Omega)$ is called contour function. pl is a probability function which can be calculated from a belief function by smets method [11].

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B); \quad \forall A \subseteq \Omega \tag{1}$$

If two BBA's m_1 and m_2 are both independent, the standard way of combining them is through conjunctive fusion of them, defined in Eq. (2). The conjunctive fusion may produce subnormal belief assignment, and then we can convert

the subnormal one into a normal one by Dempster’s rule [12] ($m^* = m_1 \oplus m_2$), defined in Eq. (3).

$$m(C) = \sum_{A \cap B = C} m_1(A)m_2(B); \quad \forall C \subseteq \Omega \tag{2}$$

$$m^*(C) = \frac{m(C)}{(1 - \kappa)}; \quad \forall C \subseteq \Omega \tag{3}$$

where $\kappa = \sum_{A \cap B = \emptyset} m_1(A)m_2(B)$

The full combination of two mass functions is quite time-consuming in very large frames of discernment. To solve this problem, combination of contour function is introduced in Eq. (4).

$$pl_1 \oplus pl_2 = \frac{pl_1 pl_2}{1 - \kappa} \tag{4}$$

2.2 EK-NNclus Algorithm

Let us consider a classification problem, a training set $X = \{x_1, x_2, \dots, x_n\}$ should be grouped into a set of classes. d_{ij} represents the distance between object x_i , to be classified, and the object x_j , one of the K -nearest neighbor of x_i . If the object x_j belongs to the class ω_k the distance d_{ij} can generate a piece of evidence which can be represented by the following mass function in Eqs. (5)–(7). According to Eqs. (5)–(7), the mass function only has two focal elements [8, 13]. The function φ in Eqs. (5) and (6) is a decreasing function represented in Eq. (8).

$$m_{ij}(\omega_k) = \alpha\varphi(d_{ij}) \tag{5}$$

$$m_{ij}(\Omega) = 1 - \alpha\varphi(d_{ij}) \tag{6}$$

$$m_{ij}(A) = 0; \quad \forall A \subseteq 2^\Omega \setminus \{\omega_k, \Omega\} \tag{7}$$

$$\varphi(d_{ij}) = e^{(-\gamma d_{ij}^\beta)} \tag{8}$$

where α , β and γ are tuning parameters.

The object x_i can obtain K pieces of evidence from the K nearest neighbors. A combined mass function can be yielded after these evidences are combined by DS rule. To make a decision, the mass function m does not need to be calculated explicitly. The contour function pl_i corresponding to m_i is

$$pl_i(\omega_k) = (1 - \alpha\varphi(d_{ij}))^{(1-s_{jl})} \tag{9}$$

where $l = 1, 2, \dots, c$. If $l = k$, $s_{jl} = 1$; otherwise, $s_{jl} = 0$.

From Eq. (4), the combined result of contour function is

$$pl(\omega_k) \propto \prod_{j=1}^K (1 - \alpha\varphi(d_{ij}))^{(1-s_{jl})} \tag{10}$$

Let us take logarithm on both sides of Eq. (10)

$$\ln(pl(\omega_k)) = \sum_{j=1}^K (s_{jl}v_j) + C \tag{11}$$

where $v_j = -\ln(1 - \alpha\varphi(d_{ij}))$ and C is a constant.

The process of EK-NNclus algorithm is introduced briefly as follows:

Initialization: The cluster number is assumed to be $n - 1$ and the objects are labelled randomly.

Iteration: By ignoring the constant term, we calculate u_{ik} the logarithms of the plausibilities of belonging to the each cluster from the Eq. (11) as

$$u_{ik} = \sum_{j=1}^K s_{jl}v_j, \quad k = 1, 2, \dots, c. \tag{12}$$

The object x_i will be assigned to the cluster with the highest plausibility. The variable s_{ik} can be updated as

$$\begin{cases} u_{ik} = 1 & \text{if } u_{ik} = \max_{l=1,2,\dots,c} u_{il} \\ u_{ik} = 0 & \text{otherwise} \end{cases} \tag{13}$$

After each iteration, the cluster number will become smaller since some clusters are disappeared. The objects are randomly re-labelled and a new iteration is started.

Decision: After the iterative process has converged, we can calculate the final mass function for every object through DS rule. Every object will be grouped into a cluster according to its mass function.

3 Parameter Selection and Experiments

Although the cluster number does not need to be fixed in the EK-NNclus algorithm, the number of nearest neighbors K is difficult to be determined. If the K is too large, some natural clusters may be merged together; if the K is too small one natural cluster may be split into two or more clusters. The α in Eq. (6) is a key parameter in the EK-NNclus algorithm, since it greatly affects the plausibility which determines the object’s cluster. For example, if α is chosen to be 0.95 as reference [13] did, the maximum of v_j in Eq. (11) is $-\ln(0.05)$. When K is relatively large, for the object x_i , the v_j belong to two adjacent natural clusters may be very close and become distinguishable. The object x_i may obtain wrong class label since the two adjacent natural clusters may be merged. To decrease the possibility that the natural clusters to be merged, α should be chosen to be very close to 1 for it can enlarge the maximum of v_j .

To illustrate the two key parameters' (K and α) affection on the clustering performance, some examples are shown as follows:

Example 1. $X = [51, 1; 2, 52; 8, 8; 222, 250; 305, 355; 309, 359; 1101, 1001; 955, 1005; 1002, 1102; 2057, 2107; 1908, 2008; 2009, 2009; -20, 6; 7, -20; 357, 317; 250, 310; 1026, 1106; 1110, 1060; 2112, 2052; 2213, 2113]$; $Y = [1; 1; 1; 2; 2; 2; 3; 3; 3; 3; 4; 4; 4; 1; 1; 2; 2; 3; 3; 4; 4]$;

X , plotted in Fig. 1, contains 20 objects described by two attributes. Y is the true clustering result of the dataset X . The number of nearest neighbor K is 19.

The v_j of first object (51,1) is shown in Fig. 2 when α is 0.9, 0.95, 0.99 and 0.999999, respectively. As shown in Fig. 2, when α is 0.9, 0.95 and 0.99, v_j belongs to class 1 is very close to that belongs to class 2(The first four nearest neighbors belong to class1 and the fifth to ninth nearest neighbors belong to class 2). Therefore, the objects in class 1 and class 2 will be merged easily on this occasion. However, when α is 0.999999, v_j in class 1 is quite different from that in class 2. Therefore, class 1 and class 2 will not be merged easily.

The v_j of last object (2213,2113) is shown in Fig. 3 when α is 0.9, 0.95, 0.99 and 0.999999, respectively. When α is 0.9, v_j belongs to class 4 is close to that belongs to class 3, therefore, objects belong to the two classes are easy to be merged. However, the parameter α is closer to 1, v_j belongs to class 4 is more distinguishable to that belongs to class 3.

We do $N = 50$ trials on the dataset X when α is 0.9, 0.95, 0.99 and 0.999999, respectively. And then we define probability of clustering number (c) as $p = n_c/N$, where n_c is the number of obtaining the clustering number (c). The results are shown in Table 1. With the increasing of parameter α , the probability of clusters merged is decreasing. Furthermore, when α is 0.999999, the probability of obtaining the right clustering number is 0.6 while it is 0 in other cases.

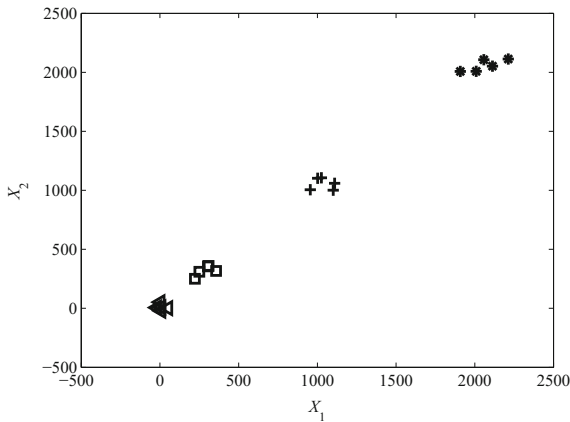


Fig. 1. X dataset

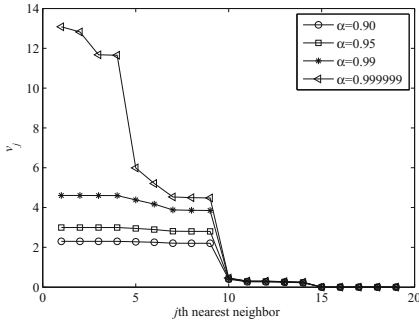


Fig. 2. v_j of the first object

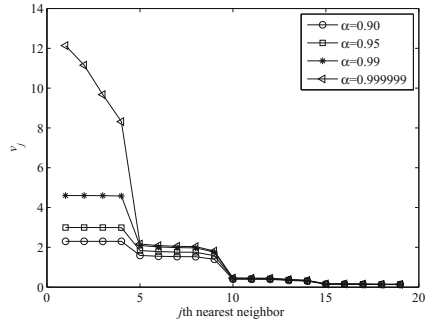


Fig. 3. v_j of the last object

Table 1. Probability of clustering number c

Clustering number	α			
	0.90	0.95	0.99	0.999999
1	0.2	0.12	0	0
2	0.68	0.62	0.36	0
3	0.12	0.26	0.64	0.4
4	0	0	0	0.6

Example 2. Take the “Fourclass” dataset as an example, it contains 400 objects described by two attributes. The dataset can be divided into four classes, which is shown in Fig. 4.

When K changes from 20 to 400 with the interval 20, and the last K is 399, we do $N=50$ trials on every K . And we choose $p = n_r/N$ as the index of clustering performance. n_r is the number of obtaining the right clustering number(c) among N trials. p vs. K is shown in Fig. 5.

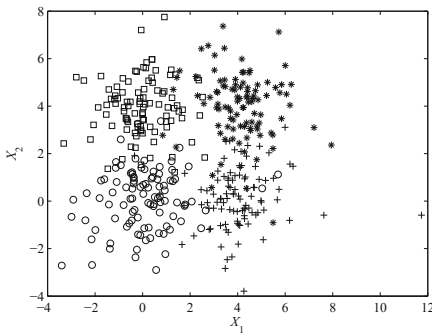


Fig. 4. Fourclass dataset

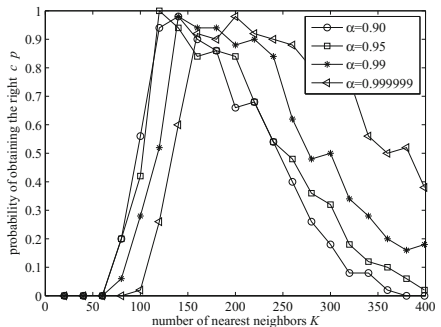


Fig. 5. p vs. K (Fourclass)

As shown in Fig. 5, in the case $\alpha = 0.9$, p is more than 50% when K varies from 100 to 240, and its interval is 140; in the case $\alpha = 0.95$, p is more than 50% when K varies from 120 to 240 and its interval is 120; in the case $\alpha = 0.99$, p is more than 50% when K varies from 120 to 260, and its interval is 140; in the case $\alpha = 0.999999$, p is more than 50% when K varies from 140 to 380 and its interval is 240. Therefore, in the case $\alpha = 0.999999$, K is the easiest to be determined. Moreover, when K is large than 200, the case $\alpha = 0.999999$ shows the best clustering performance among the four cases.

Example 3. Take the ‘‘R15’’ dataset as an example, it is composed of 600 objects described by two attributes. The dataset can be divided into 15 classes. R15 dataset is shown in Fig. 6. The same as the Example 2, when K changes from 20 to 600 with the interval 20, and the last K is 599. We do $N = 50$ trials on every K . p vs. K is shown in Fig. 7.

As shown in Fig. 7, in the case $\alpha = 0.9$, p is more than 50% when K varies from 60 to 80, and its interval is 20; in the case $\alpha = 0.95$, p is more than 50% when K varies from 60 to 120 and its interval is 60; in the case $\alpha = 0.99$, p is more than 50% when K varies from 80 to 140, and its interval is 60; in the case $\alpha = 0.999999$, p is more than 50% when K varies from 80 to 280 and its interval is 200. When α is equal to 0.999999, K is the easiest to be chosen among the four cases. Moreover, when K is more than 100, the case $\alpha = 0.999999$ shows the best clustering performance. Compared with the Example 2, K , in the Example 3, is difficult to be determined since its interval is more narrow. The reason is the dataset in the Example 3 is more complex: some clusters are close to each other, some are far from each other.

Remark 1. When K is relatively large (more than 200 in Example 2), enlarging the maximum of v_j , by selecting an α close to 1, will get better clustering performance.

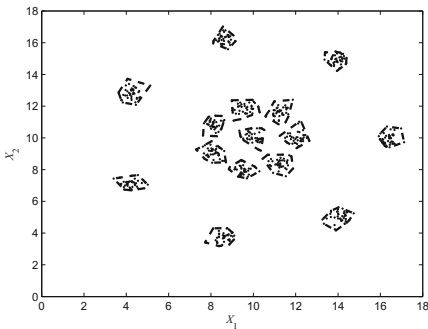


Fig. 6. R15 dataset

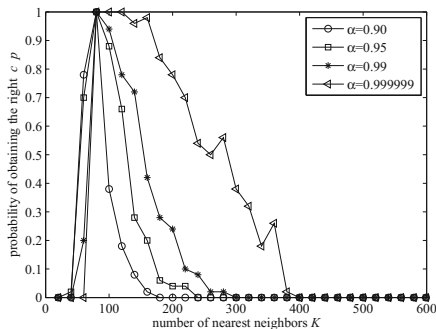


Fig. 7. p vs. K (R15)

4 Conclusion

The parameter α in discounting function is a key parameter in EK-NNclus. It can improve the clustering performance if α is well chosen, especially when K is relatively large, since it can take full advantage of the distances between the object and its nearest neighbors. However, it cannot guarantee that the right clustering number will be found, since it just could prevent natural clusters from being merged. Therefore, further study should be done to improve the clustering performance.

References

1. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c -means clustering with spatial information for image segmentation. *Comput. Med. Imaging Graph.* **30**(1), 9–15 (2006)
2. Zhang, X., Wang, G., Su, G., Guo, Q., Zhang, C., Chen, B.: An improved Fuzzy algorithm for image segmentation using peak detection, spatial information and reallocation. *Soft Comput.* **21**(8), 2165–2173 (2017)
3. Ikonen, E., Selek, I., Najim, K.: Process control using finite Markov chains with iterative clustering. *Comput. Chem. Eng.* **93**, 293–308 (2016)
4. Yiakopoulos, C.T., Gryllias, K.C., Antoniadis, I.A.: Rolling element bearing fault detection in industrial environments based on a K-means clustering approach. *Expert Syst. Appl.* **38**(3), 2888–2911 (2011)
5. Yin, S., Huang, Z.: Performance monitoring for vehicle suspension system via fuzzy positivistic C-Means clustering based on accelerometer measurements. *IEEE/ASME Trans. Mechatron.* **20**(5), 2613–2620 (2015)
6. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* **3**(3), 32–57 (1974)
7. Masson, M.H., Denoeux, T.: ECM: an evidential version of the fuzzy c -means algorithm. *Pattern Recogn.* **41**(4), 1384–1397 (2008)
8. Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S.: EK-NNclus: a clustering procedure based on the evidential K-Nearest neighbor rule. *Knowl. Based Syst.* **88**(3), 57–69 (2015)
9. Galan-Marin, G., Munoz-Perez, J.: Design and analysis of maximum Hopfield networks. *IEEE Trans. Neural Netw.* **12**(2), 329–39 (2001)
10. Hopfield, J.: Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**(8), 2554–2558 (1982)
11. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. *Int. J. Approximate Reasoning* **38**, 133–147 (2005). Elsevier Science Inc
12. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
13. Denoeux, T.: A K-Nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (2005)



An Empirical Study to Determine the Optimal k in Ek-NNclus Method

Yiru Zhang^(✉), Tassadit Bouadi, and Arnaud Martin

Univ Rennes 1, CNRS, IRISA, Rennes, France
{yiru.zhang,tassadit.bouadi,arnaud.martin}@irisa.fr
<http://www-druid.irisa.fr/>

Abstract. Ek-NNclus is a clustering algorithm based on the evidential k -nearest-neighbor rule. It has the advantage that the number of clusters can be detected. However, the parameter k has crucial influence on the clustering results, especially for the number of clusters and clustering quality. Thus, the determination of k is an important issue to optimize the use of the Ek-NNclus algorithm. The authors of Ek-NNclus only give a large interval of k , which is not precise enough for real applications. In traditional clustering algorithms such as c -means and c -medoid, the determination of c is a real issue and some methods have been proposed in the literature and proved to be efficient. In this paper, we borrow some methods from c determination solutions and propose a k determination strategy based on an empirical study.

Keywords: Ek-NNclus · k determination · Clustering

1 Introduction

In cluster analysis, choosing the optimal number of clusters is a well-known problem [1, 7]. For many clustering algorithms (such as c -means, c -medoids, etc.), the number of clusters noted by c must be pre-defined¹. The correct choice of c is not simple, needing most of the time a subjective interpretation of some criterion directly linked with the structure of data and the wanted clustering resolution.

The Ek-NNclus method, proposed by [3], does not need the pre-definition of the parameter c and is able to detect the number of clusters. However, as Ek-NNclus is based on the k -nearest neighbors, the parameter k , given the size of neighborhood, should be set. Different k may result in various clustering results and often with different number of clusters. Therefore, Ek-NNclus has replaced the problem of c determination by the problem of finding a proper value for k . In [3], the authors concluded that the results of clustering are mostly conducted

¹ In many articles, the number of clusters is denoted by k . To avoid ambiguity with another parameter k of k -nearest neighbors in Ek-NNclus algorithm, we use c in this article.

by the parameter k . Following the rule of thumb, such as the determination k in the k -nearest neighbors classifier, the authors also give an empirical suggestion on the determination of k , which is *two or three times* \sqrt{n} where n denotes the number of all objects. The range between two or three times \sqrt{n} is sometimes too wide and even within this range, the clustering results are still quite different. Besides, in Ek-NNclus, the existence of some random processes makes the method not perfectly reproducible (*i.e.* on one dataset, with identical k , clustering results may not be even close).

Moreover, the optimal k varies with the scale of the data, making the determination of k necessary for every clustering analysis problem. The determination of k is two-fold. An optimal k in Ek-NNclus should:

1. Cluster the data into the correct number of clusters;
2. Return a result with high quality, close to the real partitions of objects.

There are already some often-applied methods to determine c , such as evaluation criteria (*e.g.* silhouette coefficient [10]) optimization, elbow method and information criterion approach. In this article, we borrow and test these methods to evaluate if they are still applicable for the determination of k in Ek-NNclus. We also propose a determination strategy based on these methods.

In the following parts, we briefly introduce the Ek-NNclus algorithm as well as some criteria for c determination in c -means in Sects. 2 and 3. In Sect. 4, we introduce the proposed k determination strategy. We illustrate this strategy on synthetic data and real-world data in Sect. 5 and give a conclusion in Sect. 6.

2 Ek-NNclus Algorithm

Ek-NNclus is a clustering algorithm based on the evidential k -nearest-neighbor classifier. It requires only the pairwise metric for k -nearest-neighbor searching. Ek-NNclus starts from an initial random partition, and reassigns objects to clusters iteratively using Ek-NN classifier [2]. The algorithm converges to a stable partition. For each object, its membership to clusters is described by a mass function in a framework of each cluster and the whole set of clusters (*i.e.* ignorance). Given a matrix of pairwise distances $D = (d_{ij})$, where d_{ij} denotes the distance between object o_i and object o_j , according to [3], the procedure of EkNNclus can be briefly divided into the following parts:

- **Preparation.** Calculate the mass value α_{ij} of the event: o_j is in the k -nearest neighbors of o_i based on d_{ij} by a non-increasing mapping function $\phi(d_{ij})$. Naturally, $\alpha_{ij} = 0$ if o_j does not belong to the k -nearest neighbors of o_i .
- **Initialization.** Initialize the labels of each object randomly. The authors of [3] suggest that the number of clusters c can be set to the number of objects n if n is not too large.
- **Iteration.** Randomly reorder all objects. Then, for every object $o_{i'}$ in the new order, calculate the plausibility of belonging to each cluster. Assign $o_{i'}$ to the cluster with the highest plausibility.

- **Convergence condition.** The iterations stop when the labels of all objects are stable.

In this procedure, the number of k at the preparation step has a vital impact on the clustering results. If k is too small, the matrix of α becomes sparse. In this case, the number of iterations is small and the clustering result highly depends on the initialization step, which is usually random. If k is too large, two objects far away from each other may be considered as in the same neighborhood. This may have two consequences:

1. The computation time becomes important;
2. Objects naturally in different clusters may be targeted as in the same one, causing an underestimation of number of clusters.

Therefore, the determination of k is important to guarantee a good quality of clustering.

3 Some Methods in c Determination

Some solutions from c determination for c -means algorithm are borrowed to help the determination of k in Ek-NN algorithm. In this section, we introduce how Adjusted Rand Index (ARI), silhouette coefficient and elbow method are applied for the determination of optimal c .

Adjusted Rand Index (ARI)

Rand index (RI) [8] is a measure of similarity between two data clustering. Developed from RI, Adjusted RI (ARI) is adjusted for chance grouping of objects in clusters [8, 12]. We use ARI as the priority criterion for the evaluation of the clustering result with the knowledge of the ground truth given. Thus, the cluster number c that returns the highest ARI value is determined as the optimal one.

Silhouette coefficient

Silhouette coefficient is useful in determining the natural number of clusters [1, 10]. The silhouette coefficient is an evaluation criterion, in which the calculation is only based on the intra-class and inter-class distances of each object pair. A higher silhouette coefficient score relates to a model with better defined clusters. Thus, the problem of optimal c determination can be transferred to a silhouette coefficient maximization problem [1]. Another advantage of silhouette coefficient is that only pairwise distances are needed and the calculating of centers is avoided. Indeed, independent to centroid is a good property. For some metrics where only pairwise distances are given, the calculation of centroid is a metric k -center problem, proved to be NP-hard [4].

Elbow method

The elbow method [11] applies the distortion as a criterion for clustering result. The rule is simple: among different number of clusters \mathcal{C} , one should choose a number $c \in \mathcal{C}$, such that $c+1$ clusters do not give a much better modeling of the data. Given n objects in c clusters, we denote the objects by x_1, x_2, \dots, x_n and

the center of clusters by $\mu_1, \mu_2, \dots, \mu_c$. The quality of the modeling is measured by the distortion J of the clustering, calculated by:

$$J(c, \mu) = \frac{1}{n} \sum_{i=1}^n \left(\min_{j=1}^c (x_i - \mu_j)^2 \right) \quad (1)$$

Therefore, c can be subjectively determined with the help of a distortion plot helps, illustrated in the experiment part of Sect. 5.2.

A disadvantage of this method is that *the “elbow” cannot always be unambiguously identified* [5]. The observation of the “elbow” is subjective because “a cluster that does not give a much better modeling of the data” cannot be justified quantitatively. Another inconvenience of the elbow method is that the calculation of distortion is based on the centroid of each cluster. This jeopardizes the property that Ek-NNclus is independent of the calculation of centroid.

4 A k Determination Strategy

The idea of k determination is simple: an optimal k in EkNNclus should return a high quality clustering result. Given a dataset, the quality of clustering can be easily evaluated if knowledge of ground truth is provided. A high value of ARI between clustering result and the ground truth implies a good clustering quality. However, in most cases, the ground truth is absent. The results of clustering are often evaluated by how well different clusters are separated. Silhouette coefficient is such a criteria and it is often strongly correlated with ARI. The correlation is plotted in the Sect. 5.1. However, to determine k only by silhouette coefficient is still risky. Fewer clusters may sometimes return a higher silhouette coefficient (example illustrated in Sect. 5.1 and Fig. 4b). Thus, other conditions are needed. Elbow method is used as the second criterion to avoid that too few clusters are detected. The strategy is straightforward. From the intersection of the set of k (\mathcal{K}_c) corresponding to the best c and the set of k (\mathcal{K}_{sil}) corresponding to relatively high silhouette coefficient, the interval of values of k is obtained. We denote a set of all possible k by \mathcal{K} . A proper subset of k is therefore refined by: $\mathcal{K}_{refine} = \mathcal{K}_c \cap \mathcal{K}_{sil}$. We define a silhouette efficient function $f_{sc}(k)$, implying the silhouette coefficient of the clustering result with k in Ek-NNclus algorithm. Thus, the optimal k is given by:

$$k = \arg \max_{k \in \mathcal{K}_{refine}} (f_{sc}(k)). \quad (2)$$

Note that the elbow method is subjective and that “relatively high silhouette coefficients” are also subjectively defined, both \mathcal{K}_c and \mathcal{K}_{sil} are not definite sets. Thus, if $\mathcal{K}_{refine} = \emptyset$, we can extend \mathcal{K}_c by softer condition or \mathcal{K}_{sil} by lower threshold to obtain a non empty \mathcal{K}_{refine} .

5 Experimentation Results

In this section, we study the correlation between ARI and silhouette coefficient, and then applied our strategy on toy datasets. The synthetic data are generated by Gaussian distributions. For the sake of better visualization, the synthetic data are always generated in a 2 dimensional space.

5.1 Correlation Between ARI and Silhouette Coefficient

We generate synthetic datasets for this experiment. The procedure is as follows:

1. Given a set of standard deviation (noted *std*) and the number of clusters denoted by n_{clus} , we generate a set of datasets $\mathcal{S}_{data} = \{X_1, X_2, \dots, X_D\}$ with ground truth. Datasets with 8 clusters and with $std = 0.5, 1.0, 2, 2.5$ are illustrated in Fig. 2.
2. On one dataset $X_d \in \mathcal{S}_{data}$, given a set of parameter values $\mathcal{K} = \{k_1, k_2, \dots, k_{|K|}\}$, calculate ARI and silhouette coefficient of each $k \in \mathcal{K}$. A set of ARIs and silhouette coefficients are obtained corresponding to different k , respectively denoted as \mathcal{S}_{ARI} and \mathcal{S}_{sil} . The Pearson correlation coefficient [9] $\rho(\mathcal{S}_{ARI}, \mathcal{S}_{ARI})$ is calculated for dataset X_d , denoted by ρ_d .

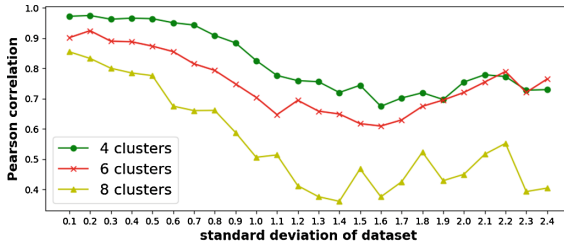


Fig. 1. Pearson correlation coefficient between ARI and silhouette vs data sets with different *std*.

Figure 1 illustrates the variation of the correlation between ARI and silhouette coefficient via different standard deviations. We observe that the correlation declines while data are distributed more sparsely. From a certain standard deviation, the correlation has a tendency to increase. These are datasets used in the experiment of Fig. 1. While *std* is small, data are obviously clustered. Thus a clustering result regrouping objects nearby is consistent with the knowledge of the ground truth, which returns a high correlation. With *std* increasing, different clusters overlap and the correlation decreases. When *std* is high enough that data distribution converges to random, the clustering returns lows values on both ARI and silhouette coefficient, making them “correlated” again.

However, the strong correlation cannot guarantee that silhouette coefficient is enough for k determination. The ARI and silhouette coefficient obtained from

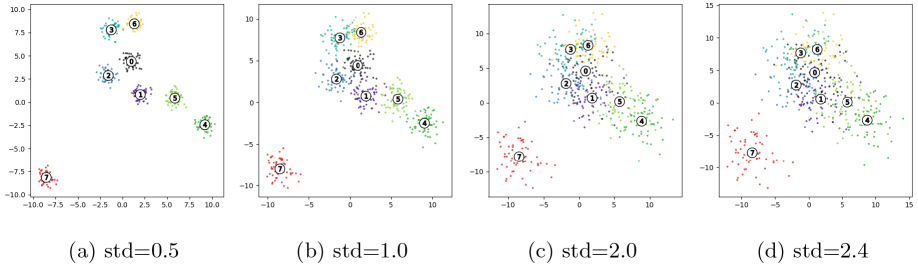


Fig. 2. Data distributions with different values of standard deviation.

different k on data in Fig. 2 are respectively plotted in Fig. 3. We observe that a high silhouette coefficient does not always correspond to a high ARI when value of k is large, even if objects in different clusters are naturally well separated (e.g. dataset with $std = 0.5$). This has been explained in Sect. 2 that a high value on k may cause underestimation of the number of clusters c , which may result in a satisfying silhouette coefficient. Elbow method determining the c helps to provide a constraint condition.

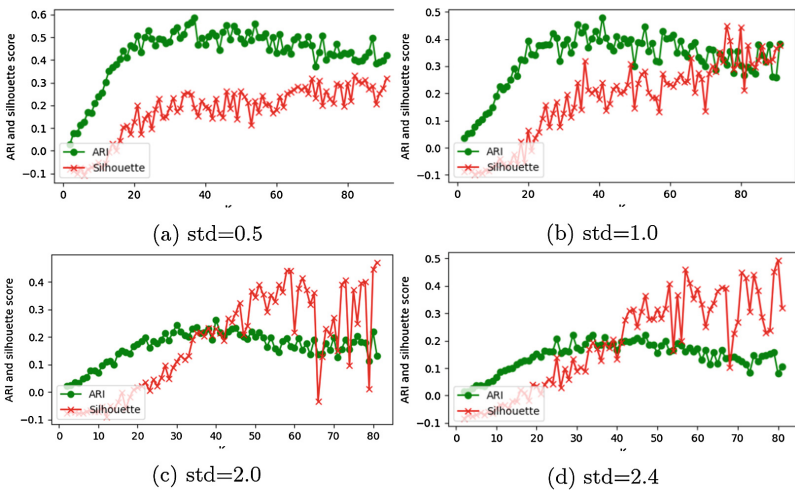


Fig. 3. ARI and Silhouette coefficient *via* k on different datasets.

5.2 Optimal k Determination Strategy on Real Toy Datasets

We applied the strategy in Sect. 4 on real toy datasets: Iris and Wine datasets from UCI² to help to refine the interval of k .

² Iris: <https://archive.ics.uci.edu/ml/datasets/Iris>.
 Wine: <https://archive.ics.uci.edu/ml/datasets/wine>.

Toy dataset *Iris*: Fig. 4 illustrates the plot supporting k determination strategy for *Iris* toy data. Results are obtained with a cross validation of 10 experiments. We still observe that the values of ARI, silhouette coefficient and number of clusters have large fluctuation, which proves that the determination of k is risky.

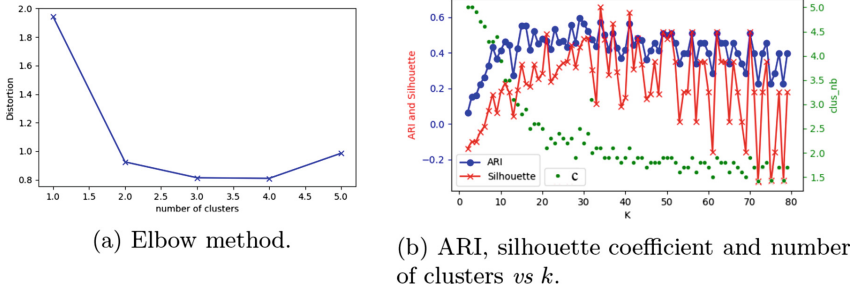


Fig. 4. Results on Iris dataset.

Without knowledge of c , from the silhouette coefficient plot in Fig. 4b, one may conclude that $k \in [30, 50]$ is the best value. With elbow method, we can figure that $c = 2$ or 3 is a reasonable value, so $k \in [15, 40]$ is more reasonable. Taking the intersection of both intervals, we focus on a refined interval $k \in [30, 40]$. In this interval, $k = 35$ returns the highest silhouette coefficient (given by the abscissa of Fig. 4b). Thus, finally we determine $k = 32$ by Eq. (2). With the ARI plot (given by the ordinate of Fig. 4b), we can verify that $k \approx 35$ is the proper value, so the proposed strategy is adapted.

Toy dataset *Wine*: The elbow method and clustering criteria plot are illustrated in Fig. 5. It is tricky to determine the number c of clusters by Elbow method for this dataset. Different observers may give different decisions on the best number of clusters. Therefore, 3 or 4 can both be concluded as c . According to Fig. 5b, $c \in \{3, 4\}$ corresponds approximately to $k \in [20, 50]$. A high silhouette

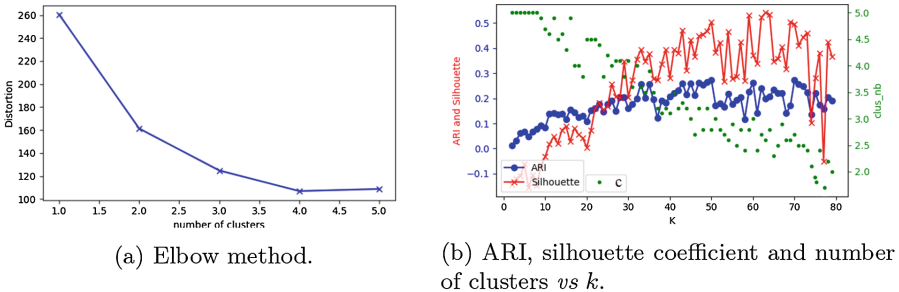


Fig. 5. Results on Wine dataset.

coefficient value corresponds to the interval $k \in [40, 70]$. By taking the intersection of both intervals, we conclude that a proper k should be in the interval $[40, 50]$ and we obtain $k = 49$ such as the optimal value by Eq. (2).

According to Fig. 5b, with only silhouette coefficient, we may arbitrarily choose a high value $k \in [60, 70]$. However, this value gives an underestimation of the c value. The elbow method fixing a proper number of clusters helps to determine a k that returns the highest ARI.

6 Conclusion

In this article, we discuss a practical problem encountered in the application of EkNNclus algorithm: the determination of the optimal number of nearest neighbors k . Based on some methods borrowed from determination of the number c of clusters in c -means, we proposed a combined strategy. In this strategy, silhouette coefficient is applied to evaluate the clustering quality and elbow method is used as an extensive procedure for over-fitting. Comparing with an empirical suggestive interval for k determination given by [3], the proposed strategy gives a more refined selection of k and guarantees a relative high quality of clustering.

The strategy has some short-comings conducted by elbow method. Firstly, the determination of c by elbow method is subjective and can be sometimes ambiguous. Besides, the distortion requires the calculation of centroids of clusters, which neutralizes an advantage of Ek-NNclus: Ek-NNclus is centroid independent. In the future, we can replace elbow method by centroid-independent c determination method, making the strategy more adaptable.

References

1. De Amorim, R.C., Hennig, C.: Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci.* **324**, 126–145 (2015)
2. Denoeux, T.: A k -nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (1995)
3. Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S.: Ek-NNclus: a clustering procedure based on the evidential k -nearest neighbor rule. *Knowl. Based Syst.* **88**, 57–69 (2015)
4. Hsu, W.L., Nemhauser, G.L.: Easy and hard bottleneck location problems. *Discret. Appl. Math.* **1**(3), 209–215 (1979)
5. Ketchen Jr., D.J., Shook, C.L.: The application of cluster analysis in strategic management research: an analysis and critique. *Strateg. Manag. J.* **17**, 441–458 (1996)
6. Lletí, R., Ortiz, M.C., Sarabia, L.A., Sánchez, M.S.: Selecting variables for k -means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Anal. Chim. Acta* **515**(1), 87–100 (2004)
7. Pham, D.T., Dimov, S.S., Nguyen, C.D.: Selection of k in k -means clustering. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **219**(1), 103–119 (2005)
8. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)

9. Rodgers, J.L., Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. *Am. Stat.* **42**(1), 59–66 (1988)
10. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
11. Thorndike, R.L.: Who belongs in the family? *Psychometrika* **18**(4), 267–276 (1953)
12. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010)



Evidential Community Detection Based on Density Peaks

Kuang Zhou¹(✉), Quan Pan¹, and Arnaud Martin²

¹ Northwestern Polytechnical University,
Xi'an 710072, Shaanxi, People's Republic of China
kzhoumath@163.com, quanpan@nwpu.edu.cn

² DRUID, IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France
Arnaud.Martin@univ-rennes1.fr

Abstract. Credal partitions in the framework of belief functions can give us a better understanding of the analyzed data set. In order to find credal community structure in graph data sets, in this paper, we propose a novel evidential community detection algorithm based on density peaks (EDPC). Two new metrics, the local density ρ and the minimum dissimilarity δ , are first defined for each node in the graph. Then the nodes with both higher ρ and δ values are identified as community centers. Finally, the remaining nodes are assigned with corresponding community labels through a simple two-step evidential label propagation strategy. The membership of each node is described in the form of basic belief assignments, which can well express the uncertainty included in the community structure of the graph. The experiments demonstrate the effectiveness of the proposed method on real-world networks.

Keywords: Community detection · Theory of belief functions
Density peaks · Evidential clustering

1 Introduction

Community structure is one of the primary features in graphs which can gain us a better understanding of organizations and functions in the real networked systems. As a result, community detection, which can extract specific structures from complex networks, has attracted considerable attention in many areas.

In 2014, Rodriguez and Laio have proposed a density peak clustering method (DPC) in Science [6]. It is an effective and powerful tool for the task of clustering, as neither optimization nor iteration is required in the algorithm. DPC only provides us with a hard partition of the analyzed data set. However, many real-world networks contain uncertain community structure, such as bridge nodes and outliers. Credal partitions in the framework of belief functions can give us a better understanding of the uncertain class structures of the analyzed data set.

In Ref. [7], an evidential label propagation algorithm was introduced, where only the whole frame is used to express the uncertainty of the class structure but

the partial ignorance is not considered. In this paper, an algorithm for detecting credal community structure, which can well describe both the total and partial ignorance about nodes' community, is proposed based on the concept of density peaks. Two new metrics, the local density ρ and the minimum dissimilarity δ , are first defined for each node in the graph. Then the nodes with both higher ρ and δ values can be identified as community centers. Finally, the rest of the nodes are assigned with corresponding community labels with a simple two-step evidential label propagation strategy. The experiments show that meaningful partitions of the graph could be obtained by the proposed detection approach and it indeed could provide us more informative information of the graph structure.

The remainder of this paper is organized as follows. The density peak based clustering is briefly introduced in Sect. 2. The proposed community detection approach is presented in detail in Sect. 3. Some experiments on graph data sets are conducted to show the performance in Sect. 4. Conclusions are drawn in the final section.

2 Density Peak Based Clustering

Rodriguez and Laio [6] proposed a fast clustering approach by finding density peaks, denoted by DPC. The idea is that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from any points with higher densities [6]. From this point of view, the cluster center selection problem can be converted into the problem of detecting outliers through a defined decision graph using two delicately designed measures:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \quad (1)$$

and

$$\delta_i = \begin{cases} \max_j (d_{ij}), & \text{if } \rho_i = \max_k (\rho_k) \\ \min_{j:\rho_j > \rho_i} (d_{ij}), & \text{otherwise} \end{cases} \quad (2)$$

The value ρ_i is called the local density of point i . In Eq. (1), d_{ij} is the distance between points i and j , d_c is a cut-off distance. $\chi(x)$ is an indicator function which equals to 1 when $x < 0$, and 0 otherwise.

The decision graph is then generated by taking ρ_i as x axis and δ_i as y axis. Those points with both relatively large ρ_i and δ_i , which are located in the upper right corner of the graph and far away from other points, are chosen as the centers of classes. The rest patterns can be assigned into the same cluster as its nearest neighbor of higher density in a single step.

3 Evidential Density-Based Community Detection

Inspired by the idea of density peaks, in this section we will introduce a fast evidential community detection approach based on density peaks of graphs (denoted

by EDPC). Consider the network $G(V, E)$, where $V = \{n_1, n_2, \dots, n_N\}$ is the set of N nodes, and E is the set of edges. Denote the adjacency matrix by $\mathbf{A} = (a_{ij})_{N \times N}$, where $a_{ij} = 1$ indicates that there is a direct edge between nodes n_i and n_j . Let $a_{ii} = 1$.

3.1 The Dissimilarity Between Nodes

In the task of community detection, the available information is often the adjacency matrix, representing the topological structure of the graph. The similarities or dissimilarities between nodes can be determined based on the graph structure.

In this work, the dissimilarity measure based on signaling propagation process in the network is adopted, as it can map the topological structure into N -dimensional vectors in the Euclidean space [4]. For a network with N nodes, every node is viewed as an excitable system which can send, receive, and record signals. Initially, a node is selected as the source of signal. Then the source node sends a signal to its neighbors and itself first. Afterwards, the nodes with signals can also send signals to their neighbors and themselves. After a certain T time steps, the amount distribution of signals over the nodes could be viewed as the influence of the source node on the whole network.

Naturally, compared with nodes in other communities, the nodes of the same community have more similar influence on the whole network. Therefore, dissimilarities between nodes could be obtained by calculating the differences between the amount of signals they have received.

3.2 The Density Peaks

In DPC clustering, the local density of point i describes the number of points which is very close to this pattern (with a distance to pattern i smaller than d_c). In social networks, the person who is the center of a community may have the following characteristics: she/he has relation with most of the members of the group; she/he may directly contact with other persons who also play an important role in their own communities. Therefore, the centers of communities should be such nodes that not only with high degree, but also with neighbors who also have high degree. Thus we can define the local degree of node n_i as:

$$\rho_i^{(d)} = k_i + \sum_{\{j:a_{ij}=1\}} k_j, \quad (3)$$

where k_i denotes the degree of node n_i , which can be defined as:

$$k_i = \sum_{j=1}^N a_{ij}. \quad (4)$$

In graphs, some bridge nodes which have connections with many groups may also have high degree centrality. In order to distinguish these bridge nodes

with the centers, we propose a new local density measure to consider both the dissimilarities with neighbors and the centralities:

$$\rho_i = \exp\left(-\frac{1}{k_i} \sum_{j:a_{ij}=1} d_{ij}^2\right) + \rho_i^{(d)}. \tag{5}$$

For some networks with fuzzy community structure, the local density measure and the minimum dissimilarities can be regularized to distinguish cores more accurately [5]:

$$\rho_i^* = \frac{\rho_i}{\max_i\{\rho_i\}}, \quad \delta_i^* = \frac{\delta_i}{\max_i\{\delta_i\}}. \tag{6}$$

The minimum dissimilarity of nodes defined as Eq. (2) is adopted to measure the degree of dispersion among center nodes. Similar to the idea of DPC clustering, the initial centers of the graph can be set to the nodes with high ρ_i and large δ_i . Through the 2-dimensional decision graph where one dimension is ρ_i and the other is δ_i , nodes that are located right upper in the decision graph are figured out as the centers.

3.3 Allocation of Other Nodes

Assume that the set of centers obtained in the last step is $V_c \subset V$. Thus there are c communities in the graph, and let the frame of discernment be $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$. The credal partition defined on the power set allows to gain a deeper insight into the community structure. The nodes located in the overlapping areas between communities will be grouped into some imprecise classes such as $\{\omega_1, \omega_2\}$, which indicates the indistinguishability of the membership. The outliers will be assigned to a special class O^* . We use O^* instead of Ω in order to distinguish between the total ignorance class in an open world and the imprecise class $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$ for overlapping nodes. The communities of the nodes can be determined by the label propagation process, which can be implemented as follows.

Initialization. All the center nodes are assigned with one unique community label. As there is not any uncertainty for the communities of these centers, the Bayesian categorical mass function can be adopted to describe its membership. For example, if the center node $n_i \in V_c$ is assigned to community ω_j , we can get:

$$m^i(A) = \begin{cases} 1, & \text{if } A = \{\omega_j\} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

For the rest of nodes, as there is no information about their membership at this time, the total ignorant mass function can be used to show their membership:

$$m^j(A) = \begin{cases} 1, & \text{if } A = O^* \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

One Round Expansion. In this step, the nodes sharing a direct link with only one center node will be first considered. Suppose that node n_i has only linked with center $n_j \in \omega_t$, and does not link with any other centers. Similar to the principle of the label determination process in EK-NNclus [1], the mass function of the node n_i 's membership can be constructed as:

$$m^i(A) = \begin{cases} \alpha, & \text{if } A = \{\omega_t\} \\ 1 - \alpha, & \text{if } A = O^* \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where α is the discounting parameter such that $0 \leq \alpha \leq 1$, and it can be determined by the dissimilarity between nodes n_i and n_j . If the dissimilarity between the two nodes is small, that is to say, the two nodes are very close, they are most probably in the same community. Thus α can be set as a decreasing function of d_{ij} . In this work, we suggest to use:

$$\alpha = \exp \left\{ -\gamma d_{ij}^\beta \right\}, \tag{10}$$

where parameter β can be set to be 2 as default, and γ can be set to:

$$\gamma = 1/\text{median} \left(\left\{ d_{ij}^\beta, i = 1, 2, \dots, n, j \in N_i \right\} \right). \tag{11}$$

If one node shares a direct edge with more than one center nodes, it may be located in the overlap between/among these communities. Suppose that node n_i links with centers $n_{j_1}, n_{j_2}, \dots, n_{j_t}$, and the communities of the t centers are $\omega_{j_1}, \omega_{j_2}, \dots, \omega_{j_t}$ respectively. The mass function for node n_i can be defined as:

$$m^i(A) = \begin{cases} w & \text{if } A = \{\omega_{j_1}, \omega_{j_2}, \omega_{j_t}\} \\ 1 - w, & \text{if } A = O^* \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where w should be in inverse proportion to the variation of dissimilarities between nodes n_i and the corresponding centers. If the variation is small, it indicates that there is a large amount of uncertainty for the membership of node n_i and the belief assigned to the imprecise class is large. In this paper, we use:

$$w = \exp \left\{ -\text{Var}(d_{ij_1}, \dots, d_{ij_t}) \right\}. \tag{13}$$

Diffusion in the Whole Network. The unlabeled nodes will be assigned to the existing communities based on their neighbors. The labeled nodes in the neighbors can be seen as a source of evidence. The more labeled neighbors, the more information for the node's membership. Therefore, the update order of the unlabeled nodes should be determined by labeled rate [2], which is defined as:

$$\psi_i = \frac{|N_i^L|}{|N_i|}, \tag{14}$$

where $|N_i|$ denote the number of neighbors of node n_i , and $|N_i^L|$ denote the number of labeled neighbors. The unlabeled node with highest ψ_i are first chose for assigning a community label. Suppose that node n_i is the one with highest labeled rate, the evidence provided by its $|N_i|$ neighbors are in the form of BBAs, $m_1^i, m_2^i, \dots, m_{|N_i|}^i$, the BBA for node n_i 's community membership can be obtained by combining the N_i pieces of evidence from its neighbors.

The combination process can be proceeded in two steps. The first step is to divide the BBA into different groups based on the focal element except O^* , and then to combine the BBAs in each group. As there is no conflict at all among these BBAs in the same group, we can use the Dempster's rule directly for the inner group combination. The next step is to combine the fused BBA in different groups. Each group can be regarded as a source for the outer combination. The reliability of one source is related to the proportion of BBAs in this group. The larger the number of BBAs in one group, the more reliable the source is. Then the reliability discounting factor can be defined as:

$$\alpha_k = \frac{s_k}{\sum_i s_i}, \tag{15}$$

where s_k denotes the number of BBAs in each group. The discounted BBAs in different groups are combined using the Dubois and Prade rule [3] to represent the partial ignorance. Finally, after the mass functions for all the nodes' credal membership are determined, each node can be partitioned into the community with maximal mass assignment among all the focal elements.

4 Experiments

Experiment 1. In order to show the process of EDPC algorithm clearly, in the first experiment, we will consider a small illustrative graph with 11 nodes displayed in Fig. 1-a. As can be seen from the figure, there are obviously two communities in the graph, and nodes 5 and 10 are the cores of the group, and node 11 serves as a bridge between two communities. From the decision graph in Fig. 1-b, we can see that both center nodes can be easily detected.

Table 1. The BBAs for the 11 nodes after the first round expansion.

Node	ω_1	ω_2	$\Omega = \{\omega_1, \omega_2\}$	O^*
1,2,3,4	0.6065	0	0	0.3935
5	1	0	0	0
6,7,8,9	0	0.6065	0	0.3935
10	0	1	0	0
11	0	0	0	1

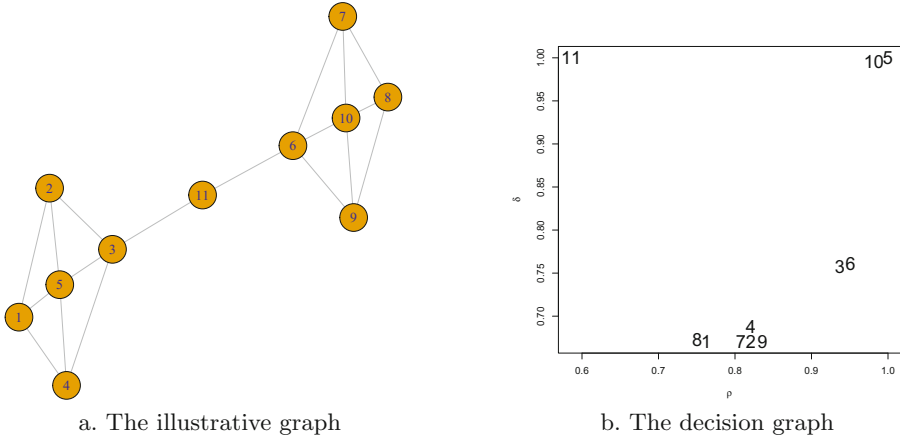


Fig. 1. An illustrative graph with 11 nodes.

In the first round expansion, according to the principle to determine the BBA, the membership for nodes 1, 2, 3, 4 and 6, 7, 8, 9 can be identified using Eq. (9). After the expansion, the BBA for 10 nodes in the graph have already been determined, which can be found in Table 1.

From this table we can see, nodes 1–4 are partitioned into the community of center node 5, while nodes 6–9 are grouped into the community of center node 10. As node 11 has no connection with both center nodes, we have not any information for its membership after the first round expansion. Thus the total ignorance mass function is still used to expression its membership.

In the diffusion process, the BBA for node 11 can be determined. The evidence for updating the membership of node 11 is from its neighbors, node 3 and node 6. Using the combination rule presented in Sect. 3, we can get the BBA for node 11 which is listed in Table 2.

As can be seen from the table, node 11 is assigned with the largest belief to imprecise class $\{\omega_1, \omega_2\}$. It reflects the indistinguishability of its membership and its bridge role between the two communities.

Table 2. The BBA for node 11 after the diffusion.

Node	ω_1	ω_2	$\Omega = \{\omega_1, \omega_2\}$	O^*
11	0.2387	0.2387	0.3678	0.1548

Experiment 2. To further test our proposed method, EDPC was applied to four real networks¹: Karate Club, American college football, Dolphin and Books about US politics, which have been widely used as test networks. Two commonly

¹ <http://www-personal.umich.edu/~mejn/netdata/>.

used community detection methods, the label propagation algorithm (LPA), the modularity-based optimization method and the median evidential c means clustering (MECM) based approach, are used for comparison. The parameters in EDPC are all set as default. The NMI values of the obtained community structure by different methods are reported in Table 3. It is noted here for EDPC, each node is partitioned into the specific community with maximal belief assignment among all the singleton focal elements. The results show EDPC performs best in most of the data sets. It is noted that MECM based community detection method also provides credal partitions. The behavior of MECM and EDPC is similar, but EDPC is more efficient as it does not require iterative optimization.

Table 3. Comparison of EDPC and other algorithms by NMI in UCI graphs.

	Karate	Football	Dolphins	Books
EDPC	1.0000	0.9346	1.0000	0.6428
MMO	0.6873	0.8550	0.4617	0.5121
LPA	0.8255	0.9095	0.8230	0.5485
MECM	1	0.9042	1	0.7977

5 Conclusion

In this paper, a novel evidential community detection approach, named EDPC, was presented inspired from the idea of density peak based clustering. The local density of each node was defined based on its centrality and the dissimilarities with its neighbors. The centers were identified according to the density and the minimum dissimilarity with the nodes with larger densities. A simple two-step evidential label propagation strategy was designed for grouping the rest of nodes. EDPC can provide us the credal community structure of the network, which enables us to gain a better insight into the graph structure. The experimental results have shown the effectiveness of the proposed method.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Nos. 61701409, 61403310, 61672431, 61135001).

References

1. Denceux, T., Kanjanatarakul, O., Sriboonchitta, S.: EK-NNclus: a clustering procedure based on the evidential k -nearest neighbor rule. *Knowl. Based Syst.* **88**, 57–69 (2015)
2. Ding, J., He, X., Yuan, J., Chen, Y., Jiang, B.: Community detection by propagating the label of center. *Phys. A Stat. Mech. Appl.* **503**, 675–686 (2018)
3. Dubois, D., Prade, H.: Representation and combination of uncertainty with belief functions and possibility measures. *Comput. Intell.* **4**(3), 244–264 (1988)

4. Hu, Y., Li, M., Zhang, P., Fan, Y., Di, Z.: Community detection by signaling on complex networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **78**(1), 139–143 (2008)
5. Li, Y., Jia, C., Yu, J.: A parameter-free community detection method based on centrality and dispersion of nodes in complex networks. *Phys. A Stat. Mech. Appl.* **438**, 321–334 (2015)
6. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. *Science* **344**(6191), 1492–1496 (2014)
7. Zhou, K., Martin, A., Pan, Q., Liu, Z.: SELP: semi-supervised evidential label propagation algorithm for graph data clustering. *Int. J. Approx. Reason.* **92**, 139–154 (2017)

Author Index

- Abdelkhalek, Raoua 1
Alshamaa, Daniel 10
Antoine, Violaine 14
- Ben Abdallah, Nadia 199
Ben Ayed, Safa 22
Bouadi, Tassadit 260
Boudaren, Mohamed El Yazid 103
Boukersoul, Houdaifa 103
Boukezzoula, Reda 129
Boukhris, Imen 1
Bronevich, Andrey G. 31
- Chebbah, Mouna 121
Cherfaoui, Véronique 95
Coquin, Didier 129
Côte, Philippe 73
Cuzzolin, Fabio 39, 48
- Debicha, Islam 103
Delmotte, François 208
Denoëux, Thierry 57, 155
Destercke, Sébastien 65
Dezert, Théo 73
Dubois, Didier 77
- Elouedi, Zied 1, 22
- Fargier, Yannick 73
Faux, Francis 77
Fundo, Akli 86
- Geng, Xiaojiao 137
Ghédira, Khaled 217
Gravouil, Kevin 14
Guyard, Romain 95
- Habbouchi, Ahmed 103
Hamache, Ali 103
Honeine, Paul 10
- Imakhlaf, Ayyoub 112
- Jendoubi, Siwar 121, 129
Jiao, Lianmeng 137
Jiroušek, Radim 146
Jousselme, Anne-Laure 199
- Kanjanatarakul, Orakanya 155
Klein, John 65
Koki, Constandina 163
Kuson, Siwarat 155
- Labroche, Nicolas 14
Lefevre, Eric 22
Leonardos, Stefanos 163
Lepskiy, Alexander 172
Liu, Liping 181
- Martin, Arnaud 121, 260, 269
Melolidakis, Costis 163
Merouani, Omar 103
Miranda, Enrique 190
Montes, Ignacio 190
Mourad-Chehade, Farah 10
- Nace, Dritan 86
- Palma Lopes, Sérgio 73
Pan, Quan 137, 269
Pichon, Frédéric 65, 199, 208
Pis Vigil, Elisa 243
Prade, Henri 77
- Ramel, Sébastien 208
Rouahi, Aouatef 217
Rozenberg, Igor N. 31
- Sadouk, Hamza 103
Salah, Kais Ben 217
Salamanca, Juan J. 226
Sallak, Mohamed 112
Shenoy, Prakash P. 146
Su, Zhi-gang 234, 252
- Terán, Pedro 243

Vicig, Paolo 190

Wang, Chenghao 86

Wang, Pei-hong 234

Wang, Qian 252

Zhang, Yiru 260

Zhao, Gang 234

Zhao, Ming 234

Zhou, Hong-yu 234

Zhou, Kuang 269

Zibani, Rezki 103