ADVANCES IN INTELLIGENT AND SOFT COMPUTING 164

Thierry Denœux
Marie-Hélène Masson (Eds.)

Belief Functions: Theory and Applications

Springer

Advances in Intelligent and Soft Computing 164

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/4240

Thierry Denœux and Marie-Hélène Masson (Eds.)

Belief Functions: Theory and Applications

Proceedings of the 2nd International Conference on Belief Functions, Compiègne, France 9–11 May 2012

Springer

Editors
Thierry Denœux
Université de Technologie de Compiègne
Heudiasyc
Compiègne
France

Marie-Hélène Masson
Université de Picardie Jules Verne
Heudiasyc
Compiègne
France

ISSN 1867-5662 e-ISSN 1867-5670
ISBN 978-3-642-29460-0 e-ISBN 978-3-642-29461-7
DOI 10.1007/978-3-642-29461-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012937041

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

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 to other frameworks such as probability, random set, possibility and imprecise probability theories.

This edited volume contains the proceedings of the 2nd International Conference on Belief Functions that was held in Compiègne, France on 9–11 May, 2012, under the auspices of the Belief Functions and Applications Society.

The book starts with an invited contribution by Prof. Hung T. Nguyen, outlining the connections between belief functions and random sets. This connection, which was first pointed out in Prof. Nguyen's seminal paper published in 1978, still has important implications today in relation to decision making, among other topics.

The remaining 50 chapters are selected peer-reviewed papers describing recent developments both on theoretical issues (including approximation methods, conflict management, combination rules, continuous belief functions, graphical models, causality and independence concepts) and applications in various areas including classification, image processing, statistics and intelligent vehicles. Overall, the large number of high quality contributions to this volume demonstrates the vitality and topicality of this research area.

The editors would like to thank Professor Janusz Kacprzyk for his kind invitation to prepare this book. We also thank all those who have contributed with their papers to this volume, as well as the program committee members for reviewing the large number of received submissions.

April 2012
Compiègne, France

Thierry Denœux
Marie-Hélène Masson

Contents

Invited Paper

Classification

Computational Issues

Conflict Management

Image Processing and Classification

Independence, Causality, Graphical Models

Information Fusion

Intelligent Vehicles

Statistics

Applications

On Belief Functions and Random Sets

Hung T. Nguyen

Dedicated to Lotfi Zadeh

Abstract. We look back at how axiomatic belief functions were viewed as distributions of random sets, and address the problem of joint belief functions in terms of *copulas*. We outline the axiomatic development of belief functions in the setting of *incidence algebras*, and some aspects of decision-making with belief functions.

1 Introduction

Just few weeks after I arrived at the University of California, Berkeley, in the late winter of 1975, Professor Lotfi Zadeh handed to me two interesting research documents. The first one is a handwritten letter of I.R. Goodman, later appeared in Goodman (1982), showing that fuzzy sets can be viewed as equivalence classes of random sets. The second one is a fresh Ph.D. thesis of G. Shafer entitled "A mathematical theory of evidence" which appeared a year later as a book (Shafer, 1976). Also, the very first book treating rigorously the theory of random sets appeared in 1975 (Matheron, 1975).

Perhaps the appearance of random sets in Goodman's letter and in Matheron's book was on my mind when I read Shafer's thesis! I reported to Professor Zadeh few weeks later that the concept of belief functions in Shafer's thesis is nothing else than the distribution function, not of a random variable or vector, but of a random set (on a finite space). Professor Zadeh clearly reacted that he was not at all happy with my remark, since being working at that time on his *theory of possibility* (for some background, see, e.g., Nguyen and Walker, 2006), which is a kind of uncertainty different than randomness modeled quantitatively by probability, he expected Shafer's concept of belief should be somewhat related to possibility, but in any case, there should be no randomness around the concept of belief. I explained that from the mathematical definition of a belief function, it is a bona fide distribution function operating on sets rather than on points, and which can be rigorously interpreted as the probability law of a random set (which is a random element), so that Shafer's theory can be placed within the standard framework of probability theory, but at the level of random sets, i.e., random elements taking sets as values.

Hung T. Nguyen
New Mexico State University (USA) and Chiang Mai University (Thailand)

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 1–19.
springerlink.com

Just few days later, Professor Zadeh walked into my office in Cory Hall and said "Could you write up what you told me the other day about Shafer's belief functions?". Seeing the surprise on my face, Professor Zadeh said "I just come back from a seminar at Stanford University where Patrick Suppes was presenting something very similar to what you told me". What Professor Zadeh referred to was a talk given by Professor Suppes, later appeared in Suppes and Zanotti (1977).

So I wrote a memorandum (an internal publication forum) for the Electronics Research Laboratory, UCB, in 1976, later appeared as "On random sets and belief functions", Nguyen (1978).

Thirty six years later, I was asked to speak about the connections between random sets and belief functions as well as their implications. In the history of science in general, and mathematics in particular, connections between two different fields should not be just formal relationships, but should provide benefits to both fields. This is exemplified by the important work of G. Hunt relating Potential Theory to Markov Processes in 1957. Here, the relation between belief functions and random sets ("the point of view is everything") does not have that grandeur. The hope was that by viewing belief functions as distributions of random sets, inference based on belief functions could gain some firm footing within probability and statistics theories. Anyway, I will speak about it, but while in 1976, the focus was on "random sets" so the title was "On random sets and belief functions", but this time, the focus is on "belief functions", so that the title of my present lecture is the other way around, namely "On belief functions and random sets"!

After all these years, both the theories of belief functions and random sets have gone a long way, in theoretical developments as well as in applications. In this lecture, I restrict myself to only one topic, namely, decision-making with belief functions, where I think the connections with random sets are somewhat significant.

2 Belief Functions

Let's start out with the standard framework from which the concept of belief functions was introduced (Dempster,1967; Shafer 1976). A "true state of nature" u_o is known to be in some *finite* set U, although it is not known which element of U is that true state. For each subset $A \subseteq U$, we express our "belief" that A contains u_o by a number, denoted as $F(A)$. Such a number $F(A)$ could come from some "evidence". We are talking about modeling/quantifying information provided by evidence, i.e., some mathematical theory of evidence. Here we are talking about information of localization.

A belief function on a finite set U is a set-function $F : 2^U \to [0,1]$ such that

(i) $F(\varnothing) = 0, F(U) = 1$
(ii) For any $n \geq 2$, and any $A_1, A_2, ..., A_n \in 2^U$

$$F(\cup_{i=1}^{n} A_i) \geq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} F(\cap_{j \in I} A_j)$$

where $|I|$ denotes the cardinality of the set I.

Remarks

a) The interpretation of the concept of belief has been widely discussed in the literature.

b) When $F(\cup_{i=1}^{n} A_i) \geq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} F(\cap_{j)\in I} A_j)$, for a given n, we say that F is monotone of order n. The property (ii) is referred to as monotonicity of infinite order. It is a weakening of the Poincaré equality of probability measures.

c) Note that, here we take the range of F to be $[0,1]$ and $F(\varnothing) = 0$ which is the minimum value of F. As such, $F(.)$ is monotone, i.e., $A \subseteq B \Longrightarrow F(A) \leq F(B)$. If $F : 2^U \to \mathbb{R}$, then monotonicity of F should be added to the set of axioms, unless $F(\varnothing)$ is the minimum value of F. In fact, if F is monotone of order 2, then it is monotone if and only if $F(\varnothing)$ is the minimum value of $F(.)$. Indeed, if $F(.)$ is monotone, then clearly $F(\varnothing)$ is its minimum value. Conversely, suppose that $F(\varnothing)$ is its minimum value, for $A \subseteq B$, we write $B = A \cup (B \backslash A)$. By $2-$ monotonicity, we have $F(B) = F(A \cup (B \backslash A)) \geq F(A) + F((B \backslash A)) - F(\varnothing)$. But, by hypothesis, $F(\varnothing) \leq F((B \backslash A))$, we have $F(B) \geq F(A)$.

Example 1 (belief functions induced by probabilities)

Suppose the true probability measure $P_o : 2^U \to [0,1]$ is only known by its values on a partition $\Theta_1, \Theta_2, \ldots, \Theta_k$ of the finite set U, say, $P_o(\Theta_i) = \alpha_i$, $i = 1,2,\ldots,k$. What is the uncertainty of an arbitrary $A \subseteq U$? Suppose we quantify the uncertainty of any A by

$$F(A) = \inf\{P(A) : P \in \mathscr{P}\}$$

where $\mathscr{P}$ denotes the set of all probability measures P on U satisfying the constraints $P(\Theta_i) = \alpha_i$, $i = 1,2,\ldots,k$. Then, formally, $F(.)$ is a belief function. Indeed, clearly $F(\varnothing) = 0$ and $F(U) = 1$, by construction. For $A \in 2^U$, we approximate it by $A^* = \cup_{\Theta_i \subseteq A} \Theta_i$. Clearly, $P(A^*)$ is the same for any $P \in \mathscr{P}$, so that we can consider the function $G : 2^U \to [0,1]$, $G(A) = P(A^*)$ for any $P \in \mathscr{P}$. Now, $G(\varnothing) = 0$, $G(U) = 1$, and infinitely monotone:

$$G(\cup_{i=1}^{n} A_i) = P(\cup_{i=1}^{n} A_i)^* \geq P(\cup_{i=1}^{n} A_i^*) = \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} P(\cap_{j \in I} A_j^*)$$

$$\sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} P((\cap_{j \in I} A_j)^*) = \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} G(\cap_{j \in I} A_j)$$

since $\cup(A_i)^* \subseteq (\cup A_i)^*$ and $\cap_{j \in I} A_j^* = (\cap_{j \in I} A_j)^*$.

Now, for any $A \in 2^U$, there is a $P_A \in \mathscr{P}$ such that $P_A(A) = P_A(A^*)$ and hence $G(.) = F(.)$. Note also that $\mathscr{P} = \{P : G \leq P\}$.

Example 2 (belief functions as distributions of random sets)

Let $(\Omega, \mathscr{A}, P)$ be a probability space, and $(V, \mathscr{V})$ be an arbitrary measurable space. A map $X : \Omega \to V$ is called a random element if $X^{-1}(\mathscr{V}) \subseteq \mathscr{A}$, and its probability law is the probability measure $P_X = PX^{-1}$ on $\mathscr{V}$. For U a finite set, and $V = 2^U$, $\mathscr{V}$ being the power set of 2^U, $X : \Omega \to 2^U$ is called a *non empty* random set whose probability law is completely determined by its *distribution function* $F : 2^U \to [0,1]$, defined by

$$F(A) = P(X \subseteq A)$$

Now, clearly $F(\varnothing) = 0$ and $F(U) = 1$. Moreover, $F(.)$ is infinitely monotone. Indeed, for $B \in 2^U$, and $A_i \in 2^U$, $i = 1, 2, ..., n$, let

$$J(B) = \{i : \quad \text{such that } \ B \subseteq A_i\}$$

We have

$$F(\cup_{i=1}^n A_i) = \sum_{B \subseteq \cup_{i=1}^n A_i} F(B) \geq \sum_{B \subseteq U, J(B) \neq \varnothing} F(B)$$

Now observe that, when $J(B) \neq \varnothing$, $\sum_{B \subseteq J(B)} (-1)^{|B|+1} = 1$, we can write

$$\sum_{B \subseteq U, J(B) \neq \varnothing} F(B) = \sum_{B \subseteq U, J(B) \neq \varnothing} [\sum_{\varnothing \neq I \subseteq J(B)} (-1)^{|I|+1}] F(B)$$

$$= \sum_{\varnothing \neq I \subseteq J(B)} (-1)^{|I|+1} \sum_{B \subseteq U, I \subseteq J(B)} F(B)$$

$$= \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} \sum_{B \subseteq \cap_{j \in I} A_j} F(B) = \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} F(\cap_{j \in I} A_j)$$

As expected, as in the case of random vectors, the properties of belief functions can be used as axioms for distribution functions of (finite) random sets: if F is a belief function on a finite set U, then it must be the distribution of some non empty random set, i.e., there exist a probability space $(\Omega, \mathscr{A}, P)$ and a non empty random set $X : \Omega \to 2^U$ such that $F(A) = P(X \subseteq A)$. For that, it suffices to show that there exists a function $f : 2^U \to [0,1]$ with $\sum_{A \subseteq U} f(A) = 1$ (called the *density function* of the random set X) such that $F(A) = \sum_{B \subseteq A} f(B)$.

For that purpose, define

$$f(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} F(B)$$

where $A \setminus B = A \cap B^c$.

$f(.)$ is nonnegative, indeed, $f(\varnothing) = F(\varnothing) = 0$, and by construction, $f(\{u\}) = F(\{u\}) \geq 0$. For $A \in 2^U$ with $|A| \geq 2$, say, $A = \{u_1, u_2, ..., u_k\}$, let $A_i = A \setminus \{u_i\}$, $i = 1, 2, ..., k$. Then,

$$f(A) = F(A) - \sum_{i=1}^{k} F(A_i) + \sum_{i<j} F(A_i \cap A_j) + \ldots + (-1)^{k-1} \sum_{i=1}^{k} F(\cap_{j \neq i} A_j) \geq 0$$

by infinite monotonicity of F, noting that $\cap_{i=1}^{k} A_i = \varnothing$, and $A = \cup_{i=1}^{k} A_i$.

Next,

$$\sum_{B \subseteq A} f(B) = \sum_{B \subseteq A} \sum_{C \subseteq B} (-1)^{|B \setminus C|} F(C) = \sum_{C \subseteq B \subseteq A} (-1)^{|B \setminus C|} F(C)$$

If $C = A$, the last term is $F(A)$. If $C \neq A$, then $A \setminus C$ has $2^{|A \setminus C|}$ subsets, so there are an even number of subsets B with $C \subseteq B \subseteq A$, exactly half of which have an even number of elements. The half of the numbers $(-1)^{|B \setminus C|}$ are 1 and half are -1. Thus, for each $C \neq A$, we have

$$\sum_{C \subseteq B \subseteq A} (-1)^{|B \setminus C|} F(C) = 0$$

with the summation taken over B. Hence, $\sum_{B \subseteq A} f(B) = F(A)$. In particular,

$$1 = F(U) = \sum_{B \subseteq U} f(B)$$

Remarks

a) In the context of finite random sets,

$$f(A) = P(X = A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} F(B)$$

which is just a fact from combinatorial theory (Möbius transforms). For a comprehensive study of belief functions as set-functions, we outline, in the next section the setting of *incidence algebras.*

b) The probability law of a random set X on finite U can be also characterized by a dual concept of distribution function, namely, *capacity functional*: $T : 2^U \to [0,1]$,

$$T(A) = P(X \cap A \neq \varnothing) = 1 - F(A^c)$$

While $T(\varnothing) = 0$ and $T(U) = 1$, T is *alternating of infinite order*, i.e.,

$$T(\cap_{i=1}^{n} A_i) \leq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} T(\cup_{j \in I} A_j)$$

Set functions $\varphi : 2^U \to \mathbb{R}$ which are *maxitive,* i.e.,

$$\varphi(A \cup B) = \max\{\varphi(A), \varphi(B)\}$$

are alternating of infinite order. For general random sets, the concept of capacity functionals is more convenient to work with. Capacity functionals play the role of distribution functions of random vectors, characterizing probability laws of random

sets via the well-known Choquet theorem (counterpart of the classical Lebesgue-Stieltjes theorem characterizing probability measures on Euclidean spaces by distribution functions).

c) A quantitative concept of degrees of belief can be also justified in the context of *coarse data* in statistics. Coarse data are data with low quality, due, e.g., to its imprecision. If $Y : \Omega \to U$ is a random variable, then a *coarsening* of Y is a non empty random set $X : \Omega \to 2^U$ such that $P(Y \in X) = 1$, i.e., Y is an almost sure selector of X. Without observing the (latent) random variable of interest Y, we rely on the observable X to conduct statistical inference. The distribution F of X is a belief function, and any possible probability law Q of Y on U should be compatible with F, i.e., $Q \geq F$, i.e., is in the *core* of F, namely $\mathscr{C}(F) = \{Q : Q \geq F\}$. This is so, since, for any $A \in 2^U$, $\{\omega \in \Omega : X(\omega) \subseteq A\} \subseteq \{\omega \in \Omega : Y(\omega) \in A\}$ and hence

$$F(A) = P(X \subseteq A) \leq P(Y \in A) = PY^{-1}(A)$$

It is interesting to note that humans often use *coarsening schemes* in decision-making, a fact which can be attributed to "intelligence". I have once "argued" with Professor Zadeh that fuzziness in perception information appears as consequences of using fuzzy coarsening schemes, i.e., fuzzy partitions, in order to make decisions. When facing a decision, or a question, with not enough information to act, humans use a coarsening of the domain, such as in if...then rules in fuzzy control (see, e.g., Nguyen and Walker, 2006).

d) In current literature, when referring to belief functions, you see typical statements as follows. Let Θ be a parameter space in a statistical model. The Bayesian probability theory treats the parameter θ as a random variable with a prior probability distribution over Θ. In the Dempster-Shafer belief functions theory, information about the unknown true value of the parameter is described by the probability distribution of a non empty random set on Θ.

The theory of belief functions was introduced as a mathematical theory of evidence. Since in a given problem, there might exist several sources of evidence, each represented by a belief function, there is a need to combine them. In random set language, if we have two random sets X and Y on the same finite set U, then the random set $X \cap Y$ is a natural candidate for a combined evidence. Clearly, the distribution of $X \cap Y$ depends on the joint distribution of (X,Y). From the knowledge of the marginal distributions, say, F_X, F_Y, we seek some possible joint distribution H for (X,Y). This sounds like an old problem of Maurice Fréchet! If X and Y are random *vectors*, then the problem is solved by A. Sklar (1959) via the concept of *copulae* (see also, Nelsen (1999)). However, here, not only we are facing discrete variables, but also these variables are not random vectors, they are random sets. The extension of Sklar's work from random vectors to multivariate random sets (random sets in n dimensions) is *an open problem.* Some efforts in extending Sklar's results from distribution functions on euclidean spaces to more general infinitely dimensional Polish spaces (for probability measures and Choquet capacities) have been partially carried out by Scarsini (1989, 1996), but the specific case of random sets, finite or not (especially *random closed sets* on Hausdorff, locally compact spaces) has not been touched upon.

While X and Y are non empty random sets, $X \cap Y$ might not be a non empty random set, i.e., $P(X \cap Y = \varnothing) > 0$. However, the "conditional random set" $X \cap Y \mid (X \cap Y \neq \varnothing)$ is a non empty random set. Indeed, its density is

$$\psi(A) = P(X \cap Y = A | X \cap Y \neq \varnothing) = \frac{P(X \cap Y = A, X \cap Y \neq \varnothing)}{P(X \cap Y \neq \varnothing)}$$

from which we see that $\psi(\varnothing) = 0$ (since then $(X \cap Y = \varnothing, X \cap Y \neq \varnothing) = \varnothing$).

The approach to combination of evidence, known as the Dempster's rule of combination, assumes in addition that the random sets X and Y are independent, i.e., $P(X = A, Y = B) = P(X = A)P(Y = B)$, for any A, B in 2^U, so that $\psi(.)$ is reduced to, for $A \neq \varnothing$,

$$\psi(A) = \frac{\sum_{C \cap D = A} P(X = C)P(Y = D)}{\sum_{S \cap T = \varnothing} P(X = S)P(Y = T)}$$

We refer the reader to discussions concerning the independence assumption of X, Y in this rule of combination and its incompatibility with the condition $X \cap Y \neq \varnothing$. But if we drop the independence assumption on X, Y, then we face Frechet's problem: how to specify a joint distribution from its marginals? (here in the context of random sets!). Using *maximum entropy principle?* Then we need to consider the concept of *entropy of random sets.* See Section 4 below.

Example 3 (belief functions on arbitrary sets)

The definition of belief functions on *finite* sets can be kept for arbitrary sets. Let U be an arbitrary set (finite or not). Since a priori, there is nothing random around (!), we can "commit" our belief to any subsets of U (subjective assignments of degrees of belief to subsets of U are possible from an intuitive viewpoint), so that, $F : 2^U \to [0,1]$ as in the finite case. Here is an example.

Let $(U, \mathscr{U}, P)$ be a probability space. P induces a belief function $F : 2^U \to [0,1]$ as follows. Define

$$F(A) = \sup\{P(B) : B \in \mathscr{U}, B \subseteq A\}$$

Clearly, $F(\varnothing) = 0$ and $F(U) = 1$. The fact that F is infinitely monotone can be seen as follows. First, the above sup is attained, i.e., for each $A \in 2^U$, there is an $B \in \mathscr{U}$ such that $B \subseteq A$ and $F(A) = P(B)$. Indeed, for each positive integer n, let $B_n \subseteq A$ with $B_n \in \mathscr{U}$ and $P(B_n) \geq P(A) - \frac{1}{n}$. It follows readily that $B = \cup_n B_n \in \mathscr{U}$, $B \subseteq A$ and $F(A) = F(B)$. Such B is called a *measurable kernel* of A. Now, observe that if B_i is a measurable kernel of A_i, then $\cap_i B_i$ is a measurable kernel of $\cap_i A_i$. Thus,

$$F(\cup_{i=1}^n A_i) \geq P(\cup_{i=1}^n A_i) = \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} P(\cap_{i \in I} A_i)$$

$$= \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} P(\cap_{i \in I} B_i) = \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} F(\cap_{i \in I} B_i)$$

Remarks

a) For relations between belief functions and general random sets, see Matheron (1975), Wasserman (1987), Molchanov (2005), and Nguyen (2006).

b) When U is infinite, a question of independent interest is whether there exist Möbius tranforms (in the non locally finite case)? Marinacci (1996) gave a formulation of a such counterpart.

Example 4 (Belief functions and possibility measures)

We discuss now some relations between belief functions and Zadeh's *possibility measures*. Let U be a nonempty set. A set function $\pi : 2^U \to [0,1]$ is called a possibility measure on U when it satisfies

(i) $\pi(\varnothing) = 0$, $\pi(U) = 1$
(ii) For any collection of subsets $S \subseteq 2^U$, $\pi(\cup_{A\in S} A) = \sup\{\pi(A) : A \in S\}$

The dual of π is defined as $\pi^*(A) = 1 - \pi(A^c)$. Then $\pi^*(\varnothing) = 0$, $\pi^*(U) = 1$. Clearly, π^* satisfies: $\pi^*(\cap_{A\in S} A) = \inf\{\pi^*(A) : A \in S\}$, and hence monotone of infinite order, i.e., the dual π^* of a possibility measure is a belief function of a special type. The special type of belief functions F satisfying $F(A \cap B) = \min\{F(A), F(B)\}$ can be characterized as follows, a result due to Dubois and Prade (1986), where we rephrase as: A belief function F on a finite U satisfies $F(A \cap B) = \min\{F(A), F(B)\}$ if and only if the support of $F * \mu$ is a chain, i.e., if and only if the support of the Möbius inverse of F is a chain (see a proof in the next section).

3 Belief Functions and Incidence Algebras

In the above section we simply recall the definition of a belief function and provide some typical examples of belief functions. We outline now an axiomatic development of belief functions in a convenient setting, and develop some methods for computations with them.

Let U be a finite set. We will study belief functions in the context of the set $\mathscr{F}$ of all functions from 2^U to the real line $\mathbb{R}$. Our goal is to establish some mathematical facts of use in the study and application of various kinds of set-functions in reasoning under uncertainty.

Let $\mathscr{F} = \{f : 2^U \to \mathbb{R}\}$. With addition and scalar ($\mathbb{R}$) multiplication pointwise, $\mathscr{F}$ is a vector space over $\mathbb{R}$:

$$(f+g)(A) = f(A) + g(A), \quad (rf)(A) = rf(A)$$

One basis of $\mathscr{F}$ is

$$\{f_A : A \subseteq U : f_A(A) = 1, f_A(B) = 0 \text{ for } B \neq A\}$$

so that $\mathscr{F}$ has dimension $2^{|U|}$.

Let $\mathscr{I}$ be the set of functions $\alpha : \{(A,B) : A \subseteq B \subseteq U\} \to \mathbb{R}$. Addition on $\mathscr{I}$ is defined pointwise, and multiplication is defined as a *convolution*

$$(\alpha * \beta)(A,B) = \sum_{A \subseteq C \subseteq B} \alpha(A,C)\beta(C,B)$$

$\mathscr{I}$ is an algebra called the *incidence algebra* over the field $\mathbb{R}$ in combinatorial theory (from the work of Gian-Carlo Rota beginning in 1964, for locally finite posets). Here, we are in the simple case of the poset of subsets of a finite set U, with set inclusion as partial order relation.

The following facts are useful.

(i) The incidence algebra $\mathscr{I}$ is a ring with identity which is

$$\delta(A,B) = \begin{cases} 1 \text{ if } A = B \\ 0 \text{ if } A \subset B \end{cases}$$

Remark

Thus, $\mathscr{I}$, with addition $+$ is an Abelian group. However, while the ring $\mathscr{I}$ has an identity, not every nonzero element has an inverse.

(ii) An element $\alpha \in \mathscr{I}$ has an inverse if and only if for all $A \in 2^U$, $\alpha(A,A) \neq 0$. In this case, the inverse of α is given by

$$\alpha^{-1}(A,B) = \begin{cases} \frac{1}{\alpha(A,A)} \text{ if } A = B \\ \frac{-1}{\alpha(A,A)} \sum_{A \subseteq C \subseteq B} \alpha(A,C)\alpha^{-1}(C,B) \text{ if } A \subset B \end{cases}$$

Proof

a) *Necessity.* If α has β as inverse, then $(\alpha * \beta)(A,A) = \alpha(A,A)\beta(A,A) = \delta(A,A) = 1$, implying that $\alpha(A,A) \neq 0$.

b) *Sufficiency.* Suppose for all $A \in 2^U$, $\alpha(A,A) \neq 0$. We seek $\beta \in \mathscr{I}$ such that $\alpha * \beta = \beta * \alpha = \delta$. Define $\beta(A,B)$ inductively on the number of subsets C between A and B, donoted as $\#(A,B)$. If $\#(A,B) = 1$ (i.e., when $A = B$), we let $\beta(A,A) = \frac{1}{\alpha(A,A)}$. Assume $\beta(A,D)$ has been defined for A,D for which $\#(A,D) < n$. Then, for A,B with $\#(A,D) = n$ (> 1), we want

$$0 = (\alpha * \beta)(A,B) = \sum_{A \subseteq C \subseteq B} \alpha(A,C)\beta(C,B)$$

$$= \alpha(A,A)\beta(A,B) + \sum_{A \subset C \subseteq B} \alpha(A,C)\beta(C,B)$$

which can be solved for $\beta(A,B)$ since $\alpha(A,A) \neq 0$, yielding

$$\beta(A,B) = \frac{-1}{\alpha(A,A)} \sum_{A \subseteq C \subseteq B} \alpha(A,C)\alpha^{-1}(C,B)$$

Similarly, there is $\gamma \in \mathscr{I}$ such that $\gamma * \alpha = \delta$, and hence,

$$(\gamma * \alpha) * \beta = \delta * \beta = \beta = \gamma * (\alpha * \beta) = \gamma * \delta = \gamma$$

For example, elements of the ring $\mathscr{I}$ which have an inverse are called units, e.g., the Möbius function $\mu(A,B) = (-1)^{(B\setminus A)}$, and the Zeta function $\zeta(A,B) = 1$. Note that, convolution with δ is analogous to integration, whereas "multiplying" by μ is analogous to differentiation (resulting what is called the Möbius inversion).

(iii) The Möbius and Zeta functions are inverses of each other.

(iv) There is a natural operation on the elements of the vector space $\mathscr{F}$ by the elements of the incidence algebra $\mathscr{I}$ which is common in combinatorics and will simplify some of the computations with belief functions. For $f \in \mathscr{F}$ and $\alpha \in \mathscr{I}$, define, for each $A \in 2^U$,

$$(f * \alpha)(A) = \sum_{B \subseteq A} f(B)\alpha(B,A)$$

With this operation, $\mathscr{F}$ is a right module over $\mathscr{I}$. Note that, $(f * \alpha) \in \mathscr{F}$. The maps $f \in \mathscr{F} \to \mathscr{F} : f \to f * \mu$, and $f \to f * \zeta$ are one to one maps and are inverses of one another. The set-function $f * \mu$ is referred to as the Möbius inverse of f.

We focus now on belief functions in this setting. A density on 2^U is a function $f : 2^U \to [0,1]$ such that $\sum_{A \subseteq U} f(A) = 1$. Then, it can be checked that $(f * \zeta)(\varnothing) = 0$, $(f * \zeta)(U) = 1$, and $f * \zeta$ is monotone of infinite order (which is equivalent to $f(A) \geq 0$ for $|A| \geq 2$).

The precise correspondence between belief functions and densities is this.

g is a belief function if and only if $g * \mu$ is a density such that $(g * \mu)(\varnothing) = 0$.

Proof

If $g * \mu$ is a density such that $(g * \mu)(\varnothing) = 0$, then $(g * \mu) * \zeta = g$ is a belief function. Conversely, if g is a belief function, then

$$\sum_{A \subseteq U} (g * \mu)(A) = ((g * \mu) * \zeta)(U) = g(U) = 1$$

It remains to check that $g * \mu \geq 0$.

$$(g * \mu)(\varnothing) = g(\varnothing)\mu(\varnothing,\varnothing) = g(\varnothing) = 0$$

For $A = \{u\}$,

$$(g * \mu)(\{u\}) = g(\varnothing)\mu(\varnothing,\{u\}) + g(\{u\})\mu(\{u\},\{u\}) = g(\{u\}) \geq 0$$

Finally, since g is monotone of infinite order, $(g * \mu)(A) \geq 0$ for $|A| \geq 2$.

As a consequence, there is a one to one correspondence between densities with value 0 at $\varnothing$ and belief functions $(f \to f * \zeta)$ with inverse μ $(g \to g * \mu)$.

There is a natural way to construct densities on U from a density f on 2^U with $f(\varnothing) = 0$. A function $\tau : U \times 2^U \to [0,1]$ is called an *allocation* of f if $\sum_{u \in A} \tau(u,A) = f(A)$ for all $A \in 2^U$. Clearly, the function $u \in U \to \sum_{\{A : u \in A\}} \tau(u,A)$ is a density on U.

Examples

(i) Let g be a belief function on U and $f = g * \mu$ its Möbius inverse. For $A \neq \varnothing$, let $\tau(u,A) = \frac{f(A)}{|A|}$ for $u \in A$. Then τ is an allocation of f.

(ii) For $|U| = n$, let $u_1, u_2, ..., u_n$ be an ordering of U. Let f be a density on 2^U. Consider the allocation τ of f such that $\tau(u,A) = f(A)$ if u is the largest element in A, and zero otherwise. The associated density on U is described in terms of g as

$$f_\tau(u_i) = g(\{u_1, u_2, ..., u_i\}) - g(\{u_1, u_2, ..., u_{i-1}\})$$

for $i = 1, 2, ..., n$ (noting that, for $i = 1$, $g(\{u_1, u_2, ..., u_{i-1}\}) = g(\varnothing) = 0$). The density f_τ on U gives rise to the probability measure Q_τ on 2^U. Since there are $n!$ orderings of U, and hence $n!$ such probability measures. The average of the densities on U obtained this way is the Shapley value in coalition games.

The *core of a belief function* g on U, denoted as $\mathscr{C}(g)$, is the set of probability measures Q on $(U, 2^U)$ such that $g \leq Q$. If g is is a probability measure, then $\mathscr{C}(g) = \{g\}$. If $g(A) = 0$ for $A \neq U$, and $g(U) = 1$, then $\mathscr{C}(g)$ is the set of all probability measures on 2^U.

We are going to show that $g = \inf\{Q : Q \in \mathscr{C}(g)\}$.

Let $f = g * \mu$. Let τ be an allocation of a density f on 2^U, and let Q_τ be the probability measure on 2^U induced by τ, i.e., $Q_\tau(A) = \sum_{u \in A} \sum_{\{B : u \in B\}} \tau(u,B)$. We have

$$g(A) = (f * \zeta)(A) = \sum_{B \subseteq A} f(B) = \sum_{\{B : B \subseteq A\}} \sum_{\{u : u \in B\}} \tau(u,B) \leq Q_\tau(A)$$

Next, for each $A \in 2^U$, let τ_A be an allocation of f such that for $u \in A$, and for all B not contained in A, allocate 0 to u. Then, $g(A) = Q_A(A)$, where $Q_A(.)$ is the probability measure induced by τ_A. It follows that $g(A) = \inf\{Q(A) : Q \in \mathscr{C}(g)\}$.

From the above, we see that if τ is an allocation of the density (on 2^U) $f = g * \mu$, then $Q_\tau \in \mathscr{C}(g)$. An elementary proof of the converse seems difficult to find: if $Q \in \mathscr{C}(g)$, then $Q = Q_\tau$ for some τ. However, it can be obtained by the following considerations. Write $U = \{u_1, u_2, ..., u_n\}$. We identify $\mathscr{C}(g)$ with the subset of the simplex $S = \{(x_1, x_2, ..., x_n) \in [0,1]^n : \sum_{i=1}^n x_i = 1\}$ consisting of the n-tuple $(x_1, x_2, ..., x_n)$ corresponding to $x_i = Q(\{u_i\})$ for $Q \in \mathscr{C}(g)$. It can be shown that, with this identification, $\mathscr{C}(g)$ is a closed, convex subset of S whose extreme points are those $n!$ densities in example (ii) above. Elements of $\mathscr{C}(g)$ are convex combinations of its extreme points.

We close this section with the following information.

Let g be a belief function on a finite set U. Then $g(A \cap B) = \min\{g(A), g(B)\}$ if and only if the support of $g * \mu$ is a chain.

Proof

a) *Necessity.* Suppose $g(A \cap B) = \min\{g(A), g(B)\}$ for any A, B. Let A, B be such that $A \subsetneq B$ and $B \subsetneq A$ (so that $A \cap B \neq A$), and $g(A \cap B) = \min\{g(A), g(B)\} = g(A)$. Then

$$g(A\cap B)=\sum_{D\subseteq A\cap B}(g*\mu)(D)=g(A)=\sum_{D\subseteq A}(g*\mu)(D)$$

implying that $(g*\mu)(A)=0$. Thus, if $(g*\mu)(A)\neq 0$ and $(g*\mu)(B)\neq 0$, then either $A\subseteq B$ or $B\subseteq A$, i.e., the support of $g*\mu$ is a chain.

b) *Sufficiency.* Suppose the support of $g*\mu$ is a chain.

From

$$g(A)=\sum_{D\subseteq A}(g*\mu)(D),\quad g(B)=\sum_{D\subseteq B}(g*\mu)(D)$$

we see that those D above such that $(g*\mu)(D)\neq 0$ must all be contained in A or all contained in B. Thus, $g(A\cap B)=\min\{g(A),g(B)\}$ for any A,B.

4 Decision-Making with Belief Functions

In the context of incomplete information, a standard framework for decision-making is this. Consider decision problems consisting of choosing an action in a set $\mathbb{A}$ to maximize some utility function $u:\mathbb{A}\times U\to\mathbb{R}$. When the probabilistic information on the set of "states" of nature U is given by a probability Q on it, then $E_Q u(a,.)$ is used as a criterion. We address this decision-making problem when the probabilistic information is in a weaker form, namely, it is given by a belief function (the knowledge about U is supported by some evidence whose mathematical representation is a belief function on it).

4.1 Entropy

Entropy is a mysterious term in Thermodynamics used by L. Boltzmann (Vorlesungen uber Gastheorie, 2 vol. Leipzig, 1895-1898):

$$H=-\int\int\int f(u,v,w)\log f(u,v,w)dudvdw$$

to define the entropy of a gas, when the velocities of molecules are distributed according to a probability density f.

Nowadays, thanks to N. Wiener and Cl. Shannon's works on information theory, we know its meaning! "The uncertainty on the state of a monoatomic gas, when it is maximal, is equal to the entropy of that gas, computed by elementary methods of Thermodynamics; this maximum of uncertainty corresponds precisely to the statistical equlibrium of the gas, where the distribution of the velocities is the distribution of J.C. Maxwell (1859)" (from *Théorie de l'Information*, Cours de Monsieur Joseph Kampé de Feriet, Publ. Laboratoire de Calcul, Univ. de Lille, France, 1961-1962).

Specifically, entropy is used to designate a *macro aspect of uncertainty* of a stochastic system. If X is a random element (e.g., a random variable, or a random set) taking values in a *finite* set V ($V=U=\{u_1,u_2,...,u_n\}$ for random variables; $V=2^U$ for random sets), with probability density f (on U for random variables; on 2^U, for random sets), then the (macro) uncertainty about X is taken to be

$$H(X) = H(f) = -\sum_{v \in V} f(v) \log f(v)$$

Note that, like Shannon, we use the letter H to denote uncertainty, since it was the letter used by Boltzmann, to remind us of the relation with entropy in Thermodynamics, although, the concept of uncertainty here came from probability theory, and has nothing to do with Thermodynamics!

Since the function $x \in [0,1] \to h(x) = x \log x$ is convex (since $h''(x) = \frac{1}{x}$, noting that we set $x \log x = 0$ when $x = 0$, since $\lim_{x \to 0}(x \log x) = 0$), $0 \leq H(f) \leq \log n$. Now, $\log n = -n[\frac{1}{n} \log \frac{1}{n}]$, the uncertainty (entropy) of X is maximum when f is the uniform density. This corresponds to Laplace's insufficiency principle: if there is no additional information about the distribution of a random element, then the states $v \in V$ should be treated equally, so that the uniform distribution (which has maximum entropy) is plausible to use.

The analogue of entropy for continuous variables (not derivable from the discrete case through a limiting process) is

$$H(f) = -\int f(x) \log f(x) dx$$

For example, if $f(x) = \frac{1}{b-a} 1_{[a,b]}(x)$, then $H(f) = \log(b-a)$ which could be negative if $b - a < 1$.

Consider the maximization problem

$$\max H(f) = \max[-\int f(x) \log f(x) dx]$$

over $f \in \mathscr{F}$. It is well-known that, using the calculus of variation,

a) if $\mathscr{F}$ consists of all probability densities $f : \mathbb{R} \to \mathbb{R}^+$ with support $[a,b]$, then the above maximum is attained at the uniform density on $[a,b]$,

b) if $\mathscr{F}$ consists of distributions with mean μ and variance σ^2, then $N(\mu, \sigma^2)$ has maximum entropy.

In view of the above, and just like the Maximum Likelihood principle, E.T. Jaynes avocated, in 1957, the *Maximum Entropy Principle* (*Maxent)* as a statistical inference principle: Subject to known constraints, the probability distribution which best represents the current state of knowledge is the one with maximum entropy. See Jaynes, (1957).

Remark on principles

MLE and Maxent are among commonly used principles of inference. Another principle related to random set data is Hartigan's *excess mass* (Hartigan, 1987).

Let X be a random vector with values in $\mathbb{R}^d$ with unknown density f. Having only some appropriate analytic properties of f, we estimate f (pointwise) nonparametrically, using, say, a random sample $X_1, X_2, ..., X_n$ drawn from X. When the dimension d is high, conventional methods such as kernel and orthogonal functions seem inefficient. For each $\alpha > 0$, the $\alpha-$ level set of f is

$$A_\alpha = \{x \in \mathbb{R}^d : f(x) \geq \alpha\}$$

Since,

$$f(x) = \int_0^\infty 1_{A_\alpha}(x) d\alpha$$

where $1_{A_\alpha}(.)$ is the indicator function of the set A_α, we could first estimate the *sets* A_α, say by a *random set estimator* $A_{\alpha,n}(X_1, X_2, ..., X_n)$, then use the plug-in estimator

$$f_n(x, X_1, X_2, ..., X_n) = \int_0^\infty 1_{A_{\alpha,n}(X_1,X_2,...,X_n)}(x) d\alpha$$

to estimate $f(x)$.

The question is: which $A_{\alpha,n}(X_1, X_2, ..., X_n)$ to use? Hartigan proposed a method in which the concept of likelihood is replaced by the concept of excess mass.

Let dF, $\lambda(dx)$ denote the probability law of X and the Lebesgue measure on the Borel $\sigma-$field $\mathscr{B}(\mathbb{R}^d)$, respectively. Then, $(dF - \alpha\lambda)(A_\alpha)$ is the "excess mass" of the level set A_α. If we consider the signed measure $\varepsilon_\alpha(.) = (dF - \alpha\lambda)(.)$ on $\mathscr{B}(\mathbb{R}^d)$, then, for any $A \in \mathscr{B}(\mathbb{R}^d)$, and $\alpha > 0$, we have $\varepsilon_\alpha(A) \leq \varepsilon_\alpha(A_\alpha)$. Thus, the excess mass principle is this. A_α has the largest excess mass at level α. This suggests to estimate A_α by using the empirical counterpart of the signed measure $\varepsilon_\alpha(.)$, namely $\varepsilon_{\alpha,n}(.) = (dF_n - \alpha\lambda)(.)$ where $dF_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, with δ_{X_i} being the Dirac (random) probability measure at X_i. Specifically, $A_{\alpha,n}(X_1, X_2, ..., X_n)$ is taken to be the solution of the maximization problem of the (set-function) objective function $\varepsilon_{\alpha,n}(A)$ over $A \in \mathscr{B}(\mathbb{R}^d)$. It is easier said than done! This is an optimization problem, neither for vector nor for functions, but for sets. For a suggested variational calculus of set-functions, see Nguyen and Kreinovich (1999).

The entropy of a random set X describes the uncertainty about X, and hence can be used to "compare" different sources of evidence. Yager (1983) defined entropy of a belief function, slightly different than Nguyen (1987), as well as specificity measures in order to judge the quality of evidence. In the context of *sampling designs* (e.g., Hajek, 1981), a random set X in a finite population U is a sampling design. Thus, to design a sampling plan, it suffices to specify a density f on 2^U, whose entropy is used as a *measure of spread* for sampling probabilities and it is well-known that every conditional Poisson sampling design maximixes the entropy in the class of designs having the same carrier and same covering function, i.e., $\pi_X : U \to [0,1]$

$$\pi_X(u) = P(u \in X) = \sum_{u \in A} f(A)$$

As we will see, in general, the constraints $\mathscr{F}$ in entropy maximization of random sets are sets of densities on U. But here is a maximization problem where, as in the case of random variables, the constraints are in the form of known moments.

Just like the case of random vectors, it is often possible to know some "moments" of a random set, either by statistical estimation or other computing methods. By "moments" of a random set X we really mean the moments of its measure, here, in the finite case, the counterpart of the first moment is the expected value of its cardinality

$$E(|X|) = \sum_{u \in U} \sum_{\{A: u \in A\}} f(A) = \sum_{u \in U} \pi_X(u)$$

The above expression for $E(|X|)$ is a special case of Robbins' formula (Robbins, 1944) for random closed sets in $\mathbb{R}^d$. Here are the details. Let $\mathscr{F}$ denote the space of all closed subsets of $\mathbb{R}^d$, and $\mathscr{B}(\mathscr{F})$ its Borel $\sigma-$ field, generated by the hit-or-miss topology (see Matheron, 1975). A *random closed set* X is a map from Ω to $\mathscr{F}$ such that $X^{-1}(\mathscr{F}) \subseteq \mathscr{A}$, and its probability measure on $\mathscr{B}(\mathscr{F})$ is $P_X = PX^{-1}$, as usual.

Remark

Just like in probability theory, the case of finite sets is a simple starting point. Evidence can induce belief functions on more general parameter spaces, where it is more convenient to work with their duals, namely, Choquet capacity functionals (on locally compact, Hausdorff spaces, but not on infinitely dimensional Polish spaces, however, see Nguyen and Nguyen, 1998). It is also possible to treat belief functions, or equivalently, random sets (including random fuzzy sets), in a unified manner using the setting of *continuous lattices* (see Gierz et al, 1980; Nguyen and Tran, 2008).

Let $\varphi : \mathbb{R}^d \times \mathscr{F} \to [0,1]$ be

$$\varphi(x,F) = \begin{cases} 1 \text{ if } x \in F \\ 0 \text{ if } x \notin F \end{cases}$$

Then the restriction of the Lebesgue measure λ on $\mathbb{R}^d$ to $\mathscr{F}$ is

$$\lambda(F) = \int_{\mathbb{R}^d} \varphi(x,F)\lambda(dx)$$

so that, by Fubini's theorem, the map $\lambda : F \in \mathscr{F} \to \bar{\mathbb{R}}$ is $\mathscr{B}(\mathscr{F}) - \mathscr{B}(\bar{\mathbb{R}})-$ measurable. It follows that $\lambda \circ X = \lambda(X)$ is $\mathscr{A} - \mathscr{B}(\bar{\mathbb{R}})-$ measurable, i.e., the Lebesgue measure of the random closed set X is a bona fide (nonnegative) random variable. On the other hand, $\varphi(.)$ is measurable since

$$\varphi^{-1}(\{0\}) = \{(x,F) : x \notin F\} = \cup_{B \in \mathbb{B}}(B \times \mathscr{F}^B) \in \mathscr{F}(\bar{\mathbb{R}}) \otimes \mathscr{B}(\mathscr{F})$$

where $\mathbb{B}$ is a countable base for the topology of $\mathbb{R}^d$, and $\mathscr{F}^B = \{F \in \mathscr{F} : F \cap B = \varnothing\}$.

As such, by Fubini,

$$\int_{\mathbb{R}^d \times \mathscr{F}} \varphi(x,F)d(\lambda \otimes P_X) = \int_{\mathbb{R}^d} \int_{\mathscr{F}} \varphi(x,F)dP_X(F)d\lambda(x)$$

$$= \int_{\mathscr{F}} \int_{\mathbb{R}^d} \varphi(x,F)d\lambda(x)dP_X(F) = \int_{\mathbb{R}^d} P(x \in F)d\lambda(x)$$

$$= \int_{\mathscr{F}} \lambda(F)dP_X(F) = E[\lambda(X)]$$

which yields Robbins' formula

$$E[\lambda(X)] = \int_{\mathbb{R}^d} \pi_X(x) d\lambda(x)$$

where $\pi_X(x) = P(x \in X)$, the one-point coverage function of X. Robbins' formula is interesting since, as far as the expected value of the measure of a random set is concerned, there is no need to derive the distribution of $\lambda(X)$ from that of X (not an easy task!), but it suffices to derive its one-point coverage function (a much easier task).

Since Fubini theorem is valid for $\sigma-$ finite measures, we have $E(|X|) = \sum_{u \in U} \pi_X(u)$ for a random set X on a finite set U, with the counting measure $|.|$.

Consider the maximization problem

$$\text{Maximize} - \sum_{A \subseteq U} f(A) \log f(A)$$

subject to

(i) f is a density on 2^U
(ii) $E_f(|X|) = \sum_{u \in U} \sum_{\{A: u \in A\}} f(A) = \theta \in (1, |U|)$

Observe that $E(|X|) = \sum_{j=1}^{n} j q_j$, where $n = |U|$, $q_j = \sum f(A)$ where the sum is over $A \subseteq U$ such that $|A| = j$. If we let $p_i, i = 1, 2, ..., 2^n - 1 = m$, be $f(A)$, $A \subseteq U$ (excluding $\varnothing$) and let $a_i \in \{1, 2, ..., n\}$ be such that $E(|X|)$ is written as $\sum_{i=1}^{m} a_i p_i$, then the above problem becomes

$$\text{Maximize} - \sum_{i=1}^{m} p_i \log p_i$$

subject to (i) $p_i \geq 0$, $\sum_{i=1}^{m} p_i = 1$ and (ii) $\sum_{i=1}^{m} a_i p_i = \theta$

Using Lagrange multiplier technique, the solution is found to be $p_i = \frac{1}{\phi(\beta)} e^{-\beta a_i}$ where $\phi(\beta) = \sum_{i=1}^{m} e^{-\beta a_i}$, and β is the unique solution of the equation $\phi'(\beta) + \theta\phi(\beta) = 0$.

For another maximum entropy problem related to belief functions, see Jaffray (1997).

4.2 Maximum Entropy Principle

In the context of belief functions, we address the problem of maximizing the entropy $H(f)$ over the the class $\mathscr{F}$ of densities on U compatible with a given belief function g. Specifically, given g, its Möbius inverse $g * \mu$ is a density on 2^U. If α is an allocation of $g * \mu$, then

$f_\alpha(u) = \sum_{\{A: u \in A\}} \alpha(u, A)$ is a density on U. This density on U gives rise to a probability measure $Q_\alpha(A) = \sum_{u \in A} f_\alpha(u)$, defined on 2^U, such that $Q_\alpha \geq g$. In fact,

all probability measures Q on U (i.e., on $(U, 2^U)$), such that $Q \geq g$ come from allocations. In other words,

$$\mathscr{C}(g) = \{Q_\alpha \geq g : \alpha \text{ allocations}\}$$

Then,

$$\mathscr{F} = \{f_\alpha : \alpha \text{ allocations}\}$$

We seek

$$\max - \sum_{u \in U} f(u) \log f(u)$$

subject to $f \in \mathscr{F}$.

The solution to this general problem is given in Meyerowitz et al. (1994).

Remark

The approach to expected utility in decision-making can be also carried out, in the context of belief functions, by using expectation of a function of a random set X on U. Specifically, let $\varphi : 2^U \to \mathbb{R}$, then $E(\varphi(X)) = \sum_{A \subseteq U} \varphi(A) P(X = A)$. Now, for each $f \in \mathscr{F}$, we can find many $\varphi : 2^U \to \mathbb{R}$ such that $E_f(u) = E(\varphi(X))$. Indeed, for any $\varphi : 2^U \to \mathbb{R}$, we modify it to φ_A, for some chosen $A \subseteq U$ with $P(X = A) = (g * \mu)(A) \neq 0$, as

$$\varphi_A(B) = \begin{cases} \varphi(B) \text{ for } B \neq A \\ \frac{E_f(u) - \sum_{B \neq A} \varphi(B)(g*\mu)(B)}{(g*\mu)(A)} \text{ for } B = A \end{cases}$$

The point is this. Selecting φ and considering $E(\varphi(X))$ as expected utility seems to be a more general procedure. For more details, see Nguyen and Walker (1994).

4.3 Minimax

Let $u = \mathbb{A} \times U \to \mathbb{R}$ be a utility function, and $\mathscr{P}$ be the set of probability measures on U compatible with a given belief function F on U, i.e., $\mathscr{P} = \{P : P \geq F\}$. The minimax procedure consists of choosing the action $a \in \mathbb{A}$ to maximize $\inf\{E_P u(a, .) : P \in \mathscr{P}\}$.

It turns out that the above infimum (for each fixed a) is attained and is equal to the Choquet integral of $u(a, .)$ with respect to the belief function F. This can be seen as follows.

Suppose $U = \{u_1, u_2, ..., u_n\}$ with $u(u_1) \leq u(u_2) \leq ... \leq u(u_n)$ (we drop a for simplicity). Then

$$\sum_{i=1}^{n} u(u_i)[F(\{u_i, u_{i+1}, ..., u_n\}) - F(\{u_{i+1}, u_{i+2}, ..., u_n\})] = E_F(u)$$

where

$$E_F(u) = \int_0^\infty F(u > t)dt + \int_{-\infty}^0 [F(u > t) - 1]dt$$

If we let

$$f(u_i) = F(\{u_i, u_{i+1}, ..., u_n\}) - F(\{u_{i+1}, u_{i+2}, ..., u_n\})$$

then f is a density on U and $f \in \mathscr{F}$. Indeed, let $A_i = \{u_i, u_{i+1}, ..., u_n\}$, then

$$f(u_i) = F(A_i) - F(A_i \backslash \{u_i\}) =$$
$$\sum_{B \subseteq A_i} (F * \mu)(B) - \sum_{B \subseteq A_i \backslash \{u_i\}} (F * \mu)(B) = \sum_{u_i \in B \subseteq A_i} (F * \mu)(B)$$

so that $f \in \mathscr{F}$. Next, for each $t \in \mathbb{R}$ and $g \in \mathscr{F}$, it can be checked that $P_f(u > t) \leq P_g(u > t)$ since $(u > t)$ is of the form $\{u_i, u_{i+1}, ..., u_n\}$. Hence, $E_{P_f}(u) \leq E_{P_g}(u)$.

Remark

For a comprehensive treatment of Choquet integral, see Sriboonchitta et al (2010).

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Probability 38, 325–339 (1967)
2. Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. Intern. J. General Systems 12, 193–226 (1986)
3. Gierz, G., Hofman, K.H., Keimel, K., Lawson, J.D., Mislove, M., Scott, D.S.: A Compendium of Continuous Lattices. Springer (1980)
4. Goodman, I.R.: Fuzzy sets as equivalence classes of random sets. In: Yager, R. (ed.) Fuzzy Sets and Possibility Theory, pp. 327–343 (1982)
5. Hajek, K.: Sampling from a Finite Population. Marcel Dekker, New York (1981)
6. Hartigan, J.A.: Estimation of a convex densoty contour in two dimensions. JASA 82, 267–270 (1987)
7. Jaffray, J.Y.: On the maximum of conditional entropy for upper/lower probabilities generated by random sets. In: Goutsias, J., et al. (eds.) Random Sets: Theory and Applications, pp. 107–127. Springer (1997)
8. Hartigan, J.A.: Information theory and statistical mechanics. Phys. Rev. 106(4), 620–630 (1957)
9. Marinacci, M.: Decomposition and representation of coalitional games. Math. Oper. Res. 21, 1000–1015 (1996)
10. Matheron, G.: Random Sets and Integral Geometry. John Wiley (1975)
11. Meyerowitz, A., Richman, F., Walker, E.A.: Calculating maximum entropy probability densities for belief functions. Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems 2, 377–390 (1994)
12. Molchanov, I.: Theory of Random Sets. Springer (2005)
13. Molchanov, I.: An Introduction to Copulas. LNCS. Springer (1999)

14. Nguyen, H.T.: On random sets and belief functions. J. Math. Anal. Appl. 65, 531–542 (1978); reprinted in Yager, R., Liu, L. (eds.): Classical Works of the Dempster-Shafer Theory of belief Functions, pp. 105–116. Springer (2008)
15. Nguyen, H.T.: On the entropy of random sets and possibility distributions. In: Bezdek, J. (ed.) The Analysis of Fuzzy Information, pp. 145–156. CRC Press (1987)
16. Nguyen, H.T., Walker, E.A.: On decision -making using belief functions. In: Yager, R., et al. (eds.) Advances in the Dempster-Shafer Theory of Evidence, pp. 312–330. John Wiley (1994)
17. Nguyen, H.T., Nguyen, N.T.: A negative version of Choquet theorem for Polish spaces. East- West J. Math. 1, 61–71 (1998)
18. Nguyen, H.T., Kreinovich, V.: How to divide a territory? A new simple differential formalism for optimization of set-functions. Intern. J. Intell. Systems 14, 223–251 (1999)
19. Nguyen, H.T., Walker, E.A.: A First Course in Fuzzy Logic, 3rd edn. Chapman and Hall/CRC (2006)
20. Nguyen, H.T.: An Introduction to Random Sets. Chapman and Hall/CRC (2006)
21. Nguyen, H.T., Tran, H.: On a continuous lattice approach to modeling of coarse data in system analysis. J. Uncertain Systems 1(1), 62–73 (2007)
22. Robbins, H.E.: On the measure of a random set. Ann. Math. Statist. 14, 70–74 (1944)
23. Scarsini, M.: Copulae of probability measures on product spaces. J. Multi. Anal. 31, 201–219 (1989)
24. Scarsini, M.: Copulae of capacities on product spaces. In: Distributions with Fixed Marginals and Related Topics. IMS Lecture Notes, vol. 28, pp. 307–318 (1996)
25. Shafer, G.: A Mathematical Theory of Evidence. Princeton Univ. Press (1976)
26. Sklar, A.: Fonctions de repartition a n dimensions et leur marges. Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)
27. Sriboonchitta, S., Wong, W.K., Dhompongsa, S., Nguyen, H.T.: Stochastic Dominance and Applications to Finance, Risk and Economics. Chapman and Hall/CRC (2010)
28. Suppes, P., Zanotti, A.: On using random relations to generate upper and lower probabilities. Synthese 36, 427–440 (1977)
29. Wasserman, L.A.: A Some Applications of Belief Functions to Statistical Inference. Ph.D. Thesis, University of Toronto, Canada (1987)
30. Yager, R.: Entropy and specificity in a mathematical theory of evidence. Intern. J. General Systems 9(4), 249–269 (1983)

Evidential Multi-label Classification Using the Random *k*-Label Sets Approach

Sawsan Kanj, Fahed Abdallah, and Thierry Denœux

Abstract. Multi-label classification deals with problems in which each instance can be associated with a set of labels. An effective multi-label method, named RA*k*EL, randomly breaks the initial set of labels into smaller sets and trains a single-label classifier in each of this subset. To classify an unseen instance, the predictions of all classifiers are combined using a voting process. In this paper, we adapt the RA*k*EL approach under the belief function framework applied to set-valued variables. Using evidence theory makes us able to handle lack of information by associating a mass function to each classifier and combining them conjunctively. Experiments on real datasets demonstrate that our approach improves classification performances.

1 Introduction

Multi-label classification considers problems in which an object may belong simultaneously to multiple classes [4, 5, 10]. Several applications may be subscribed under the multi-label classification problem. In semantic scene classification, each image can be separated into semantic classes as beaches, sunsets or parties [1]. In text categorization, each document may belong to multiple categories such as government, arts and health [6]. In music classification, each song can evoke more than one emotion at the same time, such as amazed, happy, excited, etc. [7].

A lot of algorithms have been proposed for multi-label learning. The existing methods can be categorized into two groups: the *indirect* methods and the *direct* ones [8]. The former one transforms the multi-label classification problem into one or more single-label classification problems, while the latter handles directly the multi-label classification problem.

This paper focuses on an effective multi-label learning method introduced in [9]. This method, named RA*k*EL (RAndom-*k*-labEL sets), aims at solving the multi-label classification problem while taking into consideration the correlation between labels. It randomly breaks the set of labels into smaller sets and learns a single-label

Sawsan Kanj · Fahed Abdallah · Thierry Denœux
Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France
e-mail: firstname.lastname@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 21–28.
springerlink.com

classifier for each subset. To make a decision, the different predictions for each label are aggregated via voting. In this approach, the user has to identify the number of random label sets, the size of these sets and an adequate threshold in the voting process.

Our goal in this paper is to alleviate the loss of information inherent in the RA*k*EL method (as each base classifier only considers a subset of labels) while accounting for label correlation in a more efficient way. For this purpose, we propose to retain the basic principle of the RA*k*EL approach but to combine the different classifiers in the belief function framework. In [3], a formalism for representing uncertain information has been proposed for manipulating knowledge about set-valued variables. We use this formalism in order to represent and combine information about an unseen instance and to predict its set of labels. To show the effectiveness of this strategy even when using simple classifiers structure, we use Linear Discriminant Analysis (LDA) as the base-level learning method for each classifier. In LDA, each classifier provides information about the object to classify on the form of estimated posterior probabilities. Due to the fact that these outputs can be expressed as set-valued variables, we encode them as mass functions and combine them conjunctively. To make a final decision, we compute the belief function for each label or the maximum of commonality in order to find the whole set of labels to be assigned. The proposed method, called *Evidential-Rakel-LDA* has the advantage of reducing the number of parameters since the decision making process is automatically performed under the belief function framework.

The rest of this paper is organized as follows. Section 2 recalls the background on belief functions for set-valued variables. Section 3 introduces the *Rakel-LDA* method. Section 4 presents experiments on two real datasets and discusses the results. Finally, section 5 concludes the paper.

2 Belief Functions on Set-Valued Variables

Let X be a variable taking zero, one or several values in a finite set Ω. Such a variable is said set-valued [3].

To express partial knowledge about a set-valued variable X, we may specify a set A of values that are *certainly* taken by X and a set B of values that are *certainly not* taken by X. The set of subsets of Ω that contain A and have an empty intersection with B is denoted by $\varphi(A,B)$. Let $C(\Omega)$ be the set of all subsets of $\Theta = 2^{\Omega}$ of the form $\varphi(A,B)$, completed by the empty set of Θ.

The theory of belief functions can be applied to describe partial knowledge about set-valued variables by defining a mass function on $\Theta = 2^{\Omega}$. It is clear that the cardinality of $C(\Omega)$ is equal to $3^K + 1$.

The belief and commonality functions are defined, respectively, as follows:

$$bel(A,B) = \sum_{\varphi(C,D)\subseteq\varphi(A,B)} m(C,D) - \emptyset_{\Theta}, \tag{1}$$

$$q(A,B) = \sum_{\varphi(C,D)\supseteq\varphi(A,B)} m(C,D), \tag{2}$$

where $m(A,B)$ is a notation for $m(\varphi(A,B))$.

As shown in [3], Dempster's rule can be expressed as follows:

$$(m_1 \oplus m_2)(A,B) = \frac{\sum_{\varphi(C,D)\cap\varphi(E,F)=\varphi(A,B)} m_1(C,D)m_2(E,F)}{\sum_{\varphi(C,D)\cap\varphi(E,F)\neq\emptyset_\Theta} m_1(C,D)m_2(E,F)}. \tag{3}$$

Even if the evidential approach reduces the number of focal elements to 3^K+1, this method still has high complexity for large numbers of labels. As an example, if we have 20 labels in the multi-label problem, we may have to handle up to $3.4868e+009$ focal elements. The method proposed in the next section aims to overcome this problem by applying the Evidential formalism to several partitions of the label set and to combine the results under the belief functions framework.

3 Evidential-Rakel-LDA

Let $\mathscr{X} = \mathbb{R}^d$ denote the input space, and let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_Q\}$ be the finite set of labels. The multi-label classification problem can be described as follows. Given a training set $\mathscr{D} = \{(x_1, Y_1), \ldots, (x_N, Y_N)\}$, of N instances drawn from $\mathscr{X} \times 2^\Omega$, and identically distributed, where x_i is a feature vector describing instance i, and $Y_i \subseteq \Omega$ is the set of labels for that instance, the goal of the multi-label learning is to find a multi-label classifier $\mathscr{H} : \mathscr{X} \to 2^\Omega$ that can associate a set of labels to each unseen instance.

As in the standard RAkEL method, we randomly split the initial set of labels Ω into a number of smaller label sets Ω_j. For each one, the training set of instances, denoted $\mathscr{D}_j$, is deduced from the original dataset $\mathscr{D}$ by replacing the label sets of training instances by their intersections with Ω_j. Inside $\mathscr{D}_j$, each combination of labels is considered as a new class (or group of classes).Using $\mathscr{D}_j$, we train an LDA classifier, denoted h_j (here h_j is a single-label classifier). Note that LDA is used to generate a set of linear functions, one for each group. These functions are built by maximizing the ratio of the between-class variance to the within-class variance. In order to make a decision for an unseen instance x, LDA estimates the posterior probability for each group of the set Ω_j.

In the frame of discernment Ω, the individual classifier outputs are considered as items of evidence. Each output is represented by a mass function on a focal set, noted by $\varphi(A_q, B_q)$ where $A_q, B_q \subseteq \Omega_j$. In other words, A_q is the set of labels assigned to one group and B_q is its complement in Ω_j.

After considering all the items of evidence as items on Ω, we combine them using the Dempster's rule (3) to form the resulting BBA m for an unseen instance. To determine the set of estimated label $\widehat{Y}$ of the unseen instance, we compare the two degrees of belief $bel(\omega,\emptyset)$ and $bel(\emptyset,\omega)$ for each label in Ω [3]:

$$\widehat{Y} = \{\omega \in \Omega / bel(\{\omega\},\emptyset) \geq bel(\emptyset,\{\omega\})\}. \tag{4}$$

Note here that the decision making process is automatically performed without having to define threshold. As shown by Denœux and Masson [2], we can also calculate the communality function and the maximum of this function can be determined by solving an integer programming problem with non-linear constraints. In this case, another way to calculate $\widehat{Y}$ is to select the set of labels with the largest communality.

4 Experiments

4.1 Evaluation Metrics

To evaluate the performance of our method, we calculate different metrics used in the multi-label literature [8].

Hamming Loss: The Hamming Loss metric refers to the percentage of labels that are misclassified, i.e., incorrect labels that are predicted or true labels that are not predicted:

$$\mathscr{H}Loss = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \triangle \widehat{Y}_i|}{Q}, \tag{5}$$

where $\triangle$ denotes the symmetric difference between two sets.

Accuracy: Accuracy measures the degree of closeness between the predicted and the ground truth label sets:

$$\mathscr{A}ccuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}. \tag{6}$$

F_1 *measure*: The F_1 measure is defined as the harmonic mean of two other metrics called precision and recall. *Precision* is the fraction of predicted labels that are true, while *recall* is the fraction of true labels that are predicted.

$$\mathscr{P}recision = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \widehat{Y}_i|}{|\widehat{Y}_i|}, \tag{7}$$

$$\mathscr{R}ecall = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \widehat{Y}_i|}{|Y_i|}, \tag{8}$$

and

$$\mathscr{F}_1 = 2.\frac{\mathscr{P}recision.\mathscr{R}ecall}{\mathscr{P}recision + \mathscr{R}ecall}. \tag{9}$$

The smaller the value of the *Hamming Loss*, the better the performance. For the other metrics, higher values correspond to better classification quality.

4.2 Datasets

Our method was experimented using the emotions and scene datasets [1].

The Emotion dataset contains 593 songs described by eight rhythmic features and 64 timbre features. There are six classes, and each song can belong to more than one label according to the emotions generated.

The Scene dataset consists of 2407 natural scene images. There are six different semantic classes. Spatial color moments are used as features. Each image is divided into 49 blocks using 7×7 grid. The mean and variance of each band are computed corresponding to a low-resolution image and to computationally inexpensive texture features, respectively. Each image is then described by $49 \times 2 \times 3$ features [11].

4.3 Results and Discussions

We compared our method to the classical RA*k*EL approach based on the LDA method with different threshold values. The number k of labels in each subset was fixed to three for all experiments and the number of classifiers was ranging from 2 to $2 * Q$. Experiments on *Rakel-ADL* were done with all meaningful values for the threshold (0.1, 0.5 and 0.9).

Due to randomization of label space, results are very sensitive to the selected combination of labels. To deal with this negative aspect, we grouped results in batches of 10 classifiers calculated for the same value of k, and we computed the average.

Figures 1 to 3 show the box plots for the different metrics obtained for the emotion and scene datasets. From Figure 1, we can notice that our method performs

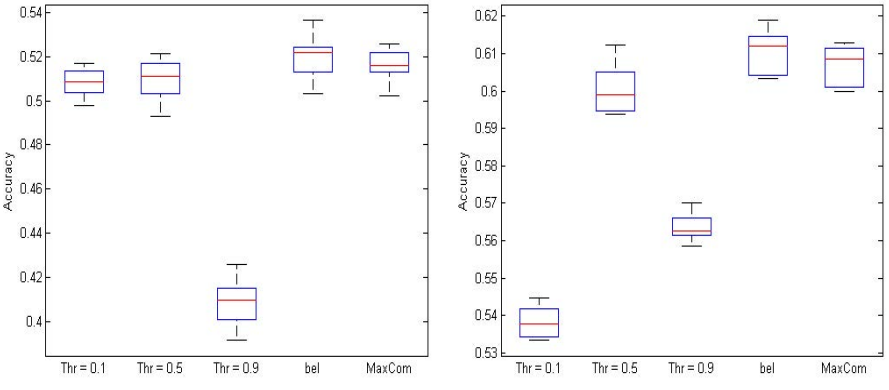


Fig. 1 *Accuracy* box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

[1] http://mulan.sourceforge.net/datasets.html

better than *Rakel-ADL* for different values of threshold in term of *Accuracy* on the two datasets.

Figure 2 shows the performance of the F_1 measure metric. As we can see on the scene dataset, the proposed method yields good performances and it is competitive with the two versions of decision. On the emotion dataset, *Rakel-ADL* performs better for a threshold value equal to 0.1. This is due to the fact that the emotion dataset is more labelled than the scene one (the average number of labels per instance is 1.87 for the former, while it is 1.07 for the latter). Decreasing the threshold value can result in taking into account all positive true labels and increasing the value of the *recall* metric.

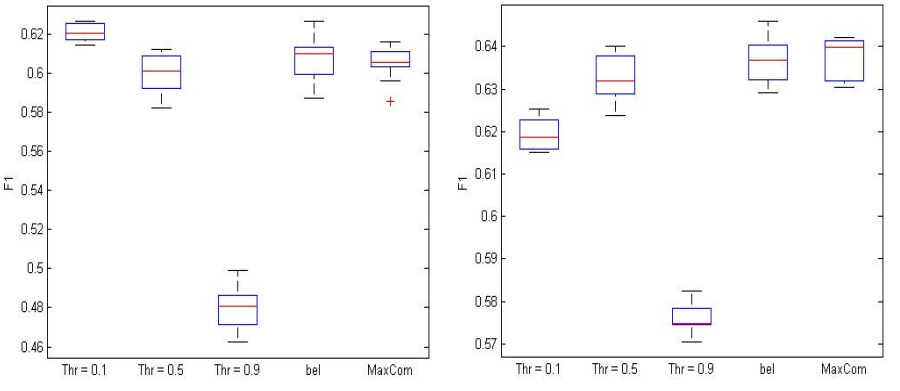


Fig. 2 F_1 box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

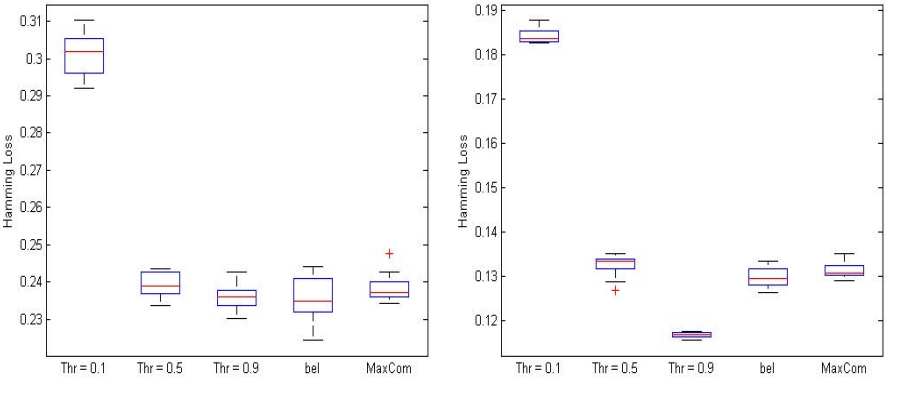


Fig. 3 *Hamming Loss* box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

Table 1 Experimental results (mean±std) of the compared algorithms on the emotions dataset

	Rakel-LDA **thr =** 0.1	***Rakel-LDA*** **thr =** 0.5	***Rakel-LDA*** **thr =** 0.9	***E-Rakel-LDA*** ***bel***	***E-Rakel-LDA*** ***max of com***
Accuracy	$0.508 \pm 0.006^{(4)}$	$0.509 \pm 0.009^{(3)}$	$0.409 \pm 0.009^{(5)}$	$0.519 \pm 0.009^{(1)}$	$0.516 \pm 0.007^{(2)}$
F_1	$0.621 \pm 0.004^{(1)}$	$0.598 \pm 0.011^{(4)}$	$0.479 \pm 0.011^{(5)}$	$0.607 \pm 0.012^{(2)}$	$0.605 \pm 0.009^{(3)}$
HLoss	$0.301 \pm 0.006^{(5)}$	$0.239 \pm 0.003^{(4)}$	$0.236 \pm 0.003^{(2)}$	$0.235 \pm 0.006^{(1)}$	$0.238 \pm 0.004^{(4)}$

Table 2 Experimental results (mean±std) of the compared algorithms on the scene dataset

	Rakel-LDA **thr =** 0.1	***Rakel-LDA*** **thr =** 0.5	***Rakel-LDA*** **thr =** 0.9	***E-Rakel-LDA*** ***bel***	***E-Rakel-LDA*** ***max of com***
Accuracy	$0.538 \pm 0.004^{(5)}$	$0.601 \pm 0.006^{(3)}$	$0.564 \pm 0.004^{(4)}$	$0.611 \pm 0.005^{(1)}$	$0.607 \pm 0.005^{(2)}$
F_1	$0.612 \pm 0.004^{(4)}$	$0.632 \pm 0.006^{(3)}$	$0.576 \pm 0.004^{(5)}$	$0.636 \pm 0.005^{(2)}$	$0.637 \pm 0.005^{(1)}$
HLoss	$0.184 \pm 0.002^{(5)}$	$0.132 \pm 0.003^{(4)}$	$0.117 \pm 0.001^{(1)}$	$0.129 \pm 0.002^{(2)}$	$0.131 \pm 0.002^{(3)}$

Figure 3 shows the box plot of the minimum *Hamming Loss* for different methods. On the emotion dataset, our approach shows good performances, while on the scene dataset and for a threshold equal to 0.9 we get the best result. This is due to the fact that increasing the threshold is followed by reducing the number of prediction errors (number of incorrect predicted labels), especially with the scene dataset (80% of instances have a single label).

Tables 1 and 2 show that our approach is suitable to multi-label classification problems under the Rakel approach where we have missing information due to lack of knowledge given by each classifier. Note that the intuitive threshold ($t = 0.5$) gives in average better performances on the *Rakel-ADL* over different values of threshold.

5 Conclusion

A variant of the RAkEL method for multi-label classification has been proposed, based on the theory of belief functions. Our approach uses the formalism developed in [3] to define belief functions for set-valued variables. This framework allows us to combine the outputs from base classifiers in a more efficient way than the voting process used in the reference method. Experimental results demonstrate the effectiveness of the approach.

References

1. Boutell, M.R., Shen, J., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
2. Denœux, T., Masson, M.-H.: Evidential reasoning in large partially ordered sets. Application to multi-label classification, ensemble clustering and preference aggregation. Annals of Operations Research (2011) (accepted for publication), doi:10.1007/s10479-011-0887-2

3. Denoeux, T., Younes, Z., Abdallah, F.: Representing uncertainty on set-valued variables using belief functions. Artificial Intelligence 174, 479–499 (2010)
4. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: 14th ACM International Conference on Information and Knowledge Management (2005)
5. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Proc. of the 20th European Conference on Machine Learning, ECML 2009 (2009)
6. Schapire, R., Singer, Y.: Boostexter: a boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
7. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 325–330 (2008)
8. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
9. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. In: Proc. 18th European Conference on Machine Learning, September 17-21 (2007)
10. Younes, Z., Abdallah, F., Denoeux, T., Snoussi, H.: A dependent multilabel classification method derived from the k-nearest neighbor rule. EURASIP Journal on Advances in Signal Processing, Article ID 645964, 14 (2011), doi:10.1155/2011/645964

An Evidential Improvement for Gender Profiling

Jianbing Ma, Weiru Liu, and Paul Miller

Abstract. CCTV systems are broadly deployed in the present world. To ensure in-time reaction for intelligent surveillance, it is a fundamental task for real-world applications to determine the gender of people of interest. However, normal video algorithms for gender profiling (usually face profiling) have three drawbacks. First, the profiling result is always uncertain. Second, for a time-lasting gender profiling algorithm, the result is not stable. The degree of certainty usually varies, sometimes even to the extent that a male is classified as a female, and vice versa. Third, for a robust profiling result in cases were a person's face is not visible, other features, such as body shape, are required. These algorithms may provide different recognition results - at the very least, they will provide different degrees of certainties. To overcome these problems, in this paper, we introduce an evidential approach that makes use of profiling results from multiple algorithms over a period of time. Experiments show that this approach does provide better results than single profiling results and classic fusion results.

1 Introduction

During the last decade, there has been massive investment in CCTV technology in the UK, e.g., e.g., the First Glasgow Bus Surveillance [10], Intelligent Surveillance Project [3, 4, 5, 6, 7, 8], Airport Corridor Surveillance [9], etc. Currently, there are approximately four million CCTV cameras operationally deployed. Despite this, the impact on anti-social and criminal behaviour has been minimal. For example, assaults on bus and train passengers are still a major problem for transport operators. Although most incidents, also called events, are captured on video, there is no response because very little of the data is actively analyzed in real-time. Consequently, CCTV operates in a passive mode, simply collecting enormous volumes of video data. For this technology to be effective, CCTV has to become active by alerting security analysts in real-time so that they can stop or prevent the undesirable

Jianbing Ma · Weiru Liu · Paul Miller
School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast BT7 1NN, UK
e-mail: {jma03,w.liu}@qub.ac.uk, p.miller@ecit.qub.ac.uk

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 29–36.
springerlink.com

behaviour. Such a quantum leap in capability will greatly increase the likelihood of offenders being caught, a major factor in crime prevention.

A key requirement for active CCTV systems is to automatically determine the threat posed by each individual to others in the scene. Most of the focus of the computer vision community has been on behaviour/action recognition. However, experienced security analysts profile individuals in the scene to determine their threat. Often they can identify individuals who look as though they may cause trouble before any anti-social behaviour has occurred. From criminology studies, the vast majority of offenders are young adolescent males. Therefore, key to automatic threat assessment is to be able to automatically profile people in the scene based on their gender and age. In this paper, we focus on the former.

The most obvious cue in determining a person's gender is the appearance of their face. However, for automatic classifiers this usually requires cooperative subjects who are directly looking at the camera and at close range. For most security scenarios one cannot assume this, as the person's face may not be visible as they are facing away from the camera, or they may be too far away - the resulting low resolution making gender discrimination difficult or impossible. Another obvious cue that can help overcome these issues is that of body shape. However, generally automatic classifiers of body shape are a less reliable indicator of gender than face-based classifiers. Furthermore, for both types of classifiers, the output result always has some degree of uncertainty. Secondly, when such classifiers are applied to video sequences, their output can vary significantly with time - even to the extent that a person's gender is incorrectly classified. Thirdly, the key to a robust solution is to use both face and body shape classifiers. Ideally, we would like to use the face classifier result, provided it is detected, otherwise we should resort to using the body shape result. However, this raises the issue of what to do when the outputs of both classifiers are different.

Imperfect information frequently occurs in video analytic processes. For example, a person may be classified as male with a certainty of 85% by a gender profiling algorithm. However, this does not imply that the person is female with a 15% certainty, rather, we say that the 15% represents what is unknown about the gender, i.e., we do not know how to distribute the remaining 15% between male and female. From probability theory, this information can only be represented as $p(male) \geq 0.85$ and $p(female) \leq 0.15$ (or interval probabilities), which is difficult to use for reasoning. Imperfect information is usually caused by ignorance or unreliability of the information sources. For example, a camera may have a faulty gain control setting, illumination could be poor, or the classifier training set may be unrepresentative. Any, or all, of these can result in imperfect information which cannot be represented by probability measures. On the other hand, such imperfect information can be easily handled using an evidential approach, namely, the Dempster-Shafer (DS) theory of evidence.

To address all of the above issues, we investigate whether a DS framework can combine uncertain profiling results from face and body shape classifiers over an extended time period, to provide robust gender profiling of subjects in video. Experiments show that this approach provides better results than a probabilistic approach.

DS theory [1, 11] is a popular framework to deal with uncertain or incomplete information from multiple sources. This theory is capable of modelling incomplete information through ignorance. For combining difference pieces of information, DS theory distinguishes two cases, i.e., whether pieces of information are from distinct, or non-distinct, sources. Many combination rules are proposed for information from distinct sources, among which are the well-known Dempster's rule [11] and Smets' rule [12]. In [2], two combination rules, i.e., the cautious rule and the bold disjunctive rule, for information from non-distinct sources are proposed. Thus, we view gender profiling results from the same classifier, e.g. face-based, at different times as being from non-distinct sources. For profiling results from different classifiers, they are naturally considered as being from distinct sources. Therefore, all of the problems mentioned above can be handled within the DS framework.

To the best of our knowledge, our approach is the first that addresses imperfect information from multiple sources for gender profiling. We demonstrate the significance and usefulness of our framework with experimental results on sample videos and by comparison to a probabilistic approach.

The rest of the paper is organized as follows. Section 2 provides the preliminaries on Dempster-Shafer theory. In Section 3, we discuss the difficulties in gender profiling in terms of scenarios. Section 4 provides experimental results which shows our method is better than a classic fusion approach and single profiling approaches. Finally, we conclude the paper in Section 5.

2 Dempster-Shafer Theory

For convenience, we recall some basic concepts of Dempster-Shafer's theory of evidence. Let Ω be a finite, non-empty set called the frame of discernment, denoted as, $\Omega = \{w_1, \cdots, w_n\}$.

Definition 1. *A basic belief assignment(bba) is a mapping* $m : 2^{\Omega} \rightarrow [0,1]$ *such that* $\sum_{A \subseteq \Omega} m(A) = 1$.

If $m(\emptyset) = 0$, then m is called a mass function. If $m(A) > 0$, then A is called a focal element of m. Let $\mathscr{F}_m$ denote the set of focal elements of m. A mass function with only a focal element Ω is called a *vacuous* mass function.

From a bba m, *belief* function (*Bel*) and *plausibility* function (*Pl*) can be defined to represent the lower and upper bounds of the beliefs implied by m as follows.

$$Bel(A) = \sum_{B \subseteq A} m(B) \text{ and } Pl(A) = \sum_{C \cap A \neq \emptyset} m(C). \tag{1}$$

One advantage of DS theory is that it has the ability to accumulate and combine evidence from multiple sources by using *Dempster's rule of combination*. Let m_1 and m_2 be two mass functions from two distinct sources over Ω. Combining m_1 and m_2 gives a new mass function m as follows:

$$m(C) = (m_1 \oplus m_2)(C) = \frac{\sum_{A \cap B = C} m_1(A) m_2(B)}{1 - \sum_{A \cap B = \emptyset} m_1(A) m_2(B)} \tag{2}$$

In practice, sources may not be completely reliable, to reflect this, in [11], a *discount rate* was introduced by which the mass function may be discounted in order to reflect the reliability of a source. Let r $(0 \leq r \leq 1)$ be a discount rate, a discounted mass function using r is represented as:

$$m^r(A) = \begin{cases} (1-r)m(A) & A \subset \Omega \\ r+(1-r)m(\Omega) & A = \Omega \end{cases} \tag{3}$$

When $r = 0$ the source is absolutely reliable and when $r = 1$ the source is completely unreliable. After discounting, the source is treated as totally reliable.

Definition 2. *Let m be a bba on Ω. A pignistic transformation of m is a probability distribution P_m over Ω such that $\forall w \in \Omega, P_m(w) = \sum_{w \in A} \frac{1}{|A|} \frac{m(A)}{1-m(\emptyset)}$ where $|A|$ is the cardinality of A.*

Let $\oplus$ be the conjunctive combination operator (or Smets' operator [12]) for any two bbas m, m' over Ω such that

$$(m \oplus m')(C) = \sum_{A \subseteq \Omega, B \subseteq \Omega, A \cap B = C} m(A)m'(B), \forall\, C \subseteq \Omega. \tag{4}$$

A simple bba m such that $m(A) = x, m(\Omega) = 1-x$ for some $A \neq \Omega$ will be denoted as A^x. The vacuous bba can thus be noted as A^0 for any $A \subset \Omega$. Note that this notation, i.e., A^x, is a bit different from the one defined in [2] in which A^x in our paper should be denoted as A^{1-x} in [2].

Similarly, for two sets $A, B \subset \Omega$, $A \neq B$, let $A^x B^y$ denote a bba m such that $m = A^x \oplus B^y$ where $\oplus$ is the conjunctive combination operator defined in Equation (4). For these kinds of bbas, we call them *bipolar* bbas. A simple bba A^x could be seen as a special bipolar bba $A^x B^0$ for any set $B \subseteq \Omega$, $B \neq A$.

It is easy to prove that any $m = A^x B^y$ is:

$$m(\emptyset) = xy, m(A) = x(1-y), m(B) = y(1-x), m(\Omega) = (1-x)(1-y) \tag{5}$$

In addition, when normalized, m in Equation 5 is changed to m' as follows.

$$m'(A) = \frac{x(1-y)}{1-xy}, m'(B) = \frac{y(1-x)}{1-xy}, m'(\Omega) = \frac{(1-x)(1-y)}{1-xy} \tag{6}$$

For two bipolar bbas $A^{x_1} B^{y_1}$ and $A^{x_2} B^{y_2}$, the cautious combination rule proposed in [2] is as follows.

Lemma 1 *(Denœux's Cautious Combination Rule). Let $A^{x_1} B^{y_1}$ and $A^{x_2} B^{y_2}$ be two bipolar bbas, then the combined bba by Denœux's cautious combination rule is also a bipolar bba $A^x B^y$ such that: $x = max(x_1, x_2), y = max(y_1, y_2)$.*

Also, according to [2], for $m_1 = A^{x_1} B^{y_1}$ and $m_2 = A^{x_2} B^{y_2}$, the combined result by Equation (2) is

$$m_{12} = A^{x_1 x_2} B^{y_1 y_2} \tag{7}$$

3 Gender Recognition Scenario

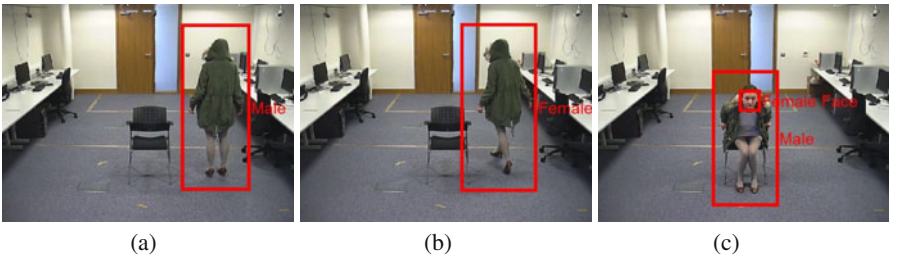
In this section, we provide a detailed description of a gender profiling scenario, which lends itself naturally to a DS approach.

Figure 1 shows three images taken from a video sequence that has been passed through a video analytic algorithm for gender profiling. In this sequence, a female wearing an overcoat with a hood enters the scene with her back to the camera. She walks around the chair, turning, so that her face becomes visible, and then sits down.

Fig. 1(a) shows that the subject is recognised by the full body shape profiling as a male. Note that her face is not visible. In Fig. 1(b), the subject is classified as female by the full body shape profiling algorithm. In Fig. 1(c), as she sits down, with her face visible, the face profiling algorithm classifies her as female, whilst the full body profiling classifies her as male. Note that the full body profiling algorithm is not as reliable as the face profiling algorithm. Conversely, full body profiling is always possible whilst the face information can be missing. That is why these two profiling algorithms should be considered together. In addition, as full body profiling is not as robust, discount operations should be performed on the algorithm output (cf. Equation (3)). The discount rate is dependent on the video samples and the training efficiency. For every video frame in which a body (face) is detected, gender recognition results are provided. The full body profiling algorithm and the face profiling algorithm, provided a person's face is detected, report their recognition results for every frame of the video, e.g., male with 95% certainty.

For a frame with only a body profiling result, for instance Fig. 1(a), the corresponding mass function m for body profiling will be M^x where M denotes that *the person is classified as a male* and x is the mass value of $m(\{M\})$. The corresponding mass function for face profiling is M^0F^0 where F denotes that *the person is classified as a female*, or the vacuous mass function. Alternatively, we can refer to this as the vacuous mass function.

Similarly, for a frame with both body profiling and face profiling, for instance Fig. 1(c), the corresponding mass function for body profiling will be M^x (or in a bipolar form M^xF^0) and the mass function for face profiling is F^y (or in a bipolar form M^0F^y) where x, y are the corresponding mass values. As time elapses, fusion of bipolar bbas by the cautious rule is reduced, as shown by Lemma 1. And when



(a) (b) (c)

Fig. 1 Three images taken from a video sequence

it comes to present the final profiling result, we use Dempster's rule to combine the two fused bipolar mass functions from the two recognition algorithms, respectively. Namely, for the two bipolar bbas $m_1 = M^{x_1}F^{y_1}$ and $m_2 = M^{x_2}F^{y_2}$, it is easy to get that the combined result m_{12} by Dempster's rule is (normalized from the result of Equation 7):

$$m_{12}(\{M\}) = \frac{m_1(\{M\}m_2(\{M\}))(1-m_1(\{F\})m_2(\{F\}))}{1-m_1(\{M\})m_2(\{M\})m_1(\{F\})m_2(\{F\})},$$
$$m_{12}(\{F\}) = \frac{m_1(\{F\}m_2(\{F\}))(1-m_1(\{M\})m_2(\{M\}))}{1-m_1(\{M\})m_2(\{M\})m_1(\{F\})m_2(\{F\})},$$
$$m_{12}(\Omega) = \frac{(1-m_1(\{M\})m_2(\{M\}))(1-m_1(\{F\})m_2(\{F\}))}{1-m_1(\{M\})m_2(\{M\})m_1(\{F\})m_2(\{F\})}.$$

Finally, we use the pignistic transformation (Def. 2) for the final probabilities. That is, $p(\{M\}) = m_{12}(\{M\}) + m_{12}(\Omega)/2$ and $p(\{F\}) = m_{12}(\{F\}) + m_{12}(\Omega)/2$

Example 1. *Let us illustrate the approach by a simple scenario with two frames. In the first frame, we have both body profiling (m_b^1) and face profiling (m_f^1) results as $m_b^1 = M^{0.7}F^{0.3}$ and $m_f^1 = M^{0.4}F^{0.6}$. In the second frame, we have the body profiling (m_b^2) result only, where $m_b^2 = M^0.8F^0.2$. By Lemma 1, the fusion results by the cautious rule is $m_b = M^{0.8}F^{0.3}$ and $m_f = M^{0.4}F^{0.6}$. Then by Equation 7, we get $m_{bf} = M^{0.32}F^{0.18}$, which, when normalized, is equivalent to $m_{bf}(\{M\}) = \frac{0.32(1-0.18)}{1-0.32*0.18} = 0.28$, $m_{bf}(\{F\}) = \frac{0.18(1-0.32)}{1-0.32*0.18} = 0.13$, $m_{bf}(\Omega) = \frac{(1-0.32)(1-0.18)}{1-0.32*0.18} = 0.59$. And finally we get $p(\{M\}) = 0.58$ and $p(\{F\}) = 0.42$.*

4 Experimental Results

In this section we compare fusion results obtained by Dempster-Shafer theory and a classic approach. As there are no benchmark datasets for both body and face profiling, we simulate the output of both body and face classifiers on a sequence containing a male subject. For the body classifier, the probability of any frame being correctly classified as male/female is roughly 60-90%. For the face classifier, only 75% of the available frames are randomly allocated as containing a face. For each of these frames the probability of the frame being correctly classified as being male/female is 85-100%. In both cases the values for $m(\{M\})$ and $m(\{F\})$ are uniformly sampled from the ranges 0.6-0.9 and 0.85-1.0 for the body and face classifiers outputs respectively.

As mentioned before, for gender profiling results from the same classifier at different time points, we use the cautious rule (Lemma 1) to combine them. For profiling results from different classifiers (i.e., face profiling and full body profiling), we use Dempster's rule (Equation (2)) to combine them. And finally, we apply the pignistic transformation (Def. 2) to get the probabilities of the subject being male or female.

Classic fusion in the computer vision community [13] takes the degrees of certainty as probabilities, i.e., they consider the face profiling and the full body

profiling output p_f^t and p_b^t indicating the probabilities of faces and full bodies being recognized as males at time t. Then it uses $p_{b,f}^t = c_f^t p_f^t + c_b^t p_b^t$ to calculate the final probability $p_{b,f}^t$ at time t, where c_f^t and c_b^t are the weights of the face and full body profiling at time t, proportional to the feasibility of the two algorithms in the last twenty frames. As full body profiling is always feasible, suppose face profiling can be applied n times in the last twenty frames, then we have:

$$c_b = \frac{20}{20+n}, c_f = \frac{n}{20+n}.$$

For this experiment, the performance of the DS and classic fusion schemes were characterised by the true positive rate:

$$T_{PR} = \frac{N_{PR}}{N}$$

where N_{PR} is the number of frames in which the gender has been correctly classified and N is the total number of frames in which the body/face is present. According to the training on the sample videos, the discount rate r for the full body profiling is set to 0.3. For comparison, we calculate the T_{PR} value for the body classifier alone, the face classifier, the DS fusion scheme and the classic fusion scheme.

When applying the methods on the randomly-generated simulation data, the comparison results are presented as follows.

Table 1 Comparison of T_{PR} for body classification, face classification, DS fusion and classic fusion

Methods	TotalFrame	N	N_{PR}	T_{PR} (%)
Full Body	3100	3100	1872	60.4
Face	3100	2321	2178	93.8
Classic Method	3100	3100	2658	85.7
DS Approach	3100	3100	3014	97.2

From Table 1, we can see that the DS fusion scheme gives an increase of approximately 11% in T_{PR} compared to the classic fusion scheme.

5 Conclusion

In this paper, we have proposed how to combine gender profiling classifier results by utilizing DS theory. We have used the cautious rule to combine gender profiling results from the same classifier at different time points and used Dempster's rule to combine profiling results from different classifiers. Experimental results show that the introduction of the DS theory indeed improves profiling performance.

We have mentioned that there are three problems that a classic gender profiling system should deal with, i.e., uncertain profiling results, unstable results over

time for a gender profiling classifier, and different classifiers capturing different features. We have shown that a DS-based approach handles these three issues in a seamless way.

For future work, we plan to apply the fusion schemes to profiling classifier results generated from real video sequences.

Acknowledgements. This research work is sponsored by the EPSRC projects EP/G034303/1 and EP/H049606/1 (the CSIT project).

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. The Annals of Statistics 28, 325–339 (1967)
2. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artifical Intelligence 172(2-3), 234–264 (2008)
3. Liu, W., Miller, P., Ma, J., Yan, W.: Challenges of distributed intelligent surveillance system with heterogenous information. In: Procs. of QRASA, Pasadena, California, pp. 69–74 (2009)
4. Ma, J., Liu, W., Miller, P.: Event Modelling and Reasoning with Uncertain Information for Distributed Sensor Networks. In: Deshpande, A., Hunter, A. (eds.) SUM 2010. LNCS, vol. 6379, pp. 236–249. Springer, Heidelberg (2010)
5. Ma, J., Liu, W., Miller, P.: Belief change with noisy sensing in the situation calculus. In: Procs. of UAI (2011)
6. Ma, J., Liu, W., Miller, P.: Handling Sequential Observations in Intelligent Surveillance. In: Benferhat, S., Grant, J. (eds.) SUM 2011. LNCS, vol. 6929, pp. 547–560. Springer, Heidelberg (2011)
7. Ma, J., Liu, W., Miller, P., Yan, W.: Event composition with imperfect information for bus surveillance. In: Procs. of AVSS, pp. 382–387. IEEE Press (2009)
8. Miller, P., Liu, W., Fowler, F., Zhou, H., Shen, J., Ma, J., Zhang, J., Yan, W., McLaughlin, K., Sezer, S.: Intelligent sensor information system for public transport: To safely go... In: Procs. of AVSS (2010)
9. ECIT Queen's University of Belfast. Airport corridor surveillance (2010), http://www.csit.qub.ac.uk/Research/ResearchGroups/IntelligentSurveillanceSystems
10. Gardiner Security. Glasgow transforms bus security with ip video surveillance, http://www.ipusergroup.com/doc-upload/Gardiner-Glasgowbuses.pdf
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
12. Smets, P.: Non-standard logics for automated reasoning. In: Smets, P., Mamdani, A., Dubois, D., Prade, H. (eds.) Belief Functions, pp. 253–286 (1988)
13. Zhou, H., Miller, P., Zhang, J., Collins, M., Wang, H.: Gender classification using facial and full body features. Technical Report, CSIT, Queen's University Belfast, UK (2011)

An Interval-Valued Dissimilarity Measure for Belief Functions Based on Credal Semantics

Alessandro Antonucci

Abstract. Evidence theory extends Bayesian probability theory by allowing for a more expressive model of subjective uncertainty. Besides standard interpretation of belief functions, where uncertainty corresponds to probability masses which might refer to whole subsets of the possibility space, *credal* semantics can be also considered. Accordingly, a belief function can be identified with the whole set of probability mass functions consistent with the beliefs induced by the masses. Following this interpretation, a novel, set-valued, dissimilarity measure with a clear behavioral interpretation can be defined. We describe the main features of this new measure and comment the relation with other measures proposed in the literature.

1 Introduction

Evidence theory [4, 7] generalizes classical Bayesian theory of probability by providing a more robust, and hence reliable, model of subjective uncertainty. While the Bayesian framework models uncertainty with probability masses assigned to single outcomes of a variable, evidence theory allows these masses to be associated to whole, not necessarily disjoint, sets of outcomes. The probabilities for the single states might be therefore not precisely specified, being only characterized by their lower and upper bounds, corresponding to *beliefs* and *plausibilities*. In other words, in general, there are multiple probability mass functions consistent with a single belief function specification. This is an equivalent characterization of a belief function, which can be identified with the *credal set* of its consistent mass functions. This provides a clear behavioral interpretation, based on Walley's theory of *imprecise probability* [8], where de Finetti's fair prices (associated to single mass functions) are extended to maximum buying/minimum selling prices.

Alessandro Antonucci
IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Manno-Lugano, Switzerland
e-mail: alessandro@idsia.ch

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 37–44.
springerlink.com

Although already present in the first formalization of evidence theory [4], these *credal semantics* received relatively little attention.[1] In this paper we exploit these semantics to define a novel, interval-valued, *dissimilarity measure* for belief functions.[2] Given a distance for probability mass functions, we evaluate the bounds when the two mass functions vary in the credal sets consistent with the belief functions to be compared. Notably, with the Manhattan (one-norm) distance, the evaluation of these bounds maps to linear programming and, the bounds can be equivalently evaluated by only comparing the extreme mass functions of the credal sets. Besides such a computational advantage, the behavioral semantics of credal sets can be used to provide a clear interpretation of the proposed measure.

Many dissimilarity measures for belief functions have been proposed [5], and some of them have been already based on comparisons of probability mass functions (e.g., the pignistic). The novelty of our approach consists in taking an interval-valued descriptor, which might provide a more cautious, and hence reliable, model of the (dis)similarity for belief functions.[3]

The paper is organized as follows. In Section 2, we review the basics of evidence theory and set the notation. Section 3 details the credal semantics of belief functions, while the interval-valued measure we propose is described in Section 4. Conclusions and outlooks are finally summarized in Section 6.

2 Basics

Let X denote a variable taking values in a finite set $\mathscr{X} := \{x_1, \ldots, x_n\}$. We consider two models of the uncertainty about the actual state of X.

A *probability mass function* P over X is a map $P : \mathscr{X} \to \mathbb{R}$, such that $P(x) \geq 0$ for each $x \in \mathscr{X}$ and $\sum_{x \in \mathscr{X}} P(x) = 1$. This models subjective uncertainty according to the following behavioral interpretation: the number $P(x)$ is regarded as the highest price a subject is willing to pay for buying a gamble which pays one unit if $X = x$ and zero otherwise (or equivalently the lowest price for which he/she sells it).

A *basic belief assignment* m over X is a map $m : 2^{\mathscr{X}} \to \mathbb{R}$, such that $m(A) \geq 0$ for each $A \in 2^{\mathscr{X}}$ and $\sum_{A \in 2^{\mathscr{X}}} m(A) = 1$.[4] Given $A, B \in 2^{\mathscr{X}}$, $inc(B,A)$ and $int(B,A)$ are indicator functions which are equal to one, being zero otherwise, if B is, respectively, included in A or has non-empty intersection with A. For each $A \in 2^{\mathscr{X}}$, the *belief* and *plausibility* of A corresponding to the mass m are:

[1] A remarkable exception is the work of Cuzzolin (e.g., [2]), where these semantics have has been exploited to define new, consistent, Bayesian approximations of belief functions.

[2] Although the focus of the paper is on the special class of credal sets associated to belief functions, the measure we present can be considered also for general credal sets.

[3] We agree with [6] in emphasizing the difficulties of capturing the level of dissimilarity between two belief functions with a single scalar indicator.

[4] The set of all the possible subsets of $\mathscr{X}$ is denoted by $2^{\mathscr{X}}$. Notation $|\cdot|$ will be used to denote the cardinality of the set in the argument. E.g., $|\mathscr{X}| = n$, and $|2^{\mathscr{X}}| = 2^n$.

$$b_m(A) := \sum_{B \in 2^{\mathscr{X}}} inc(B,A) \cdot m(B), \tag{1}$$

$$pl_m(A) := \sum_{B \in 2^{\mathscr{X}}} int(B,A) \cdot m(B). \tag{2}$$

It is trivial to check that beliefs and plausibilities are conjugated by the relation $b(A) = 1 - pl(\mathscr{X} \setminus A)$, for each $A \in 2^{\mathscr{X}}$. Similarly, the masses can be obtained from the beliefs through the so-called Möbius transform:

$$m(A) = \sum_{B \in 2^{\mathscr{X}}} mob(B,A) \cdot inc(B,A) \cdot b_m(B), \tag{3}$$

where $mob(B,A)$ is minus one if the difference between the cardinality of A and B is odd and one otherwise. Masses, beliefs and plausibilities can be therefore regarded as equivalent specifications of a single uncertainty model. In the following we refer to this model as a *belief function* (BF), independently of the particular way this has been specified. Given a BF, a probability distribution $P_m(X)$ can be obtained by simply considering the *pignistic* transformation:

$$P_m(x) := \sum_{B \in 2^{\mathscr{X}}} inc(\{x\},B) \frac{m(B)}{|B|}. \tag{4}$$

Finally note that a probability mass function can be regarded as a special belief function whose masses are defined only on the singletons. Note that, in this case, (4) returns the original mass function.

3 Credal Semantics of Belief Functions

Classical BFs semantics can be easily reduced to the interpretation of probability mass functions provided in the previous section. Exactly as P assigns mass $P(x)$ to event $X = x$, m assigns mass $m(A)$ to event $X \in A$. Yet, as the different elements of $2^{\mathscr{X}}$ are not exclusive, the masses associated to two or more subsets can contribute to determine the total amount of probability of an event. In particular, the sum as in (1) can be regarded as the minimum amount of probability associated to event $X \in A$ (and the sum in (2) the maximum). Multiple probability mass functions can be therefore consistent with a BF specification.

We denote by $K_m(X)$ the set of probability mass functions consistent with m:[5]

$$K_m(X) := \left\{ P(X) \left| \begin{array}{l} \sum_{x \in \mathscr{X}} P(x) = 1 \\ \sum_{x \in A} P(x) \geq b_m(A) \quad \forall A \in 2^{\mathscr{X}} \end{array} \right. \right\}. \tag{5}$$

[5] As a consequence of the conjugation between beliefs and plausibility, this set of probability mass functions can be equivalently defined in terms of plausibilities (with the inequalities inverted).

As a trivial consequence of (1), the pignistic as in (4) satisfies constraints in (5), being therefore included in $K_m(X)$. This implies that $K_m(X)$ cannot be empty. Similarly, the inequality constraints in (5) are tight, i.e., for each $A \in 2^{\mathscr{X}}$, a probability distribution satisfying the strict equality always exists. Different BFs should therefore induce different sets and *vice versa*. In other words, $K_m(X)$ is an equivalent specification for BFs.

Being defined by linear constraints, $K_m(X)$ is a closed and convex set of probability mass functions, i.e., a *credal set*.[6] Accordingly, Walley's behavioral interpretation of credal sets [8] can refer to BFs: the bounds with respect to $K_m(X)$ of the probability for an event A, which are respectively to $b(A)$ and $pl(A)$, can be regarded as the lowest selling price and the maximum buying price a subject is willing to pay for a gamble which pays one if $X \in A$ and zero otherwise.[7]

The *credal semantics* of m based on $K_m(X)$ also provides a direct geometric interpretation (see Figure 1). Being defined by linear constraints, $K_m(X)$ is a polytope over the probabilistic simplex, which can be equivalently described by the set $\text{ext}[K_m(X)]$ of its (finite-number) extreme points. These can be obtained from the plausibilities by a simple combinatorial formula. Let σ denote a permutation of the first n integers and $(x_{\sigma(1)},\ldots,x_{\sigma(n)})$ the corresponding permutation of $\mathscr{X}$; the corresponding extreme point $P^{\sigma}(X)$ of $K_m(X)$ is such that:[8]

$$P^{\sigma}(x_{\sigma(j)}) = \text{Pl}(\{x_{\sigma(1)},\ldots,x_{\sigma(j)}\}) - \text{Pl}(\{x_{\sigma(1)},\ldots,x_{\sigma(j-1)}\}), \tag{6}$$

for each $j = 2,\ldots,n$, while $P^{\sigma}(x_{\sigma(1)}) = \text{Pl}(\{x_{\sigma(1)}\})$. Being indexed by the permutations of the first n integers, the number of extreme probability mass functions in $K_m(X)$ cannot exceed the factorial of $n = |\mathscr{X}|$. Yet, most of the times, this is only an upper bound to the actual number of extremes: the less are the focal elements (i.e., events with non-zero mass), the less will be the distinct extreme mass functions returned by (6). As an example, if the non-zero masses are only the singletons and the universe, the distinct extremes will be only n (see Figure 1.a and 1.b). Finally, let us note that the average in the definition of the pignistic (4) corresponds to the computation of the center of mass of $K_m(X)$ (see references in [2]).[9]

As an example, the *vacuous* BF m_0 assigning all the mass to the universe (i.e., $m_0(\mathscr{X}) = 1$ and, hence, zero on any other subset) models a complete lack of information. The corresponding credal set coincides with the whole probability simplex, which will be denoted by $K_{m_0}(X)$ and its pignistic is uniform (see Figure 1.a).

[6] Note that there are credal sets which cannot be associated BFs. In this sense, credal set are a more general class of models of uncertainty.

[7] The behavioral counterpart of the non-emptiness and bounds tightness of $K_m(X)$, which has been proved in the previous paragraph, is that the subject obeys the rationality criteria of *avoiding sure loss* and *coherence* [8].

[8] This formula, formalized in [3], was rewritten in terms of the masses in [2]. Yet, such a characterization was already implicitly present in [4].

[9] This is true even if we consider only the extreme mass functions.

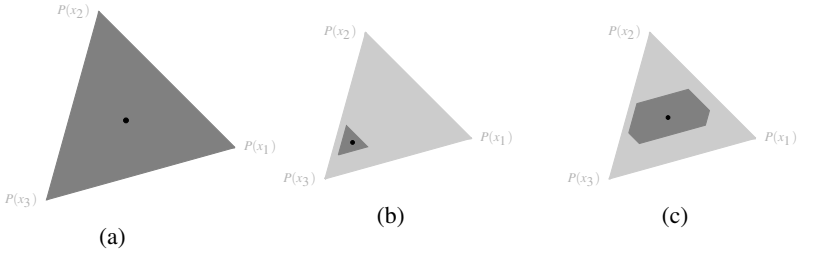


Fig. 1 Credal sets associated to BFs over a ternary variable X (dark gray) and their pignistic mass functions (black points). We consider: (a) the vacuous specification m_0; (b) a specification whose focal elements are only the singletons and on the universe with $m(\{x_1\}) = .05$, $m(\{x_2\}) = .15$, $m(\{x_3\}) = .6$, $m(\mathscr{X}) = .2$, and (c) a generic specification with $m(\{x_1\}) = .05$, $m(\{x_2\}) = .2$, $m(\{x_3\}) = .1$, $m(\{x_1,x_2\}) = .1$, $m(\{x_1,x_3\}) = .35$, $m(\{x_2,x_3\}) = .1$, $m(\{\mathscr{X}\}) = .1$.

4 A New Dissimilarity Measure for Belief Functions

The credal semantics introduced in the previous section is exploited here to define a new dissimilarity measure for BFs. This problem has been studied by many authors, and we point the reader to [5] for a survey. Yet, as emphasized by [6], scalar descriptors generally used for that can be unable to properly model the (dis)similarity between two BFs. This supports our idea of using an interval-valued measure.

Consider BFs m_1 and m_2 modeling two subjects' uncertainty about X. Our goal is define a measure of the (dis)similarity between the two subjects' beliefs based on the corresponding credal sets $K_{m_1}(X)$ and $K_{m_2}(X)$. Following a sensitivity analysis approach, we might assume that a *true* probability mass function (or, in behavioural terms, a true fair price), modeling the subjective uncertainty about X, exists for both subjects. Yet, due to partial lack of information, the subjects are only able to identify that these mass functions belong to their corresponding credal sets. As an example, the two credal sets in Figure 2.a partially overlap, and we cannot exclude that the two subjects' uncertainty corresponds to the same mass function. Yet, it could also be that they refer to completely different mass functions. To characterize this maximal dissimilarity case (and maximal similarity when the credal sets do not overlap) a measure to compare probability mass functions is needed.

To formalize these ideas, let us therefore consider a distance $\delta(P_1,P_2)$ modeling the level of (dis)similarity for any pair of mass functions $P_1(X)$ and $P_2(X)$. In particular, we consider a non-degenerate measure, i.e., the minimum distance $\delta(P_1,P_2) = 0$ is achieved if and only if $P_1 = P_2$, while its maximum value is normalized to one. The maximal dissimilarity should refer to a situation where both functions are deterministic, i.e., all the mass is assigned to a single outcome, which is different for the two functions. These desirable properties are, among others, satisfied by the so-called "Manhattan" distance, i.e., the one-norm measure:

$$\delta(P_1,P_2) := \frac{1}{2}\sum_{x\in\mathscr{X}} |P_1(x)-P_2(x)|. \tag{7}$$

We also provide an interpretation for this measure. Given variables X_1 and X_2, both with possibility space $\mathscr{X}$, we generate two samples of size m based on $P_1(X_1)$ and $P_2(X_2)$. The elements common to both samples are removed, and k elements remain. Then $\delta(P_1,P_2)$ coincides with k/m in the limit of large m. E.g., if both P_1 and P_2 are deterministic and referred to different outcomes, the two samples cannot have common elements and the distance should be one, while, if the two mass functions coincide, k should tend to zero.

Following the above discussion, we extend δ to cope with BFs by simply considering the bounds:[10]

$$\underline{\delta}(m_1,m_2) := \min_{P_1(X)\in K_{m_1}(X),P_2(X)\in K_{m_2}(X)} \delta(P_1,P_2), \tag{8}$$

$$\overline{\delta}(m_1,m_2) := \max_{P_1(X)\in K_{m_1}(X),P_2(X)\in K_{m_2}(X)} \delta(P_1,P_2). \tag{9}$$

With overlapping credal sets, (8) is zero, which means that the two models of uncertainty can refer to the same probability mass functions, while, because of the non-degeneracy, (9) is zero only if both the credal sets are made of a single mass function, this being the same for both. In fact, we cannot exclude that the two subjects refer to different mass functions. This is the case even when we compare a BF with itself. The result is a (scalar) descriptor of the level of Bayesianity for BFs, which we call *radius*:

$$\rho(m) := \overline{\delta}(m,m). \tag{10}$$

Only probability mass functions have zero radius, while the maximum value of one is reached by credal sets include (at least) two degenerate probability mass functions (i.e., we could sample completely disjoint data from mass functions consistent with the two BFs) corresponding to BFs assigning mass one to a non-singleton. Note also that, if the mass are assigned only to the singletons and to the universe, the radius is the mass of the universe.

5 Computational Issues and Preliminary Tests

Consider the optimization tasks required to compute (8) and (9), when based on (7). The feasible region is defined by linear constraints as in (5). Apart from the absolute values in (7), this is a linear program. Yet, the problem can be reduced to a linear task by introducing $2n$ auxiliary variables. Let us show this for the minimum of a single term of the objective function, say $|P_1(x)-P_2(x)|$. Introduce two nonnegative

[10] Notably (8) has been already proposed in [1] as a possible descriptor of the level of similarity for credal sets. Yet, we emphasize here the novelty and importance of considering both the bounds for a reliable modeling of the similarity level.

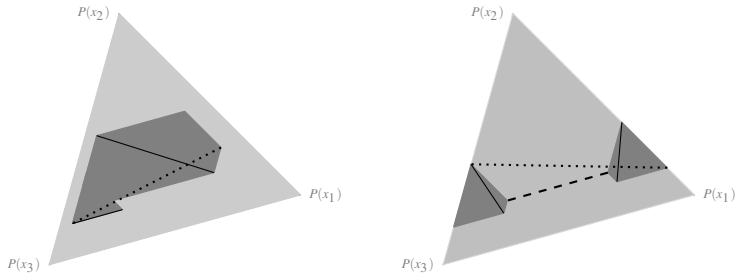


Fig. 2 Lower (dotted) and upper (dashed) distances and radiuses (continuous). Lines connect the extremes corresponding to the optima. As the distance is not Euclidean, lengths are not proportional to the actual distances. (a) compares the BFs in Figures 1, (b) a BF with $m(\{x_2\}) = .15$, $m(\{x_3\}) = .6$, $m(\{x_2, x_3\}) = .05$, $m(\mathscr{X}) = .2$ and the BF obtained by swapping x_3 and x_1.

variables Δ_+ and Δ_- such that:

$$\Delta_+ + \Delta_- = |P_1(x) - P_2(x)|. \tag{11}$$

This allows to rewrite the objective function in a linear form. We also set:

$$\Delta_+ - \Delta_- = P_1(x) - P_2(x), \tag{12}$$

this being an additional (linear) constraint. Let $(P_1(x)^*, P_2(x)^*, \Delta_+^*, \Delta_-^*)$ denote the solution of the corresponding linear task. It should be $\Delta_+^* \cdot \Delta_-^* = 0$, because otherwise it would be possible to subtract $\min\{\Delta_+^*, \Delta_-^*\} > 0$ to both the (nonnegative) auxiliary variables without violating (12), and thus obtain a smaller minimum. But if $\Delta_-^* = 0$, $\Delta_- = 0$ can be assumed in the problem, which therefore coincides with the original one (similarly with $\Delta_+^* = 0$). Overall, the computation of (8) and (9) maps to a linear program, whose solution is known to be on the extremes:

$$\underline{\delta}(m_1, m_2) := \min_{P_1(X) \in \text{ext}[K_{m_1}(X)], P_2(X) \in \text{ext}[K_{m_2}(X)]} \delta(P_1, P_2). \tag{13}$$

The measure we presented can be therefore computed by pairwise comparison of the extremes as in (13) or by solving the above derived linear program. Regarding complexity, linear programming is (roughly) cubic in the number of constraints/variables, which is at most 2^n, while the evaluation based on the extreme points is quadratic in the number of vertices (which are at most $n!$). Thus, for worst case scenarios, linear programming is faster for large n, while pairwise comparison is faster for small values (the threshold being around $n = 6$).

Some preliminary numerical tests on randomly generated BFs were performed to compare our interval-valued measure with other, singly-valued, descriptors. The results, summarized in Figure 3, suggests that our intervals are seemingly

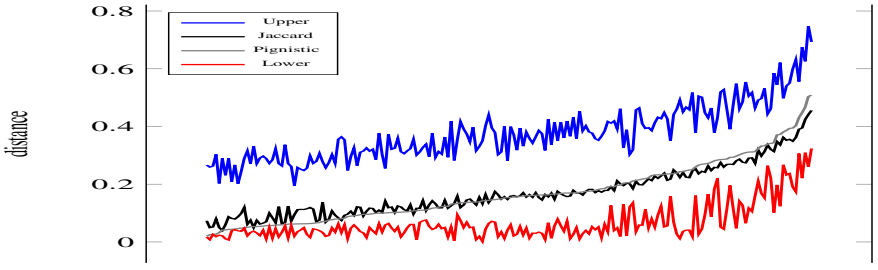


Fig. 3 Comparison between the bounds of the interval valued measure proposed in this paper, the Manhattan distance of the pignistic distributions and the distance based on the inner product of the masses with the *Jaccard index* [5]. The distances are computed on 1000 randomly generated pairs BFs defined over a ternary variable and with radius smaller than .3. Results are sorted by increasing values of the pignistic distance.

effective in including the single-valued descriptors we consider, without increasing too much their size. Thus, the desired cautiousness in the estimates is achieved without compromising the informativeness of the results.

6 Conclusions and Outlooks

A new interval-valued dissimilarity measure, together with a measure of the level of Bayesianity, has been proposed within the framework of evidence theory. The development of similar results for measures other than the Manhattan distance (KL and Euclidean in particular) should be regarded as a necessary future work. A more systematic experimental comparison with other measures should be also considered. Finally, we want to extend k-NN classification to interval-valued distances, and then apply the ideas developed in this paper to classifiers modeling instances by BFs.

References

1. Abéllan, J., Gómez, M.: Measures of divergence on credal sets. Fuzzy Sets and Systems 157, 1514–1531 (2006)
2. Cuzzolin, F.: On the Credal Structure of Consistent Probabilities. In: Hölldobler, S., Lutz, C., Wansing, H. (eds.) JELIA 2008. LNCS (LNAI), vol. 5293, pp. 126–139. Springer, Heidelberg (2008)
3. de Campos, L.M., Bolaños, M.J.: Characterization of fuzzy measures through probabilities. Fuzzy Sets and Systems 31, 23–36 (1989)
4. Dempster, A.: Upper and lower probabilities induced by multi-valued mapping. Ann. Math. Stat. 38, 325–339 (1967)
5. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53, 118–145 (2012)
6. Liu, Z., Dezert, J., Pan, Q.: A new measure of dissimilarity between two basic belief assignments. In: Proceedings of Fusion 2010. IEEE (2010)
7. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
8. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall (1991)

An Evidential Pattern Matching Approach for Vehicle Identification

Anne-Laure Jousselme and Patrick Maupin

Abstract. In this paper, we propose a novel pattern matching approach for vehicle identification based on belief functions. Distances are computed within a belief decision space rather than directly in the feature space as traditionally done. The main goal of the paper is to compare performances obtained when using several distances between belief functions recently introduced by the authors. Belief functions are modeled using the outputs of a set of modality-based 1-NN classifiers, two distinct uncertainty modeling techniques and are combined with Dempster's rule. Results are obtained on real data gathered from sensor nodes with 4 signal modalities and for 4 classes of vehicles (pedestrian, bicycle, car, truck). Main results show the importance of the uncertainty technique used and the interest of the proposed pattern matching approach in terms of performance and expressiveness.

1 Introduction

Sensor networks are widely used for monitoring and surveillance applications in which the vehicle identification task plays a crucial role (*e.g.* [3]). Although acoustic sensors are often used [1, 3] other modalities can also be considered and in particular seismic sensors. The local fusion of these modalities can be done within the evidence theory as for instance in [5]. Nevertheless, one rarely finds in the literature papers in which more than 3 modalities are used.

Pattern matching techniques have successfully been applied to vehicle identification using for instance a one-class classifier as in [7] or an evidence theory based approach as in [6]. Other examples of similar approaches can also be found in the recent survey compiled by [1]. The advantage of a pattern matching approach over a standard pattern recognition scheme is that it retrieves the closest individual objects (observed in the past) to the one currently observed. Hence, besides the class

Anne-Laure Jousselme · Patrick Maupin
Defence Research & Development Canada-Valcartier, Québec (Qc), Canada
e-mail: {Anne-Laure.Jousselme,Patrick.Maupin}@drdc-rddc.gc.ca

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 45–52.
springerlink.com

estimation of the observed object, the user has access to contextual information which has not necessary been considered in the classification process. For instance, among the k individuals retrieved by the pattern matching algorithm, 90% were white cars. An higher-level analysis would then be possible based on the retrieved cases rather than on the estimated class alone.

In this paper, we propose an evidential pattern matching (EPM) scheme for vehicle identification. The idea is to evaluate the distances between objects within the class label space considering the classification uncertainty rather than directly computing the distances in the feature space as traditionally done in pattern matching. Uncertainty patterns are then compared rather than feature patterns directly, meaning that two quite distinct feature patterns may have a quite similar uncertainty regarding their belonging to a given class or set of classes. The idea behind the proposed approach is to abstract away the measures together with their possible errors and rather consider the uncertainty patterns these measures induce within the class space.

We are particularly interested in this paper in comparing distances' behaviour on the practical use case of pattern matching for vehicle identification. The aim is to highlight some common and distinct characteristics of distances between belief functions according to two dimensions that are (1) the weighting matrices and (1) the family to which the distances belong. We test several distances recently proposed in [4]. The choice of a specific distance measure is in general not trivial as it can be governed by either axomatic or semantic properties, or by an optimization process aiming at maximizing a given system performance. Depending on the application, some formal properties are required while others are superfluous. Our position is that none of the dissimilarity measures is better than an another in the absolute, but rather that its choice should be directed by the practical use.

Section 2 sets the basic notation for evidence theory as well as the general formulations of distances proposed in [4]. Section 3 first presents the real dataset of vehicle features obtained by a sensor node with 4 modalities followed by the description of the classification scheme based on a pattern matching approach. Finally Section 3.3 presents some preliminary results obtained for the Area Under Curve (AUC) measure of performance. Section 4 draws some conclusions and outlines some future work.

2 Background

In the upcoming presentation, we use a geometrical interpretation of evidence theory together with matrix-vector notation. X is a frame of discernment of N distinct objects, x denoting any element of X. 2^X is the power set of X and $\mathscr{E}_X$ is the associated 2^N-dimensional Cartesian space. Basic Probability Assignments, belief, plausibility and communality functions are all vectors of $\mathscr{E}_X$ with specific properties, that we will denote by **m**, **Bel**, **Pl** and **q** respectively. In particular, **m** is a vector which first coordinate is 0 and its coordinates sum up to 1.

The intersection and inclusion indexes Int and Inc are defined respectively as $\text{Int}(A,B) = 1$ if $A \cap B \neq \varnothing$ and 0otherwise and $\text{Inc}(A,B) = 1$ if $A \subseteq B$ and 0 otherwise. The dual index of Int is $1 - \text{Int}$ which is such that $1 - \text{Int}(A,B) = 1$ if $A \cap B = \varnothing$ and 1 otherwise. If we denote by **Int** and **Inc** the $2^N - 1$ binary matrices whose elements are respectively $\text{Int}(A,B)$ and $\text{Inc}(A,B)$, A in rows and B in columns, we can then define the belief, plausibility and communality by their matrix notation as $\mathbf{Bel} = \mathbf{Inc}'.\mathbf{m} \quad \mathbf{Pl} = \mathbf{Int}.\mathbf{m} \quad \mathbf{q} = \mathbf{Inc}.\mathbf{m}$ where $\mathbf{Inc}'$ is the transpose matrix of **Inc**. In a recent survey paper of the distances between belief functions, we identified 3 main families of distances namely (1) the Minkowski family L_p including Manhattan ($p = 1$), Euclidean ($p = 2$) and Chebyshev distances ($p = \infty$), (2) the inner product family including the direct inner product itself (IP) and the cosine measure (cos), and (3) Fidelity family based on the Bhattacharyya coefficient which only considered measure is the Hellinger distance. We also proposed general formulations for each of these families:

$$\begin{array}{ll}
\text{Minkowski} & d_W^{(p)}(m_1,m_2) = \left(\left[(\mathbf{Um}_1 - \mathbf{Um}_2)^{\frac{p}{2}} \right]' \left[(\mathbf{Um}_1 - \mathbf{Um}_2)^{\frac{p}{2}} \right] \right)^{\frac{1}{p}} \\
\text{Inner product} & \otimes_W(m_1,m_2) = a - \mathbf{m}_1'\mathbf{W}\mathbf{m}_2 \\
\text{Cosine} & \cos_W^{(d)}(m_1,m_2) = 1 - \frac{\mathbf{m}_1'\mathbf{W}\,\mathbf{m}_2}{||\mathbf{m}_1||_W \cdot ||\mathbf{m}_2||_W} \\
\text{Fidelity 0.5} & d_W^{(H)}(m_1,m_2) = \left(b - \otimes_W^{\frac{1}{2}}(m_1,m_2) \right)^{\frac{1}{2}}
\end{array}$$

where $\mathbf{W}$ is a weighting matrix as exemplified in the second column of Table 1. a and b maximum values of the general inner product and Bhattacharyya coefficient respectively guaranteeing that the measure is positive. $||\mathbf{m}||_W = \sqrt{\otimes_W(m,m)}$ is the norm of $\mathbf{m}$.

Table 1 summarizes the distances between belief functions defined so far (gray cells) and provides the natural generalizations (white cells) hence new distances[1]. The Minkowski L_2 family is the most populated while L_1 and L_∞ have been less used. Extending the L_2 distances to the study of L_1 and L_∞ distances requires in some cases a Cholesky decomposition of the weighting matrices $\mathbf{W}$. A single cosine

Table 1 Distances between belief functions in their respective family according to several definitions of the weighting matrix $\mathbf{W}$. The distances defined so far are in gray cells while new ones are in white cells. See [4] for the complete table.

		L_p			Inner product		Fidelity
$\mathbf{W} = \mathbf{U}'\mathbf{U}$	$W(A,B)/U(A,B)$	$p=1$	$p=2$	$p=\infty$	IP	cos	(Hellinger)
I	1 iff $A = B$ (W)	$d_I^{(1)}$	$d_I^{(2)}$	$d_I^{(\infty)}$	$\otimes_I^s$	$\cos_I$	$d_I^{(H)}$
Jac	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$ (W)	$d_J^{(1)}$	$d_J^{(2)}$	$d_J^{(\infty)}$	$\otimes_J^s$	$\cos_J$	$d_J^{(H)}$
IncInc$'$	1 iff $A \subseteq B$ (U)	$d_{Inc}^{(1)}$	$d_{Inc}^{(2)}$	$d_{Inc}^{(\infty)}$	$\otimes_{Inc}^s$	$\cos_{Inc}$	$d_{Inc}^{(H)}$
Int$'$**Int**	1 iff $A \cap B \neq \varnothing$ (U)	$d_{Int2}^{(1)}$	$d_{Int2}^{(2)}$	$d_{Int2}^{(\infty)}$	$\otimes_{Int2}^s$	$\cos_{Int2}$	$d_{Int}^{(H)}$
Bet$'_x$**Bet**$_x$	$\frac{\lvert x\cap A\rvert}{\lvert A\rvert}$ (U)	$d_{Betx}^{(1)}$	$d_{Betx}^{(2)}$	$d_{Betx}^{(\infty)}$	$\otimes_{Betx}^s$	$\cos_{Betx}$	$d_{Betx}^{(H)}$
Bet$'$**Bet**	$\frac{\lvert A\cap B\rvert}{\lvert B\rvert}$ (U)	$d_{Bet}^{(1)}$	$d_{Bet}^{(2)}$	$d_{Bet}^{(\infty)}$	$\otimes_{Bet}^s$	$\cos_{Bet}$	$d_{Bet}^{(H)}$

[1] A more complete table is proposed in [4] where other weighting matrices are considered.

measure has been defined so far but a multitude of measures of this kind remains to be explored. This comment also applies to the Hellinger distance and to other distances of the Fidelity family which are based on Bhattacharyya's coefficient involving the squared-root of the BPAs.

3 Evidential Pattern Matching for Vehicle Identification

3.1 Data Description

The sensor network considered is the SASNet sensor network, designed at DRDC-Valcartier [8] (see Fig. 1). At each sensing node location, four modalities come into play. The acoustic sensor (noise detector) helps mainly in distinguishing between motorized and non-motorized objects. Among the class of non-motorized objects, the seismic sensor (vibration detector) helps to distinguish between jogging and walking pedestrians, or between either tracked or wheeled vehicles. The magnetic sensor (ferric metals detector) helps to distinguish bicycles from pedestrians or within a pedestrian group, dismounted soldiers from civilians. Generally a single sensing modality will not be sufficient to design an efficient surveillance system aimed at the detection and identification of passing by objects and a combination of them will rather be required. For instance, a high value for the three modalities generally corresponds to a truck (high noise, high vibrations, and high magnetic response). An experiment has been conducted in which 17 types of vehicles, ranging from pedestrians to trucks, transited through the sensor network for a total of 372 recorded events.

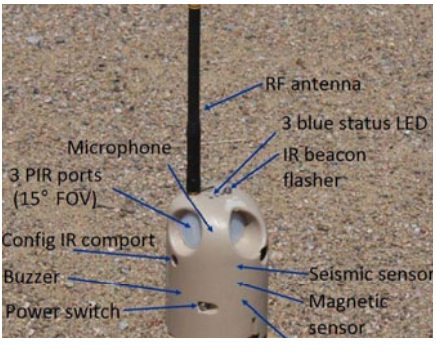


Fig. 1 The SASNet sensor node developed by DRDC-Valcartier provided the data exploited in this paper [8]. Each sensor node is sensing in four modalities: acoustic, seismic, magnetic and pyroelectric passive infrared (PIR, a motion detector). On each of the 4 types of signal, a series of 28 energy and Fourier transform-based features are computed for a total of 112 features.

3.2 Classification Scheme

Let X be the frame of discernment, *i.e.* the set of class labels possibly assigned to an observed sample. We consider the following partition as the frame of

discernment $X = \{P, B, C, T\}$ where P stands for Pedestrian, B for Bicycle, C for Car and T for Truck. The basic classification architecture is composed of 4 parallel 1-Nearest Neighbour (1-NN) classifiers g_a, g_s, g_m and g_p differing in their set of 28 features extracted from the acoustic (a), seismic (s), magnetic (m) and PIR (p) modality respectively. The output of each classifier is transformed into a belief function Bel_a, Bel_s, Bel_m and Bel_p and then combined through Dempster's rule, *i.e.* $\text{Bel} = \text{Bel}_a \otimes \text{Bel}_s \otimes \text{Bel}_m \otimes \text{Bel}_p$. Two different uncertainty modelers (*i.e.* belief function constructors) are considered at the output of the classifiers, namely the Consonant Likelihood Based model (CLB) proposed by Shafer in [9], and the q-Least Committed belief function (q-LC) (*e.g.* [2]). The dataset of samples is split into a training set X_R which will serve as reference, as well as a test set X_Q. According to the model above, each sample $\mathbf{x}_r$ of X_R and $\mathbf{x}_q$ of X_Q is assigned with a belief function Bel_r for $r = 1, \ldots, |X_R|$ and Bel_q for $q = 1, \ldots, |X_Q|$. For each sample $\mathbf{x}_q$ of X_Q to be classified, the purpose of the pattern matching approach is then to find the best fit between its representative belief function Bel_q and one of the reference set Bel_r. The decision function is then:

$$\text{Class}(\mathbf{x}_q) = \arg \min_{r=1,\ldots,|X_R|} d(\text{Bel}_r, \text{Bel}_q) \tag{1}$$

where d is a distance measure. Note that rather than a minimum function in (1) a majority voting could be considered, the observed object being thus assigned the class represented as a majority among k known objects. Further analysis may be then performed on the set of the retrieved individuals, no decision being required in this case.

3.3 Results

All the results provided in the three tables below show the Area Under Curve (AUC) transformed into a measure of error (*i.e.* the lower the better) obtained over 30 iterations of an hold-out procedure with 90% of data for training and 10% for testing. As a reference, standard classification results are presented in Table 2[2]: Each of the first four columns of the top most part of the table corresponds to a 1-NN classifier built upon each of the four modalities; the 6 following bottom most columns correspond to the outputs of these 4 previous classifiers combined according to 6 classical fusion schemes; the last column corresponds to an Oracle[3]. Finally, the last column of the top most part of the table is a raw combination of the 112 features through a 1-NN and corresponds to the pattern matching scheme in the feature space (PM

[2] Other classification schemes could have been considered such as evidential ones but this kind of study is out of the scope of this paper which main purpose is to setup the bases of a pattern matching approach.

[3] The Oracle is an ideal combiner which outputs the true class as soon as the latter appears in at least one of the decisions of the 4 classifiers to be combined. These results should thus not read as genuine classifier results, but rather used for comparison purposes to an ideal situation.

Table 2 AUC in standard classification. The best mean value is in bold.

	Single modalities				All features
	Acoustic	Seismic	Magnetic	PIR	(PM in FS)
Minimum	0.259	0.197	0.237	0.429	0.247
Maximum	0.363	0.326	0.342	0.489	0.345
Mean	0.299	0.255	0.306	0.459	0.304

	Combiners						
	Product	Mean	Median	Maximum	Minimum	Majority	ORACLE
Minimum	0.183	0.176	0.182	0.300	0.213	0.161	0.025
Maximum	0.317	0.314	0.333	0.386	0.317	0.249	0.063
Mean	0.257	0.253	0.267	0.323	0.267	**0.200**	0.044

in FS) which should be considered as the pattern matching results of reference. The best result (in bold) is obtained by combining the outputs of the 4 modality-based classifiers by a majority vote.

Tables 3 and 4 show results for the EPM scheme, one table for each of the two uncertainty modelisation CLB and q-LC respectively. The results are obtained over 30 iterations of an hold-out procedure with 90% of data for training (*i.e.* the reference set) and 10% for testing (set of queries). In the tables, the minimum, maximum and mean values are provided for each of the 36 distances considered in Table 1. We highlighted in bold the best weighting matrix (according to the mean) for a fixed kind of distance, while underlined the best kind of distance for a fixed weighting matrix.

In the light of these statistics ventilated into Tables 3 and 4, it seems that the uncertainty modeling method plays a crucial role in the improvement of the AUC performance figures. The best result of Table 3 is 0.272 obtained for both $d_J^{(2)}$ and $\cos_J$, while for Table 4 it is 0.265 obtained for both $\otimes_{Betx}^{(d)}$ and $\otimes_{Bet}^{(d)}$.

If we adopt an optimistic analysis of the results above and look at the minimum values of AUC obtained over 30 replications, we observe that the overall minimum value for the CLB uncertainty modeling (Tab. 3) is obtained for $\otimes_{Inc}^{(d)}$ with 0.241,

Table 3 AUC for the EPM scheme with CLB uncertainty modelisation.

	Minkowski family								
	L_1			L_2			L_∞		
	min	max	mean	min	max	mean	min	max	mean
I	0.266	0.333	0.291	0.266	0.333	0.302	0.266	0.350	0.308
Jac	0.249	0.333	0.291	0.249	0.315	**0.272**	0.269	0.315	**0.285**
IncInc′	0.252	0.315	**0.279**	0.252	0.333	0.285	0.249	0.333	**0.285**
Int′Int	0.252	0.315	**0.279**	0.252	0.333	0.285	0.249	0.333	**0.285**
Bet$'_\gamma$Bet$_\gamma$	0.271	0.315	0.291	0.271	0.315	0.291	0.271	0.315	0.291
Bet′Bet	0.271	0.315	0.291	0.271	0.315	0.291	0.271	0.315	0.291

	Inner product family						Fidelity family		
	IP			cos			Hellinger		
	min	max	mean	min	max	mean	min	max	mean
I	0.281	0.368	0.335	0.249	0.333	0.285	0.252	0.333	**0.279**
Jac	0.259	0.381	0.325	0.249	0.315	**0.272**	0.253	0.333	0.285
IncInc′	0.241	0.354	0.283	0.252	0.333	0.285	0.252	0.333	**0.279**
Int′Int	0.384	0.512	0.454	0.269	0.315	0.285	0.315	0.464	0.406
Bet$'_\gamma$Bet$_\gamma$	0.252	0.315	**0.274**	0.253	0.315	0.285	0.252	0.333	**0.279**
Bet′Bet	0.252	0.315	**0.274**	0.271	0.315	0.291	-	-	-

Table 4 AUC for the EPM scheme with q-LC uncertainty modelisation.

	Minkowski family								
	L_1			L_2			L_∞		
	min	max	mean	min	max	mean	min	max	mean
I	0.269	0.299	0.289	0.286	0.299	**0.294**	0.286	0.320	0.303
Jac	0.278	0.337	0.301	0.269	0.354	0.308	0.249	0.299	**0.273**
IncInc′	0.249	0.333	**0.278**	0.266	0.354	0.297	0.266	0.303	0.279
Int′Int	0.249	0.333	**0.278**	0.266	0.354	0.297	0.266	0.303	0.279
Bet$_\lambda'$Bet$_\chi$	0.249	0.354	0.291	0.266	0.354	0.297	0.249	0.354	0.285
Bet′Bet	0.249	0.354	0.291	0.266	0.354	0.297	0.249	0.354	0.291
	Inner product family						**Fidelity family**		
	IP			cos			Hellinger		
	min	max	mean	min	max	mean	min	max	mean
I	0.347	0.498	0.398	0.266	0.303	**0.283**	0.252	0.354	**0.292**
Jac	0.347	0.410	0.387	0.269	0.320	0.296	0.283	0.336	0.309
IncInc′	0.241	0.315	0.276	0.266	0.354	0.297	0.305	0.340	0.317
Int′Int	0.401	0.485	0.454	0.266	0.354	0.297	0.430	0.460	0.447
Bet$_\lambda'$Bet$_\chi$	0.227	0.315	**0.265**	0.249	0.354	0.291	0.297	0.304	0.300
Bet′Bet	0.227	0.315	**0.265**	0.249	0.354	0.291	-	-	-

whereas the overall minimum value for the q-LC uncertainty modeling (Tab. 4) is 0.227 obtained for $\otimes^{(d)}_{Betx}$ and $\otimes^{(d)}_{Bet}$. Again, the q-LC uncertainty modeling together with the $\otimes^{(d)}_{Betx}$ and $\otimes^{(d)}_{Bet}$ outperform the results of Tab. 3.

When comparing results from tables 3 and 4 to the ones of Table 2, one might argue that a simple majority vote on individual modality trained classifiers (0.2) outperforms the best EPM result (0.265). But we should rather compare two pattern matching approaches one in the feature space, and the other in class labels belief space, and we indeed observe an improvement of the performance.

4 Conclusions and Future Works

We presented a preliminary study of distances' behaviour on an evidential pattern matching (EPM) scheme performed on real data. These preliminary results are encouraging as the basic scheme for the EPM approach may be improved by for instance (1) considering other individual classifiers than 1-NNs, (2) improving the feature selection part, (3) exploring other uncertainty modeling methods as we observed its impact on the performances. Beyond the proposed pattern matching scheme, we highlighted the fact (1) that the choice of a distance measure is application-dependent and that no prior evaluation is really valid; and (2) that other measures than the ones traditionally used may be of interest, in particular the inner product family would be worth to be studied more deeply, as its computational cost is lower than the other families.

Besides the maybe not so convincing classification results, the EPM offers the advantage to retrieve past known cases on the basis of their associated uncertainty profile together with their contextual information for a higher-level analysis. Considering that humans are monitoring wide areas through the SASNet sensor network, providing them a richer uncertainty representation together with contextual information is of great interest for building a human-machine interface with the SASNet

sensor system. In future works, we intend to apply this EPM scheme to information retrieval (IR) problems and use IR performance measures to assess the distances' behaviour.

References

1. Aljaafreh, A., Al-Fuqaha, A.: Multi-target classification using acoustic signatures in wireless sensor networks: A survey. Signal Processing - An International Journal (SPIJ) 4(4), 175–246 (2010)
2. Aregui, A., Denoeux, T.: Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. International Journal of Approximate Reasoning 49, 575–594 (2008)
3. Guo, B., Nixon, M.S., Damarla, T.: Improving acoustic vehicle classification by information fusion. Pattern Analysis and Applications 15(1), 29–43 (2012)
4. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53(2), 118–145 (2012)
5. Liu, C.T., Huo, H., Fang, T., Li, D.R., Shen, X.: Classification fusion in wireless sensor networks. Acta Automatica Sinica 32, 947–955 (2006)
6. Mercier, D., Lefèvre, E., Jolly, D.: Object association with belief functions, an application with vehicles. Information Sciences 181(24), 5485–5500 (2011)
7. Munroe, D.T., Madden, M.G.: Multi-class and single-class classification approaches to vehicle model recognition from images. In: Proceedings of AICS 2005: Irish Conference on Artificial Intelligence and Cognitive Science, Portstewart (2005)
8. Ricard, B., Fournier, J.: The SASNet system: Military UGS development in Canada. In: NATO Conference SET-176 - Multi-Sensor Integration for ISR Applications (2011)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)

A Comparison between a Bayesian Approach and a Method Based on Continuous Belief Functions for Pattern Recognition

Anthony Fiche, Arnaud Martin, Jean-Christophe Cexus, and Ali Khenchaf

Abstract. The theory of belief functions in discrete domain has been employed with success for pattern recognition. However, the Bayesian approach performs well provided that once the probability density functions are well estimated. Recently, the theory of belief functions has been more and more developed to the continuous case. In this paper, we compare results obtained by a Bayesian approach and a method based on continuous belief functions to characterize seabed sediments. The probability density functions of each feature of seabed sediments are unimodal and estimated from a Gaussian model and compared with an α-stable model.

1 Introduction

The theory of belief functions, introduced by Dempster [4] and formalized by Shafer [13], has found in these recent years many applications especially in pattern recognition. The Bayesian approach performs well provided that once the probability density functions (pdfs) are well estimated. However, the Bayesian approach introduces the notion of prior probabilities. It is possible to avoid this problem by using the theory of belief functions. The theory of belief functions is often presented as an extension of the probability theory. However, the theory of belief functions is not often been used in problem of estimation. Recently, many papers [5, 16] have been proposed to extend the theory of belief functions in discrete domain to

Anthony Fiche · Jean-Christophe Cexus · Ali Khenchaf
ENSTA Bretagne, 2 rue François Verny, 29806 Brest Cedex 9, France
e-mail: {anthony.fiche,jean-chirstophe.cexus,
ali.khenchaf}@ensta-bretagne.fr

Arnaud Martin
UMR 6074 IRISA, Université de Rennes 1, rue Édouard Branly BP 30219,
22302 Lannion Cedex, France
e-mail: arnaud.martin@univ-rennes1.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 53–60.
springerlink.com

continuous domain. In [1, 11], the authors proposed solutions to solve problem of pattern recognition from continuous belief functions.

We propose a supervised classification of seabed sediments based on a Bayesian approach and compared with a method based on the theory of continuous belief functions. The pdfs of each seabed sediment are bell-shaped [1]. Many distributions can have this property: Gaussian, Weibull, K .... However, the pdfs from seabed sediments have the properties of skewness and heavy tails. A distribution is said to have heavy tails if the tails decays slower than the tail of the Gaussian distribution. Therefore, the property of skewness means that it is impossible to find a mode where the curve is symmetric. It is possible to consider these contraints from α-stable distribution. Consequently, we use two models of estimation during the classification: Gaussian and α-stable distributions.

The remainder of this paper is organized in the following manner. In section 2, we introduce the theory of continuous belief functions. In section 3, we describe the data set, the model of estimation and compare results between the Bayesian approach and the method based on continuous belief functions.

2 Background on Continuous Belief Functions

2.1 Basic Belief Density

Recently, Smets [16] extended the definition of belief functions to the set of reals $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and basic belief assignment (bbd) are only attributed to intervals of $\overline{\mathbb{R}}$. Let us consider $\mathscr{I} = \{[x,y],(x,y],[x,y),(x,y);x,y \in \overline{\mathbb{R}}\}$ as a set of closed, half-opened and opened intervals of $\overline{\mathbb{R}}$. A bbd $m^{\mathscr{I}}(x,y)$ linked to a specific *pdf* is a non negative function on $\mathscr{I}$ such that $m^{\mathscr{I}}(x,y) = 0$ if the interval defined by (x,y) is not closed in $\mathscr{I}$. The closed intervals $[x,y]$ which satisfy the relation $m^{\mathscr{I}}(x,y) > 0$ are called focal elements. From the definition of the bbd, it is possible to define others belief functions [16] as in the discrete case credibility function $bel^{\overline{\mathbb{R}}}$, plausibility function $pl^{\overline{\mathbb{R}}}$ and communality function $q^{\overline{\mathbb{R}}}$. A bbd is said to be "consonant" when focal elements are nested. Focal elements I_u can be labeled as an index u such that $I_u \subseteq I'_u$ with $u' > u$.

2.2 Least Commitment bbd Induced by an Unimodal pdf

The definition of pignistic probability [14] for $a < b$ is:

$$Bet f([a,b]) = \int_{x=-\infty}^{x=+\infty} \int_{y=x}^{y=+\infty} \frac{\min(y,b) - \max(x,a)}{y-x} m^{\mathscr{I}}(x,y)dxdy \qquad (1)$$

It is possible to calculate pignistic probabilities to have basic belief densities. However, many basic belief densities exist for one same pignistic probability. To resolve

[1] *i.e.* the probability density function is unimodal with a mode μ, continuous and strictly monotonous increasing (decreasing) at left (right) of the mode

this issue, we can use the consonant basic belief density. This definition is used to apply the least commitment principle [15], which consists in choosing the least informative belief function when a belief function is not totally defined and is only known to belong a family of functions. The function $Bet f$ can be induced by a set of isopignistic belief functions $\mathscr{B}\mathbf{iso}(Bet f)$. Many papers [12, 16, 1] deal with the particular case of continuous belief functions with nested focal elements. The least commitment principle proposes to choose the least informative mass function, *i.e.* the mass functions must be ordered. An order relation is given in equation 2, but there are other order relations.

$$(\forall A \subseteq \overline{\mathbb{R}}, q_1^{\overline{\mathbb{R}}}(A) \leq q_2^{\overline{\mathbb{R}}}(A)) \Rightarrow (m_1^{\overline{\mathbb{R}}} \leq m_2^{\overline{\mathbb{R}}}) \tag{2}$$

For example, Smets [16] proved that the basic belief assignment $m^{\overline{\mathbb{R}}}$ attributed to an interval $I = [x,y]$ with $y > \mu$ related to a bell-shaped pignistic probability function with a mode μ is determined by [2]:

$$m^{\overline{\mathbb{R}}}([x,y]) = \theta(y)\delta(x - \gamma(y)) \tag{3}$$

with $x = \gamma(y)$ satisfying $Bet f(\gamma(y)) = Bet f(y)$ and $\theta(y)$:

$$\theta(y) = (\gamma(y) - y)\frac{dBet f(y)}{dy} \tag{4}$$

The build basic belief assignment $m^{\overline{\mathbb{R}}}$ is consonant and belongs to the set $\mathscr{B}\mathbf{iso}(Bet f)$.

2.3 Link between Pignistic Probability Function and Plausibility Function in $\overline{\mathbb{R}}$

The available information are the conditioned pignistic density $Bet f[C_i]$ with $C_i \in \Theta$, where Θ is called the frame of discernement. The function $Bet f[C_i]$ is supposed to be bell-shaped. The plausibility function from a bbd $m^{\overline{\mathbb{R}}}$ with $x > \mu$ is obtained by an integral of equation (4) between $[x, +\infty[$:

$$pl^{\overline{\mathbb{R}}}[C_i](I) = \int_x^{+\infty} (\gamma(t) - t)\frac{dBet f(t)}{dt}dt \tag{5}$$

By assuming that $Bet f$ is symmetrical, an integration by parts can simplified the equation (5):

$$pl^{\overline{\mathbb{R}}}[C_i](I) = 2(x - \mu)Bet f(x) + 2\int_x^{+\infty} Bet f(t)dt \tag{6}$$

We can calculate $\int_x^{+\infty} Bet f(t)dt$ in a particular case of symmetrical $Bet f$ by using the Chasles' theorem. Consequently, the equation (6) can be simplified [7]:

[2] δ refers to the Dirac's measure.

$$pl^{\overline{\mathbb{R}}}[C_i](I) = 2(x-\mu)pdf(x) + 2(1-cdf(x)) \tag{7}$$

If $x < \mu$, we use the variable modification $x = 2\mu - y$. In the particular case of Gaussian pdf, Caron *et al.* [1] propose the plausibility function:

$$pl^{\overline{\mathbb{R}}}[C_i](I) = 1 - F_3((x-\mu)(\Sigma)^{-1}(x-\mu)) \tag{8}$$

The function F_{d+2} is a cumulative density function of the χ^2 distribution with 3 degrees of freedom, μ the mean and Σ the standard-deviation of a Gaussian pdf. It is difficult to generalize in the case of asymmetric pdf because the function $\gamma(y) = x$ satisfying $Betf(\gamma(y)) = Betf(y)$ is not trivial. The plausibility function related to an interval $I_1 = [x_1, y_1]$ is defined by the area defined under the α-cut such as $\alpha = Betf(x_1)$ (Figure 1):

$$pl^{\overline{\mathbb{R}}}[C_i](I_1) = \int_{-\infty}^{x_1} Betf(t)dt + (y_1 - x_1)Betf(x_1) + \int_{y_1}^{+\infty} Betf(t)dt \tag{9}$$

In general, we know only one point y_1. We estimate numerically x_1 such that $pdf(y_1) = pdf(x_1)$. Finally, the plausibility function related to the interval I_1 is:

$$pl^{\overline{\mathbb{R}}}[C_i](I_1) = 1 + cdf(x_1) - cdf(y_1) + (y_1 - x_1)pdf(x_1) \tag{10}$$

In classification, we assume that we have several pdfs associated to a class C_i. We can calculate a plausibility function related to its pdfs by using the least commitment principle. Several plausibility functions can be combined by using the general Bayes theorem [15, 3] to calculate mass functions allocated to A of an interval I:

$$m^{\overline{\mathbb{R}}}[x](A) = \prod_{C_j \in A} pl_j(x) \prod_{C_j \in A^c} (1 - pl_j(x)) \tag{11}$$

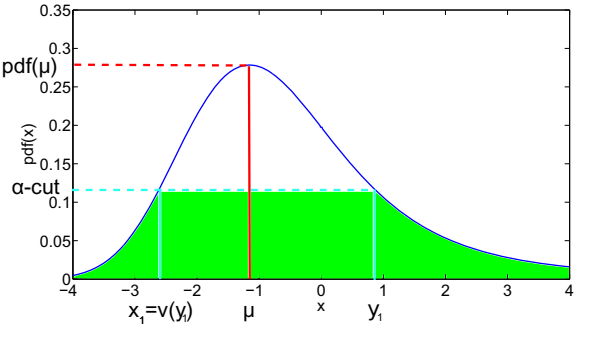


Fig. 1 Plausibility function in the case of asymmetric pdf.

3 Application to Pattern Recognition

3.1 Data Set

The data set are picked up by the Service Hydrographique et Océanique de la Marine (SHOM) with the Daurade Autonomous Underwater Vehicle (AUV) from the Atlas DESO 35 mono-beam echo sounder in the Mediterranean Sea off the coast of Toulon. Raw data represents an echo signal amplitude according to time. These data are processed to obtain some features, which have been normalized between [0,1] (defined and used in the Quester Tangent Corporation (QTC) software [2]). The frame of discernment is $\Theta = \{\text{rock}, \text{sand}, \text{silt}\}$, with 6017 samples from rock, 7338 samples from sand and 4853 samples from silt. From the data, we choose the features called the "third quantile calculated on echo signal amplitude" and the "75th quantile calculated on cumulative energy". The authors would like to thank the Service Hydrographique et Océanique de la Marine (SHOM) for the data and G. Le Chenadec for his advices about the data.

3.2 Models of Estimation

We use two models of estimation: Gaussian and α-stable distributions. The Gaussian distribution is a particular case of α-stable distribution [10]. Several equivalent definitions have been suggested in the literature to parametrize an α-stable distribution from its characteristic function [17, 18]. Zolotarev [18] proposed the following:

$$\phi(t) = \begin{cases} \exp(it\nu - |\gamma t|^{\alpha}[1 + i\beta \tan(\frac{\pi\alpha}{2}) sign(t)(|t|^{1-\alpha} - 1)]) & \text{if } \alpha \neq 1 \\ \exp(it\nu - |\gamma t|[1 + i\beta \frac{2}{\pi} sign(t) \log|t|]) & \text{if } \alpha = 1 \end{cases} \quad (12)$$

with $\alpha \in]0,2]$ is the characteristic exponent, $\beta \in [-1,1]$ is the skewness parameter, $\gamma \in \mathbb{R}^{+*}$ represents the scale parameter and $\nu \in \mathbb{R}$ is the location parameter. In general, the notation $S_{\alpha}(\beta, \gamma, \nu)$ refers to α-stable distributions.

The α-stable pdf, noticed pdf_{α}, is obtained by calculating the Fourier transform of its characteristic function (*cf.* [9] for the implementation). An α-stable random variable can be estimated by using methods based on quantiles or moments. For the rest of the paper, we use a method based on moments developed by Koutrouvelis [8] in order to estimate the parameters α, β, γ and ν.

To implement the classification with the belief functions, we firstly need to estimate the parameters of distribution from the learning base. For each feature of vectors belonging to the test base, the plausibility functions for each class are then calculated from equation (10). These plausibility functions are combined from equation (11) to obtain two mass functions. These two mass functions are combined by the conjunctive combination (we stay in open-world). Indeed, m_1 and m_2 and $\forall X \in 2^{\Theta}$:

$$m(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{13}$$

The decision is finally made by using the maximum of the pignistic probabilities.

3.3 Results

The two features are considered as a source of information. 5000 samples are randomly selected for the data set. Half the samples are used for the learning base and the rest for the test base. For the two approaches, the parameters of each model are estimated from the learning base. For the Bayesian approach, we need to estimate the prior probabilities $p(C_i)$ from the learning base approach. For each seabed sediment, the prior probabilities correspond to the proportion of seabed sediments in the learning base. The application of Bayes theorem gives posterior probabilities:

$$p(C_i/x) = \frac{p(x/C_i)p(C_i)}{\sum_{i=1}^{n} p(x/C_i)p(C_i)} \tag{14}$$

Finally, the decision is chosen by using the maximum of the posterior probabilities.

We can observe that the assumption of the α-stable model can easily accommodate the data compared to the Gaussian model (Figure 2). For each model and each method, we can observe that there is confusion between sand and silt (Table 1,2,3,4). Indeed, these sediments have similar properties. With the Gaussian models, we can observe that the theory of belief functions (Table 2)

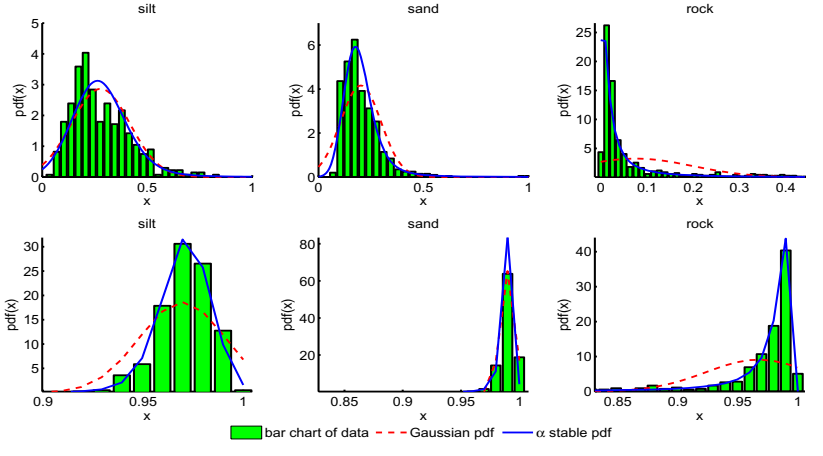


Fig. 2 Empirical pdfs and its estimations (The first row corresponds to the feature called "third quantile calculated on echo signal amplitude" and the second row corresponds to the feature called "25th quantile calculated on cumulative energy").

Table 1 Confusion matrix of seabed classification results based on the Bayesian approach with the Gaussian model

Ground truth seabed type	Predicted Seabed Type rock	sand	silt
rock	8.48 %	23.00 %	1.28 %
sand	0.00 %	37.32 %	2.80 %
silt	0.36 %	11.32 %	15.44 %

Table 2 Confusion matrix of seabed classification results based on the theory of belief functions with the Gaussian model

Ground truth seabed type	Predicted Seabed Type rock	sand	silt
rock	32.40 %	0.00 %	0.36 %
sand	12.44 %	20.92 %	6.76 %
silt	7.20 %	2.32 %	17.60 %

Table 3 Confusion matrix of seabed classification results based on the Bayesian approach with the α-stable model

Ground truth seabed type	Predicted Seabed Type rock	sand	silt
rock	28.28 %	0.04 %	4.44 %
sand	0.00 %	34.88 %	5.24 %
silt	0.84 %	6.76 %	19.52 %

Table 4 Confusion matrix of seabed classification results based on the theory of belief functions with the α-stable model

Ground truth seabed type	Predicted Seabed Type rock	sand	silt
rock	26.48 %	0.00 %	6.28 %
sand	0.00 %	29.84 %	10.28 %
silt	0.52 %	2.48 %	24.12 %

(classification accuracy of 70.92 %) give better results compared to the Bayesian approach (Table 1) (classification accuracy of 61.24 %). The belief functions take into account the imprecision of data introduced by the Gaussian model. The α-stable model gives better results compared to the Gaussian model because the α-stable can easily accommodate the data compared the Gaussian model. However, the Bayesian approach (Table 3) (classification accuracy of 82.68 %) gives better results than the belief functions (Table 4) (classification accuracy of 80.44 %) with the α-stable model but not significantly. We can explain these phenomena by the fact we introduce more information with the prior probability. The Bayesian approach performs well provided that once the probability density functions are well estimated. However, the probability density functions are poorly estimated. The theory of belief functions takes into account of imprecision/uncertainty during the learning step.

3.4 Conclusion

In this paper, we show the interest in using the theory of belief functions compared to a Bayesian approach in classification, especially to model imprecision of data. The problem with the Bayesian approach is that we introduce the *prior* probability We show the interest to use the α-stable model compared to the Gaussian model to estimate data from a mono-beam echo sounder. However, the proposed approach is limited to the unimodal case. In [6], the authors deal with the problem of the belief functions linked to a multimodal pdf.

References

1. Caron, F., Ristic, B., Duflos, E., Vanheeghe, P.: Least Committed basic belief density induced by a multivariate Gaussian pdf. International Journal of Approximate Reasoning 48(2), 419–436 (2008)
2. Caughey, D., Prager, B., Klymak, J.: Sea bottom classification from echo sounding data. Quester Tangent Corporation, Marine Technology Center, British Columbia, V8L 3S1, Canada (1994)
3. Delmotte, F., Smets, P.: Target identification based on the transferable belief model interpretation of Dempster–Shafer model. IEEE Transactions on Systems, Man, and Cybernetics 34(4), 457–471 (2004)
4. Dempster, A.: Upper and Lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Denœux, T.: Extending stochastic ordering to belief functions on the real line. Information Science 179(9), 1362–1376 (2009)
6. Doré, P.E., Martin, A., Khenchaf, A.: Constructing of a consonant belief function induced by a multimodal probability density function. In: COGnitive Systems with Interactive Sensors (COGIS 2009), Paris (2009)
7. Fiche, A., Martin, A., Cexus, J.C., Khenchaf, A.: Continuous belief functions and α-stable distributions. In: International Conference on Information Fusion, Edinburgh, United Kingdom (2010)
8. Koutrouvelis, I.A.: An iterative procedure for the estimation of the parameters of stable laws. Communications in Statistics-Simulation and Computation 10(1), 17–28 (1981)
9. Nolan, J.P.: Numerical calculation of stable densities and distribution functions. Communications in Statistics-Stochastic Models 13(4), 759–774 (1997)
10. Nolan, J.P.: Stable Distributions - Models for Heavy Tailed Data, ch. 1 (in progress, 2012), academic2.american.edu/~jpnolan
11. Ristic, B., Smets, P.: Belief function theory on the continuous space with an application to model based classification. In: Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU, pp. 4–9 (2004)
12. Ristic, B., Smets, P.: Target classification approach based on the belief function theory. IEEE Transactions on Aerospace and Electronic Systems 42(2), 574–583 (2005)
13. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
14. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. Uncertainty in Artificial Intelligence 5, 29–39 (1990)
15. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1), 1–35 (1993)
16. Smets, P.: Belief functions on real numbers. International Journal of Approximate Reasoning 40(3), 181–223 (2005)
17. Taqqu, M.S., Samorodnisky, G.: Stable non-gaussian random processes. Chapman and Hall (1994)
18. Zolotarev, V.M.: One-dimensional stable distributions. Translations of Mathematical Monographs, vol. 65. American Mathematical Society (1986); Translation from the original 1983 Russian edition

Prognostic by Classification of Predictions Combining Similarity-Based Estimation and Belief Functions

Emmanuel Ramasso, Michèle Rombaut, and Noureddine Zerhouni

Abstract. Forecasting the future states of a complex system is of paramount importance in many industrial applications covered in the community of Prognostics and Health Management (PHM). Practically, states can be either continuous (the value of a signal) or discrete (functioning modes). For each case, specific techniques exist. In this paper, we propose an approach called EVIPRO-KNN based on case-based reasoning and belief functions that jointly estimates the future values of the continuous signal and of the future discrete modes. A real datasets is used in order to assess the performance in estimating future break-down of a real system where the combination of both strategies provide the best prediction accuracies, up to 90%.

1 Introduction

Forecasting the future states of a complex system is a complicated task that arised in many industrial applications covered in the community of Prognostics and Health Management (PHM) such as locomotive's health prediction [1], analysis of fleet of vehicles [2] and turbofan engine monitoring [3]. Continuous states generally represent the value of a signal (an observation or a feature) and their prediction can be made by Kalman-like procedures or by neural networks [4, 5], Discrete states generally depict functioning modes reflecting the current degradation and its prediction can be performed by state machines such as Hidden Markov Models [6]. In

E. Ramasso · N. Zerhouni
FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, Automatic Control and Micro-Mechatronic Systems Department, 24 rue Alain Savary, F-25000 Besançon, France
e-mail: emmanuel.ramasso@femto-st.fr, noureddine.zerhouni@ens2m.fr

M. Rombaut
GIPSA-lab, UMR CNRS 5216 - UJF, Signal and Images Department, 38000 Grenoble, France
e-mail: michele.rombaut@gipsa-lab.inpg.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 61–68.
springerlink.com

both cases, data-driven prognostics generally involves a training procedure where statistical models of the degradation are built. To cope with the problem of lack of knowledge in PHM, case-based reasoning (CBR) was proposed as an alternative to perform prognostics. For example, the method described in [3] demonstrated better performance than neural network for continuous state prediction in a turbofan engine. For that, historical instances of the system - with condition data and known failure time - were used to create a library of degradation models. Then, for a test instance of the same system, the similarity between it and the degradation models was evaluated generating a set of *Remaining Useful Life* (RUL) estimates which were finally aggregated by a density estimation method. The main problem with the approach described in [3] is the number of parameters that has to be estimated in order to apply it. Moreover, several parts of the algorithm relied on statistical learning procedures requiring large amount of data.

In this paper, we propose an algorithm called EVIPRO-KNN that requires a training dataset composed of trajectories (historical information) plus uncertain knowledge about the possible states and has the following characteristics:

EVIPRO-KNN is a new prognostics approach based on belief functions: A trajectory similarity-based approach based on belief functions is proposed for prognostics. Belief functions were justly proposed to cope with lack of data in data representation, combination and decision-making [7, 8, 9].

EVIPRO-KNN takes into account partial labelling on states: In some applications, the training dataset is composed of continuous trajectories and of a set of labels reflecting the current system state. If these labels are known only partially, then belief functions can be used [10].

EVIPRO-KNN manages trajectories with different temporal length: The weighted sum of trajectories used to compute the prediction of observations requires trajectories with the same length, that is generally false in most of applications. We described two approaches to solve it.

EVIPRO-KNN is able to predict jointly continuous and discrete states: The prediction of the future sequence of states is performed jointly with the prediction of continuous observations. These sequences allow the user to have access to the online segmentation of the current observed data and generate accurate estimate of the Remaining Useful Life (RUL) of the system. As far as we know, the joint prediction of discrete states and of continuous observations was not considered jointly in PHM applications nor in CBR-based prediction.

2 Background

At each time t, an observation vector X_t can be extracted from the observed system. This system can be in one of the possible discrete states ω belonging to a set of S exhaustive and exclusive states $\Omega = \{\omega_1, \ldots, \omega_S\}$. The states can be imprecise and uncertain due to *aleatory uncertainty* induced by the variability in observations and to *epistemic uncertainty* induced by lack of knowledge. For that, we describe the knowledge of states at time t by a belief function [7, 8, 9].

The basis in the theory of belief functions is the basic belief assignment (BBA) defined by: $m_t : 2^{\Omega} \rightarrow [0,1]$, $S \mapsto m_t(S)$, with $\sum_{A \subseteq \Omega} m_t(A) = 1$. The belief mass $m_t(A)$ represents the uncertainty (since $m_t(A) \in [0,1]$) and imprecision (since A is a subset with cardinality $|A| \geq 1$) about the possible state of the system at t. Subset A is composed of unions of singletons ($\omega \in \Omega$) and thus represents explicitly the *doubt* concerning the value of the state.

The training dataset used in EVIPRO-KNN is denoted $\mathscr{L} = \{T_i\}_{i=1}^{N}$ and is composed of N trajectories T_i defined by both a sequence of Q-dimensional observation vectors $X_t \in \Re^Q$ and their associated states $T_i = \{(X_t^i, m_t^i)\}_{t=t_i}^{t_i+|T_i|}$. The i-th (continuous) trajectory begins at time t_i and finishes at time $t_i + |T_i|$ where $|T_i|$ is the length of T_i. With each trajectory T_i is associated a set of blocks $\mathbb{B}_i$ where each block B_i^j in this set corresponds to a sub-trajectory of length W: $B_j^i = \{(X_t^i, m_t^i)\}_{t=c_j}^{c_j+W}$, where $c_j \in [t_i, (t_i + |T_i| - W)]$ is the starting time of the j-th block. The number of blocks (and the range of index j) in the i-th trajectory depends on the length of the latter.

In some applications, the training dataset is composed of features and of a set of labels reflecting the current system's state. If the labels are known only partially, then belief functions can be used [10]. The state can thus be known with uncertainty and imprecision and can be described by a belief mass denoted $m_t^i, \forall i = 1 \ldots N$ and defined on the set of states Ω.

3 EVIPRO-KNN Algorithm

Let now consider that a block of data $Y_t \in \Re^Q$ of length W is available (obtained from sensors located on the system). Given the training dataset and this observation, the goal is to predict an observation trajectory $\hat{T}_t = \{(\hat{X}_{t'}, \hat{m}_{t'})\}_{t'=t}^{t+H}$ where H is an horizon of prediction. The value of H will be set automatically as shown in the sequel.

Step 1 - K-best Trajectories Determination: In this step, the K nearest trajectories to observations Y_t are determined. For that, all trajectories in the training dataset $\mathscr{L}$ are scanned. For each trajectory T_i, the nearest block $B_{j^*}^i \in \mathbb{B}_i$ to the observation block Y_t is found. Index j^* of the best block $B_{j^*}^i$ in the i-th trajectory is given by: $j^* = \text{argmin}_{j, B_j^i \in \mathbb{B}_i} \mathscr{D}(Y_t, B_j^i)$. Note that all distances $\mathscr{D}$ are measured using the Euclidean one as in most of the KNN-based algorithms [3]. Let denote c_i^* the starting time of best block $B_{j^*}^i$ in the i-th trajectory. When the best block in each trajectory has been found, all best blocks are sorted by ascending order according to their distance: $\mathscr{D}_{j^*}^i \equiv \mathscr{D}(Y_t, B_{j^*}^i)$. Let $\mathscr{D}_{j^*}^{(i)}$ denote one element of this partial ordering with $\mathscr{D}_{j^*}^{(1)} \leq \mathscr{D}_{j^*}^{(2)} \leq \ldots \mathscr{D}_{j^*}^{(i)} \leq \ldots \mathscr{D}_{j^*}^{(N)}$. Finally, the K best trajectories $T_k, k = 1 \ldots K$ are simply the ones associated to the K best and sorted blocks: $\mathscr{D}_{j^*}^{(1)} \leq \mathscr{D}_{j^*}^{(2)} \leq \ldots \mathscr{D}_{j^*}^{(k)} \leq \ldots \mathscr{D}_{j^*}^{(K)}$. The K selected trajectories $T_k = \{(X_t^k, m_t^k)\}_{t=c_k}^{|T_k|}, k = 1 \ldots K$ are composed of both a set of features $X_t \in \Re^Q$ and

knowledge m_t about the state. The next steps of the algorithm consists in aggregating trajectories $T_k, k = 1 \dots K$ where two problems arised: 1) How to aggregate the features $\{X_t^k\}_{t=c_k}^{|T_k|}, k = 1 \dots K$ in order to obtain a predicted set of features $\hat{X}_t$ (Step 2)?, and 2) How to aggregate the knowledge about states $\{m_t^k\}_{t=c_k}^{|T_k|}, k = 1 \dots K$ in order to obtain a predicted knowledge $\hat{m}_t$ (Step 3)?

Step 2 - Predicted Observation Trajectory: A simple and usual way to define a predicted observation trajectory $\hat{X}_t$ linked to the observation block Y_t is to compute the weighted average of the K sets of features:

$$\hat{X}_{t+h} = \sum_{k=1}^{K} F^k \cdot X_l^k, l = c_k \dots |T_k|, h = 1 \dots \mathscr{P} \tag{1}$$

where $\mathscr{P} = |T_k| - c_k + 1$ defines the set of instants of prediction. The normalized weights F^k are obtained by the softmax function of the sorted distances:

$$F^k = \frac{\exp(-\mathscr{D}_{j^*}^{(k)})}{\sum_{k'=1}^{K} \exp(-\mathscr{D}_{j^*}^{(k')})}, k = 1 \dots K \tag{2}$$

Equations 1 and 2 are directly used if the length of trajectories $T_k, k = 1 \dots K$ are the same. If it is not the case (and generally it is not), one can use a strategies consisting in selecting an horizon of prediction equal to the length of the smallest trajectory. For that, first, the trajectory with the smallest size is found: $H_t = \min_{k=1}^{K} |T_k|$, where H_t can be seen as the horizon of prediction at time t. Then, for all trajectories, only samples from c_k to H_t are kept. After removal of samples located beyond H_t, Equations 1 and 2 can be directly used:

$$\hat{X}_{t+h}^{CS} = \sum_{k=1}^{K} F^k \cdot X_l^k, l = c_k \dots H_t, h = 1 \dots H_t \tag{3}$$

where CS stands for "Cautious Strategy" and X_h^k is the value of features in trajectory T_k taken at time h. The value of F^k is given by Eq. 2. The main advantage of this strategy is simplicity and efficiency since the horizon is gene rally shortened (to the smallest trajectory) and thus providing more reliable predictions. The main drawback is that the horizon of prediction is justly made shorter and therefore reducing forecasting capability.

At the end of step 2, the prediction of observation trajectory $\hat{X}_t$ is known according to the observation block Y_t and to the training dataset $\mathscr{L}$. Note that exponential smoothing using past prediction ($\hat{X}_{t-1}$) can be performed to improve temporal consistency [1] (not used in this paper).

Step 3 - Predicted Sequence of States: It is concerned by the prediction of future states. Two strategies are proposed: 1) Classification of predictions (CPS) and 2) Direct projection of future state sequence (DPS).

Classification of predictions strategy (CPS): This strategy consists in classifying the predicted observations given by step 2 into states. It requires the training of classifiers able to discriminate the different states. For the sake of simplicity, we consider the multiclass classifier called Evidential K-nearest neighbours (EvKNN) [11]. This classifier is able to generate a belief mass on the possible states in Ω given an observation. The main feature of this classifier is the possibility to manage belief functions m_t^i provided in the training dataset $\mathscr{L}$ (partially-supervised classification).

Given both a block of data $\hat{B}_h$ centered around the predicted observation $\hat{X}_{t+h}$ and the training dataset $\mathscr{L}$, the classifier provides a belief mass on the possible states:

$$m_{t+h}^{CPS} \leftarrow \text{EvKNN classifier}(\mathscr{L}, \hat{B}_h) \tag{4}$$

From this belief mass, a *hard* decision can be made to estimate the state of the current block by using the pignistic transform [9] which computes a probability distribution (suited for decision-making) from the belief mass m_{t+h}^{CPS}. Repeating this process on blocks composing the predicted observation $\hat{X}_t$, one simply obtains a sequence of states.

Direct projection of future state sequence (DPS): In order to avoid the dependency between state sequence prediction to observation prediction as in CPS, we propose to exploit another strategy that is the *direct projection of future state sequence*. This second strategy draws benefits directly from the training dataset. The main idea is to apply a similar reasoning as for features X_t but now for belief mass m_t. To go further in details, let consider the set of belief masses for the K nearest neighbours, i.e. $m_t^k, k = 1 \dots K, t = c_k \dots |T_k|$. These K belief masses can be considered as coming from distinct pieces of evidence so that the conjunctive rule of combination $\oplus$ can be used:

$$\hat{m}_{t+h}^{DPS} = \oplus_{k=1}^{K} m_l^k, l = c_k \dots |T_k|, h = 1 \dots \mathscr{P} \tag{5}$$

where *DPS* stands for "direct projection strategy" and $\mathscr{P} = |T_k| - c_k + 1$. To decrease the amount of conflict during the fusion process, we propose to use a *discounting* using the weights estimated in the KNN. The highest the weight, the less the discount, meaning that the related BBA is trusted. Once the BBAs have been discounted, the estimated belief mass at time t in DPS is given by Eq. 5.

Step 4 - Remaining Useful Life Estimation: *CPS and DPS fusion:* To draw benefits from both CPS and DPS approaches, the BBAs m_{t+h}^{CPS} (Eq. 4) and m_{t+h}^{DPS} (Eq. 5) are combined and the resulting BBA is converted into a probability distribution from which a decision can be made [12]. Dempster's rule is not adapted for the fusion of CPS and DPS's BBAs because m_{t+h}^{CPS} and m_{t+h}^{DPS} can not be considered as coming from distinct bodies of evidence. Indeed: 1) CPS is a classification of predictions resulting from the weighted combination of continuous predictions, and 2) DPS generates belief masses discounted by the weights, and therefore, both approaches depend on the weights. Moreover, both rely on the BBAs in the training dataset $\mathscr{L}$. Thus, the fusion may be performed using the cautious rule [13]:

$$\hat{m}_{t+h} = m^{CPS}_{t+h} \textcircled{\wedge} m^{DPS}_{t+h} \tag{6}$$

from which a decison concerning the state at time $t+h$ can be made and the result is the estimation of a sequence of states $\hat{\omega}_{t+h}$. Note that the neutral element is not always the vacuous BBA [13], except for separable BBAs as the ones used in evidential KNN exploited in CPS. In this case, if BBAs in the training dataset are vacuous, then the fusion equals CPS.

RUL estimates: Let consiser this sequence of states but also all previous predicted sequences. Since each sequence is composed of possible transitions between some states q and r, the set of time instants of transitions between both states is: $I_{q\to r} = \{t : \hat{\omega}_{t-1} = q \text{ and } \hat{\omega}_t = r\}$. To estimate the Remaining Useful Life (RUL) of the system, it is sufficient to determine the location of the critical transition from state $q =$ "degrading state" to state $r = q+1 =$ "fault state":

$$\text{transition } q \to r \text{ critical } \Rightarrow RUL = \mu_{q,r} - t \tag{7}$$

where $\mu_{q,r}$ is the estimated time from t to the transition between the degrading state q and the faulty state r that can be computed by a median. It can be associated to a dispersion $\sigma_{q\to r}$ that we computed using the interquartile range:

$$\begin{aligned} \mu_{q\to r} &= \text{median } (I_{q\to r}) \\ \sigma_{q\to r} &= Q_3 - Q_1 \end{aligned} \tag{8}$$

where Q_i is the i-th quartile and $n_I = |I_{q\to r}|$ is the number of elements in the set of time instants of transition $I_{q\to r}$.

Therefore, both methods for sequence prediction, CPS (classification) and DPS (direct projection), assume that each trajectory in the training dataset is made of *at least* two states, say "normal state" and "abnormal state", and knowledge on these states can be uncertain and imprecise and represented by belief functions.

4 First Results, Conclusion and Further Work

Illustration : We considered the PHM'08 challenge data [14] that we segmented into four states (available at `http://www.femto-st.fr/~emmanuel.ramasso/PEPS_INSIS_2011_PHM_by_belief_functions.html`). The first features and the segmentation are depicted in Fig. 1 which underlines the difficulty of using a statistical approach based on durations for degradation modelling [15].

Figure 2 depicts the sensitivity of the EVIPRO-KNN algorithm with respect to the parameters K (number of neighbours) and W (window's size). With $K = 3$ and $W = 30$, one can expect results close to 90% on the considered dataset. The prediction was considered as correct when falling in the interval $[-10, +13]$ around the ground truth, and the beginning of the prediction was taken as the time-instant corresponding to 75% of the length of the analysed trajectory (e.g. if the trajectory's length is equal to 240 then the starting time of the prediction was set to 180).

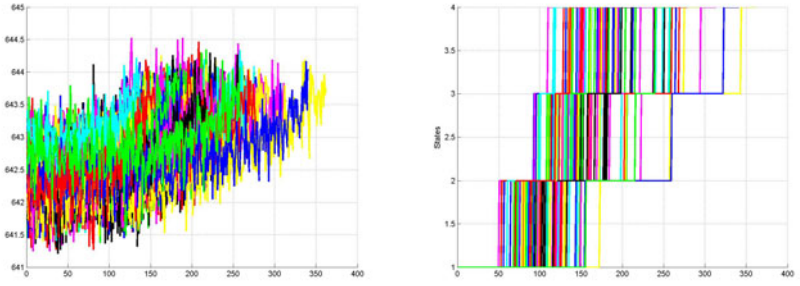


Fig. 1 Left: Evolution of the first feature for all trajectories in the training dataset, and right: the state sequences after decision-making based on the belief masses.

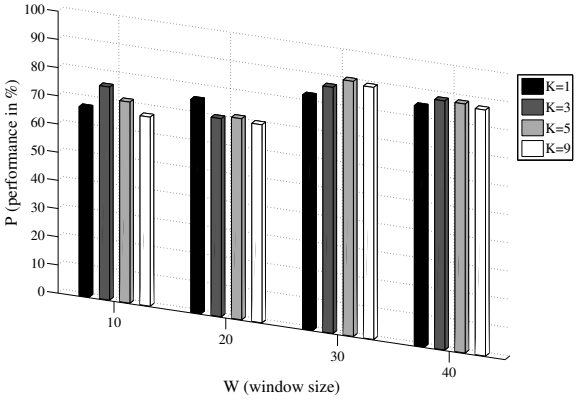


Fig. 2 Left: Sensitivity to W and K.

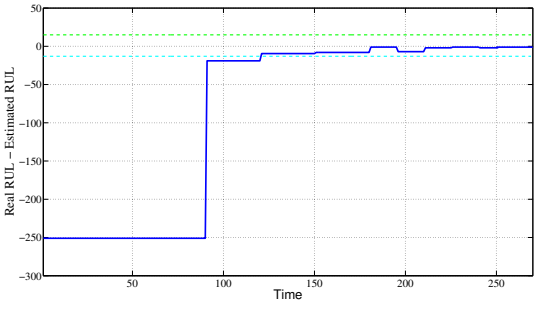


Fig. 3 The prediction appears in continuous bold line, while the real value is in dashed line.

Figure 3 illustrates the evolution of the differences at each time-step between the estimated RUL and the real RUL for $W = 30$ and $K = 3$ where a convergence to the real value is observed as expected: a good estimate of the RUL (in interval $[-10, 13]$) is obtained at $t = 90$, so 180 time-units in advance.

Conclusion and further work : EVIPRO-KNN is an online algorithm for prognostics and health detection working as case-based reasoning but managing uncertain knowledge about the states that could be provided as belief functions in the training dataset. EVIPRO-KNN can predict sequence of continuous observations jointly with discrete states enabling the user to have access to the online segmentation of the current observed data and of predictions which is then used to estimate the RUL.

Acknowledgment. This work is supported by a PEPS-INSIS-2011 grant from the French National Center for Scientific Research (CNRS) under the administrative authority of France's Ministry of Research.

References

1. Bonissone, P., Varma, A., Aggour, K.: A fuzzy instance-based model for predicting expected life: A locomotive application. In: IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications, pp. 20–25 (2005)
2. Saxena, A., Wu, B., Vachtsevanos, G.: Integrated diagnosis and prognosis architecture for fleet vehicles using dynamic case-based reasoning. In: Autotestcon, pp. 96–102 (2005)
3. Wang, T.: Trajectory similarity based prediction for remaining useful life estimation. Ph.D. dissertation, University of Cincinnati (2010)
4. Bishop, C.: Pattern Recognition and Machine Learning. Springer (August 2006)
5. Murphy, K.P.: Dynamic Bayesian networks: Representation, inference and learning. Ph.D. dissertation, UC Berkeley (2002)
6. Ramasso, E., Gouriveau, R.: Prognostics in switching systems: Evidential Markovian classification of real-time neuro-fuzzy predictions. In: IEEE Int. Conf. on Prognostics and System Health Management, Macau, China, pp. 1–10 (2010)
7. Dempster, A.: Upper and lower probabilities induced by multiple valued mappings. Annals of Mathematical Statistics 38, 325–339 (1967)
8. Shafer, G.: A mathematical theory of Evidence. Princeton University Press, Princeton (1976)
9. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence 66(2), 191–234 (1994)
10. Come, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recognition 42(3), 334–348 (2009)
11. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. on Systems, Man and Cybernetics 5, 804–813 (1995)
12. Smets, P.: Decision making in the TBM: The necessity of the pignistic transformation. Int. Jour. of Approximate Reasoning 38, 133–147 (2005)
13. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. Artificial Intelligence 172, 234–264 (2008)
14. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: Int. Conf. on Prognostics and Health Management, Denver, CO, USA, pp. 1–9 (2008)
15. Dong, M., He, D.: A segmental hidden semi-markov model (HSMM)-based diagnostics and prognostics framework and methodology. Mechanical Systems and Signal Processing 21, 2248–2266 (2007)

Adaptive Initialization of a EvKNN Classification Algorithm

Stefen Chan Wai Tim, Michèle Rombaut, and Denis Pellerin

Abstract. The establishment of the learning data base is a long and tedious task that must be carried out before starting the classification process. An Evidential KNN (EvKNN) has been developed in order to help the user, which proposes the "best" samples to label according to a strategy. However, at the beginning of this task, the classes are not clearly defined and are represented by a number of labeled samples smaller than the k required samples for EvKNN. In this paper, we propose to take into account the available information on the classes using an adapted evidential model. The algorithm presented in this paper has been tested on the classification of an image collection.

1 Problem Positionning

1.1 Classification Problem

The classification process needs some a priori knowledge for the class definition. This knowledge can be modeled for the classes (neural network, bayesian classifier) or can be limited to a learning set composed of labeled samples (KNN, SVM). In any case, the classifier needs a learning set to manage the classification of unlabeled samples from the collection and this learning set must be representative of the classes. When it is the case, the classical approaches are very efficient and are used in numerous applications. However, setting up such learning database can be a laborious task for the user.

We proposed in a previous paper [1], an assistance system for image collection classification presented Fig. 1. The first part of the system, based on Evidential KNN (EvKNN), models all available knowledge provided by the already labeled images in order to structure the unlabeled ones. The second part is a user assistance system (based on active learning) that proposes an ordered list of images to be labeled

Stefen Chan Wai Tim · Michèle Rombaut · Denis Pellerin
GIPSA-Lab/DIS, CNRS - UJF, Grenoble FRANCE
e-mail: firstname.author@gipsa-lab.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 69–76.
springerlink.com

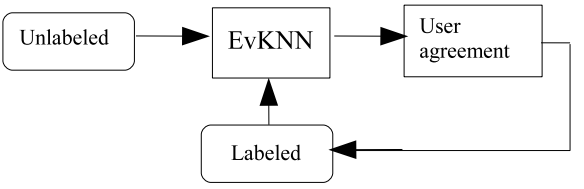


Fig. 1 Labeling process of the training set. At the beginning, the training set is almost empty. The EvKNN classifier takes all available labeled samples to propose to the user a label for an unlabeled sample. With agreement of the user, the new labeled sample is stored in the labeled set.

according to a specific strategy and assign a possible label. Using a suitable interface, the user agrees or disagrees with the proposal, and the global knowledge is updated.

This paper deals with the beginning of the first part of the labeling process, when the training set is almost empty, with only some labeled samples. In this case, there are generally less than k samples that belong to each known class and the samples are not completely representative of a class. Therefore, EvKNN algorithm cannot be used directly without adaptations. The adaptations are presented in Section 2, and the adapted algorithm is tested on an image collection (Section 3).

1.2 Evidential KNN

In [2], T. Denœux explains that "voting KNN" procedures show several limitations and he proposes to take into account the distance from the neighbors to model uncertainty and imprecision in class labels. It is assumed that the set of training samples is composed of enough samples for each class of decision. In the KNN algorithm, when there are at least k known samples of each class, there are enough training neighbors to model the membership of every incoming unlabeled sample to each class. T. Denœux proposes to model these memberships by belief functions (see Eq. 1).

We assume that x^s is the incoming unlabeled sample, and x_q^i is a labeled sample belonging to class C_q, one of the Q known classes. $d^{s,i}$ is the distance between these two samples in the parameter space. The knowledge of the x_q^i label gives information about the class of x^s. The basic belief assignment (BBA) $m_i^{\Omega_q}$ is defined on $\Omega_q = \{H_q, \overline{H_q}\}$, where hypothesis H_q means "sample x^s belongs to class C_q", whereas $\overline{H_q}$ is the opposite hypothesis:

$$\begin{aligned} m_i^{\Omega_q}(H_q) &= \alpha_q . e^{-\left(\frac{d^{s,i}}{\sigma_q}\right)^{\beta}} \\ m_i^{\Omega_q}(\Omega_q) &= 1 - m_i^{\Omega_q}(H_q) \end{aligned} \tag{1}$$

This model is very interesting when a class is represented by several samples in the parameter space. It means that two distant samples in this space can still belong to

the same class. It can be noticed that for a particular class C_q, the proposed BBA form will not cause conflict.

If there are k neighbors x_q^i, we can define k BBAs on the same frame of discernment Ω_q that can be conjunctively combined to give the BBA m^{Ω_q} concerning the sample x^s on membership to class C_q. In our previous paper [1], we proposed some adaptations in the combination. Contrary to Denœux's propositions in [2], we assumed that the Q classes are not exclusive. The combination of the Q BBAs m^{Ω_q} extended to the space $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_Q$ gives one BBA with possible multi-labeling. The combination architecture is described in [1].

1.3 Initialisation Step

The EvKNN method is very efficient if the number of classes Q is known, and if the training set is representative enough. If not, the performance of the classifier is reduced. In the later case, the goal is to model the poor information efficiently and possibly to ask an expert to validate the decision. It is also important to take into account the difference of available samples for each class, as well as the relative properties of the classes. The Belief Function Model is particularly well adapted to model such poor information, and given a large mass of belief for sets Ω_q.

In this paper, we describe an adaptive method to propose a decision to an expert. At each step, the choice of the expert is used to improve the knowledge to get a labeled sample and to adapt the information model for the class C_q. At the begining, the training set is only composed of some labeled samples, for instance less than k samples for each known class. The problem is to model this knowledge about the belonging of x^s to a known class. Then, a proposition is made that is validated by the operator.

2 Adaptive Model of Knowledge

The labeled neighbor x_q^i gives information on the belonging of x_s to the class C_q that can be modeled by the equations 1. The parameter σ_q weights the distance $d_q^{s,i}$ between the sample x_s and the labeled sample x_q^i. The parameter α_q is the discounting parameter that models the unreliability of the source of information. In the classification step, if the distance between two samples is null then it is not completely sure that x_s belongs to the same class C_q of x_q^i. Generally, the two parameters σ_q and α_q are constant, at least for each class. We propose to adapt them using the knowledge from known classes C_q, that is to adapt them according to number and position of labeled samples in the parameter space.

2.1 Adaptation of σ_q

For one unlabeled sample x_s, the k neighbors x_q^i of each class C_q are extracted if they exist. If not, all labeled samples of the class C_q are used. We propose to adapt

the distance $d_q^{s,i}$ between x_s and x_q^i by defining a relative distance $\left(\frac{d^{s,i}}{\sigma_q}\right)$. The idea is to take into account the mean distance $d_{q'}^{s,i}$ of x_s from all samples $x_{q'}^i \in C_{q'}$ for all classes $C_{q'}$. We propose to define σ_q where $C_{q'}$ and C_q are known classes and γ is a tuning parameter:

$$\sigma_q = \gamma \cdot min_{q'}(mean_{q' \neq q}(d_{q'}^{s,i})) \tag{2}$$

Therefore in equation 1, the distance $d_q^{s,i}$ is weighted by mean distance to the nearest class $C_{q'}$. The consequence of this definition is :

- if the near class $C_{q'}$ has a mean distance comparable to the distance $d_q^{s,i}$, the doubt is high. This can be modeled with a large mass attributed to each $m_i^{\Omega_q}(\Omega_q)$, given a small value to σ_q.
- if the near class $C_{q'}$ has a mean distance higher than the distance $d_q^{s,i}$, the doubt is low. This can be modeled with a larger mass attributed to $m_i^{\Omega_q}(H_q)$, given a large value to σ_q.

2.2 *Adaptation of* α_q

The number of known neighbors has a great influence on the BBA's values. If one class C_q contains a lot of labeled samples (more than k), due to the definition of the BBA (Eq. 1), the conjonctive combination of k BBAs reinforce the $m_i^{\Omega_q}(H_q)$. On the contrary, if the class C_q is underepresented ($k_q < k$), then BBA is less informative. This can induce an imbalance between the classes.

We propose to adapt the parameter α_q to the number of known neighbors for each class C_q. The idea is to reinforce the mass $m_i^{\Omega_q}(H_q)$ when $k_q < k$. The definition of α_q is:

$$\alpha_q = \alpha_0^{\frac{1}{1+k-kq}} \tag{3}$$

where $\alpha_0 = 0.8$. In equation 3, $\alpha_q > \alpha_0$ when the number of neighbors k_q is less than k, to reinforce the mass of the H_q hypothesis. It is equal to α_0 when $k_q = k$.

3 Application to Image Classification

The automatic classification problem is very complex for image (and video) collections because the user interprets the semantic content. The extracted attributes from the images are not directly connected to the classes wished by the user. During the labeling process of the learning set, the classification system must take into account the knowledge of the user in order to "learn" the classes C_q. In the KNN approach, the system requires samples of images (or videos) that are labeled by the user. The operation is long and tedious. In a previous work [1], we developed an assistance

classification system based on the fact that it is difficult for a user to a priori define all the classes, and manage all the images from the database simultaneously.

3.1 Global Architecture of the Classification System

It could be difficult for a user to classify a set of images, particularly when the set is large and the classes are not defined a priori. This is the case, for instance, when somebody wants to store his holiday images, not only by time stamp, but also by themes (actions: visit, drive..., locations: at home, outdoor...). The images can be multi-labeled. Rather than submitting all the images simultaneously, or one by one in random order, the idea is to propose an ”adequate” order following a sampling strategy by an active learning process, rarely used for multi-labeling [3]. We retain the main elements of the developed system. The main idea is to select images for the user which are ”interesting” to classify according to a specific strategy and to propose a label. The user can accept the proposed label, or change the label or create a new class. The automatic image selection is carried out from the accumulated knowledge from the previous image classification.

The framework is divided into two main parts [4]: a fully automatic part for “modeling the knowledge” presented in this paper, and another part that concerns the user interactions in order to select the images to be labeled via a graphic user interface. The entire framework is presented as three modules in Fig. 2.

3.2 Sampling Strategies

A small set of chosen images is proposed to the user to classify. These could be very similar to labeled images (Most Positive unlabeled images) or very different from labeled images (Most Rejected unlabeled images). We chose the Most Positive strategy for the test because it introduces an imbalance of number of neighbors between classes during the process.

We define a positive hypothesis $\omega_p^q \in \Omega_P \subset \Omega$ composed of only one local positive hypothesis such as H_q, the others corresponding to local negative ones such as $\overline{H_n}$:

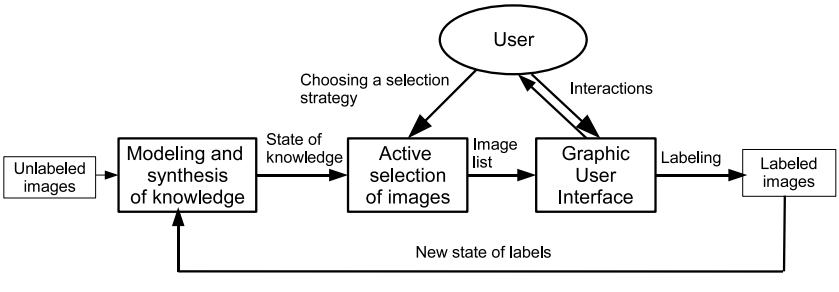


Fig. 2 Architecture of the system of classification

$$\omega_p^q = (H_q, \overline{H}_{n_1}, \overline{H}_{n_2}, \cdots, \overline{H}_{n_N}) \tag{4}$$

This positive hypothesis ω_p^q means that the unlabeled image belongs to the single class C_q. The strategy, sometimes named "most relevant" [5], selects the unlabeled images that obtain the highest pignistic probability [6] computed on Ω_P, subset of Ω made up of only positive hypotheses ω_p^q (Eq. 4). It corresponds to the selection of "easy to classify" images, because the visual content is very similar to already labeled images.

3.3 Results

The classification algorithm has been tested on a Corel database of 321 images (Examples in Fig. 3). The database contains 9 classes ('Monuments', 'Bus', 'Dinausors', 'Elephants', 'Mountains', 'Flowers', 'Horses', 'Meals', 'Faces'), and each class has between 15 to 46 images. Some classes are very heterogeneous from the color point of view.

For each image, two kinds of features (color and orientation) have been extracted. For color, classic 3D histograms in HSV domain have been used with 8 bins in each dimension, giving 512 components. For orientation, we used horizontal and vertical gradient filters that give a histogram of 64 bins.

At any time, an unlabeled image is proposed to the user according to the chosen strategy (here the Most Positive) as well as a proposed label. The user can accept the proposed label or reject it. In the later case the proposal is recognized as false proposal. The objective is to limit such false proposals in order to make the task easier for the user. The test is performed automatically since the ground truth is known.

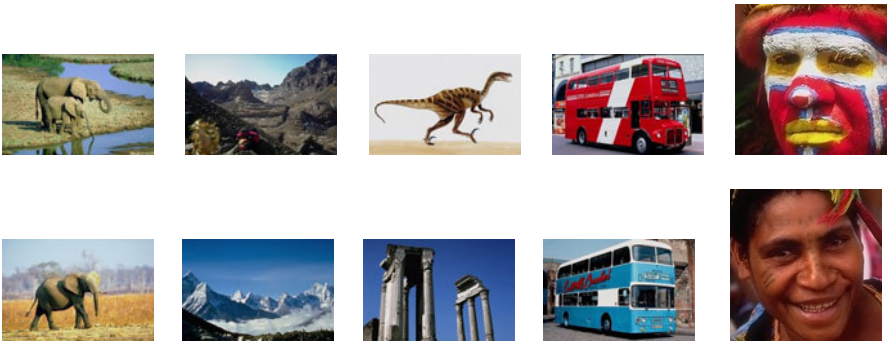


Fig. 3 Examples of color images belonging to the collection

3.3.1 Effect of the Parameter σ_q

An example of comparison is given in Fig. 4. We chose $\sigma = 0.5$ (best result) in the constant case, whereas σ_q adapted case follows Eq. 2.

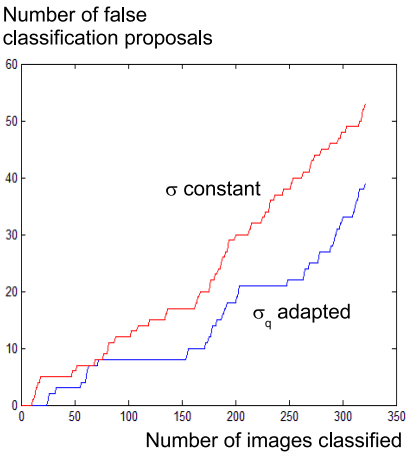


Fig. 4 Comparison of classification for σ constant and σ_q adapted (α constant)

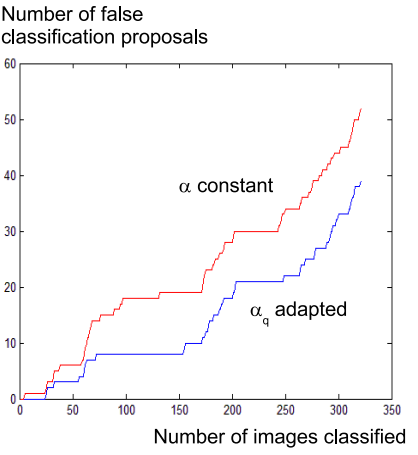


Fig. 5 Comparison of classification results for α constant and α_q adapted (σ constant)

Compared to σ constant, σ_q adapted resulted in a reduced number of false classification proposals. Indeed, if σ_q is too small, the mass goes to the doubt and part of information disappears. If σ_q is too large, the mass $m(H_q)$ tends towards α_q. Here we are too categorical comparatively to the complexity of the content. For σ_q adapted, if the class is far from any other one then σ_q is large, otherwise σ_q is small.

3.3.2 Effect of the Parameter α_q

An example of comparison is given in Fig. 5. We chose $\alpha = 0.8$ (best result) in the constant case, whereas α_q adapted follows Eq. 3 with $\alpha_0 = 0.8$.

Compared to α constant, α_q adapted resulted in reduced number of false classifications proposals. This result is due to the reduction of imbalance on the masses during the search of neighbors. The value of α_q is close to 1 when the number of neighbors is 1, giving more mass to the H_q hypothesis. It is equal to α_0 when $k_q = k$.

4 Conclusion

The adapted EvKNN proposed in this paper makes the task of the user easier during long and tedious labeling of the training set. The algorithm takes into account the real known neighbors (less than k) and the relative distances of the classes. Because the user is in the loop, a new class can be added when a sample arrives, and in this case, the proposed adapted EvKNN is particularly efficient. The algorithm has been tested on an image collection. The image classification process is very complex because the user attaches semantic interpretation for an image that an automatic system can not manage using simple image attributes.

Acknowledgements. We thank the Rhône-Alpes region for its support with the LIMA project.

References

1. Goëau, H., Rombaut, M., Pellerin, D., Buisson, O.: Multi-labeled image classification by TBM active learning. In: Workshop on the Theory of Belief Function, April 1-2, Brest (2010)
2. Denœux, T.: A k-nearest neighbor classification rule based on dempster-shafer theory. IEEE Trans. on SMC, 804–813 (1995)
3. Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Zhang, H.-J.: Two-dimensional multi-label active learning with an efficient online adaptation model for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(1) (2008)
4. Goëau, H.: Structuration de collections d'images par apprentissage actif crédibiliste, PhD, Université Joseph Fourier, Grenoble, France (May 2009)
5. Crucianu, M., Ferecatu, M., Boujemaa, N.: Relevance feedback for image retrieval: a short survey. Report of the DELOS2 European Network of Excellence (FP6) (2004), citeseer.ist.psu.edu/crucianu04relevance.html
6. Smets, P.: Decision making in tbm: the necessity of the pignistic transformation. Journal of Approximate Reasonning 38(2), 133–147 (2005)
7. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proceedings of the 20th International Conference on Machine Learning, pp. 59–66. AAAI Press (2003)

Classification Trees Based on Belief Functions

Nicolas Sutton-Charani, Sébastien Destercke, and Thierry Denœux

Abstract. Decision tree classifiers are popular classification methods. In this paper, we extend to multi-class problems a decision tree method based on belief functions previously described for two-class problems only. We propose three possible extensions: combining multiple two-class trees together and directly extending the estimation of belief functions within the tree to the multi-class setting. We provide experiment results and compare them to usual decision trees.

1 Introduction

Decision trees [2] (classification trees for categorical labels and regression trees for numerical ones) are popular classifiers, due to their simplicity, efficiency and readability. The construction of usual decision trees relies on probability theory. However, classical methods are not always fully adequate to deal with some problems. Among these problems are (1) the fact that all kinds of uncertainties (either in inputs or outputs) cannot be modeled faithfully by classical probabilities and (2) the fact that frequencies of occurrence are only sensible to proportions in a sample and not to its size.

Beyond the fact that the relationship between inputs and outputs may be non-deterministic, a classifier may have to deal with three different possible levels of uncertainty: in inputs, in outputs, and uncertainty due to the fact that the trained classifier is an estimation of the ideal one, due to a limited amount of knowledge or data. In this work, we mainly address the third issue, where the estimation quality translates into imprecision of belief functions.

Belief function theory [13] offers a convenient framework to deal with all these problems. For instance, Elouedi *et al.* [9] propose different ways to adapt decision

Nicolas Sutton-Charani · Sébastien Destercke · Thierry Denœux
UMR CNRS 7253 Heudiasyc Université Technologique de Compiègne, BP 20529 - F-60205 Compiègne cedex - France
e-mail: nicolas.sutton-charani@hds.utc.fr, sebastien.destercke@hds.utc.fr, thierry.denoeux@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 77–84.
springerlink.com

trees in the Transferable Belief model (TBM) framework to deal with uncertain outputs during the tree construction. In this work, we extend another approach also using belief functions proposed by Denœux and Skarstein Bjanger [8] that can cope with uncertain outputs and imprecision arising from limited sample size. In this sense, this approach is closer to some imprecise probabilistic approaches [1] that naturally integrate sample size information in their construction.

As Skarstein Bjanger's method only concerns two-class problems, we extend this methodology to any number of classes. For multi-class problems, we propose three ways of doing such an extension:

- combining belief functions provided by sets of two-class trees [12];
- building multinomial belief functions using the Imprecise Dirichlet Model (IDM) [14];
- building multinomial predictive belief functions using Denœux's approach [6].

Section 2 presents the needed background about decision trees and Skarstein Bjanger's method. Section 3 then extends this methodology to the multi-class case. Finally, in Section 4 we compare new classifiers with the usual CART algorithm and discuss the effects of parameters on experiment results.

2 Background

2.1 Decision Trees

Let (X,Y) be a random vector where $X = (X_1,...,X_J) \in \mathscr{X} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_J$ represents the features (continuous or discrete) and $Y \in \mathscr{Y} = \{Y_1,\ldots,Y_K\}$ the class to predict. From a sample $E = \{(X^1,Y^1),\ldots,(X^n,Y^n)\}$, decision tree methods build iteratively a model of (X,Y) by building a partition of $\mathscr{X}$. Here, we consider binary trees (i.e., CART-like models), where each split provides two children.

The method works as follows: from a root node containing the whole learning sample, the optimal split (among all the variables and their values) in term of information gain is searched. The information gain IG corresponding to splitting on variable X_k with value α is computed as follows:

$$IG(k,\alpha) = i(t_0) - p_L i(t_1) - p_R i(t_2), \tag{1}$$

where $i(t)$ is an impurity measure of a node t, t_0 the root node, t_1 and t_2 its child nodes, p_L is the proportion of the samples in t_0 verifying the condition $X_k < \alpha$ (i.e., $p_L = {}^{n_L}/_{n}$ where n is the sample size in t_0 and n_L the number of cases such that $X_k < \alpha$). $p_R = 1 - p_L$ is the sample proportion not verifying it. The selected splitting value (k,α) is then the one maximizing IG (resulting in a gain in purity).

The method is then applied recursively to each child nodes until no possible information gain greater than a pre-established threshold can be made. In this case, the node becomes a leaf predicting the most frequent class of the leaf sample.

The information gain (or impurity measure) is calculated using the Gini-index for the CART algorithm or Shanon entropy for C4.5's (Quinlan [11]). Both of these

functions measure the homogeneity in term of classes. They both use the frequencies of the different classes in the node samples; however, these frequencies do not depend on the sample size (provided class proportions remain the same). In contrast, Skarstein Bjanger's method impurity measure do change with the sample size.

2.2 Skarstein Bjanger's Method for Two-class Datasets

This method shares CART principles, but differs in the computation of information gain: it uses mass functions instead of simple frequencies and the used impurity measure combines nonspecifity (imprecision) and conflict (variability).

To build the mass functions, Dempster's inference method applied to Bernoulli trials [5] induces the following mass function:

$$\begin{cases} m_{DaBt}(\{Y_1\}) = \frac{n_1}{n+1} \\ m_{DaBt}(\{Y_2\}) = \frac{n_2}{n+1} \\ m_{DaBt}(\mathscr{Y}) = \frac{1}{n+1}, \end{cases} \tag{2}$$

where n is the number of samples and n_1, n_2 are the number of samples whose class is Y_1, Y_2, respectively. Denœux and Skarstein Bjanger then propose to use the following impurity measure [10], applied to m_{DaBt}:

$$U_\lambda(m) = (1-\lambda)N(m) + \lambda D(m) \tag{3}$$

where $N(m) = \sum_{A \subseteq \mathscr{Y}} m(A) \log_2 |A|$ measures the non-specificity and

$$D(m) = -\sum_{A \subseteq \mathscr{Y}} m(A) \log_2 BetP(A)$$

the variability. The two parts are weighted by hyperparameter $\lambda \in [0,1]$. Note that as the size n of the sample increases, $m(\mathscr{Y})$ (the imprecision) decreases. When using U_λ as impurity measure $i(t)$, the information gain (1) can be negative. This gives a natural stopping criterion when building the tree, that is, no split is done if all possible information gains are negative. Usually, λ can be fixed by cross-validation (see Section 4).

Table 1 shows results obtained with CART-classification trees and with classification trees based on Skarstein Bjanger's method. The stopping criteria was the following: keep splitting while $IG > \beta$ for usual CART-trees ($IG > 0$ for the one based on U_λ) and while the children nodes of the split contains a minimum of 10 samples. The usual CART procedure and the U_λ-based algorithm were optimized with respect to the threshold β and parameter λ, respectively, using 10-fold cross-validation. Results show that the methods achieve comparable accuracies.

Dempster's method of inference cannot be easily extended from the binomial to the multinomial case. Therefore, we propose three ways to handle multiple classes: break up the classification problem containing K classes (K $\geq$ 3) into C_K^2 two-class problems using Quost's method for combining binary classifiers [12] and use the Imprecise Dirichlet Model (IDM) approach or Denœux's multinomial model.

Table 1 Error rates of trees depending of the used impurity measure

Data set	Number of features	standard CART	trees based on U_λ
Blood transfusion	4	23.5%	24.2%
Statlog heart	13	28%	25.7%
Tic-tac	9	21.5%	11.5%
Breast-cancer	10	5.9%	4.7%
Pima	8	27.3%	25.1%

3 Multi-class Cases

3.1 Combinations of Binary Classifiers

In [12], Quost presents a method to handle multi-class classification problems by combining classifiers built on sub-samples containing only two classes. He proposes to learn (from the corresponding sub-sample) a conditional belief function for each pair $\{Y_i, Y_j\}$, $1 \leq i < j \leq K$ of classes and to combine them into a global belief function over $\mathscr{Y}$ using an optimisation procedure.

Here, we propose to use this method with decision trees issued from Skarstein Bjanger's method, using the latter as base classifier to learn conditional belief functions. This method is different from the one proposed by Vannoorenberghe and Denœux [15] in which K two-class trees corresponding to a "one vs all strategy" are built, their output being then combined by an averaging of obtained massess.

Decision trees are well adapted to this kind of combination, since they are simple classifiers. However, note that the optimization of λ becomes an issue, as $K(K-1)/2$ classifiers have to be learned at each optimization step.

3.2 IDM

The IDM was introduced in the "imprecise probability" framework by Walley [16]. Note that, although belief functions can be interpreted as imprecise probabilities, it is not their only possible interpretation. However, the IDM turns out to yield a belief function as output, hence it can be used in our framework. The IDM imprecision is controlled by a hyperparameter $s \in \mathbb{R}^+$. From a random sample $Y^1, ..., Y^n$, Walley showed that the lower predictive probability distribution on $\mathscr{Y}$ is $\underline{P}(Y_k|N,s) = {}^{n_k}/_{n+s}$ where n_k is the number of times Y_k has been observed. The corresponding mass function is such that:

$$\begin{cases} m_{IDM}(Y_j) = n_j/(n+s) & j = 1, ..., K \\ m_{IDM}(\mathscr{Y}) = s/(n+s) \end{cases} \tag{4}$$

Note that we recover equation (2) for $K = 2$ and $s = 1$. Using m_{IDM}, U_λ can be applied to measure the impurity in a node and multi-class trees can thus be created. The analytical form of U_λ applied to m_{IDM} can be derived as:

$$U_\lambda(m_{IDM}) = \frac{(1-\lambda)s}{n+s}\log_2(K) - \frac{\lambda}{n+s}\sum_{k=1}^{K} n_k \log_2\left[\frac{Kn_k+S}{K(n+s)}\right] \quad (5)$$

However, even if this model is simple, it is not easy to interpret it within the belief function framework. Also, the IDM imprecision only depends on the sample size n, and not on its distribution over $\mathscr{Y}$. This is not the case for Denœux's multinomial predictive belief function that offers an interesting alternative.

3.3 Denœux's Multinomial Model

Denœux [6] proposes to use Goodman's confidence intervals to build a predictive belief function. The first step is to build probability intervals [4] (probability lower and upper bounds over singletons) and then to transform them into belief functions.

Let $(X^1,Y^1),...,(X^n,Y^n)$ be an *iid* sample where $Y^k \in \mathscr{Y} = \{Y_1,...,Y_K\}$, those probability intervals $[P_k^-, P_k^+]$ are given, for Y_k (k=1,...,n), as:

$$P_k^- = \frac{q+2n_k-\sqrt{\Delta_k}}{2(n+q)} \quad and \quad P_k^+ = \frac{q+2n_k+\sqrt{\Delta_k}}{2(n+q)} \quad , \quad (6)$$

where q is the quantile of order $1-\alpha$ of the chi-square distribution with one degree of freedom, and where $\Delta_k = q(q+\frac{4n_k(n-n_k)}{n})$. As shown in [6], the lower confidence measure (i.e., $P^-(A) = \max(\sum_{Y_k\in A} P_k^-, 1-\sum_{Y_k\notin A} P_k^-)$) built using these regions in the case where $K=2$ or 3 is a belief function.

Note that the built belief functions follow Hacking's principle (see [6] for details), but the solution for $K=2$ is not equivalent to that of Eq. (2).

In the case $K>3$, the Möbius inverse of P^- may take negative values, so P^- is not a belief function in general. Different methods involving linear programming are proposed in [6] to approximate it into a belief function. Also, in the special case where the classes are ordinal, Denœux proposes an algorithm restricted to a certain set of focal elements. A valid predictive *bba* is obtained. These belief functions can then be used with U_λ to build multi-class trees.

4 Experiments

We start by comparing the classifier performances, and then discuss the effect of λ.

4.1 Comparison between classifiers

We compares the three proposed extensions with the usual CART algorithm. Table 2 shows three multi-class UCI datasets characteristics. Table 3 presents experimental results on the previous datasets comparing the accuracy of four types of classifiers:

- Standard CART trees based on Gini index (CART);
- Trees based on U_λ with m_{IDM} (IDM);

Table 2 UCI data sets used in experiments

Data set	Number of features	Number of classes	learning sets size	test sets size
Iris	4	3	113	37
Balance scale	4	3	469	156
Wine	13	3	134	44
Car	6	4	1152	576
Page blocks	10	5	3649	1824
Forest-fires	12	6	345	172

Table 3 Accuracies (R=error rate T=time computation in seconds) of trees depending of the masses assignment model

	CART		IDM		Combi		Multinomial	
datasets	R	T	R	T	R	T	R	T
iris	2.0%	0	2.0%	0	2.0%	1	2.0%	6
balance-scale	20.2%	0	25.0%	0	17.8%	2	15.9%	29
wine	11.9%	0	8.5%	0	13.6%	1	13.6%	19
car	17.7%	1	17.7%	1	15.6%	9	32.3%	8
pageblocks	4.8%	53	4.7%	38	5.0%	140	5.2%	1801
forests-fire	43.6%	1	43.0%	1	43.0%	15	43.0%	81

- Combination of two-class trees based on U_λ (combination);
- Trees based on U_λ with $m_{Multinomial}$ (multi).

The tree growing strategy is the following: keep splitting while $IG > \beta$ for CART and $IG > 0$ for the tree based on U_λ, the children nodes sample size is greater than 10 and the depth of the tree is smaller or equal to 5.

Because the aim of this experiment was to compare the different methods, none of the trees were optimized: for CART we fixed the threshold $\beta = 0$ and for trees based on U_λ we fixed $\lambda = 0.5$. None of the trees were post-pruned, as we are only interested in accuracies of each model, and not in their simplicity (defining a proper pruning strategy for U_λ based decision trees remains the matter of further research).

For the datasets with 3 classes we used the belief function induced by P^- whereas the linear programming and the ordinal approaches were used for *Page blocks* and $Forest - fires$, respectively.

The classifiers are competitive; however, as expected, computation times are longer with the multinomial model, due to its higher complexity.

4.2 Discussion about λ

Figure 1 shows the impact of λ in terms of tree complexity (using the usual number of leaf criterion) and in terms of accuracy on the UCI dataset "Pima". We can see that this complexity increases with λ, confirming that $1 - \lambda$ can be interpreted as the importance given to the lack of samples in a node (i.e., to non-specifity $N(m)$)

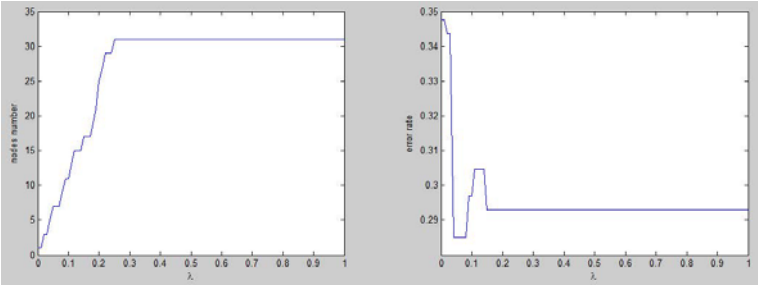


Fig. 1 Number of nodes as a function of λ (left) and error rate as a function of λ (right) for the PIMA dataset

and to the propensity of *IG* to be negative. This suggests that optimization (here, a 10-fold cross-validation) should also integrate tree complexity as a criterion. The parameter λ seems to have only a small influence on accuracy.

5 Conclusion

In this paper, we have extended Skarstein Bjanger's method for building decision trees to the multi-class case, proposing three ways to do so. The IDM is not really based on the belief function theory and may result in too simple belief functions; Denœux's multinomial model is more elaborated, fits better with a belief function approach, but requires heavier computational efforts; two-class decomposition is efficient, but makes the interpretation of results possibly harder (and, in any case, longer), as it builds a quadratic number of decision trees.

We have shown that the presented methods have a prediction power comparable to usual methods. However the present work is only a starting point with many perspectives: one of the major interest of using belief functions is the ability to handle uncertain data in inputs or outputs, a feature we shall integrate to the present methods in future works (using, for example, extensions of EM-algorithm to learn trees [3] [7]). Another interesting extension would be to adapt this model to continuous outputs and to regression problems.

References

1. Abellan, J., Moral, S.: Upper entropy of credal sets. Applications to credal classification. International Journal of Approximate Reasoning 39(2-3), 235–255 (2005)
2. Breiman, Friedman, Olshen, Stone: Classification And Regression Trees. Wadsworth, Belmont (1984)
3. Ciampi, A.: Growing a tree classifier with imprecise data. Pattern Recognition Letters 21(9), 787–803 (2000)
4. de Campos, L., Huete, J., Moral, S.: Probability intervals: a tool for uncertain reasoning. Int. J. Uncertainty Fuzziness Knowledge-Based Syst. 1, 167–196 (1994)

5. Dempster, A.P.: New methods for reasoning towards posterior distributions based on sample data. Annals of Mathematical Statistics 37, 355–374 (1966)
6. Denoeux, T.: Constructing belief functions from sample data using multinomial confidence regions. International Journal of Approximate Reasoning 42(3), 228–252 (2006)
7. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans. on Know. and Data Eng. (2011) (to appear), doi:10.1109/TKDE.2011.201
8. Denœux, T., Bjanger, M.S.: Induction of decision trees from partially classified data using belief functions. In: 2000 IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, pp. 2923–2928. IEEE (2000)
9. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. International Journal of Approximate Reasoning 28(2-3), 91–124 (2001)
10. Klir, G.J.: Uncertainty and information: foundations of generalized information theory. Wiley-IEEE Press (2006)
11. Quinlan, J.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
12. Quost, B., Denœux, T.: Pairwise Classifier Combination using Belief Functions. Pattern Recognition Letters 28, 644–653 (2006)
13. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
14. Utkin, L.V.: Extensions of belief functions and possibility distributions by using the imprecise dirichlet model. Fuzzy Sets and Systems 154(3), 413–431 (2005)
15. Vannoorenbergue, P., Denœux, T.: Handling uncertain labels in multiclass problems using belief decision trees. In: IPMU 2002, Annecy, France, vol. 3, pp. 1919–1926 (2002)
16. Walley, P.: Inferences from multinomial data: Learning about a bag of marbles. Journal of the Royal Statistical Society, Series B, 3–57 (1996)

Combination of Supervised and Unsupervised Classification Using the Theory of Belief Functions

Fatma Karem, Mounir Dhibi, and Arnaud Martin

Abstract. In this paper, we propose to fuse both clustering and supervised classification approach in order to outperform the results of a classification algorithm. Indeed the results of the learning in supervised classification depend on the method and on the parameters chosen. Moreover the learning process is particularly difficult which few learning data and/or imprecise learning data. Hence, we define a classification approach using the theory of belief functions to fuse the results of one clustering and one supervised classification. This new approach applied on real databases allows good and promising results.

1 Introduction

Behind the term of classification, one distinguishes two types of classification: the supervised and unsupervised one. The unsupervised classification is also called *clustering*. In clustering, from given data representing some object, we try to find groups or *clusters* which are the most compact and separated as possible. Then, we can try to affect one of the found cluster to a new observed object [2]. Generally, we make such decision based on the analysis of the dispersion of the objects in the data set. In the supervised context, the process can also be divided in two steps: the learning one and the classification. The learning step build a discriminate function based on labeled data, an unknown information in clustering. From this function, in the classification step, a new observed object is affected to one of the classes given by the fixed labels. Whatever the type of classification, we face up to many problems.

Fatma Karem · Mounir Dhibi
Research Unit PMI 09/UR/13-0, Zarouk College Gafsa 2112, Tunisie
e-mail: fatoumacy@yahoo.fr, mounir.dhibi@ensta-bretagne.fr

Arnaud Martin
University of Rennes 1, UMR 6074 IRISA, rue Edouard Branly, BP 30219,
22302 Lannion Cedex, France
e-mail: Arnaud.Martin@univ-rennes1.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 85–92.
springerlink.com

We are always looking for the appropriate method for a given problem without to be sure to achieve it. Indeed, the obtained results depend on the method and on parameters; the no-free lunch theorem assures us that there is no better algorithm. Therefore, the choice of the appropriate method and parameters is not easy for a given application. Furthermore in the supervised context, the learning data do not generally represent perfectly the real data we have to classify. For example, all real classes are not systematically well represented in the learning database. As a result, a possible solution to some of these classification problems is the fusion of clustering and supervised classification. The goal of this fusion is to reduce the imprecision of results by trying to make a compromise between both classifications.

Studying classification fusion approaches, most of them are dealing with the fusion of either supervised [9, 12] or unsupervised classification [3, 4, 11, 7, 8]. The unsupervised classification fusion approaches are more complex due to the absence of class labels: an association between the clusters coming from the different algorithms must be found. The researches made on the fusion between the clustering and the classification were used essentially in order to deploy the unsupervised in the learning of the supervised classification [6, 10, 13].

In this article, we propose a fusion approach combining supervised and unsupervised classification results. As framework, we choose the theory of belief functions which have been used with success to fuse supervised classification results [12]. This framework allows to represent the uncertainty and imprecision of the results of the clustering and supervised classification and to combine the results managing the conflict.

This paper is organized as follow: in the next section, we present the clustering and the supervised classification principles. In the third section, we explain the fusion based on the theory of belief functions. In section four, we present the proposed fusion approach and finally the last section presents the results given by an experimental study on real data.

2 Classification

The goal of the classification task is to identify the classes to which belong the objects representing by theirs characteristics or attributes. We distinguish two types of classification: supervised and unsupervised one.

2.1 Unsupervised Classification or Clustering

In the clustering, we want to group the similar objects of a population in clusters. Let's assume, we dispose of an ensemble of objects noted by $X = \{x_1, x_2, ..., x_N\}$ characterized by an ensemble of descriptors D. Therefore, the data are D-multidimensional. The aim is to find the groups (or cluster) to which each object x belongs. Hereafter, the clusters are noted by $C = \{C_1, C_2, ..., C_n\}$. The clustering can be formalized by a function noted by $Y_{\bar{s}}$, that associates each element of X to one or more elements of C. Generally, the clustering is essentially based on the

dispersion analysis to find the real clusters. Many difficulties can arise in this task. The main difficulty is to find the borders of the clusters. To evaluate the results, we have to find some evaluation criteria measuring the quality of results. Usually, we use indexes called validity indexes. There is no standard or general index. Among the clustering methods, we mention for example K-means and the hierarchical classification.

2.2 Supervised Classification

In the supervised context, the classification is based on two steps: the learning step and the classification step. In the learning step, we consider the objects in X already labeled, *i.e.* each object is associated to a known label belonging to an ensemble of classes noted by $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$. This is the conceptual difference with the clustering. The goal of the learning step is to find the best discriminate function Cl associating each data of the learning database x using the descriptor set (noted by D) to the correct class in Θ. The classification step consists to predict the class of a new object based on the learning function. Among the classification methods, we mention the k-nearest neighbors (k-NN), the decision tree, the neural network, the support vector machine (SVM) [2]. In the supervised context, the lack of learning data or the availability of inappropriate one make problems. In this case, we can consider that the learning function to discriminate data is imprecise and uncertain and leads to bad results. The confusion matrices are generally used to evaluate supervised classification results.

In this paper, a new approach is proposed to overcome the classification problems identified previously. This approach is based on the fusion between the supervised classification and the clustering results using the theory of belief functions.

3 Information Fusion Using the Theory of Belief Functions

The fusion of classifiers can be made in three levels of the classification process: data, characteristic and decision. The third level is the level of the classification results and is the most interesting for our study. Many framework have been used for information fusion, such as vote theory, theory of possibilities or theory of belief functions. The last one, also called Dempster-Shafer theory, allows to represent both imprecision and uncertainty through two functions: plausibility and belief. Both functions are derived from a function called mass function defined on all the subsets of the frame of discernment Θ, noted 2^{Θ}. That is the difference with the theory of probabilities where only singletons are considered. Let's design by m_j the mass function associated to the source S_j. The mass functions are defined on 2^{Θ} and affect a value from $[0,1]$. Moreover, the mass functions verify the constraint:

$$\sum_{A \in 2^{\Theta}} m_j(A) = 1 \tag{1}$$

Hence, the power set 2^{Θ} is the set of all the disjunctions of decisions θ_i in the classification context: $2^{\Theta} = \{\emptyset, \{\theta_1\}, \{\theta_2\}, \{\theta_1 \cup \theta_2\}, \ldots, \Theta\}$. The decisions or classes θ_i must be exclusive but not necessarily exhaustive. The definition of mass functions depend on the context, but generic approaches can be used. We will use here a model based on probabilities proposed in [1]. There are many rules of combination in the theory of belief functions such as the conjunctive and the disjunctive one. The conjunctive combination, introduced by Dempster in its normalized form, combines the mass functions considering the intersections between the elements of 2^{Θ} [9, 12]. This combination is formulated as follows for M mass functions, $\forall A \in 2^T heta$:

$$m(A) = \sum_{B_1 \cap B_2, \ldots \cap B_M = A} \prod_{j=1}^{M} m_j(B_j)$$

The obtained mass is the combination of the mass functions of each different sources. Form this mass function, the decision to find the best class θ_i for the considered observation can be made with the pignistic probability. The pignistic probability is defined by:

$$bet(\theta_i) = \sum_{A \in 2^{\Theta}, \theta_i \in A} \frac{m(A)}{|A|(1 - m(\emptyset))} \tag{2}$$

where $|A|$ is the cardinality of A. This criterion is employed in a probabilistic context of decision. In the next section, we present the proposed approach to fuse the results of clustering and supervised classification.

4 Fusion of Supervised Classification and Clustering Results Using the Theory of Belief Functions

The most researches made in fusion are dealing with unsupervised classification (such as in [3]) or with supervised classification (such as in [12]). For the fusion of supervised classification and clustering was essentially done to deploy the unsupervised classification to make the learning of the supervised one [6, 10]. That is not the goal in this paper.

The proposed approach in this article fuses both types of classification to improve the results. Our approach is based on two main steps: the first one is to apply the clustering and the supervised classification on the learning database separately; the second step consists to fuse the results of classification approaches. Based on the two different outputs, we try to make a compromise between both classifiers. We must take into account the bad representation of the cluster's borders in clustering and the bad learning in supervised classification. We model that through the theory of belief functions. Therefore, as inputs of our process, we must define the mass functions of both sources: supervised and unsupervised classification. How to model these mass functions? First, to define the mass function for the supervised source, we choose the probabilistic model of Appriou [1] previously used with success.

Therefore, we define a mass function for each object x belonging to a class θ_j, we have n classes. We have for each class θ_j:

$$m_s^j(\theta_j) = \frac{\alpha_{sj} R_s p(\theta_j^f|\theta_i)}{1 + R_s p(\theta_j^f|\theta_i)} \tag{3}$$

$$m_s^j(\theta_j^c) = \frac{\alpha_{sj}}{1 + R_s p(\theta_j^f|\theta_i)} \tag{4}$$

$$m_s^j(\Theta) = 1 - \alpha_{sj} \tag{5}$$

We note by θ_j^f the class affected by the supervised classifier to the object x, by θ_i the real class and by α_{sj} the reliability coefficient of the supervised classification for the class θ_j^f. The conditional probabilities are estimated through the confusion matrix on the learning database:

$$\alpha_{sj} = \max p(\theta_j^f|\theta_i) \forall i = \{1, ..., n\} \tag{6}$$

and

$$R_s = \max_{\theta_j^f} (p(\theta_j^f|\theta_i))^{-1} \tag{7}$$

For the unsupervised source, mass functions must also be defined on the discernment space Θ. However, the classes of Θ are unknown in clustering. We only dispose of clusters without any labels. Therefore the definition of mass function is made by measuring the similarities between clusters and classes found by the supervised classification. If the found clusters are more similar to the classes, the clustering and supervised classification agree with each other. The similarity is calculated using recovery between clusters and classes. A class is considered similar to a cluster if it is recovered totally by the cluster. Therefore the biggest is the number of objects in common the biggest is the similarity. We look for the proportions of found classes $\theta_1^f, \ldots, \theta_n^f$ by the supervised classifier in each cluster [4, 3]. $\forall x \in C_i$ with c the number of clusters found. The mass function for an object x to be in the class θ_j is as follows:

$$m_{ns}(\theta_j) = \frac{|C_i \cap \theta_j^f|}{|C_i|} \tag{8}$$

where $|C_i|$ is the number of elements in the cluster C_i and $|C_i \cap \theta_j^f|$, the number of elements in the intersection between C_i and θ_j^f. Then we discount the mass functions as follows, $\forall A \in 2^\Theta$ by:

$$m_{ns}{}^{\alpha_i}(A) = \alpha_i m_{ns}(A) \tag{9}$$

$$m_{ns}^{\alpha_i}(\Theta) = 1 - \alpha_i (1 - m_{ns}(\Theta)) \tag{10}$$

The discounting coefficient α_i depends on objects. We can not discount in the same way all the objects. An object situated in the center of cluster is considered more representative of the cluster than another one situated on the border for example. The coefficient α_i is defined as (v_i is the center of cluster C_i):

$$\alpha_i = e^{-\|x-v_i\|^2} \tag{11}$$

After calculating the mass functions for the two sources, we can combine using the conjunctive rule and we adopt as decision criterion the maximum of pignistic probability. Based on the construction of our mass functions for the non-supervised classifier, both mass functions cannot be considered cognitively independent. Other combination rules could be used. In our problem we look for known singletons thanks to the use of supervised classification. Each object is affected to a precise class. The pignistic probability is employed because we are in probabilistic context.

5 Experimental Study

In this section we present the obtained results for our fusion approach between supervised classification and unsupervised classification. We conduct our experimental study on different databases coming from generic databases without missing values obtained from the U.C.I repository of Machine Learning databases. The aim is to demonstrate the performance of the proposed method and the influence of the fusion on the classification results. The experience is based on three unsupervised methods such as the fuzzy C-Means (FCM), the *k*-Means and the Mixture Model. For the supervised classification, we use the *k*-Nearest Neighbors and the Bayes Classifier. We show in the Tables 1 and 2 the obtained classification rates on the data before and after the fusion respectively for the *k*-NN with the FCM, the *k*-Means and the mixture model and the Bayes classifier with the FCM, the *k*-Means and the mixture model.

The number of clusters may be equal to the number given by the supervised classification or fixed by the user. The values shown in both tables 1 and 2 are obtained after cross-validation with ten trials of experiments. In each trial, we test

Table 1 Results obtained with *k*-NN and FCM, *k*-Means and Mixture Model. NbC: number of classes, NbCl: number of clusters, NbA: Number of attributes, CR-BF: classification rate obtained before fusion, CR-AF classification rate obtained after fusion

Data	NbC	NbCl	NbA	CR-BF	CR-AF		
					FCM	*k*-Means	Mixture Model
Iris	3	3	5	96.67	100.00	100.00	100.00
Breast- Cancer wisconsin	2	2	11	64.52	80.00	80.00	80.00
Sensor-readings-24	4	4	5	84.00	100.00	100.00	100.00
Habeman	2	2	4	75.17	100.00	100.00	99.34
Abalone	2	2	8	53.10	61.70	61.27	59.42

Table 2 Results obtained with Bayes Classifier and FCM, *k*-Means and Mixture Model. NbC: number of classes, NbCl: number of clusters, NbA: Number of attributes, CR-BF: Classification rate obtained before fusion, CR-AF Classification rate obtained after fusion

Data	NbC	NbCl	NbA	CR-BF	CR-AF		
					FCM	*k*-Means	Mixture Model
Iris	3	3	5	95.33	100.00	100.00	100.00
Breast- Cancer wisconsin	2	2	11	96.00	100.00	100.00	100.00
Sensor-readings-24	4	4	5	52.57	100.00	100.00	100.00
Habeman	2	2	4	73.83	77.74	77.74	77.41
Abalone	2	2	8	51.95	73.08	73.59	66.62

with a test database taken from 10 databases. The fusion effect is remarkable in the table 1. In fact, we obtain a rate greater than 90% for the databases: iris, sensor-readings24 and haberman, a rate equal to 80% for breast-cancer and a rate about 60% for abalone database. In the table 2, we obtain a rate of 100% for iris, breast-cancer, sensor-readings24 and a rate greater than 70% for abalone and haberman. The error rate does not exceed 30% after fusion. We note that the obtained rate after fusion are better than before fusion.

6 Conclusion

This paper proposes a new approach allowing the fusion between supervised classification and clustering. Both methods have limits and problems. The fusion is established to improve the performance of the classification. We make the fusion with the belief function theory. The proposed approach showed encouraging results on classical and real databases. This work can be spread by studying results on imprecise and uncertain databases and on database with missing data. The final goal of this work is to apply the approach in very difficult applications such as sonar and medical images where the learning is difficult due to an incomplete knowledge of the reality.

References

1. Appriou, A.: Discrimination multisignal par la théorie de l'évidence. In: Décision et Reconnaissance des Formes en Signal. Hermes Science Publication (2002)
2. Campedel, M.: Classification supervisée. Telecom Paris (2005)
3. Forestie, G., Wemmert, C., Gançarski, P.: Multisource Images Analysis Using Collaborative Clustering. EURASIP Journal on Advances in Signal Processing 11, 374–384 (2008)
4. Gançarski, P., Wemmert, C.: Collaborative multi-strategy classification: application to per-pixel analysis of images. In: Proceedings of the 6th International Workshop on Multimedia Data Mining: Mining Integrated Media and Complex Data, vol. 6, pp. 595–608 (2005)

5. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer Theory. IEEE Transactions on Systems, Man and Cybernetics 25(5), 904–913 (1995)
6. Guijarro, M., Pajares, G.: On combining classifiers through a fuzzy multi-criteria decision making approach: Applied to natural textured images. Expert Systems with Applications 39, 7262–7269 (2009)
7. Masson, M., Denoeux, T.: Clustering interval-valued proximity data using belief functions. Pattern Recognition 25, 163–171 (2004)
8. Masson, M., Denoeux, T.: Ensemble clustering in the belief functions framework. International Journal of Approximate Reasoning 52(1), 92–109 (2011)
9. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics 22(3), 418–435 (1992)
10. Urszula, M.K., Switek, T.: Combined Unsupervised-Supervised Classification Method. In: Proceedings of the 13th International Conference on Knowledge Based and Intelligent Information and Engineering Systems: Part II, vol. 13, pp. 861–868 (2009)
11. Wemmert, C., Ganarski, P.: A Multi-View Voting Method to Combine Unsupervised Classifications. In: Proceedings of the 2nd IASTED International Conference on Artificial Intelligence and Applications, vol. 2, pp. 447–453 (2002)
12. Martin, A.: Comparative study of information fusion methods for sonar images classification. In: Proceeding of the 8th International Conference on Information Fusion, vol. 2, pp. 657–666 (2005)
13. Prudent, Y., Ennaji, A.: Clustering incrémental pour un apprentissage distribué : vers un système volutif et robuste. In: Conférence CAP (2004)

Continuous Belief Functions: Focal Intervals Properties

Jean-Marc Vannobel

Abstract. The set of focal elements resulting from a conjunctive or disjunctive combination of consonant belief functions is regretfully not consonant and is thus very difficult to represent.

In this paper, we propose a graphical representation of the cross product of two focal sets originating from univariate Gaussian pdfs. This representation allows to represent initial focal intervals as well as focal intervals resulting from a combination operation. We show in case of conjunctive or disjunctive combination operations, that the whole domain can be separated in four subsets of intervals having same properties. At last, we focus on identical length focal intervals resulting from a combination. We show that such intervals are organized in connected line segments on our graphical representation.

1 Introduction

1.1 Sources of Information

Consider a source of information $\mathcal{S}_i$ with knowledge modeled by a univariate convex (unimodal and consonant) probability density function $betf_i$ of a continuous random variable X. The support of $betf_i$ is called $\Omega_i = [\Omega_i^-, \Omega_i^+]$ with $\Omega_i^-, \Omega_i^+ \in \mathcal{R}$ [4] . The mode and the variance of $betf_i$ are respectively noted μ_i and σ_i^2. Suppose now $\mathcal{E}_i = (\mathcal{F}_i, m_i)$, the *piece of evidence* deduced from $betf_i$. $\mathcal{E}_i$ is totally described by the pair composed of m_i, the Least Commited isopignistic *basic belief density (bbd)* deduced from $betf_i$ [1] [4] and $\mathcal{F}_i = \{I \subseteq \Omega_i | m_i(I) > 0\}$, the focal set of intervals with elements in Ω_i.

Jean-Marc Vannobel
LAGIS, Université Lille1
e-mail: jean-marc.vannobel@univ-lille1.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 93–100.
springerlink.com

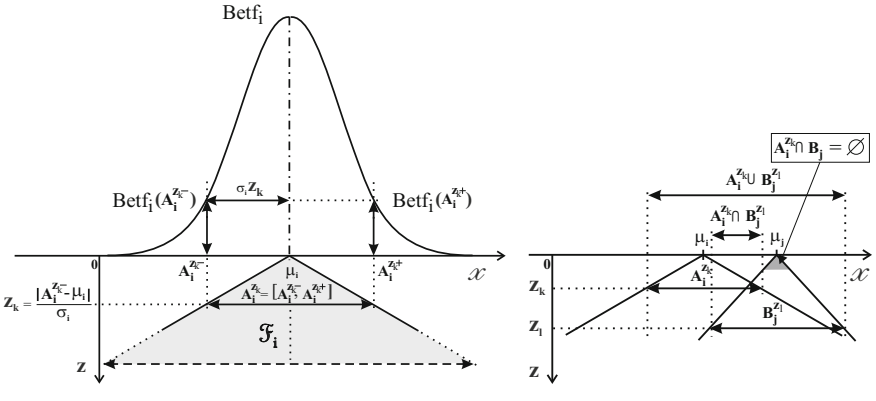
1.2 Focal Intervals

An interval $A = [A^-, A^+]$ with $A^-, A^+ \in \mathcal{R}$ such as $m_i(A) \neq 0$ is called *focal interval* of $\mathcal{E}_i$ thus $A \in \mathcal{F}_i$. All elements of $\mathcal{F}_i$ are nested intervals in case of a consonant pdf $betf_i$ and correspond to horizontal cuts of $betf_i$ as shown in figure 1(a). It is convenient to label the elements of $\mathcal{F}_i$ according to their inclusion order by a continuous index. This can be done for instance wrt the pdf value at focal interval bounds [2] or wrt the half-length of the focal interval [3]. This last option allows in general to define a single bbd's expression for a whole family of pdfs [6]. In case of symmetrical pdfs like Gaussian ones as well as Laplace ones, focal intervals can be labeled by an index z such as $A^z = [A^{z-}, A^{z+}]$ with:

$$z = \frac{|x - \mu|}{\sigma}, \; z \in \mathcal{R}^+, \tag{1}$$

$$A^{z-} = \mu - \sigma z, A^{z-} \in [\Omega^-, \mu], \tag{2}$$

$$A^{z+} = \mu + \sigma z, A^{z+} \in [\mu, \Omega^+]. \tag{3}$$



(a) Focal intervals domain $\mathcal{F}_i$ resulting from a Gaussian pdf.

(b) Intersection and union of focal intervals resulting from two Gaussian pdfs.

Fig. 1 Focal intervals graphical representation relatively to the z label value.

2 Focal Sets Graphical Representations

2.1 General Considerations

We consider in what follows two pieces of evidence $\mathcal{E}_i = (\mathcal{F}_i, m_i)$ and $\mathcal{E}_j = (\mathcal{F}_j, m_j)$ deduced respectively from the Gaussian pdfs $betf_i(x; \mu_i, \sigma_i^2)$ and $betf_j(x; \mu_j, \sigma_j^2)$ with $\mu_i \leq \mu_j$. Focal intervals are denoted by $A_i^{z_k}$ with z_k the value taken by z_i the label obtained using relation (1). We assume the use of Gaussian pdfs since many sensors model uncertainty by such pdfs but any other symmetrical bell shaped pdfs like Laplace or Cauchy ones would also serve the purpose. $\mathcal{F}_{i,j}$ is the focal set resulting from a conjunctive (resp. disjunctive) combination of $\mathcal{E}_i$ and $\mathcal{E}_j$. Elements of $\mathcal{F}_{i,j}$ correspond to the non empty intersection (resp. union) of pairs in $\mathcal{F}_i \times \mathcal{F}_j$. The content of $\mathcal{F}_{i,j}$ and the length dependencies between its elements depend of course on the chosen combination rule.

2.2 Bell Shaped Probability Density Functions

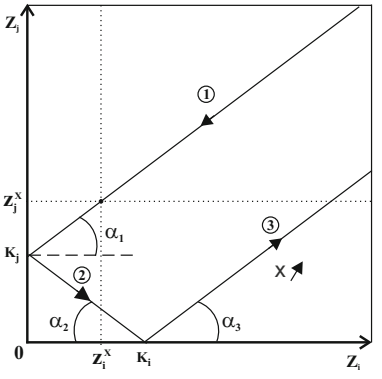
The graphical representation proposed in figure 1(a) shows the focal set $\mathcal{F}_i$ obtained from a Gaussian pdf $Betf_i$. Elements of $\mathcal{F}_i$ are ordered wrt label z, differing in that point from the graphical representation proposed by Strat [5]. When labeling focal intervals wrt their length as defined in (1), the focal set $\mathcal{F}_i$ is encompassed by two symmetrical half-lines defining an isosceles triangle. Equations of these half-lines are deduced from relations (2) and (3) and correspond to:

$$\begin{cases} z = \frac{|A_i^{z^-} - \mu_i|}{\sigma_i}, \\ z = \frac{|A_i^{z^+} - \mu_i|}{\sigma_i}. \end{cases} \tag{4}$$

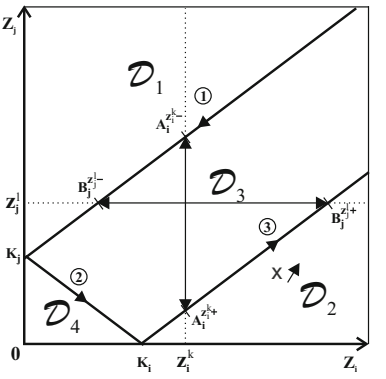
As shown in figure 1(b), this is a convenient way to graphically compare focal intervals coming from different focal sets. One can see in this figure the result of the intersection or union of two intervals $A_i^{z_k} \in \mathcal{F}_i$ and $B_j^{z_l} \in \mathcal{F}_j$ which are indexed resp. by z_k and z_l. For instance, it also allows to show the domain of intervals $B_j \in \mathcal{F}_j$ that do not intersect with $A_i^{z_k}$ (if any). It is obvious from relation (1) that the label value of these B_j intervals is in $[0, \frac{|A_i^{z_k+} - \mu_j|}{\sigma_j})$.

2.3 Bounds Relations of $\mathcal{F}_i \times \mathcal{F}_j$ Elements

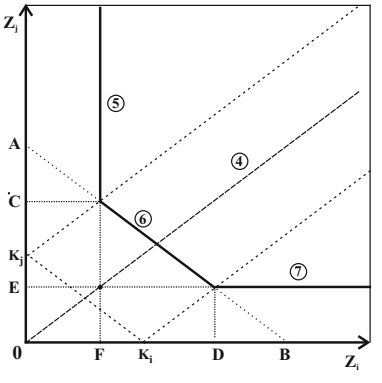
Figure 2(a) shows the pairs of intervals $(A_i^{z_i^x}, B_j^{z_j^x}) \in \mathcal{F}_i \times \mathcal{F}_j$ having a common bound $x \in \Omega$. The index pairs $(z_i^x, z_j^x) \in \mathcal{R}^{+2}$ corresponding to



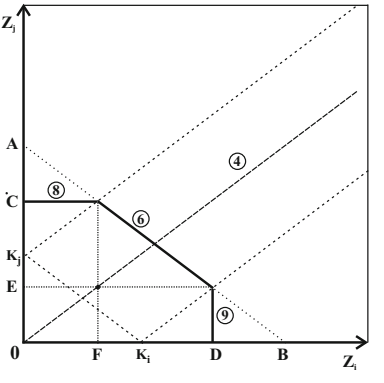
(a) Pairs of intervals having a common lower or upper bound



(b) Domains of identical properties



(c) Same length intervals resulting from a conjunctive combination



(d) Same length intervals resulting from a disjunctive combination

Fig. 2 Graphical representations of focal intervals properties.

$(A_i^{z_i^x}, B_j^{z_j^x})$ draw the lines ①, ② and ③ as shown in Figure 2(a). These lines are defined by[1]:

$$\begin{cases} ① : z_j^x = \frac{|\mu_i - \mu_j|}{\sigma_j} + \frac{\sigma_i}{\sigma_j} z_i^x, & x \in [-\infty, \mu_i], \\ ② : z_j^x = \frac{|\mu_i - \mu_j|}{\sigma_j} - \frac{\sigma_i}{\sigma_j} z_i^x, & x \in [\mu_i, \mu_j], \\ ③ : z_j^x = \frac{-|\mu_i - \mu_j|}{\sigma_j} + \frac{\sigma_i}{\sigma_j} z_i^x, & x \in [\mu_j, +\infty]. \end{cases} \tag{5}$$

[1] Proof is not given here due to lack of space.

Pairs (z_i^x, z_j^x) on the half line called ① lead to $x \leq \mu_i$ such as:

$$\begin{cases} A_i^{z_i^x -} = B_j^{z_j^x -} = x, \\ A_i^{z_i^x +} \leq B_j^{z_j^x +}. \end{cases} \tag{6}$$

Pairs (z_i^x, z_j^x) on line segment ② correspond to $x \in [\mu_i, \mu_j]$ such as:

$$A_i^{z_i^x +} = B_j^{z_j^x -} = x. \tag{7}$$

At last, pairs (z_i^x, z_j^x) on the half line ③ correspond to $x \geq \mu_2$ such as:

$$\begin{cases} A_i^{z_i^x +} = B_j^{z_j^x +} = x, \\ A_i^{z_i^x -} \leq B_j^{z_j^x -}. \end{cases} \tag{8}$$

To outline existing partial conflict $k_{i,j}$ between the agents $\mathcal{E}_i$ and $\mathcal{E}_j$ [6], the z label values of the modes are denoted by K_i and K_j:

$$\begin{cases} K_i = \frac{|\mu_j - \mu_i|}{\sigma_i}, \\ K_j = \frac{|\mu_j - \mu_i|}{\sigma_j}. \end{cases} \tag{9}$$

Relations (5) show that the absolute value of line directions of ①, ② and ③ are equal to $|\text{arctg}(\frac{\sigma_i}{\sigma_j})|$ and correspond to angles α_1, α_2 and α_3 in figure 2(a).

2.4 Focal Intervals Intersection or Union Overview

As shown in figure 2(b), focal intervals $A_i^{z_k} \in \mathcal{F}_i$ and $B_j^{z_l} \in \mathcal{F}_j$ can directly be drawn on the chart by respectively vertical and horizontal line segments since ①, ② and ③ correspond to the focal intervals bounds. The path defined by half-lines ①, ② and ③ covers Ω. It becomes thus easy to analyze pairs in $\mathcal{F}_i \times \mathcal{F}_j$ to deduce their intersection or union. For instance, intervals shown in figure 2(b) are such as $A_i^{z_k} \cap B_j^{z_l} = \left[B_j^{z_l -}, A_i^{z_k +}\right]$ and $A_i^{z_k} \cup B_j^{z_l} = \left[A_i^{z_k -}, B_j^{z_l +}\right]$. When considering z_k and z_l respectively as horizontal and vertical cursors it is also possible to find the z limits of intervals intersecting or including one another or not.

2.5 Particular Domains

Figure 2(b) shows also that the ①, ② and ③ lines separate $\mathcal{F}_i \times \mathcal{F}_j$ in four domains called $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ ans $\mathcal{D}_4$. For $\mu_i < \mu_j$, $z_i \geq 0$, $z_j \geq 0$, defining pairs $(A_i^{z_i}, B_j^{z_j}) \in \mathcal{F}_i \times \mathcal{F}_j$, we have:

$$\begin{cases} z_j > K_j + \frac{\sigma_i}{\sigma_j} z_i & \Rightarrow A_i^{z_i} \subset B_j^{z_j} & (\mathcal{D}_1), \\ z_j < -K_j + \frac{\sigma_i}{\sigma_j} z_i & \Rightarrow A_i^{z_i} \supset B_j^{z_j} & (\mathcal{D}_2), \\ z_j < K_j + \frac{\sigma_i}{\sigma_j} z_i, \quad z_j > K_j - \frac{\sigma_i}{\sigma_j} z_i, \quad z_j > -K_j + \frac{\sigma_i}{\sigma_j} z_i & & \\ \qquad \Rightarrow A_i^{z_i} \cap B_j^{z_j} \notin \left\{ \emptyset, A_i^{z_i}, B_j^{z_j} \right\} & & (\mathcal{D}_3), \\ z_j < K_j - \frac{\sigma_i}{\sigma_j} z_i, & \Rightarrow A_i^{z_i} \cap B_j^{z_j} = \emptyset & (\mathcal{D}_4). \end{cases} \tag{10}$$

When $\mu_1 = \mu_2$ only $\mathcal{D}_1$ and $\mathcal{D}_2$ exist and are separated by the half-line $z_2 = \frac{\sigma_1}{\sigma_2} z_1$.

2.6 Consonant Subsets of $\mathcal{F}_{i,j}$

The focal set $\mathcal{F}_{i,j}$ obtained after a conjunctive or a disjunctive combination operation of $\mathcal{E}_i$ and $\mathcal{E}_j$ is composed of an infinite number of nested focal intervals subsets. These consonant subsets appear both in $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$ shown in figure 2(b) and are partially represented in figure 1(b) with the dark gray area. These subsets are not disjoint and consequently if the same interval belongs to several of them, it is necessary to integrate to get its total weight. The domain $\mathcal{D}_4$ differs from the other ones as it is empty in case of a conjunctive combination operation and composed of nested non convex intervals in case of a disjunctive one.

3 Same Length Intervals Resulting from Intersection and Union Operations of Focal Ones

3.1 Intersection of Focal Intervals

The length l of the intersection of the pairs $(A_i^{z_i}, B_j^{z_j}) \in \mathcal{F}_i \times \mathcal{F}_j$ can be deduced from the characteristics of $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ and $\mathcal{D}_4$ given by relations (10). l only depends on the pdfs $bet f_i$ and $bet f_j$ parameters and the focal intervals indexes z_i and z_j:

$$\begin{cases} \mathcal{D}_1 : A_i^{z_i} \cap B_j^{z_j} = A_i^{z_i}, & l(A_i^{z_i} \cap B_j^{z_j}) = 2\sigma_i z_i, \\ \mathcal{D}_2 : A_i^{z_i} \cap B_j^{z_j} = B_j^{z_j}, & l(A_i^{z_i} \cap B_j^{z_j}) = 2\sigma_j z_j, \\ \mathcal{D}_3 : A_i^{z_i} \cap B_j^{z_j} = [B_j^{z_j -}, A_i^{z_i +}], & l(A_i^{z_i} \cap B_j^{z_j}) = -|\mu_2 - \mu_1| + \sigma_i z_i + \sigma_j z_j, \\ \mathcal{D}_4 : A_i^{z_i} \cap B_j^{z_j} = \emptyset, & l(A_i^{z_i} \cap B_j^{z_j}) = 0. \end{cases} \tag{11}$$

We can find the elements of $\mathcal{F}_i \times \mathcal{F}_j$ having an identical intersection length $L = l(A_i^{z_i} \cap B_j^{z_j})$ with the help of relations (11). For instance, the z indexes of these elements in D_3 are:

$$\sigma_j z_j = L + |\mu_2 - \mu_1| - \sigma_i z_i \tag{12}$$

thus:

$$z_j = A - \frac{\sigma_i}{\sigma_j} z_i \text{ with } A = \frac{L}{\sigma_j} + K_j. \tag{13}$$

This is graphically represented in figure 2(c) by the line segment ⑥ bounded by points (F, C) and (D, E) and crossing the point $(z_i, z_j) = (0, A)$ with A as defined in relation (13). Pairs $(A_i^{z_i}, B_j^{z_j})$ in $\mathcal{D}_1$ having also an intersection length equal to L correspond to the points of the half-line ⑤ crossing (F, C) since $A_i^{z_i} \cap B_j^{z_j} = A_i^{z_i}$ in $\mathcal{D}_1$. Pairs of intervals on the half line ⑦ crossing (D, E) in figure 2(c) have an intersection length equal to L too since $A_i^{z_i} \cap B_j^{z_j} = B_j^{z_j}$ in $\mathcal{D}_2$.

At last, values A to F are quite interdependent depending on K_i and K_j values. This is due to the symmetry properties of straight lines ① to ⑦ in figures 2(b) and 2(c). Many relations depending on these parameters lead to the value of L such as $L = 2\sigma_j(C - K_j) = 2\sigma_i(D - K_i)$ for instance.

3.2 Union of Focal Intervals

One can deduce from the relations (10) related to the domains $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$ that for the union operation of pairs $(A_i^{z_i}, B_j^{z_j}) \in \mathcal{F}_i \times \mathcal{F}_j$, we have:

$$\begin{cases} \mathcal{D}_1 : A_i^{z_i} \cup B_j^{z_j} = B_j^{z_j}, & l(A_i^{z_i} \cup B_j^{z_j}) = 2\sigma_j z_j, \\ \mathcal{D}_2 : A_i^{z_i} \cup B_j^{z_j} = A_i^{z_i}, & l(A_i^{z_i} \cup B_j^{z_j}) = 2\sigma_i z_i, \\ \mathcal{D}_3 : A_i^{z_i} \cup B_j^{z_j} = [B_j^{z_j -}, A_i^{z_i +}], & l(A_i^{z_i} \cup B_j^{z_j}) = |\mu_2 - \mu_1| + \sigma_i z_i + \sigma_j z_j, \\ \mathcal{D}_4 : A_i^{z_i} \cap B_j^{z_j} = \emptyset, & l(A_i^{z_i} \cup B_j^{z_j}) = 2(\sigma_i z_i + \sigma_j z_j). \end{cases} \tag{14}$$

As expressed in (14), concerning $\mathcal{D}_3$, one can write:

$$l(A_i^{z_i} \cup B_j^{z_j}) - |\mu_2 - \mu_1| = \sigma_i z_i + \sigma_j z_j. \tag{15}$$

From (15), pairs $(A_i^{z_i}, B_j^{z_j}) \in \mathcal{D}_3$ having a constant union length $L = l(A_i^{z_i} \cup B_j^{z_j})$ satisfy thus:

$$z_j = A - \frac{\sigma_i}{\sigma_j} z_i \text{ with } A = \frac{L}{\sigma_j} - K_j. \tag{16}$$

This relation corresponds in figure 2(d) to the line segment ⑥ bounded by points (F, C) and (D, E) and crossing the point $(z_i, z_j) = (0, A)$ with A as defined in relation (16). Pairs in $\mathcal{D}_1$ having a union length equal to L correspond to the points of the line segment ⑧. This is also the case in $\mathcal{D}_2$ for the points of the line segment ⑨. As in the case of intersection operation, many relations link the value of L to the parameters of pdfs $betf_i$ and $betf_j$. For instance we have $L = 2\sigma_j C = 2\sigma_i D$.

4 Conclusion and Acknowledgment

As we have seen, the focal set $\mathcal{F}_{i,j}$ obtained in case of a conjunctive (resp. disjunctive) combination of two pieces of evidence $\mathcal{E}_i$ and $\mathcal{E}_j$ with consonant focal domains is not as heterogeneous as it seems to be. Intervals belonging to $\mathcal{F}_{i,j}$ are sorted into only four domains. In each of these domains, pairs $(A_i \in \mathcal{F}_i, B_j \in \mathcal{F}_j)$ of focal intervals share common properties regarding intersection $A_i \cap B_j$ and union $A_i \cup B_j$. These four domains can be graphically represented in a linear space where they are separated by straight lines when the focal sets $\mathcal{F}_i$ and $\mathcal{F}_j$ are composed of centered and consonant intervals.

At last, elements of $\mathcal{F}_{i,j}$ having a same length are linked by linear relations. This can be useful in problems where interval lengths have to be taken into account.

Authors are indebted to J. Klein for the review of this work.

References

1. Caron, F., Ristic, B., Duflos, E., Vanheeghe, P.: Least commited basic belief density induced by a multivariate gaussian: formulation with applications. International Journal on Approximate Reasoning 48(2), 419–436 (2008)
2. Doré, P.-E., Martin, A., Khenchaf, A.: Constructing of least commited basic belief density linked to a multimodal probability density. In: COGIS, Paris (Fr) (2009)
3. Ristic, B., Smets, P.: Belief function theory on the continuous space with an application to model based classification. In: Information Processing and Management of Uncertainty in Knowledge Based Systems, IPMU 2004, Paris (Fr), pp. 1119–1126 (2004)
4. Smets, P.: Belief functions on real numbers. International Journal of Approximate Reasoning 40(3), 181–223 (2005)
5. Strat, T.H.: Continuous belief functions for evidential reasoning. In: National Conference on Artificial Intelligence, AAAI 1984, pp. 308–313 (1984)
6. Vannobel, J.-M.: Continuous belief functions: singletons plausibility function in conjunctive and disjunctive combination operations of consonant bbds. In: Proceedings of Workshop on the Theory of Belief Functions, CDROM, Brest (Fr), 6 p (2010)

Game-Theoretical Semantics of Epistemic Probability Transformations

Fabio Cuzzolin

Abstract. Probability transformation of belief functions can be classified into different families, according to the operator they commute with. In particular, as they commute with Dempster's rule, relative plausibility and belief transforms form one such "epistemic" family, and possess natural rationales within Shafer's formulation of the theory of evidence, while they are not consistent with the credal or probability-bound semantic of belief functions. We prove here, however, that these transforms can be given in this latter case an interesting rationale in terms of optimal strategies in a non-cooperative game.

1 Introduction

The theory of evidence (ToE) [21] extends classical probability theory through the notion of *belief function* (b.f.), a mathematical entity which independently assigns probability values to *sets* of possibilities rather than single events. A belief function $b : 2^\Theta \to [0,1]$ on a finite set or *frame* Θ has the form $b(A) = \sum_{B \subseteq A} m_b(B)$, where the function $m_b : 2^\Theta \to [0,1]$ (called *basic probability assignment* or *basic belief assignment* b.b.a.) is both non-negative $m_b(A) \geq 0 \ \forall A \subseteq \Theta$ and normalized $\sum_{A \subseteq \Theta} m_b(A) = 1$. Subsets $A \subset \Theta$ associated with non-zero basic probabilities $m_b(A) \neq 0$ are called *focal elements*. Different operators have been proposed for the combination of two or more belief functions, starting from the orthogonal sum originally formulated by A. Dempster [15]. Special belief functions assigning non-zero masses to singletons only ($m_b(A) = 0$ whenever $|A| > 1, A \subseteq \Theta$) are called *Bayesian* b.f.s, and are in 1-1 correspondence with probability distributions on Θ.

Belief functions possess a number of alternative semantics in terms of multi-valued mappings, random sets [19], inner measures [17], transferable beliefs [25] or hints [18]. In some of his papers [15], Dempster claimed that the mass $m_b(A)$ associated with a non-singleton event $A \subseteq \Theta$ could be understood as a "floating probability

Fabio Cuzzolin
Oxford Brookes University, Oxford, UK
e-mail: fabio.cuzzolin@brookes.ac.uk

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 101–108.
springerlink.com

mass". This has originated a popular but controversial interpretation of belief functions b as convex sets $\mathscr{P}[b]$ of probabilities (often called *consistent* with b) determined by sets of lower and upper bounds on their probability values: $\mathscr{P}[b] \doteq \{p \in \mathscr{P} : b(A) \leq p(A) \leq pl_b(A)\ \forall A \subseteq \Theta\}$, where the plausibility function $pl_b : 2^\Theta \to [0,1]$, $pl_b(A) = 1 - b(A^c)$ carries the same evidence as b. In [22] Shafer disavowed any probability-bound interpretation, also criticized by Walley as incompatible with Dempster's rule of combination [31], a position later seconded by Dempster [14].

Probability Transformation of Belief Functions. Nevertheless, the relation between belief and probability in the theory of evidence has been an important subject of study, and a number of papers have been published on the issue of probability transform [13]. A decision based approach to the problem is the foundation of Smets' "Transferable Belief Model" [25], in which belief functions are defined directly in terms of basis belief assignments ("credal" level), while decisions are made via the *pignistic probability* $BetP[b](x) = \sum_{A \supseteq \{x\}} \frac{m_b(A)}{|A|}$, generated by what he calls the *pignistic transform*: $BetP : \mathscr{B} \to \mathscr{P}$, $b \mapsto BetP[b]$. The pignistic probability is the result of a redistribution process in which the mass of each focal element A is re-assigned to all its elements $x \in A$ on an equal basis, and is perfectly compatible with the upper-lower probability semantics of b.f.s, as it is the center of mass of the polytope $\mathscr{P}[b]$ of consistent probabilities [4].

Other proposals have been recently brought forward by Dezert et al. [16], Burger [3], Sudano [27] and others, based on redistribution processes similar to that of the pignistic transform. In addition, two new Bayesian approximations of belief functions have been derived from purely geometric considerations [7] in the context of the geometric approach to the ToE [8], in which belief and probability measures are represented as points of a Cartesian space.

Relative Plausibility and Belief Transforms. Originally developed by Voorbraak [29] as a probabilistic approximation intended to limit the computational cost of operating with belief functions in the Dempster-Shafer framework, the *plausibility transform* [5] has later been supported by Cobb and Shenoy in virtue of its commutativity properties with respect to Dempster's sum. Initially defined in terms of commonality values, the plausibility transform $\tilde{pl} : \mathscr{B} \to \mathscr{P}$, $b \mapsto \tilde{pl}[b]$ maps each belief function b to the probability distribution $\tilde{pl}[b] = \tilde{pl}_b$ obtained by normalizing the plausibility values $pl_b(x)$[1] of the element of Θ: $\tilde{pl}_b(x) = \frac{pl_b(x)}{\sum_{y \in \Theta} pl_b(y)}$.

We call the output $\tilde{pl}_b$ of the plausibility transform *relative plausibility of singletons*. Voorbraak proved that his (in our terminology) relative plausibility of singletons is a perfect representative of b when combined with other probabilities $p \in \mathscr{P}$ through Dempster's rule $\oplus$: $\tilde{pl}_b \oplus p = b \oplus p$ for all $p \in \mathscr{P}$.

Dually, a *relative belief transform* $\tilde{b} : \mathscr{B} \to \mathscr{P}$, $b \mapsto \tilde{b}[b]$ mapping each belief function to the corresponding *relative belief of singletons* $\tilde{b}(x) = \frac{b(x)}{\sum_{y \in \Theta} b(y)}$ can be defined.

The notion of relative belief transform (under the name of "normalized belief of singletons") has first been proposed by Daniel [13]. Some preliminary analyses of

[1] With a harmless abuse of notation we denote the values of b.f.s and pl.f.s on a singleton x by $b(x), pl_b(x)$ rather than $b(\{x\}), pl_b(\{x\})$.

the relative belief transform and its close relationship with the (relative) plausibility transform have been presented in [9, 10]. A detailed discussion of the geometrical properties of $\tilde{b}$ and $\tilde{pl}$ has been given in [11]. In [10], in particular, the author has shown that plausibility and belief transforms both commute with Dempster's rule of combination, and meet a number of dual properties with respect to the orthogonal sum, therefore forming what we call the "epistemic" family of transforms. In opposition, an "affine" family can be defined which groups together those transforms which commute with affine combination, and fit in the probability-bound interpretation of belief functions.

Paper Contribution. In this paper, instead, we point out that, even though they are not consistent with the credal set of probabilities dominated by the original belief function, plausibility and belief transforms can be provided in this interpretation with an interesting betting semantics within an adversarial game theory scenario [28]. In this scenario, inspired by Wald's minimax/maximin model [30], an opponent representing the uncertainty encoded by a b.f. is free to pick any probability function in the set determined by the latter: the decision maker's goal is to maximize their minimal expected reward (or minimize their maximal expected loss).

2 A Game/Utility Theory Interpretation

It can be proven that a probability distribution on Θ is consistent with a belief function b iff it is the result of a *redistribution process*, in which the mass of each focal element is shared between its elements in an arbitrary proportion [12]. However, neither the relative belief of singletons nor the relative plausibility of singletons (unlike Smets' pignistic function) are consistent in this sense: indeed, it is easy to prove that they are not the result of such a redistribution process [12]. Nevertheless, an interesting interpretation for them under the probability-bound semantic can be provided in a game/utility theory context [28, 26].

Strat's Carnival Wheel Scenario. Consider the following scenario, inspired by Strat's expected utility approach to decision making with belief functions [26, 20].

In a country fair, by paying a fixed fee c, people get the chance to spin a carnival wheel divided into a number of sectors labeled, say, $\Theta = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. In return, they get an amount $r(x)$ which varies with the label $x \in \Theta$ of the sector that stops at the top, so that the gain or "utility" of each outcome for the player is $u(x) = r(x) - c$, while their loss is, dually, $l(x) = -u(x) = c - r(x)$.

The game amounts to a "lottery" (probability distribution), in which the probability of each outcome is proportional to the area covered on the wheel. People are asked to make a binary decision: to play/not to play. A rational behavior on the player's side consists on computing their expected utility $\sum_{x \in \Theta} u(x)p(x)$ and decide to play if the latter is positive: the decision, lacking any uncertainty, is trivial.

Cloaked Carnival Wheel. Strat therefore introduces a more challenging scenario, in which the fair's manager decides to make the game more interesting by covering part of the wheel. People are still asked whether they want to spin the wheel or not, knowing that the manager is allowed to rearrange the hidden sector of the wheel as they pleases (see Figure 1). Clearly, this new situation amounts to

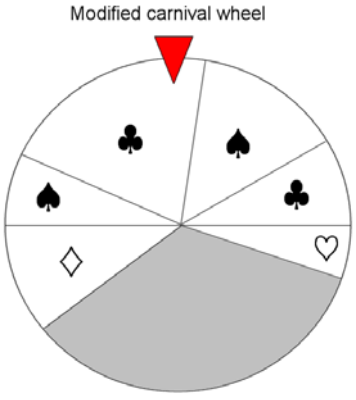


Fig. 1 The modified carnival wheel: part of the spinning wheel is cloaked.

a set of possible lotteries which can be described as a belief function, in particular one in which the fraction of area associated with the hidden sector is assigned as mass to the whole decision space $\{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. If additional (partial) information is provided, for instance that $\diamondsuit$ cannot appear in the hidden sector, different belief functions must be chosen instead. Regardless the particular belief function b (seen as a set of probabilities) at hand, the rule allowing the manager to pick an arbitrary distribution of outcomes in the hidden section mathematically translates into allowing them to choose *any* probability distribution $p \in \mathscr{P}[b]$ consistent with b in order to damage the player. Strat uses this situation as a way of introducing upper and lower bounds to the expected utility [28]

$$E(u) = \sum_{x \in \Theta} u(x)p(x),$$

of the player, induced by the upper and lower bounds to probabilities associated with the belief function describing the set of lotteries [26].

A Modified Carnival Wheel Scenario. Let us consider, instead, a modified scenario in which players are asked (after paying the usual fee c) *to bet on a single outcome* $x \in \Theta$. What is the expected utility of the player in this case? Clearly:

$$E(u) = \sum_{y \in \Theta} p(y)u'(y), \qquad u'(y) = \begin{cases} 0 & y \neq x \\ u(x) = r(x) - c & y = x \end{cases}$$

so that $E(u) = \sum_{y \in \Theta} p(y)u'(y) = p(x)u(x)$.

Suppose that the aim of the player is to play conservatively, and maximize their *worst case* expected utility $p(x)u(x)$, under the uncertainty given by the countermove by the fair's manager: which outcome (singleton) should they pick?

Wald's Minimax Model. This situation can be naturally described by *Wald's maximin model* [30], a non-probabilistic, robust decision making model in which

the optimal decision is one whose worst outcome is at least as good as the worst outcome in any other case. Mathematically, it reads as follows:

$$f^* = \max_{a \in \mathscr{A}} \min_{s \in \mathscr{S}(a)} f(a,s) \tag{1}$$

where $\mathscr{A}$ denotes the set of alterative actions/decisions/strategies, $\mathscr{S}(a)$ denotes the set of states associated with action s, and $f(a,s)$ denotes the return of strategy a taking place in the state s. The model represents a 2-person game in which the max player plays first, making a move a: in response, the second (min) player selects the available state ($s \in \mathscr{S}(a)$) which minimizes the return for the first player.

Wald's model (1) represents a major simplification of the classic 2-person zero sum game [2], in which the two players decide without being aware of the other's choice, while in this case the players choose sequentially.

A Minimax Model of the Carnival Wheel, and Relative Beliefs. Clearly, our scenario can be described by a maximin model (1), in which: the set of possible actions corresponds to the set of outcomes of the lottery $\mathscr{A} = \Theta$; the set of possible states the second player can pick from does not depend on $a = x$, and is the set $\mathscr{S}(a) = \mathscr{S} = \mathscr{P}[b]$ of probability distributions consistent with b; and finally, the return is the player's expected utility $E(u) = p(x)u(x)$ under the constraint of having to pick a single outcome: $f(a,s) = f(x,p) = p(x)u(x)$ which is a function of the lottery outcome only. The problem may therefore be described as:

$$x_{maximin} = \arg\max_{x \in \Theta} \min_{p \in \mathscr{P}[b]} u(x)p(x). \tag{2}$$

Now, in the probability-bound interpretation of belief functions, the belief value of each singleton $x \in \Theta$ measures the minimal support x can receive from a distribution of the family associated with b: $b(x) = \min_{p \in \mathscr{P}[b]} p(x)$. Therefore

$$\begin{aligned} x_{maximin} &= \arg\max_{x \in \Theta} \min_{p \in \mathscr{P}[b]} u(x)p(x) = \arg\max_{x \in \Theta} \left(u(x) \min_{p \in \mathscr{P}[b]} p(x) \right) \\ &= \arg\max_{x \in \Theta} u(x)b(x) = \arg\max_{x \in \Theta} u(x)\tilde{b}(x) \end{aligned}$$

is the optimal decision for the player since, by normalizing $b(x)$ to obtain the relative belief of singletons, the maximal decision is obviously preserved.

If, in particular, the utility function is constant (i.e., no element of Θ can be preferred over the others), the best possible defensive strategy $x_{maximin}$ aimed at maximizing the minimal return of the possible outcomes is/are the peak(s) of the relative belief of singletons. In the example of Figure 1, as ♣ is the outcome which occupies the largest share of the visible part of the wheel, the safest bet (the one which guarantees the best expected return in the worst case) is indeed ♣.

Dual Maximin Model, and Relative Plausibilities. The dual *maximin* model describes the case in which the player moves first again, but this time to minimize the worst possible expected loss. In the modified carnival wheel scenario, once again, when people are asked to bet on a single outcome, their expected loss is $E(l) = l(x)p(x)$ so that:

$$x_{minimax} = \arg\min_{x\in\Theta} \max_{p\in\mathscr{P}[b]} l(x)p(x) = \arg\min_{x\in\Theta} l(x) \max_{p\in\mathscr{P}[b]} p(x) = \arg\min_{x\in\Theta} l(x)pl_b(x), \tag{3}$$

as $pl_b(x) = \max_{p\in\mathscr{P}[b]} p(x)$ measures the maximal possible support to x by a distribution consistent with b. Since $l(x) = c - r(x) = -u(x)$, and after noting that normalizing the plausibility of singletons does not alter the above optimization problem, the outcome/action which minimizes the maximal expected loss is:

$$x_{minimax} = \arg\min_{x\in\Theta} -u(x)\tilde{pl}_b(x) = \arg\max_{x\in\Theta} u(x)\tilde{pl}_b(x).$$

Once again, if in particular the loss (utility) function is constant, then the elements whose relative plausibility is maximal are the best possible defensive strategies aimed at minimizing the maximum possible loss.

In both the maximin and the minimax scenarios, relative belief and plausibility of singletons play a crucial role in determining the safest betting strategy in an adversarial game in which the decision maker has to minimize their maximal expected loss/maximize their minimal expected return under uncertainty representable as a belief function, interpreted as a set of lower/upper bounds to probability values.

The Role of Expected Utility in Pignistic Transform. It can be useful to compare our scenario based on the maximin/minimax model with classical expected utility theory [28]. There, a decision maker can choose between a number of "lotteries" (probability distributions) $p_i(x)$, in order to maximize the expected return or utility $E(p_i) = \sum_x u(x)p_i(x)$ of the lottery. Here, the "lottery" is chosen by their opponent (given the available partial evidence), and the decision maker is left with betting on the safest strategy (element of Θ).

However, a look at how expected utilities are employed in the justification of Smets' pignistic transform provides a useful hint on a natural generalization of the proposed scenario. In [24], the author proves the necessity of the linearity axiom (and therefore of the pignistic transform) by maximizing the following expected utility (our notation), where $p = BetP$ is the pignistic function: $E[u] = \sum_{x\in\Theta} u(a,x)p(x)$. In this case, the set of possible actions (decision) $\mathscr{A}$ and the set Θ of possible outcomes of the problem are distinct, and the utility function is defined on $\mathscr{A}\times\Theta$.

A Generalization of the Proposed Scenario. We can then wonder what happens if we generalize our scenario to the more general case in which the second player still impersonates the uncertainty on the lottery represented by a belief function, but the set of actions $\mathscr{A}$ is fully distinct from Θ, so that a utility function $u : \mathscr{A}\times\Theta \to \mathbb{R}^+$ can be defined. Let us focus on the *maximin* form, while forgetting the carnival wheel situation to move to a more abstract setting.

In this case, once again, the max player moves first and picks an action $\bar{a} \in \mathscr{A}$. This fixes a utility profile $u(\bar{a},x)$, $x \in \Theta$ for the elements of Θ: the first player now has a non-zero utility $u(\bar{a},x)$ for any possible outcome x of the problem, so that their expected utility is obviously given by $\sum_{x\in\Theta} u(\bar{a},x)p(x)$, which depends on the actual probability distribution describing the problem. The min player at this point selects the admissible probability distribution $p \in \mathscr{P}[b]$ which minimizes the expected return of the max player. The overall model is in this more general case:

$$a_{maximin} = \arg\max_{a \in \mathscr{A}} \min_{p \in \mathscr{P}[b]} \Big(\sum_{x \in \Theta} u(a,x) p(x) \Big). \tag{4}$$

We can notice that, in this new situation, $\arg\max_{a\in\mathscr{A}} \min_{p\in\mathscr{P}[b]} (\sum_{x\in\Theta} u(a,x)p(x)) \neq \arg\max_{a\in\mathscr{A}} \big(\sum_{x\in\Theta} u(a,x) \min_{p\in\mathscr{P}[b]} p(x)\big) = \arg\max_{a\in\mathscr{A}} \big(\sum_{x\in\Theta} u(a,x)b(x)\big) = \arg\max_{a\in\mathscr{A}} \big(\sum_{x\in\Theta} u(a,x)\tilde{b}(x)\big)$ for the worst case probability distribution $p^*(x) = \arg\min_{p\in\mathscr{P}[b]} p(x)$ is, in general, different for each $x \in \Theta$, and we cannot simply swap the min and $\sum$ operators. As a consequence, the generalization of Wald's maximin/minimax model to the case in which the second player represents the uncertainty associated with a belief function (in the probability-bound interpretation), but actions/decisions are distinct from the outcomes of the problem is no more a function of belief and plausibility values on singletons, and cannot be solved by using only the knowledge encoded by relative plausibilities and beliefs of singletons. A deeper study of this and more general settings is in order.

3 Conclusions

Epistemic transforms commute with Dempster's rule but they are not consistent with the probability bound interpretation of belief functions. Nevertheless, in this paper we proposed an interesting, novel interpretation of relative belief and plausibility of singletons as tools to provide optimal conservative strategies in a maximin/minimax 2-person game scenario derived from Wald's model, in which a player has to optimize their minimal expected gain/maximal expected loss under epistemic uncertainty in the form of a belief function. The study of more general models will be the goal of further research in the near future.

References

1. Bogler, P.: Shafer-Dempster reasoning with applications to multisensor target identification systems. IEEE Trans. on Systems, Man and Cybernetics 17(6), 968–977 (1987)
2. Bowles, S.: Microeconomics: Behavior, institutions, and evolution. Princeton University Press (2004)
3. Burger, T.: Defining new approximations of belief functions by means of Dempster's combination. In: Proc. of BELIEF, Brest, France (2010)
4. Chateauneuf, A., Jaffray, J.: Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Mathematical Social Sciences 17, 263–283 (1989)
5. Cobb, B., Shenoy, P.: A comparison of Bayesian and belief function reasoning. Information Systems Frontiers 5(4), 345–358 (2003)
6. Cuzzolin, F.: Geometry of Dempster's rule of combination. IEEE Trans. on Systems, Man and Cybernetics B 34(2), 961–977 (2004)
7. Cuzzolin, F.: Two new Bayesian approximations of belief functions based on convex geometry. IEEE Trans. on Systems, Man, and Cybernetics B 37(4), 993–1008 (2007)
8. Cuzzolin, F.: A geometric approach to the theory of evidence. IEEE Trans. on Systems, Man, and Cybernetics C 38(4), 522–534 (2008)

9. Cuzzolin, F.: Dual properties of the relative belief of singletons. In: Proc. of PRICAI, Hanoi, Vietnam, pp. 78–90 (2008)
10. Cuzzolin, F.: Semantics of the relative belief of singletons. In: Workshop on Uncertainty and Logic, Kanazawa, Japan (2008)
11. Cuzzolin, F.: The geometry of consonant belief functions: simplicial complexes of necessity measures. Fuzzy Sets and Systems 161(10), 1459–1479 (2010)
12. Cuzzolin, F.: On the relative belief transform. International Journal of Approximate Reasoning (in press, 2012)
13. Daniel, M.: On transformations of belief functions to probabilities. Int. J. of Intelligent Systems 21(3), 261–282 (2006)
14. Dempster, A.P.: Lindley's paradox: Comment. Journal of the American Statistical Association 77(378), 339–341 (1982)
15. Dempster, A.P.: A generalization of Bayesian inference. In: Classic Works of the Dempster-Shafer Theory of Belief Functions, pp. 73–104 (2008)
16. Dezert, J., Smarandache, F.: A new probabilistic transformation of belief mass assignment. In: Proc. of the 11th International Conference of Information Fusion, pp. 1–8 (2008)
17. Fagin, R., Halpern, J.: Uncertainty, belief and probability. In: Proc. of IJCAI, pp. 1161–1167 (1989)
18. Kohlas, J., Monney, P.-A.: A Mathematical Theory of Hints. An Approach to Dempster-Shafer Theory of Evidence. Springer (1995)
19. Nguyen, H.: On random sets and belief functions. Journal of Mathematical Analysis and Applications 65, 531–542 (1978)
20. Schubert, J.: On 'rho' in a decision-theoretic apparatus of Dempster-Shafer theory. International Journal of Approximate Reasoning 13, 185–200 (1995)
21. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
22. Shafer, G.: Constructive probability. Synthese 48, 309–370 (1981)
23. Shenoy, P.: No double counting semantics for conditional independence. Working Paper No. 307. School of Business, University of Kansas (2005)
24. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. International Journal of Approximate Reasoning 38(2), 133–147 (2005)
25. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66(2), 191–234 (1994)
26. Strat, T.M.: Decision analysis using belief functions. International Journal of Approximate Reasoning 4, 391–417 (1990)
27. Sudano, J.: Equivalence between belief theories and nave Bayesian fusion for systems with independent evidential data. In: Proc. of ICIF, vol. 2, pp. 1239–1243 (2003)
28. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944)
29. Voorbraak, F.: A computationally efficient approximation of Dempster-Shafer theory. International Journal on Man-Machine Studies 30, 525–536 (1989)
30. Wald, A.: Statistical decision functions. Wiley, New York (1950)
31. Walley, P.: Belief function representations of statistical evidence. The Annals of Statistics 15, 1439–1465 (1987)
32. Xu, H., Hsia, Y.-T., Smets, P.: The transferable belief model for decision making in the valuation-based systems. IEEE Trans. on Systems, Man, and Cybernetics 26, 698–707 (1996)
33. Zadeh, L.: A simple view of the Dempster-Shafer theory of evidence and its implications for the rule of combination. AI Magazine 7(2), 85–90 (1986)

Generalizations of the Relative Belief Transform

Fabio Cuzzolin

Abstract. Probability transformation of belief functions can be classified into different families, according to the operator they commute with. In particular, as they commute with Dempster's rule, relative plausibility and belief transforms form one such "epistemic" family, and possess natural rationales within Shafer's formulation of the theory of evidence. However, the relative belief transform only exists when some mass is assigned to singletons. We show here that relative belief is only a member of a class of "relative mass" mappings, which can be interpreted as low-cost proxies for both plausibility and pignistic transforms.

1 Introduction

The theory of evidence (ToE) [14] extends classical probability theory through the notion of *belief function* (b.f.), a mathematical entity which independently assigns probability values to *sets* of possibilities rather than single events. A belief function $b : 2^\Theta \to [0,1]$ on a finite set or *frame* Θ has the form $b(A) = \sum_{B \subseteq A} m_b(B)$, where the function $m_b : 2^\Theta \to [0,1]$ (called *basic probability assignment* or *basic belief assignment* b.b.a.) is both non-negative $m_b(A) \geq 0 \; \forall A \subseteq \Theta$ and normalized $\sum_{A \subseteq \Theta} m_b(A) = 1$. Subsets $A \subset \Theta$ associated with non-zero basic probabilities $m_b(A) \neq 0$ are called *focal elements*. A basic probability assignment m_b can be uniquely recovered from a belief function b by Moebius transform: $m_b(A) = \sum_{B \subseteq A} (-1)^{|A-B|} b(B)$. Special belief functions assigning non-zero masses to singletons only ($m_b(A) = 0$ whenever $|A| > 1$, $A \subseteq \Theta$) are called *Bayesian* b.f.s, and are in 1-1 correspondence with probability distributions on Θ. Different operators have been proposed for the combination of two or more belief functions, starting from the orthogonal sum originally formulated by A. Dempster [10].

Fabio Cuzzolin
Oxford Brookes University, Oxford, UK
e-mail: fabio.cuzzolin@brookes.ac.uk

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 109–116.
springerlink.com

Probability Transformation of Belief Functions. The relation between belief and probability, in particular, has been an important subject of study in the theory of evidence, and a number of papers have been published on the issue of probability transform [9]. Many of these proposals, such as [13] or [18], seek efficient implementations of the rule of combination. A different, decision based approach to probability transformation is the foundation of Smets' "Transferable Belief Model" [15], in which decisions are made via the *pignistic probability* $BetP[b](x) = \sum_{A \supseteq \{x\}} \frac{m_b(A)}{|A|}$, justified via a number of rationality principles. Other proposals have been recently brought forward by Dezert et al. [11], Burger [1], Sudano [17] and others, based on redistribution processes similar to that of the pignistic transform. Two new Bayesian approximations of belief functions have been derived from purely geometric considerations [4] in the context of the geometric approach to the ToE [5], in which belief and probability measures are represented as points of a Cartesian space.

Relative Plausibility and Belief Transforms. Following the efficient implementation approach, Voorbraak [19] has developed a probabilistic approximation intended to limit the computational cost of operating with belief functions in the Dempster-Shafer framework, the *plausibility transform*. Initially defined in terms of commonality values, the plausibility transform $\tilde{pl} : \mathscr{B} \to \mathscr{P}$, $b \mapsto \tilde{pl}[b]$ maps each belief function b onto the probability distribution $\tilde{pl}[b] = \tilde{pl}_b$ obtained by normalizing the plausibility values $pl_b(x)$[1] of the element of Θ:

$$\tilde{pl}_b(x) = \frac{pl_b(x)}{\sum_{y \in \Theta} pl_b(y)}. \tag{1}$$

We call the output (1) of the plausibility transform *relative plausibility of singletons* (r.pl.s.). Voorbraak proved that his (in our terminology) relative plausibility of singletons $\tilde{pl}_b$ is a perfect representative of b when combined with other probabilities $p \in \mathscr{P}$ through Dempster's rule $\oplus$: $\tilde{pl}_b \oplus p = b \oplus p \quad \forall p \in \mathscr{P}$.

Dually, a *relative belief transform* $\tilde{b} : \mathscr{B} \to \mathscr{P}$, $b \mapsto \tilde{b}[b]$ mapping each belief function to the corresponding *relative belief of singletons* (r.b.s.) $\tilde{b}[b] = \tilde{b}$ [6, 8, 12, 9]

$$\tilde{b}(x) = \frac{b(x)}{\sum_{y \in \Theta} b(y)} \tag{2}$$

can be defined. Unlike (1), however, (2) exists iff b assigns some mass to singleton focal sets: $\sum_{x \in \Theta} m_b(x) \neq 0$. The notion of relative belief transform (under the name of normalized belief of singletons) has first been proposed by Daniel [9]. Some preliminary analyses of the relative belief transform and its close relationship with the (relative) plausibility transform have been presented in [6, 8]. A detailed discussion of the geometrical properties of $\tilde{b}$ and $\tilde{pl}$ has been given in [7].

The Epistemic Family of Probability Transforms. Cobb and Shenoy [3] have argued in favor of the plausibility transform as a link between Shafer's theory of evidence (endowed with Dempster's rule) and Bayesian reasoning. They have proved

[1] With a harmless abuse of notation we denote the values of b.f.s and pl.f.s on a singleton x by $b(x), pl_b(x)$ rather than $b(\{x\}), pl_b(\{x\})$.

[2] that the plausibility transform commutes with Dempster's rule, and meets a number of additional properties which they claim "allow an integration of Bayesian and D-S reasoning that takes advantage of the efficiency in computation and decision-making provided by Bayesian calculus while retaining the flexibility in modeling evidence that underlies D-S reasoning":

$$b \oplus p = \tilde{pl}_b \oplus p \; \forall p; \quad \tilde{pl}_b[b_1 \oplus b_2] = \tilde{pl}_b[b_1] \oplus \tilde{pl}_b[b_2];$$
$$b \oplus b = b \vdash \tilde{pl}[b] \oplus \tilde{pl}[b] = \tilde{pl}[b].$$

On our side, we have proved [8] that a similar set of (dual) properties hold for the relative belief transform:

$$pl_b \oplus p = \tilde{b} \oplus p \; \forall p; \quad \tilde{b}[pl_{b_1} \oplus pl_{b_2}] = \tilde{b}[pl_{b_1}] \oplus \tilde{b}[pl_{b_2}];$$
$$pl_b \oplus pl_b = pl_b \vdash \tilde{b}[pl_b] \oplus \tilde{b}[pl_b] = \tilde{b}[pl_b],$$

where $pl_b\oplus$ denotes the extension of Dempster's rule to plausibility measures [8] (seen as pseudo belief functions, i.e., sum functions $pl_b(A) = \sum_{B \subseteq A} \mu_b(B)$ on 2^Θ whose Moebius transform $\mu_b(B)$ can be negative for some $B \subset \Theta$). This supports the existence of a family of probability transformations strongly linked to Shafer's interpretation of the theory of evidence via Dempster's rule, which includes relative belief and relative plausibility transforms, and which we call *epistemic* family, in opposition to the *affine* family of mappings which commute with affine combination [4] (a property that Smets calls "linearity" [15]).

Paper Contribution and Outline. The symmetry/duality between (relative) plausibility and belief is, unfortunately, broken, as the existence of the relative belief of singletons is subject to a strong condition. This stresses the issue of its applicability for, in practice, the situation in which the mass of all singletons is nil is common. However, in Section 2 we point out that relative belief is only a member of a class of *relative mass* transformations which generalize it, are computable even when relative belief is not, and can be interpreted as low-cost proxies for both plausibility and pignistic transforms (Section 3). We discuss their applicability as approximate transformations in two significant scenarios (Section 4).

2 Generalizing the Relative Belief Transform

No matter its semantics and that of its sister plausibility transform, a serious issue with the relative belief of singletons is its applicability. In opposition to relative plausibility, $\tilde{b}$ does not exist for a large class of belief functions (those which assign no mass to singletons). Indeed, in many practical applications there is a bias towards some particular models which are the most exposed to the problem. For example, in "consonant" belief functions [14] at most one focal element is a singleton, therefore the vast majority of the useful information in the b.b.a. is contained in the non-singleton focal elements.

Relative belief is in fact only one element of an entire family of probability transformations. Indeed, $\tilde{b}$ can be thought of as the transform which, given a b.f. b:

1. retains the focal elements of size 1 only, yielding an unnormalized b.f.;
2. computes (indifferently) the latter's relative plausibility/pignistic transformation:

$$\tilde{b}(x) = \frac{\sum_{A \supseteq x, |A|=1} m_b(A)}{\sum_y \sum_{A \supseteq x, |A|=1} m_b(A)} = \frac{m_b(x)}{k_{m_b}} = \frac{\sum_{A \supseteq x, |A|=1} \frac{m_b(A)}{|A|}}{\sum_y \sum_{A \supseteq x, |A|=1} \frac{m_b(A)}{|A|}}.$$

Accordingly, a family of natural generalizations of the relative belief transform is obtained by, given an arbitrary b.f. b:

1. retaining the focal elements of size s only;
2. computing either the resulting relative plausibility ...
3. ... or the associated pignistic transformation.

Now, both alternatives 2) or 3) *yield the same probability distribution*. Indeed, the application of the relative plausibility transform yields: $p(x) = \frac{\sum_{A \supseteq \{x\}: |A|=s} m_b(A)}{\sum_{y \in \Theta} \sum_{A \supseteq \{y\}: |A|=s} m_b(A)} = \frac{\sum_{A \supseteq \{x\}: |A|=s} m_b(A)}{\sum_{A \subseteq \Theta: |A|=s} m_b(A)|A|} = \frac{\sum_{A \supseteq \{x\}: |A|=s} m_b(A)}{s \sum_{A \subseteq \Theta: |A|=s} m_b(A)}$, while applying the pignistic transform yields:

$$p(x) = \frac{\sum_{A \supseteq \{x\}: |A|=s} \frac{m_b(A)}{|A|}}{\sum_{y \in \Theta} \sum_{A \supseteq \{y\}: |A|=s} \frac{m_b(A)}{|A|}} = \frac{s \sum_{A \supseteq \{x\}: |A|=s} m_b(A)}{s \sum_{y \in \Theta} \sum_{A \supseteq \{y\}: |A|=s} m_b(A)}, \tag{3}$$

i.e., the same result. The following natural extension of the relative belief operator is therefore well defined.

Definition 1. *Given any b.f. $b : 2^\Theta \to [0,1]$ with b.b.a. m_b, we call* relative mass transformation *of level s the transform $\tilde{M}_s[b]$ which maps b to the probability (3). We denote by $\tilde{m}_s$ the output of the relative mass transform of level s.*

3 Approximation of Pignistic and Plausibility Transform

It is easy too see that both relative plausibility of singletons and pignistic probability are *convex combinations of all the (n) relative mass probabilities* $\{\tilde{m}_s, s = 1, ..., n\}$. Namely, let us we denote by:

$$k_{b,s} = \sum_{A \subseteq \Theta: |A|=s} m_b(A), \quad pl_b(x;s) = \sum_{A \supseteq \{x\}: |A|=s} m_b(A)$$

the total mass of focal elements of size s, and the contribution to the plausibility of x of size-s focal elements, respectively. Immediately: $\sum_y pl_b(y) = \sum_y \sum_{A \supseteq \{y\}} m_b(A) = \sum_{A \subseteq \Theta} m_b(A)|A| = \sum_{r=1}^n r(\sum_{A \subseteq \Theta, |A|=r} m_b(A)) = \sum_{r=1}^n r k_{b,r}$. Therefore, we obtain for the relative plausibility of singletons the following convex decomposition into relative mass probabilities $\tilde{m}_s$: $\tilde{pl}_b(x) =$

$$= \frac{pl_b(x)}{\sum_y pl_b(y)} = \frac{\sum_s pl_b(x;s)}{\sum_r r k_{b,r}} = \sum_s \frac{pl_b(x;s)}{\sum_r r k_{b,r}} = \sum_s \frac{pl_b(x;s)}{s k_{b,s}} \frac{s k_{b,s}}{\sum_r r k_{b,r}} = \sum_s \alpha_s \tilde{m}_s(x), \tag{4}$$

as $\tilde{m}_s(x) = \frac{pl_b(x;s)}{sk_{b,s}}$, with coefficients $\alpha_s = \frac{sk_{b,s}}{\sum_r rk_{b,r}} \propto sk_{b,s} = \sum_y pl_b(y;s)$ measuring for each level s the total plausibility contribution of the focal elements of size s.

In the case of the pignistic probability we get:

$$\begin{aligned} BetP[b](x) &= \sum_{A \supseteq \{x\}} \frac{m_b(A)}{|A|} = \sum_s \sum_{A \supseteq \{x\}, |A|=s} \frac{m_b(A)}{s} = \sum_s \frac{1}{s} \sum_{A \supseteq \{x\}, |A|=s} m_b(A) \\ &= \sum_s \frac{1}{s} pl_b(x;s) = \sum_s k_{b,s} \frac{pl_b(x;s)}{sk_{b,s}} = \sum_s k_{b,s} \tilde{m}_s(x), \end{aligned} \tag{5}$$

with coefficients $\beta_s = k_{b,s}$ measuring for each level s the mass contribution of the focal elements of size s.

Accordingly, the relative mass probabilities can be seen as basic components of both the pignistic and the plausibility transform, associated with the evidence carried by focal elements of a specific size.

As such transforms can be computed just by considering size-s focal elements, they can also be thought of as low-cost proxies for both relative plausibility and pignistic probability, since only the $\binom{n}{s}$ size-s focal elements (instead of the initial 2^n) have to be stored, while all the others can be dropped without further processing.

We can think of two natural criteria for such an approximation of $\tilde{pl}$, $BetP$ via the relative mass transforms.

- (C1) in the convex decompositions (4) and (5) associated with $\tilde{pl}$ and $BetP$, respectively, we retain the component s whose coefficient (α_s in the first case, β_s in the second) is the largest;
- (C2) we retain the component associated with the *minimal size* focal elements.

Clearly, the relative belief transformation coincides with the approximation produced by (C2) if $\sum_x m_b(x) \neq 0$. When the mass of singletons is nil, instead, the second criterion delivers a natural extension of the relative belief operator:

$$\tilde{b}^{ext}(x) \doteq \frac{\sum_{A \supseteq \{x\}: |A| = min} m_b(A)}{|A|_{min} \sum_{A \subseteq \Theta: |A| = min} m_b(A)}. \tag{6}$$

The two approximation criteria favor different aspects of the original belief function. (C1) focuses on the strength of the evidence carried by focal elements of equal size, by selecting those whose cardinality s is such that the total plausibility contribution of the focal elements of size s, $k_{b,s} = \sum_y pl_b(y;s)$, is the greatest. Note that, however, the optimal (C1) approximations of plausibility or pignistic transform are in principle quite distinct, as: $\hat{s}[\tilde{pl}] = \arg\max_s sk_{b,s}$, while $\hat{s}[BetP] = \arg\max_s k_{b,s}$. The best approximation for the pignistic probability will not necessarily be the best approximation of the relative plausibility of singletons. Criterion (C2) favors instead the *precision* of such pieces of evidence, measured by the size of the corresponding focal elements. Let us compare these two approaches in two simple scenarios.

4 Two Scenarios

While C1 is (at least superficially) a sensible, rational principle (the selected proxy must be the greatest contributor to the actual classical probability transformation), C2 seems harder to justify. Why should one retain only the smallest focal elements, regardless their mass?

The attractive feature of the relative belief of singletons, among C2 approximations, is its simplicity: the original mass is directly re-distributed onto the singletons. What about the "extended" operator (6)?

4.1 Scenario 1

Consider a scenario in which we want to approximate the plausibility/pignistic transform of a b.f. $b : 2^\Theta \to [0,1]$, with b.b.a. $m_b(A) = m_b(B) = \varepsilon$, $|A| = |B| = 2$, and $m_b(\Theta) = 1 - 2\varepsilon \gg m_b(A)$ (Figure 1-left). Its relative plausibility of singletons is given by:

$$\tilde{pl}_b(x) \propto m_b(A) + m_b(\Theta), \;\; \tilde{pl}_b(y) \propto m_b(A) + m_b(B) + m_b(\Theta),$$
$$\tilde{pl}_b(z) \propto m_b(B) + m_b(\Theta), \;\; \tilde{pl}_b(w) \propto m_b(\Theta) \; \forall w \neq x, y, z.$$

Its pignistic probability values are:

$$BetP(x) = \frac{m_b(A)}{2} + \frac{m_b(\Theta)}{n}, \;\; BetP(y) = \frac{m_b(A)+m_b(B)}{2} + \frac{m_b(\Theta)}{n},$$
$$BetP(z) = \frac{m_b(B)}{2} + \frac{m_b(\Theta)}{n}, \;\; BetP(w) = \frac{m_b(\Theta)}{n} \forall w \neq x, y, z.$$

Assuming $m_b(A) > m_b(B)$, both transformations have a profile as in Figure 1-right.

Now, according to (C1), the best approximation (among all relative mass transormations) of both $\tilde{pl}_b$ and $BetP[b]$ is given by selecting the focal elements of size n, i.e., Θ, as the greatest contributor to both the convex sums (4) and (5).

However, it is easy to see that this yields as an approximation the average probability $\tilde{m}_1(w) = 1/n \; \forall w \in \Theta$, which carries no information at all. In particular, the fact that the available evidence supports to a limited extent the singletons x, y and z is completed discarded, and no decision is possible.

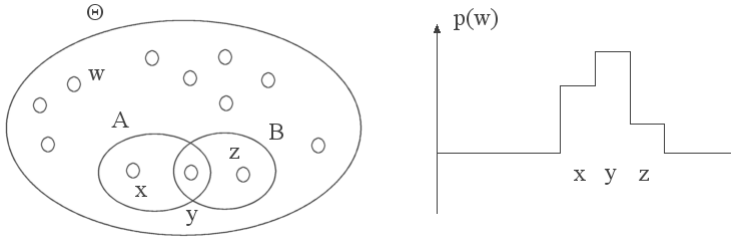


Fig. 1 Left: the original b.f. in the first scenario discussed in the text. Right: corresponding profile of both relative plausibility of singletons and pignistic probability.

If, on the other hand, we operate according to the criterion (C2), we end up selecting the size-2 focal elements A and B. The resulting approximation is

$$\tilde{m}_2(x) \propto m_b(A),\ \tilde{m}_2(y) \propto m_b(A) + m_b(B),\ \tilde{m}_2(z) \propto m_b(B),$$

$\tilde{m}_2(w) = 0\ \forall w \neq x, y, z$. This has the same profile as that of $\tilde{pl}_{b_.}$ or $BetP[b]$ (Figure 1-right): the decision made corresponds to that made based on $\tilde{pl}_b$ or $BetP[b]$.

In a decision-making sense, therefore, $\tilde{m}_2 = \tilde{b}^{ext}$ is the best approximation of both plausibility and pignistic transforms. We end up making the same decision, at a much lower (in general) computation cost.

4.2 Scenario 2

Consider however a second scenario, in which a b.f. has only two focal elements A and B, with $|A| > |B|$ and $m_b(A) \gg m_b(B)$ (Figure 2-left). Both relative plausibility and pignistic probability have the following values:

$$\tilde{pl}_b(w) = BetP(w) \propto m_b(A)\ w \in A, \quad \tilde{pl}_b(w) = BetP(w) \propto m_b(B)\ w \in B,$$

and correspond to the profile of Figure 2-right.

In this second case, (C1) and (C2) generate the uniform probability on elements of A (as $m_b(A) \gg m_b(B)$) and the uniform probability on elements of B (as $|B| < |A|$), respectively. Therefore, it is (C1) that yields the best approximation of both plausibility and pignistic transforms in a decision-making perspective.

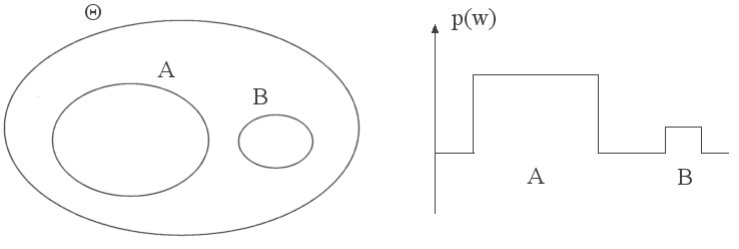


Fig. 2 Left: the b.f. of the second scenario. Right: corresponding profile of both relative plausibility of singletons and pignistic probability.

5 Conclusions

In this paper we tried and enrich our understanding of the family of epistemic transforms of belief functions. We showed that relative belief is only a member of a class of *relative mass* transformations which generalize it, are computable even when the mass of singletons is nil, and can be interpreted as low-cost proxies for both

plausibility and pignistic transforms. We discussed their applicability as approximate transformations in two significant scenarios.

References

1. Burger, T.: Defining new approximations of belief functions by means of Dempster's combination. In: Proc. of BELIEF 2010 (2010)
2. Cobb, B.R., Shenoy, P.P.: A Comparison of Methods for Transforming Belief Function Models to Probability Models. In: Nielsen, T.D., Zhang, N.L. (eds.) ECSQARU 2003. LNCS (LNAI), vol. 2711, pp. 255–266. Springer, Heidelberg (2003)
3. Cobb, B., Shenoy, P.: A comparison of Bayesian and belief function reasoning. Information Systems Frontiers 5(4), 345–358 (2003)
4. Cuzzolin, F.: Two new Bayesian approximations of belief functions based on convex geometry. IEEE Transactions on Systems, Man, and Cybernetics - Part B 37(4), 993–1008 (2007)
5. Cuzzolin, F.: A geometric approach to the theory of evidence. IEEE Transactions on Systems, Man, and Cybernetics - Part C 38(4), 522–534 (2008)
6. Cuzzolin, F.: Dual properties of the relative belief of singletons. In: Proc. of the Pacific Rim International Conference on AI, pp. 78–90 (2008)
7. Cuzzolin, F.: Geometry of relative plausibility and relative belief of singletons. Annals of Mathematics and Artificial Intelligence, 1–33 (2010)
8. Cuzzolin, F.: Semantics of the relative belief of singletons. In: Workshop on Uncertainty and Logic, Kanazawa, Japan (2008)
9. Daniel, M.: On transformations of belief functions to probabilities. Int. J. of Intelligent Systems 21(3), 261–282 (2006)
10. Dempster, A.P.: A generalization of Bayesian inference. Journal of the Royal Statistical Society, Series B 30, 205–247 (1968)
11. Dezert, J., Smarandache, F.: A new probabilistic transformation of belief mass assignment. In: Proc. of the 11th International Conference of Information Fusion, pp. 1–8 (2008)
12. Haenni, R.: Aggregating referee scores: an algebraic approach. In: 2nd International Workshop on Computational Social Choice, COMSOC 2008, pp. 277–288 (2008)
13. Lowrance, J., Garvey, T., Strat, T.: A framework for evidential-reasoning systems. In: Proc. of the National Conference on Artificial Intelligence, pp. 896–903 (1986)
14. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
15. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66(2), 191–234 (1994)
16. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. IJAR 38(2), 133–147 (2005)
17. Sudano, J.: Equivalence between belief theories and nave Bayesian fusion for systems with independent evidential data. In: Proc. of the 6th International Conference of Information Fusion, vol. 2, pp. 1239–1243 (2003)
18. Tessem, B.: Approximations for efficient computation in the theory of evidence. Artificial Intelligence 61(2), 315–329 (1993)
19. Voorbraak, F.: A computationally efficient approximation of Dempster-Shafer theory. International Journal on Man-Machine Studies 30, 525–536 (1989)

Choquet Integral as Maximum of Integrals with Respect to Belief Functions

Mikhail Timonin

Abstract. We study the problem of representing the Choquet integral w.r.t. an arbitrary capacity as maximum of integrals w.r.t. belief functions. We propose an algorithm and prove that for 2-additive capacities it allows to obtain a decomposition with the lowest number of elements.

1 Introduction

In applications of the Choquet integral to decision making problems it is often desirable to find the solution of the following optimization problem

$$\begin{aligned} & C(\nu, f) \quad \to \max_{f} \\ & f \in \mathscr{F}, \end{aligned} \tag{1}$$

where $C(\nu, f)$ is the Choquet integral with respect to some capacity ν, and $\mathscr{F}$ is the set of admissible "acts" or "alternatives". The solution can thus be interpreted as the "optimal" decision. According to the theorem of Lovász (1983) the Choquet integral is concave iff the capacity ν is 2-monotone. In the opposite case, the integral is not concave, can have several local maxima on $\mathscr{F}$ and is therefore hard to optimize. One potential solution to this problem is to find a partition of ν into a set of 2-monotone measures $\nu = \bigvee_i \beta^{T_i}$ such that $C(\nu, f) = \bigvee_i C(\beta^{T_i}, f)$. Maxima of the integrals $C(\beta^{T_i}, f)$ (which are concave) can then be easily found(Timonin 2011). It turns out that it is actually easier to partition the capacity into totally monotone measures (belief functions). Moreover, at least for 2-additive capacities doing so does not increase the number of elements in the partition.

Mikhail Timonin
National Nuclear Research University MEPhI, 31 Kashirskoe Shosse, Moscow 115409, Russian Federation
e-mail: mikhail.timonin@gmail.com

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 117–124.
springerlink.com

2 Basic Definitions

Definition 1. *Let N be a finite set and 2^N its power set. Capacity (non-additive measure, fuzzy measure) is a set function $\nu : 2^N \to \mathbb{R}_+$ such that:*

1. $\nu(\varnothing) = 0$;
2. $A \subseteq B \Rightarrow \nu(A) \leq \nu(B),\ \forall A, B \in 2^N$.

In this article, it is assumed that capacity is normalized, i.e. $\nu(N) = 1$.

Definition 2. *The Choquet integral of a function $\Phi : N \to \mathbb{R}_+$ with range $\{f_1, \ldots, f_n\}$ with respect to a capacity ν is defined as*

$$C(\nu, (f_1, \ldots, f_n)) = \sum_{i=1}^{n} (f_{(i)} - f_{(i-1)}) \nu(j | \Phi(j) \geqslant f_{(i)})$$

where $f_{(1)}, \ldots, f_{(n)}$ is a permutation of $f_1, \ldots, f_n$ such that $f_{(1)} \leq f_{(2)} \leq \cdots \leq f_{(n)}$, and $f_{(0)} = 0$.

Definition 3. *The Mobius transform m^ν of a capacity ν is given by:*

$$m^\nu(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} \nu(B).$$

Accordingly, the reverse of this operation is (zeta-transform):

$$\nu(A) = \sum_{B \subset A} m^\nu(B), \tag{2}$$

Definition 4. *The capacity ν is called k-monotone for some $k \geqslant 2$, if, for all families k of subsets $A_1, \ldots, A_k$, it holds that*

$$\nu\left(\bigcup_{i=1}^{k} A_i\right) \geqslant \sum_{\emptyset \neq I \subset 1, \ldots, k} (-1)^{|I|+1} \nu\left(\bigcap_{i \in I} A_i\right).$$

The capacity is called totally monotone if it is k-monotone for all $k \geqslant 2$. A 2-monotone capacity is called supermodular. If the inequality is reversed, the capacity is called 2-alternating (respectively infinitely alternating) or submodular.

An important property of totally monotone capacities (also called *belief functions*) is that their Mobius transform includes only non-negative coefficients. A belief function whose Mobius transform coefficients are distinct from zero only on elements of some maximal chain $\emptyset \subset Y_1 \subset \cdots \subset Y_n = N$ is called a *necessity measure*.

Definition 5 (Grabisch (1997)). *Capacity ν is called k-additive, if its Mobius coefficients $m(A) = 0$ for all $A \subset N, |A| > k$, and there exists $A \subset N, |A| = k$ such that $m(A) \neq 0$.*

Note, that for 2-additive capacities total monotonicity is equivalent to 2-monotonicity.

3 Disjunctive Representations of the Choquet Integral

For a capacity ν denote the set all pairs $\{i,j\}, i,j \in N$ such that $m^\nu(\{i,j\}) < 0$ as K. Construct an undirected graph from these pairs, connecting two vertices i and j with an edge if $m^\nu(\{i,j\}) < 0$. The following theorem holds:

Theorem 1. *The Choquet integral w.r.t. a 2-additive capacity ν can be represented as maximum of*

$$|\mathscr{B}| = (-1)^p \chi(-1) \tag{3}$$

Choquet integrals w.r.t. totally monotone measures $\beta^{T_i} \in \mathscr{B}, \nu = \bigvee_i \beta^{T_i}$, where p is the number of vertices in the graph, and $\chi(\cdot)$ is its chromatic polynomial. This decomposition is optimal, i.e. it is not possible to construct a representation with number of partitions less then $|\mathscr{B}|$. Moreover, it is not possible to obtain a lower number even when using 2-monotone measures.

The Algorithm

The decomposition of an arbitrary capacity into disjunction of $n!$ necessity measures has been proposed by Denneberg (2000). These measures are defined as:

$$m^{\mathscr{N}} = \begin{cases} m(A) = 0, & A \notin \mathscr{C} \\ m(K_i) = \nu(Y_i) - \nu(Y_{i-1}), & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathscr{C} = \{\emptyset \subset Y_1 \subset \cdots \subset Y_n = N\}$ is a maximal chain. Note that

$$\mathscr{N}(A) = \begin{cases} \nu(A), & A \in \mathscr{C} \\ \bigvee_{B \subsetneq A, B \in \mathscr{C}} \nu(B), & A \notin \mathscr{C}. \end{cases} \tag{5}$$

Also, $C(\nu, f) = C(\bigvee_i \mathscr{N}_i, f) = \bigvee_i C(\mathscr{N}_i, f)$ where $\mathscr{N}_i$ belong to the set of all necessity measures corresponding to maximal chains and are given by (5). Such representation fulfils our requirements but is not very convenient, since the number of elements in the partition is always $n!$. In order to find a coarser partition we elaborate on the bijection between three sets of power $n!$:

- Permutations $f_{(1)} \leq \cdots \leq f_{(n)}$;
- Maximal chains $\emptyset \subset Y_1 \subset \cdots \subset Y_n = N$;
- Necessity measures $\mathscr{N}_i : 2^N \to [0,1]$ such that $\nu = \bigvee_{i=1}^{n!} \mathscr{N}_i$.

The Choquet integral w.r.t capacity ν can be also expressed using its Mobius transform:

$$C(\nu, f) = \sum_{A \subseteq X} m(A) \bigwedge_{i \in A} f_i. \tag{6}$$

An important property of totally monotone measures is the non-negativity of their Mobius transform coefficients, $m(A) \geq 0, \forall A \subset X$. This property allowed to create an algorithm presented in Fig. 1. The principal idea of the algorithm is in locating the negative Mobius coefficients $m(A)$ in (6) and eliminating them. For a set of pairs

```
Input: Results = [ ]                                   // Results array (global)
Input: T = ∅                                           // Constraints f_i < f_j
begin Split(m,T,S)  // The function forms t.m. ''reduced'' capacities
    for f ∈ S do
        T = T ∩ {⋂_j (f_i < f_j), j ∈ A \ i}                          // f_i < f_j
        T = Closure(T)
        m^T = Reduced(m,T)
        S = MinNegSet(m^T)
        if S ≠ NULL then
            Split(m^T,T,S)                                 // Recursive call
        end
        else
            Results ← m^T,T                // Save the capacity and the set
        end
    end
end
begin MinNegSet(m)                  // Search for A: m(A) < 0 with the least
cardinality
    for i ∈ 2,...,n do
        for A ⊂ N,|A| == i do
            if m(A) < 0 then
                return S
            end
        end
    end
    return NULL
end
S = MinNegSet(m)
if S ≠ NULL then
    Split(m,T,S)
end
for T ∈ Results do
    BetaT(v,T)
end
```

Fig. 1 Capacity decomposition algorithm

$f_i < f_j, i, j = 1,\ldots,k$ the function **Closure**(T) finds all pairs which follow from transitivity of the relation "$<$".[1] The function **MinNegSet**(m) searches for the least-cardinality subset of N with a negative Mobius coefficient. Note the use of a stronger relation "$<$". This is done in order to simplify the definitions and theorems to follow by removing ambiguity in simplification of $\bigwedge_{i\in A} f_i$ terms of (6). Since the integral values for two adjacent partitions coincide at points where $f_i = f_j$, i.e.

[1] This is required since in numerical algorithms transitivity is not "automatically" enforced. For practical implementation we have used the Floyd-Warshall algorithm (see e.g. Korte and Vygen (2008)). Also, it is convenient to store constraints $f_i < f_j$ as pairs (i, j).

$$C(\nu,(f_{(1)},\ldots,f_{(i)},f_{(i+1)},\ldots,f_{(n)}))=C(\nu,(f_{(1)},\ldots,f_{(i+1)},f_{(i)},\ldots,f_{(n)}))$$

whenever $f_{(i)}=f_{(i+1)}$, nothing is lost due to such change, and partition elements T_i can be trivially extended afterwards so that their union is equal to $\mathbb{R}$.

Consider the following example. Let the integral be given as:

$$0.1f_1+0.1f_2+0.1f_3+0.6(f_1\wedge f_2)+0.6(f_1\wedge f_3)+0.6(f_2\wedge f_3)-1.1(f_1\wedge f_2\wedge f_3). \tag{7}$$

The expression contains an element with a negative coefficient: $m(\{1,2,3\})=-1.1<0$. To eliminate it the algorithm forms the sets $(f_1<f_2)\cap(f_1<f_3)$, $(f_2<f_1)\cap(f_2<f_3)$, $(f_3<f_1)\cap(f_3<f_2)$, which allows to transform the expression (7):

$$\begin{aligned}
&\nu^{T_1}:0.2f_1+0.1f_2+0.1f_3+0.6(f_2\wedge f_3) &&\sim f_1<f_2,f_1<f_3;\\
&\nu^{T_2}:0.1f_1+0.2f_2+0.1f_3+0.6(f_1\wedge f_3) &&\sim f_2<f_1,f_2<f_3;\\
&\nu^{T_3}:0.1f_1+0.1f_2+0.2f_3+0.6(f_1\wedge f_2) &&\sim f_3<f_1,f_3<f_2
\end{aligned}$$

Observe, that the obtained expressions can be viewed the Choquet integrals with respect to some new capacities ν^{T_i}. In Fig. 1 the generation of these capacities is performed by the function **Reduced**(m,T). In the following definitions and theorems for a set T formed as an intersection of some open hyperplanes of form $f_i<f_j, i,j\in N$, we denote as N_T^2 a set of ordered pairs (i,j) such that $(i,j)\in N\times N$, and $T=\bigcap_{(i,j)\in N_T^2}(f_i<f_j)$.

Definition 6 (Reduced(m,T)). *For a set $T=\bigcap_{(i,j)\in N_T^2}(f_i<f_j)\neq\emptyset$, the reduced capacity ν^T is given by:*

$$m^T(A)=\begin{cases}0, & \exists(i,j)\in N_T^2: i\in A, j\in A\\ \sum\limits_{B\subset\{j\mid i\in A, j\notin A,(i,j)\in N_T^2\}} m^\nu(A\cup B) & \nexists(i,j)\in N_T^2: i\in A, j\in A.\end{cases} \tag{8}$$

Theorem 2. *$C(\nu,f)=C(\nu^T,f)$ for all $f\in T$.*

Proof. Follows directly from (6).

The obtained capacities ν^{T_i} are totally monotone, but do not allow to obtain the required disjunctive decomposition. However, they are still very useful as Theorem 5 will show.

Definition 7 (BetaT(ν,T)). *For a set $T=\bigcap_{(i,j)\in N_T^2}(f_i<f_j)\neq\emptyset$ define the β^T-measure in a following way. Its coefficients are computed iteratively starting with the singletons of 2^N:*

$$\beta^T(A)=\begin{cases}\bigvee\limits_{B\subsetneq A}\beta^T(B), & \exists(i,j)\in N_T^2: i\in A, j\notin A\\ \nu(A), & \textit{otherwise}.\end{cases} \tag{9}$$

Theorem 3 (Properties of β^T-measure).

1. *For a set T corresponding to a permutation $f_{(1)} \leq \ldots\ldots f_{(n)}$, β^T-measure coincides with a necessity measure defined in (4).*
2. *Let $\mathscr{C}_1$ and $\mathscr{C}_2$ be two maximal chains corresponding to (single permutation) sets T_1 and T_2, and necessity measures $\mathscr{N}_1$ and $\mathscr{N}_2$. Then, the β^T-measure, corresponding to the set $T = T_1 \cup T_2$ is equal to:*

$$\beta^T(A) = \mathscr{N}_1(A) \vee \mathscr{N}_2(A), \quad \forall A \subset N. \tag{10}$$

3. *For some partition [2] $\bigcup_i T_i = \mathbb{R}^n$, where the sets T_i are unions of some single permutation subsets $\nu(A) = \bigvee_i \beta^{T_i}(A), \quad \forall A \subset N$.*
4. *Relation to the Choquet integral w.r.t. ν:*

$$\begin{aligned} C(\nu, f) &= C(\beta^T, f), \quad f \in T \\ C(\nu, f) &\geq C(\beta^T, f), \quad f \notin T. \end{aligned} \tag{11}$$

Proof. Properties 1 and 2 follow directly from the definition (9) (see also (5)), while 3 and 4 can be easily derived therefrom. □

Theorem 4. *The Mobius transform coefficients of β^T, where $T = \bigcap_{(i,j)\in N_T^2}(f_i < f_j)$ are given by:*

$$m^\beta(A) = \begin{cases} 0, & \exists (i,j \in N_T^2) : i \in A, j \notin A \\ \sum\limits_{B \subset \{j | i \in A, j \in A, (i,j) \in N_T^2\}} m^\nu(A \setminus B), & \nexists (i,j) \in N_T^2 : i \in A, j \notin A. \end{cases} \tag{12}$$

Proof. Sketch: show that for m^β thus defined zeta-transform (2) allows to obtain (9).For full proof refer to (Timonin 2011).

Theorem 5. *If for some set $T = \bigcap_{(i,j)\in N_T^2}(f_i < f_j)$ the reduced capacity ν^T is totally monotone, then β^T is also totally monotone.*

Proof. Show that (12) is a permutation of (8).

k-Additive Case

A formal characterization of resulting decomposition in the k-additive case is left for the future research. Here we would only point out some differences with the results above. For $k > 2$ the algorithm in Fig. 1 can lead to a suboptimal result. This happens when the integral retains concavity within non-convex unions of permutation regions. This causes the algorithm to perform undesirable "over-splitting", since it is only capable of producing convex sets T_i. Note, that over-splitting does not affect the correctness of the results, i.e. a desired disjunctive decomposition is still obtained, albeit with a larger number of partition elements. More details and examples can be found in (Timonin 2011).

[2] For equality to hold T_i must be accordingly extended by switching back to "$\leq$" from "$<$".

4 Proof of Theorem 1

A full proof of Theorem 1 can be found in (Timonin 2011). The proof is built upon three fundamental lemmas which we present here. Capacities $\mathcal{N}_i$ are those from (4) and form a disjunctive decomposition of the capacity ν.

Definition 8. *We will call capacities $\mathcal{N}_1$ and $\mathcal{N}_2$ non-$\vee_{2m}$-joinable, if the capacity $\mathcal{N}_1 \vee \mathcal{N}_2$ is not 2-monotone, and there do not exist capacities $\mathcal{N}_i, i = 1,\dots,k$ such that $(\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\dots,k} \mathcal{N}_i$ is 2-monotone.*

Since we analyze 2-additive capacities, their Mobius transform can have negative coefficients only for sets of the form $\{a,b\}$, i.e. for sets of power 2. Denote the set of all unordered pairs $\{a,b\}$ having a negative Mobius coefficient as K. In other words,

$$K = \{\{a,b\} | m^{\nu}(\{a,b\}) < 0\}. \tag{13}$$

To simplify the notation we will write ab instead of $\{a,b\}$ and Ya instead of $Y \cup \{a\}$. Denote the subset of N comprised of elements that are included in at least one pair from K as N_K. In the following proofs we will denote partial orders over N_K induced by various combinations of orderings (with relation $<$) on each of the pairs from K as P_i. Note, that not every combination induce a partial order, but only those which correspond to acyclic orientations of the corresponding graph(see Theorem 1). The total number of such orientations is due to Stanley (1973). The elements of N not included into at least one pair in K do not influence the 2-monotonicity of the capacity, and therefore will be excluded from the analysis. We prove that it is possible to find at least $(-1)^p \chi(-1)$ necessity measures which are pairwise non-$\vee_{2m}$-joinable. In particular, it will be shown that it is possible to pick a necessity measure corresponding to each partial order so that these measures are pairwise non-$\vee_{2m}$-joinable. We assume here that the graph made of the pairs from K does not contain disconnected parts. Otherwise, the proof below can be applied to each part separately. In this case the number of resulting measures would be equal to product of numbers generated by each part (recall also Theorem 1 and the fact that the chromatic polynomial of a disconnected graph equals to the product of chromatic polynomials of its connected parts (e.g. (Read 1968))).

Lemma 1. *Capacities $\mathcal{N}_1$ and $\mathcal{N}_2$ are non-$\vee_{2m}$-joinable, if there exist $Ya \in \mathcal{C}_1, Yb \in \mathcal{C}_2$, where $\mathcal{C}_1, \mathcal{C}_2$ - are maximal chains such that $\mathcal{N}_1 \sim \mathcal{C}_1, \mathcal{N}_2 \sim \mathcal{C}_2$, $\{a,b\} \in K$.*

Proof. The 2-monotonicity condition is

$$\nu(A \cup B) - \nu(A) - \nu(B) + \nu(A \cap B) \geq 0, \quad \forall A, B \subset N. \tag{14}$$

We will write it down for the element Yab:

$$\nu(Yab) - \nu(Ya) - \nu(Yb) + \nu(Y). \tag{15}$$

Since ν is 2-additive for all $A \subset N$ holds(Grabisch 1997):

$$\nu(A) = \sum_{i,j \in A} \nu(ij) - (|A| - 2) \sum_{i \in A} \nu(i) \tag{16}$$

Thus,

$$v(Yab) - v(Ya) - v(Yb) + v(Y) = v(ab) - v(a) - v(b) = m(ab) < 0. \quad (17)$$

And according to the definition of the $\mathcal{N}$ measures (see (5)) and the conditions of the lemma $(\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\ldots,k} \mathcal{N}_i$, i.e. $(\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\ldots,k} \mathcal{N}_i(Yab) - (\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\ldots,k} \mathcal{N}_i(Ya) - (\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\ldots,k} \mathcal{N}_i(Yb) + (\mathcal{N}_1 \vee \mathcal{N}_2) \bigvee_{i=1,\ldots,k} \mathcal{N}_i(Y) < 0$

Note: since $\{a,b\} \in K$, conditions of the lemma imply that $\mathscr{C}_1$ and $\mathscr{C}_2$ correspond to different partial orders.

Lemma 2. *$\mathcal{N}_1 \vee \mathcal{N}_2$ is not 2-monotone if there exist $Ya \in \mathscr{C}_1, Yb \in \mathscr{C}_2, Yab \notin \mathscr{C}_1, Yab \notin \mathscr{C}_2, \{a,b\} \notin K$, where $\mathcal{N}_1 \sim \mathscr{C}_1, \mathcal{N}_2 \sim \mathscr{C}_2, \mathscr{C}_1 \sim P_1, \mathscr{C}_2 \sim P_2$, and P_1 and P_2 are different partial orders.*

Proof. Similar to the proof of Lemma 1.

Lemma 3. *$\mathcal{N}_1$ and $\mathcal{N}_2$ are non-$\vee_{2m}$-joinable if there exist $Ya \in \mathscr{C}_1, Yb \in \mathscr{C}_2, Yab \notin \mathscr{C}_1, Yab \notin \mathscr{C}_2, \{a,b\} \notin K$, where $\mathcal{N}_1 \sim \mathscr{C}_1, \mathcal{N}_2 \sim \mathscr{C}_2, \mathscr{C}_1 \sim P_1, \mathscr{C}_2 \sim P_2$, and P_1 and P_2 are different partial orders.*

Proof. Sketch: Show that either Lemma 1 or 2 can be applied. If Lemma 2 applies then it is possible to find $\mathcal{N}_3 : Yab \in \mathscr{C}_3$ such that $\mathcal{N}_1 \vee \mathcal{N}_2 \vee \mathcal{N}_3$ is 2-monotone. However, conditions of lemmata 1,2 would then be met for at least one of pair of sets (Yab, Yax_1) or (Yab, Ybz_1), where Yax_1 and Ybz_1 are the next elements in chains, corresponding to capacities $\mathcal{N}_1$ and $\mathcal{N}_2$. Continuing to add more necessity measures we will eventually obtain a distributive lattice generated by joins of all elements from chains $Y \subset Ya \subset \ldots \subset N_K$ (sub-chain of $\mathscr{C}_1$) and $Y \subset Yb \subset \ldots \subset N_K$(sub-chain of $\mathscr{C}_2$) and Lemma 1 applies.

References

Denneberg, D.: Totally monotone core and products of monotone measures. International Journal of Approximate Reasoning 24(2-3), 273–281 (2000) ISSN 0888-613X

Grabisch, M.: k-order additive discrete fuzzy measures and their representation. Fuzzy Sets and Systems 92(2), 167–189 (1997) ISSN 0165-0114

Korte, B.H., Vygen, J.: Combinatorial optimization: theory and algorithms. Springer (2008) ISBN 3540718435

Lovász, L.: Submodular functions and convexity. Mathematical Programming: the State of the Art, 235–257 (1983)

Read, R.C.: An introduction to chromatic polynomials*. Journal of Combinatorial Theory 4(1), 52–71 (1968)

Stanley, R.P.: Acyclic orientations of graphs. Discrete Mathematics 5(2), 171–178 (1973)

Timonin, M.: Maximization of the Choquet integral over a convex set and its application to resource allocation problems. Annals of Operations Research (2011) (under review)

Consonant Approximations in the Belief Space

Fabio Cuzzolin

Abstract. In this paper we solve the problem of approximating a belief measure with a necessity measure or "consonant belief function" by minimizing appropriate distances from the consonant complex in the space of all belief functions. Partial approximations are first sought in each simplicial component of the consonant complex, while global solutions are obtained from the set of partial ones. The L_1, L_2 and L_∞ consonant approximations in the belief space are here computed, discussed and interpreted as generalizations of the maximal outer consonant approximation. Results are also compared to other classical approximations in a ternary example.

1 Introduction

The theory of evidence [14] is a popular approach to uncertainty description in which probabilities are replaced by *belief functions* (b.f.s), functions $b : 2^\Theta \to [0,1]$ on the power set $2^\Theta = \{A \subseteq \Theta\}$ of the sample space Θ of the form $b(A) = \sum_{B \subseteq A} m_b(B)$, where $m_b : 2^\Theta \to [0,1]$ is a non-negative, normalized set function called "basic probability assignment" (b.p.a.) or "mass assignment", and $pl_b(A) \doteq 1 - b(A^c)$ is the *plausibility function* (pl.f.) associated with b. Belief functions assign values $b(A)$ between 0 and 1 to subsets of the sample space Θ rather than to single elements. Possibility theory [8], instead, studies *possibility measures*, i.e., functions $Pos : 2^\Theta \to [0,1]$ on the power set such that $Pos(\bigcup_i A_i) = \sup_i Pos(A_i)$ for any family of subsets $\{A_i | A_i \in 2^\Theta, i \in I\}$, where I is an arbitrary set index. Given a possibility measure Pos, the dual *necessity* measure is defined as $Nec(A) = 1 - Pos(A^c)$.

Interestingly, necessity measures have as counterparts in the theory of evidence *consonant* belief functions (co.b.f.s), i.e., b.f.s whose non-zero mass subsets $m_b(A) \neq 0$ or "focal elements" (f.e.s) are nested [14] and form a chain (totally ordered collection) of subsets $A_1 \subset \cdots \subset A_m$, $A_i \subseteq \Theta$, in which case $Pos(\{x\}) = pl_b(\{x\})$.

Fabio Cuzzolin
Oxford Brookes University, Oxford, UK
e-mail: fabio.cuzzolin@brookes.ac.uk

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 125–133.
springerlink.com

Approximating a b.f. with a necessity measure amounts therefore to mapping it to a consonant b.f. [9, 11, 2]. As possibilities are completely determined by their values on the singletons ($Pos(\{x\}), x \in \Theta$), they are less computationally expensive than b.f.s, making the approximation process interesting for many applications. Applications to the approximate computation of belief functions on Cartesian products and combinations by Dempster's rule have indeed been proposed in [9], while arguments for inferring consonant belief functions from data available in the form of likelihoods have been brought forward by Shafer [14].

A Geometric Approach to Consonant Approximation. Dubois and Prade, in particular, have proposed the notion of "outer consonant approximations" [9] of belief functions. Their work has been later extended by Baroni [2] to capacities, while, in [6], the author has provided a description of the geometry of the set of outer consonant approximations. A different "isopignistic" approximation has been proposed as the unique consonant b.f. whose pignistic probability $BetP(x) = \sum_{A \supseteq \{x\}} m_b(A)$ is identical to that of the original b.f. b [10, 17, 1]. In more recent times, the opportunity of seeking probability or consonant approximations / transformations of belief functions by minimizing appropriate distance functions has been explored [3, 4]. Any dissimilarity measure could be in principle employed to define conditional b.f.s, or to approximate b.f.s by necessity or probability measures [12, 15, 13]. We focus here on L_p norms, which have been successfully applied in the past [5].

Contribution. The goal of this paper is to conduct an analytical study of all the consonant approximations induced by minimizing L_1, L_2 or L_∞ distances between the original belief function and the consonant region, in the vector space they form or *belief space* $\mathscr{B}$, as a stepping stone of a more extensive theoretical study of the nature of consonant approximations induced by geometric distance minimization.

As it turns out, all "partial" L_p consonant approximations in $\mathscr{B}$ (having a desired maximal chain of subsets $A_1 \subsetneq \cdots \subsetneq A_n$, $n = |\Theta|$ as focal elements) amount to picking different representatives from the n lists of belief values: $\mathscr{L}^i = \{b(A), A \supseteq A_i, A \not\supset A_{i+1}\}\ \forall i = 1,...,n$, as they have mass $m'(A_i) = f(\mathscr{L}^i) - f(\mathscr{L}^{i-1})$, where f is a simple function of the belief values in the list, such as max, average, or median. Classical maximal outer and "contour-based" approximations can also be expressed in the same way. As they would all reduce to the maximal outer approximation $m'(A_i) = \min(\mathscr{L}^i) - \min(\mathscr{L}^{i-1}) = b(A_i) - b(A_{i-1})$ if the power set was totally (rather than partially) ordered, all these consonant approximations can be considered as generalization of the latter. Sufficient conditions on their admissibility can be given in terms of the (partial) plausibility values of the singletons. Due to lack of space, the reader is referred to [7] for the proofs of all main results.

2 Geometry of Consonant Belief Functions

Given a domain Θ, each belief function $b : 2^\Theta \to [0,1]$ is completely specified by its $N-2$ belief values $\{b(A), \emptyset \subsetneq A \subsetneq \Theta\}$, $N \doteq 2^n$ ($n \doteq |\Theta|$), (as $b(\emptyset) = 0$, $b(\Theta) = 1$ for all b.f.s), and can therefore be represented as a vector $\mathbf{b} = [b(A), \emptyset \subsetneq A \subsetneq \Theta]'$ of $\mathbb{R}^{N-2}$. We can prove that [4] the set of points of $\mathbb{R}^{N-2}$ which correspond to a b.f. or *belief space* $\mathscr{B}$ is the convex closure $\mathscr{B} = Cl(\mathbf{b}_A, \emptyset \subsetneq A \subseteq \Theta)$, where b_A is

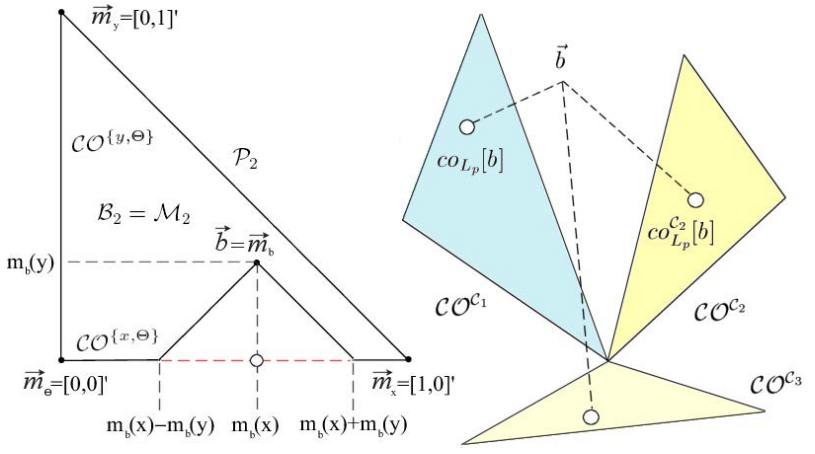


Fig. 1 Left: the belief space $\mathscr{B}_2$ for a binary frame is a triangle in $\mathbb{R}^2$ whose vertices are the vectors $\mathbf{b}_x = [1,0]'$, $\mathbf{b}_y = [0,1]'$, $\mathbf{b}_\Theta = [0,0]'$ associated with the categorical belief functions focused on $\{x\}, \{y\}$ and Θ, respectively. Consonant b.f.s live in the union of the segments $\mathscr{CO}^{\{x,\Theta\}}$ and $\mathscr{CO}^{\{y,\Theta\}}$. The unique $L_1 = L_2$ consonant approximation (circle) and the set of L_∞ consonant approximations (dashed segment) on $\mathscr{CO}^{\{x,\Theta\}}$ are shown. Right: To minimize the distance of a point from a simplicial complex, we need to find all the partial solutions on all the simplices in the complex (empty circles), and compare them to select a global one (black circle).

the *categorical* [18] belief function assigning all the mass to a single subset $A \subseteq \Theta$ and Cl denotes the convex closure operator: $Cl(\mathbf{b}_1, ..., \mathbf{b}_k) = \{\mathbf{b} \in \mathscr{B} : \mathbf{b} = \alpha_1 \mathbf{b}_1 + \cdots + \alpha_k \mathbf{b}_k, \sum_i \alpha_i = 1,\ \alpha_i \geq 0\ \forall i\}$. The belief space $\mathscr{B}$ is a simplex[1] [4], and each vector $\mathbf{b} \in \mathscr{B}$ representing a belief function b can be written as a convex sum as: $\mathbf{b} = \sum_{\emptyset \subsetneq A \subseteq \Theta} m_b(A) \mathbf{b}_A$. The set $\mathscr{P}$ of all "Bayesian" b.f.s (assigning non-zero masses to singletons only: $m_b(A) = 0$ if $|A| > 1$) is the simplex[2] $\mathscr{P} = Cl(\mathbf{b}_x, x \in \Theta)$.

In the case of a domain $\Theta_2 = \{x, y\}$ of cardinality 2, each b.f. b is completely determined by its mass values $m_b(x)$, $m_b(y)$, as $m_b(\Theta) = 1 - m_b(x) - m_b(y)$ and $m_b(\emptyset) = 0$, and is represented by a vector $\mathbf{b} = [b(x) = m_b(x), b(y) = m_b(y)]' \in \mathbb{R}^2$.

Since $m_b(x) \geq 0$, $m_b(y) \geq 0$, and $m_b(x) + m_b(y) \leq 1$, the set $\mathscr{B}_2$ of all the possible b.f.s on Θ_2 can be depicted as the triangle in the Cartesian plane of Figure 1-left. The region $\mathscr{P}_2$ of all Bayesian b.f.s on Θ_2 is the diagonal line segment $Cl(\mathbf{b}_x, \mathbf{b}_y)$. On $\Theta_2 = \{x, y\}$ consonant belief functions can have as chain of focal elements either $\{\{x\} \subset \Theta_2\}$ or $\{\{y\} \subset \Theta_2\}$. Therefore, they live in the union of two segments (see Figure 1-left): $\mathscr{CO}_2 = \mathscr{CO}^{\{x,\Theta\}} \cup \mathscr{CO}^{\{y,\Theta\}} = Cl(\mathbf{b}_x, \mathbf{b}_\Theta) \cup Cl(\mathbf{b}_y, \mathbf{b}_\Theta)$.

Approximation in the Consonant Complex. In the general case the region $\mathscr{CO}$ of consonant belief functions in the belief space is a *simplicial complex*

[1] The convex closure $Cl(x_1, ..., x_{n+1})$ of $n+1$ (affinely independent) points $x_1, ..., x_{n+1}$ of $\mathbb{R}^n$ [4].

[2] We will use here the notation x to denote both an element $x \in \Theta$ of the frame and the set $\{x\}$.

[6], i.e., the union of a collection of (maximal) simplices, each associated with a maximal chain $\mathscr{C} = \{A_1 \subset \cdots \subset A_n\}$, $|A_i| = i$, $A_n = \Theta$ of subsets of Θ: $\mathscr{CO} = \bigcup_{\mathscr{C}} \mathscr{C}\mathscr{O}^{\mathscr{C}} = \bigcup_{\mathscr{C}=A_1\subset\cdots\subset A_n} Cl(\mathbf{b}_{A_1}, \cdots, \mathbf{b}_{A_n})$. Given a belief function b, we call *consonant approximation of b induced by a distance function d* in $\mathscr{B}$ the b.f.(s) $\mathscr{C}\mathscr{O}_d[b]$ which minimize(s) the distance $d(\mathbf{b}, \mathscr{CO})$ between b and the consonant simplicial complex in $\mathscr{B}$. We use the notation $co_d[b]$ when the solution is unique, or to denote the barycenter of the set of solutions $\mathscr{C}\mathscr{O}_d[b]$. As the consonant complex $\mathscr{CO}$ is a *collection* of simplices which generate distinct linear spaces, solving the approximation problem involves finding first a number of partial solutions: $co^{\mathscr{C}}_{L_p}[b] = \arg\min_{\mathbf{co}\in\mathscr{C}\mathscr{O}^{\mathscr{C}}} \|\mathbf{b} - \mathbf{co}\|_{L_p}$ (see Figure 1-right), one for each maximal chain $\mathscr{C}$ of subsets of Θ. Then, the distance of b from all partial solutions has to be assessed in order to select a global optimum. L_p norms have been recently employed in the probability transformation problem [3] and for conditioning [5]. For vectors $\mathbf{b}$, $\mathbf{b}' \in \mathscr{B}$ representing two belief functions b, b', such norms read as: $\|\mathbf{b} - \mathbf{b}'\|_{L_1} \doteq \sum_{\emptyset\subsetneq B\subsetneq\Theta} |b(B) - b'(B)|$; $\|\mathbf{b} - \mathbf{b}'\|_{L_2} \doteq \sqrt{\sum_{\emptyset\subsetneq B\subsetneq\Theta}(b(B) - b'(B))^2}$, and $\|\mathbf{b} - \mathbf{b}'\|_{L_\infty} \doteq \max_{\emptyset\subsetneq B\subsetneq\Theta} |b(B) - b'(B)|$. Clearly, however, a number of other norms can be picked [12]: this paper is as just a first step of a long line of research.

3 Consonant Approximation in the Belief Space

3.1 Calculation of L_p Approximations in the Belief Space

L_1 approximation. The set of partial L_1 consonant approximations in $\mathscr{B}$ can be expressed in terms of a list of belief values very much related to the *maximal (partial) outer consonant approximation* [9] with maximal chain $\mathscr{C}$:

$$m_{co^{\mathscr{C}}_{max}[b]}(A_i) = \sum_{B\subseteq A_i, B\not\subset A_{i-1}} m_b(B) = b(A_i) - b(A_{i-1}). \tag{1}$$

Theorem 1. *Given a b.f. $b : 2^\Theta \to [0,1]$, its partial L_1 consonant approximations $\mathscr{C}\mathscr{O}^{\mathscr{C}}_{L_1}[b]$ in $\mathscr{B}$ with maximal chain of focal elements $\mathscr{C} = \{A_1 \subset \cdots \subset A_n, |A_i| = i\}$ are the co.b.f.s co whose mass vectors $[m_{co}(A_1), ..., m_{co}(A_n)]'$ live in:*

$$Cl\Big(\big[b^1, b^2 - b^1, \cdots, b^i - b^{i-1}, \cdots, 1 - b^{n-1}\big]' \;\Big|\; b^i \in \big\{\gamma^i_{int1}, \gamma^i_{int2}\big\}\ \forall i\Big), \tag{2}$$

where $\gamma^i_{int1}, \gamma^i_{int2}$ are the innermost (median) elements of the list of belief values:

$$\mathscr{L}_i = \{b(A), A \supseteq A_i, A \not\supset A_{i+1}\}. \tag{3}$$

As $b^{n-1} = \gamma^{n-1}_{int1} = \gamma^{n-1}_{int2} = b(A_{n-1})$, (2) is a polytope of 2^{n-2} vertices. Note that we present our results in terms of mass assignments, as they are simpler and easier to interpret. Due to the nature of partially ordered set of 2^Θ, the innermost values of the above lists (3) cannot be analytically identified in full generality (even though

they can be easily computed numerically), but can be derived in some simple (e.g. ternary) cases. As for the *global* L_1 approximation(s):

Theorem 2. *Given a belief function $b : 2^\Theta \to [0,1]$, its global L_1 consonant approximations $\mathscr{CO}_{L_1}[b]$ in $\mathscr{B}$ live in the collection of partial such approximations associated with the maximal chain(s) $A_1 \subset ... \subset A_n$ which maximize the cumulative lower halves of the lists of belief values $\mathscr{L}_i$ (3): $\arg\max_{\mathscr{C}} \sum_i \sum_{b(A)\in\mathscr{L}_i, b(A)\leq \gamma^i_{int_1}} b(A)$.*

L_2 **approximation.** To find the partial consonant approximation(s) at minimal L_2 distance from b in $\mathscr{B}$ we need to impose the orthogonality of the difference vector $\mathbf{b}-\mathbf{co}$ with respect to any given simplicial component $\mathscr{CO}^{\mathscr{C}}$ of the complex $\mathscr{CO}$: $\langle \mathbf{b}-\mathbf{co}, \mathbf{b}_{A_j}-\mathbf{b}_\Theta\rangle = \langle \mathbf{b}-\mathbf{co}, \mathbf{b}_{A_j}\rangle = 0 \ \forall A_j \in \mathscr{C}, 1 \leq j \leq n-1$, as $\mathbf{b}_\Theta = \mathbf{0}$ is the origin of the Cartesian space in $\mathscr{B}$, and $\mathbf{b}_{A_j}-\mathbf{b}_\Theta$ for $j = 1,...,n-1$ are the generators of $\mathscr{CO}^{\mathscr{C}}$ (compare the binary case of Figure 1-left). The L_2 partial approximation of b is unique, and a function of the list of belief values (3) as well.

Theorem 3. *Given a b.f. $b : 2^\Theta \to [0,1]$, its partial L_2 consonant approximation $co^{\mathscr{C}}_{L_2}[b]$ in $\mathscr{B}$ with maximal chain $\mathscr{C} = \{A_1 \subset \cdots \subset A_n\}$ is unique, and has b.p.a.:*

$$m_{co^{\mathscr{C}}_{L_2}[b]}(A_i) = ave(\mathscr{L}_i) - ave(\mathscr{L}_{i-1}) \qquad \forall i = 1,...,n, \tag{4}$$

where $ave(\mathscr{L}_i) = \frac{1}{2^{|A^c_{i+1}|}} \sum_{A\supseteq A_i, A\not\supset A_{i+1}} b(A)$ is the average of the list $\mathscr{L}_i$ (3), $\mathscr{L}_0 \doteq \{0\}$.

The problem of finding the global L_2 approximation is not trivial, and has not been addressed yet. L_∞ **approximations** also form a polytope, with 2^{n-1} vertices.

Theorem 4. *Given a b.f. $b : 2^\Theta \to [0,1]$, its partial L_∞ consonant approximations $\mathscr{CO}^{\mathscr{C}}_{L_\infty}[b]$ in $\mathscr{B}$ with maximal chain of focal elements $\mathscr{C} = \{A_1 \subset \cdots \subset A_n, |A_i| = i\}$ are the co.b.f.s co whose mass vectors $[m_{co}(A_1),...,m_{co}(A_n)]'$ live in:*

$$Cl\Big(\big[b^1, \cdots, b^i - b^{i-1}, \cdots, 1-b^{n-1}\big]' \,\Big|\, b^i = \tfrac{b(A_i)+b(\{x_{i+1}\}^c)}{2} + \{-b(A_1^c), b(A_1^c)\} \ \forall i\Big). \tag{5}$$

The barycenter $co^{\mathscr{C}}_{L_\infty}[b]$ of (5) has b.p.a.: $m_{co^{\mathscr{C}}_{L_\infty}[b]}(A_1) = \frac{b(A_1)+b(\{x_2\}^c)}{2}$, $m_{co^{\mathscr{C}}_{L_\infty}[b]}(A_i) = \frac{b(A_i)-b(A_{i-1})}{2} + \frac{pl_b(x_i)-pl_b(x_{i+1})}{2}$, $2 \leq i \leq n-1$, while $m_{co^{\mathscr{C}}_{L_\infty}[b]}(A_n) = 1 - b(A_{n-1})$.

Now, let us call *contour-based* consonant approximation of a b.f. b with maximal chain of focal elements $\mathscr{C} = \{A_1 \subset \cdots \subset A_n\}$ the co.b.f. with mass assignment: $m_{co_{con}[b]}(A_1) = 1 - pl_b(x_2)$, $m_{co_{con}[b]}(A_i) = pl_b(x_i) - pl_b(x_{i+1})$ for $i = 2,...,n-1$, and $m_{co_{con}[b]}(A_n) = pl_b(x_n)$, where $\{x_i\} \doteq A_i \setminus A_{i-1}$ for all $i = 1,...,n$. Such an approximation uses the (unnormalized) contour function of an arbitrary b.f. b to generate a consonant b.f., *as if* it was a possibility distribution. Then, by (1) and the above definition, it is clear that the barycenter of the partial L_∞ approximations in $\mathscr{B}$ is the *average of the maximal outer consonant approximation and what we called "contour-based" consonant approximation*.

As the distance from b of the partial solutions (5) is $b(A_1^c)$ (see the proof of Theorem 4, [7]), the global L_∞ consonant approximations of b in $\mathscr{B}$ are

associated with the chains of focal elements: $\arg\min_{\mathscr{C}} b(A_1^c) = \arg\min_{\mathscr{C}} (1 - pl_b(A_1)) = \arg\max_{\mathscr{C}} pl_b(A_1)$, which are *nested around the maximal plausibility singleton*.

3.2 Interpretation as Generalized Maximal Outer Approximations

From Theorems 1, 3 and 4, the b.p.a.s of *all L_p partial approximations in the belief space are differences of simple functions of belief values taken from the list* (3):

$$\begin{aligned}
&m_{co^{\mathscr{C}}_{max}[b]}(A_i) = \min(\mathscr{L}_i) - \min(\mathscr{L}_{i-1}); \quad m_{co^{\mathscr{C}}_{con}[b]}(A_i) = \max(\mathscr{L}_i) - \max(\mathscr{L}_{i-1});\\
&m_{co^{\mathscr{C}}_{L_1}[b]}(A_i) = (int_1(\mathscr{L}_i) + int_2(\mathscr{L}_i))/2 - (int_1(\mathscr{L}_{i-1}) + int_2(\mathscr{L}_{i-1}))/2;\\
&m_{co^{\mathscr{C}}_{L_2}[b]}(A_i) = ave(\mathscr{L}_i) - ave(\mathscr{L}_{i-1});\\
&m_{co^{\mathscr{C}}_{L_\infty}[b]}(A_i) = (\max(\mathscr{L}_i) + \min(\mathscr{L}_i))/2 - (\max(\mathscr{L}_{i-1}) + \min(\mathscr{L}_{i-1}))/2.
\end{aligned} \tag{6}$$

The maximal outer approximation $co^{\mathscr{C}}_{max}[b]$ is obtained by picking as representative $\min(\mathscr{L}_i)$, $co^{\mathscr{C}}_{con}[b]$ amounts to picking $\max(\mathscr{L}_i)$, the barycenter of the L_1 approximations to choosing the average innermost (median) value, the barycenter of the L_∞ approximations to the average outermost value, L_2 to picking the overall average value of the list. Each vertex of the L_1 solution set (2) amounts to selecting, for each component, either one of the innermost values; each vertex of the L_∞ polytope (5), either one of the outermost values.

Belief functions are defined on a partially ordered set, the power set $2^\Theta = \{A \subseteq \Theta\}$, of which a maximal chain is a maximal totally ordered subset. Therefore, given two elements of the chain $A_i \subset A_{i+1}$, there are a number of "intermediate" focal elements A which contain the latter but not the former. If 2^Θ were to be a totally ordered set, the list $\mathscr{L}_i$ would contain a single element $b(A_i)$ and all the L_p approximations (6) would reduce to the function $co^{\mathscr{C}}_{max}[b]$ (1): they can all be seen as different *generalizations of the maximal outer consonant approximation*. It should be noted, however, that such approximations are not, in general, outer approximations in the sense of the former (as it is confirmed by the following example).

3.3 Graphical Comparison in a Ternary Example

It can be useful to compare the different approximations in the toy case of a ternary frame, $\Theta = \{x, y, z\}$. Let the desired maximal chain be $\mathscr{C} = \{\{x\} \subset \{x, y\} \subset \Theta\}$. Figure 2 illustrates the different partial L_p consonant approximations in $\mathscr{B}$ in the simplex of consonant b.f.s with chain $\mathscr{C}$, for a b.f. b with masses: $m_b(x) = 0.2$, $m_b(y) = 0.3$, $m_b(x,z) = 0.5$. The analogous L_p approximations *in the mass space* $\mathscr{M}$ [7] (in which b.f.s are represented by their mass vectors) for the same b.f. are depicted for comparison. Its isopignistic approximation $m_{co_{iso}[b]}(A_i) = i \cdot \big(BetP[b](x_i) - BetP[b](x_{i+1})\big)$, $\{x_i\} \doteq A_i \setminus A_{i-1} \forall i$ [10] is also plotted. For the comparison to be homogeneous, we plot both sets of approximations (in $\mathscr{B}$ and

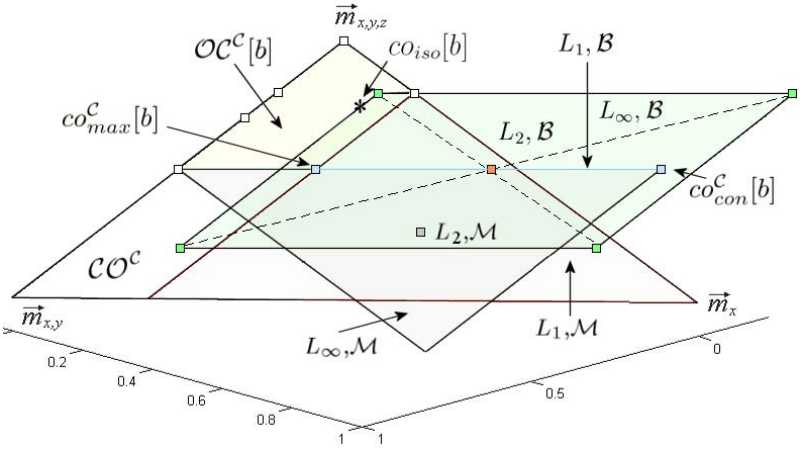


Fig. 2 Comparison between L_p partial consonant approximations in the mass $\mathscr{M}$ and belief $\mathscr{B}$ spaces for the b.f. of the example. The $L_2, \mathscr{B}$ approximation is plotted as a red square, as the barycenter of both the sets of $L_1, \mathscr{B}$ (blue segment) and $L_\infty, \mathscr{B}$ (green quadrangle) approximations. Contour-based and maximal outer approximations are in this example the extreme of the segment $L_1, \mathscr{B}$ (blue squares). The partial outer consonant approximations (yellow), the isopignistic approximation (star) and the various L_p partial approximations in $\mathscr{M}$ (in gray levels) are also drawn.

$\mathscr{M}$) as vectors $\mathbf{m}$ of mass values. As for the approximations (6) in $\mathscr{B}$, we have $\mathscr{L}_1 = \{b(x), b(x,z)\}$ and $\mathscr{L}_2 = \{b(x,y)\}$, so that $\min(\mathscr{L}_1) = int_1(\mathscr{L}_1) = b(x)$, $\max(\mathscr{L}_1) = int_2(\mathscr{L}_1) = b(x,z)$, $ave(\mathscr{L}_1) = \frac{b(x)+b(x,z)}{2}$, while $\min(\mathscr{L}_2) = int_1(\mathscr{L}_2) = \max(\mathscr{L}_2) = int_2(\mathscr{L}_2) = ave(\mathscr{L}_2) = b(x,y)$. Therefore, the set of L_1 partial consonant approximations is, by Equation (2), a segment with vertices: $[b(x), b(x,y) - b(x), 1 - b(x,y)]'$, $[b(x,z), b(x,y) - b(x,z), 1 - b(x,y)]'$ (the blue segment in Figure 2). The partial L_2 approximation in $\mathscr{B}$ is, by Equation (6), unique (red square) and coincides (in this special case) with the barycenter of the set of partial L_∞ approximations (green quadrangle): $\mathbf{m}_{co^{\mathcal{C}}_{L_2}[b]} = \mathbf{m}_{co^{\mathcal{C}}_{L_\infty}[b]} = [(b(x) + b(x,z))/2, b(x,y) - (b(x) + b(x,z))/2, 1 - b(x,y)]'$. The set of partial L_∞ approximations has the following four vertices (5): $[(b(x) + b(x,z))/2 - b(y,z), b(x,y) - (b(x) + b(x,z))/2, 1 - b(x,y) + b(y,z)]'$, $[(b(x) + b(x,z))/2 - b(y,z), b(x,y) - (b(x) + b(x,z))/2 + 2b(y,z), 1 - b(x,y) - b(y,z)]'$, $[(b(x) + b(x,z))/2 + b(y,z), b(x,y) - (b(x) + b(x,z))/2 - 2b(y,z), 1 - b(x,y) + b(y,z)]'$, $[(b(x) + b(x,z))/2 + b(y,z), b(x,y) - (b(x) + b(x,z))/2, 1 - b(x,y) - b(y,z)]'$.

Admissibility. Geometric approximation in the belief space generates solutions which are in general only partially admissible, i.e., they may contain approximations with negative masses. However, sufficient conditions on the desired maximal chain under which they are indeed admissible can be given in terms of the list of belief values (3). As $\min(\mathscr{L}_{i-1}) = b(A_{i-1}) \leq b(A_i) = \min(\mathscr{L}_i)$, the maximal partial outer approximation co_{max} is admissible for all maximal chains $\mathscr{C}$.

As for the contour-based approximation co_{con}, $\max(\mathscr{L}_i) = b(A_i + A_{i+1}^c) = b(x_{i+1}^c) = 1 - pl_b(x_{i+1})$ (when once again $x_i \doteq A_i \setminus A_{i-1}$), while $\max(\mathscr{L}_{i-1}) = 1 - pl_b(x_i)$, so that $\max(\mathscr{L}_i) - \max(\mathscr{L}_{i-1}) = pl_b(x_i) - pl_b(x_{i+1})$, which is guaranteed non-negative *if and only if the chain $\mathscr{C}$ is generated by singletons sorted by their plausibility values*. As a consequence, the barycenter of the set of L_∞ approximations is also admissible on the same chain(s). A similar condition holds in the L_1, L_2 cases [7].

4 Conclusions

From the example of Figure 2 geometric approximations in mass and belief spaces do not appear to be strongly linked. Indeed, their semantic is different, as in the mass space [7] L_p consonant approximations are associated with different but related *mass redistribution* processes: the mass outside the desired chain of focal elements is re-assigned in some way to the elements of the chain. As for the isopignistic approximation, it naturally fits in the context of the Transferable Belief Model and is quite unrelated to approximations in both the mass and the belief space. It would be interesting, in this respect, to study the property of geometric consonant approximations (which seem to be related the plausibilities of the singletons) with respect to other major probability transforms, such as the intersection probability or relative plausibility and belief of singletons. In conclusion then, isopignistic, mass-space and belief-space consonant approximations form three distinct families of approximations, with fundamentally different rationales: which approach to use will therefore vary according to the chosen framework, and the problem at hand.

References

1. Aregui, A., Denoeux, T.: Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. International Journal of Approximate Reasoning 49(3), 575–594 (2008)
2. Baroni, P.: Extending consonant approximations to capacities. In: Proceedings of Information Processing and Management of Uncertainty, IPMU, pp. 1127–1134 (2004)
3. Cuzzolin, F.: Two new Bayesian approximations of belief functions based on convex geometry. IEEE Trans. on Systems, Man, and Cybernetics B 37(4), 993–1008 (2007)
4. Cuzzolin, F.: A geometric approach to the theory of evidence. IEEE Trans. on Systems, Man, and Cybernetics C 38(4), 522–534 (2008)
5. Cuzzolin, F.: Geometric conditioning of belief functions. In: Proceedings of the First Workshop on the Theory of Belief Functions (2010)
6. Cuzzolin, F.: The geometry of consonant belief functions: simplicial complexes of necessity measures. Fuzzy Sets and Systems 161(10), 1459–1479 (2010)
7. Cuzzolin, F.: L_p consonant approximations of belief functions. IEEE Trans. on Fuzzy Systems (under review)
8. Dubois, D., Prade, H.: Possibility theory. Plenum Press (1988)
9. Dubois, D., Prade, H.: Consonant approximations of belief functions. International Journal of Approximate Reasoning 4, 419–449 (1990)
10. Dubois, D., Prade, H., Sandri, S.: On possibility-probability transformations. In: Fuzzy Logic: State of the Art, pp. 103–112. Kluwer Academic Publisher (1993)

11. Joslyn, C., Klir, G.: Minimal information loss possibilistic approximations of random sets. In: Proc. of the FUZZ-IEEE Conference, pp. 1081–1088 (1992)
12. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. Int. Journal of Approximate Reasoning 53(2), 118–145 (2012)
13. Khatibi, V., Montazer, G.: A new evidential distance measure based on belief intervals. Scientia Iranica - Transactions D 17(2), 119–132 (2010)
14. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
15. Shi, C., Cheng, Y., Pan, Q., Lu, Y.: A new method to determine evidence distance. In: Proceedings of CiSE, pp. 1–4 (2010)
16. Smets, P.: The nature of the unnormalized beliefs encountered in the transferable belief model. In: Proceedings of Uncertainty in Artificial Intelligence, pp. 292, 229 (1992)
17. Smets, P.: Belief functions on real numbers. International Journal of Approximate Reasoning 40(3), 181–223 (2005)
18. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)

Controling the Number of Focal Elements

Some Combinatorial Considerations

Christophe Osswald

Abstract. A basic belief assignment can have up to 2^n focal elements, and combining them with a simple conjunctive operator will need $\mathcal{O}(2^{2n})$ operations. This article proposes some techniques to limit the size of the focal sets of the bbas to be combined while preserving a large part of the information they carry.

The first section revisits some well-known definitions with an algorithmic point of vue. The second section proposes a matrix way of building the least committed isopignistic, and extends it to some other bodies of evidence. The third section adapts the k-means algorithm for an unsupervized clustering of the focal elements of a given bba.

Keywords: Basic belief assignments, Combinatorial complexity, Focal elements, k-means, Pignistic probability, Body of evidence, Least commitment.

1 General Considerations on Basic Belief Assignments

Let the finite set $X = \{x_1, \ldots, x_n\}$ be our frame of discernment. The size of X will be noted $n = |X|$. The set of all the subsets of X will be noted 2^X.

Definition 1. *[Shafer(1976)] The application m from 2^X to $[0, 1]$ is a* basic belief assignment *(bba) if :*

$$\sum_{A \subseteq X} m(A) = 1 \tag{1}$$

The constraint of *closed world* is modeled by $m(\emptyset) = 0$. If $m(\emptyset)$ is greater than 0, we either have an *open world* or a *conflict* within the information.

Christophe Osswald
ENSTA Bretagne, Lab-STICC UMR 3192
e-mail: Christophe.Osswald@ensta-bretagne.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 135–143.
springerlink.com

Definition 2. *Let m be a bba on X. $A \subseteq X$ is a* focal element *of m if $m(A) > 0$. The* focal set *of m is composed of all its focal elements :*

$$F(m) = \{A \subseteq X \mid m(A) > 0\} \tag{2}$$

The size of m is noted $|m| = \mathrm{Card}(F(m))$.

Of course, $|m| \leqslant 2^n$. In most applications, $|m|$ will be very small compared to 2^n when a bba is constructed from a source's information, but after some steps of combination, this limit can be reached.

Definition 3. *Let m be a bba on X. The most usual* bodies of evidence *are :*

- *The* belief*:*
$$\mathrm{bel}(A) = \sum_{\substack{B \subseteq A,\\ B \neq \emptyset}} m(B) = \sum_{\substack{B \subseteq A,\\ B \neq \emptyset,\\ B \in F(m)}} m(B) \tag{3}$$

- *The* plausibility*:*
$$\mathrm{pl}(A) = \sum_{B \cap A \neq \emptyset} m(B) = \sum_{\substack{B \cap A \neq \emptyset,\\ B \in F(m)}} m(B) \tag{4}$$

- *The* commonality*:*
$$\mathrm{q}(A) = \sum_{B \supseteq A} m(B) = \sum_{\substack{B \supseteq A,\\ B \in F(m)}} m(B) \tag{5}$$

- *The* pignisitic probability*, which is additive (knowing* $\mathrm{betP}(\{x\})$ *for all $x \in X$ is sufficient):*

$$\mathrm{betP}(A) = \frac{1}{1 - m(\emptyset)} \sum_{B \subseteq X} \frac{|A \cap B|}{|B|} m(B) = \frac{1}{1 - m(\emptyset)} \sum_{B \in F(m)} \frac{|A \cap B|}{|B|} m(B) \tag{6}$$

When the context is not obvious, the bba used to define the body of evidence will be placed as an index : $\mathrm{betP}_m(A)$ *instead of* $\mathrm{betP}(m)$.

In the definition 3, the first expression concerns all the subsets of X, and the second expression concerns only the focal elements. Therefore, if f is either of the bodies of evidence, and A a subset of X, a natural implementation of the equation brings an algorithm which calculates $f(A)$ in $\mathcal{O}(2^n)$ operations with the first expression. As the second expression only browses the focal set of m, its complexity is $\mathcal{O}(|m|)$, for the same result.

The most popular combination operator is the non-normalized conjunctive rule, also known as Smet's rule. It is a quite simple operator to implement; it is associative, and therefore allows to combine many sources.

Definition 4. *Let m_1 and m_2 be two bbas on X. The* conjunctive combination *of m_1 and m_2 is a bba on X, $m_1 \oplus m_2$, defined by :*

$$(m_1 \oplus m_2)(A) = \sum_{\substack{B \subseteq X, \\ C \subseteq X, \\ B \cap C = A}} m_1(B) m_2(C) = \sum_{\substack{B \in F(m_1), \\ C \in F(m_2), \\ B \cap C = A}} m_1(B) m_2(C) \quad (7)$$

The cost for calculating $B \cap C$ is $\mathcal{O}(n)$. The first expression brings an algorithm in $\mathcal{O}\left(n 2^{2n}\right)$ operations for calculating $(m_1 \oplus m_2)(A)$, and $\mathcal{O}\left(n 2^{3n}\right)$ for determining $m_1 \oplus m_2$. The second expression brings an algorithm in $\mathcal{O}\left(n|m_1||m_2|\right)$ operations for calculating $(m_1 \oplus m_2)(A) = (m_1 \oplus m_2)(A)$, and $\mathcal{O}\left(n 2^n |m_1||m_2|\right)$ for determining $m_1 \oplus m_2$.

Smets [Smets(2002)] proposed a nice implementation in $\mathcal{O}(n 2^n)$ operations for transformations between bba and commonality. The conjunctive combination of the commonality functions is a simple multiplication, which is linear, but on vectors having a size of 2^n.

The expression (7), nor the commonality, can prevent us from making operations on non-focal elements of $m_1 \oplus m_2$. Let the bba be implemented by an adaptive structure that contains information only for its focal elements. A hashtable is a convenient way for it. The algorithm 1 uses only $\mathcal{O}(n|m_1||m_2|)$ to build $m_1 \oplus m_2$.

The size of $m_\cap$ is at most $|m_1||m_2|$. The algorithm coming from (7) needs to be executed for all the subsets of X, but the algorithm 1 only works on the focal elements of $m_\cap$, and does not compute useless intersections [Smets(1994)]. Using a hashtable for the focal elements, with a hashcode calculation in $\mathcal{O}(n)$ operations, the conjunctive combination takes $\mathcal{O}(n|m_1||m_2|)$ operations.

Data: bbas m_1, m_2
Result: bba $m_\cap$
forall the $B \in m1$ **do**
  forall the $C \in m_2$ **do**
    if $B \cap C \in m_\cap$ **then**
      $m_\cap(B \cap C) \leftarrow m_\cap(B \cap C) + m_1(B) m_2(C)$
    else
      Add $B \cap C$ to $m_\cap$
      $m_\cap(B \cap C) \leftarrow m_1(B) m_2(C)$

Algorithm 1: Conjunctive combination

However, the very nature of the combination operator brings a combinatorial explosion of the focal set. Let m_i be the bba defined by $m_i(X) = \frac{1}{2}$ and $m_i(X \backslash \{x_i\}) = \frac{1}{2}$: $|m_i| = 2$. Let $m_\cap$ be the conjunctive combination of all those bbas : $m_\cap = m_1 \oplus \ldots \oplus m_n$. For any $A \subseteq X$, $m_\cap(A) = \frac{1}{2^n}$. Therefore, $F(m_\cap) = 2^X$ and $|m_\cap| = 2^n$.

The objective of the following sections will be to guarantee that the size of a bba cannot be too large, and to respect its nature as much as possible.

2 Linear Algebra for bbas

The definition 3 builds the bodies of evidence bel, pl, betP and q as linear transformations of m. Considering a bba m on X and an integer K, our objective will be to build an bba m' on X such that $|m'| \leqslant K$ and $f_{m'}(A) = f_m(A)$ for some bodies of evidence f and some subsets A of X.

Within this section, we forbid $\emptyset$ to be a focal element of m, and we do not allow it to become a focal element of m'. As convenient consequences, we have $\mathrm{bel}(A) \leqslant \mathrm{betP}(A) \leqslant \mathrm{pl}(A)$, $\mathrm{bel}(X) = 1$, and $\mathrm{pl}(X) = 1$.

A popular and efficient way to build a bba from a probability or another source of uncertain information is to build a least committed bba having the same pignistic probability than the source [Smets(1990)].

Definition 5. *Let m be a bba on X. A bba m' is an* isopignisitic *of m if*

$$\forall x \in X,\ \mathrm{betP}_m(x) = \mathrm{betP}_{m'}(x) \tag{8}$$

The bba m' is the least committed isopignistic *of m if for any isopignistic m'' of m and for any $A \subseteq X$, $\mathrm{pl}_{m'}(A) \geqslant \mathrm{pl}_{m''}(A)$.*

The algorithm 2 builds the least committed isopignistic in $\mathcal{O}(n^2 + n|m|)$ operations. It contains at most n focal elements.

Data: bba m on X
Result: bba m' on X
forall the $x \in X$ **do**
  Calculate $p[i] = \mathrm{betP}(x)$
$A \leftarrow X$; $k \leftarrow |X|$
while $\max(p) \neq 0$ **do**
  $i \leftarrow \mathrm{argmin}(p)$
  $m'(A) \leftarrow kp[i]$
  forall the $j \in p$ **do**
    $p[j] \leftarrow p[j] - p[i]$
  Delete element i from p
  $A \leftarrow A \backslash \{x_i\}$; $k \leftarrow k - 1$

Algorithm 2: Building the least committed isopignistic

If we calculate $\mathrm{betP}(x)$ for all $x \in X$, and order the elements of X such that $p_i = \mathrm{betP}(x_i) \geqslant \mathrm{betP}(x_{i+1}) = p_{i+1}$, the focal elements of the least committed isopignistic are a subset of the $A_i = \{x_1, \ldots, x_i\}$.

We have

$$p_i = \mathrm{betP}(x_i) = \sum_{k=i}^{n} \frac{1}{k} m'(A_k) \tag{9}$$

Let p be the vector of the p_i and y be the vector of the $m'(A_i)$. We have $p = \mathrm{Bet}y$ with Bet a $n \times n$ matrix, triangular and inversible. Therefore $y = \mathrm{Bet}^{-1}p$, with

$$\mathrm{Bet} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n-1} & \frac{1}{n} \\ 0 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n-1} & \frac{1}{n} \\ \vdots & 0 & \frac{1}{3} & \cdots & \frac{1}{n-1} & \frac{1}{n} \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \frac{1}{n-1} & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{n} \end{pmatrix}, \quad \mathrm{Bet}^{-1} = \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 2 & -2 & 0 & & \vdots \\ 0 & 0 & 3 & -3 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & (n\text{–}1) & -(n\text{–}1) \\ 0 & \cdots & \cdots & \cdots & 0 & n \end{pmatrix} \tag{10}$$

As Bet^{-1} is a triangular band matrix, we can compute all the $m'(A_i)$ from p_i in $\mathcal{O}(n)$ operations.

With $\mathcal{O}(n|m|)$ operations for computing betP, $\mathcal{O}(n \ln n)$ operations for sorting X, $\mathcal{O}(n)$ operations for building the sets A_i (with an adapted data structure) and $\mathcal{O}(n)$ operations for solving the linear system, building the least committed isopignistic costs $\mathcal{O}(n(\ln n + |m|))$ operations. Usually, $|m| \gg \ln n$, and the cost of the least committed isopignistic is not greater than the cost of computing $\mathrm{betP}(x)$ for the elements of X.

The interval $[\mathrm{bel}(A), \mathrm{pl}(A)]$, containing $\mathrm{betP}(A)$, can be interpretated as an uncertainty on A [Janez and Appriou(1996)]. For singletons, bel is trivial: $\mathrm{bel}(x) = m(x)$. For sets of size $n-1$, pl is trivial: $\mathrm{pl}(X \backslash \{x\}) = 1 - m(\{x\})$. Considering the non-trivial bodies of evidence on the sets of interest $\{x_1\}$, $\ldots$, $\{x_n\}$, $B_1 = X \backslash \{x_1\}$, $\ldots$, $B_n = X \backslash \{x_n\}$, we search a bba m' with those focal elements, forming a vector

$$\mathbf{y} = (m'(\{x_1\}), \ldots, m'(\{x_n\}), m'(B_1), \ldots, m'(B_n))^T \tag{11}$$

which verifies:

$$\forall i \in \lfloor 1, n \rfloor, \ \ \mathrm{pl}_{m'}(\{x_i\}) = \mathrm{pl}_m(\{x_i\}) \tag{12}$$

$$\forall i \in \lfloor 1, n \rfloor, \ \ \mathrm{bel}_{m'}(B_i) = \mathrm{bel}_m(B_i) \tag{13}$$

We have:

$$\mathrm{pl}_{m'}(\{x_i\}) = m'(\{x_i\}) + \sum_{j \neq i} m'(B_j) \tag{14}$$

$$\mathrm{bel}_{m'}(B_i) = \sum_{j \neq i} m'(\{x_j\}) + m'(B_i) \tag{15}$$

As $\forall i$, $\mathrm{pl}_{m'}(\{x_i\}) + \mathrm{bel}_{m'}(B_i) = \sum_i m'(\{x_i\}) + \sum_i m'(B_j)$, there are only $n+1$ independent equations among the $2n$ listed above: we cannot guarantee to kep at the same time $\mathrm{pl}_m(\{x_i\})$ and $\mathrm{bel}_m(B_i)$ on those $2n$ focal elements.

As $\mathrm{q}(B_i) = m(B_i) + m(X)$ and $\mathrm{q}(\{x_i\}) = \mathrm{pl}(\{x_i\})$, introducing commonality does not bring any new independent equation.

2.1 *Mixing* Bet *with Other Bodies of Evidence*

Here we search a bba with $2n$ focal elements which is an isopignistic of m and respects an other body of evidence on some focal elements. In the following examples, we allow the A_i obtained in section 2 to be focal elements, and we complete them with $(\{x_i\})_{i\in\lfloor 1,n\rfloor}$ or the $(B_i)_{i\in\lfloor 1,n\rfloor}$.

With plausibility, we should use the focal elements $(\{x_i\})_{i\in\lfloor 1,n\rfloor}$. We build a vector

$$\mathbf{y} = (m'(\{x_1\}), \ldots, m'(\{x_n\}), m'(A_1), \ldots, m'(A_n))^T \tag{16}$$

The constraints are:

$$\mathrm{betP}(x_i) = m'(\{x_i\}) + \sum_{k=i}^{n} \frac{1}{k} m'(A_k) \tag{17}$$

$$\mathrm{pl}(\{x_i\}) = m'(\{x_i\}) + \sum_{k=i}^{n} m'(A_k) \tag{18}$$

As $A_1 = \{x_1\}$, we cannot have $m'(A_1) \neq m'(\{x_1\})$; we have only $2n-1$ focal elements. We drop the term $m'(\{x_1\})$ in y, and the constraint on $\mathrm{pl}(\{x_i\})$ to obtain a matrix P such that

$$P\mathbf{y} = (\mathrm{pl}_m(\{x_2\}), \ldots, \mathrm{pl}_m(\{x_2\}), \mathrm{betP}(x_1), \ldots, \mathrm{betP}(x_n))^T \tag{19}$$

The matrix P_4 and more generally P_n are:

$$P_4 = \begin{pmatrix} 0&0&0&1&\frac{1}{2}&\frac{1}{3}&\frac{1}{4} \\ 1&0&0&0&\frac{1}{2}&\frac{1}{3}&\frac{1}{4} \\ 0&1&0&0&0&\frac{1}{3}&\frac{1}{4} \\ 0&0&1&0&0&0&\frac{1}{4} \\ 1&0&0&0&1&1&1 \\ 0&1&0&0&0&1&1 \\ 0&0&1&0&0&0&1 \end{pmatrix}, \quad P_n = \left(\begin{array}{c|c} \begin{matrix}0\cdots 0\end{matrix} & \\ \hline I_{n-1} & \mathrm{Bet}_n \\ \hline I_{n-1} & \begin{matrix}0\\ \vdots \\ 0\end{matrix} \; U_{n-1} \end{array}\right) \tag{20}$$

where Bet_n is matrix obtained in the section 2 and U_{n-1} the upper triangular $(n-1)\times(n-1)$ matrix full of 1.

The matrix P_n is inversible, and we can solve this system in $\mathcal{O}(n^3)$ operations. Overall, we can reduce the focal set of m to $2n-1$ focal elements in $\mathcal{O}(n(n^2+|m|))$ operations, respecting betP and pl on the singletons.

With commonality, we obtain the same results : $\mathrm{q}(\{x_i\}) = \mathrm{pl}(\{x_i\})$.

With belief, we should use $(B_i)_{i\in\lfloor 1,n\rfloor}$ as focal elements instead of $(\{x_i\})$. As $\mathrm{bel}(B_i)+\mathrm{pl}(\{x_i\})=1$, we obtain another – but similar – $(2n-1)\times(2n-1)$ inversible matrix.

3 Optimatization by *k*-Means

Denoeux and Yaghlane(2002) proposed to reduce a bba by adapting the single linkage hierarchical clustering algorithm to coarsen its focal set. Another interesting family of unsupervized clustering algorithm are the k-means techniques, born from the ISODATA method of Ball and Hall(1965). One can adapt this method to find a subset $\mathcal{K}$ of 2^X limited in size: $|\mathcal{K}|\leqslant k$.

Usual k-means does not guarantee an optimal choice of centers: finding them is equivalent to the MINIMUM-k CENTER, which is a NP-Complete problem Garey and Johnson(1979). The convergence of the k-means algorithm is guaranteed, but only to a local minimum of the intra-cluster variance.

Data: bba m, integer k with $k\leqslant|m|$
Result: bba m_k
Let $C[1],\ldots,C[k]$ be k focal elements of m [1]
repeat
  forall the $j\leqslant k$ **do** $\mathcal{C}[j]\leftarrow\emptyset$
  forall the $A\in m$ **do** [2]
    $\mathcal{C}[\mathrm{argmin}(\mathrm{dist}(A,C_j))]\leftarrow\mathcal{C}[\mathrm{argmin}(\mathrm{dist}(A,C_j))]\cup\{A\}$
  forall the $j\leqslant k$ **do** $C[j]\leftarrow$ center of $\mathcal{C}[j]$; [3]
until *ending condition reached* [4]
forall the $j\leqslant k$ **do**
  $m_k(C[j])\leftarrow\sum_{A\in\mathcal{C}[j]}m(A)$

Algorithm 3: k-means, in a general way that applies to focal elements.

[1] It is natural to initialize the algorithm with the k focal elements with the greatest masses. But, as the algorithm converges – if it converges – to a local minimum, it should be a good idea to execute various instances, with random starting sets.

[2] The focal element A is affected to the center $C[j]$ such that

$$\mathrm{dist}(A,C[j])=\left|\left(A\cap\overline{C[j]}\right)\cup\left(\overline{A}\cap C[j]\right)\right| \tag{21}$$

is minimal. It corresponds to a natural L_1 distance based on an exclusive OR. In case of equal distances to different centers, it is possible to:

- choose a random one *(the algorithm is no longer deterministic)*
- use a lexicographical order *(elements are no longer equivalent)*
- try to build balanced clusters *(the underlying problem is NP-complete)*

[3] The usual k-means technique uses the geometrical barycenter of the focal sets of $\mathcal{C}[j]$ seen as points of $[0,1]^n$: $C[j] \leftarrow \sum_{A\in\mathcal{C}[j]} m(A)A$.
It would build fuzzy focal elements, which is not the way the definition 2 accepts them. Therefore, we put x in the new $C[j]$ if and only if :

$$\sum_{A\in\mathcal{C}[j],x\in A} m(A) > \sum_{A\in\mathcal{C}[j],x\notin A} m(A) \tag{22}$$

[4] As we "move" the centers of the classes to the nearest sharp subset of X, the total intra-cluster variance is not necessarily decreasing. Therefore, the ending condition must include a maximum steps number, and/or test the cycles it should encounter.

A step of the algorithm 3 costs $\mathcal{O}(kn|m|)$ operations. A reasonable number of steps before ending the loop is k, and we obtain an algorithm in $\mathcal{O}(k^2n|m|)$ operations. If we want to compare this approach with the ones of the section 2.1, we should use $k = 2n-1$, and get an algorithm in $\mathcal{O}(n^3|m|)$ operations.

4 Conclusion

In a general way, dealing with basic belief assignments on large frames of discernment need a proper encoding of the focal sets. We propose to use hashtables for this purpose, but this not the only way. We propose two categories of methods for restricting any bba to a bba modest in focal set size.

We extend the principle of isopignistic to other bodies of evidence to build a bba with only $2n$–1 focal elements, respecting both the pignistic probability and another body of evidence of the original bba. We first determine the value of the bodies of evidence on some simple elements, and then determine the restricted focal set. A linear equation gives the restricted bba.

Trying to restrict the focal set to a number of respresentative elements leads to a NP-Complete problem. We adapt the k-mean algorithm to build a heuristical solution. It is more expensive, but it does not need to define *a priori* a focal set, and can adapt to more situations.

References

[Ball and Hall(1965)] Ball, G.H., Hall, D.J.: Isodata, a novel methed of data analysis and pattern classification. Tech. rep., Stanford Research Institute (1965)

[Denoeux and Yaghlane(2002)] Denoeux, T., Yaghlane, A.B.: Approximating the combination of belief functions using the fast moebius transform in a coarsened frame. International Journal of Approximate Reasoning 31(1-2), 77–101 (2002)

[Garey and Johnson(1979)] Garey, M.R., Johnson, D.S.: Computers and intractability – a guide to the theory of NP-Completeness. Freeman (1979)

[Janez and Appriou(1996)] Janez, F., Appriou, A.: Théorie de l'Evidence et cadres de discernement non exhaustifs. Traitement du Signal 13(3), 237–250 (1996)
[Shafer(1976)] Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
[Smets(1990)] Smets, P.: Constructing the pignistic probability function in a context of uncertainty. Uncertainty in Artificial Intelligence 5, 29–39 (1990)
[Smets(1994)] Smets, P.: The transferable belief model. Artificial Intelligent 66, 191–234 (1994)
[Smets(2002)] Janez, F., Appriou, A.: Théorie de l'Evidence et cadres de discernement non exhaustifs. Traitement du Signal 13(3), 237–250 (1996)

Random Generation of Mass Functions: A Short Howto

Thomas Burger and Sébastien Destercke

Abstract. As Dempster-Shafer theory spreads in different applications fields involving complex systems, the need for algorithms randomly generating mass functions arises. As such random generation is often perceived as secondary, most proposed algorithms use procedures whose sample statistical properties are difficult to characterize. Thus, although they produce randomly generated mass functions, it is difficult to control the sample statistical laws. In this paper, we briefly review classical algorithms, explaining why their statistical properties are hard to characterize, and then provide simple procedures to perform efficient and controlled random generation.

1 Introduction

In this paper, we concentrate on the question of simulating and sampling belief functions, or equivalently mass distributions, which are the central elements of Dempster-Shafer theory [1,2]:

Let Ω be a finite space, $\mathcal{P}(\Omega)$ its power set, and let m be a *mass function* on Ω, i.e. an application from $\mathcal{P}(\Omega) \mapsto [0,1]$ such that $\sum_{A\in\mathcal{P}(\Omega)} m(A) = 1$. Let $\mathcal{M}_\Omega$ be the set of mass functions defined on Ω. Classically, mass functions are distinguished according to their *support* $\mathcal{F}_m$, i.e., the set of all its *focal elements*. A focal element is a subset $A \subseteq \Omega$ such that $m(A) > 0$.

Although it is often overlooked (for the reason that the theory has to face many other algorithmic issues), the problem of random generation of mass functions is useful and necessary in many problems: First, mass functions are complex uncertainty representations often summarized by some descriptors (information measures [3], distances [4], conflict, ...) in practical methods, and the efficiency of these

Thomas Burger
CNRS (FR3425), CEA (iRTSV/BGE), INSERM (EDyP, U1038), Grenoble, France
e-mail: thomas.burger@cea.fr

Sebastien Destercke
CNRS, UMR Heudiasyc, Centre de recherche de Royallieu, 60205 COMPIEGNE
e-mail: sebastien.destercke@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 145–152.
springerlink.com

descriptors has to be evaluated from a statistical point of view (see, e.g. [4]). Second, simulations can be helpful to test some conjectures prior to demonstrating it. Third, simulation and sampling are useful tools to produce data sets to test and calibrate data fusion methods or learning algorithms [5,6].

The problem of simulating mass functions is equivalent to randomly sampling elements out of $\mathcal{M}_\Omega$, according to a particular distribution $\mathcal{D}$. In theory, $\mathcal{D}$ could be any distribution, but in practice such generality is seldom required. Three main situations occur:

1. the behavior of some descriptor over $\mathcal{M}_\Omega$ has to be tested. In this case, it is necessary to perform a uniform sampling over all $\mathcal{M}_\Omega$;
2. mass functions are assumed to be restricted to some specific form, i.e. an expert providing consonant mass functions, a classifier providing simple mass assignments such as 2-additive, etc. In this case, the restriction describes a subregion $\mathcal{S}$ of $\mathcal{M}_\Omega$, and it is necessary to sample uniformly from $\mathcal{S}$.
3. mass functions can be general (belong to $\mathcal{M}_\Omega$) but should follow some tendency while still being pervaded with some randomness. This can happen, for instance, when one wants to simulate a training set of data with uncertain labels from a training set with known labels. In such a case, sampling procedure should (on the long run) give higher masses to sets containing the true label but still allow wrong masses to be sampled with lower probability.

From a mathematical point of view, the first case is the simplest, and the two others are generalizations of the first (the second requires a selection of focal elements and the third requires the definition of $\mathcal{D}$). Note that depending on $\mathcal{S}$, the second may be rather complicated, and so is the combination of the second and third cases (non-uniform sampling on arbitrary domain). Thus, in this paper, we completely deal with the first and third cases, while only giving clues for the more complex simulation problems. The paper is organized as follow: Section 2 is a state of the art, where one recalls the main method used to sample mass functions and where one explains why its statistical properties are hard to characterize. Section 3 details the mathematical framework needed to develop a more controllable algorithm. Finally, in Section 4 we derive algorithms for the various cases stated above.

2 State of the Art

In practice, the most used (see e.g. [7]) algorithm is based on the following intuitive procedure (see Algorithm 1): (1) select a set of N elements of $\mathcal{P}(\Omega)$ (possibly the entire power set, i.e., $N = |\mathcal{P}(\Omega)|$) to form $\mathcal{F}_m$, (2) uniformly and independently sample N values in $[0, 1]$ corresponding to the N focal elements, and (3) perform a normalization enforcing the constraint $\sum_{A\in\mathcal{P}(\Omega)} m(A) = 1$. Usually, the binary representation is used to order elements of $\mathcal{P}(\Omega)$ and sampling of subsets consists in drawing a number between 0 and $2^{|\Omega|} - 1$ and associate to it the subset corresponding to its binary conversion.

The main problem with Algorithm 1 is that the distribution $\mathcal{D}$ it generates on $\mathcal{M}_\Omega$ is difficult to characterize[1]. At first sight, one may think that it generates a

[1] This is also the case in [6], even if the author does not aim to control the distributions.

Algorithm 1: A classical algorithm to randomly generate a mass function.

Input: frame Ω, number of focal elements N
Output: random mass function m

1 $\mathcal{P} \leftarrow$ `generatePowerSetof`(Ω);
2 $\mathcal{F} \leftarrow$ `getFocalElements`$(N, \mathcal{P})$;
3 **foreach** $1 \leq i \leq N$ **do**
4 $\quad m_i \leftarrow$ `randomlySample`$(\mathcal{U}([0,1]))$;
5 **foreach** $1 \leq i \leq N$ **do**
6 $\quad m(\mathcal{F}_i) \leftarrow {}^{m_i}/_{\sum_{k=1}^{N} m_k}$;

uniform distribution, however this is not what happens. The reason is that although (at Line 5) values are independently sampled from $\mathcal{U}([0,1])$ (i.e. the uniform law on $[0,1]$), they are normalized afterward (Line 6). Since the normalization of each value involves all the other values, they are not realizations of i.i.d. random variables. Thus, if uniform sampling is sought, other algorithms are required. These latter are based on well-known probability distributions which are presented in the next section.

3 Mathematical Tools

In this section, we recall the mathematics behind the algorithms: we briefly explain the correspondence between mass functions and categorical distributions (Section 3.1), which are well-known in Bayesian statistics. Then (Section 3.2), we recall that Dirichlet distributions can be used to sample categorical distributions (hence mass functions in $\mathcal{M}_\Omega$). Finally, we use the relation between Dirichlet and gamma distributions to build efficient algorithms (Section 3.3).

3.1 The Categorical Family and $\mathcal{M}_\Omega$

The *k-way categorical distribution* is just a (discrete) probability distribution defined on k exhaustive and exclusive outcomes $\mathcal{O}_k = \{O_1, \dots O_k\}$ with probabilities $p_1, \dots, p_k$, the well-known *Bernouilli distribution* corresponding to $k = 2$. As well as the *binomial distribution* is the probability of the number of successes among n Bernouilli trials, the *multinomial distribution* is the probability distribution that describes the repartition amongst categories $O_1, \dots O_k$ of n categorical trials.

A categorical distribution is defined by the probabilities $p_i = \mathbb{P}(X = O_i)$, $i = 1, \dots, k$, with $\sum_{i=1}^{k} p_i = 1$ and $p_i \geq 0$. This means the vector $(p_1, \dots, p_k)$ is in the $(k-1)$-dimensional simplex, denoted $\mathcal{C}_k$ and called here the *k-way categorical family*. Clearly, the probabilities p_i can act as masses given to focal elements, as illustrates the example.

Example 1. *Consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, and a mass function $m \in \mathcal{M}_\Omega$ such that $m(\{\omega_1\}) = 0.2$, $m(\{\omega_2\}) = 0.3$ and $m(\{\omega_1, \omega_3\}) = 0.5$. The vector modelling $m \in \mathcal{M}_\Omega$ is $\{0, 0.2, 0.3, 0, 0, 0.5, 0, 0\}$. It is equivalent to the 8-way categorical*

distribution defined on $\mathcal{O}_8$ and formalized as a vector $(p_1, \ldots, p_8)$ such that outcome O_2, O_3 and O_6 have probabilities of 0.2, 0.3 and 0.5 of happening, respectively.

Theorem 1. *$\mathcal{M}_\Omega$ is isomorphic to $\mathcal{C}_{(2^{|\Omega|})}$.*

Proof (sketch). There are two ways to show this simple theorem, each shedding different light on the related problem.

In the first demonstration, let us simply show that the elements of $\mathcal{M}_\Omega$ and the elements of $\mathcal{C}_{(2^{|\Omega|})}$ are in one to one correspondence. A compact notation for a mass function m on Ω is the binary ordered list $\{p_1, \ldots, p_{2^{|\Omega|}}\}$ of $2^{|\Omega|}$ values such that the jth value p_j is $m(A_j)$, where A_j has j for binary representation. Moreover, as $m(A_{2^{|\Omega|}}) = 1 - \sum_{\ell=1}^{2^{|\Omega|}-1} m(A_\ell)$, any mass function m is uniquely identified by $\{p_1, \ldots, p_{2^{|\Omega|}-1}\}$, which uniquely defines a $(2^{|\Omega|})$-way categorical distribution. Conversely, any $(2^{|\Omega|})$-way categorical distribution can be seen as a mass function, hence $\mathcal{M}_\Omega$ is isomorphic to $\mathcal{C}_{|\mathcal{P}(\Omega)|} = \mathcal{C}_{(2^{|\Omega|})}$.

The second demonstration relies on more geometric arguments: $\mathcal{C}_{(2^{|\Omega|})}$ is known to be the *standard* $2^{|\Omega|}-1$ *simplex*, while Cuzzolin [8] established in his geometrical interpretation of Dempster-Shafer theory that $\mathcal{M}_\Omega$ is also the standard $2^{|\Omega|} - 1$ simplex. □

Thus sampling uniformly a mass function m out of $\mathcal{M}_\Omega$ is equivalent to uniformly sampling $(2^{|\Omega|})$-way categorical distributions out of $\mathcal{C}_{(2^{|\Omega|})}$. This is particularly useful, as the distribution of categorical distributions over $\mathcal{C}_{(2^{|\Omega|})}$ is the well-known Dirichlet distribution.

3.2 The Dirichlet Distribution

The Dirichlet law $\mathcal{D}ir(\pi_1, \ldots, \pi_k)$ of order $k \geq 2$ with parameters $\pi = (\pi_1, \ldots, \pi_k) \in [0, 1]^k$ describes a random variable over $\mathcal{C}_k$ that has the following probability density function

$$\mathbf{x} \to f_{\mathcal{D}ir}(\mathbf{x}; \pi) = \frac{\Gamma(\pi_0)}{\prod_{i=1}^{k} \Gamma(\pi_i)} \prod_{i=1}^{k-1} x_i^{\pi_i - 1}$$

where $\mathbf{x} = \{x_1, \ldots, x_k\} \in \mathcal{C}_k$ with $x_k = 1 - \sum_{i=1}^{k-1} x_i$, $\pi_0 = \sum_{i=1}^{k} \pi_i$ and Γ is the gamma function. $f_{\mathcal{D}ir}$ is the *conjugate prior* of the categorical distributions, as it is a probability density over all the k-way categorical distributions. Hence, a trial according to the Dirichlet distribution results in a k-way categorical distribution that can then be translated in a mass function (Theorem 1). Parameters $\pi = (\pi_1, \ldots, \pi_k) \in [0, 1]^k$ of the distribution determine the behavior of the probability density function over $\mathcal{C}_k$, that will govern sampling behavior.

Property 1. Let X be a random vector following a Dirichlet distribution of parameter $\pi = (\pi_1, \ldots, \pi_k) \in [0, 1]^k$ (or, $X \sim \mathcal{D}ir(\pi_1, \ldots, \pi_k)$ for short). We have:

$$\mathbb{E}[X] = \left(\frac{\pi_1}{\pi_0}, \cdots, \frac{\pi_i}{\pi_0}, \cdots, \frac{\pi_k}{\pi_0} \right)$$

In other words, the expected mass of the ith focal element, with mass sampled according to a Dirichlet distribution of parameters $\pi = (\pi_1, \ldots, \pi_{2^{|\Omega|}}) \in [0,1]^{2^{|\Omega|}}$ will be $\pi_i / \sum_{j=1}^{2^{|\Omega|}} \pi_j$. This property is useful to set the long-run behaviour of simulated masses. Similarly, if m is sampled uniformly from $\mathcal{M}_\Omega$, then every focal element of m should have the same expected mass. Indeed, we have a stronger result:

Property 2. If $\pi = \mathbf{1}_k$ (i.e. $\pi_i = 1 \; \forall i$), the Dirichlet distribution corresponds to the uniform distribution on $\mathcal{C}_k$

This property leads us to the following theorem (direct from Th. 1 and above prop.), which defines the uniform probability on the set of mass functions:

Theorem 2. *The uniform distribution on $\mathcal{M}_\Omega$ is given by the Dirichlet distribution of order $2^{|\Omega|}$ and of parameters $\pi_i = 1 \; \forall i \in (1, \ldots, 2^{|\Omega|})$.*

3.3 Links with the Gamma Distribution

From a theoretical point of view, we now know how to uniformly generate mass functions. However dealing directly with Dirichlet law is not practical. To solve this, we can use the link between the Dirichlet and the gamma distributions.

The gamma distribution $\mathcal{G}(\alpha, \beta)$ with shape parameter $\alpha \in \mathbb{R}_+^*$ and scale parameter $\beta \in \mathbb{R}_+^*$ has the following probability density function over $\mathbb{R}$:

$$x \to f_{\mathcal{G}}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Recall that the gamma distribution can be seen as a generalized exponential distribution, as $\mathcal{G}(\alpha = 1, \beta) = \mathcal{E}xp(\beta)$. The link between the gamma and the Dirichlet distributions is given by the following property:

Property 3. Let $X_1, \ldots, X_k$ be k independent random variables such that $X_i \sim \mathcal{G}(\alpha = \pi_i, \beta)$. The random vector $Y = (\frac{X_1}{\sum_{j=1}^k X_j}, \ldots, \frac{X_i}{\sum_{j=1}^k X_j}, \ldots, \frac{X_k}{\sum_{j=1}^k X_j})$ follows a Dirichlet law of parameters $(\pi_1, \ldots, \pi_k)$.

This means that to generate a Dirichlet distribution, we can use independent realizations of gamma distributions with identical scale parameter.

4 Algorithms

4.1 Uniform and Non-uniform Sampling on $\mathcal{M}_\Omega$

We can now use the tools of Section 3 to study the sampling of mass functions with a statistically known distribution. As generating mass functions on $\mathcal{M}_\Omega$ involves all focal elements, it comes down to sample Dirichlet laws on $\mathcal{C}_{|\mathcal{P}(\Omega)|}$. Algorithm 2 summarizes how to achieve this. We notice two main differences with Algorithm 1: First, no selection of focal elements is performed, as the entire set $\mathcal{M}_\Omega$ is considered. Second, gamma distributions are used instead of uniform ones. Note that most

Algorithm 2: Algorithm to sample from $\mathcal{M}_\Omega$.

Input: frame Ω, parameter $\pi = (\pi_1, \ldots, \pi_{2^{|\Omega|}})$
Output: random mass function m

$\mathcal{P} \leftarrow$ `generatePowerSetof(`Ω`)`;
foreach $1 \leq i \leq |\mathcal{P}|$ **do**
  $m_i \leftarrow$ `randomlySample(`$\mathcal{G}(\pi_i, 1)$`)`;
foreach $1 \leq i \leq |\mathcal{P}|$ **do**
  $m(\mathcal{P}_i) \leftarrow {}^{m_i}/_{\sum_{k=1}^{|\mathcal{P}|} m_k}$;

coding language have native functions allowing for such sampling: For instance, with `R` language, one writes `> rgamma(k, shape = 1, scale = pi_i)`.

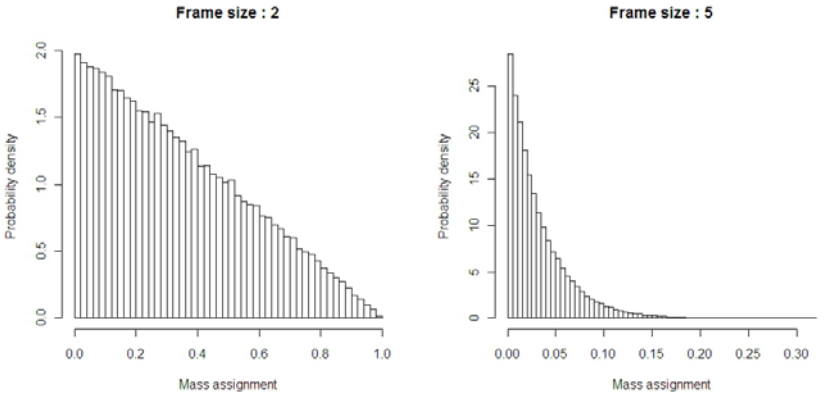
Uniform sampling is achieved by choosing $\pi_i = 1$ for $i = 1, \ldots, |\mathcal{P}(\Omega)|$, or alternatively by generating i.i.d. realizations of exponential law $\mathcal{E}xp(\beta = 1)$, as $\mathcal{G}(1, 1) = \mathcal{E}xp(\beta = 1)$. Figures 1 illustrates mass distributions obtained by Algorithm 1 and 2 for $|\Omega| = 2$ and 5, respectively. The difference is obvious, and the uniformity of Algorithm 2 can be checked in the case $|\Omega| = 2$, as we obtain the 2-dimensional simplex (i.e. a triangle with a straight angle, [8]).

Non-uniform sampling is achieved by choosing the parameters π_i proportionally to the "average" mass focal element $\mathcal{P}_i$ ($\mathcal{P}_i$ being the subset with binary encoding i) should receive, and by sampling i.i.d. realizations of $\mathcal{G}(1, \pi_i)$ distributions. For instance, to induce noisy mass functions around a true known value ω_j, the parameter π_i for all subsets such that $\omega_j \in \mathcal{P}_i$ should be higher than the ones of other subsets.

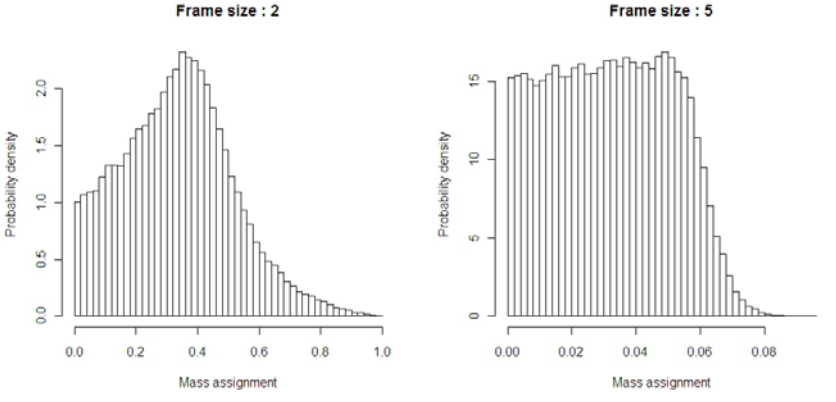
4.2 Uniform Sampling on Sub-Domains of $\mathcal{M}_\Omega$

In many situations, mass functions to be sampled have a particular structure of support (consonant, k-additive, simple support, etc.), and thus live in a subregion $\mathcal{S}$ of $\mathcal{M}_\Omega$. In the case where $\mathcal{S}$ is a single determined hyperface of $\mathcal{M}_\Omega$ (i.e. a simplex), this is not really difficult, even in case of non-uniform sampling (simply apply Algorithm 2 on the adapted simplex). However, if $\mathcal{S}$ is a randomly chosen sub-simplex or a collection of hyperfaces, non-uniform sampling may be very complicated, even impossible (how to define a distribution on the focal elements if these latter are not specified yet?). Worst, if the domain is completely arbitrary, even uniform sampling may be rather tricky. Thus we focus on uniform sampling on domains of $\mathcal{M}_\Omega$ which correspond to well-known particular mass functions: consonant, consistant, k-additive, k-intolerant, simple support, categorical, or with restrictions on the cardinality of the focal elements.

In this case, and if, in addition, the number $N \leq 2^\Omega$ of focal elements is fixed, the characterization of $\mathcal{S}$ comes down to the characterization of $\mathfrak{F}$, the set of all the possible supports in $\mathcal{S}$: $\mathfrak{F} = \{\mathcal{F}_m \subset \mathcal{P}(\Omega)/m \in \mathcal{S}, |\mathcal{F}_m| = N\}$. At this point, there are two strategies. First, if $\mathfrak{F}$ can be automatically enumerated, it is possible to sample uniformly $\mathfrak{F}$, and then, to perform N i.i.d. $\mathcal{E}xp(1)$ trials. If $\mathfrak{F}$ can not be explained, it is always possible to apply the *Acceptance-Rejection method*, which is based on the following idea: Repeat uniform samples on $\{\mathcal{F}_m \subset \mathcal{P}(\Omega)\}$, until a



Simulations with Algorithm 2



Simulations with Algorithm 1

Fig. 1 Uniform sampling, illustration

Algorithm 3: Algorithm to sample on subspaces.

Input: frame Ω, number of focal elements N, constraints defining $\mathcal{S}$
Output: random mass function m from $\mathcal{S}$

if $\mathcal{S}$ *enumerated* **then**
  $\mathcal{F} \leftarrow$ `randomlySample`$(\mathcal{U}(\mathcal{S}))$
else
  $\mathcal{P} \leftarrow$ `generatePowerSetof`(Ω);
  while $\mathcal{F} \notin \mathcal{S}$ **do**
    $\mathcal{F} \leftarrow$ `getFocalElements`$(N, \mathcal{P})$
foreach $1 \leq i \leq N$ **do**
  $m_j \leftarrow$ `randomlySample`$(\mathcal{E}xp(1))$;
foreach $1 \leq i \leq N$ **do**
  $m(\mathcal{F}_i) = {}^{m_i}/_{\sum_{k=1}^{N} m_k}$;

support in $\mathfrak{F}$ is found. The sampling law of such an algorithm is proved to correspond to $\mathcal{U}(\mathfrak{F})$. This couple of strategies are implemented in Algorithm 3, which allows random generation of most types of particular mass functions, while autorizing to select the number of focal elements.

5 Conclusion

In this article, we have presented mathematical and algorithmic tools to randomly generate mass functions with controlled statistical distributions. Further works should study various strategies for specific mass functions, with an objective of efficiency from a computational point of view. For instance, the Acceptance-Rejection method may be non optimal, as potential many rejects occur before the acceptance. In addition, we look forward to studying non-uniform sampling on such specific mass fucntions.

References

1. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
2. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)
3. Klir, G.J.: Uncertainty and Information. John Wiley & Sons, Inc., Hoboken (2005)
4. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning (2011) (accepted manuscript) (in press)
5. Shafer, G., Shenoy, P.P.: Local computation on hypertrees. Working paper No. 201, School of Business, University of Kansas (1988)
6. Bauer, M.: Approximation algorithms and decision making in the dempster-shafer theory of evidence - an empirical study. Int. J. Approx. Reasoning 17(2-3), 217–237 (1997)
7. Jousselme, A.-L., Maupin, P.: On some properties of distances in evidence theory. In: Proceedings of the Workshop on Theory of Belief Functions (2010)
8. Cuzzolin, F.: A geometric approach to the theory of evidence. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38(4), 522–534 (2008)

Revisiting the Notion of Conflicting Belief Functions

Sébastien Destercke and Thomas Burger

Abstract. The problem of conflict measurement between information sources knows a regain of interest. In most works related to this issue, Dempter's rule plays a central role. In this paper, we propose to revisit conflict from a different perspective. We do not make a priori assumption about dependencies and start from the definition of conflicting sets, studying its possible extensions to the framework of belief functions.

Keywords: Consistency, Fusion, Contour Function, Dependence.

1 Introduction

In this paper, we revisit the notion of conflict and its quantification in Dempster-Shafer theory (DST), in which it plays an essential role. In particular, its uses in merging rules is the matter of lively debates [1]. Recently, some researchers have questioned the validity of the usual conflict measure (i.e., the mass attributed to the empty set after combination) [2,3]. To solve the issue, they have mostly proposed to complement the usual measure with others. In this work, we take a rather different approach. Two main ideas have motivated this study:

1. First, the idea that conflict between belief functions should be an extension of conflict between sets: when belief functions reduce to sets, the conflict measure should be a binary value that is maximum in case of disjoint sets, minimum otherwise.
2. Second, the idea that conflict between sources should not *a priori* depend on a specific independence assumption between the sources. This is coherent with the *least commitment* principle.

Sebastien Destercke
CNRS, UMR Heudiasyc, Centre de recherche de Royallieu, 60205 COMPIEGNE
e-mail: sebastien.destercke@hds.utc.fr

Thomas Burger
CNRS (FR3425), CEA (iRTSV/BGE), INSERM (EDyP, U1038), Grenoble, France
e-mail: thomas.burger@cea.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 153–160.
springerlink.com

After recalling some basics (Section 2), Section 3 investigates how consistency degree of a single mass assignment can be defined. Then, in Sections 4 and 5, we investigate the case of conflict between sets, and the case of conflict between mass functions. This study leads us to two different propositions of conflict measures, whose differences are briefly discussed in Section 6.

2 Preliminaries

We assume the reader to be familiar with DST [4, 5], and we only present notations and unusual definitions. A *mass assignment* m over Ω is a mapping $m : \wp(\Omega) \to [0,1]$, with $\wp(\Omega)$ the power set of Ω and s.t. $\sum_{A\in\wp(\Omega)} m(A) = 1$. $\mathcal{M}_\Omega$ denote the set of all mass assignments over Ω. A subset $A \subseteq \Omega$ is a *focal element* of m if $m(A) \neq 0$. The set of focal elements of m is noted $\mathcal{F}$. m is *normalised* if $m(\emptyset) = 0$. From m, in addition to the classical *belief*, *plausibility* and *commonality* functions [4], respectively denoted Bel, Pl and Q we use the *contour function* $pl : \Omega \to [0,1]$ of a mass assignment that corresponds to its plausibility on singletons. Recall that m can be associated to a probability set $\mathcal{P}_m := \{Pr(.) \mid \forall A \subseteq \Omega, Bel(A) \leq Pr(A)\}$.

Among the existing interpretations of belief functions, we focus on Shafer's view [4], extensively taken over by Smets in his Transferable Belief Model [5]. In this view, $m(A)$ is the mass of belief exactly committed to the hypothesis $\{\omega_0 \in A\}$, where ω_0 is the true value of an ill-known variable $\mathcal{W}$. A difference between Shafer's view and the TBM is that the latter allows $m(\emptyset) \neq 0$. Note that in the TBM original exposure, $m(\emptyset)$ is not related to conflict itself, but to the open-world assumption in which $m(\emptyset)$ quantifies the belief that the true value does not lie in Ω.

A main source of conflict comes from the conjunctive combination of information coming from not fully agreeing sources. The most classical conjunctive combination is the conjunctive rule [5], or Dempster's [6] unnormalised rule, that assumes that the sources of information are independent. In this paper, we consider a more general framework [7] where other dependency structures are considered. Given two mass assignments m_1 and m_2 defined on Ω, we consider that a conjunctive combination is achieved in two steps:

1. A joint mass assignment $\mathbf{m} : \wp(\Omega) \times \wp(\Omega) \to [0,1]$ is built s.t.

$$\sum_{B\subseteq\Omega} \mathbf{m}(A \times B) = m_1(A)\ ; \ \sum_{A\subseteq\Omega} \mathbf{m}(A \times B) = m_2(B)\ \ \forall A, B \in \wp(\Omega). \quad (1)$$

2. A mass $m_\cap : \wp(\Omega) \to [0,1]$ such that $m_\cap(C) = \sum_{A\cap B=C} \mathbf{m}(A \times B)$.

The joint mass $\mathbf{m}$ encodes the dependence structure between the two sources m_1, m_2. The conjunctive rule, whose result is denoted $m_\oplus$, corresponds to choose $\mathbf{m}(A\times B) = m_1(A)m_2(B)$ in step 1. We denote by $\mathcal{M}_{12}$ the set of all mass $m_\cap$ obtainable by a conjunctive combination of m_1 and m_2. Note that all mass assignments in $\mathcal{M}_{12}$ are specialisations of both m_1 and m_2. Recall that a mass m with $\mathcal{F} = \{E_1, \ldots, E_q\}$ is a specialisation of m' with $\mathcal{F}' = \{E'_1, \ldots, E'_p\}$ if and only if there exists a non-negative matrix $G = [g_{ij}]$ such that for $j = 1, \ldots, p$, $\sum_{i=1}^{q} g_{ij} = 1, g_{ij} > 0 \Rightarrow E_i \subseteq E'_j$, and for $i = 1, \ldots, q$, $\sum_{j=1}^{p} m'(E'_j)g_{ij} = m(E_i)$, where

g_{ij} is the proportion of E'_j that "flows down" to E_i. In other words, m_1 is s-included in m_2 ($m_1 \sqsubseteq_s m_2$) if the mass of any focal element E_j of m_2 can be redistributed among subsets of E_j in m_1. In fact, s-inclusion is a direct extension of the relation of inclusion between sets. As for set inclusion, s-inclusion can therefore be used to compare informative contents, $m_1 \sqsubseteq_s m_2$ meaning that m_1 is less informative than m_2.

3 Consistent Mass Assignments

We first define the notion of consistent set, before extending it to mass assignment. When information is provided as a single set $\omega_0 \in A$, this information is consistent if and only if $A \neq \emptyset$. A can be seen, for instance, as the set of models of a logic base that could be inconsistent. In this case, either a set is consistent (i.e. non-empty) or it is not, and a degree of consistency ϕ can only takes two values. Moreover, it should obey the following properties:

Property 1 (Bounded). *ϕ should be bounded.*

Property 2 (Extreme consistency). *ϕ should be maximal iff information is totally consistent, and minimal iff information is totally inconsistent.*

For simplicity, we assume that the bounds are $[0, 1]$. In the case of sets, we define the consistency degree as $\phi : \wp(\Omega) \rightarrow \{0, 1\}$ such that

$$\phi(A) = 1 \text{ if } A \neq \emptyset, 0 \text{ otherwise} \tag{2}$$

which satisfies Properties 1 and 2. We now extend it to generic mass functions. We consider first extreme cases of totally consistent and totally inconsistent mass functions: It is natural to associate totally inconsistent information with the mass $m(\emptyset) = 1$. On the other hand, the totally consistent information on sets can be extended in two main different ways. A first definition of consistent belief functions (see [7,8]) is the following:

Definition 1. A mass assignment m is said to be *logically consistent* if and only if $\bigcap_{E \in \mathcal{F}} E \neq \emptyset$.

That is, a (normalized) mass m whose focal elements have a non-empty intersection. Next lemma characterizes these masses in terms of contour function.

Lemma 1. $\bigcap_{E \in \mathcal{F}} E \neq \emptyset \quad \Leftrightarrow \quad \exists \omega \in \Omega \text{ s.t. } pl(\omega) = 1$

m is *logically consistent* iff its contour function is normalized. This form of consistency is in accordance with the TBM interpretation, as a source is logically consistent if it considers at least one state of the world to be totally plausible. Among logically consistent mass assignments, *consonant* ones play a particular role, displaying an even stronger form of consistency: the intersection of any two focal sets is still a focal set of this mass assignment (since if $A \subset B$, $A \cap B = A$), which is not the case for general logically consistent mass assignments. The next definition provides a weaker form of consistency:

Definition 2. A mass assignment m is said to be *probabilistically consistent* if and only if $m(\emptyset) = 0$.

The name probabilistic consistency comes from the fact that requiring $m(\emptyset) = 0$ is equivalent to requiring that the probability set $\mathcal{P}_m$ induced by m is non-empty. It is also in accordance with logic-based interpretation of belief functions [9].

Definitions 1 and 2 each suggests a different measure of consistency. The following measures ϕ_{pl}, ϕ_m from $\mathcal{M}_\Omega$ to $[0, 1]$, such that:

$$\phi_{pl}(m) = \max_{\omega \in \Omega} pl(\omega), \tag{3}$$

$$\phi_m(m) = 1 - m(\emptyset) \tag{4}$$

do satisfy Property 2 for totally inconsistent information and for Definitions 1 and 2 of totally consistent information, respectively. When $\exists A \in \Omega / m(A) = 1$, then both ϕ_m and ϕ_{pl} reduce to Eq. (2).

Although Definition 2 and Eq. (4) appear less adapted to the TBM interpretation than Definition 1, we will see in further sections that Eq. (4) can be useful in the TBM interpretation as well. Also, let us note that the inequality $\phi_{pl} \leq \phi_m$ always holds, and $\phi_{pl} = \phi_m$ if and only if $\bigcap_{E \in \mathcal{F} \setminus \emptyset} E \neq \emptyset$. Moreover, for consonant masses ϕ_{pl}, ϕ_m are the consistency degree of possibility theory [10].

4 Conflict between Sets

We can now study conflict between sources, starting with sets. Similar to possibility theory [10], we measure conflict as the inconsistency (inconsistency being the inverse of consistency) resulting from the conjunctive merging of information. Considering two sources of information (extension $N > 2$ is straightforward), we define the conflict of sets as $\kappa : \wp(\Omega) \times \wp(\Omega) \to \{0, 1\}$ embedding the combination step.

In the case of sources assessing that $\omega_0 \in A$ and $\omega_0 \in B$, two extreme cases may occur: they are conflicting ($A \cap B = \emptyset$) or not ($A \cap B \neq \emptyset$). As for the consistency measure, a (bounded) measure of conflict κ should take its maximal / minimal values in such cases, giving

Property 3 (Extreme conflict). *A conflict measure should be maximal value iff sources are totally conflicting, and minimal iff sources are non-conflicting.*

In other words, conflict κ for sets should be such that

$$\kappa(A, B) = 1 - \phi(A \cap B) = 1 \text{ if } A \cap B = \emptyset,\ 0 \text{ otherwise} \tag{5}$$

Other desirable properties may be formulated by observing sets. A first property should be symmetry, as we consider the two sources of equal importance.

Property 4 (Symmetry). *A measure of conflict should be symmetric.*

This translates into $\kappa(A, B) = \kappa(B, A)$. The other properties concern the behaviour of the measure with respect to some changes in the information.

Property 5 (Imprecision monotonicity). *A measure of conflict should be non-increasing if a source becomes less informative.*

If $A \cap B \neq \emptyset$, then considering $A' \supseteq A$ implies $A' \cap B \neq \emptyset$, hence κ should not increase. In contrast, we may have $A \cap B = \emptyset$ but $A' \cap B \neq \emptyset$, in which case κ should decrease. This translates by the constraint $\kappa(A', B) \leq \kappa(A, B)$.

Property 6 (Ignorance is bliss). *A measure of conflict should be insensitive to combination with ignorance.*

If $B = \Omega$, then $A \cap B \neq \emptyset$ unless $A = \emptyset$, and a state of ignorance should not conflict with any information, unless the latter is inconsistent. This translates by the constraint $\kappa(A, \Omega) = 1 - \phi(A)$.

5 Conflict between Mass Assignments

In the case of mass assignments m_1, m_2, the conjunctive combination is no longer unique (Eq. (1)), unless a specific (in)dependence structure is given. In our opinion, conflict measurement should reflect our knowledge of dependence. In particular, $m_\oplus$ **should not be used** to measure conflict, unless independence assumption between sources holds. This results in the following property.

Property 7 (Independence to dependence). *A conflict measure should not depend on a dependence assumption not supported by evidence.*

5.1 Characterising Total Conflict and Conflict Absence

It is natural to say that two sources are totally conflicting if none of their focal elements intersect (i.e., only $\emptyset$ can have positive mass after merging). Let $\mathcal{D}_i = \cup_{A \in \mathcal{F}_i} A$, then

Definition 3. m_1 and m_2 are *totally conflicting* when $D_1 \cap D_2 = \emptyset$.

If $m_1(A) = 1$ and $m_2(B) = 1$, we retrieve the set definition. To extend the notion of non-conflicting sets, we see two main ways fitting the TBM interpretation, given here from the most to the least constraining.

Definition 4. m_1, m_2 are strongly non-conflicting iff $\bigcap_{A \in \mathcal{F}_{m_1} \cup \mathcal{F}_{m_2}} A \neq \emptyset$.

Definition 5. m_1, m_2 are non-conflicting iff $\forall (A, B)$ such that $A \in \mathcal{F}_{m_1}$, $B \in \mathcal{F}_{m_2}$, we have $A \cap B \neq \emptyset$.

Definition 4 requires all focal elements to have a non-empty intersection, and is stronger than requiring that all pairs of focal elements from m_1 and m_2 have a non-empty intersection (Definition 5). If $m_1(A) = 1$ and $m_2(B) = 1$, the two definitions reduce to non-empty intersecting sets. The next proposition shows that strongly non-conflicting masses are related to plausibility measures, hence to consistency given by Eq. (3).

Proposition 1. $\bigcap_{A\in\{\mathcal{F}_{m_1}\cup\mathcal{F}_{m_2}\}} A \neq \emptyset$ *iff* $\forall m_\cap \in \mathcal{M}_{12}, \exists \omega \in \Omega$ *s.t.* $pl_{m_\cap}(\omega) = 1$

This suggests to use the contour function to evaluate the conflict when conflict absence corresponds to Definition 4 (Strong non-conflict). Proposition 1 says that two sources are strongly non-conflicting iff there is at least one state of the world ω that they both consider "normal" or totally plausible. This is in agreement with the TBM interpretation and similar to Daniel [3] proposal. Definition 5, on the other hand, is related to the consistency measure given by Eq. (4) and we have

Proposition 2. $A \cap B \neq \emptyset \ \ \forall A \in \mathcal{F}_{m_1}, \ \forall B \in \mathcal{F}_{m_2}$ *iff* $m_\cap(\emptyset) = 0 \ \forall m_\cap \in \mathcal{M}_{12}$

This suggests to use $m_\cap(\emptyset)$ to measure conflict under Definition 5 (Non-conflict). It is by far the most common value used to estimate conflict between information sources in Dempster-Shafer theory.

5.2 *Measuring Conflict between Mass Assignments*

We now propose different measure of conflicts corresponding to each notion of conflict absence, some of them being imprecise (reflecting a possible lack of knowledge about source dependencies). First, we reformulate some properties of conflict measurement κ in the vocabulary of mass assignments:

- **Prop. 3 (Extreme conflict):** $\kappa(m_1, m_2) = 0$ if and only if m_1 and m_2 are non-conflicting (according to the considered definition);
- **Prop. 4 (Symmetry):** $\kappa(m_1, m_2) = \kappa(m_2, m_1)$;
- **Prop. 5 (Imprecision monotonicity):** if $m_1 \sqsubset_s m_1'$, then $\kappa(m_1', m_2) \leq \kappa(m_1, m_2)$;
- **Prop. 6 (Ignorance is bliss):** if $m_2(\Omega) = 1$, then $\kappa(m_1, m_2) = 1 - \phi(m_1)$;

Measures for strong non-conflict: Given Proposition 1, it is natural to use ϕ_{pl} (Eq. (3)) to measure conflict from strong non-conflict. We propose to distinguish three cases:

- the case where dependence is unknown, and where one accepts imprecise conflict. In this case, if $\mathcal{I}([0,1])$ denote intervals of $[0,1]$, the measure of conflict is an application $\kappa^1_{pl} : \mathcal{M}_\Omega \times \mathcal{M}_\Omega \to \mathcal{I}([0,1])$ such that

$$\begin{aligned}\kappa^1_{pl}(m_1, m_2) &= [\min_{m_\cap\in\mathcal{M}_{12}} 1 - \phi_{pl}(m_\cap), \max_{m_\cap\in\mathcal{M}_{12}} 1 - \phi_{pl}(m_\cap)] \\ &= [\min_{m_\cap\in\mathcal{M}_{12}} 1 - \max_{\omega\in\Omega} pl_\cap(\omega), \max_{m_\cap\in\mathcal{M}_{12}} 1 - \max_{\omega\in\Omega} pl_\cap(\omega)];\end{aligned} \tag{6}$$

- the case where dependence is unknown, but the least commitment principle is followed to get a unique conflict value. In this case, we propose to select the minimal conflicting situation and $\kappa^2_{pl} : \mathcal{M}_\Omega \times \mathcal{M}_\Omega \to [0,1]$ is such that

$$\kappa^2_{pl}(m_1, m_2) = \min_{m_\cap\in\mathcal{M}_{12}} 1 - \phi_{pl}(m_\cap) = \min_{m_\cap\in\mathcal{M}_{12}} 1 - \max_{\omega\in\Omega} pl_\cap(\omega) \tag{7}$$

- the case where dependence is known (i.e., a joint mass $\mathbf{m}$ is specified) and where the result of conjunction is a single $m_\cap$: We propose to simply use

$$\kappa^3_{pl}(m_1, m_2) = 1 - \phi_{pl}(m_\cap) = 1 - \max_{\omega \in \Omega} pl_\cap(\omega) \tag{8}$$

They all satisfy properties 3- 6, and can deal with unknown dependence. Note that both κ^3_{pl} and κ^2_{pl} are straightforward to compute (the latter using results from [7]), and only the upper bound of κ^1_{pl} requires the use of linear programming techniques.

Measures for non-conflict: As Proposition 2 is linked to Definition 2, we use ϕ_m (Eq.(4)) to derive three measures under non-conflict:

$$\kappa^1_m(m_1, m_2) = [\min_{m_\cap \in \mathcal{M}_{12}} 1 - \phi_m(m_\cap), \max_{m_\cap \in \mathcal{M}_{12}} 1 - \phi_m(m_\cap)] \tag{9}$$

$$\kappa^2_m(m_1, m_2) = \min_{m_\cap \in \mathcal{M}_{12}} 1 - \phi_m(m_\cap) = \min_{m_\cap \in \mathcal{M}_{12}} m_\cap(\emptyset) \tag{10}$$

$$\kappa^3_m(m_1, m_2) = 1 - \phi_m(m_\cap) = m_\cap(\emptyset) \tag{11}$$

$\kappa^1_m(m_1, m_2)$, $\kappa^2_m(m_1, m_2)$ corresponding to unknown dependence (without and with least commitment principle, respectively) and $\kappa^3_m(m_1, m_2)$ corresponding to known dependence. They all satisfy properties 3- 6 and can deal with unknown dependence. Classical conflict measure $m_\oplus(\emptyset)$ is captured by $\kappa^3_m(m_1, m_2)$ when independence between sources can be assumed. Computing the two bounds of κ^1_m require the use of linear programs, while κ^3_m remains straightforward to evaluate.

6 Short Exemplified Discussion

Let us take two different examples, showing that the proposed measures of conflict behave differently, and each have their own interest.

First, let us consider m_1, m_2 on $\Omega = \{\omega_1, \omega_2, \omega_3\}$ such that $m_1(\{\omega_1, \omega_2\}) = 0.6$, $m_1(\{\omega_1, \omega_3\}) = 0.4$ and $m_2(\{\omega_2, \omega_3\}) = 0.5$, $m_2(\Omega) = 0.5$. Both are logically and probabilistically consistent, and we have $\kappa^1_{pl}(m_1, m_2) = [0.4, 0.4] = 0.4$ while $\kappa^1_m(m_1, m_2) = [0, 0] = 0$. According to the measure based on the contour functions, there is some conflict, whereas according to the one based on $m(\emptyset)$ there is not. While each source is consistent, they disagree on which state of the world is the most plausible (ω_1 for m_1 and ω_2 or ω_3 for m_2). Hence, in some sense (meaningful in a TBM interpretation), the two sources can be considered as conflicting. Clearly, only the measure based on contour functions is able to detect it.

As a second example, consider two identical masses on $\Omega = \{\omega_1, \omega_2\}$ such that $m_1(\{\omega_1\}) = m_2(\{\omega_1\}) = 0.5$ and $m_1(\{\omega_2\}) = m_2(\{\omega_1\}) = 0.5$. First, note that $\phi_{pl}(m_i) = 0.5$ for $i = 1, 2$, a rather low score indicating some internal inconsistency for each source. Also, the conflict measures are $\kappa^1_{pl}(m_1, m_2) = [0.5, 1]$ and $\kappa^1_m(m_1, m_2) = [0, 1]$. The highest and lowest conflict value being obtained for the combination $\mathbf{m}(\omega_1 \times \omega_2) = 0.5$ and $\mathbf{m}(\omega_2 \times \omega_1) = 0.5$ and for the combination $\mathbf{m}(\omega_1 \times \omega_1) = 0.5$ and $\mathbf{m}(\omega_2 \times \omega_2) = 0.5$ (idempotent merging), respectively. Note that every possible dependency between these extremes may be considered. This example shows that some conflict is generated from the combination, but that contour-function based measures tend to mix it with some initial inconsistency, while κ_m does detect that sources can totally agree in case of dependence. Hence, contrarily to the first example, here, measures based on $m(\emptyset)$ provide some interesting information which are not captured by measures based on contour functions. This

short discussion shows that the measures have different behaviors, and that an extended discussion would be interesting. A first quick conclusion is that $m(\emptyset)$ based measures identify conflict arising from combination only, while contour-function based measures also capture some internal inconsistency. Hence, $m(\emptyset)$ seems better fitted to measure conflict **between** sources.

7 Conclusion

We have considered conflict as the inconsistency resulting from conjunctive combination. Starting from sets, we have derived a number of results regarding consistency and conflict on mass assignments. Then, we have proposed several conflict measurements not relying on Dempster's rule and able to cope with unknown (or partially known) dependencies. Our findings show that using the contour function may be a better conflict measure within the TBM interpretation, but that using $m(\emptyset)$ may be useful to characterise conflict between mass assignments.

The next step is to relate this study with other works. For instance, how it can be used to differentiate between internal and external conflict [3]. Our approach should also be compared to conflict measurements based on distances [2,11], however we can already notice that dissimilarities based on distances do not generally satisfied the properties required here (e.g., Prop. 3 and 5), hence the two approaches are likely to give different conclusions in some cases.

References

1. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion 8, 387–412 (2006)
2. Liu, W.: Analyzing the degree of conflict among belief functions. Artif. Intell. 170(11), 909–924 (2006)
3. Daniel, M.: Conflicts within and between Belief Functions. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 696–705. Springer, Heidelberg (2010)
4. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
5. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)
6. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
7. Destercke, S., Dubois, D.: Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory. Information Sciences 181(18), 3925 (2011), doi:10.1016/j.ins.2011.05.007
8. Cuzzolin, F.: On Consistent Approximations of Belief Functions in the Mass Space. In: Liu, W. (ed.) ECSQARU 2011. LNCS, vol. 6717, pp. 287–298. Springer, Heidelberg (2011)
9. Cattaneo, M.: Combining belief functions issued from dependent sources. In: Proc. Third International Symposium on Imprecise Probabilities and Their Application (ISIPTA 2003), Lugano, Switzerland, pp. 133–147 (2003)
10. Dubois, D., Prade, H.: Possibility theory and data fusion in poorly informed environments. Control Eng. Practice 2, 811–823 (1994)
11. Martin, A., Jousselme, A.-L., Osswald, C.: Conflict measure for the discounting operation on belief functions. In: The 11th International Conference on Information Fusion, pp. 1003–1010 (2008)

About Conflict in the Theory of Belief Functions

Arnaud Martin

Abstract. In the theory of belief functions, the conflict is an important concept. Indeed, combining several imperfect experts or sources allows conflict. However, the mass appearing on the empty set during the conjunctive combination rule is generally considered as conflict, but that is not really a conflict. Some measures of conflict have been proposed, we recall some of them and we show some counter-intuitive examples with these measures. Therefore we define a conflict measure based on expected properties. This conflict measure is build from the distance-based conflict measure weighted by a degree of inclusion introduced in this paper.

1 Introduction

The theory of belief functions was first introduced by [2] in order to represent some imprecise probabilities with *upper* and *lower probabilities*. Then [13] proposed a mathematical theory of evidence with is now widely used for information fusion. Combining imperfect sources of information leads inevitably to conflict. One can consider that the conflict comes from the non-reliability of the sources or the sources do not give information on the same observation. In this last case, one must not combine them.

Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ be a frame of discernment of exclusive and exhaustive hypothesis. A mass function m is the mapping from elements of the power set 2^Θ onto $[0,1]$ such that:

$$\sum_{X \in 2^\Theta} m(X) = 1. \tag{1}$$

Arnaud Martin
University of Rennes 1, IRISA, rue E. Branly, 22300 Lannion
e-mail: Arnaud.Martin@univ-rennes1.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 161–168.
springerlink.com

A focal element X is an element of 2^Θ such that $m(X) \neq 0$. If the focal elements are nested, the mass functions is *consonant*. Constraining $m(\emptyset) = 0$ corresponds to a closed-world assumption [13], while allowing $m(\emptyset) \geq 0$ corresponds to an open world assumption [15]. Smets interprets this mass on the empty set such as an non-expected hypothesis and normalizes it in the pignistic probability defined for all $X \in 2^\Theta$, with $X \neq \emptyset$ by:

$$\text{BetP}(X) = \sum_{Y \in 2^\Theta, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m(Y)}{1 - m(\emptyset)}. \tag{2}$$

The first combination rule has been proposed by Dempster [2] and is defined for two mass functions m_1 and m_2, for all $X \in 2^\Theta$, with $X \neq \emptyset$ by:

$$m_{\text{DS}}(X) = \frac{1}{1-k} \sum_{A \cap B = X} m_1(A) m_2(B), \tag{3}$$

where $k = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$ is the inconsistence of the combination and generally called conflict. We call it here the *global conflict*.

To stay in an open world, Smets [15] proposes the non-normalized conjunctive rule given for two mass functions m_1 and m_2 and for all $X \in 2^\Theta$ by:

$$m_{\text{Conj}}(X) = \sum_{A \cap B = X} m_1(A) m_2(B) := (m_1 \textcircled{\cap} m_2)(X). \tag{4}$$

These both rules allow to reduce the imprecision of the focal elements and to increase the belief on concordant elements. The main assumptions to apply these rules are the cognitive independence and the reliability of the sources.

Based on the results of these rules, the problem enlightened by the famous Zadeh's example [20] is the repartition of the global conflict. Indeed, consider $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and two experts opinions given by $m_1(\theta_1) = 0.9$, $m_1(\theta_3) = 0.1$, and $m_2(\theta_2) = 0.9$, $m_1(\theta_3) = 0.1$, the mass function resulting in the combination using Dempster's rule is $m(\theta_3) = 1$ and using conjunctive rule is $m(\emptyset) = 0.99$, $m(\theta_3) = 0.01$. Therefore, several combination rules have been proposed to manage this global conflict [16, 9].

As observed in [8, 10], the weight of conflict given by $k = m_{\text{Conj}}(\emptyset)$ is not a conflict measure between the mass functions. Indeed, the conjunctive-based rules are not idempotent (as the majority of the rules defined to manage the global conflict): the combination of identical mass functions leads generally to a positive value of k. Hence, new kind of conflict measures are defined in [10].

In the following section 2, we recall the different measures of conflict in the theory of belief functions. Then, on the bases of wanted properties we propose a new conflict measure based on a degree of inclusion that we define in section 3. The last section 4 presents the interest of the proposed conflict measures on numerical example and gives uses of this measure.

2 Conflict Measures

First of all, we should not mix up conflict measure and contradiction measure. The measures defined by [7, 17] are not conflict measures, but some discord and specificity measures (to take the terms of [6]) we call contradiction measures. We define the contradiction and conflict measures by the following definitions:

Definition *A* contradiction *in the theory of belief functions quantifies how a mass function contradicts itself.*

Definition (C1) *The* conflict *in the theory of belief functions can be defined by the contradiction between two or more mass functions.*

Therefore, is the mass on the empty set or the functions of this mass (such as $-\ln(1-m_{\text{Conj}}(\emptyset))$ proposed by [13]) a conflict measure? It seems obvious that the property of the non-idempotence is a problem to use this as a conflict measure. However, if we define a conflict measure such as $\text{Conf}(m_1,m_2)=m_{\text{Conj}}(\emptyset)$, we note that $\text{Conf}(m_1,m_\Theta)=0$ where $m_\Theta(\Theta)=1$ is the ignorance. Indeed, the ignorance is the neutral element for the conjunctive combination rule. This propertiy seems to be reached by a conflict measure.

Other conflict measures have been defined. In [5], a conflict measure is given by:

$$\text{Conf}(m_1,m_2)=1-\frac{\mathbf{pl}_1^T.\mathbf{pl}_2}{\|\mathbf{pl}_1\|\|\mathbf{pl}_2\|} \tag{5}$$

where $\mathbf{pl}$ is the plausivity function and $\mathbf{pl}_1^T.\mathbf{pl}_2$ the vector product in 2^n space of both plausibility functions. However, generally $\text{Conf}(m_1,m_\Theta)\neq 0$, that seems couter-intuitive.

Auto-conflict

Introduced by [11], the auto-conflict of order s for one expert is given by:

$$a_s=\left(\bigcirc\!\!\!\!\!\!\cap_{i=1}^{s} m\right)(\emptyset). \tag{6}$$

where $\cap\!\!\!\!\bigcirc$ is the conjunctive operator of Equation (4). The following property holds: $a_s\leq a_{s+1}$, meaning that due to the non-indempotence of $\cap\!\!\!\!\bigcirc$, the more m is combined with itself the nearer to 1 k is, and so in a general case, the more the number of experts is high the nearer to 1 k is. The behavior of the auto-conflict was studied in [10] and show that we should take into account the auto-conflict in the global conflict in order to really define a conflict. In [19], the auto-conflict was defined and called the plausibility of the belief structure with itself. The auto-conflict is a kind of measure of the contradiction, but depends on the order. A measure of contradiction independent on the order has been defined in [14].

Conflict Measure Based on a Distance

The definition of the conflict (C1) involves firstly to measure it on the bba's space and secondly that if the opinions of two experts are far from each other, we consider

that they are in conflict. That suggests a notion of distance. That is the reason why in [10], we give a definition of the measure of conflict between experts assertions through a distance between their respective bba's. The conflict measure between 2 experts is defined by:

$$\text{Conf}(1,2) = d(m_1, m_2). \tag{7}$$

We defined the conflict measure between one expert i and the other $M-1$ experts by:

$$\text{Conf}(i, \mathscr{E}) = \frac{1}{M-1} \sum_{j=1, i \neq j}^{M} \text{Conf}(i, j), \tag{8}$$

where $\mathscr{E} = \{1, \ldots, M\}$ is the set of experts in conflict with i. Another definition is given by:

$$\text{Conf}(i, M) = d(m_i, \overline{m_M}), \tag{9}$$

where $\overline{m_M}$ is the bba of the artificial expert representing the combined opinions of all the experts in $\mathscr{E}$ except i.

We use the distance defined in [3], which is for us the most appropriate. See [4] for a comparison of distances in the theory of belief functions. This distance is defined for two basic belief assignments $\mathbf{m}_1$ and $\mathbf{m}_2$ on 2^Θ by:

$$d(m_1, m_2) = \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T \underline{\underline{D}} (\mathbf{m}_1 - \mathbf{m}_2)}, \tag{10}$$

where $\underline{\underline{D}}$ is an $2^{|\Theta|} \times 2^{|\Theta|}$ matrix based on Jaccard distance whose elements are:

$$D(A,B) = \begin{cases} 1, \text{if} A = B = \emptyset, \\ \\ \dfrac{|A \cap B|}{|A \cup B|}, \forall A, B \in 2^\Theta. \end{cases} \tag{11}$$

This measure is called a *total conflict* measure. An interesting property of the total conflict is given by $\text{Conf}(m,m) = 0$. That means that there is no conflict between a source and itself (that is not a contradiction). However, we generally do not have $\text{Conf}(m, m_\Theta) = 0$, where $m_\Theta(\Theta) = 1$ is the ignorance.

3 Towards Efficient Conflict Measures

We have seen that we cannot use the mass on the empty set as a conflict measure because of the non-idempotence of the conjunctive rule. We also have seen that the conflict measure based on the distance is not null in general for the ignorance mass.

The conjunctive rule does not transfer mass on the empty set if the mass functions are *included*.

Definition *We say that the mass function m_1 is* included *in m_2 if all the focal elements of m_1 are included in all focal elements of m_2. We note this inclusion by $m_1 \subseteq m_2$. The mass functions are* included *if m_1 is included in m_2 or m_2 is included in m_1.*

Therefore these two conflict measures have not an intuitive and expected behavior. Hereafter, we define a new conflict measure having expected properties presented in the following axioms.

Axioms

Let note $\text{Conf}(m_1, m_2)$ a conflict measure between the mass functions m_1 and m_2. We present hereafter essential properties that must verify a conflict measure.

1. Non-negativity: $\text{Conf}(m_1, m_2) \geq 0$
   A negative conflict does not make sens. This axiom is for us necessary.
2. Identity: $\text{Conf}(m_1, m_1) = 0$
   Two equal mass functions are not in conflict. This property is not reached by the global conflict, but seems natural.
3. Symmetry: $\text{Conf}(m_1, m_2) = \text{Conf}(m_2, m_1)$
   The conflict measure must be symmetric. We do not see any case where the non-symmetry can make sens.
4. Normalization: $0 \leq \text{Conf}(m_1, m_2) \leq 1$
   This axiom is may be not necessary to define a conflict measure, but the normalization is very useful in many applications of conflict measure.
5. Inclusion: $\text{Conf}(m_1, m_2) = 0,$ iif $m_1 \subseteq m_2$ or $m_2 \subseteq m_1$
   This axiom means if the focal elements of two mass functions are not conflicting (the intersection is never empty), the mass functions are not in conflict and the mass functions cannot be in conflict if they are included. This property is not reached by a distance based conflict measure.

If a conflict measure verifies these axioms that is not necessary a distance. Indeed, we only impose the identity and not the definiteness ($\text{Conf}(m_1, m_2) = 0 \Leftrightarrow m_1 = m_2$). The axiom of inclusion is less restrictive and make more sens for a conflict measure. Moreover, we do not impose the triangle inequality ($\text{Conf}(m_1, m_2) \leq \text{Conf}(m_1, m_3) + \text{Conf}(m_3, m_2)$). It can be interesting to have $\text{Conf}(m_1, m_2) \geq \text{Conf}(m_1, m_3) + \text{Conf}(m_3, m_2)$ meaning that an expert given the mass function m_3 can reduce the conflict. He reach a kind of consensus. Therefore, a distance cannot be used directly to define a conflict measure as before.

Degree of Inclusion

We see that the axiom of inclusion seems very important to define a conflict measure. This is the reason why we define here a degree of inclusion measuring how two mass functions are included. Let the inclusion index: $Inc(X_1, Y_2) = 1$ if $X_1 \subseteq Y_2$ and 0 otherwise, where X_1 and Y_2 are two focal elements of m_1 and m_2 respectively.

Let $d_{inc}(m_1,m_2)$ a degree of inclusion of m_1 **in** m_2. We can define it by:

$$d_{inc}(m_1,m_2) = \frac{1}{|\mathscr{F}_1||\mathscr{F}_2|} \sum_{X_1 \in \mathscr{F}_1} \sum_{Y_2 \in \mathscr{F}_2} Inc(X_1,Y_2) \tag{12}$$

where $\mathscr{F}_1$ and $\mathscr{F}_2$ are the set of focal elements of m_1 and m_2 respectively, and $|\mathscr{F}_1|$, $|\mathscr{F}_2|$ are the number of focal elements of m_1 and m_2.

Let $\delta_{inc}(m_1,m_2)$ a degree of inclusion of m_1 **and** m_2 define by:

$$\delta_{inc}(m_1,m_2) = \max(d_{inc}(m_1,m_2), d_{inc}(m_2,m_1)) \tag{13}$$

This degree gives the maximum of the proportion of focal elements from one mass function included in another one. Therefore, $\delta_{inc}(m_1,m_2) = 1$ if and only if m_1 and m_2 are included, and the axiom of inclusion is reached for $1 - \delta_{inc}(m_1,m_2)$.

A Conflict Measure

We define a conflict measure between two mass functions m_1 and m_2 by:

$$\text{Conf}(m_1,m_2) = (1 - \delta_{inc}(m_1,m_2))d(m_1,m_2) \tag{14}$$

where d is the distance defined by the equation (10). All the previous axioms are reached. Indeed the axiom of inclusion is reached by $1 - \delta_{inc}(m_1,m_2)$ and the distance d verify the other axioms. Moreover $0 \leq \delta_{inc}(m_1,m_2) \leq 1$, by the product of $1 - \delta_{inc}$ and d, all the axioms are verified.

For more than two mass functions, the conflict measure between one expert i and the other $M-1$ experts can be defined from the equations (8) or (9).

4 Illustration

Comportment of the Proposed Conflict Measure

We can first note $\text{Conf}(m_1,m_1) = 0$ and $\text{Conf}(m_1,m_\Theta) = 0$ as expected. We have even: if m_1 and m_2 are included then $\text{Conf}(m_1,m_2) = 0$, because the degree of inclusion gives the axiom of inclusion. For example, let's consider:

$$m_1(\theta_1) = m_1(\theta_2) = m_1(\theta_1 \cup \theta_2) = 1/3, \tag{15}$$

$$m_2(\theta_1 \cup \theta_2) = m_2(\theta_1 \cup \theta_2 \cup \theta_3) = 1/2. \tag{16}$$

On this example, $d(m_1,m_2) = 0.3727$. $d_{inc}(m_1,m_2) = 1$ and $d_{inc}(m_2,m_1) = 0.17$, therefore $\delta_{inc}(m_1,m_2) = 1$ and $\text{Conf}(m_1,m_2) = 0$

Note we have $d_{inc}(m_1,m_1) = 0.56$ and $d_{inc}(m_2,m_2) = 0.75$, we only have $d_{inc}(m,m) = 1$ if m is categoric ($m(X) = 1, X \in 2^\Theta$).

To illustrate the comportment of the proposed conflict measure we consider:

$$m_3(\theta_3) = m_3(\theta_1 \cup \theta_2 \cup \theta_3) = 0.5. \tag{17}$$

We have $d_{inc}(m_1,m_3) = d_{inc}(m_2,m_3) = 0.5$, but $\text{Conf}(m_1,m_3) = 0.3815$ and $\text{Conf}(m_2,m_3) = 0.3571$. Hence, we obtain: $\text{Conf}(m_1,m_3) \geq \text{Conf}(m_1,m_2) + \text{Conf}(m_2,m_3)$. m_2 reduce the conflict between m_1 and m_3. If we consider two categorical mass functions such as $m_4(\theta_1) = 1$, $m_5(\theta_2) = 1$ we obtain the maximum of the conflict measure: $\text{Conf}(m_4,m_5) = 1$. That means the most conflicting mass functions are two different categorical mass functions.

On the Use of Conflict Measures

The role of conflict is essential in information fusion. Different ways can be use to manage and reduce the conflict. The conflict can come from the low reliability of the sources. Therefore, we can use this conflict to estimate the reliability of the sources if we cannot learn it on databases as proposed in [10]. Hence, we reduce the conflict before the combination, but we can also directly manage the conflict in the rule of combination as generally made in the theory of belief functions such as explained in [16, 9]. The proposed conflict measure could also use to define combination rules.

According to the application, we do not search always to reduce the conflict. For example, we can use the conflict measure such as an indicator for example in databases [1]. Conflict information can also be an interesting information in some applications such as presented in [12].

5 Conclusion

We propose in this paper an analysis of existing conflict measure. On the base of the drawbacks of these measures, we propose a conflict measure in order to outperform existing ones. This measure is based on the definition of a degree of inclusion. This degree is introduced here in order to quantify how the focal elements of two mass functions are included together. Indeed, we can consider that two mass functions are not in conflict if its are included. The proposed conflict measure, based on five axioms, is then the product of this degree of inclusion and a distance between two mass functions. We see for example this conflict measure can be use to reduce the conflict before or in the combination or as enrichment in databases.

References

1. Chebbah, M., Ben Yaghlane, B., Martin, A.: Reliability estimation based on conflict for evidential database enrichment. In: Belief, Brest, France (2010)
2. Dempster, A.P.: Upper and Lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 83, 325–339 (1967)
3. Jousselme, A.-L., Grenier, D., Bossé, E.: A new distance between two bodies of evidence. Information Fusion 2, 91–101 (2001)
4. Jousselme, A.-L., Maupin, P.: On some properties of distances in evidence theory. In: Belief, Brest, France (2010)
5. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning (2011)

6. Klir, G.J.: Measures of uncertainty in the Dempster-Shafer theory of evidence. In: Yager, R.R., Fedrizzi, M., Kacprzyk, J. (eds.) Advances in the Dempster-Shafer Theory of Evidence, pp. 35–49. John Wiley and Sons, New York (1994)
7. George, T., Pal, N.R.: Quantification of conflict in Dempster-Shafer framework: a new approach. International Journal of General Systems 24(4), 407–423 (1996)
8. Liu, W.: Analyzing the degree of conflict among belief functions. Artificial Intelligence 170, 909–924 (2006)
9. Martin, A., Osswald, C.: Toward a combination rule to deal with partial conflict and specificity in belief functions theory. In: International Conference on Information Fusion, Québec, Canada (2007)
10. Martin, A., Jousselme, A.-L., Osswald, C.: Conflict measure for the discounting operation on belief functions. In: International Conference on Information Fusion, Cologne, Germany (2008)
11. Osswald, C., Martin, A.: Understanding the large family of Dempster-Shafer theory's fusion operators - a decision-based measure. In: International Conference on Information Fusion, Florence, Italy (2006)
12. Rominger, C., Martin, A.: Using the conflict: An application to sonar image registration. In: Belief, Brest, France (2010)
13. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
14. Smarandache, F., Martin, A., Osswald, C.: Contradiction measures and specificity degrees of basic belief assignments. In: International Conference on Information Fusion, Boston, USA (2011)
15. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. Uncertainty in Artificial Intelligence 5, 29–39 (1990)
16. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion 8(4), 387–412 (2007)
17. Wierman, M.J.: Measuring Conflict in Evidence Theory. In: IFSA World Congress and 20th NAFIPS International Conference, vol. 3(21), pp. 1741–1745 (2001)
18. Yager, R.R.: Entropy and Specificity in a Mathematical Theory of Evidence. International Journal of General Systems 9, 249–260 (1983)
19. Yager, R.R.: On Considerations of Credibility of Evidence International. Journal of Approximate Reasoning 7, 45–72 (1992)
20. Zadeh, L.A.: A mathematical theory of evidence (book review). AI Magazine 5, 81–83 (1984)

The Internal Conflict of a Belief Function*

Johan Schubert

Abstract. In this paper we define and derive an internal conflict of a belief function We decompose the belief function in question into a set of generalized simple support functions (GSSFs). Removing the single GSSF supporting the empty set we obtain the base of the belief function as the remaining GSSFs. Combining all GSSFs of the base set, we obtain a base belief function by definition. We define the conflict in Dempster's rule of the combination of the base set as the internal conflict of the belief function. Previously the conflict of Dempster's rule has been used as a distance measure only between consonant belief functions on a conceptual level modeling the disagreement between two sources. Using the internal conflict of a belief function we are able to extend this also to non-consonant belief functions.

1 Introduction

In this paper we define and derive an internal conflict of a belief function within Dempster-Shafer theory [1–3, 14]. We decompose the belief function in question into a set of generalized simple support functions (GSSFs). Removing the single GSSF supporting the empty set we obtain the base of the belief function as the remaining GSSFs. Combining all GSSFs of the base set, we obtain a base belief function by definition. We define the conflict in Dempster's rule of this combination as the internal conflict of the belief function. We propose that the base belief function is a better measure than the original belief function which can be obtained by combining the base belief function with pure conflict, i.e., $\{[m_1(\varnothing), \varnothing],\ [m_1(\Theta), \Theta]\}$.

There are several different ways to manage a high conflict in combination of belief functions within Dempster-Shafer theory. For an overview of different

Johan Schubert
Department of Decision Support Systems, Division of Information and Aeronautical Systems, Swedish Defence Research Agency, SE-164 90 Stockholm, Sweden
e-mail: schubert@foi.se,
http://www.foi.se/fusion

* This work was supported by the FOI research project "Real-time Simulation Supporting Effects-based Planning", which is funded by the R&D programme of the Swedish Armed Forces.

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 169–177.
springerlink.com

alternatives to manage the combination of conflicting belief functions, see articles by Smets [16] and Liu [7]. For a recent survey of alternative distance between belief functions, see Jousselme and Maupin [5].

In section 2 we review a method for decomposing a belief function into a set of GSSFs [15]. In section 3 we derive the base set of a belief function and construct a base belief function from the base set corresponding to the belief function under decomposition. In section 4 we derive the internal conflict of the belief function and show how this extends the conflict from being a distance measure only for consonant belief functions to a functioning distance measure also between non-consonant belief functions. In section 5 we provide an example. Finally, conclusions are drawn (section 6).

2 Decomposing a Belief Function

All belief functions can be decomposed into a set of GSSFs on a frame of discernment Θ using the method developed by Smets [15]. A GSSF is either a traditional simple support function (SSF) [14] or an inverse simple support function (ISSF) [15]. Let us begin by defining an ISSF:

Definition 1. An inverse simple support function on a frame of discernment Θ is a function $m: 2^{\Theta} \rightarrow (-\infty, \infty)$ characterized by a weight $w \in (1, \infty)$ and a focal element $A \subseteq \Theta$, such that $m(\Theta) = w$, $m(A) = 1 - w$ and $m(X) = 0$ when $X \notin \{A, \Theta\}$.

Let us recall the meaning of SSFs and ISSFs [15]: An SSF $m_1(A) \in [0, 1]$ represents a state of belief that "You have some reason to believe that the actual world is in A (and nothing more)". An ISSF $m_2(A) \in (-\infty, 0)$ on the other hand, represents a state of belief that "You have some reason *not* to believe that the actual world is in A". Note that *not* believing in A is different than believing in A^c.

A simple example is one SSF $m_1(A) = 1/4$ and $m_1(\Theta) = 3/4$, and one ISSF $m_2(A) = -1/3$ and $m_2(\Theta) = 4/3$. Combining these two functions yields a vacuous belief function $m_{1 \oplus 2}(\Theta) = 1$.

The decomposition method is performed in two steps eqs. (1) and (2). First, for any non-dogmatic belief function Bel_0, i.e., where $m_0(\Theta) > 0$, calculate the commonality number for all focal elements A by eq. (1). We have

$$Q_0(A) = \sum_{B \supseteq A} m_0(B) \tag{1}$$

For dogmatic belief functions assign $m_0(\Theta) = \varepsilon > 0$ and discount all other focal elements proportionally.

Secondly, calculate $m_i(C)$ for all decomposed GSSFs, where $C \subseteq \Theta$ including $C = \varnothing$, and i is the ith GSSF. There will be one GSSF for each subset C

of the frame unless $m_i(C)$ happens to be zero. In the general case we will have $\left|2^{\Theta}\right|$ GSSFs. We get for all $C \subseteq \Theta$ including $C = \varnothing$

$$\begin{aligned} m_i(C) &= 1 - \prod_{A \supseteq C} Q_0(A)^{(-1)^{|A|-|C|+1}} \\ m_i(\Theta) &= 1 - m_i(C) \end{aligned} \tag{2}$$

where $i \in [1, 2^{|\Theta|} - 1]$.

Here, $C \subseteq \Theta$ of $m_i(C)$ is the ith subset of Θ in numerical order[2] which also includes $C = \varnothing$, i.e., $\{[m_1(\varnothing), \varnothing],\ [m_1(\Theta), \Theta]\}$ is the first decomposed GSSF of eq. (2).

3 Transforming a Belief Function into a Base Belief Function

Using eqs. (1) and (2) we may decompose a belief function m_0 into a set of GSSFs. We call the non-conflict GSSFs of the decomposition the base of m_0.

Definition 2. The base of a belief function m_0 is the set of decomposed simple support and inverse simple support function

$$\{m_i\}_{i=2}^{|2^{\Theta}|-1} \tag{3}$$

deliberately excluding m_1 that supports only $\{\varnothing, \Theta\}$, where

$$\{m_i\}_{i=1}^{|2^{\Theta}|-1} \tag{4}$$

is the full set of $\left|2^{\Theta}\right| - 1$ simple support and inverse simple support function from the decomposition of m_0 by eqs. (1) and (2).

Definition 3. A base belief function m_{00} of a belief function m_0 is the belief function resulting from the unnormalized combination of the base of m_0, i.e.,

$$m_{00} = \bigcirc\!\!\!\!\!\!\bigcirc \{m_i\}_{i=2}^{|2^{\Theta}|-1}. \tag{5}$$

Definition 4. A base conflict of a belief function m_0 is the obtained conflict $m_1(\varnothing)$ of the first GSSF that supports $\{[m_1(\varnothing), \varnothing],\ [m_1(\Theta), \Theta]\}$ of the decomposition of a belief function of m_0 by eqs. (1) and (2).

---

[2] With $\Theta = \{a, b, c\}$ the numerical order of all subsets in Θ including $\varnothing$ is $\Theta = \{\varnothing, \{a\}, \{b\}, \{a, b\}, \{c\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.

Theorem 1. *A belief function m_0 can be recovered by combination of its corresponding base belief function m_{00} with the base conflict m_1, i.e.,*

$$m_0 = m_{00} \textcircled{\cap} m_1 \tag{6}$$

Proof. Immediate by definition 2 and 3. □

4 The Internal Conflict of a Belief Function

The conflict received from a combination of belief functions by Dempster's rule is not a measure of dissimilarity between the combined belief functions. Indeed, belief functions can be quite different and yet have zero conflict as intersection of their focal elements are non-empty. Instead the conflict of Dempster's rule is best viewed as a different kind of distance measure; a measure of conceptual disagreement between sources. When they disagree highly it is a sign that something is wrong. It should be noted that there is at least two possible sources of conflict other than measurement errors. We may have modeling errors or faulty sources [4]. Faulty sources are corrected by appropriate discounting (e.g., [6, 12, 16]) while modeling errors are corrected by adopting an appropriate frame of discernment [13, 14].

In this section we define and investigate an internal conflict of a belief function. We further devise a way to obtain the internal conflict.

Definition 5. The internal conflict of a belief function m_0 is the conflict received in the unnormalized combination of the base of m_0 to obtain the base belief function m_{00}, i.e., $m_{00}(\varnothing)$ where

$$m_{00}(\varnothing) = \textcircled{\cap}\{m_i\}_{i=2}^{|2^\Theta|-1}(\varnothing)\,. \tag{7}$$

For simplicity, view the intersection of eq. (7) as taking place in a $|2^\Theta|-2$ hyper cube of all $|2^\Theta|-2$ GSSFs. Note that $m_{00}(\varnothing)$ can take both positive and negative values.

Theorem 2. The internal conflict within a base belief function is strictly a function of conflicts between different GSSFs supporting subsets of the frame.

Proof. Immediate by observation of the combination in eq. (7) as m_1 with body of evidence $\{[m_1(\varnothing), \varnothing],\ [m_1(\Theta), \Theta]\}$ is not included in the combination. □

Theorem 3. There exist an infinite size family of unnormalized belief functions $\{\, m_0^p \,|\, p = m_1(\varnothing) \in (-\infty, \infty),\ p \neq 1 \,\}$ with an identical base belief function m_{00} and identical internal conflict.

Proof. Let us generate a family from the base: Take any belief function m_0 on a frame Θ of size $n = |\Theta|$. Decompose m_0 into its 2^n GSSFs. Combine $\{m_i\}_{i=2}^{2^n}$ using eq. (7) into the base belief function. Let us ignore the value obtained for $m_1(\varnothing)$ in the decomposition. Instead, let $m_1(\varnothing) \in (-\infty, \infty)$, $m_1(\varnothing) \neq 1$. The family of belief functions $\{m_0^p\}$ is generated by combining the base belief function m_{00} with each of the $\{m_1\}$. The family is of infinite size. □

When going in the other direction from family to base: Note, that in the special case of a normalized base belief function it can be recovered from any family member by normalization.

If we combine a non-consonant belief function with itself we should not be surprised that we receive a conflict. A non-consonant belief function can be expressed as a construct from the base set of that belief function. If the belief function combined with itself is constructed in two steps by first combining all GSSFs pairwise with themselves, these $\left|2^\Theta\right| - 2$ combinations of GSSFs with identical focal sets are conflict free (excluding $\varnothing$ and Θ). Combining the resulting $\left|2^\Theta\right| - 2$ GSSFs in the second step obviously has empty intersections among their focal elements resulting in conflict. Thus, the internal conflict received is a function of conflicts from different GSSFs (excluding $m_1(\varnothing)$) that are used to construct the non-consonant belief function. Thus, the conflict noticed in the combination exists internally within non-consonant belief functions before combination and is a consequence of the scattering of mass within the distribution. This makes the internal conflict appropriate as a conceptual distance measure also between non-consonant belief functions as it measures the internal conflict in the combination of GSSFs from two different base sets corresponding to the two different base belief functions without the added pure conflict of m_1 (supporting only $\varnothing$ and Θ) that is always included in the conflict obtained by Dempster's rule.

Thus, from theorem 2 and 3 follows that the internal conflict is an appropriate distance measure for all belief functions as it excludes the pure conflict of $m_1(\varnothing)$ (i.e., also for non-consonant belief functions), where this distance measure sought after is a measure of conceptual disagreement between sources.

When calculating the conceptual distance measure based on internal conflict between two belief functions we first transform the two belief functions m_0 and m_0' to their base belief function form using eqs. (1) and (2) to find the base set, this is followed by eq. (5) to construct the base belief function, m_{00} and m_{00}'. We perform a conjunctive combination $m_{00}'' = m_{00} \textcircled{\cap} m_{00}'$ and find the internal conflict $m_{00}''(\varnothing)$ of the resulting base belief function using eq. (7).

This measure $m''_{00}(\varnothing)$ of internal conflict is the most objective conflict measure since it excludes pure conflict and is immune to normalizations and incoming belief functions from sources without any information on conflicts in earlier combinations.

In the problem of partitioning mixed-up belief functions into subsets that correspond to different subproblems [8–11] we may use the distance measure of internal conflict for all belief functions (i.e., also for non-consonant belief functions).

5 An Example

Let us study a simple example. In this example we will represent all belief functions using numerical ordering of focal elements.

We assume a frame of discernment $\Theta = \{a, b\}$ and a belief function m_0 that is build up by combination of two SSFs m_2 and m_3 that are yet unknown to us, where

$$m_2 = [0, 0.4, 0, 0.6], \qquad m_3 = [0, 0, 0.4, 0.6]. \tag{8}$$

We have,

$$m_0 = m_2 \textcircled{\cap} m_3 = [0.16, 0.24, 0.24, 0.36]. \tag{9}$$

Using eqs. (1) and (2), m_0 can be decomposed into the base SSFs m_2 and m_3. Here, the base conflict is 0, i.e., m_1 in the decomposition of m_0 is a vacuous SSF; $m_1(\Theta) = 1$. From the base set eq. (8) we can construct the base belief function m_{00} which in this case is identical to the belief function m_0.

If m_0 is normalized then the situation is different. Let us call this normalization m_{0n}. We have,

$$m_{0n} = [0, 0.2857, 0.2857, 0.4286]. \tag{10}$$

Decomposing m_{0n} we get

$$\begin{aligned} &m_{1n} = [-0.1905, 0, 0, 1.1905], \qquad m_{2n} = [0, 0.4, 0, 0.6], \\ &m_{3n} = [0, 0, 0.4, 0.6] \end{aligned} \tag{11}$$

which is the same base for m_{0n} as in the decomposition of m_0. Thus, m_0 and m_{0n} has the same base belief function, which is $m_{00} = m_0$. However, we obtain an inverse base conflict of $m_{1n}(\varnothing) = -0.1905$ when decomposing m_{0n} compared to $m_1(\Theta) = 1$ in the decomposition of m_0.

Furthermore, let us assume that we have a second belief function m'_0 which is build up by combination of two SSFs m_2 and m_3 that are also unknown to us, where

$$m'_2 = [0, 0.3, 0, 0.7], \qquad m'_3 = [0, 0, 0.3, 0.7]. \tag{12}$$

We have

$$m_0' = m_2' \textcircled{\cap} m_3' = [0.09, 0.21, 0.21, 0.49]. \quad (13)$$

As above, using eqs. (1) and (2) m_0' can be decomposed into the base m_2' and m_3' (m_1' is vacuous). Using eq. (5) we construct the base belief function m_{00}'.

Finally, given both m_{00} and m_{00}' we combine them to obtain

$$m_{00}'' = m_{00} \textcircled{\cap} m_{00}' = [0.3364, 0.2436, 0.2436, 0.1764]. \quad (14)$$

We notice a conflict of 0.3364 in the combination of m_{00} and m_{00}'.

Assuming instead that we receive m_{00}'' from a source (let us then call it m_0'') we can decompose it to obtain a pure base without any base conflict;

$$\begin{aligned} m_1'' &= [0, 0, 0, 1], \qquad m_3'' = [0, 0.58, 0, 0.42], \\ m_3'' &= [0, 0, 0.58, 0.42]. \end{aligned} \quad (15)$$

We should notice that the two base SSFs m_2'' and m_3'' are themselves conflict free combinations $m_2 \textcircled{\cap} m_2'$ and $m_3 \textcircled{\cap} m_3'$ resulting in m_2'' and m_3'', respectively.

Recombining m_2'' and m_3'' yields a base belief function m_{00}'' identical to m_0''. Thus, the conflict of m_0'' and the internal conflict of m_{00}'' are identical in this case.

Had m_0'' been normalized the situation is somewhat different. Let us call the normalization m_{0n}'''. We have,

$$m_{0n}''' = [0, 0.3671, 0.3671, 0.2658]. \quad (16).$$

It can be decomposed into

$$\begin{aligned} m_{1n}'' &= [-0.3082, 0, 0, 1.3082], \qquad m_{2n}'' = [0, 0.58, 0, 0.42], \\ m_{3n}'' &= [0, 0, 0.58, 0.42]. \end{aligned} \quad (17)$$

We observe that m_0'' and m_{0n}'' have the same base set in that $m_2'' = m_{2n}''$ and $m_3'' = m_{3n}''$.

Finally, let us study a combination of a belief function with itself. We combine m_{0n}''' with itself. We have,

$$m_{0n}'''' = m_{0n}''' \textcircled{\cap} m_{0n}''' = [0.2695, 0.3299, 0.3299, 0.0706] \quad (18)$$

where $m''''_{0n}(\varnothing) = 0.2695$ is the internal conflict distance measure between the two belief function.

Decomposing m''''_{0n} we get,

$$\begin{aligned} m''''_{1n} &= [-1.2711, 0, 0, 2.2711] \qquad m''''_{2n} = [0, 0.8236, 0, 0.1764], \\ m''''_{3n} &= [0, 0, 0.8236, 0.1764]. \end{aligned} \tag{19}$$

As before we have a base set of two SSFs. Here the base set of m''''_{0n} is m''''_{2n} and m''''_{3n}, where

$$m''''_{2n} = m''_{2n} \textcircled{\cap} m''_{2n}, \qquad m''''_{3n} = m''_{3n} \textcircled{\cap} m''_{3n}. \tag{20}$$

6 Conclusions

We conclude that the internal conflict of a non-consonant belief function is actually a function of conflicts between different GSSFs in the base set of that belief function. Here all GSSFs that have identical focal elements have zero conflict. Thus, the internal conflict between two belief functions is an appropriate distance measure on a conceptual level that measures disagreement between sources.

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multiple valued mapping. The Annals of Mathematical Statistics 38, 325–339 (1967)
2. Dempster, A.P.: A generalization of Bayesian inference. Journal of the Royal Statistical Society B 30, 205–247 (1968)
3. Dempster, A.P.: The Dempster-Shafer calculus for statisticians. International Journal of Approximate Reasoning 48, 365–377 (2008)
4. Haenni, R.: Shedding new light on Zadeh's criticism of Dempster's rule of combination. In: Proceedings of the Seventh International Conference on Information Fusion, pp. 879–884 (2005)
5. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53, 118–145 (2012)
6. Klein, J., Colot, O.: Automatic discounting rate computation using a dissent criterion. In: Proceedings of the Workshop on the Theory of Belief Functions, pp. 1–6 (paper 124) (2010)
7. Liu, W.: Analyzing the degree of conflict among belief functions. Artificial Intelligence 170, 909–924 (2006)
8. Schubert, J.: On nonspecific evidence. International Journal of Intelligent Systems 8, 711–725 (1993)
9. Schubert, J.: Clustering belief functions based on attracting and conflicting metalevel evidence using Potts spin mean field theory. Information Fusion 5, 309–318 (2004)

10. Schubert, J.: Managing decomposed belief functions. In: Bouchon-Meunier, B., Marsala, C., Rifqi, M., Yager, R.R. (eds.) Uncertainty and Intelligent Information Systems, pp. 91–103. World Scientific Publishing Company, Singapore (2008)
11. Schubert, J.: Clustering decomposed belief functions using generalized weights of conflict. International Journal of Approximate Reasoning 48, 466–480 (2008)
12. Schubert, J.: Conflict management in Dempster-Shafer theory using the degree of falsity. International Journal of Approximate Reasoning 52, 449–460 (2011)
13. Schubert, J.: Constructing and evaluating alternative frames of discernment. International Journal of Approximate Reasoning 53, 176–189 (2012)
14. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
15. Smets, P.: The canonical decomposition of a weighted belief. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1896–1901 (1995)
16. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion 8, 387–412 (2007)

Plausibility in DSmT*

Milan Daniel

Abstract. Preparing for generalization of results on conflicts of classic belief function to DSm approach, we need normalized plausibility of singletons also in DSmT. To enable this, plausibility of DSm generalized belief functions is analyzed and compared on entire spectrum of DSm models for various types of belief functions; from simple uniform distribution, through general classic belief function, to general generalized belief function in full generality. Both numeric and comparative variability with respect to particular DSm models has been observed and described. This comparative study enables deeper understanding of plausibility in DSm approach and also underlines the sensitivity to selection of particular DSm models.

Figure of elements of DSm domain — DSm hyper-power set — and figures representing particular DSm models (the free DSm model, hybrid DSm models, and Shafer's model) throughout the text enable better understanding of DSm principles.

Further, a notion of non-conflicting DSm model is introduced and characterized towards the end of the study.

1 Introduction

Investigating nature of conflicts of classic belief functions (BFs) [5, 6], plausibility function, specially normalized plausibility of singletons was utilized. To generalize/transform the classic results of [6] to DSm approach we need plausibility function also in DSmT. Plausibility was defined already in [7, 9] on DSm free model; nevertheless it was not studied in detail as it has value 1 for all elements of a frame of discernment; thus it is considered as not interesting and is often ignored in DSmT.

Milan Daniel
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, CZ – 182 07 Prague 8, Czech Republic
e-mail: milan.daniel@cs.cas.cz

* This research is supported by the grant P202/10/1826 of the Grant Agency of the Czech Republic. Partial support by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications" is also acknowledged.

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 179–187.
springerlink.com

On the other hand, we can observe that constant plausibility value 1 holds true in DSm free model, but it does not hold true in general; simple counter-example is Shafer's DSm model (i.e. classic Shafer's approach).

The free DSm model is presented as the most general DSm model in DSmT. Considering classic approach [3, 4], the free model is one of special models, where plausibility is constant and equal to 1 for all elements of the corresponding frame of discernment. Using this, we investigate plausibility in different DSm models in this contribution. After that, the notion of non-conflicting DSm model is introduced and characterized in the study.

2 Preliminaries

Let us suppose classic belief functions according to Shafer's book [8]. Further we will use DSm approach [9] in the notation of Chapter 3 of volume 2 [10]. DSmT supposes non-empty intersections of all elements of a frame of discernment in general. For bibliography of DSmT and free download of [9, 10] see its web page [11].

Application of non-existential constraints is analogous to usage of smaller frame of discernment (the original one without constrained elements). Hence, we will not deal with non-existential constraints in this study.

Analogously to the classic Shafer's approach, belief and plausibility functions are defined on DSm hyper-power set D^Θ as follows: $Bel(A) = \sum_{\emptyset \neq X \subseteq A, X \in D^\Theta} m(X)$, $Pl(A) = \sum_{\emptyset \neq X \cap A} m(X)$. Hyper-power set D^Θ is the set containing $\emptyset$ and all unions and intersections of elements θ_i of the frame of discernment Θ.

3 Plausibility of Belief Functions on Frame $\Theta_3 = \{\theta_1, \theta_2, \theta_3\}$

3.1 *Plausibility of Uniform Distribution of Belief Masses to Elements of Frame of Discernment* $\Theta_3 = \{\theta_1, \theta_2, \theta_3\}$

Let us start with a simple BF U_3 which assigns $1/3$ to any $\theta_i \in \Theta_3$, i.e. $m(\theta_1) = m(\theta_2) = m(\theta_3) = 1/3$. We will compute plausibility of U_3 in particular DSm models starting from the free DSm model without any constraint, finishing by Shafer's model with all possible exclusivity constraints $\theta_1 \cap \theta_2 \equiv \theta_1 \cap \theta_3 \equiv \theta_2 \cap \theta_3 \equiv \theta_1 \cap \theta_2 \cap \theta_3 \equiv \emptyset$ in the following subsections.

3.1.1 Plausibility of U_3 in the Free DSm Model $\mathcal{M}^f$

The DSm free model $\mathcal{M}^f$ has not any constraint, see Fig. 2. It corresponds to entire DSm hyper-power set D^Θ (i.e., Dedekind lattice extended with $\emptyset$); it contains $\emptyset$ and 18 non-empty elements for 3-element frame Θ_3, see [9] Chap. 2 and Fig. 1:

$$\begin{array}{llll} \alpha_0 = \emptyset, & \alpha_5 = (\theta_1 \cup \theta_2) \cap \theta_3), & \alpha_{11} = \theta_3, & \alpha_{15} = \theta_1 \cup \theta_2, \\ \alpha_1 = \theta_1 \cap \theta_2 \cap \theta_3, & \alpha_6 = (\theta_1 \cup \theta_3) \cap \theta_2, & \alpha_{12} = (\theta_1 \cap \theta_2) \cup \theta_3, & \alpha_{16} = \theta_1 \cup \theta_3, \\ \alpha_2 = \theta_1 \cap \theta_2, & \alpha_7 = (\theta_2 \cup \theta_3) \cap \theta_1, & \alpha_{13} = (\theta_1 \cap \theta_3) \cup \theta_2, & \alpha_{17} = \theta_2 \cup \theta_3, \end{array}$$

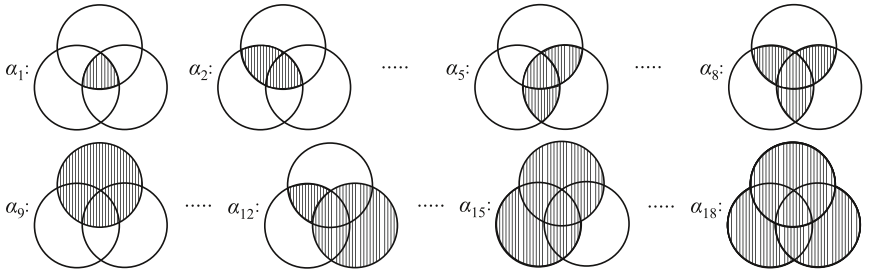


Fig. 1 Non-empty elements of hyper-power set D^{Θ_3}.

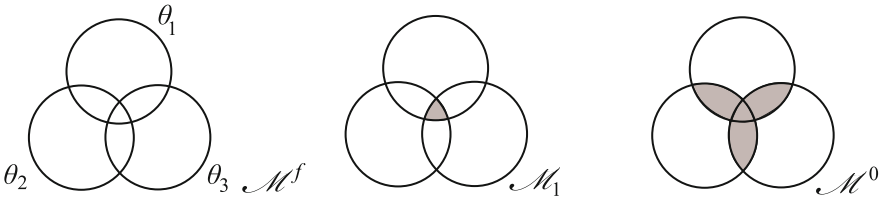


Fig. 2 DSm models $\mathscr{M}^f$, $\mathscr{M}_1$, and $\mathscr{M}^0$. Grey parts are excluded by constraints in $\mathscr{M}_1$ and $\mathscr{M}^0$.

$\alpha_3 = \theta_1 \cap \theta_3$, $\alpha_9 = \theta_1$, $\alpha_{14} = (\theta_2 \cap \theta_3) \cup \theta_1$, $\alpha_{18} = \theta_1 \cup \theta_2 \cup \theta_3$,
$\alpha_4 = \theta_2 \cap \theta_3$, $\alpha_{10} = \theta_2$, $\alpha_8 = (\theta_1 \cap \theta_2) \cup (\theta_1 \cap \theta_3) \cup (\theta_2 \cap \theta_3)$.

We can compute plausibility according to definition for every α_i, $i = 1, ..., 18$. Or we can compute $Pl(\alpha_1) = \sum_{\emptyset \neq X \cap \alpha_1} m(X) = \sum_{i=1}^{18} \alpha_i = m(\theta_1) + m(\theta_2) + m(\theta_3) = 1$, and further use that $\alpha_1 \subset \alpha_i$ for $i = 2, 3, 4, ..., 18$, thus $Pl(\alpha_i) = 1$ for any $\alpha_i \in D^{\Theta_3}$.

3.1.2 Plausibility of U_3 in Hybrid DSm Model $\mathscr{M}_1$

The simplest hybrid DSm model is $\mathscr{M}_1$ where only the conjunction of all 3 elements of the frame Θ_3 is excluded, i.e., $\alpha_1 = \theta_1 \cap \theta_2 \cap \theta_3 \overset{\mathscr{M}_1}{\equiv} \emptyset$. We have 17 non-empty elements of the constrained hyper-power set $D^{\Theta_3}_{\mathscr{M}_1}$ in this DSm model.

We can compute plausibility according to definition for every α_i, $i = 2, ..., 18$ again. Or we can notice, that θ_1 has non-empty intersection with any α_i, $i = 2, 3, 5, 6, 7..., 18$, analogously θ_2 has non-empty intersection with all α_i except for $\alpha_3 = \theta_1 \cap \theta_3$ and similarly θ_3 has non-empty intersection with all α_i except for α_2. Hence $Pl(\alpha_i) = \sum_{i=1}^{3} m(\theta_i) = 1$ for any α_i for $i = 5, 6, 7, ..., 18$ and $Pl(\alpha_2) = m(\theta_1) + m(\theta_2) = 2/3 = Pl(\alpha_3) = Pl(\alpha_4)$ in $D^{\Theta_3}_{\mathscr{M}_1}$.

3.1.3 Plausibility of U_3 in Hybrid DSm Models $\mathscr{M}_2$ – $\mathscr{M}_4$

Futher exclusion is exclusion of intersection of two elements, e.g. $\theta_1 \cap \theta_2$ is excluded in hybrid DSm model $\mathscr{M}_2$, see Fig. 3, thus $\alpha_2 = \theta_1 \cap \theta_2 \overset{\mathscr{M}_2}{\equiv} \emptyset$. As $\alpha_1 = \theta_1 \cap \theta_2 \cap \theta_3 \subset \theta_1 \cap \theta_2 = \alpha_2$, we have $\alpha_1 \overset{\mathscr{M}_2}{\equiv} \alpha_2 \overset{\mathscr{M}_2}{\equiv} \emptyset$. Further $\alpha_6 = (\theta_1 \cup \theta_3) \cap \theta_2 \overset{\mathscr{M}_2}{\equiv} \theta_2 \cap \theta_3 = \alpha_4$, and

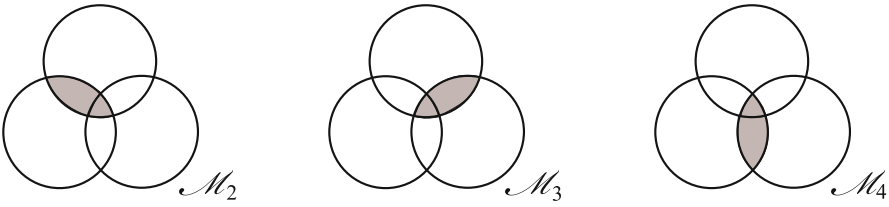


Fig. 3 DSm models $\mathscr{M}_2 - \mathscr{M}_4$. Grey parts are excluded by constraints.

analogically $\alpha_7 = (\theta_2 \cup \theta_3) \cap \theta_1 \overset{\mathscr{M}_2}{\equiv} \theta_1 \cap \theta_3 = \alpha_3, \alpha_8 = (\theta_1 \cap \theta_2) \cup (\theta_1 \cap \theta_3) \cup (\theta_2 \cap \theta_3) \overset{\mathscr{M}_2}{\equiv} (\theta_1 \cup \theta_2) \cap \theta_3 = \alpha_5$ and $\alpha_{12} = (\theta_1 \cap \theta_2) \cup \theta_3 \overset{\mathscr{M}_2}{\equiv} \theta_3$. Thus we have only 12 different non-empty elements of constrained $D^{\Theta_3}_{\mathscr{M}_2}$: $\alpha_3 - \alpha_5$, $\alpha_9 - \alpha_{11}$, and $\alpha_{13} - \alpha_{18}$.

Plausibility function in $\mathscr{M}_2$ has the following values: $Pl(\alpha_3) = Pl(\theta_1 \cap \theta_3) = Pl(\theta_1) = Pl(\alpha_9) = m(\theta_1) + m(\theta_3) = 2/3$ and $Pl(\alpha_4) = Pl(\theta_2 \cap \theta_3) = Pl(\theta_2) = Pl(\alpha_{10}) = m(\theta_2) + m(\theta_3) = 2/3$. There is $Pl(\alpha_i) = \sum_{j=1}^{3} m(\theta_j) = 1$ for any α_i for $i = 5, 11, 13, 14, ..., 18$ in $\mathscr{M}_2$.

And similarly for hybrid DSm model $\mathscr{M}_3$ (resp. $\mathscr{M}_4$), where $\theta_1 \cap \theta_3$ (resp. $\theta_2 \cap \theta_3$) is excluded.

3.1.4 Plausibility of U_3 in Hybrid DSm Models $\mathscr{M}_5$ – $\mathscr{M}_7$

Greater exclusion does mean to exclude two intersections of couples of elements of Θ, e.g. $\alpha_2 = \theta_1 \cap \theta_2 \overset{\mathscr{M}_5}{\equiv} \alpha_3 = \theta_1 \cap \theta_3 \overset{\mathscr{M}_5}{\equiv} \emptyset$ (and implicitly also $\alpha_1 = \theta_1 \cap \theta_2 \cap \theta_3 \overset{\mathscr{M}_5}{\equiv} \emptyset \overset{\mathscr{M}_5}{\equiv} (\theta_2 \cup \theta_3) \cap \theta_1$. Further $\alpha_5 = (\theta_1 \cup \theta_2) \cap \theta_3 \overset{\mathscr{M}_5}{\equiv} \alpha_6 = (\theta_1 \cup \theta_3) \cap \theta_2 \overset{\mathscr{M}_5}{\equiv} \alpha_4 = \theta_2 \cap \theta_3$. $\alpha_{12} = (\theta_1 \cap \theta_2) \cup \theta_3 \overset{\mathscr{M}_5}{\equiv} \alpha_{11} = \theta_3$. $\alpha_{13} = (\theta_1 \cap \theta_3) \cup \theta_2 \overset{\mathscr{M}_5}{\equiv} \alpha_{10} = \theta_2$. Thus we have only 9 different non-empty elements of constrained $D^{\Theta_3}_{\mathscr{M}_5}$ in hybrid DSm model $\mathscr{M}_5$, see Fig. 4: α_4, $\alpha_9 - \alpha_{11}$, and $\alpha_{14} - \alpha_{18}$.

Plausibility function in $\mathscr{M}_5$ has the following values: $Pl(\alpha_9) = Pl(\theta_1) = m(\theta_1) = 1/3$, $Pl(\alpha_{10}) = Pl(\theta_2) = Pl(\alpha_{11}) = Pl(\theta_3) = Pl(\alpha_4) = Pl(\theta_2 \cap \theta_3) = Pl(\alpha_{17}) = Pl(\theta_2 \cup \theta_3) = m(\theta_2) + m(\theta_3) = 2/3$, and there is $Pl(\alpha_i) = \sum_{j=1}^{3} m(\theta_j) = 1$ for $i = 14, 15, 16, 18$ in $\mathscr{M}_5$.

And similarly for hybrid DSm model $\mathscr{M}_6$ (resp. $\mathscr{M}_7$), where $\theta_1 \cap \theta_2$ and $\theta_2 \cap \theta_3$ (resp. $\theta_1 \cap \theta_3$ and $\theta_2 \cap \theta_3$) are excluded.

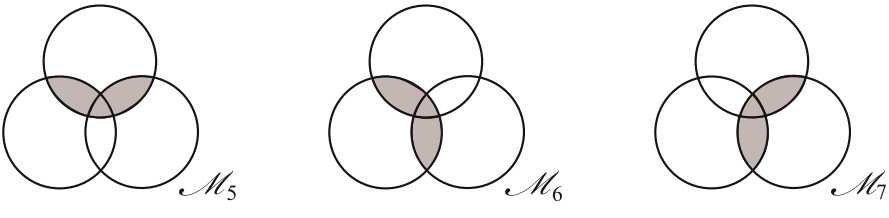


Fig. 4 DSm models $\mathscr{M}_5 - \mathscr{M}_7$. Grey parts are excluded by constraints.

3.1.5 Plausibility of U_3 on Shafer's DSm Model $\mathscr{M}^0$

The greatest exclusion on Θ_3 is exclusion of all 3 intersections of couples of elements and consequently intersection of 3 elements and unions of excluded intersections, i.e. exclusion of all $\alpha_1, \alpha_2, ..., \alpha_8$. Further $\alpha_{12} = (\theta_1 \cap \theta_2) \cup \theta_3 \overset{\mathscr{M}^0}{\equiv} \alpha_{11} = \theta_3$, $\alpha_{13} = (\theta_1 \cap \theta_3) \cup \theta_2 \overset{\mathscr{M}^0}{\equiv} \alpha_{10} = \theta_2$, $\alpha_{14} = (\theta_2 \cap \theta_3) \cup \theta_1 \overset{\mathscr{M}^0}{\equiv} \alpha_9 = \theta_1$, Thus we have just 7 different non-empty elements of constrained $D^{\Theta_3}_{\mathscr{M}^0}$: $\alpha_9 = \theta_1$, $\alpha_{10} = \theta_2$, $\alpha_{11} = \theta_3$, $\alpha_{15} = \theta_1 \cup \theta_2$, $\alpha_{16} = \theta_1 \cup \theta_3$, $\alpha_{17} = \theta_2 \cup \theta_3$, $\alpha_{18} = \theta_1 \cup \theta_2 \cup \theta_3$, which correspond to non-empty elements of classic power set as in classic Dempster-Shafer approach.

Hence we have $Pl(\alpha_9) = Pl(\theta_1) = m(\theta_1) = 1/3, Pl(\theta_2) = m(\theta_2) = 1/3, Pl(\theta_3) = m(\theta_3) = 1/3, Pl(\alpha_{15}) = Pl(\theta_1 \cup \theta_2) = m(\theta_1) + m(\theta_2) = 2/3, Pl(\theta_1 \cup \theta_2) = m(\theta_1) + m(\theta_3) = 2/3, Pl(\theta_2 \cup \theta_3) = m(\theta_2) + m(\theta_3) = 2/3$, and $Pl(\alpha_{18}) = Pl(\theta_1 \cup \theta_2 \cup \theta_3) = m(\theta_1) + m(\theta_2) + m(\theta_3) = 1$, as in classic Dempster-Shafer approach.

3.1.6 Summary of Plausibility of U_3 in DSm Models

The presented simple example displays that plausibility function related to the same belief assignment has different values in different DSm models. This also characterize particular DSm models.

3.2 *Plausibility of Non-Uniform Distribution of Belief Masses to Elements of a Frame of Discernment* $\Theta_3 = \{\theta_1, \theta_2, \theta_3\}$

Let us assume Bayesian BF *Bel* which assigns belief masses as it follows $m(\theta_1) = 0.6, m(\theta_2) = 0.3, m(\theta_3) = 0.1$.

In the same way as in subsection 3.1.1, we can compute $Pl(\alpha_i) = 1$ for any $\alpha_i \in D^{\Theta_3}$ in $\mathscr{M}^f$. Analogously $Pl(\alpha_1) = 0$, $Pl(\theta_i \cap \theta_j) = m(\theta_i) + m(\theta_j)$, and $Pl(\alpha_i) = \sum_{i=1}^{3} m(\theta_i) = 1$ for any α_i for $i = 5, 6, 7, ..., 18$ in $\mathscr{M}_1$.

In Shafer's model we have $Pl(\theta_1) = 0.6$, $Pl(\theta_2) = 0.3$, $Pl(\theta_3) = 0.1$, $Pl(\theta_1 \cup \theta_2) = 0.9$, $Pl(\theta_1 \cup \theta_2) = 0.7$, $Pl(\theta_2 \cup \theta_3) = 0.4$, and $Pl(\theta_1 \cup \theta_2 \cup \theta_3) = 1$, as in classic Dempster-Shafer approach.

3.2.1 Plausibility of Non-uniform Distribution in Models $\mathscr{M}_2$ – $\mathscr{M}_4$

Plausibility function has the following values in hybrid DSm model $\mathscr{M}_2$: $Pl(\alpha_3) = Pl(\theta_1 \cap \theta_3) = Pl(\theta_1) = Pl(\alpha_9) = m(\theta_1) + m(\theta_3) = 0.7$, $Pl(\alpha_4) = Pl(\theta_2 \cap \theta_3) = Pl(\theta_2) = Pl(\alpha_{10}) = m(\theta_2) + m(\theta_3) = 0.4$, and $Pl(\alpha_i) = \sum_{j=1}^{3} m(\theta_j) = 1$ for any α_i for $i = 5, 11, 13, 14, ..., 18$ in $\mathscr{M}_2$. Analogously, there is $Pl(\theta_1) = m(\theta_1) + m(\theta_2) = 0.9$, $Pl(\theta_2) = m(\theta_1) + m(\theta_2) + m(\theta_3) = 1$, $Pl(\theta_3) = m(\theta_2) + m(\theta_3) = 0.4$, in $\mathscr{M}_3$. Similarly, $Pl(\theta_1) = 1$, $Pl(\theta_2) = 0.9$, $Pl(\theta_3) = 0.7$, in $\mathscr{M}_4$.

Thus for $m(\theta_1) > m(\theta_2) > m(\theta_3)$, where $m(\theta_1) + m(\theta_2) + m(\theta_3) = 1$ we have $Pl(\theta_2) > Pl(\theta_1)$ in $\mathscr{M}_3$ and $Pl(\theta_3) > Pl(\theta_1), Pl(\theta_2)$ in $\mathscr{M}_2$.

3.2.2 Plausibility of Non-uniform Distribution in Models $\mathcal{M}_5$ – $\mathcal{M}_7$

Plausibility function has the following values in $\mathcal{M}_5$: $Pl(\theta_1) = m(\theta_1) = 0.6$, $Pl(\alpha_{10}) = Pl(\theta_2) = Pl(\alpha_{11}) = Pl(\theta_3) = Pl(\alpha_4) = Pl(\theta_2 \cap \theta_3) = Pl(\alpha_{17}) = Pl(\theta_2 \cup \theta_3) = m(\theta_2) + m(\theta_3) = 0.4$, and there is $Pl(\alpha_i) = \sum_{j=1}^{3} m(\theta_j) = 1$ for $i = 14, 15, 16, 18$.

Analogously, there is $Pl(\theta_1) = Pl(\theta_3) = m(\theta_1) + m(\theta_3) = 0.7$, and $Pl(\theta_2) = m(\theta_2) = 0.4$ in $\mathcal{M}_6$. Similarly $Pl(\theta_1) = Pl(\theta_2) = 0.9$, and $Pl(\theta_3) = 0.1$ in $\mathcal{M}_7$.

Thus for $m(\theta_1) > m(\theta_2) > m(\theta_3)$, where $m(\theta_1) + m(\theta_2) + m(\theta_3) = 1$ we have $Pl(\theta_2) = Pl(\theta_3)$ in $\mathcal{M}_5$, $Pl(\theta_1) = Pl(\theta_1)$ in $\mathcal{M}_7$, and $Pl(\theta_1) = Pl(\theta_3) > Pl(\theta_2)$ in $\mathcal{M}_6$.

3.3 Plausibility of General Classic and General Generalized BFs

3.3.1 Plausibility of a General Classic BF in DSm Models

Let us consider a general classic BF Bel_c defined by m_c as it follows $m_c(\theta_1) = 0.40$, $m_c(\theta_2) = 0.20$, $m_c(\theta_3) = 0.05$, $m_c(\theta_1 \cup \theta_2) = 0.15$, $m_c(\theta_1 \cup \theta_3) = 0.10$, $m_c(\theta_2 \cup \theta_3) = 0.05$, $m_c(\theta_1 \cup \theta_2 \cup \theta_3) = 0.05$[1]. We will compute plausibility corresponding to Bel_c in particular DSm models again, for results see Tab. 1. Notice, that $Pl(\theta_2) = Pl(\theta_3)$ in $\mathcal{M}_5$, $Pl(\theta_1) = Pl(\theta_2)$ in $\mathcal{M}_7$, and $Pl(\theta_1) = Pl(\theta_3) > Pl(\theta_2)$ in $\mathcal{M}_6$.

Table 1 Plausibility of hyper-power set elements α_j for general classic BF Bel_c in DSm model $\mathcal{M}_i$

Model	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}
$m_c(\alpha_i)$:									0.4	0.2	0.05				0.15	0.10	0.05	0.05
$\mathcal{M}^f$:	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_1$:	$\emptyset$	0.95	0.8	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_2$:	$\emptyset$	$\emptyset$	0.8	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_3$:	$\emptyset$	0.95	$\emptyset$	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_4$:	$\emptyset$	0.95	0.8	$\emptyset$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_5$:	$\emptyset$	$\emptyset$	$\emptyset$	0.6	(α_4)	(α_4)	$\emptyset$	(α_4)	0.7	0.6	0.6	(α_{11})	(α_{10})	1.0	1.0	1.0	0.6	1.0
$\mathcal{M}_6$:	$\emptyset$	$\emptyset$	0.6	$\emptyset$	(α_3)	$\emptyset$	(α_3)	(α_3)	0.8	0.55	0.8	(α_{11})	1.0	(α_9)	1.0	0.8	1.0	1.0
$\mathcal{M}_7$:	$\emptyset$	0.95	$\emptyset$	$\emptyset$	$\emptyset$	(α_2)	(α_2)	(α_2)	0.95	0.95	0.25	1.0	(α_{10})	(α_9)	0.95	1.0	1.0	1.0
$\mathcal{M}^0$:	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.7	0.45	0.25	(α_{11})	(α_{10})	(α_9)	0.95	0.8	0.6	1.0

[1] Note that $\theta_1 \cup \theta_2$ corresponds to $\{\theta_1, \theta_2\}$ in classic Shafer's notation. Similarly, $\alpha_{18} = \theta_1 \cup \theta_2 \cup \theta_3$ corresponds to $\Theta_3 = \{\theta_1, \theta_2, \theta_3\}$.

Table 2 Plausibility of hyper-power set elements α_j for generalized BF Bel_g in DSm model $\mathcal{M}_i$

Model	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}
$m_g(\alpha_i)$:	0.05	0.09	0.04	0.02	0.02	0.03	0.05	0.05	0.18	0.09	0.03	0.01	0.03	0.06	0.08	0.04	0.03	0.10
$\mathcal{M}^f$:	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{M}_1$:	$\emptyset$	0.84	0.72	0.59	0.86	0.86	0.86	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
$\mathcal{M}_2$:	$\emptyset$	$\emptyset$	0.72	0.59	0.86	(α_4)	(α_3)	(α_5)	0.72	0.59	0.86	(α_{11})	0.86	0.86	0.86	0.86	0.86	0.86
$\mathcal{M}_3$:	$\emptyset$	0.84	$\emptyset$	0.59	(α_4)	0.91	(α_2)	(α_6)	0.84	0.91	0.59	0.91	(α_{10})	0.91	0.91	0.91	0.91	0.91
$\mathcal{M}_4$:	$\emptyset$	0.84	0.72	$\emptyset$	(α_3)	(α_2)	0.93	(α_7)	0.93	0.84	0.72	0.93	0.93	(α_3)	0.93	0.93	0.93	0.93
$\mathcal{M}_5$:	$\emptyset$	$\emptyset$	$\emptyset$	0.59	(α_4)	(α_4)	$\emptyset$	(α_4)	0.46	0.59	0.59	(α_{11})	(α_{10})	0.77	0.77	0.77	0.59	0.77
$\mathcal{M}_6$:	$\emptyset$	$\emptyset$	0.72	$\emptyset$	(α_3)	$\emptyset$	(α_3)	(α_3)	0.72	0.33	0.72	(α_{11})	0.81	(α_9)	0.81	0.72	0.81	0.81
$\mathcal{M}_7$:	$\emptyset$	0.84	$\emptyset$	$\emptyset$	$\emptyset$	(α_2)	(α_2)	(α_2)	0.84	0.84	0.21	0.87	(α_{10})	(α_9)	0.84	0.87	0.87	0.87
$\mathcal{M}^0$:	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.46	0.33	0.21	(α_{11})	(α_{10})	(α_9)	0.61	0.53	0.41	0.65

3.3.2 Plausibility of a General Generalized BF in DSm Models

Analogously, we can compute plausibility values for a general generalized BF Bel_g given by generalized bba m_g, see Tab. 2[2]. Notice that, $Pl(\theta_3) > Pl(\theta_1), Pl(\theta_2)$ in $\mathcal{M}_2$, $Pl(\theta_2) > Pl(\theta_1)$ in $\mathcal{M}_3$, $Pl(\theta_3) = Pl(\theta_2) > Pl(\theta_1)$ in $\mathcal{M}_5$, and $Pl(\theta_1) = Pl(\theta_3) > Pl(\theta_2)$ in $\mathcal{M}_6$.

4 Summary

We have observed variability of plausibility values for different DSm models for large spectrum of BFs from simple uniform distribution U_3 through general classic BFs to general generalized BFs. We have observed not only numeric variability, but also the comparative one: e.g., $Pl(\theta_1) = Pl(\theta_3) > Pl(\theta_2)$ in $\mathcal{M}_6$ for all investigated types of BFs, and $Pl(\theta_3) > Pl(\theta_1), Pl(\theta_2)$ in $\mathcal{M}_2$ for generalized BFs whereas $Pl(\theta_1) > Pl(\theta_2) > Pl(\theta_3)$ for all investigated BFs in $\mathcal{M}^0$.

This characterizes both plausibility functions and hybrid DSm models. It underlines sensitivity of DSm approach to selection of hybrid models and stresses out the necessity to be careful when selecting a DSm model for real-world applications.

[2] Fully general generalized bba m_g is out of DSm models $\mathcal{M}_1$ — $\mathcal{M}_7$ and $\mathcal{M}^0$. To tune m_g with particular DSm models it should be normalized over non-constrained α_i's in particular models. Thus there should be 8 different bba's $m_{\mathcal{M}^f} \equiv m_g$, $m_{\mathcal{M}_1}$ — $m_{\mathcal{M}_7}$ and $m_{\mathcal{M}^0}$. For better comparison of behaviour of particular DSm models, we use the only generalized bba m_g, thus the values in Tab. 2 should be normalized (i.e., divided by particular values for α_{18}, i.e., divided by 1 minus sum of bbas of excluded elements) to represent correct plausibility values.

5 Non-conflicting DSm Models

BF *Bel* is non-conflicting (without internal conflict) when $Bel \textcircled{\cap} Bel$ do not assign any belief mass to $m(\emptyset)$, [5, 6]. All elements of entire (non-constrained) hyper-power set D^Θ have non-empty intersection in the free DSm model ($\alpha_i \cap \alpha_j \not\equiv \emptyset$ in $\mathscr{M}^f$). Thus any BF in $\mathscr{M}^f$ is non-conflicting and any two BFs are mutually non-conflicting in $\mathscr{M}^f$. Hence $\mathscr{M}^f$ is *non-conflicting DSm model* and moreover it is the only non-conflicting model for generalized BFs in full generality.

Theorem 1. *The free DSm model $\mathscr{M}^f$ is the only non-conflicting DSm model in full generality.*

Let us note that considering some special class of BFs, e.g. classic BFs as inputs, the issue of non-conflictness of DSm models is more complicated.

6 Conclusion

Plausibility functions for various types of belief functions (BFs) were analyzed and compared on entire spectrum of DSm models on Θ_3. Both numeric and comparative variabilities were observed. Finally, the notion of non-conflicting DSm model was introduced.

The original purpose of this study was just a preparation for generalizations of results on conflicts of BFs from [6]. As a side-effect, a nature of particular DSm models was displayed for classic BF audience; further a variability and importance of plausibility in DSmT for DSmT researchers; and also significant sensitivity of results to selection of a specific DSm model was shown.

References

1. Cholvy, L.: Using Logic to Understand Relations between DSmT and Dempster-Shafer Theory. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS(LNAI), vol. 5590, pp. 264–274. Springer, Heidelberg (2009)
2. Daniel, M.: A Generalization of the Classic Combination Rules to DSm Hyper-power Sets. Information & Security. An International Journal 20, 50–64 (2006)
3. Daniel, M.: The DSm Approach as a Special Case of the Dempster-Shafer Theory. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 381–392. Springer, Heidelberg (2007)
4. Daniel, M.: Contribution of DSm Approach to the Belief Function Theory. In: Magdalena, L., Ojeda-Aciego, M., Verdegay, J.L. (eds.) Proc. of IPMU 2008, Málaga, pp. 417–424 (2008)
5. 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)
6. Daniel, M.: Non-conflicting and Conflicting Parts of Belief Functions. In: Coolen, F., de Cooman, G., Fetz, T., Oberguggenberger, M. (eds.) Proceedings of the 7th ISIPTA 2011, pp. 149–158. Studia Universitätsverlag, Innsbruck (2011)

7. Dezert, J.: Foundations for a New Theory of Plausible and Paradoxical Reasoning. Information and Security, An International Journal 9 (2002)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton Univ. Press, New Jersey (1976)
9. Smarandache, F., Dezert, J.: Advances and Applications of DSmT for Information Fusion. American Research Press, Rehoboth (2004)
10. Smarandache, F., Dezert, J.: Advances and Applications of DSmT for Information Fusion, vol. 2. American Research Press, Rehoboth (2006)
11. www.gallup.unm.edu/~smarandache/DSmT.htm (cited, January 28, 2012)

A Belief Function Model for Pixel Data

John Klein and Olivier Colot

Abstract. Image data *i.e.* pixel values are notably corrupted with uncertainty. A pixel value can be seen as uncertain because of additional noise due to acquisition conditions or compression. It is possible to represent a pixel value in a more imprecise but less uncertain way by considering it as interval-valued instead of a single-valued. The Belief Function Theory (BFT) allows to handle such interval-based pixel representations. We provide in this paper a model describing how to define belief functions from image data. The consistency of this model is demonstrated on edge detection experiments as conflicting pixel-based belief functions lead to image transitions detection.

1 Introduction

The Belief Function Theory (BFT) [3, 8], also known as evidence theory or Dempster-Shafer theory, provides a framework for processing uncertain and imprecise data. As image data can be considered as such, an evidential model leading to a new representation of pixel values can be introduced. Existing evidential image processing approaches are mainly dedicated to information fusion on multiple image components or neighbor pixels as part of pixel classification algorithms [1, 10]. In this article, we intend to process images using the BFT under a new perspective by introducing a model that translates directly each raw pixel value into a belief function. Indeed, a pixel value equals x with probability p. It is also possible to consider that the pixel value belongs to the interval $[x-q, x+q]$ with a probability $p' > p$ and $q > 0$. This piece of information can be easily encoded using a belief function.

John Klein
Lille1 University, LAGIS FRE CNRS 3303
e-mail: john.klein@univ-lille1.fr

Olivier Colot
Lille1 University, LAGIS FRE CNRS 3303
e-mail: olivier.colot@univ-lille1.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 189–196.
springerlink.com

In section 2, BFT fundamental concepts necessary for our approach are recalled. Section 3 presents a methodology to represent pixel values as belief functions. Section 4 introduces results for pixel conflict computation. Finally, in section 5, the consistency of the model is demonstrated through edge detection experiments on synthetic gray-scale images.

2 Belief Functions Fundamentals

In this section, BFT fundamentals are briefly recalled. Suppose a finite set of mutually exclusive solutions denoted by $\Omega = \{\omega_1, ..., \omega_K\}$ and called the **frame of discernment**. The set of all subsets of Ω is denoted by 2^Ω. The mass of belief assigned to A by a source S_i is denoted by $m_i(A)$. The function $m_i : 2^\Omega \rightarrow [0,1]$ is called **basic belief assignment** (**bba**) and is such that: $\sum_{A \subseteq \Omega} m_i(A) = 1$.

A set A such that $m_i(A) > 0$ is called a **focal element** of m_i. A bba is denoted by ${}^A m^x$ if it has two focal elements: Ω and $A \subsetneq \Omega$, and if:

$$ {}^A m^x(A) = 1 - x \text{ and } {}^A m^x(\Omega) = x. \tag{1} $$

with $x \in [0,1]$. Such bbas are called **simple bbas** (sbbas).

To combine bbas issued by reliable sources, the conjunctive rule $\textcircled{\cap}$ can be used:

$$ \forall X \in 2^\Omega, m_{1\textcircled{\cap}2}(X) = \sum_{B \cap C = X, B, C \subset \Omega} m_1(B)\, m_2(C). \tag{2} $$

The mass $m(\emptyset)$ is denoted by κ and called the **degree of conflict**. This mass is given support when S_1 and S_2 advocate respectively for non-intersecting solutions. It is thus an indication on how much the two sources disagree.

Furthermore, it is possible to bring down a source of information using an operation called discounting [8]. Discounting m_i with rate $\alpha \in [0,1]$ is defined as:

$$ m_i^\alpha(X) = (1-\alpha) m_i(X) + \alpha \mathbf{1}_{X=\Omega} \tag{3} $$

with $\mathbf{1}$ the indicator function. The higher α is, the stronger the discounting. One may remark that a sbba ${}^A m^x$ is ${}^A m^0$ discounted with rate x.

3 A Model for Pixel Representation Using Belief Functions

In this section, our evidential pixel representation (EPR) model is introduced. Let us denote a pixel $\mathbf{p} = (p_x, p_y)$ with p_x and p_y its coordinates along the two image axes. An 8-bit gray-scale image can be represented by a function $I(\mathbf{p})$ such that $0 \leq I(\mathbf{p}) \leq 255$.

The measured image $\tilde{I}$ can be viewed as the sum of the true image and a random noise B: $\tilde{I} = I + B$, and $B(\mathbf{p}) \sim f_b$ the noise density function. Suppose one is able to find a symmetrical cumulative distribution function F such that $\forall q, F(q) \geq F_b(q) = \int_{-\infty}^{q} f_b(v)\, dv$, then the following assertion holds:

$$1-2F(-q) \leq \mathbb{P}\left(\text{true pixel value } I(\mathbf{p}) \in A_{\mathbf{p},q} = \left[\tilde{I}(\mathbf{p})-q, \tilde{I}(\mathbf{p})+q\right]\right) \leq 1 \quad (4)$$

with $\mathbb{P}$ the probability measure on the pixel value set. In case of a Gaussian centered noise, F is simply the cumulated density of a centered Gaussian function with a greater standard deviation than that of f_b. As pixel values are integers ranging from 0 to 255, the set of possible values for q is $Q = \{0.5, 1.5, ..., 255.5\}$. Consequently, the set generated by intersections and unions of all possible $A_{\mathbf{p},q}$ is $\Omega = \{-255, -254, ..., 512\}$. Now, the information on a pixel value can be represented by a parametrized bba $m_{\mathbf{p},q}$ defined on 2^Ω:

$$m_{\mathbf{p},q} = {}^{A_{\mathbf{p},q}}m^{2F(-q)} \quad (5)$$

Multiple bbas may be defined by this mean, as it is difficult to determine what value for q to choose. To cumulate all pieces of evidence, we propose to define the bba representing pixel $\mathbf{p}$ as the conjunctive combination of parametrized bbas for all possible values of q:

$$m_{\mathbf{p}} = \bigcirc\!\!\!\!\!\!\cap_{q\in Q}\, m_{\mathbf{p},q}. \quad (6)$$

The focal elements of $m_{\mathbf{p}}$ are nested: $\{A_{\mathbf{p},\mathbf{q}}\}_{q=0.5}^{255.5}$. A frequent criticism addressed to the BFT is the computational load induced by large frames of discernment. In this paper, the cardinal of 2^Ω is 2^{255}, thus computing the above bba using equation (2) is infeasible. Yet, since the set of bbas to combine has some particular properties, this computation can be easily done using the following proposition:

Theorem 1. *Let* $\left\{m_i = {}^{A_i}m^{\alpha_i}\right\}_{i=1}^{N}$ *be a set of sbbas with nested focal elements such that* $A_1 \subsetneq A_2 \subsetneq ... \subsetneq A_N \subsetneq \Omega$. *Let us denote* $m_{\bigcirc\!\!\!\!\!\!\cap}$ *the conjunctive combination of these sbbas. We have:*

$$m_{\bigcirc\!\!\!\!\!\!\cap}(X) = (1-\alpha_i)\prod_{j=1}^{i-1}\alpha_j \mathbf{1}_{X=A_i} + \prod_{j=1}^{N}\alpha_j \mathbf{1}_{X=\Omega} \quad (7)$$

Proof. The focal elements of $m_{\bigcirc\!\!\!\!\!\!\cap}$ are the sets $\{A_i\}_{i=1}^N$ and Ω. For $\{A_i\}_{i=1}^N$, we have:

$$m_{\bigcirc\!\!\!\!\!\!\cap}(A_i) = \sum_{(\cap_{j=1}^N B_j)=A_i, B_j \in \left\{A_j,\Omega\right\}} {}^{A_1}m^{\alpha_1}(B_1)...{}^{A_N}m^{\alpha_N}(B_N)$$

The condition under the sum can only be verified if $\forall j < i, B_j = \Omega$ and $B_i = A_i$. Since ${}^{A_i}m^{\alpha_i}(A_i) = (1-\alpha_i)$ and $\forall j < i$, ${}^{A_j}m^{\alpha_j}(\Omega) = \alpha_j$, we have:

$$m_{\bigcirc\!\!\!\!\!\!\cap}(A_i) = (1-\alpha_i)\prod_{j=1}^{i-1}\alpha_j \sum_{(\cap_{j=i+1}^N B_j)\cap A_i = A_i, B_j \in \left\{A_j,\Omega\right\}} {}^{A_{i+1}}m^{\alpha_{i+1}}(B_{i+1})...{}^{A_N}m^{\alpha_N}(B_N)$$

Because $\forall j > i$, $A_i \cap B_j = A_i$, we have:

$$m_{\textcircled{\cap}}(A_i) = (1-\alpha_i)\prod_{j=1}^{i-1}\alpha_j \sum_{B_j \in \{A_j,\Omega\}} {}^{A_{i+1}}m^{\alpha_{i+1}}(B_{i+1}) \ldots {}^{A_N}m^{\alpha_N}(B_N)$$

and since $m_j(A_j) + m_j(\Omega) = 1$, we get $m_{\textcircled{\cap}}(A_i) = (1-\alpha_i)\prod_{j=1}^{i-1}\alpha_j$.
Finally, the mass allocated to Ω is easily obtained from more general results about the conjunctive rule: $\forall m_i, m_j$, $m_i \textcircled{\cap} m_j(\Omega) = m_i(\Omega)\, m_j(\Omega)$. □

It is important to note that for two different pixels $\mathbf{p}$ and $\mathbf{p}'$, we have $\forall q, m_{\mathbf{p}}(A_{\mathbf{p},q}) = m_{\mathbf{p}'}(A_{\mathbf{p}',q}) = \beta_i$ although $A_{\mathbf{p},q} \neq A_{\mathbf{p}',q}$. All pixels have the same masses β_i but their focal elements are potentially different.

4 Pixel Conflict Computation

In the previous section, we have obtained bbas $m_{\mathbf{p}}$ representing uncertain and imprecise values of each pixel. We present now a simple way to compute pixel-based degree of conflict.

As mentioned before, the cardinality of Ω makes it hard to compute the degree of conflict using equation (2). To overcome this difficulty, we propose to use a result from [4]. It is shown in that article that if there is at least one pairwise positive degree of conflict among a set of bbas $\{m_i\}_{i=1}^{M}$, then the global conflict of a set of identically discounted bbas can be approximated by the sum of pairwise degrees of conflict. Consequently, it makes sense to compute the sum of pairwise degrees of conflict instead of the global degree of conflict, as identically discounting bbas preserves their relative prevalences. The pairwise conflict of two bbas $m_{\mathbf{p}}$ and $m_{\mathbf{p}'}$ can be easily computed using the following result:

Theorem 2. *Let $m_{\mathbf{p}}$ and $m_{\mathbf{p}'}$ be two bbas obtained from the process described in section 3. Then their pairwise conflict $\kappa_{\{\mathbf{p},\mathbf{p}'\}}$ is a function of $\Delta = \left|\tilde{I}(\mathbf{p}) - \tilde{I}(\mathbf{p}')\right|$ and*

$$\kappa_{\{\mathbf{p},\mathbf{p}'\}}(0) = 0, \tag{8}$$

$$\kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta > 0) = \begin{cases} \kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta-1) + 2\sum_{k=1}^{\Delta/2} \beta_k \beta_{\Delta-k} \text{ if } \Delta \text{is even}, \\ \kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta-1) + 2\sum_{k=1}^{(\Delta-1)/2} \beta_k \beta_{\Delta-k} + \beta_{(\Delta+1)/2}^2 \text{ if } \Delta \text{is odd}. \end{cases} \tag{9}$$

Proof. If one denotes by k the index of a focal element of $m_{\mathbf{p}}$ and by l the index of a focal element of $m_{\mathbf{p}'}$, those with empty intersections are such that $k + l \leq \Delta$. Consequently, we obtain $\kappa_{\{\mathbf{p},\mathbf{p}'\}} = \sum_{k+l\leq\Delta} \beta_k \beta_l$. It can be seen that $\kappa_{\{\mathbf{p},\mathbf{p}'\}}$ is a function of Δ which can be recursively computed. Indeed, we have

$\kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta) - \kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta - 1) = \sum_{k+l=\Delta} \beta_k \beta_l$. In the end, some elements of $\sum_{k+l=\Delta} \beta_k \beta_l$ are counted twice that is why two cases are distinguished corresponding odd and even values of Δ. □

The values of β_i for all i and of $\kappa_{\{\mathbf{p},\mathbf{p}'\}}(\Delta)$ for all Δ can be stored in a lookup table, making it easy and fast to compute pixel pairwise degrees of conflict. The method appears to be based only on pixel value differences. The function applied to these differences is entirely justified using the BFT and is based on conflict.

5 Experiments on Edge Detection

The degree of conflict of bbas belonging to the neighborhood $\mathcal{V}_\mathbf{p}$ of pixel $\mathbf{p}$ is obtained as follows: $\kappa(\mathbf{p}) = \sum_{\mathbf{p}' \in \mathcal{V}_\mathbf{p}} \kappa_{\{\mathbf{p},\mathbf{p}'\}}$. It is likely to be a relevant feature for assessing the presence of an edge at pixel $\mathbf{p}$. Consequently, edge detection was chosen to demonstrate the consistency of EPR. For using the EPR, one must first define function F. The unknown noise f_b is supposed to be centered and Gaussian. The function F is thus defined likewise with a greater standard deviation σ_{EPR} than that of the noise. A pixel neighborhood $\mathcal{V}_\mathbf{p}$ is defined as follows: $\mathcal{V}_\mathbf{p} = \left\{\mathbf{p}' | \sqrt{(p_x - p'_x)^2 + (p_y - p'_y)^2} \leq h_{EPR}\right\}$. An edge detector yields a binary edge image whereas $\kappa(\mathbf{p})$ corresponds to an image containing edge probabilities (if normalized). If $\kappa(\mathbf{p})$ appraises correctly image edges, then it should be compliant with output edge probability distributions drawn from classical edge detection algorithms. The algorithms retained for the experiments are : Roberts [7], Prewitt [6], Sobel [9], Canny [2] and LoG [5] edge detectors. Roberts, Prewitt and Sobel detectors are based on image first derivatives whereas LoG is based on second order derivatives. Canny [2] introduced a filter as an optimal solution in terms of detection of step edges, edge localization and uniqueness. In addition to filtering, his approach also comprises two other steps helping to obtain thin edges and to remove false edges. To allow a fair comparison of the methods, we only use in the experiments the filtering part of Canny's approach.

Roberts, Sobel and Prewitt are parameter-free, but LoG, Canny and $\kappa(\mathbf{p})$ are depending on two parameters each: a filter spread h_{LoG}, h_{Canny}, h_{EPR} respectively and a standard deviation σ_{LoG}, σ_{Canny}, σ_{EPR} respectively. For Canny and LoG, the filter spread is usually greater than at least three times the standard deviation. Concerning h_{EPR}, its value was set to 2 for all experiments. σ_{LoG}, σ_{Canny}, σ_{EPR} are hand-tuned in each experiment. The value yielding the lowest Kullback-Leibler D_{KL} divergence is retained. This criterion is defined as:

$$D_{KL}(I_e || GT) = \sum_{\mathbf{p}} I_e(\mathbf{p}) \log\left(\frac{I_e(\mathbf{p})}{GT(\mathbf{p})}\right) \tag{10}$$

where I_e is an edge probability image and GT the ground truth. When these distributions are identical $D_{KL}(I_e || GT) = 1$. Higher values are obtained when the distributions are different. This criterion is adapted to our purpose as it penalizes wrongly

located edges, thick edges and partially detected edges. Note that a contrast enhancement is used on some of the images displayed in this section in order to help the reader to perceive some image details.

5.1 Omni-Directional Edges

In this experiment, the dependence on edge direction is examined. A synthetic image I_1 containing a ramp-edge in shape of a circle is used. The edge is made of two transitions: from 0 to gray level 125 and from 125 to 255. When using a step-edge with a single transition from 0 to 255, the edge is located at a sub-pixel precision which makes it harder to define a ground truth.

I_1, its corresponding ground truth GT as well as the output edge probability distributions produced by several approaches are presented in Figure 1. The D_{KL} obtained in this experiment are gathered in Table 1. $\kappa(\mathbf{p})$ produces the smallest divergence because the edge distribution is thinner.

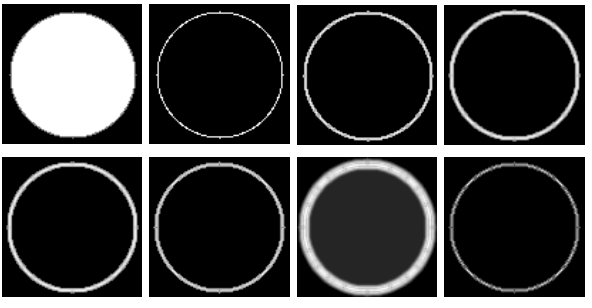


Fig. 1 From top-left to down-right: input image I_1, ground truth GT, output edge distribution using Roberts, Sobel, Prewitt, Canny, LoG and $\kappa(\mathbf{p})$. $\sigma_{Canny} = 0.3$, $\sigma_{LoG} = 0.9$ and $\sigma_{EPR} = 1e5$

Table 1 Performances of several edge detection methods on synthetic image I_1.

Method	Roberts	Sobel	Prewitt	Canny $\sigma_{Canny} = 0.3$	LoG $\sigma_{LoG} = 0.9$	$\kappa(\mathbf{p})$ $\sigma_{EPR} = 1e5$
D_{KL}	8.61	8.32	8.52	7.88	11.56	3.81

5.2 Edges with Varying Contrast

In this experiment, the dependence on edge contrast is examined. The input image I_2 is obtained by shading I_1. I_2, GT and the edge probability distributions are presented in Figure 2. The corresponding D_{KL} are gathered in Table 2. $\kappa(\mathbf{p})$ produces the smallest divergence because the transitions inside the circle are filtered out. The divergence is more stringent on this aspect than on detecting the whole circle. Smaller values of σ_{EPR} leads to performances close to Canny's ones.

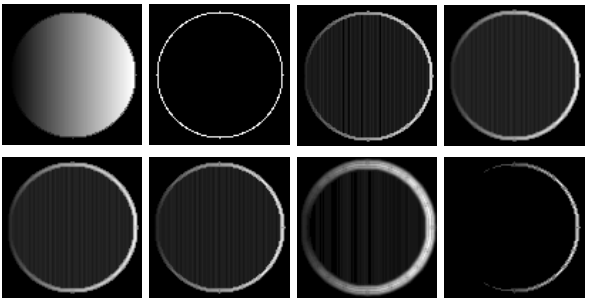


Fig. 2 From top-left to down-right: input image I_2, ground truth GT, output edge distribution using Roberts, Sobel, Prewitt, Canny, LoG and $\kappa(\mathbf{p})$. $\sigma_{Canny} = 0.4$, $\sigma_{LoG} = 0.9$ and $\sigma_{EPR} = 1e5$

Table 2 Performances of several edge detection methods on synthetic image I_2.

Method	Roberts	Sobel	Prewitt	Canny $\sigma_{Canny} = 0.4$	LoG $\sigma_{LoG} = 0.9$	$\kappa(\mathbf{p})$ $\sigma_{EPR} = 1e5$
D_{KL}	9.98	9.70	9.84	9.40	11.82	4.53

5.3 Robustness to Gaussian Noise

In this experiment, the robustness to additive Gaussian noise is examined. The input image I_3 is obtained by adding to I_2 such a noise with standard deviation $\sigma_b = 50$. I_3, GT and the edge probability distributions are presented in Figure 3. The corresponding D_{KL} are gathered in Table 3. Again, $\kappa(\mathbf{p})$ produces the smallest divergence because non-relevant transitions are filtered out. Obviously, if the edge distributions were thresholded, Canny's approach would detect a larger part of the circle than $\kappa(\mathbf{p})$. Smaller values of σ_{EPR} lead to output images close to Canny's. It is important to remind that **it is only intended to validate EPR and not to introduce an edge detector**. For such a purpose, additional experiments involving image thresholding, more evaluation criteria and natural images are needed.

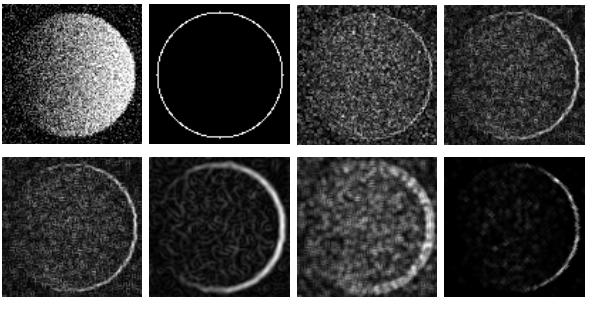


Fig. 3 From top-left to down-right: input image I_3, ground truth GT, output edge distribution using Roberts, Sobel, Prewitt, Canny, LoG and $\kappa(\mathbf{p})$. $\sigma_{Canny} = 1.5$, $\sigma_{LoG} = 1.2$ and $\sigma_{EPR} = 1e5$

Table 3 Performances of several edge detection methods on synthetic image I_3.

Method	Roberts	Sobel	Prewitt	Canny $\sigma_{Canny} = 1.5$	LoG $\sigma_{LoG} = 1.2$	$\kappa(\mathbf{p})$ $\sigma_{EPR} = 1e5$
D_{KL}	13.31	12.85	12.83	12.28	13.13	12.02

6 Conclusion

In this paper, a model for pixel representation (EPR) is proposed. This model is based on the belief function theory. The consistency of this model was proved through preliminary edge detection experiments. Indeed, the degree of conflict of neighbor pixels appears to be a relevant feature to assess the presence of an edge.

The approach is easy to implement and and does not require a heavy computation load. The goal behind this paper is to pave the way for future evidential image processing developments like denoising or pixel classification. Some additional processes and experiments will be investigated to introduce potentially a full edge detector. Furthermore, the EPR should also be extended to multi-component images and other belief masses than the degree of conflict may be exploited. The possibility to extend the model to non-additive noise should also be considered.

References

1. Bloch, I.: Some aspects of dempster-shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. Pattern Recognition Letters 17(8), 905–919 (1996)
2. Canny, J.: Finding edges and lines in images. Tech. rep., Cambridge, MA, USA (1983)
3. Dempster, A.: A generalization of bayesian inference. Journal of Royal Statistical Society B 30, 205–247 (1968)
4. Klein, J., Colot, O.: Singular sources mining using evidential conflict analysis. International Journal of Approximate Reasoning 52, 1433–1451 (2011)
5. Marr, D., Hildreth, E.: Theory of edge detection. Proceedings of the Royal Society of London, Series B, Biological Science 207(1167), 187–217 (1980)
6. Prewitt, J.: Object enhancement and extraction. Picture Processing and Psychopictorics, pp. 75–149. Academic Press, New York (1970)
7. Roberts, L.: Machine perception of 3-D solids. Optical and Electro-optical Information Processing. MIT Press (1965)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
9. Sobel, I.: Camera model and machine perception. Ph.D. thesis, Stanford University (1970)
10. Vannoorenberghe, P., Macaire, L., Colot, O.: Evidence-based pixel labeling for color image segmentation. In: Computer Vision Research Progress, ch. 11, pp. 279–296. Nova Science (2008)

Using Belief Function Theory to Deal with Uncertainties and Imprecisions in Image Processing

Benoît Lelandais, Isabelle Gardin, Laurent Mouchard, Pierre Vera, and Su Ruan

Abstract. In imaging, physical phenomena and acquisition system often induce an alteration of the information. It results in the presence of noise and partial volume effect corresponding respectively to uncertainties and imprecisions. To cope with these different imperfections, we propose a method based on information fusion using Belief function theory. First, it takes advantage of neighborhood information and combination rules on mono-modal images in order to reduce uncertainties due to noise while considering imprecisions due to partial volume effect on disjunctions. Imprecisions are then reduced using information coming from multi-modal images. Results obtained on simulated images using various signal to noise ratio and medical images show its ability to segment multi-modal images having both noise and partial volume effect.

1 Introduction

In imaging, two distinct problems lead to ambiguities from a spatial point of view: uncertain information due to noise and imprecise information due to lack of knowledge at the transition between areas. At this transition, the information carried by voxels is more ambiguous than the one suffering from noise. Both uncertainties and imprecisions have a negative effect on image processing.

Benoît Lelandais · Laurent Mouchard · Su Ruan
University of Rouen, LITIS EA 4108 - QuantIF, 22 bd Gambetta, 76183 Rouen, France
e-mail: benoit.lelandais@univ-rouen.fr, laurent.mouchard@univ-rouen.fr, su.ruan@univ-rouen.fr

Isabelle Gardin · Pierre Vera
Centre Henri-Becquerel, Department of nuclear medicine, 1 rue d'Amiens, 76038 Rouen, France & University of Rouen, LITIS EA 4108 - QuantIF, 22 bd Gambetta, 76183 Rouen, France
e-mail: isabelle.gardin@chb.unicancer.fr, pierre.vera@chb.unicancer.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 197–204.
springerlink.com

Belief function theory (BFT) [1, 2, 3] is particularly well suited to represent information from partial and unreliable knowledge. In [4, 5], authors propose to use BFT to reduce uncertainties and imprecisions using conjunctive combination of neighboring voxels. On one hand, it allows to reduce noise and on the other hand, to highlight conflicting areas mainly present at the transition between areas where PVE occurs due to the fact that information is extremely ambiguous in a spatial context. Therefore, results obtained by these authors allow to represent both segmented regions and contours.

BFT has the advantage to manipulate not only singletons but also disjunctions. This gives the ability of explicitly representing both uncertainties and imprecisions. One of the difficulties resides in the modeling of disjunctions, while they make it possible to take into consideration the lack of knowledge. In [6], author proposes to use fuzzy morphological operators to transfer for each voxel a part of belief on disjunctions according to its neighborhood. This method is interesting, but considers uncertainties and imprecisions in the same way.

By using BFT, our aim is two-fold: first, we reduce uncertainties due to noise, then imprecisions due to Partial Volume Effect (PVE) which corresponds to the lack of knowledge at the transition between areas. At first, our method operates a disjunctive combination followed by a conjunctive combination of neighboring information on mono-modal images. The disjunctive combination allows to transfer both uncertain and imprecise informations on disjunctions. Then, the conjunctive combination is applied to reduce uncertainties due to noise while maintaining representation of imprecise information at the boundaries between areas on disjunctions. In order to remove some imprecise informations, a multi-modal image fusion is also proposed. We take benefit from the complementarity of images to reduce imprecisions.

The method is used for the fusion of multi-modal PET (Positron Emission Tomography) medical images of the same patient using three radiotracers which give respectively information on tumor glucose metabolism, cell proliferation, and hypoxia (inadequate supply of oxygen). These images are of major interest for the treatment of lung cancer by radiotherapy, but need a relevant treatment considering both their important noise and partial volume effect.

First, we present our method, based on the fusion in mono-modal images, followed by the multi-modal fusion of informations. Then, the validation of the method is done on simulated data. Finally, the method is applied on multi-modal PET images.

2 Information Fusion Using Belief Function Theory for Reducing Uncertainties and Imprecisions

2.1 Fusion for Reducing Uncertainties While Considering Imprecision on Mono-Modal Images

Partial knowledge as uncertainties and imprecisions are taken into account by assigning Basic Belief Assignments (BBA) m over different subsets of the considered

frame of discernment $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$. m is defined as a mapping from 2^Ω to [0,1] verifying $\sum_{A \subseteq \Omega} m(A) = 1$.

From initial BBA, to reduce uncertainties while considering imprecisions, we propose to take benefit from both neighborhood information and combination operators [4, 5]. The way in which the neighborhood contribution has been carried out is as follow: let $\Phi(V_i)$ be a set of P voxels V_k (with $k \in \{1, \ldots, P\}$), surrounding a voxel V_i, and including V_i. Because of the different distances separating V_k and V_i, we propose to associate for each voxel in $\Phi(V_i)$ a coefficient α_k that depends on the distance separating it from V_i. It is computed by: $\alpha_k = \exp\left((V_k - V_i)^2/\sigma^2\right)$, with $\text{FWHM} = 2\sqrt{2\log 2}\sigma$ the Full Weight at Half Maximum corresponding to the spatial resolution of our images.

2.1.1 Disjunctive Combination of Neighboring Voxels

First, the influence of a voxel V_k from $\Phi(V_i)$ is weighted by the coefficient α_k. The BBA on $A \neq \varnothing$ and on $\varnothing$ can be calculated using the expressions:

$$\begin{aligned} m'_{V_k}(A) &= \alpha_k m_{V_k}(A), \qquad \forall A \neq \varnothing \\ m'_{V_k}(\varnothing) &= 1 - \alpha_k + \alpha_k m_{V_k}(\varnothing) \end{aligned} \tag{1}$$

Thus, the further away from V_i the voxel V_k is, the lower its contribution to the computation will be. The transfer to the empty set is interpreted as a non-commitment toward all the other hypotheses, and allows, before applying a disjunctive combination, to reduce the influence of $m_{V_k}(A)$ proportionally to α_k. The Fuzzy C-Means algorithm (FCM) [7] is used in order to initialize $m_{V_k}(A)$.

After this step, and in order to transfert uncertain and imprecise data to disjunctions, the following disjunctive combination is performed for each voxel:

$$\mathcal{M}_{V_i}(.) = \bigcirc\!\!\!\!\!\!\cup_{V_k \in \Phi(V_i)} m'_{V_k}(.) \tag{2}$$

It follows that nonzero masses are assigned to disjunctions. Higher they are, more different the informations carried by the neighboring voxels are. It is especially true on the edges between areas where PVE occurs and for voxels located in a very noisy environment. This operator, usually used when at least one source is reliable, can be used at that time. It is reasonable to assume that at least one of the voxels in the neighborhood gives a reliable information. After the disjunctive combination, we can assume that all sources become reliable since operator acts as a discounting of spatially ambiguous sources.

2.1.2 Estimation Using Disjunctive Combination in FCM

Disjunctive step allows to transfert, from initial BBA computed using FCM, uncertain and imprecise information on disjunctions. We also propose to integrate our disjunctive combination of neighboring voxels inside FCM algorithm. This process

is used for updating centroids and computing membership degrees with adapted data. After the fusion, the obtained data is less ambiguous.

2.1.3 Conjunctive Combination of Neighboring Voxels

To reduce uncertainties without impacting the ambiguities brought by the imprecisions, the opposite operation has been proposed. It consists in the conjunctive combination of neighboring voxels V_k discounted. First, the discounting is done according to the coefficient α_k by transferring a part of belief on the set Ω:

$$\begin{aligned} \mathcal{M}''_{V_k}(A) &= \alpha_k \mathcal{M}_{V_k}(A), \qquad \forall A \neq \Omega \\ \mathcal{M}''_{V_k}(\Omega) &= 1 - \alpha_k + \alpha_k \mathcal{M}_{V_k}(\Omega) \end{aligned} \tag{3}$$

The discount process allows to reduce the influence of voxels which are far from V_i before doing the conjunctive combination using Dempster's rule given by:

$$M_{V_i}(.) = \bigoplus_{V_k \in \Phi(V_i)} \mathcal{M}''_{V_k}(.) \tag{4}$$

Since all sources are reliable (thanks to disjunctive combination), it is appropriate to use this operator. This step allows to remove ambiguities due to noise, by transferring their belief on the singletons, while the voxels at the boundaries between areas remain represented on disjunctions.

2.2 *Fusion for Reducing Imprecisions Using Multi-modal Informations*

Having different information coming from other sources, we propose to take benefit from this information in order to reduce imprecisions due to PVE. Information coming from mono-modal image does not allow to reduce imprecisions. If two sources of information are available, reducing the imprecision with BFT is possible by using the conjunctive rule of combination. Let m_1 and m_2 be two fully reliable BBA. Their fusion is defined as follow:

$$m_1 \textcircled{\cap} m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \tag{5}$$

We propose an information fusion method based on conjunctive rule to deal with multi-modal images. Furthermore, we choose to use an external contextual knowledge to reduce imprecisions for our application on PET images (section 4).The external knowledge is learned from PET phantom images containing spheres whose volumes are known. It consists in applying the conjunctive combination of the current source with the external knowledge whose BBA has two focal elements: the subset A of Ω to reinforce and Ω.

3 Validation

The efficiency of our method is evaluated and compared to the method proposed in [6] on simulated images by measuring the recognition rates (rates of pixels correctly labelled) according to several Signal to Noise Ratio (SNR) varying from 1.5 to 6. The simulated images (Fig. 1) consist in a square surrounding by a background (images with two classes). The simulated images are blurred with a Gaussian filter whose FWHM vary according to the SNR, and noised with a Gaussian filter whose standard deviation is inversely proportional to the SNR. Note that BBA presented on Fig. 1(a) and (b) correspond to a simulated image with SNR of 5.

The proposed method is applied on the simulated images. At the end of the disjunctive combination of each pixel with its neighborhood, the belief masses are spread over the hypotheses $\{\omega_1\}$, $\{\omega_2\}$ and $\{\omega_1, \omega_2\}$ (Fig. 1(f), (g) and (h)). The belief of each pixel for which the information is ambiguous is mainly represented on the hypothesis $\{\omega_1, \omega_2\}$. The result, using the conjunctive combination of each pixel with its neighborhood is presented Fig. 1(i), (j) and (k). Inside areas, the method provides high beliefs in favor of $\{\omega_1\}$ and $\{\omega_2\}$. The uncertainties due to noise are therefore reduced. Within the transitions between areas, the belief is mainly represented on the hypothesis $\{\omega_1, \omega_2\}$, highlighting the imprecision.

For comparison, we present Fig. 1(c), (d) and (e) the result of the modeling using the method proposed in [6]. This method considers both noisy and fuzzy information

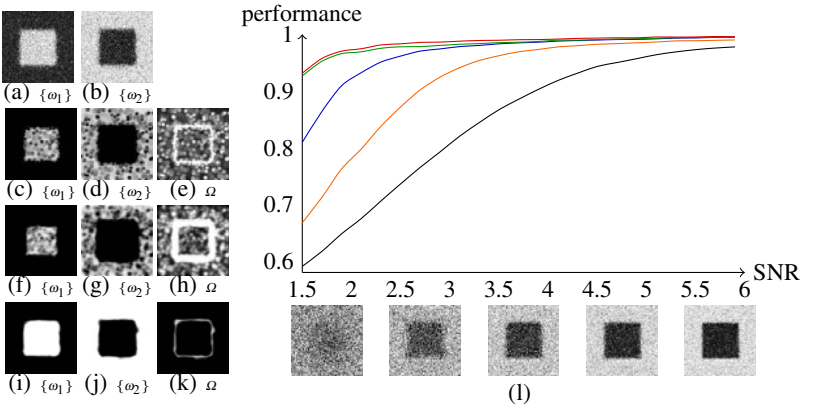


Fig. 1 Results of the fusion of mono-modal simulated images. (a) and (b) are initial BBA. (c), (d) and (e) show the BBA with the method proposed in [6]. (f), (g) and (h) present the BBA after the disjunctive combination. (i), (j) and (k) show the BBA applying then the conjunctive combination (our method). On (l) are presented the recognition rates according to different SNR on simulated images using only FCM (black curve), both FCM and the method proposed in [6] (orange curve), the conjunctive rules (blue curve), the disjunctive and the conjunctive rules (green curve), and using disjunctive rule integrated in FCM followed by conjunctive rule (red curve).

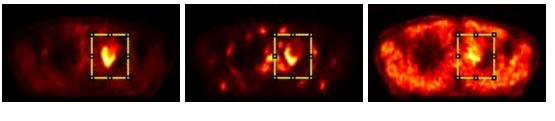
as imprecision. Note that the conjunctive combination of neighboring pixels is an important step of our method to reduce noise and transfer fuzzy information on disjunctions.

Fig. 1(1) presents the recognition rates using FCM, and applying our method, and the method proposed in [6] according to different Signal to Noise Ratio (SNR). To measure recognition rates for each SNR, the methods are here applied on two images having the same SNR, and results are then fused using the Dempster's rule in order to reduce both uncertainties and imprecisions. As we can see, our method gives better recognition rates than the method proposed in [6]. In addition, note that the conjunctive combination of neighboring pixels allows to improve the results. Moreover, when disjunctive rule is integrated in FCM, performances are lightly improved.

4 Application to Multi-modal PET Images for Functional Tumor Localization

The proposed method is also applied for multi-modal fusion of PET functional medical images (Fig. 2) to localize the tumor. These images are obtained after injection of a tracer specific to a studied function. From tracer FDG, FLT and FMISO, three type of PET images are obtained for a patient. Their characteristics are respectively glucose metabolism, cell proliferation and hypoxia (inadequate supply of oxygen). The FDG provides a good definition of the tumor target volume, especially ganglionic [8]. The FLT has a better tumor specificity than FDG [9] and lets us to envisage increasing the frequency of radiation therapy sessions on hyper-proliferative lesions. Finally, FMISO defines hypoxic tumors for which an irradiation dose escalation can be envisaged to improve the treatment [10].

The three PET images allow the distinction of areas that can be represented by five singletons, namely healthy tissue $\{N\}$ (*N*ormal), those with an important glucose *M*etabolism $\{M\}$, an important cell *P*roliferation $\{P\}$, a significant hypoxia $\{H\}$, and tissues with a *F*ull uptake $\{F\}$: where tissues need an increasing of both the radiation therapy frequency and the dose. For each image, estimation step as proposed has been applied in order to obtain degrees of belief over two hypotheses and their union as presented in Table 1.

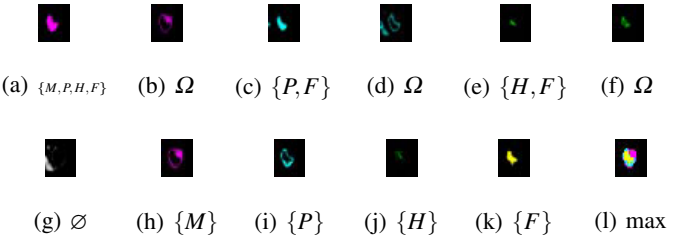


(a) FDG (b) FLT (c) FMISO

Fig. 2 Transverse slices for one patient with lung cancer. (a) Glucose metabolism PET image, (b) cell proliferation PET image, (c) hypoxia PET image. The area of interest (tumor lesion) is located in the rectangle.

Table 1 Hypotheses considered for PET images in order to fuse them coherently.

Image	Low uptake	High uptake	Ω
FDG	$\{N\}$	$\{M,P,H,F\}$	$\{N,M,P,H,F\}$
FLT	$\{N,M,H\}$	$\{P,F\}$	$\{N,M,P,H,F\}$
FMISO	$\{N,M,P\}$	$\{H,F\}$	$\{N,M,P,H,F\}$



(a) $\{M,P,H,F\}$ (b) Ω (c) $\{P,F\}$ (d) Ω (e) $\{H,F\}$ (f) Ω

(g) $\varnothing$ (h) $\{M\}$ (i) $\{P\}$ (j) $\{H\}$ (k) $\{F\}$ (l) max

Fig. 3 Images showing results of our method of fusion on multi-modal PET images. (a) to (f) show the BBA assigned to each voxel after the estimation step. (g) to (l) are the results applying the Dempster's rule of combination. (g) correspond to the areas of conflict. (h) to (k) are the plausibility corresponding to our hypotheses of interest. Finally, (l) is the segmented image using the maximum of plausibility.

On PET images, partial volume effect depends both on the size of the high uptake area and the contrast between areas. It becomes also important to reinforce BBA of high uptake areas according to its contrast and its volume. We chose to apply a reinforcement with a knowledge depending on contrast and volume. A learning is first carried out on PET images for which the high uptake volume is known in order to determine the parameters of reinforcing. This step allows to reduce imprecisions on mono-modal PET images.

Results obtained from multi-modal PET images (Fig. 2) are presented in Fig. 3. From Fig. 3(a) to (f) are presented BBA corresponding to high uptake tissues and imprecise information after applying our fuzzy clustering method on each image followed by the conjunctive combination of neighboring voxels. On one hand, we observe that noisy information is removed from areas corresponding to high uptake tissues. On the other hand, we can see that areas corresponding to partial volume effect and medium uptake are mainly assigned to the vacuous BBA. Fig. 3(g) to (k) present the result of the multi-modal PET image fusion using first the reinforcing and then the Dempster's rule [1]. They present both the BBA corresponding to the conflict, and the plausibility corresponding to our hypotheses of interest: $\{M\}$, $\{P\}$, $\{H\}$ and $\{F\}$. Note that the conflict corresponds to a high uptake in FLT and FMISO and a low uptake in FDG. Finally, Fig. 3(l) correspond to the segmented tumor using the maximum of plausibility. This image is of great interest for the radiotherapist in order to adapt dose deliverance according to the functional tumor tissues.

5 Discussion-Conclusion

Currently, in medical images, very few authors consider both spatial uncertainties and imprecisions in the information modeling with the BFT [6]. We propose to perform a fusion of neighboring information by a disjunctive combination followed by a conjunctive combination. This method allows to deal with both types of imperfection. In addition, we suggest to integrate the disjunctive combination in FCM in order to compute centroids only with certain and precise information. Finally, we propose to take benefit from prior knowledge in order to reduce imprecisions.

As shown from the results on simulated and medical images, the interest of our method is that the uncertainties due to noise are largely removed, and that the imprecision at the boundaries between regions is taken into account in the modeling. Moreover, considering large amount of noise, our method outperform a simple FCM and the method proposed in [6].

The method is generic since it can be applied whatever the distribution of initial beliefs is. In future work, we will test our method on a larger database to assess the robustness of the method, and on other types of images to confirm its genericity.

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. 38, 225–339 (1967)
2. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
3. Smets, P., Kennes, R.: The Transferable Belief Model. Artif. Intell. 66, 191–234 (1994)
4. Capelle, A.S., Colot, O., Fernandez-Maloigne, C.: Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information. Inf. Fusion 5(3), 203–216 (2004)
5. Zhang, P., Gardin, I., Vannoorenberghe, P.: Information fusion using evidence theory for segmentation of medical images. In: Int. Colloq. on Inf. Fusion, vol. 1, pp. 265–272 (2007)
6. Bloch, I.: Defining belief functions using mathematical morphology - Application to image fusion under imprecision. Int. J of Approx. Reason. 48(2), 437–465 (2008)
7. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
8. Gould, M.K., Kuschner, W.G., Rydzak, C.E., et al.: Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. Ann. Int. Med. 139, 879–892 (2003)
9. Xu, B., Guan, Z., Liu, C., et al.: Can multimodality imaging using 18F-FDG/18F-FLT PET/CT benefit the diagnosis and management of patients with pulmonary lesions? Eur. J. Nucl. Med. Mol. Imaging 38(2), 285–292 (2011)
10. Choi, W., Lee, S.W., Park, S.H., et al.: Planning study for available dose of hypoxic tumor volume using fluorine-18-labeled fluoromisonidazole positron emission tomography for treatment of the head and neck cancer. Radiother. Oncol. 97(2), 176–182 (2010)

Belief Theory for Large-Scale Multi-label Image Classification

Amel Znaidia, Hervé Le Borgne, and Céline Hudelot

Abstract. Classifier combination is known to generally perform better than each individual classifier by taking into account the complementarity between the input pieces of information. Dempster-Shafer theory is a framework of interest to make such a fusion at the decision level, and allows in addition to handle the conflict that can exist between the classifiers as well as the uncertainty that remains on the sources of information. In this contribution, we present an approach for classifier fusion in the context of large-scale multi-label and multi-modal image classification that improves the classification accuracy. The complexity of calculations is reduced by considering only a subset of the frame of discernment. The classification results on a large dataset of 18,000 images and 99 classes show that the proposed method gives higher performances than of those classifiers separately considered, while keeping tractable computational cost.

Keywords: Demspster-Shafer theory, multi-label classification, multi-modal classification, classifier fusion.

1 Introduction

Image annotation consists in describing an image content according to a finite number of concepts. This problem is usually posed as a set of binary classification tasks,

Amel Znaidia
CEA, LIST, Laboratory of Vision and Content Engineering
e-mail: amel.znaidia@cea.fr

Hervé Le Borgne
CEA, LIST, Laboratory of Vision and Content Engineering
e-mail: herve.le-borgne@cea.fr

Céline Hudelot
MAS Laboratory, Ecole Centrale de Paris
e-mail: celine.hudelot@ecp.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 205–212.
springerlink.com

which means to address both image description and visual concept learning. Concerning the first step, images are commonly described using only visual content such as color, texture or shape etc. However, in practice an important gap remains between visual descriptors and the semantic content of images [12].

Therefore, the use of multiple classifiers trained on different modalities (visual, textual ...) and features becomes more popular due to the fact that classifiers are different and informative [5, 7]. Thus, the fusion of their decisions can yield to higher performance than the best individual classifier [4].

Most commonly, straightforward fusion approaches, such as majority voting, maximum and averaging [13] have been used in the literature. According to Tax *et al.* [13] simple average is the optimal linearly combining rule, only if the individual classifiers exhibit both identical performances and correlations between estimation errors. Otherwise, Dempster-Shafer theory [11] is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers. However, it suffers from a high computational cost, in particular when the number of classes (*i.e* the frames of discernment) is large. To encounter this limitation, Denoeux *al.* [2] proposed a method to reduce the complexity of manipulating and combining mass functions, when belief functions are defined over a suitable subset of the frame of discernment equipped with a lattice structure. This approach was applied for multi-label classification based on the Evidential KNN classifier. For a problem with $\mathbf{C}$ classes, this method reduces the complexity from $2^{2^{\mathbf{C}}}$ to $3^{\mathbf{C}}+1$. Althougth such a reduction is impressive, the problem remains intractable when C is above 10, that is quite common for a multimedia classification problem, for which $\mathbf{C}$ can reach 100 or 1000.

The most similar prior work is [9], which combine textual and visual classifiers based on Dempster's rule to improve the classification accuracy. However, their system was applied for single-label classification task, for a small dataset ($\approx 1,200$ images) and only for *six* classes of emotions.

In this work, we aim at improving the classification accuracy based on classifier fusion in the Dempster-Shafer theory to handle the uncertainty and the conflict that can exist between different classifiers and to assess the discrepancy between various sources of information. The major difference between our work and aforementioned efforts is that we address the problem of combination in a multi-label classification task for a large problem: to the best of our knowledge, this is the first attempt to apply Dempster theory for a multimodal multi-label image classification for a large dataset ($\approx 18,000$ images) and a large variety of categories simultaneously (scene, event, objects, image quality and emotions ≈ 99 concepts). First, we convert the classifier output probabilities into consonant mass functions using the inverse pignistic transform [3]. Secondly, these mass functions are combined in the belief theory using Dempster's rule [11]. Since Average rule has been widely used in the literature, and it outperforms other conventional methods (Maximum, Product, Majority voting), we use it as a baseline to compare with the Dempster's rule.

The remainder of the paper is organized as follows. The background on belief functions is first recalled in section 2. The proposed approach for large scale

multi-label image classification is presented in section 3, and experimental results are reported and discussed in section 4. Section 5 concludes this paper.

2 Basics of Dempster-Shafer Theory

In Dempster-Shafer (DS) theory [11], a *frame of discernment* Ω is defined as the set of all hypothesis in a certain domain. A basic belief assignement (BBA) is a function m that defines the mapping from the power set of Ω to the interval $[0,1]$ and verifies:

$$m : 2^{\Omega} \rightarrow [0,1] \tag{1}$$

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \tag{2}$$

The quantity $m(A)$ can be interpreted as a measure of the belief that is commited exactly to A, given the available evidence. A subset $A \in 2^{\Omega}$ with $m(A) > 0$ is called a *focal element* of m. In DS theory, two functions of evidence can be deduced from m and its associated focal elements, belief function Bel and plausibility function Pl. $Bel(A)$ is the measure of the total belief committed to a set A. The belief function is defined as a mapping $Bel : 2^{\Omega} \rightarrow [0,1]$ that satisfies $Bel(\emptyset) = 0, Bel(\Omega) = 1$ and for each focal element A, we have:

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \tag{3}$$

The *plausibility* of A, $Pl(A)$, represents the amounts of belief that could potentially placed in A and defined as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{4}$$

2.1 Dempster's Combination Rule

When there are many sources of information defined on the same frame of discernment, the mass functions from different sources are combined under the normalized Dempster's combination rule [11].

$$m_{1-2}(A) = m_1 \oplus m_2 = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}, & \forall\ A \subseteq\ \Omega,\ A \neq \emptyset \\ 0 & if\ \ A = \emptyset \end{cases} \tag{5}$$

where $k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ represents the degree of conflict between the two sources. If $k = 1$ the two evidences are in conflict and they can not be combined.

3 Proposed Multi-label Classification System

In the context of multi-label and multi-modal classification problem, each image can belongs to one or more than one class. Formally, let $\Omega = \{w_1, ..., w_C\}$ be the set of labels or classes. The frame of discernment of the multi-label extended DS theory is not the set of all possible single hypotheses but its power set $\Theta = 2^{|\Omega|}$. Given a training set $T = \{(X_1, Y_1), ..., (X_N, Y_N)\}$ of N labelled images, where $X_i = \{x_i^1 ... x_i^L\}$ represents the feature vector of image I_i extracted from L modalities and Y_i the corresponding set of labels, our goal is to predict the set of lables that describe the image content. The flowchart of the proposed system is presented in Figure 1. Assume that we have Q classifiers, denoted by $\psi_1, \psi_2, ...\psi_Q$ to be combined. Given an input image I, each classifier ψ_i produced an output $\psi_i(I)$ defined as :

$$\psi_i(I) = [s_{i1}, ..., s_{iC}] \tag{6}$$

where s_{ij} indicates the degree of confidence in saying that 'image I belongs to class w_j according to classifier ψ_i'. First, classifier output are normalized to obtain a probability distribution p_i over Ω as follows:

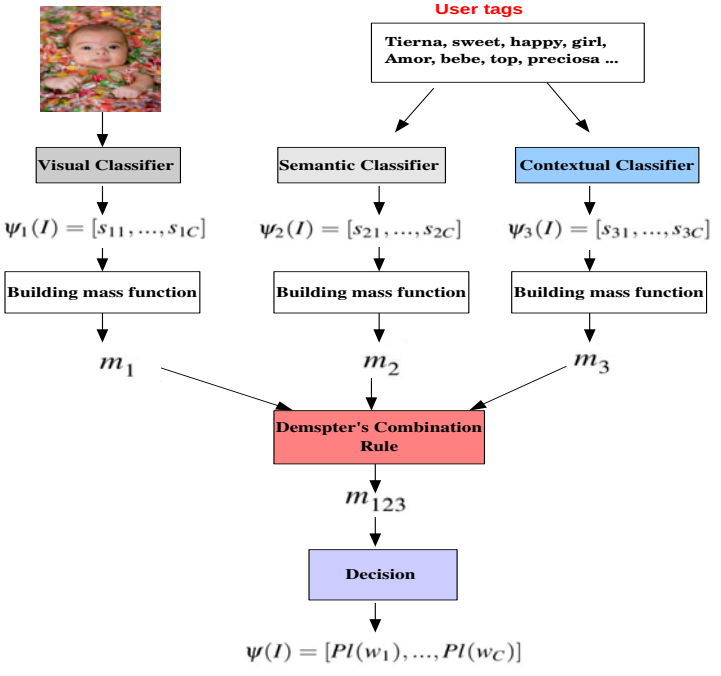


Fig. 1 Flowchart of the proposed system. First, the classifier output scores ψ_i are normalized to sum to one. Secondly, the obtained probabilities are transformed into mass function using the inverse pignistic transform. A combination is performed to obtain the final mass function, used to compute the plausibility fo decision making.

$$p_i(w_j) = \frac{s_{ij}}{\sum_{k=1}^{C} s_{ik}}, \quad for \;\; j = 1, ..., C \tag{7}$$

For each classifier ψ_i, the element of Ω are ranked by decreasing probabilities such that $p(w_1) \geq p(w_2) \geq ... \geq p(w_{|\Omega|})$. The class label of an instance may be represented by a variable Y taking values in $\Theta = 2^{|\Omega|}$. Thus, expressing partial knowledge of Y in the Dempster-Shafer framework may imply storing 2^{2^C} numbers. Based on this ordering, instead of considering the whole power set of Θ, we will focus on a smaller subset $R(\Omega)$ defined by:

$$R(\Omega) = \{A_k = \{w_1, ..., w_{k+1}\}, \forall\, k = 1, ..., |\Omega| - 1\} \tag{8}$$

The size of this subset is $|\Omega| - 1$, it is thus much smaller than 2^{2^C} while being rich enough to express evidence because we consider only the most probable subsets. Secondly, we convert the obtained probabilities into consonant mass functions using the inverse pignistic transform [3]. The consonant mass function derived from these probabilities verifies :

$$m : 2^{\Omega} \rightarrow [0, 1], \quad \sum_{A_k \in 2^{\Omega}} m(A_k) = 1 \tag{9}$$

$$\begin{aligned} m(\{w_1, w_2, ..., w_i\}) &= i \times \; [p(w_i) - p(w_{i+1})] \; \forall \; i \; < \; |\Omega| \\ m(\{w_1, w_2, ..., w_{|\Omega|}\}) &= |\Omega| \times p(w_{|\Omega|}) \\ m(X) &= 0 \;\; \forall \, X \notin R(\Omega). \end{aligned} \tag{10}$$

In this work, we choose to combine the obtained consonant mass functions from different classifiers using the normalized Dempster's rule [11]. Other combination rules can be used [10]. Let m_i be the mass function of the source i, the combination of n mass function (corresponding to n classifiers) is defined according to Dempster's combination rule as follows:

$$m_{1-n}(A) = \begin{cases} \dfrac{\sum\limits_{\cap_{k=1}^{n} b_k = A} \prod\limits_{i=1}^{n} m_i(b_i)}{1 - \sum\limits_{\cap_{k=1}^{n} b_k = \emptyset} \prod\limits_{i=1}^{n} m_i(b_i)}, \; \forall A \subseteq \; \Omega, \; A \neq \emptyset, b_k \in R_k(\Omega) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad if \;\; A = \emptyset \end{cases} \tag{11}$$

Let $\hat{Y}$ be the predicted label set for instance x. To decide whether to include each class or not, we compute the degree of plausibility $Pl(w_j)$ that the true label set Y contains the label w_j, and the degree of plausibility $Pl(\bar{w_j})$ that it does not contain the label w_j using formula (4). We then define $\hat{Y}$ as:

$$\hat{Y} = \{w_j \in \Omega | Pl(w_j) \geq Pl(\bar{w_j})\} \tag{12}$$

4 Experimental Results

4.1 Dataset and Experimental Setup

The Dataset used in our experiments is the MIR Flickr dataset [6] containing 8,000 images for training and 10,000 for testing belonging to 99 concept classes. These concepts describe the scene 'indoor, outdoor, landscape...', depicted objects 'car, animal, person...', the representation of image content 'portrait, graffiti, art...', events 'travel, work...', or quality issues 'overexposed, underexposed, blurry...' and emotions 'funny, cute, nice, scary ... '. Figure 2 shows samples of images taken from the dataset with their annotated concepts.

Features We used two textual descriptors and one visual descriptor. The textual descriptor is based on semantic similarity between tags and visual concepts. Two distances were used: one based on the Wordnet ontology and one based on social networks. Each feature vector is of size 99 (the number of concepts). The visual component considers various local and global features, such as colour and edge features. The visual feature vector is of size 890. More details about the used features can be found in [14]. Each feature vector was used to train a classifier using the Fast Shared Boosting algorithm [8]. Three measures are used to test the performance of the individual classifiers and the different combinations: Mean Average Precision (MAP), Equal Error Rate (ERR) and Area Under Curve (AUC).

4.2 Results and Discussions

Table 1 displays the performances of individual classifiers and the two considered combination rules in terms of MAP, ERR and AUC. These results show that individual classifiers exhibit identical performances with a small superiority to the contextual classifier. Since Average rule has been widely used in the literature, and it outperforms other conventional methods (Maximum, Product, Majority voting), we will use it as a baseline to compare to the Dempster's rule.

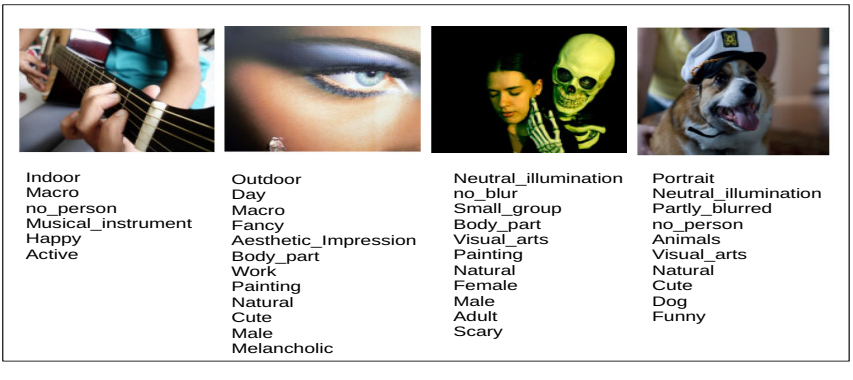


Fig. 2 Samples of images taken from the dataset with their annotated concepts.

Table 1 Comparative Performance of individual classifiers in terms MAP, ERR and AUC.

Classifier	Visual Classifier	Contextual Classifier	Semantic Classifier	Dempster's rule	Average rule
MAP	29.86	32.13	29.24	39.05	**40.21**
EER	28.93	31.50	35.69	26.21	**24.64**
AUC	77.59	74.32	68.44	80.79	**82.29**

Table 2 Comparative Performance of individual classifiers, Dempster, Average and the ImageClef 2011 Winner [1] for some challenging classes in terms of Mean Average Precision (MAP).

Classes	Visual	Contextual	Semantic	Dempster	Average	ImageClef 2011 Winner [1]
Travel	18.85	14.78	17.55	**22.12**	14.57	16.72
Technical	08.19	06.37	04.52	**12.85**	07.24	08.51
Boring	07.28	07.78	07.63	**15.88**	08.79	09.94
Bird	17.55	51.71	56.08	**61.52**	58.77	58.71
Insect	14.26	47.84	46.44	**58.08**	53.12	45.21
Airplane	05.36	44.36	42.53	**61.66**	59.32	22.93
Skateboard	00.27	10.29	21.54	**28.42**	11.46	00.56
Scary	18.46	08.31	14.10	**19.02**	11.29	16.39

By comparing these results, we can see that the combination of classifiers for both Dempster's rule and average rule gives better results than the best individual classifier. We obtain a gain of $\approx 10\%$ in terms of classification accuracy and consequently, reducing the classification error by $\approx 9\%$. For this dataset, we observe that the average rule achieve slightly better performances. These results may be explained by the performance of the individual classifiers which exhibit both identical performances and correlations between estimation errors. In addition, we train individual classifiers with unbalanced data over classes which can generate unreliable confidences (*e.g.* caused by a small training set or by overtraining).

The average rule is hardly ever theoretically optimal, but performs sometimes surprisingly good except for some classes as shown in Table 2. For these challenging classes, Dempster'rule performs much better than the average rule especially when considering ensembles of 'good' and 'bad' classifiers, then using the average rule to combine the classification results will not be a good choice. We compare Dempster's rule to the ImageClef 2011 Winner [1] for these classes. The proposed method outperforms the state of art [1] for such type of classes. We can notice that the Belief theory seems to offer a significant advantage to such situations. It is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers.

5 Conclusion

In this paper, we presented a system for combining classifiers using Belief theory for large-scale multi-label image classification. When individual classifiers present

similar performances, results have shown that using simple rules such as averaging can be a good choice. While, for conflicting classifiers, the Belief theory seems to be an interesting framework to handle the uncertainty and the conflict that can exist between different classifiers. One direction for future research is to take into account the classifier reliability while combining. An additional direction is to construct mass functions directly in the classifiers.

References

1. Binder, A., Samek, W., Kloft, M., Müller, C., Müller, K.-R., Kawanabe, M.: The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef 2011 photo annotation task. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
2. Denoeux, T., Masson, M.: Evidential reasoning in large partially ordered sets. Annals of Operations Research (May 2011)
3. Dubois, D., Prade, H., Smets, P.: New Semantics for Quantitative Possibility Theory. In: Benferhat, S., Besnard, P. (eds.) ECSQARU 2001. LNCS (LNAI), vol. 2143, pp. 410–421. Springer, Heidelberg (2001)
4. Duin, R.P.W.: The combining classifier: To train or not to train? In: ICPR (2), pp. 765–770 (2002)
5. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 902–909 (June 2010)
6. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: MIR 2008: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval. ACM, New York (2008)
7. Kawanabe, M., Binder, A., Muller, C., Wojcikiewicz, W.: Multi-modal visual concept classification of images via markov random walk over tags. In: Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision, WACV (2011)
8. Le Borgne, H., Honnorat, N.: Fast shared boosting for large-scale concept detection. Multimedia Tools and Applications, 1–14 (2010)
9. Liu, N., Dellandréa, E., Tellez, B., Chen, L.: Associating textual features with visual ones to improve affective image classification. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 195–204. Springer, Heidelberg (2011)
10. Quost, B., Masson, M.-H., Denoeux, T.: Classifier fusion in the dempster–shafer framework using optimized t-norm based combination rules. Int. J. Approx. Reasoning 52, 353–374 (2011)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1349–1380 (2000)
13. Tax, D.M., van Breukelen, M., Duin, R.P., Kittler, J.: Combining multiple classifiers by averaging or by multiplying? Pattern Recognition 33(9), 1475–1485 (2000)
14. Znaidia, A., Borgne, H.L., Popescu, A.: Cea list's participation to visual concept detection task of imageclef 2011. In: CLEF (Notebook Papers/Labs/Workshop) (2011)

Facial Expression Classification Based on Dempster-Shafer Theory of Evidence

Mohammad Shoyaib[1], M. Abdullah-Al-Wadud[2], S.M. Zahid Ishraque[1], and Oksam Chae[1]

Abstract. Facial expression recognition is a well discussed problem. Several machine learning methods are used in this regard. Among them, Adaboost is popular for its simplicity and considerable accuracy. In Adaboost, decisions are made based on the weighted majority vote of several weak classifiers. However, such weighted combination may not give expected accuracy due to the lack of proper uncertainty management. In this paper, we propose to adopt the Dempster Shafer theory (DST) of Evidence based solution where mass values are calculated from *k*-nearest neighboring feature information based on some distance metric, and combined together using DST. Experiments on a renowned dataset demonstrate the effectiveness of the proposed method.

1 Introduction

Facial expression generally conveys information, from which the state of the mind of a person may be inferred. The rapid development of technologies facilitates the consumer devices to incorporate different types of applications related to face images. Among them, facial expression recognition (FER) has become an active research area over the last two decades due to its diversified application areas. The main focuses of FER based researches are the appropriate representation of different expressions and their proper classification. Selection of appropriate and small number of features to represent the facial expressions improves the classification accuracies. Again, a suitable classifier also increases the overall performance.

Several promising facial expression recognition systems (FERSs) have already been proposed [16], [11]. Based on the features used to represent the expression, FERSs can generally be categorized into geometrical feature-based approaches

Mohammad Shoyaib · S.M. Zahid Ishraque · Oksam Chae
Kyung Hee University, Seocheon, Yongin, Gyonggi, Korea
e-mail: shoyaib@khu.ac.kr, zahidishraque@khu.ac.kr, oschae@khu.ac.kr

M. Abdullah-Al-Wadud
Hankuk University of Foreign Studies, Yongin, Gyonggi, Korea
e-mail: wadud@hufs.ac.kr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 213–220.
springerlink.com

and appearance-based approaches [18]. The geometrical feature-based approaches rely on the detection of a set of fiducial points [13], face feature contours [7] or active shape model [10] which is usually followed by tracking of the detected points or shapes. However, the computational cost is very high for these methods [2]. Further, they usually require accurate facial feature detection and tracking, which may not be feasible in many situations [16]. On the other hand appearance-based approaches usually use partial or whole facial textures to identify different expressions. Among the appearance based methods Gabor and local binary pattern (LBP) based FERS are popular for their better performances. In this paper, we adopt LBP based method due to its success in various texture based classification and face analysis work. Further, LBP is computationally simple and robust in monotonic illumination change. We extract LBP features from the whole face image and analyze its performances under evidential theory.

For facial expression classification Adaboost [6] is largely used [8] [14]. It is a well-known ensemble learning algorithm and can be used either as a classifier or as a feature selector. In Adaboost, every weak classifier offers binary decisions (1 or -1) regarding all the classes. Generally, it constructs a strong classifier, $H(x)$, as a linear combination of T weak classifiers, $h_t(x)$, and is given by

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right), \tag{1}$$

where $t = 1, 2, \ldots, T$, and α_t is a weight that indicates the importance of the corresponding $h_t(x)$. The final decision is thus made based on the weighted majority vote of the T weak classifiers.

A facial expression may not always be a perfectly distinguishable one. It may rather be a combination of different expressions. A well-designed statistics-based method might offer a good solution in this case. Again, a set of features generated from a given image may not always be adequate to express all the variations in expressions. This may also lead to uncertainty in expression detection. However, the weighted majority voting scheme of Adaboost may fail to handle these uncertainties. To solve this problem, we propose to use the Dempster-Shafer theory of evidence, which offers a powerful and flexible framework for representing and handling uncertainties and thus helps to overcome the aforementioned limitations.

The rest of the paper is organized as follows. Section 2 includes a short overview of the Dempster-Shafer theory of evidence. The Proposed method is described in Section 3. Section 4 shows some comparative results and Section 5 concludes the paper.

2 Dempster-Shafer Theory of Evidence

The Dempster-Shafer theory of evidence (DST) [15] uses a *frame of discernment*, which is defined as a set of mutually exclusive and collectively exhaustive hypotheses denoted by Θ. The power set of all possible subsets of Θ, including itself and the empty set ϕ, is 2^{Θ}. A mass function m: $2^{\Theta} \rightarrow [0,1]$ is a function satisfying (2).

$$\left.\begin{aligned} &m(\phi)=0 \\ &m(S)\geq 0, \forall_{S\subseteq\Theta} \\ &\sum_{S\subseteq\Theta} m(S)=1 \end{aligned}\right] \tag{2}$$

Here, $m(S)$ represents the belief reflecting how strongly S is supported. The mass values assigned to Θ is called the degree of ignorance, and the subsets S of Θ with non-zero mass values are called the focal elements. Belief (*bel*) and plausibility (*pl*) are two other common evidential measures, which are derived in (3) and (4), respectively.

$$bel(S)=\left.\sum_{T\subseteq S, T\neq\phi} m(T)\right] \tag{3}$$

$$pl(S)=\left.\sum_{S\cap T\neq\phi} m(T)\right] \tag{4}$$

where S and T are subsets of Θ.

Dempster's rule of combination can fuse the mass functions m_i obtained from n Sources of information according to (5).

$$m(S)=\frac{\sum_{S_1\cap\ldots\cap S_n=S}\prod_{i=1}^{n} m_i(S_i)}{1-K}, \tag{5}$$

where K represents the degree of conflict given by

$$K=\sum_{S_1\cap\ldots\cap S_n=\phi}\prod_{i=1}^{n} m_i(S_i).$$

There are several ways of taking the final decision using DST framework. For instance, decision can be made by choosing the hypothesis with the maximum mass, belief, plausibility or using pignistic probability distribution [17, 4].

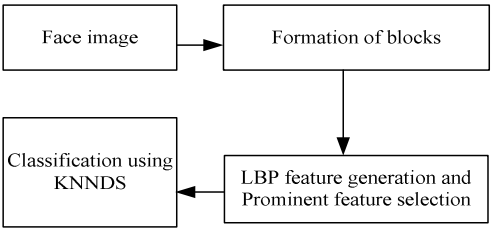


Fig. 1 The overview of the proposed expression recognition method

3 The Proposed Method

The proposed FERS consists of two main steps: feature generation – using LBP and selection of the prominent features by Adaboost; classification – using a DST based framework. The overall proposed framework is depicted in Fig. 1.

3.1 Feature Generation

In our proposal, we use local binary patterns as the feature. A local binary pattern (LBP) is a binary code defined at a pixel, c, with respect to its neighboring pixels in a grayscale image [12]. Therefore, the LBP at c for n neighbors, which are located at uniform distance on a circle centered at c with radius r, is given by

$$LBP_{n,r}(x_c, y_c) = \sum_{l=0}^{n-1} q(g_l - g_c)2^l, \quad q(a) = \begin{cases} 1 & if\ a \geq 0 \\ 0 & otherwise \end{cases} \tag{6}$$

where (x_c, y_c) is the pixel co-ordinate of c, and g_c and g_l are the intensities of c and the l^{th} neighboring pixel, respectively. An LBP code thus encodes local micro-patterns at a pixel such as edge, corner and line-end

These LBP codes of an image are then used to build a histogram which represents a feature vector of the image. The k^{th} component of the histogram, H^k, for an image of size $M \times N$ is derived by

$$H^k = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \psi_{i,j}^k, \quad \psi_{i,j}^k = \begin{cases} 1 & if\ LBP_{n,r}(i,j) = k \\ 0 & otherwise. \end{cases} \tag{7}$$

It is observed that all the variations in LBP codes are not necessary to represent the most of the available patterns in the nature. Thus following [12] we use uniform LBP patterns. In the histogram, each of the uniform LBP codes is placed in separate bin, and all non-uniform LBP codes are placed in a single bin.

For generating the patterns, we follow the similar method describe in [1]. For this we first divide the whole face into few blocks and then generate LBP histogram for each block using (7). For different facial expressions, the changes in facial features in all the parts of the face are not usually same. Hence, they cannot equally contribute for discriminating facial expressions. To reflect the significance of the features coming from different blocks, we multiply the accumulations in histogram bins of every block with some predefined weights as suggested in [1]. We then concatenate all the resultant histograms to form the feature vector to represents an expression.

During the training phase, the feature vectors of all the expressions are fed into the multiclass Adaboost, which then selects N prominent bins having better discrimination ability among all the bins in the feature vector. We use these selected bins to form the final feature vector for classification.

3.2 Feature Classification

For classification, we use the Dempster-Shafer theory of evidence-based approach. The reason behind choosing this approach is its capability to handle the uncertainties that might arise due to the similarity of different expression features, and the distributions of these features for different expression may overlap.

For using DST based method one of the foremost challenge is to find a way to calculate the mass values. In this case, we adopt an approach proposed by Denoeux [3] (here we named as KNNDS). In KNNDS, the mass values are calculated from the *k*-nearest neighboring patterns according to a distance metric for a given test pattern. The evidences from the neighbors are then combined using DS rule of combination.

Let $\Theta = \{C_1, C_2, \ldots, C_M\}$ denote M different expression classes. Consider that $\mathbf{Y}$ is the feature vector to be classified and $\mathbf{N_k}$ is the set of its k-nearest neighbors in the training set. Here any $\mathbf{Y_i} \in \mathbf{N_k}$ may belongs to a class $C_q \in \Theta$. This membership of Y_i can provide piece of evidence to increase our belief that $\mathbf{Y}$ belongs to C_q. This evidence is represented by mass value m_i as follows.

$$\begin{aligned} m_i(\{C_q\}) &= \alpha_q^i \\ m_i(\Theta) &= 1 - \alpha_q^i \end{aligned} \tag{8}$$

All the other values in m_i are 0. The mass α_q^i is chosen as a decreasing function of the Euclidean distance d^i between $\mathbf{Y}$ and $\mathbf{Y_i}$

$$\alpha_q^i = \alpha_0 \exp(-\gamma_q^2 (d^i)^2) \tag{9}$$

where γ_q is a parameter associated to class C_q and α_0 is fixed. Such pieces of evidences are combined using Dempster's rule of combination, and $\mathbf{Y}$ is classified to the class, for which the pignistic probability is maximum.

4 Experimental Results

In this section, we first discuss about the data that we use for our experiments, and then present the experimental results followed by a discussion.

4.1 Source of Experimental Data

According to Ekman and Friesen, six basic emotions, namely joy, disgust, sadness, anger, fear and surprise, can be universally recognized [5] (including neutral expression it becomes seven). To evaluate the performances, we take 408 image sequences of 96 subjects the Cohn-Kanade (C-K) database [9]. For six-class recognition, we pick three peak expression images from every sequence, which results in 1224 images. For seven-class, we include the first neutral expression frame, and thus there are 1632 images in total. For generating the results, we

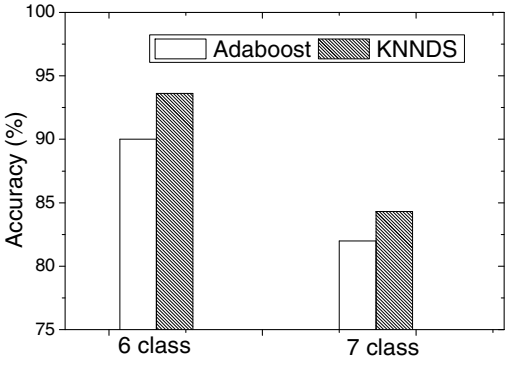


Fig. 2 Recognition accuracies (%) on C-K database

Table 1 Confusion Matrix for Six-Class Expression (KNNDS)

	A	**D**	**F**	**J**	**S**	**G**
Anger (A)	89.66	3.44	0	0	6.9	0
Disgust (D)	0	99	0	1	0	0
Fear (F)	0	4	84	8	0	4
Joy (J)	0	0	3.7	96.3	0	0
Sadness (S)	1	1	0	0	98	0
Surprise (G)	0	3.33	0	0	0	96.67
Mean	93.94					

perform 10-fold cross validation on our dataset. For LBP codes, we take eight neighbors at two pixels apart ($r = 2$). For KNNDS, we take $k = 10$.

4.2 Results

In this section, we analyze the performances of KNNDS for expression classification. We first compare the performance of KNNDS with Adaboost, and the results for both six and seven class expression recognition are shown in Fig. 2.

Table 2 Confusion Matrix for Six-Class Expression (Adaboost)

	A	**D**	**F**	**J**	**S**	**G**
Anger (A)	86.21	0	0	0	13.79	0
Disgust (D)	5	85	0	5	0	5
Fear (F)	12	0	76	8	0	4
Joy (J)	0.14	0	2.3	97.56	0	0
Sadness (S)	10.53	0	0	0	89.47	0
Surprise (G)	0	0	0	0	3.33	96.67
Mean	88.49					

From Fig. 2, we can observe that in both six and seven class expression recognition, KNNDS outperforms the Adaboost. To analyze it in more detail, we build up the confusion matrices as shown from Tables 1 to 4. The data in the tables show that KNNDS gives much better outcomes as compared to the Adaboost. The principle reason behind this better outcome is that the uncertainties in classifications by weak classifiers are handled well in Dempster-Shafer-based approach.

Table 3 Confusion Matrix for Seven Class Expression on C-K Database (KNNDS)

	A	**D**	**F**	**J**	**S**	**G**	**N**
Anger (A)	76.47	0	0	0	0	11.76	11.76
Disgust (D)	0	80	0	0	0	0	12
Fear (F)	0	0	86.96	4.35	0	0	8.7
Joy (J)	0	0	3.23	86.12	0	3.34	7.31
Neutral (N)	0	0	0	0	83.3	0	16.7
Sadness(S)	0	0	0	0	0	91.2	8.8
Surprise(G)	3.9	0	3.7	3.7	5.3	0	83.4
Mean	84.93						

Table 4 Confusion Matrix for Seven Class Expression on C-K Database (Adaboost)

	A	**D**	**F**	**J**	**S**	**G**	**N**
Anger (A)	52.94	0	0	0	5.88	5.88	35.29
Disgust (D)	10	80	0	0	0	0	10
Fear (F)	4.35	0	82.61	4.35	0	0	8.70
Joy (J)	0	0	0	96.77	0	0	3.23
Neutral N	5.88	0	0	0	76.47	5.88	11.76
Sadness (S)	4	0	0	0	0	96	0
Surprise(G)	0	0	3.7	0	3.7	0	92.59
Mean	82.48						

From the aforementioned tables, we can observe that the individual class accuracies of KNNDS are acceptable. However, in case of Adaboost, there are many variations in the accuracies (for example, 96.77% for Joy, but 52.94% for anger in Table 4).

5 Conclusion

In this paper, we have investigated an appearance based facial expression recognition method using Dempster-Shafer theory of evidence. Here we use *k*-nearest neighbors to calculate the mass values and combine them using DST to identify the most probable expression present in an image. This method can also be extended to expression recognition in video images.

We use an existing mass generation method in this paper. By incorporating uncertainty management with the weak classifiers of Adaboost, it might be possible

to find an improved mass generation methodology to achieve better accuracy in expression recognition. We leave this as our future work.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0017151).

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004, Part I. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
2. Bacivarov, I., Corcoran, P., Ionita, M.: Smart cameras: 2D affine models for determining subject facial expressions. IEEE Trans. on Consumer Electronics 56(2) (2010)
3. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster–Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25(05), 804–813 (1995)
4. Denoeux, T.: A neural network classifier based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics A 30(2), 131–150 (2000)
5. Ekman, P.: Facial expressions. In: Dalgleish, T., Power, M. (eds.) Handbook of Cognition and Emotion, pp. 301–320. Wiley, New York (1999)
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proc. of the European Conference on Computational Learning Theory, Euro-COLT (1995)
7. Hammal, Z., Couvreur, L., Caplier, A., Rombaut, M.: Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. International Journal of Approximate Reasoning 46(3), 542–567 (2007)
8. Jung, S.U., Kim, D.H., An, K.H., Chung, M.J.: Efficient rectangle feature extraction for real-time facial expression recognition based on adaboost. In: IEEE/RSJ International Conference on Intelligent Robotics and Systems, IROS (2005)
9. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proc. of IEEE Inter. Conference on Automatic Face & Gesture Recognition, pp. 46–53 (2000)
10. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Trans. Image Processing 16(1), 172–187 (2007)
11. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J.: Dynamics of facial expression extracted automatically from video. Image and Vision Comp. 24, 615–625 (2006)
12. Ojala, T., Pietikainen, M., Maenpaa, T.T.: Multiresolution gray-scale and rotation invariant texture classification with local binary pattern. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 971–987 (2002)
13. Pantic, M., Patras, I.: Dynamics of facial expressions—recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Systems, Man, and Cybernetics. 36(2), 433–449 (2006)
14. Peng, Y., Qingshan, L., Metaxas, D.N.: Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition. In: CVPR 2007, pp. 1–6 (2007)
15. Shafer, G.: A Mathematical Theory of Evidence. P.U. Press, Princeton (1976)
16. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image and Vision Computing 27(6), 803–816 (2009)
17. Smets, P.: Decision Making in the TBM: the Necessity of the Pignistic Transformation. Int. J. Approximate Reasoning 38, 133–147 (2005)
18. Xiao, R., Zhao, Q., Zhang, D., Shi, P.: Facial expression recognition on multiple manifolds. Pattern Recognition 44, 107–116 (2011)

Compositional Models in Valuation-Based Systems

Radim Jiroušek and Prakash P. Shenoy

Abstract. Compositional models were initially described for discrete probability theory, and later extended for possibility theory, and Dempster-Shafer (D-S) theory of evidence. Valuation-based systems (VBS) can be considered as a generic uncertainty framework that has many uncertainty calculi, such as probability theory, a version of possibility theory where combination is the product t-norm, Spohn's epistemic belief theory, and D-S belief function theory, as special cases. In this paper, we describe compositional models for the VBS framework using the semantics of no-double counting. We show that the compositional model defined here for belief functions differs from the one studied by Jiroušek, Vejnarová, and Daniel. The latter model can be described in the VBS framework, but with a combination operation that is different from Dempster's rule.

1 Introduction

Compositional models were initially described for discrete probability theory [4, 5]. They were later extended by Vejnarová [14] for possibility theory, and in [6] for belief functions in the Dempster-Shafer (D-S) belief function theory. In this paper, we use the valuation-based systems (VBS) framework [10] to extend compositional models to all uncertainty calculi captured by the VBS framework, which includes calculi such as probability theory, a version of possibility theory with the product t-norm, Spohn's epistemic belief theory, and D-S belief function theory.

Radim Jiroušek
Faculty of Management, University of Economics, Jindřichův Hradec and Prague, Czech Republic
e-mail: radim@utia.cas.cz

Prakash P. Shenoy
School of Business, University of Kansas, Lawrence, KS, USA
e-mail: pshenoy@ku.edu

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 221–228.
springerlink.com

We start by recalling the necessary basic notions of the VBS framework (most of the material is taken from [10]).

2 Valuation-Based Systems

VBS consists of two parts — a static part that is concerned with representation of knowledge, and a dynamic part that is concerned with reasoning.

The static part consists of objects called variables and valuations. Let Φ denote a finite set whose elements are called *variables*. Elements of Φ are denoted by upper-case Roman alphabets such as X, Y, Z, etc. Subsets of Φ are denoted by lower-case Roman alphabets such as r, s, t, etc.

Let Ψ denote a finite set whose elements are called *valuations*. Elements of Ψ are denoted by lower-case Greek alphabets such as ρ, σ, τ, etc. Each valuation is associated with a subset of variables, and represents some knowledge about the variables in the subset. Thus, we say that ρ is a valuation for r, where $r \subseteq \Phi$ is the subset associated with ρ.

We identify a subset of valuations $\Psi_n \subset \Psi$, whose elements are called *normal* valuations. Normal valuations are valuations that are coherent in some sense. In D-S belief function theory, normal valuations are basic probability assignment potentials whose values for all non-empty subsets add to one.

The dynamic part of VBS consists of three operators — combination, marginalization, and removal — that are used to make inferences from the knowledge encoded in a VBS. We define these operators using axioms.

Combination. The first operator is the *combination* operator $\oplus\colon \Psi \times \Psi \to \Psi_n$, which represents aggregation of knowledge. It must satisfy the following three axioms:

1. (*Domain*) If ρ is a valuation for r, and σ is a valuation for s, then $\rho \oplus \sigma$ is a normal valuation for $r \cup s$.
2. (*Commutativity*) $\rho \oplus \sigma = \sigma \oplus \rho$.
3. (*Associativity*) $\rho \oplus (\sigma \oplus \tau) = (\rho \oplus \sigma) \oplus \tau$.

The domain axiom expresses the fact that if ρ represents some knowledge about variables in r, and σ represents some knowledge about variables in s, then $\rho \oplus \sigma$ represents the aggregated knowledge about variables in $r \cup s$. The commutativity and associativity axioms reflect the fact that the sequence in which knowledge is aggregated makes no difference in the aggregated result.

The set of all normal valuations with the combination operator $\oplus$ forms a commutative semigroup. We let $\iota_\emptyset$ denote the (unique) identity valuation of this semigroup. Thus, for any normal valuation ρ, $\rho \oplus \iota_\emptyset = \rho$.

The set of all normal valuations for $s \subseteq \Phi$ with the combination operator $\oplus$ also forms a commutative semigroup (which is different from the semigroup discussed in the previous paragraph). Let ι_s denote the (unique) identity for this semigroup. Thus, for any normal valuation σ for s, $\sigma \oplus \iota_s = \sigma$.

Notice that in general $\rho \oplus \rho \neq \rho$. Thus, it is important to ensure that we do not double count knowledge when double counting matters, i.e., it is okay to double

count knowledge ρ that is idempotent, i.e., $\rho \oplus \rho = \rho$. In representing our knowledge as valuations in Ψ, we have to ensure that there is no double counting of non-idempotent knowledge.

Marginalization. Another operator is marginalization $-X\colon \Psi \to \Psi$, which allows us to coarsen knowledge by marginalizing X out of the domain of a valuation. It must satisfy the following four axioms:

1. (*Domain*) If ρ is a valuation for r, and $X \in r$, then ρ^{-X} is a valuation for $r \setminus \{X\}$.
2. (*Normal*) ρ^{-X} is normal if and only if ρ is normal.
3. (*Order does not matter*) If ρ is a valuation for r, $X \in r$, and $Y \in r$, then $(\rho^{-X})^{-Y} = (\rho^{-Y})^{-X}$, which is denoted by $\rho^{-\{X,Y\}}$.
4. (*Local computation*) If ρ and σ are valuations for r and s, respectively, $X \in r$, and $X \notin s$, then $(\rho \oplus \sigma)^{-X} = (\rho^{-X}) \oplus \sigma$.

The domain axiom is self-explanatory. Marginalization preserves normal (and non-normal) property of valuations. The order does not matter axiom dictates that when we coarsen knowledge by marginalizing out several variables, the order in which the variables are marginalized does not matter in the final result. Occasionally, we let $\rho^{\downarrow r\setminus\{X,Y\}}$ denote $\rho^{-\{X,Y\}}$.

Removal. The removal operator $\ominus\colon \Psi \times \Psi_n \to \Psi_n$ represents removing knowledge in the second valuation from the knowledge in the first valuation. It must satisfy the following three axioms:

1. (*Domain*): Suppose σ is a valuation for s and ρ is a normal valuation for r. Then $\sigma \ominus \rho$ is a normal valuation for $r \cup s$.
2. (*Identity*): For each normal valuation ρ for r, $\rho \oplus \rho \ominus \rho = \rho$. Thus, $\rho \ominus \rho$ acts as an identity for ρ, and we denote $\rho \ominus \rho$ by ι_ρ. Thus, $\rho \oplus \iota_\rho = \rho$.
3. (*Combination and Removal*): Suppose π and θ are valuations, and suppose ρ is a normal valuation. Then, $(\pi \oplus \theta) \ominus \rho = \pi \oplus (\theta \ominus \rho)$.

We call $\sigma \ominus \rho$ the valuation resulting after removing ρ from σ. The identity axiom defines the removal operator as an inverse of the combination operator.

In [10], a number of properties of combination, marginalization, and removal operators are proved. For example, suppose π, σ, θ are valuations for p, s, and t, respectively, ρ is a normal valuation for r, $X \in s$, and $X \notin r$. Then, $(\pi \oplus \theta) \ominus \rho = (\pi \ominus \rho) \oplus \theta$, and $(\sigma \ominus \rho)^{-X} = \sigma^{-X} \ominus \rho$.

3 VBS for D-S Belief Function Theory

In D-S belief function theory, we can use either basic probability assignments, or belief functions, or plausibility functions, or commonality functions, to represent knowledge. Here, we use only basic probability assignments.

Basic Probability Assignment. A *basic probability assignment* (bpa) μ for s is a function $\mu : 2^{\Omega_s} \to \mathbb{R}$ such that $\mu(\mathbf{a}) \geq 0$ for all $\mathbf{a} \in 2^{\Omega_s}$, and $\sum\{\mu(\mathbf{a}) \mid \mathbf{a} \in 2^{\Omega_s}\} = 1$.

B-Valuations. A b-valuation σ for s is a function $\sigma : 2^{\Omega_s} \to \mathbb{R}$. We say σ is *normal* if $\sum\{\sigma(\mathbf{a}) \mid \mathbf{a} \in 2^{\Omega_s}\} = 1$, and we say σ is *proper* if $\sigma(\mathbf{a}) \geq 0$ for all $\mathbf{a} \in 2^{\Omega_s}$.

Proper normal b-valuations represent bpa functions. Normal b-valuations that are not proper are called *pseudo-bpa* .

Set Operations. Suppose r, s, and t are sets of variables, $r \subseteq s$. For $x \in \Omega_s$, $x^{\downarrow r}$ denotes the projection of x into Ω_r. Similarly, for $\mathbf{a} \in 2^{\Omega_s}$, the projection of $\mathbf{a}$ to r, denoted by $\mathbf{a}^{\downarrow r}$, is given by $\mathbf{a}^{\downarrow r} = \{x^{\downarrow r} \mid x \in \mathbf{a}\}$. Also, if $\mathbf{a} \subseteq \Omega_s$, and $\mathbf{b} \subseteq \Omega_t$, then the join of $\mathbf{a}$ and $\mathbf{b}$, denoted by $\mathbf{a} \bowtie \mathbf{b}$ is given by:

$$\mathbf{a} \bowtie \mathbf{b} = \{x \in \Omega_{s \cup t} \mid x^{\downarrow s} \in \mathbf{a}, x^{\downarrow t} \in \mathbf{b}\}. \tag{1}$$

Combination. Suppose ρ and σ are b-valuations for r and s, respectively. Let K denote $\sum\{\rho(\mathbf{b}) \cdot \sigma(\mathbf{c}) \mid \mathbf{b} \subseteq \Omega_r, \mathbf{c} \subseteq \Omega_s \text{ s.t. } \mathbf{b} \bowtie \mathbf{c} = \emptyset\}$. The combination $\rho \oplus \sigma$ is a normal b-valuation for $r \cup s$ given for all $\mathbf{a} \subseteq \Omega_{r \cup s}$ by

$$(\rho \oplus \sigma)(\mathbf{a}) = \begin{cases} K^{-1} \sum\{\rho(\mathbf{b}) \cdot \sigma(\mathbf{c}) \mid \mathbf{b} \subseteq \Omega_r, \mathbf{c} \subseteq \Omega_s \text{ s.t. } \mathbf{b} \bowtie \mathbf{c} = \mathbf{a}\} & \text{if } K \neq 0 \\ 0 & \text{if } K = 0. \end{cases} \tag{2}$$

If $K \neq 0$, then K is a the normalization constant that ensures that $\rho \oplus \sigma$ is a normal b-valuation. It is evident that if ρ and σ are bpa's (proper normal b-valuations), and $K \neq 0$, then $\rho \oplus \sigma$ is a bpa. It can be shown that the definition of combination in Equation (2) satisfies the three axioms of combination.

Marginalization. Suppose σ is a b-valuation for s, and suppose $X \in s$. The marginal σ^{-X} is a b-valuation for $s \setminus \{X\}$ given by

$$\sigma^{-X}(\mathbf{a}) = \sum\{\sigma(\mathbf{b}) \mid \mathbf{b} \in 2^{\Omega_s} \text{ s.t. } \mathbf{b}^{\downarrow s \setminus \{X\}} = \mathbf{a}\} \quad \text{for all } \mathbf{a} \in 2^{\Omega_{s \setminus \{X\}}}. \tag{3}$$

It can be shown that the definition of marginalization in Equation (3) satisfies the four axioms of marginalization.

Removal. Removal is inverse of combination. It is not easy to define removal in terms of b-valuations. For readers familiar with commonality functions, $\oplus$ reduces to pointwise multiplication of commonality functions followed by normalization. Thus, $\sigma \ominus \rho$ is pointwise division of commonality functions corresponding to σ and ρ, followed by normalization. It can be shown that this definition satisfies the three axioms of removal.

Notice that if σ and ρ are proper b-valuations, it is possible that $\sigma \ominus \rho$ is a pseudo-bpa. This may be true even if $r \subseteq s$ and ρ is a marginal of σ.

Convention. For the sake of simplicity, in the rest of this paper we assume that whenever the operator $\oplus$ or $\ominus$ is applied, then the result does not result in the zero valuation, a valuation whose values are identically 0.

4 Compositional Models in VBS

Suppose we have marginals for two overlapping subsets of variables, say for $\{D, G\}$ and $\{D, B\}$. How do we construct a joint distribution for $\{D, G, B\}$ that is consistent with the two marginals (assuming that it exists)? In [4], the operation of "composing" the two marginals to obtain a joint distribution is introduced. One way to view

the composition operator is in terms of no double counting. Notice that the two marginals are not distinct since the knowledge of $\{D\}$ is included in both marginals. So, the composition operator should aggregate the knowledge in the two marginals while adjusting for the double counting of knowledge of $\{D\}$.

In practice, it is extremely unlikely we would find marginals on non-disjoint subsets of variables with common marginals. In this case, there does not exist a joint that agrees with both marginals. So we relax the requirements so that the joint distribution that is constructed is required to agree only with the first marginal.

Composition. A general definition of composition is as follows. Suppose ρ and σ are normal valuations for r and s, respectively. The composition of ρ and σ, written as $\rho \rhd \sigma$, is defined as follows:

$$\rho \rhd \sigma = \rho \oplus \sigma \ominus \sigma^{\downarrow r \cap s} \tag{4}$$

It can be seen directly from the definition in Equation (4) that the composition operator is, in general, neither commutative nor associative. Its most important properties are summarized in the following lemma.

Lemma. *Suppose ρ and σ are normal valuations for r and s, respectively. Then the following statements hold.*

1. Domain: *$\rho \rhd \sigma$ is a normal valuation for $r \cup s$.*
2. Composition preserves first marginal: $(\rho \rhd \sigma)^{\downarrow r} = \rho$.
3. Commutativity under consistency: *If ρ and σ have a common marginal for $r \cap s$, i.e., $\rho^{\downarrow r \cap s} = \sigma^{\downarrow r \cap s}$, then $\rho \rhd \sigma = \sigma \rhd \rho$.*
4. Associativity under a special condition: *Suppose τ is a normal valuation for t, and suppose $s \supset (r \cap t)$. Then, $(\rho \rhd \sigma) \rhd \tau = \rho \rhd (\sigma \rhd \tau)$.*
5. Composition of marginals: *Suppose t is such that $(r \cap s) \subseteq t \subseteq s$. Then*

$$(\rho \rhd \sigma^{\downarrow t}) \rhd \sigma = \rho \rhd \sigma.$$

5 Comparison with an Alternative Compositional Model

For belief functions in the D-S theory, the operator of composition was originally introduced in [6]. Since, as it will be shown in a simple example, it differs from the operator introduced here in Equation (4), we will use for the original operator a slightly different symbol.

Definition. Suppose ρ and σ are normal b-valuations for r and s, respectively. The old-composition of ρ and σ, written here as $\rho \trianglerighteq \sigma$, is defined for each $\mathbf{a} \subseteq \Omega_{r \cup s}$ by one of the following expressions:

[1] if $\sigma^{\downarrow r \cap s}(\mathbf{a}^{\downarrow r \cap s}) > 0$ and $\mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}$ then $(\rho \trianglerighteq \sigma)(\mathbf{a}) = \dfrac{\rho(\mathbf{a}^{\downarrow r}) \cdot \sigma(\mathbf{a}^{\downarrow s})}{\sigma^{\downarrow r \cap s}(\mathbf{a}^{\downarrow r \cap s})}$;

[2] if $\sigma^{\downarrow r \cap s}(\mathbf{a}^{\downarrow r \cap s}) = 0$ and $\mathbf{a} = \mathbf{a}^{\downarrow r} \times \Omega_{s \setminus r}$ then $(\rho \trianglerighteq \sigma)(\mathbf{a}) = \rho(\mathbf{a}^{\downarrow r})$;

[3] in all other cases $(\rho \trianglerighteq \sigma)(\mathbf{a}) = 0$.

Example. Consider the Studený's example [1]. Suppose X, Y and Z are variables with state spaces $\Omega_X = \{x, \bar{x}\}$, $\Omega_Y = \{y, \bar{y}\}$, and $\Omega_Z = \{z, \bar{z}\}$. Consider two b-valuations ρ and σ for $\{X, Z\}$ and $\{Y, Z\}$, respectively, each having only two non-zero values: $\rho(\{x\bar{z}, \bar{x}z\}) = \rho(\{x\bar{z}, \bar{x}\bar{z}\}) = 0.5$ and $\sigma(\{y\bar{z}, \bar{y}z\}) = \sigma(\{y\bar{z}, \bar{y}\bar{z}\}) = 0.5$.

In [7], it is shown that $\rho \trianglerighteq \sigma$ has also only two non-zero values: $(\rho \trianglerighteq \sigma)(\{xy\bar{z}, \bar{x}\bar{y}z\}) = (\rho \trianglerighteq \sigma)(\{xy\bar{z}, x\bar{y}\bar{z}, \bar{x}y\bar{z}, \bar{x}\bar{y}\bar{z}\}) = 0.5$. Thus, we see that $\rho \trianglerighteq \sigma$ is a proper normal b-valuation.

Also, $\rho \oplus \sigma$ is a normal b-valuation with value 0.25 for the following four sets: $\{xy\bar{z}, x\bar{y}\bar{z}\}, \{xy\bar{z}, \bar{x}y\bar{z}\}, \{xy\bar{z}, \bar{x}\bar{y}z\}, \{xy\bar{z}, x\bar{y}\bar{z}, \bar{x}y\bar{z}, \bar{x}\bar{y}\bar{z}\}$. In contrast, $\rho \triangleright \sigma = \rho \oplus \sigma \ominus \sigma^{-Y}$ is a pseudo-bpa since $(\rho \triangleright \sigma)(\{\bar{x}\bar{y}z\}) = -0.25$ (the following are the remaining non-zero values of $\rho \triangleright \sigma$: $(\rho \triangleright \sigma)(\{xy\bar{z}, x\bar{y}\bar{z}\}) = 0.25$, $(\rho \triangleright \sigma)(\{xy\bar{z}, \bar{x}y\bar{z}\}) = 0.25$, $(\rho \triangleright \sigma)(\{xy\bar{z}, \bar{x}\bar{y}z\}) = 0.5$, $(\rho \triangleright \sigma)(\{xy\bar{z}, x\bar{y}\bar{z}, \bar{x}y\bar{z}, \bar{x}\bar{y}\bar{z}\}) = 0.25$).

It is worth mentioning that the same result as $\rho \trianglerighteq \sigma$ is obtained also by the Srivastava-Cogger algorithm [13], but it need not be the case for different values of the ρ and σ b-valuations in this example.

To understand the differences between the two operators of composition, recall that a close connection exists between the combination operator $\oplus$ and a notion of independence. Namely, after combining ρ for X and σ for Y, we get the valuation $\rho \oplus \sigma$ for $\{X, Y\}$, with respect to which variables X and Y are independent. Similarly, if ρ is a valuation for $\{X, Z\}$, and σ is a valuation for $\{Y, Z\}$, with respect to the valuation $\rho \oplus \sigma$ for $\{X, Y, Z\}$, variables X and Y are conditionally independent given Z. However, several other concepts of independence and conditional independence for belief functions exists in the literature. For a non-exhaustive survey, see [1, 2].

In their seminal papers, Dempster [3] and Walley and Fine [15] considered a type of independence that hold for variables X and Y with respect to bpa μ for $\{X, Y\}$ if

$$\mu(\mathbf{a}) = \begin{cases} \mu^{\downarrow X}(\mathbf{a}^{\downarrow X}) \cdot \mu^{\downarrow Y}(\mathbf{a}^{\downarrow Y}) & \text{if } \mathbf{a} = \mathbf{a}^{\downarrow X} \times \mathbf{a}^{\downarrow Y} \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } \mathbf{a} \in \Omega_{\{X,Y\}}. \tag{5}$$

Generalizing this idea, we define an alternative operation of combination, denoted by $\underline{\oplus}$, for b-valuations ρ and σ (for r and s, respectively) as follows. Suppose K denotes $\sum\{\rho(\mathbf{a}^{\downarrow r}) \cdot \sigma(\mathbf{a}^{\downarrow s}) \mid \mathbf{a} \in \Omega_{r \cup s} \text{ s.t. } \mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}\}$. The combination $\rho \,\underline{\oplus}\, \sigma$ is the b-valuation for $r \cup s$ given for all $\mathbf{a} \in \Omega_{r \cup s}$ by

$$(\rho \,\underline{\oplus}\, \sigma)(\mathbf{a}) = \begin{cases} K^{-1} \rho(\mathbf{a}^{\downarrow r})\, \sigma(\mathbf{a}^{\downarrow s}) & \text{if } K > 0, \mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

It is obvious that $\rho \,\underline{\oplus}\, \sigma$ defined in Equation (6) is a proper normal b-valuation for $r \cup s$, and that $\underline{\oplus}$ satisfies all the three axioms of combination.

In a similar way, we define an alternative removal operator $\underline{\ominus}$. Suppose ρ and σ are b-valuations for r and s, respectively, and suppose that ρ is normal. Let K denote $\sum\{\frac{\sigma(\mathbf{a}^{\downarrow s})}{\rho(\mathbf{a}^{\downarrow r})} \mid \mathbf{a} \in \Omega_{r \cup s} \text{ s.t. } \mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}, \rho(\mathbf{a}^{\downarrow r}) > 0\}$. $\sigma \,\underline{\ominus}\, \rho$ is the b-valuation for $s \cup r$ given for all $\mathbf{a} \in \Omega_{s \cup r}$ by

$$(\sigma \underline{\oplus} \rho)(\mathbf{a}) = \begin{cases} K^{-1}\left(\frac{\sigma(\mathbf{a}^{\downarrow s})}{\rho(\mathbf{a}^{\downarrow r})}\right) & \text{if } K > 0, \mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}, \rho(\mathbf{a}^{\downarrow r}) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Thus, together with marginalization defined as in Section 3, we get an alternative VBS for belief functions in the D-S theory. Let two normal b-valuations ρ and σ for r and s, respectively, be such that

$$\sigma^{\downarrow r\cap s}(\mathbf{x}) = 0 \Longrightarrow \rho^{\downarrow r\cap s}(\mathbf{x}) = 0.$$

Consider $\mathbf{a} \subseteq \Omega_{r\cup s}$ for which $\mathbf{a} = \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}$. Then,

$$(\rho \underline{\oplus} \sigma \underline{\ominus} \sigma^{\downarrow r\cap s})(\mathbf{a}) = \begin{cases} k\left(\frac{\rho(\mathbf{a}^{\downarrow r})\sigma(\mathbf{a}^{\downarrow s})}{\sigma^{\downarrow r\cap s}(\mathbf{a}^{\downarrow r\cap s})}\right) & \text{if } \sigma^{\downarrow r\cap s}(\mathbf{a}^{\downarrow r\cap s}) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

which, due to the definition of old-composition, can be rewritten as

$$(\rho \underline{\oplus} \sigma \underline{\ominus} \sigma^{\downarrow r\cap s})(\mathbf{a}) = k\,(\rho \trianglerighteq \sigma)(\mathbf{a}).$$

Notice that because of the above assumption, when computing $\rho \trianglerighteq \sigma$, whenever case [2] of the definition of old composition applies, the value $\rho(\mathbf{a}^{\downarrow r}) = 0$.

Since for all $\mathbf{a} \neq \mathbf{a}^{\downarrow r} \bowtie \mathbf{a}^{\downarrow s}$, $(\rho \underline{\oplus} \sigma \underline{\ominus} \sigma^{\downarrow r\cap s})(\mathbf{a}) = (\rho \trianglerighteq \sigma)(\mathbf{a}) = 0$, we get

$$(\rho \underline{\oplus} \sigma \underline{\ominus} \sigma^{\downarrow r\cap s})(\mathbf{a}) = k\,(\rho \trianglerighteq \sigma)(\mathbf{a}), \quad \text{for all } \mathbf{a} \subseteq \Omega_{r\cup s}.$$

Since we know that both $\rho \underline{\oplus} \sigma \underline{\ominus} \sigma^{\downarrow r\cap s}$ and $\rho \trianglerighteq \sigma$ are normal b-valuations (for the former, it follows from the lemma presented in Section 4; for the latter, it is proved in [6]), it follows that $k = 1$.

Thus, we have shown that the operator of composition defined in [6] can be considered as a special case of composition in a VBS where combination is $\underline{\oplus}$, removal is $\underline{\ominus}$, and marginalization is the same as in the D-S theory.

6 Summary and Conclusions

We have described the VBS framework in general, and described the composition model in the VBS framework using the semantics of no double counting of knowledge. We have compared the compositional model defined in this paper for D-S belief function *theory* with the one described in [6] for belief functions. Our conclusion is that although both of these compositional models are defined for belief functions and its alternative representations (bpa, commonality, etc.), the former is defined for *the* D-S belief function theory (that necessarily entails Dempster's rule of combination), and the latter for a belief function theory that has $\underline{\oplus}$ as the rule of combination. Both of these theories fit in the VBS framework, but they have different semantics, different notions of conditional independence, etc.

Acknowledgment. This work has been supported in part by funds from grant GAČR 403/12/2175 to the first author, and from the Ronald G. Harper Distinguished Professorship at the University of Kansas to the second author. We are grateful to M. Studený for valuable discussions and comments. This paper is derived from a longer version in [7].

References

1. Ben-Yaghlane, B., Smets, P., Mellouli, K.: Belief function independence: II. The conditional case. International Journal of Approximate Reasoning 31(1-2), 31–75 (2002)
2. Couso, I., Moral, S., Walley, P.: Examples of independence for imprecise probabilities. In: Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications, ISIPTA 1999 (1999)
3. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
4. Jiroušek, R.: Composition of probability measures on finite spaces. In: Geiger, D., Shenoy, P.P. (eds.) Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI 1997), pp. 274–281. Morgan Kaufmann, San Francisco (1997)
5. Jiroušek, R.: Foundations of compositional model theory. International Journal of General Systems 40(6), 623–678 (2011)
6. Jiroušek, R., Vejnarová, J., Daniel, M.: Compositional models of belief functions. In: de Cooman, G., Vejnarová, J., Zaffalon, M. (eds.) Proceedings of the 5th Symposium on Imprecise Probabilities and Their Applications (ISIPTA 2007), Prague, Czech Republic, pp. 243–252. Charles University Press (2007)
7. Jiroušek, R., Shenoy, P.P.: Compositional models in valuation-based systems. Working Paper No. 325, School of Business, University of Kansas, Lawrence, KS (2011)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
9. Shenoy, P.P.: On Spohn's rule for revision of beliefs. International Journal of Approximate Reasoning 5(2), 149–181 (1991)
10. Shenoy, P.P.: Conditional independence in valuation-based systems. International Journal of Approximate Reasoning 10(3), 203–234 (1994)
11. Shenoy, P.P.: Binary join trees for computing marginals in the Shenoy-Shafer architecture. International Journal of Approximate Reasoning 17(2-3), 239–263 (1997)
12. Shenoy, P.P.: No double counting semantics for conditional independence. In: Cozman, F.G., Nau, R., Seidenfeld, T. (eds.) Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications (ISIPTA 2005), pp. 306–314. Society for Imprecise Probabilities and Their Applications (2005)
13. Srivastava, R.P., Cogger, K.: Beliefs on Individual Variables from a Single Source to Beliefs on the Joint Space under Dempster-Shafer Theory: An Algorithm. In: Filipe, J., Fred, A., Sharp, B. (eds.) Proceedings of the First International Conference on Agents and Artificial Intelligence, ICAART 2009, pp. 191–197. INSTICC Press, Porto (2009)
14. Vejnarová, J.: Composition of possibility measures on finite spaces: Preliminary results. In: Bouchon-Meunier, B., Yager, R.R. (eds.) Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 1998), pp. 25–30 (1998)
15. Walley, P., Fine, T.L.: Towards a frequentist theory of upper and lower probability. Annals of Statistics 10, 741–761 (1982)

Ascribing Causality from Interventional Belief Function Knowledge

Imen Boukhris, Salem Benferhat, and Zied Elouedi

Abstract. In many Artificial Intelligence applications, causality is an important issue. Interventions are external manipulations that alter the natural behavior of the system. They have been used as tools to distinguish causal relations from spurious correlations. This paper proposes a model allowing the detection of causal relationships under the belief function framework resulting from acting on some events. Facilitation and justification in the presence of interventions, concepts complementary to the concept of causality, are also discussed in this paper.

1 Introduction

Causal knowledge simplifies decision-making. In fact, it enables to choose the right actions to achieve the goals. Accordingly, discovering causal relations is a task of crucial importance in many applications. Three kinds of causal reasoning may exist: by abduction for diagnosis problems, by deduction to deal with simulation problems or by induction for ascribing causal links [2].

Usually, an agent identifies causal links from its background knowledge about the normal course of things and a set of *observed* events. Some of these reported events are considered as exceptional ones. Therefore, the concept of abnormality plays an important role for ascribing causality.

Observational data provide some information about the statistical relations among events. This means that they might be correlations that do not necessarily follow a causal process. To tackle this problem, interventions are used [11, 13]. They consist in external actions that perturb the spontaneous behavior of the system by forcing some variables to take specific values. Through these experimentations, the effects of all direct (and undirected) causes related to the variable of interest will be

Imen Boukhris · Salem Benferhat
CRIL-Université d'Artois-Faculté Jean Perrin-Lens-France
e-mail: imen.boukhris@hotmail.com, benferhat@cril.univ-artois.fr

Imen Boukhris · Zied Elouedi
LARODEC-Université de Tunis-ISG-Tunis-Tunisie
e-mail: zied.elouedi@gmx.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 229–237.
springerlink.com

ignored. Therefore, given two dependent variables A and B, if an action on an event A has no impact on an event B, then A cannot be the cause of event B, but if a manipulation of event A leads to a change in B, then we can conclude that A is a cause of B.

While in the context of observations, any representation of the background knowledge is suitable, in the context of interventions, the "graphical structure" is needed. Interventions will be represented on this causal structure by the mean of the "do" operator. This tool was originally introduced by [10] for the ordinal conditional functions of Spohn [19] and proposed after that in [13] for Bayesian causal networks. To fit several kinds of imperfect knowledge, counterparts of the do-operator were proposed in possibilistic causal networks [1, 3, 5] to deal with pure qualitative knowledge when only the ordinal handling is important, and in belief function causal networks [9].

Since information is is almost always tainted with various kinds of imperfection, in this paper we are interested in ascribing causality when the agent's background knowledge is formalized with belief functions [14, 18]. It is an appropriate framework to handle imperfect causal data [15]. In fact, the belief function theory has an expressive power to model different forms of uncertainty including full knowledge, partial ignorance, total ignorance and even probabilistic knowledge. It also better manages ignorance situations [20].

A very preliminary work has been addressed in [8] in the context of observations only. In this paper, we propose to ascribe causality when observing abnormal events and also in presence of *interventions*. Our introduced model is based on belief function causal networks. The advantage of these networks comparing to the Bayesian ones, is that they allow the description of uncertain effects including situations of total ignorance after making an intervention.

The rest of the paper is organized as follows: In Section 2, we recall the basics of the belief function theory. Section 3 presents a new definition of the concept of acceptance. The latest is used to ascribe causality. Our model for causality ascription under the belief function framework in presence of observations or interventions is introduced in Section 4. We also define its related notions namely, facilitation, justification introduced in [6] and also the concepts of confirmation and attenuation proper to the belief function framework. Section 5 concludes the paper.

2 Belief Function Theory: Basic Concepts

In the following, we recall some of the basics of belief function theory. More details can be found in [18]. In the belief function theory [14], beliefs are expressed on propositions belonging to the powerset of Θ. The basic belief assignment (bba), denoted by m, is a mapping from 2^{Θ} to $[0,1]$ such that:

$$\sum_{A \subseteq \Theta} m(A) = 1. \tag{1}$$

where $m(A)$ is a basic belief mass (*bbm*) assigned to $A \subseteq \Theta$, and represents the part of belief *exactly* committed to the event A. The subsets of Θ such that $m(A) > 0$ are called focal elements.

A bba m can be equivalently represented by a plausibility function pl: $2^{\Theta} \rightarrow [0,1]$, defined as:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \text{ and } pl(\emptyset) = 0. \tag{2}$$

The value $pl(A)$ quantifies the maximum amount of belief that could be given to a subset A.

The combined effect of two distinct sources, providing two bba's m_1 and m_2, is computed by Dempster's rule of combination [16], defined as:

$$m_1 \oplus m_2(A) = K^{-1} \sum_{B \cap C = A} m_1(B) \cdot m_2(C), \forall B, C \subseteq \Theta. \tag{3}$$

where the normalization factor: $K = 1 - \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C)$.

Conditioning consists in revising the agent belief originally defined on A, following the arrival of a new information $B \subseteq A$. Indeed, the mass that was specifically allocated to A is transferred to $A \cap B$ using Dempster's rule of conditioning. $pl(.|B)$ denotes the conditional plausibility function obtained after revising the corresponding pl using a new piece of evidence B (where $pl(B) > 0$) and is defined as [17]:

$$pl(A|B) = \frac{pl(A \cap B)}{pl(B)}. \tag{4}$$

3 Acceptance and Ascribing Causality in Presence of Observations

In this section, we introduce our causal model which is a counterpart of the qualitative model proposed in [6, 7]. It also overcomes the limitation of the original model in which the representation of events is restrained to binary variables. Therefore, an agent will identify an unknown causal relation from three components:

- his non-causal uncertain background knowledge about the natural course of the world formalized within the belief function theory;
  For instance, in an intrusion detection system [12], one may express the effects of legitimates actions on the system.
- a sequence of observations occurring in his environment, $O = \{f_1, \ldots, f_n\}$. We define an observation f_i as a subset of the frame of discernment Θ which is the cartesian product of all variable domains;
  In intrusion detection systems, this may refer to log files (for legitimate actions) and alerts that report a set of actions that are considered as malicious by some security policies.

- an event from the set of observed events that contradicts his judgment about the normal course of things: an abnormal event e_i. In fact, the agent will ascribe the causes of this abnormal event.
  For instance, there exists an illegal access to some data.

The agent will discriminate from the set of observed events between potential causes. For that we will use the concepts of acceptance/rejection instead of changes in uncertainty to ascribe causality. We define potential causes, which are rejected events, as *a partition* representing exhaustive and mutually exclusive events of the cartesian product of the domain of some n-ary variables.

The set of possible events is denoted by $\Theta_E = \{e_1, e_2, \ldots, e_n\}$ satisfying these properties:

1- Exhaustibility: $e_1 \cup e_2 \cup \ldots \cup e_n = \Theta$,
2- Exclusivity: $\forall i, j \;\; e_i \cap e_j = \emptyset$.

The complement of e_i w.r.t. Θ, denoted by $\overline{e_i}$ is defined as: $\overline{e_i} = \bigcup_{e_j, e_j \neq e_i} e_j$.

An abnormal event can be an atomic event representing any instance of some n-ary variable $A_i = a_{ij}$. In this case $\Theta_E = \{[a_{i1}], \ldots, [a_{in}]\}$ where $[a_{ij}]$ is a set of all elements $\theta \in \Theta$ such that A_i in θ has the value a_{ij}.
Note that when there is no ambiguity, we will use e_i to denote normal and abnormal events.

In this paper, we consider that an event is accepted if it is likely enough to be considered as it holds. An event e_i has different possible status, in particular it said to be:

- accepted: if the confidence in this event is strictly greater than the confidence in its complement according to Θ_E: $pl(e_i) > pl(\overline{e_i})$;
- rejected: if the confidence in this event is strictly less than the confidence in its complement according to Θ_E: $pl(e_i) < pl(\overline{e_i})$;
- ignored: if the confidence in this event is the same than the confidence in its complement according to Θ_E: $pl(e_i) = pl(\overline{e_i})$.

Of course, one may consider different levels of acceptance (rejection) in order to associate a strength to causality ascriptions. This topic is not addressed here due to space limitations.

Based on these definitions of acceptance/rejection, observed events are seen to be linked by :

- causation: If an event e_i is rejected and after observing an event e_j it becomes accepted, e_j is said to be a cause of e_i. Namely, $pl(\overline{e_i}, e_j) < pl(e_j, e_i) \leq pl(e_i) < 1$;
- facilitation: If an event e_j is rejected and after observing an event e_i it becomes ignored then e_i is said to facilitate the occurrence of e_j. Namely, $0 < pl(e_i, e_j) = pl(\overline{e_i}, e_j) \leq pl(e_j) < pl(\overline{e_j})$;
- justification: Given a sequence of events, e_j is said to justify e_i, if e_i is ignored and becomes accepted after the observation of e_j. Namely, $pl(\overline{e_i}, e_j) < pl(e_j) \leq pl(e_i) = pl(\overline{e_i})$.

These three concepts are the counterpart of the ones proposed in [4, 6] in possibility theory framework.

4 Ascribing Causality Based on Belief Function Causal Networks

4.1 Needs of Interventions

Causality plays an important role in many applications such as policy analysis or decision making. Thus, a spurious correlation should be well distinguished from a causal connection. In fact, two events may be wrongly inferred as causally related, due to either the coincidence of their occurrence or the presence of a common cause which is a hidden event. Finding the cause of an event will be much better and easier and if it is based on data collected via active interventions rather than passive observations.

Interventions can be seen as experimentations that force some variables to have some specific values. They are represented with the "do" operator. A manipulation on a variable A_i is an external action that forces it to take the specific value a_{ij} without modifying our beliefs over its direct causes. It is denoted $do(A_i = a_{ij})$ or $do(a_{ij})$.

Example 1. An agent learns that someone took up drugs, that he has dilated pupils. He notices that this person's heart rate has increased. The agent believes that generally, it is abnormal to be a drug-consumer, to have dilated pupils, and to have an accelerated heart rate: ($pl(\{\overline{Drugs}\}) > pl(\{Drugs\})$; $pl(\{\overline{Dilated}\}) > pl(\{Dilated\})$; $pl(\{\overline{Accelerated}\}) > pl(\{Accelerated\})$).

From the observation: a person who has dilated pupils, has also an accelerated heart rate, the agent will conclude that when pupils are dilated, it causes an increase in heart rate ($pl(\{(Dilated, Accelerated)\}) > pl(\{(Dilated, \overline{Accelerated})\})$).

Tropicamide shortly acts on the dilation of the pupil. When it is applied as eyes drops, it forces the eyes to be dilated ($do(Dilated)$). The agent notes that his action has no effect on the speed of the heartbeat. Accordingly, he will not be able to infer that there is a causal relation between these two events.

The agent believes that it is normal for a drug-consumer to have dilated pupils ($pl(\{(Drugs, Dilated)\}) > pl(\{(Drugs, \overline{Dilated})\})$) and to have an accelerated heart rate ($pl(\{(Drugs, Accelerated)\}) > pl(\{(Drugs, \overline{Accelerated})\})$).

After forcing someone to take drugs ($do(Drugs)$), he observes that his pupils are dilated and the speed of his heartbeat is altered. Therefore, he concludes that the hidden event, namely taking drugs, is their common cause.

4.2 Extending Causality Ascription to Deal with Interventions

To ascribe causal relations between elements of the system, we propose to use belief function causal networks [9]. The distinction between observations and interventions, to identify causal relationships, is therefore made by the use of the "do" operator.

A belief function causal network is a directed acyclic graph where nodes represent variables and arcs represent not only dependence relations but also cause-effect relationships. The set of parents of A_i is denoted by Pa(A_i). Quantitatively, a set of bba's is associated with each node in the graph. For each root node A_i ($Pa(A_i) = \emptyset$) having a frame of discernment Θ_{A_i}, an a priori mass $m(a)$ is defined on the powerset $2^{\Theta_{A_i}}$. For other nodes, a conditional bba $m(a|Pa_j(A_i))$ is specified for each subset of A_i knowing an instance of $Pa(A_i)$.

An intervention on a variable A_i forces it to take the specific value a_{ij}, $do(a_{ij})$. This action makes the original causes of the manipulated variable no more responsible of its state. However, our beliefs on the parents set $Pa(A_i)$ will not be affected. Graphically, it can be represented by the deletion of arcs relating the variable of interest with its parents. The resulting graph is called a mutilated graph. The effect of an intervention $do(a_{ij})$ corresponds to observing the value a_{ij} on this graph.

While an observation is encoded as a conditional bba computed as m(.|a_{ij}), the effect of an intervention on a variable A_i forcing it to take the value a_{ij}, is given by m(.|do(a_{ij})). If m is compactly represented by a belief network, then $m(.|do(a_{ij}))$ is obtained from the network after removing the links between A_i and its parents (see [9] for more details). The following subsection extends causality ascription for handling such interventions.

4.3 Causality Ascription

An agent will ascribe causality from four components, namely:

- his background knowledge represented with a graphical structure: a belief function causal network;
- a sequence of observations;
- a sequence of interventions;
- an event that contradicts his judgment about the normal course of things: an abnormal event.

Two events are perceived as causally related, if the agent starts believing that one of them is rejected and after an action on the other one, he changes his beliefs and accepts it.

Definition 1. Belief function causality ascription: If an event e_i is rejected, i.e. $pl(\overline{e_i}) > pl(e_i)$, and after acting on an event e_j it becomes accepted, i.e. $pl(e_i|do(e_j)) > pl(\overline{e_i}|do(e_j))$. An intervention do($e_j$) is said to be a cause of e_i, namely

$$pl(\overline{e_i}, do(e_j)) < pl(e_i, do(e_j)) \leq pl(e_i) < 1. \quad (5)$$

Example 2. Having dilated pupils is a rejected event: $pl(\{\overline{Dilated}\}) > pl(\{Dilated\})$; After forcing someone to take drugs ($do(Drugs)$), the agent observes that his pupils are dilated: $pl(\{(do(Drugs), Dilated)\}) > pl(\{(do(Drugs), \overline{Dilated})\})$. He will conclude that forcing the event Drugs to be taken caused the dilation of the pupils.

4.4 Facilitation Ascription

Facilitation is a concept that is very close to causality. It is used when an agent is cautious in his causal interpretation of the sequence of events: he starts not believing in the occurrence of an event under the normal course of things and he changes his beliefs after acting on an another event. However, this change consists to not believe in the event neither in its complement instead of accepting it as it is the case for causality.

Definition 2. Belief function facilitation ascription: If an event e_i is rejected, i.e. $pl(\overline{e_i}) > pl(e_i)$, and after acting an event e_j it becomes ignored, i.e. $pl(e_i|do(e_j)) = pl(\overline{e_i}|do(e_j))$, then an intervention do(e_j) is said facilitate the occurrence of e_i. Namely,

$$0 < pl(do(e_j), e_i) = pl(do(\overline{e_j}), e_i) \leq pl(e_i) < pl(\overline{e_i}). \tag{6}$$

Example 3. The frequency of seizures is represented with a variable epilepsy, $\Theta_E = \{low, medium, high\}$. Beliefs are expressed on subsets of Θ_E. Having a high frequency of epileptic seizure is an event strongly rejected, i.e. $pl(\{\overline{high}\}) > pl(\{high\})$. By administering a drug to someone, namely $do(Drug)$, the risk of having many crises becomes unsurprising, i.e. $pl(\{high\}|do(Drug)) = pl(\{\overline{high}\}|do(Drug))$. The intervention $do(Drug)$ is therefore seen as facilitating having epileptic seizures.

4.5 Justification Ascription

If an agent judges that forcing the occurrence of an event e_j gave reason to expect the occurrence of observing e_i, we deal then with justification. Acting on e_j caused the agent to start believing e_i, and that it should not be surprised of having e_i reported afterwards.

Definition 3. Belief function justification: Given a sequence of events, an intervention do(e_j) is said to justify e_i, if e_i is ignored, i.e. $pl(e_i) = pl(\overline{e_i})$, and becomes accepted after an action on e_j, i.e. $pl(e_i|do(e_j)) > pl(\overline{e_i}|do(e_j))$. Namely,

$$pl(\overline{e_i}, do(e_j)) < pl(do(e_j)) \leq pl(e_i) = pl(\overline{e_i}). \tag{7}$$

Example 4. The risk of heart failure is an ignored event: $pl(\{\overline{Failure}\}) = pl(\{Failure\})$ In context of high level of alcohol in the blood, after forcing someone to take cocaine, the risk that this person has a heart failure afterward is very strongly accepted $pl(\{(High, Failure)\}|do(Drug)) > pl(\{(High, \overline{Failure})\}|do(Drug))$. Accordingly, intervening on cocaine in context of alcohol strongly justifies heart failure risk.

4.6 Confirmation and Attenuation

In the quantitative belief function framework acceptance, rejection can be confirmed or attenuated upon intervening on some variables.

Definition 4. Belief function confirmation: An intervention on e_j, do(e_j), is said to confirm another event e_i if the plausibility of observing e_i after acting on e_j is greater than the plausibility of observing e_i alone, i.e. $pl(e_i|do(e_j)) > pl(e_i)$. Namely,

$$pl(e_i) \cdot pl(do(\overline{e_j})) < pl(e_i, do(e_j)) < pl(e_j) \tag{8}$$

Definition 5. Belief function attenuation: An intervention do(e_j) is said to attenuate e_i if the plausibility of observing e_i after acting on e_j is smaller than the plausibility of observing e_i alone, i.e. $pl(e_i|do(e_j)) < pl(e_i)$. Namely,

$$pl(e_i, do(e_j)) < pl(e_i) \cdot pl(do(e_j)) < pl(do(e_j)). \tag{9}$$

Example 5. Suppose that an agent expresses his beliefs about the state of drunkenness of a person: $pl(\{Drunk\}) > pl(\{\overline{Drunk}\})$. By administering cocaine, $(do(Drug))$, the agent changes his beliefs. Indeed, $pl(\{Drunk\}|do(Drug)) < pl(\{Drunk\})$ and $pl(\{\overline{Drunk}\}|do(Drug)) > pl(\{\overline{Drunk}\})$. Thus, forcing a person to take cocaine is seen as attenuating drunkenness and confirming the acceptance of $\overline{Drunk}$.

5 Conclusion

In this paper, we proposed a model able to identify causal links between events in a sequence when external actions are experienced. This is done using the do-operator. We showed that making interventions allows to better identify causal links by distinguishing between correlation and causation. After forcing some n-ary variables to take a specific value, we showed the impact of this action to differentiate between events related in a causal way and those when facilitation or justification are involved according to the definitions of acceptance and rejection that we have proposed. As future works, we intend to include other definitions of acceptance and rejection to define several strength of causal links.

References

1. Benferhat, S.: Interventions and belief change in possibilistic graphical models. Artif. Intell. 174(2), 177–189 (2010)
2. Benferhat, S., Bonnefon, J.-F., Chassy, P., Da Silva Neves, R., Dubois, D., Dupin de Saint-Cyr, F., Kayser, D., Nouioua, F., Nouioua-Boutouhami, S., Prade, H., Smaoui, S.: A Comparative Study of Six Formal Models of Causal Ascription. In: Greco, S., Lukasiewicz, T. (eds.) SUM 2008. LNCS (LNAI), vol. 5291, pp. 47–62. Springer, Heidelberg (2008)
3. Benferhat, S., Smaoui, S.: Possibilistic causal networks for handling interventions: A new propagation algorithm. In: AAAI, pp. 373–378. AAAI Press (2007)
4. Benferhat, S., Smaoui, S.: Quantitative Possibilistic Networks: Handling Interventions and Ascribing Causality. In: Gelbukh, A., Morales, E.F. (eds.) MICAI 2008. LNCS (LNAI), vol. 5317, pp. 720–731. Springer, Heidelberg (2008)

5. Benferhat, S., Smaoui, S.: Inferring interventions in product-based possibilistic causal networks. Fuzzy Sets and Systems 169(1), 26–50 (2011)
6. Bonnefon, J., Da Silva Neves, R., Dubois, D., Prade, H.: Background default knowledge and causality ascriptions. In: ECAI, pp. 11–15 (2006)
7. Bonnefon, J.F., Da Silva Neves, R., Dubois, D., Prade, H.: Predicting causality ascriptions from background knowledge: model and experimental validation. Int. J. Approx. Reasoning 48(3), 752–765 (2008)
8. Boukhris, I., Benferhat, S., Elouedi, Z.: A belief function model for ascribing causality. In: EPIA, pp. 342–356 (2011)
9. Boukhris, I., Elouedi, Z., Benferhat, S.: Modeling interventions using belief causal networks. In: FLAIRS, pp. 602–607 (2011)
10. Goldszmidt, M., Pearl, J.: Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In: KR, pp. 661–672 (1992)
11. Halpern, J., Pearl, J.: Causes and explanations: A structurel model approach. In: UAI, pp. 194–202 (2001)
12. Morin, B., Mé, L., Debar, H., Ducassé, M.: A logic-based model to support alert correlation in intrusion detection. Information Fusion 10(4), 285–299 (2009)
13. Pearl, J.: Causality: Models, Reasonning and Inference. Cambridge University Press (2000)
14. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
15. Shafer, G.: The Art of Causal Conjecture. The MIT Press (1997)
16. Smets, P.: The combination of evidence in the transferable belief model. IEEE Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)
17. Smets, P.: About updating. In: UAI, pp. 378–385 (1991)
18. Smets, P.: The transferable belief model for quantified belief representation, vol. 1, pp. 267–301. Kluwer Academic Publisher (1998)
19. Spohn, W.: Ordinal conditional functions: a dynamic theory of epistemic states causation in decision. In: Belief Changes and Statistics, pp. 105–134 (1988)
20. Wakker, P.: Dempster belief functions are based on the principle of complete ignorance. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 8(3), 271–284 (2000)

About Sources Dependence in the Theory of Belief Functions

Mouna Chebbah, Arnaud Martin, and Boutheina Ben Yaghlane

Abstract. In the theory of belief functions many combination rules are proposed in the purpose of merging and confronting several sources opinions. Some combination rules are used when sources are cognitively independent whereas others are specific to dependent sources. In this paper, we suggest a method to quantify sources degrees of dependence in order to choose the more appropriate combination rule. We used generated mass functions to test the proposed method.

1 Introduction

Decision making is more and more difficult when using imperfect data, however information can be imprecise, uncertain and even not available. Usually decision is made using precise and certain data but available information are not always so. Many theories manage uncertainty such as the *theory of probabilities*, the *theory of fuzzy sets*, the *theory of possibilities* and the *theory of belief functions*. Within imperfect environment, combining several imperfect information helps users and decision makers to reduce the degree of uncertainty by confronting several opinions. The theory of belief functions presents a strong framework for combination.

Mouna Chebbah
LARODEC Laboratory, ISG Tunis, 41 Rue de la liberté, Cité Bouchoucha 2000 Le Bardo, Tunisia
IRISA, University of Rennes 1, rue E. Branly, 22300 Lannion
e-mail: Mouna.Chebbah@gnet.tn

Arnaud Martin
IRISA, University of Rennes 1, rue E. Branly, Lannion
e-mail: Arnaud.Martin@univ-rennes1.fr

Boutheina Ben Yaghlane
LARODEC Laboratory, IHEC Carthage, Carthage Présidence 2016, Tunisia
e-mail: boutheina.yaghlane@ihec.rnu.tn

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 239–246.
springerlink.com

To combine uncertain information many combination rules can be used. Some of these combination rules are used when sources are cognitively independent like [6, 7, 9, 10, 13] but the cautious rule [5] is applied when sources are dependent. A source is assumed to be cognitively independent towards another one when the knowledge of the belief of that source does not affect the belief of the first one. In some cases, like when a source is completely dependent on another source, the user can decide to discard the dependent source and its mass functions from the combination.

Some researches are focused on the sources statistical dependence such as [1, 2] and others [12, 11] tackled the cognitive dependence between variables. This paper is focused on sources dependence measuring. Thus, we suggest a method to estimate the dependence between sources.

In the following, we introduce preliminaries of the theory of belief functions in the second section. In the third section, the independence measure is presented. This independence is estimated in three steps, in the first step a clustering technique is applied then similar clusters are matched in the second step and finally a weight is affected to matched clusters. This method is tested on random mass functions in the fourth section. Finally, conclusions are drawn.

2 Theory of Belief Functions

The theory of belief functions was introduced by [4] and [12] and so called *Dempster-Shafer theory* to model imperfect information held by a source (an expert, a belief holder, ...). In this section, we will remind some basic notions of this theory as seen in the transferable belief model [10].

The *frame of discernment* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is a set of n elementary and mutually exclusive and exhaustive hypotheses. These hypotheses are all the possible and eventual solutions of the problem under study. The *power set* 2^Ω is the set of all subsets made up of hypotheses and union of hypotheses from Ω. The *basic belief assignment (bba)* also called *mass function* is a function defined on the power set 2^Ω and affects a value from $[0,1]$ such that: $\sum_{A \subseteq \Omega} m(A) = 1$. We can also assume that: $m(\emptyset) = 0$. A subset A having a strictly positive mass is called *focal element*. The mass allocated to this focal element A is the source's degree of belief that the solution of the problem under study is in A. In the theory of belief functions, a great number of combination rules [6, 7, 9, 10, 13] are used to summarize all combined mass functions into only one mass function reflecting all the sources beliefs. The first combination rule was proposed by Dempster in [4] and is defined for two distinct mass functions m_1 and m_2:

$$m_{1\textcircled{\cap}2}(A) = (m_1 \textcircled{\cap} m_2)(A) = \begin{cases} \dfrac{\sum_{B \cap C = A} m_1(B) \times m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) \times m_2(C)} & \forall A \subseteq \Omega,\ A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \quad (1)$$

Dempster's rule of combination together with other rules [6, 7, 9, 10, 13] are used to combine independent mass functions. In the case of dependent sources, the cautious rule [5] can be applied. After the combination, the pignistic probability $\mathrm{BetP}(X)$ is generally used to decide.

3 Independence

Independence concept was first introduced in probability theory in the purpose of studying dependent statistical variables. In the probability theory, two variables A and B are assumed to be independent if one of these equivalent conditions is satisfied: $P(A \cap B) = P(A) * P(B)$ or $P(A|B) = P(A)$. *Statistical independence* is generalized from probability theory to the theory of belief functions [1, 2]. Mass functions can be seen as subjective probabilities held by sources (experts, belief holders, ...) who can communicate, thus *cognitive independence* is specially defined in the theory of belief functions. A definition of cognitive independence was first proposed by Shafer ([12], page 149) as "two frames of discernment may be called cognitively independent with respect to the evidence if new evidence that bears on only one of them will not change the degree of support for propositions discerned by the other". Smets [11] claims that two variables are independent when the knowledge of the value taken by one of them does not affect our belief about the other. This paper is not focused on variables independence but on sources independence.

Definition 1. Two sources are independent when the knowledge of the belief provided by one source does not affect the belief of the other source, otherwise these sources are dependent.

Not only communicating sources are considered to be dependent but also sources having the same background of knowledge since their beliefs are similar. In this paper, mass functions provided by two sources are studied in order to reveal any dependence between them. Therefore, the aim is to find dependence between sources if it exists. In the following, we define an independence measure $\mathrm{I_d}$, ($\mathrm{I_d}(s_1, s_2)$ is the independence of s_1 towards s_2) verifying the following axioms:

1. Non-negativity: The independence of a source s_1 on an another source s_2, $\mathrm{I_d}(s_1, s_2)$ cannot be negative, it is a positive or null degree.
2. Normalization: Source independence $\mathrm{I_d}$ is a degree on $[0, 1]$, it is null when the source is dependent from the other one, equal to 1 when it is completely independent and a degree in $[0, 1]$ otherwise.
3. Non-symmetry: If a source s_1 is dependent on a source s_2, s_2 is not necessarily dependent on s_1. Even if s_1 and s_2 are mutually dependent, degrees of dependence are not the same.
4. Identity: $\mathrm{I_d}(s_1, s_1) = 0$. A source is completely dependent from it self.

If two sources s_1 and s_2 are dependent, there will be a relation between their belief functions. The main idea of this paper is to classify mass functions provided by each source, then a study of the similarities between cluster repartitions can reveal any dependence between sources. Once clustering is performed, the idea is to study the

sources overall behavior. The proposed method is in three steps, in the first step mass functions of each source are classified then in the second step similar clusters are matched and finally the weights of the linked clusters are quantified in the third step.

3.1 Clustering

In this paper, we use a modified C-means algorithm with the distance on belief functions given by [8] such as in [3] to classify mass functions of one source. The number of clusters C has to be also known, a set T contains n objects $o_i : 1 \leq i \leq n$ which values m_i are belief functions defined on a frame of discernment Ω. For example, a doctor is diagnosing the disease of n patients and giving each time a mass function as an uncertain diagnostic. In that case, patients are considered as these objects o_i to be classified, the frame of discernment Ω contains all the possible diseases and m_i is the mass function provided by the doctor when diagnosing each patient o_i. In this section a clustering technique is performed on mass functions m_i provided by the same source in order to study the overall behavior of a source.

This clustering technique is based on a dissimilarity measure which is used to quantify the dissimilarity of an object o_i towards a cluster Cl_k. The dissimilarity D of the object o_i towards the cluster Cl_k is the mean of distances between m_i the mass function value of the object o_i and all the n_k mass functions classified into the cluster Cl_k as follows:

$$D(o_i, Cl_k) = \frac{1}{n_k} \sum_{j=1}^{n_k} d(m_i^{\Omega}, m_j^{\Omega}) \tag{2}$$

$$d(m_1^{\Omega}, m_2^{\Omega}) = \sqrt{\frac{1}{2}(m_1^{\Omega} - m_2^{\Omega})^t \underline{\underline{D}}(m_1^{\Omega} - m_2^{\Omega})}, \underline{\underline{D}}(A,B) = \begin{cases} 1 & \text{if } A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} & \forall A, B \in 2^{\Omega} \end{cases} \tag{3}$$

Each object is affected to the most similar cluster in an iterative way until reaching an unchanged cluster partition. It is obvious that the number of clusters C has to be fixed. In this paper, we suppose that C is the cardinality of the frame of discernment. In a classification problem, the cardinality of the frame of discernment is the number of classes that is why we choose $C = |\Omega|$ in this paper.

3.2 Cluster Matching

Clustering technique, given in section 3.1, is used to classify mass functions provided by both sources s_1 and s_2, the number of clusters is assumed to be the cardinality of the frame of discernment. After the classification, both mass functions provided by s_1 and s_2 are distributed on C clusters. Once clustering performed the most similar clusters have to be linked, a cluster matching is performed for both clusters of s_1 and that of s_2. The dissimilarity between two clusters Cl_{k_1} of s_1 and Cl_{k_2} of s_2 is the mean of distances between objects $o_l \in Cl_{k_1}$ and $o_w \in Cl_{k_2}$:

$$\delta^1(Cl_{k_1}, Cl_{k_2}) = \frac{1}{n_{k_1}} \sum_{l=1}^{n_{k_1}} \frac{1}{n_{k_2}} \sum_{w=1}^{n_{k_2}} d(o_l, o_w) \tag{4}$$

We note that n_{k_1} is the number of objects on the cluster Cl_{k_1}, δ^1 is the dissimilarity towards the source s_1 and d is the distance defined by equation (3). It is obvious that $d(o_l, o_w) \in [0,1]$. $\delta^1(Cl_{k_1}, Cl_{k_2})$ is the mean of pairwise distances between objects of Cl_{k_1} and Cl_{k_2}, thus $\delta^1(Cl_{k_1}, Cl_{k_2}) \in [0,1]$.

A dissimilarity matrix M_1 containing dissimilarities of clusters of s_1 according to clusters of s_2, and M_2 the dissimilarity matrix between clusters of s_2 and clusters of s_1 are defined as follows:

$$M_1 = \begin{pmatrix} \delta^1_{11} & \delta^1_{12} & \dots & \delta^1_{1C} \\ \dots & \dots & \dots & \dots \\ \delta^1_{k1} & \delta^1_{k2} & \dots & \delta^1_{kC} \\ \dots & \dots & \dots & \dots \\ \delta^1_{C1} & \delta^1_{C2} & \dots & \delta^1_{CC} \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} \delta^2_{11} & \delta^2_{12} & \dots & \delta^2_{1C} \\ \dots & \dots & \dots & \dots \\ \delta^2_{k1} & \delta^2_{k2} & \dots & \delta^2_{kC} \\ \dots & \dots & \dots & \dots \\ \delta^2_{C1} & \delta^2_{C2} & \dots & \delta^2_{CC} \end{pmatrix} \tag{5}$$

We note that $\delta^1_{k_1 k_2}$ is the dissimilarity between Cl_{k_1} of s_1 and Cl_{k_2} of s_2 and $\delta^2_{k_1 k_2}$ is the dissimilarity between Cl_{k_2} of s_2 and Cl_{k_1} of s_1 and $\delta^1_{k_1 k_2} = \delta^2_{k_2 k_1}$. The dissimilarity matrix M_2 of s_2 is the transpose of the dissimilarity matrix of s_1 noted M_1. Therefore, a unique matrix M_1 can be used to store dissimilarities between all clusters of s_1 and that of s_2. Clusters of s_1 are matched to the nearest clusters of s_2, a cluster Cl_{k_1} of s_1 is matched to the cluster having the minimal dissimilarity $\delta^1_{k_1 .}$ and a cluster Cl_{k_2} of s_2 is matched to the cluster having the minimal dissimilarity $\delta^2_{k_2 .} = \delta^1_{. k_2}$.Two clusters of s_1 can be linked to the same cluster of s_2. The output are C cluster matchings of s_1, C different cluster matchings of s_2 and $2 \times C$ dissimilarity values of each matched clusters.

3.3 Cluster Independence

Once cluster matching is obtained, the degree of independence and dependence between sources are quantified in this step. A set of matched clusters is obtained for both sources and a mass function can be used to quantify the independence between each couple of clusters. Suppose that the cluster Cl_{k_1} of s_1 is matched to Cl_{k_2} of s_2, a mass function m defined on the frame of discernment $\Omega_I = \{Dependent\ \ Dep, Independent\ \ Ind\}$ describes how much this couple of clusters is independent or dependent as follows:

$$\begin{cases} m^{\Omega_I}_{k_1 k_2}(Dep) = \alpha\,(1 - \delta^1_{k_1 k_2}) \\ m^{\Omega_I}_{k_1 k_2}(Ind) = \alpha\,\delta^1_{k_1 k_2} \\ m^{\Omega_I}_{k_1 k_2}(Dep \cup Ind) = 1 - \alpha \end{cases} \tag{6}$$

The coefficient α is used to take into account the number of mass functions in each cluster. Mass functions defining sources dependence are not provided by any source

whereas they are estimations of the sources dependence. α is not the reliability of any source but it can be seen as the reliability of the estimation. Therefore, the more a cluster contains mass functions the more our dependence measure estimation of that cluster is reliable. For example, let us take two clusters the first one containing only one mass function and the second one containing 100 mass functions, it is obvious that the dependency estimation of the second cluster is more precise and significant than the dependency estimation of the first one.

The obtained mass functions quantify the independence of each matched clusters according to each source. Therefore, C mass functions are obtained for each source such that each mass function quantifies the independence of each couple of matched clusters. The combination of C mass functions for each source using Dempster's rule of combination defined by equation (1) is a mass function m^{Ω_I} defining the whole dependence of one source towards the other one: $m^{\Omega_I} = \textcircled{\cap}\, m^{\Omega_I}_{k_1 k_2}$.

Two different mass functions $m^{\Omega_I}_{s_1}$ and $m^{\Omega_I}_{s_2}$ are obtained for s_1 and s_2 respectively. We note that $m^{\Omega_I}_{s_1}$ is the combination of C mass functions representing the dependence of matched clusters defined using equation (6). These mass functions are different since cluster matchings are different which verifies the axiom of non-symmetry. $\delta^1_{k_1 k_2}, \delta^2_{k_2 k_1} \in [0,1]$ which verifies the non-negativity and the normalization axioms. Finally, pignistic probabilities are computed from these mass functions in order to decide about these sources independence $\mathrm{I_d}$ such that $\mathrm{I_d}(s_1,s_2) = \mathrm{BetP}(Ind)$ and $\overline{\mathrm{I_d}}(s_1,s_2) = \mathrm{BetP}(Dep)$, if $\mathrm{BetP}(Ind) > 0.5$ we can claim that the corresponding source is independent from the other one otherwise it is dependent.

4 Examples on Generated Mass Functions

To test this method we used generated mass functions. Thus, two sets of mass functions are generated for two sources s_1 and s_2. We note that the number of sources is always two (s_1 and s_2) because the dependence is a binary relationship. Thus a source is dependent or independent according to another one. For the sake of simplicity, we take here the discounting factor $\alpha = 1$, thus mass functions are not discounted. To generate bbas, some information are needed: the cardinality of the frame of discernment $|\Omega|$, the number of mass functions. Mass functions are generated as follows:

1. The number of focal elements F is chosen randomly from $[1, |2^{\Omega}|]$. The F focal elements are also chosen randomly from the power set.
2. The interval $[0,1]$ is divided randomly into F continuous sub intervals.
3. A random mass from each sub interval is attributed to focal elements. Masses are attributed to focal elements chosen in the first step. The complement to 1 of the attributed masses sum is affected to the total ignorance $m(\Omega)$.

This method is used to generate a random mass function, thus the number of focal elements and masses are attributed randomly. Using the pignistic transformation, the decided class is not known from the beginning. In some cases generated mass functions are corrected in order to correct the classification result as follows:

i Generate a mass function as described above,
ii to change the classification result of the generated mass function, masses affected to each focal element are transfered to its union with the decided class.

Dependent sources: When sources are dependent, they are either providing similar belief functions with the same decided class (using the pignistic transformation) or one of the sources is saying the opposite of what says the other one. In the case of sources deciding the same class, the decided class of one source is directly affected by that of the other one. To test this case, we generated 100 mass functions on a frame of discernment of cardinality 5. Both sources are classifying objects in the same way. Applying the method described above, the obtained mass function defined on the frame $\Omega_I = \{Ind,\ Dep\}$ and describing the independence of s_1 towards s_2 is $m(Ind) = 0.0217$, $m(Dep) = 0.9783$ meaning that $\mathrm{I_d}(s_1,s_2) = 0.0217$ and $\overline{\mathrm{I_d}}(s_1,s_2) = 0.9783$. Thus s_1 is highly dependent on s_2.
The mass function of the independence of s_2 according to s_1 is $m(Ind) = 0.022$, $m(Dep) = 0.978$. It proves that s_2 is also dependent on s_1 because $\overline{\mathrm{I_d}}(s_2,s_1) = 0.978$.

When sources are indirectly dependent, one of them is saying the opposite of the other one. In other words, when the decision class of the first source is a class A, the second source may classify this object in any other class but not A. In that case, the obtained mass function for the dependence of s_1 according to s_2 is $m(Ind) = 0.0777$, $m(Dep) = 0.9223$ meaning that s_1 is dependent on s_2 because $\overline{\mathrm{I_d}}(s_1,s_2) = 0.9223$.
The mass function of the independence of s_2 according to s_1 is $m(Ind) = 0.0805$, $m(Dep) = 0.9195$, thus s_2 is also highly dependent on s_1 and $\overline{\mathrm{I_d}}(s_2,s_1) = 0.9195$.
Thus s_1 is dependent towards s_2 with a degree 0.978 and s_2 is dependent towards s_1 with a degree 0.9195. s_1 and s_2 are mutually dependent.

Independent sources: We generated randomly 100 mass functions for both sources s_1 and s_2. The number of focal elements is randomly chosen on the interval $[1, \frac{2^\Omega}{4}]$ rather than $[1, 2^\Omega]$ to reduce the number of focal elements. The obtained mass function of the independence of s_1 according to s_2 is $m(Ind) = 0.7211$, $m(Dep) = 0.2789$. The mass function of the independence of s_2 according to s_1 is $m(Ind) = 0.6375$, $m(Dep) = 0.3625$. Thus s_1 is independent towards s_2 because $\mathrm{I_d}(s_1,s_2) = 0.7211$ and s_2 is independent towards s_1 because $\mathrm{I_d}(s_2,s_1) = 0.6375$. s_1 and s_2 are mutually independent.

5 Conclusion

Combining mass functions provided by different sources is helpful when making decision. The choice of the combination rule is conditioned on the sources dependence, thus the cautious rule is especially used when sources are dependent but other rules can be applied with independent sources. In this paper, we suggested a method estimating the dependence degree of one source towards another one. As a future work, we may try to estimate the dependence of one source according to many other sources and not only one source. When one source is dependent on another one, this dependence can be direct (positive dependence) or indirect (negative dependence). Thus, we will also quantify the positive and negative dependence in the case of

dependent sources. We will also define the discounting factor which will be a function of the number of mass functions. Finally, we will use the discounting operator in order to take into account the number of provided mass functions because we cannot decide on the sources independence if they do not provide a sufficient number of mass functions.

References

1. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief function independence: I. The marginal case. International Journal Approximate Reasoning 29(1), 47–70 (2002)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief function independence: II. The conditional case. International Journal Approximate Reasoning 31(1-2), 31–75 (2002)
3. Ben Hariz, S., Elouedi, Z., Mellouli, K.: Clustering Approach Using Belief Function Theory. In: Euzenat, J., Domingue, J. (eds.) AIMSA 2006. LNCS (LNAI), vol. 4183, pp. 162–171. Springer, Heidelberg (2006)
4. Dempster, A.P.: Upper and Lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Denoeux, T.: The cautious rule of combination for belief functions and some extensions. In: Proceedings of the Ninth International Conference on Information Fusion (FUSION 2006), Florence, Italy (2006)
6. Dubois, D., Prade, H.: Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence 4, 244–264 (1988)
7. Martin, A., Osswald, C.: Toward a combination rule to deal with partial conflict and specificity in belief functions theory. In: Int. Conf. on Information Fusion, Québec, Canada (2007)
8. Jousselme, A.-L., Grenier, D., Bossé, E.: A new distance between two bodies of evidence. Information Fusion 2, 91–101 (2001)
9. Murphy, C.K.: Combining belief functions when evidence conflicts. Decision Support Systems 29, 1–9 (2000)
10. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligent 66, 191–234 (1994)
11. Smets, P.: Belief Functions: The Disjunctive Rule of Combination and the Generalized Bayesian Theorem. Classic Works of the Dempster-Shafer Theory of Belief Functions, 633–664 (2008)
12. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
13. Yager, R.R.: On the Dempster-Shafer Framework and New Combination Rules. Information Sciences 41, 93–137 (1987)

On Random Sets Independence and Strong Independence in Evidence Theory

Jiřina Vejnarová

Abstract. Belief and plausibility functions can be viewed as lower and upper probabilities possessing special properties. Therefore, (conditional) independence concepts from the framework of imprecise probabilities can also be applied to its sub-framework of evidence theory. In this paper we concentrate ourselves on random sets independence, which seems to be a natural concept in evidence theory, and strong independence, one of two principal concepts (together with epistemic independence) in the framework of credal sets. We show that application of strong independence to two bodies of evidence generally leads to a model which is beyond the framework of evidence theory. Nevertheless, if we add a condition on resulting focal elements, then strong independence reduces to random sets independence. Unfortunately, it is not valid no more for conditional independence.

1 Introduction

Imprecise probabilities is a general concept comprising different theories dealing with imprecise information. These theories can be partially ordered with respect to their generality and evidence theory belongs to the most specific ones. More precisely, belief and plausibility functions can be viewed as lower and upper probabilities, respectively, possessing special properties.

Independence belongs to the most important concepts within any theory dealing with uncertainty and therefore it has been studied in the evidential framework from the very beginning [11]. Because of reasons stated above, the application of independence concepts from imprecise probabilities to belief plausibility functions is, in principle, possible and their relationship to "natural" independence concepts in evidence theory is an interesting question, as already suggested in [5, 6, 8].

Jiřina Vejnarová
Institute of Information Theory and Automation of the AS CR, Pod Vodárenskou věží 4, Prague, Czech Republic
e-mail: vejnar@utia.cas.cz

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 247–254.
springerlink.com

In this paper we confine ourselves to random sets independence and strong independence and will not deal with epistemic irrelevance and independence, as they are based on conditional probabilities/beliefs and there does not exist a uniquely accepted conditioning rule [7] in the framework of evidence theory.

The paper is organized as follows. Section 2 is an overview of basic concepts from evidence theory and form credal sets and in Section 3 random sets independence and strong independence are introduced and their relationship in the framework of evidence theory is studied.

2 Basic Concepts

In this section we will briefly recall basic concepts from evidence theory [11] concerning sets and set functions and from the framework of credal sets [10].

2.1 Set Projections and Joins

For an index set $N = \{1,2,\ldots,n\}$ let $\{X_i\}_{i\in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$. In this paper we will deal with *multidimensional frame of discernment* $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$, and its *subframes* (for $K \subseteq N$)

$$\mathbf{X}_K = \bigtimes_{i\in K} \mathbf{X}_i.$$

When dealing with groups of variables on these subframes, X_K will denote a group of variables $\{X_i\}_{i\in K}$ throughout the paper.

A *projection* of $x = (x_1,x_2,\ldots,x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1,i_2,\ldots,i_k\}$

$$x^{\downarrow K} = (x_{i_1},x_{i_2},\ldots,x_{i_k}) \in \mathbf{X}_K.$$

Analogously, for $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of A into $\mathbf{X}_M$:

$$A^{\downarrow M} = \{y \in \mathbf{X}_M \mid \exists x \in A : y = x^{\downarrow M}\}.$$

In addition to the projection, in this text we will also need an opposite operation, which will be called a join. By a *join* [1] of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K,L \subseteq N)$ we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K\cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Let us note that for any $C \subseteq \mathbf{X}_{K\cup L}$ naturally $C \subseteq C^{\downarrow K} \bowtie C^{\downarrow L}$, but generally $C \neq C^{\downarrow K} \bowtie C^{\downarrow L}$, i.e., a join is, in a sense, a generalization of a rectangle — so called $X^{\downarrow K\cap L}$-*layered* rectangle [3].

2.2 Set Functions

In evidence theory [11] (or Dempster-Shafer theory) two dual measures are used to model the uncertainty: belief and plausibility measures. Both of them can be defined with the help of another set function called a *basic (probability* or *belief) assignment m* on $\mathbf{X}_N$, i.e., $m : \mathscr{P}(\mathbf{X}_N) \longrightarrow [0,1]$, where $\mathscr{P}(\mathbf{X}_N)$ is power set of $\mathbf{X}_N$ and $\sum_{A \subseteq \mathbf{X}_N} m(A) = 1$. Furthermore, we assume that $m(\emptyset) = 0$. A set $A \in \mathscr{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$.

Belief and *plausibility measures* are defined for any $A \subseteq \mathbf{X}_N$ by the equalities

$$Bel(A) = \sum_{B \subseteq A} m(B), \qquad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B),$$

respectively. It is well-known (and evident from these formulae) that for any $A \in \mathscr{P}(\mathbf{X}_N)$

$$Bel(A) \leq Pl(A), \qquad Pl(A) = 1 - Bel(A^C), \tag{1}$$

where A^C is the set complement of $A \in \mathscr{P}(\mathbf{X}_N)$.

Because of (1) belief and plausibility functions may be viewed as lower and upper probabilities, respectively. Furthermore, basic assignment can be computed from belief function via Möbius inversion:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B), \tag{2}$$

i.e. any of these three functions is sufficient to define values of the remaining two.

For a basic assignment m on $\mathbf{X}_K$ and $M \subset K$, a *marginal basic assignment* of m on $\mathbf{X}_M$ is defined (for each $A \subseteq \mathbf{X}_M$):

$$m^{\downarrow M}(A) = \sum_{\substack{B \subseteq \mathbf{X}_K \\ B^{\downarrow M} = A}} m(B).$$

Analogously we will denote by $Bel^{\downarrow M}$ marginal belief measure on $\mathbf{X}_M$.

2.3 Credal Sets

A *credal set* $\mathscr{M}(X)$ about a variable X is defined as a closed convex set of probability measures about the values of this variable. In order to simplify the expression of operations with credal sets, it is often considered [10] that a credal set is the set of probability distributions associated to the probability measures in it. Under such consideration a credal set can be expressed as a *convex hull* of its extreme distributions

$$\mathscr{M}(X) = \mathrm{CH}\{\mathrm{ext}(\mathscr{M}(X))\}.$$

Any lower probability $\underline{P}$ can be associated with a credal set of probabilities dominating it:

$$\mathscr{M}(\underline{P}) = \mathrm{CH}\{P : P(A) \geq \underline{P}(A), A \subseteq \mathbf{X}\}.$$

As belief measure is a lower probability, this association can be done also for it, as suggested in both examples in the next section.

3 Independence Concepts

3.1 Random Sets Independence

Let us start this section by recalling the notion of random sets independence [4].

Definition 1. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be disjoint. We say that groups of variables X_K and X_L are *independent with respect to basic assignment* m if

$$m^{\downarrow K\cup L}(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L}) \tag{3}$$

for all $A \subseteq \mathbf{X}_{K\cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise.

Example 1. Consider two basic assignments m_X and m_Y on $\mathbf{X} = \{x, \bar{x}\}$ and $\mathbf{Y} = \{y, \bar{y}\}$, respectively, specified in Table 1 together with their beliefs and plausibilities. Under the assumption of random sets independence we get the joint basic assignment m, values of which are contained in the second column of Table 2. In third and fourth columns one can find beliefs and plausibilities of the corresponding sets, respectively. ♢

Table 1 Basic assignments m_X and m_Y.

$A \subseteq \mathbf{X}$	$m_X(A)$	$Bel_X(A)$	$Pl_X(A)$
$\{x\}$	0.3	0.3	0.8
$\{\bar{x}\}$	0.2	0.2	0.7
$\mathbf{X}$	0.5	1	1

$A \subseteq \mathbf{Y}$	$m_Y(A)$	$Bel_Y(A)$	$Pl_Y(A)$
$\{y\}$	0.6	0.6	0.9
$\{\bar{y}\}$	0.1	0.1	0.4
$\mathbf{Y}$	0.3	1	1

There exist numerous generalizations [3, 9, 12] of this notion to the conditional case. For the reasons presented e.g. in [9], we use the following one.

Definition 2. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. We say that groups of variables X_K and X_L are *conditionally independent given* X_M *with respect to* m (and denote it by $K \perp\!\!\!\perp L | M\ [m]$), if the equality

$$m^{\downarrow K\cup L\cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K\cup M}(A^{\downarrow K\cup M}) \cdot m^{\downarrow L\cup M}(A^{\downarrow L\cup M}) \tag{4}$$

holds for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$ such that $A = A^{\downarrow K\cup M} \bowtie A^{\downarrow L\cup M}$, and $m(A) = 0$ otherwise.

Table 2 Results of application of random sets independence (Col. 2–4) and strong independence (Col. 5–7).

$C \subseteq \mathbf{X} \times \mathbf{Y}$	$m_R(C)$	$Bel_R(C)$	$Bel_R(C)$	$\underline{P}_{XY}(C)$	$\overline{P}_{XY}(C)$	$\tilde{m}_{XY}(C)$
$\{xy\}$	0.18	0.18	0.72	0.18	0.72	0.18
$\{x\bar{y}\}$	0.03	0.03	0.32	0.03	0.32	0.03
$\{\bar{x}y\}$	0.12	0.12	0.63	0.12	0.63	0.12
$\{\bar{x}\bar{y}\}$	0.02	0.02	0.28	0.02	0.28	0.02
$\{x\} \times \mathbf{Y}$	0.09	0.3	0.8	0.3	0.8	0.09
$\{\bar{x}\} \times \mathbf{Y}$	0.06	0.2	0.7	0.2	0.7	0.06
$\mathbf{X} \times \{y\}$	0.3	0.6	0.9	0.6	0.9	0.3
$\mathbf{X} \times \{\bar{y}\}$	0.05	0.1	0.4	0.1	0.4	0.05
$\{xy, \bar{x}\bar{y}\}$	0	0.2	0.85	0.34	0.74	0.14
$\{x\bar{y}, \bar{x}y\}$	0	0.15	0.8	0.26	.66	0.11
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}\bar{y}\}$	0	0.72	0.98	0.72	0.98	−0.11
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}y\}$	0	0.37	0.88	0.37	0.88	−0.14
$\mathbf{X} \times \mathbf{Y} \setminus \{x\bar{y}\}$	0	0.68	0.97	0.68	0.97	−0.14
$\mathbf{X} \times \mathbf{Y} \setminus \{xy\}$	0	0.28	0.82	0.28	0.82	−0.11
$\mathbf{X} \times \mathbf{Y}$	0.15	1	1	1	1	0.4

3.2 Strong Independence

From numerous definitions of independence for credal sets [4] we have chosen strong independence, as it seems to be most proper for multidimensional models.

We say that X_K and X_L are *strongly independent* with respect to $\mathscr{M}(X_K X_L)$ iff (in terms of probability distributions)

$$\mathscr{M}(X_K X_L) = \mathrm{CH}\{P_1 \cdot P_2 : P_1 \in \mathscr{M}(X_K), P_2 \in \mathscr{M}(X_L)\}. \tag{5}$$

Again, there exist several generalizations of this notion to conditional independence, see e.g. [10], but as the following definition is suggested by the authors as the most appropriate for the marginal problem, it seems to be a suitable counterpart of random sets independence.

Given three variables X, Y and Z we say that X and Y are *independent on the distribution* given Z under global set $\mathscr{M}(X,Y,Z)$ iff

$$\mathscr{M}(X,Y,Z) = \{(p_1 \cdot p_2)/p_1^{\downarrow Z} : p_1 \in \mathscr{M}(X,Z), p_2 \in \mathscr{M}(X,Z), p_1^{\downarrow Z} = p_2^{\downarrow Z}\}.$$

From the term "strong independence" one could deduce that it should imply random sets independence. Nevertheless, it is not true, as can be seen from the following simple example.

Example 1. (Continued) From values contained in Table 1 we obtain credal sets about variables X and Y:

$$\mathscr{M}(X) = \mathrm{CH}\{(0.3,0.7),(0.8,0.2)\}, \quad \mathscr{M}(Y) = \mathrm{CH}\{(0.6,0.4),(0.9,0.1)\}.$$

Under the assumption of strong independence we get

$$\mathscr{M}(XY) = \mathrm{CH}\{(0.18,0.12,0.42,0.28),(0.27,0.03,0.63,0.07),\\ (0.48,0.32,0.12,0.08),(0.72,0.08,0.18,0.02)\}.$$

Let us compute lower and upper probabilities of all nonempty subsets of $\mathbf{X} \times \mathbf{Y}$. Their values can be found in fifth and sixth columns of Table 2.

In the last column one can find hypothetical values of basic assignment corresponding the these lower and upper probabilities taken as beliefs and plausibilities computed via formula (2). From this column one can see that X and Y do not satisfy random set independence, as m_{XY} assigns positive values also to subsets which are not of the form $A = B \times C$. Furthermore, negative values are assigned to some sets, which violates the nonnegativity of basic assignment, i.e. we are beyond the limits of evidence theory. ♢

This result led us to the conclusion that strong independence cannot be applied in the framework of evidence theory. Nevertheless, under specific conditions it can be done as the following theorem[1] holds true.

Theorem 1. *Let X_K and X_L ($K \cap L \neq \emptyset$) be two groups of variables with basic assignments $m^{\downarrow K}$ and $m^{\downarrow L}$, respectively. Let $Bel^{\downarrow K\cup L}$ and $\underline{P}^{\downarrow K\cup L}$ denote the joint belief function under random sets independence and joint lower probability under strong independence, respectively, and let A be a subset of $\mathbf{X}_K \times \mathbf{X}_L$ such that $A = A^{\downarrow K} \times A^{\downarrow L}$. Then*

$$Bel^{\downarrow K\cup L}(A) = \underline{P}^{\downarrow K\cup L}(A). \tag{6}$$

Proof. It is well-known[2] that for random sets independence the following equality holds true for any $A = A^{\downarrow K} \times A^{\downarrow L}$:

$$Bel^{\downarrow K\cup L}(A) = Bel^{\downarrow K}(A^{\downarrow K}) \cdot Bel^{\downarrow L}(A^{\downarrow L}).$$

Taking into account the fact that

$$Bel^{\downarrow K}(A^{\downarrow K}) = \underline{P}^{\downarrow K}(A^{\downarrow K}), \qquad Bel^{\downarrow L}(A^{\downarrow L}) = \underline{P}^{\downarrow L}(A^{\downarrow L}),$$

to get (6) it is enough to prove that for any $A \subseteq \mathbf{X}_K \times \mathbf{X}_L$ such that $A = A^{\downarrow K} \times A^{\downarrow L}$ the equality

[1] Let us note that the content of this theorem was already mentioned (without a proof) in [4].

[2] In [2] equality (7) together with an analogous one for plausibilities is used as a definition of *evidential independence* and Definition 1 is presented as an equivalent characterization.

$$\underline{P}^{\downarrow K\cup L}(A) = \underline{P}^{\downarrow K}(A^{\downarrow K}) \cdot \underline{P}^{\downarrow L}(A^{\downarrow L})$$

is satisfied.

Generally,

$$\underline{P}^{\downarrow K\cup L}(A) = \min_{P\in\mathscr{M}}\{\sum_{x\in A} P(x)\},$$

but as $\mathscr{M} = \{P_1 \cdot P_2 : P_1 \in \mathscr{M}_K, P_2 \in \mathscr{M}_L\}$, and $A = A^{\downarrow K} \times A^{\downarrow L}$ then

$$\begin{aligned}\underline{P}^{\downarrow K\cup L}(A) &= \min_{P\in\mathscr{M}}\{\sum_{x\in A} P(x)\} = \min_{P=P_1\cdot P_2, P_1\in\mathscr{M}_K, P_2\in\mathscr{M}_L}\{\sum_{x_K\in A^{\downarrow K}} P(x_K)\cdot \sum_{x_L\in A^{\downarrow L}} P(x_L)\}\\ &= \min_{P_1\in\mathscr{M}_K}\{\sum_{x_K\in A^{\downarrow K}} P(x_K)\}\cdot \min_{P_2\in\mathscr{M}_L}\{\sum_{x_L\in A^{\downarrow L}} P(x_L)\} = \underline{P}^{\downarrow K}(A^{\downarrow K})\cdot\underline{P}^{\downarrow L}(A^{\downarrow L}),\end{aligned}$$

as requested (where the last but one equality holds thanks to the fact that we deal with non-negative numbers.) □

Unfortunately, for conditional independence an analogous result does not hold.

Example 2. Let X, Y and Z be three binary variables with values in $\mathbf{X} = \{x, \bar{x}\}$, $\mathbf{Y} = \{y, \bar{y}\}$ and $\mathbf{Z} = \{z, \bar{z}\}$, respectively, and m_{XZ} and m_{YZ} two basic assignments, both of them having only two focal elements:

$$\begin{aligned} m_{XZ}(\{(x,\bar{z}),(\bar{x},\bar{z})\}) &= 0.5, & m_{XZ}(\{(u,\bar{z}),(\bar{x},z)\}) &= 0.5,\\ m_{YZ}(\{(y,\bar{z}),(\bar{y},\bar{z})\}) &= 0.5, & m_{YZ}(\{(y,\bar{z}),(\bar{y},z)\}) &= 0.5.\end{aligned}$$

Applying Definition 2 one can easily obtain the following joint assignment:

$$m(\mathbf{X}\times\mathbf{Y}\times\{\bar{z}\}) = 0.5, \qquad m(\{(x,y,\bar{z}),(\bar{x},\bar{y},z)\}) = 0.5.$$

From the values of the basic assignments m_{XZ} we will obtain credal set

$$\mathscr{M}(XZ) = \mathrm{CH}\{(0,1,0,0),(0,.5,0,.5),(0,.5,0.5,0),(0,0,.5,.5)\},$$

and credal set $\mathscr{M}(YZ)$ is identical. We can see, that the first two probability distributions are projective and the remaining two as well. Therefore under the assumption of strong conditional independence we will get the following joint credal set

$$\begin{aligned}\mathscr{M}(XYZ) = \mathrm{CH}\{&(0,1,0,0,0,0,0,0),(0,.5,0,.5,0,0,0,0),(0,.5,0,0,0,.5,0,0),\\ &(0,.25,0,.25,0,.25,0,.25),(0,.5,0,0,0,0,.5,0),\\ &(0,.0,0,.5,0,0,.5,0),(0,0,0,0,0,.5,.5,0),(0,0,0,0,0,0,.5,.5)\}.\end{aligned}$$

From $\mathscr{M}(XYZ)$ we can easily get values of lower and upper probabilities of all singletons as well as values of bigger subsets. For example, for the above mentioned focal elements we have

$$\underline{P}(\mathbf{X}\times\mathbf{Y}\times\{v\}) = 0.5, \quad \underline{P}(\{(u,u,v),(v,v,u)\}) = 0.25,$$

i.e., the latter is different from that obtained under random sets independence. ♢

4 Conclusions

The aim of this paper was to clarify the relationship between random sets independence and strong independence in the framework of evidence theory. Although evidence theory can be viewed as a special case of imprecise probabilities, application of strong independence may lead to models which are beyond the framework of evidence theory. If we confine ourselves to rectangles, values of joint belief function (under random sets independence) and those of joint lower probability (under strong independence) coincide. Nevertheless, an analogous result does not hold in the conditional case.

The problem of (epistemic) irrelevance was not discussed here, as the properties of irrelevance are dependent on the conditioning rule in question, and the problem of conditioning in evidence theory has not yet been satisfactorily solved.

Acknowledgements. The support of Grant GAČR P402/11/0378 is gratefully acknowledged.

References

1. Beeri, C., Fagin, R., Maier, D., Yannakakis, M.: On the desirability of acyclic database schemes. J. of the Association for Computing Machinery 30, 479–513 (1983)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief functions independence: I. the marginal case. Int. J. Approx. Reasoning 29, 47–70 (2002)
3. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief functions independence: II. the conditional case. Int. J. Approx. Reasoning 31, 31–75 (2002)
4. Couso, I., Moral, S., Walley, P.: Examples of independence for imprecise probabilities. In: de Cooman, G., Cozman, F.G., Moral, S., Walley, P. (eds.) Proceedings of ISIPTA 1999, Ghent, June 29-July 2, pp. 121–130 (1999)
5. Couso, I., Moral, S.: Independence Concepts in Evidence Theory. Int. J. Approx. Reasoning 51, 748–758 (2010)
6. Destercke, S.: Independence concepts in evidence theory: some results about epistemic irrelevance and imprecise belief functions. In: Proceedings of BELIEF 2010 (2010)
7. Fagin, R., Halpern, J.Y.: A new approach to updating beliefs. In: Bonissone, et al. (eds.) Uncertainty in Artifiacial Intelligence, vol. VI, pp. 347–374. Elsevier (1991)
8. Fetz, T.: Sets of joint probability measures generated by weighted marginal focal sets. In: de Cooman, G., Cozman, F.G., Fine, T., Moral, S. (eds.) Proceedings of ISIPTA 2001, Ithaca, June 26-29, pp. 201–210 (2001)
9. Jiroušek, R., Vejnarová, J.: Compositional models and conditional independence in Evidence Theory. Int. J. Approx. Reasoning 52, 316–334 (2011)
10. Moral, S., Cano, A.: Strong conditional independence for credal sets. Ann. of Math. and Artif. Intell. 35, 295–321 (2002)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Shenoy, P.P.: Conditional independence in valuation-based systems. Int. J. Approx. Reasoning 10, 203–234 (1994)

Combining Linear Equation Models via Dempster's Rule

Liping Liu

Abstract. This paper proposes a concept of imaginary extreme numbers, which are like traditional complex number $a + bi$ but with $i = \sqrt{-1}$ being replaced by $e = 1/0$, and defines usual operations such as addition, subtraction, and division on the numbers. It applies the concept to representing linear equations in knowledge-based systems. It proves that the combination of linear equations via Dempster's rule is equivalent to solving a system of simultaneous equations or finding a least-square estimate when they are overdetermined.

1 Introduction

The concept of linear belief functions unifies the representation of a diverse range of linear models in expert systems [Liu et al., 2006]. These linear models include linear equations that characterize linear deterministic relationships of continuous or discrete variables and stochastic models such as linear regressions, linear time series, and Kalman filters in which some variables are deterministic while others stochastic. They also include normal distributions that describe probabilistic knowledge on a set of variables, a lack of knowledge such as ignorance and partial ignorance, and direct observations or observations with missing values. Despite the varieties, the concept of linear belief functions unifies them as manifestations of a single concept, represents them as matrices with the same semantics, and combine them by a single mechanism, the matrix addition rule, which is consistent with Dempster's rule of combination [Shafer, 1976].

What makes the unification possible is the sweeping operator. Nevertheless, when the operator is applied to knowledge representation, a division-by-zero enigma often arises. For example, when two linear models are combined, their matrix representations must be fully swept via the old matrix addition rule [Dempster, 2001] or partially swept via the new matrix addition rule [Liu, 2011b]. This poses no issue

Liping Liu
The University of Akron

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 255–265.
springerlink.com

to linear models with a positive definite covariance matrix [Liu, 2011a]. However, for deterministic linear models such as linear equations, sweeping points are often zero, and a sweeping, if needs to be done, will have to divide regular numerical values by zero, a mathematical operation that is not defined. The division-by-zero issue has been a challenge that hinders the development of intelligent systems that implements linear belief functions.

In this paper, I propose a notion of imaginary extreme numbers to deal with the division-by-zero problem. An imaginary extreme number is a complex number like $3+4e$ with extreme number $e=\frac{1}{0}$. On these imaginary numbers, usual operations can be defined. The notion of imaginary extreme numbers makes it possible to represent linear equations as knowledge in intelligent systems. As we will illustrate, a linear equation is transformed into an equivalent one by a sweeping from a zero variance and a reverse sweeping from an extreme inverse variance. The notion also makes it possible to combine linear equations as independent pieces of knowledge via Dempster's rule of combination. We will show that the combination of linear equations corresponds to solving the equations or finding the least-square estimate when the equations are over-determining.

2 Matrix Sweepings

Sweeping is a matrix transformation that starts from *a sweeping point*, a square submatrix, and iteratively spreads the change across the entire matrix:

Definition 1. Assume real matrix A is made of submatrices as

$$A=(A_{ij})$$

and assume A_{ij} is a square submatrix. Then a forward (reverse) sweeping of A from A_{ij} replaces submatrix A_{ij} by its negative inverse $-(A_{ij})^{-1}$, any other submatrix A_{ik} in row i and any submatrix A_{kj} in column j are respectively replaced by $(-)(A_{ij})^{-1}A_{ik}$ and $(-)A_{kj}(A_{ij})^{-1}$, and the remaining submatrix A_{kl} not in the same row or column as A_{ij}, i.e., $k\neq i$ and $j\neq l$, by

$$A_{kl}-A_{kj}(A_{ij})^{-1}A_{il}.$$

Note that forward and reverse sweepings defined above operationally differ only in the sign for the elements in the same column or row as the sweeping point. Yet the difference is significant in that forward and reverse sweepings cancel each other's effects, and thus the modifiers "forward" and "reverse" are justified. Both forward and reverse sweeping operations may be also defined to sweep from a square submatrix as a sweeping point. If a sweeping point is positive definite such as a covariance matrix, then a sweeping from the submatrix is equivalent to a series of successive sweepings from each of the leading diagonal elements of the submatrix [Liu, 2011a].

When applied to a moment matrix that consists of a mean vector and a covariance matrix, sweeping operations can transform a normal distribution to its various forms,

each with interesting semantics. Assume X has mean vector μ and covariance matrix Σ. Then in general the moment matrix is

$$M(X) = \begin{bmatrix} \mu \\ \Sigma \end{bmatrix}$$

and its fully swept form

$$M(\overrightarrow{X}) = \begin{bmatrix} \mu\Sigma^{-1} \\ -\Sigma^{-1} \end{bmatrix}$$

represents the density function of X. Note $M(\overrightarrow{X})$ symbolizes that $M(X)$ has been swept from the covariance matrix of X, or to be brief, that $M(X)$ has been swept from both X. It is interesting to imagine that, if the variances of X are so huge that their inverse covariance matrix $\Sigma^{-1} \longrightarrow 0$, then $M(\overrightarrow{X}) = 0$. Thus, a zero fully swept matrix is the representation of ignorance; intuitively, we are ignorant about X if its variances are infinite. A partial sweeping has more interesting semantics. For example, for the normal distribution of X, Y, and Z with moment matrix:

$$M(X,Y,Z) = \begin{bmatrix} 3 & 4 & 2 \\ 4 & 2 & 0 \\ 2 & 5 & 2 \\ 0 & 2 & 6 \end{bmatrix},$$

its sweeping from the variance terms for X and Y is a partially swept matrix

$$M(\overrightarrow{X},\overrightarrow{Y},Z) = \begin{bmatrix} 0.4375 & 0.625 & 0.75 \\ -0.3125 & 0.125 & -0.25 \\ 0.125 & -0.25 & 0.5 \\ -0.25 & 0.5 & 5 \end{bmatrix}.$$

This contains two distinct pieces of information about the variables [Liu, 2011a]. First, the submatrix corresponding to variables X and Y,

$$M(\overrightarrow{X},\overrightarrow{Y}) = \begin{bmatrix} 0.4375 & 0.625 \\ -0.3125 & 0.125 \\ 0.125 & -0.25 \end{bmatrix}$$

represents the density function of X and Y. Second, the remaining partial matrix

$$\begin{bmatrix} & & 0.75 \\ & & -0.25 \\ & & 0.5 \\ -0.25 & 0.5 & 5 \end{bmatrix}$$

represents a regression model $Y = 0.75 - 0.25X + 0.5Y + \varepsilon$ with $\varepsilon \sim N(0,5)$. Since this regression model alone casts no information on independent variables X and Y, the missing elements in the above partial matrix shall be zero. Furthermore,

when the conditional variance of Z vanishes, the conditional distribution reduces to a regular linear equation model $Z = 0.75 - 0.25x + 0.5y$ as represented by the matrix:

$$M(\overrightarrow{X}, \overrightarrow{Y}, Z) = \begin{bmatrix} 0 & 0 & 0.75 \\ 0 & 0 & -0.25 \\ 0 & 0 & 0.5 \\ -0.25 & 0.5 & 0 \end{bmatrix}.$$

Here $M(\overrightarrow{X}, \overrightarrow{Y}, Z)$ represents a generic moment matrix of X, Y, and Z with X and Y being swept. Note that it has been long realized that a linear model such as a regression model or a linear equation is a special case of a multivariate normal distribution [Khatri, 1968].What is new, however, is that with sweeping operations, it can be uniformly represented as a moment matrix or its partially swept form.

3 Imaginary Numbers

In this section I propose a new type of imaginary numbers, called *extreme numbers*, and use it to resolve the division-by-zero issue. Just as a usual imaginary number uses i for non-existent $\sqrt{-1}$, we use e for $\frac{1}{0}$, which also does not exist. Also, as a usual imaginary number consists of two parts, a real part and an imaginary part, an imaginary extreme number consists of the same two parts. For example, $3 - 2e$ is an extreme number with 3 as real part and -2 as imaginary part. When imaginary part vanishes, an extreme number reduces to a real one. When its imaginary part is nonzero, we call an extreme number *true extreme number*. When its real part is zero, we call the extreme number *pure extreme*. When both real and imaginary parts are zero, the extreme number is zero, i.e., $a + be = 0$ if and only if $a = 0$ and $b = 0$. Thus, the system of extreme numbers includes real numbers as a subset.

Extreme numbers may be added, subtracted, or scaled as usual imaginary numbers. For any extreme number $a + be$ and a real number c, their multiplication, or scaling of $a + be$ using scale c is defined as $c(a + be) = (a + be)c = ac + bce$.For any two extreme number $a_1 + b_1e$ and $a_2 + b_2e$, their addition is defined as $(a_1 + b_1e) + (a_2 + b_2e) = (a_1 + a_2) + (b_1 + b_2)e$.Clearly, the system of extreme numbers is closed under the operation of scaling, addition, and subtraction.

Unlike usual imaginary numbers, the multiplication of two extreme numbers is not defined because it is not closed operationally. However, division can be defined here: for any two extreme number $a_1 + b_1e$ and $a_2 + b_2e$, their division is defined as follows:

$$\frac{a_1 + b_1e}{a_2 + b_2e} = \frac{b_1}{b_2}$$

if $b_2 \neq 0$. If the denominator is a nonzero real number, then division reduces to scaling. If the denominator is zero and the numerator is one, i.e., $b_1 = 0$ and $a_1 = 1$,

the division is $e = 1/0$ via definition. Also, $0/0$ is defined to be 0 to be consistent with scaling, i.e., $0(0+1e) = 0+0e = 0$.

Because division generally cancels out imaginary parts, the operation of multiplication followed by division, called *crossing*, can be defined. For any three extreme numbers a_1+b_1e, a_2+b_2e, and a_3+b_3e, their crossing is defined as follows:

$$\frac{(a_1+b_1e)(a_2+b_2e)}{a_3+b_3e} = \frac{a_1b_2+a_2b_1}{b_3} + \frac{b_1b_2}{b_3}e$$

if $b_3 \neq 0$. Crossing reduces to division if one of the multiplicants a_1+b_1e and a_2+b_2e is real, i.e., $b_1b_2 = 0$. If at the same time the denominator is a nonzero real number, i.e., $b_3 = 0$ and $a_3 \neq 0$, it is reduced to scaling. It is consistent with the definition of extreme numbers if the dividor $a_3+b_3e = 0$, and $b_1 = 0$, $b_2 = 0$.

Extreme numbers may be extended to extreme matrices with the inverse of zero matrix being defined as $0^{-1} = Ie$, where I is an identity matrix. In general, $A+Be$ with real part A and imaginary part B, where both A and B are of the same dimensions. Operations on extreme matrices can be adopted from those for extreme numbers with slight modifications on division and crossing. For any two extreme matrices A_1+B_1e and A_2+B_2e, if B_2 is nonsingular, then

$$(A_1+B_1e)(A_2+B_2e)^{-1} = B_1(B_2)^{-1}$$
$$(A_2+B_2e)^{-1}(A_1+B_1e) = (B_2)^{-1}B_1$$

For any three extreme matrices A_1+B_1e, A_2+B_2e, and A_3+B_3e, if B_3 is nonsingular, then their crossing is defined as

$$A_1(B_3)^{-1}B_2 + B_1(B_3)^{-1}A_2 + B_1(B_3)^{-1}B_2e.$$

4 Equation Combination

Intuitively, a linear equation carries partial knowledge on the values of some variables through a linear relationship with other variables. If each of such equations is considered an independent piece of knowledge, its combination with other similar knowledge will render the values more certain. When there exist sufficient number of linear equations, their combination may jointly determine a specific value of the variables with complete certainty. Therefore, the combination of linear equations should correspond to solving a system of simultaneous equations. In this section, we will prove this statement.

In genearl, a linear equation may be expressed explicitly as

$$X_n = b + a_1X_1 + a_2X_2 + ... + a_{n-1}X_{n-1} \tag{1}$$

or implicitly as

$$a_1X_1 + a_2X_2 + ... + a_{n-1}X_{n-1} + a_nX_n = b. \tag{2}$$

The matrix representation for the explicit expression is straightforward:

$$M(\overrightarrow{X_1},...,\overrightarrow{X_{n-1}},X_n) = \begin{bmatrix} 0 & ... & 0 & b \\ 0 & ... & 0 & a_1 \\ ... & ... & ... & ... \\ 0 & ... & 0 & a_{n-1} \\ a_1 & ... & a_{n-1} & 0 \end{bmatrix}.$$

This partially swept matrix indicates that we have ignorance on the values of X_1, X_2, ..., and X_{n-1}; thus they correspond to a zero submatrix in the fully swept form. While given X_1, X_2, ..., and X_{n-1}, the value of X_n is b for sure; thus its conditional mean and variance are respectively b and zero. Of course, in algebra, a variable on the right-hand-side can be moved to the left-hand-side through a linear transformation. For example, if $a_1 \neq 0$, Equation 1 can be equivalently turned into

$$X_1 = -\frac{b}{a_1} - \frac{a_2}{a_1}X_2 - ... - \frac{a_{n-1}}{a_1}X_{n-1} + \frac{1}{a_1}X_n.$$

This transformation can be also done through the sweepings of matrix representations first by a foward sweeping from X_n and then a backward sweeping from X_1.

An implicit expression like Equation 2 may be represented as two separate linear equations in explicit forms:

$$a_1X_1 + a_2X_2 + ... + a_{n-1}X_{n-1} + a_nX_n = U$$

and $U = b$. Their matrices are respectively

$$M_1(\overrightarrow{X_1},...,\overrightarrow{X_n},U) = \begin{bmatrix} 0 & ... & 0 & 0 \\ 0 & ... & 0 & a_1 \\ ... & ... & ... & ... \\ 0 & ... & 0 & a_n \\ a_1 & ... & a_n & 0 \end{bmatrix}$$

and

$$M_2(U) = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

To combine them via Dempster's rule, we sweep both matrirces from U respectively into $M_1(\overrightarrow{X_1},...,\overrightarrow{X_n},\overrightarrow{U})$ as

$$\begin{bmatrix} 0 & ... & 0 & 0 \\ -(a_1)^2e & ... & -a_1a_ne & a_1e \\ ... & ... & ... & ... \\ -a_na_1e & ... & -(a_n)^2e & a_ne \\ a_1e & ... & a_ne & -e \end{bmatrix}$$

and

$$M_2(\overrightarrow{U}) = \begin{bmatrix} be \\ -e \end{bmatrix},$$

and then add the results position-wise into $M(\overrightarrow{X_1},...,\overrightarrow{X_n},\overrightarrow{U})$ as

$$\begin{bmatrix} 0 & ... & 0 & be \\ -(a_1)^2 e & ... & -a_1 a_n e & a_1 e \\ ... & ... & ... & ... \\ -a_n a_1 e & ... & -(a_n)^2 e & a_n e \\ a_1 e & ... & a_n e & -2e \end{bmatrix}.$$

To remove the auxiliary variable U, we shall unsweept $M(\overrightarrow{X_1},...,\overrightarrow{X_n},\overrightarrow{U})$ from U into $M(\overrightarrow{X_1},...,\overrightarrow{X_n},U)$ and then remove U by projecting the result to the variables X_1, X_2, ..., and X_n. We will obtain a fully swept matrix representation $M(\overrightarrow{X_1},...,\overrightarrow{X_n})$ for the implicit linear equation 2 as

$$\begin{bmatrix} \frac{1}{2} b \left(a_1 \ ... \ a_n \right) e \\ -\frac{1}{2} \begin{pmatrix} a_1 \\ ... \\ a_n \end{pmatrix} \left(a_1 \ ... \ a_n \right) e \end{bmatrix} \tag{3}$$

Assume coefficient $a_n \neq 0$, we can then unswept it from X_n and obtain $M(\overrightarrow{X_1},...,\overrightarrow{X_{n-1}},X_n)$ as

$$\begin{bmatrix} 0 & ... & 0 & b/a_n \\ 0 & ... & 0 & -a_1/a_n \\ ... & ... & ... & ... \\ 0 & ... & 0 & -a_{n-1}/a_n \\ -a_1/a_n & ... & -a_{n-1}/a_n & 0 \end{bmatrix},$$

which is the matrix representation for an explicit form for equation 2:

$$X_n = \frac{b}{a_n} - \frac{a_1}{a_n} X_1 - ... - \frac{a_{n-1}}{a_n} X_{n-1}.$$

Now let us study the representation and combination of multiple linear equations. For explicit expressions, without loss of generality, assume two linear equations are respectively $Y = b_1 + XA_1$ and $Y = b_2 + XA_2$, where Y is a single variable, X is n dimensional horizontal vector, $b_1$and b_2 are constant values, and A_1 and A_2 are n dimensional vertical vectors. Their matrix representations are

$$M_1(\overrightarrow{X},Y) = \begin{bmatrix} 0 & b_1 \\ 0 & A_1 \\ (A_1)^T & 0 \end{bmatrix},$$

$$M_2(\overrightarrow{X},Y) = \begin{bmatrix} 0 & b_2 \\ 0 & A_2 \\ (A_2)^T & 0 \end{bmatrix}.$$

To combine them, we need to sweep both matrices from Y and then add them position-wise into $M(\overrightarrow{X},\overrightarrow{Y})$ as

$$\begin{bmatrix} -b_1(A_1)^T e - b_2(A_2)^T e & (b_1+b_2)e \\ -A_1(A_1)^T e - A_2(A_2)^T e & (A_1+A_2)e \\ (A_1+A_2)^T e & -2e \end{bmatrix}.$$

Now unsweeping $M(\overrightarrow{X}, \overrightarrow{Y})$ from Y, we obtain $M(\overrightarrow{X}, Y)$ as

$$\begin{bmatrix} \frac{1}{2}(b_2-b_1)(A_1-A_2)^T e & (b_1+b_2)/2 \\ -\frac{1}{2}(A_1-A_2)(A_1-A_2)^T e & (A_1+A_2)/2 \\ (A_1+A_2)^T/2 & 0 \end{bmatrix}.$$

Comparing to Equation 3, the above matrix represents the implicit linear equation:

$$X(A_1 - A_2) = b_2 - b_1$$

for X along with the conditional knowledge of Y given X. It is trivial to note that the combination is equivalent to solving linear equations $Y = b_1 + XA_1$ and $Y = b_2 + XA_2$ by substitution:

$$b_1 + XA_1 = b_2 + XA_2.$$

When linear equations are expressed implicitly, their combination is equivalent to forming a larger system of linear equations. Assume $XA = U$ and $XB = V$ are two systems of linear equations on a vector of variables X, U, and V, where U and V are distinct vectors of auxiliary variables, and A and B are appropriate coefficient matrices. Their matrix representations are

$$M(\overrightarrow{X}, U) = \begin{bmatrix} 0 & 0 \\ 0 & A \\ A^T & 0 \end{bmatrix},$$

$$M(\overrightarrow{X}, V) = \begin{bmatrix} 0 & 0 \\ 0 & B \\ B^T & 0 \end{bmatrix}.$$

Since both matrices have been swept from common variables X, they can be directly summed according to the new generalized rule of combination [Liu, 2011b]:

$$M(\overrightarrow{X}, U, V) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A & B \\ A^T & 0 & 0 \\ B^T & 0 & 0 \end{bmatrix},$$

which corresponds to

$$X \begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} U & V \end{bmatrix}.$$

In words, the combination of $XA = U$ and $XB = V$ is identical to a system of linear equations joining both $XA = U$ and $XB = V$.

To understand what it really means by combining linear equations, let us perform sweepings on the matrix representation for a system of m equations, $XA = U$, where X is a vector of n variables, and U is a vector of m variables, and A is a $n \times m$ coefficient matrix. First, assume $n \geq m$ and all linear equations are independent, i.e., none is linear combination of others, and thus there is a subvector of X that can be solved in terms of other variables. Without loss of generality, assume $X = (X_1, X_2)$ with X_1 being any subvector of m variables that can be solved and A is split vertically into two submatrices A_1 and A_2 with A_1 being a nonsingular $m \times m$ matrix. Then we have

$$X_1 A_1 + X_2 A_2 = U,$$

which is represented as

$$M(\overrightarrow{X_1}, \overrightarrow{X_2}, U) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & A_1 \\ 0 & 0 & A_2 \\ A_1^T & A_2^T & 0 \end{bmatrix}.$$

Apply a forward sweep to $M(\overrightarrow{X_1}, \overrightarrow{X_2}, U)$ from U:

$$M(\overrightarrow{X_1}, \overrightarrow{X_2}, \overrightarrow{U}) = \begin{bmatrix} 0 & 0 & 0 \\ -eA_1A_1^T & -eA_1A_2^T & eA_1 \\ -eA_2A_1^T & -eA_2A_2^T & eA_2 \\ eA_1^T & eA_2^T & -eI \end{bmatrix}$$

and unsweep $M(\overrightarrow{X_1}, \overrightarrow{X_2}, \overrightarrow{U})$ from X_1. Noting that A_1 is nonsingular and

$$(A_1 A_1^T)^{-1} = (A_1^T)^{-1}(A_1)^{-1},$$

we can easily verify that $M(X_1, \overrightarrow{X_2}, \overrightarrow{U})$ is

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & -(A_1^T)^{-1}A_2^T & (A_1^T)^{-1} \\ -A_2(A_1)^{-1} & 0 & 0 \\ (A_1)^{-1} & 0 & 0 \end{bmatrix},$$

which is the matrix representation of

$$X_1 = -X_2 A_2 (A_1)^{-1} + U(A_1)^{-1}.$$

Therefore, sweeping from U and unsweeping from X_1 is the same as solving for X_1 in terms of U.

Second, assume the system $XA = C$ contains m equations and n variables with $n \leq m$, C being an n dimensional vector, and A has rank n. Using auxiliary variable U, the system is equivalent to the combination of

$$M(\overrightarrow{X},U) = \begin{bmatrix} 0 & 0 \\ 0 & A \\ A^T & 0 \end{bmatrix}$$

with

$$M(U) = \begin{bmatrix} C \\ 0 \end{bmatrix}$$

or via extreme numbers,

$$M(\overrightarrow{X},\overrightarrow{U}) = \begin{bmatrix} 0 & Ce \\ -AA^Te & Ae \\ A^Te & -2Ie \end{bmatrix}.$$

Unsweeping $M(\overrightarrow{X},\overrightarrow{U})$ from the inverse covariance matrix of U, we obtain

$$M(\overrightarrow{X},U) = \begin{bmatrix} \frac{1}{2}CA^Te & C/2 \\ -\frac{1}{2}AA^Te & A/2 \\ A^T/2 & 0 \end{bmatrix}.$$

Since A has rank n, AA^T is positive definite. Thus, we can unsweep $M(\overrightarrow{X},U)$ from the inverse covariance matrix of X and obtain $M(X,U)$ as

$$\begin{bmatrix} CA^T(AA^T)^{-1} & \frac{1}{2}C[I+A^T(AA^T)^{-1}A] \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

implying that, after combination, variable X takes on value

$$X = CA^T(AA^T)^{-1}$$

with certainty. Note that this solution is the least-square estimate of X from regression model $XA = C$ with A being the observation matrix for independent variables and C being the observations for a dependent variable. In addition, the auxiliary variable U takes on the value

$$U = \frac{1}{2}C[I+A^T(AA^T)^{-1}A] \tag{4}$$

with certainty. This seems to be in conflict with initial component model $U = C$. However, one shall realize that, when $m > n$, there exist only n independent linear equations. Thus, only n variables of U can take independent observations, and the remaining $n-m$ variables take on the values as derived from those observations. Otherwise, $U = C$ will have conflicting observations on some or all variables. Equation 4 represents values that are closest to the observations if there is any conflict. In fact, in the special case when $m = n$, we have

$$(AA^T)^{-1} = (A^T)^{-1}A^{-1}.$$

Thus

$$M(X,U) = \begin{bmatrix} CA^{-1} & C \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

implying that $X = CA^{-1}$ and $U = C$ with certainty. This is simply the solution to $XA = C$.

5 Conclusion

In knowledge-based systems, extreme numbers arise whenever a deterministic linear model like a linear equation exists in the knowledge base. A linear model is represented as a marginal or conditional normal distribution. For a linear equation, its conditional variance is zero, and its matrix sweeping from such a zero variance turns the matrix into an extreme one. This paper studied the application of extreme numbers to representing and transforming linear equations and combining them as belief functions via Dempster's rule. When a number of linear equations are under-determined, their combination corresponds to solving the equations for some variables in terms of others. When they are just determined, their combination corresponds to solving the equations for all the variables. When they are over-determined, their combination corresponds to finding the least-square estimate of all the variables.The meaning of the combination in such a case should be studied by future research.

References

[Dempster, 2001] 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)

[Khatri, 1968] Khatri, C.G.: Some results for the singular normal multivariate regression nodels. Sankhya A 30, 267–280 (1968)

[Liu, 2011a] Liu, L.: Dempster's rule for combining linear models. Technical report, Department of Management, The University of Akron, Akron, Ohio (2011a)

[Liu, 2011b] Liu, L.: A new rule for combining linear belief functions. Technical report, Department of Management, The University of Akron, Akron, Ohio (2011b)

[Liu et al., 2006] Liu, L., Shenoy, C., Shenoy, P.P.: Knowledge representation and integration for portfolio evaluation using linear belief functions. IEEE Transactions on Systems, Man, and Cybernetics, Series A 36(4), 774–785 (2006)

[Shafer, 1976] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

Reliability in the Thresholded Dempster-Shafer Algorithm for ESM Data Fusion

Melita Hadzagic, Marie-Odette St-Hilaire, and Pierre Valin

Abstract. The effectiveness of a multi-source information fusion process for decision making highly depends on the quality of information that is received and processed. This paper proposes methods for incorporating reliability, as one of the attributes of the quality of information, into the Thresholded Dempster-Shafer fusion algorithm for Electronic Support Measure (ESM) data fusion and delivers its quantitative assessment by evaluating statistically the performance of the fusion algorithm. The results suggest that accounting for the reliability of information in the fusion algorithm will lead to an improved decision making.

1 Introduction

An information fusion system for decision making involves gathering and fusing a large amount of imperfect heterogeneous information obtained from geographically distributed sources. The effectiveness of a fusion algorithm in use highly depends on the quality of received and processed information, which may be characterized by its attributes such as uncertainty, reliability, relevance, completeness, and others [5].

This paper addresses the reliability as one of the attributes of the quality of information and attempts to assess its impact on an information fusion process. For this purpose, the reliability is considered as a higher order of uncertainty [4].

Melita Hadzagic
Centre de Recherches Mathémathiques, Université de Montréal, 2920 Chemin de la tour, Montréal, QC, H3T 1J4, Canada
e-mail: hadzagic@crm.umontreal.ca

Marie-Odette St-Hilaire
OODA Technologies Inc., 4891 Av. Grosvenor Montreal, QC, H3W 2M2, Canada
e-mail: marie-odette.st-hilaire@ooda.ca

Pierre Valin
C2 Decision Support Systems Section, Defence R&D Canada Valcartier, 2459 Pie-XI Blvd. Nord Québec, QC, G3J 1X5, Canada
e-mail: pierre.valin@drdc-rddc.gc.ca

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 267–274.
springerlink.com

There exist two approaches to define reliability as a higher order of uncertainty. In the first approach, reliability is considered as the relative stability of the first order of uncertainty, i.e. the reliability is measured by the sensor's performance in terms of amount of countermeasures acting on the sensor. In the second approach, the goal is to represent reliability by measuring the accuracy of predicted beliefs. In the latter case, the following situations can be identified, [4]: *(i)* assigning a numerical degree of reliability to each source, *(ii)* a subset of sources is reliable but it is not known which one, and *(iii)* reliabilities of respective sources are known up to an order, however no precise reliability values are known.

Addressing each of these situations requires implementing one or all of the following strategies while incorporating the reliability into the fusion process: *(i)* strategies for identifying reliability of sources and discarding data coming from a source of poor reliability prior to the fusion process, *(ii)* strategies for modifying beliefs by considering their reliability before fusion, *(iii)* strategies for modifying the fusion process to account for reliability of the input.

This paper proposes methods for incorporating reliability into the Thresholded Dempster-Shafer (DS) fusion algorithm [6] for Electronic Support Measure (ESM) data fusion, and delivers its quantitative assessment by evaluating statistically the performance of the fusion algorithm in terms of two measures of the fusion algorithm's performance, the stability and the latency. The results obtained will facilitate evaluating how well the fusion product represents the reality, which will further contribute to improved decision making and situation awareness.

The rest of the paper is organized as follows. Section 2 introduces the reliability and general approaches for its incorporation into the information fusion process, and also presents the proposed methods. In Section 3, the results obtained by numerical simulations are presented, while Section 4 delivers the conclusion remarks.

2 Reliability in a Fusion Process

Following from Section 1, the reliability is considered as the adequacy of belief models with respect to reality where a numerical degree of reliability is assigned to each source, while the strategies for modifying beliefs prior to fusion process are adopted. Generally for this approach, expert-based and/or context-based methods may be used for modifying beliefs. In the expert-based methods, the reliability coefficients are assumed to have fixed values. For references on the context-based methods, see, e.g. [1], [2], and [3].

The information fusion process is considered to be a process of combining successive ESM measurements, i.e. fusing data of a single ESM sensor over time. The fusion algorithm used is the Thresholded-DS algorithm [6]. The approach proposed here is to assign a numerical degree of reliability to each source while applying the strategies of modifying beliefs before fusion. Prior to each combination of beliefs in the Thresholded-DS algorithm, three situations of interest, when assigning a numerical degree of reliability, are distinguished: (1) assuming a fixed value of the

reliability coefficient during the fusion process, (2) assuming a variable value of the reliability coefficient during the fusion process in the form of a step function, and (3) assuming different values of reliability coefficients for different ESM reported allegiances.

With the basic assumptions of the DS theory (e.g. see Chapter 4 in [4]), it is assumed that for each source i the beliefs $m_i(A)$ are not equally reliable, $m_i(A)$ being the belief of the source i that the observed data belongs to a subset $A \in 2^{\Theta}$, from a frame of discernment Θ. For the reliability factors R_i, $i = 1,\ldots,I$, and I number of sources, the combination rule can be written as, [4],

$$m(A) = \sum_{i=1}^{I} R_i m_i(A) \tag{1}$$

Since only a single sensor is considered here, i denotes the index of ESM reports ordered in time, where $i \in [1,\ldots,N]$, N being the total number of the ESM reports.

2.1 Constant Sensor Reliability

For a single sensor ESM data fusion process, i.e. the combining successive ESM measurements (different allegiances) obtained from a single sensor, it is possible to assign to each sensor's declaration a reliability coefficient that is assumed constant for all reports i, $i = 1,\ldots,N$, during the whole duration of the fusion process, i.e.

$$R_i = R \tag{2}$$

The assumption on the existence of countermeasures acting on the ESM sensor is allowed.

2.2 Variable Reliability Coefficient

For a single sensor ESM data fusion process, the value of the reliability coefficient may vary for different intervals of sensor's reporting time, i.e. the reliability coefficient (e.g. for three intervals) can be defined as

$$R = \begin{cases} R_1, & i \in [0, i_1] \\ R_2, & i \in [i_1 + 1, i_2] \\ R_3, & i \in [i_2 + 1, N] \end{cases} \tag{3}$$

where i_1, i_2 and i_3 define the intervals of interest, and N is the total number of the ESM reports (or total number of reporting time instants). The beliefs of each new sensor report are modified using (1) where the reliability coefficient, R_i, has the corresponding interval value. The assumption on the existence of countermeasures acting on the ESM sensor is also allowed.

2.3 Allegiance-Based Reliability

The values of the reliability coefficients may depend on the allegiance, i.e. one can assume that an ESM report *friend* (F) is more reliable than the report *hostile* (H). In this case, for reported allegiances F, *neutral* (N) and H different fixed values of reliability coefficients R_F, R_N, and R_H, respectively, may be used in (1) to update the beliefs of the ESM sensor. The assumption on the existence of countermeasures acting on the ESM sensor is allowed.

3 Numerical Simulations

For the purpose of the algorithm's performance evaluation, the data obtained from a simulated ground truth scenario, were used. The scenario assumes total number of N reports obtained at discrete times t_i, $i = 1,\ldots,N$. For the simplicity of notation, let $t_i = i$. It is assumed that for $i \in [1, i_{sw}]$ the ESM sensor reports are *friend* (F), while for $i \in [i_{sw}+1, N]$, the reported allegiances are *hostile* (H). The default values of the scenario are $N = 100$ and the time of the switch of allegiance $i_{sw} = 50$. Additionally, it is assumed that there exists a percentage of countermeasures acting on the reporting ground-truth, which appear $1 - Acc\%$ of time, where $Acc\%$ represents the percentage of total number of correct reports of allegiance. The number of false reports for the true allegiance F as the result of the countermeasures is equally distributed between the declarations N and H, while for the true allegiance H is equally distributed between the declarations N and F. For the Thresholded-DS algorithm, it is assumed that the basic probabilistic assignment BPA (or mass m) for the ESM sensor has a value m, while the rest $(1-m)$ is assigned to ignorance. The default value for the ESM mass m is $m = 0.7$. The value of the ignorance threshold I_{min}, defined as the value below which the ignorance cannot be lower after a fusion step [6], is assumed to be $I_{min} = 0.0325$ for all MC simulations. Both m and I_{min} values can be user defined. For all numerical simulations, it was assumed that 20% of countermeasures have acted on the ESM sensor.

Stability is defined in terms of error of the fusion product i.e. the error on the allegiance decision of the fusion algorithm at each report time instant [6]. It is defined as the averaged standard deviation of the statistical error, $\bar{\sigma}_i(e)$, over all fusion steps at discrete time instants $i \in I$, $I = [15, i_{sw}-5] \cup [i_{sw}-5, N-15]$, i_{sw} being the time of the allegiance switch, i.e., $\bar{\sigma}_i(e) = \frac{1}{I}\sum_{i\in I}\sigma_i(e)$ where $\sigma_i(e) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(e_i^j - \mu_i(e))^2}$, and $\mu_i(e) = \frac{1}{M}\sum_{j=1}^{M} e_i^j$, and where M is the number of Monte Carlo (MC) simulations, N is the total number of the ESM reports, and e_i^j is the absolute decision error at the fusion step i for the j-th MC run. The choice of the set of fusion steps considered in calculating the stability is justified by the fact that the poor stability occurs (by default) in the beginning due to the initialization process of the algorithm. For the current version of the scenario, it is assumed that the total number of ESM reports $N \geq 50$.

Latency is defined as the number of fusion steps required to detect the true allegiances after the switch of allegiance occurs, i.e. Latency $= \sum_{i\in I, i\geq i_{sw}} L_i$ where

$L_i = \begin{cases} 1, & g_i < \mathrm{T} \\ 0, & \text{otherwise} \end{cases}$ and where $i \in I$ is the current index of the fusion step, I is the set of considered fusion steps, $g_i = 100\mu_i(e)$ represents the so-called *good decision rate* at the fusion step i, while T is a threshold for sufficiently good reaction time performance. The threshold T is calculated as $T = \mu(g_i) - 3\sigma(g_i)$ where the mean good decision rate, $\mu(g_i)$, and its standard deviation, $\sigma(g_i)$, are calculated as $\mu(g_i) = \frac{1}{I}\sum_{i\in I} g_i$ and $\sigma(g_i) = \sqrt{\frac{1}{|I|-1}\sum_{i\in I}(g_i - \mu(g_i))^2}$.

The results are presented next.

Figure 1 shows the output of the Thresholded-DS algorithm for one MC run and the constant reliability (or the tradeoff (*rtoff*)) approach, in which the value of the reliability coefficient, $R = 0.8$ during the total duration of the scenario. Figure 2 shows latency as the number of reports below the threshold T (see the above definition of latency) obtained from 1000 MC runs. The computed $latency = 9$ suggests relatively good performance of the fusion algorithm in presence of a lower sensor reliability, $R = 0.8$. The latency for different values of reliability coefficients in the presence of 20% of countermeasures is shown in Figure 3. It increases linearly as the values of reliability coefficients decrease, while the stability, shown in Figure 4 decreases for decreasing values of the reliability coefficients, hence both performance measures indicating the influence of the reliability on the fusion process.

With the same assumptions on the countermeasures, the BPAs obtained for the variable reliability approach, referred to as *r123*, for the specified intervals, are shown in Figure 5. It illustrates slightly poorer performance with the *r123* approach comparing to the *rtoff* one, which is expected since in *r123* the values of reliability coefficients decrease during the fusion process. The latency, see Figure 6, was not affected in the proximity of the switch. Using the same reliability approach, a situation of a low sensor reliability in the proximity of the switch was investigated. In the vicinity of the switch, $i_{sw} = 51$, between the fusion step $i = 31$ and $i = 61$ the reliability factor value was set to drop to $R_2 = 0.5$. It can be observed in Figure 7 that

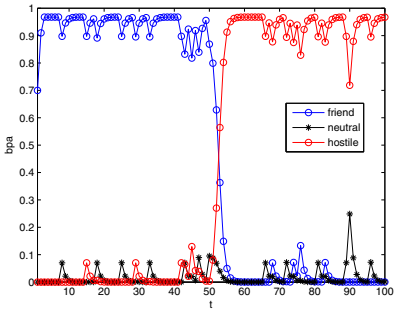


Fig. 1 BPA for the tradeoff (*rtoff*) approach, $R = 0.8$, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for one MC run. Countermeasures = 20%.

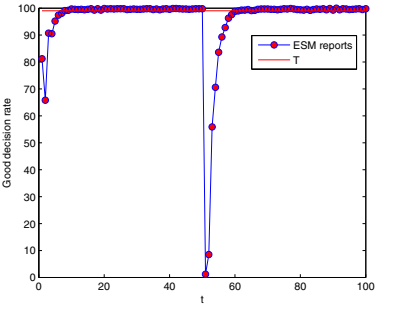


Fig. 2 Good decision rate (gdr) for the tradeoff (*rtoff*) approach, $R = 0.8$, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs. Countermeasures = 20%.

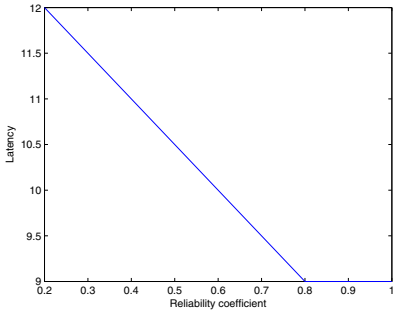


Fig. 3 Latency versus reliability for the tradeoff (*rtoff*) approach, $R = 0.8$, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs. Countermeasures = 20%.

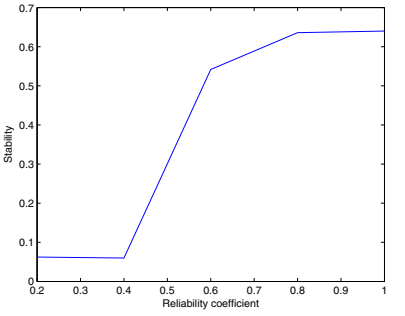


Fig. 4 Stability versus reliability for the tradeoff (*rtoff*) approach, $R = 0.8$, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs. Countermeasures = 20%.

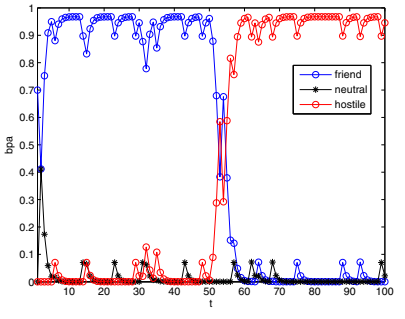


Fig. 5 BPA for the variable reliability (*r123*) approach, $R_1 = 0.8$, $R_2 = 0.7$ and $R_3 = 0.6$ for $i \in [0,30]$, $i \in [31,60]$, and $i \in [61,100]$, respectively, $I_{min} = 0.0325$, ESM mass $m = 0.7$ for one MC run. Countermeasures = 20%.

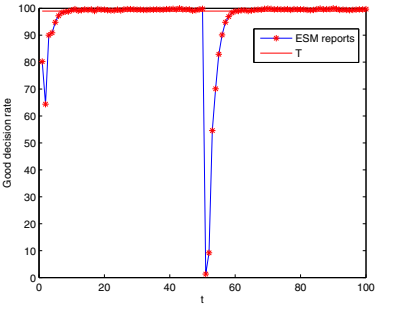


Fig. 6 Good decision rate (gdr) for the variable reliability (*r123*) approach, $R_1 = 0.8$, $R_2 = 0.7$ and $R_3 = 0.6$ for $i \in [0,30]$, $i \in [31,60]$, and $i \in [61,100]$, respectively, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs. Countermeasures = 20%.

the performance of detection of a correct allegiance decreases near the switch as the reliability decreases its value. In this case, $latency_{R=0.5} = 12 > latency_{R=0.8} = 9$, i.e. the decision on a switch of allegiance was delayed.

The allegiance-based approach (*rfhn*) allows for setting a larger value of reliability coefficient to the declarations F and N than to H. For the specified values, the results depicted in Figure 8 illustrate that the decision of a correct allegiance occurs but is delayed. This is confirmed by the calculated average latency value for 1000MC, $latency = 12$, see Figure 10. However, for another MC run an error in making a correct decision may also happen in the vicinity of the switch, as observed in Figure 9. Finally, Figure 11 shows the standard deviations for all three approaches suggesting on average (1000 MC runs) similar algorithm's stability

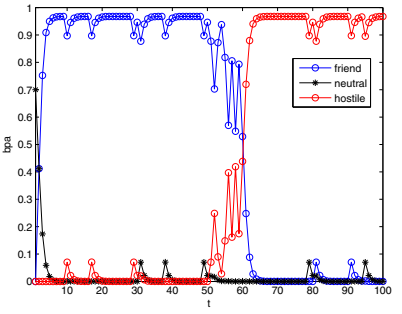


Fig. 7 BPA for the variable reliability (*r123*) approach, $R_1 = 0.8$, $R_2 = 0.5$ and $R_3 = 0.8$ for $i \in [0,30]$, $i \in [31,60]$, and $i \in [61,100]$, respectively, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for one MC run. Countermeasures = 20%.

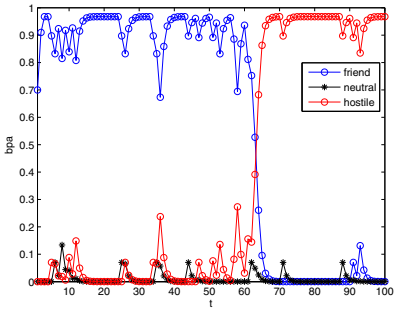


Fig. 8 BPA for the different allegiance approach ($R_F = R_N = 0.9$, $R_H = 0.5$), $I_{min} = 0.0325$, ESM mass $m = 0.7$ for one MC run. Countermeasures = 20%.

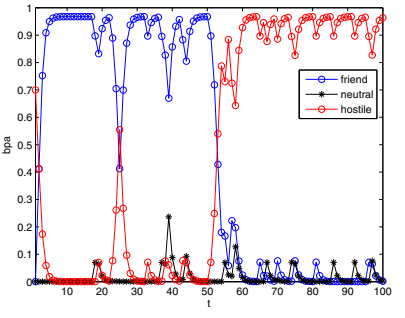


Fig. 9 BPA for different allegiance approach (*rfhn*) ($R_F = R_N = 0.9$, $R_H = 0.5$) for one MC run. Countermeasures = 20%.

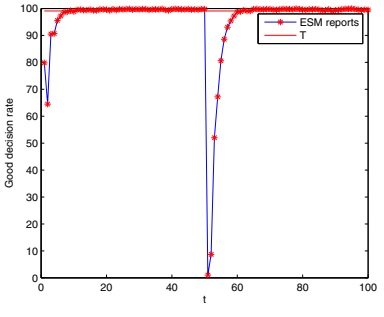


Fig. 10 Good decision rate (gdr) for the different allegiance (*rfhn*) approach, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs.

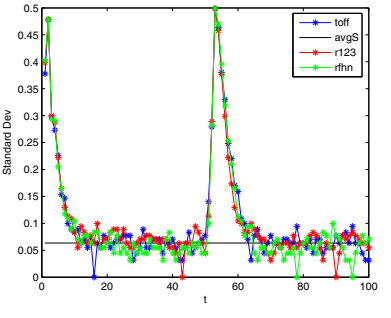


Fig. 11 Standard deviation for the tradeoff (*toff*), the variable reliability (*r123*), and the allegiance-based (*rfhn*) reliability approach, $I_{min} = 0.0325$, ESM mass $m = 0.7$, for 1000 MC runs.

performance for the tradeoff (*toff*), the variable (*r123*), and the allegiance-based (*rfhn*) reliability. The horizontal line (labeled as *AvgS*) represents the average stability.

4 Conclusions

The tradeoff approach shows on average (for 1000 MC runs) good fusion performance for a lower but still large ($R = 0.8$) reliability. However, the fusion performance decreases as the value of the reliability coefficient decreases. Similarly, the variable reliability approach suggests that low reliability highly affects the performance of the algorithm, especially in terms of detection of switch of allegiance. The allegiance-based approach also shows that the reliability significantly affects the Thresholded-DS algorithm's performance in terms of correct decision on allegiance. In conclusion, the study suggests that accounting for the reliability in the fusion algorithm will lead to an improved decision making.

References

1. Bennett, P., Dumais, S., Horwitz, E.: Probabilistic combination of text classifiers using reliability indicators: Models and result. In: Proceedings of SIGIR 2002, p. 1115 (2002)
2. Fabre, S., Appriou, A., Briottet, X.: Presentation and description of two classification methods using data fusion based on sensor management. Information Fusion 2, 47–71 (2001)
3. Nimier, V.: Supervised multisensor tracking algorithm by context analysis. In: Proceedings of the Intl. Conference on Information Fusion, pp. 149–156 (1998)
4. Rogova, G., Bossé, É.: Reliability and information quality assessment for information fusion. Tech. Rep. DRDC Valcartier, TR-2005-270, DRDCV (2008)
5. Rogova, G., Bosse, E.: Information quality in information fusion. In: Proceedings of the 13th Intl. Conference on Information Fusion, pp. 1–8 (2010)
6. Valin, P., Djiknavorian, P., Bossé, É.: A pragmatic approach for the use of Dempster-Shafer theory in fusing realistic sensor data. Journal of Advances in Information Fusion (JAIF) 5(1), 32–40 (2010)

Hierarchical Proportional Redistribution for bba Approximation

Jean Dezert, Deqiang Han, Zhunga Liu, and Jean-Marc Tacnet

Abstract. Dempster's rule of combination is commonly used in the field of information fusion when dealing with belief functions. However, it generally requires a high computational cost. To reduce it, a basic belief assignment (bba) approximation is needed. In this paper we present a new bba approximation approach called hierarchical proportional redistribution (HPR) allowing to approximate a bba at any given level of non-specificity. Two examples are given to show how our new HPR works.

1 Introduction

Dempster-Shafer Theory (DST), also called Theory of Evidence [10], has been widely used in many applications, e.g., information fusion, pattern recognition and decision making [11]. Although it is appealing in uncertainty modeling, while appearing more controversial for consistent reasoning, the high computational cost remains problematic which is often raised against its use [11]. To resolve such a problem, three major types of approaches have been proposed.

The first is to propose efficient procedures for performing exact computations [1, 8]. The second is composed of Monte-Carlo techniques [9]. The third is to

Jean Dezert
The French Aerospace Lab, F-91761 Palaiseau, France
e-mail: jean.dezert@onera.fr

Deqiang Han
Inst. of Integrated Automation, Xi'an Jiaotong University, Xi'an, 710049, China
e-mail: deqhan@gmail.com

Zhunga Liu
School of Automation, North-Western Polytechnical University, Xi'an, 710072, China
e-mail: liuzhunga@gmail.com

Jean-Marc Tacnet
Irstea, UR ETGR, 2 rue de la papeterie-B.P. 76, F-38402 St-Martin-d'Hères, France
e-mail: jean-marc.tacnet@irstea.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 275–283.
springerlink.com

approximate a belief function to a simpler one. The papers of Voorbraak [13], Dubois and Prade [5] are seminal works of this type. Other representative works include $k-l-x$ [3] and k-additive belief function [2, 6]. Denœux uses hierarchical clustering to implement the inner and outer approximation [3].

In this paper, we propose a new method called hierarchical proportional redistribution (HPR) to approximate any general basic belief assignment (bba) at a given level of non-specificity [4], up to the ultimate level 1 corresponding to a Bayesian bba [10]. The level of non-specificity can be controlled by the users through the adjustment of the maximum cardinality of remaining focal elements. For the approximated bba obtained by HPR, the maximum cardinality of the focal elements is k. Thus HPR can be considered as a generalized k-additive belief approximation. Some examples are given to show how our proposed HPR method works, and to compare it with other approximations.

2 Basics of Dempster-Shafer Theory (DST)

In DST [10], the frame of discernment (FoD) is a set Θ of mutual exhaustive and exclusive elements. $m(.): 2^{\Theta} \rightarrow [0,1]$ is a basic belief assignment (bba), also called mass function, if it satisfies

$$\sum_{A \subseteq \Theta} m(A) = 1, \; m(\emptyset) = 0. \tag{1}$$

Belief function (Bel) and plausibility function (Pl) are defined as

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad \text{and} \quad Pl(A) = \sum_{A \cap B \neq \emptyset} m(B). \tag{2}$$

Suppose that $m_1, m_2, ..., m_n$ are n bba's, then Dempster's rule of combination is defined by

$$m(A) = \begin{cases} 0, & A = \emptyset \\ \dfrac{\sum\limits_{\cap A_i = A} \prod\limits_{1 \leq i \leq n} m_i(A_i)}{\sum\limits_{\cap A_i \neq \emptyset} \prod\limits_{1 \leq i \leq n} m_i(A_i)}, & A \neq \emptyset \end{cases} \tag{3}$$

This rule is used in DST to combine pieces of evidence expressed by bba's. As referred above, Dempster's combination has high computational cost and three types of approaches have been proposed to reduce it. We prefer belief approximation approaches [2, 3, 6, 12] since they both reduce the computational cost of the combination and allow to deal with smaller-size focal elements, which is more intuitive for human to catch the meaning and interpret fusion results [2].

3 Two bba Approximation Approaches

1) $k-l-x$ **approximation:** This was proposed by Tessem [12]. The simplified bba obtained by $k-l-x$ approach satisfies: a) keep no less than k focal elements; b) keep no more than l focal elements; c) the mass assignment to be deleted is no greater than x. In $k-l-x$, the focal elements of a original bba are sorted by their masses. Such an algorithm chooses the first p focal elements such that $k \leq p \leq l$ and such

that the sum of the masses of these first p focal elements is no less than $1-x$. The deleted masses are redistributed to the other focal elements through a normalization.

2) k-additive belief function approximation: Given $m(.): 2^{\Theta} \rightarrow [0,1]$, one kind of k-additive belief function [2, 6] induced by the mass $m(.)$ is defined by

$$\begin{cases} m_k(B) = m(B) + \sum\limits_{A \supset B, A \subseteq \Theta, |A|>k} \frac{m(A)\cdot|B|}{\mathscr{N}(|A|,k)}, & \forall\, |B| \leq k \\ m_k(B) = 0, & \forall\, |B| > k \end{cases} \tag{4}$$

where $B \subseteq \Theta$ and

$$\mathscr{N}(|A|,k) = \sum_{j=1}^{k} \binom{|A|}{j} \cdot j = \sum_{j=1}^{k} \frac{|A|!}{(j-1)!(|A|-j)!} \tag{5}$$

is the average cardinality of the subsets of A of size at most k. For k-additive belief approximation, the maximum cardinality of available focal elements is no greater than k. Other bba approximation methods can be found in related references.

4 Hierarchical Proportional Redistribution Approximation

In this paper we propose a new bba approximation approach called hierarchical proportional redistribution (HPR), which provides a new way to reduce step-by-step the mass committed to uncertainties. Ultimately an approximate measure of subjective probability can be obtained if needed, i.e. a so-called Bayesian bba in [10]. Our proposed procedure can be stopped at any step in the process and thus it allows to reduce the number of focal elements of a given bba in a simple manner to diminish the size of the core [10] of a bba. Thus we can reduce the complexity (if needed) when applying also some complex rules of combinations. By using HPR, we can obtain approximate bba's at any different non-specificity level that we want. Let us first introduce two new notations for convenience and conciseness:

1. Any element of cardinality $1 \leq k \leq n$ of the power set 2^{Θ} will be denoted $X(k)$ by convention. For example, if $\Theta = \{A,B,C\}$, then $X(2)$ can denote the following partial uncertainties $A \cup B$, $A \cup C$ or $B \cup C$, and $X(3)$ denotes the total uncertainty $A \cup B \cup C$.
2. The proportional redistribution factor (ratio) of width s involving elements X and Y of the powerset is defined by (for $X \neq \emptyset$ and $Y \neq \emptyset$)

$$R_s(Y,X) \triangleq \frac{m(Y) + \varepsilon \cdot |X|}{\sum\limits_{\substack{Y \subset X \\ |X|-|Y|=s}} m(Y) + \varepsilon \cdot |X|} \tag{6}$$

where ε is a small positive number introduced here to deal with particular cases where $\sum\limits_{\substack{Y \subset X \\ |X|-|Y|=s}} m(Y) = 0$.

By convention, we will denote $R(Y,X) \triangleq R_1(Y,X)$ when we use the proportional redistribution factors of width $s=1$, as we use in this paper for this HPR method.

The HPR is a step-by-step (recursive) proportional redistribution of the mass $m(X(k))$ of a given uncertainty $X(k)$ (partial or total) of cardinality $2 \leq k \leq n$ to all the least specific elements of cardinality $k-1$, i.e., to all possible $X(k-1)$, until $k=2$ is reached. The proportional redistribution is done from the masses of belief committed to $X(k-1)$ as done classically in DSmP transformation. The "hierarchical" masses $m_h(.)$ are recursively (backward) computed as follows. Here $m_{h(k)}$ represents the approximate bba obtained at the step $n-k$ of HPR, i.e., it has the maximum focal element cardinality of k.

$$\begin{aligned} &m_{h(n-1)}(X(n-1)) = m(X(n-1)) + \sum_{\substack{X(n)\supset X(n-1),\\ X(n),X(n-1)\in 2^\Theta}} [m(X(n))\cdot R(X(n-1),X(n))]; \\ &m_{h(n-1)}(A) = m(A), \forall |A| < n-1 \end{aligned} \tag{7}$$

$m_{h(n-1)}(\cdot)$ is the bba obtained at the first step of HPR ($n-(n-1)=1$), the maximum focal element cardinality of $m_{h(n-1)}$ is $n-1$.

$$\begin{aligned} m_{h(n-2)}(X(n-2)) &= m(X(n-2)) \\ &\quad + \sum_{\substack{X(n-1)\supset X(n-2)\\ X(n-2),X(n-1)\in 2^\Theta}} [m_{h(n-1)}(X(n-1))\cdot R(X(n-2),X(n-1))] \\ m_{h(n-2)}(A) &= m_{h(n-1)}(A), \forall |A| < n-2 \end{aligned} \tag{8}$$

$m_{h(n-2)}(\cdot)$ is the bba obtained at the second step of HPR ($n-(n-2)=2$), the maximum focal element cardinality of $m_{h(n-2)}$ is $n-2$.

This hierarchical proportional redistribution process can be applied similarly (if one wants) to compute $m_{h(n-3)}(.)$, $m_{h(n-4)}(.)$, ..., $m_{h(2)}(\cdot)$, $m_{h(1)}(\cdot)$ with

$$\begin{aligned} &m_{h(2)}(X(2)) = m(X(2)) + \sum_{\substack{X(3)\supset X(2)\\ X(3),X(2)\in 2^\Theta}} [m_{h(3)}(X(3))\cdot R(X(2),X(3))] \\ &m_{h(2)}(A) = m_{h(3)}(A), \forall |A| < n-2 \end{aligned} \tag{9}$$

$m_{h(2)}(\cdot)$ is the bba obtained at the first step of HPR ($n-2$), the maximum focal element cardinality of $m_{h(2)}$ is 2.

Mathematically, for any $X(1)\in\Theta$, i.e. any $\theta_i \in \Theta$ a Bayesian belief function can be obtained by HPR method in deriving all possible steps of proportional redistributions of partial ignorances in order to get

$$m_{h(1)}(X(1)) = m(X(1)) + \sum_{\substack{X(2)\supset X(1)\\ X(1),X(2)\in 2^\Theta}} [m_{h(2)}(X(2))\cdot R(X(1),X(2))] \tag{10}$$

In fact, $m_{h(1)}(\cdot)$ is a probability transformation, called here the Hierarchical DSmP (HDSmP). Since $X(n)$ is unique and corresponds only to the full ignorance $\theta_1\cup\theta_2\cup\ldots\cup\theta_n$, the expression of $m_h(X(n-1))$ in Eq.(9) just simplifies as

$$m_{h(n-1)}(X(n-1)) = m_h(X(n-1)) + m(X(n))\cdot R(X(n-1),X(n)) \tag{11}$$

For the full proportional redistribution of the masses of uncertainties to the elements least specific involved in these uncertainties, no mass is lost during the step-by-step hierarchical process and thus at any step of HPR, the sum of masses is kept to one.

5 Examples

5.1 Example 1

Let's consider the following bba over $\Theta = \{\theta_1, \theta_2, \theta_3\}$:

$$m(\theta_1) = 0.10, \quad m(\theta_2) = 0.17, \quad m(\theta_3) = 0.03, \quad m(\theta_1 \cup \theta_2) = 0.15,$$
$$m(\theta_1 \cup \theta_3) = 0.20, \quad m(\theta_2 \cup \theta_3) = 0.05, \quad m(\theta_1 \cup \theta_2 \cup \theta_3) = 0.30.$$

We apply the HPR with $\varepsilon = 0$ in this example because there is no mass of belief equal to zero. It can be verified that the result obtained with small positive ε parameter remains (as expected) numerically very close to what is obtained with $\varepsilon = 0$.

• **Step 1:** The first step of HPR consists in redistributing back $m(\theta_1 \cup \theta_2 \cup \theta_3) = 0.30$ committed to the full ignorance to the elements $\theta_1 \cup \theta_2$, $\theta_1 \cup \theta_3$ and $\theta_2 \cup \theta_3$ only, because these elements are the only elements of cardinality 2 that are included in $\theta_1 \cup \theta_2 \cup \theta_3$. Applying the Eq. (8) with $n = 3$, one gets when $X(2) = \theta_1 \cup \theta_2$, $\theta_1 \cup \theta_3$ and $\theta_1 \cup \theta_2$ the following masses.

$$m_{h(2)}(\theta_1 \cup \theta_2) = m(\theta_1 \cup \theta_2) + m(X(3)) \cdot R(\theta_1 \cup \theta_2, X(3)) = 0.15 + (0.30 \cdot 0.375) = 0.2625$$

because $R(\theta_1 \cup \theta_2, X(3)) = \frac{0.15}{0.15+0.20+0.05} = 0.375$.
Similarly, one gets

$$m_{h(2)}(\theta_1 \cup \theta_3) = m(\theta_1 \cup \theta_3) + m(X(3)) \cdot R(\theta_1 \cup \theta_3, X(3)) = 0.20 + (0.30 \cdot 0.5) = 0.35$$

because $R(\theta_1 \cup \theta_3, X(3)) = \frac{0.20}{0.15+0.20+0.05} = 0.5$, and also

$$m_{h(2)}(\theta_2 \cup \theta_3) = m(\theta_2 \cup \theta_3) + m(X(3)) \cdot R(\theta_2 \cup \theta_3, X(3)) = 0.05 + (0.30 \cdot 0.125) = 0.0875$$

because $R(\theta_2 \cup \theta_3, X(3)) = \frac{0.05}{0.15+0.20+0.05} = 0.125$.

• **Step 2** Now, we go to the next step of HPR principle and one needs to redistribute the masses of partial ignorances $X(2)$ corresponding to $\theta_1 \cup \theta_2$, $\theta_1 \cup \theta_3$ and $\theta_2 \cup \theta_3$ back to the singleton elements $X(1)$ corresponding to θ_1, θ_2 and θ_3. We use Eq. (10) for doing this as follows:

$$\begin{aligned} m_{h(1)}(\theta_1) &= m(\theta_1) + m_h(\theta_1 \cup \theta_2) \cdot R(\theta_1, \theta_1 \cup \theta_2) \quad + m_h(\theta_1 \cup \theta_3) \cdot R(\theta_1, \theta_1 \cup \theta_3) \\ &\approx 0.10 + (0.2625 \cdot 0.3703) + (0.35 \cdot 0.7692) = 0.10 + 0.0972 + 0.2692 = 0.4664 \end{aligned}$$

because $R(\theta_1, \theta_1 \cup \theta_2) = \frac{0.10}{0.10+0.17} \approx 0.3703$ and $R(\theta_1, \theta_1 \cup \theta_3) = \frac{0.10}{0.10+0.03} \approx 0.7692$
Similarly, one gets

$$\begin{aligned} m_{h(1)}(\theta_2) &= m(\theta_2) + m_h(\theta_1 \cup \theta_2) \cdot R(\theta_2, \theta_1 \cup \theta_2) + m_h(\theta_2 \cup \theta_3) \cdot R(\theta_2, \theta_2 \cup \theta_3) \\ &\approx 0.10 + (0.2625 \cdot 0.6297) + (0.0875 \cdot 0.85) = 0.17 + 0.1653 + 0.0744 = 0.4097 \end{aligned}$$

because $R(\theta_2, \theta_1 \cup \theta_2) = \frac{0.17}{0.10+0.17} \approx 0.6297$ and $R(\theta_2, \theta_2 \cup \theta_3) = \frac{0.17}{0.17+0.03} = 0.85$. and also

$$\begin{aligned} m_{h(1)}(\theta_3) &= m(\theta_3) + m_h(\theta_1 \cup \theta_3) \cdot R(\theta_3, \theta_1 \cup \theta_3) + m_h(\theta_2 \cup \theta_3) \cdot R(\theta_3, \theta_2 \cup \theta_3) \\ &\approx 0.03 + (0.35 \cdot 0.2307) + (0.0875 \cdot 0.15) = 0.03 + 0.0808 + 0.0131 = 0.1239 \end{aligned}$$

because $R(\theta_3, \theta_1 \cup \theta_3) = \frac{0.03}{0.10+0.03} \approx 0.2307$ and $R(\theta_3, \theta_2 \cup \theta_3) = \frac{0.03}{0.17+0.03} = 0.15$
Hence, the result of final step of HPR is:

$$m_{h(1)}(\theta_1) = 0.4664, \quad m_{h(1)}(\theta_2) = 0.4097, \quad m_{h(1)}(\theta_3) = 0.1239.$$

We can easily verify that $m_{h(1)}(\theta_1) + m_{h(1)}(\theta_2) + m_{h(1)}(\theta_3) = 1$.

To compare HPR with the approach of $k-l-x$, we set the parameters of $k-l-x$ to obtain bba's with equal focal element number with HPR at each step. In Example 1, for HPR at first step, it can obtain a bba with 6 focal elements. Thus we set $k = l = 6, x = 0.4$ for $k-l-x$ to obtain a bba with 6 focal elements. Similarly, for HPR at second step, it can obtain a bba with 3 focal elements. Thus we set $k = l = 3, x = 0.4$ for $k-l-x$. Based on HPR and $k-l-x$, the results are shown in Table 1.

Table 1 Experimental results of Example 1.

Focal elements	$m_{h(k)}(\cdot)$ - approximate bba			$m(\cdot)$ obtained by $k-l-x$	
	$k=3$	$k=2$	$k=1$	$k=l=6$	$k=l=3$
θ_1	0.1000	0.1000	0.4664	0.1031	0.0000
θ_2	0.1700	0.1700	0.4097	0.1753	0.2573
θ_3	0.0300	0.0300	0.1239	0.0000	0.0000
$\theta_1 \cup \theta_2$	0.1500	0.2625	0.0000	0.1546	0.0000
$\theta_1 \cup \theta_3$	0.2000	0.3500	0.0000	0.2062	0.2985
$\theta_2 \cup \theta_3$	0.0500	0.0875	0.0000	0.0515	0.0000
$\theta_1 \cup \theta_2 \cup \theta_3$	0.3000	0.0000	0.0000	0.3093	0.4478

5.2 Example 2

Let's consider $\Theta = \{\theta_1, \theta_2, \theta_3\}$, and the bba $m(\theta_3) = 0.7$ and $m(\theta_1 \cup \theta_2 \cup \theta_3) = 0.30$. Here, the masses of all the focal elements with cardinality size 2 equal to zero. For HPR, when $\varepsilon > 0$, $m(\theta_1 \cup \theta_2 \cup \theta_3)$ will be divided equally and redistributed to $\{\theta_1 \cup \theta_2\}$, $\{\theta_1 \cup \theta_3\}$ and $\{\theta_2 \cup \theta_3\}$. Because the ratios are (taking for example $\varepsilon = 0.001$)

$$R(\theta_1 \cup \theta_2, X(3)) = R(\theta_1 \cup \theta_3, X(3)) = R(\theta_2 \cup \theta_3, X(3)) = \frac{0.00 + 0.001 \cdot 3}{(0.00 + 0.001 \cdot 3) \cdot 3} = 0.3333$$

In this case, HPR cannot work directly when $\varepsilon = 0$. This shows the necessity for the use of $\varepsilon > 0$. The bba's obtained from $\text{HPR}_{\varepsilon=0.001}$ and $k-l-x$ are listed in Table 2.

From the results of Examples 1 & 2, we can see that based on $k-l-x$, the users can control the number of focal elements but cannot control the maximum cardinality of focal elements. Although based on $k-l-x$, the number of focal elements can be reduced, the focal elements with big cardinality might also be kept. This is not good for further reducing computational cost. But with the proposed HPR method, users can easily control both the non-specificity of approximated bba's and the focal element's size.

Table 2 Experimental results of Example 2 ($\varepsilon = 0.001$)

Focal elements	$m_{h(k)}(\cdot)$ - approximate bba			$m(\cdot)$ obtained by $k-l-x$	
	$k=3$	$k=2$	$k=1$	$k=l=6$	$k=l=3$
θ_1	0.0000	0.0000	0.0503	0.0000	0.0000
θ_2	0.0000	0.0000	0.0503	0.0000	0.0000
θ_3	0.7000	0.7000	0.8994	0.7000	0.7000
$\theta_1 \cup \theta_2$	0.0000	0.1000	0.0000	0.0000	0.0000
$\theta_1 \cup \theta_3$	0.0000	0.1000	0.0000	0.0000	0.0000
$\theta_2 \cup \theta_3$	0.0000	0.1000	0.0000	0.0000	0.0000
$\theta_1 \cup \theta_2 \cup \theta_3$	0.3000	0.0000	0.0000	0.3000	0.3000

5.3 *Example 3*

In this work, an approximation method 1 (giving $m_1(.)$) is considered better than a method 2 (giving $m_2(.)$) if both conditions are fulfilled: 1) if the distance between $m_1(.)$ and original bba $m(.)$ is smaller than the distance between $m_2(.)$ and original bba $m(.)$, i.e. $d(m_1,m) < d(m_2,m)$; 2) if the approximate non-specificity value $U(m_1)$ is closer (and lower) to the true non-specificity value $U(m)$ than $U(m_2)$. We have used Jousselme's distance [7] which has been proved recently to be a strict distance metric because it is commonly used in applications. The Non-specificity [4] is given by $U(m) = \sum_{A \subseteq \Theta} m(A) \log_2 |A|$. In this example, we make a comparison between HPR (method 1) and k-additive approach (method 2). We have taken $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ and generated randomly 30 bba's using the algorithm given in [7]. We compute and plot $d(m_1,m)$, $d(m_2,m)$, $U(m)$, $U(m_1)$ and $U(m_2)$ for several

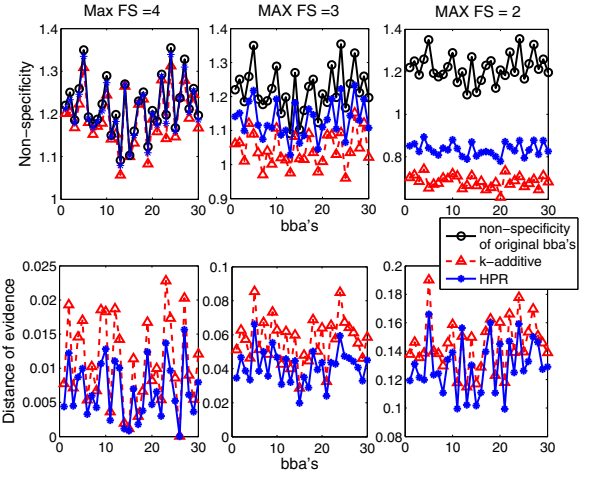


Fig. 1 Results for the Example 3. Comparison of k-additive belief function approximation with HPR approximation method. (FS means Focal element Size)

levels of approximation. The results are shown in Fig. 1 and indicate clearly the superiority of HPR over the *k*-additive approach.

6 Conclusions

In this paper, a novel bba approximation called HPR has been proposed as an interesting alternative approach to two classical ones. With this HPR, the non-specificity degree can be easily controlled by the users. Our example show its behavior and advantage in comparisons with other well-known bba approximation approaches. HPR has a low computational cost compared with *k*-additive approach, which will be discussed in a more detailed paper in future. In further works, we will also compare our proposed HPR with more bba approximation approaches available in the literature. In this paper, we have used only the distance of evidence and the non-specificity criteria, which in fact are not enough, or comprehensive to evaluate efficiently bba approximations. So in future, we will try to propose more efficient evaluation criteria to evaluate and design better bba approximations (if possible).

Acknowledgements. This work was supported by National Natural Science Foundation of China (No.61104214), Fundamental Research Funds for the Central Universities, China Postdoctoral Science Foundation (No.20100481337, No.201104670).

References

1. Barnett, J.A.: Computational methods for a mathematical theory of evidence. In: Proceedings of IJCAI 1981, Vancouver, pp. 868–875 (1981)
2. Burger, T., Cuzzolin, F.: The barycenters of the k-additive dominating belief functions and the pignistic k-additive belief functions. In: Workshop on Theory of Belief Functions, Brest, France, pp. 1–6 (2010)
3. Denœux, T.: Inner and outer approximation of belief structures using a hierarchical clustering approach. International Journal of Uncertainty, Fuzziness, and Knowledge-based Systems 9, 437–460 (2001)
4. Dubois, D., Prade, H.: A note on measures of specificity for fuzzy sets. International Journal of General Systems 10, 279–283 (1985)
5. Dubois, D., Prade, H.: An alternative approach to the handling of subnormal possiblity distributions. Fuzzy Sets and Systems 24, 123–126 (1987)
6. Grabisch, M.: Upper approximation of non-additive measures by *k*-additive measures - the case of belief functions. In: Proc. of the 1st Int. Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium (1999)
7. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53, 118–145 (2011)
8. Kennes, R.: Computational aspects of the M*ö*bius transform of graphs. IEEE Transactions on SMC 22, 201–223 (1992)
9. Moral, S., Salmerón, A.: A Monte Carlo Algorithm for Combining Dempster-Shafer Belief Based on Approximate Pre-computation. In: Hunter, A., Parsons, S. (eds.) ECSQARU 1999. LNCS (LNAI), vol. 1638, pp. 305–315. Springer, Heidelberg (1999)

10. Shafer, G.: A Mathematical Theory of Evidence. Princeton University, Princeton (1976)
11. Smets, P.: Practical uses of belief functions. In: Lskey, K.B., Prade, H. (eds.) Uncertainty in Artificial Intelligence 15 (UAI 1999), Stockholm, Sweden, pp. 612–621 (1999)
12. Tessem, B.: Approximations for efficient computation in the theory of evidence. Artificial Intelligence 61, 315–329 (1993)
13. Voorbraak, F.: A computationally efficient approximation of Dempster-Shafer theory. Int. J. Man-Machine Studies 30, 525–536 (1989)

On the α-Conjunctions for Combining Belief Functions

Frédéric Pichon

Abstract. The α-conjunctions basically represent the set of associative, commutative and linear operators for belief functions with the vacuous belief function as neutral element. Besides, they include as particular case the unnormalized Dempster's rule. They are thus particularly interesting from a formal standpoint. However, they suffer from a main limitation: they lack a clear interpretation in general. In this paper, an interpretation for these combination rules is proposed, based on a new framework that allows the integration of meta-knowledge on the various forms of lack of truthfulness of the information sources.

1 Introduction

The theory of belief functions [1, 5, 10] is a general framework for reasoning under uncertainty. Within this framework, many belief function combination rules have been proposed for the fusion of information and, in particular, the unnormalized version of Dempster's rule [1, 5] (also called conjunctive rule) introduced by Smets in the Transferable Belief Model [6], the disjunctive rule [2, 7], the exclusive disjunctive rule and its negation [2, 8].

In [8], Smets studied these four combination rules and discovered that they are actually special cases of an infinite family of rules, which he called α-junctions and that basically represent the set of associative, commutative and linear operators for belief functions with a neutral element. He further showed that there are only two possibilities for the neutral element, leading to two subfamilies of rules called α-conjunctions and α-disjunctions; the α-conjunctions being the family that has the so-called vacuous belief function as neutral element and the conjunctive rule and the negation of the exclusive disjunctive rule as particular cases. However, except for the four particular cases, he did not provide an interpretation for these rules.

Frédéric Pichon
Thales Research and Technology, Campus Polytechnique, 1 Avenue Augustin Fresnel, 91767 Palaiseau Cedex, France
e-mail: Frederic.Pichon@thalesgroup.com

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 285–292.
springerlink.com

The first effort to find an interpretation for this important family of rules was reported in [3], where we showed that the α-junctions correspond to a particular form of meta-knowledge on the truthfulness of the sources. However, the actual meaning of the α-junctions as presented in [3] remains hard to grasp.

In this paper, the investigation on these rules is pursued. An interpretation for the α-conjunctions is proposed in Section 3, based on a new framework that allows the integration of meta-knowledge on the various forms of lack of truthfulness of the information sources. This framework is introduced in the next section. Let us note that an interpretation for the α-disjunctions can also be provided using this framework. However, due to lack of space, only the case of the α-conjunctions is addressed in this paper.

2 Truthfulness

In this section, a general approach to information fusion for belief functions is proposed, where the various of forms of lack of truthfulness of the sources may be taken into account. We formalize first the notion of truthfulness, before unveiling an associated general combination rule.

2.1 Truthfulness of a Single Source

Let $\mathbf{x}$ be a parameter defined on a domain X. Let us suppose that a source S, such as a sensor or a human agent, provides a piece of information on the value taken by $\mathbf{x}$ and that this source is relevant[1], which means that it provides useful information regarding the value of $\mathbf{x}$ [4]. Let us further assume that the information provided by S takes the form $\mathbf{x} \in A$, for some $A \subseteq X$. In [4], the notion of source truthfulness is investigated and the definition that is used is the following: a source is truthful if it actually supplies the information it possesses and a non truthful source is a source that declares the contrary of what it knows. According to this definition, one must conclude that $\mathbf{x} \in A$ or $\mathbf{x} \in \overline{A}$, where $\overline{A}$ is the complement of A, depending on whether the source S is assumed to be truthful or not.

This definition correspond to the crudest description of the lack of truthfulness. Various other forms of lack of truthfulness exist: besides telling the contrary of what it knows, a source may just say less, or something different, even if consistent with its knowledge, as already remarked in [4]. We propose in the following a refined model of source truthfulness, which allows us to take into account these various forms of lack of truthfulness.

Let $\mathscr{T}_x = \{t_x, \neg t_x\}$ be the frame of the variable $\mathbf{t}_x$ used to model the truthfulness (t_x) or non truthfulness ($\neg t_x$) of a source S with respect to $x \in X$, i.e., the assumptions that the source tells what it knows or the opposite of what it knows with respect to the value $x \in X$. There are thus four possible cases:

[1] In this paper, all information sources will be assumed relevant and thus we will hereafter omit to state this assumption, for clarity of exposition.

1. Suppose the source is in state t_x.
   a. If the source tells x is possibly the actual value of $\mathbf{x}$, i.e., the information $\mathbf{x} \in A$ provided by the source is such that $x \in A$, then one must conclude that x is possibly the actual value of $\mathbf{x}$;
   b. If the source tells x is not a possibility for the actual value of $\mathbf{x}$, i.e., $x \notin A$, then one must conclude that x is not a possibility for the actual value of $\mathbf{x}$.
2. Suppose the source is in state $\neg t_x$.
   a. If the source declares x is possibly the actual value of $\mathbf{x}$, then one must conclude that x is not a possibility for the actual value of $\mathbf{x}$;
   b. If the source tells x is not a possibility for the actual value of $\mathbf{x}$, then one must conclude that x is possibly the actual value of $\mathbf{x}$.

Let $\mathscr{H}$ denote the possible states of S with respect to its truthfulness for all $x \in X$. By definition, $\mathscr{H} = \times_{x \in X} \mathscr{T}_x$. Furthermore, let h_B, $B \subseteq X$, be the state where the source tells the truth for all $x \in B$ and lies for all $x \notin B$. For instance, let $X = \{x_1, x_2, x_3, x_4\}$ and $B = \{x_3, x_4\}$, then $h_B = \left(\neg t_{x_1}, \neg t_{x_2}, t_{x_3}, t_{x_4}\right)$. We have thus $\mathscr{H} = \times_{x \in X} \mathscr{T}_x = \{\cup_{B \subseteq X} h_B\}$ and there are $2^{|X|}$ states h_B, $B \subseteq X$, of which $2^{|X|} - 1$ are distinct lies and the remaining one, h_X, corresponds to telling the truth for all $x \in X$.

Let us now consider the following question: suppose a source declares $\mathbf{x} \in A$ and is in state h_B, what must one conclude about $\mathbf{x}$? The answer follows directly from the fact that the four cases above correspond to the Boolean equivalence connective: one must conclude $\mathbf{x} \in (A \cap B) \cup (\overline{A} \cap \overline{B})$ [4]. For all $A \subseteq X$, we can define a multivalued mapping Γ_A from $\mathscr{H}$ to X that encodes this reasoning:

$$\Gamma_A(h_B) = (A \cap B) \cup (\overline{A} \cap \overline{B}), \quad \forall B \subseteq X.$$

$\Gamma_A(h)$ indicates how to interpret the information $\mathbf{x} \in A$ provided by the source in each configuration h of the source. We may also consider non elementary hypotheses H, $H \subseteq \mathscr{H}$, corresponding to subsets of possible states of the source. Let $\Gamma_A(H)$ denote the image of H under Γ_A. It is defined as $\Gamma_A(H) = \cup_{h \in H} \Gamma_A(h)$.

This framework allows us to represent the various forms of lack of truthfulness mentioned above. For instance, let $X = \{x_1, x_2, x_3, x_4\}$ and suppose a source tells $\mathbf{x} \in \{x_3, x_4\}$: if the source is in state $h_{\{x_1, x_4\}}$, this means that it actually knows $\mathbf{x} \in \{x_2, x_4\}$, i.e., it is telling something different yet consistent with what it knows; if the source is in state $h_{\{x_1, x_2, x_3\}}$, then it knows $\mathbf{x} \in \{x_3\}$ and thus is telling less. Let us also remark that the state h_X corresponds to the state of the source being truthful in the truthfulness model considered in [4] and recalled above, as it also yields the conclusion $\mathbf{x} \in A$ from a piece of information $\mathbf{x} \in A$, $A \subseteq X$. The truthfulness model of [4] is actually a particular case of the one proposed here, since the state $h_\emptyset$ (lying for all $x \in X$) corresponds to the state of non truthfulness in [4].

2.2 Uncertain Meta-knowledge and Testimony

Consider the situation where an agent's meta-knowledge on the truthfulness of S is uncertain and represented by subjective probabilities $p^{\mathscr{H}}(h)$, $h \in \mathscr{H}$. For instance, suppose the agent knows that S behave similarly for all $x \in X$, that is, for any $x \in X$ the probability of telling the truth is α and the probability of lying is $1-\alpha$. Besides, the agent knows that the behavior of the source are independent for all $x \in X$, i.e., the variables $\mathbf{t}_x$, $x \in X$, are independent. In other words, the agent knows that the truthfulness of the source for all $x \in X$ may be assimilated to a Bernoulli process. We have then $p^{\mathscr{H}}(h_A) = \alpha^{|A|}(1-\alpha)^{|\overline{A}|}$, for all $A \subseteq X$, since the probability that the source lies for all elements in $\overline{A}$ is equal to $\times_{x \in \overline{A}} p(\neg t_x) = (1-\alpha)^{|\overline{A}|}$ and the probability that it tells the truth for all elements in A is equal to $\times_{x \in A} p(t_x) = \alpha^{|A|}$.

Following Dempster's approach [1], a testimony $\mathbf{x} \in A$ provided by S will then be interpreted by a belief function [5] with associated mass function m_A on X defined by, for all $B \subseteq X$: $m_A(B) = p^{\mathscr{H}}(h)$, where $B = \Gamma_A(h)$. Formally, a mass function m on X is a probability distribution on the power set of X, hence $\sum_{A \subseteq X} m(A) = 1$. Mass functions can encode various forms of knowledge, for instance the so-called vacuous mass function $m_\top$ defined by $m_\top(X) = 1$ represents total ignorance about the actual value of $\mathbf{x}$.

More generally, the testimony of the source may be uncertain and represented by a mass function m_S on X. Each testimony $\mathbf{x} \in A$ is then allocated mass $m_S(A)$, yielding the following mass function:

$$m(B) = \sum_A m_S(A) m_A(B), \quad \forall B \subseteq X. \tag{1}$$

Let us remark that the framework introduced in this paper for modeling source truthfulness is actually a particular case of an approach to account for general source behavior assumptions proposed in [4] and that Reference [4] may readily be used to provide a formal derivation of (1).

2.3 The Case of Multiple Sources

Let us now consider that there are two sources S_1 and S_2 providing the pieces of information $\mathbf{x} \in A$ and $\mathbf{x} \in B$, respectively. Let $\mathscr{H}_1$ and $\mathscr{H}_2$ denote the set of possible state configurations of each source. The set of elementary joint state assumptions on sources is $\mathscr{H}_{12} = \mathscr{H}_1 \times \mathscr{H}_2$. Following the approach described in [4], a multivalued mapping $\Gamma_{A,B}$ from $\mathscr{H}_{12}$ to X, which assigns to each elementary hypothesis $h = (h^1, h^2), h \in \mathscr{H}_{12}$, the result of the fusion of the two pieces of information $\mathbf{x} \in A$ and $\mathbf{x} \in B$ may be defined as follows: $\Gamma_{A,B}(h) = \Gamma_A(h^1) \cap \Gamma_B(h^2)$ and, more generally, $\Gamma_{A,B}(H) = \bigcup_{(h^1,h^2) \in H} (\Gamma_A(h^1) \cap \Gamma_B(h^2))$, for all $H \subseteq \mathscr{H}_{12}$.

Now, suppose uncertain meta-knowledge on the two sources, represented by a mass function $m^{\mathscr{H}_{12}}$, and that the sources provide uncertain information m_1 and m_2, respectively, on X. Assume further that the sources are independent, where independence means the following: if we interpret $m_i(A)$ as the probability that the source

S_i provide the information $\mathbf{x} \in A$, then the probability that the source S_1 provide the information $\mathbf{x} \in A$ and the source S_2 provide conjointly the information $\mathbf{x} \in B$ is the product $m_1(A) \cdot m_2(B)$ [4]. In such a situation, one may use the so-called Behavior-Based Fusion (BBF) rule introduced in [4], to obtain the following mass function m on X:

$$m(C) = \sum_{H} m^{\mathscr{H}_{12}}(H) \sum_{A,B:C=\Gamma_{A,B}(H)} m_1(A)\, m_2(B), \quad \forall C \subseteq X. \tag{2}$$

Two important variants of this last equation are the unnormalized version of Dempster's rule [1] and the negation of the exclusive disjunctive rule [2, 8]. The former rule is recovered with $m^{\mathscr{H}_{12}}\left(\left\{(h_X^1, h_X^2)\right\}\right) = 1$, i.e., both sources are truthful, since we have $\Gamma_{A,B}(\left\{(h_X^1, h_X^2)\right\}) = A \cap B$. The latter rule is recovered with $m^{\mathscr{H}_{12}}\left(\left\{(h_X^1, h_X^2), (h_\emptyset^1, h_\emptyset^2)\right\}\right) = 1$, i.e., both or none of the sources are truthful, since we have $\Gamma_{A,B}(\left\{(h_X^1, h_X^2), (h_\emptyset^1, h_\emptyset^2)\right\}) = (A \cap B) \cup (\overline{A} \cap \overline{B})$. This latter rule will be called for short the *equivalence rule* in the remainder of this paper, since it corresponds to the Boolean equivalence connective.

3 α-Conjunctions

In this section, we first recall some basic and necessary notions on the α-conjunctions. We then proceed with the disclosure of an interpretation for these rules.

3.1 Basic Notions

Smets introduced the α-junctions in [8, 9] as follows. Let $\mathscr{M}^X$ be the set of mass functions on X. Let m_1 and m_2 be two mass functions on X. Suppose we want to build a mass function m_{12} such that $m_{12} = f(m_1, m_2)$. Smets [8] determined the operators that map $\mathscr{M}^X \times \mathscr{M}^X$ to $\mathscr{M}^X$ and that satisfy the following requirements (the origins of those requirements are summarized in [9, p.25]).

- Linearity: $f(m, pm_1 + qm_2) = pf(m, m_1) + qf(m, m_2)$, $p \in [0,1]$, $q = 1 - p$.
- Commutativity: $f(m_1, m_2) = f(m_2, m_1)$.
- Associativity: $f(f(m_1, m_2), m_3) = f(m_1, f(m_2, m_3))$.
- Neutral element: existence of a mass function m_0 such that $f(m, m_0) = m$ for any m.
- Anonymity: relabeling the elements of X does not affect the results.
- Context preservation: let pl_i be the plausibility function [10] associated to the mass function m_i and defined by $pl_i(A) = \sum_{B \cap A \neq \emptyset} m_i(B)$, for all $A \subseteq X$ (the quantity $pl_i(A)$ represents the probability that the proposition $\mathbf{x} \in A$ can not be refuted by the available information). Context presentation correspond to the requirement: if $pl_1(A) = 0$ and $pl_2(A) = 0$ for some $A \subseteq X$, then $pl_{12}(A) = 0$.

Smets [8] showed that there are two families of rules that satisfy these requirements: one for each of the only two possible solutions for m_0, which can only be, as shown by Smets, either $m_0 = m_\top$ or $m_0 = m_\bot$, with $m_\bot$ the mass function defined by $m_\bot(\emptyset) = 1$. Besides, he showed that each of these two families depend on a parameter $\alpha \in [0,1]$. He called these families the α-conjunctions and α-disjunctions, respectively. In the remainder of this paper, we focus on the α-conjunctions.

Smets provided a complex definition for these rules. In [3], a simpler definition was found. We reproduce this latter definition here. Let m_1 and m_2 be two mass functions and let $m_{1\textcircled{\cap}^\alpha 2}$ denote the mass function resulting from the α-conjunction of m_1 and m_2. We have, for all $D \subseteq X$ [3, Proposition 3]:

$$m_{1\textcircled{\cap}^\alpha 2}(D) = \sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap C)=D} m_1(A)\, m_2(B)\, m_\alpha(C), \tag{3}$$

where $m_\alpha(A) = \alpha^{|\overline{A}|}(1-\alpha)^{|A|}$, for all $A \subseteq X$. The α-conjunctions include the conjunctive rule (for $\alpha = 1$) and the equivalence rule (for $\alpha = 0$).

3.2 Interpretation

Suppose meta-knowledge on two sources S_1 and S_2 of the following form:

- For each $x \in X$, they both tell the truth with probability α and both lie with probability $1-\alpha$. Besides, their behavior for all $x \in X$ are independent.
- Or they are both truthful.

The first part of this meta-knowledge amounts to the hypotheses $\left(h_A^1, h_A^2\right)$, $A \subseteq X$, being allocated probability $\alpha^{|A|}(1-\alpha)^{|\overline{A}|}$ since for any $A \subseteq X$, the probability that both sources lie for all elements in $\overline{A}$ is equal to $\times_{x\in\overline{A}} p((\neg t_x^1, \neg t_x^2)) = (1-\alpha)^{|\overline{A}|}$ and the probability that they both tell the truth for all elements in A is equal to $\times_{x\in A} p((t_x^1, t_x^2)) = \alpha^{|A|}$. Since the assumption that both sources are truthful correspond to the hypothesis $\left(h_X^1, h_X^2\right)$, we have that the above meta-knowledge on the sources truthfulness can be represented by the following mass function on $\mathscr{H}_{12}$:

$$m^{\mathscr{H}_{12}}\left(\left\{\left(h_X^1, h_X^2\right), \left(h_A^1, h_A^2\right)\right\}\right) = \alpha^{|A|}(1-\alpha)^{|\overline{A}|}, \quad \forall A \subseteq X. \tag{4}$$

Theorem 1. *Let m_1 and m_2 be two mass functions on X provided by two independent sources S_1 and S_2. Let m be the mass function obtained by combining m_1 and m_2 using the BBF rule (2), with $m^{\mathscr{H}_{12}}$ defined by (4). We have $m_{1\textcircled{\cap}^\alpha 2} = m$.*

Proof. (Sketch) It may easily be shown that, for all $A, B, C \subseteq X$,

$$\Gamma_{A,B}\left(\left\{\left(h_X^1, h_X^2\right), \left(h_C^1, h_C^2\right)\right\}\right) = (A\cap B)\cup(\overline{A}\cap\overline{B}\cap\overline{C})$$

and thus the quantity

$$m^{\mathscr{H}_{12}}\left(\left\{\left(h_X^1, h_X^2\right), \left(h_C^1, h_C^2\right)\right\}\right)\cdot m_1(A)\cdot m_2(B) = \alpha^{|C|}(1-\alpha)^{|\overline{C}|}\cdot m_1(A)\cdot m_2(B)$$

is allocated to the subset $(A \cap B) \cup (\overline{A} \cap \overline{B} \cap \overline{C})$ by the BBF rule. Besides, from (3), we have

$$\sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap C)=D} m_1(A)\, m_2(B)\, m_\alpha(C) = \sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap\overline{C})=D} m_1(A)\, m_2(B)\, \overline{m}_\alpha(C),$$

where $\overline{m}_\alpha$ denotes the negation of m_α defined by $\overline{m}_\alpha(A) = m_\alpha(\overline{A})$, for all $A \subseteq X$ [2]. We have $\overline{m}_\alpha(A) = \alpha^{|A|}(1-\alpha)^{|\overline{A}|}$, for all $A \subseteq X$, and thus the quantity $\alpha^{|C|}(1-\alpha)^{|\overline{C}|} \cdot m_1(A) \cdot m_2(B)$ is transferred by the α-conjunctive rule to the subset $(A \cap B) \cup (\overline{A} \cap \overline{B} \cap \overline{C})$, for all $A, B, C \subseteq X$. □

Theorem 1 shows that an α-conjunction is a particular case of the BBF procedure and as such corresponds to a special meta-knowledge on the sources. This meta-knowledge, represented by (4), basically comes down to assuming that either both sources tell the truth or they commit the same lie h_A, with probability $\alpha^{|A|}(1-\alpha)^{|\overline{A}|}$. In [3], the α-conjunctions are decomposed into simple pieces of evidence on the truthfulness of the sources, using the same definition of truthfulness as the one adopted in [4]. However, even if it is easy to understand the meaning of each of the simple pieces of evidence in [3], it is difficult to capture the meaning of their combination and thus of the α-conjunctions. Comparatively, in the present paper, thanks to the new richer model of source truthfulness, we are able to provide a single mass function on the truthfulness of the sources, which admits a clear interpretation and from which the α-conjunctions can be recovered.

Let us end this section with a few comments on the behavior of the fusion scheme that the meta-knowledge $m^{\mathscr{H}_{12}}$ defined by (4) leads to. To analyze the behavior of the α-conjunctions, it is useful to consider the situation where ones receives two certain testimonies $\mathbf{x} \in A$ and $\mathbf{x} \in B$ from two sources such that $A \cap B \neq \emptyset$ (the case where the testimonies are uncertain is merely a generalization of what follows). One may show that in such a case, only the subsets D such that $A \cap B \subseteq D \subseteq (A \cap B) \cup (\overline{A} \cap \overline{B})$ will receive a non null mass after combining the two testimonies by an α-conjunctive rule. Precisely, each of those subsets D can be expressed as $(A \cap B) \cup C$ for some $C \subseteq \overline{A} \cap \overline{B}$ and one may show that each of those subsets $D = (A \cap B) \cup C$ will be allocated mass $\alpha^{|\overline{A}\cap\overline{B}|-|C|}(1-\alpha)^{|C|}$. One may then remark that as α goes from 0 to 1, masses flow from the least specific subsets D to the most specific subsets D. In particular, for $\alpha = 1$, all the mass is allocated to the subset $D = A \cap B$ (there is a probability equal to one of knowing $\mathbf{x} \in A \cap B$), i.e., the most specific subset D such that $A \cap B \subseteq D \subseteq (A \cap B) \cup (\overline{A} \cap \overline{B})$ is the result of the combination. Conversely, for $\alpha = 0$, the least specific subset D such that $A \cap B \subseteq D \subseteq (A \cap B) \cup (\overline{A} \cap \overline{B})$ is the result of the combination, i.e., all the mass is allocated to the subset $D = (A \cap B) \cup (\overline{A} \cap \overline{B})$. Another interesting particular case is $\alpha = 0.5$: all subsets D are allocated mass $1/2^{|\overline{A}\cap\overline{B}|}$, i.e., they are equiprobable. It appears thus that the α-conjunctions allow us to control the behavior of the combination, from the conjunctive rule to the equivalence rule, by ranging from the principle of maximum specificity to the principle of maximum entropy to the principle of minimum specificity with respect to the subsets D such that $A \cap B \subseteq D \subseteq (A \cap B) \cup (\overline{A} \cap \overline{B})$.

4 Conclusion

In this paper, an interpretation for the α-conjunctions was proposed. It was shown that they correspond to assuming that the sources behave similarly (they both tell the truth or commit the same lie), with some particular probability. Of special interest is the new framework that was introduced to provide this interpretation: it allows the integration of meta-knowledge on the various forms of lies the information sources may commit and it extends recent work [4] on the formalization of meta-knowledge on information sources.

Acknowledgements. The author thanks Didier Dubois and Thierry Denœux for inspiring discussions on the α-junctions. This work was supported by a grant from the French national research agency through the CSOSG research program (project CAHORS).

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
2. Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. International Journal of General Systems 12(3), 193–226 (1986)
3. Pichon, F., Denœux, T.: Interpretation and computation of α-junctions for combining belief functions. In: Sixth Int. Symp. on Imprecise Probability: Theories and Applications (ISIPTA 2009), Durham, United Kingdom (July 2009)
4. Pichon, F., Dubois, D., Denœux, T.: Relevance and truthfulness in information correction and fusion. International Journal of Approximate Reasoning 53(2), 159–175 (2012)
5. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
6. Smets, P.: The combination of evidence in the Transferable Belief Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)
7. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1), 1–35 (1993)
8. Smets, P.: The α-junctions: Combination Operators Applicable to Belief Functions. In: Nonnengart, A., Kruse, R., Ohlbach, H.J., Gabbay, D.M. (eds.) FAPR 1997 and ECSQARU 1997. LNCS, vol. 1244, pp. 131–153. Springer, Heidelberg (1997)
9. Smets, P.: The application of the matrix calculus to belief functions. International Journal of Approximate Reasoning 31(1-2), 1–30 (2002)
10. Smets, P., Kennes, R.: The Transferable Belief Model. Artif. Intell. 66, 191–243 (1994)

Improvements to the GRP1 Combination Rule

Gavin Powell, Matthew Roberts, and Dafni Stampouli

Abstract. The recursive use of belief function combination rules, as required with temporal data, is issue prone. Systems will either become unreactive, through a greedy empty set, or provide a false sense of security through applying a closed world model to an open world scenario. We improve on the previous combination rule GRP1 to enhance its ability to work with temporal data in an open world. Specifically we have progressed with the dynamic self adjustment properties of the rule, which allow it to gauge how fusion should take place dependant on the temporal information that it receives. Comparisons are made between the improved GRP1 rule and other rules which have been applied to temporal datasets.

1 Introduction

Sensors are extensively used to continually monitor the environment of interest. Measurements are taken at each time step and provide a rich stream of information about the object or event that is being sensed. Analysis of this information provides understanding of the observed object or event, its changes, and helps us manage the inherent issues related to sensing within the real world (such as sensor errors or failures). Existing belief function fusion methods tend not to take advantage of the temporal information and look at the situation at a single time step and base

Gavin Powell
EADS Innovation Works, Newport, United Kingdom
e-mail: gavin.powell@eads.com

Matthew Roberts
EADS Innovation Works, Newport, United Kingdom
e-mail: matthew.roberts@eads.com

Dafni Stampouli
EADS Innovation Works, Newport, United Kingdom
e-mail: dafni.stampouli@eads.com

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 293–300.
springerlink.com

decisions on this. Where temporal information is available, it should be utilised as much as possible in order to enhance the decision process. Analysis of this incoming stream of information will allow for an understanding of the object or state, how well the sensing medium is performing, and the changes that are occurring in the world that is being sensed. A new combination rule is being developed to utilise temporal information by dynamically accounting for changes in input information and adjusting the means of combination in order to provide a more robust output from which decisions can be made. At the same time, the new rule allows for open world scenarios, by retaining the empty set.

In this paper we examine and present findings based upon the combination of open world belief assignments and temporal information and we follow on from previous work of the authors [4, 5, 6] and address previous feedback. Section 2 presents an outline of belief functions and discusses issues with existing methods. Section 3 and 4 discuss the GRP1 rule [4] and the proposed improvements. A scenario to classify target vehicle based on their kinematic information is outlined in Section 5. This scenario is used to demonstrate disadvantages of well established combination rules and provide a comparison to the proposed method. The results are analysed and discussed in Section 6 and demonstrate the improvement achieved in terms of stability and avoidance of false sense of security.

2 Belief Functions

Belief functions are a mature technology for use in fusing or combining data and information. Their origins can be said to lie in work undertaken by Arthur Dempster and Glenn Shafer [1, 9] who defined the set theoretic means of combination of information of Dempster-Shafer Theory (DST). Difficulties in its use within an open world and with temporal data are documented [4, 5, 6] and an understanding of its abilities and disabilities are essential to its effective use. Variants and extensions to the original DST framework and its rules of combination are available where the notable extensions to the original DST framework are Dezert-Smarandache Theory (DSmT) [2] and the Transferable Belief Model (TBM) [11]. DSmT utilises an open world scenario through inclusion of an additional hypothesis that accounts for the open world. This approach allows for an open world when redistribution of the empty set takes place, providing a much more complete framework to work with. A notable issue of set based approaches is the curse of dimensionality where computational load rises rapidly as the number of possible outcomes increases creating a limit based on computing power but at present falls between 10–20 possible outcomes. DSmT, through its more complete framework, can exaggerate this issue. The TBM also allows for an open world by not redistributing the empty set after combination and using this as a marker for the ‘anything else’ outcome of the open world. In practice though the use of the unnormalised conjunctive rule of combination will degrade the system as more pieces of evidence are recursively combined.

3 Recursive Combination

Errors in the model or sensors can be hidden by normalisation and as such they become silent problems, particularly so in recursive applications where the incoming information is used to update the current estimate. Smets noted this in his work and proposed that normalisation should not take place and that the conjunctive rule (or similar) can be used without normalisation [11]. This provides the advantage of retaining the empty set value that will allow for an open world, but also allows for an understanding of the state of the system and the information that is being combined. The empty set is a marker for conflict within the information which should be used to assist in any later decision making. By removing the normalisation phase your model of the object, state or event will deteriorate as more information is combined [5]. To counter act this, the authors previously proposed the GRP1 [4] algorithm which uses a discounted mixture of the conjunctive and disjunctive combination rule. This created a dynamic system that could self adjust in terms of how much onus should be placed on the incoming information due to how certain the estimate was.

GRP1 discounts incoming information based on the distribution of mass in the current state, object, event estimate. This meant that a lack of focus was given to the actual incoming data and how that compared to the current state, object, event estimate, GRP2 addresses this.

4 Proposed Combination Rule

To fully understand the state of the (1 or many) sensors that are providing information to the system it is necessary to compare them to possibly each other at that time instant, with each other over time, or with the current state, object, event estimate. This will allow for an understanding of which sensors should be most prominent, to manage conflict within sensors and to mitigate for sensors when they are failing. As a first step we look at a single sensor and how that compares to the current state, object, event estimate. This allows for a single measure of how much in agreement the sensor is with the estimate at this instance in time thus identifying if it is likely to be failing or providing noisy or incorrect information. Through an unnormalised conjunctive combination between the incoming information and the current estimate we use the empty set to identify the amount of conflict there is. This measure of conflict is used to adjust the weighting parameter between the conjunctive and disjunctive mix, which previously in GRP1 was the arithmetic mean. If there is a great deal of conflict then we wish to be placing more of the weighting on the cautious disjunctive rule as we have less faith in that information, if there is little conflict we wish to place more weight on the conjunctive rule as we are more certain of this information and have more faith within it.

The GRP2 combination rule is

$$m_{GRP2}(A) = k\, m_{\cup}^{\alpha}(A) + (1-k)\, m_{\cap}^{\alpha}(A), \tag{1}$$

where $m^{\alpha}_{\cup}$ is the disjunctive combination rule after discounting, $m^{\alpha}_{\cap}$ is the conjunctive combination rule after discounting, and k is the conflict after conjunctive combination between the current estimate and the input from a sensor. Both $m^{\alpha}_{\cup}$ and $m^{\alpha}_{\cap}$ are the combination (after discounting) between the current estimate and the input — only the *basic belief assignment (bba)* for the input is discounted. Discounting is used to reduce the impact of the new information in a manner that is proportional to the confidence of the existing estimate. The method used for discounting depends on whether the conjunctive or disjunctive combination rule is used — this is discussed and justified in previous work [4]. The same value of α is used for both discounting methods and is calculated using:

$$\alpha = p\left(2^{\Omega}\right) = \sum \frac{|\Omega| - |A|}{|\Omega| - 1} m_c(A) \qquad \forall A \neq \emptyset, A \in 2^{\Omega}, \tag{2}$$

where $p\left(2^{\Omega}\right)$ is precision, or educatedness [4], and $m_c(A)$ is the mass assigned to A for the current estimate.

RCR [3] is similar to the above rule but mass cannot be assigned to the empty set of the resultant *bba* and normalisation is required after the weighted average is calculated. This normalisation adjusts the weighting of the symmetric version of RCR, called RCR-S, to have the same linear weighting as found in Equation 1. Assigning mass to the empty set and not requiring normalisation can be advantageous — this is shown in Section 6.

5 Scenario

The scenario used in this paper is based on previous work [7, 8] which used simulations of vehicles to compare Wireless Sensor Network (WSN) tracking and classification algorithms. The simulation used here consists of an amphibious light tank moving over road, grass, and water. There are five possible target classes: amphibious light tank (ALT); pedestrian (Ped); car; light tank (LT); and main battle tank (MBT) (i.e. $\Omega = \{\text{ALT}, \text{Ped}, \text{Car}, \text{LT}, \text{MBT}\}$). At each time step, a subset of the nodes within the WSN are used to update the kinematic state estimate of the target — this is then used within the TBM to provide a target classification which is fused over time to provide a more reliable classification.

The classification produced at each time step is a *bba* [11] which is conditional upon the target speed and the terrain beneath the target. This work uses the, non-fused, conditional assignment produced at each time step as its input. This input is fused over time using different combination rules to compare their performance (see Section 6).

6 Results

Four classification methods have been used (PCR5 [10], RCR-S, GRP1, and GRP2) and compared in their task of identifying a vehicle moving over different terrain

(see Figures 1–4, the legend can be found in Figure 1). The evaluation that follows is a qualitative one, where the different methods are compared in terms of correctness of classification, timely response, amount of confidence, stability in decision, and variation between classifications over time. Further analysis could be conducted in the future to compare the methods in terms of accuracy, which determines how close different results agree with the true class.

The algorithms which are being used to compare against are designed for a closed world and as such an open world comparison is unjust as the GRPx algorithms are designed to succeed in that environment. The open world of the GRPx algorithms is an important feature even within a closed world. Here the empty set value provides a measure of conflict and uncertainty of the sensing mediums. This can show errors in the system/model or faulty sensors/communications links and can aid in the decision process so that false positives/negatives are more readily avoided.

For the given scenario, the vehicle observed during the first 50 seconds is moving on a road, from 50 to 70 seconds it is moving on grass and then enters the water till the 95th second at which it goes onto grass again until the end of the sensing.

Looking at the results more closely, Figure 1 shows the BetPs of all the classes obtained using PCR5. The BetP for the ALT is steadily increasing while at the 75th second when the vehicle is about to enter the water, the system gives a belief of 1, that it is a car and a belief of 0 that it is an ALT. Within 10 seconds it totally

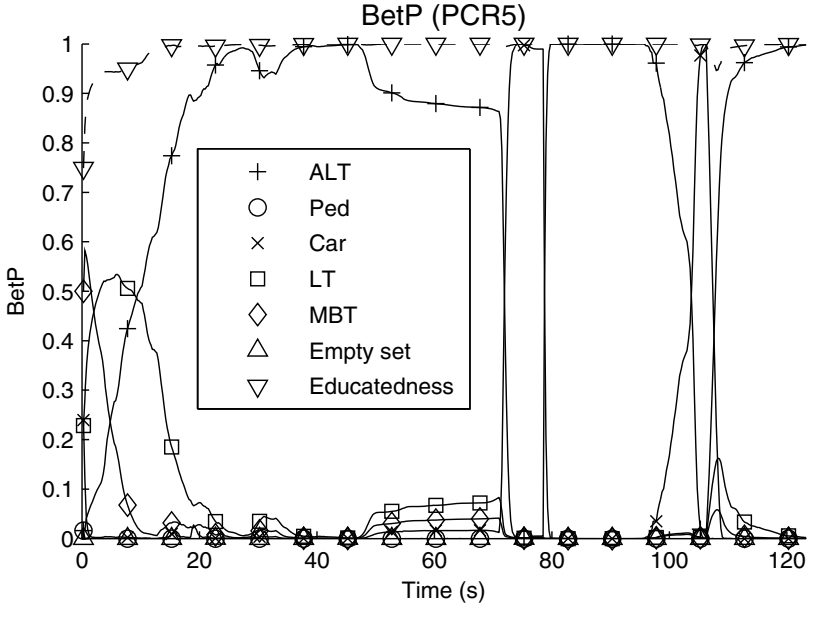


Fig. 1 BetP after applying PCR5 at each time step.

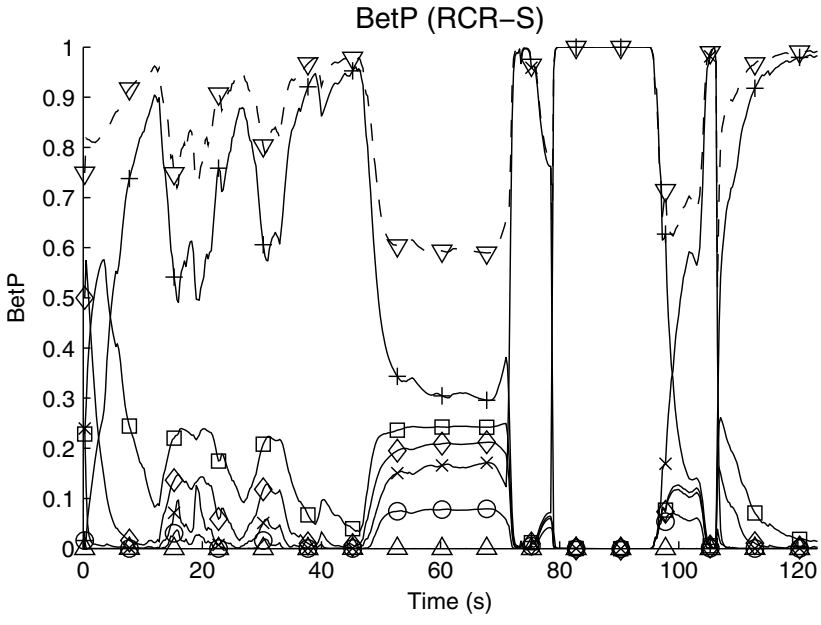


Fig. 2 BetP after applying RCR-S at each time step.

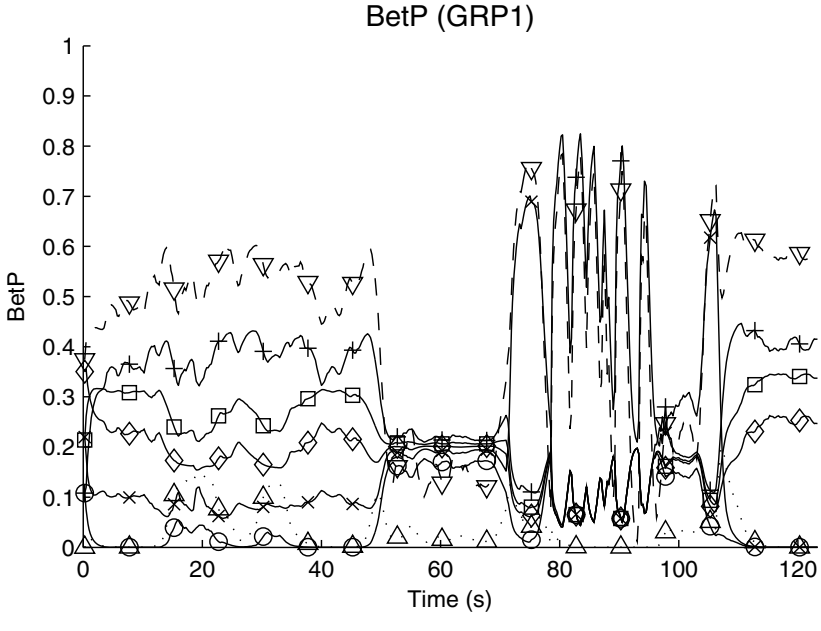


Fig. 3 BetP after applying GRP1 at each time step.

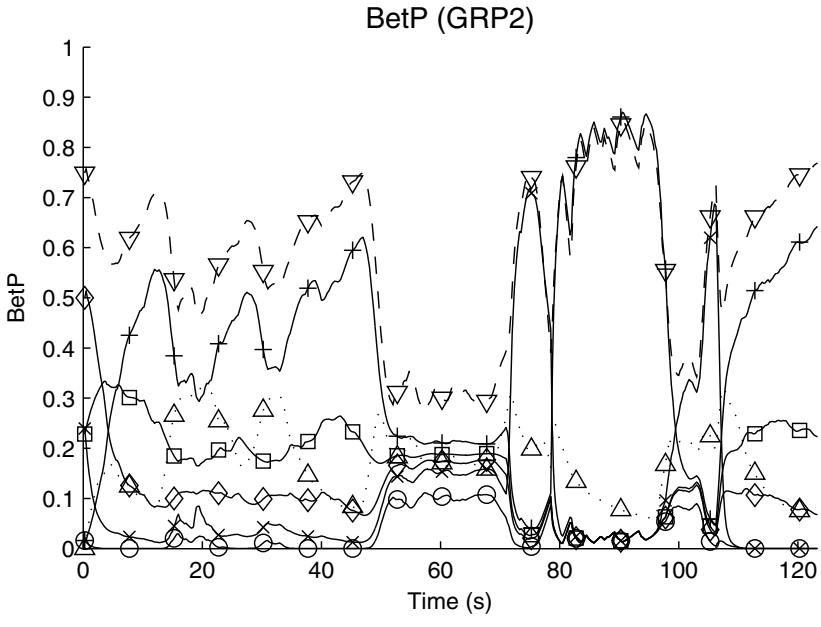


Fig. 4 BetP after applying GRP2 at each time step.

changes its mind and gives a belief of 0 to the car and a belief of 1 to the ALT, and again changes its mind on the 105th second when the vehicle moves on grass. Taking into account that incoming information is very noisy, we can conclude that PCR5 over-inflates the confidence in each class, and is unstable (is certain about one class and 10 seconds later is it certain about another). This happened due to the normalisation stage which hides problems in the sensors or the model. Similar unstable results are shown by RCR-S, Figure 2 where the system fluctuates between classes with high certainty within 20 seconds when the vehicle exits the water.

GRP1 (Figure 3), on the other hand, takes into account uncertainty and incorporates it into the empty set. The empty set has some value in this scenario which indicates that there is noise in the incoming information. GRP1, when the vehicle is in the water, presents stability issues and ripples the BetP of the ALT. GRP2 (Figure 4) on the other hand, solves this problem and classifies correctly the vehicle without over-inflating the confidence as PCR5 and RCR-S do. It also presents the most stable result among the four methods. Furthermore in terms of timely response, both PCR5 and RCR-S are too fast to respond to changing inputs. This creates large oscillations. GRP1 and GRP2 on the other hand, take into account previous states (through educatedness), which smoothes the output, balancing a timely response and stability.

7 Conclusion

We have shown that it is possible to recursively combine information using an open world model and to adapt in a logical way as the vagueness and consistency of the information stream changes over time.

Through our work on this paper it is clear that a new combination rule, that isn't reliant on the unnormalised combination rule, needs to be created. Retaining the empty set and open world are critical, as is non associativity and one that will not degrade as more sensors or information are combined. Currently the unnormalised combination rule will begin to degrade as soon as fusion begins and will continue as more information sources are added. This makes combination rules based on this unusable for recursive fusion of multiple sensors over time which is the next goal for belief functions.

References

1. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society 30(2), 205–247 (1968)
2. Dezert, J., Smarandache, F.: An introduction to DSmT. CoRR abs/0903.0279 (2009)
3. Florea, M., Jousselme, A.L., Bosséb, E., Grenier, D.: Robust combination rules for evidence theory. Information Fusion 10(2), 183–197 (2009), doi:10.1016/j.inffus.2008.08.007
4. Powell, G., Roberts, M.: GRP1. A recursive fusion operator for the transferable belief model. In: Proceedings International Conference on Information Fusion 2011 (2011)
5. Powell, G., Roberts, M., Marshall, D.: Empty set biasing issues in the Transferable Belief Model for fusing and decision making. In: Proceedings International Conference on Information Fusion 2010 (2010)
6. Powell, G., Roberts, M., Marshall, D.: Pitfalls for recursive iteration in set based fusion. In: Workshop on the Theory of Belief Functions (2010)
7. Roberts, M.: Tracking and classification with wireless sensor networks and the transferable belief model. Ph.D. thesis, Cardiff School of Computer Science & Informatics, Cardiff University (2010)
8. Roberts, M., Marshall, D., Powell, G.: Improving joint tracking and classification with the Transferable Belief Model and terrain information. In: Proceedings International Conference on Information Fusion 2010 (2010)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
10. Smarandache, F., Dezert, J.: Information fusion based on new proportional conflict redistribution rules. In: Proceedings International Conference on Information Fusion 2005, vol. 2, pp. 907–914 (2005), doi:10.1109/ICIF.2005.1591955
11. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence 66(2), 191–234 (1994)

Consensus-Based Credibility Estimation of Soft Evidence for Robust Data Fusion

Thanuka L. Wickramarathne, Kamal Premaratne, and Manohar N. Murthi

Abstract. Due to its subjective nature which can otherwise compromise the integrity of the fusion process, it is critical that *soft evidence* (generated by human sources) be validated prior to its incorporation into the fusion engine. The strategy of *discounting* evidence based on source reliability may not be applicable when dealing with soft sources because their reliability (e.g., an eye witnesses account) is often unknown beforehand. In this paper, we propose a methodology based on the notion of *consensus* to estimate the credibility of (soft) evidence in the absence of a 'ground truth.' This estimated credibility can then be used for source reliability estimation, discounting or appropriately 'weighting' evidence for fusion. The consensus procedure is set up via Dempster-Shafer belief theoretic notions. Further, the proposed procedure allows one to constrain the consensus by an estimate of the ground truth if/when it is available. We illustrate several interesting and intuitively appealing properties of the consensus procedure via a numerical example.

1 Introduction

Motivation: The recent interest in defense-related fusion community in incorporating *soft evidence* (e.g., witness statements) into the fusion process has spawned a multitude of research avenues and challenges. As a vital aspect of any robust evidence fusion strategy, one crucial question to be addressed is, how can (soft) evidence be validated when the *ground truth (GT)* is not known?

Challenges: Dempster Shafer (DS) belief theory is increasingly being used in the fusion community due to the flexibility it provides in modeling and decision making in imperfect data domains. In the DS framework, if the reliability (a measure of trustworthiness of a source based on past performance) of sources are available, one

Thanuka L. Wickramarathne · Kamal Premaratne · Manohar N. Murthi
Dept. of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146 USA
e-mail: t.wickramarathne@umiami.edu, (kamal,mmurthi)@miami.edu

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 301–309.
springerlink.com

can account for the credibility of evidence (a measure of trustworthiness of current evidence) via a procedure referred to as *discounting*. However, one often has to deal with (soft) sources whose reliability is not known beforehand.

In a typical fusion scenario, soft evidence is usually gathered from many sources whose actual reliabilities may not be known. When an adequate number of sources are considered, it is not unreasonable to assume that the truth is reflected in the majority opinion. If this majority opinion can be established via some rational aggregation procedure, the very aggregate, often referred to as a *consensus*, can in turn be used for credibility estimation. However, due to the subjective (and hence possibly inconsistent and even contradictory) nature of soft evidence [6], simple averaging operations may not be adequate enough to provide a *rational consensus* [3]. Further, in many applications, even though the GT is unknown, highly reliable rough estimates can still be generated, perhaps based on hard sensor data and/or expert opinions. Thus, one can actually 'drive' the consensus towards the GT, if the process can be constrained by an estimate of the GT.

Contributions: In this paper, we propose a consensus-based technique for credibility estimation of evidence in the absence of the GT. This estimated credibility can then be used for source reliability estimation and evidence discounting prior to fusion operations. The consensus procedure is set up via DS theoretic notions of evidence fusion, thus allowing it the flexibility to capture a variety of imperfections inherent to soft evidence. Further, the proposed procedure allows one to constrain the consensus by an estimate of the GT if/when it is available. While the detailed proofs of convergence have been omitted due to space limitations, several interesting and intuitively appealing properties of the consensus procedure are illustrated via a numerical example.

2 DS Theory Preliminaries

In DS theory, the *frame of discernment (FoD),* $\Theta = \{\theta_1, \ldots, \theta_n\}$, refers to the set of mutually exclusive and exhaustive propositions of interest; a proposition θ_i represents the lowest level of discernible information.

Definition 1. The mapping $m : 2^{\Theta} \mapsto [0,1]$ is a *basic probability assignment (BPA)* or *mass function* for the FoD Θ if $\sum_{B \subseteq \Theta} m(B) = 1$ with $m(\emptyset) = 0$. Consider the proposition $B \subseteq \Theta$. Let $\overline{B} = \Theta \setminus B$.

(i) When $m(B) > 0$, B is referred to as a *focal element* and the quantity $m(B)$ is the *mass* allocated to B.

(ii) The set of focal elements is the *core* $\mathfrak{F}$; the triplet $\mathscr{E} \equiv \{\Theta, \mathfrak{F}, m(\centerdot)\}$ is the corresponding *body of evidence (BoE).*

(iii) The mapping $\mathrm{Bl} : 2^{\Theta} \mapsto [0,1]$ where $\mathrm{Bl}(B) = \sum_{C \subseteq B} m(C)$ is the *belief* of B; the mapping $\mathrm{Pl} : 2^{\Theta} \mapsto [0,1]$ where $\mathrm{Pl}(B) = 1 - \mathrm{Bl}(\overline{B})$ is the *plausibility* of B. ■

Theorem 1 (Fagin-Halpern Conditionals). *[1] For $B \subseteq \Theta$ and a conditioning event A s.t.* $\mathrm{Bl}(A) > 0$, *the conditional belief* $\mathrm{Bl}(B|A)$ *is given by* $\mathrm{Bl}(B|A) = \mathrm{Bl}(A \cap B) / [\mathrm{Bl}(A \cap B) + \mathrm{Pl}(A \cap \overline{B})]$. ■

Evidence Combination: This refers to the process of combining BoEs $\mathscr{E}_i \equiv \{\Theta_i, \mathfrak{F}_i, m_i(.)\}$, $i = 1,2$, to arrive at a new BoE $\mathscr{E} \equiv \{\Theta, \mathfrak{F}, m(.)\}$ representing the aggregated evidence. We restrict our discussion to BoEs with identical FoDs (i.e., $\mathscr{E}_i$ s.t. $\Theta_i = \Theta, \forall i$); for the non-identical FoDs case, see [10] and references therein. Henceforth, we will use $\Re^+$ to denote the non-negative reals. Also, $\overline{1,n}$ and $\overline{1,n} \setminus i$ denote the sets $\{1,\ldots,n\}$ and $\{1,\ldots,n\} \setminus \{i\}$, respectively.

Definition 2 (Dempster's Combination Rule (DCR)). The BoE $\mathscr{E}$ generated by fusing BoEs $\mathscr{E}_1$ and $\mathscr{E}_2$ is $\mathscr{E} \equiv \mathscr{E}_1 \oplus \mathscr{E}_2$, where

$$m(B) = \sum_{C \cap D = B} \frac{m_1(C)\, m_2(D)}{1-K}, \forall B \subseteq \Theta, \text{ whenever } K = \sum_{C \cap D = \emptyset} m_1(C)\, m_2(D) \neq 1. \quad \blacksquare$$

Evidence Updating: This refers to the process of updating the evidence in a BoE $\mathscr{E}_i[k]$ with evidence received from the BoEs $\mathscr{E}_j[k]$, $j = \overline{1,n} \setminus i$, to arrive at $\mathscr{E}_i[k+1]$. Here k denote the discrete update index. We denote this as $\mathscr{E}_i[k+1] \equiv \mathscr{E}_i[k] \lhd \mathscr{E}_1[k] \ldots \mathscr{E}_{i-1}[k] \mathscr{E}_{i+1}[k] \ldots \mathscr{E}_n[k]$. The updating scheme proposed in [7] (for $n = 2$ case) provides several interesting properties applicable to the task at hand.

Definition 3 (Conditional Update Equation (CUE)). [7] The CUE that updates $\mathscr{E}_1[k]$ with the evidence in $\mathscr{E}_2[k]$ is

$$Bl_1(B)[k+1] = \alpha_1[k]\, Bl_1(B)[k] + \sum_{A \in \mathfrak{F}_2[k]} \beta_2(A)[k]\, Bl_2(B|A)[k], \ \forall k \geq 0.$$

The parameters $\alpha_1[k], \beta_2(.)[k] \in \Re^+$ satisfy $\alpha_1[k] + \sum_{A \in \mathfrak{F}_2[k]} \beta_2(A)[k] = 1$. $\blacksquare$

We can extend the CUE to get a strategy for fusion of evidence from $\mathscr{E}_1$ and $\mathscr{E}_2$:

Definition 4 (Conditional Fusion Equation (CFE)). The CFE that fuses the evidence of $\mathscr{E}_1$ and $\mathscr{E}_2$ is

$$\mathrm{Bl}(B) = K_1 \sum_{A \in \mathfrak{F}_1} \beta_1(A)\ \mathrm{Bl}_1(B|A) + K_2 \sum_{A \in \mathfrak{F}_2} \beta_2(A)\ \mathrm{Bl}_2(B|A).$$

The parameters $K_1, K_2, \beta_i(.) \in \Re^+$ satisfy $K_1 \sum_{A \in \mathfrak{F}_1} \beta_1(A) + K_2 \sum_{A \in \mathfrak{F}_2} \beta_2(A) = 1$. $\blacksquare$

Remarks:

1. Parameters $\alpha_1[.]$ and $\beta_2(A)[.]$ can be used to account for the *inertia* of $\mathscr{E}_1[.]$ and to appropriately weigh the evidence from $\mathscr{E}_2[.]$, respectively (see [7] for details).
2. The parameters K_1 and K_2 can be used to incorporate a measure of *relative importance* of the sources (e.g., evidence credibility) into fusion.
3. CUE and CFE can be generalized to handle the non-identical FoDs case [10]. $\blacksquare$

In our work, we will also use the following DS theoretic distance measure:

Definition 5. [2] The distance between two BoEs $\mathscr{E}_i$, $i = 1,2$, is given by

$$\text{dist}(\mathscr{E}_1, \mathscr{E}_2) = \sqrt{0.5\,(\mathbf{m}_1 - \mathbf{m}_2)^T\, D\,(\mathbf{m}_1 - \mathbf{m}_2)},$$

where $\mathbf{m}_i = \{m_i(\boldsymbol{.})\}$, $i = 1,2$, are $2^\Theta \times 1$ column vectors; and $D = \{d_{j\ell}\}$ is a $2^\Theta \times 2^\Theta$ matrix with $d_{j\ell} = |A_j \cap A_\ell| / |A_j \cup A_\ell|$, $A_j, A_\ell \in 2^\Theta$, $|\emptyset \cap \emptyset| / |\emptyset \cup \emptyset| \equiv 0$. ■

3 Credibility of Evidence

In this section, we present our consensus-based credibility estimation technique. The terms *credibility* and *reliability* are being used in the literature to refer to both evidence and the sources. For our purpose, we interpret these terms as follows: **(a) credibility** refers to "... the quality of being trusted and believed in (e.g., the government's loss of credibility) [8]"; **(b) reliability** refers to the notion of "... [being] consistently good in quality/performance or able to be trusted (e.g., a reliable source of information) [8]". So, credibility can be considered an *instantaneous* measure of trustworthiness (of evidence), while reliability is thought of as an *overall* measure of trustworthiness (of a source).

3.1 Credibility Estimation

A conflict-based credibility estimation method appears in [4][1].

Definition 6. Given the BoEs $\mathscr{E}_i$, $i \in \overline{1,n}$, the credibility of $\mathscr{E}_i$ is given by

$$Cr_{cf}(\mathscr{E}_i) = \left(1 - \text{conf}(\mathscr{E}_i, \mathscr{E}_{j\neq i})^\lambda\right)^{1/\lambda},$$

where $\lambda \in \Re^+$, $\mathscr{E}_{j\neq i} = \{\mathscr{E}_j \mid j \in \overline{1,n} \setminus i\}$ and $\text{conf}(\mathscr{E}_i, \mathscr{E}_{j\neq i})$ is the *conflict* between $\mathscr{E}_i$ and $\mathscr{E}_{j\neq i}$. Two variants Cr_{cf1} and Cr_{cf2} are

$$\text{conf}(\mathscr{E}_i, \mathscr{E}_{j\neq i}) = \begin{cases} \frac{1}{n-1} \sum\limits_{j \in \overline{1,n}\setminus i} \text{dist}(\mathscr{E}_i, \mathscr{E}_j), & \text{for } Cr_{cf1}; \\ \text{dist}(\mathscr{E}_i, \mathscr{E}_{\oplus\, j\neq i}), & \text{for } Cr_{cf2}, \end{cases}$$

where $\mathscr{E}_{\oplus\, j\neq i} = \mathscr{E}_1 \oplus \cdots \oplus \mathscr{E}_{i-1} \oplus \mathscr{E}_{i+1} \oplus \cdots \oplus \mathscr{E}_n$, for $i \in \overline{1,n}$. ■

However, with credibility viewed as a measure of the instantaneous trustworthiness of evidence, it makes sense to assess the credibility of a BoE by comparing it to the GT via a distance measure (such as what appears in Definition 5):

[1] The authors in [4] refer to this as a measure of relative reliability. However, to be consistent with our interpretations of the terms, we take their definition as a measure of credibility.

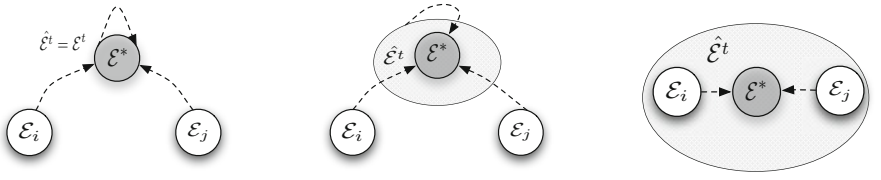
Definition 7. Let $\mathscr{E}^t$ denote the GT. Then, the credibility of the BoE $\mathscr{E}$ is given by

$$Cr_{con}(\mathscr{E}) = \left(1 - \text{dist}(\mathscr{E}, \mathscr{E}^t)^{\lambda}\right)^{1/\lambda}, \text{ where } \lambda \in \Re^{+}. \quad \blacksquare$$

As sensible as it appears, the difficulty with this strategy lies in the fact that the GT is usually absent. Is there a way to estimate the GT in such a situation? The notion of *consensus* has been used in many disciplines (e.g., social sciences, marketing/finance, engineering) and in a myriad of applications as a method to arrive at a 'general agreement' among opinions or sources (e.g., a consensus of opinion among judges). Thus, given the imprecise, unstructured and often inconsistent nature of soft evidence, we contend that a consensus provides an ideal estimate of the GT that can then be used for estimating credibility. Our approach thus differs from the work in [4] in that we first seek an estimate of the GT and use this estimate for credibility estimation.

3.2 Establishing a Consensus Among DS Theretic BoEs

For the task at hand (viz., estimation of credibility), we seek a consensus process that satisfies several desirable properties: the consensus being attained must **(P1)** be a *rational agreement* among the sources [3]; **(P2)** reach the GT (given by $\mathscr{E}^t$) when it is known (i.e., all evidence must converge to the GT); and **(P3)** be 'consistent' with a reliable estimate of the GT (given by $\hat{\mathscr{E}}^t$) when it is available (i.e., the consensus must be 'contained' within the estimate of the GT). See Figure 1.



(a) GT known (b) Estimate of GT known (c) Unknown GT

Fig. 1 The convergence behavior of BoEs to $\mathscr{E}^*$. Here, $\mathscr{E}^t$, $\hat{\mathscr{E}}^t$ and $\mathscr{E}^*$ denote the GT, an estimate of GT, and consensus, respectively. Note that $\mathscr{E}^*$ is always 'contained' within $\hat{\mathscr{E}}^t$.

We now present a consensus control strategy based on an extension to the CUE in Definition 3. This scheme mimics the process followed by humans to arrive at a consensus, viz., mutually exchange each other's opinions until all arrive at a consensus. Further, it can be shown that this process satisfies the properties **(P1)**-**(P3)**.

3.3 CUE-Based Consensus Control Strategy

Definition 8. Let $\mathfrak{E} = \{\mathscr{E}_i\}_{i=1}^{n}$ be the set of n BoEs among which a consensus is sought[2]. Update each $\mathscr{E}_i[k], \forall k \geq 1$, as

$$\mathrm{Bl}_i(B)[k+1] = \alpha_i[k]\,\mathrm{Bl}_i(B)[k] + \sum_{j \in \overline{1,n}\setminus i} \sum_{A_{ij} \in \mathfrak{F}_j} \beta_{ij}(A_{ij})[k]\, Bl_j(B|A_{ij})[k],$$

where $\mathscr{E}_i[0] = \mathscr{E}_i$, $i \in \overline{1,n}$. Here, $\alpha_i[k] = C_i[k] \in \Re^+$, $i \in \overline{1,n}$,

$$\beta_{ij}(A)[k] = \begin{cases} C_j[k]\, m_i(A)[k], & \text{for updating } \hat{\mathscr{E}}^t \text{ (i.e., } i \text{ s.t. } \mathscr{E}_i = \hat{\mathscr{E}}^t); \\ C_j[k]\, m_j(A)[k], & \text{otherwise,} \end{cases}$$

and $\alpha_i[k] + \sum_{j \in \overline{1,n}\setminus i} \sum_{A_{ij} \in \mathfrak{F}_j} \beta_{ij}(A_{ij}) = 1$, $\forall j \neq i$, $i \in \overline{1,n}$. ■

Theorem 2. *The iterative scheme in Definition 8 converges to a BoE $\mathscr{E}^*$, which we refer to as the* consensus BoE, *i.e., $\mathscr{E}_i[k] \to \mathscr{E}^*$, $\forall i \in \overline{1,n}$, as $k \to \infty$.* ■

Remarks:

1. It can be easily shown that the iterative scheme above generates a valid belief function $\mathrm{Bl}_i(.)[k], \forall i$, at each k. Further, it inherits many desirable properties from the CUE, e.g., robustness against contradictory evidence (see [7, 10] for details).
2. Theorem 2 can be established by using the convergence properties of paracontracting operators [5]. We omit the proof due to space limitations.
3. The scheme in Definition 8 updates a BoE with weighted sum of *conditionals* of the other BoEs. This agrees with the well-established *weighted average view* of consensus [3] and is also consistent with the CUE (for evidence updating).
4. The consensus BoE is guaranteed to be 'consistent' with the GT if/when an estimate $\hat{\mathscr{E}}^t$ of the GT is incorporated into the consensus process. For instance, if $\hat{\mathfrak{F}}^t = ab$, then $\mathfrak{F}^* \subseteq \{a, b, ab\}$.
5. The above parameter selection strategy combines the *cautious* and *receptive* strategies in [7, 10]: the cautious strategy applies to the estimate of the GT (i.e., $\mathscr{E}_i = \hat{\mathscr{E}}^t$); the receptive strategy applies to the other BoEs (i.e., $\mathscr{E}_i \neq \hat{\mathscr{E}}^t$).
6. The parameter $C_i \in (0,1)$ associated with each BoE $\mathscr{E}_i$ can be used to assign importance weights to BoEs. When such information is unavailable, take all C_i s to be equal.
7. In practice, the iterative scheme can be terminated either (a) at $k = K$ for some chosen K, or (b) when $dist(\mathscr{E}_i[k+1], \mathscr{E}_i[k]) \leq \varepsilon$, $i \in \overline{1,n}$, for some threshold $\varepsilon \geq 0$.

[2] $\mathfrak{E}$ is taken to contain an estimate of GT when it exists, i.e., if $\exists \hat{\mathscr{E}}^t$, then $\mathscr{E}_i = \hat{\mathscr{E}}^t$ for some $\mathscr{E}_i \in \mathfrak{E}$.

4 Numerical Example

Consider five (5) soft sources represented via the BoEs $\mathscr{E}_i$, $i \in \overline{1,5}$, with $\Theta_i \equiv \Theta = \{abcde\}$. Suppose their credibilities are unknown.

Setup: Suppose the BPAs are as follows:

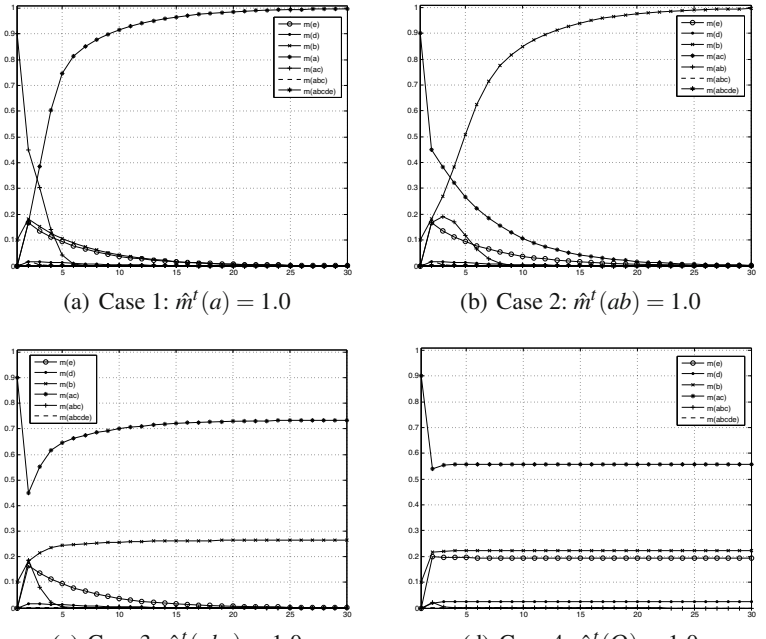
$$m_1(ac)=0.9;\ m_2(b)\ =0.9;\ m_3(ac)=0.9;\ m_4(ac)=0.9;\ m_5(e)\ =0.9;$$
$$m_1(b)\ =0.1;\ m_2(abc)=0.1;\ m_3(e)\ =0.1;\ m_4(d)\ =0.1;\ m_5(abc)=0.1.$$

We consider four cases (in decreasing order of 'preciseness' of the GT estimate):

Case 1	**Case 2**	**Case 3**	**Case 4**
$\hat{m}^t(a)=1.0;$	$\hat{m}^t(ab)=1.0;$	$\hat{m}^t(abc)=1.0;$	$\hat{m}^t(\Theta)=1.0.$

In Case 1, GT is known; in Cases 2-3, only an estimate of the GT is known; and in Case 4, the GT is completely unknown. For each case, all the five BoEs $\mathscr{E}_i$, $i \in \overline{1,5}$, reach the following consensus BoE:

Case 1: $m^*(a) = 1.00$
Case 2: $m^*(b) = 1.00$
Case 3: $m^*(b) = 0.29 \quad m^*(ac) = 0.71$
Case 4: $m^*(b) = 0.30 \quad m^*(ac) = 0.66 \quad m^*(d) = 0.02 \quad m^*(e) = 0.02$



(a) Case 1: $\hat{m}^t(a) = 1.0$ (b) Case 2: $\hat{m}^t(ab) = 1.0$

(c) Case 3: $\hat{m}^t(abc) = 1.0$ (d) Case 4: $\hat{m}^t(\Theta) = 1.0$

Fig. 2 Convergence of $\mathscr{E}_1$ to $\mathscr{E}^*$ as indicated by the the evolution of the BPA with k. All the focal elements that are not contained in the core of the estimated GT $\hat{\mathscr{E}}^t$ vanish as $\mathscr{E}_1$ reaches $\mathscr{E}^*$. This is exactly what has been referred to as being 'consistent' with $\hat{\mathscr{E}}^t$ in **(P3)**.

Table 1 Estimated credibility measures of the BoEs.

Method	Credibility																			
	Case 1. $\hat{m}^t(a)=1.0$					**Case 2.** $\hat{m}^t(ab)=1.0$					**Case 3.** $\hat{m}^t(abc)=1.0$					**Case 4.** $\hat{m}^t(\Theta)=1.0$				
	$\mathscr{E}_1$	$\mathscr{E}_2$	$\mathscr{E}_3$	$\mathscr{E}_4$	$\mathscr{E}_5$	$\mathscr{E}_1$	$\mathscr{E}_2$	$\mathscr{E}_3$	$\mathscr{E}_4$	$\mathscr{E}_5$	$\mathscr{E}_1$	$\mathscr{E}_2$	$\mathscr{E}_3$	$\mathscr{E}_4$	$\mathscr{E}_5$	$\mathscr{E}_1$	$\mathscr{E}_2$	$\mathscr{E}_3$	$\mathscr{E}_4$	$\mathscr{E}_5$
Cr_{cf1}	0.32	0.05	0.32	0.32	0.06	0.47	0.17	0.46	0.45	0.12	0.51	0.15	0.51	0.50	0.13	0.52	0.12	0.52	0.51	0.14
	5	1	5	5	2	5	2	4	3	1	5	2	5	3	1	5	1	5	3	2
Cr_{cf2}	0.32	0.05	0.32	0.32	0.06	0.32	0.05	0.32	0.32	0.06	0.90	0.07	0.90	0.90	0.08	0.90	0.07	0.90	0.90	0.08
	5	1	5	5	2	5	1	5	5	2	5	1	5	5	2	5	1	5	5	2
Cr_{con}	0.32	0.05	0.32	0.32	0.06	0.10	0.92	0.05	0.05	0.06	0.83	0.33	0.77	0.77	0.19	0.71	0.37	0.70	0.67	0.38
	5	1	5	5	2	4	5	2	2	3	5	2	4	4	1	5	1	4	3	2

Figure 2 shows the convergence of BoE $\mathscr{E}_1$ to $\mathscr{E}^*$ for each case. Note how the consensus BoE is 'consistent' or 'agrees' with $\mathscr{E}^t$. Behavior of other BoEs are similar and converge to $\mathscr{E}^*$ in each case.

Credibility Estimation: We now use the consensus BoE $\mathscr{E}^*$ in place of $\mathscr{E}^t$ in Definition 7 to get the credibility estimates Cr_{con} for each BoE. See Table 1 which also shows the two measures Cr_{cf1} and Cr_{cf2} in Definition 6. Ranked credibility values (lowest is '1') are also indicated underneath each credibility value in Table 1.

In Case 1, not surprisingly, all measures produce identical results. In Case 2, the assignment of a low credibility to $\mathscr{E}_2$ (supporting proposition b) by both Cr_{cf1} and Cr_{cf2} is surprising when GT is either a or b. Cr_{con} assigns a significantly higher credibility to $\mathscr{E}_2$ relative to other BoEs. The assignment of low credibility to $\mathscr{E}_1, \mathscr{E}_3, \mathscr{E}_4$ (mainly supporting a or b) also needs further investigation. The comparison is more difficult with decreasing exactness of the GT estimate, but they all seem to agree. Cases 3-4 are illuminating: **(a)** consensus BoE allocates higher supports for ac, which is in concordance with what a cursory glance at the BoEs reveals; **(b)** a is absent in the consensus because no BoE supports the singleton a; **(c)** d and e are absent in Case 3 consensus because they are absent in the GT estimate.

5 Concluding Remarks

The proposed consensus-based credibility estimation strategy addresses an important research question: *how can we validate evidence when the GT is unknown?* The proposed strategy can be used for purposes of (a) estimating the source reliability, (b) weighting the sources for fusion, or (c) discounting the BoEs.

The iterative process of consensus generation mimics rational agreement via the exchange of evidence among sources, and credibilities are assessed with respect to this consensus BoE. This is the major difference with conflict-based methods which are heavily dependent on the combination rule utilized[3].

However, the conflict-based measures may be computationally more efficient than the consensus-based approach. This important issue warrants further

[3] For example, when the Cr_{cf2} employs the DCR, evidence conflicts can generate null results. To avoid this, we set $m_i(\Theta)=0.0001$ (and deducted 0.0001 from the largest mass).

investigation. We plan to study error bounds so that an 'optimal' number of iterations could be determined. We are also exploring how the conditional core theorem [9] and approximation techniques can be exploited to reduce computational cost.

Acknowledgements. This work is based on research supported by the US Office of Naval Research (ONR) via grants #N00014-10-1-0140 and #N00014-11-1-0493 and the US National Science Foundation (NSF) via grant #1038257.

References

1. Fagin, R., Halpern, J.Y.: A new approach to updating beliefs. In: Bonissone, P.P., Henrion, M., Kanal, L.N., Lemmer, J.F. (eds.) Proc. Conference on Uncertainty in Artificial Intelligence (UAI 1991), pp. 347–374. Elsevier Science, New York (1991)
2. Jousselme, A.L., Grenier, D., Bosse, E.: A new distance between two bodies of evidence. Information Fusion 2(2), 91–101 (2001)
3. Lehrer, K., Wagner, C.: Rational Consensus in Science and Society. Philosophical studies series in philosophy. D. Reidel Pub. Co. (1981)
4. Martin, A., Jousselme, A.L., Osswald, C.: Conflict measure for the discounting operation on belief functions. In: 11th International Conference on Information Fusion 2008 (2008)
5. Pott, M.: On the convergence of asynchronous iteration methods for nonlinear paracontractions and consistent linear systems. Linear Algebra and its Applications 283(13), 1–33 (1998)
6. Pravia, M., Prasanth, R., Arambel, P., Sidner, C., Chong, C.Y.: Generation of a fundamental data set for hard/soft information fusion. In: 11th International Conference on Information Fusion 2008, pp. 1–8 (2008)
7. Premaratne, K., Murthi, M.N., Zhang, J., Scheutz, M., Bauer, P.H.: A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. In: Proc. International Conference on Information Fusion (ICIF 2009), Seattle, WA, pp. 2122–2129 (2009)
8. Webster, M.: Merriam Websters Collegiate Dictionary. Merriam Webster (2003)
9. Wickramarathne, T.L., Premaratne, K., Murthi, M.N.: Focal elements generated by the Dempster-Shafer theoretic conditionals: A complete characterization. In: Proc. International Conference on Information Fusion (ICIF 2010), Scotland, UK (2010)
10. Wickramarathne, T.L., Premaratne, K., Murthi, M.N., Scheutz, M.: A Dempster-Shafer theoretic evidence updating strategy for non-identical frames of discernment. In: Proc. Workshop on the Theory of Belief Functions (WTBF 2010), Brest, France (2010)

Ranking from Pairwise Comparisons in the Belief Functions Framework

Marie-Hélène Masson and Thierry Denœux

Abstract. The problem of deriving a binary relation over alternatives based on paired comparisons is studied. The problem is tackled in the framework of belief functions, which is well-suited to model and manipulate partial and uncertain information. Starting from the work of Tritchler and Lockwood [8], the paper proposes a general model of mass allocation and combination, and shows how to practically derive a complete or a partial ranking of the alternatives. A small example is provided as an illustration.

1 Introduction

The aim of the paper is to study the task of constructing a linear order, or a ranking, of n alternatives, based on paired comparisons. Paired experiments consist in presenting two objects to one or several judges and asking them to choose the best alternative among the pair. Each paired comparison is supposed to provide uncertain pieces of evidence on the ranking relation, and the derivation of a linear order is considered as an information fusion problem. Uncertain possibilistic preferences have been already considered e.g. in [2]. In this paper, the problem is tackled in the framework of belief functions.

A first work using belief functions to describe the uncertainty about the comparisons has been proposed in [8]. Unfortunately, this work remains essentially theoretical and gives very few tools for practical applications. Our paper synthesizes their main results in Section 3 and extends them in Section 4 in three ways: a more general model of mass allocation is proposed, a linear programming approach for determining the most plausible ranking is introduced, and a heuristic procedure for choosing only a partial order, starting from the most plausible ranking, is given.

Marie-Hélène Masson
UPJV, Heudiasyc, UMR CNRS 6599 BP 20529, 60205 Compiègne, France
e-mail: mmasson@hds.utc.fr

Thierry Denœux
UTC, Heudiasyc, UMR CNRS 6599 BP 20529, 60205 Compiègne, France
e-mail: thierry.denoeux@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 311–318.
springerlink.com

A small example in Section 5 illustrates the proposed method. Note that, due to space limitations, basic knowledge on belief functions will be assumed. The reader is referred, in particular, to [3, 6].

2 Basic Notions on Relations

Let $O = \{o_1, o_2, ..., o_n\}$ be a set of n alternatives. We recall that a strict order R on O is a binary relation for which the following properties hold for all o_i, o_j and $o_k \in O$:

- if $(o_i, o_j) \in R$ then $(o_j, o_i) \notin R$ (asymmetry);
- if $(o_i, o_j) \in R$ and $(o_j, o_k) \in R$ then $(o_i, o_k) \in R$ (transitivity);

If the order is complete (either $(o_i, o_j) \in R$ or $(o_j, o_i) \in R$), it is a linear (or total) order, otherwise it is a partial order. If $(o_i, o_j) \in R$ or $(o_j, o_i) \in R$ then o_i and o_j are comparable, otherwise they are said incomparable.

A linear order L is called a linear extension of a partial order P if $P \subseteq L$ ($\forall$ $(o_i, o_j) \in P$, then $(o_i, o_j) \in L$). To each partial order can thus be associated the set of its linear extensions. Conversely, any collection C of total orders defines a partial order H as follows: $(o_i, o_j) \in H$ iff (o_i, o_j) belongs to all linear order in C. One then says that H is realized by C. Note that two subsets can realize the same partial order.

Any relation R can be conveniently represented by a directed graph with nodes O. Two nodes o_i and o_j are connected by an arc in the graph if $(o_i, o_j) \in R$.

3 Pairwise Comparisons in the Framework of Belief Functions

Combining pairwise comparisons in the framework of belief functions has been already addressed by Tritchler and Lockwood [8]. This section follows their presentation and synthetizes the most useful notions. They consider that, for each pair (o_i, o_j) of alternatives in O ($1 \leq i < j \leq n$), an expert expresses its preference between o_i and o_j using a mass function $m^{\Theta_{ij}}$ quantifying the uncertainty in the evaluation. This mass function is defined on the frame of discernement $\Theta_{ij} = \{o_i \succ o_j, o_j \succ o_i\}$: the singleton $o_i \succ o_j$ means that o_i should be ranked before o_j and the singleton $o_j \succ o_i$ that o_j should be ranked first. Tritchler and Lockwood propose to use a simple support mass function: the expert chooses one of the singletons with mass α_{ij} and the rest of the mass is allocated to Θ_{ij}. The value α_{ij} is interpreted as the reliability of the choice.

Let ϕ_{ij} denote a focal element of $m^{\Theta_{ij}}$. Each focal element ϕ_{ij} has a graph representation which consists of two nodes, o_i and o_j, with one arc if ϕ_{ij} is a singleton element, and no arc if the focal element is Θ_{ij}.

The problem is to derive from the $n(n-1)/2$ mass functions a ranking of the alternatives. This task may be seen as an information fusion problem and Dempster's rule of combination [3] can be used to this end. Let I denote the set $\{(i, j) \mid 1 \leq i < j \leq n\}$ and let $\Theta(I)$ denote the product space:

$$\Theta(I) = \Theta_{12} \times \Theta_{13} \times ... \times \Theta_{(n-1)n}.$$

$\Theta(I)$ consists of all complete asymmetric relations (or graphs) defined on the set O. Before being combined, pieces of evidence from all pairs have to be expressed on the same frame of discernement, namely the product space $\Theta(I)$. This is achieved by applying the vacuous extension operation [3, 5] to each $m^{\Theta_{ij}}$. This operation, denoted $\uparrow$, transfers each mass $m^{\Theta_{ij}}(\phi_{ij})$ to $\phi_{ij} \times \Theta(I - \{(i,j)\})$. The symbol $\oplus$ representing Dempster's rule of combination, the expression of the combination can thus be formally written as:

$$m^{\Theta(I)} = m^{\Theta_{12}\uparrow\Theta(I)} \oplus m^{\Theta_{13}\uparrow\Theta(I)} \oplus \ldots \oplus m^{\Theta_{(n-1)n}\uparrow\Theta(I)}, \tag{1}$$

or, using the commonalities:

$$\mathrm{q}^{\Theta(I)} = \prod_{(i,j)\in I} \mathrm{q}^{\Theta_{ij}\uparrow\Theta(I)}. \tag{2}$$

The focal elements of $m^{\Theta(I)}$ are of the form: $\phi = \phi_{12} \times \phi_{13} \times \ldots \times \phi_{(n-1)n}$, where ϕ_{ij} is a focal element of $m^{\Theta_{ij}}$ and the mass resulting from the combination is:

$$m^{\Theta(I)}(\phi) = m^{\Theta_{12}}(\phi_{12}) m^{\Theta_{13}}(\phi_{13}) \ldots m^{\Theta_{(n-1)n}}(\phi_{(n-1)n}). \tag{3}$$

In terms of graph, each focal element ϕ of $m^{\Theta(I)}$ can be represented by a directed graph formed by the union of individual graphs. Since each ϕ_{ij} is equal either to a singleton or to Θ_{ij}, each focal element ϕ is a subset composed of complete asymmetric relations on O, whose graphs contain the arcs of ϕ.

The combination described above allocates masses on various sets of asymmetric relations defined on O. A first objective is to find a linear ordering on O that is the most compatible with the pairwise evaluations. This can be done by imposing conditions on the set in which the solution has to be found. Let $\mathscr{L}$ denote the set of all linear orders defined on O which is a subset of $\Theta(I)$. To impose the nature of the solution, it is proposed in [8] to condition the mass $m^{\Theta(I)}$ with respect to $\mathscr{L}$:

$$m^{\Theta(I)}[\mathscr{L}] = m^{\Theta(I)} \oplus m_{\mathscr{L}}, \tag{4}$$

with $m_{\mathscr{L}}$ a categorical mass function defined by $m_{\mathscr{L}}(\mathscr{L}) = 1$.

Expressed using the commonalities, the whole combination can be written as:

$$\mathrm{q}^{\Theta(I)}[\mathscr{L}] = \frac{1}{1-K}\, \mathrm{q}_{\mathscr{L}} \prod_{(i,j)\in I} \mathrm{q}^{\Theta_{ij}\uparrow\Theta(I)}, \tag{5}$$

where K is the conflict resulting from the combination of $m^{\Theta(I)}$ with $m_{\mathscr{L}}$. K can be interpreted as an index of the internal coherence of the evaluations. Its practical computation will be explained when dealing with partial orders.

4 Practical Use

We consider in this section a general form of mass allocation defined by:

$$\begin{cases} m^{\Theta_{ij}}(o_i \succ o_j) = \alpha_{ij}, \\ m^{\Theta_{ij}}(o_j \succ o_i) = \beta_{ij}, \\ m^{\Theta_{ij}}(\Theta_{ij}) = 1 - \alpha_{ij} - \beta_{ij}. \end{cases} \tag{6}$$

This mass allocation may come from a single expert who is asked to provide, for each $(i,j) \in I$, the above mass function, or from the combination of the evaluations of several experts. In that case, for each $(i,j) \in I$, several $m_k^{\Theta_{ij}}$ are available and they have to be fused to provide $m^{\Theta_{ij}}$. The choice of the combination rule depends on the hypotheses made on the dependence between the experts. If they can be considered as independent, Dempster's rule should be chosen. Otherwise, the cautious rule [1] may be preferred.

4.1 Most Plausible Ranking

Let $L \in \mathscr{L}$ be a strict linear ordering on O. L being a singleton of the frame of discernement, one has $\mathrm{q}^{\mathscr{L}}(\{L\}) = 1$ and, with the mass allocation (6), one has:

$$\begin{cases} \mathrm{q}^{\Theta_{ij}\uparrow\Theta(I)}(\{L\}) = 1 - \beta_{ij} & \text{if } (o_i, o_j) \in L, \\ \mathrm{q}^{\Theta_{ij}\uparrow\Theta(I)}(\{L\}) = 1 - \alpha_{ij} & \text{if } (o_j, o_i) \in L. \end{cases} \tag{7}$$

Let us introduce $n(n-1)/2$ binary variables $l_{ij}((i,j) \in I)$ defined by $l_{ij} = 1$ if $(o_i, o_j) \in L$ and 0 otherwise. Using (7) and (5), the commonality, or, equivalently, the plausiblity of L can be written as:

$$\mathrm{q}^{\Theta(I)}[\mathscr{L}](\{L\}) = \mathrm{pl}^{\Theta(I)}[\mathscr{L}](\{L\}) = \frac{1}{1-K} \prod_{(i,j)\in I} (1-\beta_{ij})^{l_{ij}} (1-\alpha_{ij})^{1-l_{ij}}, \tag{8}$$

where K is the conflict resulting from the combination of $m^{\Theta(I)}$ with $m^{\mathscr{L}}$. To find the most plausible ranking of the alternatives, it is not necessary to enumerate all possible linear orderings. We propose to solve the problem using a linear programming approach. Maximizing expression (8) is equivalent to maximize its logarithm so that, omitting the constant term depending on K, the most plausible ranking L can be found as the solution of the following linear program:

$$\max_{l_{ij}\in\{0,1\}} \sum_{(i,j)\in I} l_{ij} \ln\left(\frac{1-\beta_{ij}}{1-\alpha_{ij}}\right), \tag{9}$$

subject to:

$$\begin{cases} l_{ij} + l_{jk} - 1 \leq l_{ik}, & \forall i < j < k, \\ l_{ik} \leq l_{ij} + l_{jk}, & \forall i < j < k. \end{cases} \tag{10}$$

The constraints are used to insure that L belongs to $\mathscr{L}$: if $l_{ij} = 1$ and $l_{jk} = 1$ then $l_{ik} = 1$. If $l_{ij} = 0$ and $l_{jk} = 0$ then $l_{ik} = 0$.

Remark 1. *Note that the general form of mass allocation (6) allows us to take naturally into account tied evaluations. If the comparison between o_i and o_j results in a tie, we let $\alpha_{ij} = \beta_{ij}$. Then, it can be easily seen that the pair (o_i, o_j) does not appear any more in the objective function.*

4.2 Plausibility of a Partial Ranking

In some situations, it may be also interesting to compute the plausiblity of a partial order. When working in the set of asymmetric relations as defined by Tritchler and Lockwood, it not possible to provide an analytical expression. However, some results from [8] make it possible to use simple algorithms based on graph theory to compute the plausibility of any partial order. The following theorem is proved in [8]:

Theorem 1 (Tritchler and Lockwood (1991)). *Let K be the conflict between the two mass functions $m^{\Theta(I)}$ and $m^{\mathscr{L}}$.*

1. *$K = \sum m^{\Theta(I)}(\phi)$ where the summation is over every focal element of $m^{\Theta(I)}$ whose graph contains a cycle;*
2. *Let H be a partial order realized by a focal element θ_H of $m^{\Theta(I)}[\mathscr{L}]$. Then θ_H is the set of all linear extensions of H (θ_H is the largest subset of $\mathscr{L}$ which realizes H).*
3. *$m^{\Theta(I)}[\mathscr{L}](\theta_H) = \frac{1}{1-K} \sum m^{\Theta(I)}(\phi)$, where the summation is over every focal element ϕ of $m^{\Theta(I)}$ such that the transitive closure of $G(\phi)$, $G^t(\phi)$, is equal to H.*

To compute the plausiblity of any partial order, one has to sum the masses associated to all focal elements with a non null intersection with this partial order. The following lemma helps to recognize the focal elements which intersect a given partial order:

Lemma 1 (Tritchler and Lockwood (1991)). *Let C_1 and C_2 be two subsets of $\mathscr{L}$ realizing the partial orders H_1 and H_2. Then $C_1 \cap C_2 \neq \emptyset$ if and only if $H_1 \cup H_2$ is acyclic.*

Computing the plausibility of any partial order H is thus achieved by summing the masses of all focal sets $\theta_{H'}$ such that $H' \cup H$ is acyclic. .

Lemma 1 and Theorem 1 allow us to propose two simple procedures (one exact, one approximate) for computing the plausibility of a given partial order. The two procedures are detailed in Algorithms 1 and 2.

4.3 Heuristic Search for a Partial Ranking

If the plausibility of the most plausible ranking is too low, it can be preferable to provide the user with only a partial ranking of the alternatives. The algorithms

Algorithm 1. Plausibility of a partial order H

$K \leftarrow 0$
$\text{pl}(H) \leftarrow 0$
for all $\phi = \phi_{12} \times \phi_{13} \times \ldots \times \phi_{(n-1)n}$ **do**
    Compute the transitive closure $G^t(\phi)$ of $G(\phi)$
    Compute the mass $m = m^{\Theta(I)}(\phi)$ by equation (3)
    if $G^t(\phi)$ contains a cycle **then**
        $K = K + m$
    else if $G(H) \cup G^t(\phi)$ is acyclic **then** $\text{pl}(H) = \text{pl}(H) + m$
    end if
end for
$\text{pl}(H) \leftarrow \frac{1}{1-K}\text{pl}(H)$

Algorithm 2. Approximate computation by a Monte-Carlo approach

$N_{\text{pl}} \leftarrow 0$
$N_K \leftarrow 0$
for $rep \leftarrow 1, N$ **do**
    $G \leftarrow \emptyset$
    for each $(i,j) \in I$ **do**
        With probability $m^{\Theta_{ij}}(\phi_{ij})$, randomly select a focal element ϕ_{ij} from the focal elements of $m^{\Theta_{ij}}$
        if ϕ_{ij} is a singleton, add the corresponding arc to G
    end for
    Compute the transitive closure G^t of G
    if G^t contains a cycle **then**
        $N_K = N_K + 1$
    else if $G(H) \cup G^t$ is acyclic **then** $N_{\text{pl}} = N_{\text{pl}} + 1$
    end if
end for
$\hat{K} = \dfrac{N_K}{N}$
$\hat{\text{pl}}(H) = \dfrac{1}{1-\hat{K}} \dfrac{N_{\text{pl}}}{N}$

described in the previous section allow us to compute the plausibility of any partial order. Instead of exploring every possible partial orders, which would be practically intractable, we propose a heuristic procedure based on a principle of hierarchical clustering. We start from the most plausible ranking (see Section 4.1). Then, at the first step of the procedure, we compute the plausibility of every partial orders obtained by removing one pair of adjacent alternatives from the total order relation. The most plausible partial order is retained and the corresponding pair of alternatives is "merged". The process is repeated until all alternatives have been merged into a single one. It is easy to see that the sequence of plausibility values is monotonically increasing. Finally, a partial order with the desired level of plausibility can be chosen by the user.

5 Example

We illustrate the methods described above using an example inspired from [8]. In a study conducted at the Ontario Cancer Institute, subjects were asked to give their preferences about four scenarios describing ethical dilemmas in health care. The preferences for all six possible scenario pairs were obtained. The experts were also asked to rate the reliability of their evaluations. The preferences of a subject can be represented by a directed graph in which the vertices are the scenarios and the edges represent the relation "is preferred to". The values on the edges represent the reliability given by the expert. The graphs of the experts are given in Figure 1. The fact that graph 1 (left) contains a cycle (ADB) shows that the evaluations of expert 1 are not fully consistent. There is no cycle in graph 2 (right), but the degrees of belief are weaker than for expert 1. The evaluations of each expert are modelled using the mass allocation expressed by equation (6) with either α_{ij} or β_{ij} equal to zero.

Applied individually to each expert, the procedure for deriving a complete ranking of the alternatives (Section 4.1) gives the ranking $A \succ D \succ B \succ C$ with a plausibility of 0.8070 for the first expert, and the ranking $A \succ C \succ D \succ B$ with a plausibility equal to 1 for the second one. The plausiblities thus reflect the internal coherence of the experts. The evaluations of the experts can also be combined before searching for a complete ranking. The results of the combination using Dempster's rule of combination can be found in Table 1.

The most plausible total ranking derived from Table 1 is $A \succ D \succ B \succ C$ with a plausibility equal to 0.8893. Applying the heuristic procedure using the masses of Table 1 for determining a partial ranking gives the result presented in Figure 2. The dendrogram can be cut at the desired level of plausibility. For example, the partial order $A \succ D \succ \{B,C\}$ reaches a plausibility of almost 0.96.

Fig. 1 Graph representation of the evaluations; (left) : expert 1 ; (right) : expert 2.

Table 1 Mass assignment using Dempster's rule of combination

(o_i, o_j)	$o_i \succ o_j$	$o_j \succ o_i$	Θ_{ij}
(A,B)	0.3056	0.3056	0.3889
(A,C)	0.9991	0	0.0009
(A,D)	0.9964	0	0.0036
(B,C)	0.7266	0.2187	0.0547
(B,D)	0	0.9324	0.0676
(C,D)	0.0594	0.0094	0.9312

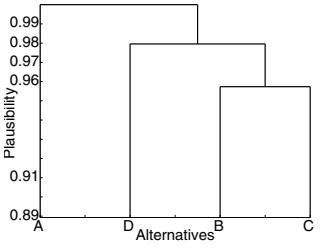


Fig. 2 Dendrogram obtained from Table 1.

6 Conclusion

In this paper, we have shown how to use the framework of belief functions to model paired comparisons and how to derive from these individual judgements a total or a partial ranking of the alternatives. The linear order is obtained by solving a linear program maximizing the plausibility of the relation. A heuristic procedure has been proposed to provide only a partial order when the plausibility of the linear order is too low. This work offers several perspectives, among which the application of the approach to machine learning problems like instance or label ranking problems.

References

1. Denoeux, T.: Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence. Artificial Intelligence 172, 234–264 (2008)
2. Dubois, D., Prade, H.: On the ranking of ill-known values in possibility theory. Fuzzy sets and Systems 43, 311–317 (1991)
3. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
4. Smets, P.: The combination of evidence in the Transferable Belief Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)
5. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9, 1–35 (1993)
6. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence 66, 191–243 (1994)
7. Smets, P.: The Transferable Belief Model for quantified belief representation. In: Gabbay, D.M., Smets, P. (eds.) Hanbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, pp. 267–301. Kluwer Academic Publishers, Dordrecht (1998)
8. Tritchler, D., Lockwood, G.: Modelling the reliability of paired comparisons. Journal of Mathematical Psychology 35, 277–293 (1991)

Dempster-Shafer Fusion of Context Sources for Pedestrian Recognition

Magdalena Szczot, Otto Löhlein, and Günther Palm

Abstract. This contribution presents the design of an image-based contextual pedestrian classifier for an automotive application. Our previous work shows that local classifiers working with image cutouts are in many cases not sufficient to achieve satisfactory results in complex scenarios. As a solution the work proposed incorporating contextual knowledge into the classification task, significantly improving the classification results. Contextual knowledge is described by a set of different and independent context sources. This paper discusses the fusion of these sources on the basis of the Dempster-Shafer theory. It presents and compares different possibilities to model the frame of discernment and the mass function to achieve optimal results. Furthermore, it provides an elegant way to take uncertainties of the context sources into account. The methods are evaluated on simulated and on real data.

1 Introduction

Recent studies in the field of driver assistance systems concern themselves with the task of detecting pedestrians in front of a vehicle. This task is usually solved by applying a local classifier to the camera images. Such classifiers regard only local image cutouts on different positions and scales and decide for each cutout

Magdalena Szczot
University of Ulm, Germany
e-mail: magdalena.szczot@uni-ulm.de

Otto Löhlein
Daimler AG, Ulm, Germany,
Department Environment Perception (GR/PAP)
e-mail: otto.loehlein@daimler.com

Günther Palm
University of Ulm, Germany,
Institute of Neural Information Processing
e-mail: guenther.palm@uni-ulm.de

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 319–326.
springerlink.com

separately whether it contains a pedestrian or not. Our previous work [5] shows the disadvantages of such an approach and greatly improves the results by incorporating contextual knowledge into the classification task. Such knowledge is hereby modeled as a set of many different hints (*context sources*) which describe the relation between a pedestrian and his surroundings. The actual challenge is finding an elegant way of describing those hints in a unified manner ([5], [4]) and fusing them into a single classifier. This contribution concentrates on the fusion of context sources with the Dempster-Shafer theory and shows the benefits of this method in comparison to the usual Bayesian approach. It discusses different possibilities for choosing the frame of discernment and modeling the mass function. Furthermore, it describes an effective way of representing the uncertainty of one context source for one pedestrian detection. Finally, the paper compares the different methods on a real data set.

For a comprehensive overview of the Dempster-Shafer theory, we would like to refer to the work by Smets [3].

2 Application Setup

The goal of the system is a robust recognition of pedestrians in camera images. Its first component is a Viola-Jones cascaded classifier [6], which for one camera image delivers a list of detections in the form of bounding boxes. This list of detections is the basis for any further processing. Furthermore, all evaluation results are given in relation to this list of detections, which can be either true positive (TP) or false positive (FP) pedestrian detections. This means that if the context classifier accepts all pedestrians previously detected by the cascade, the detection rate would be 1, even though the whole system might still have overlooked some of the pedestrians. Figure 1 shows an example of a false positive and a true positive detection.

The second component of the application is given by the context sources. All context sources share a common model which delivers a foundation for the fusion algorithms.

Fig. 1 An example of a true positive (green) and a false positive (red) detection of the cascaded classifier.

One context source represents an arbitrary piece of information which is considered to be relevant for the classification task. Each context source gets one detection D of a cascaded classifier as an input and computes one value $q(D) \in \mathbb{R}$. The computation of $q(D)$ depends on the sort of the context source. For example, if one wants to consider the position of the horizon in the image, than $q(D)$ might be defined as a normalized distance of the bottom line of the detection box to the horizon line. Additionally, each source holds two histograms over the outputs $q(D)$ of its algorithm over the training data: one for the positive and one for the negative samples. Let $\Theta^{TP} = (\Theta_1^{TP} \dots \Theta_N^{TP})$ denote the bin boundaries of the histogram over the set of true positive detections I^{TP}. The number of entries in the true positive histogram bins is then defined as:

$$V_j^{TP} = |\left\{q(D)|\Theta_j^{TP} \leq q(D) < \Theta_{j+1}^{TP}\right\}|. \tag{1}$$

The frequencies V_j^{FP} of the histogram over false positives FP are computed in the same manner. This simple representation of each contextual information as a pair of histograms together with the according algorithm for the computation of $q(D)$ allows modeling any arbitrary information source and is the foundation for the fusion algorithms presented in this contribution. The basic idea is to use the quotient $\pi = V_j^{TP}/(V_j^{TP} + V_j^{FP})$ for a detection with $q(D)$ in bin j as an estimate for $p(\text{pedestrian}|q(D))$. One possibility then is to use the naive Bayes method for the combination of these probabilities. However, this would not account for the uncertainty of these probability estimates. For this reason we propose the use of the Dempster-Shafer theory and in particular of Dempster's rule of combination ([3]).

3 Dempster-Shafer Theory for Fusion of Context Sources

There are two main design decisions to be made for a practical application of the Dempster-Shafer theory: the first one concerns the frame of discernment and the definition of the single hypotheses. The second decision regards the mass function of the sources over the propositions from 2^{Ω}.

This work presents two different approaches concerning the choice of the frame of discernment as well as the distribution of mass. The first discrimination between the methods is given by the choice of Ω. Following the most common approaches in the literature, the first method in this paper divides the frame of discernment into the proposition *pedestrian* (F) and *background* ($\neg F$) (i.e. $\Omega = \{F, \neg F, \}$). This frame of discernment will be called *symbolic*. As an alternative, this paper suggests choosing a *real-valued* frame of discernment. Specifically, we define the frame of discernment as the interval between zero and one ($\Omega = [0,1]$), where the focal elements are nested intervals in Ω.

3.1 Modeling the Uncertainty of One Context Source

Both methods share a common model of the uncertainty of a context source. This uncertainty determines the assignment of mass to the propositions. The foundation of this is provided by the histograms of the context sources.

Let both histograms be defined as described in Section 2. Let j be the index of the histogram bin for the output of q for current detection $D : \Theta_j \leq q(D) < \Theta_{j+1}$. The number of entries in each bin is modeled by a binomially distributed random variable:

$$B(n = (V_j^{TP} + V_j^{FP}), \pi = V_j^{TP}/(V_j^{TP} + V_j^{FP})). \tag{2}$$

Here the number of successes matches the number of positive samples in the j-th histogram bin. The certainty with which on source makes a statement (π) is roughly proportional to the number of entries in the bin. To model this certainty we consider the conjugate distribution to the binomial distribution - the beta distribution ([1]) over π ($\pi \sim f_{\alpha,\beta}$). The number of entries in histogram bins are taken as parameters for the beta distribution: $\alpha = V_j^p$ and $\beta = V_j^n$. The smaller the number of entries, the larger the variance of the beta distribution.

3.2 Simulation Framework

For a better understanding of each presented approach, this work gives exemplary results on artificial sources from a simulation framework. The simulation consists of three contextual sources. The output of each source is modeled by two normal distributions (for positive and negative samples). During the simulation of the training step the outputs are drawn from those distributions in order to create the positive and negative histograms. In the evaluation step each context source delivers a random output which is used to estimate the according histogram bin and the parameters of the beta distribution.

3.3 Method 1: Symbolic Frame of Discernment $\Omega = \{F, \neg F, \}$

The first method handles propositions consisting of symbolic classes: pedestrian (F), background($\neg F$), and the union of the two classes ($\{F, \neg F, \}$).

After estimating the correct histogram bin of each context source for current detection, the according beta distribution parameters are known and given by the equations in the previous sections. The expectation value of this distribution should steer the mass assessment to the propositions (F) and ($\neg F$). Furthermore, the variance of the beta distribution should have some influence on the amount of mass assigned to the union of the single hypotheses ($\{F, \neg F, \}$). In order to meet these requirements, the first method regards the integral of the beta distribution over a certain interval. Let $\gamma \in [0, 1]$ and $c = f_{\alpha,\beta}([\pi - \gamma, \pi + \gamma])$ be the integral of the beta distribution around its expectation value. The value c describes the certainty of the output of the

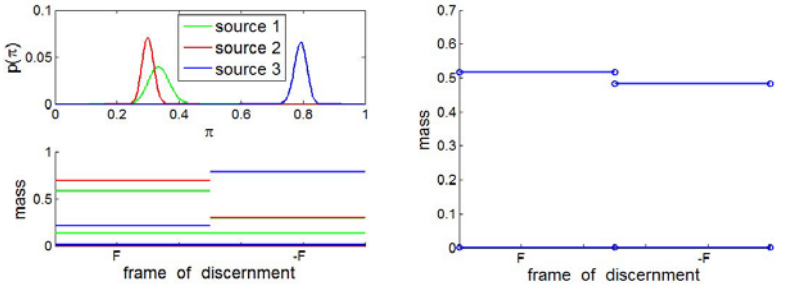


Fig. 2 Method 1. Result of a fusion of simulated context sources with $\Omega = \{F, \neg F, \}$. The figure on the left shows the three beta distributions and the mass function of each context source. The figure on the right shows the result of Dempster's rule of combination.

context source: the smaller the variance of the beta distribution, the larger c is. The value c is used as a scaling factor for the computation of the mass function.

$$m(F) = \pi \cdot c, \text{ and } m(\neg F) = (1 - \pi) \cdot c. \tag{3}$$

The remaining mass is assigned to the union of both classes ($m(\Omega) = 1 - c$). The estimation of an optimal γ is done by evaluating this method on a learnset and choosing the parameter which achieves the lowest error over all samples. Figure 2 shows the mass of the context sources and the result of the fusion for the simulation example. The presented method distributes the mass of one context source according to its posterior probability for a detection and its confidence derived from the beta distribution. The single hypotheses of the context sources are class names, and the uncertainty is modeled only in an indirect way by assigning mass to the union of both classes. The different context sources are combined by Dempster's rule of combination.

In contrast to this approach the second method assumes that the actual statement of a context source is its posterior probability π for a pedestrian. The uncertainty for this statement is given directly by the beta distribution $f_{\alpha,\beta}$ and we consider a different choice of the frame of discernment, namely $\Omega = [0,1]$ ([2]). Here one could use different combination methods, in particular different ways of handling the conflict (see [3]), but Dempster's rule works fine in our application.

3.4 Method 2: Real-Valued Frame of Discernment ($\Omega = [0,1]$)

The present realisation considers propositions which are nested intervals within Ω. The propositions B_i are taken to lie symmetrically around the expectation value of the beta distribution, as shown in figure 3. This works better than the more common use of disjoint intervals, because it reduces the conflict arising from the combination of the sources. The mass function is computed by integrating the beta-distribution

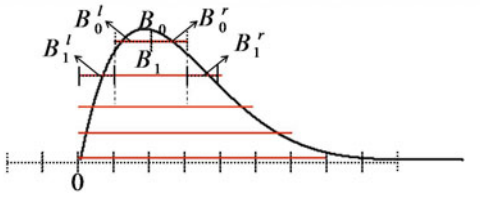


Fig. 3 Method 2. Choice of propositions over the frame of discernment. The propositions interleave and lie around the expectation value of the beta distribution.

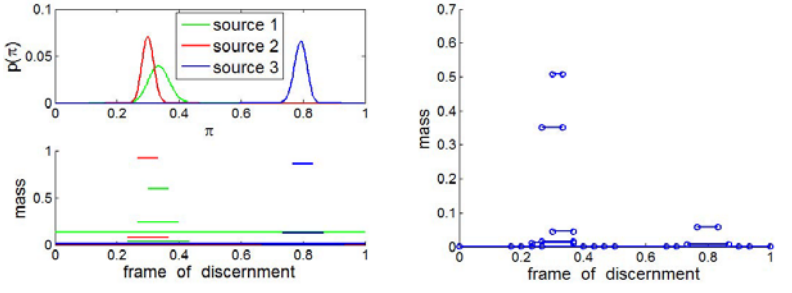


Fig. 4 Method 2. Result of the Dempster-Shafer Fusion for the interleaved focal propositions and $\Omega = [0,1]$. The figure on the left shows the beta distributions and the propositions. The result of the orthogonal sum is shown in the figure on the right.

over the differences $B_i \setminus B_{i-1} = B_i^l \cup B_i^r$ (see figure 3) and assigning the value of the integral to the nested focal elements:

$$m(B_i) = f_{\alpha,\beta}(B_i^l) + f_{\alpha,\beta}(B_i^r) \tag{4}$$

Figure 4 shows the result of the Dempster-Shafer fusion for a partitioning of Ω into 20 single intervals. In this particular example, the largest resulting mass is assigned to the intervals around 0.3 and 0.4, where the responses of the two first context sources agree. The third context source, whose posterior probability would be higher, does not have enough samples in the current bin, to overrule this decision.

For the evaluation of this method the actual output $O(D)$ of the Dempster-Shafer fusion for one detection is given by the center of the proposition with greatest mass:

$$O(D) = (x_{i+1} + x_i) \cdot 0.5 \text{ , for } i = \text{argmax}(m([x_0, x_1]), ..., m([x_{M-1}, x_M])). \tag{5}$$

4 Evaluation Results

This section presents the evaluation results for the application of the pedestrian recognition. As described above, the fusion algorithms are evaluated on the list of detections of a cascaded classifier. The cascade used consists of 30 layers. In order to gain more data (especially more false alarms) for this evaluation we regard all

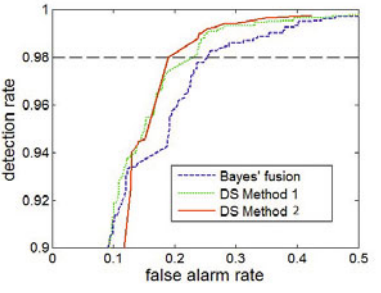


Fig. 5 ROC-curve for the Bayes' fusion and both methods from the Dempster-Shafer approach. Since the results are only relevant for high detection rates, the axis is limited to a detection rate between 0.9 and 1 for visualization purposes.

detections beginning with the 26th layer. In this way, the fusion methods are evaluated for 2900 labeled pedestrian occurrences and 900 false alarms. The detection rate is 1 if all pedestrians represented by the detections are accepted by the fusion system. Three context sources were used for the evaluation. The context sources regard the position of the street, the position of the horizon line in the image as well as the relative positions of objects in the image. A thorough description of these sources can be found in [5] and [4].

Figure 5 presents the detection and false alarm rates for the Bayes' fusion and the Dempster-Shafer fusion. The operating points of the ROC-curve are computed by comparing the output $O(D)$ of the fusion algorithm with increasing thresholds between 0 and 1. The results show an improvement achieved by incorporating uncertainties of the sources into the fusion framework. Furthermore, the usage of a real-valued frame of discernment leads to a further reduction of the false alarm rate in comparison to the second method with symbolic frame of discernment. For example, for a constant detection rate of 98% the Bayes' fusion achieves a false alarm rate of 25.2%, the Dempster-Shafer approach with symbolic Ω leads to a false alarm rate of 23.4% whereas the real-valued Ω achieves a false alarm rate of 18.9%.

5 Conclusion and Future Work

Many cognitive systems take advantage of a combination of diverse pieces of information. An optimal design of the fusion method for such systems is crucial in order to achieve satisfactory results. This contribution arises from the field of pedestrian recognition and is motivated by the desire to incorporate contextual hints into the classification process. As there are many different contextual hints, the architecture of a contextual classifier requires the fused response of all sources. A classical Bayes' approach delivers only suboptimal results since it ignores the uncertainty of the different sources. As a possible solution, we propose the use of Dempster-Shafer theory.

This paper discusses two possibilities to model the frame of discernment and the mass function. Specifically, it differentiates between a symbolic and a real-valued frame of discernment. The first method requires the sources to assign mass to symbolic class names. This is the most common approach and the evaluation shows that it already delivers better results than the Bayes' fusion. Further improvement is achieved by a real-valued frame of discernment directly containing intervals from $[0, 1]$ describing the posterior probabilities. This approach proposes an elegant way of modeling the mass functions within the Dempster-Shafer theory and incorporating the uncertainty of the evidences. The evaluation results show a significant improvement achieved through the introduction of the uncertainty into the fusion by applying the Dempster-Shafer theory.

References

1. Forbes, C., Evans, M., Hastings, N., Peacock, B.: Statistical Distributions. John Wiley & Sons (2011)
2. Ristic, B., Smets, P.: Belief function theory on the continuous space with an application to model based classification. In: Information Processing and Management of Uncertainty (2004)
3. Smets, P.: The combination of evidence in the transferable belief model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)
4. Szczot, M., Dannenmann, I., Lohlein, O.: Incorporating lane estimation as context source in pedestrian recognition task. In: ICPR (2010)
5. Szczot, M., Löhlein, O., Palm, G.: Incorporating contextual information in pedestrian recognition. In: The IEEE Intelligent Vehicles Symposium, IV (2009)
6. Viola, P., Jones, M.: Robust real-time object detection. In: Proceedings of IEEE Workshop on Statistical and Computational Theories of Vision (2001)

Multi-level Dempster-Shafer Speed Limit Assistant

Jérémie Daniel and Jean-Philippe Lauffenburger

Abstract. This paper deals with a Speed Limit Assistant (*SLA*) performing the fusion of a Geographic Information System (*GIS*) and a vision system. The present strategy is based on multi-level data fusion using Evidence Theory. In a first step, the *GIS* reliability is estimated through *GIS* criteria related to the positioning, the localization and the digital map resolution. Contextual criteria also extracted from the *GIS* define the belief masses of the speed candidates. Afterwards, a multi-criterion fusion is processed to detect potential *GIS* incoherences (difference between the *GIS* speed and the road context). The second fusion level (the multi-sensor fusion) then combines the *GIS* and vision information by considering these sensors as specialized sources. In order to manage the conflict, the Proportional Conflict redistribution Rule 5 (*PCR*5) has been chosen. The benefits of the proposed solution are shown through real experiments performed with a test vehicle.

1 Introduction

Speed Limit Assistants (*SLA*) usually refer to the combination of a Speed Limit Sign Recognition System (*SLSRS*) with a Geographical Information System (*GIS*) as they are complementary. Among the different techniques which can be used for their fusion, Evidence Theory [1, 2] showed its effectiveness. This formalism was for instance employed for *SLA*s by [3] and [4]. They proposed an approach in which the *GIS* information is processed through a combination of digital map database attributes. The selected attributes are of great help, on the one hand in the description of the current road context (giving information about context-dependent implicit speeds), and on the other hand to characterize the reliability of the *GIS*. However, in these works, no detection of the *GIS* incoherences and inaccuracies was performed

Jérémie Daniel · Jean-Philippe Lauffenburger
Modélisation Intelligence Processus Systèmes (MIPS) laboratory, 12 rue des frères Lumière, 68093 Mulhouse Cedex, France
e-mail: name.surname@uha.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 327–334.
springerlink.com

and a simple weighted sum was used to define the basic belief assignments (*bba*) w.r.t. the *GIS* criteria.

To overcome these limitations, the present *SLA* is based on a multi-level fusion approach. The first level - the multi-criterion fusion - consists in the evaluation of the navigation reliability while the second level - the multi-sensor fusion - fuses the vision and the *GIS* to define the final speed limit and its confidence. The main benefits of the multi-criterion fusion is the consideration of the *GIS* reliability (positioning, localization and digital map database quality) in its *bba*. In addition, this fusion helps to detect the *GIS* incoherences (contradiction between the speed indicated by the *GIS* and contextual data related to the driving situation) and to determine the appropriate navigation speed candidate through a local decision step. Both fusion levels are based on the *Rombaut*'s *bba* model [5] which considers the sensors as specialized sources [6]. The conflict which may be generated during this step is managed by the Proportional Conflict redistribution Rule 5 (*PCR5*) [7].

The paper is organized as follows: Section 2 describes the notations and the strategy adopted for speed limit determination. Section 3 presents the combination rules as well as the decision technique used to select the most relevant speed candidate considering the eventually generated conflict. Before the concluding remarks (Section 5), experimental results are highlighted in Section 4.

2 Combining Vision and Navigation Information

2.1 Notations

The discernment frame Θ used for the *SLA* contains the speeds S_j defined by the general legal driving rules. The corresponding referential subset 2^Θ is presented in (1) with k the number of possible speeds:

$$\Theta = \{S_1, S_2, ..., S_k\}, \qquad 2^\Theta = \{S_1, S_2, ..., S_k, \{S_1, S_2\}, ..., \{S_2, S_3\}, ..., \Theta\} \tag{1}$$

The model initiated by *Rombaut* [5] has been retained for the representation of the navigation and vision data. It considers the sources to be independent and specialized: a source can only give information about one specific speed (S_j) of the discernment frame. Moreover, the source can only say *"I believe in this speed"*, *"I do not believe in this speed"* or *"I do not know"*. Consequently, each source gives its opinion on the triplet $\left\{S_j, S_j^c, \Theta\right\}$ where $S_j^c = \left\{S_1, ..., S_{j-1}, S_{j+1}, ..., S_k\right\}$. In addition, this model is defined regarding a non-overlapping condition of the masses S_j and S_j^c. In the multi-criterion fusion, the sources are the different criteria extracted from the *GIS* (cf. Section 2.2). They give their opinion on a speed S_j of Θ through a *bba* m_j defined as follows:

$$\begin{aligned} m_j(S_j) &= \begin{cases} 0 & R_v \in [0,\tau] \\ \left(\frac{\alpha_j}{1-\tau}\right) R_v - \frac{\alpha_j \tau}{1-\tau} & R_v \in [\tau,1] \end{cases} \\ m_j\left(S_j^c\right) &= \begin{cases} -\frac{\alpha_j}{\tau} R_v + \alpha_j & R_v \in [0,\tau] \\ 0 & R_v \in [\tau,1] \end{cases} \\ m_j(\Theta) &= \begin{cases} \frac{\alpha_j}{\tau} R_v + (1-\alpha_j) & R_v \in [0,\tau] \\ -\left(\frac{\alpha_j}{1-\tau}\right) R_v + \frac{1-(1-\alpha_j)\tau}{1-\tau} & R_v \in [\tau,1] \end{cases} \end{aligned} \tag{2}$$

with τ the boundary value defining the limit between the belief in S_j and S_j^c, and R_v the reliability variable of the considered information source. α_j describes the level of coherence which links a speed S_j to a criteria through a maximum mass value. For example, a highway road coupled to a $30km.^{-1}$ speed, thus describing an incoherent situation, are linked by a low value of α_j. The multi-criteria fusion, combining *bba*s defined with different α_j, then helps to detect the incoherences of the *GIS* data.

For the vision system, the reliability R_{vis} is defined regarding the confidence in the detected speed while the navigation reliability R_{GIS} is defined regarding the accuracy of the positioning, localization and digital map database as described in Section 2.2. For briefness reasons, the model cannot be completely described here. Additional details are available in [8].

2.2 Multi-level Fusion and Navigation Reliability

The fusion strategy is based on the diagram presented in Fig.1 showing a two-level fusion. The first step consists in the local processing of the sensor data, i.e. the determination of their speed candidates and the related confidence masses. For the vision, it consists in the *SLSRS* reliability R_{vis} estimation. For the *GIS*, it consists in a fusion based on digital map criteria. This approach, used in [6] and [9] for different applications, helps to characterize a set of potential speeds w.r.t. to the knowledge of the driving context. In fact, [4] showed that the criteria initially proposed by [3] can be classified in two sets. The first set is suitable for the determination of the *GIS* reliability R_{GIS}, while the second one defines the road context informing about induced speeds (e.g. highway $\Rightarrow 130km.h^{-1}$, urban $\Rightarrow 50km.h^{-1}$ for french context). The *bba* of every candidate speed w.r.t. the criteria are defined using R_{GIS} (if it is low/high, low/high confidence will be given to the *GIS* data) and their fusion allows the determination of the road context (urban, highway, etc.) informing about implicit speed candidates. Finally, a local decision is performed to select the *GIS* best speed candidate based on the evaluated context. Contrary to [4], no extra confidence is put on the speed contained in the *GIS* due to potential out-dated data, errors, etc.

The second level is dedicated to the multi-sensor fusion in which each sensor is independent and specialized on one speed. Finally, a raw speed limit based on the combination using *Smets*' conjunctive rule [10], and the conflict redistributed speed obtained with the *PCR5* [7] are provided to the driver.

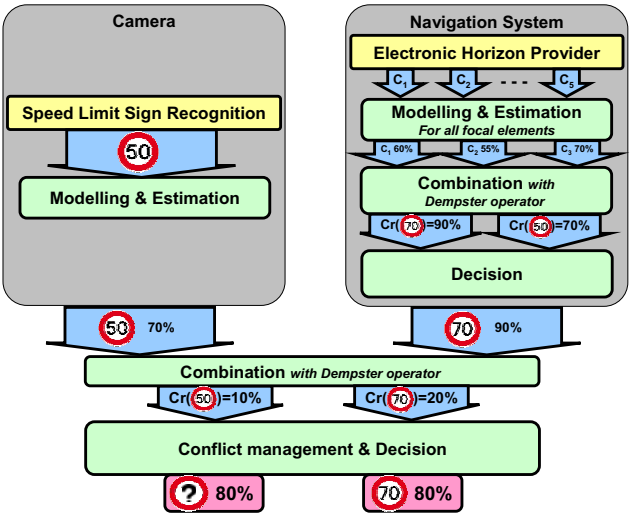


Fig. 1 Multi-level Fusion Strategy

First of all, relevant criteria describing the road context and the *GIS* reliability are selected. This approach presents interesting results as it allows to determine the level of confidence which can be given to the speed traditionally stored in the *GIS* database [3]. Five criteria are considered for the road context description and implicit speeds definition: C_1 describes the road importance in the digitalization level of the map, C_2 refers to the road type (european, highway, national, departmental or communal), C_3 informs about urban or extra urban context, C_4 gives the eventual intersection presence, and finally C_5 provides highway ramp presence/absence.

For the *GIS*, the reliability is evaluated considering its performance in positioning, localization and the resolution of the digital map database. These three elements are subject to inaccuracies, which have different origins, but may lead to false navigation information. For the *GPS* positioning, one of the reliability indicators is the satellite geometry characterized by the Horizontal Dilution Of Precision (*HDOP*). For the localization of the vehicle on the map, an indicator is the probability of the candidate locations provided by the map-matching. In fact, for each vehicle position, a set of possible locations are determined regarding the road context and the candidate with the best probability is selected. This probability, called Most Likely Candidate Probability (*MLCP*), is relevant of the quality of the map-matching. The last element, the digital map, is an approximation of the reality. The digital map database accuracy level is provided by a specific attribute named *ADASAttribute*. The latter denotes if the quality of the road representation is *ADAS*-compliant. Equation (3) presents the way the navigation reliability is finally defined using these elements.

$$R_{GIS} = \left(1 - \left(\frac{HDOP}{HDOP_{\max}}\right)\right) \cdot \left(1 - \left(\frac{MLCP}{MLCP_{\max}}\right)\right) \cdot ADASAttribute \tag{3}$$

3 Combination and Decision

3.1 Multi-criterion and Multi-sensor Fusion

The multi-criterion fusion is performed independently and sequentially for each speed S_j of Θ, consequently specialized to $\Theta = \left\{S_j, S_j^c\right\}$. The resulting power set becomes $2^\Theta = \left\{\emptyset, S_j, S_j^c, \Theta\right\}$. The objective is to evaluate the coherence between the speed candidate S_j and the driving context obtained through the criteria. The combined mass related to S_j w.r.t. criteria C_1 to C_5 is obtained using *Smets*' conjunctive operator. The latter is given for l criteria in [6] such as:

$$\begin{aligned}
m_{1\ldots l,j}(S_j) &= \prod_{i=1}^{l}\left(1 - m_{i,j}\left(S_j^c\right)\right) - \prod_{i=1}^{l} m_{i,j}(\Theta) \\
m_{1\ldots l,j}\left(S_j^c\right) &= \prod_{i=1}^{l}\left(1 - m_{i,j}(S_j)\right) - \prod_{i=1}^{l} m_{i,j}(\Theta) \\
m_{1\ldots l,j}(\Theta) &= \prod_{i=1}^{l} m_{i,j}(\Theta) \\
m_{1\ldots l,j}(\emptyset) &= 1 - \prod_{i=1}^{l}\left(1 - m_{i,j}(S_j)\right) - \prod_{i=1}^{l}\left(1 - m_{i,j}\left(S_j^c\right)\right) + \prod_{i=1}^{l} m_{i,j}(\Theta)
\end{aligned} \tag{4}$$

Contrary to the multi-criterion fusion in which all the criteria express their opinion on a speed S_j at a time, *GIS* and vision may have different points of view, i.e. give opinions on two different speeds S_{GIS} and S_{vis}. In this case, there are only two speeds in the specialized discernment frame $\Theta = \{S_{GIS}, S_{vis}\}$. Considering these elements, the multi-sensor fusion using *Smets*' operator and generalized for p sensors is [6]:

$$\begin{aligned}
m_{1\ldots p}(S_j) &= m_j(S_j) \prod_{\substack{a=1 \\ a\neq j}}^{p}\left(1 - m_a(S_a)\right) + m_j(\Theta) \prod_{\substack{a=1 \\ a\neq j}}^{p}\left(m_a\left(S_a^c\right)\right) \\
m_{1\ldots p}\left(\left\{S_j, \ldots, S_l\right\}\right) &= m_j(\Theta)\ldots m_l(\Theta) \prod_{\substack{a=1 \\ a\neq j \\ \cdots \\ a\neq l}}^{p}\left(m_a\left(S_a^c\right)\right) \\
m_{1\ldots p}\left(S_j^c\right) &= m_j\left(S_j^c\right) \prod_{\substack{a=1 \\ a\neq j}}^{p} m_a(\Theta) \\
m_{1\ldots p}(\Theta) &= \prod_{a=1}^{p} m_a(\Theta) \\
m_{1\ldots p}(\emptyset) &= 1 - \prod_{a=1}^{p}\left(1 - m_a(S_a)\right) - \sum_{a=1}^{p} m_a(S_a) \prod_{\substack{b=1 \\ b\neq a}}^{p}\left(1 - m_b(S_b)\right) + \prod_{a=1}^{p}\left(m_a\left(S_a^c\right)\right)
\end{aligned} \tag{5}$$

3.2 Conflict Management

During the multi-criterion fusion, as the sources express themselves sequentially over the same speed, no conflict can be generated. On the opposite, the multi-sensor fusion combines sources which may be confident in different speeds, thus may lead to conflict. To manage this conflict, the *Proportional Conflict redistribution Rule* 5 (*PCR*5) introduced in [7], has been chosen[1]. This operator is described by (6) with $m_{\textcircled{\cap}}(S_j)$ the mass on speed S_j after the conjunctive combination, and $m_{PCR5}(S_j)$ the mass on speed S_j after conflict redistribution.

$$\begin{aligned} &m_{PCR5}(\emptyset) = 0 \text{ and for } S_j \in 2^{\Theta} \setminus \{\emptyset\} \\ &m_{PCR5}(S_j) = m_{\textcircled{\cap}}(S_j) + \sum_{\substack{S_a \in 2^{\Theta} \setminus \{S_j\} \\ S_j \cap S_a = \emptyset}} \left[\frac{m_j(S_j)^2 m_a(S_a)}{m_j(S_j) + m_a(S_a)} + \frac{m_a(S_j)^2 m_j(S_a)}{m_a(S_j) + m_j(S_a)} \right] \end{aligned} \tag{6}$$

Sources which generate a high conflict have usually strong beliefs in their original propositions. These beliefs are greatly reduced after the combination due to conflict apparition. The use of the *PCR*5 can therefore involve a re-appearance of the original strong beliefs which caused the conflict while preserving the other information obtained from the combination (ignorance, etc.) contrary to the *Dempster* normalization which redistributes it over all the propositions of Θ. For test comparison purposes, the multi-sensor combination results without redistribution are also considered in Section 4. As for the multi-criterion fusion, the selection of the multi-sensor speed is done considering the maximum of Belief:

$$S = \left\{ arg \max_{1 \leq j \leq p} Bel(S_j) \right\} \tag{7}$$

4 Experimental Results

Contrary to [8], in this work, no focal elements are selected since the *bba*s of every speed S_j are defined using the *GIS* criteria. Consequently, the multi-criterion fusion is performed successively over every speed of the discernment frame (8) and it defines the relevant *GIS* speed candidates out of Θ. Finally, the local decision selects the most confident candidate.

$$\Theta = \{5, 10, 20, 30, 45, 50, 60, 70, 80, 90, 100, 110, 120, 130\} \tag{8}$$

This *SLA* has been implemented using Navteq's ADASRP software and ^{RT}MAPS from Intempora [8]. The focus is placed on a driving situation defined by an average *GIS* reliability ($R_{GIS} = 0.68$) based on the MLCP, HDOP and AdasAttribute but, in the same time, on an incoherency between the criteria and the *GIS* extracted speed. Indeed, the *GIS* indicates $50 km.h^{-1}$ while the criteria describe a highway

[1] The generalized Proportional Conflict redistribution Rule for n sources known as *PCR*6 yields to the *PCR*5 when two sources are considered [7].

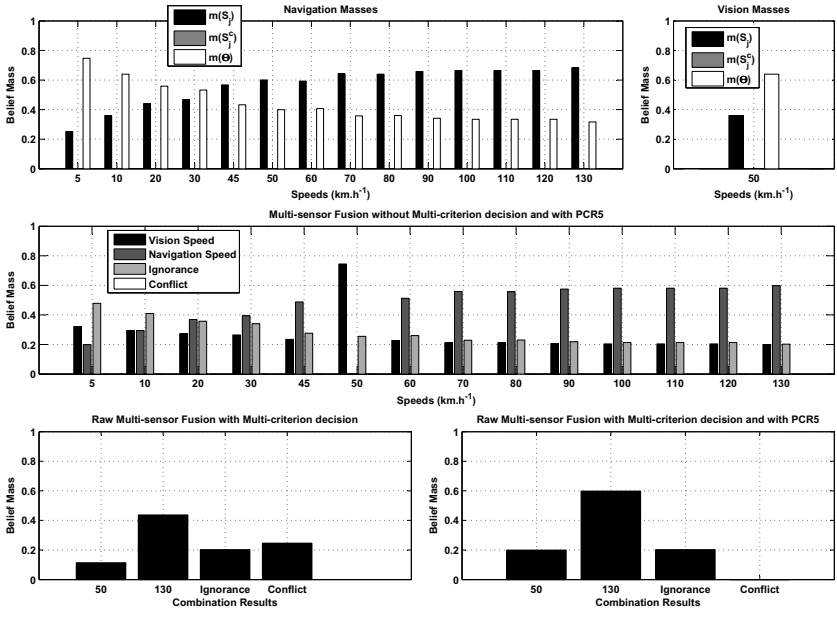


Fig. 2 Multi-level Fusion Results

situation: the vehicle is on an accurately digitalized extra-urban highway, without any intersection nor highway ramp. On the other hand, the vision has detected a speed limit sign of $50km.h^{-1}$ with an equivalent reliability R_{vis} of 0.68. This average reliability leads to the belief masses described in Fig.2 right top plot which show low confidence in the detected speed, due to the neutral *bba* model [8].

The results of the multi-criterion fusion are presented in Fig.2 left top plot. The detection of the *GIS* incoherences is shown as the multi-criterion fusion favorites high speeds, even if no focal elements related to the *GIS* extracted speed are considered, due to the context attributes. The multi-sensor fusion results (Fig.2 middle plot) obtained with *PCR*5, then show that even if the *GIS* and vision belief masses in $50km.h^{-1}$ are low, the combination yields in selecting this speed. This result is involved by the conjunctive combination which favorites the sources common propositions and cancels the benefits of the multi-criterion fusion (*GIS* incoherences detection). On the opposite, by selecting the best *GIS* candidate after multi-criterion fusion (here $130km.h^{-1}$) using the local decision, the final speed becomes $130km.h^{-1}$ as presented in Fig.2 bottom plots. Nevertheless, the selection of $130km.h^{-1}$ is difficult in the non-normalized case (Fig.2 bottom left plot) as the level of ignorance and conflict are close to the level of confidence of the retained speed. Thanks to the *PCR*5 partial conflict management, the belief in the final speed of $130km.h^{-1}$ becomes slightly higher than the belief in $50km.h^{-1}$ and the level of ignorance (cf. Fig.2 bottom right plot). The final speed considering the best *GIS*

candidate is consequently more coherent with the driving context than the final speed determined without decision step after the multi-criterion fusion.

5 Conclusion

This paper has presented an approach to the fusion of a Geographic Information System (*GIS*) and a vision system for Speed Limit Determination. A *Demspter-Shafer* multi-level data fusion composed of a multi-criterion fusion for the definition of the *GIS* reliability and a multi-sensor fusion between the *GIS* and the vision, have been proposed. The main benefit of this approach is to detect the *GIS* incoherences by an earlier evaluation of its reliability based on the positioning, the localization and the digital map performance. Then, contextual information is used to confirm or infirm the *GIS* speed during criteria fusion which is further combined with the vision information. The system has been validated through experimental results. Further studies would be focused on the vision system through the integration of contextual information such as weather conditions.

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
2. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
3. Lauffenburger, J.P., Bradai, B., Basset, M., Nashashibi, F.: Navigation and speed signs recognition fusion for enhanced vehicle location. In: IFAC World Congress (IFAC WC), Seoul, South Korea, September 6-11 (2008)
4. Puthon, A.S., Nashashibi, F.: B Bradai. Improvement of multisensor fusion in speed limit determination by quantifying navigation reliability. In: International Conference on Intelligent Transportation Systems (ITSC), Madeira, Portugal, September 19-22 (2010)
5. Rombaut, M.: Decision in multi-obstacle matching process using Dempster-Shafer's theory. In: Advances in Vehicle Control and Safety (AVCS), Amiens, France, July 1-3 (1998)
6. Royère, C.: Contribution à la résolution du conflit dans le cadre de la théorie de l'évidence: Application à la perception et à la localisation de véhicules intelligents. PhD thesis, Université de Technologie de Compiègne (2002)
7. Smarandache, F., Dezert, J.: Advances and Applications of DSmT for Information Fusion. Collected Works. American Research Press (2009)
8. Daniel, J., Lauffenburger, J.-P.: Conflict management in multi-sensor dempster-shafer fusion for speed limit determination. In: Intelligent Vehicles Symposium (IV), Baden Baden, Germany, June 3–7 (2011)
9. El Najjar, M., Bonnifait, P.: Road selection using multicriteria fusion for the road-matching problem. IEEE Transactions on Intelligent Transportation Systems 8, 264–278 (2007)
10. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)

A New Local Measure of Disagreement between Belief Functions – Application to Localization

Arnaud Roquel, Sylvie Le Hégarat-Mascle, Isabelle Bloch, and Bastien Vincke

Abstract. In the theory of belief functions, the disagreement between sources is often measured in terms of conflict or dissimilarity. These measures are global to the sources, and provide few information about the origin of the disagreement. We propose in this paper a "finer" measure based on the decomposition of the global measure of conflict (or distance). It allows focusing the measure on some hypotheses of interest (namely the ones likely to be chosen after fusion). We apply the proposed so called "local" measures of conflict and distance to the choice of sources for vehicle localization. We show that considering sources agreement/disagreement outperforms blind fusion.

1 Introduction

Multi-sensor systems are used in many applications such as classification, image processing, change detection, object trajectory localization. Usually the information provided by each sensor is prone to imperfections, such as imprecision and uncertainty, and fusion procedures aim at making better decisions by combining multi-sensor information. Belief Functions (BF) are suitable for modeling imprecision and uncertainty, and handle belief on the power set of the frame of discernment (set of hypotheses). A disagreement between sources makes the system unstable and can impact the decision. Many techniques have been developed to measure the disagreement between sources. A review can be found in [4] or [5]. One method consists in observing the so-called "Demspter's conflict" [10] resulting from the conjunctive combination of the basic belief functions. However, the non-idempotence

Arnaud Roquel · Sylvie Le Hégarat-Mascle · Bastien Vincke
Université Paris Sud, IEF, Orsay, France
e-mail: first-name.last-name@u-psud

Isabelle Bloch
Télécom ParisTech, CNRS LTCI, Paris, France
e-mail: isabelle.bloch@telecom-paristech.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 335–342.
springerlink.com

of the usual conjunctive rule can create a non-zero conflict for the combination of two equal belief functions. Other measures are based on distances between mass functions. The distances derived from L1 or L2 norms measure the inter-sources disagreement taking into account all elements of the space of discernment.

In this paper, we aim at exploiting the conflict or distance to provide a diagnosis of the system status. For this we need a more precise measurement than the "Dempster's conflict" or global dissimilarity between sources. Thus we propose a new measure which is related to the different elements of the discernment space, that we call "local" measure. After recalling some notations and basic elements on mass function decompositions and distance measures in Section 2, the proposed measure is introduced and analyzed in Section 3. It is then illustrated on a vehicle localization problem described in Section 4. Results are provided in Section 5.

2 Background

In the following, we denote by Ω the frame of discernment, by 2^Ω the power set of Ω, and by m_j a Basic Belief Assignment (BBA) on 2^Ω associated with a source S_j. Plausibility and communality are denoted by Pls and q, respectively. Smets proposed a canonical decomposition of every non-dogmatic BBA, as a unique conjunctive combination of simple support functions (SSF) [9]:

$$m_j = \textcircled{\cap}_{A\subset\Omega} A^{w_j(A)}, \tag{1}$$

where $A^{w_j(A)}$ is a SSF, i.e. a function with only two focal elements A and Ω, such that $A^{w_j(A)}(\Omega) = w_j(A)$, $A^{w_j(A)}(A) = 1 - w_j(A)$, and $A^{w_j(A)}(B) = 0, \forall B \in 2^\Omega \setminus \{A, \Omega\}$. If $w_j(A) \leq 1$ then $A^{w_j(A)}$ is a BBA, and if $w_j(A) \leq 1, \forall A \subset \Omega$, m_j in Eq. 1 is a separable BBA (SBBA). The weight $w_j \in \mathbb{R}^+$ is expressed from the commonalities as follows:

$$\forall A \subset \Omega,\ w_j(A) = \prod_{B\supseteq A} q_j(B)^{(-1)^{|B|-|A|+1}}. \tag{2}$$

For the conjunctive combination of N BBAs, two main rules are generally considered depending on whether the sources are "cognitively independent", and can be expressed using the canonical decomposition: Smets' combination [9] (sometimes simply called conjunctive rule because of its authority): $m_{\textcircled{\cap}} = \textcircled{\cap}_{A\subset\Omega} A^{\prod_{j=1}^{N} w_j(A)}$, and the cautious rule [2]: $m_{\textcircled{\wedge}} = \textcircled{\cap}_{A\subset\Omega} A^{\bigwedge_{j=1}^{N} w_j(A)}$, where $\bigwedge$ denotes the minimum operator.

The dissimilarity between BBAs is often used for computing their disagreement. It is generally estimated from a conflict or distance measure (the reader can refer to [5] or [4] for an overview). These measures involve all elements of 2^Ω.

In this study, we focus on the conflict (as a diagnostic tool of the system) between different sources. Besides, in the estimation of a disagreement (conflict), to avoid the bias due to the individual source auto-conflict, we consider sources with null auto-conflict, namely modelled using consonant BBAs, as proposed in [6].

3 Proposed Local Measures of Sources Disagreement

3.1 Local Conflict

We note $\Upsilon = \{\{A_i\}_j, A_i \subseteq \Omega,\ j \in \{1,...,N\}\}$ the multi-set containing the elements of the canonical decomposition of the BBAs to be combined, where $\{A_i\}_j$ is the set of elements of the canonical decomposition of m_j for which $w_j(A_i) \neq 1$. From Eq. 1, the mass of the empty set resulting from the combination of N SBBAs defined on Ω (with $|\Omega| > 2$) is:

$$m_{\textcircled{\cap}}(\emptyset) = 1 + \sum_{B \subseteq \Omega, B \neq \emptyset} (-1)^{|B|} \prod_{k=1}^{|\Upsilon|} \sum_{B \subseteq A} w_k(A). \tag{3}$$

In the following, we focus on the case of two consonant BBAs. If Υ is not a consonant set, then conflict appears. Now for two BBAs the conflict can be brought by different hypotheses. We propose to analyze the origin of the conflict by decomposing it on pairs of elements. For this we consider the canonical decomposition of $m_{1\,\textcircled{\cap}\,2}$ and we analyze the conflict between the pairs of elements (singletons or compound hypotheses) of this decomposition.

We introduce the following function $f_\emptyset$ on $2^\Omega \times 2^\Omega$ for conflict decomposition:

$$\forall (A,B) \ / \ A \cap B \neq \emptyset \ , \ f_\emptyset(A,B) = 0, \tag{4}$$

$$\forall (A,B) \ / \ A \cap B = \emptyset \ , \tag{5}$$

$$\begin{aligned} f_\emptyset(A,B) &= \frac{1}{2} \sum_{g=1}^{|\Upsilon|} \sum_{l=1}^{|\Upsilon|} \mu_g(A) \times \mu_l(B) \times \sum_{\substack{\{X_1,...,X_{|\Upsilon|-2}\} \\ \in \{\Upsilon_{\supseteq A} \cup \Upsilon_{\supseteq B}\}}} \prod_{\substack{k=1 \\ k \neq \{g,l\}}}^{|\Upsilon|} \mu_k(X_k), \\ &= \frac{1}{2} \sum_{g=1}^{|\Upsilon|} \sum_{l=1}^{|\Upsilon|} \mu_g(A) \times \mu_l(B) \times \prod_{\substack{k=1, \\ k \neq \{g,l\} \\ X_k \in \{\Upsilon_{\subset A} \cup \Upsilon_{\subset B}\}}}^{|\Upsilon|} \mu_k(\Omega), \end{aligned} \tag{6}$$

where $\Upsilon_{\supseteq A}$ is the set of elements of Υ including A: $\Upsilon_{\supseteq A} = \{X \in \Upsilon \ / \ A \subseteq X\}$ and $\Upsilon_{\subset A}$ is the set of elements being strictly included in A: $\Upsilon_{\subset A} = \{X \in \Upsilon / X \subset A\}$. $\mu_j(A) = A^{w_j(A)}$ if A is an element of the decomposition of m_j and $\mu_j(A)$ is the vacuous BBA (such that $m(\Omega) = 1$) if A is not an element of the decomposition of m_j.

For each element of 2^Ω we define the conflict brought by this element as:

$$\forall A_i \in 2^\Omega,\ Mes_\emptyset(A_i) = \sum_{A_j \in 2^\Omega} f_\emptyset(A_i, A_j). \tag{7}$$

The mass on the empty set (Eq. 3), is thus:

$$m_{\textcircled{\cap}}(\emptyset) = \sum_{A \in 2^{\Omega}} Mes_{\emptyset}(A). \tag{8}$$

Example: Let m_1 and m_2 be two consonant SBBAs defined on $\Omega = \{a,b,c\}$ (see the table below). Here $\Upsilon = \{\{a\},\{b\},\{a \cup c\},\{b \cup c\}\}$. After conjunction, $\forall A_i \in \Upsilon$, $A_i^{\prod_{j=1}^{2} w_j(A_i)}$ is a SSF.

	$\{a\}$	$\{b\}$	$\{a \cup b\}$	$\{c\}$	$\{a \cup c\}$	$\{b \cup c\}$	$\{\Omega\}$	$\{\emptyset\}$
m_1	0.3	0	0	0	0.6	0	0.1	0
m_2	0	0.3	0	0	0	0.6	0.1	0
$m_{1 \textcircled{\cap} 2}$	0.03	0.03	0	0.36	0.06	0.06	0.01	0.45
w_1	0.7	1	1	1	0.1429	1	1	1
w_2	1	0.7	1	1	1	0.1429	1	1
μ_1	0.3	0	0	0	0	0	0.7	0
μ_2	0	0.3	0	0	0	0	0.7	0
μ_3	0	0	0	0	0.8571	0	0.1429	0
μ_4	0	0	0	0	0	0.8571	0.1429	0

$\forall A_i \neq \Omega, \mu_i(A_i) = 1 - \prod_{j=1}^{2} w_j(A_i)$ and $\mu_i(\Omega) = \prod_{j=1}^{2} w_j(A_i)$. From Eq. 6, the decomposition of $m_{1 \textcircled{\cap} 2}(\emptyset)$ can be written as:

Decomposition of $m_{1 \textcircled{\cap} 2}(\emptyset)$	Pairs of conflicting hypotheses
$\mu_1(\{a\}) \times \mu_2(\{b\})$	$(\{a\},\{b\})$
$\mu_1(\{a\}) \times \mu_2(\{\Omega\}) \times \mu_4(\{b \cup c\})$	$(\{a\},\{b,c\})$
$\mu_1(\{\Omega\}) \times \mu_1(\{b\}) \times \mu_3(\{a \cup c\})$	$(\{b\},\{a,c\})$

The result of the conflict decomposition is:

	$\{a\}$	$\{b\}$	$\{a \cup b\}$	$\{c\}$	$\{a \cup c\}$	$\{b \cup c\}$	$\{\Omega\}$	$\{\emptyset\}$
$Mes_{\emptyset}$	0.27	0.27	0	0	0.18	0.18	0	0

For this example, we note that the conflict is mainly due to the couple of hypotheses $\{a\}$ and $\{b\}$.

3.2 Local Pseudo-distance

In Section 3.1, we introduced the notion of "local" conflict induced by a hypothesis. In a similar way, we introduce a local pseudo-distance:

$$Dist_{Pl_{1,2}}(A,B) = \frac{1}{2} \mid (Pl_1(A) - Pl_2(A)) + (Pl_2(B) - Pl_1(B)) \mid, \tag{9}$$

where Pl_j, is the plausibility function associated with $m_j, j = \{1,2\}$, and A and B denote two elements of 2^{Ω}. This defines a pseudo-metric: it is non-negative and

symmetrical by construction, $\forall A \in 2^{\Omega}, Dist_{Pl_{1,2}}(A,A) = 0$ and satisfies the triangular inequality: $\forall (A,B,C) \in (2^{\Omega})^3$, $Dist_{Pl_{1,2}}(A,C) + Dist_{Pl_{1,2}}(C,B) \geq Dist_{Pl_{1,2}}(A,B)$.

Note that the detection of a partial conflict between BBAs and the detection of a high distance have very different interpretations. In the first case, we aim at selecting the hypotheses mainly inducing conflict in order to specify the conflict (origin, type of conflict, etc.). In the second case, we aim at restricting the measure of distance to a sub-part of 2^{Ω} (pairs of elements) because our interest focuses on some hypotheses (typically those that can be selected when making the decision).

4 Application to the Localization Problem

4.1 Localization Problem

In this section, we apply the previously presented measures to the problem of vehicle localization using different sources j, here algorithms providing localization estimates from vehicle sensors (odometers, camera). Odometers provide the distance travelled by each wheel independently. Using the wheel parameters (radius, length of the rear axle, tick number) and assuming a rigid structure of the vehicle, we can compute its displacement (longitudinal and rotational components). From the camera data, features (interest points i.e. SURF, SIFT points, etc.) are tracked in several images, both to infer the scene structure (3D) and the camera movement [1]. In our experiments, the longitudinal and rotational components of displacement are estimated using three different algorithms. The first one (S_1) exploits only odometer data. The second one (S_2), FastSLAM algorithm [7], exploits both odometer and camera. Finally, the third algorithm (S_3), exploits only images. The estimates from these three algorithms are more or less precise depending on the physical world and the movement of the vehicle. A wheel sliding may induce an error in the estimates of the algorithms using odometer data; an homogeneous environment or a mismatch between features may induce an error for the algorithms using camera data.

4.2 Fusion Model

At each instant the movement is described by a couple $(\delta_s, \delta_\Theta)$ (longitudinal and rotational components), whose values are bounded by the motor vehicle features. Each hypothesis of Ω represents a pair of values $(\delta_s, \delta_\Theta)$. We denote the measurement provided by a given source at instant t by $\overrightarrow{\delta(t)} = (\delta_s(t), \delta_\theta(t))^t$, and the measurement associated to a hypothesis H by $\overrightarrow{\delta_H} = (\delta_s(H), \delta_\theta(H))^t$. The considered measure between $\delta_s(H)$ and $\delta_\theta(H)$ is the Mahalanobis distance $d^2(\overrightarrow{\delta_t}, \overrightarrow{\delta_H}) = \begin{pmatrix} \delta_s[H] - \delta_s(t) \\ \delta_\theta[H] - \delta_\theta(t) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \delta_s[H] - \delta_s(t) \\ \delta_\theta[H] - \delta_\theta(t) \end{pmatrix}$, where Σ is the covariance matrix.

We assume longitudinal and rotational components of the movement are decorrelated, and thus Σ is diagonal. We also assume that the more the acceleration is

important, the less accurate are the movement estimations by the considered algorithms, and thus the higher are the Σ terms: in our model Σ depends on the movement estimate itself. The ellipsoid centered at $\overrightarrow{\delta_t}$ models the movement of the vehicle. The probability of a hypothesis $H, H \in \Omega$, is calculated conditionally to $\overrightarrow{\delta_t}$:

$$P(H \mid \overrightarrow{\delta_t}) = \frac{1}{2\pi \times |\Sigma|^{\frac{1}{2}}} \exp\{\frac{-d^2(\overrightarrow{\delta_t}, \overrightarrow{\delta_H})}{2}\} \tag{10}$$

The higher the distance between hypothesis H and source estimate at t, the lower the probability of H. The mass allocation proposed by Dubois [3] builds a consonant BBA (the less committed BBA having given a pignistic probability) centred on the hypothesis maximizing Eq. 10. For consonant BBAs, the number of focal elements is $|\Omega|$, and the auto-conflict [6] is null.

As second main hypothesis about the data model, we assume the sampling of data ($30Hz$) is high relatively to the acceleration so that $(\delta_s(t), \delta_\theta(t))$ varies slowly versus time. This so called "regularity assumption" allows us to consider $(\delta_s(t-1), \delta_\theta(t-1))$ as sources for the estimation of the vehicle movement at t, even if less reliable than measurements at t. We will see in the next section how such $(t-1)$ sources are used in the data fusion process.

Finally, for combination, recall that S_1 and S_2, which both use odometer data, are not independent, and that S_2 and S_3, which both use camera data, are also not independent. Independence between sources can only be assumed for S_1 and S_3. In this study, our aim is to show the interest of the conflict measurement, and sources are combined at the same time. Therefore we consider that the sources are at least partially correlated and we use the cautious combination proposed by Denoeux [2].

4.3 Exploitation of Conflict

As said in Section 4.2, the precision of sources is time varying (e.g. mainly depends on the acceleration), and so is its reliability. In this work, we estimate dynamically the reliability of the sources to improve the fusion robustness. The estimation of conflict (Eq. 7) is "local" to the candidate to be chosen by the fusion. If this latter is conflictual, we try to remove the "unreliable" sources. Using three sources, we could have chosen a majority criterion to decide the reliable sources. However, sources being partially correlated, we prefer to base the detection of reliable sources on "regularity assumption", based on the local distance (Eq. 9), between successive instant measurements. It allows us to focus on the information concerning the hypotheses of interest (the ones selected by the sources).

Precisely, if we denote by $H_1, H_2, H_3, H_{\textcircled{\wedge}}$ the singleton elements maximizing the plausibility function of respectively $m_1, m_2, m_3, m_{\textcircled{\wedge}}$, where m_1, m_2, m_3 are the consonant BBAs associated with sources S_1, S_2, S_3 described in Section 4.2 and $m_{\textcircled{\wedge}}$ is the BBA after combination of m_1, m_2, m_3 by the cautious rule, the exploitation of conflict is composed of three steps:

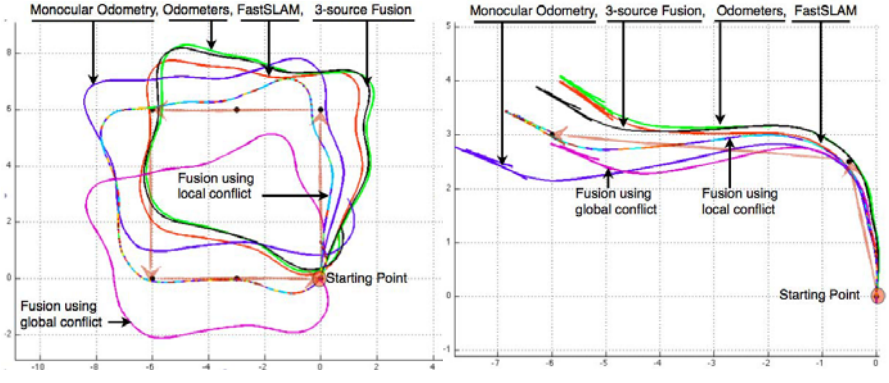
1. Compute the level of conflict introduced by the singleton element chosen by the decision step: $Mes(H_{\textcircled{\wedge}}) = \sum_{B \subseteq \Omega, H_{\textcircled{\wedge}} \subset B} Mes_{\emptyset}(B)$.
2. If $Mes(H_{\textcircled{\wedge}}) > T_M$, then search the sources which do not respect the assumption of regularity: $(Dist_{Pl_n}(H_n(t), H_n(t-1)) > T_D$, with

$$Dist_{Pl_n}(H_n(t), H_n(t-1)) = \frac{1}{2} \mid (Pl_{n_t}(H_n(t)) - Pl_{n_{t-1}}(H_n(t))) + (Pl_{n_{t-1}}(H_n(t-1)) - Pl_{n_t}(H_n(t-1))) \mid,$$

   where $n = \{1,2,3\}$ is the source index, t and $t-1$ two successive times, and Pl_{n_t} is the plausibility of source n at time t. The threshold values T_M and T_D have been fixed experimentally to 0.1 and 0.5.
3. Combine the sources which have been found as reliable using the two conditions.

5 Results and Conclusion

In this section we present the results obtained in the case of two various trajectories. The first one includes a strong acceleration at the beginning of the trajectory, inducing a sliding of the wheels. During the second trajectory, there is an acceleration at a turn. Figure 1 presents a 2D top view of the 3D physical world. On both trajectories, we remark a wrong odometer estimation either at the beginning, or at the turn, due to the sliding of the wheels. The monocular vision algorithm shows



(a) Trajectory one (b) Trajectory two

Fig. 1 Two different trajectories. On each we can observe respectively in red, green and blue the integration of the movement estimation by odometer data (S_1), FASTSLAM (S_2) and visual odometry (S_3) algorithms. The trajectory in black represents the integration of movement estimated by the fusion of sources S_1, S_2 and finally the multi-color and purple trajectories correspond to the integration of the movement estimation exploiting the local conflict and the global conflict (process derived from [8]), respectively.

also some limitations due to some imprecision in the camera mode (parameters) and some matching errors in the presence of a white wall. These causes of errors also occur for the FASTSLAM algorithm that uses both kinds of data.

We observe that the conflict as defined in Section 4.3 allows us to estimate a movement close to the ground truth even in extreme cases. We also observe that it outperforms the result of the three source fusion not considering their reliability.

In conclusion, this paper introduces a "local" measure to compute the disagreement between sources. Theoretical and experimental examples show that a global measure like "Demspter's conflict" or dissimilarity do not always allow a fine analysis of source reliability and origin of conflict, while the proposed local measure does. Further analysis of the properties of the local measures of conflict, potential extension to non-consonant BBA, more experiments on localization and other applications are planned for our future work.

Acknowledgements. This work was partially supported by a grant from Digiteo.

References

1. Bak, A., Bouchafa, S., Aubert, D.: Detection of independently moving objects through stereo vision and ego-motion extraction. In: Intelligent Vehicles Symposium (IV), pp. 863–870. IEEE (2010)
2. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artificial Intelligence 172(2-3), 234–264 (2008)
3. Dubois, D., Prade, H.: Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence 4(3), 244–264 (1988)
4. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53(2), 118–145 (2012)
5. Liu, W.: Analyzing the degree of conflict among belief functions. Artificial Intelligence 170(11), 909–924 (2006)
6. Martin, A., Jousselme, A.L., Osswald, C.: Conflict measure for the discounting operation on belief functions. In: The 11th Annual Conference on Information Fusion, pp. 1–8. IEEE, Cologne, Germany (2008)
7. Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B.: FastSLAM: A factored solution to the simultaneous localization and mapping problem. In: National Conference on Artificial Intelligence, pp. 593–598. AAAI, Menlo Park (2002)
8. Schubert, J.: Conflict management in Dempster-Shafer theory using the degree of falsity. International Journal of Approximate Reasoning 52(3), 449–460 (2011)
9. Smets, P.: The canonical decomposition of a weighted belief. In: 14th International Joint Conference on Artificial intelligence, pp. 1896–1901. Morgan Kaufmann Publishers Inc., San Francisco (1995)
10. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion 8(4), 387–412 (2007)

Map-Aided Fusion Using Evidential Grids for Mobile Perception in Urban Environment

Marek Kurdej, Julien Moras, Véronique Cherfaoui, and Philippe Bonnifait

Abstract. Evidential grids have been recently used for mobile object perception. The novelty of this article is to propose a perception scheme using prior map knowledge. A geographic map is considered an additional source of information fused with a grid representing sensor data. Yager's rule is adapted to exploit the Dempster-Shafer conflict information at large. In order to distinguish stationary and mobile objects, a counter is introduced and used as a factor for mass function specialisation. Contextual discounting is used, since we assume that different pieces of information become obsolete at different rates. Tests on real-world data are also presented.

1 Introduction

Autonomous driving has been an important challenge in recent years. Navigation and precise localisation aside, environment perception is an important on-board system of a self-driven vehicle. The level of difficulty in autonomous driving increases in urban environments, where a good scene understanding makes the perception subsystem crucial. There are several reasons that make cities a demanding environment. Poor satellite visibility deteriorates the precision of GPS positioning. Vehicle trajectories are hard to predict due to high variation in speed and direction. Also, the sheer number of mobile objects poses a problem, e.g. for tracking algorithms.

On the other hand, more and more detailed and precise geographic databases become available. This source of information has not been well examined yet, hence our approach of incorporating prior knowledge from digital maps in order to improve perception scheme. A substantial amount of research has focused on the mapping problem for autonomous vehicles, e.g. Simultaneous Localisation and Mapping (SLAM) approach, but the use of maps for perception is still understudied.

Marek Kurdej · Véronique Cherfaoui · Julien Moras · Philippe Bonnifait
UMR CNRS 6599 Heudiasyc, University of Technology of Compiégne, France
e-mail: marek.kurdej@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 343–350.
springerlink.com

In this article, we propose a data fusion method based on Dempster–Shafer theory [8] taking into account meta-knowledge obtained from a digital map. We show the advantage of including prior knowledge into an embedded perception system of an autonomous car. The vehicle environment is modelled by 2D occupancy grids proposed in [2]. This paper describes a robust and unified approach to a variety of problems in spatial representation using the theory of probability. The theory of evidence was not combined with occupancy grids until recently to build environment maps for robot perception [7]. Only recent works take advantage of the theory of evidence in the context of mobile perception [4]. Some works use 3D city model as a source of prior knowledge for localisation and vision-based perception [1], whereas our method uses maps for scene understanding.

This article is organised as follows. In section 2, we describe the details of the method. Section 3 presents the results and section 4 concludes the paper.

2 Multi-grid Fusion Approach

This section presents the proposed perception schemes. The grid construction method is described in section 2.2 and all data processing steps are detailed in section 2.4. Figure 1 presents a general overview of our approach.

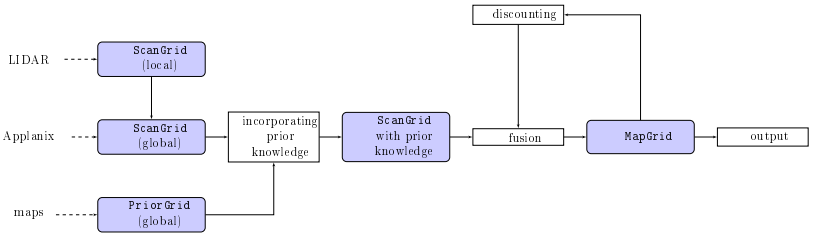


Fig. 1 Method overview (lidar: laser scanner, Applanix: inertial measurement unit).

2.1 Heterogeneous Data Sources

There are three sources in our perception system: vehicle pose, lidar range scanner point cloud and vector maps. The vehicle pose comes from the Applanix system based on a GPS, an odometer and an IMU. The system is supposed to provide precise and integral positioning. Our main source of information about the environment is an IBEO Alaska XT lidar able to provide a cloud of about 800 points 10 times per second. The digital maps that we use were provided by the French National Geographic Institute (IGN) and contain 3D building models as well as the road surface. We also performed successful tests with freely available OpenStreetMap project 2D maps [6], but here we limited the use to building data. We assume the maps to be precise and accurate.

2.2 Occupancy Grids

An occupancy grid models the world using a tessellated representation of spatial information. In general, it is a multidimensional spatial lattice with cells storing some stochastic information. In our case, each cell representing a box (a part of environment) $X \times Y$ where $X = [x_-, x_+]$, $Y = [y_-, y_+]$ stores a mass function.

- `ScanGrid` (SG) construction: In order to process the lidar data, an evidential occupancy grid is computed when a new scan arrives, this grid is called `ScanGrid`. Each cell of this grid stores a mass function on the frame of discernment (FOD) $\Omega_{SG} = \{F, O\}$, where F refers to the free space and O – to the occupied space. The basic belief assignment, which reflects the sensor model, is described in [4].
- `MapGrid` (MG): To store the results of information fusion, an occupancy grid MG has been introduced with a FOD $\Omega_{MG} = \{F, C, N, S, V\}$. Respective classes represent: free space F, mapped infrastructure (buildings) C, non-mapped infrastructure N, temporarily stopped objects S and mobile (moving) V objects. Ω_{MG} is a common frame used for information fusion. By using MG as a cumulative information storage, we are not obliged to aggregate preceding `ScanGrid`s.
- `PriorGrid` (PG) context representation: PG allows us to perform a contextual information fusion incorporating some meta-knowledge about the environment. This grid uses the same frame of discernment Ω_{MG} as MG. The grid is obtained by projection of map data, buildings and roads, onto a 2D grid with global coordinates.
  We define two sets of polygons defining the 2D position of buildings and road surface by, respectively, $\mathscr{B} = \left\{ b_i = \begin{bmatrix} x_1 x_2 \dots x_{m_i} \\ y_1 y_2 \dots y_{m_i} \end{bmatrix}, i \in [0, n_B] \right\}$ and $\mathscr{R} = \left\{ r_i = \begin{bmatrix} x_1 x_2 \dots x_{m_i} \\ y_1 y_2 \dots y_{m_i} \end{bmatrix}, i \in [0, n_R] \right\}$, $\mathscr{B} \cap \mathscr{R} = \emptyset$. Then, we attribute the mass to each cell $\{X, Y\}$ of the `PriorGrid` in the following way:
  We note that $B = \{C\}$, $R = \{F, S, V\}$, $T = \{F, N, S, V\}$ for convenience and readability only. A denotes all other strict subsets of Ω. These aliases characterise the meta-information inferred from geographic maps. For instance, on the road surface R, we *encourage* the existence of free space F as well as stopped S and moving V objects. Analogically, building information B fosters mass transfer to C. Lastly, T denotes the intermediate area, e.g. pavements, where mobile and stationary objects as well as small urban infrastructure can be present. Note that neither buildings nor roads are present, so we exclude existence of mapped infrastructure C, but we cannot omit other classes. Also, we define a level of confidence β for each map source, possibly different for each context. Let $\tilde{x} = \frac{x_- + x_+}{2}$, $\tilde{y} = \frac{y_- + y_+}{2}$.

$$
\begin{aligned}
m_{PG}\{X,Y\}(B) &= \begin{cases} \beta_B & \text{if } (\tilde{x},\tilde{y}) \in b_i \\ 0 & \text{otherwise} \end{cases} && \forall i \in [0,n_B] \\
m_{PG}\{X,Y\}(R) &= \begin{cases} \beta_R & \text{if } (\tilde{x},\tilde{y}) \in r_i \\ 0 & \text{otherwise} \end{cases} && \forall i \in [0,n_R] \\
m_{PG}\{X,Y\}(T) &= \begin{cases} 0 & \text{if } (\tilde{x},\tilde{y}) \in b_i \vee (\tilde{x},\tilde{y}) \in r_j \\ \beta_T & \text{otherwise} \end{cases} && \forall i \in [0,n_B], \forall j \in [0,n_R] \\
m_{PG}\{X,Y\}(\Omega) &= \begin{cases} 1-\beta_B & \text{if } (\tilde{x},\tilde{y}) \in b_i \\ 1-\beta_R & \text{if } (\tilde{x},\tilde{y}) \in r_i \\ 1-\beta_T & \text{otherwise} \end{cases} && \forall i \in [0,n_B], \forall j \in [0,n_R] \\
m_{PG}\{X,Y\}(A) &= 0 && \forall A \subsetneq \Omega \text{ and } A \notin \{B,R,T\}
\end{aligned}
\tag{1}
$$

2.3 Incorporating Prior Knowledge

The frame of discernment Ω_{SG} used in SG is distinct from Ω_{MG}, so in order to enable the fusion of SG and MG we define a refining $r_{SG} : 2^{\Omega_{SG}} \rightarrow 2^{\Omega_{MG}}$ such that $r_{SG}(\{F\}) = \{F\}$, $r_{SG}(\{O\}) = \{C,N,S,V\}$, $r_{SG}(A) = \bigcup_{\theta \in A} r_{SG}(\theta)$. The refined mass function can be expressed as $m_{SG}^{\Omega_{MG}}(r_{SG}(A)) = m_{SG}^{\Omega_{SG}}(A), \forall A \subseteq \Omega_{SG}$. Then, Dempster's rule is applied in order to exploit the prior information included in `PriorGrid`:

$$
m'^{\Omega_{MG}}_{SG,t} = m^{\Omega_{MG}}_{SG,t} \oplus m^{\Omega_{MG}}_{PG} \tag{2}
$$

2.4 Temporal Fusion

Computing Conflict Masses

We use the idea from [5] to distinguish between two types of conflict, which arise from the fact that the environment is dynamic. We denote $\emptyset_{FO}$ the conflict induced when a free cell in MG is fused with an occupied cell in SG. Similarly, $\emptyset_{OF}$ indicates the conflicted caused by an occupied cell in MG fused with a free cell in SG. In an error-free case, these conflicts represent, respectively, the disappearance and the appearance of an object. Conflict masses are calculated using the formulas: $m_{MG,t}(\emptyset_{OF}) = m_{MG,t-1}(O) \cdot m_{SG,t}(F)$, $m_{MG,t}(\emptyset_{FO}) = m_{MG,t-1}(F) \cdot m_{SG,t}(O)$, where $m(O) = \sum_A m(A)$, $\forall A \subseteq \{C,N,S,V\}$.

`MapGrid` Specialisation Using a Counter

Mobile object detection is an important issue in dynamic environments. We propose the introduction of a counter ζ in each cell in order to include temporal information

on the cell occupancy. For this purpose, incrementation and decrementation steps $\delta_{inc} \in [0,1]$, $\delta_{dec} \in [0,1]$, as well as threshold values γ_O, $\gamma_\emptyset$ have been defined.

$$\zeta^{(t)} = \min\left(1, \zeta^{(t-1)} + \delta_{inc}\right) \quad \text{if } m_{MG}(O) \geq \gamma_O \text{ and } m_{MG}(\emptyset_{FO}) + m_{MG}(\emptyset_{OF}) \leq \gamma_\emptyset$$
$$\zeta^{(t)} = \max\left(0, \zeta^{(t-1)} - \delta_{dec}\right) \quad \text{if } m_{MG}(\emptyset_{FO}) + m_{MG}(\emptyset_{OF}) > \gamma_\emptyset$$

Otherwise $\zeta(t)$ rests unchanged. Using ζ values, we impose a specialisation of mass functions in MG using the equation:

$$m'_{MG,t}(A) = S(A,B) \cdot m_{MG,t}(B) \tag{3}$$

where specialisation matrix $S(\cdot,\cdot)$ is defined as:

$$\begin{aligned} S(A \setminus \{V\}, A) &= \zeta && \forall A \subseteq \Omega_{MG} \text{ and } \{V\} \in A \\ S(A,A) &= 1-\zeta && \forall A \subseteq \Omega_{MG} \text{ and } \{V\} \in A \\ S(A,A) &= 1 && \forall A \subseteq \Omega_{MG} \text{ and } \{V\} \notin A \\ S(\cdot,\cdot) &= 0 && \text{otherwise} \end{aligned} \tag{4}$$

Fusion Rule

An important part of the method consists in fusing a discounted and specialized MG (see section 2.5 and preceding paragraph) with a SG combined with prior knowledge (see section 2.3).

$$m_{MG,t} =^{\alpha} m'_{MG,t-1} \circledast m'_{SG,t} \tag{5}$$

The fusion rule $\circledast$ is a modified Yager's rule [10] adapted to mobile object detection. There are of course many different rules that could be used, but in order to distinguish between moving and stationary objects some modifications had to be included. These modifications consist in transferring the mass corresponding to a newly appeared object $\emptyset_{FO}$ to the class of moving objects V as described by the equation 6. Symbol $\textcircled{\cap}$ denotes the conjunctive fusion rule.

$$\begin{aligned} (m_1 \circledast m_2)(A) &= (m_1 \textcircled{\cap} m_2)(A) && \forall A \subsetneq \Omega \wedge A \neq V \\ (m_1 \circledast m_2)(V) &= (m_1 \textcircled{\cap} m_2)(V) + (m_1 \textcircled{\cap} m_2)(\emptyset_{FO}) \\ (m_1 \circledast m_2)(\Omega) &= (m_1 \textcircled{\cap} m_2)(\Omega) + (m_1 \textcircled{\cap} m_2)(\emptyset_{OF}) \\ (m_1 \circledast m_2)(\emptyset_{FO}) &= 0 \\ (m_1 \circledast m_2)(\emptyset_{OF}) &= 0 \end{aligned} \tag{6}$$

All the above steps allow us to construct a `MapGrid` containing reach information on the environment state, including the knowledge on mobile and static objects.

2.5 Contextual Discounting

Information discounting allows to *forget* information which is no longer valid. Discounting parameter α serves to model the speed with which information becomes obsolete. Thanks to the contextual discounting [3], we make use of more detailed information regarding the confidence we have in the source in various contexts. We noticed that different pieces of information become obsolete with different speed. Hence, the coarsening used is $\Theta = \{\theta_{static}, \theta_{dynamic}, \theta_{free}\}$, with $\theta_{static} = \{C, N\}$, $\theta_{dynamic} = \{S, V\}$, $\theta_{free} = \{F\}$, and discount rates $\alpha = \{\alpha_{static}, \alpha_{dynamic}, \alpha_{free}\}$. We assign higher discount rates (lower confidence) to rapidly changing contexts such as free space, stopped and moving objects, and lower rates to the static context. The discounted mass function is obtained by the disjunctive combination of the input mass function m_{MG} and mass functions for each element of the partition Θ.

$$^{\alpha}m_{MG,t} = m_{MG,t} \textcircled{\cup} m_{static} \textcircled{\cup} m_{dynamic} \textcircled{\cup} m_{free} \tag{7}$$

where each mass function m_l ($l =$ static, dynamic, free) is defined by $m_l(\theta_l) = \alpha_l$, $m_l(\emptyset) = 1 - \alpha_l$, $m_l(A) = 0, \forall A \subseteq \Omega \wedge A \notin \{\emptyset, \theta_l\}$.

3 Results

3.1 Setup

The data set used for our experiments was acquired in cooperation with IGN in Paris. The overall length of the trajectory was about 3 km. The size of the grid cell in the occupancy grids was set to 0.5 m, which is sufficient to model a complex environment with mobile objects. The discount rates α describing the speed of information becoming obsolete were defined empirically, but they can be learnt from data, as proposed in [3]. We have defined the map confidence factor β by ourselves, but ideally, it should be given by the map provider. β describes data currentness (age), errors introduced by geometry simplification and spatial discretisation. β can also be used to depict the localisation accuracy. Other parameters, such as counter steps δ_{inc}, δ_{dec} and thresholds γ_O, $\gamma_\emptyset$ used for mobile object detection determine the sensitiveness of mobile object detection and were set by manual tuning.

3.2 Impact of Prior Knowledge

The results for a particular instant of the approach tested on real-world data are presented on figure 2. The visualisation of the MG has been obtained by calculating the pignistic probability of each class [9]. The presented scene contains two cars (only one is visible in the camera image) going in the direction opposite to the test vehicle and a bus parked on the road edge. Bus and car positions are marked on the grids by green and red boxes, respectively. The test vehicle position is shown as a blue box. Different classes of Ω_{MG} are represented by different colours: F – white,

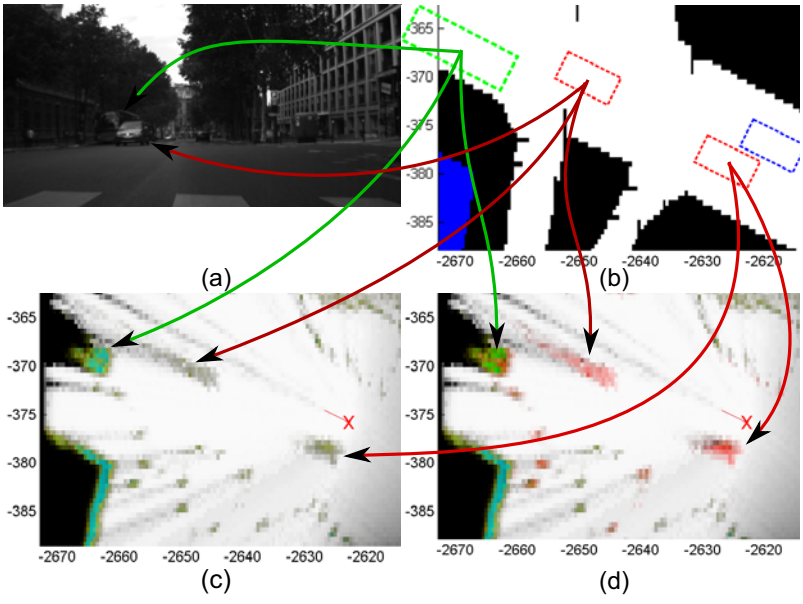


Fig. 2 (a) Scene. (b) PG. (c) MG without prior information. (d) MG with prior map knowledge.

C, N – blue, S – green and V – red. PG on figure 2(b) shows the position of the road space (white) and buildings (blue).

The principal advantage gained by using map knowledge is richer information on the detected objects. A clear difference between a moving object (red, car) and a stopped one (green, bus) is visible. Also, stopped objects are distinct from infrastructure when prior map information is available (cf. figures 2(c) and 2(d). In addition, thanks to the prior knowledge, stationary objects (cyan) such as infrastructure are distinguished from stopped objects on the road. Grids make noticeable the effect of discounting, as information on the environment behind the vehicle is being forgotten. On the other hand, the parked bus is still in evidence despite being occluded by the passing car.

4 Conclusion and Perspectives

A new mobile perception scheme based on prior map knowledge has been introduced. Geographic information is exploited to reduce the number of possible hypotheses delivered by an exteroceptive source. A modified fusion rule taking into account the existence of mobile objects has been defined. Furthermore, the variation in information lifetime has been modelled by the introduction of contextual discounting. In the future, we anticipate removing the hypothesis that the map is accurate. This approach will entail considerable work on creating appropriate

error models for the data source. Moreover, we envision differentiating the free space class into two complementary classes to distinguish navigable and non-navigable space. This will be a step towards the use of our approach in autonomous navigation. Another perspective is the use of reference data to validate the results, choose the most appropriate fusion rule and learn algorithm parameters. We envision using map information to predict object movements. It rests also a future work to exploit fully the 3D map information.

Acknowledgements. This work has been supported by ANR (French National Agency) CityVIP project under grant ANR-07_TSFA-013-01.

References

1. Cappelle, C., et al.: Virtual 3D City Model for Navigation in Urban Areas. J. Intell. Robot. Syst. (2011)
2. Elfes, A.: Using Occupancy Grids for Mobile Robot Perception and Navigation. Computer 22(6), 46–57 (1989)
3. Mercier, D., Quost, B., Denoeux, T.: Refined modeling of sensor reliability in the belief function framework using contextual discounting. J. Inf. Fusion 9(2), 246–258 (2008)
4. Moras, J., Cherfaoui, V., Bonnifait, P.: Credibilist Occupancy Grids for Vehicle Perception in Dynamic Environments. In: IEEE Int. Conf. Robot. Autom., pp. 84–89 (2011)
5. Moras, J., Cherfaoui, V., Bonnifait, P.: Moving Objects Detection by Conflict Analysis in Evidential Grids. In: Int. Veh. Symp., Baden-Baden, Germany, pp. 1120–1125 (2011)
6. OpenStreetMap project, http://www.openstreetmap.org (Cited November 9, 2011)
7. Pagac, D., Nebot, E.M., Durrant-Whyte, H.: An evidential approach to map-building for autonomous vehicles. IEEE Trans. Robot. Autom. 14(4), 623–629 (1998)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
9. Smets, P.: Decision making in the tbm: the necessity of the pignistic transformation. Int. J. Approx. Reason. 38(2), 133–147 (2005)
10. Yager, R.R.: On the Dempster-Shafer framework and new combination rules. Information sciences 4, 93–138 (1987)

Distributed Data Fusion for Detecting Sybil Attacks in VANETs

Nicole El Zoghby, Véronique Cherfaoui,
Bertrand Ducourthial, and Thierry Denœux

Abstract. Sybil attacks have become a serious threat as they can affect the functionality of VANETs (Vehicular Ad Hoc Networks). This paper presents a method for detecting such attacks in VANETs based on distributed data fusion. An algorithm has been developed in order to build distributed confidence over the network under the belief function framework. Our approach has been validated by simulation.

1 Introduction

Exchanging data in a Mobile Ad hoc NETwork (MANET) in a safe manner becomes an important issue. These networks are vulnerable to different attacks such as intrusion. The need for security requires the introduction of the notion of confidence, as each node should have confidence in other nodes or in the received data before using the exchanged information in different applications. By broadcasting messages, nodes will discover their neighborhood. These neighbors can be fake or real nodes, they can also be attackers. Different research papers have been dedicated to find a solution to these problems. Many recent works deal with reputation mechanisms ([20],[1],[9]) and trust evaluation ([16],[17]) to manage the confidence in the source of information. Others were interested in data aggregation without taking into account the source [2][3][10][13].

We propose a method to fuse data in a distributed system in order to build confidence over the network. Nodes broadcast their opinions, which are then used at the reception to evaluate other nodes. Since local opinion is uncertain and incomplete, the use of belief functions to evaluate the received messages seems appropriate. The

Nicole El Zoghby · Véronique Cherfaoui · Bertrand Ducourthial · Thierry Denœux
Heudiasyc UMR CNRS 7253, Université de Technologie de Compiègne, France
e-mail: nicole.el-zoghby@hds.utc.fr,
veronique.cherfaoui@hds.utc.fr,
bertrand.ducourthial@hds.utc.fr,
thierry.denoeux@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 351–358.
springerlink.com

fusion of a node's local knowledge with all the received messages is done by Dempster's rule. The network can suffer from cycles of data dissemination where the same information can be combined many times as it is coming from independent sources [14],[11]. To avoid that, we use the cautious rule of combination [5].

We are interested in studying the confidence in a node for the purpose of detecting sybil attacks in VANETs (Vehicular Ad Hoc NETworks). The sybil attack is the case where a single faulty entity, called a malicious node, can present multiple identities [6] known as sybil nodes or fake nodes. This attack can affect the functionality of the network for the benefit of the attacker. Several techniques have been developed to detect misbehaving or fake nodes in VANETs. Gole et al [7] represented an adversial parsimony that means finding the explanation for corrupted data. Vehicles can distinguish their neighbors by using cameras or exchanging messages in infrared light spectrum. The technique described by Xiao et al. is based on statistic signal strength analysis with the help of roadside infrastructure to detect sybil nodes [18]. Yan et al. [19] used an on-board radar to detect neighbors and to confirm their announced position. Piro et al [12] showed that the sybil attack can be detected passively through single or multiple observers. Due to the dynamics of the vehicular networks, of the number of vehicles and of the lack of permanent infrastructure access, deploying a Public Key Infrastructure in vehicular network (Vehicular PKI) is a very challenging task. As shown in [8], by simply comparing the received signal strength, half of the vehicles can detect the Sybil nodes and it is expected that cooperative techniques would decrease the number of cheated vehicles. Our work proposes such a cooperative algorithm between vehicles, based on the theory of belief functions, and could allow to avoid cryptographic schemes.

In this paper, we develop a distributed fusion technique based on the theory of belief functions. We first describe the system and how we represent the confidence using mass functions. We present the distributed data fusion approach and the proposed algorithm. We validate our approach by simulations and finally we conclude.

2 Distributed Data Fusion Approach

We consider a network composed by nodes exchanging messages. It can be modeled by a directed graph $G = (V,E)$, where V represents the set of nodes $V = \{v_1, v_2, ..., v_n\}$ and E represents the set of edges. The neighbors of each node are represented by $\Gamma(v) = \{v_j \in V, \{v_i, v_j\} \in E\}$. For the sake of simplicity, we suppose that each node knows $n = |V|$. Figure 1 shows an example of network configuration. Each node periodically sends *regular messages* composed of its true identity and geographical position. Moreover, one of the node sends both its regular messages and *fake messages* composed of a forged identity and a forged position. By receiving the fake messages, other nodes are cheated and consider a non existing node, called *fake node* or *Sybil node*. We consider a single malicious node, which creates several Sybil nodes. All nodes use the same transmission system (same antenna, same transmission power). The topology of the network is given by the transmission radio range of the nodes (unit disk graph). We propose a data fusion methodology to

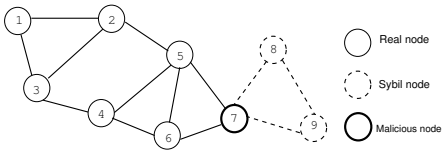


Fig. 1 Network Configuration

combine data exchanged in a mobile ad hoc network, with the aim of quantifying the confidence in the nodes of the network. For this purpose, the messages contain also the sender's confidence in the nodes of the network.

Representing the Confidence by Mass Function: Each node is able to assign a confidence to each other node of the network. This confidence is represented by a basic belief assignment (bba) denoted by m, defined from the frame of discernment $\Omega = \{0,1\}$ where 0 represents *FakeNode* and 1 represents *RealNode*.

We denoted by m_{ij} the corresponding *bba* that represents the opinion of node v_i about node v_j. The bba m_{ij} is defined in Ω by:

$$m_{ij}(\emptyset) = 0 \; ; \; m_{ij}(0) = p_{ij} \; ; \; m_{ij}(1) = q_{ij} \; ; \; m_{ij}(\Omega) = 1 - p_{ij} - q_{ij}. \tag{1}$$

Principle of the Approach: Node v_k sends a message to v_i containing its identity, its coordinates and its opinion about the network. When node v_i receives the message, it calculates, after analyzing the signal strength, what we call a direct confidence. It is a mass vector denoted by $m_{d_{ik}}$. This direct confidence is saved in a local memory called *local knowledge* or *private knowledge*.

Note that each node has two bodies of knowledge: *local knowledge* and *public knowledge*. Local knowledge represents what each node can collect from its neighborhood. It is combined with the public knowledge of other nodes in order to update the public knowledge and rebroadcast it through the network. We thus have a distributed system. Local knowledge depends only on the signal strength of the messages and not on their content: consequently, it cannot be cheated. In contrast, public knowledge is based on the combination of the content of the messages and can be cheated by fake messages. This is why we separate local and public knowledge. The internal memory of each node is thus represented by two mass vectors (arrays of $|V|$ cells initialized at $m(\Omega)$ if $i \neq j$ and $m(1)$ if $i = j$):

$$Kprivate_i(t) = [m_{l_{ij}}^{(t)}] \; ; \quad Kpublic_i(t) = [m_{p_{ij}}^{(t)}]. \tag{2}$$

Distributed Fusion Algorithm: The processing steps performed at the reception are presented in Algorithm 1, and explained hereafter.

Distributed Fusion: When node v_i receives a message, it computes the direct confidence $m_{d_{ik}}$. This confidence is independent of previous messages and it is not the result of any other combination. So we use it to update the receiver's local knowledge

Algorithm 1. Received Message Processing on node v_i

Require: message from v_k to v_i, the signal strength P, message contains $m_{p_{kj}} \forall j$

Ensure: $Kprivate_i$=$[m_{l_{ij}}^{(t)}]$ and $Kpublic_i$=$[m_{p_{ij}}^{(t)}] \forall j \in V$

$m_{d_{ik}}^{(t)} \leftarrow DirectConfidence(message, P)$

$m_{l_{ik}}^{(t)} \leftarrow UpdateLocalKnowledge(m_{l_{ik}}^{(t-1)}, m_{d_{ik}}^{(t)})$

$m_{p_{ik}}^{(t)} \leftarrow UpdatePublicKnowledge(m_{p_{ik}}^{(t-1)}, m_{l_{ik}}^{(t)})$

$\alpha \leftarrow DiscountingFactor(m_{l_{ik}})$

for *each node* $j \in V$ *such as* $j \neq i$, $j \neq k$ **do**

$\quad {}^{\alpha}m_{p_{kj}}^{(t)} \leftarrow DiscountTransmitterKnowledge(\alpha, m_{pkj}{}^{(t)}, m_{\Omega}^{(t)})$

$\quad m_{p_{ij}}^{(t)} \leftarrow UpdatePublicKnowledge(m_{p_{ij}}^{(t-1)}, {}^{\alpha}m_{p_{kj}}^{(t)})$

about the transmitter by Dempster's rule [4]. The function UpdateLocalKnowledge $(m_{l_{ik}}^{(t-1)}, m_{d_{ik}}^{(t)})$ is calculated as:

$$m_{l_{ik}}^{(t)} = m_{l_{ik}}^{(t-1)} \oplus m_{d_{ik}}^{(t)}, \tag{3}$$

where $\oplus$ denotes Dempster's rule. Since fake nodes might falsify the opinion of each node, the knowledge of other nodes is needed. To this end we use a distributed fusion to collect other opinions. As we consider that the transmitter is not totally reliable, we discount its opinion before combining it with the receiver's knowledge. The discounting factor $\alpha = 1 - m_{l_{ik}}(1)$ is defined as the plausibility that the transmitter is unreliable. The transmitter's opinion is discounted with the function DiscountTransmitterKnowledge$(\alpha, m_{kj}^{(t)}, m_{\Omega}^{(t)})$ as follows:

$${}^{\alpha}m_{p_{kj}}^{(t)} = (1-\alpha).m_{p_{kj}}^{(t)} + \alpha.m_{\Omega}^{(t)}. \tag{4}$$

To update the receiver's public knowledge, we use the cautious rule [5]. In a distributed system, the same information can be received and treated many times. While combining the knowledge, it is useful to use an idempotent rule to avoid counting the same information several times (data incest) as if it is provided by different independent sources. So the function UpdatePublicKnowledge$(m_{p_{ij}}^{(t-1)}, {}^{\alpha}m_{p_{kj}}^{(t)})$ allows us to combine the receiver's public knowledge with the transmitter's discounted knowledge about its neighbors as follows:

$$m_{p_{ij}}^{(t)} = m_{p_{ij}}^{(t-1)} \circledwedge {}^{\alpha}m_{p_{kj}}^{(t)}, \tag{5}$$

where $\circledwedge$ denotes the cautious rule.

Direct Confidence: Different methods can be used to compute the direct confidence $m_{d_{ik}}$. We propose a method that allows us to convert a real measure into a mass function. The real measure is based on signal strength analysis. Each receiver can analyze the signal strength to detect if the announced position is the real one [8]. It measures the strength of the received signal and calculates a theoretical value

in terms of the node's coordinates. The estimated value of the signal strength is calculated by the Friis formula as $\mu = P_e.G_{SR}/d_{ik}^2$, where

- P_e is the transmitted signal power, depending on the transmitter antenna;
- $G_{SR} = \frac{G_t.G_r.\lambda^2}{16.\pi^2}$ is the antenna gain, G_t and G_r are the gains of the transmit antenna and the receive antenna, respectively, and λ is the wavelength;
- d_{ik} is the distance between the transmitter node v_k and the receiver node v_i.

The comparison between the estimated power and the theoretical one allows the detection of a misbehavior. We propose to compute the plausibility that the received signal power P is equal to x, given that the transmitting node is a true node ($\omega = 1$) as follows:

$$pl(P = x/\omega = 1) = \frac{f(x/\omega=1)}{\sup_{x' \in \mathbb{R}}(f(x'/\omega=1)}\ , \tag{6}$$

where $f(x/\omega = 1)$ is the normal density function with mean μ and standard deviation σ depending on the receiver antenna.

The plausibility $pl(P = x/w = 0)$ is defined as shown in Figure 2: if the estimated and the theoretical powers are equal, we leave the possibility that the transmitter can be a fake node. Indeed, if the transmitter is a fake node but its position is near the malicious node therefore the estimated position will be approximately equal to the measured position. This result can influence the detection of the fake node.

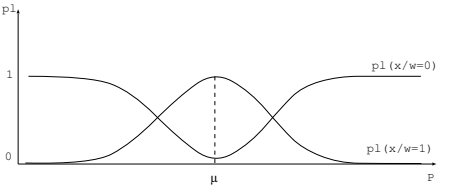


Fig. 2 Plausibility of received power values for true ($\omega = 1$) and fake ($\omega = 0$) nodes

The direct confidence is computed using the Generalized Bayes theorem [15]. It is obtained by combining the prior knowledge about the transmitter m_0^{Ω} with the plausibility that the node is a fake node knowing that it is a real $\{0\}^{pl(x/w=1)}$ and the plausibility that the node is a real node knowing that it is a fake $\{1\}^{pl(x/w=0)}$:

$$m_{d_{ik}}^{(t)} = m^{\Omega}(./x) = m_0^{\Omega} \textcircled{\cap} \{0\}^{pl(x/w=1)} \textcircled{\cap} \{1\}^{pl(x/w=0)}, \tag{7}$$

where $\textcircled{\cap}$ denotes the unnormalized Dempster's rule.

3 Results

In order to validate our approach, Algorithm 1 has been implemented in Matlab. Simulations were performed on static and dynamic network. For simplicity of analysis, we first assumed all nodes in the network to be static. We performed simulations on different random network configurations. Next we tested our approach on

a dynamic network, where nodes were moving in the same direction following a highway scenario.

Implementation: In this part we will represent an example of a network composed from six true nodes, one of them is a malicious node that creates three fake nodes. The transmitted signal Power P_e is about 600 mW and the antenna ranges is in order of 400 m. We consider that each transmitter sends its *id*, its *position* and its *public knowledge*. The receiver uses these informations to perform all the calculations and to verify if the node is true or fake. Simulations are performed until the convergence of the algorithm. We consider that the algorithm has converged when $|m_{ij}^{(t-1)} - m_{ij}^{(t)}| < \varepsilon$, where ε is a defined small threshold. The results of the simulation will be represented by gray scale matrices.

Static Network: We present in Figure 3 an example of a network configuration where the nodes are static (left figure) and the result of the simulation (right figure). The white color in the right figure corresponds to a mass equal to 1 representing true nodes. The black color correspond to a mass equal to 0 representing fake nodes. The malicious node 3 will try to convince other nodes that the fake nodes (7,8,9) are true nodes. The fake nodes have the same opinion as the malicious node. The first part of the rightmost figure represents the private knowledge. Each node has only information about its neighbors. The second part represents the public knowledge. We see that $m_{p_{ij}}(\{1\}) = 0$ for $i = \{1,2,4,5,6\}$ and $j = \{7,8,9\}$, which means that the true nodes have correctly identified nodes $\{7,8,9\}$ as untrustworthy.

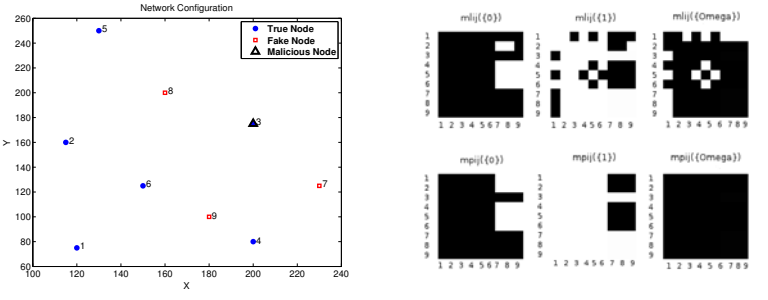


Fig. 3 Network configuration and simulation results.

To verify the convergence of the algorithm, we performed simulations on different random network configurations by changing the number of the fakes nodes. Table 1 represents the result with different proportions of fake nodes. Each iteration represents the simulation of the process of a message. It needs more time to converge when the proportion of the fake nodes is greater. Our approach can detect sybil nodes with different static configurations.

Dynamic Network: Static configurations have some limits, especially when a malicious node is not in the neighborhood of the true nodes: in that case, fake nodes cannot be detected. So, we simulated a dynamic scenario in which nodes move in

Table 1 Results with different nodes configurations

Nodes Configurations	Average of the number of iterations[a]	Standard deviation
True Nodes=6 Fake Nodes =3	207.05	7.86
True Nodes=6 Fake Nodes =4	227.55	6.89
True Nodes=6 Fake Nodes =5	255.8	6.33
True Nodes=6 Fake Nodes =6	304.7	7.55

[a] These results represent the average of 20 simulations.

Table 2 Results for dynamic networks with different node configurations

Nodes Configurations	Average of the number of iterations[a]	Standard deviation
True Nodes=6 Fake Nodes =3	119.3	45.88
True Nodes=6 Fake Nodes =4	274.4	40.96
True Nodes=6 Fake Nodes =5	361.1	54.23
True Nodes=6 Fake Nodes =6	376.3	32.05

[a] These results represent the average of 10 simulations.

the same direction as on a highway. While moving, the neighborhood of each node changes. It influences the private knowledge because it depends on the neighborhood. Thanks to public knowledge, each node can get information about the whole network and can quantify its confidence. Table 2 shows results for different dynamic network configurations. The number of iterations until convergence changes at each simulation, because the node motions and neighborhoods are random. These preliminary results suggest that true nodes can successfully detect fake nodes in the network while moving on a highway.

4 Conclusion

A distributed data fusion approach based on belief functions for detecting sybil attacks in VANETs has been developed. The method uses both Dempster's rule and cautious rule to combine information and to compute a distributed confidence over the network. The results are promising and demonstrate that we can determine the reliability of nodes and detect fake nodes in a VANET. More realistic scenarios are currently being studied using an ad hoc network simulator.

The method presented in this paper computes the confidence in the nodes without taking into account the contents of the messages exchanged in the network. The joint analysis of information and node reliability is currently being investigated. Results along these lines will be reported in future publications.

References

1. Singh, M.P., Yu, B.: An evidential model of distributed reputation management. In: First international Joint Conference on Autonomous Agents and Multi-Agents Systems, Bologna, Italy, pp. 294–301. ACM Press (2002)

2. Chen, T.M., Venkataramanan, V.: Dempster-shafer theory for intrusion detection in ad hoc networks. IEEE Internet Computing 9, 35–41 (2005)
3. Cherfaoui, V., Denoeux, T., Cherfi, Z.L.: Distributed data fusion: application to confidence management in vehicular networks. In: 11th Int. Conf. on Information Fusion, Germany, pp. 846–853 (2008)
4. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artificial Intelligence 172, 234–264 (2008)
6. Douceur, J.R.: The sybil attack. In: The International Workshop on Peer to Peer Systems, Cambridge, MA, USA, pp. 251–260 (2002)
7. Golle, P., Greene, D., Staddon, J.: Detecting and correcting malicious data in vanets. In: 1st ACM Workshop on Vehicular Ad hoc Networks (VANET), New York, NY, USA, pp. 29–37 (2004)
8. Guette, G., Ducourthial, B.: On the sybil attack detection in vanet. In: International Workshop on Mobile Vehicular Networks (MoveNet 2007), co-located with IEEE MASS 2007, Pisa (October 2007)
9. Liu, J., Issarny, V.: Enhanced reputation mechanism for mobile ad hoc networks. In: 2nd International Conference on Trust Management, Oxford, UK, pp. 48–62 (2004)
10. Lochert, C., Scheuermann, B., Mauve, M.: Probabilistic aggregation for data dissemination in vanets. In: 4th ACM international Workshop on Vehicular Ad Hoc Networks, Montreal, QC, Canada, pp. 1–8 (2007)
11. Mitchell, H.B.: Multisensor Data Fusion: An introduction. Springer (2007)
12. Piro, C., Shields, C., Levine, B.N.: Detecting the sybil attack in mobile ad hoc networks. In: IEEE/ACM Intl Conf on Security and privacy in Communication Networks (SecureComm), pp. 1–11 (August 2006)
13. Raya, M., Papadimitratos, P., Gligor, V.D., Hubaux, J.-P.: On data-centric trust establishment in ephemeral ad hoc networks. In: The 28th IEEE Conference on Computer Communications (INFOCOM), Phoenix, AZ, USA, pp. 1238–1246 (April 2008)
14. Evans, R.J., Mclaughlin, S., Krishnamurthy, V.: Bayesian network model for data incest in a distributed sensor network. In: The 7th International Conference on Information Fusion, Stockholm, Sweden, vol. 1 (2004)
15. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9, 1–35 (1993)
16. Theodorakopoulos, G., Baras, J.S.: Trust evaluation in ad-hoc networks. In: ACM Workshop Wireless Security, Philadelphia, PA, USA, pp. 1–10 (2004)
17. Wang, J., Sun, H.-J.: A new evidential trust model for open communities. Computer Standards & Interfaces 31, 994–1001 (2009)
18. Xiao, B., Yu, B., Gao, C.: Detection and localization of sybil nodes in vanets. In: The Workshop on Dependability Issues in Wireless Ad Hoc Networks and Sensor Networks, Los Angeles, CA, USA, pp. 1–8 (2006)
19. Yan, G., Choudhary, G., Weigle, M., Olariu, S.: Providing vanet security through active position detection. Computer Communications: Special Issue on Mobility Protocols for ITS/ VANET 31(12), 2883–2897 (2008)
20. Zacharia, G., Maes, P.: Trust management through reputation mechanisms. Applied Artificial Intelligence 14, 881–907 (2000)

Partially-Hidden Markov Models

Emmanuel Ramasso, Thierry Denœux, and Noureddine Zerhouni

Abstract. This paper addresses the problem of Hidden Markov Models (HMM) training and inference when the training data are composed of feature vectors plus uncertain and imprecise labels. The "soft" labels represent partial knowledge about the possible states at each time step and the "softness" is encoded by belief functions. For the obtained model, called a Partially-Hidden Markov Model (PHMM), the training algorithm is based on the Evidential Expectation-Maximisation (E2M) algorithm. The usual HMM model is recovered when the belief functions are vacuous and the obtained model includes supervised, unsupervised and semi-supervised learning as special cases.

1 Introduction

Hidden Markov Models (HMM) are powerful tools for sequential data modelling and analysis. Many applications for several decades have found solutions based on HMM such as discovering word sequences based on speech audio recordings [9], gene finding based on a DNA sequence [8], and performing prognostics and health detection of ball bearings degradation based on noisy sensors [6, 10]. In the sequel, we consider sequential data taking the form of a time-series of length T where each element is a multidimensional feature vector $x_t \in \Re^F, t = 1 \ldots T$ also called vector

Emmanuel Ramasso · Noureddine Zerhouni
FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM,
Automatic Control and Micro-Mechatronic Systems Department, 24 rue Alain Savary,
F-25000 Besançon, France
e-mail: emmanuel.ramasso@femto-st.fr,
noureddine.zerhouni@ens2m.fr

Thierry Denœux
Université de Technologie de Compiègne, Heudiasyc, U.M.R. C.N.R.S. 7253,
Centre de Recherches de Royallieu, B.P. 20529, F-60205 Compiègne Cedex,
France, Address of Institute
e-mail: Thierry.Denoeux@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 359–366.
springerlink.com

of observations [9]. The modelling part assumes that the system (a speaker, a DNA sequence or a ball bearing) generating the time-series is a Markov process with unobserved (hidden, latent) discrete states. In HMMs, the states are not visible but when the system is entering one of the states, the features follow a particular probability distribution. The sequence of observations thus provides information about the sequence of states. One of the most powerful characteristics of HMMs, accounting for its wide range of applications, is the possibility to estimate the parameters efficiently and automatically. Given a training dataset composed of the observed data $\mathbf{X} = \{x_1, \ldots, x_t, \ldots, x_T\}$ (where x_t can be continuous or discrete), and denoting by K the number of hidden states such that the state variable y_t at time t can take a value in

$$\Omega_{\boldsymbol{Y}} = \{1, \ldots, j, \ldots, K\} \ , \tag{1}$$

the following parameters have to be estimated:

- $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_j, \ldots, \boldsymbol{\phi}_K\}$ is the set of parameters characterising the probability distribution of observations given each state:

$$b_j(x_t) = P(x_t | y_t = j; \boldsymbol{\phi}_j), j = 1 \ldots K \tag{2}$$

- $\boldsymbol{A} = [a_{ij}]$ with

$$a_{ij} = P(y_t = j | y_{t-1} = i), i = 1 \ldots K, j = 1 \ldots K, \tag{3}$$

  that is the probability of the system to be in state j at time-instant t, given that the system was in state i at $t-1$, with $\sum_j a_{ij} = 1$.
- $\boldsymbol{\Pi} = \{\pi_1, \ldots, \pi_j, \ldots, \pi_K\}$, where

$$\pi_j = P(y_1 = j) \tag{4}$$

  is the probability of state j at $t = 1$, such that $\sum_j \pi_j = 1$.

In the sequel, all these parameters are aggregated in a vector θ:

$$\theta = \{\mathbf{A}, \boldsymbol{\Pi}, \boldsymbol{\Phi}\} \ . \tag{5}$$

These parameters can be estimated using an iterative procedure called the Baum-Welch algorithm [1, 9] and relying on the Expectation-Maximisation process.

There are applications where some observations x_t in the training data $\mathbf{X}$ are associated to a label that actually represents the state at time t. Instead of considering the labelling process as a binary one, where states can be known or unknown, we address the problem of partially-supervised HMM training, assuming partial knowledge about the states to be available and represented by belief functions.

The contribution of this paper holds in the development of a model called Partially-Hidden Markov Model (PHMM) that manages partial labelling of the training dataset in HMMs. Compared to [3], we take into account the temporal dependency into account, helping in time-series modelling. The proposed approach is based on the Evidential Expectation-Maximisation (E2M) algorithm introduced in [5].

2 Partially-Hidden Markov Models (PHMM)

Given the observation sequence $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$, there are three main problems of interest in connection with HMMs [9]:

Problem 1: Given a model $\theta = \{\boldsymbol{\Pi}, \boldsymbol{A}, \boldsymbol{\Phi}\}$, how to compute its likelihood $L(\theta; \mathbf{X})$?
Problem 2: Given a model θ, how to choose the state sequence $\mathbf{Y}^* = \{y_1^*, y_2^*, \ldots, y_T^*\}$ that best explains observations?
Problem 3: How to estimate parameters $\theta = \{\boldsymbol{\Pi}, \boldsymbol{A}, \boldsymbol{\Phi}\}$ of a model?

These problems have been solved in different ways for some decades for HMMs [9]. In the sequel, we present the solutions for the case where partial information on states is available in the form of a set of belief functions m defined on the set of states $\Omega_{\boldsymbol{Y}}$. States are then "partially hidden" and the case of completely hidden states is recovered when all the masses are vacuous.

The main idea behind the solutions of partially-supervised training in statistical models is to combine the probability distributions on hidden variables with the belief masses m. This combination can be computed from the contour function pl associated to m.

The next paragraph describes the main features of the E2M algorithm in order to introduce the conditioning process that plays a central role in solutions for problems 1, 2 and 3. The E2M algorithm will be used in the last paragraph dedicated to parameter estimation in PHMMs.

2.1 Generalized Likelihood Function and E2M Algorithm

The Evidential EM (E2M) algorithm [5] is an iterative procedure dedicated to maximum likelihood estimation in statistical models based on uncertain observations encoded by belief functions. As for the usual EM algorithm, the E2M algorithm does not maximise directly the observed-data likelihood function denoted here $L(\theta; \mathbf{X}, m)$ but it focuses instead on a lower bound called the auxiliary function [2], and usually denoted by Q and defined as:

$$Q(\theta, \theta^{(q)}) = \mathbb{E}_{\theta^{(q)}} \left[\log L(\mathbf{X}, \mathbf{Y}; \theta) | \mathbf{X}, pl \right] \quad , \tag{6}$$

where pl denotes the contour function associated to m, $\theta^{(q)}$ is the fit of parameter θ at iteration q and Q represents the conditional expectation of the complete-data log-likelihood. In the E-step of the E2M algorithm, the conditional expectation in the auxiliary function Q is taken with respect to $\gamma' \stackrel{\text{def}}{=} P(\cdot|\mathbf{X}, pl; \theta^{(q)}) = P(\cdot|\mathbf{X}; \theta^{(q)}) \oplus pl$, that is the combination of the expectation, denoted γ_t, with the plausibilities using Demspter's rule [4, 5]. The new expectation is then defined for each state j at time t by $\gamma_t'(j|pl; \theta^{(q)}) = P(y_t = j|\mathbf{X}, pl; \theta^{(q)})$:

$$\gamma_t'(j|pl; \theta^{(q)}) = \frac{\gamma_t(j; \theta^{(q)}) \cdot pl_t(j)}{L(\theta^{(q)}; \mathbf{X}, pl)} \tag{7}$$

and the auxiliary function becomes:

$$Q(\theta,\theta^{(q)}) = \frac{\sum_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X};\theta^{(q)}) \cdot pl(\mathbf{Y}) \cdot \log L(\mathbf{X},\mathbf{Y};\theta)}{L(\theta^{(q)};\mathbf{X},pl)} \quad . \tag{8}$$

The M-step is similar to to that of the usual EM algorithm and consists in maximising Q with respect to θ. The maximisation is ensured to increase the likelihood of observed data since E2M inherits the monotonicity of EM as for any sequence $L(\theta^{(q)};\mathbf{X},pl)$, $q = 1,2\ldots$, we have $L(\theta^{(q+1)};\mathbf{X},pl) \geq L(\theta^{(q)};\mathbf{X},pl)$.

2.2 Solution to Problem 1 in PHMM

Using a similar process as in usual HMM (see [2] for details on HMM), the marginal posterior distribution on latent variables for the set of parameters $\theta^{(q)}$ at iteration q of E2M can be rewritten as:

$$\gamma_t^{'} = P(y_t|\mathbf{X};\theta^{(q)}) \oplus pl_t = \alpha_t^{'} \cdot \beta_t \tag{9}$$

with $\alpha_t^{'} \stackrel{\text{def}}{=} P(\mathbf{X}_{1:t}, y_t|pl;\theta^{(q)})$ and $\beta_t^{'} \stackrel{\text{def}}{=} P(\mathbf{X}_{t+1:T}|y_t;\theta^{(q)})$. The definition of β remains the same as in the standard algorithm with $\beta_t(i;\theta^{(q)}) = \sum_j \beta_{t+1}(j;\theta^{(q)}) \cdot b_j(x_{t+1}) \cdot a_{ji}, \quad t = 2\ldots T$ starting from $\beta_T(i;\theta^{(q)}) = 1, \forall i$. The probability of jointly observing a sequence $\mathbf{X}_{1:t}$ up to t and state j at time t given the parameters and the uncertain data is given by the modified forward variable $\alpha_t^{'}$ such that $\alpha_t^{'}(j;\theta^{(q)}) = P(\mathbf{X}_{1:t}, y_t = j|pl;\theta^{(q)})$ with:

$$\alpha_t^{'}(j;\theta^{(q)}) = \frac{\alpha_t(j;\theta^{(q)}) \cdot pl_t(j)}{L(\theta^{(q)};\mathbf{X},pl)} \tag{10}$$

and therefore

$$\gamma_t^{'}(j;\theta^{(q)}) = \frac{\alpha_t(j;\theta^{(q)}) \cdot pl_t(j) \cdot \beta_t(j;\theta^{(q)})}{L(\theta^{(q)};\mathbf{X},pl)} \quad . \tag{11}$$

Variables α and β are the same as in HMM [2].

Summing Eq. 11 over latent variables gives the observed data likelihood. Therefore, to assess the likelihood function $L(\theta^{(q)};\mathbf{X},pl)$ at the current iteration of the E2M algorithm, we simply need to choose a time index t. A good candidate is the index T since in this case we do not need to evaluate β_T (that equals to 1) reducing the computation load:

$$L(\theta^{(q)};\mathbf{X},pl) = \sum_{j=1}^{K} \alpha_T(j;\theta^{(q)}) \cdot pl_T(j). \tag{12}$$

Practically, we can use the normalization process proposed in [9] in order to cope with the limited machine precision range.

2.3 Solution to Problem 2 in PHMM

The Viterbi algorithm [7] was defined in order to retrieve the best sequence of hidden states within the noisy observations. The best sequence is found in $K^2 \times T$ operations (instead of K^T for a greedy search) and is ensured to be the one with the highest likelihood. Given the observed data $\mathbf{X}$, the Viterbi algorithm finds the maximum a posteriori (MAP) sequence $\mathbf{Y}^* = \{y_1^*, \ldots, y_t^*, \ldots, y_T^*\}, y_t^* \in \Omega_{\mathbf{Y}}$. In PHMM, the MAP criterion is modified by taking soft labels into account, i.e., $P(\mathbf{Y}^*|\mathbf{X}, pl; \theta^{(q)})$ or, equivalently, $\log P(\mathbf{X}, \mathbf{Y}^*|pl; \theta^{(q)})$. In HMMs, the Viterbi algorithm is called the max-sum product algorithm and it is equivalent to a forward propagation with conditioning at each time-step by the potential predecessors of each state. In PHMMs, a similar reasoning can be applied where conditioning (by singletons states) naturally leads to the use of plausibilities. The MAP criterion can be written as:

$$\delta_t^{'}(j;\theta^{(q)}) = \max_i \left[\delta_{t-1}^{'}(i;\theta^{(q)}) \cdot a_{ij}\right] \cdot b_j(x_t) \cdot pl_t(j), \quad t = 2 \ldots T \tag{13}$$

starting from $\delta_1^{'}(j;\theta^{(q)}) = \pi_j \cdot pl_t(j) \cdot b_j(x_1)$. Keeping track of the argument maximising this expression as $\psi_t^{'}(j) = \text{argmax}_i \left[\delta_{t-1}^{'}(i;\theta^{(q)}) \cdot a_{ij}\right]$, the backtracking of the best state sequence ending in $y_t^* = j$ at time t is given by $y_{t-1}^* = \psi_t^{'}(y_t^*)$.

2.4 Solution to Problem 3 in PHMM

In the E2M algorithm, the auxiliary function is given by Eq. (8). In order to define the maximisation step, the Q-function has to be computed. For that purpose, we introduce the multinomial representation of variables such that $y_{tj} = 1$ if state j at time t is true, else $y_{tj} = 0$. Then, we can write:

$$P(\mathbf{Y}, \mathbf{X}; \theta) = \left(\prod_{j=1}^{K} \pi_j^{y_{1j}}\right) \cdot \left(\prod_{t=2}^{T}\prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{y_{t-1,i}, y_{tj}}\right) \cdot \left(\prod_{t=1}^{T}\prod_{j=1}^{K} b_j(x_t)^{y_{tj}}\right). \tag{14}$$

Taking the logarithm of the above expression leads to the complete-data log-likelihood. In this paper, partial knowledge on y_{tj} is assumed to be represented by a belief function (and in particular by its contour function $pl_t(j), \forall t = 1 \ldots T, j = 1 \ldots K$). The auxiliary function Q thus becomes:

$$Q(\theta, \theta^{(q)}) = \mathbb{E}_{\theta^{(q)}}\left[\log P(\mathbf{X}, \mathbf{Y}; \theta)|\mathbf{X}, pl\right] \tag{15a}$$

$$= Q_\pi(\theta, \theta^{(q)}) + Q_{\mathbf{A}}(\theta, \theta^{(q)}) + Q_{\boldsymbol{\Phi}}(\theta, \theta^{(q)}) \ , \tag{15b}$$

with $Q_\pi(\theta,\theta^{(q)}) = \sum_{j=1}^{K} \mathbb{E}_{\theta^{(q)}}\left[y_{1j}|\mathbf{X},pl\right]\cdot\log\pi_j$ given by:

$$Q_\pi(\theta,\theta^{(q)}) = \sum_{j=1}^{K} \gamma_1'(j;\theta^{(q)})\cdot\log\pi_j \quad , \tag{16}$$

$Q_{\mathbf{A}}(\theta,\theta^{(q)}) = \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{E}_{\theta^{(q)}}\left[y_{t-1,i}y_{tj}|\mathbf{X},pl\right]\cdot\log a_{ij}$ with:

$$Q_{\mathbf{A}}(\theta,\theta^{(q)}) = \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} \xi_{t-1,t}'(i,j;\theta^{(q)})\log a_{ij} \quad , \tag{17}$$

and $Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)}) = \sum_{t=1}^{T}\sum_{j=1}^{K} \mathbb{E}_{\theta^{(q)}}\left[y_{tj}|\mathbf{X},pl\right]\cdot\log b_j(x_t)$ given by:

$$Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)}) = \sum_{t=1}^{T}\sum_{j=1}^{K} \gamma_t'(j;\theta^{(q)})\cdot\log b_j(x_t) \quad . \tag{18}$$

In the above expressions we have:

$$\gamma_t'(j;\theta^{(q)}) = \frac{\gamma_t(j;\theta^{(q)})\cdot pl_t(j)}{\sum_{l=1}^{K}\gamma_t(l;\theta^{(q)})\cdot pl_t(l)} \quad , \tag{19}$$

which is the marginal posterior distribution of a latent variable y_j at t given pl, and

$$\xi_{t-1,t}'(i,j;\theta^{(q)}) = \frac{\xi_{t-1,t}(i,j;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(j)}{\sum_{l=1}^{K}\xi_{t-1,t}(i,l;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(l)} \tag{20}$$

is the joint probability of two consecutive latent variables $y_{t-1,i}$ and y_{tj} given pl. The optimal parameters at each iteration of E2M are given by using a similar reasoning as in the standard algorithm, but the posterior probability over latent variables now depends on the plausibilities:

$$\pi_j^{(q+1)} = \frac{\gamma_1(j;\theta^{(q)})\cdot pl_1(j)}{\sum_{l=1}^{K}\gamma_1(l;\theta^{(q)})\cdot pl_1(l)} \tag{21a}$$

$$a_{ij}^{(q+1)} = \frac{\sum_{t=2}^{T}\xi_{t-1,t}(i,j;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(j)}{\sum_{t=2}^{T}\sum_{l=1}^{K}\xi_{t-1,t}(i,l;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(l)}. \tag{21b}$$

The maximisation of $Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)})$ depends on the form of the distribution of observations given the latent variable j.

3 Partial Results, Conclusion and Further Work

Partial results: To illustrate this approach, we considered that observations can be modelled by mixtures of Gaussians. We proceeded as in standard HMM to derive the M-step in PHMM and to estimate the parameters of the distributions. Equations are however not reported in this paper for lack of space.

For illustration purpose, we used the dataset of the PHM'08 data challenge [12] concerning the health state of a turbofan engine. It was manually segmented into four states (to evaluate the results) such that each time-series is accompanied by a set of labels reflecting the current state of the fan, that is normal, transition, degrading or faulty mode. Each label corresponds to a mass function focused on a singleton, except in the transitions where doubt between two labels is defined[1]. The BBA were then transformed into plausibilities. For these tests, we corrupted them by additive noise: $pl_t(j) \leftarrow pl_t(j) + \sigma_k \cdot \varepsilon_t(j)$, where $\sigma_k \in \{0, 0.1, \ldots, 1\}$ and $\varepsilon_t(j) \sim \mathscr{U}_{[0,1]}$ was drawn from a uniform distribution. For each noise level, we considered the influence of the number of unlabelled data $v_k \in \{0\%, 10\%, \ldots, 100\%\}$. The partitions of time-series in the testing dataset estimated by HMM and PHMM using the Viterbi algorithm as defined in HMM (since we do not know the labels for the testing) were compared using the Folkes and Mallows index ($F \in [0,1]$) [11]. Positive values of the relative performance improvement index $G = F_{pshmm}/F_{hmm} - 1$ indicate that the proposed PHMM provided a better segmentation of the time-series into states.

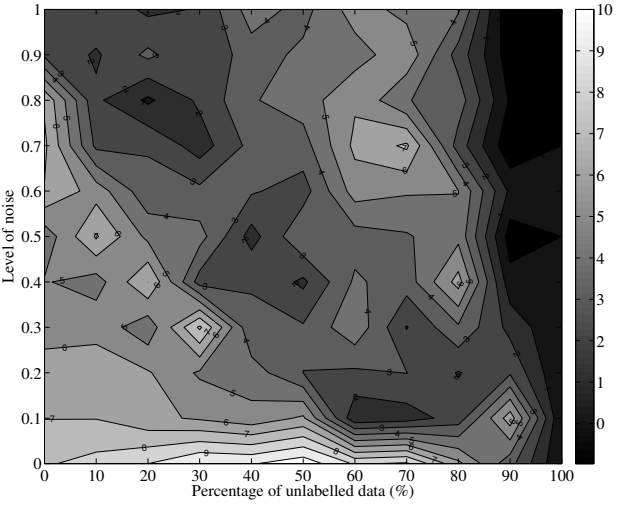


Fig. 1 Performance (G-index): median value over 10 runs with different initialisation. Positive value reflects an improvement provided by PHMM. Here almost all values are positive except darkest areas.

[1] The segmentation and the associated BBA are available at `http://www.femto-st.fr/~emmanuel.ramasso/PEPS_INSIS_2011_PHM_by_belief_functions.html`.

The plot of G as a function of the percentage of unlabelled data and noise level shown in Figure 1 shows an improvement by several percents when using the proposed PHMM (up to 12%). When all data were unlabelled and with no noise (bottom right hand-side corner), both models provided exactly the same results, as expected. When the noise increased, the performance decreased but was still higher than that of the standard HMM. The most difficult cases were encountered when the noise was high (top of figure), where PHMM improvements were between $[2\%, 5\%]$.

Conclusion and Further Work: Taking partial knowledge into account is of crucial importance in many statistical models. Encoding prior information by belief functions leads to simple modifications of the initial estimation formula while remaining theoretically sound. The statistical model considered in this paper was the Hidden Markov Models. Further work remains to be done in order to compute in developing re-estimation formula for various distributions of observations given latent states.

Acknowledgements. This work was partially supported by a PEPS-INSIS-2011 grant from the French National Center for Scientific Research (CNRS) under the administrative authority of the French Ministry of Research.

References

1. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in statistical analysis of probabilistic functions of markov chains. Ann. Math. Stat. 41, 164–171 (1970)
2. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
3. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recognition 42, 334–348 (2009)
4. Dempster, A.: Upper and lower probabilities induced by multiple valued mappings. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Transactions on Knowledge and Data Engineering (2011), doi:10.1109/TKDE.2011.201
6. Dong, M., He, D.: A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. Mechanical Systems and Signal Processing 21, 2248–2266 (2007)
7. Forney, G.: The viterbi algorithm. Proceedings of the IEEE 61(3), 268–278 (1973)
8. Murphy, K.P.: Dynamic Bayesian networks: Representation, inference and learning. Ph.D. thesis, UC Berkeley (2002)
9. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77, 257–285 (1989)
10. Ramasso, E.: Contribution of belief functions to hidden markov models. In: IEEE Workshop on Machine Learning and Signal Processing, Grenoble, France, pp. 1–6 (2009)
11. Saporta, G., Youness, G.: Comparing two partitions: Some proposals and experiments. In: COMPSTAT (2002)
12. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: Int. Conf. on Prognostics and Health Management, Denver, CO, USA, pp. 1–9 (2008)

Large Scale Multinomial Inferences and Its Applications in Genome Wide Association Studies

Chuanhai Liu and Jun Xie

Abstract. Statistical analysis of multinomial counts with a large number K of categories and a small number n of sample size is challenging to both frequentist and Bayesian methods and requires thinking about statistical inference at a very fundamental level. Following the framework of Dempster-Shafer theory of belief functions, a probabilistic inferential model is proposed for this "large K and small n" problem. Using a data-generating device, the inferential model produces probability triplet (p,q,r) for an assertion conditional on observed data. The probabilities p and q are *for* and *against* the truth of the assertion, whereas $r = 1 - p - q$ is the remaining probability called the probability of "don't know". The new inference method is applied in a genome-wide association study with very-high-dimensional count data, to identify association between genetic variants to a disease Rheumatoid Arthritis.

1 Introduction

Statistical analysis of multinomial counts with a large number K of categories and a small number n of sample size is a challenging problem for both frequentist and Bayesian methods. For Bayesian methods, it is well known that in this case Bayesian priors have tremendous effects on the final inferential results. Frequentist methods, such as the χ^2-test for contingency table, suffer from the problem of small or zero counts. It is not uncommon that frequentist methods are applied to modified contingency tables obtained by either deleting or combining categories with small observed counts (*e.g.*, zeros and ones).

This "large K and small n" problem is motivated by genome-wide association studies with very-high-dimensional count data, *i.e.*, single nucleotide polymorphisms (SNPs) data. SNPs are major genetic variants that may associate with

Chuanhai Liu · Jun Xie
Department of Statistics, Purdue University, 250 N. University St.,
West Lafayette, IN 47907
e-mail: chuanhai@purdue.edu, junxie@purdue.edu

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 367–374.
springerlink.com

common diseases such as cancer and heart disease. A SNP has three possible genotypes, wild type homozygous, heterozygous, and mutation (rare) homozygous. As a simple example, we compare differences in allele frequencies of a set of SNPs between cases and controls to identify association with a disease. Commonly used statistical analyses in genetic association studies, based on unrelated individuals, include logistic regression and χ^2 tests of association. However, the conventional statistical methods only work for a single SNP or a very small number of SNPs. If we consider a block of SNPs, for example, in exploratory data analysis and model checking or validation, even a moderate size 10 results in $3^{10} = 59,049$ possible genotypes. This number of categories is much larger than a typical study size of a few thousands subjects. Therefore, most of categories will have zero or one observation. The familiar logistic regression and χ^2 tests are not appropriate any more in such a case.

Recently, Martin, Zhang, and Liu [6, 9] proposed what they called weak beliefs for probabilistic inference, based on an extension of the framework of Dempster-Shafer theory of belief functions [7, 1]. Inferential models using weak beliefs were used to produce valid probabilities, in terms of long-run frequency, *for* and *against* the truth of assertions of interest (see, e.g., [3]). We introduce the new framework of probabilistic inference in Section 2 and develop a specific inference model for the multinomial problem in Section 3. The method is applied in a genome-wide association study to identify SNPs that are potentially associated with a given disease in Section 4. Section 5 concludes with a few remarks.

2 A New Framework of Probabilistic Inference

We start with a demonstration example. Assume that a set of observed data X is available and that model $f_\theta(X)$ for $X \in \mathscr{X}$ is specified, usually with unknown parameter $\theta \in \Theta$. We use the following example to explain the new framework of probabilistic inference. The key idea is to use an unobserved auxiliary random variable to represent $f_\theta(X)$.

Example 1. Let X be a dichotomous observation with $X \in \mathscr{X} = \{0,1\}$. Assume a Bernoulli model

$$P_\theta(X=1) = \theta \qquad \text{and} \qquad P_\theta(X=0) = 1-\theta$$

with unknown $\theta \in \Theta = [0,1]$. The problem is to infer θ from X. We consider a data generating mechanism using an auxiliary random variable $U \sim Unif(0,1)$:

$$X = \begin{cases} 1, & \text{if } U \le \theta; \\ 0, & \text{if } U > \theta. \end{cases}$$

This sampling mechanism preserves the model for X given θ. Moreover, it creates a random set for the parameter θ given the observation X, as defined below,

$$S_X = \begin{cases} U \le \theta \le 1, & \text{if } X = 1; \\ 0 \le \theta < U, & \text{if } X = 0. \end{cases} \qquad (U \sim Unif(0,1))$$

In other words, we think $\theta \in [U,1]$ if we observe $X=1$ and $\theta \in [0,U)$ if $X=0$, where U is a random variable from $Unif(0,1)$. *This relationship among the parameter of interest θ, the observation X, and the auxiliary random variable U is critical in our construction of the probabilistic inferential model, where inference about the parameter θ will be derived from prediction of the auxiliary random variable U.*

Given, for example, $X=1$, we have the random interval $S_X=[U,1]$ as the region for θ. Now consider an assertion $\mathscr{A}=\{\theta \le \theta_0\} \subseteq \Theta$ for a fixed $\theta_0 \in (0,1)$. There are two possible cases: (*i*) if $U > \theta_0$, the random set $S_X=[U,1]$ for θ provides evidence against the truth of $\mathscr{A}$; (*ii*) if $U \le \theta_0$, the random set $S_X=[U,1]$ for θ does not have any information about the truth or falsity of $\mathscr{A}$. Note that there is no realization of the random interval that provides evidence for the truth of $\mathscr{A}$, because the random set $[U,1]$ cannot be fully contained in $\mathscr{A}=\{\theta \le \theta_0\}$. As a result, the probability triplet (p,q,r) for the assertion $\mathscr{A}$ are calculated in the following

$$p=0, \quad q=P\{U>\theta_0\}=1-\theta_0, \quad \text{and } r=\theta_0. \qquad \square$$

To emphasize the fact that the (p,q,r) output is conditional on the observed data X, we write (p,q,r) as $(p_X(\mathscr{A}),q_X(\mathscr{A}),r_X(\mathscr{A}))$, that is,

$p_X(\mathscr{A})$: the probability for the truth of $\mathscr{A}$, given X
$q_X(\mathscr{A})$: the probability against the truth of $\mathscr{A}$, given X
$r_X(\mathscr{A})$: the probability neither for nor against the truth of $\mathscr{A}$, given X.

Formally, an inferential model for probabilistic inference about θ is given by a probability model with the sample space consisting of all subsets of Θ. Its probability measure is defined by an auxiliary random variable, for example the uniform variable U in Example 1. More specifically, a random set is constructed for inference about θ using the auxiliary random variable and conditional on the observed data X. Denote the random set by S_X, as in Example 1. The probability for the truth of a given assertation $\mathscr{A}$ (on the parameter θ) is computed as the probability that the random set S_X is contained in $\mathscr{A}$,

$$p_X(\mathscr{A})=P(S_X \subseteq \mathscr{A}).$$

Based on a symmetry argument, the probability against the truth of $\mathscr{A}$ or for the truth of $\mathscr{A}^c$ is computed as the probability that the random set S_X is contained in $\mathscr{A}^c$,

$$q_X(\mathscr{A})=P(S_X \subseteq \mathscr{A}^c).$$

The remaining probability

$$r_X(\mathscr{A})=1-p_X(\mathscr{A})-q_X(\mathscr{A})$$

is the probability that the random set S_X intersects with both $\mathscr{A}$ and $\mathscr{A}^c$, in which case we "don't know" the truth or falsity of $\mathscr{A}$.

In order for the probability triplet $(p_X(\mathscr{A}),q_X(\mathscr{A}),r_X(\mathscr{A}))$ to have desirable long-run frequency properties, the concept of validity is helpful.

Definition 1. The inferential model is valid for assertion $\mathscr{A}$ if for every α in $(0,1)$, both

$$P_\theta(\{X : p_X(\mathscr{A}) \geq \alpha\}) \leq 1-\alpha \quad \text{and} \quad P_\theta(\{X : q_X(\mathscr{A}) \geq \alpha\}) \leq 1-\alpha \qquad (1)$$

hold respectively for every $\theta \in \mathscr{A}^c = \Theta \setminus \mathscr{A}$ and for every $\theta \in \mathscr{A}$. The probabilities in (1) are defined with respect to the random variable X following $f_\theta(X)$.

In other words, credibility requires $p_X(\mathscr{A})$ and $q_X(\mathscr{A})$, as functions of the random variable X, to be stochastically bounded by the uniform distribution over the unit interval $(0,1)$ in repeated experiments. Thus, the triplet $(p_X(\mathscr{A}), q_X(\mathscr{A}), r_X(\mathscr{A}))$ provides strength of evidence for both $\mathscr{A}$ and $\mathscr{A}^c$ in term of long-run frequency probability. For hypothesis testing, thresholds for $p_X(\mathscr{A})$ and $q_X(\mathscr{A})$ can be used to confirm the truth and falsity of $\mathscr{A}$.

3 Inference of Multinomial Models

Now we develop an inferential model for parameters of multinomial distributions. In the following, the probabilistic inference of multinomial models is valid for data with both small and large number of categories. We start with a motivating example of genome-wide association study, where we compare SNPs frequencies of control samples and case samples. We scan the whole genome sequence using blocks of SNPs, for example, with a block size of 10 SNPs. For a given block, there are two independent multinomial distributions, corresponding to distributions of SNP genotypes of the control and case populations. These two multinomial distributions can be derived by a $2 \times K$ table of independent Poisson counts, where K is the total number of SNP genotypes in the block. More specificially, let $N_j^{(i)}$ denote a Poisson count with unknown rates $\lambda_j^{(i)} \geq 0$ for $i = 0, 1$ and $j = 1, ..., K$. It is well known that conditioning on $m_i = \sum_{j=1}^K N_j^{(i)}$ for $i = 0$ and 1, the observed data $N_j^{(i)}$ follow two independent multinomial models with

$$(N_1^{(i)}, ..., N_K^{(i)}) \sim \text{Multinomial}\left(m_i, \theta_1^{(i)}, ..., \theta_K^{(i)}\right) \qquad (i = 0, 1)$$

where $\theta_j^{(i)} = \lambda_j^{(i)} / \sum_{j=1}^K \lambda_j^{(i)}$ is the SNPs frequencies of the control ($i = 0$) and case ($i = 1$) populations. The problem of interest here is inference about the assertion that $\theta_j^{(0)} = \theta_j^{(1)}$ for $j = 1, ..., K$. In terms of $\lambda_j^{(i)}$, this assertion can be written as

$$\lambda_j^{(0)} \propto \lambda_j^{(1)} \qquad (j = 1, ..., K).$$

Alternatively, conditional on each column the Poisson counts of the $2 \times K$ table lead to K binomial distributions. Let $\phi_j = \lambda_j^{(1)} / (\lambda_j^{(0)} + \lambda_j^{(1)})$ and write $n_j = N_j^{(0)} + N_j^{(1)}$ and $X_i = N_j^{(1)}$ for $j = 1, ..., K$. Then,

$$X_j \sim \text{Binomial}(n_j, \phi_j) \qquad (j = 1, ..., K).$$

The inference about equal multinomial frequency parameters is the same as inference about

$$\mathscr{A} = \{\phi_j = \phi_0 : j = 1, ..., K \text{ for some } \phi_0 \in [0,1]\}. \tag{2}$$

For probabilistic inference of (2), we take the generalized inferential model approach [4] (See also [7, 8, 2] for examples of belief approaches based on likelihood functions). That is, inference can be made from a function of the observed data, *e.g.*, $Y = h(X)$ for some specified function $h(.)$. This can be viewed as an extension of the basic method described in Section 2. Denote $N = \sum_{j=1}^{K} n_j$, which is the total sample size of both control and case groups. We introduce a statistic

$$Y = \sum_{j=1}^{K} w_j \frac{\left(X_j - n_j \frac{\sum_{j=1}^{K} X_j}{N}\right)^2}{n_j(N - n_j)}$$

where $w_j = (n_j - 1)/(n_j + 1)$ is used to down-weight observations with small column size n_j. Note that we only consider counts with the column total of $n_j \geq 2$ when calculating Y. Let $\phi = (\phi_1, ..., \phi_K)$ denote the parameter of the assertion of interest (2) and $F_\phi(y)$ be the cdf of Y conditioning on $\sum_{j=1}^{K} X_j$. The conditional distribution $F_\phi(y)$ may be derived using the fact that X_j's follow a (multivariate) hypergeometric distribution conditioning on $\sum_{j=1}^{K} X_j$. In addition, $F_\phi(y)$ depends on ϕ only through their relative values, say, $\phi / \sum_{j=1}^{K} \phi_j$. For a data-generating device of the observable quantity Y, we know that Y can be generated by taking the inverse of $F_\phi(y)$ on a uniform random variable U. Following the idea of Example 1 in Section 2, we have a random set for ϕ,

$$S_Y = \left\{\phi : F_\phi(Y) \leq U\right\},$$

where U is the auxiliary random variable from $Unif(0,1)$. The probability triplet for the assertion $\mathscr{A}$ are obtained as

$$p(\mathscr{A}) = P(S_Y \subseteq \mathscr{A}), \quad q(\mathscr{A}) = P(S_Y \subseteq \mathscr{A}^c), \quad r(\mathscr{A}) = 1 - p(\mathscr{A}) - q(\mathscr{A}).$$

The probability for $\mathscr{A}$, $p(\mathscr{A})$, is necessarily zero, as the assertion represents a lower-dimensional space, where all components of ϕ are equal. The probability against the assertion, $q(\mathscr{A})$, is computed by using the fact that

$$\begin{aligned} q(\mathscr{A}) &= \Pr\left(S_Y \subseteq \mathscr{A}^c\right) = \Pr\left(S_Y^c \supseteq \mathscr{A}\right) \\ &= \Pr\left(\{\phi : F_\phi(Y) > U\} \supseteq \mathscr{A}\right) \\ &= \Pr\left(U < F_\phi(Y) \text{ for all } \phi \in \mathscr{A}\right) \\ &= \Pr\left(U < \min_{\phi \in \mathscr{A}} F_\phi(Y)\right) \\ &= \min_{\phi \in \mathscr{A}} F_\phi(Y). \end{aligned}$$

We compute this q value by a Monte Carlo method. Under $\mathscr{A}$, all components of ϕ are the same. Because the distribution $F_\phi(Y)$ only depends on relative values of the components of ϕ, there is only one quantity of $F_\phi(Y)$ over $\phi \in \mathscr{A}$. The minimization is in fact not necessary. For a good approximation with a small Monte Carlo sample size (e.g., $1,000$), we compute the distribution of Y using a scaled χ^2 distribution with the scale and degrees of freedom estimated from the Monte Carlo sample by the method of moments.

4 Application in Genome-Wide Association Study

We apply the methodology on the GAW16 (Genetic Analysis Workshop 16) data from the North American Rheumatoid Arthritis Consortium. This genome-wide association study aims at identifying genetic variants, more specifically single nucleotide polymorphisms, that are associated with the Rheumatoid Arthritis disease. The data consists of 2062 samples, where 868 are cases and 1194 are controls. For each sample, whole genome SNPs are observed with a total coverage of 545,080 SNPs.

We partition the entire SNP sequence on each chromosome into a sequence of m blocks of consecutive SNPs, each block consisting of, for example, 10 SNPs. For each block, indexed by $b = 1,...,m$, our proposed approach to analysis of the two-sample multinomial counts produces (p_b, q_b, r_b) output for the assertion that "the two samples, cases versus controls, are from the same population". The (p_b, q_b, r_b) output has $p_b = 0$ and q_b providing evidence against the assertion.

The validity of our (p,q,r)'s implies that most values in the collection $\{q_b : b = 1,...,m\}$ can be viewed as a sample from a common distribution, called the *null* distribution, that has more small values in the unit interval $(0,1)$ than the uniform distribution. The remaining large values provide information on the potential blocks that distinguish a case from a control.

Figure 1 displays sequences of the q-value for chromosomes 6 and 14 in terms of Z-score, $Z = \Phi^{-1}(q_b)$, where $\Phi^{-1}(.)$ stands for the cdf of the standard normal distribution. When larger than 8, the values of the Z-scores are replaced with 8 in the plots. Figure 2 displays the histograms of the q-value for chromosomes 6 and 14. Large values on the right tail in Figure 2(a) indicate that there are some blocks on chromosome 6 potentially associated with Rheumatoid Arthritis. This result is consistent with the known fact that the HLA (human leukocyte antigen) region on chromosome 6 contributes to disease risk. Figure 2(b) shows that there are hardly any blocks on chromosome 14 that are associated with Rheumatoid Arthritis. Except for these large values, the q-value in Figures 2(a) and 2(b) have very smooth distributions. This implies that we can specify a null distribution so that outliers or blocks potentially associated with Rheumatoid Arthritis can be identified. The same results on the null distribution are seen on the other chromosomes.

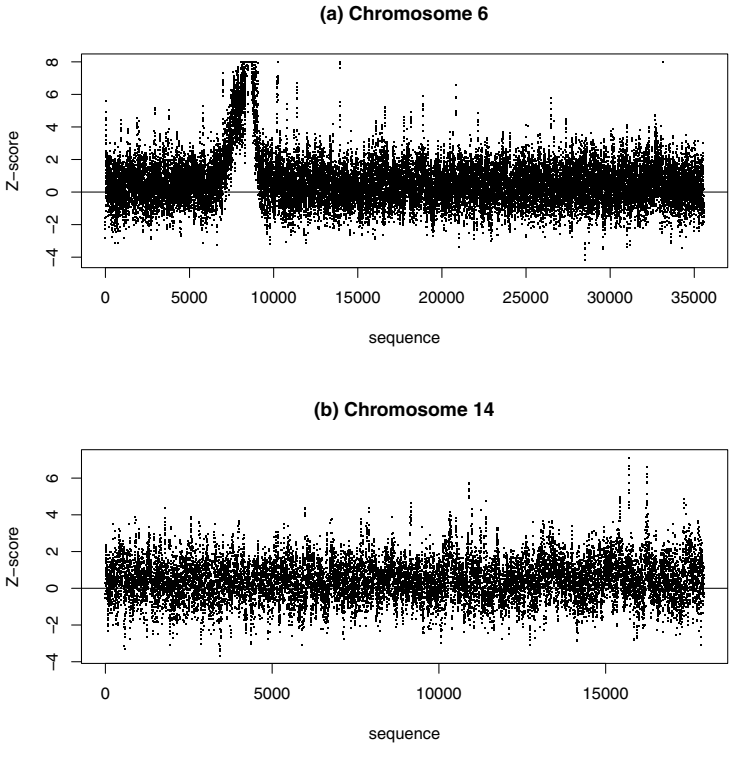


Fig. 1 The time-series plots of the Z-scores of the probabilities for the assertion that control and case populations are different, computed based on the two-multinomial model for SNPs in blocks of 10 in (a) chromosome 6 and (b) chromosome 14.

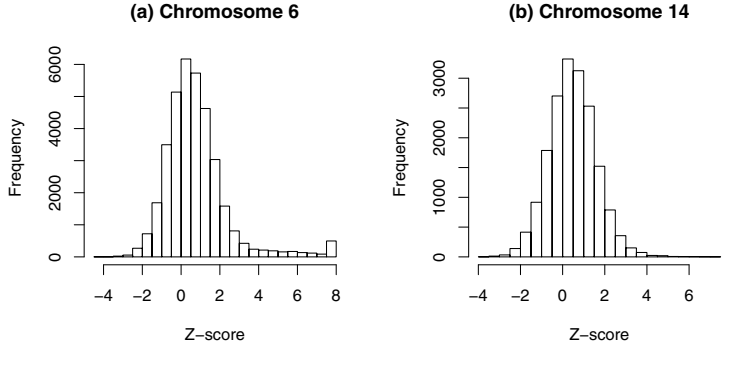


Fig. 2 Histograms of the Z-scores of the probabilities for the assertion that control and case populations are different, computed based on the two-multinomial model for SNPs in blocks of 10 in (a) chromosome 6 and (b) chromosome 14.

5 Conclusion

The difficulty of existing statistical methods for large-scale multinomial counts requires thinking about statistical inference at a very fundamental level and demands novel ideas beyond the current two dominant schools of thought, the frequentist and Bayesian. We propose a probabilistic inferential model, which uses auxiliary random variables for reasoning towards inference rather than constructing fiducial probabilities in the attempt to replace Bayesian posterior probabilities. The proposed method works for data of both small and large sample sizes. It produces inferential results that have desirable frequency properties. In addition, compared with maximum likelihood based inference with hypothetically large sample sizes, the proposed method is also efficient. In the inferential model framework, there are issues that need further investigation. These include the arbitrariness of the unobserved auxiliary random variable, specification of the predictive random sets, and choice of partial sampling model in generalized inferential models. To save space, we refer to the on-going work [3, 5] for relevant discussions. We believe that the proposed method will 1) generates useful tools for applied statisticians who are challenged by very-high-dimensional count data, and 2) call attention to fundamental research on statistical inference and problems considered by founding fathers such as Ronald Fisher and Jerzy Neyman.

Acknowledgements. This work is supported by the National Science Foundation Grant DMS-1007678.

References

1. Dempster, A.P.: The Dempster-Shafer calculus for statisticians. International Journal of Approximate Reasoning 48, 265–277 (2008)
2. Denoeux, T.: Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions. International Journal of Approximate Reasoning 42(3), 228–252 (2006)
3. Martin, R., Liu, C.: Inferential models: A framework for prior-free posterior probabilistic inference. Technical Report, Department of Statistics, Purdue University (2011), http://www.stat.purdue.edu/~chuanhai/docs/imbasics.pdf.
4. Martin, R., Liu, C.: Generalized inferential models. Technical Report, Department of Statistics, Purdue University (2011), http://www.stat.purdue.edu/~chuanhai/docs/imlik-4.pdf.
5. Martin, R., Liu, C.: Inferential Models: Reasoning with uncertainty. Chapman & Hall/CRC (2013)
6. Martin, R., Zhang, J., Liu, C.: Dempster-Shafer theory and statistical inference with weak beliefs. Statistical Science 25, 72–87 (2010)
7. Shafer, G.: A mathematical theory of evidence. New Jersey. Princeton University Press, Princeton (1976)
8. Wasserman, L.A.: Belief functions and statistical evidence. The Canadian Journal of Statistics 18(3), 183–196 (1990)
9. Zhang, J., Liu, C.: Dempster-Shafer inference with weak beliefs. Statistica Sinica 21, 475–494 (2011)

Belief Function Robustness in Estimation

Alessio Benavoli

Abstract. We consider the case in which the available knowledge does not allow to specify a precise probabilistic model for the prior and/or likelihood in statistical estimation. We assume that this imprecision can be represented by belief functions. Thus, we exploit the mathematical structure of belief functions and their equivalent representation in terms of closed convex sets of probability measures to derive robust posterior inferences.

1 Introduction

Lower and Upper probabilities induced from multivalued mappings were introduced by Dempster [1]. Shafer [2] called them belief and plausibility functions. Associated with a belief function there is a closed convex set of probability measures of which the belief function is a lower bound [1, 3, 4]. On the other hand, the lower bound of a convex set of probability measures is not necessarily a belief function, e.g., [3, Sec. 5.13.4]. Wasserman [5, 6] has shown that the mathematical structure of belief functions makes them suitable for generating classes of prior distributions to be used in robust Bayesian inference. In particular, in case the prior is expressed via a belief function and the likelihood is a precise probability measures, he has derived a closed form solution for the upper and lower bounds of the posterior probability content of a measurable subset of the parameter space (even in case of infinite spaces). In this paper, we extend this work in three directions. First, we compute upper and lower bounds of the posterior expectations for any bounded scalar function g of interest in statistical estimation. Second, we consider the case in which also the likelihood model (not only the prior) may be expressed via belief functions. By using the formalism of Walley's theory of coherent lower previsions [3], we provide

Alessio Benavoli
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA),
Galleria 1, CH-6928 Manno (Lugano), Switzerland
e-mail: alessio@idsia.ch

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 375–383.
springerlink.com

closed form solutions for the lower and upper expectations of g. Third, we show the application of this model to several cases of practical interest.

2 Belief Function

In this section we revise some properties of belief functions. Let $\mathscr{X}$ be a Polish space (e.g., Euclidean space) with Borel σ-algebra $\mathscr{B}(\mathscr{X})$ and let $\mathscr{Z}$ be a convex, compact, metrizable subset of a locally convex topological vector space with Borel σ-algebra $\mathscr{B}(\mathscr{Z})$ [5]. Let P_Z be a probability measure on $(\mathscr{Z},\mathscr{B}(\mathscr{Z}))$ and let Γ be a map taking points in $\mathscr{Z}$ to nonempty, closed subsets of $\mathscr{X}$.[1] For each $A \subseteq \mathscr{X}$, define the belief and plausibility function as [1, 5]:

$$\begin{aligned} \underline{P}(A) &= Bel(A) = P_Z(\{z_i \in \mathscr{Z} : \Gamma(z_i) \subset A\}), \\ \overline{P}(A) &= \ \ Pl(A) = P_Z(\{z_i \in \mathscr{Z} : \Gamma(z_i) \cap A \neq \emptyset\}). \end{aligned} \tag{1}$$

The fourtuple $(\mathscr{Z},\mathscr{B}(\mathscr{Z}),P_Z,\Gamma)$ is called a source for Bel. Bel and $P1$ are related by $Bel(A) = 1 - P1(A^C)$, where A^c is the complement of A. An intuitive explanation [5] of Bel and $P1$ is as follows. Draw z randomly according to P_Z. Then $Bel(A)$ is the probability that the random set $\Gamma(z)$ is contained in A and $P1(A)$ is the probability that the random set $\Gamma(z)$ hits A [7]. Here, a simple example [3, Sec. 5.13.3] that explains the construction of belief functions through multivalued mappings.

Example 1. Suppose that our information on $\mathscr{X}$ is a report from an unreliable witness that the event $B \subset \mathscr{X}$ has occurred. We might consider two possible explanations: either the witness really observed B, or he observed nothing at all. These hypotheses are represented by z_1 and z_2, with multivalued mapping $\Gamma(z_1) = B$ and $\Gamma(z_2) = \mathscr{X}$. If we assess the probability $P_Z(z_1) = p$ and $P_Z(z_2) = 1 - p$, this corresponds to the belief function $Bel(A) = p$ if $A \supseteq B$ and $A \neq \mathscr{X}$; $Bel(A) = 1$ if $A = \mathscr{X}$ and zero otherwise. □

This lack of knowledge expresses via a belief function can equivalently be represented through a set of probability measures, i.e., the set of all probabilities on X that are compatible with the bounds Bel and Pl [1]:

$$\mathscr{P}_X = \{P_X : \ Bel(A) \leq P_X(A) \leq Pl(A) \ for \ any \ A \subseteq \mathscr{X}\}. \tag{2}$$

For this reason, Bel is also called lower probability $\underline{P}$ (and Pl upper probability $\overline{P}$), since it is the lower (upper) envelope of a set of probability measures. Thus, associated to each belief function, there is a closed convex set of probability measures of which a belief function is a lower bound but, on the other hand, the lower bound $\underline{P}$ of a closed convex set of probability measures is not necessarily a belief function [3]. To be a belief function, the lower probability $\underline{P}$ has to satisfy the property of ∞-monotonicity. There are many closed convex sets of distributions that are used in practical applications that are not belief functions. By restricting closed

[1] Natural conditions, such as upper or lower semi-continuity, may be imposed on Γ to guarantee measurability [5].

convex sets of distributions to be belief functions one looses in generality but gains in tractability. In fact, because of the ∞-monotonicity property, belief functions satisfy several nice properties. Besides tractability, belief functions are also a useful source of closed convex set of probabilities. For instance, the multivalued mapping mechanism can be used to define belief functions also in the case the set $\mathscr{X}$ is continuous.

Example 2. Consider the case $\mathscr{X} = \mathscr{Z} = \mathbb{R}$ and thus $\mathscr{B}(\mathscr{Z})$ and $\mathscr{B}(\mathscr{X})$ coincide with the standard Borel σ-algebra in $\mathbb{R}$. Since $\mathscr{X} = \mathscr{Z}$, we are considering a map from $\mathscr{X}$ to itself and, thus, for simplicity we can denote z with x. Assume that $p(x)$ is the probability density w.r.t. the Lebesgue measure on $\mathbb{R}$ associated to P_Z (assuming it exists) and consider the case $p(x) = U_{[a,b]}(x)$, i.e., the uniform density on the interval $[a,b]$. Consider then the multivalued mapping $\Gamma(x) = [x-c, x+c]$ with $c > 0$ which maps each point x in the interval $[x-c, x+c]$. This originates the following lower/upper probabilities for the interval $[r,s]$ with $r < s$:

$$\begin{aligned} \underline{P}([r,s]) &= \int_{x\in[a,b]} I_{\{x:\ [x-c,x+c]\subset[r,s]\}}(u)\tfrac{1}{b-a}du, \\ \overline{P}([r,s]) &= \int_{x\in[a,b]} I_{\{x:\ [x-c,x+c]\cap[r,s]\neq\emptyset\}}(u)\tfrac{1}{b-a}du, \end{aligned} \tag{3}$$

where $I_{\{A\}}$, defined by $I_{\{A\}}(x) = 1$ if $x \in A$ and $I_{\{A\}}(x) = 0$ if $x \notin A$ is called the indicator of A. Notice that the inclusion $[x-c, x+c] \subset [r,s]$ holds for all $x \in [r+c, s-c]$, while the condition $[x-c, x+c] \cap [r,s] \neq \emptyset$ is satisfied by all $x \in [r-c, s+c]$. By setting $[r,s] = (-\infty, x]$, one can compute the lower/upper cumulate distribution function:

$$\underline{P}((-\infty,x]) = \begin{cases} 0 & x < a+c, \\ \frac{x-a-c}{b-a} & a+c \le x < b+c, \\ 1 & x \ge b+c, \end{cases} \quad \overline{P}((-\infty,x]) = \begin{cases} 0 & x < a-c, \\ \frac{x-a+c}{b-a} & a-c \le x < b-c, \\ 1 & x \ge b-c. \end{cases} \tag{4}$$

This model can be used to account for lack of information on the support of the uniform distribution. We are eliciting a support of length $b-a$ but we are not completely sure about its extremes. □

This approach can be extended to any PDF $p(x)$ (e.g., see [5] for the Gaussian case). Assume $\mathscr{X} = \mathscr{Z} = \mathbb{R}$, we discuss two other models (the first is discussed in [5]) generated by multivalued mappings.

ε-contamination: $p(x) = (1-\varepsilon)\pi'(x) + \varepsilon\delta_{\{z_0\}}(x)$, $\Gamma(x) = x$ if $x \neq z_0$ and $\Gamma(x) = \mathbb{R}$ if $x = z_0$, where $\delta_{\{z_0\}}$ is a Dirac's delta on z_0 and π' is PDF such that $\pi'(z_0) = 0$ and $\pi' = \pi$ if $x \neq z_0$, then:

$$\underline{P}(A) = (1-\varepsilon)\int_A \pi(x)dx, \quad \overline{P}(A) = (1-\varepsilon)\int_A \pi(x)dx + \varepsilon.$$

When $\varepsilon = 1$, we have a vacuous model $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$.

Heavy-tail: $p(x) = \mathcal{N}(x;0,1)$, $\Gamma(x) = [x, 1/x)$ if $x \in [0,1)$ and $\Gamma(x) = x$ if $x \geq 1$ (symmetric for the negative axis). Consider $A = (-\infty, w]$ with $w > 1$, we can then compute the lower upper distribution of X:

$$\underline{P}((-\infty,w]) = \tfrac{1}{2} + \int_1^w \mathcal{N}(x;0,1)dx + \int_{1/w}^1 \mathcal{N}(w;0,1)dx = \tfrac{1}{2}\left(\mathrm{erf}\left(\tfrac{w}{\sqrt{2}}\right) - \mathrm{erf}\left(\tfrac{1}{\sqrt{2}w}\right)\right),$$

where $\mathrm{erf}(w) = \frac{2}{\sqrt{\pi}}\int_0^w e^{-t^2}dt$, while

$$\overline{P}((-\infty,w]) = \tfrac{1}{2} + \int_0^w \mathcal{N}(x;0,1)dx = \tfrac{1}{2}\left(1 + \mathrm{erf}\left(\tfrac{w}{\sqrt{2}}\right)\right).$$

By differentiating $\underline{P}((-\infty,w])$ w.r.t. w, one gets:

$$\tfrac{d}{dw}\underline{P}((-\infty,w]) = \tfrac{1}{2}\left(\sqrt{\tfrac{2}{\pi}}e^{-\frac{w^2}{2}} + \frac{\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2w^2}}}{w^2}\right).$$

Observe that the derivative goes to zero as $1/w^2$ and, thus, it has the same tail behaviour of the Cauchy density $\gamma/[\pi((w-w_0)^2+\gamma^2)]$. This belief function can be used for instance for robustness to outliers when employed as likelihood model or as a sort of weak-informative prior (when employed as prior model).

2.1 Upper and Lower Expectation

The previous section has discussed several belief functions generated through a multivalued mappings. We have also seen that a belief function can equivalently be interpreted as a lower probability model defined on the subsets of $\mathscr{X}$ and also as a lower expectation model defined on the indicator functions over the subsets of $\mathscr{X}$, i.e., $\underline{E}(I_{\{A\}}) = \underline{P}(A)$. Assume that we know the functional $\underline{P}(A) = \underline{E}(I_{\{A\}})$ for any subset A of $\mathscr{X}$ how can we extend this lower probability model to compute $\underline{E}(g)$ for any bounded real-valued function of interest g. It can be shown that

$$\underline{E}(g) = \inf_{P_X \in \mathscr{P}_X} \int g(x)P_X(dx), \quad \overline{E}(g) = \sup_{P_X \in \mathscr{P}_X} \int g(x)P_X(dx). \tag{5}$$

Thus, the interpretation of belief functions as closed convex set of probability measures allows to computer lower and upper expectations for any bounded real valued function. Since belief function are multivalued mapping, it has been proved in [5] that (5) is equal to:

$$\underline{E}(g) = \int g_*(z)P_Z(dx), \quad \overline{E}(g) = \int g^*(z)P_Z(dz), \tag{6}$$

where $g_*(z) = \inf_{x\in\Gamma(z)} g(x)$ and $g^*(z) = \sup_{x\in\Gamma(z)} g(x)$. This fact has important implications for computation because it reduces the problem of calculating extrema

over the set of probability measures $\mathscr{P}_X$ to that of finding extrema of g over subsets of $\mathscr{X}$ followed by a single integral over Z.

Example 3. Consider for instance the ε-contamination model discussed in the previous section, then

$$\begin{aligned}\underline{E}(g) &= \int g_*(z)P_Z(dz) = \int dz\left[(1-\varepsilon)\pi'(z)+\varepsilon\delta_{\{z_0\}}(z)\right]\inf_{x\in\Gamma(z)} g(x),\\ &= \int_{\mathscr{X}-\{z_0\}} (1-\varepsilon)\pi'(z)g(z)dz+\varepsilon\inf_{x\in\mathbb{R}} g(x) = \int(1-\varepsilon)\pi(z)g(z)dz+\varepsilon\inf_{x\in\mathbb{R}} g(x).\end{aligned} \tag{7}$$

In case $\pi(z)=\mathscr{N}(z;x_0,\sigma_0^2)$ and in the case the vacuous part is restricted to $[-a,a]$ with $a>0$, one gets $\underline{E}(g)=\int(1-\varepsilon)g(z)\mathscr{N}(z;x_0,\sigma_0^2)dz+\varepsilon\inf_{x\in[-a,a]} g(x)$. □

2.2 Statistical Inference

Assume that $\mathscr{X}\subseteq\mathbb{R}$. Consider a likelihood model $p(y|x)$, where Y denotes the observation variable taking values from a sample space $(\mathscr{Y},\mathscr{B}(\mathscr{Y}))$ and $x\in\mathscr{X}$. Assume that the prior information over X is expressed through a belief function or, equivalently, through the closed convex set of probability measures associated to the belief function, how can we compute the lower/upper posterior expectation of a bounded real-valued function g given the observation $\tilde{y}$?

Theorem 1. *Assume that $p(y|x)$ is $\mathscr{B}(\mathscr{Y})\times\mathscr{B}(\mathscr{X})$-measurable and bounded. Assume that the value $\tilde{y}$ of Y is observed and that $\underline{E}_X(E_Y(\delta_{\{\tilde{y}\}}|X))=\underline{E}\big(p(\tilde{y}|x)\big)>0$, where $\delta_{\{\tilde{y}\}}$ is a degenerate limiting measure (e.g., Dirac's delta) on $\mathscr{Y}$. The lower posterior expectation $\underline{E}(g|\tilde{y})$ is the unique solution μ of the following equation:*

$$\underline{E}_X\Big(E_Y\big((g-\mu)\delta_{\{\tilde{y}\}}|X\big)\Big)=0. \ \square \tag{8}$$

This equation is called Generalized Bayes rule (GBR) [3, Ch. 6].

Proof.

$$\begin{aligned}0 &= \underline{E}_X\Big(E_Y\big((g-\mu)\delta_{\{\tilde{y}\}}|X\big)\Big) = \underline{E}\big((g-\mu)p(\tilde{y}|x)\big) = \inf_{P_X\in\mathscr{P}_X}\int(g(x)-\mu)p(\tilde{y}|x)P_X(dx)\\ &= \inf_{p\in\mathscr{P}_X}\int p(\tilde{y}|x)P(dx)\left(\frac{\int g(x)p(\tilde{y}|x)P(dx)}{\int p(\tilde{y}|x)P(dx)}-\mu\right).\end{aligned}$$

Being $\int p(\tilde{y}|x)P_X(dx)=\underline{E}\big(p(\tilde{y}|x)\big)>0$ by hypothesis, it follows that $\mu=\inf_{p_X\in\mathscr{P}_X}\frac{\int g(x)p(\tilde{y}|x)P_X(dx)}{\int p(\tilde{y}|x)P_X(dx)}$. Therefore, GBR is equivalent to apply Bayes rule to all probability measures in $\mathscr{P}_X$ and, then, take the infimum. The following proof has been derived by [3, Sec. 6.4.1.] replacing the indicator with a Dirac's delta to account for the fact that Y is a continuous variable. □

Corollary 1. *Exploiting (6) and applying (8) to belief function, it results that the lower posterior expectation* $\underline{E}(g|\tilde{y})$ *is the unique solution* μ *of the following equation:*

$$\underline{E}\big((g-\mu)p(\tilde{y}|x)\big) = \int P_Z(dz) \inf_{x\in\Gamma(z)} (g(x)-\mu)p(\tilde{y}|x) = 0. \tag{9}$$

□

For the proof, see Theorem 2. Equation (9) can be extended to the case of n i.i.d. observation, by simply replacing $p(\tilde{y}|x)$ with $\prod_{i=1}^{n} p(\tilde{y}_i|x)$. Assume now the case in which also the likelihood model is expressed through a belief function characterized by $(\mathscr{U},\mathscr{B}(\mathscr{U}),\Gamma(\cdot|x),P_{U|x})$ for each value of the conditional variable x.

Theorem 2. *Assume that the value* $\tilde{y}$ *of* Y *is observed, that* $\underline{E}_X(\underline{E}_Y(\delta_{\{\tilde{y}\}}|X)) > 0$ *and* $\underline{E}_Y(\delta_{\{\tilde{y}\}}|x)$ *is well defined for each* $x \in \mathscr{X}$*. The lower posterior expectation* $\underline{E}(g|\tilde{y})$ *is the unique solution* μ *of the following equation:*

$$\underline{E}_X\left(\underline{E}_Y\big((g-\mu)\delta_{\{\tilde{y}\}}|X\big)\right) = 0, \tag{10}$$

which for belief functions becomes:

$$0 = \int P_Z(dz) \inf_{x\in\Gamma(z)} \int P_{U|x}(du|x) \inf_{y\in\Gamma(u|x)} \delta_{\{\tilde{y}\}}(y)(g(x)-\mu). \tag{11}$$

□

This is the extension of Corollary 1 to the case also the likelihood is a belief function. The proof of this theorem can be derived from the proof of [8, Th. 2] by using the expression for the lower expectation in (6).[2] The above result is very important for practical applications as shown in the next examples.

3 ε-Contamination and Interval Estimation

Consider an ε-contamination model for $(\mathscr{U},\mathscr{B}(\mathscr{U}),\Gamma(\cdot|x),P_{U|x})$, i.e., $P_{U|x} = (1-\varepsilon_m)\mathscr{N}(u;x,\sigma^2) + \varepsilon_m\delta_{\{u_0\}}(u|x)$ and $\Gamma(u|x) = y$ and $\Gamma(u_0|x) = \mathscr{A}_Y(x) = [x-b, x+b]$ for $b > 0$. In the domain of the variable Y, this model is equivalent to: $y = x + (1-\varepsilon_m)n + \varepsilon_m v$, where n is a Gaussian noise with zero mean and variance σ^2, while v is a noise with unknown distribution. The only knowledge about v is its support $[-b,b]$ (norm bounded noise). This model can be used to account for the uncertainty in the measurement process which is due to a white noise component (n) and to the finite precision of the instrument (v), so it is very important for practical applications. Assume that also the prior over X is a ε-contamination model at the end of the Example 3. Applying (11) one gets:

[2] Observe that the proof in [8, Th. 2] has been obtained by assuming that the observation variables are discretized. Intuitively, we can see Theorem 2 as the limit of this result when the size of the discretization interval goes to zero.

$$0 = \int dz \left[(1-\varepsilon)\mathscr{N}(z;x_0,\sigma_0^2)+\varepsilon\delta_{\{z_0\}}(z)\right] \inf_{x\in\Gamma(z)} \int du \left[(1-\varepsilon_m)\mathscr{N}(u;x,\sigma^2)+\varepsilon_m\delta_{\{u_0\}}(u|x)\right] \inf_{y\in\Gamma(u|x)} \delta_{\{\tilde{y}\}}(y)(g(x)-\mu). \tag{12}$$

Consider the case where $\delta_{\{\tilde{y}\}}(y)$ is the limit for $|\Omega(\tilde{y})| \to 0$ of the following sequence of functions $\frac{1}{|\Omega(\tilde{y})|} I_{\{\Omega(\tilde{y})\}}$, where $\Omega(\tilde{y})$ is a ball centred at $\tilde{y}$ which does not depend on x and $|\Omega(\tilde{y})|$ is its Lebesgue volume [3, Sec. 6.10.4], [8]. Then, the previous integral equation can be rewritten as:

$$0 = \frac{1}{|\Omega(\tilde{y})|} \int dz \left[(1-\varepsilon)\mathscr{N}(z;x_0,\sigma_0^2)+\varepsilon\delta_{\{z_0\}}(z)\right] \inf_{x\in\Gamma(z)} \int du \left[(1-\varepsilon_m)\mathscr{N}(u;x,\sigma^2)+\varepsilon_m\delta_{\{u_0\}}(u|x)\right] \inf_{y\in\Gamma(u|x)} I_{\{\Omega(\tilde{y})\}}(y)(g(x)-\mu). \tag{13}$$

Since $|\Omega(\tilde{y})|$ is positive it can be simplified in the equation, which for $|\Omega(\tilde{y})| \to 0$ tends can be written as:

$$\begin{aligned} 0 &= \int dz \left[(1-\varepsilon)\mathscr{N}(z;x_0,\sigma_0^2)+\varepsilon\delta_{\{z_0\}}(z)\right] \inf_{x\in\Gamma(z)} \left[(1-\varepsilon_m)\mathscr{N}(\tilde{y};x,\sigma^2)(g(x)-\mu)+\varepsilon_m \inf_{y\in\mathscr{A}_Y(x)} I_{\{\Omega(\tilde{y})\}}(y)(g(x)-\mu)\right] \\ &= \int dz \left[(1-\varepsilon)\mathscr{N}(z;x_0,\sigma_0^2)+\varepsilon\delta_{\{z_0\}}(z)\right] \inf_{x\in\Gamma(z)} \left[(1-\varepsilon_m)\mathscr{N}(\tilde{y};x,\sigma^2)(g(x)-\mu)-\varepsilon_m I_{\{x:\ \tilde{y}\in\mathscr{A}_Y(x)\}}(x)(g(x)-\mu)^-\right], \end{aligned} \tag{14}$$

where $(g(x)-\mu) = -\min(g(x)-\mu,0)$ is the negative part of $g-\mu$. Simplifying the other integral and exploiting that $\Gamma(z_0)=[-a,a]$, $\mathscr{A}_Y(x)=[x-b,x+b]$ and, thus, $\tilde{y}\in\mathscr{A}(x)$ implies $x\in[\tilde{y}-b,\tilde{y}+b]$, one finally gets:

$$0 = \int (1-\varepsilon)\mathscr{N}(x;x_0,\sigma_0^2)dx \left[(1-\varepsilon_m)\mathscr{N}(\tilde{y};x,\sigma^2)(g(x)-\mu)-\varepsilon_m I_{\{x\in[\tilde{y}-b,\tilde{y}+b]\}}(x)(g(x)-\mu)^-\right] + \varepsilon \inf_{x\in[-a,a]} \left[(1-\varepsilon_m)\mathscr{N}(\tilde{y};x,\sigma^2)(g(x)-\mu)-\varepsilon_m I_{x\in[\tilde{y}-b,\tilde{y}+b]}(x)(g(x)-\mu)^-\right].$$

Notice that in case $\varepsilon=\varepsilon_m=0$ (no imprecision) and $g=X$, then $\mu=E(X|\tilde{y}) = (1/\sigma_0^2+1/\sigma^2)^{-1}(x_0/\sigma_0^2+\tilde{y}/\sigma^2)$ that is the well known expression for the posterior mean in the Gaussian case. In the vacuous case, $\varepsilon=\varepsilon_m=1$ (full imprecision), one gets:

$$0 = -\sup_{x\in[-a,a]} I_{\{x\in[\tilde{y}-b,\tilde{y}+b]\}}(x)(g(x)-\mu)^- = \sup_{x\in[-a,a]\cap[\tilde{y}-b,\tilde{y}+b]} (g(x)-\mu)^-. \tag{15}$$

For $g=X$ and assuming that the two intervals overlap, one has $\mu=\underline{E}(X|\tilde{y}) = \max(\tilde{y}-b,-a)$. Similarly, we can compute $\overline{E}(X|\tilde{y})=\min(\tilde{y}+b,a)$. This is the well known updating formula in *interval estimation*, for instance in *set-membership estimation* [9, 10]. Figure 1 shows the expression for $\underline{E}(X|\tilde{y})$ in case $\varepsilon=0$, $\varepsilon_m=0.5$, $\tilde{y}=1$, $x_0=0$, $\sigma_0^2=\sigma^2=1$ and different values of b, i.e., $b\in[0,4]$. It can be observed that for $b<(1/\sigma_0^2+1/\sigma^2)^{-1}(x_0/\sigma_0^2+\tilde{y}/\sigma^2)=0.5$, the posterior mean coincides with that of the case $\varepsilon_m=0$. The lower expectations starts to decrease when the

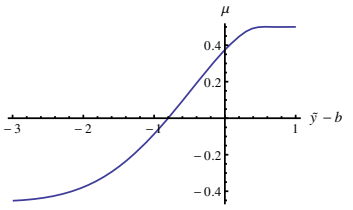


Fig. 1 Lower expectation $\mu = \underline{E}(X|\tilde{y})$ versus $\tilde{y} - b$ for $\tilde{y} = 1$ and $b \in [0,4]$.

support of the norm-bounded noise v, i.e., $[\tilde{y} - b, \tilde{y} + b]$, becomes larger than the interval $[0.5, 1.5]$.

4 Heavy-Tail Belief Function Model

Consider the model discussed at the end of Section 2 for the variable X with $p(x) = \mathcal{N}(x;0,2.19)$, i.e., variance 2.19. Assume that the measurement process is described by a normal density function $\mathcal{N}(y;x,\sigma^2)$. We can then use Corollary 1 to compute the lower posterior expectation of some function of interest g of X, i.e., $\underline{E}[g|y]$ is the unique solution μ of

$$0 = \int dx \mathcal{N}(x;0,2.19) \inf_{x' \in \Gamma(x)} (g(x') - \mu) \prod_{i=1}^{n} \mathcal{N}(y_i;x,\sigma^2), \tag{16}$$

where $\Gamma(x) = [x, 1/x)$ for $x \in (-1,1)$ and $\Gamma(x) = x$ otherwise. The above equation can be solved numerically by discretizing X. In Table 1 (last row) we have reported the lower and upper posterior mean of X computed according to (16) in case $g = X$. For the sake of comparison we have reported also the posterior means obtained by the prior $p_1(x) = \mathcal{N}(x;0,2.19)$ (Normal distribution), denoted by $E_1(X|y)$ in the table, and $p_2(x) = \mathscr{C}(x;0,1)$ (Cauchy distribution), denoted by $E_2(X|y)$. Both these two prior distributions have prior mean equal to zero and and prior quartiles equal to ± 1.[3] From Table 1 it can be noticed that at the increasing of the prior-data conflict (increasing of y) the Cauchy prior is more robust than the Normal prior, i.e., its posterior mean is closer to the value of the measurement. The third row in the table shows that the choice of a set of priors based on the heavy-tail belief function model further increases the robustness. Notice in fact that, for a small prior-data conflict, the interval $[\underline{E}(X|y), \overline{E}(X|y)]$ is tight and includes the posterior mean $E_1(X|y)$. However, at the increasing of the conflict, the interval enlarges highlighting the presence of a prior-data conflict, and its centre moves towards y similarly to the posterior mean of the Cauchy prior that moves towards y.

[3] This example has been adapted from [11, Sec. 4.7.1.].

Table 1 Posterior mean computed for the three different prior models.

y	0	1	2	4.5	10
$E_1(X\|y)$	0	0.69	1.37	3.09	6.87
$E_2(X\|y)$	0	0.55	1.28	4.01	9.80
$\underline{E}(X\|y), \overline{E}(X\|y)$	-0.26,0.26	0.68,1.46	1.35,1.93	2.78,4.52	5.42,14.01

5 Conclusions

We have derived robust inferences based on classes of priors and likelihoods generated by belief functions. As future work, we intend to apply this work to practical estimation problems and to derive more closed convex sets of probability measures by using the multivalued mapping mechanism of belief functions.

Acknowledgements. This work has been partially supported by the Swiss NSF grants n. 200020-121785/1.

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multiple-valued mapping. Ann. Math. Stat. 38, 325–339 (1967)
2. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
3. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)
4. Miranda, E., de Cooman, G., Couso, I.: Lower previsions induced by multi-valued mappings. Journal of Statistical Planning and Inference 133(1), 173–197 (2005)
5. Wasserman, L.A.: Prior envelopes based on belief functions. The Annals of Statistics 18(1), 454–464 (1990)
6. Wasserman, L.A.: Belief functions and statistical inference. The Canadian Journal of Statistics 18(3), 183–196 (1990)
7. Molchanov, I.: Theory of random sets. Springer (2005)
8. Benavoli, A., Zaffalon, M., Miranda, E.: Robust filtering through coherent lower previsions. IEEE Transactions on Automatic Control 56, 1567–1581 (2011)
9. Schweppe, F.C.: Recursive state estimation: Unknown but bounded errors and system inputs. In: Sixth Symposium on Adaptive Processes, vol. 6, pp. 102–107 (1967)
10. Bertsekas, D., Rhodes, I.: Recursive state estimation for a set-membership description of uncertainty. IEEE Transactions on Automatic Controll 16, 117–128 (1971)
11. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics, New York (1985)

Conditioning in Dempster-Shafer Theory: Prediction vs. Revision

Didier Dubois and Thierry Denœux

Abstract. We recall the existence of two methods for conditioning belief functions due to Dempster: one, known as Dempster conditioning, that applies Bayesian conditioning to the plausibility function and one that performs a sensitivity analysis on a conditional probability. We recall that while the first one is dedicated to revising a belief function, the other one is tailored to a prediction problem when the belief function is a statistical model. We question the use of Dempster conditioning for prediction tasks in Smets generalized Bayes theorem approach to the modeling of statistical evidence and propose a modified version of it, that is more informative than the other conditioning rule.

1 Introduction

Probabilistic conditioning is used both for *prediction* from observations and *revision* of uncertain information. When dealing with prediction, we have at our disposal a model of the world under the form of a probability distribution P issued for instance from a representative set of statistical data (e.g., medical knowledge). Moreover we get some new observations C on the current state of the world (e.g., test results for a patient). Then, one tries to predict some property A of the current world with its associated degree of belief (e.g. predict the disease of the patient). The conditional probability $P(A \mid C)$ (the frequency of observation of A in context C) is used for estimating the degree of belief that the current world satisfies A.

The revision scenario is different: given a probability distribution P (which is often a subjective probability), one learns that an event C occurred, which makes its subjective probability equal to 1 (and not to $P(C) < 1$ as it was supposed before).

Didier Dubois
IRIT, CNRS and Université de Toulouse, France
e-mail: dubois@irit.fr

Thierry Denœux
HEUDIASYC, CNRS and Université de Technologie de Compiègne, France
e-mail: thierry.denoeux@hds.utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 385–392.
springerlink.com

The problem is to determine the new subjective probability measure P', such that $P'(C) = 1$, which is the closest to P in some sense, in order to comply with a minimal change principle. Then, it can be shown that if we use an appropriate relative information measure, it follows that $P'(A) = P(A \mid C), \forall A$ again [3]. However, in the prediction problem, generic knowledge remains unaffected by singular evidence, which is handled apart : predictions can be computed beforehand on the basis of the statistical probability, for each possible set of observations. Note that we do not consider the question of learning a statistical model.

For belief functions, not so much has been done from a statistical point of view, because the main references (Shafer's book [8], and most papers by Smets) present the theory of evidence as an approach to the merging of unreliable testimonies, and consider the mass function at work as a form of subjective probability. Shafer's book contains a single chapter on statistical evidence, then severely criticized, idualincluding by Shafer himself [9] (but more recently rehabilitated [2]). So, the mainstream literature on belief functions is a theory of handling singular uncertain evidence, and not so much an extension of Bayesian statistical prediction. Interestingly, Dempster's pioneering works on upper and lower probabilities are motivated by statistical reasoning and the rehabilitation of ideas of Fisher. In Dempster's setting, a probability space is given that corresponds to the usual setting of random observations for statisticians, along with a random set. Dempster [1] proposes a merging rule and two conditioning rules. However, only one of them was retained in the theory of belief functions. It was originally motivated by the presence of several independent sample spaces, not by a prediction problem. While Shafer is interested by merging independent unreliable isolated testimonies, Dempster considers the problem of merging independent bodies of statistical information. What is crucial is to notice that the merged items are of the same nature, whether singular or generic. As what is known as "Dempster conditioning" is a special case of Dempster rule of combination, this conditioning, widely used in evidence theory, can be viewed as a revision process, understood as a prioritized merging of a sure piece of information with an uncertain one. However the scenario of prediction, which involves a model reflecting a population and a piece of evidence pertaining to a single situation, features of which are to be predicted, totally escapes this revision scheme. In this paper, we distinguish between revision and prediction conditionings. The latter is applied to predicting the class θ of an instance x based on a statistical model $P(\cdot|\theta)$.

- We claim that in general, prediction cannot be achieved using Dempster conditioning. The most cautious prediction method is based on sensitivity analysis on conditional probabilities.
- Using Dempster conditioning in prediction relies on a bold application of the maximum likelihood principle [6]. The two first points were previously discussed by Dubois Prade and Smets [4].
- As a consequence, the approach to statistical prediction proposed by Smets as the Generalized Bayes Theorem (GBT) is questionable, and the prediction conditioning is inefficient. We propose a trade-off approach assuming some information on the training populations.

2 Conditioning for Belief Functions and Imprecise Probabilities

In standard probability theory, even if their solution is given by the same conditioning rule, the problem of revising a probability function is different from the one of predicting on the basis of generic statistical information. In the following we examine these two problems in the setting of belief functions.

Prediction Conditioning: In the case where the generic knowledge of the agent is represented by imprecise probabilities, Bayesian prediction is generalized by performing a sensitivity analysis on the conditional probability. It represents all predictions that could have been made, had the probabilistic model been precise. Let $\mathscr{P}$ be a family of probability measures on S. For each proposition A, a lower bound $P_*(A)$ and an upper bound $P^*(A)$ of the probability degree of A are known. In presence of singular observations summarized under the form of a context C, the belief of an agent in a proposition A is represented by the interval $[P_*(A \mid C), P^*(A \mid C)]$ (already proposed in [1]) where:

$$P_*(A \mid C) = \inf\{P(A \mid C) : P(C) > 0, P \in \mathscr{P}\}; \tag{1}$$

$P^*(A \mid C)$ is obtained by replacing inf by sup. It may happen that the interval $[P_*(A \mid C), P^*(A \mid C)]$ is larger than $[P_*(A), P^*(A)]$, which corresponds to a loss of information in specific contexts. This property reflects the idea that the more singular information is available about a situation, the less informative is the application of generic information to it (since the number of statistical data that fit this situation may become very small). We see that this form of conditioning does not correspond at all to the idea of enriching a statistical model, it is only a matter of querying it.

Belief and plausibility functions in the sense of Shafer [8] are mathematically speaking important particular cases of lower and upper probabilities, although these functions were independently introduced in Shafer's book without any reference to the idea of imprecise probability. Information is supposed to be represented by the assignment of non-negative weights $m(E)$ to subsets E of S. In a generic knowledge representation perspective, $m(E)$ is, for instance, the proportion of imprecise results of the form $x \in E$, in a statistical experiment on a random variable x (in Dempster work, it stems from a known random variable relating observations and parameters). It is clear that in that kind of situation, there exists a real probabilistic model underlying the belief function representation. It contrasts with belief functions on unique events, where there is no underlying subjective probability.

In the frequentist framework, prediction in context C requires evaluating the proportion of the population lying in C, taken as the new frame, from the information on the mass function m. Three cases should be considered for a focal set E:

- $E \subseteq C$: In this case, the frequency $m(E)$ remains assigned to E.
- $E \cap C = \emptyset$: In this case, E no longer matters and $m(E)$ is eliminated.
- $E \cap C \neq \emptyset$ and $E \cap \overline{C} \neq \emptyset$: In this case, there is a proportion $\alpha_E \cdot m(E)$ of the population that satisfies $E \cap C$ and the rest, i.e., $(1 - \alpha_E) \cdot m(E)$, satisfies $E \cap \overline{C}$. But these proportions may be unknown.

It is clear that $\alpha_E = 1$ and $\alpha_E = 0$ in the first and second above cases respectively. The third case corresponds to incomplete observations E that neither confirm nor disconfirm C. Suppose that all values $\alpha = \{\alpha_E : E \subseteq S\}$ were known. Then, we can build a mass function obtained as

$$m_\alpha^C(B) = \sum_{E:B=C\cap E} \alpha_E m(E). \tag{2}$$

Note that a renormalisation of this mass function is necessary, in general, as soon as $Pl_\alpha^C(C) < 1$, letting $m_\alpha(\cdot \mid C) = \frac{m_\alpha^C(\cdot)}{Pl_\alpha^C(C)}$. If one denotes by $Bel_\alpha(A \mid C)$ and $Pl_\alpha(A \mid C)$ the corresponding conditional belief and plausibility functions, based on the allocation vector α, we can define conservative belief and plausibility degrees conditional to C by considering all possible weight vectors α. One still obtains belief and plausibility functions (Jaffray [7]). Moreover the following result holds:

$$Bel(A \mid C) = \inf_\alpha Bel_\alpha(A \mid C) = P_*(A \mid C) = \frac{Bel(A\cap C)}{Bel(A\cap C)+Pl(\overline{A}\cap C)} \tag{3}$$

$$Pl(A \mid C) = \sup_\alpha Pl_\alpha(A \mid C) = P^*(A \mid C) = \frac{Pl(A\cap C)}{Pl(A\cap C)+Bel(\overline{A}\cap C)} \tag{4}$$

It is easy to see that $Pl(A \mid C) = 1 - Bel(\overline{A} \mid C)$, and that the closed form formulas generalize probabilistic conditioning. Note that if $Bel(C) = 0$ and $Pl(C) = 1$ (complete ignorance regarding C) then all the focal sets of m overlap C without being contained in C. In this case, $Bel(A \mid C) = 0$ and $Pl(A \mid C) = 1, \forall A \neq \emptyset, A \subset C$: one loses all information in context C.

Example: Ellsberg urn Consider a bag of balls containing $1/3$ red balls, the rest being black or white. So $S = \{w, b, r\}$ and the corresponding frequentist mass function is $m(r) = 1/3$, $m(wb) = 2/3$ (we omit the brackets to denote sets). The prediction problem consists in guessing the colour of a ball x picked at random in the urn. Before knowing anything about x, $Bel(r) = Pl(r) = 1/3; Bel(w) = 0, Pl(w) = 2/3$. Suppose one hears that x is *not black* ($C = \bar{b}$). Applying the prediction conditioning yields $Bel(r|\overline{b}) = \frac{Bel(r)}{Bel(r)+Pl(w)} = 1/3$, $Pl(r|\overline{b}) = \frac{Pl(r)}{Pl(r)+Bel(w)} = 1$, $Bel(w|\overline{b}) = \frac{Bel(w)}{Bel(w)+Pl(r)} = 0$, $Pl(w|\overline{b}) = \frac{Pl(w)}{Pl(w)+Bel(r)} = 2/3$. So the piece of information "the ball is not black" does not alter our beliefs about x being white or not. One may indeed argue it should not, as, hearing x is not black, nothing forbids the urn to contain no white ball, nor no black ball. But the plausibility of the ball being red strongly increases. This is a loss of information on the probability of the ball being red or not.

Revision Conditioning: The other conditioning, called 'Dempster conditioning' systematically assumes $\alpha_E = 1$ as soon as $E \cap C \neq \emptyset$ in the above mass transfer process. It transfers the full mass of each focal set E to $E \cap C \neq \emptyset$ (followed by a renormalisation). This means that we interpret the new information C as modifying the initial belief function in such a way that $Pl(\overline{C}) = 0$: situations where C is false are considered as impossible. If one denotes by $Pl(A \,||\, C)$ the plausibility function after revision, we have:

$$Pl(A \mid\mid C) = \frac{Pl(A \cap C)}{Pl(C)} \tag{5}$$

The conditional belief is then obtained by duality as $Bel(A \mid\mid C) = 1 - Pl(\overline{A} \mid\mid C)$. Note that with this conditioning, the size of focal sets diminishes, thus information becomes more precise, and the intervals $[Bel, Pl]$ may become tighter than those obtained by prediction conditioning. Dempster conditioning thus corresponds to a process where information is enriched, which contrasts with prediction conditioning. If $Bel(C) = 0$ and $Pl(C) = 1$ (complete ignorance about C), Dempster conditioning on C will often significantly increases the precision of resulting beliefs.

In the more general framework of imprecise probabilities, the revision by a piece of information C consists in adding the extra constraint $P(C) = Pl(C)$ to the family $\mathscr{P} = \{P \geq Bel\}$. It has been shown by Gilboa and Schmeidler [6] that:

$$P_*(A \mid\mid C) = \inf\{P(A \mid C), P(C) = Pl(C), P \geq Bel\}; \tag{6}$$

$$P^*(A \mid\mid C) = \sup\{P(A \mid C), P(C) = Pl(C), P \geq Bel\}. \tag{7}$$

They indicate that Dempster conditioning comes down to applying the maximal likelihood principle in the imprecise probability setting.

In the view of Shafer and Smets, this type of conditioning is little related with the previous prediction problem, since, in their setting, the mass function m does not represent generic knowledge, but rather uncertain information collected about a particular situation (non fully reliable testimonies, more or less safe clues). It is a form of reasoning under uncertainty where generic knowledge is not taken into account, but where all the pieces of information are singular (as in the Peter Paul and Mary example [4]).

Example: Ellsberg again Hearing that there is no black ball in the urn is a piece of generic information of the same nature as the prior knowledge about the urn. There are then two independent sample spaces as assumed by Dempster [1] in his pioneering paper (one with the possibility of back balls, one without it). We then revise the content of the urn, which in turn impacts a change of belief about the colour of the picked ball x. We then legitimately conclude that since the urn does not contain black balls, the probability of x being white is $2/3$. It may look questionable to apply this conditioning rule to the problem of predicting the colour of a ball x drawn from the urn, based on the fact that it is not black. Using the maximum likelihood interpretation of Dempster conditioning, doing so comes down to assuming that since the ball is known not to be black, we restrict to the probabilistic models P making this event maximally likely (the ones such that $P(\bar{b}) = Pl(\bar{b})$). Then, $P(\bar{b})$ is viewed as the likelihood of the probabilistic model $P \geq Bel$ if the ball is not black.

There is one case when it is easy to show that the two forms of conditionings coincide:

Proposition 1. *If the conditioning event C is such that for any focal set E of m it either contains it or is disjoint from it, then $\forall A \subseteq S, Pl(A \mid\mid C) = Pl(A \mid C)$.*

3 Application to Smets Generalized Bayes Theorem

Despite the warning of Shafer regarding the fact that his theory of evidence deals with unique uncertain events, it has been applied to statistical prediction problems, and the appropriateness of Dempster conditioning for such a task is most of the time not even questioned. A typical example is the Generalized Bayes Theorem of Smets [10], but it is true as well for various other similar approaches to the estimation of parameters based on a belief function model (see [2] for more bibliography).

In the simplest setting, the parametric inference problem is stated as follows : Given a **finite** parameter space Θ and a set of parametric belief functions $Bel_X(\cdot|\theta)$, $\theta \in \Theta$, and some observation $x \in X$, compute $Bel_\Theta(\cdot|x)$.

The most usual situation is when a finite number of probabilistic likelihood functions $\{P(\cdot|\theta), \theta \in \Theta\}$, are available, each one coming from a different population representing a class θ. The GBT procedure specializes as follows:

1. **Conditional embedding:** Each likelihood function $P(\cdot|\theta)$ is modelled by a belief function Bel^θ on $X \times \Theta$ (ballooning): the associated mass function is defined by $m^\theta(\bar{\theta} \cup x) = P(x|\theta), x \in X$; Bel^θ on $X \times \Theta$ has a vacuous marginal on Θ and yields $P(\cdot|\theta)$ back when conditioned on θ.
2. **Conjunctive merging** of the belief functions $Bel^\theta, \theta \in \Theta$ on $X \times \Theta$. Consider a function $\phi : \Theta \to X$; we must assign mass $\prod_{\theta\in\Theta} P(\phi(\theta)|\theta)$ to $\bigcap_{\theta\in\Theta} \bar{\theta} \cup \phi(\theta) = \bigcup_{\theta\in\Theta} \{\theta\} \times \{\phi(\theta)\}$. So, each function ϕ is a focal element and the resulting mass function on $\Theta \times X$ is of the form $m(\phi) = \prod_{\theta\in\Theta} P(\phi(\theta)|\theta)$[1].
3. **Conditioning** of the result on the observation x using Dempster conditioning:

$$Pl_\Theta(\theta||x) = \frac{Pl(\{\theta\} \times \{x\})}{Pl(\Theta \times \{x\})} = \frac{\sum_{\phi:\phi(\theta)=x} m(\phi)}{\sum_{\phi:\phi^{-1}(x)\neq\emptyset} m(\phi)} = \frac{\sum_{\phi:\phi(\theta)=x} \prod_{\tau\in\Theta} P(\phi(\tau)|\tau)}{\sum_{\phi:\phi^{-1}(x)\neq\emptyset} \prod_{\tau\in\Theta} P(\phi(\tau)|\tau)}.$$

   The result is a general belief function not fully representable by the $Pl_\Theta(\theta||x)$'s, since the mass function $m(T||x)$ is of the form $\frac{\sum_{\phi:\phi^{-1}(x)=T} m(\phi)}{Pl(\Theta\times\{x\})}$.

Step 2 comes down to applying the disjunctive combination rule to the conditional probabilities $P(\cdot|\theta)$: $Bel(A \times T) = \prod_{\theta\in T} P(A|\theta), \forall A \subseteq X$. For $T \subseteq \Theta, Pl_X(x|T)$ is a function of elementary likelihoods $P(x|\theta), \theta \in T$. The merging rule in step 2 assumes that the likelihood functions $P(\cdot|\theta)$, $\theta \in \Theta$ have been inferred from distinct sets of empirical data obtained from independent sources. Hence, each value θ corresponds to a specific class of objects, and is not a continuous parameter.

However, x is a single observation while m is a statistical model on $X \times \Theta$. Let us then compute the prediction conditioning $Pl_\Theta(\theta|x)$, the plausibility that θ is the class of x. Since a focal element is in the form of a mapping $\phi : \Theta \to X$, and observation x is modelled by $\{x\} \times \Theta$ three situations can be met for a focal ϕ:

[1] Unions are intersections over different spaces consider cylindrical extensions of elements. In the more general case of conditional belief functions $Bel_X(\cdot|\theta)$, focal elements on $\Theta \times X$ are multimappings $\Gamma : \Theta \to 2^X$.

1. $\nexists\theta \in \Theta, \phi(\theta) = x$, which means $\phi^{-1}(x) = \emptyset$; ϕ can be eliminated.
2. $\phi^{-1}(x) \neq \emptyset$ and $\neq \Theta$, then the line $\{x\} \times \Theta$ only overlaps the graph of ϕ.
3. $\phi(\theta) = x, \forall\theta \in \Theta$ then $\phi = \{x\} \times \Theta$ supports x but it gives no clue on θ.

When observing x and considering a focal ϕ consistent with it (case 2), it is not clear what part of $m(\phi)$ should be allocated to $\phi^{-1}(x)$ and what part should be allocated to its complement in Θ. Applying the prediction conditioning would then yield an empty prediction since $\forall T \subset \Theta, Bel(T \times \{x\}) = Bel(\overline{T} \times \{x\}) = 0$, because $\phi \subseteq T \times \{x\}$ never holds ($\forall\theta \in \Theta, \phi(\theta) \neq \emptyset$). However, it is possible to propose an alternative prediction rule that is less bold than Dempster conditioning and more useful than the plain prediction conditioning.

The prior probabilities $P(\theta)$'s are unknown (for Edwards [5], they are even meaningless). It is then tempting to replace $P(\theta)$ by the number of observations $n(\theta)$ available for class θ. This number does not necessarily correspond to a prior probability (as the actual probability $P(\theta)$, if any, is different from the number of cases actually at hand). Yet, $n(\theta)$ if available reflects the reliability of the information regarding the likelihood function $P(\cdot|\theta)$.

Since $m(\phi)$ accounts for all $m_X(\phi(\theta)|\theta)$, one may consider, when observing x, sharing $m(\phi)$ between all θ such that $\phi(\theta) = x$, and those such that $\phi(\theta) \neq x$, according to $n(\theta)$. Then consider the portion $\alpha_\phi(x) = \frac{\sum_{\theta\in\phi^{-1}(x)} n(\theta)}{\sum_{\theta\in\Theta} n(\theta)}$ of the available training data that pertains to observing x. It suggests the following modified conditional belief function for prediction:

$$\forall T \subseteq \Theta, m_\alpha(T|x) = \frac{\sum_{\phi:\phi^{-1}(x)=T} m(\phi)\alpha_\phi(x)}{\sum_{\phi:\phi^{-1}(x)\neq\emptyset} m(\phi)\alpha_\phi(x)}. \tag{8}$$

Example: The simplest example of the problem (actually studied by Shafer [9]) uses a space $\mathscr{S} = \{x, \bar{x}\} \times \{\theta, \bar{\theta}\}$. The available knowledge consists in the two likelihood values $a = P(x|\theta) > b = P(x|\bar{\theta})$. So there is a majority of x's in class θ and a majority of $\overline{x}$'s in class $\bar{\theta}$. Suppose that the likelihood functions are based on independent populations and that the number of available samples of class θ is much greater than those for class $\bar{\theta}$ (say 1000 times more). There are only 4 ϕ functions with their mass assignments on $\Theta \times X$ shown in the table below.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	
θ	x	x	$\overline{x}$	$\overline{x}$	$n(\theta) = 3000$
$\overline{\theta}$	x	$\overline{x}$	x	$\overline{x}$	$n(\overline{\theta}) = 3$
$P(\cdot\|\theta)$	$a = 2/3$	$a = 2/3$	$1-a = 1/3$	$1-a = 1/3$	
$P(\cdot\|\overline{\theta})$	$b = 1/3$	$1-b = 2/3$	$b = 1/3$	$1-b = 2/3$	
$m(\phi)$	$ab = 2/9$	$a(1-b) = 4/9$	$(1-a)b = 1/9$	$(1-a)(1-b) = 2/9$	
$\alpha_\phi(x)$	1	1000/1001	1/1001	0	
$\alpha_\phi(\overline{x})$	0	1/1001	1000/1001	1	

The following results are obtained if x is observed using Dempster maximal likelihood conditioning:

$$Pl_\Theta(\theta||x) = \frac{a}{a+b-ab} = \frac{6}{7}; Pl_\Theta(\bar{\theta}||x) = \frac{b}{a+b-ab} = \frac{3}{7}. \tag{9}$$

Note that if $\overline{x}$ is observed, the figures are exchanged. However, this symmetry is surprising: since the data set of class $\overline{\theta}$ is very poor, the observation of $\overline{x}$ should suggest class $\overline{\theta}$ to a lesser extent than the one to which observing x should suggest class θ. The last two lines of the table above show the proportions of the overall available population of examples concerned when observing x and $\overline{x}$, given any focal element ϕ. It comes down to assuming that if we observe x, the portion of weight of ϕ_3 to be transferred to $\bar{\theta} \wedge x$ should be much less than the portion to be transferred to $\theta \wedge x$. Conversely if we observe $\bar{x}$, the portion of weight of ϕ_2 to be transferred to $\bar{\theta} \wedge \bar{x}$ should be much less than the portion to be transferred to $\theta \wedge \bar{x}$. So, if x is observed the modified conditional mass $m(\phi_3|\overline{x})$ is reduced to $1/9009$ and becomes negligible: $Pl_\alpha(\theta|x) \simeq 1; Pl_\alpha(\bar{\theta}|x) \simeq 1/3$. If $\overline{x}$ is observed, the modified conditional mass $m(\phi_2|\overline{x})$ is reduced to $4/9009$ so $Pl_\alpha(\theta|\overline{x}) \simeq 1; Pl_\alpha(\bar{\theta}|\overline{x}) \simeq 2/3$. It makes it clear that, as expected, we become more confident about class θ when observing x than about class $\overline{\theta}$ when observing $\overline{x}$. We even still believe in class θ in the latter case, due to the overwhelming number of θ examples. Note that having many more examples of class θ than of class $\overline{\theta}$ does not mean that class $\overline{\theta}$ is rare, but only that we could not have access to many examples of it. So we should not confuse the size of available samples with prior probabilities of classes. The above discussion also lays bare that the GBT approach presupposes not only independent populations of samples for each class, but also that such populations are (approximately) equal.

To conclude, we suggest how to improve the GBT so as to make a trade-off between prediction and revision conditioning. The case of prediction based on several pieces of observations can be addressed by merging the information coming from each observation. Other techniques for modelling statistical evidence [2] should be studied in the light of the above discussion as well.

References

1. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. 38, 325–339 (1967)
2. Dubois, D., Denoeux, T.: Statistical inference with belief functions and possibility measures: a discussion of basic assumptions. In: Advances in Intelligent and Soft Computing, SMPS 2010, vol. 77, pp. 217–225. Springer (2010)
3. Domotor, Z.: Probability kinematics - conditional and entropy principles. Synthese 63, 74–115 (1985)
4. Dubois, D., Prade, H., Smets, P.: Representing partial ignorance. IEEE Trans. on Systems, Man and Cybernetics 26(3), 361–377 (1996)
5. Edwards, W.F.: Likelihood. Cambridge University Press, Cambridge (1972)
6. Gilboa, I., Schmeidler, D.: Updating ambiguous beliefs. In: Moses, Y. (ed.) Proc. of the 4th Conf. Theor. Aspects of Reasoning About Knowledge (TARK 1992), pp. 143–162 (1992)
7. Jaffray, J.Y.: Bayesian updating and belief functions. IEEE Trans. on Systems, Man and Cybernetics 22, 1144–1152 (1992)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
9. Shafer, G.: Belief Functions and Parametric Models. Journal of the Royal Statistical Society. Series B 44, 322–352 (1982)
10. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. Int. J. Approx. Reasoning 9(1), 1–35 (1993)

Combining Statistical and Expert Evidence within the D-S Framework: Application to Hydrological Return Level Estimation

Nadia Ben Abdallah, Nassima Mouhous Voyneau, and Thierry Denœux

Abstract. Estimation of extreme sea levels and waves for high return periods is of prime importance in hydrological design and flood risk assessment. The common practice consists of inferring design levels from the available observations and assuming the distribution of extreme values to be stationary. However, in the recent decades, more concern has been given to the integration of the effect of climate change in environmental analysis. When estimating defense structure design parameters, sea level rise projections provided by experts now have to be combined with historical observations. Due to limited knowledge about the future world and the climate system, and also to the lack of sufficient sea records, uncertainty involved in extrapolating beyond available data and projecting in the future is considerable and should absolutely be accounted for in the estimation of design values.

In this paper, we present a methodology based on evidence theory to represent statistical and expert evidence in the estimation of future extreme sea return level associated to a given return period. We represent the statistical evidence by likelihood-based belief functions [7] and the sea level rise projections provided by two sets of experts by a trapezoidal possibility distribution. A Monte Carlo simulation allows us to combine both belief measures to compute the future return level and a measure of the uncertainty of the estimations.

1 Introduction

Comprehensive uncertainty analysis is a key part of design and safety assessment procedures for reliable results and optimal decision. In the hydrological field,

Nadia Ben Abdallah · Thierry Denoeux
Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc
e-mail: nadia.ben-abdallah@hds.utc.fr,
thierry.denoeux@hds.utc.fr

Nassima Mouhous-Voyneau
Université de Technologie de Compiègne, Avenues-GSU
e-mail: nassima.voyneau@utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 393–400.
springerlink.com

communicating the uncertainty about future flood risk to the decision makers is becoming the rule rather than the exception [1, 12]. If there is a general consensus about the relevant sources of uncertainty within a flood risk analysis, there is an increasing debate among risk analysts about the framework to use for quantifying it. The commonly used probabilistic framework has been strongly criticized for treating in the same way aleatory uncertainty that characterizes natural variability and epistemic uncertainty resulting from limited knowledge [1]. Given that, in environmental risk analysis, these uncertainties usually arise from different sources (statistical evidence, expert opinion...), the need for alternative frameworks to address differently both kinds of uncertainty emerged. Intensive works are investigating the appropriateness of approaches such as fuzzy set theory, imprecise probability or Dempster-Shafer theory in assessing reliability and risk analyses.

In this paper, we are interested in modeling the uncertainty pertaining to the design parameters of flood defense structures in the context of future climate change. Evidence for estimating the parameter of interest and its uncertainty originates from two sources of different natures. The first one is related to statistical evidence commonly expressed by frequentist or Bayesian approach, the relevance of which has been increasingly criticized [5, 8]. The second one concerns projections of climate change and its impacts in terms of sea level rise, which have to be assessed by climate experts. Partial disagreement about the future climate change within the climate community leads, as will be showed later, to a high level of uncertainty attached to the projections available in the literature.

We propose to represent and combine the two different sources of evidence (data and experts) using the DS framework, given its ability to address in a unified mathematical context different sources of evidence and the tools it offers to combine them.

The paper is organized as follows. In the first section, we review the use of extreme value statistics in hydrology and the characteristics of hydrologic extremes in flood design. We briefly address the issue of climate change impacts and present the main projections on the future sea level rise existing in the literature. In the second part, we justify and explain the use of likelihood-based inference to represent statistical evidence and briefly address its connection with the DS framework. Finally, the last part describes the application of the methodology and summarizes the main results.

2 Key Elements on Hydrology and Climate Change

Flood structures have to withstand exceptional sea events and their design has thus to be based on extreme sea level and waves. The main tool for modeling extreme events in environmental applications such as floods, droughts or rainfalls is Extreme Value Theory (EVT), which has emerged giving the limit of the conventional frequency analysis in fitting the tails of probability distributions. The block maxima approach is the original and best known method in EVT. It is based on the assumption that the maximum of an independent and identically distributed

(i.i.d.) sample has asymptotically a generalized extreme value (GEV) distribution [11]. The cumulative distribution function of the GEV distribution is given by:

$$F(z,\mu,\sigma,\xi) = \begin{cases} \exp\left(-\left[1-\xi\frac{z-\mu}{\sigma}\right]^{\frac{1}{\xi}}\right) & \text{for } \xi \neq 0 \\ \exp\left(-\left(\exp\left[-\frac{z-\mu}{\sigma}\right]\right)\right) & \text{for } \xi = 0, \end{cases} \quad (1)$$

where $\mu, \sigma > 0, \xi$ are, respectively, location, scale and shape parameters. According to the sign of ξ, the distribution is called Fréchet ($\xi > 0$), Weibull ($\xi < 0$) or Gumbel ($\xi = 0$).

In extreme-values studies, the probability of exceedance of a certain value z is usually expressed in terms of the return period T, defined as the average number of years between two successive exceedances of the corresponding return value z. Within the annual maxima method, the return period of a given level z is directly related to its annual non exceedance probability p by the relation: $T = 1/(1-p)$. Therefore, we get from (1) the following expression of the return level z_T associated to a given return period T:

$$z_T = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\left(-\mathrm{Log}\left(1-\frac{1}{T}\right)\right)^{-\xi}\right] & \text{for } \xi \neq 0 \\ \mu - \sigma\,\mathrm{Log}\left(-\mathrm{Log}\left(1-\frac{1}{T}\right)\right) & \text{for } \xi = 0, \end{cases} \quad (2)$$

The only available evidence when estimating extreme quantiles is derived from the historical observations.

Commonly, flood defenses in coastal areas are designed to withstand at least 100-year events. However, due to climate change, they will be subject during their life time to higher loads than the design estimations. The main impact is related to the increase of the mean sea level which affects the frequency and intensity of surges. For adaptation purposes, the present statistics of extreme sea levels derived from the observations should be combined with the projections of the future sea level rise (SLR).

Future SLR projections provided by the IPCC's (International Panel of Climate Change Experts) last Assessment Report [10] assess the likely range of values for sea-level rise over the 1990-2095 period as 0.18 to 0.79 m; it is indicated in this report that higher values should not be excluded. This range takes into account uncertainties associated to future emissions of greenhouse gases (GHGs) corresponding to the SRES (Special Report Emission Scenarios) (scenarios that cover a wide range of possible economic, technological and energetic states of the world), global circulation models used to estimate future temperature projections and impacts models (melting of the Antarctic and Greenland, oceans expansion, etc.).

Since the release of the last IPCC report, other sea level rise assessments based on semi-empirical models have been undertaken, proposing more pessimistic sea level rise scenarios for 2100. For example, based on a simple statistical model,

Rahmstorf [15] suggests [0.5m, 1.4 m] as a likely range of values for sea-level rise at the end of this century. However, recent studies showed that there is a physical limit to the sea level rise in the coming years: the threshold of 2 m could not be exceeded by the end of this century [13].

Current methods for integrating future SLR in flood risk or design analysis have considered a deterministic particular sea level rise scenario since there is no information to quantify the probability of any given sea level magnitude within the IPCC range. However, as shown by Purvis [14], who undertook a flood risk analysis under climate change, using the most plausible sea level rise scenario may significantly underestimate effective consequences and lead to erroneous decisions.

For estimating design level under climate change, we proceed in two steps: we first infer the current design level from statistical evidence (available sea level measurements). In a second step, we integrate expert judgment on future sea level rise.

3 Likelihood-Based Representation of Statistical Evidence

The estimated level is usually obtained from (2) by replacing the probability distribution parameters by their best estimates. Commonly, parameters are estimated using frequentist methods. However, these methods are based on asymptotic properties and their performance turns to be quite poor when we deal with small samples. As for the estimations, the confidence intervals supposed to inform about the level of uncertainty within the estimations are quite unreliable because of the very crude approximations in the calculation of the upper and lower bounds of the confidence interval [18]. In fact, confidence intervals are based on the repeated sampling hypothesis which consists of hypothetically repeating the particular experiment and derive accordingly the confidence bounds. In cases such that the repetition is not possible, this approach can be questioned and alternative approaches to effectively represent the available evidence are needed.

Authors such as Fisher [8], Cox [5], Barnard *et al.* [3] and Edwards [7] have criticized the frequentist approach for its inappropriate use of significance levels, confidence intervals and other repeated-sampling criteria to represent evidence and have advocated a new, more 'evidential' approach to statistical inference that uses only the likelihood function.

The likelihood-based inference approach relies on the likelihood principle, which states that given an observation X, the relevant information about an unknown parameter $\theta (\theta \in \Theta)$ (possibly a vector) is all contained in the likelihood function for the observed sample X, denoted $L(\theta; \mathrm{X})$. Recall that the likelihood is a function of the parameters of a statistical model $f(x; \theta)$ defined as follows: given some observed outcomes, the likelihood of a set of parameters is equal to the probability of the observations given those parameters. Thus $L(\theta; \mathrm{X}) = f(X; \theta)$.

The representation of statistical evidence in the belief function framework is motivated by the fact that belief functions form a richer set of functions than probability measures: it is thus expected that inference, when based on belief functions, would allow us to model a wider range of uncertainty than probabilities. Shafer [17] was the first to propose to represent likelihood information as a consonant belief function about the parameters. Shafer's method was later justified axiomatically by Wasserman [19]; additional arguments for its use in statistical inference were provided by Aickin [2]. Fisher [8] interprets the likelihood function as an expression of the relative plausibility of the parameters when no additional information, except the observations, is available. It thus seems reasonable to define the plausibility contour function (or credibility function), when the likelihood is bounded, as:

$$pl(\theta; X) = \frac{L(\theta;X)}{L(\hat{\theta};X)} \tag{3}$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ. The associated plausibility is easily computed for every subset $A \in \Theta$ as:

$$Pl(A) = \sup\{pl(\theta; X);\ \theta \in A\} \tag{4}$$

The contour function (3) has a simple interpretation: $pl(\theta; x)$ represents the probability of observing x if the true parameter value is θ, relative to the maximum probability of observing x for any value of θ. A parameter value with low plausibility, say 0.001, indicates that there are other values of θ which ensure a 1000 times higher probability to observe x.

The set $\{\theta \in \Theta / pl(\theta, X) \geq \alpha\}$, called the α-level likelihood region, allows us to characterize ranges of implausible values (for example, values ranging outside 5% likelihood region) and very plausible values.

4 Application and Results

As a case study, we applied the likelihood-based inference method described above to infer the design variable z_{100} from the sample of observations X corresponding to 15 years of hourly records of sea level (observations from tide gauges at le Havre harbor, France). This measure was estimated under the assumption that the annual maxima have a Gumbel distribution (2); here, μ is the structural parameter and σ the nuisance one. The latter was eliminated through a profile likelihood approach. The corresponding contour function is shown in Figure 1.The most plausible value characterized by a plausibility level equal to 1 corresponds to the maximum likelihood estimate.

In a second step, we integrated the uncertain effect of climate change in terms of SLR (in meters) to estimate the future return level (at the end of the century)

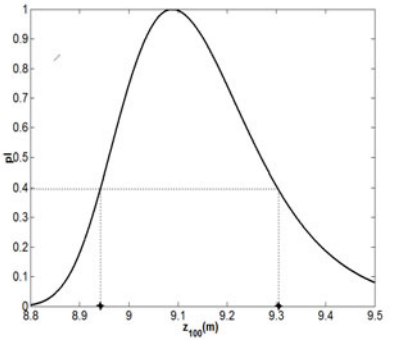


Fig. 1 Plausibility measure of the design parameter z_{100}

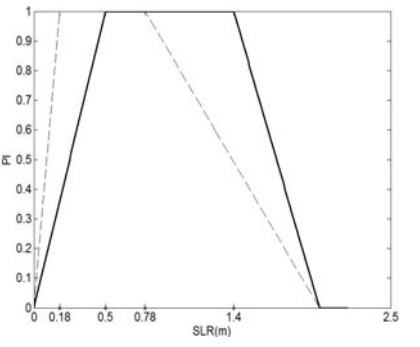


Fig. 2 SLR Trapezoidal Possibility inferred measures (in continuous bold line: inference based on Rahmstorf evidence. In dashed line: inference based on IPCC evidence)

associated to the same return period. As sources of information about the SLR, we considered projections by the IPCC and Rahmstorf [15] estimations provided above as the current best available evidence. We formalized each of these pieces of evidence by a trapezoidal possibility measure that represents our best interpretation of the expert's estimations (Figure 2). Since both sources are reliable, a conjunctive aggregation is applicable. Among the conjunctive rules, the minimum is the most appropriate when the opinions of the experts are based on a common knowledge: we thus applied this rule to derive the aggregated SLR possibility distribution.

Finally we computed the belief function on the future design level $z_{100}^f = z_{100} + SLR$ using a Monte Carlo sampling procedure. This procedure consists in randomly drawing plausibility levels α and possibility levels ω using independent uniform distributions For every random α and ω, we associate the α and ω likelihood regions $[\underline{z_{100}^{\alpha}}, \overline{z_{100}^{\alpha}}]$ and $[\underline{SLR^{\omega}}, \overline{SLR^{\omega}}]$; the corresponding design level z_{100}^f is within $[\underline{z_{100}^{\alpha}} + \underline{SLR^{\omega}}, \overline{z_{100}^{\alpha}} + \overline{SLR^{\omega}}]$. This procedure was repeated a thousand times. From these simulated intervals, we can calculate for a fixed level z_{100}^f, the cumulative plausibility and belief. The cumulative plausibility corresponds to the relative frequency, over the simulations, of the event "the lower bound is less than the fixed level", whereas the cumulative belief corresponds to the relative frequency of the event "the upper bound is less than the fixed level". Figure 3 shows the cumulative plausibility and belief functions of the current and future return level (respectively in dashed and solid line). The upper curve corresponds to the plausibility function and the lower one to the belief measure.

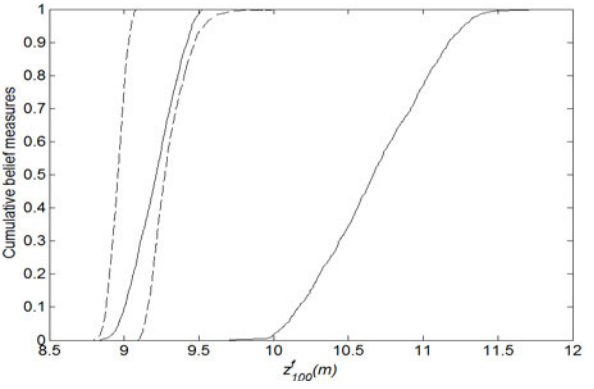


Fig. 3 Cumulative belief functions for the current (in dashed line) and future (in solid line) return level: the lower distribution is the belief; the upper one is the plausibility

The area between the belief and the plausibility dashed curves can be interpreted as a measure of the total uncertainty. When climate change is considered in the estimation of the future level, the area becomes very large, reflecting the important uncertainty associated with the SLR projections.

5 Conclusion

The Dempster-Shafer theory of belief functions places emphasis on the representation of evidence for evaluating degrees of belief. The generality and flexibility of this framework makes it suitable for representing and combining expert judgments and statistical evidence. In this paper, this approach has been applied to the estimation of the centennial sea level at a particular location, taking into account historical data and expert assessments of sea level rise by the end of the century. This work is part of a larger project that aims at defining engineering design processes for the adaptation of coastal infrastructure to climate change.

References

1. Apel, H., Thikeken, A.H.: Flood risk assessment and associated uncertainty. Natural Hazards and Earth Sysem Sciences 4, 295–308 (2004)
2. Aickin, M.: Connecting Dempster-Shafer belief functions with likelihood based inference. Synthese 123, 347–364 (2000)
3. Barnard, G.A., Jenkins, G.M., Winsten, C.B.: Likelihood inference and time series. Journal of the Royal Statistical Society 125(3), 321–372 (1962)
4. Coles, S.G., Dixon, M.J.: Likelihood based inference for extreme value models. Extremes 2, 5–23 (1999)
5. Cox, D.R.: Some problems connected with statistical inference. Ann. Math. Statistics 29, 357–372 (1958)

6. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Transactions on Knowledge and Data Engineering (to appear), doi:10.1109/TKDE.2011.201
7. Edwards, A.W.F.: Likelihood. University Press, Baltimore (1972)
8. Fisher, R.A.: Inverse Probability and the use of likelihood. Proceedings of the Cambridge Philosophical Society 28, 257–261, CP3 (1932)
9. Gumbel, E.J.: The statistics of extremes. Colombia University Press, New York (1958)
10. IPCC Forth Assessment Report (2007), http://www.ipcc.ch/publications_and_data/publications_and_data_reports.shtml
11. Jenkinson, A.F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quart. J. Roy. Meteo. Soc. 81, 158–171 (1955)
12. Merz, B., et al.: Flood risk curves and uncertainty bounds. Natural Hazards 51, 437–458 (2009)
13. Pfeffer, W., Harper, J., O'Neel, S.: Kinematic constraints on glacier contribution to 21st century sea level rise. Science 321, 1340–1430 (2008), doi:10.1126/science.1159099
14. Purvis, M.: Probabilistic methodology to estimate future coastal flood risk due to sea level rise. Coastal Engineering 55, 1062–1073 (2008)
15. Rhamstorf, S.: A semi empirical approach to projecting future sea level rise. Science 315, 368–370 (2007)
16. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press (1976)
17. Shafer, G.: Belief Functions and Parametric Models. Journal of the Royal Statistical Society. Series B. 44, 322–352 (1982)
18. Xu, P., et al.: Uncertainty analysis in statistical modeling of extreme hydrological events. Stochastic Environmental Research and Risk Assessment 24, 567–578 (2010)
19. Wasserman, L.: Belief functions and statistical inference. The Canadian Journal of Statistics 18, 183–196 (1990)

Sigmoidal Model for Belief Function-Based Electre Tri Method

Jean Dezert and Jean-Marc Tacnet

Abstract. Main decision-making problems can be described into choice, ranking or sorting of a set of alternatives or solutions. The principle of Electre TRI (ET) method is to sort alternatives a_i according to criteria g_j into categories C_h whose lower and upper limits are respectively b_h and b_{h+1}. The sorting procedure is based on the evaluations of outranking relations based firstly on calculation of partial concordance and discordance indexes and secondly on global concordance and credibility indexes. In this paper, we propose to replace the calculation of the original concordance and discordance indexes of ET method by a more effective sigmoidal model. Such model is part of a new Belief Function ET (BF-ET) method under development and allows a comprehensive, elegant and continuous mathematical representation of degree of concordance, discordance and the uncertainty level which is not directly taken into account explicitly in the classical Electre Tri.

1 Introduction

The Electre Tri (ET) method, developed by Yu [13], remains one of the most successful and applied methods for multiple criteria decision aiding (MCDA) sorting problems [5]. ET method assigns a set of given alternatives $a_i \in \mathbf{A}$, $i = 1,2,\ldots,n$ according to criteria g_j, $j = 1,2,\ldots,m$ to a pre-defined (and ordered) set of categories $C_h \in \mathbf{C}$, $h = 1,2,\ldots,p+1$ whose lower and upper limits are respectively b_h and b_{h+1} for all $h = 1,\ldots,p$), with $b_0 \leq b_1 \leq b_2 \leq \ldots \leq b_{h-1} \leq b_h \leq \ldots \leq b_p$. The assignment of an alternative a_i to a category C_h (limited by profiles b_h and $b_h + 1$) consists in four steps involving at first the computation of global concordance $c(a_i, b_h)$ and discordance $d(a_i, b_h)$ indexes[1] (steps 1 & 2), secondly their fusion into a credibility

Jean Dezert
The French Aerospace Lab, F-91761 Palaiseau, France
e-mail: jean.dezert@onera.fr

Jean-Marc Tacnet
Irstea, UR ETGR, 2 rue de la papeterie-B.P. 76, F-38402 St-Martin-d'Hères, France
e-mail: jean-marc.tacnet@irstea.fr

[1] Themselves computed from partial concordance and discordance indexes based on a given set criteria $g_j(.)$, $j \in \mathbf{J}$.

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 401–408.
springerlink.com

index $\rho(a_i,b_h)$ (step 3), and finally the decision and choice of the category based on the evaluations of outranking relations [13, 6] (step 4). The partial concordance index $c_j(a_i,b_h)$ measures the concordance of a_i and b_h in the assertion *"a_i is at least as good as b_h"*. The partial discordance index $d_j(a_i,b_h)$ measures the opposition of a_i and b_h in the assertion *"a_i is at least as good as b_h"*. The global concordance index $c(a_i,b_h)$ measures the concordance of a_i and b_h on all criteria in the assertion *"a_i outranks b_h"*. The degree of credibility of the outranking relation denoted as $\rho(a_i,b_h)$ expresses to which extent *"a_i outranks b_h"* according to $c(a_i,b_h)$ and $d_j(a_i,b_h)$ for all criteria. The main steps of ET method are described below:

1. **Concordance Index:** The concordance index $c(a_i,b_h) \in [0,1]$ between the alternative a_i and the category C_h is computed as the weighted average of partial concordance indexes $c_j(a_i,b_h)$, that is

$$c(a_i,b_h) = \sum_{j\in\mathbf{J}} w_j c_j(a_i,b_h) \tag{1}$$

where the weights $w_i \in [0,1]$ represent the relative importance of each criterion $g_j(.)$ in the evaluation of the global concordance index. They must satisfy $\sum_{j\in\mathbf{J}} w_j = 1$. The partial concordance index $c_j(a_i,b_h) \in [0,1]$ based on a given criterion $g_j(.)$ is computed from the difference of the criteria evaluated for the profil b_h, and the criterion evaluated for the alternative a_i. If the difference $g_j(b_h) - g_j(a_i)$ is less (or equal) to a given preference threshold $q_j(g_j(b_h))$ then a_i and C_h are considered as different based on the criterion $g_j(.)$ so that a preference of a_i with respect to C_h can be clearly done. If the difference $g_j(b_h) - g_j(a_i)$ is strictly greater to another given threshold $p_j(g_j(b_h))$ then a_i and C_h are considered as indifferent (similar) based on $g_j(.)$). When $g_j(b_h) - g_j(a_i) \in [q_j(g_j(b_h)), p_j(g_j(b_h))]$, the partial concordance index $c_j(a_i,b_h)$ is computed from a linear interpolation. Mathematically, the partial concordance index is obtained by:

$$c_j(a_i,b_h) \triangleq \begin{cases} 1 & \text{if} \quad g_j(b_h) - g_j(a_i) \le q_j(g_j(b_h)) \\ 0 & \text{if} \quad g_j(b_h) - g_j(a_i) > p_j(g_j(b_h)) \\ \frac{g_j(a_i) + p_j(g_j(b_h)) - g_j(b_h)}{p_j(g_j(b_h)) - q_j(g_j(b_h))} & \text{otherwise} \end{cases} \tag{2}$$

2. **Discordance Index:** The discordance index between the alternative a_i and the category C_h depends on a possible veto condition expressed by the choice of a veto threshold $v_j(g_j(b_h))$ imposed on some criterion $g_j(.)$. The (global) discordance index $d(a_i,b_h)$ is computed from the partial discordance indexes:

$$d_j(a_i,b_h) \triangleq \begin{cases} 1 & \text{if} \quad g_j(b_h) - g_j(a_i) > v_j(g_j(b_h)) \\ 0 & \text{if} \quad g_j(b_h) - g_j(a_i) \le p_j(g_j(b_h)) \\ \frac{g_j(b_h) - g_j(a_i) - p_j(g_j(b_h))}{v_j(g_j(b_h)) - p_j(g_j(b_h))} & \text{otherwise} \end{cases} \tag{3}$$

One defines by $\mathbf{V}$ the set of indexes $j \in \mathbf{J}$ where the veto applies (where the partial discordance index is greater than the global concordance index), that is

$$\mathbf{V} \triangleq \{j \in \mathbf{J} | d_j(a_i, b_h) > c(a_i, b_h)\} \tag{4}$$

Then a global discordance index can be defined [12] as

$$d(a_i, b_h) \triangleq \begin{cases} 1 & \text{if} \quad \mathbf{V} = \emptyset \\ \prod_{j \in \mathbf{V}} \frac{1 - d_j(a_i, b_h)}{1 - c_j(a_i, b_h)} & \text{if} \quad \mathbf{V} \neq \emptyset \end{cases} \tag{5}$$

3. **Global Credibility Index:** In ET method, the (global) credibility index $\rho(a_i, b_h)$ is computed by the simple discounting of the concordance index $c(a_i, b_h)$ given by (11) by the discordance index (discounting factor) $d(a_i, b_h)$ given in (5). Mathematically, this is given by

$$\rho(a_i, b_h) = c(a_i, b_h) d(a_i, b_h) \tag{6}$$

4. **Assignment Procedure:** The assignment of a given action a_i to a certain category C_h results from the comparison of a_i to the profile defining the lower and upper limits of the categories. For a given category limit b_h, this comparison relies on the credibility of the assertions a_i outranks b_h. Once all credibility indexes $\rho(a_i, b_h)$ for $i = 1, 2, \ldots, m$ and $h = 1, 2, \ldots, k$ have been computed, the assignment matrix $\mathbf{M} \triangleq [\rho(a_i, b_h)]$ is available for helping in the final decision-making process. In ELECTRE TRI method, a simple λ-cutting level strategy (for a given choice of $\lambda \in [0.5, 1]$) is used in order to transform the fuzzy outranking relation into a crisp one to determine if each alternative outranks (or not) each category. This is done by testing if $\rho(a_i, b_h) \geq \lambda$. If the inequality is satisfied, it means that indeed a_i outranks the category C_h. Based on outranking relations between all pairs of alternatives and profiles of categories, two approches are proposed in ELECTRE TRI to finally assign the alternatives into categories, see [5] for details:

   - Pessimistic (conjunctive) approach: a_i is compared with b_k, b_{k-1}, b_{k-2}, ..., until a_i outranks b_h where $h \leq k$. The alternative a_i is then assigned to the highest category C_h if $\rho(a_i, b_h) \geq \lambda$ for a given threshold λ.
   - Optimistic (disjunctive) approach: a_i is compared with b_1, b_2, ... b_h, ... until b_h outranks a_i. The alternative a_i is assigned to the lowest category C_h for which the upper profile b_h is preferred to a_i.

The objective and motivation of this paper is to develop a new Belief Function based ET method taking into account the potential of BF to model uncertainties. The whole BF-ET method is under development and will be presented and evaluated on a detailed practical example in a forthcoming publication. Due to space limitation constraints, we just present here what we propose to compute the new concordance and discordance indexes useful in our BF-ET.

2 Limitations of the Classical Electre Tri

ET method remains rather based on heuristic approach than on a theoretical one for each of its steps. Belief functions can improve ET method because of their ability to model and manage conflicting as well as uncertainty information in a theoretical framework. We only focus here on steps 1 and 2 and we propose a solution to overcome their limitations in the next section.

Example 1: Let's consider $g_j(a_i) \in [0,100]$, and let's take $g_j(b_h) = 50$ and the following thresholds: $q_j(g_j(b_h)) = 20$ (indifference threshold), $p_j(g_j(b_h)) = 25$ (preference threshold) and $v_j(g_j(b_h) = 40$ (veto threshold). Then the local concordance and discordances indexes obtained in steps 1 and 2 of ET are shown on the Fig. 1.

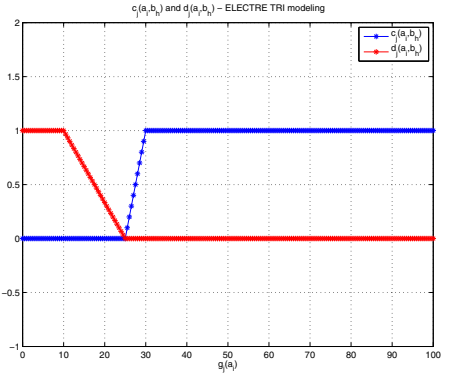


Fig. 1 Example of partial concordance and discordance indexes.

From this very simple example, one sees that ET modeling of partial concordance and discordance indexes is not very satisfactory since there is no clear (explicit and consistent) modeling of the uncertainty area where the action a_i is not totally discordant, nor totally concordant with the profile b_h. In such simplistic modeling, there exist points $g_j(a_i)$ (lying on the slope of the blue or red curves) that can be not totally concordant while being totally not discordant (and vice-versa), which is counter-intuitive and rather abnormal. This drawback will be solved using our new sigmoidal basic belief assignment (bba) modeling presented in the next section.

3 Sigmoidal Model for Concordance and Discordance Indexes

In fact, there are several ways to compute partial concordances and discordances indexes and to combine them in order to provide the global credibility indexes $\rho(a_i, b_h)$. Electre Tri proposes a simple and basic approach based on hard thresholding techniques for doing this. It can fail to work efficiently in practice in some cases, or may require a lot of experience to calibrate/tune all setting parameters in order to apply it to get pertinent results for decision-making support. Usually, a sensitivity analysis must be done very carefully before applying ET in real applications. Here,

we propose a more flexible approach based on sigmoidal modeling where no hard thresholding technique is required.

In ET approach, we are mainly concerned in the evaluation of the credibility indexes $\rho(a_i,b_h) \in [0,1]$ for $i = 1,2,\ldots,m$ and $h = 1,2,\ldots,k$ (step 3) from which the final decision (assignment) will be drawn in step 4. Step 3 is conditioned by the results of steps 1 and 2 which can be improved using belief functions. For such purpose, we consider, a binary frame of discernment[2] $\Theta \triangleq \{c,\bar{c}\}$ where c means that the alternative a_i is concordant with the assertion "a_i is at least as good as profile b_h", and $\bar{c}$ means that the alternative a_i is opposed (discordant) to this assertion. This must obviously be done with all the assertions to check in the ET framework. The basic idea is for each pair (a_i,b_h) to evaluate its bba $m_{ih}(.)$ defined on the power-set of Θ, denoted 2^Θ. Such bba's have of course to be defined from the combination (fusion) of the local bba's $m^j_{ih}(.)$ evaluated from each possible criteria $g_j(.)$ (as in steps 1 and 2). The main issue is to derive the local bba's $m^j_{ih}(.)$ defined in 2^Θ from the knowledge of the criteria $g_j(.)$ and preference, indifference and veto thresholds $p_j(g_j(b_h))$, $q_j(g_j(b_h))$ and $v_j(g_j(b_h))$ respectively. It turns out that this can be easily obtained from the new method of construction of bba presented in [4] and adapted here in the ET context as follows:

• Let $g_j(a_i)$ be the evaluation of the criterion $g_j(.)$ for the alternative a_i, following ET approach when $g_j(a_i) \geq g_j(b_h) - q_j(g_j(b_h))$ then the belief in concordance c must be high (close to one), whereas it must be low (close to zero) as soon as $g_j(a_i) < g_j(b_h) - p_j(g_j(b_h))$. Similarly, the belief in discordance $\bar{c}$ must be high (close to one) if $g_j(a_i) < g_j(b_h) - v_j(g_j(b_h))$, and it must be low (close to zero) when $g_j(a_i) \geq g_j(b_h) - p_j(g_j(b_h))$. Such behavior can be modeled directly from the sigmoid functions defined by $f_{s,t}(g) \triangleq 1/(1+e^{-s(g-t)})$ where g is the criterion magnitude of the alternative under consideration; t is the abscissa of the inflection point of the sigmoid. $s/4$ is the slope[3] of the tangent at the inflection point. It can be easily verified that the bba $m^j_{ih}(.)$ satisfying the expected behavior can be obtained by the fusion[4] of the two following simple bba's defined by: where the abscisses of inflection points are given by $t_c = g_j(b_h) - \frac{1}{2}(p_j(g_j(b_h)) + q_j(g_j(b_h)))$ and $t_{\bar{c}} = g_j(b_h) - \frac{1}{2}(p_j(g_j(b_h)) + v_j(g_j(b_h)))$ and the parameters s_c and $s_{\bar{c}}$ are given by[5] $s_c = 4/(p_j(g_j(b_h)) - q_j(g_j(b_h)))$ and $s_{\bar{c}} = 4/(v_j(g_j(b_h)) - p_j(g_j(b_h)))$.

Table 1 Construction of $m_1(.)$ and $m_2(.)$.

focal element	$m_1(.)$	$m_2(.)$
c	$f_{s_c,t_c}(g)$	0
$\bar{c}$	0	$f_{-s_{\bar{c}},t_{\bar{c}}}(g)$
$c \cup \bar{c}$	$1 - f_{s_c,t_c}(g)$	$1 - f_{-s_{\bar{c}},t_{\bar{c}}}(g)$

[2] Here we assume that Shafer's model holds, that is $c \cap \bar{c} = \emptyset$.

[3] i.e. the ratio of the vertical and horizontal distances between two points on a line; zero if the line is horizontal, undefined if it is vertical.

[4] With averaging rule, PCR5 rule, or Dempster-Shafer rule [8].

[5] The coefficient 4 appearing in s_c and $s_{\bar{c}}$ expressions comes from the fact that for a sigmoid of parameter s, the tangent at its inflection point is $s/4$.

• From the setting of threshold parameters $p_j(g_j(b_h))$, $q_j(g_j(b_h))$ and $v_j(g_j(b_h))$, it is easy to compute the parameters of the sigmoids (t_c, s_c) and $(t_{\bar{c}}, t_{\bar{c}})$, and thus to get the values of bba's $m_1(.)$ and $m_2(.)$. Once this has been done the local bba $m^j_{ih}(.)$ is computed by the fusion (denoted $\oplus$) of bba's $m_1(.)$ and $m_2(.)$, that is $m^j_{ih}(.) = [m_1 \oplus m_2](.)$. As shown in [4], the choice of a particular rule of combination (Dempster, PCR5, or hybrid rule) has only a little impact on the result of the combined bba $m^j_{ih}(.)$. But since PCR5 proposes a better management of conflicting bba's yielding to more specific results than with other rules [1], we use it to combine $m_1(.)$ with $m_2(.)$ to compute $m^j_{ih}(.)$ associated with the criterion $g_j(.)$ and the pair (a_i, b_h). In adopting such sigmoidal modeling, we get now from $m^j_{ih}(.)$ a fully consistent and elegant representation of local concordance $c_j(a_i, b_h)$ (step 1 of ET), local discordance $d_j(a_i, b_h)$ (step 2 of ET), as well as of the local uncertainty $u_j(a_i, b_h)$ by considering: $c_j(a_i, b_h) \triangleq m^j_{ih}(c) \in [0,1]$, $d_j(a_i, b_h) \triangleq m^j_{ih}(\bar{c}) \in [0,1]$ and $u_j(a_i, b_h) \triangleq m^j_{ih}(c \cup \bar{c}) \in [0,1]$. Of course, one has also $c_j(a_i, b_h) + d_j(a_i, b_h) + u_j(a_i, b_h) = 1$.

4 Example of a Sigmoidal Model

If one takes back the example 1, the inflection points of the sigmoids $f_1(g) \triangleq f_{s_c,t_c}(g)$ and $f_2(g) \triangleq f_{-s_{\bar{c}},t_{\bar{c}}}(g)$ have the following abscisses $t_c = 50 - (25+20)/2 = 27.5$ and $t_{\bar{c}} = 50 - (25+40)/2 = 17.5$ and parameters $s_c = 4/(25-20) = 4/5 = 0.8$ and $s_{\bar{c}} = 4/(40-25) = 4/15 \approx 0.2666$. The two sigmoids $f_1(g_j(a_i))$ and $f_2(g_j(a_i))$ are shown on the Fig. 2.

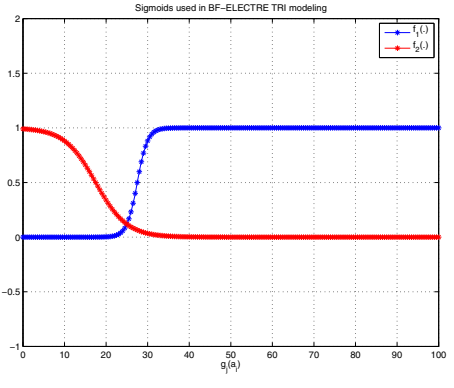


Fig. 2 $f_1(g_j(a_i))$ and $f_2(g_j(a_i))$ sigmoids.

It is interesting to note the resemblance of Fig. 2 with Fig. 1. From these sigmoids, the bba's $m_1(.)$ and $m_2(.)$ are computed according to Table 1 and shown on the Figure 3.

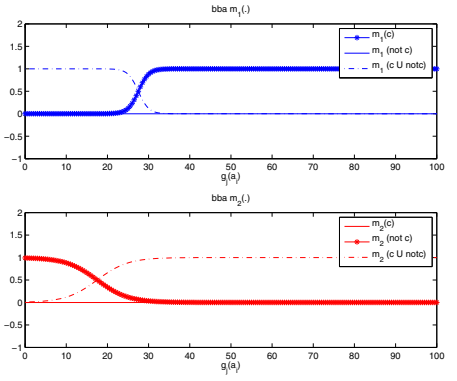


Fig. 3 Bba's $m_1(.)$ and $m_2(.)$ to combine.

The construction of the consistent bba $m^j_{ih}(.)$ is obtained by the PCR5 fusion of the bba's $m_1(.)$ and $m_2(.)$. The result is shown on Fig. 4.

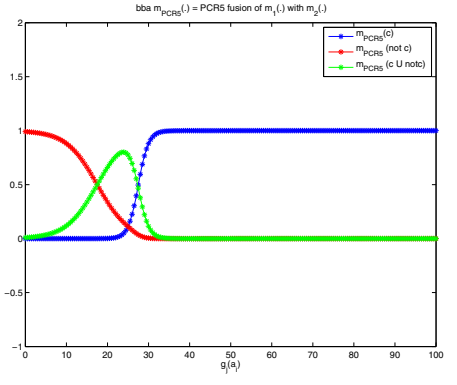


Fig. 4 $m^j_{ih}(.)$ obtained from the PCR5 fusion of $m_1(.)$ with $m_2(.)$.

From this new sigmoidal modeling, we can compute the local bba's $m^j_{ih}(.)$ derived from the knowledge of criterion $g_j(.)$ and setting parameters. This is a smooth appealing and elegant technique to build all the local bba's: no hard thresholding is necessary because of the continuity of sigmoid functions.

One can then compute the global concordance and discordance indexes of steps 1 and 2 from the computation of the combined bba $m_{ih}(.)$ resulting of the fusion of local bba's $m^j_{ih}(.)$ taking eventually into account their importance and reliability[6] (if one wants). This can be done using the recent fusion techniques proposed in [9], or by a simple weighted averaging. From $m_{ih}(.)$ we can use the same credibility index as in step 3 of ET, or just skip this third step and define a decision-making based directly on the bba $m_{ih}(.)$ using classical approaches used in belief function framework (say the max of belief, plausibility, or pignistic probability, etc).

[6] In classical ET, the reliability of criteria is not taken into account.

5 Conclusions

After a brief presentation of the classical ET method, we have proposed a new approach to model and compute the concordance and discordance indexes based on belief functions in order to overcome the limitations of steps 1 and 2 of the ET approach. The advantages of our modeling is to provide an elegant and simple way not only to compute the concordance and discordance indexes, but also the uncertainty level that may occur when information appears partially concordant and discordant. The Improvements of other steps of ET method are under development. In future reaserch works, we will evaluate and compare on real MCDA problem our BF-ET with the original ET method and with other belief functions based methods already available in MCDA frameworks [10, 11].

References

1. Dezert, J., Smarandache, F.: Proportional Conflict Redistribution Rules for Information Fusion. In: [8], vol. 2, pp. 3–68 (2006)
2. Dezert, J., Smarandache, F.: An introduction to DSmT. In: [8], vol. 3, pp. 3–73 (2009)
3. Dezert, J., Smarandache, F., Tacnet, J.-M., Batton-Hubert, M.: Multi-criteria decision making based on DSmT/AHP. In: International Workshop on Belief Functions, Brest, France (April 2010)
4. Dezert, J., Liu, Z., Mercier, G.: Edge Detection in Color Images Based on DSmT. In: Proceedings of Fusion 2011, Chicago, USA (July 2011)
5. Figueira, J., Mousseau, V., Roy, B.: ELECTRE methods. In: Multiple Criteria Decision Analysis: State of Art Surveys, ch. 4. Springer Science+Business Media Inc. (2005)
6. Mousseau, V., Slowinski, R., Zielniewicz, P.: Electre tri 2.0a - methological guide and user's manual - document no 111. In: Cahier et Documents du Lamsade, Lamsade, Université Paris-Dauphine, Paris (1999)
7. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
8. Smarandache, F., Dezert, J.: Advances and applications of DSmT for information fusion (Collected works), vol. 1-3. American Research Press (2004-2009), http://www.gallup.unm.edu/~smarandache/DSmT.htm
9. Smarandache, F., Dezert, J., Tacnet, J.-M.: Fusion of sources of evidence with different importances and reliabilities. In: Proc. of Fusion 2010 Conf., Edinburgh, UK (July 2010)
10. Tacnet, J.-M., Dezert, J.: Cautious OWA and Evidential Reasoning for Decision Making under Uncertainty. In: Proceedings of Fusion 2011, Chicago, USA (July 2011)
11. Tacnet, J.-M., Batton-Hubert, M., Dezert, J.: A two-step fusion process for multi-criteria decision applied to natural hazards in mountains. In: Proceedings of International Workshop on the Theory of Belief Functions, Belief 2010, Brest (France), April 1-2 (2010)
12. Tervonen, T., Figueira, J.R., Lahdelma, R., Dias, J.A., Salminen, P.: A stochastic method for robustness analysis in sorting problems. European Journal of Operational Research 192, 236–242 (2009)
13. Yu, W.: Aide multicritère à la décision dans le cadre de la problématique du tri: Concepts, méthodes et applications. Ph.D Thesis, University Paris-Dauphine, France (1992)

Belief Inference with Timed Evidence

Methodology and Application Using Sensors in a Smart Home

Bastien Pietropaoli, Michele Dominici, and Frédéric Weis*

Abstract. Smart Homes need to sense their environment. Augmented appliances can help doing this but sensors are also required. Then, data fusion is used to combine the gathered information. The belief functions theory is adapted for the computation of small pieces of context such as the presence of people or their posture. In our application, we can assume that a lot of sensors are immobile. Also, physical properties of Smart Homes and people can induce belief for more time than the exact moment of measures.

Thus, in this paper, we present a simple way to apply the belief functions theory to sensors and a methodology to take into account the timed evidence using the specificity of mass functions and the discounting operation. An application to presence detection in smart homes is presented as an example.

1 Introduction and Motivation

Context-aware applications have to sense the environment in order to adapt themselves and provide contextual services. This is the case of Smart Homes equipped with sensors and augmented appliances. However, sensors can be numerous, heterogeneous and unreliable. Thus the data fusion is complex and requires a solid theory

Bastien Pietropaoli · Michele Dominici
INRIA, Rennes-Bretagne Atlantique,
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
e-mail: Bastien.Pietropaoli@inria.fr, Michele.Dominici@inria.fr

Frédéric Weis
IRISA, Université de Rennes 1,
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
e-mail: Frederic.Weis@irisa.fr

* The authors would like to thank EDF that has funded this research.

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 409–416.
springerlink.com

to handle those problems. For this purpose, we adopted the belief functions theory (BFT) [7].

In our Smart Home application, we forbid the use of equipment on people [2]. Thus, we assume that a lot of sensors are immobile. Some of them should also induce belief for a certain amount of time after the measures because of the continuity of studied context. For instance, a motion sensor in a room could be able to induce a belief on the presence of someone for a longer time than the exact moment at which the measure has been obtained. It is a matter of physical system with inertia. In this example, it is easy to take into account that physical persons cannot move too fast and thus will certainly be there for some seconds before they can exit the room. Thus, this little example brings two questions: how to build evidence from raw data and how to take into account evidence over time? In this paper, we show a simple method already existing to build belief functions from raw data and propose an improvement to take into account timed evidence.

Section 2 presents the basics of the belief functions theory required to understand the proposed methodology. In section 3, the methodology used to build the belief functions from raw data given by sensors and the algorithm to take into account timed evidence and create temporization are detailed. In section 4, a very simple example of application is given. Finally in section 5 the results are discussed and our future work is presented.

2 Basics of Belief Functions Theory

In this section, we present the basics of the BFT, only considering the notions that are required to understand the methodology given in section 3.

2.1 Mass Function

In the BFT [7], the first thing that should be defined is a set of possible worlds $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ called the *frame of discernment*. These worlds have to be exclusive and if possible exhaustive. To give an example, we can define a set of the possible postures of someone by $\Omega = \{Seated, Standing, LyingDown\}$. Once the frame of discernment is created, a *mass function* (also called *basic belief assignment* or *body of evidence*) representing the degree of belief associated to each subset of Ω is defined such that:

$$\begin{aligned} m : 2^{\Omega} &\longmapsto [0,1] \\ \sum_{A \subseteq \Omega} m(A) &= 1 \end{aligned} \tag{1}$$

Every subset A with $m(A) > 0$ is called a *focal set* and may be considered as a part of belief. As the mass functions are applied to the powerset of Ω (the set of all subsets of Ω), differing from the probability theory, the beliefs may be non-specific (imprecise), i.e. accorded to a set of possible worlds. Thus, the belief functions

theory offers a double way to express the uncertainty using degrees of belief but also non-specificity.

2.2 Specificity

We saw that a mass function can be non-specific and assign belief to a set of possible worlds. Thus, there exists a tool, called *specificity* [9], to characterize mass functions and defined by:

$$S_m = \sum_{A \subseteq \Omega, A \neq \emptyset} \frac{m(A)}{|A|} \text{ where } |A| \text{ is the cardinality of } A \tag{2}$$

It can be interpreted as the degree of precision of the mass function or the inverse of the average cardinality of focal elements. For instance, if $S_m = 1$, then it means the mass function has only focal elements with a cardinality of 1 and if $S_m = 0.5$, then the average cardinality of focal elements is two. This tool will be used in section 3 as an indicator of the global precision of a mass function.

2.3 Discounting

There exist many operations on mass functions. One we found interesting to manage time is the *discounting*. It is defined by:

$$m_\alpha(A) = \begin{cases} \alpha.m(A) \text{ if } A \subseteq \Omega \text{ and } A \neq \Omega \\ \alpha.m(A) + (1-\alpha) \text{ if } A = \Omega \end{cases} \text{ with } \alpha \in [0,1] \tag{3}$$

This operation transfers a part of the mass attributed to each focal element to the set of all possible worlds (Ω) considered, in our case, as the belief given to the total ignorance. Thus, this operation always reduces the specificity of the mass function it is applied to.

The basics presented in this section are the only tools required by the methodology described in the next section.

3 Building Belief Functions

In this section, we describe a simple way to build mass functions from raw data given by sensors and also a way to add temporization to take into account timed evidence. For simplicity's sake, only linear functions are shown but any kind of function can be used to build sets of mass functions as well as temporization.

3.1 Sets of Mass Functions

To build mass functions, it is possible to use methods exploiting statistics [1] but we wanted a simple and intuitive way to build mass functions without the need for hours

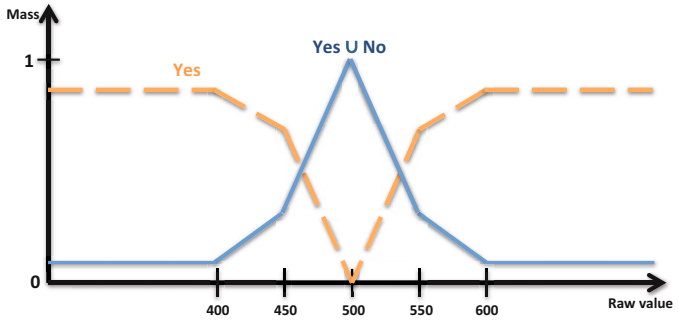


Fig. 1 Example of a set of mass functions associated to a Phidget motion sensor in the case of presence detection

of experimentation and easy to tune to adapt to any place and any situation. Thus, we chose the method used in [6]. Instead of building mass functions from previous observations, we build for each sensor a set a mass functions. Figure 1 gives an example of set of mass functions for a simple motion sensor in a case of presence detection ($\Omega = \{Yes, No\}$). The sensor used is a Phidget Motion Sensor [4]. When connected to a USB interface, it returns a measure between 0 and 1000. A measure of 500 corresponds to no motion detected and any other measure is equivalent to its symmetrical around 500. This kind of set is built on intuition and can be fine-tuned after few experiments. Once a set of mass functions is built, a projection on this set is done in order to obtain the corresponding mass function each time a raw data from that sensor is received. For instance, with the given figure 1, if the motion sensor returns a value of 450, then the resulting mass function would have two focal elements: $m(\{Yes\}) = 0.7$ and $m(\{Yes \cup No\}) = 0.3$.

A constraint to respect when building these sets of mass functions is the *least commitment principle*. In our case it can be translated by the fact that the belief induced by a sensor measure should not be too specific when it cannot be. In the given example, the motion is only a proof that somebody may be there but the gathered measure can never be a good proof that nobody is there. That is why the set $\{No\}$ never appears in the set of mass functions in figure 1.

3.2 Temporization

In physical systems such as Smart Homes, some properties can make evidence stay valid over time. In the case of presence detection, a motion sensor, if well placed, is certainly a timed evidence as a person moving in a room cannot leave it before seconds. This assumption is valid for any state that can be perceived as continuous in time. Thus, it should be possible to infer belief for a certain amount of time after the measure has been obtained. The belief should become less and less committed over time if no new clear evidence is brought. As far as we now, no tool in the BFT handles this problem directly.

In order to represent the weakening of a belief over time, we use the discounting operation (3). The specificity (2) will also help characterizing the mass function resulting of each measure and to give priority to precision instead of certainty. Thus, an old and discounted mass function m_1 will be preferred to a fresh mass function m_2 if $S_{m_1} > S_{m_2}$. As a consequence, instead of giving just the projection induced by the last measure, the specificity of both mass functions, the old one and the new one, are compared and the most specific is returned. Algorithm 1 shows the pseudo-code corresponding to this procedure.

We decided to give priority to precision instead of certainty because in the decision making process, when using classical criterion such as *credibility*, *plausibility* or the *bet on the probability* [8], it is easier to degrade the precision to gain in certainty. As a matter of fact, it is possible to increase the cardinality of the set of possible worlds chosen as the system response to gain certainty.

Algorithm 1. Take time into account

```
newMassFunction = getProjection(lastRawData)
alpha = getDiscountingFactor(oldTime, newTime)
discountedMassFunction = discounting(oldMassFunction, alpha)
if S_discountedMassFunction > S_newMassFunction then
    return discountedMassFunction
else
    oldMassFunction = newMassFunction
    oldTime = newTime
    return newMassFunction
end if
```

The function computing the discounting factor can be any function depending on time. In the example that will be given in section 4, a simple linear function has been used. This very simple algorithm can be applied easily but does require a bit of computation as discounting and specificity are in $O(2^n)$ where $n = |\Omega|$. However, n stays acceptable for the computation of small pieces of context such as the presence, the posture of someone, etc, as it is the case in our application [2, 5]. In practice, the computation of discounting and specificity is also reduced to the number of focal sets ($A \subset \Omega$ with $m(A) > 0$) of a belief function which can be smaller than the set of possible subsets of Ω. Some performance results of our implementation of the belief functions theory are presented in [5].

4 Application

In this section, we focus on the importance of taking into account the timed evidence induced by the physical properties of systems. Then, we illustrate the use of temporization in a simple example of presence detection in a room.

4.1 Presence Detection

In Smart Homes, one of the main problems is the detection of people anywhere and anytime. Sometimes, indoor geo-localization is not needed and only an indication that someone is in a room or in another is enough. To do this, simple systems often use only a motion sensor considered as a sufficient proof. Using data fusion and especially belief functions theory, it is possible to combine multiple sensors to obtain a finer result with poor evidence. The fact that someone cannot exit a room in the second after he or she has been detected can indirectly bring interesting evidence. That is why we apply temporization in the detection of presence.

In order to detect presence in a room, we use in this application three simple sensors [4]: a motion sensor, a vibration sensor on a chair and a sound sensor. The evidence gathered from the sound sensor is taken from the variation of sound level and not directly from the measure. None of these sensors can bring a good proof that nobody is in the room. Thus, the system response should always be $\{Yes \cup No\}$ or $\{Yes\}$. Figure 1 shows the model used for the motion sensor in this application.

In the presence detection described here, we used certain types of sensors but some others such as CO_2 level sensors, microphones and so on can also be used and the same methodology can be applied. The application presented here is thus just a simple example the proposed methodology is not limited to.

4.2 Application of Temporization

Figure 2 shows the results of an experimentation where both results have been computed in parallel from the same raw data. The combination rule used is the *normalized Dempster's rule of combination* but any combination could be used. As a matter of fact, this way of building mass functions does not prevent from using any combination rule.

Figure 2 is eloquent. The resulting mass functions are clearly stabilized when applying temporization on sensors. Without temporization, the system can doubt sometimes because the sensors used in our experiment are not sufficient proofs, especially if there is someone quiet and not moving at all. The first consequence of temporization is thus stabilization of the system response.

Unfortunately, the effect of noisy measures is also stabilized. In the top part of figure 2, when nobody is in the room, the mass assigned to the $\{Yes\}$ is greater than in the case where no temporization has been used. Thus, a less committed model should be used to be less sensitive to noisy measures. The use of temporization enables the use of less committed models because asynchronous proofs can be fused altogether anyway, a simple discounting factor reducing the strength of each proof. As a consequence, temporization brings more confidence in the end as multiple sensors may not be activated at the same time but, considered as timed evidence, it does not matter as long as they induce belief in the same time interval.

Temporization also enables the reduction of required measures to get an image of what is happening at a given time. The absence of measure gives a vacuous mass function ($m(\Omega) = 1$) which is always less specific than a discounted mass function.

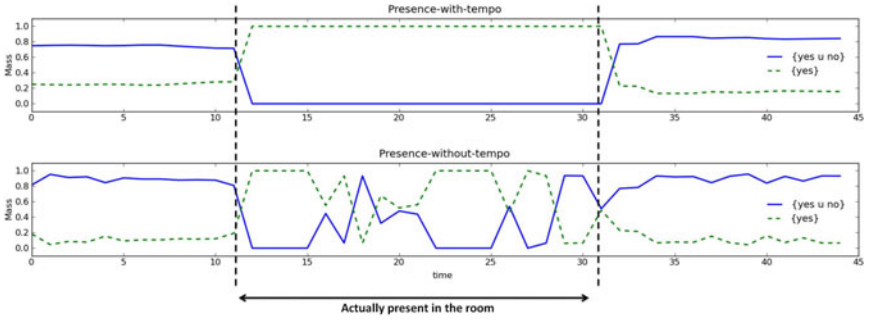


Fig. 2 Resulting mass functions from the detection of presence in a room with a motion sensor, a vibration sensor on a chair and a sound sensor. On top, the temporization has been used. In the bottom part, no temporization has been used. Both results come from the same measures and from the same sets of mass functions.

It gives the possibility to still have belief when there is no new measure. Thus, if a sensor measure induces belief for a long time, its frequency can be drastically reduced. This can also be useful in systems using wireless sensor nodes characterized by the possibility of measure loss during communication.

Even though, the measure frequency should not be reduced too much or the system may suffer severe problems of reactivity. A compromise between frequency and reactivity is needed.

5 Discussion and Future Work

In this paper, we presented a simple way to build mass functions when physical systems can induce belief over time. The complexity of the algorithm is acceptable because only fast operations are used to take into account the time. The results presented in section 4 show that the reactivity is slightly reduced. However, if the chosen model respects the least commitment principle, the system response is more stable. The algorithm also enables the use of poor evidence without degrading too much the response of the system. The use of less committed evidence is also a very good point when the measures are noisy.

The application presented in this paper is simple. Other experiments with more complex cases have been tested and seem to work in the same way. Anyway, the temporization should only be applied when the physical continuity of systems is easy to assume. Also, the least commitment principle is key to prevent from stabilizing too much belief induced by noisy measures and should be strictly respected. More work is also required to generalize this method when the conflict between the sources is high.

The methodology introduced enables taking into account that a belief can stay partly valid for a certain amount of time after the measure it comes from. Another interesting studied thing in the BFT is the time as evidence by itself [3]. For example,

in the application presented in section 4, the fact that no sensor has detected even a low activity for a very long time can be a good proof that nobody is in the room. This kind of evidence, time as evidence, requires more work in our application.

Another big question when using the belief functions theory is decision making. Promising work on result filters suggests that the temporization can help creating a natural way to doubt on what is going on. Still with the same example of presence, a natural way to doubt could be instead of directly saying $\{Yes\}$ or $\{No\}$, the system could transit from one state to the other with a $\{Yes \cup No\}$ for some seconds.

References

1. Aregui, A., Denoeux, T.: Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. International Journal of Approximate Reasoning 49, 575–594 (2008), doi:10.1016/j.ijar.2008.06.002
2. Dominici, M., Fréjus, M., Guibourdenche, J., Pietropaoli, B., Weis, F.: Towards a system architecture for recognizing domestic activity by leveraging a naturalistic human activity model. In: Workshop on Goal, Activity and Plan Recognition at the International Conference on Automated Planning and Scheduling, ICAPS 2011 (2011)
3. Mckeever, S., Ye, J., Coyle, L., Bleakley, C., Dobson, S.: Activity recognition using temporal evidence theory. J. Ambient Intell. Smart Environ. 2, 253–269 (2010)
4. Phidgets, http://www.phidgets.com/
5. Pietropaoli, B., Dominici, M., Weis, F.: Multi-sensor Data Fusion within the Belief Functions Framework. In: Balandin, S., Koucheryavy, Y., Hu, H. (eds.) NEW2AN 2011 and ruSMART 2011. LNCS, vol. 6869, pp. 123–134. Springer, Heidelberg (2011)
6. Ricquebourg, V., Delafosse, M., Delahoche, L., Marhic, B., Jolly-Desodt, A., Menga, D.: Fault Detection by Combining Redundant Sensors: a Conflict Approach Within the TBM Framework. In: Cognitive Systems with Interactive Sensors, COGIS 2007, Stanford University (2007)
7. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
8. Smets, P.: Decision making in the tbm: the necessity of the pignistic transformation. Int. J. Approx. Reasoning 38(2), 133–147 (2005)
9. Yager, R.: Entropy and Specificity in a Mathematical Theory of Evidence. In: Yager, R., Liu, L. (eds.) Classic Works of the Dempster-Shafer Theory of Belief Functions. STUDFUZZ, vol. 219, pp. 291–310. Springer, Heidelberg (2008)

Evidential Network with Conditional Belief Functions for an Adaptive Training in Informed Virtual Environment

Loïc Fricoteaux, Indira Thouvenin, Jérôme Olive, and Paul George

Abstract. Simulators have been used for many years to learn driving, piloting, steering, etc. but they often provide the same training for each learner, no matter his/her performance. In this paper, we present the GULLIVER system, which determines the most appropriate aids to display for learner guiding in a fluvial-navigation training simulator. GULLIVER is a decision-making system based on an evidential network with conditional belief functions. This evidential network allows graphically representing inference rules on uncertain data coming from learner observation. Several sensors and a predictive model are used to collect these data about learner performance. Then the evidential network is used to infer in real time the best guiding to display to learner in informed virtual environment.

1 Introduction

Virtual reality can provide, in comparison with classical training, many advantages [1]. In the case of fluvial navigation, training in virtual environment allows to simply modify environmental conditions (wind, current, etc.), which has an impact on the behavior of the ship. Another advantage of training in virtual reality is the strong coupling between the user and the virtual environment. The virtual world must credibly answers to user's actions. We use an informed virtual environment (IVE: environment including knowledge-based models and providing an action/perception coupling) for fluvial navigation training. The purpose of our work is to provide the best learner guiding (set of aids) in real time based on learner observation. We propose an adaptive system: the learner's behavior is taken into account for the choice of the aids to display [3]. On the opposite side, non-adaptive systems [4] are easier to build but the aids will not be adapted to the learner's performance. For example, novice learners will not have enough help and experienced ones will have

Loïc Fricoteaux · Indira Thouvenin · Jérôme Olive · Paul George
Heudiasyc Laboratory UMR CNRS 6599, Compiégne, France
e-mail: firstname.lastname@utc.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 417–424.
springerlink.com

too much help. As fluvial navigation is a complex task, there is no complete procedure to follow to know how to navigate (there is only a navigation code). With a procedural approach, errors can be easily detected by comparing learner's actions with good actions to perform [3]. With a non-procedural approach [6], the system is more complex to build but is adapted to the training of complex tasks. Thus, our system is based on this approach. Errors are mainly detected according to a predictive model (the future position of the ship), therefore this detection is uncertain and this has to be taken into account by the decision-making module in the choice of the best guiding. We also use physiological sensors to detect learner's state (for example the stress level with a heart rate variability sensor), which gives uncertain data about the user's state due to sensor reliability and uncertainty of data interpretation.

All data coming from learner observation has to be expressed in a common formal framework to allow making decision. We use the Dempster-Shafer (DS) theory [9] to take the uncertainty of these data into account. Comparing to the theory of probability, the DS theory allows modeling ignorance explicitly, which is useful in our case since we can have incomplete data about learner's actual situation. To represent influences between variables (i.e. variables about learner's errors and possible feedbacks to avoid these errors) and to reason on these variables, directed graphs are widely used. In the case of probabilistic inference, Bayesian networks (BN) are used [7]. With belief functions, the equivalent network is called an Evidential Network with Conditional belief functions (ENC) [11, 15]. ENC have been generalized by DEVN (Directed EVidential Network with conditional belief functions) to have n-ary relations between variables (ENC are limited to binary relations) [2]. In our case, relations between variables are only binary because it is easier to specify [11] and to update, so we use an ENC in our system. Also to be more intuitive, we represent knowledge by using conditional belief functions unlike joint belief functions as in valuation networks [10]. Contrary to ENC, BN need experimental data to be initialized (to compute conditional probabilities) because they are not intuitive to specify. Indeed, if A and B influences C, you have to specify the probability of C conditionally to A and B, whereas in an ENC you have to specify the belief of C conditionally to A and the belief of C conditionally to B (and then apply a combination rule to fusion these two results). When the number of variables increases, conditional probabilities in BN cannot be simply specified or updated by hand.

The paper is organized as follows. In section 2, some useful formulas and notations about DS theory are presented. Section 3 describes our GULLIVER system of adaptive training in IVE based on inferences in ENC.

2 Decision-Making with Conditional Belief Functions

In this section, some useful definitions and notations in the Transferable Belief Model (TBM) [12] are briefly presented.

Definition 1. Let Ω be a finite set called the frame of discernment. Ω is the domain relative to the variable X. A basic belief assignment (bba) $m^{\Omega} : 2^{\Omega} \rightarrow [0;1]$ is a set

of belief masses such that $\sum_{A\subseteq\Omega} m^{\Omega}(A) = 1$. The belief mass $m_S^{\Omega}(A)$ represents the belief of the source S in the fact "$\omega \in A$", where ω is the real state of the system observed [8].

Data fusion is used to enhance-decision-making. To combine heterogeneous data coming from several sources and representing by bba, combination rules are used. Two of them are presented in the next two definitions.

Definition 2. Let two distinct bba m_{S1}^{Ω} and m_{S2}^{Ω} defined on the same frame of discernment Ω. The sources $S1$ and $S2$ are supposed reliable and distinct [14]. The TBM conjunctive rule of combination (CRC) of m_{S1}^{Ω} and m_{S2}^{Ω} is defined as follows:

$$\forall A \subseteq \Omega,\ m_{S1\textcircled{\cap}S2}^{\Omega}(A) = \left(m_{S1}^{\Omega}\textcircled{\cap}m_{S2}^{\Omega}\right)(A) = \sum_{B\cap C=A} m_{S1}^{\Omega}(B)\, m_{S2}^{\Omega}(C) \tag{1}$$

If these belief sources are distinct and at least one of them is reliable (without being able to quantify the reliability and knowing which source is reliable), the disjunctive rule of combination (DRC) must be used [11].

Definition 3. Let two distinct bba m_{S1}^{Ω} and m_{S2}^{Ω} defined on the same frame of discernment Ω. The sources $S1$ and $S2$ are supposed distinct and one of them reliable. The disjunctive rule of combination (DRC) of m_{S1}^{Ω} and m_{S2}^{Ω} is defined as follows:

$$\forall A \subseteq \Omega,\ m_{S1\textcircled{\cup}S2}^{\Omega}(A) = \left(m_{S1}^{\Omega}\textcircled{\cup}m_{S2}^{\Omega}\right)(A) = \sum_{B\cup C=A} m_{S1}^{\Omega}(B)\, m_{S2}^{\Omega}(C) \tag{2}$$

Applying the DRC on bba, when belief sources are not reliable, produces a less informative bba [8]. When it is possible to quantify the reliability of a source S, the CRC can be used after applying a discounting [5] on the bba coming from S.

Definition 4. The Shafer discounting of a bba m_S^{Ω} coming from a source S which has a reliability of $1-\alpha$ is defined as follows [9]:

$$\begin{cases} {}^{\alpha}m_S^{\Omega}(A) = (1-\alpha)\, m_S^{\Omega}(A),\ \forall A \subset \Omega \\ {}^{\alpha}m_S^{\Omega}(\Omega) = (1-\alpha)\, m_S^{\Omega}(\Omega) + \alpha \end{cases} \tag{3}$$

To represent knowledge about influences between variables, conditional beliefs are used:

Definition 5. Let m^{Ω} be a bba about the frame of discernment Ω. The conditional bba given $B \subseteq \Omega$ is defined by the following unnormalized rule of conditioning [11]:

$$m^{\Omega}(A|B) = \begin{cases} \sum_{\{X\subseteq\overline{B}\}} m^{\Omega}(A\cup X), \text{ if } A \subseteq B \subseteq \Omega \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

The next theorem is a property of the generalized Bayesian theorem [11] to compute the belief of A with the knowledge of A given B and an a priori on B.

Theorem 1. *Suppose there exists some a priori belief m_0^{Θ} distinct from the belief induced by the conditional bba $m^{\Omega}(.|\theta)$, $\theta \subseteq \Theta$, then [11]:*

$$m^{\Omega}(\omega) = \sum_{\theta \subseteq \Theta} m_0^{\Theta}(\theta) m^{\Omega}(\omega|\theta), \ \forall \omega \subseteq \Omega \tag{5}$$

As it is easier to determine belief given singletons instead of belief given sets [11], the following formula, applying DRC on a conditional bba, can be used in (5):

$$m^{\Omega}(\omega|\theta) = \bigcirc\!\!\!\!\!\!\bigcup_{\theta_i \in \Theta} m^{\Omega}(\omega|\theta_i), \ \forall \omega \subseteq \Omega, \forall \theta \subseteq \Theta \tag{6}$$

In the TBM, decisions are made with the pignistic probabilities [13] of a bba.

Definition 6. The pignistic probability function $BetP\{m^{\Omega}\}(\omega)$ on Ω of the bba m^{Ω} is defined as follows [13]:

$$BetP\left\{m^{\Omega}\right\}(\omega) = \sum_{A \subseteq \Omega, \ \omega \in \mathrm{A}} \frac{m^{\Omega}(A)}{|A|\left(1 - m^{\Omega}(\emptyset)\right)}, \ \forall \mathrm{A} \subseteq \Omega \tag{7}$$

The decision is generally made by choosing the element ω with the highest pignistic probability [8].

To graphically represent knowledge about influences between variables, an ENC (Fig. 1) can be used. It is a directed acyclic graph where [11, 15, 2]:

- Each variable X has a set of possible values in Ω_X and a bba associated m^{Ω_X}. A variable is represented by a circle (or an oval).
- Each root node represents an a priori bba $m_0^{\Omega_X}$ on its child node X and is represented by a rectangle.
- Each edge between two variables X and Y has a diamond representing the conditional bba $m^{\Omega_X}(X|Y)$ on the child X given its parent Y.

3 Inference in ENC: The GULLIVER System for an Adaptive Training in Informed Virtual Environment

The system GULLIVER (GUiding by visuaLization metaphors for fluviaL navigation training in Informed Virtual EnviRonment) is a decision-making system which interprets data coming from user observation to infer the best guiding to display.

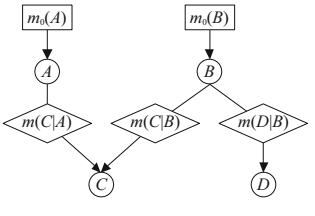


Fig. 1 Example of ENC

Knowledge is represented in an ENC which is also used to propagate belief for decision-making.

The global system operates as follows. The fluvial navigation simulator SimNav computes the position, the direction and the speed of the barge controlled by the learner thanks to the boat controls associated (Fig. 2). From these data, the position of the barge is updated in the IVE (Informed Virtual Environment).

Actions (boat movements) and events (collisions, etc.) are transmitted by the IVE to the user's activity detection module. Information about the learner's gestures is also transmitted to this module which is in charge of detecting the mistakes made by the learner. The learner's state (stress level, cognitive load, etc.) is also recognized thanks to data coming from physiological sensors.

From the learner's state and mistakes, the decision-making module, based on an ENC, activates the right aids to guide the learner. This module can also decide to trigger events. For example, if the learner does not make mistakes and feels at ease, the environment will be complexified by adding some dangers, for instance floating objects to avoid or a thick fog.

In addition to the learner's state and mistakes, the system takes also the learner's profile into account: his/her usage history (errors made before, inefficient aids, etc.) and his/her level (novice, experienced, etc.). If the learner is a novice, the guiding system must adapt to a cognitive speed compatible with the learner's perception and comprehension speed to avoid a cognitive overload.

3.1 A Priori Beliefs Deduced from User Observation

Learner's observation brings a priori beliefs on variables representing learner's errors and state. Fig. 3 presents an extract of the ENC used in GULLIVER system.

A heart rate sensor is used to determine the learner's stress level. In function of the heart rate (in beats per minute), an a priori bba is computed for the variable *stress*, which has a set of possible values in $\Omega_{stress} = \{yes; no\}$, as shown in Fig. 4.

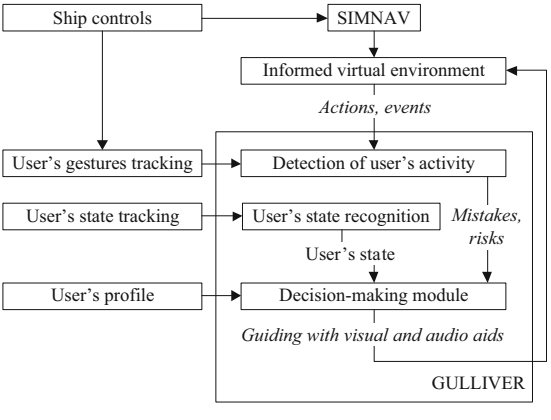


Fig. 2 Model of an adaptive training system: GULLIVER

As the heart rate sensor has a reliability of 80%, a discounting is apply on the a priori bba. For example, if the heart rate is 105 bpm, then the a priori bba is:

$$\begin{cases} m_0^{\Omega_{stress}}(yes) = 0.5 \\ m_0^{\Omega_{stress}}(no) = 0 \\ m_0^{\Omega_{stress}}(\Omega_{stress}) = 0.5 \end{cases}$$

By using the formula (3) for the discounting, we obtain:

$$\begin{cases} {}^{0.2}m_0^{\Omega_{stress}}(yes) = 0.4 \\ {}^{0.2}m_0^{\Omega_{stress}}(no) = 0 \\ {}^{0.2}m_0^{\Omega_{stress}}(\Omega_{stress}) = 0.6 \end{cases}$$

Similarly as for *stress*, we can compute an a priori bba for the variable *bridge collision* in function of the future position of the boat. The more the collision is in a near future, the more certain the collision is. In this case there is no discounting because the physical engine can reliably compute the future position of the boat.

3.2 Information Propagation in ENC

After computing the a priori beliefs, they must be propagated in the ENC (Fig. 3) to determine the useful aids for the current learner's situation. A link between two variables represents an influence of the parent variable on its child. These influences are represented by conditional bba, which are the translation of simple rules. The simplicity is a choice made so that the system can be modified by navigation trainers.

For example, we have the rules "if the learner is under stress, then the event "hiding of the next bridge" is useful at 50%" and "if the learner is not under stress, then the event "hiding of the next bridge" is useless at 50%". The two "50%" represent the degree of relevance of the rules. The variable *hiding next bridge* has a set of possible values in $\Omega_{hiding\ next\ bridge} = \{useful; useless\}$. The previous rules can be translated by the left conditional belief table in Table 1.

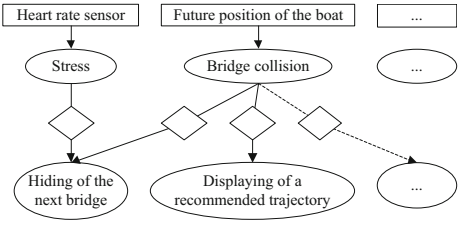


Fig. 3 Extract of the ENC used in GULLIVER system

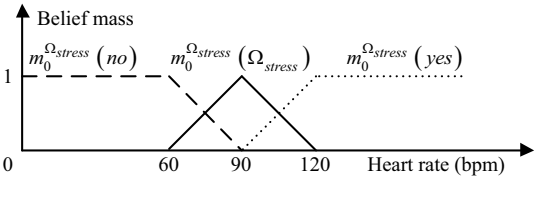


Fig. 4 A priori bba of *stress* in function of hear rate

For the aid "displaying of a recommended trajectory" (Fig. 3), we have the rule "if the boat is about to collide with a bridge, "displaying of a recommended trajectory" is useful". If the boat is not about to collide with a bridge, we cannot conclude anything on the usefulness of displaying a recommended trajectory. This is translated by the right conditional belief table in Table 1.

Table 1 Conditional belief tables

	Yes	No	Ω_{stress}
Useful	0.5	0	0
Useless	0	0.5	0
$\Omega_{hiding\ next\ bridge}$	0.5	0.5	1

	Yes	No	$\Omega_{bridge\ collision}$
Useful	1	0	0
Useless	0	0	0
$\Omega_{disp.\ recommended\ trajectory}$	0	1	1

With the conditional belief tables and the a priori bba, the belief can be propagated to the next level of the ENC by using the formula (5). In the case of *hiding next bridge*, there are two parent variables which bring beliefs (Fig. 3), so the resulting beliefs must be combined by using the CRC (formula (1)).

In order to classify the aids by priority order for decision-making, the pignistic probability of the usefulness of each aid/event (the terminal nodes in Fig. 3) is computed (formula (7)). The aids/events with the highest pignistic probability are very likely to be displayed, but a final filtering is necessary. Indeed, the aids/events composing the guiding must be moderated in order to respect some constraints. For example, the set of aids/events to trigger must not overload the screen, they must be mutually compatible (for example it is not possible to trigger at the same time "hiding of the next bridge" and "highlighting of the next bridge traffic sign"), they must be adapted to the learner's level, they must have a high pignistic probability of usefulness, etc. Some constraints must be obligatory respected (for example they must have a high pignistic probability of usefulness) and others can be not respected (but the best solution must respect as many of them as possible). This is a constraint satisfaction problem and it is solved by enumerating every possibility or by using a genetic algorithm if the combinatorics is too important. Indeed, a genetic algorithm allows computing a good solution within a time chosen, which is a very short time in our case since the guiding must be updated in real time.

Conclusion

We propose the GULLIVER system, an adaptive training system in informed virtual environment which infers a learner guiding thanks to an ENC (evidential network with conditional belief functions). User observation (errors made, stress level, etc.) is translated into a priori beliefs which are then propagated in the ENC for decision-making about the best guiding to display to the learner. The main advantage of our system is that it can be easily updated by non-expert of the system thanks to the use of conditional beliefs translated from simple rules. The system can be intuitively

initialized by hand and does not need experimental data as for Bayesian networks by example. Another advantage is the genericity of our system which can be applied, for example, in car driving assistance in augmented reality.

As perspective, we plan to enhance the system so that it will be auto-adaptive: the system will provide auto-regulation, throughout its use, for the rule relevancies (translated in conditional beliefs) and for the constraint satisfaction system. Another perspective is to reuse this system for other applications like car, train or plane driving assistance with augmented reality visualization.

Acknowledgements. This work has been funded by the European Union and Picardie region.

References

1. Amokrane, K., Lourdeaux, D., Burkhardt, J.M.: HERA: Learner Tracking in a Virtual Environment. International Journal of Virtual Reality 7(3), 23–30 (2008)
2. Ben Yaghlane, B., Mellouli, K.: Inference in directed evidential networks based on the transferable belief model. International Journal of Approximate Reasoning 48(2), 399–418 (2008)
3. Buche, C., Bossard, C., Querrec, R., Chevaillier, P.: PEGASE: A Generic and Adaptable Intelligent System for Virtual Reality Learning Environments. International Journal of Virtual Reality 9(2), 73–85 (2010)
4. Flipo, A.: TRUST: the Truck Simulator for Training. In: Driving Simulation Conference, Paris, France, 208–220 (2000)
5. Mercier, D., Quost, B., Denoeux, T.: Refined modeling of sensor reliability in the belief function framework using contextual discounting. Information Fusion 9(2), 246–258 (2008)
6. Mufti-Alchawafa, D.: Modélisation et représentation de la connaissance pour la conception d'un systéme décisionnel dans un environnement informatique d'apprentissage en chirurgie. PhD thesis, Université Joseph Fourier - Grenoble I (2008)
7. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
8. Ramasso, E., Rombaut, M., Pellerin, D.: Modéle des Croyances Transférables: Représentation des connaissances, Fusion d'informations, Décision. Technical report (2007)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
10. Shenoy, P.P.: Using Dempster-Shafer's belief-function theory in expert systems. In: Advances in the Dempster-Shafer theory of evidence, pp. 395–414. John Wiley & Sons (1994)
11. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1), 1–35 (1993)
12. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66(2) (1994)
13. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. International Journal of Approximate Reasoning 38(2), 133–147 (2005)
14. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion 8(4), 387–412 (2007)
15. Xu, H., Smets, P.: Evidential Reasoning with Conditional Belief Functions. In: Uncertainty in Artificial Intelligence, San Francisco, CA, pp. 598–605 (1994)

Using the Belief Functions Theory to Deploy Static Wireless Sensor Networks

Mustapha Reda Senouci, Abdelhamid Mellouk,
Latifa Oukhellou, and Amar Aissani

Abstract. The location of sensors is one of the fundamental design issues in wireless sensor networks. It may affect the fulfillment of the system's requirements and multiple network performance metrics. Assuming that an inherent uncertainty can be associated with sensor readings, it is very important to consider this issue in the deployment process to anticipate this sensing behavior. This paper addresses the issue of uncertainty-aware sensor networks deployment by exploiting the belief functions reasoning framework. An evidence-based coverage model is proposed and some possible extensions are discussed. The deployment problem is formulated as an optimization problem and possible solutions are discussed. Preliminary experimental analysis demonstrates very promising results of the proposed methodology.

1 Introduction

The rapid development in wireless communications and electronics have enabled the development of small-scale, low-power, low-cost sensor nodes (or sensors) that integrate processing, storage, sensing and communication capabilities. These tiny sensors leverage the idea of wireless sensor networks.

Mustapha Reda Senouci
A.I Laboratory, Ecole Militaire Polytechnique, Algiers, Algeria
e-mail: mrsenouci@gmail.com

Abdelhamid Mellouk
LiSSi Laboratory, UPEC, Paris, France
e-mail: mellouk@u-pec.fr

Latifa Oukhellou
UPE, IFSTTAR, GRETTIA, Marne-la-vallée, France
e-mail: latifa.oukhellou@ifsttar.fr

Amar Aissani
LRIA Laboratory, USTHB, Algiers, Algeria
e-mail: amraissani@yahoo.fr

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 425–432.
springerlink.com

A Wireless Sensor Network (WSN) consists of a spatially distributed sensors and one or more *sinks*. Sensors monitor environmental conditions, such as temperature, vibration, or motion and produce sensory data. A sink, on the other hand, collects data from sensors.

One of the fundamental design issues in WSNs is where to place the sensors in the Region of Interest (RoI). The location of a sensor may affect the fulfillment of the system's requirements and multiple network performance metrics. A problem which impinges upon the success of any deterministic WSN deployment is the fact that there is an inherent uncertainty associated with sensor readings. Indeed, sensors may not always provide reliable information, either due to hardware configuration or environmental conditions. As an example, for omnidirectional acoustic sensors or ultrasonic sensors, a longer distance between the sensor and the target generally implies a greater loss in the signal strength or a lower signal-to-noise ratio [1]. Therefore, it is very important to take into account this issue in the deployment process to anticipate this sensing behavior.

This paper addresses the uncertainty-aware sensor networks deployment problem (USDP) by exploiting the belief functions theory. The paper starts with a general discussion of related work. In Section 3, we present our evidence-based coverage model and some possible extensions. Section 4 formulates the USDP as an optimization problem and discuss some possible solutions. In Section 5, we present experimental results that show the effects of different parameters on the performance. Section 6 concludes the paper and discusses some future directions for our work.

2 Related Work

Usually, the deterministic deployment of static WSNs involves two components: (i) a *sensor coverage model* and (ii) a *placement algorithm*. A sensor coverage model is an abstraction model trying to quantify how well sensors can sense physical phenomena at some locations in the RoI [1]. On the other hand, a placement algorithm determines the minimum number of sensors and their locations to achieve the desired design goals. The sensors locations are computed based on a sensor coverage model. This section reviews the related work of sensor coverage models and placement algorithms.

The most widely used sensor coverage model in the literature is the *binary coverage model* [1, 2, 3], which assumes that an event happening within the sensing radius of a node is always detected, while any event outside this disk is assumed not to be detected. While this is appropriate for defining the foundation of research, the binary coverage model is overly simplistic and does not reflect reality.

Some researchers argue that the sensing quality of a sensor will decay with the increase of the distance away from the sensor, environmental conditions, hardware configurations, and other problem specific attributes [1, 4, 5]. *Probabilistic coverage models* [1, 2, 4, 5, 6] are used to capture such attenuated sensing qualities. Although these models capture the behavior of sensors more realistically, they remain limited. For example, it is not clear how to handle the reliability of sensors in the design stage

when using a probabilistic coverage model. The objective of our research described in this paper is to extend the probabilistic coverage model to an evidence-based coverage model that can provide insight into USDP.

When considering a binary coverage model, the sensor deployment problem can be formulated as the famous *art gallery problem* (AGP) addressed by the art gallery theorem [7]. Authors in [3] try to minimize costs subjects to covering all target points. Some other formulations include the *minimal disk covering problem* [8].

There has been relatively little research performed in the area of USDP. In [5], Zou et al. consider the uncertainty in sensor locations subsequent to airdropping. In literature, when sensor detection is modeled probabilistically, the sensor deployment problem is formulated as an optimization problem which is NP-complete. Therefore, proposed solutions are mainly heuristics [4, 5, 9, 10].

In this paper, we propose a more realistic approach for USDP based on the transferable belief model (TBM) [11]. Our model is generic and flexible and can be extended in many ways.

3 Our Proposal

The TBM manipulates belief functions. Thus, we need to translate the sensory data into belief functions. In this section, we define an evidence-based coverage model. In what follows, we assume the reader to be familiar with the TBM [11].

3.1 Evidence Construction

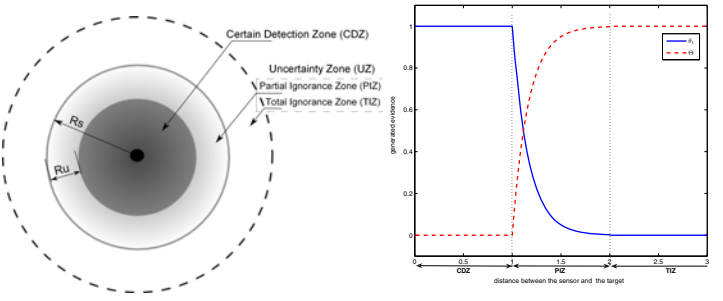
Only two states are required to specify whether a space point $p \in RoI$ is covered: $\theta_0(not\ covered)$ and $\theta_1(covered)$. Thus, the Frame of Discernment (FoD) is the set $\Theta = \{\theta_0, \theta_1\}$.

Let s be a sensor, R_s be its sensing range and R_u be a distance ($0 \leq R_u \leq R_s$) as illustrated in Fig. 1a. Each sensor s provides information on the coverage of a space point $p \in RoI$ with a belief $x_{s/p}$. The complementary information $1 - x_{s/p}$ is assigned to the whole FoD because it encodes the sensor ignorance. The output from the sensor s about a space point $p \in RoI$ can thus be represented as a basic belief assignment (*bba*) $m_{s/p}$ with two focal sets: the singleton $\{\theta_1\}$ and the FoD Θ.

Definition 1. The Certain Detection Zone (CDZ) of s is a disk centered at s with the radius of $(R_s - R_u)$. Within its CDZ, s produces a categorical belief function

Definition 2. The Uncertainty Zone (UZ) of s is the complement of its CDZ relative to the RoI; $UZ = \overline{CDZ}_{RoI}$. UZ is divided into two sub-zones: the Partial Ignorance Zone (PIZ) and the Total Ignorance Zone (TIZ).

Definition 3. The PIZ is defined as an annulus centered at s with an inner radius of $R_s - R_u$ and an outer radius of R_s. Within its PIZ, s produces a simple support function defined by equation (1).



(a) Illustration of the different zones.

(b) Illustration of the relation between the generated evidence and sensor-target distance.

Fig. 1 Evidence-based sensor coverage model.

$$\begin{aligned} &m_{s/p}(\{\theta_1\}) = x_{s/p},\ x_{s/p} \in [0,1] \\ &m_{s/p}(\Theta) = 1 - x_{s/p},\ \ m_{s/p}(\emptyset) = 0 \end{aligned} \tag{1}$$

Definition 4. The TIZ is defined as the complement of the PIZ relative to $(RoI - CDZ)$; $TIZ = \overline{PIZ_{RoI-CDZ}}$. Within its TIZ, s produces a vacuous belief function.

The parameter $1 - x_{s/p}$ reflects the sensor's degree of ignorance. One can assume, as in the truncated attenuated disk model, that $x_{s/p} = e^{-\delta(d(s,p)-(R_s-R_u))^\beta}$ where $d(s,p)$ is the Euclidean distance between a sensor s and a space point p, δ is a sensor technology related parameter and β is an event characteristic-dependent parameter. This model reflects the behavior of range sensing devices such as infrared and ultrasound sensors [1]. An example of this model is depicted in Fig. 1b.

3.2 Evidence Combination

For N sensors, the combination of the N *bbas* $m_{1/p}, \ldots, m_{N/p}$ using the unnormalized Dempster's rule yields a *bba* m_p with 2 focal sets: the singleton $\{\theta_1\}$ and the FoD Θ. This *bba* has the following expression:

$$\begin{aligned} m_p(\{\theta_1\}) &= \prod_{i=1}^{N} x_{i/p} + \underbrace{x_{j/p}x_{k/p}\ldots x_{L/p}}_{\substack{1:N-1\ \ terms \\ j,k,\ldots,L=1\ldots N \\ j\neq k\neq\ldots\neq L\neq i}} \sum_{i=1}^{N}(1 - x_{i/p}) \\ m_p(\Theta) &= \prod_{i=1}^{N}(1 - x_{i/p}) \end{aligned} \tag{2}$$

3.3 Decision Making

Relatively to a space point p, we construct the pignistic transformation (denoted by $BetP_p$) that permits the construction of the probabilities needed for decision making. The decision is based on selecting the hypothesis $\hat{\theta}$ with the largest pignistic probability: $\hat{\theta} = \max_{i=0,1} BetP_p(\{\theta_i\})$.

A space point p is covered if: $\hat{\theta} = \theta_1$ and $BetP_p(\{\theta_1\}) \geq Th_p$. The threshold ($Th_p$) value is an application-specific user-specified parameter.

3.4 Some Possible Extensions

In this section, we discuss how discounting factors [11, 12] can be included in our solution in order to handle deployment-related issues such as *sensor reliability* and *challenging environments*.

When sensors are vulnerable to misreading or malfunctioning due to their quality, we consider such sensors as only partially reliable. For $\alpha \in [0,1]$, let $(1-\alpha)$ be the degree of "confidence" we assign to the sensor. It can be encoded into a *bba* defined on the set $\{reliable, not\ reliable\}$ such that:

$$m(reliable) = 1 - \alpha \ and \ m(not\ reliable) = \alpha \tag{3}$$

Suppose that the *bba* m on Θ represents the sensor report about the actual value of Θ. The result of combining the sensor report with the *bba* given in equation (3) is a new *bba* denoted m^α, defined as:

$$m^\alpha(A) = (1-\alpha).m(A) \quad for \ A \subset \Theta$$

$$m^\alpha(\Theta) = \alpha + (1-\alpha).m(\Theta)$$

If a priori knowledge on sensor reliability is available, discounting factors associated to the sensors will be used in the evidence combination step.

Although sensors may be of good quality and provide accurate readings, external factors can greatly influence their correct functioning [12]. In this case, sensors are vulnerable to misreading or malfunctioning due to their locations in the RoI; we call such locations "*challenging locations*". A fully reliable sensor is considered as only partially reliable if it is deployed in a challenging location. Let $\beta \in [0,1]$ be the degree of "challenge" that we assign to a location p in the RoI. It can be encoded into a *bba* defined on the set $\{challenging, not\ challenging\}$ such that:

$$m(challenging) = \beta \ and \ m(not\ challenging) = 1 - \beta \tag{4}$$

Suppose that the *bba* m on Θ represents the report of the sensor (deployed at a challenging location p) about the actual value of Θ. The result of combining the

sensor report with the *bba* given in equation (4) is a new *bba* denoted m^{β}, defined as:

$$m^{\beta}(A) = (1-\beta).m(A) \quad for \ A \subset \Theta$$

$$m^{\beta}(\Theta) = \beta + (1-\beta).m(\Theta)$$

Thus, discounting factors will be associated with deployment points. The available a priori knowledge on environmental factors is used to compute the discounting factors.

Some sensors are more vulnerable to misreading or malfunctioning due to their quality and/or their location in the RoI. Suppose the *bba* m on Θ represents the sensor report about the actual value of Θ. The result of combining the sensor report with the *bba* given in equation (3) and the *bba* given in equation (4) is a new *bba* denoted $m^{\alpha\beta}$, defined as:

$$m^{\alpha\beta}(A) = (1-\alpha)(1-\beta).m(A) \quad for \ A \subset \Theta$$

$$m^{\alpha\beta}(\Theta) = \beta + \alpha.(1-\beta) + (1-\alpha)(1-\beta).m(\Theta)$$

Discounting factors will be associated with deployment and target points.

4 Problem Formalization

We define the USDP as the problem of covering a set of target points using an evidence-based coverage model. The RoI has a number of target points defining the set $T \subseteq$ RoI. We constrain the deployment of sensors in the set $D \subseteq$ RoI consisting of deployment points in the RoI.

In the simplest form of the USDP, the number of sensors should be kept to a minimum while also satisfying the coverage requirements for all points. We assume a non-uniform area coverage; thus each target point $p \in RoI$ is associated with a required minimum event detection probability threshold, denoted by th_p. The USDP can be formulated as an optimization problem:

$$\min \sum_{p \in D} x_p \tag{5}$$

$$\text{s.t.} \quad BetP(\{\theta_1\})_p \geq th_p, \ \forall p \in T \tag{6}$$

$$x_p \in \{0,1\}, \ \forall p \in D \tag{7}$$

Equation (5) defines the optimization problem where the objective is to minimize the deployment cost. The solution is constrained in equation (6), which requires that each target point is covered. Equation (7) defines x_p as a zero-one variable. If $x_p = 1$ than a sensor will be deployed at the point p. It should be pointed out that this formulation can be extended to include additional objectives such as the network connectivity. As the USDP is now clearly formulated as an optimization problem

which is NP-complete, previously proposed heuristics [4, 5, 9, 10] can be adapted to use our evidence-based coverage model.

5 Experimental Analysis

Using the proposed evidence-based coverage model, we have devised a constructive greedy-like placement algorithm to analyze its effect on the sensors placement. For a 20×20 RoI, D=RoI and a predefined T, Fig. 2 shows an example of obtained results. CDZs are represented by black disks and PIZs by gray disks.

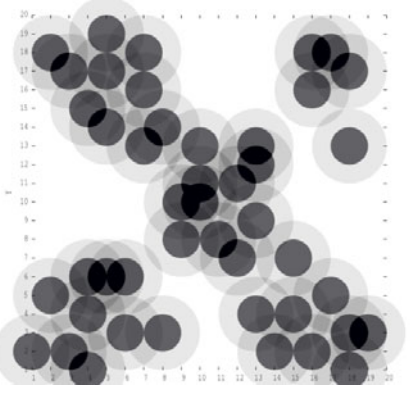


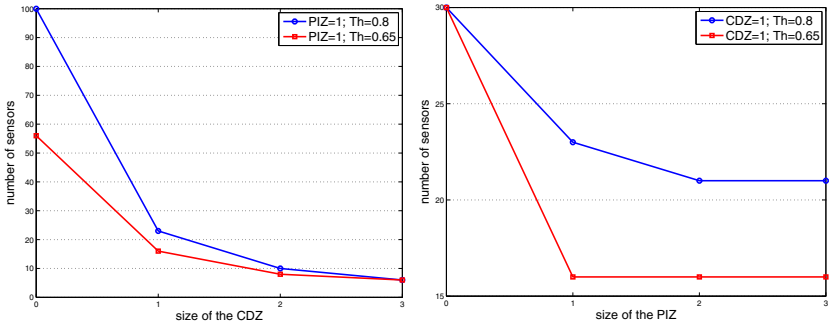
Fig. 2 An example of sensor placement in a 20×20 RoI.

Fig. 3 shows the effects of R_s, R_u, and Th (we assume a uniform coverage) on the performance. We see clearly on Fig. 3a that increasing the size of the CDZ reduces the cost of deployment. This reduction is more significant when the Th is more important. Intuitively, as the CDZ increases, more target points will be covered by the sensor resulting in better sensing quality. The same behavior is observed when increasing the size of the PIZ as depicted in Fig. 3b. These results show that considering noisy sensory data of multiple sensors (partial ignorance zones) can significantly improve sensing coverage by exploiting the collaboration among sensors.

6 Conclusion and Future Works

In this paper, a new methodology based on the transferable belief model has been proposed for handling the uncertainty-aware sensor networks deployment problem (USDP). We have first presented an evidence-based coverage model and discussed some possible extensions. Second, we have formulated the USDP as an optimization problem and we have discussed possible solutions.

Preliminary experimental analysis demonstrates very promising results of the proposed methodology. In the future, we plan to validate our proposal in real platform which will allow us to quantify the real benefit of the proposed methodology.



(a) Varying the size of the CDZ. (b) Varying the size of the PIZ.

Fig. 3 The effects of R_s, R_u, and Th on the performance.

References

1. Wang, B.: Coverage Control in Sensor Networks. Springer, New York (2010)
2. Younis, M., Akkaya, K.: Strategies and techniques for node placement in wireless sensor networks: A survey. Ad Hoc Netw. 6, 621–655 (2008)
3. Carter, B., Ragade, R.: An extensible model for the deployment of non-isotropic sensors. In: IEEE Sensors Applications Symposium, pp. 22–25 (2008)
4. Dhillon, S.S., Chakrabarty, K., Iyengar, S.S.: Sensor placement for grid coverage under imprecise detections. In: Proc. 5th Int. Information Fusion Conf., pp. 1581–1587 (2002)
5. Yi, Z., Chakrabarty, K.: Uncertainty-aware sensor deployment algorithms for surveillance applications. In: Proc. GLOBECOM 2003, vol. 5, pp. 2972–2976 (2003)
6. Clouqueur, T., Phipatanasuphorn, V., Ramanathan, P., Saluja, K.: Sensor deployment strategy for target detection. In: Proceedings of the 1st ACM WSNA, NY, USA, pp. 42–48 (2002)
7. O'Rourke, J.: Art Gallery Theorems and Algorithms. Oxford University Press (1987)
8. Kershner, R.: The number of circles covering a set. American Journal of Mathematics 61(3), 665–671 (1939)
9. Dhillon, S.S., Chakrabarty, K.: Sensor placement for effective coverage and surveillance in distributed sensor networks. In: Proc. IEEE WCNC 2003, vol. 3, pp. 1609–1614 (2003)
10. Aitsaadi, N., Achir, N., Boussetta, K., Pujolle, G.: A tabu search wsn deployment method for monitoring geographically irregular distributed events. Sensors 9, 1625–1643 (2009)
11. Smets, P., Kennes, R.: The transferable belief model. A.I. 66, 191–234 (1994)
12. Elouedi, Z., Mellouli, K., Smets, P.: Assessing sensor reliability for multisensor data fusion within the transferable belief model. IEEE Transactions on Systems, Man, and Cybernetics 34(1), 782–787 (2004)

A Quantitative Study of the Occurrence of a Railway Accident Based on Belief Functions

Felipe Aguirre, Mohamed Sallak, Walter Schön, and Fabien Belmonte

Abstract. In the field of railway systems, there is a great interest to include the human factor in the risk analysis process. Indeed, a great number of accidents are consider to be triggered by the human factors interacting in the situation. Several attempts have been made to include human factors in safety analysis, but they generally attack the problem in a qualitative way. The choice of qualitative methods arises from the difficulty to elicit human behavior and the effects on systems safety. This paper presents a first attempt to account for the human factor by using the generalized bayesian theory and fault tree analysis.

Introduction

As stated by Hale and Hovden [1], we are living in a new era of industrial safety in which three different factors are considered of major influence: technical, human and organisational factors. Indeed, they divide the era of industrial safety in three. First, there is the age of technology in which things were considered to fail because technology fails. This age started with the industrial revolution until the human factor started to take its place and the age of human factors was born. By this time, accidents were also attributed to human failures and new techniques for Human Reliability Assessment (HRA) were invented (THERP [2], HEART [3], JHEDI [4], etc.). Further on, the age of organisational factors arrived and accidents were also considered as consequences of organisational deficiencies.

Felipe Aguirre · Mohamed Sallak · Walter Schön
University of Technology of Compiègne. HEUDIASYC UMR 6599.
Centre de recherche de Royallieu BP 20529. Compiègne, France
e-mail: {faguirre,sallakmo,wschon}@utc.fr

Fabien Belmonte
Alstom Transport, 48 Rue Albert Dhalenne, 93482 Saint-Ouen cedex, France
e-mail: fabien.belmonte@transport.alstom.com

T. Denœux & M.-H. Masson (Eds.): Belief Functions: Theory & Appl., AISC 164, pp. 433–440.
springerlink.com

Human Reliability Assessment (HRA) has been often attacked as being of dubious validity [5, 6]. One of the main points of the critics is that human performance is not easy to quantify due to a large number of factors affecting it and the variability of one person over others. Nevertheless, studies have been made to prove the performance of current methods [7, 8, 9]. One of their drawbacks is that performance factors are dependent of the working context and a validation of the experimental data must be done. This suggest that there is a need for new methods of HRA [10].

Moreover, in the field of railway systems there has been some attemps to include human factors in safety analysis (i.e. [11]). However, these works are mostly of a qualitative nature mainly because of the difficulty to elicit human behaviour in a quantitative way. Indeed, it is considered that human behaviour is surrounded by epistemic uncertainties, thus needing the use of proper theories to represent and propagate the uncertainty in risk analysis. In the literature, several theories are proposed to treat epistemic uncertainties in risk analysis: Imprecise probabilities [12], Fuzzy sets [13], Monte-Carlo simulation [14], Belief functions theory [15, 16], etc. From the listed theories, Belief functions theory has been proven to be a promising one to treat epistemic uncertainty in risk analysis. However, only technical aspects are taken into account. This paper is a first attempt to integrate the human and organisational factors to risk analysis in railway accidents using belief functions theory in a quantitative way.

1 Risk Analysis Using Belief Functions Theory

Sallak et al.[15] presented a reliability model using belief functions theory. Further on, Aguirre et al. [16] improved the model by adding generalized reliability expressions that optimise the computational time. The model proposes to represent reliability using basic belief masses defined over a binary frame of discernement $\Omega = \{F_i, W_i\}$. F_i and W_i represent respectively the failing and working state of component i. Afterwards, using the concept of *minimal cut sets* and/or *minimal path sets*, the reliability masses are combined to obtain the final reliability of the system. A cut set is a set of components that induce the failure of the system if they are all in a failing state and a *minimal cut set* is a cut set that doens't contain any more cut sets as a subset. A *minimal path set* is defined in a similar way, the difference being that a path set containts a set of components that keeps the system in a working state if they are all working.

With a slight change on the definitions of the model, it can also be applied to risk analysis. Lets consider that for each basic event defined over the frame of discernement E_i, the state of belief on its ocurrence is bounded by $[bel(e_i), pl(e_i)]$ defined over $E_i = \{e_i, \overline{e}_i\}$. Now, lets say that the *minimal cut sets* contain sets of basic events that induce an undesired top event e_{top}. The belief over the undesired event is obtained using Eqs. 1 and the following notation:

N_C Number of minimal cuts in the system
C_i Index set of the i_{th} minimal cut set

$$bel(e_{top}) = \prod_{i=1}^{N_C} \left(1 - \prod_{j=1}^{size(C_i)} \left(1 - m(e_{C_i(j)})\right) \right)$$
$$pl(e_{top}) = \prod_{i=1}^{N_C} \left(1 - \prod_{j=1}^{size(C_i)} m(\overline{e}_{C_i(j)}) \right) \quad (1)$$

Therefore, the risk of arriving to the top undesired event is bounded by the interval $[bel(e_{top}), pl(e_{top})]$. The critical part of the model is the definition of the basic belief masses m^{E_i} over the events representing human actions. These events, are influenced by diffrent performance factors that modify the behavior of the agent performing the action. The key point is to use the Generalized Bayesian Theorem and the Disjunctive Rule of Combination to account for the influence of these performance factors on human behaviour in order to elicit the basic belief masses m^{E_i}.

2 Effect of Performance Factors on a Decision Taken by an Agent

To study the influence of performance factors on the decision taking process of an agent, we propose the use of a generalization of the Disjunctive Rule of Combination (DRC) introduced by Smets [17]. Lets assume there are two simple binary variables defined over the frames of discernement $\Theta = \{\theta, \overline{\theta}\}$ and $E = \{e, \overline{e}\}$

$$pl^E(e) = \sum_{\theta \subseteq \Theta} m_0^{\Theta}(\theta) pl^E(e|\theta)$$
$$= \sum_{\theta \subseteq \Theta} m_0^{\Theta}(\theta) \left(1 - \prod_{\theta_i \in \theta} \left(1 - pl^E(e|\theta_i)\right) \right) \quad (2)$$

From equation 2, it can be seen that if we have a basic belief mass over Θ (m_0^{Θ}) and conditional plausibility functions on E given θ and $\overline{\theta}$. Then, the plausibility induced on E can be computed. To better understand this, take a look at equation 2 developped for each of the subsets of E using matrix calculus:

$$\begin{bmatrix} pl(e) \\ pl(\overline{e}) \\ pl(e \cup \overline{e}) \end{bmatrix} = \begin{bmatrix} pl(e|\theta) & pl(e|\overline{\theta}) & pl(e|\theta \cup \overline{\theta}) \\ pl(\overline{e}|\theta) & pl(\overline{e}|\overline{\theta}) & pl(\overline{e}|\theta \cup \overline{\theta}) \\ pl(e \cup \overline{e}|\theta) & pl(e \cup \overline{e}|\overline{\theta}) & pl(e \cup \overline{e}|\theta \cup \overline{\theta}) \end{bmatrix} \begin{bmatrix} m(\theta) \\ m(\overline{\theta}) \\ m(\theta \cup \overline{\theta}) \end{bmatrix} \quad (3)$$

Equation 3 can be used to study the effect of a given factor Θ on the performance of an agent in a given event E. For example, let's suppose that Θ refers to a noisy environment and that E is a basic event refering to a bad decision taken by an agent. Under a noisy environment, the agent is prone to take a bad decision and this is

represented by a conditional belief mass $m^E(|\theta)$ (that we further convert into a plausibility space $pl^E(|\theta)$). Furthermore, we dispose of a basic belief mass m^Θ:

$$\begin{array}{lll} m(e|\theta) = 0.2 & pl(e|\theta) = 0.5 & m(\theta) = 0.6 \\ m(\overline{e}|\theta) = 0.5 \implies & pl(\overline{e}|\theta) = 0.8 & m(\overline{\theta}) = 0.3 \\ m(e \cup \overline{e}|\theta) = 0.3 & pl(e \cup \overline{e}|\theta) = 1 & m(\theta \cup \overline{\theta}) = 0.1 \end{array} \quad (4)$$

Additionally, lets assume that the conditional plausibilities $pl^E(|\overline{\theta})$ and $pl^E(|\theta \cup \overline{\theta})$ are vacuous—that is, we have nothing to say about E if we know that there isn't a noisy environment or if we ignore the state of Θ. After applying equation 3 it is concluded that the the risk that the agent takes a bad decision knowing that he is in a noisy environment is bounded by $[bel(e), pl(e)] = [0.12, 0.88]$

In practice, the agent is subject to more than one factor in the decision taking process. The way to proceed is to apply the DRC for each of the different factors and then combining the obtained masses over the event E. Note that if the factors are conditionaly independent, they are considered as distinct pieces of evidence and Dempster's rule can be used to combine the masses [18]

3 Case Study

The scenario takes place in a railway section between several stations. Some works were being done on one side of the railwaytrack heading east, forcing all the trains heading in this direction to take the other track in the opposite direction. In these kind of situations, protective measures are installed to avoid a potential accident and the planning of the commercial trains is slightly modified. The works extended over 10 km, they took place over the night and were programmed to end early in the morning in order to avoid unnecesary delays in the commercial trains. Three trains were involved in the night works (named $TTX1$, $TTX2$, $TTX3$ and $TTX4$). The works were supervised by a *foreman* (F) and the movement of the trains was controlled by a *signaller*.

Thanks to a series of unexpected events, the engineering train $TTX3$ encountered head to head with a commercial train going in the opposite direction. A delay on the works, forced to change the siding position of the trains $TTX2$ and $TTX3$ and the situation became complicated. When the proposed initial solution was going to be implemented, the *foreman* realized that the reserved track for trains $TTX2$ and $TTX3$ was blocked by the train $TTX1$, therefore, he was rushed into another change of plans under complicated measures. Badly advised by the *signaller*, he chose to park the trains in a manner that forced them to enter a protected zone running in an oposite direction. When the train $TTX3$ was going to engage the siding position, he encountered a closed sign that forbade him from crossing. The conductor of the train called the *signaller* who decided to authorize him to cross a signal at danger (or closed) by the application of a specific procedure without taking into account

the time table of the commercial trains. As a consequence of these series of events, the accident happened.

After a deep analysis of the circumstances, three basic events are identified as the precursors of the accident (see Fig. 1), namely:

- **Bad change of siding strategy:** The decision to change the parking plans of the engineering trains $TTX2$ and $TTX3$ was taken by the *foreman* thanks to several factors. To start with, the planning was delayed at the begining putting some extra pression to finish the work on time. In the second place, the *signaller* validated the change of plans with a lack of knowledge of the different parking positions of the train station and their state of occupancy. Finally, the *foreman* was in a very noisy environment when he took the decision.
- **Crossing permission of closed signal:** The *signaller* granted the permission to cross the closed sign to the train driver because he had the ilusion of a "safe" situation. Indeed, as all of the signals in the working area were closed, he thought that there wasn't any danger. He forgot that the signals were all closed on the moment that a train entered the protected area running in the opposite direction. As a consequence, he didn't take into account the time table of the commercial trains and didn't realized that a train was heading in the direction of the accident. This situation is considered a consequence of the lack of experience of the *signaller*. He also based his the decision on the fact that others had already granted the permission to cross the same signals. He thought, *"Somebody already verified the situation, it should work this time"* [19, 20].
- **Blocked road:** Finally, the train $TTX4$ was blocking the way of trains $TTX2$ and $TTX3$, causing more delays on their parking maneuvers. The way the intial plan was stated, this situation shouldn't have happened because the siding of the trains headed in an opposite direction.

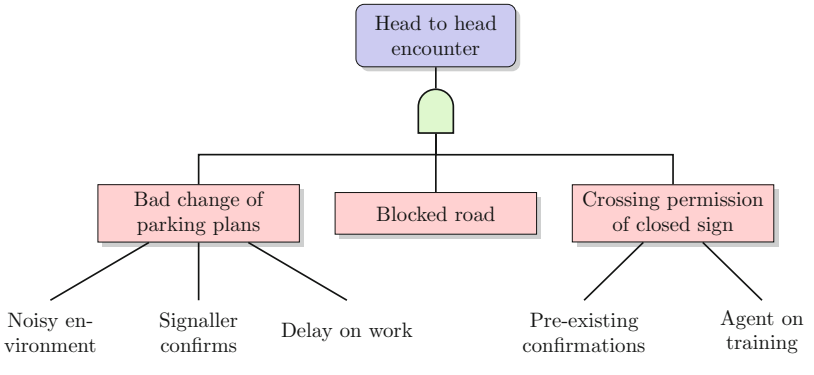


Fig. 1 Fault tree of the situation leading to a head to head encounter

Table 1 Conditional belief masses for the different factors

Θ	$m^E(e\|\theta)$	$m^E(\overline{e}\|\theta)$	$m^E(E\|\theta)$	$m^\Theta(\theta)$	$m^\Theta(\overline{\theta})$	$m^\Theta(\Theta)$
Noisy environment	0.02	0.98	0	0.6	0.3	0.1
signaller confirms	0.5	0.3	0.2	0.7	0.2	0.1
Delays on works	0.04	0.9	0.06	0.1	0.8	0.1
Pre-existing confirmations	0.1	0.8	0.1	0.9	0.5	0.5
Agent on training	0.1	0.75	0.15	0.1	0.8	0.1

3.1 Quantitative Analysis of the Accident

The three basic events identified are considered as the causes of the *head to head encounter*. One of them is considered as technical, the *blocked road* and the two others are considered as an event influenced by human and organisational factors as indicated in the following descriptions:

1. Bad change of parking plans
   - Noisy environment: The *foreman* had to take the decision in a noisy environment which made the comunication with the *signaller* complicated adn created a harsh work environment.
   - signaller confirms: The *signaller* confirmed that the new plan was feasible, but complicated. In reality it wasn't, but the *foreman* trusted in his opinion
   - Delay on works: Several delays at the beginning and during the work, forced the signaller to change the plans.
2. Crossing permission of closed sign
   - Pre-existing confirmations: The *signaller* based his decision on the fact that others had already granted the permission to cross the same signs. He thought, *"Somebody already verified the situation, it should work this time"*.
   - Agent on training: He was also in the final phase of a training period.

Fig. 1 describes the situation at hand. The fault tree consist of the three basci events related to the top event by an *and* gate. Moreover, the different factors that influence the events are also highlighted. The conditional masses as well as the basic beliefs on the different factors are showned in table 1.

First, the masses for the events *Bad change of parking plans* and *Crossing permission of closed sign* are elicited using Eq. 3. As for the *block road* event, its basic belief is considered to be bounded by $[0.5, 0.9]$. Finally, equations 1 are applied to obtain the bounds of the ocurrence of the top event. We get to the conclution that the risk of accident is bounded by $[0.00760, 0.06012]$.

4 Conclusions

This paper presents a first attemp to account for the human factor in risk analysis using belief functions theory. The generalized bayesian theorem is used to elicit the masses of the basic events when they are influenced by several factors and finally, the belief over the basic events is propagated to the top undesired event.

The advantage of the presented method is that our state of belief about the conditional relationship between the different factors and the basic beliefs doesn't have to be perfect. Indeed, the method is well suited to account for ignorance and a priori knowledge about the basic events are not needed.

Another advantage is that the method is capable of taking into account human, organisational and technical factors in the risk analysis of railway systems. Nevertheless, the elicitation method still needs to be improved and validated with studies similar to those made in other fields [7, 8, 9].

References

1. Hale, A.R., Hovden, J.: Management and culture: the third age of safety. A review of approaches to organizational aspects of safety, health and environment. In: Feyer, A.-M., Williamson, A. (eds.) Occupational Injury: Risk, Prevention, and Intervention, p. 129. Taylor & Francis (1998)
2. Swain, A., Guttmann, H.: Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report, Sandia National Labs, Albuquerque, NM (USA). Tech. Rep. (1983)
3. Williams, J.: HEART, a proposed method for assessing and reducing human error (1986)
4. Kirwan, B.: A resources flexible approach to human reliability assessment for PRA. In: Safety and Reliability Symposium, pp. 114–135. Elsevier Inc., Altrincham (1990)
5. Reason, J.: Human Error. Cambridge University Press (1990)
6. Hollnagel, E.: Human reliability analysis: Context and control. Academic Press, London (1993)
7. Kirwan, B.: The validation of three human reliability quantification techniques THERP, HEART and JHEDI: Part I technique descriptions and validation issues. Applied Ergonomics 27(6), 359–373 (1996)
8. Kirwan, B., Kennedy, R., Taylor-Adams, S., Lambert, B.: The validation of three Human Reliability Quantification techniques THERP, HEART and JHEDI: Part II Results of validation exercise. Applied Ergonomics 28(1), 17–25 (1997)
9. Kirwan, B.: The validation of three human reliability quantification techniques THERP, HEART and JHEDI: Part III Practical aspects of the usage of the techniques. Applied Ergonomics 28(1), 27–39 (1997)
10. Hovden, J., Albrechtsen, E., Herrera, I.A.: Is there a need for new theories, models and approaches to occupational accident prevention? Safety Science 48(8), 950–956 (2010)
11. Belmonte, F., Schön, W., Heurley, L., Capel, R.: Interdisciplinary safety analysis of complex socio-technological systems based on the functional resonance accident model: An application to railway traffic supervision. Reliability Engineering & System Safety 96(2), 237–249 (2011)
12. Utkin, L.V.: Imprecise Reliability: An Introductory Overview. In: Utkin, L.V., Coolen, F.P.A. (eds.) Computationa Intelligence in Reliability Engineering. SCI, vol. 40, pp. 261–306. Springer, Heidelberg (2007)

13. Tyagi, S.K., Pandey, D., Tyagi, R.: Fuzzy set theoretic approach to fault tree analysis. International Journal of Engineering, Science and Technology 2(5), 276–283 (2010)
14. Zio, E., Marella, M., Podofillini, L.: A Monte Carlo simulation approach to the availability assessment of multi-state systems with operational dependencies. Reliability Engineering & System Safety 92(7), 871–882 (2007)
15. Sallak, M., Schon, W., Aguirre, F.: Transferable belief model for reliability analysis of systems with data uncertainties and failure dependencies. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 224(4), 266–278 (2010), http://pio.sagepub.com/lookup/doi/10.1243/1748006XJRR292
16. Aguirre, F., Sallak, M., Schon, W.: Generalized expressions of reliability of series-parallel and parallel-series systems using the Transferable Belief Model. In: Berenguer, C., Grall, A., Guedes Soares, C. (eds.) Advances in Safety, Reliability and Risk Management, September 2011, p. 344. Taylor & Francis, Troyes (2011)
17. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1), 1–35 (1993)
18. Sentz, K., Ferson, S.: Combination of Evidence in Dempster- Shafer Theory. SANDIA, Albuquerque, New Mexico. Tech. Rep (April 2002)
19. Parry, G.W.: Suggestions for an improved HRA method for use in probabilistic safety assessment. Reliability Engineering & System Safety 49(1), 1–12 (1995)
20. Hollnagel, E., Green, M.: Accident Analysis and Barrier Functions. Institute for Energy Technology, Halden, Norway. Tech. Rep (1999)

Author Index