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.

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

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]

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

[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 $\mathrm{ext}[K_m(X)]$ 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]

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

for each $j = 2, \ldots, n$, while $P^\sigma(x_{\sigma(1)}) = \mathrm{Pl}(\{x_{\sigma(1)}\})$. 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.

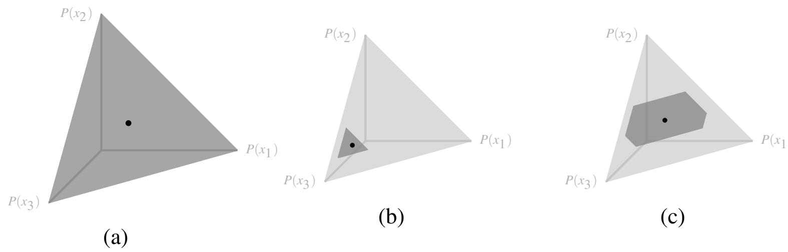

Fig. 1 Credal sets associated to BFs over a ternary variable X (dark gray) and their pignistic mass functions (black points). We consider: (a) the vacuous specification m_0; (b) a specification whose focal elements are only the singletons and on the universe with $m(\{x_1\}) = .05$, $m(\{x_2\}) = .15$,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

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

[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.

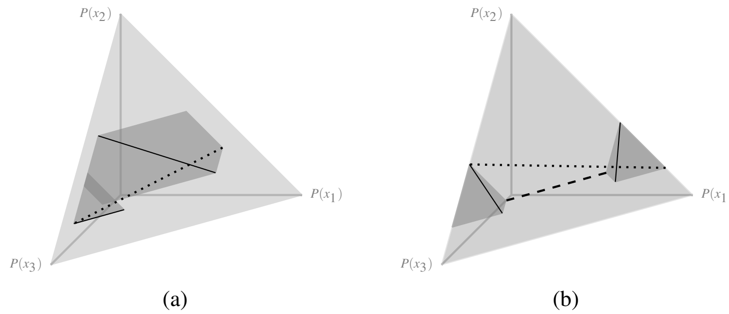

Fig. 2 Lower (dotted) and upper (dashed) distances and radiuses (continuous). Lines connect the extremes corresponding to the optima. As the distance is not Euclidean, lengths are not proportional to the actual distances. (a) compares the BFs in Figures 1, (b) a BF with $m(\{x_2\}) = .15$, $m(\{x_3\})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

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:

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

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 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.

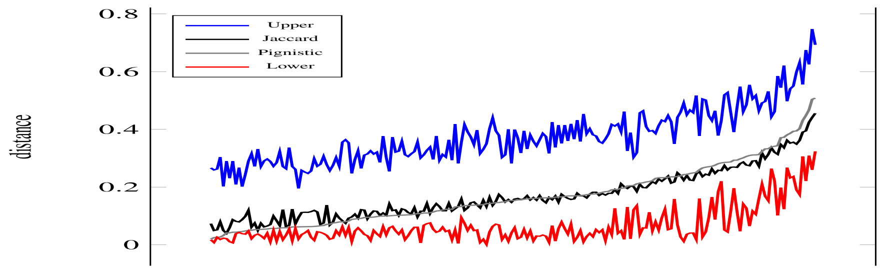

Fig. 3 Comparison between the bounds of the interval valued measure proposed in this paper, the Manhattan distance of the pignistic distributions and the distance based on the inner product of the masses with the *Jaccard index* [5]. The distances are computed on 1000 randomly generated pairs BFs defined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

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