Dealing with atypical instances in evidential decision-making

Benjamin Quost, Marie-Hélène Masson, Sébastien Destercke

Dealing with atypical instances in evidential classification

Benjamin Quost[1,2], Marie-Hélène Masson[1,3], and Sébastien Destercke[1,4]

[1] UMR UTC-CNRS 7253 Heudiasyc
[2] Sorbonne Universités, Université de Technologie de Compiègne, France
[3] Université de Picardie Jules Verne, France
[4] Centre National de la Recherche Scientifique, France

Abstract. When classifying an example on the basis of an observed population of (training) samples, at least three kinds of situations can arise where picking a single class may be difficult: high aleatory uncertainty due to the natural mixing of classes, high epistemic uncertainty due to the scarcity of training data, and non-conformity or atypicality of the example with respect to observations made so far. While the two first kinds of situations have been explored extensively, the last one still calls for a principled analysis. This paper is a first proposal to address this issue within the theory of belief function.

Keywords: Belief functions · Supervised classification · Epistemic and aleatoric uncertainty · Atypicality management · Novelty detection

1 Introduction

In a classification problem, assigning a new test instance to a class based on the set of training instances can be made difficult due to various kinds of uncertainties: aleatoric uncertainty or ambiguity (the classes being mixed, none seems to prevail), epistemic uncertainty (training data are scarce), non-conformity (or atypicality) of the test example to training observations. This last source of uncertainty, although related to epistemic uncertainty, cannot be tackled by additional training effort or by gathering additional training data. It is central in novelty, anomaly or outlier detection [2].

In this paper, we study how atypical instances can be accounted for in the framework of belief functions, in addition to situations of ambiguous or scarce data. The theory of belief functions, introduced in [3,6], and then further developed by Smets [9], provides a suitable framework for representing uncertainties. Atypicality has been already accounted for in different settings, such as distance rejection [5], or conformal predictions [7]. It was also addressed using belief functions (e.g., [1]), yet for specific kinds of atypicality.

To our knowledge, no principled, generic way to deal with atypicality has been proposed in the belief function framework. This paper can be seen as a preliminary contribution to this issue. We establish our basic setting in Section 2. Section 3 then discusses some desirable properties when accounting for atypicality, for which Section 4 proposes several strategies which are illustrated on a nearest-neighbor classification problem.

2 Basic setting

We recall here basic material on the theory of belief functions, a rich and flexible framework for managing uncertainty, which will be required in the rest of the paper.

2.1 Preliminaries on belief functions

Let us consider a variable ω taking values in a finite unordered set $\Omega = \{\omega_1, \dots, \omega_M\}$ called the frame of discernment. Partial knowledge regarding the actual value taken by ω is represented by a mass function [6] $m : 2^\Omega \to [0;1]$ such that

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

The sets $A \subseteq \Omega$ such that $m(A) > 0$ are called *focal sets* of m. If $m(A) = 1$ for some $A \subseteq \Omega$, m is said to be *categorical* and is denoted by m_A (if $A = \Omega$, m_Ω represents complete ignorance). It is often required that $m(\emptyset) = 0$; otherwise, $m(\emptyset)$ may have various interpretations, such as the degree of conflict after inconsistent pieces of information were aggregated, or the degree of belief that $\omega \notin \Omega$ (*open world assumption*).

Any mass function can be equivalently represented by a belief function *bel*, and a plausibility function *pl* defined, respectively, for all $A \subseteq \Omega$ by:

$$bel(A) = \sum_{B \subseteq A} m(B), \quad pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2}$$

Various strategies have been proposed for making decisions based on a belief function — see, e.g., [4]. Hereafter, we will denote by δ any decision operator to be applied to a mass function defined over the set of classes. For instance, the *interval dominance* operator, which may result in an imprecise decision (i.e. it may provide a set of classes) is defined as follows.

Definition 1 (Interval dominance). *Given a mass m, ω_i is said to dominate ω_j, noted $\omega_i \succ \omega_j$, if $bel(\{\omega_i\}) > pl(\{\omega_j\})$. The interval dominance rule consists in computing the set of non-dominated classes:*

$$\delta_{ID}(m) = \left\{\omega_i : pl(\{\omega_i\}) \geq bel(\{\omega_j\}) \, for \, all \; j \neq i \right\}. \tag{3}$$

2.2 Class membership model

We assume that a source provides us with information regarding the actual class of a test instance $\mathbf{x}$ to classify in the form of a mass function m. This mass function it is usually derived from a sample of N training instances $\mathbf{x}_i$ ($i = 1, \dots, \mathbf{x}_N$) observed in the same region than $\mathbf{x}$, and to which $\mathbf{x}$ is assumed to be similar. For example, when using decision trees, $\mathbf{x}$ is classified using the training data falling into the same the leaf node; in the K-NN algorithm, the decision is made based on the K closest training instances to $\mathbf{x}$.

In cautious classification, the quantity of information carried out by the training sample is taken into account in the decision process — thus allowing for cautious strategies, such as retaining a set of plausible classes, should the information be scarce. The imprecise Dirichlet model (IDM) makes it possible to provide such a cautious model of class frequencies in the form of a mass function m. In a nutshell, if $(n_1, n_2, ..., n_M)$ denote the counts of the classes in the observed sample of size N, with $\sum_k n_k = N$, the IDM produces the mass function

$$m^{IDir}(\Omega) = \frac{s}{N+s}, \quad m^{IDir}(\{\omega_i\}) = \frac{n_i}{N+s} \quad \forall i = 1, \dots, M, \tag{4}$$

where the parameter s can be interpreted as a number of additional unknown observations interfering with estimating the probabilities of the classes. This mass function produces in turn the bounds

$$\left[bel^{IDir}(\{\omega_i\}) = \frac{n_i}{N+s} \,; pl^{IDir}(\{\omega_i\}) = \frac{n_i + s}{N+s} \right] \quad \forall i = 1, \dots, M, \tag{5}$$

which account for both aleatoric uncertainty (which occurs if $n_1, \dots, n_M$ take similar values) and epistemic uncertainty (in which case the width of the intervals will increase when $s/(N+s)$ increases).

2.3 Conformity

It should be clear, however that the IDM does not take into account the typicality of a test instance of interest, that is, the extent to which it is similar to one of the training instances from which m^{IDir} is to be built. We assume here that this information is provided by a separate source, in the form of a conformity score $C \in [0;1]$: we have $C = 0$ for a completely unusual instance, and $C = 1$ for a normal one.

Figure 1 displays two situations where an instance $\mathbf{x}$ is to be classified into one of three classes $\{\omega_1, \omega_2, \omega_3\} = \Omega$, based on four training instances (with known classes). The same IDM would be built (the class counts being the same), but the level of typicality of $\mathbf{x}$ with respect to the four instances is very different. How this level of typicality may be assessed is left aside for now (for example, it may be derived from the distance of $\mathbf{x}$ to its first neighbour).

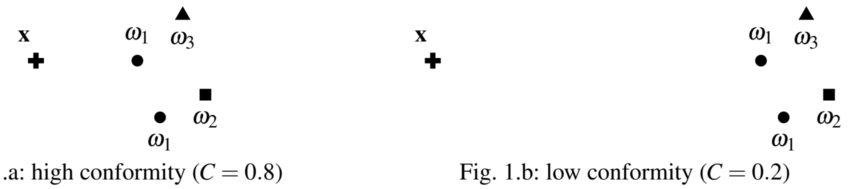

Fig. 1.a: high conformity ($C = 0.8$) Fig. 1.b: low conformity ($C = 0.2$)

Fig. 1. Two situations with identical class counts but different levels of typicality

The purpose of this paper is to determine how a mass function m related to the class of the instance $\mathbf{x}$ can be revised according to its level of typicality C. To this end, we introduce the notion of *conformity operator* Cf, which updates m into a new mass function $\mathrm{Cf}\,[m,C]$. Various properties may be desired (see Section 3), according to which different operators may be proposed (on which Section 4 focuses).

3 Desirable properties of conformity operators

Hereafter, by abuse of notation, $\mathrm{Cf}\,[pl,\cdot]$ (respectively, $\mathrm{Cf}\,[bel,\cdot]$) will stand for the plausibility function (resp., belief function) obtained from a revised mass function $\mathrm{Cf}\,[m,\cdot]$.

Property 1 (Class preference preservation). A conformity operator Cf preserves the preferential information over the classes if, for any $C \in [0;1]$,

$$pl(\{\omega_i\}) \le bel(\{\omega_j\}) \Rightarrow \mathrm{Cf}\,[pl,C]\,(\{\omega_i\}) \le \mathrm{Cf}\,[bel,C]\,(\{\omega_j\}). \tag{6}$$

Plainly put, it means that taking into account conformity does not alter interval dominance between classes (see Equation (1)).

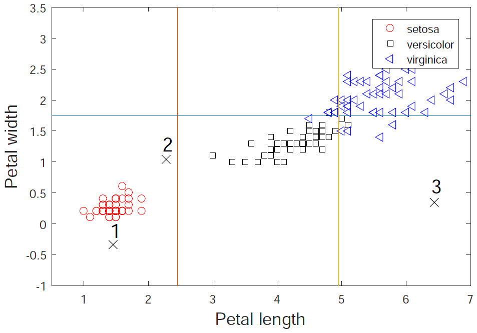

Fig. 2. Iris dataset example and some non-conformal examples

The example in Figure 2 displays the decision boundaries of a decision tree applied to the Iris dataset (two features were kept for illustrative purpose). Instances 1,2 and 3 are atypical. Class Setosa clearly dominates both others for instance 1, an information which may reasonably be kept in the revision process. However, it is more questionable for instance 2, which seems closer to class Versicolor than class Setosa once class dispersion is taken into account: then, its seems legitimate to discard the information brought by the training subset associated with the leaf of the tree. Overall, keeping the preference information inferred from the reference population seems reasonable if the model is unlikely to confuse atypicality with another source of uncertainty. Here, instance 2 is equally far from the Versicolor and Setosa classes, which the nature of the decision tree and hence the decision boundary make it impossible to detect.

Property 2 (Decision strengthening). A conformity operator Cf strengthens the decisions made with a strategy δ if

$$C \leq C' \Rightarrow \delta\left(\mathrm{Cf}\left[m,C\right]\right) \subseteq \delta\left(\mathrm{Cf}\left[m,C'\right]\right). \tag{7}$$

In other terms, the set of plausible classes for an instance should grow with its level of conformity: as it becomes atypical, classes previously deemed likely may be dropped off. This is similar to assuming an open-world, since known classes are discarded, possibly ending up with an empty set at the limit, similarly to conformal predictions.

Property 3 (Decision weakening). A conformity operator Cf weakens the decisions made with a strategy δ if

$$C \leq C' \Rightarrow \delta\left(\mathrm{Cf}\left[m,C'\right]\right) \subseteq \delta\left(\mathrm{Cf}\left[m,C\right]\right). \tag{8}$$

Contrary to Property 2, Property 3 is more in line with a closed world assumption, where Ω is assumed to necessarily contain all classes, but where the information related to an atypical example may seem too weak to provide a reliable prediction.

In the example above corresponding to Figure 2), assume that the decision for instance 3 is $\delta\left(m_3\right) = \{\text{Versicolor, Virginica}\}$. Then, requesting Property 2 would amount to discarding Versicolor, this class being too far; whereas Property 3 would rather leave us with complete ignorance, Setosa being then added to the set of plausible classes.

Note that other properties might also be proposed, for instance so as to specify the desired behaviour of the decision rule for extremely non-conformal examples (i.e. for $C \to 0$). Several will be examined in the next section.

4 Some conformity operators and their decision rule

This section investigates various operators in the light of the aforementioned properties. In a nutshell, they consist in computing a linear transformation of the initial mass according to the level of non-conformity.

4.1 Classical discounting in a closed world

A first strategy amounts to discounting[5] m according to the level $1-C$ of atypicity:

$$\mathrm{Cf}_1\left[m,C\right] = C\,m + (1-C)\,m_\Omega. \tag{9}$$

In the case of a mass function induced by the IDM, we thus have

$$\begin{cases} \mathrm{Cf}_1\left[m,C\right](\{\omega_i\}) = C\dfrac{n_i}{N+s}, & \text{for all } i = 1,\dots,M; \\ \mathrm{Cf}_1\left[m,C\right](\Omega) \quad = C\dfrac{s}{N+s} + (1-C). & \end{cases} \tag{10}$$

[5] The discounting ${}^{\varepsilon}m$ of m by a factor ε is defined by ${}^{\varepsilon}m(A) = (1-\varepsilon)m(A)$, for all $A \neq \Omega$; and ${}^{\varepsilon}m(\Omega) = (1-\varepsilon)m(\Omega) + \varepsilon$.

It should be clear that Cf_1 satisfies Property 3 (decision weakening) with respect to δ_{ID}, since discounting makes the belief-plausibility intervals wider. We have as extreme case $\delta_{ID}\left(\mathrm{Cf}_1\left[m,0\right]\right)=\Omega$. On the contrary, and for the same reason, Cf_1 does not satisfy Property 1 (class preference preservation). Such a rule therefore appears to be more consistent with a closed world assumption, where atypical instances are treated as being scarcely characterizable: complete atypicity should therefore be associated with complete ignorance.

4.2 Open world with an "unknown" class ω_u

Our second operator Cf_2 considers the open world assumption via an "unknown" class ω_u: that is, $\mathrm{Cf}_2\left[m,C\right]$ is now a mass function defined on a frame $\Theta=\Omega\cup\omega_u$:

$$\mathrm{Cf}_2\left[m,C\right]=Cm^{\uparrow\Theta}+(1-C)m_{\omega_u}, \tag{11}$$

where the *vacuous extension* $m^{\uparrow\Theta}$ of m onto Θ [8] is such that $m^{\uparrow\Theta}(A)=m(A)$ for any $A\subseteq\Omega$ and $m^{\uparrow\Theta}(A)=0$ for $A\not\subseteq\Omega$; and where $m_{\omega_u}(\{\omega_u\})=1$.

When applied to a mass function m^{IDir} generated by the IDM, this operator gives

$$\begin{cases}\mathrm{Cf}_2\left[m^{IDir},C\right](\{\omega_i\}) &= C\dfrac{n_i}{N+s} \quad \text{for all } \omega_i\in\Omega,\\ \mathrm{Cf}_2\left[m^{IDir},C\right](\{\omega_u\}) &= 1-C,\\ \mathrm{Cf}_2\left[m^{IDir},C\right](\Omega) &= C\dfrac{s}{N+s};\end{cases} \tag{12}$$

then, for any $\omega_i\in\Omega$, we have the following belief and plausibility values:

$$\mathrm{Cf}_2\left[bel^{IDir},C\right](\{\omega_i\})=C\frac{n_i}{N+s},\quad \mathrm{Cf}_2\left[pl^{IDir},C\right](\{\omega_i\})=C\frac{n_i+s}{N+s}, \tag{13}$$

and

$$\mathrm{Cf}_2\left[bel^{IDir},C\right](\{\omega_u\})=\mathrm{Cf}_2\left[pl^{IDir},C\right](\{\omega_u\})=1-C. \tag{14}$$

Applying δ_{ID} to an updated mass function $\mathrm{Cf}_2\left[m,C\right]$ (defined on Θ) satisfies Properties 1 and 2, with the extreme case $\delta_{ID}\left(\mathrm{Cf}_2\left[m,0\right]\right)=\{\omega_u\}$. Also note that

$$\delta_{ID}\left(\mathrm{Cf}_2\left[m,C\right]\right)\ni\omega_u \quad\Leftrightarrow\quad \max_{\omega_j\in\Omega}\mathrm{Cf}_2\left[bel,C\right](\{\omega_j\})\leq 1-C, \tag{15}$$

$$\delta_{ID}\left(\mathrm{Cf}_2\left[m,C\right]\right)=\{\omega_u\} \quad\Leftrightarrow\quad \max_{\omega_j\in\Omega}\mathrm{Cf}_2\left[pl,C\right](\{\omega_j\})<1-C. \tag{16}$$

As a consequence, the set of decisions will include $\{\omega_u\}$ only if the degree of support to each class is low. This inspires an alternative strategy, where ω_u is left aside when computing non-dominated classes, and added post-hoc should itself have been non-dominated.

Definition 2 (interval dominance with atypicity trigger). *Given a mass m defined on* $\Theta=\Omega\cup\{\omega_u\}$, *the* interval dominance with atypicity trigger *rule is defined by*

$$\delta_{ID:AT}(m)=\begin{cases}\delta_{ID}\left(m[\Omega]\right) & \textit{if} \displaystyle\min_{\omega_j\in\delta_{ID}(m[\Omega])} bel(\{\omega_j\})>1-C,\\ \delta_{ID}\left(m[\Omega]\right)\cup\{\omega_u\} & \textit{otherwise},\end{cases} \tag{17}$$

where the conditioning $m[\Omega]$ *of* m *on* Ω *[8] is such that* $m[\Omega](A) = \sum_{B\subseteq\Theta:B\cap\Omega=A} m(B)$*, for any* $A \subseteq \Omega$*.*

In a nutshell, the set of non-dominated classes is determined from well-identified classes (i.e., associated with an identified subpopulation), and a warning trigger is sent if the instance is deemed atypical. This strategy satisfies Property 1 and includes ω_u when $C = 1$: in particular, $\delta_{ID:AT}\left(\mathrm{Cf}_1\left[m,0\right]\right) = \Theta$, and $\delta_{ID:AT}\left(\mathrm{Cf}_2\left[m,0\right]\right) = \omega_u$.

4.3 Classical discounting in an open world

Finally, we propose a third operator where the mass m is first vacuously extended onto Θ and then discounted according to the level of atypicity C:

$$\mathrm{Cf}_3\left[m,C\right] = C\,m^{\uparrow\Theta} + (1-C)\,m_\Theta. \tag{18}$$

In the case of masses m^{IDir} obtained via the IDM, we thus obtain:

$$\begin{cases} \mathrm{Cf}_3\left[m^{IDir},C\right](\{\omega_i\}) = C\dfrac{n_i}{N+s} & \text{for all } \omega_i \in \Omega, \\ \mathrm{Cf}_3\left[m^{IDir},C\right](\Omega) = \quad C\dfrac{s}{N+s}, \\ \mathrm{Cf}_3\left[m^{IDir},C\right](\Theta) = \quad 1-C; \end{cases} \tag{19}$$

therefore, for any $\omega_i \in \Omega$, we have the following belief and plausibility values:

$$\mathrm{Cf}_3\left[bel^{IDir},C\right](\{\omega_i\}) = C\frac{n_i}{N+s}, \quad \mathrm{Cf}_3\left[pl^{IDir},C\right](\{\omega_i\}) = C\frac{n_i+s}{N+s} + 1 - C, \tag{20}$$

and

$$\mathrm{Cf}_3\left[bel^{IDir},C\right](\{\omega_u\}) = 0, \quad \mathrm{Cf}_3\left[pl^{IDir},C\right](\{\omega_u\}) = 1 - C. \tag{21}$$

Note that applying $\delta[ID]$ to $\mathrm{Cf}_3\left[m,\cdot\right]$ satisfies Property 3, since $\delta_{ID}\left(\mathrm{Cf}_3\left[m,\cdot\right]\right) = \Theta$.

Remark 1 (Open world assumption). The "unknown" class ω_u introduced above plays in spirit a role very similar to $\emptyset$ in the "canonical" open-world assumption (where $m(\emptyset)$ quantifies the belief that the instance is from a class outside Ω). However, introducing ω_u makes it possible to 1) distinguish between this degree of belief and the degree of conflict arising from combining belief masses, and 2) properly handle this degree of belief when it comes to decision making.

5 Conclusion and perspectives

Table 1 summarizes the properties of the conformity operators with their associated decision strategies presented in this paper. We recall the mass function used in each conformity operator, the set of decisions retrieved by the strategy when $C = 0$, the properties satisfied (class preference preservation, decision strengthtening, decision weakening), and the frame assumptions associated with the operator. Whether these latter assumptions should be accounted for depends on the application considered. For instance, in novelty detection, the open world assumption is clearly at work, which is not so clear in outlier or anomaly detection problems.

Future work will be conducted into two directions. First, we will study whether further properties should be required or desirable. Besides, we will compare the strategies to other approaches to dealing with atypical examples.

Conformity	mass function	associated	set of decisions	properties satisfied			type of frame	
operator	combined to m	decision rule	for $C = 0$	Prop. 1	Prop. 2	Prop. 3	open	closed
Cf$_1$	m_{Ω}	δ_{ID}	Ω			×	×	
Cf$_2$	m_{ω_u}	δ_{ID}	ω_u	×	×			×
Cf$_2$	m_{ω_u}	$\delta_{ID:AT}$	$\delta_{ID}\,(m[\Omega]) \cup \{\omega_u\}$	×		×	×	×
Cf$_3$	$m_{\Omega\cup\omega_u}$	δ_{ID}	$\Omega \cup \omega_u$			×		×

Table 1. Summary of the conformity operators and their associated decision strategies

Acknowledgements

This work was carried out in the framework of Labex MS2T (ANR-11-IDEX-0004-02), which were funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research.

References

1. Aregui, A., Denoeux, T.: Novelty detection in the belief functions framework. In: Proceedings of IPMU. vol. 6, pp. 412–419 (2006)
2. Carreño, A., Inza, I., Lozano, J.A.: Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. Artificial Intelligence Review (2019)
3. Dempster, A.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics **38**, 325–339 (1967)
4. Denoeux, T.: Decision-making with belief functions: a review. International Journal of Approximate Reasoning **109**, 87–110 (2019)
5. Dubuisson, B., Masson, M.: A statistical decision rule with incomplete knowledge about classes. Pattern Recognition **26**(1), 155 – 165 (1993)
6. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
7. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal on Machine Learning Research **9**, 371 – 421 (2008)
8. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. I. J. of Approximate Reasoning **9**, 1–35 (1993)
9. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence **66**, 191–234 (1994)