Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches

Thierry Denœux
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex, France
tdenoeux@hds.utc.fr
http://www.hds.utc.fr/~tdenoeux

Philippe Smets
IRIDIA - Université Libre de Bruxelles
50 av. Roosevelt, CP 194-6, 1050 Bruxelles, Belgium
psmets@ulb.ac.be
http://iridia.ulb.ac.be/~psmets

April 25, 2006

Abstract

The Transferable Belief Model (TBM) is a model to represent quantified uncertainties based on belief functions, unrelated to any underlying probability model. In this framework, two main approaches to pattern classification have been developed: the TBM model-based classifier, relying on the General Bayesian Theorem (GBT), and the TBM case-based classifier, built on the concept of similarity of a pattern to be classified with training patterns. Until now, these two methods seemed unrelated, and their connection with standard classification methods was unclear. We show in this paper that both methods actually proceed from the same underlying principle: the GBT, and that they essentially differ by the nature of the assumed available information. We also show that both methods collapse to a kernel rule in the case of precise and categorical learning data and for certain initial assumptions, and a simple relationship between basic belief assignments produced by the two methods is exhibited in a special case. These results shed new light on the issues of classification and supervised learning in the TBM. They also suggest new research directions, and they may help users in selecting the most appropriate method for each particular application, depending on the nature of the information at hand.

Keywords: Belief Functions, Evidential reasoning, Evidence Theory, Dempster-Shafer Theory, Classification, Supervised Learning, Pattern Recognition.

1 Introduction

The Transferable Belief Model (TBM) is a model to represent quantified uncertainties based on belief functions [7, 21], regardless of any underlying probability model [31, 28, 29]. We focus here on its application to classification problems.

In statistical pattern recognition, two main families of classifiers can be distinguished: methods that directly estimate posterior class probabilities (such as the k-nearest-neighbor rule, decision trees or multilayer perceptron classifiers), and methods based on density estimation, in which posterior probability estimates are computed from class-conditional densities and prior probabilities using Bayes' theorem (see, e.g., [15][14]). Another distinction can be made between parametric methods and nonparametric ones. Nonparametric methods are based on distribution-free estimation of posterior probabilities or class densities, using training patterns in a local region around a given attribute vector to be classified. A simple example is the *moving window classifier* [12], in which posterior probabilities are estimated by the proportions of those cases that belong to the various classes, among those that are "similar" to (i.e., within some distance of) the one to classify.

In the TBM framework, two main families of classifiers have also been defined:

- the TBM model-based classifier developed by Smets [25][27] and Appriou [2][3] based on the General Bayesian Theorem (GBT), the extension of Bayes' theorem in the TBM framework;

- the TBM case-based classifier developed by Denœux [8][10][11], a nonparametric method akin to statistical case-based classifiers such as the k-nearest neighbor (k-NN) rule.

Both produce a belief function over a set of classes given a vector of attributes, based on a learning set. This belief function is computed directly in the case-based method, whereas it is deduced from class-conditional beliefs using the model-based approach. As it will be explained in Section 4, a peculiarity of the case-based method is its ability to handle situations of partially supervised learning, in which the class membership

of each training pattern is described by a belief function over the set of classes.

Experimental comparisons between the case-based and model-based approaches to classification in the TBM framework have been presented in [4][32][6]. The goal of this paper is to study the relationships between these two methods at a conceptual level. In particular, it will be shown that:

- both methods can be derived from a single principle – the GBT;
- under some conditions, both methods have the same decision function, which corresponds to a kernel classification rule;
- the output belief functions produced by the two methods are linked by a simple relation, in a special case; both methods then collapse to the moving window classifier.

We hope that these results will help potential users in understanding the nature of the assumptions behind each of the two TBM classifiers, as well as their links with simple classification rules.

The rest of this paper is organized as follows. The necessary background information on the TBM is first recalled in Section 2. The TBM model-based and case-based classifiers are then described in Sections 3 and 4, respectively. The derivation of the case-based classifier from the GBT is then presented in Section 5, and the relationship between the two methods and kernel classifiers is studied in Section 6. Section 7 concludes the paper.

2 The TBM

2.1 Basic concepts

Belief functions quantify the beliefs held by an agent at a given time t, about the value of some unknown variable, given what the agent knows at t. The components are traditionally denoted as:

- Y for the agent, which can be any source, sensor, robot, computer program, etc;

- $EC_{Y,t}$ for what the agent Y knows at time t, also called the evidential corpus;
- $\boldsymbol{\omega}$, the variable, or unknown quantity, about which Y has some weighted opinions;
- Ω, the domain of $\boldsymbol{\omega}$, also called the frame of discernment.

For example,

$$bel^{\Omega}_{Y,t}\{\boldsymbol{\omega}\}[EC_{Y,t}](A) = 0.75$$

means that the degree of belief allocated by Y at time t, in the hypothesis that $\boldsymbol{\omega}$ belongs to $A \subseteq \Omega$, given the evidential corpus $EC_{Y,t}$, is 0.75. Whenever possible, indices are omitted, provided there is no risk of confusion.

The central element of the TBM is the basic belief assignment (bba), denoted m^{Ω}. For $A \subseteq \Omega$, $m^{\Omega}(A)$ is the part of belief allocated to A as a set of possible values for $\boldsymbol{\omega}$, and that, due to a lack of information, cannot be allocated to any strict subset of A. It satisfies $m^{\Omega}(A) \in [0, 1]$, with $\sum_{A \subseteq \Omega} m^{\Omega}(A) = 1$. The value of $m^{\Omega}(A)$ is called the basic belief mass (bbm) given to A, and A is termed a focal element of $m^{\Omega}(A)$ whenever $m^{\Omega}(A) > 0$. Two special cases are Bayesian bbas, whose focal elements are singletons, and the vacuous bba, defined by $m^{\Omega}(\Omega) = 1$. The vacuous bba encodes a belief state in which no information about the value of $\boldsymbol{\omega}$ is available. It will later be noted VBF (for vacuous belief function).

Note that we do not require $m^{\Omega}(\emptyset) = 0$. Under the so-called open-world assumption [26], the quantity $m^{\Omega}(\emptyset)$ is interpreted as the mass of belief assigned to the hypothesis that the actual value of $\boldsymbol{\omega}$ might not lie in Ω. When $m^{\Omega}(\emptyset) = 0$, the bba is said to be normal. A bba m^{Ω} such that $m^{\Omega}(\emptyset) > 0$ may be transformed into a normal bba M^{Ω} using the normalization operation defined by:

$$M^{\Omega}(A) = \frac{m^{\Omega}(A)}{1 - m^{\Omega}(\emptyset)}, \quad \forall A \subseteq \Omega, A \neq \emptyset, \tag{1}$$

and $M^{\Omega}(\emptyset) = 0$.

Related to the bba m^Ω, we define a belief function bel^Ω and a plausibility function pl^Ω as

$$bel^\Omega(A) = \sum_{\emptyset \neq B \subseteq A} m^\Omega(B), \quad \forall A \subseteq \Omega, \tag{2}$$

and

$$pl^\Omega(A) = \sum_{B \subseteq \Omega, B \cap A \neq \emptyset} m^\Omega(B) = bel^\Omega(\Omega) - bel^\Omega(\overline{A}), \quad \forall A \subseteq \Omega, \tag{3}$$

where $\overline{A}$ denotes the complement of A. The quantity $bel^\Omega(A)$ is interpreted as a degree of belief in A, taking into account the mass of belief given to A and nonempty subsets of A. In contrast, $pl^\Omega(A)$ measures to what extent one fails to believe in $\overline{A}$, i.e., to doubt A. Another interpretation is related to the conditioning operation introduced below: $pl^\Omega(A)$ is the maximum degree of belief that could potentially be assigned to A, if further evidence became available [31].

Two bbas m_1 and m_2 obtained from distinct (independent) sources may be combined using the unnormalized Dempster's rule of combination, also referred to as the unnormalized conjunctive rule, or the conjunctive rule for short, and defined as $m_{1 \textcircled{\cap} 2} = m_1 \textcircled{\cap} m_2$ with

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

Given a bba m and a nonempty subset B of Ω, a conditional bba given B, noted $m[B]$, may be defined as

$$m[B] = m \textcircled{\cap} m_B,$$

where m_B is the categorical bba defined by $m_B(B) = 1$. This operation is named the unnormalized Dempster's rule of conditioning (or unnormalized conditioning for short). Probabilistic conditioning is recovered as a special case when m is Bayesian, and the conditional bba is normalized.

A simple support function is defined as a belief function with a bba m^Ω so that there exists $A \subset \Omega$ ans $w \in [0, 1]$ such that $m^\Omega(A) = 1 - w$ and $m^\Omega(\Omega) = w$, all other masses being zero. We denote it as A^w. This notation is convenient as $A^x \textcircled{\cap} A^y = A^{xy}$.

In the TBM, decision making is based on the *pignistic transformation* [31, 30], which transforms m^Ω into the following *pignistic probability distribution* :

$$Betp(\omega) = \sum_{\{A \subseteq \Omega, \omega \in A\}} \frac{m^\Omega(A)}{(1 - m^\Omega(\emptyset))|A|}, \quad \forall \omega \in \Omega. \tag{4}$$

The simplest decision rule selects the element of Ω with the highest pignistic probability.

2.2 Operations on Product Spaces

Let $m^{\Omega \times \Theta}$ denote a bba defined on the Cartesian product $\Omega \times \Theta$ of two variables $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$. The marginal bba $m^{\Omega \times \Theta \downarrow \Omega}$ on Ω is defined, for all $A \subseteq \Omega$, as

$$m^{\Omega \times \Theta \downarrow \Omega}(A) = \sum_{\{B \subseteq \Omega \times \Theta \mid \text{Proj}(B \downarrow \Omega) = A\}} m^{\Omega \times \Theta}(B) \ , \tag{5}$$

where $\text{Proj}(B \downarrow \Omega)$ denotes the projection of B onto Ω, defined as

$$\text{Proj}(B \downarrow \Omega) = \{\omega \in \Omega \mid \exists \theta \in \Theta, (\omega, \theta) \in B\} \ . \tag{6}$$

Conversely, let m^Ω be a bba on Ω. Its vacuous extension on $\Omega \times \Theta$ is defined as:

$$m^{\Omega \uparrow \Omega \times \Theta}(B) = \begin{cases} m^\Omega(A) & \text{if } B = A \times \Theta \text{ for some } A \subseteq \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Note that $\Omega \times \Theta$ can be seen as a refining of Ω, i.e., a frame with finer granularity than Ω, such that there exists a one-to-one mapping between Ω and a partition of $\Omega \times \Theta$ [21]. The concept of vacuous extension was initially introduced by Shafer [21] in the general case of an arbitrary refining of Ω.

The conditioning operation introduced in Section 2.1 may be generalized to product spaces as follows [27]. Let $m^{\Omega \times \Theta}$ denote a bba on $\Omega \times \Theta$, and $m_B^{\Omega \times \Theta}$ the bba on $\Omega \times \Theta$ with single focal set $\Omega \times B$ with $B \subseteq \Theta$, i.e., $m_B^{\Omega \times \Theta}(\Omega \times B) = 1$. The conditional bba on Ω given $\boldsymbol{\theta} \in B$ is defined as:

$$m^\Omega[B] = (m^{\Omega \times \Theta} \textcircled{\cap} m_B^{\Omega \times \Theta})^{\downarrow \Omega} \tag{8}$$

The inverse operation is termed the *ballooning extension* [27]. Let $m^\Omega[B]$ denote the conditional bba on Ω, given $\boldsymbol{\theta} \in B \subseteq \Theta$. The ballooning extension of $m^\Omega[B]$ on

$\Omega \times \Theta$ is the least committed bba, whose conditioning on B yields $m^{\Omega}[B]$ (see [27] for detailed justification). It is obtained as:

$$m^{\Omega}[B]^{\Uparrow \Omega \times \Theta}(C) = \begin{cases} m^{\Omega}[B](A) & \text{if } C = (A \times B) \cup (\Omega \times (\Theta \setminus B)) \text{ for some } A \subseteq \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Example 1 *Consider two frames $\Omega = \{\omega_1, \omega_2\}$ and $\Theta = \{\theta_1, \theta_2\}$, each containing only two elements. Let $m^{\Omega \times \Theta}$ be a joint bba defined as:*

$$\begin{aligned} m^{\Omega \times \Theta}(\{(\omega_1, \theta_1)\}) &= 0.3 \\ m^{\Omega \times \Theta}(\{(\omega_1, \theta_2), (\omega_2, \theta_1), (\omega_2, \theta_2)\}) &= 0.2 \\ m^{\Omega \times \Theta}(\{(\omega_1, \theta_1), (\omega_1, \theta_2)\}) &= 0.5 \end{aligned}$$

Marginalization on Ω yields:

$$\begin{aligned} m^{\Omega \times \Theta \downarrow \Omega}(\{\omega_1\}) &= 0.3 + 0.5 = 0.8 \\ m^{\Omega \times \Theta \downarrow \Omega}(\Omega) &= 0.2. \end{aligned}$$

The vacuous extension of $m^{\Omega \times \Theta \downarrow \Omega}$ does not recover $m^{\Omega \times \Theta}$, as some information is lost in the marginalization process. We have:

$$\begin{aligned} m^{(\Omega \times \Theta \downarrow \Omega) \uparrow \Omega \times \Theta}(\{(\omega_1, \theta_1), (\omega_1, \theta_2)\}) &= 0.8 \\ m^{(\Omega \times \Theta \downarrow \Omega) \uparrow \Omega \times \Theta}(\Omega \times \Theta) &= 0.2. \end{aligned}$$

Conditioning $m^{\Omega \times \Theta}$ on θ_2 yields:

$$\begin{aligned} m^{\Omega}[\theta_2](\{\omega_1\}) &= 0.5 \\ m^{\Omega}[\theta_2](\Omega) &= 0.2 \\ m^{\Omega}[\theta_2](\emptyset) &= 0.3. \end{aligned}$$

Finally, the ballooning extension of $m^{\Omega}[\theta_2]$ is:

$$\begin{aligned} m^{\Omega}[\theta_2]^{\Uparrow \Omega \times \Theta}(\{(\omega_1, \theta_2), (\omega_1, \theta_1), (\omega_2, \theta_1)\}) &= 0.5 \\ m^{\Omega}[\theta_2]^{\Uparrow \Omega \times \Theta}(\Omega \times \Theta) &= 0.2 \\ m^{\Omega}[\theta_2]^{\Uparrow \Omega \times \Theta}(\Omega \times \{\theta_1\}) &= 0.3. \end{aligned}$$

2.3 Discounting

We want to build m_Y^Ω, quantifying the agent's beliefs about the actual value of Ω. The agent has no prior beliefs about Ω, and all it gets is what source S reports. Assume S provides the bba m_S^Ω.

- If the agent believes that S is reliable, it accepts S's beliefs, and equates its beliefs to S's beliefs. Hence $m_Y^\Omega[R] = m_S^\Omega$ where R denotes the hypothesis that S is reliable.

- If the agent does not accept that S is reliable, then it discards S's beliefs, and it is left with its own initial beliefs, hence $m_Y^\Omega[\neg R] = VBF$ where VBF denotes the vacuous belief function and $\neg R$ indicates the hypothesis that S is not reliable.

- Suppose the agent is not sure whether S is fully reliable, then it only partially accept S's beliefs. Its beliefs will be somewhere between m_S^Ω and the VBF.

Let $\mathcal{R} = \{R, \neg R\}$ be the space Reliable (R), not Reliable ($\neg R$). Let $m_Y^\mathcal{R}$ be the agent's bba about S's reliability. The agent's bbas given m_S^Ω and $m_Y^\mathcal{R}$ is then computed as:

$$m_Y^\Omega[m_S^\Omega, m_Y^\mathcal{R}] = \Big(\, m_Y^\Omega[R]^{\Uparrow \Omega\times\mathcal{R}} \textcircled{\cap} \, m_Y^\Omega[\neg R]^{\Uparrow \Omega\times\mathcal{R}} \textcircled{\cap} \, m_Y^{\mathcal{R}\uparrow \Omega\times\mathcal{R}} \, \Big)^{\downarrow\Omega}.$$

Let $d = 1 - m_Y^\mathcal{R}(\{R\})$. The bba $m_Y^\Omega[m_S^\Omega, m_Y^\mathcal{R}]$ depends only on m_S^Ω and d, so one can write $m_Y^\Omega[m_S^\Omega, m_Y^\mathcal{R}] = m_S^\Omega[disc = d]$ with, for all $A \subseteq \Omega$,

$$m_S^\Omega[disc = d\,](A) = \begin{cases} (1-d) \cdot m_S^\Omega(A) & \text{if } A \neq \Omega \\ d + (1-d) \cdot m_S^\Omega(\Omega) & \text{if } A = \Omega. \end{cases}$$

This operation is called *discounting*. It was introduced in [21] and explained in [27].

Example 2 *Assume that source S provides the bba m_S^Ω defined on $\Omega = \{\omega_1, \omega_2, \omega_3\}$ as*

$$\begin{aligned} m_S^\Omega(\{\omega_1\}) &= 0.5 \\ m_S^\Omega(\{\omega_2, \omega_3\}) &= 0.3 \\ m_S^\Omega(\Omega) &= 0.2, \end{aligned}$$

and the agent has a degree of belief $m_Y^{\mathcal{R}}(\{R\}) = 0.6$ *that the source is reliable. Then its belief on* Ω *is represented by the bba:*

$$\begin{aligned} m_Y^{\Omega}(\{\omega_1\}) &= m_S^{\Omega}[disc = 0.4](\{\omega_1\}) = 0.3 \\ m_Y^{\Omega}(\{\omega_2, \omega_3\}) &= m_S^{\Omega}[disc = 0.4](\{\omega_2, \omega_3\}) = 0.18 \\ m_Y^{\Omega}(\Omega) &= m_S^{\Omega}[disc = 0.4](\Omega) = 0.52. \end{aligned}$$

2.4 Maximal de-discounting

Consider a bba m^{Ω} on Ω. Let $m^{\Omega}[disc = d]$ be the result of its discounting by factor $d < 1$. If we knew d and $m^{\Omega}[disc = d]$, we could easily recover m^{Ω}:

$$\begin{aligned} m^{\Omega}(A) &= \frac{m^{\Omega}[disc = d](A)}{1 - d} \qquad \forall A \subset \Omega, \\ m^{\Omega}(\Omega) &= 1 - \sum_{A \subset \Omega} m^{\Omega}(A) \end{aligned}$$

We call this procedure the *de-discounting*. It is just the inverse of the discounting operation.

When $d = 1$, the discounted bba is vacuous, and there is no hope to recover the initial bba.

When d is unknown and we know the bba that resulted from the discounting, denoted m^{Ω}, then we can imagine all possible values for the discount rate. The discount rate must be smaller than $m^{\Omega}(\Omega)$.

Proposition 1 *Let a bba* m^{Ω}*. Let* m' *be the bba given by*

$$m'(A) = \frac{m^{\Omega}(A)}{1 - d} \qquad \forall A \subset \Omega, \tag{10}$$

$$m'(\Omega) = \frac{m^{\Omega}(\Omega) - d}{1 - d}. \tag{11}$$

The largest coefficient d *such that* m' *is a bba is* $d = m^{\Omega}(\Omega)$*.*

Proof. To get $m'(A) \geq 0$ in (10), we must have $d \leq 1$. To satisfy $m'(\Omega) \geq 0$ in (11), we must have $d \leq m^{\Omega}(\Omega)$. Thus the largest value for d is $m^{\Omega}(\Omega)$. In that case, we have:

$$\sum_{A \subseteq \Omega} m'(A) = \frac{1}{1 - m^{\Omega}(\Omega)} \left(\sum_{A \subseteq \Omega} m^{\Omega}(A) - m^{\Omega}(\Omega) \right) = 1.$$

Hence m' is a bba. □

The maximal (or boldest) de-discounting consists in de-discounting m^Ω by $m^\Omega(\Omega)$. The result of this operation will be noted $m^{\Omega*}$. We call it maximal as the resulting bba allocates a mass 0 to Ω.

This operation is in fact the dual of normalization (1):

- In normalization, the mass $m^\Omega(\emptyset)$ given to $\emptyset$ is proportionally redistributed among the masses given to the non empty sets.
- In maximal de-discounting, the mass $m^\Omega(\Omega)$ given to the Ω is proportionally redistributed among the masses given to the sets not equal to Ω.

Maximal de-discounting is also related to the negation of a bba. The negation of a bba m is a bba, denoted $\overline{m}$, such that $\overline{m}(A) = m(\overline{A}),\ \forall A \subseteq \Omega$ [13]. Given a bba m, build its negation, normalize it and take the negation of the result. The end result is the same as the one obtained with the maximal de-discounting of m.

Example 3 *Consider again the discounted bba of Example 2. Maximal de-discounting of this bba yields*

$$\begin{aligned} m'(\{\omega_1\}) &= \frac{0.3}{1-0.52} = 0.625 \\ m'(\{\omega_2,\omega_3\}) &= \frac{0.18}{1-0.52} = 0.375. \end{aligned}$$

3 The TBM model-based classifier

3.1 Problem statement

Classification problems can typically be represented according to the following schema. Let

- $\Omega = \{\omega_k\}_{k=1}^K$ be a finite set of mutually exclusive classes (categories, groups);
- $\mathcal{X}$ be the domain of the J-dimensional attribute (feature) vector $\boldsymbol{x}$;
- $\{o_i\}_{i=1}^I$ be a set of objects;

- $\boldsymbol{x}_i \in \mathcal{X}$ be the attribute vector of object o_i;
- $c_i \in \Omega$ be the class of object o_i;
- $e_i = (\boldsymbol{x}_i, c_i)$ be the data collected for object o_i;
- $\mathcal{L} = \{e_i\}_{i=1}^{I}$ denote the database, or learning set.

The problem is: given an object o with a known attribute $\boldsymbol{x}$ but an unknown class, produce the agent's belief about the actual value c of its class, based on $\mathcal{L}$.

3.2 The General Bayesian theorem

Let $m^{\mathcal{X}}[\omega_k]$ denote the bba $m^{\mathcal{X}}\{\boldsymbol{x}\}[c = \omega_k]$, for every $k = 1, \ldots, K$. It is the agent's bba about the $\boldsymbol{x}$ value of object o under the hypothesis that its actual class is ω_k.

Suppose vector $\boldsymbol{x}$ has been observed, and the agent's prior belief on c is represented by a vacuous belief function on Ω. We want to quantify the agent's bba about c, given all the available information.

The solution is given by the General Bayesian Theorem (GBT). Its origin and justification can be found in [2][25][27]. It is based on

- computing the ballooning extensions of the conditional bbas $m^{\mathcal{X}}[\omega_k]$;
- combining these bbas conjunctively, to get

$$m^{\Omega\times\mathcal{X}} = \textcircled{\cap}_{k=1}^{K} m^{\mathcal{X}}[\omega_k]^{\Uparrow\Omega\times\mathcal{X}} \; ;$$

- conditioning on $\Omega \times \{\boldsymbol{x}\}$ by Dempster's rule of conditioning,
- and marginalizing the result on Ω.

It produces $m^{\Omega}[\boldsymbol{x}]$, the a posteriori bba on Ω induced by the set of conditional bbas, a vacuous a priori bba and the observation $\boldsymbol{x}$:

$$m^{\Omega}[\boldsymbol{x}] = m^{\Omega}\{c\}\left[\left\{m^{\mathcal{X}}[\omega_k]\right\}_{k=1}^{K}, \boldsymbol{x}\right] = m^{\Omega\times\mathcal{X}}[\Omega \times \{\boldsymbol{x}\}]^{\downarrow\Omega},$$

which was shown [25][27] to be equal to

$$m^{\Omega}[\boldsymbol{x}] = \textcircled{\cap}_{k=1}^{K} \overline{\omega_k}^{\, pl^{\mathcal{X}}[\omega_k](\boldsymbol{x})}, \tag{12}$$

where $\overline{\omega_k}$ denotes the complement of $\{\omega_k\}$ in Ω. In particular, we have, for every $A \subseteq \Omega$:

$$m^{\Omega}[\boldsymbol{x}](A) = \prod_{\omega_k \in A} pl^{\mathcal{X}}[\omega_k](\boldsymbol{x}) \prod_{\omega_k \in \overline{A}} (1 - pl^{\mathcal{X}}[\omega_k](\boldsymbol{x})) \quad (13)$$

$$pl^{\Omega}[\boldsymbol{x}](A) = 1 - \prod_{\omega_k \in A} (1 - pl^{\mathcal{X}}[\omega_k](\boldsymbol{x})). \quad (14)$$

Should there be some non vacuous a priori bba on Ω, it is simply combined with $m^{\Omega}[\boldsymbol{x}]$ by the application of the conjunctive rule of combination.

Note that $m^{\Omega}[\boldsymbol{x}]$ in (13) may be subnormal, i.e., we may have $m^{\Omega}[\boldsymbol{x}](\emptyset) > 0$. The normalized form of $m^{\Omega}[\boldsymbol{x}]$ will be noted $M^{\Omega}[\boldsymbol{x}]$. All details on these relations can be found in [27].

3.3 Special case

Let us assume that the plausibility of observing $\boldsymbol{x}$ given $c = \omega_k$ is defined as:

$$pl^{\mathcal{X}}[\omega_k](\boldsymbol{x}) = \frac{N(\boldsymbol{x}, k)}{N(k)}, \quad (15)$$

where $N(\boldsymbol{x}, k)$ is the number of observations from class ω_k in a ball S_r of radius r (according to some distance measure δ), centered at $\boldsymbol{x}$, and $N(k)$ is the total number of examples in class ω_k.

Let us further assume that we have an a priori belief on c in the form of a Bayesian bba defined as the empirical class distribution in the training set:

$$m^{\Omega}\{c\}(\{\omega_k\}) = \frac{N(k)}{I},$$

where I is the total number of training examples.

If we normalize all beliefs, then the a posteriori belief on Ω is given by the Bayesian bba:

$$\begin{aligned} M^{\Omega}\{c\}[\boldsymbol{x}](\{\omega_k\}) &\propto \frac{N(k)}{I} pl^{\Omega}[\boldsymbol{x}](\{\omega_k\}) \\ &\propto \frac{N(k)}{I} pl^{\Omega}[\omega_k](\boldsymbol{x}) \\ &\propto \frac{N(k)}{I} \frac{N(\boldsymbol{x}, k)}{N(k)} \\ &\propto N(\boldsymbol{x}, k), \end{aligned}$$

in which case:

$$M^{\Omega}\{c\}[\boldsymbol{x}](\{\omega_k\}) = \frac{N(\boldsymbol{x},k)}{N(\boldsymbol{x})}, \tag{16}$$

with $N(\boldsymbol{x}) = \sum_{k=1}^{K} N(\boldsymbol{x},k)$.

Example 4 *As a simple real-world classification problem, let us consider the Cushing's syndrome dataset shown in Table 1, and taken from [1, Tables 11.1-3] (see also [20, page 11]). This dataset is represented in Figure 1. Cushing's syndrome is a hypersensitive disorder associated with over-secretion of cortisol by the adrenal gland. The dataset has three types of the syndrom: adenoma (a), bilateral hyperplasia (b), and carcinoma (c). These three classes will be noted, respectively,* ω_1, ω_2 *and* ω_3. *The attributes are urinary excretion rates (mg/24h) of the steroid metabolites tetrahydrocortisone and pregnanetriol. A logarithmic transformation was applied to both attributes. There are 25 learning examples.*

Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. *We have* $N(1) = 7$, $N(2) = 12$ *and* $N(3) = 6$. *Assume that* r *is set to 0.5, and* $\boldsymbol{x} = (0.8, 0.4)$. *There are 10 learning examples in the corresponding neighborhood of* $\boldsymbol{x}$, *with the following counts:*

$$N(\boldsymbol{x},1) = 2, \quad N(\boldsymbol{x},2) = 5, \quad N(\boldsymbol{x},3) = 3.$$

We thus have

$$\begin{aligned} pl^{\mathcal{X}}[\omega_1](\boldsymbol{x}) &= 2/7 \\ pl^{\mathcal{X}}[\omega_2](\boldsymbol{x}) &= 5/12 \\ pl^{\mathcal{X}}[\omega_3](\boldsymbol{x}) &= 3/6 = 1/2. \end{aligned}$$

label	class	tetrahydrocortisone (mg/24h)	pregnanetriol (mg/24h)
a_1	ω_1	3.1	11.70
a_2	ω_1	3.0	1.30
a_3	ω_1	1.9	0.10
a_4	ω_1	3.8	0.04
a_5	ω_1	4.1	1.10
a_6	ω_1	1.9	0.40
a_7	ω_1	2.6	0.1
b_1	ω_2	8.3	1.00
b_2	ω_2	3.8	0.20
b_3	ω_2	3.9	0.60
b_4	ω_2	7.8	1.20
b_5	ω_2	9.1	0.60
b_6	ω_2	15.4	3.60
b_7	ω_2	7.7	1.60
b_8	ω_2	6.5	0.40
b_9	ω_2	5.7	0.40
b_{10}	ω_2	13.6	1.60
b_{11}	ω_2	5.1	0.4
b_{12}	ω_2	13.0	0.8
c_1	ω_3	10.2	6.40
c_2	ω_3	9.2	7.90
c_3	ω_3	9.6	3.10
c_4	ω_3	53.8	2.50
c_5	ω_3	15.8	7.60
c_6	ω_3	12.9	5.0

Table 1: Cushing's syndrome dataset.

Hence, from (13):

$$\begin{aligned}
m^\Omega[\boldsymbol{x}](\emptyset) &= (1-2/7)(1-5/12)(1-1/2) = 0.2083 \\
m^\Omega[\boldsymbol{x}](\{\omega_1\}) &= 2/7 \times (1-5/12)(1-1/2) = 0.0833 \\
m^\Omega[\boldsymbol{x}](\{\omega_2\}) &= (1-2/7) \times 5/12 \times (1-1/2) = 0.1488 \\
m^\Omega[\boldsymbol{x}](\{\omega_3\}) &= (1-2/7)(1-5/12) \times 1/2 = 0.2083 \\
m^\Omega[\boldsymbol{x}](\{\omega_1,\omega_2\}) &= 2/7 \times 5/12 \times (1-1/2) = 0.0595 \\
m^\Omega[\boldsymbol{x}](\{\omega_1,\omega_3\}) &= 2/7 \times (1-5/12) \times 1/2 = 0.0833 \\
m^\Omega[\boldsymbol{x}](\{\omega_2,\omega_3\}) &= (1-2/7) \times 5/12 \times 1/2 = 0.1488 \\
m^\Omega[\boldsymbol{x}](\{\Omega\}) &= 2/7 \times 5/12 \times 1/2 = 0.0595.
\end{aligned}$$

After combining with the a priori bba:

$$m^\Omega\{c\}(\{\omega_1\}) = \frac{7}{25}, \quad m^\Omega\{c\}(\{\omega_2\}) = \frac{12}{25}, \quad m^\Omega\{c\}(\{\omega_3\}) = \frac{6}{25},$$

we obtain:

$$\begin{aligned}
m^\Omega[\boldsymbol{x}](\{\omega_1\}) &= (0.0833 + 0.0595 + 0.0833 + 0.0595) \times 7/25 = 0.08 \\
m^\Omega[\boldsymbol{x}](\{\omega_2\}) &= (0.1488 + 0.0595 + 0.1488 + 0.0595) \times 12/25 = 0.20 \\
m^\Omega[\boldsymbol{x}](\{\omega_3\}) &= (0.2083 + 0.0833 + 0.1488 + 0.0595) \times 6/25 = 0.12,
\end{aligned}$$

the rest being given to the empty set. After normalization, we get:

$$\begin{aligned}
M^\Omega[\boldsymbol{x}](\{\omega_1\}) &= 0.08/(0.08 + 0.20 + 0.12) = 0.2 \\
M^\Omega[\boldsymbol{x}](\{\omega_2\}) &= 0.20/(0.08 + 0.20 + 0.12) = 0.5 \\
M^\Omega[\boldsymbol{x}](\{\omega_3\}) &= 0.12/(0.08 + 0.20 + 0.12) = 0.3.
\end{aligned}$$

4 The TBM case-based classifier

4.1 Principle

This method was originally introduced in [8]. It consists in considering each example of the training set as an item of evidence regarding the class membership of a new

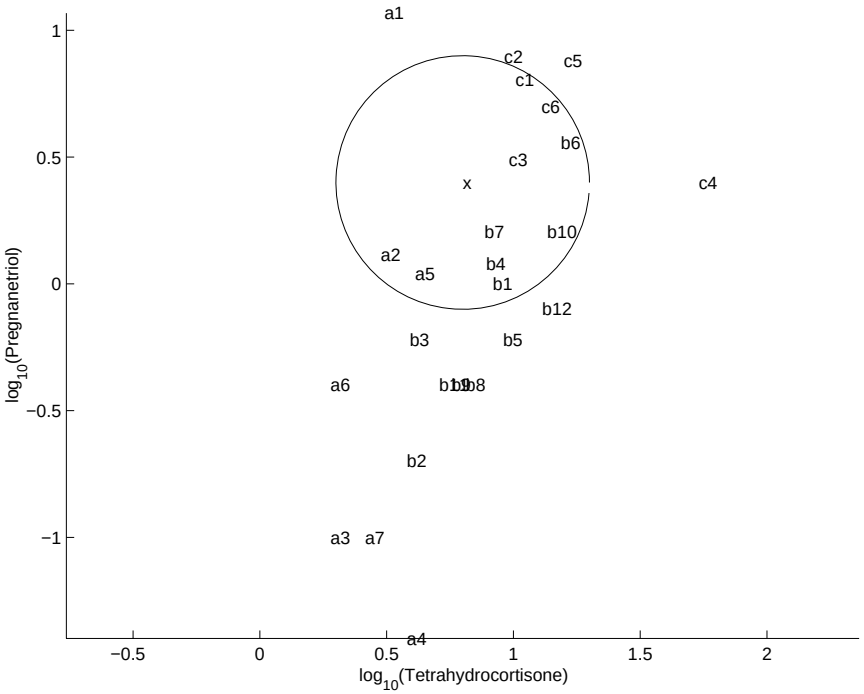

Figure 1: Cushing's syndrome data. Vector $\boldsymbol{x}$ with unknown class label is represented by x, together with the circle with center $\boldsymbol{x}$ and radius $r = 0.5$.

vector to be classified. The strength of this evidence is assumed to depend on a measure of distance between feature vectors, which is accounted for by discounting each item of evidence with a discount rate defined as a function of distance.

A characteristic of this approach resides in the possibility of taking into account partial knowledge of the class of training patterns [11]. A general form of the database is:

$$\mathcal{L} = \left\{(\boldsymbol{x}_i, m^{\Omega}\{c_i\})\right\}_{i=1}^{I},$$

where $m^{\Omega}\{c_i\}$ is a bba on the set of classes Ω expressing the agent's beliefs about the actual class c_i of object o_i. Particular cases of $\mathcal{L}$ are based on:

- *precise and categorical* knowledge when for every o_i, we have $m^{\Omega}\{c_i\}(\{\omega_k\}) = 1$ for some $\omega_k \in \Omega$ (this corresponds to the classical case);
- *imprecise and categorical* knowledge when for every o_i, we have $m^{\Omega}\{c_i\}(C_i) = 1$ for some $C_i \subseteq \Omega$, with at least one C_i such that $|C_i| > 1$;
- *probabilistic uncertainty* when for every o_i, $m^{\Omega}\{c_i\}(A) = 0$ for all $A \subseteq \Omega$ such that $|A| \neq 1$;
- *uncertainty according to the TBM theory*, where $m^{\Omega}\{c_i\}$ can be any bba on Ω.

Using data from object o_i, the agent's beliefs concerning the class c of a new vector $\boldsymbol{x}$ can be expressed as:

$$m^{\Omega}\{c\}[(\boldsymbol{x}_i, m^{\Omega}\{c_i\}), \boldsymbol{x}] = m^{\Omega}\{c_i\}[disc = d_i], \tag{17}$$

with the discount rate d_i defined as a function of the distance δ_i between $\boldsymbol{x}$ and $\boldsymbol{x}_i$ as $d_i = \varphi(\delta_i)$, where φ is a nondecreasing function taking values in $[0, 1]$. The interpretation of (17) is straightforward: if δ_i is small, then each learning example $e_i = (\boldsymbol{x}_i, m^{\Omega}\{c_i\})$ is regarded as a reliable source of information regarding the class c of the pattern to be classified, and the discount rate is small. The greater the distance from $\boldsymbol{x}$ to $\boldsymbol{x}_i$, the less informative the training example, and the higher the discount rate.

The data are assumed to be distinct pieces of evidence. Thus the overall belief is:

$$m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}] = \textcircled{\cap}_{i=1}^{I} m^{\Omega}\{c_i\}[disc = d_i]. \tag{18}$$

Decision is made using the pignistic transformation (4) [9]. The discount rates can be automatically adjusted to minimize an error criterion [37, 10]. Variants of this method have been proposed in [18], [17] and [36]. This approach has been successfully applied in many domains including bioinformatics [34, 35, 24, 23], medical image processing [6], remote sensing [36] and machine diagnosis [33]. An extension to nonparametric regression has been proposed in [19].

4.2 Special case

Suppose the class of every object is precise and categorical. It means that, for every i, we have $m^{\Omega}\{c_i\}(\{c_i\}) = 1$. It can also be written as $m^{\Omega}\{c_i\} = \{c_i\}^0$, using the notation for simple support functions introduced in Section 2.1.

The bba regarding the class c of a new object o, given case i is defined by (17). It may be written in that case:

$$m^{\Omega}\{c_i\}[disc = d_i] = \{c_i\}^{d_i}.$$

The combination over the whole training set is

$$\begin{aligned} m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}] &= \textcircled{\cap}_{i=1}^{I} \{c_i\}^{d_i} && (19)\\ &= \textcircled{\cap}_{k=1}^{K} \textcircled{\cap}_{\{i, c_i = \omega_k\}} \{\omega_k\}^{d_i} && (20)\\ &= \textcircled{\cap}_{k=1}^{K} \{\omega_k\}^{D_k}, && (21) \end{aligned}$$

with

$$D_k = \prod_{\{i, c_i = \omega_k\}} d_i.$$

We thus have:

$$m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_k\}) = (1 - D_k) \prod_{\ell \neq k} D_\ell, \quad k = 1, \ldots, K, \tag{22}$$

$$m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\Omega) = \prod_{k=1}^{K} D_k, \tag{23}$$

and the remainder is given to the empty set.

Example 5 *Let us come back to the data of Example 4. The distances between* $\boldsymbol{x}$ *and each of its 10 neighboring points are given in Table 2. For simplicity, assume that* $d_i = \delta_i$ *if* $\delta_i \leq 0.5$*, and* $d_i = 0$ *otherwise. This amounts to considering those cases at a distance greater than 0.5 as irrelevant to classify* $\boldsymbol{x}$ *(a practical consequence of this assumption is that the number of bbas to be combined is reduced, thus decreasing computation time).*

We have $D_1 = 0.1745$*,* $D_2 = 0.0048$*, and* $D_3 = 0.0401$*. Hence, the unnormalized bba is given by*

$$\begin{aligned}
m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_1\}) &= (1 - 0.1745) \times 0.0048 \times 0.0401 = 1.5972 \times 10^{-4} \\
m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_2\}) &= 0.1745 \times (1 - 0.0048) \times 0.0401 = 0.007 \\
m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_3\}) &= 0.1745 \times 0.0048 \times (1 - 0.0401) = 8.0752 \times 10^{-4} \\
m^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\Omega) &= 0.1745 \times 0.0048 \times 0.0401 = 3.3765 \times 10^{-5},
\end{aligned}$$

the rest being given to the empty set. The normalized bba is thus equal to

$$\begin{aligned}
M^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_1\}) &= 0.0200 \\
M^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_2\}) &= 0.8744 \\
M^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_3\}) &= 0.1013 \\
M^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\Omega) &= 0.0042.
\end{aligned}$$

i	1	2	3	4	5	6	7	8	9	10
c_i	ω_1	ω_1	ω_2	ω_2	ω_2	ω_2	ω_2	ω_3	ω_3	ω_3
δ_i	0.431	0.405	0.417	0.334	0.418	0.214	0.387	0.457	0.204	0.431

Table 2: Data of Example 5.

The bba computed by this method is quite different from the one obtained in Example 4 using the model-based approach with similar data (the distances, though, were not used in the model-based method). Despite these differences, we will show in the next section that the case-based method can also be derived from the GBT, just as the model-based method, but with different initial assumptions.

5 Deriving the case-based classifier from the GBT

5.1 Underlying assumptions

Let us consider two objects o and o', with classes c and c' in Ω. The distance between these two objects is measured by a quantity $\delta \in \Delta$ (quite often Δ will be equal to $\mathbb{R}_+$, but the case of dissimilarities expressed on a discrete numerical or ordinal scale can be handled as well).

The pair (c, c') takes values in $\Omega \times \Omega$. In this space, let S denote the hypothesis that the two objects belong to the same class:

$$S = \{(\omega_k, \omega_k)\}_{k=1}^{K} \subset \Omega^2.$$

Assume that we only know:

- the conditional plausibility $pl^{\Delta}[S](\delta)$ of the observed distance δ (for all $\delta \in \Delta$) given that the two objects belong to the same class, and
- the conditional plausibility $pl^{\Delta}[\overline{S}](\delta)$ of the observed distance δ given that the two objects belong to different classes.

Let us consider the following coarsening of Ω^2: $\Theta = \{S, \overline{S}\}$. The GBT allows us to express our belief on Θ, given δ, from the plausibilities over Δ given S and $\overline{S}$. Using (13), we have:

$$m^{\Theta}\{c, c'\}[\delta](\{S\}) = pl^{\Delta}[S](\delta)(1 - pl^{\Delta}[\overline{S}](\delta)) \triangleq \alpha_S \quad (24)$$

$$m^{\Theta}\{c, c'\}[\delta](\{\overline{S}\}) = pl^{\Delta}[\overline{S}](\delta)(1 - pl^{\Delta}[S](\delta)) \triangleq \alpha_{\overline{S}} \quad (25)$$

$$m^{\Theta}\{c, c'\}[\delta](\Theta) = pl^{\Delta}[\overline{S}](\delta) pl^{\Delta}[S](\delta) \triangleq \alpha_{\Theta} \quad (26)$$

$$m^{\Theta}\{c, c'\}[\delta](\emptyset) = 1 - pl^{\Delta}[\overline{S}](\delta) - pl^{\Delta}[S](\delta) + pl^{\Delta}[\overline{S}](\delta) pl^{\Delta}[S](\delta) \triangleq \alpha_{\emptyset}. \quad (27)$$

The vacuous extension $m^{\Omega\times\Omega}\{c, c'\}[\delta]$ of $m^{\Theta}\{c, c'\}[\delta]$ on $\Omega \times \Omega$ is simply defined

as:

$$m^{\Omega\times\Omega}\{c,c'\}[\delta](S) = \alpha_S \quad (28)$$

$$m^{\Omega\times\Omega}\{c,c'\}[\delta](\overline{S}) = \alpha_{\overline{S}} \quad (29)$$

$$m^{\Omega\times\Omega}\{c,c'\}[\delta](\Omega\times\Omega) = \alpha_\Theta \quad (30)$$

$$m^{\Omega\times\Omega}\{c,c'\}[\delta](\emptyset) = \alpha_\emptyset \quad (31)$$

$$m^{\Omega\times\Omega}\{c,c'\}[\delta](A) = 0 \quad \forall A \in 2^{\Omega\times\Omega} \setminus \{S,\overline{S},\Omega\times\Omega,\emptyset\}. \quad (32)$$

Remark 1 *Note that the model described in this section is related to the concept of similarity profile introduced by Hüllermeier [16]. This author considers a general model of case-based reasoning, in which a case is composed of a situation, and an associated result, or outcome. Similarity measures are defined on the set of situations, and on the set of outcomes. A similarity profile specifies conditional probabilities of the form $P(\delta'|\delta)$, where δ is the similarity between situations (objects), and δ' the similarity between the corresponding outcomes (class labels). If labels are either identical or not, as assumed here, we have the special case where δ' is 0 or 1; these values correspond to the two hypotheses S and $\overline{S}$ that objects belong to the same class or not. Thus, $P(1|\delta)$ is the probability that two objects that are similar to the degree δ have different class labels. This probability corresponds to the belief α_S in (28). Likewise, $P(0|\delta)$ corresponds to the belief $\alpha_{\overline{S}}$ in (29). In the context of classification, it may be argued that it is more natural to express one's beliefs concerning the similarity of objects from the same class or from different classes, and deduce a belief function over $\Omega\times\Omega$ given δ using the GBT, as done here. The result is slightly more general that the model considered in [16], since we obtain a general belief function instead of a probability measure. Note that the TBM cased-based regression method described in [19] could also be formalized using a similar approach.*

5.2 Classification of a new case

Let us assume that we want to determine the class c of a new object, based on its distances to each of I objects in a learning set. Let δ_i denote the distance to

object o_i, and let $m^{\Omega}\{c_i\}$ denote the bba concerning the class c_i of object o_i. Let $e_i = (\delta_i, m^{\Omega}\{c_i\})$ denote the data referring to case o_i in the learning set.

The impact of case o_i on our knowledge regarding the class of the new object may be modeled by combining $m^{\Omega\times\Omega}\{c, c_i\}[\delta_i]$ (the "generic knowledge") with $m^{\Omega}\{c_i\}$. This can be done by:

- vacuously extending $m^{\Omega}\{c_i\}$ to $\Omega \times \Omega$;
- combining the resulting bba on $\Omega \times \Omega$ with $m^{\Omega\times\Omega}\{c, c_i\}[\delta_i]$ using the conjunctive rule of combination;
- marginalizing the result to obtain a bba on Ω related to variable c.

Formally:

$$m^{\Omega}\{c\}[e_i] = \left(m^{\Omega\times\Omega}\{c, c_i\}[\delta_i] \circledcirc m^{\Omega\uparrow\Omega\times\Omega}\{c_i\}\right)^{\downarrow\Omega}\{c\}.$$

To simplify the expression of $m^{\Omega}\{c\}[e_i]$, let us use the notation introduced in (24)-(27), and let $m_i = m^{\Omega}\{c_i\}$. Consider a mass $m_i(A)$ for $A \subseteq \Omega$, $A \neq \emptyset$. Take its product with the four masses α_S, $\alpha_{\overline{S}}$, $\alpha_{\emptyset}$, and α_{Θ}:

1. the product $\alpha_S m_i(A)$ is transferred to $\text{Proj}\left((\Omega \times A) \cap S \downarrow_1 \Omega\right) = A$, where $\downarrow_1$ denotes projection on the first component of product space $\Omega \times \Omega$;
2. the product $\alpha_{\overline{S}} m_i(A)$ is transferred to

$$\text{Proj}\left((\Omega \times A) \cap \overline{S} \downarrow_1 \Omega\right) = \begin{cases} \overline{A} & \text{if } |A| = 1, \\ \Omega & \text{if } |A| > 1; \end{cases}$$

3. the product $\alpha_{\emptyset} m_i(A)$ is transferred to $\emptyset$;
4. the product $\alpha_{\Theta} m_i(A)$ is transferred to Ω.

For $A = \emptyset$, the mass $m_i(A)$ is transferred to $\emptyset$.

The resulting bba $m^\Omega\{c\}[e_i]$ is given for $A \subseteq \Omega$ by:

$$m^\Omega\{c\}[e_i](A) = \begin{cases} \alpha_\emptyset(1 - m_i(\emptyset)) + m_i(\emptyset) & \text{if } A = \emptyset, \\ \alpha_S m_i(A) & \text{if } 0 < |A| < |\Omega| - 1, \\ \alpha_S m_i(A) + \alpha_{\overline{S}} m_i(\overline{A}) & \text{if } |A| = |\Omega| - 1, \\ \alpha_\Theta(1 - m_i(\emptyset)) + \alpha_S m_i(\Omega) + & \\ \alpha_{\overline{S}} \sum_{B \subseteq \Omega, |B| > 1} m_i(B) & \text{if } A = \Omega. \end{cases} \tag{33}$$

A bba on c given all the available learning information is then obtained by combining the bbas induced by each of the I cases in the learning set $\mathcal{L}$:

$$m^\Omega\{c\}[\mathcal{L}] = \textcircled{\cap}_{i=1}^{I} m^\Omega\{c\}[e_i].$$

5.3 Special case: Vacuous belief function on Δ given S

Assume that we have no information concerning the within-class distances when the two objects belong to the same class (i.e., when S is true). Then, $pl^\Delta[S](\delta) = 1$ for all $\delta \subseteq \Delta$. We also assume that m_i is normalized, i.e., $m_i(\emptyset) = 0$. From (24)-(27), we have:

$$\begin{aligned} \alpha_S &= 1 - pl^\Delta[\overline{S}](\delta) \\ \alpha_{\overline{S}} &= 0 \\ \alpha_\Omega &= pl^\Delta[\overline{S}](\delta) \\ \alpha_\emptyset &= 0 \end{aligned}$$

and (33) becomes:

$$m^\Omega\{c\}[e_i](A) = \begin{cases} \alpha_S m_i(A) & \text{if } 0 \leq |A| < |\Omega| \\ 1 - \alpha_S + \alpha_S m_i(\Omega) & \text{if } A = \Omega, \end{cases}$$

with $\alpha_S = 1 - pl^\Delta[\overline{S}](\delta_i)$. Hence, $m^\Omega\{c\}[e_i]$ is obtained by discounting m_i with a discount rate equal to $1 - \alpha_S = pl^\Delta[\overline{S}](\delta_i)$. We obtain exactly the case-based TBM classifier defined by (17), with $d_i = pl^\Delta[\overline{S}](\delta_i)$.

We have thus shown how to derive the case-based TBM classifier from the GBT and produced an interpretation of function φ relating the discount rate d_i to the distance δ_i.

Although the case-based and model-based classifiers can be derived from the same underlying principle (the GBT), we have seen in Examples 4 and 5 that they usually produce different results. In the next section, we study conditions under which both methods yield the same decisions, and we demonstrate a simple relationship between the output bbas in a special case.

6 TBM classifiers as kernel rules

6.1 Kernel rules

Kernel classification rules are classical methods in statistical pattern recognition (see, e.g., [12, chapter 10]). The simplest of these rules is the moving window rule, which simply takes the data points within a certain distance of the point to be classified, and decides according to the majority vote. Formally, this rule may be defined by the following decision function:

$$g(\boldsymbol{x}) = \omega_k \Leftrightarrow \sum_{\{i, c_i = \omega_k\}} \kappa_r(\delta_i) \geq \sum_{\{i, c_i = \omega_\ell\}} \kappa_r(\delta_i), \quad \forall \ell \neq k, \tag{34}$$

with

$$\kappa_r(\delta_i) = \begin{cases} 1 & \text{if } \delta_i \leq r \\ 0 & \text{otherwise,} \end{cases} \tag{35}$$

r being a parameter.

By replacing κ_r in (35) by a smoother nonnegative, monotone decreasing function (called a *kernel function*), we obtain a general *kernel classification rule*. Kernel-based rules can be derived from Parzen window estimates of density functions [15].

6.2 The TBM model-based and case-based classifiers as kernel rules

Model-based classifier

Consider the TBM model-based approach with the assumptions of Section 3.3, which leads to the normalized bba given by (16). The corresponding decision function is

$$g_{MB}(\boldsymbol{x}) = \omega_k \Leftrightarrow N(\boldsymbol{x}, k) \geq N(\boldsymbol{x}, \ell), \quad \forall \ell \neq k.$$

Since

$$N(\boldsymbol{x}, k) = \sum_{\{i, c_i = \omega_k\}} \kappa_r(\delta_i),$$

with κ_r defined by (35), this rule is exactly the moving window rule (34).

Now, let us assume that we replace $N(\boldsymbol{x}, k)$ in (15) by

$$N'(\boldsymbol{x}, k) = \sum_{\{i, c_i = \omega_k\}} \kappa(\delta_i), \tag{36}$$

where κ is now an arbitrary kernel function. We then have

$$M^{\Omega}[\mathcal{L}, \boldsymbol{x}](\{\omega_k\}) = \frac{\sum_{\{i, c_i = \omega_k\}} \kappa(\delta_i)}{\sum_{i=1}^{I} \kappa(\delta_i)}, \tag{37}$$

and the corresponding decision rule is then a general kernel rule.

We have shown that the TBM model-based classifier is a kernel classifier when the plausibility of $\boldsymbol{x}$ in each class is computed from (15) and (36) using a kernel function κ, and when the class proportions in the training set are taken as prior probabilities.

Case-based classifier

The case-based classifier with precise and categorical training data is also a kernel rule, as will now be demonstrated.

Let us consider the special case considered in Section 4.2, in which the class of each training pattern in known. From the expression of the output bba (21), it is clear

that the corresponding decision function is

$$\begin{aligned} g_{CB}(\boldsymbol{x}) = \omega_k \quad &\Leftrightarrow \quad D_k \leq D_\ell, \quad \forall \ell \neq k \\ &\Leftrightarrow \quad -\ln D_k \geq -\ln D_\ell, \quad \forall \ell \neq k \\ &\Leftrightarrow \quad \sum_{\{i,c_i=\omega_k\}} -\ln d_i \geq \sum_{\{i,c_i=\omega_\ell\}} -\ln d_i, \quad \forall \ell \neq k \\ &\Leftrightarrow \quad \sum_{\{i,c_i=\omega_k\}} -\ln \varphi(\delta_i) \geq \sum_{\{i,c_i=\omega_\ell\}} -\ln \varphi(\delta_i), \quad \forall \ell \neq k. \end{aligned}$$

This is a kernel rule, with the kernel function $\kappa = -\ln \varphi(\cdot)$.

If this kernel function is used in (15) and (36) to compute the class conditional plausibilities $pl_{\mathcal{X}}[\omega](\boldsymbol{x})$, the model-based classifier therefore yields exactly the same decisions as the case-based classifier.

Note that, although the TBM case-based classifier was originally introduced in [8] as a k-nearest neighbor rule, we have shown in this section that its "probabilistic" equivalent is not the voting k-nearest neighbor rule, but a kernel rule. However, the voting k-nearest neighbor rule can be recovered as a special case if one defines $d_i = 1$ if $\boldsymbol{x}_i$ is among the k nearest neighbors of $\boldsymbol{x}$, and 0 otherwise (this amounts to choosing an adaptive kernel function). This again yields the same decisions as the TBM model-based classifier with the same adaptive kernel function.

Although the decision functions of the two methods can be made identical, the computed bbas are fundamentally different. In particular, $M^\Omega[\mathcal{L}, \boldsymbol{x}]$ in (37) is Bayesian, whereas the bba computed using the case-based method in (22)-(23) is not. However, conditions under which a simple relationship exists between the two mass functions are studied in the next section.

6.3 Relationship between mass functions

The purpose of this section is to show that, given certain requirements, the mass function (22)-(23) produced by the case-based method is closely related to the mass function (16) produced by the model-based method, involving ratios $N(\boldsymbol{x}, k)/N(\boldsymbol{x})$.

Suppose that function φ relating d_i to δ_i is defined by:

$$d_i = \varphi(\delta_i) = \begin{cases} 1-\epsilon & \text{if } \delta_i \leq r \\ 1 & \text{otherwise,} \end{cases} \tag{38}$$

where r is some constant and $\epsilon \in (0,1)$ is an arbitrarily small positive number. As shown in Section 5, d_i can be interpreted as the plausibility of observing a distance δ_i for two patterns from different classes. Hence, (38) assumes a very weak form of knowledge, in which all values of δ_i greater than r have plausibility equal to one, and all values smaller than r are only slightly less plausible.

With this choice of φ, relations (22) and (23) become:

$$m^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_k\}) = \left(1-(1-\epsilon)^{N(\boldsymbol{x},k)}\right)(1-\epsilon)^{N(\boldsymbol{x})-N(\boldsymbol{x},k)}, \quad k=1,\ldots,K \tag{39}$$

$$m^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\Omega) = (1-\epsilon)^{N(\boldsymbol{x})} \tag{40}$$

and their normalized forms are:

$$M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_k\}) = \frac{(1-\epsilon)^{-N(\boldsymbol{x},k)}-1}{1+\sum_{\ell=1}^{K}\left((1-\epsilon)^{-N(\boldsymbol{x},\ell)}-1\right)} \tag{41}$$

$$M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\Omega) = \frac{1}{1+\sum_{\ell=1}^{K}\left((1-\epsilon)^{-N(\boldsymbol{x},\ell)}-1\right)} \tag{42}$$

When $\epsilon \to 0$, $M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\omega_k)$ tends to 0, and $M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\Omega)$ tends to 1, hence $\lim_{\epsilon\to 0} m^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}] = VBF$.

However, let us consider the bba $M^{\Omega *}\{c\}[\mathcal{L},\boldsymbol{x}]$ obtained by maximally de-discounting $M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}]$ (see Section 2.4). We have $M^{\Omega *}\{c\}[\mathcal{L},\boldsymbol{x}](\Omega) = 0$, and

$$\begin{aligned} M^{\Omega *}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_k\}) &= \frac{M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_k\})}{\sum_{\ell=1}^{K} M^{\Omega}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_\ell\})} && (43)\\ &= \frac{(1-\epsilon)^{-N(\boldsymbol{x},k)}-1}{\sum_{\ell=1}^{K}\left((1-\epsilon)^{-N(\boldsymbol{x},\ell)}-1\right)} && (44)\\ &= \frac{N(\boldsymbol{x},k)+O(\epsilon)}{\sum_{\ell=1}^{K} N(\boldsymbol{x},\ell)+O(\epsilon)} && (45)\\ &= \frac{N(\boldsymbol{x},k)+O(\epsilon)}{N(\boldsymbol{x})+O(\epsilon)}. && (46) \end{aligned}$$

Consequently, we have:

$$\lim_{\epsilon\to 0} M^{\Omega *}\{c\}[\mathcal{L},\boldsymbol{x}](\{\omega_k\}) = \frac{N(\boldsymbol{x},k)}{N(\boldsymbol{x})}. \tag{47}$$

We thus face the same issue as encountered with the ϵ-belief functions described in default reasoning [5]: the mass ratios are informative, although the masses themselves are not. This translates the idea that, under weak prior knowledge regarding between-class distances, a finite amount of data is never large enough to induce non vacuous beliefs; still, the ratio of the zero's is meaningful. An immediate consequence of (47) is that, for small enough ϵ, the decision function of the case-based classifier will be identical to that of the model-based classifer in the special case studied in Section 3.3, itself identical to the moving window rule.

We thus have shown that the TBM case-based and model-based classifiers produce similar bbas, up to a maximal de-discounting, if

- the class of each object in the learning set class is known precisely and categorically;
- the discount rates are defined by (38), with $\epsilon \to 0$.

Example 6 *Let us come back to the problem of Examples 4 and 5, with the data of Table 2. The normalized masses computed from (41)-(42), as well as the maximally de-discounted masses computed from (43)-(44) are plotted in Figure 2, as a function of* $\epsilon \in (0, 0.2]$. *As expected,* $M^{\Omega}\{c\}[\mathcal{L}, \boldsymbol{x}](\Omega)$ *tends to 1, whereas* $M^{\Omega *}\{c\}[\mathcal{L}, \boldsymbol{x}](\{\omega_k\})$ *tends to* $N(\boldsymbol{x}, k)/N(\boldsymbol{x})$, *when* ϵ *tends to 0.*

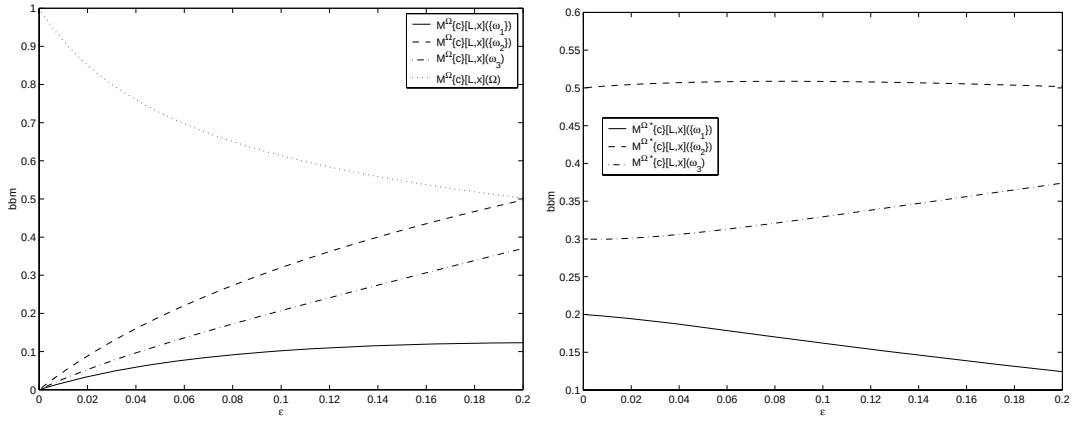

Figure 2: Convergence of bbms in Example 6.

7 Conclusions

In the TBM framework, two main approaches to pattern classification have been proposed. The model-based approach [25, 27, 2] computes a conditional belief function on the class variable c, given the feature vector $\boldsymbol{x}$, from the conditional belief functions on $\boldsymbol{x}$ given each class. It relies on the GBT as an essential inference mechanism, similar to the role played by the Bayes theorem in probability theory. The other approach, proposed more recently [8], is built around the idea of similarity between patterns, a central notion both in learning and in uncertain reasoning [22, 16]. The so-called case-based approach essentially transposes to a given pattern what is known regarding similar patterns in a database, using a discounting mechanism and conjunctive combination. Until now, these two methods seemed unrelated, and their connection to standard classification methods was unclear.

In this paper, we have been able to show that both methods actually proceed from the same underlying principle: the GBT, and that they essentially differ by the nature of the available information. The model-based approach is based on class-conditional belief functions on the observed feature vector $\boldsymbol{x}$. In contrast, the case-based approach is based on conditional belief functions regarding the distance between two patterns, given that they belong to the same class, or to different classes (what could be called within and between-class conditional belief functions on distances). In both cases, the GBT allows one to convert this information into a belief function on the class of a pattern, for which either the feature vector, or the distances to training patterns, have been observed. Although the two methods yield different results in the general case, they have been shown to collapse to a simple kernel rule (the moving window rule) in the case of precise and categorical learning data and for certain choices of the conditional belief functions in both methods.

We hope that these results will help to clarify the issues of supervised learning and classification in the belief functions framework. They also suggest new directions of research such as the generalization of the case-based approach to incorporate other information such as the conditional beliefs on the distances of two patterns taken from

each pair of classes, or the fusion of case-based and model-based classifiers. Finally, our results may help users in selecting the most appropriate method for each particular application, depending on the nature of the available information.

References

[1] J. Aitchison and I. R. Dunsmore. *Statistical Prediction Analysis*. Cambridge University Press, Cambridge, 1975.

[2] A. Appriou. Probabilités et incertitude en fusion de données multi-senseurs. *Revue Scientifique et Technique de la Défense*, (11):27–40, 1991.

[3] A. Appriou. Uncertain data aggregation in classification and tracking processes. In B. Bouchon-Meunier, editor, *Aggregation and Fusion of imperfect information*, pages 231–260. Physica-Verlag, Heidelberg, 1998.

[4] A. Bastière. Methods for multisensor classification of airbone targets integrating evidence theory. *Aerospace Science and Technology*, 6:401–411, 1998.

[5] S. Benferhat, A. Saffiotti, and P. Smets. Belief functions and default reasoning. *Artificial Intelligence*, 122(1–2):1–69, 2000.

[6] A.-S. Capelle, O. Colot, and C. Fernandez-Maloigne. Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information. *Information Fusion*, 5(3):203–216, 2004.

[7] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[8] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[9] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[10] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[11] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001.

[12] L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New-York, 1996.

[13] D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12:193–226, 1986.

[14] R. O. Duda, P. E Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, New-York, 2001.

[15] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Boston, 1990.

[16] E. Hüllermeier. Similarity-based inference as evidential reasoning. *International Journal of Approximate Reasoning*, 26(2):67–100, 2001.

[17] E. Lefèvre, O. Colot, and P. Vannoorenberghe. Belief function combination and conflict management. *Information Fusion*, 3(2):149–162, 2002.

[18] N. K. Pal and S. Gosh. Some classification algorithms integrating Dempster-Shafer theory of evidence with the rank nearest neighbor rule. *IEEE Trans. on Systems, Man and Cybernetics – Part A*, 31(1):59–66, 2001.

[19] S. Petit-Renaud and T. Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004.

[20] B. D. Ripley. *Pattern Recognition and Neural networks.* Cambridge University Press, Cambridge, 1996.

[21] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[22] G. Shafer. Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B*, 44:322–352, 1982.

[23] H. Shen and K.-C. Chou. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochemical and Biophysical Research Communications*, 337(3):752–756, 2005.

[24] H. Shen and K.-C. Chou. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochemical and Biophysical Research Communications*, 334(1):288–292, 2005.

[25] Ph. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical*. PhD thesis, Université Libre de Bruxelles, Brussels, Belgium, 1978. (in French).

[26] Ph. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[27] Ph. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[28] Ph. Smets. The normative representation of quantified beliefs by belief functions. *Artificial Intelligence*, 92(1–2):229–242, 1997.

[29] Ph. Smets. The Transferable Belief Model for quantified belief representation. In D. M. Gabbay and Ph. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 267–301. Kluwer Academic Publishers, Dordrecht, 1998.

[30] Ph. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133–147, 2005.

[31] Ph. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[32] P. Vannoorenberghe and T. Denœux. Likelihood-based vs. distance-based evidential classifier. In *The 10th IEEE International Conference on Fuzzy Systems*, volume 1, pages 320–323. IEEE, December 2001.

[33] B.-S. Yang and K. J. Kim. Application of Dempster-Shafer theory in fault diagnosis of induction motors using vibration and current signals. *Mechanical Systems and Signal Processing*, 20:403–420, 2006.

[34] N. M. Zaki, S. Deris, S. N. V. Arjunan, and R. M. Illias. Assignment of protein sequence to functional family using neural network and Dempster-Shafer theory. *Journal of Theoretics*, 5(1), 2003.

[35] W. Zhong, G. Altun, X. Tian, R. Harrison, P.C. Tai, and Y. Pan. Parallel protein secondary structure prediction based on neural networks. In *Proceedings of the 26th Annual International Conference of the Engineering in Medicine and Biology Society, 2004 (EMBC 2004)*, pages 2968 – 2971. IEEE, 2004.

[36] H. Zhu and O. Basir. An adaptive fuzzy evidential nearest neighbor formulation for classifying remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(8):1874–1888, 2005.

[37] L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. on Systems, Man and Cybernetics C*, 28(2):263–271, 1998.

Vitae

- **Thierry Denœux** graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and received a doctorate from the same institution in 1989. Currently, he is a Full Professor with the Department of Information Processing Engineering at the Université de Technologie de Compiègne, France. His research interests concern belief functions theory, fuzzy data analysis and, more generally, the management of imprecision and uncertainty in data analysis, pattern recognition and information fusion. He is the Editor-in-Chief of the *International Journal of Approximate Reasoning*.

- **Philippe Smets** (1938-2005) was a Professor at the Université Libre de Bruxelles, Bruxelles, Belgium, and created its Artificial Intelligence Laboratory (IRIDIA) in 1985. He coordinated several major research projects on uncertainty, which were funded by the European Union. He retired in 1999, but later proceeded with his researches on the transferable belief model, until his death on November 14th, 2005. Prof. Smets was a member of the editorial boards and program committees of most journals and conferences dealing with the numerical representation of uncertainty.

List of Figures

List of Tables