Handling possibilistic labels in pattern classification using Evidential reasoning [1]

Thierry Denœux and Lalla Meriem Zouhal

Université de Technologie de Compiègne
U.M.R CNRS 6599 Heudiasyc
BP 20529 - F-60205 Compiègne cedex - France
email: Thierry.Denoeux@hds.utc.fr

March 23, 2015

[1] Accepted for publication in *Fuzzy Sets and Systems*.

Abstract

A category of learning problems is considered, in which the class membership of training patterns is assessed by an expert and encoded in the form of a possibility distribution. Each example i thus consists in a feature vector $\mathbf{x}^i$ and a possibilistic label $(u_1^i, \ldots, u_c^i)$, where u_k denotes the possibility of that example belonging to class k. This problem is tackled in the framework of Evidence Theory. The evidential distance-based classifier previously introduced by one of the authors is extended to handle possibilistic labeling of training data. Two approaches are proposed, based either on the transformation of each possibility distribution into a consonant belief function, or on the use of generalized belief structures with fuzzy focal elements. In each case, a belief function modeling the expert's beliefs concerning the class membership of each new pattern is obtained. This information may then be either interpreted by a human operator to support decision-making, or automatically processed to yield a final class assignment through the computation of pignistic probabilities. Experiments with synthetic and real data demonstrate the ability of both classification schemes to make effective use of possibilistic labels as training information.

Keywords: Evidence Theory, Belief Functions, Pattern Recognition, Fuzzy Statistics and Data Analysis, Possibility Theory.

1 Introduction

In the conventional approach to pattern classification, a decision rule, or *classifier*, is typically generated from a set of training samples whose membership to each of c classes is unambiguously defined by *hard class labels* of the form:

$$\mathbf{u}^i \triangleq (u_1^i, \ldots, u_c^i) \in \{0,1\}^c$$

where $u_k^i = 1$ if pattern i belongs to class k and $u_k^i = 0$ otherwise.

In certain applications, however, such hard labels do not accurately reflect the nature of the available information. For example, it sometimes occurs that pattern categories are ill-defined and best represented as *fuzzy sets* of patterns [2]. Each training vector $\mathbf{x}^i$ is then assigned a *fuzzy label* $\mathbf{u}^i \in [0,1]^c$, whose components are the grades of membership of that pattern to each class. Each membership value u_k^i is generally interpreted as reflecting the typicality of pattern $\mathbf{x}^i$ with regard to class k, as measured by its similarity to certain prototypes. Such labels are often assumed to define a fuzzy partition [1, 14] of the training set, i.e., it is assumed that

$$\sum_{k=1}^{c} u_k^i = 1 \quad , \forall i. \tag{1}$$

In particular, the above property naturally arises when class labels are generated by some clustering algorithm such as the fuzzy c-means [2].

The problem considered in this paper is related to, but different from the previous one. In the type of applications envisaged here, classes are well-defined and each training pattern has full membership to a single class. However, *this true classification is only partially known*, for lack of complete and accurate information regarding the training set. Such a situation typically arises when learning examples are labeled *a posteriori* by one or several experts, whose subjective evaluation of pattern categories may be tainted with imprecision and uncertainty. Consider, for instance, the problem of recognizing certain transient phenomena (such as, e.g., K-complexes or delta waves) in electroencephalogram (EEG) data [15, 16]. Such phenomena are usually very difficult to detect because they may have a variety of shapes and are not always distinct from EEG background activity. For that reason, even a trained physician may not always be able to recognize certain classes of patterns with full certainty. He or she may, however, be able to assess the "likelihood", or "possibility", that a certain phenomenon is present in the data. An alternative approach to obtain such an assessment is to collect the opinions of several experts, and consider for each example the empirical distribution of expert opinions about the its class membership.

In this paper, we propose to adopt the formalism of Possibility Theory to represent such uncertainty about the class membership of training patterns. Each pattern will then be assumed to be assigned a *possibilistic labels* of the form

$$\mathbf{u}^i \triangleq (u_1^i, \ldots, u_c^i) \in [0,1]^c \tag{2}$$

where u_k^i represents the *degree of possibility* that pattern i belongs to class k. Vector $\mathbf{u}^i$ thus defines a possibility distribution over the set Ω of classes or, equivalently, the fuzzy subset of "possible" classes for pattern i. In most cases, it is reasonable

to assume that the membership of each pattern to at least one class is completely possible, which translates to the condition

$$\max_k u_k^i = 1 \quad , \forall i.$$

However, this condition may be relaxed under the open-world assumption, i.e., when it is recognized that some classes may have been omitted from the reference set Ω, so that a pattern may actually belong to none of the enumerated categories.

Possibilistic labels may be obtained in several different ways, including:

- direct elicitation from an expert, who is asked to quantify by a real number between 0 and 1 the degree of possibility that case i belongs to each of the c classes, or
- from an empirical distribution of expert opinions, using possibilistic histograms [7]. If q_k^i denotes the number of experts (out of q) who assigned pattern i to class k, we have, according to Ref. [7] :

$$u_k^i = \frac{1}{q} \sum_{j=1}^{c} \min(q_k^i, q_j^i). \tag{3}$$

A simpler, but related problem was addressed in [3, 4, 40], in which each training pattern was assumed to be labeled by a *crisp* set of categories. A solution to this problem was proposed based on the Dempster-Shafer theory of evidence [17, 27]. A classification scheme was described, in which each training pattern (with crisp set-valued class label) is treated as a piece of evidence and, as such, induces a belief function regarding the class membership of each new pattern $\mathbf{x}$ to be classified. The evidence of training patterns in a neighborhood of $\mathbf{x}$ is then pooled by means of the conjunctive sum operation, resulting in a final belief function which can be used to support decision making based on arbitrary utilities.

In this paper, the above classification scheme is extended to allow one's knowledge regarding the class of training patterns to be described by possibilistic class labels such as (2). Two directions will be successively investigated. In the first approach, the close connection between fuzzy sets, possibility distributions and consonant belief structures [6] is exploited to transform each fuzzy set

$$\widetilde{L}^i \triangleq \{\frac{u_1^i}{1}, \ldots, \frac{u_c^i}{c}\} \tag{4}$$

containing the "possible classes" for training pattern $\mathbf{x}^i$ into a consonant belief structure whose focal elements are the α-cuts of $\widetilde{L}^i$. The second approach uses a generalization of the Dempster-Shafer theory to fuzzy sets [39, 30, 35], allowing to manipulate belief structures with fuzzy focal elements.

The rest of this paper is organized as follows. First of all, the background on standard (crisp) Evidence Theory, as well as our approach to evidential pattern classification are briefly recalled in Section 2. Sections 3 and 4 then present the two extensions of this method to possibilistic class labels. Numerical examples are then presented in Section 5, and Section 6 concludes the paper.

2 Background

2.1 Evidence theory

This aim of this section is to clarify the notation and terminology of Evidence Theory, as used in the rest of the paper. The interested reader is referred to, e.g., [17, 22, 27, 26, 33] for mathematical developments and in-depth discussion on possible interpretations of the theory. The subjectivist, non probabilistic view of Smets' Transferable Belief Model (TBM) [27, 26] will be adopted here.

Let Ω denote a finite set of possible answers to a certain question, and y a variable describing the correct (but unknown) answer. The beliefs held at a certain time by a rational agent, regarding the value taken by y, are assumed to be properly described by a *belief structure* (BS), also called a basic belief assignment or a mass function, defined as a function m from 2^Ω to $[0,1]$, verifying:

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{5}$$

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. A belief structure m such that $m(\emptyset) = 0$ is said to be normal. This condition need not be imposed if one interprets the mass $m(\emptyset)$ as quantifying one's belief that $y \notin \Omega$ (open-world assumption, [20]). However, if Ω is assumed to be exhaustive (closed-world assumption), then this condition should be imposed, which is the situation initially considered by Shafer [17].

Assuming one's state of belief to be represented by m, it is possible to measure one's *total belief* in the proposition $\omega \in A$, for any $A \subseteq \Omega$, by a number, called the *credibility* of A, and defined as:

$$\mathrm{bel}(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B). \tag{6}$$

Mathematically, function $\mathrm{bel} : 2^\Omega \mapsto [0,1]$, called a *belief function*, may be shown to be a Choquet capacity of order infinite [17]. The use of such non additive measures for representing subjective degrees of belief has recently received an axiomatic justification [24]. Closely related to the notion of belief function is that of *plausibility function*, defined for each $A \subseteq \Omega$ as:

$$\mathrm{pl}(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B) \tag{7}$$

$$= \mathrm{bel}(\Omega) - \mathrm{bel}(\overline{A}) \tag{8}$$

where $\overline{A}$ denotes the complement of A. The quantity $\mathrm{pl}(A)$ may be interpreted as the degree of belief that could *potentially* be given to A, if further evidence became available [27]. The three functions m, bel and pl are in one-to-one correspondence, and actually constitute three facets of the same information [17].

The dynamic part of the Dempster-Shafer model indicates how pieces of evidence coming from two distinct sources should be combined. Let m_1 and m_2 denote the BSs induced by each of these pieces of evidence considered individually. Then, if

both sources are known to be reliable, m_1 and m_2 may be combined conjunctively by defining a new BS $m_1 \cap m_2$ as:

$$(m_1 \cap m_2)(A) \triangleq \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega. \tag{9}$$

Note that this operation may produce a subnormal BS, i.e., it is possible to have $(m_1 \cap m_2)(\emptyset) > 0$. Under the closed-world assumption, some kind of normalization thus has to be performed. The *Dempster normalization procedure* converts a subnormal BS m into a normal BS m^* defined by:

$$m^*(A) \triangleq \begin{cases} \dfrac{m(A)}{1 - m(\emptyset)} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset. \end{cases} \tag{10}$$

The conjunctive sum operation followed by Dempster normalization is the *orthogonal sum* operation (also called Demspter's rule of combination) initially studied by Shafer [17]. Note that other operations on BS's may be defined, based on different assumptions regarding the reliability of the information sources [23].

A final issue of fundamental importance for pattern recognition applications of Evidence Theory is that of decision analysis [4]. Several approaches to this problem have been proposed, based, e.g., on upper and lower expected utilities [4] or on a "pessimism index" such as the Hurwicz criterion [13, 28]. Following the TBM, the approach adopted in this paper is based on the concept of pignistic transformation [21]: for decision making purposes, a belief function (representing the agent's beliefs) is transformed into a *pignistic* probability function, defined for all $A \subseteq \Omega$ as:

$$\mathrm{BetP}(A) \triangleq \sum_{B \subseteq \Omega, B \neq \emptyset} m^*(B) \frac{|A \cap B|}{|B|}, \tag{11}$$

where m^* is the *normalized* form of m according to the Dempster procedure (10). In a decision problem involving a set of actions whose consequences are quantified by utilities, the model then prescribes the action entailing the maximum expected utility, relative to the pignistic probability function.

2.2 Evidential classification

This section presents the principles of the *evidential distance-based classifier* (EDC) introduced in [3] and refined in [4, 40]. The generalization of this method to possibilistic class labels, which is the main topic of this paper, will be considered at length in Section 3 and 4.

We consider a classification problem involving c classes. The set of classes is denoted $\Omega = \{1, \ldots, c\}$. The available information is assumed to consist in a training set $\mathcal{T}$ composed of n examples of the form $e^i \triangleq (\mathbf{x}^i, \mathbf{u}^i)$, where $\mathbf{x}^i$ denotes a vector in $\mathbb{R}^d$ describing some entity of interest, and $\mathbf{u}^i$ a vector of binary indicator variables indicating possible classes for that entity, i.e., $u_k^i = 1$ if it is possible that entity i belongs to class k, and $u_k^i = 0$ otherwise. We denote by L^i the set of possible classes for pattern i, i.e.,

$$L^i = \{k \in \Omega, u_k^i = 1\}.$$

It is a crisp subset of Ω known to contain the true class of pattern $\mathbf{x}^i$.

Let us now consider a new vector $\mathbf{x}$, which we wish to classify, based on the training set information. Each of the stored examples e^i may be regarded as a piece of evidence inducing certain beliefs regarding the class of $\mathbf{x}$. Obviously, if one learns that $\mathbf{x}^i$ is close to $\mathbf{x}$ in feature space, and $L^i = A$, it becomes more likely that the class of $\mathbf{x}$ is also contained in A. In the framework of evidence theory, such a belief may be modeled by a BS $m(\cdot|e^i)$, focused on A and Ω (and on no other subset, since the closeness of $\mathbf{x}$ to $\mathbf{x}^i$ does not point to any other hypothesis). Furthermore, the mass of belief assigned to A can reasonably be assumed to be a decreasing function of the dissimilarity (according to some relevant measure δ) between $\mathbf{x}$ and $\mathbf{x}^i$. We then arrive at the following expression for $m(\cdot|e^i)$:

$$m(A|e^i) \triangleq \begin{cases} \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)] & \text{if } A = L^i \\ 1 - \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)] & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where ϕ_θ is a decreasing function, depending on some parameter θ, verifying $\phi_\theta(0) \leq 1$ and $\lim_{\delta \to \infty} \phi_\theta(\delta) = 0$ [3]. When δ denotes the Euclidean distance, a rational choice for ϕ_θ was shown in [5] to be:

$$\phi_\theta(\delta) = \theta_1 e^{-\theta_2 \delta^2} \tag{13}$$

where $\theta = (\theta_1, \theta_2)^t$ is a two-dimensional parameter vector verifying $0 < \theta_1 \leq 1$ and $\theta_2 > 0$.

Regarding the n learning examples as "distinct" items of evidence, in the sense of Shafer [17], the corresponding BSs should be combined conjunctively, leading to the following global BS:

$$m \triangleq m(\cdot|e^1) \cap \ldots \cap m(\cdot|e^n). \tag{14}$$

Function m thus provides a description of one's belief regarding the class of $\mathbf{x}$, after considering the whole training set. Note that, for an example e^i with unknown classification, $m(\cdot|e^i)$ is the vacuous BS, and has therefore no influence of m, as it should be. The computational efficiency of this procedure may be significantly improved by taking into account only the K nearest neighbors of $\mathbf{x}$ in feature space (or, alternatively, the training patterns situated within a certain distance from $\mathbf{x}$) [3].

Finally, a decision regarding the assignment of $\mathbf{x}$ to a class, or its rejection, may be made by considering the pignistic probability function BetP induced by m, and applying Bayes decision theory. In one of the simplest cases, the $c + 1$ actions to be considered are the assignment α_k to class k, $k = 1, \ldots, c$, and rejection α_0 [4]. Assuming $\{0, 1\}$ losses, and denoting by λ_0 the fixed rejection loss, the risks are then:

$$\begin{aligned} R(\alpha_k) &= 1 - \text{BetP}(\{k\}), \quad k = 1, \ldots, c \\ R(\alpha_0) &= \lambda_0, \end{aligned}$$

leading to the assignment to the class with highest pignistic probability, provided this probability exceeds $1 - \lambda_0$. This classification scheme is illustrated by the following example.

EXAMPLE 1 To illustrate the above classification scheme, let us consider the following simple three-class problem, in which the training set consists of only three one-dimensional patterns $x^1 = 0$, $x^2 = 2$ and $x^3 = 3$ with class labels:

$$\begin{aligned} u^1 &= (1, 1, 0) \\ u^2 &= (0, 1, 1) \\ u^3 &= (0, 0, 1). \end{aligned}$$

Therefore, example 1 is known to belong to class 1 *or* 2, example 2 belongs to class 2 *or* 3, whereas example 3 belongs for sure to class 3. Leaving aside the parameter determination problem for the moment, let us assume that parameter θ has been set to the following value

$$\theta = (0.9, 1).$$

What is our belief concerning the class membership of a new sample $x = 1.2$, and what would be our decision regarding the assignment of that pattern to a class ? The answers to these questions are shown in Table 1. Belief structures $m(\cdot|e^i), i = 1, 2, 3$ induced by each of the three learning samples are computed using (12) and (13); the conjunctive sum (14) of these BS results in a global BS m, which after normalization yields a pignistic probability function BetP computed from (11). Assuming $\{0, 1\}$ losses [4], a decision rule without rejection would therefore assign that pattern to class 2. The pignistic probabilities computed for the three classes are shown as a function of x in Figure 1. □

The whole classification scheme described above depends on parameter θ. In [40], it was proposed to learn this parameter from the data by minimizing an error function. In the special case where all class labels are singletons (i.e., the class membership of all training patterns is known with certainty), this error function was defined as the mean squared error the between pignistic probabilities and the 0-1 class membership indicator variables :

$$E_{MS}(\theta) \triangleq \sum_{i=1}^{n} \sum_{k=1}^{c} \left[\mathrm{BetP}^{(i)}(\{k\}) - u_k^i \right]^2, \tag{15}$$

where $u_k^i \in \{0, 1\}$ ($u_k^i = 1$ if example i belongs to class k, and $u_k^i = 0$ otherwise), and $\mathrm{BetP}^{(i)}$ is the pignistic probability function computed for example i, using all other training patterns. Detailed numerical experiments, presented in [40], demonstrated the very good performance of the optimized evidential classification rule, as compared to several other methods such as the voting, distance-weighted and fuzzy k-NN rules.

In the more general case of arbitrary crisp class labels, however, the above error criterion cannot be blindly generalized for the following reason. Assume that we have initially a training set of size n, and a new example $e^{n+1} = (\mathbf{x}^{n+1}, L^{n+1})$ is added, with $L^{n+1} = \Omega$, and consequently $u_k^{n+1} = 1$ for all $k \in \{1, \ldots, c\}$. Its class membership being totally unknown, such a pattern obviously contains no useful information for the classification of other patterns. However, its addition to the training set modifies the error function (15), and hence the optimal parameter value.

To correct this defect of the mean squared error in the case of arbitrary class labels, we introduce the following new error function:

$$E(\theta) \triangleq n - \sum_{i=1}^{n} \mathrm{BetP}^{(i)}(L^i). \tag{16}$$

This criterion has the following two desirable properties:

- the addition of a new pattern e^{n+1} with unknown class label ($L^{n+1} = \Omega$) leaves $E(\theta)$ unchanged, and therefore does not affect the optimal parameter value;
- $E(\theta)$ is equal to the number of misclassifications if the class labels L^i are singletons and $\mathrm{BetP}^{(i)}(\{k\}) \in \{0, 1\}$ for all i and k; $E(\theta)/n$ is then the leave-one-out error rate, which makes $E(\theta)$ easy to interpret in the general case.

3 Generalization based on consonant BS's

In this and the following section, we tackle a more general class of learning problems, in which training patterns are assigned possibilistic class labels. Each training example will be assumed to be of the form

$$\widetilde{e}^i \triangleq (\mathbf{x}^i, \mathbf{u}^i) \tag{17}$$

where $\mathbf{u}^i$ is a vector of continuous variables in $[0, 1]$, indicating the degrees of possibility (measured on a continuous scale) that pattern i belongs to each of c classes. The corresponding set of possible classes for pattern i is therefore a fuzzy subset $\widetilde{L}^i$ of Ω, defined by (4). As explained in Section 1, possibilistic class labels may be obtained by direct elicitation from experts (e.g., by asking such questions as: "What is the degree of possibility that pattern i belongs to class k?"), or they may be computed from multiple expert opinions using (3).

In this section, it is proposed to transform each possibilistic label $\mathbf{u}^i$ into a BS m^i, and to generalize (12) accordingly. This approach is based on the connection between the concepts of possibility and belief measures, as explained in the sequel.

3.1 BS induced by a fuzzy set

Let $\widetilde{F}$ be a fuzzy subset of a referential Ω, and π the corresponding possibility distribution [38]. These two entities are related by the following equalities:

$$\pi(\omega) = \widetilde{F}(\omega) \quad \forall \omega \in \Omega$$

where $\widetilde{F}(\omega)$ denotes the degree of membership of ω to F. Each number $\pi(\omega)$ is interpreted as a degree of possibility that y (the variable of interest) is equal ω, based on fuzzy proposition "y is $\widetilde{F}$". The associated possibility measure Π is then defined for any arbitrary crisp subset A of Ω as:

$$\Pi(A) \triangleq \max_{\omega \in A} \pi(\omega),$$

while the corresponding necessity measure N is given by [10, 9]:

$$N(A) = \Pi(\Omega) - \Pi(\overline{A}).$$

As pointed out by Dubois and Prade [6], N is formally identical to what Shafer [17] called a *consonant belief function*, i.e., it is a belief function with nested focal elements, Π being the associated plausibility measure. The focal elements of N are the α-cuts of $\widetilde{F}$, and the corresponding belief structure $m_{\widetilde{F}}$ is defined as follows. Let $\pi_1 > \ldots > \pi_r$ be the distinct values taken by π, arranged in decreasing order, and $\pi_{r+1} = 0$ by convention. Let A_i denote the π_i-cut of $\widetilde{F}$. Then, we have, for any non-empty subset A of Ω:

$$m_{\widetilde{F}}(A) = \begin{cases} \pi_i - \pi_{i+1} & \text{if } A = A_i, i = 1, \ldots, r \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Note that

$$\sum_{A \subseteq \Omega, A \neq \emptyset} m_{\widetilde{F}}(A) = \pi_1.$$

Hence, a subnormal fuzzy set $\widetilde{F}$ induces a subnormal BS $m_{\widetilde{F}}$, with

$$m_{\widetilde{F}}(\emptyset) = 1 - \pi_1.$$

EXAMPLE 2 Let $\Omega = \{1, 2, 3, 4\}$, and $\widetilde{F} = \{\frac{0.1}{1}, \frac{0.5}{2}, \frac{0.9}{3}, \frac{0.2}{4}\}$. Application of the above formula leads to:

$$\begin{aligned} m_{\widetilde{F}}(\{3\}) &= 0.9 - 0.5 = 0.4 \\ m_{\widetilde{F}}(\{2,3\}) &= 0.5 - 0.2 = 0.3 \\ m_{\widetilde{F}}(\{2,3,4\}) &= 0.2 - 0.1 = 0.1 \\ m_{\widetilde{F}}(\Omega) &= 0.1 \\ m_{\widetilde{F}}(\emptyset) &= 1 - 0.9 = 0.1 \end{aligned}$$

Note that a subnormal BS is obtained ($m_{\widetilde{F}}(\emptyset) > 0$), as a result of the subnormality of $\widetilde{F}$. □

3.2 Application

The above approach to the construction of a consonant BS from a fuzzy set allows to reformulate each learning example of the form $\widetilde{e}^i = (\mathbf{x}^i, \widetilde{L}^i)$ as $\widetilde{e}^i = (\mathbf{x}^i, m^i)$, where $m^i = m_{\widetilde{L}^i}$ is the consonant BS induced by $\widetilde{L}^i$. This BS may be interpreted as quantifying one's belief regarding the class of the entity described by vector $\mathbf{x}^i$. This method of handling possibilistic labels will later be referred to as EDCP-I (evidential distance-based classifier with possibilistic class labels, method I).

Let $\mathbf{x}$ be a feature vector describing a new entity to be classified, and let us first assume that $\mathbf{x} = \mathbf{x}^i$. Then, if one is asked to quantity one's belief regarding the class of $\mathbf{x}$, based on information $\widetilde{e}^i$ alone, it is natural to write:

$$m(\cdot|\widetilde{e}^i) = m^i,$$

i.e., one should hold the same beliefs regarding the class membership of two entities with exactly the same description. In the more general situation in which $\mathbf{x} \neq \mathbf{x}^i$,

the information brought by $\widetilde{e}^i$ should be regarded as less relevant as the dissimilarity between vectors $\mathbf{x}$ and $\mathbf{x}^i$ increases. This may be modeled by *discounting* m^i, the discount rate α being an increasing function of the distance between the two vectors. Setting

$$\alpha = 1 - \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)],$$

where, as before, δ is a dissimilarity measure and ϕ_θ a decreasing function taking valued between 0 and 1, we than have:

$$m(A|\widetilde{e}^i) \triangleq \begin{cases} \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)] m^i(A) & \text{if } A \neq \Omega \\ \big(1 - \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)]\big) \sum_{B \neq \Omega} m^i(B) & \text{if } A = \Omega \end{cases}, \tag{19}$$

which may be regarded as a generalization of (12) if the set-valued class label L^i with $L^i = A$ is interpreted as a BS m^i such that $m^i(A) = 1$.

EXAMPLE 3 Let us return to the simple classification problem of Example 1, but let us now assume the class labels of patterns x^1, x^2 and x^3 to be

$$\begin{aligned} u^1 &= (1, 0.8, 0.1) \\ u^2 &= (0.1, 1, 0.8) \\ u^3 &= (0, 0.2, 1). \end{aligned}$$

These labels are interpreted as describing possibility distributions: for instance, it is completely possible that pattern x^1 belongs to class 1 ($u^1_1 = 1$), whereas the degrees of possibility that it belongs to classes 2 and 3 are 0.8 and 0.1, respectively. Table 2 shows

- the corresponding BS's $m^i, i = 1, 2, 3$,
- the induced BS's $m(\cdot|\widetilde{e}^i)$ computed from (19) for $x = 1.2$,
- the conjunctive sum $m = \bigcap_{i=1}^3 m(\cdot|\widetilde{e}^i)$,
- the final pignistic probability function BetP.

The pignistic probabilities are represented as a function of x in Figure 2. □

As before, parameter θ should be determined by minimizing an error function. For that purpose, the criterion defined by (16) may be generalized as:

$$E(\theta) \triangleq n - \sum_{i=1}^{n} \mathrm{BetP}^{(i)}(m^i), \tag{20}$$

where $\mathrm{BetP}^{(i)}(m^i)$ is defined as

$$\mathrm{BetP}^{(i)}(m^i) \triangleq \sum_{A \subseteq \Omega} m^i(A) \mathrm{BetP}^{(i)}(A).$$

Note that we have

$$\begin{aligned}
\mathrm{BetP}^{(i)}(m^i) &= \sum_{A\subseteq\Omega} m^i(A) \sum_{\omega\in A} \mathrm{BetP}^{(i)}(\omega) \\
&= \sum_{\omega\in\Omega} \mathrm{BetP}^{(i)}(\omega) \sum_{A\ni\omega} m(A) \\
&= \sum_{\omega\in\Omega} \mathrm{BetP}^{(i)}(\omega)\mathrm{pl}(\{\omega\}) \\
&= \sum_{\omega\in\Omega} \mathrm{BetP}^{(i)}(\omega)\widetilde{L}^i(\omega),
\end{aligned}$$

which shows that $\mathrm{BetP}^{(i)}(m^i)$ is equal to $\mathrm{BetP}^{(i)}(\widetilde{L}^i)$, the pignistic probability of fuzzy event $\widetilde{L}^i$, defined as the expectation of its membership function [37]. Note also that (16) is recovered as a special case when each m^i has a single focal element.

4 Generalization based on fuzzy BS's

This section presents an alternative approach to the above problem, based on a generalization of Evidence theory allowing to assign degrees of belief to fuzzy events. The underlying concepts will first be recalled, and then applied to the classification problem.

4.1 Fuzzy belief structures

An alternative method for handling pieces of evidence of the form given by (17) is based on the concept of *fuzzy belief structure* first introduced by Zadeh [39], and subsequently investigated and developed by several authors [18, 30, 12, 8, 35, 32].

Basically, a fuzzy BS defines the assignment of a unit mass of belief to a finite number of fuzzy focal elements. Formally, it is a mapping m from $[0,1]^\Omega$, the set of fuzzy subsets of Ω, to $[0,1]$, verifying the following two conditions:

$$m(\widetilde{F}) > 0 \Leftrightarrow \widetilde{F} \in \mathcal{F}(m) = \{\widetilde{F}_1, \ldots, \widetilde{F}_n\} \subset [0,1]^\Omega \tag{21}$$

$$\sum_{i=1}^{n} m(\widetilde{F}_i) = 1 \tag{22}$$

If, furthermore, the $\widetilde{F}_i$ are normal fuzzy sets, then m is said to be normal (this generalizes the normality condition $m(\emptyset) = 0$ defined for crisp BS's).

For any focal element $\widetilde{F}_i$ and any crisp or fuzzy subset A of Ω, the conditional possibility measure A given that y is $\widetilde{F}_i$ is [39]:

$$\Pi(A|\widetilde{F}_i) \triangleq \max_{\omega\in\Omega} A(\omega) \wedge \widetilde{F}_i(\omega). \tag{23}$$

The expected possibility of A may then be defined as:

$$\mathbb{E}\Pi(A) = \sum_{i=1}^{n} m(\widetilde{F}_i)\Pi(A|\widetilde{F}_i). \tag{24}$$

In the special case where both A and the $\widetilde{F}_i$ are crisp, the above quantity is equal to the plausibility of A, which allowed Zadeh [39] to propose (24) as a general definition for the plausibility of a fuzzy event A, induced by a fuzzy BS. Note that $\Pi(A|\widetilde{F}_i)$ may also be interpreted as a degree of intersection between fuzzy sets A and $\widetilde{F}_i$, which clarifies the relationship with (7). When the focal elements of m are crisp but A is fuzzy, then the expected possibility of A defined by (24) may also be viewed as the upper expectation $\mathbb{E}^*_m(A)$ of the membership function of A with respect to m:

$$\mathbb{E}\Pi(A) = \sum_{i=1}^{n} m(F_i) \max_{\omega \in F_i} A(\omega) = \mathbb{E}^*_m(A), \tag{25}$$

which was proposed by Smets as the definition of the plausibility of a fuzzy event induced by a crisp belief structure [18]. Note that this definition generalizes Zadeh's definition of the probability of a fuzzy event proposed in [37].

Similarly, the conditional necessity measure of A given $\widetilde{F}_i$ may be defined (taking into account the possible subnormality of $\widetilde{F}_i$) as [9]:

$$N(A|\widetilde{F}_i) \triangleq \Pi(\Omega|\widetilde{F}_i) - \Pi(\overline{A}|\widetilde{F}_i).$$

The expected necessity:

$$\mathbb{E}N(A) = \sum_{i=1}^{n} m(\widetilde{F}_i) N(A|\widetilde{F}_i) \tag{26}$$

is then a generalization of (6) for the belief in fuzzy event A, based on fuzzy BS m [39]. In the special case where the focal elements of m are crisp and A is fuzzy, $\mathbb{E}N(A)$ defined by (26) can be interpreted as the lower expectation $\mathbb{E}_{*m}(A)$ of the membership function of A with respect to m:

$$\mathbb{E}N(A) = \sum_{i=1}^{n} m(F_i) \min_{\omega \in F_i} A(\omega) = \mathbb{E}_{*m}(A), \tag{27}$$

which is Smets' definition for the belief in a fuzzy event induced by a crisp belief structure [18].

The extensions of the concepts of plausibility and belief provided by (24) and (26) seem "natural" but they are by no means unique. For instance, the min and max operations occurring in (23) may be replaced by any other pair of a t-norm and a t-conorm, as suggested by Yager [30]. Following a different path, Dubois and Prade [8] proposed to treat each fuzzy focal element $\widetilde{F}_i$ of a fuzzy BS m as a possibility distribution associated with a crisp consonant BS m_i. This leads to the following alternative definitions for the plausibility and degree of belief of a fuzzy event A relative to a fuzzy BS m:

$$\mathrm{pl}_S(A) \triangleq \sum_{i=1}^{n} m(\widetilde{F}_i) \mathbb{E}^*_{m_i}(A) \tag{28}$$

$$\mathrm{bel}_S(A) \triangleq \sum_{i=1}^{n} m(\widetilde{F}_i) \mathbb{E}_{*m_i}(A), \tag{29}$$

where $\mathbb{E}^*_{m_i}(A)$ and $\mathbb{E}_{*m_i}(A)$ denote, respectively, the upper and lower expectations of the memebership function of A, with respect to BS m_i. Some properties of these extensions are discussed in Ref. [8]. The same expressions were independently obtained by Yen [35, 36], starting from the concept of a fuzzy compatibility relation between an evidence space and the hypothesis space Ω.

As proposed by Yen [35] and Yager [31, 34], the conjunctive sum operation defined by (9) may also be extended to fuzzy BS's by replacing crisp intersection by its fuzzy counterpart:

$$(m_1 \cap m_2)(A) \triangleq \sum_{\{\widetilde{F} \in \mathcal{F}(m_1), \widetilde{F}' \in \mathcal{F}(m_2) | \widetilde{F} \cap \widetilde{F}' = A\}} m_1(\widetilde{F}) m_2(\widetilde{F}') \quad \forall A \in [0,1]^\Omega, \tag{30}$$

where $\cap$ now denotes standard fuzzy intersection.

Note that the possible subnormality of $m = m_1 \cap m_2$ may be corrected (if needed) by applying Yager's *soft normalization* procedure [32]:

$$m^*(A) \triangleq \frac{\sum_{\widetilde{F}^* = A} h_{\widetilde{F}} m(\widetilde{F})}{\sum_{\widetilde{F} \in \mathcal{F}(m)} h_{\widetilde{F}} m(\widetilde{F})} \tag{31}$$

where $h_{\widetilde{F}} = \sup \widetilde{F}$ denotes the height of $\widetilde{F}$, $\widetilde{F}^*$ is the normal fuzzy set defined by $\widetilde{F}^*(\omega) = \widetilde{F}(\omega)/h_{\widetilde{F}}$ for all $\omega \in \Omega$, and $\mathcal{F}(m)$ is the set of focal elements of m.

4.2 Application

The possibility provided by fuzzy BS's to assign degrees of support to fuzzy propositions leads to a very simple generalization of (12), in which the consideration of each learning example $\widetilde{e}^i = (\mathbf{x}^i, \widetilde{L}^i)$ induces a fuzzy BS $\widetilde{m}(\cdot|\widetilde{e}^i)$ defined for all $A \in [0,1]^\Omega$ by

$$\widetilde{m}(A|\widetilde{e}^i) \triangleq \begin{cases} \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)] & \text{if } A = \widetilde{L}^i \\ 1 - \phi_\theta[\delta(\mathbf{x}, \mathbf{x}^i)] & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases} . \tag{32}$$

The fuzzy BS's $\widetilde{m}(\cdot|\widetilde{e}^i)$ induced by each of the n learning examples may then be combined using the generalized conjunctive sum operation (30). This leads to an alternative approach of handling possibilistic labels, later referred to as the EDCP-II method.

Having summarized the whole evidence in the form of a fuzzy BS

$$\widetilde{m} = \widetilde{m}(\cdot|\widetilde{e}^1) \cap \ldots \cap \widetilde{m}(\cdot|\widetilde{e}^n),$$

some decision strategy has to be applied. To be consistent with the crisp case, such a strategy must rely on some generalization of the pignistic probability measure. We propose to generalize (11) as:

$$\mathrm{BetP}(\widetilde{A}) \triangleq \sum_{\widetilde{F} \in \mathcal{F}(\widetilde{m}^*)} \widetilde{m}^*(\widetilde{F}) \frac{|\widetilde{A} \cap \widetilde{F}|}{|\widetilde{F}|}, \tag{33}$$

where $\widetilde{A}$ is an arbitrary crisp or fuzzy subset of Ω, $|\cdot|$ denotes scalar cardinality (sigma-count), $\cap$ represents standard fuzzy set intersection, and $\widetilde{m}^*$ is the normalized form of $\widetilde{m}$ according to (31). The above equation reduces in the case of a singleton to

$$\mathrm{BetP}(\{k\}) = \sum_{\widetilde{F} \in \mathcal{F}(\widetilde{m}^*)} \widetilde{m}^*(\widetilde{F}) \frac{\widetilde{F}(k)}{|\widetilde{F}|}. \tag{34}$$

Note that we have obviously $\mathrm{BetP}(\Omega) = 1$. Furthermore, given two fuzzy events $\widetilde{A}$ and $\widetilde{B}$ such that $\widetilde{A} \cap \widetilde{B} = \emptyset$, we have:

$$\begin{aligned}
\mathrm{BetP}(\widetilde{A} \cup \widetilde{B}) &= \sum_{\widetilde{F} \in \mathcal{F}(\widetilde{m}^*)} \widetilde{m}^*(\widetilde{F}) \frac{|(\widetilde{A} \cup \widetilde{B}) \cap \widetilde{F}|}{|\widetilde{F}|} \\
&= \sum_{\widetilde{F} \in \mathcal{F}(\widetilde{m}^*)} \widetilde{m}^*(\widetilde{F}) \frac{|(\widetilde{A} \cap \widetilde{F}) \cup (\widetilde{B} \cap \widetilde{F})|}{|\widetilde{F}|} \\
&= \sum_{\widetilde{F} \in \mathcal{F}(\widetilde{m}^*)} \widetilde{m}^*(\widetilde{F}) \frac{|\widetilde{A} \cap \widetilde{F}| + |\widetilde{B} \cap \widetilde{F}|}{|\widetilde{F}|} = \mathrm{BetP}(\widetilde{A}) + \mathrm{BetP}(\widetilde{B}),
\end{aligned}$$

which proves the additivity of BetP.

For the optimization of θ in (32), the error function given by Eq. (20) may again be used, with the pignistic probabilities $\mathrm{BetP}^{(i)}$ now defined from the fuzzy BS's using (33).

EXAMPLE 4 Let us consider again the data and classification problem of Example 3. Each training sample x^i is now assumed to induce a fuzzy BS $\widetilde{m}(\cdot|\widetilde{e}^i)$ according to (32). These fuzzy BSs, together with their conjunctive sum $\widetilde{m}$ are shown in Table 3. The pignistic probabilities of each class computed from (33) are

$$\mathrm{BetP}(\{1\}) = 0.22 \quad \mathrm{BetP}(\{2\}) = 0.45 \quad \mathrm{BetP}(\{3\}) = 0.33.$$

These quantities are shown as a function of x in Figure 3. As can be seen from this figure, both methods of handling possibilistic labels yield very close results. □

5 Numerical experiments

In the type of learning problems considered in this paper, the class membership of training patterns is only partially known, and described by a possibility distribution over the set of classes. Our approach will now be illustrated by the following two experiments based on synthetic and real data.

5.1 Synthetic data

Let us consider a c-class problem, in which each class k has a conditional probability density $f(\mathbf{x}|k)$. Let $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ be a set of vectors with unknown class membership. The "expert" does not know the true class label of any of these vectors, but he has

seen (infinitely) many examples and is able to assess the *likelihood* of each class for each of the examples, through:

$$L(k|\mathbf{x}^i) = f(\mathbf{x}^i|k).$$

As argued by Smets [19] [25, p. 289] and Walley and Moral [29], among others, the likelihood $L(k|\mathbf{x}^i)$ may be interpreted, after normalization, as the degree of possibility that the actual class is k. This closeness between the concepts of likelihood and possibility allows us to define a possibility distribution $(u_1^i, \ldots, u_c^i)$ from the relative likelihoods:

$$u_k^i = \frac{L(k|\mathbf{x}^i)}{\max_{j=1,c} L(j|\mathbf{x}^i)}.$$

The "most possible" class for pattern i is then the maximum likelihood (ML) estimate $\widehat{k}^*$ verifying

$$L(\widehat{k}^*|x^i) > L(k|x^i) \quad \forall k \neq k^*.$$

It is therefore possible to associate to each training pattern three different class labels: the true class k^*, the most possible class (ML estimate) $\widehat{k}^*$, and the whole possibility distribution $\mathbf{u}^i = (u_1^i, \ldots, u_c^i)$.

The performances of various classification algorithms may then be compared, based on the corresponding three different kinds of training information:

- "conventional" classification methods with crisp labeling of training data (either based on true labels or estimated ones);
- the EDCP-I and EDCP-II methods applied to data with possibilistic labels.

Note that, when the number n of training samples tends to infinity, all three types of labeling fully determine the class-conditional probability distributions and, consequently, essentially carry the same information. Therefore, any significant difference between them is more likely to be observed for small sample size.

As an example of such an approach, we considered three Gaussian classes with following means μ_k and variance matrices Σ_k:

$$\begin{array}{lll} \mu_1 = (-1, 2)^t & \mu_2 = (0, 0)^t & \mu_3 = (3, 1)^t \\ \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} & \Sigma_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix} & \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \end{array}$$

For each trial, a training data sets of 6 vectors (2 per class) and a test set of 500 vectors (with equal prior probabilities) were generated. Five classification procedures were tested: (a) Dudani's distance-weighted 5-nearest neighbor rule [11], (b) Keller's fuzzy 5-nearest neighbor rule [14], (c) our EDC method with optimized parameters [40] and (d) methods EDCP-I and II introduced in this paper. For methods (a), (b) and (c), both kinds of crisp labeling (real and estimated) were considered. A typical training set with possibilistic class labels is shown in Figure 4.

Results from 20 trials are summarized in Table 4 and in Figure 5. For this problem, all three methods based on crisp class labels (either real, or estimated) happen to be more or less equivalent, whereas methods I and II appear to make effective use of the additional information provided by possibilistic labels.

5.2 Real data

The methods introduced in this paper were applied to the problem of detecting K-complexes in sleep EEG, using the data described in [15, 16]. The K-complex is a transient EEG pattern which plays a major role in sleep stages assessment. It has a duration of between 500 and 1500 ms, and is characterized by a sharp upward wave followed by a downward one. Its amplitude is three times background activity [16]. The discrimination of K-complexes from background activity is generally recognized as a very complex pattern recognition problem.

The data used in this experiment consisted of EEG signals encoded as 64-dimensional patterns. Some of these signals were negative examples containing paroxysmal delta bursts, a phenomenon bearing some resemblance to K-complexes. The other signals consisted of patterns which, after visual inspection by 5 physicians, had been classified as containing a K-complex by at least one of them. Among these examples, those categorized in the K-complex class by a majority of experts were considered as positive examples, the others as negative ones. Each example (positive or negative) was then assigned a possibilistic label $\mathbf{u}^i = (u_1^i, u_2^i)$ based on the number of experts (between 0 and 5) who recognized it as containing a K-complex, using (3). Figure 6 shows four examples of patterns together with their possibilistic labels.

A training set of 200 patterns and a test set of 300 patterns were considered, each one containing an equal number of positive and negative examples. Table 5 shows the test error rates obtained by 5 methods: the voting K-NN rule, Dudani's distance-weighted K-NN rule, the EDC method with optimized parameters [40] and the EDCP-I and II methods introduced in this paper, for $K \in \{9, 11, 13\}$ neighbors. These results clearly show a consistent decrease in error rate due to the consideration of the uncertainty in class labels.

6 Conclusions

When applying the pattern classification methodology in such domains as medical diagnosis or process monitoring, the situation often occurs where the class membership of training patterns is not given a priori and has to be derived from subjective expert assessments. One of the forms in which expert knowledge may be encoded is that of a possibility distribution, consisting of real numbers in the range [0,1] indicating the degree of possibility with which each pattern belongs to each class. The resulting training information may then be given as a set of training patterns with associated possibilistic class labels.

In this paper, this problem was tackled in the framework of Evidence Theory. The evidential distance-based classifier previously introduced by one of the authors was extended to handle possibilistic labeling of training data. Two approaches were proposed based either on the transformation of each possibility distribution into a consonant belief function, or on the use of generalized belief structures with fuzzy focal elements. In each case, a belief function modeling the expert's beliefs concerning the class membership of each new pattern to be classified can be obtained. This information may then be either interpreted by a human operator to support decision-making, or automatically processed to yield a final class assignment through the computation of pignistic probabilities. Numerical experiments demonstrated the ability of both

classification schemes to make effective use of possibilistic labels as training information.

Acknowledgment

The author thanks Cédric Richard and Régis Lengellé from UTT (Université de Technologie de Troyes) for kindly providing the EEG data, as well as the Foundation for Applied Neuroscience Research in Psychiatry (CHS de Rouffach, 68250 Rouffach, France) for authorizing their use.

References

[1] J. C. Bezdek, S. K. Chuah, and D. Leep. Generalized k-nearest neighbor rules. *Fuzzy Sets and Systems*, 18:237–256, 1986.

[2] J. C. Bezdek and S. K. Pal. *Fuzzy models for pattern recognition.* IEEE Press, Piscataway, NJ, 1992.

[3] 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.

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

[5] T. Denœux. Application du modèle des croyances transférables en reconnaissance de formes. *Traitement du Signal*, 14(5):443–451, 1998.

[6] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M. M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, New-York, 1982.

[7] D. Dubois and H. Prade. Unfair coins and necessity measures: towards a possibilistic interpretation of histograms. *Fuzzy sets and systems*, 10(1):15–20, 1983.

[8] D. Dubois and H. Prade. Evidence measures based on fuzzy information. *Automatica*, 21(5):547–562, 1985.

[9] D. Dubois and H. Prade. An alternative approach to the handling of subnormal possibility distributions. *Fuzzy Sets and Systems*, 24:123–126, 1987.

[10] D. Dubois and H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty.* Plenum Press, New-York, 1988.

[11] S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems Man and Cybernetics*, SMC-6:325–327, 1976.

[12] M. Ishizuka and K. S. Fu. Inference procedures under uncertainty for the problem-reduction method. *Information Sciences*, 28:179–206, 1982.

[13] J.-Y. Jaffray. Dynamic decision making using belief functions. In R. R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 331–352. Wiley, New-York, 1994.

[14] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-NN neighbor algorithm. *IEEE Trans. Syst. Man Cybern.*, SMC-15(4):580–585, 1985.

[15] C. Richard. *Une méthodologie pour la détection à structure imposée. Applications au plan temps-fréquence.* PhD thesis, Université de Technologie de Compiègne, 1998.

[16] C. Richard and R. Lengellé. Data driven design and complexity control of time-frequency detectors. *Signal Processing*, 77:37–48, 1999.

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

[18] Ph. Smets. The degree of belief in a fuzzy event. *Information Sciences*, 25:1–19, 1981.

[19] Ph. Smets. Possibilistic inference from statistical data. In *Second World Conference on Mathematics at the service of Man*, pages 611–613, Universidad Politecnica de Las Palmas, 1982.

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

[21] Ph. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R. D. Schachter, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 29–40. North-Holland, Amsterdam, 1990.

[22] Ph. Smets. Resolving misunderstandings about belief functions. *International Journal of Approximate Reasoning*, 6:321–344, 1990.

[23] Ph. Smets. The alpha-junctions: Combination operators applicable to belief functions. In *ECSQARU'97*, Bad Honnef, Germany, June 1997.

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

[25] Ph. Smets. Numerical representation of uncertainty. In D. M. Gabbay and Ph. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 3, pages 265–309. Kluwer Academic Publishers, Dordrecht, 1998.

[26] 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.

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

[28] T. M. Strat. Decision analysis using belief functions. In R. R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 275–309. Wiley, New-York, 1994.

[29] P. Walley and S. Moral. Upper probabilities based on the likelihood function. *Journal of the Royal Statistical Society B*, 161:831–847, 1999.

[30] R. R. Yager. Generalized probabilities of fuzzy events from fuzzy belief structures. *Information Sciences*, 28:45–62, 1982.

[31] R. R. Yager. Arithmetic and other operations on Dempster-Shafer structures. *Int. J. Man-Machines Studies*, 25:357–366, 1986.

[32] R. R. Yager. On the normalization of fuzzy belief structure. *International Journal of Approximate Reasoning*, 14:127–153, 1996.

[33] R. R. Yager, M. Fedrizzi, and J. Kacprzyk. *Advances in the Dempster-Shafer theory of evidence*. John Wiley and Sons, New-York, 1994.

[34] R. R. Yager and D. P. Filev. Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(8):1221–1230, 1995.

[35] J. Yen. Generalizing the Dempster-Shafer theory to fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):559–569, 1990.

[36] J. Yen. Computing generalized belief functions for continuous fuzzy sets. *International Journal of Approximate Reasoning*, 6:1–31, 1992.

[37] L. A. Zadeh. Probability measures of fuzzy events. *J. Math. Analysis and Appl.*, 10:421–427, 1968.

[38] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[39] L. A. Zadeh. Fuzzy sets and information granularity. In R. K. Ragade M. M. Gupta and R. R. Yager, editors, *Advances in Fuzzy Sets Theory and Applications*, pages 3–18. North-Holland, Amsterdam, 1979.

[40] 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.

Tables and figures

Table 1: Belief structures and resulting pignistic probabilities for $x = 1.2$ (Example 1)

A	$m(A\|e^1)$	$m(A\|e^2)$	$m(A\|e^3)$	$m(A)$	BetP(A)
$\emptyset$	0	0	0	0.01	0
$\{1\}$	0	0	0	0	0.19
$\{2\}$	0	0	0	0.10	**0.47**
$\{3\}$	0	0	0.04	0.03	0.34
$\{1,2\}$	0.21	0	0	0.11	0.66
$\{1,3\}$	0	0	0	0	0.53
$\{2,3\}$	0	0.47	0	0.36	0.81
$\{1,2,3\}$	0.79	0.53	0.96	0.40	1

Table 2: Belief structures and resulting pignistic probabilities for $x = 1.2$ (Example 3)

A	m^1	m^2	m^3	$m(A\|\widetilde{e}^1)$	$m(A\|\widetilde{e}^2)$	$m(A\|\widetilde{e}^3)$	$m(A)$	BetP(A)
$\emptyset$	0	0	0	0	0	0	0.03	0
$\{1\}$	0.2	0	0	0.04	0	0	0.02	0.22
$\{2\}$	0	0.2	0	0	0.10	0	0.14	**0.47**
$\{3\}$	0	0	0.8	0	0	0.03	0.02	0.31
$\{1,2\}$	0.7	0	0	0.15	0	0	0.08	0.69
$\{1,3\}$	0	0	0	0	0	0	0	0.53
$\{2,3\}$	0	0.7	0.2	0	0.33	0.01	0.26	0.78
$\{1,2,3\}$	0.1	0.1	0	0.81	0.57	0.96	0.45	1

Table 3: Belief structures for $x = 1.2$ (Example 4)

$\widetilde{A}$	$\widetilde{m}(\widetilde{A}\|\widetilde{e}^1)$	$\widetilde{m}(\widetilde{A}\|\widetilde{e}^2)$	$\widetilde{m}(\widetilde{A}\|\widetilde{e}^3)$	$\widetilde{m}(\widetilde{A})$
$\{1/1, 0.8/2, 0.1/3\}$	0.21	0	0	0.11
$\{0.1/1, 1/2, 0.8/3\}$	0	0.47	0	0.36
$\{0/1, 0.2/2, 1/3\}$	0	0	0.04	0.01
$\{0/1, 0.2/2, 0.1/3\}$	0	0	0	0.01
$\{0.1/1, 0.8/2, 0.1/3\}$	0	0	0	0.10
$\{0.1/1, 0.2/2, 0.8/3\}$	0	0	0	0.01
$\{1, 2, 3\}$	0.79	0.53	0.96	0.40

Table 4: Means and standard deviations (in parentheses) of test error rates for 20 trials. The methods are: the weighted 5-NN rule (w5-NN), the fuzzy 5-NN rule (f5-NN), the evidential distance-based classifier (EDC) and the evidential distance based classifiers with possibilistic labels (EDCP-I and II).

labels	w5-NN	f5-NN	EDC	EDCP-I	EDCP-II
real	0.39	0.40	0.45	–	–
	(0.08)	(0.10)	(0.10)		
ML	0.42	0.44	0.46	–	–
	(0.09)	(0.12)	(0.11)		
possibilistic	–	–	–	0.33	0.32
				(0.09)	(0.08)

Table 5: Test error rates for the EEG data, using different values of K (number of neighbors). The methods are: the voting K-NN rule, the weighted K-NN rule (w K-NN), the evidential distance-based classifier (EDC) and the evidential distance based classifiers with possibilistic labels (EDCP-I and II).

K	K-NN	w K-NN	EDC	EDCP-I	EDCP-II
9	0.30	0.30	0.31	0.28	0.27
11	0.29	0.30	0.29	0.27	0.26
13	0.31	0.30	0.31	0.27	0.26

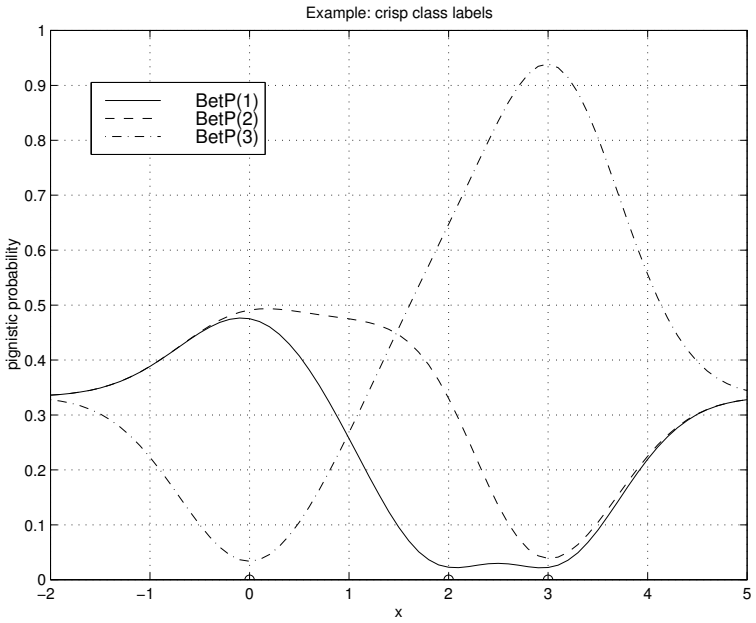

Figure 1: Pignistic probabilities as a function of x for the classification problem of Example 1.

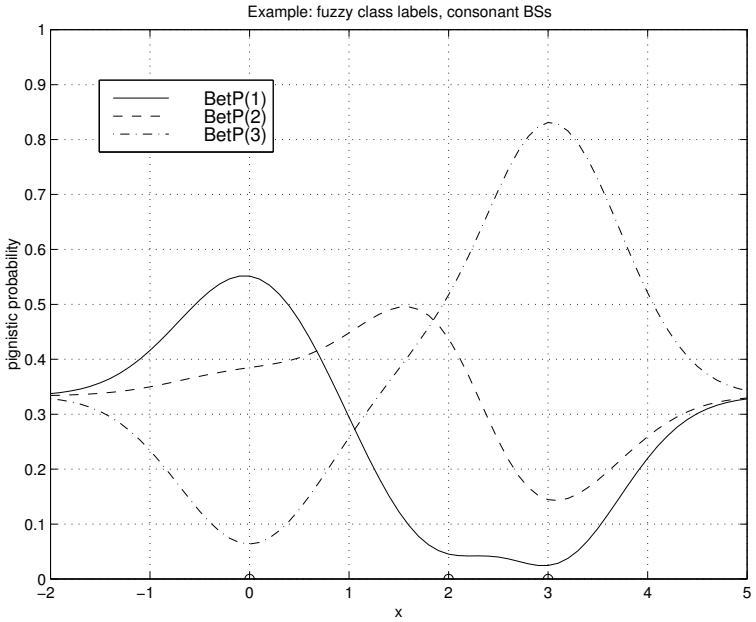

Figure 2: Pignistic probabilities as a function of x for the classification problem of Example 3.

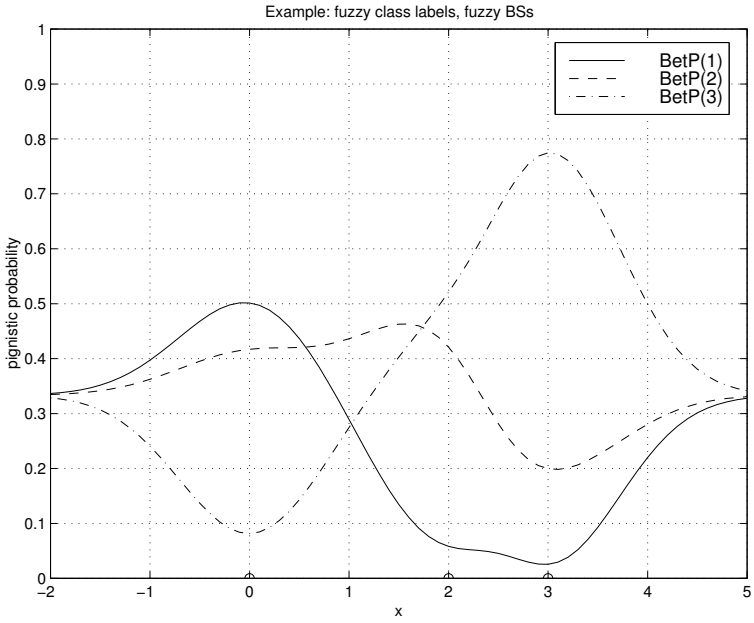

Figure 3: Pignistic probabilities as a function of x for the classification problem of Example 4.

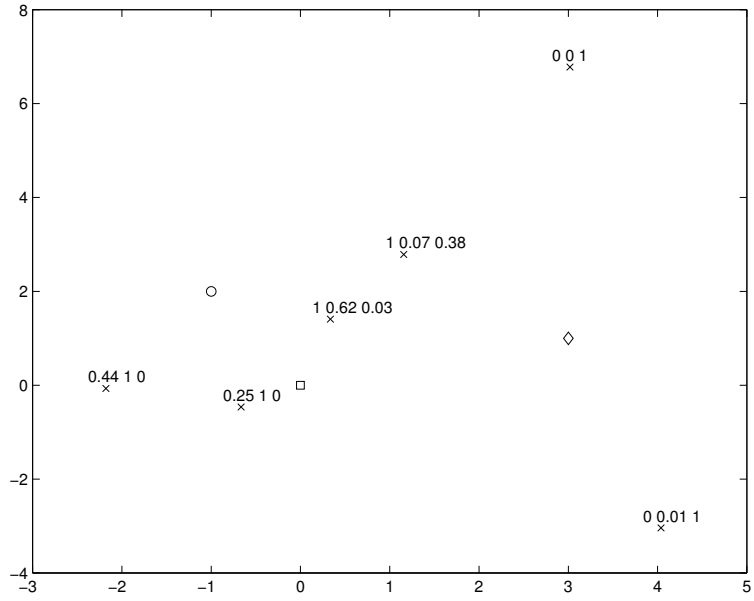

Figure 4: A training data set of 6 patterns with possibilistic class labels. The means of the three underlying distributions are shown as ∘ (class 1), □ (class 2) and ♢ (class 3).

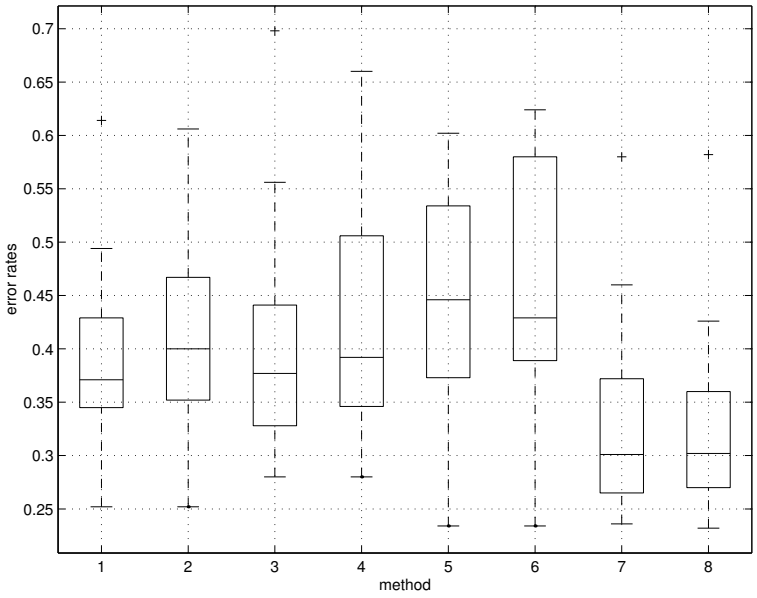

Figure 5: Box plots of misclassification error rates (1: weighted 5-NN rule with real labels; 2: weighted 5-NN rule with ML labels; 3: fuzzy 5-NN rule with real labels; 4 : fuzzy 5-NN rule with ML labels; 5: EDC with real labels; 6 : EDC with ML labels; 7: EDCP-I ; 8: EDCP-II). Boxes indicated the lower quartile, median and upper quartiles of each distribution; whiskers show the extent of the rest of the data.

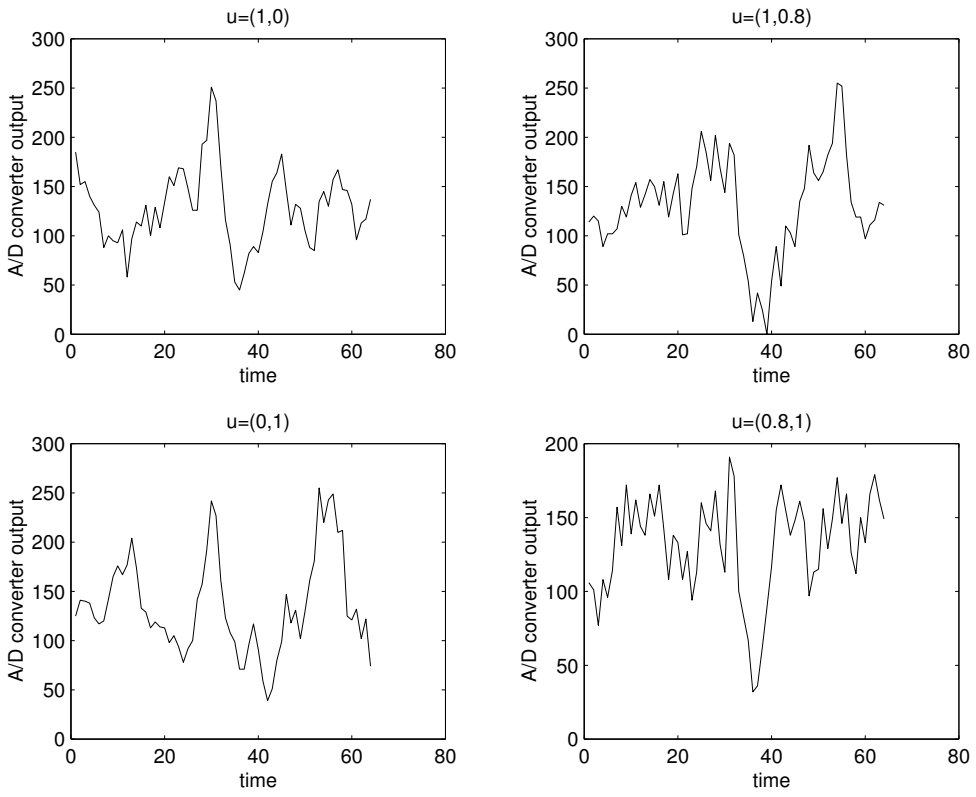

Figure 6: Four examples of EEG patterns, with corresponding possibilistic class labels: positive examples (upper two figures) and negative examples (lower two figures).