Reasoning with unlabeled samples and belief functions

Patrick Vannoorenberghe

Perception **S**ystèmes **I**nformation, FRE 2645 CNRS
Université de Rouen, Faculté des Sciences
F-76821 Mont Saint Aignan Cedex, France
Patrick.Vannoorenberghe@univ-rouen.fr

***Abstract*— This paper presents a decision rule which allows to reason with unlabeled samples in the framework of Dempster-Shafer (DS) theory of evidence. Thanks to the power of this theoretical framework to represent different kinds of knowledge (from total ignorance to full knowledge), we propose an extension of a so-called evidential classifier which allows to process learning sets whose labeling has been specified with belief functions. This kind of functions can encode partial knowledge on examples of the learning set. In this context, using unlabeled examples can significantly improve the performance of the classifier. In addition, the proposed methodology constitutes by this way a convergence point between supervised and unsupervised learning.**

I. Introduction

In supervised learning, the classification problem consists in assigning an input pattern $\mathbf{x}$ to a class, given a learning set $\mathcal{L}$ composed of n patterns $\mathbf{x}_i$ with known classification. Each pattern in $\mathcal{L}$ is represented by a feature vector $\mathbf{x}_i$ and its corresponding class label ω_i. However, in many applications, a lot of unlabeled data can be available while labeled instances are scarse. This is due to the fact that they may be difficult or impossible to obtain under given circumstances or require expensive expert knowledge. Solutions to handle this kind of information have been proposed by several authors [1], [2], [3], [4]. In addition, the learning set can be composed of examples whose class is not precisely known. Instead the knowledge of ω_i, the expert gives a subset of classes which should include the correct solution. This information sometimes describes more precisely the true state of knowledge and is generally denoted partial labeling. A solution to this problem has been proposed in the probabilistic framework [5].

In order to take into account this kind of labeling (partial labels and unlabeled instances), several solutions based on the Dempster-Shafer (DS) theory of evidence [6], have been proposed [7], [8]. Partial labeling has been investigated in the framework of Dempster-Shafer theory because this last enables to reason on beliefs expressed on subsets of Ω. Advantages of these techniques are numerous including the description of the uncertainty on the prediction, the possibility of rejecting a pattern and detecting unknown class, ... More recently, several decision tree induction methods based on belief functions [9], [8] have been introduced, giving rise to the notion of Belief Decision Tree (BDT). Thanks to the greater flexibility of DS theory to represent different kinds of knowledge (from total ignorance to full knowledge), BDT's allow to process training sets whose labeling has been specified with belief functions [10]. Semi supervised learning is a special case of partial labeled problem where all examples are either precisely labeled or unlabeled i.e. with labels belonging to Ω. However, unlabeled examples have not been explicitly used to improve the performance of the decision rule.

In this paper, we propose a decision rule which can cope with training sets whose labelling is not precisely known and consider the potential role of unlabeled data in supervised learning. We present an algorithm based on belief functions theory and experimental results demonstrating that unlabeled data can significantly improve learning accuracy in certain practical problems. This paper is organized as follows. The basic concepts of belief function theory are first briefly introduced, including the way to handle uncertain labels with belief functions (Section II). The methodology and the proposition to evaluate the classifier in this context are described in Section III. Finally, Section IV presents some experimental results.

II. Background

In this section, several concepts of the DS theory of evidence [6] are recalled, which allows to introduce uncertain labeling and notations used in this paper. Let $\Omega = \{\omega_q, q = 1, \cdots, Q\}$ denote a finite set, generally called the frame of discernment. In pattern classification, Ω is the set of Q classes to be recognize. A basic belief assignment (bba) m on Ω is defined[1] as a function from 2^Ω to $[0,1]$ verifying $m(\emptyset) = 0$ and $\sum_{A\subseteq\Omega} m(A) = 1$. Each subset $A \subseteq \Omega$ such as $m(A) > 0$ is called a focal element of m. From this, a communality function q is defined as $q(A) = \sum_{B\supseteq A} m(B)$. Note that functions m and q are in one-to-one correspondence [6], and can be seen as two facets of the same piece of information. Functions q are generally used for combination of several pieces of evidence.

A. Uncertain labeling

This paper focuses on learning from partially labeled data in the framework of belief function theory. In this context, the available learning set can be written of the form:

$$\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \cdots, n\}, \tag{1}$$

where m_i is a bba defined on Ω and represents the knowledge on the label of the i^{th} example[2]. This belief function can

[1]The notation $m^\Omega[data]$ is generally used to denote a bba defined on the domain Ω based on observed $[data]$.

[2]For the sake of simplicity and because all belief functions used in this paper are defined on the same frame Ω, the superscript is forget in the sequel.

$A \subseteq \Omega$	HL	IL	PrL	PoL
$\{\omega_1\}$	0	0	0.2	0
$\{\omega_2\}$	1	0	0.6	0
$\{\omega_1, \omega_2\}$	0	1	0	0
$\{\omega_3\}$	0	0	0.2	0.7
$\{\omega_1, \omega_3\}$	0	0	0	0.2
$\{\omega_2, \omega_3\}$	0	0	0	0
Ω	0	0	0	0.1

TABLE I

EXAMPLE OF UNCERTAIN LABELING WITH BELIEF FUNCTIONS

represent different forms of label including of course hard labels (HL), probabilistic labels (PrL), possibilistic (PoL) labels or imprecise labels (IL). Table I illustrates an example of these evidential labels on a three-class frame. Note that a possibility measure is known to be formally equivalent to a consonant belief function, i.e., a belief function with nested focal elements [7]. Unlabeled samples can be encoded using the vacuous belief function m_v defined as $m_v(\Omega) = 1$.

B. Operations on belief functions

An α-discounted bba m_α can be obtained from an original bba m as follows:

$$m_\alpha(A) = \alpha m(A) \qquad \forall\, A \subseteq \Omega, A \neq \Omega \tag{2}$$

$$m_\alpha(\Omega) = 1 - \alpha + \alpha m(\Omega) \tag{3}$$

with $0 \leq \alpha \leq 1$. The discounting operation is useful when the source of information from which m has been derived is not fully reliable, in which case coefficient α represents some form of metaknowledge about the source reliability, which could not be encoded in m.

Two pieces of evidence m_1 and m_2 can be aggregated with the Dempster's rule of combination (orthogonal sum $\oplus$), yielding to an unique belief function m defined as:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)} \quad \forall A \subseteq \Omega. \tag{4}$$

The use of this rule is possible only if m_1 and m_2 are not totally conflicting, i.e., if there exists two focal elements B and C of m_1 and m_2 satisfying $B \cap C \neq \emptyset$. Note that combining a belief function m with the vacuous belief function m_v leads to the same belief function m. The conjunctive combination of these two pieces of evidence ($m = m_1 \cap m_2$) can be computed from q_1 and q_2 as:

$$q(A) = q_1(A) q_2(A) \qquad \forall\, A \subseteq \Omega. \tag{5}$$

This rule is sometimes referred to as the (unnormalized) Dempster's rule of combination.

Based on rationality arguments developed in the TBM (**T**ransferable **B**elief **M**odel), Smets [11] proposes to transform m into a probability function p_m on Ω (called the *pignistic* probability function) defined for all $\omega_q \in \Omega$ as:

$$p_m(\omega_q) = \sum_{A \ni \omega_q} \frac{m(A)}{|A|} \tag{6}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. In this transformation, the mass of belief $m(A)$ is distributed equally among the elements of A. This pignistic probability function is used in the TBM for decision making.

III. METHODS

A. Problem

Let us suppose that the learning set $\mathcal{L}$ is composed of a partial labeled set defined as $\mathcal{P} = \{(\mathbf{x}_p, m_p), p = 1, \cdots, P\}$ and the unlabeled set $\mathcal{U} = \{(\mathbf{x}_u, m_v), u = 1, \cdots, U\}$ with m_v the vacuous belief function. The availability of this unlabeled data set poses the challenge of how to use them in order to improve generalization in semi-supervised learning. In this paper, the idea consists in re-labeling the set $\mathcal{U}$ in order to improve the performance of the algorithm learnt on the whole set $\mathcal{L}$. We then need an algorithm which can estimate an output bba based on the partial labeled set $\mathcal{P}$. Such a decision rule is proposed in [7] and briefly introduced in the next section. In this paper, this extension of the k nearest neighbors algorithm is used for simulations but another kind of classifier, generally denoted evidential classifier, has been proposed by several authors [8], [10] and can be used in this context.

B. Distance-based Method

In the method introduced by Denœux [12], a basic belief assignment is constructed directly, using as a source of information the training patterns $\mathbf{x}_i$ situated in the neighborhood of the pattern $\mathbf{x}$ to be classified. If the k nearest neighbors (according to some distance measure) are considered, we thus obtain k bba's that are combined using the Dempster's rule of combination. The initial method was later refined to allow parameter optimization [13], and a neural-network-like version was recently proposed [14]. Finally, a generalization to imprecise labelling has been proposed in [7]. Each neighbor can be viewed as a piece of evidence that influences the belief concerning the membership class of $\mathbf{x}$ according to a discounting coefficient. A belief function m_i associated to each neighbor i is then defined as:

$$m_{\mathbf{x}}[\mathbf{x}_i](A) = \phi(d_i) m_i(A) \qquad \forall\, A \subseteq \Omega, A \neq \Omega \tag{7}$$

$$m_{\mathbf{x}}[\mathbf{x}_i](\Omega) = 1 - \sum_{A \subset \Omega} m_{\mathbf{x}}[\mathbf{x}_i](A) \tag{8}$$

where d_i is the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}_i$ and $\phi(.)$ is a decreasing function defined as $\phi(d_i) = \exp[-\gamma (d_i)^2]$ with γ a positive parameter. The k belief functions $m_{\mathbf{x}}[\mathbf{x}_i](.)$ for the k nearest neighbors are then aggregated using the Dempster's rule of combination:

$$\hat{m}_{\mathbf{x}} = \bigoplus_{i=1}^{k} m_{\mathbf{x}}[\mathbf{x}_i]. \tag{9}$$

In this method, an unlabeled sample $\mathbf{x}_i$ (an example with unknown classification $m_i = m_v$) has no influence of $\hat{m}_{\mathbf{x}}$. In other terms, learning the algorithm on the set $\mathcal{L}$ leads to consider only neighbors with known classification.

C. Use unlabeled samples

Using the learning set $\mathcal{P}$, we first build a decision rule using the distance-based method previously presented. Having observed $\mathcal{P}$, the algorithm is then used to estimate output belief functions on the unlabeled data set $\mathcal{U}$. Let us denote $\hat{m}_{\mathcal{U}}[\mathcal{P}]$ these belief functions estimated on all instances of $\mathcal{U}$. By this way, all learning examples containing in $\mathcal{L}$ have been labeled and can be used to build the classifier on the final learning set. This available data set is of the form $\mathcal{P} = \{(\mathbf{x}_p, m_p), p = 1, \cdots, P\}$ and the re-labeled set $\mathcal{U} = \{(\mathbf{x}_u, \hat{m}_{\mathcal{U}}[\mathcal{P}]), u = 1, \cdots, U\}$. Note that it is possible to use different algorithms to build the first decision rule and the final classifier. For example, we can build the first decision rule used to estimate output belief functions on unlabeled data with the method presented in the section III-B and used a BDT for the final classifier.

D. Evaluation

Performance assessment is an important issue in the design of a classifier. In a decision-theoretic setting, this problem is formalized by considering a set of actions $\mathcal{A}$, and a loss function $L : \mathcal{A} \times \Omega \mapsto \mathbb{R}$, where $L(\alpha, \omega)$ is the loss incurred if one selects action α and the true state of nature is ω. Typically, each action in $\mathcal{A}$ corresponds to the choice of a class, and the loss is one for misclassification, and 0 for correct classification. The performance of a classifier $c : \mathbb{R}^d \mapsto \mathcal{A}$ can then be measured by taking the expectation of $L(c(\mathbf{x}), \omega)$ over both $\mathbf{x}$ and ω. This expectation is usually estimated by a sample average over test data. In our case, this framework needs to be extended in two directions:

- the output of an evidential classifier is a belief function: the set of actions is thus a set of belief functions; we then need to define the loss associated to an output bba $\widehat{m}$ when the true state of nature is ω;
- the test set may be of the form defined in (1), i.e., the class of test pattern may be only partially known.

A first solution was proposed in [7], [9]. This solution postulates the following loss function:

$$L(\widehat{m}, m) = 1 - \sum_{A \subseteq \Omega} m(A) p_{\widehat{m}}(A) \tag{10}$$

where $\widehat{m}$ is the output bba produced by the classifier, and m is a bba that quantifies the uncertainty concerning the true state of nature ω. A nice property of this loss function is that, when $m(\Omega) = 1$, $L(\widehat{m}, m) = 0$ whatever $\widehat{m}$, which seems reasonable. Deeper understanding of this loss function can be gained by observing that:

$$\begin{aligned} L(\widehat{m}, m) &= 1 - \sum_{A \subseteq \Omega} m(A) \sum_{B \subseteq \Omega} \widehat{m}(B) \frac{|B \cap A|}{|B|} \\ &= 1 - \sum_{A,B \subseteq \Omega} m(A) \widehat{m}(B) \text{Incl}(B, A) \end{aligned}$$

where $\text{Incl}(B, A) = |B \cap A|/|B|$ is the degree of inclusion of B in A. An alternative form of $L(\widehat{m}, m)$ is given by

$$\begin{aligned} L(\widehat{m}, m) &= 1 - \sum_{A \subseteq \Omega} m(A) \sum_{\omega \in A} p_{\widehat{m}}(\omega) && (11) \\ &= 1 - \sum_{\omega \in \Omega} p_{\widehat{m}}(\omega) \sum_{A \ni \omega} m(A) && (12) \\ &= 1 - \sum_{\omega \in \Omega} p_{\widehat{m}}(\omega) q(\{\omega\}) \ . && (13) \end{aligned}$$

We can therefore propose a criterion to evaluate the performance of a classifier on a test set of n' examples $(\mathbf{x}_i, m_i)$, $i = 1, \ldots, n'$:

$$C_1 = 1 - \frac{1}{n'} \sum_{i=1}^{n'} \sum_{\omega \in \Omega} p_{\widehat{m}_i}(\omega) q_i(\{\omega\}) \tag{14}$$

where q_i is the commonality function associated to m_i, and $\widehat{m}_i$ is the output bba for example i. This kind of cost function can be used to optimize parameters of the evidential classifier.

IV. Results

In this section, we present several simulations in order to illustrate the performance of the proposed methodology.

A. Synthetic data

Let us first consider a simple two-class problem in which the training set is composed of only four patterns in a one-dimensional space with $x_1 = 0$, $x_2 = 0.5$, $x_3 = 2$ and $x_4 = 3$. Let us suppose that the two first samples belong to class ω_1 while the other ones belong to class ω_2 but the second example x_2 has not been labeled by the expert. To illustrate the above classification procedure, we compare the pignistic probabilities obtained in learning the algorithm with the unlabeled sample itself $m_2(\Omega) = 1$, learning the algorithm with the true class $m_2(\{\omega_1\}) = 1$ and finally learning the algorithm with the re-labeled sample. The distance-based method presented in the section III-B is both used to label the second example and for the final decision rule. Figure 1 illustrates these pignistic probabilities associated to each of the two classes. In this problem, we can note that the pignistic probabilities obtained using the proposed algorithm tend to the true probability (i.e., when the algorithm is learnt with the true class). This result clearly shows that the methodology presented in this paper can give a more precise idea of the data distribution.

B. Synthetic 3-class problem

For this simulation, a learning set $\mathcal{L}$ was generated using 3 classes containing 50 bidimensional vectors each. Each vector $\mathbf{x}$ from class q was generated by first drawing a vector $\mathbf{z}$ from a Gaussian $f(\mathbf{z}|\omega_q) \sim \mathcal{N}(\mu_q, \Sigma_q)$, and applying a non linear transformation $\mathbf{z} \mapsto \mathbf{x} = \exp(0.3\ \mathbf{z})$ to obtain non-Gaussian data. The means of the 3 Gaussian distributions were taken as: $\mu_1 = (-1, -1)'$, $\mu_2 = (1, 2)'$, $\mu_3 = (-1.5, 2)'$ and the variance matrices were of the form $\Sigma_q = D_q A D_q'$ with

$$A = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{3}/3 \end{pmatrix} \quad D_q = \begin{pmatrix} \cos\theta_q & -\sin\theta_q \\ \sin\theta_q & \cos\theta_q \end{pmatrix}$$

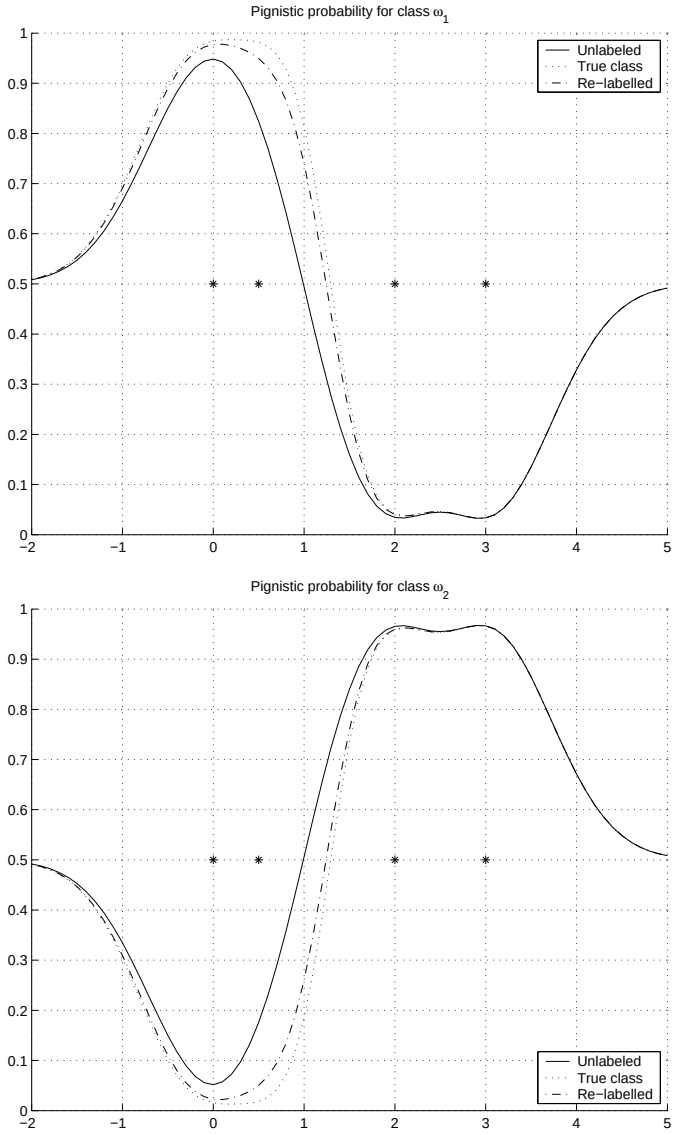

Fig. 1. Pignistic probabilities of class ω_1 and ω_2 for the three methods ($*$ = learning vectors)

and $\theta_1 = \pi/3$, $\theta_2 = \pi/2$, $\theta_3 = -\pi/3$. We need to compare the performance of three classifiers : a classifier learnt with all labeled samples, a classifier learnt with 100 unlabeled samples and finally a classifier learnt using the re-labeled data. To compare the performances of the 3 methods, a test set $\mathcal{T}$ was generated using the same distribution as $\mathcal{L}$ with $15,000$ samples. The results are given in table II. According to the mean errors rates and respective standard deviations, it is obvious to note that the proposed methodology can significantly improve the performance of the classifier learnt with re-labeled samples.

C. Increasing the number of unlabeled data

Another simulation is used to evaluate the performance of the classification rule. The goal is to demonstrate that supervised learning can be improved using unlabeled samples.

Error	All labeled	Unlabeled	Re-labeled
Mean	0.112	0.279	0.146
Sd	0.011	0.029	0.036

TABLE II
MEAN ERROR RATES AND STANDARD DEVIATIONS FOR THE 3 METHODS

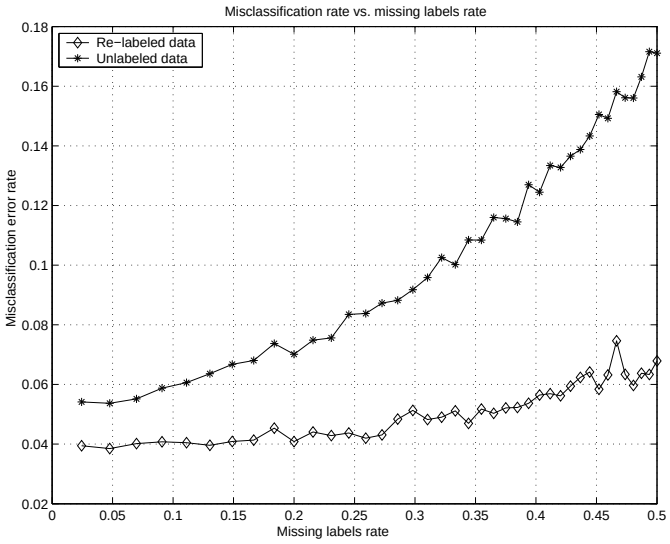

Fig. 2. Error rate vs. missing labels rate

For this simulation, a learning set $\mathcal{L}$ was generated using 2 classes containing 100 bidimensional vectors each drawing with Gaussian distributions. Missing labels are randomly chosen in the learning set. To estimate the generalization performances, a test set was generated using the same distributions as $\mathcal{L}$ with $5,000$ samples. The experiment was repeated ten times with independent training sets. The criterion presented in the equation (14) is used for optimizing the parameters of the classifier while number of neighbors is adjusted using a cross-validation set. Figure 2 shows the misclassification error rate vs. the missing labels rate. It is obvious to see that the proposed decision rule obtains better performance than the rule learnt on unlabeled examples. In fact, if we consider unlabeled samples in the learning set $\mathcal{L}$, the number of neighbors (with known classification) used to estimate the output bba $\hat{m}_{\mathcal{L}}[\mathcal{L}]$ tends to decrease. For example, with 50% of missing labels in the learning set, the probability to observe k neighbors with know classification is divided by 2. Consequently, this increases the uncertainty on the prediction and damages the performance of the classification procedure. On the contrary, using the re-labeled data leads to stabilize the misclassification error rate.

V. CONCLUSION

This paper has focused on pattern recognition techniques based on the Dempster-Shafer theory of evidence. In this context, a classification procedure has been proposed to cope with partial labeling and unlabeled samples. An extension of evidential classifiers which allows to process learning sets whose labeling has been specified with belief functions has

been presented. The idea consists in substituting unlabeled data with predictions estimated from a first decision rule learnt on labeled data. The proposed methodology can be extended to other algorithms where uncertain labels are handled. Unlabeled samples have been used to improve the performance of the decision rule which can be used in many applications of semi-supervised learning where an abundance of unlabeled data is available. Finally, the proposed methodology constitutes a convergence point between supervised and unsupervised learning. Future work concerns a validation of the method on a real-world application. Using the proposed methodology as a way of exploiting unlabeled data in Content Based Image Retrieval can offer some interesting insights. In such applications, the major difficulty for learning during relevance feedback is the relatively small numbers of labeled training samples available from the user.

REFERENCES

[1] T. Mitchell, "The role of unlabeled data in supervised learning," in *Proceedings of the Sixth International Colloquium on Cognitive Science*, San Sebastian, Spain, 1999.

[2] M. Szummer and T. Jaakkola, "Kernel expansions with unlabeled examples," in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2000.

[3] K. Bennett, A. Demiriz, and R. Maclin, "Exploiting unlabeled data in ensemble methods," in *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 2002.

[4] R. Dara, S. Kremer, and D. Stacey, "Clustering unlabeled data with SOMs improves classification of labeled real-world data," in *Proceedings of the World Congress on Computational Intelligence*, 2002, pp. 2237–2242.

[5] Y. Grandvalet, "Logistic regression for partial labels," in *Proceedings of IPMU'2002*, Annecy, France, 2002, pp. 1935–1941.

[6] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[7] T. Denœux and L. Zouhal, "Handling possibilistic labels in pattern classification using evidential reasoning," *Fuzzy Sets ans Systems*, vol. 122, no. 3, pp. 47–62, 2001.

[8] Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: Theoretical foundations," *International Journal of Approximate Reasoning*, vol. 28, pp. 91–124, 2001.

[9] T. Denœux and M. S. Bjanger, "Induction of decision trees from partially classified data using belief functions," in *Proceedings of SMC'2000*. Nashville, USA: IEEE, 2000, pp. 2923–2928.

[10] P. Vannoorenberghe and T. Denœux, "Handling uncertain labels in multiclass problems using belief decision trees," in *Proceedings of IPMU'2002*, Annecy, France, 2002, pp. 1919–1926.

[11] P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[12] T. Denœux, "A k-nearest neighbour classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[13] L. Zouhal and T. Denœux, "An evidence-theoretic k-nn rule with parameter optimization," *IEEE Transactions on Systems, Man and Cybernetics-Part C*, vol. 28, no. 2, pp. 263–271, May 1998.

[14] T. Denœux, "A neural network classifier based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics, Part A : Systems and humans*, vol. 30, no. 2, pp. 131–150, 2000.