An evidence-theoretic neural network classifier*

Thierry Denœux
Université de Technologie de Compiègne
U.R.A. CNRS 817 Heudiasyc
BP 649 F-60206 Compiègne cedex
France

ABSTRACT

A new classifier based on the Dempster-Shafer theory of evidence is presented. The approach consists in considering the similarity to prototype vectors as evidence supporting certain hypotheses concerning the class membership of a pattern to be classified. The different items of evidence are represented by basic belief assignments over the set of classes and combined by Dempster's rule of combination. An implementation of this procedure in a neural network with specific architecture and learning procedure is presented. A comparison with LVQ and RBF neural network classifiers is performed.

1. INTRODUCTION

Recently, a new approach to pattern classification based on the Dempster-Shafer (D-S) theory of Evidence was proposed [2, 3]. This approach is based on the computation for each input vector to be classified of a basic belief assignment (BBA), i.e., a function from the power set of the set Ω of classes, to $[0, 1]$ (see [10]). A BBA specifies a belief function and a plausibility function over the set of classes, which can be interpreted as defining posterior probability intervals. The width of these intervals varies between 0 (perfect knowledge) and 1 (complete ignorance), depending on the informativeness of the training data with respect to the class membership of the pattern under consideration. *Ambiguity* and *distance* reject options were introduced using the concepts of *lower* and *upper* expected costs. Simulations with artificial and real-world data sets revealed a better performance of the proposed scheme as compared to the voting k-NN procedure and, to a lesser extent, to the distance-weighted k-NN rule. In [13, 14], a method for determining the optimal set of parameters by optimizing a performance criterion was introduced. This modification was shown to significantly improve the classification accuracy of the method as well as its robustness with respect to the number of neighbors taken in consideration.

In this paper, an adaptive version of this evidence-theoretic classification rule is proposed. In this new approach, the assignment of a pattern to a class is made by computing distances to a limited number of prototypes. Based on these distances and on the degrees of membership of prototypes to each class, BBAs are computed, which are then combined using Dempster's rule. This rule can be implemented in a multilayer neural network with specific architecture comprising one input layer, two hidden layers and one output layer. The weight vector, the receptive field and the class membership of each prototype are determined by minimizing the mean squared differences between the classifier outputs and target values. Experiments demonstrate the effectiveness of this classification scheme as compared to Learning Vector Quantization (LVQ) [6] and Radial Basis Functions (RBF) [8] classifiers.

The paper is organized as follows. Section 2 starts with a brief presentation of the main concepts of D-S theory. The proposed method is introduced in Section 3, and simulation results are presented in Section 4.

2. OVERVIEW OF D-S THEORY

One of the main concepts underlying D-S theory is that of *basic belief assignment* (BBA) [10]. Let us consider a set Θ of mutually exclusive and exhaustive hypotheses about some problem domain called the "frame of discernment". A piece of evidence that influences our belief concerning which hypothesis in Θ is true, can be represented by a basic belief assignment m, which is defined as a function from 2^Θ

*This paper has appeared in the Proceedings of the *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 712–717, Vancouver, October 1995.

to $[0,1]$ verifying $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. The quantity $m(A)$ for any $A \subseteq \Theta$ can be interpreted as the "mass of belief" that one is willing to commit to A (and to none of its subsets), given the available evidence. Situations of complete ignorance therefore correspond to the case where $m(\Theta) = 1$. The information contained in m can be represented in two alternative forms: a belief or credibility function Bel, and a plausibility function Pl defined respectively as $Bel(A) = \sum_{B \subseteq A} m(B)$ and $Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) = 1 - Bel(\bar{A})$. A natural interpretation of $Bel(A)$ is that of a measure of one's degree of belief in A, while $1 - Pl(A)$ provides a measure of doubt in that hypothesis. The belief and plausibility functions can be seen as defining a set $\mathcal{C}$ of *compatible* probability distributions over Θ such that $P \in \mathcal{C}$ iif $Bel(A) \leq P(A) \leq Pl(A)$ for all $A \subseteq \Theta$. The width of the interval $[Bel(A), Pl(A)]$ can be regarded as a measure of one's ignorance of the "probability" of A, although the theory need not assume the existence of any objective probability distribution, as shown by Smets [12].

The "dynamic" part of D-S theory indicates how the evidence coming from several independent sources should be combined. Two BBAs m_1 and m_2 induced by distinct pieces of evidence are said to be *combinable* if there exist at least two non-disjoint subsets B and C of Θ such that $m_1(B) > 0$ and $m_2(C) > 0$. The orthogonal sum of m_1 and m_2, noted $m = m_1 \oplus m_2$ is then defined as $m(\emptyset) = 0$ and

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \qquad (1)$$

for $A \neq \emptyset$. The normalizing factor appearing in denominator in Eq. 1 can be omitted, resulting in an *unnormalized* orthogonal sum [11]. The orthogonal sum (either normalized or unnormalized) is commutative and associative, yielding invariance of the result with respect to the order in which the items of evidence are taken in consideration and combined.

Conventional Bayesian decision analysis is based on expected utilities (or losses) and requires uncertainty to be represented by probability distributions. Although the concept of mathematical expectation can be generalized in the framework of D-S theory [1], probability distributions can be shown to be better suited to decision making [12]. For that reason, Smets introduced the concept of "pignistic" probability $BetP$ obtained by distributing the mass $m(A)$ equally to each $\theta \in A$, for all $A \subseteq \Theta$: $BetP(\theta) = \sum_{\theta \in A} \frac{m(A)}{|A|}$, where $|A|$ denotes the cardinality of A. In this model, belief functions are proposed as a formalism for representing and updating belief ("credal level"), whereas the associated pignistic probabilities should be used to make decisions ("decision level").

3. THE METHOD

Principle

We consider a classification problem in which a pattern $\mathbf{x} \in \mathbb{R}^P$ has to be assigned to one of M classes. $\Omega = \{\omega_1, \ldots, \omega_M\}$ denotes the set of classes. The procedure consists in comparing $\mathbf{x}$ to n reference patterns or prototypes: $\mathbf{p}^1, \ldots, \mathbf{p}^n$. Each prototype i is assumed to have a degree of membership u_q^i to each class ω_q, with $\sum_{q=1}^{M} u_q^i = 1$. A decision regarding the class assignment of $\mathbf{x}$ is made as a result of a three-step procedure:

- The first step consists in computing the distances d^i between $\mathbf{x}$ and each prototype $\mathbf{p}^i$, according to some metric (e.g. the Euclidean one).

- In a second step, the information provided by each prototype, regarding the class membership of $\mathbf{x}$, is represented by a BBA m^i over Ω. The portion of belief $m^i(\{\omega_q\})$ committed to a class ω_q, given the fact that $\mathbf{x}$ is located at the distance d^i from $\mathbf{p}^i$, is a function of both d^i and the class membership of $\mathbf{p}^i$. Assuming proportionality between $m^i(\{\omega_q\})$ and the membership degree u_q^i of prototype i, we can define the mass $m^i(\{\omega_q\})$ as:

  $$m^i(\{\omega_q\}) = \alpha^i u_q^i \phi^i(d^i) \qquad (2)$$

  where ϕ^i is a decreasing function such that $\phi^i(0) = 1$ and $\lim_{d \to \infty} \phi^i(d) = 0$, and $0 < \alpha^i < 1$ is a parameter. Since all masses must sum to 1, the "uncommitted" portion of belief is assigned to the whole set of classes:

  $$m^i(\Omega) = 1 - \alpha^i \phi^i(d^i) \qquad (3)$$

  Note that m^i becomes the *vacuous* BBA when d^i becomes infinitely large. An exponential form is proposed for ϕ^i:

  $$\phi^i(d^i) = \exp(-\gamma^i (d^i)^2) \qquad (4)$$

  where γ^i is a positive parameter defining the receptive field size of prototype i.

- In the last step, the n BBAs m^i, $i = 1, \ldots, n$ obtained in step 2 are combined using either Dempster's rule of combination (Eq. 1) or the

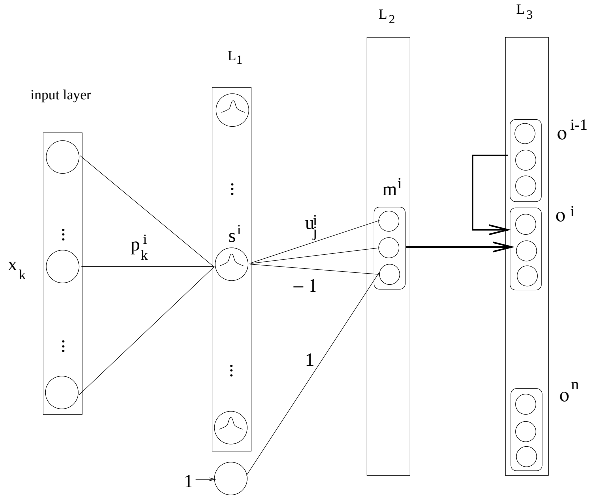

Figure 1: Neural network implementation

unnormalized rule proposed in [11]. It is straightforward to show that the unnormalized BBA $m = \bigwedge_{i=1}^{n} m^i$ and the normalized one $m' = \bigoplus_{i=1}^{n} m^i$ are linked by the following simple relation: $m' = m/K$ with $K = \sum_{q=1}^{M} m(\{\omega_q\}) + m(\Omega)$. Finally, a decision is made to assign the input vector $\mathbf{x}$ to the class $D(\mathbf{x}) = \omega_{qmax}$ with:

$$\begin{aligned} qmax &= \arg\max_q m(\{\omega_q\}) && (5) \\ &= \arg\max_q m'(\{\omega_q\}) && (6) \end{aligned}$$

Note that $D(\mathbf{x})$ also maximizes the belief, plausibility and pignistic probability functions.

Neural network implementation

The evidence-theoretic classifier just introduced can be modeled as a particular type of neural network classifier with one input layer, two hidden layers L_1 and L_2, and one output layer L_3 (Figure 1). Each layer L_1 to L_3 corresponds to one step of the classification procedure described in the former section.

- The first hidden layer L_1 is similar to the hidden layer of a RBF network. It contains n units with Gaussian activation functions

$$s^i = \alpha^i \exp(-\gamma^i \|\mathbf{p}^i - \mathbf{x}\|^2) \qquad (7)$$

- The second layer L_2 computes the BBA m^i associated to each prototype $i = 1, \ldots, n$. It is composed of n modules of $M+1$ units. The outputs from the i-th module is a vector

$$\begin{aligned} \mathbf{m}^i &= (m^i(\{\omega_1\}), \ldots, m^i(\{\omega_M\}), m^i(\Omega)) && (8) \\ &= (u_1^i s^i, \ldots, u_M^i s^i, 1 - s^i) && (9) \end{aligned}$$

- The last (output) layer performs the combination of the n BBAs from the former layer using the Dempster's rule *without normalization*. It is composed of n interconnected modules of $M+1$ sigma-pi units. The outputs $\mathbf{o}^i = (o_1^i, \ldots, o_{M+1}^i)$ from the i-th module correspond to the unnormalized orthogonal sum of $m^1, \ldots, m^i$. They are computed iteratively using the outputs from module i of layer L_2 and module $i-1$ of layer L_3:

$$o_j^i = o_j^{i-1} m_j^i + o_j^{i-1} m_{M+1}^i + o_{M+1}^{i-1} m_j^i \qquad (10)$$

for $j = 1, \ldots, M$ and

$$o_{M+1}^i = o_{M+1}^{i-1} m_{M+1}^i \qquad (11)$$

with $o_j^1 = m_j^1$ for $j = 1, \ldots, M+1$.

The final network output is the vector $\mathbf{o}^n$ of activations in the last module of layer L_3. Its values are the belief masses corresponding to the unnormalized orthogonal sum of $m^1, \ldots, m^n$. The normalized sum m is obtained by normalizing the components of $\mathbf{o}^n$ to unity:

$$\begin{aligned} \forall q \in \{1, \ldots, M\} \quad m^i(\{\omega_q\}) &= o_q^n / K && (12) \\ m^i(\Omega) &= o_{M+1}^n / K && (13) \end{aligned}$$

with $K = \sum_{j=1}^{M+1} o_j^n$. The computation of the activations in the first hidden layer requires $O(nP)$ arithmetic operations (where P denotes the number of inputs), and the combination of the outputs from the first hidden layer is performed in $O(nM)$ operations. Hence, the overall complexity for the propagation of one input pattern is not higher than in the case of RBF networks.

Training

The above network can be trained by minimizing an error function. Let us denote by $\mathbf{t} = (t_1, \ldots, t_M)^t$ the target output vector for pattern $\mathbf{x}$, with $t_j = \delta_{jq}$ if $\mathbf{x} \in \omega_q$. An output error for that pattern can be defined by comparing $\mathbf{t}$ to the classifier output vector $\mathbf{m}$ (or alternatively its normalized version $\mathbf{m}'$). Since $\mathbf{t}$ has M components and $\mathbf{m}$ is of length $M+1$, we need to define from $\mathbf{m}$ a vector of length M. A natural choice is to consider the vector of pignistic probabilities. The pignistic probability of class ω_q is defined from m as $BetP(\omega_q) = m(\{\omega_q\}) + m(\Omega)/M$. The output error $E(\mathbf{x})$ for input pattern $\mathbf{x}$ can then be defined as:

$$E(\mathbf{x}) = \frac{1}{2} \sum_{q=1}^{M} (BetP(\omega_q) - t_q)^2 \qquad (14)$$

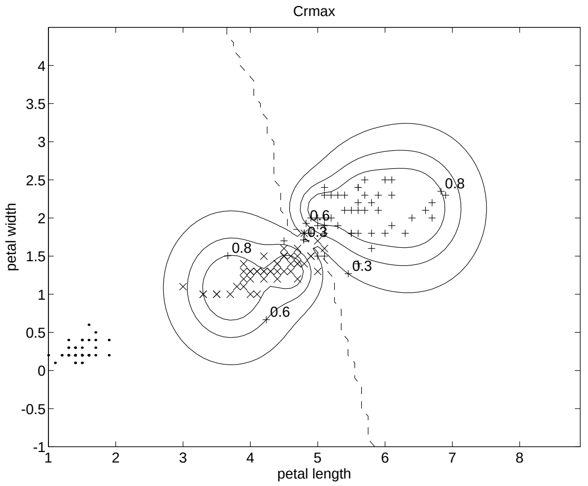

Figure 2: Contours of credibility function $Bel(D(\mathbf{x}))$ (–) and decision boundary (- -). Classes *iris virginica* (+) and *versicolor* (x) were presented during training.

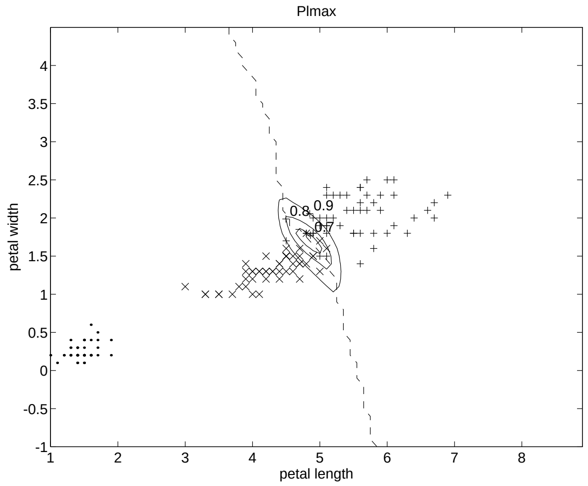

Figure 3: Contours of plausibility function $Pl(D(\mathbf{x}))$ (–) and decision boundary (- -). Classes *iris virginica* (+) and *versicolor* (x) were presented during training.

A mean output error E can be computed by averaging $E(\mathbf{x})$ for all $\mathbf{x}$ in the training set $\mathcal{X}$. Minimizing E amounts to performing least mean square approximation of the posterior probabilities of each class. The free parameters are $p_k^i (k = 1, \ldots, P)$, γ^i, α^i and $u_j^i (j = 1, \ldots, M+1)$ for $i = 1, \ldots, n$. The constraints are $\gamma^i > 0$, $0 < \alpha^i < 1$ and $\sum_{j=1}^{M} u_j^i = 1$ for all $1 \leq i \leq n$. These constraints are automatically satisfied by introducing new parameters η^i, ξ^i and β_q^i such that:

$$\gamma^i = (\eta^i)^2 \tag{15}$$

$$\alpha^i = (1 + \exp(-\xi^i))^{-1} \tag{16}$$

$$u_q^i = \frac{(\beta_q^i)^2}{\sum_{k=1}^{M} (\beta_k^i)^2} \tag{17}$$

and minimizing E with respect to p_k^i and these new parameters. Calculation of the whole gradient can be performed in linear time with respect to the input dimension P, the number M of classes and the number n of prototypes [4].

4. SIMULATIONS

Iris data

The well-known Anderson iris data consist of 150 samples belonging to three classes (*iris virginica*, *iris versicolor* and *iris setosa*), in a four-dimensional feature space. Only two features (petal width and petal length) were considered in this analysis.

This example can be used to show that the description of uncertainty contained in the classifier's outputs allows for the rejection of input vectors that are too dissimilar from training patterns. Two classes (*iris virginica* and *versicolor*) were presented to a network with five hidden units. The network was trained by minimizing the quadratic error function E. The initial prototype locations were determined by a simple clustering method.

Figures 2 to 4 represent contour lines of two different functions:

- the credibility $Bel(D(\mathbf{x}))$ represents the weight of the evidence directly supporting the assignment of the input vector $\mathbf{x}$ to its predicted class $D(\mathbf{x})$; as expected, $Bel(D(\mathbf{x}))$ decreases with the distance to prototypes (Figure 2);

- the plausibility $Pl(D(\mathbf{x}))$ can be interpreted as the weight of the evidence that does not contradict the assignment of $\mathbf{x}$ to $D(\mathbf{x})$; this quantity is minimum in the vicinity of the boundary between classes, where the available evidence points simultaneously to several hypotheses with almost equal strength (Figure 3).

- the mass $m(\Omega)$ assigned to the set Ω of classes represents the amount of belief that could not be committed to any particular class because of the weakness of available information. It is an increasing function of the distance to training patterns (Figure 4).

As already proposed in [3], $Bel(D(\mathbf{x}))$ and $Pl(D(\mathbf{x}))$

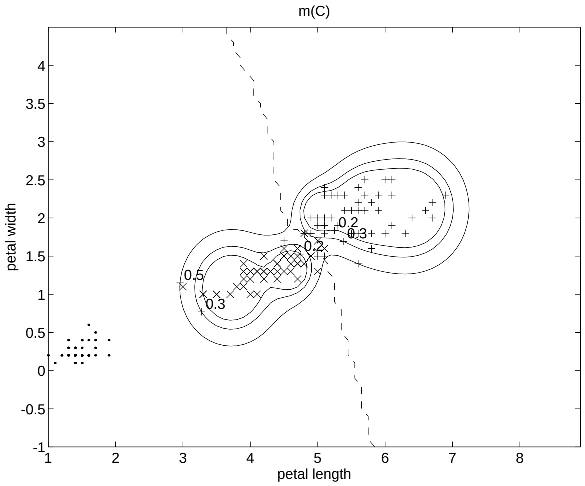

Figure 4: Contours of $m(\Omega)$ (–) and decision boundary (- -). Classes *iris virginica* (+) and *versicolor* (x) were presented during training.

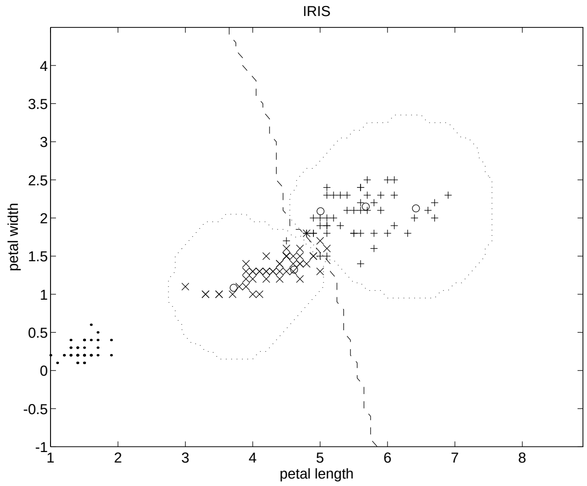

Figure 5: Distance reject boundary (:), classification boundary (- -) and prototypes (o). Classes *iris virginica* (+) and *versicolor* (x) were presented during training.

can be used to implement respectively the distance and ambiguity reject options proposed in [5]. The rejection threshold can be determined as the minimum value of $Bel(D(\mathbf{x}^s))$ for any training pattern $\mathbf{x}^s$ of class $D(\mathbf{x})$. The region of distance reject is then defined as:

$$Bel(D(\mathbf{x})) < \min_{\mathbf{x}^s \in \mathcal{X} \cap D(\mathbf{x})} Bel(D(\mathbf{x}^s)) \tag{18}$$

The contour of this region is represented as a dotted line in Figure 5. As can be seen, the class *iris setosa* that was not presented during training is correctly rejected. A similar result could be achieved by defining the distance reject rule as:

$$m(\Omega) > \max_{\mathbf{x}^s \in \mathcal{X} \cap D(\mathbf{x})} m^s(\Omega) \tag{19}$$

where m^s is the BPA computed for training pattern $\mathbf{x}^s$ using other training patterns.

Forensic glass data

This method was tested on several artificial and real-world data sets, including the "forensic glass data" [7]. This data set contains the description of 214 fragments of glass. Each fragment has a measured reflectivity index and chemical composition (weight percent of Na, Mg, Al, Si, K, Ca, Ba and Fe). As suggested by Ripley [9], 29 instances were discarded, and the remaining 185 were re-grouped in four classes: window float glass (70), window non-float glass (76), vehicle window glass (17) and other (22). The data set was split randomly in a training set of size 89 and a test set of size 96.

Three methods utilizing prototypes were compared on this data: Learning Vector Quantization (LVQ), RBF networks and our method (with normalized outputs and $\rho = 0$). Each method was provided with the same initial prototypes generated by a clustering procedure with random initial conditions. In RBF networks, the second layer of weights was initialized using a pseudo-inverse approach. The initial parameters for the evidence-theoretic method were $\gamma^i = 0.1$ and $\alpha^i = 0.5$ for $i = 1, \ldots, n$. The initial value of u_q^i for each prototype i and class ω_q was defined as the proportion of training samples belonging to class ω_q in the neighborhood of prototype i.

The RBF and evidence-theoretic networks were trained with the same optimization algorithm (gradient descent with adaptive learning rates). The number n of prototypes was varied between 5 and 9. For each value of n, each learning algorithm was run 10 times. The average test error rates and standard deviations are reported in Table 1. As can be seen, our method yields better results for small values of n, while the three methods have similar performance for higher values of n. The error rates obtained with our method are comparable to the best results reported by Ripley [9] in a comparative study of 13 classifiers.

Table 1: Test error rates for the Forensic glass example with different values of n (means of 10 runs with standard deviations). The methods are Learning Vector quantization (LVQ), Radial Basis Function (RBF) network and evidence-theoretic classifier (ETC)

n	5	6	7	8	9
LVQ	0.47	0.45	0.41	0.33	0.33
	(0.00)	(0.02)	(0.03)	(0.06)	(0.09)
RBF	0.38	0.38	0.38	0.35	0.34
	(0.01)	(0.01)	(0.01)	(0.03)	(0.02)
ETC	0.30	0.30	0.29	0.32	0.32
	(0.00)	(0.02)	(0.02)	(0.02)	(0.02)

5. CONCLUSION

The pattern classification method presented in this paper can be seen as combining ideas from supervised neural network models with local representation, and the evidence-theoretic k-NN rule introduced in [3]. As in RBF networks, classification is based on the similarity between the input pattern and a set of reference patterns. This information is converted in the form of BPAs, which are then combined using Dempster's rule of combination. The classifier output can be presented as a belief function defining for each class a posterior probability interval. The width of this interval can be used as a measure of the uncertainty attached to the classification. This approach has been shown to allow decision making with reject options, and to have good classification performance as compared to other methods.

6. REFERENCES

[1] A. P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[2] T. Denœux. Application of evidence theory to k-NN pattern classification. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice IV*, pages 13–24. Elsevier, Amsterdam, 1994.

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

[4] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *Submitted to IEEE Trans. on Systems, Man and Cybernetics*, 1995.

[5] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

[6] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[7] P. M. Murphy and D. W. Aha. *UCI Repository of machine learning databases [Machine-readable data repository]*. University of California, Department of Information and Computer Science., Irvine, CA, 1994.

[8] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, M.I.T, 1988.

[9] B. D. Ripley. Flexible non-linear approaches to classification. In V. Cherkassly, J. H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks*, pages 105–126. Springer Verlag, Berlin, 1995.

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

[11] P. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[12] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

[13] L. M. Zouhal and T. Denœux. An adaptive k-NN rule based on Dempster-Shafer theory. In *Proceedings of the 6th Int. Conf. on Computer Analysis of Images and Patterns (CAIP'95)*, Prague, September 1995.

[14] L. M. Zouhal and T. Denœux. A comparison between fuzzy and evidence-theoretic k-NN rules for pattern recognition. In *Proceedings of EUFIT'95*, Aachen, August 1995.