A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory

Thierry Denœux

***Abstract*—In this paper, the problem of classifying an unseen pattern on the basis of its nearest neighbors in a recorded data set is addressed from the point of view of Dempster-Shafer theory. Each neighbor of a sample to be classified is considered as an item of evidence that supports certain hypotheses regarding the class membership of that pattern. The degree of support is defined as a function of the distance between the two vectors. The evidence of the k nearest neighbors is then pooled by means of Dempster's rule of combination. This approach provides a global treatment of such issues as ambiguity and distance rejection, and imperfect knowledge regarding the class membership of training patterns. The effectiveness of this classification scheme as compared to the voting and distance-weighted k-NN procedures is demonstrated using several sets of simulated and real-world data.**

I. Introduction

In classification problems, complete statistical knowledge regarding the conditional density functions of each class is rarely available, which precludes application of the optimal Bayes classification procedure. When no evidence supports one form of the density functions rather than another, a good solution is often to build up a collection of correctly classified samples, called the *training set*, and to classify each new pattern using the evidence of nearby sample observation. One such non-parametric procedure has been introduced by Fix and Hodges [11], and has since become well-known in the Pattern Recognition literature as the voting k-nearest neighbor (k-NN) rule. According to this rule, an unclassified sample is assigned to the class represented by a majority of its k nearest neighbors in the training set. Cover and Hart [4] have provided a statistical justification of this procedure by showing that, as the number N of samples and k both tend to infinity in such a manner that $k/N \rightarrow 0$, the error rate of the k-NN rule approaches the optimal Bayes error rate. Beyond this remarkable property, the k-NN rule owes much of its popularity in the Pattern Recognition community to its good performance in practical applications. However, in the finite sample case, the voting k-NN rule is not guaranteed to be the optimal way of using the information contained in the neighborhood of unclassified patterns. This is the reason why the improvement of this rule has remained an active research topic in the past 40 years.

The main drawback of the voting k-NN rule is that it implicitly assumes the k nearest neighbors of a data point x to be contained in a region of relatively small volume, so that sufficiently good resolution in the estimates of the different conditional densities can be obtained. In practice, however, the distance between x and one of its closest neighbors is not always negligible, and can even become very large outside the regions of high density. This has several consequences. First, it can be questioned whether it is still reasonable in that case to give all the neighbors an equal weight in the decision, regardless of their distances to the point x to be classified. In fact, given the k nearest neighbors $x^{(1)}, \cdots, x^{(k)}$ of x, and $d^{(1)}, \cdots, d^{(k)}$ the corresponding distances arranged in increasing order, it is intuitively appealing to give the label of $x^{(i)}$ a greater importance than to the label of $x^{(j)}$ whenever $d^{(i)} < d^{(j)}$. Dudani [10] has proposed to assign to the ith nearest neighbor $x^{(i)}$ a weight $w^{(i)}$ defined as:

$$w^{(i)} = \frac{d^{(k)} - d^{(i)}}{d^{(k)} - d^{(1)}} \qquad d^{(k)} \neq d^{(1)} \tag{1}$$

$$= 1 \qquad d^{(k)} = d^{(1)}. \tag{2}$$

The unknown pattern x is then assigned to the class for which the weights of the representatives among the k nearest neighbors sum to the greatest value. This rule was shown by Dudani to be admissible, i.e. to yield lower error rates than those obtained using the voting k-NN procedure for at least one particular data set. However, several researchers, repeating Dudani's experiments, reached less optimistic conclusions [1], [16], [6]. In particular, Baily and Jain [1] showed that the distance-weighted k-NN rule is not necessarily better than the majority rule for small sample size if ties are broken in a judicious manner. These authors also showed that, in the infinite sample case ($N \rightarrow \infty$), the error rate of the traditional unweighted k-NN rule is better than that of any weighted k-NN rule. However, Macleod et al. [15] presented arguments showing that this conclusion may not apply if the training set is finite. They also proposed a simple extension of Dudani's rule allowing for a more effective use of the kth neighbor in the classification.

Apart from this discussion, it can also be argued that, because the weights are constrained to span the interval [0, 1], the distance-weighted k-NN procedure can still give considerable importance to observations that are very dissimilar to the pattern to be classified. This represents a serious drawback when all the classes cannot be assumed to be represented in the training set, as is often the case in some application areas as target recognition in noncooperative environments [5] or diagnostic problems [9]. In such situations, it may be wise to consider that a point that is far away from any

Manuscript received September 23, 1993; revised July 17, 1994. This work was partially supported by EEC funded Esprit Project 6757 EMS (Environmental Monitoring System).

The author is at the Université de Technologie de Compiègne, U.R.A. CNRS 817 Heudiasyc, BP 649 F-60206 Compiègne cedex, France.

IEEE Log Number 9409222.

0018–9472/95$04.00 © 1995 IEEE

previously observed pattern most probably belongs to an unknown class for which no information has been gathered in the training set, and should therefore be rejected. Dubuisson and Masson [9] call *distance reject* this decision, as opposed to the *ambiguity reject* introduced by Chow [3] and for which an implementation in a k-NN rule has been propoposed by Hellman [12]. Dasarathy [5] has proposed a k-NN rule where a distance reject option is made possible by the introduction of the concept of an *acceptable neighbor*, defined as a neighbor whose distance to the pattern to be classified is smaller than some threshold learnt from the training set. If there is less than some predefined number of acceptable neighbors of one class, the pattern is rejected and later considered for assignment to a new class using a clustering procedure.

Another limitation of the voting k-NN procedure is that it offers no obvious way to cope with uncertainty or imprecision in the labelling of the training data. This may be a major problem in some practical applications, as in the diagnostic domain, where the true identity of training patterns is not always known, or even defined, unambiguously, and has to be determined by an expert or via an automatic procedure that is itself subject to uncertainty. From a slightly different point of view, it may also be argued that patterns, even correctly labelled, have some degree of "typicality" depending on their distance to class centers, and that atypical vectors should be given less weight in the decision than those that are truly representative of the clusters [14]. Fuzzy sets theory offers a convenient formalism for handling imprecision and uncertainty in a decision process, and several fuzzy k-NN procedures have been proposed [13], [14]. In this approach, the degree of membership of a training vector x to each of M classes is specified by a number of u_i, with the following properties:

$$u_i \in [0, 1] \tag{3}$$

$$\sum_{i=1}^{M} u_i = 1. \tag{4}$$

The membership coefficients u_i can be given (e.g. by experts) or computed using the neighbors of each vector in the training set [14]. The membership of an unseen pattern in each class is then determined by combining the memberships of its neighbors. Keller et al. [14] have proposed a rule in which membership assignment is a function of both the vector's distance from its k nearest neighbors, and those neighbors' memberships in the possible classes. Beyond an improvement in classification performance over the crisp k-NN procedure, this approach allows a richer information content of the classifier's output by providing membership values that can serve as a confidence measure in the classification.

In this paper, a new classification procedure using the nearest neighbors in a data set is introduced. This procedure provides a global treatment of important issues that are only selectively addressed in the aforementioned methods, namely: the consideration of the distances from the neighbors in the decision, ambiguity and distance rejection, and the consideration of uncertainty and imprecision in class labels. This is achieved by setting the problem of combining the evidence provided by nearest neighbors in the conceptual framework of Dempster-Shafer (D-S) theory. As will be seen, this formalism presents the advantage of permitting a clear distinction between the presence of conflicting information—as happens when a pattern is close to several training vectors from different classes—and the scarcity of information—when a pattern is far away from any pattern in the training set, or close to patterns whose class memberships are not defined precisely. In the following section, the basics of D-S theory are recalled. The application to a new k-NN procedure is then described, and experimental results are presented.

II. Dempster-Shafer Theory

Let Θ be a finite set of mutually exclusive and exhaustive hypotheses about some problem domain, called the *frame of discernment* [19]. It is assumed that one's total belief induced by a body of evidence concerning Θ can be partitioned into various portions, each one assigned to a subset of Θ. A *basic probability assignment* (BPA) is a function m from 2^Θ, the power set of Θ, to [0, 1], verifying:

$$m(\emptyset) = 0 \tag{5}$$

$$\sum_{A \subseteq \Theta} m(A) = 1. \tag{6}$$

The quantity m(A), called a *basic probability number*, can be interpreted as a measure of the belief that one is willing to commit *exactly* to A, and not to any of its subsets, given a certain piece of evidence. A situation of total ignorance is characterized by $m(\Theta) = 1$.

Intuitively, a portion of belief committed to a hypothesis A must also be committed to any hypothesis it implies. To obtain the total belief in A, one must therefore add to $m(A)$ the quantities $m(B)$ for all subsets B of A. The function assigning to each subset A of Θ the sum of all basic probability numbers for subsets of A is called a *belief function*:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{7}$$

$Bel(A)$, also called the *credibility* of A, is interpreted as a measure of the total belief committed to A. The subsets A of Θ such that $m(A) > 0$ are called the *focal elements* of the belief function, and their union is called its *core*. The *vacuous* belief function has Θ for only focal element, and corresponds to complete ignorance. Other noticeable types of belief functions are *Bayesian* belief functions, whose focal elements are singletons, and *simple support* functions, that have only one focal element in addition of Θ.

It can easily be verified that the belief in some hypothesis A and the belief in its negation $\overline{A}$ do not necessarily sum to 1, which is a major difference with probability theory. Consequently, $Bel(A)$ does not reveal to what extent one believes in $\overline{A}$, i.e. to what extent one doubts A, which is described by $Bel(\overline{A})$. The quantity $Pl(A) = 1 - Bel(\overline{A})$, called the *plausibility* of A, defines to what extent one fails to doubt in A, i.e. to what extent one finds A *plausible*. It is straightforward to show that:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{8}$$

As demonstrated by Shafer [19], any one of the three functions m, Bel and Pl is sufficient to recover the other two. This follows from the definition of $Pl(A)$ as $1 - Bel(\overline{A})$, and:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B). \tag{9}$$

A BPA can also be viewed as determining a set of probability distributions P over 2^{Θ} satisfying:

$$Bel(A) \leq P(A) \leq Pl(A) \tag{10}$$

for all $A \subseteq \Theta$. For that reason, Bel and Pl are also called *lower* and *upper* probabilities, respectively. This fundamental imprecision in the determination of the probabilities reflects the "weakness", or incompleteness of the available information. The above inequalities reduce to equalities in the case of a Bayesian belief function.

Given two belief functions Bel_1 and Bel_2 over the same frame of discernment, but induced by two independent sources of information, we must define a way by which, under some conditions, these belief functions can be combined into a single one. Dempster's rule of combination is a convenient method for doing such pooling of evidence. First, Bel_1 and Bel_2 have to be *combinable*, i.e. their cores must not be disjoint. If m_1 and m_2 are the BPAs associated with Bel_1 and Bel_2, respectively, this condition can also be expressed as:

$$\sum_{A \cap B = \emptyset} m_1(A) m_2(B) < 1. \tag{11}$$

If Bel_1 and Bel_2 are combinable, then the function $m : 2^{\Theta} \mapsto [0, 1]$, defined by:

$$m(\emptyset) = 0 \tag{12}$$

$$m(\theta) = \frac{\Sigma_{A \cap B = \theta} m_1(A) m_2(B)}{1 - \Sigma_{A \cap B = \emptyset} m_1(A) m_2(B)} \qquad \theta \neq \emptyset \tag{13}$$

is a BPA. The belief function Bel given by m is called the orthogonal sum of Bel_1 and Bel_2, and is denoted $Bel_1 \oplus Bel_2$. For convenience, m will also be denoted $m_1 \oplus m_2$. The core of Bel equals the intersection of the cores of Bel_1 and Bel_2.

Although Dempster's rule is hard to justify theoretically, it has some attractive features, such as the following: it is commutative and associative; given two belief functions Bel_1 and Bel_2, if Bel_1 is vacuous, then $Bel_1 \oplus Bel_2 = Bel_2$; if Bel_1 is Bayesian, and if $Bel_1 \oplus Bel_2$ exists, then it is also Bayesian.

The D-S formalism must also be considered in the perspective of decision analysis [2]. As explained above, under D-S theory, a body of evidence about some set of hypotheses Θ does not in general provide a unique probability distribution, but only a set of *compatible* probabilities bounded by a belief function Bel and a plausibility function Pl. An immediate consequence is that simple hypotheses can no longer be ranked according to their probability: in general, the rankings produced by Bel and Pl will be different. This means that, as a result of lack of information, the decision is, to some extent, indeterminate. The theory does not make a choice between two distinct strategies: select the hypothesis with the greatest degree of belief—the most *credible*, or select the hypothesis with the lowest degree of doubt—the most *plausible*.

This analysis can be extended to decision with costs. In the framework of D-S theory, there is nothing strictly equivalent to Bayesian expected costs, leading unambiguously to a single decision. It is however possible to define lower and upper bounds for these costs, in the following way [7], [2]. Let M be the number of hypotheses, and U be an $M \times M$ matrix such that $U_{i,j}$ is the cost of selecting hypothesis θ_i if hypothesis θ_j is true. Then, for each simple hypothesis $\theta_i \in \Theta$, a *lower* expected cost $E_*[\theta_i]$ and an *upper* expected cost $E^*[\theta_i]$ can be defined:

$$E_*[\theta_i] = \sum_{A \subseteq \Theta} m(A) \min_{\theta_j \in A} U_{i,j} \tag{14}$$

$$E^*[\theta_i] = \sum_{A \subseteq \Theta} m(A) \max_{\theta_j \in A} U_{i,j}. \tag{15}$$

The lower (respectively: upper) expected cost can be seen as being generated by a probability distribution compatible with m, and such that the density of $m(A)$ is concentrated at the element of A with the lowest (respectively: highest) cost. Here again, the choice is left open as to which criterion should be used for the decision. Maximizing the upper expected cost amounts to minimizing the worst possible consequence, and therefore generally leads to more conservative decisions. Note that, when U verifies:

$$U_{i,j} = 1 - \delta_{i,j} \tag{16}$$

where $\delta_{i,j}$ is the Kronecker symbol, the following equalities hold:

$$E_*[\theta_i] = 1 - Pl(\{\theta_i\}) \tag{17}$$

$$E^*[\theta_i] = 1 - Bel(\{\theta_i\}). \tag{18}$$

In the case of $\{0, 1\}$ costs, minimizing the lower (respectively: upper) expected cost is thus equivalent to selecting the hypothesis with the highest plausibility (respectively: credibility).

III. The Method

A. The Decision Rule

Let $\mathcal{X} = \{x^i = (x_1^i, \cdots, x_P^i) | i = 1, \cdots, N\}$ be a collection on N P-dimensional training samples, and $\mathcal{C} = \{C_1, \cdots, C_M\}$ be a set of M classes. Each sample x^i will first be assumed to possess a class label $L^i \in \{1, \cdots, M\}$ indicating with certainty its membership to one class in $\mathcal{C}$. The pair $(\mathcal{X}, \mathcal{L})$, where $\mathcal{L}$ is the set of labels, constitutes a training set that can be used to classify new patterns.

Let x^s be an incoming sample to be classified using the information contained in the training set. Classifying x^s means assigning it to one class in $\mathcal{C}$, i.e. deciding among a set of M hypotheses: $x^s \in C_q, q = 1, \ldots, M$. Using the vocabulary of D-S theory, $\mathcal{C}$ can be called the *frame of discernment* of the problem.

Let us denote by Φ^s the set of the k-nearest neighbors of x^s in $\mathcal{X}$, according to some distance measure (e.g. the euclidian one). For any $x^i \in \Phi^s$, the knowledge that $L^i = q$ can be regarded as a piece of evidence that increases our belief that x^s also belongs to C_q. However, this piece of evidence does not by itself provide 100% certainty. In D-S formalism, this can be expressed by saying that only some part of our belief is committed to C_q. Since the fact that $L^i = q$ does not point to any other particular hypothesis, the rest of our belief cannot be distributed to anything else than $\mathcal{C}$, the whole frame of discernment. This item of evidence can therefore be represented by a BPA $m^{s,i}$ verifying:

$$m^{s,i}(\{C_q\}) = \alpha \tag{19}$$

$$m^{s,i}(\mathcal{C}) = 1 - \alpha \tag{20}$$

$$m^{s,i}(A) = 0 \quad \forall A \in 2^{\Theta} \setminus \{\mathcal{C}, \{C_q\}\} \tag{21}$$

with $0 < \alpha < 1$.

If x^i is far from x^s, as compared to distances between neighboring points in C_q, the class of x^i will be considered as providing very little information regarding the class of x^s; in that case, α must therefore take on a small value. On the contrary, if x^i is close to x^s, one will be much more inclined to believe that x^i and x^s belong to the same class. As a consequence, it seems reasonable to postulate that α should be a decreasing function of $d^{s,i}$, the distance between x^s and x^i. Furthermore, if we note:

$$\alpha = \alpha_0 \phi_q(d^{s,i}) \tag{22}$$

where the index q indicates that the influence of $d^{s,i}$ may depend on the class of x^s, the following additional conditions must be imposed on α_0 and ϕ_q:

$$0 < \alpha_0 < 1 \tag{23}$$

$$\phi_q(0) = 1 \tag{24}$$

$$\lim_{d \to \infty} \phi_q(d) = 0. \tag{25}$$

The first two conditions indicate that, even if the case of a zero distance between x^i and x^s, one still does not have certainty that they belong to the same class. This results from the fact that several classes can, in general, simultaneously have non zero probability densities in some regions of the feature space. The third condition insures that, in the limit, as the distance between x^s and x^i gets infinitely large, the belief function given by $m^{s,i}$ becomes vacuous, which means that one's belief concerning the class of x^s is no longer affected by one's knowledge of the class of x^i.

There is obviously an infinitely large number of decreasing functions ϕ verifying (24) and (25), and it is very difficult to find any a priori argument in favor of one particular function or another. We suggest to choose ϕ_q as:

$$\phi_q(d) = e^{-\gamma_q d^{\beta}} \tag{26}$$

with $\gamma_q > 0$ and $\beta \in \{1, 2, \cdots\}$. β can be arbitrarily fixed to a small value (1 or 2). Simple heuristics for the choice of α_0 and γ_q will be presented later.

For each of the k-nearest neighbors of x^s, a BPA depending on both its class label and its distance to x^s can therefore be defined. In order to make a decision regarding the class assignment of x^s, these BPAs can be combined using Dempster's rule. Note that this is always possible, since all the associated belief functions have $\mathcal{C}$ as a focal element.

Let us first consider two elements x^i and x^j of Φ^s belonging to the same class C_q. The BPA $m^{s,(i,j)} = m^{s,i} \oplus m^{s,j}$ resulting from the combination of $m^{s,i}$ and $m^{s,j}$ is given by:

$$m^{s,(i,j)}(\{C_q\}) = 1 - (1 - \alpha_0 \phi_q(d^{s,i}))(1 - \alpha_0 \phi_q(d^{s,j})) \tag{27}$$

$$m^{s,(i,j)}(\mathcal{C}) = (1 - \alpha_0 \phi_q(d^{s,i}))(1 - \alpha_0 \phi_q(d^{s,j})). \tag{28}$$

If we denote by Φ_q^s the set of the k-nearest neighbors of x^s belonging to C_q, and assuming that $\Phi_q^s \neq \emptyset$, the result of the combination of the corresponding BPAs $m_q^s = \bigoplus_{x^i \in \Phi_q^s} m^{s,i}$ is given by:

$$m_q^s(\{C_q\}) = 1 - \prod_{x^i \in \Phi_q^s} (1 - \alpha_0 \phi_q(d^{s,i})) \tag{29}$$

$$m_q^s(\mathcal{C}) = \prod_{x^i \in \Phi_q^s} (1 - \alpha_0 \phi_q(d^{s,i})). \tag{30}$$

If $\phi_q^s = \emptyset$, then m_q^s is simply the BPA associated with the vacuous belief function: $m_q^s(\mathcal{C}) = 1$.

Combining all the BPAs m_q^s for each class, a global BPA $m^s = \bigoplus_{q=1}^{M} m_q^s$ is obtained as:

$$m^s(\{C_q\}) = \frac{m_q^s(\{C_q\}) \prod_{r \neq q} m_r^s(\mathcal{C})}{K} \qquad q = 1, \cdots, M \tag{31}$$

$$m^s(\mathcal{C}) = \frac{\prod_{q=1}^{M} m_q^s(\mathcal{C})}{K} \tag{32}$$

where K is a normalizing factor:

$$K = \sum_{q=1}^{M} m_q^s(\{C_q\}) \prod_{r \neq q} m_r^s(\mathcal{C}) + \prod_{q=1}^{M} m_q^s(\mathcal{C}) \tag{33}$$

$$= \sum_{q=1}^{M} \prod_{r \neq q} m_r^s(\mathcal{C}) + (1 - M) \prod_{q=1}^{M} m_q^s(\mathcal{C}). \tag{34}$$

The focal elements of the belief function associated with m^s are the classes represented among the k-nearest neighbors of x^s, and $\mathcal{C}$. The credibility and plausibility of a given class C_q are:

$$Bel^s(\{C_q\}) = m^s(\{C_q\}) \tag{35}$$

$$Pl^s(\{C_q\}) = m^s(\{C_q\}) + m^s(\mathcal{C}). \tag{36}$$

Therefore, both criteria produce the same ranking of hypotheses concerning x^s.

If an $M \times M$ cost matrix U is given, where $U_{i,j}$ is the cost of assigning an incoming pattern to class i, if it actually belongs to class j, then lower and upper expected costs are defined for each possible decision:

$$E_*[C_q] = \sum_{A \subseteq \mathcal{C}} m^s(A) \min_{C_r \in A} U_{q,r} \tag{37}$$

$$= \sum_{r=1}^{M} m^s(\{C_r\}) U_{q,r} + m^s(\mathcal{C}) \min_{C_r \in \mathcal{C}} U_{q,r} \tag{38}$$

$$E^*[C_q] = \sum_{A \subseteq C} m^s(A) \max_{C_r \in A} U_{q,r} \tag{39}$$

$$= \sum_{r=1}^{M} m^s(\{C_r\}) U_{q,r} + m^s(\mathcal{C}) \max_{C_r \in C} U_{q,r}. \tag{40}$$

Note that minimizing the lower or upper expected cost do not necessarily lead to the same decision, as can be seen from the following example. Let us consider the problem of assigning an incoming sample x^s to one of three classes ($M = 3$). It is assumed that the consideration of the k-nearest neighbors of x^s has produced a BPA m^s such that $m^s(\{C_1\}) = 0.2, m^s(\{C_2\}) = 0, m^s(\{C_3\}) = 0.4$ and $m^s(\mathcal{C}) = 0.4$. The cost matrix is:

$$U = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix}.$$

The lower and upper expected costs are, in that case:

$$E_*[C_1] = 0.4 \quad E_*[C_2] = 0.6 \quad E_*[C_3] = 0.2$$
$$E^*[C_1] = 0.8 \quad E^*[C_2] = 1.0 \quad E^*[C_3] = 1.0.$$

Thus, C_3 minimizes E_*, while C_1 minimizes E^*.

However, in the case of $\{0, 1\}$ costs, that will exclusively be considered henceforth, minimizing the lower (resp. upper) expected cost amounts to maximizing the plausibility (resp. credibility). In that case, and under the assumption that the true class membership of each training pattern is known, both criteria therefore lead to the same decision rule D:

$$q^s_{\max} = \arg\max_p m^s(\{C_p\}) \Rightarrow D(x^s) = q^s_{\max} \tag{41}$$

where $D(x^s)$ is the class label assigned to x^s.

Note that the consideration of the distances makes the probability of a tie taking place much smaller than in the simple majority rule, whose relationship with D can also be described by the following theorem:

Theorem 1: If the k nearest neighbors of a data point x^s are located at the same distance of x^s, and if $\phi_1 = \phi_2 = \cdots = \phi_M = \phi$, then the decision rule D produces the same decision as the majority rule.

Proof: Let us denote by ℓ the distance between x^s and all of its k nearest neighbors $x^i \in \Phi^s$. For all $q \in \{1, \ldots, M\}$, m^s_q is defined by:

$$m^s_q(\{C_q\}) = 1 - (1 - \alpha_0 \phi(\ell))^{|\Phi^s_q|} \tag{42}$$

$$m^s_q(\mathcal{C}) = (1 - \alpha_0 \phi(\ell))^{|\Phi^s_q|}. \tag{43}$$

Thus:

$$m^s(\{C_q\}) = \frac{(1 - (1 - \alpha_0 \phi(\ell))^{|\Phi^s_q|})(1 - \alpha_0 \phi(\ell))^{k - |\phi^s_q|}}{K}$$
$$q \in \{1, \cdots, M\} \tag{44}$$

$$m^s(\mathcal{C}) = \frac{(1 - \alpha_0 \phi(\ell))^k}{K}. \tag{45}$$

For any p and q in $\{1, \cdots, M\}$ such that $m^s(\{C_q\}) > 0$, we have:

$$\frac{m^s(\{C_p\})}{m^s(\{C_q\})} = \frac{(1 - \alpha_0 \phi(\ell))^{k - |\phi^s_p|} - (1 - \alpha_0 \phi(\ell))^k}{(1 - \alpha_0 \phi(\ell))^{k - |\phi^s_q|} - (1 - \alpha_0 \phi(\ell))^k}. \tag{46}$$

Therefore:

$$m^s(\{C_p\}) > m^s(\{C_q\}) \Leftrightarrow k - |\phi^s_p| < k - |\phi^s_q| \tag{47}$$
$$\Leftrightarrow |\phi^s_p| > |\phi^s_q|. \tag{48}$$

□

B. Reject Options

The decision rule D can easily be modified so as to include ambiguity and distance reject options. The ambiguity reject option, as introduced by Chow [3] consists in postponing decision-making when the conditional error of making a decision given x^s is high. This situation typically arises in regions of the feature space where there is a strong overlap between classes. In that case, an incoming sample x^s to be classified will generally be close to several training vectors belonging to different classes. Hence, this can be viewed as a problem of conflicting information.

The distance reject option discussed in [9] corresponds to a different situation, where the point x^s to be classified is far away from any previously recorded sample, and is therefore suspected of belonging to a class that is not represented in the training set. The problem here no longer arises from conflict in the data, but from the weakness or scarcity of available information.

In our framework, the first case will be characterized by a BPA m^s that will be uniformly distributed among several classes. As a consequence, both the maximum plausibility $Pl^s(\{C_{q^s_{\max}}\})$ and the maximum credibility $Bel^s(\{C_{q^s_{\max}}\})$ will take on relatively low values. In the second case, most of the probability mass will be concentrated on the whole frame of discernment $\mathcal{C}$. As a consequence, only $Bel^s(\{C_{q^s_{\max}}\})$ will take on a small value; as the distance between x^s and its closest neighbor goes to infinity, $Bel^s(\{C_{q^s_{\max}}\})$ actually goes to zero, while $Pl^s(\{C_{q_{\max}}\})$ goes to one.

As a result, it is possible to introduce ambiguity and distance reject options by imposing thresholds $Pl_{\min}$ and $Bel_{\min}$ on the plausibility and credibility, respectively. The sample x^s will be ambiguity rejected if $Pl^s(\{C_{q^s_{\max}}\}) < Pl_{min}$, and it will be distance rejected if $Bel^s(\{C_{q^s_{\max}}\}) < Bel_{\min}$. Note that, in the case of $\{0, 1\}$ costs, these thresholds correspond to thresholds $E_{*\max}$ and $E^*_{\max}$ on the lower and upper expected costs, respectively:

$$E_{*\max} = 1 - Pl_{\min} \tag{49}$$

$$E^*_{\max} = 1 - Bel_{\min}. \tag{50}$$

The determination of $Pl_{\min}$ has to be based on a trade-off between the probabilities of error and reject, and must therefore be left to the designer of the system. The choice of $Bel_{\min}$ is more problematic, since no unknown class is, by definition, initially included in the training set. A reasonable approach is to compute $Bel^i(\{C_{q^i_{\max}}\})$ for each x^i in the training set using the leave-one-out method, and define a distinct threshold $Bel^q_{\min}$ for each class C_q as:

$$Bel^q_{\min} = \min_{x^i \in \mathcal{X}, L^i = q} Bel^i(\{C_{q^i_{\max}}\}). \tag{51}$$

C. Imperfect Labelling

In some applications, it may happen that one only has imperfect knowledge concerning the class membership of some training patterns. For example, in a three class problem, an expert may have some degree of belief that a sample x^i belongs to a class C_3, but still consider as possible that it might rather belong to C_1 or C_2. Or, he may be sure that x^i does not belong to C_3, while being totally incapable of deciding between C_1 and C_2. In D-S formalism, one's belief in the class membership of each training pattern x^i can be represented by a BPA m^i over the frame of discernment $\mathcal{C}$. For example, if the expert is sure that x^i does not belong to C_3, has no element to decide between C_1 and C_2, and evaluates the chance of his assessment being correct at 80%, then his belief can be represented in the form of a BPA as:

$$m^i(\{C_1, C_2\}) = 0.8 \tag{52}$$

$$m^i(\mathcal{C}) = 0.2 \tag{53}$$

with all remaining $m^i(A)$ values equal to zero.

The approach described in above can easily be generalized so as to make use of training patterns whose class membership is represented by a BPA. If x^s is a sample to be classified, one's belief about the class of x^s induced by the knowledge that $x^i \in \Phi^s$ can be represented by a BPA $m^{s,i}$ deduced from m^i and $d^{s,i}$:

$$m^{s,i}(A) = \alpha_0\phi(d^{s,i})m^i(A) \qquad \forall A \in 2^{\mathcal{C}} \setminus \mathcal{C} \tag{54}$$

$$m^{s,i}(\mathcal{C}) = 1 - \sum_{A \in 2^{\mathcal{C}} \setminus \mathcal{C}} m^{s,i}(A) \tag{55}$$

where ϕ is a monotonically decreasing function verifying (24) and (25).

As before, the $m^{s,i}$ can then be combined using Dempster's rule to form a global BPA:

$$m^s = \bigoplus_{x^i \in \Phi^s} m^{s,i}. \tag{56}$$

Note that, while the amount of computation needed to implement Dempster's rule increases only linearly with the number of classes when the belief functions given by the $m^{s,i}$ are simple support functions as considered in Section III.A, the increase is exponential is the worst general case. However, more computationally efficient approximation methods such as proposed in [21] could be used for very larger numbers of classes.

IV. Experiments

The approach described in this paper has been successfully tested on several classification problems. Before presenting the results of some of these experiments, practical issues related to the implementation of the procedure need to be addressed.

Leaving alone the rejection thresholds, for which a determination method has already been proposed, and assuming an exponential form for ϕ_q as described in (26), the following parameters have to be fixed in order to allow the pratical use of the method: $k, \alpha_0, \gamma_q, q = 1, \cdots, M$ and β.

As in the standard k-NN procedure, the choice of k is difficult to make *a priori*. Although our method seems to be far less sensitive to this parameter than the majority rule, a systematic search for the best value of k may be necessary in order to obtain optimal results.

For the choice of α_0 and γ_q, several heuristics have been tested. Good results on average have been obtained with $\alpha_0 = 0.95$ and γ_q determined seperately for each class as $1/d_q^\beta$, where d_q is the mean distance between two training vectors belonging to class C_q.[1] The value of β has been found to have very little influence on the performance of the method. A value of $\beta = 1$ has been adopted in our simulations.

The following examples are intended to illustrate various aspects of our method, namely: the shape of the decision boundaries and reject regions for simple two-dimensional data sets, the relative performance as compared to the voting and distance-weighted k-NN rules for different values of k, and the effect of imperfect labelling.

A. Experiment 1

The purpose of this experiment is to visualize the decision boundary and the regions of ambiguity and distance reject for two different two-dimensional data sets of moderate size. The first data set is taken from two Gaussian distributions with the following characteristics:

$$\mu_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

$$\Sigma_1 = 0.25I \quad \Sigma_2 = I$$

where I is the identity matrix. There are 40 training samples in each class.

The second data set consists of two non-gaussian classes of 40 samples each separated by a non-linear boundary. Both data sets are represented in the Figs. 1–4, together with the lines of equal maximum credibility $Bel^s(\{C_{q^s_{\max}}\})$ and plausibility $Pl^s(\{C_{q^s_{\max}}\})$, for $k = 9$. As expected, the region of low plausibility is concentrated in each case around the class boundary, allowing for ambiguity reject, whereas small credibility values are obtained in the regions of low probability density. The distance reject regions, as defined in Section III.B, are delimited by dotted lines.

For the first data set, the estimated error rate obtained using an independent test set of 1000 samples is 0.084, against 0.089 for the voting 9-NN rule. The corresponding results for the second data set and leave-one-out error estimation are 0.075 for both methods.

B. Experiment 2

A comparison between the performances of the voting k-NN procedure, the distance-weighted k-NN rule and our method was performed using one artificial and two real-world classification problems. In the majority rule, ties were resolved by randomly selecting one of the tied pattern classes.

[1] This heuristic was suggested to me by Lalla Meriem Zouhal.

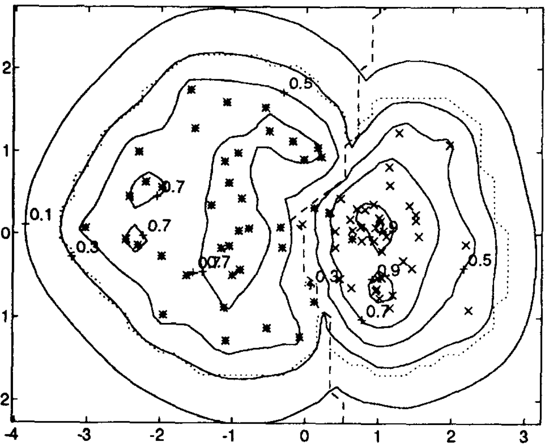


Fig. 1. Lines of equal maximum credibility $(Bel^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (Gaussian data). Samples falling outside the region delimited by the dotted line are distance rejected.

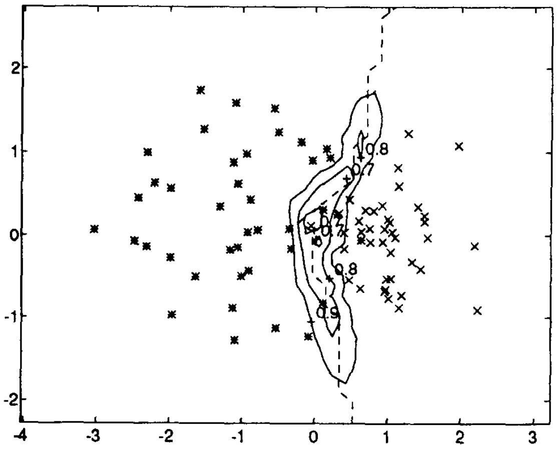


Fig. 2. Lines of equal maximum plausibility $(Pl^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (Gaussian data).

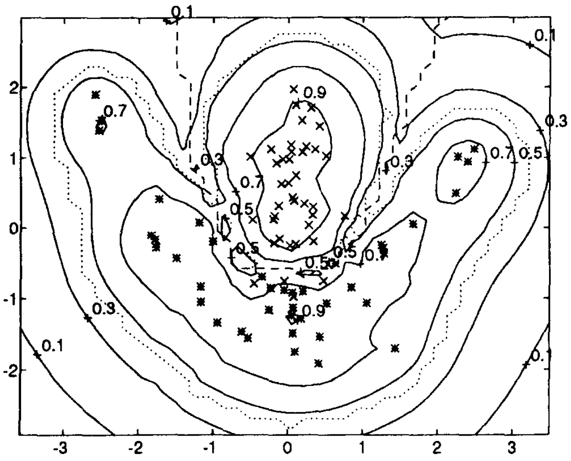


Fig. 3. Lines of equal maximum credibility $(Bel^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (non-gaussian data). Samples falling outside the region delimited by the dotted line are distance rejected.

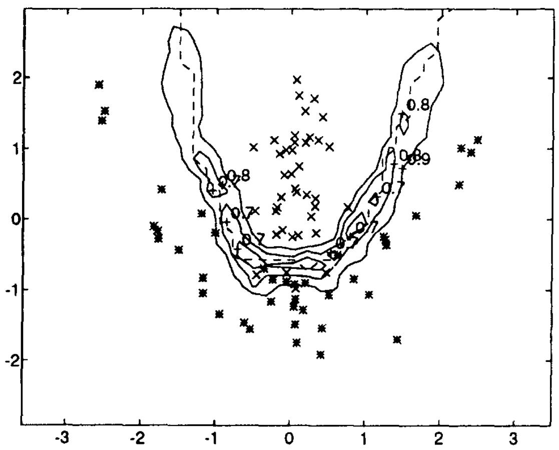


Fig. 4. Lines of equal maximum plausibility $(Pl^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (non-gaussian data).

The first problem implies three gaussian distributions in a three-dimensional space, with the following characteristics:

$$\mu_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \mu_3 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

$$\Sigma_1 = I \qquad \Sigma_2 = I \qquad \Sigma_3 = 2I$$

Training sets of 30, 60, 120 and 180 samples have been generated using prior probabilities (1/3, 1/3, 1/3). Values of k ranging from 1 to 25 have been investigated. A test set of 1000 samples has been used for error estimation. For each pair (N, k), the reported error rates are averages over 5 trials performed with 5 independent training sets. The results are presented in Table I and Figs. 5–8.

The second data set is composed of real-world data obtained by recording examples of the eleven steady state vowels of English spoken by fifteen speakers [8], [18]. Words containing each of these vowels were uttered once by the fifteen speakers. Four male and four female speakers were used to build a training set, and the other four male and three female speakers were used for building a test set. After suitable preprocessing, 568 training patterns and 462 test patterns in a 10 dimensional input space were collected. Fig. 9 shows the test error rates for the three methods with k ranging from 1 to 30.

The third task investigated concerns the classification of radar returns from the ionosphere obtained by a radar system consisting of a phased array of 16 high-frequency antennas [17], [20]. The targets were free electrons in the ionosphere. Radar returns were manually classified as "good" or "bad" depending on whether or not they showed evidence of some type of structure in the ionosphere. Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. This processing yielded 34 continuous attributes for each of the 351 training instances collected. The classification results for different values of k are described in Fig. 10. The figures shown are leave-one-out estimates of the error rates, computed using the training data.

TABLE I

RESULTS OF THE SECOND EXPERIMENT (GAUSSIAN DATA, 1000 TEST SAMPLES) FOR THE VOTING k-NN RULE (k-NN), THE DISTANCE-WEIGHTED k-NN RULE (WEIGHTED k-NN) AND OUR METHOD (D-S): BEST ERROR RATES (MEANS OVER 5 RUNS) WITH CORRESPONDING VALUES OF k (UPPER NUMBERS) AND AVERAGE ERROR RATES INTEGRATED OVER THE DIFFERENT VALUES OF k (LOWER NUMBER)

	Classification rule		
	k-NN	weighted k-NN	Dempster-Shafer
$N = 30$	0.326 (5) 0.397	0.299 (16) 0.338	0.267 (15) 0.306
$N = 60$	0.309 (8) 0.335	0.293 (21) 0.314	0.260 (23) 0.284
$N = 120$	0.296 (7) 0.306	0.277 (25) 0.300	0.254 (22) 0.280
$N = 180$	0.280 (18) 0.296	0.267 (14) 0.293	0.249 (23) 0.273

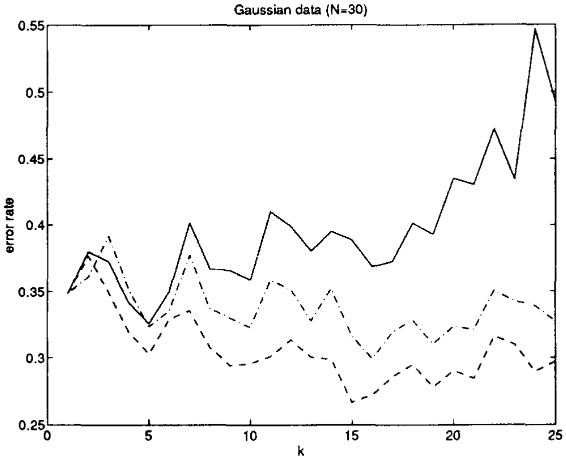


Fig. 5. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 30$).

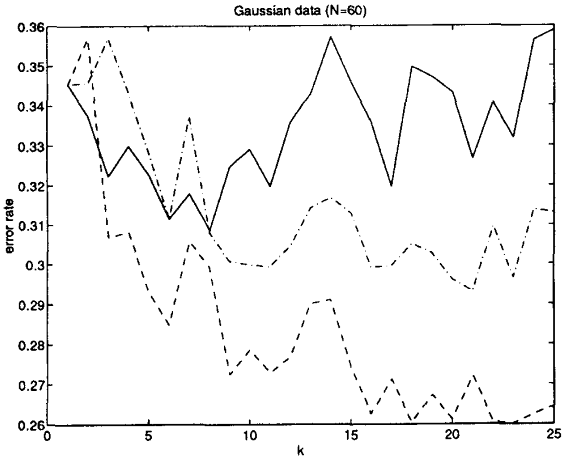


Fig. 6. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 60$).

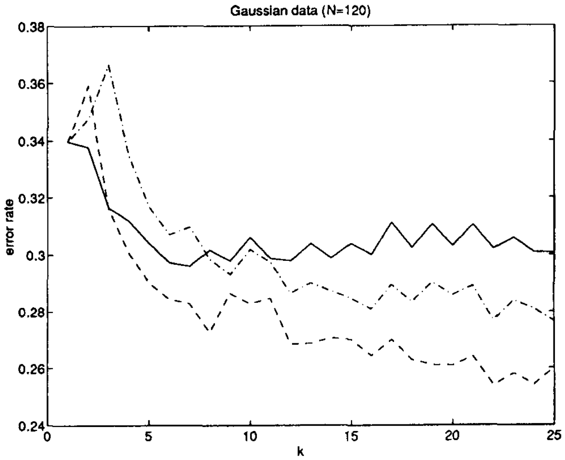


Fig. 7. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 120$).

Not surprisingly, the performances of the two methods taking into account distance information are better than that of the voting k-NN rule, for the three classification problems investigated. Whereas the error rate of the voting k-NN rule passes by a minimum for some problem-dependent number of neighbors, the results obtained by the two other methods appear to be much less sensitive to the value of k, provided k is chosen large enough. Our method clearly outperforms the distance-weighted approach on the Gaussian data sets and the vowel recognition task. Both methods are almost equivalent on the ionosphere data.

C. Experiment 3

In order to study the behavior of our method in case of imperfect labelling, the following simulation has been performed. A data set of 120 training samples has been generated using the three gaussian distributions of the previous experiment. For each training vector x^i, a number p^i has been generated using a uniform distribution on [0, 1]. With probability p^i, the label of x^i has been changed (to any of the other two classes with equal probabilities). Denoting by L^i the new class label of x^i, and assuming that $L^i = q$, then the BPA m^i describing the class membership of x^i has been defined as:

$$m^i(\{C_q\}) = 1 - p^i \tag{57}$$

$$m^i(\mathcal{C}) = p^i \tag{58}$$

and $m^i(A) = 0$ for all other $A \subseteq \mathcal{C}$. Hence, $m^i(\mathcal{C})$ is an indication of the reliability of the class label of x^i. Using the D-S formalism, it is possible to make use of this information, by giving less importance to those training vectors whose class membership is uncertain. This property can be expected to result in a distinctive advantage over the majority rule in a situation of this kind.

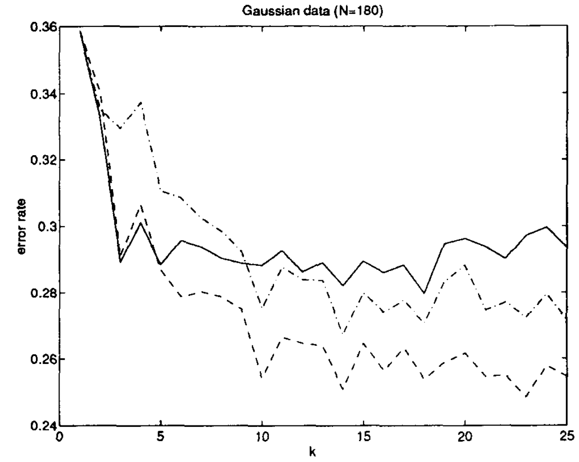


Fig. 8. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 180$).

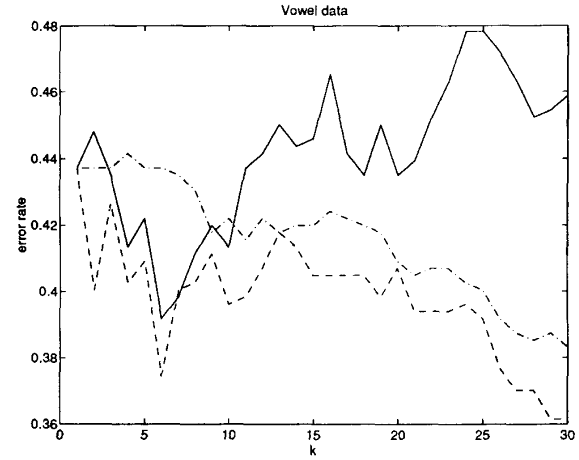


Fig. 9. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Vowel data).

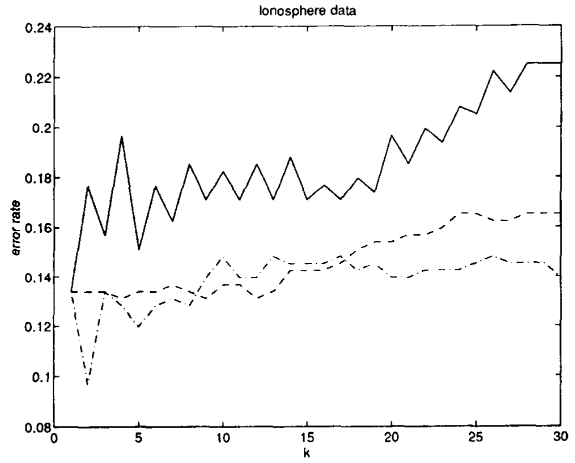


Fig. 10. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Ionosphere data).

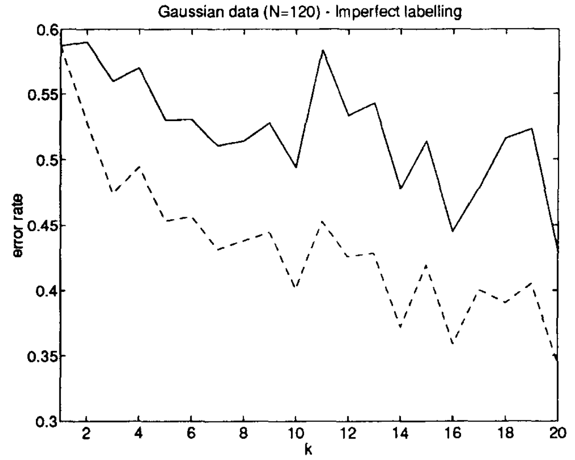


Fig. 11. Mean classification error rates for the voting k-NN rule (-) and our method with consideration of uncertainty in class labels (- -), as a function of k (Gaussian data, $N = 120$).

As can be seen from Fig. 11, our results support this assumption. The two curves correspond to the voting k-NN rule and our method with consideration of uncertainty in class labels. As before, the indicated error rates are averages over 5 trials. The lowest rates achieved, as estimated on an independent test set of 1000 samples, are 0.43 and 0.34, respectively. The percentages of performance degradation resulting from the introduction of uncertainty in the class labels are respectively 54% and 21%. These results indicate that the consideration of the distances to the nearest neighbors *and* of the BPAs of these neighbors both bring an improvement over the majority rule in that case.

V. Conclusion

Based on the conceptual framework of D-S theory, a new non parametric technique for pattern classification has been proposed. This technique essentially consists in considering each of the k nearest neighbors of a pattern to be classified as an item of evidence that modifies one's belief concerning the class membership of that pattern. D-S theory then provides a simple mechanism for pooling this evidence in order to quantify the uncertainty attached to each simple or compound hypothesis. This approach has been shown to present several advantages. It provides a natural way of modulating the importance of training samples in the decision , depending on their nearness to the point to be classified. It allows for the introduction of ambiguity and distance reject options, that receive a unified interpretation using the concepts of lower and upper expected costs. Situations in which only imperfect knowledge is available concerning the class membership of some training patterns are easily dealt with by labelling each recorded sample using basic probability numbers attached to each subset of classes. Simulations using artificial and real-world data sets of moderate sizes have illustrated these different aspects, and have revealed in each case a superiority of the proposed scheme over the voting k-NN procedure in terms of classification performance. In two cases, the results obtained with our method were also better than those obtained with the distance-weighted k-NN rule, while both methods yielded similar results in a third experiment. It should be noted that these results are obviously not sufficient to claim the superiority of our approach for all possible data sets, although no counterexample has been encountered up to now. The comparison with the weighted or unweighted k-NN rules in the infinite sample case is also an interesting, but so far unanswered question.

Another particularity of the technique described in this paper is the quantification of the uncertainty attached to the decisions, in a form that permits combination with the outputs of complementary classifiers, possibly operating at different levels of abstraction. For example, given three classes C_1, C_2 and C_3, one classifier may discriminate between class C_1 and the other two, while another one may help to discern C_2 and C_3. By combining the BPAs produced by each of these classifiers, Dempster's rule offers a way to assess the reliability of the resulting classification. This approach is expected to be particularly useful in data fusion applications, where decentralized decisions based on data coming from disparate sensor sources need to be merged in order to achieve a final decision.

ACKNOWLEDGMENT

The vowel data were obtained from the Carnegie Mellon University collection of neural net benchmarks maintained by Matt White, under the supervision of Scott E. Fahlman. The ionosphere data were obtained from the UCI Repository of machine learning databases maintained by Patrick M. Murphy and David W. Aha. The author wishes to express his thanks to the anonymous referees for their helpful comments and suggestions during the revision of this paper.

REFERENCES

[1] T. Baily and A. K. Jain, "A note on distance-weighted k-nearest neighbor rules," *IEEE Trans. Syst. Man Cyber.*, vol. 8, no. 4, pp. 311–313, 1978.

[2] W. F. Caselton and W. Luo, "Decision making with imprecise probabilities: Dempster-Shafer theory and application," *Water Resources Research*, vol. 28, no. 12, pp. 3071–3081, 1992.

[3] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 41–46, 1970.

[4] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 1, pp. 21–27, 1967.

[5] B. V. Dasarathy, "Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 1, pp. 67–71, 1980.

[6] ——, "Nearest neighbor norms: NN pattern classification techniques," *IEEE Computer Society Press*, Los Alamitos, CA, 1991.

[7] A. P. Dempster and A. Kong, "Comment," *Stat. Sci.*, vol. 2, no. 1, pp. 32–36, 1987.

[8] D. H. Deterding, "Speaker normalization for automatic speech recognition," Ph.D. thesis, University of Cambridge, 1989.

[9] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognition*, vol. 26, no. 1, pp. 155–165, 1993.

[10] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Trans. Syst. Man Cyber.*, vol. 6, pp. 325–327, 1976.

[11] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[12] M. E. Hellman, "The nearest neighbor classification rule with a reject option," *IEEE Trans. Syst. Man Cyber.*, vol. 3, pp. 179–185, 1970.

[13] A. Jozwik, "A learning scheme for a fuzzy k-NN rule," *Pattern Recognition Letters*, vol. 1, pp. 287–289, 1983.

[14] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-NN neighbor algorithm," *IEEE Trans. Syst. Man Cyber.*, vol. 15, no. 4, pp. 580–585, 1985.

[15] J. E. Macleod, A. Luk, and D. M. Titterington, "A re-examination of the distance-weighted k-nearest neighbor classification rule," *IEEE Trans. Syst. Man Cyber.*, vol. 17, no. 4, pp. 689–696, 1987.

[16] R. L. Morin and D. E. Raeside, "A reappraisal of distance-weighted k-nearest-neighbor classification for pattern recognition with missing data," *IEEE Trans. Syst. Man Cyber.*, vol. 11, no. 3, pp. 241–243, 1981.

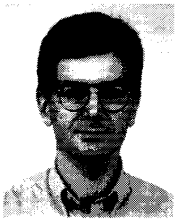
[17] P. M. Murphy and D. W. Aha, "UCI Repository of machine learning databases [Machine-readable data repository]," University of California, Department of Information and Computer Science., Irvine, CA, 1994.

[18] A. J. Robinson, "Dynamic error propagation networks," Ph.D. thesis, Cambridge University Engineering Department, 1989.

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

[20] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," in *Johns Hopkins APL Technical Digest*, vol. 10, pp. 262–266, 1989.

[21] B. Tessem, "Approximations for efficient computation in the theory of evidence," *Artificial Intelligence*, vol. 61, pp. 315–329, 1993.



Thierry Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and earned a Ph.D. from the same institution in 1989.

Until 1992, he worked at LIAC (Laboratoire d'Informatique Avancée de Compiègne), a research center of Lyonnaise des Eaux Dumez, where he was in charge of a European research project on the application of neural networks to forecasting and diagnosis. He is currently an assistant professor at the Université de Technologie de Compiègne. His research interests include artificial neural networks, statistical pattern recognition, and data fusion.