An Adaptive k-NN rule based on Dempster-Shafer theory*

Lalla Merieme Zouhal[1,2] and Thierry Denœux[1]

[1] Université de Technologie de Compiègne - U.R.A CNRS 817 Heudiasyc
BP 649 - F-60206 Compiègne cedex - France
email: tdenoeux@hds.univ-compiegne.fr

[2] Lyonnaise des Eaux
Laboratoire d'Informatique Avancée de Compiègne (LIAC)

Abstract. A new classifier using neighborhood information in the framework of the Dempster-Shafer theory of evidence has recently been introduced. This approach consists in considering each neighbor of a pattern to be classified as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. In this paper, an adaptive version of this method is proposed, in which the parameters used to define the basic probability assignments are learnt from the data by minimizing the mean squared error between the classifier outputs and target values. Based on the evidence-theoretic concepts of degree of conflict and ignorance, new reject rules are introduced. Several sets of artificial and real-world data are used for comparison with the voting and distance-weighted classifiers.

1 Introduction

Since its introduction in the 1950's, the k-nearest neighbor (k-NN) method has been used extensively in statistical pattern recognition for non parametric discrimination. This procedure consists in assigning a pattern to the class represented by a majority of its k nearest neighbors. An important characteristic of this rule is its good asymptotic performance. In [1], it was shown that the asymptotic error rate of the nearest neighbor rule is at most twice the Bayes error. As an improvement to the k-NN voting procedure, Dudani [6] proposed to consider the closeness of each neighbor. The basic idea of the *distance-weighted* k-NN rule is to weight more heavily those training patterns that are the closest to the pattern under consideration. This method was shown to be admissible for small or moderate sized data sets. For small sample problems, it often proves more advantageous to use the distance-weighted rule than the voting procedure with resolved tie [11].

In [2, 3], the classification problem was addressed from the point of view of the Dempster-Shafer (D-S) theory of evidence. In this approach, each of the

* This paper has appeared in the *Proceedings of the 6th International Conference on Computer Analysis of Images and Patterns (CAIP'95)*, pages 310–317, Prague, September 1995. Springer Verlag.

nearest neighbors of a pattern to be classified is considered as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. This information is represented by a basic belief assignment (BBA) defined as a function of the distance between the pattern and its neighbor. Classification is based on the combination of the BBAs associated with the k nearest neighbors, using the Dempster's rule of combination.

In this paper, we propose an adaptive method to tune automatically the parameters of this evidence-theoretic k-NN rule. The approach is based on minimizing the mean squared error between the outputs of the classifier and target values. In the following section, we provide a brief overview of the Dempster-Shafer theory of evidence and the evidence-theoretic k-NN rule. The proposed improvement to this rule is then described, and new reject options are introduced. Finally, several sets of artificial and real-world data are used to compare the performances of the voting, distance-weighted and evidence-theoretic k-NN rules.

2 The evidence-theoretic k-NN rule

2.1 Dempster-Shafer theory

In Dempster-Shafer theory, a problem is represented by a set Θ of mutually exclusive and exhaustive hypotheses called the *frame of discernment* [12]. A BBA is a function $m : 2^\Theta \longmapsto [0,1]$ verifying $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. The quantity $m(A)$ can be interpreted as a measure of the belief that is committed exactly to A, given the available evidence. Two evidential functions derived from the BBA are the credibility function Bel and the plausibility function Pl defined respectively as $Bel(A) = \sum_{B \subseteq A} m(B)$ and $Pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$ for all $A \subseteq \Theta$. A probability distribution P such that $Bel(A) \leq P(A) \leq Pl(A)$ for any subset A of Θ is said to be *compatible* with m. An example of such a probability distribution is the pignistic probability distribution $BetP$ defined for each single hypothesis $\theta \in \Theta$ as [14]:

$$BetP(\theta) = \sum_{\theta \in A} \frac{m(A)}{|A|} \tag{1}$$

Any subset A of Θ such that $m(A) > 0$ is called a focal element of m. Given two BBAs m^1 and m^2 representing two independent sources of evidence, the Dempster's rule defines a new BBA $m = m^1 \oplus m^2$ verifying $m(\emptyset) = 0$ and:

$$m(A) = \frac{\sum_{A_1 \cap A_2 = A} m^1(A_1)\ m^2(A_2)}{\sum_{A_1 \cap A_2 \neq \emptyset} m^1(A_1)\ m^2(A_2)} \tag{2}$$

for all $A \neq \emptyset$.

2.2 Application to k-NN classification

In a classification problem, the frame of discernment consists in the set Ω of classes [2, 3]. Let us consider a pattern $\mathbf{x}^s$ to be classified, and ϕ_k^s the set of its k nearest neighbors in a training set $\mathcal{X}$. For any $\mathbf{x}^i \in \phi_k^s$, the knowledge that $\mathbf{x}^i$ belongs to ω_q can be considered as a piece of evidence that increases our belief that $\mathbf{x}^s$ also belongs to ω_q. This item of evidence can be represented by a BBA m^i assigning a fraction α_q^i of the unit mass to the singleton $\{\omega_q\}$ and the rest to Ω:

$$m^i(\{\omega_q\}) = \alpha_q^i \quad (3)$$

$$m^i(\Omega) = 1 - \alpha_q^i \quad (4)$$

with $m^i(A) = 0$ for all $A \in 2^\Omega \backslash \{\Omega, \{\omega_q\}\}$. The mass α_q^i is chosen as a decreasing function of the distance d^i between $\mathbf{x}^s$ and $\mathbf{x}^i$:

$$\alpha_q^i = \alpha_0 \exp(-\gamma_q^2 \, (d^i)^2) \quad (5)$$

where γ_q is a parameter associated to class ω_q and α_o a fixed parameter. The BBAs $m^1, m^2, \ldots, m^k$ corresponding to the k nearest neighbors of $\mathbf{x}^s$ can then be combined using Dempster's rule: $m = m^1 \oplus \ldots \oplus m^k$. The focal elements of m associated to pattern $\mathbf{x}^s$ are the classes of its k nearest neighbors, and Ω. The credibility and plausibility of each class ω_q are $Bel(\{\omega_q\}) = m(\{\omega_q\})$ and $Pl(\{\omega_q\}) = m(\{\omega_q\}) + m(\Omega)$. The pignistic probability of ω_q is $BetP(\{\omega_q\}) = m(\{\omega_q\}) + m(\Omega)/M$.

A decision is made by assigning a pattern to the class ω_{qmax} with maximum credibility (or equivalently, with maximum plausibility or pignistic probability):

$$qmax = \arg\max_q m(\{\omega_q\}) \quad (6)$$

Ambiguity and distance reject options can be introduced by considering minimum acceptable values for, respectively, the plausibility and credibility of ω_{qmax}.

3 The adaptive evidence-theoretic k-NN rule

In [3], the following heuristic was proposed to determine the parameters α_0 and $\gamma_q, \; q = 1, \ldots, M$:

$$\alpha_0 = 0.95 \;\text{ and }\; \gamma_q = 1/d_q \quad (7)$$

where d_q is the mean distance between training patterns belonging to ω_q. Although these values yield good results on average, the classification performance may vary significantly as a function of $\gamma_q, \; q = 1, \ldots, M$. The fine tuning of these parameters is therefore an important issue.

In this paper, a method is proposed to determine the $\gamma_q, \; q = 1, \ldots, M$ automatically from the training set. Let us consider a training pattern $\mathbf{x}$ belonging to class ω_q. A class label vector $\mathbf{t} = (t_1, \ldots, t_M)^t$ can be defined as $t_i = \delta_{i,q}$ if $\mathbf{x} \in \omega_q$. As a result of classification by the evidence-theoretic k-NN rule, pattern

$\mathbf{x}$ is assigned a BBA m. Based on m, an output vector $\mathbf{P} = (P_1, \ldots, P_M)^t$ is defined as $P_q = m(\{\omega_q\}) + \frac{m(\Omega)}{M}$.

We try to determine the values of γ_q ($q = 1, \ldots, M$) for which $\mathbf{P}$ is as close as possible to the label vector $\mathbf{t}$. Closeness is defined according to the squared error $E(\mathbf{x})$:

$$E(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{M}(P_i - t_i)^2 \tag{8}$$

The mean error for the whole training set $\mathcal{X}$ of size N is defined by $E = \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{X}} E(\mathbf{x})$. The derivative of E with respect to γ_q is:

$$\frac{\partial E(\mathbf{x})}{\partial \gamma_q} = -2\ \gamma_q \sum_{x^j\in\omega_q} \alpha_q^j\ (d^j)^2\ \frac{\partial E(\mathbf{x})}{\partial \alpha_q^j} \tag{9}$$

$$\begin{aligned}\frac{\partial E(\mathbf{x})}{\partial \alpha_q^j} = \frac{1}{K^2}\sum_{i=1}^{M}(P_i - t_i)[-(m'(\{\omega_i\}) + \frac{m'(\Omega)}{M})\frac{\partial K}{\partial \alpha_q^j} - \frac{K}{M}\bar{m}^j(\Omega) + \\ K((\bar{m}^j(\{\omega_i\}) + \bar{m}^j(\Omega))\delta_{iq} - \bar{m}^j(\{\omega_i\}))]\end{aligned} \tag{10}$$

$$\frac{\partial K}{\partial \alpha_q^j} = \sum_{i=1}^{M}((\bar{m}^j(\{\omega_i\}) + \bar{m}^j(\Omega))\delta_{iq} - \bar{m}^j(\{\omega_i\})) - \bar{m}^j(\Omega) \tag{11}$$

where

- m' is the *unnormalized* [13] orthogonal sum of $m^1, \ldots, m^k$;
- $\bar{m}^j$ is the *unnormalized* orthogonal sum of $m^\ell, \ell \in \{1, \ldots, k\} \setminus \{j\}$
- $K = \sum_{q=1}^{M} m'(\{\omega_q\}) + m'(\Omega)$

The parameters γ_q can be determined by gradient descent of E.

3.1 Ambiguity and distance reject

In some applications such as process fault detection, reject options have to be considered [5, 9, 8]. *Ambiguity* reject is required when the pattern under consideration is located close to class boundaries. In this case, the probability of misclassification is high, and it may be preferable to postpone decision-making. *Distance* reject is applied when the pattern is situated far from all known classes. In such a situation, it may be assumed that this pattern actually belongs to an unknown class that is not represented in the training set. When several such patterns have been identified, a new class can sometimes be created using a clustering procedure. By definition, we note:

$$conf = \sum_{A_1\cap A_2=\emptyset} m^1(A_1)\ m^2(A_2) = 1 - K \tag{12}$$

where K is a normalizing factor appearing in Dempster's rule (Equation 2). If m^1 and m^2 have the same focal elements, then $K = 1$ and $conf = 0$; if m^1 and m^2 are not combinable, i.e, $m = m^1 \oplus m^2$ does not exist, then $K = 0$ and $conf = 1$.

The quantity $conf$ can therefore be used as a measure of the degree of conflict between the different BBAs. Given a vector $\mathbf{x}^s$ to be classified, ambiguity reject can be decided by comparing $conf^s$ to a user-defined threshold k_r $(0 \leq k_r \leq 1)$:

$$\text{if} \quad conf^s \;\; \geq \; k_r \quad \text{then } \mathbf{x}^s \text{ is ambiguity-rejected} \tag{13}$$

The Dempster-Shafer theory of evidence has the advantage over Bayesian probability theory that it allows explicit representation of ignorance [12]. The width of the interval $[Bel(\{\omega_q\}), Pl(\{\omega_q\})]$ can be regarded as a measure of the "ignorance" regarding the uncertainty attached to class ω_q. This quantity is constant for all classes, and is equal to the mass associated to the whole frame of discernment Ω. The case where $m(\Omega) = 1$ corresponds to complete ignorance regarding the class of the pattern under consideration. A distance reject option can be based on this concept:

$$\text{if} \;\; m(\Omega) \; \geq \; I_r \qquad \text{then } \mathbf{x}^s \text{ is distance-rejected} \tag{14}$$

The threshold I_r can be determined as the maximum value of $m(\Omega)$ obtained after classification of all training patterns using the leave-one-out method:

$$I_r = \max_{\mathbf{x}^s \in \mathcal{X}} m(\Omega) \tag{15}$$

4 Results

The above method was tested on several sets of synthetic and real data. A comparison was performed with the voting, distance-weighted and evidence-theoretic k-NN (with fixed γ_q) rules. The characteristics of the data sets are summarized in Table 1. The synthetic data sets B1 and B2 were generated using the method proposed in [7]. Each of the four distributions contains three Gaussian classes, which diagonal covariance matrices D_1, D_2, D_3. The i-th diagonal element D_{qi} of D_q is defined as a function of two parameters a and b:

$$D_{1i}(a,b) = a + (b-a)\frac{(i-1)}{(d-1)}, \quad a \; < \; b \;\; i = 1, \ldots, d \tag{16}$$

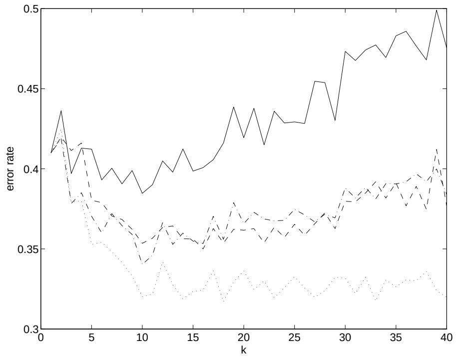
$$D_{2i}(a,b) = b + (a-b)\frac{(d-i)}{(d-1)}, \quad b \; < \; a \;\; i = 1, \ldots, d \tag{17}$$

$$D_{3i}(a,b) = a + (b-a) \min\left(i, \frac{d-i}{(d/2)}\right), \quad a \; < \; b \;\; i = 1, \ldots, d \tag{18}$$

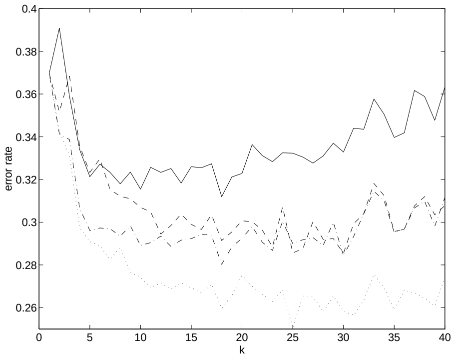
where d is the input dimension. The mean vectors and covariance matrices (with $d = 10$) for the three classes were:

$$\begin{array}{ccc} m_1 = (0, \ldots, 0) & m_2 = (1, \ldots, 1) & m_3 = (1, -1, \ldots, 1, -1) \\ D_1(1, 10) & D_2(10, 1) & D_3(1, 10) \end{array}$$

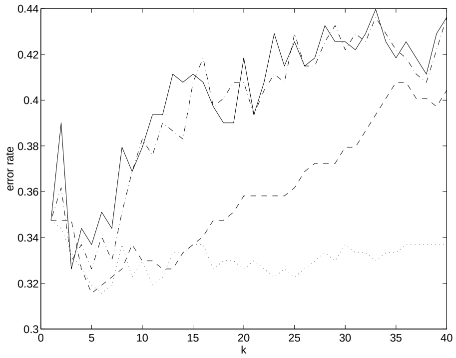
The ionosphere data set (Ion) was collected by a radar system and consists of phased array of 16 high-frequency antennas with a total transmitted power of



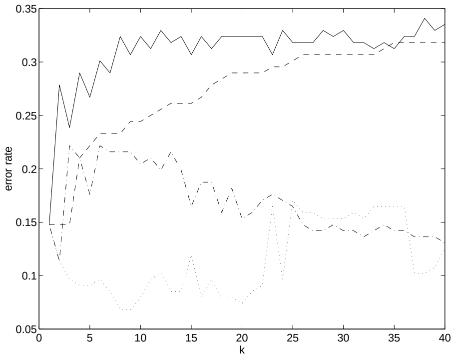
training set size Napp = 90



training set size Napp = 300



vehicle data



ionosphere data

Fig. 1. Test error rates on the four data sets as a function of k for the voting (-), evidence-theoretic with fixed γ_q (-.), adaptive evidence-theoretic (:) and distance-weighted (- -) k-NN rules

Table 1. Data set description.

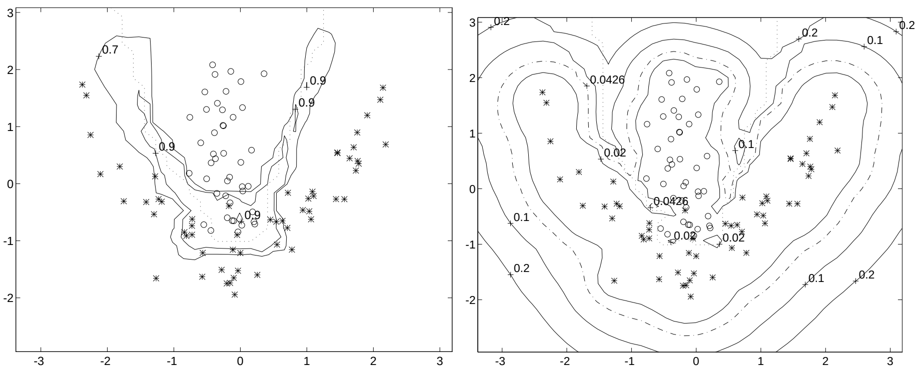
data set	M	training set	test set	dimension
B1	3	90	1000	10
B2	3	300	1000	10
Ion	2	175	176	34
Veh	4	564	282	18

the order of 6.4 kilowatts [10]. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not.

The vehicle data set (Veh) was collected from silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS [10]. Four model vehicles were used for the experiment: bus, Chevrolet van, Saab 9000 and Opel Manta 400. The data was used to distinguish 3D objects within a 2D silhouette of the objects.

For each data set, the values of γ_q $(q = 1, \ldots, M)$ were determined by minimizing the error function E computed on the training set. The error rates were then estimated on the test set. The Euclidean metric was used as a distance measure. The tie problem in the voting k-NN rule was resolved randomly. For synthetic data, the test classification error rate represents the mean for 6 independent training sets. Figure 1 shows the test error rates as a function of k for the four methods. The overall best performance was obtained with the adaptive evidence-theoretic method for the four problems.

In order to test the new reject rules, the regions of ambiguity and distance reject were represented for a two-dimensional data set. Each class contains 50 vectors. As shown in Figure 2, the ambiguity reject region (high degree of conflict) corresponds to the region located between the different classes, while the distance reject region (high degree of ignorance) is located far from existing classes, in regions of low input probability density.



ambiguity reject regions   distance reject regions

Fig. 2. Contour lines of degree of conflict (left) and ignorance $m(\Omega)$ (right) for $k = 6$.

5 Conclusion

In this paper, an adaptive method has been proposed for automatic determination of optimal parameter values to be used in the evidence-theoretic k-NN rule. The method is based on minimizing an error function by gradient descent. For a number of different data sets, the adaptive evidence-theoretic rule outperformed the voting and distance-weighted k-NN rules, as well as the evidence-theoretic k-NN rule with fixed parameters originally proposed. The method also appears to be more robust with respect to the number k of neighbors, which considerably facilitates the design of the classifier.

References

1. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", *IEEE Trans. Inform. Theory*, vol. IT-13(1), 1967,pp. 21-27.
2. T. Denoeux, "Application of evidence theory to k-NN pattern classification". In E. S. Gelsema and L-N Kanal (Editors), *Pattern Recognition in Practice IV*, Elsevier, Amsterdam, 1994, pp. 13-24.
3. T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory", *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-25(5), 1995, pp. 804-813.
4. 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(1), 1980, pp. 67-71.
5. B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes", *Pattern Recognition*, vol. 26(1), 1993, pp. 155-165.
6. S. A. Dudani, "The distance-weighted k-nearest neighbor rule", *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-6(4), 1976, pp. 325-327.
7. J. H. Friedman, "Regularized discriminant analysis", *J. Am. Statist. Ass.*, vol. 84, 1989, pp. 165-175.
8. M. E. Hellman, "The k-nearest neighbor classification rule with a reject option", *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-6(3), 1970, pp. 155-165.
9. A. Jousselin and B. Dubuisson, "A link between k-nearest neighbor rules and knowledge based systems by sequence analysis", *Pattern Recognition Letters*, vol. 6, 1987, pp. 287-295.
10. P. M. Murphy and D. W. Aha, *UCI Reposition of machine learning databases [Machine-readable data repository]*, Irvine, CA, 1994.
11. G. Parthasarathy and B. N. Chatterji, "A class of new k-NN methods for low sample problems", *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-6(4), 1990, pp. 715-718.
12. G. Shafer, *A mathematical theory of evidence*, Princeton N.J, Princeton University Press, 1976.
13. P. Smets, "The combination of evidence in the transferable belief model", *IEEE Trans. Pattern Anal. Machine. Intell.* vol. PAMI-12(5), 1990, pp. 447-458.
14. P. Smets and R. Kennes, "The transferable belief model", *Artificial Intelligence* vol. AI-66, 1994, pp. 191-234.

This article was processed using the LaTeX macro package with LLNCS style