An evidence-theoretic k-NN rule with parameter optimization[1]

L. M. Zouhal[*,†] and T. Denœux[*]

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

[†] Centre International des Techniques Informatiques
Lyonnaise des Eaux

[1] Technical Report Heudiasyc 97/46. To appear in *IEEE Transactions on Systems, Man and Cybernetics.*

Abstract

This paper presents a learning procedure for optimizing the parameters in the evidence-theoretic k-nearest neighbor rule, a pattern classification method based on the Dempster-Shafer theory of belief functions. In this approach, each neighbor of a pattern to be classified is considered as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. Based on this evidence, basic belief masses are assigned to each subset of the set of classes. Such masses are obtained for each of the k nearest neighbors of the pattern under consideration and aggregated using the Dempster's rule of combination. In many situations, this method was found experimentally to yield lower error rates than other methods using the same information. However, the problem of tuning the parameters of the classification rule was so far unresolved. In this paper, we propose to determine optimal or near-optimal parameter values from the data by minimizing an error function. This refinement of the original method is shown experimentally to result in substantial improvement of classification accuracy.

1 Introduction

In the classical approach to statistical pattern recognition, the entities to be classified are assumed to be selected by some form of random experiment. The feature vector describing each entity is then a random vector with well-defined – though unknown – probability density function depending on the pattern category. Based on these densities and on the prior probability of each class, posterior probabilities can be defined, and the optimal Bayes decision rule can then theoretically be used for classifying an arbitrary pattern with minimal expected risk. Since the class-conditional densities and prior probabilities are usually unknown, they need to be estimated from the data. A lot of methods have been proposed for building consistent estimators of the posterior probabilities under various assumptions. However, for finite sample size, the resulting estimates generally do not provide a faithful representation of the fundamental uncertainty pertaining to the class of a pattern to be classified. For example, if only a relatively small number of training vectors is available, and a new pattern is encountered that is very dissimilar from all previous patterns, the uncertainty is quite high and this situation of near-ignorance is not reflected by the outputs of a conventional parametric or non parametric statistical classifier, whose principle fundamentally relies on asymptotic assumptions. This problem is particularly acute in situations in which decisions need to be made based on weak information, such as commonly encountered in system diagnosis applications, for example.

As an attempt to provide an answer to the above problem, it was recently suggested to re-formulate the pattern classification problem by considering the following question: Given a training set of finite size containing feature vectors with known (or partly known) classification, and a suitable distance measure, how to characterize the uncertainty pertaining to the class of a new pattern ? In a recent paper [2], an answer to this question was proposed based on the Dempster-Shafer theory of evidence [12]. The 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. Based on this evidence, basic belief masses are assigned to each subset of the set of classes. Such masses are obtained for each of the k nearest neighbors of the pattern under consideration and aggregated using the Dempster's rule of combination. Given a finite set of actions and losses associated to each action and each class, decisions can then be made by using some generalization of the Bayes decision theory.

In many situations, this method was found experimentally to yield lower error rates than other methods based on the same information. However, the problem of optimizing the parameters involved in the classification rule was so far unresolved. In this paper, we propose to determine optimal or near-optimal parameter values from the data by minimizing a certain error function. This refinement of the original method is shown experimentally to result in substantial improvement of classification accuracy.

The paper is organized as follows. The evidence-theoretic k-NN rule is first recalled in Section 2. The basic concepts of the Dempster-Shafer theory are assumed to be

known to the reader who is invited to refer to [12] and [14] for detailed presentations, and to [2] for a short introduction. Section 3 describes the learning procedure, as well as an approximation to it allowing to near-optimize the error function very efficiently. Simulation results are then presented in Section 4, and Section 5 concludes the paper.

2 The evidence-theoretic k-NN rule

We consider the problem of classifying entities into M categories or classes. The set of classes is denoted by $\Omega = \{\omega_1, \ldots, \omega_M\}$. The available information is assumed to consist in a training set $\mathcal{T} = \{(\boldsymbol{x}^{(1)}, \omega^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \omega^{(N)})\}$ of N n-dimensional patterns $\boldsymbol{x}^{(i)}, i = 1, \ldots, N$ and their corresponding class labels[1] $\omega^{(i)}, i = 1, \ldots, N$ taking values in Ω. The similarity between patterns is assumed to be correctly measured by a certain distance function $d(\cdot, \cdot)$.

Let $\boldsymbol{x}$ be a new vector to be classified on the basis of the information contained in $\mathcal{T}$. Each pair $(\boldsymbol{x}^{(i)}, \omega^{(i)})$ constitutes a distinct item of evidence regarding the class membership of $\boldsymbol{x}$. If $\boldsymbol{x}$ is "close" to $\boldsymbol{x}^{(i)}$ according to the relevant metric d, then one will be inclined to believe that both vectors belong to the same class. On the contrary, if $d(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ is very large, then the consideration of $\boldsymbol{x}^{(i)}$ will leave us in a situation of almost complete ignorance concerning the class of $\boldsymbol{x}$. Consequently, this item of evidence may be postulated to induce a basic belief assignment (BBA) $m(\cdot|\boldsymbol{x}^{(i)})$ over Ω defined by:

$$m(\{\omega_q\}|\boldsymbol{x}^{(i)}) = \alpha\phi_q(d^{(i)}) \quad (1)$$

$$m(\Omega|\boldsymbol{x}^{(i)}) = 1 - \alpha\phi_q(d^{(i)}) \quad (2)$$

$$m(A|\boldsymbol{x}^{(i)}) = 0, \quad \forall A \in 2^\Omega \setminus \{\Omega, \{\omega_q\}\} \quad (3)$$

where $d^{(i)} = d(\boldsymbol{x}, \boldsymbol{x}^{(i)})$, ω_q is the class of $\boldsymbol{x}^{(i)}$ $(\omega^{(i)} = \omega_q)$, α is a parameter such that $0 < \alpha < 1$ and ϕ_q is a decreasing function verifying $\phi_q(0) = 1$ et $\lim_{d\to\infty} \phi_q(d) = 0$. Note that $m(\cdot|\boldsymbol{x}^{(i)})$ reduces to the vacuous belief function $(m(\Omega|\boldsymbol{x}^{(i)}) = 1)$ when the distance between $\boldsymbol{x}$ and $\boldsymbol{x}^{(i)}$ tends to infinity, reflecting a state of total ignorance. When d denotes the Euclidean distance, a rational choice for ϕ_q was shown in [4] to be:

$$\phi_q(d) = \exp(-\gamma_q d^2) \quad (4)$$

γ_q being a positive parameter associated to class ω_q.

As a result of considering each training pattern in turn, we obtain N BBAs that can be combined using the Dempster's rule of combination to form a resulting BBA m synthesizing one's final belief regarding the class of $\boldsymbol{x}$:

$$m = m(\cdot|\boldsymbol{x}^{(1)}) \oplus \ldots \oplus m(\cdot|\boldsymbol{x}^{(N)}) \quad (5)$$

[1] In this paper, we assume for simplicity the class of each training vector to be known with certainty. The more general situation in which the training set is only imperfectly labeled has been introduced in [2]. However, the problem of optimizing the parameters in the general case is not completely solved yet (see Section 5).

Since those training patterns situated far from $\boldsymbol{x}$ actually provide very little information, it is sufficient to consider the k nearest neighbors of $\boldsymbol{x}$ in this sum. An alternative definition of m is therefore:

$$m = m(\cdot|\boldsymbol{x}^{(i_1)}) \oplus \ldots \oplus m(\cdot|\boldsymbol{x}^{(i_k)}) \tag{6}$$

where $I_k = \{i_1, \ldots, i_k\}$ contains the indices of the k nearest neighbors of $\boldsymbol{x}$ in $\mathcal{T}$.

Adopting this latter definition, m can be shown [2] to have the following expression:

$$\begin{aligned} m(\{\omega_q\}) &= \frac{1}{K}\left(1 - \prod_{i \in I_{k,q}} (1 - \alpha\phi_q(d^{(i)}))\right) \prod_{r \neq q} \prod_{i \in I_{k,r}} (1 - \alpha\phi_r(d^{(i)})) \qquad (7) \\ & \quad \forall q \in \{1, \ldots, M\} \\ m(\Omega) &= \frac{1}{K} \prod_{r=1}^{M} \prod_{i \in I_{k,r}} (1 - \alpha\phi_r(d^{(i)})) \qquad (8) \end{aligned}$$

where $I_{k,q}$ is the subset of I_k corresponding to those neighbors of $\boldsymbol{x}$ belonging to class ω_q and K is a normalizing factor. Hence, the focal elements of m are singletons and the whole frame Ω. Consequently, the credibility and the plausibility of each class ω_q are respectively equal to:

$$\mathrm{bel}(\{\omega_q\}) = m(\{\omega_q\}) \tag{9}$$

$$\mathrm{pl}(\{\omega_q\}) = m(\{\omega_q\}) + m(\Omega) \tag{10}$$

The pignistic probability distribution as defined by Smets [13] is given by:

$$\mathrm{BetP}(\{\omega_q\}) = \sum_{\{A \subseteq \Omega | \omega_q \in A\}} \frac{m(A)}{|A|} = m(\{\omega_q\}) + \frac{m(\Omega)}{M} \tag{11}$$

for $q = 1, \ldots, M$. Let us now assume that, based on this evidential corpus, a decision has to be made regarding the assignment of $\boldsymbol{x}$ to a class, and let us denote by α_q the action of assigning $\boldsymbol{x}$ to class ω_q. Let us further assume that the loss incurred in case of a wrong classification is equal to 1, while the loss corresponding to a correct classification is equal to 0. Then, the lower and the upper expected losses [3] associated to action α_q are respectively equal to:

$$R_*(\alpha_q|\boldsymbol{x}) = 1 - \mathrm{pl}(\{\omega_q\}) \tag{12}$$

$$R^*(\alpha_q|\boldsymbol{x}) = 1 - \mathrm{bel}(\{\omega_q\}) \tag{13}$$

The expected loss relative to the pignistic distribution is:

$$R_{bet}(\alpha_q|\boldsymbol{x}) = 1 - \mathrm{BetP}(\{\omega_q\}) \tag{14}$$

Given the particular form of m, the three strategies consisting in minimizing R_*, R^* and R_{bet} lead to the same decision in that case: the pattern is assigned to the class with maximum belief assignment. Other decision strategies including the possibility of pattern rejection as well as the existence of unknown classes are studied in [3, 5].

3 Parameter optimization

3.1 The approach

In the above description of the evidence-theoretic k-NN rule, we left open the question of the choice of parameters α and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^t$ appearing in Equations 1 and 4. Whereas the value of α proves in practice not to be too critical, the tuning of the other parameters was found experimentally to have significant influence on classification accuracy. In [2], it was proposed to set $\alpha = 0.95$ and γ_q to the inverse of the mean distance between training patterns belonging to class ω_q. Although this heuristic yields good results on average, the efficiency of the classification procedure can be improved if these parameters are determined as the values optimizing a performance criterion. Such a criterion can be defined as follows.

Let us consider a training pattern $\boldsymbol{x}^{(\ell)}$ belonging to class ω_q. The class membership of $\boldsymbol{x}$ can be encoded as a vector $\boldsymbol{t}^{(\ell)} = (t_1^{(\ell)}, \ldots, t_M^{(\ell)})^t$ of M binary indicator variables $t_j^{(\ell)}$ defined by $t_j^{(\ell)} = 1$ if $j = q$ and $t_j^{(\ell)} = 0$ otherwise. By considering the k nearest neighbors of $\boldsymbol{x}^{(\ell)}$ in the training set, one obtains a "leave-one-out" BBA $m^{(\ell)}$ characterizing one's belief concerning the class of $\boldsymbol{x}^{(\ell)}$ if this pattern was to be classified using other training patterns. Based on $m^{(\ell)}$, an output vector $\boldsymbol{P}^{(\ell)} = (\mathrm{BetP}^{(\ell)}(\{\omega_1\}), \ldots, \mathrm{BetP}^{(\ell)}(\{\omega_M\}))^t$ of pignistic probabilities can be computed, $\mathrm{BetP}^{(\ell)}$ being the pignistic probability distribution associated to $m^{(\ell)}$. Ideally, vector $\boldsymbol{P}^{(\ell)}$ should as "close" as possible to vector $\boldsymbol{t}^{(\ell)}$, closeness being defined, for example, according to the squared error $E(\boldsymbol{x}^{(\ell)})$:

$$E(\boldsymbol{x}^{(\ell)}) = (\boldsymbol{P}^{(\ell)} - \boldsymbol{t}^{(\ell)})^t (\boldsymbol{P}^{(\ell)} - \boldsymbol{t}^{(\ell)}) = \sum_{q=1}^{M} (P_q^{(\ell)} - t_q^{(\ell)})^2 \tag{15}$$

The mean squared error over the whole training set $\mathcal{T}$ of size N is finally equal to:

$$E = \frac{1}{N} \sum_{\ell=1}^{N} E(\boldsymbol{x}^{(\ell)}) \tag{16}$$

Function E can be used as a cost function for tuning the parameter vector $\boldsymbol{\gamma}$. The analytical expression for the gradient of $E(\boldsymbol{x}^{(\ell)})$ with respect to $\boldsymbol{\gamma}$ can be calculated, allowing the parameters γ_q to be determined iteratively by a gradient search procedure (see Appendix A). Alternatively, the minimum of function E can be approximated in one step for large N using the approach described in the sequel.

3.2 One-step procedure

For an arbitrary training pattern $\boldsymbol{x}^{(\ell)}$ and fixed parameters, vector $\boldsymbol{P}^{(\ell)}$ can be regarded as a function of two vectors:

1. a vector $\boldsymbol{d}^2 = (d^{(i_1)2}, \ldots, d^{(i_k)2})^t$ of squared distances between $\boldsymbol{x}^{(\ell)}$ and its k nearest neighbors, and

2. a vector $\boldsymbol{\omega}$ containing the class labels of these neighbors.

For small k and large N, $\boldsymbol{d}^2$ can be assumed to close to zero[2], allowing each component $P_q^{(\ell)}$ to be approximated by Taylor series expansion around 0 up to the first order:

$$P_q^{(\ell)}(\boldsymbol{d}^2) \cong P_q^{(\ell)}(\mathbf{0}) + \nabla_{\boldsymbol{d}^2} P_q^{(\ell)t}\Big|_{\boldsymbol{d}^2=\mathbf{0}} \boldsymbol{d}^2 \tag{17}$$

The first term in this expression can be readily obtained from Equations 7 and 8 by setting $d^{(i)}$ to 0 for all i, which leads to:

$$P_q^{(\ell)}(\mathbf{0}) = \frac{(1-\alpha)^k}{K}\left((1-\alpha)^{-k_q} - 1 + \frac{1}{M}\right) \tag{18}$$

where k_q is the number of neighbors of $\boldsymbol{x}^{(\ell)}$ in class ω_q and

$$K = \frac{(1-\alpha)^k}{\sum_{q=1}^{M}(1-\alpha)^{-k_q} - k + 1} \tag{19}$$

The computation of the first order term:

$$\nabla_{\boldsymbol{d}^2} P_q^{(\ell)t}\Big|_{\boldsymbol{d}^2=\mathbf{0}} \boldsymbol{d}^2 = \sum_{i=1}^{k} \frac{\partial P_q^{(\ell)}}{\partial d^{(i)2}}(\mathbf{0}) d^{(i)2} \tag{20}$$

is more involved (see appendix B). This term can be shown to be of the form $\boldsymbol{A}^{(\ell)}\boldsymbol{\gamma}$, where $\boldsymbol{A}^{(\ell)}$ is a square matrix of size M. As a consequence, both $E(\boldsymbol{x}^{(\ell)})$ and E can be approximated by quadratic forms of $\boldsymbol{\gamma}$, which allows the minimum to be approached directly by solving a system of linear equations.

Figures 1 and 2 show the quality of this approximation in the case of two Gaussian classes with mean vectors $\boldsymbol{\mu}_1 = (2\ 0)^t$ and $\boldsymbol{\mu}_2 = (-2\ 0)^t$, respectively, and covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$. The data set contained 500 samples of each class. Displayed are the mean squared error as a function of (γ_1, γ_2) (Figure 1) and its quadratic approximation (Figure 2), for $k = 5$. The minima of the two functions differ by less than 0.0007 %, which proves the relevance of the approximation in that case. Note that the quality of the approximation depends on both k and N.

4 Numerical experiments

The performances of the above methods were compared to those of the voting k-NN rule with randomly resolved ties, the distance-weighted k-NN rule [6], the fuzzy k-NN rule proposed by Keller [9], and the evidence-theoretic rule without parameter optimization [2]. Experiments were carried out on a set of standard artificial and

[2]This assumption is justified by the following result [1]: Regarding the training set as a sample drawn from some probability distribution, the k-th nearest neighbor of $\boldsymbol{x}^{(\ell)}$ converges to $\boldsymbol{x}^{(\ell)}$ with probability one as the sample size increases with k fixed.

real-world benchmark classification tasks. The main characteristics of the used data sets are summarized in Table 1.

Data sets B_1 and B_2 were generated using a method proposed in [7]. The data consists in three Gaussian classes in 10 dimensions, which diagonal covariance matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$. The i-th diagonal element D_{qi} of $\boldsymbol{D}_q$ is defined as a function of two parameters a and b:

$$\begin{aligned} D_{1i}(a,b) &= a + (b-a)\frac{i-1}{n-1} \\ D_{2i}(a,b) &= a + (b-a)\frac{n-i}{n-1} \\ D_{3i}(a,b) &= a + (b-a)\min\left(i, \frac{n-i}{n/2}\right) \end{aligned}$$

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

$$\begin{array}{ccc} \boldsymbol{\mu}_1 = (0,\ldots,0) & \boldsymbol{\mu}_2 = (1,\ldots,1) & \boldsymbol{\mu}_3 = (1,-1,\ldots,1,-1) \\ \boldsymbol{D}_1(1,10) & \boldsymbol{D}_2(10,1) & \boldsymbol{D}_3(1,10) \end{array}$$

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 the order of 6.4 kilowatts [11]. 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. 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 [11].

The sonar data were used by Gorman and Sejnowski in a study of the classification of sonar signals using a neural network [8]. The task is to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.

Test error rates are represented as a function of k in Figures 3 to 12. The results for synthetic data are averages over 6 independent training sets. Test error rates obtained for the values of k yielding the best classification of training vectors are presented in Table 2.

As can be seen from these results, the evidence-theoretic rule with optimized $\boldsymbol{\gamma}$ presented in this paper always performed as well or better than the four other rules tested, and significantly improved at the 90 % confidence level over the evidence-theoretic rule with fixed $\boldsymbol{\gamma}$ on data sets B_1 and B_2 when considering the best results obtained for $1 \leq k \leq 40$. However, the most distinctive feature of this rule seems to be its robustness with respect to the number k of neighbors taken in consideration (Figures 3 to 12). By optimizing $\boldsymbol{\gamma}$, the method learns to discard those neighbors whose distance to the pattern under consideration is too high. Practically, this property is of great importance since it relieves the designer of the system from the burden of

searching for the optimal value of k. When the number of training patterns is large, then the amount of computation may be further reduced by adopting the approximate one-step procedure for optimizing $\boldsymbol{\gamma}$ which gives reasonably good results for small k (Figures 3, 5, 7, 9 and 11). However, the use of the exact procedure should be prefered for small and medium-sized training sets.

5 Concluding remarks

A technique for optimizing the parameters in the evidence-theoretic k-NN rule has been presented. The classification rule obtained by this method has proved superior to the voting, distance-weighted and fuzzy rules on a number of benchmark problems. A remarkable property achieved with this approach is the relative insensitivity of the results to the choice of k.

The method can be generalized in several ways. First of all, one can assume a more general metric than the Euclidean one considered so far, and apply the principles described in this paper to search for the optimal metric [15]. For instance, let Σ_q be a positive definite diagonal matrix with diagonal elements $\gamma_{q,1}, \ldots, \gamma_{q,n}$. The distance between an input vector $\boldsymbol{x}$ and a learning vector $\boldsymbol{x}^{(i)}$ belonging to class ω_q can be defined as:

$$\begin{aligned} d(\boldsymbol{x}, \boldsymbol{x}^{(i)}) &= (\boldsymbol{x} - \boldsymbol{x}^{(i)})^t \Sigma_q (\boldsymbol{x} - \boldsymbol{x}^{(i)}) \\ &= \sum_{j=1}^{n} \gamma_{q,j} (x_j - x_j^{(i)})^2 \end{aligned}$$

The parameters $\gamma_{q,j}$ for $1 \leq q \leq M$ and $1 \leq j \leq n$ can then be optimized using exactly the same approach as described in this paper, which may in some cases result in further improvement of classification results. A more general form could even be assumed for Σ_q, with however the risk of a dramatic increase in the number of parameters for large input dimensions.

More fundamentally, the method can also be extended to handle the more general situation in which the class membership of training patterns is itself affected by uncertainty. For example, let us assume that the class of each training pattern $\boldsymbol{x}^{(i)}$ is only known to lie in a subset $A^{(i)}$ of Ω (such a situation may typically arise, e.g., in medical diagnosis problems in which some records in a database are related to patients for which only a partial or uncertain diagnosis is available). A natural extension of Equations 1–3 is then:

$$\begin{aligned} m(A^{(i)}|\boldsymbol{x}^{(i)}) &= \alpha\phi(d^{(i)}) \\ m(\Omega|\boldsymbol{x}^{(i)}) &= 1 - \alpha\phi(d^{(i)}) \\ m(B|\boldsymbol{x}^{(i)}) &= 0, \quad \forall B \in 2^{\Omega} \setminus \{\Omega, A^{(i)}\} \end{aligned}$$

with $\phi(d) = \exp(-\gamma d^2)$, γ being a positive parameter (note that we cannot a define a separate parameter for each class in this case, since the class of $\boldsymbol{x}^{(i)}$ is only partially known). The BBAs defined in that way correspond to simple belief functions and can

be combined in linear time with respect to the number of classes. For optimizing γ, the error criterion defined in Equation 15 has to be generalized in some way. With the same notations as in Section 3.1, a possible expression for the error concerning pattern $\boldsymbol{x}^{(\ell)}$ is:

$$E(\boldsymbol{x}^{(\ell)}) = \left(\mathrm{BetP}^{(\ell)}(A^{(\ell)}) - 1\right)^2,$$

which reflects the fact that the pignistic probability of $\boldsymbol{x}^{(\ell)}$ belonging to $A^{(\ell)}$, given the other training patterns, should be as high as possible. The value of γ minimizing the mean error may then be determined using an iterative search procedure. Experiments with this approach are currently under way and will be reported in future publications.

A Computation of the derivatives of E w.r.t. γ_q

Let $\boldsymbol{x}^{(\ell)}$ be a training pattern and $m^{(\ell)}$ the BBA obtained by classifying $\boldsymbol{x}^{(\ell)}$ using its k nearest neighbors in the training set. Function $m^{(\ell)}$ is computed according to Equations 7 and 8:

$$\begin{aligned} m^{(\ell)}(\{\omega_q\}) &= \frac{1}{K}\left(1-\prod_{i\in I_{k,q}^{(\ell)}}(1-\alpha\phi_q(d^{(\ell,i)}))\right)\prod_{r\neq q}\prod_{i\in I_{k,r}^{(\ell)}}(1-\alpha\phi_r(d^{(\ell,i)})) \quad (21)\\ &\quad \forall q\in\{1,\ldots,M\}\\ m^{(\ell)}(\Omega) &= \frac{1}{K}\prod_{r=1}^{M}\prod_{i\in I_{k,r}^{(\ell)}}(1-\alpha\phi_r(d^{(\ell,i)})) \quad (22)\end{aligned}$$

where $I_{k,r}^{(\ell)}$ denotes the set of indices of the k nearest neighbors of pattern $\boldsymbol{x}^{(\ell)}$ in class ω_r, $d^{(\ell,i)}$ is the distance between $\boldsymbol{x}^{(\ell)}$ and $\boldsymbol{x}^{(i)}$, and K is a normalizing factor. In the following, we shall assume that:

$$\phi_q(d^{(\ell,i)}) = \exp(-\gamma_q d^{(\ell,i)2}) \quad (23)$$

which will simply be denoted by $\phi_q^{(\ell,i)}$. To simplify the calculations, we further introduce the *unnormalized* orthogonal sum [13] $m^{(\ell)'}$ defined as $m^{(\ell)'}(A) = Km^{(\ell)}(A)$ for all $A\subseteq\Omega$. We also denote as $\bar{m}^{(\ell,i)}$ the unnormalized orthogonal sum of the $m(\cdot|\boldsymbol{x}^{(j)})$ for all $j\in I_k^{(\ell)}$, $j\neq i$, that is $m^{(\ell)'} = \bar{m}^{(\ell,i)}\wedge m(\cdot|\boldsymbol{x}^{(i)})$, where $\wedge$ denotes the unnormalized orthogonal sum operator. More precisely, we have:

$$\begin{aligned} m^{(\ell)'}(\{\omega_q\}) &= m^{(\ell)}(\{\omega_q\}|\boldsymbol{x}^{(i)})(\bar{m}^{(\ell,i)}(\{\omega_q\})+\bar{m}^{(\ell,i)}(\Omega))+ \quad (24)\\ &\quad m^{(\ell)}(\Omega|\boldsymbol{x}^{(i)})\bar{m}^{(\ell,i)}(\{\omega_q\}) \quad \forall q\in\{1,\ldots,M\}\\ m^{(\ell)'}(\Omega) &= \bar{m}^{(\ell,i)}(\Omega)m(\Omega|\boldsymbol{x}^{(i)}) \quad (25)\end{aligned}$$

The error for pattern $\boldsymbol{x}^{(\ell)}$ is:

$$E(\boldsymbol{x}^{(\ell)}) = \sum_{q=1}^{M}(P_q^{(\ell)} - t_q^{(\ell)})^2 \quad (26)$$

where $\boldsymbol{t}^{(\ell)}$ is the class membership vector for pattern $\boldsymbol{x}^{(\ell)}$ and $P_q^{(\ell)}$ is the pignistic probability of class ω_q computed from $m^{(\ell)}$ as $P_q^{(\ell)} = m^{(\ell)}(\{\omega_q\}) + \frac{m^{(\ell)}(\Omega)}{M}$.

The derivative of $E(\boldsymbol{x}^{(\ell)})$ with respect to each parameter γ_q can be computed as:

$$\frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial\gamma_q} = \sum_{i\in I_{k,q}^{(\ell)}}\frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial\phi_q^{(\ell,i)}}\frac{\partial\phi_q^{(\ell,i)}}{\partial\gamma_q} \quad (27)$$

with

$$\frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial \phi_q^{(\ell,i)}} = \sum_{r=1}^{M} \frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial P_r^{(\ell)}} \frac{\partial P_r^{(\ell)}}{\partial \phi_q^{(\ell,i)}} \tag{28}$$

$$= \sum_{r=1}^{M} 2(P_r^{(\ell)} - t_r^{(\ell)}) \left[\frac{\partial m^{(\ell)}(\{\omega_r\})}{\partial \phi_q^{(\ell,i)}} + \frac{1}{M} \frac{\partial m^{(\ell)}(\Omega)}{\partial \phi_q^{(\ell,i)}} \right] \tag{29}$$

and

$$\frac{\partial \phi_q^{(\ell,i)}}{\partial \gamma_q} = -d^{(\ell,i)2} \phi_q^{(\ell,i)} \tag{30}$$

The derivatives in Equation 29 can be computed as:

$$\frac{\partial m^{(\ell)}(\{\omega_r\})}{\partial \phi_q^{(\ell,i)}} = \frac{\partial}{\partial \phi_q^{(\ell,i)}} \left(\frac{m^{(\ell)'}(\{\omega_r\})}{K} \right) \tag{31}$$

$$= \frac{1}{K^2} \left[K \frac{\partial m^{(\ell)'}(\{\omega_r\})}{\partial \phi_q^{(\ell,i)}} - m^{(\ell)'}(\{\omega_r\}) \frac{\partial K}{\partial \phi_q^{(\ell,i)}} \right] \tag{32}$$

$$\frac{\partial m^{(\ell)}(\Omega)}{\partial \phi_q^{(\ell,i)}} = \frac{1}{K^2} \left[K \frac{\partial m^{(\ell)'}(\Omega)}{\partial \phi_q^{(\ell,i)}} - m^{(\ell)'}(\Omega) \frac{\partial K}{\partial \phi_q^{(\ell,i)}} \right] \tag{33}$$

$$\frac{\partial K}{\partial \phi_q^{(\ell,i)}} = \sum_{r=1}^{M} \frac{\partial m^{(\ell)'}(\{\omega_r\})}{\partial \phi_q^{(\ell,i)}} + \frac{\partial m^{(\ell)'}(\Omega)}{\partial \phi_q^{(\ell,i)}} \tag{34}$$

Finally:

$$\frac{\partial m^{(\ell)'}(\{\omega_r\})}{\partial \phi_q^{(\ell,i)}} = (\bar{m}^{(\ell,i)}(\{\omega_r\}) + \bar{m}^{(\ell,i)}(\Omega))\alpha\delta_{rq} - \alpha \bar{m}^{(\ell,i)}(\Omega) \tag{35}$$

where δ is the Kronecker symbol, and

$$\frac{\partial m^{(\ell)'}(\Omega)}{\partial \phi_q^{(\ell,i)}} = -\bar{m}^{(\ell,i)}(\Omega) \tag{36}$$

which completes the calculation of the gradient of $E(\boldsymbol{x}^{(\ell)})$ w.r.t. γ_q.

To account for the constraint $\gamma_q \geq 0$, we introduce new parameters μ_q ($q = 1, \ldots, M$) such that:

$$\gamma_q = (\mu_q)^2 \tag{37}$$

and we compute $\frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial \mu_q}$ as:

$$\frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial \mu_q} = \frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial \gamma_q} \frac{\partial \gamma_q}{\partial \mu_q} = 2\mu_q \frac{\partial E(\boldsymbol{x}^{(\ell)})}{\partial \gamma_q} \tag{38}$$

B Linearization

We consider the expansion around $\boldsymbol{d}^2 = \mathbf{0}$ of $P_q^{(\ell)}$ by a Taylor series up to the first order:

$$P_q^{(\ell)}(\boldsymbol{d}^2) \cong P_q^{(\ell)}(\mathbf{0}) + (\nabla_{\boldsymbol{d}^2} P_q^{(\ell)}(\mathbf{0}))^t \boldsymbol{d}^2 \tag{39}$$

where $P_q^{(\ell)}(\mathbf{0})$ is given by Equations 18. In the following, we shall compute the first order term in the above equation, and deduce from that result a method for determining an approximation to the optimal parameter vector. To simplify the notation, the superscript (ℓ) will be omitted from the following calculations.

As a result of the definition of the pignistic probability, we have:

$$\frac{\partial P_q}{\partial d^{(i)2}} = \frac{\partial m(\{\omega_q\})}{\partial d^{(i)2}} + \frac{1}{M}\frac{\partial m(\Omega)}{\partial d^{(i)2}} \tag{40}$$

The derivatives of $m(\{\omega_q\})$ and $m(\Omega)$ can be more conveniently expressed as a function of the unnormalized BBA m':

$$\frac{\partial m(\{\omega_q\})}{\partial d^{(i)2}} = \frac{1}{K^2}\left(K\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}} - m'(\{\omega_q\})\frac{\partial K}{\partial d^{(i)2}}\right) \tag{41}$$

$$\frac{\partial m(\Omega)}{\partial d^{(i)2}} = \frac{1}{K^2}\left(K\frac{\partial m'(\Omega)}{\partial d^{(i)2}} - m'(\Omega)\frac{\partial K}{\partial d^{(i)2}}\right) \tag{42}$$

To compute $\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}}$ and $\frac{\partial m'(\Omega)}{\partial d^{(i)2}}$ we need to distinguish two cases:
Case 1: $i \in I_{k,q}$. We then have:

$$\begin{aligned}\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}} &= -\alpha\gamma_q \exp(-\gamma_q d^{(i)2}) \prod_{j\in I_{k,q}, j\neq i}(1-\alpha\exp(-\gamma_q d^{(j)2})) \\ &\quad \times \prod_{r\neq q}\prod_{j\in I_{k,r}}(1-\alpha\exp(-\gamma_q d^{(j)^2})\end{aligned} \tag{43}$$

$$\frac{\partial m'(\Omega)}{\partial d^{(i)2}} = \alpha\gamma_q \exp(-\gamma_q d^{(i)2}) \prod_{r=1}^{M}\prod_{j\in I_{k,r}, j\neq i}(1-\alpha\exp(-\gamma_r d^{(j)2})) \tag{44}$$

Setting all distances to 0 in the above equations, we have:

$$\left.\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} = -\alpha\gamma_q(1-\alpha)^{k-1} \tag{45}$$

$$\left.\frac{\partial m'(\Omega)}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} = \alpha\gamma_q(1-\alpha)^{k-1} \tag{46}$$

Case 2: $i \in I_{k,l}, l \neq q$. We have:

$$\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}} = \alpha\gamma_l \exp(-\gamma_l d^{(i)2})(1 - \prod_{j\in I_{k,q}}(1-\alpha\exp(-\gamma_q d^{(j)2}))$$

$$\times \prod_{r \neq q} \prod_{j \in I_{k,r}, j \neq i} (1 - \alpha \exp(-\gamma_r d^{(j)2})) \tag{47}$$

$$\frac{\partial m'(\Omega)}{\partial d^{(i)2}} = \alpha \gamma_l \exp(-\gamma_l d^{(i)2}) \prod_{r=1}^{M} \prod_{j \in I_{k,r}, j \neq i} (1 - \alpha \exp(-\gamma_r d^{(j)2})) \tag{48}$$

Setting the distances to zero in the above equations:

$$\left.\frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} = \alpha \gamma_l (1 - (1-\alpha)^{k_q})(1-\alpha)^{k-k_q-1} \tag{49}$$

$$\left.\frac{\partial m'(\Omega)}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} = \alpha \gamma_l (1-\alpha)^{k-1} \tag{50}$$

where $k_q = |I_{k,q}|$.

The derivatives of K are simply obtained as follows:

$$\frac{\partial K}{\partial d^{(i)2}} = \sum_{q=1}^{M} \frac{\partial m'(\{\omega_q\})}{\partial d^{(i)2}} + \frac{\partial m'(\Omega)}{\partial d^{(i)2}} \tag{51}$$

Hence:

$$\left.\frac{\partial K}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} = \alpha \gamma_q \sum_{r=1, r \neq q}^{M} (1 - (1-\alpha)^{k_r})(1-\alpha)^{k-k_r-1} \tag{52}$$

It follows from the preceding calculations that, for $i \in I_{k,r}$, the derivatives of $m'(\{\omega_q\})$, $m'(\Omega)$ and K for $\boldsymbol{d}^2 = \mathbf{0}$ are proportional to γ_r. Since $m'(\{\omega_q\})$, $m'(\Omega)$ and K do not themselves depend on $\boldsymbol{\gamma}$ for $\boldsymbol{d}^2 = \mathbf{0}$, the derivative of P_q is also proportional to γ_r. Hence, we have:

$$\sum_{i \in I_{k,r}} \left.\frac{\partial P_q}{\partial d^{(i)2}}\right|_{\boldsymbol{d}^2=\mathbf{0}} d^{(i)2} = A_{q,r} \gamma_r \tag{53}$$

for all $r \in \{1, \ldots, M\}$, $A_{q,r}$ being some constant not depending on $\boldsymbol{\gamma}$. Consequently, we can write:

$$P_q = P_q(\mathbf{0}) + \sum_{r=1}^{M} A_{q,r} \gamma_r \tag{54}$$

and, expressing this result in matrix form:

$$\boldsymbol{P} \cong \boldsymbol{P}(\mathbf{0}) + \boldsymbol{A}\boldsymbol{\gamma} \tag{55}$$

with $\boldsymbol{A} = (A_{i,j})$ is a square matrix of size M.

The above calculations have been performed for an arbitrary training pattern $\boldsymbol{x}$. Reintroducing the pattern index ℓ, we have:

$$\boldsymbol{P}^{(\ell)} \cong \boldsymbol{P}^{(\ell)}(\mathbf{0}) + \boldsymbol{A}^{(\ell)}\boldsymbol{\gamma} \tag{56}$$

Introducing these terms into the mean squared error, we have:

$$
\begin{aligned}
E &= \frac{1}{N}\sum_{\ell=1}^{N}(\boldsymbol{P}^{(\ell)} - \boldsymbol{t}^{(\ell)})^t(\boldsymbol{P}^{(\ell)} - \boldsymbol{t}^{(\ell)}) \\
&= \frac{1}{N}\sum_{\ell=1}^{N}(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)} + \boldsymbol{A}^{(\ell)}\boldsymbol{\gamma})^t(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)} + \boldsymbol{A}^{(\ell)}\boldsymbol{\gamma}) \qquad (57) \\
&= \frac{1}{N}\sum_{\ell=1}^{N}(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)})^t(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)}) + 2\boldsymbol{\gamma}^t\boldsymbol{A}^{(\ell)t}(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)}) + \qquad (58) \\
&\quad \boldsymbol{\gamma}^t\boldsymbol{A}^{(\ell)t}\boldsymbol{A}^{(\ell)}\boldsymbol{\gamma}
\end{aligned}
$$

The gradient E with respect to $\boldsymbol{\gamma}$ is therefore given by:

$$\nabla_\gamma E = \frac{1}{N}\left[\sum_{\ell=1}^{N}\boldsymbol{A}^{(\ell)t}(\boldsymbol{P}^{(\ell)}(\boldsymbol{0}) - \boldsymbol{t}^{(\ell)}) + \sum_{\ell=1}^{N}\boldsymbol{A}^{(\ell)t}\boldsymbol{A}^{(\ell)}\boldsymbol{\gamma}\right] \qquad (59)$$

Minimizing E under the constraint $\boldsymbol{\gamma} \geq 0$ is a nonnegative least squares problem that may be solved efficiently using, for instance, the algorithm described in [10, page 161].

References

[1] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1):21–27, 1967.

[2] 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.

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

[4] T. Denœux. Application du modèle des croyances transférables en reconnaissance de formes. *Traitement du Signal (In press)*, 1997.

[5] T. Denœux and G. Govaert. Combined supervised and unsupervised learning for system diagnosis using Dempster-Shafer theory. In P. Borne et al., editor, *CESA'96 IMACS Multiconference. Symposium on Control, Optimization and Supervision*, volume 1, pages 104–109, Lille, July 1996.

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

[7] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[8] R. P. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89, 1988.

[9] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-NN neighbor algorithm. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-15(4):580–585, 1985.

[10] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-hall, 1974.

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

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

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

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

[15] L. M. Zouhal. *Contribution à l'application de la théorie des fonctions de croyance en reconnaissance des formes.* PhD thesis, Université de Technologie de Compiègne, 1997.

Biography

Lalla Meriem Zouhal

Lalla Meriem Zouhal received a M.S. in electronics from the Faculté des Sciences Hassan II, Casablanca, Morocco, in 1990, a DEA (Diplôme d'Etudes Approfondies) in System Control from the Université de Technologie de Compiègne, France, in 1992, and a PhD from the same institution in 1997. Her research interests concern pattern classification, Dempster-Shafer theory and Fuzzy logic.

Thierry Denœux

Thierry Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and earned a PhD from the same institution in 1989. He obtained the "Habilitation à diriger des Recherches" from the Institut National Polytechnique de Lorraine in 1996. From 1989 to 1992, he was employed by the Lyonnaise des Eaux water company, where he was in charge of research projects concerning the application of neural networks to forecasting and diagnosis. Dr. Denœux joined the Université de Technologie de Compiègne as an assistant professor in 1992. His research interests include artificial neural networks, statistical pattern recognition, uncertainty modeling and data fusion.

List of Tables

1 Main characteristics of data sets: number of classes (M), training set size (N), test set size (N_t) and input dimension (n). 18

2 Test error rates obtained with the voting, distance-weighted, fuzzy and evidence-theoretic classification rules for the best value of k (in brackets), with 90 % confidence intervals. ETF: evidence-theoretic classifier with fixed γ; ETO: evidence-theoretic classifier with optimized γ. . . . 18

List of Figures

1 Contour lines of the error function for different values of γ, using the gradient-descent method. The optimal value is $\gamma = (0.84, 0.46)^t$. . . . 19

2 Contour lines of the error function for different values of γ, using the linearization method for pignistic probabilities vectors. The optimal value is $\gamma = (0.76, 0.6)^t$. 19

3 Test error rates on data sets B_1 as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method. 20

4 Test error rates on data sets B_1 as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules. 20

5 Test error rates on data sets B_2 as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method. 21

6 Test error rates on data sets B_2 as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules. 21

7 Test error rates on ionosphere data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method. 22

8 Test error rates on ionosphere data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules. 22

9 Test error rates on vehicle data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method. 23

10 Test error rates on vehicle data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules. 23

11 Test error rates on sonar data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method. 24

12 Test error rates on sonar data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules. 24

Table 1: Main characteristics of data sets: number of classes (M), training set size (N), test set size (N_t) and input dimension (n).

data set	M	N	N_t	n
B_1	3	60	1000	10
B_2	3	300	1000	10
Ion	2	175	176	34
veh	4	564	282	18
Son	2	104	104	60

Table 2: Test error rates obtained with the voting, distance-weighted, fuzzy and evidence-theoretic classification rules for the best value of k (in brackets), with 90 % confidence intervals. ETF: evidence-theoretic classifier with fixed γ; ETO: evidence-theoretic classifier with optimized γ.

data set	voting	ETF	ETO	weighted	fuzzy
B_1	0.41 ± 0.03 (8)	0.37 ± 0.03 (13)	0.32 ± 0.02 (27)	0.37 ± 0.03 (21)	0.40 ± 0.03 (5)
B_2	0.31 ± 0.02 (14)	0.29 ± 0.02 (13)	0.24 ± 0.02 (17)	0.28 ± 0.02 (15)	0.29± 0.02 (16)
Ion	0.15 ± 0.04 (1)	0.11 ± 0.04 (2)	0.07 ± 0.03 (8)	0.15 ± 0.04 (1)	0.15 ± 0.04 (1)
Veh	0.33 ± 0.05 (3)	0.34 ± 0.05 (4)	0.33 ± 0.05 (3)	0.32 ± 0.05 (6)	0.34 ± 0.05 (3)
Son	0.17 ± 0.06 (1)	0.17 ± 0.06 (1)	0.15 ± 0.06 (6)	0.17 ± 0.06 (1)	0.17 ± 0.06 (1)

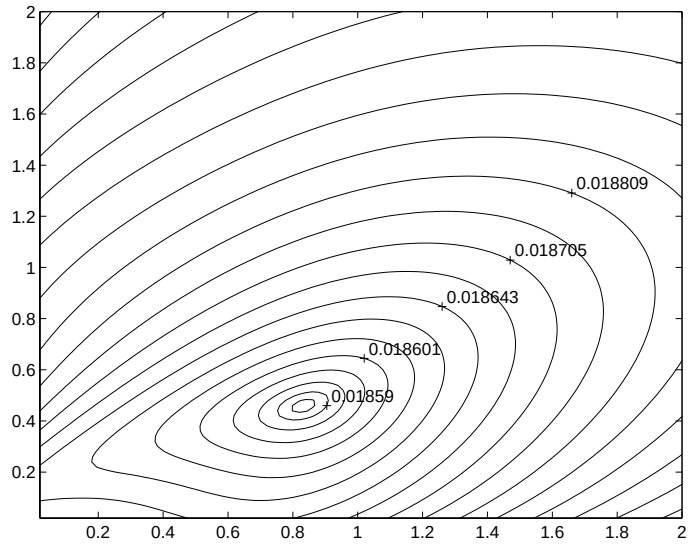

Figure 1: Contour lines of the error function for different values of $\boldsymbol{\gamma}$, using the gradient-descent method. The optimal value is $\boldsymbol{\gamma} = (0.84, 0.46)^t$.

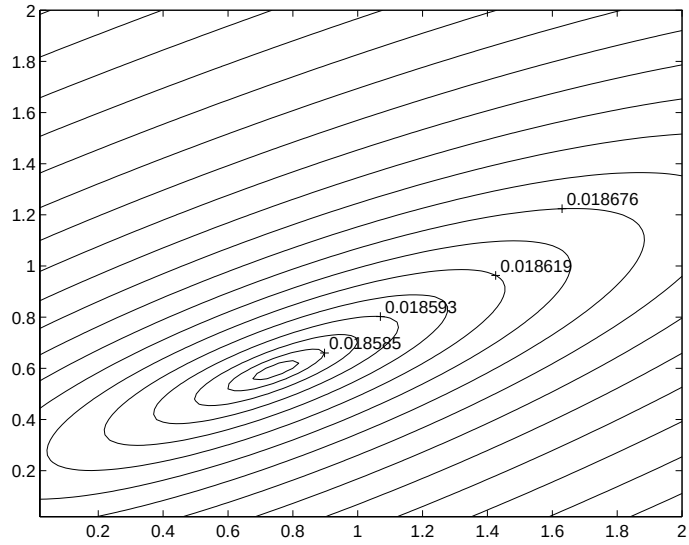

Figure 2: Contour lines of the error function for different values of $\boldsymbol{\gamma}$, using the linearization method for pignistic probabilities vectors. The optimal value is $\boldsymbol{\gamma} = (0.76, 0.6)^t$.

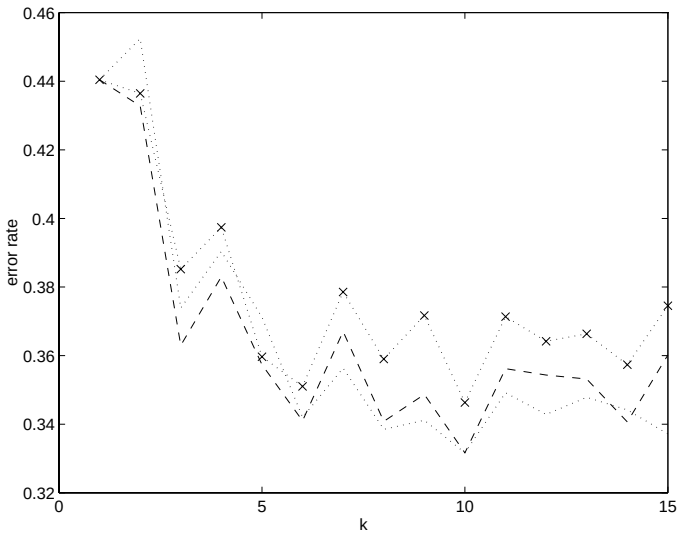

Figure 3: Test error rates on data sets B_1 as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method.

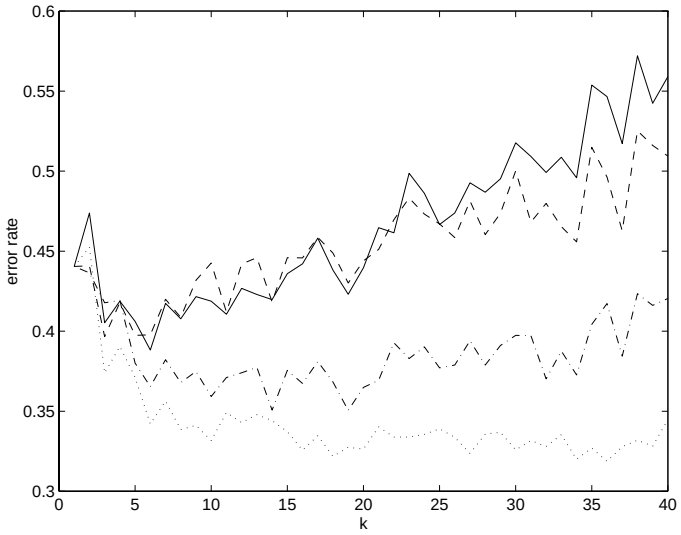

Figure 4: Test error rates on data sets B_1 as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules.

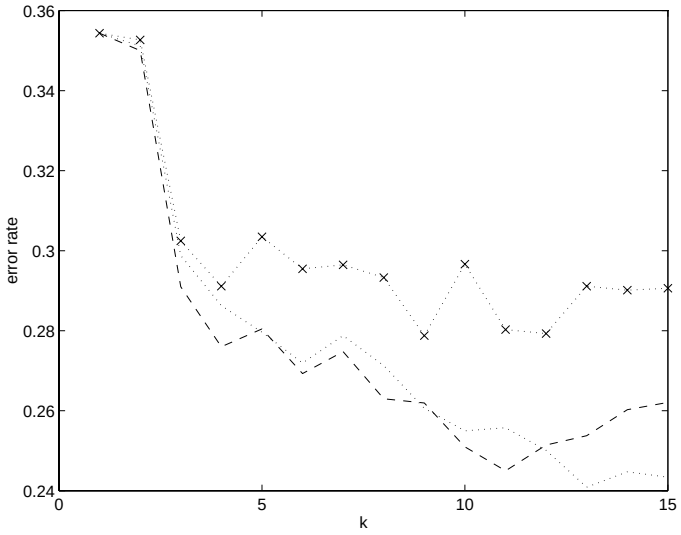

Figure 5: Test error rates on data sets B_2 as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method.

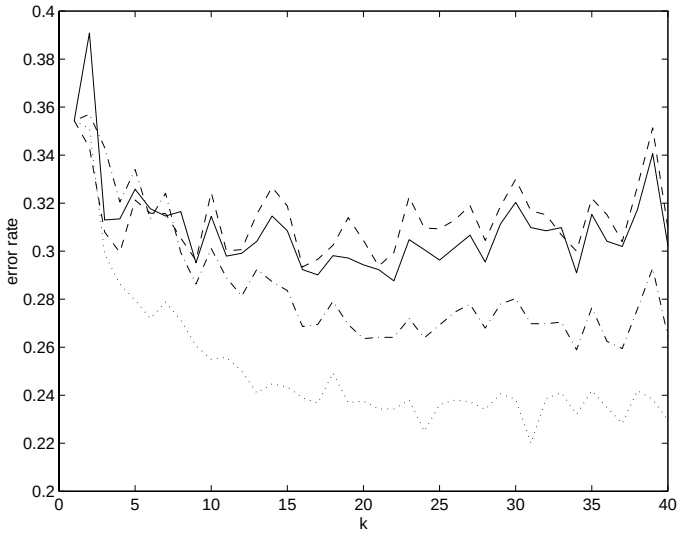

Figure 6: Test error rates on data sets B_2 as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules.

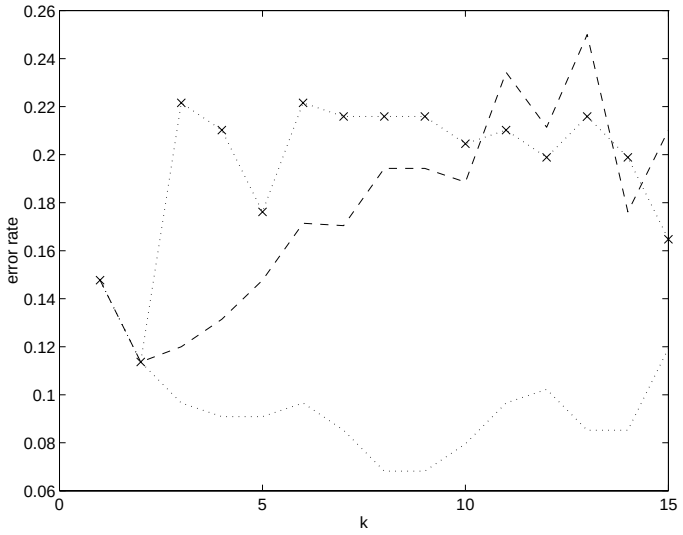

Figure 7: Test error rates on ionosphere data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method.

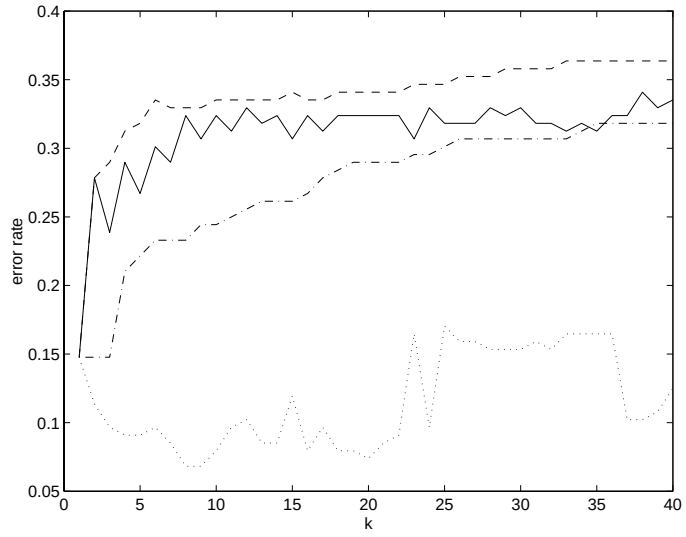

Figure 8: Test error rates on ionosphere data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules.

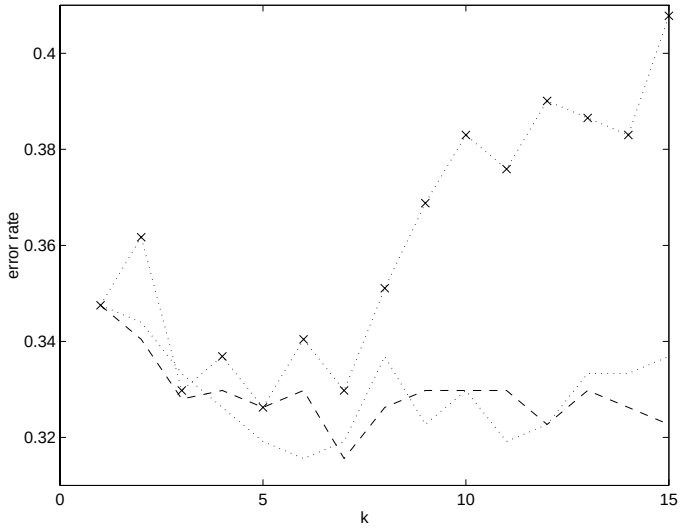

Figure 9: Test error rates on vehicle data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method.

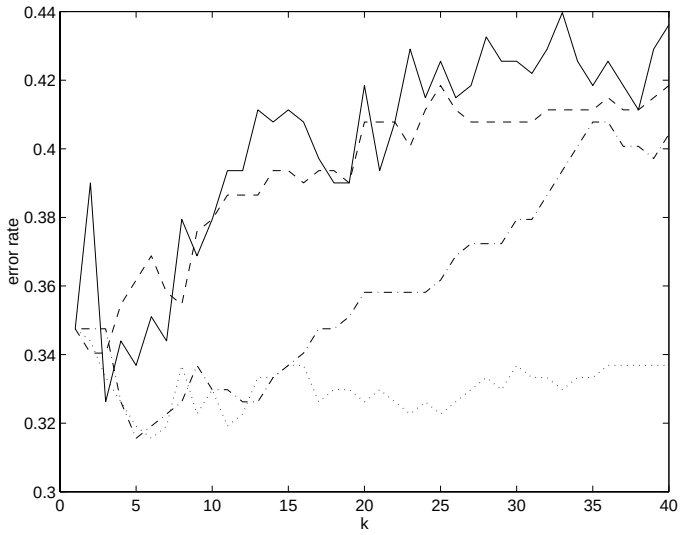

Figure 10: Test error rates on vehicle data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules.

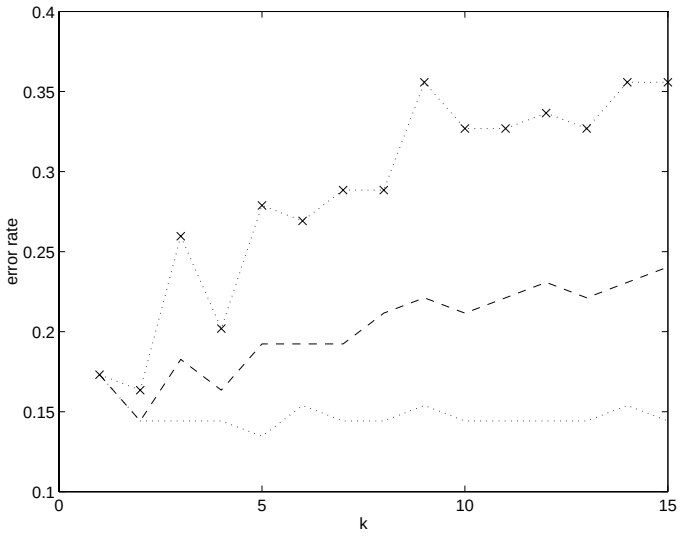

Figure 11: Test error rates on sonar data as a function of k, for the ETF (x.) and ETO k-NN rules: (:) gradient method and (- -) linearization method.

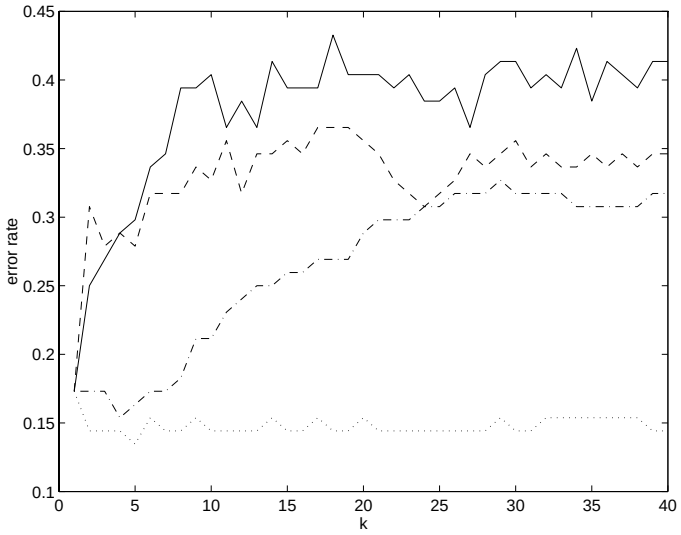

Figure 12: Test error rates on sonar data as a function of k, for the voting (-), ETO(:), fuzzy (- -) and distance-weighted (-.) k-NN rules.