Resample and Combine:
An Approach to Improving Uncertainty Representation in Evidential Pattern Classification

J. François[†‡], Y. Grandvalet[†*], T. Denœux[†] and J.-M. Roger[‡]

[†] Heudiasyc, UMR CNRS 6599,
Université de Technologie de Compiègne,
Centre de Recherches de Royallieu,
F-60205 Compiègne Cedex, France

[‡] Cemagref, GIQUAL Research Unit,
361 rue Jean-François Breton,
F-34033 Montpellier, France

[*]corresponding author, e-mail: Yves.Grandvalet@utc.fr, fax: 33 3 44 23 44 77, phone: 33 3 44 23 49 28

Abstract

Uncertainty representation is a major issue in pattern recognition. In many applications, the outputs of a classifier do not lead directly to a final decision, but are used in combination with other systems, or as input to an interactive decision process. In such contexts, it may be advantageous to resort to rich and flexible formalisms for representing and manipulating uncertain information. This paper addresses the issue of uncertainty representation in pattern classification, in the framework of the Dempster-Shafer theory of evidence. It is shown that the quality and reliability of the outputs of a classifier may be improved using a variant of bagging, a resample-and-combine approach introduced by Breiman in a conventional statistical context. This technique is explained and studied experimentally on simulated data and on a character recognition application. In particular, results show that bagging improves classification accuracy and limits the influence of outliers and ambiguous training patterns.

Keywords: Supervised pattern recognition, K-Nearest Neighbor rule, Decision fusion, Dempster-Shafer theory, Evidence theory, Bootstrap, Bagging, Character recognition.

1 Introduction

Supervised pattern recognition, or classification, is concerned with the design of decision rules whereby entities, described by feature vectors, are assigned to predefined categories. Whereas classification systems are sometimes used directly to trigger specific actions, it is often the case that the outputs from a classifier are used in combination with other sources of information, or are presented to a human decision maker via an interactive decision-aid system. Such situations occur, for example, in medical or technical diagnosis, weather forecasting, financial decision making, and even in certain character recognition applications in which ambiguous patterns are rejected for further interactive processing. In such contexts, it is particularly important to provide not only an indication of the most plausible class, but also a faithful description of the plausibility (taken here in a broad sense) of various hypotheses regarding the class of the pattern under consideration. Uncertainty representation and management thus play an important role in pattern recognition.

In the last thirty years, the issue of uncertainty representation has received considerable attention in the computer science and electrical engineering communities. New theoretical frameworks such as possibility theory [27] and evidence theory [17] have been proposed as alternatives to Bayesian probability theory to describe, manipulate, and reason with partial knowledge and unreliable information. In particular, the so-called Dempster-Shafer (D-S) theory of evidence, first proposed by Shafer [17] and further elaborated by many authors (see, e.g., reviews in Refs. [18, 21, 23]) has been shown to constitute a rich and flexible framework, in which the concepts of a probability and possibility measures are recovered as special cases of the more general concept of *belief function*. This theory has been successfully applied in many areas such as diagnosis [22], sensor fusion [2, 12] and pattern classification [4, 8, 16, 26].

When applying D-S theory to classification tasks, the construction of belief functions from observation data is a crucial step. Typically, a training set of patterns $\{x_i\}_{i=1}^N$ with known classification is given, and one wishes to quantify one's beliefs concerning the category of a new pattern x submitted to the system. A method for inferring a belief function in this context is the evidential K-NN rule previously introduced by one of the authors [4, 15, 28]. In this method, each training example x_i is treated as an item of evidence regarding the unknown class of the pattern x under consideration. The strength of this evidence is assumed to be a decreasing function of the distance between x and x_i. A belief function is constructed by pooling the evidence from the K nearest neighbors of x in the training set.

In this paper, it is proposed to improve this method using a variant of a technique proposed by Breiman [3] in a conventional statistical context to improve the stability of classification rules. In this technique, known as "bagging", B "bootstrap" samples are generated by drawing instances with replacement from the original data set. Each of these samples is then used separately as a training set, resulting in the construction of B distinct classifiers which are then combined using the majority rule. In the present paper, a modification of this technique is proposed, in which each bootstrap sample yields a belief function, and the B belief functions are combined in an appropriate way before a decision is made. This method is shown experimentally to provide a more "realistic" description of the uncertainty pertaining to the classification task, leading to improved classification performances.

The paper is organized as follows. After an introduction to the main concepts of evidence theory and their use in pattern recognition (Section 2), the central idea of this paper, i.e., the adaption of the bagging approach to evidential classifiers, is explained in Section 3. The rest of the paper is then devoted to the presentation and discussion of experimental results obtained in an artificial learning task (Sections 4-6) and in an optical character recognition application (Section 7). In particular, the latter experiment investigates the effect of bagging in an information fusion context, the classifier outputs being combined with a rule expressing prior knowledge. Finally, Section 8 concludes the paper and presents directions for further research.

2 Background

2.1 Theory of Belief Functions

Only the main concepts of the Dempster-Shafer theory of belief functions will be recalled here. The reader is referred to Shafer's book [17] for a detailed exposition of the mathematical background, and to more recent papers such as, e.g., Refs. [23, 24, 25] for up-to-date presentations of the latest developments in both the theoretical aspects and practical applications of belief functions. Note that debates on the relevance of the Dempster-Shafer model, and particularly on its relationship with probability theory have sometimes been obscured by misunderstandings regarding the nature of belief functions at the semantic level [20, 21]. Although our approach is not tied to a particular interpretation of belief functions, we shall adopt the non-probabilistic view of Smets' Transferable Belief Model (TBM), which constitutes a coherent and justified approach [23, 25].

In short, the main assumptions underlying the TBM are that (1) degrees of belief are quantified by numbers between 0 and 1; (2) there exists a two-level structure composed of a *credal level* where beliefs are entertained, and a *pignistic level* where decisions are made; (3) beliefs at the credal level are quantified by belief functions, while decisions at the pignistic level are based on probability functions; (4) when a decision has to be made, beliefs are transformed into probabilities using the so-called *pignistic transformation*.

The credal level

Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be a finite possibility space containing *all* the possible answers to a certain question (the truth lies necessarily somewhere in Ω). In the type of applications envisaged here, Ω is the set of possible classes for an object with unknown class membership. It is assumed that any item of evidence can be represented by a *belief structure, or* basic belief assignment, defined as a function m from 2^Ω (the power set of Ω) to the [0,1] interval, verifying

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

and $m(\emptyset) = 0$. The value of $m(A)$ can be interpreted as the "mass" of belief that is given to A and that cannot be given to any other subset without further information. In particular, $m(\Omega) = 1$ represents total ignorance (m is then called the vacuous belief structure), and $m(\{\omega_1, \omega_2\}) = 1$ means complete certainty that either hypothesis 1 or hypothesis 2 is true (with no evidence in favor of any one of them individually).

The information conveyed by a new source of belief can be incorporated to the current belief structure by use of the Dempster's rule of combination [19]. This can be done only if the sources of belief are independent and non-totally contradictory (that is, two belief structures m_1 and m_2 can be combined if there is $A \subseteq \Omega$ and $B \subseteq \Omega$ with $A \cap B \neq \emptyset$, such that $m_1(A) > 0$ and $m_2(B) > 0$). This combination creates a new belief structure m on Ω that represents the new state of knowledge, defined, for each $C \subseteq 2^\Omega \setminus \emptyset$, as:

$$m(C) = \frac{1}{1-\kappa} \sum_{A \cap B = C} m_1(A) m_2(B) \tag{2}$$

$$\kappa = \sum_{A \cap B = \emptyset} m_1(A) m_2(B). \tag{3}$$

The normalizing factor κ is interpreted as a degree of conflict between the two sources: when $\kappa = 1$, the conflict is total and the sources cannot be combined.

The pignistic level

Given a belief structure, different criteria can be used to choose one hypothesis, such as the maximum of plausibility [2], or the minimization of some given risk. We will use here the pignistic risk minimization as defined and justified by Smets [25] on an axiomatic basis.

Let P_{Bet} be the so-called pignistic probability distribution, defined by uniformly distributing the mass of belief given to each subset of Ω among its elements:

$$P_{\text{Bet}}(\omega) = \sum_{\{A \subseteq \Omega | \omega \in A\}} \frac{m(A)}{|A|} \qquad \forall \omega \in \Omega, \tag{4}$$

where $|A|$ is the number of elements in A.

In the TBM, the pignistic probability function is used for decision making according to the Bayes decision theory. Let $\mathcal{A}$ denote a set of actions, and $\lambda(\alpha|\omega)$ the loss incurred if action $\alpha \in \mathcal{A}$ is selected, $\omega \in \Omega$ being the true state of nature. Then, the expected cost (or risk) when choosing action α, relative to the pignistic distribution, is:

$$R_{\text{Bet}}(\alpha) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega)\ P_{\text{Bet}}(\omega) \tag{5}$$

$$= \sum_{A \subseteq \Omega} \frac{m(A)}{|A|} \sum_{\omega \in A} \lambda(\alpha|\omega). \tag{6}$$

The Bayes decision rule then recommends the action α with the lowest expected cost $R_{\text{Bet}}(\alpha)$.

2.2 Application to pattern classification

In the first applications of D-S theory to pattern recognition problem, the outputs from conventional, probabilistic classifiers were converted into belief structures for more effective combination [13, 16]. This was usually done through the use of classification error rates [26], distances to class centers [13], or class-conditional density estimates [2].

More recently, Denœux proposed an evidence-theoretic pattern recognition scheme, named the evidential K-NN rule [4, 8], although it may be more accurately described as an evidential kernel classifier. It takes fully advantage of the extensive representation of beliefs, without resorting to any intermediate probabilistic representation. The outline of this approach is summarized below.

Let x be the sample to be classified, $\Omega = \{\omega_1, \ldots, \omega_M\}$ the set of classes, and $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{N}$ the learning set of known patterns, where $y_i \in \Omega$ is the class of pattern x_i. Each example x_i is considered as an item of evidence about the class of x. If $y_i = \omega_q$, this evidence induces a belief structure m_i with focal elements $\{\omega_q\}$ and Ω:

$$m_i(A) = \begin{cases} \alpha \exp(-\gamma_q \|x_i - x\|^2) & \text{if } A = \{\omega_q\} \\ 1 - \alpha \exp(-\gamma_q \|x_i - x\|^2) & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\|x_i - x\|$ is the Euclidean distance between x_i and x, and α and γ_q ($q = 1, \ldots, M$) are positive parameters.

This basic belief assignment is thus defined by a radially symmetric function centered on x_i. Each parameter $\gamma_q \in \mathbb{R}^+$ adjusts the influence of the patterns of class q according to their distance to x, while the certainty expressed by training patterns is limited by parameter $\alpha \in [0, 1]$ setting the minimum belief mass given to Ω. These coefficients can be determined from data by a fully automatic procedure [28].

The belief induced by the training examples far from x is almost vacuous (knowing the label of examples far away from the query point is not informative). Hence, for computational reasons, only the belief structures provided by the K-nearest neighbors of x are evaluated. As they are independent from each other, these K belief structures are simply combined into a single structure by means of Dempster's rule

(2-3). This structure represents the available information about the class of x. It is used to compute pignistic probabilities $P_{\mathrm{Bet}}(\omega_j|x)$, from which class assignment can be performed, using the approach described in Section 2.1 [5]. In this context, the set of actions may be defined as $\mathcal{A} = \{\alpha_0,\ \alpha_1, \ldots,\ \alpha_M\}$, where α_i for $i = 1, \ldots, M$ is the decision to classify x in class ω_i, and α_0 denotes rejection. In this paper, the loss is assumed to be 1 in case of a wrong classification and 0 for correct classification. The rejection loss is assumed to be constant, and equal to some value $\lambda_0 \in [0, 1]$. We thus have:

$$\lambda(\alpha_i|\omega_j) = 1 - \delta_{ij} \quad \forall i, j \in \{1, \ldots, M\} \tag{8}$$

$$\lambda(\alpha_0|\omega_j) = \lambda_0 \quad \forall j \in \{1, \ldots, M\}, \tag{9}$$

where δ_{ij} is the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$, and 0 otherwise).

With these costs, the risks are defined, for each action, as follows :

$$R_{\mathrm{Bet}}(\alpha_i|x) = 1 - P_{\mathrm{Bet}}(\omega_i|x), \quad i = 1, \ldots, M \tag{10}$$

$$R_{\mathrm{Bet}}(\alpha_0|x) = \lambda_0. \tag{11}$$

Each pattern is thus assigned to the class with highest pignistic probability, provided this probability is greater that $1 - \lambda_0$. Otherwise, it is rejected. Consequently, parameter λ_0 allows to control the rejection rate of the classifier.

3 Sampling, Learning and Uncertainty

3.1 Problem

The basic belief assignment defined by Eq. (7) handles the uncertainty that stems from the possibly novel characteristics of the query sample. However, additional causes of uncertainty exist. First, the known instances x_i are usually not "prototypical" patterns, such as measurement vectors obtained from some careful experimental design.

They are records of past solved cases, which are supposed to be representative of future unsolved cases. In probabilistic terms, they may be considered as randomly sampled from the distribution of future cases. This random sampling is responsible for some uncertainty in the global belief assignment. This "sampling" uncertainty cannot be represented by a basic belief assignment conditioned on a single realization of the training set.

Additionally, when the parameters of the basic belief assignment are tuned by minimizing some performance criterion on the training set, the learned parameters are also random variables, whose variability is responsible for another part of uncertainty. This is why we propose here the use of bagging, introduced in the probabilistic framework by Breiman to limit the effects of sampling on a learned decision rule.

3.2 Bagging Decision Rules

Bagging is a procedure for improving a classification procedure using a resample-and-combine technique [3]. Breiman argues that its main effect is to decrease the variance of the estimator, and advocates its use for unstable classification methods, i.e. methods which are sensitive to perturbations of the training set.

"Bagging" is an acronym for "bootstrap aggregating". From the original decision rule, the bagged estimator is produced by aggregating, using a majority vote, several replicates of the rule, trained on bootstrap resamples of the learning set. A bootstrap sample [11] is created by drawing with replacement N examples from the learning set $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{N}$. It has thus the same size as the original sample but may contain replicates of some given examples, while other ones are not represented. The drawing with replacement in $\mathcal{L}$ simulates the original sampling from the distribution that generated $\mathcal{L}$. Several empirical evaluations showed that the method almost systematically improves the original predictor [3, 9, 10]. In situations with substantial noise, its performance is also comparable to other ensemble methods such as boosting or randomization [9].

3.3 Bagging in the TBM

In pattern classification, bagging is usually applied to the decisions. In this paper, however, we propose to use it upstream, at the credal level. The main goal is to better take into account the uncertainty attached to the finite training set, in order to allow steadier decisions and, consequently, to improve the result of further combinations when new sources are available.

Practically, B bootstrap learning sets $\mathcal{L}_b$ $(b = 1, \ldots, B)$ are obtained by drawing with replacement N examples from the original learning set $\mathcal{L}$. Here, the bootstrap is balanced, which means that each sample (x_i, y_i) is globally drawn B times over the B resamples. Then, for a given unknown sample x, each training set $\mathcal{L}_b$ produces a belief structure m_b through a given evidential K-NN classifier. These are finally aggregated into the average structure $m_{\mathbf{B}}$, defined as:

$$\forall A \subseteq \Omega, \quad m_{\mathbf{B}}(A) = \frac{1}{B} \sum_{b=1}^{B} m_b(A). \tag{12}$$

The usual bagging combines votes by the majority rule on the B decision rules. Since we are interested in uncertainty representation, aggregation takes place here at the credal level, using the average operator. Note that the Dempster's rule of combination cannot be used here, because the belief sources are obviously not independent.

Although other operators could be used (this is a subject of on-going research), averaging seems to be a good candidate as it is idempotent, commutative and linear: first, getting B times the same structure should lead to this same structure after aggregation (idempotency), second, the resulting structure should be independent from the aggregation order (commutativity), and third, the linear relationship between credal and probabilistic levels, introduced by Smets [25] in the decision process, also supports linear aggregation (linearity).

4 Experimental Settings

4.1 The Problem

In a first attempt to investigate the benefits of bagging, we will focus on an artificial learning task. For easy problems, with well-separated classes and large training sets, many different algorithms usually yield similar results. A learning task of interest should therefore involve overlapping class distributions and a small learning set. Additionally, it should contain outliers as these are frequently encountered in real data sets. Finally, we chose a bidimensional problem so as to easily represent and interpret the results.

In the experiments reported in the sequel, we considered three bidimensional Gaussian distributions with common covariance matrix $\Sigma = 2.25I$ and mean vectors $(0,0)$, $(3,0)$ and $(0,5)$. Each training set $\mathcal{L}$ was constructed by drawing 15 points from each distribution. Additionally, to simulate the contamination of the training set by outliers, 6 points with randomly selected class labels were drawn from a uniform distribution on $[-5,9] \times [-3,8]$.

To exhibit general trends, 15 training sets were generated from the same distribution. Fig. 1 shows an example of such a generated set.

4.2 Evaluation

For each training set, the decision rule was evaluated on a single independent test set $\mathcal{T}$ generated from the same distribution as $\mathcal{L}$ with $N_{\mathcal{T}} = 2000 \times 3 + 800$ items: 2000 patterns in each class and 800 “outliers”. The mean classification cost $\mathbf{C}$ was estimated by the average of the classification costs on the $N_{\mathcal{T}}$ test points of $\mathcal{T}$:

$$\mathbf{C} = \frac{1}{N_{\mathcal{T}}} \sum_{(x,y)\in\mathcal{T}} \lambda(D(x)|y) \tag{13}$$

where $D(x) \in \mathcal{A} = \{\alpha_0, \dots, \alpha_M\}$ denotes the decision made by the classifier for pattern x. The costs were defined according to Eqs (8) and (9): zero for correct classification, one for wrong classification and λ_0 for rejection.

The classification error rate $\mathbf{E}$ was estimated by the proportion of bad predictions (rejection is not an error) and the rejection rate $\mathbf{R}$ was defined as the proportion of rejected items. We thus have the following relation between $\mathbf{C}$, $\mathbf{E}$ and $\mathbf{R}$:

$$\mathbf{C} = \mathbf{E} + \lambda_0 \mathbf{R} \tag{14}$$

The mean classification cost was also computed for the Bayes classifier, whose optimal solution provides a baseline to compare results with and without bagging. Its performances also characterize the intrinsic difficulty of the task.

4.3 Implementation

The evidential K-NN rule described in Section 2.2 requires the setting of $M + 2$ parameters: K (number of neighbors), α and $\gamma = (\gamma_1, \dots, \gamma_M)$. The bagged estimate requires an additional parameter B for the number of bootstrap resamples of the learning set.

In the evidential K-NN rule, the influence of a neighboring vector decreases with its distance to the query point. Setting $K = 8$ was found to result in near-asymptotic behavior while limiting the computational expense. The influence of training patterns depends on parameters α and γ (see Eq. 7). As the influence of α on the classification is low, it was set to the default 0.95 value [4]. Regarding γ, we will proceed here in two steps. First, all γ_q are fixed (Section 5); they are set to the same value (0.5) since the three classes have the same shape and the same number of items. Then, different learning strategies are tested in Section 6.

Finally, the average structure $m_{\mathbf{B}}$ estimates the expected structure over training sets. The expectation over training samples is ideally estimated by the expectation over bootstrap samples. Hence, the number B of bootstrap samples should tend towards infinity. In fact, the effect of bagging is quite visible for values as low as $B = 10$. We used $B = 50$, as the small improvement achieved by higher values is not worth the computation cost. Note that Breiman recommends values around 25.

5 Results without Learning

In this section, the results with and without bagging, for the problem described in Section 4.1, will be compared from three successive viewpoints: (1) the quality of the decisions, (2) the closeness of the pignistic probabilities to the class posterior probabilities, and (3) the ability of the output belief structures to adequately represent the classification uncertainty.

5.1 Decision Level

Figure 2 shows mean classification costs vs. rejection costs for the 15 experiments. The horizontal segments in boxplots represent the lower quartile, median, and upper quartile over the 15 simulations. Minimal and maximal values are indicated by the whiskers, and the plotted curve itself is the average over experiments. Bagging clearly improves classification for low classification costs, which correspond to higher rejection rates (the difference in the average mean classification cost is significant to the 5% level for $0 < \lambda_0 \leq 0.5$ according to the exact Wilcoxon signed ranks test for matched samples). Its cost is half-way between the original algorithm and the Bayes classifier. However, this benefit vanishes for high values of λ_0 (low rejection rates).

The improvement due to bagging is thus linked to its higher capacity to reject truly ambiguous patterns. In fact, the class with maximum pignistic probability $P_{\mathrm{Bet}}(\omega_j|x)$ is generally not modified by the bagging procedure, but the pignistic probabilities values may be significantly modified so that rejection is more frequent.

In agreement with what intuition suggests, taking into account the uncertainty due to the finite size of the training sample hardly modifies the rank of the highest pignistic probability. Its value is however properly lowered, which is interpreted as a more uncertain outcome. Bagging is thus beneficial when the values attached to belief assignments are of interest. Besides rejection, all applications where a measure of uncertainty should be attached to the decision are concerned.

Remark: In our method, each bootstrap resample of the training set generates a belief structure for each x. These B structures are first aggregated by averaging, and the decision is then based on this average belief structure. The faithful transposition of the original proposition of Breiman would have been to perform a majority vote between the decisions provided by the B classifiers. Experimental results (not shown here) show that this strategy is a poor choice in the TBM framework. This suggests that the evidential K-NN procedure already provides stable decision rules, a finding in agreement with Breiman's results concerning the standard K-NN [3].

5.2 Pignistic Level

While it may be possible to display the effect of bagging at the credal level, there is no satisfactory criteria for measuring the relevance of a belief structure. We thus resort to the study of pignistic probabilities which give more information on beliefs than the decisions themselves.

The results regarding mean classification cost suggest that, with bagging, the pignistic probabilities $P_{\mathrm{Bet}}(\cdot|x)$ should be closer to the posterior probabilities $p(\cdot|x)$.

The latter can be computed exactly from the densities

$$f(x|\omega_j) = \frac{15}{17} f_{\mathcal{N}}(x; \mu_j, \Sigma) + \frac{2}{17} f_{\mathcal{U}}(x) \tag{15}$$

where $f_{\mathcal{N}}(\cdot; \mu, \Sigma)$ is the Gaussian density of mean μ and covariance matrix Σ and $f_{\mathcal{U}}(\cdot)$ is the uniform distribution on $[-5, 9] \times [-3, 8]$ (see Section 4.1). Knowing that all priors $p(\omega_j)$ are equal to $1/3$, the posterior probabilities are directly obtained by Bayes' rule.

The overall mean quadratic error on posterior class probabilities is 40% lower when bagging is applied. As our two-dimensional example allows us to visualize probability surfaces, it is possible to characterize situations where bagging incurs significant modifications of probabilities. An example is given in Fig. 3, which shows that bagging performs a data-dependent smoothing, highly effective in regions where data is scarce, and otherwise less marked. Hence, the main differences occur at class boundaries and for outliers (one is situated at the left-hand side of the graph).

In terms of estimation errors, the result is beneficial, as displayed in Fig. 4. Bagging thus yields a better representation of uncertainties, stemming either from ambiguity (where classes overlap) or from lack of information (in regions of low density of training patterns).

5.3 Credal Level

At each point x, the aggregated belief structure is the average of 50 belief structures. The distribution of these structures indicates the relevance of the average operator for aggregating beliefs regarding uncertainty representation.

At the credal level, the effect of bagging is again visible in the regions where outliers were present in the learning set or where classes overlap. Fig. 5 shows the mass distributions of the B structures associated to two test examples. These are given to each of the four hypotheses ω_1, ω_2, ω_3 and to the reference set Ω.

In the low-probability density regions, the masses on the three hypotheses ω_j are small because the neighbors are far from x. Much of the mass then goes to Ω, which is always fully compatible with any more precise hypothesis; the remaining mass is usually given to the nearest neighbor class. In the absence of conflict in the neighborhood, the average structure is a good summary of the distribution, providing a good representation of uncertainty.

In ambiguous regions, some belief structures that were produced on bootstrapped training sets assigned most of the mass to one hypothesis or another (in our example ω_1 or ω_3) because of high heterogeneity in the neighborhood. The resulting mass distributions $m(\{\omega_1\})$ and $m(\{\omega_3\})$ are bimodal. Bagging through averaging distributes the belief mass between the classes in conflict, and provides a good compromise at the pignistic level. However, the average is not a faithful summary of multimodal distributions. As a consequence, no trace of the individual conflicts remains at the aggregated credal level. Possible answers to this problem will be mentioned in the concluding section.

6 Results with Learning

In the previous section, the parameters α and γ of the basic belief assignments were set to arbitrary values. The effect of bagging regarding uncertainty due to the finite sample size was thus isolated. This section depicts the effect of bagging regarding the uncertainty pertaining to the learning of parameters. In the following simulations, α was fixed at 0.95 as it was shown to have only marginal influence on the classification results [4, 28].

6.1 Influence of γ

As explained in Section 2.2, the influence regions of training patterns are controlled by γ (Eq. (7)). Fig. 6 shows the mean classification cost as a function of γ for the original classifier and its bagged version. Note that these curves, computed on the test set, could not have been drawn in a real problem. Our goal here is to understand why bagging works, not to propose a method for choosing γ.

The bagged K-NN mean classification cost is always lower than that of the original algorithm, for all values of γ and all rejection costs. Thus, the results presented in the previous sections are representative of what would be obtained for any value of γ. The comparison of the two plots in Fig. 6 also shows that the differences between the two methods are larger for small rejection costs, regardless of γ.

Looking now at both plots in Fig. 6, we see that bagging is more effective in improving the original method for small values of γ, i.e., when all neighbors have almost the same influence, regardless of their distance to the query sample. In this case, the resulting belief is too confident, and bagging neatly corrects it.

In comparing the two graphs, it may be noted that, for the bagged algorithm, the optimal γ value is identical for both rejection costs, while it depends on λ_0 for the standard algorithm. Indeed, these two values should ideally not interact, as beliefs should not be affected by the consequences of actions. These consequences should only be taken into account in the decision process.

Finally, the lower variability of $\mathbf{C}$ provides a steadier optimal γ value and a lower sensitivity to errors in γ, in terms of misclassification cost. This stability results in an improvement of γ estimation methods, as shown in the sequel.

6.2 Estimation of γ

Although the fine tuning of γ is less important with bagging, we need a practical way of estimating a relevant value. Here, we use the learning scheme of Zouhal and Denœux [28], which minimizes the leave-one-out cross-validation estimate of the mean quadratic error on posterior probabilities.

Fig. 7 shows the mean classification cost as a function of the rejection cost for the 15 experiments. As in the fixed-γ case, bagging is beneficial mostly for low classification costs (the difference in the average mean classification cost is significant to the 5% level for $0 < \lambda_0 \leq 0.35$ according to the exact Wilcoxon signed ranks test for matched samples). The improvement is higher, which means that the outcome of bagging regarding learning is also beneficial, and that it does not counteract the effect regarding sampling. The comparison of box sizes here and in Fig. 2 also illustrates that the learning of γ induces an additional variability of performances which is lowered, and even almost suppressed with bagging.

The mean quadratic difference between pignistic probabilities and true posterior probabilities confirms the benefit of bagging at this level. Bagging significantly reduces the average error from 0.61 to 0.30. The variability with respect to the learning sets is also lowered (the standard deviation drops from 0.22 to 0.09).

7 Combination of Beliefs: an Application

We now turn to real data in order to illustrate and study the benefits of bagging, from the point of view of combination with external sources of beliefs. Indeed, there does not seem to be any direct way to measure how well a belief structure *represents* the available information (for example, pignistic probabilities do not allow the representation of ignorance, wich can be coded in a belief structure). We thus tackle the problem by combining both the bagged and non-bagged K-NN structures with the same simple rule, decide, and only then compare the results to assess the effect of bagging at the credal level.

7.1 The problem

Alpaydin and Kaynak [1] proposed a multistage recognition method, which was applied to a handwritten digit recognition problem. Their database consists of scanned digits (0 to 9) represented as 32×32 normalized black-and-white bitmap images [14]. A group of 30 subjects contributed to the 3823 images of the training set and another group of 13 subjects was used to generate the 1797 test images. The images were reduced to 8×8 gray scale bitmaps using a low-pass filter. A standard 1-NN classifier based on a simple distance between images then leads to 98% correct classification[1].

With such a large training set, there is no room for significant improvement using more sophisticated procedures such as the evidential K-NN rule. To assess the usefulness of bagging when combining with external sources of beliefs, the original training set was therefore subsampled to 10 items by class, resulting in a total of 100 items (Fig. 8). The test set was left unchanged. Given the small number of items per class, we chose $K = 4$. Here again, the experiments were repeated 15 times, resulting in 15 different learning sets.

7.2 Combination with a rule

The evidence-theoretic framework allows the combination of different sources of information as long as they are represented by belief structures. This example is intended to illustrate that information stemming from pattern recognition systems and complementary sources of belief such as rules can easily be combined by Dempster's rule. The benefits of bagging at the credal level in the pattern recognition system are then

[1]More elaborate distances are usually proposed, but this is not the point in this paper.

highlighted by the performances at the decision level of the combined classifier.

Let $\Omega = \{0, \ldots, 9\}$ be the hypothesis space and $H = \{0, 6, 8, 9\}$ the set of digits whose handwritten representation has usually at least one hole. Let R be the simple rule:

> If the bitmap image x of a digit has at least one hole, then it is highly probable that it represents a digit of H, $y \in H$, and $y \notin H$ otherwise.

We wish to use R as an additional source of belief concerning the class of bitmap images. The presence of a hole in the bitmap representation can easily be computed by applying mathematical morphology operators to the original 32×32 binary images.

In the TBM, $y \in H$ translates to $m(H) = 1$ and $m(\overline{H}) = 0$, and $y \notin H$ translates to $m(H) = 0$ and $m(\overline{H}) = 1$. However, R cannot be completely trusted, as some digits may have both holed and non-holed representations (e.g. some people write digit 4 like it is typeset, with a hole). Let P_{hole} and $P_{\overline{\text{hole}}}$ be respectively the proportion of bitmap images with and without hole. The rule error rate $\mathbf{E}$ can be decomposed in two parts $\mathbf{E} = P_{\text{hole}}\mathbf{E}_{\text{hole}} + P_{\overline{\text{hole}}}\mathbf{E}_{\overline{\text{hole}}}$, where $\mathbf{E}_{\text{hole}}$ is the classification error rate for bitmaps with a hole, and $\mathbf{E}_{\overline{\text{hole}}}$ is the error rate for bitmaps without holes.

In this regard, $\mathbf{E}_{\text{hole}}$ and $\mathbf{E}_{\overline{\text{hole}}}$ can be considered as measures of distrust in R. For example, $\mathbf{E}_{\text{hole}} = 0.5$ means that R is completely useless in predicting digits for bitmaps with holes (as random guess achieves the same error rate). This should be represented by the vacuous belief structure $m_R(\Omega) = 1$. On the other hand, $\mathbf{E}_{\overline{\text{hole}}} = 0$ means that R is fully reliable concerning bitmaps without holes, and should then lead to the belief structure $m_R(\overline{H}) = 1$. Consequently, we define the belief structure m_R associated to R as shown in Table 1. Note that a similar method for defining belief settings based on error rates was proposed by Xu [26]. The values of $\mathbf{E}_{\text{hole}}$ and $\mathbf{E}_{\overline{\text{hole}}}$ can directly be computed with the unused items of the training set.

The belief structure m_R is a distinct source of belief: it can therefore be combined with the belief structures produced by the evidential K-NN rule and its bagged version.

7.3 Combination results

The error rates computed from unused samples in the original training set are $\mathbf{E}_{\text{hole}} = 5.6\%$ and $\mathbf{E}_{\overline{\text{hole}}} = 1.7\%$. Note that these low error rates should not be compared to the ones obtained by evidential K-NN technique, since only two subsets of classes are discriminated by the rule. The belief structure produced by the evidential K-NN rule is combined with m_R, by use of the Dempster's rule of combination (Eqs. 2-3).

Table 2 gives the mean classification error rates averaged over 15 different learning sets. These are given with and without bagging, with and without the use of rule R. The multiplicative coefficients associated to the horizontal arrows give the improvement rate when rule R is taken into account. The factors corresponding to vertical arrows are the improvement linked to the use of bagging (all differences in mean error rates are significant, with p-values smaller than 0.05% according to the exact McNemar test for matched samples).

7.4 Discussion

Looking at horizontal arrows in Table 2, we see that both the bagged and unbagged K-NN rules are improved when combining with R. The reduction of the mean error rate by a factor of 2/3 illustrates the usefulness of such a simple classification rule in this context. The improvement related to bagging is shown by factors associated to vertical arrows. Bagging also leads to significant improvements.

The observation of vertical arrows on a single particular dataset can also be interesting, as depicted in Table 3. In this example, the improvement due to bagging before applying the rule is not significant (at the 5% level). However, it becomes important (significant up to the 0.2% level) when the rule is used.

As the rule is the same with and without bagging, this can only be explained by better belief representation before combination. The bagged method yields roughly the same ranking of pignistic probabilities (as shown by the similar error rates before combination), but its belief structure is less confident and, consequently, it may be highly improved by additional information.

This application demonstrates that the combination of a pattern recognition technique with an external source of belief is more profitable when uncertainty is faithfully represented. On the one hand, when a query example is very similar to a known prototype, the output of the case-based classifier should be able to contradict the imperfect rule-based classifier. Hence, digit 4 may be recognized as being a 4 with or without a hole. On the other hand, when the query point is far from all prototypes, the final decision process should be more trustful in the rule classifier. Once the classifiers are constructed, Dempster's rule of combination entails the weighting between more or less confident opinions. We have presented empirical evidence that resampling and combination techniques provide a fully automatic, yet very efficient means to correct overconfident beliefs, thus improving the performances of evidence-based multi-source classification schemes.

8 Conclusion

In the framework of pattern recognition, belief structures allow to represent uncertainty stemming from lack of information (small sample size) or from doubtful items of information (unvalidated data). As for probabilistic classifiers, evidential classifiers predict the plausibility of each outcome. Besides, their ability to provide imprecise predictions can be used as a reliability index by the final decision process. This feature is extremely attractive in information fusion.

Bagging combines B belief structures given by the evidential classifier applied to bootstrap samples. This modification of the belief structure construction process aims at improving uncertainty representation when the sample size is small.

The method was tested on controlled artificial datasets. Classification error was shown to be significantly reduced when rejection was allowed. The improvements were even higher when the belief assignment parameters were estimated, due to the stabilization of the estimation process. The influence of bagging was also visible when looking at pignistic probabilities, which estimate posterior probabilities. Among all quantities which can be computed and evaluated objectively in the TBM, pignistic probabilities are the closest we can get to belief structure. There is thus evidence that the belief structures provided by bagging are more relevant.

Another clue supporting this conjecture was provided by an application to handwritten character recognition, where the pattern recognition classifier was combined with another source of belief expressed as a rule. After combination of the two information sources, error rates were reduced, even when bagging had no perceptible effect before combination. Bagging thus turns out to be beneficial at the credal level, since the relevance of a belief structure can be defined by its capacity to be specified by additional trustful pieces of evidence.

Beyond the evidential K-NN, this paper illustrates the necessity to build generic tools for inferring accurate beliefs. It provides, up to our knowledge, one of the first attempts to take into account the uncertainty due to the presence or absence of an information source upon which beliefs are constructed. In the classical pattern recognition paradigm, in which information sources are data points assumed to be sampled from some fixed distribution, resample and combine techniques provide a fully automatic means to correct undue certainty in inferred beliefs. In our experiments, this correction was shown to have more important outcomes for classifiers making a more intensive use of data (with learned parameters). The improvements should thus be more effective with more sophisticated inference methods such as the neural-network based evidential classifier described in [8]. This should be confirmed in a further experimental study.

Another extension of this work concerns the investigation of other operators to combine the belief structures in the bagging procedure. More general mathematical objects such as interval-valued or fuzzy belief structures [6, 7] could even be used to keep track of the discord within the B structures. This could further improve the quality of belief representation at the credal level, which was shown to be an important issue in an information fusion context.

References

[1] E. Alpaydin and C. Kaynak. Cascading classifiers. *Kybernetika*, 34(4):369–374, 1998.

[2] A. Appriou. Uncertain data aggregation in classification and tracking processes. In B. Bouchon-Meunier, editor, *Aggregation and Fusion of imperfect information*, pages 231–260. Physica-Verlag, Heidelberg, 1998.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

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

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

[6] T. Denœux. Reasoning with imprecise belief structures. *International Journal of Approximate Reasoning*, 20:79–111, 1999.

[7] T. Denœux. Modeling vague beliefs using fuzzy-valued belief structures. *Fuzzy Sets and Systems*, 116(2):167–199, 2000.

[8] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[9] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40(2):1–19, 2000.

[10] P. Domingos. Why does bagging work? a Bayesian account and its implications. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 155–158. AAAI Press, 1997.

[11] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, New York, 1993.

[12] S. Fabre, A. Appriou, and X. Briottet. Presentation and description of two classification methods using data fusion based on sensor management. *Information Fusion*, 2(1):49–71, 2001.

[13] E. Mandler and J. Schürmann. Combining the classification results of independent classifiers based on the Dempster-Shafer theory of evidence. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition and Artificial Intelligence*, pages 381–393. North-Holland, Amsterdam, 1988.

[14] 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.

[15] N. R. Pal and S. Ghosh. Some classification algorithms integrating dempster-shafer theory of evidence with the rank nearest neighbor rules. *IEEE Trans. on Systems, Man and Cybernetics A*, 31(1):59–66, 2001.

[16] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

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

[18] G. Shafer. Perspectives in the theory and practice of belief functions. *Intern. J. Approx. Reasoning*, 4:323–362, 1990.

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

[20] P. Smets. Resolving misunderstandings about belief functions. *International Journal of Approximate Reasoning*, 6:321–344, 1990.

[21] P. Smets. What is Dempster-Shafer's model? In R. R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 5–34. Wiley, New-York, 1994.

[22] P. Smets. The application of the Transferable Belief Model to diagnosis problems. *International Journal of Intelligent Systems*, 13:127–158, 1998.

[23] P. Smets. The Transferable Belief Model for quantified belief representation. In D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 267–301. Kluwer Academic Publishers, Dordrecht, 1998.

[24] P. Smets. Practical uses of belief functions. In *Proc. of UAI'99*, Stockholm, Sweden, August 1999.

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

[26] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[27] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[28] L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. on Systems, Man and Cybernetics C*, 28(2):263–271, 1998.

Tables

Table 1: Belief defined by R for a digit d.

Case	$m_R(H)$	$m_R(\overline{H})$	$m_R(\Omega)$
hole	$1 - 2\mathbf{E}_{\text{hole}}$	0	$2\mathbf{E}_{\text{hole}}$
no hole	0	$1 - 2\mathbf{E}_{\overline{\text{hole}}}$	$2\mathbf{E}_{\overline{\text{hole}}}$

Table 2: Averaged misclassification rate (%) over 15 different training sets, $K = 4$. All differences in misclassification rates are all significant up to the 0.05% level.

K-NN	bare		with R
Original	12.9	$\xrightarrow{\times 0.67}$	8.6
	$\downarrow\times 0.88$		$\downarrow\times 0.83$
Bagged	11.3	$\xrightarrow{\times 0.64}$	7.2

Table 3: Example of misclassification rate changes (%) for one training set, $K = 4$. The improvement due to bagging before applying the rule is not significant at the 5% level; after the rule is applied the difference is significant up to the 0.2% level.

K-NN	bare		with R
Original	11.3	$\xrightarrow{\times 0.70}$	8.0
	$\downarrow\times 0.95$		$\downarrow\times 0.68$
Bagged	10.7	$\xrightarrow{\times 0.51}$	5.4

Figures

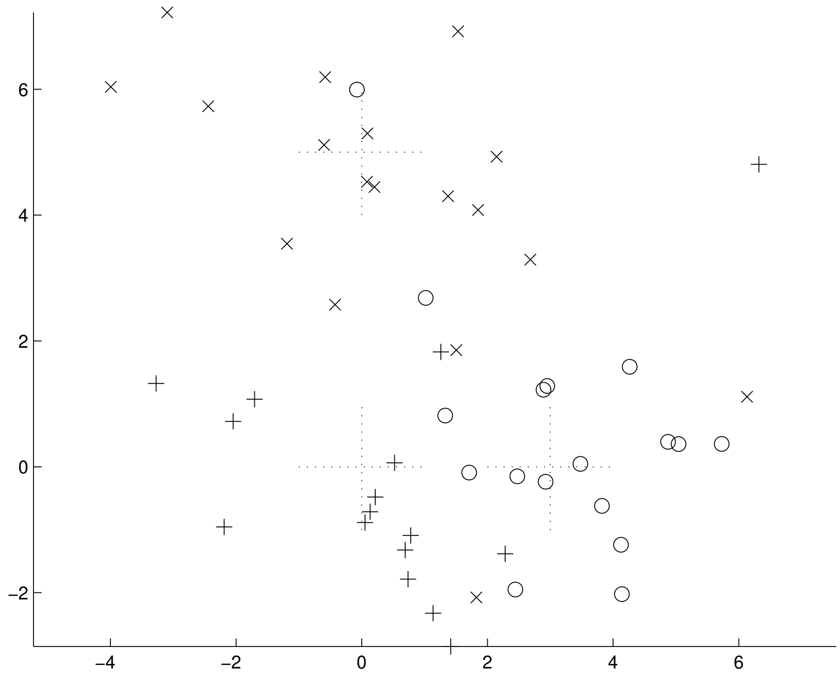

Figure 1: Example of a generated learning set. The intersections of dotted lines indicate the class means.

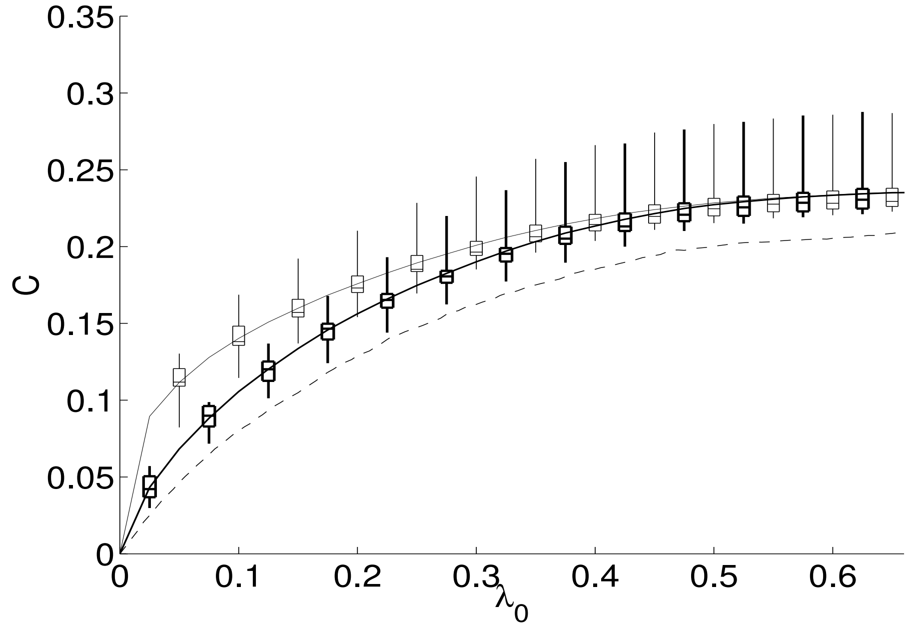

Figure 2: Mean classification cost $\mathbf{C}$ as a function of rejection cost λ_0 for original (thin line) and bagged (bold line) methods (γ fixed). The dotted line corresponds to the minimum cost.

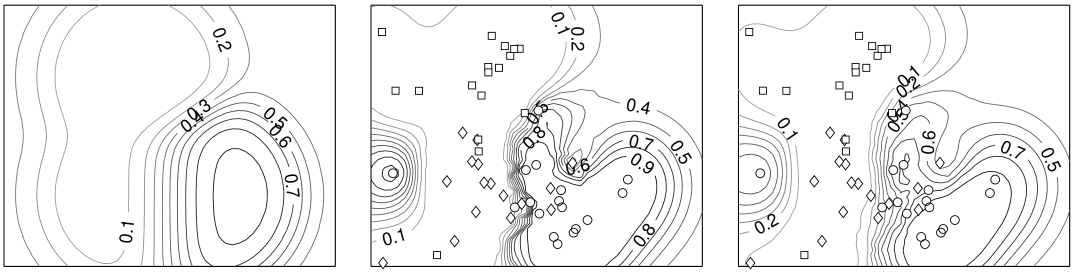

Figure 3: Contour plots of true posterior probabilities $p(\omega_2|x)$ (left) and estimated pignistic probabilities $P_{\mathrm{Bet}}(\omega_2|x)$ without bagging (center) and with bagging (right). Each symbol represents a training example of class 1 ($\diamond$), class 2 ($\circ$) or class 3 ($\square$).

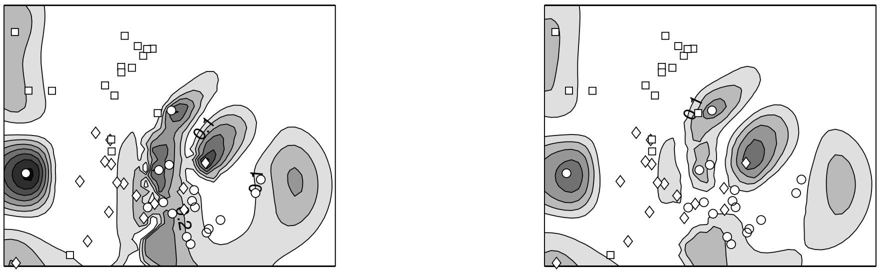

Figure 4: Contour plots of absolute error on posterior probabilities $|P_{\mathrm{Bet}}(\omega_2|x) - p(\omega_2|x)|$ without bagging (left) and with bagging (right). Each symbol represents a training example of class 1 ($\diamond$), class 2 ($\circ$) or class 3 ($\square$).

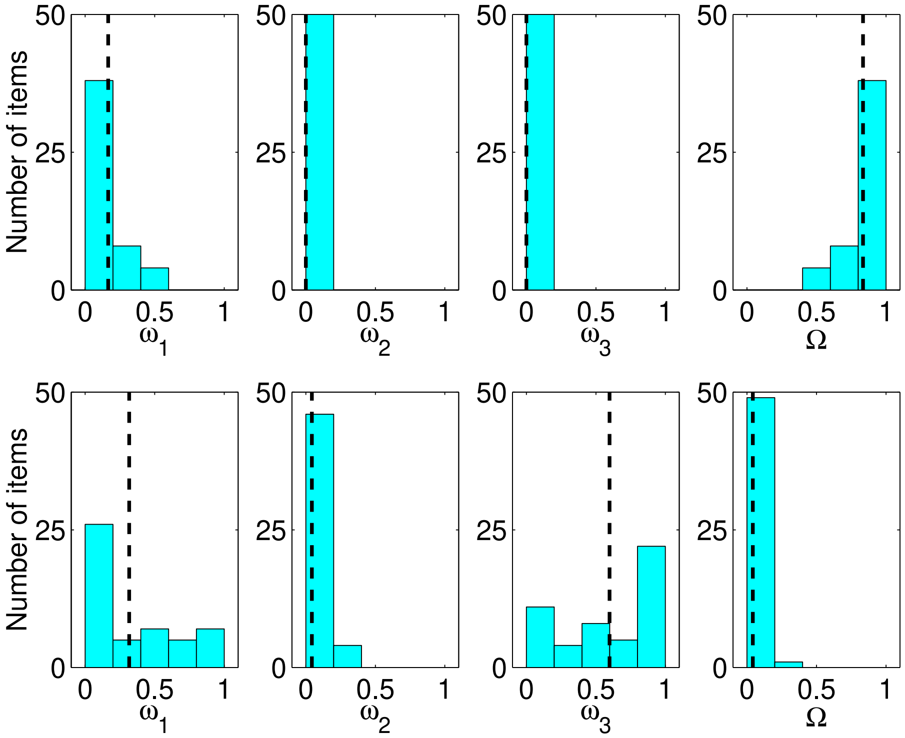

Figure 5: Histograms of the 50 belief structures obtained before averaging: for an outlier (top) and a point in ambiguous ω_1–ω_3 region (bottom). The vertical dotted line indicates the value of the average (combined structure).

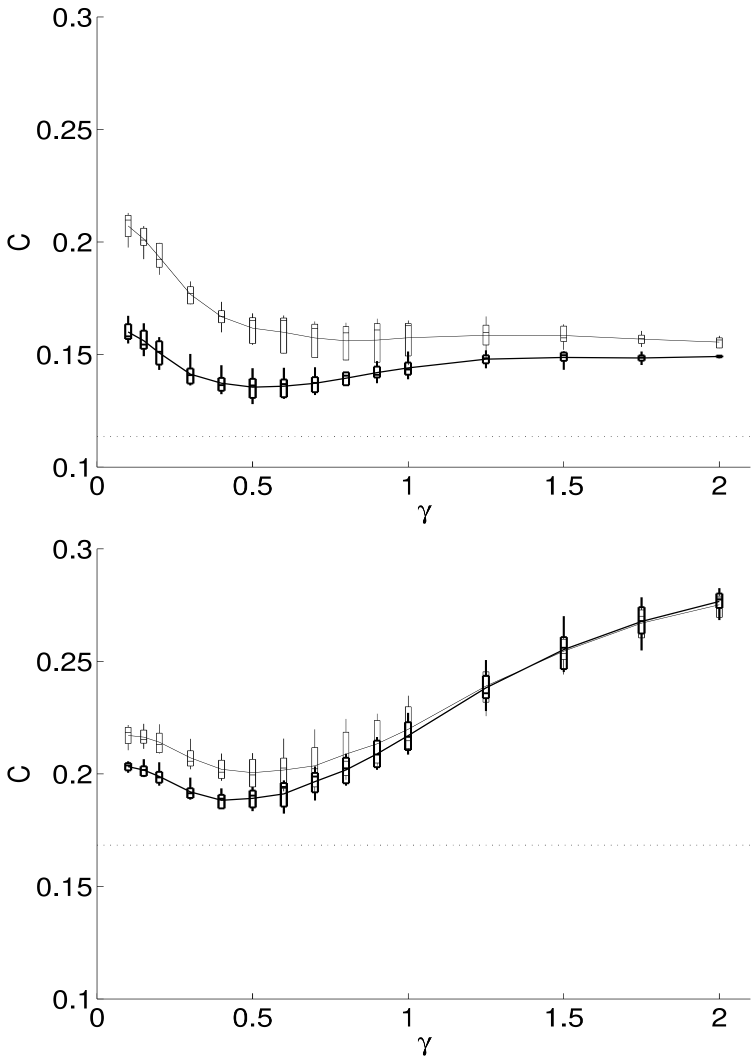

Figure 6: Mean classification cost $\mathbf{C}$ as a function of γ for $\lambda_0 = 0.15$ (top) and $\lambda_0 = 0.3$ (bottom) for original (thin line) and bagged (bold line) methods. The dotted line represents Bayes' classification cost.

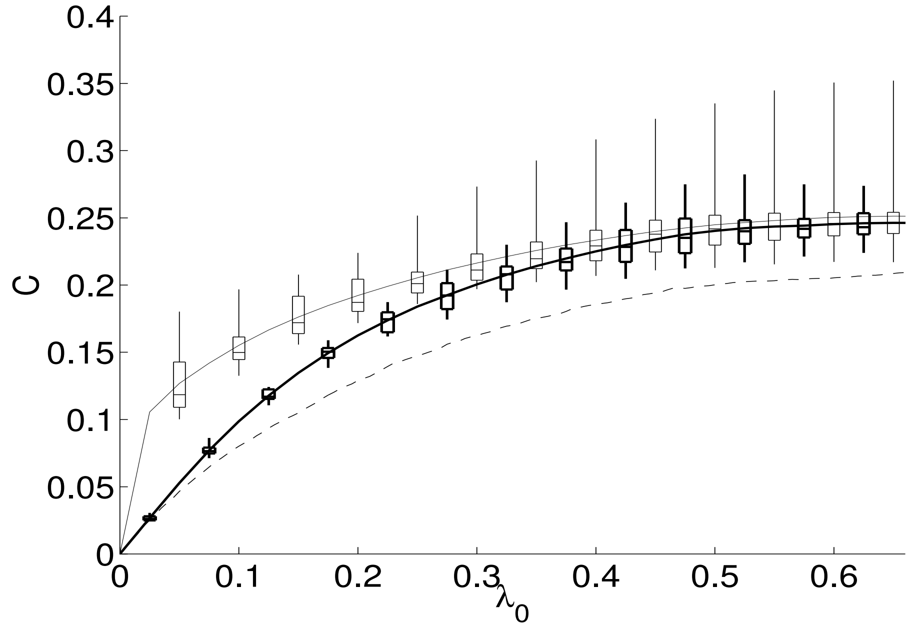

Figure 7: Mean classification cost **C** as a function of rejection cost λ_0 for original (thin line) and bagged (bold line) methods (γ learned). The dotted line represents the minimum cost.

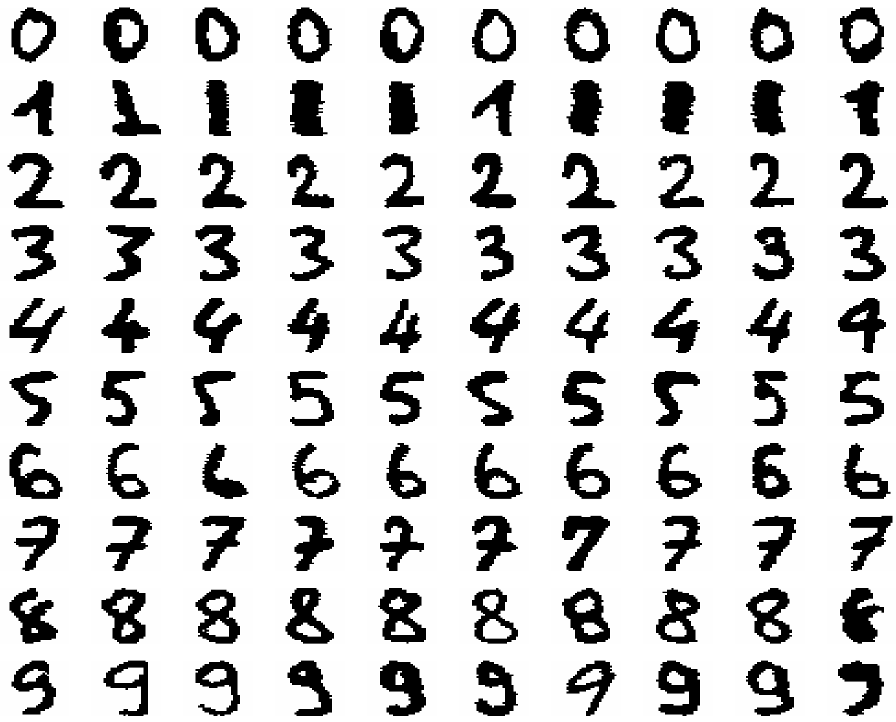

Figure 8: Example of a learning set of size $N = 100$.