Bagging belief structures in Dempster-Shafer K-NN rule

J. François†‡ Y. Grandvalet† T. Denœux†
† Heudiasyc, UMR CNRS 6599,
Université de Technologie de Compiègne,
F-60205 Compiègne, France
{francois,grandvalet,denoeux}@hds.utc.fr

J.-M. Roger‡
‡ Cemagref, GIQUAL Research Unit,
361 rue Jean-François Breton,
F-34033 Montpellier, France
jean-michel.roger@cemagref.fr

Abstract

This paper introduces bagging in the evidence-theoretic K-nearest neighbor rule (K-NN). It is known that bagging decreases the variability of classifiers, so the main idea here is to build stable belief structures associated to query samples, before decisions are made. In order to compare bagged K-NN with the classical algorithm in a precisely controlled environment, data sets were generated according to known distributions. Results show that bagging improves classification, especially for ambiguous cases and outliers. Moreover, bagged belief structures give pignistic probabilities closer to the *a posteriori* class probabilities than the original method.

Keywords: Non-additive Measures, Nearest Neighbor Rule, Estimations of beliefs, Bagging.

1 Introduction

The transferable belief model was introduced by Smets [11] as a non probabilistic interpretation of the Dempster-Shafer theory. Denœux [2] used this model to derive a Dempster-Shafer K-nearest neighbor decision rule. This paper deals with bagging as a way to improve the robustness of this classifier.

For small learning sets, the standard K-nearest neighbors rule is very sensitive to ambiguous training items, *e.g.* points that are isolated or badly labelled. Indeed, the number of neighbors is set by a bias/variance trade-off: small K values yield nearly unbiased but variable rules, whereas large values provide biased but stable classifiers. As this is a global setting, the trade-off usually results in a highly variable rule in low density regions. Regarding this point, the Dempster-Shafer K-nearest neighbor is better than the original rule, as the conjunctive combination weakens the influence of ambiguous items. In this paper we discuss bagging [1] as a means to further reduce the sampling-induced noise.

The next section of this paper deals with the transferable belief model, used to formalize the K-NN classifier. Section 3 explains bagging. The experimental settings are detailed in section 4. Results and discussions can be found in section 5.

2 The K-NN rule in the Transferable Belief Model

The theory of belief functions was introduced by Dempster and developed by Shafer [9]. It is a numerical method that modelizes reasoning under uncertainty. On its basis, Smets [11] created the transferable belief model, within which he formalized a decision process.

2.1 Belief functions

Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the set of all possible distinct and exclusive hypotheses (closed-world assumption). Any piece of information is represented by a belief structure m:

$$m : 2^{\Omega} \to [0,1] \quad \begin{cases} m(\emptyset) = 0 \\ \sum_{A \subseteq \Omega} m(A) = 1 \end{cases} \tag{1}$$

The value of $m(A)$ is the quantity of belief that is given to A and that cannot be given to any other subset without further information. Consequently, $m(\Omega) = 1$ represents total ignorance (vacuous structure), $m(\{\omega_1\}) = 0.5$ stands for a moderated belief in hypothesis 1, and $m(\{\omega_1, \omega_2\}) = 1$ means absolute belief in hypothesis 1 or hypothesis 2 (with no information about each of them separately).

A new trustable piece of information can be incorporated by use of the conjunctive Dempster rule of combination [10]. 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 and B in Ω such that $A \cap B \neq \emptyset$, $m_1(A) > 0$ and $m_2(B) > 0$). This combination creates a new belief structure m on Ω that represents the new state of knowledge, defined as $m(\emptyset) = 0$ and, for each $\Theta \subseteq 2^{\Omega}$, $\Theta \neq \emptyset$:

$$m(\Theta) = \frac{1}{1-k_\Theta} \sum_{A_1 \cap A_2 = \Theta} m_1(A_1)\ m_2(A_2) \tag{2a}$$
$$k_\Theta = \sum_{B_1 \cap B_2 = \emptyset} m_1(B_1)\ m_2(B_2) \tag{2b}$$

k_Θ is called the degree of conflict between the two sources ($k_\Theta = 1$ means total conflict, the sources cannot then be combined).

All such belief representation and combination is called the credal level. As we show it in next section, decision will be done at another level, called the pignistic level[1]. This approach clearly separate belief representation from the decision based on the beliefs where and only where probabilities are introduced.

2.2 Decision

Given a belief structure, different criteria can be used to choose one hypothesis, such as the maximum of plausibility, or the minimization of some given risk. We will use here the pignistic risk minimization as defined and justified by Smets [11] on an axiomatic principle.

[1] The word was coined after the latin *pignus (a bet)*.

Let $\texttt{BetP}$ be the so-called pignistic probability distribution that is defined by uniformly distributing the belief masses given to any set of Ω among the elements of the corresponding sets:

$$\texttt{BetP}(\omega) = \sum_{A \ni \omega} \frac{m(A)}{|A|} \qquad \forall \omega \in \Omega \tag{3}$$

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

Now let $\{\alpha_0,\ \alpha_1, \ldots,\ \alpha_M\}$ be the set of all possible actions, where $\alpha_{i,\ i=1,\ldots,\ M}$ is the decision to classify x in class ω_i, and α_0 is rejection. We name $\lambda(\alpha_i|\omega_j)$ the cost of deciding α_i when the class of x is ω_j. In this paper, the loss incurred in case of a wrong classification is set to one, while the loss corresponding to correct classification is equal to zero. Rejection costs λ_0 (with $0 \leq \lambda_0 \leq 1$).

$$\lambda(\alpha_i|\omega_j)_{i=1,\ldots,\ M} = \begin{cases} 0 \text{ if } x \text{ is of class } \omega_i, \\ 1 \text{ otherwise} \end{cases} \tag{4a}$$
$$\lambda(\alpha_0|\omega_j) = \lambda_0 \tag{4b}$$

For any given α_i, $\texttt{Rbet}(\alpha_i|x)$ denotes the expected loss relative to the pignistic distribution and is computed as:

$$\texttt{Rbet}(\alpha_i|x) = \sum_{\omega_j \in \Omega} \lambda(\alpha_i|\omega_j)\ \texttt{BetP}(\omega_j) \tag{5a}$$
$$= \sum_{A \subseteq \Omega} \frac{m(A)}{|A|} \sum_{\omega_j \in A} \lambda(\alpha_i|\omega_j) \tag{5b}$$

The Bayesian decision theory then recommends the action α_i which has the lowest expected loss $\texttt{Rbet}(\alpha_i|x)$.

2.3 The Dempster-Shafer K-NN rule

Many K-NN rule extensions have been proposed in the literature, among them lies its extension in the scope of the Dempster-Shafer theory by Denœux [2, 3, 4] and Zouhal and Denœux [12]. We briefly give the outline below.

Let x be the sample to be classified, and $\mathcal{P} = \{(x_i,\ y_i)\}_{i=1}^{N}$ be the population of known patterns, where $y_i \in \Omega$ is the class of pattern x_i. First, the K-nearest neighbors of x in $\{x_i\}_{i=1}^{N}$ are selected according to the Euclidean distance. Each neighbor x_k is then considered as an item of evidence about the class of x, and it defines a corresponding belief structure m_k [2]:

$$\begin{aligned} m_k(\{y_k\}) &= \alpha \exp(-\gamma_q |x_k - x|^2) && \text{(6a)}\\ m_k(\Omega) &= 1 - \alpha \exp(-\gamma_q |x_k - x|^2) && \text{(6b)}\\ m_k(A) &= 0 \quad \forall A \subsetneq \Omega,\ A \neq \{y_k\} && \text{(6c)} \end{aligned}$$

where y_k is the class label of x_k and $|x_k - x|$ is the Euclidean distance between x_k and x.

This basic belief assignment has been discussed in great detail by Denœux [4]. Parameter $\alpha \in [0, 1]$ limits the certainty expressed by the prototypes by setting the minimum belief mass given to Ω. Parameters $\gamma_q \in \mathbb{R}^+$ adjusts the influence of the prototypes of class q according to their distance to x (the farther the prototype, the weaker the structure it gives [9]). The settings of α coefficient and γ vector have been the subject of a learning scheme [12].

As they are independent from each other, the K belief structures m_k can be combined into a single structure m by means of Dempster's rule. This structure can be regarded as a synthesis of the data provided by the neighborhood about the class of x. Finally m is used to compute pignistic probabilities $\mathtt{BetP}(\omega_k|x)$, from which class affectation can be made (see section 2.2).

3 Bagging the K-NN rule

Bagging is a procedure that builds an estimator by a resample-and-combine technique [1]. 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 that are sensitive to perturbations of the training set.

From the original estimation method, a bagged estimator is produced by averaging several replicates trained on bootstrapped learning sets [7]. The method often compares favorably with the original predictor, and also, in situations with substantial noise, with other ensemble methods such as boosting or randomization [6].

In discrimination, bagging is usually applied to the estimate of the decision rule. In this paper however we propose to use it upstream at the so-called credal level, in order to get a better estimate of the belief structure induced by data. The main goal is not only to allow steadier decisions but also to improve further combinations when new sources are available.

Practically, B bootstrap learning sets $\mathcal{P}_b$ are obtained by drawing with replacement N examples from the original learning set $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^N$ (some samples will be missing and some will be represented more than once in each set). This sampling simulates the original sampling that generated $\mathcal{P}$ from the true distribution. 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 $\mathcal{P}_b$ produces a belief structure m_b through a given K-NN classifier. These structures are finally aggregated into the average structure $m_{\mathbf{B}}$, defined as:

$$\forall A \subseteq \Omega, \quad m_{\mathbf{B}}(A) = \frac{1}{B} \sum_{b=1,B} m_b(A) \qquad (7)$$

The average aggregation operator is used for combining whereas majority vote on the B decisions would usually be prefered in the context of discrimination. Note that the Dempster rule of combination cannot be used because the belief sources are not independent, and thus, neither are the belief structures. Other choices could be made here, such as using the median operator or choosing the most precise belief structure among the B. However, averaging is a good candidate as it is idempotent, symmetrical 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 of the aggregation order (symmetry), and third, the linear relationship between credal and probabilistic levels advocates for a linear aggregation (linearity).

4 Experimental settings

4.1 Problem characteristics

For easy problems, with large training sets and well-separated clusters corresponding to classes, many different algorithms yield similar results. The learning set should then contain items that can be encountered in real applications, including ambiguous cases and outliers. Furthermore, it's size must be small for the classification to be difficult.

The learning set $\mathcal{P}$ is drawn using three bidimensional Gaussian classes of standard deviation 1.5, respectively centered on $(0,0)$, $(3,0)$ and $(0,5)$. This learning set is completed by some outliers, whose positions are uniformly distributed on $[-5,\ 9] \times [-3,\ 8]$ and classes are randomly chosen with equal probabilities. $\mathcal{P}$ contains respectively 15, 15, 15 and 6 items for the four underlying distributions ($N_{\mathcal{P}} = 51$). To exhibit general trends, the experiment is repeated 15 times with training sets generated from the same distribution.

4.2 Evaluation

For each training set, the evaluation is done on an independent test set $\mathcal{T}$ generated with respect to the same distribution than $\mathcal{P}$ with $N_{\mathcal{T}} = 2000 \times 3 + 800$ items.

Decision with reject option can be made according to two strategies. On one hand it can be based on a rejection cost λ_0 which should correspond to the labour cost for another system, such as for a human expert to decide. On the other hand, some applications such as postal services require that a given classification error rate is met. This can be achieved by setting λ_0 so as to reach this rate. For the optimal Bayes classifier, a rejection rate of at least 23% (with $\lambda_0 < 0.34$) is needed to achieve a classification error rate lower than 10%.

The mean classification cost $\mathbf{C}$ is estimated by the average of the classification costs on the $N_{\mathcal{T}}$ test points of $\mathcal{T}$. Costs are defined as in section 2.3: zero for correct classification, one for wrong classification and λ_0 for rejection. The classification error rate $\mathbf{E}$ is estimated by the proportion of bad predictions (rejection is not an error). Finally, the rejection rate $\mathbf{R}$ is estimated by the percentage of rejected items.

4.3 Implementation

The bagged Dempster-Shafer K-NN requires the setting of four parameters: K for the number of neighbors, α and γ to parametrize the basic belief assignments made by the prototypes, and finally B for the number of bootstrapped learning sets.

The method proposed by Denœux resembles Parzen's estimates when the neighborhood is constitued of the whole training set ($K = N_{\mathcal{P}}$), because the influence of a prototype is decreasing with its distance to the query point. Setting $K = 8$ achieves a near-asymptotical behavior while limiting the computational expense.

The basic belief assignment made by the prototypes according to their respective distance to x depends on the two parameters α and γ (see equ. 6). As the influence of α on the classification is low and in order to reduce the complexity of the analysis, its value is set to the standard 0.95 value [2]. In a first approach, an estimation for the optimum values of vector γ was proposed by Denœux as the mean intraclass distances. As the three classes have the same distribution and the same number of items, all γ_q were set to the same value (0.5).

Lastly, the average structure $m_{\mathbf{B}}$ is the bootstrap estimate of the expected structure over the bootstrapped training sets. In the bootstrap world, the expectation over training samples is 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 computing cost. Note that Breiman recommends values around 25.

5 Results

5.1 Decision level

5.1.1 Belief structure bagging

The first graph of figure 1 shows mean classification cost according to rejection cost for multiple data sets. The horizontal segments in boxplots respectively represent the lower quartile, median, and upperquartile results. Minimal and maximal values are shown by the whiskers, and the plotted curve itself is computed as the average of the different results.

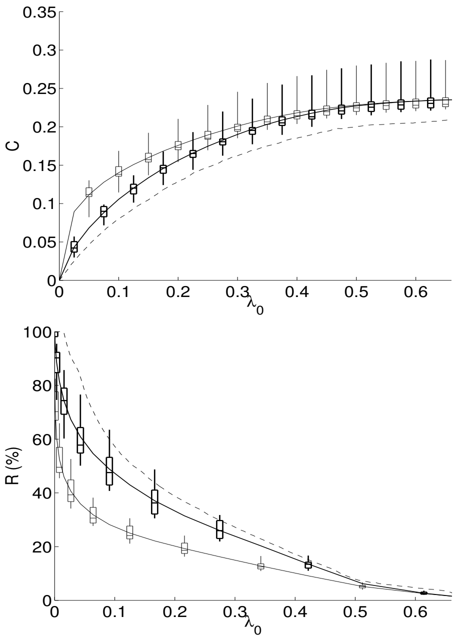

Figure 1: Mean classification cost **C** (top) and rejection rate **R** (bottom) according to rejection cost λ_0 for classic (thin line) and bagged (bold line) methods. Dotted lines show Bayes' optimum classifier.

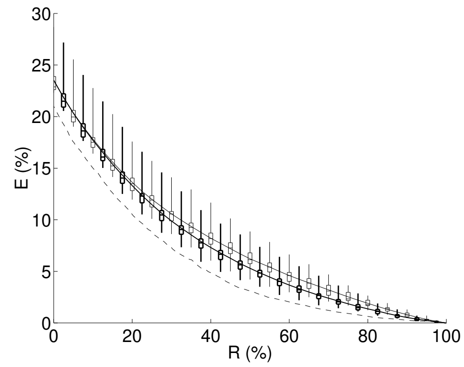

Figure 2: Classification error rate **E** according to rejection rate **R**

Table 1: Classification error **E** and rejection cost λ_0 for some given rejection rates **R** (single data set). **S** is the 5%-level significative difference threshold on **E**.

	Without bagging		With bagging		
R (%)	λ_0	**E** (%)	λ_0	**E**(%)	**S** (%)
15	0.273	14.8	0.390	14.5	1.3
20	0.158	12.7	0.310	12.4	1.3
25	0.086	10.9	0.236	10.4	1.2
30	0.046	9.7	0.174	8.6	1.2
35	0.025	8.8	0.133	7.4	1.1

Bagging neatly improves classification for $\lambda_0 < 0.3$ which corresponds to higher rejection rates. It halves the cost difference when compared to Bayes' optimum classifier. However, these differences vanish for low rejection rates (high values of λ_0): the improvement due to bagging seems to be linked with its higher capacity to reject ambiguous patterns (second graph of figure 1).

Regarding fixed cost decision strategy, classification is significatively improved for rejection rates around 30-70% (figure 2 and table 1).

For a proportion e of badly classified samples and a rejection rate r, the standard deviation of the mean test error rate is $\sqrt{\frac{e(1-e)}{(1-r)N_{\mathcal{T}}}}$. For all tested values of e and r, it is always lower than 0.5%. For a given learning set, two mean test error rates p_1 and p_2 differ significantly (at the 5% level), if $|p1 - p2| > 1.96\sqrt{p(1-p)\frac{2}{(1-r)N_{\mathcal{T}}}}$, where p is the proportion upon the null hypothesis $p_1 = p_2 = p$ [8].

Bagging is less sensitive to outliers and points in highly mixed regions: they are rejected more frequently. It gives better results for both fixed rejection cost and fixed rejection rate classifications, provided the latter is high enough (between 30% and 70% of the test set). Otherwise the differences are not significative, even if bagging never performs worse than the classic algorithm.

5.1.2 Decision Bagging

Given x, each bootstrapped training set generates a belief structure. These B structures are aggregated by averaging and the decision is based on this average belief structure. The faithful transposition of the original proposition of Breiman would be to perform a majority vote between the decisions provided by B classifiers.

This aggregation is a poor choice in the transferable belief model framework, as uncertainties are lost once decisions are made and also because combination with future items of evidence would become impossible. In fact, decision bagging does not even yield any improvement compared to the standard algorithm. This suggests that the latter already provides stable decision rules, which is in agreement with Breiman results on the standard K-NN [1]. Efforts should thus rather be directed towards reducing bias (systematic errors) rather than variability.

5.2 Credal level

It would be of great interest to study the effect of bagging at the credal level. However there is no satisfactory criteria for measuring the relevance of a belief structure. We thus resort to the study of pignistic probabilities which are a better summary of beliefs than the decisons themselves.

5.2.1 Influence regions of prototypes

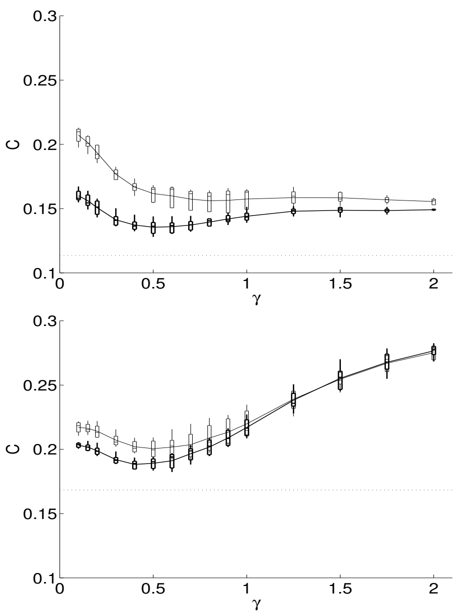

Figure 3: Mean classification cost $\mathbf{C}$ according to γ for $\lambda_0 = 0.15$ (top) and $\lambda_0 = 0.3$ (bottom). Dotted lines represent Bayes' optimum classification.

As presented in section 2.3, the influence of the prototypes in the neighborhood is controlled by γ. Figure 3 presents the mean classification cost according to γ for the two algorithms. Note that these curves, computed on the test set, could not be drawn in a real problem. Our goal here is to gain insight on the reasons of bagging success, not to propose a method for choosing γ.

The bagged K-NN mean classification cost according to γ is always lower than that of the original algorithm for any given rejection cost. Thus, for given values of γ, the mean classification cost versus rejection cost (top of figure 1) would also always be favourable to bagging.

The differences between the two methods are larger for small rejection costs, which correspond both to high rejection rates and to the highest differences in rejection rates (bottom of figure 1). These differences are also larger for small values of γ, where all neighbors have the same influence regardless of their distance to the query sample. This is a logical consequence of less precise probability estimates.

The two methods give similar results as $\gamma \to \infty$ where classification tends towards systematic rejection ($\mathbf{C} \to \lambda_0$). In fact, high γ values mean narrow kernels that lead to increased mass transfer towards Ω (see equ. 6).

It is clear from the boxplots that bagging reduces the variability of the optimal γ value. This may be useful when trying to learn its value. Moreover, the optimal γ value depends on the rejection cost λ_0 for the standard algorithm, but not for its bagged version. Ideally, these two values should not interact, as beliefs (representation) should be independent on the costs associated to actions (decision).

Thus, the pignistic probabilities $\mathtt{BetP}$ should be closer to the posterior probabilities P. Indeed, mean quadratic errors on *a posteriori* class probabilities[2] are 40% lower with bagging. Figure 4 show the probabilities meshes, where the main improvements are more faithful class boundaries and reduced outlier influence. This supports the hypothesis that more relevant belief structures are generated.

[2] Defined as $\int \sum_i (\mathtt{BetP}(w_i|x) - P(w_i|x))^2 dp(x)$

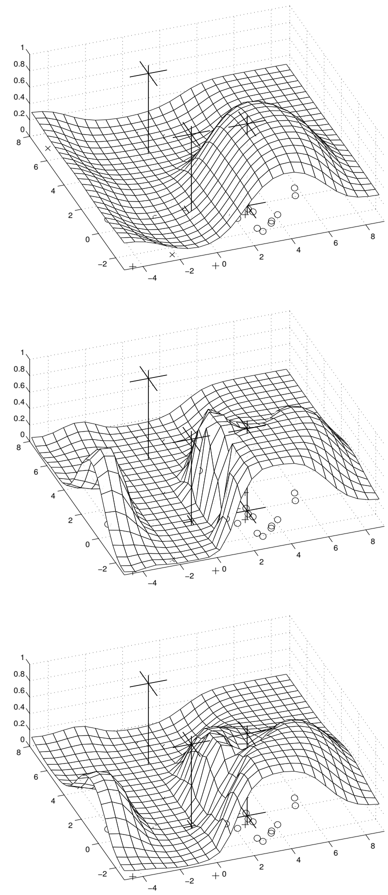

Figure 4: Posterior probability $P(\omega_2|x)$ (top) and pignistic probability $\mathtt{BetP}(\omega_2|x)$ without bagging (middle) and with bagging (bottom). Vertical lines locate the centers of the three class distributions.

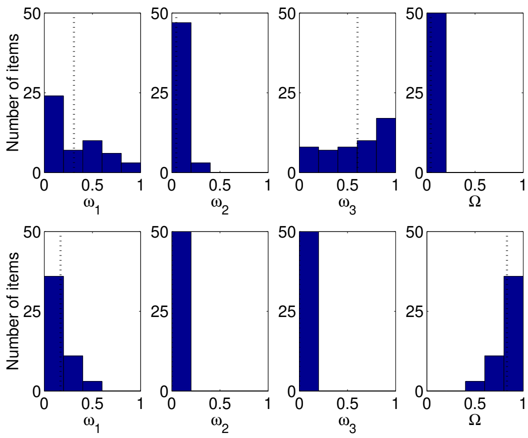

Figure 5: Mass distributions for a point in ambiguous 3-1 region (up) and an outlier (down). Vertical dotted lines indicates the average mass value.

5.2.2 Belief distributions

Figure 5 shows the mass distributions of the B structures associated to two points in the test set. These are given for each of the four hypotheses ω_1, ω_2, ω_3 and the vacuous hypothesis Ω for which no support is given to the other hypothesis.

For outliers in low-probability density regions, the masses on the three hypotheses ω_i 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 such absence of conflict in the neighborhood, the average structure is a good representation of the B belief structures both at the pignistic and credal levels.

In ambiguous zones, some belief structures that were produced on bootstrapped training sets gave a near-full mass on one hypothesis or another because of high variability in the neighborhood. Bagging through average distributes the belief mass between the classes in conflict, and provides a good compromise at the pignistic level. On the other hand however, no trace of this conflict remains at the credal level.

Extensions of the Dempster-Shafer theory such as fuzzy belief structures or interval-valued belief structures [5] could be used to keep a measure of discordance within the B structures. It could improve the quality of belief representation at the credal level, which is an important issue if combinations with other belief sources is foreseen.

6 Conclusion

In this paper we introduced a bagged version of the Dempster-Shafer K-NN rule. Classification is done on the aggregation of the generated belief structures.

The mean classification error is significantly reduced for high to intermediate rejection rates, and is otherwise always lower than that of the non-bagged K-NN rule.

Bagging makes the pignistic probabilities closer to the posterior probabilities. This is a clue that the relevance of the belief structures is improved. Further work on the combination with other sources may also contribute to this conclusion.

Finally, the estimation of α and γ belief assigment parameters can itself be computed on each of the bootstrapped training sets. Our results suggest that the learning algorithm proposed by Zouhal and Denœux [12] can be made more stable with respect to the dataset while giving better results.

References

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

[2] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):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*, 14(5):444–451, 1997.

[5] T. Denœux. Reasoning with imprecise belief structures. Technical report, UMR CNRS 6599 Heudiasyc 97/44, 1997.

[6] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. Technical report, Oregon State University, Corvallis, 1998.

[7] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993.

[8] G. Saporta. *Probabilités, analyse des données et statistique*. Edition Technip, Paris, 1990.

[9] G. Shafer. A mathematical theory of evidence. Technical report, Princeton University Press, 1976.

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

[11] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

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