Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules

Benjamin Quost, Marie-Hélène Masson, Thierry Denoeux

Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules[1]

Benjamin Quost, Marie-Hélène Masson and Thierry Denœux
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France

November 16, 2010

[1]This paper is a revised and extended version of [47].

Abstract

When combining classifiers in the Dempster-Shafer framework, Dempster's rule is generally used. However, this rule assumes the classifiers to be independent. This paper investigates the use of other operators for combining non independent classifiers, including the cautious rule and, more generally, t-norm based rules with behavior ranging between Dempster's rule and the cautious rule. Two strategies are investigated for learning an optimal combination scheme, based on a parameterized family of t-norms. The first one learns a single rule by minimizing an error criterion. The second strategy is a two-step procedure, in which groups of classifiers with similar outputs are first identified using a clustering algorithm. Then, within- and between-cluster rules are determined by minimizing an error criterion. Experiments with various synthetic and real data sets demonstrate the effectiveness of both the single rule and two-step strategies. Overall, optimizing a single t-norm based rule yields better results than using a fixed rule, including Dempster's rule, and the two-step strategy brings further improvements.

Keywords: Classification, Pattern Recognition, Machine-Learning, classifier ensemble, Dempster-Shafer theory, theory of evidence, Transferable Belief Model, belief functions, cautious rule.

1 Introduction

The use of multiple classifiers, also called classifier ensembles, is now recognized as a practical and efficient solution for solving complex pattern recognition problems [29, 4, 18, 49, 62]. The idea behind classifier ensembles is that different classifiers may potentially offer complementary information about patterns to be classified, allowing for potentially higher classification accuracy. Optimizing a classifier ensemble generally involves two main tasks [2]: creating a pool of classifiers, and combining their outputs. Before focusing on the latter issue, which is the main topic of this paper, we will first provide a brief survey of work related to the former one, which has received a lot of attention until now.

A lot of studies (e.g. [54, 33, 34]) have provided experimental evidence that ensembles could be more accurate than individual classifiers when the predictions of their members share a low level of dependence, or at least reflect some level of *diversity*. This concept of diversity is generally thought as the ability of the classifiers to make different errors on new data points [17, 25]. From a theoretical point of view, Tumer and Ghosh [60] have shown that reducing the correlation among classifiers that are combined increases the accuracy of the ensemble. For ensembles of classification trees, Breiman found that an upper bound of the ensemble error depends on the average pairwise correlation between members of the ensemble [5]. Measuring the diversity of an ensemble has thus become a challenging issue and several measures have been proposed. Most of them are reviewed in [34]. These measures are used to select members of the ensemble, using forward algorithms that add one classifier at a time, or using backward algorithms, which prune classifiers from a large set if the removal is not harmful [61]. Another approach consists in clustering the classifiers according to their diversity and retaining only one representative classifier in each cluster [22]. Until now, however, it is not quite clear what is the most suitable diversity measure and how diversity measures relate to the overall performance of the ensemble.

Whereas the measurement of diversity is still an open question, there is a general agreement on the ways of enforcing the diversity in an ensemble, among them [60, 18, 17, 7]:

- Using different classifiers: *hybrid* ensembles composed of various types of classifiers (e.g., neural networks with various architectures, k-nearest neighbors, decision trees, quadratic Bayes classifier) are likely to produce classifiers with different specialties and accuracies in different regions of the space;

- Resampling the training data: the most popular techniques are bagging [4], boosting [52] and cross-validation. All these methods operate by taking a base classifier and training it with different data sets obtained by resampling the original data set. The instability of the base classifier, i.e., the property that small changes in the training set will cause large changes in the learned classifier, is usually required to expect some diversity.

- Using different features: in some problems, it is possible to extract different groups of features and to train a separate classifier for each group. If the features from different groups are not too correlated, the combined classifiers can be expected to have high diversity. Another approach proposed in [17] is called

input decimation. It consists in selectively pruning some input features according to their effect of the output of the classifier.

- Injecting randomness: another way for generating diversity is to inject randomness into the learning algorithm [18, 3]. For example, in neural networks, the initial configuration of weights is chosen at random. If the algorithm is applied with the same training data but different initial weights, the resulting classifiers can be quite different. For decision trees, Dietterich proposed a procedure introducing randomness in the selection of the best split at each internal node, thereby introducing diversity in an ensemble of decision trees.

The second task in classifier ensembles, which is the focus of the paper, is the combination of the outputs from a pool of classifiers. Depending on the form of the information delivered by individual classifiers, a variety of schemes has been proposed for deriving a combined decision from individual ones, such as majority voting [51], Bayes combination [63, 32], fuzzy integrals [8, 44], multilayered perceptrons [64], or the Dempster-Shafer theory of belief functions [36, 63, 49, 39, 35, 45, 2, 46, 48]. We have chosen this latter formalism because it provides, as will be shown, powerful tools for representing and combining uncertain information. The starting point of our work is the following: since there is no real way to quantify the level of dependence between the members of the ensemble, it seems desirable to *optimize* the combination rule so as to automatically adapt to the level of dependence between the classifiers.

Most of the works based on belief functions use Dempster's rule of combination [59, 56] for fusing individual classifier outputs. Indeed, Dempster's rule plays a central role in the theory of belief functions. However, a major limitation of this rule comes from the requirement that the combined items of evidence be independent, or distinct [59]. As remarked by Dempster [9], the real-world meaning of this notion is difficult to grasp. The general idea is that, in the combination process, no elementary item of evidence should be counted twice. Thus, non overlapping random samples from a population are clearly distinct items of evidence, whereas "opinions of different people based on overlapping experiences could not be regarded as independent sources" [9]. Classifiers trained on non-overlapping data sets and based on independent features can thus be considered as independent. In contrast, classifiers trained on the same or overlapping datasets (using, e.g., different learning algorithm and/or resampling techniques) as well as classifiers based on correlated features cannot be considered as independent sources of information. Consequently, Dempster's rule may not be the best suited to combine the outputs from such classifiers.

The need for a rule allowing the combination of information coming from dependent sources has led to the introduction of the cautious rule of combination [13, 14], which avoids double counting the same information provided by overlapping bodies of evidence. It was also pointed out that both Dempster's rule and the cautious rule, when restricted to separable mass functions, may be seen as extreme elements of infinite families of combination rules based on triangular norms, or t-norms for short [14, 42, 43]. A parameterized family of combination rules can be defined, based on a corresponding family of t-norms. In this paper, we propose to select a rule among such a family by optimizing the classification performance of the ensemble. This approach is clearly in line with the conclusions of Ruta and Gabrys [50] for whom the classification accuracy of an ensemble is the only adequate measure of diversity. Additionally,

using an approach similar to that proposed by Gatnar [22], a two-step procedure is also proposed, in which classifiers are clustered according to the similarity of their outputs; a within-cluster rule and a between-cluster rule are then determined simultaneously.

The rest of this paper is organized as follows. The background on belief functions is first recalled in Section 2. The combination rule optimization methods are then presented in Section 3, and experimental results are reported in Section 4. Section 5 concludes the paper.

2 Background on Belief Functions

This section presents the necessary notions of the theory of belief functions used in the rest of the paper. The basic definitions are first recalled in Section 2.1. Sections 2.2 and 2.3 then present, respectively, the canonical decomposition of a belief function and the Least Commitment Principle. These two notions are at the origin of the cautious rule and its extensions, introduced in Sections 2.4 and 2.5, respectively.

2.1 Basic Definitions

Let Ω denote a finite set of answers to some question, called the *frame of discernment*. A body of evidence about the question under consideration may be quantified by a *mass function* m, defined as a mapping from 2^Ω to $[0,1]$ such that $\sum_{A\subseteq\Omega} m(A) = 1$ (here, 2^Ω denotes the power set of Ω). Any subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a *focal set* of m. A mass function is said to be *normalized* if $\emptyset$ is not a focal set. Any mass function m such that $m(\emptyset) < 1$ can be normalized by the following transformation:

$$\begin{cases} m^*(A) &= \dfrac{m(A)}{1-m(\emptyset)}, \quad \forall A \subseteq \Omega, A \neq \emptyset; \\ m^*(\emptyset) &= 0. \end{cases} \tag{1}$$

A mass function m has several equivalent representations [53]. Two of those are the *plausibility* and *commonality* functions defined, respectively, as:

$$pl(A) = \sum_{B\cap A\neq\emptyset} m(B), \quad \forall A \subseteq \Omega,$$

$$q(A) = \sum_{B\supseteq A} m(B), \quad \forall A \subseteq \Omega.$$

Conversely, m can be recovered from pl or q. For instance, the following equality holds:

$$m(A) = \sum_{A\subseteq B} (-1)^{|B|-|A|} q(B), \quad \forall A \subseteq \Omega. \tag{2}$$

If the focal sets of m are nested, m is said to be *consonant*. The following relation then holds [53]:

$$pl(A \cup B) = \max(pl(A), pl(B)), \quad \forall A, B \subseteq \Omega.$$

In particular,

$$pl(A) = \max_{\omega\in A} pl(\{\omega\}), \quad \forall A \subseteq \Omega.$$

The function $\omega \rightarrow pl(\{\omega\})$ is referred to as the *contour function* [53].

Two mass functions m_1 and m_2, provided by independent sources may be combined using the *conjunctive rule of combination*, also referred to as the *unnormalized Dempster's rule of combination* $\textcircled{\cap}$ [56]. This rule is defined as follows:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega. \tag{3}$$

Let $q_1 \textcircled{\cap} q_2$ denote the commonality function corresponding to $m_1 \textcircled{\cap} m_2$. It can be computed from q_1 and q_2, the commonality functions associated to m_1 and m_2, as follows:

$$(q_1 \textcircled{\cap} q_2)(A) = q_1(A) \cdot q_2(A), \quad \forall A \subseteq \Omega.$$

Let us now assume we learn that m_2 was provided by a spurious source of information, so that it should be *subtracted* from $m_1 \textcircled{\cap} m_2$. If m_2 is nondogmatic, i.e., if $m_2(\Omega) > 0$ or, equivalently, $q_2(A) > 0$ for all $A \subseteq \Omega$, then q_1 can be recovered as follows:

$$q_1(A) = \frac{(q_1 \textcircled{\cap} q_2)(A)}{q_2(A)}, \quad \forall A \subseteq \Omega.$$

Following Smets [58], we may write:

$$m_1 = (m_1 \textcircled{\cap} m_2) \textcircled{\emptyset} m_2,$$

where $\textcircled{\emptyset}$ is a "decombination" operator. It should be noted, however, that $m_1 \textcircled{\emptyset} m_2$ may not be a mass function, depending on the choice of m_1 and m_2.

In [59], Smets proposed a two-level model, called the *Transferable Belief Model* (TBM), in which items of evidence are quantified by mass functions and combined at the *credal* level, while decisions are made at the *pignistic* level (from the Latin word *pignus* meaning a bet). Once a decision has to be made, a mass function m is thus transformed into a *pignistic probability distribution* $BetP$. The pignistic transformation consists in normalizing m, and then distributing each mass $m^*(A)$ equally between the atoms $\omega_k \in A$:

$$BetP(\omega_k) = \sum_{\{A \subseteq \Omega, \omega_k \in A\}} \frac{m^*(A)}{|A|}, \quad \forall \omega_k \in \Omega. \tag{4}$$

2.2 Canonical Decomposition of a Belief Function

According to Shafer [53], a mass function is said to be *simple* if it has the following form

$$\begin{aligned} m(A) &= 1 - w_0 \\ m(\Omega) &= w_0, \end{aligned}$$

for some $A \subset \Omega$ and some $w_0 \in [0, 1]$. Let us denote such a mass function as A^{w_0}. The vacuous mass function may thus be noted A^1 for any $A \subset \Omega$. It is clear that

$$A^{w_0} \textcircled{\cap} A^{w_0'} = A^{w_0 w_0'}.$$

A mass function may be called *separable* if it can be obtained as the result of the conjunctive combination of simple mass functions. It can then be written:

$$m = \textcircled{\cap}_{A \subseteq \Omega} A^{w(A)},$$

with $w(A) \in [0, 1]$ for all $A \subset \Omega$.

Smets [58] showed that any non dogmatic mass function m may be uniquely expressed as the decombination of two separable mass functions:

$$m = \left(\textcircled{\cap}_{A \subset \Omega} A^{w_C(A)}\right) \textcircled{\varnothing} \left(\textcircled{\cap}_{A \subset \Omega} A^{w_D(A)}\right) \tag{5}$$

with $w_C(A) \in (0, 1]$, $w_D(A) \in (0, 1]$ and $\max(w_C(A), w_D(A)) = 1$ for all $A \subset \Omega$. Equation (5) is referred to as the *(conjunctive) canonical decomposition* of m. Let w denote the mapping from $2^\Omega \setminus \{\Omega\}$ to $(0, +\infty)$ defined as

$$w(A) = \frac{w_C(A)}{w_D(A)}, \quad \forall A \subset \Omega.$$

If m is separable, then $w_D(A) = 1$ and $w(A) \leq 1$ for all $A \subset \Omega$. Function w is called the *conjunctive weight function* associated to m [14]. It is a new equivalent representation of a non dogmatic mass function, which may be computed directly from m as follows:

$$w(A) = \prod_{A \subseteq B} q(B)^{(-1)^{|B|-|A|+1}}, \quad \forall A \subseteq \Omega, \tag{6}$$

or, equivalently:

$$\ln w(A) = -\sum_{A \subseteq B} (-1)^{|B|-|A|} \ln q(B), \quad \forall A \subseteq \Omega. \tag{7}$$

We notice the similarity with (2). Hence, as pointed out in [14], any procedure suitable for transforming q to m can be used to compute $\ln w$ from $-\ln q$.

Function w may have a simpler expression if m has a special form. For instance, let us consider a consonant mass function m. Let us denote $pl_k = pl(\{\omega_k\})$ for $k = 1, \ldots, K$, and let us assume, without loss of generality, that the ω_k are ordered in such a way that

$$1 \geq pl_1 \geq pl_2 \geq \cdots \geq pl_K > 0.$$

Then the corresponding weight function w was shown in [14] to be defined by

$$w(A) = \begin{cases} pl_1 & \text{if } A = \emptyset, \\ \dfrac{pl_{k+1}}{pl_k} & \text{if } A = \{\omega_1, \ldots, \omega_k\}, 1 \leq k < K, \\ 1 & \text{otherwise.} \end{cases} \tag{8}$$

Finally, we note that function w has a simple property with respect to the unnormalized Dempster's rule. Let w_1 and w_2 be two weight functions, and let $w_1 \textcircled{\cap} w_2$ denote the result of their $\textcircled{\cap}$-combination (i.e., the weight function corresponding to $m_1 \textcircled{\cap} m_2$). Then the following relation holds:

$$(w_1 \textcircled{\cap} w_2)(A) = w_1(A) w_2(A), \quad \forall A \subset \Omega. \tag{9}$$

2.3 Least Commitment Principle

The Least Commitment Principle (LCP) plays a central role in the theory of belief functions [57]. This principle states that, if several mass functions are compatible with some constraints, then the least committed (informative) one should be selected. To apply this principle, some informational ordering between mass functions has to be chosen. Several such orderings have been defined [19]. For instance, the q-ordering is defined as follows: we say that m_1 is q-more committed than m_2, and we note $m_1 \sqsubseteq_q m_2$, if

$$q_1(A) \leq q_2(A), \quad \forall A \subseteq \Omega. \tag{10}$$

In [14], an alternative ordering, called w-ordering, was defined based on the conjunctive weight function: m_1 is w-more committed than m_2 (noted $m_1 \sqsubseteq_w m_2$) if

$$w_1(A) \leq w_2(A), \quad \forall A \subset \Omega. \tag{11}$$

This ordering is stronger than the q-ordering, i.e.,

$$m_1 \sqsubseteq_w m_2 \Rightarrow m_1 \sqsubseteq_q m_2$$

for all m_1 and m_2; this implication is strict.

As an illustration of the LCP, let us assume that we want to guess an unknown mass function m from its pignistic probability distribution $BetP$. Obviously, there exist infinitely many solutions. However, using the LCP, we may consider the q-least committed element in the set of mass functions m whose pignistic probability distribution is $BetP$. As shown in [20], this problem admits a unique solution, which is the consonant mass function with the following contour function:

$$pl(\{\omega_k\}) = \sum_{\ell=1}^{K} p_k \wedge p_\ell, \tag{12}$$

where $\wedge$ denotes the minimum operator, and $p_k = BetP(\omega_k)$, $k = 1, \ldots, K$. The corresponding mapping from probability distributions to mass functions is referred to as the *inverse pignistic transformation*. The corresponding weight function can be computed from (8).

Another important application of the LCP using the w-ordering is recalled in the next section.

2.4 Cautious Rule

Let us assume that we receive two non dogmatic mass functions m_1 and m_2 from two information sources considered to be reliable. Our state of belief, after receiving these two pieces of information, should then be represented by a mass function m_{12} more informative than both m_1 and m_2.

Let us assume that the w-ordering is chosen to compare the information content of two mass functions. Let us denote by $\mathcal{S}_w(m)$ the set of mass functions m' such that $m' \sqsubseteq_w m$. We should then have $m_{12} \in \mathcal{S}_w(m_1)$ and $m_{12} \in \mathcal{S}_w(m_2)$ or, equivalently, $m_{12} \in \mathcal{S}_w(m_1) \cap \mathcal{S}_w(m_2)$. According to the LCP, the w-least committed element in $\mathcal{S}_w(m_1) \cap \mathcal{S}_w(m_2)$ should be chosen, if it exists. It was shown in [14] that this element

exists and is unique. It is the non dogmatic mass function m_{12} with the following weight function:

$$w_{12}(A) = w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega. \tag{13}$$

This defines a new rule, called the *cautious rule* and noted $\textcircled{\wedge}$. We have

$$m_{12} = m_1 \textcircled{\wedge} m_2 = \textcircled{\cap}_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)}.$$

As shown in [14], this rule is commutative, associative, and idempotent: for all m,

$$m \textcircled{\wedge} m = m.$$

Additionally, Dempster's rule $\textcircled{\cap}$ is distributive over $\textcircled{\wedge}$, i.e.,

$$(m_1 \textcircled{\cap} m_2) \textcircled{\wedge} (m_1 \textcircled{\cap} m_3) = m_1 \textcircled{\cap} (m_2 \textcircled{\wedge} m_3), \tag{14}$$

for all m_1, m_2 and m_3. This property explains why the cautious rule can be considered to be more relevant than Dempster's rule when combining overlapping items of evidence: if two sources provide mass functions $m_1 \textcircled{\cap} m_2$ and $m_1 \textcircled{\cap} m_3$ having some evidence m_1 in common, the shared evidence is not counted twice.

2.5 T-norm Based Rules

By comparing Equations (9) and (13), we notice that the unnormalized Dempster's rule is based on the *product* of weights, whereas the cautious rule is based on the *minimum*. In $[0, 1]$, these two operators are *t-norms* [31]. If we consider only separable mass functions, for which $w(A) \in [0, 1]$ for all $A \subset \Omega$, it is thus possible to generalize both the $\textcircled{\cap}$ and $\textcircled{\wedge}$ rules by using any t-norm instead of the product or the minimum [13, 14, 42]. As the minimum is the largest t-norm, the cautious rule is the w-least committed of all these rules when combining separable mass functions.

A family of combination operators generalizing Dempster's rule and the cautious rule was recently proposed in [28], based on a generalized discounting process. Another approach, which will be adopted here, is to consider a parameterized family of t-norms containing both the product and the minimum as special cases [14]. For instance, we may consider Frank's family of t-norms [31, page 108]:

$$x \top_s y = \log_s \left(1 + \frac{(s^x - 1)(s^y - 1)}{s - 1} \right), \tag{15}$$

where $\log_s$ defines the logarithm function with base $s > 0$. Here, each value of parameter s defines a t-norm: the minimum is retrieved in the limit as $s \to 0$, and the product as $s = 1$. To each value of s corresponds a t-norm $\top_s$ and a combining rule $\textcircled{\top}_s$ defined by:

$$m_1 \textcircled{\top}_s m_2 = \textcircled{\cap}_{A \subset \Omega} A^{w_1(A) \top_s w_2(A)}, \tag{16}$$

where m_1 and m_2 are separable mass functions. Obviously, $\textcircled{\top}_0 = \textcircled{\cap}$ and $\textcircled{\top}_1 = \textcircled{\wedge}$. All these rules inherit important properties from t-norms: they are commutative and associative, and they admit the vacuous mass function as neutral element (if we consider only their restriction to separable mass functions).

As pointed out in [14], classifiers often provide separable belief functions in real-world applications. While this property is assumed in this paper, our approach is

not limited to this particular case. A non-separable mass function is characterized by a canonical decomposition with some weights $w(A) > 1$. In this case, combination rules may still be defined. In [43], the notion of t-norm is extended to $(0; +\infty)$. Such an operator may then be applied to the conjunctive weights of non-separable bbas to combine them.

3 Application to Classifier Combination

In this section, we come back to the classifier combination problem introduced in Section 1. We assume that we have q classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(q)}$. When presented with an input pattern, each classifier computes a mass function $m_{(j)}$. This mass function is directly available if classifier $\mathcal{C}_{(j)}$ is an evidential classifier as introduced in [10, 12]. In the case of probabilistic classifiers, we propose to convert the output probabilities into consonant belief functions using the inverse pignistic transform (12). This approach is supported by both practical and theoretical arguments:

1. The cautious rule and other t-norm based rules cannot be directly applied to probabilities because they are dogmatic belief functions; applying the inverse pignistic transformation yields non dogmatic, separable belief functions. Furthermore, the method does not depend on any parameter.

2. If the outputs from probabilistic classifier are interpreted as betting probabilities, then under the TBM any belief function whose pignistic probability distribution equates the classifier output can be considered consistent with this output. The LCP can then be invoked to select the least informative of these belief functions (see [1] for a more detailed analysis of this argument).

Once the outputs of all classifiers have been converted into separable belief functions, an overall mass function m^s is finally computed by combining the q classifier output mass functions using a t-norm based rule $\textcircled{\scriptsize T}_s$ defined by (15) and (16):

$$m^s = m_{(1)} \textcircled{\scriptsize T}_s m_{(2)} \textcircled{\scriptsize T}_s \ldots \textcircled{\scriptsize T}_s m_{(q)}.$$

This scheme is represented graphically in Figure 1.

If the classifiers are assumed to be independent, then Dempster's rule should be chosen, corresponding to $s = 0$. If the classifiers are not independent, then other rules, such as the cautious rule or other t-norm based rules as introduced in Section 2.5, could yield better performances. In the affirmative, the question arises of how to optimize the rule so as to obtain the best performances. These questions are addressed in this section.

In Section 3.1, we will first present a preliminary experiment showing that Dempster's rule may indeed be outperformed by the cautious rule or other t-norm based rules when combining non independent classifiers. A method for learning a single combination rule will then be introduced in Section 3.2. Finally, a more complex two-step combination scheme involving two rules will be described in Section 3.3.

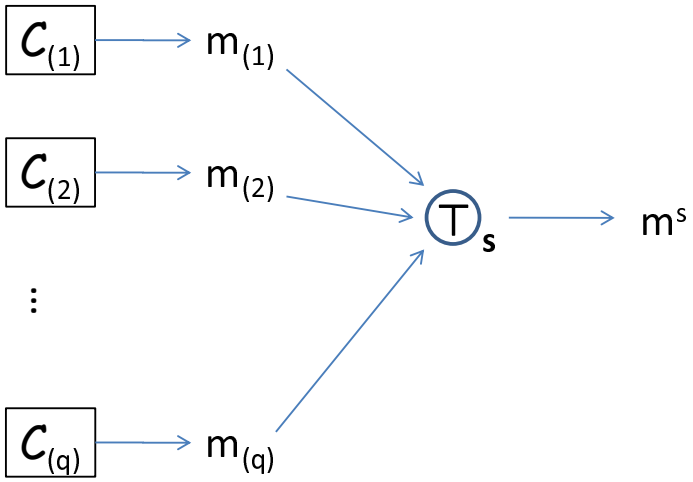

Figure 1: Classifier fusion using a single t-norm based rule.

3.1 Preliminary Experiment

To study the influence of classifier dependencies on the relative performances of various combination rules, we considered a classification problem with $K = 2$ classes and 10 features. Each feature is used as input to a separate classifier, so that we have 10 single-input classifiers. The conditional distribution of feature vector $(X_1, \ldots, X_{10})$ in class ω_k was assumed to be multivariate normal with mean $\mu_1 = (0, \ldots, 0)$ in class ω_1 and $\mu_2 = (1, \ldots, 1)$ in class ω_2, and with common variance matrix:

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \ldots & \rho & 0 \\ \rho & 1 & \rho & \ldots & \rho & 0 \\ \rho & \rho & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \rho & \vdots \\ \rho & \rho & \ldots & \rho & 1 & 0 \\ 0 & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix},$$

with $\rho \in [0, 1]$. Conditionally on each class, the last feature X_{10} was thus assumed to be independent from all other features, whereas the correlation coefficient between any two features X_i and X_j, $i, j \in \{1, \ldots, 9\}$ was equal to ρ.

This experimental framework is intended to mimic a real-world situation where we have $q - 1$ dependent classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(q-1)}$ and a q-th classifier $\mathcal{C}_{(q)}$ independent from the others. As Dempster's rule assumes independence between the first $q - 1$ classifiers, it is likely to give them too much weight in the decision.. Hence, using this rule is likely to give too much weight to the first $q - 1$ classifiers. In contrast, the cautious rule gives more importance, relatively, to the q-th independent classifier, which can be expected to result in better performance when the degree of dependence between classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(q-1)}$ is high. The purpose of this experiment is to seek an experimental confirmation of these intuitions.

For that purpose, we used a separate logistic regression classifier for each feature. We thus defined $q = 10$ classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(10)}$. For each input, the output probability distribution from each classifier was converted to a consonant mass function using the inverse pignistic transform (12). The 10 resulting mass functions were then combined using the $\textcircled{T}_s$ combination rule defined by (15) and (16).

The errors were measured as follows. Let m_i^s denote the combined mass function for example i and $BetP_i^s$ the corresponding pignistic probability distribution. The error for example i was defined as:

$$E_i^s = \sum_{k=1}^{K} (BetP_i^s(\omega_k) - \delta_{i,k})^2, \tag{17}$$

where $\delta_{i,k} = 1$ if pattern i belongs to class ω_k, and 0 otherwise. The average error over n patterns is then

$$E^s = \frac{1}{n} \sum_{i=1}^{n} E_i^s. \tag{18}$$

Each classifier was trained on a learning set of 2000 examples, and the error was evaluated on a test set of the same size. The simulations were repeated ten times, and the average error over the ten repetitions was computed. Figures 2 (a-c) display this error plotted as a function of s, for datasets generated using correlation coefficients $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. In these figures, Dempster's rule and the cautious rule correspond to the rightmost and the leftmost points of the x axis, respectively.

We can see that the best results were obtained for a rule close to Dempster's rule in the case where classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(9)}$ have low correlation ($\rho = 0.1$), whereas the cautious rule is optimal in the case of highly dependent classifiers ($\rho = 0.9$). When $\rho = 0.5$, the smallest error is obtained for an intermediate rule. Overal, no single rule is optimal in all cases, which points to the necessity of adapting the combination rule to the data. This problem will be addressed in the next subsection.

3.2 Learning a Combination Rule

The previous experiment has shown that, depending on the degree of dependence between classifiers, Dempster's rule may not be the best suited among t-norm based rules, and other operators such as the cautious rule or intermediate rules may have better classification performances. Assuming that the performances of different classifiers can be assessed on common data, it may be possible to *learn* a combination rule by minimizing an error criterion such as (17)-(18). This idea is investigated in this paper.

More specifically, assume that we have q classifiers $\mathcal{C}_{(1)}, \ldots, \mathcal{C}_{(q)}$ already trained using some learning sets, and let E^s be the error of the combined rule with t-norm parameter s evaluated on an independent *validation set*. Then, we choose the value $\widehat{s}$ of s with minimum validation error, i.e.,

$$\widehat{s} = \arg \min_{0<s\leq 1} E^s. \tag{19}$$

As the minimization of E^s is performed with respect to a single parameter in a bounded domain, a very simple search procedure can be used.

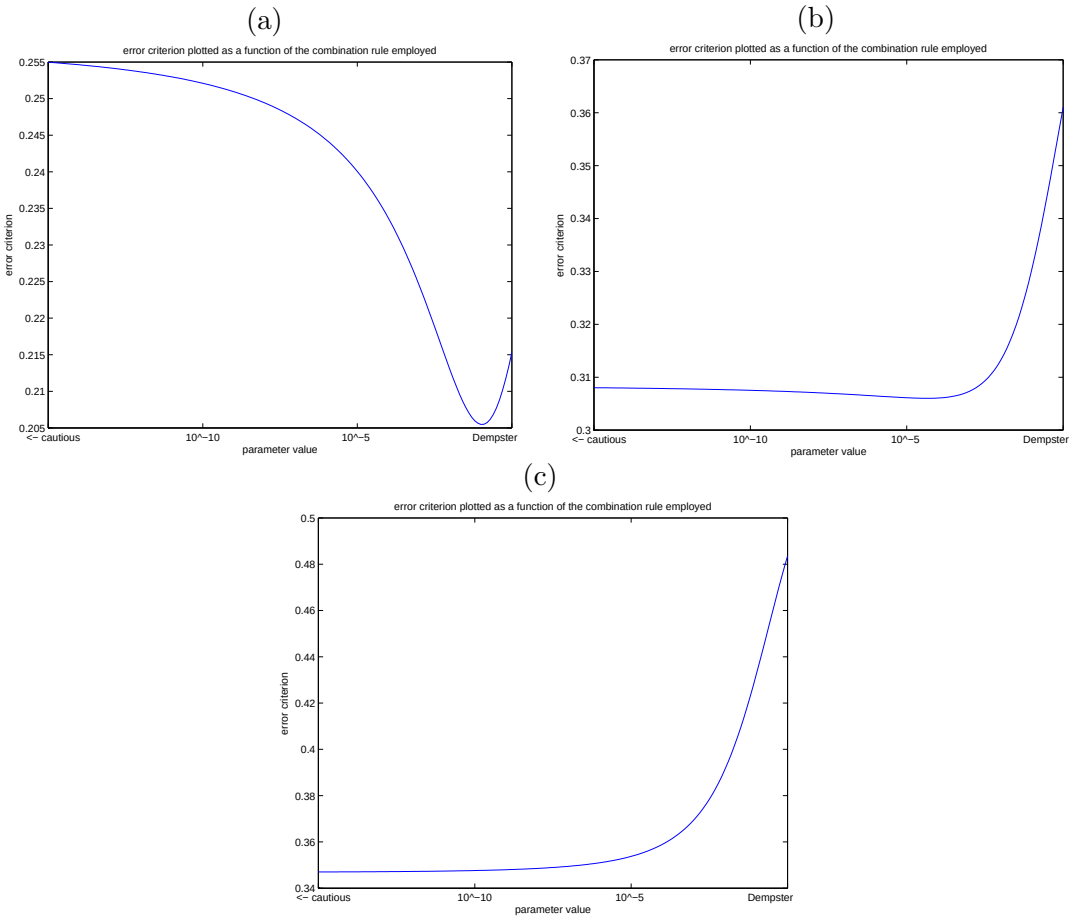

Figure 2: Error as a function of s for $\rho = 0.1$ (a), $\rho = 0.5$ (b) and $\rho = 0.9$ (c).

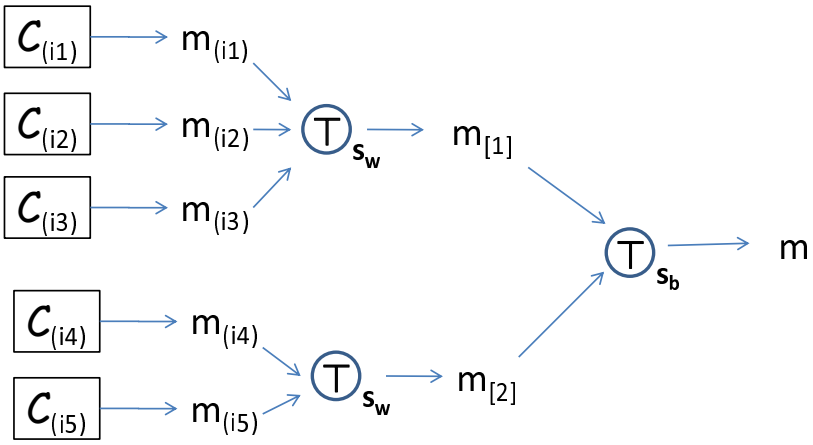

Figure 3: Classifier fusion using within-cluster and between-cluster t-norm based rules.

If no validation set is available or if there are too few learning examples to partition the data into a learning set and a validation set, we propose to estimate the error for each value of s using cross-validation. In that case, we partition the data into C subsets. Each of the C subsets then plays in turn the role of a validation set, while the union of the remaining $C-1$ subsets plays the role of a learning set and is used to train the classifiers. The cross-validation error for each value of s is then defined as the average of the C validation errors. As before, the value $\widehat{s}$ of s minimizing the cross-validation estimate of the error is finally selected.

Before presenting experimental results with this single-rule learning strategy in Section 4, a refined combination strategy that attempts to identify groups of dependent classifiers will now be introduced.

3.3 Two-step Combination Procedure

The method presented in Section 3.2 relies on a single rule for combining q classifiers. However, we have seen in Section 3.1 that Dempster's rule yields better results in the case of independent classifiers, whereas an operator close to the cautious rule is preferable in the case of highly dependent classifiers. Therefore, using a single rule may be too restrictive. A better strategy might be to identify clusters of "dependent" classifiers, and to use two rules: a *within-cluster* rule for combining dependent classifiers inside each cluster, and a *between-cluster* rule for pooling the combined outputs from each cluster. This fusion architecture with two clusters of three and two classifiers is depicted in Figure 3.

The idea of using a hierarchical fusion scheme based on a grouping of information sources has been explored by other researchers. In a study about climate sensitivity, Ha-Duong [24] proposed to combine expert opinions within given "schools of thought" using the cautious rule, and to use a disjunctive combination rule across different groups. Klein et al. [30] also employed two different rules in a computer vision application based on a grouping of sensors. None of these authors, however, considered the problem of automatically learning the optimal pair of rules from data.

The proposed clustering procedure will first be described in Section 3.3.1, and the learning procedure for learning the within and between-cluster rules will be presented in Section 3.3.2.

3.3.1 Clustering Classifiers

Meaningful groups of classifiers may be identified in different ways. Intuitively, classifiers should be grouped in such a way that there is more diversity between groups than there is inside each group. This brings us back to the issue of measuring diversity, which was discussed in Section 1. In this study, two approaches have been compared.

A first approach is to use a pairwise measure of diversity, as reviewed in [34]. In the experiments reported below, we have used the *disagreement* measure proposed in [55]. The disagreement $Dis_{k,\ell}$ between two classifiers $\mathcal{C}_{(k)}$ and $\mathcal{C}_{(\ell)}$ is defined as the percentage of observations for which one classifier is correct and the other one is incorrect.

Assuming that classifiers yielding similar outputs are more likely to be based on overlapping information, another approach consists in computing a distance measure between classifier output mass functions. The most widely used distance measure between mass functions was proposed by Jousselme [27]. It is defined as follows:

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2} \sum_{\substack{\emptyset \neq A \subseteq \Omega \\ \emptyset \neq B \subseteq \Omega}} \frac{|A \cap B|}{|A \cup B|} \left(m_1(A) - m_2(A)\right) \left(m_1(B) - m_2(B)\right)}, \tag{20}$$

where m_1 and m_2 are two normalized mass functions. The distance $\mathcal{D}_{k,\ell}$ between two classifiers $\mathcal{C}_{(k)}$ and $\mathcal{C}_{(\ell)}$ may be defined as the average distance between the output mass functions computed for the training patterns:

$$\mathcal{D}_{k,\ell} = \frac{1}{n} \sum_{i=1}^{n} d_J(m_{(k),i}, m_{(\ell),i}), \tag{21}$$

where $m_{(k),i}$ and $m_{(\ell),i}$ denote the mass functions for example i provided by classifiers k and ℓ, respectively.

Once pairwise dissimilarities between classifiers have been computed, a clustering algorithm can be used to identify groups of classifiers providing similar outputs. In this paper, we have used a hierarchical clustering algorithm (see, e.g. [26]), because this approach makes it possible to determine the number of clusters in a relatively easy way. However, other relational clustering techniques such as, e.g., the EVCLUS and RECM algorithms [15, 37] could be used.

As an illustration, Figure 4 shows the dendrogram [26] representing a hierarchy for the glass dataset. Choosing a cut value then allows us to find a partition of the set of classifiers. For instance, in Figure 4, cutting at level 0.3 yields three clusters: $\{\mathcal{C}_{(1)}, \mathcal{C}_{(5)}, \mathcal{C}_{(6)}, \mathcal{C}_{(7)}, \mathcal{C}_{(9)}\}$, $\{\mathcal{C}_{(2)}, \mathcal{C}_{(8)}\}$, and $\{\mathcal{C}_{(3)}, \mathcal{C}_{(4)}\}$.

We may notice that the dendrogram representation makes it possible to detect outlying classifiers quite easily. If such an outlier is included in the pool of classifiers and if it performs poorly, then it may degrade the overall performance. The impact of such classifiers on the global performance should thus be studied.

3.3.2 Learning Within and Between-Cluster Rules

After the classifiers have been clustered, we propose to combine the classifier outputs in two steps. First, their outputs are combined in the various clusters, and then the resulting bodies of evidence are pooled together.

Thus, two combination rules need now be learnt: a within-cluster rule for processing combination within each cluster, defined by a parameter value s_w; and a between-cluster rule for computing the final mass function, defined by a parameter value s_b. Taking the glass data as an example (Figure 4), the outputs of $\mathcal{C}_{(1)}$, $\mathcal{C}_{(5)}$, $\mathcal{C}_{(6)}$, $\mathcal{C}_{(7)}$, $\mathcal{C}_{(9)}$ are first combined using $\widehat{s_w}$, as well as those of $\mathcal{C}_{(2)}$, $\mathcal{C}_{(8)}$, and those of $\mathcal{C}_3$, $\mathcal{C}_4$; the three resulting mass functions are then pooled using $\widehat{s_b}$.

We propose to compute the pair of values $(\widehat{s_w}, \widehat{s_b})$ that minimizes the cross-validation error as follows. For each dataset, C-fold cross-validation is used to form training/validation sets from the original training set. Candidate values $a_1, \ldots, a_r$ equally spaced on a logarithmic scale, with $a_1 \approx 0$ and $a_r \approx 1$ are picked for s_b; for each a_i, candidate values $a_1, \ldots, a_i$ are considered for s_w, so that the resulting within-cluster rule is w-less committed than the between-cluster rule associated with s_b. Finally, we retain the pair of parameter values $(\widehat{s_b}, \widehat{s_w}) = (a_{i^*}, a_{j^*})$ that minimizes the error criterion (19) (averaged over the C validation sets). The number of evaluations of the error using C-fold cross-validation is thus $Cr(r-1)/2$. In our simulations, we used $C = 5$ and $r = 17$.

Note that more sophisticated learning schemes could be considered, such as learning a distinct rule within each cluster. This could be done either by maximizing a global performance criterion, or by maximizing the performances of each cluster independently. While the former approach would be very time consuming and data demanding, the latter is easier to implement but may be sub-optimal. The investigation of such fusion schemes and learning strategies is left for further research.

4 Experimental Results

We performed three series of experiments. First, the performances of various combination rules were compared to those of classifiers trained using a single feature as input. We then studied the behavior of the same rules applied to classifiers trained using randomly selected subsets of features. In a third series of experiments, we addressed the problem of hybrid classifier fusion, by combining three different classification algorithms, each one trained using the entire training set. The experimental setup will be described in Section 4.1, and the results will be presented and discussed in Sections 4.2 to 4.4.

4.1 Experimental setup

The datasets[1] used in these experiments are summarized in Table 1. Throughout the experiments, for each test pattern $\mathbf{x}$, each classifier $\mathcal{C}_{(k)}$ provided a probability distribution $p_{(k)}$ that was transformed into a mass function $m_{(k)}$ using the inverse

[1] These datasets may be found in the UCI Machine Learning repository at http://archive.ics.uci.edu/ml.

Table 1: Description of the datasets used in the experiments.

dataset	# classes	# features	number of patterns	
	K	p	training	test
glass	6	9	139	75
optdigits	10	64	2908	1797
pageblocks	5	10	3284	2189
pendigits	10	16	5640	3498
satimage	6	36	2921	2573
segment	7	19	1400	910
vowel	11	10	528	462
waveform	3	21	1491	3509

pignistic transformation (12). This output transformation has the advantage of producing separable mass functions that can be easily combined using the t-norm based rules introduced in Section 2.5.

For the single rule scheme described in Section 3.2 (hereafter referred to as OPT1), 5-fold cross validation was used to determine the optimal t-norm parameter. To implement the two-step combination procedure introduced in Section 3.3, two different measures were used for computing the dissimilarity between two classifiers: the average Jousselme distance (21) and the disagreement measure, both presented in Section 3.3.1. Hereafter, the corresponding rules will be referred to as OPT2 and OPT3, respectively. The classifiers were then grouped as explained in Section 3.3.1 using hierarchical clustering with Ward's criterion [26]. Optimal parameter values were selected by testing a grid of candidate values for each parameter.

The mass functions were combined using the average operator:

$$m_{\text{mean}} = \frac{1}{q} \sum_{k=1}^{q} m_{(k)}$$

as well as the conjunctive operators studied in this paper: Dempster's rule, the cautious rule, the single rule OPT1, and the two-step rules OPT2 and OPT3. Decisions were made based on the combined mass functions using the rule of maximum pignistic probability [11], except for the optdigits and pendigits datasets, where the rule of maximum plausibility was used because of computational issues. Finally, the probabilities provided by the classifiers were also combined using the average operator.

4.2 Classifiers trained using a single feature

In this section, we compare the performances of the various combination rules for classifiers trained using a single feature as input. There were thus as many classifiers as features. In this case, we employed logistic regression (see, e.g., [38]) as base classification method. Clustering results based on Jousselme's distance are displayed as dendrograms in Appendix A (Figures 4 to 11). Test error rates are shown in Table 2 for the glass, letter, optdigits and pageblocks datasets, and in Table 3 for the pendigits,

Table 2: Test error rates for the glass, optdigits, pageblocks and pendigits datasets. The best results are underlined; results that are not significantly different are printed in bold. For the OPT1, OPT2 and OPT3 methods, optimal parameter values are indicated in parentheses under the error rates.

data	Glass	Optdigits	Pageblocks	Pendigits
Dempster	**49.33**	**11.96**	10.19	**18.58**
OPT1 ($\widehat{s}$)	**45.33**	**11.13**	**8.59**	**18.55**
	(0.0e+00)	(0.0e+00)	(0.0e+00)	(0.0e+00)
OPT2 ($\widehat{s_w}, \widehat{s_b}$)	**45.33**	**11.13**	**8.59**	**18.55**
	(0.0e+00,0.0e+00)	(0.0e+00,0.0e+00)	(0.0e+00,1.0e-11)	(0.0e+00,1.0e-15)
OPT3 ($\widehat{s_w}, \widehat{s_b}$)	**45.33**	**11.13**	**8.59**	**18.52**
	(0.0e+00,0.0e+00)	(0.0e+00,0.0e+00)	(0.0e+00,1.0e-11)	(1.0e-15,1.0e+00)
cautious	**45.33**	**11.13**	**8.59**	**18.55**
averaging	**52.00**	14.69	10.23	23.24
proba. averaging	**52.00**	21.09	10.23	32.56
vote	66.67	85.70	10.23	58.06

segment, vowel and waveform datasets. The significance of the results was evaluated using a McNemar test [16] at the 5% level: the best result over all rules is underlined, and printed in bold together with results that are not significantly different. The analysis of the results presented in Tables 2 and 3 leads to the following comments.

First, we observe that Dempster's rule never gives the best results as a single rule for combining all the classifiers, whereas the cautious rule yields the best results in four cases: for the glass, optdigits, pageblocks and segment data sets. The OPT1 method recovers the cautious rule for all datasets but waveform where an intermediate rule (with $0 < \widehat{s} < 1$) is obtained. Overall, the OPT1 method never performed significantly worse than Dempster's rule or the cautious rule. These results show that the OPT1 method is generally a good strategy if a single combination rule is sought.

The results presented in Tables 2 and 3 also show that the OPT2 and OPT3 two-rule schemes generally yield the best results over the seven fusion methods investigated. In three cases (for the Pendigits, Satimage and Waveform data), the best results are obtained using a hierarchical combination scheme. The OPT2 and OPT3 schemes generally give almost identical results. Overall, the OPT1 method thus performs better than the other single rule schemes investigated, and the OPT2 and OPT3 methods bring further improvement (although the differences with OPT1 are not significantly different). This demonstrates the usefulness of the classifier combination approach introduced in this paper.

We may wonder how the classifier combination scheme studied in this section compares with single classifiers trained using all the features. Table 4 shows the test error rates obtained using logistic regression, the CART decision tree generation algorithm [6], and the evidential neural network [12]. By comparing these results with those reported in Tables 2 and 3, we can see that the single-feature combination strategy yields higher error rates than those obtained by the classifiers trained using all the features at once. However, a potential advantage of the fusion scheme investigated

Table 3: Test error rates for the satimage, segment, vowel and waveform datasets. The best results are underlined; results that are not significantly different are printed in bold. For the OPT1, OPT2 and OPT3 methods, optimal parameter values are indicated in parentheses under the error rates.

data	Satimage	Segment	Vowel	Waveform
Dempster	24.45	17.91	**56.06**	16.90
OPT1 ($\widehat{s}$)	**21.80**	**15.16**	**56.93**	**15.22**
	(0.0e+00)	(0.0e+00)	(2.4e-05)	(5.4e-02)
OPT2 ($\widehat{s_w}, \widehat{s_b}$)	**21.69**	**15.49**	**56.49**	**15.10**
	(1.0e-14,1.0e-07)	(0.0e+00,1.0e-03)	(1.0e-14,1.0e-04)	(1.0e-03,1.0e-03)
OPT3 ($\widehat{s_w}, \widehat{s_b}$)	**21.69**	**15.16**	**56.49**	**15.19**
	(1.0e-14,1.0e-07)	(0.0e+00,0.0e+00)	(1.0e-14,1.0e-03)	(1.0e-04,1.0e-01)
cautious	**21.80**	**15.16**	**56.93**	16.59
averaging	28.92	23.74	**52.60**	20.52
proba. averaging	28.92	23.74	**52.60**	20.52
vote	42.32	44.29	75.97	27.02

here is better robustness to missing feature values: when only a subset of features is available, we may only combine the classifiers corresponding to available features. To study this effect, the following experiment was carried out.

For each dataset, we randomly selected a chosen amount of the test data that were considered as missing. For the single classifiers trained using all the features at once, each missing value was replaced by the average of the corresponding feature, computed over the training set. For each dataset, the procedure was repeated 100 times.

Results obtained with the combination rules and individual classifiers are presented in Tables 5-6 (25% of missing test data), and in Tables 7-8 (50% of missing data). Confidence intervals computed over the 100 trials are reported.

When 25% of the data are missing, the best results are still obtained using a single classifier in six cases. The performances of Dempster's rule, the cautious rule, or the OPT1, OPT2 or OPT3 schemes are affected by the missing data, but the decrease is overall less important than for the other methods. When 50% of the data are missing, the best results are obtained using the cautious rule or the OPT1, OPT2 or OPT3 strategies in six cases, using probability averaging in one case (Vowel dataset) and using a single decision tree in one case (Pageblocks dataset). These results definitely confirm the interest of using the cautious rule, the single t-norm based rule or the two-step combination procedure introduced in this paper in a multiple sensor fusion scheme.

4.3 Classifiers trained using random subsets of features

We now study the combination of decision trees trained using subsets of the input data. Table 9 reports the total number of classifiers trained for each dataset. We used the CART algorithm [6] to train the decision trees. We randomly selected the

Table 4: Test error rates of the individual classifiers trained with all the features: logistic regression (LR), CART and evidential neural network (ENN). The best results are underlined; results that are not significantly different are printed in bold.

data	Glass	Optdigits	Pageblocks	Pendigits
LR	**44.00**	**7.96**	4.02	**7.43**
CART	**41.33**	16.42	**3.06**	9.61
ENN	**45.33**	**8.85**	10.23	16.64

data	Satimage	Segment	Vowel	Waveform
LR	**14.26**	18.02	**51.30**	13.94
CART	**14.34**	**6.70**	55.19	24.68
ENN	**13.91**	16.92	**47.84**	**13.28**

Table 5: Test error rates for the Glass, Optdigits, Pageblocks and Pendigits datasets; 25% of the test data are missing. The best results are underlined; results that are not significantly different are printed in bold.

data	Glass	Optdigits	Pageblocks	Pendigits
Dempster	53.32	14.08	10.14	22.24
	[52.67 ; 53.97]	[14.00 ; 14.16]	[10.13 ; 10.15]	[22.17 ; 22.31]
OPT1	**48.85**	13.44	9.03	22.01
	[48.19 ; 49.52]	[13.35 ; 13.52]	[9.01 ; 9.05]	[21.94 ; 22.08]
OPT2	**48.85**	13.44	9.03	22.03
	[48.19 ; 49.52]	[13.35 ; 13.52]	[9.01 ; 9.05]	[21.96 ; 22.10]
OPT3	**48.85**	13.44	9.03	22.13
	[48.19 ; 49.52]	[13.35 ; 13.52]	[9.01 ; 9.05]	[22.06 ; 22.20]
cautious	**48.85**	13.44	9.03	22.01
	[48.19 ; 49.52]	[13.35 ; 13.52]	[9.01 ; 9.05]	[21.94 ; 22.08]
averaging	53.09	16.84	10.23	26.19
	[52.47 ; 53.71]	[16.74 ; 16.94]	[10.23 ; 10.23]	[26.11 ; 26.26]
proba. averaging	53.09	22.68	10.23	34.51
	[52.47 ; 53.71]	[22.58 ; 22.78]	[10.23 ; 10.23]	[34.43 ; 34.58]
vote	64.05	85.13	10.23	60.04
	[63.34 ; 64.76]	[85.08 ; 85.19]	[10.23 ; 10.23]	[59.95 ; 60.13]
LR	56.33	17.29	8.64	38.61
	[55.28 ; 57.39]	[17.17 ; 17.42]	[8.56 ; 8.72]	[38.47 ; 38.74]
CART	54.13	42.52	**5.05**	40.78
	[53.23 ; 55.04]	[42.32 ; 42.72]	[4.99 ; 5.11]	[40.63 ; 40.92]
ENN	**49.83**	**12.42**	10.23	**21.28**
	[49.04 ; 50.61]	[12.31 ; 12.52]	[10.23 ; 10.23]	[21.20 ; 21.35]

Table 6: Test error rates for the Satimage, Segment, Vowel and Waveform datasets; 25% of the test data are missing. The best results are underlined; results that are not significantly different are printed in bold.

data	Satimage	Segment	Vowel	Waveform
Dempster	24.69	21.52	62.37	18.29
	[24.66 ; 24.73]	[21.35 ; 21.70]	[62.03 ; 62.71]	[18.22 ; 18.36]
OPT1	21.93	**19.04**	62.93	**17.48**
	[21.89 ; 21.97]	[18.86 ; 19.21]	[62.59 ; 63.27]	[17.41 ; 17.55]
OPT2	21.90	**19.34**	62.85	**17.51**
	[21.86 ; 21.94]	[19.16 ; 19.51]	[62.50 ; 63.20]	[17.43 ; 17.58]
OPT3	21.90	**19.04**	62.78	17.56
	[21.86 ; 21.94]	[18.86 ; 19.21]	[62.44 ; 63.12]	[17.48 ; 17.63]
cautious	21.93	**19.04**	62.65	18.55
	[21.89 ; 21.97]	[18.86 ; 19.21]	[62.30 ; 62.99]	[18.47 ; 18.64]
averaging	29.53	24.92	59.91	21.64
	[29.48 ; 29.58]	[24.75 ; 25.09]	[59.56 ; 60.25]	[21.57 ; 21.70]
proba. averaging	29.53	24.92	59.91	21.64
	[29.48 ; 29.58]	[24.75 ; 25.09]	[59.56 ; 60.25]	[21.57 ; 21.70]
vote	42.57	46.47	78.61	29.86
	[42.50 ; 42.64]	[46.31 ; 46.62]	[78.39 ; 78.83]	[29.76 ; 29.96]
LR	36.10	73.84	62.59	18.31
	[35.93 ; 36.26]	[73.55 ; 74.12]	[62.22 ; 62.96]	[18.20 ; 18.41]
CART	32.26	29.39	65.49	32.37
	[32.12 ; 32.40]	[29.11 ; 29.67]	[65.15 ; 65.83]	[32.25 ; 32.50]
ENN	**16.95**	33.92	**55.83**	**17.34**
	[16.88 ; 17.02]	[33.67 ; 34.16]	[55.51 ; 56.15]	[17.25 ; 17.44]

Table 7: Test error rates for the Glass, Optdigits, Pageblocks and Pendigits datasets; 50% of the test data are missing. The best results are underlined; results that are not significantly different are printed in bold.

data	Glass	Optdigits	Pageblocks	Pendigits
Dempster	55.60	18.17	10.08	28.95
	[54.76 ; 56.44]	[18.04 ; 18.29]	[10.06 ; 10.10]	[28.84 ; 29.07]
OPT1	**53.16**	**17.69**	9.44	**28.79**
	[52.27 ; 54.05]	[17.56 ; 17.82]	[9.41 ; 9.47]	[28.68 ; 28.90]
OPT2	**53.16**	**17.69**	9.45	**28.81**
	[52.27 ; 54.05]	[17.56 ; 17.82]	[9.42 ; 9.48]	[28.70 ; 28.91]
OPT3	**53.16**	**17.69**	9.44	**28.88**
	[52.27 ; 54.05]	[17.56 ; 17.82]	[9.41 ; 9.48]	[28.77 ; 29.00]
cautious	**53.16**	**17.69**	9.44	**28.79**
	[52.27 ; 54.05]	[17.56 ; 17.82]	[9.41 ; 9.47]	[28.68 ; 28.90]
averaging	54.91	21.08	10.27	31.99
	[54.05 ; 55.77]	[20.95 ; 21.22]	[10.25 ; 10.28]	[31.88 ; 32.10]
proba. averaging	54.91	26.67	10.27	39.31
	[54.05 ; 55.77]	[26.53 ; 26.81]	[10.25 ; 10.28]	[39.20 ; 39.41]
vote	62.47	84.04	10.26	63.91
	[61.59 ; 63.34]	[83.95 ; 84.13]	[10.24 ; 10.28]	[63.80 ; 64.02]
LR	58.36	34.41	10.69	62.40
	[57.29 ; 59.43]	[34.19 ; 34.64]	[10.60 ; 10.77]	[62.24 ; 62.56]
CART	60.04	64.54	**6.93**	63.07
	[59.06 ; 61.02]	[64.35 ; 64.72]	[6.88 ; 6.99]	[62.91 ; 63.22]
ENN	56.16	24.73	10.23	36.73
	[55.33 ; 56.99]	[24.57 ; 24.89]	[10.23 ; 10.23]	[36.59 ; 36.87]

Table 8: Test error rates for the Satimage, Segment, Vowel and Waveform datasets; 50% of the test data are missing. The best results are underlined; results that are not significantly different are printed in bold.

data	Satimage	Segment	Vowel	Waveform
Dempster	24.99	26.30	68.72	21.67
	[24.93 ; 25.05]	[26.12 ; 26.48]	[68.36 ; 69.08]	[21.57 ; 21.77]
OPT1	**22.42**	**25.27**	68.83	**21.33**
	[22.36 ; 22.47]	[25.10 ; 25.44]	[68.44 ; 69.23]	[21.22 ; 21.43]
OPT2	**22.46**	**25.37**	68.89	**21.49**
	[22.40 ; 22.52]	[25.19 ; 25.54]	[68.50 ; 69.28]	[21.38 ; 21.60]
OPT3	**22.46**	**25.27**	68.84	**21.43**
	[22.40 ; 22.52]	[25.10 ; 25.44]	[68.46 ; 69.22]	[21.32 ; 21.54]
cautious	**22.42**	**25.27**	68.82	22.16
	[22.36 ; 22.47]	[25.10 ; 25.44]	[68.44 ; 69.20]	[22.04 ; 22.28]
averaging	30.14	28.96	**67.52**	23.99
	[30.07 ; 30.21]	[28.78 ; 29.13]	[67.18 ; 67.87]	[23.90 ; 24.08]
proba. averaging	30.14	28.96	**67.52**	23.99
	[30.07 ; 30.21]	[28.78 ; 29.13]	[67.17 ; 67.87]	[23.90 ; 24.08]
vote	42.81	50.11	81.45	34.27
	[42.73 ; 42.90]	[49.92 ; 50.30]	[81.20 ; 81.71]	[34.16 ; 34.39]
LR	54.96	82.41	73.07	25.83
	[54.79 ; 55.13]	[82.18 ; 82.64]	[72.76 ; 73.38]	[25.70 ; 25.96]
CART	47.53	50.18	75.82	41.56
	[47.36 ; 47.69]	[49.91 ; 50.46]	[75.55 ; 76.09]	[41.43 ; 41.69]
ENN	29.86	55.00	69.36	24.86
	[29.75 ; 29.97]	[54.72 ; 55.28]	[68.99 ; 69.73]	[24.72 ; 25.00]

Table 9: Number of features and of classifiers for each dataset (classification trees).

dataset	# features	# classifiers
glass	9	5
optdigits	64	32
pageblocks	10	5
pendigits	16	8
satimage	36	18
segment	19	10
vowel	10	5
waveform	21	11

same amount of features to train each classifier, so that each feature was used at least once. We first trained decision trees from three features each, and then from seven features each. The classifiers were clustered automatically. For each dendrogram, the inconsistency coefficient of each link was computed. This coefficient characterizes the link by comparing its height with the average height of the links below it in the dendrogram. If the difference is high, the link is said to be inconsistent with the links below it. We chose to cut a link if its consistency was higher than 0.75.

The experiments were conducted on the Pageblocks, Satimage, Segment and Waveform datasets. For each dataset and each number of features per classifier, the experiment was repeated ten times. The average error rates are shown in Tables 10 (three features per classifier) and 11 (seven features per classifier). The significance of the results was evaluated using confidence intervals, in order to take into account the randomness in training the classifiers.

When the classifiers are trained from three features each, the best results are obtained by the hierarchical combination schemes in two cases out of four (for the Segment and Waveform datasets). In the other two cases, the averaging or voting operators achieve slightly better performances. When seven features are selected for each decision tree, the best results are always obtained using the averaging operator, although this is only statistically significant in one case.

These results may be explained by the nature of the combination strategies compared here. The rules in the conjunctive family are well suited to combine complementary information, which is obviously the case in the previous experiment. When the number of input features for each classifier increases, the accuracy of each classifier and the degree of overlap between the training data of the classifiers also increase. In such case, the conjunctive operators studied in this paper do not seem to offer any significant advantage over consensus operators such as averaging or majority voting.

4.4 Hybrid classifier fusion

In this last experiment, we combined three different learning algorithms (logistic regression, CART, and the evidential neural network [12]), each one trained using the whole sets of features. Here, the diversity in the ensemble stems from the nature of the algorithms employed.

Table 10: Test error rates and 95% confidence intervals for the pageblocks, satimage, segment and waveform datasets (decision trees trained using three features each). The best results are underlined; results that are not significantly different are printed in bold.

data	Pageblocks	Satimage	Segment	Waveform
Dempster	4.71	**13.85**	**7.77**	**21.18**
	[4.43 ; 4.99]	[13.60 ; 14.10]	[6.95 ; 8.58]	[20.56 ; 21.80]
OPT1	4.74	**13.95**	**7.73**	**21.29**
	[4.44 ; 5.04]	[13.68 ; 14.22]	[6.93 ; 8.54]	[20.66 ; 21.93]
OPT2	4.72	**14.13**	**7.78**	**21.21**
	[4.44 ; 4.99]	[13.74 ; 14.51]	[7.07 ; 8.48]	[20.45 ; 21.98]
OPT3	4.72	**14.04**	**7.77**	**21.17**
	[4.44 ; 5.00]	[13.74 ; 14.34]	[7.07 ; 8.46]	[20.45 ; 21.90]
cautious	5.09	14.78	**7.82**	25.80
	[4.81 ; 5.38]	[14.39 ; 15.18]	[7.04 ; 8.61]	[24.57 ; 27.04]
averaging	**4.27**	**13.64**	**8.88**	**21.89**
	[4.02 ; 4.52]	[13.29 ; 13.99]	[8.01 ; 9.74]	[21.02 ; 22.75]
proba. averaging	**4.27**	**13.64**	**8.88**	**21.89**
	[4.02 ; 4.52]	[13.29 ; 13.99]	[8.01 ; 9.74]	[21.02 ; 22.75]
vote	**4.00**	14.43	10.73	24.68
	[3.75 ; 4.25]	[14.08 ; 14.78]	[9.46 ; 11.99]	[22.97 ; 26.40]

The test error rates obtained using the various single combination rules as well as the individual classifiers are presented in Tables 12 and 13. Again, the significance of the results was evaluated using a McNemar test [16] at the 5% level: the best result over all rules is underlined, and printed in bold together with results that were not judged significantly different.

The best results are obtained using the average operator in three cases, the voting strategy in one case, Dempster's rule in one case and a single classifier in three cases. The differences between the various methods are not significant for the Glass dataset. Dempster's rule, the cautious rule, and the OPT1 combination strategy do not perform significantly worse that the decision tree for the Pageblocks and Segment datasets; and the OPT1 scheme does not perform significantly worse than Dempster's rule for the Satimage dataset.

These results confirm the observations made in the previous section. When different classifiers are trained using the same data, differences between their outputs only occasionally arise in some particular regions of the input space. Complex fusion schemes such as proposed in this paper may then not be justified, as compared to simple consensus operators such as averaging or majority voting.

5 Conclusion

The problem of combining classifiers (or, more generally, information sources) has been addressed within the framework of Dempster-Shafer theory. Although Dempster's rule

Table 11: Test error rates and 95% confidence intervals for the pageblocks, satimage, segment and waveform datasets (decision trees trained using seven features each). The best results are underlined; results that are not significantly different are printed in bold.

data	Pageblocks	Satimage	Segment	Waveform
Dempster	**3.37**	12.68	**6.31**	**20.90**
	[3.16 ; 3.57]	[12.31 ; 13.05]	[5.43 ; 7.18]	[19.94 ; 21.87]
OPT1	**3.36**	12.82	**6.36**	**20.87**
	[3.17 ; 3.55]	[12.43 ; 13.21]	[5.49 ; 7.23]	[19.91 ; 21.83]
OPT2	**3.51**	13.07	**6.04**	**20.89**
	[3.18 ; 3.84]	[12.67 ; 13.47]	[5.07 ; 7.02]	[20.00 ; 21.77]
OPT3	**3.48**	13.08	**6.18**	**20.78**
	[3.14 ; 3.82]	[12.70 ; 13.45]	[5.25 ; 7.12]	[19.82 ; 21.73]
cautious	3.74	13.28	**6.44**	23.34
	[3.46 ; 4.03]	[12.86 ; 13.70]	[5.32 ; 7.56]	[22.59 ; 24.08]
averaging	**3.17**	**12.06**	**5.79**	**20.00**
	[3.00 ; 3.33]	[11.88 ; 12.23]	[4.65 ; 6.93]	[19.26 ; 20.74]
proba. averaging	**3.17**	**12.06**	**5.79**	**20.00**
	[3.00 ; 3.33]	[11.88 ; 12.23]	[4.65 ; 6.93]	[19.26 ; 20.74]
vote	**3.20**	**12.12**	**6.13**	**20.45**
	[3.06 ; 3.35]	[11.91 ; 12.32]	[4.85 ; 7.42]	[19.67 ; 21.23]

Table 12: Test error rates (hybrid classifier ensemble) for the Glass, Optdigits, Pageblocks and Pendigits datasets. The best results are underlined; results that are not significantly different are printed in bold. For the OPT1 method, optimal parameter values are indicated in parentheses under the error rates.

data	Glass	Optdigits	Pageblocks	Pendigits
Dempster	**38.67**	12.30	**3.65**	6.86
OPT1 ($\widehat{s}$)	**38.67**	12.35	**3.70**	6.58
	(1.0e+00)	(8.8e-01)	(8.5e-01)	(1.0e+00)
cautious	**36.00**	12.91	**3.65**	6.98
averaging	**34.67**	**7.07**	**3.65**	**5.40**
proba. averaging	**34.67**	**7.51**	**3.65**	5.77
vote	**37.33**	**7.23**	3.75	6.49
LR	**44.00**	**7.96**	4.02	7.43
CART	**41.33**	16.42	**3.06**	9.61
ENN	**45.33**	8.85	10.23	16.64

Table 13: Test error rates (hybrid classifier ensemble) for the Satimage, Segment, Vowel and Waveform datasets. The best results are underlined; results that are not significantly different are printed in bold. For the OPT1 method, optimal parameter values are indicated in parentheses under the error rates.

data	Satimage	Segment	Vowel	Waveform
Dempster	**11.97**	**6.81**	51.52	15.05
OPT1 ($\widehat{s}$)	**12.01**	**6.81**	51.52	14.14
	(6.1e-01)	(8.2e-01)	(1.0e+00)	(0.0e+00)
cautious	12.86	**6.92**	51.73	14.14
averaging	**12.16**	**6.92**	**48.05**	14.53
proba. averaging	**12.16**	**6.92**	**48.05**	14.53
vote	**12.32**	8.57	**44.37**	**13.71**
LR	14.26	18.02	51.30	13.94
CART	14.34	**6.70**	55.19	24.68
ENN	13.91	16.92	**47.84**	**13.28**

plays a central role in this theory, it is well known that it relies on the assumption of independence, or distinctness, of the items of information, a condition rarely met in classification problems.

The cautious rule was recently introduced as an alternative to Dempster's rule, for combining non distinct items of evidence. If we restrict ourselves to the combination of separable mass functions, both Dempster's rule and the cautious rule may be seen as particular members of a family of rules based on t-norms. By considering a parameterized family of t-norms, it is thus possible to define a corresponding parameterized family of rules for combining separable mass functions. The problem of learning such rules from data has been investigated in this paper.

Two strategies have been studied. In the first one, a single rule is determined by minimizing an error criterion, computed either from validation data, or using a cross-validation procedure. In the second strategy, classifiers are partitioned using a hierarchical clustering algorithm, so that classifiers producing similar outputs belong to the same clusters. Classifiers inside each cluster are then combined using a within-cluster rule, and the combined results within each cluster are finally pooled using a between-cluster rule. Both rules are taken from the same t-norm based family and optimized simultaneously by minimizing an error criterion.

The strategies proposed in this article was compared to various combination rules through numerous experiments. When the classifiers provide complementary information, results demonstrate the effectiveness of the proposed scheme for learning a single rule, the optimized rule often providing better results than any of the fixed rules investigated, including Dempster's rule, the cautious rule, and simple averaging. The two-step strategy was shown to bring further improvements and was found to be the best of the fusion schemes studied. Additionally, experiments clearly demonstrated the robustness of the cautious rule, the t-norm-based combination strategy and the hierarchical combination schemes to missing data. When combining highly redundant information, such as the outputs of classifiers trained using highly overlapping or iden-

tical data, then the sophisticated fusion rules investigated in this paper do not seem to offer any significant advantage, in most cases, over simple consensus operators such as averaging or majority voting.

Although Dempster-Shafer theory was recently enriched with new combination rules, including the cautious rule and its extension, their interest had remained, until now, mainly theoretical, and the practical usefulness of these rules remained to be investigated. This paper has filled this gap by showing that these new rules can indeed be used to develop more efficient classifier combination strategies.

This work could be expanded in several directions. More complex models involving a separate within-cluster rule for each group of classifiers could be investigated. Beside combination rules, the discounting operation [53] is an efficient mechanism within Dempster-Shafer theory for taking into account the reliability of sources in information fusion problems. Method for learning discount rates were studied in [21] and [23], and the discounting operation was generalized in [41], [28] and [40]. More sophisticated classifier fusion schemes could be devised by optimizing both the combination rules and discount rates attached to each of the classifiers, making it possible to automatically discard uninformative classifiers. Research in these directions will be reported in future publications.

References

[1] A. Aregui and T. Denœux. Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, 49(3):575–594, 2008.

[2] Y. Bi, J. Guan, and D. Bell. The combination of multiple classifiers using an evidential reasoning approach. *Artificial Intelligence*, 172(15):1731–1751, October 2008.

[3] P. Bonissone, J. M. Cadenas, M. C. Garrido, and R. Andrés Díaz-Valladares. A fuzzy random forest. *International Journal of Approximate Reasoning*, 51(7):729–747, 2010.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6:5–20, 2005.

[8] S. Cho and J. Kim. Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Transactions on System, Man and Cybernetics*, 25(2):380–384, 1995.

[9] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[10] 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.

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

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

[13] T. Denœux. The cautious rule of combination for belief functions and some extensions. In *Proceedings of the 9th International Conference on Information Fusion*, Florence, Italy, 2006.

[14] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

[15] T. Denœux and M.-H. Masson. EVCLUS: Evidential clustering of proximity data. *IEEE Trans. on Systems, Man and Cybernetics B*, 34(1):95–109, 2004.

[16] T. G. Dietterich. Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[17] T. G. Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems, LNCS 1857*, pages 1–15. Springer-Verlag, 2000.

[18] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000.

[19] D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12(3):193–226, 1986.

[20] D. Dubois, H. Prade, and P. Smets. A definition of subjective possibility. *International Journal of Approximate Reasoning*, 48(2):352–364, 2008.

[21] Z. Elouedi, K. Mellouli, and Ph. Smets. Assessing sensor reliability for multisensor data fusion within the Transferable Belief Model. *IEEE Transactions on Systems, Man and Cybernetics B*, 34(1):782–787, 2004.

[22] E. Gatnar. Cluster and select approach to classifier fusion. In *Advances in Data Analysis*, pages 59–66. Springer-Verlag, Berlin, Heidelberg, 2007.

[23] H. Guo, W. Shi, and Y. Deng. Evaluating sensor reliability in classification problems based on evidence theory. *IEEE Transactions on Systems, Man and Cybernetics B*, 36(5):970–981, 2006.

[24] M. Ha-Duong. Hierarchical fusion of expert opinions in the Transferable Belief Model, application to climate sensitivity. *International Journal of Approximate Reasoning*, 49(3):555–574, 2008.

[25] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1000, 1990.

[26] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Englewood Cliffs, NJ., 1988.

[27] A.-L. Jousselme, D. Grenier, and É. Bossé. A new distance between two bodies of evidence. *Information Fusion*, 2:91–101, 2001.

[28] A. Kallel and S. Le Hégarat-Mascle. Combination of partially non-distinct beliefs: The cautious-adaptive rule. *International Journal of Approximate Reasoning*, 50(7):1000–1021, 2009.

[29] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[30] J. Klein, C. Lecomte, and P. Miché. Hierarchical and conditional combination of belief functions induced by visual tracking. *International Journal of Approximate Reasoning*, 51(4):410–428, 2010.

[31] E. P. Klement, R. Mesiar, and E. Pap. *Triangular norms*. Kluwer Academic Publishers, Dordrecht, 2000.

[32] L. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001.

[33] L. I. Kuncheva, M. Skurichina, and R.P.W. Duin. An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion*, 3:245–258, 2002.

[34] L. I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.

[35] C.A. Le, V.-N. Huynh, A. Shimazu, and Y. Nakamori. Combining classifiers for word sense disembiguation based on Dempster-Shafer theory and OWA operators. *Data and Knowledge Engineering*, 63(2):381–396, 2007.

[36] E.J. Mandler and J. Schurmann. Combining the classification results of independent classifiers based on Dempster-Shafer theory of evidence. *Pattern Recognition and Artificial Intelligence*, 10:381–393, 1988.

[37] M.-H. Masson and T. Denoeux. RECM: relational evidential c-means algorithm. *Pattern Recognition Letters*, 30(11):1015–1026, 2009.

[38] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, New-York, 1992.

[39] D. Mercier, G. Cron, T. Denoeux, and M. Masson. Fusion of multi-level decision systems using the Transferable Belief Model. In *Proceedings of FUSION'2005*, Philadelphia, USA, 2005.

[40] D. Mercier, E. Lefèvre, and F. Delmotte. Belief functions contextual discounting and canonical decompositions. *International Journal of Approximate Reasoning*, 2010. To appear.

[41] D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246–258, 2008.

[42] F. Pichon and T. Denœux. T-norm and uninorm-based combination of belief functions. In *Proceedings of the International Conference of the North American Fuzzy Information Processing Society (NAFIPS '08)*, pages 19–22, 2008.

[43] F. Pichon and T. Denœux. The unnormalized Dempster's rule of combination: a new justification from the least commitment principle and some extensions. *Journal of Automated Reasoning*, 45(1):61–87, 2010.

[44] N. J. Pizzi and W. Pedrycz. Aggregating multiple classification results using fuzzy integration and stochastic feature selection. *International Journal of Approximate Reasoning (in press)*, 2010. doi:10.1016/j.ijar.2010.05.003.

[45] B. Quost, T. Denœux, and M.-H. Masson. Pairwise classifier combination using belief functions. *Pattern Recognition Letters*, 28(5):644–653, April 2007.

[46] B. Quost, T. Denœux, and M.-H. Masson. Adapting a combination rule to non-independent information sources. In L. Magdalena, M. Ojeda-Aciego, and J.L. Verdegay, editors, *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'08)*, pages 448–455, Málaga, Spain, 2008.

[47] B. Quost, M.-H. Masson, and T. Denœux. Refined classifier combination using belief functions. In *Proceedings of the 10th International Conference on Information Fusion (Fusion'08)*, pages 776–782, Cologne, Germany, 2008.

[48] M. Reformat and R. Yager. Building ensemble classifiers using belief functions and OWA operators. *Soft Computing*, 12(6):543–558, 2008.

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

[50] D. Ruta and B. Gabrys. New measure of classifier dependency in multiple classifier systems. In *Multiple classifier systems,LNCS 2364*, pages 127–136. Springer-Verlag, Berlin, Heidelberg, 2002.

[51] D. Ruta and G. Gabrys. Classifier selection for majority voting. *Information Fusion*, 6(1):63–81, 2005.

[52] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of statistics*, 26(5):1651–1686, 1998.

[53] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[54] A. J. C. Sharkey and Noel E. Sharkey. Combining diverse neural nets. *The Knowledge Engineering Review*, 12:231–247, 1997.

[55] D. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *Proceedings of the American Association for Artificial Intelligence (AAAI'96), Integrating Multiple Learned Models Workshop*, pages 120–125, 1996.

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

[57] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9(1):1–35, 1993.

[58] P. Smets. The canonical decomposition of a weighted belief. In *Proceedings of the International Joint Conferences in Artificial Intelligence*, pages 1896–1901, Montréal, Canada, 1995.

[59] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.

[60] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 3-4(8):385–404, 1996.

[61] A. Ulas, M. Semerci, O.T. Yildiz, and E. Alpaydin. Incremental construction of classifier and discriminant ensembles. *Information Sciences*, 179:1298–1318, 2009.

[62] D.H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

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

[64] C.Y. Suen Y.S. Huang, K. Liu. The combination of multiple classifiers by a neural network approach. *Int. J. Pattern Recognition and Artificial Intelligence*, 9(3):579–597, 1995.

A Dendrograms of single-feature classifiers

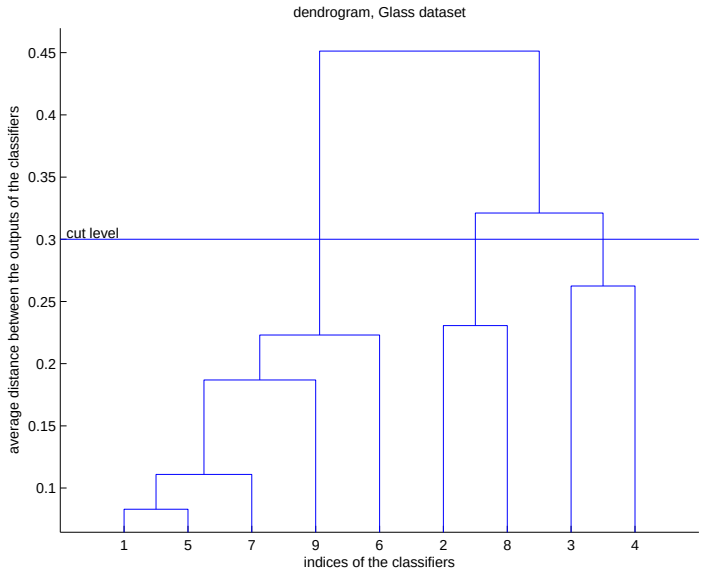

Figure 4: Dendrogram: glass data set (logistic regression).

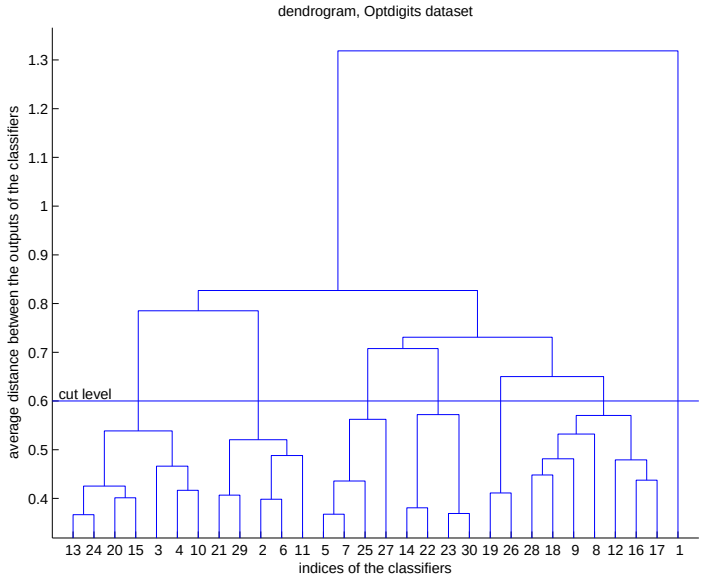

Figure 5: Dendrogram: optdigits data set (logistic regression).

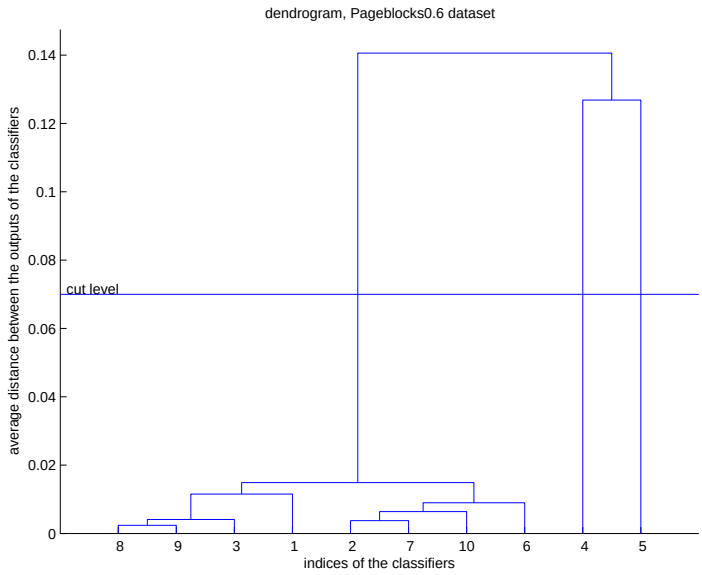

Figure 6: Dendrogram: pageblocks data set (logistic regression).

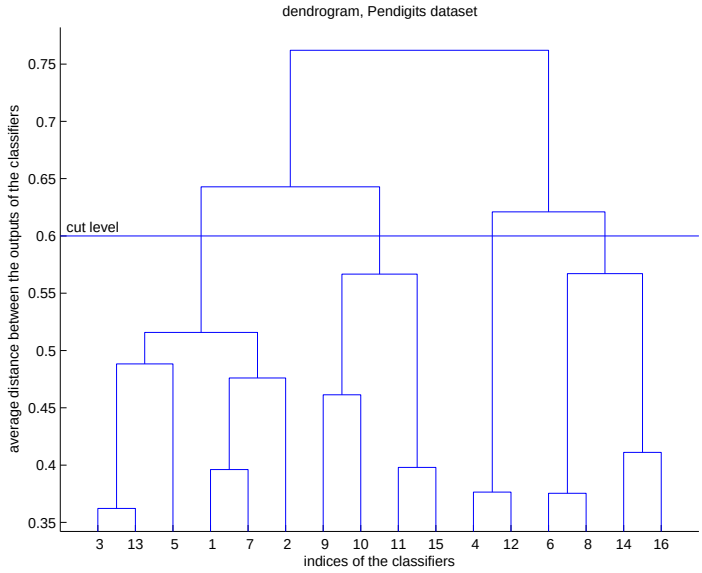

Figure 7: Dendrogram: pendigits data set (logistic regression).

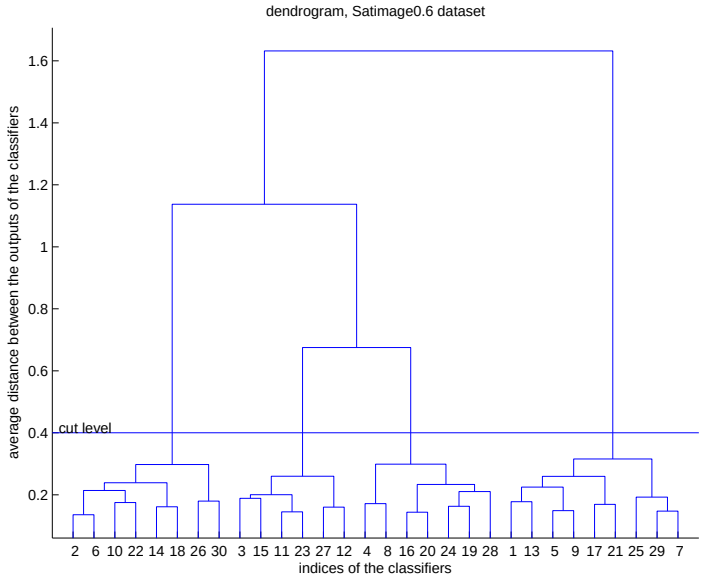

Figure 8: Dendrogram: satimage data set (logistic regression).

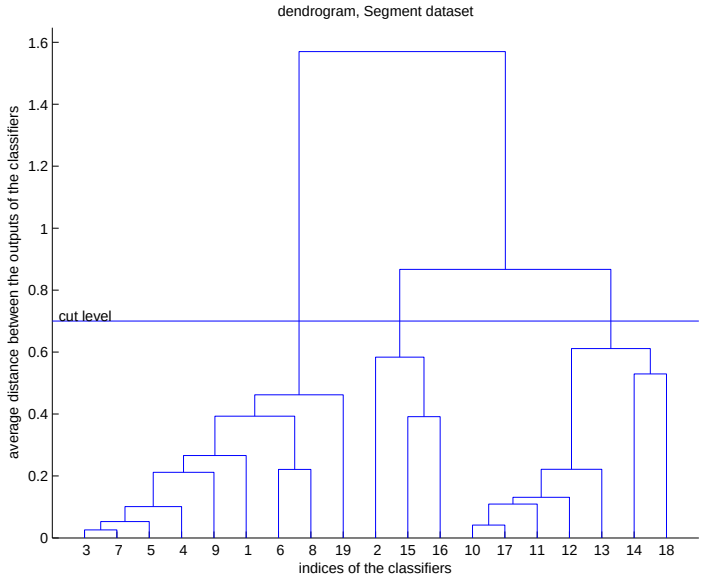

Figure 9: Dendrogram: segment data set (logistic regression).

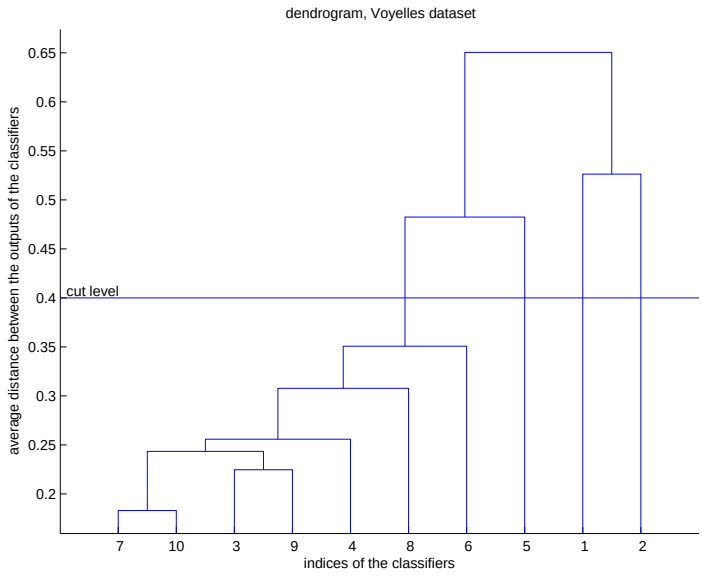

Figure 10: Dendrogram: vowel data set (logistic regression).

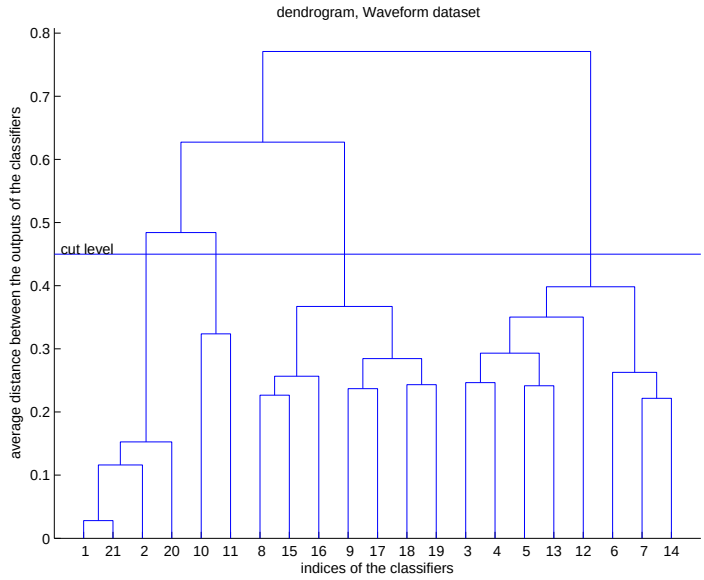

Figure 11: Dendrogram: waveform data set (logistic regression).