Refined classifier combination using belief functions

Benjamin Quost, Marie-Hélène Masson, Thierry Denoeux

Refined classifier combination using belief functions

Benjamin Quost
HeuDiaSyC laboratory
Compiègne University of Technology
Compiègne, France
Email: quostben@hds.utc.fr

Marie-Hélène Masson
HeuDiaSyC laboratory
Université de Picardie Jules Verne
Amiens, France
Email: massonmar@hds.utc.fr

Thierry Denœux
HeuDiaSyC laboratory
Compiègne University of Technology
Compiègne, France
Email: tdenoeux@hds.utc.fr

***Abstract*—We address here the problem of supervised classification using belief functions. In particular, we study the combination of non-independent sources of information. In a companion paper [1], we showed that the cautious rule of combination [2], [3] may be best suited than the widely used Dempster's Rule to combine classifiers in the case of real data. Then, we considered combination rules intermediate between the cautious rule and Dempster's rule. We proposed a method for choosing the combination rule that optimizes the classification accuracy over a set of data. Eventually, we mentioned a generalized approach, in which a refined combination rule best suited to complex dependencies of the sources to combine is learnt.**

Here, we extensively study this latter approach. It consists in clustering the sources according to some measure of similarity; then, one rule is learnt for combining the sources within the clusters, and another for combining the results thus obtained. We conduct experiments on various real data sets that show the interest of this approach.

Keywords: Classification, classifier combination, information fusion, theory of belief functions, Dempster-Shafer theory.

I. Introduction

The theory of belief functions [4], [5] has been accepted as a powerful tool for solving classification problems [6]–[8]. In this framework, experts express their belief on the value taken by an unknown variable using belief functions. Mathematical tools have been proposed for manipulating and aggregating such items. Dempster's rule of combination, also known as the conjunctive rule of combination (CoRC) [4], [5], [9] plays a central role in the theory of belief functions.

As pointed out in [3], this rule requires that the belief functions combined be distinct. Hence, a cautious rule of combination (CaRC) was proposed in [2], [3], allowing the combination of information coming from non distinct sources, by counting each elementary piece of information only once during the combination. It was also pointed out that both the CoRC and the CaRC may be seen as elements of infinite families of combination rules.

In a companion paper [1], we considered a supervised classification problem. Given a set of classifiers $\mathcal{C}_1, \dots \mathcal{C}_p$ giving partial information on the actual class of a test pattern $\mathbf{x}$, we proposed a method to learn the rule of combination that gives the best classification results over a set of patterns. We then proposed to generalize this method, by clustering and combining sources according to their dependency. We showed the interest of this approach on a series of synthetic toy problems.

In this paper, we focus on this latter approach. We consider a set of classifiers that provide information on a pattern to be classified. We propose a measure of similarity between these sources; using this measure, clusters of sources may be identified. Then, we learn two rules of combination: one is used to combine information within the clusters; the other, to combine together the results obtained in the previous step.

In Section II, we recall the notions on belief functions that will be needed, and fix the notations. In Section III, we detail our method for clustering the sources and learning adequate combination rules. Results on real data sets are reported and commented in Section IV. Section V concludes the paper.

II. Fundamentals of Belief Functions

In this article, we adopt the Transferable Belief Model (TBM) [4], [5], [10] as an interpretation of the theory of belief functions. The main notions are recalled in this section.

A. Basic Definitions

1) Representing Items of Evidence with Belief Functions: Let $\mathcal{C}$ be a classifier giving information on the actual class of a test pattern $\mathbf{x}$. This information may be represented by a basic belief assignment (bba) m, defined as a mapping from 2^Ω to $[0;1]$ satisfying $\sum_{A\subseteq\Omega} m(A) = 1$ (the notation 2^Ω denotes the powerset of Ω). A subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal set of m, and the basic belief mass (bbm) $m(A)$ quantifies the belief that the actual class of $\mathbf{x}$ is in A: this belief cannot be given to more precise hypotheses $B \subseteq A$ due to lack of information. The bbm $m(\emptyset)$ quantifies the belief that $\mathbf{x}$ belongs to none of the classes of the set Ω. A bba is said to be:

- dogmatic, if Ω is not a focal set;
- simple, if it has at most two focal sets, including Ω;
- categorical, if it is simple and dogmatic (therefore, if it has only one focal element that is not Ω);
- normal, if $\emptyset$ is not a focal set, and subnormal otherwise;
- consonant, if all its focal sets $A_1, \dots, A_N$ are nested: $\emptyset \subseteq A_1 \subseteq \dots \subseteq A_N \subseteq \Omega$.

Any subnormal bba m can be normalized; the normalization operation is defined by:

$$m^*(A) = \frac{m(A)}{1 - m(\emptyset)}, \quad \forall A \subseteq \Omega. \tag{1}$$

A bba m may also be represented by its associated plausibility, belief, commonality, or implicability functions, denoted by pl, bel, q, and b, respectively. Note that these functions are in one-to-one correspondence; they may be obtained from each other through linear transformations.

2) Conjunctive Combination and Decision Making: Two bbas m_1 and m_2, provided by distinct sources of information $\mathcal{C}_1$ and $\mathcal{C}_2$, may be combined using the conjunctive rule of combination (CoRC) $\textcircled{\cap}$ [9], also known as (the unnormalized) Dempster's rule:

$$m_1 \textcircled{\cap}_2(A) = \sum_{X \cap Y = A} m_1(X) m_2(Y), \ \forall A \subseteq \Omega. \tag{2}$$

The resulting bba $m_1 \textcircled{\cap}_2$ summarizes all the information provided by $\mathcal{C}_1$ and $\mathcal{C}_2$. Several other combination rules have been defined [11], [12]; however, although these rules may prove useful in a variety of practical applications, they have often been criticized for lacking theoretical justification.

When a decision needs to be taken, the bba m that quantifies knowledge of the actual class of $\mathbf{x}$ is transformed into a pignistic probability distribution [5]: m is first normalized, and then each bbm $m^*(A)$ is divided equally between the $\omega_k \in A$:

$$BetP(\omega_k) = \sum_{\omega_k \in A} \frac{m^*(A)}{|A|}, \ \forall \omega_k \in \Omega. \tag{3}$$

Let us define the operator Bet by: $BetP = \mathrm{Bet}(m)$. This operator is clearly nonlinear. It should also be remarked that a same $BetP$ generally corresponds to various bbas; we may then define:

$$\mathrm{Bet}^{-1}(BetP) = \{m : \mathrm{Bet}(m) = BetP\}.$$

In Section II-C, we will present a method for selecting a bba in the set $\mathrm{Bet}^{-1}(BetP)$, according to additional requirements.

B. Weights of Belief

Any non dogmatic bba may be represented by its *weight function (wf)* w [4], [13]. Transforming any representation of m into w is non-linear: for example, w may be computed from q by:

$$w(A) = \prod_{A \subseteq B} q(B)^{(-1)^{|B|-|A|+1}}. \tag{4}$$

The weights of belief satisfy $w(A) \geq 0$, for all $A \subset \Omega$. If $w(A) \leq 1$, $\forall A \subset \Omega$, the bba is said to be separable. The smaller is the weight $w(A) < 1$, the higher our confidence in A; weights $w(A) > 1$ may be interpreted as degrees of diffidence to A. In the case of consonant bbas, computing the wf becomes simpler [3]. Let the values $pl_k = pl(\{\omega_k\})$ be ordered by decreasing order: $1 \geq pl_1 \geq pl_2 \geq \cdots \geq pl_K > 0$. Then, using notations $A_k = \{\omega_1, \ldots, \omega_k\}$, we have:

$$w(A) = \begin{cases} pl_1 & A = \emptyset, \\ \dfrac{pl_{k+1}}{pl_k} & A = A_k, 1 \leq k < K \\ 1 & \text{otherwise}. \end{cases} \tag{5}$$

Only the subsets A_k ($1 \leq k \leq K$) and $\emptyset$ may be focal elements of the resulting weight function: thus, it is consonant.

The $\textcircled{\cap}$-combination of two weight functions w_1 and w_2 may be processed by:

$$w_1 \textcircled{\cap}_2(A) = w_1(A) w_2(A), \quad \forall A \subset \Omega. \tag{6}$$

It becomes here obvious that the CoRC is commutative and associative. However, it is not idempotent: in particular, combining a separable wf with itself results in decreasing all the weights $w(A) \neq 1$. More generally, $\textcircled{\cap}$-combining the outputs of two non-independent classifiers generally results in counting several times the identical items of evidence.

1) Partial Orderings on Bbas: The informational content of two bodies of evidence may be partially ordered by comparing their corresponding wfs [3]. Let m_1 and m_2 be two non-dogmatic bbas; m_1 is w-more committed than m_2, which we write $m_1 \sqsubseteq_w m_2$, iff:

$$w_1(A) \leq w_2(A), \text{for all } A \subset \Omega.$$

This property is satisfied iff there exists a separable bba m_3 such that $m_1 = m_2 \textcircled{\cap} m_3$ [3].

The q-ordering [14] is obtained by replacing w with q: m_1 is q-more committed than m_2 ($m_1 \sqsubseteq_q m_2$) iff $q_1(A) \leq q_2(A)$, for all $A \subset \Omega$. It is weaker than the w-ordering, as we have $m_1 \sqsubseteq_w m_2 \Rightarrow m_1 \sqsubseteq_q m_2$ while the converse is generally not true.

2) The Cautious Rule of Combination: Two bodies of evidence may proceed from common information; thus, a cautious approach to combining them should count each item only once [3], [15], [16]. In the most extreme case where the two bodies are identical, the result should be the body itself — equivalently, the combination rule should be idempotent.

The cautious rule of combination (CaRC) $\textcircled{\wedge}$ consists in applying the $\min$ to the wfs, instead of the product [3]:

$$w_1 \textcircled{\wedge}_2(A) = w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega, \tag{7}$$

where $a \wedge b$ stands for $\min(a, b)$, and where $w_1 \textcircled{\wedge}_2 = w_1 \textcircled{\wedge} w_2$. Thus, it is straightforward that $w_1 \textcircled{\wedge}_2$ is the least w-committed of the weight functions that are more w-committed than both w_1 and w_2.

The CaRC is associative, commutative, and idempotent, as is the $\min$ operator. Whereas $w_1 \textcircled{\cap}_2(A)$ depends on both $w_1(A)$ and $w_2(A)$, the CaRC retains only the smallest weight to compute $w_1 \textcircled{\wedge}_2(A)$ (that is, for separable bbas, the strongest support to A).

C. Least q-committed Bba Induced by a Probability Distribution

The q-ordering may be used to reverse the pignistic transform. To avoid giving unjustified support to any $A \subseteq \Omega$, we may select $\widehat{m} = \mathrm{Bet}^{-1}_{qlc}(BetP)$, the least q-informative bba $\widehat{m}$ in $\mathrm{Bet}^{-1}(BetP)$:

$$\begin{cases} \widehat{m} \in \mathrm{Bet}^{-1}(BetP), \\ m \sqsubseteq_q \widehat{m}, \ \text{for all } m \in \mathrm{Bet}^{-1}(BetP). \end{cases}$$

In [17], it was shown that $\widehat{m}$ is a consonant bba. It may be obtained by first computing $pl(\{\omega_i\})$ for all $1 \leq i \leq K$ and

then deducing $pl(A)$, for all $A \subseteq \Omega$ with $|A| > 1$:

$$pl(\{\omega_i\}) = \sum_{j=1}^{K} \min(p_i, p_j), \quad (8)$$

$$pl(A) = \max_{\omega_k \in A} pl(\{\omega_k\}). \quad (9)$$

Remark that using (5), the wf may be computed directly with (8): we need not use (9).

This transformation will help us build bbas from probabilistic classifiers. Indeed, interpreting the output of such a classifier as a pignistic probability distribution over Ω, we may then define $\widehat{m} = \mathrm{Bet}_{qlc}^{-1}(BetP)$ as its credal output.

III. Learning the Combination Rule

Both the CoRC and the CaRC may be seen as "extremal" rules of combination: the former combines distinct evidence, the latter evidence that could have been induced by the same information. We stress here that information may be non-distinct without being exactly the same. Thus, one could select a combination rule that best suits a set of data.

Both the product and the min operator are *triangular norms* (t-norms) on $[0; 1]$ [18]. Let us consider a parameterized family of t-norms counting both these operators as members [3]. Thus, in the case of separable bbas, selecting a t-norm by picking a parameter value defines implicitly a combination rule that is intermediate between the CoRC and the CaRC. In this article, we consider Frank's family:

$$x \top_s y = \log_s \left(1 + \frac{(s^x - 1)(s^y - 1)}{s - 1}\right), \quad (10)$$

where $\log_s$ defines the logarithm function with base s. Here, parameter s defines the t-norm: the $\min$ operator is retrieved as $s \to 0$, and the product as $s = 1$.

Remark that the parameter s does not represent a degree of dependency between the information to be merged. Thus, fitting the combination rule to the sources necessits first to estimate their degree of dependency, and then to find a corresponding parameter value. The relationship between both may be complex, and obviously depends on the t-norm used. We stress that the notion of dependency depends on the information considered. It is well known that estimating the degree of dependency between variables may be delicate: one should first identify its nature, and then evaluate the degree using a proper measure. For example, the correlation coefficient is inefficient for modelling nonlinear dependency.

In this paper, we consider supervised classification problems. Therefore, rather than considering the inputs of the classifiers (the variables), we propose to fit parameter values by maximizing the classification accuracy achieved after combination of their outputs. Thus, our method allows to find the combination rule that will give the best classification results, even if the dependency between the information sources is arbitrarily complex.

A. Fitting a Single Combination Rule to Training Data

In a companion paper [1], we proposed a simple procedure for learning a combination rule that gives the best classification results over a set of data. It consists in parameterizing the combination rule, and then mining the parameter space for a value that minimizes some error criterion. We briefly describe this procedure in the following.

Given a set of training patterns, we train a classifier on each of the q variables. For each pattern $\mathbf{x}$ being classified, we thus have q bbas $m_i\{\mathbf{x}\}$. Combining the $m_i\{\mathbf{x}\}$ for any parameter value s gives a bba $m\{\mathbf{x}\}$, from which the pignistic probability distribution $BetP\{\mathbf{x}\}$ used to classify $\mathbf{x}$ is then obtained.

Given a value of s, we compute the average squared difference between the pignistic probabilities and binary indicator variables encoding class membership:

$$\mathcal{E} = \sum_{j=1}^{n} \|\mathbf{betp}\{\mathbf{x}_j\} - \delta\{\mathbf{x}_j\}\|^2; \quad (11)$$

here, for any pattern $\mathbf{x}_j$ from a validation set, pignistic probabilities $\mathbf{betp}\{\mathbf{x}_j\} = (BetP\{\mathbf{x}_j\}(\omega_1), \ldots, BetP\{\mathbf{x}_j\}(\omega_K))$ are obtained from $m\{\mathbf{x}_j\}$, and the crisp memberships of $\mathbf{x}_j$ to the classes are given by $\delta\{\mathbf{x}_j\} = (\delta_1\{\mathbf{x}_j\}, \ldots, \delta_K\{\mathbf{x}_j\})$ (we have $\delta_k\{\mathbf{x}_j\} = 1$ if $\mathbf{x}_j \in \omega_k$, and 0 otherwise). We thus select the value $\widehat{s}$ that minimizes $\mathcal{E}$ using a dichotomic search algorithm, stopping the search when the width of the interval to search is less than some constant (here, 10^{-10}).

B. Refined Combination of the Sources

The method presented in Section III-A consists in learning a single rule to combine classifiers with the best classification accuracy. It relies on the implicit assumption that all the sources share the same pairwise dependency. However, this assumption may be too simplistic. Indeed, some source $\mathcal{C}_1$ may have a tendency to be highly redundant with some other source $\mathcal{C}_2$, yet being uncorrelated with another source $\mathcal{C}_3$. Choosing a single rule close to the CoRC could imply giving too much weight to the outputs of $\mathcal{C}_2$, while being too close to the CaRC could lead to ignore important information provided by $\mathcal{C}_3$.

To evaluate the dependency between two sources, we propose to compute a measure of discrepancy between their outputs. A distance d_J between two (normal) bbas m_1 and m_2 was defined in [19], by:

$$d_J(m_1, m_2) = \sqrt{\sum_{\substack{A \subseteq \Omega \\ B \subseteq \Omega}} \frac{(m_1(A) - m_2(A))(m_1(B) - m_2(B))}{2/p_{A,B}}}, \quad (12)$$

where the weights $p_{A,B}$ are defined by:

$$p_{A,B} = \frac{|A \cap B|}{|A \cup B|}, \ \forall A \neq \emptyset, B \neq \emptyset. \quad (13)$$

We propose to define the distance $\mathcal{D}_{i,j}$ between two classifiers $\mathcal{C}_i$ and $\mathcal{C}_j$ as the average distance between the (paired) bbas they provided when evaluating a set of patterns:

$$\mathcal{D}_{i,j} = \sum_{k=1}^{n} d_J(m_i\{\mathbf{x}_k\}, m_j\{\mathbf{x}_k\}). \quad (14)$$

Thus, $\mathcal{D}_{i,j}$ may be used as an indicator of the redundancy between the outputs of $\mathcal{C}_i$ and $\mathcal{C}_j$. Remark that although d_J appears to be perfectly suited to our purpose — as it was defined for evaluating the performance of a classification algorithm, other distances between two bbas may be used. We did not evaluate our method using such other measures.

Let $\mathcal{C}_1, \ldots, \mathcal{C}_q$ be our set of classifiers. For each classifier, we evaluate all training patterns, and we compute the distance $\mathcal{D}_{i,j}$ for each pair $(\mathcal{C}_i, \mathcal{C}_j)$. A hierarchy on the classifiers may be built upon these distances. Figure 1 shows the dendrogram representing this hierarchy, for the ecoli dataset (Section IV-B2 details how we generated this dendrogram). We then choose

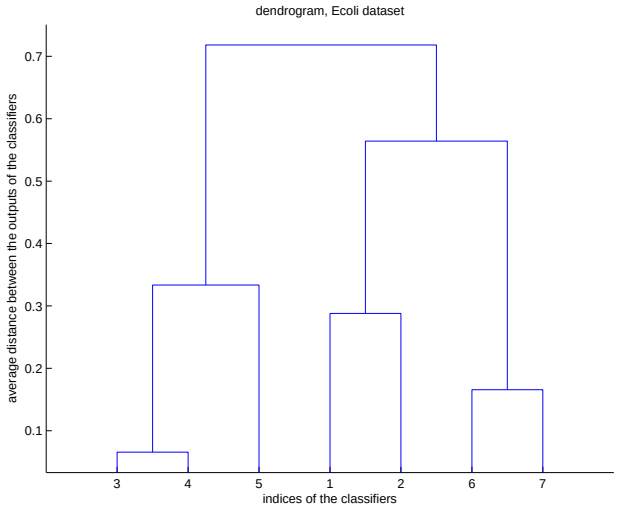

Figure 1. Dendrogram: distances between classifiers, ecoli data set.

a cut value to cluster the set of classifiers. For instance, in Figure 1, cutting at a level 0.4 would give 3 clusters: $\{\mathcal{C}_1, \mathcal{C}_2\}$, $\{\mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}$, and $\{\mathcal{C}_6, \mathcal{C}_7\}$. Remark that this clustering step may be automated.

After the classifiers have been clustered according to these distances, we propose to first combine their outputs in the various clusters, and then to combine the resulting bbas together. Thus, two combination rules have now to be learnt: a within-cluster rule for processing the combination within the clusters, defined by a parameter value s_1; and a between-cluster rule for computing the final bba, defined by a parameter value s_2. We propose to use a method similar to that described in Section III-A; except that we now select the pair of values $(\widehat{s_1}, \widehat{s_2})$ that minimizes $\mathcal{E}$. Taking the ecoli data as an example, the outputs of $\mathcal{C}_1$ and $\mathcal{C}_2$ would first be combined using $\widehat{s_1}$, as well as those of $\mathcal{C}_3$, $\mathcal{C}_4$, and $\mathcal{C}_5$, and those of $\mathcal{C}_6$ and $\mathcal{C}_7$; the three resulting bbas would then be combined using $\widehat{s_2}$.

Our method necessits to combine the bbas for various parameter values; hence, the overall complexity depends pricipally on the size of the bbas, which is generally exponential in the number K of classes. However, the bbas considered here being consonant, they have $K+1$ focal elements. Although this is not the case after the combination step, the complexity may be limitated by considering only the subsets of Ω with nonzero belief mass. Methods for reducing the combination complexity may be found in [20] and in the references therein.

IV. Results

A. Description of the Data

We ran experiments on various real datasets, whose characteristics (number of classes, of features, and the numbers of patterns in the training and test sets) are presented in Table I. These datasets may be found in the UCI Machine Learning repository at http://archive.ics.uci.edu/ml/. We simplified the letter, optdigits, and pendigits datasets by selecting six classes in the former case (the 4th, 5th, 12th, 21st, 22nd and 24th ones), and five classes in the two latter (the 1st, 2nd, 5th, 6th and 10th).

Table I
Description of the datasets employed in the experiments.

dataset	# classes	# features	number of patterns	
			training	test
ecoli	8	7	201	135
glass	6	9	139	75
letter	6	16	1800	2400
optdigits	5	64	1910	903
pageblocks	5	10	3284	2189
pendigits	5	16	3778	1762
waveform	3	21	1491	3509

B. Learning the Classifiers and the Refined Combination Rules

1) Training the Classifiers: For each dataset, we trained a classifier (logistic regression) on each variable. For any test point $\mathbf{x}$, we are thus able to provide q probability distributions p_i on Ω. Then, we computed the q-least committed bbas m_i whose pignistic probabilities are p_i, using (5) and (8). These bbas were combined using the CoRC, the CaRC, an intermediate rule learnt as defined in Section III-A, and a refined combination involving two rules such as described in Section III-B.

2) Clustering the Sources: The dendrograms representing the hierarchies between classifiers were built upon the pairwise average distances $\mathcal{D}_{i,j}$ using Ward's aggregation criterion. They are reported in Figures 1 to 7. The set of classifiers was then clustered: Table II provides the cut levels that were chosen by the user, in order to form between 3 and 5 clusters.

Basically, the dendrograms provide information about the relative pairwise dependencies of the classifiers. Taking the pageblocks dataset as an example, it may be remarked that all the classifiers except $\mathcal{E}_4$ and $\mathcal{E}_5$ have a tendency to provide bbas that are very close to each other. On the other hand, the dependencies between the classifiers of the letter or optdigits datasets are quite varied.

Note that interpreting the average distance between two classifiers remains delicate: indeed, the bbas they provide obviously depend on the classification algorithm employed. Moreover, d_J may suffer (just like any other distance) from the curse of dimensionality, as the size of the frame Ω grows.

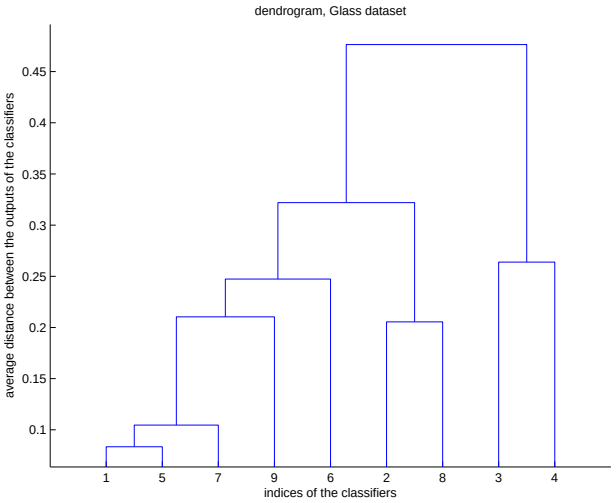

Figure 2. Dendrogram: distances between classifiers, glass data set.

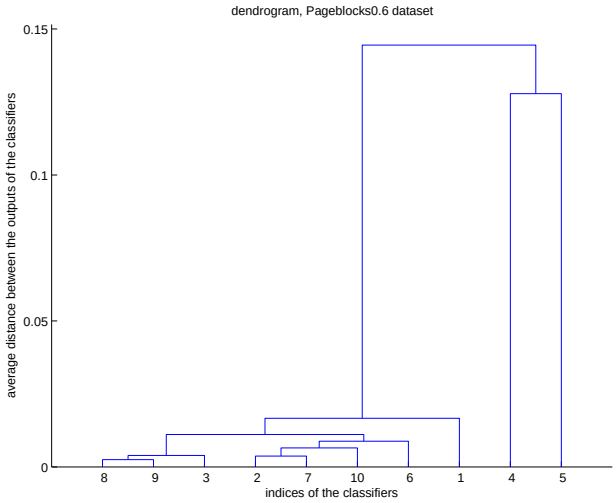

Figure 5. Dendrogram: distances between classifiers, pageblocks data set.

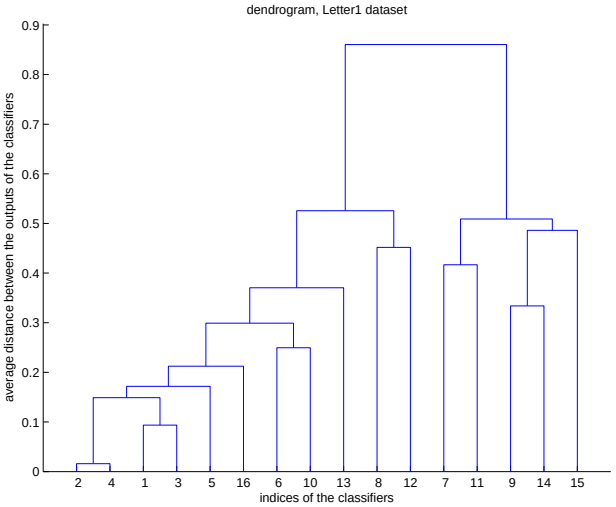

Figure 3. Dendrogram: distances between classifiers, letter data set.

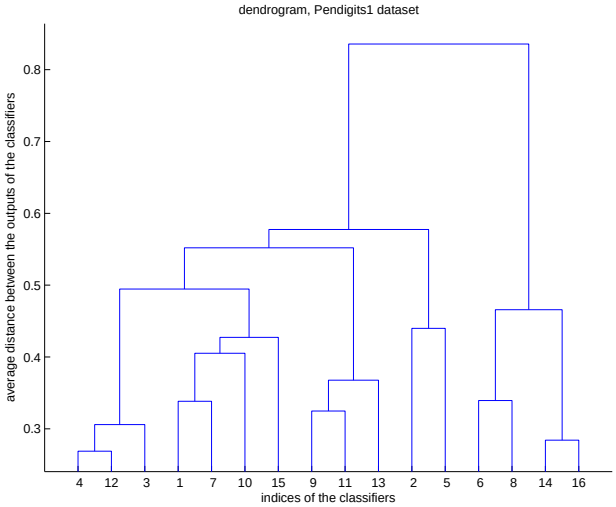

Figure 6. Dendrogram: distances between classifiers, pendigits data set.

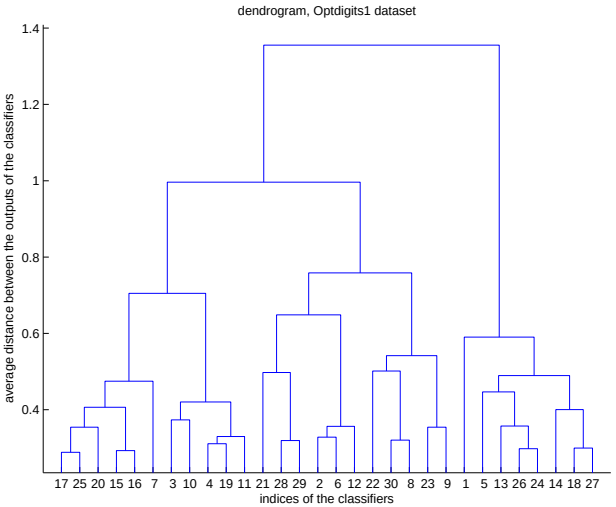

Figure 4. Dendrogram: distances between classifiers, optdigits data set.

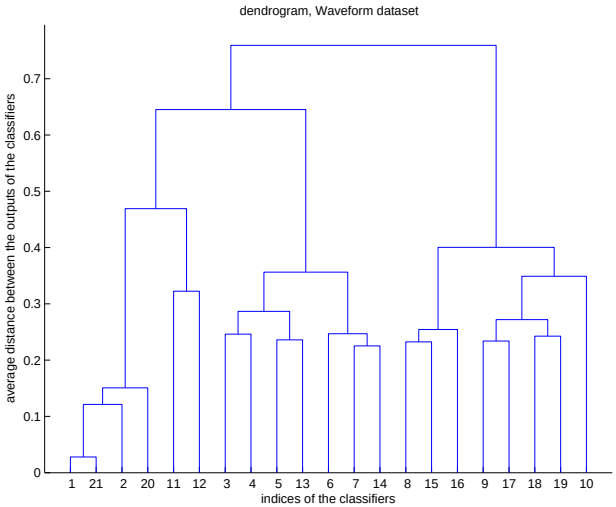

Figure 7. Dendrogram: distances between classifiers, waveform data set.

Table II
CUT LEVELS USED; NUMBER OF CLUSTERS THUS OBTAINED; TOTAL NUMBER OF SOURCES.

dataset	cut level	# clusters	# sources
ecoli	0.4	3	7
glass	0.3	3	9
letter	0.5	4	16
optdigits	0.7	5	64
pageblocks	0.05	3	10
pendigits	0.5	4	16
waveform	0.45	4	21

3) Learning the Within-Cluster and Between-Cluster Rules: The values $\widehat{s_1}$ and $\widehat{s_2}$ were then determined as follows. For each dataset, 5×1 cross-validation was used to form training/validation sets from the original training set. Candidate values were picked for s_1 and s_2; for each pair (s_1, s_2), the error criterion defined by Eq (11) was computed over the 5 validation sets and averaged. The values $\widehat{s_1}$ and $\widehat{s_2}$ minimizing this average were chosen. They are reported in Table III.

C. Numerical Results

Error rates obtained with the CoRC, the CaRC, the intermediate combination rule (InRC) described in Section III-A corresponding to the optimal parameter value $\widehat{s}$, and the refined combination strategy (ReRC) obtained with values $\widehat{s_1}$ and $\widehat{s_2}$, are provided in Table III, together with 95% confidence intervals. The best result is underlined, and printed in bold as well as results that were not judged significantly different by a McNemar test [21] at level 5%.

Table III
ERROR RATES OF THE CONJUNCTIVE (CORC), INTERMEDIATE (INRC), REFINED (RERC), AND CAUTIOUS (CARC) RULES OF COMBINATION, TOGETHER WITH 95% CONFIDENCE INTERVALS. BEST RESULTS ARE UNDERLINED; RESULTS THAT ARE NOT SIGNIFICANTLY DIFFERENT ARE PRINTED IN BOLD.

data	CoRC	InRC ($\widehat{s}$)	ReRC ($\widehat{s_1}, \widehat{s_2}$)	CaRC
ecoli	44.44 [36.06;52.83]	(0) **37.04** [28.89;45.18]	(0,1e-15) **37.04** [28.89;45.18]	**37.04** [28.89;45.18]
glass	49.33 [38.02;60.65]	(0) **45.33** [34.07;56.60]	(0,1e-15) **44.00** [32.77;55.23]	**45.33** [34.07;56.60]
letter	**16.54** [15.06;18.03]	(6e-1) **16.46** [14.97;17.94]	(1e-5,1e-2) **15.92** [14.45;17.38]	16.83 [15.34;18.33]
optdg.	**5.43** [3.95;6.90]	(6.25e-2) **5.09** [3.66;6.53]	(1e-2,0) **4.87** [3.47;6.28]	**4.98** [3.56;6.40]
pageb.	10.23 [8.96;11.50]	(0) **8.54** [7.37;9.71]	(0,0) **8.54** [7.37;9.71]	**8.54** [7.37;9.71]
pendg.	**18.73** [16.91;20.55]	(0) 20.49 [18.60;22.37]	(1,0) **18.79** [16.96;20.61]	20.49 [18.60;22.37]
wavef.	16.93 [15.69;18.17]	(2.5e-1) 16.13 [14.91;17.35]	(1e-3,1e-2) **15.25** [14.06;16.44]	16.53 [15.30;17.76]

Analyzing the results presented in Table III yields the following remarks. The interest of learning the rule of combination is clearly assessed by the good results given by the refined combination strategy: they are generally the best over the four combination rules evaluated (the only exception being the pendigits dataset). This suggests that the dependency between the combined classifiers may be complex enough to justify learning a complex, refined combination strategy, involving determining a within-cluster and a between-cluster rule instead of a single one.

The CoRC gives the best classification results for the pendigits dataset only, and it does not perform significantly worse that the best rule for the letter and optdigits datasets. Let us further remark that learning the combination rule (be it a single rule or a pair of within-cluster and between-cluster rules) never yielded the CoRC. This should draw attention to the fact that assuming independence between classifiers may be unreasonable. Henceforth, other fusion strategies may be preferable to Dempster's rule when combining information coming from various sources.

In addition, we may observe that the combination rules learnt for the ecoli, glass, and pageblocks datasets are very close to the cautious rule. Quite surprisingly, in the case of the pendigits and optdigits datasets, the within-cluster combination rule is closer to the CoRC than the between-cluster combination rule (in the former case, the within-cluster rule is the CoRC, and the between-cluster rule is the CaRC). This phenomenon is not yet entirely clear; a possible explanation could be that information combined within the clusters are rich and diverse, and hence the information resulting of the various within-cluster combinations are close to each other.

V. CONCLUSIONS

In this paper, we addressed the problem of supervised classification by classifier fusion, in the framework of belief functions theory. We presented a method for adapting a combination rule to a set of data. First, the discrepancy between each pair of classifiers was measured by computing the average distance between their outputs. Then, classifiers are clustered based on these average pairwise distances. The classifiers within the various clusters are combined using a within-cluster rule, and within-cluster combined outputs are then pooled using a between-cluster rule. Both rules are learnt by minimizing an error criterion over validation data.

We evaluated four combination rules on seven real data sets: the conjunctive rule of combination (also known as the unnormalized Dempster's rule), the cautious rule of combination, a single rule learnt to fit the data processed, and the refined combination scheme presented in this paper. Numerical results obtained show the interest of learning a complex combination strategy adapted to the dependency of the data.

Future work may focus on three points. First of all, the impact of the clustering phase on the classification results may be studied, and the clustering step may be automated. Learning the coefficients defining the combination rules requires searching the parameter space. Although this learning phase needs to be done only once, a fast procedure for determining accurately these values would be a plus. Eventually, one could also imagine refining further the combination strategy, by

proposing a hierarchy between the classifiers: at each level of the hierarchy, classifiers could then be combined using an adequate combination rule. Such a combination strategy should allow fitting even more precisely the dependency of the classifiers.

REFERENCES

[1] B. Quost, T. Denœux and M.-H. Masson. Adapting a Combination Rule to Non-independent Information Sources. submitted to *12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'08)*, 2008.

[2] T. Denœux (2006). The cautious rule of combination for belief functions and some extensions. In *Proceedings of the Ninth International Conference on Information Fusion (FUSION 2006)*, Florence, Italy, July 2006.

[3] T. Denœux (2008). Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. *Artificial Intelligence*, vol. 172, no. 2-3, pp. 234-264, 2008.

[4] G. Shafer (1976). A mathematical theory of evidence. Princeton University Press, Princeton, NJ, 1976.

[5] P. Smets and R. Kennes (1994). The Transferable Belief Model. *Artificial Intelligence*, vol. 66, no. 2, pp. 191-234, 1994.

[6] S. Fabre, A. Appriou and X. Briottet (2001). Presentation and description of two classification methods using data fusion based on sensor management. *Information Fusion*, vol. 2, no. 1, pp. 49-71, 2001.

[7] I. Bloch (2008). Information Fusion in Signal and Image Processing. Major Probabilistic and Non-Probabilistic Numerical Approaches (Digital signal and image processing series). Wiley, 2008.

[8] B. Quost, T. Denœux and M.-H. Masson (2007). Pairwise Classifier Combination using Belief Functions. *Pattern Recognition Letters*, vol. 28, no. 5, pp. 644-653, 2007.

[9] P. Smets (1990). The combination of evidence in the transferable belief model. *IEEE Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447-458, 1990.

[10] P. Smets (1993). Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, vol. 9, no. 1, pp. 1-35, 1993.

[11] D. Dubois and H. Prade (1986). On the unicity of Dempster's rule of combination. *International Journal of Intelligent Systems*, vol. 1, no. 2, pp. 133-142, 1986.

[12] R. Yager (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, vol. 41, no. 2, pp. 93-137, 1987.

[13] P. Smets (1995). The canonical decomposition of a weighted belief. In *Proceedings of the 14th International Joint Conferences in Artificial Intelligence (IJCAI'95)*, pp. 1896-1901, Montréal, Canada, 1995.

[14] D. Dubois and H. Prade (1987). The principle of minimum specificity as a basis for evidential reasoning. In *Uncertainty in Knowledge-based Systems*, Springer Verlag, Berlin, pp. 75-84, 1987.

[15] P. Smets (1986). Combining non-distinct evidences. In *Int. Conf. of the North American Fuzzy Information Processing Society (NAFIPS'86)*, pp. 544-549, New Orleans, USA, 1986.

[16] P. Smets (2000). Data fusion in the Transferable Belief Model. In *Proceedings of the Third International Conference on Information Fusion (FUSION 2000)*, PS 20/33, Paris, France, 2000.

[17] D. Dubois, H. Prade and P. Smets (2008). A definition of subjective possibility. *International Journal of Approximate Reasoning*, in press, 10.1016/j.ijar.2007.01.005.

[18] E. Klement, R. Mesiar and E. Pap (2000). Triangular norms. Kluwer Academic Publishers, Dordrecht, 2000.

[19] A.-L. Jousselme, D. Grenier and É. Bossé (2001). A new distance between two bodies of evidence. *Information Fusion*, vol. 2, no. 2, pp. 91-101, 2001.

[20] T. Denoeux and A. Ben Yaghlane. Approximating the Combination of Belief Functions using the Fast Moebius Transform in a coarsened frame. *International Journal of Approximate Reasoning*, vol. 31, no. 1-2, pp. 77-101, 2002.

[21] T. Dietterich (1998). Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation*, vol. 10, no. 7, pp. 1895-1923, 1998.

[22] T. Hastie, R. Tibshirani and J. Friedman (2001). The elements of statistical learning: data mining, inference and prediction. Springer Verlag, New York, 2001.