Adapting a Combination Rule to Non-Independent Information Sources

Benjamin Quost[1]
benjamin.quost@hds.utc.fr

Thierry Denœux[1]
thierry.denoeux@hds.utc.fr

Marie-Hélène Masson[1,2]
mylene.masson@hds.utc.fr

[1] UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529, F-60205 Compiègne cedex - France

[2] Université de Picardie Jules Verne
Chemin du Thil
80025 Amiens

Abstract

In this article, we address the combination of non-independent sources to solve classification problems, within the theory of belief functions. We show that the cautious rule of combination [1, 2] is well-suited to such problems. We propose a method to learn the combination rule from training data, and we generalize it in the case of complex dependence of the sources. We demonstrate the validity of our approach on several synthetic and real-data sets.

Keywords: Classification, classifier combination, sensor fusion; theory of belief functions.

1 Introduction

The theory of belief functions [11, 15] is a powerful tool for modelling and manipulating knowledge. In this framework, beliefs held by experts are quantified by belief functions. Various mathematical tools have been proposed for manipulating such items; in particular, the conjunctive rule of combination (CoRC) [15, 13] plays a central role in the theory of belief functions.

As pointed out in [2], a major limitation of this rule comes from the requirement that the items of evidence combined be distinct. A cautious rule of combination (CaRC) was recently proposed [1, 2], allowing the combination of information coming from non distinct sources. In a nutshell, it consists in 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 this paper, we address a supervised classification problem, in which several classifiers $\mathcal{C}_1, \dots \mathcal{C}_p$ provide partial information on the actual class of a test pattern $\mathbf{x}$. To classify $\mathbf{x}$, we need to combine the information given by each of those classifiers. We propose a method to learn the rule of combination that will give the best classification results over a set of patterns. We then propose to generalize this method, by clustering and combining sources according to their dependency. We demonstrate the interest of our approach by appling it to several synthetic and real datasets.

In Section 2, we recall basic knowledge of belief functions and fix notations. In Section 3, we propose a simple procedure to adapt the combination rule to a classification problem, and results are reported in Section 4. The method is generalized in Section 5, where numerical results are also presented. Section 6 concludes the paper.

2 Fundamentals of Belief Functions

Our approach is based on the Transferable Belief Model (TBM) [11, 15, 14], the main notions of which are recalled in this section.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 448–455
Torremolinos (Málaga), June 22–27, 2008

2.1 Basic Definitions

2.1.1 Representing Items of Evidence with Belief Functions

Let $\mathcal{C}$ be a classifier that provides information on the actual class of a test pattern $\mathbf{x}$. We suppose here that this information may be quantified by a basic belief assignment (bba) m, defined as a mapping from 2^Ω to $[0;1]$ which satisfies $\sum_{A\subseteq\Omega} m(A) = 1$ (here, 2^Ω denotes the powerset of Ω). A subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal set of m. The empty set $\emptyset$ may be a focal set: $m(\emptyset)$ 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 ;
- normal, if $\emptyset$ is not a focal set, and subnormal otherwise;
- consonant, if all its focal sets $A_1, \ldots, A_N$ are nested: $\emptyset \subseteq A_1 \subseteq \cdots \subseteq A_N \subseteq \Omega$.

Note that any subnormal bba m can be normalized; the resulting bba is defined by:

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

We may represent m by its associated plausibility, belief, commonality, or implicability function, denoted by pl, bel, q, and b, respectively. All are in one-to-one correspondence, and may be obtained from each other through linear transformations. For instance, we have:

$$m(A) = \sum_{A\subseteq B} (-1)^{|B|-|A|} q(B). \qquad (2)$$

2.1.2 Conjunctive Combination and Decision Making

Two bbas m_1 and m_2, provided by independent classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, may be combined using the conjunctive rule of combination (CoRC) $\textcircled{\cap}$ [13] : for all $A \subseteq \Omega$,

$$m_{1\textcircled{\cap}2}(A) = \sum_{X\cap Y=A} m_1(X)m_2(Y). \qquad (3)$$

The resulting bba $m_{1\textcircled{\cap}2}$ embeds all the information provided by $\mathcal{C}_1$ and $\mathcal{C}_2$. Other combination rules have been defined [4, 18]; however, they are often criticized for lacking theoretical justification, though they may prove useful in a variety of practical applications.

Once a decision has to be made, a bba m is transformed into a pignistic probability distribution [15]. The pignistic transform consists in normalizing m, and then dividing each mass $m^*(A)$ equally between the $\omega_k \in A$:

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

We may then write $BetP = \mathrm{Bet}(m)$. This transform 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 2.2.2, we will present a method for selecting a bba in $\mathrm{Bet}^{-1}(BetP)$, according to additional requirements.

2.2 Weights of Belief

Any non dogmatic bba may be represented by its *weight function (wf)* w [11, 16]; for example, w may be computed from q by:

$$w(A) = \prod_{A\subseteq B} q(B)^{(-1)^{|B|-|A|+1}}. \qquad (5)$$

The weights of belief satisfy $w(A) \geq 0$, for all $A \subset \Omega$. If $w(A) \leq 1$, $\forall A \subset \Omega$, the wf 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 [2]. Let the $pl(\{\omega_k\})$ be ordered by decreasing order: $1 \geq pl_1 \geq pl_2 \geq \cdots \geq pl_K > 0$. Then:

$$w(A) = \begin{cases} pl_1 & A = \emptyset, \\ \dfrac{pl_{k+1}}{pl_k} & A = A_k, 1 \leq k < K \\ 1 & \text{otherwise} . \end{cases} \qquad (6)$$

The nested focal sets of the resulting bba are thus $A_k = \{\omega_1, \ldots, \omega_k\}$, for $1 \leq k \leq K$, or $\emptyset$.

Let w_1 and w_2 be two wfs, and $w_{1 \odot 2}$ denote the result of their $\odot$-combination; then:

$$w_{1 \odot 2}(A) = w_1(A) w_2(A), \quad \forall A \subset \Omega. \quad (7)$$

The CoRC is commutative and associative. However, it is not idempotent: combining the wf corresponding to a separable bba with itself results in decreasing all the weights $w(A) \neq 1$. More generally, $\odot$-combining the outputs of two non-independent classifiers generally results in counting several times same items of evidence.

2.2.1 Partial Orderings on Bbas

A partial ordering on the informational content of two non-dogmatic bbas m_1 and m_2 may be built, by comparing their corresponding wfs [2]. The bba 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 $m_1 = m_2 \odot m_3$ [2].

The q-ordering [5] 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 $A \subset \Omega$. This latter ordering is weaker than the former: indeed, we have $m_1 \sqsubseteq_w m_2 \Rightarrow m_1 \sqsubseteq_q m_2$, while the converse is generally not true.

2.2.2 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 $A \subseteq \Omega$, it was proposed in [6] to select $\widehat{m} = \mathrm{Bet}_{qlc}^{-1}(BetP)$, 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 [7], it was shown that $\widehat{m}$ is a consonant bba, that 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 (6), the wf may be computed directly with (8): we need not use (9).

2.3 The Cautious Rule of Combination

A cautious approach in combining two bodies of evidence would consist in counting each item only once [12, 17, 2], considering that they may have been built on common information. In the most extreme case where the two bodies are identical, the result should be the body itself — hence, we seek an idempotent rule. As bringing new evidence aims at having more precise knowledge of the actual class of pattern $\mathbf{x}$, the result should not be less committed than the inputs.

The cautious rule of combination (CaRC) $\owedge$ consists in applying the min operator, instead of the product, to the wfs [2]:

$$w_{1 \owedge 2}(A) = w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega, \ (10)$$

where $a \wedge b$ stands for $\min(a, b)$, and where $w_{1 \owedge 2} = w_1 \owedge w_2$. The CaRC is associative, commutative, and idempotent, as is the min operator. Provided that m_1 and m_2 be separable, $w_{1 \owedge 2}$ is obviously at least as w-committed as both w_1 and w_2.

Thus, whereas $w_{1 \odot 2}(A)$ depends on both $w_1(A)$ and $w_2(A)$, the CaRC retains only the smallest to compute $w_{1 \owedge 2}(A)$ (that is, for separable bbas, the strongest support to A).

3 Finding a Compromise Between the CoRC and the CaRC

3.1 Motivations

Both the CoRC and the CaRC may be seen as "extremal" rules of combination: the former combines independent bodies of evidence, the latter, wfs that could have been induced by identical information. However, information may be non-distinct, without being identical: for example, two sensors may observe different but correlated variables.

Therefore, an intermediate between the CoRC and the CaRC may be best suited to combine such information. If we restrict ourselves to the combination of separable bbas [2], they may be defined as follows. Both the product and the min operators being *triangular norms* (t-norms) on $[0; 1]$ [10], one may consider a parameterized family of t-norms, counting both the product and the minimum as members [2]. Selecting a t-norm, by picking a parameter value, would implicitly define a combination rule, intermediate between the CoRC and the CaRC in the case of separable bbas. One may consider, for instance, Frank's family:

$$x \top_s y = \log_s \left(1 + \frac{(s^x - 1)(s^y - 1)}{s - 1} \right), \quad (11)$$

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 $s = 1$.

3.2 Preliminary Results

We created several two-class datasets, by generating numbers from a 10-dimensional multivariate Gaussian distribution, as follows:

- each marginal distribution has mean $\mu_1 = 0$ in class ω_1, $\mu_2 = 1$ in class ω_2, and variance $\sigma_1^2 = \sigma_2^2 = 1$ in both classes;
- in each class, the first 9 variables are pairwise linearly correlated with coefficient ρ, the last one is independent of the others.

We trained a classifier (logistic regression [8]) on each variable. For each test point $\mathbf{x}$, we are thus able to provide $p = 10$ probability distributions p_i on Ω. Then, we computed the q-least committed bbas m_i whose pignistic probabilities are p_i, using (6) and (8). We combine the m_i, and eventually we classify $\mathbf{x}$.

For various values of s, we computed the average misclassification rate on ten test sets, using the corresponding rule of combination. Figures 1, 2 and 3 show the misclassification rate as a function of s, together with 95% confidence intervals, for datasets generated using correlation coefficients $\rho = 0.1$, $\rho = 0.3$ and $\rho = 0.9$. In these figures, the CoRC and the CaCR correspond to the rightmost and the leftmost points of the x axis, respectively.

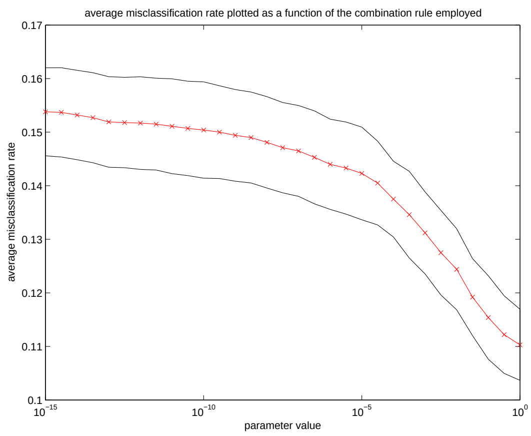

Figure 1: Misclassification rate plotted as a function of s; correlation coefficient $\rho = 0.1$.

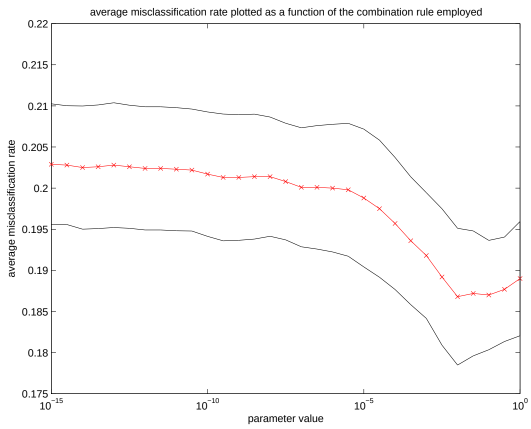

Figure 2: Misclassification rate plotted as a function of s; correlation coefficient $\rho = 0.3$.

For $\rho = 0.1$ and $\rho = 0.9$, the best results are obtained, respectively, using the CoRC and the CaRC. The former case shows that the CaRC is sensible to errors: it only considers the strongest support to each focal set, and thus may lead to a wrong decision if this support is erroneous. When $\rho = 0.9$, the first 9 classifiers have a tendency to provide erroneous evidence simultaneously; thus, using the CaRC enables the 10^{th} classifier to correct such errors, as its output weights as much as all the others. When $\rho = 0.3$, the best results are obtained for a value of s that corresponds to an intermediate rule of combination.

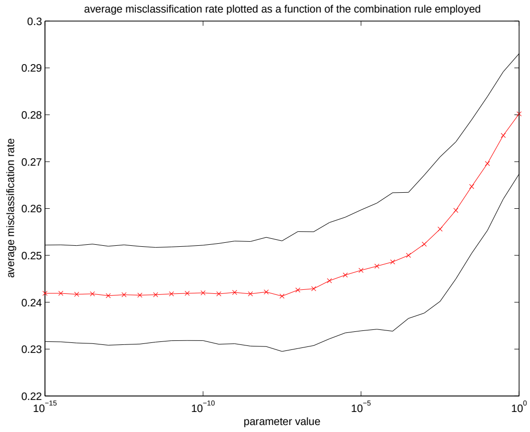

Figure 3: Misclassification rate plotted as a function of s; correlation coefficient $\rho = 0.9$.

3.3 Learning the Combination Rule

Given a set of classifiers, we want to learn the combination rule that will give the best classification results on new data. We propose here a method to learn this rule, by selecting the parameter value optimizing the performances of the classifiers' ensemble on a set of data whose actual class is known.

Rather than computing error rates, we used a more sensitive criterion based on pignistic probabilities to estimate these performances. Let $\textcircled{T}_s$ be the parameterized rule of combination (thereafter abbreviated as PaRC) defined by some parameter value s. Then, a bba m, quantifying our belief on the actual class of a test pattern $\mathbf{x}$, is obtained by combining the outputs m_i of the classifiers evaluating $\mathbf{x}$:

$$m = m_1 \textcircled{T}_s \dots \textcircled{T}_s m_p.$$

Then, finding an adequate value for parameter s can be done by solving:

$$\widehat{s} = \arg\min_{0 \le s \le 1} \sum_{j=1}^{n} \left\| \mathbf{betp}_j - \mathbf{d}_j \right\|^2 ; \qquad (12)$$

here, given any pattern $\mathbf{x}_j$ from the validation set, $\mathbf{betp}_j = (BetP_j(\omega_1), \dots, BetP_j(\omega_K))$ is the pignistic probability distribution obtained from m, and $\mathbf{d}_j = (d_{j1}, \dots, d_{jn})$ encodes its crisp membership to the classes ($d_{jk} = 1$ if $\mathbf{x}_j \in \omega_k$, and 0 otherwise).

We propose to search the parameter space (restricted to $]0; 1]$) for $\widehat{s}$. For example, we can use a dichotomic search algorithm, stopping when the width of the interval to search is less than some constant (e.g., 10^{-10}).

4 Results

We considered five synthetic datasets, generated as described in Section 3.2, characterised by $\rho = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the real-data sets Ecoli, Glass, Segment, Vowel, and Waveform from the UCI Machine Learning Repository. To compute smoother estimates, we processed the mean of several values of the error defined by (12) obtained on various validation sets. These sets were either generated apart when possible — on each synthetic dataset, 5000 validation patterns were generated in each class — or obtained from the training sets via 5×2 cross-validation.

Table 1 gives the number of classes, of features, and the numbers of patterns in the training and test sets.

Table 1: Description of the datasets.

dataset	# classes	# features
Synthetic	2	10
Ecoli	8	7
Glass	6	9
Segment	7	19
Vowel	11	10
Waveform	3	21

dataset	number of patterns	
	training	test
Synthetic	2000	10000
Ecoli	201	135
Glass	139	75
Segment	1400	910
Vowel	528	462
Waveform	1491	3509

For each dataset, we trained a classifier (logistic regression) on each variable, by the procedure described in Section 3.2. Error rates obtained with the CoRC, the CaRC, and the PaRC corresponding to the optimal parameter value, are provided in Table 2, 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 [3] at level 5%.

Table 2: Error rates of the CoRC, the CaRC and the PaRC, and 95% confidence intervals.

data	CoRC	PaRC ($\hat{s}$)	CaRC
Synth.	<u>**11.45**</u>	**11.46** (0.75)	15.68
$\rho = 0.1$	[11.00;11.89]	[11.01;11.90]	[15.18;16.19]
Synth.	**18.64**	<u>**18.58**</u> (0.5)	20.23
$\rho = 0.3$	[18.11;19.18]	[18.04;19.12]	[19.67;20.79]
Synth.	23.17	<u>**22.61**</u> (9.8e-4)	22.98
$\rho = 0.5$	[22.59;23.75]	[22.03;23.18]	[22.40;23.56]
Synth.	25.77	**24.51** (1.2e-10)	<u>**24.50**</u>
$\rho = 0.7$	[25.16;26.38]	[23.91;25.11]	[23.91;25.10]
Synth.	27.51	<u>**24.23**</u> (0)	<u>**24.23**</u>
$\rho = 0.9$	[26.89;28.13]	[23.64;24.83]	[23.64;24.83]
Ecoli	44.44	<u>**37.04**</u> (0)	<u>**37.04**</u>
	[36.06;52.83]	[28.89;45.18]	[28.89;45.18]
Glass	**49.33**	<u>**45.33**</u> (0.25)	<u>**45.33**</u>
	[38.02;60.65]	[34.07;56.60]	[34.07;56.60]
Segm.	18.35	<u>**15.71**</u> (0)	<u>**15.71**</u>
	[15.84;20.87]	[13.35;18.08]	[13.35;18.08]
Vowel	<u>**55.84**</u>	**56.71** (7.8e-3)	**56.71**
	[51.32;60.37]	[52.19;61.23]	[52.19;61.23]
Wave.	16.93	<u>**16.13**</u> (0)	**16.53**
	[15.69;18.17]	[14.91;17.35]	[15.30;17.76]

We may remark on the synthetic datasets that all error rates increase with the correlation of the features. This is not surprising, as less information becomes then available. The CoRC yields obviously the best results when the correlation is low, and the CaRC when it is high. The PaRC is close to the CoRC for $\rho = 0.1$; when $\rho = 0.3$ or $\rho = 0.5$, it is truly intermediate between the CoRC and the CaRC; when $\rho = 0.7$ and $\rho = 0.9$, learning the combination rule yields the CaRC itself.

On the real data sets, the CaRC was always learnt, except on the Waveform dataset for which an intermediate rule was learnt. The optimized rule gave the best results, except on the Vowel dataset, for which the CoRC performed better. The reason may be that the training and test distributions differ.

5 Refined Combination of Sources

The method presented above consists in adapting a unique combination rule to non-independent data. It relies on the implicit assumption that all the classifiers have the same pairwise dependency, which may be too simplistic. For example, in the synthetic datasets we generated, all the sources share the same dependency *but one*, which is conditionally independent on the others.

A distance d_J between two (normal) bbas m_1 and m_2 was defined in [9], by:

$$d_J(m_1, m_2) = \sqrt{\frac{(m_1 - m_2)\, D\, (m_1 - m_2)^\top}{2}}, \tag{13}$$

with an element $D_{A,B}$ of matrix D defined by:

$$D_{A,B} = \frac{A \cap B}{A \cup B}, \ \forall A \neq \emptyset, B \neq \emptyset. \tag{14}$$

We computed the average distance between the outputs of each pair of classifiers, using (13)-(14). Figures 4 to 6 show the dendrograms representing the hierarchy built upon the shortest distance: at distance $\widehat{d_J}$, two sets of classifiers are aggregated iff we can find a classifier in each set such that their distance to each other is $\widehat{d_J}$. While for $\rho = 0.1$, the average distances range quite uniformly, for $\rho = 0.3$ and $\rho = 0.9$ we can see that the tenth classifier is clearly separated from the others. Clustering the classifiers would lead to group the first nine, and leave the tenth one alone.

Hence, we can define a hierarchical rule of combination: the classifiers are clustered according to their pairwise average distances; their outputs are first combined in each cluster, and the resulting bbas are then combined.

We applied this combination strategy of the classifiers to the five synthetic data sets generated. We combined the first nine classifiers using a combination rule learnt as described in Section 3.3, and then we combined the tenth classifier using the CoRC. Table 3 shows the error rates (again, together with 95% confidence intervals) thus obtained. These rates were printed in bold when judged significantly

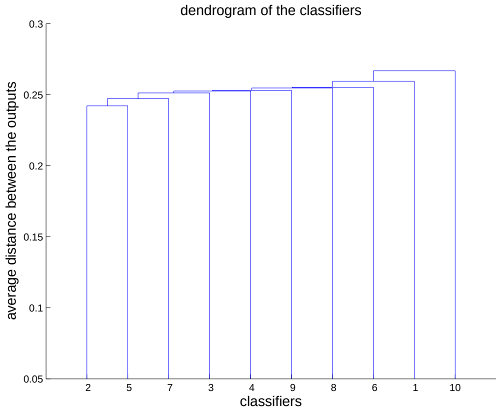

Figure 4: Dendrogram of the average distances between the classifiers; $\rho = 0.1$.

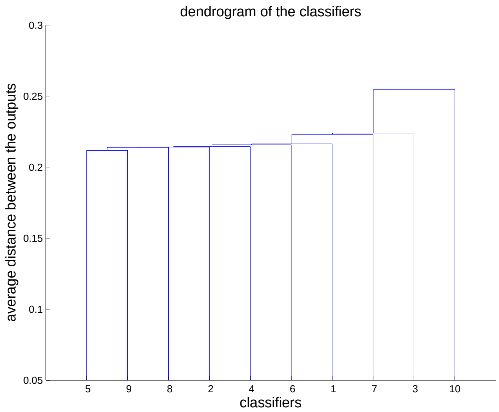

Figure 5: Dendrogram of the average distances between the classifiers; $\rho = 0.3$.

lower (by a McNemar test at level 5%) than the best rate given in Table 2.

These results are better than those obtained with other rules; the differences are significant except for $\rho = 0.1$. That demonstrates the interest of modeling all the knowledge on the dependency of the variables. One may also remark that the rule learnt is closer to the CaRC than previously: the reason is that the 10^{th} classifier (which is independent) is not considered in the learning step any more.

6 Conclusions

In this article, we addressed the problem of combining multiple sources for solving a classification problem. We concentrated on problems where the information sources are not independent. We studied the influence of the combination rule on classification results. We considered the commonly used Dempster's rule of combination [11], and the cautious rule of combination [1, 2]. The former requires that the combined sources be independent; the latter pools sources that may be provide identical information. We proposed to define a family of rules for separable bbas enclosing both as extremal cases, and to learn the rule that best suits the data processed.

On real data, the cautious rule often yields the best classification results, although we showed that Dempster's rule, or some rule intermediate between Dempster's rule and the cautious rule, may also give the best performances. Having demonstrated the validity of this method, we refined it by clustering classifiers using the average distance between

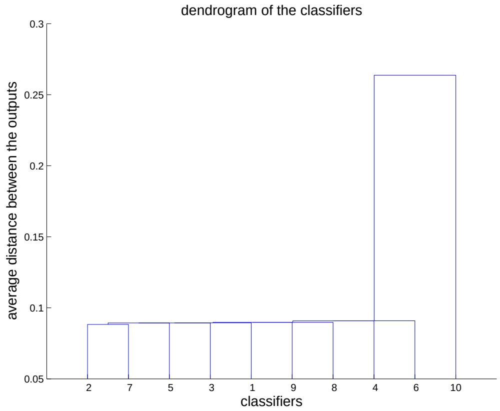

Figure 6: Dendrogram of the average distances between the classifiers; $\rho = 0.9$.

Table 3: Error rates and confidence intervals given by a hierarchical rule of combination.

dataset	results ($\hat{s}$)
Synth. $\rho = 0.1$	11.34 (0.5) [10.91;11.78]
Synth. $\rho = 0.3$	**17.66** (0.0625) [17.14;18.19]
Synth. $\rho = 0.5$	**20.87** (4.75e-7) [20.31;21.43]
Synth. $\rho = 0.7$	**22.68** (0) [22.10;23.27]
Synth. $\rho = 0.9$	**23.91** (0) [23.32;24.51]

their outputs. Thus, combination may be processed first within each cluster, the resulting bbas being then combined together.

Future work will focus on two points. First, we may cluster automatically the set of sources. In addition, searching the parameter space for several optimal values becomes computationally demanding when several rules have to be learnt. Efforts should be thus directed on deducing the combination rule of two (clusters of) classifiers using distances between classifier outputs.

References

[1] T. Denœux (2006). The cautious rule of combination for belief functions and some extensions. In *Proceedings of FUSION'2006*, Florence, Italy, July 2006.

[2] T. Denœux (2007). Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. *Artificial Intelligence*, vol. 172, pages 234-264, 2008.

[3] T. Dietterich (1998). Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation*, vol. 10 (7), p. 1895-1923, 1998.

[4] D. Dubois and H. Prade (1986). On the unicity of Dempster's rule of combination. *International Journal of Intelligent Systems*, vol. 12, p. 133-142, 1986.

[5] 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, p. 75-84, 1987.

[6] D. Dubois, H. Prade, and Ph. Smets (2001). New semantics for quantitative possibility theory. In *Proceedings of ECSQARU'2001*, pages 410-421, Toulouse, France, 2001.

[7] D. Dubois and H. Prade (2007). A definition of subjective possibility. *International Journal of Approximate Reasoning*, in press, 10.1016/j.ijar.2007.01.005.

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

[9] A.-L. Jousselme, D. Grenier and É. Bossé (2001). A new distance between two bodies of evidence. *Information Fusion*, vol. 2, p. 91-101, 2001.

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

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

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

[13] P. Smets (1990). The combination of evidence in the transferable belief model. *IEEE Pattern Analysis and Machine Intelligence*, vol. 12, p. 447-458, 1990.

[14] P. Smets (1993). Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, vol. 9 (1), p. 1-35, 1993.

[15] P. Smets and R. Kennes (1994). The Transferable Belief Model. *Artificial Intelligence*, vol. 66, p. 191-234, 1994.

[16] P. Smets (1995). The canonical decomposition of a weighted belief. In *International Joint Conferences in Artificial Intelligence*, p. 1896-1901, Montréal, Canada, 1995.

[17] P. Smets (1998). Data fusion in the Transferable Belief Model. In *International Conference on Information Fusion*, p. 21-33, Paris, France, 1998.

[18] R. Yager (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, vol. 41, p. 93-137, 1987.