Fusion of multi-level decision systems using the Transferable Belief Model

David Mercier
Université de Technologie de Compiègne / SOLYSTIC
UMR CNRS 6599 Heudiasyc, BP20529
F-60205 Compiègne Cedex, France
Email: dmercier@hds.utc.fr
david.mercier@solystic.com

Geneviève Cron
SOLYSTIC
14 avenue Raspail
F-94257 Gentilly Cedex, France
Email: genevieve.cron@solystic.com

Thierry Denœux
Université de Technologie de Compiègne
UMR CNRS 6599 Heudiasyc, BP20529
F-60205 Compiègne Cedex, France
Email: tdenoeux@hds.utc.fr

Mylène Masson
Université de Picardie Jules Verne
UMR CNRS 6599 Heudiasyc, BP20529
F-60205 Compiègne Cedex, France
Email: mmasson@hds.utc.fr

***Abstract*— In this paper, we are interested in the fusion of classifiers providing decisions which are organized in a hierarchy, i.e., for each pattern to classify, each classifier has the possibility to choose a class, a set of classes, or a reject option.**

We present a method to combine these decisions based on the Transferable Belief Model (TBM), an interpretation of the Dempster-Shafer theory of evidence. The TBM is shown to provide a powerful and flexible framework, well suited to this problem. Special emphasis is put on the construction of basic belief assignments, an important issue which has not yet been fully explored in the literature. We propose an approach extending a former proposal made by Xu, Krzyzak and Suen (1992) in a simpler context. A rational decision modelling allowing different levels of decision is also presented.

Finally, the proposed combination is compared experimentally to several simpler alternatives.

***Index Terms*— Decision Fusion, multi-level decisions, belief functions, Dempster-Shafer theory, Evidence theory, classification.**

I. Introduction

Building highly reliable classifiers is an important objective in pattern recognition. An interesting way to achieve this goal consists in the combination of already existing classifiers. Indeed, experimental results ([1], [2]) show that methods based on multiple classifiers generally outperform each individual classifier. As explained by Xu, Krzyzak and Suen in [3], the combination of multiple classifiers includes several problems: selecting the classifiers to combine (types, algorithms, number, ...), choosing an architecture for the combination (parallel, cascade, mixtures of both, ...), and combining the classifier outputs in order to achieve better performance than each classifier individually.

In this paper[1], we focus on the problem of combining classifiers providing decisions which are organized as a hierarchy: decisions can be expressed at different levels. For each pattern to classify, each classifier has the possibility to choose either a class, or a set of classes, or rejection.

We assume that the decisions are not associated with any scoring vector, or posterior probabilities. It is a common situation in real world applications.

Classifiers providing only class labels are called *abstract level classifiers* or *classifiers of Type 1* in [3]. To combine such classifiers, various combination techniques were proposed such as voting-based systems [3], [4], [5], plurality [6], Bayesian theory [3], Dempster-Shafer theory [3], [7] or classifier local accuracy [8].

Inspired by a former proposal by Xu, Krzyzak and Suen (1992) [3], a combination of these decisions based on the Transferable Belief Model (TBM) ([9], [10]) is proposed. Like all Dempster-Shafer approaches, the assignment of masses is an important task which often determines the success of the combination. Therefore, different assignments are discussed. A decision process allowing different levels of decision is also presented.

This paper is organized as follows. The key points of the TBM, an interpretation of the Dempster-Shafer theory of evidence [11] well suited to information fusion, are recalled in Section II. In Section III, we come back to an existing method for combining belief functions in the case of non hierarchical decisions. Then, in Section IV, a model based on the TBM is presented for the combination of multi-level decisions. Finally, Section V describes experimental results and compares the proposed combination with voting-based schemes.

[1]This work is the result of a cooperation agreement between the Heudiasyc laboratory at the Université de Technologie de Compiègne and the SOLYSTIC company.

II. The Transferable Belief Model (TBM): foundations

A. Information representation

Let X be a variable taking values in a finite set Ω, called the *frame of discernment* (or *frame*). Ω is composed of mutually exclusive elements $\omega_1, \ldots, \omega_K$ called *atoms*. The knowledge held by a rational agent Y, regarding the actual value ω_0 taken by X, can be quantified by a belief function defined from the power set 2^Ω to $[0, 1]$.

Belief functions can be expressed in several forms: the *basic belief assignment (BBA)* m, the *credibility function* bel and the *plausibility function* pl, which are in one-to-one correspondance. We recall that $\mathrm{m}(A)$ quantifies the *part* of belief that is restricted to the proposition $\omega_0 \in A \subseteq \Omega$ and satisfies:

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

Thus, a BBA can support a set $A \subseteq \Omega$ without supporting any subproposition of A, which allows to account for partial knowledge.

Some particular belief functions often used, are defined as follows:

Definition 1: The *vacuous belief function* quantifies total ignorance:

$$\mathrm{m}^\Omega(\Omega) = 1.$$

Bayesian belief functions quantify perfect knowledge on X's value:

$$\mathrm{m}^\Omega(A) \neq 0 \Rightarrow |A| = 1.$$

B. Handling the knowledge

Two distinct pieces of evidence, quantified by BBAs m_1 and m_2, may be combined, using a suitable operator. The most common are the *conjunctive rule of combination* (CRC) and the *disjunctive rule of combination* (DRC), defined, respectively, as:

$$\mathrm{m}_1 \textcircled{\cap} \mathrm{m}_2(A) = \sum_{B \cap C = A} \mathrm{m}_1(B)\,\mathrm{m}_2(C), \forall A \subseteq \Omega;$$
$$\mathrm{m}_1 \textcircled{\cup} \mathrm{m}_2(A) = \sum_{B \cup C = A} \mathrm{m}_1(B)\,\mathrm{m}_2(C), \forall A \subseteq \Omega.$$

If the two distinct pieces of evidence are trustful enough, the CRC is used. Otherwise, if at least one piece of evidence is reliable, the DRC can be used.

C. Decision making

When an agent has to select an optimal action among an exhaustive set of actions, *rationality principles* [12], [13] justify the strategy that consists in choosing the one that minimizes the *expected risk* (or *expected cost*). This principle leads to the use of a probability measure $P^\Omega : 2^\Omega \rightarrow [0; 1]$ and a cost function $c : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$, where $\mathcal{A}$ is the set of possible actions. The optimal action is then the one that minimizes the expected cost (risk) defined by:

$$\rho(\alpha) = \sum_{\omega \in \Omega} c(\alpha, \omega) P^\Omega(\{\omega\}). \tag{2}$$

Therefore, when a decision has to be made, the BBA obtained after the combination must be transformed into a probability measure. One solution proposed in [14] consists in using the *pignistic transformation* [15], [16] to compute the *pignistic probability*:

$$BetP(\{\omega\}) = \sum_{\{A \subseteq \Omega, \omega \in A\}} \frac{\mathrm{m}(A)}{\mid A \mid \ (1 - m(\emptyset))}. \tag{3}$$

Most of the time, classification algorithms do not directly compute a BBA. Thus, to apply the TBM or any model based on the Dempster-Shafer theory, each classifier's output has to be converted in the form of a BBA. This task is very important as each BBA is supposed to represent all the knowledge provided by a classifier. In particular, BBAs should reflect each classifier's strenghs and weaknesses. The following section aims at representing the information produced by each classifier through the best possible BBA.

III. Mass assignment through the TBM

In this paper, decisions are assumed to be the only pieces of information available on each individual classifier. In particular, we will not use the feature vector of pattern x used in others approaches [7], [17]. The BBAs representing the knowledge on each classifier can be built from the decisions already proposed in a learning set. For this task, confusion matrices will be used. First, some definitions are given.

A. Definitions

Let $\mathcal{C} = \{C_1, \ldots, C_N\}$ be a set of N classifiers, and let $\Omega = \{\omega_1, \ldots, \omega_K\}$ be a set of K class labels.

In this section, a classifier is viewed as a function C taking as input a pattern x from a set of patterns $\mathcal{P}$ and outputting a class label $C(x) = \omega_k \in \Omega \cup \{\omega_{K+1}\}$, where by convention ω_{K+1} denotes the rejection class.

Definition 2: The *confusion matrix* $\mathcal{M}_i = (n^i_{kl})_{\{k \in \{1,\ldots,K+1\}\ l \in \{1,\ldots,K\}\}}$ (Table I) of classifier C_i, computed from test data, allows to sum up the correct answers and the errors of classifier C_i for each class ω_l. Each row k corresponds to the decision $C_i(x) = \omega_k$. Each column l corresponds to the actual class ω_l. n^i_{kl} is the number of patterns of actual class ω_l which have been classified by C_i in class ω_k. For all $k \in \{1, \ldots, K+1\}$, let $n^i_{k\cdot} = \sum_{l=1}^{K} n^i_{kl}$ be the number of patterns classified by C_i in ω_k. For example, $n^i_{(K+1)\cdot}$ is the number of rejections made by classifier C_i. Let $n^i = \sum_{k=1}^{K+1} \sum_{l=1}^{K} n^i_{kl}$ be the total number of patterns classified by C_i.

Definition 3 (performance rates): The performance of each classifier C_i will be measured by the *recognition rate* (correct answer rate), the *error rate* (or *substitution rate*) and the *rejection rate* noted, respectively, R_i, S_i and T_i. Performance rates of classifier C_i are computed from its confusion matrix:

- the *recognition rate* of classifier C_i

$$R_i = \frac{\sum_{k=1}^{K} n^i_{kk}}{n^i}, \tag{4}$$

TABLE I

ILLUSTRATION OF A CONFUSION MATRIX M_i.

		ACTUAL		
		ω_1	$\dots$	ω_K
D	ω_1	n^i_{11}	$\dots$	n^i_{1K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
E	ω_K	n^i_{K1}	$\dots$	n^i_{KK}
C	ω_{K+1}	$n^i_{(K+1)1}$	$\dots$	$n^i_{(K+1)K}$
I				

- the *substitution rate* of classifier C_i

$$S_i = \frac{\sum_{k=1}^{K} \sum_{l=1;l\neq k}^{K} n^i_{kl}}{n^i}, \tag{5}$$

- the *rejection rate* of classifier C_i

$$T_i = \frac{n^i_{(K+1)\cdot}}{n^i}. \tag{6}$$

Thus $\forall i \in \{1, \dots, N\}$, $R_i + S_i + T_i = 1$.

Another rate, allowing to measure the reliability of a classifier without regarding the rejection rate, is defined by:

$$\mathcal{R}_i = \frac{\sum_{k=1}^{K} n^i_{kk}}{\sum_{k=1}^{K} \sum_{l=1}^{K} n^i_{kl}} = \frac{R_i}{1 - T_i}, \tag{7}$$

and is called the *reliability rate* of classifier C_i.

Definition 4 (classifiers comparison): One classifier C_i is said to *outperform* another classifier C_j if and only if :

- C_i has a better recognition rate than C_j and a lower substitution rate: $R_i > R_j$ and $S_i < S_j$.
- Or, C_i has a better recognition rate than C_j with the same substitution rate: $R_i > R_j$ and $S_i = S_j$.
- Or, C_i has a lower substitution rate than C_j with the same recognition rate: $S_i < S_j$ and $R_i = R_j$.

This relation defines a partial order.

The performances of each classifier can be represented in a graph with the recognition rate on the x-axis and the error rate on the y-axes. This graph allows to visualize in a simple way which classifier has the best performance and which classifiers are not comparable.

Example 1: Figure 1 represents the performances of 4 classifiers C_1, C_2, C_3 and C_4. Classifier C_2 outperforms all others. Classifiers C_1 and C_3 are not comparable.

B. Mass assignment for the combination of non-hierarchical decisions

1) Bayesian assignment: The confusion matrix allows to take into account each classifier's performance with respect to each class. When $C_i(x) = \omega_k$, the Bayesian BBA m_i representing information coming from classifier C_i, supports each $\omega_l \in \Omega$ with a mass equal to the ratio of number of patterns in class ω_l which have been classified by C_i in class ω_k, to the total number of patterns classified by C_i in class ω_k:

$$\forall \omega_l \in \Omega,\ \mathrm{m}_i(\{\omega_l\}) = \frac{n^i_{kl}}{\sum_{j=1}^{K} n^i_{kj}} = \frac{n^i_{kl}}{n^i_{k\cdot}}. \tag{8}$$

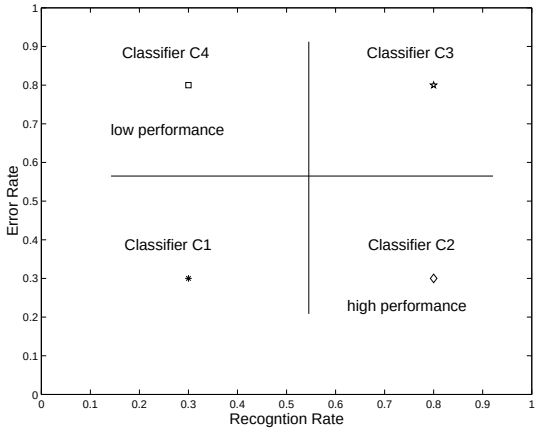

Fig. 1. Representation of the performances of classifiers.

Assignment (8) is called *Bayesian assignment* as it leads to Bayesian BBAs. With such an assignment, even a rejection brings information on the actual class (it does not lead to a vacuous BBA). However, for the ratios $n^i_{kl}/n^i_{k\cdot}$ to be statistically significant, the learning set must be well chosen and large enough. In particular, when the number of classes is high, it is generally not possible to exploit the whole confusion matrix. In this case, it is preferable to group some n^i_{kl} and to build less precise BBAs.

2) Xu's assignment: When $C_i(x) = \omega_k$ with $k \in \{1, \dots, K\}$, it is proposed in [3] to define m_i by:

$$\begin{array}{rccc} \mathrm{m}_i\ : & 2^{\Omega} & \longrightarrow & [0,1] \\ & \{\omega_k\} & \longmapsto & R_i \\ & \Omega \setminus \{\omega_k\} & \longmapsto & S_i \\ & \Omega & \longmapsto & T_i \end{array} \tag{9}$$

When $C_i(x) = \omega_{K+1}$, $\mathrm{m}_i(\Omega) = 1$ which means: when C_i makes a rejection, assignment (9) leads to the vacuous belief function.

This assignment is based on the following idea: the higher the recognition rate, the greater the confidence on the classifier decision. This assignment was tested in [3] on a digit recognition problem and provided good results. However, as shown by the following example, the confidence in the classifier decision should not depend only on the recognition rate.

Example 2: Let us consider two classifiers C_1 and C_2 with the following performance rates:

	R_i	S_i	T_i
C_1	90%	1%	9%
C_2	20%	0.1%	$79,9\%$

Let us assume that $C_1(x) = \omega_k$ and $C_2(x) = \omega_l$, assignment (9) yields $m_1(\{\omega_k\}) = 0.9$ and $m_2(\{\omega_l\}) = 0.2$. This assignment is not what could be expected since classifier C_2's decisions, when different from a rejection, are correct most of the time. In fact C_2 is a very cautious classifier which makes many rejections to have a minimum of errors. Thus, as decisions coming from classifier C_2 are more reliable than C_1's decisions, $m_2(\{\omega_l\})$ should be higher than $m_1(\{\omega_k\})$.

3) Reliability Assignment: Example 2 shows what can happen when the behaviors of two classifiers are different. To overcome this problem, we propose to use the *reliability rate* $\mathcal{R}_i$ (7) of classifier C_i. Indeed, $\mathcal{R}_i$ represents the percentage of "good" classification knowing C_i has decided a class label different from a rejection. When $C_i(x) = \omega_k$ with $k \in \{1, \ldots, K\}$, we propose the assignment:

$$\begin{array}{rccc} \mathrm{m}_i : & 2^\Omega & \longrightarrow & [0,1] \\ & \{\omega_k\} & \longmapsto & \mathcal{R}_i \\ & \Omega & \longmapsto & \mathcal{U}_i, \end{array} \tag{10}$$

where $\mathcal{U}_i$ is the *unreliability rate* of the classifier C_i:

$$\mathcal{U}_i = 1 - \mathcal{R}_i = 1 - \frac{R_i}{1 - T_i} = \frac{S_i}{1 - T_i}, \tag{11}$$

When C_i makes a rejection, $\mathrm{m}_i(\Omega) = 1$ like assignment (9).

Assignment (10) only takes into account the reliability of the classifiers when a class is chosen.

This assignment is a particular case of a more general assignment that will be defined in the following section to represent information coming from classifiers when decisions are organized in a hierarchy.

Remark 1: Assignment (10) corresponds to the least committed mass [18] in agreement with the incomplete plausibility function :

$$\begin{array}{rcl} \mathrm{pl}_i(\{\omega_k\}) & = & 1 \\ \mathrm{pl}_i(A) & = & \mathcal{U}_i,\ \forall A\ s.t.\ \omega_k \notin A. \end{array}$$

Before having information coming from classifier C_i, all propositions are totally plausible. Then, when classifier C_i outputs a class ω_k, the proposition "the actual class is ω_k" remains entirely plausible, but each set which does not contain ω_k, becomes less plausible depending on the classifier reliability.

A variant of the previous assignment consists in building a more committed assignment:

$$\begin{array}{rccc} \mathrm{m}_i : & 2^\Omega & \longrightarrow & [0,1] \\ & \{\omega_k\} & \longmapsto & \mathcal{R}_i \\ & \Omega \setminus \{\omega_k\} & \longmapsto & 1 - \mathcal{R}_i = \mathcal{U}_i, \end{array} \tag{12}$$

Assignment (12) corresponds to the least committed mass in agreement with the incomplete plausibility function

$$\begin{array}{rcl} \mathrm{pl}_i(\{\omega_k\}) & = & \mathcal{R}_i \\ \mathrm{pl}_i(A) & = & \mathcal{U}_i,\ \forall A\ s.t.\ \omega_k \notin A. \end{array}$$

Unlike Assignment (10), the classifier reliability here influences the degree of belief in $\{\omega_k\}$, too. If the classifier is unreliable, we are tempted to think the answer is not the classifier decision. This assignment brings more conflicts than the first one, that's why we choose to generalize the flexible Assignment (10) in the following section.

IV. A MODEL FOR THE COMBINATION OF MULTI-LEVEL DECISIONS (CMLD) IN THE TBM

A. Problem formalization

From now on, we consider a set of N classifiers C_i, $i \in \{1, \ldots, N\}$, selecting for each pattern x from a set of patterns $\mathcal{P}$, either a class or a set of classes, according to a hierarchy of $\Omega = \{\omega_1, \ldots, \omega_K\}$. This hierarchy is assumed to be common to all the classifiers. For the sake of simplicity, only three levels in the hierarchy are considered, but our approach could be easily extended to more levels.

As classifiers can now select a set of classes, the rejection class is equivalent to a decision for the whole universe Ω. Consequently, rejection will from now on be noted $C_i(x) = \Omega$, instead of $C_i(x) = \omega_{K+1}$ as before.

Example 3: Let us consider sensors that recognize flying objects in $\Omega = \{A_1, A_2, H_1, H_2, R_1, R_2, R_3\}$ of three types: airplanes $A = \{A_1, A_2\}$, helicopters $H = \{H_1, H_2\}$, and rockets $R = \{R_1, R_2, R_3\}$.

According to the difficulty of the recognition task, each sensor can either recognize an object (an element of Ω) or a type of object (A, H, or R). In case of high uncertainty it can also select the reject option, which amounts to choosing the whole universe Ω.

The corresponding hierarchical decision space is depicted in Figure 2.

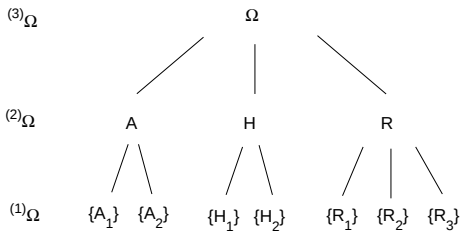

Fig. 2. A hierarchy of classifiers of example 3.

$^{(1)}\Omega$ is the set of decisions of level 1: $\{\{A_1\}, \{A_2\}, \{H_1\}, \{H_2\}, \{R_1\}, \{R_2\}, \{R_3\}\}$.

$^{(2)}\Omega$ is the set of decisions of level 2: $\{A, H, R\}$ with $A = \{A_1, A_2\}$, $H = \{H_1, H_2\}$, and $R = \{R_1, R_2, R_3\}$.

$^{(3)}\Omega$ is the set of decisions of level 3: $\{\Omega\}$.

The problem is then to fuse several decisions expressed at different levels in the hierarchy. This problem is refered to as *combination of multi-level decisions* and noted CMLD.

Example 4 (continuation of Example 3): Let us have 4 classifiers C_1, C_2, C_3, and C_4. Knowing:

- C_1 outputs "x is an airplane of model 1": $C_1(x) = \{A_1\}$,
- C_2 outputs "x is an airplane": $C_2(x) = A = \{A_1, A_2\}$ (i.e. x is an airplane of any model),
- C_3 outputs "x is a rocket": $C_3(x) = \{R_1, R_2, R_3\}$ (i.e. x is a rocket of any model),
- C_4 make a rejection "I don't know the type of x": $C_3(x) = \Omega$ (i.e. x is a flying object of any types).

What decision at which level should be undertaken by the combination of these classifiers?

We propose to model each classifier ouput by a belief function computed from a confusion matrix, using a generalization of Assignment (10) introduced in Section III.

B. Mass assignment for the CMLD

The proposed assignment is based on the use of reliability rates at each decision level.

Let u be a function assigning to each class or set of classes (other than Ω) the set of classes just above in the hierarchy. For instance, in Example 3, $u(\{A_1\}) = A$, $u(A) = \Omega$. Let us denote the elements at each level as indicated in Figure 3.

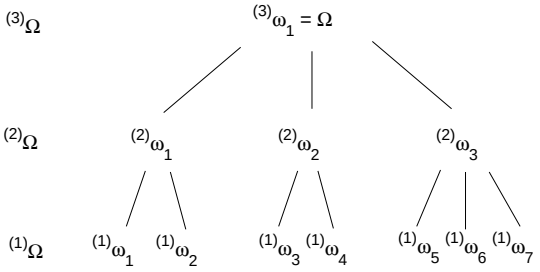

Fig. 3. Notation for a hierarchy.

Let ω_0 denote the actual value of the precise class of pattern x.

Assume that classifier C_i outputs a decision other than rejection. This decision is at a level p of the hierarchy ($p = 1$ or $p = 2$ and $C_i(x) \in {}^{(p)}\Omega$). Consider the following cases:

1) The actual value ω_0 of x is in $C_i(x)$. $C_i(x)$ is then the correct answer. The percentage of good recognition at level p is noted $\mathcal{R}_i[{}^{(p)}\Omega]$:

$$\mathcal{R}_i[{}^{(p)}\Omega] = \frac{\sum_{k=1}^{{}^{(p)}K} \sum_{\omega_l \in {}^{(p)}\omega_k} {}^{(p)}n_{kl}^i}{{}^{(p)}n^i}, \quad (13)$$

where ${}^{(p)}K$ is the number of decisions at level p. In particular ${}^{(1)}K = K$, ${}^{(2)}K$ is the number of set of classes at level 2. ${}^{(p)}n_{kl}^i$ is the number of patterns of actual class ω_l which have been classified by C_i in class ${}^{(p)}\omega_k$. ${}^{(p)}n^i = \sum_{k=1}^{{}^{(p)}K} \sum_{l=1}^{K} {}^{(p)}n_{kl}^i$ is the total number of patterns classified by C_i at level p.

2) $C_i(x)$ contains only one class ($C_i(x)$ is a decision of level 1) and ω_0 is in $u(C_i(x)) \setminus C_i(x)$ with $u(C_i(x)) \subset \Omega$; this means that the class selected by the classifier is not correct but is in the right set of classes. For instance, the sensor has decided the wrong model of airplane but the flying object was actually an airplane. The percentage of these errors, considered as errors of type 1, is noted $\mathcal{U}_i^1[{}^{(1)}\Omega]$.

$$\mathcal{U}_i^1[{}^{(1)}\Omega] = \frac{\sum_{k=1}^{K} \sum_{\omega_l \in u(\omega_k);\, l \neq k} {}^{(1)}n_{kl}^i}{{}^{(1)}n^i}. \quad (14)$$

3) The true class ω_0 is in $\Omega \setminus C_i(x)$ if $C_i(x)$ is a set of classes, or in $\Omega \setminus u(C_i(x))$ if $C_i(x)$ is a decision of level 1, which means that the actual value of x is not in the set of classes containing the output of the classifier. The percentage of these errors considered as errors of type 2 is noted $\mathcal{U}_i^2[{}^{(p)}\Omega]$:

$$\mathcal{U}_i^2[{}^{(1)}\Omega] = 1 - \mathcal{R}_i[{}^{(1)}\Omega] - \mathcal{U}_i^1[{}^{(1)}\Omega], \quad (15)$$
$$\mathcal{U}_i^2[{}^{(2)}\Omega] = 1 - \mathcal{R}_i[{}^{(2)}\Omega]. \quad (16)$$

Consequently, when classifier C_i outputs a decision at level 1, $C_i(x) = \{\omega_k\} \in {}^{(1)}\Omega$, the assignment m_i is defined by:

$$\begin{array}{rccc} \mathrm{m}_i : & 2^\Omega & \longrightarrow & [0,1] \\ & C_i(x) & \longmapsto & \mathcal{R}_i[{}^{(1)}\Omega] \\ & u(C_i(x)) & \longmapsto & \mathcal{U}_i^1[{}^{(1)}\Omega] \\ & \Omega & \longmapsto & \mathcal{U}_i^2[{}^{(1)}\Omega]. \end{array} \quad (17)$$

When classifier C_i outputs a set of classes different from Ω, $C_i(x)$ is a decision of level ${}^{(2)}\Omega$, the assignment is the following:

$$\begin{array}{rccc} \mathrm{m}_i : & 2^\Omega & \longrightarrow & [0,1] \\ & C_i(x) & \longmapsto & \mathcal{R}_i[{}^{(2)}\Omega] \\ & \Omega & \longmapsto & \mathcal{U}_i^2[{}^{(2)}\Omega] \end{array} \quad (18)$$

If classifier C_i makes a rejection then m_i is the vacuous belief function.

Example 5 (Continued from Example 3): Let us assume that the confusion matrix of classifier C_1 is the one shown in Figure 4.

		ACTUAL						
		A_1	A_2	H_1	H_2	R_1	R_2	R_3
${}^{(1)}\Omega$	$\{A_1\}$	36	4	0	0	1	1	2
	$\{A_2\}$	2	24	0	0	0	0	4
	$\{H_1\}$	0	0	19	5	0	0	0
	$\{H_2\}$	0	0	1	21	0	0	0
	$\{R_1\}$	0	0	0	0	30	2	4
	$\{R_2\}$	1	0	0	0	2	15	8
	$\{R_3\}$	1	0	0	0	0	2	15
${}^{(2)}\Omega$	$\{A_1, A_2\}$	35	45	2	1	4	5	10
	$\{H_1, H_2\}$	0	0	42	20	0	0	1
	$\{R_1, R_2, R_3\}$	2	4	0	0	25	12	21
${}^{(3)}\Omega$	Ω	2	5	1	0	1	2	10

Fig. 4. Confusion matrix of classifier C_1.

For instance, the number of patterns of actual class A_1 which have been classified by C_1 in class of level 1 $\{A_1\}$ is equal to 36. The number of patterns of actual class A_1 which have been classified by C_1 in class of level 2 $\{A_1, A_2\}$ is equal to 35. And, the number of patterns of actual class A_1 which have been classified by C_1 in Ω is equal to 2, which means that classifier C_1 rejected 2 patterns from class A_1.

Assume that classifier C_1 outputs a decision of level 1 in the hierarchy (Figure 2), with $C_1(x) = \{A_1\} \in^{(1)} \Omega$. We have:

$$\begin{array}{rcl} \mathcal{R}_1[{}^{(1)}\Omega_1] & = & (36 + 24 + 19 + 21 + 30 + 15 + 15)/200 \\ & = & 0.80 \\ \mathcal{U}_1^1[{}^{(1)}\Omega_1] & = & (2 + 4 + 1 + 5 + 2 + 2 + 2 + 4 + 8)/200 \\ & = & 0.15 \\ \mathcal{U}_1^2[{}^{(1)}\Omega_1] & = & (1 + 1 + 1 + 1 + 2 + 4)/200 \\ & = & 0.05 \end{array} \quad (19)$$

Then:

$$\begin{array}{rcl} \mathrm{m}_1(\{A_1\}) & = & 0.80 \\ \mathrm{m}_1(\{A_1, A_2\}) & = & 0.15 \\ \mathrm{m}_1(\{\Omega\}) & = & 0.05 \end{array}$$

This use of the performances of classifiers to assign the mass is close to the approach in [8], where the purpose was to estimate the local class accuracy of each classifier. The

percentage of patterns x correctly assigned when $C_i(x) = \omega_k$ indicates the strengh of the belief in the fact that the class of x is actually ω_k. We generalize this approach to the case of a hierarchical decision space. However, in [8], the decision with the maximum local class accuracy is selected, whereas in our model all the BBAs coming from each classifier are combined.

Remark 2 (No set of classes): With no decision composed of sets of classes, the hierarchy contains two levels: ${}^{(2)}\Omega_1 = \{\Omega\}$ and ${}^{(1)}\Omega_1 = \Omega$, if $C(x)$ is not a rejection (then $C(x) = \{\omega_k\},\ k \in \{1, \ldots, K\}$), the assignment is the following:

$$\begin{array}{rccl} m\ : & 2^{\Omega} & \longrightarrow & [0,1] \\ & C(x) & \longmapsto & \mathcal{R}[{}^{(1)}\Omega_1] \\ & \Omega & \longmapsto & 1 - \mathcal{R}[{}^{(1)}\Omega_1] \end{array} \tag{20}$$

where $\mathcal{R}[{}^{(1)}\Omega_1]$ is equal to the reliability rate of C, thus this assignment is the same as (10).

C. Combining the BBAs

Assuming that the classifiers constitute distinct reliable pieces of evidence, the BBAs can be combined conjunctively.

Example 6 (Example 3 continuation): Let us consider these results after the assignment:

$$\begin{array}{lllllll} \mathrm{m}_1(\{A_1\}) & = & 0.8 & \mathrm{m}_2(\{A_1, A_2\}) & = & 0.7 \\ \mathrm{m}_1(\{A_1, A_2\}) & = & 0.15 & \mathrm{m}_2(\{\Omega\}) & = & 0.3 \\ \mathrm{m}_1(\{\Omega\}) & = & 0.05 & & & \\ & & & & & \\ \mathrm{m}_3(\{R_1, R_2, R_3\}) & = & 0.6 & m_4(\{\Omega\}) & = & 1 \\ \mathrm{m}_3(\{\Omega\}) & = & 0.4 & & & \end{array}$$

Then, with $\mathrm{m} = \mathrm{m}_1 \textcircled{\cap} \mathrm{m}_2 \textcircled{\cap} \mathrm{m}_3 \textcircled{\cap} \mathrm{m}_4$:

$$\begin{array}{lllllll} \mathrm{m}(\{A_1\}) & = & 0.320 & \mathrm{m}(\{R_1, R_2, R_3\}) & = & 0.009 \\ \mathrm{m}(\{A_1, A_2\}) & = & 0.074 & \mathrm{m}(\{\Omega\}) & = & 0.006 \\ \mathrm{m}(\{\emptyset\}) & = & 0.591 & & & \end{array} \tag{21}$$

D. Rational multi-level decision

When a decision has to be made, the combination of multi levels decisions (CMLD) has to compute an optimal action a among a set of actions $\mathcal{A}$. In our problem, the set of possible actions is:

$$\mathcal{A} = {}^{(3)}\ \Omega\ \cup\ {}^{(2)}\Omega\ \cup\ {}^{(1)}\Omega \tag{22}$$

where "*decide* ω_k" is identified to "$\{\omega_k\}$", and "*decide* Ω" means rejection.

The optimal action is computed according to the following costs:

- $\forall k \in [1, K]$, $c(\Omega, \omega_k)$ represents the cost to decide Ω (i.e. rejection) knowing that the actual class is ω_k. This is the price to pay for total rejection, it will be called *total rejection cost* or *general class rejection cost* and noted $\mathcal{C}_{R_{GC}}$.
- $\forall k \in [1, K]\ \forall l \in [1, L]$, $c({}^{(2)}\omega_l, \omega_k)$ represents the cost to decide a set of classes ${}^{(2)}\omega_l$ knowing that the actual class is ω_k. If $\omega_k \in {}^{(2)}\omega_l$, this is the price to pay for the decision of a set of classes instead of the precise class contained in this set of classes, it will be called a *precise class rejection cost* and noted $\mathcal{C}_{R_{PC}}$. Otherwise, this is the price to pay for having committed an error of set of classes, it will be called a *general class error cost* and noted $\mathcal{C}_{E_{GC}}$.
- $\forall \omega_k \in \Omega$, $c(\omega_j, \omega_k)$ represents the cost to decide class ω_j knowing that the actual class is ω_k. If $\omega_j = \omega_k$, this is the price to pay for having the good answer and this price is the lowest, so it assumed to be null. Otherwise, this is the price to pay for having committed an error of class, it will be called a *precise class error cost* and noted $\mathcal{C}_{E_{PC}}$.

The following ordering between theses costs is assumed:

$$0 \leq \mathcal{C}_{R_{PC}} \leq \mathcal{C}_{R_{GC}} \leq \mathcal{C}_{E_{PC}} \leq \mathcal{C}_{E_{GC}}. \tag{23}$$

Remark 3: It is natural to assume that $\mathcal{C}_{R_{PC}} \leq \mathcal{C}_{R_{GC}}$, $\mathcal{C}_{E_{PC}} \leq \mathcal{C}_{E_{GC}}$, $\mathcal{C}_{R_{PC}} \leq \mathcal{C}_{E_{PC}}$, $\mathcal{C}_{R_{GC}} \leq \mathcal{C}_{E_{GC}}$, and $\mathcal{C}_{R_{PC}} \leq \mathcal{C}_{E_{GC}}$. In this paper, the assumption $\mathcal{C}_{R_{GC}} \leq \mathcal{C}_{E_{PC}}$ is made, in accordance with the application of the last section. However, this assumption is problem-dependent, in another application it can be false. The main idea consists in choosing the right costs according to the problem to be solved.

The risk associated with action Ω is $\mathcal{C}_{R_{GC}}$, indeed:

$$\begin{array}{rcl} \rho(\Omega) & = & \displaystyle\sum_{\omega_k \in \Omega} c(\Omega, \omega_k) BetP(\{\omega_k\}) \\ & = & \mathcal{C}_{R_{GC}} \displaystyle\sum_{\omega_k \in \Omega} BetP(\{\omega_k\}) \\ & = & \mathcal{C}_{R_{GC}}. \end{array}$$

Example 7 (Continued from Example 3): From (21) and (3), the pignistic probability is computed:

$$\begin{array}{lllllll} BetP(\{A_1\}) & = & 0.8750 & BetP(\{R_1\}) & = & 0.0094 \\ BetP(\{A_2\}) & = & 0.0926 & BetP(\{R_2\}) & = & 0.0094 \\ BetP(\{H_1\}) & = & 0.0021 & BetP(\{R_3\}) & = & 0.0094 \\ BetP(\{H_2\}) & = & 0.0021 & & & \end{array} \tag{24}$$

Then:

$$\begin{array}{rcl} \rho(A_1) & = & \displaystyle\sum_{\omega \in \Omega} c(A_1, \omega) BetP(\{\omega\}) \\ & = & \mathcal{C}_{E_{PC}} BetP(\{A_2\}) \\ & & + \mathcal{C}_{E_{GC}} \displaystyle\sum_{\omega \in \Omega \setminus A} BetP(\{\omega\}). \end{array} \tag{25}$$

As $BetP(\{A_1\}) > BetP(\{\omega\})$, $\forall \omega \in \Omega \setminus \{A_1\}$ (24) and $\mathcal{C}_{E_{PC}} \leq \mathcal{C}_{E_{GC}}$ (23), we can show that for all $\omega \in \Omega \setminus \{A_1\}$:

$$\rho(A_1) \quad \leq \quad \rho(\omega). \tag{26}$$

Which means that if a decison of level 1 must be made, the decision will be A_1. However possible actions are also $A = \{A_1, A_2\}$, $H = \{H_1, H_2\}$, $R = \{R_1, R_2, R_3\}$ and Ω.

$$\begin{array}{rcl} \rho(A) & = & \mathcal{C}_{R_{PC}}(BetP(\{A_1\}) + BetP(\{A_2\})) \\ & & + \mathcal{C}_{E_{GC}} \displaystyle\sum_{\omega \in \Omega \setminus A} BetP(\{\omega\}). \end{array}$$

$$\begin{aligned}\rho(H) &= \mathcal{C}_{R_{PC}}(BetP(\{H_1\}) + BetP(\{H_2\})) \\ &\quad + \mathcal{C}_{E_{GC}} \sum_{\omega \in \Omega \setminus H} BetP(\{\omega\}).\end{aligned}$$

$$\begin{aligned}\rho(R) &= \mathcal{C}_{R_{PC}} \sum_{\omega \in R} BetP(\{\omega\}) \\ &\quad + \mathcal{C}_{E_{GC}} \sum_{\omega \in \Omega \setminus R} BetP(\{\omega\}).\end{aligned}$$

Likewise by (24) and (23) ($\mathcal{C}_{R_{PC}} \leq \mathcal{C}_{E_{GC}}$), $\rho(A) \leq \rho(R)$, and $\rho(A) \leq \rho(H)$. As already seen, $\rho(\Omega) = \mathcal{C}_{R_{GC}}$. At last, the value of costs will decide which action at which level will be undertaken. With (24):

$$\begin{aligned}\rho(A_1) &= 0.0926\, \mathcal{C}_{E_{PC}} + 0.0324\, \mathcal{C}_{E_{GC}} \\ \rho(A) &= 0.9676\, \mathcal{C}_{R_{PC}} + 0.0324\, \mathcal{C}_{E_{GC}} \\ \rho(\Omega) &= \mathcal{C}_{R_{GC}}. \end{aligned} \quad (27)$$

Thus with $0.0926\mathcal{C}_{E_{PC}} \leq 0.9676\mathcal{C}_{R_{PC}}$ and $0.0926\mathcal{C}_{E_{PC}} + 0.0324\mathcal{C}_{E_{GC}} \leq \mathcal{C}_{R_{GC}}$, i.e. with a low error cost or a high rejection cost, the decision will be made at level 1. Otherwise, if the error cost is high and the rejection cost is low, a decision of level 2 or 3 will be made.

Ideally, these costs are provided by experts of the considered application, and reflect financial costs. They can also be learnt from training data to obtain an expected behaviour of the CMLD.

V. Application

In this application, three classifiers C_1, C_2 and C_3 are available. They are considered as black boxes which provide hard decisions in a hierarchical decision space. The aim of this section is to compare the performances of CMLD with those of two different voting schemes, for a particular application.

A. Voting schemes

When all classifiers have relatively good performances and express their decisions on the same frame of discernment, majority voting is a good candidate [5]. In this application, only three classifiers C_1, C_2 and C_3 are available and classifier C_2 is known to have the best performances. Thus a good voting based strategy consists in selecting the decision of the best classifier unless the other two models agree. In that case, the output of the two other classifiers is chosen. This method will be called *Maj*C_2. In order to achieve a better recognition rate, a variant of the previous method consists in choosing the majority decision only if it is different from rejection. This method will be called *Maj*C_2++.

Example 8: If $C_1(x) = \Omega$, $C_2(x) = {}^{(2)}\omega_l$ and $C_3(x) = \Omega$, then $MajC_2(x) = \Omega$ and $MajC_2++ = {}^{(2)}\omega_l$.

If $C_1(x) = {}^{(1)}\omega_k$, $C_2(x) = {}^{(2)}\omega_l$ and $C_3(x) = \Omega$, then $MajC_2(x) = MajC_2++(x) = {}^{(2)}\omega_l$.

B. Performance measures

Since classifiers produce decisions at different levels, new definitions of recognition rates and error rates have to be introduced. The performances of each individual and combined classifier will be measured by recognition and substitution (error) rates at two levels.

The *recognition rate* of classifier C_i at level 1, noted ${}^{(1)}R_i$, is defined as the ratio of the number of good recognition at level 1, to the total number n of classified patterns:

$${}^{(1)}R_i = \frac{\sum_{k=1}^{K} {}^{(1)}n_{kk}^i}{n}. \quad (28)$$

The *substitution rate* of classifier C_i at level 1, noted ${}^{(1)}S_i$, is defined as the proportion of misclassifications at level 1, plus the proportion of misclassifications at level 2 (i.e. the proportion of decisions at level 2 which do not contain the actual class):

$$\begin{aligned}{}^{(1)}S_i &= \frac{\sum_{k=1}^{K} \sum_{l=1;l\neq k}^{K} {}^{(1)}n_{kl}^i}{n} \\ &\quad + \frac{\sum_{k=1}^{{}^{(2)}K} \sum_{l=1;u(\omega_l)\neq {}^{(2)}\omega_k}^{K} {}^{(2)}n_{kl}^i}{n}. \end{aligned} \quad (29)$$

The *recognition rate* of classifier C_i at level 2, noted ${}^{(2)}R_i$, is defined the proportion of decisions at level 1 which are included in the same set of classes of the actual class,plus the proportions of decisions at level 2 which contain the actual class:

$$\begin{aligned}{}^{(2)}R_i &= \frac{\sum_{k=1}^{K} \sum_{l=1;u(\omega_l)=u(\omega_k)}^{K} {}^{(1)}n_{kl}^i}{n} \\ &\quad + \frac{\sum_{k=1}^{{}^{(2)}K} \sum_{l=1;u(\omega_l)={}^{(2)}\omega_k}^{K} {}^{(2)}n_{kl}^i}{n}. \end{aligned} \quad (30)$$

Each decision at level 1 is thus considered as a decision at the upper level in the hierarchy. Finally, the *substitution rate* of classifier C_i at level 2, noted ${}^{(2)}S_i$, is defined as the proportion of decisions at level 1 whose set of classes above in the hierarchy does not contain the actual class, added to the proportion of decisions at level 2 which do not contain the actual class:

$$\begin{aligned}{}^{(2)}S_i &= \frac{\sum_{k=1}^{K} \sum_{l=1;u(\omega_l)\neq(\omega_k)}^{K} {}^{(1)}n_{kl}^i}{n} \\ &\quad + \frac{\sum_{k=1}^{{}^{(2)}K} \sum_{l=1;u(\omega_l)\neq {}^{(2)}\omega_k}^{K} {}^{(2)}n_{kl}^i}{n}. \end{aligned} \quad (31)$$

In this application, the costs were learnt from a learning set containing half of the data, in order to achieve the best recognition rate while maintaining the error rate inside an interval centered around the error rate of classifier C_2.

C. Results

All individual and combined classifiers are represented Figure 5, in the two performance spaces $({}^{(1)}R, {}^{(1)}S)$ and $({}^{(2)}R, {}^{(2)}S)$. The representation in the same figure allows to compare the classifier perormances at different levels.

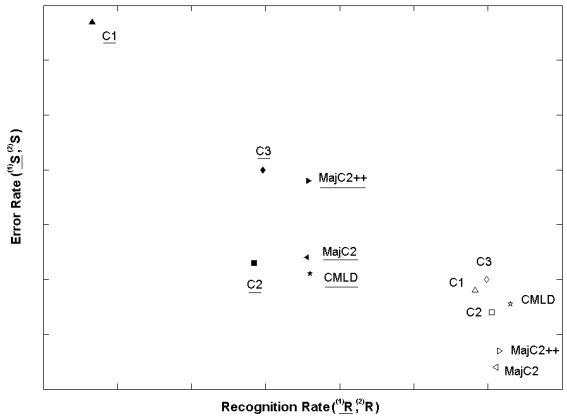

Fig. 5. Classifier performances at levels 1 (space $({}^{(1)}R, {}^{(1)}S)$, filled symbols, underlined classifiers' names) and 2 (space $({}^{(2)}R, {}^{(2)}S)$, blank symbols).

At level 1, CMLD outperform all individual classifiers, as well as majority based combinations. The combination $MajC_2 + +$ increases the recognition but the price on error rate is high.

At level 2, the performances of CMLD cannot be compared to those of the voting schemes: CMLD, although having a better recognition rate, also has a higher error rate. Indeed, CMLD makes fewer rejections than majority voting schemes: it can decide a solution proposed by only one classifier when the two others make a rejection, or it can find a compromise between two differents solutions provided by classifiers C_1 and C_3 while classifier C_2 makes a rejection.

At both decision levels, the error rate of CMLD was controlled to remain close to that of classifier C_2. Such a behavior cannot by obtained with the majority based combinations.

VI. Conclusion

In this paper, we tackled the problem of combining multi level decisions and presented an approach based on the Transferable Belief Model. The proposed approach allows to express the output from each classifier (in a hierarchical decision space) in the form a basic belief assignment computed from a confusion matrix. Experiments with data from a real application demonstrated the effectiveness of this approach, as compared to simple voting schemes.

The presented method could be applied to single-level decision problems by building a hierarchy on the set of classes. Such a hierarchy could be based on proximities between classes as revealed by the confusion matrix.

When the size of the universe is very large, it is also possible to compute several reliability rates at each level of the hierarchy, based on a partition of decisions at that level. Such a model is under construction. The selection of a "good" partition remains to be studied.

Finally, individual classifiers usually provide, together with a hard decisions, additional information in the form of scores (e.g., estimated posterior probabilities, degrees of membership, etc.). Combining such scores with the confusion matrix to define more informative belief functions is an interesting problem which is left for further research.

Acknowledgment

The authors would like to thank Philippe Smets and two anonymous referees for their helpful comments.

References

[1] Kitler J., Hatef M., Duin R.P.W., and Matas J., *On combining classifiers*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 20, pp. 226–239, 1998.

[2] Jain A.K., Duin R.P.W., and Mao J., *Statistical Pattern Recognition: a Review*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 22, pp. 4–37, 2000.

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

[4] Lam L., and Suen C. Y., *Optimal combination of pattern classifiers*, Pattern Recognition Letters, vol 16, pp. 945–954, 1995.

[5] Lam L., and Suen C. Y., *Application of majority voting to pattern recognition: An analysis of its behaviour and performance*, IEEE Transactions on systems, Man and Cybernetics, vol 27, pp. 553–568, 1997.

[6] Hansen L.K., and Salomon P., *Neural network ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 12, pp. 993–1001, 1990.

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

[8] Woods K., Kegelmeyer Jr. W. P., and Bowyer K., *Combination of multiples classifiers using local accuracy estimates*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 19, pp. 405–410, 1997.

[9] Smets Ph., Kennes R., *The transferable belief model*, Artificial Intelligence, vol 66, pp. 191–234, 1994.

[10] Smets Ph., *The transferable belief model for quantified belief representation*, Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol 1, pp. 267–301, 1998.

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

[12] Savage L. J., *Foundation of statistics*, Wiley, New York, 1954.

[13] DeGroot M. H., *Optimal statistical decision*, McGraw-Hill, New York, 1970.

[14] Denœux T., *Analysis of evidence-theoretic decision rules for pattern classification*, Pattern Recognition, vol 30, pp. 1095–1107, 1997.

[15] Smets Ph., *Decision making in a context where uncertainty is represented by belief functions*, Belief functions in business decisions, T.J. (ed.) Physica-Verlag, pp. 17–61, 2002.

[16] Smets Ph., *Decision making in the TBM: the Necessity of the Pignistic Transformation*, International Journal of Approximate Reasoning, pp. 133–147, 2005.

[17] Rogova G., *Combining the results of several neural network classifiers*, Neural Networks, vol 7, pp. 777–781, 1994.

[18] Smets Ph., *Belief functions: the disjunctive rule of combination and the generalized bayesian theorem*, International Journal of Approximate Reasoning, vol 9, pp. 1–35, 1993.