Combinaison crédibiliste de classifieurs binaires

Combinaison crédibiliste de classifieurs binaires

Benjamin Quost[1], Thierry Denœux[2] et Marie-Hélène Masson[1,2]

[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

Manuscrit reçu le 9 mai 2006

Résumé et mots clés

Nous étudions dans cet article le problème de la combinaison de classifieurs binaires. Cette approche consiste à résoudre un problème de discrimination multi-classes, en combinant les solutions de sous-problèmes binaires ; nous nous intéressons aux stratégies opposant chaque classe à chaque autre, et chaque classe à toutes les autres. La combinaison est considérée ici du point de vue de la théorie de Dempster-Shafer : les sorties des classifieurs sont ainsi interprétées comme des fonctions de croyance, conditionnelles ou exprimées dans un cadre plus grossier que le cadre initial. Elles sont combinées en calculant une fonction de croyance consistante avec les informations disponibles. Les performances des deux approches sont comparées à celles d'autres méthodes et illustrées sur divers jeux de données.

Classification multi-classes, théorie de Dempster-Shafer, théorie des fonctions de croyance, classification supervisée, fusion de classifieurs.

Abstract and key words

The problem of binary classifier combination is adressed in this article. This approach consists in solving a multi-class classification problem by combining the solutions of binary sub-problems. We consider two strategies in which each class is opposed to each other, or to all others. The combination is considered from the point of view of the theory of evidence. The classifier outputs are interpreted either as conditional belief functions, or as belief functions expressed in a coarser frame. They are combined by computing a belief function that is consistent with the available information. The performances of the methods are compared with those of other techniques and illustrated on various datasets.

Polychotomous classification, Dempster-Shafer theory, Belief Functions Theory, Classification, Classifier Fusion.

1. Introduction

Dans un problème classique de reconnaissance des formes, on dispose d'un ensemble de P vecteurs d'apprentissage $\mathbf{x}_p$, $p \in \{1,\ldots,P\}$, associés à des étiquettes y_p à valeur dans $\Omega = \{\omega_1,\ldots,\omega_K\}$ indiquant leur classe d'appartenance. Un classifieur peut être entraîné à associer l'espace des vecteurs à l'espace Ω des classes, dans le but de prédire la classe de tout vecteur $\mathbf{x}$ inconnu. La complexité d'un classifieur doit être adaptée à celle du problème abordé : les problèmes avec de nombreuses classes, aux frontières non-linéaires, nécessitent généralement des classifieurs complexes, dont le coût d'apprentissage peut devenir important.

Les problèmes de classification multi-classes (pour lesquels $K > 2$) peuvent être résolus directement en considérant simultanément l'ensemble des K classes, qui sont autant d'hypothèses mutuellement exclusives sur l'appartenance de $\mathbf{x}$. Une alternative consiste à combiner plusieurs classifieurs binaires [14,8], ne considérant qu'un ensemble de deux classes mutuellement exclusives, défini à partir de Ω. Chaque classifieur ayant ainsi une connaissance partielle de Ω, leur combinaison a pour objectif de tirer parti de leur complémentarité pour construire un système global. Un problème binaire est plus facile à résoudre qu'un problème multi-classes ; certains algorithmes (comme la régression logistique, les arbres de décision ou les séparateurs à vaste marge) étant de plus bien adaptés à la résolution de problèmes binaires, la combinaison de classifieurs constitue une approche intéressante pour résoudre un problème multi-classes.

Dans cet article, nous proposons deux méthodes pour combiner des classifieurs binaires dans le cadre du Modèle des Croyances Transférables (MCT) [23], une interprétation de la théorie des fonctions de croyance. Ce cadre théorique permet de représenter la connaissance partielle et de modéliser différents types d'imprécision ; il semble donc adéquat pour formaliser la combinaison de classifieurs binaires.

Dans un premier temps, le problème de la combinaison de classifieurs est exposé au paragraphe 2, et plusieurs méthodes de combinaison existantes sont présentées. Puis le Modèle des Croyances Transférables est brièvement décrit au paragraphe 3. Deux méthodes de combinaison des classifieurs binaires sont alors proposées aux paragraphes 4 et 5 ; les performances de ces méthodes sont comparées à celles de méthodes existantes, et les résultats analysés, au paragraphe 7. Enfin, le paragraphe 8 conclut l'article.

2. Combinaison de classifieurs binaires

2.1. Différents schémas de décomposition d'un problème multi-classes

Il existe différentes manières de décomposer un problème multi-classes en sous-problèmes binaires. La décomposition un-contre-un ou 1-1 consiste à construire un ensemble de C_K^2 classifieurs $\mathcal{E}_{ij}$, définis par convention pour tout $j > i$ (pour chaque paire de classes). Le classifieur $\mathcal{E}_{ij}$, entraîné à séparer ω_i de ω_j, a une connaissance incomplète de Ω en ce sens qu'il ignore tout des classes ω_k, $k \notin \{i,j\}$. La décomposition un-contre-tous ou 1-T consiste à construire un ensemble de K classifieurs (un pour chaque classe). Le classifieur $\mathcal{E}_k$ est entraîné à séparer ω_k de l'ensemble des autres classes ; il a une connaissance incomplète de Ω en ce sens qu'il est incapable de discerner les classes ω_l, $l \neq k$. Les classifieurs sont moins nombreux dans le cas 1-T, mais peuvent être plus coûteux à entraîner. Une comparaison du coût d'apprentissage global de ces deux méthodes a été proposée par Fürnkranz [9].

Les codes correcteurs d'erreurs (CCE) [7,1] constituent une approche générale pour la décomposition d'un problème. Des classifieurs binaires $\mathcal{E}_i$ $(i = 1 \ldots N)$ sont entraînés à séparer deux ensembles de classes $A_i \subseteq \Omega$ et $B_i \subseteq \Omega$. Le code $\mathbf{m}_k$ d'une classe ω_k indique son rôle dans l'apprentissage des différents classifieurs : $m_{ki} = 1$ si $\omega_k \in A_i$, $m_{ki} = -1$ si $\omega_k \in B_i$, $m_{ki} = 0$ si $\omega_k \notin A_i \cup B_i$. Lorsqu'on évalue $\mathbf{x}$, un code $\mathbf{m}_\mathbf{x}$ est déterminé. On affecte $\mathbf{x}$ à la classe ω_k telle que $\mathbf{m}_k$ est le plus proche de $\mathbf{m}_\mathbf{x}$, au sens d'une mesure de similarité à définir. Remarquons que ce cadre généralise les décompositions 1-1 et 1-T présentées ci-dessus.

Dans cet article, nous nous intéressons aux approches 1-1 et 1-T.

2.2. Différentes stratégies de combinaison des classifieurs binaires

Il existe différentes méthodes de combinaison, selon la nature des sorties des classifieurs. Si les classifieurs binaires fournissent des estimations de probabilités conditionnelles d'appartenance aux classes, on peut les combiner pour calculer des estimations des probabilités a posteriori $p_i = \mathbb{P}(y = \omega_i|\mathbf{x})$. Par la suite, on notera $\mathbf{p}$ le vecteur des p_i. Pour des raisons de place, nous avons choisi de ne présenter ici que les méthodes de combinaison de tels classifieurs, dans le cas de décompositions 1-1 et 1-T. Nous avons ainsi omis les méthodes de combinaison ne permettant de déterminer que des décisions [1,7,8,14,18] et les méthodes proposées dans le cadre de la fusion de données multicapteurs [2,13].

2.2.1. Combinaison dans le cas 1-1

Soit n_k le nombre de vecteurs de la classe ω_k, et $n_{ij} = n_i + n_j$. Chaque classifieur $\mathcal{E}_{ij}$ fournit une estimation r_{ij} de la probabilité conditionnelle $\mu_{ij} = \mathbb{P}(\omega_i|\omega_i \text{ ou } \omega_j,\mathbf{x})$; on peut donc estimer les p_i, en exploitant les relations $\mu_{ij} = p_i/(p_i + p_j)$. Le calcul direct des p_i vérifiant $0 \leqslant p_i \leqslant 1$, $\sum_i p_i = 1$, $\mu_{ij} = r_{ij}$ pour tout $j > i$, est un problème surdéterminé à $K-1$ inconnues pour C_K^2 contraintes d'égalité, qui n'a généralement pas de solution exacte. On peut alors estimer les p_i, telles que les μ_{ij} soient proches des r_{ij} au sens d'un critère d'erreur.

La méthode exposée dans [10] consiste à calculer des estimations $\widehat{p_i}$ des p_i, en minimisant la distance de Kullback-Leibler négative pondérée $\mathcal{L}$ entre les μ_{ij} et les r_{ij} fournies par les classifieurs, par une procédure itérative de descente de gradient. Soit $\widehat{\mathbf{p}}$ le vecteur des $\widehat{p_i}$:

$$\widehat{\mathbf{p}} = \arg\min \mathcal{L}(\mathbf{p}), \tag{1}$$

sous les contraintes :

$$\sum_i p_i = 1, \tag{2}$$

$$p_i \geqslant 0, \quad \text{pour tout } i \in \{1,\ldots,K\}; \tag{3}$$

où le critère $\mathcal{L}$ est défini par :

$$\mathcal{L}(\mathbf{p}) = \sum_{i<j} n_{ij} \left(r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1-r_{ij}) \log \frac{1-r_{ij}}{1-\mu_{ij}} \right). \tag{4}$$

Cette méthode sera appelée par la suite méthode PCpl. Hastie et Tibshirani remarquent que les p_i peuvent être écrits sous la forme :

$$p_i = \sum_{j \neq i} \left(\frac{p_i + p_j}{K-1} \right) \left(\frac{p_i}{p_i + p_j} \right). \tag{5}$$

En remplaçant $p_i + p_j$ par $2/K$ dans le premier terme et chacun des seconds termes par les r_{ij} correspondants, on obtient des estimations simples des p_i, notées $p_i^\star$, à partir des r_{ij} :

$$p_i^\star = \frac{2}{K(K-1)} \sum_{j \neq i} r_{ij}. \tag{6}$$

Il est montré dans [10] que les $p_i^\star$ sont dans le même ordre que les $\widehat{p_i}$: pour tout (i,j), $p_i^\star \geqslant p_j^\star$ si et seulement si $\widehat{p_i} \geqslant \widehat{p_j}$. Ils peuvent donc servir de valeurs de départ pour la procédure de minimisation itérative, ou lorsque seules des décisions sont requises.

Dans [24], deux méthodes non-itératives sont proposées pour estimer les p_i en minimisant l'écart entre les probabilités conditionnelles μ_{ij} et les estimations r_{ij}. La première (qui sera appelée par la suite méthode PEst1) consiste à ne remplacer que μ_{ij} par r_{ij} dans l'équation (5). La théorie des chaînes de Markov permet de montrer que la solution des équations ainsi obtenues, auxquelles on adjoint les contraintes (3)-(2), est l'unique solution d'un système linéaire. La seconde méthode (appelée par la suite méthode PEst2) consiste à remplacer la fonction de coût (1)-(4) du problème d'optimisation par :

$$\widehat{\mathbf{p}} = \arg\min_{\mathbf{p}} \frac{1}{2} \sum_{i=1}^{K} \sum_{j:j\neq i} \left(r_{ji} p_i - r_{ij} p_j \right)^2. \tag{7}$$

Les auteurs montrent que les contraintes de positivité (3) sont implicitement vérifiées par la solution de (7)-(2) ; le problème se réduit donc à un problème quadratique convexe avec une contrainte d'égalité linéaire. D'après les conditions d'optimalité de Kuhn et Tucker, le minimum global $\widehat{\mathbf{p}}$ de (7)-(2) peut donc être déterminé en résolvant un système linéaire.

Comme cela est souligné dans [10], un classifieur 1-1 peut fournir des estimations r_{ij} erronées, s'il n'a pas été entraîné à reconnaître la classe réelle du vecteur $\mathbf{x}$ évalué. Dans [16], il est donc proposé d'entraîner des classifieurs supplémentaires, dits correcteurs, à séparer les classes $\{\omega_i,\omega_j\}$ de l'ensemble des autres, et donc à calculer des estimations q_{ij} des probabilités $\mathbb{P}(\{\omega_i,\omega_j\}|\mathbf{x})$. En remplaçant dans (5) le numérateur $p_i + p_j$ par q_{ij}, on obtient des estimations $p_i^\dagger$ des p_i :

$$p_i^\dagger = \frac{1}{K-1} \sum_{j>i} r_{ij} q_{ij}. \tag{8}$$

Bien que susceptible d'améliorer la précision du processus de classification, cette méthode (appelée par la suite méthode PEstCorr) nécessite d'entraîner C_K^2 classifieurs supplémentaires sur tout l'ensemble d'apprentissage, ce qui peut être coûteux lorsque K est grand.

2.2.2. Combinaison dans le cas 1-T

Dans [25] et [2], la méthode PCPL est étendue au cas des codes correcteurs d'erreurs : chaque classifieur $\mathcal{E}_i$ $(i = 1 \ldots N)$, entraîné à séparer deux groupes de classes A_i et B_i, fournit une estimation r_i de la probabilité conditionnelle $\mathbb{P}(A_i|A_i \cup B_i,\mathbf{x})$. Une méthode est proposée dans [2] pour calculer $\widehat{\mathbf{p}}$, en minimisant la distance de Kullback-Leibler entre les probabilités conditionnelles $\mu_i = p_{A_i}/(p_{A_i} + p_{B_i})$ et les r_i par une procédure itérative. Dans le cas 1-T, les auteurs proposent une méthode non-itérative pour déterminer $\widehat{\mathbf{p}}$:

- si $\sum_{i=1}^K r_i = 1$, alors $\widehat{p_k} = r_k$ pour tout $k \in \{1,\ldots,K\}$;
- sinon, il faut trouver la solution $\widehat{\delta}$ de l'équation non-linéaire suivante :

$$\sum_{k=1}^{K} \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4r_k\delta}}{2\delta} - 1 = 0,$$

et la solution optimale est définie, pour tout $k \in \{1,\ldots,K\}$, par :

$$\widehat{p_k} = \frac{(1+\widehat{\delta}) - \sqrt{(1+\widehat{\delta})^2 - 4r_k\widehat{\delta}}}{2\widehat{\delta}}. \tag{9}$$

Cette méthode, appliquée au cas 1-T, sera appelée par la suite méthode PCpl1-T.

Dans [17], il est proposé de calculer les p_k ($k \in \{1,\ldots,K\}$) à partir des estimations r_i ($i \in \{1,\ldots,N\}$) fournies par les classifieurs binaires associés à une décomposition par codes correcteurs d'erreurs, par une procédure non-itérative. Dans le cas particulier d'une décomposition 1-T, la probabilité a posteriori p_k de chaque classe ω_k peut être estimée par :

$$p_k^\star = r_k \prod_{\substack{l=1 \\ l \neq k}}^{K} (1 - r_l) + \alpha, \tag{10}$$

où α est une constante garantissant $\sum_k p_k = 1$. Le vecteur $\mathbf{x}$ est affecté à la classe de probabilité maximale. Cette méthode, appliquée au cas 1-T, sera appelée par la suite méthode PEst1-T.

3. Le Modèle des Croyances Transférables (MCT)

3.1. Quantification et manipulation des connaissances

La nécessité de distinguer divers types d'ignorance est à l'origine de plusieurs formalismes de représentation des connaissances. La théorie de Dempster-Shafer ou théorie des fonctions de croyance [22] permet ainsi de représenter et manipuler diverses formes de connaissance partielle. Le Modèle des Croyances Transférables (MCT) [23] est une interprétation subjectiviste de cette théorie, dans laquelle une fonction de croyance modélise la connaissance partielle de la valeur d'une variable y. Le MCT semble donc adéquat pour modéliser les sorties de classifieurs à combiner, qui n'ont qu'une connaissance incomplète de l'ensemble Ω.

La variable y est définie sur un cadre de discernement ou domaine $\Omega = \{\omega_1,\ldots,\omega_K\}$. En classification, y correspond à la classe du vecteur $\mathbf{x}$ évalué. La croyance relative à la valeur de y peut être quantifiée par une fonction de masse de croyance (FMC) $m^\Omega : 2^\Omega \rightarrow [0;1]$, qui vérifie :

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{11}$$

La quantité $m^\Omega(A)$ est interprétée comme une masse de croyance allouée spécifiquement à l'hypothèse A, sur la base d'un élément d'évidence. Tout sous-ensemble $A \subseteq \Omega$ tel que $m^\Omega(A) > 0$ est appelé élément focal de m^Ω. L'exposant Ω peut être omis lorsqu'il n'y a aucune ambiguïté sur le domaine de m^Ω. Une fonction de croyance catégorique n'a qu'un élément focal : $m(A) = 1$; pour $A = \Omega$, on obtient la fonction de croyance vide, modélisant l'ignorance totale. Les fonctions de masse Bayésiennes n'ont que des éléments focaux singletons.

Si le domaine Ω est considéré comme exhaustif (hypothèse du monde clos), l'hypothèse de normalité $m(\emptyset) = 0$ est généralement acceptée et m est alors dite normale. Dans le cas contraire (hypothèse du monde ouvert), m est dite sous-normale : la masse $m(\emptyset)$ est alors interprétée comme la croyance que $y \notin \Omega$. Une FMC sous-normale peut être transformée en FMC normale en divisant chaque masse $m(A)$ ($A \neq \emptyset$) par $1 - m(\emptyset)$; le résultat de cette normalisation est alors noté m^*.

Le conditionnement $m[B]$ de m sur un sous-ensemble $B \subseteq \Omega$ peut être calculé par :

$$m[B](A) = \sum_{C \cap B = A} m(C) \quad \text{si } A \subseteq B. \tag{12}$$

Ainsi, toute masse de croyance initialement associée à $C \subseteq \Omega$ est transférée à $C \cap B$. La FMC $m[B]$ quantifie la croyance concernant la valeur de y, en supposant que $y \in B$; la masse $m[B](\emptyset)$ représente la part de croyance donnée par m aux hypothèses incompatibles avec B. L'opération définie par (12) correspond à la règle de conditionnement non normalisée. Sa version normalisée (correspondant à l'hypothèse du monde clos) est obtenue en ajoutant une étape de normalisation :

$$m^\Omega[B]^*(A) = \begin{cases} \dfrac{m[B](A)}{1 - m[B](\emptyset)} & \text{si } A \neq \emptyset, \\ 0 & \text{si } A = \emptyset. \end{cases} \tag{13}$$

Les fonctions de croyance *bel* et de plausibilité *pl* peuvent être calculées à partir de m. On a, pour tout $A \subseteq \Omega$:

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B), \tag{14}$$

$$pl(A) = \sum_{A \cap B \neq \emptyset} m(B). \tag{15}$$

La quantité $bel(A)$ représente le degré total de croyance en A, au regard des informations disponibles ; la quantité $pl(A)$ est interprétée comme une borne supérieure sur le degré de croyance qui pourrait être alloué à A après conditionnement : on a en effet $pl(A) = bel[A](A) = \max_B bel[B](A)$.

Deux fonctions de croyance m_1 et m_2 définies sur le même domaine Ω, induites par des informations distinctes, peuvent être combinées par la règle de combinaison conjonctive, notée $\textcircled{\cap}$:

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

Dans le MCT, on distingue deux niveaux : le *niveau crédal* où les croyances sont représentées et combinées au moyen de fonctions de croyance, et le *niveau pignistique*, où les fonctions de croyance sont transformées en probabilités pignistiques pour prendre une décision. La transformation pignistique [23] consiste à partager équitablement chaque masse de croyance normalisée $m^*(A)$:

$$Bet\,P^*(\omega) = \sum_{A \subseteq \Omega : \omega \in A} \frac{m^*(A)}{|A|}, \quad \forall \omega \in \Omega. \tag{16}$$

3.2. Grossissements et raffinements

Soient deux cadres Ω et Θ et une application $\rho : 2^{\Theta} \rightarrow 2^{\Omega}$ telle que :
1. l'ensemble $\{\rho(\{\theta\}), \theta \in \Theta\} \subseteq 2^{\Omega}$ est une partition de Ω ;
2. pour chaque $A \subseteq \Theta$, $\rho(A) = \bigcup_{\theta \in A} \rho(\{\theta\})$.
L'application ρ est alors appelée raffinement de Θ vers Ω ; par extension, Ω est appelé un raffinement de Θ, et Θ un grossissement de Ω [22].
L'application ρ n'est généralement pas surjective : il peut exister des $B \subseteq \Omega$ qui ne sont pas image par ρ d'un $A \subseteq \Theta$. Il y a donc plusieurs manières d'associer un tel $B \subseteq \Omega$ à un sous-ensemble de Θ ; deux d'entre elles présentent un intérêt particulier [5]. L'élément B pourrait ainsi être associé au plus grand élément $\underline{\theta}(B) \subseteq \Theta$ dont l'image par ρ est incluse dans B. Formellement,

$$\underline{\theta}(B) = \{\theta \in \Theta \, : \, \rho(\{\theta\}) \subseteq B\}.$$

Le sous-ensemble $\underline{\theta}(B)$ est appelé *réduction intérieure* de B sur Θ. La réduction intérieure de B sur Ω est définie par $\rho(\underline{\theta}(B))$. Par ailleurs, B pourrait aussi être associé au plus petit élément $\overline{\theta}(B) \subseteq \Theta$ dont l'image par ρ inclut B. Formellement,

$$\overline{\theta}(B) = \{\theta \in \Theta \, : \, \rho(\{\theta\}) \cap B \neq \emptyset\}.$$

Le sous-ensemble $\overline{\theta}(B)$ est appelé *réduction extérieure* de B sur Θ. La réduction extérieure de B sur Ω est définie par $\rho(\overline{\theta}(B))$.

Exemple 1 : Soit $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5\}$; soit $\Theta = \{\theta_1,\theta_2\}$ un grossissement de Ω défini par $\rho(\{\theta_1\}) = \{\omega_1\}$, $\rho(\{\theta_2\}) = \{\omega_2,\omega_3,\omega_4,\omega_5\}$. Le tableau 1 montre les réductions intérieures et extérieures de certains $A \subseteq \Omega$. □

Tableau 1. Réductions intérieures et extérieures, pour certains $A \subseteq \Omega$.

$A \subseteq \Omega$	$\emptyset$	$\{\omega_1\}$	$\{\omega_1,\omega_3\}$	$\{\omega_2,\omega_4\}$	$\{\omega_2,\ldots,\omega_5\}$	Ω
$\underline{\theta}(A)$	$\emptyset$	$\{\theta_1\}$	$\{\theta_1\}$	$\emptyset$	$\{\theta_2\}$	Θ
$\rho(\underline{\theta}(A))$	$\emptyset$	$\{\omega_1\}$	$\{\omega_1\}$	$\emptyset$	$\{\omega_2,\ldots,\omega_5\}$	Ω
$\overline{\theta}(A)$	$\emptyset$	$\{\theta_1\}$	Θ	$\{\theta_2\}$	$\{\theta_2\}$	Θ
$\rho(\overline{\theta}(A))$	$\emptyset$	$\{\omega_1\}$	Ω	$\{\omega_2,\ldots,\omega_5\}$	$\{\omega_2,\ldots,\omega_5\}$	Ω

Ces définitions peuvent être aisément étendues aux fonctions de masse ; les réductions intérieure et extérieure d'une FMC m^{Ω} sur Θ sont définies respectivement par :

$$\underline{m}^{\Theta}(A) = \underline{\theta}(m^{\Omega}) = \sum_{B \subseteq \Omega, \underline{\theta}(B) = A} m^{\Omega}(B), \quad \forall A \subseteq \Theta ;$$

$$\overline{m}^{\Theta}(A) = \overline{\theta}(m^{\Omega}) = \sum_{B \subseteq \Omega, \overline{\theta}(B) = A} m^{\Omega}(B), \quad \forall A \subseteq \Theta.$$

La réduction d'une FMC m^{Ω} sur Θ s'accompagne généralement d'une perte d'information.

Pour transférer une FMC m^{Θ} sur Ω, on utilise généralement l'extension vide [22], définie pour tout $B \subseteq \Omega$ par :

$$m^{\Theta\uparrow\Omega}(B) = \begin{cases} m^{\Theta}(A) & \text{si } B = \rho(A), A \subseteq \Theta \\ 0 & \text{sinon.} \end{cases} \tag{17}$$

Par commodité, l'extension vide $m^{\Theta\uparrow\Omega}$ d'une FMC m^{Θ} sur Ω sera notée m^{Ω}.

Exemple 2 : Soient Ω et Θ les cadres définis dans l'exemple 1. Soit m^{Ω} la FMC définie par :

$$m^{\Omega}(\emptyset) = 0.1 \qquad m^{\Omega}(\{\omega_1\}) = 0.3$$
$$m^{\Omega}(\{\omega_1,\omega_3\}) = 0.2 \qquad m^{\Omega}(\{\omega_2,\omega_4\}) = 0.1$$
$$m^{\Omega}(\{\omega_2,\ldots,\omega_5\}) = 0.2 \qquad m^{\Omega}(\Omega) = 0.1$$

Le tableau 2 montre les réductions intérieure et extérieure de m^{Ω} sur Θ et sur Ω. □

Tableau 2. Réductions intérieure et extérieure de m^{Ω} sur Θ (haut) et sur Ω (bas).

$A \subseteq \Theta$	$\emptyset$	$\{\theta_1\}$	$\{\theta_2\}$	Θ
$\underline{m}^{\Theta}(A)$	0.2	0.5	0.2	0.1
$\overline{m}^{\Theta}(A)$	0.1	0.3	0.3	0.3

$B \subseteq \Omega$	$\emptyset$	$\{\omega_1\}$	$\{\omega_2,\ldots,\omega_5\}$	Ω
$\underline{m}^{\Omega}(B)$	0.2	0.5	0.2	0.1
$\overline{m}^{\Omega}(B)$	0.1	0.3	0.3	0.3

4. Combinaison de classifieurs 1-1 dans le cadre du MCT

4.1. Les sorties des classifieurs vues comme des fonctions de croyance conditionnelles normales

Soit $\{\mathcal{E}_{ij}\}$ un ensemble de classifieurs binaires obtenu pour une décomposition 1-1 de Ω. Chaque classifieur $\mathcal{E}_{ij}$ a appris à séparer la classe ω_i de la classe ω_j ; par conséquent, l'information qu'il fournit est conditionnelle à l'appartenance de $\mathbf{x}$ au domaine restreint $\Omega_{ij} = \{\omega_i,\omega_j\}$. De plus, ce cadre est supposé exhaustif : en effet, peu d'algorithmes intègrent une phase de détection de nouveauté, qui permettrait ici de déterminer si $\mathbf{x}$ appartient ou non à l'une des classes que $\mathcal{E}_{ij}$ sait reconnaître.
Soit m^*_{ij} la FMC normale, définie sur le cadre de discernement Ω_{ij}, fournie par $\mathcal{E}_{ij}$. Cette FMC peut être considérée comme le résultat du conditionnement d'une FMC inconnue m^{Ω} sur Ω_{ij},

au moyen de la règle de conditionnement normalisée (13) : pour tout $A \subseteq \Omega_{ij}$, $m^*_{ij}(A) = m[\Omega_{ij}]^*(A)$. De manière équivalente,

$$m^\Omega[\Omega_{ij}](A) = m^*_{ij}(A)\left(1 - m^\Omega[\Omega_{ij}](\emptyset)\right), \quad \forall\, A \subseteq \Omega_{ij}, A \neq \emptyset, \forall\, j > i, \tag{18}$$

où $m[\Omega_{ij}]$ est la FMC conditionnelle non normalisée, obtenue à partir de m^Ω au moyen de (12). Remarquons que l'opérateur de conditionnement est une application linéaire ; chaque terme $m[\Omega_{ij}](A)$ $(A \subseteq \Omega_{ij})$ est donc une combinaison linéaire de masses $m^\Omega(C)$ $(C \subseteq \Omega)$. Tout domaine Ω_{ij} comptant trois sous-ensembles non vides ($\{\omega_i\}$, $\{\omega_j\}$, et Ω_{ij}), la relation (18) définit un système de $3 \times C_K^2$ équations linéaires et $2^K - 1$ inconnues (le cadre Ω comptant 2^K sous-ensembles A, liés entre eux par la contrainte (11)). Remarquons que la FMC catégorique définie par $m(\emptyset) = 1$ est solution triviale de (18) : elle vérifie $m[\Omega_{ij}](\emptyset) = 1$ et $m[\Omega_{ij}](A) = 0$ pour tout $A \neq \emptyset$, pour tout $j > i$. Cette solution peut être écartée en introduisant une contrainte supplémentaire de normalité :

$$m^\Omega(\emptyset) = 0. \tag{19}$$

Le système défini par (18)-(19) n'a souvent pas de solution ; les FMCs m^*_{ij}, étant calculées par différents classifieurs binaires, ne sont en général pas consistantes : il n'existe pas de FMC normale m^* dont les conditionnements normalisés sur les Ω_{ij} correspondent exactement aux m_{ij}, pour tout $j > i$. En outre, ces FMCs ne sont pas distinctes, les classifieurs binaires partageant certaines données d'apprentissage ; elles ne peuvent donc être combinées par la règle de combinaison conjonctive. Une solution approchée du système peut néanmoins être calculée en résolvant un problème d'optimisation quadratique :

$$\widehat{m}^* = \arg\min_{m^\Omega} \sum_{\Omega_{ij} \subseteq \Omega} \sum_{\substack{A \subseteq \Omega_{ij} \\ A \neq \emptyset}} \left[m^\Omega[\Omega_{ij}](A) - m^*_{ij}(A)\left(1 - m^\Omega[\Omega_{ij}](\emptyset)\right)\right]^2, \tag{20}$$

sous les contraintes :

$$m^\Omega(A) \geqslant 0, \quad \forall A \subseteq \Omega,\ A \neq \emptyset, \tag{21}$$

$$m^\Omega(\emptyset) = 0, \tag{22}$$

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{23}$$

Ce problème peut être résolu au moyen d'un algorithme usuel d'optimisation. La méthode de combinaison ainsi définie sera appelée par la suite MCT1-1.

Exemple 3 : Soit un problème à $K = 3$ classes : $\Omega = \{\omega_1, \omega_2, \omega_3\}$. On dispose donc de trois classifieurs binaires $\mathcal{E}_{12}$, $\mathcal{E}_{13}$ et $\mathcal{E}_{23}$. Supposons que, lors de l'évaluation d'un vecteur $\mathbf{x}$, ces classifieurs ont fourni les FMCs suivantes :

- Classifieur $\mathcal{E}_{12}$: $m^*_{12}(\{\omega_1\}) = 0.439$, $m^*_{12}(\{\omega_2\}) = 0.498$, $m^*_{12}(\Omega_{12}) = 0.063$;
- Classifieur $\mathcal{E}_{13}$: $m^*_{13}(\{\omega_1\}) = 0.849$, $m^*_{13}(\{\omega_3\}) = 0.014$, $m^*_{13}(\Omega_{13}) = 0.137$;
- Classifieur $\mathcal{E}_{23}$: $m^*_{23}(\{\omega_2\}) = 0.666$, $m^*_{23}(\{\omega_3\}) = 0.010$, $m^*_{23}(\Omega_{23}) = 0.324$.

En utilisant la méthode de combinaison MCT1-1, on obtient la FMC $\widehat{m}^\Omega$ suivante :

$$\widehat{m}^*(\{\omega_1\}) = 0.448, \qquad \widehat{m}^*(\{\omega_2\}) = 0.414,$$

$$\widehat{m}^*(\{\omega_2, \omega_3\}) = 0.056, \quad \widehat{m}^*(\Omega) = 0.082.$$

En conditionnant $\widehat{m}^*$ sur chaque Ω_{ij} par la règle normalisée (13), on trouve :
- $\widehat{m}^*[\Omega_{12}]^*(\{\omega_1\}) = 0.448$, $\widehat{m}^*[\Omega_{12}]^*(\{\omega_2\}) = 0.470$, $\widehat{m}^*[\Omega_{12}]^*(\Omega_{12}) = 0.082$;
- $\widehat{m}^*[\Omega_{13}]^*(\{\omega_1\}) = 0.766$, $\widehat{m}^*[\Omega_{13}]^*(\{\omega_3\}) = 0.095$, $\widehat{m}^*[\Omega_{13}]^*(\Omega_{13}) = 0.139$;
- $\widehat{m}^*[\Omega_{23}]^*(\{\omega_2\}) = 0.751$, $\widehat{m}^*[\Omega_{23}]^*(\{\omega_3\}) = 0.000$, $\widehat{m}^*[\Omega_{23}]^*(\Omega_{23}) = 0.249$,

qui sont les meilleures approximations de m^*_{12}, m^*_{13} et m^*_{23} au sens du critère (20). □

4.2. Estimation de la pertinence des classifieurs

La méthode MCT1-1 permet de calculer une FMC m^Ω la plus consistante possible avec les estimations fournies par les classifieurs. Cependant, de même que pour les méthodes PCpl, PEst1 et PEst2 (paragraphe 2.2.1), la FMC m^*_{ij} fournie par un classifieur $\mathcal{E}_{ij}$ sans capacité de détection de nouveauté peut être erronée si le vecteur $\mathbf{x}$ n'appartient ni à ω_i ni à ω_j. Nous proposons par conséquent d'estimer la masse $m[\Omega_{ij}](\emptyset)$: une valeur élevée de $m[\Omega_{ij}](\emptyset)$ indique que $\mathbf{x}$ n'appartient vraisemblablement pas à Ω_{ij} et donc que le classifieur $\mathcal{E}_{ij}$ n'est pas apte à le classer. Remarquons que si $m[\Omega_{ij}](\emptyset)$ est élevée pour tout $i < j$, on peut penser que $\mathbf{x}$ n'appartient pas à Ω, et la solution du problème vérifiera $m^\Omega(\emptyset) \approx 1$; la contrainte de normalité (19) doit donc être abandonnée. Cette information additionnelle joue un rôle similaire à celui des classifieurs correcteurs [16] définis dans un cadre probabiliste. Tandis que cette méthode prévoit d'entraîner un classifieur correcteur à séparer chaque paire $\{\omega_i, \omega_j\}$ de l'ensemble des autres classes pour estimer la probabilité que $\mathbf{x}$ appartienne à ω_i ou ω_j, nous proposons de déterminer la *plausibilité* que $\mathbf{x}$ appartienne à ω_i ou ω_j. Cette information est directement reliée à la croyance $m[\Omega_{ij}](\emptyset)$ que $\mathbf{x}$ n'appartient pas à Ω_{ij} : en effet,

$$m[\Omega_{ij}](\emptyset) = \sum_{A \cap \Omega_{ij} = \emptyset} m^\Omega(A) \tag{24}$$

$$= 1 - \sum_{A \cap \Omega_{ij} \neq \emptyset} m^\Omega(A) \tag{25}$$

$$= 1 - pl^\Omega(\Omega_{ij}). \tag{26}$$

Soient $pl_{ij} = pl^{\Omega}(\Omega_{ij})$ supposée connue, et m_{ij} la FMC obtenue en « dénormalisant » la FMC m^*_{ij} fournie par le classifieur $\mathcal{E}_{ij}$:

$$m_{ij}(A) = pl_{ij} m^*_{ij}(A), \quad \text{pour tout } A \subseteq \Omega_{ij}, A \neq \emptyset ; \tag{27}$$

$$m_{ij}(\emptyset) = 1 - pl_{ij}. \tag{28}$$

En substituant $m_{ij}(A)$ à $m^*_{ij}(A)(1 - m[\Omega_{ij}](\emptyset))$ dans (20), et en abandonnant la contrainte de normalité (19), on obtient un problème d'optimisation permettant d'approcher m^{Ω} par la FMC la plus consistante possible avec les FMCs dénormalisées m_{ij} [19] :

$$\widehat{m}^{\Omega} = \arg\min_{m^{\Omega}} \sum_{\Omega_{ij} \subseteq \Omega} \sum_{A \subseteq \Omega_{ij}} \left[m^{\Omega}[\Omega_{ij}](A) - m_{ij}(A) \right]^2, \tag{29}$$

sous les contraintes :

$$m^{\Omega}(A) \geqslant 0, \quad \forall A \subseteq \Omega, \tag{30}$$

$$\sum_{A \subseteq \Omega} m^{\Omega}(A) = 1. \tag{31}$$

Cette méthode sera appelée par la suite MCTCorr1-1.
Le calcul des plausibilités pl_{ij} au moyen de classifieurs correcteurs séparant Ω_{ij} des autres classes, de manière similaire à [16], nécessiterait d'entraîner C^2_K classifieurs additionnels, chacun à partir des n vecteurs de l'ensemble d'apprentissage. Nous préférons déterminer les pl_{ij} en utilisant des classifieurs à une classe. Pour chaque classe $\omega_i \in \Omega$, un tel classifieur peut être entraîné à évaluer la plausibilité $pl^{\Omega}(\{\omega_i\})$ que $\mathbf{x}$ appartienne à ω_i (le paragraphe 7.1 détaille comment cela peut être fait au moyen de séparateurs à vaste marge à une classe). On propose ensuite de déterminer pl_{ij} en combinant les $pl^{\Omega}(\{\omega_k\})$, au moyen d'une conorme triangulaire telle que la t-conorme probabiliste :

$$pl_{ij} = pl^{\Omega}(\{\omega_i\}) + pl^{\Omega}(\{\omega_j\}) - pl^{\Omega}(\{\omega_i\}) pl^{\Omega}(\{\omega_j\}). \tag{32}$$

Exemple 4 : Revenons à l'exemple 3, et supposons que des classifieurs à une classe permettent à présent d'estimer les plausibilités d'appartenance aux classes :

$$pl^{\Omega}(\{\omega_1\}) = 0.555, \quad pl^{\Omega}(\{\omega_2\}) = 0.418, \quad pl^{\Omega}(\{\omega_3\}) = 0.$$

En utilisant (32), on obtient les plausibilités suivantes pour les paires de classes $\{\omega_i, \omega_j\}$:

$$pl_{12} = 0.741, \quad pl_{13} = 0.555, \quad pl_{23} = 0.418.$$

La dénormalisation des FMCs m^*_{ij} fournies par les classifieurs binaires (relations (27)-(28)) donne :
- $m_{12}(\emptyset) = 0.259$, $m_{12}(\{\omega_1\}) = 0.325$, $m_{12}(\{\omega_2\}) = 0.369$, $m_{12}(\Omega_{12}) = 0.047$;
- $m_{13}(\emptyset) = 0.445$, $m_{13}(\{\omega_1\}) = 0.471$, $m_{13}(\{\omega_3\}) = 0.008$, $m_{13}(\Omega_{13}) = 0.076$;
- $m_{23}(\emptyset) = 0.582$, $m_{23}(\{\omega_2\}) = 0.279$, $m_{23}(\{\omega_3\}) = 0.004$, $m_{23}(\Omega_{23}) = 0.135$.

La combinaison de m_{12}, m_{13} et m_{23} par la méthode MCTCorr1-1 donne la FMC suivante :

$$m^{\Omega}(\emptyset) = 0.218 \qquad m^{\Omega}(\{\omega_1\}) = 0.372,$$
$$m^{\Omega}(\{\omega_2\}) = 0.254, \qquad m^{\Omega}(\{\omega_1,\omega_2\}) = 0.039,$$
$$m^{\Omega}(\{\omega_2,\omega_3\}) = 0.068, \quad m^{\Omega}(\Omega) = 0.049,$$

les autres masses étant nulles. En conditionnant $\widehat{m}^{\Omega}$ sur chaque Ω_{ij} par la règle non-normalisée (12), on obtient :
- $\widehat{m}^{\Omega}[\Omega_{12}](\emptyset) = 0.218$, $\widehat{m}^{\Omega}[\Omega_{12}](\{\omega_1\}) = 0.372$, $\widehat{m}^{\Omega}[\Omega_{12}](\{\omega_2\}) = 0.322$, $\widehat{m}^{\Omega}[\Omega_{12}](\Omega_{12}) = 0.088$;
- $\widehat{m}^{\Omega}[\Omega_{13}](\emptyset) = 0.472$, $\widehat{m}[\Omega_{13}](\{\omega_1\}) = 0.411$, $\widehat{m}[\Omega_{13}](\{\omega_3\}) = 0.068$, $\widehat{m}[\Omega_{13}](\Omega_{13}) = 0.049$;
- $\widehat{m}^{\Omega}[\Omega_{23}](\emptyset) = 0.590$, $\widehat{m}[\Omega_{23}](\{\omega_2\}) = 0.293$, $\widehat{m}[\Omega_{23}](\{\omega_3\}) = 0$, $\widehat{m}[\Omega_{23}](\Omega_{23}) = 0.117$,

qui sont les meilleures approximations de m_{12}, m_{13} et m_{23} au sens du critère (29). □

4.3. Variante de la méthode dans le cas de classifieurs probabilistes

Un classifieur binaire probabiliste fournit une distribution de probabilité définie sur le domaine restreint Ω_{ij}. On peut l'interpréter soit comme une FMC Bayésienne, soit comme une estimation de la distribution de probabilité pignistique $BetP^*_{ij}$ associée à $m[\Omega_{ij}]$, définie par (16). On peut aussi définir $BetP^*_{ij}$ par :

$$BetP^*_{ij}(\omega_i) = \frac{BetP_{ij}(\omega_i)}{1 - m[\Omega_{ij}](\emptyset)} \tag{33}$$

$$BetP^*_{ij}(\omega_j) = \frac{BetP_{ij}(\omega_j)}{1 - m[\Omega_{ij}](\emptyset)}, \tag{34}$$

où $BetP_{ij}$ peut être vue comme la « probabilité pignistique non normalisée » calculée à partir de $m[\Omega_{ij}]$:

$$BetP_{ij}(\omega_i) = m[\Omega_{ij}](\{\omega_i\}) + \frac{m[\Omega_{ij}](\Omega_{ij})}{2} \tag{35}$$

$$BetP_{ij}(\omega_j) = m[\Omega_{ij}](\{\omega_j\}) + \frac{m[\Omega_{ij}](\Omega_{ij})}{2}. \tag{36}$$

Remarquons que $BetP_{ij}$ n'est pas une distribution de probabilité : en effet, $BetP_{ij}(\omega_i) + BetP_{ij}(\omega_j) < 1$ lorsque $m[\Omega_{ij}](\emptyset) > 0$.
Soit r^*_{ij} la probabilité de la classe ω_i fournie par le classifieur $\mathcal{E}_{ij}$, et $r^*_{ji} = 1 - r^*_{ij}$ (les notations du paragraphe 2.2.1 ont été adaptées pour souligner le fait que les sorties d'un classifieur probabiliste sont normalisées). On a donc :

$$r^*_{ij} = BetP^*_{ij}(\omega_i) \tag{37}$$

$$r^*_{ji} = BetP^*_{ij}(\omega_j), \tag{38}$$

ou encore :

$$r_{ij}^{*}(1-m[\Omega_{ij}](\emptyset)) = BetP_{ij}(\omega_i) \tag{39}$$

$$r_{ji}^{*}(1-m[\Omega_{ij}](\emptyset)) = BetP_{ij}(\omega_j). \tag{40}$$

De même que précédemment, ce système n'a pas de solution, les informations n'étant généralement pas consistantes ; on peut approcher m^{Ω} par la FMC la plus consistante possible avec les informations disponibles, en résolvant :

$$\widehat{m}^{\Omega} = \arg\min_{m^{\Omega}} \sum_{\Omega_{ij}\subseteq\Omega} \big[(BetP_{ij}(\omega_i) - r_{ij})^2 + (BetP_{ij}(\omega_j) - r_{ji})^2 + (m^{\Omega}[\Omega_{ij}](\emptyset) - 1 + pl_{ij})^2\big], \tag{41}$$

sous les contraintes (30) et (31). La méthode de combinaison définie par les relations (41),(30)-(31) sera appelée par la suite MCTProb1-1.

Exemple 5 : Revenons à l'exemple 3, et supposons que nous avons à présent des classifieurs probabilistes, qui donnent les estimations suivantes des probabilités pignistiques : $r_{12}^{*} = 0.843$, $r_{13}^{*} = 0.810$ et $r_{23}^{*} = 0.665$.
Supposons en outre que des classifieurs à une classe ont fourni les estimations suivantes d'appartenance aux classes :

$$pl^{\Omega}(\{\omega_1\}) = 1, \quad pl^{\Omega}(\{\omega_2\}) = 0.444, \quad pl^{\Omega}(\{\omega_3\}) = 0.624.$$

En utilisant (32), on obtient les plausibilités suivantes pour les paires de classes :

$$pl_{12} = 1, \quad pl_{13} = 1, \quad pl_{23} = 0.791,$$

desquelles nous déduisons $r_{12} = r_{12}^{*}$, $r_{21} = 1 - r_{12}^{*}$, $r_{13} = r_{13}^{*}$, $r_{31} = 1 - r_{13}^{*}$, et

$$r_{23} = r_{23}^{*} pl_{23} = 0.526, \quad r_{32} = (1 - r_{23}^{*}) pl_{23} = 0.265.$$

La combinaison des r_{ij}^{*} et des pl_{ij} par la méthode MCTProb1-1 donne la FMC suivante :

$$\widehat{m}^{\Omega}(\{\omega_1\}) = 0.223,\ \widehat{m}^{\Omega}(\{\omega_1,\omega_2\}) = 0.464,$$

$$\widehat{m}^{\Omega}(\{\omega_1,\omega_3\}) = 0.313,$$

toutes les autres masses étant nulles. En conditionnant $\widehat{m}^{\Omega}$ sur chaque Ω_{ij} par la règle non-normalisée (12), et en calculant les probabilités pignistiques correspondantes, on obtient $BetP_{12}^{*}(\{\omega_1\}) = 0.768$, $BetP_{13}^{*}(\{\omega_1\}) = 0.843$, et $BetP_{23}^{*}(\{\omega_2\}) = 0.597$, qui approchent les sorties des classifieurs binaires probabilistes. □

5. Combinaison de classifieurs 1-T dans le cadre du MCT

5.1. Les sorties des classifieurs vues comme des fonctions de croyance définies sur des cadres grossiers

Soit $\mathcal{E}_k$ un classifieur entraîné à séparer la classe ω_k de l'ensemble des autres. Sa connaissance du domaine Ω est partielle : il est incapable de discerner les classes ω_l, $l \neq k$. Les informations qu'il fournit peuvent donc être modélisées par une FMC $m_k^{\Theta_k}$, définie sur un grossissement $\Theta_k = \{\theta_k^{+}, \theta_k^{-}\}$ de Ω tel que $\rho_k(\{\theta_k^{+}\}) = \{\omega_k\}$, $\rho_k(\{\theta_k^{-}\}) = \overline{\{\omega_k\}}$, ρ_k étant le raffinement qui permet de transformer Θ_k en Ω. Dans le cas de classifieurs probabilistes, les estimations r_k des probabilités conditionnelles $\mathbb{P}(\omega_k|\mathbf{x})$ sont transformées en FMCs Bayésiennes $m_k^{\Theta_k}$.
La FMC $m_k^{\Theta_k}$ peut alors être interprétée comme une réduction d'une FMC m^{Ω} sur Θ_k. Les tableaux 3 et 4 montrent les correspondances entre les éléments focaux $A \subseteq \Omega$ de m^{Ω} et ceux de $\underline{m}^{\Theta_k}$ et $\overline{m}^{\Theta_k}$, respectivement. On constate notamment que la masse $m^{\Omega}(C)$, $C \subset \overline{\{\omega_k\}}$, est transférée à $\emptyset$ par la réduction intérieure $\underline{m}^{\Theta_k}$ et à $\{\theta_k^{-}\}$ par la réduction extérieure $\overline{m}^{\Theta_k}$; et que la masse $m^{\Omega}(\{\omega_k\} \cup C)$, $C \subset \overline{\{\omega_k\}}$, est transférée à $\{\theta_k^{+}\}$ par la réduction intérieure $\underline{m}^{\Theta_k}$ et à Θ_k par la réduction extérieure $\overline{m}^{\Theta_k}$.
La réduction intérieure semble trop peu conservative pour correspondre à l'affectation des masses faite par le classifieur $\mathcal{E}_k$. Il semble par contre raisonnable d'interpréter $m_k^{\Theta_k}$ comme la réduction extérieure d'une FMC m^{Ω} sur Θ_k.

Tableau 3. Correspondance entre les éléments focaux de m^{Ω} et $\underline{m}^{\Theta_k}$.

éléments focaux B de $\underline{m}^{\Theta_k}$	éléments focaux de m^{Ω} dont provient $\underline{m}^{\Theta_k}(B)$
$\emptyset$	$\{A \subset \Omega : \omega_k \notin A, \overline{\omega_k} \not\subseteq A\}$
θ_k^{+}	$\{A \subset \Omega : \omega_k \in A, \overline{\omega_k} \not\subseteq A\}$
θ_k^{-}	$\{A \subset \Omega : \omega_k \notin A, \overline{\omega_k} \subseteq A\} = \overline{\{\omega_k\}}$
Θ_k	$\{A \subset \Omega : \omega_k \in A, \overline{\omega_k} \subseteq A\} = \Omega$

Tableau 4. Correspondance entre les éléments focaux de m^{Ω} et $\overline{m}^{\Theta_k}$.

éléments focaux B de $\overline{m}^{\Theta_k}$	éléments focaux de m^{Ω} dont provient $\overline{m}^{\Theta_k}(B)$
$\emptyset$	$\{A \subset \Omega : \omega_k \notin A, \overline{\omega_k} \cap A = \emptyset\} = \emptyset$
θ_k^{+}	$\{A \subset \Omega : \omega_k \in A, \overline{\omega_k} \cap A = \emptyset\} = \{\omega_k\}$
θ_k^{-}	$\{A \subset \Omega : \omega_k \notin A, \overline{\omega_k} \cap A \neq \emptyset\}$
Θ_k	$\{A \subset \Omega : \omega_k \in A, \overline{\omega_k} \cap A \neq \emptyset\}$

Soit $\overline{\theta}_k$ l'opérateur de réduction extérieure sur le cadre Θ_k. On peut donc écrire :

$$m_k^{\Theta_k} = \overline{\theta}_k(m^\Omega), \tag{42}$$

c'est-à-dire, pour tout $A \subseteq \Theta_k$:

$$m_k^{\Theta_k}(A) = \sum_{B\subseteq\Omega : \rho_k(A)\cap B \neq \emptyset} m^\Omega(B). \tag{43}$$

5.2. Combinaison des fonctions de masse grossières

Ainsi, pour chaque élément $A \subseteq \Theta_k$, un classifieur m_k fournit une masse $m_k^{\Theta_k}(A)$, qui peut être exprimée comme une combinaison linéaire de masses $m^\Omega(B)$. L'équation (42) définit donc un système linéaire de $4K$ équations à $2^K - 1$ inconnues. De même que dans le cas de la combinaison de classifieurs 1-1, les FMC $m_k^{\Theta_k}$ fournies par les différents classifieurs ne sont généralement pas consistantes : il n'existe alors pas de FMC m^Ω dont la réduction extérieure sur le grossissement Θ_k corresponde exactement à $m_k^{\Theta_k}$, pour tout k. Ces FMCs n'étant pas distinctes, elles ne peuvent être combinées par la règle de combinaison conjonctive. On peut donc calculer une solution approchée du système (42), en résolvant un problème d'optimisation quadratique [20] :

$$\widehat{m} = \arg\min_{m^\Omega} \sum_{\Theta_k} \sum_{A\subseteq\Theta_k} \left(\overline{\theta}_k(m^\Omega)(A) - m_k^{\Theta_k}(A)\right)^2, \tag{44}$$

sous les contraintes :

$$m^\Omega(A) \geqslant 0, \quad \forall\, A \subseteq \Omega,$$

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

Ce problème peut être résolu au moyen d'un algorithme usuel d'optimisation quadratique. Cette méthode sera appelée par la suite MCT1-T.

Exemple 6 : Considérons le problème à trois classes décrit dans l'exemple 3. On dispose à présent de trois classifieurs binaires $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$, séparant respectivement $\{\omega_1\}$ de $\{\omega_2,\omega_3\}$, $\{\omega_2\}$ de $\{\omega_1,\omega_3\}$ et $\{\omega_3\}$ de $\{\omega_1,\omega_2\}$. Supposons que, lors de l'évaluation d'un vecteur $\mathbf{x}$, ces classifieurs ont fourni les FMCs normales suivantes :
- $\mathcal{E}_1 : m_1^{\Theta_1}(\{\theta_{11}\}) = 0.008$, $m_1^{\Theta_1}(\{\theta_{12}\}) = 0.991$, $m_1^{\Theta_1}(\Theta_1) = 0.001$;
- $\mathcal{E}_2 : m_2^{\Theta_2}(\{\theta_{21}\}) = 0.652$, $m_2^{\Theta_2}(\{\theta_{22}\}) = 0.338$, $m_2^{\Theta_2}(\Theta_2) = 0.010$;
- $\mathcal{E}_3 : m_3^{\Theta_3}(\{\theta_{31}\}) = 0.436$, $m_3^{\Theta_3}(\{\theta_{32}\}) = 0.556$, $m_3^{\Theta_3}(\Theta_3) = 0.008$.

En les combinant par la méthode MCT1-T, on obtient la FMC $\widehat{m}^\Omega$ suivante :

$$\widehat{m}^\Omega(\{\omega_2\}) = 0.604, \qquad \widehat{m}^\Omega(\{\omega_3\}) = 0.387,$$

$$\widehat{m}^\Omega(\{\omega_2,\omega_3\}) = 0.004, \quad \widehat{m}^\Omega(\Omega) = 0.005.$$

Les réductions extérieures de cette FMC sur les différents Θ_k donnent :
- $\overline{\widehat{m}}^{\Theta_1}(\{\theta_{11}\}) = 0$, $\overline{\widehat{m}}^{\Theta_1}(\{\theta_{12}\}) = 0.995$, $\overline{\widehat{m}}^{\Theta_1}(\Theta_1) = 0.005$;
- $\overline{\widehat{m}}^{\Theta_2}(\{\theta_{21}\}) = 0.604$, $\overline{\widehat{m}}^{\Theta_2}(\{\theta_{22}\}) = 0.387$, $\overline{\widehat{m}}^{\Theta_2}(\Theta_2) = 0.009$;
- $\overline{\widehat{m}}^{\Theta_3}(\{\theta_{31}\}) = 0.387$, $\overline{\widehat{m}}^{\Theta_3}(\{\theta_{32}\}) = 0.604$, $\overline{\widehat{m}}^{\Theta_3}(\Theta_3) = 0.009$,

qui sont les meilleures approximations de $m_1^{\Theta_1}$, $m_2^{\Theta_2}$ et $m_3^{\Theta_3}$ au sens du critère (44). □

6. Réduction de la complexité

Le nombre de sous-ensembles de Ω augmente exponentiellement avec K, ce qui constitue un obstacle à la résolution de problèmes de classification comptant un grand nombre de classes. Ainsi, pour $K = 26$, la FMC m^Ω peut avoir jusqu'à $2^{26} = 67108864$ éléments focaux. Nous proposons de réduire cette complexité en limitant le nombre d'éléments focaux de m^Ω.

Cela peut être fait en identifiant les $L \leqslant K$ classes ω_i de plus grande plausibilité $pl^\Omega(\{\omega_i\})$, et en traitant les $K - L$ autres comme une classe unique. Soient $\omega_{(1)},\ldots,\omega_{(K)}$ les classes ordonnées par plausibilités décroissantes, c'est-à-dire vérifiant :

$$pl^\Omega(\{\omega_{(1)}\}) \geqslant \ldots \geqslant pl^\Omega(\{\omega_{(K)}\}). \tag{45}$$

Soient $\pi_i = \{\omega_{(i)}\}$, pour $i \in \{1,\ldots,L\}$, et $\pi_{L+1} = \{\omega_{(L+1)},\ldots,\omega_{(K)}\}$. L'ensemble $\Pi = \{\pi_1,\ldots,\pi_{L+1}\}$ constitue une partition de Ω. Les classes de π_{L+1} n'ont pas besoin d'être discernées les unes des autres : il est peu plausible que l'une d'elles soit la classe de $\mathbf{x}$. Nous définissons ainsi les sous-ensembles de Π comme étant les éléments focaux potentiels de m^Ω ; le nombre de variables du problème d'optimisation est donc réduit de 2^K à 2^{L+1}.

Dans le cas d'une décomposition 1-1, les plausibilités $pl(\{\omega_k\})$ sont estimées, pour tout $k \in \{1,\ldots,K\}$, pour déconditionner les fonctions de masse fournies par les classifieurs binaires. Dans le cas d'une décomposition 1-T, on propose de les calculer à partir des fonctions de masse m_k. On a :

$$m_k^\Omega(\rho_k(A)) = m_k^{\Theta_k}(A), \quad \forall A \subseteq \Theta_k, \forall \Theta_k.$$

En utilisant l'équation (15), on en déduit :

$$pl_k^\Omega(\{\omega_k\}) = m_k^\Omega(\{\omega_k\}) + m_k^\Omega(\Omega) = m_k^{\Theta_k}(\{\theta_k^+\}) + m_k^{\Theta_k}(\Theta_k) ;$$

$$pl_l^\Omega(\{\omega_k\}) = m_l^\Omega(\overline{\{\omega_l\}}) + m_l^\Omega(\Omega) = m_l^{\Theta_l}(\{\theta_l^-\}) + m_l^{\Theta_l}(\Theta_l),$$
$$\forall\, l \neq k.$$

Pour prendre en compte toutes les fonctions de masse m_k lors du calcul des $pl(\{\omega_k\})$, on propose de les combiner par la somme conjonctive. Bien que l'utilisation de cette règle ne soit pas conseillée (les sources n'étant pas indépendantes), elle per-

met tout de même de déterminer un ordre sur les plausibilités $pl(\{\omega_k\})$. Les plausibilités des classes peuvent être obtenues par :

$$pl(\{\omega_k\}) = \left(\textcircled{\cap}_{l=1}^{K}\ pl_l\right)(\{\omega_k\}) = \prod_{l=1}^{K} pl_l(\{\omega_k\}),$$

$$= \left(m_k^{\Theta_k}(\{\theta_k^+\}) + m_k^{\Theta_k}(\Theta_k)\right) \prod_{l \neq k} \left(m_l^{\Theta_l}(\{\theta_l^-\}) + m_l^{\Theta_l}(\Theta_l)\right).$$

On peut ainsi identifier les $L \leqslant K$ classes ω_k de plus grande plausibilité $pl^{\Omega}(\{\omega_k\})$, et traiter les $K - L$ autres comme une classe unique, de manière à réduire le nombre de variables du problème d'optimisation de 2^K à 2^{L+1}.

Exemple 7 : Considérons un problème $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5\}$. Soient $pl^{\Omega}(\{\omega_1\}) = 0$, $pl^{\Omega}(\{\omega_2\}) = 0.1$, $pl^{\Omega}(\{\omega_3\}) = 0.7$, $pl^{\Omega}(\{\omega_4\}) = 0.3$ et $pl^{\Omega}(\{\omega_5\}) = 0.05$ les plausibilités calculées en évaluant un vecteur $\mathbf{x}$. Supposons que l'on souhaite conserver les deux classes pour lesquelles la plausibilité d'appartenance est la plus élevée.
On définit donc $\pi_1 = \{\omega_3\}$, $\pi_2 = \{\omega_4\}$ et $\pi_3 = \{\omega_1,\omega_2,\omega_5\}$, et on autorise $\widehat{m}^{\Omega}$ à attribuer de la masse aux huit éléments suivants : $\emptyset$, π_1, π_2, π_3, $\pi_1 \cup \pi_2 = \{\omega_3,\omega_4\}$, $\pi_1 \cup \pi_3 = \{\omega_1,\omega_2,\omega_3,\omega_5\}$, $\pi_2 \cup \pi_3 = \{\omega_1,\omega_2,\omega_4,\omega_5\}$, et Ω. □

7. Expériences

7.1. Implémentation des méthodes 1-1 et 1-T

7.1.1. Méthodes de combinaison comparées

Cinq méthodes de combinaison 1-1 ont été comparées expérimentalement : la méthode MCTCorr1-1 présentée au paragraphe 4, ou le cas échéant sa variante MCTProb1-1 présentée au paragraphe 4.3 permettant de combiner des classifieurs probabilistes, en interprétant leurs sorties comme des estimations de probabilités pignistiques ; et les méthodes PCpl, PEst1, PEst2 et PEstCorr, présentées au paragraphe 2.2.1. De même, trois méthodes de combinaison 1-T ont été comparées : la méthode MCT1-T présentée au paragraphe 5, et les méthodes PCpl1-T et PEst1-T présentées au paragraphe 2.2.2.

7.1.2. Classifieurs binaires utilisés

Trois méthodes de classification binaires ont été utilisées : la *régression logistique* [11], les *arbres de décision* binaires [3] et la méthode des *réseaux de neurones évidentiels* [4]. La régression logistique et les arbres de décision binaires fournissent des estimations des probabilités conditionnelles d'appartenance aux classes, tandis que les réseaux de neurones évidentiels calculent des FMCs conditionnelles non bayésiennes m_{ij}^*.

La régression logistique a servi à comparer les méthodes de combinaison de classifieurs dans le cadre d'une décomposition 1-1 : MCTProb1-1, PCpl, PEst1 et PEst2, et PEstCorr. Elle a également été utilisée pour l'analyse visuelle des propriétés de la méthode MCTCorr1-1.
Les arbres de décision ont servi à évaluer les méthodes de combinaison de classifieurs 1-T : MCT1-T, PCpl1-T et PEst1-T. L'algorithme CART [3] a été employé. Les arbres ont été élagués en calculant une séquence d'arbres $\mathcal{A}_t$ de taille croissante, pour lesquels l'erreur ϵ_t et son écart-type $\text{std}(\epsilon_t)$ ont été estimées en utilisant la validation croisée à 10 coupes. L'arbre sélectionné était le plus petit arbre vérifiant $\epsilon_t - \min_t(\epsilon_t) \leqslant \text{std}(\epsilon_t)$.
Les réseaux de neurones évidentiels ont servi à évaluer les méthodes de combinaison de classifieurs 1-1 et 1-T : MCTCorr1-1, PCpl, PEst1 et PEst2, PEstCorr ; et MCT1-T, PCpl1-T et PEst1-T. En outre, les résultats servant de base à l'analyse visuelle de la méthode MCT1-T ont été obtenus au moyen de ces classifieurs. Pour évaluer les méthodes de combinaison de classifieurs probabilistes (PCpl, PEst1 et PEst2 et PEstCorr dans le cas 1-1, PCpl1-T et PEst1-T dans le cas 1-T), les FMCs fournies par les classifieurs ont été transformées en probabilités pignistiques, qui ont été utilisées comme données initiales. Les réseaux de neurones évidentiels caractérisent les classes par des prototypes[1]. Dans le cas 1-1, chaque classe est décrite par trois prototypes ; dans le cas 1-T, les classes positives sont décrites par trois prototypes et les classes négatives par $3(K-1)$ prototypes, sauf pour le jeu de données `Vowel` où $3(K-1)$ prototypes ont été nécessaires pour décrire les classes positives et négatives. Les autres paramètres ont été fixés à leur valeur par défaut.
Les tableaux 5 et 6 résument les expériences menées, pour les décompositions 1-1 et 1-T respectivement.

7.1.3. Algorithmes employés : classifieurs à une classe, classifieurs correcteurs

Dans la méthode MCT1-1, les plausibilités $pl^{\Omega}(\{\omega_k\})$ ont été calculées au moyen de séparateurs à vaste marge à une classe (1-SVM) [21]. Un 1-SVM estime le support d'une distribution, et le décrit au moyen d'un sous-ensemble choisi de vecteurs issus de cette distribution, appelés vecteurs de support. La distance signée $f_k(\mathbf{x})$ d'un vecteur $\mathbf{x}$ au support de ω_k peut alors être calculée. Un paramètre ν choisi arbitrairement permet de fixer une borne inférieure sur la fraction de vecteurs de support et une borne supérieure sur la fraction de points aberrants ; ici, ν a été fixé à 0.2. La plausibilité $pl^{\Omega}(\{\omega_k\})$ a été calculée par :

$$pl^{\Omega}(\{\omega_k\}) = \frac{f_k(\mathbf{x}) + \rho}{\rho}, \tag{46}$$

où ρ est un paramètre obtenu lors de l'apprentissage du 1-SVM (détails dans [21]). Un tel algorithme se distingue donc d'un classifieur binaire 1-T dans la mesure où il prend en compte des

1. Les codes sources MatLab de cette méthode sont disponibles à l'URL `http://www.hds.utc.fr/~tdenoeux/software.htm`.

Tableau 5. Résumé des expériences pour les méthodes de combinaison de classifieurs 1-1.

Type de classifieur binaire utilisé	Méthodes de combinaison crédales : MCTCorr1-1, MCTProb1-1	Méthodes de combinaison probabilistes : PCpl, PEst1, PEst2, PEst
régression logistique sorties : probabilités r_{ij}^* normalisées	dénormalisation par les plausibilités pl_{ij} : calcul des r_{ij}, puis combinaison (MCTProb1-1)	combinaison directe des probabilités normalisées r_{ij}^* (correction par les q_{ij} : PEstCorr)
réseaux évidentiels sorties : FMCs m_{ij}^* normales	dénormalisation par les plausibilités pl_{ij} : calcul des m_{ij}, puis combinaison des m_{ij} (MCTCorr1-1)	transformation pignistique des m_{ij}^* en probabilités conditionnelles r_{ij}^*, puis combinaison (PEstCorr : correction par q_{ij})

Tableau 6. Résumé des expériences pour les méthodes de combinaison de classifieurs 1-T.

Type de classifieur binaire utilisé	Méthodes de combinaison crédale : MCT1-T	Méthodes de combinaison probabilistes : PCpl1-T, PEst1-T
arbres de décision sorties : probabilités r_k	transformation en FMCs Bayésiennes m_k, puis combinaison des m_k	combinaison directe des probabilités r_k
réseaux évidentiels sorties : FMCs m_k	combinaison directe des FMCs m_k	transformation pignistique des m_k en probabilités r_k, puis combinaison

données non étiquetées : il s'agit d'un outil de description des données et non de discrimination.

Dans la méthode PEstCorr, les probabilités correctrices q_{ij} ont été calculées à partir d'estimations g_i de la densité de probabilité de chaque classe ω_i par une méthode de noyaux :

$$q_{ij} = \frac{n_i g_i + n_j g_j}{\sum_{k=1}^{K} n_k g_k}.$$

Dans les deux cas (1-SVM et méthode des noyaux), une estimation $\widehat{\sigma}_k$ de la largeur de bande des noyaux a été obtenue par la méthode proposée dans [15], pour chaque classe ω_k. Une largeur de bande utilisée pour toutes les classes a ensuite été déterminée par :

$$\widehat{\sigma}_{opt} = 1.5 \left(\frac{1}{K} \sum_{k=1}^{K} \widehat{\sigma}_k \right).$$

Pour les jeux de données de plus de $K = 6$ classes, la complexité a été réduite en sélectionnant les cinq classes de plus grande plausibilité $pl_k^{\Omega}(\{\omega_k\})$.

7.2. Exemple qualitatif

7.2.1. Décomposition 1-1

Dans ce paragraphe, nous illustrons les résultats obtenus sur un jeu de données synthétiques de $\mathbb{R}^2$ comptant $K = 3$ classes avec les méthodes MCTCorr1-1 et MCT1-T. Les résultats présentés concernent les classes ω_1 et ω_2.

Les figures 1 à 4 illustrent les résultats obtenus pour la méthode MCTCorr1-1 (paragraphe 4.2). La figure 1 montre les données initiales de la méthode : la régression logistique permet de déterminer une frontière linéaire entre les classes ω_1 et ω_2, et de calculer des estimations r_{12} des probabilités conditionnelles transformées ensuite en FMCs Bayésiennes m_{12}^* ; les 1-SVMs déterminent une région de l'espace ou les classes sont plausibles, et permettent de calculer les plausibilités pl_{12}. En combinant ces deux sources d'information, on obtient les FMCs conditionnelles non normalisées m_{12} représentées sur la figure 2. Le classifieur $\mathcal{E}_{12}$ étant probabiliste, on a $m_{12}(\{\omega_1,\omega_2\}) = 0$. On remarquera que les courbes de niveau des masses $m_{12}(\{\omega_1\})$ et $m_{12}(\{\omega_2\})$ sont linéaires entre les classes ω_1 et ω_2 : elles correspondent donc dans cette région de l'espace aux informations fournies par le classifieur binaire $\mathcal{E}_{12}$. Les masses $\widehat{m}^{\Omega}(\emptyset)$, $\widehat{m}^{\Omega}(\{\omega_1\})$, $\widehat{m}^{\Omega}(\{\omega_2\})$ et $\widehat{m}^{\Omega}(\{\omega_1,\omega_2\})$ obtenues en combinant les trois FMCs m_{12}, m_{13} et m_{23} par la méthode MCTCorr1-1 sont représentées sur la figure 3. On constate que la masse $\widehat{m}^{\Omega}(\{\omega_1,\omega_2\})$ est maximale autour de la frontière entre ω_1 et ω_2, tandis que la masse $\widehat{m}^{\Omega}(\emptyset)$ est significative dans les régions de faible densité de données. Les probabilités pignistiques correspondantes $BetP^*(\omega_1)$ et $BetP^*(\omega_2)$ sont représentées sur la figure 4.

7.2.2. Décomposition 1-T

Les figures 5 à 8 illustrent les résultats obtenus pour la méthode MCT1-T. Les figures 5 et 6 montrent respectivement les masses Bayésiennes $m_1^{\Theta_1}$ et $m_2^{\Theta_2}$, vues comme des estimations de réductions extérieures sur Θ_1 et Θ_2, calculées au moyen de réseaux

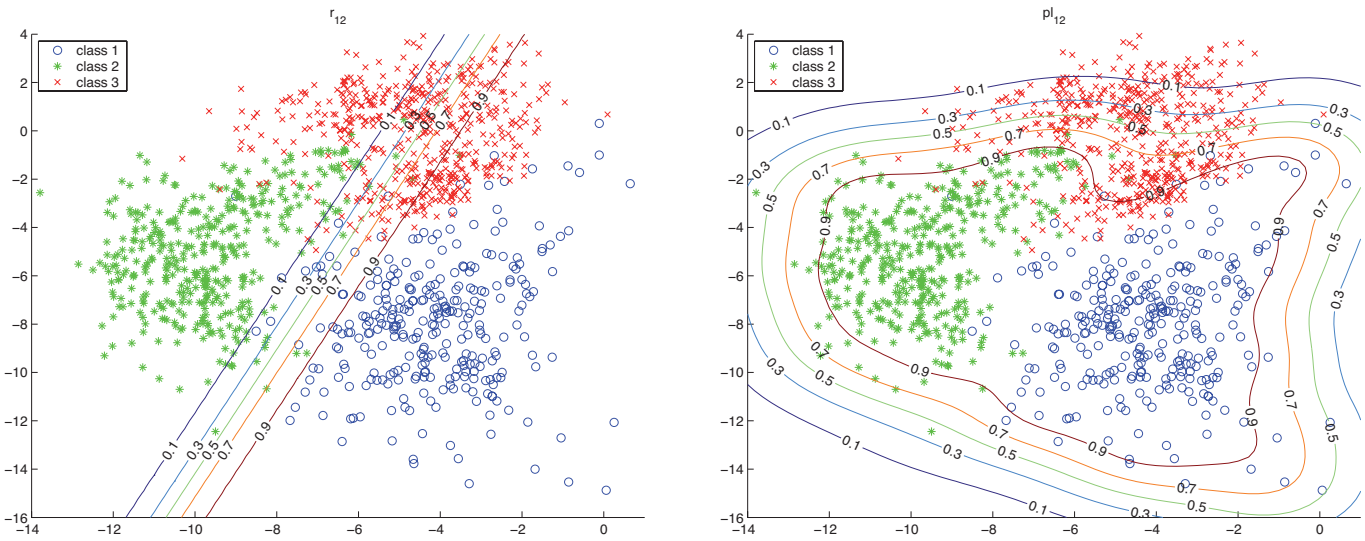

Figure 1. Estimations r_{12} des probabilités conditionnelles obtenues avec la régression logistique (gauche) ; estimation des plausibilités pl_{12} obtenues avec des 1-SVM (droite).

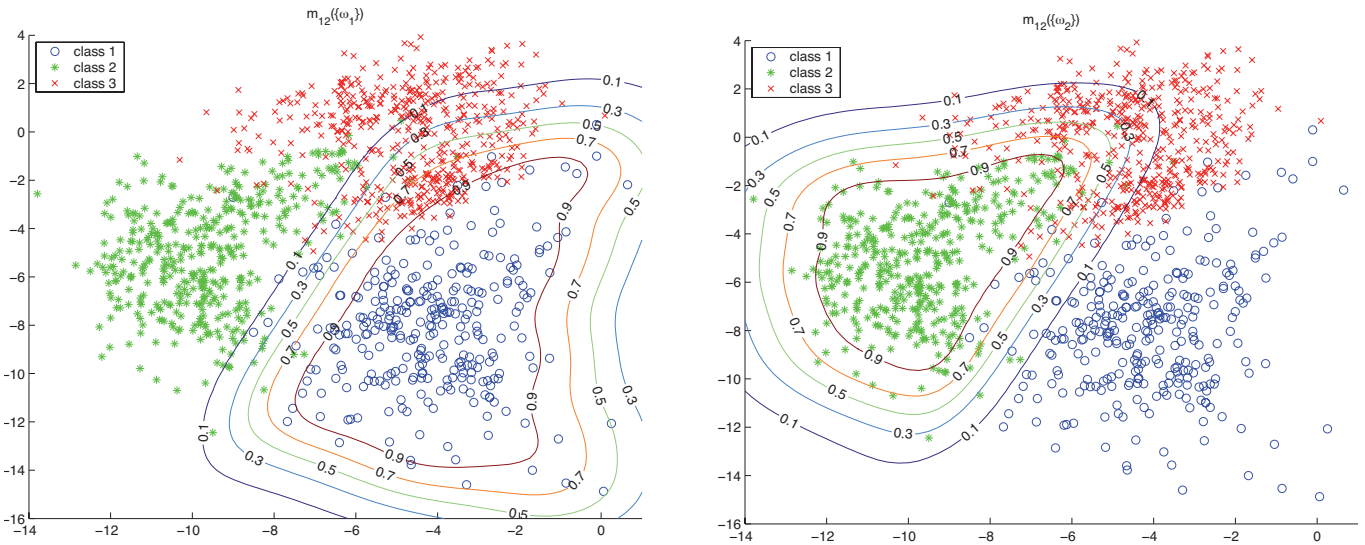

Figure 2. Masses conditionnelles dénormalisées $m_{12}(\{\omega_1\})$ (gauche) et $m_{12}(\{\omega_2\})$ (droite) obtenues en combinant r_{12} et pl_{12}.

de neurones évidentiels. Les masses $\widehat{m}^{\Omega}(\{\omega_1\})$, $\widehat{m}^{\Omega}(\{\omega_2\})$ et $\widehat{m}^{\Omega}(\{\omega_1\omega_2\})$, obtenues en combinant les masses $m_k^{\Theta_k}$ par la méthode MCT1-T (paragraphe 5.2), sont représentées sur la figure 7. Les probabilités pignistiques correspondantes sont représentées sur la figure 8. À la frontière entre les trois classes, les courbes de niveau de la masse $\widehat{m}(\{\omega_1\})$ sont très proches de celles de la masse $m_1^{\Theta_1}(\{\theta_1^+\})$; les masses $\widehat{m}(\{\omega_2\})$, au contraire, présentent certaines différences avec les masses $m_2^{\Theta_2}(\{\theta_2^+\})$. Cela est dû au fait que les classes ω_2 et ω_3 se recouvrent partiellement, tandis que la classe ω_1 est relativement bien séparable des deux autres. Dans cette région, la combinaison donne une masse qui reflète un certain compromis entre les informations fournies par $\mathcal{E}_2$ et $\mathcal{E}_3$, mais qui est consistante avec celles fournies par $\mathcal{E}_1$.

7.3. Résultats quantitatifs

Le tableau 7 présente les caractéristiques des jeux de données utilisés pour ces expériences. Tous les jeux de données proviennent de la base de données du département d'Apprentissage Statistique de l'UCI, excepté le jeu de données `Synth` (similaire à celui utilisé au paragraphe 7.2, avec une classe additionnelle), généré par mélange de Gaussiennes.

Les tableaux 8 et 9 présentent les taux de vecteurs bien classés, obtenus après combinaison par les méthodes TBMProb1-1 (tableau 8) et TBMProb1-1 (tableau 9), PCpl, PEst1, PEst2 et PEstCorr ; les classifieurs binaires utilisés sont la régression logistique (tableau 8) et les réseaux de neurones évidentiels (tableau 9). Les tableaux 10 et 11 présentent les taux de vecteurs

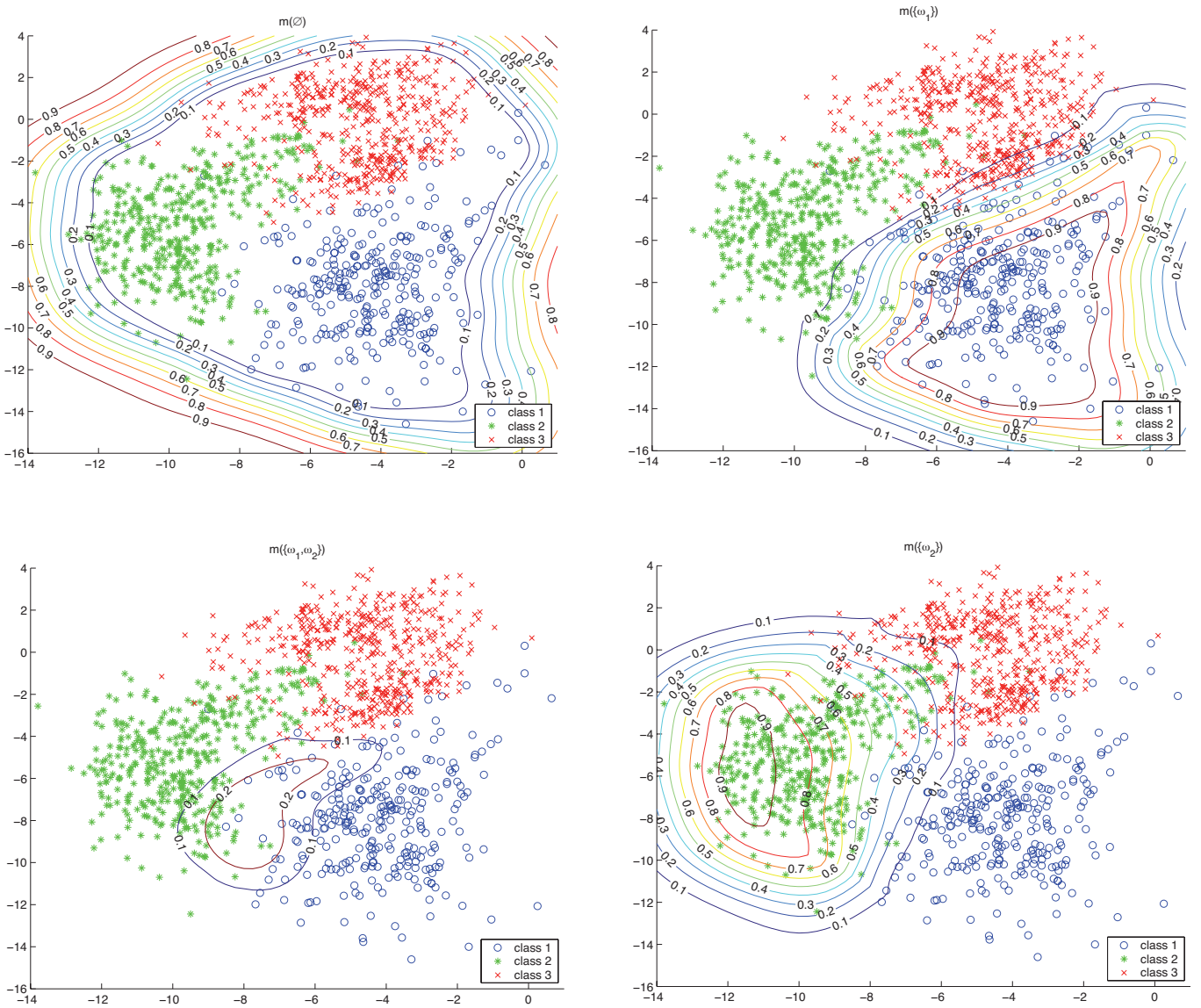

Figure 3. Masses de croyance $\widehat{m}^{\Omega}(\emptyset)$ (haut-gauche), $\widehat{m}^{\Omega}(\{\omega_1\})$ (haut-droite), $\widehat{m}^{\Omega}(\{\omega_2\})$ (bas-gauche) et $\widehat{m}^{\Omega}(\{\omega_1,\omega_2\})$ (bas-droite) obtenues en combinant les sorties de tous les classifieurs binaires et à une classe, par la méthode MCTCorr1-1 (paragraphe 4.2).

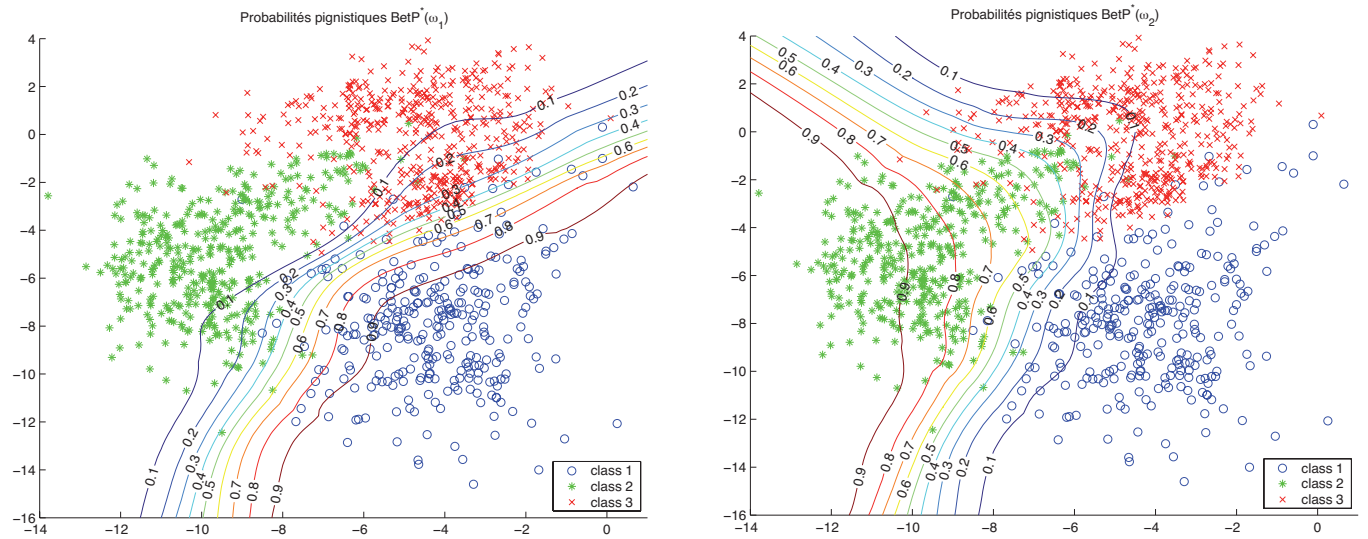

Figure 4. Probabilités pignistiques $Bet\,P^*(\omega_1)$ (gauche) et $Bet\,P^*(\omega_2)$ (droite) calculées à partir de $\widehat{m}^{\Omega}$.

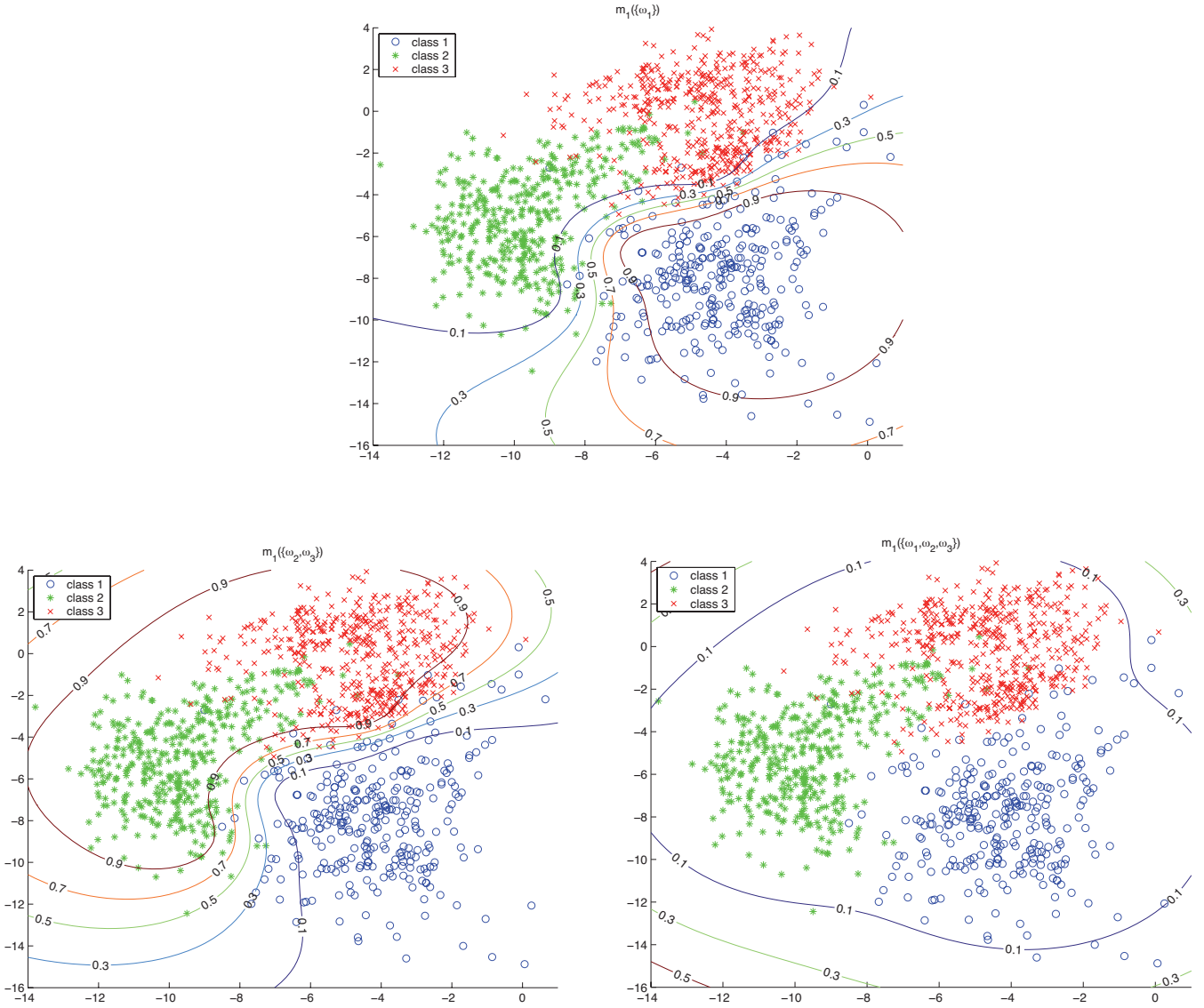

Figure 5. Masses estimées par le classifieur $\mathcal{E}_1$: $m_1^{\Theta_1}(\{\theta_1^+\})$ (haut), $m_1^{\Theta_1}(\{\theta_1^-\})$ (bas-gauche) et $m_1^{\Theta_1}(\Theta_1)$ (bas-droite).

bien classés, obtenus après combinaison par les méthodes TBM1-T, PCpl1-T et PEst1-T ; les classifieurs binaires 1-T utilisés sont les arbres de décision binaires (tableau 10) et les réseaux de neurones évidentiels (tableau 11). La significativité des différences entre les taux a été évaluée au moyen du *test de Mc Nemar* [6] au niveau 5 %. Pour chaque jeu de données, le meilleur résultat est souligné, et ceux qui ne sont pas significativement inférieurs apparaissent en gras.

7.3.1 Décomposition 1-1

Sur les six jeux de données, la méthode MCT1-1 donne les meilleurs résultats pour cinq jeux de données en utilisant la régression logistique comme classifieur binaire, et pour quatre jeux de données en utilisant les réseaux de neurones évidentiels. Les autres résultats ne sont pas significativement moins bons

Tableau 7. Caractéristiques des jeux de données.

données	dimension	nb. classes	nb. vecteurs / appr.	nb. vecteurs / test
`Ecoli`	7	8	201	135
`Glass`	9	6	139	75
`Letter`	16	26	7 800	10 400
`Satimage`	36	6	2 573	3 862
`Segment`	19	7	1 400	910
`Synth`	2	4	1 700	340
`Vowel`	10	11	528	462
`Waveform`	21	3	1 500	3 500

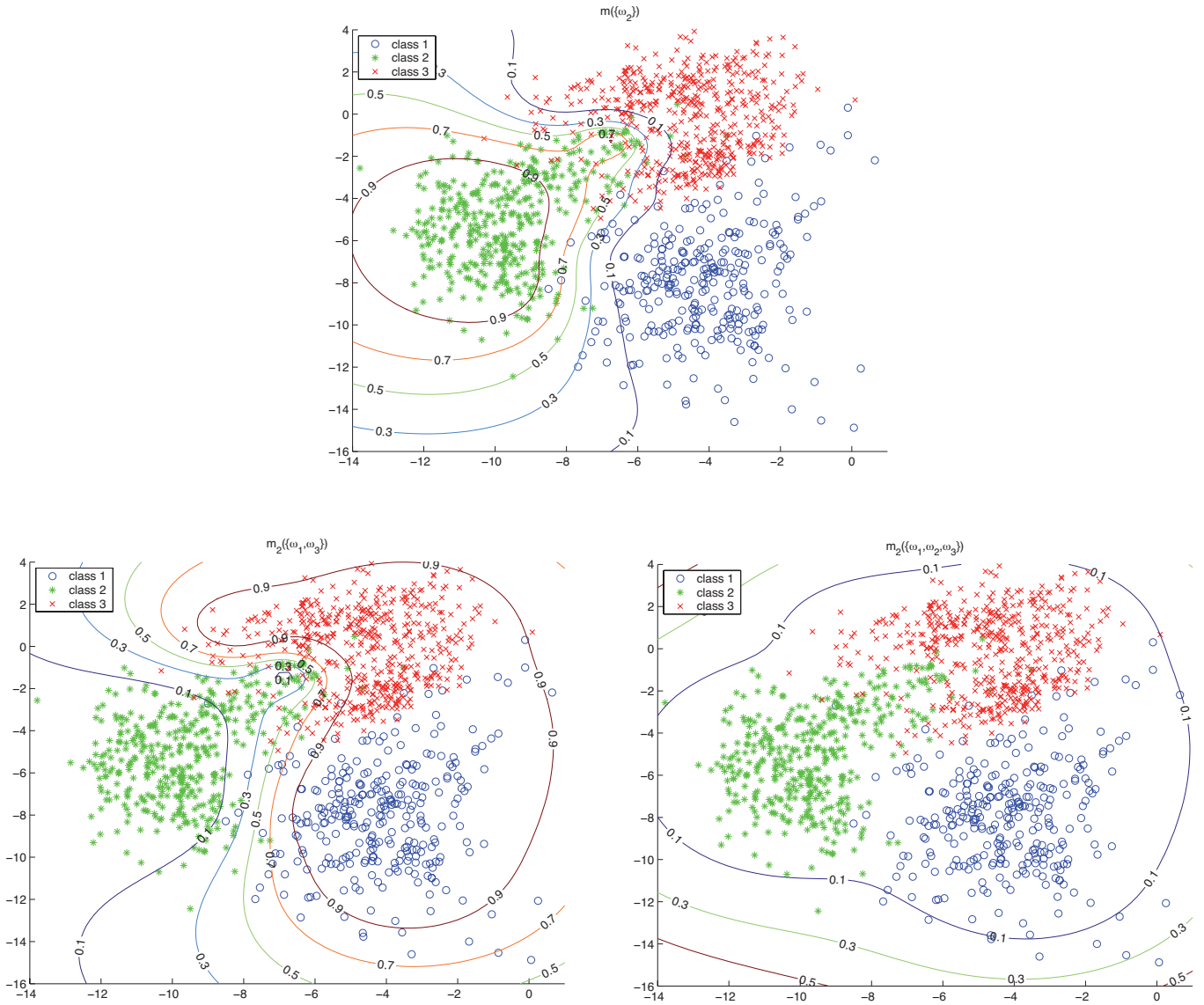

Figure 6. Masses estimées par le classifieur $\mathcal{E}_2$: $m_2^{\Theta_2}(\{\theta_2^+\})$ (haut), $m_2^{\Theta_2}(\{\theta_2^-\})$ (bas-gauche) et $m_2^{\Theta_2}(\Theta_2)$ (bas-droite).

Tableau 8. Taux de bonne classification après combinaison (%), décomposition 1-1, régression logistique.

Méthode	Glass	Satimage	Segment	Synth	Vowel	Waveform
TBMPROB1-1	**56.0**	**93.1**	**96.0**	**95.6**	**66.0**	**85.3**
PCPL	**58.7**	87.1	**96.0**	**94.4**	51.3	**85.1**
PEST1	**58.7**	86.9	**95.6**	**94.4**	50.9	**85.1**
PEST2	**60.0**	86.9	**95.6**	**94.4**	52.6	**85.1**
PESTCORR	**60.0**	90.8	90.3	**95.6**	60.6	**85.2**

que les meilleurs résultats obtenus. L'intérêt de prendre en compte la pertinence des classifieurs 1-1 lors de leur combinaison est donc démontrée empiriquement par ces résultats. Cette correction des classifieurs donne généralement de meilleurs résultats avec les méthodes MCTProb1-1 et MCTCorr1-1 qu'avec la méthode PEstCorr. Les deux premières méthodes semblent être moins sensibles au calcul des plausibilités pl_{ij}, que la dernière au calcul des probabilités correctrices q_{ij}.

Remarquons enfin qu'il n'a pas été possible de déterminer un type de problème particulier pour lequel les méthodes de combinaison TBMCorr1-1 et TBMProb1-1 donnent de meilleurs résultats que les autres.

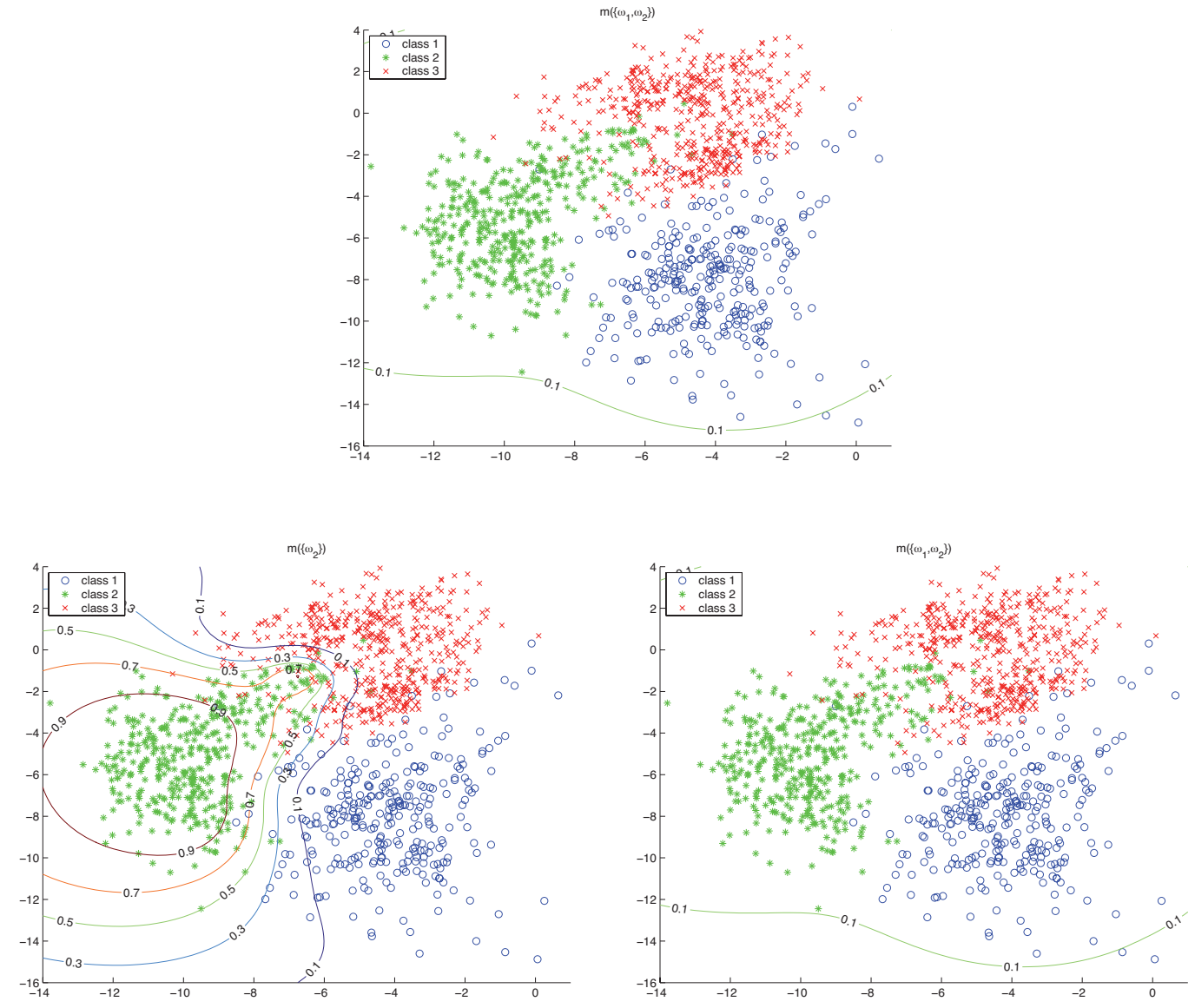

Figure 7. Masses de croyance $\widehat{m}^{\Omega}(\{\omega_1\})$ (haut), $\widehat{m}^{\Omega}(\{\omega_2\})$ (bas-gauche) et $\widehat{m}^{\Omega}(\{\omega_1,\omega_2\})$ (bas-droite) obtenues en combinant toutes les FMCs $m_k^{\Theta_k}$ fournies par les classifieurs binaires, par la méthode MCT1-T (paragraphe 5.2).

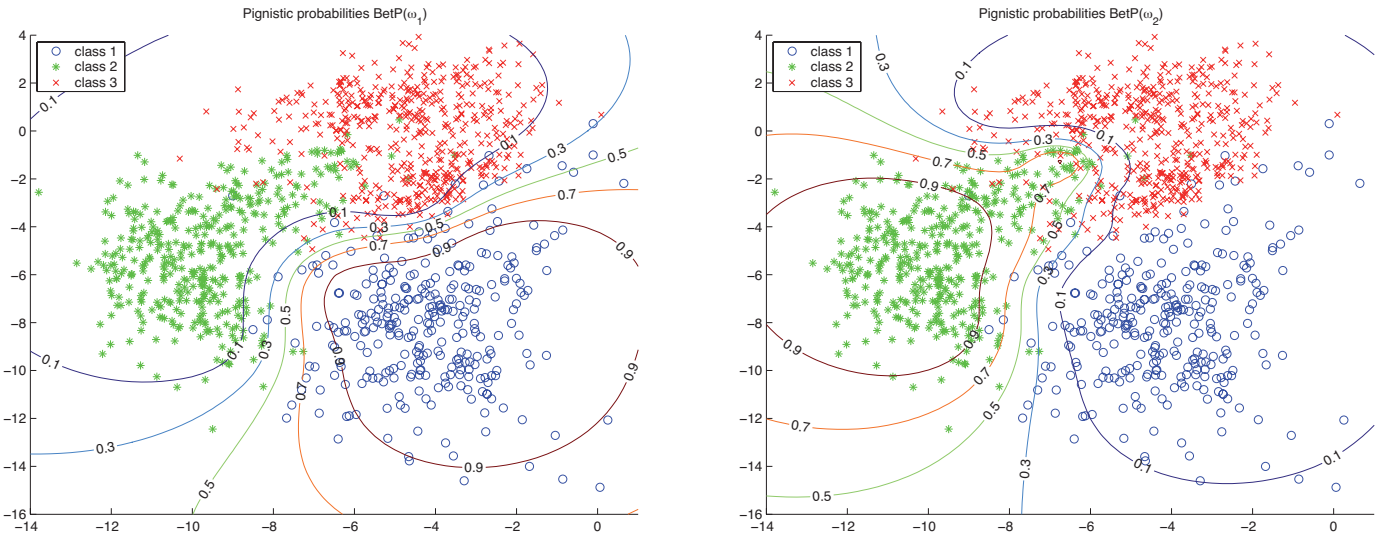

Figure 8. Probabilités pignistiques $BetP^*(\omega_1)$ (gauche) et $BetP^*(\omega_2)$ (droite) calculées à partir de $\widehat{m}^{\Omega}$.

Tableau 9. Taux de bonne classification après combinaison (%), décomposition 1-1, réseaux de neurones évidentiels.

Méthode	Glass	Satimage	Segment	Synth	Vowel	Waveform
TBMPROB1-1	**58.7**	**92.9**	**90.1**	**96.2**	**67.1**	**85.8**
PCPL	**50.7**	86.8	85.2	**95.3**	**64.7**	**86.3**
PEST1	**50.7**	87.0	86.0	**95.6**	**66.7**	**86.2**
PEST2	**52.0**	87.3	86.0	**95.9**	**66.5**	**86.0**
PESTCORR	**61.3**	89.9	82.4	**95.9**	61.7	**85.9**

Tableau 10. Taux de bonne classification après combinaison (%), décomposition 1-T, arbres de décision.

Méthode	Glass	Letter	Satimage	Segment	Synth	Vowel	Waveform
TBM1-T	53.3	**76.2**	85.2	**95.3**	95.3	**35.9**	76.1
PCPL1-T	53.3	**76.2**	85.2	**95.1**	95.3	33.8	76.1
PEST1-T	53.3	76.0	85.2	**94.9**	95.3	34.0	76.1

Tableau 11. Taux de bonne classification après combinaison (%), décomposition 1-T, réseaux de neurones évidentiels.

Méthode	Glass	Satimage	Segment	Synth	Vowel	Waveform
TBM1-T	56.0	**84.3**	**76.8**	95.3	64.1	86.2
PCPL1-T	56.0	**84.2**	**76.9**	95.3	64.1	86.2
PEST1-T	56.0	**84.2**	**76.6**	95.3	64.1	86.2

7.3.2. Décomposition 1-T

On peut constater que les résultats obtenus sur les différents jeux de données, pour les méthodes MCT1-T, PCPL1-T et PEST1-T, sont presque identiques. La méthode MCT1-T donne les meilleurs résultats pour trois des jeux de données, lorsque les classifieurs binaires sont des arbres de décision ; les différences sont significatives dans deux cas. Globalement, les différentes méthodes de combinaison semblent être de précision équivalente. Dans le cas de la décomposition 1-T, une classe ω_k n'est discernée des autres que pour un seul des K classifieurs binaires, et seuls les classifieurs $\mathcal{E}_k$ et $\mathcal{E}_l$ servent à séparer les classes ω_k et ω_l. Les informations utilisées pour déterminer la frontière de ω_k sont donc moins nombreuses que dans le cas d'une décomposition 1-1 (la classe ω_k est alors discernée des autres par $K-1$ des C_K^2 classifieurs binaires). En outre, la corrélation des classifieurs est importante dans le cas 1-T : pour tout $k \neq l$, les classifieurs $\mathcal{E}_k$ et $\mathcal{E}_l$ ont même ensemble d'apprentissage à l'exception des vecteurs $\mathbf{x} \in \omega_k$ et $\mathbf{x} \in \omega_l$. On peut donc s'attendre à ce que les classifieurs 1-T commettent des erreurs simultanées, plus souvent que ne le font les classifieurs 1-1.

8. Conclusion

La combinaison de classifieurs binaires est une approche intéressante pour la résolution de problèmes de classification multi-classes. Dans cet article, nous avons présenté deux méthodes de combinaison de classifieurs, dans le cadre de la théorie des fonctions de croyance, pour deux schémas de décomposition du domaine Ω couramment utilisés.

Dans le cas d'une décomposition un-contre-un, les sorties des classifieurs sont interprétées comme des FMCs m_{ij}^* normalisées, définies sur des conditionnements Ω_{ij} correspondant à des paires de classes. Lors de l'évaluation d'un vecteur $\mathbf{x}$, la pertinence du classifieur binaire séparant ω_i et ω_j est quantifiée en estimant la plausibilité que $\mathbf{x}$ appartienne à ω_i ou ω_j, au moyen de classifieurs à une classe. Cette information permet de calculer les FMCs m_{ij} sous-normales associées à chaque domaine Ω_{ij}. Pour chaque $\mathbf{x}$, une fonction de masse définie sur Ω est calculée, de telle manière que ses conditionnements sur les Ω_{ij} soient les plus proches possibles des m_{ij}. Dans le cas d'une décomposition un-contre-tous, les sorties des classifieurs sont interprétées comme des FMCs $m_k^{\Theta_k}$ définies sur des grossissements Θ_k de Ω. Ces informations sont combinées en calculant une masse définie sur Ω, dont les réductions extérieures sur les différents Θ_k sont les plus proches possibles des $m_k^{\Theta_k}$.

La méthode de combinaison un-contre-un donne globalement de meilleurs résultats de classification que la méthode de combinaison un-contre-tous. Pour chaque classe ω_k, les informations disponibles sont plus nombreuses et moins dépendantes les unes des autres dans le premier cas. Cette méthode reste cependant plus coûteuse que la méthode un-contre-tous, notamment de par la nécessité d'estimer la pertinence des classifieurs. Notons que la méthode proposée dans le cas un-contre-un donne de meilleurs résultats de classification que d'autres méthodes de combinaison déjà existantes, tandis que la méthode proposée dans le cas un-contre-tous donne des résultats très similaires à ceux obtenus avec d'autres méthodes de combinaison. La généralisation de ces deux méthodes de combinaison au cas des codes correcteurs d'erreurs [7,1] est actuellement en cours de réalisation.

Références

[1] E. L. ALLWEIN, R. E. SCHAPIRE, and Y. SINGER, Reducing multiclass to binary : a unifying approach for margin classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 9-16. Morgan Kaufmann, San Francisco, CA, 2000.

[2] A. APPRIOU, Approche générique de la gestion de l'incertain dans les processus de fusion multisenseur, *Revue Traitement du Signal*, 22(4): 307-319, 2005.

[3] L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, and C. J. STONE, *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

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

[5] T. DENŒUX and A. B. YAGHLANE, Approximating the combination of belief functions using the fast möbius transform in a coarsened frame, *International Journal of Approximate Reasoning*, 31(1), 2002.

[6] T. G. DIETTERICH, Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation*, 10(7): 1895-1923, 1998.

[7] T. G. DIETTERICH and G. BAKIRI, Solving multiclass learning problems via errorcorrecting output codes. *Journal of Artificial Intelligence Research*, 2: 263-286, 1995.

[8] J. FRIEDMAN, Another approach to polychotomous classification. Technical report, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, 1996.

[9] J. FÜRNKRANZ, Round robin rule learning, In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning* (ICML-01), pages 146-153, Williamstown, MA, Morgan Kaufmann Publishers, 2001.

[10] T. HASTIE and R. TIBSHIRANI, Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, The MIT Press, 1998.

[11] T. HASTIE, R. TIBSHIRANI and J. FRIEDMAN, *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, Springer Verlag, New York, 2001.

[12] T.-K. HUANG, R. C. WENG and C.-J. LIN, Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7 :85–115, 2006.

[13] F. JANEZ and A. APPRIOU, Theory of evidence and non-exhaustive frames of discernment : Plausibilities correction methods, *International Journal of Approximate Reasoning*, 18(1-2): 1-19, 1998.

[14] S. KNERR, L. PERSONNAZ and G. DREYFUS, Single-layer learning revisited : a stepwise procedure for building and training a neural network. In F. Fogelman-Soulie and J. Herault, editors, *Neurocomputing: Algorithms, Architectures and Applications – NATO ASI Series*, volume F68, Springer Verlag, Berlin, Germany, 1990.

[15] W. KOONTZ and K. FUKUNAGA. Asymptotic analysis of a nonparametric clustering technique. *IEEE Transactions on Computers*, C-21(9) : 967-974, 1972.

[16] M. MOREIRA and E. MAYORAZ, Improved pairwise coupling classification with correcting classifiers, In *European Conference on Machine Learning*, pages 160-171, 1998.

[17] A. PASSERINI, M. PONTIL, and P. FRASCONI. New results on error correcting output codes of kernel machines. *IEEE Transactions on Neural Networks*, 15(1): 45-54, 2004.

[18] J. PLATT, N. CRISTIANINI, and J. SHAWE-TAYLOR, Large margin dags for multiclass classification. In S. Solla, T. Leen, and K.-R. Mueller, editors, *Advances in Neural Information Processing Systems 12*, pages 547–553, 2000.

[19] B. QUOST, T. DENŒUX and M. MASSON, Pairwise classifier combination in the framework of belief functions. In *Proceedings of FUSION'2005*, Philadelphia, PA, July 2005.

[20] B. QUOST, T. DENŒUX and M. MASSON, One-against-all classifier combination in the framework of belief functions. In *Proceedings of IPMU'2006*, volume 1, pages 356-363, Paris, 28, July 2006.

[21] B. SCHÖLKOPF, J. PLATT, J. SHAWE-TAYLOR, A. SMOLA and R. WILLIAMSON, Estimating the support of a high-dimensional distribution, Technical report, Microsoft Research, 1999.

[22] G. SHAFER, *A Mathematical Theory of Evidence*, Princeton Univ. Press. Princeton, NJ, 1976.

[23] P. SMETS and R. KENNES, The transferable belief model, *Artificial Intelligence*, 66 : 191- 234, 1994.

[24] T.-F. WU, C.-J. LIN and R. C. WENG, Probability estimates for multi-class classi- fication by pairwise coupling, *Journal of Machine Learning Research*, 5 : 975-1005, 2004.

[25] B. ZADROZNY, Reducing multiclass to binary by coupling probability estimates. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press.

Benjamin **Quost**

B. Quost est ingénieur (2003) de l'Université de Technologie de Compiègne. Il est actuellement doctorant au laboratoire HeuDiaSyC de l'Université de Technologie de Compiègne, sous la direction de T. Denœux et M.-H. Masson. Ses recherches portent sur la théorie des fonctions de croyance, et plus particulièrement sur son application à la combinaison de classifieurs.

Marie-Hélène **Masson**

M.-H. Masson est ingénieur et docteur HDR de l'Université de Technologie de Compiègne. Elle est Maître de Conférences à l'Université de Picardie Jules Verne et membre du laboratoire Heudiasyc de l'Université de Technologie de Compiègne. Ses recherches portent sur l'analyse et la fusion de données par des théories non probabilistes de gestion de l'incertain (théorie des possibilités, théorie des fonctions de croyance).

Thierry **Denœux**

T. Denœux est ingénieur civil (1985) et docteur (1989) de l'École Nationale des Ponts et Chaussées. Il est actuellement professeur à l'Université de Technologie de Compiègne, co-responsable du thème Apprentissage Statistique, Reconnaissance de formes, Image et Décision au sein du laboratoire Heudiasyc (UMR CNRS 6599), et rédacteur en chef de la revue International Journal of Approximate Reasoning publiée par Elsevier. Ses recherches portent principalement sur la théorie des fonctions de croyance et son application à la représentation des incertitudes en reconnaissance de formes et en fusion d'informations.