Un modèle d'extraction de masses de croyance à partir de probabilités a posteriori pour une amélioration des performances en classification supervisée

Teh Amouh*, Monique Noirhomme-Fraiture*, Benoît Macq**

*Université de Namur,
Faculté d'informatique,
rue grandgagnage, 21, 5000 Namur, Belgique
{tam, mno}@info.fundp.ac.be
http://www.info.fundp.ac.be
**Université catholique de Louvain,
Laboratoire de Télécommunications et Télédétection,
Place Stevin, 2, 1348 Louvain-la-neuve, Belgique
benoit.macq@uclouvain.be
http://www.tele.ucl.ac.be

Résumé. L'objectif de cet article est de montrer que l'utilisation de la règle de décision du *maximum de masse de croyance* en lieu et place de celle du *maximum de probabilité a posteriori* peut permettre de réduire le taux d'erreur en classification supervisée. Nous proposons une technique efficace pour extraire, à partir d'un vecteur de probabilités a posteriori, un vecteur de masses de croyance sur lequel baser la décision par le maximum de masse de croyance. L'application de notre méthode dans le domaine de la classification automatique en stades de sommeil montre une amélioration des performances pouvant atteindre 80% de réduction du taux d'erreur de classification.

1 Introduction

En classification supervisée, l'information en sortie d'un classifieur se présente généralement sous la forme d'un vecteur de probabilités a posteriori dont chacune des composantes se rapporte à l'une des classes connues. La règle de décision par le maximum de probabilité a posteriori semble la plus naturelle pour atteindre les meilleures performances. Mais l'application d'une telle règle de décision suppose que l'on aie entièrement confiance dans les probabilités a posteriori produites par le classifieur. Or, dans la plupart des cas, l'on ne peut raisonnablement pas avoir une confiance totale dans le classifieur. Notre processus de décision devrait donc judicieusement tenir compte de la confiance (ou croyance) que nous mettons dans les résultats produits par le classifieur. Cette croyance doit d'abord être mesurée de sorte que l'on puisse en avoir une valeur numérique utilisable dans les calculs. La théorie de Dempster-Shafer fournit un cadre puissant de mesure de la croyance en proposant des concepts permettant de modéliser l'information imparfaite.

L'imperfection de l'information englobe l'imprécision, l'incertitude et l'incomplétude. A travers le concept de *fonction de masses de croyance*, la théorie de Dempster-Shafer nous permet de modéliser ces trois dimensions de l'information imparfaite, reflétant ainsi notre confiance dans la source d'information. Une fonction de masses de croyance (ou fonction de masses tout court) est une fonction tabulaire qui partage une quantité unitaire entre tous les sous-ensembles de l'ensemble des classes connues, contrairement à une fonction de probabilités qui, elle, partage la quantité unitaire entre les classes (sous-ensembles singletons). La valeur attribuée à un sous-ensemble de classe correspond à la masse de croyance en ce sous-ensemble. La principale difficulté dans cette théorie réside dans la manière de distribuer les masses de croyance. Les méthodes disponibles dans la littérature ont été dévoloppées dans un contexte de fusion de classifieurs, ce qui suppose que leur application nécessite d'avoir au moins deux classifieurs. Dans cet article, nous proposons une méthode de calcul des valeurs de masse de croyance qui répond aux deux préoccupations suivantes :

– prendre en compte toute information disponible a priori succeptible d'influencer la confiance que l'on pourrait avoir dans le classifieur,
– pouvoir appliquer la méthode pour améliorer les performances de n'importe quel classifieur unique donné, hors du contexte de fusion de classifieurs.

Notre approche est schématisée sur la figure 1. Le classifieur donné reçoit en entrée un vecteur de caractéristiques $\mathbf{x}$ et produit en sortie un vecteur de probabilités $\mathbf{y}$. Ce vecteur de probabilités est ensuite transformé par notre méthode en un vecteur de masses de croyance $\mathbf{z}$ sur lequel va se baser la décision d'assignation de l'objet S à l'une des classes connues, ou éventuellement la décision de rejet.

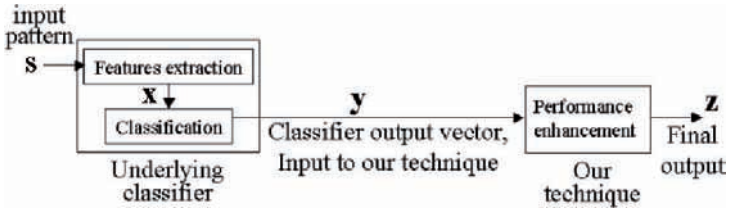

FIG. 1 – *Schéma de la transformation d'un vecteur de probabilités a posteriori en un vecteur de masses de croyance pour une amélioration des performances en classification supervisée.*

Outre la fonction de masses, la théorie de Dempster-Shafer propose également un certain nombre de règles de décision. Seule la règle du maximum de masse est utilisée dans l'approche que nous développons ici. Rappelons qu'avec la règle du maximum de probabilité à posteriori, un rejet n'a lieu que s'il n'y a pas de maximum unique ou si le maximum n'atteint pas un certain seuil, et l'on est dans l'impossibilité de quantifier la confiance que l'on a dans ce rejet. Non seulement notre approche permet de réduire le taux d'erreur, elle offre également l'opportunité de mesurer notre confiance en cas de rejet. Cette confiance correspond à la valeur de croyance attribuée à l'ensemble complet des classes.

Après avoir rappelé dans la section 2 la définition de la notion de fonction de masses, nous discutons dans la section 3 quelques méthodes proposées dans la littérature pour le calcul des valeurs de masse de croyance. Dans la section 4, nous détaillons notre technique puis la testons

dans la section 5 avec six classifieurs différents dans le domaine de la classification en stades de sommeil. Nous discutons nos résultats dans la section 6 et concluons dans la section 7.

2 Fonction de masses de croyance

La notion de fonction de masses est l'une des notions introduites par la théorie de l'évidence de Dempster-Shafer généralisant la théorie des probabilités dans le cas des variables discrètes. Dans cette section, nous donnons la définition formelle de cette fonction avant d'introduire la manière dont nous l'utilisons pour améliorer les performances d'un classifieur donné. Le lecteur est invité à se référer aux travaux originaux de Shafer (Shafer, 1976) pour un discours détaillé sur la théorie de l'évidence.

Soit $\Theta = \{\theta_1, \theta_2, \cdots, \theta_N\}$ un ensemble de N classes. On note 2^Θ l'ensemble des sous-ensembles de Θ. On définit sur 2^Θ une fonction de masses m de la manière suivante :

$$\left\{ \begin{array}{lcll} 0 & \leq & m(\mathcal{A}) \leq 1 & \forall \mathcal{A} \in 2^\Theta / \emptyset \quad \text{et } \sum_{\mathcal{A} \in 2^\Theta} m(\mathcal{A}) = 1 \\ 0 & = & m(\emptyset) & \end{array} \right. \tag{1}$$

$m(\mathcal{A})$ est une mesure de la confiance que l'on est disposé à accorder exactement au sous-ensemble $\mathcal{A}$ étant donné l'information disponible. Autrement dit $m(\mathcal{A})$ correspond à la part de croyance qui soutient l'hypothèse composite $\mathcal{A}$ sans soutenir aucun sous-ensemble strict de $\mathcal{A}$. Contrairement à la théorie des probabilités, où la quantité unitaire est divisée entre plusieurs hypothèses atomiques $\{\theta_i\}$, en théorie de l'évidence, pour toute hypothèse composite $\mathcal{A}$ tel que $m(\mathcal{A}) > 0$, $m(\mathcal{A})$ reflète une certaine ignorance, car elle représente une croyance qui ne peut être subdivisée en sous-ensembles plus fins. Ceci conduit à l'inégalité suivante : $m(\mathcal{A}) + m(\overline{\mathcal{A}}) \leq 1$, où $\overline{\mathcal{A}}$ est le complément de $\mathcal{A}$ dans Θ. Cette inégalité contraste avec le postulat d'additivité dans la théorie des probabilités. Si $m(\mathcal{A}) > 0$ alors $\mathcal{A}$ est appelé *élément focal* de m.

Comme indiqué dans l'introduction, les composantes du vecteur de sortie $\mathbf{z}$ dans notre méthode (voir figure 1) sont considérées comme étant les valeurs d'une fonction de masses m. Cette fonction de masses a $N + 1$ éléments focaux, N étant le nombre total de classes dans l'application de classification supervisée.

$$\left\{ \begin{array}{rcl} z_i & = & m(\{\theta_i\}) \quad \forall i \in \{1, \cdots, N+1\} \\ \{\theta_{(N+1)}\} & = & \Theta \qquad \text{par convention} \end{array} \right. \tag{2}$$

où z_i représente la $i^{\text{ème}}$ composante du vecteur $\mathbf{z}$. Les éléments focaux de m sont $\{\theta_1\}$, $\{\theta_2\}$, $\cdots$, $\{\theta_N\}$ et Θ. Chaque valeur z_i est la masse de croyance accordée à la classe θ_i étant donné le vecteur de probabiltés $\mathbf{y}$ produit par le classificateur. La règle de décision par le maximum de masse de croyance s'énonce :

$$\text{affecter l'objet S à } \theta_i,\ i \in \{1, \cdots, N+1\},\ \text{ si } z_i > z_j,\ \forall j \in \{1, \cdots, N+1\} \ \text{ et } \ j \neq i \tag{3}$$

S'il n'y a pas de composante z_i de $\mathbf{z}$ telle que $z_i > z_j\ \forall j \in \{1, \cdots, N+1\}$ et $j \neq i$ (par exemple lorsque deux composantes différentes atteignent une valeur maximale), la décision est alors un rejet. L'affectation de l'objet S à $\theta_{(N+1)}$ correspond également à un rejet puisque $z_{(N+1)}$ représente la masse de croyance associée à l'ensemble complet Θ. Avant de détailler notre approche, nous allons discuter quelques méthodes de calcul de la fonction de masses que l'on rencontre dans la littérature.

3 Méthodes de calcul de la fonction de masses

La stratégie que nous développons dans cet article pour le calcul des valeurs de masse de croyance peut s'appliquer dans deux contextes différents :

contexte 1 (classifieur unique) : l'on souhaite améliorer les performances d'un classifieur quelconque donné dont la sortie est un vecteur de probabilités a posteriori ;

contexte 2 (fusion de classifieurs) : l'on dispose de $K(\geq 2)$ classifieurs et l'on cherche à exploiter simultanément les K réponses différentes.

L'originalité de notre approche réside dans le fait que toutes les méthodes existantes pour le calcul des valeurs de masse de croyance dans des applications de classification ont été conçues pour la fusion de plusieurs classifieurs. Elles ne s'appliquent donc pas telles quelles, contrairement à notre méthode, dans le contexte d'un classifieur unique dont on voudrait améliorer les performances.

Dans la méthode de Xu (Xu et al., 1992), pour chacun des K classifieurs à fusionner, les taux de reconnaissance, d'erreur et de rejet (respectivement ϵ_r, ϵ_s et $1-\epsilon_r-\epsilon_s$) sont assimilés à des masses de croyance et la fonction de masses m est définie comme suit :

- si la décision par le maximum de probabilité a posteriori mène à un rejet alors $m(\Theta)$=1,
- si la décision par le maximum de probabilité a posteriori mène au choix de la classe θ_n alors m distribue la quantité unitaire sur trois élements focaux $\{\theta_n\}$, $\overline{\{\theta_n\}}$, et Θ de la manière suivante : $m(\{\theta_n\})=\epsilon_r$, $m(\overline{\{\theta_n\}})=\epsilon_s$, et $m(\Theta)=1-\epsilon_r-\epsilon_s$

Cette méthode de calcul des valeurs de masse de croyance pourrait bien s'appliquer dans le contexte d'un classifieur unique mais elle n'est guère théoriquement satisfaisante. Si les taux de reconnaissance, d'erreur et de rejet peuvent sans difficulté être considérés comme des probabilités, ils ne peuvent être assimilés à des masses de croyance que sous certaines conditions puisque la croyance dans les réponses d'un classifieur est nécessairement affectée par la confiance que l'on fait au classifieur. Les masses de croyance et les probabilités ne coïncident que lorsque l'on est tout à fait confiant quant aux réponses du classifieur.

La méthode de Rogova (Rogova, 1994) inclut une étape d'apprentissage pour la fusion des classifieurs f^k, $k \in \{1, \cdots, K\}$. Cet apprentissage consiste à calculer, pour chaque paire (f^k, θ_n), un vecteur de référence $\mathbf{E}_n^k$ comme étant la moyenne des vecteurs de sortie lorsque f^k est appliqué sur les données d'apprentissage appartenant à la classe θ_n. On considère ensuite une fonction de proximité ϕ qui varie entre 0 et 1, et n'atteint la valeur de 1 que lorsque ses deux arguments sont égaux. Pour une donnée inconnue $\mathbf{x}$ quelconque, l'ensemble des valeurs prises par $\phi(\mathbf{E}_n^k, f^k(\mathbf{x}))$ permet de générer un ensemble de fonctions de masses dont la combinaison (à l'aide d'opérateurs proposés par la théorie de l'évidence) résulte en une autre fonction de masses m. La décision de classification est alors basée sur m.

La technique proposée par Denoeux (Denoeux, 2000) pour le calcul des valeurs de masse de croyance ne s'inscrit ni dans le contexte 1 ni dans le contexte 2 puisqu'elle ne part pas d'un (ou plusieurs) classifieur(s) existant(s). C'est une technique d'apprentissage qui permet de construire un classifieur en utilisant des notions (dont celle de fonction de masses) introduites par la théorie de l'évidence. Denoeux propose d'extraire à partir de l'ensemble d'apprentissage, J vecteurs prototypes $\mathbf{p}^1, \cdots, \mathbf{p}^J$, un prototype $\mathbf{p}^j$ étant un vecteur situé dans l'espace des caractéristiques et dont on connaît la probabilité u_n^j d'appartenir à la classe θ_n. Pour une donnée inconnue, chaque prototype $\mathbf{p}^j$ génère une fonction de masses m^j à $N+1$ éléments focaux : $m^j(\Theta)=1-\alpha^j \exp(-\gamma^j \|\mathbf{x}-\mathbf{p}^j\|^2)$, $m^j(\{\theta_n\})=\alpha^j u_n^j \exp(-\gamma^j \|\mathbf{x}-\mathbf{p}^j\|^2)$, $\forall n$,

où $\mathbf{x}$ est le vecteur de caractéristiques extrait de la donnée inconnue, et γ^j et α^j sont des paramètres associés au prototype $\mathbf{p}^j$. Le paramètre α^j varie entre 0 et 1, et $\gamma^j > 0$. La combinaison des J fonctions m^j résulte en une autre fonction de masses m sur laquelle se base la décision de classification.

La méthode d'Al-Ani (Al-Ani et Deriche, 2002) est similaire à celle de par Rogova. La principale différence entre les deux méthodes réside dans la manière de déterminer les vecteurs de référence. Al-Ani propose de les contruire par apprentissage alors que Rogova les calculait par moyennage. A l'instar de la méthode de Rogova, la méthode d'Al-Ani définit une approche de fusion de classifieurs et nécessite au moins deux classificateurs pour s'appliquer.

Dans un contexte de fusion de classifieurs f^k à sorties nominales, la méthode d'Appriou (Appriou, 2002) propose d'extraire d'abord, sous la forme de fonctions de probabilités conditionnelles $p(f^k/\theta_n)$, de la connaissance au sujet de chacun des K classifieurs. Un facteur de confiance $q_n^k \in [0,1]$ est associé à chaque fonction de probabilités conditionnelles et reflète notre confiance dans l'estimation. Si $p(f^k/\theta_n)$ est parfaitement représentative de la population des données concernant la paire (f^k, θ_n) alors $q_n^k = 1$. Par la suite, lorsque pour une donnée inconnue le classifieur f^k produit le label θ_c, N fonctions de masses m_n^k ayant chacune trois éléments focaux sont calculées :

$$\begin{aligned} m_n^k(\{\theta_n\}) &= q_n^k \times \left[R^k \times p(f^k{=}\theta_c/\theta_n)\right] / \left[1 + R^k \times p(f^k{=}\theta_c/\theta_n)\right] \\ m_n^k(\overline{\{\theta_n\}}) &= q_n^k / \left[1 + R^k \times p(f^k{=}\theta_c/\theta_n)\right] \\ m_n^k(\Theta) &= 1 - q_n^k \end{aligned}$$

où $R^k \geq 0$ est un facteur de normalisation. Notons qu'Appriou a également proposé un modèle de fonction de masses ayant seulement deux éléments focaux. La combinaison (grâce à la règle de la somme orthognonale de Dempster) de toutes les fonctions de masses m_n^k ainsi calculées résulte en une autre fonction de s m. La décision de classification est alors basée sur m.

4 Détail de l'approche proposée

4.1 Une procédure en deux étapes

Dans la suite de cet article, nous désignerons par *hypothèses* les $N+1$ éléments focaux dans (2). En considérant que chaque hypothèse est représentée par un vecteur de référence à N composantes, le vecteur de sortie $\mathbf{z}$ dans notre méthode (voir figure 1) peut être obtenu par un calcul en deux étapes, étant donné un objet inconnu S :

Etape 1 : Pour chaque hypothèse $\{\theta_j\}$, la dissimilarité a_j entre son vecteur de référence $\mathbf{r}_j$ et le vecteur $\mathbf{y}$ correspondant à S est mesurée grâce à une fonction de distance ϕ (par exemple la distance euclidienne) :

$$a_j = \phi(\mathbf{y}, \mathbf{r}_j)$$

Une fonction monotone décroissante g (telle que l'exponentielle décroissante) permet de mesurer la proximité $g(a_j)$ entre $\mathbf{r}_j$ et $\mathbf{y}$.

Etape 2 : La normalisation des valeurs de proximité obtenues à l'étape 1 conduit au vecteur $\mathbf{z}$, interprété comme une fonction de masses de croyance se rapportant à l'information véhiculée par $\mathbf{y}$:

$$z_i = m(\{\theta_i\}) = \frac{g(a_i)}{\sum_{j=1}^{N+1} g(a_j)}$$

Notons que les composants de $\mathbf{z}$ satisfont à l'expression (1).

4.2 Implémentation connexionniste

Notre approche exposée ci-dessus présente quelques similitudes avec les réseaux à fonctions radiales de base (Ghosh et Nag, 2001) qui sont des réseaux de neurones composés d'une couche d'entrée, une couche cachée et une couche de sortie. Etant donné un vecteur d'entrée, la réponse d'une unité cachée est définie comme étant une fonction décroissante de la distance entre le vecteur d'entrée et un vecteur relatif à l'unité considérée. La réponse d'une unité de sortie est définie comme étant la somme pondérée des réponses des unités cachées.

A l'instar des réseaux à fonctions radiales de base, la technique que nous proposons peut également être représentée dans le formalisme connexionniste avec une seule couche cachée (voir la figure 2). La couche cachée (couche numéro 1) et la couche de sortie (couche numéro 2) correspondent respectivement aux étapes 1 et 2 de la procédure décrite ci-dessus. Il y a $N+1$ unités dans chacune de ces couches.

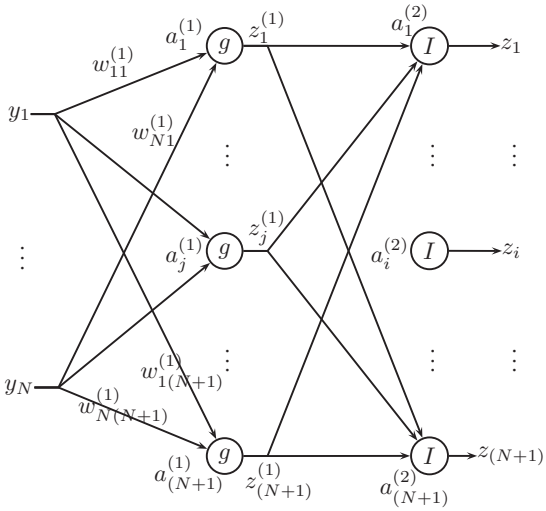

FIG. 2 – *Schéma connexionniste de notre méthode, avec N unités dans la couche d'entrée, $N+1$ unités dans la couche cachée et $N+1$ unités dans la couche de sortie.*

y_n représente la $n^{\text{ème}}$ composante ($n \in \{1, \cdots, N\}$) du vecteur d'entrée $\mathbf{y}$ produit par le classifieur dont nous voulons améliorer les performances. Ce vecteur satisfait à la contrainte suivante (puisqu'il s'agit d'un vecteur de probabilités) :

$$0 \leq y_n \leq 1 \quad \forall n \in \{1, \cdots, N\} \quad \text{et} \sum_{n=1}^{N} y_n = 1 \tag{4}$$

Dans les couches numéro 1 et 2, la $n^{\text{ème}}$ unité est associée à l'hypothèse $\{\theta_n\}$ et l'unité $N+1$ est associée à l'hypothèse Θ. On note $a_j^{(l)}$, $j \in \{1, \cdots, N+1\}$ et $l \in \{1, 2\}$, la

valeur d'activation de l'unité j dans la couche l. Une même fonction d'activation g est utilisée dans toutes les unités cachées. La fonction d'activation utilisée dans les unités de sortie est la fonction identité I. La réponse de l'unité i dans la couche l est notée $z_i^{(l)}$.

Dans la couche numéro 1, le paramètre $w_{nj}^{(1)}$ correspond à la $n^{\text{ème}}$ composante du vecteur de référence $\mathbf{w}_j^{(1)}$ de l'hypothèse $\{\theta_j\}$. La valeur d'activation $a_j^{(1)}$ qui mesure la dissimilarité entre $\mathbf{y}$ et $\mathbf{w}_j^{(1)}$ est fournie par le carré de la distance euclidienne entre les deux vecteurs.

$$a_j^{(1)} = \|\mathbf{y} - \mathbf{w}_j^{(1)}\|^2 = \sum_{n=1}^{N} (y_n - w_{nj}^{(1)})^2 \quad \forall j \in \{1, \cdots, N+1\} \tag{5}$$

La réponse $z_j^{(1)}$ à une activation $a_j^{(1)}$ mesure la proximité entre les deux vecteurs $\mathbf{y}$ et $\mathbf{w}_j^{(1)}$. Cette proximité est supposée directement proportionnelle à la croyance qui supporte l'hypothèse $\{\theta_j\}$. Elle est maximale lorsque $\mathbf{y} = \mathbf{w}_j^{(1)}$ (i.e. $a_j^{(1)} = 0$) et décroît lorsque $\mathbf{y}$ s'éloigne de $\mathbf{w}_j^{(1)}$. Un tel comportement peut être engendré par une fonction exponentielle décroissante de la dissimilarité entre les deux vecteurs.

$$z_j^{(1)} = g(a_j^{(1)}) = \exp(-a_j^{(1)}) \quad \forall j \in \{1, \cdots, N+1\} \tag{6}$$

La couche numéro 2 permet de normaliser les valeurs de proximité $z_j^{(1)}$. Ces valeurs normalisées correspondent aux composantes z_i du vecteur de sortie z :

$$z_i = z_i^{(2)} = a_i^{(2)} = \frac{z_i^{(1)}}{\sum_{j=1}^{N+1} z_j^{(1)}} \quad \forall i \in \{1, \cdots, N+1\} \tag{7}$$

4.3 Apprentissage des vecteurs de référence

Soit Err le critère d'erreur :

$$Err = \sum_{p=1}^{P} E_p \quad \text{avec} \quad E_p = \|\mathbf{z}_p - \mathbf{t}_p\|^2 \tag{8}$$

où P est le nombre total de données dans l'ensemble d'apprentissage, $\mathbf{z}_p$ le vecteur de masses de croyance produit par notre méthode pour la donnée d'apprentissage p, et $\mathbf{t_p}$ le vecteur de masses désiré pour p. Les composantes de $\mathbf{t_p}$ étant inconnues, nous proposons le raisonnement suivant pour leur calcul : nous considérons les sorties désirées comme étant les valeurs de masse de croyance que l'on devrait obtenir si l'on était dans un "monde parfait", i.e. un monde où le classifieur serait parfait et ne ferait aucune erreur. Dans ce cas l'on ferait entièrement confiance au classifieur puisque pour tout objet appartenant à la classe θ_c, $c \in \{1, \cdots, N\}$, le classifieur produirait un vecteur $\mathbf{y}$ ayant la forme

$$y_c = 1 \quad \text{et} \quad y_n = 0 \quad \forall n \neq c \tag{9}$$

et les composantes des vecteurs de référence auraient la forme

$$\begin{cases} w_{nj}^{(1)} = 1 & \text{si} \quad n = j \\ w_{nj}^{(1)} = 0 & \text{si} \quad n \neq j \quad \text{et} \quad j \neq (N+1) \\ w_{n(N+1)}^{(1)} = \alpha & \forall n \in \{1, \cdots, N\} \end{cases} \tag{10}$$

où $\alpha \in \mathbb{R}$.

Afin de déterminer les valeurs appropriées de α, on observe que sous l'hypothèse du monde parfait, le vecteur de référence du rejet $\mathbf{w}_{(N+1)}^{(1)}$ (dont toutes les composantes valent α) est à égale distance (euclidienne) de tout autre vecteur de référence $\mathbf{w}_j^{(1)}$. En notant β le carré de cette distance, on a :

$$\beta = \|\mathbf{w}_{(N+1)}^{(1)} - \mathbf{w}_j^{(1)}\| = (y_c - \alpha)^2 + \sum_{\substack{n=1 \\ n \neq c}}^{N} (y_n - \alpha)^2$$

En remplaçant y_n par ses valeurs indiquées dans (9), on obtient :

$$\beta = N\alpha^2 - 2\alpha + 1$$

Le paramètre β ainsi calculé se retrouvera dans les expressions des composantes des vecteurs $\mathbf{t}_p$ d'où l'on pourra déduire les valeurs appropriées de α.

En phase d'apprentissage et sous l'hypothèse du monde parfait, pendant que les paramètres $w_{nj}^{(1)}$ gardent leurs formes indiquées dans (10), lorsque l'on propage un vecteur $\mathbf{y}$ (voir (9)) dans le réseau et que l'on calcule les valeurs d'activation et de sortie grâce aux équations (5), (6) et (7), le vecteur de sortie $\mathbf{z}$ obtenu correspond au vecteur de masses désiré $\mathbf{t}_p$. Ainsi, pour une donnée d'apprentissage p appartenant à la classe θ_c, on obtient (en propageant le $\mathbf{y}$ fourni par le classifieur pour la donnée p) :

$$\left\{ \begin{array}{rcl} t_c & = & z_c = \exp(\beta+2)/den \\ t_i & = & z_i = \exp(\beta)/den \quad \forall i \neq c \text{ et } i \neq (N+1) \\ t_{(N+1)} & = & z_{(N+1)} = \exp(2)/den \end{array} \right. \tag{11}$$

où $den = \exp(\beta+2) + \exp(2) + (N-1)\exp(\beta)$, et t_j représente la $j^{\text{ème}}$ composante du vecteur de masses désiré $\mathbf{t}_p$. L'on peut alors déduire les valeurs appropriées de α en considérant que $\mathbf{t}_p$ étant une fonction de masses sensée représenter une situation parfaite, ses valeurs doivent satisfaire les inéquations suivantes :

$$t_{(N+1)} < t_i < t_c \quad \forall c, i \in \{1, \cdots, N\}, \quad i \neq c \tag{12}$$

En résolvant les inéquations (12) on trouve : $\alpha \in]-\infty, A[\cup]B, +\infty[$
où $A = (1-\sqrt{1+N})/N$ et $B = (1+\sqrt{1+N})/N$. En observant que $-1 < A \quad \forall N > 0$, on déduit que $\alpha = -1$ est un choix approprié puisque N est nécessairement positif. L'on pourra donc remplacer β par $N+3$ dans (11).

Afin de trouver les valeurs optimales de $w_{nj}^{(1)}$ en minimisant le critère d'erreur, nous opérons une descente de gradient stochastique à partir de l'équation 8. Cela nous amène à évaluer la dérivée suivante :

$$\frac{\partial E_p}{\partial w_{nj}^{(1)}} = 2 \times \delta_j^{(1)} \times \left(w_{nj}^{(1)} - y_n \right) \tag{13}$$

avec

$$\delta_j^{(1)} = \left(\delta_j^{(2)} f_1(z_j^{(1)}) + \sum_{j \neq i = 1}^{N+1} \delta_i^{(2)} f_2(z_i^{(1)}) \right) f_3(a_j^{(1)})$$

où

$$\delta_i^{(2)} = 2 \times (a_i^{(2)} - t_i)$$

et les fonctions f_1, f_2 et f_3 définies comme suit :

$$f_1(z_j^{(1)}) = \frac{\sum_{l=1}^{N+1} z_l^{(1)} - z_j^{(1)}}{\left(\sum_{l=1}^{N+1} z_l^{(1)}\right)^2} \quad f_2(z_i^{(1)}) = \frac{-z_i^{(1)}}{\left(\sum_{l=1}^{N+1} z_l^{(1)}\right)^2} \quad f_3(a_j^{(1)}) = -\exp(-a_j^{(1)})$$

4.4 Généralisation

Une fois que les vecteurs de référence $\mathbf{w}_j^{(1)}$ sont appris $\forall j$, l'on peut calculer un vecteur de masses $\mathbf{z}$ pour chaque vecteur d'entrée $\mathbf{y}$ inconnu et prendre une décision de classification par la règle du maximum de masse de croyance (voir (3)).

5 Tests de l'approche proposée

5.1 Cotation en stades de sommeil

Dans les laboratoires d'analyse du sommeil humain, la cotation en stades de sommeil est une activité de classification des pages successives d'un enregistrement polysomnographique. Un enregistrement polysomnographique est constitué de signaux bio-électriques tels que des électro-encéphalogrammes (EEG), des électro-occulogrammes (EOG), et des électro-myogrammes (EMG), enregistrés tout au long de la nuit (plus ou moins 8 heures) à l'aide de capteurs positionnés sur le corps d'un patient. Afin de procéder à la cotation en stades, l'enregistrement polysomnographique est logiquement segmenté en plusieurs pages successives de 30 secondes. La figure 3 illustre une page où les deux premiers signaux sont des sections de signaux EEG, les troisième et quatrième sont des sections de signaux EOG et le dernier est une section d'un signal EMG.

La cotation en stades de sommeil consiste à identifier, pour chaque page, le stade de sommeil correspondant parmi les 5 stades prédéfinis : l'éveil (A), le sommeil paradoxal (REM), le stade 1 (S1), le stade 2 (S2) et le stade 3 (S3). L'exposé de l'application de notre approche au problème de la cotation en stades de sommeil ne nécessite pas d'expliquer en détail chacun de ces stades. Il suffit de noter que ces stades correspondent aux classes (donc $N = 5$) et que les pages successives correspondent aux objets à classer.

Afin de tester notre approche dans plusieurs cas différents, six classifieurs ont été développés indépendamment de notre technique d'amélioration de performance. Chacun de ces six classifieurs produit en sortie un vecteur de probabilités a posteriori qui deviendra l'entrée de notre méthode.

5.2 Ensembles de données d'apprentissage et de test

Nous avons disposé d'un ensemble de données d'apprentissage $\mathcal{E}$ contenant 47 enregistrements polysomnographiques fournissant un total de 48579 pages classées par un expert humain du domaine. Ces 47 enregistrements ont été partagés en 3 sous-ensembles :

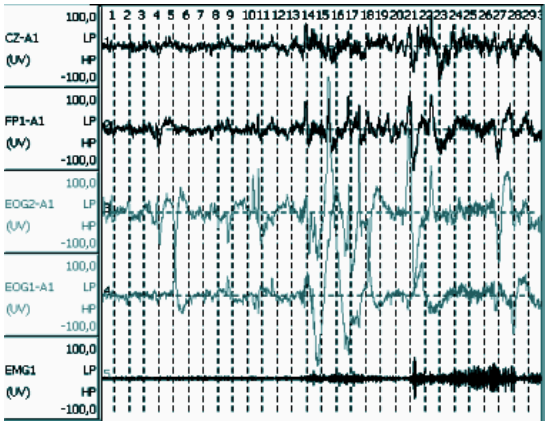

FIG. 3 – *Une page de 30 secondes.*

- le sous-ensemble $\mathcal{E}_1$ servait au développement des classifieurs indépendamment de notre méthode ;
- le sous-ensemble $\mathcal{E}_2$ servait à l'apprentissage de notre méthode ;
- le sous-ensemble $\mathcal{E}_3$ servait au test.

Les sous-ensembles $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$ forment ainsi une partition de $\mathcal{E}$. Afin de vérifier l'efficacité de notre technique dans différents scénarios, nous avons aléatoirement généré trois partitions $\mathcal{P}_1$, $\mathcal{P}_2$ et $\mathcal{P}_3$, chacune constituée des sous-ensembles $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$.

6 Résultats et discussion

$N = 5$ dans le problème de la cotation en stades de sommeil. Nous avons également déjà établi que $\alpha = -1$ est un choix approprié $\forall N > 0$. Nous implémentons l'algorithme suivant pour générer des graphiques de comparaison de performances :

- pour chaque partition $\mathcal{P}_i$ ($i = 1, \cdots, 3$)
  - fixer $N = 5$ et $\alpha = 1$, et utiliser le sous-ensemble $\mathcal{E}_2$ pour entraîner notre algorithme d'amélioration de performance
  - pour chaque classifieur (préalablement entraîné sur le sous-ensemble $\mathcal{E}_1$)
    - pour chaque page appartenant au sous-ensemble test $\mathcal{E}_3$
      - calculer le vecteur $\mathbf{y}$, prendre une décision sur base de la règle du maximum de probabilités a posteriori et comparer la classe obtenue à la vraie classe de la page
      - à partir de $\mathbf{y}$, calculer le vecteur $\mathbf{z}$, prendre une décision sur base de la règle du maximum de masse de croyance et comparer la classe obtenue à la vraie classe de la page
    - end
    - dessiner un graphe ayant en abscisse les patients de $\mathcal{E}_3$ et en ordonnées les performances du classifieur avant et après application de notre méthode.
  - end

– end

Un total de 18 graphes de comparaison (3 partitions $\times$ 6 classifieurs) ont ainsi été générés. Ces 18 graphes sont tout à fait similaires à celui de la figure 4. Il y apparaît clairement qu'en ce qui concerne le problème de la classification en stades de sommeil, l'application de notre technique permet systématiquement d'améliorer les performances en réduisant le taux d'erreur. Cette amélioration des performances peut être étonnamment élevée. Par exemple, pour le patient

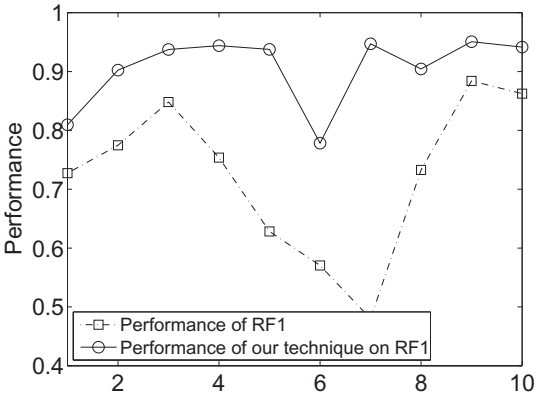

FIG. 4 – *Graphe de comparaison des performances avant et après application de notre méthode.*

représenté par la valeur 7 en abscisse, le taux de reconnaissance des pages est passé de moins de 50% (avant application de notre méthode) à plus de 90% (après application de notre méthode). Un mauvais résultat avant application de notre méthode ne nuit donc pas nécessairement au résultat final.

7 Conclusion et perspectives

En classification supervisée, l'utilisation de la règle de décision du maximum de masse de croyance en lieu et place de celle du maximum de probabilité a posteriori permet de réduire le taux d'erreur de classification. Dans cet article, une technique permettant de transformer un vecteur de probabilités a posteriori en un vecteur de masses de croyance (dans le cadre de la théorie de l'évidence de Dempster et Shafer) à été proposée et testée dans le domaine de la cotation en stades de sommeil. Nous avons pu observer une évolution du taux de reconnaissance pouvant aller de 50% à 90%, soit une réduction du taux d'erreur de 80%.

Il est à noter que notre méthode ne fait usage d'aucune caractéristique propre au domaine d'application et est totalement indépendante des classifieurs utilisés. Dans un prochain article, nous en testerons l'efficacité dans d'autres problèmes de classification supervisée.

Références

Al-Ani, A. et M. Deriche (2002). A New Technique for Combining Multiple Classifiers Using the Dempster-Shafer Theory of Evidence. *Journal of Artificial Intelligence Research 17*, 333–361.

Appriou, A. (2002). Discrimination Multisignal par la Théorie de l'Evidence. In Lavoisier (Ed.), *Décision et Reconnaissance de Formes en Signal*, pp. 219–257. 11, rue Lavoisier, 75008 Paris.

Denoeux, T. (2000). A Neural Network Classifier Based on Dempster-Shafer Theory. *IEEE Transaction on Systems, Man, and Cybernetics 30*(2), 131–150.

Ghosh, J. et A. Nag (2001). An Overview of Radial Basis Function Network. In R. Howlett et L. Jain (Eds.), *Basis Function Network*. Physica-Verlag.

Rogova, G. (1994). Combining the Results of Several Neural Network Classifiers. *Neural Networks 7*(5), 777–781.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton : Princeton University Press.

Xu, L., A. Krzyzack, et C. Y. Suen (1992). Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE Trans. Syst., Man, Cybern. 22*(3), 418–435.

Summary

We want to show that the performance of any given measurement level classifier can be enhanced when maximum posterior probability decision rule is replaced by maximum belief mass decision rule in the framework of the Dempster-Shafer theory of evidence. This shift in decision rule raises the need on a method for extracting class belief mass values from output posterior probabilities. The aim of this paper is to propose an effective method for calculating class belief mass values on which to base class assignment decision in order to improve the accuracy and reliability of any given measurement level classifier. The method can allow 80% reduction of misclassification error when applied to given measurement level classifiers in the automatic sleep stages scoring application domain.