Classification paramétrique multi-classes à croyance

Daniel ALSHAMAA[1], Farah MOURAD-CHEHADE[1], Paul HONEINE[2]

[1]Institut Charles Delaunay, ROSAS, LM2S, Université de Technologie de Troyes, UMR 6281, CNRS, Troyes, France

[2]LITIS Lab, Normandie Université, Université de Rouen, Rouen, France

daniel.alshamaa@utt.fr, farah.chehade@utt.fr, paul.honeine@univ-rouen.fr

Résumé – En classification paramétrique, l'estimation de la classe d'appartenance d'un échantillon non étiqueté nécessite une connaissance des distributions sous-jacentes des différentes classes. L'inconvénient majeur de cette approche est l'incertitude due à l'erreur de modélisation des échantillons d'apprentissage. Le présent article propose une approche prenant en compte les incertitudes. Pour ce faire, la méthode proposée repose sur la théorie des fonctions de croyance pour attribuer un poids de confiance à chaque classe pour tout nouvel échantillon. Cette approche génère une méthode de classification paramétrique à confiance pondérée pour des problèmes multi-classes. Les performances de la méthode proposée sont validées par des expérimentations sur des données réelles pour la localisation *Indoor* et pour la reconnaissance faciale.

Abstract – The aim of parametric classification is to predict the target class of a new sample, under the hypothesis of known fitted distribution. A major drawback of this approach is the uncertainty due to the imprecise modeling of the training samples. For this purpose, a belief functions framework is provided to take into account uncertainties. The proposed method investigates the belief functions theory to assign a confidence weight to each class for any new sample. This approach yields a confidence-weighted parametric classification method for multi-class problems. The performance of the proposed method is validated by experiments on real data for indoor localization and for facial image recognition.

1 Introduction

Le problème de classification vise à identifier la classe à laquelle un nouvel échantillon appartient, en utilisant un ensemble d'échantillon d'apprentissage. Les méthodes de classification se divisent en deux grandes familles, paramétriques et non paramétriques. Les méthodes paramétriques, dites aussi probabilistes, utilisent une hypothèse sur la distribution des échantillons de chaque classe ; le problème réside alors à estimer les paramètres de ces distributions et à déterminer à quelle classe les nouveaux échantillons ont le plus de chances d'appartenir. On y retrouve le classifieur bayésien naïf qui suppose une forte indépendance des hypothèses [1] et la régression logistique (multinomiale) qui repose sur l'estimation de la probabilité conditionnelle [2]. Dans la famille des méthodes non paramétriques, on retrouve les réseaux de neurones artificiels [3], les séparateurs à vaste marge (SVM) [4], mais aussi l'algorithme des k plus proches voisins [5].

Les méthodes paramétriques sont souvent plus simples, plus rapides, et nécessitent moins de données d'apprentissage que les méthodes non paramétriques. De plus, elles produisent des estimations plus précises lorsque les hypothèses sur les distributions des données sont correctes [6]. Toutefois, un obstacle majeur rencontré en classification paramétrique est la modélisation des échantillons d'apprentissage, ne menant jamais à un modèle exact. La non prise en compte de cette erreur introduira des imprécisions au niveau des résultats de la classification.

Le présent article propose une méthode de classification paramétrique pour des problèmes multi-classes. L'approche proposée utilise la théorie des fonctions de croyance pour remédier aux erreurs de modélisation des données d'apprentissage. Une sélection de caractéristiques est d'abord appliquée dans le but d'augmenter la capacité discriminatoire des données. Pour cela, les échantillons d'apprentissage de chaque classe sont ajustés à une distribution paramétrique. Cet ajustement consiste à identifier la distribution statistique la plus adaptée aux données parmi les distributions connues et à estimer ses paramètres en se basant sur les échantillons d'apprentissage. Un poids de confiance est ensuite attribué à chaque classe en prenant en compte l'erreur de modélisation. La méthode fournit enfin une croyance rangée par ordre décroissant concernant l'appartenance du nouvel échantillon à chacune des classes.

Cet article est organisé comme suit. La section suivante présente la méthode proposée, en décrivant la technique de sélection des caractéristiques et la mise en oeuvre de la théorie des fonctions de croyance pour la classification. Puis, la méthode est illustrée au travers d'expérimentations pour la localisation indoor et pour reconnaître des images faciales.

2 La classification pondérée

Le problème de classification à confiance pondérée est formulé comme suit. Soient

- $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ l'ensemble des échantillons d'apprentis-

sage, où $\mathbf{x}_i \in \mathbb{R}^p$ et p le nombre de caractéristiques ;

- $F = \{f_1, \ldots, f_p\}$ l'ensemble complet des p caractéristiques;
- $\{y_1, ..., y_m\}$ l'ensemble des m classes en compétition ;
- Q_{F',y_a} la distribution ajustée des échantillons de la classe y_a selon l'ensemble des caractéristiques $F' \subseteq F$.

L'objectif de la méthode proposée est de déterminer une fonction $\mathbf{w} : \mathbb{R}^p \to [0,1]^m$ telle que $\mathbf{w}(\mathbf{x}) = (\mathrm{w}(y_1), ..., \mathrm{w}(y_m))$, où $\mathrm{w}(y_i)$ désigne le poids associé à la classe y_i pour tout nouvel échantillon $\mathbf{x} \in \mathbb{R}^p$. Pour ce faire, une technique de sélection de caractéristiques permet d'abord de choisir les caractéristiques les plus indépendantes qui maximisent la capacité discriminatoire. Puis, la théorie des fonctions de croyance fournit un cadre pour la classification à confiance pondérée.

2.1 Sélection de caractéristiques

L'objectif de la sélection de caractéristiques est de déterminer le meilleur ensemble parmi les $2^p - 1$ sous-ensembles candidats de F qui satisfait les deux objectifs : la minimisation de l'erreur de classification et la réduction de la dépendance entre les caractéristiques [7].

D'une part, l'erreur de classification est inversement liée à la capacité discriminatoire des caractéristiques [8]. Soit $F' \subseteq F$ un sous-ensemble considéré. La capacité discriminatoire de F' est définie par

$$DisC(F') = \sum_{a=1}^{m} \sum_{b=1}^{m} D_{KL}(Q_{F',y_a} || Q_{F',y_b}), \qquad (1)$$

où $D_{KL}(Q_{F',y_a} || Q_{F',y_b})$ est la divergence de Kullback-Leibler [9] mesurée entre les distributions des échantillons appartenant aux classes y_a et y_b, en utilisant les caractéristiques de F'. L'erreur associée à F' est alors donnée par

$$\mathcal{E}(F') = 2^{-DisC(F')}. \qquad (2)$$

D'autre part, la dépendance est un facteur essentiel de la sélection pour avoir un sous-ensemble réduit. Le coefficient de corrélation multiple permet de mesurer le degré de dépendance d'une caractéristique par rapport aux autres. Le coefficient de la corrélation multiple d'une caractéristique f_j de F' par rapport aux autres éléments de $F' \setminus \{f_j\}$ est défini par

$$R_j^2 = c_j^T R_{xx,j}^{-1} c_j, \qquad (3)$$

où c_j est un vecteur colonne d'éléments $r_{f_i f_j}$ pour $f_i \in F' \setminus \{f_j\}$, $r_{f_i f_j} = \frac{cov(f_i, f_j)}{\sigma_{f_i} \sigma_{f_j}}$ étant la corrélation entre les caractéristiques f_i et f_j, tels que cov est la covariance entre les échantillons et σ l'écart-type, c_j^T est le vecteur transposé de c_j, et $R_{xx,j}^{-1}$ est l'inverse de la matrice d'éléments $r_{f_i f_{i'}}$ pour f_i et $f_{i'} \in F' \setminus \{f_j\}$. La dépendance $\mathcal{R}(F')$ entre tous les éléments de l'ensemble F' est la moyenne des coefficients de corrélation multiple pour tout $f_j \in F'$.

L'objectif de la sélection de caractéristiques est de trouver l'ensemble $F_\star \subseteq F$ tel que $\mathcal{E}(F_\star)$ et $\mathcal{R}(F_\star)$ sont simultanément minimisés. Une recherche exhaustive étant très coûteuse, un algorithme glouton est utilisé avec une stratégie d'élimination régressive. Commençant par l'ensemble complet des caractéristiques, on supprime successivement la caractéristique la moins utile. Soit F_k le sous-ensemble choisi à l'itération $k \geq 1$, avec $F_0 = F$ et le cardinal de F_k est égal à $|F_k| = p - k$. A chaque itération $k \geq 1$, tous les sous-ensembles de F_{k-1} ayant $p - k$ éléments sont considérés. Soit $F_k^{(\ell)}, \ell = 1, \ldots, p - k + 1$, le terme désignant ces sous-ensembles. On définit la fonction bi-objective $g_k\left(F_k^{(\ell)}\right)$ comme suit,

$$g_k(F_k^{(\ell)}) = \alpha \frac{\mathcal{E}(F_{k-1}) - \mathcal{E}(F_k^{(\ell)})}{\max(\mathcal{E}(F_{k-1}), \mathcal{E}(F_k^{(\ell)}))} + (1-\alpha) \frac{\mathcal{R}(F_{k-1}) - \mathcal{R}(F_k^{(\ell)})}{\max(\mathcal{R}(F_{k-1}), \mathcal{R}(F_k^{(\ell)}))}, \qquad (4)$$

où $\alpha \in [0,1]$ est le paramètre qui contrôle le compromis entre les deux fonctions objectives. L'ensemble $F_k^{(\ell)}$ qui a la plus grande valeur positive de $g_k(F_k^{(\ell)})$ est sélectionné à l'itération k. Si toutes les valeurs sont négatives, alors il n'y a plus d'amélioration possible, et donc l'ensemble final est $F_\star = F_{k-1}$. Le sous-ensemble optimal obtenu sera utilisé dans la suite pour la classification.

2.2 Fonctions de croyance

La théorie des fonctions de croyance (TFC) permet la fusion de données et la prise de décision en utilisant toute information disponible, même incertaine [10].

A Association des masses

Soient y une variables discrète prenant des valeurs dans $Y = \{y_1, \ldots, y_m\}$ et 2^Y l'ensemble de tous les sous-ensembles non vides de Y, i.e., $2^Y = \{\{y_1\}, \ldots, Y\}$. Le cardinal de 2^Y est égal a $2^{|Y|} = 2^m - 1$. Une fonction fondamentale de la TFC est la fonction de masse. Pour une source d'information $F_\star$, une fonction de masse $m_{F_\star}(\cdot) \colon 2^Y \to [0,1]$ satisfait :

$$\sum_{A \in 2^Y} m_{F_\star}(A) = 1. \qquad (5)$$

La masse $m_{F_\star}(A)$ attribuée à $A \in 2^Y$ représente la proportion des preuves, apportée par la source $F_\star$, indiquant que la variable appartient à A. Ayant une observation $\mathbf{x}$, la masse $m_{F_\star}(A)$ est calculée selon

$$m_{F_\star}(A) = \frac{Q_{F_\star,A}(\mathbf{x})}{\sum_{A' \in 2^Y, A' \neq \emptyset} Q_{F_\star,A'}(\mathbf{x})}, \quad A \in 2^Y. \qquad (6)$$

Il est à noter qu'une erreur de modélisation persiste en utilisant les distributions ajustées $Q_{F_\star,A}$. En prenant tous les sous-ensembles de Y et pas seulement les singletons, l'algorithme proposé profite de toutes les preuves disponibles, même si une erreur de modélisation existe concernant certain élément. On considère l'exemple de la Fig. 1, où $Y = \{y_1, y_2, y_3\}$. L'ensemble Y a $2^3 - 1$ sous-ensembles non vides. Seuls $\{y_1\}$, $\{y_2\}$, $\{y_3\}$, $\{y_1, y_2\}$, et $\{y_1, y_2, y_3\}$ sont représentés par leurs distributions. Un échantillon $x^{(1)}$ par exemple appartient plus

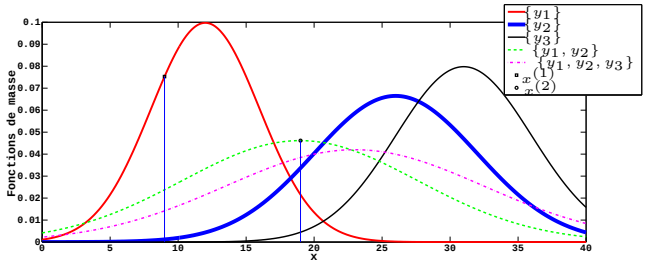

FIG. 1: Distribution des masses selon une source.

probablement à l'ensemble y_1, ce qui se traduit par une masse de $\{y_1\}$ plus grande que celle des autres en utilisant Eq. (6). Par ailleurs, pour l'échantillon $x^{(2)}$, les distributions de y_1 et y_2 sont trop proches, créant une sorte d'incertitude au niveau de l'affectation à l'un ou l'autre des singletons. En prenant le sous-ensemble $\{y_1, y_2\}$ en compte, une masse plus élevée est associée à $\{y_1, y_2\}$, ce qui permettra de couvrir cette incertitude et la corriger le plus probablement. Cet exemple illustre l'efficacité de l'utilisation des sous-ensembles non unitaires et motive l'emploi de la théorie des fonctions de croyance.

B Sélection de sous-ensembles

Il est en effet bénéfique de considérer les sous-ensembles de Y, toutefois les prendre tous demande un calcul prohibitif surtout lorsque le nombre des classes est grand. En conséquence, il est important de réduire leur nombre. Plus on monte en effectif dans un sous-ensemble, plus l'écart-type de la distribution est grande, et donc plus elle devient plate, étant donné que les échantillons des classes sont regroupés. En conséquence, un sous-ensemble A n'est pas pris en compte si

$$Q_{F_\star,A}(\mathbf{x}) \leq \max_{A' \in 2^Y,\ |A'| \leq |A|} Q_{F_\star,A'}(\mathbf{x}),\ \forall \mathbf{x}. \tag{7}$$

De plus, tous les sur-ensembles de A sont automatiquement non considérés. Dans l'exemple de la Fig. 1, le sous-ensemble $\{y_1, y_2\}$ sera gardé, tandis que le sous-ensemble $\{y_1, y_2, y_3\}$ n'est pas instructif parce que sa masse est toujours plus petite que celle de tous les autres sous-ensembles.

C Prise de la décision par la TFC

Une notion appropriée de la TFC pour prendre la décision est le niveau pignistique [11]. Il se définit comme suit,

$$BetP(A) = \sum_{A \subseteq A'} \frac{m_{F_\star}(A')}{|A'|}. \tag{8}$$

Le poids de confiance attribué à chaque classe peut se calculer,

$$\mathrm{w}(y_i) = BetP(\{y_i\}), i \in \{1, \ldots, m\}. \tag{9}$$

D Affaiblissement contextuel

Suite aux erreurs de modélisation, la source d'information n'est pas totalement fiable. Dans ce cas, il est possible de corriger ses masses au travers d'un affaiblissement classique ou contextuel. Ici, une technique d'affaiblissement contextuel est appliquée. L'idée principale est que la fiabilité d'une source peut varier en fonction de l'objet à reconnaître [12]. Pour ce faire, une matrice de confusion de taille $m \times m$ décrivant les performances de la méthode est d'abord définie, avec les éléments c_{ij} représentant le nombre d'échantillons de l'ensemble d'apprentissage appartenant réellement à la classe y_j mais classifiés dans la classe y_i. Ce nombre est calculé en appliquant la méthode de classification sur tous les échantillons et en comptabilisant les faux résultats. Le poids modifié ${}^{\beta}\mathrm{w}(y_i)$ d'une classe y_i est défini comme suit,

$${}^{\beta}\mathrm{w}(y_i) = \sum_{k=1}^{m} \frac{c_{ki}}{\sum_{j=1}^{m} c_{kj}} \mathrm{w}(y_k). \tag{10}$$

Ainsi, les masses calculées au préalable seront redistribuées selon la fiabilité des estimations. Ce calcul permet de contourner la nécessité d'un calcul réel, infaisable avec les distributions multidimensionnelles. Ces nouveaux poids adaptés seront utilisés pour la classification. En effet, la classe ayant le poids modifié le plus grand est alors sélectionnée, les autres étant rangées dans l'ordre décroissant de leur poids.

3 Expérimentations

Dans cette section, les performances de la méthode proposée sont étudiées sur deux domaines applicatifs distincts.

Localisation dans les réseaux de capteurs sans fil

L'objectif est de déterminer, en temps réel, la zone où se trouve un capteur mobile dans un environnement intérieur équipé d'un réseau de capteur sans fil (*e.g.*, bornes Wi-Fi conventionnelles). Des expérimentations ont été réalisées au département Recherche Opérationnelle, Statistiques Appliquées et Simulation de l'Université de Technologie de Troyes, France. La région considérée, d'environ 200 m^2, est divisée en onze zones (bureaux, salles et couloir). Douze points d'accès peuvent être détectés, et leurs puissances de signal (RSS) sont mesurées. Il s'agit d'un problème de classification à $m = 11$ classes et $p = 12$ caractéristiques pour chaque échantillon. Dans chaque zone, 45 mesures sont prises, dont 30 constituant l'ensemble d'apprentissage de chaque classe, et les autres pour le test. Chaque ensemble d'apprentissage est ajusté, selon un seuil de signification de 0.01, à une distribution d'une famille de distributions statistiques conventionnelles.

La technique de la sélection des caractéristiques est appliquée pour déterminer le meilleur sous-ensemble des caractéristiques. L'influence du paramètre α sur les performances en classification est examinée pour $\alpha \in \{0.75, 0.5, 0.25\}$. Le TAB. 1 montre l'influence du paramètre α au nombre de caractéristiques sélectionnés, la précision en classification, et le temps d'exécution. Une estimation est considérée correcte si l'algorithme attribue la plus grande confiance à la zone où se trouve le capteur. En comparant les résultats de la méthode

TAB. 1: Influence du paramètre de compromis α sur la dimensionalité, les précisions et le temps d'exécution.

Paramètre α	Nombre de caractéristiques sélectionnées	Précision (%)		Temps (s)	
		Apprentissage	Test	Apprentissage	Test
-	12	93.33	91.51	16.2	0.177
0.75	9	**97.57**	**95.75**	21.7	0.154
0.5	6	90.91	88.48	24.2	0.118
0.25	4	87.87	86.06	29.4	0.066

sans appliquer la technique de sélection des caractéristiques (première ligne du tableau), cette dernière peut améliorer les performances jusqu'à 95.75%.

Le TAB. 2 compare les précisions et les temps d'exécution de la méthode proposée à des méthodes de l'état de l'art en classification multi-classes. La méthode proposée surpasse les autres méthodes au niveau de la précision, avec un temps d'exécution compétitif.

Reconnaissance faciale

Les images de 10 individus de la base "Extended Yale B" [13] sont considérées selon diverses conditions d'éclairage. Pour l'extraction de caractéristiques, on considère la technique présentée dans [14] qui fournit une représentation de Gabor et la technique dans [15] qui produit une représentation parcimonieuse. Les caractéristiques de Gabor sont moins sensibles à la variation de l'illumination, l'expression, et les poses que les autres caractéristiques comme les Eigenface et Randomface. La moitié des images (32 par individu) est retenue pour l'apprentissage, et l'autre moitié pour le test. Toutes les images sont normalisées à 192×168. Le TAB. 3 compare les résultats de la méthode proposée avec les SVM, la représentation parcimonieuse (SRC), la classification régressive linéaire (LRC), les plus proches voisins (NN), et nearest subspace (NS) [14, 15]. Les résultats montrent que la méthode proposée surpasse les méthodes existantes.

4 Conclusion et perspectives

Cet article a présenté une méthode de classification paramétrique à confiance pondérée pour des problèmes multi-classes. La méthode a visé à diminuer l'erreur, réduire la dimensionalité des caractéristiques, et prendre en compte l'erreur de modélisation dans la prise de décision. Des expérimentations pour localiser un capteur mobile dans un réseau de capteurs sans fil, et pour reconnaître des images faciales ont montrées l'efficacité de la méthode proposée, surpassant les méthodes de l'état de l'art. Les futurs travaux se concentreront sur l'utilisation d'un modèle non paramétrique, comme l'estimation par noyau, pour l'attribution de masse dans les cas où le modèle paramétrique ne peut pas être utilisé. En outre, une approche hiérarchique sera considérée pour résoudre le problème de nombreuses classes.

TAB. 2: Comparaison de la méthode proposée aux méthodes de classification, en termes de précision et temps d'exécution.

	Précision (%)		Temps (s)	
Méthode	**Apprentissage**	**Test**	**Apprentissage**	**Test**
k plus proches voisins	82.22	77.78	3.2	0.115
Bayésien naïf	84.44	82.42	14.7	0.095
Régression logistique multinomiale	86.67	83.03	46.6	0.149
Réseaux de neurones	86.06	84.24	48.6	0.156
SVM	90.91	87.27	54.2	0.169
Méthode proposée	**97.57**	**95.75**	21.7	0.154

TAB. 3: La précision de la reconnaissance faciale (%) en utilisant les représentations de Gabor et parcimonieuse.

		Méthode de classification					
Représentation	**Dimension**	**NN**	**SRC**	**SVM**	**LRC**	**NS**	**Proposée**
Gabor	**56**	84.3	93.57	94.2	95.0	83.5	**96.6**
	120	91.8	96.4	96.7	95.7	86.3	**97.4**
	300	94.7	97.8	97.2	96.2	94.8	**98.3**
	504	94.8	98.3	97.7	96.6	95.2	**98.9**
Parcimonieux	**30**	75.1	79.9	81.7	83.4	81.8	**86.2**
	56	78.5	86.5	83.9	84.7	85.9	**89.5**
	120	83.4	95.3	90.3	90.8	93.1	**97.1**
	504	85.9	96.8	91.2	92.0	93.8	**98.7**

Références

[1] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive Bayes model," in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 194–201, Springer, 2013.

[2] D. Liu, T. Li, and D. Liang, "Incorporating logistic regression to decision-theoretic rough sets for classifications," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 197–210, 2014.

[3] R. Rojas, *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[4] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 6, 2012.

[5] R. Souza, L. Rittner, and R. Lotufo, "A comparison between k-optimum path forest and k-nearest neighbors supervised classifiers," *Pattern recognition letters*, vol. 39, pp. 2–10, 2014.

[6] T. Hoskin, "Parametric and nonparametric: Demystifying the terms," in *Mayo Clinic*, 2012.

[7] S. Tabakhi and P. Moradi, "Relevance–redundancy feature selection based on ant colony optimization," *Pattern recognition*, vol. 48, no. 9, pp. 2798–2811, 2015.

[8] O. S. Jahromi, *Multirate statistical signal processing*. Springer Science & Business Media, 2007.

[9] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[10] W. Zhang and Z. Zhang, "Belief function based decision fusion for decentralized target classification in wireless sensor networks," *Sensors*, vol. 15, no. 8, pp. 20524–20540, 2015.

[11] P. Smets, "Data fusion in the transferable belief model," in *Information Fusion, 2000. FUSIon 2000. Proceedings of the Third International Conference on*, vol. 1, pp. PS21–PS33, IEEE, 2000.

[12] D. Mercier, É. Lefèvre, and F. Delmotte, "Belief functions contextual discounting and canonical decompositions," *International Journal of Approximate Reasoning*, vol. 53, no. 2, pp. 146–158, 2012.

[13] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[14] M. Yang, L. Zhang, S. C. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with gabor occlusion dictionary," *Pattern Recognition*, vol. 46, no. 7, pp. 1865–1878, 2013.

[15] R. Khorsandi and M. Abdel-Mottaleb, "Classification based on weighted sparse representation using smoothed L^0 norm with non-negative coefficients," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 3131–3135, IEEE, 2015.