Apprentissage supervisé de règles de décision multiclasses avec contraintes de performances évolutives

Nisrine JRAD, Edith GRALL-MAËS, Pierre BEAUSEROY

Université de Technologie de Troyes ICD (FRE CNRS 2848), LM2S
12 rue Marie Curie, BP 2060, 10010 Troyes cedex - France
{nisrine.jrad,edith.grall,pierre.beauseroy}@utt.fr

Résumé – Des règles de décision multiclasses avec rejet sélectif et contraintes de performance évolutives sont déterminées à l'aide d'un apprentissage supervisé. Les distributions conditionnelles à chacune des classes sont sélectionnées au sein d'une famille d'estimateurs. La règle optimale minimisant un coût et respectant les contraintes évolutives est déterminée à chaque instant de modification de contraintes en utilisant les densités estimées indépendantes du temps. Une étude expérimentale sur un problème synthétique à deux dimensions et trois contraintes de performance variables, montre l'efficacité de l'approche et sa capacité à déterminer une règle de décision adaptative.

Abstract – This paper proposes a method using labeled data to learn a decision rule for multiclass problems with class-selective rejection and evolutionary performance constraints. The class-conditional densities are selected from a parameterized family of estimators by maximizing the likelihood criterion. The optimal decision rule, minimizing the loss function and satisfying each of the time evolutionary constraints, is determined using the time-independent densities. A two dimensional toy problem with three evolutionary constraints is carried out to study the efficiency of the proposed method.

1 Introduction

Les problèmes de classification consistent à trouver une règle de décision optimale au sens d'un critère défini selon le problème traité. Pour les problèmes multiclasses classiques, la règle de décision est élaborée par optimisation d'un risque qui n'est autre qu'une probabilité d'erreur ou une fonction coût. Les problèmes multiclasses avec contraintes permettent en plus d'exiger des performances sur la règle. Il s'agit par exemple de borner ou d'interdire les erreurs sur une classe donnée. À titre d'illustration, on peut citer le contrôle automatique de qualité où les contraintes sur les performances des produits peuvent varier au cours du temps selon plusieurs facteurs comme les conditions d'utilisation, la situation économique, les exigences de l'utilisateur ou les matières premières.

Les problèmes avec des contraintes fixées ont été traités dans [1]. Un formalisme général a été proposé et la solution dans le cadre des tests d'hypothèses statistiques a été déterminée. Elle repose sur la connaissance a priori des densités de probabilité et correspond au point selle d'une fonction risque définie par le dual lagrangien du problème d'optimisation initial de la fonction coût sous les contraintes exigées.

Dans un cadre supervisé, une approche possible consiste à trouver le point selle d'une estimée empirique du lagrangien en considérant une famille paramétrée d'estimateurs de densités de probabilité. L'approche présentée dans [2] consiste à optimiser les paramètres de la famille au sens du risque Lagrangien. Cette méthode est performante puisqu'elle satisfait au mieux les objectifs de la règle mais elle convient mal aux réévaluations nécessaires dans le cas des contraintes évolutives.

L'objectif de ce travail est de construire une règle de décision multiclasses qui soit capable de suivre les évolutions des contraintes. La méthode doit éviter de traiter chaque évolution des contraintes comme un problème entièrement nouveau afin de réduire la complexité du problème d'optimisation. Une approche envisageable consiste à découpler le problème d'apprentissage en deux étapes. La première cherche à optimiser les paramètres de la fonction d'estimation des probabilités conditionnelles au sein d'une famille de distribution considérée. Ces paramètres sont alors optimisés au sens d'un critère tel que la vraisemblance. La seconde consiste à optimiser la fonction risque déterminée pour chaque valeur des contraintes à partir des estimées retenues. Par la suite, l'évolution des contraintes n'amène pas à réoptimiser les paramètres de l'estimateur. La famille d'estimateurs paramétriques considérée est celle des mélanges gaussiens [3, 5, 4] en raison de son caractère universel et sa capacité à modéliser des distributions complexes.

Le problème de classification à contraintes évolutives est décrit au paragraphe 2. Ce paragraphe définit le risque utilisé et présente la règle optimale quand les densités de probabilité sont connues. Le paragraphe 3 expose l'approche envisagée. Les résultats d'essais expérimentaux sur un problème synthétique à deux dimensions et trois contraintes, dont une est évolutive, sont présentés au paragraphe 4 suivi par une conclusion.

2 Règle de décision à contraintes évolutives

Le problème de classification à contraintes évolutives est décrit par les options de décision, la fonction coût à minimiser et les contraintes de performance exigées. La solution dans le cadre des tests d'hypothèse statistiques est présentée.

2.1 Problème à contraintes évolutives

Dans un problème multiclasses à contraintes constantes, comportant N classes w_1,...,w_N, une observation x est issue d'une classe unique w_j ($j = 1, ..., N$). Une règle de décision avec rejet sélectif [6, 7] consiste à affecter x à une option de décision D_i ($i = 1, ..., I$ le nombre de décisions possibles). Chaque option de décision est définie par une classe ou un sous-ensemble de classes ou encore l'ensemble de toutes les classes. Par exemple, affecter x à $\{w_1, w_5\}$ signifie que x est classée dans w_1 et w_5 avec ambiguïté. Une règle de décision dans $\Re^d$, où d est la dimension de l'espace de représentation, est définie telle que $Z(x) = i$ quand x est affectée à la $i^{ème}$ décision. La probabilité de décider D_i sur l'ensemble des observations de w_j est notée $P(D_i/w_j)$, et définie par $\int_{\{x|Z(x)=i\}} P(x/w_j)dx$.

La fonction coût associée à une règle Z est définie par :

$$c(Z) = \sum_{i=1}^{I}\sum_{j=1}^{N} c_{i,j}P_jP(D_i/w_j)$$

où $c_{i,j}$ est le coût d'affecter une observation à la $i^{ème}$ option de décision quand elle appartient à w_j et P_j est la probabilité a priori de la classe w_j. Les contraintes de performance à respecter pour une règle Z sont exprimées formellement par K inégalités à seuils fixes, où la $k^{ème}$ contrainte est définie par :

$$e^{(k)}(Z) = \sum_{i=1}^{I}\sum_{j=1}^{N} \alpha_{i,j}^{(k)}P_jP(D_i/w_j) \leq \gamma^{(k)}$$

où

- $e^{(k)}(Z)$ est l'expression de la $k^{ème}$ contrainte,
- $\alpha_{i,j}^{(k)}$ est le coût d'affecter une observation issue de w_j à la $i^{ème}$ option dans la $k^{ème}$ contrainte,
- $\gamma^{(k)}$ est le seuil correspondant à la $k^{ème}$ contrainte.

La règle de décision recherchée doit minimiser le coût tout en respectant les contraintes. Elle est la solution d'un problème d'optimisation sous contrainte d'inégalités.

Dans le cadre des problèmes à contraintes évolutives, les seuils $\gamma^{(k)}$ évoluent en fonction du temps. Par conséquent, la règle de décision doit s'adapter aux changements des contraintes exigées. Le problème consiste à déterminer la règle optimale $Z^*(t)$ à chaque instant t qui minimise le coût $c(Z)$ tout en respectant les contraintes dépendant du temps selon $\boldsymbol{\gamma}^{(k)}(t)$:

$$\min_{Z} c(Z)$$
$$\text{sous } e^{(k)}(Z) \leq \gamma^{(k)}(t) \quad \forall k = 1, ..., K.$$

2.2 Solution dans le cadre théorique

Lorsque les densités conditionnelles $P(x/w_j)$ et les probabilités a priori P_j sont connues et indépendantes du temps, la solution peut être obtenue en déterminant le point selle $(Z^*(t), \boldsymbol{\mu}^*(t))$ du lagrangien $L(Z, \boldsymbol{\mu}, \boldsymbol{\gamma}(t))$ défini par :

$$L(Z, \boldsymbol{\mu}, \boldsymbol{\gamma}(t)) = \sum_{i=1}^{I}\int_{\{x|Z(x)=i\}} \lambda_i(x, \boldsymbol{\mu})dx - \boldsymbol{\mu}^T\boldsymbol{\gamma}(t)$$

avec

$$\lambda_i(x, \boldsymbol{\mu}) = \sum_{j=1}^{N} P_jP(x/w_j)\left(c_{i,j} + \boldsymbol{\mu}^T\boldsymbol{\alpha}_{i,j}\right)$$

et $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]^T$ le vecteur des multiplicateurs de Lagrange et $\boldsymbol{\gamma}(t) = [\gamma^{(1)}(t), \gamma^{(2)}(t), \ldots, \gamma^{(K)}(t)]^T$ et $\boldsymbol{\alpha}_{i,j} = [\alpha_{i,j}^{(1)}, \alpha_{i,j}^{(2)}, \ldots, \alpha_{i,j}^{(K)}]^T$.

Cela revient à résoudre, à chaque instant t, un problème d'optimisation donné par le dual Lagrangien :

$$\max_{\boldsymbol{\mu}\in\Re^{K+}} \{\min_{Z} L(Z, \boldsymbol{\mu}, \boldsymbol{\gamma}(t))\} \tag{1}$$

La règle optimale pour un vecteur $\boldsymbol{\mu}$ est :

$$\widetilde{Z}_{\boldsymbol{\mu}}(x) = i \quad \text{si} \tag{2}$$
$$\lambda_i(x, \boldsymbol{\mu}) < \lambda_l(x, \boldsymbol{\mu}), \forall i = 1, \ldots, I, l = 1, \ldots, I, l \neq i$$

La règle optimale à l'instant t est définie par $Z^*(t) = \widetilde{Z}_{\boldsymbol{\mu}^*(t)}$ où $\boldsymbol{\mu}^*(t) = \text{argmax} \quad (L(\widetilde{Z}_{\boldsymbol{\mu}}, \boldsymbol{\mu}, \boldsymbol{\gamma}(t)))$. Pour les problèmes à contraintes évolutives les multiplicateurs de lagrange optimaux doivent être déterminés pour chaque valeur des bornes sur les contraintes.

3 Apprentissage supervisé d'une règle de décision à contraintes évolutives

Dans le cas de l'apprentissage supervisé, les probabilités conditionnelles $P(x/w_j)$ et les probabilités a priori P_j ne sont pas connues et la règle de décision doit être déterminée à partir d'échantillons étiquetés. L'approche proposée consiste à construire la règle de décision en utilisant une estimée des densités de probabilité, correspondant à des mélanges gaussiens.

3.1 Modalité d'optimisation

Contrairement à l'approche présentée dans [2], qui optimise conjointement les paramètres de la famille d'estimateurs des densités et les multiplicateurs de lagrange au sens du

critère discriminant, l'approche proposée dans ce travail consiste à découper le problème en deux. Premièrement, les densités de probabilités conditionnelles à chacune des classes sont modélisés à l'aide d'un modèle paramétré, et les paramètres optimaux sont déterminés à l'aide d'un critère lié à l'adéquation aux données. Deuxièmement, la règle de décision à contraintes évolutives est déterminée, pour chaque instant t, sur la base des distributions estimées. Cette approche est sous-optimale du fait que les paramètres de l'estimateur ne sont pas optimisés en fonction du critère discriminant. Cependant, deux raisons plaident pour son emploi : sa flexibilité vis à vis des variations des contraintes exigées et son temps de calcul réduit.

3.2 Estimation de la densité de probabilité

La famille des mélanges gaussiens est utilisée compte tenu de son pouvoir représentatif pour un grand nombre de distributions. Ainsi, la probabilité conditionnelle $\hat{P}_M(x/w)$ d'une observation x de w, avec w une des N classes, est donnée par :

$$\hat{P}_M(x/w) = \sum_{m=1}^{M} \pi_m \hat{P}_m(x/w) \qquad (3)$$

$$\hat{P}_m(x/w) = (2\pi)^{\frac{-d}{2}} \mid \boldsymbol{S}_m \mid^{-0.5} \exp[\frac{-(x-\boldsymbol{m}_m)^T \boldsymbol{S}_m^{-1}(x-\boldsymbol{m}_m)}{2}]$$

où M est le nombre des composantes gaussiennes de la classe considérée, π_m, $\boldsymbol{m}_m$ et $\boldsymbol{S}_m$ sont respectivement le poids, la moyenne et la matrice de covariance de la $m^{ème}$ composante dans le mélange de la classe w. Ces paramètres sont estimés selon l'algorithme EM [3] pour un nombre connu de composantes relatif à chaque classe. Le nombre optimal M^* de composantes par classe doit être déterminé.

Le nombre des composantes est estimé à l'aide de la méthode "Greedy EM" [5, 4], itérativement et indépendamment dans chaque classe sur un critère lié à la performance de l'estimation des données. Le principe consiste à augmenter le nombre M itérativement pour atteindre M^* maximisant la vraisemblance estimée sur un ensemble de données de validation. Le choix de cet algorithme repose sur sa capacité de convergence globale et son aspect itératif permettant de déterminer le nombre inconnu des composantes du mélange gaussien.

3.3 Détermination de la règle de décision

Les densités estimées sont intégrées dans le cadre des tests d'hypothèses statistiques pour résoudre le problème d'optimisation (1). A chaque vecteur des contraintes $\boldsymbol{\gamma}(t)$, le problème d'optimisation à résoudre change. Les probabilités conditionnelles estimées $\widehat{P}_{M_j}(x/w_j)$ restent inchangées. La règle de décision optimale $Z^*(t)$, définie à partir des multiplicateurs de Lagrange optimaux $\boldsymbol{\mu}^*(t)$, dépend du temps. La recherche du $\boldsymbol{\mu}^*(t)$ qui optimise le risque empirique est effectuée à chaque instant par la méthode de la descente du gradient.

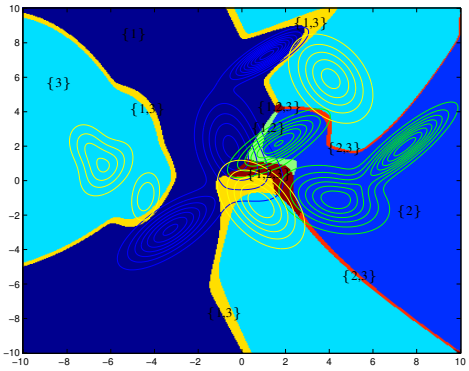

FIGURE 1 – Exemple de partition pour le seuil $\boldsymbol{\gamma}(t), t \geq t_1$ pour un ensemble de 200 échantillons par classe.

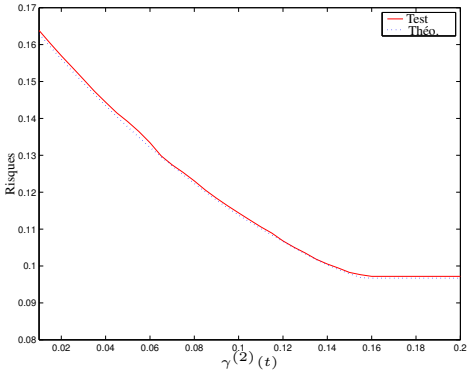

FIGURE 2 – $\widehat{L}(\widehat{Z}^*(t), \boldsymbol{\mu}^*(t), \boldsymbol{\gamma}(t))$ (continu) et $L(Z^*(t), \boldsymbol{\mu}^*(t), \boldsymbol{\gamma}(t)$ (interrompu) fonction de $\gamma^{(2)}(t)$.

4 Étude expérimentale

Un problème synthétique, défini dans $\Re^2$ a été considéré. Il est défini par 600 observations issues de 3 classes équiprobables. Ce sont des densités trimodales à supports non forcément connexes. Chaque classe est formée par 3 composantes gaussiennes. Les fonctions estimées des densités sont illustrées par les courbes de niveau de la figure 1. Les probabilités a priori sont supposées connues. Sept options de décision sont considérées. Elles correspondent aux sous-ensembles $\{w_1\}, \{w_2\}, \{w_3\}, \{w_1, w_2\}, \{w_1, w_3\}, \{w_2, w_3\}$ et $\{w_1, w_2, w_3\}$. Trois contraintes, dont le seuil pour l'une d'elles dépend du temps, sont définies :

- $P_E \leq 0.05$,
- $P_I \leq \gamma^{(2)}(t)$,
- $P(E/w_1) - 0.5P(E/w_2) \leq 0 \Leftrightarrow 2P(E/w_1) \leq P(E/w_2)$.

avec P_E la probabilité d'erreur et P_I la probabilité d'indistinction (probabilité de classer un élément dans un sous-ensemble contenant sa classe d'appartenance).

Dans un premier temps, un problème avec rupture de contraintes à l'instant t_1 est étudié :

	$\boldsymbol{\gamma}(t) = [0.05 \quad 0.03 \quad 0], \quad t < t_1$			$\boldsymbol{\gamma}(t) = [0.05 \quad 0.08 \quad 0], \quad t \geq t_1$		
	Apprentissage	Test	Théorique	Apprentissage	Test	Théorique
$\widehat{c}$	0.147	0.151	0.150	0.115	0.118	0.122
$\widehat{P_E}$	0.050	0.049	0.049	0.049	0.050	0.050
$\widehat{P_I}$	0.029	0.033	0.028	0.079	0.082	0.079
$\widehat{P}(E/w_1) - 0.5\widehat{P}(E/w_2)$	0.000	0.007	0.000	0.000	0.005	0.000
$\widehat{L}$	0.141	0.150	0.149	0.131	0.123	0.122

TABLE 1 – Valeurs théoriques et estimées du coût, des performances et du risque lagrangien pour deux ensembles de contraintes.

$$\boldsymbol{\gamma}(t) = \begin{cases} [0.05 \quad 0.03 \quad 0] & \text{pour } t < t_1 \\ [0.05 \quad 0.08 \quad 0] & \text{pour } t \geq t_1 \end{cases}$$

Les valeurs théoriques de c, P_E, P_I, $P(E/w_1) - 0.5P(E/w_2)$ calculées à partir des distributions théoriques et de la règle théorique, les valeurs estimées de ces grandeurs $\widehat{c}$, $\widehat{P_E}$, $\widehat{P_I}$, $\widehat{P}(E/w_1) - 0.5\widehat{P}(E/w_2)$ et du lagrangien $\widehat{L}$ sur l'ensemble d'apprentissage ainsi que sur un ensemble test sont reportées dans le tableau 1. Ces estimés sont calculées en utilisant les multiplicateurs de Lagrange théoriques pour éliminer les imprécisions que pourra introuduire l'estimation de ces multiplicateurs. Une partition obtenue avec 200 échantillons par classe et quand $\boldsymbol{\gamma}(t)$ avec $t \geq t_1$ est illustré par la figure 1.

Dans un second temps, l'évolution des contraintes est supposée continue. Le seuil $\gamma^{(2)}(t)$ varie entre 0.01 et 0.2. La figure 2 montre la variation du risque théorique $L(Z^*(t), \boldsymbol{\mu}^*(t), \boldsymbol{\gamma}(t))$ et estimé sur un ensemble de test $\widehat{L}(\widehat{Z}^*(t), \boldsymbol{\mu}^*(t), \boldsymbol{\gamma}(t))$. Les résultats montrent que l'approche proposée peut être considérée comme efficace. Les performances obtenues s'approchent des valeurs théoriques. Dans l'exemple choisi, l'estimation obtenue est adaptée aux distributions. Pour un usage avec des données réelles, le choix de l'estimateur conditionne en partie les performances du système. Un choix judicieux est donc une condition nécessaire à l'efficacité de cette approche.

5 Conclusion

Une méthode d'apprentissage supervisé est proposée pour résoudre les problèmes multiclasses avec contraintes de performance évolutives en fonction du temps. Elle consiste à déterminer une estimation des distributions optimale au sein d'une famille d'estimateurs puis à déterminer la règle qui minimise le risque empirique, correspondant à chaque contrainte, sur la base du modèle optimal retenu. C'est une stratégie simple et à temps de calcul réduit du fait que les distributions sont apprises indépendamment des contraintes. Sa performance, fortement liée au modèle choisi, exige un bon choix du modèle de densités. Les mélanges gaussiens sont choisis compte tenu de leur aspect universel et leur pouvoir de modéliser des données complexes. Les résultats obtenus par simulation montrent que cette approche est adaptée aux problèmes des contraintes évolutives. Les performances obtenues expérimentalement diffèrent légèrement de celles théoriques. Dans le cas des contraintes évoluant de façon continue au cours du temps, il serait intéressant d'adapter conjointement l'estimation des densités et la règle de décision. Des reflexions pour aller dans ce sens sont en cours.

Références

[1] E. Grall and P. Beauseroy, *Optimal Decision Rule with Class-Selective Rejection and Performance Constraints*, IEEE Transactions on Pattern Analysis and Machine Intelligence.To appear in 2009.

[2] E. Grall and P. Beauseroy, A. Bounsiar, *Quality assessment of a supervised multilabel classification rule with performance constraints*, in proccedings of EUSIPCO'06, Italy,2006.

[3] A.P. Dempster, N.M. Laird and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, J. Roy. Statist, 39 :1-38, 1977.

[4] N. Vlassis, J.J. Verbeek and B. Kröse, *Efficient greedy learning of gaussian mixture models*, Neural Computation, 15 :469–485, 2003.

[5] N. Vlassis and A. Likas, *A Greedy EM Algorithm for Gaussian Mixture Learning*, Neural Processing Letters, 15 :77–87, Springer, 2002.

[6] T. M. Ha. *The optimum class-selective rejection rule*. IEEE Trans. Pattern Anal. Mach. Intell., 19(6) :608–615, 1997.

[7] T. Horiuchi.*Class selective rejection rule to minimize the maximum distance between selected classes*. Pattern Recognition, 31(10) :1579–1588, October 1998.