Classification supervisée de tumeurs cancéreuses avec rejet sélectif

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 – Ce papier présente le problème de classification des tumeurs cancéreuses en se basant sur leur expression génétique. Le rejet sélectif est introduit afin de réduire les erreurs de classification et d'identifier les échantillons les plus ambigüs. La procédure proposée construit une règle de décision minimisant un coût défini dans le cadre du rejet sélectif à partir du ν-1-SVM en exploitant le chemin de régularisation. Un banc de classifieurs à rejet sélectif est également proposé pour rendre le diagnostic plus fiable. Deux études expérimentales, dans le cas Bayésien et dans le cas du rejet sélectif, portant sur cinq bases de tumeurs cancéreuses montrent que l'approche proposée est prometteuse et compétitive.

Abstract – A supervised rule for cancer diagnosis, based on selected gene profiles, is proposed. This rule introduces class selective rejection to get more accurate diagnosis. It is based on ν-1-SVM coupled with its regularization path and minimizes a general loss function defined in the class selective rejection scheme. A cascade of classifiers with class selective rejections is also suggested to ensure a higher reliability. Two experiments were carried out in the Bayesian and the class selective rejection frameworks. Five genes selected datasets are used to assess the performance of the proposed method. Results show that these approaches provide an improved cancer diagnosis.

1 Introduction

Récemment, un nombre croissant de travaux de recherche a accordé une importance capitale aux problèmes de classification des tumeurs cancéreuses basés sur leur expression génétique. Généralement, le problème est décrit par une petite population de tumeurs disponible avec une énorme quantité de gènes, ce qui rend la classification difficile. Pour remédier à ce problème, une solution consiste à utiliser les tests statistiques pour mettre en évidence les gènes dont l'expression semble la plus caractéristique [1]. Ces gènes sont ensuite utilisés pour l'apprentissage supervisé des différentes catégories de tumeurs. Cette classification a pour but de compléter les diagnostics cliniques sans les remplacer. Il est donc important de déterminer une règle de diagnostic la plus fiable possible pour éviter les confusions lors du diagnostic final.

L'approche proposée dans ce papier est basée sur l'apprentissage supervisé multiclasses avec rejet sélectif [2, 3]. Elle consiste à limiter les erreurs de diagnostic en donnant la possibilité au classifieur de classer un patient dans un sous-ensemble de classes formé de plusieurs classes. Le classifieur proposé cherche à minimiser un coût qui dépend des décisions correctes mais ambigües ou incorrectes, conditionnellement à chacune des classes. Le classifieur Bayésien correspond au classifieur proposé dans le cas particulier où les options de décision sont limitées aux classes admissibles. Le classifieur repose sur un ensemble de SVMs monoclasses ou ν-1-SVMs [4] et les chemins de régularisation [5]. Leurs paramètres sont optimisés de façon à minimiser une fonction coût autorisant le rejet sélectif. Le choix des SVMs monoclasses est justifié par la nature déséquilibrée des bases de données où certaines classes sont majoritairement représentées. Deux expérimentations sont présentées ; l'une dans le cas Bayésien et l'autre dans le cas du rejet sélectif. Elles portent sur cinq bases de gènes cancéreux : LEUKEMIA72, OVARIAN, NCI, LUNG CANCER et LYMPHOMA [6]. Dans le cas Bayésien, les résultats sont comparés à ceux donnés dans la littérature et obtenus par les algorithmes de Bayes, plus proche voisin, perceptron linéaire, perceptron multicouches et SVM.

Le paragraphe 2 présente la sélection des variables à l'aide des tests statistiques. Le paragraphe 3 décrit le problème du diagnostic multiclasses dans le cadre du rejet sélectif. Le paragraphe 4 expose l'approche utilisée. Les études expérimentales sont données dans le paragraphe 5 qui précède la conclusion.

2 Sélection de gènes

Les données des tumeurs cancéreuses sont décrites par un très grand nombre d'attributs (gènes) et un faible nombre de patients ou d'observations. Par conséquent, il est nécessaire de les traiter avant de les intégrer dans un algorithme de classification. Le but est d'obtenir un ensemble de gènes pertinents en éliminant les gènes considérés comme peu influents selon un certain critère.

Il s'agit d'un problème de sélection de variables pour lequel plusieurs techniques ont été proposées dans la littérature. Une méthode présentée dans [1] repose sur les tests statistiques. Elle consiste tout d'abord à calculer un niveau d'expression relatif à chaque gène en utilisant tous les échantillons du gène, puis à calculer le résultat d'une statistique. Ce résultat est utilisé pour tester l'hypothèse H_0 : les moyennes de chacune des classes sont égales, contre l'hypothèse H_1 : elles ne sont pas toutes

égales. Quand la valeur du test est élevée, l'hypothèse H_1 est acceptée, le pouvoir discriminant du gène est attesté et le gène est retenu.

Les six tests statistiques présentés dans [1] sont considérés : ANOVA F (F), Brown-Forsythe test (B), Welch test (W), Adjusted Welch test (W^*), Cochran test (C) et Kruskal-Wallis test (H). Ces tests permettent de sélectionner un certain nombre de gènes considérés comme pertinents. Les gènes sélectionnés sont ensuite utilisés pour faire la classification. Les performances obtenues dépendent donc à la fois du test et de la méthode de classification.

3 Classification avec rejet sélectif

Le problème repose sur l'hypothèse de N classes de tumeurs et que $x \in \Re^d$ représente un patient appartenant à une classe unique w_j ($j = 1, ..., N$). La dimension d de l'espace de représentation correspond au nombre de gènes sélectionnés selon chacune des six stratégies évoquées ci-dessus.

Dans le cadre classique, une règle de décision consiste à affecter x à une option de décision D_i parmi I décisions possibles correspondant à chacune des N classes. Dans le cadre du rejet sélectif [2] chaque option de décision est définie par un sous-ensemble de classes qui comporte une seule classe, une sélection de classes ou toutes les classes. Par exemple affecter x à $\{w_1, w_5\}$ signifie que x est classé dans w_1 et w_5 avec ambiguïté. Une règle Z est définie par $Z(x) = i$ quand x est affecté à la $i^{ème}$ décision. Elle définit une partition de $\Re^d$ en I régions correspondant à chacune des décisions. Trouver une règle de décision consiste à minimiser la fonction coût $c(Z)$:

$$\sum_{i=1}^{I} \sum_{j=1}^{N} c_{i,j} P_j P(D_i/w_j) \tag{1}$$

où $c_{i,j}$ est le coût d'affecter x à la $i^{ème}$ option de décision quand il appartient à w_j. P_j est la probabilité a priori de la classe w_j et $P(D_i/w_j)$ est la probabilité de décider D_i sur l'ensemble des échantillons de w_j définie par :

$$P(D_i/w_j) = \int_{\{x/Z(x)=i\}} P(x/w_j) dx.$$

4 Apprentissage supervisé d'une règle de classification

Les probabilités conditionnelles relatives aux classes et les probabilités a priori n'étant pas connues, il est nécessaire d'estimer ces grandeurs qui interviennent dans l'expression du coût (1). La solution proposée consiste à construire un classifieur basé sur les ν-1-SVMs et à optimiser ses paramètres de façon à minimiser une estimée empirique de la fonction coût $c(Z)$ donné par (1).

4.1 ν-1-SVM

La méthode du ν-1-SVM [4] consiste à déterminer la région $\mathcal{R}^\lambda$ de volume minimal qui contient tous les éléments à l'exception d'une fraction ν ($0 \leq \nu < 1$) d'éléments appelés outliers. Le paramètre $\lambda = \nu n$ contrôle le nombre des outliers parmi les n élements. La méthode du ν-1-SVM détermine une fonction $f^\lambda(.)$ et un réel b^λ tel que $f^\lambda(x) - b^\lambda \geq 0$ si $x \in \mathcal{R}^\lambda$ et $f^\lambda(x) - b^\lambda < 0$ sinon. Pour déterminer $\mathcal{R}^\lambda$, l'espace initial des fonctions $f^\lambda(.)$ est projeté dans un espace de fonctions de Hilbert (RKHS) par un noyau $K(.,.)$. Ce noyau induit l'espace de Hilbert à travers la fonction de projection ϕ. La propriété du noyau reproduisant implique que $< \phi(x_p), \phi(x_q) >= K(x_p, x_q)$ avec x_p et x_q deux échantillons dans $\Re^d$. Dans le cas du noyau gaussien RBF (Radial Basis Function) paramétré par σ, $K(x_p, x_p) = 1$, tous les x sont projetés sur une hypersphère de rayon 1 et centrée à l'origine. Le ν-1-SVM consiste à trouver l'hyperplan $\mathcal{W}^\lambda$ qui sépare les x projetés dans RKHS de l'origine en se situant le plus loin possible de ce dernier. Chercher $\mathcal{W}^\lambda$ revient à maximiser la marge $b^\lambda / \parallel \mathrm{w}^\lambda \parallel$ (w^λ est le vecteur normal de $\mathcal{W}^\lambda$). Cela revient à trouver la fonction f^λ telle que $f^\lambda(x) - b^\lambda =< \mathrm{w}^\lambda, \phi(x) > -b^\lambda$ sont positives pour la proportion d'éléments $(1 - \nu)$. La fonction $f^\lambda(.)$ est donnée par la solution du problème d'optimisation quadratique convexe défini par :

$$\min_{\mathrm{w}^\lambda, b^\lambda, \xi_p} \sum_{p=1}^{n} \xi_p - \lambda b^\lambda + \frac{\lambda}{2} \parallel \mathrm{w}^\lambda \parallel^2 \quad \text{sous contraintes} \tag{2}$$

$$< \mathrm{w}^\lambda, \phi(x_p) >\geq b^\lambda - \xi_p \quad \text{et} \quad \xi_p \geq 0 \quad \forall p = 1, ..., n$$

ξ_p sont les variables d'écart. Le problème dual de (2) est obtenu en introduisant les multiplicateurs de Lagrange α_p^λ :

$$\min_{\alpha_1, ..., \alpha_n} \frac{1}{2\lambda} \sum_{p=1}^{n} \sum_{q=1}^{n} \alpha_p^\lambda \alpha_q^\lambda K(x_p, x_q) \tag{3}$$

$$\text{avec} \quad \sum_{p=1}^{n} \alpha_p^\lambda = \lambda \quad \text{et} \quad 0 \leq \alpha_p^\lambda \leq 1 \quad \forall p = 1, ..., n.$$

Le paramètre ν ou λ contrôle les erreurs à la marge. Pour éviter de résoudre ce problème pour chaque valeur de λ souhaitée, il est possible d'utiliser une technique appelée méthode du chemin de régularisation du ν-1-SVM [5] qui permet d'apprendre, très rapidement, les ν-1-SVM pour toutes les valeurs de λ comprises entre 0 et n.

4.2 Classifieur multiclasses avec rejet sélectif

Le classifieur multiclasses considère N ν-1-SVMs et affecte un élément donné à une option de décision. Le principe de la méthode proposée consiste à affecter un patient x à l'option i ssi :

$$\sum_{j=1}^{N} c_{ij} P_j \beta_j d^{\lambda_j}(x) \leq \sum_{j=1}^{N} c_{lj} P_j \beta_j d^{\lambda_j}(x), \forall l = 1, ..., I, l \neq i$$

où :

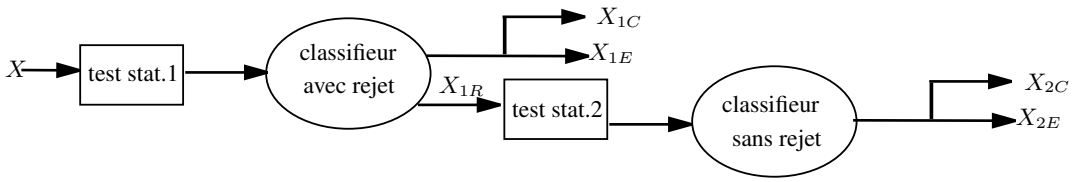

FIGURE 1 – Banc de classifieurs avec rejet sélectif

– $d^{\lambda_j}(x)$ est la "mesure" de distance entre x et une classe donnée w_j, proposée par [7] :

$$d^{\lambda_j}(x) = \frac{\frac{1}{\lambda_j}\sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K(x_p, x)}{b^{\lambda_j}}, \qquad (4)$$

– β_j est le poids associé à d^{λ_j} pour ajuster la mesure d^{λ_j} en faveur de la classe la moins pénalisante,
– n_j est le nombre d'observations de la classe w_j.

Le calcul de d^{λ_j} est simple et rapide car les $\alpha_p^{\lambda_j}$ sont obtenus par le chemin de régularisation pour toute valeur de λ_j. La règle de décision dépend donc des paramètres σ_j, λ_j et β_j, $j = 1, ..., N$. Ces paramètres sont optimisés par validation croisée afin de minimiser une estimée empirique $\hat{c}(Z)$ de la fonction coût $c(Z)$ définie par (1).

L'algorithme proposé permet d'introduire du rejet, c'est à dire de ne pas prendre une décision ou de prendre une décision avec une certaine ambiguïté, permettant ainsi de limiter les erreurs sur les échantillons les plus difficiles à classer. La probabilité d'erreur est ainsi réduite. Cet algorithme présente l'intérêt d'avoir une formulation générale, adaptée aussi bien aux problèmes sans rejet qu'à ceux avec rejet. Il n'est pas biaisé en faveur de la classe la plus représentée puisque chaque classe est apprise séparément.

4.3 Banc de classifieurs avec rejet

L'idée du banc de classifieurs est basé sur deux concepts. Premièrement, en introduisant du rejet sélectif, le nombre d'erreurs est réduit. Deuxièmement, les gènes pertinents issus de deux tests statistiques différents enrichissent les informations fournies pour une observation. Pour améliorer la performance du diagnostic à partir de l'ensemble initial X, un banc de classifieurs représenté par la figure 1 est proposé. Il est formé de deux ensembles constitués chacun d'un test statistique et d'un classifieur. Le premier autorise le rejet afin de classer les observations les plus ambigües X_{1R} dans un sous-ensemble de classes. Ces observations sont ensuite définies suivant un deuxième test statistique (test stat.2) et classées par le second classifieur sans rejet. Le nombre d'observations mal classées correspond au cumul des erreurs des deux classifieurs X_{1E} et X_{2E}. Les observations X_{1C} et X_{2C} correspondent aux individus correctement classés par les deux classifieurs.

5 Résultats expérimentaux

Deux études expérimentales ont été réalisées sur cinq bases de données décrites dans le tableau 1 [6]. La sélection des gènes effectuée par les auteurs de [1], a été reprise pour mener cette étude. Cinquante gènes caractéristiques sont choisis selon les six tests statistiques : ANOVA F (F), Brown-Forsythe test (B), Welch test (W), Adjusted Welch test (W^*), Cochran test (C) et Kruskal-Wallis test (H). Les erreurs de classification ont été calculées en utilisant la méthode du Leave-One-Out (LOO). Deux expérimentations, dans le cas Bayésien et dans le cas du rejet sélectif ont été considérées.

5.1 Classifieur Bayésien

La première étude expérimentale porte sur les erreurs obtenues en minimisant le coût $\widehat{c}(Z)$ correspondant à la probabilité d'erreur avec un nombre de décisions égal au nombre de classes. Le tableau 1 présente les performances obtenues pour chacune des bases utilisées ainsi que la valeur médiane et la valeur moyenne des erreurs de classification obtenues par les cinq algorithmes testés : Bayes, plus proche voisin, perceptron linéaire, perceptron multicouches et SVM rapportées dans [1].

Les erreurs obtenues avec la méthode proposée sont inférieures à la moyenne et à la valeur médiane données par les cinq algorithmes de [1]. De plus, pour la base de donnée LYMPHOMA qui présente un déséquilibre important (le rapport du nombre d'observations de la classe la plus peuplée sur la classe la moins représentée est approximativement de 23), la méthode proposée apporte une amélioration considérable au niveau de la précision. La méthode proposée est donc pertinente.

5.2 Performance des tests statistiques

Du point de vue performance des tests statistiques, conformément aux résultats de [1], les tests B, W, W^* et C sont plus performants que F et H pour une stratégie de classification donnée. Pour les données issues du LUNG CANCER pour lesquelles le ratio entre le nombre d'observations et la dimension de représentation est le plus faible, la qualité de la prédiction est presque la même quel que soit le test utilisé pour la selection des variables.

5.3 Banc de classifieurs avec rejet

La deuxième expérimentation traite une base de donnée, LUNG CANCER, dans le cas du rejet sélectif. Après avoir

			F	B	W	W^*	C	H
LEUKEMIA	# Gènes 6817	Algorithme Proposé	4	3	5	5	3	2
	# Patients 72	Moyenne	3.4	2.4	2.8	2.8	3.2	3.0
	# Classes 3	Médiane	3	2	3	3	3	3
OVARIAN	# Gènes 7129	Algorithme Proposé	0	0	0	0	0	0
	# Patients 39	Moyenne	0.2	0.0	0.0	0.0	0.0	0.0
	# Classes 3	Médiane	0	0	0	0	0	0
NCI	# Gènes 9703	Algorithme Proposé	31	26	27	27	27	33
	# Patients 60	Moyenne	36.0	32.0	27.4	26.0	27.0	35.4
	# Classes 9	Médiane	35	29	27	27	27	35
LUNG CANCER	# Gènes 918	Algorithme Proposé	14	16	16	16	16	15
	# Patients 73	Moyenne	17.6	17.0	17.6	17.6	18.0	18.0
	# Classes 7	Médiane	17	17	18	18	18	18
LYMPHOMA	# Gènes 4026	Algorithme Proposé	18	16	9	10	9	15
	# Patients 96	Moyenne	23.8	19.8	14.0	14.0	12.8	22.0
	# Classes 9	Médiane	23	19	12	12	13	20

TABLE 1 – Description des bases de données utilisées. Nombres d'observations mal classées par le classifieur proposé et valeurs moyennes et médianes des observations des observations mal classées par les 5 classifieurs rapportées [1].

étudié la nature du problème [8], nous avons considéré 10 options de décisions dont 7 correspondent aux classes admissibles et les 3 restantes correspondent à des rejets partiels et total. Les résultats montrent qu'avec la caractérisation W^*, l'erreur diminue de 16 observations mal classées à 10 mal classées avec 8 décisions ambigües. Les 8 éléments rejetés sont ensuite classés en considérant leurs attributs issus du test H et l'erreur sur l'ensemble total des individus est alors réduite à 13. Cela s'explique facilement puisque deux caractérisations distinctes contiennent des informations différentes. Par la suite, le banc de classifieurs proposé, avec rejet sélectif et différentes caractérisations, est une solution prometteuse pour construire un système de diagnostic des tumeurs cancéreuses plus fiable.

6 Conclusion

Une approche a été proposé afin d'apprendre une règle de décision basée sur le ν-1-SVM avec rejet sélectif pour les problèmes de classification des tumeurs cancéreuses. Elle possède plusieurs avantages : sa fiabilité, sa formulation générale et sa robustesse face aux problèmes déséquilibrés. Contrairement aux algorithmes qui ont tendance à favoriser la classe dominante, l'approche proposée traite toutes les classes d'une manière équitable du fait qu'elle apprend chaque classe individuellement. Deux études expérimentales sur cinq bases d'expressions génétiques cancéreuses permettent de valider la pertinence de l'approche proposée. Dans le cas Bayésien, les résultats obtenus, comparés à ceux des classifieurs de la littérature, sont compétitifs. Dans le cas du rejet, un banc de classifieurs avec rejet sélectif, basé sur des gènes issus de différents tests statiques, montre également que la procédure proposée permet d'aboutir à une règle plus fiable.

Références

[1] D. Chen and Z. Liu and X. Ma and D. Hua, *Selecting Genes by Test Statistics*, Journal of Biomedicine and Biotechnology. 132–138, 2005.

[2] 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.

[3] N. Jrad, E. Grall-Maës, and P. Beauseroy. *A supervised decision rule for multiclass problems minimizing a loss function*, Seventh International Conference on Machine Learning and Applications, 48–53, 2008.

[4] D. Tax, *One-class classification : concept learning in the absence of counter-examples*, Technische Universiteit Delft. 2001.

[5] A. Rakotomamojy and M. Davy, *One-class SVM regularization path and comparison with alpha seeding*, ESANN, 2007.

[6] Lawrence Berkeley National Laboratory, California, USA.

[7] M. Davy and F. Desobry and A. Gretton and C. Doncarli, *An online support vector machine for abnormal events detection*. Signal Process 86(8),2009–2025, 2006.

[8] M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G. Rosen, C. Perou, R. Whyte, R. Altman, P. Brown, D. Botstein, and I. Petersen. *Diversity of gene expression in adenocarcinoma of the lung*. In Proc Natl Acad Sci, volume 98, 13784–13789, USA, 2001.