Classification mixte pour l'Analyse de Données d'Expression

Hybrid Classification for Analyzing Expression Data

Pierre CHIAPPETTA[‡] Marie-Christine ROUBAUD[‡] Bruno TORRÉSANI[‡]

[‡] LATP–CMI, 31 rue Joliot-Curie, 13013 Marseille Cedex 13

Courriel : chiapeta@cpt.univ-mrs.fr, {mcroubau,Bruno.Torresani}@cmi.univ-mrs.fr

Résumé

Nous décrivons une approche pour l'analyse de grands ensembles de données issues du transcriptome, basée sur la classification mixte qui combine l'agrégation autour de centres mobiles et la classification hiérarchique. Cette méthode appliquée à un jeu de données sur le cancer du sein fournit un regroupement robuste d'un grand nombre de gènes ayant la même fonction biologique dans une même classe. Parmi les classes obtenues celles du recepteur d'estrogène et du récepteur de la prolactine fournissent à leur tour une séparation des tumeurs en deux groupes.

Mots clés : *données d'expression, classification.*

Abstract

We present an approach for analyzing transcriptome data, based upon hybrid clustering. This approach combines the partition method and the hierarchical method. This method used on a data set of breast carcinomas gives a robust clustering of many genes participating the same biological process in the same cluster. Genes co-expressed with the estrogen receptor and with the prolactin receptor allow a separation of breast tumors in two groups.

Keywords: *expression data, clustering.*

1 Introduction

Pour l'analyse du transcriptome, différentes méthodes de classification ont été appliquées pour classer les gènes ou les conditions basées soit sur une agrégation autour de centres mobiles, soit sur une classification hiérarchique ascendante [1]. Les méthodes du type agrégation autour de centres mobiles affectent chaque donnée à la classe la plus "proche". Elles offrent l'avantage d'obtenir à moindre coût une partition d'un ensemble volumineux de données mais nécessitent de fixer a priori le nombre de classes et fournissent des partitions dépendant de l'initialisation. Les algorithmes de classification hiérarchique sont mal adaptés aux grands jeux de données et ne remettent jamais en cause les classes obtenues aux différentes étapes de la procédure. Nous décrivons ici une approche basée sur une classification mixte qui combine les deux techniques précédentes afin de tirer parti de leurs avantages et d'amoindrir leurs inconvénients [2]. A notre connaissance les classifications précédemment obtenues [4] utilisaient l'une des deux techniques mais ne combinaient pas les deux.

2 Méthode de classification mixte.

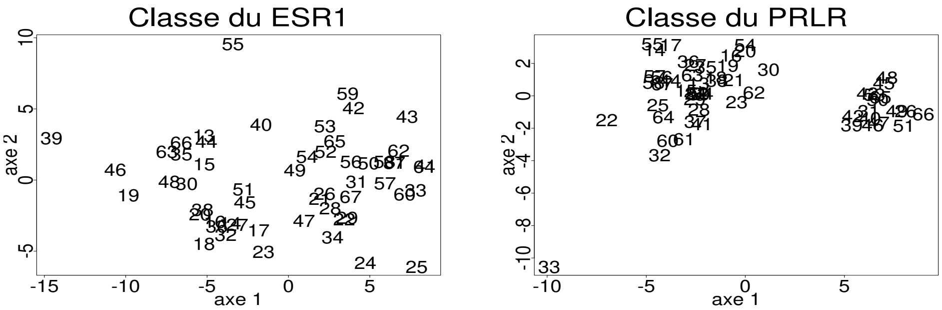
Première étape : recherche des groupements stables. La procédure d'initialisation est effectuée au moyen de l'algorithme de Kohonen (SOM) qui projette les données dans un espace à deux dimensions,de sorte que deux observations voisines dans l'espace d'entrée appartiennent à la même classe ou à des classes voisines à la sortie [5]. Les centres des classes issus de cet algorithme sont ensuite utilisés comme initialisation de l'algorithme des centres-mobiles (K-means). Cependant ce type d'algorithme converge vers un optimum local. Pour remédier partiellement à cet inconvénient, une méthode de *groupements stables* suggérée principalement par E. Diday est utilisée. Cette technique consiste à appliquer plusieurs fois la même procédure et à ne retenir que les groupements stables c'est à dire les ensembles d'objets qui ont toujours été affectés à une même classe.

Deuxième étape : recherche de hiérarchies des groupements. La deuxième étape consiste à effectuer une classification ascendante hiérarchique selon le critère de Ward sur les centres des groupements stables.

Troisième étape : partition finale. L'algorithme des centres-mobiles est de nouveau utilisé pour réaffecter l'ensemble des objets à classer. On obtient ainsi une partition finale constituée de *super-classes*.

3 Prétraitement des données

Cette étape est essentielle car la plupart des méthodes sont sensibles aux valeurs extrêmes. Pour l'étude de la covariation des gènes, il s'avère aussi nécessaire de centrer et réduire les données.



Transformation logarithmique : Ce premier prétraitement pose toutefois le problème des valeurs nulles. Pour y remédier la procédure suivante est employée : pour chaque condition c on sélectionne la plus faible valeur x^c_{min} non nulle. On remplace ensuite toutes les valeurs nulles des niveaux d'expression par des valeurs pseudo-aléatoires uniformément distribuées entre 0 et αx^c_{min} où α est une constante.

Méthode non paramétrique : Le second prétraitement consiste à transformer les données d'expression en données ordinales (rangs). Cette méthode a l'avantage d'être insensible aux valeurs extrêmes mais entraîne la perte de l'information quantitative.

4 Application aux données d'expression.

Cette procédure a été appliquée à un jeu de données sur le cancer du sein obtenu par l'équipe TAGC du CIML de Marseille [3]. On dispose des profils d'expression de 1045 gènes sur 12 lignées et 55 tumeurs. Les objectifs de cette étude s'inscrivent principalement dans deux problématiques. La première est la recherche de covariations de l'expression des gènes afin d'essayer d'en extraire des informations sur leurs interactions. La seconde est la comparaison des conditions dans le but d'identifier des gènes permettant de séparer des groupes de tumeurs.

Classification des gènes. On aboutit à des regroupements distincts de gènes suivant leur fonction biologique en une dizaine de *super-classes*. Les deux prétraitements n'induisent pas de différences importantes en particulier pour les *super-classes* de l'estrogène (ESR1), de la prolactine (PRLR) et du système immunitaire.
Le gène ESR1 est regroupé avec les gènes ANG, C4A, EGF, GLI3, MUC1, KRT19, BCL2, GATA3, ILF1, MYB, SMARCA2, THBS1, VIL2, XBP1, IGFBP2, AMFR, PLAT,FOS. Le gène PRLR est regroupé avec ABCB1, API1, BAK1, BCL2, CDH15, CDKN3, CIDEA, CTSB, DAP3, GJA1, FAT, TC21. Toutefois les gènes de réponse immédiate sont regroupés avec l'ESR1. Le gène BRCA1 est regroupé avec APC, CASP6, CREBBP, ITGAL, ITGB4, RARA, RBL1, TUBB1, CDKN2A, CDK2, NOS3, GZMA, IL5RA.

Classification des tumeurs. Partant des *super-classes* des gènes co-exprimés avec ESR1 et PRLR, une ACP fournit dans chaque cas une séparation des tumeurs en deux groupes (voir figures).
La confrontation aux données cliniques montre en ce qui concerne l'ESR1, que la séparation est liée au taux de récepteur d'estrogène présent dans les tumeurs. Les tissus " ER+" se trouvent essentiellement à droite de $x = 0$ et les "ER-" à gauche.

Références

[1] L. Lebart, A. Morineau, M. Piron (2000) : Statistique exploratoire multidimensionnelle. *Dunod* **3ème édition.**

[2] M. A. Wong (1982) : A hybrid clustering method for identifying high density clusters. *Amer. Statist. Assoc.* **77** pp. 841-847.

[3] R. Houlgatte et al (2001) : Unpublished results. Voir aussi F. Bertucci, R. Houlgatte et al (2000) : Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Molecular Genetics* **9** :20, pp. 2981-2991.

[4] M. Eisen et al, P. T. Spellman, P. O. Brown and D. Botstein (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**, pp. 14863-14868.

[5] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen (1995) : *SOM-PAK : The Self-Organizing Map Program Package Version 3.1*, Helsinki University of Technology, Finland.