

Nouvelle méthode de classification adaptée aux données de grandes dimensions : application aux données de biopuces

Doulaye Dembélé

IGBMC, CNRS-IMSERM-ULP
1 rue Laurent Fries, BP 10142
Parc d'Innovation
67404 Illkirch Cedex, France
doulaye@titus.u-strasbg.fr

Résumé

Nous proposons dans cet article une nouvelle méthode de classification adaptée aux données de grandes dimensions. Pour ces données la distance de Chebyshev semble intéressante, de plus elle nécessite moins de temps de calcul comparée à la distance Euclidienne, plus utilisée en raison de ses bonnes propriétés géométriques. La méthode proposée combine les méthodes de regroupement hiérarchique et par partition pour obtenir le nombre de classes dans les données. Des données issues d'expériences de biopuces sont utilisées pour illustrer les performances de la méthode proposée.

1 Introduction

Nous proposons dans ce papier une nouvelle méthode de classification adaptée aux données de grandes dimensions. Pour ces données le problème d'espace vide est connu [2]. D'autres faiblesses sont liées à l'utilisation de la distance Euclidienne, parmi celles-ci signalons le fait que les distances entre toutes les paires de points des données ont tendance à être identiques quand la dimension augmente [1]. Dans ces conditions il sera impossible de discriminer les classes, s'il y en a, dans les données. Il est aussi montré dans [4] qu'en augmentant l'ordre de la métrique de Minkowski, il est possible d'augmenter la dimension de la matrice des distances des données prises deux à deux. Le maximum de dimension pour la matrice des distances sera obtenu pour un ordre infini, c'est-à-dire en utilisant la distance de Chebyshev. Notons que l'ordre de la matrice des distances définit le degré de contraste permettant d'obtenir des classes dans les données. La distance de Chebyshev semble alors intéressante, de plus elle nécessite moins de temps de calcul comparée à la distance Euclidienne la plus utilisée en raison de ses bonnes propriétés géométriques. Ceci n'est pas négligeable pour les données de grande dimension comme celles générées par les biologistes dans le cadre de l'étude de l'expression des gènes à l'aide de la technologie de biopuces.

Nous nous intéressons ici aux méthodes de classification non paramétriques et non supervisées ou clustering. Ces méthodes peuvent être regroupées dans deux grandes familles : les méthodes hiérarchiques et les méthodes par partition. Les méthodes hiérarchiques ne nécessitent pas la connaissance *a priori* du nombre de classes dans les données. Leur résultat est représenté sous forme d'un arbre ou dendrogramme dans lequel les branches contiennent les échantillons similaires du point de vue du critère utilisé pour les construire [3, 7, 6]. Cette méthode est intéressante mais elle ne permet pas de re-examiner un échantillon déjà placé dans une branche. Les méthodes par partition consistent à trouver le meilleur regroupement des N échantillons des données en K classes de manière à optimiser un critère de qualité défini *a priori*. Pour résoudre ce problème combinatoire, on utilise dans la pratique une heuristique pour avoir une solution en un temps raisonnable. Dans cette heuristique,

les échantillons sont initialement réparties en K classes puis itérativement, on recherche la meilleure combinaison locale qui améliore la qualité du critère prédéfini en changeant la classe de certains échantillons. Cette étape prend fin quand il n'y a plus d'amélioration possible du critère [3, 7, 6]. Le principal inconvénient des méthodes par partition est la nécessité de connaître *a priori* le nombre de classes à former. Dans cet article, nous proposons une nouvelle stratégie qui combine les méthodes hiérarchiques et par partition. Dans un premier temps la matrice des distances des données est calculée et utilisée pour former un nombre maximum de classes (étape par partition sans connaissance *a priori* du nombre des classes), puis un regroupement est effectué pour réduire le nombre des classes (étape hiérarchique ascendante avec critère de validation). L'idée de combiner les méthodes de classification hiérarchique et par partition n'est pas nouvelle, voir par exemple [10]. Toutefois la procédure présentée dans le paragraphe suivant est originale.

Nous présentons la méthode de classification proposée puis des résultats obtenus avec des données de biopuces sont ensuite présentés.

2 Nouvelle méthode de classification

La méthode de clustering non paramétrique par partition la plus utilisée est la méthode K-Means. Soient K le nombre des classes à trouver, \mathbf{c}_k le centre de la classe k ($k = 1, 2, \dots, K$) et \mathbf{x}_i l'échantillon i des données. La méthode K-Means permet d'obtenir la répartition des données après la minimisation de la fonction suivante :

$$J(\mathbf{c}_k) = \sum_{k=1}^K \sum_{i=1}^N u_{ik} d(\mathbf{x}_i, \mathbf{c}_k) \quad (1)$$

où $d(\mathbf{x}_i, \mathbf{c}_k)$ désigne la distance entre l'échantillon \mathbf{x}_i et le centre \mathbf{c}_k de la classe k alors que u_{ik} vaut 1 si l'échantillon \mathbf{x}_i appartient à la classe k et 0 sinon.

À partir d'une partition initiale, la fonction (1) est itérativement améliorée en changeant la classe des échantillons, jusqu'à l'obtention d'une solution stable. Dans la relation (1), il y a deux paramètres à choisir avant le début des calculs, la distance $d(.,.)$ et le nombre K des classes.

La distance Euclidienne est la plus utilisée à cause de ses bonnes propriétés géométriques. Elle définit toutefois implicitement une forme sphérique pour les classes à trouver. Pour obtenir des classes de forme ellipsoïdale, la distance de Mahalanobis est utilisée. Pour les données de grande dimension la distance de Chebyshev qui est équivalente à la métrique de Minkowski à l'ordre infini offre plus de contraste dans la matrice des distances [4]. C'est pour cette raison que notre choix s'est porté sur la distance de Chebyshev.

Étant donné qu'il est souvent difficile de connaître *a priori* le nombre des classes, nous nous proposons de les déterminer directement à partir des données. L'idée dans la fonction (1) est de former des classes telles que les échantillons membres d'une classe soient plus proches que ceux d'une autre classe. Cette proximité peut être définie par un seuil sur les distances [8]. Soit d_{seuil} cette distance seuil, toutes les distances des échantillons d'une classe doivent être alors plus petites que d_{seuil} . Le problème revient alors à déterminer ce seuil à partir des distances des données. La recherche du seuil est examinée plus loin. À partir d'un seuil approprié sur les distances, nous répartissons les données pour obtenir une valeur maximale pour K , puis un regroupement de certaines classes est enfin effectué. La procédure de classification proposée se résume comme suit :

1. Calculer toutes les distances entre les échantillons des données,
2. Rechercher un seuil pour les distances des échantillons d'une classe,
3. Repartir les données en utilisant le seuil trouvé,
4. Déterminer le nombre maximum K_{max} de classes sans prendre en compte les classes singletons,
5. Calculer les K_{max} centres des classes et les utiliser pour avoir une partition initiale des données,
6. Utiliser une méthode de regroupement et un critère de validation pour obtenir K classes.

La première étape consiste à calculer les $\frac{N(N-1)}{2}$ distances des N échantillons. La médiane de ces distances est utilisée pour obtenir le seuil recherché (étape 2). Ce seuil est utilisé dans la troisième étape conjointement avec les distances pour affecter un index à chaque échantillon. Cela est fait en comparant le premier échantillon aux autres, puis le second échantillon non indexé est comparé aux autres, et ceci est poursuivi jusqu'à l'avant dernier échantillon. Le même index est associé aux échantillons de distances inférieures ou égales au seuil. L'examen des différents index permet enfin d'obtenir K_{max} (étape 4). La cinquième étape consiste à calculer les centres des classes retenues puis à répartir les échantillons entre celles-ci. Dans la dernière étape, une méthode hiérarchique ascendante peut être utilisée. Il est également possible d'utiliser l'algorithme K-Means dans lequel le nombre de classes est décroissant.

2.1 Conditionnement des données

Avant d'utiliser un algorithme de classification, les données sont souvent standardisées. Cela consiste en général à transformer les données pour avoir une moyenne nulle et un écart type égal à un pour chaque échantillon. La standardisation est particulièrement utile pour les données de biopuces qui peuvent avoir des amplitudes très différentes alors que l'on est intéressé par la variation des profils. La transformation ci-dessus rend toutefois les données sphériques. Une autre transformation peut consister à ramener toutes les valeurs des données entre 0 et 1. Cela est obtenu en otant la valeur minimale de chaque valeur et en divisant le résultat par l'étendue, c'est-à-dire, la différence entre les valeurs maximale et minimale.

2.2 Détermination du seuil des distances

Le nombre maximum K_{max} dépend de la valeur de d_{seuil} . Si cette valeur est élevée, K_{max} sera faible et inapproprié, inversement si la valeur de d_{seuil} est faible, K_{max} sera élevé et générera beaucoup de classes singletons. Après des tests sur des données synthétiques, la médiane des distances fournit un bon compromis si les données sont transformées pour avoir des valeurs comprises entre 0 et 1. Nous utilisons cette solution heuristique pour d_{seuil} en attendant les résultats d'autres études sur le sujet.

2.3 Critère d'arrêt

Dans l'étape de réduction du nombre des classes, les critères comparés dans [9] peuvent être utilisés. En utilisant la transformation $[0, 1]$ des données, des échantillons de profils identiques peuvent être affectés à des classes différentes en fonction de leur valeur absolue moyenne. Pour regrouper les classes, nous utilisons l'information de leur forme. Ceci permet d'identifier les classes de profils voisins. Nous définissons et utilisons la co-variation des profils comme le coefficient de corrélation des écarts non centrés observés pour les valeurs successives de chaque profil (pour plus de détails, voir la version longue de ce papier).

3 Résultats

Pour illustrer les performances de la méthode proposée, nous avons utilisé les données de biopuces. Les biopuces sont de petits supports (lames de verre) sur lesquels des milliers de séquences d'ADN correspondant chacune à un gène sont attachées à des adresses connues (spots). L'ARN des échantillons à analyser est marqué avec une molécule fluorescente puis hybridé (par appariement entre séquences d'ADN complémentaires) sur les biopuces. Les biopuces sont ensuite scannées. Le niveau d'expression des gènes est représenté par une intensité de fluorescence. La quantification de l'image (mesure de l'intensité de fluorescence pour chacun des spots) fournit des données numériques qui servent à l'analyse. Nous nous sommes servi des données d'une étude de réponses de fibroblastes humains à des concentrations de sérum au cours du temps [5]. Nous avons utilisé les données correspondant à une sélection de 517 gènes. Ces données peuvent être récupérées à l'adresse suivante : <http://www.sciencemag.org/feature/data/984559.shl>.

La valeur de distance seuil obtenue pour ces données est $d_{seuil} = 0.223$. En utilisant cette valeur, nous avons obtenu un nombre maximum de classes $K_{max} = 23$. Les profils de ces classes sont représentés sur la figure 1. La matrice des co-variations des 23 profils initiaux des données de sérum

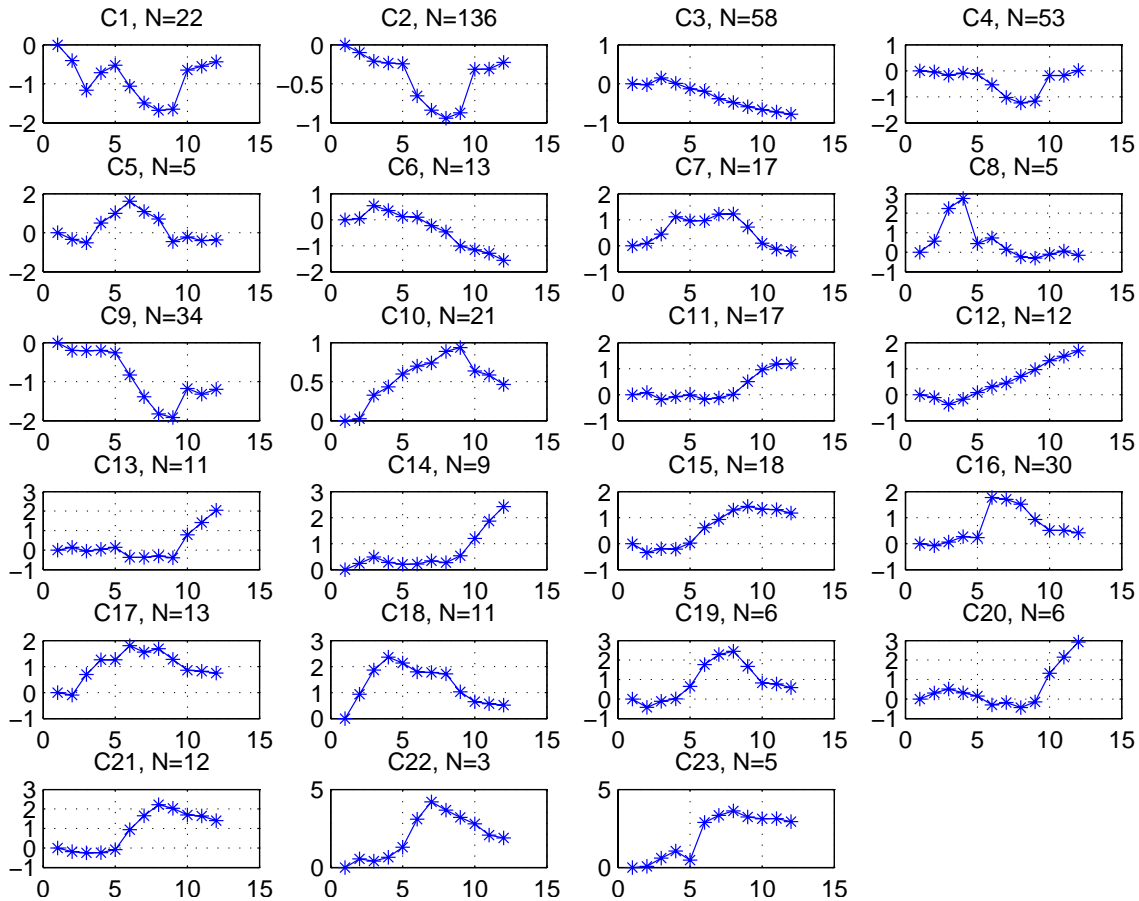


FIG. 1 – Données de serum, les 23 classes obtenues dans la phase initiale. Chaque profil est identifié par un numero et le nombre des échantillons qui la forme, e.g. la première classe $C1$ contient 22 échantillons ($N = 22$).

a ensuite été calculée. Cette matrice (non présentée ici) montre des coefficients de co-variation élevés, ≥ 0.75 , pour les classes 13, 14 et 20 qui ont des profils similaires. Le coefficient de co-variation est la plus petite, -0.87 , pour les classes 6 et 12 montrant une opposition de profils pour ces classes. La figure 1 contient tous les profils obtenus dans [5] avec des redondances. Avec la méthode hiérarchique

ascendante les classes ont été regroupées.

4 Conclusion

Une nouvelle méthode de classification de données est présentée. La distance de Chebyshev est utilisée. Cette distance semble plus appropriée pour les données de grandes dimensions. La stratégie proposée consiste dans un premier temps à rechercher un nombre maximum de classes dans les données. Ceci est fait après examen de la matrice des distances des données. Puis une réduction du nombre de classes est effectuée. Pour cela une méthode hiérarchique ascendante peut être utilisée. La méthode K-Means qui offre la possibilité de re-affecter un échantillon à une autre classe peut être aussi utilisée. Pour la standardisation des données, la méthode qui permet de ramener toutes les valeurs entre 0 et 1 a été utilisée. Un travail future consiste à étudier le choix du seuil d_{seuil} des distances d'une classe. Une interface conviviale pourra également faciliter l'exploitation des résultats de classification.

Remerciements

Merci à Bernard Jost qui a pris le temps de lire cet article. Je remercie également deux rapporteurs anonymes pour leurs remarques sur ce papier. Ce travail a bénéficié du soutien du Centre National de la Recherche Scientifique (CNRS), de l'Institut National de la Recherche Médicale (INSERM), de l'Hôpital Universitaire de Strasbourg et du Centre National de Recherche en Génomique (CNRG).

Références

- [1] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, TIRF, INPG, Grenoble, France, novembre 1994.
- [2] D. L. Donoho. High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. In *Am. Math. Soc. Conf. "Math Challenges of the 21st Century"*, Los Angeles, www-stat.stanford.edu/~donoho, 2000.
- [3] B. S. Everitt. *Cluster Analysis*. Arnold, London, 3rd edition, 1993.
- [4] J. Hérault, A. Guérin-Dugué, and P. Villemain. Searching for the Embedded Manifolds in High-Dimensional Data, Problems and Unsolved Questions. In *SANN'2002 Proceedings - European Symposium on Artificial Neural Networks 24-26 April, Bruges, Belgium*, pages 173–184, 2002.
- [5] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. H. Jr, M. S. Bogoski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The Transcriptional Program in the Response of Human Fibroblast to Serum. *Science*, 283 :83–87, january 1 1999.
- [6] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition : A Review. *IEEE trans. PAMI*, 22(1), January 2000.
- [7] J. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliff, New Jersey, 1988.
- [8] A. V. Lukashin and R. Fuchs. Analysis of Temporal Gene Expression Profiles : Clustering by Simulated Annealing and Determining the Optimal Number of Clusters. *Bioinformatics*, 17(5) :405–414, 2001.
- [9] G. W. Milligan and M. C. Cooper. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2) :159–179, June 1985.
- [10] M. A. Wong. A hybrid clustering method for identifying high-density clusters. *Journal of American Statistical Association*, 77(380) :841–847, December 1982.