Nouvelle méthode de classification adaptée aux données de grande dimension : 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.Dembele@igbmc.u-strasbg.fr,
http://www-microarrays.u-strasbg.fr

Résumé. Nous proposons une nouvelle méthode de classification adaptée aux données de grande dimension. Pour ces données la distance de Chebyshev semble intéressante, car 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

La classification permet de représenter des données sous une forme plus aisée à interpréter, à visualiser ou à manipuler. Nous proposons une nouvelle méthode de classification adaptée aux données de grande dimension. Pour ces données le problème d'espace vide est connu (Dohono, 2000). D'autres faiblesses sont liées à l'utilisation de la distance Euclidienne : sous certaines hypothèses sur la distribution des échantillons, les distances entre toutes les paires de points des données sont identiques quand la dimension augmente (Beyer et al., 1998). Dans ces conditions il est impossible de discriminer les classes, s'il y en a, dans les données. Il est aussi montré dans (Herault et al., 2002) qu'en augmentant l'ordre de la métrique de Minkowski, il est possible d'augmenter le rang 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 le rang de la matrice des distances définit le degré de contraste permettant d'obtenir des classes dans les données, c'est-à-dire le degré de redondance dans les données. La distance de Chebyshev semble alors intéressante. De plus, elle nécessite moins de temps de calcul comparé à la distance Euclidienne généralement utilisée en raison de ses bonnes propriétés géométriques. Le gain en temps de calcul 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 des biopuces.

Nous nous intéressons ici aux méthodes de classification heuristiques automatiques et non supervisées également appelées clustering. Ces méthodes se divisent en 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 (Jain et Dubes, 1988; Everitt, 1993; Jain et al., 2000). Cette méthode est intéressante mais elle ne permet pas de réexaminer 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 habituellement une heuristique pour obtenir un résultat en un temps raisonnable. Dans cette heuristique, les échantillons sont initialement repartis 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 (Jain et Dubes, 1988; Everitt, 1993; Jain et al., 2000). 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 les méthodes 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 d'arrêt). L'idée de combiner les méthodes de classification hiérarchique et par partition n'est pas nouvelle (Wong, 1982). Toutefois la procédure présentée dans cet article est originale.

Dans la suite la nouvelle méthode de classification est décrite puis des résultats obtenus avec des données de biopuces sont ensuite présentés.

2 Données de grande dimension

On assiste de nos jours à la génération de grandes masses de données. Ceci est particulièrement vrai dans le domaine de la biologie moléculaire où les techniques de biopuces permettent par exemple d'étudier simultanément la transcription de plusieurs milliers (e.g. 50 000) de gènes. Si une telle expérience est répétée pour des conditions (e.g. 100) différentes, on se retrouve avec un tableau contenant des millions de valeurs ou composantes. En plus de l'aspect stockage, il y a un besoin d'outils efficaces pour la gestion de ces données et l'accession à l'information qu'elles contiennent. C'est pour cette raison que les outils de classification et de fouille de données sont utiles. Ces méthodes permettent de rechercher des échantillons similaires à une référence ou de regrouper un ensemble d'échantillons en des sous-ensembles cohérents du point de vue d'un critère prédéfini. Le mot échantillon est utilisé ici pour désigner un gène dans une expérience de biopuces.

La plupart des méthodes de classification sont basées sur la distance Euclidienne. Quand la dimension n des données augmente, le phénomène d'espace vide apparaît (Dohono, 2000). Ceci se caractérise par des propriétés qui sont résumés dans (Jimenez et Landgrebe, 1995; Herault et al., 2002). Une solution à ce problème consiste à rechercher les classes dans un espace de dimension faible comparé à l'espace initial. La réduction de la dimension des données et la recherche des classes peuvent être faites simultanément ou dans des étapes séparées. Les données de dimension réduite sont obtenues après projection dans un espace de dimension plus faible que l'espace original. Un outil très utilisé pour réduire la dimension des données est l'Analyse en Composantes Principales (ACP) (Golub et Loan, 1996; Holter et al., 2000; Bonnet et al., 2002; Horn et Axel, 2003). Dans certaines méthodes, une projection non linéaire

(curviligne) est utilisée simultanément avec la recherche des classes (Demartines, 1994; Herault et al., 2002). Bien que la réduction de la dimension soit particulièrement utile pour la visualisation, il est montré dans (Yeung et Ruzzo, 2001) que les premières composantes d'une analyse ACP ne capturent pas nécessairement l'information sur les groupes présents dans les données. Il est alors souvent utile de rechercher les groupes dans l'espace initial.

2.1 Normes et distances

Il est montré dans (Demartines, 1994) que pour des échantillons de composantes indépendantes et identiquement distribuées, leur déviation standard tend à devenir constante quand la dimension n augmente. Il est aussi montré dans (Beyer et al., 1998) pour ces mêmes données que si l'écart type des distances entre toutes les paires de points tend vers zéro, ces distances ont tendance à être identiques quand la dimension n augmente. Ces deux résultats combinés expliquent en partie la détérioration des performances des algorithmes de classification pour les données de grande dimension. Notons toutefois que dans la pratique, l'hypothèse d'indépendance des échantillons n'est pas toujours satisfaite.

Soient $\mathbf{x}_i$ et $\mathbf{x}_j$ deux échantillons des données ayant chacun n valeurs. La distance Euclidienne est obtenue en prenant $p = 2$ dans la distance de Minkowski définie par :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^{n}(x_{ik} - x_{jk})^p\right]^{\frac{1}{p}} \tag{1}$$

La norme L_p ou de Minkowski de l'échantillon $\mathbf{x}_i$ est définie par :

$$L_p(\mathbf{x}_i) = \left[\sum_{k=1}^{n}(x_{ik})^p\right]^{\frac{1}{p}} \tag{2}$$

Soit M^p la matrice formée avec les distances de tous les échantillons et définie par :

$$M^p = [d^p(\mathbf{x}_i, \mathbf{x}_j)], \;\; i, j = 1, \ldots, N \tag{3}$$

où N est le nombre total des échantillons dans les données.

La matrice M^p est carrée d'ordre N et la borne supérieure de son rang est donnée par $r = n(p-1)+2$ (Herault et al., 2002). Plus r sera élevé, plus grand sera le nombre de colonnes indépendantes dans la matrice M^p, i.e. moins il y a de redondance dans les N échantillons $\mathbf{x}_i$. La relation donnant r montre que le plus grand rang sera obtenu pour p infini, c'est-à-dire en utilisant la distance de Chebyshev définie par :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max_k |x_{ik} - x_{jk}| \tag{4}$$

En comparaison avec la distance Euclidienne, la distance de Chebyshev nécessite moins de charge de calcul. Ceci peut s'avérer utile quand la dimension des données est élevée. La faible charge de calcul combinée avec la possibilité d'avoir un maximum de contraste dans les données nous ont conduit à utiliser cette distance dans la nouvelle méthode que nous proposons dans le paragraphe suivant. Pour les données de biopuces exprimées en rapport d'expression ou ratios, au lieu d'utiliser la norme L_2 comme c'est le cas habituellement, nous avons opté pour une autre normalisation qui consiste à transformer les valeurs de façon à ce qu'elles soient toutes comprises entre 0 et 1.

3 Nouvelle méthode de classification

La méthode heuristique de clustering 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 ($k = 1, \ldots, K$) et $\mathbf{x}_i$ l'échantillon i ($i = 1, \ldots, N$) des données. La méthode K-Means permet d'obtenir la répartition des données après la minimisation de la fonction suivante :

$$J(\mathbf{c}_k) = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik} d(\mathbf{x}_i, \mathbf{c}_k) \tag{5}$$

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.

La minimisation de la fonction (5) peut être obtenue avec l'algorithme suivant (Jain et al., 2000) :

1. Sélectionner K échantillons pour représenter les centres des groupes à former,
2. Repartir les échantillons entre les groupes suivant leur proximité par rapport aux centres des groupes,
3. Calculer les centres $\mathbf{c}_k$ des groupes (moyennes des échantillons des groupes),
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Dans la relation (5), 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 du fait 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 une matrice de distances de rang maximum (Herault et al., 2002). Un rang élevé pour la matrice des distances indiquera l'absence de forte redondance dans les échantillons et par conséquent la possibilité d'obtenir un grand nombre de classes. C'est pour cette raison que notre choix s'est porté sur la distance de Chebyshev. La distance de Chebyshev est utilisée avec succès dans l'algorithme K-Means (Estlick et al., 2001).

Étant donné qu'il est souvent difficile de connaître *a priori* le nombre des classes, nous proposons de le déterminer directement à partir des données. L'idée est de former dans la fonction (5) des classes telles que les échantillons membres d'une classe soient plus proches entre eux que par rapport à ceux des autres classes. Cette proximité peut être définie par un seuil sur les distances (Lukashin et Fuchs, 2001). Soit d_{seuil} cette distance seuil, toutes les distances des échantillons d'une classe doivent être au plus égales à d_{seuil}. Le problème consiste 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, les données sont réparties pour obtenir une valeur maximale pour K, puis un regroupement de certaines classes est effectué.

3.1 Algorithme

La procédure de la nouvelle méthode de classification 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.

Soit D le nombre total des distances de toutes les paires d'échantillons. La première étape consiste à calculer les $D = \frac{N(N-1)}{2}$ distances. La médiane de ces distances (voir plus loin), permet d'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. La figure 1 illustre les détails de l'étape 3. 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 on fait varier le nombre des classes, $k = 2, \ldots, Kmax$, puis on retient la valeur de k pour laquelle un critère de validation est optimal (Milligan et Cooper, 1985).

3.2 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 (normalisation L_2). La standardisation est particulièrement utile pour les données d'expression absolues de biopuces, celles-ci peuvent avoir des amplitudes très différentes alors que l'on est intéressé par la variation des profils. La normalisation L_2 rend toutefois les données sphériques. Une autre transformation peut consister à ramener toutes les valeurs des données dans un intervalle compris entre 0 et 1. Cela est obtenu en otant la valeur minimale des données de chaque composante et en divisant le résultat par l'étendue, c'est-à-dire, la différence entre les valeurs maximale et minimale. Un inconvénient de cette transformation est sa sensibilité aux échantillons aberrants, c'est-à-dire ceux qui ont des valeurs éloignées comparées à la majorité. Les échantillons aberrants (ayant des valeurs très faibles ou très élevées) conduiront à l'obtention d'une étendue grande et par conséquent à un tassement de la majorité des valeurs des données. Ceci doit être pris en compte par un prétraitement où les échantillons ayant des valeurs aberrantes sont corrigés ou écartés de l'analyse. Un exemple de prétraitement pour des données de biopuces consiste à éliminer toutes les valeurs inférieures à un seuil pour tous les échantillon (Lukashin et Fuchs, 2001).

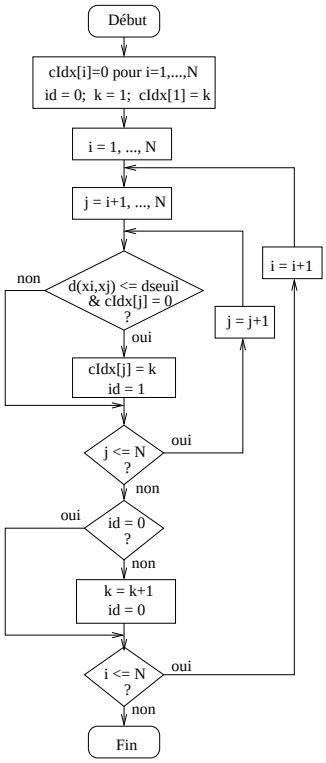

FIG. 1 – *Répartition des données en utilisant le seuil d_{seuil} (étape 3). Initialement l'index du groupe de chaque échantillon est mis à zéro sauf le premier qui est supposé appartenir au groupe 1 (k=1). Étant donné un échantillon x_i appartenant au groupe k, nous plaçons dans le même groupe tous les échantillons x_j non encore classés et pour lesquels $d(\mathbf{x}_j, \mathbf{x}_i) \leq d_{seuil}$. Un indicateur (id) permet de contrôler l'incrémentation du nombre de groupes.*

3.3 Détermination du seuil des distances

Le nombre maximal de classes K_{max} dans les données dépend de la valeur du seuil d_{seuil} utilisé. 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 conduira à la génération de beaucoup de classes singletons. Nous avons fait des tests sur des données synthétiques et réelles pour pouvoir choisir la valeur de d_{seuil}.

Considérons des données contenant deux groupes dans l'espace de dimension 2. Supposons que les échantillons dans chaque groupe ont une distribution normale de moyennes μ_1 et μ_2 et de même variance σI (I = matrice identité de dimension 2). Soient a, b, et ϵ respectivement la plus grande distance entre deux échantillons d'un même groupe (étendue du groupe), la distance entre les centres des deux groupes et la plus petite distance entre un échantillon du groupe 1 et un autre du groupe 2 (distance entre les groupes). Si $b > \epsilon > a$, alors l'histogramme des D distances de Chebyshev des données possède deux pics, figure 2.A. Quand la distance entre

les groupes diminue ($\epsilon \rightarrow 0$), on observe un rapprochement des deux pics de l'histogramme des distances, figures 2.B, 2.C. Notons que le nombre de pics n'indique pas le nombre de

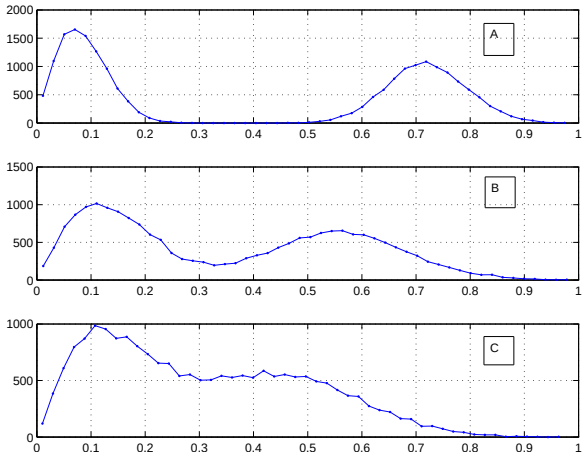

FIG. 2 – *Histogrammes des distances de données synthétiques contenant 2 groupes dans l'espace de dimension 2. Les échantillons dans chaque groupe ont une distribution normale de moyennes* $\mu_1 = [0\ 0]$ *et* $\mu_2 = [4\ 0]$. *La variance est la même pour les 2 groupes et a pour valeur* $\sigma_1 = 0.1$, $\sigma_2 = 0.5$ *et* $\sigma_3 = 0.7$ *respectivement pour les données de (A), (B) et (C).*

groupes dans les données. Cela est montré dans la figure 3.A pour des données de dimensions 10 contenant 14 groupes bien séparés. Les figures 3.B, 3.C et 3.D montrent les histogrammes des distances de Chebyshev pour les données de l'iris, du serum et de la levure respectivement. Les données de l'iris sont des mesures de la longueur et de la largeur des sépales et des pétales de trois espèces d'iris : Setosa, Virginica et Versicolor. Elles peuvent être récupérées à l'adresse suivante : *ftp ://ftp.ics.uci.edu/pub/machine-learning-databases/*. Les données du sérum et de la levure sont présentées dans le paragraphe 4.

Les figures 2 et 3 montrent que pour des données contenant des groupes tels que $b > \epsilon > a$, la valeur de d_{seuil} peut être choisie directement à partir de l'histogramme. Par exemple d_{seuil} vaut respectivement 0.4 et 0.3 pour les données des figures 2.A et 3.A. Ceci n'est pas possible quand la plus petite distance entre les groupes est plus petite que la plus grande étendue des groupes ($\epsilon \leq a$). Cette situation semble le cas pour les données de biopuces. Dès que la distance ϵ entre les groupes devient plus petite que la plus grande étendue a des groupes, la simple lecture de l'histogramme des distances ne permet pas d'avoir d_{seuil}. La présence de plus d'un pic dans l'histogramme indique la présence de groupes dans les données. Toutefois, quand un seuil pic est présent, sa position par rapport à l'origine peut servir à tester la présence de groupes dans les données, voir le test basé sur la moyenne des distances dans (Bock, 1985). De façon générale, le premier pic de l'histogramme correspond aux faibles distances c'est-à-dire des distances qui proviennent des échantillons proches des centres de leurs groupes respectifs. Ainsi en supposant que ces échantillons correspondent à $\tau\%$ des données, nous pouvons estimer le nombre de distances donnant le premier pic de l'histogramme. Cela a été fait pour différentes valeurs de τ et pour différentes données et représenté dans le tableau 1. Le choix

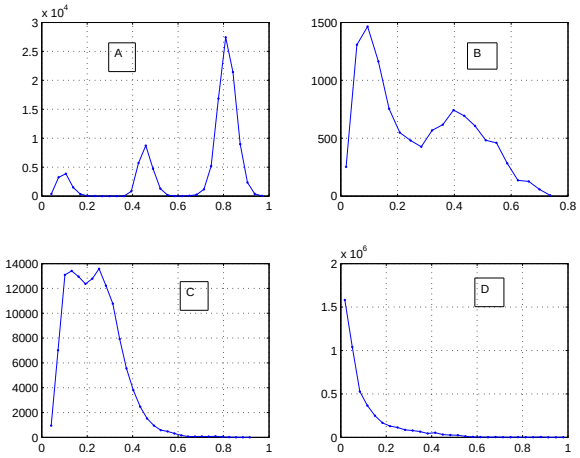

FIG. 3 – *Histogrammes des distances de données (A) synthétiques contenant 14 groupes bien séparés et de dimension 10, (B) des données de l'iris, (C) du sérum et de (D) la levure, voir texte.*

retenu consiste à choisir τ de manière à avoir un seuil d_{seuil} se situant à la moitié des D distances ordonnées. Ceci est obtenu après résolution de l'équation $\frac{\tau N(\tau N-1)}{2} = \frac{N(N-1)}{4} = \frac{D}{2}$ où $\tau \in]0\ 1[$. On obtient ainsi $\tau = \frac{1+\sqrt{1+2N(N-1)}}{2N}$, cette expression est voisine de 0.707 dès que N dépasse 500. Ainsi le seuil d_{seuil} est estimé par la médiane des D distances. Les résultats du tableau 1 montrent que le premier pic de l'histogramme des distances des données synthétiques $y14C$ correspond à un plus faible pourcentage d'échantillons. Les données de l'iris contiennent 3 groupes dont 2 ayant des échantillons qui se chevauchent. Quant aux données de biopuces du sérum et de la levure les nombres de 10 et 30 groupes ont été utilisé dans la littérature, voir (Iyer et al., 1999) et (Tavazoie et al., 1999). Les résultats du tableau 1 pour les données réelles montrent que 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 avoir une valeur du seuil d_{seuil}.

3.4 Critère d'arrêt

Les performances d'une trentaine de procédures permettant de déterminer le nombre de classes dans des données sont comparées dans (Milligan et Cooper, 1985). Pour apprécier la qualité de l'affectation de l'échantillon $\mathbf{x}_i$ dans la classe $\mathcal{C}_k$, Rousseeuw a proposé une mesure appelée silhouette (Rousseeuw, 1987; Kaufman et Rousseeuw, 1990). La valeur de la silhouette est comprise entre -1 et 1. Une valeur négative correspond à une mauvaise affectation de l'échantillon considéré. Le nombre de classes est égal à la valeur de K pour laquelle la moyenne des silhouettes de tous les échantillons des données est maximale. Récemment, une nouvelle méthode basée sur une statistique appelée "Gap" est proposée dans (Tibshirani et al.,

	$\tau\%$	50%	70.7%	75%	80%	90%
y14C	d_{seuil}	0.4987	0.7974	0.8058	0.8158	0.8397
(480,10)	K_{max}	3	1	1	1	1
Iris	d_{seuil}	0.1025	0.2307	0.2948	0.3461	0.4615
(150,4)	K_{max}	10	5	5	5	5
Serum	d_{seuil}	0.1444	0.2233	0.2428	0.2663	0.3247
(517,13)	K_{max}	45	23	14	12	7
Yeast	d_{seuil}	0.0256	0.0548	0.0667	0.0873	0.1653
(2945,15)	K_{max}	86	50	46	35	13

TAB. 1 – *Valeurs estimées de d_{seuil} et de K_{max} en fonction du pourcentage $\tau\%$ des données supposées utilisées dans le premier pic de l'histogramme. Les données synthétiques y14C et réelles de l'iris, du sérum et de la levure sont utilisées.*

2001). Toutes ces méthodes ont été proposées pour des algorithmes de classification basés sur la distance Euclidienne.

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 alors la co-variation des profils comme le coefficient de corrélation des écarts non centrés observés pour les valeurs successives de chaque profil. Le mot coefficient de co-expression (*ce*) est utilisé dans (Moller-Levet et Yin, 2005) pour des données temporelles de biopuces d'expression. Pour ce type de données en temps continu, la variation des profils est obtenue en prenant la dérivée. En temps discrétisé, le vecteur $\Delta\mathbf{c}_i$ obtenu avec les écarts associés au profil $\mathbf{c}_i$ est :

$$\Delta\mathbf{c}_i = c_{ik} - c_{i(k-1)} \;\; ; \;\; k = 2, \ldots, n \;\; \text{avec} \;\; \Delta c_{k1} = 0 \tag{6}$$

Pour deux profils $\mathbf{c}_i$ et $\mathbf{c}_j$, le coefficient de co-variation (ccv) est défini par :

$$ccv(\mathbf{c}_i, \mathbf{c}_j) = \frac{\sum_{k=1}^{n} \Delta c_{ik} \Delta c_{jk}}{\left(\sum_{k=1}^{n} \Delta c_{ik}^2 \sum_{k=1}^{n} \Delta c_{jk}^2 \right)^{\frac{1}{2}}} \tag{7}$$

Les ccv ont été convertis en distances qui sont ensuite utilisées dans la méthode de classification hiérarchique ascendante. La distance de saut minimal (*single linkage*) a été utilisée pour aggréger les classes.

4 Résultats

Pour illustrer les performances de la méthode proposée, nous avons utilisé des données issues de la technologie des biopuces pour l'étude de l'expression des gènes. Les biopuces sont des petits supports (lames de verre) sur lesquels plusieurs milliers de séquences d'ADN (Acide DésoxyriboNucléique) correspondant chacune à un gène sont attachées à des adresses connues

(spots). L'ARN (Aide RiboNucléique) 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.

4.1 Données utilisées

4.1.1 Données de sérum

Ces données sont décrites et utilisées dans (Iyer et al., 1999). Elles proviennent d'une étude sur la réponse de fibroblastes humains au sérum au cours du temps ($n = 13$ conditions). Nous avons utilisé les données correspondant à une sélection de $N = 517$ gènes. Ces données peuvent être récupérées à l'adresse suivante : *http ://www.sciencemag.org/ feature/data/984559.shl*.

4.1.2 Données de la levure

Nous nous sommes aussi servi des données issues de l'étude du profil d'expression de 6200 gènes chez la levure (Cho et al., 1998) réalisé sur deux cycles mitotiques au cours desquels les valeurs ont été mesurées toutes les 10 minutes (soit 17 hybridations différentes). Nous avons utilisé les données d'une sélection de $N = 2945$ gènes, c'est-à-dire la sélection faite dans (Tavazoie et al., 1999). Dans cette sélection les valeurs qui correspondent aux instants 90 et 100 minutes sont exclues, soit $n = 15$.

4.2 Résultats obtenus

Avant de présenter les résultats obtenus avec la procédure proposée, nous avons déterminé le rang de la matrice des distances pour les données du sérum et de la levure. Le rang de la matrice des distances est donné par le nombre de ses valeurs singulières supérieures à 0.001. Le tableau 2 montre le rang de la matrice des distances de Minkowski et de Chebyshev pour les données du sérum et de la levure. Le rang obtenu est systématiquement plus grand pour la distance de Chebyshev.

	Minkowski				Chebyshev
Norme	L_2	L_4	L_8	L_{20}	[0,1]
Sérum	13 *(15)*	38 *(41)*	90 *(93)*	217 *(249)*	514
Levure	15 *(17)*	44 *(47)*	104 *(107)*	269 *(287)*	2943

TAB. 2 – *Rang de la matrice des distances pour la distance de Chebyshev et pour différentes normes de la distance de Minkowski. La valeur maximale théorique du rang est indiquée entre parenthèses pour les données L_p normées.*

4.2.1 Données de sérum

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 4.

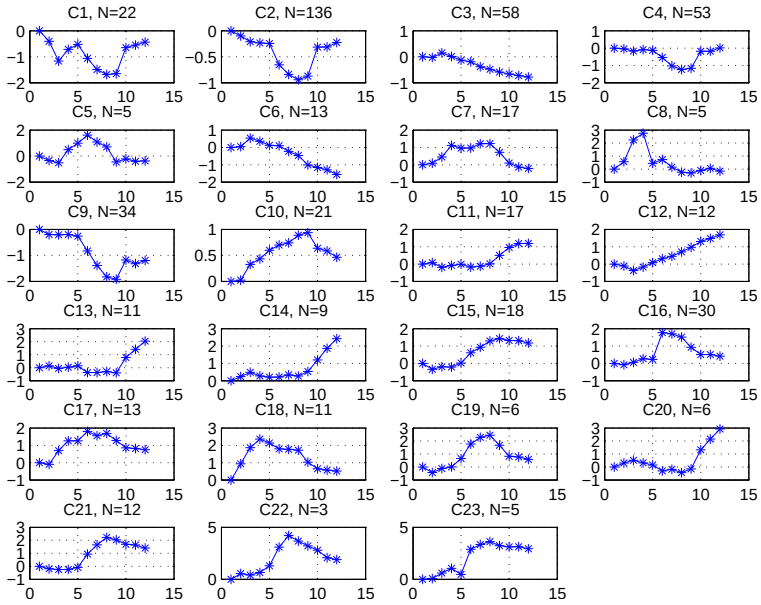

FIG. 4 – *Données de sérum, les* 23 *classes obtenues dans la phase initiale. Chaque profil est identifié par un numéro et le nombre des échantillons qui la forme, e.g. la première classe* $C1$ *contient* 22 *échantillons (*$N = 22$*).*

Un extrait de la matrice des ccv des 23 profils initiaux des données de sérum est donnée dans les tableaux 3 et 4. La similitude des profils des classes 13, 14 et 20 est indiquée dans le tableau 4 par des ccv élevés, ≥ 0.75. La valeur du ccv est la plus petite, -0.87, pour les classes 6 et 12 montrant une opposition de profils pour ces classes. La figure 4 contient tous les profils obtenus dans (Iyer et al., 1999) avec des redondances. Avec la méthode hiérarchique ascendante les classes ont été regroupées.

4.2.2 Données de la levure

En appliquant l'algorithme proposé aux données des 2945 gènes, nous avons obtenu une distance seuil $d_{seuil} = 0.0549$ et un nombre maximum de classes $K_{max} = 50$. Parmi les 50 profils obtenus, certains sont connus pour ces données (Tavazoie et al., 1999). Les profils de certaines classes sont représentés sur les figures 5 et 6.

Nous avons calculé la matrice des ccv des 50 classes obtenues puis cette matrice est utilisée dans la méthode hiérarchique ascendante à saut minimal. Les classes 14 et 31 ont été fusionnées

	1	2	3	4	5	6	7	8	9	10
2	0.84	0								
3	-0.20	0.13	0							
4	0.84	0.95	0.11	0						
5	0.31	-0.05	-0.06	0.13	0					
6	-0.35	-0.23	0.48	-0.14	0.26	0				
7	-0.44	-0.60	-0.03	-0.56	0.39	0.45	0			
8	-0.28	-0.05	0.48	0.07	-0.06	0.51	0.30	0		
9	0.75	0.91	0.36	0.92	0.18	0.11	-0.41	0.11	0	
10	-0.64	-0.72	-0.16	-0.69	0.03	0.35	0.62	0.07	-0.62	0
11	0.61	0.59	-0.45	0.56	-0.34	-0.74	-0.65	-0.22	0.28	-0.44
12	0.41	0.29	-0.31	0.28	0.02	-0.87	-0.48	-0.39	0.02	-0.38
13	0.68	0.75	-0.25	0.77	0.02	-0.33	-0.47	-0.08	0.57	-0.60
14	0.17	0.36	-0.25	0.42	-0.26	-0.17	-0.44	0.19	0.21	-0.27
15	-0.22	-0.22	0.14	-0.27	-0.04	-0.47	-0.17	-0.18	-0.35	0.04

TAB. 3 – *Extrait de la matrice des ccv des* 23 *classes initialement obtenues pour les données de sérum.*

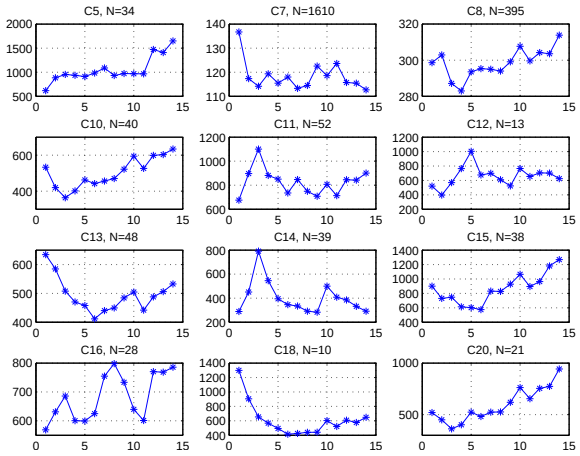

FIG. 5 – *Données de la levure.*

les premières, à cette nouvelle classe, la classe 32 a été ajoutée à la huitième étape puis la classe 12 à la douzième étape de la méthode hiérarchique.

4.2.3 Discussion

Avec les données de la levure, nous avons obtenu une classe, $C7$, contenant plus de la moitié des échantillons dans les données. En observant les niveaux d'expression des gènes de cette classe, on note que le profil d'expression moyen varie entre 110 et 120. Cette valeur est

	12	13	14	15	16	17	18	19
13	0.43	0						
14	0.23	0.75	0					
15	0.66	-0.24	-0.22	0				
16	0.05	-0.47	-0.25	0.44	0			
17	-0.08	-0.42	-0.29	0.38	0.67	0		
18	-0.44	-0.12	-0.08	-0.20	0.08	0.55	0	
19	-0.06	-0.56	-0.38	0.42	0.76	0.59	0.08	0
20	0.38	0.90	0.87	-0.14	-0.42	-0.38	-0.08	-0.62
21	0.31	-0.44	-0.27	0.69	0.67	0.36	-0.15	0.78
22	0.19	-0.42	-0.28	0.49	0.73	0.33	0.08	0.68
23	0.43	-0.27	-0.13	0.78	0.78	0.60	0.10	0.47

TAB. 4 – *Extrait de la matrice des ccv des* 23 *classes initialement obtenues pour les données de sérum (suite).*

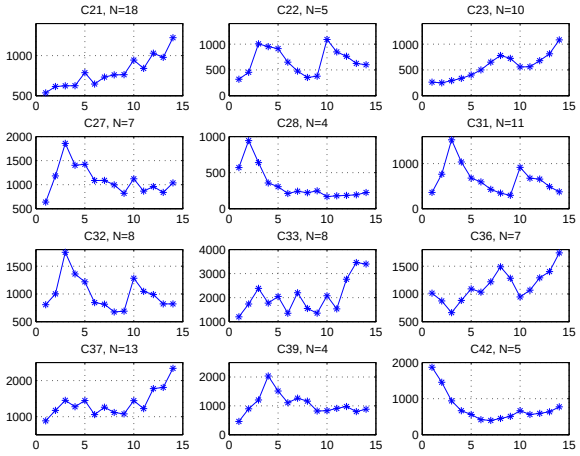

FIG. 6 – *Données de la levure (suite), sélection de* 24 *classes obtenues dans la phase initiale. Chaque profil est identifié par un numéro et le nombre des échantillons qui la forme, e.g. la première classe* $C14$ *contient* 39 *échantillons (*$N = 39$*).*

très faible comparée à la variation globale $(0 - 9500)$. Ce type de résultat n'est pas obtenu pour les données du sérum qui sont des ratios d'expression dont les valeurs varient entre -3 et 3. La classe $C7$ contient tous les échantillons de niveau d'expression moyen faible. La normalisation $[0, 1]$ a ramené ces valeurs au voisinage de zéro. Nous avons extrait les 1610 échantillons de la classe $C7$. La procédure proposée est ensuite appliquée à cette nouvelle sélection. Ceci nous a permis d'obtenir des profils similaires à certains présents dans les figures 5 et 6.

La normalisation $[0, 1]$ permet d'obtenir des classes contenant des échantillons de profil moyen similaire. Nous pouvons ainsi séparer des gènes fortement exprimés de ceux faiblement

exprimés. Ceci est masqué par la normalisation L_2. Le ccv permet ensuite de regrouper les classes de profils similaires indépendamment des niveaux des profils moyens observés. Le ccv n'est pas parfait, voir sa valeur (cas des données du sérum) pour les classes 3 et 6 qui sont assez similaires puis pour les classes 4 et 13 qui ne sont pas similaires. Cependant il a donné des résultats semblables à ceux présentés dans (Iyer et al., 1999) où le nombre de classes a été fixé *a priori* à 10. La procédure proposée est très rapide. L'étape de calcul des distances, la plus coûteuse en temps est 6 fois plus rapide avec la distance de Chebyshev qu'avec la distance Euclidienne. Pour les données avec N très élevé, une sélection aléatoire de quelques milliers d'échantillons peut être suffisant pour obtenir une estimation de la valeur du seuil d_{seuil}.

5 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 grande dimension. 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 réaffecter 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 futur consistera à étudier le coefficient de co-variation plus en détail. Une interface conviviale pourra également faciliter l'exploitation des résultats de classification.

Remerciements

Merci à Philippe Kastner, Bernard Jost et Christelle Thibault qui ont pris le temps de lire cet article. Ce travail a bénéficié du soutien du Centre National de la Recherche Scientifique, de l'Institut National de la Recherche Médicale, de l'Hôpital Universitaire de Strasbourg et du Centre National de Recherche en Génomique.

Références

Beyer, K., J. Goldstein, R. Ramakrishnan, et U. Shaft (1998). When Is "Nearest Neighbor" Meaningful ? In P. B. Catriel Beeri (Ed.), *Proc. of the ICDT'99*, Volume LNCS 1540, pp. 217–235. Springer-Verlag Berlin Heidelberg.

Bock, H. H. (1985). On Some Significance Tests in Cluster Analysis. *Journal of Classification 2*, 77–108.

Bonnet, N., M. Herbin, J. Cutrona, et J.-M. Zahm (2002). A New Clustering Approach, Based on the Estimation of the Probability Density Function, for Gene Expression Data. In *IFCS symposium, july 15-20, Kraköw, Poland*, pp. pp–pp+7.

Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, et R. W. Davis (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell 2*, 65–73.

Demartines, P. (1994). *Analyse de données par réseaux de neurones auto-organisés*. Ph. D. thesis, TIRF, INPG, Grenoble, France.

Dohono, D. L. (2000). 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*.

Estlick, M., M. Leeser, J. Theiler, et J. J. Szymanski (2001). Algorithmic Transformations in the Implementation of K- Means Clustering on Reconfigurable Hardware. In *FPGA*, pp. 103–110.

Everitt, B. S. (1993). *Cluster Analysis* (3rd ed.). Arnold, London.

Golub, G. H. et C. F. V. Loan (1996). *Matrix Computations*. The Johns Hopkins University Press, Third Edition.

Herault, J., A. Guérin-Dugué, et P. Villemain (2002). 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*, pp. 173–184.

Holter, N. S., M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, et N. V. Fedoroff (2000). Fundanmental Patterns Underying Gene Expression Profiles : Simplicity from Complexity. *PNAS 97*(15), 8409–8414.

Horn, D. et I. Axel (2003). Novel Clustering Algorithm for Microarray Expression Data in a Truncated SVD Space. *Bioinformatics 19*(9), 1110–1115.

Iyer, V. R., 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, et P. O. Brown (1999). The Transcriptional Program in the Response of Human Fibroblast to Serum. *Science 283*, 83–87.

Jain, A. K. et R. C. Dubes (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliff, New Jersey.

Jain, A. K., R. P. W. Duin, et J. Mao (2000). Statistical Pattern Recognition : A Review. *IEEE trans. PAMI 22*(1), 4–37.

Jimenez, J. O. et D. Landgrebe (1995). High Dimension Feature Reduction Via Projection Pursuit. Technical Report TR-ECE 96-5, School of Electical and Computer Engineering, Purdue University.

Kaufman, L. et P. Rousseeuw (1990). *Finding Group in Data : an Introduction to Cluster Analysis*. Wiley, New York.

Lukashin, A. V. et R. Fuchs (2001). Analysis of Temporal Gene Expression Profiles : Clustering by Simulated Annealing and Determining the Optimal Number of Clusters. *Bioinformatics 17*(5), 405–414.

Milligan, G. W. et M. C. Cooper (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika 50*(2), 159–179.

Moller-Levet, C. S. et H. Yin (2005). Modeling and Analysis of Gene Expression Time-Series Based on Co-Expression. *Int J of Neural Syst. 15*(4), 311–322.

Rousseeuw, J. P. (1987). Silhouettes : a Graphical Aid to the Interpration and Validation of Cluster Analysis. *J. Comp. Appl. Math. 20*, 53–65.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. I. Cho, et G. M. Church (1999). Systematic Determination of Genetic Network Architecture. *Nature Genetic 22*, 281–285.

Tibshirani, R., G. Walther, et T. Hastie (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal Royal Stat Society (B) 63*(2), 441–423.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *Journal of American Statistical Association 77*(380), 841–847.

Yeung, K. Y. et W. L. Ruzzo (2001). Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics 17*(9), 763–774.

Summary

In this paper, we proposed a new clustering method especially suitable for high dimensional data. This method combines partitioning and hierarchical clustering algorithms for getting groups in a given data set. We also used Chebyshev distance which seems interesting for high dimensional. Indeed, less computationally load is required in comparison with Euclidean distance frequently used because of its nice geometrical properties. Microarray data are used for illustrating performances of our findings.