Classification en référence à une matrice stochastique

Stéphane Verdun, Véronique Cariou, El Mostafa Qannari

Unité de Sensométrie et Chimiométrie
ENITIAA, rue de la Géraudière - BP82 225
44322 Nantes Cedex 3 - France
{prénom.nom}@enitiaa-nantes.fr

Résumé. Étant donné un tableau de données portant sur un ensemble d'objets et une matrice stochastique qui peut être assimilée à une matrice de passage d'une marche aléatoire, nous proposons une méthode de partitionnement consistant à identifier les différents états stationnaires associés à la matrice stochastique. Les classes formant la partition sont déterminées à partir de ces états stationnaires. De manière pratique, la matrice stochastique peut être définie à partir d'une matrice de mesure de ressemblance entre les objets du tableau de données. Différentes mesures sont étudiées (plus proches voisins, noyaux de densité...). L'approche est présentée sur la base des propriétés des graphes et illustrée sur des données simulées. La méthode de classification proposée est également comparée à une autre approche de classification, la classification spectrale.

1 Introduction

Une méthode de classification basée sur une matrice stochastique est proposée. La matrice stochastique reflète une affinité entre les objets à classer et sera généralement déduite d'une matrice de mesure de ressemblance. L'intérêt de cette méthode de classification est sa grande flexibilité du fait de la prise en compte de différentes mesures de ressemblance (mesures de similarités, mesures de densité, ...). Il est ainsi possible de l'adapter à différents contextes (e.g. classification d'individus, classification de variables) et à plusieurs configurations (séparations linéaires ou non linéaires des classes). Cependant, cette méthode de classification exige que les classes soient bien séparées et qu'il n'y ait pas d'objets qui créent un effet de chaînage. C'est pourquoi une étape consistant à rechercher les individus isolés s'avère généralement nécessaire.

La démarche de classification proposée est intimement liée aux marches aléatoires sur un graphe (Lovasz, 1993) car les classes sont déterminées en considérant la matrice de convergence de la suite des puissances de la matrice stochastique. En effet, cette suite converge vers une matrice qui permet d'identifier les différentes classes. De plus, la méthode de classification permet d'associer à chaque objet une probabilité qui indique son degré de centralité dans la classe à laquelle il appartient. Il faut également souligner que tous les objets ne sont pas affectés de manière déterministe aux différentes classes. En effet, certains objets qui ne constituent pas les "noyaux forts" des classes sont caractérisés par un vecteur de probabilité qui

indique leur degré d'appartenance aux différentes classes, reflétant ainsi leur positionnement particulier par rapport à l'ensemble des classes.

La classification proposée présente une certaine similitude avec la méthode de classification dite classification spectrale (von Luxburg, 2007). Cette méthode vise, en effet, à partitionner les objets en identifiant les groupes à partir de l'analyse spectrale d'une matrice de similarités entre les objets. Dans certains travaux (Shi et Malik, 2000; Ng et al., 2001), les auteurs proposent une normalisation de la matrice analysée. En particulier dans l'algorithme de Shi et Malik, la normalisation consiste à rendre cette matrice stochastique. Dans ce cas, le type de matrice analysée en classification spectrale est le même que celui considéré dans l'article.

Le plan du reste de l'article est le suivant. Dans la partie 2, la méthode est présentée en se référant à la théorie des graphes. Une méthode d'obtention de la matrice stochastique est décrite dans la partie 3 sur la base de différentes mesures de ressemblance. Dans la partie 4, les liens avec la classification spectrale sont discutés et enfin dans la partie 5, la méthode est illustrée sur la base de données simulées.

2 Méthode

2.1 Données et notations

Soit X une matrice comprenant p variables qui sont mesurées sur un ensemble Ω constitué de N individus. Soit $P = (p_{ij})_{1 \leq i,j \leq N}$ une matrice carrée, stochastique en ligne, à N lignes et N colonnes qui mesure l'affinité entre les objets. Chaque ligne de P est un vecteur de probabilité (composantes positives de somme égale à 1). La i^e ligne mesure les affinités de l'objet i aux autres objets. Il faut souligner que la matrice P n'est pas nécessairement symétrique. Différents modes d'obtention de P à partir de X sont détaillés dans la partie 3.

La norme utilisée dans la suite de l'article est notée $||.||$. Il s'agit de la norme euclidienne usuelle pour les vecteurs et de la norme de Frobenius pour les matrices.

2.2 Principe général de la méthode

Par souci pédagogique, la méthode est présentée sur la base d'un exemple simple. Dans la Table 1, la matrice P décrit l'affinité entre 8 objets. P peut être considérée comme la matrice

$P =$

	A	B	C	D	E	F	G	H
A	0,5	0,125	0,125	0,25	0	0	0	0
B	0,125	0,5	0,125	0,25	0	0	0	0
C	0,125	0,125	0,5	0,25	0	0	0	0
D	0,2	0,2	0,2	0,4	0	0	0	0
E	0	0	0	0	0,6	0,2	0,2	0
F	0	0	0	0	0,2	0,6	0,2	0
G	0	0	0	0	0,2	0,2	0,6	0
H	0	0,25	0	0	0	0	0,75	0

TAB. 1: affinités entre les 8 objets.

$P^{\infty} =$

	A	B	C	D	E	F	G	H
A	0,24	0,24	0,24	0,28	0	0	0	0
B	0,24	0,24	0,24	0,28	0	0	0	0
C	0,24	0,24	0,24	0,28	0	0	0	0
D	0,24	0,24	0,24	0,28	0	0	0	0
E	0	0	0	0	0,33	0,33	0,33	0
F	0	0	0	0	0,33	0,33	0,33	0
G	0	0	0	0	0,33	0,33	0,33	0
H	0,06	0,06	0,06	0,07	0,25	0,25	0,25	0

TAB. 2: degrés de centralité des 8 objets.

d'incidence d'un graphe valué orienté composé de 8 sommets. Deux groupes ou composantes fortement connexes peuvent être aisément identifiés : {A,B,C,D} et {E,F,G}. Par ailleurs, dans la mesure où aucun objet ne possède d'affinité avec H, ce sommet est assimilé à un point isolé qui possède, cependant, des arcs vers les sommets de {A,B,C,D} et {E,F,G}. Considérons une marche aléatoire sur le graphe ainsi défini. Celle-ci est caractérisée par la matrice de passage P telle que la probabilité de passer du sommet i au sommet j est donnée par p_{ij}. Elle conduit à définir deux groupes {A,B,C,D} et {E,F,G} correspondant à des classes finales et le groupe {H} correspondant à une classe transitoire. Ce point sera discuté plus en détail dans la partie suivante (2.3).

Il est possible de calculer les différentes puissances de la matrice P. Cette suite de matrices converge vers une matrice stochastique P^{∞}, donnée dans la Table 2. À partir de P^{∞}, les différentes classes peuvent être déduites. Une classe correspond à l'ensemble des individus possédant une ligne identique dans P^{∞}. Le vecteur de probabilité associé à chaque classe reflète le degré de centralité des objets dans la classe. Ainsi, dans la première classe, l'objet D a un degré de centralité plus élevé que les autres objets. Ceci est cohérent du fait que D a le plus d'affinité avec les autres objets de sa classe. Dans le deuxième groupe, tous les objets ont le même degré de centralité dans la mesure où ils ont les mêmes affinités les uns vis à vis des autres. La ligne correspondant à H contient des valeurs qui indiquent que cet objet peut-être affecté aux deux classes {A,B,C,D} et {E,F,G}. De manière plus précise, nous pouvons définir la probabilité d'appartenir à la première classe comme étant la somme des affinités de H avec les éléments de cette classe, soit 0,25. De manière similaire, la probabilité d'appartenance de H à la deuxième classe est 0,75.

Appliquée au tableau X, la matrice stochastique P^{∞} conduit à $Y^{(\infty)} = P^{\infty}X$. La matrice $Y^{(\infty)}$ est la matrice dans laquelle chaque individu est remplacé par le prototype de sa classe ; ce prototype étant le barycentre des individus pondérés par leur degré de centralité. Pour les éléments transitoires, l'interprétation est légèrement différente. Dans l'exemple, la ligne correspondant à H de la matrice P^{∞} est une combinaison linéaire des vecteurs des degrés de centralité de chaque classe, les coefficients dans la combinaison linéaire de ces vecteurs étant les poids d'affectation de H à ces classes (i.e. 0,25 pour la première classe et 0,75 pour la deuxième classe). De ce fait, dans la matrice $Y^{(\infty)}$, la ligne correspondant à H est la moyenne pondérée par les poids d'affectation des prototypes des classes {A,B,C,D} et {E,F,G}. Plus un individu transitoire sera proche d'une classe, plus la ligne correspondante de $Y^{(\infty)}$ sera proche

du prototype de la classe.

2.3 Identification des classes à partir d'une matrice stochastique

La démarche qui a été illustrée sur la base de l'exemple peut s'étendre aisément au cas général où nous disposons d'une matrice stochastique P et d'un tableau de données X portant sur un ensemble Ω comportant N objets. Une justification de la démarche de classification est présentée en se basant sur une marche aléatoire sur un graphe. Le lecteur souhaitant approfondir les propriétés des marches aléatoires peut se référer à Lovasz (1993), ou, plus généralement sur la théorie des chaînes de Markov, à Norris (1998).

Nous considérons le graphe $G(\Omega, V)$ où Ω est l'ensemble des sommets et V l'ensemble des arcs. Le graphe possède donc N sommets correspondant aux différents individus. Un arc est créé du sommet i au sommet j si l'affinité de l'individu i à l'individu j est non nulle. Une marche aléatoire sur ce graphe est un processus stochastique qui modélise un déplacement aléatoire de sommet en sommet. Elle est définie en prenant la matrice P comme matrice de passage. Cela signifie que la probabilité de passer du sommet i au sommet j est donnée par le coefficient p_{ij}. Parmi les propriétés connues des matrices de passage, notons que pour un entier naturel n ($n > 0$) la matrice $P^n = (p_{ij}^{(n)})$ donne la probabilité de passer d'un sommet à un autre en n étapes. Le sommet j est dit accessible à partir de i s'il existe un chemin de i à j, c'est-à-dire s'il existe un entier $n > 0$ tel que $p_{ij}^{(n)} > 0$. Les sommets i et j communiquent s'il existe un chemin de i à j et un chemin de j à i, c'est-à-dire s'il existe deux entiers $n > 0$ et $m > 0$ tels que $p_{ij}^{(n)} > 0$ et $p_{ji}^{(m)} > 0$. La relation de communication est une relation d'équivalence. Elle définit des classes d'équivalence de sommets, où chaque classe d'équivalence est un ensemble maximal de sommets qui communiquent. De plus, les classes d'équivalence correspondent aux composantes fortement connexes du graphe.

Partant de ces propriétés, nous pouvons distinguer deux types de classes d'équivalence. Une classe est dite finale s'il n'y a aucun chemin partant d'un de ses sommets vers un sommet extérieur. Une marche aléatoire partant d'un sommet de cette classe ne pourra atteindre que les sommets de cette classe. Une classe est dite transitoire s'il y a un chemin partant d'un de ses sommets vers un sommet extérieur à la classe. Une marche aléatoire partant d'un sommet de cette classe atteindra un sommet extérieur à la classe et ne pourra plus y revenir.

Le principe de la méthode est de trouver des ensembles d'individus (sommets) qui correspondent aux classes d'équivalence, et de distinguer classes finales et classes transitoires. Pour chaque classe finale, il existe une loi de probabilité asymptotique donnée par le vecteur de probabilité stationnaire qui précise les probabilités asymptotiques d'atteindre les différents sommets de la classe en partant d'un sommet de cette classe. Ces probabilités définissent les degrés de centralité des individus au sein de la classe. Dans la mesure où le nombre d'individus est fini, ce vecteur est de plus unique.

Pour un individu appartenant à une classe transitoire, les probabilités d'atteindre chacune des classes finales en partant de ce sommet peuvent également être calculées. Ces probabilités peuvent être assimilées à des poids d'affectation de l'individu considéré aux différentes classes.

2.4 Recherche des classes finales et transitoires

Si nous supposons que les individus sont organisés de telle sorte qu'ils sont regroupés par classe et que les éléments transitoires sont en dernières positions, la matrice stochastique P peut être mise sous une forme dite canonique :

$$P = \begin{bmatrix} B & 0 \\ R & Q \end{bmatrix} \tag{1}$$

avec B et Q des matrices carrées. De plus, B est une matrice diagonale par blocs :

$$B = \begin{bmatrix} P_1 & 0 & 0 \\ 0 & \ddots P_k \ddots & 0 \\ 0 & 0 & P_K \end{bmatrix}$$

où les matrices P_k sont des matrices stochastiques correspondant aux classes finales C_k ($k = 1, \cdots, K$). La matrice R est la matrice des probabilités de passage des éléments transitoires vers les éléments des classes finales. Notons r_{ij} la probabilité pour un élément transitoire i de passer à un élément j appartenant à une classe finale quelconque.

La matrice de passage en n étapes est la matrice P^n. Elle peut se calculer en fonction des différents blocs de la matrice P :

$$P^n = \begin{bmatrix} B^n & 0 \\ \sum_{i=0}^{n-1} Q^i R B^{n-i-1} & Q^n \end{bmatrix} \tag{2}$$

avec

$$B^n = \begin{bmatrix} P_1^n & 0 & 0 \\ 0 & \ddots P_k^n \ddots & 0 \\ 0 & 0 & P_K^n \end{bmatrix}$$

et $\lim_n Q^n = 0$.

Le vecteur de probabilité stationnaire π_k de la matrice P_k est le vecteur tel que $\pi_k = \pi_k P_k$, dont tous les coefficients sont positifs et dont la somme vaut 1. Pour une marche aléatoire débutant dans cette classe, le vecteur stationnaire donne les probabilités d'atteindre les différents éléments lorsque le nombre d'étapes n de la marche tend vers l'infini. Cela implique notamment que la suite de matrices $(P_k^n)_{n>0}$ converge vers sa probabilité invariante π_k : $\lim_{n\to+\infty} P_k^n = \begin{pmatrix} \pi_k \\ \pi_k \\ \vdots \\ \pi_k \end{pmatrix}$. Le vecteur π_k donne les degrés de centralité des éléments dans la classe C_k ($k = 1, ..., K$).

Pour obtenir les poids d'affectation des individus transitoires aux différentes classes, un graphe réduit est construit. Dans ce graphe, tous les individus d'une classe finale sont agrégés en un seul sommet. La matrice de transition de ce graphe est la matrice :

$$\tilde{P} = \begin{bmatrix} I_K & 0 \\ \tilde{R} & Q \end{bmatrix} \tag{3}$$

avec I_K la matrice identité de dimension K, Q la même matrice que celle définie dans l'équation 1 et $\tilde{R} = (\tilde{r}_{ik})$, où $\tilde{r}_{ik} = \sum_{j \in C_k} r_{ij}$. Le coefficient $\tilde{r}_{ik}$ correspond à la probabilité de passage de l'élément transitoire i à n'importe quel sommet de la classe k. En reprenant l'équation 2, le terme donnant les probabilités de passage des éléments transitoires aux éléments des classes en n étapes devient :

$$\sum_{i=0}^{n-1} Q^i \tilde{R} \rightarrow_{n \rightarrow \infty} (I - Q)^{-1} \tilde{R}. \tag{4}$$

La matrice $(I - Q)^{-1}\tilde{R}$ donne, ainsi, les poids d'affectation des individus transitoires aux différentes classes.

2.5 Algorithme

La trame de l'algorithme de classification est la suivante :

- Rechercher les classes d'équivalence associées à la matrice P. Cela peut se faire par la recherche des composantes fortement connexes du graphe (par exemple en utilisant l'algorithme de Tarjan (1972)).
- Pour chacune des classes, vérifier s'il s'agit d'une classe finale. La classe C est finale si $p_{ij} = 0$, pour tout i dans la classe et pour tout j hors de la classe.
- Pour chaque classe finale C_k, rechercher le vecteur de probabilité stationnaire π_k de la matrice P_k.
- Calculer la matrice $\tilde{P}$ selon l'équation (3).
- Calculer $(I - Q)^{-1}\tilde{R}$ et affecter les éléments transitoires aux différentes classes selon leurs coefficients.
- Calculer les prototypes des classes $Y^{(\infty)} = P^{\infty}X$.

3 Détermination de la matrice stochastique

3.1 Mesures de ressemblance

La détermination de la matrice stochastique est une étape importante. Les problèmes de chaînage qui peuvent survenir doivent être identifiés a priori car ils empêchent la détection des classes. Dans le cadre de ce travail, la matrice stochastique P est déterminée à partir d'une matrice S de mesure de ressemblance entre individus calculées à partir du tableau X. À la différence d'une mesure de similarité, cette mesure n'est pas nécessairement symétrique. Pour calculer P, nous procédons à la normalisation : $P = D^{-1}S$, où D est la matrice diagonale dont le i^e élément diagonal est égal à la somme de la i^e ligne de S. Des exemples de mesures qui nous semblent appropriées pour la classification sont discutés ci-après.

1. **Plus Proches Voisins :** $s_{ij} = \begin{cases} 1 & \text{si } x_j \text{ est dans les } k \text{ plus proches voisins de } x_i, \\ 0 & \text{sinon .} \end{cases}$

   k est un paramètre à fixer a priori.

2. **Boule :** $s_{ij} = \begin{cases} 1 & \text{si } ||x_i - x_j|| \leq r, \\ 0 & \text{sinon.} \end{cases}$
   r est un paramètre à fixer a priori.
3. **Gaussienne :** $s_{ij} = \begin{cases} \exp(-\frac{1}{2\sigma^2}||x_i - x_j||^2) & \text{si } ||x_i - x_j|| \leq r, \\ 0 & \text{sinon.} \end{cases}$
   σ et r sont deux paramètres à fixer a priori. Il est toutefois possible de se ramener à un seul paramètre en liant r à σ. En considérant la distribution gaussienne associée à cette fonction. Cela revient à éliminer les queues de la distribution en prenant par exemple $r = 1.96\sigma$.
4. **Voisinage :** Notons V_i le voisinage de x_i, défini par les k plus proches voisins de x_i, et V_j celui de x_j. La similarité de voisinage est définie par :

   $$s_{ij} = \begin{cases} \frac{card(V_i \cap V_j)}{card(V_i \cup V_j)} & \text{si } x_j \in V_i \text{ et si ce rapport est supérieur à une proportion } p_0, \\ 0 & \text{sinon.} \end{cases}$$

   k et p_0 sont deux paramètres à fixer a priori. Le principe sous-jacent à cette mesure de ressemblance qui est, à notre connaissance, originale est que la mesure entre deux individus est d'autant plus grande que ces individus partagent une proportion importante de voisins (supérieure à p_0).

Le choix des paramètres des mesures de ressemblance est important. À cet effet, nous pouvons définir une mesure d'homogénéité de la partition par $||Y^{(\infty)} - X||$. Rappelons que la matrice $Y^{(\infty)}$ est la matrice où chaque individu est remplacé par son prototype. Cette mesure est proche de l'inertie intra-classe. En effet, si nous considérons le cas particulier où il n'y a pas d'éléments transitoires et qu'il y a uniquement K classes finales, nous avons :

$$||Y^{(\infty)} - X||^2 = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - m_k||^2$$

où m_k est le prototype de la classe k.

Le choix d'un paramètre de la mesure de ressemblance considérée peut s'effectuer à partir du graphe représentant pour différentes valeurs du paramètre, l'indicateur d'homogénéité $||Y^{(\infty)} - X||$.

Si une variation du paramètre ne fait pas varier la partition en ce sens que le nombre de classes finales, K, et l'indicateur $||Y^{(\infty)} - X||$ sont stables, alors nous pouvons considérer qu'il s'agit d'une plage de valeurs judicieuses.

3.2 Recherche des individus isolés

Comme nous l'avons déjà souligné, la méthode de classification en référence à une matrice stochastique est sensible aux effets de chaînage. Un bon choix des paramètres des mesures de ressemblance peut limiter ces effets. De même, l'identification des individus responsables de ces problèmes permet également d'y remédier. Le principe est de localiser les individus qui créent un lien entre deux classes. Typiquement, un tel individu sera un individu pour lequel peu d'individus auront une ressemblance strictement positive. La recherche de ces individus

peut s'effectuer en considérant pour chaque point la moyenne des ressemblances que les autres individus ont avec lui. Si nous notons la matrice de ressemblances $S = (s_{ij})_{1 \leq i,j \leq N}$, il faut calculer pour tout j ($j = 1, ..., N$) $\overline{s_{.j}} = \frac{1}{N} \sum_{i=1}^{N} s_{ij}$. Un individu j qui a une moyenne $\overline{s_{.j}}$ faible est susceptible de créer des problèmes de chaînage. Nous proposons de faire en sorte que cet individu apparaisse comme un individu transitoire. Pour cela, il suffit d'imposer que les ressemblances s_{ij} ($i = 1, ..., N$ et $i \neq j$) soient nulles.

Deux stratégies peuvent être envisagées pour déterminer le seuil sur la moyenne des ressemblances en dessous duquel les individus seront considérés comme étant isolés. La première consiste à partir d'une matrice de ressemblance S fixée (c'est-à-dire que les paramètres de la mesure de ressemblance sont fixés). Pour tous les individus, les moyennes des ressemblances $\overline{s_{.j}}$ sont calculées. Ensuite, à partir de l'histogramme de ces valeurs, il s'agit de repérer le premier mode principal, qui contient beaucoup d'individus. Ce mode correspond généralement aux individus d'une classe. Par conséquent, le seuil sur $\overline{s_{.j}}$ doit être fixé avant ce premier mode. Une deuxième stratégie pour l'identification de points isolés qui peut être appliquée indépendamment de la variation du paramètre de la mesure de ressemblance, consiste à retirer un pourcentage fixe des individus (10% ou 15%, par exemple). Les individus ayant les plus faibles moyennes sont retirés. Cette méthode comporte le risque de prendre un seuil trop élevé et donc de supprimer des individus de classes. Toutefois, comme les individus seront identifiés à l'issue de la classification comme des individus transitoires, ils seront, de ce fait, affectés de poids qui indiquent leurs proximités aux différentes classes. Ainsi, il est tout à fait possible de réaffecter un individu, écarté à tort, à sa propre classe. Ce traitement a un rapport avec les techniques de classification par niveau de densité (Wishart, 1969; Hartigan, 1975). Brièvement, ces méthodes supposent que les groupes correspondent à des modes de la densité $p(x)$. Le but est de trouver ces modes et d'affecter chaque individu à son domaine d'attraction. Les classes de haute densité définies par Hartigan sont les composantes connexes de $L(\lambda; p) = \{x | p(x) > \lambda\}$. Les points où la densité est inférieure à ce seuil ne sont pas des points de classe. Le parallèle avec la stratégie proposée ici est évident, mais plutôt que d'utiliser une densité et un estimateur pour cette densité, la moyenne des mesures de ressemblance avec chaque individu est choisie comme alternative.

4 Comparaison de méthodes

Une méthode de classification appelée classification spectrale a récemment été développée (Shi et Malik, 2000; Meila et Shi, 2000; Ng et al., 2001). Partant d'une similarité déterminée à partir d'un tableau de donnée X, le principe général de la classification consiste à changer d'espace de représentation et à appliquer un algorithme de classification sur la base des nouvelles coordonnées. De manière plus précise, soit S une matrice de similarité déterminée à partir de X. S est apparentée à la matrice d'adjacence pondérée d'un graphe de similarité. Soit D la matrice diagonale représentant les degrés des sommets du graphe ; D est une matrice diagonale dont le i^e élément diagonal est $d_i = \sum_j s_{ij}$. La matrice Laplacienne non normée associée au graphe est définie par $L = D - S$. Par la suite, deux normalisations peuvent être adoptées :

$$\begin{aligned} L_{sym} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} S D^{-1/2}, \\ L_{RW} &= D^{-1} L = I - D^{-1} S. \end{aligned}$$

La classification spectrale basée sur la matrice L_{RW} est celle se rapprochant le plus de la méthode que nous proposons ici. En effet, la matrice $P = D^{-1}S$ qui apparaît dans l'expression de L_{RW} définit une matrice stochastique similaire à celle que nous avons préconisé de calculer à partir de la matrice de ressemblance correspondant à la matrice de passage d'une marche aléatoire sur un graphe. Par la suite, certaines démarches de classification spectrale consistent à extraire les vecteurs propres de la matrice L_{RW} (ou L_{sym}) et effectuer un algorithme de classification sur la base de ces vecteurs propres. Des justifications de cette démarche peuvent être trouvées dans Shi et Malik (2000), Ng et al. (2001) et von Luxburg (2007).

Il est clair que les principes de base de la classification spectrale et la classification que nous proposons ici sont proches. En effet, dans les deux cas, il s'agit d'assimiler les points aux sommets d'un graphe de similarité et, par la suite, de déterminer les composantes connexes de ce graphe. Cependant, dans notre démarche l'effort est porté sur la création d'une matrice d'affinité P telle que les composantes connexes (classes finales associées à la matrice stochastique) soient déterminées par simple réarrangement des objets sur lesquels porte la classification. Cela passe par un choix pertinent des mesures de ressemblance et par la recherche des individus isolés. A contrario, pour la classification spectrale, l'effort est davantage porté sur le choix du nombre de vecteurs propres de la matrice Laplacienne à retenir et sur l'algorithme de classification effectué sur ces vecteurs propres. La complexité de calcul des vecteurs propres étant en $O(N^2)$ par vecteur propre, ces algorithmes deviennent très coûteux pour de grands jeux de données. Dans la classification en référence à une matrice stochastique discutée dans ce papier, il est également préconisé de déterminer des vecteurs propres. Cependant, il faut souligner qu'il s'agit de calculer les vecteurs dominants des matrices stochastiques associées aux différentes classes qui sont, par conséquent, de dimensions plus petites que dans le cas de la classification spectrale. De plus, ce calcul est optionnel et ne s'impose que si l'utilisateur est intéressé par la détermination des degrés de centralité qui confèrent aux prototypes des classes un caractère robuste, présentant, en effet, un autre avantage par rapport à la classification spectrale.

5 Applications

La méthode est illustrée sur la base de données simulées. Dans le premier exemple, il s'agit de classes sphériques dans un espace à 10 dimensions, tandis que dans le deuxième exemple, deux classes formant des cercles concentriques ont été simulées.

5.1 Premier exemple

Le premier tableau de données a été traité par Yan et Ye (2007) dans le contexte du choix du nombre de classes d'une partition. Il contient 4 classes d'individus dans un espace de 10 dimensions. Les moyennes de chaque classe sont tirées suivant une loi multinormale centrée et de matrice de covariance $3.6I_{10}$, où I_{10} représente la matrice identité de taille 10. Par la suite, les individus sont générés pour chaque classe suivant une loi multinormale de matrice de covariance I_{10} centrée autour de ces moyennes. L'effectif de chaque classe est tiré aléatoirement entre 25 et 50 individus. Les résultats de la méthode sont comparés à ceux obtenus par la méthode de classification par partitionnement autour de centres mobiles (Lebart et al., 2006).

Dans le cadre de la classification autour de centres mobiles, une méthode fréquemment utilisée pour le choix du nombre de classes est l'étude de l'évolution de l'inertie intra-classe en fonction du nombre de classes. Il s'agit d'identifier un coude dans la courbe qui dépeint cette évolution. Pour l'exemple considéré ici, ceci conduit au choix de 4 classes (données non présentées).

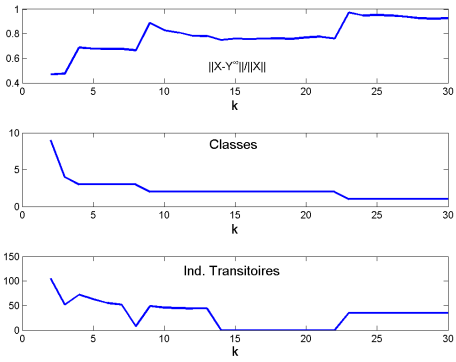
La méthode de classification présentée dans ce papier est appliquée en utilisant la ressemblance du voisinage. Cette mesure présente l'avantage de ne pas nécessiter la suppression d'individus isolés si les données ne sont pas très bruitées du fait du seuillage directement effectué sur la mesure. Les deux paramètres à déterminer sont le nombre de plus proches voisins k constituant un voisinage et le seuil de la ressemblance, p_0. Pour différentes valeurs de p_0 (de 0 à 0,3), des indicateurs ont été calculés en fonction de k : la mesure d'homogénéité $\frac{||Y^{(\infty)}-X||}{||X||}$ (la division par $||X||$ permet d'avoir des échelles similaires pour différentes données), le nombre de classes finales et le nombre d'éléments transitoires. Les courbes correspondantes sont représentées dans la Figure 1. La première étape consiste à déterminer le paramètre p_0 à conserver. Il faut le choisir pour que les indicateurs soient stables et qu'ils ne laissent pas apparaître d'effets de chaînage. Par exemple, si nous analysons les courbes correspondant à $p_0 = 0$ (Fig. 1a), il y a plusieurs plages de stabilité qui apparaissent pour la norme et le nombre de classes : pour k aux alentours de 5, entre 10 et 20, et supérieur à 25. Toutefois, le nombre d'individus transitoires correspondant laisse entrevoir des problèmes de chaînages. Les partitions obtenues comportent, en effet, beaucoup d'individus transitoires, qui parfois disparaissent (pour $k = 15$ par exemple), ce qui démontre la présence de problèmes de chaînage importants. Pour $p_0 = 0.1$ (Fig. 1b), l'analyse est assez semblable ; ces deux valeurs de p_0 sont trop faibles et ne suffisent pas à gérer les problèmes de chaînage.

Pour p_0 valant 0,2 ou 0,3 (Fig. 1c et 1d), tous les indicateurs sont globalement stables et le nombre d'éléments transitoires est faible pour k supérieur à 5. La plage principale de stabilité se situe pour k compris entre 7 et 20. Le seuil $p_0 = 0,2$ a été conservé, mais nous aurions pu tout aussi bien choisir $p_0 = 0,3$. Le paramètre k a été choisi dans la zone de stabilité. Nous avons conservé $k = 12$ car c'est la plus petite valeur pour laquelle il n'y a plus d'éléments transitoires. Il est intéressant de noter que les individus qui sont transitoires avec un paramètre $k = 10$ ou 11 ont chacun un poids d'affectation élevé pour une classe spécifique et, lorsque nous passons à $k = 12$, ces individus sont affectés à ces classes et ils apparaissent avec des degrés de centralité faibles. La partition obtenue est représentée sur le premier plan factoriel de l'ACP dans la Figure 2. Cette partition est identique à celle obtenue par l'algorithme de partitionnement des centres mobiles. En effet, l'indice de Rand qui mesure le degré de recouvrement de deux partitions (Rand, 1971) est égal à 1 (valeur maximale).

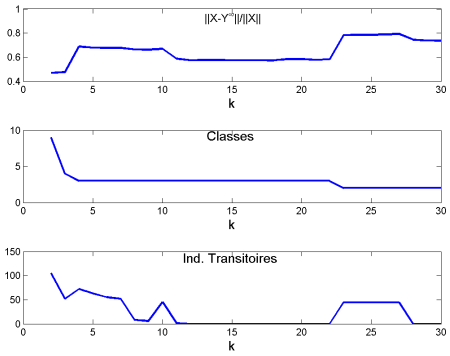
5.2 Deuxième exemple

La deuxième simulation consiste en un jeu de données en deux dimensions comportant deux cercles centrés sur l'origine. Le premier cercle a un rayon de 1 et le deuxième un rayon de 3. Un ensemble de 200 points est généré pour chaque cercle. Par la suite, les données sont bruitées en ajoutant 100 points aléatoirement (Fig. 3).

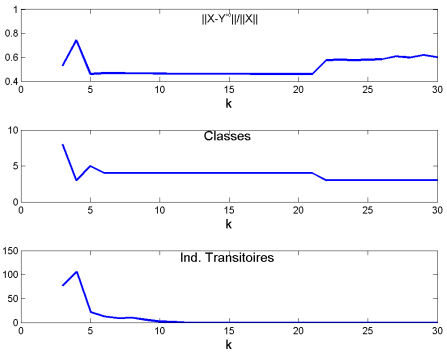
Nous avons considéré la mesure du voisinage sans imposer de seuil. Le paramètre p_0 est donc fixé 0. Les deux paramètres qui restent à fixer sont le nombre, k, de plus proches voisins constituant le voisinage d'un individu et le pourcentage d'individus devant être traités



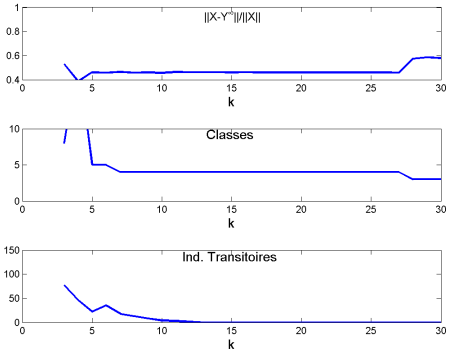
(a) Indicateurs avec un seuil p_0=0 sur la mesure du voisinage.



(b) Indicateurs avec un seuil p_0=0.1 sur la mesure du voisinage.



(c) Indicateurs avec un seuil p_0=0.2 sur la mesure du voisinage.



(d) Indicateurs avec un seuil p_0=0.3 sur la mesure du voisinage.

FIG. 1: Premier exemple : pour différentes valeurs du seuil p_0 de la mesure du voisinage, valeur de $\frac{||Y^{(\infty)}-X||}{||X||}$ (en haut), nombre de classes finales (au milieu) et nombre d'éléments transitoires (en bas) obtenus en fonction du paramètre k.

comme des individus isolés. Pour déterminer ces paramètres, les courbes des indicateurs (mesure d'homogénéité $\frac{||Y^{(\infty)}-X||}{||X||}$, nombre de classes finales et nombre d'éléments transitoires) en fonction de k sont représentées pour différentes proportions d'individus traités comme isolés (0%, 10%, 15% et 20%) dans la Figure 4. Le choix de la proportion adéquate est ici aisé. En effet, si tous les individus sont conservés (Fig. 4a), une seule classe apparaît qui absorbe environ 300 individus transitoires quand le paramètre k augmente. Si 10% ou 20% (Fig. 4b et 4d) des individus sont isolés, il y a beaucoup d'éléments transitoires et une grande instabilité dans la courbe de $\frac{||Y^{(\infty)}-X||}{||X||}$. À l'inverse, pour une proportion de 15% (Fig. 4c), il y a peu d'individus transitoires et une large plage de stabilité, s'étendant de $k = 11$ à $k = 24$. Le paramètre k doit être choisi dans cette plage de valeurs ($k = 12$, par exemple). Les classes obtenues sont représentées sur la Figure 5.

Classification en référence à une matrice stochastique

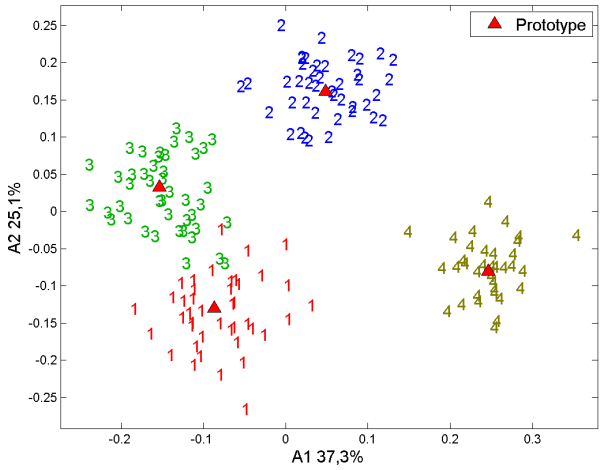
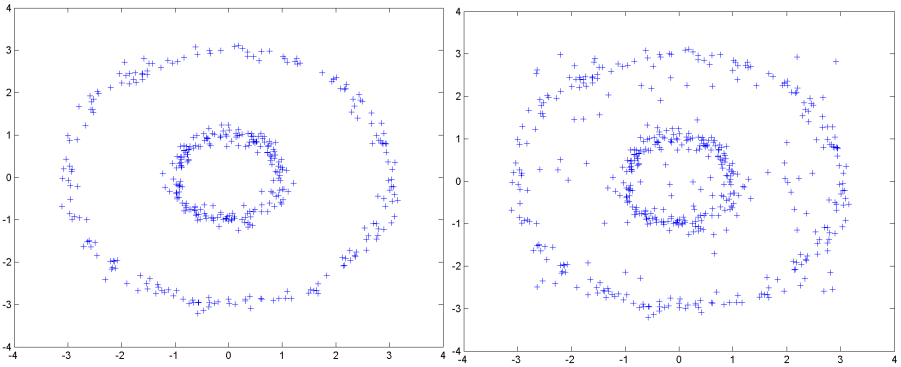


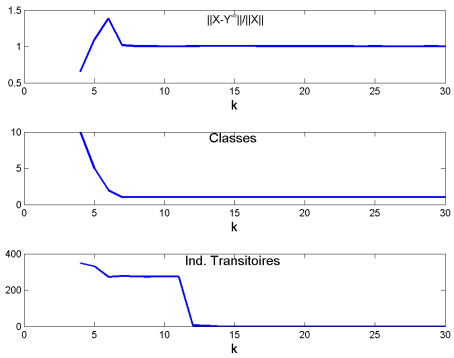
FIG. 2: Premier exemple : représentation des classes obtenues avec la mesure du Voisinage de paramètre $p_0 = 0.2$ et $k = 12$ sur le premier plan factoriel (62% de l'inertie restituée).



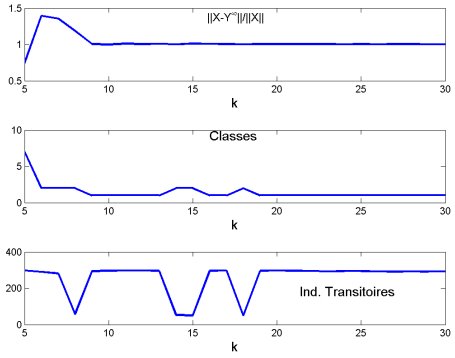
(a) Deux cercles de même centre. (b) Données bruitées.

FIG. 3: Deuxième exemple : représentation des données

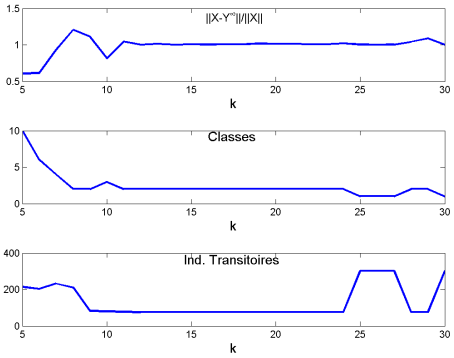
Les individus ont aussi été partitionnées en utilisant une méthode de classification spectrale. Cette méthode est basée sur la similarité Gaussienne qui, comme cela est indiqué dans le paragraphe 3.1, dépend d'un paramètre σ. Nous avons exploré toute la plage de variation de σ et identifié une zone qui permet d'obtenir la vraie partition originale. Cette plage, correspondant à l'intervalle $[0,25 \;\; 0,35]$, est relativement réduite et en dehors de cette plage l'algorithme conduit à des résultats peu satisfaisants (séparation linéaire des classes). Pour la valeur $\sigma = 0,30$, la partition obtenue se recoupe parfaitement avec celle que nous avons obtenue en utilisant la classification en référence à une matrice stochastique.



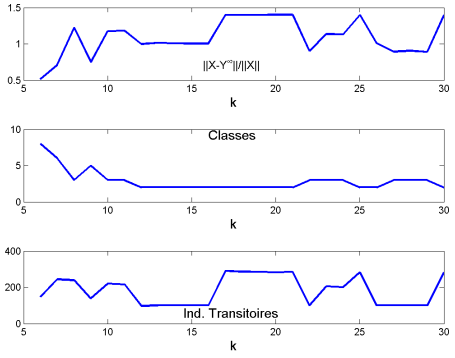
(a) Indicateurs avec un pourcentage de 0%.



(b) Indicateurs avec un pourcentage de 10%.



(c) Indicateurs avec un pourcentage de 15%.



(d) Indicateurs avec un pourcentage de 20%.

FIG. 4: Deuxième exemple : pour différents pourcentage d'individus traités comme isolés, valeur de $\frac{||Y^{(\infty)}-X||}{||X||}$ (en haut), nombre de classes finales (au milieu) et nombre d'éléments transitoires (en bas) obtenus en fonction du paramètre k.

5.3 Quelle mesure de ressemblance choisir ?

Dans les deux exemples présentés, nous avons utilisé la mesure du voisinage. Par expérience, nous nous sommes aperçus que les deux ressemblances non symétriques (k plus proches voisins et voisinage) sont les moins vulnérables aux problèmes de chaînage. De plus, la mesure du voisinage permet d'obtenir des degrés de centralité qui représentent bien la structure des classes. Cette propriété a été également décrite par von Luxburg (2007). L'auteur recommande d'utiliser les k plus proches voisins comme première mesure de ressemblance, du fait de sa robustesse vis à vis du choix des paramètres. De plus, elle apparait comme particulièrement adaptée si les classes ont des densités différentes.

La similarité de la boule et la similarité Gaussienne (paragraphe 3.1) conduisent à des résultats qui se recoupent. Elles sont toutes les deux très sensibles aux problèmes de chaînage. La recherche d'individus isolés s'avère presque toujours nécessaire. Cependant, il faut noter que la similarité Gaussienne présente l'avantage par rapport à la similarité de la boule de conduire

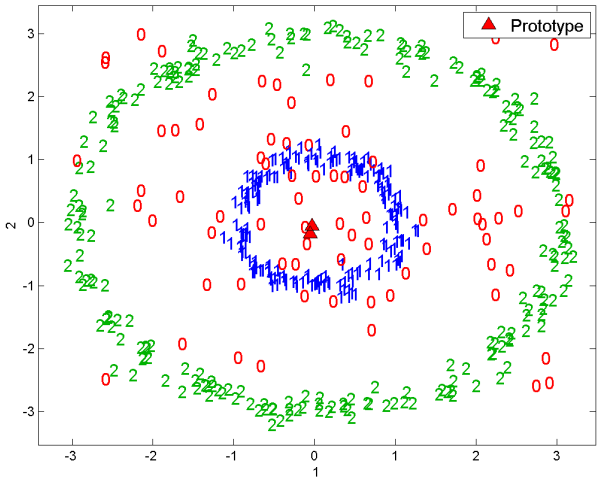


FIG. 5: Deuxième exemple : représentation des deux classes obtenues et des individus transitoires pour un paramètre $k = 12$ et 15% des individus traités comme isolés.

à des degrés de centralité plus fins du fait de la prise en compte d'une pondération et non d'une mesure binaire.

6 Conclusion

Une nouvelle méthode de classification reposant sur une matrice d'affinités entre les individus est présentée. Le fondement de la méthode repose sur la théorie des marches aléatoires sur un graphe. Les liens avec la classification spectrale ont été mis en évidence. Tout comme pour la classification spectrale, le cadre idéal est celui où les différentes classes sont bien séparées. Toutefois, des choix pertinents pour les valeurs des paramètres et une stratégie de recherche d'individus isolés permettent de limiter les effets de chaînage entre les classes.

Dans ce papier, nous avons considéré le cas d'une matrice d'affinité calculée à partir du tableau de données X mais la méthode s'applique à un cadre plus général où l'utilisateur dispose d'une mesure de ressemblance définie sur un ensemble d'objets. Ceci permet, par exemple, de prendre en compte des contraintes de contiguïté sur les objets, ou d'autres informations telles que l'appartenance des objets à des groupes définis a priori, ouvrant ainsi la voie à des problématiques de discrimination. Parmi les différentes perspectives à ce travail, deux nous paraissent essentielles. Il s'agit d'étudier les propriétés des mesures de ressemblance proposée et d'explorer davantage des stratégies de détermination des paramètres liés aux mesures de ressemblance et à la recherche de points isolés.

Références

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc. New York, NY, USA.

Lebart, L., M. Piron, et A. Morineau (2006). *Statistique exploratoire multidimensionnelle : Visualisation et inférence en fouille de données*. Dunod.

Lovasz, L. (1993). *Random walks on graphs : A survey*, Volume 2 of *Combinatorics, Paul Erdos is Eighty*. Budapest : Bolyai Society mathematical studies.

Meila, M. et J. Shi (2000). Learning segmentation by random walks. *Advances in Neural Information Processing Systems*, 873–879.

Ng, A., M. Jordan, et Y. Weiss (2001). On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pp. 849–856. MIT Press.

Norris, J. R. (1998). *Markov chains*. Cambridge : Cambridge Univ Pr.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Shi, J. et J. Malik (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22*(8), 888–905.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing 1*, 146.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*(4), 395–416.

Wishart, D. (1969). Mode analysis : A generalization of nearest neighbor which reduces chaining effects. In A. J. Cole (Ed.), *Numerical Taxonomy,*, pp. 282–311. Academic Press, New York.

Yan, M. et K. Ye (2007). Determining the number of clusters using the weighted gap statistic. *Biometrics 63*(4), 1031–1037.

Summary

We consider a data table measured on a set of individuals, and a stochastic matrix. We propose a method for partitioning the individuals by identifying the different stationnary states associated with the stochastic matrix. A partition of the individuals is set up from the stationnary points. In practice, the stochastic matrix can be derived from a similarity matrix that can be determined from the dataset at hand. Different similarities and density functions are studied and compared (nearest neighbors, kernels,...). The general approach of analysis is illustrated using simulated data. It is also compared to other clustering analysis approaches, including spectral clustering.