Classification en référence à une matrice stochastique

Stéphane Verdun, Véronique Cariou, El Mostafa Qannari

Classification en référence à une matrice stochastique

Verdun S., Cariou V. et Qannari E.M.

ENITIAA / INRA, Unité de Sensométrie et de Chimiométrie
Rue de la Géraudière, BP 82225, 44322 Nantes cedex 3, France

Résumé

Étant donné un tableau de données X portant sur un ensemble de n individus, et une matrice stochastique S qui peut être assimilée à une matrice de transition d'une chaîne de Markov, nous proposons une méthode de partitionnement consistant à appliquer la matrice S sur X de manière itérative jusqu'à convergence. Les classes formant la partition sont déterminées à partir des états stationnaires de la matrice stochastique. Cette matrice stochastique peut être issue d'une matrice de similarité entre les objets ; similarité qui peut être déterminée à partir du tableau X ou bien à partir de données externes. La matrice stochastique peut également refléter la densité de points autour des objets considérés. Différentes similarités et fonctions de densité sont étudiées et comparées (plus proches voisins, noyaux de densité...). La démarche sera illustrée sur la base de données simulées et de données réelles.

Abstract

We consider a data table X measured on a set of n individuals, and a stochastic matrix S wich can be assimilated to a transition matrix of a Markov chain. We propose a method for partitioning the individuals by iteratively applying S on X until convergence. A partition of the individuals is set up from the stationary points. In practice, the stochastic matrix can be derived from a similarity matrix that can be determined from the table X or from external data. The stochastic matrix can also reflect the density around the objects under consideration. Different similarities and density functions are studied and compared (nearest neighbors, kernels,...). The general approach of analysis is illustrated using simulated and real data.

Mots-clés : Analyse des Données - Data Mining, Classification.

Introduction

Nous proposons une méthode de partitionnement qui se caractérise par sa flexibilité. En effet, cette méthode repose essentiellement sur une matrice S qui peut refléter la structure (similarités, densités, ...) des objets à classer ou encore un lien avec des données externes (graphes, contraintes, variables à expliquer,...). De manière concrète, la méthode consiste à appliquer de façon itérative la matrice stochastique S sur le tableau de données jusqu'à convergence. Par la suite, les classes formant la partition sont déterminées à partir des points stationnaires associés à la matrice S.

Données et méthodes

Soit X un tableau de données décrivant la mesure de p variables sur n individus. Dans la suite, X est supposé centré. Soit S une matrice stochastique de dimension (n, n) (S vérifie les conditions suivantes : $S_{ij} \geq 0, \forall i, j = 1, \cdots, n$ et $\sum_{j=1}^{n} S_{ij} = 1, \forall i = 1, \cdots, n$). Le tableau $Y = SX$ est alors assimilé à un tableau de prototypes associés aux individus. En effet, la i^e ligne peut être considérée comme étant le barycentre de tous les points obtenu avec les pondérations $S_{ij}, (j = 1, \cdots, n)$.

Par exemple, si nous disposons d'une variable qualitative z ayant q modalités, et si nous définissons S_{ij} par :

$$S_{ij} = \begin{cases} \frac{1}{n_k} & \text{si i et j ont pris la } k^e \text{ modalité de z dont la fréquence est } n_k\,; \\ 0 & \text{si i et j ont pris des modalités différentes.} \end{cases}$$

Dans ce cas, il est facile de vérifier que $Y = SX$ est tout simplement la matrice qui associe à chaque individu le barycentre correspondant à la modalité prise par cet individu.

Un autre exemple consiste à considérer la similarité suivante entre les individus :

$$S^*_{ij} = \begin{cases} 1 & \text{si j est parmi les k plus proches voisins de i}\,; \\ 0 & \text{sinon.} \end{cases}$$

Par la suite, S^* peut être normalisée de manière à avoir la somme des lignes égale à 1, conduisant ainsi à une matrice stochastique S. Dans ce cas, les prototypes consistent en des barycentres locaux tenant compte des voisinages de chaque individu.

De manière générale, nous pouvons montrer que :

$$E(X) = E(Y),$$
$$||Y|| \leq ||X||.$$

Ces propriétés étendent les propriétés connues de l'espérance conditionnelle.

La méthode de partitionnement consiste à procéder de manière itérative en appliquant à chaque fois la matrice S : $Y = S \cdot X, Y^{(2)} = S \cdot Y = S^2 \cdot X, \cdots, Y^{(n)} = S^n \cdot X$. Ce processus converge vers une matrice $Y^{(\infty)}$ qui comporte $k \leq n$ lignes distinctes, k étant l'ordre de multiplicité de la valeur propre égale à 1 de la matrice S. De fait, chaque individu est associé à une classe représentée par un prototype (qui est une ligne de $Y^{(\infty)}$). La qualité de la partition peut être évaluée à l'aide de l'indice $C(k) = \frac{||Y^{(\infty)}||}{||X||}$. Cet indice s'apparente à un rapport d'inertie interclasses sur l'inertie totale.

Les situations intéressantes que nous avons identifiées concernent le cas où la matrice stochastique est associée à une matrice de similarité entre individus, ou une matrice

reflétant la densité de points autour des individus. Par exemple, nous pouvons considérer :

$$S^*_\sigma(i,j) = e^{-\frac{1}{2*\sigma^2}||x_i - x_j||^2}$$

Afin d'éviter des chaînages qui pourraient conduire à la formation d'un seul point stationnaire, et par conséquent d'une seule classe, il convient d'effectuer un seuillage ramenant à 0 toutes les similarités en deçà d'un seuil fixé. La matrice S_σ est obtenue par une normalisation en ligne de S^*_σ. Lorsque σ augmente, le nombre de classes k tend à diminuer, ainsi que le rapport $C(k)$. Il est possible à partir de là de déterminer le nombre de classes en fonction de l'évolution de $C(k)$. D'autres choix de la matrice stochastique, basés notamment sur les voisins réciproques, seront également discutés et comparés.

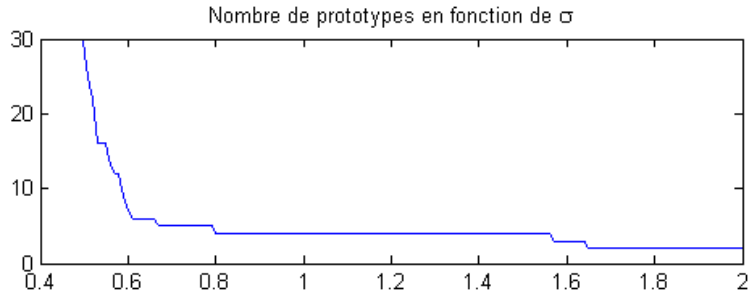
Application

Nous allons présenter ici une application de cette méthode de partitionnement sur un jeu de données simulées. Celui-ci a été proposé par Yan et Ye (2007) dans le contexte du choix du nombre de classes. Les données ont été simulées à partir d'une loi multinormale en faisant varier les paramètres de manière à définir quatre groupes a-priori. Les valeurs des variables sont distribuées suivant une loi normale $\mathcal{N}(\mu_k, I_{10})$. Les moyennes μ_k de chacune des classes sont générées aléatoirement à partir d'une loi $\mathcal{N}(0_{10}, 3.6I_{10})$, où I_{10} est la matrice identité de taille $(10, 10)$. Enfin, l'effectif de chaque classe est tiré aléatoirement entre 25 et 50 individus. Il apparaît après avoir effectué une ACP que les deux premiers axes factoriels représentent 78% de l'inertie totale du nuage. Une représentation des données sur ces deux axes donnera donc une configuration très peu déformée du nuage original dans l'espace total.

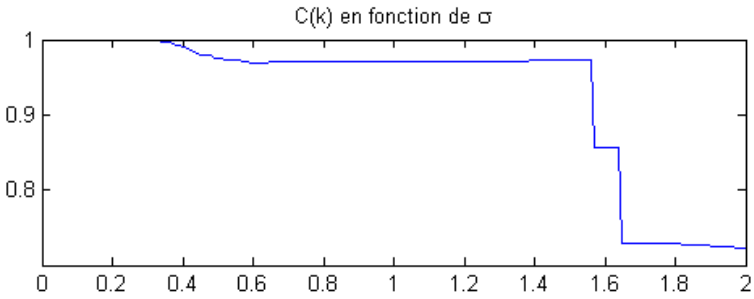
Nous avons considéré la similarité suivante :

$$S_{ij} = \begin{cases} e^{-\frac{1}{2*\sigma^2}||x_i - x_j||^2} \text{ si } x_j \in \mathcal{B}(x_i, 1.96\sigma) \\ 0 \text{ sinon} \end{cases}$$

Où $\mathcal{B}(x_i, r)$ est la boule fermée de centre x_i et de rayon r. Comme cela est indiqué ci-dessus, ce choix permet de limiter les effets de chaînage en tranchant les queues de la distribution. Dans les graphiques de la figure 1, on peut voir l'évolution du nombre de classes k (fig. 1a) et du rapport $C(k)$ (fig. 1b) en fonction de la valeur de σ. Sur le graphique de la fig. 1b, la décroissance brutale de $C(k)$ au point $\sigma = 1.56$ indique qu'il serait approprié de choisir une partition en 4 classes. Celle-ci est visualisée sur le premier plan factoriel de l'ACP de X (fig. 2). La partition obtenue identifie bien les quatre classes définies a priori.



(a) Évolution du nombre de classes.



(b) Évolution de $C(k)$.

FIGURE 1 – Evolution du nombre de classes k (fig.1a) et du critère $C(k)$ (fig.1b) en fonction du paramètre σ.

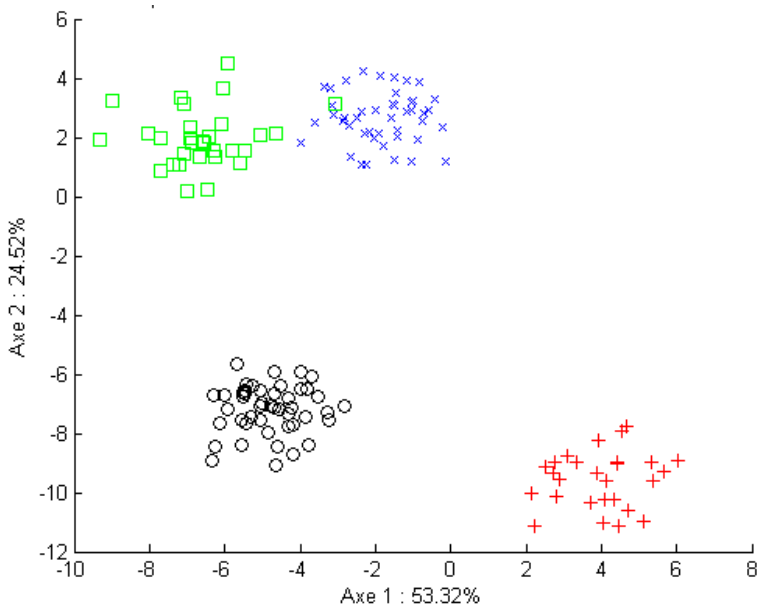


FIGURE 2 – Représentation de la partition obtenue sur le premier plan factoriel, avec $\sigma = 1.56$.

Bibliographie

[1] Yan, M. et Ye, K. (2007) Determining the Number of Clusters Using the Weighted Gap Statistic. *Biometrics*, 63, 1031–1037.