Classification de distributions par décomposition de mélange de copules archimédiennes : choix de la dimension des copules par visualisation

Cuvelier, Etienne; Fraiture, Monique Noirhomme

Published in:
4e Atelier Visualisation et extraction de connaissances, EGC'2006, Lilles, 2006

Publication date:
2006

Document Version
Première version, également connu sous le nom de pré-print

Classification de distributions par décomposition de mélange de copules archimédiennes: choix de la dimension des copules par visualisation

Etienne Cuvelier*, Monique Noirhomme-Fraiture*

*Institut d'Informatique, Facultés Universitaires Notre-Dame de la Paix,
21 rue Grandgagnage, 5000 Namur, Belgique
etienne.cuvelier@info.fundp.ac.be, monique.noirhomme@info.fundp.ac.be

Résumé. En analyse symbolique, un objet complexe peut être décrit par une variable s'exprimant comme une distribution de probabilité. La classification d'un ensemble d'objets symboliques décrits par ce type de variable, peut être obtenue en appliquant une décomposition de mélange de copules archimédiennes sur les valeurs des distributions calculées en un nombre q de points distincts, appelés coupures. Jusqu'à présent ces coupures ont été choisies arbitrairement.
Dans cet article nous montrons d'abord de façon empirique sur quelques exemples que le taux d'erreur de classement varie avec le nombre de coupures et leur position. Nous proposons ensuite de fixer ces deux paramètres grâce à une interaction visuelle.

1 Introduction

En analyse symbolique (Bock et Diday, 2000) une variable peut, entre autre, être décrite par une distribution de probabilité continue. Dans ce cas, la classification en K groupes de N objets symboliques décrits par cette variable, peut être obtenue en appliquant une décomposition de mélange aux valeurs obtenues par l'échantillonnage des distributions calculées en q endroits distincts, appelés coupures. L'estimation des composantes du mélange est réalisée à l'aide de copules archimédiennes. Cette appproche a déjà été utilisée avec succès par (Vrac et al., 2001) sur des données atmosphériques avec deux coupures. Nous étendons cette approche en permettant d'utiliser un nombre plus grand de coupures (Cuvelier et Noirhomme-Fraiture, 2005).
Dans tout les cas, le choix des coupures se révèle déterminant pour la qualité de la classification, mais actuellement aucun critère de décision automatique n'existe pour effectuer ce choix. Nous proposons donc, à l'instar de (Poulet, 2003) et de (Guo, 2003), une coopération entre une technique automatique de classification et technique visuelle de choix des coupures. Ce choix des coupures par visualisation étant basé sur des heuristiques inspirées des résultats de simulations.
Dans le paragraphe 2 nous rappellerons brièvement l'algorithme des nuées dynamiques appliqué aux distributions de probabilité, et nous introduirons la notion de copule. Le paragraphe 3 traitera de l'influence du nombre et du choix des coupures dans la qualité de la classification. Nous détaillerons ensuite dans le paragraphe 4 l'intérêt du choix visuel des coupures.

2 Classification de distributions de probabilités

2.1 Distributions de distributions

Nous supposerons que nous avons comme base de travail un tableau T de n lignes et p colonnes, et que la j^{me} colonne contient des distributions de probabilités , c'est-à-dire que si nous notons Y^j la j^{eme} variable alors Y_i^j est une distribution $F_i(.)$ pour tout $i \in \{1, ..., n\}$. Dans ce qui suit nous noterons ω_i le concept décrit par l'objet symbolique de la i^{eme} ligne, et $F_{\omega_i}(.)$ la distribution associée. Pour effectuer la classification en K classes nous commençons par échantilloner les distributions en q valeurs $T_1, ..., T_q$, et donc pour chaque $i \in \{1, ..., n\}$ nous calculerons $F_i(T_1), ..., F_i(T_q)$. Si nous appelons Ω l'ensemble de tous les concepts, la distribution conjointe des valeurs $F_i(T_j)$ est définie par :

$$H_{T_1,...,T_q}(x_1, ..., x_q) = P\left(\omega \in \Omega : \{F_\omega(T_1) \leq x_1\} \cap ... \cap \{F_\omega(T_q) \leq x_q\}\right) \tag{1}$$

et est appelée distribution de distributions.
La méthode, classique, de décomposition de mélange consiste à considérer cette distribution comme étant la résultante d'un mélange de distributions :

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \tag{2}$$

avec $\forall\, i \in \{1, ..., K\} : 0 < p_i < 1$ et $\sum_{i=1}^{K} p_i = 1$.
La distribution de la i$^{\text{eme}}$ classe étant donnée par $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$, avec $\beta_i \in R^d$, et p_i étant la probabilité qu'un élément appartienne à cette classe.
La densité de chaque distribution est donnée par :

$$h(x_1, ..., x_q) = \frac{\partial^q}{\partial x_1 ... \partial x_q} H(x_1, ..., x_q) \tag{3}$$

2.2 Algorithme des nuées dynamiques

L'algorithme utilisé (Diday, 2002) est en fait une extension de la méthode des nuées dynamiques (Diday et al., 1974) dans le cas d'un mélange. L'idée principale est, alternativement, d'estimer au mieux la distribution de chaque classe, et ensuite de vérifier que chaque objet symbolique appartient à la classe de densité maximale. L'étape d'estimation est réalisée en maximisant un critère de qualité, ici la log-vraisemblance :

$$lvc(P, \beta) = \sum_{i}^{K} \sum_{\omega \in P_i} log(h(w)) \tag{4}$$

avec

$$h(w) = h_{T_1,...,T_q}(F_\omega(T_1), ..., F_\omega(T_q)) \tag{5}$$

La classification commence avec une partition initiale aléatoire, et les deux étapes suivantes sont donc répétées jusqu'à stabilisation de la partition :

- **Etape 1 : Estimation des paramètres**
  Déterminer le vecteur $(\beta_1, ..., \beta_K)$ qui maximise le critère de qualité.

– **Etape 2 : Distribution des objets symboliques dans les classes**
Les classes $(P_i)_{i=1,...,K}$, dont les paramètres ont été calculés à l'étape 1, sont construites comme suit

$$P_i = \{\omega : h(\omega, \beta_i) \geq h(\omega, \beta_m) \forall m\} \tag{6}$$

2.3 Copules

L'estimation de distributions multivariées et de leurs densités n'est pas toujours chose aisée, alors que l'estimation univariée pose moins de problèmes.
Ainsi dans notre cas, les marges des distributions $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$ définies par

$$G^i_{T_j}(x) = P\{\omega \in P_i : F_\omega(T_j) \leq x\} \tag{7}$$

peuvent être facilement estimées par

$$\widehat{G}^i_{T_j}(x) = \frac{card\{\omega \in P_i : F_\omega(T_j) \leq x\}}{card(P_i)} \tag{8}$$

et les densités associées par la méthode des noyaux (Silverman, 1986)

$$\widehat{g}_{T_i}(x) = \frac{1}{card(P_i) \cdot h} \sum_{\omega \in P_i} K\left(\frac{x - F_\omega(T_i)}{h}\right) \tag{9}$$

où K est une fonction noyaux (Gaussienne, Epanechnikov, Triangulaire,...) et h est le paramètre de lissage, qui peut être calculé à l'aide de l'Erreur Quadratique Moyenne Intégrée (Mean Integrated Square Error - MISE).
La notion de copule (Nelsen, 1999) permet d'utiliser ces estimations des marges pour reconstruire les distributions multivariées.
Par définition

$$C(u_1, ..., u_n) \textit{ est une copule}$$
$$\textit{ssi}$$
$$C \textit{ est une distribution multivariée dont toutes les marginales sont uniformes sur } [0, 1]$$

Les copules sont des outils précieux dans la modélisation des structures de dépendance grâce au théorème de Sklar :
Si $H(x_1, , x_n)$ est une distribution multivariée de marginales $F_1(x_1), ..., F_n(x_n)$
alors il existe une copule C telle que

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{10}$$

de plus, si $F_1, ..., F_n$ sont toutes continues, alors C est unique ; sinon C est unique seulement sur $domF_1 \times ... \times domF_n$ Les copules capturent la structure de dépendance et permettent de séparer la modélisation de celle-ci de la modélisation des marges.
Plusieurs techniques de construction des copules existent (Joe, 1997; Nelsen, 1999), et parmi celles-ci une méthode génère une classe importante de copules : les copules archimédiennes.

2.4 Copules Archimédiennes

Les copules archimédiennes sont définies par

$$C(\underline{u}) = C(u_1, ..., u_n) = \phi^{-1}\left(\sum_{i=1}^{n} \phi(u_i)\right) \tag{11}$$

où ϕ est une fonction, appelée générateur, de $[0, 1]$ vers $[0, \infty]$ telle que
- ϕ est une fonction continue strictement décroissante
- $\phi(0) = \infty$ et $\phi(1) = 0$
- ϕ^{-1} est complétement monotonique sur $[0, \infty[$ c-à-d que

$$(-1)^k \frac{d^k}{dt^k}\phi^{-1}(t) \geq 0 \tag{12}$$

quel que soit $t \in [0, \infty[$ et pour tout k.

Il existe quatre familles de copules archimédiennes bivariées qui possèdent une extension intéressante en dimension quelconque. Il s'agit des copules de Clayton, Gumbel, Frank et Joe. Dans le cadre de cet article nous n'utiliserons que la première : la copule de Clayton :

$$C_\theta(\underline{u}) = \left(\sum_{i=1}^{n}(u_i^{-\theta}) - n + 1\right)^{-\frac{1}{\theta}} \tag{13}$$

Une des propriétés des copules archimédiennes est que pour k fixé (avec $2 \leq k \leq n$) toutes les marges de dimension k d'une copule sont identiques. Ainsi les marges bidimensionnelles, obtenues à partir de l'expression (11) de la façon suivante,

$$\phi^{-1}\left(\phi(u_i) + \phi(u_j) + \sum_{k \neq i,j} \phi(1)\right) = \phi^{-1}\left(\phi(u_i) + \phi(u_j)\right) = C(u_i, u_j) \tag{14}$$

sont toutes modèlisées de la même façon, c'est-à-dire avec le même générateur ϕ.

3 Influence du nombre et du choix des coupures

Pour illustrer notre propos nous utiliserons un ensemble artificiel de données (cf. Figure 1). Cet ensemble est constitué de 4 groupes de distributions. En parcourant le graphe de gauche à droite, nous trouvons 45 distributions exponentielles (exp1), 45 distributions normales (norm1), 45 distributions exponentielles (exp2), et enfin 50 distributions beta (beta1).

Pour chaque distribution, 500 nombres aléatoires ont été générés suivant la loi choisie, ensuite on a estimé la distribution empirique.

Nous avons ensuite généré tous les ensembles de 2 à 7 coupures équidistantes (d'un multiple de 0.25) et ayant comme première coupure au moins -4, et comme dernière coupure au plus 4 :

$$\{\{T_k = d + k.s : k \in \{0, ..., q-1\}\} : q \in \{2, ..., 7\}; -4 \leq T_0; T_{q-1} \leq 4; s = k * 0.25\} \tag{15}$$

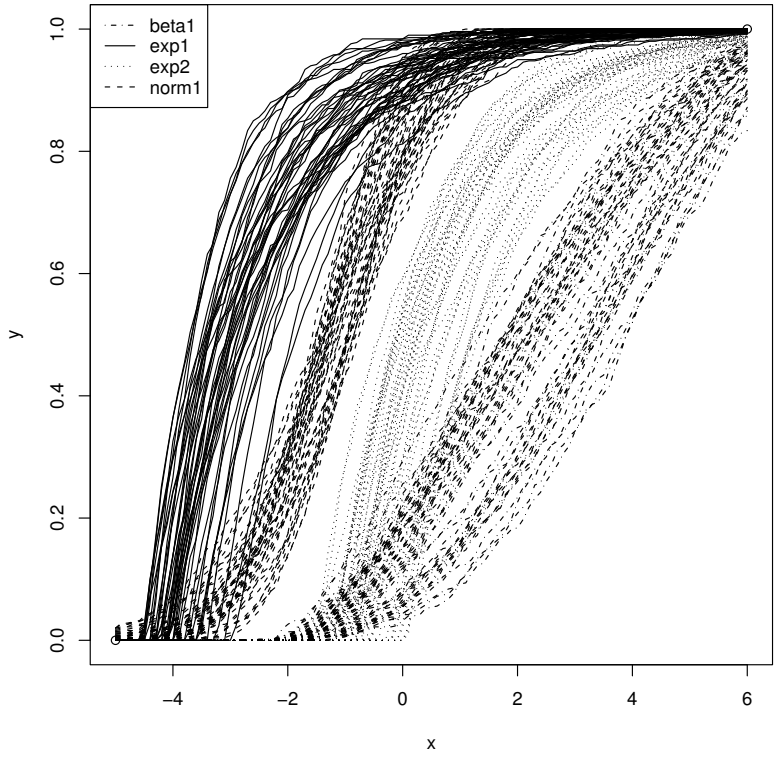

FIG. 1 – *Exemple d'ensemble de donnée*

Ensuite nous avons effectué les classifications sur base de chacun de ces ensembles de 2 à 7 coupures (820 ensembles). Ces classifications ont été réalisées en utilisant toujours la même partition initiale, et ce afin de pouvoir comparer les résultats des classifications.
Sur ces 820 ensembles de coupures, seuls 9 ensembles permettent de classer sans erreur les distributions. Ces ensembles sont les suivants : $\{-1.50, 0.50\}$, $\{-1.50, 1.00\}$, $\{-1.50, 1.25\}$, $\{-1.50, 1.50\}$, $\{-1.75, 1.50\}$, $\{-1.50, 2.00\}$, $\{-1.50, 2.25\}$, $\{-1.75, 2.25\}$, $\{-1.50, 2.50\}$.

Comme on peut le voir dans le tableau 1, si on choisit les coupures de manière arbitraire on peut obtenir des cas favorables et avoir un taux d'erreur proches de 0, mais aussi obtenir des cas très défavorables et avoir jusqu'à 51% d'erreur. Nous avons dès lors besoin d'heuristiques pour bien choisir les coupures.

Concentrons-nous maintenant sur les ensembles de coupures qui permettent d'effectuer la classification avec un taux d'erreur acceptable (10% maximum). En comparant les résultats par dimension des intervalles de la première et de la dernière coupure (Tableau 2) et le graphique nous pouvons observer les comportements suivants :

1. la longueur des intervalles des valeurs intéressantes pour les coupures varie de façon décroissante par rapport au nombre de coupures ;

q	Erreur moyenne	Erreur minimale	Erreur maximale
2	28,8%	0%	49,1%
3	33,1%	1,6%	51,3%
4	31,3%	4,3%	48,6%
5	28,5%	7%	48,6%
6	29%	5,4%	48,6%
7	28%	6,4%	49,1%

TAB. 1 – *Minima, maxima et moyennes par dimension*

q	Erreur moyenne	T_1	T_q
2	0.0334	[-2.50, -0.25]	[-1.75, 4.00]
3	0.0529	[-1.75, -0.75]	[0.25, 2.25]
4	0.0639	[-2.50, -0.50]	[0.25, 2.75]
5	0.0828	[-2.75, -1.75]	[-0.50, 1.25]
6	0.0567	[-2.75, -2.50]	[0.00, 2.25]

TAB. 2 – *Intervalles des premières et dernières coupures pour les meilleurs ensembles.*

2. l'augmentation du nombre de coupures n'améliore pas nécessairement le taux d'erreur ;

Le premier comportement est dû au fait que les copules archimédiennes modélisent de la même façon toutes les relations entres les marges (cf. supra). Cela permet de laisser "tomber" le début et la fin de l'ensemble des valeurs des distributions aux comportements différents de la partie centrale.
Le second comportement veille à maximiser le nombre d'informations par coupure compte tenu du nombre de coupures.

4 Intérêt du choix visuel

Les coupures ayant une influence directe sur la qualité des résultats, nous ne pouvons nous satisfaire d'un choix aléatoire. En nous inspirant des comportements ci-dessus nous pouvons émettre les heuristiques suivantes :

1. minimiser le nombre de coupures choisies ;
2. choisir des coupures qui maximisent le nombre de groupes discernables de valeurs le long de ces coupures, en veillant à ce que chaque groupe soit discerné un maximum de fois sur l'ensemble des coupures ;

Ainsi sur notre exemple, on peut visuellement repérer les zones suivantes dans l'espace possible des coupures :

- un faible intervalle situé autour de -0.5 où l'on peut presque distinguer 4 groupes distincts de valeurs ;
- un intervalle allant de -2.25 à 1.75 où l'on peut en chaque point distinguer au moins 3 groupes distincts de valeurs ;

– un intervalle allant de -3.25 à 4 où l'on peut en chaque point distinguer au moins deux distincts de valeurs ;

En suivant les heuristiques énoncées on peut par exemple choisir deux coupures, la première $-2 \leq T_1 \leq -1$ et la seconde $1 \leq T_2 \leq 1.75$. Dans nos tests les classifications qui ont utilisé des coupures respectant ces conditions ont les résultats suivants pour les taux d'erreurs : moyenne = 14,2%, min = 0% , max = 27%, ce qui représente une substantielle augmentation par rapport au choix aléatoire.

5 Conclusions et perspectives

Dans cet article nous avons montré comment le choix des coupures nécessaires à la classification pouvait se faire aisément visuellement. Nous avons aussi suggéré deux heuristiques qui doivent guider ce choix visuel. Dans le futur, plusieurs axes de recherche sont à développer,notamment :

– la déternimation du nombre optimal de coupures ;
– l'aide à la détermination visuelle du nombre de classes distincts le long d'une coupure.

Références

Bock, H. et E. Diday (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag.

Cuvelier, E. et M. Noirhomme-Fraiture (2005). Clayton copula and mixture decomposition. In *ASMDA 2005*, pp. 699–708.

Diday, E. (2002). Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pp. 297–310.

Diday, E., A. Schroeder, et Y. Ok (1974). The dynamic clusters method in pattern recognition. In *IFIP Congress*, pp. 691–697.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*(4), 232–246.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman and Hall.

Nelsen, R. (1999). *An introduction to copulas*. London: Springer.

Poulet, F. (2003). Interactive decision tree construction for interval and taxonomical data. In *Third international workshop on visual data mining - ICDM 2003*, pp. 183–194.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Vrac, M., E. Diday, A. Chédin, et P. Naveau (2001). Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pp. 348–355.

Classification de distributions : choix de la dimension par visualisation

Summary

In symbolic data analysis, a complex object can be provided in the form of a continuous distribution. The classification of a set of symbolic objects described by this type of variable, can be obtained by applying a mixture decomposition of archimedean copulas to the values of the distributions calculated in a number Q of distinct points, called cuts. Until now these cuts were arbitrarily selected. In this article we show initially in an empirical way on some examples that the error rate of classification varies with the number of cuts and their positions. We then propose to chose these two parameters thanks to a visual interaction.