Une nouvelle méthode divisive en classification non supervisée pour des données symboliques intervalles

Nathanaël Kasoro*, André Hardy**

*Université de Kinshasa
Département de Mathématique et d'Informatique
B.P. 190, Kinshasa, République Démocratique du Congo
kasoro.mulenda@yahoo.fr
**Université de Namur
Unité de Statistique - Département de Mathématique
8 Rempart de la Vierge - B - 5000 Namur - Belgique
andre.hardy@fundp.ac.be

Résumé. Dans cet article nous présentons une nouvelle méthode de classification non supervisée pour des données symboliques intervalles. Il s'agit de l'extension d'une méthode de classification non supervisée classique à des données intervalles. La méthode classique suppose que les points observés sont la réalisation d'un processus de Poisson homogène dans k domaines convexes disjoints de R^p. La première partie de la nouvelle méthode est une procédure monothétique divisive. La règle de coupure est basée sur une extension à des données intervalles du critère de classification des Hypervolumes. L'étape d'élagage utilise un test statistique basé sur le processus de Poisson homogène. Le résultat est un arbre de décision. La seconde partie de la méthode consiste en une étape de recollement, qui permet, dans certains cas, d'améliorer la classification obtenue à la fin de la première partie de l'algorithme. La méthode est évaluée sur un ensemble de données réelles.

1 Introduction

Le but de la classification non supervisée est de décomposer un groupe d'objets, sur lesquels on mesure un ensemble de variables, en un nombre relativement restreint de sous-groupes d'objets semblables. De nombreuses méthodes de classification ont été publiées dans la littérature scientifique. La plupart d'entre elles utilisent un critère de classification basé sur une mesure de dissimilarité. Pour éviter ce choix (bien souvent arbitraire) d'une dissimilarité nous utilisons un modèle statistique pour la classification basé sur le processus de Poisson homogène (Hardy (1983)). De ce modèle est issue la méthode de classification des Hypervolumes (Hardy (1983)). Pirçon (2004) a développé une nouvelle méthode divisive de classification basée sur le critère de classification des Hypervolumes. Notre objectif est d'étendre cette méthode à des données intervalles. Une variable Y dont le domaine d'observation est $\mathcal{Y}$ est appelée à valeurs d'ensemble si $\forall x_i \in E, \quad Y : E \to \mathcal{B} : x_i \longmapsto Y(x_i)$ où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\}$.

Une variable à valeurs d'ensemble Y est appelée une variable intervalle si $\mathcal{Y} = R$ et $\forall x_i \in E$, il existe $\alpha, \beta \in R$, tels que $Y(x_i) = [\alpha, \beta]$.

2 Un modèle statistique pour la classification basé sur le processus de Poisson homogène

2.1 Définition : le processus de Poisson homogène

N est un processus de Poisson homogène d'intensité q ($q \in R$) sur un ensemble $D \subset R^p$ $(0 < m(D) < \infty)$ si les deux conditions suivantes sont satisfaites (Cox et Isham (1980)) :

- $\forall\ A_1, \ldots, A_k \subset D, \quad \forall i \neq j \in \{1, \ldots, k\}$ où $A_i \cap A_j = \emptyset, \quad N(A_i) \perp\!\!\!\perp N(A_j)$.
  Les variables aléatoires qui comptent le nombre de points dans des régions disjointes de l'espace sont indépendantes.
- $\forall\ A \subset D, \quad \forall k > 0, \quad P(N(A) = k) = \dfrac{(q\, m(A))^k}{k!}\, e^{-q\, m(A)}$.
  La variable aléatoire $N(A)$ a une distribution de Poisson de moyenne $m(A)$ où $m(.)$ est la mesure de Lebesgue multidimensionnelle.

2.2 Le critère de classification des hypervolumes

La méthode de classification des Hypervolumes (Hardy et Rasson (1982), Hardy (1983)) suppose que les n observations p-dimensionnelles $x_1, \ldots, x_n$ représentent un échantillon aléatoire simple d'un processus de Poisson homogène N dans un ensemble D inclus dans l'espace Euclidien R^p (avec $0 < m(D) < \infty$). L'ensemble D est l'union de k domaines convexes compacts disjoints $D_1, \ldots, D_k$. Le problème statistique consiste à estimer les domaines inconnus D_i dans lesquels les points ont été générés. On désigne par $C_i \subset \{x_1, .., x_n\}$ l'ensemble des points appartenant à $D_i \ \ (1 \leq i \leq k)$. Les estimations du maximum de vraisemblance des k domaines inconnus $D_1, \ldots, D_k$ sont les k enveloppes convexes $H(C_i)$ des k sous-groupes de points C_i telles que la somme des mesures de Lebesgue des enveloppes convexes disjointes $H(C_i)$ est minimale. Le critère de classification des Hypervolumes est donc défini par $W_k = \sum_{i=1}^{k} m(H(C_i))$. Le problème de classification des Hypervolumes consiste donc à trouver la partition P^* telle que $P^* = arg\min_{P_k \in \mathcal{P}_k} \sum_{i=1}^{k} m(H(C_i))$ où $\mathcal{P}_k$ représente l'ensemble de toutes les partitions de C en k classes. Par exemple, dans le plan, la mesure de Lebesgue d'un domaine D est l'aire de ce domaine. Donc si on mesure sur chacun des n objets la valeur de deux variables quantitatives, la méthode de classification des Hypervolumes recherche les k groupes C_i, contenant tous les points, tels que la somme des aires des enveloppes convexes des ensembles C_i est minimale.

2.3 Un test statistique pour le nombre de classes : le Gap test

Grâce au modèle statistique pour la classification basé sur le processus de Poisson homogène, on peut définir un test du quotient de vraisemblance pour le nombre de classes (Kubushishi (1996)). On teste H_0 : les $n = n_1 + n_2$ points observés sont la réalisation d'un processus

de Poisson homogène dans D contre l'alternative H_1 : n_1 points sont la réalisation d'un processus de Poisson homogène dans D_1 et n_2 points dans D_2 où $D_1 \cap D_2 = \emptyset$. Les ensembles D, D_1, D_2 sont inconnus. La statistique du test est donnée par $Q(x) = \left(1 - \frac{m(\triangle)}{m(H(C))}\right)^n$ où $\triangle = H(C) \setminus (H(C_1) \cup H(C_2))$ est l'espace vide (*Gap space*) entre les classes et m la mesure de Lebesgue multidimensionnelle. La règle de décision est la suivante (Kubushishi (1996)) : on rejette H_0, au niveau α, si (distribution asymptotique)

$$\frac{nm(\triangle)}{m(H(C))} - \log n - (p-1)\log\log n \geq -\log(-\log(1-\alpha)).$$

3 La méthode de classification HOPP

HOPP (Pirçon (2004)) est une méthode de classification non supervisée divisive issue du modèle statistique décrit ci-dessus. La première étape consiste à couper successivement les noeuds de l'arbre en deux sous-noeuds, jusqu'à ce qu'un critère d'arrêt soit vérifié (le nombre de points dans un noeud). A chaque coupure on recherche la bipartition de la classe C en deux sous-classes C_1 et C_2, qui minimise le critère de classification des Hypervolumes $W_2 = m(H(C_1)) + m(H(C_2))$. La méthode est monothétique ; on choisit chaque fois le noeud et la variable tels que W_2 est minimal.

A la fin du processus de coupure, on obtient un arbre de grande taille. La deuxième étape permet d'élaguer l'arbre. Afin de vérifier si les coupures effectuées sont valides, on utilise le Gap test. On teste donc à chaque noeud les hypothèses suivantes : H_0 : les points sont distribués dans un seul domaine D contre l'hypothèse alternative H_1 : les points sont distribués dans deux domaines D_1 et D_2 ($D_1 \cap D_2 = \emptyset$). Lorsque l'hypothèse nulle n'est pas rejetée, on conclut que la coupure est mauvaise. Par contre si l'hypothèse nulle est rejettée, on décide que la coupure est bonne. A la fin du procédé on utilise la règle suivante : élaguer toutes les branches qui ne contiennent que des mauvaises coupures.

Dans certain cas la structure naturelle des données n'est pas obtenue à la fin de l'étape d'élagage. La troisième étape est un outil de recollement. Des tests sont effectués uniquement sur les classes qui ne sont pas issues du même noeud au niveau précédent. Pour ce faire nous utilisons à nouveau le Gap test. Si au moins un regroupement est effectué dans l'étape de recollement, HOPP perd son caractère hiérarchique. Elle devient alors une méthode de partitionnement.

4 La méthode de classification symbolique SPART

Dans ce paragraphe nous présentons l'extension de la méthode HOPP à des données intervalles (Bock et Diday (2000)). Pour ce faire, on représente chaque intervalle par son centre et sa demi-longueur, donc par un point dans un espace bidimensionnel.

Comme dans la méthode classique, la première étape consiste à trouver la meilleure bipartition d'une classe C en deux sous-classes C_1 et C_2. Comme SPART est une méthode monothétique, nous travaillons dans les p (m, ℓ) espaces, dans lesquels les intervalles deviennent des points. On considère donc toutes les bipartitions d'une classe C en deux classes $\{C_1, C_2\}$, en respectant l'ordre des centres des intervalles. Il s'agit donc des bipartitions générées par des droites verticales (figure 1). On définit une extension à des données intervalles de

la mesure de l'espace vide Δ entre les classes de la façon suivante : $m_E(\triangle) = (m_{i+1} - m_i) + (max(l_i, l_{i+1}) - min(l_i, l_{i+1}))$. On choisit l'intervalle $]m_i, m_{i+1}[$ tel que $m_E(\Delta)$ est maximal. Une valeur de coupure c est prise arbitrairement dans l'intervalle $]m_i, m_{i+1}[$. Habituellement on choisit le centre de l'intervalle (figure 1).

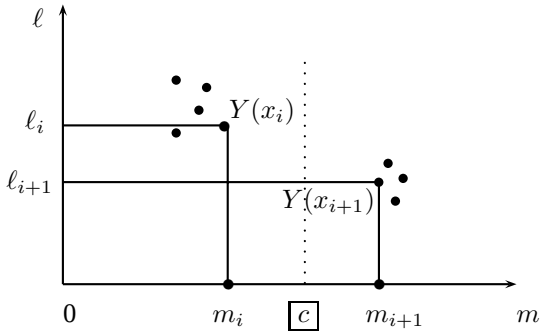

FIG. 1 - Bipartition d'une classe.

Si $x_\ell \in C$, $Y(x_\ell) = [\alpha_\ell, \beta_\ell]$ et $m_\ell = \frac{\alpha_\ell + \beta_\ell}{2}$. Une classe C est divisée en deux grâce à une question binaire de la forme "$m_\ell \leq c$?" où c est la valeur de coupure. On définit une fonction binaire $q_c : C \longrightarrow \{0, 1\}$ par $q_c(x_\ell) = 0$ si $m_\ell \leq c$ et 1 sinon. On obtient alors la bipartition souhaitée : $C_1 = \{x \in C : q_c(x) = 0\}$ et $C_2 = \{x \in C : q_c(x) = 1\}$.

Les étapes d'élagage et de recollement sont effectuées de la même façon que dans le cas classique. On utilise ici aussi une extension symbolique du Gap test à des données intervalles en représentant chaque intervalle par son centre et sa demi-longueur.

5 Application

On applique la méthode SPART à un jeu de données réelles. Nous comparerons les résultats donnés par SPART avec ceux obtenus par deux autres méthodes de classification non supervisées monothétiques divisives pour des variables intervalles : SCLASS (Rasson et al. (2007)) est une méthode de classification hiérarchique monothétique divisive basée sur une extension à des variables intervalles du critère de classification généralisé des Hypervolumes. DIV (Chavent (1998)) est quant à elle une méthode de classification hiérarchique monothétique divisive basée sur une extension du critère de l'inertie intra-classe.

Le jeu de données *"cars"* est constitué de 33 voitures disponibles en 2001, sur lesquelles ont été mesurées 8 variables intervalles. Il est répertorié dans les bases du logiciel SODAS 2 (SODAS2 (2004)). Les objets sont repris sur la figure 2. Les variables sont les suivantes : prix, empattement, cylindrée, longueur, vitesse maximale, largeur, accélération maximale, hauteur. Une partition en 4 classes est obtenue après l'étape d'élagage. L'étape de recollement ne modifie pas cette partition en 4 classes.

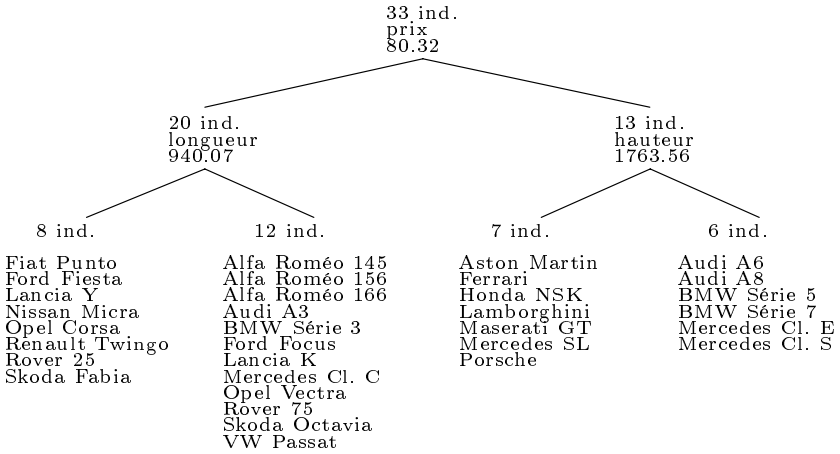

FIG. 2 - Arbre pour l'ensemble de données *"Cars"*

La première variable de coupure est le prix des voitures (cher - bon marché). Pour les voitures bon marché, la deuxième variable de coupure est la longueur de la voiture, tandis que pour les voitures chères, il s'agit de la hauteur de la voiture.

La figure 2 montre l'arbre hiérarchique produit par SPART. Les 4 classes peuvent être étiquettées de la manière suivante : classe 1 : voitures citadines, classe 2 : voitures berline, classe 3 : modèles sport et classe 4 : voitures limousines.

La méthode DIV donne la même partition en 4 classes. SCLASS produit une autre partition dont les classes ne semblent pas correspondre à une structure utile de l'ensemble des 33 voitures.

6 Conclusion

SPART est une nouvelle méthode de classification non supervisée pour des données intervalles. Elle est basée sur une extension aux données intervalles du critère de classification des Hypervolumes et du Gap test. L'originalité de l'approche est double. D'une part le modèle sous-jacent à la méthode n'utilise pas de mesure de dissimilarité ; le critère de classification est déduit d'un modèle statistique basé sur le processus de Poisson homogène. D'autre part la méthode inclut une étape de recollement qui lui permet, dans le cas de structures particulières, de retrouver les classes naturelles d'un ensemble de données multidimensionnelles.

SPART et DIV donnent souvent des résultats semblables. SPART produit cependant des résultats meilleurs que DIV lorsqu'on est en présence de classes allongées. Ceci s'explique principalement par le fait que DIV utilise une extension à des données intervalles du critère de la variance intra-classe, et que ce critère est biaisé par rapport aux classes de forme ellipsoïdale. SCLASS (Rasson et al. (2007)) utilise un modèle statistique pour la classification basé sur le processus de Poisson non homogène. Le critère à minimiser est l'intensité intégrée du

processus de Poisson sur les enveloppes convexes des classes. Cette méthode exige donc l'estimation de l'intensité du processus de Poisson non homogène. Elle est donc plus complexe d'un point de vue temps calcul. De plus dans sa version actuelle SCLASS ne comporte pas d'étape d'élagage ni d'étape de recollement. Les résultats obtenus par SCLASS sont donc généralement qualitativement moins bons que ceux produits par SPART. Enfin la procédure SPART détermine automatiquement le nombre de classes, qui doit être fixé au préalable dans SCLASS et DIV.

Références

Bock, H. et E. Diday (2000). *Analysis of Symbolic Data - Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin - Heidelberg: Springer-Verlag.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters 19*, 989–996.

Cox, D. et V. Isham (1980). *Point Processes*. London: Chapman and Hall.

Hardy, A. (1983). *Statistique et classification automatique : un modèle, un nouveau critère, des algorithmes, des applications*. Thèse de doctorat, FUNDP - Université de Namur.

Hardy, A. et J.-P. Rasson (1982). Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données 7 (2)*, 41–56.

Kubushishi, T. (1996). *On some Applications of Point Process Theory in Cluster Analysis and Pattern Recognition*. Thèse de doctorat, FUNDP - Université de Namur.

Pirçon, J.-Y. (2004). *La classification et les processus de Poisson pour de nouvelles méthodes monothétiques de partitionnement*. Thèse de doctorat, FUNDP - Université de Namur.

Rasson, J.-P., J.-Y. Pirçon, P. Lallemand, et S. Adans (2007). Unsupervised divisive classification. In E. Diday et M. Noirhomme (Eds.), *Symbolic Data Analysis and the Sodas 2 Software*. Wiley.

SODAS2 (2004). *Logiciel (http://www.info.fundp.ac.be/asso)*.

Summary

We present a new clustering method for symbolic interval data. It is an extension to interval data of a classical clustering method. The classical method assumes that the observed data points are a realisation of a homogeneous Poisson point process in k disjoint domains of R^p. The first part of the new method is a monothetic divisive procedure. The cut rule is based on an extension to interval data of the Hypervolumes clustering criterion. The pruning step uses a statistical hypothesis test based on the homogeneous Poisson process. The output is a decision tree. The second part of the method is a merging process, that allows in particular cases to improve the classification obtained at the end of the first part of the algorithm. The method is applied to a real data set.