**Mathématiques et sciences humaines**
Mathematics and social sciences

**187 | Automne 2009**
**Journée 2007 de la Société Francophone de Classification**

# Une nouvelle méthode de classification pour des données intervalles

*A new clustering method for interval data*

**André Hardy et Nathanael Kasoro**

**Édition électronique**
URL : http://journals.openedition.org/msh/11138
DOI : 10.4000/msh.11138
ISSN : 1950-6821

**Éditeur**
Centre d'analyse et de mathématique sociales de l'EHESS

**Édition imprimée**
Date de publication : 30 décembre 2009
Pagination : 79-91
ISSN : 0987-6936

**Référence électronique**
André Hardy et Nathanael Kasoro, « Une nouvelle méthode de classification pour des données intervalles », *Mathématiques et sciences humaines* [En ligne], 187 | Automne 2009, mis en ligne le 15 décembre 2009, consulté le 23 juillet 2020. URL : http://journals.openedition.org/msh/11138 ; DOI : https://doi.org/10.4000/msh.11138

*Math. & Sci. hum. / Mathematics and Social Sciences* (47e année, n° 187, *2009*(3), p. 79–91)

# UNE NOUVELLE MÉTHODE DE CLASSIFICATION POUR DES DONNÉES INTERVALLES

André HARDY[1], Nathanaël KASORO[2]

RÉSUMÉ – *Cet article propose une nouvelle méthode de classification automatique pour des données intervalles. C'est une extension d'une méthode de classification classique à des données intervalles. La procédure classique est basée sur la théorie des processus ponctuels, et plus particulièrement sur le processus de Poisson homogène. La première partie de la nouvelle méthode est une procédure de classification monothétique divisive. La règle de coupure utilise une extension à des données intervalles du critère de classification des Hypervolumes. L'étape d'élagage utilise deux tests statistiques du quotient de vraisemblance basés sur le processus de Poisson homogène : le test des Hypervolumes et le Gap test. Nous obtenons alors un arbre de décision. La seconde partie de la méthode est une procédure 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 des données générées et sur des données réelles. Elle est comparée à d'autres méthodes de classification disponibles pour des données intervalles.*

MOTS CLÉS – Arbre de décision, Classification automatique, Critère des Hypervolumes, Maximum de vraisemblance, Processus de Poisson

SUMMARY – A new clustering method for interval data

*This paper presents a new clustering method for interval data. It is an extension of a classical clustering method to interval data. The classical procedure is based on the theory of point processes, and more particularly on the homogeneous Poisson process. The first part of the new method is a monothetic divisive procedure. The cut rule is an extension to interval data of the Hypervolumes clustering criterion. The pruning step uses two statistical likelihood ratio tests based on the homogeneous Poisson process : the Hypervolumes test and the Gap test. 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 generated data set and to a real data set. It is compared with other clustering methods available for interval data.*

KEYWORDS – Clustering, Decision tree, Hypervolumes criterion, Maximum likelihood estimation, Poisson process

## 1. INTRODUCTION

Le but de cet article est de présenter une nouvelle méthode de classification monothétique divisive pour des données intervalles. La section deux introduit le modèle statistique sous-jacent à cette méthode qui se base sur le processus de Poisson homogène et le critère de

[1]FUNDP-Université de Namur, Département de Mathématique, 8 Rempart de la Vierge, B-5000 Namur, Belgique, andre.hardy@fundp.ac.be

[2]Département de Mathématique et Informatique, Université de Kinshasa, B.P. 190, Kinshasa, République Démocratique du Congo, kasoro.mulenda@yahoo.fr

classification des Hypervolumes. La section trois décrit deux tests statistiques qui permettent de tester si une classe doit être divisée en deux. Ces tests utilisent également le processus de Poisson homogène. La quatrième section présente la méthode de classification classique HOPP dont l'extension à des données intervalles fait l'objet de cet article. La section cinq détaille les différentes étapes de la nouvelle méthode SPART (la construction de l'arbre, la procédure d'élagage et le processus de recollement). Dans la sixième section un ensemble de données artificielles est utilisé afin de mettre en évidence les spécificités de la nouvelle méthode ; elle est ensuite appliquée à un jeu de données réelles. La section sept présente quelques conclusions.

## 2. UN MODÈLE STATISTIQUE POUR LA CLASSIFICATION

Un processus ponctuel $N(.)$ sur un ensemble $D$ inclu dans l'espace euclidien $R^p$ est une distribution aléatoire de points dans $D$ telle que seul un nombre fini d'entre eux sont répartis dans tout sous-ensemble borné $A \subset D$. Le processus ponctuel le plus simple est le processus de Poisson [Cox, Isham, 1980 ; Karr, 1991]. Le modèle statistique pour la classification sur lequel s'appuie la nouvelle méthode de classification présentée dans cet article, repose sur le Processus de Poisson homogène.

### 2.1. DÉFINITION : LE PROCESSUS DE POISSON HOMOGÈNE

$N(.)$ est un processus de Poisson homogène de taux $\lambda$ ($\lambda \in R$) sur l'ensemble $D \subset R^p$ $(0 < m(D) < \infty)$ si pour des ensembles compacts, bornés et disjoints $A_1, A_2, \cdots, A_k \subset D$, les variables aléatoires $N(A_1), N(A_2), \cdots, N(A_k)$ qui comptent le nombre de points dans les ensembles $A_i$ $(i = 1, \cdots, k)$ sont indépendantes et ont respectivement des distributions de Poisson de paramètres $\lambda m(A_1), \lambda m(A_2), \cdots, \lambda m(A_k)$ où $m(.)$ est la mesure de Lebesgue. Les processus de Poisson sont totalement déterminés par leur taux $\lambda$ [Rasson, 1976]. Pour un processus de Poisson homogène, ou stationnaire, le taux $\lambda$ est constant.

### 2.2. PROPRIÉTÉ D'UNIFORMITÉ CONDITIONNELLE

Cette propriété du processus de Poisson homogène [Cox, Isham, 1980] est à la base du modèle statistique sur lequel repose la nouvelle méthode. Elle va nous permettre d'écrire des fonctions de vraisemblance, et donc de calculer des estimateurs du maximum de vraisemblance.

PROPRIÉTÉ 1. *Si $N(D)$, le nombre de points générés par le processus de Poisson homogène dans $D$, est fixé à $n$, alors ces $n$ points sont distribués indépendamment, suivant une loi uniforme dans $D$.*

*Par conséquent, si $x = (x_1, \cdots, x_n)$ est une réalisation d'un processus de Poisson homogène dans $D$, la fonction de vraisemblance $L$ de l'échantillon s'écrit*

$$L(D; x_1, \cdots, x_n) = \frac{1}{(m(D))^n} \prod_{i=1}^{n} I_D(x_i) \qquad (1)$$

*où $I_D$ est la fonction indicatrice de l'ensemble $D$.*

### 2.3. PROBLÈME DE DÉPART : L'ESTIMATION D'UN ENSEMBLE CONVEXE

Le point de départ de notre modèle est le problème suivant : « Étant donné la réalisation d'un processus de Poisson homogène $N(.)$ d'intensité $\lambda$ dans un ensemble convexe compact $D$ du plan, estimer $D$ en utilisant des méthodes d'inférence statistique ». Le paramètre à estimer est le domaine $D$, qui est un paramètre de dimension infinie. La solution de ce problème fut proposée par [Ripley, Rasson, 1977]. La fonction de vraisemblance (1) peut s'écrire sous la forme

$$L(D; x_1, \cdots, x_n) = \frac{1}{(m(D))^n} I_D(H(x_1, \cdots, x_n))$$

où $H(x_1, \cdots, x_n)$ est l'enveloppe convexe des points appartenant à l'ensemble $D$. Le théorème de factorisation en statistique inférentielle [Lejeune, 2004] permet d'affirmer que $H(X_1, \cdots, X_n)$ est une statistique exhaustive pour le paramètre $D$. Cet estimateur maximise la fonction de vraisemblance. C'est donc aussi l'estimateur du maximum de vraisemblance du domaine $D$. L'enveloppe convexe $H(x_1, \cdots, x_n)$ sera toujours incluse dans le domaine $D$. L'estimateur est donc biaisé. Un estimateur non biaisé de $D$ est obtenu en prenant une dilatation homothétique de l'enveloppe convexe $H(X_1, \cdots, X_n)$ à partir de son centre de gravité. Une estimation du coefficient de dilatation $c$ est donnée par

$$c = \frac{n}{\sqrt{n - V_n}}$$

où $n$ est le nombre d'observations dans l'échantillon, et $V_n$ le nombre de sommets de l'enveloppe convexe $H(x_1, \cdots, x_n)$ [Moore, 1984].

### 2.4. LE MODÈLE STATISTIQUE ET LE CRITÈRE DES HYPERVOLUMES

Nous supposons que nous traitons un problème de classification lorsque les $n$ observations $p$-dimensionnelles $x_1$, ..., $x_n$ sont générées par un processus de Poisson homogène $N$ dans un ensemble $D$ de l'espace Euclidien $R^p$ ($0 < m(D) < \infty$ ) où $D$ est l'union de $k$ domaines convexes compacts disjoints $D_1, \cdots, D_k$. Le problème statistique consiste à estimer les domaines inconnus $D_i$ dans lesquels les points ont été générés. La fonction de vraisemblance $L$ s'écrit

$$L(D_1, \cdots, D_k; x_1, \cdots, x_n) = \frac{1}{(m(D_1 \cup \cdots \cup D_k))^n} \prod_{i=1}^{n} I_{D_1 \cup \cdots \cup D_k}(x_i).$$

Notons par $C_i \subset \{x_1, \cdots, x_n\}$ l'ensemble des points appartenant à $D_i$ $(1 \leq i \leq k)$. Maximiser la fonction de vraisemblance $L$ revient à minimiser $\sum_{i=1}^{k} m(H(C_i))$ sur l'ensemble $\mathcal{P}_k$ des partitions $\{C_1, \cdots, C_k\}$ de l'ensemble $C$ en $k$ classes, où $C$ est l'ensemble des points appartenant au domaine $D$, $H(C_i)$ l'enveloppe convexe des points appartenant à $C_i$ et $m(H(C_i))$ la mesure de Lebesgue multidimensionnelle de cette enveloppe convexe [Hardy, 1983].

Le critère de classification des Hypervolumes est donc défini par

$$W_k = \sum_{i=1}^{k} m(H(C_i)) = \sum_{i=1}^{k} \int_{H(C_i)} m(dx).$$

Le problème de classification revient donc à trouver les $k$ sous-groupes $C_i$ contenant tous les points et tels que la somme des mesures de Lebesgue des enveloppes convexes disjointes $H(C_i)$ est minimale. Dans le contexte classificatoire, la partition optimale $P^*$ est donc donnée par

$$P^* = arg \min_{P_k \in \mathcal{P}_k} \sum_{i=1}^{k} \int_{H(C_i)} m(dx).$$

Un algorithme de complexité polynomiale a été écrit afin de trouver un optimum local intéressant de ce critère de classification [Hardy, 1996].

## 3. TESTS DU QUOTIENT DE VRAISEMBLANCE POUR LA DÉTERMINATION DU NOMBRE DE CLASSES

Nous décrivons deux tests statistiques du quotient de vraisemblance pour la détermination du nombre de classes basés sur le processus de Poisson homogène.

### 3.1. LE TEST DES HYPERVOLUMES [HARDY, 1996]

Nous nous plaçons toujours dans le cadre du modèle statistique pour la classification basé sur le processus de Poisson homogène. Notons par $C = \{C_1, C_2, ..., C_k\}$ la partition optimale (au sens du critère des Hypervolumes) de l'échantillon en $k$ classes et $B = \{B_1, B_2, ..., B_{k-1}\}$ la partition optimale en $k-1$ classes. Si $t$ représente le nombre de classes "naturelles", on teste l'hypothèse nulle $H_0 : t = k$ contre l'hypothèse alternative $H_1 : t = k - 1$ ($k \geq 2$). La statistique du test, obtenue en appliquant la méthode du quotient de vraisemblance, est donnée par ([Hardy, 1996])

$$S(x_1, \cdots, x_n; k) = \frac{W_k}{W_{k-1}}$$

où $W_k$ (respectivement, $W_{k-1}$) est la valeur du critère de classification des Hypervolumes calculé pour la meilleure partition en $k$ (respectivement, $k-1$) classes.

Malheureusement la loi de la statistique $S$ n'est pas connue. Mais, comme $W_k \geq 0, \forall k \geq 1$ et $W_k \leq W_{k-1}, \forall k \geq 2$, on a toujours la propriété suivante : $S(x_1, \cdots, x_n; k) \in [0, 1[$. Nous pouvons donc utiliser la règle de décision empirique suivante : rejeter $H_0$ lorsque $S$ est « proche » de 1. Cette règle est appliquée de manière séquentielle : si $k_0$ est la plus petite valeur de $k \geq 2$ pour laquelle on rejette $H_0$, $k_0 - 1$ sera choisi comme le nombre adéquat de classes « naturelles ». En pratique, lorsqu'une structure existe dans les données, cette valeur $k_0$ est facilement détectable par la présence d'un coude dans le graphe de $S$ en fonction de $k$. Néanmoins, plus formellement, des tests de permutation ont été utilisés de manière à calculer des $p$-valeurs pour cette statistique de test [Hardy, Blasutig, 2007]. Les deux approches conduisent habituellement aux mêmes conclusions.

### 3.2. LE GAP TEST

Le Gap test [Kubushishi, 1996] utilise le même modèle statistique pour la classification. On teste $H_0$ : les $n$ 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 $D_1 \cap D_2 = \emptyset$ et $n_1 + n_2 = n$. Les ensembles $D, D_1, D_2$ sont inconnus. Notons par $C$ (respectivement $C_1$, $C_2$) l'ensemble des points appartenant à $D$ (respectivement $D_1$, $D_2$). La statistique du test est donnée par [Kubushishi, 1996]

$$Q(x_1, \cdots, x_n) = \left(1 - \frac{m(\triangle)}{m(H(C))}\right)^n$$

où $H(C)$ est l'enveloppe convexe des points appartenant à $C$, $\triangle$ l'espace vide (Gap space) entre les classes et $m$ la mesure de Lebesgue multidimensionnelle. $\triangle$ est l'ensemble $H(C)$ duquel on a retiré les sous-ensembles $H(C_1)$ et $H(C_2)$. La statistique du test dépend donc de la mesure de Lebesgue de l'espace vide entre les classes. La règle de décision est la suivante [Kubushishi, 1996] : on rejette $H_0$, au niveau $\alpha$, si

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

## 4. LA MÉTHODE DE CLASSIFICATION CLASSIQUE HOPP (HOmogeneous Poisson Process)

HOPP [Pirçon, 2004] est une méthode de classification pour des données quantitatives classiques élaborée à partir du modèle statistique pour la classification basé sur le processus de Poisson homogène. La nouvelle méthode SPART décrite dans la section 5 est une extension de la méthode HOPP à des données intervalles. Nous supposons que les points observés sont générés par un processus de Poisson homogène dans un domaine $D \subset R^p$, où $D$ est l'union de $k$ domaines convexes compacts disjoints $D_1$, ..., $D_k$. La méthode HOPP comporte trois parties.

La première étape concerne la construction de l'arbre. HOPP est une méthode de classification monothétique divisive. Nous utiliserons, comme critère de coupure, le critère des Hypervolumes. On commence par couper le premier noeud $C$ (la racine) en deux parties. La méthode étant monothétique, pour chacune des variables, les ensembles convexes de points sont des intervalles de points. La mesure de Lebesgue d'un intervalle est sa longueur $l$. On recherche donc, pour le groupe $C$ et pour chaque variable, les deux intervalles $I_1$ et $I_2$, contenant tous les points, tels que la somme de leurs longueurs est minimale. On retient alors la meilleure variable (celle pour laquelle cette somme est minimale), les intervalles correspondants, et la bipartition de $C$ en deux sous-classes $C_1$ et $C_2$ obtenue à partir de ces intervalles. On recommence ensuite le processus sur les sous-classes obtenues à l'itération précédente en choisissant, à chaque fois, le meilleur noeud et la meilleure variable. La construction se termine lorsqu'un critère d'arrêt est vérifié. Il s'agira du nombre de points dans un noeud. On peut, par exemple, demander que chaque classe contienne au moins cinq pourcents des données de l'échantillon. Ce paramètre doit être fixé par l'utilisateur.

À la fin de la première étape, on obtient un arbre assez volumineux. Remarquons qu'aucun test statistique n'a été utilisé pour vérifier si les coupures qui ont été faites sont statistiquement valides. La deuxième étape de la méthode est une procédure d'élagage qui permet d'obtenir un arbre réduit. Pour la réaliser on utilise le test des Hypervolumes ou le Gap test. Ces tests sont appliqués à chaque noeud afin de voir si chacune des coupures

faites dans l'étape de construction de l'arbre est justifiée. On teste donc à chaque noeud les hypothèses suivantes :

- $H_0$ : les points sont distribués dans un seul domaine
- $H_1$ : les points sont distribués dans deux domaines disjoints.

Lorsque l'hypothèse nulle n'est pas rejetée, on conclut que la coupure est mauvaise. Par contre si l'hypothèse nulle est rejetée, on décide que la coupure est bonne. À la fin du processus on utilise la règle suivante : on élague toutes les branches qui ne contiennent que des mauvaises coupures. Illustrons ce procédé. Supposons que nous obtenions l'arbre suivant à la fin de l'étape de construction de l'arbre (cf. Figure 1).

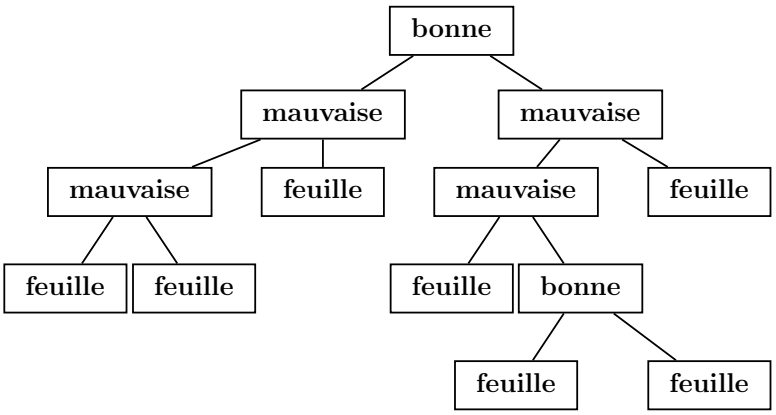


FIGURE 1. Arbre avant élagage

Après avoir appliqué la règle d'élagage, on obtient l'arbre de décision suivant :

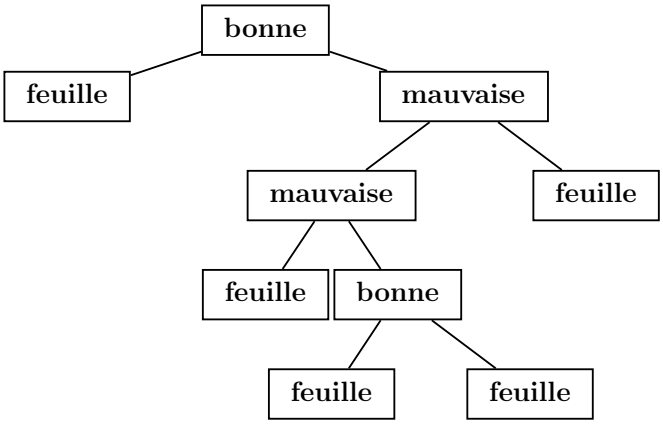


FIGURE 2. Arbre après élagage

Dans certain cas la structure naturelle des données ne peut être obtenue à la fin de l'étape d'élagage (cf. Figure 2). C'est pourquoi une étape de recollement a été ajoutée à la procédure. Ainsi des tests supplémentaires sont effectués uniquement sur les classes

qui ne proviennent pas d'un même noeud. Par exemple, dans l'arbre qui suit (Figure 3), il s'agira de tester si les classes $C_{12}$ et $C_{21}$ doivent être considérées comme deux classes distinctes, ou comme une seule classe. Pour cela nous utilisons à nouveau le test des Hypervolumes ou le Gap test.

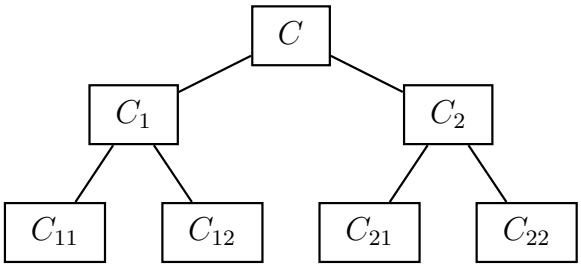


FIGURE 3. Arbre avant recollement

Si au moins un regroupement est effectué dans l'étape de recollement, HOPP perd son caractère hiérarchique. Elle devient alors une méthode de partitionnement.

## 5. EXTENSION À DES VARIABLES INTERVALLES : LA MÉTHODE SPART

Une variable à valeurs d'ensemble $Y$ est une variable intervalle si $\forall x_i \in E$, $Y(x_i) = [\alpha, \beta]$ où $[\alpha, \beta]$ est un intervalle de $R$.

Afin d'adapter la méthode HOPP à des données intervalles, on utilise une modélisation. Chaque intervalle est représenté par son milieu et sa demi-longueur, donc par un point dans un espace bidimensionnel. Dans l'exemple de la Figure 4, l'intervalle $[\alpha_1, \beta_1]$ (respectivement, $[\alpha_2, \beta_2]$, $[\alpha_3, \beta_3]$) est représenté par le point $I_1$ (respectivement, $I_2$, $I_3$).

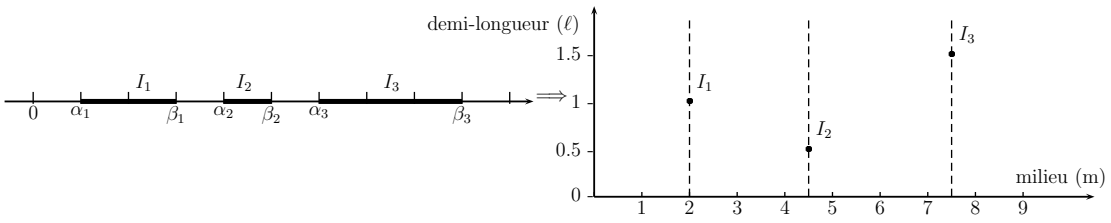


FIGURE 4. Modélisation « milieu - demi-longueur »

SPART est une méthode de classification monothétique. Dans chacun des $p$ espaces « milieu - demi-longueur », les intervalles sont des points

$$Y(x_i) = [\alpha_i, \beta_i] \longrightarrow (m_i, \ell_i).$$

La règle de coupure est la suivante : on considère toutes les bipartitions d'une classe $C$ en deux classes $\{C_1, C_2\}$, qui respectent l'ordre des milieux des intervalles. Il s'agit donc des bipartitions générées par des droites verticales.

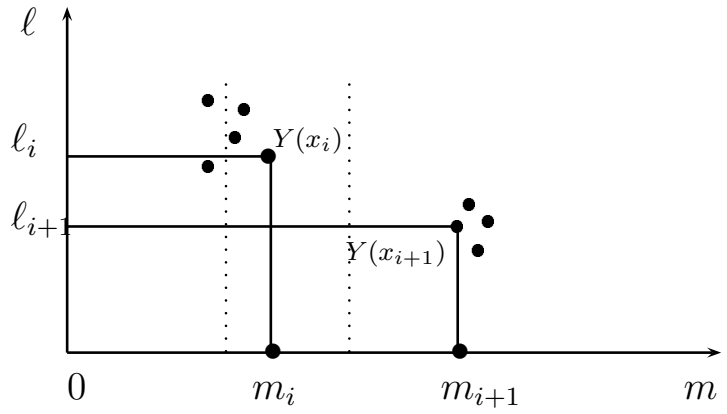


FIGURE 5. Bipartition d'une classe

On définit une extension $m_E(\triangle)$ de la mesure de l'espace vide $\Delta$ entre les classes, pour des données intervalles, de la façon suivante :

$$m_E(\Delta) = \int_{m_i}^{m_{i+1}} dx + \int_{min(l_i,l_{i+1})}^{max(l_i,l_{i+1})} dy$$
$$= (m_{i+1} - m_i) + (max(l_i, l_{i+1}) - min(l_i, l_{i+1})).$$

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 milieu de l'intervalle.

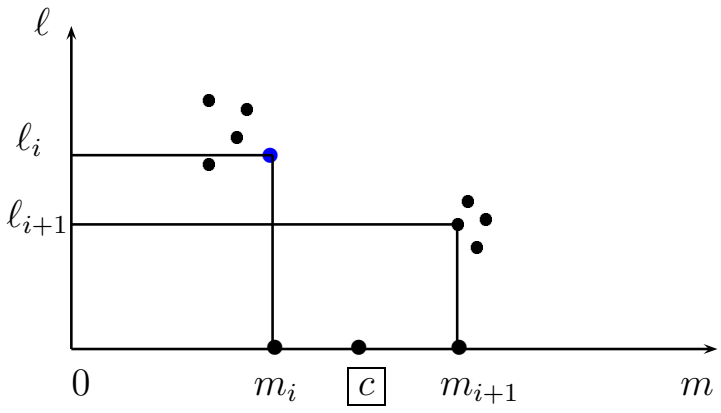


FIGURE 6. Valeur de coupure

Dès que l'on a déterminé la valeur de coupure, qui est basée sur le critère $m_E(\Delta)$, la procédure de bipartition est identique à celle utilisée dans la méthode DIV.

Soit $x_\ell \in C$. On a $Y(x_\ell) = [\alpha_\ell, \beta_\ell]$ 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

$$q_c(x_\ell) = \begin{cases} 0 & \text{si} \quad \mathrm{m}_\ell \leq \mathrm{c} \\ 1 & \text{sinon.} \end{cases}$$

On obtient alors la bipartition souhaitée :
– $C_1 = \{x \in C : q_c(x) = 0\} = q_c^{-1}(\{0\})$
– $C_2 = \{x \in C : q_c(x) = 1\} = q_c^{-1}(\{1\})$.

Le critère d'arrêt est le même que celui utilisé pour la méthode classique HOPP : le nombre d'objets dans un noeud. Ce paramètre est fixé par l'utilisateur.

## 6. APPLICATIONS

On considère tout d'abord une illustration basée sur des données générées, afin de mettre en évidence d'une part le fonctionnement de la méthode, et d'autre part l'utilité de l'étape de recollement. On appliquera ensuite 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.

### 6.1. ILLUSTRATION

Dans le premier exemple, nous avons 100 objets sur lesquels on mesure deux variables intervalles. Chacun des objets peut donc être représenté par un rectangle dans un espace bidimensionnel. Les objets ont été générés de manière à obtenir une structure hyperellipsoïdale en 3 classes naturelles, telles qu'aucune d'entre elles ne peut être séparée des deux autres par un hyperplan (une droite). Ces données sont visualisées sur la Figure 7.

À la fin de l'étape d'élagage, nous obtenons la partition suivante en 4 classes (cf. Figure 8).

Après l'étape de recollement, les trois classes hyperellipsoïdales naturelles sont retrouvées. Dans cet exemple nous avons utilisé le Gap test dans les procédures d'élagage et de recollement.

D'une manière évidente, SCLASS et DIV ne peuvent restituer la partition naturelle des données en trois classes, car ces méthodes sont monothétiques, et effectuent des coupures parallèles aux axes. L'étape de recollement de SPART a donc pour avantage de permettre à cette méthode de retrouver, dans certains cas, la structure naturelle des données, lorsque celle-ci n'apparaît pas à la fin de l'étape d'élagage. Par contre elle a l'inconvénient de faire perdre à la méthode son caractère hiérarchique, et par conséquent, pour certaines classes (par exemple, celles qui ont été regroupées), leur interprétation monothétique.

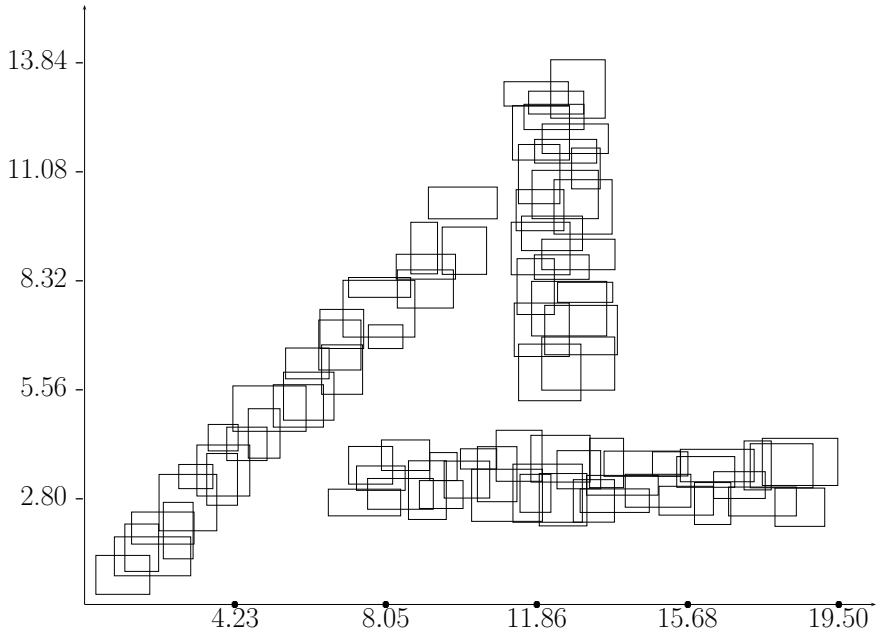


FIGURE 7. Trois classes hyperellipsoïdales

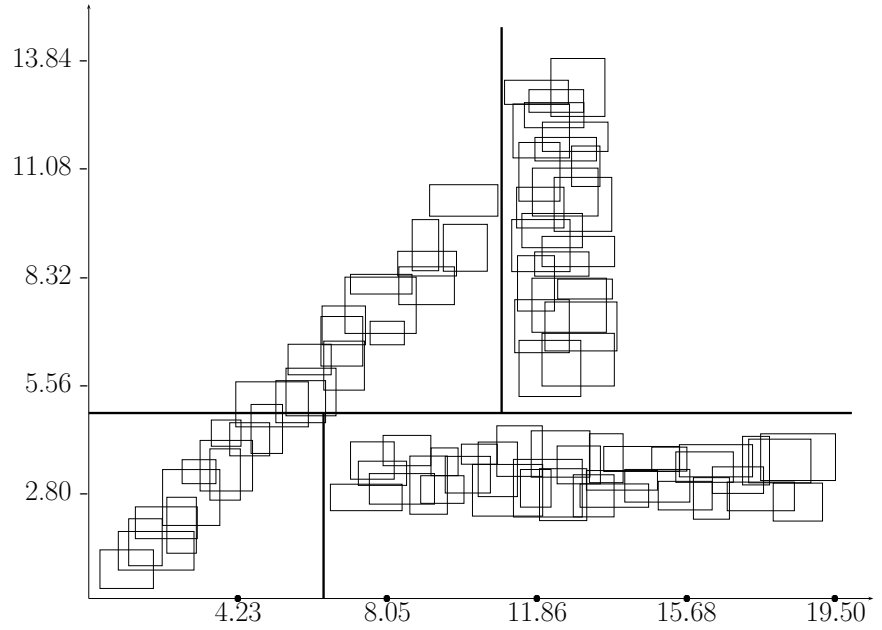


FIGURE 8. Quatre classes obtenues à la fin de l'élagage

## 6.2. APPLICATION RÉELLE

L'ensemble de données « cars » est constitué de 33 voitures disponibles en 2001, sur lesquelles ont été mesurées 8 variables intervalles.

| | | | | | |
|---|---|---|---|---|---|
| 0 | AlfaRomo145 | 11 | FiatPunto | 22 | MercedesClasseS |
| 1 | AlfaRomo156 | 12 | FordFiesta | 23 | NissanMicra |
| 2 | AlfaRomo166 | 13 | FordFocus | 24 | OpelCorsa |
| 3 | AstonMartinDB7 | 14 | HondaNSK | 25 | OpelVectra |
| 4 | AudiA3 | 15 | LamborghiniDiablo | 26 | Porsche |
| 5 | AudiA6 | 16 | LanciaY | 27 | RenaultTwingo |
| 6 | AudiA8 | 17 | LanciaK | 28 | Rover25 |
| 7 | BMWSrie3 | 18 | MaseratiGT | 29 | Rover75 |
| 8 | BMWSrie5 | 19 | MercedesSL | 30 | SkodaFabia |
| 9 | BMWSrie7 | 20 | MercedesClasseC | 31 | SkodaOctavia |
| 10 | Ferrari | 21 | MercedesClasseE | 32 | VolkswagenPassat |

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.

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.

Les 4 classes peuvent être étiquettées de la manière suivante.
- Classe 1 : voitures citadines
- Classe 2 : voitures berline
- Classe 3 : modèles sport
- Classe 4 : voitures limousines

La méthode DIV donne la même partition en 4 classes. La méthode SCLASS, quant à elle, donne un résultat différent, dont les classes sont difficilement interprétables. La classification produite par SPART est reprise à la Figure 9.

## 7. CONCLUSION

SPART est une nouvelle méthode de classification monothétique divisive pour des données intervalles. Elle est basée sur une extension aux données intervalles du critère de classification des Hypervolumes, du test des Hypervolumes et du Gap test. Elle 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 produisent souvent des résultats semblables. SPART donne généralement de meilleurs résultats que DIV lorsqu'on est en présence de classes hyperellisoïdales. Ceci s'explique principalement par le fait que DIV est basée sur une extension du critère de la variance intra-classe, et que ce critère est biaisé par rapport aux classes de forme hyperellipsoïdale. SPART se comporte aussi mieux que DIV lorsque des coupures parallèles aux axes ne permettent pas de mettre en évidence la structure naturelle des données. D'un point de vue théorique, la principale différence entre SCLASS et SPART réside dans le fait que SCLASS utilise un modèle pour la classification basé sur le processus de Poisson non homogène tandis que SPART utilise le processus de Poisson homogène. SCLASS est donc

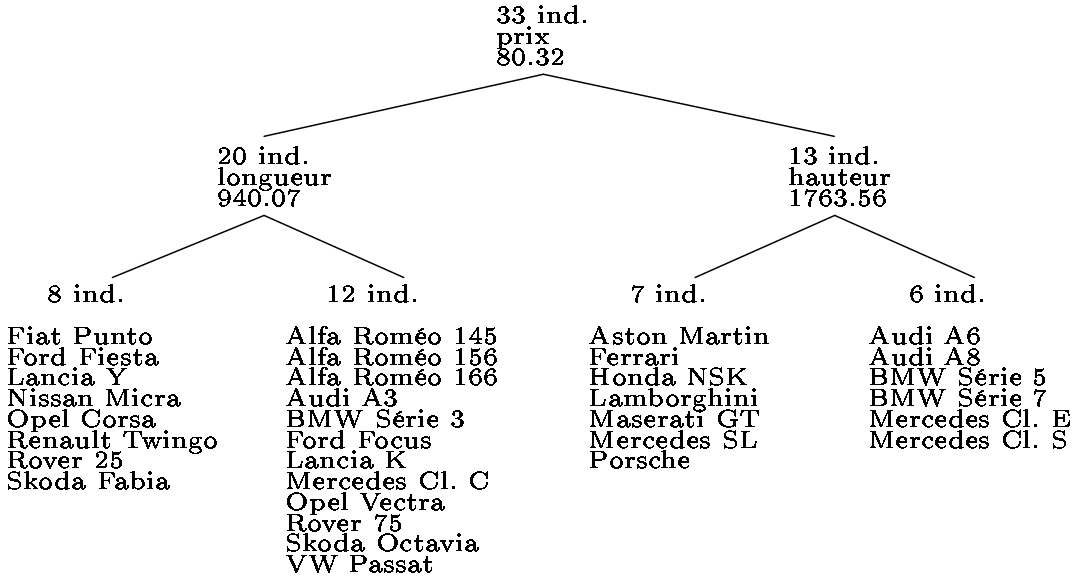


FIGURE 9. Classification de l'ensemble de données « Cars »

plus complexe que SPART d'un point de vue algorithmique car SCLASS nécessite l'estimation de l'intensité du processus de Poisson non homogène. Pour les mêmes raisons, le critère associé à SCLASS est également plus complexe. D'autre part, dans sa version actuelle, SCLASS ne comporte pas d'étape d'élagage ni d'étape de recollement. Enfin la procédure SPART détermine automatiquement le nombre de classes. Il doit être fixé au préalable dans SCLASS et DIV.

## BIBLIOGRAPHIE

BOCK H.H., DIDAY E. (eds), *Analysis of Symbolic Data : Exploratory methods for extracting statistical information from complex data*, Springer, 2000.

CHAVENT M., "A monothetic clustering method", *Pattern Recognition Letters*, 1998, p. 989-996.

COX D.R., ISHAM V., *Point Processes*, London, Chapman and Hall, 1980.

DIDAY E., NOIRHOMME M. (eds), *Symbolic data analysis and the SODAS 2 software*, Wiley & Sons, 2007.

HARDY A., RASSON J.P., "Une nouvelle approche des problèmes de classification automatique", *Statistique et analyse des données* 7(2), 1982, p. 41-56.

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

HARDY A., "On the number of clusters", *Computational Statistics and Data Analysis*, 1996, p. 83-96.

HARDY A., "A heuristic approach for the hypervolumes method in cluster analysis", *Jorbel*, 1996, p. 43-55.

HARDY A., BLASUTIG L., *Les tests de permutation pour la statistique des hypervolumes*, Rapport de Recherche, FUNDP, Université de Namur, 2007.

KARR A.F., *Point Processes and their Statistical Inference*, New York, Marcel Dekker, 1991.

KUBUSHISHI T., *On some applications of point process theory in cluster analysis and pattern recognition*, Thèse de Doctorat, FUNDP, Université de Namur, Namur (Belgique), 1996.

LEJEUNE M., *Statistique. La théorie et ses applications*, Springer, 2004.

MOORE M., "On the estimation of a convex set", *Annals of Statistics*, vol. 12(3), 1984, p. 1090-1099.

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

RASSON J.P., *De quelques problèmes d'entropie et d'inférence pour des processus ponctuels*, Thèse de Doctorat, FUNDP, Université de Namur, Namur (Belgique), 1976.

RASSON J.P., LALLEMAND P. PIRÇON J.Y., ADANS S., "Unsupervised divisive classification", in E. Diday and M. Noirhomme (eds), *Symbolic Data Analysis and the Sodas 2 Software* , Wiley, 2008.

RIPLEY B.D., RASSON J.P., "Finding the edge of a Poisson forest", *Journal of Applied Probability*, 1977, p. 483-491.