S-Class, A Divisive Clustering Method, and Possible "Dual" Alternatives

Jean-Paul Rasson, François Roland, Jean-Yves Pirçon, Séverine Adans, and Pascale Lallemand

Facultés Universitaires Notre-Dame de la Paix
Unité de Statistiques
5000 Namur, Belgique
(e-mail: jean-paul.rasson@fundp.ac.be, francois.roland@fundp.ac.be, severine.adans@fundp.ac.be)

Abstract. A new partitioning method based on the non-homogeneous Poisson processes is presented. The principle of construction is of hierarchical divisive monothetic type. A variable is selected at each stage to cut a group into two subsets in a recursive way. The criterion consists in maximizing the 'gap' between the data. This last-one is deduced from the maximum likelihood criterion. A pruning phase, that is a simplification of the tree structure, based on the Gap test is then performed. An application of this algorithm on the well-know Ichino's oils dataset (interval data) is described.

Keywords: Clustering trees, Non-homogeneous Poisson processes, Gap test, Symbolic data.

1 Introduction

One of the most common tasks in data analysis is the detection and construction of groups of objects in a population E such that objects from the same group show a high similarity whereas objects from different groups are typically more dissimilar. Such groups are usually called 'clusters' and must be constructed on the basis of the data which were recorded for the objects. This problem is know as clustering.

The present method is a divisive monothetic clustering method for a symbolic $n \times p$ data array $\underline{X}$.

The resulting classification structure is a k-partition.

2 Input Data: Interval Data

This algorithm studies the case where n symbolic objects are described by p interval variables $Y_1, \ldots, Y_p$.

The interval-valued variable $Y_j (j = 1, \ldots, p)$ is measured for each element of the basic set $E = \{1, \ldots, n\}$. For each element $x \in E$, we denote the interval $Y_j(x)$ by $[\underline{y}_{jx}, \bar{y}_{jx}]$, thus $\underline{y}_{jx}$ (resp. $\bar{y}_{jx}$) is the lower (resp. the upper) bound of the interval $Y_j(x) \subseteq \mathcal{R}$.

An example is given by table 1.

3 The Clustering Tree Method

The proposed algorithm is a recursive one intended to divide a given population of symbolic objects into classes. According to the clustering tree method, nodes are split recursively by choosing the best interval variable.

The original contribution of this method lies in the way of splitting a node. The cut will be based on the only assumption that the distributions of points can be modeled by non-homogeneous Poisson process, where the intensity will be estimated by the kernel method. The cut will be made in order to maximize the likelihood function.

3.1 General Hypothesis: Non-Homogeneous Poisson Process

The only assumption on which the clustering problem rests is that the observed points are generated by a non-homogeneous Poisson process with intensity $q(.)$ and are observed in E, where E is the union of k disjoint convex fields.

The likelihood function, for the observations $\underline{x} = (x_1, x_2, \ldots, x_n)$ with $x_i \in R^d, i = 1, \ldots, n$ is:

$$f_E(\underline{x}) = \frac{1}{(\rho(E))^n} \prod_{i=1}^{n} \mathbb{1}_E(x_i).q(x_i)$$

where

- $\rho(E) = \int_E q(x)dx$ is the integrated intensity;
- $q(.)$ is the process intensity ($q(x) > 0 \ \forall x$).

Consequently, if the intensity of the process is known, the solution of the maximum likelihood will correspond to k disjoint convex fields containing all the points and for which the sum of the integrated intensities is minimal. For an homogenous Poisson process on the line, this gives exactly the N-N rule. When the intensity is unknown, it will be estimated.

3.2 Kernel Method

To estimate the intensity of a non-homogeneous Poisson process, the non-parametric kernel method is used. Because this algorithm proceeds in a monothetic way, formulas dont't need to be extended beyond one dimension. The kernel estimator, which is a sum of 'bumps', each of these centered on an observation, is defined by:

$$\hat{q}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big)$$

where

- K is the kernel and is a positive continuous symmetric function satisfying $\int K(x)dx = 1$. The kernel determines the shapes of the bumps.
- h is the window width, also called the smoothing parameter and determines the width of the bumps.

The choice of the smoothing parameter is important. If it is too small, the estimator degenerates into a succession of peaks located at each point of the sample. If it's too large, the estimation approaches an uniform law and then we will have a loss of details.

3.3 Bumps and Multi-modalities

Within the clustering context, Silverman ([Silverman, 1981], [Silverman, 1986]) defined a mode in a density f as a local maximum while a bump is characterized by an interval, in such way that the density is concave on this interval but not on a larger interval.

In the framework of density estimation by the kernel method, the number of modes will be determined by the smoothing parameter, following Silverman's assertion : the number of modes is a decreasing function of the smoothing parameter h ([Silverman, 1981],[Silverman, 1986]).

This has been proved at least for the normal kernel defined by :

$$K_{\mathcal{N}}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Consequently, this one was prefered to perform estimation of the intensity of the non-homogeneous Poisson process.

Because of this choice, there is a critical value h_{crit} of the smoothing parameter for which the estimation changes from unimodality to multimodality. The split criterion will seek this value.

3.4 Splitting Criteria

For each variable, a dichotomic process computes the highest value of parameter h, giving a number of modes of the associated intensities strictly larger than 1. Once this h determined, E is split into two convex disjoint fields E_1 and E_2, such that $E = E_1 \cup E_2$, for which the likelihood function

$$f_{E_1,E_2}(\underline{x}) = \frac{1}{(\rho(E_1)+\rho(E_2))^n} \prod_{i=1}^{n} 1\!\!1_{E_1 \cup E_2}.\hat{q}(x_i)$$

is maximum, i.e. for which the integrated density $\rho(E_1) + \rho(E_2)$ is smallest.

Since the algorithm proceeds variable by variable, the best variable, i.e. the one which generates the "largest gap" (the density integrated on this gap is the largest), is selected.

This procedure is recursively performed until some stopping rule is fulfilled: the number of points in a node must be under a cut-off value.

3.5 Pruning Method

At the end of the splitting process, a large tree is obtained. A pruning method to select the best subtree was then developped. This pruning method takes the form of a classical hypothesis test: the Gap test ([Kubushishi, 1996], [Rasson and Kubushishi, 1994]).

The principle is the following: each cut is examined to determine if it is a good one (Gap Test satisfied) or a bad one (Gap Test unsatisfied). In the case of two classes D_1 and D_2, with $D_1 \cup D_2 = D$, the hypotheses are:

$$H_0\text{: there are } n = n_1 + n_2 \text{ points in } D_1 \cup D_2$$

$$\text{VS}$$

$$H_1\text{: there are } n_1 \text{ points in } D_1 \text{ and } n_2 \text{ points in } D_2, \text{ with } D_1 \cap D_2 = \emptyset.$$

This pruning method crosses the tree branch by branch, from its root to its leaves, in order to index the good cuts and the bad cuts. The ends of the branches for which there are only bad cuts are pruned.

3.6 Application to Interval Data

The current problem is to apply this new method to symbolic data of interval type. Let an interval set

$$I = \{[a_i, b_i],\ i = 1, \ldots, n,\ a_i \leq b_i\}.$$

The usual distance used for interval variables is the Hausdorff distance:

$$d_H([a_1, b_1], [a_2, b_2]) = Max(|a_1 - a_2|, |b_1 - b_2|)$$

or ([Chavent and Lechevallier, 2002], [Chavent, 1997])

$$d([a_1, b_1], [a_2, b_2]) = |M_1 - M_2| + |L_1 - L_2|$$

where $M_i = \frac{a_i + b_i}{2}$ is the middle point of the interval $[a_i, b_i]$ and $L_i = \frac{b_i - a_i}{2}$ is its half-length.

So each interval can be represented by its coordinates *(middle,half-length)*, on the space $(M, L) \subseteq I\!R \times I\!R^+$.

It is clear that separations must respect the order of the classes centers and thus, in the half-plane $I\!R \times I\!R^+$, only partitions invariant in relation to M are considered.

In the most general case of a non-homogeneous Poisson process, the integrated intensity has to be minimized:

$$\int_{M_i}^{M_{i+1}} \rho_1(m)dm + \int_{Min(L_i, L_{i+1})}^{Max(L_i, L_{i+1})} \rho_2(l)dl. \qquad (1)$$

Any bipartition generated by a point being located inside the interval which maximizes (1) is appropriate.

3.7 Set of Binary Questions for Interval Data

In the framework of the divisive clustering method, the split of a node C is performed on the basis of one single variable (suitably chosen) and answers ([Chavent and Lechevallier, 2002], [Chavent, 1997]) to a specific binary question of type *'Is $Y_j \leq c$?'*, where c is called the cut value.

To the binary question *'Is $Y_j \leq c$?'*, an object $x \in C$ answers 'yes' or 'no' according to a binary function $q_c : E \rightarrow \{true, false\}$. The bipartition (C_1, C_2) of C induced by the binary question is as follows :

- $C_1 = \{x \in C \mid q_c(x) = true\}$
- $C_2 = \{x \in C \mid q_c(x) = false\}$

Consider the case of interval variables: Let $Y_j(x) = [\alpha, \beta]$, the middle of $[\alpha, \beta]$ is $m_x = \frac{\alpha+\beta}{2}$.

1. The binary question is "Is $m_x \leq c$?".
2. The function q_c is defined by:
    - $q_c(x) = true$ if $m_x \leq c$
    - $q_c(x) = false$ if $m_x > c$
3. The bipartition (C_1, C_2) of C induced by the binary question is :
    - $C_1 = \{x \in C \mid q_c(x) = true\}$
    - $C_2 = \{x \in C \mid q_c(x) = false\}$

3.8 Output Data and Results

After the tree-growing algorithm and the pruning procedure, the final clustering tree is obtained.

The nodes of the tree represent the binary questions selected by the algorithm and the k leaves of the tree define the k-partition. Each cluster is characterized by a rule, i.e, the path in the tree which provided it. The clusters therefore become new symbolic objects defined according to the binary questions leading from the root to the corresponding leaves.

4 Example on the Oils and Fats Data

The above clustering method has been examined with the well-known Ichino's oils dataset. The data set (Table 1) is composed of 8 oils described in terms of four interval variables.

This divisive algorithm yields the 3-cluster partition represented in the tree given in figure 1.

Two binary questions correspond to two binary functions $E \rightarrow \{true, false\}$, given by q_1 = [Spec. Grav.$(x) \leq 0.89075$] and q_2 = [Iod. Val.$(x) \leq 148.5$].

Each cluster corresponds to a symbolic object, e.g. a query assertion:

Sample	Specific Gravity	Freezing point	Iodine Value	Saponification Value
linseed oil	[0.930;0.935]	[-27;-18]	[170;204]	[118;196]
perilla oil	[0.930;0.937]	[-5;-4]	[192;208]	[188;197]
cottonseed oil	[0.916;0.918]	[-6;-1]	[99;113]	[189;198]
sesam oil	[0.920;0.926]	[-6;-4]	[104;116]	[187;193]
camelia oil	[0.916;0.917]	[-21;-15]	[80;82]	[189;193]
olive oil	[0.914;0.919]	[0;6]	[79;90]	[187;196]
beef tallow	[0.860;0.870]	[30;38]	[40;48]	[190;199]
hog fat	[0.858;0.864]	[22;32]	[53;77]	[190;202]

Table 1. *Table of oils and fats*

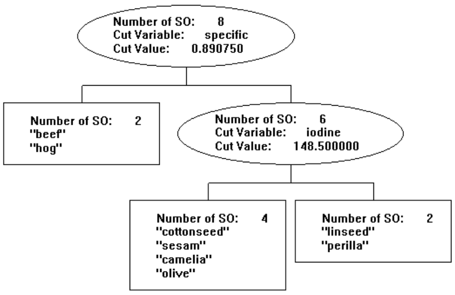

Fig. 1. Clustering tree obtained on the Ichino'oils dataset.

- $C_1 = [\text{Spec. Grav.}(x) \leq 0.89075]$,
- $C_2 = [\text{Spec. Grav.}(x) > 0.89075] \wedge [\text{Iod. Val.}(x) \leq 148.5]$,
- $C_3 = [\text{Spec. Grav.}(x) > 0.89075] \wedge [\text{Iod. Val.}(x) > 148.5]$.

Then, the resulting 3-cluster partition is: $C_1 = \{\text{beef, hog}\}$, $C_2 = \{\text{cottonseed, sesam, camelia, olive}\}$, $C_3 = \{\text{linseed, perilla}\}$.

5 Further Works and Conclusions

Following that work, a new clustering method was conceived. It's also a hierarchical clustering method but a multivariate agglomerative one. The basic idea was to find a merging criterion which would have been dual and complementary to the splitting one. But the strictly dual criterion, consisting in measuring the area sustended by the density between 2 points (or groups of points) and then merging the 2 points (or groups) which are the closest in that sense, presents a risk: gathering 2 points (or groups) which are obviously in different groups.

If a model in dimension d is used, the real criterion (the maximum likelihood criterion) for the divisive method, e.g. between two convex clusters consists in finding the two clusters such that the difference of the hypervolumes sustended by the density between the global convex hulls of the two clusters is the largest. In an agglomerative way, this difference should be the smallest.

Computing hypervolumes causes computational problems. But, if all the sustended areas (on each axis) between the respective coordinates of the two points are small, then the hypervolume in dimension d will be small (This implication is not reversible).

Therefore for each couple of points $x_i = (x_{i1}, \cdots, x_{id})$ and $x_j = (x_{j1}, \cdots, x_{jd})$, the following quantities are computed

$$\mathrm{diss}(x_i, x_j) = \max_{1 \le k \le d} | \int_{x_{ik}}^{x_{jk}} \hat{f}_k(x) dx | \tag{2}$$

where $\hat{f}(\cdot)$ is an estimation of the density function for the variable k:

$$\hat{f}_k(x) = \frac{1}{nh_k} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\dfrac{(x-x_i)^2}{2h_k^2}} .$$

The value of h_k, the smooting parameter is chosen following Silverman ([Silverman, 1986]) as $h_k = 1,06 \cdot \min(\sigma_k, R_k/1,34) \cdot n^{-0,2}$, where σ_k (respectively R_k) is the standard deviation (respectively the interquartil range) of the n values $x_{1k}, \cdots, x_{nk}$.

It can be shown easily that (2) is a dissimilarity measure. For two clusters C_i and C_j, there exist many ways to define $\mathrm{diss}(C_i, C_j)$. For example:

- the single linkage method where $\mathrm{diss}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \mathrm{diss}(x, y)$,
- the complete linkage method where $\mathrm{diss}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \mathrm{diss}(x, y)$.

Based on these definitions, the merging criterion consists in grouping the two objects X and Y (either points or clusters) for which $\mathrm{diss}(X, Y)$ is the smallest.

The method proceeds from the situation where all the points are in separate clusters until they all form a unique cluster. Consecutive merging can be represented by a dendrogram (figure 2).

The resulting algorithm based on these concepts was implemented and seems to be very powerful. The first results obtained are promising. For example, the structure of the dendrogram (figure 2) constructed by the method on the Ichino's Oils dataset is very good if compared with the tree obtained with the first method or those presented in ([Chavent, 1997], page 139).

A new hierarchical divisive monothetic method was first developped. The only hypothese needed was that the observed points are generated by a non-homogeneous Poisson process. The algorithm performed in two steps : splitting and pruning. The splitting rule was deduced from a maximum likelihood criterion; the pruning method was based on the Gap test. An application of this algorithm was presented on a well-known interval dataset. The splitting criterion also gave the idea to develop a new dissimilarity measure for

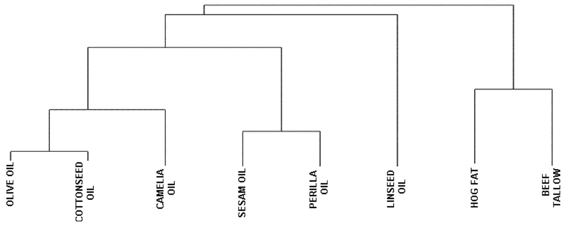

Fig. 2. Dendrogram obtained on the Ichino'oils dataset, complete linkage method.

hierarchical agglomerative clustering. The resulting algorithm was briefly described. Applied on the same dataset, it produced very interesting results. All these ways will be thorough in the future.

References

[Chavent and Lechevallier, 2002]M. Chavent and Y. Lechevallier. Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In K. Jujuga, A. Sokolowski, and H.H. Bock, editors, *Classification, Clustering, and Data Analysis*, pages 53–60, 2002.

[Chavent, 1997]M. Chavent. *Analyse des Données Symboliques: Une méthode divisive de classification.* PhD thesis, Université Paris IX-Dauphine, 1997.

[Kubushishi, 1996]T. Kubushishi. *On some Applications of the Point Process Theory in Cluster Analysis and Pattern Recognition.* PhD thesis, Facultés Universitaires Notre-Dame de la Paix, Namur, 1996.

[Rasson and Kubushishi, 1994]J.-P. Rasson and T. Kubushishi. The gap test: an optimal method for determining the number of natural classes in cluster analysis. In E. Diday, Y. Lechevallier, M. Shader, P. Bertrand, and B. Burtschy, editors, *New approaches in Classification and Data Analysis*, pages 186–193, 1994.

[Silverman, 1981]B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of The Royal Statistic Society, B*, pages 97–99, 1981.

[Silverman, 1986]B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.