Validation in unsupervised symbolic classification

André Hardy

Department of Mathematics
University of Namur,
8 Rempart de la Vierge,
B - 5000 Namur, Belgium
(e-mail: andre.hardy@fundp.ac.be)

Abstract. One important topic in unsupervised classification is the objective assessment of the validity of the clusters found by a clustering algorithm. The determination of the "best" number of "natural" clusters has often been presented as the central problem of cluster validation. In this paper we investigate the problem of the determination of the number of clusters for symbolic objects described by interval, multi-valued and modal variables. We consider five classical methods for the determination of the number of clusters and two hypothesis tests based on the Poisson point process, and we show how these methods can be extended to symbolic data. We present applications of these symbolic methods to real data sets.
Keywords: Validation, Number of clusters, Poisson process, Symbolic data.

1 Introduction

The aim of cluster analysis is to identify a structure within a data set. When hierarchical algorithms are used, an important problem is then to choose one solution in the nested sequence of partitions of the hierarchy. On the other hand, optimization methods for cluster analysis usually require the a priori specification of the number of classes. So most clustering procedures demand the user to fix the number of clusters, or to determine it in the final solution.

Some studies have been proposed to compare procedures for the determination of the number of clusters. For example, Milligan and Cooper [Milligan and Cooper, 1985] conducted a Monte Carlo evaluation of thirty indices for determining the number of clusters. [Hardy, 1996] compared three methods based on the Hypervolumes clustering criterion with four other methods available in the Clustan software. [Gordon, 1996] modified the five stopping rules whose performance was best in the Milligan and Cooper study in order to detect when several different, widely-separated values of c, the number of clusters, would be appropriate, that is, when a structure is detectable at several different scales.

In this paper we consider two hypothesis tests for the number of clusters based on the Hypervolumes clustering criterion: the Hypervolumes test and the Gap test. These statistical methods are based on the assumption that

the points we observe are generated by a homogeneous Poisson process [Karr, 1991] in k disjoint convex sets. We consider also the five best stopping rules for the number of clusters analysed by [Milligan and Cooper, 1985]. We show how these methods can be extended in order to be applied to symbolic objects described by interval, multi-valued and modal variables [Bock and Diday, 2000].

2 The clustering problem

The clustering problem we are interested in is the following.
$E = \{x_1,\ x_2,\ ...,\ x_n\}$ is a set of objects. On each of the n objects we measure the value of p variables $Y_1,\ Y_2,\ ...,\ Y_p$. The objective is to find a "natural" partition $P = \{C_1,\ C_2,\ ...,\ C_k\}$ of the set E into k clusters.

3 Statistical models based on the Poisson process

3.1 The Hypervolumes clustering method

The Hypervolumes clustering method [Hardy and Rasson, 1982] assumes that the n p-dimensional observation points $x_1,\ x_2,\ ...,\ x_n$ are generated by a homogeneous Poisson process in a set D included in the Euclidean space R^p. The set D is supposed to be the union of k disjoint convex domains $D_1,\ D_2,\ ...,\ D_k$. We denote by $C_i \subset \{x_1,\ x_2,\ ..,\ x_n\}$ the subset of the points belonging to $D_i\ \ (1 \leq i \leq k)$. The Hypervolumes clustering criterion is deduced from that statistical model, using maximum likelihood estimation. It is defined by

$$W(P,k) := \sum_{i\ =\ 1}^{k} m(H(C_i))$$

where $H(C_i)$ is the convex hull of the points belonging to C_i and $m(H(C_i))$ is the multidimensional Lebesgue measure of that convex hull. That clustering criterion has to be minimised over the set of all the partitions of the observed sample into k clusters.

3.2 The generalised Hypervolumes clustering method

The generalised Hypervolumes clustering method [Rasson and Granville, 1996] assumes that the n p-dimensional points $x_1,\ x_2,\ ...,\ x_n$ are generated by a nonhomogeneous Poisson process in a set D. D is the union of k disjoint convex domains $D_1,\ D_2,\ ...,\ D_k$. The generalised Hypervolumes clustering criterion is deduced from that statistical model, using maximum likelihood estimation. It is defined by

$$W(P,k) := \sum_{i\,=\,1}^{k} \int_{H(C_i)} q(x)m(dx)$$

where q(x) is the intensity of the nonhomogeneous Poisson process.

4 Statistical tests for the number of clusters based on the Poisson point process

4.1 The Hypervolumes test

The statistical model based on the Poisson process allows us to define a likelihood ratio test for the number of clusters [Hardy, 1996]. Let us denote by $C = \{C_1, C_2, ..., C_\ell\}$ the optimal partition of the sample into ℓ clusters and $B = \{B_1, B_2, ..., B_{\ell-1}\}$ the optimal partition into $\ell - 1$ clusters. We test the hypothesis H_0: $t = \ell$ against the alternative H_A: $t = \ell - 1$, where t represents the number of "natural" clusters ($\ell \geq 2$). The test statistics is defined by

$$S(x) := \frac{W(P,\ell)}{W(P,\ell-1)}.$$

Unfortunately the sampling distribution of the statistics S is not known. But $S(x)$ belongs to $[0,1[$. Consequently, for practical purposes, we can use the following decision rule: reject H_0 if S is close to 1. We apply the test in a sequential way: if ℓ_0 is the smallest value of $\ell \geq 2$ for which we reject H_0, we choose $\ell_0 - 1$ as the best number of "natural" clusters.

4.2 The Gap test

The Gap test [Kubushishi, 1996] [Rasson and Kubushishi, 1994] is based on the same statistical model (homogeneous Poisson process). We test H_0 : the $n = n_1 + n_2$ observed points are a realisation of a Poisson process in D against H_A: n_1 points are a realisation of a homogeneous Poisson process in D_1 and n_2 points in D_2 where $D_1 \cap D_2 = \emptyset$. The sets D, D_1, D_2 are unknown. Let us denote by C (respectively C_1 , C_2) the set of points belonging to D (respectively D_1, D_2). The test statistics is given by

$$Q(x) = \left(1 - \frac{m(\triangle)}{m(H(C))}\right)^n$$

where $\triangle = H(C) \setminus (H(C_1) \cup H(C_2))$ is the "gap space" between the clusters. The test statistics is the Lebesgue measure of the gap space between the clusters.

The decision rule is the following [Kubushishi, 1996]. We reject H_0, at level α, if (asymptotic distribution)

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

5 Other methods for the determination of the number of clusters

We consider the best methods from the [Milligan and Cooper, 1985] study: the Calinski and Harabasz index [Calinski and Harabasz, 1974], the Duda and Hart rule [Duda and Hart, 1973], the C index [Hubert and Levin, 1976], the γ index [Baker and Hubert, 1975] and the Beale test [Beale, 1969]. The Calinski and Harabasz, Duda and Hart, and Beale indices use various forms of sum of squares within and between clusters. The Duda and Hart rule and the Beale test are statistical hypothesis tests on the number of clusters.

6 Symbolic data analysis

Symbolic data analysis [Bock and Diday, 2000] is concerned with the extension of classical data analysis and statistical methods to complex data called symbolic data. We will consider sets of objects described by interval, multi-valued and modal variables.

6.1 Interval, multi-valued and modal variables

This paper is based on the following definitions [Bock and Diday, 2000].

A variable Y is termed set-valued with the domain $\mathcal{Y}$, if for all $x_k \in E$,

$$\begin{aligned} Y : E &\rightarrow \mathcal{B} \\ x_k &\longmapsto Y(x_k) \end{aligned}$$

where $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\}$.

A set-valued variable is called multi-valued if its values $Y(x_k)$ are all finite subsets of the underlying domain $\mathcal{Y}$; so $|Y(x_k)| < \infty$, for all elements $x_k \in E$.

A set-valued variable Y is called **categorical multi-valued** if it has a finite range $\mathcal{Y}$ of categories and **quantitative multi-valued** if the values $Y(x_k)$ are finite sets of real numbers.

A modal variable Y on a set $E = \{x_1, ..., x_n\}$ with domain $\mathcal{Y}$ is a mapping

$$Y(x_k) = (U(x_k), \pi_k), \quad \text{for all } \mathrm{x_k} \in \mathrm{E}$$

where π_k is, for example, a frequency distribution on the domain $\mathcal{Y}$ of possible observation values and $U(x_k) \subseteq \mathcal{Y}$ is the support of π_k in the domain $\mathcal{Y}$.

Y is an interval variable if for all $x_k \in E$,

$$Y : E \rightarrow \mathcal{B} : x_k \mapsto Y(x_k) = [\alpha_k, \beta_k] \subset \mathcal{R}$$

where $\mathcal{B}$ is the set of all closed bounded interval of $\mathcal{R}$.

7 Symbolic clustering procedures

In order to generate partitions, we consider several symbolic clustering methods. SHICLUST [Hardy, 2004] is a module containing the symbolic extensions of four well-known hierarchical clustering methods: the single link, complete link, centroid and Ward methods. SCLUST [Verde *et al.*, 2000] is a partitioning clustering method; it is a symbolic extension of the well-known Dynamic clouds clustering method [Celeux *et al.*, 1989]. DIV [Chavent, 1997] is a symbolic hierarchic monothetic divisive clustering procedure based on the extension of the within class sum-of-squares criterion. SCLASS [Pirçon, 2004] is a symbolic hierarchic monothetic divisive method based on the generalised Hypervolumes clustering criterion. The first part of HIPYR [Brito, 2000] is also a module including four hierarchical symbolic clustering methods.

8 Determination of the number of clusters

8.1 Methods based on a dissimilarity matrix

In order to apply the five best methods for the determination of the number of clusters from the Milligan and Cooper [Milligan and Cooper, 1985] study, it is necessary to define a dissimilarity matrix for symbolic objects described by interval, multi-valued and modal variables.

Let us consider the case of n objects described by p interval variables

$$Y_j : E \to \mathcal{B}_j : x_i \mapsto Y_j(x_i) = x_{ij} = [\alpha_{ij}, \beta_{ij}].$$

We first define p dissimilarity indices δ_1, ..., δ_p on the sets $\mathcal{B}_j$. Let $x_{uj} = [\alpha_{uj}, \beta_{uj}]$ and $x_{vj} = [\alpha_{vj}, \beta_{vj}]$. We consider three distances for interval variables

The Haussdorff distance:

$$\delta_j\,(x_{uj}, x_{vj}) = \max\{\,|\,\alpha_{uj} - \alpha_{vj}\,|, |\,\beta_{uj} - \beta_{vj}\,|\,\}$$

The L_1 distance:

$$\delta_j\,(x_{uj}, x_{vj}) = \,|\,\alpha_{uj} - \alpha_{vj}\,| + |\,\beta_{uj} - \beta_{vj}\,|$$

The L_2 distance:

$$\delta_j\,(x_{uj}, x_{vj}) = \,(\alpha_{uj} - \alpha_{vj})^2 + (\beta_{uj} - \beta_{vj})^2.$$

We combine the p dissimilarity indices δ_1, ..., δ_p in order to obtain a global dissimilarity measure on E.

$$d : E \times E \longrightarrow R^+ : (x_u, x_v) \longmapsto d(x_u, x_v) = \left(\sum_{j=1}^{p} \delta_j^2(x_{uj}, x_{vj})\right)^{1/2}.$$

For multi-valued and modal variables, we define suitable L_1 and L_2 distances and we use also the de Carvalho distance [Hardy, 2004].

Concerning the four hierarchical procedures included in SHICLUST, the five indices for the determination of the number of clusters are computed at each level of the hierarchies. For SCLUST, we select the best partition into ℓ clusters, for each value of ℓ ($\ell = 1, \cdots, K$) (K is a reasonably large integer fixed by the user) and we compute the indices available for nonhierarchical classification. The analysis of theses indices should provide the "best" number of clusters.

8.2 Tests based on the Poisson point processes

The Hypervolumes test and the Gap test are now available only for classical quantitative and for interval data. These tests are not based on the existence of a dissimilarity matrix, but only on the positions of the points. For interval data, we use the following modelisation. We represent an interval by two numbers: its middle and its lenght. So each interval can be represented by a point in a two-dimensional space, and an object by a point in a $2p$-dimensional space. We first determine the best number of clusters for each interval variable. A synthesis is then made in order to precise the actual structure of the set of symbolic data.

9 Examples

9.1 Merovingian buckles - VI-VIII a.c. Century

The set of symbolic data is constituted by 58 buckles described by six symbolic multi-valued variables. These variables and the corresponding categories are presented in Table 1. The complete data set is available at http://www-rocq.inria.fr/sodas/WP6/data/data.html.

Variables	Categories
Fixation	iron nail; bronze bump; none
Damascening	bichromate; predominant veneer; dominant inlaid; silver monochrome
Contours	undulations; repeating motives; geometric frieze
Background	silver plate, hatching; geometric frame
Inlaying	filiform; hatching banner; dotted banner; wide ribbon
Plate	arabesque; large size; squared back; animal pictures; plait; circular

Table 1. Merovingian buckles: six categorical multi-valued variables

The 58 buckles have been examined by archeologists. They identified two natural clusters. SCLUST and the four hierarchical clustering methods

included in SHICLUST have been applied to that data set in order to generate partitions. The true structure has been detected by most of the stopping rules.

9.2 e-Fashion stores

That data set describes the sales in a group of stores (items of clothing and accessories), belonging to six different countries. These sales concern the years 1999, 2000 and 2001. The 13 objects are the stores (Paris 6th, Lyon, Rome, Barcelona, Toulouse, Aix-Marseille, Madrid, Berlin, Milan, Brussels, Paris 15th, Paris 8th, London). Eight modal variables are recorded on each of the 13 objects, describing the items sold in these stores. For example, the variable "family product" has 13 categories (dress, sweater, T-shirt, ...). The proportion of sales in each store is associated with all these categories. The variable "month" describes the proportion of sales for each month of the year.

9.3 Fats and oils

The data set contains eight fats and oils described by four quantitative features of interval type: specific gravity, freezing point, iodine value and saponification [Ichino and Yaguchi, 1994] [Gowda and Diday, 1994].

References

[Baker and Hubert, 1975]F.B. Baker and L.J. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, pages 31–38, 1975.

[Beale, 1969]E.M.L. Beale. Euclidean cluster analysis. *Bulletin of the International Statistical Institute*, pages 92–94, 1969.

[Bock and Diday, 2000]H.-H. Bock and E. Diday. *Analysis of Symbolic Data.* Springer Verlag, 2000.

[Brito, 2000]P. Brito. Hierarchical and pyramidal clustering with complete symbolic objects. In H.H. Bock and E. Diday, editors, *Analysis of Symbolic Data Analysis*, pages 312–323, 2000.

[Calinski and Harabasz, 1974]T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, pages 1–27, 1974.

[Celeux *et al.*, 1989]G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralanbondrainy. *Classification automatique des données.* Bordas, 1989.

[Chavent, 1997]M. Chavent. *Analyse des données symboliques - Une méthode divisive de classification.* Thèse. Université Paris Dauphine, 1997.

[Duda and Hart, 1973]R.O. Duda and P.E. Hart. *Classification and Scene Analysis.* Wiley, 1973.

[Gordon, 1996]A.D. Gordon. How many clusters? an investigation of five procedures for detecting nested cluster structure. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock, and Y. Baba, editors, *Data Science, Classification, and Related Methods*, pages 109–116, 1996.

[Gowda and Diday, 1994]K.C. Gowda and E. Diday. Symbolic clustering algorithms using similarity and dissimilarity measures. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, *Data Science, Classification, and Related Methods*, pages 414–422, 1994.

[Hardy and Rasson, 1982]A. Hardy and J.P. Rasson. Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données*, pages 41–56, 1982.

[Hardy, 1996]A. Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, pages 83–96, 1996.

[Hardy, 2004]A. Hardy. Les méthodes de classification et de détermination du nombre de classes: du classique au symbolique. In M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, and B. Patouille, editors, *Comptes rendus des 11èmes Rencontres de la Société Francophone de Classification*, pages 48–55, 2004.

[Hubert and Levin, 1976]L.J. Hubert and J.R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, pages 1073–1080, 1976.

[Ichino and Yaguchi, 1994]M. Ichino and H. Yaguchi. Generalized minkowsky metrics for mixed feature type data analysis. *IEEE Transactions System, Man and Cybernetics*, pages 698–708, 1994.

[Karr, 1991]A.F. Karr. *Point Processes and their Statistical Inference*. Marcel Dekker, 1991.

[Kubushishi, 1996]T. Kubushishi. *On some Applications of the Point Process Theory in Cluster Analysis and Pattern Recognition*. PhD Thesis, University of Namur, Belgium, 1996.

[Milligan and Cooper, 1985]G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, pages 159–159, 1985.

[Pirçon, 2004]J.Y. Pirçon. *Le clustering et les processus de Poisson pour de nouvelles méthodes monothétiques*. PhD. thesis, University of Namur, Belgium, 2004.

[Rasson and Granville, 1996]J.P. Rasson and V. Granville. Geometrical tools in classification. *Computational Statistics and Data Analysis*, pages 105–123, 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. Schader, P. Bertrand, and Butschy B., editors, *New Approaches in Classification and Data Analysis*, pages 186–193, 1994.

[Verde *et al.*, 2000]R. Verde, F. de Carvalho, and Y. Lechevallier. A dynamical clustering algorithm for multi-nominal data. In H. Kiers, J.P. Rasson, P. Groenen, and M. Schader, editors, *Data Analysis, Classification, and Related Methods*, pages 387–393, 2000.