Symbolic clustering interpretation and visualization

Rosanna Verde[1], Yves Lechevallier[2], and Marie Chavent[3]

[1] Dip. di Strategie Aziend. e Metod. Quantitative, Seconda Universitá di Napoli, Piazza Umberto I, 81043 Capua, Italy
[2] INRIA Rocquencourt B.P. 105 - 78153 Le Chesnay Cedex, France
[3] MAB-Mathématiques Appliquées de Bordeaux, Université Bordeaux1, 351 cours de la libération, 33405 Talence cedex, France

Abstract. In this paper we propose some new tools for a symbolic clustering interpretation. In the framework of Symbolic Data Analysis the algorithms to cluster a set E of symbolic data (Verde, De Carvalho, Lechevallier, 2000) are based on different types of assignment functions and different kinds of prototypes which represent the classes. The classical interpretative aids are usually based on inertia criterion. In our approach we propose to generalize this criterion to symbolic data, where the barycenter is replaced by the prototype of the clusters. The several indexes measure the improvement of a partition in k clusters with respect to the global cluster E. This comparison allows to evaluate the contribution and the discrimination of the symbolic objects and variables to the partition. Further suitable measures are considered to value the homogeneity of the clusters of the partition

1 Introduction

In (Verde, De Carvalho, Lechevallier, 2000) and (De Carvalho, Lechevallier, Verde, 2001) we have proposed a generalization of the dynamical clustering algorithm in order to cluster a set of symbolic data (Bock and Diday, 2000). According to the classical dynamical clustering algorithm, the symbolic clustering algorithm is based on a criterion of the best fitting between the obtained partition at each step and the representation of the clusters. Let's D be the representation space of a set E of symbolic objects described by a set of p multi-valued variables $Y_1, \ldots, Y_p$ which can be of different type: intervals, multi-categorical or modal (Bock and Diday, 2000, pages 42–48). The description of every object s of E is a vector x_s containing the values $(x_s^1, \ldots, x_s^j, \ldots, x_s^p)$ observed on the variables $Y_1, \ldots, Y_j, \ldots, Y_p$. In such context, the clusters of symbolic data can be represented by means of a model of prototype which is consistent with the assignment function chosen to assign the objects of E to the different clusters. Hereafter we proposed some indexes as aids to the interpretation of the clusters obtained by the algorithm. They are based on the notion of prototype associate to the representation of each cluster. Whereas the prototype G of a cluster C belongs to the same space of representation D of the set of symbolic data to be classified, it is defined

as follows:

$$G = \arg\min_{x \in D} \sum_{s \in C} \Psi(x_s, x) \tag{1}$$

where: Ψ is a dissimilarity function defined on D. Different kind of Ψ are considered according to the type of multi-valued variable, for instance *Hausdorff's distance* (Chavent, 1997) for intervals ; norm-2 distance, Φ^2, two component distance, based on distributions, for multi-nominal variables (De Carvalho, 1994) and (Bock and Diday, 2000, pages 153–165).

2 Dynamical partitioning algorithm

The proposed clustering algorithm, according to the dynamical clustering algorithm, looks for simultaneously a *partition* P of E in k classes and a *vector* L of k prototypes $(g_1, \ldots, g_i, \ldots, g_k)$ associated to the classes $(C_1, \ldots, C_i, \ldots, C_k)$ of the partition P that minimizes a criterion Δ:

$$\Delta(P^*, L^*) = \min \{\Delta(P, L) / \ P \in P_k, \ L \in D^k\} \tag{2}$$

with P_k the *set of partitions* of E in k classes no-empty. Such criterion Δ expresses the fitting between the partition P and the vector L of the k prototypes. That is defined as the sum of the distances between all the objects s of E and the prototypes g_i of the nearest class C_i:

$$\Delta(P, L) = \sum_{i=1}^{k} \sum_{s \in C_i} \Psi(x_s, g_i) \quad C_i \in P, \ g_i \in D \tag{3}$$

The algorithm alternates a step of **representation** to a step of **assignment**, as follows :

a) **initialization step** Let's be chosen a random partition $P = (C_1, \ldots, C_i, \ldots, C_k)$ or alternatively a set k of random prototypes $(g_1, \ldots, g_i, \ldots, g_k)$ of E (in this last case a step of assignment will be run)

b) **representation step** `FOR i=1 TO k DO` the prototype g_i of D is obtained by minimizing the criterion: $\sum_{s \in C_i} \Psi(x_s, g_i)$

c) **assignment step**
   $test \leftarrow 0$
   `FOR s=1 TO n DO` (having indicated with m the class of assignment of the object s)
      Looks for the new class C_l to assign s with $l = \arg\min_{i=1,\ldots,k} \Psi(x_s, g_i)$
      `IF` $l \neq m$ `THEN` $test \leftarrow 1$, $C_m \leftarrow C_m - \{s\}$ and $C_l \leftarrow C_l \cup \{s\}$

d) **convergence step** `IF` $test = 0$ `THEN` stop `OTHERWISE` go to b)

The convergence of the criterion D to a stationary point is obtained under the following conditions:

- Uniqueness of the class of assignment of each object of E
- existence and uniqueness of the prototype g_C minimizing the criterion $\sum_{s \in C} \Psi(x_s, g)$ for all the clusters C of E.

3 Interpretation of the clusters

In (Celeux *et al.* 1989) the indexes proposed to describe a partition are based on the decomposition of the total inertia in the within and between inertia. These indexes furnish a suitable aid to the interpretation of the clusters obtained by partition methods like "*Nuées Dynamiques*" method. In our approach, the barycenter of the clusters is replaced for a prototype and the *total inertia* is generalized by the homogeneity of the partition in one class and the *within inertia* by the homogeneity criterion of the description of each cluster; while the *between inertia* is defined as the difference between the homogeneity measure of the partition in one cluster and the partition in k clusters.

The **quality measures** of a partition and of their clusters, hereafter proposed, can be interpreted as the gain between the null hypothesis "No structure = Partition into one cluster" and the solution carried out a classification algorithm into K clusters optimizing the fitting criterion between a partition P and the corresponding vector of prototypes of D. Such criterion can be written as follows:

$$\Delta(P,L) = \sum_{i=1}^{k} \sum_{s \in C_i} \Psi(x_s, g_i) = \sum_{i=1}^{k} \sum_{s \in C_i} \sum_{j=1}^{p} \Psi_j(x_s^j, g_i^j)\ C_i \in P,\ g_i \in D = \prod_{j=1}^{p} D_j$$

where (Ψ_j, D_j) is a metric space defined on the representation space of the variable j. Let's denote : $S(C_i, x) = \sum_{j=1}^{p} S_j(C_i, x) = \sum_{j=1}^{p} \sum_{s \in C_i} \Psi_j(x_s^j, g_i^j)$the variability of the class C_i with respect to x belonging to the space of description.

For construction we have : $\Delta(P, g) = \sum_{i=1}^{k} S(C_i, g_i) \leq \sum_{i=1}^{k} S(C_i, g_E) = S(E)$ as the difference between the two values representing the gain to replace the prototype g_E for the k prototypes $g = (g_1, \ldots, g_k)$.

The values of the indicators that we propose ranging between 0 and 1 with 0 in absence of structure.

The *quality of the partition*, globally is measured by the normalized deviation of the gain obtained replacing the class unique for the partition : $Q(P) = 1 - \frac{\Delta(P,g)}{S(E)}$. In this case, if the value is equal to 1 it means that all the objects belonging to the clusters are coincident with the corresponding prototypes. If the value is equal to 0 the prototypes are all equal. We propose this general quality index for every kind of class prototype representing a cluster in the symbolic partition algorithm.

The *quality of the partition* can be decomposed on the variable j by :

$$Q_j(P) = 1 - \frac{\Delta(P, g_j)}{S_j(E)} = 1 - \frac{\Delta(\sum_{i=1}^{k} S_j(C_i, g_i^j)}{\sum_{i=1}^{k} S_j(C_i, g_E^j)}.$$

This index measure the participation of the variable j to the partition in this meaning as more this value is great more the variable plays an important

role in the building of the partition because it contributes to the reduction of the variability within the classes.

Moreover, a *quality measure* for each class is given by:

$$Q(C_i) = 1 - \frac{S(C_i, g_i)}{S(C_i, g_E)}.$$

If the value of $Q(C_i)$ is near to 1 then the elements of the classes are very different from the general prototype g_E.

In other way a **contribution measure** of the class or the variable is giving the ratio between the homogeneity criterion computed on this variable or the class and the global one. The sum of all the contributions is equal to 1.

We can measure the *contribution of the class* C_i to the global variability by:$K(C_i/p) = \frac{S(C_i,g_i)}{\Delta(P,g)}$. Thus, it is possible to make a comparison with respect to the ratio $\frac{S(C_i,g_E)}{S(E)}$, that is the contribution of the class C_i having taken as prototype g_E (the prototype of the partition in one class). Whereas this value is more than 1 then the prototype g_i of the class C_i is very far from g_E.

The *contribution of the variable* j to the partition P allows to evaluate the role of this variable to the building of such partition, that is measured by: $K_j(P) = \frac{\sum_{i=1}^{k} S_j(C_i,g_i^j)}{\Delta(P,g)}$. Similarly than above, an evaluation of the contribution of j can be made on the basis of the following ratio: $\frac{\sum_{i=1}^{k} S(C_i,g_E)}{S(E)}$. If this value is larger than 1 then the variable j furnishes a strength contribution to the reduction of the within variability of the classes of the partition.

In conclusion, we furnish a measure to evaluate the *contribution of each object* to the variability of the belonging class. It is given by the following index: $K(s) = \frac{\Psi(x_s,g_i)}{S(C_i)} \cdot \frac{1}{n_i - 1}$. More this value is near to 0 more the representation of this object is similar to the prototype of the class.

4 Interpretation of the partition of 60 meteorological stations in China

The proposed aids to the interpretation of a partition of symbolic data has been realized on a set of data by Long-Term Instrumental Climatic Data Base of the People's Republic of China (http://dss.ucar.edu/datasets/ds578,5/data). This set of data contains the monthly temperatures observed in 60 meteorological stations of China. According a natural representation of the temperatures, they are coded in a table as the interval of the minima and maxima for each month. For our example we have considered the temperatures of the year 1988 and we have built a table of dimension 60 rows and 12 columns, corresponding to the number of stations and to the number of months of the year. The different quality and contribution indexes have been computed on

an example of partition in 5 classes obtained by a dynamical partitioning algorithm () on symbolic data described by interval variables. For instance, the station "ChangSha" is described by the 12 intervals of the monthly temperatures:

```
  [January = [2.7:7.4]]      ^[February = [3.1:7.7]]
^[March = [6.5:12.6]]        ^[April = [12.9:22.9]]
^[May = [19.2:26.8]]         ^[June = [21.9:31]]
^[July = [25.7:34.8]]        ^[August = [24.4:32]]
^[September = [20:27]]       ^[October = [15.3:22.8]]
^[November = [7.6:19.6]]     ^[December = [4.1:13.3]]
```

In the table 1 are collected the descriptions of the 60 meteorological stations:

Meteorological stations	January	February	...	December
AnQing	[1.8 : 7.1]	[5.2 : 11.2]	...	[4.3 : 11.8]
BaoDing	[-7.1 : 1.7]	[-5.3 : 4.8]	...	[-3.9 : 5.2]
BeiJing	[-7.2 :2.1]	[-5.3 : 4.8]	...	[-4.4 : 4.7]
BoKeTu	[-23.4 : -15.5]	[-24 : -14]	...	[-21.1 : -13.1]
ChangChun	[-16.9 : -6.7]	[-17.6 : -6.8]	...	[-15.9 : -7.2]
ChangSha	[2.7:7.4]	[3.1:7.7]	...	[4.1:13.3]
ZhiJiang	[2.7 : 8.2]	[2.7 : 8.7]	...	[5.1 : 13.3]

Table 1. Minima and maxima monthly temperatures recorded by the 60 meteorological stations

Fixed the number of classes to 5, the algorithm is reiterated 50 times and the best solution is found for the minimum value of the criterion equal to: D=3848.97. It is worth to noting that the obtain partition of the 60 elements follows the geographical contiguity of the stations.

Fig. 1. Visualization on the Map of the China of the 5 Classes of the partition of the 60 meteorological stations

According to the kind of representation of the classes by intervals proposed in the partitioning algorithm on interval data, the prototype of each class is the interval which minimizes the Hausdroff distances from all the elements belonging to the class.

In the table 2 we have indicated the values of the different indices of quality and contribution proposed in the present paper.

Variable	Quality	Contribution with P	Contribution with E
January	**69.50**	**13.76**	12.74
February	**66.18**	**12.63**	12.28
March	**64.52**	9.30	9.27
April	**64.36**	6.74	6.73
May	61.68	6.15	6.42
June	53.36	4.56	5.50
July	46.31	4.05	5.63
August	47.19	3.73	5.08
September	**61.10**	6.05	6.37
October	**70.41**	8.97	8.19
November	**70.63**	**10.79**	9.83
December	**71.33**	**13.26**	11.96

Table 2. Quality and contribution measures (times 100) of the intervals of temperatures observed in the 12 months to the partition of the stations in 5 classes.

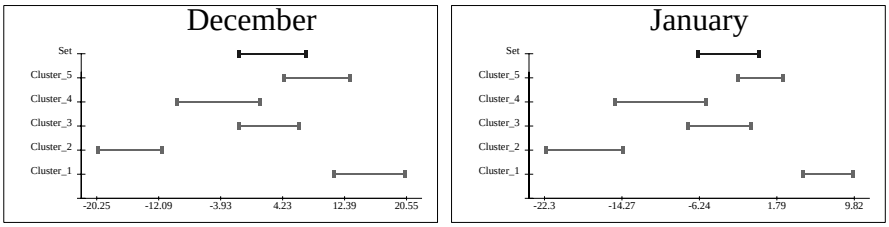

Fig. 2.

We can observe that the wintering months are more discriminant of the cluster (high value of the quality index) than the summering ones. In the figure 2 the prototypes of the classes computed on the interval values of January and December are much more separated than the ones in figure 3 corresponding to the prototype of temperatures of June and September.

5 Conclusion

The proposed quality measures furnish an interpretation of a partition of multi-valued data. However, each class can be modelling as a symbolic object:

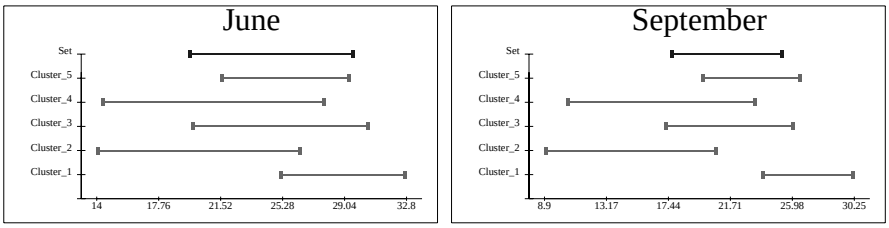

Fig. 3.

its description is given by the prototype associated to the class and its extent can be obtained on the basis of the assignment function considered in the algorithm, as following:

$$Ext(C_i/E) = \{s \in E/\Psi(x_s, g_i) < \Psi(x_s, g_m)\ \forall m \neq i\}$$

For construction all the elements belonging to the class Ci are elements of the extent of this class and any element of the others classes take part to this extent, therefore the obtained partition by the proposed dynamical algorithm furnish in output a complete classification.

References

BOCK, H. H., DIDAY, E. (eds.), (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

CELEUX, G., DIDAY, E., GOVAERT, G., LECHEVALLIER, Y., RALAMBONDRAINY, H. (1989): *Classification Automatique des Données*. Bordas, Paris.

CHAVENT, M. (1997): *Analyse des Données Symboliques. Une méthode divisive de classification*. Thèse de l'Université de PARIS-IX Dauphine.

DE CARVALHO, F.A.T. (1994): Proximity coefficients between Boolean symbolic objects. In: E. Diday *et al* (Eds.): *New Approaches in Classification and Data Analysis*. Springer Verlag, Heidelberg, 387–394.

DE CARVALHO, F.A.T., SOUZA, R. M. C. (1998): Statistical proximity functions of Boolean symbolic objects based on histograms. In: A. Rizzi, M. Vichi and H.-H. Bock, (Eds.): *Advances in Data Science and Classification*. Springer-Verlag, Heidelberg, 391–396

DE CARVALHO, F.A.T, VERDE, R., LECHEVALLIER, Y. (2001): Deux nouvelles méthodes de classification automatique d'ensembles d'objets symboliques décrits par des variables intervalles. *SFC'2001*, Guadeloupe.

DIDAY, E. (1971): La méthode des Nuées dynamiques *Revue de Statistique Appliquée, Vol 19-2, 19–34*.

VERDE, R., DE CARVALHO, F.A.T., LECHEVALLIER, Y. (2000) : A Dynamical Clustering Algorithm for Multi-Nominal Data. In : H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen and M. Schader (Eds.): *Data Analysis, Classification, and Related Methods*, Springer-Verlag, Heidelberg, 387-394.

VERDE, R., LECHEVALLIER, Y., DE CARVALHO, F.A.T. (2001): A dynamical clustering algorithm for symbolic data. *Tutorial Symbolic Data Analysis, GfKl Conference*, Munich.