Crossed Clustering method on Symbolic Data tables

Rosanna Verde[1] and Yves Lechevallier[2]

[1] Seconda Università di Napoli - P.zza Umberto I, 81043 Capua - Italy
Email: rosanna.verde@unina2.it
[2] INRIA - Rocquencourt, Domaine de Voluceau - Rocquencourt - B. P. 105
78153 Le Chesnay Cedex - France, Email: Yves.Lechevallier@inria.fr

Abstract. In this paper we propose a crossed clustering algorithm in order to partition a set of symbolic objects in a predefined number of classes and to determine, in the same time, a structure (taxonomy) on the categories of the object descriptors. The procedure is an extension of the classical simultaneous clustering algorithms proposed on binary and contingency tables. Our approach is based on a dynamical clustering algorithm on symbolic objects. The criterion optimized is the chi-square distance computed between the description of the objects by modal variables (distributions) and the prototypes of the classes, represented by marginal profiles of the objects set partitions. The convergence of the algorithm is guaranteed at the best partitions of the symbolic objects in r classes and of the categories of the symbolic descriptors in c groups, respectively. An application on the web log data from INRIA web server allows to validate the proposed procedure and to suggest it as an useful tool in the Web Usage Mining framework.

Keywords: Dynamic Clustering, Symbolic Data Analysis, Web Mining.

1 Introduction

A generalization of the clustering dynamic algorithms (Diday, 1971, Celeux et al. 1989) has been proposed (Chavent, 1997; Chavent et al., 2003; De Carvalho et al., 2001; Verde et al. 2000) in the Symbolic Data Analysis (SDA) in order to partition a set E of symbolic objects (hereafter denoted SO's) in a predefined number k of homogeneous classes. As in the classical clustering algorithm the criterion optimized is based on the best fitting between classes of objects and their representation.
According to nature of the symbolic data, the first phase of the proposed algorithms consists in choosing a suitable representation of the classes of objects. In the context of SDA, we propose to represent the classes by prototypes which summarize the whole information of the SO's belonging to each of them. Each prototype is even modelling as a SO described by multi-values variables: intervals, multi-categories, with associated distributions. Furthermore, related to the representation of the clusters, every element of E is assigned to a class according to its proximity to the prototype.

In the context of SDA, several distances and dissimilarity functions have been proposed as assignment. In particular, whereas either the SO's to cluster and the prototypes are described by interval variables, the most suitable distance, defined between intervals, is given by the Hausdroff distance (Chavent et al., 2003); while, if they are describe by modal variables, the dissimilarity measure can be chosen as a classical distance between distributions (e.g. chi-squared) or, as one of the context dependent measures (De Carvalho et al., 2001). Moreover, if the SO's descriptors are of different nature (intervals, multi-categories, distributions) they can be retrieved in modal ones.
The convergence of the algorithm to a stationary value of the criterion is guaranteed by the best fitting between the type representation of the classes and the properties of the allocation function. Different algorithms, even referred the same scheme, has been proposed according to the type of SO's descriptors and to the choice of the allocation function.
The generalized dynamic algorithm on symbolic objects has been proposed in different contexts of analysis, for example: to cluster archaeological data, described by multi-categorical variables; to look for typologies of waves, characterized by intervals values; to analyze similarities between the different shapes of micro-organism, described by both multi-categorical and intervals; to compare social-economics characteristics in different geographical areas with respect to the distributions of some variables (e.g.: economics activities; income distributions; worked hours; etc).
The main advantage to use a symbolic cluster algorithm is to get a tool for comparing and clustering aggregated and structured data. In this perspective, we generalize a crossed clustering algorithm (Govaert, 1977, 1995) to symbolic data. Such algorithm performs iteratively a cluster on the rows and on the variables of a symbolic data table.

2 General scheme of dynamical clustering algorithm

Let E a set of symbolic objects s described by p symbolic variables y_j $(j = 1, \ldots, p)$and a weight $\mu_s > 0$. According to the standard dynamic clustering algorithm (Celeux et al.,1989) we look for the partition $P \in P_k$ of E in k classes, among all the possible partitions P_k, and the vector $L \in L_k$ of k prototypes $(g_1, \ldots, g_i, \ldots, g_k)$ representing the classes in P, such that, a criterion Δ of fitting between L and P is minimized:

$$\Delta(P^*, L^*) = Min\{\Delta(P, L) \mid P \in P_k, L \in L_k\}$$

This criterion is defined as the weighted sum of the dissimilarities $\delta(x_s, g_i)$ between the descriptions of the SO's of E collected in a vector x_s and the prototype g_i representing the cluster C_i, for all the clusters C_i $(i = 1, ..., k)$ and for all the objects s of C_i:

$$\Delta(P, L) = \sum_{i=1}^{k} \sum_{s \in C_i} \mu_s . \delta^2(x_s, g_i)\ C_i \in P, g_i \in \Lambda$$

The dynamic algorithm is performed by the following steps:

a) *Initialization*; a partition $P = (C_1, \ldots, C_k)$ of E is randomly chosen.
b) *representation step:*
   for j=1 to k, find g_h associated to C_h such that $\sum_{x_s \in C_h} \mu_s.\delta^2(x_s, g_h)$ is minimized
c) *allocation step:*
   test $\longleftarrow$ 0
   for all x_s do
      find m such that C_m is the class of s of E
      find l such that: $l = \arg\min_{h=1,\ldots,k} \delta(x_s, g_h)$
      if $l \neq m$
         test $\longleftarrow$ 1; $C_l \longleftarrow C_l \cup \{x_s\}$ and $C_m \longleftarrow C_m - \{x_s\}$
d) if $test = 0$ then stop, else go to b)

Then, the first choice concerns with the representation structure by prototypes $(g_1, \ldots, g_k)$ for the classes $\{C_1, \ldots, C_k\} \in P$.
The criterion $\Delta(P, L)$ is an additive function of the k clusters and of the N SO's of E. Therefore, the criterion Δ decreases under the following conditions:

- *uniqueness* of the affectation cluster for each element of E;
- *uniqueness* of the prototype g_h which minimizes the criterion Δ for all the cluster C_h for $(h = 1, \ldots, k)$ of the partition P of E.

3 Crossed dynamical clustering algorithm

In order to find a structure in the symbolic data, we perform a crossed dynamical clustering algorithm. The data are described in a symbolic data table $\mathbf{X}$=$[X^1, \ldots, X^v, \ldots, X^p]$. Along the rows of $\mathbf{X}$ we find the descriptions of the symbolic objects x_s $(s = 1, ..., N)$ of E, the columns of X^v contain the distributions of the symbolic variables y_v $(v = 1, ..., p)$. The set of categories of the symbolic variable y_v is denoted V_v and $V = \bigcup_{v=1}^{p} V_v$.
The general scheme of the dynamical algorithm, described above, is followed in order to cluster the rows of the symbolic data table $\mathbf{X}$ in a set of homogeneous classes, representing typology of SO's or groups of categories.

According to our aim to obtain rows partition, a classification of the symbolic descriptors is accomplished. Some authors (Govaert, 1977, Govaert and Nadif, 2003) proposed the maximization of the χ^2 criterion between rows and columns of a contingency table.

In our context we extent the crossed clustering algorithm to look for the partition P of the set E in r classes of objects and the partitions Q in c column-groups of V, according to the Φ^2 criterion on symbolic modal variables. It worth to notice that the criterion optimized in such algorithm is additive:

$$\Delta(P,(Q^1,\ldots,Q^p)) = \sum_{v=1}^{p} \Phi^2(P,Q^v|Q)$$

where Q^v is the partition associated to the modal variable y_v and $Q = (Q_1,\ldots,Q_c) = (\bigcup_{v=1}^{p} Q_1^v,\ldots,\bigcup_{v=1}^{p} Q_k^v,\ldots,\bigcup_{v=1}^{p} Q_c^v,)$.

The cells of the crossed tables can be modelling by marginal distributions (or profiles) summarizing the classes descriptions of the rows and columns.

The criterion $\Delta(P,Q,G)$ optimized in the crossed algorithm is consistent with the clustering one and iteratively optimizes the two partitions P and Q and the related representation $\mathbf{G}$.

$$(\mathbf{X^1},\ldots,\mathbf{X^v} = \begin{pmatrix} x_{1v_1} & \cdots & x_{1v_j} & \cdots & x_{1v_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{sv_1} & \cdots & x_{sv_j} & \cdots & x_{sv_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{Nv_1} & \cdots & x_{Nv_j} & \cdots & x_{Nv_m} \end{pmatrix} \ldots,\mathbf{X^p}) \Rightarrow \mathbf{G} = \begin{pmatrix} g_{11} & \cdots & g_{1c} \\ \cdots & g_{ik} & \cdots \\ g_{r1} & \cdots & g_{rc} \end{pmatrix}$$

The value g_{ik} of the matrix $\mathbf{G}$ represents the aggregation of the set of rows belonging to the class P_i with the set of columns belonging to the class Q_k is computed by the following formula:

$$g_{ik} = \sum_{v=1}^{p} \sum_{s\in P_i} \sum_{j\in Q_k^v} x_{sj} = \sum_{s\in P_i} \sum_{j\in Q_k} x_{sj} \tag{1}$$

We denote the marginal profiles of the $\mathbf{G}$ matrix as follows:

$$g_{.k} = \sum_{i=1}^{r} g_{ik} = \sum_{i=1}^{r} \sum_{s\in P_i} \sum_{j\in Q_k} x_{sj} = \sum_{j\in Q_k} \sum_{s=1}^{N} x_{sj} = \sum_{j\in Q_k} x_{.j}$$

The ϕ^2 distance between a row vector of $\mathbf{X}$ and the row vector $g_i = (g_{i1},\ldots,g_{ic})$ of $\mathbf{G}$ is computed with respect to the aggregated $\tilde{x}_s$'s rows $\tilde{x}_s^v = (\tilde{x}_{s1}^v,\ldots,\tilde{x}_{sc}^v)$ belonging to the partition Q^v, for each variable y_v, where:

$$\phi^2(\tilde{x}_s^v, g_i) = \sum_{k=1}^{c} \frac{1}{g_{.k}} \left(\frac{\tilde{x}_{sk}^v}{\tilde{x}_{s.}^v} - \frac{g_{ik}}{g_{i.}} \right)^2 \qquad \text{with} \tag{2}$$

$$g_{i.} = \sum_{k=1}^{c} g_{ik} \qquad \tilde{x}_{sk}^v = \sum_{j\in Q_k^v} x_{sj} \qquad \tilde{x}_{s.}^v = \sum_{j=1}^{c} \sum_{j\in Q_k^v} x_{sj}$$

$$d^2(\tilde{x}_s, g_i) = \sum_{v=1}^{p} \phi^2(\tilde{x}_s^v, g_i) \tag{3}$$

The ϕ^2 distance between a column vector of $\mathbf{X}$ and the column vector $g^k = (g_{1k}, \ldots, g_{rk})$ of $\mathbf{G}$ is computed with respect to the aggregated $\tilde{x}^j$'s column $\tilde{x}^j = (\tilde{x}^{1j}, \ldots, \tilde{x}^{rj})$ belonging to the partition P, where:

$$\phi^2(\tilde{x}^j, g^k) = \sum_{i=1}^{r} \frac{1}{g_{i.}} \left(\frac{\tilde{x}^{ij}}{\tilde{x}^{\cdot j}} - \frac{g_{ik}}{g_{.k}} \right)^2 \qquad \text{with} \tag{4}$$

$$g_{.k} = \sum_{i=1}^{r} g_{ik} = \sum_{s=1}^{N} \sum_{j \in Q_k} x_{sj} \qquad \tilde{x}^{ij} = \sum_{s \in P_i} x_{sj} \qquad \tilde{x}^{\cdot j} = \sum_{i=1}^{r} \sum_{s \in P_i} x_{sj} = x_{.j}$$

The Crossed Dynamic Algorithm is performed by the following steps:

a) *Initialization*; a partition $P = (P_1, \ldots, P_r)$ of E and p partitions ($Q^v = (Q_1^v, \ldots, Q_c^v), v = 1, \ldots, p$) are randomly chosen.
b) *Block model representation step:*
   The prototype table G is computed by the formula (1).
c) *Row allocation step:*
   test_row $\longleftarrow$ 0;
   for all objects s of E do
      Such that P_i is the class of s, find i^* which verifies :
      $i^* = \arg\ \min_{i=1,\ldots,r} d(\tilde{x}_s, g_i)$ where d is defined by (3)
      if $i^* \neq i$
         test_row $\longleftarrow$ 1; $P_{i^*} \longleftarrow P_{i^*} \cup \{s\}$ and $P_i \longleftarrow P_i - \{s\}$
d) *Block model representation step:*
   The prototype table G is computed by the formula (1).
e) *Column allocation step:*
   test_column $\longleftarrow$ 0
   for all variables y_v do
   for all categories j of V^v do
      Such that Q_k^v is the class of j, find j^* which verifies :
      $j^* = \arg\ \min_{k=1,\ldots,c} \phi(\tilde{x}^j, g^k)$ where ϕ is defined by (4)
         if $j^* \neq j$
         test_column $\longleftarrow$ 1; $Q_{k^*}^v \longleftarrow Q_{k^*}^v \cup \{j\}$ and $Q_j^v \longleftarrow Q_k^v - \{j\}$
f) if $test\_row = 0$ and $test\_column = 0$ then stop, else go to b)

Using the theorem of the decomposition of the inertia we have the following relations :

$$\Phi^2(E, Q) = \Delta(P, Q, G) + \Phi^2(P, Q) \tag{5}$$

$$\Phi^2(P, V) = \Delta(P, Q, G) + \Phi^2(P, Q) \tag{6}$$

For the row allocation step b) the partition Q and the prototype block model are fixed also the criterion $\Delta(P, Q, G) = \sum_{i=1}^{r} \sum_{s \in P_i} x_{s.} d^2(\tilde{x}_s, g_i)$ decreases during this step. By the relation (5) the criterion $\phi^2(P, Q)$ increases.

For the column allocation step e) the partition P and the prototype block model are fixed also the criterion $\Delta(P, Q, G) = \sum_{k=1}^{c} \sum_{j \in Q_k^v} x_{.j} \phi^2(\tilde{x}^j, g^k)$ decreases during this step. By the relation (6)the criterion $\Phi^2(P, Q)$ increases.

Globally the criterion $\phi^2(P, Q)$ increases in each step of this process.

4 Application

A direct extension of the dynamical algorithms is hereafter proposed in the context of the Web Usage Mining (Sauberlich and Huber, 2001). In particular, the application has performed on the Web Logs Data, coming from the HTTP log files by the INRIA web server (Lechevallier et al., 2003). This study aims to detect the behavior of the users and, in the same time, to check the efficacy of the structure of the site. Behind the research of typologies of users, we have defined a hierarchical structure (taxonomy) over the pages at different levels of the directories. The analyzed data set has concerned the set of *page views* by visitors that were connected to the INRIA site from the 1^{st} to the 15^{th} of January, 2003. Globally, the database contained 673.389 clicks (like *page views* in an user session), which have been already filtered from robot/spider entries and accesses of graphic files.
A very important aspect in the analyzing of logfiles is the *navigation* which is a set of clicks belonging to the same user. A further cleaning of the logfile has been performed in order to keep the navigations on both URL: *www.inria.fr* and *www-sop.inria.fr*. Moreover, only *long navigations* (duration $\geq$ 60s, the ratio duration/number of clicks $\geq 4sec.$ and number of visited pages ≥ 10) has been taken into account for the analysis. Therefore, the number of selected navigations was 2639, corresponding to 145643 clicks. For sake of brevity, in this context, we have restrained our analysis just to two web sites at the highest level. The visited pages were collected in semantic topics according to the structure of the two web sites. In particular the clicks on the web site *www.inria.fr* were referred to 44 topics; while the clicks on the web site *www-sop.inria.fr*, to 69 topics. Thus, we have consider the 2639 as symbolic objects described by two symbolic multi-categorical variables: *www.inria.fr* and *www-sop.inria.fr* having 44 and 69 categories respectively. The data are collected in a symbolic tables where each row contains the descriptions of a symbolic object (navigation), that is the distribution of the visited topics on the two websites. Following our aim to study the behavior of the INRIA web users, we have performed symbolic clustering analysis to identify an homogeneous typology of users according to the sequence of the visited web pages, or better, according to the occurrences of the visited pages of the several semantic topics.

The results of the navigation set partition in 12 classes and of the topics one in 8 classes, constituted by the two partitions Q^1 and Q^2, are shown in the Table 1.

For the example, the *Topic_5* associated to the group Q^5 is composed by two subgroups, one for each website, Q_1^5={*travailler, formation, valorisation*} for the website *www* and Q_1^5={*formation, recherche*} for the website *sop*.

It is worth to notice as the 8 topics groups correspond to different typology of information. In particular, the 8 groups can be identify as follows:

T-group 1 $\rightarrow$ *INTRANET*; T-group 2 $\rightarrow$ *Scientific information: Conferences, project activities*; T-group 3 $\rightarrow$ *Dissemination*; T-group 4 $\rightarrow$ *dias*; T-group 5 $\rightarrow$

Topic_1	Topic_2	Topic_3	Topic_4
/www/partenaires	/www/projets	/www/presse	/www/dias
/www/agos-sophia	/www/rrrt	/www/actualites-siege	/sop/dias
/www/modeles	/www/w3c	/www/multimedia	
/sop/partenaires	/www/manifestations	/www/icons	
/sop/agos-sophia	/sop/projets	/www/fonctions	
/sop/color	/sop/sophia	/sop/chir	
/sop/interne-sophia	/sop/site-eng	/sop/direction	
/sop/wiki	/sop/externe		
/sop/modeles	/sop/colloquium	Topic_7	Topic_8
/sop/sapr	/sop/horde		
/sop/didacticiel	/sop/manifestations	/www/recherche	/www/sophia
/sop/ctime	/sop/international	/www/accueil-siege	/www/site-old
/sop/freesoft		/www/personnel	/sop/cgi-bin
		/www/intro-inria	/sop/commun
		/www/publications	/sop/accueil-sophia
Topic_5	Topic_6	/www/cgi-bin	/sop/intro-sophia
		/www/ra	/sop/actualites-sophia
/www/travailler	/www/rapports	/www/interne-siege	/sop/rev
/www/formation	/www/semir	/www/international	/sop/intech
/www/valorisation	/sop/rapports	/www/site-beta	/sop/services
/sop/formation	/sop/semir	/www/sophia-antipolis	/sop/challengeTV
/sop/recherche	/sop/rmi	/www/thesauria	/sop/xml

Table 1. topic descriptions groups

Training; T-group 6 → *Research activity*; T-group 7 → *Headquarter (www.inria.fr)*; T-group 8 → *Headquarter - Sophia research activity.*

From the classification Table 2 we can remark as: the topics-group 2 represents the set of the most visited topics by the users; mainly the users of the class 8 visited this group attentively; the topics-group 1 represents the set of topics specially visited by the users of the class 3; the topic group 1 contains the internal internet users of INRIA. Then, analyzing the classes of navigations: the class 3 contains the navigations with an high number of pages visited; the users of this class visited different topic groups (1,2,6 and 7); the class 4 contains the navigations which visited only the topics group 5. This topics group represents the general topics of INRIA (training, researchers, scientific manifestations, etc.)

	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	pages
Navigation_ 1	222	1470	587	34	611	80	18757	143	21904
Navigation_ 2	78	2381	254	7	3094	80	2055	249	8198
Navigation_ 3	8578	7767	425	309	448	2749	2091	1386	23753
Navigation_ 4	29	280	115	7	3387	7	347	91	4263
Navigation_ 5	209	242	9	26	23	2544	221	55	3329
Navigation_ 6	29	1185	3204	28	1247	19	2670	82	8464
Navigation_ 7	43	140	22	795	39	47	218	636	1940
Navigation_ 8	288	35742	920	90	594	308	2174	1101	41217
Navigation_ 9	186	1040	136	106	283	72	370	3739	5932
Navigation_10	24	39	6	2786	2	25	49	210	3141
Navigation_11	175	7630	606	87	574	326	10227	257	19882
Navigation_12	4	231	3088	4	96	8	179	10	3620
Total pages	9865	58147	9372	4279	10398	6265	39358	7959	145643

Table 2. Contingence table of the navigations and topic groups

The achieved results by the proposed algorithm must be considered just as a bref example of an automatic clustering procedure to structure complex data to perform simultaneously typologies of navigation and groups of topics, homogenous from a semantic point of view.

An extension of our approach to more web sites, or in general to more symbolic variables, allows to take in account a hierarchical structure of the complex data descriptors. According to our example, if we had to take into account rubriques at lower level of the web architecture, in the grouping of the topics, their belonging to a higher level rubriques of the web site must be considered in the clustering process.

In conclusion, the most relevant difference of the crossed clustering algorithm on complex data with respect to the one on classical data, is surely in its extension to multi-valued categorical variables with a hierarchical structure associated.

References

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

CHAVENT, M., DE CARVALHO, F.A.T., LECHEVALLIER, Y., VERDE, R. (2003): Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistique Appliquées*, n. 4.

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, 19, 2, 19–34.*

GOVAERT, G. (1977): Algorithme de classification d'un tableau de contingence. In Proc. of *first international symposium on Data Analysis and Informatics*, INRIA, Versailles, 487–500.

GOVAERT, G. (1995): Simultaneous clustering of rows and columns. *Control Cybernet.*, 24, 437–458

GOVAERT, G., NADIF M. (2003): Clustering with block mixture models. *Pattern Recognition*, Elservier Science Publishers, 36, 463-473

LECHEVALLIER, Y., TROUSSE, B., VERDE, R., TANASA, D. (2003): *Classification automatique: Applications au Web-Mining.* In: Proceeding of SFC2003, Neuchatel, 10–12 September.

SAUBERLICH, F., HUBER K.-P. (2001) : A Framework for Web Usage Mining on Anonymous Logfile Data. In : Schwaiger M. and Opitz O.(Eds.): *Exploratory Data Analysis in Empirical Research*, Springer-Verlag, Heidelberg, 309–318.

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.