Pattern Recognition Letters 31 (2010) 430–443

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Unsupervised pattern recognition models for mixed feature-type symbolic data

Francisco de A.T. de Carvalho [*], Renata M.C.R. de Souza

Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n, Cidade Universitária, CEP 50740–540, Recife (PE), Brazil

ARTICLE INFO

Article history:
Received 30 October 2008
Received in revised form 6 June 2009
Available online 13 November 2009

Communicated by R.C. Guido

Keywords:
Symbolic data analysis
Clustering analysis
Mixed feature-type symbolic data
Adaptive distances
Partition interpretation indices
Cluster interpretation indices

ABSTRACT

Unsupervised pattern recognition methods for mixed feature-type symbolic data based on dynamical clustering methodology with adaptive distances are presented. These distances change at each algorithm's iteration and can either be the same for all clusters or different from one cluster to another. Moreover, the methods need a previous pre-processing step in order to obtain a suitable homogenization of the mixed feature-type symbolic data into histogram-valued symbolic data. The presented dynamic clustering algorithms have then as input a set of vectors of histogram-valued symbolic data and they furnish a partition and a prototype to each cluster by optimizing an adequacy criterion based on suitable adaptive squared Euclidean distances. To show the usefulness of these methods, examples with synthetic symbolic data sets as well as applications with real symbolic data sets are considered. Moreover, various tools suitable for interpreting the partition and the clusters given by these algorithms are also presented.

1. Introduction

Clustering methods seek to organize a set of items (usually represented as a vector of quantitative values in a multidimensional space) into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity. These methods have been widely applied in various areas such as taxonomy, image processing, information retrieval, data mining, etc. and they may be divided into hierarchical and partitioning methods (Jain et al., 1999; Gordon, 1999): hierarchical methods yield complete hierarchy, i.e., a nested sequence of partitions of the input data, whereas partitioning methods seek to obtain a single partition of the input data in a fixed number of clusters, usually by optimizing an objective function.

The partitioning dynamical cluster algorithms (Diday, 1971; Diday and Simon, 1976) are iterative two-step relocation algorithms involving the construction of clusters at each iteration and the identification of a suitable representation or prototype (means, axes, probability laws, groups of elements, etc.) for each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding representations. The adaptive dynamic clustering algorithm (Diday and Govaert, 1977) also optimize a criterion based on a measure of fitting between the clusters and their prototypes, but there are distances to compare clusters and their prototypes that change at each iteration. These distances are not determined once and for all, and moreover, they can be different from one cluster to another. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes.

In classical clustering analysis, the patterns to be grouped are usually represented as a vector of quantitative or qualitative measurements where each column represents a variable. Each particular pattern takes a single value for each variable. In practice, however, this model is too restrictive to represent complex data. In order to take into account variability and/or uncertainty inherent to the data, variables must assume sets (or ordered lists) of categories or intervals, possibly even with frequencies or weights. Symbolic Data Analysis (SDA), a domain in the area of knowledge discovery and data management related to multivariate analysis, pattern recognition and artificial intelligence, has provided suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by multi-valued variables, where the cells of the data table contain sets (or ordered lists) of categories, intervals, or weight (probability) distributions (Bock and Diday, 2000; Billard and Diday, 2007; Diday and Noirhome-Fraiture, 2008).

In SDA the clustering methods for symbolic data differ in the type of the considered symbolic data, in their cluster structures and/or in the considered clustering criteria. With hierarchical methods, Gowda and Diday (1991) introduced an agglomerative approach that forms composite symbolic objects using a join operator whenever mutual pairs of symbolic objects are selected for agglomeration based on minimum dissimilarity. Ichino and

[*] Corresponding author. Tel.: +55 81 21268430; fax: +55 81 21268438.
E-mail addresses: fatc@cin.ufpe.br (F.A.T. de Carvalho), rmcrs@cin.ufpe.br (R.M.C.R. de Souza).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.11.007

Yaguchi (1994) defined generalized Minkowski metrics for mixed feature variables and presents dendrograms obtained from the application of standard linkage methods for data sets containing numeric and symbolic feature values. Gowda and Ravi (1995a,b), respectively, presented divisive and agglomerative algorithms for symbolic data based on the combined usage of similarity and dissimilarity measures. These proximity (similarity or dissimilarity) measures are defined on the basis of the position, span and content of symbolic objects. Chavent (2000) proposed a divisive clustering method for symbolic data that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterization of each cluster in the hierarchy. Gowda and Ravi (1999) presented a hierarchical clustering algorithm for symbolic objects based on the gravitational approach, which is inspired on the movement of particles in space due to their mutual gravitational attraction. Guru et al. (2004) and Guru and Kiranagi (2005) introduced agglomerative clustering algorithms based, respectively, on similarity and dissimilarity functions that are multi-valued and non-symmetric.

A number of authors have addressed the problem of non-hierarchical clustering for symbolic data. Diday and Brito (1989) used a transfer algorithm to partition a set of symbolic objects into clusters described by weight distribution vectors. Ralambondrainy (1995) extended the classical k-means clustering method in order to manage data characterized by numerical and categorical variables, and complemented this method with a characterization algorithm to provide a conceptual interpretation of the resulting clusters. Gordon et al. (2000) presented an iterative relocation algorithm to partition a set of symbolic objects into classes so as to minimize the sum of the description potentials of the classes. Verde et al. (2001) introduced a dynamic clustering algorithm for symbolic data considering context-dependent proximity functions, where the cluster representatives are weight distribution vectors. Bock (2003) has proposed several clustering algorithms for symbolic data described by interval variables, based on a clustering criterion and has thereby generalized similar approaches in classical data analysis.

Concerning partitional dynamic clustering algorithms for symbolic data, Chavent and Lechevallier (2002) proposed a dynamic clustering algorithm for interval data where the cluster representatives are defined by an optimality criterion based on a modified Hausdorff distance. Souza and De Carvalho (2004) proposed partitioning clustering methods for interval data based on city-block distances, also considering adaptive distances. De Carvalho et al. (2006a) proposed an algorithm using an adequacy criterion based on adaptive Hausdorff distances and De Carvalho et al. (2006b) presented dynamical clustering algorithms based on non-adaptive Euclidean distances for interval data. More recently, De Carvalho and Lechevallier, 2009 presented dynamic clustering algorithms based on single adaptive (city-block and Hausdorff) distances that change at each iteration, but are the same for all clusters. However, none of these former dynamic clustering models are able to manage mixed feature-type symbolic data.

In this paper, we introduce dynamic clustering methods for mixed feature-type symbolic data based on suitable adaptive squared Euclidean distances used for compare clusters and their respective prototypes that change at each iteration: adaptive distances for each cluster, which are different from one cluster to another, and single adaptive distances, which are the same for all clusters. To be able to manage mixed feature-type symbolic data, these methods assume a previous pre-processing step the aim of which is to obtain a suitable homogenization of mixed feature-type symbolic data into histogram-valued symbolic data.

This paper is organized as follows. Section 2 first describes mixed feature-type symbolic data. and then introduces dynamical clustering algorithm for mixed feature-type symbolic data which uses, respectively, a single adaptive squared Euclidean distance and an adaptive squared Euclidean distance for each class. Section 3 presents various tools for cluster interpretation according to these adaptive clustering models: indices for evaluating the quality of a partition, the homogeneity and eccentricity of the individual clusters and the role played by the different variables in the cluster formation process. To show the usefulness of these clustering algorithms and the merit of these cluster interpretation tools, experiments with simulated data in a framework of a Monte Carlo schema as well as applications with real symbolic interval-valued data sets are considered in Section 4. Finally, Section 5 gives the concluding remarks.

2. Dynamic clustering algorithms for mixed feature-type symbolic data

In classical data analysis, an individual is described by a row of a data matrix whose columns are single-valued variables, i.e., variables that assumes only one value from their domain. According to its domain, a variable may be quantitative (discrete or continuous) or qualitative (ordinal or nominal). However, this type of data is too restrictive to represent complex data which may comprehend, for instance, variability and/or uncertainty. It is why different types of symbolic variables and symbolic data have been introduced in symbolic data analysis.

A symbolic variable (Bock and Diday, 2000) X_j is set-valued if, given an item $i, X_j(i) = x_i^j \subseteq A_j$ where $A_j = \{t_1^j, \ldots, t_{H_j}^j\}$ is a set of categories. A symbolic variable X_j is ordered list-valued if, given an item $i, X_j(i) = x_i^j$, where x_i^j is a sub-list of a ordered list of categories $A_j = [t_1^j, \ldots, t_{H_j}^j]$. A symbolic variable X_j is an interval-valued variable when, given an item $i, X_j(i) = x_i^j = [a_i^j, b_i^j] \in [a, b]$, where $[a, b] \in \Im$ and $\Im$ is the set of closed intervals defined from $\Re$. Finally, a symbolic variable X_j is histogram-valued variable if, given an item $i, X_j(i) = x_i^j = (S^j(i), \mathbf{q}^j(i))$ where $\mathbf{q}^j(i) = (q_{i1}^j, \ldots, q_{iH_{ij}}^j)$ is a vector of weights defined in $S^j(i)$ such that a weight $q(m)$ corresponds to each category $m \in S^j(i)$. $S^j(i)$ is the support of the measure $\mathbf{q}^j(i)$.

Table 1 shows a mixed feature-type symbolic data table describing four cities. Symbolic variables X_1 (number of inhabitants in thousands), X_2 (spectrum of political parties: Democrats, Conservatives, Socialists, Nationalists), X_3 (Consulates: France, Italy, Spain, Great-Britain, Germany, Belgium) are, respectively, interval-valued, histogram-valued and set-valued.

Let a generic data table representing the values of p symbolic variables $X_1, \ldots, X_p$ on a set $\Omega = \{1, \ldots, n\}$ of n objects each one represented as a vector of mixed feature-type symbolic data $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ $(i = 1, \ldots, n)$. This means that $x_i^j = X_j(i)$ can be a set or a (ordered) list of categories, an interval or a weight distribution according to the type of the corresponding symbolic variable. Table 2 shows a mixed feature-type symbolic data table where X_1 is an interval-valued variable, X_j is a set-valued variable and X_p is a histogram-valued variable.

The standard dynamical clustering algorithm (Diday and Simon, 1976) aims to provide a partition $P = (C_1, \ldots, C_K)$ of Ω in a fixed number K of clusters and their corresponding prototypes $L = (L_1, \ldots, L_K)$ by locally minimizing a criterion W that evaluates the fit between the clusters and their corresponding representatives.

Table 1
Mixed feature-type symbolic data table describing four cities.

City	X_1	X_2	X_3
1	[70,100]	$((D, C, S, N), (0.4, 0.3, 0.2, 0.1))$	$\{F, I\}$
2	[50,70]	$((D, C, S, N), (0.3, 0.3, 0.3, 0.1))$	$\{S, G\}$
3	[20,40]	$((D, C, S, N), (0.2, 0.2, 0.2, 0.4))$	$\{GB, G\}$
4	[60,100]	$((D, C, S, N), (0.1, 0.3, 0.4, 0.2))$	$\{B, GB\}$