Partitioning Clustering Algorithms for Interval-Valued Data based on City-Block Adaptive Distances

Francisco de A.T. de Carvalho[1], Yves Lechevallier[2]

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

[2]INRIA – Rocquencourt
Domaine de Voluceau – Rocquencourt
B.P. 105 – 78153 Le Chesnay Cedex, France

fatc@cin.ufpe.br, yves.lechevallier@inria.fr

Abstract. *The recording of interval-valued data has become a common practice nowadays. This paper presents some partitioning clustering algorithms for interval-valued data. The proposed methods furnish a partition of the input data and a corresponding prototype (a vector of intervals) for each cluster by optimizing an adequacy criterion which is based on suitable adaptive city-block distances between vectors of intervals. Experiment with a real interval-valued data set shows the usefulness of the proposed method.*

1. Introduction

Clustering is one of the most popular tasks in knowledge discovery and is applied in various fields, including data mining, pattern recognition, computer vision, etc. These methods seek to organize a set of items 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. Partitioning clustering methods [Everitt 2001], [Gordon 1999], [Jain et al 1999] seek to obtain a single partition of the input data into a fixed number of clusters. Such methods often look for a partition that optimizes (usually locally) an adequacy criterion function.

The partitioning dynamic cluster algorithms [Diday and Simon 1976] are iterative two steps relocation clustering algorithms involving at each iteration the construction of the clusters and the identification of a suitable prototype (means, factorial axes, probability laws, etc.) of each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding prototypes.

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.

Often, objects to be clustered are represented as a vector of quantitative features. However, the recording of interval data has become a common practice in real world

applications and nowadays this kind of data is often used to describe objects. Symbolic Data Analysis (SDA) is an area related to multivariate analysis, data mining and pattern recognition, which has provided suitable data analysis methods for managing objects described as a vector of intervals [Bock and Diday 2000].

Concerning dynamical cluster algorithms for symbolic interval data, [Chavent and Lechevallier 2002] proposed an algorithm using an adequacy criterion based on Hausdorff distances. [Souza and De Carvalho 2004] presented a dynamic cluster algorithm for symbolic interval data based on City-Block distances. [De Carvalho et al 2006] proposed an algorithm using an adequacy criterion based on adaptive Hausdorff distances for each cluster. In this paper, we introduce partitioning dynamic clustering methods based on single adaptive city-block distances. These adaptive distances change at each iteration but are the same for all clusters.

2. Partitioning dynamic clustering algorithms for interval-valued data based on city-block adaptive distances

This section presents partitioning dynamic clustering methods for interval-valued data based on single adaptive city-block distances. These adaptive distances are defined by one weight vectors. The main idea of these methods is that there is a distance to compare clusters and their representatives (prototypes) that changes at each iteration but is the same for all clusters.

Let Ω be a set of n objects described by p interval-valued variables. An interval-valued variable [Bock and Diday 2000] is a mapping from Ω in $\Re$ such that for each $i \in \Omega$, $X(i) = [a, b] \in \Im$, in which $\Im$ is the set of closed intervals defined from $\Re$. Each object i $(i=1,\ldots,n)$ is represented as a vector of intervals $\boldsymbol{x}_i = ([a_i^1 , b_i^1], \ldots , [a_i^p , b_i^p])$.

This adaptive clustering method looks for a partition of Ω into K clusters $P_1,\ldots,P_K$ and their corresponding prototypes $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K$ such that an adequacy criterion J measuring the fitting between the clusters and their prototypes is locally minimized. Assuming that each cluster P_k is also represented as a vector of intervals $\boldsymbol{y}_k = ([\alpha_k^1,\beta_k^1], \ldots , [\alpha_k^p,\beta_k^p])$, the criterion J is defined as:

$$J = \sum_{k=1}^{K} \sum_{i \in P_k} d(\mathbf{x}_i, \mathbf{y}_k) \text{ in which } d(\mathbf{x}_i, \mathbf{y}_k) = \sum_{j=1}^{p} \lambda^j \left[\max\{ | a_i^j - \alpha_k^j |, | b_i^j - \beta_k^j | \} \right]$$

is an adaptive city-block distance measuring the dissimilarity between an object $\boldsymbol{x}_i$ $(i = 1, \ldots , n)$ and a cluster prototype $\boldsymbol{y}_k$ $(k = 1, \ldots ,K)$ that is parameterized by the weight vector $\boldsymbol{\lambda} = (\lambda^1, \ldots , \lambda^p)$, which changes at each iteration but is the same for all clusters.

This algorithm sets an initial partition and alternates three steps until convergence, when the criterion J reaches a stationary value representing a local minimum.

2.1. Step 1: definition of the best prototypes

The partition of Ω into K clusters and the weight vector $\boldsymbol{\lambda}$ are fixed. The prototype $\boldsymbol{y}_k = ([\alpha_k^1,\beta_k^1], \ldots , [\alpha_k^p,\beta_k^p])$ of cluster P_k, which minim**izes** the clustering criterion J, has the boundaries of the interval $[\alpha_k^j,\beta_k^j]$ $(j = 1, \ldots , p)$ calculated **a**ccording to:

α_k^j is the median of the set $\{ a_i^j : i \in P_k \}$ and β_k^j is the median of the set $\{ b_i^j : i \in P_k \}$

2.2. Step 2: definition of the best adaptive distances

The partition of Ω into K clusters and the prototypes $\mathbf{y}_k$ $(k=1,\ldots,K)$ are fixed. The vector of weights $\boldsymbol{\lambda} = (\lambda^1, \ldots, \lambda^p)$, which minimizes the clustering criterion J under $\lambda^j > 0$ and $\prod_{j=1}^{p} \lambda^j = 1$, has its weights λ^j $(j=1,\ldots,p)$ calculated according to the following expression:

$$\lambda^j = \frac{\left\{\prod_{h=1}^{p}\left(\sum_{k=1}^{K}\left[\sum_{i\in P_k}\left(\left|a_i^h - \alpha_k^h\right| + \left|b_i^h - \beta_k^h\right|\right)\right]\right)\right\}^{\frac{1}{p}}}{\sum_{k=1}^{K}\left[\sum_{i\in P_k}\left(\left|a_i^j - \alpha_k^j\right| + \left|b_i^j - \beta_k^j\right|\right)\right]}$$

2.3. Step 3: definition of the best partition

T the prototypes $\mathbf{y}_k$ $(k=1,\ldots,K)$ and the vector of weights $\boldsymbol{\lambda} = (\lambda^1, \ldots, \lambda^p)$ are fixed. The clusters P_k $(k = 1, \ldots, K)$, which minimizes the clustering criterion J, are updated according to the following allocation rule:

$$P_k = \{i \in \Omega : d(\mathbf{x}_i, \mathbf{y}_k) \leq d(\mathbf{x}_i, \mathbf{y}_h), \forall h \neq k\}$$

3. City Temperature Interval-Valued Data Set

The city temperature interval-valued data set gives the average minimal and average maximal monthly temperatures of cities in degrees centigrade. This data set consists of a set of 503 cities described by 12 interval-valued variables (see Figure 1).

	January	February	...	November	December
Amsterdam	[−4, 4]	[−5, 3]	...	[1, 10]	[-1, 4]
Athens	[6, 12]	[6, 12]	...	[11, 18]	[8, 14]
...	...	...	...	...	...
Mauritius	[22, 28]	[22, 29]	...	[19, 27]	[21, 28]
...	...	...	...	...	...
Vienna	[−2, 1]	[−1, 3]	...	[2, 7]	[1, 3]
Zurich	[−11, 9]	[−8, 15]	...	[0, 19]	[−11, 8]

Figure 1. City Temparature Interval-Valued Data Set

With this city temperature interval-valued data set, the clustering algorithm was run until the convergence to a stationary value of the criterion J 100 times and the best result according to the adequacy criterion was selected. The main characteristics of the 6-cluster partition furnished by algorithm were:

- Cluster 1: the cities have very cold temperatures in winter similar to that of northern and eastern Europe;

- Cluster 2: the cities have temperatures similar to that of cities located in the southern hemisphere.

- Cluster 3: the cities have temperatures similar to that of western and central Europe.

- Cluster 4: the cities have a tropical climate and warm to hot temperatures

- Cluster 5: the cities have temperatures similar to that of southern Europe

- Cluster 6: the cities have a sub-tropical climate and warm to hot temperatures

Figure 2 shows the most discriminant months for the prototypes.

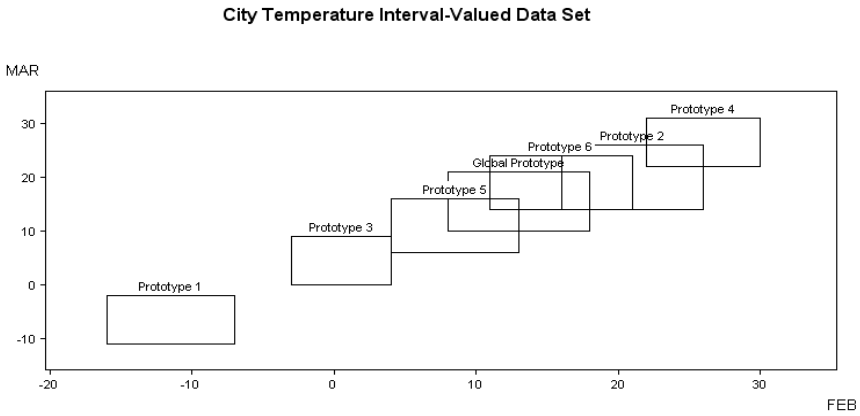

Figure 2. The most discriminant months

7. Final Remarks

In this paper, a partitioning clustering algorithm for interval-valued data was presented. This method furnishes a partition of the input data and a corresponding prototype for each cluster by optimizing an adequacy criterion which is based on adaptive city-block distances between vectors of intervals. An application with a city temperature interval-valued data set illustrates the usefulness of the proposed approach.

References

Bock, H.H. and Diday, E. (2000), Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data., Springer, Berlin Heidelberg.

Chavent, M. and Lechevallier, Y. (2002) "Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance", In: Classification, Clustering and Data Analysis, Edited by A. Sokolowski and H.-H. Bock, Springer, Heidelberg, 53–59.

De Carvalho, F.A.T, Souza, R.M.C.R., Chavent, M. and Lechevallier, Y. (2006), Adaptive Hausdorff distances and dynamic clustering of symbolic data, Pattern Recognition Letters, 27 (3), 167–179.

Diday, E. and Govaert, G. (1977), Classification Automatique avec Distances Adaptatives, R.A.I.R.O. Informatique Computer Science, 11 (4), 329–349.

Diday, E. and Simon, J.C. (1976) "Clustering analysis", In Digital Pattern Classification, Edited by K.S. Fu, Springer, Berlin et al, 47–94.

Everitt, B. (2001), Cluster Analysis, Halsted, New York.

Gordon, A. D. (1999), Classification, Chapman and Hall/CRC, Boca Raton, Florida.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), Data Clustering: A Review, ACM Computing Surveys, 31 (3), 264-323.

Souza, R. M. C. R. and De Carvalho, F. A. T. (2004), Clustering of interval data based on city-block distances, Pattern Recognition Letters, 25 (3), 353–365