A dynamic cluster algorithm based on L_r distances for quantitative data

Francisco de A. T. de Carvalho[1], Yves Lechevallier[2], and Renata M.C.R. de Souza[1]

[1] Centro de Informatica - CIn / UFPE, Av. Prof. Luiz Freire, s/n - Cidade Universitaria, CEP: 50740-540 - Recife - PE - Brasil, Email: {fatc,rmcrs}@cin.ufpe.br
[2] INRIA - Rocquencourt, Domaine de Voluceau - Rocquencourt - B. P. 105 78153 Le Chesnay Cedex - France, Email: Yves.Lechevallier@inria.fr

Summary. Dynamic cluster methods aim to obtain both a single partition of the input data into a fixed number of clusters and the identification of a suitable representation of each cluster simultaneously. In its adaptive version, at each iteration of these algorithms there is a different distance to the comparison of each cluster with its representation. In this paper, we present a dynamic cluster method based on L_r distances for quantitative data.

1 Introduction

Clustering (Bock (1993), Jain et al. (1999)) is an exploratory data analysis method the aim of which is to group a set of items into clusters such that items within a given cluster have a high degree of similarity, while items belonging to different clusters have a high degree of dissimilarity. The most popular cluster analysis techniques are hierarchical and partitional methods (Spaeth (1980), Gordon (1999), Everitt (2001)).

Hierarchical methods yields complete hierarchy, i.e., a nested sequence of partitions of the input data. Hierarchical methods can be either agglomerative or divisive. Agglomerative methods starts with trivial clustering, where each item is in a unique cluster, and ending with the trivial clustering, where all items are in the same cluster. A divisive method starts with all items in the same cluster and performs divisions until a stopping criterion is met.

Partitional methods aim to obtain a single partition of the input data into a fixed number of clusters. Often, these methods look for a partition that optimizes (usually locally) a criterion function. To improve the cluster's quality, the algorithm is run multiple times with different starting points, and the best configuration obtained from all the runs is used as the output clustering.

The dynamic cluster algorithms (Diday and Simon (1976)) are iterative two steps relocation algorithms involving at each iteration the construction of the clusters and the identification of a suitable representative or exemplar (means, axes, probability laws, groups of elements, etc.) of each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding representatives. The k-means algorithm with class representative updated after all objects have been considered for relocation, is a particular case of dynamical clustering with adequacy criterion equal to variance criterion such that class exemplar equal to cluster centers of gravity.

In these algorithms, the optimization problem can be stated as follow. Let Ω be a set of n objects indexed by $i = 1, ..., n$ and described by p quantitative variables. Then each object i is described by a vector $\mathbf{x}_i = (x_i^1, \ldots, x_i^p) \in \Re^p$. The problem is to find the partition $P = (C_1, ..., C_K)$ of Ω in K clusters and the system $Y = (\mathbf{y}_1, ..., \mathbf{y}_K)$ of class exemplars, which minimizes a partitioning criterion $g(P, Y)$ that measures the fitting between the clusters and their representatives.

This optimization process starts from a set of representatives or an initial partition and interactively applies an "allocation" step (the exemplars are fixed) in order to assign the individuals to the classes according to their proximity to the exemplars. This is followed by a "representation" step (the partition is fixed) where the exemplars are updated according to the assignment of the individuals in the allocation step, until achieving the convergence of the algorithm, when the adequacy criterion reaches a stationary value.

The dynamic cluster algorithm converges and the partitioning criterion decreases at each iteration if the class exemplars are properly defined at each representation step. Indeed, the problem is to find the exemplar $\mathbf{y}_k = (y_k^1, \ldots, y_k^p) \in \Re^p$ of each cluster $C_k (k = 1, \ldots, K)$, which minimizes an adequacy criterion $f(\mathbf{y}_k)$ measuring the dissimilarity between the exemplar $\mathbf{y}_k$ and the cluster C_k.

The adaptive dynamic clusters algorithms (Diday and Govaert (1977)) also optimize a criterion based on a measure of fitting between the clusters and their representation, but at each iteration there is a different distance to the comparison of each cluster with its representative. The idea is to associate each cluster with a distance which is defined according to the intra-class structure of the cluster. These distances are not determined once and for all, and moreover, they are different from one class to another. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes.

In these algorithms, the optimization problem is now to find the partition $P = (C_1, ..., C_K)$ of Ω in K clusters, its corresponding set of K exemplars $Y = (\mathbf{y}_1, ..., \mathbf{y}_K)$ and a set of K distances $d = \{d^1, \ldots, d^k\}$ each one associated with a cluster, which minimizes a partitioning criterion $g(P, Y, d)$ that measures the fitting between the clusters and their representatives.

The initialization, the allocation step and the stopping criterion are nearly the same in the adaptive and non-adaptive dynamic cluster algorithm. The

main difference between these algorithms occurs in the representation step which has two stages: a first stage, where the partition and the distances are fixed and the exemplars are updated, is followed by a second one, where the partition and their corresponding representatives are fixed and the distances are updated. The adaptive dynamic cluster algorithm converges and the partitioning criterion decreases at each iteration if the class exemplars and the distances are properly defined at each representation step.

The aim of this paper is to present a dynamic cluster method based on L_r distances for quantitative data. Sections 2 and 3 present, respectively, the non-adaptive and the adaptive version of this method. An example concerning adaptive and non-adaptive dynamic cluster method based on L_2 distance is given in section 4 and the concluding remarks are given in section 5.

2 A dynamic cluster method based on non-adaptive L_r distance

Let $\mathbf{x}_i = (x_i^1, ..., x_i^p)$ and $\mathbf{x}_{i'} = (x_{i'}^1, , ..., x_{i'}^p)$ be two quantitative features vectors, representation of objects i and i' belonging to class $C_k (k = 1, \ldots, K)$, respectively. We consider the L_r distance function to measure the dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$:

$$d_r(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} (|x_i^j - x_{i'}^j|)^r, \; r \geq 1 \tag{1}$$

In equation (1), $r = 1$ and $r = 2$ gives, respectively, L_1 and L_2 distances.

2.1 The optimization problem for class exemplar

As presented in the introduction, the exemplar $\mathbf{y}_k$ of a cluster C_k is defined in the framework of the dynamic cluster algorithm by optimizing an adequacy criterion f measuring the dissimilarity between the cluster and its representative. Here, we search the vector $\mathbf{y}_k = (y_k^1, ..., y_k^p)$ which minimizes the following adequacy criterion:

$$f(\mathbf{y}_k) = \sum_{i \in C_k} d_r(\mathbf{x}_i, \mathbf{y}_k) = \sum_{i \in C_k} \sum_{j=1}^{p} (|x_i^j - y_k^j|)^r, \; r \geq 1 \tag{2}$$

where d_r is the distance between two vectors of quantitative data as given by equation (1).

The criterion (2) can also be written:

$$f(\mathbf{y}_k) = \sum_{j=1}^{p} \overbrace{\sum_{i \in C_k} (|x_i^j - y_k^j|)^r}^{\bar{f}(y_k^j)} \tag{3}$$

and we search, for $j = 1, ..., p$, the quantity y_k^j which minimizes:

$$\tilde{f}(y_k^j) = \sum_{i \in C_k} (|x_i^j - y_k^j|)^r \tag{4}$$

When $r \in \{1, 2\}$ the minimum of the equation (4) has an analytical solution: for $r = 1$, y_k^j is the median of $\{x_i^j, i \in C_k\}$; for $r = 2$, y_k^j is the average of $\{x_i^j, i \in C_k\}$.

For $r > 2$, let $\{x_i^j | i \in C_k\}$ and let $X_{C_k}^j = \{x_{(1)}^j, \ldots, x_{(\#C_k)}^j\}$ be the set of increasing ordered values of $\{x_i^j | i \in C_k\}$, i.e., $\forall l \in \{1, \ldots, \#C_k\}, x_{(l)}^j \in \{x_i^j | i \in C_k\}$ and $\forall l \in \{1, \ldots, \#C_k - 1\}, x_{(l)}^j \leq x_{(l+1)}^j$.

Let the functions $\tilde{f}_t, t = 0, \ldots, \#C_k$, be

$$\tilde{f}_t(y_k^j) = \sum_{l=1}^{\#C_k} (|x_{(l)}^j - y_k^j|)^r,\ y_k^j \in \mathcal{B}_t \tag{5}$$

where $\mathcal{B}_0 =]-\infty, x_{(1)}^j]$, $\mathcal{B}_t = [x_{(l)}^j, x_{(l+1)}^j]$, $t, l = 1, \ldots, \#C_k - 1$, and $\mathcal{B}_{\#C_k} = [x_{(\#C_k)}^j, \infty[$

In that case, the quantity y_k^j which minimizes (4) belongs to the union of two sets: the set $\{x_i^j | i \in C_k\}$ and the set of roots of the derivatives, for $y_k^j \in \mathcal{B}_t$, of the functions $\tilde{f}_t$ $(t = 0, \ldots, \#C_k)$.

2.2 The dynamic cluster algorithm

The dynamic cluster algorithm search for the partition $P = (C_1, ..., C_K)$ of Ω and the system $Y = (\mathbf{y}_1, ..., \mathbf{y}_K)$ of class exemplars which locally minimizes the following partitioning criterion based on the distance d_r defined in (1):

$$g(P, Y) = \sum_{k=1}^{K} \sum_{i \in C_k} d_r(\mathbf{x}_i, \mathbf{y}_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (|x_i^j - y_k^j|)^r \tag{6}$$

This algorithm proceeds by iteratively repeating an "allocation" step and a "representation" step. During the "representation" step, the partition is fixed and the algorithm computes for each cluster C_k its representative $\mathbf{y}_k$ which minimizes the adequacy criterion given in (2). During the "allocation step", the exemplars are fixed and the algorithm performs a new partition by reassigning each object i to the closest class exemplar $\mathbf{y}_{k*}$ where $k* = arg\ min_{k=1,...,K}\ d(\mathbf{x}_i, \mathbf{y}_k)$.

Finally the algorithm is the following:

(a) Initialization

Choose a partition $(C_1, \ldots, C_K)$ of the data set Ω or choose K distinct

objects $\mathbf{y}_1, ..., \mathbf{y}_K$ among Ω and assign each object i to the closest exemplar $\mathbf{y}_{k*}$ ($k* = arg\ min_{k=l,...,K}\ d(\mathbf{x}_i, \mathbf{y}_k)$) to construct the initial partition $(C_1, \ldots, C_K)$.

(b) "Representation" step (the partition is fixed)
For k in 1 to K compute the exemplar $\mathbf{y}_k = (y_k^1, ..., y_k^p)$

(c) 'Allocation' step (the exemplars are fixed)
$test \leftarrow 0$
For i in 1 to n do
define the cluster C_{k*} such that $k* = arg\ min_{k=l,...,K}\ d(\mathbf{x}_i, \mathbf{y}_k)$
if $i \in C_k$ and $k* \neq k$
$test \leftarrow 1$
$C_{k*} \leftarrow C_{k*} \cup \{i\}$
$C_k \leftarrow C_k \backslash \{i\}$

(d) If $test = 0$ END, else go to (b)

3 A dynamic cluster method based on adaptive L_r distance

Let again $\mathbf{x}_i = (x_i^1, ..., x_i^p)$ and $\mathbf{x}_{i'} = (x_{i'}^1, , ..., x_{i'}^p)$ be two quantitative features vectors, representation of objects i and i' belonging to class C_k, respectively. We consider the following adaptive L_r distance function, which is parameterized by the weight vector $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p))$, to measure the dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$:

$$d_r^k(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \lambda_k^j (|x_i^j - x_{i'}^j|)^r,\ r \geq 1 \tag{7}$$

In equation (7), $r = 1$ and $r = 2$ gives, respectively, adaptive L_1 (Diday and Govaert (1977)) and adaptive L_2 distances.

3.1 The optimization problem for class exemplar

We search the vectors $\mathbf{y}_k = (y_k^1, ..., y_k^p)$ and $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p)$ which minimizes the following adequacy criterion:

$$f(\mathbf{y}_k, \lambda_k) = \sum_{i \in C_k} d_r^k(\mathbf{x}_i, \mathbf{y}_k) = \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j (|x_i^j - y_k^j|)^r,\ r \geq 1 \tag{8}$$

where d_r^k is the adaptive distance between two vectors of quantitative data given in (7).

The criterion (8) can also be written:

$$f(\mathbf{y}_k, \lambda_k) = \sum_{j=1}^{p} \lambda_k^j \overbrace{\sum_{i \in C_k} (|x_i^j - y_k^j|)^r}^{\tilde{f}(y_k^j)} \tag{9}$$

and the optimization problem will be solved in two stages.

In the first step the vector $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p)$ is fixed and we search, for $j = 1, ..., p$, the quantity y_k^j which minimizes:

$$\tilde{f}(y_k^j) = \sum_{i \in C_k} (|x_i^j - y_k^j|)^r \tag{10}$$

The solution is the same as pointed out in section 2.1.

In the second step, the vector $\mathbf{y}_k = (y_k^1, ..., y_k^p)$ is fixed and we search for $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p)$ which minimizes the adequacy criterion $f(\mathbf{y}_k, \lambda_k)$. According to the standard adaptive method (Diday and Govaert (1977)), we look for the coordinates λ_k^j $(j = 1, \ldots, p)$ of the vector λ_k that satisfies the following restrictions:

$$\lambda_k^j > 0 \ (j = 1, \ldots, p) \text{ and } \prod_{j=1}^{p} \lambda_k^j = 1 \tag{11}$$

These coordinates, which are calculated according to the Lagrange multipliers method (Govaert (1975)), are:

$$\lambda_k^j = \frac{\left[\prod_{h=1}^{p} (\sum_{i \in C_k} (|x_i^h - y_k^h|)^r)\right]^{\frac{1}{p}}}{\sum_{i \in C_k} (|x_i^j - y_k^j|)^r}, \ j = 1, \ldots, p \tag{12}$$

3.2 The adaptive dynamic cluster algorithm

The dynamic cluster algorithm searches for the partition $P = (C_1, ..., C_K)$ of Ω in K clusters, its corresponding set of K exemplars $Y = (\mathbf{y}_1, ..., \mathbf{y}_K)$ and a set of K distances $d = \{d_r^1, \ldots, d_r^k\}$ each one associated with a cluster, which locally minimizes the following partitioning criterion based on the distance d_r^k defined in (7):

$$g(P, Y, d) = \sum_{k=1}^{K} \sum_{i \in C_k} d_r^k(\mathbf{x}_i, \mathbf{y}_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j (|x_i^j - y_k^j|)^r \tag{13}$$

As in the standard dynamic cluster algorithm, this method performs an "allocation" step (the partition and the exemplars are fixed) in order to assign the individuals to the classes according to their proximity to the class representative, followed by a two-stages "representation" step where, according to the assignment of the individuals in the allocation step, in the first

stage the partition and the distances are fixed and the class exemplars are updated, whereas in the second stage the partition and the exemplars are fixed and the distances are updated. The algorithm iterates these steps until the convergence when the partitioning criterion reaches a stationary value. The algorithm schema is the following:

(a) Initialization
Choose a partition $(C_1, \ldots, C_K)$ of the data set Ω or choose K distinct objects $\mathbf{y}_1, ..., \mathbf{y}_K$ among Ω and assign each object i to the closest exemplar $\mathbf{y}_{k*}$ $(k* = arg\ min_{k=l,...,K}\ d_r^k(\mathbf{x}_i, \mathbf{y}_k))$ to construct the initial partition $(C_1, \ldots, C_K)$.
(b) "Representation" step
a) (The partition P and the distances d_r^k are fixed)
For $k = 1$ to K compute the exemplar $\mathbf{y}_k$
b) (the partition P and the exemplars $\mathbf{y}_k$ are fixed)
For $j = 1, \ldots, p$ and $k = 1, \ldots, K$, compute λ_k^j
(c) "Allocation" step
$test \leftarrow 0$
For i in 1 to n do
define the cluster C_{k*} such that $k* = arg\ min_{k=l,...,K}\ d_r^k(\mathbf{x}_i, \mathbf{y}_k)$
if $i \in C_k$ and $k* \neq k$
$test \leftarrow 1$
$C_{k*} \leftarrow C_{k*} \cup \{i\}$
$C_k \leftarrow C_k \backslash \{i\}$
(d) If $test = 0$ END, else go to (b)

4 Experimental evaluation with artificial data sets

The adaptive dynamic cluster method based on L_1 distance has been studied by Diday and Govaert (1977). As an example of the adaptive and non-adaptive dynamic cluster based on L_r distances, we consider here the comparison between the adaptive and the non adaptive methods for the case of L_2 distance. To accomplish this comparison we cluster quantitative data sets scattered in $\Re^2$ using both methods and we evaluate the clustering results based on a Monte Carlo experience.

The basic data set considered here is described in Diday and Govaert (1977). It has 150 points scattered among three clusters of size 50 and unequal shapes: two clusters with ellipsis shapes and one cluster with spherical shape. Figure 1 shows an example of this data set. The data points of each cluster were drawn according to a bi-variate normal distribution with correlated components according to the following parameters:

a) Class 1: $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 4$, $\sigma_{12} = 1.7$, $\sigma_2^2 = 1$ and $\rho_{12} = 0.85$;
b) Class 2: $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1^2 = 0.25$, $\sigma_{12} = 0.0$, $\sigma_2^2 = 0.25$ and $\rho_{12} = 0.0$;
c) Class 3: $\mu_1 = 4$, $\mu_2 = 3$, $\sigma_1^2 = 4$, $\sigma_{12} = -1.7$, $\sigma_2^2 = 1$ and $\rho_{12} = -0.85$;

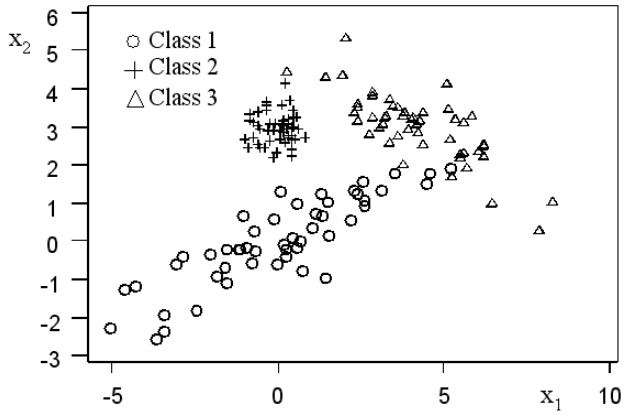

Fig. 1. Quantitative data set showing three clusters

The clustering results are evaluated based on a external index in the framework of a Monte Carlo experience with 100 replications. In each replication a L_2 clustering method (non-adaptive or adaptive) is run (until the convergence to a stationary value of the partitioning criterion) 50 times and the best result, according to the partitioning criterion is selected.

The average of the corrected Rand (CR) index (Hubert and Arabie (1985)) among these 100 replications is calculated. The CR index assesses the degree of agreement (similarity) between an a priori partition and a partition furnished by the clustering algorithm. The CR index takes values in the interval [-1,1], where the value 1 indicates a perfect agreement between the partitions, whereas values near 0 (or negatives) correspond to cluster agreements found by chance (Milligan (1996)).

The CR indices according to the clustering results are 0.64 and 0.61 for the methods with adaptive and non-adaptive L_2 distances , respectively. The comparison between these results is achieved by a paired Students' t-tests at a 5% significance level. The observed value of the test statistic following a t Student distribution with 99 degrees of freedom were 3.3. From this observed value, we reject the hypothesis that the average performance of the adaptive L_2 method is inferior to the non-adaptive L_2 method. The results of this experiment show that the performance of the adaptive methods is superior to the non-adaptive method at least for this data set.

5 Conclusion

Finally, our approach proposes a framework which permit to generalize easily the dynamic cluster method for the case of the adaptive and non-adaptive L_r distances. If r is equal to 1 and 2 we rediscover the usual exemplars (median and mean, respectively) of the clusters but now the difficulty is to find a

realistic interpretation for the cluster representatives when $r > 2$. Moreover, the adaptive dynamic cluster method based on L_1 distance has been studied by Diday and Govaert (1977). In this work we accomplished a similar study concerning the adaptive dynamic cluster method based on L_2 distance. In this study, the accuracy of the results furnished by the adaptive and non-adaptive methods for the L_2 distance have been assessed by an external index in the framework of a Monte Carlo experience. These results clearly show that the adaptive method outperforms the non-adaptive one concerning the quality of the clusters as measured by the corrected Rand index.

Acknowledgments: The authors would like to thank CNPq (Brazilian Agency) for its financial support.

References

Bock, H. H. (1993). Classification and Clustering: Problems for the Future. In *New Approaches in Classification and Data Analysis* (E. D. et al, ed.), 3–24. IFCS93, Springer.

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

Diday, E. and Simon, J. J. (1976). Clustering Analysis. In *Digital Pattern Recognition* (K. S. Fu, ed.), 47–94.

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

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

Govaert, G. (1975). *Classification automatique et distances adaptatives*. Ph.D. dissertation, Thèse de 3ème cycle, Mathématique appliquée, Université Paris VI.

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification* **2** 193–218.

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

Milligan, G. W. (1996). Clustering Validation: results and implications for applied analysis. In *Clustering and Classification* (P. A. et al, ed.), 341–375. Word Scientific, Singapore.

Spaeth, H. (1980). *Cluster analysis algorithms*. Wiley, New York.