Dynamical clustering of interval data : optimization of an adequacy criterion based on Hausdorff distance

Marie Chavent[1] and Yves Lechevallier[2]

[1] Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux 1 - 351, Cours de la libération,
33405 Talence Cedex, France

[2] INRIA- Institut National de Recherche en Informatique et en Automatique,
Domaine de Voluceau- Rocquencourt B.P. 105, 78153 Le Chesnay Cedex, France

Abstract. In order to extend the dynamical clustering algorithm to interval data sets, we define the prototype of a cluster by optimization of a classical adequacy criterion based on Hausdorff distance. Once this class prototype properly defined we give a simple and converging algorithm for this new type of interval data.

1 Introduction

The main aim of this article is to define a dynamical clustering algorithm for data tables where each cell contains an interval of real values (Table 1 for instance). This type of data is a particular case of a symbolic data table where each cell can be an interval, a set of categories or a frequency distribution (Diday (1988), Bock and Diday (2000)).

	Pulse Rate	Systolic pressure	Diastolic pressure
1	[60,72]	[90,130]	[70,90]
2	[70,112]	[110,142]	[80,108]
3	[54,72]	[90,100]	[50,70]
4	[70,100]	[130,160]	[80,110]
5	[63,75]	[60,100]	[140,150]
6	[44,68]	[90,100]	[50,70]

Table 1. A data table for $n = 6$ patients and $p = 3$ interval variables

Dynamical clustering algorithms (Diday (1971), Diday and Simon (1976)) are iterative two steps relocation algorithms involving at each iteration the identification of a prototype (or center) for each cluster by optimizing an adequacy criterion. The k-means algorithm with class prototypes 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 prototypes equal to cluster centers of gravity (MacQueen (1967), Späth (1980)).

In dynamical clustering, the optimization problem is the following. 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 $x_i \in \Re^p$. The problem is to find the partition $P = (C_1, ..., C_K)$ of Ω in K clusters and the system $Y = (y_1, ..., y_K)$ of class prototypes, optimum with respect to a partitioning criterion $g(P, Y)$. Two classical partitioning criteria are:

$$g(P,Y) = \sum_{k=1}^{K} \sum_{i \in C_k} d^2(x_i, y_k) \tag{1}$$

where $d(x, y) = ||x - y||_2$ is the L_2 distance, and:

$$g(P,Y) = \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, y_k) \tag{2}$$

where $d(x, y) = ||x - y||_1$ is the L_1 distance.

More precisely, the dynamical clustering algorithm converges and the partitioning criterion decreases at each iteration if the class prototypes are properly defined at each 'representation' step. Indeed, the problem is to find the prototype y of each cluster $C \subset \{1, ..., n\}$ which minimizes an adequacy criterion $f(y)$ measuring the "dissimilarity" between the prototype y and the cluster C. The two adequacy criteria corresponding to the partitioning criteria (1) and (2) are respectively:

$$f(y) = \sum_{i \in C} d^2(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} (x_i^j - y^j)^2 \tag{3}$$

and:

$$f(y) = \sum_{i \in C} d(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} |x_i^j - y^j| \tag{4}$$

The coordinates of the class prototype y minimizing criterion (3) are:

$$y^j = mean\{x_i^j \mid i \in C\} \tag{5}$$

and the coordinates of the class prototype y minimizing criterion (4) are:

$$y^j = median\{x_i^j \mid i \in C\} \tag{6}$$

In this latter case, the solution y^j is not always unique. If there is an interval of solutions, we usually choose y^j as the midpoint of this interval.

In this paper, we define the prototype y of a cluster C in the particular case of p-dimensional interval data. Each object i is now described on each variable j by an interval

$$x_i^j = [a_i^j, b_i^j] \in I = \{[a,b] \mid a, b \in \Re\ ,\ a \leq b\}$$

and the coordinates of the class prototype y are also intervals of I noted $y^j = [\alpha^j, \beta^j]$. In other words, the vector x_i representing an object i and the class prototype y are vectors of intervals, i.e., (hyper-)rectangles in the euclidean space $\Re^p$.
The distance d between two vectors of intervals x_i and $x_{i'}$ will be based on the Hausdorff distance between two sets. This distance is given section 2. Then we focus on the optimization problem for class prototypes and on its solution in section 3. Once the new class prototypes properly defined, a dynamical clustering algorithm of interval data is presented in section 4.

2 A distance measure between two vectors of intervals

There are several methods for measuring dissimilarities between interval data or more generally between symbolic objects (Chapters 8 and 11.2.2 of Bock and Diday (2000), De Carvalho (1998), Ichino and Yaguchi (1994)).

From our point of view, it is a natural approach to use Hausdorff distance, initially defined to compare two sets, to compare two intervals.

The Hausdorff distance d_H between two sets $A, B \in \Re^p$ is (Aubin, (1994)):

$$d_H(A,B) = \max(h(A,B), h(B,A)) \tag{7}$$

with

$$h(A,B) = \sup_{a \in A} \inf_{b \in B} ||b - a|| \tag{8}$$

By using L_2 norm in (8), the Hausdorff distance d_H between two intervals $A_1 = [a_1, b_1]$ and $A_2 = [a_2, b_2]$ is:

$$d_H(A_1, A_2) = \max(|a_1 - a_2|, |b_1 - b_2|) \tag{9}$$

In this paper, the distance d between two vectors of intervals

$$x_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])$$

and

$$x_{i'} = ([a_{i'}^1, b_{i'}^1], ..., [a_{i'}^p, b_{i'}^p])$$

representing two objects i and i' is defined as the sum for $j = 1, ..., p$ of the Haussdorf distance (9) between the two intervals $[a_i^j, b_i^j]$ and $[a_{i'}^j, b_{i'}^j]$.

Finally, the distance d is defined by:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} d_H(x_i^j, x_{i'}^j) = \sum_{j=1}^{p} \max(|a_i^j - a_{i'}^j|, |b_i^j - b_{i'}^j|) \tag{10}$$

In the particular case of intervals reduced to single points, this distance is the well-known L_1 distance between two points of $\Re^p$.

3 The optimization problem for class prototype

As presented in the introduction, the prototype y of a cluster C is defined in dynamical clustering by optimizing an adequacy criterion f measuring the “dissimilarity” between the prototype and the cluster. Here, we search the vector of intervals y noted:

$$y = (y^1, ..., y^p) = ([\alpha^1, \beta^1], ..., [\alpha^p, \beta^p])$$

which minimizes the following adequacy criterion:

$$f(y) = \sum_{i \in C} d(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} d_H(x_i^j, y^j) \tag{11}$$

where d is the distance between two vectors of intervals given in (10).

The criterion (11) can also be written:

$$f(y) = \sum_{j=1}^{p} \overbrace{\sum_{i \in C} d_H(x_i^j, y^j)}^{\tilde{f}(y^j)} \tag{12}$$

and the problem is now to find for $j = 1, ..., p$ the interval $y^j = [\alpha^j, \beta^j]$ which minimizes:

$$\tilde{f}(y^j) = \sum_{i \in C} d_H(x_i^j, y^j) = \sum_{i \in C} \max(|\alpha^j - a_i^j|, |\beta^j - b_i^j|) \tag{13}$$

We will see how to solve this minimization problem by transforming it into two well-known L_1 norm problems.

Let m_i^j be the midpoint of an interval $x_i^j = [a_i^j, b_i^j]$ and l_i^j be an half of its length:

$$m_i^j = \frac{a_i^j + b_i^j}{2}$$
$$l_i^j = \frac{b_i^j - a_i^j}{2}$$

and let μ^j and λ^j be respectively the midpoint and the half-length of the interval $y^j = [\alpha^j, \beta^j]$. According to the following property defined for x and y in $\Re$:

$$\max(|x - y|, |x + y|) = |x| + |y| \tag{14}$$

the function (13) can be written:

$$\tilde{f}(y^j) = \sum_{i \in C} \max(|(\mu^j - \lambda^j) - (m_i^j - l_i^j)|, |(\mu^j + \lambda^j) - (m_i^j + l_i^j)|)$$

$$= \sum_{i \in C} \max(|(\mu^j - m_i^j) - (\lambda^j - l_i^j)|, |(\mu^j - m_i^j) + (\lambda^j - l_i^j)|)$$
$$= \sum_{i \in C} (|\mu^j - m_i^j| + |\lambda^j - l_i^j|) = \sum_{i \in C} (|\mu^j - m_i^j| + \sum_{i \in C} |\lambda^j - l_i^j|) \quad (15)$$

This yields two well-known minimization problems in L_1 norm: Find $\mu^j \in \Re$ which minimizes:

$$\sum_{i \in C} |\mu^j - m_i^j| \quad (16)$$

and find $\lambda^j \in \Re$ which minimizes:

$$\sum_{i \in C} |\lambda^j - l_i^j| \quad (17)$$

The solutions $\hat{\mu^j}$ and $\hat{\lambda^j}$ are respectively the median of $\{m_i^j, i \in C\}$, the midpoints of the intervals $x_i^j = [a_i^j, b_i^j]$, $i \in C$, and the median of the set $\{l_i^j, i \in C\}$ of their half-lengths. Finally, the solution $\hat{y^j} = [\hat{\alpha^j}, \hat{\beta^j}]$ is the interval $[\hat{\mu^j} - \hat{\lambda^j}, \hat{\mu^j} + \hat{\lambda^j}]$.

4 The dynamical clustering algorithm

Iterative algorithms or dynamical clustering methods for symbolic data have already been proposed in Bock (2001), De Carvhalo et al. (2001) and Verde et al (2000). Here, we consider the problem of clustering a set $\Omega = \{1, ..., i, ..., n\}$ of n objects into K disjoint clusters $C_1, ..., C_K$ in the particular case of objects described on each variable j by an interval $x_i^j = [a_i^j, b_i^j]$ of $\Re$.

The dynamical clustering algorithm search for the partition $P = (C_1, ..., C_K)$ of Ω and the system $Y = (y_1, ..., y_K)$ of class prototypes which are optimum with respect to the following partitioning criterion based on the distance d defined in (10):

$$g(P, Y) = \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, y_k)$$
$$= \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \max(|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|) \quad (18)$$

This algorithm proceeds like classical dynamical clustering by iteratively repeating an 'allocation' step and a 'representation' step.

4.1 The 'representation step'

During the 'representation' step, the algorithm computes for each cluster C_k the prototype y_k which minimizes the adequacy criterion given in (11). We have defined in section 3 the 'optimal' prototype y_k for this criterion. It is described on each variable j by the interval $[\alpha_k^j, \beta_k^j] = [\mu_k^j - \lambda_k^j, \mu_k^j + \lambda_k^j]$ where:

$$\mu_k^j = median\{m_i^j \mid i \in C_k\} \tag{19}$$

is the median of the midpoints of the intervals $[a_i^j, b_i^j]$ with $i \in C_k$ and

$$\lambda_k^j = median\{l_i^j \mid i \in C_k\} \tag{20}$$

is the median of their half-lenghts.

4.2 The 'allocation' step

During the 'allocation step', the algorithm performes a new partition by reassigning each object i to the closest class prototype y_{k*} where:

$$k* = arg \min_{k=1,\ldots,K} d(x_i, y_k)$$

and d is defined in (10).

4.3 The algorithm

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 $y_1, ..., y_K$ among Ω and assign each object i to the closest prototype y_{k*} ($k* = arg \min_{k=l,\ldots,K} \sum_{j=1}^{p} max(|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|)$ to contruct the initial partition $(C_1, \ldots, C_K)$.

(b) 'Representation' step
For k in 1 to K compute the prototype $y_k = (y_k^1, ..., y_k^p)$ with $y_k^j = [\alpha_k^j, \beta_k^j] = [\mu_k^j - \lambda_k^j, \mu_k^j + \lambda_k^j]$ and:

$$\mu_k^j = median\{m_i^j \mid i \in C_k\}$$
$$\lambda_k^j = median\{l_i^j \mid i \in C_k\}$$

(c) 'Allocation' step
$test \leftarrow 0$
For i in 1 to n do

```
    define the cluster $C_{k*}$ such that
        $k* = arg \min_{k=l,\ldots,K} \sum_{j=1}^{p} max(|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|)$
    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)
```

5 Conclusion

We have proposed a dynamical clustering algorithm for interval data sets. The convergence of the algorithm and the decrease of the partitioning criterion at each iteration, is due to the optimization of the adequacy criterion (11) at each 'representation' step. The implementation of this algorithm is simple and the computationnal complexity is in $nlog(n)$.

References

AUBIN, J.P. (1994): *Initiation à l'analyse appliquée*, Masson.

BOCK H.H. (2001): Clustering algorithms and kohonen maps for symbolic data. *Proc. ICNCB*, Osaka, 203–215.

BOCK H.H. and DIDAY, E. (eds.) (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg.

DE CARVALHO, F.A.T. (1998): Extension based proximities coefficients between boolean symbolic objects. In: C. Hayashi et al. (eds): *Data Science, Classification and Related Methods*, Springer Verlag, 370–378.

DE CARVALHO, F.A.T, DE SOUZA, R.M., VERDE, R. (2001): Symbolic classifier based on modal symbolic descriptions. *Proc. CLADAG2001*, Univ. de Palermo.

DIDAY, E. (1971): La méthode des nuées dynamiques. *Rev. Stat. Appliquées* , *XXX (2)*, 19–34.

DIDAY, E. (1988): The symbolic approach in clustering and related methods of data analysis: The basic choice. In: H.H. Bock (ed.): *Classification and related methods of data anlysis. Proc. IFCS-87*, North Holland, Amsterdam, 673-684.

DIDAY, E., and SIMON, J.C. (1976): Clustering analysis. In: K.S. Fu (ed.): *Digital Pattern Clasification.* Springer Verlag, 47–94.

ICHINO, M. and YAGUCHI, H. (1994): Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics, 24 (4)*, 698–708.

MACQUEEN, J. (1967): Some methods for classification and analysis of multivariate observations. In: L.M. LeCam et al. (eds.): *Proc. 5th Berkeley Symp. on Math. Stat. Proba.*, University of California Press, Los Angeles, vol 1, 281–297.

SPÄTH, H. (1980): *Cluster analysis algorithms*, Horwood Publishers/Wiley, New York.

VERDE, R., DE CARVALHO, F.A.T., LECHEVALLIER, Y. (2000): A dynamical clustering algorithm for multi-nominal data. In: H.A.L. Kiers et al. (eds.): *Data Analysis, Classification and Related methods*. Springer Verlag, 387–394.