An Hausdorff distance between hyper-rectangles for clustering interval data

Marie Chavent

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux 1 - 351, Cours de la libération,
33405 Talence Cedex, France
chavent@math.u-bordeaux.fr

Summary. The Hausdorff distance between two sets is used in this paper to compare hyper-rectangles. An explicit formula for the optimum class prototype is found in the particular case of the Hausdorff distance for the L_∞ norm. When used for dynamical clustering of interval data, this prototype will ensure that the clustering criterion decreases at each iteration.

1 Introduction and notations

Symbolic Data Analysis (SDA) deals with data tables where each cell is not only a single value but also an interval of values, a set of categories or a frequency distribution. SDA generalizes well-known methods of multivariate data analysis to this new type of data representations (Diday, 1988), (Bock and Diday, 2000).

Throughout this paper, we consider the problem of clustering a set $\Omega = \{1, ..., i, ..., n\}$ of n objects into K disjoint clusters $\{C_1, ..., C_K\}$ by dynamical clustering (Diday and Simon, 1976). Iterative algorithms or dynamical clustering methods for symbolic data have already been proposed in Bock (2001), Verde et al. (2000).

Here, we consider the particular case of objects i described on each variable j by an interval $x_i^j = [a_i^j, b_i^j]$ of $\Re$. In other words an object i is an hyper-rectangle in the euclidean space $\Re^p$ noted:

$$x_i = \prod_{j=1}^{p} \underbrace{[a_i^j, b_i^j]}_{x_i^j}$$

In: Classification, Clustering an Data Mining Application, D. Banks et al. (Eds.), Proceedings of the IFCS'04, Springer, 333-340 (2004)

In dynamical clustering, the prototype y of a cluster C is defined by optimizing an adequacy criterion f measuring the "dissimilarity" between the prototype and the cluster. In the particular case of interval data, this prototype is an hyper-rectangle (see Fig. 1).

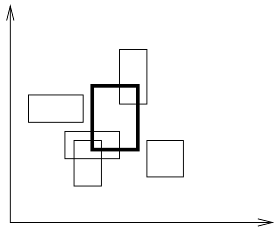

Fig. 1. A prototype y (thick line) of a set of rectangles (thin lines)

Here, the distance chosen to compare two p-dimensional hyper-rectangles is the Hausdorff distance d_H. This distance, defined two compare two sets of objects, depends on the distance chosen two compare two objects, here two points of $\Re^p$. In the particular case where the Hausdorff distance is based on the L_∞ distance in $\Re^p$, we are able to give an explicit formula for the prototype which minimizes:

$$f(y) = \max_{i \in C} d_H(x_i, y) \tag{1}$$

In the case of Hausdorff distances based on Euclidean or Manhattan distance between points, explicit formulas seem to be more difficult to find.

Chavent and Lechevallier (2002), give however an explicit formula of the prototype $\hat{y}$ which minimizes the adequacy criterion:

$$f(y) = \sum_{i \in C} d(x_i, y) \tag{2}$$

where d is not the Haudorff distance between two hyper-rectangles but the sum on each variable j of the one-dimensional Hausdorff distance d_H between two intervals.

In practice, the hyper-rectangle prototype defined in this article will probably be more sentitive to extreme values than the one defined in Chavent and Lechevallier (2002) but the distance used is a "real" Hausdorff distance on $\Re^p$-set.

2 The L_∞ Hausdorff distance between two hyper-rectangles

The Hausdorff distance (Nadler, 1978), (Rote, 1991), often used in image processing (Huttenlocher et al., 1993), is defined to compare two sets A and

B of objects. This distance depends on the distance d chosen to compare two objects u and v respectively in A and B.
We consider here, the L_∞ distance d_∞ between two points u and v of $\Re^p$:

$$d_\infty(u,v) = \max_{j=1,\ldots,p} |u_j - v_j| \tag{3}$$

and call "L_∞ Hausdorff distance" the Hausdorff distance associated to d_∞. Given A and B two hyper-rectangles of $\Re^p$ noted:

$$A = \prod_{j=1}^{p} A_j,\ B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\Re$, the L_∞ Hausdorff distance $d_{H,\infty}$ between A and B is defined by:

$$d_{H,\infty}(A,B) = \max(h_\infty(A,B), h_\infty(B,A)) \tag{4}$$

where

$$h_\infty(A,B) = \sup_{u\in A} \inf_{v\in B} d_\infty(u,v) \tag{5}$$

Remark 1. In the one dimensional case i.e. $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$, we can drop the ∞ subscript:

$$h(A_j, B_j) = \sup_{u_j\in A_j} \inf_{v_j\in B_j} |u_j - v_j| \tag{6}$$

and formula (4) simplifies to:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|) \tag{7}$$

This remark will be used in the proof of properties 1 and 2, which are the basis for the explicit formulas of the optimum class prototype in section 3.

Property 1. With the L_∞ distance, we have the following relation between the asymetrical functions h in p dimensions and in one dimension:

$$h_\infty(A,B) = \max_{j=1,\ldots,p} h(A_j, B_j) \tag{8}$$

Proof:

$$\begin{aligned}
h_\infty(A,B) &= \sup_{u\in A}\{ \inf_{v\in B} \max_{j=1,\ldots,p} |u_j - v_j| \} \\
&= \sup_{u\in A}\{ \max_{j=1,\ldots,p} \{ \inf_{v_1\in B_1} |u_1 - v_1|, ..., \inf_{v_p\in B_p} |u_p - v_p| \}\} \\
&= \max_{j=1,\ldots,p} \underbrace{\sup_{u_j\in A_j} \inf_{v_j\in B_j} |u_j - v_j|}_{h(A_j,B_j)}
\end{aligned}$$

Property 2. With the L_∞ distance, we have the following relation between the Haudorff distances d_H in p dimensions and in one dimension:

$$d_{H,\infty}(A,B) = \max_{j=1,\ldots,p} d_H(A_j,B_j) \tag{9}$$

Proof: From (8) we have :

$$\begin{aligned} h_\infty(A,B) &= \max_{j=1,\ldots,p} h(A_j,B_j) \\ h_\infty(B,A) &= \max_{j=1,\ldots,p} h(B_j,A_j) \end{aligned}$$

Then:

$$\begin{aligned} d_{H,\infty}(A,B) &= \max\{h_\infty(A,B), h_\infty(B,A)\} \\ &= \max_{j=1,\ldots,p} \max\{h(A_j,B_j), h(B_j,A_j)\} \\ &= \max_{j=1,\ldots,p} \underbrace{\max\{|a_j-\alpha_j|, |b_j-\beta_j|\}}_{d_H(A_j,B_j)} \end{aligned}$$

3 The optimum class prototype

We denote by y and $x_i \in C$ the hyper-rectangles which describe respectively the prototype and an object in cluster C:

$$\begin{aligned} y &= \prod_{j=1}^{p} \underbrace{[\alpha^j, \beta^j]}_{y^j} \\ x_i &= \prod_{j=1}^{p} \underbrace{[a_i^j, b_i^j]}_{x_i^j} \end{aligned}$$

We measure the “dissimilarity” between the prototype y and the cluster C by mean of the function f defined in (1), in the particular case of the L_∞ Hausdorff distance (4):

$$f(y) = \max_{i\in C} d_{H,\infty}(x_i, y) \tag{10}$$

We define our prototype $\hat{y}$ as an hyper-rectangle which minimizes f. We see that:

$$\begin{aligned} f(y) &= \max_{i\in C} \max_{j=1,\ldots,p} d_H(x_i^j, y^j) && (11) \\ &= \max_{j=1,\ldots,p} \underbrace{\max_{i\in C} d_H(x_i^j, y^j)}_{\tilde{f^j}(y^j)} && (12) \end{aligned}$$

The equality (11) is due to property 2.

Denote now $\hat{y}^j$ the minimizer of $\tilde{f}^j$ (see (12)), for $j = 1, ..., p$. Obviously, $\hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f, but for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions. Hence, the minimizer of f is not unique.

In the sequel, we will use the minimizer $\hat{y} = \prod_{j=1}^{p} \hat{y}^j$, computable by the following explicit formulas (17) and (18).

We know from (12) and (7) that:

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

i.e:

$$\tilde{f}^j(y^j) = max\{\max_{i \in C} |a_i^j - \alpha^j|, \max_{i \in C} |b_i^j - \beta^j|\} \tag{14}$$

i.e. minimizing $\tilde{f}^j$ is equivalent to:

$$\min_{\alpha^j \in \Re} \max_{i \in C} |a_i^j - \alpha^j| \tag{15}$$

and

$$\min_{\beta^j \in \Re} \max_{i \in C} |b_i^j - \beta^j| \tag{16}$$

The solutions $\hat{\alpha}^j$ and $\hat{\beta}^j$ are:

$$\hat{\alpha}^j = \frac{\max_{i \in C} a_i^j + \min_{i \in C} a_i^j}{2} \tag{17}$$

$$\hat{\beta}^j = \frac{\max_{i \in C} b_i^j + \min_{i \in C} b_i^j}{2} \tag{18}$$

An example of the construction of this optimum prototype $\hat{y}$ is given Fig.2 and Fig.3.

4 Application to dynamical clustering

Dynamical clustering algorithm proceeds by iteratively determining the K class prototypes y_k and then reassigning all objects to the closest class prototype. The advantage of using the L_∞ Hausdorff distance and the adequacy criterion defined above is that we can get explicit and simple formulas for the prototypes. As for the convergence of the algorithm, the prototypes which minimize the adequacy criterion, ensure the decrease of any of the two clustering criteria (19) and (20), independantly of the choice of the minimizer.

$$g(\{C_1, ..., C_K\}) = \sum_{k=1}^{K} \max_{i \in C_k} d_{H,\infty}(x_i, y_k) \tag{19}$$

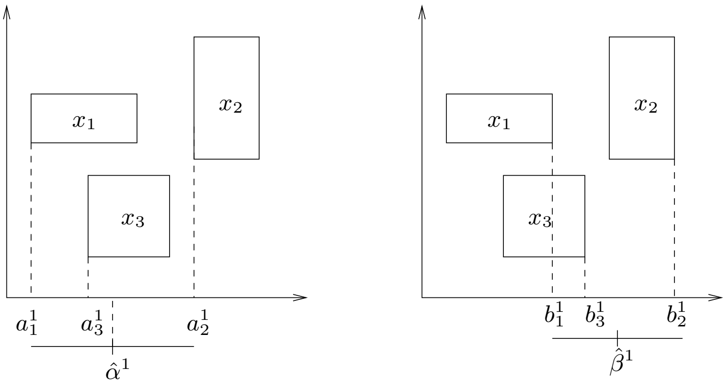

Fig. 2. Construction of a prototype $\hat{y}$ for $j = 1$ i.e. $\hat{y}^1 = [\hat{\alpha}^1, \hat{\beta}^1]$

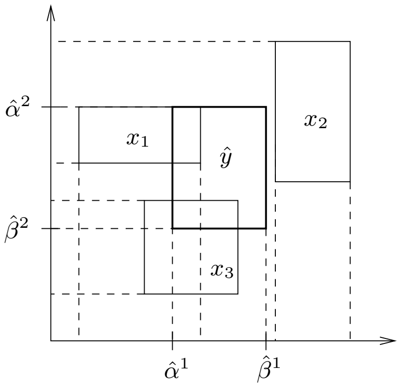

Fig. 3. Construction of a prototype $\hat{y}$

$$g(\{C_1, ..., C_K\}) = \max_{k=1}^{K} \max_{i \in C_k} d_{H,\infty}(x_i, y_k) \tag{20}$$

Hence, according to Celeux et al. (1989), this implies the convergence of the algorithm.

Acknowledgements

The author thanks a referee for pointing out the non uniqueness of the minimizing hyper-rectangle.

References

Bock, H.-H. (2001). Clustering algorithms and kohonen maps for symbolic data. In *ICNCB Proceedings*, 203–215. Osaka.

Bock, H.-H. and Diday, E., eds. (2000). *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Studies in classification, data analysis and knowledge organisation. Springer Verlag, Heidelberg.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y. and Ralambondrainy, H. (1989). *Classification automatique des donnes*. Dunod.

Chavent, M. and Lechevallier, Y. (2002). Dynamical clustering of interval data. Optimization of an adequacy criterion based on hausdorff distance. In *Classification, Clustering, and Data Analysis* (K. Jajuga, A. Sokolowski and H.-H. Bock, eds.), 53–60. Sringer Verlag, Berlin.

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

Diday, E. and Simon, J. C. (1976). Clustering analysis. In *Digital Pattern Classification* (K. S. Fu, ed.), 47–94. Springer Verlag.

Huttenlocher, D. P., Klanderman, G. A. and Rucklidge, W. J. (1993). Comparing images using the Hausdorff Distance. *IEE Transaction on Pattern Analysis and Machine Intelligence* **15** 850–863.

Nadler, S. B. J. (1978). *Hyperspaces of sets*. Marcel Dekker, Inc., New York.

Rote, G. (1991). Computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters* **38** 123–127.

Verde, R., De Carvalho, F. A. T. and Lechevallier, Y. (2000). A dynamical clustering algorithm for multi-nominal data. In *Data Analysis, Classification and Related methods* (H. A. L. K. et al., ed.), 387–394. Springer Verlag.