Normalized k-means clustering of hyper-rectangles

Marie Chavent

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux 1 - 351, Cours de la libération,
33405 Talence Cedex, France
(e-mail: `chavent@math.u-bordeaux1.fr`)

Abstract. Interval variables can be measured on very different scales. We first remind a general methodology used for measuring the dispersion of a variable from an optimal center and we define two measures of dispersions associated to two optimal "centers" for interval variables. Then we study the relations between the standardization of a data table and the use in clustering of a normalized distance. Finally we define two normalized distances between hyper-rectangles and their use in two normalized k-means clustering algorithms.
Keywords: Interval data, Standardization, Normalized Hausdorff distance, Clustering.

1 Introduction

A classical quantitative data table $(x_i^j)_{n \times p}$ describes n objects $\{1, ..., i, ..., n\}$ by p quantitative variables $\{1, ..., j, ..., p\}$ which may be defined on different scales. This phenomenon is measured by the dispersion (standard deviation, range, percentile ranges...) of each variable.

Dealing with variables measured on very different scales is a problem when comparing two objects globally on all the variables. For instance the Euclidean distance or more generaly the L_p-distance will give implicitly more importance to variables of strong dispersion and the comparison between objects will only reflect their differences on those variables. This phenomenon has then an incidence on the clustering into classes of homogeneous objects (i.e. objects highly similar to each other): only variables with strong dispersion will have an important contribution in the construction of clusters. A natural way to avoid this effect is either to normalize the data table or to use normalized distances.

Recently, several clustering methods have been proposed in the field of symbolic data analysis [Diday, 1988], [Bock and Diday, 2000]. Several works on k-means clustering of interval data sets have been published [Bock, 2001], [Chavent and Lechevallier, 2002], [De Carvalho *et al.*, 2003], [Chavent *et al.*, 2003], [De Souza and De Carvalho, 2004] and [Chavent, 2004].

The problem of the standardization of this new type of data is now naturally arising. In [Chavent, 1997], the symbolic data set was not directly

normalized but normalized distances between symbolic objects were used:

$$d(i,i') = (\sum_{j=1}^{p} \frac{1}{(\sigma^j)^\alpha} d(x_i^j, x_{i'}^j)^\alpha)^{1/\alpha} \tag{1}$$

where d was a measure of comparison between two symbolic descriptions (two intervals for instance) and σ^j a measure of dispersion of a variable j defined by:

$$\sigma^j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(x_i^j, x_{i'}^j) \tag{2}$$

The use of a double sum in [2] was not really appropiate for computing σ^j on voluminous data sets.

This question of the standarization of symbolic data has also been clearly raised for interval data in [De Carvalho et $al.$, 2003] where the authors proposed measures of dispersion based on the dispersion of the centers, the lower bounds or the upper bounds of the intervals.

In [Chavent and Lechevallier, 2002] and [Chavent, 2004], two k-means clustering algorithms of hyper-rectangles with Hausdorff distances where proposed. The idea here is to use the explicit formula of the optimum class prototype given in those two papers in order to define two "mean" intervals optimizing two measures of dispersion (see section 2). Those two measures of dispersion are called the "star" and the "radius" of an interval variable (see sections 2.1 and 2.2). After a few words on the relation between standardizing an interval data table and using a normalized distance between hyper-rectangles (see section 3), the two k-means algorithms given in [Chavent and Lechevallier, 2002] and in [Chavent, 2004] are "normalized" (see section 4).

In the rest of this paper we will consider an interval data table $(x_i^j)_{n \times p}$ where each object i is described for each variable j by an interval

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

Each object i is then described by an hyper-rectangle of $\Re^p$:

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

2 Measure of centrality and dispersion

For a classical quantitative variable j the mean squared deviation measures the dispersion from the mean $\bar{x}^j$ which is the optimal solution $\hat{y}$ of the following minimization problem:

$$\min_{y \in \Re} \sum_{i=1}^{n} (x_i^j - y)^2 = \min_{y \in \Re} \underbrace{\sum_{i=1}^{n} d^2(x_i^j, y)}_{f(y)} \tag{3}$$

In the same way the mean absolute deviation measures the dispersion from the median x_M^j which is the optimal solution $\hat{y}$ of the following minimization problem:

$$\min_{y \in \Re} \sum_{i=1}^{n} |x_i^j - y| = \min_{y \in \Re} \underbrace{\sum_{i=1}^{n} d(x_i^j, y)}_{f(y)} \qquad (4)$$

In both cases, $f(\hat{y})$ is a measure of dispersion.

For an interval variable j we have $x_i^j = [a_i^j, b_i^j]$ and the "measures" of centrality are not real values like the mean or the median values but an interval of values noted $y = [\alpha, \beta]$. We have seen that the mean and the median are optimal centers of two different dispersion measures f. Our aim is then to define optimal centers $\hat{y} = [\hat{\alpha}, \hat{\beta}]$ for functions f chosen to measure the dispersion. Those functions are based on a distance d between intervals.

The distance chosen here to compare two intervals is the Hausdorff distance. This set-distance d_H is simplified in the particular case of two intervals to:

$$d_H([a_i^j, b_i^j], [a_{i'}^j, b_{i'}^j]) = \max(|a_i^j - a_{i'}^j|, |b_i^j - b_{i'}^j|) \qquad (5)$$

In the next sections we will define two different optimal "centers" $\hat{y} = [\hat{\alpha}, \hat{\beta}]$ and two different measures of dispersion $f(\hat{y})$.

2.1   The "star"

We consider the following measure of dispersion from $\hat{y}$:

$$f(\hat{y}) = \sum_{i=1}^{n} d_H(x_i^j, \hat{y}) \qquad (6)$$

where d_H is the Hausdorff distance between the intervals x_i^j and $\hat{y}$ and where $\hat{y}$ is defined by:

$$\hat{y} = arg \min_{y \in I} \sum_{i=1}^{n} d_H(x_i^j, y) \qquad (7)$$

We use a result of [Chavent and Lechevallier, 2002] to define an explicit formula for the optimal "central" interval $\hat{y} = [\hat{\alpha}, \hat{\beta}]$: by a simple rewriting of the intervals $x_i^j = [a_i^j, b_i^j]$ according to their middle point m_i^j and their half-length l_i^j, the authors proved that the middle point $\hat{\mu}$ and the half-length $\hat{\lambda}$ of the interval $\hat{y}$ minimizing $\sum_{i=1}^{n} d_H(x_i^j, y)$ is:

$$\hat{\mu} = median\{m_i^j \mid i = 1, ..., n\} \qquad (8)$$
$$\hat{\lambda} = median\{l_i^j \mid i = 1, ..., n\} \qquad (9)$$

The following measure of dispersion σ^j is defined:

$$\sigma^j = \sum_{i=1}^{n} \max(|a_i^j - \hat{\mu} + \hat{\lambda}|, |b_i^j - \hat{\mu} - \hat{\lambda}|) \tag{10}$$

Because the formulation of f given in (6) is close to the measure of homogeneity of a cluster C called the "star":

$$\min_{i \in C} \sum_{j \in C} d_{ij}$$

we will call σ^j defined in (10) the "star" of the interval variable j.

2.2   The "radius"

We consider the following measure of dispersion from $\hat{y}$:

$$f(\hat{y}) = \max_{i=1...n} d_H(x_i^j, \hat{y}) \tag{11}$$

where d_H is once again the Hausdorff distance between the intervals x_i^j and y and where $\hat{y}$ is defined by:

$$\hat{y} = arg \min_{y \in I} \max_{i=1...n} d_H(x_i^j, y) \tag{12}$$

We use here a result of [Chavent, 2004] to define an explicit formula for the optimal "central" interval $\hat{y} = [\hat{\alpha}, \hat{\beta}]$: the author proved that the lower and upper bounds of interval $\hat{y}$ minimizing $\max_{i=1...n} d_H(x_i^j, y)$ are:

$$\hat{\alpha}^j = \frac{\max_{i=1...n} a_i^j + \min_{i=1...n} a_i^j}{2} \tag{13}$$

$$\hat{\beta}^j = \frac{\max_{i=1,...,n} b_i^j + \min_{i=1,...,n} b_i^j}{2} \tag{14}$$

The following measure of dispersion σ^j can then be defined:

$$\sigma^j = \max_{i=1...n} \max(|a_i^j - \hat{\alpha}^j|, |b_i^j - \hat{\beta}^j|) \tag{15}$$

Because the formulation of f given in (11) is close to the measure of homogeneity of a cluster C called the "radius":

$$\min_{i \in C} \max_{j \in C} d_{ij}$$

we will call σ^j defined in (15) the "radius" of the interval variable j.

3 Standardization, distance and clustering

For a classical quantitative data table $(x_i^j)_{n \times p}$, standardizing is a technique for removing location and scale attributes. The standardized variables z^j have mean equal to 0 and standard deviation equal to 1 when the variables x^j are centered by their mean $\bar{x}^j$ and normalized (reduced) by their standard deviation σ^j. The Euclidean distance between two objects i and i' of the standardized matrix $(z_i^j)_{n \times p}$ is then:

$$d(z_i, z_{i'}) = \sqrt{\sum_{j=1}^{p} (\frac{x_i^j - \bar{x}^j}{\sigma^j} - \frac{x_{i'}^j - \bar{x}^j}{\sigma^j})^2} \qquad (16)$$

$$= \sqrt{\sum_{j=1}^{p} \frac{1}{(\sigma^j)^2}(x_i^j - x_{i'}^j)^2} \qquad (17)$$

$$= d_M(x_i, x_{i'}) \qquad (18)$$

where d_M is the weighed Euclidean distance and $M = D_{1/\sigma^2}$. This weighed distance is also sometimes called the normalized Euclidean distance.

We can then notice that:

- the clustering obtained from the initial data table $(x_i^j)_{n \times p}$ is similar to the clustering obtained from the centered data table $(x_i^j - \bar{x}^j)_{n \times p}$ (because the distances are equal). Indeed we are not directly concerned in this article with the problem of centering interval data even if we have defined a "central" interval previously in this article.
- the clustering performed with the initial data table $(x_i^j)_{n \times p}$ and the normalized Euclidean distance d_M is similar to the clustering performed with the standardized (or simply normalized) data table $(z_i^j)_{n \times p}$ and the "simple" Euclidean distance.

We have of course the same kind of results with the Minkowsky distance.

The questions are now: do we have the same kind of results for interval data ? Is it equivalent to "normalize" the intervals $x_i^j = [a_i^j, b_i^j]$ and to use a "normalized" distance ? What does "normalizing" an interval or "normalizing" a distance between hyper-rectangles mean ?

Here we will try to answer those questions in the particular case of two distances between hyper-rectangles of $\Re^p$ used in [Chavent and Lechevallier, 2002] and [Chavent, 2004]. We consider

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

and

$$x_{i'} = \prod_{j=1}^{p} \underbrace{[a_{i'}^j, b_{i'}^j]}_{x_{i'}^j}$$

The first distance d_1 is not a real $\Re^p$-set Hausdorff distance but a sum of Hausdorff distances d_H between intervals:

$$d_1(x_i, x_{i'}) = \sum_{j=1}^{p} d_H(x_i^j, x_{i'}^j) \qquad (19)$$

The second distance d_2 is a real $\Re^p$-set Hausdorff distance called the L_∞-Hausdorff distance which can be written in the particular case of hyper-rectangles as a maximum of Hausdorff distances d_H between intervals:

$$d_2(x_i, x_{i'}) = \max_{j=1...p} d_H(x_i^j, x_{i'}^j) \qquad (20)$$

If we consider now that "normalizing" an interval $x_i^j = [a_i^j, b_i^j]$ consists in dividing its lower and upper bounds by the same measure of dispersion σ^j, the "normalized" interval of x_i^j is $z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$.

The Hausdorff distance between two "normalized" intervals is then:

$$d_H(z_i^j, z_{i'}^j) = \max(|\frac{a_i^j}{\sigma^j} - \frac{a_{i'}^j}{\sigma^j}|, |\frac{b_i^j}{\sigma^j} - \frac{b_{i'}^j}{\sigma^j}|) = \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \qquad (21)$$

and the distances d_1 and d_2 between the two "normalized" hyper-rectangles z_i and $z_{i'}$ can then be written as:

$$d_1(z_i, z_{i'}) = \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \qquad (22)$$

and

$$d_2(z_i, z_{i'}) = \max_{j=1...p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \qquad (23)$$

The normalized distance is then defined for d_1 by:

$$d_1(x_i, x_{i'}) = ||\frac{(d_H(x_i^j, x_{i'}^j)_{j=1,...,p}}{\sigma^j}||_{L_1} \qquad (24)$$

and for d_2 by:

$$d_2(x_i, x_{i'}) = ||\frac{(d_H(x_i^j, x_{i'}^j)_{j=1,...,p}}{\sigma^j}||_{L_\infty} \qquad (25)$$

Finally, we have once again the result that the clustering performed with the initial interval data table $(x_i^j)_{n \times p}$ and the normalized distances d_1 or d_2 (given in (24) and 25)) is similar to the clustering performed with the "normalized" interval data table $(z_i^j)_{n \times p}$ and the "simple" distances d_1 or d_2 (given in (19) and (20)).

4   Normalized k-means of hyper-rectangles

Dynamical clustering [Diday and Simon, 1976] called here for simplification k-means clustering, proceeds by iteratively determining K class prototypes and then reassigning all objects to the closest class prototype. If the prototype $\hat{y}$ of a cluster C is properly defined by optimization of an adequacy criterion f (measuring the "dissimilarity" between the prototype and the cluster), the algorithm converges and the partitioning criterion decreases at each iteration.

For classical quantitative data, when the prototype $\hat{y}$ of a cluster C is the mean-vector, the adequacy criterion minimized is:

$$f(y) = \sum_{i \in C} d^2(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} (x_i^j - y^j)^2 \tag{26}$$

When a standardization is necessary, the columns x^j are usually normalized by $\sigma^j = \sqrt{\sum_{i=1}^{n}(x_i^j - \bar{x^j})^2}$ or the normalized Euclidean distance d_M with $M = D_{1/\sigma^2}$ is used. The adequacy criterion measured on $\Omega = \{1, ..., n\}$ is then equal to p, the number of variables.

In the same way when the prototype $\hat{y}$ of a cluster C is the median-vector x_m, the adequacy criterion minimized is:

$$f(y) = \sum_{i \in C} d(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} |x_i^j - y^j| \tag{27}$$

When a standardization is necessary, the columns x^j are normalized by $\sigma^j = \sum_{i=1}^{n} |x_i^j - x_m^j|$ or the normalized Euclidean distance d_M with $M = D_{1/\sigma}$ is used. The adequacy criterion measured on $\Omega = \{1, ..., n\}$ is then once again equal to p, the number of variables.

In the particular case of interval data the optimal prototype of a cluster is an hyper-rectangle. We can repeat the previous reasoning for "normalizing" any k-means clustering algorithm of hyper-rectangles when the prototypes are properly defined by optimization of an adequacy criterion. Here we use:

- the normalized distance (24) with σ^j the "star" defined in (10) for "normalizing" the k-means method of [Chavent and Lechevallier, 2002]
- the normalized distance (25) with σ^j the "radius" defined in (15) for "normalizing" the k-means method of [Chavent, 2004]

5   Conclusion

In this paper we have proposed a general approach for the "normalization" of dynamical clustering algorithms. We have seen that if the prototype of a

cluster is properly defined by optimization of an homogeneity criterion, this result can also be used to define a measure of dispersion and then to normalize either the data or the distances. We have applied this methodology in the particular case of two k-means clustering algorithms of hyper-rectangles. The first one uses a "star" homogeneity criterion and a distance between hyper-rectangles which is a sum of Hausdorff distances between intervals. The second one uses a "radius" homogeneity criterion and the L_∞ Hausdorff distance between hyper-rectangles. The two corresponding dispersion measures of interval variables called here the "star" and the "radius" are then simply used to "normalize" those two algorithms.

References

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

[Bock, 2001]H.-H. Bock. Clustering algorithms and Kohonen maps for symbolic data. In *ICNCB Proceedings*, pages 203–215, Osaka, 2001.

[Chavent and Lechevallier, 2002]M. Chavent and Y. Lechevallier. Dynamical clustering of interval data. Optimization of an adequacy criterion based on Hausdorff distance. In K. Jajuga, A. Sokolowski, and H.-H. Bock, editors, *Classification, Clustering, and Data Analysis*, pages 53–60, Berlin, 2002. Springer Verlag.

[Chavent *et al.*, 2003]M. Chavent, F.A.T. De Carvalho, Y. Lechevallier, and R. Verde. Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistiques Appliquées*, LI(4), 2003.

[Chavent, 1997]M. Chavent. *Analyse des données symboliques. Une méthode divisive de classification.* PhD thesis, Université Paris-IX Dauphine, 1997.

[Chavent, 2004]M. Chavent. An Hausdorff distance between hyper-rectangles for clustering interval data. In D. Banks and al., editors, *Classification, Clustering and Data Mining Applications*, pages 333–340. Springer, 2004.

[De Carvalho *et al.*, 2003]F.A.T. De Carvalho, P. Brito, and H.-H. Bock. Une méthode type nuées dynamiques pour les données symboliques quantitatives. In Y. Dodge and G. Melfi, editors, *Méthodes et perspectives en Classification*, pages 79–81. Presses Académiques Neuchâtel, 2003.

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

[Diday and Simon, 1976]E. Diday and J. C. Simon. Clustering analysis. In K. S. Fu, editor, *Digital Pattern Classification*, pages 47–94. Springer Verlag, 1976.

[Diday, 1988]E. Diday. The symbolic approach in clustering and related methods of data analysis: The basic choices. In H.-H. Bock, editor, *Classification and related methods of data analysis*, pages 673–684, Amsterdam, 1988. North Holland.