Hierarchical Clustering for Boxplot Variables

Javier Arroyo[1], Carlos Maté[2], and Antonio Muñoz-San Roque[2]

[1] Departamento de Sistemas Informáticos, Universidad Complutense de Madrid, Profesor García-Santesmases s/n, 28040 Madrid, Spain
[2] Instituto de Investigación Tecnológica, ETSI (ICAI), Universidad Pontificia Comillas, Alberto Aguilera 25, 28015 Madrid, Spain

Abstract. Boxplots are well-known exploratory charts used to extract meaningful information from batches of data at a glance. Their strength lies in their ability to summarize data retaining the key information, which also is a desirable property of symbolic variables. In this paper, boxplots are presented as a new kind of symbolic variable. In addition, two different approaches to measure distances between boxplot variables are proposed. The usefulness of these distances is illustrated by means of a hierarchical clustering of boxplot data.

1 Introduction

Symbolic data analysis (SDA) proposes an alternative approach to deal with large and complex datasets as it allows the summary of these datasets into smaller and more manageable ones retaining the key knowledge. In a symbolic dataset, items are described by symbolic variables such as lists of (categorical or quantitative) values, intervals and distributions (Bock and Diday (2000)). However, new types of symbolic variables can be considered.

We believe that boxplots (henceforth BPs) can be proposed as a new kind of quantitative symbolic variable, which have interesting features. BPs, developed by Tukey (1977) and also called box-and-whiskers plots, are an univariate data display of a summary of an empirical distribution which, conventionally, consists of five values: the extremes, the lower and upper quartiles (hinges) and the median. BPs are an extremely useful exploratory tool, which is widely used in data analysis. The reasons are given by Benjamini (1988):

- BPs offer an excellent summary for a batch of data available at a glance.
- Good compromise among detailed description and a condensed display of summaries.
- Easy implementation and computation.
- Allowing comparison between many batches of data.
- Easy explanation to non-statisticians.

These properties also explain why BPs can be considered as an interesting type of symbolic variable. The main feature of symbolic variables is that they provide more information than classical ones, where only one single number or category is allowed as value. They also are well suited for characterizing

the properties of classes obtained after an aggregation process, time-varying patterns or complex scenarios. BP variables fit in the symbolic context, as BPs represent empirical distributions in a shortened but explanatory way, reporting about their location, spread, skewness and normality.

BP variables offer an intermediate point between the simplicity of interval variables and the detailed information provided by histogram variables. Interval variables do not provide information about the central area of an empirical distribution, while BP variables do this by means of the three quartiles. In the other hand, histogram variables report in detail about the empirical distribution, but their structure is more complex: a set of consecutive intervals with associated weights is needed. The structure of the BP variable is simpler, but it is enough to report about the shape of the distribution. In addition, BP variables do not suffer from the subjectivity of histogram variables, that is, a BP is always defined by five definite components, while the components of the histogram depends on the analyst criterion.

This paper defines the concept of boxplot variable in the SDA framework,proposes two ways to measure distances between BP variables and illustrates the approach by a hierarchical clustering of BP data.

2 Definition of boxplot variable

Let Y be a variable defined for all elements of a finite set $E = \{1, ..., N\}$, Y is termed a BP variable with domain of values $\mathcal{Y}$, if Y is a mapping from E to a range $\mathcal{B}(\mathcal{Y})$ with the structure $\{m, q, Me, Q, M\}$, where $-\infty < m \leq q \leq Me \leq Q \leq M < \infty$, and m represents the minimum, q the lower quartile, Me the median, Q the upper quartile, and M the maximum.

This definition follow the style of the definitions of other symbolic variables given by Billard and Diday (2003). Equivalently, a BP variable can be considered as an special case of an interval-valued modal variable. Interval-valued modal variables are variables where each individual take as value one or more intervals, each one with a weight attached (weights usually represent frequency, probability or possibility). BP variables are a kind of modal interval-valued variable with the following restrictions: the number of intervals is always four, intervals are consecutive, the weights represent frequency or probability, and the weight of each interval is 0.25. Consequently, BP variables can also be defined by: a lower-whisker interval $\xi_1 = [m, q)$, a lower-mid-box interval $\xi_2 = [q, Me)$, an upper-mid-box interval $\xi_3 = [Me, Q)$, and an upper-whisker interval $\xi_4 = [Q, M)$, with $p_i = 0.25$, $i = 1, ..., 4$.

In a SDA context, BP variables can describe the quantitative properties of the classes obtained after an aggregation process, i.e. a BP variable summarizes the batch of values of a quantitative feature for each class. In these cases, determining the value of the elements describing each BP is needed.

Regarding the quartiles, Frigge et al (1989) show that different definitions are applied in the statistical software packages. It has been verified that the

lack of standardization still prevails nowadays. Frigge et al (1989) suggested the definition proposed by Tukey (1977). Tukey's definition for the lower quartile in terms of the ordered observations $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ is

$$q = (1-g)x_{(j)} + gx_{(j+1)}, \tag{1}$$

where $[(n+3)/2]/2 = j + g$ (note that g=0 or g=1/2).

In the case of extreme values, outliers require special attention. According to Hoaglin et al (1986), $[q - k(Q-q), Q + k(Q-q)]$ with $k = 1.5$ or $k = 2$ is a resistant rule for flagging outliers in exploratory contexts. In practice, deciding wether outliers should be removed or not depends on issues such as the problem domain, the properties of the data set, the aims of the analysis and so on.

BP variables can also describe properties of symbolic individuals where the theoretical distribution is known. In those cases, if the minimum (resp. maximum) value of the distribution is $-\infty$ (resp.$+\infty$) we recommend the use of the 0.01- or 0.05-quantile (resp. 0.99- or 0.95-quantile). Trenkler (2002) offers further information about turning theoretical distributions into boxplots.

3 Distances for BP variables

In this section, two approaches for measuring distances for BP data are considered: the first entails considering that a BP variable is a symbolic variable described by five single-valued variables and the second entails considering that a BP variable is a special case of interval-valued modal variable.

3.1 The classical approach

If a BP variable is considered as a variable compound of five sub-variables (m, q, Me, Q, M) the distance between a pair of BPs can be measured by means of a classical metric in $\mathbb{R}^5$. If a Minkowski metric with a given order $q = 1$ is considered, we obtain the Manhattan distance between a pair of BPs, BP_1 and BP_2:

$$\begin{aligned} d_{q=1}(BP_1, BP_2) = &|m_1 - m_2| + |q_1 - q_2| + |Me_1 - Me_2| + \\ &|Q_1 - Q_2| + |M_1 - M_2|. \end{aligned} \tag{2}$$

Other Minkowski metrics such as Euclidean or Chebychev or other kinds of distances can be applied. As the approach is straightforward, further details will not be given.

3.2 The Ichino-Yaguchi distance function

Ichino and Yaguchi (1994) define a generalized Minkowski metric for a multidimensional space of mixed variables (quantitative, qualitative and structural variables), which is based on the Cartesian join and meet operators.

Let $A = [\underline{A}, \overline{A}]$ and $B = [\underline{B}, \overline{B}]$ be a pair of intervals, the Cartesian meet and join operations for intervals are defined as $A \otimes B = A \cap B$ and $A \oplus B = [\min(\underline{A}, \underline{B}), \max(\overline{A}, \overline{B})]$, respectively. The Ichino-Yaguchi dissimilarity measure is defined by:

$$\phi(A, B) := |A \oplus B| - |A \otimes B| + \gamma(2|A \otimes B| - |A| - |B|), \tag{3}$$

where $|X| = \overline{X} - \underline{X}$ and $\gamma \in [0, 0.5]$ is a parameter which controls the effects of the inner-side nearness and the outer-side nearness between A and B. If $\gamma = 0.5$, the resulting dissimilarity function is:

$$\phi(A, B)_{\gamma=0.5} = |A \oplus B| - 0.5(|A| + |B|). \tag{4}$$

Ichino and Yaguchi consider that (4) is a suitable distance function. It can be seen that (4) is equivalent to:

$$\phi(A, B)_{\gamma=0.5} = \frac{|\underline{A} - \underline{B}| + |\overline{A} - \overline{B}|}{2}. \tag{5}$$

In order to measure distances between a pair of individuals, X and Y, described by p variables of different kinds, Ichino and Yaguchi (1994) proposed the generalized Minkowski metric of order q, with $q \geq 1$:

$$\delta(X, Y) = \left(\sum_{i=1}^{p} \phi(X_i, Y_i)^q \right)^{1/q}. \tag{6}$$

If BPs are considered as elements compound by four consecutive intervals, (6) can be applied to measure distances between them. Given (5) and (6) with order $q = 1$, the Ichino-Yaguchi distance function between a pair of BPs, BP_1 and BP_2, is defined as follows:

$$\begin{aligned} \delta_{q=1}(BP_1, BP_2) = & \frac{1}{2}(|m_1 - m_2| + 2|q_1 - q_2| + 2|Me_1 - Me_2| + \\ & 2|Q_1 - Q_2| + |M_1 - M_2|). \end{aligned} \tag{7}$$

This distance assigns more weight to the differences between quartiles than to the difference between the extremes. Therefore, it is not greatly conditioned by the behavior of the extremes, which is usually less stable than the behavior of the quartiles. Hence (7) seems suitable in contexts where the interest mainly lies in the behavior of the central area of the considered distributions.

As the BP distance in (7) is defined as the combination of the distances between the four intervals that characterizes each BP, it has to be divided into four in order to be integrated in the Ichino-Yaguchi metric for symbolic data given in (6). If not, BP variables would have quadruple weight than the other of the variables considered. The resulting BP distance is defined as follows:

$$\eta(BP_1, BP_2) = \frac{\delta(BP_1, BP_2)}{4}. \tag{8}$$

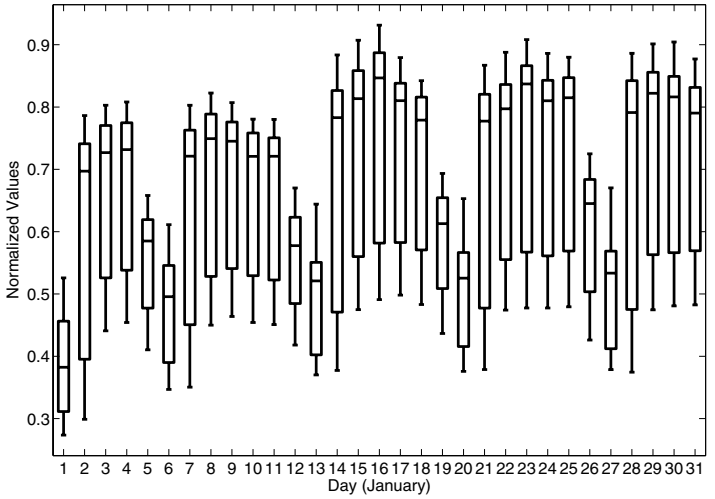


Fig. 1. Electric daily demand represented by a BP variable.

4 Hierarchical BP clustering: an example based on the daily electric demand

In order to illustrate our approach, an agglomerative clustering using the distances proposed is carried out in a real-life dataset which is available at: http://www.iit.upcomillas.es/∼cmate/. This dataset describes the hourly electric demand in a Spanish distribution area during 31 days, from 1-Jan-1991 until 31-Jan-1991, i.e. 31 individuals (days) described by 24 quantitative variables (hours) which records the value of the demand (the value of the demand is normalized). The dataset is summarized by means of a BP variable which represents the empirical distribution of the 24 values of the hourly demand along each day. The resulting dataset, shown in Fig. 1, is described by one BP variable.

An agglomerative clustering algorithm with the complete linkage method is applied to the distance matrix obtained with the Ichino-Yaguchi distance for BP data. The resulting dendrogram is shown in Figure 2. It can be seen that at 0.9 the dataset is clearly partitioned into two clusters: the first one with the holidays (Sundays and New Year's Day) and the second one with the working-days. In the second cluster, a subcluster containing all the January's Saturdays and other subcluster containing Mondays can be seen. This dendrogram is quite similar to the one shown in Figure 4, which is obtained by clustering the original dataset (described by 24 classical variables) using the Manhattan distance matrix. This strong resemblance reinforces the proposed approach and the Ichino-Yaguchi distance for BP variables.

In addition, Figure 3 shows the dendrogram yielded by the clustering algorithm of the distance matrix obtained by applying the Manhattan distance

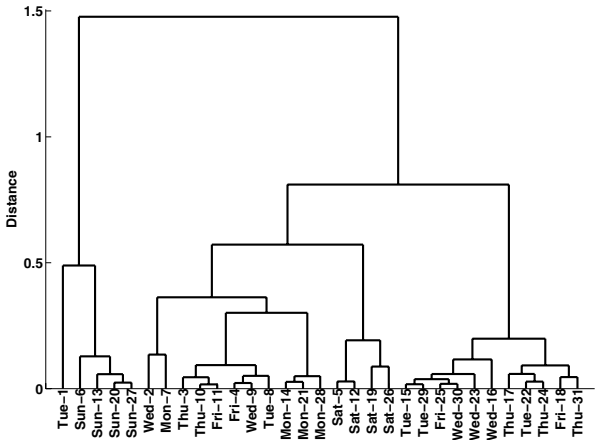


Fig. 2. Dendrogram obtained yielded by the Ichino-Yaguchi BP distance.

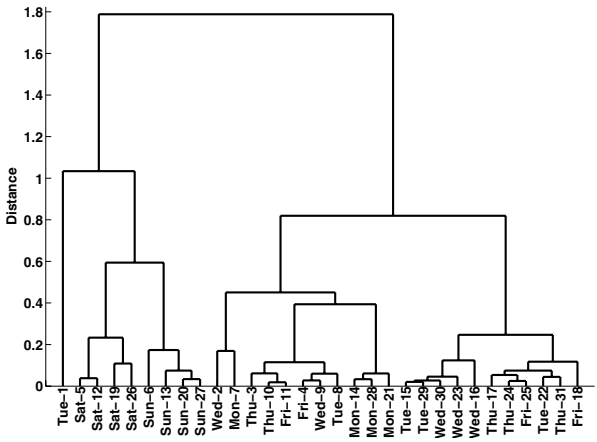


Fig. 3. Dendrogram yielded by the Manhattan BP distance.

for BP variables in the symbolic dataset. The main difference with the other dendrograms is that the subcluster of Saturdays belongs to the cluster of the Holidays. This makes sense as Saturdays are holidays for certain sectors.

Finally, the classical dataset has been summarized into a symbolic dataset described by a histogram variable representing the daily demand. The histogram variable has the same structure for all the 31 elements (days) of the set. Each histogram is composed by 10 fixed intervals partitioning the range of the variable, each interval with an attached weight representing its frequency. The distance matrix has been estimated with the quadratic-form distance proposed by Niblack et al. (1993) for image retrieval:

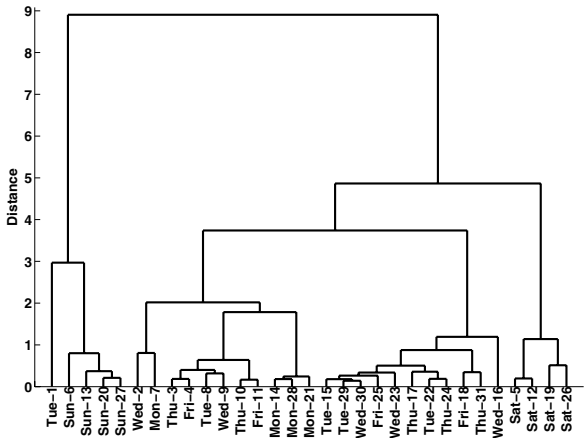


Fig. 4. Dendrogram yielded by the Manhattan distance in the classical dataset.

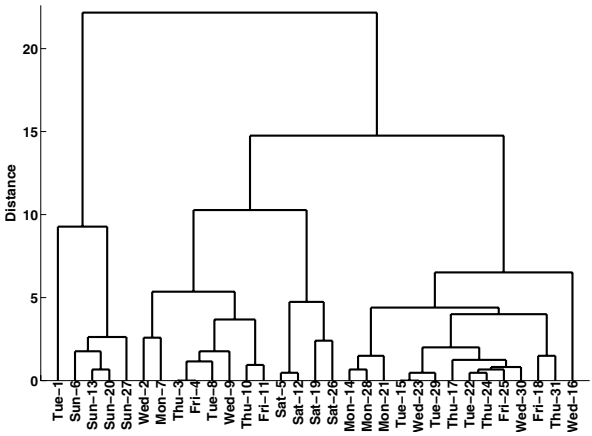


Fig. 5. Dendrogram obtained using the quadratic-form distance in the dataset described by a histogram variable.

$$d(H, K) = \sqrt{(h-k)'A(h-k)}, \tag{9}$$

where h and k are vectors representing the interval weights of the histograms H and K, respectively; matrix $A = [a_{ij}]$ denote the similarity between intervals i and j; $a_{ij} = 1 - d_{ij}/\max(d_{ij})$ where d_{ij} is the ground distance between the midpoints of intervals i and j. The dendrogram obtained with the quadratic-form distance and shown in Figure 5 generally resembles with those shown in Figures 2 and 4.

The example has shown that the summarization by means of BP variables and the subsequent clustering based on the distances proposed obtains analogous results to those obtained with histogram variables and with the clustering of the classical dataset. This fact endorses the approach proposed and shows that BP variables are a kind of variable worth considering in SDA contexts.

5 Conclusions

This paper presents BP variables as a new kind of symbolic variable with a great potential in SDA. BP variables enable the efficient summary of quantitative data providing information about the location, the spread,the skewness and the normality of the summarized empirical distribution. The proposed distances enhance the role of BPs as an exploratory tool, because, thus far, the comparison between BPs is made visually and cannot be quantified in an objective way.

Our approach to cluster BP variables is just the first step, but other clustering methods and distances for BP variables can be proposed. The adaptation of other statistical methods to BP variable also are interesting extensions to the present work.

References

BENJAMINI, Y. (1988): Opening the Box of a Boxplot. *American Statistician, 42/4, 257-262.*

BILLARD, L., and DIDAY, E. (2002): From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association, 98/462, 991-999.*

BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information ¿From Complex Data.* Springer-Verlag, Heidelberg.

FRIGGE, M., HOAGLIN, D. C., and IGLEWICZ, B. (1989): Some Implementations of the Boxplot. *American Statistician, 43/1, 50-54.*

HOAGLIN, D. C., IGLEWICZ, B., and TUKEY, J. W. (1986): Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association, 81/396, 991-999.*

ICHINO, M., and YAGUCHI, H. (1994): Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man and Cybernetics, 24/1, 698-708.*

NIBLACK, W., BARBER, R., EQUITZ, W., FLICKNER, M.D., GLASMAN, E.H., PETKOVIC, D., YANKER, P., FALOUTSOS, C., TAUBIN, G., and HEIGHTS, Y. (1993): Querying images by content, using color, texture, and shape. *SPIE Conference on Storage and Retrieval for Image and Video Databases, 1908, 173-187.*

TRENKLER, D. (2002): Quantile-Boxplots. *Communications in Statistics: Simulation and Computation, 31/1, 1-12.*

TUKEY, J. W. (1977): *Exploratory Data Analysis.* Addison-Wesley, Reading.