Interval-valued Data Clustering Based on Range Metrics

Sérgio Galdino
Polytechnic School
UPE
Recife, PE, Brazil
sergio.galdino@ieee.org

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
UFPE
Recife, PE, Brazil
wellington.santos@ufpe.br

Ricardo Paranhos Pinheiro
Polytechnic School
UPE
Recife, PE, Brazil
paranhos@gmail.com

***Abstract*—This paper introduces new approach to Data Clustering on interval-valued data. We introduce the Width of Range Distance (City Block and Euclidean) matrix dissimilarity to process hierarchical clustering using Ward method. Both width of Range Distances has good clustering results while preserving the topology of the data.**

***Index Terms*—Clustering, Interval-valued Data, Interval Arithmetic**

I. Introduction

Typically, data we analyse are classical data, which means each observation is a single point in a n-dimensional space. However, sometimes the data are represented intervals [1]. Traditional clustering techniques can be easily applied to interval data types by replacing each interval with a representative (e.g, the median of the points in the interval). However, we have limitations of using representative centroids to replace intervals [2]. We introduce the Width of Range distance (City Block and Euclidean) matrix dissimilarity to process hierarchical clustering using Ward method.

II. Range Distances

The set of real numbers x satisfying $\underline{x} \leq x \leq \overline{x}$ is the closed interval $[x] = [\underline{x},\ \overline{x}]$. The width of $[x]$ are defined as,

$$w([x]) = (\overline{x} - \underline{x}) \tag{1}$$

The magnitude and the mignitude can both be calculated using the end points of $[x]$,

$$mag([x]) = max\left\{|\underline{x}|, |\overline{x}|\right\}, \tag{2}$$

$$mig([x]) = \begin{cases} min\left(|\underline{x}|, |\overline{x}|\right) & \text{if } 0 \notin [x] \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

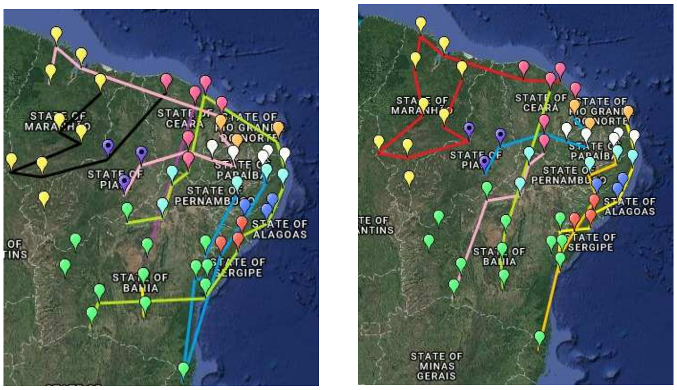
1) Range City Block Distance:

$$\begin{aligned} range[d_1([\mathbf{x}], [\mathbf{y}])] &= \sum_{i=1}^{n} |[x_i] - [y_i]| \\ &= \left[\sum_{i=1}^{n} mig([x_i] - [y_i]), \quad \sum_{i=1}^{n} mag([x_i] - [y_i])\right] \end{aligned} \tag{4}$$

2) Range Euclidean Distance:

$$d_2([\mathbf{p}], [\mathbf{q}])) = d_2([\mathbf{q}], [\mathbf{p}])) = \sqrt{\sum_{i=1}^{n}([q_i] - [p_i])^2} \tag{5}$$

DOI reference number: 10.18293/SEKE2018-113



(a) Width of Range City Block Dissimilarity.

(b) Width of Range Euclidean Dissimilarity.

Fig. 1: Interval-valued Data Hierarchical Clustering - *R hclust function (method = "Ward.D").*

III. Interval-valued Clustering

The data set concerns minimal and maximal of monthly temperatures observed in 50 meteorological stations mounted all over Brazil Northeast from 2017 year (January to November).http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep (2018/12/02). Figure 1 represents the Brazil Northeast map containing the 50 stations over 9 clusters. All stations of the same cluster are drawn with the same line colour.

IV. Conclusions

Wards method was used with the dissimilarities matrix using the width from Range City Block Distance Matrix and from Range Euclidean Distance Matrix. We can conclude that the stations located near each other geographically tend to be assigned to the same cluster or to a neighbour cluster.

References

[1] Billard, L. and Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester, (2007)

[2] S. M. L. GALDINO: *Interval-valued Data Clustering Based on the Range City Block Metric.* In: SMC 2016, 2016, Budapest. The 2016 IEEE International Conference on Systems, Man, and Cybernetics, 2016.