Multidimensional Temporal Clustering: geometrical similarity measures analysis in k-means

Ramona Stoica
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1, Cluj-Napoca, Romania
ramona@cs.ubbcluj.ro

Mihaela Ola
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1, Cluj-Napoca, Romania

Mihai Paraschivescu
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1, Cluj-Napoca, Romania

I. INTRODUCTION

Time series data is a sequence of real numbers that represent the measurements of a real variable at equal time intervals. A data stream is an ordered sequence of points $x_1, , , , , ,x_n$. These data can be read or accessed only once or a small number of times. A time series is a sequence of real numbers, each number indicating a value at a time point. Data flows continuously from a data stream at high speed, producing more examples over time in recent real world applications.

Most of the time series encountered in cluster analysis are discrete time series. When a variable is defined at all points in time the time series is continuous. Clustering of time series data has applications in an extensive assortment of fields and has attracted a large amount of research ([2][4][5][6][7][8][12]).

Multidimensional time series are an extension and generalization of regular time series. They have more impact nowadays as most of the data consists of more parameters which are measured over time and decision has to be made considering the behavior of all these parameters together. We propose to investigate in this paper the behavior of the k-means algorithm for several multidimensional time series data. We compare versions of k-means for several distance measures. The paper is organized as follows: Section II describes the multidimensional time series data, Section III presents the k-means for multidimensional time series data clustering and the distance measures used, Section IV contains experiments and comparisons and Section V presents the conclusions.

II. MULTIDIMNESIONAL TIME SERIES

A time series is defined as an array $X = (x_1, x_2, \ldots, x_n)$ of measurements in time for a given parameter (or variable).

A *multidimensional time series* [9] is defined as:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

where each X_i, $1 \le i \le N$ is a time series on its on. The size of these time series can vary.

In the multidimensional case, clustring involves grouping entities of the form X. Figure 1 shows an example of hierarchical clustering for 2-dimensional time series (there are 5 entries or instances that are clustered).

Multi-dimensional time series appear if one deals with multiple measurements on some objects, phenomena, or variables.

Many times, the multidimensional time series data are converted into a single time series by concatenating all the time series into a sinle one. But this will conduct to loss of generality. The advantage of dealing with a multi-dimensional time series as such without transforming them is that, on the one hand, it offers a global point of view and shows some critical pathologies arising from evident discrepancies, whereas, on the other hand, it permits to integrate the information contained in each one-dimensional time series of X and therefore it is useful when each array is sparse and short [1].

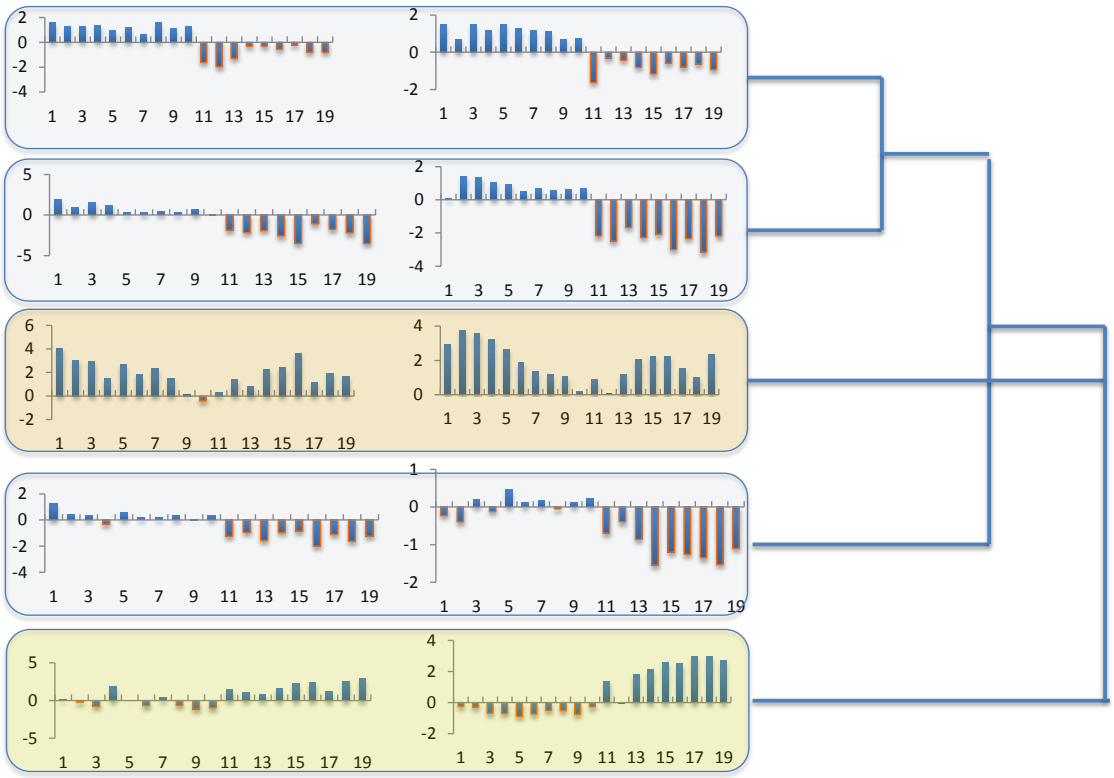

Figure 1 Two dimensional time series: example of hierarchical clustering

III. K-MEANS FOR MULTIDIMENSIONAL CLUSTERING

The similarity between two time series is usually calculated using a distance or a similarity measure. In this chapter we consider the difference between each time series (of a multidimensional time series instance) as an objective function which has to be minimized. Thus, the goal is to compare how similar two object X and Y are, where X and Y are given by:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}, \qquad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$$

A. *Similarity measures*

We define an N dimensional objective function F=($f_1, f_2, \ldots, f_N$) as:

$$F = \begin{pmatrix} f_1 = d(X_1, Y_1) \\ f_2 = d(X_2, Y_2) \\ \vdots \\ f_N = d(X_N, Y_N) \end{pmatrix}$$

where d(·) defines a similarity measure.

We use k-means [10] for clustering multidimensional time series data. In our case, each item is assigned to a cluster based on the values of the F function. We consider a weighted combination of all f_i, $1 \le i \le N$ as a result of the similarity and denote this by d_{sim}:

$$d_{sim} = \sum_{i=1}^{N} w_i f_i$$

where w is a vector of weigths denoting the importance of that particulat time series in the clustering. For our experiments we considered all time series as having equal importance and in this case $w_i = 1$, $1 \le i \le N$.

We implemented four different distances d(·):

- Euclidian distance
- Manhattan distance
- Maximum distance
- Average distance

B. *Learning the k value*

One of the four distance measures (Euclidian distance, Manhattan distance, Maximum distance, Average distance) is selected from the main menu, and sent as parameter for the algorithm to use while computing. Also a Maximum Distance Percent can be introduced before running the algorithm; the default value for this variable is 0.6 in our experiments.

The algorithm starts with a large k (equal to the no. of items to cluster) which is decreased step-by-step (by moving data, if convenient, from initial clusters - containing only one item from the data set - to new clusters - containing similar items) until it reaches a value that satisfies the stability of each cluster (small distance between data belonging to same cluster, large distance between data belonging to distinct clusters).

IV. EXPERIMENTS

We perform experiments considering three datasets from various domains. Silhouette coefficient [11] is used to compare the performance of k-means for various distance measures.

A. First dataset

This dataset contains data about countries with respect to temperature, precipitation level, atmospheric pressure and humidity. The countries have to be clustered based on the records over time for all these parameters together.
These are the details of the dataset:

- 14 (Countries);
- No. of parameters: 5 (Precipitations Level (L/m^2), Wind Speed (m/s), Temperature (grC), Atm. Pressure (mmHg), Humidity (%RH));
- No. of time points: 77.

The results obtained by k-means are presented in Table 1.

Table 1. k-means results for the first dataset.

Algorithm	Number of clusters	Silhouette coefficient
k-means with Euclidian distance	8	0.2105
k-means with Manhattan distance	10	0.1574
k-means with Maximum distance	12	0.0200
k-means with Average distance	8	0.2105

From the experiments we observe that:

- Best average silhouette coefficient: Euclidian distance and Average distance;
- Better average silhouette coefficient for cluster 0 is obtained using Manhattan distance (0.745) not Euclidian/Average distance (0.181) or Maximum distance (0.240);
- Better average silhouette coefficient for cluster 1 is obtained using Euclidian distance or Average distance (0.674);
- Best average silhouette coefficient obtained for a cluster is 0.828 using Euclidian, Average or Manhattan distance.

B. Second dataset

This dataset if from the Machine Learning Repository [13]. The files contains 19 activities (like sitting, lying on back and on right side, ascending and descending stairs, running on a treadmill with a speed of 8 km/h, etc). Data is acquired from one of the sensors (T_xacc) of one of the units (T) over a period of 5 sec, for each subject and for each of the activities.
Results obtained by k-means are presented in Table 2. In this case we tested the algorithm with two values for the maximum Distance Percent parameter (used to decide which k (number of clusters) is best): 0.6 and 0.9.

Table 2. k-means results for the second dataset.

Algorithm	Number of clusters	Silhouette coefficient
Max Distance Percent = 0.6		
k-means with Euclidian distance	18	0.028
k-means with Manhattan distance	18	0.028
k-means with Maximum distance	17	0.028
k-means with Average distance	18	0.028
Max Distance Percent = 0.9		
k-means with Euclidian distance	17	0.028
k-means with Manhattan distance	16	0.038
k-means with Maximum distance	17	0.028
k-means with Average distance	18	0.028

We observed that:

- Best average silhouette coefficient: Manhattan distance using Max Distance Percent 0.9;
- The same average silhouette coefficient for cluster 1 is obtained using Manhattan distance, Euclidian distance or Average distance and the default Max Distance Percent (0.6) or Average distance and a Max Distance Percent = 0.9 (0.521)
- The same average silhouette coefficient for cluster 0 is obtained using Maximum distance and the default Max Distance Percent or Euclidian distance or Average distance and a Max Distance Percent = 0.9 (0.491);
- Best average silhouette coefficient obtained for a cluster is 0.613 using Manhattan distance and Max Distance Percent 0.9;
- For Max Distance Percent lower than default (0.6) worse clustering results have been obtained.

C. *Third dataset*

The third dataset if from the KEGG database [14] and is not a time series dataset. We wanted to test the algorithm for this kind of data as well in order to validate the findings. The data is a Metabolic Relation Network (Directed) Data Set. It has 8 attributes such as: Nodes (min:2, max:116), Edges (min:1, max:606), Connected Components (min:1, max:13), Network Diameter (min:1, max:30), Network Radius (min:1, max:2), Shortest Path (min:1, max:3277), Characteristic Path Length (min:1), Average number of Neighbors (min:1))
The data set has 1,000 instances.
The results obtained by k-means are given in Table 3.

Table 3. k-means results for third dataset.

Algorithm	**Number of clusters**	**Silhouette coefficient**
k-means with Euclidian distance	618	0.0014
k-means with Manhattan distance	618	0.0014
k-means with Maximum distance	618	0.0014
k-means with Average distance	618	0.0014

We can observe that:

- The same average silhouette coefficient is obtained for all distance measures (0.0014);
- Using different values for Max Distance Percent (0.2, 0.6, 0.9) hasn't improved the results;
- The best average silhouette coefficient obtained for a cluster is 0.920.

V. CONCLUSIONS

The paper investigates the role of various distance measures in k-means algorithm for clustering multidimensional time series data. Euclidian distance is the most frequent used and most common measure. Our experiments – three different datasets – reveal that Manhattan distances (and sometimes the average distance) are better candidates for similarity between two multidimensional time series instances. This work only investigates geometrical distances, but as future work, geometric distances presented here will be compared with other similarity measures (such as descriptive measures, pattern finding measures, etc.).

REFERENCES

[1]. Marco Franciosi, Giulia Menconi: Multi-dimensional sparse time series: feature extraction. CoRR abs/0803.0405 (2008)

[2]. Tsay, R. (2002). *Analysis of Financial Time Series*.Wiley Series in Probability and Statistics. New York: JohnWiley & Sons.

[3]. Shadbolt, J. and Taylor, J., editors (2002). *Neural Networks and the Financial Markets*. London: Springer.

[4]. Azoff, E. (1994). *Neural Network Time Series Forecasting of Financial Markets*. New York: JohnWiley & Sons.

[5]. Gunopulos, D., and Das, G. (2000). Time series similarity measures (tutorial PM-2). In *Tutorial notes of the sixth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 243–307. Boston, MA: ACM Press.

[6]. Bollobás, B., Das, G., Gunopulos, D., and Mannila, H. (1997). Time-series similarity problems and well-separated geometric sets. In *SCG '97: Proceedings of the thirteenth annual symposium on computational geometry*, pages 454–456. New York: ACM Press.

[7]. Das, G., Gunopulos, D., and Mannila, H. (1997). Finding similar time series. In *Proceedings of the first European symposium on principles of data mining and knowledge discovery*, pages 88–100. New York: Springer-Verlag.

[8]. T. Warren Liao, Clustering of time series data—a survey, Pattern Recognition, Volume 38, Issue 11, Pages 1857-1874,2005

[9]. Marco Franciosi, Giulia Menconi: Multi-dimensional sparse time series: feature extraction. CoRR abs/0803.0405 (2008)

[10]. Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, pp. 281–297. Berkeley, CA: University of California Press.

[11]. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics* **20**: 53–65

[12]. Kendall, S. and Ord, J. (1990). *Time Series*, 3rd edition. Seven Oaks, U.K.: Edward Arnold.

[13]. http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities+Dataset

[14]. http://www.genome.jp/kegg/pathway.html