JOURNAL LA MULTIAPP

VOL. 01, ISSUE 03 (001-007), 2020
DOI: 10.37899/journallamultiapp.v1i3.191

Time Series Clustering Based on the K-Means Algorithm

Oleg Kobylin[1], Vyacheslav Lyashenko[1]

[1]*Department of Informatics, Kharkiv National University of RadioElectronics, Ukraine*

Corresponding Author: Vyacheslav Lyashenko
Email: lyashenko.vyacheslav@gmail.com

Article Info

Article history:
Received 27 August 2020
Received in revised form 06 September 2020
Accepted 19 September 2020

Keywords:
Clustering
Time Series
K-Means

Abstract

Time series is one of the forms of data presentation that is used in many studies. It is convenient, easy and informative. Clustering is one of the tasks of data processing. Thus, the most relevant currently are methods for clustering time series. Clustering time series data aims to create clusters with high similarity within a cluster and low similarity between clusters. This work is devoted to clustering time series. Various methods of time series clustering are considered. Examples are given for real data.

Introduction

Primary data is the base that allows you to understand and predict the processes that are studied and analyzed. Therefore, data processing and analysis is the basis for any research (Matarneh, Maksymova, Lyashenko & Belova, 2017; Lyashenko et al., 2016). The amount of such data can be very large. This makes it necessary to use various methods for the analysis and interpretation of primary data (Khan, Joshi, Ahmad & Lyashenko, 2015; Baranova, Sergienko, Stepurina & Lyashenko, 2020; Kang, 2019). Among these methods, data clustering should be distinguished. This approach allows you to split the general data set into separate groups, where each group has some common characteristics.

Thus, clustering is a way of preprocessing data for more convenient subsequent analysis. Having received the necessary groups, as well as their centroids, you can continue to work with specific representatives, and not with the entire data set. This reduces the processing time and the time to obtain results. This approach also allows for a better understanding of the data; to carry out their compression in conditions of unprofitable data. It should also be noted that raw data can be presented in different ways. Time series is one of the forms of data presentation. A time series is a time-oriented sequence of data on a certain subject area that is of interest. It is a way of presenting statistics. Time series data is used in various spheres of human activity (Baranova, Sergienko, Stepurina & Lyashenko, 2020; Baranova et al., 2020). Therefore, this form of data presentation is of particular interest. Some issues of processing such data are considered in our work.

Some Features of Time Series Processing

When processing a time series, you can encounter typical difficulties: large dimension of input data, presence of noise and missing data. Considering clustering of time series, one should also pay attention to the fact that rows can contain a different number of samples; there are more

ISSN: 2716-3865 (Print), 2721-1290 (Online)

degrees of freedom to determine the similarity of one object to another; when choosing metrics and statistics, you should pay attention to the local dependence of the data.

An important task when processing a time series is also to determine the proximity of data. It can be closeness in time, closeness in form, closeness in structure (Maharaj, D'Urso & Caiado, 2019; Ali, Alqahtani, Jones & Xie, 2019). Time series can also contain anomalous values, which requires pre-processing and series smoothing. If this is not done, abnormal data may distort the results to be obtained.

Irwin's criterion is used to analyze anomalous data; methods such as moving average, exponential smoothing are used to smooth data (Zou et al., 2019; Walker, Curtis & Goldacre, 2019). All this must be taken into account when clustering data that is presented in the form of time series.

K-Means Based Time Series Clustering Methods

Let's consider the most common time series clustering methods that use the k-means algorithm. These methods include: Euclidean k-means, DBA k-means and Soft-DTW k-means. One of the common method for clustering time series is the k-means approach, where Euclidean distance is used as a measure of proximity (Steinley, 2006; Khachumov, 2012):

$$d_{E1}(X_i, X_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

where X_i and X_j are two time series of length m.

The k-means algorithm is that k arbitrary centers are selected first. Then the rest of the elements are grouped around these centers, which must be divided into classes. At the next step, new centers are calculated for the resulting clusters so that the square of the Euclidean distance from the cluster element to its centroid is less than the distance to the centroids of the remaining clusters.

At the same time, the algorithm places the centers of the clusters (centroids) so that the average values for the lists of elements within the constructed clusters differ as much as possible. Thus, the Euclidean k-means method divides time series of sample length m into k groups (clusters). This separation occurs by minimizing the total squared deviation of cluster points from the centroids of these clusters:

$$\min\left[\sum_{i=1}^{k}\sum_{x^{(j)}mS_i}\left\|x^{(j)} - \mu_i\right\|^2\right]$$

where $x^{(j)} \in R^n, \mu_i \in R^n$; μ_i – cluster centroid S_i.

Using the Euclidean k-means method has several disadvantages: it is necessary to determine in advance the number of resulting clusters, which may not always be advisable; the method is sensitive to the choice of the initial cluster centers – this leads to an increase in the probability of error and the possibility of obtaining results that differ from each other when the algorithm is restarted.

There are also cases when an object can belong to different clusters. Despite the shortcomings, Euclidean k-means is a simple algorithm, well suited for understanding the general clustering processes and a good basis for building extended new algorithms on its basis. When clustering time series, it is essential to take into account the fact that some series can be almost the same, but at the same time these series can be shifted in time (along the time axis). Therefore, it is advisable to use a metric that is implemented in the dynamic timeline transformation (DTW) algorithm.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Consider two time series X_i with length n_i and X_j with length n_j :

$$X_i = \{x_i\}_{i=1}^{n_i},$$

$$X_j = \{x_j\}_{j=1}^{n_j}.$$

Then the implementation of the DTW method can be described in the following steps (Kate, 2016; Hu, Mashtalir, Tyshchenko & Stolbovyi, 2018).

At the first step, we construct the distance matrix $D = \{d_{i,j}\}$.

At the second step, we construct a transformation matrix $D_{DTW} = \{r_{i,j}\}$, where each element is determined using the formula:

$$r_{i,j} = d_{i,j} + \min(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1}).$$

After filling in the transformation matrix, we move on to the final stage. This stage consists in building the optimal transformation path and DTW distance. The transformation path W is a set of contiguous elements of the D_{DTW} matrix that matches the series X_i and X_j and minimizes the total distance between these time series. Thus, the last step is to build the optimal transformation path and DTW distance.

The transformation path between X_i and X_j is determined by the formula (Kate, 2016):

$$W = \{w_k\}_{k=1}^{K}, \quad w_k = (i,j)_k, \max(n_i, n_j) \le K \le (n_i + n_j)$$

where K – path length.

Then DTW the distance between two time series is determined by the formula (Kate, 2016; Hu, Mashtalir, Tyshchenko & Stolbovyi, 2018):

$$DTW(X_i, X_j) = \min\left(\frac{\sum_{k=1}^{K} d(w_k)}{K}\right)$$

A modification of the DTW method is the soft-DTW k-means algorithm, in which the DTW distance is determined as (Montgomery, Jennings & Kulahci, 2015):

$$DTW_\gamma(X_i, X_j) = -\gamma \log \sum_{k=1}^{K} e^{\left(\frac{d(w_k)}{K \cdot \gamma}\right)}$$

for different values of the smoothing parameter (γ) of the time series.

Also in the Euclidean k-means method, we can estimate the distance between the «centers of weight» of each group of time series (Okawa, 2019):

$$d_{ct}(x_i, x_j) = d(\bar{x_i}, \bar{x_j})$$

Then the corresponding method for determining the distance between time series (clustering them) is called the DBA-k-means method (DTW Barycenter Averaging). Let's conduct a comparative analysis of clustering time series using the methods that we discussed above.

Results and Discussion

For the analysis, we will look at the time series that represent medical data. In particular, these are the data of the electrocardiogram of the heartbeat (ECG). Thus, the time series correspond to the forms of the electrocardiogram of heart contractions for the normal case and cases of lesions by various arrhythmias and myocardial infarction. These signals were preprocessed and segmented, with each segment corresponding to one heartbeat.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

An example of such time series is shown in Figure 1. These time series are included in the database that is used for fundamental research and is described in (Moody & Mark, 2001).

The main characteristics that are used for clustering time series (Figure 1) are: the number of series – 87554; the number of values in each row is at least 187; sampling rate – 125 Hz; the number of classes that we are considering is 4. To implement the methods discussed above, to carry out the clustering procedure, the Python environment was chosen.

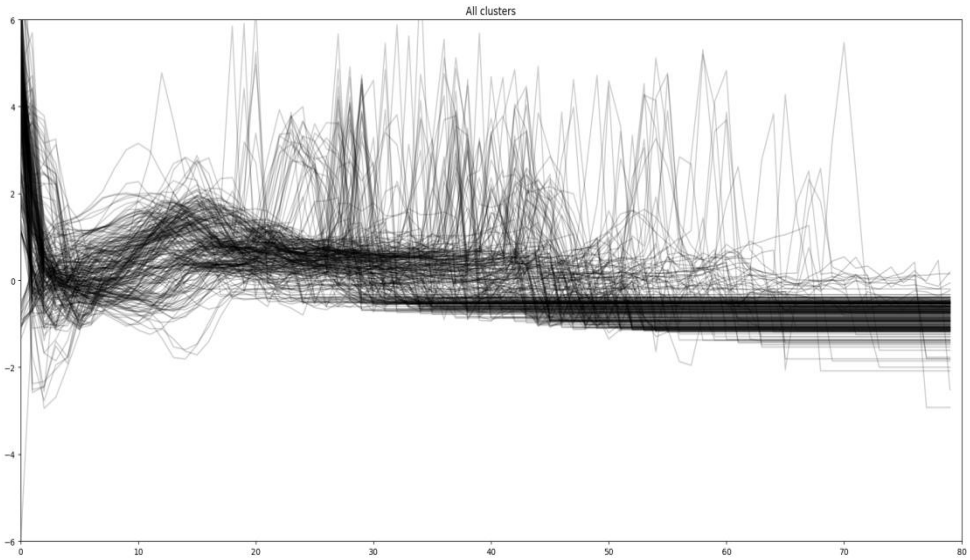

Figure 1. Time series used to cluster them

Figure 2 - Figure 4 shows the results of clustering by different methods. The red line is the centroid of each cluster.

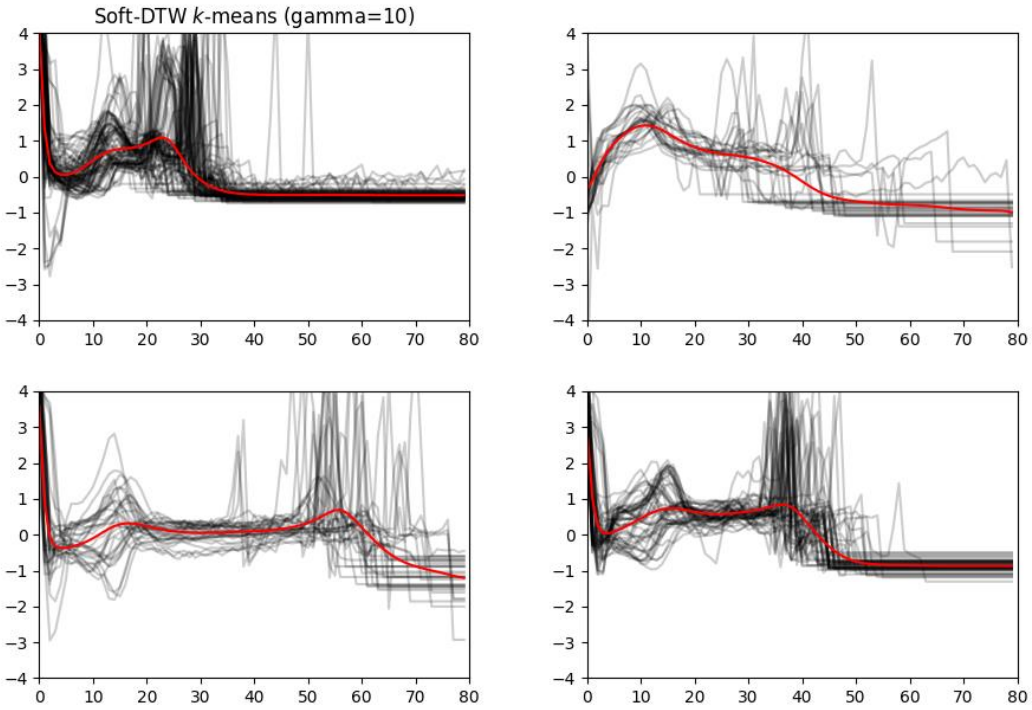

Figure 2. Results of clustering by the Soft-DTW k-means method

ISSN: 2716-3865 (Print), 2721-1290 (Online)

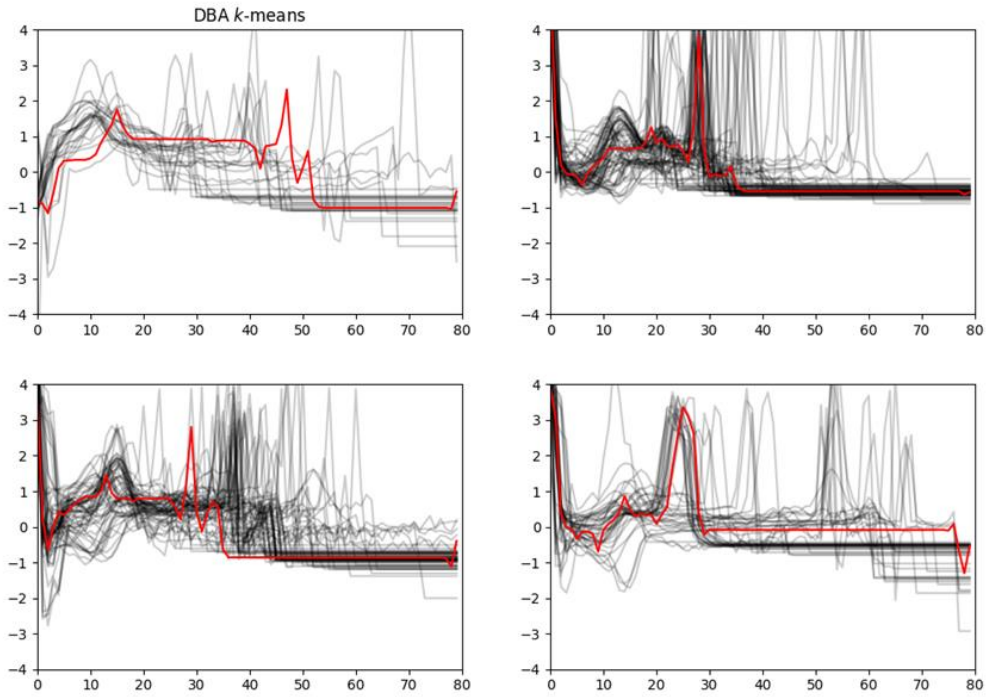

Figure 3. Results of clustering by the DBA k-means method

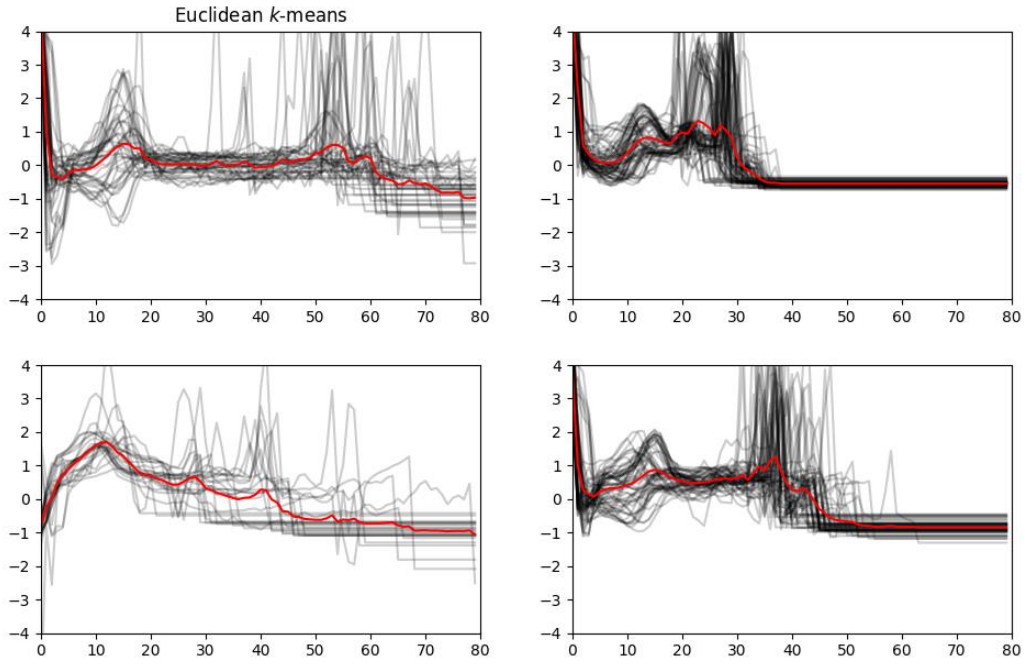

Figure 4. Results of clustering by the Euclidean k-means method

From these figures (Figure 2 - Figure 4), it can be seen that the above methods of clustering time series for the presented sample (Figure 1) give approximately the same result. This is due to the fact that the time sequence was previously divided into segments in accordance with one heartbeat. This made it possible to level out minor deviations of readings along the time axis. Nevertheless, the presented results make it possible to build systems for automatic data analysis, which are presented in the form of time series.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Conclusion

The paper provides an overview of the main clustering methods that are used to analyze time series. These methods include: Euclidean k-means, DBA k-means and Soft-DTW k-means. The advantages and disadvantages of each method are noted. Some features of data analysis are also considered, which are presented in the form of a time series. For experimental studies, a dataset was selected, formed from the ECG database of heart beats. The task of clustering and classifying this data assists in processing and identifying anomalies in humans to diagnose cardiovascular problems. The results of clustering are presented.

References

Ali, M., Alqahtani, A., Jones, M. W., & Xie, X. (2019). Clustering and Classification for Time Series Data in Visual Analytics: A Survey. *IEEE Access*, 7, 181314-181338.

Baranova, V., Orlenko, O., Vitiuk, A., Yakimenko-Tereschenko, N., & Lyashenko, V. (2020). Information System for Decision Support in the Field of Tourism Based on the Use of Spatio-Temporal Data Analysis. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 6356-6361.

Baranova, V., Sergienko, O., Stepurina, S., & Lyashenko, V. (2020). Price Environment for Gold and Silver in the Context of the Development of COVID-19. *Journal of Asian Multicultural Research for Economy and Management Study*, 1(2), 25-32.

Hu, Z., Mashtalir, S. V., Tyshchenko, O. K., & Stolbovyi, M. I. (2018). Clustering matrix sequences based on the iterative dynamic time deformation procedure. *International Journal of Intelligent Systems and Applications*, 10(7), 66-73.

Kang, Z., Xu, H., Wang, B., Zhu, H., & Xu, Z. (2019). Clustering with similarity preserving. *Neurocomputing*, 365, 211-218.

Kate, R. J. (2016). Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2), 283-312.

Khachumov, M. V. (2012). Distances, metrics and cluster analysis. *Scientific and Technical Information Processing*, *39*(6), 310-316.

Khan, A., Joshi, S., Ahmad, M. A., & Lyashenko, V. (2015). Some Effect of Chemical Treatment by Ferric Nitrate Salts on the Structure and Morphology of Coir Fibre Composites. *Advances in Materials Physics and Chemistry*, 5, 39-45.

Lyashenko, V., Lyubchenko, V., Mohammad, A., Alveera, K., & Kobylin, O. (2016). The Methodology of Image Processing in the Study of the Properties of Fiber as a Reinforcing Agent in Polymer Compositions. *International Journal of Advanced Research in Computer Science*, 7(1), 15-18.

Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). *Time series clustering and classification*. CRC Press.

Matarneh, R., Maksymova, S., Lyashenko, V., & Belova, N. (2017). Speech Recognition Systems: A Comparative Review. *Journal of Computer Engineering (IOSR-JCE)*, 19(5), 71-79.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45-50.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Okawa, M. (2019). Template matching using time-series averaging and DTW with dependent warping for online signature verification. *IEEE Access*, 7, 81010-81019.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1), 1-34.

Walker, A. J., Curtis, H. J., & Goldacre, B. (2019). Impact of Chief Medical Officer activity on prescribing of antibiotics in England: an interrupted time series analysis. *Journal of Antimicrobial Chemotherapy*, 74(4), 1133-1136.

Zou, Y., Donner, R. V., Marwan, N., Donges, J. F., & Kurths, J. (2019). Complex network approaches to nonlinear time series analysis. *Physics Reports*, 787, 1-97.

ISSN: 2716-3865 (Print), 2721-1290 (Online)