Comparative Analysis of Noisy Time Series Clustering

Lyudmyla Kirichenko [1] [0000-0002-2780-7993], Tamara Radivilova[1] [0000-0001-5975-0269], Anastasiia Tkachenko[1] [0000-0002-1683-4662]

[1] Kharkiv National University of Radio Electronics, Kharkiv, 61166, Ukraine
lyudmyla.kirichenko@nure.ua

Abstract. A comparative analysis of the clustering of sample time series was performed. The clustering sample contained time series of various types, among which atypical objects were present. In the numerical experiment, white noise with different variance was added to the time series. Clustering was performed by k-means and DBSCAN methods using various similarity functions of time series. The values of the quality functionals were quantitative measures of the quality of clustering. The best results were shown by the DBSCAN method using the Euclidean metric with a Complexity Invariant Distance. The method allows to separate a cluster with atypical series at different levels of additive noise. The results of the clustering of real time series confirmed the applicability of the DBSCAN method for detecting anomaly.

Keywords: Time Series Clustering, DBSCAN Method, Atypical Time Series, Noisy Time Series Clustering

1 Introduction

Nowadays, a lot of datasets of time series are created and constantly replenished, for example, purchase and sale series, stock prices, exchange rates, weather data, biomedical measurements and others. Processing of such datasets require new approaches, in particular, machine learning approaches [1-4]. One of the impotent tasks of machine learning is objects clustering, information about which is presented in the form of time series. The time series clustering is used as an independent research technique, and as a part of more complex data mining methods, such as rule detection, classification, anomaly detection, and so on [5-7].

In the task of cluster analysis of time series, it is required to split the set of objects (time series) into a relatively small number of clusters so that the group quality criterion takes on the best value. The quality criterion is usually understood as a certain functional depending on the scatter of objects within clusters and the distances between them. A ways to specify distances or similarity measure between objects are also various [5].

There are several approaches to time series clustering. In [1,6] three basic approaches of time series clustering are proposed: raw-data-based, feature-based, model-based. In case of raw-data-based clustering the objects are raw data in the frequency or time domain.

Time series clustering algorithms are mainly classified into the same categories as for other data (k-means algorithms, hierarchical methods, density based methods, mesh based methods, etc.) [1,5,6,8]. The choice of the measure of difference plays an important role in clustering, since an unsuccessfully selected measure would entail incorrect results. When choosing a specific similarity measure, the researcher often relies on his knowledge and experience in solving similar problems. Many works are devoted specifically to the application and development of measures for the various of time series [9-11].

One of the tasks of time series clustering is the separation anomalous objects into a separate cluster [12,13]. This is not simple, especially in a noisy time series. In the present paper, the problem of clustering noisy time series containing anomalous objects is considered.

2 Statement of the problem and methods of solution

2.1 The problem formulation

Consider the object space X, whose elements are time series. Let $X^l = \{x_i\}_{i=1}^l$ is training set, $\rho: X \times X \to [0, +\infty)$ is similarity function between two time series.

It is necessary to determine: Y is set of clusters and $a: X \to Y$ is an algorithm that allows to construct non-intersecting subsets (clusters), such that within each cluster objects are close according to similarity function ρ and the objects of different clusters differ significantly among themselves.

The aim of the work is to conduct a comparative analysis of the noisy time series clustering with anomalous objects using several clustering methods and various similarity functions.

2.2 Clustering methods

The following clustering methods were chosen for the numerical experiment: the k-means method which is actually most often used for clustering time series and Density-based spatial clustering of applications with noise (DBSCAN) method because it works well with noisy data.

K-means method is a one of iterative clustering algorithms. At first, we need to initialize number of clusters and centroids for each cluster. Centroids are main objects on which we distribute initial data in clusters. They can be obtained randomly or we can choose some objects from initial data. Then we distribute all the objects in clusters according to proximity to the centroids. We should distribute objects again after centroids recalculation and repeat process until centroids stop changing.

In this case objects are time series with length n, i. e. they consist of n components. Let it be k clusters and centroids for this clusters look like:

$$\mu_l^{(i)} = (\mu_{l1}^{(i)}, \mu_{l2}^{(i)}, \dots, \mu_{ln}^{(i)}), \quad (1)$$

where $\mu_l^{(i)}$ is time series which is the centroid for the cluster number l on the iteration number i. To recalculate centroids, we need to find the average by each component:

$$\mu_{lj}^{(i+1)} = \frac{\sum_{j=1}^{m_l} x_j^{(l)}}{m_l}, j = \overline{1,n}, \quad (2)$$

where $\mu_{lj}^{(i+1)}$ is the component number j for the cluster number l; $x_j^{(k)}$ is the component number j of object which belongs to the cluster number l; m_l is objects amount of the cluster number l.

Continue the process until the next condition is true:

$$\rho(\mu_l^{(i)}, \mu_l^{(i+1)}) \leq \varepsilon. \quad (3)$$

In terms of computational complexity this algorithm is simple. There is one disadvantage that amount of clusters does not change and the result depends on initial centroids. It means that it is possible to get dissimilar objects in one cluster.

The DBSCAN method is typically not used with time series. One of the main features is possibility to define atypical objects from initial data.

The key idea is to distribute similar objects in clusters regarding the density. At first, we define a closeness radius and objects number, which should be located within this closeness radius. Pretty similar or densely located are objects, which are located at the distance less or equal to defined closeness radius.

Let it be data set which contains n objects; $\rho(x, y)$ is defined similarity function; r is the closeness radius; m is the minimum number of objects, which should be within the radius. There is a function M which defines objects number that located within the closeness radius:

$$M(x_i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} c(x_i, x_j), \quad (4)$$

where $c(x_i, x_j)$ is membership function:

$$c(x_i, x_j) = \begin{cases} 1, \rho(x_i, x_j) \leq r, \\ 0, \rho(x_i, x_j) > r. \end{cases} \quad (5)$$

Kernel objects are objects for which the conditional $M(x_i) \geq m$ is true.

Boundary objects are objects for which the conditional $M(x_i) < m$ is true and there is such kernel object y_i, that the condition $\rho(x_i, y_i) \leq r$ is true.

Noise objects are objects for which the condition $M(x_i) < m$ is true but such kernel object y_i doesn't exist for which the condition $\rho(x_i, y_i) \leq r$ is true. In terms of the DBSCAN method, the noise is a set of atypical objects.

Kernel objects determine the main clusters. Then we distribute boundary objects by these clusters. Noise objects are distributed by clusters in the next way:

- if there is no any other noise object within the radius r, we distribute this object in a separate cluster;

- if there is at least one noise object within the radius r, we combine them in a common cluster;
- if there is at least one kernel object or boundary object within the radius r, we add noise object in a cluster, which contains this kernel object or boundary object.

One of the disadvantages of this distribution way is possibility to define atypical objects like boundary objects.

2.3 Similarity functions

The initial set consists of time series. Using special similarity functions is required for time series clustering.

One of the most popular similarity functions is the Euclidian distance. In the case with time series the distance is calculated by the formula:

$$E(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}, \tag{6}$$

where X, Y are time series of length n.

There is a situation when two time series are similar in common, but they are very different in some points. If it is necessary to classify these time series like similar, it is possible to use the main similarity function with a Complexity Invariant Distance:

$$\begin{gathered} CID(X,Y) = D(X,Y) \times CF(X,Y), \\ CF(X,Y) = \frac{max\{CE(X),CE(Y)\}}{min\{CE(X),CE(Y)\}}, \\ CE(X) = \sum_{t=2}^{n}\sqrt{(X_i - X_{i-1})^2}, \end{gathered} \tag{7}$$

where X, Y are time series of length n, $D(X,Y)$ is the main similarity function.

If it is necessary to compare time series of different lengths, it is possible to use similarity function Minimum Jump Cost (MJC). The key idea of MJC is to find sum of the minimum "jumps" between time series.

Let it be two time series x and y of lengths N and M respectively, i. e. they consist of N and M components. We start from the component $x(i),\ i = 0$. Then we find such component $y(j),\ j > i$ that the condition $(x(i) - y(j))^2 \to min$ is true. On the next iteration $i = j + 1$:

$$MJC(X,Y) = \sum_i c_{min}^i, \tag{8}$$

$$c_{min}^i = min\left(c_{t_x}^{t_y}, c_{t_x}^{t_{y+1}}, \dots, c_{t_x}^{t_{y+N}}\right), \tag{9}$$

where $c_{t_x}^{t_y}$ is all kinds of "jumps" which is calculated by formula

$$c_{t_x}^{t_{y+\Delta}} = \left(x_{t_x} - y_{t_{y+\Delta}}\right)^2. \tag{10}$$

MJC function is asymmetrical, that's why it is necessary to use the value

$$min\{MJC(X,Y), MJC(Y,X)\}. \quad (11)$$

In some cases, it is possible to improve the result accuracy by using similarity function Dynamic Time Warping (DTW). To calculate this distance, it is necessary to do the next steps.

- Create the matrix d of local distances between each time series components:

$$\{d_{ij}\} = |X_i - Y_j|,\ i = \overline{1,n}, j = \overline{1,m}, \quad (12)$$

where *X*, *Y* are time series of length *n* and *m* respectively.

- Create the transformation matrix D:

$$D_{11} = d_{11},\ \{D_{ij}\} = d_{ij} + min\{D_{i-1\ j-1}, D_{i-1\ j}, D_{i\ j-1}\}, i = \overline{1,n}, j = \overline{1,m}. \quad (13)$$

- Create the path from D_{MN} to D_{11}:

$$D_{next} = min\{D_{i-1\ j}, D_{i-1\ j-1}, D_{i\ j-1}\}, i = \overline{1,n}, j = \overline{1,m}. \quad (14)$$

The result is $DTW(X,Y) = \frac{\sum_i^K d_{ij}}{K}$, where *K* is the number of values which are contained in the minimum path.

3 Clustering Quality Check

To check the clustering result it is necessary to verify as far as objects are close in the same cluster and as far as objects are differ in different clusters. Quality functional is the function which defines how the clustering result is close to a perfect solution. There are rules for calculation of quality functional.

Let's find the sum of intra-cluster distances. It is the sum of distances between objects which are located in the same cluster:

$$F_0 = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho(x_i,x_j)u(x_i,x_j,y_i,y_j)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}u(x_i,x_j,y_i,y_j)}, \quad (15)$$

$$u(x_i,x_j,y_i,y_j) = \begin{cases} 1, & y_i = y_j | x_i \in y_i, x_j \in y_j \\ 0, & y_i \neq y_j | x_i \in y_i, x_j \in y_j \end{cases}, \quad (16)$$

where n is the number of objects; $\rho(x_i,x_j)$ is defined similarity function; $u(x_i,x_j,y_i,y_j)$ is the membership function of object x_i to the cluster y_i.

Let's find the sum of inter-cluster distances. This is the sum of distances between objects which are located in different clusters:

$$F_1 = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho(x_i,x_j)\left(1-u(x_i,x_j,y_i,y_j)\right)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(1-u(x_i,x_j,y_i,y_j)\right)}. \quad (17)$$

If it is needed to compare several clustering results, the best result has minimal value of the functional F_0/F_1.

If it is possible to initialize centroids, the additional quality functional can be used:

$$\Phi_0 = \sum_{j=1}^{m} \sum_{\substack{i=1 \\ x_i \in y_j}}^{n} \rho^2(x_i, \mu_j), \tag{18}$$

where n is the number of objects; m is the number of clusters; $\rho(x_i, x_j)$ is the defined similarity function; μ_j is the centroid of the cluster j.

$$\Phi_1 = \sum_{j=1}^{m} \rho^2(\mu_j, \mu), \tag{19}$$

where μ is the center mass of the data. Similar to functional F_0/F_1, if it is necessary to compare several clustering results, the best one has the minimum value of functional Φ_0/Φ_1.

4 Description of the experiment

The experiment to research the use of k-means and DBSCAN for model time series with additive white noise was conducted. A set on which clustering had performed, consisted of m time series of different types. The types were harmonic realizations, parabolas and «bursts». All realizations had random shift on the Y axis. The typical realizations of clustering are shown on Fig. 1a.

The white noise realizations were used for adding noise to time series that shown on Fig. 1 b. White noise was an independent values of a random variable with normal distribution $N(0, \sigma)$. The variance of noise had values σ^2={0.5, 0.75, 1.0, 1.25}. Thus, each time clustering was carried out for the same time series, but with a different level of additive noise.

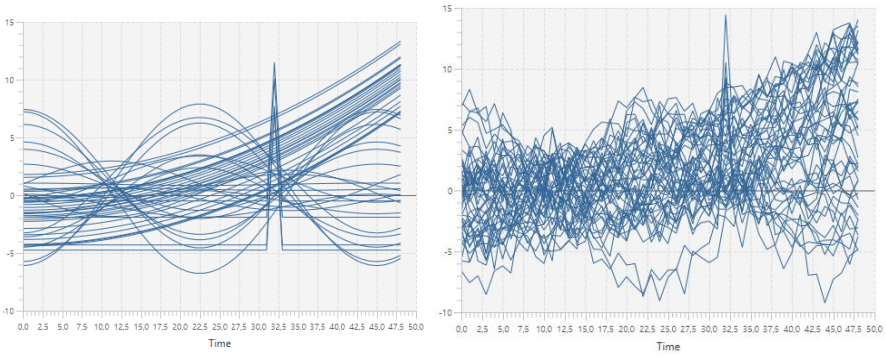

Fig. 1. Typical realizations for clustering a) «clean» realizations; b) with noise, σ=1.

We expected that we get at least 3 clusters as a good clustering result. The first one should consist of harmonic realizations, the second one should consist of parabolas and the third one should consist of atypical objects with «bursts».

DBSCAN and K-means methods for clustering were chosen. K-means input parameters were 3 centroids. DBSCAN input parameters were selected experimentally. To compare time series similarity, the similarity functions MJC, DTW, CID and Euclidean distance were used.
Thus, the next combination of clustering methods and similarity functions for each sample set of time series with noise were used:

- DBSCAN with Euclidean with CID function (Euclidean + CID);
- DBSCAN with MJC with CID function (MJC + CID);
- DBSCAN with DTW;
- K-means with Euclidean with CID function (Euclidean + CID);
- K-means with MJC with CID function (MJC + CID);
- K-means with DTW.

The results of the clustering were evaluated both visually by checking whether the objects hit the desired cluster, and using the functions of clustering quality F_0/F_1 and Φ_0/Φ_1. Thus, the similarity functions and the clustering method, which give the best results, were defined. The final part of our experiment is clustering of time series which contain a real data taken from [14].

5 Research results

5.1 Clustering of "clean" time series.

The feature of the sample data is having of atypical realizations («bursts»). According to the clustering results the separation of atypical realizations in the same cluster was successful only using DBSCAN method despite the fact that one of initial centroids for k-means method was atypical object.

Clusters obtained k-means method had time series of different shape and atypical time series in same cluster. The typical distribution in clusters by k-means method shown on Fig. 2.

The best time series clustering result was obtained by using Euclidean distance with CID function. This similarity function has the lowest iterations number and thus it has high performance. For MJC with ACID function it is necessary to calculate the value twice for each time series pair because of asymmetry property. This has impact on performance. The clusters obtained by the DBSCAN method and Euclidean distance with CID function are shown on Fig.3.

The quantitative indicators of clustering quality are represented on Table 1. It should be noted the experiment has shown that low values of F_0/F_1 for k-means method can comply with incorrect distribution in clusters. Therefore, Φ_0/Φ_1 functional was chosen to compare clustering results obtained by k-means method.

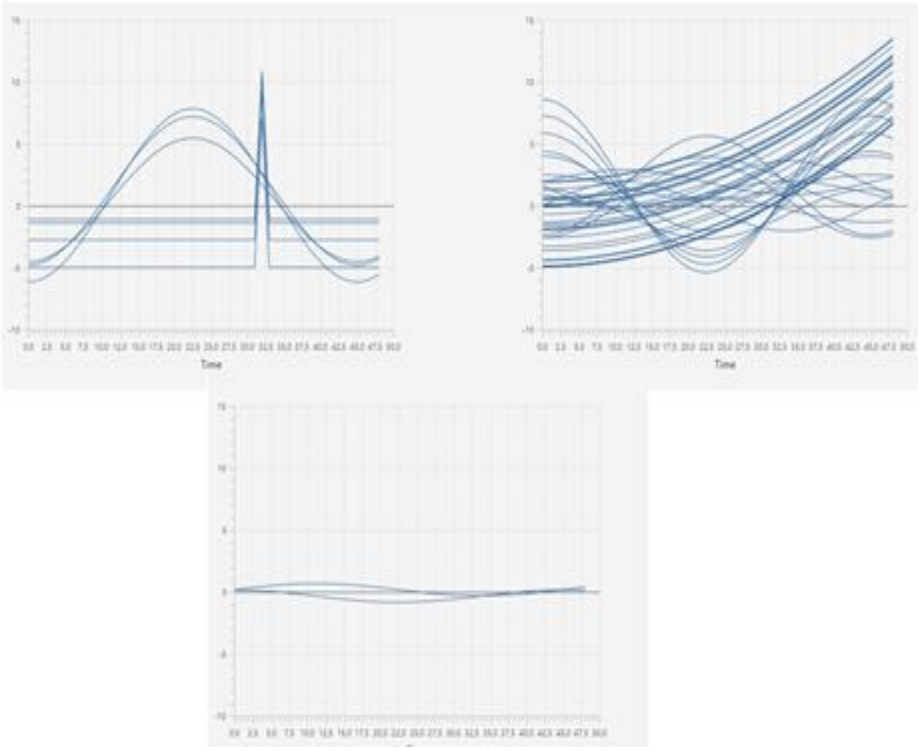

Fig. 2. Splitting into clusters by k-means method.

The incorrect splitting refers to time series with different shape in same cluster. Thus, the best results were obtained by using Euclidean distance with CID function.

Table 1. Values of quality functionals for clustering of "clean" time series.

Measure	Methods	F_0/F_1	Φ_0/Φ_1
Euclidean+ CID	K-means	0.415	0.222
	DBSCAN	0.326	
MJC + CID	K-means	0.184	0.197
	DBSCAN	0.480	
DTW	K-means	0.748	1.095
	DBSCAN	0.954	

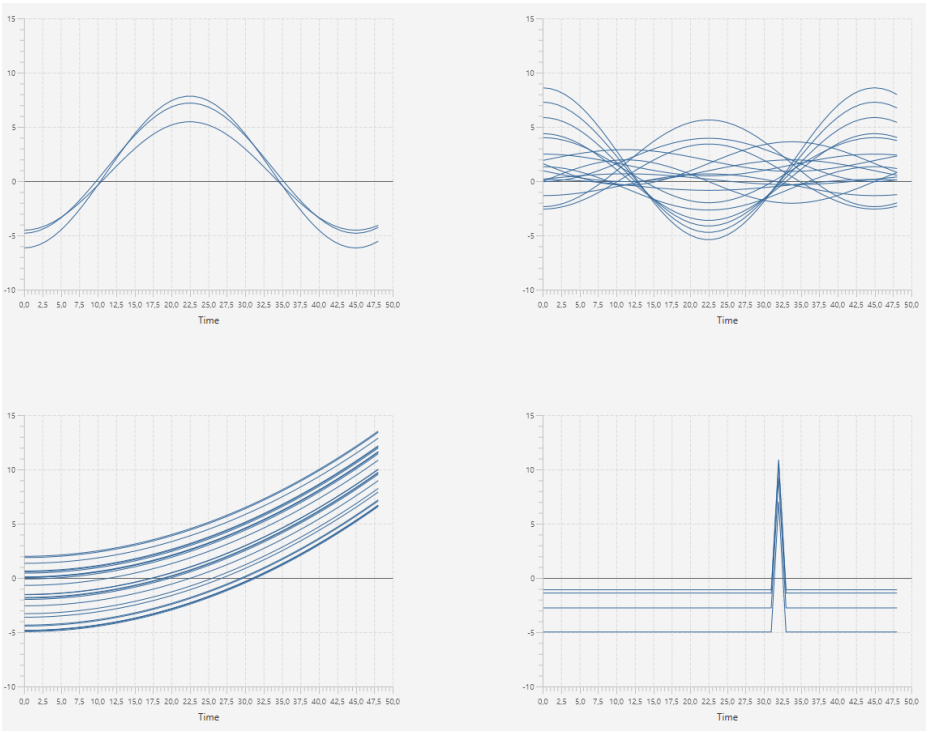

Fig. 3. Splitting into clusters by DBSCAN method.

5.2 Clustering of time series with noise

The results of clustering time series with noise are similar to results of clustering of "clean" time series. The best results are obtained by using DBSCAN method with Euclidean + CID function. The distribution in cluster using k-means method with MJC + CID is incorrect despite low values of quality functional. The time series with significantly different shape were in same cluster. The clustering for noise level σ^2=1.25 by DBSCAN method are shown on Fig.4. In this case time series, which have "bursts" are located in separate cluster. The noise also caused the harmonics with small amplitudes to be separated into separate clusters.

The Table 2 represents quantitative values of clustering quality. If we compare DTW function with Euclidean + CID and MJC + CID, it should be noted that DTW has caused to time series with different shape were in same cluster despite low values of Φ_0/Φ_1 functional. Such splitting is incorrect for the set.

Table 2. Values of quality functionals for clustering of time series with noise.

Measure	Methods	F_0/F_1	Φ_0/Φ_1
Euclidian + CID	K-means		7.736
	DBSCAN	0.684	
MJC + CID	K-means		12.401
	DBSCAN	1.065	
DTW	K-means		2.448
	DBSCAN	0.988	

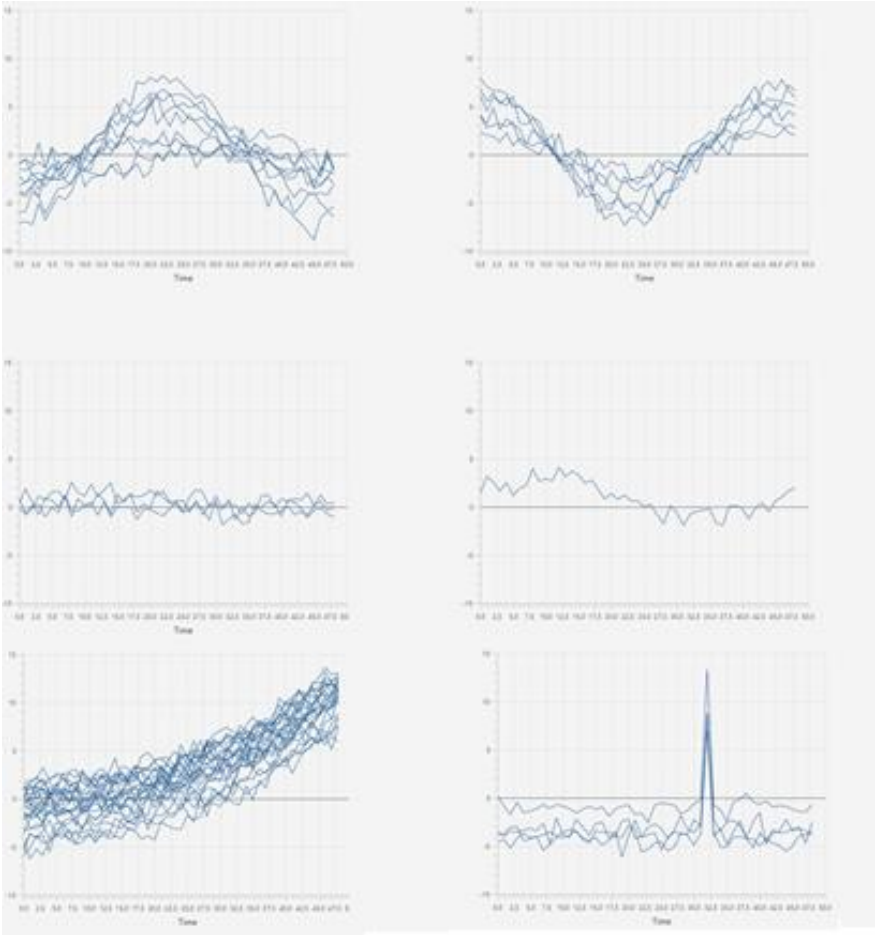

Fig. 4. Clustering of time series with noise by using DBSCAN method with Euclidean distance and CID function.

The Table 3 represents changes of quantitative rates of clustering quality with increasing noise variance for the DBSCAN method. The value $\sigma^2 = 0$ corresponds to sample with "clean" realizations.

Table 3. Changes of quantitative rates with increasing noise variance.

σ^2	F_0/F_1
0	0.326
0.5,	0.617
0.75	0.657
1.0	0.684
1.25	0.764

DBSCAN method showed good results defining atypical objects despite high level of noise. This method is also sustainable to different levels of noise. It allows to use this method for clustering real data, in which there are always various noises.

5.3 Clustering of real data

Consider the clustering of data that is the result of a medical research study of brain activity [14]. The data is a set of time series, which have some atypical objects. The initial sample of implementations for which clustering was performed are shown on Fig. 5.

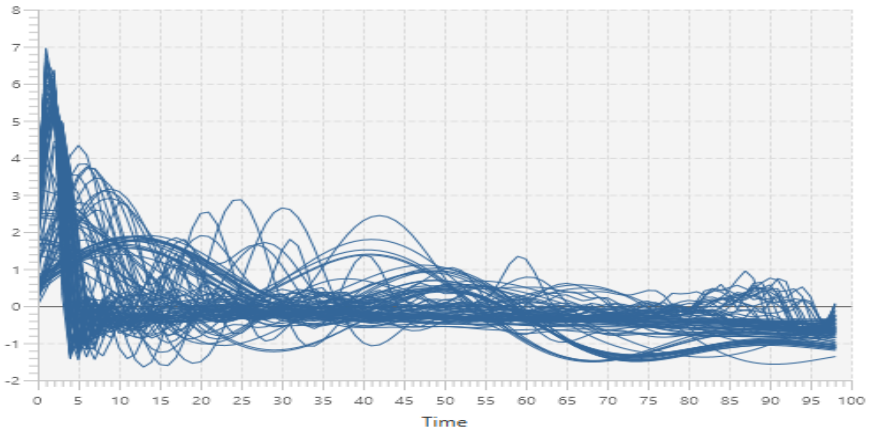

Fig. 5. Original sample of medical data.

We have used clustering methods and similarity functions described above. According to the values of quality functionals (Table 4) DBSCAN method with Euclidean + CID function has the best results.

Table 4. The result of clustering of real time series.

Measure	Methods	F_0/F_1	Φ_0/Φ_1
Euclidian + CID	K-means	0.492	1.102
	DBSCAN	0.478	
MJC + CID	K-means	0.568	1.325
	DBSCAN	0.534	
DTW	K-means	0.891	0.843
	DBSCAN	0.974	

As a result, 3 clusters were received. One of them is atypical object. The splitting into clusters are shown on Fig. 6.

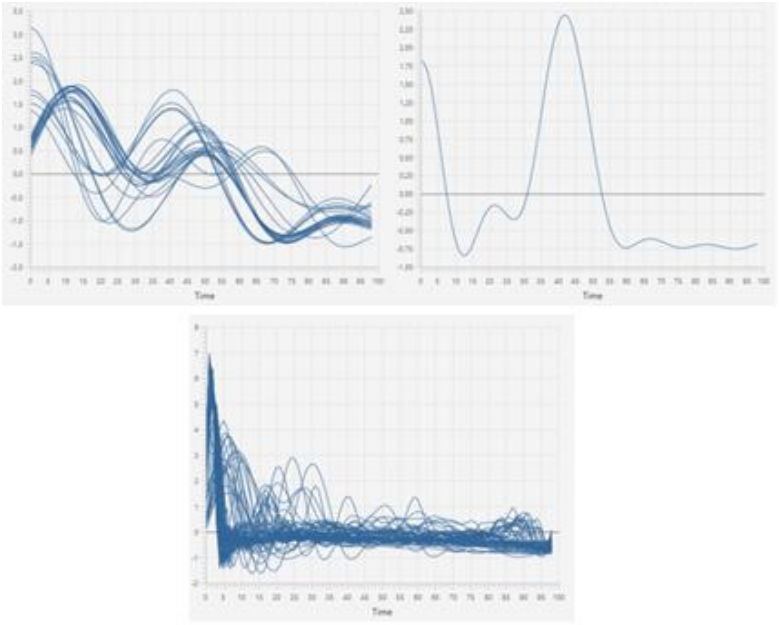

Fig. 6. The result of clustering by using the DBSCAN method with Euclidean + CID.

6 Conclusion

In this work the clustering of time series of different types with noise components was performed. The DBSCAN and k-means methods with different similarity functions for time series have been used. The DBSCAN method with Euclidean and CID function showed the best results. The results of clustering real time series confirmed the good application of the DBSCAN method for time series.

The analysis of clustering results allows to define the key differences between the k-means and DBSCAN methods for clustering time series: if it is possible to

determine number of clusters or their centroids and it is not required to separate of atypical objects, k-means method shows pretty good results; if there is no information about number of clusters and there is the task of separating atypical objects, it is possible to use the DBSCAN method.

References

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.J.: Time-series clustering. A Decade Review Information systems 53, 16-38 (2015).
2. Kirichenko, L., Radivilova, T., Zinkevich, I. Comparative Analysis of Conversion Series Forecasting in E-commerce Tasks. Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, Springer, Cham 689, 230-242 (2018). doi: 10.1007/978-3-319-70581-1_16
3. Lyudmyla, K., Vitalii, B., Tamara, R. Fractal time series analysis of social network activities. In: 2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, Ukraine, 456-459 (2017). doi: 10.1109/INFOCOMMST.2017.8246438
4. Bulakh, V., Kirichenko, L., Radivilova, T. Time Series Classification Based on Fractal Properties. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 198-201 (2018). doi: 10.1109/DSMP.2018.8478532
5. Aggarwal, C., Reddy, C.: Data Clustering: Algorithms and Applications. CRC Press (2013).
6. Liao, T.W. Clustering of time series data – a survey. Pattern Recognition, 38 (11), 1857-1874 (2005). doi: https://doi.org/10.1016/j.patcog.2005.01.025
7. Rani, S., Sikka, G.: Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications 52 (15), 1-9 (2012). doi: 10.5120/8282-1278
8. Grabusts, P., Borisov, A.: Clustering methodology for time series mining (2009). Scientific Journal of Riga Technical University 40, 81-86 (2009). doi: 10.2478/v10143-010-0011-0.
9. Serrà J., Arcos J.L.: A Competitive Measure to Assess the Similarity between Two Time Series. In: Agudo B.D., Watson I. (eds) Case-Based Reasoning Research and Development. ICCBR 2012. Lecture Notes in Computer Science, 7466. Springer, Berlin, Heidelberg, 414-427 (2012). doi: https://doi.org/10.1007/978-3-642-32986-9_31
10. Gustavo, E.A., Batista, P.A., Wang, X., Keogh, E. J.: A Complexity-Invariant Distance Measure for Time Series. Data Mining and Knowledge Discovery 28(3), 1-36 (2013). doi: 10.1007/s10618-013-0312-3
11. Marteau, P.F.: Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching. In: IEEE Transactions on Pattern Analysis and Machine In-telligence 31(2), 306 – 318 (2009)
12. Barreto, G., Aguayo, L.: Time Series Clustering for Anomaly Detection Using Competitive Neural Networks. In: Proceeding WSOM '09 Proceedings of the 7th International Workshop on Advances in Self-Organizing Maps, St. Augustine, FL, USA, 28-36 (2009). doi: 10.1007/978-3-642-02397-2_4.
13. Nascimento, E.S., Tavares, O.L., Souza, A.F.: A Cluster-based Algorithm for Anomaly Detection in Time Series Using Mahalanobis. In: ICAI'2015 International Conference on Artificial Intelligence 2015, Las Vegas, USA 622-628 (2015).
14. Time series classification, http://www.timeseriesclassification.com last accessed 2019/28/02