Development and Study of a Combined Algorithm for Temporal Series Clustering

Sibirev Ivan,
Afanasyeva Tatyana
Ulyanovsk State Technical University,
Ulyanovsk, Russia
Email: ivan.sibirev@yandex.ru
Email: tv.afanasjeva@gmail.com

Abstract—**This paper offers a combined algorithm for temporal series clustering based on the primary clustering of the points of each temporal series, and then on the secondary clustering of a set of temporal series by a set of parameters that are statistical characteristics of the primary clusters.**

The paper describes the experiments to reveal the minimum set of parameters for the temporal series clustering, and compares the results of the algorithm operation for various methods of clustering.

The proposed algorithm allows the clustering of temporal series with various numbers of points, with various time scales; it unites into one cluster the temporal series with a consistent similarity of parts of the graphs with the accuracy up to their contraction, extension, shift along the OX and OY axes, and symmetries.

Keywords—clustering, temporal series, parametric clustering, centroid method, Ward's method, single linkage method.

I. Introduction

The study of time series, their semantics , extract knowledge is interest because time series have many applications in various fields of science and technic, economy, medicine, etc. Often study of data sets are using as unlabeled time series. A common task is defining groups of homogeneous time series. The results of analysis can be used to monitor multiple processes in different time periods.

The time series clustering increased interest in order to analyze data dynamics. Works about dynamic clustering (clustering of time series) is more less than – static clustering. However, the interest in this topic grows.

Initially, cluster analysis was developed as methods designed to work with static data, for which values remain constant over time or change slightly. The target of cluster analysis is identification of structure in studied data set by organizing homogeneous groups for which intergroup differences and intragroup similarities are minimal. Clustering is useful when data sets is not structured and belong to one of the following types: binary, numeric, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia or their combination.

Modern software is using for on clustering static data. The software exists in form of application programs or as part of software packages for data processing or data mining.

II. An overview of the approaches clustering time series

The clustering methods are the divisive clustering, agglomerative clustering, density based clustering, mesh methods, and model methods [1]. The paper [2] gives the description of each clustering method category. The following clustering methods are frequently used: single-linkage clustering [3], average linkage clustering [14], complete-linkage clustering [5], Ward's method [6], centroid linkage clustering [7], [8], [9], [10], K-mean clustering [11], [12], and others.

The hierarchical methods of clear clustering have a weak point which is the correction of elements belonging to the clusters after the clusters are formed. The papers by Karypis, Han, Kumar [13] and Guha, Rastogi, Shim [14] show the linkage analysis at each phase of the hierarchical splits. The paper by Zhang, Ramakrishnan, and Livny [15] uses the iteration permutation to specify the results obtained using the hierarchical agglomerative clustering more correctly.

In the clear clustering algorithms, each object belongs to one cluster; in the fuzzy clustering algorithms, each object belongs to several clusters. FCM is the fuzzy clustering method [16] used together with one of the clear clustering algorithms, the choice of which determines the result. In papers [17], [18], the fuzzy C-method is described. This is a heuristic algorithm successfully used for finding spherical clusters on small and medium data arrays.

The data extraction from temporal series using the cluster analysis methods attracts heightened interest. More often, the temporal series clustering also known as dynamic clustering is discussed for the dynamic data analysis. At the temporal series clustering, some measure of similarity, based on the representation of the temporal series in the form of a model or a set of parameters, is required. The existing dynamic series clustering methods can be classified by the manner the input data is treated: directly – with input data, frequency-temporal characteristics, area data, models or functions obtained from the input data.

The dynamic clustering conditionally includes two stages: obtaining the set of parameters by which clustering is to be performed; selection of the static clustering method and its application. Obtaining the set of parameters, and their application are the subject of today's intense study and discussions. At that, the issues of choosing the adequate set of parameters, the choice of the static clustering method responsible for the problem setting, the assessment of the clustering results compliance with the problem setting and the objectives of the study still remain open.

Static clustering methods are used at the temporal series clustering. Special literature concerning this topic shows the interest to the following methods:

- The density based clustering (the paper [19] develops the idea of "a cluster as the density that does not exceed some schedule in a certain area").

- The methods based on the quantum mesh with a finite number of cells on which clustering is performed. The typical example of such method is given in paper [20]. It uses several levels of rectangular cells that correspond to different levels of resolution. Statistical information is calculated. According to the attributes in each cell, the mesh cells parameters are considered starting from the largest allowable scale. For each cell in the current layer, the confidence interval reflecting the correspondence of the cell to the given (current) request is calculated. The cells that do not agree with the request are removed from the consideration. The request process continues on the next level – the lower one – for the selected cells, until the lower level is reached.

- The model based methods that compare a model with each cluster. At that, the set of models is compared with the data array. Both can be primary. At present, static and neural network approaches to the model description are offered. An example of the static data description approach is given in the paper [21], where the Bayesian static analysis for the evaluation of the cluster number. Examples of the neural network approach are given in [22], [23].

III. Problem setting

We have conducted a series of experiments for the study of temporal series clustering where the lists of standard parameters of cluster statistical description obtained by separate clustering of each temporal series were used as the data for clustering.

The task was the obtaining of visually similar temporal series clusters with the accuracy up to the normalization along the OX and OY axes.

The objectives were: the development and testing of the algorithm of combined temporal series clustering; the study of the set of parameters (compliance of the selection to the set task, minimization of their number); the study of the static clustering method applicability in the proposed algorithm.

The outcome of the experiment was the obtaining, basing on the statistics and cluster analysis methods, of a minimum set of parameters that characterizes the temporal series. This set is used for the temporal series reclustering.

Input data description. 72 temporal series are used as the input data. The temporal series belong to different types, i.e. they have different numbers of points, global tendencies, spans of values, etc. For each separate series, the coordinates of the points were normalized.

IV. Description of the proposed algorithm

We are going to cluster a temporal series into N1 clusters; at the clustering, the data is normalized along the reference axis. For each cluster, we get an ordered set of values. The studied parameters were the dispersion, maximum, minimum, span and arithmetic mean value for the X and Y coordinates and the temporal series points. We sort the clusters in the chronological order concerning the clusters center X coordinate. In this order, we write down the parameters that characterize the first cluster, the second one, ... the N1 cluster. For each temporal series, we get a list of parameters of dimensionality which is the product of the number of studied parameters multiplied by the number of N1 clusters. We form the lists of parameters obtained using the above described method for each studied temporal series. These lists are used for the clustering into N2 clusters.

The algorithm uses the clustering procedure twice. First, we perform the points clustering for each temporal series. Second, we perform clustering of the set of temporal sequences by the parameters that are the temporal series. It is appropriate to minimize the set of parameters due to the exclusion of the mutually depending quantities. At that, the results of the second clustering remain almost unchanged. The experiments have shown that the temporal series clustering results utterly do not change if we go from the set of parameters {X dispersion, X maximum, X minimum, X span, X arithmetic mean value, Y dispersion, Y maximum, Y minimum, Y span, Y arithmetic mean value} to the set {X dispersion, X arithmetic mean value, Y dispersion, Y arithmetic mean value} for each cluster after primary clustering.

V. Experimental study of the proposed algorithm

We have conducted a series of experiments for the above described algorithm using the following clustering methods: centroid linkage clustering, Ward's method, single linkage method. 72 temporal series were clustered into N2=10,20,30,40,50 clusters; at that, the clustering into N1=2,3,4,5,6,7,8,9 clusters was applied for obtaining the list of parameters. All in all, we conducted a series of 120 experiments. We wrote a program in the C# programming language for automated experiment conduction.

During the automated experiment, each temporal series was normalized (the normalization was made over the OX and OY axes); the results for each cluster were recorded in a png. file in the normalized form.

By setting N1, we specify the number of clusters into which each temporal series is to be clustered, i.e. what number of multi-dimensional points with the coordinates {X dispersion, X maximum, X minimum, X span, X arithmetic mean value, Y dispersion, Y maximum, Y minimum, Y span, Y arithmetic mean value] to the set {X dispersion, X arithmetic mean value, Y dispersion, Y arithmetic mean value] we were going to use to represent the temporal sequence. We have the possibility to compare the temporal series over these multi-dimensional points. When N1=2, the series were compared over two points. This was a too rough comparison. When N1=9, we obtained an exceedingly sensitive tool of comparison requiring a high degree of conversion for the temporal series; at that, each sequence was found in a separate cluster. We have discovered experimentally that the values close to 6 were the most appropriate ones.

By using in turn N2 = 10,20,. . . ,50, we could follow the clustering progress for all temporal series. The obtained clusters can be conditionally split into 3 groups by the degree of similarity: groups of series similar by behaviour (see Fig. 1) – "strongly similar"; groups of series with relative similarity (see Fig. 2) – "moderately similar"; and the clusters that include one or, less frequently, two temporal series – "solitary" without similar ones (see Fig. 3). Let us mention that the source temporal series had absolutely different areas of determinations and values and sometimes differed by 5-6 orders.

In the case of N2=10, a significant part of the temporal series got into the second group. As N2 was increased from 10 to 50, we observed the growth of the number of clusters belonging to the first group. For all values of N2, about a half of the clusters were referred to the third group.

We discovered that the most successful clustering turned out to be that with N2=20 and 50. At N2=50, we can observe the initial stages of the agglomerative clustering when the "strongly similar" temporal clusters are grouped. At N2=20, for this set of temporal series, we managed to obtain the most informative and expressive clustering results.

We have discovered that the result of the algorithm operation depends on the chosen clustering methods.

The centroid method showed the best results out of the used methods (see Fig. 1). The experimental results provided the separation of the temporal series into three above mentioned groups of clusters, at that the separation turned out to be quite sharp. The presence of one or two multi-component clusters and big amounts of fringe clusters is characteristic for the method. For the centroid method, in the series of conducted experiments the discard of the major part of the temporal series for which no similar series were found, into solitary clusters was characteristic (see Fig. 3).

At the comparison of temporal series, the issue is the identification of the different-length areas where the graphs are similar with the accuracy up to the graph contraction or extension along the OX and OY axes with the accuracy up to the shifts and symmetry. The paper [24] is devoted to this issue. Here the issue of identification of the graph areas similar with the accuracy up to contractions, extensions and shifts is solved automatically, as the secondary parameters at the reclustering are also subject to normalization. For each temporal series, the OX axis is split into various-length sections D1, D2, ..., DN1. For all series the parameters associated with the sections D1, D2, ..., DN1, respectively, are compared. They undergo normalization at the parameters reclustering (i.e. the contraction or extension along the OX and OY axes is performed separately for all series in the D1 sections, then in the D2 sections, etc.). Thus, the proposed algorithm is sensitive to the presence, in all compared series, of sequential sections of graphs similar to the accuracy up to contraction, extension and shifts along the OX and OY axes.

The same effect is observed at the use of the single linkage method (see Fig. 4). The single linkage method has shown itself quite well: quite similar series are separated (see Fig.4). This algorithm is able to group the temporal series similar to the accuracy up to the symmetries of separate sections (see Figs. 5-6).

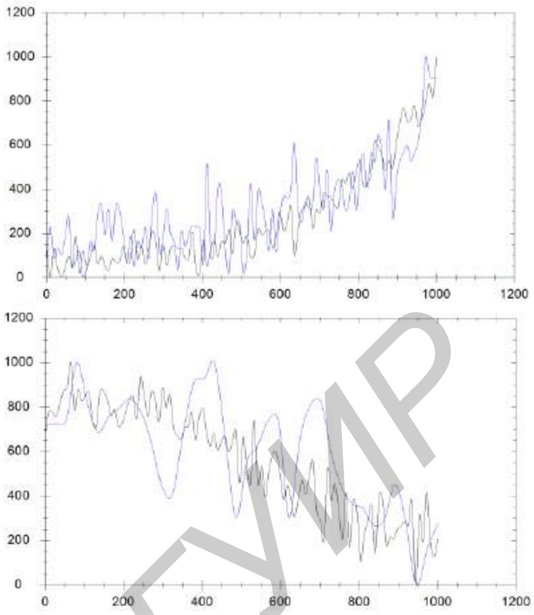

Figure 1. N1=6, N2=20, centroid method, "strongly similar".

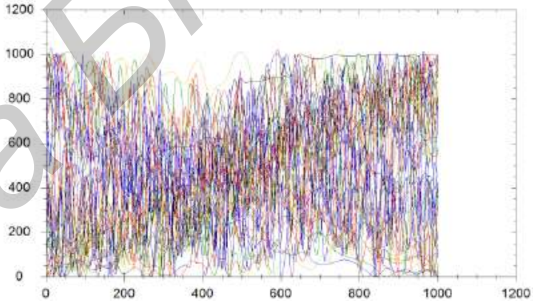

Figure 2. N1=6, N2=20, centroid method, "moderately similar".

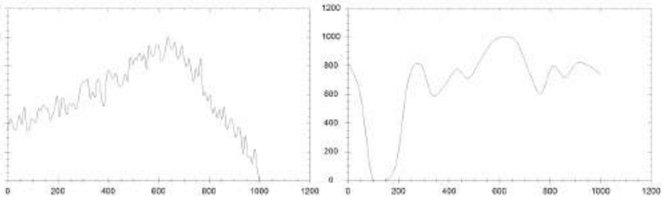

Figure 3. N1=6, N2=20, centroid method, "weakly similar".

Ward’s method is based on the aggregate dispersion minimization. The results of the algorithm application with the use of the Ward’s method are given in Figs. 7-9; no pure visual similarity is observed there. This means that this method meets the set task least of all the used methods.

A particular case of the combined temporal series clustering algorithm is the clustering at N1=1 or one-stage clustering. A set of statistical parameters is imposed to correspond to each temporal series; these sets are used for clustering all temporal series. The series are compared "over one point". The combined temporal series clustering provides better result, since it allows comparison over N1 point. The results at N1=6 and N1=1 are shown in Figs. 1-3 and 10-11, respectively.

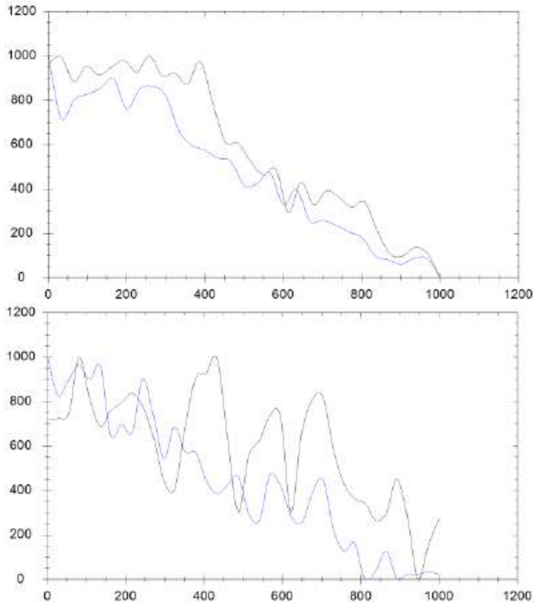

Figure 4. N1=6, N2=20, single linkage method, "strongly similar".

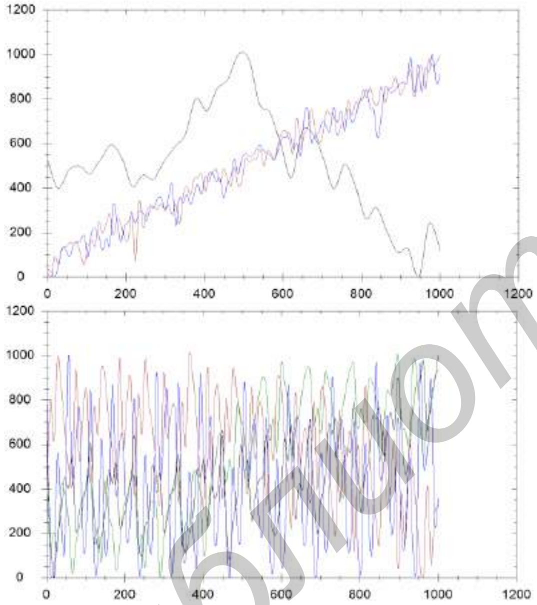

Figure 5. N1=6, N2=20, single linkage method, "moderately similar".

In practice, the clustering with initially set centers, e.g. using the K-mean method, can be helpful. One and the same cluster analysis method shall not be obligatorily used both for obtaining the temporal series parameters and for clustering over these parameters. The clustering methods can be combined.

Fuzzy clustering methods can be used in this algorithm, but with account for their application specifics. Our team develops such algorithm modifications, but they go beyond the scope of this paper.

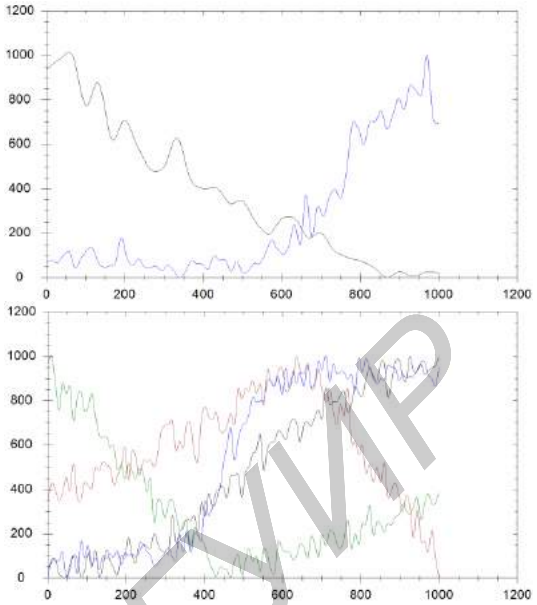

Figure 6. N1=6, N2=20, single linkage method, "weakly similar".

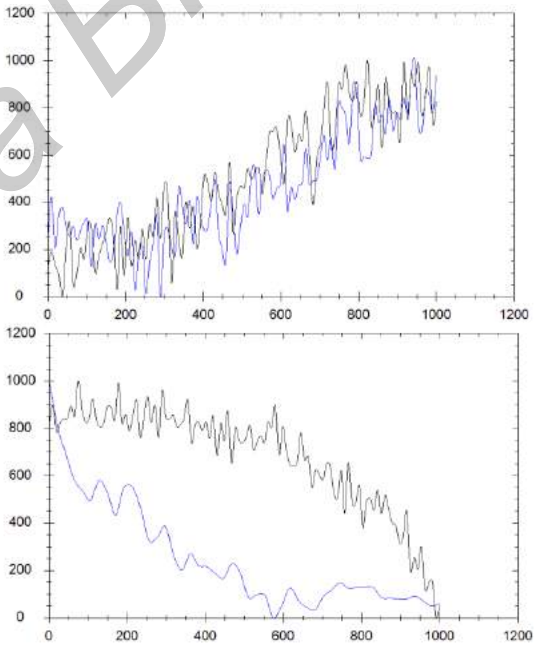

Figure 7. N1=6, N2=20, Ward's method, "strongly similar".

VI. Conclusions

The proposed algorithm can be applied for temporal series clustering for the purposes of identification of series with visually similar behaviours with the normalization accuracy along the OX and OY axes.

The developed temporal series clustering program has shown its performance and allowed us to conduct a series of 120 experiments in an automated manner within rather short time.

The algorithm is applicable for the comparison of various-length temporal series. Other activities, for the purposes of

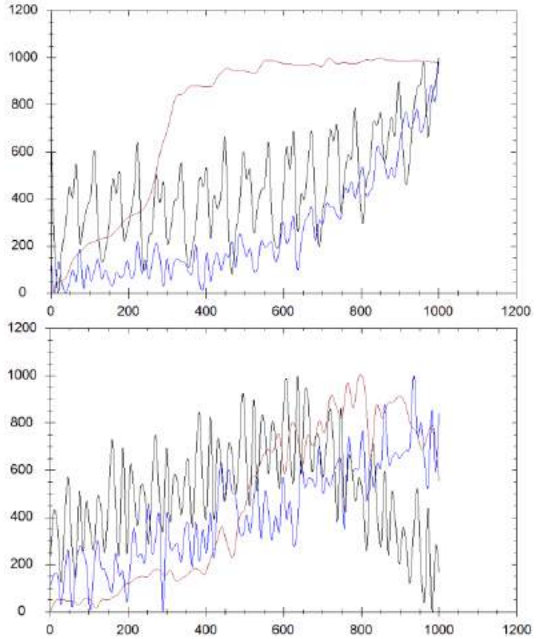

Figure 8. N1=6, N2=20, Ward's method, "moderately similar".

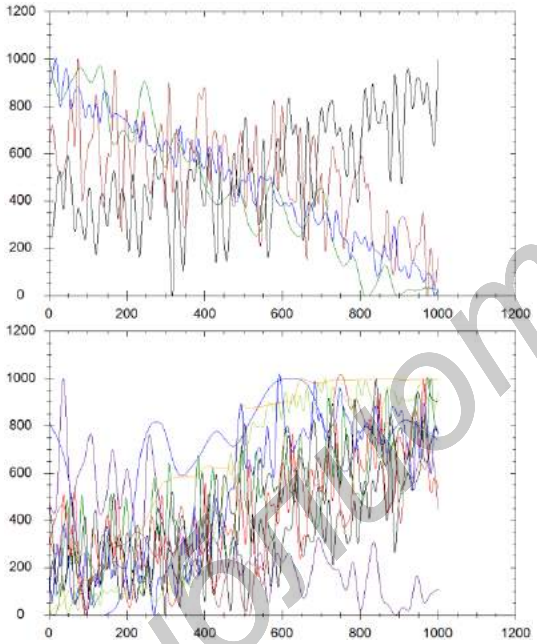

Figure 9. N1=6, N2=20, Ward's method, "weakly similar".

temporal series comparison, use artificial methods (arithmetic, statistical, reduction) for bringing the temporal series to one number of studied points. When our algorithm is used, there is no need to do this.

The algorithm compares the series with different temporal scales.

The proposed algorithm is sensitive to the presence, in all compared series, of sequential sections of graphs similar to the accuracy up to contraction, extension and shifts along the OX and OY axes. The issues of temporal series splitting into various-length homogeneous sections, the issues of contraction

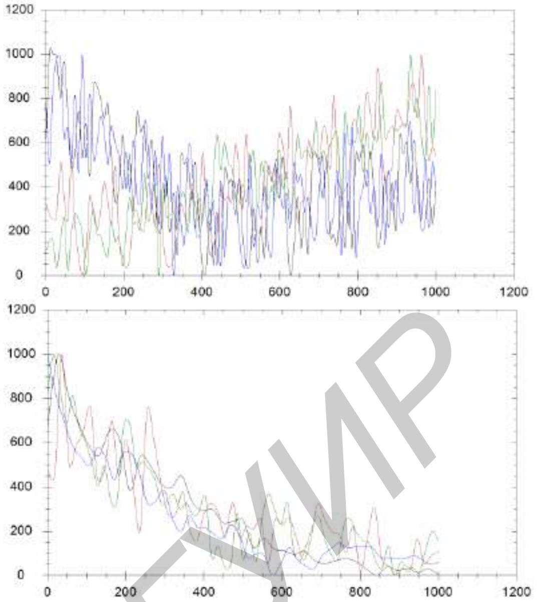

Figure 10. N1=1, N2=20, centroid method, "strongly similar".

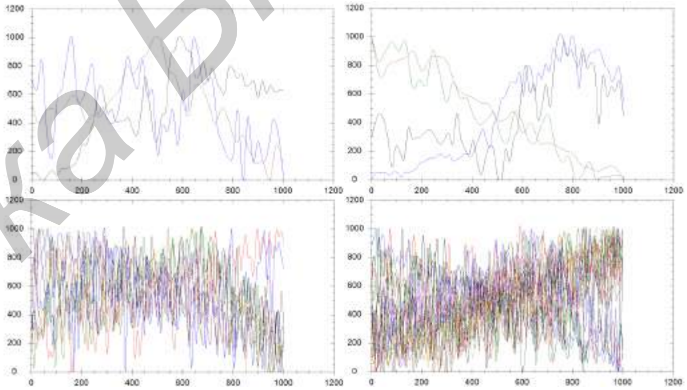

Figure 11. N1=1, N2=20, centroid method, "moderately similar".

and extension of these sections are solved on the level of normalization during cluster analysis in automated manner. The use of single linkage method in the algorithm as the clustering method allows the identification of the series similar to each other with the accuracy up to the symmetry of separate sections of graphs.

In our algorithm, different clustering methods can be used depending on the study objective and traditions of the field. Clustering is used twice (for the points of each temporal series and for the sets of parameters corresponding different temporal series). Different clustering methods can be combined.

At the comparison of experimental results, when different clustering methods were used, the centroid method showed itself quite well. At the reclustering, the single linkage method provides clusters with approximately similar order with quite similar temporal series as the components. The single linkage method identifies the temporal series similar to each other with the accuracy up to the symmetry of separate sections of graphs; with some modification the separation of graphs with symmetrical areas into one cluster can be eliminated. Ward's method in this algorithm showed unsatisfactory results.

Temporal series are clustered by a set of parameters which are the statistical characteristics of the primary clustering clusters for each temporal series. Our experiments have shown that the reduction of the number of parameters down to {X dispersion, X arithmetic mean value, Y dispersion, Y arithmetic mean value} for each primary clustering cluster utterly does not change the clustering results.

We have obtained a temporal series clustering algorithm with a high degree of modularity. By selecting the problem-oriented sets of mutually independent parameters of primary clustering of each temporal series, using classical clustering methods or those specially developed for applied problem we can gain the temporal series clustering for a given problem.

We are working on the fuzzy modification of this algorithm; we study its properties and application to gain knowledge on the sets of temporal series.

This work was supported by the Russian Foundation for Basic Research grant 16-47-732112 "Study and development of forecast methods for temporal series based on multi-model approach".

References

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001 p. 346–389.

[2] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001 p. 346–389.

[3] Sibson R. SLINK: An optimally efficient algorithm for the single-link cluster method, The Computer Journal, Vol. 16, 1973, p. 30-34

[4] Мандель И.Д. Кластерный анализ /И.Д. Мандель. – М.: Финансы и статистика. 1988. – 176с.

[5] Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content // Kongelige Danske Videnskabernes Selskab. Biol. krifter., Bd V, № 4, 1968, p. 1-34

[6] Ward J.H. Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc., Vol. 58, 1963, p. 236-244

[7] Болч, Б. Многомерные статистические методы для экономики / Б. Болч, К. Дж. Хуань. / Пер. с англ. – М.: Статистика, 1979. – 317с

[8] Портер, М. Международная конкуренция. Конкурентные преимущества стран / М. Портер. – М.: Междунар. отношения, 1993. – 896 с.

[9] Сошникова, Л.А. Многомерный статистический анализ в экономике / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер. – М.: ЮНИТИ-ДАНА, 1999. – 598 с.

[10] Типология и классификация в социологических исследованиях. Отв. ред. В.Г. Андреенков, Ю.Н. Толстова. –М.: Наука, 1982. – 296с.

[11] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: L.M. LeCam, J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 281–297.

[12] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

[13] G. Karypis, E.-H. Han, V. Kumar, Chameleon: hierarchical clustering using dynamic modeling, Computer August (1999) 68–75. T. Warren Liao / Pattern Recognition 38 (2005) 1857 – 1874 1873

[14] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, Proceedings of the 1998 ACMSIGMOD International Conference on Management of Data, Seattle, WA, June 1998, p. 73–84.

[15] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, Proceedings of the 1996 ACM-SIGMOD International Conference on Management of Data, Montreal, Canada, June 1996, p. 103–114.

[16] Ярушкина, Н. Г. Интеллектуальный анализ временных рядов : учебное пособие / Н. Г. Ярушкина, Т. В. Афанасьева, И. Г. Перфильева. – Ульяновск : УлГТУ, 2010. – 320 с.

[17] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York and London, 1987.]

[18] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Lowcomplexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Systems 9 (4) (2001) 595–607.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A densitybased algorithm for discovering clusters in large spatial databases, Proceedings of the 1996 International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, 1996, p. 226–231.

[20] W. Wang, J. Yang, R. Muntz, R., STING: a statistical information grid approach to spatial data mining, Proceedings of the 1997 International Conference on Very Large Data Base (VLDB'97), Athens, Greek, 1997, p. 186–195.

[21] P. Cheeseman, J. Stutz, Bayesian classification (AutoClass): theory and results, in: U.M. Fayyard, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Cambridge, MA, 1996.

[22] G.A. Carpenter, S. Grossberg, A massively parallel architecture for a self-organizing neural pattern recognition machine, Comput. Vision Graphics Image Process. 37 (1987) 54–115.

[23] T. Kohonen, The self organizing maps, Proc. IEEE 78 (9) (1990) 1464–1480.

[24] Tassos Argyros, Charis Ermopoulos, Efficient Subsequence Matching in Time Series Databases Under Time and Amplitude Transformations, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece.

РАЗРАБОТКА И ИССЛЕДОВАНИЕ КОМБИНИРОВАННОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ ВРЕМЕННЫХ РЯДОВ

Сибирев И.В., Афанасьева Т.В.

В данной работе предложен комбинированный алгоритм кластеризации временных рядов, основанный на первичной кластеризации точек каждого временного ряда, затем – на вторичной кластеризации множества временных рядов по набору параметров, являющихся статистическими характеристиками первичных кластеров.

Описаны эксперименты по выявлению минимального набора параметров для кластеризации временных рядов, сравниваются результаты работы алгоритма для разных методов кластеризации. Предлагаемый алгоритм позволяет кластеризовать временные ряды с разным количеством точек, разных временных масштабов, объединяет в один кластер временные ряды с последовательным сходством участков графиков с точностью до их сжатия, растяжения, сдвигов вдоль осей OX и OY и симметрий. Написанная нами программа позволила автоматически произвести 120 серий эксперимента в сравнительно короткие сроки. В качестве входных данных используется набор из 72 разнотипных временных рядов.