A System for Clustering of Time Series

Dr. Ventsislav Nikolov, Aleksander Krastev

Eurorisk Systems Ltd., 31, General Kiselov, Varna 9010, Bulgaria

vnikolov at eurorisksystems dot com, akrustev at eurorisksystems dot com

Abstract: The clustering is an important subtask of many other actions like parallel processing, local data analysis, dimensionality reduction, etc. In this paper some aspects are considered about clustering of time series. The main accent is directed toward modification of the known clustering algorithms taking into account of the weights in both the time series values and the time series as a whole in the process of clustering as well as determining the optimal number of clusters and assessment of clustering quality statistics.

Keywords: Clustering, Time series, Clustering quality, Number of clusters, Self-organizing map

1. Introduction

Clustering of objects is an important task in many fields and it is a base for modeling and many processing techniques. It can be used for: classification, patterns recognition, dimensionality reduction, vectors quantization, data compression [2][4][5], etc. The main purpose of the clustering is creating of groups of patterns in such a way that the patterns with similar feature values to be in a same group. This could provide many advantages, some of which are the following:

- Parallel data processing. The data in a given cluster can be processed independently from the other clusters [3][7].
- Possibility for analysis only on specific parts of the whole available data [6][11].
- Better interpretability. Additional statistical analysis can be done about the clusters, like: dispersion, average linkage, etc. [21][23].
- Additional flexibility of data processing by further clustering of some clusters or merging of clusters. Thus hierarchical models can be developed [12].

Here the objects of clustering are time series and some modifications of the well-known approaches are proposed related to their characteristics: determination of the optimal number of clusters, different priorities for the time series according to their importance, different weights of the values in the series according to their position in the time axis, statistics about the clustering quality and interpretation of the cluster centers.

2. Clustering of time series

In the general case the objects that should be clustered are represented by temporal or spatial features. In our case these objects are time series and the time ordering of the observations is taken into account. The initial input data are:

- A set of time series that have to be clustered;
- The number of clusters in which the series must be grouped.

The aim is to obtain clusters of time series and every cluster consists of:

- A set of series belonging to the cluster;
- A generated synthetic series, called prototype series, considered as a representative of the cluster. This prototype series is with the same dimensionality as all other series and it is usually near to the center of the cluster.

There are many clustering techniques, for example: k-means [22], hierarchical clustering that can be divisive or agglomerative [21], self-organizing maps (SOM) [13], adaptive resonance theory (ART) [8][9], iterative self-organizing data analysis technique (ISODATA) [21], etc. Here the clustering experiments are done by self-organizing map, which is an unsupervised trained neural network proposed by Teuvo Kohonen [13].

In addition to the clustering procedure, wich groups time series and generates prototypes series, clustering quality statistics are also generated [21][22]: inter-cluster and intra-cluster distances, R^2, adjuster R^2, etc. Some of these statistics can be used as critaria to determine the optimal number of clusters.

One important application of the clustering module developed by the authors is to reduce large sets of time series to small sets of synthetic prototype series. Thus the prototype series can be used instead of the real series for time and memory consuming operations, such as calculations of correlation matrices, with as minimal error as possible. In this way the number of time series is reduced as any calculation that should be done with a given real time series is actually performed with the cluster prototype to which the real series is classified. The number of all available series is reduced to the number of clusters and as a consequence the calculation operations that must be performed with the series decrease. This should be done because the huge data causes too many calculations, which often cannot be finished within acceptable time period.

3. Proposed solution

The procedure of clustering performed by self-organizing map is a procedure of the unsupervised neural network training. During this process the similarity of the time series behaviour is examined and this is done not only for the current values but for all other historical values as well. That is why it is possible a given series, for example with high current values to be classified into a cluster containing lower current series values. This may occur when a series with different actual values to the other series for the current time moment, but similar historical behaviour to other series in the past.

Series weights

Individual series may have different importance expressed by coefficients called weights. If the weights are all equal to ‘one’, then all series are equally treated in the clustering process. Some of the series however may be considered more important than the others. If the weight of a series is two times bigger than the weight of another series, the former is considered as two series with the same series values in the clustering alorithm. The practical signifficance of this property is that the weights can be not only integers but any arbitrary real values as well.

The effect of the usage of weights is shown in fig. 1.

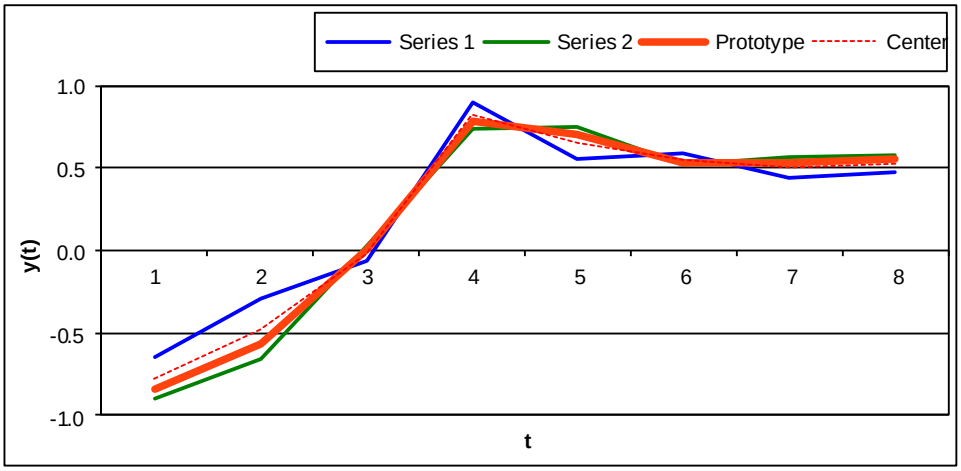


Fig. 1 The effect of a series weight to the cluster prototype

The ceneter between Series 1 and Series 2 is shown with a dashed line. If these two series are with the same weight, the center will be very close to the prototype series. However, in this example, Series 2 is with greater weight and thus it attracts the prototype shown with bold line.

Weights of series values

In the clustering algorithm the series must be compared one another by a chosen criterion. One of the most often used such criterion is the Euclidean distance (1) where compared series must be with the same length [4][10].

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1}$$

where x and y are the series that are compared;
i is the consequent index of x and y.

The Euclidean distance can take into account the importantce of the values by using a decay factor (2) [17].

$$d = \sqrt{\frac{1}{\sum_{i=1}^{N}\lambda^{i-1}}\sum_{i=1}^{N}\lambda^{N-i}(x_i - y_i)^2} \tag{2}$$

where λ is the decay factor which takes values between 0 and 1.

The closer to the 0 the decay factor is, the more important values are to the end of the series. When the decay factor is 1, then the formula coincides to the original Euclidean distance.

Clustering quality criteria

In order to assess how good the clustering is, a suitable criterion or criteria must be applied [14][17][20]. Some such criteria with their formulas are shown below.

- Adjusted R^2. In fig.2 the main parts of this statistics are shown.

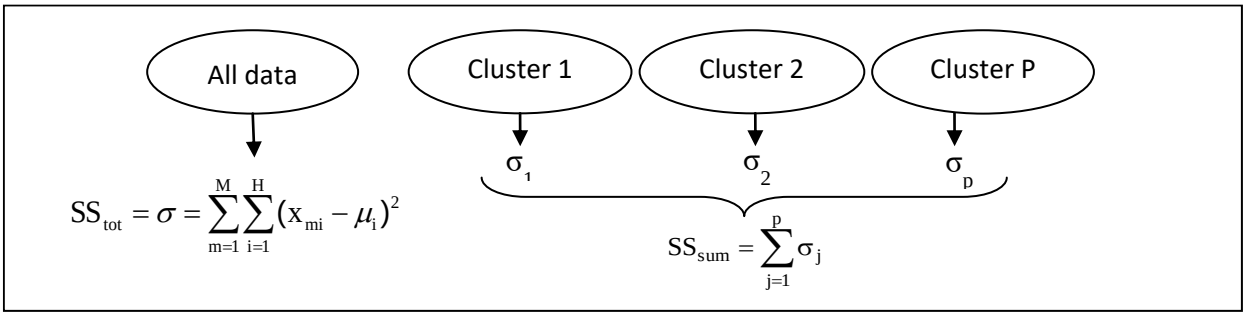


Fig. 2 Calculation of squared distances for Adjusted R^2

The following steps must be performed:

1) For all data the sum squared distance of each series to the center is calculated:

$$SS_{tot} = \sigma = \sum_{m=1}^{M}\sum_{i=1}^{H}(x_{mi} - \mu_i)^2 \quad (3)$$

where *M* is the number of all available series;
H is the number of values in the series;
x is current time series;
μ is center of all available series.

2) Similarly, the sum squared distance is calculated for each cluster *j*:

$$\sigma_j = \sum_{m=1}^{M_j}\sum_{i=1}^{H}(x_{jmi} - \mu_{ji})^2 \quad (4)$$

mhere M_j is the number of time series series in cluster *j*;
μ_j is the center of the series in cluster *j*.

3) The sum of the squared distances calculated in step 2 is calculated:

$$SS_{sum} = \sum_{j=1}^{P}\sigma_j \quad (5)$$

where *P* is the number of clusters.

4) R^2 is calculated as follows:

$$R^2 = 1 - \frac{SS_{sum}}{SS_{tot}} \quad (6)$$

5) And finally the adjusted R^2 is calculated as:

$$\text{Adjusted}R^2 = 1-(1-R^2)\frac{M-1}{M-P-1} = 1-\frac{SS_{sum}}{SS_{tot}}\frac{M-1}{M-P-1} \quad (7)$$

- Euclidean distances between the cluster prototype series – fig 3. Calculated Euclidean distances can be regarded as other useful statistics about the clustering quality and they can be further processed to find the average distance, maximal distance, etc.

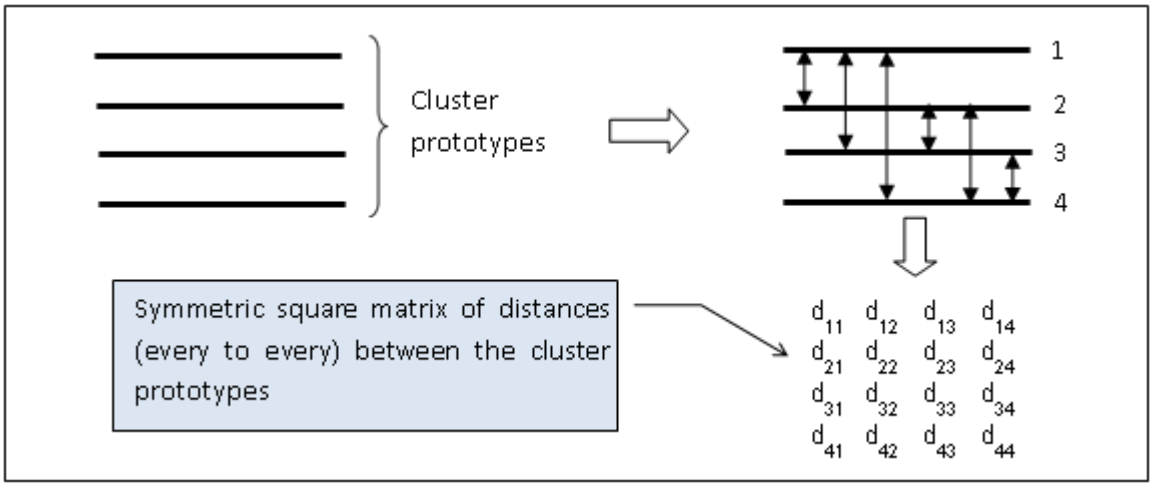


Fig. 3 Euclidean distances between cluster prototypes

- Euclidean distances can be calculated also for every cluster – fig.4. In this case the distances from every time series in the cluster to the center of the cluster are considered. The center is with the same dimensionality as the time series in the cluster and every value of the center is defined as the average value of the corresponding values of time series classified to that cluster.

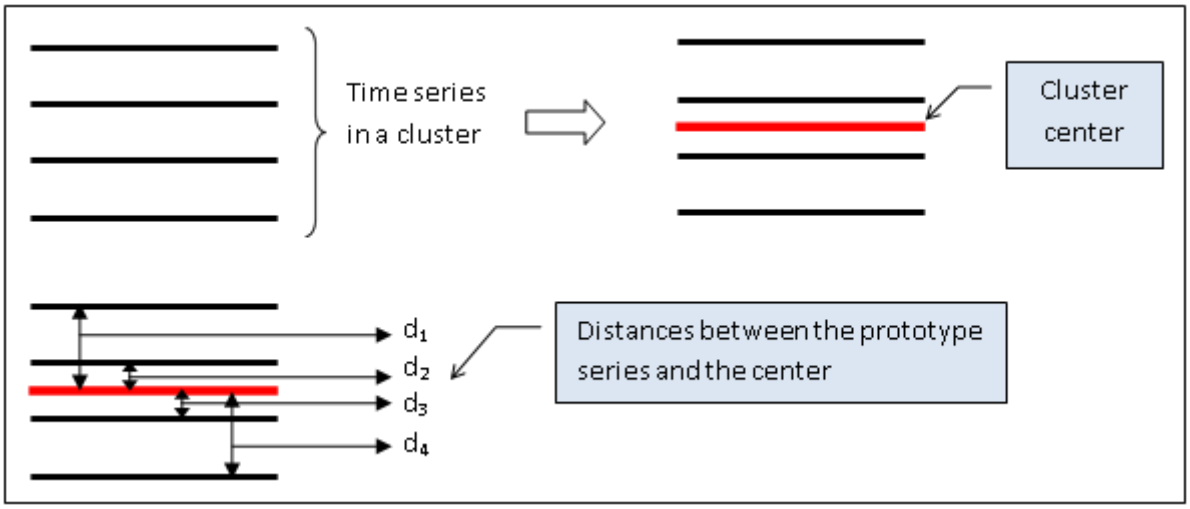


Fig. 4 Euclidean distances between the time series in a cluster and its center

- Average linkage. This statistic is based on calculation of the distances between clusters. The distance between two clusters is defined as the average distance between the series classified to these clusters. The distance between cluster X and cluster Y is calculated in the following way:

$$d_{xy} = \frac{1}{N_x N_y} \sum_{i=1}^{Nx} \sum_{j=1}^{Ny} d(x_i, y_j), \quad x_i \in X, y_j \in Y \tag{8}$$

where N_x and N_y are the number of series in cluster X and Y respectively;
$d(x_i, y_j)$ is the Euclidean distance between series i from cluster X and series j from cluster Y.

Determination of the optimal number of clusters

Finding of the optimal number of clusters is based on an appropriate criterion or criteria that must be defind beforehand. After that one possible way to proceed is to cluster available data for all possible number of clusters (1, 2,…, n) and for each of thesee clustering attempts the chosen criterion is calculated. The best quality value of the criterion determines the best number of clusters.

More often in order to find the optimal number of clusters in practical applications the criterion for optimal clustering is plot in two-dimensional space where X axis represents the number of clusters and Y axis the clustering quality. The sharp drop end point in the graphic, determines the optimal number of clusters. In fig. 5 an example is shown where the criterion is adjusted R^2.

Num Clusters	Adjusted R squared	Error	Time (sec.)
2	0.7888814	0.2111186	20
3	0.8856307	0.1143693	32
4	0.9281010	0.0718990	40
5	0.9351225	0.0648775	60
6	0.9360977	0.0639023	70
7	0.9361842	0.0638158	95
8	0.9647925	0.0352075	109
9	0.9543623	0.0456377	122
10	0.9544117	0.0455883	144
11	0.9758081	0.0241919	154
12	0.9757913	0.0242087	173
13	0.9572335	0.0427665	180
14	0.9572007	0.0427993	194
15	0.9571655	0.0428345	218
16	0.9573212	0.0426788	225
17	0.9572855	0.0427145	260
18	0.9861978	0.0138022	276
19	0.9861863	0.0138137	287
20	0.9861746	0.0138254	305
21	0.9861818	0.0138182	314
22	0.9861700	0.0138300	326
23	0.9861583	0.0138417	362
24	0.9861466	0.0138534	381
25	0.9861356	0.0138644	379
26	0.9861230	0.0138770	429

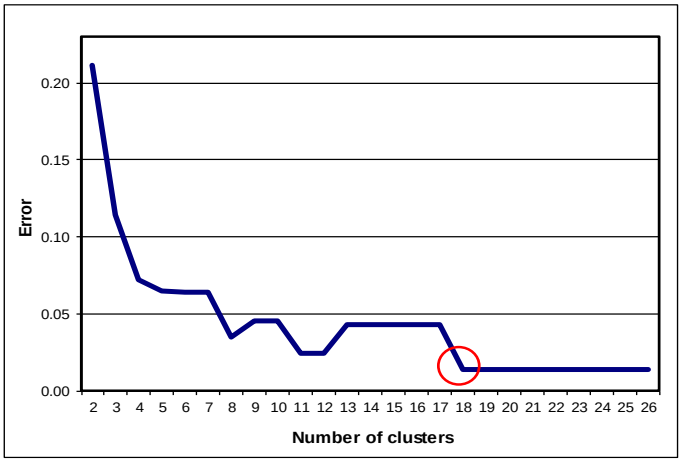


Fig. 5 Clustering of 1200 time series in consequent numbers of clusters and clustering quality calculation based on an adjusted R^2 and Euclidean distance (Error) to find the optimal number of clusters

Another more effective procedure regarding the number of calculation operations is to perform all possible clustering not in consequntive numbers of clusters but over some fixed number k [15][18] as it is shown in fig. 6.

Num Clusters	Adjusted R Squared	Error
5	0.9351225	0.0648775
10	0.9544117	0.0455883
15	0.9571655	0.0428345
20	0.9861746	0.0138254
25	0.9861356	0.0138644
30	0.9868221	0.0131779
35	0.9875898	0.0124102
40	0.9878886	0.0121114

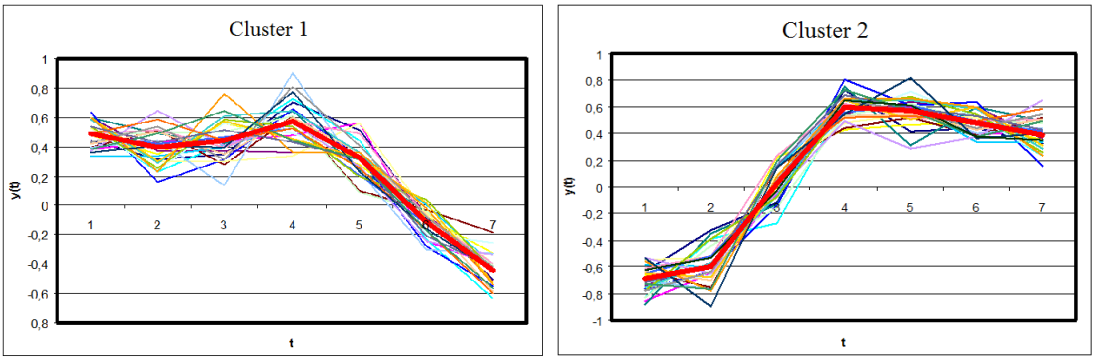
Num Clusters	Adjusted R Squared	Error
15	0.9571655	0.0428345
16	0.9573212	0.0426788
17	0.9572855	0.0427145
18	0.9861978	0.0138022
19	0.9861863	0.0138137
20	0.9861746	0.0138254

Fig.6 Two level of searching of optimal number of clusters - the first level is through 5 clusters and the second level is performed by consequtive number of clusters

In this example k = 5 and clustering is performed in 5, 10, 15, …, N number of clusters and for all of these attempts the clustering quality is plot to find the sharp drop end point i. After that additional clusterings in i+1, i+2, …, i+k-1 clusters are performed in order to find the optimal number of clusters more precisely. This is two levels searching of optimal number of clusters.

Conclusion and future work

In fig. 7 an example of 4 clusters are shown with their time series and their prototypes in bold line. It can be seen that the similar time series are grouped in the same cluster. The current clustering module is used mainly for reduction of the number of series in case of simulation calculations [1] [16]. For example, in case of correlation matrix calculations, where the correlations between all possible couples of series must be calculated, the time needed even for powerful machine is too much [19]. By clustering the number of series is reduced to the number of clusters as when a given series is needed for correlation calculation it is represented by the prototype of the cluster in which the series is classified.



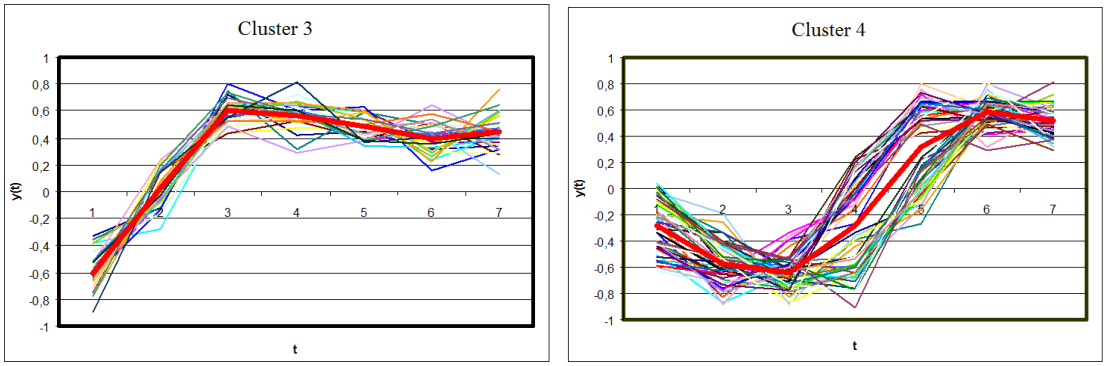


Fig. 7 Examples of clusters, their prototypes illustrated as a bold red line

It is possible alternative clustering approaches to be developed and compared to the developed one using the same or other additional criteria for clustering quality.

References

[1]. Bury, K. Statistical Distributions in engineering. Cambridge University Press, 1999.

[2]. Cameron, C. Trivedi, P. Regression Analysis of Count Data. Cambridge university press, 1998.

[3]. Carpenter, G., S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics and Image Processing, no. 37, Academic Press Professional Inc., 1987, pp. 54-115.

[4]. Chatfield, C. The analysis of time series. An introduction. Fifth edition. Chapman & Hall/CRC, 1996.

[5]. Draper, N. Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics, 1998.

[6]. Forgy, E. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics no. 21, 1965. pp. 768-780.

[7]. Grossberg, S. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics, no. 23, 1976. pp. 121-134.

[8]. Grossberg, S. Adaptive pattern recognition and universal encoding II: Feedback, expectation, olfaction, and illusions. Biological Cybernetics no. 23, 1976. pp. 187-202.

[9]. Grossberg, S. Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, 11, 1987, pp. 23-63

[10]. Hamilton, J. Time Series Analysis. Princeton University Press, ISBN: 0-691-04289-6, 1994.

[11]. Hansen, P., B. Jaumard. Cluster analysis and mathematical programming. Mathematical Programming (79), 1997, pp. 191-215.

[12]. Jain, A., R Dubes. Algorithms for Clustering Data. Prentice-Hall, New Jersey, 1998.

[13]. Kohonen, T. Self-Organizing Maps. Springer, 2001.

[14]. Krishnamoorthy, K. Handbook of statistical distributions with applications. Chapman & Hall, 2006.

[15]. Maiorana, F. Performance improvements of a Kohonen self organizing classification algorithm on sparse data sets. Proceedings of the 10th WSEAS international conference on Mathematical methods, computational techniques and intelligent systems, Corfu, Greece, 2008. pp. 347-352.
[16]. Mun, J. Modeling Risk. Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization Techniques. Wiley, 2006.
[17]. O'Neill, B. Elementary Differential Geometry. Academic Press, 2006.
[18]. Palit, A. K., D. Popovic. Computational intelligence in time series forecasting. Theory and engineering applications. Springer-Verlag, 2005.
[19]. Robert, C., Casella, G. Monte Carlo statistical methods. Springer-Verlag, 1999.
[20]. Vallentin, M. Probability and Statistics Cookbook, 2011.
[21]. Xu, R. Wunsch, C. Clustering. Wiley, 2009.
[22]. Zhang, R., A. Rudnicky. A large scale clustering scheme for kernel K-means. Proceedings of the 16th International Conference on Pattern Recognition, Vol. 4, 2002, pp. 289-292.
[23]. Мандель, И. Кластерный анализ. Финансы и статистика, Москва, 1988.