ISSN 1691-5402
ISBN 978-9984-44-071-2
Environment. Technology. Resources
Proceedings of the 8th International Scientific and Practical Conference. Volume I1

CLUSTERING-BASED BEHAVIOURAL ANALYSIS OF BIOLOGICAL OBJECTS

Arnis Kirshners
Riga Technical University, Institute of Information Technology
Kalku street 1, Riga, LV 1658, Latvia, e-mail: arnis.kirsners@rtu.lv

Abstract. *The article examines the problem of processing short time series for bioinformatics tasks using data mining methods in the field of pharmacology. The experiments were conducted using heart contraction (contraction and relaxation) power data that were obtained in experiments with laboratory animals with the goal of registering the power changes of heart contractions in different stages of experiment in a given period of time. The selected data were treated using data preprocessing technologies. The short time series were compared using various time-point similarity search methods using agglomerative hierarchical clustering, k-means clustering, modified k-means clustering and expectation-maximization clustering algorithms. Based on the clustering result evaluation the most suitable algorithm was chosen and the optimal number of clusters was determined for the least clustering error. The acquired clusters were used for to create cluster prototypes that aggregate the groups of similar heart contraction power objects. The article offers an examination of the errors produced by algorithms and methods as well as a discussion of the obtained clustering results using different evaluation methodologies. It also gives conclusions about the application of data mining methods in solving bioinformatics tasks and outlines further research directions.*

Keywords: *clustering short time series, clustering algorithms, cluster prototypes*.

Introduction

This article proposes a solution for bioinformatics experts for effective short time series processing using data mining methods and algorithms. In bioinformatics and pharmacology in particular the obtained data during large and time-consuming experiments are specific and voluminous. This data can be used to acquire new knowledge using data mining technologies, which can be used as a basis for result interpretation and improve the accuracy of the research results as well as obtain results in a shorter period of time and using less experiments that involve animals (laboratory rats).

The study uses the results acquired in experiments with animals [1, 2] that resulted in heart contraction power measurements. The animals were fed for eight weeks a specially prepared food that contained certain supplements that stimulated action of the heart. Each group of the animals was fed one of the supplement types for a certain period of time. The biological goal of this experiment was the creation of a pharmacological drug that prevents heart failure. Experiments with animal hearts that conform to European Union animal protection ethics rules [2] were carried out to achieve this goal. This resulted in measurements of heart contraction power in a given period of time or in "Isolated heart" [2] experiments. This experiment has three stages: preparation stage when a heart is placed into a special machine for life support and prepared for further experiments; occlusion – closure of blood vessels; reperfusion – opening of blood vessels. The goal of the pharmacological experiment is to determine the percentage of the dead tissue (necrosis) that depends on the types of supplements that were fed to the animals.

The data used in this study describe the occlusion period. The measurements of heart contraction power over a certain period of time can be viewed as short time series because the period of occlusion lasts for 40 minutes in each experiment with each heart. Using pharmacological equipment the intensity of measurement readings was set to 1 minute intervals because what matters is the structural changes over a period of time and not the frequency of the measurements. *ADInstruments* software used in experiments provides this

type of data readings, dismissal of noisy data values and calculation of mean values of heart contraction power in a chosen period of time. As a result a data set vas obtained that included certain number of period for each animal heart used in the experiment. For an obtained subset a set of data mining experiments was carried out to determine object membership to cluster and create prototypes that describe mean values of cluster in each moment of time based on the clusters.

Objective statement

This study is a part of an extensive research that is intended to develop a system to determine heart failure that would help bioinformatics and pharmacology experts to carry out experiments to determine the percentage of necrosis in 'Isolated Heart' experiments based on individual descriptive parameters like weight, food supplements and the parameters that describe the blood plasma. The development of this system would increase the speed of experiments while decreasing their costs and the number of animals used in them. This study is set to prove that application of data mining methods and algorithms is useful in solution of bioinformatics and pharmaceutical tasks. The goal of the study is to find groups of similar objects in the heart contraction data and patterns in them using data mining methods and algorithms so that these groups can be used in later experiments enabling classification of individuals based on their descriptive parameters. The following tasks have to be solved to achieve the goal: carry out the selection of equal numbers of periods for each individual; choose the most appropriate data preprocessing approach to the problem to be solved; find the most appropriate clustering approach; determine the most appropriate number of clusters based on the evaluation of the clustering error; create prototypes that describe similar objects in a cluster.

A series of researches involving choosing of a data normalization approach and using clustering algorithms used in this research has to be conducted to complete the set tasks. The restrictions of the research are related to the processing of short time series. When solving this type of tasks there are several problems associated with choosing methods and algorithms. They are mainly related to the issue of inability of traditional time series data processing methods and algorithms to solve such tasks related to the insufficient length of the time series. Therefore this type of tasks are solved using data mining technologies that provide the acquisition of new knowledge from the analyzed data.

The scientific novelty of the study is based on providing bioinformatics experts with the chance of analyzing heart contraction power data using data mining methods and algorithms by determining groups of similar objects.

The methods

A. Data normalization

This study uses two approaches for data normalization: Z-score normalization by standard deviation (ZSSD) [3] and demand normalization by life curve (LC) [4].

ZSSD is based on the calculations of time series mean value and standard deviation; the normalized value of each object T_i' is calculated using the following formula (1):

$$T_i' = \frac{T_i - \overline{T}_i}{\sigma_{T_i}}, \tag{1}$$

where T_i – value of the time series T in the moment i;

$\overline{T}_i$ – mean value of the time series;

σ_{T_i} – standard deviation.

When ZSSD is used the range of the normalized values includes both the negative and the positive values [3]. LC is calculated using the following formula (2):

$$y_i = \frac{x_i}{\sum_{j=1}^{n} x_j}, \tag{2}$$

where y_i – the normalized attribute value at the time moment i;
x_i – the value of time series x at the time moment i;
x_j – the value of time series y in the period n;
n – the time series duration (number of periods).

This approach is specifically suited for time series normalization. The normalized values belong to the positive range only [4, 5].

B. Clustering algorithms

The data set analyzed in this study does not contain any information about classes and their connection with the data. Therefore there is the necessity to apply data analysis, i.e. to find knowledge in the data that would help to determine groups of similar objects in the data and describe the relationships in these groups. The aggregation of the objects into clusters is implemented in a way that allows grouping the time series with similar structure based on one of the distance metrics.

K-means algorithm

The data set that consists of n objects is split based on predefined parameter, which determines the number of clusters in the data set, into k clusters where the similarity among objects of one cluster is larger than that of objects from different clusters. Each cluster has its own centroid or center of gravity that is calculated using the arithmetic average of the objects that belong to the same cluster using each attribute. In each iteration the algorithm assigns each object to the nearest centroid obtaining a cluster. Then the centroid is recalculated based on objects belonging to the cluster and the algorithm calculates objects' membership to the clusters. This process continues until centroids stop changing their memberships. To assign an object to the nearest centroid a distance metric *Euclidean distance* [6] is used. The only imperfection of the *k-means* algorithm is its dependency on the initial division because the centroids are assigned randomly and the initial position of clusters can be less than optimal. To avoid imprecision the given algorithm is run several times and the results are compared to each other to choose the best result. The result evaluation is carried out using sum of squares error that is calculated using the following formula (3):

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} d^2(x, m_i), \tag{3}$$

where x – given multi-dimensional object;
m_i – multi-dimensional centroid of cluster c_i;
$d(\cdot)$ – distance function;
k – number of clusters.

The results can be interpreted when the best result of the algorithm is determined and the sum of squares is calculated [6, 7].

Modified k-means algorithm

At the beginning of the algorithm the range of the clusters has to be determined to make the process of clustering more effective and less time consuming. The optimal number of clusters has to be found in the range from 2 to maximum number of clusters. This maximum number should be large enough to precisely cluster the data set but also small enough to avoid the

influence of noise on the results. Therefore the maximum number of clusters $Max_{quantityofclusters}$ [5, 7] is calculated according to the theoretical assumption that the maximum number of clusters equals the square root of the number of records in the data set (4):

$$\mathrm{Max}_{\mathbf{quantity\ of\ clusters}} = \sqrt{\mathrm{n}} \tag{4}$$

The calculation of the maximum number of clusters improves the speed of obtaining the desired results by saving time and calculation resources. Then the *modified k-means* algorithm works like *k-means* algorithm until it reaches the calculation of sum of squares error. The algorithm determines the number of clusters that produces the least error that is calculated by dividing the sum of distances *d* between each object and centroid (see Table 1) by the number c_n of records in the cluster (see Formula (5)). It produces the mean absolute error AE_i for each cluster.

Table 1.

Object distances to the centroids used for calculating mean clustering absolute error

Number of object	*Periods*				*Distance measure*
c	***T1***	***T2***	***…***	***T12***	***d***
c_1					d_1
c_2					d_2
…					…
c_n					d_n

Then the sum of mean absolute errors AE_n is divided by number of clusters C_n obtaining the mean absolute error *MeanAE* [5, 10] of clustering using the formula (6):

$$AE_i = \frac{d_1 + d_2 + \cdots + d_n}{c_n}, \tag{5}$$

where d_1, d_2, d_n – the distances from the corresponding record to the centroid;
c_n – the number of records in a cluster;
AE – the mean absolute error in a cluster.

$$MeanAE = \frac{AE_1 + AE_2 + \cdots + AE_n}{C_n}, \tag{6}$$

where C_n – the number of clusters;
MeanAE – the mean clustering absolute error.

This approach provides the analysis of distances between each object of the data set and centroids in each cluster and provides the calculation of average absolute error of clustering. This error is used for finding the number of clusters that is needed to cluster the data set.

Expectation maximization

Expectation maximization algorithm uses probability measure instead of a firm distance measure and analyzes distribution curves for each dimension where each point belongs to a specific cluster with a certain probability. This approach is called soft clustering, which means that clusters can overlap because they do not have strict borders. The algorithm is intended to use when borders of clusters are fuzzy. *Expectation maximization* algorithm calculates the expected value of a hidden variable for each record and then recalculates these values if they were of the observed variables that use the expected values. The work of the algorithm consists of two steps (E and M). In the *E-step* the algorithm calculates the sum of an expression that includes the expected logarithmic credibility value of the appended data against probability distribution. The *M-step* maximizes the expected logarithmic credibility value according to parameters [8, 9].

Agglomerative hierarchical algorithm

Agglomerative clustering algorithm belongs to the hierarchical clustering algorithms. The main working principle of the algorithm consists of stepwise object merging into groups. At the beginning each object represents a separate cluster. The nearest objects are merged first based on the minimum distance between them and later between farther objects and object groups. The process is run until all objects belong to the same group or cluster that signals the algorithm to stop. To determine the distance between two clusters and compose a distance matrix the "nearest neighbor" principle is used (7):

$$d_{\min}\left(S_l, S_m\right) = \min_{\substack{x_i \in S_l \\ x_j \in S_m}} d\left(x_i, x_j\right) \tag{7}$$

where $d(\cdot)$ – distance function;

S_l, S_m – two separate clusters;

x_i, x_j– two objects of different clusters.

Agglomerative clustering algorithm is often used to construct taxonomies because it uses a hierarchical approach. The result is a dendrogram that graphically demonstrates the succession of object merges in each step of the algorithm iteration [6, 7].

C. Prototypes of clusters

Prototypes of clusters are created based on clustering results. Each cluster is represented by its own prototype that characterizes mean values of the cluster in each period of time.

Results and discussion

The following software was used in experiments: *LabChart 7 View*, *Microsoft Excel 2010*, *Weka 3.6.3*, *OrangeCanvas 2.0* un *Statistica 8.0*.

Data and experiments

This study uses a data set that consists of 92 records; each object was described by 38 time periods of time that were read from the pharmacological apparatus using *ADInstruments* and *LabChart 7 View* software. The readings were made every 60 seconds over a period of 40 minutes; the first and the last readings were excluded. The mean values of heart contraction power in each period of time were expressed using mercury height (mmHg). The obtained short time series values were normalized using various normalization approaches to determine the most appropriate for the used data. The result of the readings were two data sets with different normalization results. Both data sets were clustered using each clustering algorithm. The obtained results were analyzed which provided the best clustering approach for the given task. The clustering results served as a basis to create prototype for each cluster.

Data preprocessing

Heart contraction data of 40 minutes were sampled for each animal using *LabChart 7 View*. Heart contraction power measurement graphs are shown in Fig. 1; the vertical dashed lines show the beginning and the end of the measurements. The beginning phase indicates occlusion and the end phase shows reperfusion. The tool provided data readings in the given period of time and exclusion of noise using a built-in mathematical analyzer. The prepared data were normalized using ZSSD and LC approach providing two data sets for further experiments.

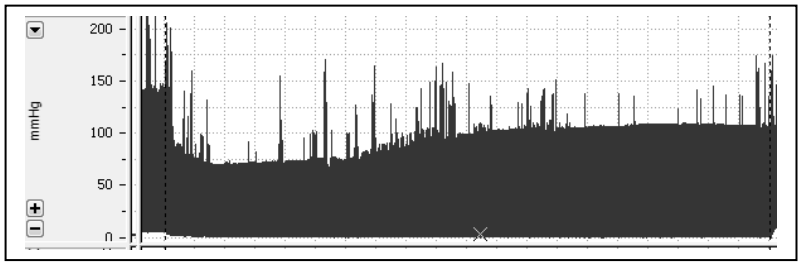

Fig. 1. **Heart contraction power (mmHg) measurements in a given period of time**

Data clustering

The experiments with each data set were repeated 20 times varying centroids based on the randomness principle. The results were evaluated using the distances between objects and centroids and distances between clusters. The sums of squares were calculated for different numbers of clusters as shown in Table 2.

Table 2.

Sums of squares errors for *k-means* algorithm

	Number of clusters							
Approach	2	3	4	5	6	7	8	9
ZSSD	141,67	138,42	128,36	125,89	122,64	120,56	116,34	113,92
LC	115,73	112,99	101,81	96,36	92,54	89,44	87,39	85,79

When *modified k-means* algorithm was used the absolute errors were calculated using formula (5) and the mean absolute errors of each cluster were calculated using formula (6); see Table 3 for results. The number of clusters was determined using evaluation of these results. The optimal number of clusters for this data set is 5 (see Fig. 3), which is represented by the minimum mean absolute error. The normalization approach was also chosen based on these errors.

Table 3.

Mean absolute errors using *modified k-means* algorithm

	Number of clusters							
Approach	2	3	4	5	6	7	8	9
ZSSD	1,233	1,192	1,158	1,150	1,128	1,095	1,093	0,946
LC	1,026	1,017	1,026	**0,59**	0,644	0,91	0,824	0,615

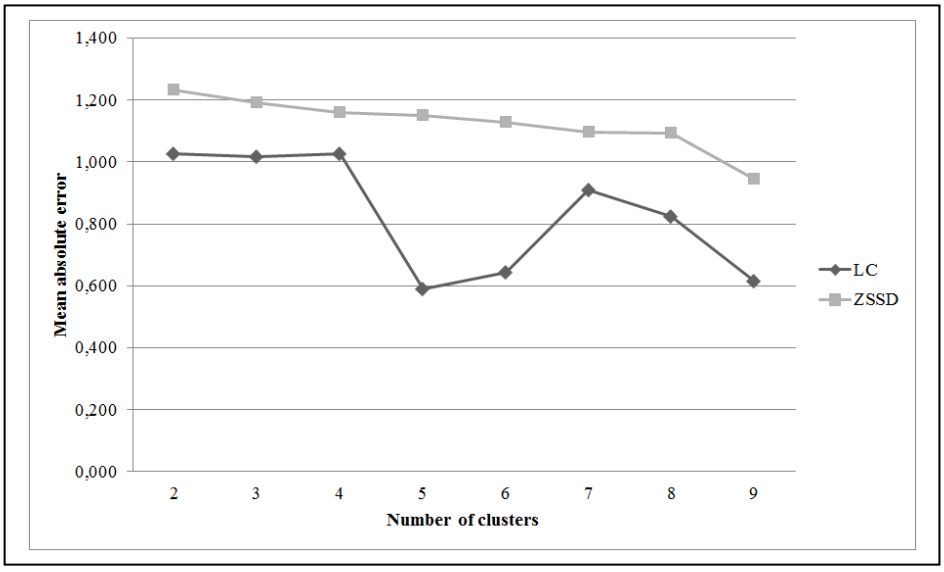

Fig. 2. **The choice of normalization approach and number of clusters using mean absolute error**

Expectation maximization algorithm found two clusters for each approach in the clustering process. The obtained logarithmic probabilities that evaluate the performance of the algorithm are shown in Table 4. The agglomerative hierarchical algorithm found three clusters for each approach using *Ward's linkage*. This Ward's linkage is based on increase in sum of squares of object distances between clusters and their centroids during merges. In the first case

(see Fig. 4, left) when the ZSSD was used the stopping criterion is distance in the range from 4.2 to 5 because this value is the maximum in the analyzed range of merges. There are three clusters expressed that describe this division.

Table 4.

Logarithmic probabilities for *expectation maximization* algorithm

	Number of clusters							
Approach	2	3	4	5	6	7	8	9
ZSSD	-49,64	-48,64	-47,93	-47,39	-46,36	-45,47	-45,25	-44,13
LC	144,49	145,40	148,42	149,06	150,06	150,32	151,13	150,87

In the approach that uses demand normalization using life curve (see Fig. 4, right) the stopping criterion is reached in the distance range between 3.8 and 5.8. There are two expressed clusters for this division.

Based on the clustering results the best results were shown by *modified k-means* algorithm using mean absolute error estimation. The results of this algorithm were used as basis for prototype construction (see Fig. **Error! Reference source not found.**). The obtained prototypes will be used in further research that will be based on merging heart's descriptive parameters (like weight of the heart, blood count etc.) with the clustering results.

The merging of the results is necessary to develop a system that would diagnose heart failure in the future.

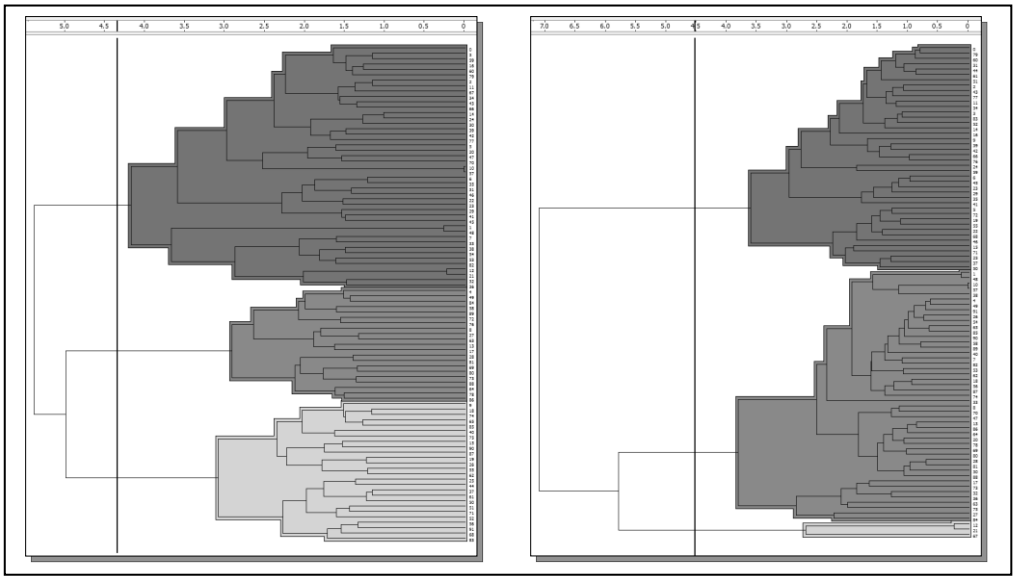

Fig. 3. **Clustering of the data set using hierarchical agglomerative algorithm using *Z-score normalization by standard deviation* (left) and *demand normalization by life curve* (right)**

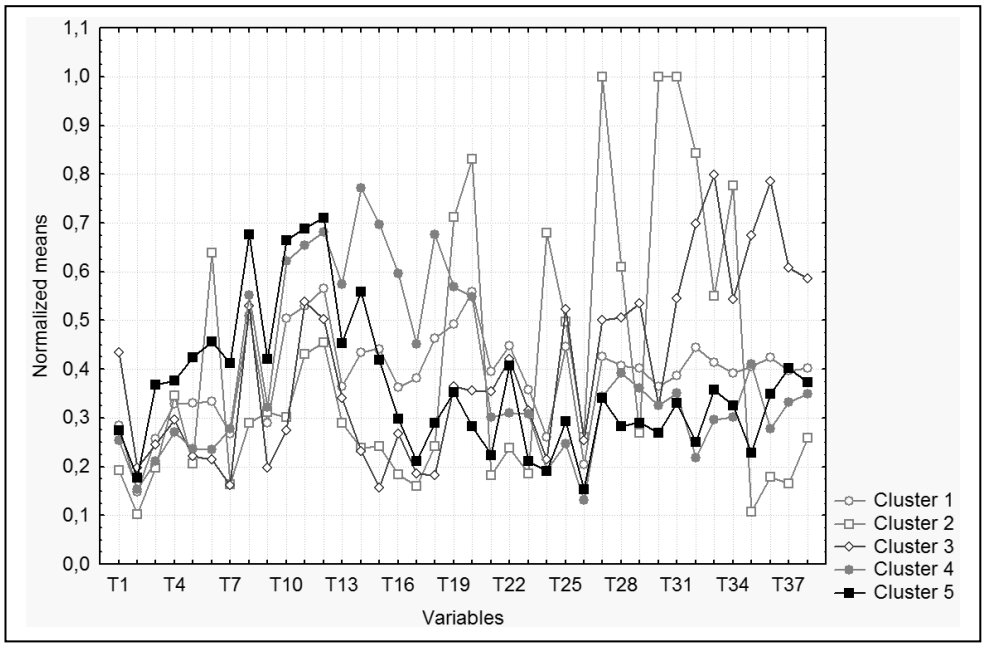

Fig. 4. **The prototypes obtained using *modified k-means* algorithm and demand normalization by life curve for five clusters**

Conclusions

The results of this study show that data mining methods and algorithms are useful when solving bioinformatics and pharmacology tasks that involve short time series.

The experiments show that the best results in normalizing heart contraction power data are acquired by normalization by life curve approach that can be explained by the fact that Z-score normalization by standard deviation normalizes values in positive and negative value ranges whereas normalization by life curve uses only values in the positive range.

Clustering algorithms showed different results. Expectation maximization algorithm showed a unitary trend when clustering both data sets: when the number of clusters grew the logarithmic probability, which is used when calculating the error of the algorithm, increases. The obtained results show that this algorithm is not suited for this type of tasks.

K-means algorithm also showed a unified trend in error evaluation. Two evaluation methods were used in this study: *distance to centroids* where the maximum (9) number of clusters was chosen in both data sets; and *distance between clusters* where the minimum (2) number of clusters was chosen. The results are not clearly interpretable therefore this algorithm is not suited for this type of tasks.

Agglomerative hierarchical algorithm showed similar trends assigning the same number of clusters using both approaches. Visual analysis of the dendrograms shows that the results of clustering are informative and visually interpretable. The analysis also shows that Z-score normalization by standard deviation has more advantages because the cluster distribution among objects is more even.

The results and the mean absolute error show that *modified k-means* algorithm is the most adequate of the analyzed clustering algorithms for this task. This algorithm showed the best results using normalization by life curve judging by mean absolute clustering error evaluation. Therefore the data set used to build prototypes was clustered using this approach and five clusters.

Based on the prototypes found in the data for each cluster the further experiments will revolve around determining connections between clustering results and the parameters describing animals like weight, blood count etc. There are plans to build a system that would help pharmacology experts to conduct experiments in heart failure diagnostics. The use of such

system would decrease the number of needed experiments, increase the efficiency of the results and save human and time resources.

Acknowledgements

The author thanks the lead researcher of the pharmaceutical pharmacology laboratory Dr. phrm. Edgars Liepins and researcher Janis Kuks of Latvian Institute of Organic Synthesis for the expressed interest and the provided research data on heart contraction power that were used for experiments in this study.
This work has been supported by the European Social Fund within the project «Support for the implementation of doctoral studies at Riga Technical University».

References

1. Liepinsh E., Vilskersts R., Skapare E., Svalbe B., Kuka J., Cirule H. et al. Mildronate decreases carnitine availability and up-regulates glucose uptake and related gene expression in the mouse heart. Life Sci, 2008, 83: 613–619.
2. Liepinsh E., Vilskersts R., Zvejniece L., Svalbe B., Skapare E., Kuka J. et al. Protective effects of mildronate in an experimental model of type 2 diabetes in Goto-Kakizaki rats. British Journal of Pharmacology, 2009, **157**: 1549–1556.
3. Kirshners A., Sukov A. Rule induction for forecasting transition points in product life cycle data. Scientific Proceedings of Riga Technical University, Information Technology and Management Science, Issue 5, Vol.36, RTU, Riga, 2008, p. 170-177.
4. Kirshners A., Parshutin S., Borisov A. Combining clustering and a decision tree classifier in a forecasting task. Automatic Control and Computer Sciences, Vol.44, N3, 2010, p. 124-132.
5. Thomassey S., Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems, Vol.42, Issue 1, 2006, p. 408-421.
6. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining. – Boston: Addison-Wesley, 2006, 769 p.
7. Witten I.H., Frank E. Data mining: Practical Machine Learning Tools and Techniques, 2nd edition. – Amsterdam etc.: Morgan Kaufman, 2005, 525 p.
8. Dellaert F. The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology, Technical Report number GIT-GVU-02-20, Feb., 2002.
9. McLachlan, G. and Krishnan, T. The EM Algorithm and Extensions. Wiley series in probability and statistics. John Wiley & Sons, 1997.
10. Montgomery D. C., Jennings C. L., Kulachi M. Introduction to Time Series Analysis and Forecasting. Wiley-interscience, 2008, 472 p.

Anotācija. *Darbā tika aplūkota problēma, kas saistīta ar īsu laika rindu apstrādi, risinot bioinformātikas uzdevumu farmakoloģijas nozarē, pielietojot datu ieguves metodes. Eksperimentiem tika izmantoti dati par sirds kontrakcijas (saraušanās un atslābšanas) spēku, kas iegūti eksperimentos ar laboratorijas dzīvniekiem, ar mērķi reģistrēt sirds kontrakcijas spēka izmaiņas dažādās eksperimenta stadijās noteiktā laika periodā. Atlasītie dati tika apstrādāti, izmantojot datu pirmapstrādes tehnoloģijas. Īsu laika rindu salīdzināšanai tika izmantotas dažādas punktveida līdzības meklēšanas metodes, izmantojot aglomeratīvo hierarhisko, k-vidējo, modificētu k-vidējo un maksimālās līdzības (EM) algoritmus. Balstoties uz klasterizācijas rezultātu novērtējumu, tika izvēlēts atbilstošākais algoritms un noteikts piemērotākais klasteru skaits, kas nepieciešams šīs datu kopas klasterizācijai, balstoties uz klasterizācijas vidējās absolūtās kļūdas novērtējumu. Pamatojoties uz iegūtajiem klasteriem, tika izveidoti paraugmodeļi klasteros, kuri apvieno līdzīgu sirds kontrakcijas spēka objektu grupas. Darbā tika izvērtētas pielietoto algoritmu un metožu radītās kļūdas. Salīdzināti iegūtie klasterizācijas rezultāti, pielietojot dažādu novērtēšanas metodiku. Izdarīti secinājumi par datu ieguves metožu pielietošanu bioinformātikas uzdevumu risināšanai, iezīmētas nākotnes vīzijas turpmākajiem pētījumiem.*