SIMPLE FREE-SHARE PACKAGE FOR VISUAL ANALYSIS OF MULTIDIMENSIONAL DATA SETS4

Alina Nasibullina, Mikhail Alexandrov, Alexander Kovaldji

Abstract: *We describe a simple free share package to study a structure of multidimensional object sets and to classify objects on these structures. The package includes 4 methods: visualization of inter-object distance distribution, object presentations in 2D subspaces of parameters, 2D subspaces of principal components (factors), 2 different alternatives. The first method is a new one: it allows to evaluate the possible number of clusters (compact groups) in data. The second and the third methods are well-known: they give possibility to see structures in data. The forth method is the new one as well: it is a convenient way to see in 2D subspace the relation to alternative multidimensional objects. A user can mark some objects as the 'good' or 'bad' ones. Their position on structures allows to implement a visual binary classification using the principal of neighborhood. All methods are managed in the interactive mode. The functionality of the system is shown on analysis of a real data set reflecting business activity of Russian companies of mobile communication.*

ACM Classification Keywords*: I.2 Artificial Intelligence*

Keywords*: visual cluster analysis, visual binary classification, software*

Introduction

Analysis of large data sets of multidimensional objects needs application of automatic data processing. But automatic procedures are practically useful if an expert can interpret results and justify the revealed regularities. Visualization is one of the effective ways for such interpretation.

In this paper we describe the developed software system that includes 4 methods of visual analysis. The first method uses inter-object distance distribution to predict the possible number of clusters (compact groups) in data. It is a new method. The second and the third methods provide object presentations in 2D subspaces of object parameters and principal components related with these parameters. It is the traditional methods. The forth method is a convenient way to see in a plane the relation to alternative multidimensional objects. It is a new method. A user can mark some objects as the 'good' or 'bad' ones. Their position on structures allows to implement visual binary classification using the principal of neighborhood. All methods are managed in the interactive mode.

By the moment we could not find simple programs for multidimensional data visualization, which would include convenient manipulations with subspaces and marked objects. We mean here such power tools for Data Mining as R, Rapid Miner, Weka [R, http], [Rapid Miner, http], [Weka, http]. From the other hand we would like to make accessible the new methods of data visualization mentioned above. These circumstances defined the actuality of the completed work.

The paper is structured by the following way. In the section 2 we describe traditional and new algorithms. In the section 3 we present the software package and show the result of experiments on real data set. Section 4 includes conclusions.

[4] Work done under partial support of the British Petroleum grant (RPANEPA - 2013).

Algorithms and software

Preprocessing

The first operation an expert should use is normalization and clearing. Speaking clearing we mean determination of outliers. We use the interval, statistical and logarithmic normalization for positive numbers:

$$x_{i\,new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad x_{i\,new} = \frac{x_i - M(x)}{\sqrt{D(x)}} \qquad x_{i\,new} = log_a(x_i + 1)$$

Here: x_i is the parameter value in the i-th observation, $M(x)$ is the expectation of parameter, $D(x)$ is the variance of parameter.

We find oulliers using Chebychev's inequality [Cramer, 1999]. Parameter *k* is selected by the user:

$$P\{|x - M(x)| < k\sigma\} \geq 1 - {}^{1}/_{k^2}$$

Traditional methods of visualization

If the source data contains more than three parameters, data representation in the space of four or more coordinates in the usual sense is impossible. There are special ways to visual data in multidimensional space, for example: parallel coordinates, Chernoff faces and flap position.

This software package uses the traditional way of visualizing multidimensional objects in 2D space of any two selected parameters. The only difficulty is how to choose these two parameters for expert could see groups of objects, because cluster structure will not be observed in each sub-space of the parameters.

Another way to visualize multidimensional data is a reduction of dimension and data presentation in 2D space of principal components. Using the principal components allows to reduce the dimension of the original data and to distort the geometric characteristics of clouds of points in the parameter space as little as it is possible.

To calculate new coordinates in the space of principal components it is necessary to find the eigenvectors and the eigenvalues of the covariance matrix of object parameters. Such a matrix is calculated as

$$M = A^T A$$

Here: *M* is a covariance matrix (*p,p*), *A* is matrix objects/attributes (*n,p*), *n* is number of objects, *p* is number of parameters. Parameter values should have the dimensionless form, i.e. it should be normalized in accordance with the problem to be solved. Traditionally the correlation matrix is used, where all the parameters have zero mean and unit variance. Matrix A is supposed to have normalized parameters. With this the covariance matrix is the correlation matrix.

Simultaneously with visualization of multidimensional data in 2D spaces the program uses the method of so-called "labeled atoms". Some specially selected points are marked on the plane.

Here the certain amount of representatives of "good" and "bad" classes is supposed to be known. Belonging of some objects within of selected structures to these classes can be determined on the basis of their neighborhood. For example, objects that are closer to "good labeled atoms" can be considered as the "good" ones.

New methods of visualization

In addition to traditional methods the program realizes 2 new methods. The first one is building diagram of inter-object distance distribution. It allows to determine the lower limit of the number of clusters

[Nasibullina, 2014]. The method was proposed 10 years ago by A.Kovaldji but by the moment it was not published.

Consider the set of N objects in a multidimensional space. Description of the algorithm includes the following steps:

1) Calculation of distances between all objects and determination of scattering (minimum and maximum distance). For *N* objects we have *N* (*N*-1) / 2 distances.

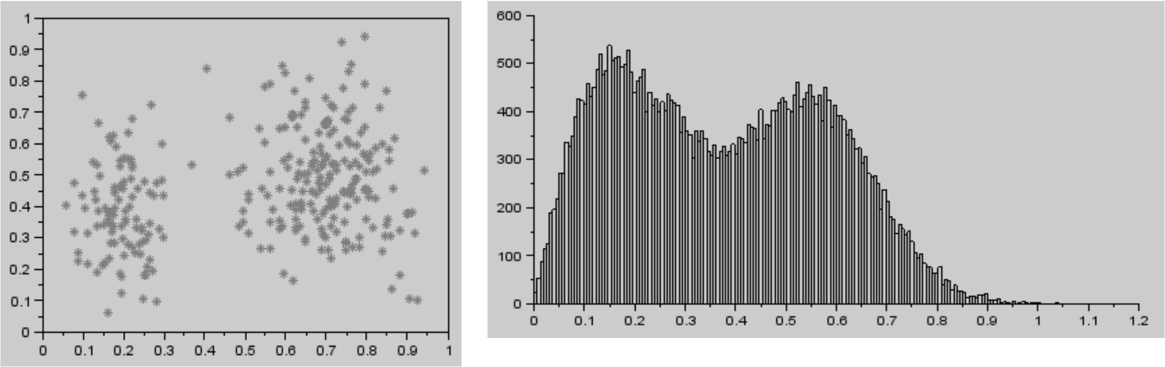
2) Building the histogram with the step equal to the several minimum inter-object distances. There is a distance on the X-axis and frequency of interval occurrence on the Y-axis. Thus, we obtain a function of the distance distribution.

3) If it is necessary, a diagram is smoothed.

Remote objects or groups of objects give far removed peaks from the origin. On the contrary, close objects or groups of objects give peaks that are near to the origin. The number of local maxima is equal to *K*=*n* (*n*+1) / 2, where *n* is the number of clusters (Figure 1). Such a formula is valid if scatter of objects in clusters is different and intercluster distances are different too. Otherwise some peaks will coincide. If scatter in clusters is more or less equal and all intercluster distances are different then number of peaks is equal *K*= *n* (*n*-1) / 2 + 1. Just figure 1 illustrates last formula.

Smoothing is necessary in order to make it easier to reveal essential peaks and to eliminate minor peaks. There are various methods of smoothing, but here we use the simple moving average.

Diagram of inter-object distances has 2 restrictions: a) intergroup distances are supposed to be greater than the inter-object ones; b) the method is suitable for clusters in the form of clouds, but no chains! Thus, this method is well suited to clearly separable clusters in the form of approximately round clouds.



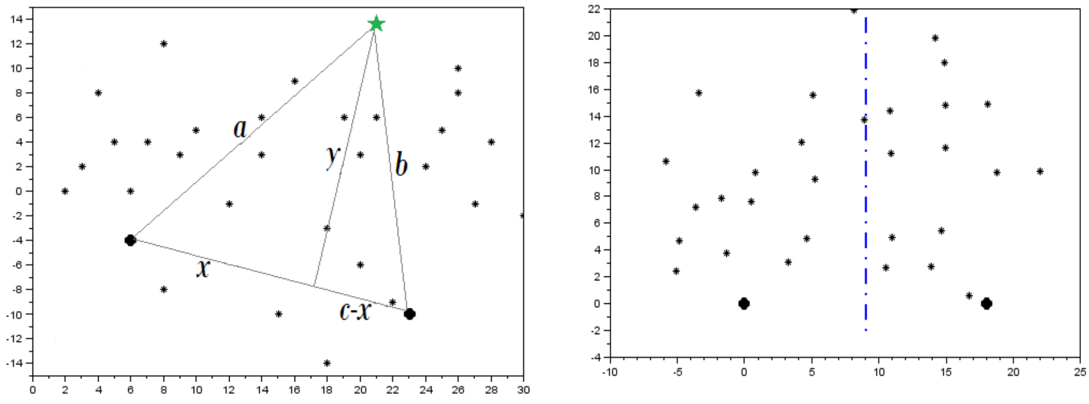
(a) (b)

Fig.1. Example on synthetic data: (a) Sets of points on the plane, (b) Diagram of inter-object distance distribution

The second new method is visualization of objects on the diagram of two alternatives. It allows to determine the proximity of objects to two points, which are known to have alternative properties. Speaking alternatives we mean an expert himself/herself assigns such objects. The principal idea is to represent multidimensional objects in 2D space. New coordinates can be calculated by the following formula:

$$x = \frac{a^2 - b^2 + c^2}{2c} \qquad y = \sqrt{a^2 - \frac{(a^2 - b^2 + c^2)^2}{4c^2}}$$

Here: x, y are new coordinates in 2D space; a, b are distance from the object to the first and to the second alternatives in multidimensional space; c is a distance between two alternatives in multidimensional space (Figure 2). The perpendicular at the center on Figure 2(b) gives possibility to say what alternative object is closer to a given object.



(a) (b)

Fig.2. (a) Points in the original space (2D as an example), (b) Points on the diagram of two alternatives

Software and example of application

Short package description

The program was developed on the platform SciLab [SciLab, http], [Alekseev, 2008]. It contains 4 modules: the computing one, man-machine interface (including visualization), and interface with data base (including preprocessing). These modules are schematically presented in Figure 3. HM-interface and BD-interface mean here human-machine interface and interface with data base respectively.

The main procedures of the computing module are:

- Calculation of distances between all objects and identification of scattering (minimum and maximum distance)
- Building a histogram with a step, which is equal to the several minimum inter-object distances (in the next version of package the step will be given as a part of maximum inter-object distance)
- Smoothing the histogram
- Calculation of geometric coordinates to represent objects in parameter sub-space for their subsequent visualization in the form of points in the selected window
- Calculation of eigenvectors and eigenvalues of the correlation matrix related to parameters of the original data
- Calculation of the principal components (PC)
- Calculation of geometric coordinates to represent objects in principal component sub-space for their subsequent visualization in the form of points in the selected window
- Calculation of the new coordinates of the objects relatively to two alternatives and their subsequent visualization in the form of points in the diagram "the worst – the best".

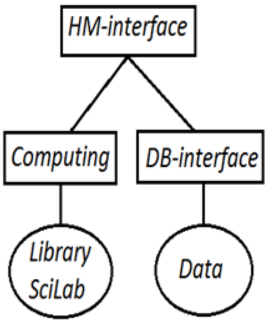
Interface allows to point source of information, to complete preprocessing, assign subspaces and labeled objects, select method of visualization. We demonstrate interface in the process of analysis of data related to business activity of Russian companies of mobile communication.

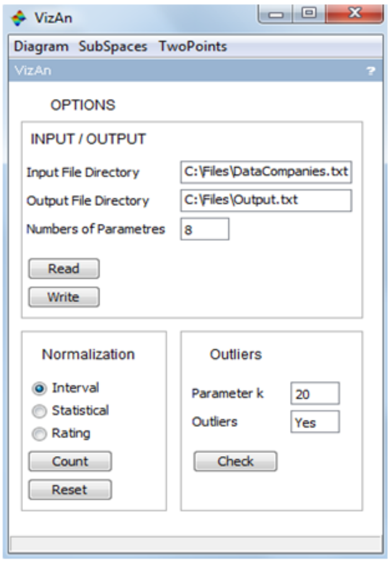
Source data and preprocessing

As real data to test the program we have taken data by 89 mobile companies of SPARK database. SPARC system is the largest database of companies in Russia, Ukraine and Kazakhstan. In this paper we used the following indicators of mobile companies: 1) current liquidity ratio, 2) quick liquidity ratio, 3) cash liquidity ratio, 4) equity ratio, 5) gearing ratio, 6) ratio of maneuverability, 7) average number of employees, 8) normalized net assets. The part of data are presented in the Table 1. We have conducted the interval data normalization and search for outliers using the parameter K=20 (it was recommended us). Here we used the interface shown in Figure 4.

Table 1. Data of Russian mobile companies

	Current liquidity ratio	Quick liquidity ratio	Cash liquidity ratio	Equity ratio	Gearing ratio	Ratio of maneuverability	Average number of employees	Normalized net assets
	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010	2010, RUR
1	0,621	0,617	0,490	0,351	4,391	0,879	149,4	21,520
2	0,563	0,464	0,062	0,055	23,194	-7,518	162,6	13,035
3	0,747	0,686	0,379	0,439	2,056	0,231	39,5	31,974
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
89	0,217	0,211	0,110	0,341	12,165	0,701	154,7	-1,880





Input File Directory C:\Files\DataCompanies.txt
Output File Directory C:\Files\Output.txt
Numbers of Parametres 8

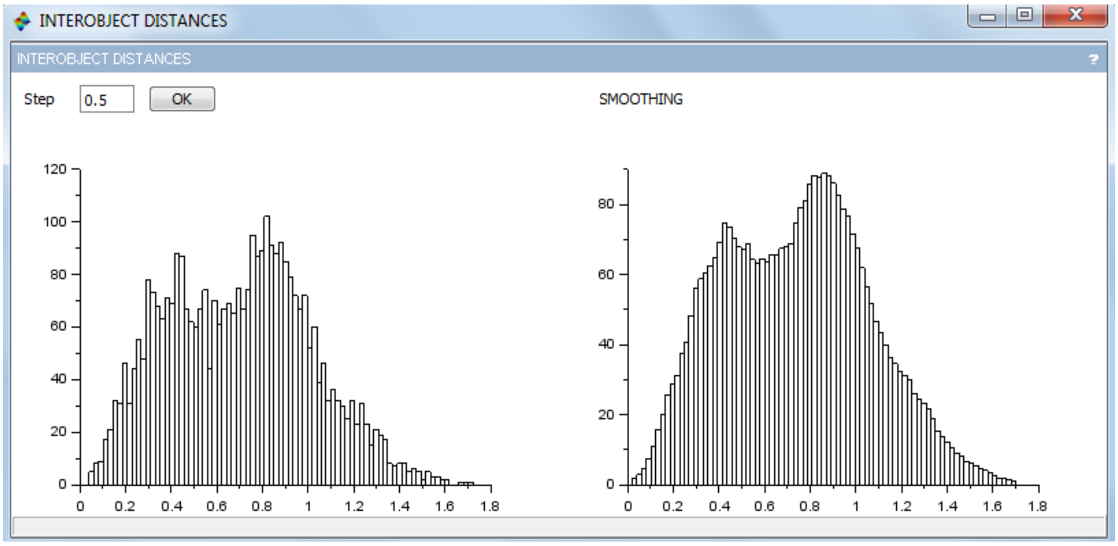
● Interval
○ Statistical
○ Rating

Parameter k 20
Outliers Yes

Fig.3. Structure of the system *Fig.4. Window for preprocessing*

Application of diagram of inter-object distance distribution

Generally a user tries to evaluate the approximate number of possible compact groups in data. For this the user builds the diagram of inter-object distance distribution The user can vary the step of the chart and complete smoothing with simple moving average method. The results are presented on Figure 5. We can distinguish two local maxima, so the lower bound of the number of clusters is two and we expect to see 2 compact groups of objects in one of the sub-spaces.



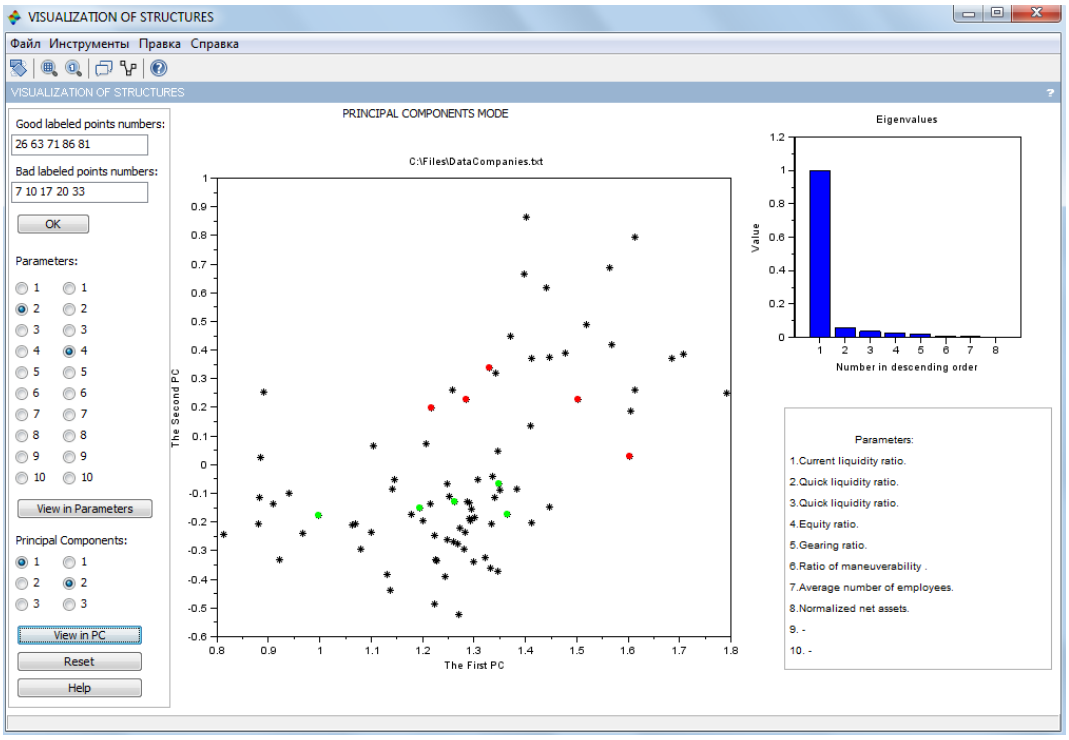
Step 0.5

Fig.5. Window with diagram of inter-object distance distribution

Visualization in the 2D parameter space and in the space of principal components

The package allows to handle objects with no more than 10 parameters. The number of PC does not exceed 3. It is enough for almost all problems being met in practice. As the procedures of choice and the presentation of objects in the space of parameters and in the space of PC are is similar then we consider only the second case.

A user specifies a pair of principal components being interested to him/her using interface on Figure. 6. The diagram of eigen-values at the corner shows that the first component can explain variations of object parameters almost completely. With the second and the third ones such a presentation will be more than enough. On Figure 6 one can see object distribution in the space of the first and the second PCs. The user can assign “good objects” and “bad objects” using fields at the upper left corner. They are presented as green and red points in the same window. The number of good and bad objects should not exceed 5 respectively. It is clearly seen that the labeled points spaced apart from each other. It allows to see the set of good and bad points on the basis of neighborhood.



VISUALIZATION OF STRUCTURES

Good labeled points numbers: 26 63 71 86 81

Bad labeled points numbers: 7 10 17 20 33

Parameters:

○ 1 ○ 1
● 2 ○ 2
○ 3 ○ 3
○ 4 ● 4
○ 5 ○ 5
○ 6 ○ 6
○ 7 ○ 7
○ 8 ○ 8
○ 9 ○ 9
○ 10 ○ 10

Principal Components:

● 1 ○ 1
○ 2 ● 2
○ 3 ○ 3

PRINCIPAL COMPONENTS MODE

C:\Files\DataCompanies.txt

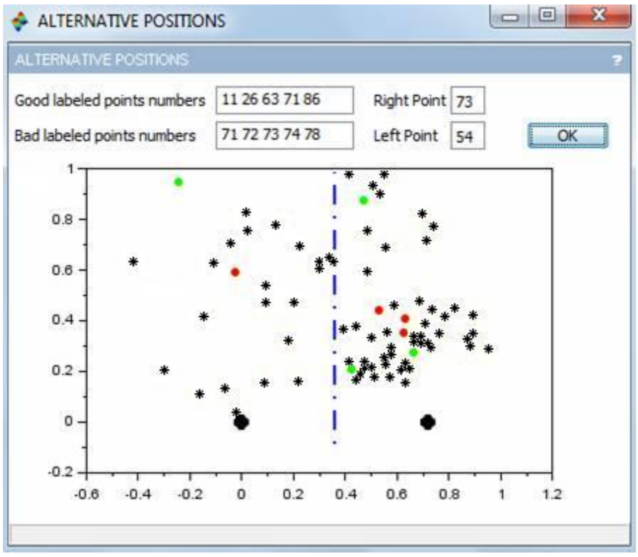
Parameters:

1.Current liquidity ratio.
2.Quick liquidity ratio.
3.Quick liquidity ratio.
4.Equity ratio.
5.Gearing ratio.
6.Ratio of maneuverability .
7.Average number of employees.
8.Normalized net assets.
9. -
10. -

Fig.6. Visualization window in the parameter space and in the space of principal components

Visualization of objects on the diagram of two alternatives

To use such a presentation one should select two objects. One of them is considered as a good one, and the other object as the bad one. Here the user can assign "good objects" and "bad objects" as he/she did it early. The number of good and bad objects should not exceed 5 respectively. The results are presented on Figure 7. Objects located right of the central line are closer to the good object. And objects located left of the central line are closer to the bad object.



ALTERNATIVE POSITIONS

Good labeled points numbers 11 26 63 71 86

Bad labeled points numbers 71 72 73 74 78

Right Point 73

Left Point 54

Fig.7. Companies in the diagram of two alternatives

One can see that the majority of labeled objects (the red and the green ones) are located right of the line. It means that both left and right alternative objects are not too contradictive.

Conclusion

This article presents possibilities of software package that allows to reveal structures in data set and select objects on these structures. We shortly described algorithms included to the package and demonstrated their work on a real data of Russian mobile phone companies.

Identifying structure in data sets is one of the problems to be resolved in Data Mining [Manning, 2008] and just inside this area Visual Data Mining occupies an essential niche [Ankerst, 2000]. So, the developed package can be considered as a tool of Data Mining.

Autumn-winter this year we suppose to release a second version of the package, which will include:

1) The algorithms providing more effective visual analysis. For this we plan to propose measures / metrics that allow to use the diagram of inter-object distance distribution for chain-like structures. Also we hope to develop algorithms to select the most interesting projections.
2) Typical algorithms of cluster analysis providing an automatic search for groups of different structure [Alexandrov, 2007]. We will use here various modifications of K-means, nearest neighbors and MajorClust methods. In these methods, we intend to use traditional measures: Euclidean, cosine linear, binary. We will use also untraditional measures, taking into account probabalistic and diffuse nature of the data parameters.

Bibliography

[Alekseev, 2008] E. Alekseev, O. Chesnokov, E. Rudchenko. Scilab. Solving engineering and mathematical problems. BINOM. 2008 (rus).

[Alexandrov, 2007] M. Alexandrov, P. Makagonov. Introduction to Technique of Clustering. In: Proc. of 3-rd Intern. Summer School on Computational Biology, Masaryk Univ. of Brno, Czech Rep., 2007, pp. 55-80.

[Ankerst , 2000] V.M. Ankerst. Visual Data Mining, Master thesis, 2000.

[Cramer, 1999] H.Cramer. Mathematical Methods of Statistics, Princeton University Press, 1999.

[Manning, 2009] C. Manning, P. Raghavan, H. Schutze. An introduction to information retrieval. Online edition. Cambridge UK, 2009.

[Nasibullina, 2014] A. Nasibullina. Evaluation of cluster number on the basis of diagram of inter-object distance distribution (this proceedings).

[RapidMiner, http] electronic resource, http://rapid-i.com .

[R-studio, http] electronic resource, http://www.rstudio.com, http://www.r-project.org.

[SciLab, http] electronic resource, http://www.scilab.org.

[Weka, http] electronic resource, http://www.cs.waikato.ac.nz/ml/weka/.

Authors' Information

Alina Nasibullina – M.Sc.Student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia e-mail: *nasibullina.alinka@yandex.ru*

Major Fields of Scientific Research: Visual Data Mining

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: *MAlexandrov@mail.ru*

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Alexander Kovaldji - Deputy director for science, Director of multidisciplinary evening school, Moscow mathematical lyceum “Vtoraya Shkola”; St. Fotiyeva 18, Moscow, 119333, Russia; e-mail: *koval-dji@yandex.ru*

Major Fields of Scientific Research: mathematics, mathematical modeling