SRSTI 50.47.00:20.23.19 https://doi.org/10.58805/kazutb.v.2.19-96

DATAMINING CAPABILITIES FOR CLUSTERING CONCRETE MIX FORMULATIONS

K.M. Akishev[1*],V.I. Karpov[2], L.A. Akisheva[3], A.D.Tulegulov[1]

[1]Kazhach University of Technology and Business,

[2]Moscow State University of Technology and Management named after K.G. Razumovsky,

[3]Nazarbayev Intellectual School

e-mail: Akmail04cx@mail.ru

The article discusses the possibilities of using DataMining technology for clustering concrete mixtures. In practice, it is often necessary to face tasks in which it is necessary to choose the formulations closest in quality characteristics from a large number of formulations of concrete mixtures. The distribution of formulations of concrete mixtures by classes is provided on the basis of specified criteria such as strength, as well as the composition of the ingredients of the concrete mixture. Earlier in work [1], clustering of concrete mix formulations was carried out with the help of the program "Comprehensive quality assessment and classification of multidimensional objects", which made it possible to distribute formulations into classes with the closest characteristics and collect the highest quality concrete mix formulations into the appropriate classes. The result of using DataMining technology for clustering concrete mix formulations allowed us to create classes in which the distribution using the DBSCAN algorithm is quite high-quality, however, there is a need for more detailed training of this algorithm, since clustering using the program "Integrated Quality Assessment and classification of multidimensional objects" turned out to be more optimal.

Keywords : clustering, datamining, concrete mix formulations, quality criteria

БЕТОН ҚОСПАЛАРЫНЫҢ РЕЦЕПТЕРІН КЛАСТЕРЛЕУГЕ АРНАЛҒАН DATA MINING МҮМКІНДІКТЕРІ

К.М.Акишев[1*], В.И.Карпов[2], Л.Ақышева[3], А.Д.Тулеғулов[1]

[1]Қазақ технология және бизнес университеті,

[2] К. Г. Разумовский атындағы Москеу мемлекеттік технология және басқару университеті

[3] Назарбаев Зияткерлік мектебі,

e-mail: Akmail04cx@mail.ru

Мақалада бетон қоспаларын кластерлеу үшін DataMining технологиясын қолдану мүмкіндіктері қарастырылады. Іс жүзінде сіз көбінесе бетон қоспаларының көптеген рецептерінен сапалық сипаттамалары бойынша ең жақын рецептілерді таңдау қажет болатын міндеттерге тап болуыңыз керек. Бетон қоспаларының рецептураларын сыныптар бойынша бөлу беріктік, сондай-ақ бетон қоспасының ингредиенттерінің құрамы сияқты берілген критерийлер негізінде қамтамасыз етіледі. [1] -жұмыстың басында бетон қоспаларының рецептураларын кластерлеу "сапаны кешенді бағалау және көпөлшемді объектілерді жіктеу" бағдарламасының көмегімен жүзеге асырылды, бұл рецептураларды ең жақын сипаттамалары бар сыныптарға бөлуге және бетон қоспаларының ең сапалы рецептураларын тиісті сыныптарға жинауға мүмкіндік берді. Бетон қоспаларының формулаларын кластерлеу үшін DataMining технологиясын қолданудың нәтижесі DBSCAN алгоритмін қолдана отырып бөлу өте сапалы болатын сыныптар құруға мүмкіндік берді, дегенмен бұл алгоритмді егжей-тегжейлі оқыту қажет, өйткені "сапаны кешенді бағалау және көп өлшемді объектілерді жіктеу" бағдарламасын қолдана отырып кластерлеу оңтайлы болды.

Түйін сөздер: кластерлеу, datamining, бетон қоспаларының формулалары, сапа критерийлері.

ВОЗМОЖНОСТИ DATAMINING ДЛЯ КЛАСТЕРИЗАЦИИ РЕЦЕПТУР БЕТОННЫХ СМЕСЕЙ

К.М.Акишев[1*], В.И.Карпов[2], Л.Акишева[1], А.Д.Тулегулов[1]

[1]Казахский университет технологии и бизнеса,

[2]Московский Государственный университет технологии и управления им. К.Г. Разумовского,

[3]Назарбаев интеллектуальная школа,

e-mail: Akmail04cx@mail.ru

В статье рассматривается возможности использования технологии DataMining для кластеризации бетонных смесей. На практике часто приходится сталкиваться с задачами в которых необходимо из большого количества рецептур бетонных смесей выбрать наиболее близкие по качественным характеристикам рецептуры. Распределение рецептур бетонных смесей по классам обеспечивается на основании заданных критериев таки, как прочность, а также состав ингредиентов бетонной смеси. Раннее в работе [1], кластеризация рецептур бетонных смесей была осуществлена с помощью программы «Комплексная оценка качества и классификация многомерных объектов», позволившая распределить рецептуры по классам с наиболее близкими характеристиками и собрать наиболее качественные рецептуры бетонных смесей в соответствующие классы. Результат использования технологии DataMining для кластеризации рецептур бетонных смесей позволил создать классы в которых распределение с помощью алгоритма DBSCAN достаточно качественно, тем не менее существует необходимость более детального обучения данного алгоритма, так как кластеризация с использованием программы «Комплексная оценка качества и классификация многомерных объектов» оказалось более оптимальной.

Ключевые слова: кластеризация, datamining, рецептуры бетонных смесей, критерии качества

Introduction. To date, the active promotion of information technology (datamining) provides the necessary and sufficient opportunities to obtain reliable and high-quality results, in particular for solving clustering and forecasting problems.

Data analysis depends on efficient data collection, storage and computer processing. Data Mining allows you to analyze large volumes of heterogeneous data of various scientific fields.

The international market of DataMining systems has a dynamic growth. Firms such as SAS, IBM, Microsoft, Oracle, provide investments of $56.2 billion by 2027 [2].

Promising trends in DataMining allow us to develop methods of virtual and augmented reality analysis, statistical data analysis, and data protection.

It is known for certain that data mining in the future for big data analysis using corporate databases.

At the same time, the main criterion for DataMining technology is the system time required to complete the tasks. At the same time, the main difficulty lies in the limitations that arise during the search of decision trees, which affect the efficiency and performance of the search.

Solving this problem remains the main goal of DataMining product developers[2].

When performing clustering tasks, there are a number of typical, standard stages for DataMining, which include:

-statement of the task, which includes the analysis of requirements, the definition of the problem area, the metrics for which the assessment will be performed, as well as the definition of tasks for the analysis project;

-preparation of data, evaluation criteria;

-research and evaluation of data;

-building analytical dependencies;

-research, verification of the accuracy of solutions.

Today, everyone who deals with problems in the field of big data processing must have skills in the field of mathematical statistics, programming languages, machine learning techniques, statistical analysis, predictive decisions, including managerial decision-making, depend on the accuracy and in-depth data analysis.

To date, there is a wide selection of programs on the market for solving Data Mining problems. Let's look at the main ones:

SAS EnterpriseMiner - used mainly for fraud detection, financial risk assessment, market forecasting, etc. It has a fairly high performance when working with big data;

-MicrosoftAnalysisServices - used to create analytical reports;

-SAS CustomerIntelligence 360 - used as a tool for information business, evaluation of marketing campaigns, real-time data analysis;

-SAS CreditScoring - used as a risk management tool for financial institutions;

-Board - nothing remarkable, for business analytics, corporate governance, evaluation of the effectiveness of projects.

-SAS RevenueOptimization - a use case as an intellectual business tool, mainly in retail;

-RapidMiner - is used for text analysis, machine learning, and the creation of analytical reports.

The aim of the study is to evaluate the DataMining technology for clustering concrete mix formulations and compare the results with the data obtained using the program "Comprehensive quality assessment and classification of multidimensional objects" with the possibility of using DataMining technology in further studies of clustering problems of technogenic deposits.

In [1], prior to the start of cluster analysis, each object under study or the formulation of a concrete mixture is a separate cluster, and the proximity between clusters is assumed by accepted metrics.

The most optimal way to solve the research problem, determining proximity (distances) between clusters in the studied space, or as they say in many sources, Euclidean distances. "Euclidean distance is a general type of distance used from ancient times to the present day, it is a geometric distance in multidimensional space"[3] and is used in various methods.

As a result of the solution [1], a mathematical formulation of the problem of clustering concrete mixtures using technogenic waste formula 1 was obtained.

$$F(C_k, C_l) = \frac{1}{N_k \cdot N_l} \sum_{SiCk} \sum_{SjCl} f(S_j, S_j) \quad (1)$$

That is, each cluster contains formulations of concrete mixtures that are closest in their characteristics and the formulations of each class differ from each other.

For the software implementation of the mathematical formulation of the solution of the problem of clustering concrete mixtures, the program "Comprehensive quality assessment and classification of multidimensional objects" in Russian was used. Data on the metrics of the recipe of concrete mixtures are presented in Table 1.

Table 1. Recipe of concrete mixtures

№ recipes	Composition of the concrete mix			
	Ravr. Compr.Mpa	ash g/%	Metall. Slag g/%	bauxite sludge g/%
1	3.11			337/3
2	3.7			505/4
3	8.32			674/5
4	3.5			842/7
5	3.6			1011/8
6	2.84			944/7
7	2.56			910/7
8	2.73	574/4.4		
9	2.3	246/2		337/3
10	4.4	574/4.4		
11	2.5	656/5		
12	4.5	3370/26		
13	2.51	492/4		
14	2.8	574/4.4		
15	3.45			410/3.2
16	4.54			410,3.2
17	3.9	246/2	798/6	
18	2.55	328/2.5	798/6%	
19	3.5	328/2.5	640/5	
20-15	1.47	107/1.3	1066/13	
21	9.74	328/2.5	798//6	246/2
22	3.18		399/3	328/2.5
23	3.23	164/1.3	798/6	164/1.3

24-$_{15}$	1.6		1020/12	357/4.2
25	4.77	328/2.5	399/3	337/2.5
26	4.33	164/1.2	798/6	337/2.5
27	3.3	337/2.5	798/6	164/1.2
28	9	337/2.5	798/6	505/4
29	10	337/2.5	1596/12	
30	3.22	328/.5	798/6	674/5.2
31	4.62	246/2		505/3.2
32	2.3			1685/13
33	9.1	164/1.3		1685/13
34	8.18	164/1.3		2022/16
35	4.37	1011/8		
36	1.35	7798/80		
37	4.43	337/1.5	798/6	674/5.2
38	4.13	337/2.5	798/6	337/2.5
39	22.3	505/4	1197/9.	337/2.5

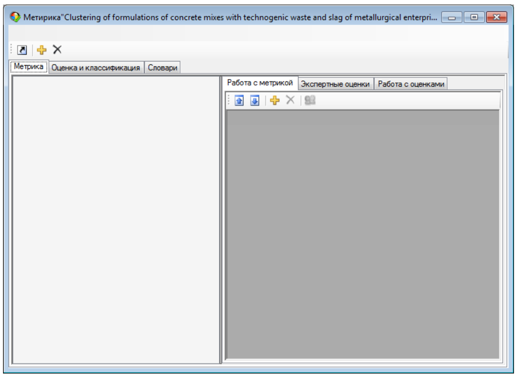

Fig. 1- Program interface ""Comprehensive quality assessment and classification of multidimensional objects"

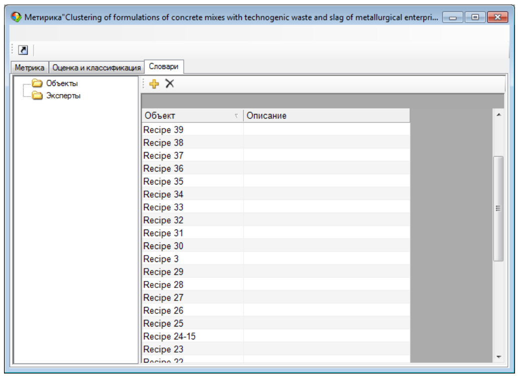

Fig. 2 - Shows the input of metrics for compounding concrete mixtures

Fig.1-Shows the interface of the program "Comprehensive quality assessment and classification of multidimensional objects". The data is entered directly from the keyboard**.**

As a result of the program, the distribution of formulations of concrete mixtures of Table 1, Figure 3 was obtained. In each cluster there are recipes that are closest in terms of metrics.

From the clustering result obtained, it can be seen that the formulations of concrete mixtures are distributed in 6 clusters. In each cluster there are formulations of concrete mixtures with the closest characteristics in composition and strength. The formulations of concrete mixtures with low strength indicators are located in clusters 1-4. Table 2 shows the formulations, the composition of concrete mixtures with the highest strength indicators of 5 and 6 clusters.

Table 2. Recipe of cluster 5-6

Cluster №	№ recipes	Ravr.compr. Mpa	ash g/%	Metall. Slagg/%	bauxite sludge g/%
5	RBS3	8.32			674/5
	RBS34	8.18			842/7
	RBS28	9	337/2.5	798/6	505/4
	RBS29	10	337/2.5	1596/12	
	RBS33	9.1	164/1.3		1685/13

Table 2. Recipe of cluster 5-6

Cluster №	№ recipes	Ravr.compr. Mpa	ash g/%	Metall. Slagg/%	bauxite sludge g/%
	RBS21	9.74	328/2.5	798//6	246/2
6	RBS39	22.3	505/4	1197/9.2	337/2.5

Methods and materials. As a software tool for clustering concrete mix formulations, we use the DBSCAN clustering algorithm [4-11].

DBSCAN (Density-based spatial clustering of applications with noise), a density algorithm for spatial clustering with the presence of noise), as the name implies, operates with data density.

We use the same data as in [1].Ideally, DBSCAN can achieve good results, but it's not worth hoping for. Many versions of the algorithm are able to work from cluster to cluster.

At the same time, the algorithm considers clusters as high-density areas separated by low-density areas.

Because of this, clusters obtained in DBSCAN come in any shape, unlike k-means, which assume that clusters are convex.

In practice, an important component of DBSCAN is a cluster consisting of a set of core samples, "each of which is close to each other (measured using some distance measurement measure) and a set of non-core samples that are close to the core sample (but are not core samples themselves)" [4].

2 parameters are important for the DBSCAN algorithm: min_samples and eps, high min_samples or lower eps, provide the high density necessary for cluster organization.

The algorithm is formed in this way:

1. All points are marked as core, boundary or noise;
2. Interference will be eliminated;
3. The face between all the main points located inside at the distance of the Eps parameter from each other is marked;
4. Each group is placed in a separate cluster;
5. The boundary points of one of the clusters associated with this boundary point are specified.

At the same time, the base sample is part of a cluster, a sample that is not a core sample, and exists, at a distance of eps from any core sample, is considered an anomaly of the algorithm.

The algorithm executed (DensitybasedClusteringAlgorithm) allows the cluster to grow until the density in the neighboring cluster exceeds a certain threshold.

The DBSCAN algorithm is deterministic, always generating the same clusters when providing the same data in the same order. However, the results may differ if the data is provided in a different order. First, even though core samples will always be assigned to the same clusters, the labels of these clusters will depend on the order in which these samples occur in the data. Secondly, and more importantly, the clusters to which the non-core samples are assigned may differ depending on the order of the data.

The clustering algorithm is effective when, as a rule, the "compactness hypothesis" can be implemented, while splitting objects into classes, the distances between objects from the same class (intra-clusterdistances) will be less than some value ϵ>0�>0, and between objects from different classes (cross-clusterdistance) will be greater than ϵ�.

For our case, a table of distances between classes is formed in Figure 4.

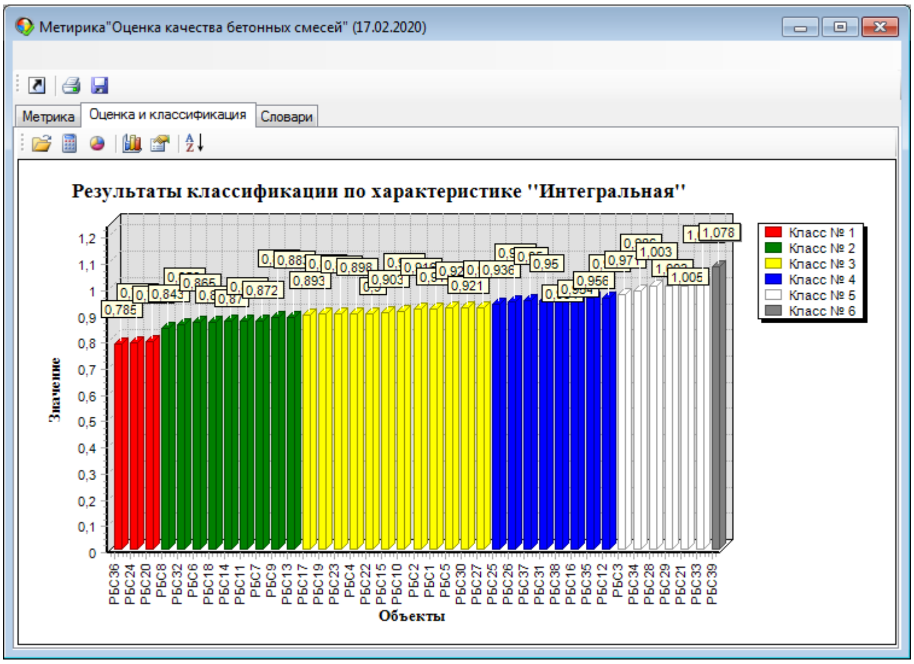

Fig. 3- Clustering of concrete mix formulations

```
library('reshape2')
n <- dim(df.stand)[[1]]
euc.dist <- as.matrix(dist(df.stand))
dist = melt(euc.dist)
df.stand$cluster <- clus$cluster
pairs <- data.frame(dist = dist,
                    ca = as.vector(outer(1:n, 1:n,
                                   function(a, b) df.stand[a, 'cluster'])),
                    cb = as.vector(outer(1:n, 1:n,
                                   function(a, b) df.stand[b, 'cluster'])))
dcast(pairs, ca ~ cb, value.var = 'dist.value', mean)
```

```
##   ca         1        2        3        4        5
## 1  1 0.8192236 3.740748 4.651124 2.865534 1.589380
## 2  2 3.7407483 1.283606 2.556633 2.986930 3.136207
## 3  3 4.6511236 2.556633 1.584229 2.777835 3.656428
## 4  4 2.8655340 2.986930 2.777835 1.293573 1.822823
## 5  5 1.5893797 3.136207 3.656428 1.822823 1.030227
```

В полученной таблице по главной диагонали приведены средние внутрикластерные расстояния, которые очевидно меньше, чем межкластерные расстояния (недиагональные элементы таблицы).

Поскольку объекты таблицы `USArrests` многомерны, то для вывода ординационной диаграммы выполним свертку информации по 4 имеющимся показателям к двум главным компонентам. После этого можно осуществить визуализацию групп (см. рис. 2.13) и “на глаз” оценить качество кластеризации.

```
c.pca <- prcomp(USArrests, center = TRUE, scale = TRUE)
d <- data.frame(x=c.pca$x[, 1], y=c.pca$x[, 2])
```

Fig. 4 - Table of distances between classes

```
c.pca <- prcomp(USArrests, center = TRUE, scale = TRUE)
d <- data.frame(x=c.pca$x[, 1], y=c.pca$x[, 2])
d$cluster <- clus$cluster
library('ggplot2')
library('grDevices')
h <- do.call(rbind, lapply(unique(clus$cluster),
                    function(c) { f <- subset(d,cluster==c); f[chull(f),]}))
ggplot() + geom_text(data = d,
                     aes(label = cluster, x = x, y = y, color = cluster),
                     size = 3) +
    geom_polygon(data = h,
                 aes(x = x, y = y, group = cluster, fill = as.factor(cluster)),
                 alpha = 0.4, linetype = 0) +
    theme(legend.position = "none")
```

Fig. 5 - Dependencies of distances inside and outside clusters

From Figure 5, we see that the average intracluster distances are significantly smaller than the intercluster distances.

As a result, using the DBSCAN algorithm, the distribution of concrete mix formulations by class is obtained, as shown in Figure 6.

Figure 7 shows the clustering of concrete mix formulations.

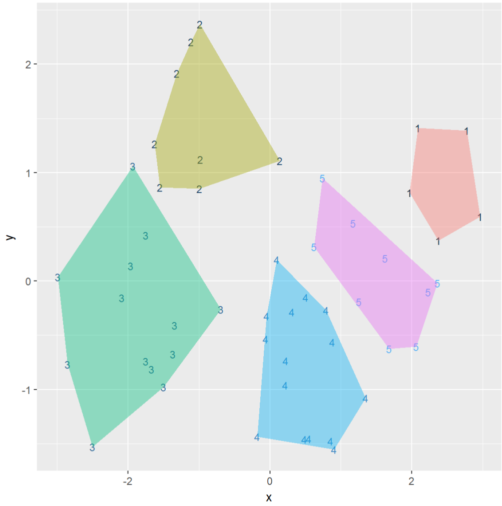

Fig. 6 - Results of the distribution of concrete mix recipe by class

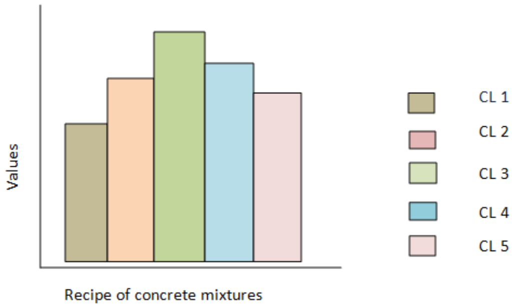

Fig. 7 - Distribution of concrete mix formulations by clusters

Discussion and results. As can be seen from Figure 7, all the formulations of concrete mixtures are combined into 5 clusters. The largest is cluster 3, the smallest is cluster 1.

The task of describing in detail which formulations of concrete mixtures were included in each cluster was not set.

Clustering technology using datamining is quite complex, the difficulty lies in the uncontrolled decision-making process. It is not completely clear whether the correct solution has been achieved.

"Deep artificial neural networks are very good at classification, but clustering is still an open question"[2].

Conclusions. We see that DataMining is quite a promising tool for clustering, including formulations of concrete mixtures. Although the results leave much to be desired compared to those previously obtained in [1], nevertheless, with fairly constant machine learning and training, there is a potential prospect that we will get good clustering results. In the future, we plan to use one of the software tools presented above for clustering man-made waste of the Pavlodar region.

Referenses

1. Akishev K and other. Mathematical formulation and the problem solution of clustering recipes of concrete mixtures using technogenic waste waste and slags of metallurgical enterprises.- Metallurjia. - 2022. - 61(1)213-216. Zagreb, p.321.
2. Electronicresource URL: https://trends.rbc.ru -Date of application 12.05.2023.
3. Electronic resource:URL: https://cyberleninka.ru - Date of application 12.05.2023.
4. Electronic resource: URL:https://etu.ru. -date of application 12.05.2023.
5. Electronic resource: URL:https://elibrary.ru - Date of application 12.05.2023.
6. Electronic resource: URL:https://habr.com -Date of application 12.05.2023.
7. Electronic resource:URL: http://pzs.dstu.dp.ua/DataMining - Date of application 12.05.2023.
8. Electronicresource: URL: https://www.math.spbu.ru -Date of application 12.05.2023.
9. Electronicresource:URL: https://core.ac.uk - date of application 12.05.2023.
10. Electronic resource :URL: Ortiz-Arroyo D. Discovering Sets of Key Players in Social Networks // Computational Social Networks Analysis. - 2010 - C. 27-47 [Date of application 12.05.2023.
11. Electronic resource:URL: https://kpfu.ru/portal/docs/F_1980845423/161_3_phys_mat_8.pdf - Date of application 12.05.2023.

Information about the authors

- Akishev Karshyga Maksutovich, Candidate of Technical Sciences, Ass. Professor, Department of "Information Technology", Kazakh University of Technology and Business, Astana, Republic of Kazakhstan

  email:akmail04cx@mail.ru
- Karpov Valery Ivanovich, Doctor of Technical Sciences, Professor, Department of Information Technology, Moscow State University of Technology and Management named after K.G. Razumovsky, Moscow, Russia email:Vikarp@mail.ru
- Akisheva Lenara Karshygaevna, 7th grade student of Nazarbayev Intellectual School, Astana, Republic of Kazakhstan email:LK@mail.ru
- Tulegulov Amandos Dabysovich, Ph.D., Ass. Professor, Head of the Department of "Information Technology", Kazakh University of Technology and Business, Astana, Republic of Kazakhstan email:tud62@yandex.ru

Сведения об авторах

- Акишев Каршыга Максутович, к.т.н., асс. профессор, кафедра «Информационные технологии», Казахский университет технологии и бизнеса, г. Астана, Республика Казахстан: email:akmail04cx@mail.ru
- Карпов Валерий Иванович, доктор технических наук, профессор, кафедра «Информационные технологии» Московский Государственный универститет технологии и управления имени К.Г. Разумовского,г. Москва, Россия:email:Vikarp@mail.ru
- Акишева Ленара Каршыгаевна, ученица 7 классса Назарбаев Интеллектуальная школа, г. Астана, Республика Казахстан LK@mail.ru
- Тулегулов Амандос Дабысович, к.ф.м.н., асс. профессор, зав. кафедрой «Информационные технологии», Казахский университет технологии и бизнеса, г. Астана, Республика Казахстан: tud62@yandex.ru