Pre-processing of data on the behavior of users of the Moscow Electronic School service

Y V Frolov and T M Bosenko
Moscow City Pedagogical University, 28, Sheremetyevskaya street, Moscow, 127521, Russia

E-mail: boss-ua@yandex.ru

Abstract. The article presents the results of preliminary processing of data on the behavior of users of the Moscow Electronic School with a volume of about 50 thousand observations. The relevance of the study is due to the increase in the volume of distance and mixed education in the context of digitalization of space. The aim of this study was to identify groups of teachers (users) exhibiting differing behaviors when interacting with educational content on the Moscow Electronic School platform. According to the results of the study, the relationships between the behavioral parameters of users were determined and 5 behavioral types of the Moscow Electronic School users were identified, which will further allow using this approach in increasing the efficiency of the Moscow Electronic School platform as a recommendation system.

1. Introduction
Moscow Electronic School (MES) is a cloud-based Internet platform that contains educational materials, tools for creating and editing them, as well as designers of educational scenarios and programs [1]. Currently, the MES system: hosts more than 520,000 educational materials; approximately 55,000 teachers use MES resources; presents more than 41540 interactive lesson scenarios, 348 e-textbooks, and 1,270 e-workbooks created by Moscow school teachers in various subjects.

The purpose of this study was to pre-process data for performing cluster analysis-grouping of MES users (secondary school teachers) who demonstrate different behavior when interacting with educational content on the MES platform.

The scientific problem of the work was to study the MES user database and assess its suitability for cluster analysis. This kind of knowledge, according to the authors, will make it possible to manage the complex processes of choosing educational content and technologies for working in the system by new users, in particular young specialists – teachers, more effectively.

The relevance of the research is determined by the increasing volume of distance and mixed education in the context of its digitalization [2]. In terms of the issue statement and the tools used, the relevance is also determined by the possible use of cases on pre-processing of big data in the educational process of training business analysts, whose competencies should integrate not only knowledge and skills in machine learning techniques, but also the ability to interpret the results of analytical research [3-5].
2. Related work
Analysis of publications on this problem [6-8] shows that most clustering methods and algorithms are focused on a small dimension of the feature vector of the studied object. In [9] the clustering process is presented as a model that is created based on three stages:

- "data research", the goal is to answer the question: are there any significant groups of units based on measures for a set of variable features or responses;
- creating and / or testing a hypothesis about the cluster structure;
- the performed cluster analysis should confirm previously published clustering results.

This approach allows us to transfer the multi-stage search for the best clustering to the theory of cluster analysis; a criterion approach for evaluating the quality of clustering, etc. As a result, there is a need to develop effective methods for this type of data preprocessing and reducing the dimension of the feature space without significant loss of information about the objects under study.

The study was intended to offer measurable indicators that would allow users to be divided into clusters with uniform behavior and that could be used in the future in the process of designing a recommendation system.

3. Problem statement
The task of the study included a step-by-step preprocessing of the source data, including filling in the missing values of object attributes, identifying and removing anomalies, normalizing the values of features [10], and conducting a cluster analysis of the behavior of users of the MES platform. The study used the following parameters: user actions with objects in the MES system (table 1).

| № of category | Name                                           |
|---------------|------------------------------------------------|
| 1             | Number of views of a specific user's content   |
| 2             | Number of favorites added by a specific user   |
| 3             | The number of user ratings given to a specific content author |
| 4             | The number of uses of scripts in a specific user's homework |
| 5             | The number of copies of content created that belong to a specific user |
| 6             | Number of user-created scenarios in the MES    |

The study analyzed the database in the context of the requirements for data from cluster analysis methods.

4. Methods
To perform cluster analysis, the following criteria should be followed [11-13]:

- indicators (measurements, data) should not be correlated with each other;
- indicators (measurements, data) must be dimensionless;
- their distribution should be close to normal;
- the data sample must be uniform and contain no "outliers".

The most common clustering method is the k-means method. It was created in the 50s of the twentieth century by mathematicians Hugo Steingauz and Stuart Lloyd [14]. The algorithm is designed to minimize the total square deviation of cluster points from the centers of these clusters.
\[ V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2, \]

where \( k \) is the number of clusters, \( S_i \) - obtained clusters, \( i = 1..k \), \( \mu_i \) - centers of mass of the \( x_j \in S_i \) vectors. The algorithm splits the set of elements of a vector space into a pre-known number of clusters \( k \). The main idea is that at each iteration, the center of mass is recalculated for each cluster obtained in the previous step, then the vectors are divided into clusters again according to which of the new centers was closer according to the selected metric. The algorithm terminates when the center of mass of the clusters does not change at some iteration.

Preliminary data analysis and construction of clustering models using the k-means method were performed in the environment of the IBM SPSS STATISTICS package [15] and the Python numpy package. To train neural networks with a multi-layer perceptron configuration, the STATISTICA 13. TIBCO Software Inc. package was used [16]. Data visualization was performed in Tableau Desktop [17].

5. Results
Based on the results of the study, the relationships between the behavioral parameters of users were determined. Figure 1 illustrates the results of scattering and histograms of the studied values (indicators of user behavior).

![Figure 1. Scattering and histograms of MES user behavior indicators.](image-url)
From the analysis of figure 1, we can conclude that the indicators that characterize user behavior in the MES system have three types of interaction: 1) linear correlation (for example, "the number of copies created" and "the number of scenarios created" are related); 2) non-linear relationship (this type of relationship shows the indicator "the number of uses of scenarios in the HW" in relation to all other indicators of user behavior); 3) there is no correlation or weak connection (for example, between the indicator "number of additions to favorites" with three parameters – "number of copies created", "number of scenarios created", "number of ratings issued").

In the context of studying the data structure, an analysis of "unusual" observations was also performed, which includes outliers in the values of the studied behavioral indicators (table 2).

**Table 2. The number of unusual observations by categories of behavioral indicators (user actions).**

| № of category | Name of input parameters | Number of "unusual observations" |
|---------------|-------------------------|----------------------------------|
| 2             | Number of views         | 272                              |
| 3             | Number of favorites added | 134                             |
| 4             | The amount of the proposed assessment | 80                             |
| 5             | Number of uses of scenarios in the H/W | 231                            |
| 6             | Number of copies created | 220                              |
| 7             | Number of scenarios created | 215                            |
| TOTAL         |                         | 1152                             |

As unusual observations cannot be clustered, but are submitted for study in classification theory, such data were excluded from consideration during further cluster analysis and detailing of clustering features.

Figures 2 and 3 show the results of visualization of clusters of MES user databases by two parameters - the average value of the number of views and the average value of the number of created scenarios.

**Figure 2.** Visualization of clusters of the base MES by the K-Means method: the average number of views “A” – the area of a rectangle; the average number of created scenarios “B” - coloring.
From figure 2 we can see that cluster 4 is the cluster with the highest average value of the "number of created scenarios" indicator, and cluster 3 is the cluster with the highest average value of the number of views.

The results shown in figure 3 indicate that clusters with low average values of the two studied parameters "merge" into spots.

The simplified two-dimensional visualization shown in figure 3 is just one of the ways to analyze data and allows you to more clearly divide the clusters of the MES user base by the limited number of clustering factors selected by researchers. However, the results of the study show that in each cluster, user behavior depends on a larger number of parameters.

6. Conclusion
Based on the results of the study, the MES database was analyzed according to the criterion of its suitability for cluster analysis. A linear correlation was established between the behavioral indicators "number of copies created" and "number of scenarios created". A non-linear relationship is shown by the indicator "the number of uses of scenarios in the H/W" in relation to all other indicators of user behavior. Visualization of clusters of the MES user base is performed.

References
[1] Frolova E V, Ryabova T M and Rogach O V 2019 Digital Technologies in Education: Problems and Prospects for Moscow Electronic School Project Implementation European Journal of Contemporary Education (Slovakia, Academic Publishing House Researcher) 8(4) 779-89
[2] Skrypnikova N N 2018 Technology of mixed training: relevance and problems Professional education and labor market 3 74-8
[3] Frolov Y V and Bosenko T M 2020 Training of personnel for the development of innovative entrepreneurship Academy of Entrepreneurship Journal 26(327) 1-6
[4] Frolov Y V 2001 Effective Technologies of Economic Choice of MSFU p 293
[5] Frolov Y V 2020 Knowledge Management: Textbook for Bachelor's and Master's Degree (2nd ed., cor. and add) (Moscow: Youwrite publishing house)
[6] Saracli S, Dogan N and Dogan I 2013 Comparison of hierarchical cluster analysis methods by cophenetic correlation Journal of Inequalities and Application 203 1-8
[7] Rezankova H 2014 Cluster analysis of economic data Statistica 94(1) 73-86
[8] Jain A K, Murty M N and Flynn P J 1999 Data clustering: A review ACM Computing Surveys 31(3) 264-323
[9] Smart J C et al 2005 Higher Education: Handbook of Theory and Research Springer Printed in Great Britain XX 437-57
[10] Jiawei Han, Micheline Kamber and Jian Pei 2011 Data Mining: Concepts and Techniques (3rd) (Edition Elsevier Science LtdPublication) p 703
[11] Tyurin A G, Zuev I O 2014 Cluster analysis, clustering methods and algorithms Bulletin of MGTU
[12] Ayvazyan S A, Buchstaber V M, Enyukov I S and Meshalkin L D 1989 *Applied Statistics: Classification and Dimension Reduction* (Moscow: Finance and statistics) p 607

[13] Mandel And D 1988 Cluster Analysis (Moscow: Finance and statistics) p 176

[14] J Van Raisin 1980 *Classification and Cluster* (Moscow: Mir) p 390

[15] Software platform for statistical analysis IBM SPSS STATISTICS 26 ([https://www.ibm.com/ru-ru/products/spss-statistics](https://www.ibm.com/ru-ru/products/spss-statistics))

[16] STATISTICA TIBCO ® Data Science software platform for statistical analysis and machine learning ([https://www.tibco.com/products/data-science/downloads](https://www.tibco.com/products/data-science/downloads))

[17] Tableau Desktop software platform ([https://www.tableau.com/products/desktop](https://www.tableau.com/products/desktop))