Application of Big Data Methods in E-Learning Systems

Natalia Sharonova[a], Iryna Kyrychenko[b] and Glib Tereshchenko[b]

[a] *National Technical University "KhPI", Kyrpychova str. 2, Kharkiv, 61002, Ukraine*
[b] *Kharkiv National University of Radioelectronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine*

Abstract

Analytics and Big Data play an important role in the future of higher education. This paper analyzes and practices the use of e-learning technology tools to provide relevant information for teachers and students trying to optimize the learning process. The combination of data processing and analytical training is an aid that will greatly enhance higher education and determine the path for further development in the new educational era.

Keywords 1

E-learning technological tools, curriculum analysis, educational data acquisition, Big Data, learning management system

1. Introduction

Currently, there is already a large volume of data from pupils. who have access to LMS [1]. The growth of students in new distance education systems is driving in the educational field, there is a new trend. The recent rise in popularity of MOOC is an example of the new expectations that are being offered to university students. This tendency leads to a change in the role of behavior in different educational roles, where both teachers and students have to respond to new methods and change their traditional teaching methods. This phenomenon is not limited to public schools, which must adapt their structure and information structures to accommodate the demands of students in order for them to gain access to their academic programs.

Mining data is widely used in the education industry to find problems in the industry. Student performance is of great concern in educational institutions where performance may be influenced by several factors. There are three necessary components to forecasting: parameters that affect student performance, data mining methods, and a third, data mining tools. These parameters can be psychological, personal and environmental. The research conducted in this paper is aimed at supporting the quality of the education of the institute, minimizing the diverse impact of these factors on student success.

Big Data is the automated fusion of organized data stored in libraries with unstructured data from emerging outlets such as social media, electronic devices, cameras, smart meters, and financial systems. Big Data, on the other hand, is described by the McKinsley Global Institute as "data sets that transcend the capacity of traditional database software to record, process, handle, and analyze." Today, this approach enables businesses to collect and interpret all data, regardless of the type, volume, or speed of transmission, and make more informed decisions based on that data.

It has been decided that there is so much to discover about how to manage Big Data in the same way that everyone else does. But one thing is certain: conventional data-processing methods would not lead to Big Data research performance [2]. The number of data sources, the volume of data, the processing time, and even the key business models all contribute to a broad data space. Recommendations to use the same old tools under these new conditions are not suitable for data analysis.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: nvsharonova@ukr.net (N. Sharonova); iryna.kyrychenko@nure.ua (I. Kyrychenko); hlib.tereshchenko@nure.ua (G. Tereshchenko)
ORCID: 0000-0002-8161-552X (N. Sharonova); 0000-0002-7686-6439 (I. Kyrychenko); 0000-0001-8731-2135 (G. Tereshchenko)


Data mining is the study and discovery of secret information by "machines" (algorithms, artificial intelligence) in raw data that was previously unknown, non-trivial, functional, and interpretable by humans [3].

The main stages of solving problems using Data Mining methods are:

1. setting the task of analysis;
2. data collection;
3. data preparation (filtering, supplementing, coding);
4. choice of model (data analysis algorithm);
5. selection of model parameters and training algorithm;
6. model training (automatic search of other model parameters);
7. analysis of the quality of training, if unsatisfactory, then move to item 5) or item 4);
8. analysis of identified patterns, if unsatisfactory, then move to paragraph 1), 4) or 5).

The choice of data analysis method is based on some features of the source data. In our case, we can distinguish the following features:

- No prior knowledge of the data being analyzed, since we are in the initial stages of analysis;
- The number of groups to which each sample object will be assigned is unknown in advance;
- Object partitioning must take place on a whole set of features, not on a single dimension.

Based on these features, a clustering method was selected for this study using a mathematical apparatus for cluster analysis.

Cluster analysis is a set of mathematical methods designed to form relatively "distant" friends of groups of "related" objects based on distances or relationships between them.

Clustering differs from classification in that the solution of the problem is possible without any prior knowledge of the analyzed data.

Cluster analysis has the advantage of allowing you to split objects not only by a single parameter, but by a whole range of attributes, as well as viewing a large amount of raw data of almost any kind.

The task of clustering is to divide the studied set of objects into groups of "similar" objects, which are called clusters [4]. A cluster in English means a bunch, a bundle, a group.

Classification tasks involve assigning each data object to one (or more) of predefined classes, and in a clustering task assigning each of the data objects to one (or more) of previously unknown classes.

Note a number of features inherent in the problem of clustering. The decision depends heavily on the nature of the data objects and their attributes, i.e. they can be uniquely defined objects, accurately quantified objects, and may be objects that have a plausible or fuzzy description.

The decision also depends heavily on the representation of the clusters and the predicted relationships between the data objects and the clusters. That is, you need to consider such features as the ability or inability to attach objects to multiple clusters. It is also necessary to define the very concept of cluster membership: unambiguous (belonging / not belonging), probabilistic (belonging probability), fuzzy (degree of belonging).

2. Choosing the algorithm

Cluster analysis divides a group of objects G into m (m - integer) clusters (subsets) Q_1, Q_2, ..., Q_m based on data found in a large number of X, such that Gj belonged to one and only one subset of the partition, and objects belonging to the same cluster were identical, while objects belonging to different clusters were heterogeneous.

When clustering, the number of clusters generated is crucial. Clustering is designed to detect natural object thickening on a local level. As a consequence, the number of clusters is a parameter that, if considered undefined, can significantly complicate the form of algorithm and, if understood, can significantly affect the consistency of the result.

Usually nothing is known at the beginning of a data survey, so clustering algorithms are usually built as a way to sort through the number of clusters and determine its optimal value.

The number of methods of splitting a set into clusters is quite large. All of them can be divided into hierarchical and non-hierarchical [4].

Hierarchical clustering combines small clusters into large clusters or splits large clusters into small clusters. Hierarchical algorithms are in turn divided into agglomerative and divisible ones.

Agglomerative methods are characterized by sequential integration of the original elements and a corresponding decrease in the number of clusters. At the beginning of the algorithm, all objects are separate clusters. Initially, the most similar objects are clustered. The merge then proceeds until all the objects form a single cluster.

The sequential division of the initial cluster composed of all items, as well as the resulting increase in the number of clusters, define divided approaches. Both objects belong to a single cluster at the start of the algorithm, which is then separated into smaller clusters in subsequent stages, resulting in a series of splitting sets.

Non-hierarchical algorithms try to group the data into clusters so that the objective function of the partition algorithm reaches the extremum (minimum).

A very important issue is the problem of selecting the required number of clusters. Sometimes m number of clusters can be chosen a priori. However, in the general case, this number is determined in the process of splitting the set into clusters. Studies were conducted by Fortier and Solomon, and it was found that the number of clusters should be accepted to achieve the probability α that the best fit was found. Thus, the optimal partition number is a function of a given fraction β of the best, or in some sense permissible, partitions in many of all possible [5]. The total scattering will be greater than the fraction β of the admissible partitions.

Based on the information obtained for the experiment were selected algorithms COBWEB, DBSCAN, hierarchical clustering algorithm, XMEANS and EM algorithm.

Cluster analysis of model data using the selected algorithms showed good results, which is why these algorithms are used for further clustering of real data.

The DBSCAN algorithm is an algorithm for clustering spatial data with the presence of noise, proposed by Martin Esther, Hans-Peter Krigel, and colleagues in 1996 as a solution to the problem of splitting data into arbitrary clusters. It is density-based: for a given set of points in some space, it assigns to one group the points that are closest (the points with many neighbors) and marks the points that lie in areas of low density (whose neighbors are too far apart) as emissions. DBSCAN is one of the most common clustering algorithms as well as the most cited in the scientific literature. The authors of DBSCAN have experimentally shown that the algorithm is able to recognize clusters of different shapes, such as in Figure 1.

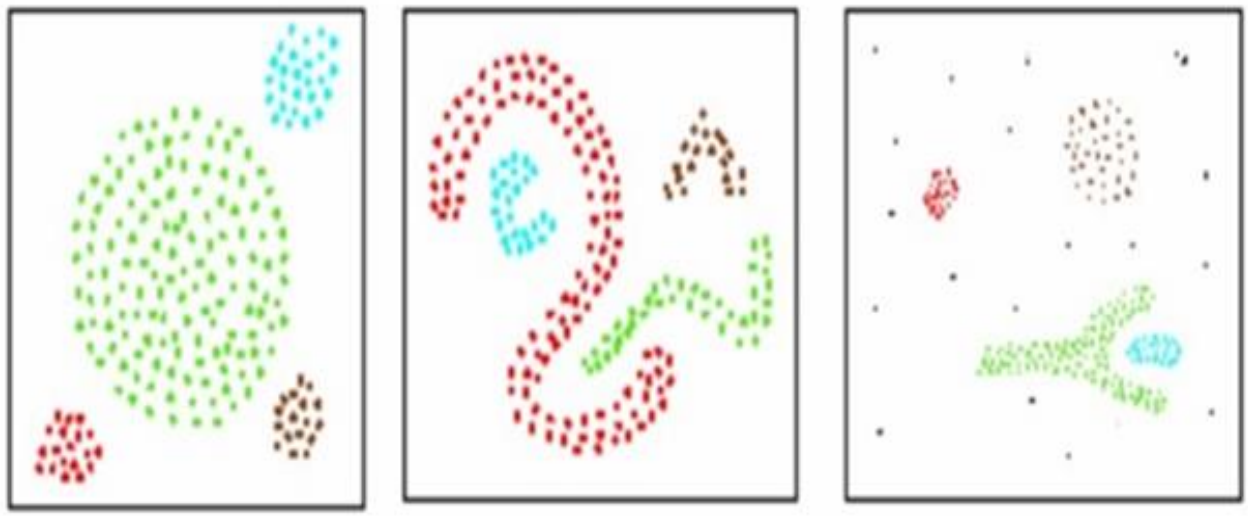

Figure 1: Examples of arbitrary shape clusters recognized by DBSCAN

The algorithm's theory is that each cluster has a normal density of points (objects) that is significantly higher than the density outside the cluster, as well as a density in areas of noise that is lower than all of the clusters' density. More specifically, each point of the cluster must have at least a certain number of points in its neighborhood of a given radius, which is determined by a limit value.

This algorithm investigates the cluster for given parameter values as follows: first it selects as a seed a random point which is a nucleus, then places in the cluster the bait and all points densely reachable from it.

The EM algorithm is based on the calculation of distances, ie the identification of areas that are more "populated" than others. In the process of the algorithm there is an iterative improvement of the solution, and stopping is carried out at the moment when the required level of accuracy of the model is reached.

The basis of the EM algorithm is the assumption that the investigated set of data can be modeled using a linear combination of multidimensional normal distributions. It is assumed that the data in each cluster is subject to a certain distribution law, namely, the normal distribution.

The EM algorithm is an iterative algorithm, each iteration consists of two steps: a step of mathematical expectation (E-step) and a maximization step (M-step) [4].

The E-step calculates the expected value of the likelihood function, with hidden variables being considered as observable. In the M-step, the maximum likelihood estimate is calculated, thus increasing the expected likelihood calculated in the E-step. This value is then used for the E-step in the next iteration. The algorithm runs to convergence. Here are the steps from a mathematical point of view. To do this, consider the function:

$$F(q,\theta) = E_q[logL(\theta;x,Z)] + H(q) = -D_{KL}\left(q \parallel p_{Z|X}(\cdot\,|x;\theta)\right) + logL(\theta;x), \quad (1)$$

where q is the probability distribution of the unobserved variables Z, $p_{Z|X}(\cdot\,|x;\theta)$ – conditional distribution of unobservable variables at fixed observable x and parameters θ, H – entropy, D_{KL} – distance from Kulbak-Leibler.

Then the steps of the EM algorithm can be represented as:

a) E (xpectation) step: Select q to maximize F:

$$q^{(t)} =* arg\ max_q F(q,\theta^{(t)});$$

b) M (aximization) step: We choose θ to maximize F:

$$q^{(t+1)} =* arg\ max_\theta F(q^{(t)},\theta).$$

The X-Means algorithm is one of the most popular clustering methods. It is also a generalization of the k-means algorithm and uses it in its implementation. One of the main differences of this algorithm can be called the absence of the requirement of the exact number of required clusters, only the required range of values for the number of clusters is specified.

The basic idea of the algorithm is that at each iteration the center of mass for each cluster obtained in the previous step is recalculated, then the vectors are broken down into clusters again according to which of the new centers turned out to be closer to the chosen metric.

When no iteration shifts the clusters' center of mass, the algorithm stops. This occurs for a finite number of iterations since the number of possible partitions of the finite set is normally finite, and the cumulative square deviation of V does not increase with each step, making looping impossible.

This algorithm uses the Bayesian model selection criterion [6]. It follows from the principle of maximum likelihood. This criterion is determined by the formula:

$$BIC\ =\ -2ln(L) + kln(n)\,, \quad (2)$$

where L is the maximum value of the likelihood function of the observed sample with a known number of parameters, k is the number of (estimated) parameters used, n is the number of objects in the sample.

Hierarchical clustering is a set of algorithms that use the division of large clusters into smaller clusters or the merging of smaller clusters into larger ones. Accordingly, allocate divided and agglomerative clustering. In this work, the agglomerative Lance-Williams clustering was used.

To calculate the distance R (W, S) between clusters W = U ∪ V and S, knowing the distances R (U, S), R (V, S), R (U, V), we need to use a formula that generalizes most reasonable ways:

$$R(U \cup V,S) = \alpha_U \times R(U,S)\ + +\alpha_V \times R(V,S) + \beta \times \quad (3)$$
$$\times R(U,V)\ + +\gamma \times |R(U,S) - R(V,S)|\ ,$$

where αU, αV, β, γ - numerical parameters.

The COBWEB algorithm is a classic incremental conceptual clustering method that defines clusters as groups of objects that belong to one concept - a specific set of attribute-value pairs. It creates a hierarchical clustering in the form of a tree: each node of this tree references a concept and contains a distribution of all the descriptions of that concept, which includes the probability of the concept's belonging to a given node and the conditional probabilities of the species:

$$CU\ = \sum\sum\sum P(A\ =\ U\_ij\ |\ C\_k)P(C\_k\ |\ A\ =\ U\_ij)P(A\ =\ U\_ij)\ k\_ij. \quad (4)$$

The values are summed across all Ck categories, all Aj properties, and all Uij property values. The value of P (Aj = Uij | Ck) is called predictability. This is the probability that the object for which the

property Aj takes the value Uij belongs to the category Ck. The higher the value, the more likely the properties of two objects in the same category have the same values. The value of P (Ck | A = Uij) is called predictiveness. This is the probability that for Ck objects, the Aj property takes the value Uij. The greater the value, the less likely it is for objects that do not belong to this category to take the specified value.

The value of P (A = Uij) is a weighting factor that enhances the influence of the most common properties. By sharing these values together, the high utility of a category means a high probability that objects in one category have the same properties, and a low likelihood of having those qualities in objects in other categories.

The algorithm for constructing a tree uses a heuristic measure of estimation, called category utility - an increase in the expected number of correct assumptions about the value of attributes while knowing their belonging to a certain category relative to the expected number of correct assumptions about the value of attributes without this knowledge. To embed a new object in a tree, the COBWEB algorithm iteratively scans the entire tree in search of the "best" node to which that object is assigned.

Selecting a node is based on the placement of the object in each node and calculating the usefulness of the category of the slice obtained. It also calculates the usefulness of a category for the case when an object belongs to a newly created node. As a result, the object refers to a node for which the usefulness of the category is greater.

As a result of the study of existing clustering algorithms, the EM algorithm was selected for the experiment.

3. Choosing software

One of the objectives of the study is the choice of software for the process of clustering and subsequent visualization of the results.

For this purpose, a great deal of work was done to search for existing statistical packages. As you can see, all existing programs can be divided into three main categories: private research, implemented using popular software math packages, expensive commercial solutions focused on corporate statistical research, and a small proportion of statistical packages that are freely available. In order to select the mathematical package for the study, a number of existing statistical processing tools were considered.

The Fuzzy Clustering and Data Analysis Toolbox - software package for Matlab provides three categories of functions:

- clustering algorithms that break data into clusters with different approaches: K- means and K-medoid - algorithms for stable clustering; FCMclust, GKclust and GGclust unstable clustering algorithms;
- analysis functions that evaluate each fixed partition performed by an algorithm based on indices (Xie and Beni's, Dunn, Alternative Dunn, Partition index);
- visualization features that implement Sammon's modified method of displaying data in a smaller space.

This program is installed as a plug-in, does not provide a ready-made interface for analysis, but allows you to further use the functions described above when developing applications on Matlab [7].

Cluster Validity Analysis Platform (CVAP) is a software tool, also implemented on Matlab. Based on a user-friendly graphical interface, it includes several algorithms for cluster analysis (K-means, hierarchical, SOM, PAM), as well as the most widely used indexes of their performance. By working in this application, the user is not only able to download their data, but also save the results of work. The undoubted advantage is that the graphical part allows to analyze several algorithms for one index at a time.

SPSS Statistics is a paid modular, fully integrated software package that covers all stages of the analytical process, focused on solving business problems and related research problems. The intuitive interface has many statistics management features. It has clustering algorithms [8].

RapidMiner is a machine learning and data processing environment that protects the user from the grunt work. Instead, he is asked to "draw" the entire desired data processing method as a chain (graph) of operators and then execute it. RapidMiner displays the operator chain as an interactive graph and an XML expression (the main language of the system).

Now more than 400 operators are implemented in the system. Of them:

- operators of precedent training, which implements clustering, classification, regression and association search algorithms;
- pre-processing operators (filtering, sampling, filling in the gaps, reducing the dimension;
- operators of work with signs (selection and generation of signs);
- meta-operators (for example, multi-parameter optimization operator);
- operators of quality assessment (sliding control);
- visualization operators;
- data downloading and storage operators (including working with special formats: arff, C4.5, csv, bibtex, databases, etc.).

WEKA is written in Java at the University of Waikato (New Zealand) and provides the user with the ability to pre-process data, solve clustering, classify, regress and search for associative rules, as well as visualize data and results [9]. The program is very easy to learn (probably has the most intuitive interface among all programs of this type), is free and can be supplemented with new means of pre-processing and data visualization.

The output can be represented as a matrix of feature descriptions. WEKA provides access to SQL databases through Java Database Connectivity (JDBC) and can accept SQL query results as output.

WEKA has the Explorer UI, but the same functionality is available through the Knowledge Flow Component Interface and from the command line [10]. There is a separate Experimenter application to compare the very root of the ability of machine learning algorithms on a given set of tasks.

Explorer has several panels:

- Preprocess panel allows you to import data from a database, CSV file, etc., and apply filtering algorithms to them, for example, translate quantitative characters into discrete ones, delete objects and features by a given criterion;
- Classify panel allows you to apply;
- Classification and regression algorithms for data sampling, estimating the predicted ability of algorithms, visualizing erroneous predictions, ROC curves, and the algorithm itself, if possible (in particular, decision trees);
- The Associate panel search bar is concerned with identifying any meaningful relationships between features;
- Cluster panel Cluster panel gives access to K-Means algorithm, EM algorithm, COBWEB, DBSCAN and others;
- Select attributes panel gives access to feature selection methods;
- Visualize visualization panel builds scatter plot matrix, allows you to select and enlarge graphs, etc.

The disadvantage of The Fuzzy Clustering and Data Analysis Toolbox and CVAP lies primarily in their inaccessibility and inability to analyze their own algorithms. These non-commercial applications are implemented mainly on Matlab, which automatically imposes a number of restrictions:

- applications depend on version and additional libraries supplied;
- it is necessary to know its internal structure and rules of operation;
- graphical and computational implementations are fixed;
- analysis of our own algorithms, if possible, it is necessary to create additional systems of interaction.

CVAP is only supported by the Matlab application, despite its user-friendly graphical interface [11]. In order to use The Fuzzy Clustering and Data Analysis Toolbox, you need to write additional functions that relate the algorithm and analysis features.

In most cases, all such programs are personal research, often designed to demonstrate specific methods and are therefore limited in functionality.

Matlab itself is a commercial product that needs to be purchased and installed, which in itself is a long, time-consuming process. Participation in commercial projects, including SPPP Statistics, in turn are successfully developed, but since they are focused mainly on statistical surveys in business, they include clustering algorithms as part of statistical methods, so specialized tools for analyzing the work of the algorithms themselves, such as usually do not have. The cost of such developments is quite large. For example, the licensed program SPSS Statistics for one private user is currently worth about forty

thousand rubles. In addition, the implementation and analysis of their algorithms in such programs is not provided.

Free software complexes (RapidMiner, WEKA) also impose a number of restrictions on data processing [12]. These programs do not have the ability to embed their own algorithms, and the number and variety of existing clustering algorithms is also negligible.

RapidMiner has very good rendering tools: there are many rendering methods and all the graphics look great. But the only downside to rejecting this software is the lack of connection to the FireBird database and the lack of algorithms selected at the beginning of the study.

Thus, after examining several statistical packages, WEKA software package was selected for further clustering process, which contains the selected algorithms and has the possibility to connect to the database via URL. In addition, WEKA is one of the few products that has an intuitive interface and translated technical literature.

4. Selection of objects for clustering of data of the remote workshop and definition of their signs

Based on the analysis of the conceptual scheme of the database checks the systems of the remote workshop, the main essences were identified - these are the tasks, users and solutions. Therefore, it was decided to select the following clustering objects:

- students (workshop users);
- tasks of the workshop;
- pairs "student - task".

In order to determine the set of features for clustering, the attributes of the selected entities available in the database were investigated. The following attributes are stored for tasks in the database:

- task identifier;
- the limit of CPU time and operational of this task;
- the minimum percentage of unique code at which the solution of the problem is considered unique;
- expert complexity of the task;
- the number of users who have solved this task;
- the number of users who tried to solve this problem;
- the number of decisions received for this task.

Each user of the system in the database is assigned an ID, login and password to log in, also stored calculated data, the number of tasks solved by the user and the number of tasks that the user tried to solve. Each attempt to solve the problem by the student is recorded in the database, while storing the following information:

- student ID and task number;
- date and time of receipt of the solution of the problem by the inspection system;
- used compiler;
- attempt status (correct decision or error code);
- characteristics of the correct decision - the execution time of the program (query, script), the amount of memory used, the percentage of plagiarism.

After analyzing the attributes of clustering objects stored in the database and selecting the most significant features, a set of features for clustering was determined for each stage of the study.

To cluster users of the verification system, the following attributes were selected, which will allow to select groups of students by level of training:

- user ID;
- relative indicator of the student's level of preparation;
- the average number of attempts to solve problems;
- the average complexity of the tasks;
- year of study (1-5, students of previous years are considered as one course "-1").

To cluster the tasks of the testing system, the following attributes were selected, which will allow to select groups of tasks by level of complexity:

- task identifier;
- a relative indicator of the degree of complexity of the task;
- the number of non-unique solutions;
- the number of partially correct decisions;
- the average number of attempts to solve the problem.

For clustering of pairs "student-task" the following attributes were chosen, which will allow to determine a suitable student task or not:

- complexity of the task;
- relative indicator of the degree of complexity of the task;
- relative indicator of the student's level of preparation; the average complexity of the tasks solved by the student;
- the relative indicator with which the problem is solved;
- the number of days between the first and last attempt to resolve.

5. Conclusions

We should draw a number of conclusions from this research that will assist teachers in improving both the learning process and the teaching process. It's worth noting that the Collaborative and Storage resources are the most widely used, since they provide a vast volume of data when processed by both teachers and students. To optimize efficiency and receive new developments and enhancement of learning processes, this data must be intelligently ordered. Furthermore, we can infer from the usage of the Evaluation instruments that teachers need to find alternative ways of assessment in order to improve access to these tools.

This portion of the visualization is built on two attributes: SEM / GRANDTOTAL, as well as the Y axis and the X axis corresponding to the outcome. However, the other two properties can only be used to create one aspect of the visualization.

The visualization diagram (Figure 2) is shown together with the details of each cluster.

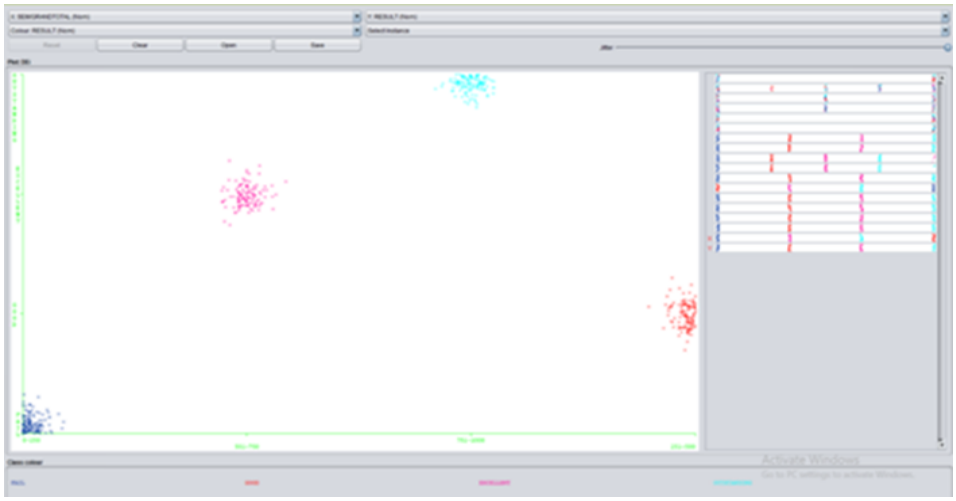

Figure 2: Visualize a chart between two attributes

As we can see in Figure 3, most boys were absent from the two tests and therefore had poor grades in all subjects. Most of the girls with average grades who went to the first test but missed the next one also got into the cluster of section B.

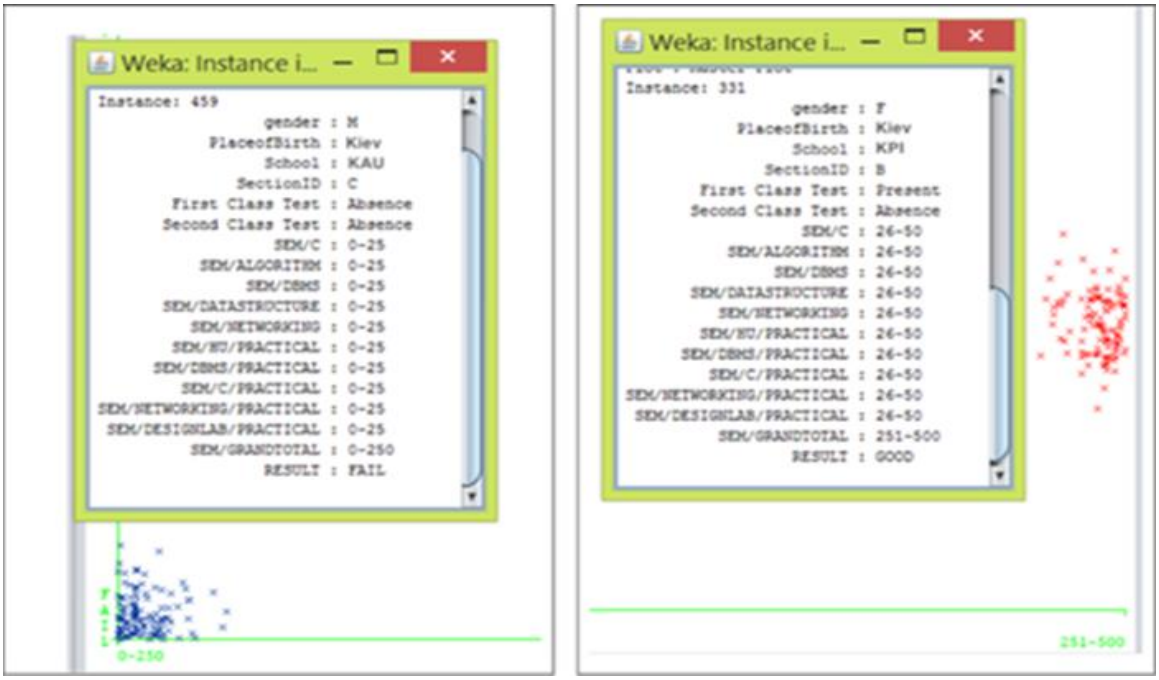

Figure 3: Visualization of clustering result for unsatisfactory and satisfactory results

Figure 4 shows that most guys had excellent grades in subjects and wrote the second test.

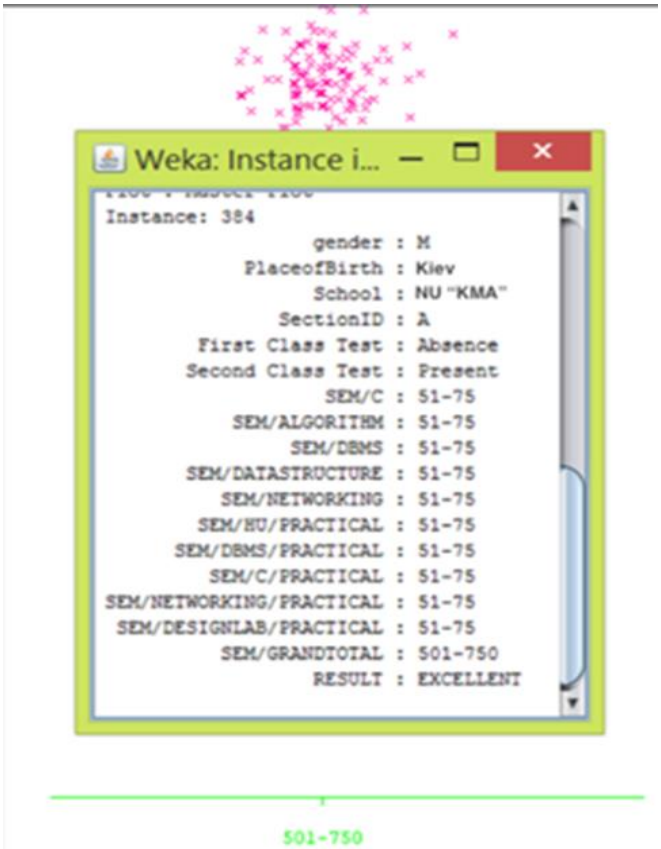

Figure 4: Visualize the result of clustering for great results

Data mining in an instructional setting is presented in this article, which uses associative rule extraction strategies to identify patterns of student failure. To examine pupil performance, association law research was extended to educational systems. The Association Rules extraction methodology is

used in this research to uncover elusive dynamics and assess student success and trends. To find connections between attributes, the EM algorithm is used.

Student success was measured using academic and personal data gathered over the course of one semester. After that, J48 classification algorithms were used. WEKA 3.8.2 was the data processing software used in the trial. We can infer that the J48 classification system was the most suitable algorithm for the data set based on the accuracy and classification errors.

WEKA was used to apply the EM algorithm to the dataset in order to find an interpretation of average student success based on some of the best rules. Data can be extended to include any of a student's extracurricular activities and technical abilities, and various classification algorithms can be used to forecast student success.

6. References

[1] Learning management system. Wikipedia, 2021. URL: https://en.wikipedia.org/wiki/ Learning_ management_system.

[2] K. Cook, K. Kukier, V. Shteinberg, Big Data: A Revolution That Will Transform How We Live, Work, and Think, 2013.

[3] R. Baker, G. Siemens, Educational data mining and learning analytics, The Cambridge handbook of the learning sciences, 2014.

[4] Expectation-maximization algorithm. Wikipedia, 2020. URL: https://en.wikipedia.org/wiki/ Expectation%E2%80%93maximization_algorithm.

[5] S. Rzheutsk, Experience in the application of clustering methods for analyzing the results of distance learning, in: Proceedings of the International Scientific and Practical Conference, Informatization of Engineering Education, 56, 2016, pp. 617–620.

[6] K-means clustering. Wikipedia. 2019. URL: https://en.wikipedia.org/wiki/K-means_clustering.

[7] O. Dashkevich, Analysis of Apache Kafka capabilities within the framework of secure Big Data streaming, in: Proceedings of the 7th. International scientific and technical conference, Information Systems and Technologies, 12, 2018, pp. 34–35.

[8] C. Lam, Hadoop in Action, 2010.

[9] Data Mining with WEKA MOOC – Material, Machine Learning at Waikato University, 2019. URL: https://www.cs.waikato.ac.nz/ml/WEKA/mooc/dataminingwithWEKA.

[10] Weka Tutorial, Tutorials Point, 2018. URL: https://www.tutorialspoint.com/weka/weka_ quick_guide.htm.

[11] K. Smelyakov, M. Shupyliuk, V. Martovytskyi, D. Tovchyrechko, O. Ponomarenko, Efficiency of Image Convolution, in: Proceedings of the 8th IEEE International Conference on Advanced Optoelectronics and Lasers, CAOL'2019, Sozopol Bulgaria. 2019, pp. 578-583.

[12] K. Smelyakov, O. Ponomarenko, A. Chupryna, D. Tovchyrechko, I. Ruban, Local Feature Detectors Performance Analysis on Digital Image, in: Proceedings of the IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology, PIC S&T'2019, Kyiv Ukraine, 2019, pp. 644-648.