An application of analytical data research in e-learning system

Daniela Orozova¹ and Nely Hristova²

¹Department of Computer Science, Burgas Free University, Burgas, Bulgaria
²Department of Mathematics, Valencia College, Orlando, Florida, USA

e-mails: orozova@bfu.bg, nhristova@valenciacollege.edu

Abstract. Algorithms for extracting information from big data today are very useful in e-learning systems. The main functions of these algorithms are the application of various methods to detect and retrieve models of stored data. Different machine learning techniques are used and rules for association, classifications, clusters, etc. are derived. Open knowledge is used to make predictions about: admission of students in a particular course, forecasting grades in online tests, finding inaccurate grades, predictions for graduation of students and more.

1. Introduction
With the growing access to the Internet, the number of users of e-learning systems is multiplying. This in turn leads to a greater amount of accumulated data in these systems, which must be systematized in an appropriate way in order to be used in training. This makes e-courses more difficult to configure and maintain. Users need to be classified in order to be provided with relevant information and their training to be adequately assessed.

In the website www.educationaldatamining.org data extraction in training systems is defined as a new discipline that deals with the development of methods for researching types of data in the fields of training and the use of these methods to achieve better understanding of the learners and the ways in which they are trained. Learner models in modern systems assess not only whether a student has a skill or understands a concept, but also store data on the motivation of the learner and a number of cognitive and behavioural characteristics.

In education, websites containing a large amount of data such as YouTube, Facebook, Tumblr, Flickr and Instagram can help learners. Using web blogs to write magazines instead of the standard way is another approach to developing skills among learners, which makes the learning process more interesting [1]. Graphics, maps, tables, photos and other multimedia information that learners use to draw conclusions and reasoning can be used in the training. Solving such tasks helps learners to develop data skills as well as to learn critical thinking [5].

The data, accumulated from the work in the educational environments, are constantly increasing. As a result, the data analysis goals can vary: monitoring, forecasting, individualization, intervention in the learning process, assessment and recommendations of the learner and feedback, etc. The aim of the present study is to show approaches to use of data analytics in the field of educational space.

Popular means of extracting knowledge that can be used for data analysis are, for example, systems:

- WEKA – Waikato Environment for Knowledge Analysis is a system developed at the University of Waikato in New Zealand. The software is available at http://www.cs.waikato.ac.nz/ml/weka for free use, licensed under the GNU General Public License. The system is written in Java. WEKA contains modules for data visualization (49 tools), 76 algorithms for classification and regression, 8 algorithms for clustering and others. [11].
RapidMiner is a software platform that provides an integrated environment for machine learning, data retrieval, text retrieval, forecasting and business analysis, results visualization, validation and optimization. RapidMiner uses a client/server model, with the server being offered by software as a service in cloud infrastructures. RapidMiner is written in Java and provides a graphical user interface for designing and executing workflows. They consist of multiple operators, each operator performing a single task within the process and an input for the next operator being formed at the output. [10].

- KNIME is an open source data analysis platform. Integrates various machine learning and data retrieval components, allowing users to visually create streams (or pipelines). A graphical user interface allows the installation of data processing units (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization. Since 2006, KNIME has been used in pharmaceutical research as well as in other areas such as CRM (customer relationship software), customer data analysis, financial data analysis and training systems. KNIME is software written in Java. The basic version includes hundreds of modules for data integration (file input and output, database connections), data conversion (filter, converter, combiner), as well as the most commonly used data analysis and visualization methods [6].

The integration of the education spaces with Data Mining tools is necessary for the process of personalization of courses for electronic learning. On the basis of the obtained results, additional measures for analysis and change of learning courses can be introduced. This, on its part, is a path towards the increasing of the quality of learning.

2. Experiment for assessment of learners

Different forms and methods of testing are used to assess students’ knowledge, such as: online or offline tests, answers to open-ended questions, code testing and reading, problem-solving, etc. [3,4]. Depending on the purpose and application, there are different classifications of approaches for assessing knowledge [8,9]. Problems related to predictive modeling for student recruitment are discussed in [1]. Multi-criteria models in cluster design may prove useful/applicable when dividing students into groups for the purpose of their evaluation [13]. Coefficients on the importance of experts’ opinion, may also be taken into account in determining the final decision [5].

The data presented in this paper is based on studies, experiments and analytics taken from the course: “Mathematics”- led by Prof. Hristova at Valencia College, Orlando, Florida, United States of America. First experiments were made by a small data sample limited by initially available real data for students (365 records) for two years.

The final assessment of the discipline is often complex and includes several components involved in its formation with different weights. It must reflect the different aspects of the training of trainees (theoretical knowledge or practical skills), be consistent with the specifics of the assessed, with the age and individual characteristics of the assessed, etc. For the purposes of the analysis, we consider a specific approach to student assessment, in which the grade at the end of the semester is defined as the weighted mean ($\bar{x}$) of all grades during the semester. Each assessment has a weight. The function we use has the form:

$$\bar{x} = \frac{\sum wx}{\sum w}$$

where $x$ is the grade, $w$ is the weight of each grade.

Students’ grades in Mathematics at the end of the semester are determined on the basis of the following tests: Test 1 Sets, Test 2 Logic, Test 3 Geometry, Test 4 Statistics, Test 5 Probability, Final Exam, Quizzes, Homework, Projects and Class Activities. Each criterion has a degree of importance such as: The weight of Test 1 Sets is 9%, the weight of Test 2 Logic is 9%, the weight of Test 3 Geometry is 9%, the weight of Test 4 Statistics is 9% and the weight of Test 5 Probability is also 9% of the overall grade at the end of the semester. The weight of Final Exam is 15%, the weight of Quizzes is 15% and the weight of Homework is also 15% of the grade of the grade at the end of the semester. The weight of Projects is 5% of the grade and the weight of Class Activities category is 5% of the grade at the end of the semester. Tests and the Final exam have a total weight of 60%, which is crucial for the assessment. We combine Homework, Quizzes, Projects, and Class Activities into a category called Activity, which determine 40% of the grade. The function we use is:
\[ \bar{x} = \frac{\sum T_1 \times 0.09 + T_2 \times 0.09 + T_3 \times 0.09 + T_4 \times 0.09 + T_5 \times 0.09 + \text{Fin.Ex} \times 0.15 + \text{Act} \times 0.40}{\sum w} \]

where \( \bar{x} \) is the grade at the end of the semester, \( w \) is the weight of each grade. We consider how the grade obtained for the individual tests affects the assessment and how the activities in class and out of class affect the assessment at the end of the semester.

In addition to the points accumulated by the individual components, we also store data for each student: Age, Gender, New Student Experience (NSE), Full_Time/Part_Time, Student Program type of training, Times Taken Course.

3. Using Excel for analytical research of student data
Excel offers effective functions for summarizing data trends. Filters provide options for selecting records. Conditional formatting colors the data that meet certain criteria and helps us detect deviations and trends in the data. Charts visually show deviations and trends. Pivot tables make it easier for users to view data.

![Figure 1](image-url)  
**Figure 1.** Graphs on the average grades for Test 2, Test 3, Activity Average and Final Exam.

The large number of ready-made functions in Excel and the possibilities for creating formulas allow analyzes on the data such as: finding a linear function showing a trend and finding a correlation coefficient. Macros act as a set of functions and commands that we can use to automate tasks. They are created with the help of Visual Basic for Applications (VBA) and their capabilities are very large.

In the following graphs (Fig. 1) it can be seen that the average scores for Test 2 Logic, Test 3 Geometry, Activity Average and Final Exam have increased from Term 2016 to Term 2018. Creating line diagrams for these four data groups, we find a linear equation for them and the value of \( R^2 \), where \( R \) is the correlation coefficient. Using macros, we can create line graphs for the four groups on the average grade for each semester.

For the distribution of the ratings for term 2017 we can say that there is a big variation in the ratings of Test 1 Sets and Test 5 Probability. On these tests, half of the trainees performed worse than 25% of the trainees in all other groups. During this semester there is a small number of students whose
success is much lower than the success of the others. There is an improvement in the success of Test 2 Logic.

The following graph (Fig. 2) shows that there is no big difference in the assessments of students on all tests in term 201810. This shows a good absorption of the learning material. The lowest scores are in Test 4 Statistics, where 75% of students have scores lower than 50% of students in Test 1 Sets, Test 2 Logic, Test 3 Geometry and Test 5 Probability.

![Figure 2. Graph on the distribution of grades by modules.](image)

Excel is used successfully to prepare data, to visualize results through graphs, and to find linear functions that show trends in the distribution of scores across categories. Excel’s statistical capabilities make it possible to determine students whose grades are much lower than the majority, what is the minimum and maximum grade, and more. The use of macros allows fast and multiple processing of groups of data. In the experiment, we continue to work on the data through the Orange system, because the system’s tools allow us to create models on the data to be used to predict new data.

4. **Orange Data Mining System for create models and predict new data for students**

*Orange* is open source software for machine learning and data retrieval written in Python. The program is maintained and developed by the Bioinformatics Laboratory of the Faculty of Computer and Information Sciences at the University of Ljubljana [7]. Orange is supported on OS X, Windows and Linux. Applying the Orange system for data analysis of the described experiment, we can look for dependencies between the individual evaluation components, between theoretical and practical knowledge and skills. The main input data are the points obtained from the assessment components, and the outputs are the respective final grades set by the teacher.

![Figure 3. Workflow for creating models.](image)
To solve the classification problem, we apply several different techniques through the Orange Data Mining System application. Through the tool File we load the data for students’ grades on the various assessment components in number of points. The data is pre-cleared and in “Orange” format. We consistently create models using the tools of the system: Tree, Random Forest, Logistic Regression, Naïve Bayes, SVM (Support Vector Machines) and Neural Network.

The created models can be used to predict new data. For this purpose we prepare a table with the same structure as the initial data table, but the column with the final score is not set. The result of the operation of the Test & Score tool is a table with evaluations on efficiency, precision, Recall (sensitivity) and F1 Score for the created models, shown in Fig. 4.

![Figure 4. Result of the work of the Test&Score tool on the created models.](image)

Summarizing the results on the average score for the forecast accuracy on the rating classes (Poor, Medium, Good, Very good, Excellent) we can choose a forecast model suitable for the specific area [9].

Analyzing the data accumulated from eLearning courses makes it possible for the student assessment model to be innovated, and educators to further design modules that meet students’ individual needs.

5. Conclusion

Based on the accumulated data from the work of an e-learning system with different users, applying tools in the field of data science, different decisions can be made about training [8]. The analysis of the data accumulated in the e-learning courses makes it possible to change the test model and to design modules that meet the individual needs of the learner to search for the necessary information. The interest of the learners, their active participation in the process of acquiring knowledge and acquiring skills can be greatly influenced by the quality of the learning environment used.

Acknowledgments

The work is funded by the SRF of the Burgas Free University as a part of project “Data Science in the Learning Space for a Blue Career” D-9/4.03.2020.

References

[1] Egger N et al 2014 Themes in Science & Technology Education vol 7 issue 1 3-17
[2] Etheridge D 2010 Excel Programming: Your Visual Blueprint for Creating Interactive Spreadsheets, John Willey & Sons, Inc.
[3] Hadzhikolev E, Yotov K, Trankov M and Hadzhikoleva S 2019 Use of Neural Networks in Assessing Knowledge and Skills of University Students, Proceedings of ICERI2019 Conference Seville, Spain pp 7474-7484
[4] Hadzhikoleva S, Hadzhikolev E and Kasakliev N 2019 Using Peer Assessment to Enhance Higher Order Thinking Skills, Tem Journal vol 8 no 1 pp 242-247 http://www.temjournal.com/content/81/TEMJOURNALFebruary2019_242_247.pdf
[5] Kastens K and Turrin M 2010 Earth Science Puzzles: Making Meaning from Data, National Science Teachers Association
[6] KNIME – Open for Innovation [Oonline] https://www.knime.com/, last access: September 24 2020
[7] Orange system [Online] https://orange.biolab.si/training/introduction-to-data-mining/ last access: September 25 2020
[8] Popchev I and Orozova D 2020 Towards a multistep method for assessment in e-learning of emerging technologies Cybernetics and Information Technologies vol 20 no 3 116-129
[9] Popchev I and Orozova D 2019 Towards Big Data Analytics in the E-learning Space Cybernetics and Information Technologies vol 19 (3) pp 16-25
[10] RapidMiner - Depth for Data Scientists https://rapidminer.com/ last access: November 2 2020
[11] Waikato Environment for Knowledge Analysis http://www.cs.waikato.ac.nz/ml/weka last access: September 4 2020
[12] Walkenbach J 2016 Microsoft Excel Bible.: Wiley, Indianapolis, IN
[13] Radeva I 2013 Multi-Criteria Models for Clusters Design Cybernetics and Information Technologies vol 13 (1) pp 18-33