Mining Student Data Using Clustering Expectation-Maximization Algorithm

Gabrijela Dimic[1], Petar Spalevic[2] and Kristijan Kuk[3]

Abstract – **In this paper data mining technique named clustering is applied to analyze student's learning behavior u e-learning system Moodle. Study showed what kind of data can be collected, how to perform the previous data preparation, how to apply appropriate methods on the data, and how to apply the discovered knowledge.**

Keywords – **Data mining, Educational data mining, Clustering, Moodle, E-learning**

I. Introduction

Over the past years, several thousand e-learning systems have been developed, which represents a basis of the modern learning technology concept [1].

E-learning systems enables autoidentification of the user, access and use of learning material, simple editing and layout of documents, leading of discussions and communication with other participants of the course, carrying out of exercises, testing and surveying of users, grades recording etc. These systems provide database in which all system information on users is being recorded.

The analysis of activity data and user interaction e-learning system can be measured its effectiveness and then used as a support for decision making in the organization of online courses for distance learning. Majority of analysis and evaluation of e-learning system based on the use of questionnaires and surveys that are widely used to assess the impact and usability of interactive systems [2].

An alternative to traditional data analysis is the use of data mining as an inductive approach for automatic detection of hidden information present in the data. Data mining, unlike the traditional analysis, is based on discovering knowledge in the sense that an assumption is automatically derived from data, i.e. that it is driven by data rather than based on research or driven by a man.

Data mining is a process of extracting samples or models from observed data. It is defined as a process of identifying valid, new, potentially useful and understandable patterns in data [3]. Data mining techniques can be applied to a wide range of data. It is often used as a method for discovering knowledge on commercial websites for identification of key buyers and increasing efficiency of online sale on commercial websites.

These aspects can be transmitted to the field of knowledge and management systems. The use of data mining in education implies integration of data discovery processes and methods for knowledge discovery in educational environment, i.e. development of methods for extracting a unique type of data coming from the educational environment, and the use of those methods in order to better understand students. This new developing field, known as Educational Data Mining, refers to selection of best methods for discovery of knowledge on the basis of data available in the educational environment.

The data needed for research in this field are collected from operational data that are kept in databases of educational institutions and represent personal and academic data on students, and from the systems that support electronic learning and have a huge quantity of information.

On the basis of discovered knowledge an analytical model is constructed, which reveals interesting patterns and guidelines based on student user information that the teacher uses to improve the course performance and efficiency, and thus students' learning, their results and final marks as well. The use of data mining in e-learning systems is a constant cyclical process in which the discovered knowledge should realize and record the cyclical loop of the system, and on the basis of that state, enable and improve learning as a whole, not only by putting data into knowlegde but also by filtering the discovered knowlegde for making a final decision.

II. Case Study

For the purpose of this study we collected and analyzed students data in the course Computer graphics, held in the summer semester of the school year 2009/2010 at the College of Electrical Engineering and Computer Science Applied Studies in Belgrade. The course was chosen by 130 students from different study programs. Students were not obliged to attend the lectures, and the lecture material was also available in the Moodle course. Before every laboratory exercise, students took a knowledge check test referring to the current exercise in Moodle, which contained questions from the theoretical part needed for performing tasks from the exercises.

[1] Gabrijela Dimic is with the College of Electrical Engineering and Computer Science Applied Studies, Vojvode Stepe 283, 11000 Belgrade,Serbia, E-mail: gdimic@viser.edu.rs.
[2] Petar Spalevic is with the Faculty of Technical Sciences, Kneza Miloša 7, 38000 Kosovska Mitovica, Serbia,
E-mail: petarspalevic@yahoo.com
[3] Kristijan Kuk is with the College of Electrical Engineering and Computer Science Applied Studies, Vojvode Stepe 283, 11000 Belgrade,Serbia, E-mail: kkristijan@viser.edu.rs.

The course user environment was organized in the following way:

- Forum: Notifications by which students were delivered information related to lectures by the course teacher.
- Learning material: Lessons, Self-check tests, Interactive tasks.
- Knowledge check tests for each laboratory exercise.
- Final knowledge check tests.

The main objective of this study was grouping students into clusters according to their characteristics and analyzing of their activities. In order to perform the necessary research, was identified set of basic questions for the analysis course e-learning system:

- What materials do they access most often?
- Is there a connection between activities and results?
- How do results of data mining methods correspond to final marks?
- How to form the groups – on the basis of what characteristics and activities of students?
- What are the main characteristics within a group, and what are differences between them?

III. Application of Data Mining Methods

A data mining process comprises the following stages [4]:

- collection of data to be analyzed,
- preparation and pre-procession of data,
- usage of data mining algorithms,
- interpretation and evaluation of results.

A.Data collection

For this study, we collected the information activities of students in the Computer Graphics course from Moodle database. We collected information on students access to the course and excluded the students who didn’t access the course at all, and thus the number of analyzed students decreased from 130 to 123. Data were collected about usage of learning material, results of the tests for individual knowledge evaluation, reading of messages sent via forums.

B. Data preprocessing and discretization

This phase implies gathering of relevant and exlusion of irrelevant data, transformation of original data to an adequate form for implementation data mining algorithm and discretization and adjustment of data in regard to usage of data mining algorithm. All information on access to the Moodle system are recorderd into database and log file on the server where Moodle is installed. Anyway, the database is much more reliable and powerful data source about all activities in comparison to log file and data gathered from the e-learning system database require less clean up and less processing.

In the preprocessing phase, data are prepared for the application of data mining techniques. This phase implies previous procession and cleaning of data, removal of inapplicable and unimportant information, collection of information necessary for modelling, selection of methods for handling the missing data fields, and finally discretization.

Atributes are determined in a way to integrate information regarding all student activities at the analyzed course and namely number of lessons read, number of interactive tasks performed, number of read forum posts, number of self-check tests, an average result achieved on self-check for all try outs and all of these in a single record.

By the application of discretization method [5] numerical values of an atribute have been transformed into discrete thus becoming more comprehansive and clearer for further analysis. Analysis report about all student's activities was showed key parameters on which the new table is created with the appropriate attributes (see Table I).

TABLE I
USED ATTRIBUTES FOR EACH INSTANCE OF STUDENT

n_lessons	Number of read lessons	Low/ Medium/High
n_int_tasks	Number of solved interactive tasks	Low/ Medium/High
n_read_forum_post	Number of read forum posts	Low/ Medium/High
n_self_check_tests	Number of taken self-check tests	Low/ Medium/High
mark_ self_check_tests	An average result achieved on tests	Failed/ Passed
mark_exam	Mark obtained in the exam	Failed/ Passed/ Excellent

Percentage of usage of learning material has been established (lessons, interactive tasks, read forum posts, self_check_tests) and labeled like this:

- Usage of up to 25% – Low
- Usage of up to 55% – Medium
- Usage of up to 100% – High

After the use of dicretization filter, the resulting file with data has all nominal attributes, and it is transformed into a text file of the ARFF format (Attribute-Relation File Format), which is an ASCII text file and describes the list of instances assigned within the collection of attributes [4] and is ready for the use of data mining methods.

C. Clustering

Clustering is one of the basic techniques often used in analyzing data sets. The main task of this method is to put items into groups so that the similarity of items within a group is maximized, while similarity of items in various groups is minimized [4]. With the use of clustering in e-learning systems the implemented algorithms can find clusters of students with similar learning characteristics, group students in order to obtain various environments based on their capabilities and other characteristics [6].

There are a lot of algorithms for generating clustering methods and in this case study was applied Expectation-Maximization (EM) algorithm . The EM algorithm is a mixture of the basic algorithm that finds maximum probabilities of parameter evaluation in a possible model [7].

This method is an efficient iterative procedure to compute the Maximum Likelihood estimate in the presence of missing or hidden data [8]. Generally, this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps: first, the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fixpoint or until the log-likelihood function, which measures the quality of clustering, reaches its maximum.

In this study, we used data mining EM algorithm for predicting the final grade on the basis of analyzing the behavior of students in the course. Students were grouped into three clusters according to their learning activities. The algorithm assigned to attributes the maximum probability of affiliation with each cluster (see Table II).

TABLE II
RESULTS OF GROUPING WITH THE USE OF EM ALGORITHM

Attribute	Cluster 1	Cluster 2	Cluster 3
n_lessons			
Low	1.2397	6.9088	**21.8515**
Medium	9.232	**25.5078**	1.26
High	**33.2252**	7.7296	1.0552
n_int_tasks			
Low	2.0727	**19.8781**	**16.0491**
Medium	11.1044	13.7918	6.1038
High	**30.5152**	6.4762	2.0138
n_read_forum_posts			
Low	15.1014	**25.9356**	**18.9629**
Medium	10.8257	6.907	3.2673
High	**17.76**	7.3035	1.9365
n_self_check_tests			
Low	3.9294	**19.7804**	**20.2902**
Medium	3.3961	9.6714	1.9324
High	**36.3616**	10.6943	1.9441
mark_ self_check_tests			
Failed	1.0001	1.005	**13.9949**
Passed	**41.687**	**38.1411**	9.1719
mark_exam			
Failed	1.0278	2.1752	**21.797**
Passed	4.2375	**36.407**	1.3555
Excellent	**38.4219**	1.5639	1.0143

Observation of maximum values of the probability of affiliation of the atribute mark_exam, in Cluster 1 we have grouped students which achieved excellent results at the exam, in Cluster 2 those who passed the exam and in Cluster 3 those who failed it.

Cluster 1 constists of 21% of students which in large extens used all kinds of offered learning material and had mostly passed self-check tests of knowledge. Grade of students in this cluster is *excellent*. Cluster 2 consists of 40% of students which moderately used lessons, interactive tasks, read forum posts and self-check tests in less extent, but on average have passed self-check tests. Students are grouped in this cluster achieved the grade *passed*. Cluster 3 has 39% of students which poorly used learning material and have in average failed self-check tests. Final grade of students in this cluster is *failed*.

How to use results from application EM algorithm? The biggest issue are definately students from Cluster 3 (failed). By following student activities in Moodle, during semester, teacher can certainly mark students which do not use any of learning materials, to give them timely warning, or to organize some other model of learning which will be more efficient for them.

For students in Cluster 2 teacher could create new learning material which will be combination of lessons, interactive tasks and self-check tests. That way these students can prepare and achieve even better results at the exam.

IV. CONCLUSION

This paper describes application one of data mining methods for analysis of behavior and activities of students in the e-learning systems. An advantage of data mining methods lies in the fact that a student’s low mark can be predicted on time. The teacher can predict what students have a tendency to fail the exam and can work with them on improvement of their characteristics and achievements prior to the end of semester and the final exam.

Based on the detected information from the student activities, the teacher can create a new kind of activities and learning materials from which the students receive higher grades, decide to eliminate some of the activities related to the low grade or to offer a new way of learning, which will improve passing the exam.

Continuation of work in this field will refer to Establishing a model of student knowledge in the e-learning systems using data mining methods.

ACKNOWLEDGEMENT

This paper was produced under the research project TR35026 funded by the Ministry of Education and Science Republic of Serbia.

REFERENCES

[1] B. Erol , Y. Li, "An overview of technologies for e-meeting and e-lecture", IEEE International Conference on Multimedia and Expo, pp. 6 pp, 2005.

[2] P. Zaharias and A.Poylymenakou, "Developing a usability evaluation method for e-learning applications: Beyond Functional Usability", *International Journal of Human Computer Interaction,* Vol. 25, Issue 1, 76-79, 2009.

[3] U.Fayyad, "Data mining and knowledge discovery in databases: implications for scientific databases", 9th International Conference on Scientific and Statistical Database Management, 1996.

[4] I.H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques", Morgan Kaufmann, 2005.

[5] J. Dougherty, M. Kohavi and M. Sahami, "Supervised and unsupervised discretization of continuous features", Internacional Conference Machine Learning Tahoe City, CA, 194–202, 1995.

[6] W. Hamalainen, J. Suhonen, E. Sutinen and H. Toivonen, "Data mining in personalizing distance education courses", World Conference on Open Learning and Distance Education, Hong Kong , 1-11, 2004.

[7] D. Frank, "The Expectation-Maximization algorithm", Technical Report GIT-GVU-02-20, Georgia Institute of Technology, 2002.

[8] B.Sean, "The Expectation Maximization Algorithm - A short tutorial", Last updated, January, 2009.