Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning

Mohammad Hassan Falakmasir, Jafar Habibi
falakmasir@ce.sharif.edu, jhabibi@sharif.edu
Computer Engineering Department, Sharif University of Technology

Abstract. In the past few years, Iranian universities have embarked to use e-learning tools and technologies to extend and improve their educational services. After a few years of conducting e-learning programs a debate took place within the executives and managers of the e-learning institutes concerning which activities are of the most influence on the learning progress of online students. This research is aimed to investigate the impact of a number of e-learning activities on the students' learning development. The results show that participation in virtual classroom sessions has the most substantial impact on the students' final grades. This paper presents the process of applying data mining methods to the web usage records of students' activities in a virtual learning environment. The main idea is to rank the learning activities based on their importance in order to improve students' performance by focusing on the most important ones.

1 Introduction

During the past decades, because of the significant benefits it brings for all participants, the use of information and communication technologies in the educational domain has become widespread all around the world. Particularly in Iran, according to the effective role of education in the national development plan, these types of training drew more attention from the major and prestigious universities. Thus, they began to set up e-learning departments one after another. Due to the shortcomings and deficiencies of the e-learning platforms in the early years, there were lots of unresolved problems which affecting both students and teachers performance. First of all, lack of collaboration and communication facilities caused students to feel lonely and unsupported. In addition, their educational tendency to instructor-led learning caused them face new challenges understanding self-paced learning materials. These problems inspired e-learning departments to use web conferencing and virtual collaboration tools to satisfy the students' demands. This paper presents the major findings that resulted from studying the e-learning activities and their impacts on students' final grades. The structure of the paper is organized as follows: Section 2 outlines a literature review and the related works. Section 3 presents a background of the e-learning platform and the students' learning activities. Section 4 describes the methodology used to conduct the study and analyses the results using a decision tree. Finally, conclusions and future works are presented in Section 5.

2 Related Work

Sometimes the term *Virtual Classroom* is referred as the whole e-learning process or all the teacher-student interactions. However, in this paper the term is used for online synchronous virtual meetings which are conducted by the participation of the teacher and students using audio and video conferencing technologies. There is a limited amount of research concentrating on the impact of these technologies upon the learning effectiveness. In contrast, there are a significant numbers of studies which have examined the role of other activities in the process of learning using data mining methods. In what follows a review on the most prominent studies is presented.

Bower and Richards [3] studied the impact of virtual classroom laboratories in computer science education. The main purpose of this research was to study the pedagogical aspects of these technologies; as a result they showed that such virtual laboratories are helpful particularly in the field of computer science. Also, another research by Redferm and Naughton [11] discusses and approves the positive role of video conferencing technologies in online education. This study particularly concentrates on the creation of brainstorming style-discussions and small group meeting which are fundamental to many of modern educational techniques.

From another perspective, some researchers have used web usage mining to look into students' activities in online learning environments. One of the leading studies conducted by Zaïane and Lou [18], employed these methods to improve the features of an e-learning environment. In [16], web usage data of students is used to cluster them based on 'Expectation Maximization' (EM) algorithm. Each cluster represents a group of students with similar behavior. Moreover, the results are used to give the students suitable advices according to their group. The EM algorithm is also used in [15] to extract similar behavioral pattern of students in a collaborative unstructured e-learning environment.

Minaei-Bidgoli and Punch [8], used genetic algorithms and a combination of multiple classifiers to predict students' final grades. In [5] and [9], several machine learning and classification techniques were applied in order to predict the students' final score; the relevance of each feature is also assessed. This work was extended in [4] and Artificial Neural Networks were used to predict students' final grades. Beck and Mostow [2], had a different approach toward studying the students' performance data. They used a method called 'Learning Decomposition' to evaluate students' success ratio based on the amount of pedagogical support they received. The most important point in all the studies is that predicting students' final score based on their online activities is the leading approach to examine the effectiveness of e-learning.

3 The Platform

The E-Learning Department of Iran University of Science and Technology (IUST) started its services in the spring semester of 2004 with about 700 students and is currently serving about 1,800 students in two Bachelors' and three Masters' programs. The instructional plan in this department is designed in a way that the learning materials are mainly developed in the form of multimedia courseware which can be accessed by students in a weekly manner. In addition, the teacher can add supplementary resources to the learning content and evaluate the process of learning by providing the students with assignments and online quizzes. Having gained proper perception about the course concepts, the students participate in a virtual classroom session so as to discuss the lessons with the teacher and other students. The teacher can also present complementary information and gain feedback about the students' learning progress.

After the three years of using commercial products as virtual learning environment, the E-Learning Department of IUST started to build its own e-learning platform based on Moodle, a free open-source Learning Management System (LMS). Moodle is designed to support the learning style of *Social Constructivism*, in which the process of learning is performed by a set of interactions between students, teacher, and learning materials [12]. This style is not mandatory in Moodle but is what it supports best. There are several kinds of activities which students can perform in a course such as: viewing courseware, uploading assignments, posting messages in forums, writing messages to teachers and other students, etc. The system keeps detailed information about the way students interact with the system which have inspired lots of researchers to use these data to apply knowledge discovery and data mining methods to extract useful information about the students learning behavior [7,13,14].

Although there are several activities such as messaging, forums, and text chat to support collaboration of teachers and students, a sophisticated synchronous collaboration tools in Moodle is still missing. Consequently, the e-learning center of IUST added a new module to its Moodle to conduct virtual classrooms. The module was developed based on Adobe Flash Platform [1] considering its attractive interface and low bandwidth requirements, making it suitable for students connecting from small towns in different parts of the country. For each course there are 16 two-hour online sessions in a semester which are conducted on the specified time every week. During the session, different levels of interaction such as using video, audio, document sharing, whiteboard, and text chat can be used depending on the requirements of the lesson. For example, the teachers can share a power point slide or simply use a virtual whiteboard to present the content as well as broadcasting their own voice and video. Students primarily use text chat to interact with the teacher and ask questions. It is also possible for teachers to permit the students send their voices. To support the “any-time, any-where” promise of e-learning, all the sessions are recorded and archived for the students who cannot participate online sessions. The students who attend the class can also review the parts of lessen they didn’t follow or understand. In fact, these recorded sessions can be used as a permanent learning resource and the students can review them as many times they want.

4 Methodology Design

4.1 Main Idea

Although Moodle presents several reports on the students’ activities, they are not flexible enough to satisfy the instructors’ needs for observing their interactions with the system [6]. Additionally, there is no way for educational technologists and training managers to indicate the value of each activity in success of students. As it was mentioned before, this research aims to rank online learning activities based on their impact on the students’ final grades. For this reason, some variables have been defined as key performance indicators (KPIs) of students. Then the impact of each variable has been evaluated based on its influence on the score of students in the final exams. Particularly, data mining techniques have been employed to analyze the web usage logs of the virtual learning environment to infer some rules about the importance of each activity in the performance of students.

4.2 Participants

The current study have been conducted on the web usage logs of the system in Fall Semester 2008, when about 1,300 students were enrolled in almost 100 courses. However, the research is limited to 824 students in 11 courses; the instructors used most of the learning activities; and, the final grades of students were also available. The total population of students under investigation for this research is larger than similar studies. In addition, the students were completely remote from the university and had to learn most of the concepts and practices just by using the system via the Internet. In previous studies [8,13,14] the e-learning platform was used to facilitate teacher-student interactions and the online activities of students were not assumed as the most essential part of the learning process.

4.3 Procedure

The general process of educational data mining consists of four steps: *Collecting Data, Preprocessing, Applying data mining,* and *Interpreting the results* [13]. Here, a similar process has been used which follows slightly different methods in data collection and preprocessing steps. The two steps are integrated into a single extended stage of building a data warehouse from the activity logs of students. This approach makes it possible to monitor and study the learning behavior of the students and its relevant trends more in depth. The use of Data Warehouse and On-Line Analytical Processing (OLAP) tools in e-learning is gaining popularity among educational institutes and virtual universities [19]. In this section the whole process of applying data mining methods on the students' usage information is described.

4.3.1 Building the Data Warehouse

As it was mentioned before, Moodle keeps detailed records of students' activities. The teacher has access to summarized reports about students such as the date of their first and last logins, and the number of pages visited by them. The information about each learning activity is also available according to the categories specified by the system, not by the professor. Consequently, we designed a model and built a data warehouse to monitor the students' activities in precise detail. The activities are classified into nine categories: *resource view, virtual classroom participation, archive view, assignment view, assignment upload, forum read, forum post, discussion read,* and *discussion post*. In addition, according to our interviews with instructional technologists and training managers of IUST, a list of data elements and analytical dimensions along with the students' KIPs have been defined. Then, corresponding information was extracted from the Moodle database to answer their questions. Anyhow, the details of the dimensional modeling are beyond the scope of this paper.

For this study an information model is being used which gathers the information about the identified business requirements in the form of a summary table. Each column of the table represents a dimension important according to our objectives. Table 1 shows the design of the summarized table. A brief description of each dimension is also included. The structure is quite similar to the one which was used in [13] but it contains some other attributes based on the KPIs extracted. To promote the level of interpretation and to facilitate the comprehensibility, the grades are stored in a discrete format. There are four categories of grades: A, if the value is equal or above 16.6; B, if the value is between 13.3 and 16.6; C, if the value is between 10 and 13.3; and F, if the value is less than 10.

Table 1: Summarization table of students activities

Name	Description
UserName	Name of User
CourseName	Name of the Course
ResourceView	Number of Coursware and Other Supporting Materials Views
VirtualClassroom	Number of Virtual Classroom Participations
ArchiveView	Number of Archive Views
ForumRead	Number of Forum Reads
ForumPost	Number of Forum Posts
DiscussionRead	Number of Discussion Reads
DiscussionPost	Number of Discussion Responses
AssignmentView	Number of Assignments Views
AssignmentUpload	Number of Assignment Answer Uploads
FinalGrade	Final Grade

4.3.2 Applying Data Mining Methods

The main group of data mining algorithms used in this study is ‘Feature Selection’. These methods, also known as ‘Attribute Evaluation’ algorithms, try to select the most relevant features according to a target concept. Several feature ranking and attribute selection methods have been proposed in the machine learning literature which use different metrics to discard irrelevant features and select the important ones including: information gain, gain ratio, symmetrical uncertainty, relief-F, one-R, and chi-squared. Each metric has its own bias. For example, the information gain measure is biased toward attributes with many values. Here, we use *gain ratio* [10] as the main evaluation metric since there are various number of records in the table regarding to each activity. The results of ranking based on the other methods are also presented and can be compared.

In this project, the data mining software package used to rank the attributes is Weka [17]. The reasons are that it is a free open-source application which implements several methods for attribute evaluation. Table 2 presents the results obtained from applying *gain ratio* attribute evaluation method on the summarized table of students’ activities. As the table shows the **virtual classroom participation** plays the most prominent role in this ranking while the second place belongs to the **archive views**.

Table 2: The results of ranking activities based on gain ratio metric

Attribute	Gain Ratio
Virtual Classroom	0.0839
Archive View	0.0694
Forum Read	0.052
Assignment View	0.0517
Assignment Upload	0.0497
Discussion Read	0.0364
Resource View	0.0324
Forum Post	0
Discussion Post	0

To confirm the results obtained from this evaluation some other methods are applied on the dataset and the attributes are ranked using other metrics. The results are outlined in Table 3. It can be perceived from the table that virtual classroom participation is the most influential feature affecting students' final grades among all other attribute evaluation methods.

Table 3: The The results of ranking activities based on other methods

Attribute	χ^2	Info-Gain	Symmetric Uncertainty	One-R	Relief-F	SVM
Virtual Classroom	1	1	1	2	2	2
Archive View	3	3	2	1	3	7
Forum Read	2	2	3	4	7	1
Assignment View	7	7	6	5	4	6
Assignment Upload	4	4	4	3	5	5
Discussion Read	5	5	5	7	6	4
Resource View	6	6	7	8	1	9
Forum Post	8	8	8	9	9	8
Discussion Post	9	9	9	6	8	3

4.3.3 Interpreting the Results

To illustrate and explain the results obtained from the research, a decision tree was created based on the C4.5 algorithm [10]. This algorithm uses the *gain ratio* metric to select the attributes and to build the tree. Figure 1 shows the first two levels of the tree. As depicted, the number of virtual classroom participation comes in the first level separating the students into two groups. Students who have participated fewer than 11 virtual classroom sessions, will probably (with the probability of about %55) fail in their exam. In contrast, Students with more than 11 participations might (with the probability of about %42) pass the exam with a C.

In addition, each node of the tree can be used to extract a rule to predict students' final grades based on their activities. For example, as highlighted in the figure, students with more than 11 virtual classroom participation and 17 archive views would get an A in the final exam. The coverage of this rule is about %25 and the accuracy is almost %41. These rules may help the teachers to identify the most important activities to focus on in order to improve their teaching style. The rules can also be employed by training managers and executives to provide with helpful information in resource planning and decision making.

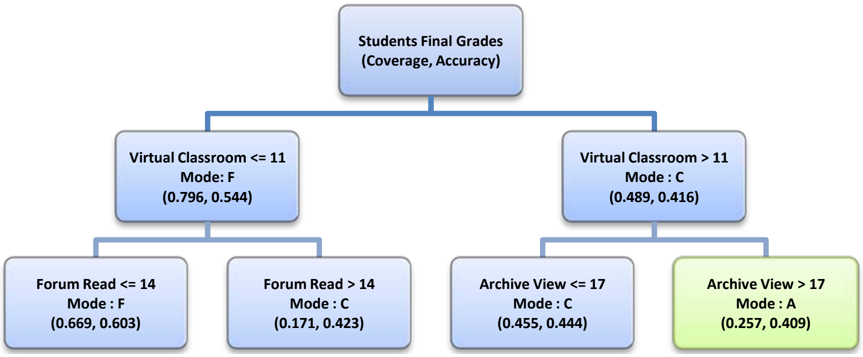

Figure 1: The first two levels of the decision tree model generated to predict students' final grades

5 Conclusions and Future Work

In this work we described the process of applying data mining methods in order to rank the students activities based on their impact on the performance of students in final exams. We used a number of ‘Feature Selection’ and ‘Attribute Evaluation’ methods together with real usage data of students picked up from the Moodle LMS in order to perform the case study. The results indicated that participation in virtual classroom sessions has the greatest impact on the effectiveness of learning in the particular settings of the IUST e-learning center. As a result, this fact motivated the managers and instructors to pay more attention to virtual classrooms and encourage the students to participate in these sessions. In the future, the effect of virtual classroom will be studied more profoundly considering some variables other than just the number of participation and archive views. It is also possible to analyze the students’ behavior in virtual classrooms more deeply considering the activities performed by students. Finally, the teachers’ instructional model in the virtual classroom will be studied in order to find the best methods that fulfill students’ demands which might have a great impact on their learning performance.

Acknowledgement

The authors gratefully acknowledge the managers, executives, instructors, and officials of the E-Learning Department of Iran University of Science and Technology for their valuable suggestions and comments.

References

[1] (2010) Adobe Flash Platform. [Online]. www.adobe.com/flashplatform/

[2] J. E. Beck and J. Mostow, "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of student," in *the 9th International Conference on Intelligent Tutoring Systems*, 2008, pp. 353-362.

[3] M. Bower and D. Richards, "The Impact of Virtual Classroom Laboratories in Computer Science Education," in *Thirty-Sixth SIGCSE Technical Symposium of Computer Science Education*, St. Louis, Missouri, USA, 2005, pp. 292-296.

[4] A. T. Etchells, A. Nebot, A. Vellido, P. J. Lisboa, and F. Mugica, "Learning What is Important: Feature Selection and Rule Extraction in a Virtual Course," in *The 14th European Symposium on Artificial Neural Networks, ESANN*, Bruges, Belgium, 2006, pp. 401–406.

[5] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting Students’ Performance in Distance Learning Using Machine Learning Techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[6] R. Mazza and V. Dimitrova, "CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses," *International Journal of Human-Computer Studies*, vol. 65, no. 2, pp. 125–139, 2007.

[7] A. Merceron and Kalina Yacef, "Mining student data captured from a web-based

tutoring tool: Initial exploration and results," *Journal of Interactive Learning Research*, vol. 15, no. 4, pp. 319–346, 2004.

[8] B. Minaei-Bidgoli and B. Punch, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System," *Genetic and Evolutionary Computation*, vol. 2, pp. 2252–2263, 2003.

[9] A. Nebot, F. Castro, A. Vellido, and F. Mugica, "Identification of Fuzzy Models to Predict Students Performance in an e-Learning Environment," in *The Fifth IASTED International Conference on Web-Based Education*, Puerto Vallarta, Mexico, 2006, pp. 74–79.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning.*: Morgan Kaufmann, 1993.

[11] Sam Redferm and Neil Naughton, "Collaborative Virtual Environments to Support Communication and Community in Internet-Based Distance Education," *Journal of Information Technology Education*, vol. 1, no. 3, pp. 201-211, 2002.

[12] W. H. Rice, *Moodle e-learning course development. A complete guide to successful learning using Moodle.*: Packt Publishing, 2006.

[13] C. Romero, S. Ventura, and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education* , vol. 51, no. 1, pp. 368–384, 2008.

[14] C. Romero, S. Ventura, P. G. Spejo, and C. Hervas, "Data Mining Algorithms to Classify Students," in *the 1st International Conference on Educational Data Mining*, Montral, Canada, 2008, pp. 8-17.

[15] L. Talavera and E. Gaudioso, "Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces," in *Workshop in Artificial Intelligence in Computer Supported Collaborative Learning in conjuntion with 16th European Conference on Artificial Intelligence, ECAI'2003.*, Valencia, Spain, 2004, pp. 17–22.

[16] C. Teng, C. Lin, S. Cheng, and J. Heh, "Analyzing User Behavior Distribution on e-Learning Platform with Techniques of Clustering," in *Society for Information Technology and Teacher Education International Conference*, 2004, pp. 3052–3058.

[17] (2010) Weka. [Online]. http://www.cs.waikato.ac.nz/~ml/weka/

[18] O. Zaïane and J. Luo, "Web usage mining for a better web-based learning environment," in *the conference on advanced technology for education*, Banff, Alberta, 2001, pp. 60-64.

[19] M. E. Zorilla, "Data Warehouse Technology for E-Learning," in *Methodologies and Supporting Technologies for Data Analysis*. Berlin, Heidelberg: Springer-Verlog, 2009, pp. 1-20.