Analyzing blended based learning systems

Tanja Mohar, Dejan Sraka, Branko Kaučič
Faculty of Education
University of Ljubljana
Kardeljeva ploščad 16, 1000 Ljubljana, Slovenia
{tanja.mohar, dejan.sraka, branko.kaucic}@pef.uni-lj.si

Abstract. *Nowadays as more and more teachers beside face-to-face education take online courses, at least partly in electronic learning environment, the usability of e-learning systems is one of the important questions in education. These e-learning systems accumulate a vast amount of information which is very valuable for analyzing students' behavior. However it is difficult to manage this kind of data manually.*

The paper describes educational data mining and student modeling as part of this area. We collected real educational data from programming courses in e-learning system Moodle. The objective is to preprocess the collected data and build a model, which will give the course facilitator information about possible course adaptation and learning recommendations based on the students learning behavior.

Keywords. blended learning, student modeling, educational data mining, Moodle

1 Introduction

The rapid emergence of technological innovations over the last half century (particular digital technologies) has had a huge impact on the possibilities for learning in the distributed environment.

The widespread adoption and availability of digital learning technologies has led to increased levels of integration of computer-mediated instructional elements into the traditional face-to-face learning experience. Today more educators use blended learning rather than traditional face-to face learning.

1.1 What is blended learning?

The idea of blended learning is based on theory of constructivism, where students build their knowledge based on their past and present knowledge. By this theory, educator's role is not delivering knowledge anymore but he/she assumes the role of a consultant [9]. In literature we can find three most common definitions of blended learning [4]:

1. blended learning is combining instructional modalities (or delivery media),
2. blended learning is combining instructional methods, and
3. blended learning is combining online and face-to-face instruction.

Most widespread use of the term blended learning in literature is that blended learning combines face-to-face education and e-learning. It tries to integrate technology to improve the learning process in terms of depth and scope [6].

The fact that neither students nor teachers are bound to a specific location and that this form of a computer-based education is virtually independent of any specific hardware platforms is one advantage of e-learning. Learning course/content management systems (in the continuation we will use L(C)MS) like Moodle in blended learning context support learning by offering a great variety of channels and workspaces to facilitate information sharing and communication between participants in a course, to let educators distribute information to students, direct students to information sources, produce content material [7], prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. [12].

1.2 Blended learning issues

Educator for his/her work requires different skills - social, didactical and technical skills. Many of them lack time, didactical know-how, technical expertise, incentives, and flexibility, to use e-learning platforms for more than convenient repositories of slides [6].

Psychological and pedagogical theories highly agree on viewing lectures that serve only to transmit information onto several students as not being very effective in the long run [14]. Because this kind of

knowledge does not take into account any individual needs, interests and learning style, it tends to be forgotten very quickly.
Blended based approach to learning offers what is the best in both worlds: face-to-face and web-based learning. Educator can organize his/her class by encouraging social learning or individual learning, with elements of frontal learning and later with group or individual practice.

Blended based learning systems accumulate a great amount of data which can be useful for analyzing student's behavior. In general, they accumulate a large quantity of log data about student activities, such as reading, writing, taking tests, downloading course material, cooperating in forums and wikis, communicating with peers, etc.

Although some systems offer their own reporting tools, they fail to meet the educator expectations and needs, due to a great amount of daily accumulated data and the number of all students enrolled in a course. They do not provide specific tools which would allow educators to thoroughly track and assess all the activities performed by their learners. The problem of analyzing student's behavior for better learning process is well covered with educational data mining (EDM) area.

The paper is oriented toward "how to" point of view of analyzing educational data with emphasize on the learning management system Moodle and is arranged in the following way: section 2 describes the background of EDM; section 3 continues to explain the area of EDM and how it can be used in blended learning, categorizes research areas of EDM and presents more details on student modeling; section 4 describes a case study of a programming courses for academic years 2009/10 and 2010/11 where we gathered data in order to find out which course obligations and activities contribute the most to the success of students in programming courses. In section 5 final conclusions are outlined.

2 Backgrounds

First articles about EDM and their mining task in an educational system are listed in [5] and [11]. They list the articles that were most influential in early years of educational data mining research. Zaïane [16] suggested an application for data mining, for using it to study online courses. In 2002 he wrote an article about how educational data mining methods (specifically association rules and clustering to support collaborative filtering) can support the development of more sensitive and effective e-learning system [17][1]. In [1] we can find some new research area (at that time): study of gaming the system (attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material). In [3] Beck and Woolf already wrote about how educational data mining prediction methods can be used to develop student models. They used a variety of parameters to predict whether a student will make a correct answer. After this work, student modeling has become a key theme in modern EDM and the paradigm of testing EDM models' ability to predict future correctness has become very common [2].

3 Educational data mining

The Educational Data Mining community on their website www.educationaldatamining.org defines educational data mining as »an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in«.

Data mining methods are often different from standard data mining methods, due to the needs to explicitly account for (and the opportunities to exploit) the multi-level hierarchy and non-independence in educational data. For this reason, it is increasingly common to see the use of models drawn from the psychometrics literature in educational data miming publications [2].

Educational data mining consists of three different research areas [10]:

- offline education,
- e-learning and learning management systems (LCMSs), and
- intelligent tutoring systems (ITSs) and adaptive hypermedia systems (AEHSs), which are trying to personalize teaching and learning by taking into account the needs of each particular student.

3.1 Usefulness of educational data mining in blended learning

For good evaluation of blended learning, the data must come from two different types of educational systems. In blended learning we have to take into account face-to-face learning (traditional learning in classroom environment) and web-based education (LCMSs like Moodle, Blackboard, etc.). Traditional classroom only have information about student attendance, course and curriculum goals, while web-based educational systems record more information. These systems contain all the information about students' actions, their interactions and record them into log files and databases.

In the last years, researchers wrote about different data mining methods to help educators improve learning process. Some of the main e-learning problems to which EDM methods were applied are [12]:

- dealing with the assessment of student's learning performance,
- providing course adaptation and learning recommendations based on the students' learning behavior,
- dealing with the evaluation of learning materials and educational web-based courses,
- providing feedback to both educators and students of e-learning courses, and
- detection of atypical student's learning behavior.

Blended learning can be more student-oriented and can get into account students' individual needs and their learning styles, where that is possible. For better personalization of e-learning and face-to-face learning, educator can recommend learner specific learning activities, resources, suggests path pruning or shortening learners learning path if necessary. Educator can get more objective feedback for his/her instruction and can evaluate the structure of course content, can classify learners into groups based on their needs for guidance and monitoring, etc.

3.2 Educational data mining process and categorization

EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice [10]. The e-learning data mining process in LCMS is not significantly different as the general data mining process. It consists of four steps:

- collecting the LCMS usage data,
- preprocessing the data,
- applying data mining algorithms, and
- interpretation, evaluation and deployment of the results.

In LCMSs we collect the data. Such systems automatically collect the usage data in logs in a relational database (e.g. MySQL). As an alternative we can use the existing data, e.g. in 2008 the Pittsburgh Science of Learning Center has opened a public data repository, the PSLC DataShop, which makes substantial quantities of data from a variety of online learning environments available, for free, to any researcher worldwide [2]. In the next step we have to decide which aspects we wish to observe and preprocess the data, build a summarization table, discretize (categorical values are more friendly for educator than precise magnitudes and ranges) and transform the data to a required format of the data mining algorithm. Further we apply data mining algorithms to build and execute the model that discovers and summarizes the knowledge of interest to the educator, student or administrator of the LCMS. In the final, fourth step, we obtain the results or model, which can be interpreted and used for further actions.

In literature we can find two different categories or taxonomies of educational data mining. Romero and Ventura [11] categorize EDM into the following categories:

- statistics and visualization,
- web mining:
  - clustering, classification, and outlier detection,
  - association rule mining and sequential pattern mining, and
  - text mining.

While Romero and Ventura focused on categorization based on development of EDM, Baker [2] proposed another viewpoint of EDM:

- prediction:
  - classification,
  - regression, and
  - density estimation.
- clustering,
- relationship mining:
  - association rule mining,
  - correlation mining,
  - sequential pattern mining, and
  - casual data mining.
- distillation of data for human judgment, and
- discovery with models.

Both taxonomies involve data mining tasks and methods. Romero [13] states that classification is one of the most useful tasks of data mining in e-learning. It predicts a value of attribute based on the values of other attributes. The use of association rule mining discovers relationships among attributes in databases (IF-THEN rules). Sequential pattern mining is a bit more restrictive as association rule mining, while it takes into account items order. It tries to discover if the presence of a set of items is followed by another item in a time-ordered set of events.

Baker and Yacef [2] also defined four key applications of EDM methods: the improvement of student models; discovering or improving models of a domain's knowledge structure; studying pedagogical support (both in learning software, and in other domains, such as collaborative learning), towards discovering which types of pedagogical support are most effective, either overall or for different groups of students or in different situations; looking for empirical evidence to refine and extend educational theories and well-known educational phenomena, towards gaining deeper understanding of the key factors impacting learning often with a view to design better learning systems.

3.3 Student modeling

We already stated that student modeling soon became a key theme of educational data mining. McCalla [8] defines student modeling as: »student modeling involves the construction of a qualitative representation that accounts for student behavior in

terms of existing background knowledge about a domain and about learning the domain. Such a representation, called a student model, can assist an intelligent tutoring system, an intelligent learning environment, or an intelligent collaborative learner in adapting to specific aspects of student behavior«.

Sision and Shimura [15] write further that student modeling is a construction of a qualitative representation, called a *student model* that accounts for *student behavior* in terms of a system's *background knowledge*.

We build student model over a certain domain. This domain is an expert knowledge and can have a different structure and complexity of knowledge. For good defined domain (like algebra in math) we can prepare a sequence of practical exercises and tests.

In general, we can define student modeling like:

student model = student behavior + background knowledge

In the continuation, these three components of student modeling are presented according to Sison and Shimura [15].

3.2.1 Student behavior

The term "student behavior" is used for a student observable response to a particular stimulus in a given domain. The stimulus with the response serves as the main input into a student modeling system.

Moodle offers us a variety of different usage data of student activities that can be observed. We can observe student actions like writing, reading etc. For an input in student modeling system we can also use the results of an action (post on a forum, solved quiz, etc.) that Moodle offers.

3.2.2 Background knowledge

The background knowledge includes:

- the correct facts, procedures, concepts, principles, schemata and/or strategies of a domain (called the theory of a domain or a domain knowledge), and
- the misconceptions held and other errors made by population of students in the same domain (called the bug library).

We can also mention the quality and quantity of information about student (data gathered between solving problems in a given domain) and an open model of a student based on the different data about student, his/her wishes, evaluation of knowledge from other peers and cooperation with the educator.

3.2.3 Student model

A student model is an output of a student modeling process. The student model is an approximate, possibly partial, primarily qualitative representation of student behavior about a particular domain, or a particular topic or skill in that domain, that can fully or partially account for specific aspects of student behavior. Model describes objects and processes in terms of spatial, temporal or causal relations.

With built student model, we can:

- identify specific relationship between the input behavior and the system's background knowledge,
- search for mismatches between actual and desired behaviors, and
- recognize misconceptions (incorrect or inconsistent facts, procedures, concepts, principles, schemata or strategies that result in behavioral errors) and other classes of knowledge errors (inconsistent, missing or incomplete knowledge).

There are three approaches to construct a student model. The most basic approach is where a student model is assumed to be a subset of the expert model. However, this approach necessarily excludes the misconception and incorrect knowledge diagnosis. This is so called overlay model. The other two approaches are trying to deal with the problem of exclusion. The analytic or transformational approach tries to use the background knowledge to transform the student behavior to the problem given and to verify if the problem given and the desired behavior are equivalent or not. The third, synthetic approach, tries to obtain a set of behaviors and compute a generalization of these by synthesizing elements from the background knowledge or input data.

4 Case study: programming courses in Moodle

For building an example of a student model we used real data from a course Programming, which took place in traditional classroom and in web-based educational system Moodle.

For academic year 2009/10 we observed two courses (two modules in Moodle): *Programming 1* in first year of Bologna undergraduate study programme "The two-subject teacher" for students that took Computer science for one of the subjects (42 enrolled students), and course *Programming* in second year of older (before Bologna reform) study programme "Math and computer science teacher" (41 enrolled students).

In academic year 2010/11 course Programming was replaced within Bologna reform with courses *Programming 1* and *Programming 2*. In academic year 2010/11 we observed two courses (again two

modules in Moodle): *Programming 1* (first year of Bologna undergraduate study programme, 46 enrolled students) and *Programming 2* (second year of the same Bologna undergraduate study programme, 38 enrolled students).

We gathered the data about students' presence in all courses (*Programming*, *Programming 1* and *Programming 2*) and activities conducted in Moodle. The course lectures of observed courses (modules) took place in classroom with the presence of educator and students, while course exercises were divided in exercises with teacher assistant and exercises where students had to do assignments, cooperate in forums discussions, etc. in Moodle.

As part of the course *Programming* students were obliged to submit homework after each set of exercises or to submit three seminar works at the end of semester. Seminar works or 80% of all home works was one of the conditions for taking oral exam.

As part of the course *Programming 1* students were obliged to submit at least five of six seminar works and defend the theoretical part of their big seminar work. These were also one of the conditions for students to attend the oral exam.

As part of the course *Programming 2* students were obliged to carry out the seminar work that covered the idea concept of the application, implementation of application and instructions how to use the application.

The total number of all students from Faculty of education University of Ljubljana that we used in this study is 167. At the beginning of both academic years students were informed that data gathered in Moodle will be used for the educational data mining research.

The programming knowledge domain is not very well defined. Applications and small programs that students submitted may not be solved in unique way (there exist many different solution approaches to one problem). So the main goal is to detect which activities had the biggest influence on students' programming knowledge. Besides obliged activities in Moodle we also looked at other optional activities (module attendance, forums, wiki and resources).

4.1 Analysis

For each student we gathered his/her interaction data scattered over several tables for selected course unique identifier. Summarization table (Table 1) for each enrolled student was used as input data; there was no missing or incomplete data.

Instead of precise value of grade (in percentage) we used categorical values. We performed a discretization of grade in the way that we defined the value "fail" if grade < 50% and the value "pass" if grade >= 50%. In general, models obtained using categorical values are more comprehensible than when using numerical data because categorical values are easier for an educator to interpret than precise magnitudes and ranges [13].

Table 1: summarization table

Name	**Description (per student)**
ID_course_view	Number of accesses to the course
N_forum_view	Number of all views of all forums in a course
N_discussion_view	Number of discussion views in course
N_discussion_add	Number of discussions added in course
N_post_add	Number of posts added
N_assignments_all	Number of all assignments (all possible)
N_assignments	Number of all submitted assignments
N_resource_view	Number of all resources viewed
N_wiki_edit	Number of all wiki edits
N_wiki_view	Number of all wiki views
Grade	Pass if grade>=50%, fail if grade<50%

For analyzing the collected data from Moodle, we used Weka software. First we had to transform the data into required format ARFF (a text file that describes a list of instances sharing a set of attributes). Then we had to decide which algorithms we will use. We decided for classification algorithms. For an educator a user friendly model is very important, so we used decision trees and rule induction algorithms. From an educators' point of view decision trees are considered easily understood. However, if tree is very large (with a lot of nodes and leaves) then they are less comprehensive. Decision tree can be directly transformed into a set of IF-THEN rules that are one of the most popular forms of knowledge representation, due to their simplicity and comprehensibility. Rule induction algorithms are also considered to produce comprehensible models because they also discover a set of IF-THEN classification rules that can be used directly for decision making [13]. We used Ridor rules (Ripple down rule learner rules) because they expose the most likely option of a class.

4.2 Experimental results

In course Programming in academic year 2009/10 we discovered that the most important activity was assignments. Pruned decision tree (J48) suggests that if student in general submitted more than 17 assignments then he/she had 81% chance that he/she will pass (Fig. 1).

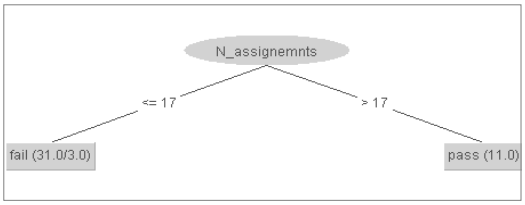

Figure 1: Programming, 2009/10

Ridor rules give us even more detailed explanation. Student will pass, except if he/she did not submit at least 14 assignments or at least 17,5 assignments. If all the conditions are fulfilled then student had 85% chance to pass. Here we have to consider that we used cross validation, where data was not equally represented (the power of classes fail and pass were not equal).

The course *Programming 1* in study year 2009/10 gives us a slightly different result. Among all activities, the J48 decision tree model exposes the number of resources view. Student that were reviewing the studying material more regularly, have been more successful (Fig. 2).

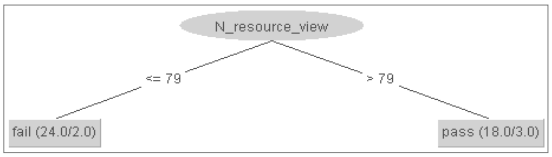

Figure 2: Programming 1, 2009/10

The Ridor rule gives us even stricter rule. The student fails, except if he/she reviewed the study material at least 82 times. This shows the importance of theoretical background knowledge.

In academic year 2010/11 in course *Programming 1* we can observe that beside assignments also cooperation in the forum was important. The J48 decision tree model is only 69% accurate, while Ridor rules give us more specific conditions with 82% certainty (Fig. 3).

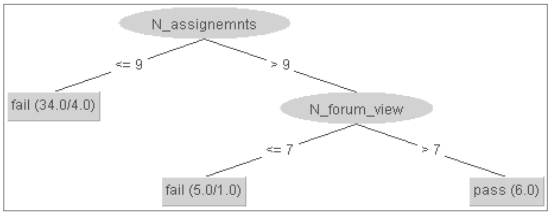

Figure 3: Programming 1, 2010/11

Student will fail except if number of submitted assignments is bigger than nine and forums were viewed at least seven times.

For course *Programming 2* in academic year 2010/11 we get J48 decision tree with three nodes: number of assignments, number of resource views and number of viewed forum topics. These were the most important elements that improved in understanding students' knowledge about a domain of programming (Fig. 4).

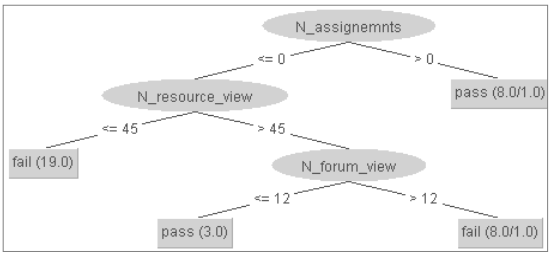

Figure 4: Programming 2, 2010/11

The J48 tree model is a model with 89,5% of correctly classified instances. Ridor rules in this case are not better than the decision tree, while their prediction of correctly classified instances was only 73,7% correct. But it highlights some elements that are left out in decision tree. One of the findings is that the number of course views is not among the nodes in decision tree, but it is still important for student.

Ridor rules for *Programming 2* in academic year 2010/11 are:
class = fail (38,0/11,0)
Except (N_assignments > 1) => class = pass

Except (N_resource_view > 46) and (N_course_view <= 233,5) => class = pass

These results do not take into account differences among students that have already attended the course in previous year and between the students that are attending the course for the first time. Some achievements of these students were acknowledged from previous year, so they were relieved from some obligations of the course, due to the fact that they have already done it in the past year.

5 Conclusions

In the paper we have discussed how educational data mining and student modeling can contribute to education in blended learning. As an example of semi blended based learning approach we took the programming courses where students some of their obligations perform outside classes. Educational data mining approach and student modeling we conducted is appropriate also for real blended learning courses.

Our student model(s) was/were built on real data. We used only decision trees and Ridor rules, because they are easy to interpret and easy to understand for

course facilitators. From these examples we can conclude that theoretical background is very important for meeting the course objectives that allow student to pass the course. Along with theoretical background very important are also practical exercises with strong inner motivation. A lot of students passed if they've done their homework or seminar work, but even more students passed if they did exercises at home. The third important aspect of good knowledge is peer cooperation, which is shown within the importance of forums views.

In similar way, analyzing any blended learning course that uses LCMS that tracks log of students' activity can result in student models that can pose valuable information for course facilitators.

Among EDM methods and tasks we intentionally left out to mention fuzzy rules, although they are also very useful from educator's point of view. They would be useful if among input data there would also be oral and written evaluation of students' knowledge and criteria for evaluation. In addition, we pointed out that definition of a programing knowledge domain is neither simple nor good. To build a good model over this domain we have to narrow modeling scope.

Educational data mining is an emerging discipline and we have made only the first step towards improvement and adapting our blended courses in order to enable students to achieve the required standards of knowledge and progress in their study.

References

[1] Baker R.S., Corbett A.T., Koedinger K.R. Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, pages 531-540, 2004.

[2] Baker R.S.J.D., Yacef K. The State of Educational Data Mining in 2009: A review and Future visions. *Journal of Educational Data Mining*, 2009.

[3] Beck J., Woolf B. *Lecture Notes in Computer Science: Intelligent Tutoring systems*. High-Level Student Modeling with Machine Learning, 1839:584-593, 2000.

[4] Bonk C.J., Graham C.R. (eds.). Pfeiffer Publishing. *Handbook of Blended Learning: Global perspectives, local designs.* San Francisco, 2006.

[5] Castro F., Vellido A., Nebot A., Mugica F. *Evolution of Teaching and Learning Paradigms in Intelligent Environment.* Applying Data Mining Techniques to e-Learning Problems, 62:183-221, 2007.

[6] Derntl M., Motsching-Pitrik R. The role of structure, patterns, and people in blended learning. *The Internet and Higher Education*. 8(2):111-130, 2005.

[7] Kaučič B., Krašna M. Blended learning as implementation of Bologna process – learning resources issues through a case study. *Proceedings of 4th International Conference of Education, Research and Innovation*, pages 3616-3625, 2011.

[8] McCalla G. The central importance of student modeling to intelligent tutoring. *New Directions for Intelligent Tutoring Systems*, 1992.

[9] Mei L., Yuhua N., Peng Z., Yi Z. Pedagogy in the information age: Moodle-based Blended Learning approach. *International Forum on Computer Science-Technology and Applications*, 3: 38-40, 2009.

[10] Romero C., Ventura S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Par C: Applications and Review*, 40(6): 601-618, Spain, 2010.

[11] Romero C., Ventura S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1): 135-146, 2007.

[12] Romero C., Ventura S., Garcia E. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1): 368-384, 2008.

[13] Romero C., Ventura S., Hervs C., Gonzales P. Data mining algorithms to classify students. *Proceedings - International Educational Data Mining Society*, 2008.

[14] Salmon G. Kogan Page. *E-Moderating – The Key to Teaching and Learning Online*. London, 2000.

[15] Sison R., Shimura M. Student Modeling and Machine Learning. *International Journal of Artificial Intelligence in Education*, 9:128-158, 1998.

[16] Zaïane O. Web usage mining for a better web-based learning environment. *Proceedings of Conference on Advanced Technology for Education*, pages 60-64, 2001.

[17] Zaïane O. Building a recommender agent for e-learning system. *Proceedings of the international Conference on Computers in Education*, pages 55-59, 2002.