Predicting Student Outcome in Moodle

Igor Felix
INF-UFG,
igormoreira@inf.ufg.br

Ana Paula Ambrosio
INF-UFG,
apaula@inf.ufg.br

Jacques Duilio
Dep. Computação-UEL,
jacques@uel.br

Eduardo Simões
INF-UFG,
eduardo@inf.ufg.br

Abstract – **Preventing student failure and dropout is an important aspect of academic success. This paper presents MoodlePredicta, an educational data mining tool used to predict students' outcome and performance. It was developed for the Moodle environment, and uses data generated by all user actions, including their interactions with the system, content and other users. This dataset reflects students' behavior, allowing analysis to monitor their progress and predict their outcome in the course. Predictions are conducted using Naïve Bayes classifier algorithm. The training dataset is composed of 13 different cohorts, with 1.307 students, and for each student 40 attributes were included. The tool was tested and evaluated by a group of specialists in distance education in a Brazilian educational institution, that assigned higher grades to almost all analyzed items.**

Keywords - Educational Data Mining, Moodle, Prediction, Tool, Virtual Learning Environment.

INTRODUCTION

The virtual learning environment Moodle is the most widely used distance learning platform (Educause, 2014). It offers, within the system, several pedagogical and administrative tools and instruments, that can be used to promote teaching and learning experiences. As the user interacts with the environment, Moodle automatically records all users' activities. These records contain useful and relevant information that can be used to improve distance learning (Romero & Ventura, 2010). However, analyzing data and extracting information from Moodle can be a complicated task, since the environment in its basic version (no added plugins or other changes) has hundreds of tables in its database. To manipulate them, a prerequisite knowledge of data mining is necessary, and not always the distance-learning teams have a professional with such experience.

Educational Data Mining (EDM) and Learning Analytics (LA) are new research domains that focus on the analysis of data originated in educational environments. Published papers in these domains mainly present the result of the application of data mining techniques on specific databases to predict student performance or activities within the environment (Peña-Ayala, 2014). Very few include the development of tools that could contribute to distance education (S. Khatwani & Arya, 2013). MoodlePredicta is a system developed to allow visualization of user behavior and student progress within the Moodle environment, and the prediction of student outcome using datamining techniques.

METHODOLOGY

The methodology adopted in the research project was divided in several steps:

• Theoretical foundation: survey of related literature, including educational data mining and the Moodle architecture.

• Data Extraction: selection and pre-processing of data from Moodle tables. The data used in this process included an online undergraduate course in Biology and an online tutor training course offered at the Federal University of Goias (UFG). These courses were chosen because they represent two distinct and relevant environments for the analysis of contexts and mapping of different behaviors.

• Survey of profiles: tracing of behavioral profiles and characteristics of the students who failed or dropped out from online courses, using data mining tasks and techniques. In this step several algorithms were tested to verify the one best adapted to the problem. The algorithm that obtained higher accuracy rates was Naive Bayes, an algorithm that uses conditional probability, based on Bayes theorem (defines how to find the probability of an event occurring, given the probability that another event has already occurred).

• Implementation of MoodlePredicta: development of tool that allows the visualization of data from the Moodle environment and prediction of student outcome, alerting the teacher to the students who are at risk of failing or dropping out of the discipline.

• Validation: the tool was tested and evaluated by a group of distance education specialists and tutors that are active users of the Moodle platform. Criteria used in the evaluation of the tool include: its operation, access to database, response and process-time, the results presented.

FINDINGS

The main objective of this research project was the definition and implementation of the MoodlePredicta environment. It allows direct connection to Moodle tables, performing the collection, processing and analysis of a varied set of data, generating a series of reports.

MoodlePredicta was divided into two modules: Visualization and Prediction Modules. Underpinning these

two modules, is the preprocessing basis of the environment that establishes a database connection, collects information from tables, then prepares the data, formatting it, and applying the necessary transformations. The resulting dataset can then be visualized in the Visualization module or analyzed by the WEKA data mining tool (Hall et al. 2009) that has been integrated in the Prediction module to predict student outcome using the Naive Bayes classification algorithm. This prediction aims to indicate if a given student may be at risk of failure or dropping out of the course.

Data collected from Moodle include detailed information related to student interactions in forums, chats, quizzes and activities, as well as time students were logged on and their grades. Examples of data collected from forums include: total of forums the user participated in, number of posts in the forums, number of posts that have been updated by the user, amount of posts the user has read, total of readings received from user's publications, number of discussions the user participated, number of discussions the user has created, number of replies received by the user, number of characters the user posted, total of words posted by the user, number of phrases sent by the user. From quizzes: total of quizzes that the user answered, number of submissions user made, number of questions the user answered, total of questions with correct answer, number of questions with incorrect answer, total time spent to answer the quiz. In total, 40 attributes are collected that give a good summary of the students' participation in class through the learning environment. All this information can be visualized in MoodlePredicta, organized by cohort or by student.

For selection of the prediction algorithm, a training dataset, composed by 13 different cohorts, with 1.307 students, was used. In datamining, this data is used to build a model that will be used for prediction. Different algorithms define different models. Several algorithms were tested to identify the one that presented the best results for the given problem. Tested algorithms included Decision Trees, Multilayer Perceptron, Regression, among others. The one that presented higher accuracy was Naïve Bayes, with an accuracy higher than 87%, that was then used in the implementation of the Prediction Module.

MoodlePredicta analyzes student data incrementally along the course. As the course advances, the Prediction Module incorporates the data generated by the students, generating a prediction of the student's performance for each week. In this context, performance is divided in two groups: positive outcome, when the student is expected to complete the course with passing grade, and negative outcome, when the student is expected to dropout or fail the course.

Results can be visualized by cohort, where the tool presents the percentage of students in each performance group, and the lists of students in each group. Individual student performance can also be visualized, showing a graph where each week of the course is colored in green, if the student has a positive predicted outcome, or in red, if the predicted outcome is negative, based on the student's behavior data up to that week. This analysis allows teachers to follow and evaluate student performance along the course, giving support to early intervention strategies that can be put into place to aid and recover students at risk.

The tool was validated by a group of 10 people directly involved with Moodle and online courses with intermediate or advanced knowledge of the Moodle environment. They were presented to MoodlePredicta and allowed to manipulate the tool as desired. They were given a questionnaire collecting information about their profile, their degree of satisfaction with the tool, and evaluation of its usability. These aspects were evaluated using a Likert scale of 1-6. They were also invited to propose additions and modifications they deemed appropriate, including new data visualization graphs. 90% thought outcome prediction is important and would use the tool in their courses.

CONCLUSIONS

MoodlePredicta is a tool that focuses on the prediction of students' outcome, based on their behavior within the virtual learning environment. This type of tool helps identify students at risk, understanding the variables that impact student performance, allowing teachers to propose early intervention strategies, that may increase academic success.

The prototype was well evaluated and now must be tested in a working environment. However, the Moodle installation has recently changed, and the tool must be modified to adapt to these changes and be installed in the university's system.

Furthermore, MoodlePredicta is being expanded to analyze how teacher behavior impacts student outcome. This will allow teachers to reflect on their approach and take it into account when designing and presenting the course.

REFERENCES

Educause (2014) Center for Analysis and Research. The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009) The weka data mining software: An update. SIGKDD Expl. Newsl., 11(1):10–18.

Khatwani, S. & Arya, A. (2013) A novel framework for envisaging a learner's performance using decision trees and genetic algorithm. In 2013 International Conference on Computer Communication and Informatics, 1–8.

Peña-Ayala, A. (2014) Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4, Part 1):1432 – 1462.

Romero, C. & Ventura, S. (2010) Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6):601–618.