World Transactions on Engineering and Technology Education
Vol.12, No.3, 2014

Research based on data mining of an early warning technology for predicting engineering students' performance

Ke Zhu

Henan Normal University
Xin Xiang, People's Republic of China

ABSTRACT: There is a crisis in engineering students' performance and the best way to eliminate it or to nip it in the bud is through a pre-control mechanism, or *warning*. It is considered most important to adopt the appropriate technology and methods of such an early warning. Based on theories and methods, some measures can be taken to deal with this crisis. Therefore, the use of different data mining approaches for improving the prediction of students' performance is proposed in this article, starting from the test data drawn from quantitative, qualitative data and also the learning materials. The objective is to determine how the selection of attributes, the use of different classification algorithms, and the date and time when data is gathered, affect the accuracy and comprehensibility of the prediction. The results of this study show that the method proposed here, the early warning technology of students' performance, has had a positive influence on their learning effectiveness.

INTRODUCTION

During the past few years, e-learning has grown significantly. Contemporary Web-based courses take advantage of Internet capabilities to support and improve effective traditional education while, at the same time, offering greater innovative possibilities [1]. With the arrival of the knowledge-based economy, university students - and especially engineering students - are required to have more knowledge and ability. However, the current situation is that contemporary university students' practical ability is generally low [2].

In the face of global competition in science and technology, it is imperative that China produces a high number of highly qualified, innovative people [3]. Higher education in engineering must produce a great many qualified engineers. Cultivating engineering students is an urgent requirement for economic development in the new century as is the need for ceaseless innovation in higher engineering education.

Numerous different warning systems have been implemented and validated in emergency medical admissions [4]. Many institutions have begun to turn to information technology, i.e. early warning technology, to provide predictors of developments in business or marketing. For example, IBM provides a service to analyse past data from the stock market to predict when clients should buy or sell stock.

Analytics are now being used in higher education to identify and even predict students at risk of failure, by studying demographic and performance data of former students on the same course, subject and college [5].

The learning management system (LMS) is an early warning technology used by instructors to build and maintain course Web sites. Web site maintenance includes posting course content, updating events, and managing interactive communication with students via messages, forums and surveys [6]. To justify the investment in LMS, it is important to study the pattern of student use of LMS and students who are at risk.

Due to complex effects and interactions between the organisational variables and LMS, as evident by diverse findings in the past, no hypotheses were raised in advance regarding the correlations among the variables [7]. Rather, two research questions are raised in this exploratory study:

1. To what extent is LMS used by students and to what extent are students at risk in their performance?
2. To what extent are students' learning performance (dependent variables) correlated with the instructor, course, course year, course size, staff size, content on the course Web site, and existence of forums on the course Web site (independent variables).

METHODOLOGY

The early warning technology was implemented for all engineering students in the summer of 2013. In this research, the author designed a framework for categorising key milestones in any institutional application of analytics:

- Stage 1 - Extraction and reporting of resource data;
- Stage 2 - Analysis and monitoring of operational performance;
- Stage 3 - What-if decision support (such as scenario building);
- Stage 4 - Predictive modelling and simulation;
- Stage 5 - Automatic triggers of business processes (such as alerts).

Data Collection and Procedures

In this study, the early warning technology was implemented in three colleges. In total, 210 students responded to the survey, resulting in a response rate of 86.4%, while only 170 responses were valid. Each course consists of six learning activities and is offered twice a quarter. During the first learning activity, the educational material of each activity is delivered to the students and their knowledge is assessed through multiple choice tests to determine their knowledge, and projects are carried out to test the application of this knowledge in practice. The final examinations for each course are conducted during the last activity. The structure of each educational activity is predetermined, using the same number of multiple choice tests every year.

When designing the monitoring module, the first design decision was to select only a subset of the tools available for monitoring. Exhaustive monitoring, although feasible, would have made the collection of all the observations in a central server very difficult. As a result, only the tools most commonly used by students to work on course activities were selected. This collection included browser, text editor, command interpreter and several additional programs related to the course (see Table 1). The author collected the educational data mining date using log data from the learning management system. However, a problem encountered by the author was reflected in the question *which features should one select?* Finally, the author also produced data about:

- clicks, logins;
- postings in forums, chats and dialogs;
- minutes on-line;
- total and average dwell time.

With these data, the author defined additional metrics: activities per session and clicks per session. Following the approach of Herzog et al [8], the author discretised these numerical attributes.

Table 1: Course tracking variables selected for the early warning technology.

Total number of on-line sessions	Number of uses of the *Compile* tool	Number of assessments started
Total time on-line	Number of uses of the *Search* function	Number of assessments finished
Number of mail messages read	Number of visits to MyGrades tool	Time spent on assessments
Time spent on assignments	Number of visits to MyProgress tool	Number of assignments read
Number of discussion messages read	Number of uses of the *Who is on-line* viewer	Number of assignments submitted
Total number of discussion messages posted	Number of visits to the course chat area	Number of mail messages sent
Number of new discussion messages posted	Number of files viewed	Number of Web links viewed

Integration and Processing Data

Databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources [9]. Low-quality data will lead to low-quality mining results. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Careful integration can help reduce and void redundancies and inconsistencies in the resulting data set. This can help improve the accuracy of the subsequent data mining process.

The author defines $D_{Training}$ as training data set, D_{Test} as test data set, $R=\{r_1,r_2,\cdots,r_n\}$ as early warning rules data set, $P(r)\ r:X\rightarrow Y$, as accuracy, $PR(R)$ as early warning rule’s accuracy rate.

Then, the accuracy rate and the early warning rule’s accuracy rate are:

$$P(r) = \frac{\left|\{T : X \cup Y \subseteq T, T \in D_{Test}\}\right|}{\left|T : X \subseteq T, T \in D_{Test}\right|} * 100\% \tag{1}$$

$$PR(R) = \frac{\sum_{r \in R} P(r)}{|R|} \tag{2}$$

The early warning technology architecture was based on the reference framework of Chen et al [7]. Figure 1 depicts the architecture for the early warning technology. The early warning technology framework consists of 12 modules. Data mod (first layer) gathers the data mining service repository and other data sources, which stores the data to be processed. The data access uses a wrapper, which mediates between calls from client application components to the data sources by transforming incoming requests into a message format that is understandable to the enterprise components.

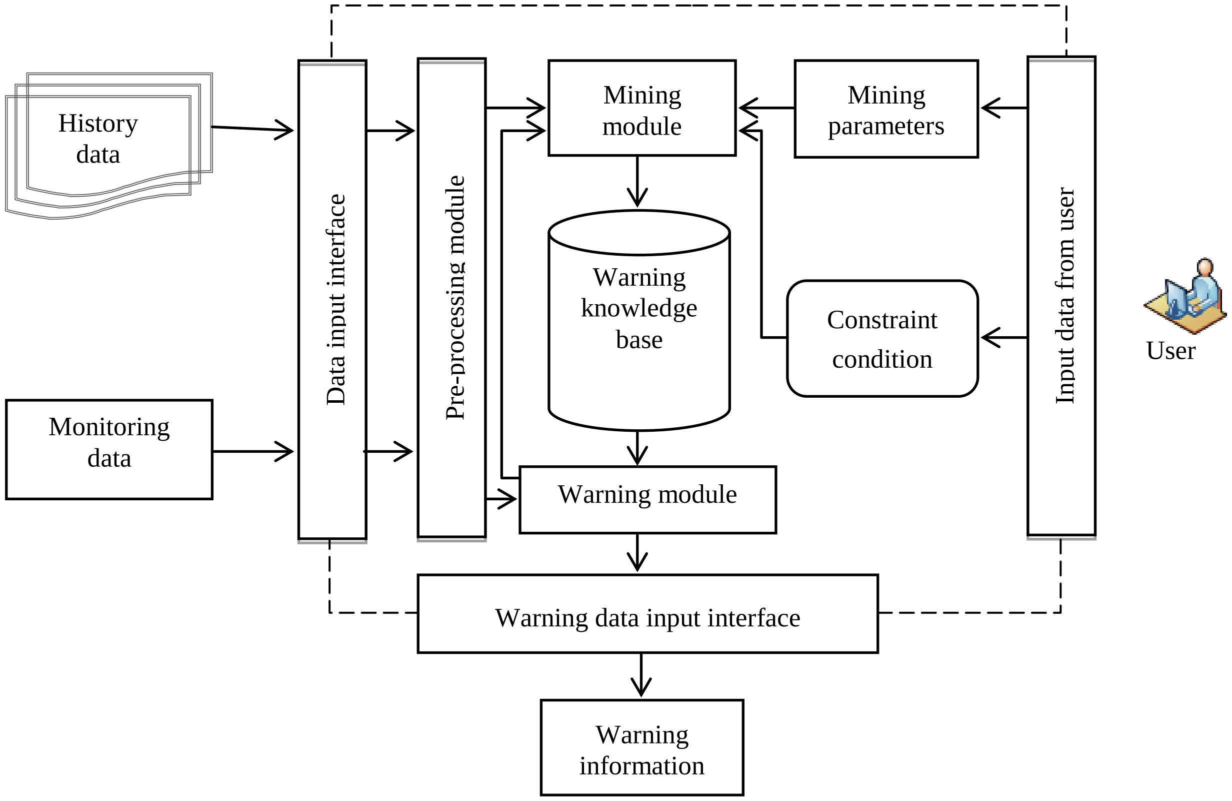
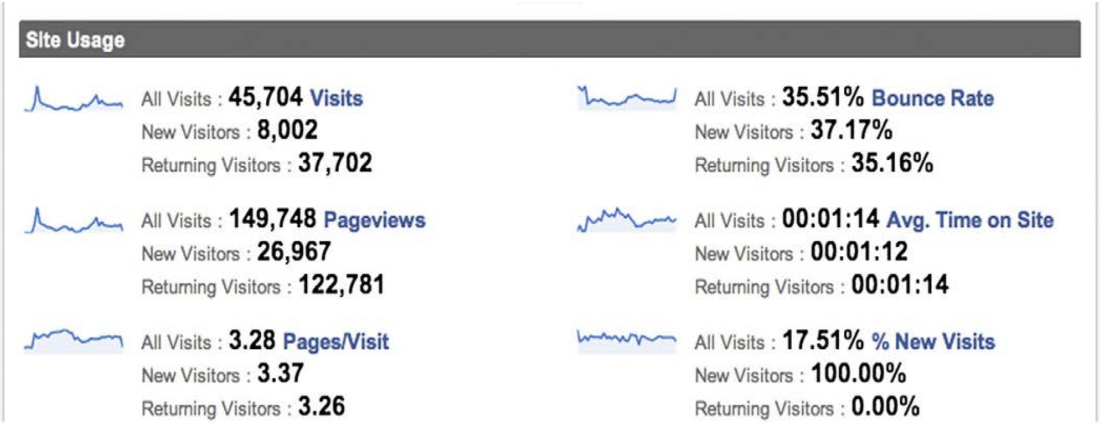


Figure 1: Framework of the early warning technology.

According to the early warning technology tracking reports, user accesses related to learning increased dramatically, from a daily average of 30 visits in February (the month before the campaign) to a daily average of 88 visits during November, followed by a daily average of 102 visits during December (the month after the campaign). After barely recording any user accesses 18 months after the early warning technology was actually launched, once students had easy, direct access to it, nearly 18,000 visits were recorded within a two-month window (Figure 2).



Site Usage

All Visits : 45,704 Visits
New Visitors : 8,002
Returning Visitors : 37,702

All Visits : 35.51% Bounce Rate
New Visitors : 37.17%
Returning Visitors : 35.16%

All Visits : 149,748 Pageviews
New Visitors : 26,967
Returning Visitors : 122,781

All Visits : 00:01:14 Avg. Time on Site
New Visitors : 00:01:12
Returning Visitors : 00:01:14

All Visits : 3.28 Pages/Visit
New Visitors : 3.37
Returning Visitors : 3.26

All Visits : 17.51% % New Visits
New Visitors : 100.00%
Returning Visitors : 0.00%

Figure 2: Data analytics graph of usage statistics for a sample from the learning management system.

CONCLUSIONS

A method for an early warning of at-risk students in e-learning courses was presented in this article. The study draws on detailed student logs extracted from the learning management system that hosts the e-learning courses, to dynamically make estimations and relate them to student progress throughout the course [10]. The goal of this research, which used data mining, was to explore which factors influenced the students' performance. The empirical study validates the proposed research model. Finally, the causal relationships resulting in these phenomena were identified.

The method uses data mining technology. To overcome inaccuracies in identifying at-risk students, the author combined their estimations using different schemes. The proposed method is expected to facilitate instructors to promptly identify at-risk students and focusing on their needs; thus, increasing e-learning retention rates.

As an observation, time-invariant student data were found to be less accurate predictors of a student's decision to drop out compared to time-varying data, obtained as the course progresses. Using time-invariant data, the overall accuracy, sensitivity and precision rates achieved were 43-52%, 61-64% and 63%, while higher experimental results were acquired using time-varying attributes, that correspond to 75-82%, 70-74% and 64-88% in early course stages and 97%, 95-100% and 100% in latter stages, respectively.

Using the scheme, the proposed method achieved a 75-85% overall student classification rate of four courses, to reach a 97-100% rate of two courses.

The results concerning the sensitivity and precision criteria were also high, indicating that the scheme was accurate in both correctly identifying dropouts and avoiding completer misclassifications.

In future, the possible application of the proposed at-risk prediction method on other types of courses besides e-learning, such as blended learning, m-learning (mobile learning) and MOOC (massive open on-line course) will be investigated. Another issue, which should be investigated, is the potential for achieving better results using different student attributes. Moreover, more techniques also could be examined in terms of individual and combined performance for at-risk prediction. Finally, another issue to be investigated in the future, is the potential for incorporating the proposed method in student-retention strategies at educational institutions.

ACKNOWLEDGEMENT

This work is supported by the soft-scientific project of the Henan Province (No. 142400410425).

REFERENCES

1. Liaw, S.S., Huang, H.M. and Chen, G.D., An activity–theoretical approach to investigate learners' factors toward e-learning systems. *Computers in Human Behavior*, 23, **4**, 1906-1920 (2007).
2. Costu, B., Comparison of students' performance on algorithmic problems. *J. of Science Educ. and Technol.*, 16, **5**, 379-386 (2007).
3. Vare, J.W., Dewalt, M.W. and Dockery, R.E., Predicting student retention in teacher education programs. *Proc. Annual Meeting of the American Assoc. of Colleges for Teacher Educ.*, Chicago, USA (2000).
4. Ueno, M., On-line statistical outlier detection of irregular learning processes for e-learning. *Proc. World Conf. on Educational Multimedia, Hypermedia and Telecommunications*, 227-234 (2003).
5. Herzog, S., Estimating student retention and degree-completion time: decision trees and neural networks regression. *New Directions for Institutional Research*, 2006, **131**, 17-33 (2006).
6. Ayed, M.B., Ltifi, H., Kolski, C. and Alimi, A.M., A user-centered approach for the design and implementation of KDD-based DSS: a case study in the healthcare domain. *Decision Support System*, 50, **1**, 64-78 (2010).
7. Chen, Y., Spangler, S., Kreulen, J., Boyer, S., Griffin, T., Alba, A., Behal, A., He, B., Kato, L., Lelescu, A., Kieliszewski, C., Wu, X. and Zhang, L., SIMPLE: a strategic information mining platform for licensing and execution. *Proc. 2009 IEEE Inter. Conf. on Data Mining Workshops*, 270-275 (2009).
8. Hosseini, M.S., Maleki, A. and Gholamian, M.R., Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37, **7**, 5259-5264 (2010).
9. Romero, C., Ventura, S., De Bra, P. and De Castro, C., Discovering prediction rules in AHA courses. *Proc. User Modelling Conf.*, 35-44 (2003).
10. Holder, B., An investigation of hope, academics, environment, and motivation as predictors of persistence in higher education online programs. *The Internet and Higher Educ.*, 10, **4**, 245-260 (2007).