Mining Event Logs of Web-Based Educational Systems for Predicting Student Performance

Amlan Mukherjee
University of Washington
amlan@hitl.washington.edu

Submitted in partial completion of *CSE590D (Winter '04)*

March 7, 2004

1 Introduction

In this brief note, I have discussed a paper authored by Minaei-Bidgoli et. al. (2003) that investigates data mining techniques used to predict student performance, by identifying user patterns in the vast quantities of student interaction data collected from web-based educational systems. The research reported here was performed on a part of the latest online educational system developed at the Michigan State University, the *Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA).*

The LON-CAPA system consists of two main databases. The first one contains educational resources (web pages, demonstrations, simulations, individualized home-work assignments etc.) and the second one contains information about student interaction with the system. This includes log files containing information about how often students access resources, their usage statistics and their pattern of correct and incorrect responses. In this study, the data mining techniques were applied to these databases.

2 Research Goals

There were two main goals of this research:

- Classify students into groups based on how similarly they use the available educational resources. This will allow instructors to appropriately classify individual students early on and identify "at risk" students and help them use the resources better
- Classify problems that have been used by students and allow instructors to identify develop course material more efficiently and effectively

In the following section I have briefly mentioned the methodology used in the research, the results and finally discussed the pertinence of this research to systems like INFACT.

3 Methodology and Results

This paper uses classification techniques to classify students based on how they interact with the LON-CAPA system. To start with, a list of ten features were identified that could be used in the classification process. The list included factors like student success rates at answering questions, time taken to answer questions correctly etc. (check page 2 of paper for complete list of factors).

The understanding that no single classifier can classify all aspects of the dataset with an acceptable level of accuracy, and also, that different classifiers are more effective at classifying different aspects of the training datasets, lead the researchers to use a combination of multiple classifiers (CMC). The range of classifiers used included Quadratic Bayesian classifier, 1-nearest neighbor, k-nearest neighbor, Parzen Window, Multi-Layer Perception and Decision Tree. The *online* CMC method (the class getting the maximum votes from the individual classifiers is assigned to the test sample) was used. Using CMC significantly increased the accuracy in the classification process.

The researchers used Genetic Algorithms (Randomized algorithms used for finding optimal/most fit solutions from a set of possible solutions given a particular criteria) to optimize the combination of classifiers. Using a population size of 200, and a Simple Genetic Algorithm (SGA), they developed a fitness function that measures the error rate (percentage of examples misclassified) achieved by the CMC at classifying the samples. The objective was to minimize the error rate.

Experimental results showed that there was a significant (at least 10%) increase in prediction accuracy, in all classification categories, when using GA optimized CMC for student classification.

4 Discussion

The take from this paper pertinent to INFACT is as follows:

- It is a pointer to how existing INFACT event logs can be used to evaluate and predict student performance by classifying their interaction patterns. There already exist very detailed time-stamped student interaction data and it would be interesting to come up with a list of features similar to the ones described in this paper to classify students. For example, how often do the students erase an entity, length of time interval between when a student starts a drawing and when they first erase an entity and so on. This could be very useful in classifying and evaluating student performance and also for predicting future performance.
- The GA optimized CMC classification method described in this paper can be implemented for the INFACT event logs. Given that there are many similarities in the nature of the LON-CAPA and INFACT event log datasets, this may prove to be an appropriate technique.

Reference

Minaei-Bidgoli, B., Kashy, A. D., Kortemeyer, G. and Punch, F. W. (2003) "Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA" 33rd ASEE/IEEE Frontiers in Education Conference, November 5-8, 2003, Boulder,CO.