Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100

Title, Authors, Source, Year

Title: Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100

Author: Erkan

Source: http://www.ijmlc.org/show-32-132-1.html

Year: 2012

Summary:

This paper discusses a potential way to predict student performance and identify at risk students in an online course. The was complied from the course IS100 and contained only time varying data from their online learning environment. Time invariant data was not used in the training to stage to train the models. The study uses three learning algorithms, including instance-based learning Classifier (K-star), decision tree (C4.5) and naïve bayes. Three additional decision schemes were used to combine the results of the different models. Each student is classified into two groups, successors and failures. The study used a step-based approach to train the various models allowing for the most accurate classifications as early as possible. The steps included, 1st step: Attendance information for first four weeks, grade of 1st assignment, 2) 2nd step: Attendance information for first seven weeks, grade of 1st, 2nd assignments, midterm grade 3) 3rd step: Attendance information for first ten weeks, grade of 1st, 2nd and 3rd assignments, final exam grade, midterm grade. The three decision trees varied by the number of machine learning models required to consider a student a failure. DS1 required 1 model, DS2 required 2 models and DS3 required 3.

The results of the experiment were evaluated on overall accuracy, sensitivity and precision. Overall, the results showed DS3 achieved the highest accuracy after the first step, with an accuracy of 65%. Both naïve bayes and K-star achieved an accuracy of 60%. In step 2 DS3 continued to perform well at an accuracy of 75% but DS2 achieved a higher accuracy of 78%. DS2 continued to achieve the highest accuracy of 85%. K-star achieved the highest accuracy among the ML models.

The paper shows overall that the exclusion of time invariant data has no significant impact on the final result and time varying data is enough to get accurate results.

Why did I read this paper?

I read this paper to see the impact time invariant data has on the overall results of the classifications and if time varying data can be used solely instead.

Personal view:

In my opinion I thought this was a good paper which covered an interesting research question. The paper helped me understand the requirement of time invariant data to classify students and how effective different ML models are for this problem.

What problem does this paper address?

The paper helps to solve the problem on predicting students' performance early to allow appropriate help to be given. It also asks the question on how important time invariant data is in classifying students and if time variant can be used instead.

Is it an important problem?

This problem is hugely important right now. With class sizes increasing, less personal support can be given requiring systems like these in place to prevent students falling behind. It also shows the significance of the type of data you use to train models, especially if time invariant data will unlikely be given to create similar systems.

What is the significance of the result and its solution?

The results show time invariant data has little impact on the overall result.

What are the claimed novel contributions of the paper?

The use of time invariant data to predict the risk of students falling behind.

What previous work is the basis for this research?

One of the initial papers that utilised ML in solving this problem. S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques.

Another paper this study takes influences from is I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, 2009. "Dropout prediction in e-learning courses through the combination of machine learning techniques. This paper reports that using decision scheme increased the accuracy in student classification to 97-100%.