Predicting Students’ Performance by ML

Subjects: Computer Science, Artificial Intelligence

Contributor: Juan A. Gómez-Pulido

Predicting students' performance is one of the most important topics for learning contexts such as schools and universities, since it helps to design effective mechanisms that improve academic results and avoid dropout, among other things. These are benefited by the automation of many processes involved in usual students' activities which handle massive volumes of data collected from software tools for technology-enhanced learning. Thus, analyzing and processing these data carefully can give us useful information about the students' knowledge and the relationship between them and the academic tasks. This information is the source that feeds promising algorithms and methods able to predict students' performance.

Keywords: machine learning ; recommender systems ; technology-enhanced learning ; prediction ; forecasting ; e-learning ; education ; predicting students' performance ; collaborative filtering ; artificial neural networks

1. Introduction

There is often a great need to be able to predict future students' behavior in order to improve curriculum design and plan interventions for academic support and guidance on the curriculum offered to the students. This is where Data Mining (DM) [1] comes into play. DM techniques analyze datasets and extract information to transform it into understandable structures for later use. Machine Learning (ML), Collaborative Filtering (CF), Recommender Systems (RS) and Artificial Neural Networks (ANN) are the main computational techniques that process this information to predict students' performance, their grades or the risk of dropping out of school.

Nowadays, there is a considerable amount of research and studies that follow along the lines of predicting students' behaviour, among other related topics of interest in the educational area. Indeed, many articles have been published in journals and presented in conferences on this topic.

2. Techniques

The application of techniques such as ML, CF, RS, and ANN to predict students' behavior take into account different types of data, for example, demographic characteristics and the grades from some tasks. A good starting point was the study conducted by the Hellenic Open University, where several machine-supervised learning algorithms were applied to a particular dataset. This research found that the Naïves Bayes (NB) algorithm was the most appropriate for predicting both performance and probability of student dropout [2]. Nevertheless, each case study has its own characteristics and nature, hence different techniques can be selected as the best option to predict students' behaviour.

We have gathered the different techniques into main four groups: supervised ML, unsupervised ML, CF and ANN. An additional group dealing with other DM techniques is added in order to include some works where similar objectives were tackled. Figure 1 shows the weight amount of each of these groups of techniques in the literature, which can indicate the number of problems and cases where each technique is more suitable. In this sense, supervised ML makes up almost half of the cases, followed by CF with a quarter. On the contrary, unsupervised ML has been applied in very few cases.

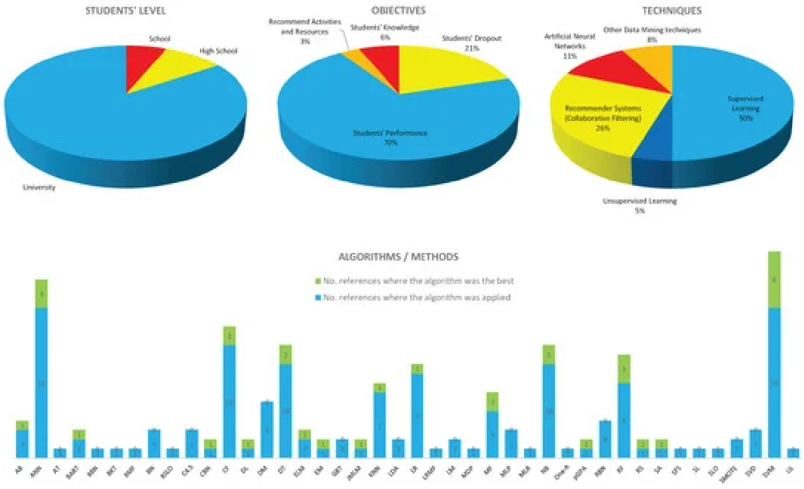

Figure 1. Basic statistics about the techniques, objectives and algorithms tackled in the literature review.

3. Objectives

We have gathered the different objectives into four wide groups: student dropout, students' performance, recommend recommended activities and resources, and students' knowledge. Figure 1 shows the weight of each of these objectives in the literature, which can indicate their importance or interest for research. In this sense, students' performance collect the majority of the prediction efforts (70%), followed by student dropout (21%). Students' knowledge and recommend activities and resources were low-demand objectives (6% and 3% respectively).

4. Discussion

We have noted that there is a strong tendency to predict student performance at the university level, as around 70% of the articles included are intended for this purpose. This may encourage us to consider complementary research efforts to fill gaps in other areas. Thus, we consider that it would be interesting to promote working lines to apply these predictions at school level, which would contribute to identify the low performance of students at early ages. The analysis of student dropout during the early stages of their levels is very interesting, as there are still opportunities to research about helpful predictive tools to enable prevention mechanisms. In this sense, a good approach to research would be to apply the same predictive techniques used for academic performance (and other novel ones) to this case, in addition to considering non-university levels.

Based on the data collected in this review, the most widely used technique for predicting students' behavior was supervised learning, as it provides accurate and reliable results. In particular, the SVM algorithm was the most used by the authors and provided the most accurate predictions. In addition to SVM, DT, NB and RF have also been well-studied algorithmic proposals that generated good results.

Recommender systems, in particular collaborative filtering algorithms, have been the next successful technique in this field. However, it should be clarified that success has been more in recommending resources and activities than in predicting student behavior.

As for the neural networks, they are a less used technique, but they obtain a great precision in predicting the students' performance. We believe that a good line of research with these techniques would be to apply them to other related types of predictions in the educational field, different from the strict students' performance.

We emphasize that unsupervised learning is an unattractive technique for researchers, due to the low accuracy of predicting students' behavior in the cases studied. However, this fact can be an incentive for research, as it provides the opportunity to further improve these techniques in order to obtain more reliable and accurate results.

References

1. Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, 3rd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2011.
2. Kotsiantis, S.; Pierrakeas, C.; Pintelas, P. Predicting Students’ Performance in Distance Learning using Machine Learning Techniques. Appl. Artif. Intell. 2004, 18, 411–426.