Towards Freshmen Performance Prediction

Hana Bydžovská
CSU and KD Lab Faculty of Informatics
Masaryk University, Brno
bydzovska@fi.muni.cz

ABSTRACT

In this paper, we deal with freshmen performance prediction. We analyze data from courses offered to students at Faculty of Informatics, Masaryk University. We supposed that the success rate of our predictions increases when we omit freshmen from our experiments as we have no study-related data about them. However, we disproved this hypothesis because there was generally no significant difference in prediction of freshmen and non-freshmen students. We also presented the attributes that were important for freshmen performance prediction.

Keywords

Student Performance, Prediction, Freshmen, Social Network Analysis, Educational Data Mining.

1. INTRODUCTION

Universities are faced with the problem of a high number of students' drop outs. Thus, researches explore what influences students' performance, and identify weak students in order to help them to improve their achievements. It is important to predict student failure as soon as possible. The task is difficult because the less data about students we have the less accurate the prediction we obtain is.

Data mining techniques represent a typical way for discovering regularity in data [3]. It allows us to build predictive models by defining valid and exact corresponding rules. Authors in [2] explored the drop-out prediction after the first year at Electrical Engineering department. Their data contained the study results of students enrolled in selected courses or the average grades gained in different groups of courses. Their results showed that decision trees belong to the most suitable algorithms. They also demonstrated that the cost-sensitive learning methods helped to bias classification errors towards preferring false positives to false negatives. Authors in [4] also investigated the prediction after the first year. They used questionnaires to get more detailed information about student habits.

We are interested in a similar problem but our task involves the prediction of student success in a course not in the whole study. Our aim is to identify the combinations of students and courses that could be predicted with a high accuracy. Due to the lack of data, we supposed that omitting freshmen (students in the first semester in their first study at the faculty) from the investigation should significantly increase the prediction accuracy. We also investigated how accurately we are able to predict the success or failure of freshmen.

2. DATA

The data used in our experiment originated from the Information System of Masaryk University. Our aim was to reveal useful attributes characterizing students in order to predict student performance in every particular course. Our data comprised of study-related and social behavior data about students. We explored the freshmen performance prediction and the observations were verified on 62 courses offered to students of the Faculty of Informatics of Masaryk University. The data sets comprised of students enrolled in courses in the years 2010-2012 and their grades. We constructed three data sets: (1) All students – 3,862 students with 42,677 grades, (2) Without freshmen – 2,927 students with 32,945 grades, (3) Only freshmen – 935 students with 9,732 grades.

2.1 Study-related data

This kind of data contained personal attributes (e.g. gender, year of birth, year of admission at the university) and data about study achievements (e.g. the number of credits to gain for enrolled, but not yet completed courses, the number of credits gained from completed courses, the number of failed courses). This data contained 42 different attributes in total.

2.2 Social Behavior Data

This kind of data described students' behavior and co-operation with other students. In order to get additional social attributes, we created sociograms. The nodes denoted users and the edges represented ties among them. The ties were calculated from the communication statistics, students' publication co-authoring, and comments among students. Particularly, we applied social network analysis methods on the sociograms to compute the values of attributes that represent the importance of each user in the network, e.g. centrality, degree, closeness, and betweenness. We also calculated the average grades of students and their friends. Finally, the social behavior data contained 131 attributes in total. We already proved that this type of data increases the accuracy of student performance prediction [1].

3. EXPERIMENT

Hypothesis. The accuracy of the student success prediction will significantly increase when we omit freshmen.

Evaluation. We utilized nine different classification algorithms implemented in Weka. We built a classifier for each investigated course because courses differ in their specialization, difficulty, and student occupancy rate. In the first place, we had to select suitable methods and compare the results of data sets with and without freshmen. We used the accuracy and coverage for comparing the results. Generally, the accuracy represents the percentage of correctly classified students. The coverage represents the amount of students for whom we can predict the success or failure.

Observations. In all cases, SMO reached the highest accuracy (with and without freshmen). We computed also baseline (the prediction into the majority class) in order to compute the percentage of successful grades. In all cases, we used 10-fold cross-validation for evaluation the results. The results comparison can be seen in Table 1. Surprisingly, the results indicate that there is no significant improvement when we omit the freshmen. We improved the results only by 1% but for almost 10,000 grades we did not give any prediction.

Table 1. Comparison of results with and without freshmen

ALL COURSES	Accuracy		Coverage
	SMO	Baseline	
All students	80.04%	73.45%	100%
Without freshmen	81.26%	75.79%	77.2%

Naturally, the increase can be distorted by the large amount of non-freshmen students. No freshman has enrolled in 8 courses. Less than 10 freshmen were enrolled in 22 courses. Moreover, freshmen did not constitute 10% of all students in the next 18 courses. For the next investigation we selected only 14 courses where the number of freshmen is not negligible.

The results of selected 14 courses can be seen in Table 2. As can be seen, the improvement was 3.3%. However, there was a significant difference in baseline – about 7%. SMO was the most suitable method again but the results were difficult to interpret. For this reason, we also presented the accuracy using J48 for the purpose of comparison the success rate of the both approaches. We considered the J48 model to be similar enough for indication the attributes that influenced the results.

Table 2. Comparison of results for 14 courses

14 COURSES	Accuracy			Coverage
	SMO	J48	Baseline	
Without freshmen	82.07%	80.24%	77.82%	59.27%
All students	78.77%	77.48%	70.66%	100%
Only freshmen	76.56%	75.10%	67.11%	40.73%

When comparing the results presented in Table 1 and Table 2, freshmen decreased the overall accuracy in all cases. However, the difference was insignificant. The model based on J48 algorithm was explored for each course. We also investigated trees built only for the freshmen. The classifiers classified the data based on using the following attributes:

Known study-related attributes: field of study, programme of the study, if the student passed the entrance test or the student was accepted without taking any entrance test, score of the entrance test, if the course is mandatory, elective, or voluntary for the student.

Social behavior attributes: degree, centrality, betweenness, number of friends / average of grades of friends that already passed investigated course, number of friends / average of grades of friends that are enrolled in the course with the investigated student.

It was very interesting that the freshmen can be characterized by social attributes. They got the access to the system in June during the enrollment to their studies. During the enrollment of courses in September when we investigated their probability to pass the courses, we already had some data about their activity in the system. In order to measure the influence of the social behavior data on the freshmen performance prediction, we removed different types of data from the data set. The comparison can be seen in Table 3. SMO reached all presented results. The accuracy obtained by mining social behavior attributes was surprisingly slightly better than by mining only known study-related attributes. The best result was obtained when we used the both data types together.

Table 3. Freshmen performance prediction using different types of data

Data set	Accuracy
All attributes	76.56%
Only known study-related attributes	73.95%.
Only social behavior attributes	74.72%

Decision. The results indicated that the accuracy of the prediction was almost the same for all students regardless the status of freshmen. The freshmen passed through the similar classification paths as the non-freshmen. When we consider only the courses with a high proportion of the freshmen, the difference is higher but not significant. As a result, the hypothesis was not confirmed.

4. CONCLUSION

In this paper, we were dealing with the freshmen performance prediction. The hypothesis was that the success rate of the predictions will increase when we omit the freshmen. We disproved this hypothesis because the results sustained almost the same. The freshmen passed through the similar classification path as the non-freshmen. When we inspected the possibility of estimation only the freshmen grades, surprisingly, mining the social behavior data collected from students in the information system only in two months reached better results than mining data about results in the entrance test, course category, and the basis of the study specialization.

5. REFERENCES

[1] Bydžovská H. and Popelínský L. 2014. The Influence of Social Data on Student Success Prediction. *In Proceedings of the 18th International Database Engineering & Applications Symposium*, pp. 374-375 (2014)

[2] Dekker, G.W. and Pechenizkiy, M. and Vleeshouwers, J.M. 2009. Predicting students drop out: a case study. In T. Barnes et al. (eds.), *Proceedings of the 2nd International Conference on Educational Data Mining* (EDM'09), pp. 41-50.

[3] Marquez-Vera, C. Romero, C. and S. Ventura. 2011. Predicting school failure using data mining. *In Proceedings of the 4th International Conference on Educational Data Mining (EDM'11)*, pp. 271-276.

[4] Vandamme, J.P. and Superby, J.F. and Meskens, N. 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop*, pp. 37-44.