Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education

Conference Paper · April 2015

CITATIONS
9

READS
1,864

5 authors, including:

Vlatko Nikolovski
Ss. Cyril and Methodius University in Skopje
5 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Riste Stojanov
Ss. Cyril and Methodius University in Skopje
35 PUBLICATIONS **111** CITATIONS
SEE PROFILE

Igor Mishkovski
Ss. Cyril and Methodius University in Skopje
68 PUBLICATIONS **383** CITATIONS
SEE PROFILE

Ivan Chorbev
Ss. Cyril and Methodius University in Skopje
114 PUBLICATIONS **846** CITATIONS
SEE PROFILE

Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education

Vlatko Nikolovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
vlatko.nikolovski@finki.ukim.mk

Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
riste.stojanov@finki.ukim.mk

Igor Mishkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
igor.mishkovski@finki.ukim.mk

Ivan Chorbev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
ivan.chorbev@finki.ukim.mk

Gjorgji Madjarov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
gjorgji.madjarov@finki.ukim.mk

***Abstract*— reducing the student dropout rate in higher education is one of the challenges that universities are dealing with. By providing enriched study programs and qualified teams of professors, universities aim to enroll more students. Improving working conditions at the laboratories and other university resources aims to attract both ambitious students as well as high quality staff. After enrollment, the main goal of the faculty is to guide students into successful completion of their studies with the appropriate knowledge and skills acquired. Nowadays however, the development and deployment of Student Information Systems at the universities provides an appropriate infrastructure for student's data organization and storage as well as data acquisition and deeper analyses. This data can help model the behavior of dropouts, and predict future dropouts, therefore giving chance to counselors to advise and guide students into success.**

This paper presents various data mining experiments and results obtained from data for the students from one of the faculties at the University Ss. Cyril and Methodius in Skopje. Initially, we give an overview of several data mining algorithms suitable for analysis of students' data and dropout prediction. Furthermore, we explain modifications and applications of the algorithms over the existing student data. Finally, we provide a predictive model which will identify a subset of students who tend to drop out of the studies after the first year. The classification task aims to identify a pattern among students who tend to drop out.

Keywords*— *Educational Data Mining, Student Dropout Prediction, Machine Learning Algorithms, Classification

I. INTRODUCTION

The use of Data Mining techniques in Educational data sets, known as Educational Data Mining [1] is a relatively new field of research. Educational Data Mining provides tasks for clustering, prediction, relationship mining (subset of sequential mining, association and correlation), social related mining for human behavior and discovery with pattern models [2]. In addition, the methodology of Educational Data Mining is not yet clearly defined and there are no clear standards about which data mining algorithms are preferable in this context. Clustering and classification techniques of data sets for building a predictive model are used in [3]. Another variation of improved data classification with cost-sensitive learning is presented in [4]. Popular association analysis is presented in [5], while neural networks and Bayesian networks analyses are used in [6]. Classification and data retention techniques similar to ours are presented in [7, 8, and 9] to predict the dropout rate of the students in their first year of study.

In our case study, we used data collected over the period of three years, starting from 2011 to 2014, containing details about the students, their course retention and grade evaluation. Divided into three subcategories, based on the study year of enrolment in the faculty, each subset contains about 680 students. The classification techniques were applied over the three datasets, based on different attributes, such as: nationality of the students, sex, city of living, high school grades, study program enrolled, number of earned credits in the first year of study, an average grade in the first year of study. In addition,

we divided the first year courses into two categories, a subset of mathematics courses and a subset of programming courses. So, the attributes pool for data classification was enriched with details for average grade and number of exams applied for each subset of courses.

II. Background and Related Work

Data mining techniques are becoming an essential way of transforming data into linked usable information, extracting unexpected knowledge and discovering numerous patterns among large data sets. Due to the omnipresent implementation of various Information Systems by the faculties for logging and processing students` data, Data mining techniques can be applied to estimate unanticipated relationships among attributes of students, correlation between learning strategies and assessments. Aside from the traditional statistical methods for extracting and processing most valuable information from the large datasets, Data mining techniques provide a huge potential for knowledge discovery since they embrace numerous disciplines such as machine learning and artificial intelligence into an advanced technique for estimating large datasets. Furthermore, these combined techniques produce predictive analysis for identifying interconnections and variables regarding the context of the study [10].

Back at the beginning of this research area, Tinto [12] proposed a model of a theoretical framework for considering factors in academic success. Tinto made a correlation between the students and the faculties, considering the process of student enrolment as a sociological interplay between the characteristics of the student and the experience at the faculty. Furthermore, this interaction between the students` past and present environment leads to a degree of integration of the student into the faculty environment. Based on this model, the integrity of the institution directly depends on the quality of teachers and studies, providing an environment for the new students.

Since the methodology of educational Data mining is not yet transparent, researchers have used various techniques for estimating preferable algorithms in this context. Clustering the datasets in a manner of classification and transformation techniques to provide a considerable predictor is presented in [3]. As presented in [6], using neural networks and Bayesian networks over small datasets are outperformed by decision tree algorithms.

Beside the traditional approach, Diego [13] proposed a meta-algorithm for pre-processing the data before classification, which improves the accuracy of the model. Different techniques are presented in [11] for reducing inaccuracies in prediction of the students` dropout. In addition, the comparison of the three techniques that were used, namely neural networks, support vector machines and probabilistic fuzzy ARTMAP maintained that the most successful technique in predicting students` dropout is the decision scheme.

III. Methodology

This paper presents the usage of classification data mining techniques over the students` data to analyze and extract the important attributes affecting the dropout of students in higher education. In addition, two classifier algorithms were used, J48 [15] and Native Bayes [15] implemented in Weka (Data Mining Software) [14].

The Naïve Bayes classifier is based on the Bayes rule of conditional probability. It analyzes all the contained attributes individually as though they are equally important and independent of each other. In the process of classification, each attribute works independently from the other attributes contained in the model.

Besides the Naïve Bayes independent treatment of the attributes, J48 is a predictive machine-learning model that predicts the attribute as a dependent variable from the values of all other attributes. In order to classify a new item, J48 first creates a decision tree based on the attributes of the training data in order to gain balance, flexibility and accuracy.

IV. Data Mining process

Figure 1 shows the process of acquisition of the students` data, the process of transformation and evaluation of the attributes extracted from the data. It also shows the estimation and evaluation of the classified data and the improvement of the model for prediction and decrease of the dropout rate.

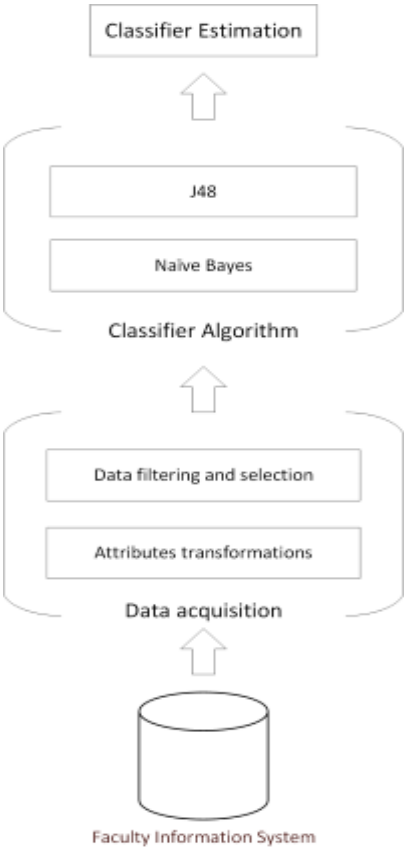

Fig. 1. Data retention process

A. Data acquisition

The dataset used in this case study was collected from the Students` Information System at one of the faculties at the Ss. Cyril and Methodius University – Skopje. In addition, the collected data over the period from 2011 to 2014 contains information about all students being enrolled. In addition, a target dataset of 2029 students was selected from three different generations, with attribute values only for their first year of study at the faculty.

B. Data preparation and attribute selection

In this step, three main datasets were considered shown in table 1: a dataset with students who were enrolled at 2011, containing 665 instances; a dataset with students who were enrolled at 2012 with 679 instances and a dataset with students enrolled at 2013 with 685 instances. Each dataset contains same attributes described in table 2.

TABLE I. DATASETS OF STUDENTS

2011	2012	2013
665 students	679 students	685 students

[a.] Total of 2029 instances

Estimation assumed that the first year of study is the most critical for the students' dropout, so the attribute values are based only on the first year of study of the students. Table 2 shows the attribute retention mechanism for grouping students into categories based on several criteria.

TABLE II. ATTRIBUTES IN THE DATASET

Attribute	Type	Possible values
PreviousGrade	Numeric	{2,3,4,5} – pre-university education curriculum
AvgGrade	Numeric	{5,6,7,8,9,10} – average grade from exams passed in first year of study (FYoS)
SumCredits	Numeric	{0-70} – number of credits enrolled in FYoS
MathematicsAvg	Numeric	{5,6,7,8,9,10} – average grade from mathematics courses enrolled in FYoS
MathematicsCount	Numeric	{0-36} – number of exam applications for mathematics courses enrolled in FYoS
ProgrammingAvg	Numeric	{5,6,7,8,9,10} – average grade from programming courses enrolled in FYoS
ProgrammingCount	Numeric	{0-36 } – number of exam applications for programming courses enrolled in FYoS
LivingPlace	Nominal	{east,west,skopje} – estimated regions based on a geographic basis of the cities in Republic of Macedonia
Nationality	Nominal	{mk,notmk} – estimated values based on nationality of the student
StudyProgramme	Nominal	{Kni,Pet,Mt,Knia,Ke,Iki,Asi,Info,Pit} – study programme of enrolment
Sex	Nominal	{m,f}

[b.] Total of 2029 instances

- Mandatory courses from the first year of study are divided into two categories, mathematical and programming. Based on this classification, each student dataset contains values for average grade and number of exam applications for each category of courses.
- Living place of the students provides another group of students, organizing the geographic locations of the cities in Republic of Macedonia to east region, west region and the region of Skopje – the capital.
- Another group of classification defines the nationality of the students, divided into students who have Macedonian nationality and those who are not ethnic Macedonians.
- The last group defines the study programme in which the student was enrolled at the first year of study at the faculty.

C. Implementation of mining model

The mining model was created based on the attributes described earlier. Initially, a training set was created from the subsets of students enrolled in 2011 and 2012. After the model was evaluated in Weka, the subset of students enrolled in 2013 was supplied as a test set. After the model was completely evaluated and predictions were estimated, we compared the predicted data with the real data obtained from the Students` Information System. The estimation of the model and predictions were made with both J48 and Naïve Bayes classifier algorithms.

D. Result analysis and discussion

Classification accuracies for the dataset containing all three subsets of student enrolment retentions are shown in table 3. The presented results indicate that the Data mining with J48 algorithm is more accurate than the Data mining with Naïve Bayes classifier algorithm.

TABLE III. ACCURACY AND RATES OF TOTAL DATASET

Classifier	J48		Naïve Bayes	
Accuracy	81.1679 %		76.7833 %	
Class signedOut	Yes	No	Yes	No
TP Rate	0.959	0.771	0.918	0.727
FP Rate	0.229	0.041	0.273	0.082
Precision	0.534	0.986	0.479	0.97
Recall	0.959	0.771	0.918	0.727
F-Measure	0.686	0.865	0.629	0.831
ROC Area	0.933	0.933	0.915	0.915

The same classification techniques were applied to the datasets containing both the real data and predicted implications. This comparison aims to achieve the reasons for students' dropout and to predict the circumstances in which the student needs attention.

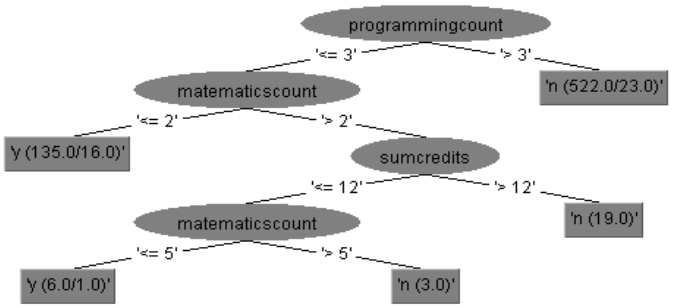

Fig. 2. Decision tree of critical variable values for students` dropout

The presented results in fig 2, emphasize the significant fact that the most critical variables for estimating the students` dropout in their first year of study are the number of exam applications for both mathematics and programming courses.

To get a better insight over the dropout analyses, Appendix A shows the correlation between the nominal attributes that we had extracted from the data and the social and demographic characteristics of the student. From the results, the dropout rate is higher among the students from other nationalities than Macedonian. Also, the dropout rate is higher among the students that are from the western regions of Macedonia.

V. CONCLUSION AND FUTURE WORK

Student dropout prediction is an important and challenging task, yet not clearly defined. This paper presents the accuracy of both J48 and Naïve Bayes classifier algorithms for data mining over the educational data collected from the Students` Information system at one of the faculties at the Ss. Cyril and Methodius University – Skopje.

Our case study shows that the accuracy of the different classifier algorithms notably depends on the quality of the attributes extracted from the data. The accuracy of the classifiers is tied closely with the quality and sophistication of the data model.

According to the results, the most valuable attributes for the prediction are the number of exam applications for both mathematical and programming courses. Since the results reveal a pattern among the number of exam applications between the mathematical and programming courses, equally important are the demographic characteristics of the students.

The model evaluation techniques of classification presented in this paper point to two major improvements that can be noticed. First, the courses grouping into subsets based on the field of study brings a big improvement into the data mining process. In addition, a better encoding grades scheme is required for the students not involved into a specific course. The second and final remark points out the quality and size of the students` dataset. Furthermore, the results are better if the dataset is bigger and well organized.

REFERENCES

[1] Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P. Mining the student assessment data: Lessons drawn from a small scale case study. In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187-191, 2008.

[2] Romero C and Ventura S, "Educational Data Mining: A survey from 1995 to 2005" expert system with Application 33(2007) 135-146

[3] Luan, J. Data mining and its applications in higher education. New Directions For Institutional Research, p. 17-36, Spring 2002.

[4] Romero, C.; Ventura, S.; Espejo, P.G.; Hervas, C. (2008) Data Mining Algorithms to Classify Students. In Proceedings of the First International Conference on Educational Data Mining, (pp. 8-17).

[5] Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P. Mining the student assessment data: Lessons drawn from a small scale case study. In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187-191, 2008.

[6] Herzog, S. Estimating student retention and degreecompletion time: Decision trees and neural networks vis-avis regression, New Directions for Institutional Research, p. 17-33, 2006.

[7] Gerben W. Dekker , Mykola Pechenizkiy, Jan M. Vleeshouwers, Predicting Students Drop Out: A Case Study, Educational Data Mining, 2009

[8] Dr. Saurabh Pal, Mining Educational Data Using Classification to Decrease Dropout Rate of Students, INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, VOL. 3, NO. 5, MAY 2012

[9] Mario Jadrić, Željko Garača, Maja Ćukušić, STUDENT DROPOUT ANALYSIS WITH APPLICATION OF DATA MINING METHODS, 2010

[10] Su, J.-M.; Tseng, S.-S.; Wang, W.; Weng, J.-F.; Yang, J.T.D. and Tsai, W.-N. (2006). Learning Portfolio Analysis and Mining for SCORM Compliant Environment. In Educational Technology & Society, 9(1), (pp.262-275).

[11] Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G. and Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. In Computers & Education, 53(3), (pp. 950-965).

[12] Tinto, V. Limits of theory and practice in student attrition, Journal of Higher Education 53, p. 687-700, 1982.

[13] D. Carcia-Saiz and M.E. Zorrilla, Comparing Classification Methods for Predicting Distance Students' Performance, Journal of Machine Learning Research – Proceedings Track, Vol.17, 2011, pp.26-32.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[15] J. Han and M. Kamber, Data Mining: Concepts and Techniques, second edition, Morgan Kaufmann, 2006.

APPENDIX A. DECISION TREE