Using Classification Methods in Higher Education

Alba Çomo, Ilia Ninka, Brisilda Munguli

Abstract— currently, in Albania, there is an increased interest in data mining and educational system, making educational data misning a new growing research community. Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict which students will or will not graduate. The university can use this information to concentrate on those students that are most at risk. In this paper we attempt to use data mining process to help in enhancing the quality of the higher education system by evaluating student data. For this purpose we have collected data from a public faculty in Tirana from 2011 to 2014, covering 1300 students. The classification process is based on the decision tree as a classification method where the generated rules are studied. We aim to build a system that will facilitate the use of generated rules, which will allow students to predict the average grade of the next year in university.

Index Terms—student data, data mining, decision trees, higher education, classification process

I. INTRODUCTION

Today, one of the biggest challenges that educational institutions face is the growth of educational data, so the institutions need to use this data to improve the quality of managerial decisions. The prediction of the student success is crucial for these institutions, because the quality of the study and learning process is the ability to satisfy and fulfill the students' needs. The data collected from different systems require proper method of extracting knowledge for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1].

There is increased interest in using data mining in education, also called Educational Data mining. Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting, and using those methods to better understand students, and the settings which they learn in [2]. Educational data mining can mine student grades, student address, enrollment data, and student demographic/social/economic data. Also with data mining the university can predict which students will graduate and those that will not. So the university can use this result to improve the education behavior and student performance.

The main objective of this paper is to use data mining process (Classification process) to study student's performance in different courses. In this research, the decision tree method is used to evaluate student's performance. The data used in this research is collected from the student's management system of a public faculty in Tirana, Albania. Information like student's demographic data, course attendance, course grades, etc. was collected from this system, to predict student performance at the end of each year. In this paper we will use the decision tree method for predicting student performance.

Revised Version Manuscript Received on August 14, 2015.

MSc. Alba Çomo Department of Computer Science, Faculty of Natural Science, University of Tirana, Tirana, Albania.

Prof. Dr. Ilia Ninka Department of Computer Science, Faculty of Natural Science, University of Tirana, Tirana, Albania.

MSc. Brisilda Munguli Department of Computer Science, Faculty of Natural Science, University of Tirana, Tirana, Albania.

II. RELATED WORK

In the last 20 years educational data mining emerged as a new application area for data mining, becoming well established with its own journal. Romero & Ventura [3] provided a survey of educational data mining from 1995-2005 and Baker & Yacef [7] extended their survey covering the latest development until 2009. There is an increasing number of data mining applications in education, from enrollment management, graduation, academic performance, gifted education, web-based education, retention and other areas (Nandeshwar & Chaudhari, [8]). This section reviews only research where the main focus is on study outcome, i.e. successful or unsuccessful course completion.

Romeo & Ventura [3] introduced a survey of the specific application of data mining in learning management systems. Its objective is to introduce it both theoretically and practically to all users interested in the new research area, they describe the full process of how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering and association rule mining of Moodle data. They have describe how different data mining techniques can be used in order to improve the course and students' learning.

Kalles and Pierrakeas [5] discussed different machine learning methods (such as decision tree, neural networks, Naïve Bayes, logic regression, etc.) and compared them with genetic algorithm based induction of decision trees. They analyze students' academic performance through the academic years, as measured by the student's homework assignments, and attempted to derive short rules that explain and predict success and failure in the final exam students.

The research by Honogjie [6] describe how data mining techniques can be used to determine the student learning result evaluation system is an essential tool and approach for monitoring and controlling the learning quality.

The research of Cortez & Silva [9] predicted the secondary student grades of two core classes using past school grades, demographics, social and other school related data. The results were obtained using decision trees, random forests, neural networks and support vector machines.

In other data mining studies, based on enrollment data, the following factors were found to be significant: faculty and

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

nationality and the secondary school science marks.
In conclusion, there is mixed evidence on whether the contribution of background information to the early prediction of student success is significant or not. It depends on the list of variables included, students population and classification methods used.

III. DATA AND METHODOLOGY

In this study we have used data mining process to study students' performance in the courses. Three different classification methods have been tested. The data used in this research is collected from the student's management system of a public faculty in Tirana, Albania. Information collected from this system, is used to predict student performance at the end of each year.

A. Problem definition

The success of the students is a main factor in the progress of the education in Albania. But the statistics show that even though the students win the right to attend the university, they have difficulties to succeed during the academic years. In this regard, we find it necessary to put in categories the characteristics of the students in order to undertake the necessary measurements aiming the students' performance prediction in higher education.

B. The study goal

The main goal of this study is summarized in one question: "How can we use data mining process to help in enhancing the quality of the higher education system by evaluating student data". To find the best solution of this problem, first we have to collect and process the data and then we have to implement data mining techniques to make the most accurate prediction.

C. The importance of selecting the variables

To predict the student performance it is necessary to evaluate many parameters. This study aims to give accurate prediction of students' performance, based on students' academic performance and demographic characteristics (such age, gender and ethnicity).
Demographic characteristics are commonly included in studies of student success. They are easily measured and consistently documented. They facilitate the division of students into groups with some shared identity. Factors, such as age, gender and residence are important; the exclusion of these variables would threaten the validity of the study.
Academic performance – the interface of education (such as grades) creates the construct of the academic environment, which directly/indirectly affects the degree completion or time degree completion.

D. Methods used

Classification is the process of finding a set of models (or functions, rules) which describe and distinguish data classes and concepts, for the purpose of being able to use the method to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification model(if-then-else), decision trees, etc.
We have examined the impact of three data mining algorithms C4.5, Naive Bayes and AdaBoostM1. To estimate the power of the classifier, the usual methodology is to execute K-Cross-Validation method, and to measure the performance of the classifier on the test set, the classification accuracy or error rate are usually used for these purpose. To obtain accuracy of the classification model the WEKA [4], [10] toolkit is used.

IV. DATA MINING PROCESS

The student management system does not provide data in a format ready for an easy and direct statistical analysis and modeling. Therefore a data preparation and cleaning as well as creation of variables of analysis were undertaken.
To build a reliable classification model we used a methodology that consists mainly in five steps: collecting data, prepare the data, building classification model, evaluating the model and finally using the model for future prediction of student performance. These five steps are presented below.

A. Data collection

Initially the faculty provided us with 68300 records corresponding to 2100 students. The data supplied was student data at 2 bachelor study programs: Computer Science and Information Technology enrolled through the academic years 2011 to 2014.

B. Data Preprocessing

Initially more than 15 attributes have been collected and some attributes has been manually eliminated since they are considered as irrelevant to the study. The collected data were prepared in a table in a format that it is suitable for the used data mining system. The data are cleansed by removing inconsistent values using the same values for all the data. Finally, we have data for 1321 students.
In this study we have used 8 conditional attributes and one attribute class. The attributes along with their descriptions and possible values are presented in Table I. The class attribute is: ST_AG3 – student average grade in third academic year.
The next step is using the WEKA toolkit, *selection attribute*, for analyzing the chosen attributes, and for these we have executed four tests: Chi-Square, Gain Ratio, OneR and Info Gain. The purpose of this analysis is to determine the importance of each attribute. Table II shows the result of these tests, as we can see the attributes that affects mostly the result are ST_AG1, ST_AG2 and SP. Attribute AC_YEAR and ST_AGE gives the lowest performance. Based on this rank we have eliminated AC_YEAR and ST_ID attribute. Finally, the most significant attributes list contains the following attributes: ST_AG1, ST_AG2, ST_GEN, SP and ST_CIT.

A. Classification model – J48 algorithm

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Table I
The chosen attributes

Attributes	Description	Values
ST_ID	Student id	
Demographic data		
ST_AGE	Student Age	20,21,22
ST_GEN	Student Gender	M, F
REZ	Student residence	A ->Tirana, B-> Other City, C-> Village
ST_CIT	Student citizenship	A -> Albanian, B -> Others
Academic data		
SP	Student study program	CS, IT
AC_YEAR	Student Academic Year	3
ST_AG1	Student Average grade in the first year	A –> 9.00 - 10, B -> 8.00 – 8.99, C -> 7.00 – 7.99, D -> < 6.99
ST_AG2	Student Average grade in the second year	
ST_AG3	Student Average grade in the third year	

Table II
Result of four tests performed

Attribute/Test	Gain Ratio	OneR	Info Gain	Chi Square
ST_AG1	1	1.3	3	2
ST_AG2	2	1.7	2	3
ST_ID	3	9	1	1
SP	4.1	8	4	4
ST_CIT	5.1	7	6.6	6.7
REZ	5.8	6	5	5
ST_GEN	7	5	6.4	6.3
ST_AGE	8	3	9	8
AC_YEAR	9	4	8	9

Decision trees are commonly used for gaining information for the purpose of decision-making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. [4]

In this study we have used C4.5 (J48 in WEKA) algorithm. C4.5 is an algorithm used to generate a decision tree, and is an extension of ID3 algorithm. The decision trees generated by C4.5 can be used for classification and for this reason, is often referred to as statistical classifier.

For class ST_AG3, the attribute that has the highest gain ration was the ST_AG1 (student average grade in first year). This attribute is considered as the root node of the decision tree. The process is repeated for the remaining attributes to build the next level of the tree, and so on. After building the complete decision tree, the classification rules are generated by following all the paths of the tree. The decision tree generated has classification rules, some of them are given in table III. In the first column is given the rule description, in the second column is given the number of instances associates with the leaf and in the third is given the number of instances incorrectly classified.

Table III
Rules generated from decision tree

Rule	#NI	#IIC
IF ST_AG1 = D, ST_AG2 = B, SP = CS, (REZ = A OR (REZ = B, ST_AGE > 20) THEN ST_AG3 = C	11	2
IF ST_AG1 = C, ST_AG2 = C OR ST_AG2 = B THEN ST_AG3 = C	87	9
IF ST_AG1 = B, ST_AG2 = C OR (ST_AG2 = A, SP = IT) THEN ST_AG3 = A	30	5
IF ST_AG1 = D, ST_AG2 = B, OR ST_AG2 = C OR (ST_AG2 = D, SP = IT) THEN ST_AG3 = D	900	62

We have executed the same steps as above, for class ST_AG2 (by removing first the ST_AG3 attributes). This is study is done to predict the average grade of the second year. Below are illustrated some classification rules generated from decision tree build (showed in fig 1).

Rule

IF ST_AG1 = C, SP = IT, ST_GEN = M, REZ = A, ST_AGE > 21 THEN ST_AG2 = A

IF ST_AG1 = A OR (ST_AG1 = C, SP = IT, ST_GEN = F, (REZ = C OR REZ = B) THEN ST_AG2 = A

IF ST_AG1 = B OR ST_AGI = D OR ST_AG1 = A THEN ST_AG2 = ST_AG1

V. RESULTS AND EXPERIMENT EVALUATION

The classification error rate and accuracy are usually used to measure the classification performance. We have used K-Cross-Validation method to measure these values. Table IV shows the evaluation results of three classification methods used: Naïve Bayes, C4.5 (J48 in WEKA) and AdaBoost1 Perceptron. The performances of these models were evaluated based on these criteria: the accuracy of the prediction, the time to build the model and the error rate. Based on the comparisons made (referring to IV) the J48 algorithm offers the best accuracy, the MLP algorithm offers the lowest accuracy.

A good classifier must be both accurate and understandable to its users, whoever they might be. In figure 1, is shown the decision tree resulting from implementation of J48 algorithm on the collected data using ST_AG2 attribute class.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

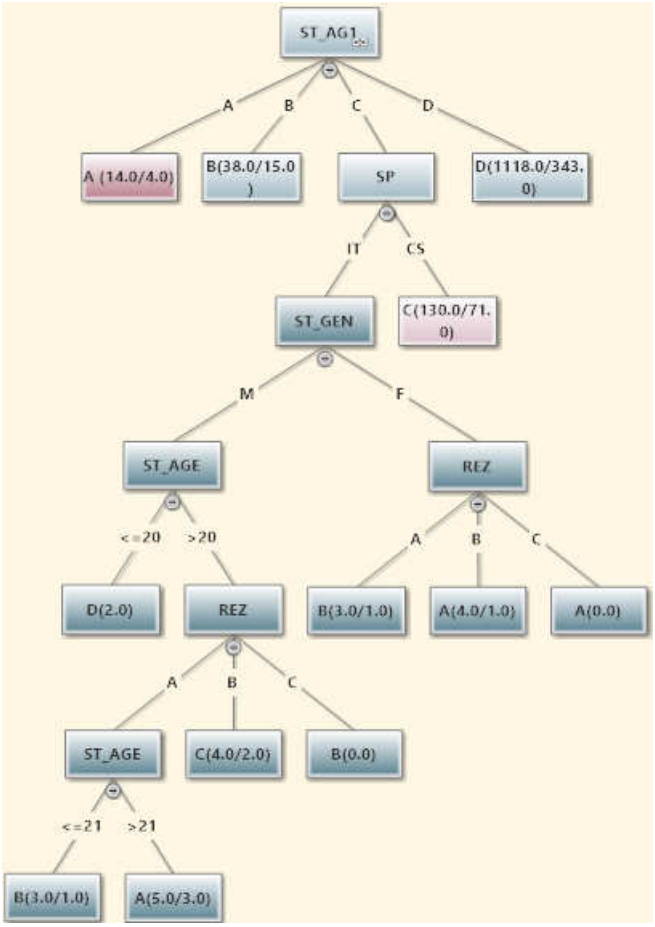

Figure 1 Decision Tree of ST_AG2 class

Table IV
The evaluation of the performance algorithms

Evaluation Criteria	J48	Naïve Bayes	AdaBoostM1
Time taken to build the model (Sec)	0.02	0.19	0.42
Correctly Classified Instances	1203	1154	1120
Incorrectly Classified Instances	118	167	201
Accuracy of prediction	91.07	87.35	84.78%

This methodology may give students/ teachers interesting information about students learning and provides a guideline for students to choose a suitable way, by analyzing the experiences of students with similar academic achievement. We can conclude that J48 algorithm offers the best accurate results and also can express the result in more comprehensible way for the users.

VI. CONCLUSION AND FUTURE WORK

In this study, three important Data mining algorithms were inspected and executed, to evaluate the pre process of the data and the performance of the learning methods, where the evaluation was based in the accuracy. As there are many approaches that are used for data classification, the decision tree method is used in this study. We have collected data from a student management system, using information like student residence, age, gender, average grade in first, second and third year to predict the performance at the end of each academic year.

This study may give to students and teachers interesting information about students learning and provides a guideline for students to choose a suitable way, by analyzing the experiences of students with similar academic achievement.

In order to achieve the goals of this study, we intended to build a system which can allows students or other interested persons to predict the average grade of a specific academic year.

REFERENCES

1. H. Manila, “Data mining: machine learning, statistics and databases” IEEE, 1996.
2. "EducationalDataMining.org", http://www.educationaldatamining.org/ , 2010
3. C. Romero, S. Ventura, Educational Data Mining: a Survey from 1995 to 2005, Expert Systems with Applications, Elsevier, pp. 135-146. 2007
4. H. Witten et al. “Data Mining: Practical machine learning tools and techniques”, 2nd ed. Morgan Kaufmann, San Francisco, 2005.
5. D. Kalles, C. Pierrakeas, Analyzing student performance in distance learning with genetic algorithms and decision trees, Hellenic Open University, Patras, Greece, 2004.
6. S. Hongjie, “Research on Student Learning Result System based on Data Mining”, IJCSNS International Journal of Computer Science and Network Security, Vol.10, No. 4, April 2010.
7. R. S. J. D. Baker, K. Yacef. The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1. 2009
8. A. Nandeshwar, S. Chaudhari. Enrollment prediction models using data mining. 2009 http://nandeshwar.info/wpcontent/uploads/2008/11/DMWVU_Project.pdf
9. P. Cortez, A. Silva. Using data mining to predict secondary school student performance. In the Proceedings of 5th Annual Future Business Technology Conference, Porto, Portugal. 2008
10. I. Witten, E. Frank WEKA Machine Learning Algorithms in Java, Morgan Kaufmann Publishers, 2000

MSc. Alba Çomo is a lecturer in the Computer Science Department, Faculty of Natural Science, University of Tirana. She obtained the MSc. Degree in Computer Science (2011). She is currently doing research in Educational Data mining, for her PhD degree. She has authored 6 research papers in international/national Conference/journals.

Prof. Dr. Ilia Ninka is Professor in the Computer Science Department, Faculty of Natural Science, University Of Tirana. He obtained the grade "Doctor in Sciences"(1984). He received the Title “Professor” (1994). He is the Computer Science Department (since 2008). His research field is Data mining and Knowledge Discover and Artificial Intelligence. He has authored a commendable number of research papers in international/national Conference/journals and also guides research scholars in Computer Science/Applications

MSc. Brisilda Munguli is a lecturer in the Computer Science Department, Faculty of Natural Science, University of Tirana. She obtained the MSc. Degree in Computer Science (2013). Her research field is Data mining and Knowledge Discover and Artificial Intelligence.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.