Mining Student Data Using Decision Trees

Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar

Department of Computer Information Systems
Faculty of Information Technology and Computer Science
Yarmouk University, Irbid 21163, Jordan

Abstract

Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. This paper is an attempt to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. For this purpose, the CRISP framework for data mining is used for mining student related academic data. The classification rule generation process is based on the decision tree as a classification method where the generated rules are studied and evaluated. A system that facilitates the use of the generated rules is built which allows students to predict the final grade in a course under study.

Key Words*: Data Mining, Classification, Decision Trees, Student Data, Higher Education*

1. Introduction

Data mining techniques have been applied in many application domains such as banking, fraud detection, and telecommunications [1]. Recently the data mining methodologies were used to enhance and evaluate the higher education tasks. Some researchers have proposed some methods and architectures for using data mining for higher education [2-5].

In this direction, some models have been proposed and implemented. The authors of [2] have proposed a model to represent how data mining can be used in a higher educational system to improve the efficiency and effectiveness of the traditional processes. In the model, several processes are proposed to be enhanced through data mining functions. The model is also presented as a guideline for higher educational system to improve the decision-making processes.

The research by [3] has used Rough Set theory as a classification approach to analyze student data where the Rosetta toolkit was used to evaluate the student data to describe different dependencies between the attributes and the student status. The discovered patterns are explained in plain English. The data set used in their experiments is the student data of Suranaree University of technology (SUT) during the academic year 2001-2002.

The research by [4] describes the results of analyzing data from a large collection of the so called concurrent version system (CVS) created by many students working on a small set of identical projects (course assignments) in the second year undergraduate computer science course. The proposed model is used to extract all information of student behavior in writing the code of assignments and to find some statistical patterns or predicators that can be used to enhance students' performance in writing the code. The results obtained have suggested that aspects such as student work habits, and even code quality, have little bearing on the student's performance.

The model of Delavari *et al.* in [5] is a motivation toward enhancing the proposed analysis model presented in [2] and that is used as a roadmap for the application of data mining in higher educational system. The enhanced model is named Data Mining for Higher Education (DM_EDU). To prove the model correctness, one of the sub processes proposed by [2] has been implemented and evaluated. The sub-process is student assessment in the computer programming II course. The model allows the decision makers to better predict

which students are less likely to perform well in that specific course, or those who are less likely to be successful in it.

The research by Kalles and Pierrakeas in [6] discussed different machine learning techniques (decision trees, neural networks, Naive Bayes, instance-based learning, logistic regression and support vector machines) and compared them with genetic algorithm based induction of decision trees. They have discussed why the approach has a potential for developing into an alert tool. They have embarked in an effort to analyze students' academic performance through the academic years, as measured by the students home work assignments, and attempted to derive short rules that explain and predict success or failure in the final exams. Students' data are collected from the available data of the academic year 2000-2001. The latest version of WEKA machine learning toolkit [7] was used to evaluate and to experiment the proposed model.

The main objective of this paper is an attempt to use data mining methodologies to study students' performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. The data used in this research is restricted to those students who took the C++ course in Yarmouk University in the year 2005.

2. The Proposed Model

To build a reliable classification model, the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) [8] is adopted. The methodology consists mainly of five steps: Collecting the relevant features of the problem under study, preparing the data, building the classification model, evaluating the model using one of the evaluation methods, and finally using the model for future prediction of the student performance. These steps are presented in the next subsections.

2.1 Collecting the Relevant Features

In this step the relevant features are collected using a questionnaire that was passed among undergraduate students of the Information Technology & Computer Science Faculty, Yarmouk University / Jordan who took the Programming I course (C++). Initially more than 20 attributes have been collected and some of the attributes have been manually eliminated since they are considered as irrelevant to the study. Finally only 12 conditional attributes and one class attribute have been considered. The attributes along with their descriptions and possible values are presented in Table 1. The class attribute is the student grade in the C++ course and named (grade101).

2.2 Preparing the Data and Selecting the Relevant Attributes

For this step, the collected data were prepared in tables in a format that it is suitable for the used data mining system. The data are cleansed by removing the various inconsistent values using the same standard value for all the data. The cleaning also includes filling out the missing values using the most majority data approach. Since the collected attributes may have some irrelevant attributes that may degrade the performance of the classification model, a feature selection approach is used to select the most appropriate set of features. For this purpose the WEKA toolkit is used and the attributes are ranked and then 3 attributes are eliminated by the feature selection approach. Finally, the most significant attributes list contains the following attributes presented in descending order according to their ranks: **HSGrade, Fund, TDept, TDegree, HKind, Study-Type, T-Gender, St-Depart, St-Gender.**

Table 1: The Symbolic Attribute Description:

Attribute	Description	Possible Values
St-Gender	Student Gender	M, F
St-Age	Student Age	18, 19, 20, 21, 22
St-Depart	Student Department	CIS ,CS , MIS
HSMajor	High School Major	SCIENCE, ART
HSGrade *	High School Grade	A, B, C, D
Study-Type	Study Type	NORMAL, PARALLEL, INTERNATIONAL
Fund	Funding	PRIVATE, SCHOLARSHIP, LOCAL
HKind	Place of Residency	FAMILY, FRIENDS, ALONE
T-Degree	Lecturer Degree	PhD, MS, BS
T-Gender	Lecturer Gender	M, F
T-Dept	Lecturer department	CIS, CS, MIS
Repeat	Number of repetitions	0, 1, 2, 3
Grade101 *	The Grade of C++ course (the Class)	A, B, C, D

Notes : *HSGrade and Grade101 : A= 90-100, B= 80-89, C= 70-79, D= 50-69.

2.3 Building the Classification Model

The next step is to build the classification model using the decision tree method. The decision tree is a very good and practical method since it is relatively fast, and can be easily converted to simple classification rules. The decision tree method depends mainly on using the information gain metric which determines the attribute that is most useful. The information gain depends on the entropy measure.

The gain ratio is used to rank attributes and to build the decision tree where each attribute is located in according to its gain ratio. For the course under study in this paper, the attribute that has the highest gain ration was the HSGrade (The high school grade). 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. After building the complete decision tree, the set of classification rules are generated by following all the paths of the tree where the decision tree has generated 41 classification rules. Some of the generated rules are given in Table 2 in a form that is understandable by humans.

In Table 2, the first column represents the rule number, the generated rules are presented in the second column, the number of the students who successfully satisfy the rules is given in the third column, and the number of attributes contained in the rule is given in the last column. The table shows the rules in a descending order depending on the number of the students who successfully have satisfied the rule. This ordering helps in determining the most significant rule. For the generated rules, the longest rule consists of 9 attributes while the shorter rule contained only 2 attributes. Some of the discovered interesting rules are:

- IF Student gender is Male and his grade in High School was A, then the predicted grade in the C++ course is C.
- The lecturer is a Female, and the student funding is a university employee (LOCAL) and the High school grade is C, then the predicted grade is A.
- If the Students department is Computer Information Systems and the study type is Parallel and the lecturer department is Computer Information Systems or Computer Science, and the student high school grade is B or D, then the predicted grade is D.

2.4 Using the model for future prediction of the student performance

In order to achieve the goals set by this research, a system that facilitates the usage of the generated rules is built which allows students to predict the final grade in the C++ course.

Table 2: Sample of the Generated Rules.

Rule #	Rules	# Obj	# Attrib
18	IF St-Depart = CS, T-Degree = MS, St-Gender = F, T-Gender = M, Fund = SCHOLARSHIP OR PRIVATE, HKind = ALONE or FAMILY, Study-Type = NORMAL, T-Dept = CIS or CS, HSGrade = B or D THEN *Grade101* = C	13	9
25	IF St-Depart = CS , St-Gender = F , T-Degree = MS , T-Gender = F , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade =B or D THEN *Grade101* =C	9	9
21	IF St-Depart = CIS , T-Degree =BS or MS , St-Gender = M , T-Gender = M , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade = B or D THEN *Grade101* =D	9	9
17	IF T-Degree =BS or PhD , St-Gender = F , T-Gender = M , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS, HSGrade =BorD THEN *Grade101* =C	8	8
26	IF St-Depart = CIS , St-Gender = F , T-Degree = MS , T-Gender = F , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade =B or D THEN *Grade101* =C	6	9
24	IF St-Gender = M , T-Degree = MS , T-Gender = F , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade =B or D THEN *Grade101* =D	4	8
20	IF St-Depart = CS , T-Degree =BS or MS , St-Gender = M , T-Gender = M , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade =B or D THEN *Grade101* =B	4	9
41	IF T-Dept = CS , St-Depart = CIS , T-Gender = F , T-Degree = MS , Fund =SCHOLARSHIP or PRIVATE , HSGrade = C THEN *Grade101* =C	3	6
38	IF T-Gender = m , T-Degree = MS , Fund = SCHOLARSHIP or PRIVATE , HSGrade = C THEN *Grade101* =C	3	4
37	IF T-Dept = CS or MIS , St-Depart = CS , T-Degree =PhD , Fund =SCHOLARSHIP or PRIVATE , HSGrade = C THEN *Grade101* =C	3	5
32	IF T-Gender = f , Fund = LOCAL , HSGrade = C THEN *Grade101* =A	3	3
30	IF T-Dept = MIS , HSGrade = B or D THEN *Grade101* =D	3	2
23	IF T-Degree =PhD , T-Gender = F , Fund =SCHOLARSHIP or PRIVATE , HKind =ALONE or FAMILY , StudyType =NORMAL , T-Dept = CIS or CS , HSGrade =B or D THEN *Grade101* =D	3	7

3. Experiments and Evaluation

As described in [9], in order to measure the performance of a classification model on the test set, the classification accuracy or error rate are usually used for this purpose. The classification accuracy is computed from the test set where it can also be used to compare the relative performance of different classifiers on the same domain. However, in order to do so, the class labels of the test records must be known. Moreover an evaluation methodology is needed to evaluate the classification model and compute the classification accuracy. Mainly there are two methods for the evaluation named: The Holdout method and the *K*-Cross-Validation method (*k*-CV) [10].

To obtain the accuracy of the classification model the WEKA toolkit is used. Three different classification methods have been tested, the ID3, C4.5, and the Naïve Bayes. Table 3 shows the evaluation result as a percentage of the correctly classified instances using the aforementioned three different algorithms.

Table 3: Classification Accuracy of the 3 different algorithms.

Algorithm	Hold out	10-CV
ID3	38.4615 %	28.3186 %
C4.5	35.8974 %	38.0531 %
Naive Bayes	33.3333 %	38.0531 %

From the obtained results, we can notice that the classification accuracy for the three different classification algorithms is not so high. This can indicate that the collected samples and attributes are not sufficient to generate a classification model of high quality.

4. Conclusion

This research is a starting attempt to use data mining functions to analyze and evaluate student academic data and to enhance the quality of the higher educational system. The higher managements can use such classification model to enhance the courses outcome according to the extracted knowledge. Such knowledge can be used to give a deeper understanding of student's enrollment pattern in the course under study, and the faculty and managerial decision maker in order to utilize the necessary actions needed to provide extra basic course skill classes and academic counseling. On the other hand, using such knowledge the management system can improve their policies, enhance their strategies, and improve the quality of management system.

One of the most attractive future works is to collect a real and large data set from the university student database and apply the model using such data. Moreover, several other classification methods can also be applied to test the most suitable method that suit the structure of the student data and give a better classification accuracy.

References

[1] Han J, Kamber M. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2001.

[2] Delavari N, Beikzadeh M. R. *A New Model for Using Data Mining in Higher Educational System*, 5th International Conference on Information Technology based Higher Education and Training: ITEHT ’04, Istanbul, Turkey, 31st May-2nd Jun 2004.

[3] Varapron P. *et al*. *Using Rough Set theory for Automatic Data Analysis. 29th Congress on Science and Technology of Thailand*. 2003.

[4] Mierle K, Laven K, Roweis S, Wilson G, *Mining Student CVS Repositories for Performance Indicators*, 2005.

[5] Delavari N, Beikzadeh M. R, Amnuaisuk S. *Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System.* 6th Annual International Conference: ITEHT July 7-9, 2005,Juan Dolio,Dominican Republic.

[6] Kalles D., Pierrakeas C., *Analyzing student performance in distance learning with genetic algorithms and decision trees*, Hellenic Open University, Patras, Greece,2004.

[7] Witten I. Frank E. *WEKA Machine Learning Algorithms in Java*, Morgan Kaufmann Publishers, 2000.

[8] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. *CRISP-DM 1.0: Step-by-step data mining guide*, 2000.

[9] Tan P., Steinbach M., Kumar V. *Introduction to DATA MINING*. Pearson Education, 2006.

[10] Al-Radaideh Q., Sulaiman M., Selamat M, Ibrahim H. *Evaluation of Rough Sets Based Classification*. Symposium of Intelligence Systems and Information Technology (ISITS04), ITMA,UPM, Malaysia. Feb 2004 .