ISSN:2278-7690

Using Data Mining Technique to Analyze Student's Performance

Abeer Badr El Din Ahmed*, Ibrahim Sayed Elaraby
Lecturer at Sadat Academy, Computer Science Department, Cairo, Egypt
Demonstrator at Higher Institute for GPecific Studies, Management Information System Department, Cairo, Egypt

ABSTRACT

Educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. With the help of Data Mining, which is an emerging technique, one can efficiently learn from historical data and use that obtained knowledge for predicting future behaviour of concern areas. Growth of current education system is surely enhanced if Data Mining has been adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of student performance, course development and finally the development of nation's education related standards. This paper focuses on capabilities of data mining in context of education, Also it compare three of the most common data mining techniques (ID3, C4.5 and REPTree).

Indexing terms/Keywords

Higher Education, Knowledge Discovery, Data Mining, ID3, C4.5 and REPTree.

Academic Discipline And Sub-Disciplines

Computer Science; Education;

SUBJECT CLASSIFICATION

Data Mining

TYPE (METHOD/APPROACH)

Practical Study

Council for Innovative Research
Peer Review Research Publishing System
International Journal of Research in Education Methodolgy

Vol.5, No.2
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

During past few years, there is explosive growth in the educational data which contains valuable information [2]. Usually, organizations generate data about student, staff, faculty, which contains information of management system, employees, lecturers, organizational personal and so on. These strategic resources are helpful for improving quality of higher educational institutes. Traditional statistical techniques are not much adequate for analysis of this datasets. The knowledge discovered by data mining techniques would enable the higher education systems in making better decisions, having more advanced planning [1].

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the student's performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve student's academic performance and trim down failure rate, to better understand student's behaviour, to assist instructors, to improve teaching and many other benefits [3]. Educational data mining uses many techniques such as decision tree, rule induction, neural networks, k-nearest neighbour, naïve Bayesian and many others. By using these techniques, many kinds of knowledge can be discovered such as association rules, classifications and clustering [4] [5].

RELATED WORK

Cristóbal, Sebastián and García [5]. Used educational data mining in Moodle course management system. They used each step in data mining process for mining e-learning data. Also, educational data mining used by [11] to predict student,s final grade using data collected from Web based system.

Mohammad and Naeimeh [9]. Used educational data mining to identify and then enhance educational process in higher educational system which can improve their decision making process.

Chandra and Nandhini [7], applied the association rule mining analysis based on student's failed courses to identifies student's failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity student's performances. The extracted association rules reveal some hidden patterns of students' failed courses which could serve as a foundation stone for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate.

Agathe and Kalina [10]. used educational data mining to identify behaviour of failing student's to warn students at risk before final exam.

P. K. Srimani and Malini M. Patil [14], discusses on the application of data mining algorithms and techniques on academic data to potentially increase some of the aGPects of education system by developing a method called Edu-mining which is a novel approach. Academic data is also referred as educational data i.e. using REPtree algorithm.

El-Halees [6], used educational data mining to analyze student's learning behaviour. The goal of his study is to show how useful data mining can be used in higher education to improve student's performance. He used students' data from database course and collected all available data including personal records and academic records of students, course records and data came from e-learning system. Then, he applied data mining techniques to discover many kinds of knowledge such as association rules and classification rules using decision tree. Also he clustered the student into groups using EM clustering, and detected all outliers in the data using outlier analysis. Finally, he presented how can we benefited from the discovered knowledge to improve the performance of students.

RESEARCH OBJECTIVES

In this paper we used educational data mining to improve graduate student's performance, and overcome the problem of low grades of graduate student's. In our case study we try to extract useful knowledge from graduate student's data using ID3, C4.5 and REPTree.

DATA MINING

Data mining is the process of extracting useful knowledge and information including, patterns, associations, changes, anomalies and significant structures from a great deal of data stored in databases, data warehouses, or other information repositories [13, 14].

DATA MINING IN HIGHER EDUCATIONAL SYSTEM

Today the important challenge that higher education faces, is reaching a stage to facilitate the universities in having more efficient, effective and accurate educational processes. Data mining is considered as the most suited technology appropriate in giving additional insight into the lecturer, student, alumni, manager, and other educational staff behaviour and acting as an active automated assistant in helping them for making better decisions on their educational activities [9].

ISSN:2278-7690

THE GRADUATE STUDENT'S DATA SET AND PREPROCESSING

The data set used in this paper contains graduate student's information collected from one of the educational institutions for a period of five years in period from 2005 to 2010. The graduate student's data set consists of 1038 record and 18 attribute.Table 2 presents the attributes and their descriptions that exist in the data set as taken from the source database.

Table 1. Students Dataset Description

Attribute	Description	Possible Values	Selected
Student_ID	The Student _id		
Student_Name	The Name of the Students		
HS	High School of Students	{Literary, Scientific Mathematics, Scientific Science}	√
HSD	High School Degree	{Good-Acceptable}	√
G	Gender of the Students	{Male – Female}	
Date_of_Birth	Date of Birth of the Students		
Place_of_Birth	Place of Birth of the Students		
Address	Address of the Students		
Telephone	Telephone of the Students		
Course	Identification of the Course	MIS - c# - Network - VB.net	
Midterm	Midterm Marks	{Excellent >=85% , Very Good >=75 & <85% Good >=65 & <75%, Acceptable >=50 & <65%, Fail < 50%}	√
CTG	Class Test Grade	{Poor , Average, Good}	√
SEM	Seminar Performance	{Poor , Average, Good}	√
ASS	Assignment	{Yes, No}	√
GP	General Proficiency	{Yes, No}	√
ATT	Attendance	{Poor , Average, Good}	√
HW	Lab Work	{Yes, No}	√
FG	Final Grade of Student	{Excellent >=85%, Very Good >=75 & <85% Good >=65 & <75%, Acceptable >=50 & <65%, Fail < 50%}	√

As part of the data preparation and pre-processing of the data set and to get better input data for data mining techniques, we did some pre-processing for the collected data before loading the data set to the data mining software, irrelevant attributes should be removed. The attributes marked as selected as seen in Table 1 are processed via the rapid miner software to apply the data mining methods on them. The attributes such as the Student_Name or Student_ID, etc. are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the student's, also they have very large variances or duplicates information which make them irrelevant for data mining.

After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the distribution of values, GPecifically the grade of students. Figure 1 depicts the distribution of the graduate students in period from 2005 to 2010 according to their grades, it is apparent from the figure 1.

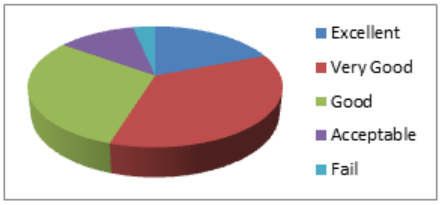

Fig 1: The distribution of graduate student's according to their grades

ISSN:2278-7690

RESULTS AND DISCUSSION

The data set used in this study was obtained from a student's database used in one of the educational institutions, on the sampling method of Information system department from session 2005 to 2010. Initially size of the data is 1038 records are given in Table 2 where Excellent (EX), Very Good (VG), Good (G), Acceptable (ACC) and Fail (FL), Average (AV), Poor (P), Yes (Y), No (N), Female (F), Male (M).

Table 2. Students Data Set

Midterm	CTG	SEM	ASS	GP	ATT	HW	FG	HSD	HS
EX	G	G	Y	Y	G	Y	EX	G	Literary
VG	G	G	Y	N	G	N	VG	G	Scientific Mathematics
VG	G	G	Y	N	G	N	VG	G	Scientific Mathematics
VG	G	G	Y	N	G	N	VG	ACC	Scientific Mathematics
FL	P	AV	Y	Y	G	Y	VG	ACC	Scientific Mathematics
FL	P	AV	Y	Y	G	Y	VG	G	Scientific Mathematics
ACC	P	AV	Y	N	G	N	G	ACC	Scientific Mathematics
.	.	.	.	.	.	.	.	.	.

Before applying the data mining techniques on the data set, there should be a methodology that governs our work. Figure 2 depicts the work methodology used in this paper. The methodology starts from the student field, then pre-processing and the data set and pre-processing sections, then we come to the data mining methods which are classification, followed by the evaluation of results and patterns, finally the knowledge representation process.

In this section, we describe the results of applying the data mining techniques to the data of our case study using data mining task: classification and how we can benefited from the discovered knowledge.

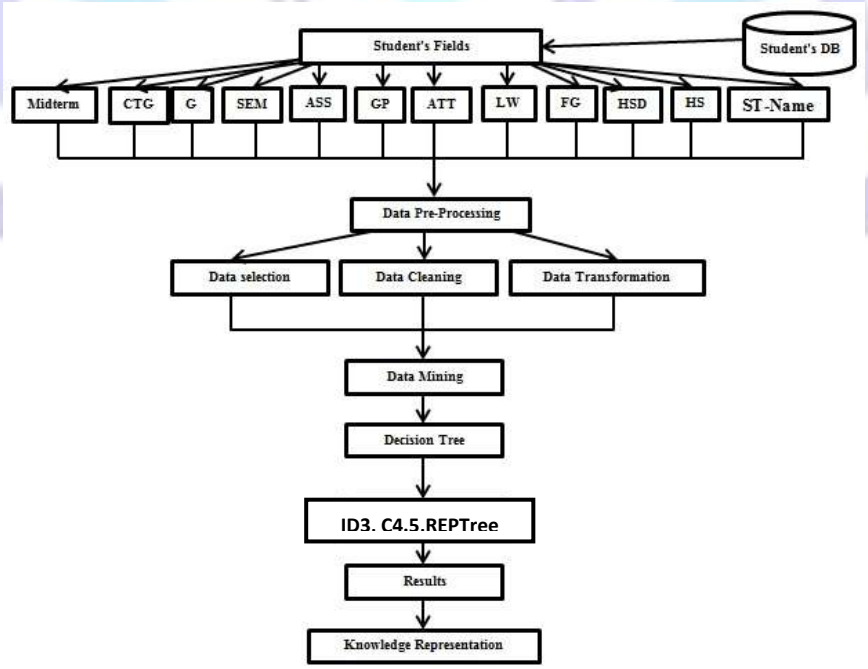

Fig 2: Data Mining Work Methodology

CLASSIFICATION

In this paper, the classification approaches are used to predict the Grade of the graduate student and there are five grades (Excellent, Very good, Good, Acceptable and Fail) and how other attributes affect them.

Two classification methods are used which are decision tree (ID3, C4.5 and REPTree). A decision tree classifier extracts a set of rules that show relationships between attributes of the data set and the class label. It uses a set of IF-THEN rules for classification. Rules are easier for humans to understand, given a class, from training data and to then allow the use of these probabilities to classify new entities.

Table 3 depicts the rules that resulted from applying the decision tree (ID3) classification algorithm on the Grade of the graduate student as a target class

Table 3. The Resulting Classification Rules

IF MEDTERM = "Excellent" AND CTG = "Good" THEN ESM ="Excellent"
IF MEDTERM = "Excellent" AND CTG = "Very Good" THEN ESM ="Excellent"
IF MEDTERM = "Excellent" AND CTG = "Poor" AND HW = "Yes" THEN ESM ="Very Good"
IF MEDTERM = "Excellent" AND CTG = "Poor" AND HW = "No" THEN ESM ="Good"
IF MEDTERM = "Excellent" AND CTG = "Poor"AND SEM = "Average" THEN ESM ="Very Good"

Table 4 depicts the rules that resulted from applying the decision tree (C4.5) classification algorithm on the Grade of the graduate student as a target class.

Table 4. The Resulting Classification Rules

IF ATT = "Good" AND CTG = "Good" AND HW = "No" AND SEM = "Good" THEN ESM ="Very Good"
IF ATT = "Good" AND CTG = "Good" AND HW = "Yes" AND SEM="Good" THEN ESM ="Excellent"
IF ATT = "Good" AND CTG = "Good" AND HW = "Yes" AND SEM="Poor" AND ASS = "No" THEN ESM ="Very Good"
IF ATT = "Good" AND CTG = "Good" AND HW = "Yes" AND SEM="Average" AND ASS = "Yes" THEN ESM ="Excellent"
IF ATT = "Good" AND CTG = "Good" AND HW = "Yes" AND SEM="AVERAGE" AND ASS = "No" THEN ESM ="Very Good"

Table 5 depicts the rules that resulted from applying the decision tree (REPTree) classification algorithm on the Grade of the graduate student as a target class.

Table 5. The Resulting Classification Rules

IF MEDTERM = "Excellent" AND CTG = "Good" AND GP = "No" AND HS=" Scientific Mathematics" THEN ESM ="Very Good"
IF MEDTERM = "Excellent" AND CTG = "Good" AND GP = "No" AND HS=" Scientific Science" AND HW="No" THEN ESM ="Very Good"
IF MEDTERM = "Excellent" AND CTG = "Good" AND GP = "No" AND HS=" Scientific Science" AND HW="Yes" THEN ESM ="Excellent"
IF MEDTERM = "Excellent" AND CTG = "Good" AND GP = "No" AND HS=" Literary" THEN ESM ="Very Good"
IF MEDTERM = "Excellent" AND CTG = "Good" AND GP = "Yes" THEN ESM ="Excellent"

EXPERIMENTAL RESULT

1. Accuracy (Correctly)

Table 6. Accuracy Comparison Between Id3, C4.5and Reptree Algorithm In Percentage

Algorithm			
Size of Data Set	ID3	C4.5	REPTREE
500	75.2941 %	82.3529 %	77.5346 %
750	73.5294 %	78.5294 %	73.8798 %
1038	75.4753 %	80.038 %	76.5764 %

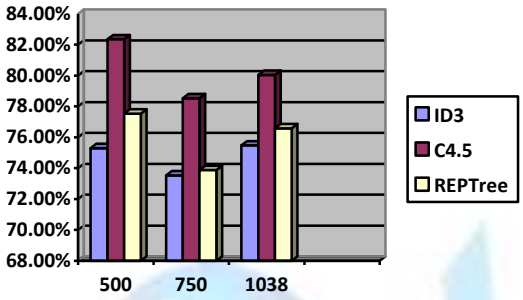

Fig 3: Accuracy Comparison between ID3, C4.5 AND REPTREE Algorithm in Percentage

2. Incorrectly

Table 7. Incorrectly Comparison Between Id3, C4.5and Reptree Algorithm In Percentage

Algorithm			
Size of Data Set	ID3	C4.5	REPTREE
500	20.7843 %	17.6471 %	20.8983 %
750	23.5294 %	21.4706 %	24.3246 %
1038	21.6732 %	19.9621 %	22.8591 %

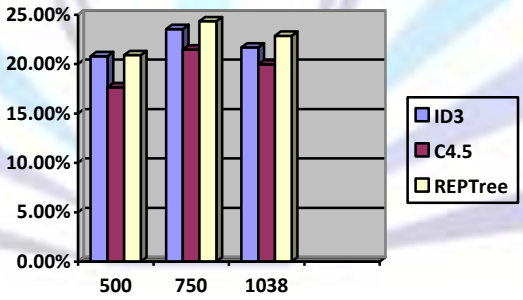

Fig 4: Incorrectly Comparison between ID3, C4.5 AND REPTREE Algorithm in Percentage

3. Model Build Time

Table 8. Models Build Time Comparison between ID3, C4.5and REPTree Algorithm in Sec

Algorithm			
Size of Data Set	ID3	C4.5	REPTREE
500	0.02 sec	0.01 sec	0.03 sec
750	0.02 sec	0.01 sec	0.03 sec
1038	0.03 sec	0.02 sec	0.04 sec

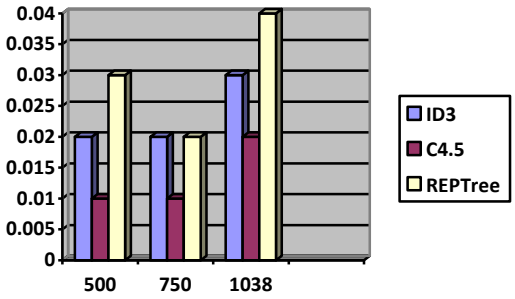

Fig 5: Models Build Time Comparison between ID3, C4.5and REPTree Algorithm in Sec

CONCLUSIONS

In this paper, we gave a case study in the educational data mining. It showed how useful data mining can be used in higher education particularly to improve graduate student's performance. We used graduate student's data collected from the one of the educational institutions. The data include five year's period [2005-2010]. We used ID3, C4.5 and REPTree techniques and compare the result, Though C4.5 gives better accuracy than ID3 and REPTree, and C4.5 builds a tree and prunes it to get a more efficient tree. the performance of J48 (implementation of C4.5 in Weka) is much better than ID3 and REPTree, tree built by C4.5 seems to give high prediction accuracies, C4.5 gave faster and better results, tree generated by C4.5 algorithm was much higher than the rest.

REFERENCES

[1] Delavari.N, Beikzadeh.M.R and Phon-Amnuaisuk.S. "Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System".2005.

[2] Bala. M and Ojha. D. B. "Study of Applications of Data Mining Techniques in Education ".2012.

[3] Kumar. V and Chadha. A. "An Empirical Study of the Applications of Data Mining Techniques in Higher Education".2011.

[4] C. Romero and S. Ventura. (2007) " Educational Data Mining: A Survey from 1995 to 2005 ".2007.

[5] Romero. C, Ventura. S and García. E. " Data Mining in Course Management Systems: Moodle Case Study and Tutorial".

[6] El-halees. A. " Mining Students Data to Analyze Learning Behaviour: A Case Study ".2005.

[7] Sundar.P.V." A Comparative Study for Predicting Student's Academic Performance Using Bayesian Network Classifiers ".2013.

[8] Ramaswami.M and Bhaskaran.R. " A Chaid Based Performance Prediction Model in Educational Data Mining ".2010.

[9] Beikzadeh.M.R and Delavari.N. " A New Analysis Model for Data Mining Processes in Higher Educational Systems".2004.

[10] Merceron.A and Yacef.K. "Educational Data Mining: A Case Study ".2005.

[11] Minaei-Bidgoli. B, Kashy. D.A , Kortmeyer. G and Punch, W.F." Predicting Student Performance: An Application of Data Mining Methods with An Educational Web-Based System ".2003.

[12] Han.J and Kamber.M." Data Mining: Concepts and Techniques ".2006.

[13] Chen.M, Hun.J and Philip.S." Data Mining: An Overview from Database PerGPective ".1996.

[14] P. K. Srimani and Malini M. Patil. "A Classification Model for Edu-Mining".2012.