OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Data Mining Techniques in Higher Education an Empirical Study for the University of Palestine

Rasha Ragheb Atallah[1], Professor Samy S. Abu Naser[2]

[1](*Faculty of Information Technology, The University of Palestine, Gaza, Palestine*)

[2] *(Department of Information Technology, Faculty of Engineering & Information Technology, Al Azhar)*

Abstract: *Nowadays, ones of the biggest challenges that educational institutions face is the explosive growth of educational data. and how to use these data to improve the quality of managerial decisions. Data mining, as an analytical tools that can be used to extract meaningful knowledge from large data sets, can be used to achieve this goal.*

This paper addresses the applications of Educational Data Mining (EDM) to extract useful information from registration information of student at university of Palestine in Gaza strip. The data include five years period [2005-2011] by providing analytical tool to view and use this information for decision making processes by taking real life example such as grade and GPA for the students. abstract should summarize the content of the paper.

Keywords: *Association, Classification, Clustering, Data mining, Higher education, Knowledge discovery, rules, Outlier analysis*

I. Introduction

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational setting ,and using those methods to better understand students, and the settings which they learn in [1]. Educational data mining can mine student GPA, student address, enrollment data, and applying it any college. Also with data mining the university can predicate which student will graduated and whose will not., So the university can use this result for improving the education behavior and improving student performance.

This paper applicant at University of Palestine which is a Palestinian private institution of higher education located in Al-Zahra' (south of Gaza City), the university was established in 2005.

In the paper, we try to extract useful knowledge from the registration system at University of Palestine during 5 years from [2005-2011]. After data preprocessing, we applied the association rules, Classification rules, and outlier analysis mining techniques.

II. Related Works

Hsu and Schombert in [3] analyze a data set comprised of academic records of undergraduates at the University of Oregon from 2000-2004. They found correlations of roughly 0.35 to 0.5 between SAT(predictive power of tests for college admissions) scores and upper division, in-major GPA (henceforth, GPA).

Interestingly, low SAT scores do not preclude high performance in most majors. The paper hypothesizes that over achievers overcome cognitive deficits through hard work, and discusses to what extent they can be identified from high school records. Only a few majors seem to exhibit a cognitive threshold – such that high GPA (mastery of the subject matter) is very unlikely below a certain SAT threshold (i.e., no matter how dedicated or hard working the student). There results suggest that almost any student admitted to university can achieve academic success, if they work hard enough. Also the paper found that the best predictor of GPA is a roughly equally weighted sum of SAT and high school GPA, measured in standard deviation units. Finally, the paper observe that one SAT scores fluctuate little on retest (very high reliability), two SAT and GRE scores (where available) correlate at roughly 0.75 (consistent with the notion that both tests measure a stable general cognitive ability) and three SAT distribution of students that obtained a degree does not differ substantially from that of the entering class.

Romero and Ventura in [4] introduced a survey of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system. It is objective is to introduce it both theoretically and practically to all users interested in this new research area, and in particular to online instructors and e-learning administrators, they describe the full process for mining e-learning data step by step as well as how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering and association rule mining of Moodle data. They have described how different data mining techniques can be used in order to improve the course and the students' learning. All these techniques can be applied separately in a same system or together in a hybrid system.

Vialardi et. al in [5] focused on how university students can take the right decision in relation to their academic itinerary based on available information such as courses, schedules, section. This paper proposes use the recommendation system based on data mining techniques to help students to take decision on their academic itineraries, as example how many and which courses to enroll on ,having as basis the experience of previous students with similar academic achievements, they analyzed real data corresponding to seven years of student enrolment at the school of system Engineering at Universidad de Lima ,According to this analysis they developed a system.

Hongjie in [6] described 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. From the perspective of data analysis, this paper conducts a research on student learning result based on data mining.

III. CONCEPTS IN DATA MINING

The objective of this research is to identify the data mining techniques which can be applied in the field of Higher education.

3.1. Data Mining Definition and Techniques

Data mining refers to extracting or "mining" knowledge from large amounts of data [7]. Data mining techniques are used on large volumes of data to discover hidden patterns and relationships to help in decision making. The sequences of steps identified in extracting knowledge from data as shown in figure 1.

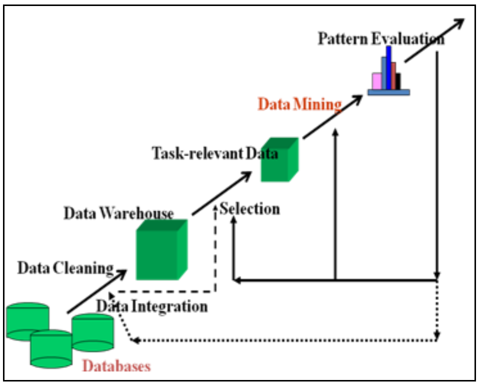


Fig 1: The steps of extracting knowledge from data

3.2. Data Collection

Initially the university provided us with 65536 records corresponding to 2493 students. The data supplied was students data at 5 faculties: Engineering, Information Technology, Management, Low and Media enrolled through the years of 2005 to 2011.

3.3. Data Preprocessing

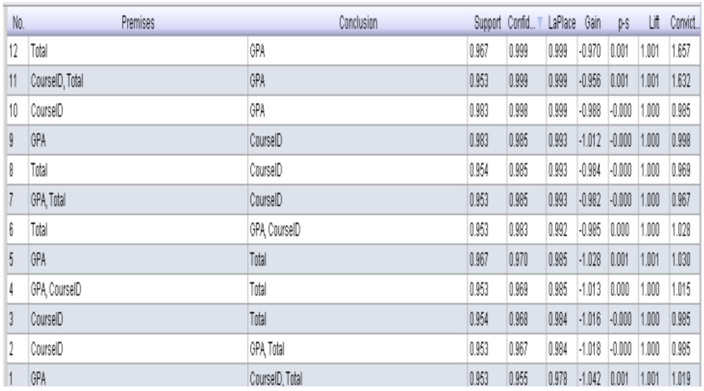
In the entire data mining process it is of great relevance the data cleaning process in order to eliminate irrelevant item, as section number. ID of student , mid grade and final grade because in this paper the grade of the course is the needed also to fill the missing values in the data.

IV. Case Study Application

This section represents the various techniques of data mining which applicant in the data of the registration at University of Palestine.

4.1. Association

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association rules used to discover the relationship between the grade of the course and the GPA of the student using FP-Growth operator. Figure 2 shows some resulting rules by association rules model with their evaluation factors support, confidence, and lift.



No.	Premises	Conclusion	Support	Confid...	LaPlace	Gain	p-s	Lift	Convict...
12	Total	GPA	0.967	0.999	0.999	-0.970	0.001	1.001	1.657
11	CourseID, Total	GPA	0.953	0.999	0.999	-0.956	0.001	1.001	1.632
10	CourseID	GPA	0.983	0.998	0.999	-0.988	-0.000	1.000	0.985
9	GPA	CourseID	0.983	0.985	0.993	-1.012	-0.000	1.000	0.998
8	Total	CourseID	0.954	0.985	0.993	-0.984	-0.000	1.000	0.969
7	GPA, Total	CourseID	0.953	0.985	0.993	-0.982	-0.000	1.000	0.967
6	Total	GPA, CourseID	0.953	0.983	0.992	-0.985	0.000	1.000	1.028
5	GPA	Total	0.967	0.970	0.985	-1.028	0.001	1.001	1.030
4	GPA, CourseID	Total	0.953	0.969	0.985	-1.013	0.000	1.000	1.015
3	CourseID	Total	0.954	0.968	0.984	-1.016	-0.000	1.000	0.985
2	CourseID	GPA, Total	0.953	0.967	0.984	-1.018	-0.000	1.000	0.985
1	GPA	CourseID, Total	0.953	0.955	0.978	-1.042	0.001	1.001	1.019

Fig. 2 Some Resulting Rules from Association Rules Model

Table 1: Associations rules for grade of the student data

Rule		Confidence
1#	[CourseID, Total] --> [GPA]	0.999
2#	[Total] --> [GPA]	0.999
3#	[GPA, Total] --> [CourseID]	0.985
4#	[Total] --> [CourseID]	0.985
5#	[GPA] --> [CourseID]	0.985
6#	[GPA, CourseID] --> [Total]	0.969

According to the rules in table1 for example from rule 1# if the course ID and total founded so the GPA founded. [not clear], also according to rule 3# from the total for the course and the GPA we can know the course because some course know that the general path for this course are high mark as Arabic but as principle of law it is mark as low.

Rule 6 help us in extract the grade of the course if we know the GPA and the Course that help the Academic advisor in the Faculty which courses registered to the student and how many hours he can registered.

4.2. Classification

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects.

In this work classification used to label is the student pass or fail according to the grade of the student, the grade type was the label in the operators which used to implemented the classification, In this paper two methods for the classification were implemented: the K_NN (k=10) with training data 60% and k =10 the result of K_NN is accuracy: 99.82%. In Decision Tree operator, it achieved 99.97% accuracy as shown in figure 3 below, the Decision tree of pass or fail student.

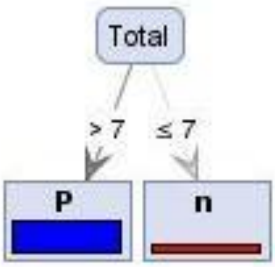


Fig. 3: Resulting Decision Tree

4.3. Outlier Analysis

A Database can contain data objects that do not obey the general behavior of the data and are identified as outliers[9]. The analysis of these outliers may help in fraud detection and predicting abnormal values.

In the EDM, outlier can be used to detect whether a student cheated in the exams according to his grade in the course and his GPA or detect the student phenomenon that he act in some courses better than other, or to guide the student in choosing the proper major in the university.

In this paper outlier founded to show the relation between the grade of the course and the GPA for the student, it was used to detect outlier (distance) operator.

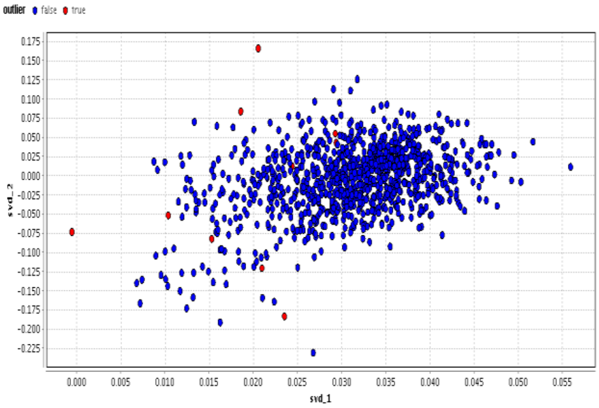


Fig. 4: Outliers Distribution

As we noted in Table 2 that there were three type of outliers :

1- The Grade was 99 and the GPA =60 here we can note that this student may be interested in this course so he can change his major.
2- The GPA was 95.95 and his grade 0, in this case the grade was not entered so must be entered by the teacher that help the Registration department in solving this problem.
3- The GPA = 0 and grade 99, in this situation the student registered zero credit hour course only.

Table 2: presented the analysis for the outliers

Number of outlier	Analysis
2	One of the outlier that student grade was 99 and his GPA 60
6	One of the outlier that student grade zero and this GPA 95.95 because here grade not entered
2	One of the outlier that student grade 99 and this GPA 0

V. CONCLUSION

In this study, a discussion of various data mining techniques which can support education system were presented. Since the application of data mining have a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like guide student in chose his faculty. Furthermore, it enhances the performance of student in terms of his grade, detect the fraud of student in grade also can detect if student grades were changed in the database.

Acknowledgements

We would like to thank all people who provided any type of help in this paper, especially Mr. Sameh Abu Hassera from the University of Palestine.

REFERENCES

[1] "EducationalDataMining.org". 2010 http://www.educationaldatamining.org/. Retrieved 26-4-2014

[2] Han Jiawei, Micheline Kamber, Data Mining: Concepts and Technique. Morgan Kaufmann Publishers, 2000

[3] Data Mining the University: College GPA Predictions from SAT Scores. Stephen D.H. Hsu, James Schombert

[4] C. Romero, S. Ventura, E. Garcia, "Datamining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008.

[5] César Vialardi, Javier Bravo, Leila Shafti, Êlvaro, Recommendation in Higher Education Using Data Mining Techniques, Educational Data Mining, 2009.

[6] Sun 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] http://en.wikipedia.org/wiki/Data_mining 25/4/2014.

[8] http://en.wikipedia.org/wiki/Educational_data_mining, 26-4-2014.

[9] Sakthi Nathiarasan A , Algorithm for Outlier Detection Based onUtility and Clustering (ODUC), International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013.