MINING STUDENTS DATA TO ANALYZE LEARNING BEHAVIOR: A CASE STUDY

ALAA EL-HALEES

Department of Computer Science, Islamic University of Gaza P.O.Box 108 Gaza, Palestine
alhalees@iugaza.edu.ps

ABSTRACT

Educational data mining concerns with developing methods for discovering knowledge from data that come from educational environment. In this paper we used educational data mining to analyze learning behavior. In our case study, we collected students' data from DataBase course. After preprocessing the data, we applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, we extracted knowledge that describes students' behavior.

***Keywords**: Educational Data Mining, E-Learning, Learning Management Systems.*

1. INTRODUCTION

There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data come from educational environments [15]. The data can be colleted form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a vast amount of information used by most institutes [8][13]. Educational data mining used many techniques such as decision trees, neural networks, k-nearest Neighbor, Naive Bayes, support vector machines and many others. Using these methods many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used to better understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems , to improve curriculums and many other benefits [14] [15].

Romero and Ventura in [14] concluded that work should be oriented towards educational domain of data mining. This paper investigates the educational domain of data mining using a case study from Database class. It showed what kind of data could be collected, how could we preprocess the data, how to apply data mining methods on the data, and finally how can we benefited from the discovered knowledge. There are many kinds of knowledge can be discovered from data. In this work we investigated the most common ones which are association, classification, clustering and outlier detection.

The rest of the paper is organized as follows: Section 2 summaries related works in educational data mining. Section 3 gives a general description of the data we used in our case study. Section 4 describes the preprocess stage of the used data. Section 5 reports our experiments about applying data mining methods on the educational data. Finally we conclude this paper with a summary and an outlook for future work.

2. RELATED WORK

Although, using data mining in higher education is a recent research field, there are many works in this area. That is because of its potentials to educational institutes. [14] have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards educational domain of data mining.

[10] gave a case study that used educational data mining to identify behavior of failing students to warn students at risk before final exam. [15] gave another case study of using 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 students' final grade using data collected from Web-based system. [3] used educational data mining to identify and then enhance educational process in higher educational system which can improve their decision-making process. Finally, [17] used data mining to assist in development of new curricula, and to help engineering students to select an appropriate major.

3. DATA COLLECTION

In our case study we collected the students data from data based management system course held at the Islamic University of Gaza in the first semester of 2007/2008. The number of students was 151. The sources of collected data were: personal records and academic records of students, course records and data came from e-learning system. For e-learning system the course used Moodle which is a well known open source course management system [12]. From Moodle, first, we collected information about student accessing e-learning, where it appeared that some students did not access the system at all. Then, we got information about how much student benefited from resources, such as using ebooks, research papers and old exams available on the system. Also, we got the results of students' grades in solving exercises available in the system.

4. DATA PREPARATION AND PREPROCESSING

To get better input data for data mining techniques, we did some preprocessing for the collected data. After we integrated the data into one file, to increase interpretation and comprehensibility, we discretized the numerical attributes to categorical ones. For example, we grouped all grades into five groups *excellent, very good, good, poor* and *failure*. In the same way, we discretized other attributes such as attendance and resource access.

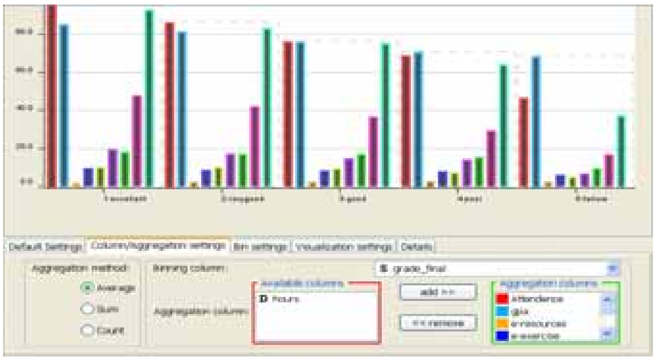


Figure 4.1: visualizing data used in the case study using Knim data mining system

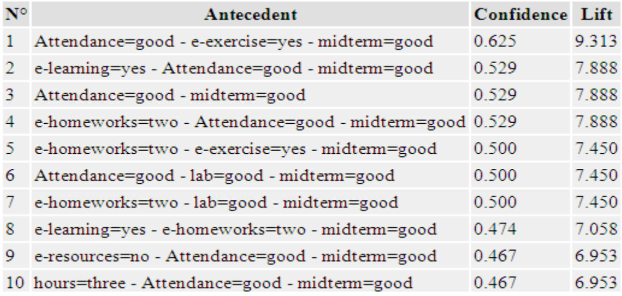
By using some preprocessing techniques, such as visualization, we can get some primary useful knowledge. For example, using Knim, which is an open source data mining system from university of Konstanz, Germany, we visualized students' data [6]. From this visualization some useful knowledge has been drawn about the attributes before applying data mining methods. By using histogram of the data as in graph (figure 4.1), we discovered that attendance, students' GPAs, and lab grades has a positive relationship with the final grade. However, e-learning facilities such as e-recourses used by student, exercises and assignment hardly affected the final grade of student.

5. DATA MINING TASKS IN EDUCATIONAL SYSTEMS

Data mining used advanced techniques to discover patterns from data. The data mining tasks are the kinds of patterns that can be mined. There are many tasks in data mining, the most common ones are: Association, classification, clustering and outlier detections. In the following sections describes the results of applying data mining techniques to the data of our case study for each of the four tasks.

5.1 Association Rules

Mining association rules searches for interesting relationships among items in a given dataset [1]. It allows finding rules of the form *If antecedent then (likely) consequent* where *antecedent* and *consequent* are itemsets [7]. Itemsets are sets of one or more items. In our dataset an example of item is: *attendance = good.* Because, we are looking for items that characterize the final grade of students, *consequent* has one item which is *final_grade= z* where *z* is one value of the final grade such as *excellent*, *very good*,…etc. Figure 5.1 is sample of association rules discovered from data for *excellent* final grade students.



N°	Antecedent	Confidence	Lift
1	Attendance=good - e-exercise=yes - midterm=good	0.625	9.313
2	e-learning=yes - Attendance=good - midterm=good	0.529	7.888
3	Attendance=good - midterm=good	0.529	7.888
4	e-homeworks=two - Attendance=good - midterm=good	0.529	7.888
5	e-homeworks=two - e-exercise=yes - midterm=good	0.500	7.450
6	Attendance=good - lab=good - midterm=good	0.500	7.450
7	e-homeworks=two - lab=good - midterm=good	0.500	7.450
8	e-learning=yes - e-homeworks=two - midterm=good	0.474	7.058
9	e-resources=no - Attendance=good - midterm=good	0.467	6.953
10	hours=three - Attendance=good - midterm=good	0.467	6.953

Figure 5.1: Associations rules for student data

These rules are sorted by *lift* metric. The *lift* value is the ratio of the confidence of the rule and the expected confidence of the rule [16]. The *lift* is measured as the ratio of the probability of antecedent and consequent occurring together to the probability of antecedent and consequent occurring independently. The *lift* value of greater than 1 indicates a positive correlation between antecedent and consequent. For example the first rule with *lift* is 9.313 means there is a high positive correlation between the antecedent good attendance, doing exercise in e-learning and has good midterm grade, and the consequent final grade *excellent*. With the *lift* value, we can interpret the importance of a rule. The first rule, with the highest lift which means highest correlation is the most important, and so on.

For more understanding of the association rules, a graph can be constructed. For example, figure 5.2 represents rules for final grade *fail*. From the graph we can see some attributes happens more frequent than others such as failing in midterm.

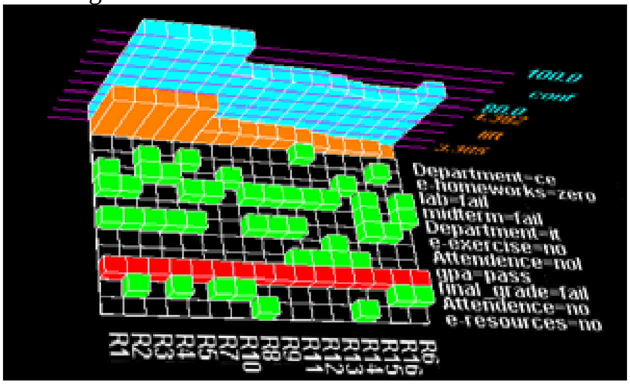


Figure 5.2 Associations rules Graph for students with *grade* fail using Arviewer [2]

5.2 CLASSIFICATION

Classification is a data mining task that predicts group membership for data instances [7]. In educational data mining, given works of a student, one may predicate his/her final grade [15]. In our case study we used J48 decision tree to represent logical rules of student final grade. The represented tree is large, some of the strong rules in the tree are:

If midterm =good and attendance=good then final grade= excellent
If midterm = fail and lab = fail the final grade = fail
If midterm = average and gpa=pass then final grade = pass
If midterm = average and gpa= good and e-homework=one then final grade=pass
If midterm = average and gpa= good and e-homework=two attendance = average then final grade=good
If midterm=average and gpa=verygood the final grade=very good

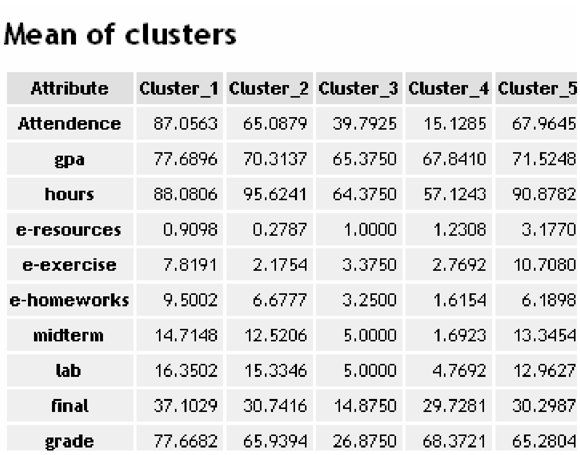
The benefit of this method is that it can predict low grades on time. For example the instructor can predict *fail* students before the end of the semester and he may work on them to improve their performance before the final.

It is important to know that classification rules are different than rules generated from association. Association rules are characteristic rules (it describes current situation), but classification rules are prediction rules (it describes future situation).

5.3 CLUSTERING

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group [7]. In educational data mining, clustering has been used to group students according to their behavior. For example, Romero in [18] used clustering to distinguish *active* students from *non-active* according to their performance in activates. According to this clustering, instructor groups *active* students with *non-active* students for better students' performance.

In our case we used Expectation-Maximization Algorithm (EM-clustering) to cluster the given data. An EM algorithm [5][4] is a mixture based algorithm that finds maximum likelihood estimates of parameters in probabilistic models. In our case, we used EM-clustering to group students according to their performance. Figure (5.3) gives Mean of each cluster for each attribute. Using these results we can divide students into five groups and guide them according to their behavior.



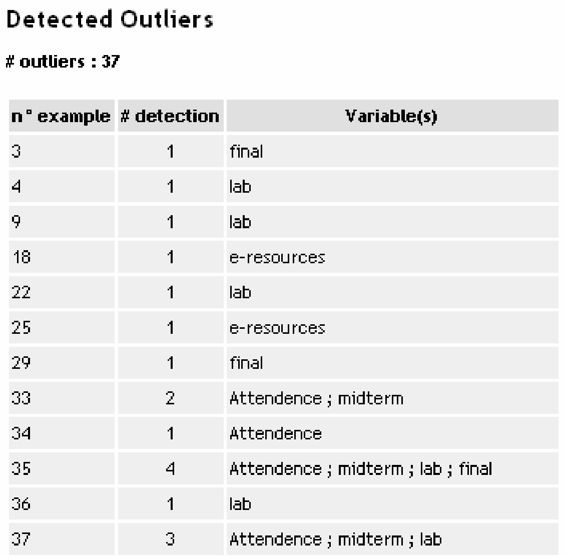
Mean of clusters

Attribute	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5
Attendence	87.0563	65.0879	39.7925	15.1285	67.9645
gpa	77.6896	70.3137	65.3750	67.8410	71.5248
hours	88.0806	95.6241	64.3750	57.1243	90.8782
e-resources	0.9098	0.2787	1.0000	1.2308	3.1770
e-exercise	7.8191	2.1754	3.3750	2.7692	10.7080
e-homeworks	9.5002	6.6777	3.2500	1.6154	6.1898
midterm	14.7148	12.5206	5.0000	1.6923	13.3454
lab	16.3502	15.3346	5.0000	4.7692	12.9627
final	37.1029	30.7416	14.8750	29.7281	30.2987
grade	77.6682	65.9394	26.8750	68.3721	65.2804

Figure 5.3: Clustering students into five groups using EM-Clustering Algorithm

5.4 OUTLIER DETECTION

Outlier detection discovers data points that are significantly different than the rest of the data [9]. In educational data mining outlier analysis can be used to detect students with learning problems [14]. In our case study, we used outlier analysis to detect outliers in the student data. The system detected 37 outliers in our data. Figure (5.4) is a sample of instances which detected as an outlier and the attribute where the outlier occurred. For each case instructor can look at the outlier behavior of the student and try to find and understand why the irregularity happened and then resolve the problem if there is any.



Detected Outliers

# outliers : 37

n° example	# detection	Variable(s)
3	1	final
4	1	lab
9	1	lab
18	1	e-resources
22	1	lab
25	1	e-resources
29	1	final
33	2	Attendence ; midterm
34	1	Attendence
35	4	Attendence ; midterm ; lab ; final
36	1	lab
37	3	Attendence ; midterm ; lab

Figure 5.4: Outlier analysis of student data

6. CONDUCTION AND FUTURE WORK

In this paper, we gave a case study in educational data mining. It showed how useful data mining can be in higher education in particularly to improve student performance. We used students' data from database course. We collected all available data including their usage of Moodle e-learning facility. We applied data mining techniques to discover knowledge. Particularly we discovered association rules and we sorted the rules using lift metric then we visualized the rules. Then we discovered classification rules using decision tree. Also we clustered the student into group using EM-clustering. Finally, using outlier analysis we detected all outliers in the data. Each one of these knowledge can be used to improve the performance of student.

For future work, a way to generalize the study to more diverse courses to get more accurate results. Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbor, Naive Bayes, support vector machines and others. Finally, the used preprocess and data mining algorithms could be embedded into e-learning system so that any one using the system can benefited from the data mining techniques.

REFERENCES:

[1] Agrawal, R. Imielinski, T. Swami, A., "Mining Association Rules between Sets of Items in large database". In proceedings of the ACM SIGMID Conferences on Management of Data, Page 207-216, Washington, D.C. May. 1993

[2] Arviewer, http://www2.lifl.fr/~jourdan/download/arv.html, 2008

[3] Beikzadeh,M. and Delavari, N., "A New Analysis Model for Data Mining Processes in Higher Educational Systems". On the proceedings of the 6th Information Technology Based Higher Education and Training 7-9 July 2005.

[4] Bradley, P. Fayyad, U. and Renia C., "Scaling EM clustering to large databases". Technical Report. Microsoft Research. 1999

[5] Dempster ,A. Larid N., Rubin,D. "Maximum Likehood estimation from incomplete data via EM Algorithm". Journal of the Royal Statistics Society, 39 (1) : 1- 38. 1977.

[6] knime <www. Knime.com> , 2008

[7] Han,J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor. 2006.

[8] Machado, L. and Becker, K. "Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites". Third IEEE International Conference on Advanced Learning Technologies (ICALT'03), 2003.

[9] Mansur, M. O. and Sap, M. Noor , M. "Outlier Detection Technique in Data Mining: A Research Perspective". In Postgraduate Annual Research Seminar. 2005

[10] Merceron, A. and Yacef, K.,"Educational Data Mining: a Case Study" In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press. 2005

[11] Minaei-Bidgoli B., Kashy, D. Kortemeyer G., Punch W., "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System". In the Processing of 33rd ASEE/IEEE conference of Frontiers in Education. 2003

[12] Moodle, <www.Moodle.com> 2008

[13] Mostow,J and Beck , J., "Some useful tactics to modify , map and mine data from intelligent tutors". Natural Language Engineering 12(2), 195-208. 2006

[14] Romero,C. and Ventura, S. ,"Educational data Mining: A Survey from 1995 to 2005".Expert Systems with Applications (33) 135-146. 2007

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

[16] Sheikh,L Tanveer B. and Hamdani,S., "Interesting Measures for Mining Association Rules". IEEE-INMIC Conference December. 2004.

[17] Waiyamai,K. "Improving Quality of Gradate Students by Data Mining" Department of Computer Engineering. Faculty of Engineering. Kasetsart University , Bangkok, Thailand. 2003.