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 collected from 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](image1)

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.

![Figure 5.1: Associations rules for student data](image2)

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](image3)

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:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>If midterm = good and attendance = good then final grade = excellent</td>
<td></td>
</tr>
<tr>
<td>If midterm = fail and lab = fail then final grade = fail</td>
<td></td>
</tr>
<tr>
<td>If midterm = average and gpa = pass then final grade = pass</td>
<td></td>
</tr>
<tr>
<td>If midterm = average and gpa = good and e-homework = one then final grade = good</td>
<td></td>
</tr>
<tr>
<td>If midterm = average and gpa = good and e-homework = two attendance = average then final grade = good</td>
<td></td>
</tr>
<tr>
<td>If midterm = average and gpa = very good final grade = very good</td>
<td></td>
</tr>
</tbody>
</table>

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] 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.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Cluster_1</th>
<th>Cluster_2</th>
<th>Cluster_3</th>
<th>Cluster_4</th>
<th>Cluster_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td>87.058</td>
<td>65.087</td>
<td>39.792</td>
<td>15.128</td>
<td>67.964</td>
</tr>
<tr>
<td>gpa</td>
<td>77.608</td>
<td>70.313</td>
<td>65.276</td>
<td>67.041</td>
<td>71.524</td>
</tr>
<tr>
<td>hours</td>
<td>60.000</td>
<td>95.624</td>
<td>64.376</td>
<td>57.124</td>
<td>90.078</td>
</tr>
<tr>
<td>e-resources</td>
<td>0.909</td>
<td>0.279</td>
<td>1.000</td>
<td>1.239</td>
<td>3.173</td>
</tr>
<tr>
<td>e-exercise</td>
<td>2.819</td>
<td>2.176</td>
<td>3.276</td>
<td>2.769</td>
<td>10.700</td>
</tr>
<tr>
<td>e-homeworks</td>
<td>9.503</td>
<td>6.677</td>
<td>3.293</td>
<td>1.615</td>
<td>6.189</td>
</tr>
<tr>
<td>midterm</td>
<td>14.714</td>
<td>12.831</td>
<td>5.000</td>
<td>1.892</td>
<td>13.345</td>
</tr>
<tr>
<td>lab</td>
<td>15.392</td>
<td>15.334</td>
<td>5.000</td>
<td>4.769</td>
<td>12.967</td>
</tr>
<tr>
<td>final</td>
<td>37.102</td>
<td>30.741</td>
<td>14.871</td>
<td>29.728</td>
<td>30.297</td>
</tr>
<tr>
<td>grade</td>
<td>77.662</td>
<td>65.594</td>
<td>25.876</td>
<td>68.372</td>
<td>65.380</td>
</tr>
</tbody>
</table>

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.

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:


