Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns

Suhem Parack#, Zain Zahid#, Fatima Merchant#

Computer Engineering, M.H. Saboo Siddik College of Engineering

[1]sparack@gmail.com

[2]zain.zahid27@gmail.com

[3]fatimaamerchant@gmail.com

***Abstract*— Data mining is a process of identifying and extracting hidden patterns and information from databases and data warehouses. There are various algorithms and tools available for this purpose. Data mining has a vast range of applications ranging from business to medicine to engineering. In this paper, we discuss the application of data mining in education for student profiling and grouping. We make use of Apriori algorithm for student profiling which is one of the popular approaches for mining associations i.e. discovering co-relations among set of items. The other algorithm used, for grouping students is K-means clustering which assigns a set of observations into subsets. In the field of academics, data mining can be very useful in discovering valuable information which can be used for profiling students based on their academic record. We apply Apriori algorithm to the database containing academic records of various students and try to extract association rules in order to profile students based on various parameters like exam scores, term work grades, attendance and practical exams. We also apply K-means clustering to the same set of data in order to group the students. The implemented algorithms offer an effective way of profiling students which can be used in educational systems.**

***Keywords*— Data Mining, Apriori Algorithm, Student Profiling, K-means Clustering, Weka**

I. INTRODUCTION

The data in any given educational institution is growing rapidly. There is a need to transform this data into useful information and knowledge; hence we make use of data mining. Data mining is the process of extracting hidden, unknown and potentially useful information and patterns from databases, data warehouses or other such data repositories. Educational data mining is the area of science where various methods are being developed for making discoveries within data. This data is obtained from an educational background. These methods provide an insight into a student's behavioral patterns and the environment in which they learn [1].

Data mining can be applied to educational databases to identify undesirable student behavior which was previously unknown. We can construct coursework, plan and schedule classes, model students, predict their performance and provide recommendations for students, using data mining techniques [2].

In this paper, we make use of a widely used data mining tool called Weka. We apply Apriori algorithm to the academic records of a group of students and obtain the best association rules which will help in student profiling. We also use K-means clustering to group the students categorically and efficiently. In the first section, we discuss the Apriori and K-means algorithms. Then, we discuss how we implement these algorithms on the academic data. Finally, we study the results and derive useful information from them. In the end the conclusion is stated.

II. ALGORITHMS USED

As discussed earlier, data mining can be used for extracting previously unknown patterns like groups of data records as well as to extract dependencies between set of data items. The algorithm used to find these dependencies between the data items is Apriori algorithm, while the algorithm used to extract groups of records is K-means clustering.

A. Apriori Algorithm

The Apriori algorithm is a traditional data mining algorithm that is used to mine association rules from the given data [3]. The aim is to extract a set of strong association rules of the form X => Y i.e. items that satisfy condition X are most likely to satisfy Y also.

For example, in our case, we may find the association "if a student scores between 80-100 and if his term work grade is A, practical marks between 21-30 and his attendance is high then his academic profile is most likely to be good".

Suppose we have a set of items $I = \{i_1,i_2,\ldots,i_m\}$. Let D be a set of transactions where each transaction T be set of items such that $T \subseteq I$. A transaction T contains A i.e. a set of items if $A \subseteq T$. So, an association rule is of the form A => B where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. The rule X => Y has support s, in the transaction set D, if s percent of transactions in D contain X => Y. The confidence c means that c percent of transactions in D that contain X also contain Y. The steps involved in the Apriori algorithm are shown below [4].

```
Ck: Candidate itemset of size k
Lk: frequent itemset of size k

L1= {frequent items};
    for(k= 1; Lk!=∅; k++) do begin
Ck+1= candidates generated from Lk;
    for each transaction tin database do:
increment the count of all candidates in Ck+1that are contained in t
    Lk+1= candidates in Ck+1with min_support
end
return ∪kLk;
```

B. K-means Clustering

Given a database of n objects and k is the number of clusters to form a partitioning algorithm that organizes the objects into k partitions where k ≤ n and each partition represents a cluster. The main reason for forming clusters is, so that all the objects in a cluster are similar to each other, whereas objects of other clusters are dissimilar in terms of database attributes. The K-means does exactly this function. It separates clusters into a set of n objects where the inter cluster similarity is low and the intra cluster similarity is high. Cluster similarity is measured based on the mean value of the objects in the cluster.

The algorithm proceeds as follows. First, it randomly selects k of the objects where each represents a cluster mean or centre. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the clusters mean. It then computes a new mean for each cluster. This process iterates until the criterion function converges [5]. Typically the squared-error criterion is used, defined as follows:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}^{n}|x-m_i|^2$$

Where x is the point in space representing the given object, and m_i is the mean of the cluster C_i (both x and $m_i$are multidimensional). This criterion tries to make the resulting k clusters as compact and as separate as possible. The K-means procedure is summarized in the following steps [5].

INPUT: This is given as the number of clusters k and the database containing n objects.

OUTPUT: This is the set of k clusters which minimize the squared error criterion.

METHOD: The K-means algorithm is implemented as follows

1) k objects are chosen arbitrarily as the initial clusters
2) repeat
3) assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster
4) update the cluster means(calculate mean value of the objects in the cluster)
5) until no change in values

The algorithm attempts to determine k partitions that minimize the squared error function. It works well when all the clusters are compact clouds rather than well separated from each other. The method is relatively scalable and efficient in processing large datasets because the computational complexity of the system is $O(nkt)$ where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, k<<*n* and *t*<<*n*. The method often terminates at a local optimum.

However, the K-means method, can be applied only when the mean of the cluster is defined. This may not be the case in some applications, such as when data with categorical attributes is involved. The necessity for users to specify *k*, the number of clusters, in advance can be seen as a disadvantage. The K-means is not suitable for discovering clusters with non convex shapes, or clusters of very different size. Moreover it is sensitive to noise and outlier data points since a small number of such data can substantially influence the mean value.

III. WORKING AND RESULTS

We make use of a student academic record file. The various parameters and corresponding possible values are shown in table 1.

We have made use of Weka to mine best association rules which will help us extract information based on which we can profile the student's performance as Good, Satisfactory or Poor. Weka proves to be an efficient tool for analyzing and predicting the student behavior [6].

TABLE I
INPUT FILE OF STUDENT RECORD

Attribute	Possible values
Exam marks	80-100 60-79 40-69 0-40
Term work grades	A B C D
Attendance	High Low
Practical marks	0-10 11-20 21-30

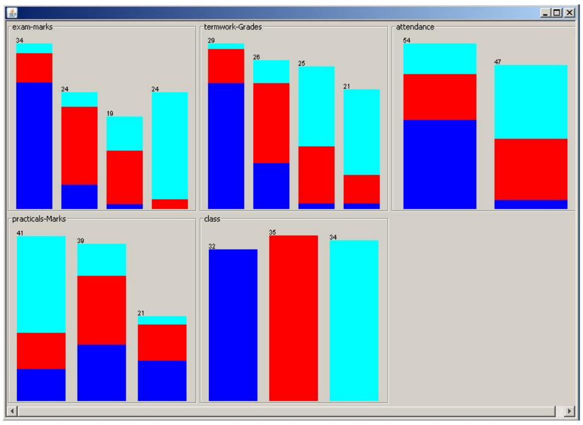
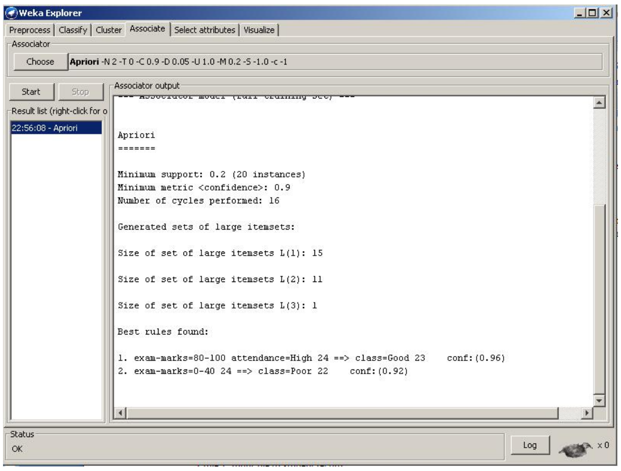


Fig. 1 Visualization of input data

The data can be better understood in the visualized manner as shown in figure 1.

We use Apriori algorithm on this set of data to obtain best association rules based on confidence with minimum support 20%. Since the data used to demonstrate this is not very large, we have made use of a low minimum support value. The higher the minimum support we provide, the stronger the association rules we obtain. Based on our input file, the best obtained rule states that if the exam marks are 80-100 and if the attendance is High, then the student can conveniently be profiled as Good. The output of this is shown in figure 2. along with the best association rules obtained that we can use to profile the students performance.



Apriori -N 2 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1

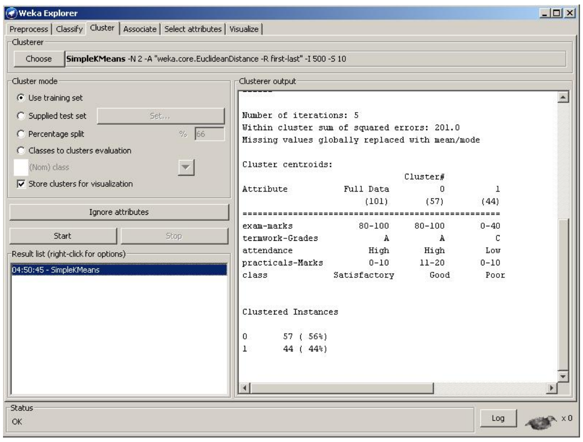
22:56:08 - Apriori

```
Apriori
=======

Minimum support: 0.2 (20 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 11

Size of set of large itemsets L(3): 1

Best rules found:

1. exam-marks=80-100 attendance=High 24 ==> class=Good 23    conf:(0.96)
2. exam-marks=0-40 24 ==> class=Poor 22    conf:(0.92)
```

OK

Fig. 2 Best obtained association rules

Clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Figure 3. shows the implementation of K-means clustering using Weka.



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) class
☑ Store clusters for visualization

04:50:45 - SimpleKMeans

```
Number of iterations: 5
Within cluster sum of squared errors: 201.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                   Cluster#
Attribute           Full Data          0          1
                        (101)       (57)       (44)
===================================================
exam-marks             80-100     80-100       0-40
termwork-Grades             A          A          C
attendance               High       High        Low
practicals-Marks         0-10      11-20       0-10
class             Satisfactory      Good       Poor


Clustered Instances

0       57 ( 56%)
1       44 ( 44%)
```

OK

Fig. 3 K-means clustering in Weka

In our observations we can see that there are totally 101 instances. These 101 instances have been partitioned into two clusters with 57 and 44 instances respectively. Clustering is mainly done to find the centers of the natural clusters in the data.

We can see from Figure 3. that the overall result of our student academic data is Satisfactory with exam-marks as 80-100, termwork-grades as A, attendance as High, and practical-marks as 0-10. The first cluster i.e. cluster 0 has overall result as Good with exam-marks as 80-100, termwork-grades as A, attendance as High, and practical-marks as 11-20. However, the second cluster i.e. cluster1 shows overall result as Poor with exam-marks as 0-40, termwork-grades as C, attendance as Low, and practical-marks as 0-10. Thus, based on the results we can easily group the students especially in cases where large academic records are present.

IV. Conclusions

Thus, based on the results of Apriori algorithm and K-means clustering in Weka, on the academic record file, we can deduce the important role that data mining can play in the field of education and teaching. It would be very difficult to manually go through the huge set of academic records to identify the student trends and behavior and the pattern in which they learn. Instead, if we make use of data mining techniques on the large amount of academic record, we can easily group the students, identify hidden patterns about their learning styles, find undesirable student behavior and perform student profiling. In this manner, data mining can certainly be an important tool and part of technologically advanced educational techniques.

V. Future Work

Further applications of the techniques we have used can be used to consider a higher number of parameters which will help us to predict more detailed information about the student under consideration. Possible applications of further prediction can be used to advise a student on the choice of his/her major based on parameters like interest in a certain field, number of marks scored, and hobbies. However on the other hand we could also help in predicting potentially violent behavior amongst students. We used the Apriori algorithm to find the association rules for a student and predict the performance. Similarly instead of considering only academic parameters we could also consider personal characteristics, behavior, family history, past records in order to predict if a student is prone to violence. Once we have done this, the K-means clustering method can be used to group the students into various clusters. After the clustering step is completed the Education Institutes can use special counseling methods in order to curb violence.

References

1. Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer, William F. Punch, ”PREDICTING STUDENT PERFORMANCE: AN APPLICATION OF DATA MINING METHODS WITH AN EDUCATIONAL WEB-BASED SYSTEM,” 33rd ASEE/IEEE Frontiers in Education Conference, 2003.

2. Cristobal Romero, Sebastian Ventura. “Educational Data Mining: A Review of the State of the Art,” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 40, NO. 6, NOVEMBER 2010.

3. WANG Pei-ji, SHI Lin, BAI Jin-niu, ZHAO Yu-lin. “Mining Association Rules Based on Apriori Algorithm and Application,” 2009 International Forum on Computer Science-Technology and Applications.

4. Anita Wasilewska. (undated). [Online]. APRIORI Algorithm Available: http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf

5. Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, 1st edition, Morgan Kaufmann Publishers, 2000.

6. Vasile Paul Bresfelean. “Analysis and Predictions on Students’ Behavior Using Decision Trees in Weka Environment,” in Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28, 2007, Cavtat, Croatia.