Predicting students performance in higher education: A Data Mining Approach

Ms.Khyati Manvar, Ms. Madhuri Rao

Abstract— To maximize the academic output of the students who are pursuing higher education, data mining is useful for finding valid pattern and extracting useful data. Here methodology is developed to predict the final achievement status of students based on their continuous assessment (test) and attendance status. Based on various data mining techniques (DMT) rules are derived that enable the classification of students in their predicted classes.

Index Terms— Apriori, k-mean, prediction, trends and patterns

.

1 INTRODUCTION

The objective of this paper is to predict the academic trends and patterns. The academic data is increasing day by day and they are used only for doing statistical analysis [1]. It is not used much whereas it contains lot of information and we can gain knowledge from these huge databases by interpreting and applying data mining techniques.

One of the major problems that the education system facing is predicting the behavior of students from large databases. To enhance the education system, we need to learn behavior of students who are obtaining higher education.

Data Mining provides various tools and techniques which can be used for improving education.It provides clustering method with which we can categorise performance of the students as Excellent, Good or Poor. The classification algorithm of data mining can be used to profile student based on some parameters like exam score, attendance etc.For students profiling best suitable association rules can be formulated using apriori algorithm.

2 LITERATURE REVIEW

2.1 A Novel Approach for Upgrading Indian Education by Using Data Mining Techniques by Banumathi.A, Pethalakshmi.A

Here the researcher has taken k-mean approach of data mining and proposed a new approach of clustering UCAM. They applied these algorithms on student database. Their input parameter were

Table 2.1: Students Sample Data[1]

Stud-no	Age	Mark1	Mark2	Mark3
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	89
S7	20	70	65	60
S8	21	53	56	59
S9	19	82	82	60
S10	47	75	76	77

1. Result of k-mean algorithm with three clusters is as below.
C1 = {S1, S9}
C2 = {S2, S5, S6, S10}
C3= {S3, S4, S7, S8}
2. Result of implementing UCAM algorithm with five clusters is as below.
C1 = {S1, S3, S7}
C2 = {S2, S5, S6}
C3= {S4, S8}
C4 = {S9}
C5= {S10}

They have given comparative analysis between k-mean and UCAM as below.

	Initial Seeds	Centroid	Threshold Value	Cluster Result	Cluster Error
k-mean	K	Initial seeds	-	Depends on initial seeds	Yes, if wrong seeds
UCAM	-	-	T	Depends on threshold value	-

Table 2.2: Comparative analysis on k-means and UCAM

2.2 Application of data mining in educational database for predicting academic trends and patterns by Parack, S., Zahid Z., Merchant, F

Here researcher has taken the approach of applying k-mean and Apriori algorithm on student database. They analyzed and predict the student's performance as Good, Satisfactory and Poor using association rule and k-means algorithm with the help of Weka.

Table 2.3: Input File of Student Records[2]

Attribute	Possible Values
Exam Marks	80-100 60-79 40-69 0-40
Termwork grades	A B C D
Attendance	High Low
Practical Marks	0-10 11-20 21-30

1. Result based on input parameters by implementing Apriori Algorithm

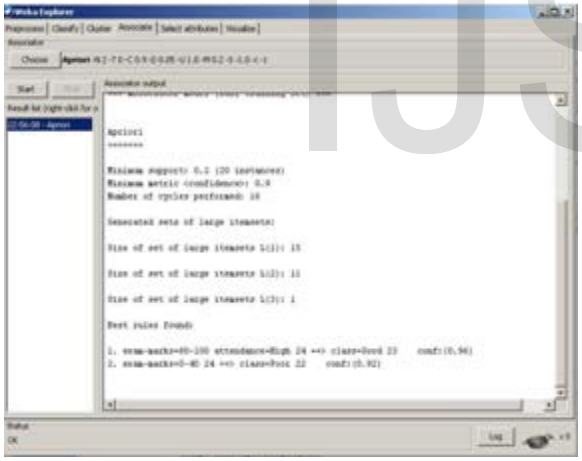

Fig 2.1 Best Obtained Association Rules[2]

2. Result based on input parameters by implementing K-means Algorithm

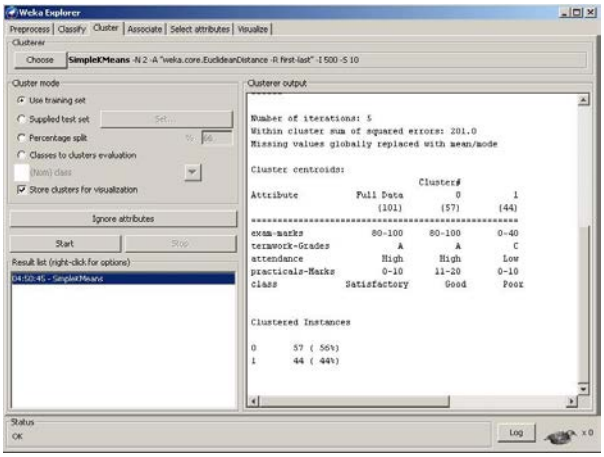

Fig 2.2 K-means clustering in Weka[2]

3 IMPLEMENTATION STRATEGY

The sample size of data of more than 500 students of MCA course is collected from Hiray College, Bandra (E) of academic year 2009-2012.During the analysis of this quantitative data, the following block diagram is evolved.

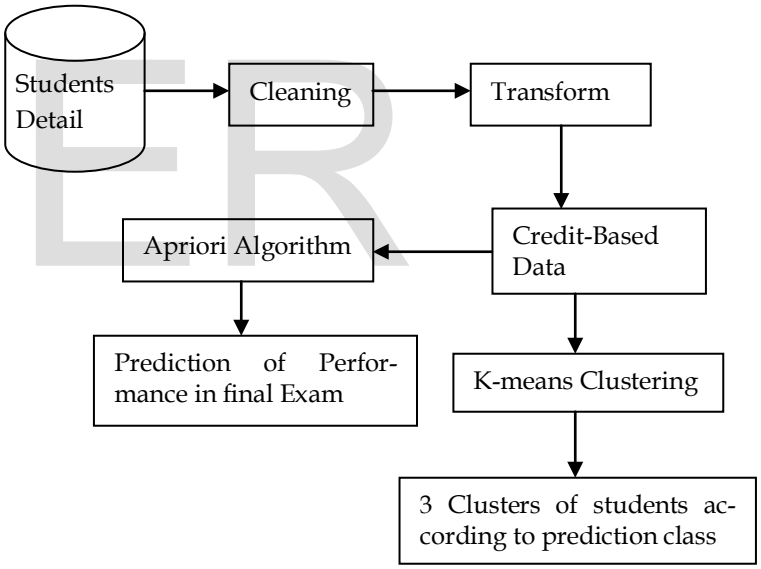

Fig. 3.1 Block Diagram of the System

Above Fig 3.1 shows the working of the system. After collecting the student details the unwanted tuples are removed during data cleaning stage. This data is then transformed into credit-based data. During transformation stage, test marks are converted into SGPA (Semester Grade Point Average) for making data credit-based. To calculate SGPA each subject has assigned a credit.
They are as below.

Subject 1	4
Subject 2	4
Subject 3	4
Subject 4	2
Subject 5	2

The formula for the SGPA is as follows:

$$SGPA = \frac{\sum_{j=1}^{n} C_j P_j}{\sum_{j=1}^{n} C_j} \quad (1)$$

Where

Cj= Number of credits of the jth course
Pj=Grade Point in the jth course, and
N=Number of courses in which a student is registered in the concerned semester.

On this credit-based data, 2 algorithms had been applied. One is k-means clustering which is used to make groups or clusters of data and other is Apriori algorithm, which is used to find strong association rule in the data.
In K-means clustering, 3 clusters have been formed based on SGPA. These clusters are Poor, Good and Excellent. K-means gives the number of the students and their percentage in each cluster from the student's database.

The input for these algorithms is as below.

Table - dbo.Student_data | Summary

id	Attendence	ClassTest 1	ClassTest2	Performance
1	High	A	B	Average
2	Medium	B	B	Average
3	Medium	A	B	Excellent
4	Medium	A	B	Excellent
5	High	B	A	Excellent
6	High	A	A	Average
7	High	A	A	Average
8	High	A	A	Excellent

Fig 3.2 Student Data Table

Fig 3.2 shows the Student Data Table which contains data like Attendance, Class test 1 marks, Class test 2 Marks and Performance.

3.1 Clustering Algorithm

For implementing clustering algorithm on student database, K-means clustering algorithm is used. According to this algorithm the data is subdivided into three clusters those are Excellent, Good and Poor. It helps to predict performance of students on the basis of Attendance and Class test marks.

3.2 Association Rule

For the implementation of association rule Apriori algorithm is used. This algorithm is predicting the performance of the students on the basis of minimum support 0.2 and confidence 0.8. It generates best six rules (refer figure 4.2).

4 Research Finding

4.1 Analysis of Clustering

K-means tab shows a button Get Clusters, on click of which 3 clusters are being generated. Each cluster shows range of SGPA and total number of tuples having that range. And finally a summary shows that
32students are in the first cluster (Poor) comprising 12.17% of overall students.
101 students are in the second cluster (Good) comprising 38.40% of overall students.
130 students are in the third cluster (Excellent) comprising 49.43% of overall students.
The overall result of the student is Excellent.

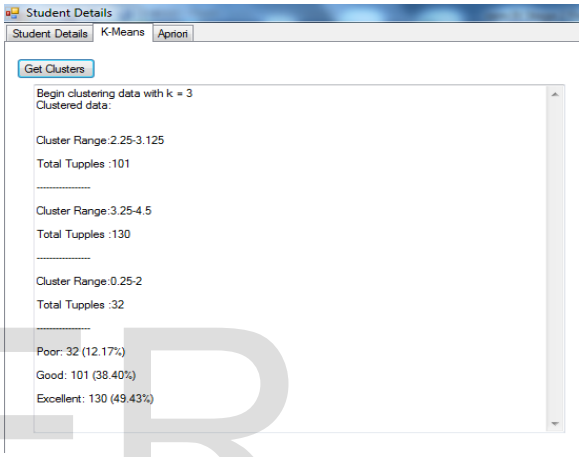

Fig 4.1 Output of K-means

4.2 Analysis of Association Rules

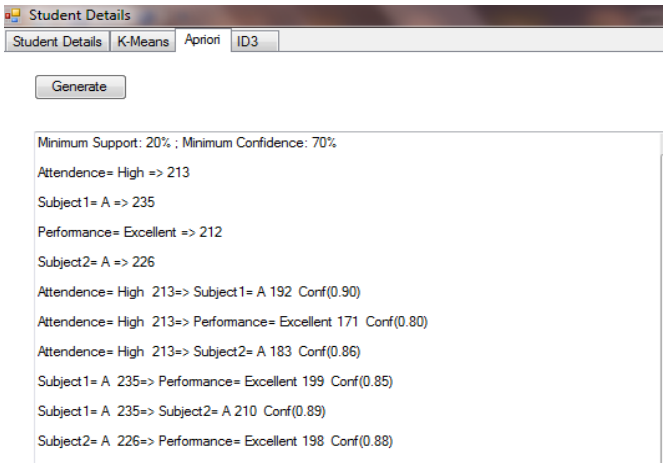

Fig 4.2 Output of Apriori

This shows the result of the application of Apriori algorithm. Here minimum support is 20% and minimum confidence is 70%. The rules says that

If attendance is High then confidence of having A grade in

class test 1 is 90%

If student has A grade in classtest1 then confidence of having Excellent performance is 85%.

If student has A grade in classtest1 and A grade in classtest2 then confidence of having Excellent performance is 89%. Etc.

5 CONCLUSION

Hence we can conclude that with the use of k-means and Apriori we can find patterns in students based on the academic records. Going through the huge database manually is difficult. With the use of data mining, we can easily find useful patterns and can predict the behavior of students.

ACKNOWLEDGMENT

I wish to thank the Admin department of Hiray College, for providing the data. Without their cooperation this study would not have been possible.

My sincerest gratitude goes to my project-guide, Ms. Madhuri Rao, for her useful guidance and strong support. But for her guidance and advice this project could not have been completed successfully and within the time limits.

REFERENCES

[1] Bhanumathi A, Pethalakshmi A, "A Novel Approach for Upgrading Indian Education by Using Data Mining Techniques", pp. 1-5, 2012.

[2] Parack S., Zahid Z., Merchant, F, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns", IEEE Conference on Technology Enhanced Education (ICTEE), pp. 1-4, 2012.

[3] Emmanuel N. Ogor, "Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques", Department of Natural Sciences Turks & Caicos Islands Community College Turks & Caicos Islands, pp. 354-359, 2007.

[4] Liu Kan, Xiao Xingyuan, Liu Ping "DMCMS: A Data Mining Based Course Management System", Second International Workshop on Education Technology and Computer Science, pp. 145-148, 2010.

[5] G.K Gupta, Introduction to Data Mining with case studies , PHI Learning Private Limited Y.

[6] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Pearson Education