Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 7, July 2013, pg.273 – 279

RESEARCH ARTICLE

Predicting Students Academic Performance Using Education Data Mining

Suchita Borkar[1], K. Rajeswari[2]
[1]MCA, Pune University, PCCOE, India
[2]Computer Engineering, Pune University, PCCOE, India

[1] *suchitaborkar1@gmail.com;* [2] *raji.pccoe@gmail.com*

Abstract— Education Data Mining is a promising discipline which has an imperative impact on predicting students' academic performance. In this paper, student's performance is evaluated using association rule mining algorithm. Research has been done on assessing student's performance based on various attributes. In our study important rules are generated to measure the correlation among various attributes which will help to improve the student's academic performance. Experiment is conducted using Weka and real time data set available in the college premises.

Key Terms: - Educational Data Mining; Apriori algorithm; Association Rule Mining; Correlation coefficient

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into important information so as to identify hidden patterns from a large data set. **Educational Data Mining** (EDM) is an emerging discipline, concerned with data from academic field to develop various methods and to identify unique patterns which will help to explore student's academic performance. EDM can be considered as learning science, as well as an feature of data mining [11].Assessing students learning process is a very complex issue .Education Data mining helps in predicting students' performance in order to recommend improvements in academics.

The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a tool by which academic institutions extract useful unknown information in the student result repositories in order to improve students' learning processes [10].The main objective of this paper is prediction of student's performance in university result on the basis of their performance in Unit test, assignment, graduation percentage and attendance.

II. PREVIOUS STUDIES

A number of studies have been made in education data mining for discovering different pattern to improve the students' performance.

Ali Buldu and Kerem Üçgüna studied the use of data mining techniques using apriori algorithm on set of students of Istanbul Eyup I.M.K.B.Vocational Commerce High School, to reveal the relation between the courses that the students failed. They have taken the dataset of 28 students for 74 courses for minimum support rate 9 and as minimum confidence rate 85%.In their study they have revealed that if a student failed in particular subject in class 9th then he will fail in next year as well. The study discovers the rate of successful students by finding the rate of unsuccessful students which will help the student for choosing the right subject. [1]

Mahendra Tiwari, Randhir Singh, Neeraj Vimal, have conducted studies on engineering students by using various data mining methods for decision making. They have applied data mining techniques for discovering knowledge, association rules, and classification rules to predict the engineering students' performance, also they have clustered the students in groups using k-means clustering algorithm. [2]

Baha Sen and Emine Ucar compared the achievements of Computer Engineering Department students in Karabük University according to various factors such as age, gender, type of high school graduation and the students studying in distance education or regular education through data mining techniques. They have taken the dataset of 3047 records. In their study they have used NN architecture called multilayer perceptron (MLP) with back propagation type supervised-learning algorithm to produce both classification and regression type prediction models and decision tree for achieving the highest possible prediction accuracy. [3]

Brijesh Kumar Baradwaj and Saurabh Pal have discussed on how to achieve high quality in higher education. They have made use of various data mining algorithms like classification algorithm to estimate the accuracy of data. Clustering algorithm is used to cluster the objects which are used as a preprocessing approach for attributes. Regression technique is used as one of the prediction method to interpret the relationship between one or more independent variables and dependent variables. Association rules are used to find the correlation between frequent item set with confidence value less than one. Neural Network is used to derive patterns from complicated or imprecise data. Automatic Interaction Detection (CHAID).Through this study Brijesh Kumar Baradwaj and Saurabh Pal tried to identify weak students needing special attention. [4]

M. Ramaswami and R. Bhaskaran developed a predictive data mining model to identify academically weak students and attributes that affect their performance using CHAID prediction model. The attributes are selected on the basis of chi-square values. If chi-square values of attributes are greater than 100 they are given due considerations and the highly influencing variables with high chi-square values. [7]

In our review we have collected the data of MCA students to predict the performance of the students in university examination. The study was done on the dataset of 60 students. In our study we have identified the attributes which impact student's academic performance.

Year	Author	Significant Work
2010	Ali Buldu, Kerem Üçgün [1].	Ali and Kerem study the dataset of students and try to find the association between the student performance and course. In their finding they have generated rule that shows if a candidate is unsuccessful in numerical course in 9^{th} class then those students are likely to be unsuccessful in 10^{th} class. Such results are generated for different course. Through this study student can be helped to choose the profession by revealing the relation between their concern fields.
2013	Mahendra Tiwari, Randhir Singh, Neeraj Vimal. [2]	Authors conducted a study on engineering student to evaluate the performance by applying data mining techniques which will help for decision making as well as cluster the students by using K-Means algorithm. The result predicted in the study was if students are poor in attendance and assignment then there is 75% probability that their grades are poor.
2012	Baha Sen, Emine Ucar.[3]	Baha Sen and Emine Ucar evaluate the engineering student's performance of distance education using data mining technique. The results reveals that as the age of the student increases the success score decreases and students success rate is much better in distance than in formal education, students come from vocational high school are more successful in cultural lessons than vocational lesson.
2011	Brijesh Kumar Baradwaj and Saurabh Pal[4]	Brijesh Kumar Baradwaj and Saurabh Pal describes students' assessment by using various data mining methods .The study facilitate students and teacher to identify students needing special attention which will reduce the fail percentage and help to take appropriate measure for next semester.
2010	M. Ramaswami and R. Bhaskaran	M. Ramaswami and R. Bhaskaran explain the student's performance by using CHAID prediction model. The study also shows the classification model that will show comparative study of observed data and data predicted. The result reveals CHAID model classifies correctly 345 students from 772.

TABLE 1: SURVEY ON EDUCATION DATA MINING

III. DATA COLLECTION AND PREPARATION

In our study, we have considered student's data that are pursuing Master of Computer Application (MCA) degree from Pune University. On the basis of the data collected some attributes have been considered to predict student's performance in the university examination. The Variables used for judging the students' performance in university results are Graduation%, Attendance%, Assignment%, UnitTest% and UniversityResult%.

Attributes	Description	Values
Graduation%	Percentage of marks obtained in graduation.	Good, Avg, Poor
Attendance	Attendance of the student.	Good, Avg, Poor
Assignment	Assignment performance given during the semester.	Good, Avg, Poor
Unit Test Performance	Percentage marks obtained by a student in Unit Test.	Good, Avg, Poor
University Result	Percentage marks obtained by the student in university examination.	Good, Avg, Poor

TABLE 2. ATTRIBUTES AND ITS POSSIBLE VALUES

IV. DATA PREPROCESSING

One of the important steps of Data Mining process is data pre-processing. Data Pre-processing is used in identifying the missing values, noisy data and irrelevant and redundant information from dataset. We use the data in percentage for the above mentioned attributes.

TABLE 3.CATERGORIZATION OF ATTRIBUTES

Attribute	Range
Graduation%	Graduation% >=70% = Good. 60% <=Graduation% <70% =Avg. Graduation% >60% =Poor.
Attendance%	Attendance% >=70% = Good. 60% <=Attendance% <70% Avg. Attendance%60% =Poor.
Assignment %	Assignment% >=70% = Good. 60% <=Assignment% <70% Avg. Assignment %>60% =Poor.
Unit Test%	Unit Test% >=70% = Good. 60% <= UnitTest% <70% =Avg. UnitTest %> 60% =Poor.
University Result%	University Result% >=70% = Good. 60% <= UniversityResult% <70% =Avg. University Result %> 60% =Poor.

V. METHODOLOGY

In this paper, we have used Weka tool free software, implemented in Java language which uses the same dataset external representation format (ARFF files). So, they can easily be obtained from Internet, used without data format problems and, if necessary, modified using the same programming language. Weka [10] is open source software that offers a collection of machine learning and data mining algorithms for data pre-processing, classification, regression, clustering, and association rules.

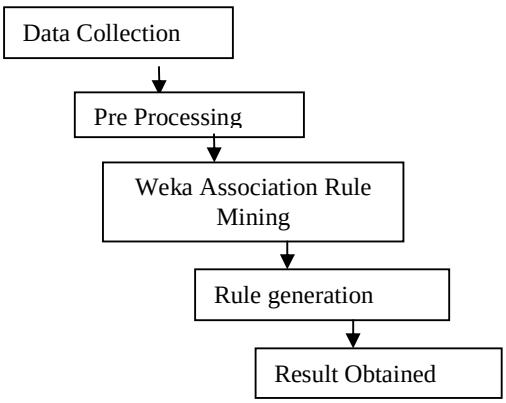

Fig 1.Work Methodology

This is the work methodology of our work, data collected from MCA students are pre-processed and using weka software important associated rules are generated for different values of confidence and support. On the basis of those rules some results are obtained. The result discussed in research and discussion.

A. Association Rule Mining

In education data mining, **association rule learning** is a conventional and well researched method for determining interesting relations between attributes in large databases [11]. Association rule Mining is mainly intended to recognize strong rules from databases using different measures support and confidence.

The preliminaries necessary for performing data mining on any data are discussed below.

Let I= $\{I_1, I2, I3_{\ldots}I_m\}$ be a set items. Let *D*, the task relevant data, be a set of database transactions where each transaction $T \subseteq I$. Each transaction is an association with an identifier, called transaction identification (TID). Let *A* be a set of items. A transaction *T* is said to contain *A* if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cup B = \emptyset$. [11]

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an item set. An item set that contains k items is a k-item set. The occurrence frequency of an item set is the number of transactions that contain the item set. If the relative support of an item set I satisfies a prescribed minimum support threshold, then I is a frequent item set. The association rule mining can be viewed as a two-step process:

***1)* Find all frequent item sets: Each of these item sets will occur at least as frequently as a predetermined minimum support count.**

***2)* Generate strong association rules from the frequent item sets: The rules must satisfy minimum support and confidence. These rules are called strong rules**. [11]

B. Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent item set in Apriori algorithm. [3]

Let Ck be a candidate item set of size k and Lk as a frequent item set of size k. The main steps of iteration are:

- Find frequent set Lk-1
- Join step: Ck is generated by joining Lk -1 with itself (Cartesian product Lk-1 x Lk-1)
- Prune step (apriori property): Any (k − 1) size item set that is not frequent cannot be a subset of a frequent k size item set, hence should be removed
- Frequent set Lk has been achieved [11].

VI. RESULTS AND DISCUSSION

The dataset of 60 students from MCA course was obtained from M.C.A department of Pimpri Chinchwad College of Engineering, Pune University. In this paper we find various association rules between attributes like students graduation percentage, Attendance, Assignment work, Unit test Performance and how these attributes affect the student's university result. Number of association rule can be found for different confidence values.

The analysis for generated association rules is as follows:

The rules generated for 90% confidence and 0.1 supports are:

1.Attendance=Good Assignment=Poor ==> UnitTest=Poor
<conf :(1)> lift :(1.28) lev :(0.02) [1] conv :(1.3)
2. Attendance=Good Assignment=Poor ==> UniversityResult=Poor
<conf :(1)> lift :(1.76) lev :(0.04) [2] conv :(2.6)

Rules for confidence 87% confidence and 0.1 supports are:

1. Attendance=Poor Assignment=Good ==> UnitTest= Poor UniversityResult=Poor
<conf :(0.86)> lift :(1.9) lev :(0.05) [2] conv :(1.93)
2. Assignment= Poor ==> UnitTest= Poor UniversityResult= Poor
<conf :(0.82)> lift :(1.82) lev :(0.07) [4] conv :(2.02)

Rules for confidence 70% confidence and 0.1 supports are:

1. Attendance=Good Assignment=Poor ==> UniversityResult= Poor
<conf :(1)> lift :(1.76) lev :(0.04) [2] conv :(2.6)
2. Attendance=good Assignment= Poor UnitTest= Poor ==> UniversityResult= Poor
<conf :(1)> lift :(1.76) lev :(0.04) [2] conv :(2.6)
3. Attendance=good Assignment= Poor ==> UnitTest= Poor UniversityResult= Poor
<conf :(1)> lift :(2.22) lev :(0.06) [3] conv :(3.3)
4. Graduate= Poor ==> UnitTest= Poor
<conf :(0.94)> lift :(1.2) lev :(0.04) [2] conv :(1.84)
5. Attendance=Avg ==> UnitTest= Poor
<conf :(0.92)> lift :(1.18) lev :(0.03) [1] conv :(1.41)
6. Assignment= Poor ==> UnitTest= Poor <conf :(0.91)> lift :(1.16) lev :(0.02) [1] conv :(1.19)
7. Assignment= Poor ==> UniversityResult= Poor
<conf :(0.91)> lift :(1.6) lev :(0.06) [3] conv :(2.38)
8. Assignment= Poor UnitTest=poor ==> UniversityResult= Poor
<conf :(0.9)> lift :(1.59) lev :(0.06) [3] conv :(2.17)
9. Attendance= Poor ==> UniversityResult= Poor
<conf :(0.89)> lift :(1.57) lev :(0.05) [2] conv :(1.95)
10. Attendance= Poor UnitTest=Poor ==> UniversityResult= Poor
<conf :(0.89)> lift :(1.57) lev :(0.05) [2] conv :(1.95)
11. Attendance= Poor ==> UnitTest= Poor UniversityResult= Poor
<conf :(0.89)> lift :(1.98) lev :(0.07) [3] conv :(2.48)

The interpretation of the above association rules for different confidence values depicts that the students' performance will be poor in unit test if either their attendance is poor or assignment is poor or both. Also their university performance will be affected by the poor performance in unit test. So we can interpret that to get the good university performance student have to be good in their assignment, attendance and Unit Test. Also graduation performance will also have an impact on the student's Unit Test performance.

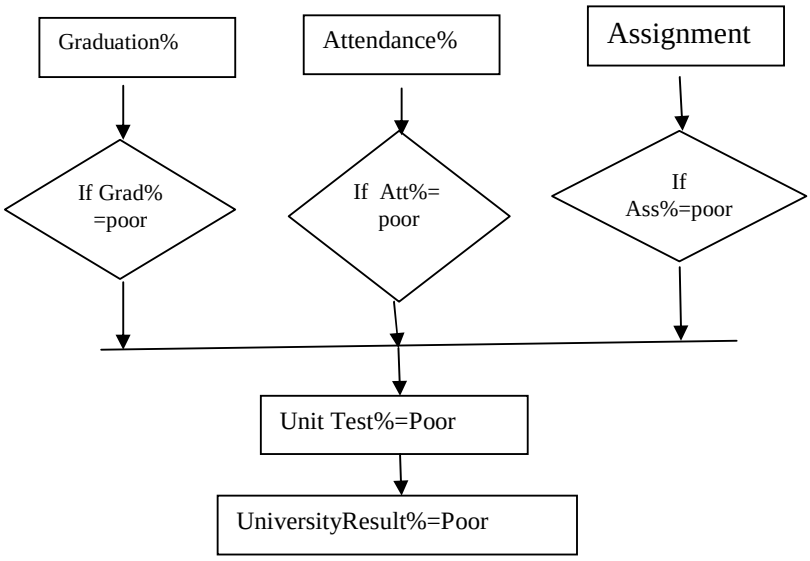

Fig 2. Interpretation of Association Rule Mining

Above figure shows the interpretation of rules generated from our data using apriori algorithm. The result shows that if a student is score poor in graduation and perform poor in attendance and assignment then there are chances that he/she will perform low in unit test. This will result in poor performance in University result. So to improve the student's performance in university result students should be perform good in graduation, attendance, assignment and unit test.

The above best associated attributes are also tested with statistical measures like correlation. **Correlation** refers the statistical relationships involving dependence. The correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The correlation coefficient, typically denoted *r* measures how strongly two numerical variables are linked to each other. The following is the formula for *r*, with a brief description of the variables:

- The number of data points is given by *n*.
- The capital sigma denotes addition of several terms.
- The data points are represented by x_i and y_i.
- The mean of the x_i is denoted by x with a horizontal bar over it.
- The standard deviation of the x_i is denoted by s_x.
- The mean of the y_i is denoted by y with a horizontal bar over it.
- The standard deviation of the y_i is denoted by s_y. [12]

The correlations obtained on statistical measure are as follows:

TABLE 4.CORRELATION RESULT

Attributes	Correlation Coefficient	Result
Graduation% Vs UnitTest%	0.20779	Weak Correlation
Attendance% Vs UnitTest%	0.564654	Moderate Correlation
Assignment% Vs UnitTest%	0.622952	Moderate Correlation
UnitTest% Vs UniversityResult%	0.193403	Weak Correlation
Attendance% Vs University Result%	0.206344	Weak Correlation
Assignment% Vs UniversityResult%	0.45224	Moderate Correlation

VII. CONCLUSIONS

The paper presented the potential use of education data mining using association rule mining algorithm in enhancing the quality and predicting students' performances in university result. The analysis revealed that student's university performance is dependent on Unit test, Assignment, Attendance and graduation percentage. The results reveal that the student's performance level can be improved in university result by identifying students who are poor unit Test, Attendance, Assignment and graduation and giving them additional guidance to improve the university result. From the above correlation coefficient values, we observed that the associations that we are getting from apriori algorithm are not identical with the correlation values of the attributes. In our future work this will be further analysed. Also our next paper will include fuzzy logic in categorized values as good, average and poor for better results.

ACKNOWLEDGEMENT

The authors wish to acknowledge MCA department of Pimpri Chinchwad College of Engineering for their support in providing the necessary data.

REFERENCES

[1] Ali Buldua, Kerem Üçgün,. Data mining application on students' data. Procedia Social and Behavioral Sciences 2 5251–5259, 2010.
[2] Singh, Randhir. An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education, 2013.
[3] Baha Sen, Emine Ucar. Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology 1 262 – 267, 2012.
[4] Baradwaj, Brijesh Kumar, and Saurabh Pal. Mining Educational Data to Analyze Students' Performance. arXiv preprint arXiv: 1201.3417, 2012.
[5] Castro, Félix, et al. Applying data mining techniques to e-learning problems. Evolution of teaching and learning paradigms in intelligent environment. Springer Berlin Heidelberg, 183-221, 2007.
[6] Huebner, Richard A. "A survey of educational."
[7] Ramaswami, M., and R. Bhaskaran. A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv: 1002.1144, 2010.
[8] Kumar, Varun, and Anupama Chadha. Mining Association Rules in Student's Assessment Data. International Journal of Computer Science Issues 9.5: 211-216, 2012.
[9] Cristo´bal Romero, Sebastia´n Ventura, Enrique Garcı´a, 2007. Data mining in course management systems: Moodle casestudy and tutorial. Received 5 March 2007; received in revised form 19 May 2007; accepted 25 May 2007.
[10] http://en.wikipedia.org/wiki/Weka
[11] Anwar, M. A., and Naseer Ahmed. "Knowledge Mining in Supervised and Unsupervised Assessment Data of Students' Performance." 2011 2nd International Conference on Networking and Information Technology IPCSIT vol. Vol. 17. 2011.
[12] http://statistics.about.com/od/Formulas/ss/Correlation-Coefficient.htm

