Prediction of Students Academic Performance using Data Mining: Analysis

Sayana T S
St.Joseph's College,
Irinjalakkuda

Abstract--Data mining is a process of extracting knowledge from huge amount of data. It also refers to a way of finding important and useful information from a data base and it is used in different areas including the educational environment. Educational data mining is used to developing methods for discovering knowledge from data that come from educational environment. Educational Data Mining has an important role in students' academic performance prediction. In this paper, there is a evaluation between student's association rule algorithm, K-means clustering algorithm and Decision tree. This survey for assessing student's performance is based on different attributes. Attributes include class quizzes, mid and final exam marks, assignment, lab –works etc. We discuss the procedure based on Decision Tree of data mining method, Data Clustering that enable academicians to predict student's performance and instructor can take essential step to develop student academic performance.

In association rule mining algorithms, the rules generated are used to measure the connection between various attributes which will help to develop the student's academic performance. But result is better in the case of clustering methods. These are also included in this paper. I would like to propose a new strategy for the prediction of student's academic performance. That is a slight modification in K-means clustering method can produce better results in the prediction of students' performance.

I. INTRODUCTION

Data mining is data analysis method used to recognize unknown patterns in a large data set. It has been effectively used in diverse areas including the educational field. Educational data mining is an interesting research area which extracts useful unknown patterns from educational database for better understanding, improved educational performance and evaluation of the student learning process. Evaluating students' performance is a difficult problem. The main goal of this paper is, compare the prediction of student's academic performance on the basis of their performance in assignment; unit test graduation per data mining is a process of extracting knowledge from huge amount of data. It also refers to a way of finding important and useful information from a data base and it is used in different areas including the educational environment.

Educational data mining is used to developing methods for discovering knowledge from data that come from educational environment. Educational Data Mining

has an important role in students' academic performance prediction. In this paper, there is a evaluation between student's association rule mining algorithm, K-means clustering algorithm and Decision tree.

There are many technique used to predict student's performance. It includes K-means clustering, decision trees, association rule mining algorithm etc. But these all have different efficiencies.

Clustering algorithm and decision tree of data mining technique is helpful for future prediction of students' performance. Data clustering is a method of extracting valid, unknown, useful and hidden patterns from large data sets. It is most widely used technique for future prediction and the main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. This study makes of cluster use analysis to segment students into groups according to their character and use decision tree for making meaningful decision for the student's. Association Rule Mining is a predictable and well researched process for determining interesting relationships between attributes in large databases.

II. EXISTING SYSTEM

There are many data mining methods for the prediction of student's academic performance. It includes the following types:

- Decision trees
- Clustering
- Association rule

A. Decision trees

A decision tree is a predictive way used in clustering, classification and prediction tasks. It is appropriate for small data sets because it is suitable for discovering knowledge derived from the experimental data. Decision tree is faster to make and easier to know.

Decision tree technique uses tree arrangement to build classification models. In this technique dataset is divided into smaller subsets and at the same time an associated decision tree is incrementally developed. Tree's root and each internal node are labeled with a question. The arcs from each node represent each possible answer to the associated question. That result in a tree having decision and leaf nodes. A decision node is one which has two or more branches. Leaf node represents a prediction, decision or classification. The root node known as a best predictor is

Special Issue - 2015

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
RTPPTDM-2015 Conference Proceedings

the uppermost decision node in a tree. Decision trees hold both numerical data and categorical data. Decision tree is a well-known and effective technique which builds classification models in the form of a tree. A decision tree is developed through recursive methods which break down the set of training data into separate groups with the objective to maximize space among groups. The final result is a tree with leaf nodes and decision nodes where the leaf represents a prediction, decision or classification.

Classification and prediction also play a very interesting function in the field of education. Different decision tree algorithm can be applied to predict and identify the failure risk of student at education level. Predicting the academic result of a student desires lots of parameters to be considered. Predicted student's academic performance using the CGPA grade system where the data set comprised of the students gender, his parental education details, his financial background etc. In [8] the author has explored the various variables to predict the students who are at risk to fail in the exam. The solution strongly suggests that the earlier academic result strongly plays a main role in predicting their current result. In accordance with [9], the marks obtained by the students during the internal examination will play an essential role in predicting the outcome of the student in the main examination. The internal marks for the subjects MCA11, MCA12, MCA13, MCA14, MCA15 for a maximum of 100 marks and a result of Pass/Fail depending upon a minimum of 50 marks from each subject is fed as input and a decision tree is obtained using C4.5.The output should compared with the original marks received and result obtained by the student in the university examination.

Data mining can be applied on the raw data so that useful information can be obtained to take decisions. A decision tree depicts rules by dividing data into groups. C4.5 builds decision trees, using the concept of information entropy. The training data is a set $S = s1, s2,..$ of already classified samples. Every sample $si = x1,x2,...$ is a vector where $x1,x2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c1,c2,...$ where $c1,c2,...$ represent the class to which each sample belongs. The marks obtained by the students during the internal examination will play a vital role in predicting the outcome of the student in the main examination.

The internal marks for different subjects (conducted for 100 marks) and a result of Pass/Fail is fed as input to construct decision tree using C4.5 algorithm. The output is then compared with the original marks received i.e. result obtained by the student in the university examination (final exam). From internal marks out of 100,decision tree is drawn. Supposed values have been categorized as (0_44) where the students are categorized as Fail, in range of (45_54) the students are considered to be on border line hence categorized as pass or fail and (54_100) is classified as pass.

Another decision tree is drawn from external marks(out of 100).Marks between (0_39) indicates a fail in the result of the student and (40_100) indicates a pass in the result of the student. In external result pass/fail class is not present. The outputs of both decision trees constructed using Internal and external marks are compared. From [7] we can conclude the results that are:

1. The students who have been predicted as pass have been declared pass in the university exam also.
2. The students who were predicted to be pass/fail in the decision tree of internal marks were declared pass in the university exam.
3. The students who have been predicted as fail, some students have actually failed and some other students have improved their studies and passed in the examination.

The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the tutor to identify the weak students and improve their Performance. The analysis of the result declared from the university is a proof for the same. Since the application of data mining brings a lot of advantages in higher learning institution.

B. Clustering

In clustering, no attribute is selected as a target but the relationship between attribute can be discovered based on the formed clusters. The most popular method for prediction in clustering is K-means. K-means clustering algorithm is used to automatically cluster the students. This method is used to classify the students' performance according to learning .However; the cluster model has a drawback that is no clear rules to define each cluster [8]. This drawback can be solved by combining several methods which are statistical methods, decision trees and association rules. Statistical methods are used to define weight for each selected parameter that is useful in developing decision tree. In K-means clustering we cluster the students according to different attributes and for each cluster we calculate the overall performance. Clustering in higher education means it classifies the student by their academic performance. Data clustering methodology can help bridging this knowledge gaps in higher education system.

Procedure for clustering student Database

If the student database containing 180 records as considered for mining. The total number of attributes considered here is 65. The working procedure to calculate the mean and clustering of the dataset is given as follows:

Step 1: Read the given student Database.

Step 2: Preprocess the dataset by avoiding duplicate records and remove irrelevant data.

Step 3: Calculate the File size, total number of records and the attributes from the database.

Step 4: Split and cluster the total records into three different clusters namely 1, 2, 3.

Step 5: Form the clusters based using on the series order three clusters now holds 60 records each.

Step 6: The individual cluster group is again clustered, or partitioned into two ways.

Step 7: Find students those with high CGPA.

Step 8: Partition the data by CGPA based on their low and high mark values.

Step 9: Calculate mean of high CGPA and partition the data based on mean value.

Step 10: Determine the sum of individual subject marks based on their cluster groups.

Step 11: Clusters are generated for CGPA.

By this way we can determine the clusters are of same properties and by this way we can calculate the overall prediction of student's performance.

C. Association rule mining algorithm

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%, Unit Test% and University Result%. 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.

In education data mining, association rule learning is a predictable and well researched method for form interesting connections between attributes in huge databases. Association rules are making based on various attributes.

These rules relating attributes like students graduation percentage, Attendance, Assignment work, Unit test Performance and how these attributes influence the student's university result. Number of association rule can be found for different confidence values. The explanation of the 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.

For example,

1. Attendance=Good Assignment=Poor ==> Unit Test=Poor

2. Attendance=Good Assignment=Poor ==> University Result=Poor

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. 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.

II. LITERATURE SURVEY

Educational Data mining for Prediction of Student Performance Using Clustering Algorithms.

K-Means clustering algorithm is a resourceful way of predicting the pass percentage and fail percentage of the students appeared for a particular examination. The results show the students' performance and it is seems to be accurate. The comparison between Navie bayes algorithm and decision tree technique shows that the Navie bayes techniques produce accurate result than the other and it is measured using confusion matrix. The results are predicted within 0 seconds. But Navie bayes algorithm is more complex than decision tree. So it is better use decision tree other than Navie bayes .

Predicting Students Academic Performance Using Education Data Mining.

Association rule mining algorithm 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. 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. It is also a better strategy, but there is a chance of occurring of error when we connect with different attributes

An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree.

Data mining process in student's database using k-means clustering algorithm and decision tree technique to predict student's learning activities. In this paper, the information generated after the implementation of data mining and data clustering technique may be helpful for instructor as well as for students. This work may improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. It is done by categorizing student database into different clusters.K - means clustering is more efficient than the decision trees.

Application of k-Means Clustering algorithm for prediction of Students 'Academic Performance.

k-Means Clustering algorithm is a simple and qualitative methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. We demonstrated our technique using k-means clustering algorithm. We divide the student data base into different clusters and finally predict the result of student's performance.

Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm.

In this paper, by using the K means clustering algorithm we divide the students into clusters according to their performance. Performance is based on various attributes like percentage of marks, attendance, class test etc.

IV. PROPOSED SYSTEM

Prediction of students' academic performance is better in the case of K-means clustering. It is robust, very fast and easier to understand and gives best result when data set are distinct or well separated from each other. But it has some

Special Issue - 2015

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
RTPPTDM-2015 Conference Proceedings

limitations also. That is the number of cluster, must be determined earlier. It is difficult to determine the number of clusters when there is slight change in the data. That is for a college, or a university student data base is increasing day to day. So it is not an efficient way. Then we have to make a more proficient system to assign number of clusters automatically. That is, I would like to propose a dynamic system for the determination of number of clusters according to their various properties in student data base. This dynamic system can provide more accurate results in K means clustering. This makes the system more efficient and less time consuming. So we will have to find a system that work more effectively.

V. CONCLUSION

In this paper we analyze different techniques like decision trees-means clustering, association rule mining based on various attributes, for the prediction of students' academic performance. But K- means clustering is better than other types because, it is very easy to understand for the large amount of data. Decision trees and association rules are efficient in the case of small amount of data. But when there is more change in student data base, it is better to use the proposed system for K-means clustering. The proposed system for K-means clustering is time consuming and efficient in the prediction of student's performance.

REFERENCES

[1] An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree.- Md. Hedayetul Islam Shovon, Mahfuza Haque, International Journal of Advanced Computer Science and Applications.

[2] Application of k-Means Clustering algorithm for prediction of Students' Academic Performance.- Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C,- International Journal of Computer Science and Information Security,Vol. 7, _o. 1, 2010

[3] Predicting Students Academic Performance Using Education Data Mining-Suchita Borkar, K. Rajeswari, International Journal of Computer Science and Mobile Computing.

[4] Performance Of Students Evaluation In Education Sector Using Clustering K-Means Algorithms- S. Ganga, Dr. T. Meyyappan, International Journal Of Computer Science And Mobile Computing.

[5] Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm- Md. Hedayetul Islam Shovon, Mahfuza Haque

[6] Educational Data mining for Prediction of Student Performance Using Clustering Algorithms- M. Durairaj , C. Vijitha- International Journal of Computer Science and Information Technologies International Journal of Computer Science and Information Technologies

[7] Efficiency of decision trees in predicting student's academic performance-S. Anupama Kumar and Dr. Vijayalakshmi M.N

[8] .Zlatko J. Kovacic, John Steven Green, Predictive working tool for early identification of 'at risk' students , Newzealand

[9] S. Anupama Kumar, Dr. Vijayalakshmi M. N., "Prediction of the students recital using classification Technique", IFRSA's International journal of computing (IIJC) , Volume 1, Issue 3 July 2011, pp305-30