THOMSON REUTERS
ENDNOTE

[Kundariya., 3(11): November 2016]
DOI- 10.5281/zenodo.168107

ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A CASE STUDY FOR STUDENT PERFORMANCE ANALYSIS BASED ON EDUCATIONAL DATA MINING (EDM)

Daxa Kundariya[*1] and Prof. Vaseem Ghada[2]
[*1]MTech Student, B.H.Gardi College of Engineering & Technology, Rajkot - India
[2]Assistant Professor, Computer Engineering, B.H.Gardi College of Engineering & Technology, Rajkot - India

ABSTRACT

Educational Data Mining (EDM) is a study methodology and an application of data mining techniques related to student's data from academic database. Like other domain, educational domain also produce vast amount of studying data. To enhance the quality of education system student performance analysis plays an important role for decision support. This paper elaborates a study on various Educational data mining technique and how they could be used to educational system to analysis student performance. It helps the teacher to identify students who need special attention and allow the teacher to provide appropriate guidance. This paper showcases the importance of Classification based data mining algorithms in the field of education system and also presents some promising future success.

Keywords: *Educational Data Mining, Classification, Decision Tree, ID3, C4.5.*

I. INTRODUCTION

Data Mining is the process to discover meaningful patterns from huge amount of data. Data mining is also popularly known as *knowledge Discovery in Database*, refers to 'extracting' or 'mining' knowledge from Database. Data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

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 originating from educational environments. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on.

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students' performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and trim down failure rate, to better understand students' behavior and many other benefits.

In today's world classification is an important technique in data mining. For classification the data use decision tree. The decision tree is important tool in data mining to do. Compare with the other, decision tree is a faster and more accurate. As a very important and widely used technology in data mining, data classification is currently used in many fields. The purpose of data classification is to construct a classification model, which can be mapped to a particular subclass through the data list in the databank. The decision tree algorithm is a more general data classification function approximation algorithm based on machine learning. A decision tree is a tree structure which classifies an input sample into one of its possible classes. Decision trees are used to extract knowledge by making

[Kundariya., 3(11): November 2016] ISSN 2348 – 8034
DOI- 10.5281/zenodo.168107 Impact Factor- 4.022

decision rules from the large amount of available information. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. This paper reviews a case study of different algorithms to classify the student's data using decision tree.

II. DATA MINING TECHNIQUE

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

Association Rule

Mining association rules searches for interesting relationships among items in a given data set. It allows finding rules of the form if antecedent then (likely) consequent where antecedent and consequent are item sets. Item set is set of one or more items. In our data set an example of item is: Grade = Average. Because, we are looking for items that characterize the grade of students, consequent has one item which is Grade = z where z is one value of the student grade such as Excellent, Very good, Good, Average. As part of association method, FP-Growth algorithm is applied to the data set.

Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. In this paper, we used classification based algorithm for research methodology that will show in III section.

Decision Trees

A decision tree is a tree structure which classifies an input sample into one of its possible classes. Decision trees are used to extract knowledge by making decision rules from the large amount of available information. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. This paper reviews different algorithms to classify the data using decision tree.

Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

III. DECISION TREE CLASSIFICATION (ALGORITHM / METHODOGOGY)

Decision tree is considered to be one of the most popular data-mining techniques for knowledge discovery. It systematically analyzes the information contained in a large data source to extract valuable rules and relationships and usually it is used for the purpose of classification/prediction. There is various kind of decision tree algorithm to

[Kundariya., 3(11): November 2016] ISSN 2348 – 8034
DOI- 10.5281/zenodo.168107 Impact Factor- 4.022

use for classified the data but here we using ID3 and C4.5 algorithm to analyze academic database. The process of ID3 and C4.5 elaborates as per below section.

ID3

ID3 (Iterative Dichotomized) is the most important decision tree generation algorithm, which is proposed in 1986 by Quinlan. This algorithm recursively partitions the training dataset till the record sets belong to the class label using depth first **greedy technique**. In growth phase of the tree construction, this algorithm uses **information gain**, an **entropy based measure**, to select the best splitting attribute, and the attribute with the **highest information gain is selected as the splitting attribute.**

C4.5

C4.5 algorithm is an improved version of ID3, this algorithm uses Gain Ratio as a splitting criteria, instead of taking gain in ID3 algorithm for splitting criteria in tree growth phase. Hence C4.5 is an evolution of ID3. This algorithm handles both continuous and discrete attributes- In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute.

Steps-

1. Check for base case.
2. Check whether all instances belong to the same class attribute (IF yes, THEN create a leaf ELSE begin to choose the splitting node).
3. For each attribute a -
   - ✓ IF attribute a is a nominal variable, calculate its entropy ratio.
   - ✓ IF attribute a is a continuous variable, calculate the distinct values' information gain first and calculate the entropy ratio of the one with the biggest gain.
   - ✓ Compare all these ratios and select the variable with the largest ratio as splitting node.
4. Create a decision node that splits on a_largest.
5. Recursively split the subsists obtained by splitting on a_largest.

IV. RESULT & ANALYSIS

A. Study on whether Attendance is enough or not (for the Performance of Students)

Explain the various ways in which performance of students can be studied, one of the ways being Correlation. Pearson Product Moment Correlation (PPMC) is used to test for a relationship between two numerical variables. It represents the linear relationship between the two sets of data. Pearson Correlation are represented by the Greek letter rho (p) for a given population and the letter "r" for a sample under consideration. The following formula is used to calculate PPMC:

Formulae:

Right at the inception, a scatter-plot should be made and studied which would unveil the possibility of a relationship between the two variables. The value of the Pearson correlation coefficient, R, ranges from values + 1 to -1. A value Closer to + 1 would indicate a stronger relationship. We can determine the category of correlation by seeing what effect one variable's increment in value has on the other. The expected categories would be:

- Positive correlation: the value of the variable naturally increases (R>O)
- Negative correlation: the value of the variable naturally decreases (R<O)
- No correlation: the value of the variable stays constant (R=O)

[Kundariya., 3(11): November 2016] ISSN 2348 – 8034
DOI- 10.5281/zenodo.168107 Impact Factor- 4.022

B. Predict Student Performance using Algorithm

For higher education institutions whose goal is to contribute to the improvement of quality of higher education, Human capital creation is assessed continuously. This makes prediction of students' success crucial for higher education institutions, to verify that the quality of teaching process is sufficiently meeting students' needs When the task of converting raw educational data into knowledge is carried out, the gratification of all stakeholders is ensured: students, professors or teachers, supporting management and the community and society of which we are a part of them. Prediction focuses on the estimation of the value of the variable describing the student, which is not known. This estimated value can be a numerical/categorical value. In our research, Classification using decision trees and Prediction using Multiple Regression has been used. Using Classification Algorithms, we aim to predict the Final Result (PASS or FAIL) of students in our university using decision trees and Predict Semester End Grade Point using Multiple Regression to help all stakeholders as mentioned above like Students, Professors, Administration, Supporting administration.
To making IF/THEN rule for analyze student's performance, parameters will be taken in algorithm as per table-1

Table 1. Parameters for ID3/C4.5 Algorithm

No	Name
1.	Previous semester Marks (PSM)
2.	Attendance (ATT)
3.	Technical Activity Score (TAS)
4.	Non-Technical Score (NTS)
5.	Behavior (BH)
6.	Communication Skill (CS)

To determine the best attribute for a particular node in the tree we use the measure called Information Gain. The information gain, Gain (S, A) of an attribute A, relative to a collection of examples S,

$$Gain(S, PSM) = Entropy(S) - \frac{|S_{First}|}{|S|} Entropy(S_{First}) - \frac{|S_{Second}|}{|S|} Entropy(S_{Second}) - \frac{|S_{Third}|}{|S|} Entropy(S_{Third}) - \frac{|S_{Fail}|}{|S|} Entropy(S_{Fail})$$

Previous Semester Marks/Grade obtained in any course. It is split into five class values: *First – >60%, Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%.*
PSM has the highest gain, therefore it is used as the root node as shown in figure –

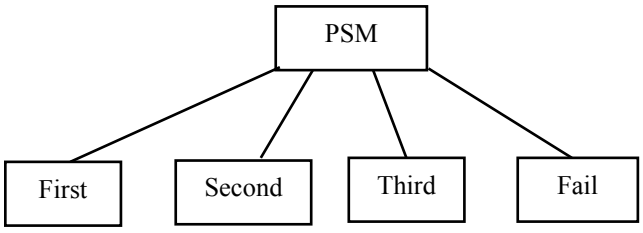

Figure 1. PSM as root node

This process goes on until all data classified perfectly or run out of attributes. The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules. FG is final grade and that will be output of analysis to student's performance.

ISSN 2348 – 8034
Impact Factor- 4.022

1. F PSM = "First" AND ATT = "Good" AND TAS = "Good" OR "Average" THEN FG = First
2. F PSM = "First" AND CTG = "Good" AND ATT = "Good" OR "Average" THEN FG = "First"
3. F PSM = "Second" AND ATT = "Good" AND TAS= "Yes" THEN FG = "First"
4. F PSM = "Second" AND TAS = "Average" AND NTS= "Yes" THEN FG = "Second"

Figure 2. Rule Set generated by Decision Tree

V. CONCLUSION

Data Mining techniques like Regression and Decision Trees to predict academic performance are studied and executed and are found to effectively predict student performance and also, to predict academic failure. Clustering is successfully used to group the students into clusters according to their academic strengths and weaknesses. These methods will greatly help the university teachers to know what changes need to be made, provide remedial courses to weak students, identify weak students at risk of failure or year drops and to make learning a better experience for their students; and it would also help the students and the placement committee in getting to know which job profiles they could apply to on the basis of their skill set.
This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

VI. ACKNOWLEDGEMENTS

For this project, we would like to thank my all faculties for their continuous support and faith in us.

REFERENCES

1. *Brijesh Kumar Baradwaj , Saurabh Pal ,"Mining Educational Data to Analyze Students' Performance " (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.*
2. *Shahrukh Teli , Prashasti Kanikar ,"A Survey on Decision Tree Based Approaches in Data Mining " International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 4, 2015 .*
3. *Mohammed M. Abu Tair, Alaa M. El-Halees ,"Mining Educational Data to Improve Students' Performance: A Case Study ",International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.*
4. *Parneet Kaura,Manpreet Singhb,Gurpreet Singh Josanc "Classification and prediction based data mining algorithms to predict slow learners in education sector" ELSEVEIR-2015.*
5. *John Jacob,Kavya JhaPaarth Kotak,Shubha Puthran "Educational Data Mining Techniques and their Applications" IEEE-2015.*
6. *Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby2, "Data Mining: A prediction for Student's Performance Using Classification Method ", World Journal of Computer Application and Technology 2(2): 43-47, 2014*