Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Student Performance Analysis, Visualization and Prediction Using Data Mining Techniques

Akshay Deshpande[1], Prashant Pimpare[2], Shashank Bhujbal[3], Abhishek Kommwar[4] & Prof.Jagruti Wagh[5]
Computer Engineering Department, Savitribai Phule Pune University

Abstract: *There are various types of evaluation methods in education systems to know how well students studied the subjects & their command over them. At present students' performance is being evaluated on percentage or credits system. But these systems are not sufficient for overall evaluation of students. To overcome these issues we propose a system in which different data mining techniques are applied on students' performances in tests, extra-curricular activities, sports to grade students for campus placement activity. ID3 decision tree algorithm is used to predict performance in end semester exam. Previous year performances, current year performances in class tests are used for prediction. This evaluation helps to know strong & weak areas of students. More guidance can be given on weak areas by teachers.*

Keywords: *Educational Data Mining (EDM), ID3 Algorithm, Clustering, Knowledge Discovery in Database (KDD).*

1. Introduction:

The involvement of information technology in various fields has led the large volumes of data storage in various formats like records, files, documents, images, scientific data and many new data formats. The data collected from different applications require a proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD) is used to discovery of useful information from large collections of data. The data mining applies various methods and algorithms in order to discover and extract patterns of stored data. Data mining techniques have been introduced into new fields of Statistics, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

Educational Data Mining (EDM) uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. These techniques are used to discover knowledge such as classifications, association rules and clustering. The discovered knowledge can be used for prediction regarding enrollment of students in a particular course, prediction about students' performance and so on.

In this paper data mining methodologies are used to study students' performance in the various courses.

The classification task is used to evaluate student's performance. There are many approaches that are used for data classification, the decision tree method is used here. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester.

2. Related Work:

Researches on the use of data mining techniques in an education system are going on all over the world. These researches are mainly focused on two areas - performance analysis & performance prediction. Different techniques are being applied to get more accurate results.

2.1 Performance Analysis:

It is very important to analyze performances of students in exams. As it helps to the teacher to determine weak areas of students, so they can focus on that areas. Researchers applied various data mining techniques for analysis.

OTOBO Firstman Noah, BAAH Barida, TAYLOR Onate Egerton of Nigeria used apriori algorithm & k-means algorithm. They considered 3 attributes – type of admission, gender, degree obtained for analysis. The outcome of this analysis used to decide how to distribute seats for a course [2].

Mohammed M. Abu Tair, Alaa M. El-Halees of University of Gaza, Palestine used basic data mining techniques such as association rule, classification, clustering, and outlier detection for analysis. They used Naïve Bayes classifier & k-means clustering algorithm. Variable used for it were gender, city, matriculation GPA, secondary school type, grade. They used 15 years' student data for mining. The output of this analysis helped them to conclude that previously well-performed student continued to perform well in the graduate course [3].

2.2 Performance Prediction:

Prediction of marks is very useful for students, as it will help them to preparation for the exam. Also, teachers can analyze how much students are preparing for exams. A lot of researches are going on to increase an accuracy of prediction.

Dorina Kabakchieva of Sofia University, USA used 10,330 students' data to generate prediction models. 20 parameters including personal information, previous education type and previous education performance are considered for analysis. She used different classifiers for predictions. Classifiers used were Naïve- Bayes, J48, Rule Learner, KNN. Prediction is given as bad, average, good, very good, excellent. But all these classifiers lacked in the accuracy of prediction. All these classifiers found accuracy in between 52%- 67% [7].

Edin Osmanbegovic & Mirza Suljic of Tuzla University, USA used Naïve Bayes classifier & Multilayer perceptron neural network to generate a prediction. Gender, family information, distance from school, high school GPA entrance exam marks, scholarship score, time of the study, material used for the study, earning were attributes considered by them. They give binary prediction only i.e. student will pass or fail in final exam. They got up to 70% accuracy in prediction [8].

Paulo Cortez and Alice Silva of Portugal used prediction model for predicting performance in Portuguese & Mathematics. They only considered past school grades, demographic information and school related data for this. They used decision trees, random forest algorithm, neural networks and support vector machine for generating a prediction. They got 78% accuracy in prediction [4].

Emmanuel N. Ogor of Turks & Caicos Islands combined artificial neural network, k-means & c5.0. They also increased no. variable affecting students' performance. They used total 78 variables. Fail, marginal, satisfactory, good, excellent were the outcome of predictions. They got good accuracy up to 85% [6].

Abeer Badr El-Din Ahmed, Ibrahim Sayed Elaraby of Egypt and Brijesh Kumar Bhardwaj & Saurabh Pal of Rajasthan used the ID3 algorithm for prediction. They used midterm marks, class test, seminar, assignment marks, attendance, lab work grades for prediction. They got accuracy up to 90% accuracy [1] [5].

By referring above papers it is necessary to use more variables & large data set to more accurate prediction.

3. Data Mining Definition and Techniques:

Data mining is process of discovering interesting patterns and knowledge from large amounts of data. Data mining, also popularly known as Knowledge Discovery in Database. There are following techniques of Data mining:

3.1 Classification:

Classification is the form of data analysis that extract models describing important data classes. This approach frequently uses decision tree classification algorithms. The data classification process includes learning and classification. In learning method, the training data sets are analyzed by the classification algorithm. In classification test data sets are used to find the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. In our paper, we used ID3 decision tree to represent logical rules of student final grade.

3.2 Clustering

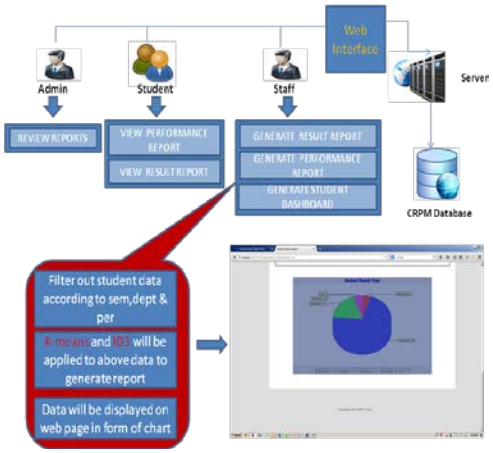
Clustering is a process of grouping a collection of objects into classes of similar object. Using clustering methods, dense and sparse regions in object space can be identified. In educational data mining, clustering has been used to group students according to their performance. According to clustering, clusters distinguish performance of student according to their behavior. In this paper, students are clustered into groups according to their academics, exams and soon.

3.3. Decision Trees

A decision tree is a tree shaped structure, where each internal node denotes a test on the attribute, each

branch represents an outcome of test and each leaf node holds a class label. Decision tree represent sets of decisions. Using these decisions rules are generated for the classification of a dataset. Decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

4. System Architecture:



4.1 Administrator: Administrator is required to assign teachers to classes, add students to different classes according to year & branch. Admin is responsible for maintaining basic information of students & users.

4.2 Teacher: Teacher can add marks, attendance of students. Also teacher can generate report of students' performance.

4.3 Student: Student can view his performance & prediction of term work, end semester exam marks.

5. Algorithm Used:

5.1 ID3 Algorithm:

Terminologies used in ID3 Algorithm:

- Establish Classification Attribute
- Compute Classification Entropy.
- Calculate Information Gain using classification attribute.
- Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
- Remove Node Attribute, creating reduced table.
- Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows

5.2 K-Means Algorithm:

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select '*c*' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

6. Data Mining Process:

In educational system, a student's performance is determined by the term work, attendance and end semester examination. The term work is carried out by the teacher based upon student's performance in educational activities such as class test, assignments, attendance. The end semester examination is one that is scored by the student in semester examination. Student has to get minimum marks to pass a semester in internal as well as end semester examination.

6.1. Data Preparations:

The data set used in this study was obtained from MMCOE, Pune University on the sampling method of computer Applications department of course BE (Bachelor of Engineering) from session 2015 to 2016.This data is stored in different tables.

6.2. Data selection and transformation:

In this process, we were selected those field which were required for data mining. We were select derived variables. With the help of data mining we can extract variables information from the database. In following tables we are showing all the predictor and response variables which were derived from the database.

TABLE I. STUDENT RELATED VARIABLES

Variable	Description	Possible Values
CTG	Class Test Grade	{ Poor , Average, Good}
ASS	Assignment	{Poor , Average, Good
ATT	Attendance	{Poor , Average, Good}
ESM	End Semester Marks	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36%}

- **CTG:**
  In this Class test grade can be obtained. Here in each semester has conducted class tests and average of two class test are used to calculated marks. Class test grade is divided into three classes: Poor – 1, Average – > 2 and < 3, Good –>5.
- **ASS :**
  In this Assignment performance can be calculated. In semester assignments are given to students by each teacher. Assignment performance is divided into three classes: Poor – 1, Average – > 2 and < 3, Good –>5.
- **ATT:**
  In this Attendance of Student can be calculated. Minimum 75% attendance is compulsory to participate in End Semester Examination. Attendance is divided into three classes: Poor - <50%, Average - > 60% and <75%, Good - >75%

7. Conclusion:

From all these analysis we can conclude that, it is possible to predict student performance with more accuracy by combining k-means & ID3 algorithm. For more accurate prediction we need to take more number of variables which affects students' performance. Also the analysis reports will be very useful for teachers as well as students.

8. References:

1. Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology 2(2): 43-47, 2014 DOI: 10.13189/wjcat.2014.020203 Egypt

2. OTOBO Firstman Noah, BAAH Barida, TAYLOR Onate Egerton ,"Evaluation of Student Performance Using Data Mining Over a Given Data Space", (IJRTE) ISSN: 2277-3878, Volume-2, Issue-4, September 2013 Nigeria.

3. Mohammed M. Abu Tair, Alaa M. El-Halees ,"Mining Educational Data to Improve Students' Performance: A Case Study", (IJICTR) ISSN 2223-4985, Volume 2 No. 2, February 2012, Palestine.

4. Paulo Cortez and Alice Silva ,"Using Data Mining to Predict Secondary School Student Performance" Portugal.

5. Brijesh Kumar Bhardwaj & Saurabh Pal,"Mining Educational Data to Analyze Students Performance", (IJACSA), Vol. 2, No. 6, 2011, Rajasthan.

6. Emmanuel N. Ogor,"Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques Fourth Congress of Electronics, Robotics and Automotive Mechanics", DOI 10.1109/CERMA.2007.78 Turks & Caicos Islands, Caribbean.

7. Dorina Kabakchieva ,"Predicting Student Performance by Using Data Mining Methods for Classification", (CAIT) Volume 13 no. 1 Print ISSN: 1311-9702; Online ISSN: 1314-4081 DOI: 10.2478/cait-2013-0006 USA.

8. Edin Osmanbegovic & Mirza Suljic ,"Data Mining Approach For Predicting Student Performance Economic Review Journal of Economics and Business", Vol. X, Issue 1, May 2012 USA.