STUDENT PERFORMANCE PREDICTION SYSTEM USING DATA MINING CLASSIFICATION AND CLUSTERING

Trupti Bagad, Dhanashree Gaikwad, Sayali Kedari, Sandhya Walunj

Under the guidance of Prof. Rahul Patil

Department of Computer Engineering,
PIMPRI CHICHWAD COLLEGE OF ENGINEERING, PUNE

ABSTRACT

The System aims towards prediction of academic performance of students for academic year on the basis of different parameters related to previous year on the basis of different parameters related to previous year using data mining classification and clustering. An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. This helps them to identify promising students and also provides them an opportunity to pay attention to and improve those who would probably get lower grades. As a solution, we will develop a system which can predict the performance of students from their previous performances using concepts of data mining techniques under classification and clustering. By applying the K-Means and its improved K-Means, CART (Classification and Regression Tree) and its improved R2-CART Algorithm on this data, we have predicted the general and individual performance of freshly admitted students in future examinations. The student academic performance is usually stored in student management system, in different formats such as files, document, records, images and other formats. These available students' data could be extracted to produce useful information. However, the increasing amount of students' data becomes hard to be analyzing by using traditional statistic techniques and database management tools. Thus, a tool is necessary for universities to extract the useful information. This useful information could be used to predict the students' performance.

Keywords: *Classification, Clustering, K-Means, improved K- means, CART, R2-CART*

---

1. INTRODUCTION

The ability to monitor the progress of students' academic performance is a critical issue to the academic community of higher learning. Analyzing the past performance of admitted students would provide a better perspective of the probable academic performance of students in the future. This can very well be achieved using the concepts of data mining. We will propose system with better accuracy for predicting student performance using data mining classification algorithms. System will also help the weaker student to improve and bring out betterment in the result. Knowledge discovery in academic institution becomes more critical and crucial in terms of identifying the students' performance. In the extraction of actionable knowledge from a large database the data mining plays a vital role. The actionable knowledge extraction provides an interestingness and meaning to the mined data. This paper focuses on the prediction of the students' academic performance from the large student database. The mining algorithm like clustering and classification algorithm is revisited to predict the performance after initial mining of raw data. The main scope of this paper is to reveal the outcome of the performance analysis of a student .This work will help the university to reach betterment in providing the quality input to the student community and impart the knowledge effectively.

2. PROPOSED SYSTEM

There are a few features from the existing systems that are employed during the design and implementation phase of the proposed system. These features and functionalities include the user interface, students' performance prediction, illustration displays and report generation. Good user interface provides a user-friendly interface as

It is easy to be navigating and not complicated. Meanwhile, the students' performance prediction is included into the proposed system to make sure the objectives are achieved. Furthermore, the generation of reports in Portable Document Format (PDF) and illustration display such as charts in PDF makes student performance analysis easier.

From these features found in proposed system, all the user requirements would be fulfilled.

Our Proposed system will provide following features:

i. Able to help lecturers to automatically predict students' performance in course "3rd year engineering"
ii. Able to keep track and retrieve students' performance in a particular course and semester
iii. Able to view the factors that affect the students' Prediction result
iv. Able to generate students' reports

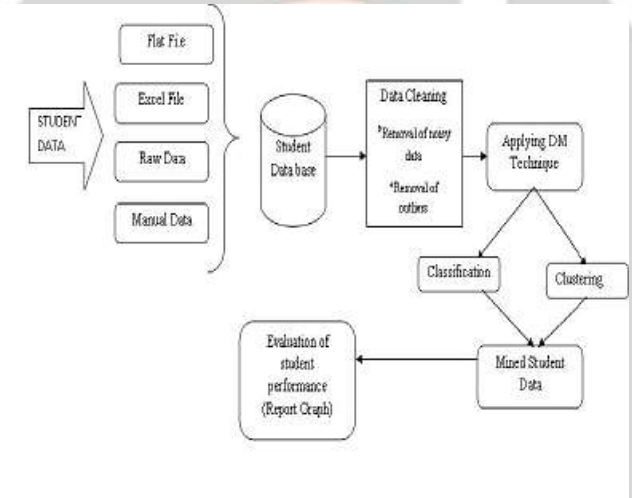

Fig-1: System Architecture

Given a student information file which is uploaded by teacher, our system will predict the performance of student mentioned in the file using data mining classification algorithms result will be displayed.

2.1 PROPOSED SYSTEM FEATURES

1. Data set is given to the system through the uploaded file.
2. System will apply classifications algorithm.
3. Student will be classified into pass or fail class.
4. Result will be displayed.

2.2 IMPLEMENTATION OF PROPOSED SYSTEM

We had divided the entire implementation into five stages:-

1. In the first stage, information about students who have been admitted to the third year was collected. This included the details submitted tithe college at the time of enrolment.

2. In the second stage, extra information was removed from the collected data and the relevant information was fed into a database.

3. The third stage involved applying the K-Means, Improved K-Means and CART and Improved CART (R2-CART) algorithms on the training data to obtain decision trees of both the algorithms.

4. In the next stage, the test data, i.e. information about students currently enrolled in the third year, was applied to the decision trees.

5. The final stage consisted of displaying of results.

The user can interact with the system by using different tags provided in the GUI these are

- Upload file to the system.
- Take result from the system.
- Analyze the results.

2.3 DATABASE

We will take a training dataset consisting of information about students in third year Computer Department. This data was in the form of a spreadsheet and had details of each student such as full name, enrollment ID, gender, percentage of marks obtained in second year percentage of marks obtained in the entrance examination, admission type, etc. For ease of performing data mining operations, the data was filled into a database. Database also consists of list of the teachers associated with department.

3. IMPLEMENTATION

The implementation work is based on the collected data which possess various data mining aspects. The Student data is taken into account for the performance prediction. The proposed research work is categorized into various modules. This research work is carried out with the inclusion of data mining technique and implementation software.

The proposed work speculated as the useful application where, student's performance can be viewed and placement criteria for various concerns are listed efficiently. The student performance is generated as a report graph which can be used as a survey to improve the student performance in the future. The student's and other user can login with their given credentials onto the desired application which modeled to predict the student performance and their placement criteria. The students, faculty and other user can view the past and recent performance of the respective courses to reexamine their current performance.

3.1. K-MEANS

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or
- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters

- Each record is assigned to the nearest cluster using admeasure of distance (e.g. Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean

of the clusters in the dataset.

3.2. IMPROVED K-MEANS

The original k-means algorithm is modified to improve the accuracy and reduce execution time.

Step 1: Input: In this step take input from the user the dataset and pass it to the algorithm.
Step 2: Apply the normalization technique to the given dataset
Step 3: Apply the sorting technique to the given dataset
Step 4: Apply the algorithm to find initial centred from the dataset
Step 5: Assign data object to the centroids (repeat until convergence occur)
Step 6: Recalculate centroid
Step 7: Check for the convergence

In the modified algorithm first initial cluster size is calculated by using total attributes divide by number of cluster. Then normalization technique is used to normalize the dataset to scale up the values is the range. In the next step the sorting technique is used on the dataset because processing the sorted array is faster than unsorted array. Then calculate the initial centroid by mean of the cluster. In the next step assign the data object to the initial centroid by calculating Euclidean distance. Check for the convergence criteria. Repeat the steps until no more changes in the last centroids and updated centroid. Because of the initial centroid is generated by calculation the number of iterations is fixed, the initial centroids are determined systematically so as to produce clusters with better accuracy.

3.3. CART

CART stands for Classification and Regression Trees. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. CART produces binary splits. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

Example: A small training set is shown in Table 1. There are 3 attributes which are a1; a2; and a3. a3 is a continuous attribute. The target set t has 2 classes which are P,N (i.e. Positive & Negative).

Table 1: Training set of Example

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	P
2	T	T	6.0	P
3	T	F	5.0	N
4	F	F	4.0	P
5	F	T	7.0	N
6	F	T	3.0	N
7	F	F	8.0	N
8	T	F	7.0	P
9	F	T	5.0	N

Solution: The gini index of a_1, a_2 is:

$$Gini(t) = 1 - [(\frac{4}{9})^2 + (\frac{5}{9})^2] = 40/81$$

$$Gini(a_1 = T) = 1 - [(\frac{3}{4})^2 + (\frac{1}{4})^2] = 3/8$$

$$Gini(a_1 = F) = 1 - [(\frac{1}{5})^2 + (\frac{4}{5})^2] = 8/25$$

$$Gini(a_2 = T) = 1 - [(\frac{2}{5})^2 + (\frac{3}{5})^2] = 12/25$$

$$Gini(a_2 = F) = 1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2] = 1/2$$

$$GiniGain(a_1) = Gini(t) - [\frac{4}{9} \cdot Gini(a_1 = T) + \frac{5}{9} \cdot Gini(a_1 = F)] = 0.149$$

$$GiniGain(a_2) = Gini(t) - [\frac{5}{9} \cdot Gini(a_2 = T) + \frac{4}{9} \cdot Gini(a_2 = F)] = 0.005$$

According to the gini gain, the best split between a1 and a2 is a1 due to it has the higher gini gain. For a3 which is a continuous attribute, firstly sort all values in increasing order (i.e. 1.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0) and then partition each position (i.e. 0.5, 2.0, 3.5,4.5, 5.5, 6.5, 7.5, 8.5) into 2 divisions which becomes a binary tree. The following calculation shows the gini gain on No.3 split position, that is, 3.5.

$$GiniGain(SplitNo.3) = Gini(t) - [\frac{2}{9} \cdot Gini(SplitNo.3left) + \frac{7}{9} \cdot Gini(SplitNo.3right)]$$

$$= \frac{40}{81} - \{\frac{2}{9} \cdot [1 - (\frac{1}{2})^2 - (\frac{1}{2})^2] + \frac{7}{9} \cdot [1 - (\frac{3}{7})^2 - (\frac{4}{7})^2]\} = 0.002$$

The results of every possible split are shown in Table 2.

Table 2: Results of continuous attribute a_3

Split No.	1	2	3	4	5	6	7	8
Split Positions	0.5	2.0	3.5	4.5	5.5	6.5	7.5	8.5
GiniGain	0	0.077	0.002	0.049	0.005	0.012	0.049	0

According to the results of a1; a2; a3, the best split is a1 due to its highest value offing gain.

3.4. IMPROVED CART(R2-CART)

Although the CART algorithm could process analysis of the sample data and build the classification tree model, the tree often has redundant tree leaf nodes. So that the description for the classification is not clear and simple enough, sometimes the classification results are also difficult to understand. In order to solve the problem of existing redundant tree leaf nodes in an classification tree of CART algorithm, this paper proposes an new improved classification and regression tree algorithm named R2-CART. At first, the algorithm uses Attribute Reduction Algorithm based on rough set to reduce the attributes in the sample set; secondly, run the CART algorithm on the reduced attribute set of sample data and get the classification tree; then change paths of the classification tree root to the end of each leaf node into the classification rules; finally, use the Rule Reduction Algorithm to reduce the rule of classification tree and express the reduction classification rules in the form of classification tree.

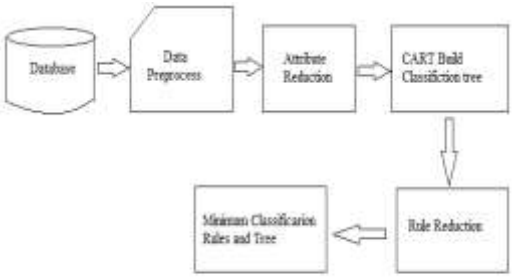

Fig. The basic process of the improved algorithm R2-CART

4. CONCLUSION

In this study, student's performance is predicted and graph is generated to improve efficiency and make the student apt for the better placement. This study will also work to identify those students who in need of special attention to increase their performance. Predicting students' academic performance is great concern to the higher education. With the help of classification and clustering technique the performance of student is identified to a maximum extent, and the result obtain through this research work reveals the positive outcome of student involvement in improving university quality. Classification technique is used to classify the student according to their academic results. They classified as average performer, intermediate performer and better performer. This experimental study can be further expanded which meets lot more academic constraints which creates effective impact in the overall outcome of the student and institution

5. REFERENCES

[1] "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", Oyelade, O. J, Oladipupo, O. O, IJCSIS) International Journal of Computer Science and Information Security 2010.

[2] "A Novel Approach for Data Clustering using Improved K- means Algorithm", Rishikesh Suryawanshi, Shubha Puthran, International Journal of Computer Applications (0975 – 8887) Volume 142 – No.12, May 2016.

[3] "An Improved Algorithm for CART based on the Rough Set Theory", Weiguang Wang, Cong Wang, Wanlin Gao – IEEE Fourth Global Congress on Intelligent Systems 2013.

[4]"Data Mining: A prediction for performance improvement using classification", Bhardwaj, B. K. & Pal, S., International Journal of ComputerScience and Information Security, 9(4), 136-140(2011).

[5] Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan. "An efficient enhanced k-means clustering algorithm." Journal of Zhejiang University SCIENCE A 7, no. 10 (2006): 1626-1633.

[6]Steinberg, Dan, and Phillip Colla, CART: classification and regression trees, The Top Ten Algorithms in Data Mining (2009):179-201.

[7] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique, "European Journal of Scientific Research, vol. 34(4), 2009.