PREDICTION SYSTEM FOR STUDENT PERFORMANCE USING DATA MINING CLASSIFICATION

Rahul Patil
Dept. Of Computer Engg.
PCCOE, Nigdi, Pune.
rahulpatilpink@gmail.com

Sagar Salunke
Dept. Of Computer Engg.
PCCOE, Nigdi, Pune.
sagargsalunke@gmail.com

Madhura Kalbhor
Dept. Of Computer Engg.
PCCOE, Nigdi, Pune.
kalbhormadhura11@gmail.com

Rajesh Lomte
Dept. Of Computer Engg.
PCCOE, Nigdi, Pune.
rajulomte1@gmail.com

Abstract— In the education system, highest level of quality can be achieved by exploring the knowledge regarding prediction about student's performance. In an analysis of data, data mining techniques play an important role. In order to predict performance of students in future academics, it's good if an educational institution have an approximate prior knowledge of all enrolled students in their institute. This prior knowledge becomes an important tool for educational institute to improve students those who would likely to get less marks and also identify bright student. . As a solution we are trying to develop a system which will help an educational institute to prefigure the performance of students from their former functioning. In order to achieve this we will use concepts of data mining techniques under Classification. Also for solution to get developed we are trying to prepare the data set containing information about students in terms of their gender, marks and rank in entrance examinations and results in Third year of the former batch of students. These data sets have been analyzed to prepare final solution. The mapping of data into predefined groups or classes is done in Classification data mining technique. It is a supervised learning method in which to generate rules for classifying test data into predetermined groups or classes labeled training data is required. The process is divided into two-phases. The first phase which is learning phase, the training data is examined and classification rules are begot in this phase. The second phase the Classification one in which test data is classified into classes in accordance with training data set. General and individual performance of third year students in future examinations is prefigured by using the ID3 (Iterative Dichotomize 3), C4.5, Improved weighted modified ID3 classification algorithms on this data.

Keywords— Data Mining, Decision Tree, Classification, ID3, C4.5, Predicting Performance.

I. INTRODUCTION

Students under diverse classes from various locations, educational backcloth and with varying scores in entrance examinations like CET, JEE take admissions in educational institutes every year. Moreover engineering colleges may be connected to various universities. Each university has vast variety of subjects in their curriculum. In order to provide a better view of most probable functioning of students in future, past performance of admitted students are required to be analyzed. By using data mining classification algorithms this can very well be achieved by our system. It has been observed that for academic community of higher learning, it is critical issue to monitor the progress of student's academic performance. We will propose system with better accuracy for predicting student's performance using data mining classification algorithms. In order to improve and bring out betterment in result for weaker students this system will be beneficial. For various universities students are main asset. In producing graduates of high quality, students play significant role with the help of their overall academic performance achievement. Through examination, judgment and other form of measurements the overall academic functioning of the students can be measured and tested.

The users of our system are teachers of particular department of an educational institute. The user is assuming to have basic knowledge of computers and basic technical knowledge about web application. DBA can add and remove number of teachers associated with the system the proper user interface and user manual provided to teachers to provide help regarding system operations and working of system.

Different format such as file, papers, records, images and relevant formats can be used to store student academic performance in the system. In order to produce useful information this data will be extracted. Student performance will be predicted using this useful information.

II. PROPOSED SYSTEM

Few features like forecasting of student performance, exemplification displays, report generation and graphical user interface from the previously implemented systems are during the design and effectuation stage of the proposed system. Meanwhile, in order to achieve the objectives forecasting of students' functioning is included into the purport system. Moreover, exemplification displays such as charts in PDF and the reports are generated in PDF in order to make analysis of student performance easier.

Every user requirement would be fulfilled with the help of all these features set up in proposed system,.

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

Our Proposed system will provide following features:

i. Provide ability for teachers to automatically forecast students' performance in course "3rd year engineering".
ii. Students' functioning in a particular course and semester can be tracked and recovered.
iii. Provides an ability to survey the various components that impact the forecasting of students' outcome.
iv. Provides an ability to generate reports on student's performance.

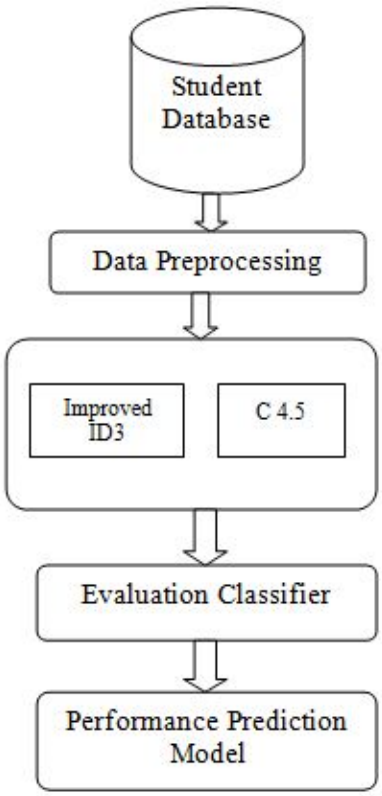

Fig 1: System Architecture.

Given 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 and results will be displayed.

III. IMPLEMENTATION OF PROPOSED SYSTEM

The implementation of the proposed system is divided in five stages. In stage one, information of the students currently in third year is collected which contains the details of students submitted during the enrolment time. Pre -processing of information is done to remove the irrelevant data and database is created with the relevant information in stage two. In stage three, Machine learning algorithms C4.5, ID3, and improved ID3 is applied on training database and decision trees are obtained. On the created decision tree the created third year student's database is applied in stage four. Stage five displays the result. For the ease of the user, interactive GUI is implemented for this various stages.

A. Database

Experimentation will be done using training database. Training database consist the information of Computer department third year students. Database is formed consisting the information like application ID, name, second year percentage, entrance exam marks, gender, admission type, caste etc. Data is filled into a database to perform data mining operations. Database also consists of list of the teachers associated with department.

B. Data Preprocessing

The training dataset is segmented further, once we had details of all the students, for this various feasible splitting attributes, i.e. the attributes which would have affected the performance of a student are considered.

Attributes or information collected which is unwanted for further processing will be removed from the database.

C. Functions

Different functions are performed by the proposed system. The proposed functions are as follow.

i. FileUploading()

This function will take data of student to be processed by our system and data will be taken in two formats bulk evaluation and single student evaluation. For bulk evaluation the data can be batch wise, division wise or department wise.

ii. DataProcessing

Extra information will be distant from collected data. Relevant information will be added into database and final format of database will be decided in this function.

IV. ALGORITHMS

A. ID3 algorithm

The ID3 algorithm is based on recursive process. Using the concept of information gain at each step there is an rating of a subset and initiation of decision node process, until the subset in rating if specified by same compounding of properties and its values. Decision tree is created using ID3 algorithm from given set and. Metric used to generate decision tree is information gain and top-down greedy search is used to check each attribute at every tree node. In each step best attribute is selected using information gain. To create a tree from top down ID3 uses entropy and information gain concepts.(1)

i .Entropy

Entropy is calculated as given in equation(1).

$$H(p1, p2, \ldots, ps) = \_ - (pi \log pi) \quad (1)$$

for given probabilities p1, p2, …, ps, where _pi = 1, In a given database entropy searches the amount of order. A perfectly classified set is identified by the value of H = 0 .The greater the entropy, the greater the potential to improve the classification process.

ii. Information Gain

An information gain is nothing but the deviation between how much information is needed after the split. Depending upon the highest gain in information ID3 chooses the splitting attributes. The calculations are done by deciding the variations between the selective information of the master dataset and the leaden sum of the selective information from each of the subdivided datasets. The formula used is:

$$G(D, S) = H(D) - \_P(Di)H(Di) \quad (2)$$

c. Shortcomings of ID3 algorithm

The Id3 algorithm selects the attribute having more number of values which is not necessarily the best attribute. This is the drawback of it. For the small test sample data may be over-classified or over-fitted. To make a decision only one attribute is tested at a time. Identifying class of continuous data is computationally expensive as many trees must be generated to see where to break the continuum.

B. Improved ID3

Inclination towards the attributes with more values is one of the primary drawbacks of the ID3 algorithm. This leads to faulty selection and hence, as a result, the tree generated is may not be very much effective. In this paper to remove the inclination of traditional ID3 algorithm an enhanced ID3 (wID3) algorithm is proposed. In proposed algorithm the attribute with maximum Gain Ratio is multiplied with the weight, which gives it a new value. From the new values, attribute with maximum Gain Ratio is selected as a node of the tree. Information gain is substituted by additional normalized Gain ratio as given in equation (3,6).

$$\text{Gain Ratio (A)} = \text{Gain(S,A)} / \text{Entropy(S)} \quad (3)$$

wID3 Algorithm Steps:

1. Begin
2. Generate a node N
3. If "All samples belongs to same class" Return node as leaf with the class name;
4. If "attribute lists are empty" Return node as leaf node labelled with most common class;
5. Calculate the weight of each attribute
6. Select test attribute (attributes having highest gain ratio)
7. Label node N with test attribute
8. For each known value of a of test attribute, grow branches
9. from node N for the condition test attribute = a;
10. Let S i be set of samples for which test attribute = a I;
11. If (S i is empty) then attach the leaf labeled with most common class in sample.
12. Else attach the node returned by
13. generate decision tree(Si; attribute list test attribute)
14. End

C. C4.5

Decision trees can be generated using a C4.5 algorithm which is extension ofID3 algorithm. The disadvantages of ID3 algorithm is overcome by C4.5. To predict the class, decision trees are generated by using C4.5 algorithm. It is referred as a statistical classifier. To improve ID3 algorithm there are number changes made which resulted as C4.5 algorithm. Few of the changes are as follow:

i. Managing breeding data with absent values of attributes.
ii. Control dowering cost properties.
iii. After creation of decision tree its Pruning is done.
iv. Handling attributes with continuous and discrete values.

C4.5 algorithm general working steps are as follows.

i. Assume all samples in list fit into the same category. If condition is true, it will create a leaf node of the decision tree to select a particular class.
ii. None of the features provide any information gain C4.5 creates a decision node higher up the tree using the expected value of the class.
iii. Instance of previously-unseen class is encountered and then C4.5 creates a decision node higher up the tree using the expected value.

V. COMPARISON BETWEEN ID3 AND IMPROVED ID3

To compare the performance of ID3 and improved ID3 algorithms, a sample loan application data set is considered. The sample training data set is represented in Table 1. The category attribute of the sample set is "Class", which will predict whether to approve or not the new customer's loan application.(3)

ID3 algorithm is applied on the sample loan dataset to generate the decision tree. The generated decision tree by the ID3 algorithm is shown in Figure 1. Figure 2 shows the decision tree when the improved ID3 algorithm is used on the same training sample loan dataset. (1)

TABLE I. TRAINING SAMPLE

Index	Age	Has_Job	Own_House	Credit	Class
1	Young	False	False	Good	No
2	Young	True	False	Good	Yes
3	Young	True	True	Fair	Yes
4	Young	False	False	Fair	No
5	Middle	False	False	Good	No
6	Middle	True	True	Good	Yes
7	Middle	False	True	Excellent	Yes
8	Low	False	True	Excellent	Yes
9	Low	False	True	Good	Yes
10	Low	True	False	Good	Yes

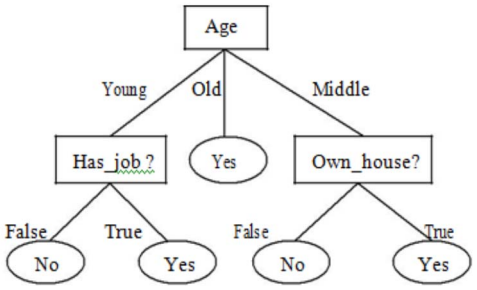

Fig 2. ID3 Decision Tree

Comparison of the decision trees generated by both the algorithm shows that the improved ID3 algorithm generates more optimized decision tree than the traditional ID3 algorithm. Better classification rules are provided by the improved ID3 algorithm.

In the proposed system to predict the performance of students, ID3 and improved ID3 algorithm is applied on the student database. The database consists of fifty student's information based on which classification is done. Classification accuracy given by the algorithms is shown in table 2. Improved ID3 algorithm has given more classification accuracy.

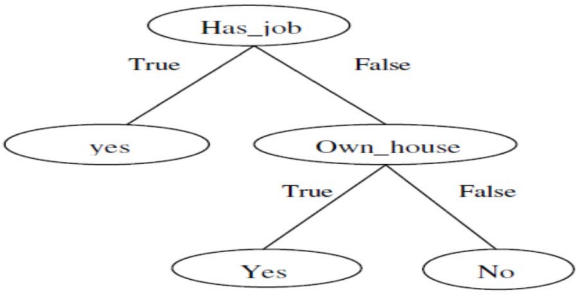

Fig 3: Decision tree using Improved ID3

Table II. Percentage classification accuracy of Proposed System

Algorithm	Total Students	Students whose results are correctly predicted	Accuracy	Execution Time
ID3	50	35	70%	47.6
Improved ID3	50	37	74%	49.4
C4.5	50	36	72%	39.1

VI. CONCLUSION

The paper illustrates the proposed system to predict the fourth year result of third year students based on their current and previous performance. Overall the student performance analysis system is proposed using data mining technique of classification to predict the performance of current students. Comparison of decision tree generation algorithms C4.5, ID3 and improved ID3 algorithm is carried out. Improved ID3 algorithm gives better performance as compared to traditional ID3 & C4.5 algorithm.

REFERENCES

[1] "Prediction Students Performance using ID3 and C4.5 Classification Algorithms" - International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013.

[2] "Student's Performance Prediction Using Weighted Modified ID3 Algorithm" - International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 4, Issue 5, May 2015.

[3] "Predicting Students' Performance using Modified ID3 Algorithm"- ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013

[4] Student Performance Analysis System (SPAS)" - Chew Li Sa, Dayang Hanani bt. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossin Faculty of Computer Science and Information System Universiti Malaysia Sarawak (UNIMAS) 94300, Kota Samarahan, Sarawak, Malaysia.

[5] " Efficient Processing of Decision Tree Using ID3 & improved C4.5 Algorithm "- Sonal Patil. et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1956-1961.

[6] " Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree. International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 11, Issue 02 (February 2015), PP.44-47

[7] " Implementation of ID3 Algorithm" International Journal of Advanced Research in Computer Science and Software Engineering