STUDENT PERFORMANCE PREDICTION SYSTEM USING DATA MINING CLASSIFICATION

Aniket Katkar ,Jalindar Veer, Ajim Shaikh, Sharvari Shravage

[1] Aniket Katkar ,Computer Engg,PCCOE, Maharashtra,India
[2] Jalindar Veer, Computer Engg,PCCOE, Maharashtra,India
[3] Ajim Shaikh, Computer Engg,PCCOE, Maharashtra,India
[4] Sharvari Shravage, Computer Engg,PCCOE, Maharashtra,India

ABSTRACT

In education system, highest level of quality can be achieved by exploring the knowledge regarding prediction about student's performance. Data mining techniques play an important role in data analysis. 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. We have analyzed the data set containing information about students, such as gender, marks and rank in entrance examinations and results in Third year of the previous batch of students. Classification is a data mining technique that maps data into predefined groups or classes. It is a supervised learning method which requires labeled training data to generate rules for classifying test data into predetermined groups or classes. It is a two-phase process. The first phase is the learning phase, where the training data is analyzed and classification rules are generated. The next phase is the classification, where test data is classified into classes according to the generated rules. By applying the ID3 (Iterative Dichotomiser 3), C4.5, Improved weighted modified ID3 classification algorithms on this data, we have predicted the general and individual performance of third year students in future examinations.

Keyword: - Data *Mining, Decision Tree, Classification, ID3, C4.5,Predicting Performance..*

1 Introduction

Every year educational institutes admits students under various courses from differernt locations, educational background and with varying merit

scores in entrance examinations.Morever engineering colleges may be affiliated to different universities, each university having different subject in their syllabus and also different level depth in

their subjects. Analyzing past performance of admitted students would provide a better perspective of propable performance of students in future. This can very well be achieved by our system which will be using data mining classification algorithms. The ability to monitor the progress of students academic performance is critical issue to the academic cumminity of higher learning. We will propose system with better accuracy for predicting students performance using data mining classification algorithms. System also helps weaker students to improve and bring out betterment in result. Students are main asset for various universities and students play and important role in producing graduates of high quality with its academic performance achievement. Academic performance achievements is the level of achievement of the students educational goal that can be measure and tested through examination, assessments and other form of measurements.

The users of our system are teachers of particular department of the institute. The user is assume 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.

Student academic performance is stored in the system in different format such as file, document, records, images and other formats. This data will be extracted to produce useful information. Thus our system will provide a tool necessary for institute to extract useful information. This useful information will be use to predict student performance.

illustration displays and report generation. A good user interface provides an user-friendly interface as

it is easy to be navigate 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

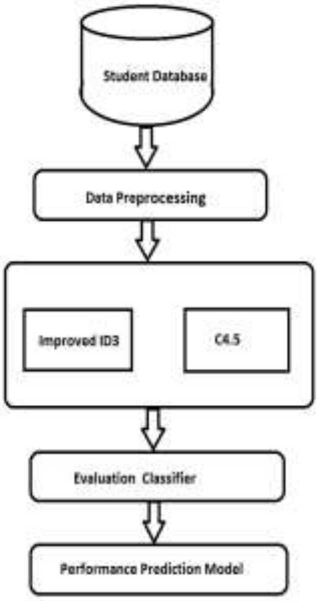

Fig 1: System Architecture

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 algorithams result will be displayed

1.1 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. 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 to the 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 ID3, Improved ID3 and C4.5 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.

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, application ID, gender, caste, percentage of marks obtained in second year percentage of marks obtained in the entrance examination, admission type, etc. Forease of performing data mining operations, the data was filled into a database. Database also consists of list of the teachers associated with department.

3.2. Data Preprocessing

Once we had details of all the students, we then segmented the training dataset further, considering various feasible splitting attributes, i.e. the attributes which would have a higher impact on the performance of a student.

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

3.3 Functions

FileUploading()-

This function will take data of student to be processed by our system and data will be taken in two format.

1.For bulk evaluation:

Batchwise evaluation

Divisionwise evaluation

Departmentwise evaluation

2.Single student evalution

DataProcessing()-

1.Extra information will be removed from collected data.

2.Relevant information will be fed into database.

3.Final format of database will be decided in this function.

4.Algorithms

4.1 ID3 algorithm

The ID3 algorithm is a recursive procedure, which in each step there is an evaluation of a subset and there is the creation of decision node, based on metric called Information Gain, until the subset in evaluation if specified by the same combination of attributes and its values. ID3 algorithm is used to create a decision tree from given set, by using top-down greedy search to check each attribute at every tree node & information gain is used as metric to generate tree.Id3 algorithm uses the information gain as a metrics to select the best attribute in each step.

The ID3 decision makes use of two concepts when creating a tree from top-down:

1. Entropy
2. Information Gain

4.1.1.Entropy

Given probabilities p1, p2, …, ps, where _pi = 1, Entropy is defined as

H(p1, p2, …, ps) = _ - (pi log pi)

Entropy finds the amount of order in a given database state. A value of H = 0 identifies a

perfectly classified set. In other words, the higher the entropy, the higher the potential to improve

the classification process

4.1.2. Information Gain

ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The formula used for this purpose is: G(D, S) = H(D) - _P(Di)H(Di)

4.1.3. Shortcomings of ID3 algorithm

There exist one problem with this approach; this means that id3 selects the attribute having more no. of values, which are not necessarily the best attribute Data may be over-fitted or over-classified, if a small sample is tested. Only one attribute at a time is tested for making a decision.

Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

4.2. Improved ID3

One of the main drawbacks of the ID3 algorithm is that it is inclined towards the attributes with more values. This can be a wrong selection and hence, as a result, the tree generated is may not be very much e_cient. For removing the inclination of traditional ID3 algorithm, an improved weighted ID3 (wID3) algorithm is proposed in this paper. In this, the attribute with highest Gain Ratio is multiplied with a weight which gives it a new value, and among the new values, attribute with highest Gain Ratio is selected as a node of the tree. Also, information gain is replaced by gain ratio, which in more normalized.

Gain Ratio (A) = Gain(S,A) / Entropy(S)

wID3 Algorithm:

1.Begin

2.Create a node N

3.If (All samples are in same class) Return node as leaf with class name;

4.If (attribute list is empty) Return node as leaf node labeled with most common class;

5.Calculate the weight of each attribute

6.Select test attribute i.e. 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 insample.

12.Else attach the node returned by 13.generatedecisiontree(Si; attributelisttestattribute)

14.End

4.3. C4.5

C4.5 is a well-known algorithm used to generate a decision trees. It is an extension of the ID3 algorithm used to overcome its disadvantages. The

decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classfier The C4.5 algorithm made a number of changes to improve ID3 algorithm

Some of these are :

1.Handling training data with missing values of attributes.

2.Handling di_ering cost attributes

3. Pruning the decision tree after its creation

4. Handling attributes with discrete and continuous values General working steps of algorithm is as follows.

5. Assume all the samples in the list belong to the same class. If it is true, it simply creates a leaf node for the decision tree so that particular class will be selected.

6. None of the features provide any information gain. If it is true, C4.5 creates a decision node higher up the tree using the expected value of the class.

7. Instance of previously-unseen class encountered. Then, C4.5 creates a decision node higher up the tree using the expected value.

5.Comparison between ID3 and Improved ID3

A simple loan application dataset is shown in below Table. The category attribute of the sample set is "Class", which will predict whether the new customer’s loan application should be approved or not

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

Table 1. Training Sample

Both improved ID3 algorithm and ID3 algorithm are applied on this dataset to construct decision trees and comparison is made. Figure 1 and figure 2 show the generated decision trees using the ID3 algorithm and the improved ID3 algorithm, respectively

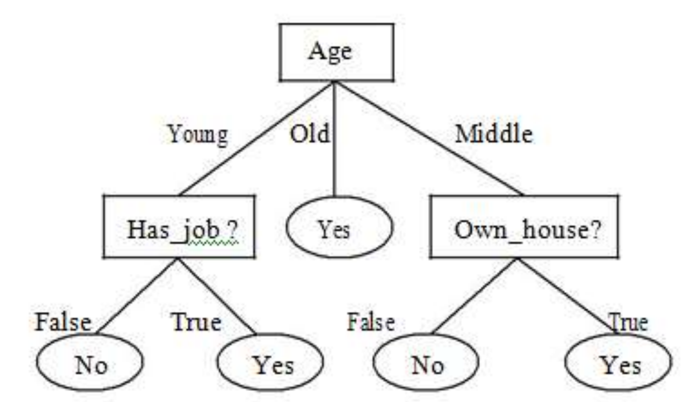

Fig 1.ID3 Decision Tree

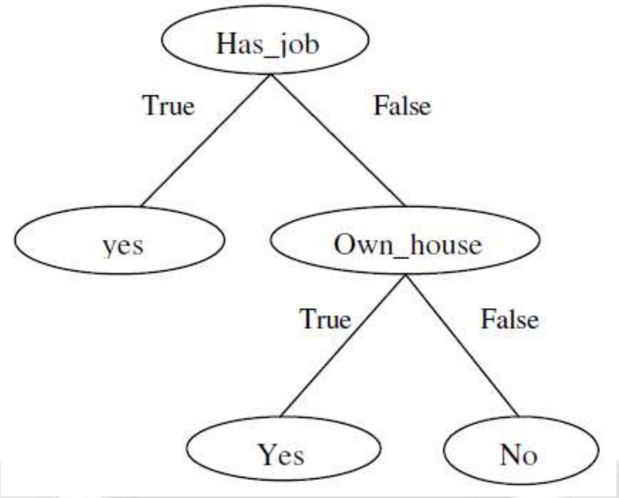

Fig 2: Decision tree using Improved ID3

Thus Decision tree generated by improved ID3

Algorithm is more optimized than decision tree generated by ID3 Algorithm and provides better classification rules.

6. CONCLUSION

In this paper, we have explained the system we have used to predict the results of students currently in the Third year of engineering, based on the results obtained by students currently in the Fourth year of engineering during their Third year. the project concentrates on the development of a system for student performance analysis. A data mining technique, classification algorithm is applied in this project to ensure the prediction of the student performance in course final year engineering is possible.

7. 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