IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Student Information System and Performance Prediction in Educational Data Mining

Pratik Nanavati[1], Abhishek Masurkar[2], Chaitanya Shinde[3], Aaryaman Singh[4], Prof. Sumitra Sadhukhan[5]

Student, Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Mumbai, India [1,2,3,4]

Asst. Professor, Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Mumbai, India[5]

Abstract: The main focus of data mining is to collect different data from databases or data warehousesand the information collected that had never been known before, it is valid and operational. Educational institutes can use this to maintain all the information of student academics easily which is critically important. The performance of students in their academics is a turning point for their brightest career. Predicting student academic performance has been an important research topic in Educational Data Mining (EDM) which uses machine learning and data mining techniques to explore data from educational settings. Measuring student academic performance is challenging since it depends on various factors. Classification and Prediction are among the major techniques in data mining and plays a vital role in EDM. The need for this is to enable the university to intervene and provide assistance to low achievers as early as possible. In this study we develop a classification model usingC4.5 algorithm for domain wise performance evaluation system for engineering students. It also brings connectivity between teachers, students and parents by keeping them updated with their child performance regularly. The whole system will be available through a secure, online interface embedded in college website.

Keywords: Educational Data mining, Student information system, Performance prediction, Decision tree, C4.5 algorithm.

I. INTRODUCTION

A. Student information system

The main purpose of designing and implementation of a comprehensive student information system and user interface is to replace the current paper records [9]. The system provides staff UI which can access all aspects of of a student's academic progress through a secure, online interface embedded in the college's website. In addition to this it also provides student and parent UI , allowing users to access information. The system uses user authentication, displaying only relevant data. Each sub-system has authentication allowing authorized users to create or update data in that sub-system.

All data is stored securely on SQL servers managed by college administrator. All the authentication details like username, password etc. are sent through an email to the registered id. Previously, this records were stored on papers which had many drawbacks. This system provides a simple interface for the maintenance of student information. Achieving this is difficult using traditional method as the data is scattered, redundant and time consuming. This system focuses on proving data in an easy and intelligent manner to students, college staff and parents.

B. Data mining

Data mining also called as Knowledge Discovery from Databases (KDD) is defined as the activity of extracting knowledge from huge amount of data and identifying patterns accordingly. This field combines tools from statistics and artificial intelligence with database management to analyse huge digital thing, known as data sets. Data mining tasks has two types supervised learning and unsupervised learning.

In supervised learning, Classification is a basic task in data analysis that need for the construction of classifier. In supervised learning classification is there and it uses decision tree method in which its rule set is used to predict class of data. Different classification algorithms are NavieBayes, Bayes Net, OneR and J48, random decision tree, random forest algorithm.
In unsupervised learning, Clustering and Association are there which are used to find hidden data in data sets to make decisions. Clustering is a technique to group similar data in one cluster to find common patterns from set of data without labels. Different clustering algorithms are Density Based, Simple K-Means, Hierarchical clustering algorithm. Association is a technique for discovering relations between variables in databases. It is often used to identify well built rules discovered in databases using different measures of interestingness [4].

C. Educational data mining (EDM)

EDM uses machine learning and data mining techniques to explore data from educational settings to find out predictions that characterize students behaviour and performance. This is a great concern to the management as

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

academic performance depends on diverse factors. Data mining techniques can be applied on student data sets to predict their performance. Predicting student performance helps to improve their academic results.

By this management can find the weak point of students and directly target that point to improve which can help to create better workforce for country's development. Nowadays quality of a student matters the most so, other than teaching this type of extra attention is essential

II. RELATED WORK

EDM has emerged as a very active research area as many thing in this field are not exposed. Work connected to student performance, faculty performance and impact on this on students final performance needs attention.

[1] The paper depicts the users, components as well as the various approaches in EDM.
[2] In this paper a strategy to improve the student's performance is mentioned by mapping the student's record using K-mean clustering algorithm and grouping datasets into cluster but there is no future performance prediction.
[3]Bharadwaj & Pal proposed ID3 decision tree algorithm as a classification model to predict the students division, the previous information such as attendance, class test, seminar and assignment marks were collected from the student's previous databases to predict the performance at the end of semester. All this helped the students and the teachers to improve the division of the students.
[4] Uses two classification algorithm J48 and Random tree to predict performance of MCA student in this work random has better accuracy and it consume less time than J48 algorithm. So prediction of student performance using random tree classification algorithm can be more efficient than J48.
[5] Presented a deep learning architecture for predicting performance by automatically learning multiple levels of representation.
[7] Uses classification algorithm ID3 and C4.5 to identify various categories of students performance. It also concludes that the performance of C4.5 algorithm is high compared to other.
[8] The model predicts student's future learning outcomes using data sets of senior students using J48 algorithm which proved to be accurate than other .
[10] A web based system is made which manages all the student related data. This system is made using HTML, CSS, Java script, PHP, SQL.
[11] This paper proposes an android application for the management of student information which automates the existing manual system.
[12] Decision tree can be used on students past performance data to generate the model and this model can be used to predict the performance. It will enable to identify the students in advance who are at risk. Giving warning to the students those are on risk of failing the students can improve their performance.

III.CLASSIFICATION

Classification is a supervised learning method with two steps.(1)By analyzing the data tuples from training data a model is built. (2) Test data is used to check accuracy. Here we use Decision tree classification algorithm.

A. Classification by Decision tree
The decision tree systems is a non-backtracking top-down approach. Following algorithm is used.

1) Create a node N.
2) If all the tuples in the partition are of the same class then return N as a leaf node labeled with that class.
3) If attributes list is empty then return N as a leaf node labeled with the most common class in samples. 4) Identify the splitting attribute so that resulting partitions at each branch are as pure as possible.
5) Label node N with splitting criterion which serves as test at that node.
6) If splitting attribute is discrete valued then remove splitting attribute from attribute list.
7) Let Pi be the partitions created based on the i outcomes on splitting criterion.
8) If any Pi is empty then attach a leaf with the majority class in the partition to node N.
9) Else recursively apply the complete process on each partition.
10) Return N. [12]

IV. DECISION TREE ALGORITHMS

There are many decision tree algorithms such as CART, ID3, C4.5 can be used for predicting students performance. In this paper C4.5 has been used to create decision tree.

A. C4.5 (j48) algorithm
This algorithm is a successor to ID3 developed by Quinlan Ross. The Hunt's algorithm.C4.5 algorithm generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy. C4.5 has the concept of Gain Ratio as an attribute selection measure to create a decision tree. C4.5 uses pessimistic pruning to get rid of unessential branches within the decision tree to enhance the accuracy of classification.

B. Fold cross-validation
When the data is less then fold cross validation can be used. In this the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is taken as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross validation process is then repeated k times with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or combined) to produce a single judgment.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

V. PROPOSED FRAMEWORK

A. System overview

EDM has a large amount of data that has to be arranged in consistent manner. To enhance existing system the proposed model is designed by collecting data and classifying them based on student performance in a particular domain. The performance is classified as

- Poor
- Average
- excellent

the system framework is as shown below which provides an efficient analysis on student performance by data collection and prediction.

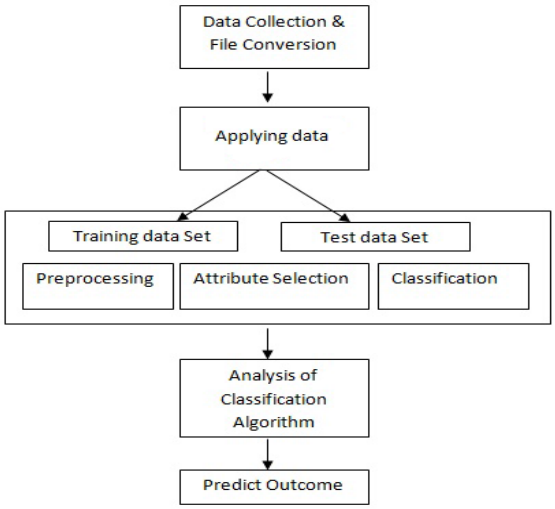

Fig 1. System Framework

VI. METHODOLOGY

In this section the steps of traditional KDD process would be followed. The process starts from data collection and data pre-processing followed by classification model construction and ends with model evaluation and interpretations. Steps in KDD process is shown in Fig 2.

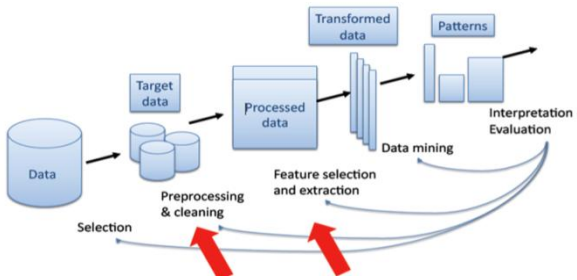

Fig 2. Steps of KDD process

A. Information gathering

The data set for study has been collected from examination cell. The study considers the academic performance of students from semester 1 to semester 7 of B.E program. Initially data is collected in an excel sheet which collects all the grades of all semesters.

B. Pre-processing

1. Selection of Attributes:

There may be many attributes available but all the attribute are not useful for prediction. So only the necessary attributes are taken into consideration to decide the values.

2. Deciding values for Attributes:

It is necessary to decide values for attributes to avoid continuous data. This can be done by using discrete data. E.g. if average marks of the domain is from 32 to 49 then Poor. If between 50 to 59 then average. If between 60 to 80 then excellent.

TABLE I FONT SIZES FOR PAPERS

Attributes	Possible values
Theory & Concepts domain	Poor, average, excellent
Programming domain	Poor, average, excellent
Networking domain	Poor, average, excellent
Maths & Logic domain	Poor, average, excellent
S/w development domain	Poor, average, excellent
1st year grade	Pass, 1stclass, 2nd class, distinction, Fail
2nd year grade	Pass, 1stclass, 2nd class, distinction, Fail
3rd year grade	Pass, 1stclass, 2nd class, distinction, Fail

C. Model building

In this stage, decision tree has been selected as a classifier under cross validation method. For model construction C4.5 decision tree method is used based on attributes selected. The attribute having maximum gain ratio value is selected for splitting the node. This process continues till the entire tree is built. Fig 3 shows the decision tree construction. Leaf nodes are represented by rectangle and root node by oval.

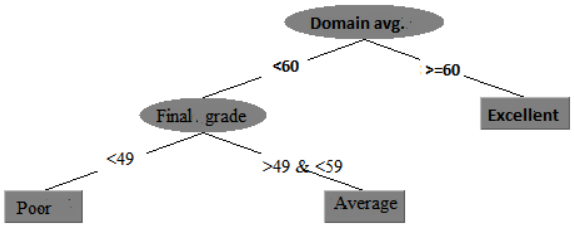

Fig 3. Decision Tree

C. Interpretation

The prediction of the result can be done for all the domains.

1.Prediction for Programming Domain:

The result shown here is for programming domain. Taking into consideration the marks obtained in programming domain subjects like Structured Programming approach

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

(spa), OOPM, data structures, analysis of algorithm (AOA) from different semesters the prediction of the student performance in that particular domain can be done. The decision tree for programming domain is shown in the Fig 4.

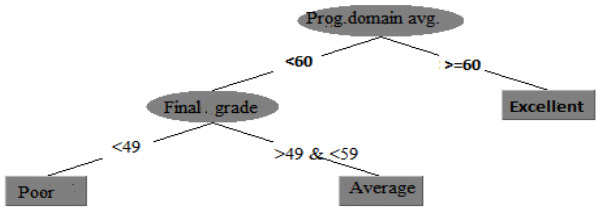

Fig 4. Decision tree for Programing Domain

2.Rules generated for Programming Domain:

- If the average domain marks is greater than or equal to 60 (>=60) then the performance of student is excellent.
- if average domain marks is less than 60 (<60) and final marks are in the range of 49-59 then the performance is average
- if average domain marks is less than 60 (<60) and final marks are less than 49 then the performance in that domain is poor.

As we can see from the above decision tree, the root attribute is average marks of programming domain. If the average marks is more than 60 than the student is excellent in programming subjects. If he gets average marks less than 49 than he is poor in that domain.

Similarly different rules can be generated for different domains like Theory & Concepts domain, Networking domain, Maths & Logic domain, S/w development domain by categorizing subjects into this domains and taking the average of marks of all the subjects in that particular domain.

VII. CONCLUSION AND FUTURE WORK

Educational data mining main focus is to analyse the education system. This paper demonstrates classification method to predict student performance. It also assist in automating the existing manual system by providing the Web Based Information System. It creates connectivity between parents and college. All the stakeholders, faculty and management can get the required information without delay. This whole model can be useful in educational system like MCT Rajiv Gandhi Institute of Technology, Mumbai. Thus improving their standards and performance.The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform.

The work can be further extended out by designing the student model analysing records of students extra-curricular skills and provide a suggestions on communication and technical skill development by which students can be built in professional aspect of talents.

ACKNOWLEDGMENT

We would like to thank our guide, **Prof. Sumitra Sadhukhan**, Department of Computer Engineering, for all the advice, Encouragement and constant support that she has given throughout our project work. We would also like to thank Dr. Satish Ket, HOD and Examination cell for their support and valuable suggestion has made our project achievable

REFERENCES

[1] Parneet Kaura, Manpreet Singhb, Gurpreet Singh Josanc "Classification and Prediction based Data Mining Algorithms to Predict Slow Learners in Education Sector" Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC-2015).

[2] Baradwaj Brijesh Kumar and Pal Saurabh (2011).Mining Educational Data to Analyze Student Performance. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.

[3] Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA

[4] Bo Guo, Rui Zhang, Guang Xu, Chuangming Shi, Li Yang, "Predicting Students Performance in Educational Data Mining", 2015 International Symposium on Educational Technology , 978-1-4673-7370-8/15©2015 IEEE.

[5] Krina Parmar Prof. Dineshkumar Vaghela Dr Priyanka Sharma, "Performance Prediction of students using Distributed Data mining", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems, 978-1-4799-6818-3/15 © 2015 IEEE

[6] C. Anuradha, T. Velmurugan "A Data Mining based Survey on Student Performance Evaluation System." 2014 IEEE International Conference on Computational Intelligence and Computing Research, 978-1-4799-3975-6/14 ©2014 IEEE

[7] R. Sumitha, E.S. Vinothkumar "Prediction of Students Outcome Using Data Mining Techniques"International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6,June 2016

[8] Zhibing Liu, Huixia Wang,Hui Zan "Design and implementation of student information management system." 2010 International symposium on intelligence information processing and trusted computing. 978-0-7695- 4196-9/10 IEEE

[9] S.R.Bharamagoudar, Geeta R.B., S.G.Totad, "Web Based Student Information Management System",International Journal of Advanced Research in Computer and Communication EngineeringVol.2,Issue6,June2013

[10] Manasi Kawathekar, Kirti K. Bhate and Pankaj Belgoankar "An Android Application for Student Information System" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 9, Sept 2015

[11] R. R. Kabra And R. S. Bichkar(2011),‖Performance Prediction Of Engineering Students Using Decision Trees‖, International Journal Of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011.