Student Performance Prediction System with Educational Data Mining

Karishma B. Bhegade
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

Swati V. Shinde, PhD
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

ABSTRACT

In this paper we apply data mining tools to predict college failure and dropout. In Current year the researcher focuses on the new area of analysis like Educational data mining (EDM). Educational data mining techniques drawn from varied literatures which have data mining and machine learning. In this paper we are collecting the student's information from Pimpri Chinchwad College of Engineering which comes under Pune University. We have preprocessed the information that we have collected for removal of unwanted information. Based on the classification rules student dropout and failure is being predicted. By using all available features, the experiments are conducted for improving the accuracy to predict which student has failed. In this paper C4.5 decision tree algorithm is proposed for prediction of students. C4.5 is the popular decision tree classifier in data mining. Accuracy of this classification algorithm is compared in order to check best performance. After tree building the ranking of the student is calculated on the basis of the student's internal assessment. And then the frequent patterns are generated by using FP growth algorithm.

General Terms

Student failure prediction, Classification.

Keywords

Educational data mining (EDM), Data mining, Decision Tree, C4.5 algorithm, rule generation.

1. INTRODUCTION

Educational data mining is wide area that provides machine learning, statistical information as well as different kinds of data mining algorithm to find out educational datasets. Schools as well as colleges have necessity to judge the academic efficiency of students by grades or external and internal marks [3].Future details such as carrier option are predicted about students using different kinds of prediction models and probability of teenagers to gain brutal future For this purpose different kinds of methods such as clustering, association data mining and diverse classifications are used. In proposed system different types of classifier are utilized as ID3 and C4.5 algorithm [4].Observation of all above information is created diverse decision trees for searching out suitable algorithm.

Educational data mining is recently developed trend and interesting method that provides diverse predictions in all educational levels. Numerous methods of data mining are presented following.

A. Calculation of Students academic performance

B. Predicting School dropouts

C. Students behavioral prediction

A. Calculation of Students academic performance

We present information analysis of datasets to predict the student's academic marks as well as student's placements were accomplished depends on previous record [3]. To increase in the academic performances of graduate students, we additionally provided numerous kinds of data mining methods such as clustering, association rule mining and classification and outlier detection.

B. Predicting School dropouts

A method additionally proposed for checking list of students who dropouts the college [6]. In this diverse attributes such as attendance, family background and gender determined for data mining [7]. Prediction of dropout is also done with use of decision tree.

C. Student's behavioral prediction

We proposed technique with Behavior Attitude Relationship Clustering (BARC) Algorithm. With help of algorithm we present the improvement in student's performance as well as relationship with faculty members and attitude for predicting their behavior [8].Machine learning method additionally demonstrates to search out if a student utilizes an intelligent tutoring framework is off stack. Mining and latent response models were used statically on the basis of the system log files [9].

2. LITERATURE SURVEY

The author proposed data mining process for evaluation of school dropout and failure [10]. Experiment done on real information of 670 school students from Zacatecas, Mexico and white box classification strategies such as decision tree and induction rules were used. Analysis of the state of craftsmanship with related to EDM and observation of the strong performance of this kind of date. Every student is classified on the basis of type of data and DM methods and resolved through type of instructional occupation [11]. Author analyzed use of data mining in training for student's profile [12] and collections. Author used apriori algorithm of data mining for student's profile. K-means clustering algorithm used for students for transfer a set of analysis within subset. Author additionally demonstrates use of data mining method for classification and helps to students in selection of UG programs. This paper additionally explores analysis on educational structure in Thailand as well as base of data mining process [13].

3. PROPOSED WORK

3.1 System Overview

Figure 1 shows the system architecture. For the test we are collecting the student's information from Pimpri Chinchwad Collage of Engineering which comes under Pune University. We preprocess the information we collected for deletion of unwanted information. Based on the rule student dropout and failure is being predicted. In proposed work we will use C4.5 algorithm to predict student failure. Accuracy of these classification algorithms is compared in order to check best performance. Student ranking is done on the basis of student's internal assessment. The ranking of student will be decided by average percentage calculated and by sorting average percentage in descending order. And then the frequent patterns are generated by using FP growth algorithm.

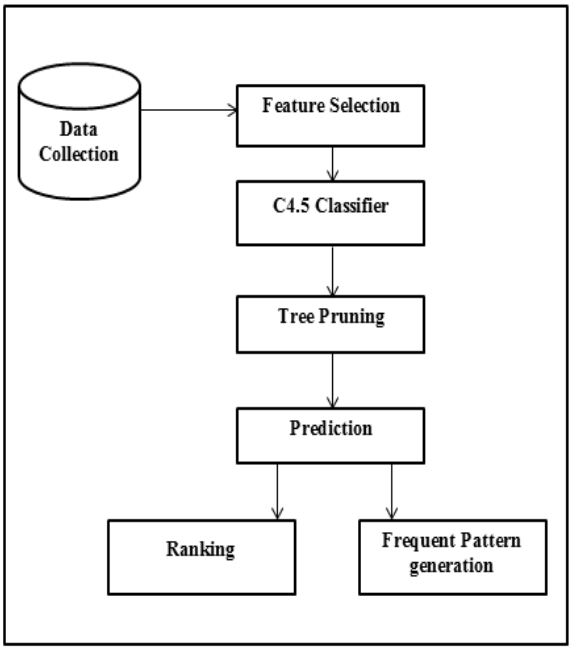


Fig.1 Proposed System Architecture

3.2 C4.5 Algorithm

Step 1: Read trained data instances.

Step 2: Calculate Overall entropy

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

Step 3: Calculate entropy of every attribute

$$M_i = Entropy(N_i) = -\sum_j p(j|N_i) \log p(j|N_i)$$

Step 4: Calculate information gain of every attribute

$$Gain_{split} = Entropy_{(p)} - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i))$$

Step 5: Build Tree

Step 6: build prune trees

3.3 Mathematical Model

Let, System S is represented as:
S = {A, B, C, D, E}

- **Data Gathering:**

Let, A is a set of student's dataset
A= {a1, a2…}
Where,
a1, a2… are the students data gathered.

- **Attribute Selection:**

Let B is a set attribute selection
B = {f1, f2, f2, f3, f4, f5….fn}
Where,
f1, f2, f3, f4, f5 are the selected attributes.

- **Classifier:**

Let, C is a classifier C = {e1; e2}
Where,
e1, e2 are the classification process.

- **Ranking:**

Let, D is a ranking process,
D = {d1, d2, d3….dn}
Where,
d1, d2… are number marks obtained by students.

- **Frequent pattern generation:**

Let, E is the set of frequent pattern generation,
E= {R1, R2, R3, R5, R6…. Rn}
Frequent Patterns:
R1= {r1, r2, r3}
R2= {r2, r3, r4}
R3= {r3, r 2}
R4= {r4, r5, r2}

Calculating Support Value:

Let S is the support value of each rule.

$$S = \frac{\sum records\ in\ complete\ database}{|No\ of\ transactions|}$$

Generating patterns:

If $S \geq T$ then rule is frequent else rule is infrequent.

Where,

T = Threshold Value

S = Support Value

4. RESULTS AND DISCUSSION

The real student dataset used taken from the pimpri chinchwad college of engineering to calculate the results. The dataset contains 10 attributes and 2 classes. The total size of the dataset is 200.

Following are the attributes information:

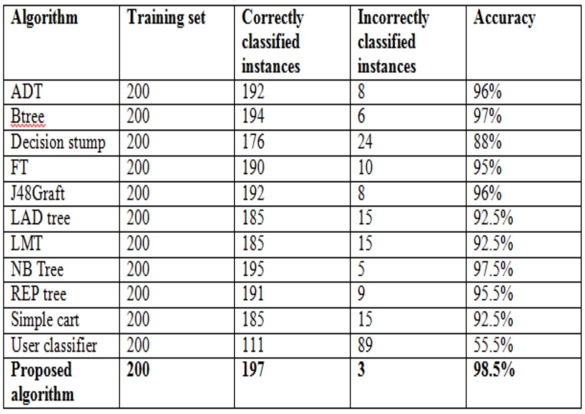
Data Structure (DS), Computer organization (CO), Digital Electronic and logic design (DELD), Fundamentals of data structure (FDS), Problem solving and object oriented programming (PSOOP), Engineering math's III (m3), Computer Graphics (CG), Processor architecture and interfacing (PAI), Data structure & files (DSF), Attendance (%).

Two classes are used:

GOOD, AVERAGE

Table 1 represents the comparison of various classifiers with proposed algorithm. Table clearly demonstrates that the proposed algorithm is better than other classifier in terms of accuracy. It correctly classifies the instances with higher accuracy up to 98.5 percent.

Table 1: Comparison of Classifier



Algorithm	Training set	Correctly classified instances	Incorrectly classified instances	Accuracy
ADT	200	192	8	96%
Btree	200	194	6	97%
Decision stump	200	176	24	88%
FT	200	190	10	95%
J48Graft	200	192	8	96%
LAD tree	200	185	15	92.5%
LMT	200	185	15	92.5%
NB Tree	200	195	5	97.5%
REP tree	200	191	9	95.5%
Simple cart	200	185	15	92.5%
User classifier	200	111	89	55.5%
Proposed algorithm	**200**	**197**	**3**	**98.5%**

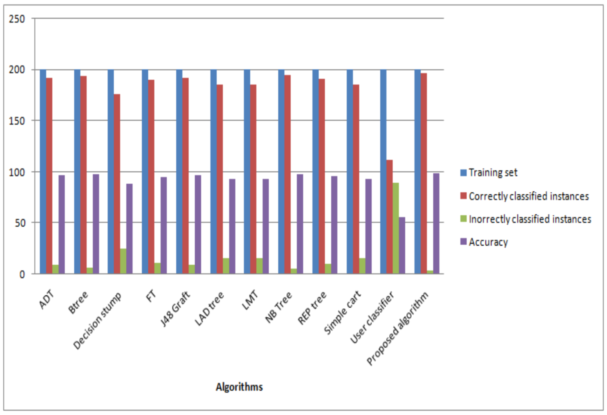
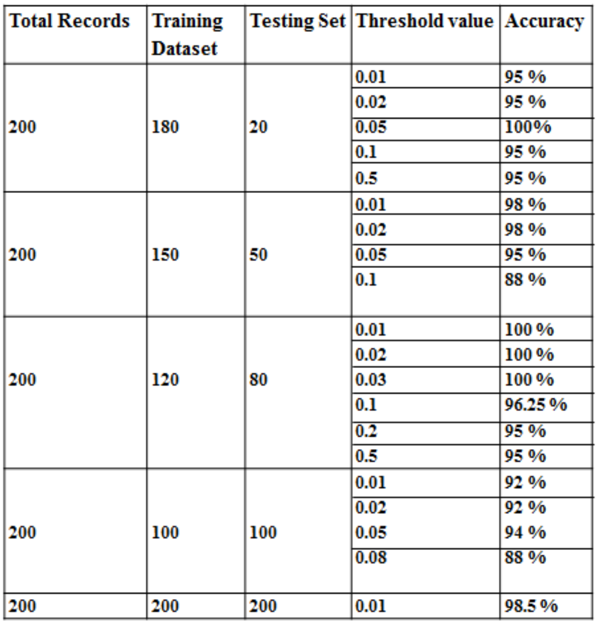


Fig. 2: Performance Measure Graph

As shown in table 2 the C4.5 classifier is used to calculate the results. Here the 200 training samples are used. While calculating, the total samples are divided into number of training and testing sets. And then the accuracy is calculated.

Table 2: C4.5 Classifier Results with different training and testing files



Total Records	Training Dataset	Testing Set	Threshold value	Accuracy
200	180	20	0.01	95 %
			0.02	95 %
			0.05	100%
			0.1	95 %
			0.5	95 %
200	150	50	0.01	98 %
			0.02	98 %
			0.05	95 %
			0.1	88 %
200	120	80	0.01	100 %
			0.02	100 %
			0.03	100 %
			0.1	96.25 %
			0.2	95 %
			0.5	95 %
200	100	100	0.01	92 %
			0.02	92 %
			0.05	94 %
			0.08	88 %
200	200	200	0.01	98.5 %

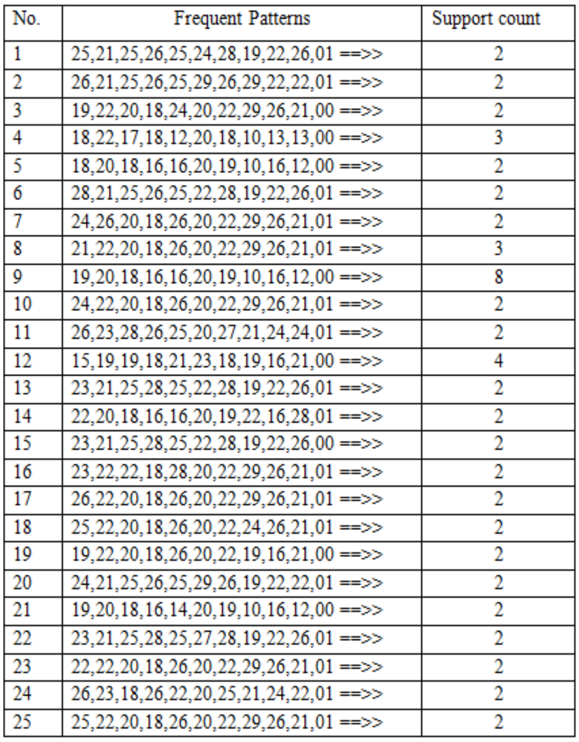
FP Growth results:

FP Growth is the basic algorithm use to generate Association rules. FP growth is an approach based on divide and conquers method. The main purpose behind this technique is to produce frequent item sets by using the combination of data attributes. It basically works on to generate frequent item set without candidate set generation

Generated frequent patterns:

The frequent patterns are generated on threshold value 2. Only those patterns are considered whose having support count 2 or >= 2. And those patterns are the frequent patterns

Table3. Frequent patterns



No.	Frequent Patterns	Support count
1	25,21,25,26,25,24,28,19,22,26,01 ==>>	2
2	26,21,25,26,25,29,26,29,22,22,01 ==>>	2
3	19,22,20,18,24,20,22,29,26,21,00 ==>>	2
4	18,22,17,18,12,20,18,10,13,13,00 ==>>	3
5	18,20,18,16,16,20,19,10,16,12,00 ==>>	2
6	28,21,25,26,25,22,28,19,22,26,01 ==>>	2
7	24,26,20,18,26,20,22,29,26,21,01 ==>>	2
8	21,22,20,18,26,20,22,29,26,21,01 ==>>	3
9	19,20,18,16,16,20,19,10,16,12,00 ==>>	8
10	24,22,20,18,26,20,22,29,26,21,01 ==>>	2
11	26,23,28,26,25,20,27,21,24,24,01 ==>>	2
12	15,19,19,18,21,23,18,19,16,21,00 ==>>	4
13	23,21,25,28,25,22,28,19,22,26,01 ==>>	2
14	22,20,18,16,16,20,19,22,16,28,01 ==>>	2
15	23,21,25,28,25,22,28,19,22,26,00 ==>>	2
16	23,22,22,18,28,20,22,29,26,21,01 ==>>	2
17	26,22,20,18,26,20,22,29,26,21,01 ==>>	2
18	25,22,20,18,26,20,22,24,26,21,01 ==>>	2
19	19,22,20,18,26,20,22,19,16,21,00 ==>>	2
20	24,21,25,26,25,29,26,19,22,22,01 ==>>	2
21	19,20,18,16,14,20,19,10,16,12,00 ==>>	2
22	23,21,25,28,25,27,28,19,22,26,01 ==>>	2
23	22,22,20,18,26,20,22,29,26,21,01 ==>>	2
24	26,23,18,26,22,20,25,21,24,22,01 ==>>	2
25	25,22,20,18,26,20,22,29,26,21,01 ==>>	2

5. CONCLUSION

Classification in data mining is wide area which attracts the researchers and exact and effectively sort out the search of information. This paper demonstrates classification methods to propose well behaved carrier for student. Undisciplined and violent student affects their carrier. Classification rules produced by decision tree are famous due to easy interpretation. Various kinds of classifiers are tried for calculation of accuracy as well as performance and well behaved classifier is selected. So, possibility of the student become violent in future prediction is accomplished. Real data from Pimpri Chinchwad College of Engineering in Pune University is used. The performance of C4.5 classifier is measured in terms of correctly classified instances and incorrectly classified instances. This prediction helpful for institution organizes counseling to appropriate student on the basis of evaluation of violence in beginning stages. Various kinds of classifications are utilized as predictive tool within data mining and compared performances. After that the ranking of the student is calculated on the basis of student academic assessment. And then the frequent patterns are generated by using FP growth algorithm.

6. REFERENCES

[1] Miss. Trupti Diwan, Prof. Bharati Dixit, "Analysis of Classification Algorithms for Prediction of Student Failure Using EDM", Fourth Post Graduate Conference, 25th March 2015.

[2] SuhemParack, ZainZahid, Fatima Merchant, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns".

[3] Shreenath Acharya, Madhu N, "Discovery of student's academic patterns using data mining techniques" IJCSE, Vol. 4 No. 06 June 2012.

[4] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao "Predicting Students' Performance Using Id3 and C4.5 Classification Algorithms", IJDKP, Vol.3, No.5, September 2013.

[5] Mohammed M. Abu Tair and Alaa M. El-Halees,"Mining Educational Data to improve Student's performance", JICT, Volume 2 No. 2, February 2012.

[6] Mr. M. N. Quadri1, Dr. N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", GJCST, Vol. 10 Issue 2 (Ver 1.0), April 2010.

[7] Gerben W. Dekker "Predicting students drop out: a case study" 2nd International Conference on Educational Data Mining, April 10, 2009.

[8] M.Sindhuja et al. "Prediction and Analysis of students Behaviour using BARC Algorithm", IJCSE, Vol. 5 No. 06 Jun 2013.

[9] Baker R., "Modeling and understanding students' off-task behavior in intelligent tutoring systems". In Conference on Human Factors in Computing Systems, San Jose, 2007 California, 1059-1068

[10] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013

[11] Carlos Márquez-Vera, Cristabal Romero Morales, and Sebastian Ventura Soto-Predicting School Failure and Dropout by Using Data Mining Techniques", Ieee Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013.

[12] Suhem Parack, Zain Zahid, Fatima Merchant, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns".

[13] Weraporn Jirapanthong,"Classification Model for Selecting Undergraduate Programs" 2009 Eighth International Symposium on Natural Language Processing.

[14] Abdullah Saad Almalaise Alghamdi, **"**Efficient Implementation of FP Growth Algorithm-Data Mining on Medical Data**",** IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.12, December 2011.