International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-5, Issue-1, January 2017
DOI: 10.21276/ijircst.2017.5.1.4

Analyzing Student's Academic Performance Based on Data Mining Approach

S. Kalaivani, B. Priyadharshini, B. Selva Nalini

Abstract— **Career building is the most cherished part of every engineering student. For an engineering graduate it is necessary to have immense knowledge in their domain to get placed in a reputed company. Data Mining is used to gain knowledge, find the hidden information and also this system applies data mining techniques to the academic dataset. The Academic data includes the Internal (CCET 1, CCET2 and CCET3) marks and the Assignment marks. The final semester marks are predicted from the analyzed result of each student. In order to increase the accuracy, this system introduces reweight enhanced boosting algorithm.**

***Keywords*— Data mining, Balanced Boosting, Educational data mining, Re-weight enhanced boosting, Ada-boosting**

I. INTRODUCTION

A. *Overview*

Data mining is extraction of interesting patterns or knowledge from huge amount of data. It is a step in knowledge discovery process. The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

B. *Educational Data Mining*

Educational Data Mining(EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data and using those methods to better understand the educational field entities.

C. *Data Preprocessing*

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

D. *Class Imbalance Problem*

Class imbalance problem exist in most of the real time data set. In imbalanced datasets, the classes having more examples are majority classes and the ones having fewer examples are minority classes. The class imbalance problem typically occurs in a classification problem. When the algorithm is applied only the majority class instances are considered for classification neglecting the minority class. This reduces the overall accuracy of the result which affects the future predictions. Hence this problem, existing in the datasets must be solved for obtaining high accuracy.

A simple example for class imbalance problem. Consider the two class dataset with ratio of two classes is 10:90. We apply two classification algorithms. The first classification algorithm gives overall prediction accuracy as 90% with 13 accuracy of each class in which 0% in first class and 100% in second class. The second classification algorithm gives overall prediction accuracy as 78% with accuracy for each class in which 60% in first class and 80% in second class. Here, the first classification algorithm gives higher overall prediction accuracy than second classification algorithm. But first classification algorithm misclassifies all the instances in class1. So in these cases, algorithms like first classification algorithm cannot be considered as a good classification algorithms.

i) Two class Imbalance

In two class imbalance, there are only two class i.e., Majority class and Minority class. By apply classification algorithms on it, only Majority class instances are correctly classified, whereas Minority class instances are neglected.

ii) Multi class Imbalance

Multi class imbalance is a two class imbalance but in which majority class is easy to identify whereas minority class is difficult to identify.

II. RELATED WORKS

Data mining techniques that have been used to predict students performance and also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data[1] The dataset consist of 203 records and 57 attributes. A classification based on an association rules algorithm is used to build a classifier to help evaluate the student's performance in the programming course. [2] The data set used are admission test scores results, the options for enrolment and some socio-demographic attribute. This paper presents the results of applying an educational data mining approach to model academic attrition (loss of academic status) at the Universidad National de Colombia [3] Data collected from questionnaires which are delivered to student between the 8th and 10th week of courses. The purpose of this study is to suggest a way to support the administration of a HEI by providing new knowledge related to the educational processes using data mining techniques. [4]

The dataset consist of 300 records. This paper focus on mapping students using K-mean Cluster algorithm to reveal the hidden pattern and classifying students based on their demographic.[5] The data used for this study come

Manuscript received January 20, 2017.

Kalaivani. S, Information Technology, Dr Mahalingam College of Engineering and Technology, Pollachi, India, 9487443674.

Priyadharshini. B, Information Technology, Dr Mahalingam College of Engineering and Technology, Pollachi, India, 9892345135

Selva Nalini. B Information Technology, Dr Mahalingam College of Engineering and Technology, Pollachi, India,,9600462446

from students questionnaires distributed in the classes This paper suggest a way to support the administration of a (Higher Educational Institution)HEI by providing new knowledge related to the educational processes using data mining techniques.[6]The dataset consist of 181 records. A classification based on an association rules algorithm is used to build a classifier to help evaluate the student's performance in the programming course. [7] The data used for this study come from the scores of students from past performance assessments This paper proposes an algorithm that predicts the final grade of each student in a class. [8] Some works, which also exclusively use the data from the course itself. First, they collect data in online education or Massive Open Online Course (MOOC) systems such as information about video-watching behaviour, time spent on specific questions or activity in forum.[9].

III. PROBLEM DEFINITION

In the real time the assessment questions provided by the institution are same to all the students, hence their performance doesn't show any improvement , which means if the students are low performing they end up their studies as low performer similarly for high and medium level , their performance are static. Similarly, there is a problem of class imbalance which results in low result of student performance.

To overcome all these issues, here the performances of the students are analyzed from their CCET marks and the assignment marks. According to the analyzed result the assignments and the course outcome are provided to the students. Practicing to these assignments results in increased performance of the students and also the class imbalance problem are overcome.

IV. IMPLEMENTATION

A. Proposed Algorithm: Re-weight Enhanced boosting
Dataset Used: Student Academic Data
Working Principle: Ensemble of classifiers using J48
Input: *D*, a set of *d* class-labelled training tuples;
k, the number of rounds (one classifier is generated per round);
a classification learning scheme.
Output: Confusion Matrix
Method:

STEP 1: Initialize the weight of each tuple in *D* to 1=*d*;
STEP 2: for *i* D 1 to *k* do // for each round:
STEP 3: Use training set *Di* to J48, *Mi*;
STEP 4: Compute *error*. *MI*/, the error rate of *Mi*
STEP 5: Calculate reweight using reweight=Sum of weight/count
STEP 6: For the next following iteration weight is calculated using above formula
STEP 7: Calculate Confusion Matrix
STEP 8: Calculate the Overall Accuracy by using the formula Overall Accuracy =
Truly Predicted Class Labels/Total Number of Class Labels
STEP 9: Calculate the class wise Accuracy by using the formula Class 1 wise
Accuracy = Truly Predicted Class 1 Labels/Total Number of Class 1 Labels

B. Modules

1. Data Preparation
1.1. Data selection
1.2. Data preprocessing
2. Implementation of balanced boosting algorithm
3. Predicting end semester grades

i) Modules Description

ii) Data Selection

Dataset: GRADE SHEET

Table 4.1 Grade sheet

Number of records	1938
Number of attributes	5
List of attributes	Subcode, CCET-1, CCET-2, CCET-3, Semester grade
Maximum value	S
Minimum value	RA

iii) Data Pre-Processing

Data pre-processing is an important step in the data mining process. Data pre-processing, includes cleaning, normalization, transformation, feature extraction and selection, etc. Data transformation is converting of data from one format to other format.

iv) Formula

=IF(A1<20,"RA",
IF(A1<=22,"E", IF(A1<=24,"D",
IF(A1<=28,"C", IF(A1<=32,"B",
IF(A1<=36,‖A‖,"S"))))))

This formula is used to convert marks into grades.

Table 4.2 Pre-Processing

Before Pre-Processing	After Pre-Processing
36-40	S
32-35	A
28-31	B
24-27	C
22-23	D
20-21	E
Below 20	RA

v) Implementation Of Re-weight enhanced Boosting Algorithm

Re-weight enhanced Boosting algorithm is implemented using Java in Netbeans. The algorithms such as J48 and Decision stump algorithms are also implemented in java. These algorithms are analyzed and the accuracy is found efficient in balanced boosting algorithm.

vi) Performance Metrics

Performance metrics is used for evaluating the performance of the classifier used for the classification.

Table 4.3: Confusion Matrix

Actual class / Predictive class	Predicted positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

True Positive Rate

Proportion of positive instances that are correctly classified in majority class. It is also called sensitivity.

$$TPrate = TP/(TP+FN)$$

True Negative Rate

Proportion of negative instances that are correctly classified in minority class. It is also called as specificity.

$$TNrate = TN/(TN+FP)$$

False Positive Rate

Proportion of negative instances that are misclassified in the minority class.

$$FP\ rate = FP/(FP+TN)$$

False Negative Rate

Proportion of positive instances that are misclassified in the majority class.

$$FN\ rate = FN/(TP+FN)$$

Overall Accuracy

Overall accuracy is calculated using all the four performance metrics mentioned above.

$$Acc = (TP+TN)/(TP+TN+FP+FN)$$

V. RESULTS

Adaptive Boosting is a machine earning algorithm which can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier.

A decision stump is a learning model consisting of a one-level decision tree. It is a decision tree with one internal node which is immediately connected to the terminal nodes . A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules. J48 is the java implementation of c4.5 algorithm in data mining.

These algorithms are applied to the dataset and the accuracy has been calculated for final, third and second year as follows:

Final Year:

Table 5.1: Final Year Accuracy Comparison

Algorithm	**Ada boosting (decision stump)**	**Ada boosting(j48)**	**Re-weight enhanced Boosting**
Overall	37.59	58.34	68.42
S	0.00	0.00	100.00
A	0.00	54.11	96.47
B	94.05	59.40	76.23
C	94.00	70.80	41.19
D	0.00	40.42	90.42
E	0.00	29.16	95.83
RA	0.00	0.00	100.00

Third Year:

Table 5.2: Third Year Accuracy Comparison

Algorithm	**Ada boosting (decision stump)**	**Ada boosting(j48)**	**Re-weight enhanced Boosting**
Overall	39.74	58.33	62.50
S	0.00	0.00	100.00
A	0.00	0.00	100.00
B	93.66	79.63	27.14
C	0.00	62.18	76.10
D	0.00	18.80	96.96
E	63.07	76.92	64.61
RA	0.00	6.25	100.00

Second Year:

Table 5.3: Second Year Accuracy Comparison

Algorithm	**Ada boosting (decision stump)**	**Ada boosting(j48)**	**Re-weight enhanced Boosting**
Overall	33.64	60.80	69.5
S	0.00	0.00	100.00
A	0.00	73.77	77.04
B	0.00	64.59	83.22
C	77.25	72.51	50.71
D	56.70	44.32	82.47
E	0.00	63.63	36.36
RA	0.00	25.92	100.00

A. Predicting End Semester Grades

Association rules are used to predict the the final grades of the students using cumulative test mark and assignment marks. Weights are given such that good=3, average=2, poor=1.Hence by calculating for three cumulative exams and assignments the total is obtained as 18. Thus our system predicts the grade by the following rule. The total weights of students above 15 is classified as good performing students,

those above 12 is graded as average performing students and those below 12 is graded as poor students.

B. Discussions

Datasets of different attributes are taken like marks only, marks and subcode only and marks, subcode and rollno. All these datasets are applied in Decision Stump, J48 and Balanced Boosting algorithms. By analyzing the results it is clearly known that comparing Decision Stump and J48, J48 provides more accuracy. While comparing J48 and Balanced Boosting it is identified that Balanced Boosting provides high accuracy.

VI. CONCLUSION

Most of the times the student performance does not show any improvement till the end of their studies. This system applies data mining techniques to the academic dataset. The final semester marks are predicted from the internal marks of each student. When compared to existing algorithms like Adaboost(Decision Stump), Adaboost(J48) proposecd system Re-weight enhanced boosting provides much better accuracy.

ACKNOWLEDGMENT

Kalaivani S , Priyadharshini B, Selva Nalini B are the authors who thank Prabakaran E for the long lasting support

REFERENCES

[1] Amirah Mohamed Shahiria, WahidahHusaina, Nur'aini Abdul Rashida,"A Review on Predicting Student's Performance using Data Mining Techniques", Science Direct,pp. 414 – 422, 2015.

[2] AsmaaElbadrawy, AgoritsaPolyzou, ZhiyunRen, Mackenzie Sweeney, George Karypis, HuzefaRangwala,"Predicting Student Performance Using Personalized Analytics", IEEE , pp. 61-69, April 2016.

[3] Camilo Ernesto LópezGuarín, Elizabeth León Guzmán, and Fabio A. González,"A Model to Predict Low Academic Performanceat a Specific Enrollment Using Data Mining", IEEE RevistaIberoamericana De Tecnologias Del Aprendizaje, vol. 10, pp. 3, August 2015.

[4] GhadaBadra,b*, AfnanAlgobaila, HanadiAlmutairia, ManalAlmuterya, "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department", Science Direct,pp. 80-89, 2016

[5] Harwatia*,ArditaPermataAlfiania, FebrianaAyuWulandari, "Mapping Student's Performance Based on Data Mining Approach", Science Direct,pp. 173 – 177, 2015

[6] ManolisChalaris*,StefanosGritzalis, ManolisMaragoudakis, Cleo Sgouropoulou and AnastasiosTsolakidis,"Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques", Science Direct,pp. 390 – 397,2014

[7] Syed TanveerJishan, Raisul Islam Rashu, NaheenaHaque and Rashedur M Rahman*, "Improving accuracy of students' final grade prediction model using optimal equal widthbinning and synthetic minority over-sampling technique",Springer, pp. 2:1,2015

[8] Yannick Meier, JieXu, OnurAtan, and Mihaela van der Schaar**,"**Predicting Grades", IEEE, pp. 20-29, 2014

[9] C.Romeo, M-I Lopez, J-M Luna and S.Venture, "Classification via clustering for predicting final marks based on student participication in forums", Comput Ed., vol 68, pp 458-472, 2015

Kalaivani S, Pursuing final year B.Tech Information Technology in Dr.Mahalingam College of engineering and Technology, Pollachi and the author of the paper "Analyzing Student's Academic Performance Based On Data Mining Approach**"**

Priyadharshini B, Pursuing final year B.Tech Information Technology in Dr.Mahalingam College of engineering and Technology, Pollachi and the author of the paper "Analyzing Student's Academic Performance Based On Data Mining Approach**"**

Nalini B, Pursuing final year B.Tech Information Technology in Dr.Mahalingam College of engineering and Technology, Pollachi and the author of the paper "Analyzing Student's Academic Performance Based on Data Mining Approach**"**.

