IJCES International Journal of Computer Engineering Science , Volume1 Issue2, November 2011
ISSN : 2250:3439
https://sites.google.com/site/ijcesjournal
http://www.ijces.com/

Envision of student's concert using supervised learning techniques

S. Anupama Kumar[1] and Dr. Vijayalakshmi M.N[2]
[1] Research Scholar, PRIST University, [1]Assistant Professor, Dept of M.C.A.
[2]Associate Professor, Dept of MCA,[1,2] R.V.College of Engineering, Bangalore, India.
[1]kumaranu.0506@gmail.com , [2] mnviju74@gmail.com

Abstract: Educational data mining is an emerging technology concerned with developing methods for exploring the various unique data that exists in the educational settings and uses them to understand the students as well as the domain in which they learn. Educational domain consists of a lot of data related to students, teachers and other learning strategies. Classification algorithms can be used on various educational data to mine the academic records. It can be used to predict student's outcome based on their previous academic performance. The various predictive algorithms like, C4.5, Random tree are applied on student's previous academic results to predict the outcome of the students in the university examination. The prediction would help the tutor in understanding the progress and attitude of the student towards the studies. It would also help them to identify the students who are constantly improving in their studies and help them to achieve a higher percentage. It also helps them to identify the underperformers so that extra effort can be taken to achieve a better result. The algorithms are analyzed based on their accuracy of predicting the result, the recall and the precision values. The accuracy of the algorithm is predicted by comparing the output generated by the algorithm with the original result obtained by the students in the university examination.

KEYWORDS: Educational EDM, Prediction, Decision trees, Recall, Precision.

.

1 Introduction

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. It can be applied on educational data to bring out new information hidden in the data. Enormous data pertaining to student information is available in educational institutions which can be mined to improve the quality of the students as well as educational institutions. EDM is the process of transforming raw data compiled by education systems in useful information that could be used to take informed decisions and answer research questions [2]. The data can be taken from student's learning environments, academic data pertaining to their scores in the examination, grade obtained by them, previous academic achievement etc. These data often has multiple levels of meaningful hierarchy depending on time and sequence. The meaningful context of the data play important role in the study and analysis of the students. The various data mining techniques like classification, clustering and association rule mining can be efficiently used on educational data to bring out new knowledge from it [1]. Classification techniques like decision trees, naïve bayes algorithm, support vector machines etc can be efficiently used on student assessment data to predict their outcome in the university examination using the intermediate marks obtained in the internal examination [6].

Examination and assessment plays a major role in all the stages of a student. The result of a student depends on the marks obtained by the student. Therefore predicting student's result is an important issue for any institution. If the result of a student can be predicted at an early stage, it would help the institution to bring out betterment in the result. From [6] it is clear that decision trees can be used to predict the student's performance using the marks obtained by them in the internal examination. Apart from the internal marks various other parameters can also be used to predict the student's performance in the examination. This paper aims to mine the student's academic data available in the records from their school till fourth semester. The paper aims to predict the result which the student would obtain in the V semester. Classification algorithms like C4.5 and Random Tree are used to mine the records and predict the outcome of the students in the V semester. These algorithms are analyzed using the following parameters.

1. The number of instances predicted correctly by the miner and the algorithm
2. The accuracy of the algorithm is analyzed by comparing it with the original result
3. The recall value of the algorithm
4. The precision value of the algorithm

This paper is organized into Introduction, Application of supervised learning techniques on educational data, Data collection and visualization, Implementation of the decision tree algorithms, result and conclusion.

2 APPLICATIONS OF SUPERVISED LEARNING TECHNIQUES IN EDUCATIONAL DATA MINING

Machine learning is one of the most popular techniques used in the field of data mining. It can be classified as Supervised and Unsupervised learning. In supervised learning the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. In unsupervised learning all the variables are treated in the same way, there is no distinction between explanatory and dependent variables. Predictive modelling is the process by which a model is created or chosen to best predict the probability of an outcome. In many cases the model is chosen on the basis of detection theory and tries to guess the probability of an outcome given a set amount of input data. Classification is a predictive data mining technique which makes predication about values of data using known results found from different data [4]. Predictive models have the specific aim of allowing us to predict the unknown value of a variable of interest given known values of other variables. Predictive modelling can be thought of as learning a mapping from an input set of vector measurements x to a scalar output y [5]. Classification maps data into predefined groups called as classes. It is often referred to as supervised learning because the classes are determined before examining the data. They often describe these classes by looking at the characteristic of data already known to belong to the classes [4]. The various classification techniques which come under supervised learning techniques include decision trees, naïve bayes networks, neural networks etc. They can be used to

- Predict Academic success for students
- Predict the Course Outcomes
- Succeeding the next task
- Meta cognitive skills, habits and motivation

The author in [7] has predicted the students who are at risk by building a predictive model and implementing that model for a group of students in a particular course. This paper concentrates on applying the various decision tree algorithms like C4.5, Random tree in predicting the academic success.

3 DATA COLLECTION AND VISUALIZATION:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples Education system can be either traditional or conventional. In case of higher education, the teaching – learning process is carried out by using traditional way of teaching and using the ICT's for the same. In [8] ,the authors has explained how clustering techniques can be used in an online education system to predict the student's drop out ratio. Behrouz Minaei-Bidgoli and et al [10] have explained how prediction techniques can be used in an virtual environment using LONCAPA system. The Indian education system is more traditional where the student's are expected to be physically present in the class and the examination is conducted in the class only using traditional method. The papers are assessed by the teachers and the scores are recorded. The percentage of marks and grades obtained by the students of a particular course is taken as a data set for the implementation of the algorithms. The data set consists of 16 attributes and 55 instances of a particular batch of students of a course. The details of the data set are given below.

Table 1 : Attributes of the dataset

S.No	Attribute	Information	Category	Information
1	UID	University id	Discrete	55 values
2	X percentage	%of marks	Continuous	-
3	X class	Class obtained	Discrete	3 values
4	XII percentage	%of marks	Continuous	-
5	XII class	Class obtained	Discrete	3 values
6	Degree percentage	%of marks	Continuous	-
7	Degree Class	Class obtained	Discrete	3 values
8	I sem Percentage	%of marks	Continuous	-
9	I sem class	Class obtained	Discrete	3 values
10	II sem Percentage	%of marks	Continuous	-
11	II sem Class	Class obtained	Discrete	3 values
12	III sem percentage	%of marks	Continuous	-
13	III sem class	Class obtained	Discrete	2 values
14	IV sem percentage	% of marks	Continuous	-
15	IV sem class	Class obtained	Discrete	2 values
16	Predicted class	Class expected	Discrete	2 values

The attributes UID consists the university id no the student, the other attributes constitutes the percentage of marks and class obtained by the students in X std, XII std, Degree and four semesters of a particular course. The description of the percentage and the class obtained by the student is described below:

Table 2: Description of Data set

Percentage of marks obtained	Class obtained	Description
40_59	SC	Second class
60_70	FC	First Class
70_100	FCD	First Class with Distinction

These attributes are used to predict the outcome of the student in the fifth semester. Figure 1 explains how the percentage of the students are split from X std to their percentage in the forth semester. From the graph it is clear that there is a growth in the graph from X std percentage to IV semester percentage. The student's percentage has started with 45% in X std where the IV semester percentage begins with 62%. Therefore it is clear that there is increase in the percentage of marks the students are obtaining in the college as compared with schools

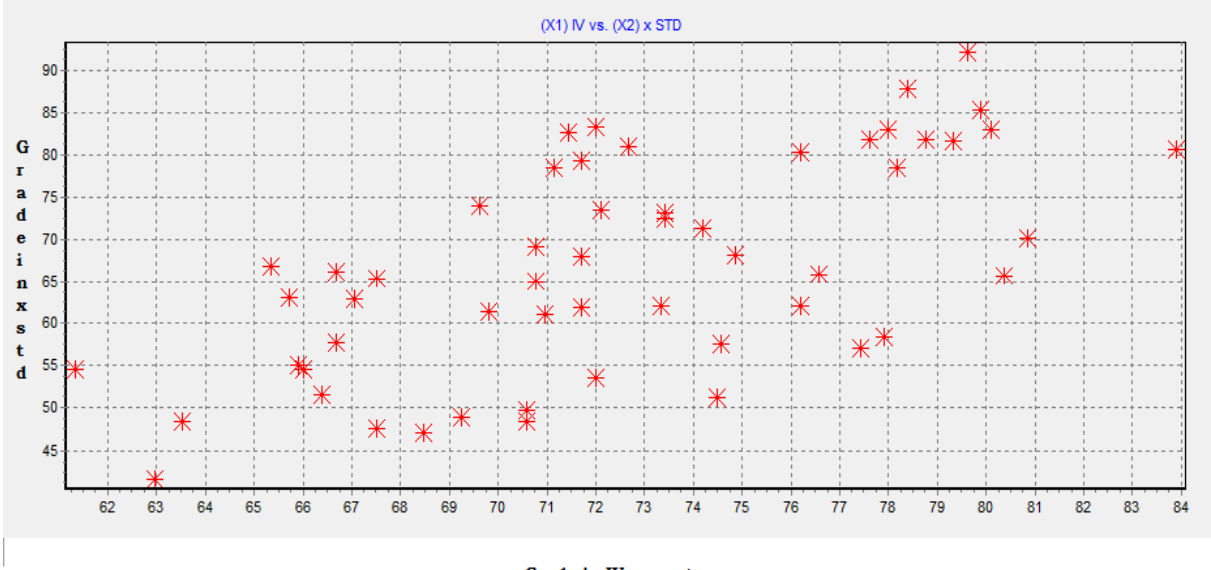

Fig 1: Data visualization of X std percentage v/s. IV semester percentage

4 IMPLEMENTATION OF DECISION TREE ALGORITHM

Supervised learning algorithms can be classified as decision tree algorithms, rule based algorithms, naïve bayes function algorithms etc. In [3] the author has explained how rule based algorithms can be used to predict the behavior of a student in an e learning system. Zlatko J. Kovačić [7] has explained how CART trees can be used as a classification technique to devise a predictive model to identify the at risk students so as to yield a better result. The author of [9] has explained how decision trees can be effectively used to estimate the motivational level of the students using various parameters. Since the decision algorithms can work efficiently on a various types of attributes, they are implemented for the above set of educational data.

4.1 C4.5 Decision Tree algorithm:

C4.5 classification algorithm builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The example has several attributes and belongs to a discrete class. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch being a possible value of the attribute. Prediction class attribute is defined as a target attribute

(discrete value FC/FCD) and all other attributes are constituted as input attributes. The output of the algorithm can be studied in the form of a confusion matrix. Table3
displays the output produced by the C4.5 algorithm for the given data set in the form of a confusion matrix.

Table 3 : Comparison of instances FCD and FC by C4.5

	No. of students in FCD	No. of students in FC	Total No of students
No. of students in FCD	35	1	36
No.of students in FC	3	16	19
Total No of students	38	17	55

From the confusion matrix it is clear that, Out of the 55 instances, 51 instances are correctly predicted and 4 instances are incorrectly predicted by the miner.
The number of instances who are actually in FCD are 38 and predicted are 36. The number of instances who are actually in FC are 17 and predicted are 19.
The rule followed by the algorithm for prediction is given below:
Rule 1: Class iv in [FCD] then prediction class = FCD (92.11%of 38 examples).

Rule 2 : Class iv in [FC] then prediction class = FC (94.12% of 17 examples)

4.2 Random Tree:

The same data set is implemented using the random tree classification technique. Table 4 explains the confusion matrix created by the random tree algorithm. From table 4 it is clear that the output predicted by the miner and the algorithm are same ie all the instances are correctly predicted by the miner as well as the algorithm

Table 4 : Comparison of instances FCD and FC by RnD tree

	No. of students in FCD	No. of students in FC	Total No of students
No. of students in FCD	36	0	36
No. of students in FC	0	19	19
Total No of students	36	19	55

The decision outcomes created by the algorithm for deriving the tree are as follows:

If class iv in [FCD] then If class ii in [FCD] then

If II < 73.3300 then
If III < 69.7150 then

If xiistd < 72.8800 then prediction = FC (100.00 % of 1 examples)

if xiiSTD > = 72.8800 then prediction = FCD (100.00 % of 1 examples)
if III >= 69.7150 then prediction = FCD (100.00 % of 4 examples)

if II >= 73.3300 then prediction = FCD (100.00 % of 29 examples)if class ii in [FC] then

if xii STD < 59.8350 then prediction = FC (100.00 % of 2 examples)

if xii STD >= 59.8350 then prediction = FCD (100.00 % of 1 examples)

if class ii in [SC] then prediction = FCD (0.00 % of 0 examples)

if class iv in [FC] then

if II < 74.2350 then prediction = FC (100.00 % of 16 examples)

if II >= 74.2350 then prediction = FCD (100.00 % of 1 examples)

From the above rules it is clear that the tree works on all the given examples and all the attributes are considered for deriving a confusion matrix

5 Results

The results obtained by the algorithms are analyzed using the following criteria.

1. The number of instances predicted correctly by the miner and the algorithm
2. The accuracy of the algorithm is analyzed by comparing it with the original result
3. The recall and precision value of the algorithm

subject using their roll number easily and give them extra coaching so as to get better results. This will also help the educational institution to bring out quality results from the students. In future, various other attributes like attendance, the marks scored in the previous semester can also be incorporated to predict the student's outcome.

5.1 : Comparison using Instances:

Table5 explains the comparison of the algorithms in terms of the number of instances

Table 5: Comparison of C4.5 and Rnd tree

Algorithm	No. of instances correctly predicted	No. of instances incorrectly predicted
C 4.5	51	4
Random Tree	55	0

C4.5 algorithm has predicted 51 instances correctly out of 55 instances where there is a variation 1 instance predicted as FCD by the miner and predicted as FC by the algorithm. Similarly, 3 instances are predicted FC by the miner and FCD by the algorithm. Random tree algorithm has not brought out any change in the prediction by the miner as well as the algorithm.

5.2 : Comparison using Accuracy:

The accuracy of the algorithms is verified using the result obtained from the university. Table5 shows the comparison of the results obtained by the algorithms and the university results.

Table 6: Comparison of results of C4.5 and Random tree

Algorithm	No. of instances in FCD	No. of instances in FC
C 4.5	38	17
Random Tree	36	19
University	40	15

From table6 it is clear that C4.5 algorithm could predict better than the Random tree since the results are 95% accurate towards the results obtained from the university.

5.3 Comparison using Recall and precision:

Any classification algorithm holding a lower precision and recall value are considered better than the other which has a higher precision and recall values. It can be calculated using the true and false positive as well as negative values obtained from the confusion matrix. Table6 shows the Recall and -precision obtained by the c4.5 algorithfor the target attribute.

Table 7: Recall and Precision values by C4.5

Value	Recall	Precision
FCD	0.9722	0.0789
FC	0.8421	0.588

Table7 gives the recall and precision values of both FCD and FC produced by the C4.5 algorithm. It is clear that the values are less than 1. The lesser the values indicates that the algorithm is more accurate.

The table8 gives the recall and precision values obtained by random tree algorithm.

Table 8: Recall and Precision values by Random tree

Value	Recall	Precision
FCD	1.00	0.0
FC	1.00	0.0

Since the algorithm could not produce any incorrect instances the true negative and false negative values are nil. Therefore the precision values of the attributes are zero. Even though the precision values are zero, the recall value is positive. Therefore Random tree algorithm is not more effective than the C4.5 algorithm.

6. Conclusion:

The application of classification algorithms on educational data can effectively help the tutor to understand the student's interest towards academics and can bring improvement in their result. The percentage of marks and the class obtained by the student from his school till he reaches a higher education is definitely an efficient attribute in predicting the outcome of a student in the form of a grade. C4.5 can be used to accurately predict the outcome of the student in the university exam. This is proved true for the given data set. The analysis of the algorithm using the recall and precision values clearly states that the C4.5 algorithm performs better than the random tree algorithm. The accuracy of the algorithm is also proved by comparing it with the original result. This also states that C4.5 algorithm is 5% better than the random tree for the given data set. Therefore for a given data set C4.5 algorithm predicts the student's outcome in the examination better than the random tree algorithm.

References

1. S.Anupama Kumar, Dr.Vijayalakshmi M.N, "A Novel Approach in Data Mining Techniques for Educational Data", Proc 2011 3rd Internal Conference on Machine Learning and Computing (ICMLC 2011) , Singapore, Feb 2011,ISBN 978-1-4244-9252-7,pp V4-152-154.

2. Cecily Heiner, Ryan Baker y Kalina Yacef, -Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems Jhongli, Taiwan.,2006.

3. Félix Castro & Àngela Nebot and el al ,"Extraction of Logical rules to predict student's behaviour", Proceedings of IASTED International conference on web education, pp 164-170

4. Margret H. Dunham, "Data Mining: Introductory and advance topic".

.5. David Hand, Heikki, Mannil Padraic smyth,"Principles of Data Mining" PHI

- 6. S.Anupama Kumar and et al , "Prediction of the student's recital using classification Technique ", IFRSA's International Journal Of Computing,Vol1,issue 3,July. ISSN (Online) : 2230-9039 ,ISSN: 2231-2153, 2011 pp 305-309

7. Zlatko J. Kovačić and et al, "Predictive working tool for at risk students", Creative Commons 3.0 New Zealand Attribution Non-commercial Share Alike Licence (BY-NC-SA).

8. Tuomas Tanner and Hannu Toivonen ," Predicting and preventing student failure – using the *k*-nearest neighbour method to predict student performance in an online course environment",
University of Helsinki.

9. Mihaela Cocea & Stephan Weibelzahl, Can Log Files Analysis Estimate Learners' Level of Motivation?, Proceedings of Lernen,Wissensentdeckung-Adaptivitat pp 32-35, Germany.

10. Behrouz Minaei-Bidgoli and et al ,Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System Lon-Capa", 33rd ASEE/IEEE Frontiers in Education Conference.

11. Dille, B., & Mezack, M.: Identifying predictors of high risk among community college telecourse students. The American Journal of Distance Education, (1991) 5(1), 24-35.