EFFICIENCY OF DECISION TREES IN PREDICTING STUDENT'S ACADEMIC PERFORMANCE

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 used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees, rule mining, Bayesian network etc can be applied on the educational data for predicting the students behavior, performance in examination etc. This prediction will help the tutors to identify the weak students and help them to score better marks. The C4.5 decision tree algorithm is applied on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass. The result is given to the tutor and steps were taken to improve the performance of the students who were predicted to fail. After the declaration of the results in the final examination the marks obtained by the students are fed into the system and the results were analyzed. The comparative analysis of the results states that the prediction has helped the weaker students to improve and brought out betterment in the result. To analyse the accuracy of the algorithm, it is compared with ID3 algorithm and found to be more efficient in terms of the accurately predicting the outcome of the student and time taken to derive the tree.

KEYWORDS

Assessment, Prediction, Educational data mining, Decision tree, C4.5algorithm, ID3 algorithm

1. INTRODUCTION

Data mining concepts and methods can be applied in various fields like marketing, stock market, real estate, customer relationship management, engineering, medicine, web mining etc. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. 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 [1].The various techniques of data mining like classification, clustering and rule mining can be applied to bring out various hidden knowledge from the educational data.

D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335–343, 2011.
 DOI: 10.5121/csit.2011.1230

Prediction can be classified into: classification, regression, and density estimation. In classification, the predicted variable is a binary or categorical variable. Some popular classification methods include decision trees, logistic regression and support vector machines. In regression, the predicted variable is a continuous variable. Some popular regression methods within educational data mining include linear regression, neural networks, and support vector machine regression. Classification techniques like decision trees, Bayesian networks etc can be used to predict the student's behavior in an educational environment, his interest towards a subject or his outcome in the examination.

Examination plays a vital role in any student's life. The marks obtained by the student in the examination decide his future. Therefore it becomes essential for any tutor to predict whether the student will pass or fail in the examination. If the prediction says that a student tends to fail in the examination prior to the examination then extra efforts can be taken to improve his studies and help him to pass the examination.

This paper is an extension of [13] where the result of the students of I semester MCA are predicted depending upon their performance in the internal examination. We have used C4.5 (J48 in WEKA) to do the prediction analysis. The outcome of the internal marks is used in this paper for finding the efficiency of the algorithm towards educational data and the accuracy of predicting the result. This paper analyses the accuracy of the algorithm in the following ways

- Comparing the result of the tree with the original marks obtained by the student in the university examination
- Comparing C4.5 algorithm with ID3 algorithm in terms of the efficiency in building the tree and time taken to build the tree.

The paper is divided into the following sections.
Section II describes the background investigation, Section III describes the data collection, IV explain the methodology, V explains the findings of the research and VI consists of the conclusion and future
Enhancements.

2. BACKGROUND INVESTIGATION

Predicting the academic outcome of a student needs lots of parameters to be considered. Data pertaining to student's background knowledge about the subject, the proficiency in attending a question, the ability to complete the examination in time etc will also play a role in predicting his performance. M.N. Quadri and Dr. N.V. Kalyankar [3] have predicted student's academic performance using the CGPA grade system where the data set comprised of the students gender, his parental education details, his financial background etc. In [2] the author has explored the various variables to predict the students who are at risk to fail in the exam. The solution strongly suggests that the previous academic result strongly plays a major role in predicting their current outcome. In accordance with [13] **,** the marks obtained by the students during the internal examination will play a vital role in predicting the outcome of the student in the main examination. The internal marks for the subjects MCA11, MCA12, MCA13, MCA14, MCA15 for a maximum of 100 marks and a result of Pass/Fail depending upon a minimum of 50 marks from each subject is fed as input and a decision tree is obtained using C4.5 (J48 in WEKA) .The

output should compared with the original marks received and result obtained by the student in the university examination.

3. DATA COLLECTION

The internal marks obtained by the students of I semester M.C.A has been considered as a source of data in [13] and a decision tree was drawn using the same. A slight modification has been done in defining the nominal values for the purpose of analyzing the accuracy in this paper. Here the nominal values have been categorized as (0_44) where the students are predicted as Fail, (45_54) where the students are considered to be on border line where they may pass or fail and (54_100) where the students are sure to pass. The results of I semester MCA declared by the university is the major source of data in this paper . The declared result consists of a university seat no in the alphanumeric form , which is the unique identifier and marks obtained (internal marks obtained out of 50, external marks obtained out of 100, total out of 150) in five subjects in the form of integers and a result field (containing pass/fail) in the form of string values. Among these data, the internal marks obtained by the student are already used in [paper] and a decision tree is obtained accordingly. For the purpose of research, the external marks (obtained out of 100) are considered. The marks are converted into nominal values according to the following condition:

1.(0_39) indicates a fail in the result of the student

2.(40_100) indicates a pass in the result of the student.

The obtained data is preprocessed according to the need of the system. The unique identifier is removed and the integer values are then converted into nominal values and stored in the .CSV format. It is then converted into the .ARFF format so that it is accessible in WEKA

4. METHODOLOGY

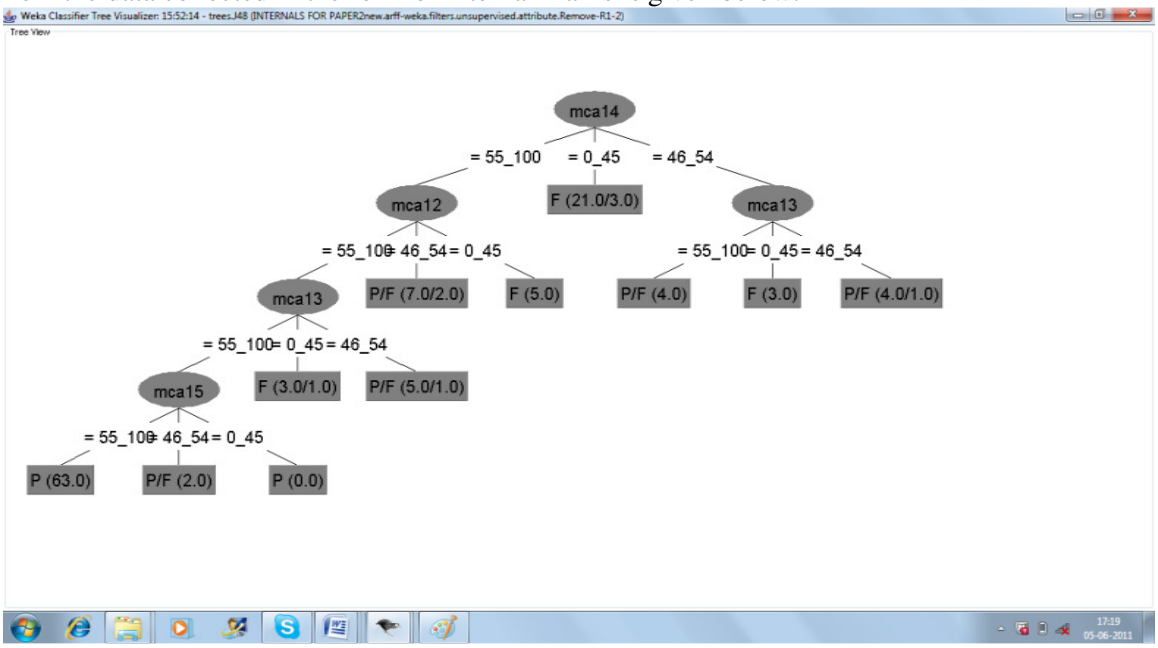
A decision tree depicts rules for dividing data into groups. J48 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs.

At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The J48 algorithm then recurs on the smaller sub lists.

4.1. Implementation of J48 Algorithm on Internal Marks

The internal marks obtained by the students of I semester MCA has been used as a source of data in paper [13] so as to predict the outcome of the student in the university exam. However a slight

modification has been done in the same data for a better prediction. The resultant tree obtained from the data collected in the form of internal marks is given below:



Decision tree 1: Tree obtained from internal marks

The following observations can be made using the above mentioned decision tree:

1. Out of the 5 subject attributes, MCA11 was not considered in the tree, because the instance corresponding the result pass was very low.

2. The other attributes were combined to form a pruned tree.

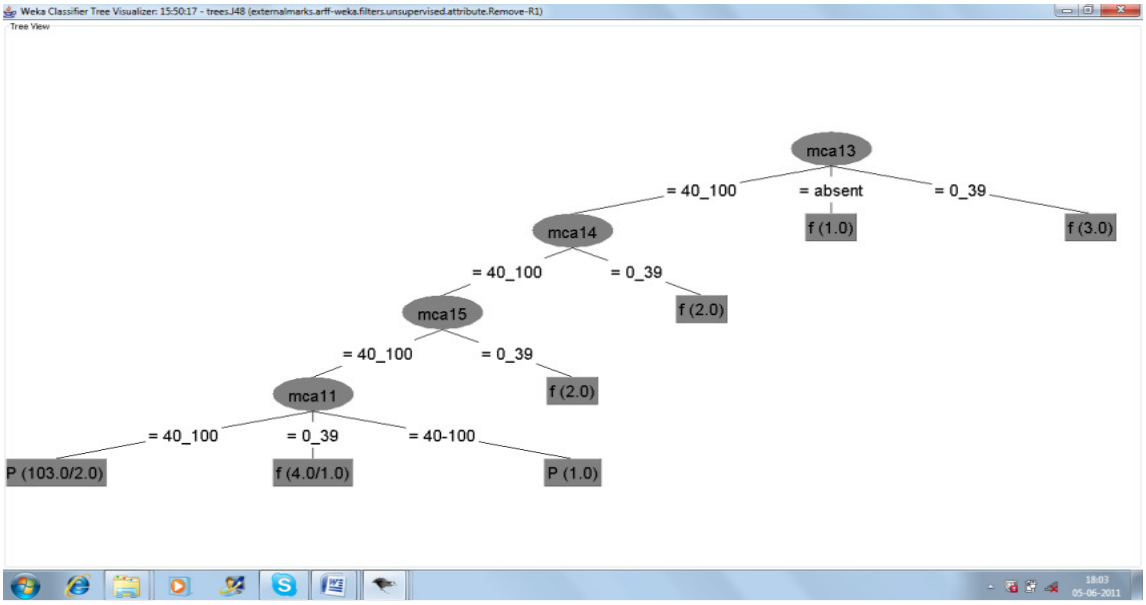
3. Instances of MCA13 and MCA14 share almost equal number of failure students out of which MCA 14 has been considered as the root of the tree since the subject was holding maximum number of students in the range of (0_44) . Beyond this MCA12 with next less number of failures has been taken a leaf and so on.

4. The total number of instances considered for deriving the tree is 117.

The main aim for deriving such a tree is to improve the performance of the students and bring out better results from them. The above derived predictions are given to the tutors and are advised to give extra coaching to the students who were in the category of Pass/Fail and Fail.

4.2. Implementation of J48 Algorithm on External Marks

The accuracy of the above result is now compared with the original result declared by the university in the month of March'11. The original result is then converted to the nominal form and a decision tree is drawn using the WEKA J48 algorithm. The decision tree obtained from the data is given below



Decision tree 2: Tree obtained from External marks

From the tree it is clear that there is a change in the result obtained by the student in the university examination. The following observations are made from the tree:

1. The subject MCA12 was not considered to form a tree since the number of failures was very less in the subject.

2. The subject MCA13 which has been considered as root, it has got three distinct leaf nodes where the node has depicted the student was absent for the examination. Therefore it is clear that the system is accepting a string value also.

3. The subject with more failures' is taken in the root and the leaves constitute the failures less than the root.

4.3 Comparison of Prediction analysis of Internal and External Marks

From the results obtained from the J48 algorithm, table 1 gives an overview of the prediction analysis made using the internal marks and the original result obtained by the students

	Total No. of Instances	Instances Classified Correctly	Instances classified Incorrectly
Result Obtained from Internal Marks	117	104	13
Result Obtained from External Marks	116	107	9

Table 1: Comparison of Insances for internal and external marks

The algorithm has classified the students as pass/fail for both the correct and incorrect instances. In the case of internal marks, out of the 117 instances, 104 instances are classified as correct and 13 instances have been incorrectly classified. The table 2 describes the confusion matrix achieved through the instances.

No. of Instances predicted Pass	No. of Instances predicted Pass/Fail	No. of Instances predicted Fail
63	1	1
2	3	16
0	25	6

Table2: Confusion Matrix obtained for Internal Marks

In case of External marks students cannot fall under the category of pass/fail since the result is declared by the university and it is compared with the prediction made. From the table 3 it is clear that out of the 104 correctly identified instances, 65 have been predicted as pass, 29 instances as fail and 23 instances can either be pass or fail. These instances are practically important and the tutors are advised to concentrate more on the Pass/fail and fail instances.

Instances classified as Pass	Instances classified as Fail
102	8
1	5

Table 3: Confusion Matrix obtained for External Marks

From table 3it is clear that out of 107 correctly classified instances, 102 students have passed and 5 students have failed. Out of 9 incorrect instances 1student passed and 5 students failed. Out of the incorrect instances one instance belongs to the student who has been marked ab ie ABSENT in the examination. To analyze the accuracy of the algorithm, the results obtained from both the internal and external marks are compared. The following inferences are made from the results obtained:

1. The students who have been predicted to be passing have been declared pass in the university exam also.

2. The students who were predicted to be pass/fail in the decision tree 1 were declared pass in the university exam.

3. Out of the 28 students predicted to be Fail, 13 students have actually failed and 15 other students have improved their studies and passed in the examination.

From the above inferences it is clear that the prediction algorithm has helped the tutors to improve the performance of the students.

5. COMPARISON OF J48 AND ID3 ALGORITHM

The main aim of the prediction analysis is to improve the academic performance of the students. The J48 prediction algorithm is analyzed using the following methods.

1. The accuracy of the algorithm is measured using the comparison of the internal and external marks obtained by the students in the university examination.

2. The efficiency of the algorithm is measured by comparing the J48 algorithm with ID3 algorithm.

The data received from the university result is fed into the ID3 algorithm for analyzing the efficiency of the J48 algorithm. The algorithm is analysed in the following terms:

1. The number of instances predicted as Pass/Fail

2. The time taken to derive the tree

The confusion matrix obtained by the ID3 algorithm is given below:

Instances classified as Pass	Instances classified as Fail
101	1
2	10

Table 4: Confusion Matrix obtained for External Marks from ID3

By keeping all the instances common, the above table clearly specifies that the number of instances declared pass is equal in both the algorithms and the number of instances declared fail is not classified accurately. The data differes by 2 students out of which one is marked absent and the other one is unpredictable. Therefore it is clear that J48 algorithm is more accurate than ID3 algorithm. Table 4 gives the comparative analysis of both the algorithms

Algorithm	ID3	J48(C4.5)
Instances classified as Pass	103	103
Instances classified as Fail	12	13
Time Taken	0.02 seconds	0 seconds

Table 5: Comparison of ID3 and J48 algorithms

6. CONCLUSION

The various data mining techniques can be effectively implemented on educational data. From the above results it is clear that classification techniques can be applied on educational data for predicting the student's outcome and improve their results. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the tutor to identify the weak students and improve their Performance. The analysis of the result declared from the university is a proof for the same. Since the application of data mining brings a lot of advantages in higher learning institution, these techniques can be applied in the other areas of education to optimize the resources, to predict the retainment of faculties in the institution, to predict the number of students who are likely to get a placement, to predict the feed back of the tutor etc.

REFERENCES

[1] 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.

[2] Zlatko J. Kovačić, John Steven Green, Predictive working tool for early identification of 'at risk' students , Newzealand

[3] M. N. Quadri1 and Dr. N.V. Kalyanka- Drop Out Feature of Student Data for Academic Performance Using Decision Tree, Global Journal of Computer Science and Technology Vol. 10 Issue 2 (Ver 1.0), April 2010.

[4] Cristóbal Romero and etel, Computer Science Department, Córdoba University, Spain,"Data Mining Algorithms to Classify Students"

[5] Zlatko J. Kovačić, Associate Professor, John Steven Green, Senior Lecturer, School of Information and Social Sciences, Open Polytechnic, NewZealand"Predictive *working tool for early identification of 'at risk' students"*, published in Creative Commons 3.0 New Zealand Attribution Non-commercial Share Alike Licence (BY-NC-SA).

[6] E.Chandra and K.Nandhini, "Predicting Student Performance using Classification Techniques", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India,p.no83-87.

[7] S. B. Kotsiantis and etel , ."Efficiency of machine learning techniques in predicting students' performance in distance learning systems",proc, Recent advances in mechanics and related fields,p.no 297-305.

[8]. Mykola Pechenizkiy and etel, "Mining the Student Assessment Data: Lessons Drawn from a Small Scale" Case Study"

[9] S.Anupama Kumar, Dr.M.N.Vijayalakshmi,"A Novel Approach in Data Mining Techniques for Educational Data" , Proc 2011 3rd International Conference on Machine Learning and Computing" (ICMLC 2011) , Singapore, 26th-28th Feb 2011,pp V4-152-154.

[10]. R. Messeguer and et al, Department of Computer Architecture, Technical University of Catalonia , Spain,"Group Prediction in Collaborative Learning".

[11] Alaa el-halees ,mining students data to analyze learning behavior:,a case study, Department of Computer Science, Islamic University of Gaza P.O.Box 108 Gaza, Palestine

[12] Dr.Varun Kumar, Anupama Chanda, An Empirical Study of the Applications of Data Mining Techniques in Higher Education, (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 2, No.3, March 2011

[13] S.Anupama Kumar, Dr.Vijayalakshmi M.N.,"Prediction of the students recital using classification Technique ", IFRSA's International journal of computing (IIJC) , Volume 1, Issue 3 July 2011,pp305-309.

[14] Shaeela Ayesha and et al," Data Mining Model for Higher Education System", European Journal of Scientific Research, Vol.43 No.1 (2010), pp.24-29

Authors Profile

Ms.S.Anupama Kumar has 12 years of teaching experience. She has completed her Master of Philosophy from Alagappa University and her Masters from Bharathidasan University. She has published research papers in national and international conferences. She also has a publication in internal journal to her credit. Her research interests are in the area of data mining and artificial intelligence. She is a member of IAENG and IACSIT.

Dr. Vijayalakshmi M.N. had completed her PhD from Mother Teresa Women's university, Kodaikanal in 2010. She has 12 years of teaching experience and 5 years of Research experience. She is a recognised research guide in VTU and Prist University. She has published many research papers in the national and international conferences and journals. She has got many research projects to her credit funded by different agencies. Her research interests are Pattern recognition, data mining , neural networks, Image Processing. She is a life member of ISTE , CSI, IACSIT.