Mining Previous Marks Data to Predict Students Performance in Their Final Year Examinations

G. Naga Raja Prasad[1] Dr. A. Vinaya Babu[2]

1 Associate professor, Dept. of MCA, CBIT, Gandipet, Hyderabad, E-mail:gnrp@cbit.ac.in
2 Professor, Dept. of CSE, JNTU – H, Hyderabad.

ABSTRACT

Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set/ data warehouse. Higher education faces a new era as a result of changes in the way people view colleges and universities. Expectations for better performance in terms of teaching and producing competent college graduates are increasing. Educational data mining is used to study the student data available in the university data base and bring out the useful information / knowledge from it. Classification methods like decision trees, rule mining, Bayesian network etc., can be applied on the student data for predicting the students behavior, performance in examination etc. This prediction will help the tutors to identify the weak / failed students and help them to pass / score better marks. The C4.5 decision tree algorithm is applied on student's first and second year 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 professor 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 failed / weaker students to improve their performance and brought out betterment in the result. To analyze 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.

1.0 INTRODUCTION

Data Mining is a process of extracting previously unknown, valid, potentional useful and hidden patterns from large data sets. The mount of data stored in educational databases is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Data mining software allow the users to analyze data from different dimensions categorize it and a summarized the relationships, identified during the mining process.

Different data mining techniques are used in various fields of life such as medicine, statistical analysis, engineering, education, banking, marketing, sale, etc. Cluster analysis used to segment a large set of data into subsets called clusters. Each cluster is a collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters.

1.1 Data mining in Higher Education System

Education is an essential element for the betterment and progress of a country. It enables the people of a country civilized and well mannered. Mining in educational environment is called Educational Data Mining, concern with developing new methods to discover knowledge from educational databases(Galit, 2007) (Erdogan and Timor 2005), in order to analyze students trends and behaviors toward education(Alaa el-Halees , 2009). Lack of deep and enough knowledge in higher educational system may prevent system management to achieve quality objectives, data mining methodology can help bridging this knowledge gaps in higher education system.

2. Related Work

Data mining is an emerging methodology used in educational field to enhance our understanding

of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students (Alaa el-Halees 2009).

(Kifaya, 2009) K-means clustering is a widely used method that is easy and quite simple to understand. Cluster analysis describes the similarity between different cases by calculating the distance. These cases are divided into different clusters due to their similarity.

In Galit, 2007 gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

(Han and Kamber, 2006) explained that k-means is a well known clustering algorithm tends to uncover relations among variables already presented in dataset.

(Erdogan and Timor 2005) used educational data mining to identify and enhance educational process which can improve their decision making process. Finally (Henrik ,2001) concluded that clustering was effective in finding hidden relationships and associations between different categories of students.

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 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 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 previous year examination will play a vital role in predicting the outcome of the student in the final year examination. The marks for the subjects MCA I year 10 theory subjects and MCA II year 10 theory subjects for a maximum of 100 marks and a result of Pass/Fail depending upon a minimum of 50 marks from each subject is field 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 Ist and 2nd year marks obtained by the students of M.C.A has been considered as a source of data in 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 to 44) where the students are predicted as Fail,
- (45 to 54) where the students are considered to be on border line where they may pass or fail and
- (54 to 100) where the students are sure to pass.

The results of MCA declared by the university is the major source of data in this paper . The declared result consists of a university hall ticket no in the alphanumeric form , which is the unique identifier and marks obtained 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 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:

- (0 to 39) indicates a fail in the result of the student
- (40 to 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

4.0 METHODOLOGY

C4.5 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 of already classified samples. Each sample is a vector where represent attributes or features of the sample. The training data is augmented with a vector where represent the class to which each sample belongs.

At each node of the tree, C4.5 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 C4.5 algorithm then recursive on the smaller sublists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Pseudocode

In pseudocode, the general algorithm for building decision trees is:

- Check for base cases
- For each attribute
- Find the normalized information gain from splitting on
- Let a_best be the attribute with the highest normalized information gain
- Create a decision node that splits on a_best
- Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node

4.1. Implementation of J48 Algorithm

The following observations can be made using the above mentioned decision tree:
1. The total number of subject attributes are 20.
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 58.
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.

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 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.

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 performance 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. This
work may improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. For future work, we hope to refine our technique in order to get more valuable and accurate outputs, useful for instructors to improve the students learning outcomes. Some different software may be utilized while at the same time various factors will be used.

REFERENCES

[1] Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.
[2] Behrouz.et.al., (2003) Predicting Student Performance: An Application of Data Mining Methods With The Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO.
[3] Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687
[4] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)
[5] Galit.et.al (2007)Examining online learning processes based on log files analysis: a case study. Research, Refelection and Innovations in Integrating ICT in Education.
[6] Henrik (2001) Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery: © 2001
[7] Han,J. and Kamber, M., (2006) "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.
[8] Kifaya(2009) Mining student evaluation using associative classification and clustering. Communications of the IBIMA vol. 11 IISN 1943-7765.
[9] ZhaoHui. Maclennan.J, (2005). Data Mining with SQL Server 2005 Wihely Publishing, Inc.