I.J. Education and Management Engineering, **2016, 2, 24-30**
Published Online March 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2016.02.03

Available online at http://www.mecs-press.net/ijeme

Edifice an Educational Framework using Educational Data Mining and Visual Analytics

Dr S Anupama Kumar

Associate Professor, Department of MCA, R V College of Engineering, Bangalore, India

Abstract

Educational Data Mining and Visual analytics are two emerging trends in the industry that plays a major role in bringing out changes in the educational institutions. This paper discusses about building an educational framework that suits the higher education in India using the above mentioned technologies. Educational data mining comprises of various technologies and tasks which can applied on educational data to bring out useful information. In this research work, a data ware house is built to store the student data, two data mining tasks classification and association rule mining are applied over the student data set to analyse their performance in the examination. Decision tree algorithm is used to predict the course and program outcome. Association mining is used to analyze the outcome and understand technical capability of the students. The algorithms were found very accurate in predicting and analyzing the performance. Visual analytics is used in the framework to depict the analysis of the student's performance.

Index Terms: Education data mining, Classification, Association mining, Visual analytics.

1. Introduction

In recent years, universities are operating very complex and highly competitive environment, the main challenge of modern universities being to analyze the student performance and identify their uniqueness to build a strategy for further development. Indian education system is built using 10+2+3/10+2+4 depending on the graduation the student selects to study. Students who wish to study higher education enroll themselves in colleges / institutions affiliated to a recognized university by the University Grants Commission. Today one of the biggest challenges, the educational institutions face, is the explosive growth of educational data and to use this data to improve the quality of managerial decisions to deliver quality education. This system hopes to improve the quality of education by analyzing the data and discover the factors that affect the academic result so as to increase success chances of students. Student data sets play a vital role in predicting the performance of the student in near future and the analysis of the data helps the institution to understand the skills of the student.

This paper deals building a data warehouse to store the data, apply data mining and visual analytics to the

* Corresponding author.
E-mail address:

data set to forecast the student performance and bring out useful information from that to help the institutions to take meaningful decision. The major challenge of the work lies in identifying the parameters for building the framework. A data warehouse is built to store the student data retrieved from various sources. Data mining techniques like Decision tree is used to predict the student performance and association rule mining is used to analyze the student performance. Learning analytics is applied over the output of the performance and new knowledge has been extracted from the same. The paper is organized as: section I gives a brief introduction about the work, section II explains the background work, section III describing the methodology adopted and section IV gives an overview of the results and discussions.

2. Background Work

Educational research can be carried out using statistics, psychology, psychometrics and cognitive psychology [1]. Statistics deals with the collection, interpretation and presentation of the historic data. Statistics deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments [2]. Psychometrics is the quantitative branch of education which when combined with EDM brings out new knowledge on the student data. The concept of Psychometrics has been used since nineteenth century by Sir Francis Galton, who focussed on measuring latent quantities of knowledge and ability in the human mind (Pearson, 1914). Cognitive Psychology involves the study of attention, perception, memory, language, and learning.

The author in [3] has explained the various tools and techniques available in EDM which can be used to analyze the student datasets to meet the educational objectives and help the various stakeholders to make effective decisions. Igor and et al [4] explains the architecture for building a data warehouse to store student data and how effectively ETL tools can be used to extract the data for better understanding. In [5], the author has explored the various socio-demo graphic variables as a data source and implemented classification techniques to predict the performance of the students at an early stage. He concluded that the data mining techniques were efficient to predict the performance of the students. He also suggested that including the academic performance of the students would yield better accuracy along with the attributes he has used. In data mining classification models are build models for the prediction of the class of an object on the basis of its attributes [6]. Classification techniques were used to predict student performance by using features extracted from log data and marks obtained in the final exam and the prediction was found more accurate than other techniques [7]. It helps the tutors to identify students at risk and help the students to prepare well for the final exam. A decision tree is a set of conditions organized in a hierarchical structure [6]. The different classification techniques have been explored by Cristobal Romero and et al [8] to predict the student's performance in the exam and concluded that the decision tree can be directly transformed into a set of IF-THEN rules that are one of the most popular forms of knowledge representation, due to their simplicity and comprehensibility. The different classification techniques that can be used to predict the student performance and concluded that classification techniques are powerful in predicting the same has been discussed by the authors in [9]. When student mistakes are recorded, association rules algorithms can be used to find mistakes often associated together [10]. The potential use of education data mining using association rule mining algorithm in enhancing the quality and predicting students' performances in university result [6]. The analysis revealed that student's university performance is dependent on Unit test, Assignment, Attendance and graduation percentage. The student's performance level can be improved in university result by identifying students who are poor unit Test, Attendance, Assignment and graduation and giving them additional guidance to improve the university result.

3. Methodology and Implementation

The project work has been carried out in various phases viz building a data ware house by collecting and preprocessing the data, implementing decision tree algorithm to predict the course outcome, apply association rule mining algorithm to interpret the performance and visualizing the outcomes using visual analytics for

better understanding. The following figure 3.1 gives the integrated framework built for the higher education framework.

Phase 1: Building a Data Warehouse

A Student data set comprising various personal and academic details of the students of a particular course is collected for the purpose of research. These data are available in an unstructured format along with noise and disturbance. Therefore it is transformed into the desired format for further work. An ETL tool Apatar is used to extract, transfer and load the data. The attributes necessary for predicting the course outcome and analysing the same are identified during this phase. They are structured, formatted and grouped together as per the needs. The transformed data is loaded into the system for further analysis.

Phase 2: Data Pre-processing

Data pre-processing prepares raw data for further processing. Data pre-processing contains two processes such as data cleaning and data transforming. The data pre-processing removes the irrelevant data and identify the required data to predict the course outcome. It involves data cleaning, data transformation and data reduction. The data cleaning is necessary because data in real world is dirty, incomplete therefore it is necessary to remove the null values and noise and also replace the missing values with zeros. The data transformation is the convert the dataset from one format to other format [.xls -> .csv->. arff]. The system is desgined to pre-process the data set dynamically whenever a data is entered so as to ensure integrity. The figure 2.2 shows the pre-processed data set which is ready for implementation.

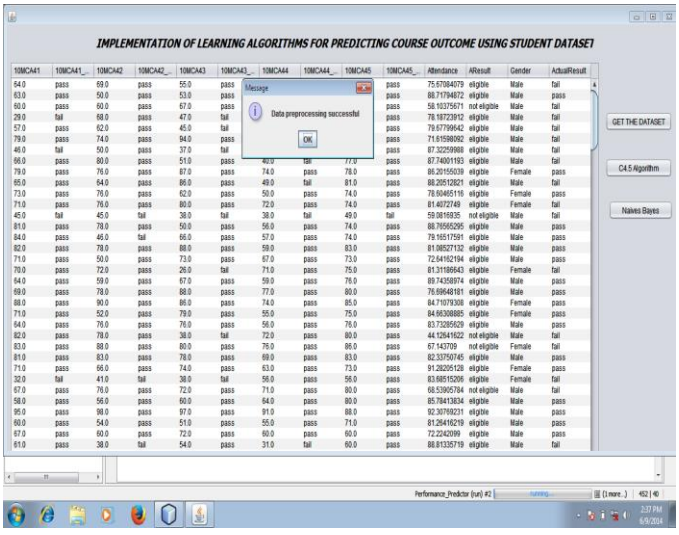

Fig.3.1. Pre Processed Data Set

Phase 3: Prediction of Course Outcome

The outcome of a particular course should be able to define what the student would achieve at the completion of the course. The outcome of any program appears in the form of course description or content. The description establishes the parameters of the program or course and defines the broad scope of knowledge, skills, and/or values that a student will experience. To predict the course outcome of a particular course, the

academic details of the students including the Usn No, name, the attendance percentage of students, internal marks of the students in eight subjects, the expected result (Pass/Fail) are used. Choosing the right algorithm to predict the course outcome is one of the challenges in this research work. In [11], the author described the results of a study conducted in 2000 which aimed at finding weak students and involving them in additional courses for advanced support by extracting association rules from data. M.Ramaswami [12] and et al have developed a predictive model for identifying slow learners among school students considering the CHAID classification algorithm and found the accuracy of the prediction to be 44.69%. The author in [13] have used Bayesian classifier to predict the student performance and used the same to predict the graduate employability statistics for Malaysian government. The author in [14] has identified ethnicity, course level, secondary school, age, course programme and course block as important factors to identify successful students in a course. In this work, decision trees are used to predict the course outcome of the students. The attributes are fed to the C4.5 decision tree algorithm and the output is depicted as a tree along with a confusion matrix explaining the true and false positive, negative values of the target and the time taken to implement the algorithm. The following figure 3.1 shows the pictorial representation of the course outcome.

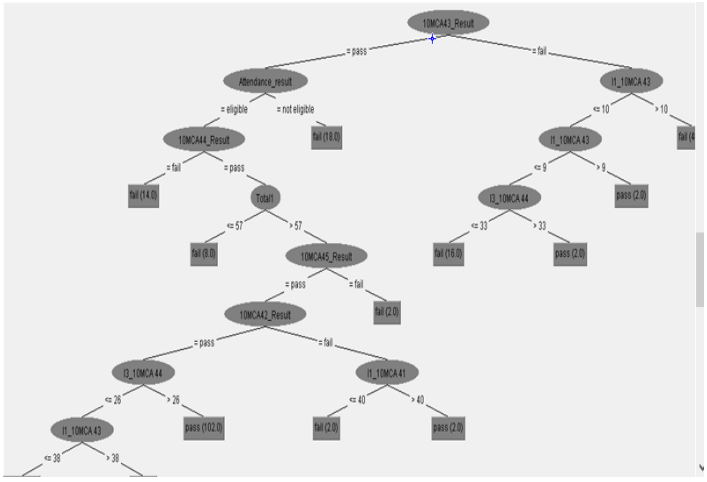

Fig.3.2. Decision tree of Prediction of Course Outcome

For the given data set, for a sample of 240 records, 206 records were predicted correctly as true positive value, 14 records were predicted as false negative, 10 records were predicted incorrectly as true negative and 10 records were predicted incorrectly as false positive. The following results were observed from the tree.

- The number of students whose attendance was eligible, and who were predicted as Pass in five or more subjects were predicted as Pass for the course otherwise they were predicted Fail.
- The students whose attendance was Not eligible were predicted as Fail only
- The number of students whose attendance was eligible, but predicted fail in three or more subjects were predicted Fail

Phase 4: Analysing the Performance

After predicting the course outcome of the students, the results are stored in the warehouse for future use. A huge dataset consisting of marks of students of a particular course for V semesters are collected and pre-processed to analyse the performance of the course at the end of the course. Data Marts are created by grouping the subjects across the five semesters based on the Programming Skills, Management, Computer Core and

Computer Applications. Then the data stored in the data marts are fed to the apriori algorithm to analyze the performance of students in that area. The algorithm generates the frequent item sets for all the data and helps to understand the performance of the students as excellent, good and poor. It also helps to understand the behaviour of the students since the rules helps the miner to understand the behaviour of the students towards the subjects.It is understood from the mart that all the subjects related to programming skills are grouped together and the the values in yellow color indicates that the missing values are replaced by zero's

USN	Name	10MCA11	10MCA14	10MCA16	10MCA17	10MCA21	10MCA22	10MCA26	10MCA28	10MCA33	10MCA37	10MCA41	10MCA43	10MCA46	10MCA47	10MCA57
1RV11MCA01	AAKRITI CHOUDHARY	109	102	67	80	76	84	86	66	93	82	96	83	91	77	91
1RV11MCA02	ABHISEK DEY	107	107	81	92	0	0	0	0	0	0	0	0	0	0	0
1RV11MCA03	ABHISHEK KUMAR	108	90	93	91	94	89	84	65	89	82	77	92	92	69	88
1RV11MCA04	ABHISHEK RATAN	102	115	85	85	81	98	64	57	113	78	85	98	83	68	70
1RV11MCA05	ADARSH MAHESH TUBACHI	113	93	95	94	96	113	93	87	102	93	96	102	91	89	83
1RV11MCA06	AKASH ANBASTI	105	107	90	97	110	107	91	88	110	98	107	113	97	94	93
1RV11MCA07	AKSHAY V GADAG	109	113	97	97	84	99	96	90	101	97	87	92	96	94	93
1RV11MCA08	ALOK RAJ	105	104	93	90	95	109	77	77	98	92	81	97	92	69	88
1RV11MCA09	ANKIT JOSHI	85	77	78	86	70	75	61	53	0	0	0	0	0	0	0
1RV11MCA10	ANTONY FELIX P	108	106	94	87	104	110	87	97	110	87	92	97	92	93	86
1RV11MCA11	ANUP BASTI	113	106	94	98	105	124	95	93	111	94	104	107	95	94	95
1RV11MCA12	ARADHANA UPADHAYA	97	81	65	76	96	114	78	75	108	91	92	85	79	73	80
1RV11MCA13	ARGHA KAMAL MAITY	97	85	78	61	89	82	56	53	89	82	76	81	87	75	85
1RV11MCA14	ARUN R PATIL	116	100	89	93	102	104	86	90	106	87	100	92	92	78	88
1RV11MCA15	ARUN D C	80	79	90	91	76	84	96	72	83	80	77	79	84	79	98
1RV11MCA16	ARUN M	91	84	67	95	88	83	84	89	67	89	75	67	80	77	86
1RV11MCA17	ARVIND AUDICHYA	99	100	68	88	104	95	61	85	102	91	93	98	89	55	70
1RV11MCA18	ASHWAJIT	93	77	60	50	95	80	57	57	88	64	82	94	78	58	53
1RV11MCA19	BALAJI V	90	86	68	92	88	95	64	79	87	83	82	84	91	64	58
1RV11MCA20	BASAVARAJ ANEHOSUR	76	67	35	40	90	87	50	65	81	76	91	84	71	58	88
1RV11MCA21	BIKASH DUTTA	98	93	57	93	0	0	0	0	0	0	0	0	0	0	0
1RV11MCA22	CHAITANYA M G	118	103	94	92	121	118	94	82	126	94	115	114	96	95	96

Fig.3.3. Data mart for Programming Skills

The following figure 3.4 shows frequent items and best rule generated from computer core subjects. The total number Input configurations are 1210 items and 1210 transactions and Found 47 frequents. From this we infer that student have scored more marks in computer core subjects.

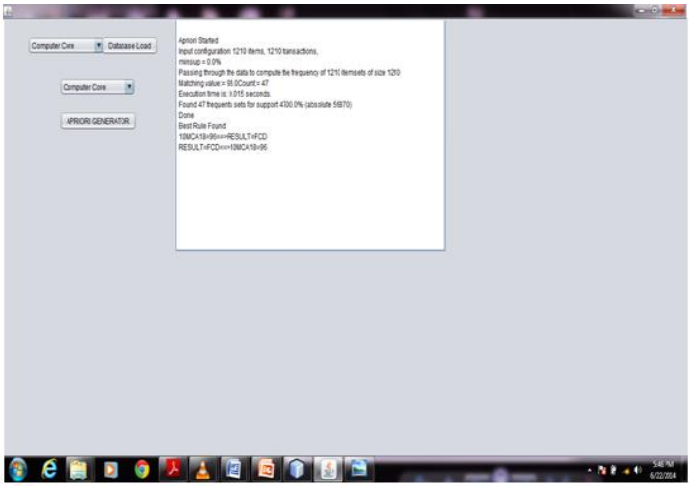

Fig.3.4. Frequent Item set of Computer Core

The outputs are then fed to the visual analytics tool and the studied. It is clear from the VA that the students who scored more marks in Computer core Subject compared to other subjects , from this we can infer the student are excellent in computer core subject, good in computer applications and average in programming skills and management.

4. Result and Discussion

An integrated framework to predict the student's performance of a higher education programme has been built in this project using data warehousing, educational data mining and visual analytics. For the given data set, C4.5 prediction algorithm has been implemented to predict the course outcome of a particular higher education programme and the algorithm was found to perform accurate and efficient. The total time taken to build the module is 0.08 seconds. The tutors were informed about the outcome and advised to improvise the performance of the students by coaching them more. The course outcome a higher education programme has been analysed using apriori algorithm to find out the behaviour of the students towards four different domains of the course by understanding the frequent item sets formed. The visual analytics of the performance of the students helped us to understand the ability of the student in each area of his programme. The system can be further used to take more decisions and bring betterment in education.

Acknowledgment

I wish to thank Computer Society of India for sponsoring for this Minor Research Project and assist me in completing the project work.

References

[1] Titus DE Lafayette Winters, A dissertation work submitted on Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment, 2006.

[2] Dodge, Y. (2006), "The Oxford Dictionary of Statistical Terms" http://www.amazon.com/The-Oxford-Dictionary-Statistical-Terms.

[3] Rajnijindal, A Survey of Educational Data mining and Research Trends, International journal of database management System (IJDMS), vol.5, No.3 June 2013.

[4] Igor, LjiljanaBrkić, Mirta, Improving the ETL process and maintenance of Higher Education Information System Data Warehouse Issue 10, Volume 8, October 2009.

[5] Zlatko J. Kovačić, Early Prediction of Student Success: Mining Students Enrolment Data, Proceedings of Informing Science & IT Education Conference (In SITE) 2010 pp 647-665.

[6] J. Ross Quinlan. "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[7] Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer, & W.F. Punch. "Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA" in Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE (2003).

[8] Cristóbal Romero and et al, Data Mining Algorithms to Classify Students, Computer Science Department, Córdoba University, Spain.

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

[10] Merceron, A. & K. Yacef. "A Web-based Tutoring Tool with Mining Facilities to Improve Learning and Teaching" in Proceedings of 11th International Conference on Artificial Intelligence in Education., F. Verdejo and U. Hoppe (Eds), pp 201-208, Sydney: IOS Press (2003).

[11] Mr. M. N. Quadri and Dr. N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", GJCST Computing Classification H.2.8 & K.3.m, P a g e | 2 Vol. 10 Issue 2 (Ver 1.0), April 2010 Global Journal of Computer Science and Technology.

[12] M. Ramaswami and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", *IJCSI International Journal of Computer Science Issues*, Vol. 7, Issue 1, No. 1, January 2010.

[13] Myzatul Akmam Sapaat, Aida Mustapha and et al, "A Data Mining Approach to Construct Graduates Employability Model in Malaysia", *International Journal on New Computer Architectures and Their Applications (IJNCAA)* 1(4): 1086-1098, The Society of Digital Information and Wireless Communications, 2011 (ISSN: 2220-9085).

[14] Zlatko J. Kovačić, John Steven Green, "Predictive working tool for early identification of 'at risk' students", Published under Creative Commons 3.0 New Zealand Attribution Non-commercial Share Alike Licence (BY-NC-SA) Licensed copy.

[15] "Assessing and Evaluating Student Learning", Atlantic Canada English Language Arts Curriculum: K–3 263, ttp://www.ed.gov.nl.ca/edu/k12/curriculum/guides.

Authors' Profiles

S Anupama Kumar, presently working as an Associate Professor, completed her Doctoral in Educational Data Mining and is currently guiding two Ph D students under Visvesvaraya Technological University, India. She has 15 years of teaching experience and has 18 publications to her credit. Her research interest includes Data Mining and Parallel Computing.

How to cite this paper: S Anupama Kumar,"Edifice an Educational Framework using Educational Data Mining and Visual Analytics", International Journal of Education and Management Engineering(IJEME), Vol.6, No.2, pp.24-30, 2016.DOI: 10.5815/ijeme.2016.02.03