ISSN:2231-0711

Lavannya Varghese et al | IJCSET(www.ijcset.net) | January 2016 | Vol 6, Issue 1, 66-69

Survey on Mining Educational Data and Recommending Best Engineering College

Lavannya Varghese
Student
Computer Science Department, Sahrdaya College of Engineering and Technology
Thrissur (District), Kodakara, Kerala- 680683, INDIA

Ms.Christina Joseph
Assistant Professor
Computer Science Department, Sahrdaya College of Engineering and Technology
Thrissur (District), Kodakara, Kerala- 680683, INDIA

Mr.Vince Paul
Assistant Professor
Computer Science Department, Sahrdaya College of Engineering and Technology
Thrissur (District), Kodakara, Kerala- 680683, INDIA

***Abstract*— In this present era, Engineering colleges are increasing day by day. Good education comes from good colleges and everyone is in search of the best to enhance their future and to live the best of it. Thus finding out one from a 1000 is a difficult task.In India itself around 1077 engineering colleges and in Kerala 145 engineering colleges. So selecting one engineering college is very difficult task.In this project, using data mining techniques and neural network, it help to predict the best engineering college on the basis of the number of students with highest passing out rate and also an account to placement happening in each college and that would definitely help the student community to get the best education and make the life worth living. Using c4.5 algorithm,student performance of different colleges will predict and using that pass/fail ratio and all other attributes like placement rate, rank holders etc recommending the best engineering college. This web application enables the colleges to enter their details. And also it can shows the analysis of their own college. Then each colleges see other registered colleges analysis based on above mentioned factors. This will help the college to improve on their facilities to become best among all.**

***Keywords*— c4.5,neural network,datamining**

I.INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospec-tive analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hid-den patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data.

These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as: Forecasting: Estimating sales, predicting server loads or server downtime Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes Recommendations: Determining which products are likely to be sold together, generating recommendations Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities

In the past years, the number of engineering colleges were very less.Even though the student applicant were high. So only the best deserving student could get their seats in the desired college. Now the scenario has changed.The colleges has increased with the increase in students and so we have a lot of vacancies in each college. Thus parents find it really confusing to find which college suits the best for their wards. Around 17,907 have remained vacant, according to a leading daily national newspaper. According to the statistics released by the Kerala Tech-nological University(KTU), 70 percentage of the seats have been filled so far.

The main challenge in this project is data collection. In the past years, the number of engineering colleges were very less. Now the scenario has changed. The colleges has increased with the increase in students. There are 152 engineering colleges in kerala. So collection of student academic, personal and college related attributes are very difficult task. So inorder to reduce the difficulty, selected 8 nearby engineering colleges and collected the data. The experiments are conducted in weka machine learning software tool, so data is converted to arff (attribute relation file format) that is another challenge in this project.

II. LITERATURE SURVEY

Data mining in higher education is a recent research field and this area of research is gain-ing popularity because of its potentials to educational institutes. Data Mining can be 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 as described by Alaa el-Halees [4]. Mining in educational environment is called Educational Data Mining.

Han and Kamber [3] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Pandey and Pal [5] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Hijazi and Naqvi [6] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as Students attitude towards attendance in class, hours spent in study on daily basis after college, students family income, students mothers age and mothers education are significantly related with student performance was framed. By means of simple linear regression analysis, it was found that the factors like mothers education and studentâĂ§s family income were highly correlated with the student academic performance.

Khan [7] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, person-ality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socioeconomic status had relatively higher academic achievement in general.

Galit [8] 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. Al-Radaideh, et al [9] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the NaÃŕveBayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Kanakana and Olanrewaju [11] used Artificial Neural Network and linear regression mod-els to predict student performance after access to higher education. Data received from the Tshwane University of Technology was utilized for the study. The total Average Point Scores students obtained in grade 12 was employed as input variable. The results indicated a better agreement between ANN model prediction and observed values compared to those in the linear regression.

Kyndt et al. in their study predicted general academic performance in the first bachelor year educational sciences, based on students motivation, approaches to learning, working mem-ory capacity and attention using a neural network analysis. Participants in this study were 128 university students. Results showed that working memory capacity and attention are both good predictors of academic performance, especially for the best and weakest performers of the group. Students motivation and approaches to learning were good predictors for the group of students whose performance was in the middle 60percentage [12].

Mukta and Usha carried out an analysis to predict the academic performance of business school graduates using neural networks and traditional statistical techniques and the results were compared to evaluate the performance of these techniques. The underlying constructs in a traditional business school curriculum were also identified and its relevance with the various elements of admission process were presented [13].

Stamos and Andreas presented a model using an artificial neural for predicting student grad-uation outcomes. The network was developed as a three-layered perceptron and was trained using the backpropagation principles. For training and testing various experiments were ex-ecuted. In these experiments, a sample of 1,407 profiles of students was used. The sample represented students at Waubonsee College and it was divided into two sets. The first set of 1,100 profiles was used for training and the remaining 307 profiles were used for testing. The average predictability rate for the training and test sets were 77percentage and 68percentage respectively [14].

Cesar et al. [6] proposed the use of a recommendation system based on data mining tech-niques to help students to make decisions related to their academic track. The system provided support for students to better choose how many and which courses to enroll on. As a result, the authors developed a system that is capable to predict the failure or success of a student in any course using a classifier obtained from the analysis of a set of historical data related to the academic field of other students who took the same course in the past.

Baha Sen and Emine Ucar compared the achievements of Computer Engineering Depart-ment students in KarabÃijk University according to various factors such as age, gender, type of high school graduation and the students studying in distance education or regular education through data mining techniques. They havetaken the dataset of 3047 records. In their study they have used NN architecture called multilayer perceptron (MLP) with back propagation type supervised-learning algorithm to produce both classification and regression type predic-tion models and decision tree for achieving the highest possible prediction accuracy. [3]

Tiwari et al., conducted a study on engineering students to evaluate their performance by applying data mining

techniques to assist them in decision making. They used K-Means al-gorithm to cluster students. The result predicted that if students are poor in attendance and assignment then there is 75percenatge probability that their grades are poor. [2]

Ali and Kerem studied the data set of students of Istanbul Eyup I.M.K.B.Vocational Com-merce High School and found the relationship between the student performance and course. In their finding they have generated a rule that shows if a candidate is unsuccessful in numerical course in 9 the class then those students are likely to be unsuccessful in 10 the class. Such results were generated for different courses. This study can facilitate students to choose their appropriate profession by revealing the relation between their concern fields. [1]

III.DESIGN

In this project, using data mining techniques and neural network, it help to predict the best engineering college on the basis of the number of students with highest passing out rate and also an account to placement happening in each college and that would definitely help the student community to get the best education and make the life worth living . Using c4.5 algorithm, student performance of different colleges will predict and using that pass/fail ratio and all other attributes like placement rate, rank holders etc predict the best engineering college.

Inorder to collect data , i have to decide what all are the attributes that contribute more informations. So for that i conducted survey on first year computer science students of sahrdaya college. Survey is mainly based on what all are the factors that depends to select sahrdaya engineering college for their education. Survey consists of 15 yes or no questions.Based on survey result and analysis of survey attributes related best engineering college prediction is determined.

Level 0 is mainly concerned with attribute selection, datacollection and data preprocessing.In attribute selection, the attributes which provide more information regarding student performance and best col-lege is identified. After start collect the students academic,personal and college details from different colleges. Next step is data preprocessing. Some attributes are not relevant and not provide any useful information to predict hidden information. So remove all unwanted attributes. The experiments are conducted in weka machine learning software tool, so data is converted to arff.

In first level of project,attributes for predicting students are pass or fail in future exam is identified. Students academic and personal informations are collected.Then apply c4.5 algorithm and mine hidden predictIve information like student pass or fail in future exam. In the level 2,using naive baysian and neural network predicting best engineering college.For this, take result form level 1 as one of the attribute to predict best engineering college.

A. .Attribute Selection

Attribute is a piece of information which determines the properties of a field or tag in a database or a string of characters in a display. An attribute is property or characteristic of an object. Attribute values are numbers or symbols assigned to an attribute.Same attribute can be mapped to different attribute values.In this project,inorder to mine data,attributes related to personal information,academic information and college related attributes are needed. in first part of project,using c4.5 algorithm predicting student pass or fail in future exam based on personal and academic attributes.academic attributes like assignment marks,university marks etc and personal in-formation attributes are like hostler,marital status etc. In second part,take this pass or fail ratio as one of the attributes to predict best engineering college using neural network and naive bayesian .

B.Data Collection

Collection of details from all colleges in kerala is very difficult. Data collection is the main challenge in this project.Students personal,academicinformations and college related details are collected from nearby 8 colleges. So for my project ,selected nearby colleges.

C.Data Preprocessing

Data pre-processing is among the common steps prior to applying any data mining tech-nique. We applied the following steps to prepare the data:

Eliminating the records of students who withdrew from the course because some of their relevant values were consequently missing.

Discretizing the total grade attribute into five categories: A, B, C, D, and F. Discretizing all attributes of the semester into four categories: excellent, good, average, and poor. After pre-processing the data, we ran the Waikato Environment for Knowledge Analysis (Weka) toolkit to apply the classification algorithms. Weka was developed at the University of Waikato in New Zealand, and is very popular data mining software that con-tains a wide range of algorithms implemented in Java. The experiments are conducted in weka machine learning software tool, so data is converted to arff (attribute relation file format).

D.Analysis And Prediction

After appliying algorithms ,Analysis and Prediction phase is analyse and compare the efficiency of both neural network and Naive bayesian algorithm.also it analyse and compare 8 engineering colleges that we selected for predicting best engineering college .Recommending the best engineering college using data mining algorithms and neural network.

IV. METHODS

A.C4.5

C4.5 Algorithm is a decision tree technique which is enhanced by ID3 algorithm. It is one of the most popular algorithm for rule base classification. Here an attributes can be split into two partition based on the selected threshold value, all the value satisfied by the constraint it will be assigned in one child and remaining values can be store in another child respectively. It also handles missing values. Here it can be gather of all binary tests through entropy

gain and the values are sorted based on the values in continuous attribute values which are calculated in one scan. This process is repeated for each continuous attributes when the process is terminated. uses continue data. It avoids over fitting of data. It improves computational efficiency and it handles training data with missing and numeric value .inorder to predict whether the particular student is pass or fail in future exam C4.5 algo-rithm is used. This is the first part of the project.Using students academic and personal informations mine the data by appliying c4.5 algorithm.Output of this algorithm, that is pass/fail rato is given as one of the attributes for next part inorder to predict best engi-neering college.

B.Naive Baysian

In probability theory, Naive Bayes classifier checking the condition rule and it can be classified by learning phase and testing phase. Bayesian reasoning is applied to decision making that deals with probability inference which is used to gather the knowledge of prior events by predicting events through rule base. The rule to estimate of a property given the set of data as evidence or input Bayes rule theorem. Naive bayesian is used to predict best engineering college by using college related at-tributes.and also compare the efficiency of naive bayesian and neural network

C.Neural Network

The area of neural networks probably belongs to the border line between the artificial in-telligence and approximation algorithm. A neural network is a collection of neurons like processing units with weighted connection between the units. It composes of many ele-ments, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed. A classification model can be represented in different forms like neural network and deci-sion tree.There are many advantages of neural networks such as adaptive learning ability, self-organization, real time operation and insensitivity to noise. Neural networks are used to identifying patterns or trends in data and well suited for prediction or forecasting needs. There are several neural network algorithms such as Back Propagation, NN Supervised Learning, and Radial Base Function (RBF) Network etc. In this project,Neural network is used to predict best engineering college using college re-lated attributes and result from the c4.5 ,ie pass/fail ratio.for this purpose backpropogation algorithm is used and input layer consist of attributes.Hidden layer performing the activa-tion function and output layer produce the final output.

V. CONCLUSION

The process of educational data mining is an iterative,knowledge discovery process which consist of hypothesis formulation, testing and refinement. Hypothesis is developed from various educational environments.It create large volume of data.Good education comes from good colleges and everyone is in search of the best to enhance their future and to live the best of it. Thus finding out one from a 1000 is a difficult task.In my project, using data mining techniques and neural network, it help to predict the best engineering college on the basis of the number of students with highest passing out rate and also an account to placement happening in each college and that would definitely help the student community to get the best education and make the life worth living. Using c4.5 algorithm, student performance of different colleges will predict and using that pass/fail ratio and all other attributes like placement rate, rank holders etc. We aim to help students in the decision making process through our project.

ACKNOWLEDGMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this seminar report. I am thankful for their aspiring guidance, invaluably constructive criticism and friendy advice during the project work.I deeply express my sincere thanks to my seminar guide Ms.Christina Joseph for her support and guidance.

REFERENCES

[1]] Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier .

[2] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Educa-tion Inc.

[3] Kantardzic, M., (2011) Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press.

[4] Ming, H., Wenying, N. and Xu, L., (2009) An improved decision tree classification algo-rithm based on ID3 and the application in score analysis, Chinese Control and Decision Conference (CCDC), pp1876-1879

[5] Mohammed M. Abu Tair,mining educational data to improve students performance of Information and Communication Technology Research, Volume 2 No. 2, February 2012 ISSN 2223-4985.

[6] M. Al-Razgan, A. Al-Khalifa, and H. Al-Khalifa, Educational Data Mining: A Systematic Review of the Published Literature 2006-2013, in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), vol. 285, T. Herawan, M. M. Deris, and J. Abawajy, Eds. Springer Singapore, 2014, pp. 711-719.

[7] T. Tanner and H. Toivonen, Predicting and preventing student failure using the k-nearest neighbour method to predict student performance in an online course environment, International Journal of Learning Technology, vol. 5, no. 4, pp. 356-377, 2010.

[8] C. MÃČÃąrquez-Vera, A. Cano, C. Romero, and S. Ventura, Predicting student failure at school using genetic programming and different data mining approaches with high dimen-sional and imbalanced data, Applied Intelligence, vol. 38, no. 3, pp. 315-330, 2013.

[9] Ogor Emmanuel. N, Student Academic Performance: Monitoring and Evaluation Using Data Mining Techniques. Fourth Congress of Electronics, Robotics and Automotive Me-chanics. 2007. I EEE Computer Society.

[10] Alaa el-Halees,Mining Students Data to Analyze e-Learning Behavior: A Case Study, 2009.

[11] Mohammed M. Abu Tair,mining educational data to improve students performance of Information and Communication Technology Research, Volume 2 No. 2, February 2012 ISSN 2223-4985.

[12] Data Mining in Educational System using WEKA, Sunita B Aher, Mr. LOBO L.M.R.J. International Conference on Emerging Technology Trends (ICETT) 2011 Proceedings pub-lished by International Journal of Computer Applications (IJCA)

[13] Ali Buldua, Kerem . Data mining application on students data. Procedia Social and Be-havioral Sciences 2 5251-5259, 2010.

[14] Romero ,Cristobel. Educational Data Mining: A Review of the State-of-theArt,Member, IEEE, Sebastian Ventura, Senior Member, IEEE2010Ali Buldua, Kerem .