Study of influencing factors of academic performance of students: A data mining Approach

V.Ramesh, P.Thenmozhi, Dr.K. Ramar

Abstract - The main concerns of any higher educational system is evaluating and enhancing the educational organization so as to improve the quality of their services and satisfy their customer's needs. This is an attempt to find suitable prediction techniques using data mining tool WEKA to help in enhancing the quality of the higher educational system by evaluating student data to predict the student performance in courses during early period of study. This will help the educational institutions to identify the students who are at risk and to tale necessary steps to reduce failing ratio at right time to improve the quality of education. The process of finding a suitable prediction algorithm was also described.

Keywords - *Classification, Prediction, Student performance*

1. INTRODUCTION

The highly inter-disciplinary field of Educational Data Mining (EDM) has resulted from a fusion of many different areas, some of which include Machine Learning, Cognitive Science and Psychometrics. The main task in EDM is to construct computational models and tools to mine data that originated in an educational setting. With rapidly increasing data repositories from different educational contexts (paper tests, e-learning, Intelligent tutoring system etc) good practices in EDM can potentially answer important research questions about student learning. Recently educational institutions targets activities within its organizations with ERP tools to handle and store huge data available in educational process for hidden patterns. The objective of this study are (i) prediction of first year engineering student's performance (ii) find out association between the different factors influencing grades and (iii) compare different prediction algorithms for classifying students. This study is more useful for identifying weak students and the identified students can be individually assisted by the educators so that their performance is better in future.

2. REVIEW OF LITERATURE

The application of Data mining is widely spread in Higher Education system. Many researchers and authors have been explored and discussed various applications of data mining in higher education. Guan Li [1] has compared the accuracy of data mining methods to classifying students in order to predicting student's class grade. These predictions are more useful for identifying weak students and assisting management to take remedial measures at early stages to produce excellent graduate that will graduate at least with second class upper.

J.F. Superby [2] conducted a study to investigate to determine the factors to be taken into account we will use a model adapted from that of Philippe Parmentier (1994). In other words the idea is to determine if it is possible to predict a decision variable using the explanatory variables which we retained in the model.

Bray[3] in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed that there was an enhcancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Ramaswami[4] in his study on CHAID based performance prediction model, observed that the CHAID prediction model was useful to analyse the interrelation between variables that are used to predict the outcome on the performance at higher secondary school education.

V.O.Oladokun[5] in his study on predicting student's academic performance using artificial neural network, observed that Multilayer Perception Topology was best to predict the performance of more than 70% of prospective students.

3. METHODOLOGY

Through extensive search of the literature and discussion with experts on student performance, a number of socio-economic, environmental, academic, and other related factors that are considered to have influence on the performance of a

university student were identified. These factors were carefully studied and harmonized into a manageable number suitable for computer coding within the context of the familiar algorithms. These influencing factors were categorized as input variables. The output on the other hand represents some possible levels of performance of a candidate in terms of the present college grading system.

3.1 Data Collection

For this study real world data's are collected from first year engineering students. Two colleges are selected from Kancheepuram district of Tamil Nadu. A sample of 464 students was taken from a group of colleges. Students were grouped in a classroom they were briefed clearly about the questionnaire and it took on average half an hour to fill this questionnaire. Selection of students was at random.

The primary data was collected using a questionnaire. Which include questions (i.e. with predefined options) related to several personal, socio-economic, psychological and school and college related variables that were expected to affect student performance. The questionnaire was reviewed by professionals and tested on a small set of 50 students in order to get a feedback. The final version contained 23 questions in a single A4 sheet and it was answered by more than 700 students. Latter we selected a sample of 464 from the whole. All questionnaires were filled with the response rate of 100% out of which 316 were females and 184 were males.

The secondary data such as semester mark details, attendance percentage, and class test performance were collected from the college and from the directed website. All the predictor and response variables which were derived from the questionnaire are given in Table 3.1 for reference.

Table 3.1: Student Related Variables

Variable Name	Description	Domain
SEX	student's sex	{male, female}
COMM	student's community	{OC, BC, MBC, SC, ST}
REL	student's religion	{ hindu, christian, muslim }
SA	student's living area	{urban, rural}
SD	student's department	{ mechanical, computer science, EEE, ECE, CIVIL, IT }
PQ	parent's qualification	{ illiterate, schooling, degree/diploma }
F-OCC	father occupation	{daily wages, farmer, weaver, ex serviceman, government, business, private}
M-OCC	mother occupation	{house wife, daily wages, farmer, weaver, private, government }
MOS	student's schooling medium of study	{ tamil, english }
SSLC-TOTAL	student's 10^{th} total	{ <250, 251 – 350, 351 – 450, >451 }
SSLC-PERCEN	student's 10^{th} percentage	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
HSC-TOTAL	student's 12^{th} total	{ <250, 251 – 350, 351 – 450, >451 }
HSC-PERCEN	student's 12^{th} percentage	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
HSC-CUT	student's 12^{th} cut off	{ <50, 51-100, 101 -150, 151 – 200 }
SQ	student's quota	{ management , counselling }
SSP	student's staying place	{ day scholar, hostel }
SATT	student's college attendance percentage	{<50, 50, 60, 70, 80, 90}
G-OBT	grade obtained	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}

3.2 Algorithms

Although many classification models exist, only some have been selected within the scope of this study. The selected algorithms are Naïve Bayesian algorithm, MLP, SMO, J48, REP tree, RANDOM tree and Decision table are used. The Naïve Bayesian model defines the classification problem with respect to probabilistic idioms, and supplies statistical methods to classify the instances based on probabilities [8]. Multilayer perceptron is a type of artificial neural network algorithm which regards the human brain as the modelling tool [8]. It provides a generic model for learning real, discrete and vector target values. The ability to understand the hidden model is hard and training times may be long. In decision tree algorithms, the classification process is summarised by a tree. After the model is built, it is applied to the database.

3.3 Implementation

First, data cleaning was applied on the datasets. According to the missing data analysis, missing data have been removed from the datasets. Other than missing data analysis, datasets were also cleaned to remove noisy data. Unnecessary space characters or other spelling mistakes were also cleaned in the datasets. Another usual step in data pre-processing is data discretisation. Although some algorithms are said to perform better when the numerical input variables are discretised, in this study numerical variables have not been put into binned intervals in order to maintain the same conditions for all algorithms.

Once the data pre-processing steps have been completed, the dataset have been used to run the classification algorithms Naïve Bayesian algorithm, MLP, SMO, J48, REP tree, RANDOM tree and Decision table. For all algorithms, splitting the data into train and test splits has been selected as the validation method. 66% of the data has been set as the training part and the rest has been set as the testing part. .

3.4 Weka Data Mining Software

WEKA is open source software issued under the GNU General Public License. WEKA has been utilized as the tool to run different classification algorithms. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

This paper is an attempt to use classification techniques to analyze and evaluate student academic data and to enhance the quality of the higher educational system. Findings from the factors influencing academic performance were very significant. When pointing at the performance results of the classifier, its classification accuracy is actually measured. Accuracy is calculated by determining the percentage of instances correctly classified. Costs for wrong assignment can also be applied in classification problems; however, misclassification costs are not within the scope of this study. The accuracy values of the multiple dataset implementations according to each classifier can be seen in Table 4.1 (in percentage)

Table 4.1: Comparison of classification accuracy

	NAIVE BAYES	MLP	SMO	J48	REPTREE	RANDOM TREE	DECISION TABLE
Accuracy	55.8	69.5	58.8	65.8	54.4	56.2	51.7

The different classification algorithm predicts the grade more accurately. From this study Multi Layer Percepton predict the all the grades more accurately. The results of accuracy of different classifier are given below:

Table 4.2: Grade wise accuracy of classifiers

Algorithms/ Grades	NAIVE BAYES	MLP	SMO	J48	REPTREE	RANDOM TREE	DECISION TREE
A	40.00	70.00	30.00	44.28	14.28	14.2	40.00
B	31.81	61.81	27.27	27.27	36.36	27.27	45.45
C	3.77	70.00	31.88	52.83	43.77	30.94	45.66
D	42.30	79.23	36.15	36.15	40.00	33.07	49.23
E	20.77	87.79	40.38	30.38	37.79	32.59	30.38
U	60.00	85.00	50.00	60.00	62.5	40.00	50.00
U+	50.00	60.00	40.00	60.00	40.00	30.00	40.00

To verify the relationships among the attributes, hypothesis is formed and tested. The results of hypothesis test is given in Table 4.3

Table 4.3: Testing Hypothesis

Attributes	Hypothesis
Student quota through they joined, Grade obtained at semester examination	Ho rejected
Student area they are belonging, Grade obtained at semester examination	Ho accepted
Student staying place, Grade obtained at semester examination	Ho accepted
Parent's qualification, Grade obtained at semester examination	Ho rejected
Student department, Grade obtained at semester examination	Ho rejected
Medium of schooling, Grade obtained at semester examination	Ho rejected
Student sex, Grade obtained at semester examination	Ho rejected

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong. In our case, we get several strong rules in our association since they satisfy these requirements. However, not all of them are useful for us. Because of that, we have to choose what the rules we need are and we should apply to. By selecting those useful rules, we can use them to do the prediction like we did before. Of course, even some attributes and rules are not our interests now; we might still need them later. When we want to predict other different attributes, we just need to repeat the same processes. In this anaysis following are the strong association rules.

1. Department=Mechanical 84 ==> Sex=M
2. Area=urban Quota=management 66 ==> Staying place=day scholar
3. 10th/12th medium=english Staying place=hostel 55 ==> Quota=counseling
4. Department=Mechanical Parent qualification=Schooling 54 ==> Sex=M
5. Department=Mechanical 10th/12th medium=tamil 53 ==> Sex=M

5. CONCLUSION

In this paper we have analyzed various factors influencing the academic performance of the students at engineering college level and predict the grade of the student if these factors are given as input. Performance of different classification algorithms are compared for classifying students using a Weka mining tool. We have shown that some algorithms improve their classification performance. We have also indicated that a good classifier model has to be both accurate and comprehensible for instructors. These included students' staying place, whether they stayed in hostel or day scholar in their first year of study and the area him/her belonging. This study will give a timely and an appropriate warning to students at risk. This work may improve student performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. Therefore, it seems to us that data mining has a lot of potential for education.

REFERENCES

[1] Guan Li, Liang Hongjun. Data warehouse and data mining. Microcomputer Applications. 1999, 15(9): 17-20.
[2] J.F. Superby , J.-P. Vandamme & N. Meskens of Catholic University of Mons "Determination of factors influencing the achievement of the first-year university students using data mining methods".
[3] Rosemary Win and Paul W. Miller: The Effects of Individual and School Factors on University Students' Academic Performance
[4] Adriaans P, Zantinge D. Data mining [M]. Addision_Wesley Longman, 1996.
[5] Chen Rong, BP arithmetic and its structure optimization tactics. Journal of Autoimmunization. 1997, 23(1), 43-49.
[6] Alcalá, J., Sánchez, L., García, S., del Jesus, M. et. al. KEEL A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing, 2007.
[7] Baker, R., Corbett, A., Koedinger, K. Detecting Student Misuse of Intelligent Tutoring Systems. Intelligent Tutoring Systems. Alagoas, 2004. pp.531–540.
[8] Barandela, R., Sánchez, J.S., García, V., Rangel, E. Strategies for Learning in Class Imbalance Problems. Pattern Recognition 2003, 36(3), pp.849-851.
[9] Breiman, L. Friedman, J.H., Olshen, R.A., Stone, C.J. Classification and Regression Trees. Chapman & Hall, New York, 1984.

AUTHORS PROFILE

V.Ramesh is Assistant Professor in the Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Kanchipuram, Tamil Nadu, India. He received his M.Phil. in the area of Data Mining from Madurai Kamaraj University, Madurai. He is doing PhD degree in the area of Data Mining in Agriculture. He has published more than 5 research papers in National, International journals and conferences. His research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks and Software Engineering.

P.Thenmozhi Mphil Research Scholar in the Department of Computer Science & Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Enathur, Kanchipuram. She received the degree in Master of Computer Applications from Arulmigu Meenakshi Amman Engineering College in 2009. At present she is working as Professor at Department of Computer Sciences in Thirumalai Engineering College, kanchipuram, Tamil Nadu. She has published 3 research papers in National, International Journals and conferences. Her research interest lies in the area of Data Mining, Neural Networks.

Dr.K.Ramar received his B.E. (Electronics and Communication Engineering) in the year 1986 from Government College of Engineering - Tirunelveli, Madurai Kamaraj University, Tamilnadu, India, M.E. (Computer Science and Engineering) in the year 1991 from PSG College of Technology, Bharathiyar University, Coimbatore Tamilnadu, India, and Ph.D in the year 2001 from Manomaniam Sundaranar University, Tamilnadu, India. He has more than 25 years of teaching experience. Currently he is working as Principal, Einstein College of Engineering, Tirunelveli, Tamilnadu, India. He published many papers in International journals and conferences. He organized many National and International conferences. His research interests include Pattern Recognition, Image Processing, Computer Networks and Fuzzy Logic based Systems.