Machine Learning Techniques for Prediction of Subject Scores: A Comparative Study

[1] Mamta Singh, [2] Dr. Jyoti Singh

[1] Sai Mahavidyalaya, Sector-6, Bhilai, Chhattisgarh,490006, India

[2] Vyavasaik Pariksha Mandal Raipur, Chhattisgarh, 490006, India

Abstract

In this paper, a novel method is proposed so as to predict the subject wise academic performance of the Engineering students. This study describes the prediction of subject scores in ongoing courses by analyzing subject preludes of previous semester. In this study we try to predict the individual subject scores for ongoing courses while comparing two classification techniques i.e. Naive Bayesian and C4.5 Decision tree classifier. This piece of work adheres to most critical aspect of Quality objectives of Academia i.e. finding students' academic performances for their ongoing courses well before they face their End semester Examination. Unlike the recent research trends that focused on predicting overall grading of students during their studies, this paper orients itself in identifying students grasping levels subject wise. It was found that from study, that obtained accuracy figure was higher in C4.5 Decision tree classifier than Naïve Bayes.

Keywords: *Academic Performance, C4.5, Naïve Bayesian classification, Analytics, Subject prelude.*

1. Introduction

Academic analytics is a new area that was introduced in higher education with quality higher education objective. It is a buzz word often used to describe the application of data mining technique to develop predictive model that can help monitor and anticipate student performance and take action in issues related to student teaching and learning. Results of student's academics can be used by various managerial levels of education system. While teachers can also use this information to predict their students subject wise performance. The most striking features of data mining technique are clustering & prediction. The clustering do the comprehensive characteristics analysis of students while the predicting function estimate the different types of outcomes like transferability persistence, retention and success in study.

2. Related Studies

Many mining experts have attempted to investigate various methodologies for improving students' academic performance, particularly in higher education courses. The evaluation parameters are found varying i.e. their forthcoming end semester scores in form of (pass/fail, grades or percentages). Others have tried to identify slow learners, so that they can be counseled with remedial measures', before the commencement of their forthcoming examinations.

Still, others have pursued the similar mining logistics for identifying student's dropouts, well beforehand.S. K Yadav., B Bharadwaj and S Pal (2011-2012) have focused their detailed research studies in formulating appropriate prediction models for predicting students' academic performance in variety of dimensions. In one of their surveys, Bharadwaj B and Pal S (2011) identified high potential variables like medium, cast category; division, gender, father's qualification and mothers' qualification ,which were also found to influence higher grades apart from their past academic score. They used these parameters as predictor attribute upon Naïve Bayesian classifier model [1].

Yadav and Pal (2012) also studied the contribution of student enrollment data for predicting the students' academic scores in forthcoming examinations [2].

R.R Kabra. and R.S Bichker. (2011) worked with Decision Tree technique to predict their overall performance [3], their work was inspired from still earlier works done by Bresfelcan (2007),Cortez and Silva(2008)and Kovacic Z.J. (2010). All the three groups of data miners predicted the students' academic performance with different classification techniques: decision tree, CHAID and CART. Interestingly their research objective also varied in the sense that ID3 technique was used to identify UG students', likelihood for contribute their PG courses, Cortez and Silva did also consider socio-economic factors, along with past academic performance a set of predictors .While Kovacic unfolded the query, upon to what extent, enrollment data should be used to predict students' success.

Although their data collection also included demographic statistics like fathers' qualification, mothers' qualification, parents occupation, living location, gender and medium of secondary education, yet they realized, it was more of past academic performance figures that helped in drafting IF-THEN decision tree rules.

Quadril M.N and Kalyankar N.V [2009] conducted study upon student's academic performance contributed by their cumulative grade point average (CGPA) [4].

Kruk S.E, Diane Lending [2003] developed a model to predict academic performance of students pursuing Information system (IS) course at introductory college level. In this study they analyzed the data, using regression analysis which checked the model and also for gender moderating effects [5].

Shymala K. and Rajgopalan S.P. (2006) presented and justified the capabilities of data mining in the context of higher education by offering a data mining model for higher educational system in the colleges. They survey upon one hundred and eighty students of Dr.Ambedkar Govt. College. The algorithm C5.0 was used for this model, the internal assessment and previous semester grades were the basic attribute for predicting current semester grade or result [6].

Pandey and Pal (2011) presented a case study that used Bayes classification method to predict the student division on the basis of previous year data base. They took sample data of six hundred students from PGDCA course of session 2009-10. The contributed attributes for this study were cast category, medium in which student passed his\her graduation program, Class is the stream which shows that a student passed to get admission in PGDCA.

3. Data Mining Process

The ability to predict a student's performance accurately is a very crucial aspect in any educational environment. Predicting the academic outcome of a student in a subject needs lots of parameters to a thinking process upon the students' understanding levels in that subject. It was agreed to one of thought that grasping a subject of a level needs complete understanding of the concepts of its subject preludes. For instance, it was decided that to understand concept of Artificial intelligence & Expert System in final year of Engineering curriculum, The students should be well versed with programming concepts, Data Structures, Data Base Management systems, 'C' languages, 'C++' languages.

A blend of such attributes was expected to possibly influence the AI subject scores of students and word selected as described in table below.

Table 1: Attributes and their domain

Variable	Description	Positive domain value
C	Programming in 'C' language	< 45 45-54 55-64 65-80
C++	Programming in C++ language	< 45 45-54 55-64 65-80
DS	Data Structure	< 45 45-54 55-64 65-80
DBMS	Data Base Management System	< 45 45-54 55-64 65-80
SP	System Programming	< 45 45-54 55-64 65-80

A formal stage of preprocessing the data was done by looking the data set at a glance cleaning out the spurious tuples. The spurious tuples here comprised of those Student's scores that are leading to either withheld grade sheets, supplementary or detained scenarios. The next vital preprocessing step was mapping the discrete values of the subject scores to four point nominal scale for the four different levels of mark ranges defined as: <45 as 1, 45-54 mapped to 2, 55-64 to 3 and 65-80 to 4. This was done adhering to the input data constraints that reside as a part of C4.5 algorithm and Naïve Bayesian classifier.

3.1 Data Set Construction

The data set used in this study was obtained from Chhattisgarh Swami Vivekananda University. These data were analyzed using classification method to predict the students' performance in a subject of their current semester.

3.1.1 Training Data Set

The training dataset was used to train or build a model. In the data set provided, each batch comprised of approximately 60 students, there by information about a total of 120 students from two passed-out batches was considered in input collection. For a while, only one batch was selected upon whom the experiments were performed with training-test bed ratios of 15-45, 30-30 & 40-20 split statistics.

3.1.2 Validation Data Set

Once a model was built using the training dataset, the performance of the model must be validated using new data. If the training data itself was utilized to compute the accuracy of the model fit, the result would be an overly optimistic estimate of the accuracy of the model. This is because the training or model fitting process ensures that the accuracy model for the training data is as high as possible. Estimate of how would perform with unseen data, must set aside of the original as the validation dataset.

3.1.3 Test Set

The validation dataset is often used to fine-tune models. The present study has taken up various ratio combinations of dataset viz. 50-50, 60-40, 70-30 and 80-20 (training data and test data) in order to obtain highest accuracy from among dataset. The accuracy of the model on the test data gives a realistic estimate of the performance.

4. Experimental Setup

XL-Miner is open source software that implements a large collection of machine learning algorithm and is widely used in data mining applications. From the students data of Engineering College using standard data partition are partitioned our data 60% training and 40% validation (60-40) similarly 80-20, 70-80 and 50-50% partitioning and apply Naïve Bayes classifier on them then we had found in every case maximum accuracy of data is 56%. In training set accuracy of data is more than validation set.

5. Results and Discussion

The results obtained in tables 1, 2, and 3 respectively show that the logistics behind research interest in such a direction is expected to assess the students 'at risk' more closely, i.e. with reference to each subject as an evolutionary dimension. In this way appropriate an remedial action can be taken adapting to different strategies for different subjects, and for different ranges of weakly identified students. More over, this study acts as a stepping stone towards an extended proposal of improving the prediction accuracy figures.

Table 2- Confusion matrix showing classification accuracy upon predicted validated trained AI scores with 70-30 training validation ratio.

Classification Confusion Matrix				
	Predicted Class			
Actual Class	<45	45-54	55-64	65-80
<45	14	1	1	0
45-54	10	13	5	3
65-80	0	2	2	5

Error Report			
Class	***#Class***	***#Error***	***%Error***
<45	16	2	12.50
45..54	31	18	58.06
55..64	28	12	42.86
65..80	9	4	44.44
Overall	84	36	42.86

Table 3- Confusion matrix showing classification accuracy upon predicted validated AI scores with 70-30 training validation ratio.

Classification confusion Matrix				
	Predicted Class			
Actual Class	***<45***	***45..54***	***55..64***	***65.80***
<45	5	1	1	0
45..54	6	3	4	1
55..64	1	2	6	3
65..80	0	0	1	2

Error Report			
Class	***#class***	***#Error***	***%Error***
<45	7	2	28.57
45..54	14	11	78.57
55..64	12	6	50.00
65..80	3	1	33.33
Overall	36	20	55.56

Table 4: Confusion matrix showing classification accuracy upon predicted test AI scores with 70-30 training validation ratio.

Classification Confusion Matrix				
	Predicated Class			
Actual class	<45	45..54	55..64	65..80

<45	31	4	1	0
45..54	6	4	0	0
55.64	3	0	1	0
65..80	1	0	0	1

Error Report			
Class	***#Class***	***#Error***	***%Error***
<45	36	5	13.89
45..54	10	6	60.00
55..64	4	3	75.00
65..80	2	1	50.00
Overall	52	15	28.85

6. Conclusion

Unlike other prior related works done that focused upon investigating overall students' academic performance; the classification model like Naïve Bayesian method was used to predict the subject wise student performance for further semester examination on the basis of previous semester subject scores.

References

[1] S. K. Yadav, B.K Bharadwaj. and S Pal., "Data Mining Application: A comparative study for Predicting Student's Performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19.

[2] S.K Yadav. and S. Pal, "A prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal, (2012) (WCSIT) ISSN: 2221-0741, Vol.2, 51-56, 2012.

[3] R.R Kabra. and R.S Bichker., "Performance Prediction of Engineering Students using Decision Tree", International Journal of computer Application December 2011 (0975-8887)Vol. 36 No. 11.

[4] M..N Quadril. and N. V., Kalyankar. "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Global Journal of Computer Science and Technology, 2010, Vol. 101Issue 2, pp.2-5, April.

[5] S.E Kruk., Diane Lending, "Predicting Academic Performance in an Introductory College –Level IS Course: Information Technology, Learning, and Performance Journal, 2003 Vol. 21, No. 2.

[6] K Shymala. and S.P Rajgopalan., "Data Mining Model for a Better Higher Educational System", Information Technology Journal, 2006, 5(3)560-564.

[7] U. K Pandey., and S Pal., "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, 2011, Vol. 2(2), pp.686-69.

[8] B.K. Bharadwaj and S.Pal. "Data mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), 2011, Vol. 9, No. 4, pp. 136-140.

First Author- Mamta Singh received her MCA degree from Maharshi Dayanand University, Rohtak in 2005. She has also received her MPhil in Computer Science from Periyar University, Salem. She is working presently with Sai College as Assistant Professor and head in Computer Science Department.

Second Author- Dr. (Prof.) Jyoti Singh received the MCA degree from Banasthali Vidyapith, Rajasthan in 1990. She has also done PhD in Computer Science and Application from Pt. Ravi Shankar Shukla University, Raipur in 2007.She worked as Dy. Registrar in Chhattisgarh Swami Vivekanand University(CSVTU) Bhilai since 2005. Act as Resource person in computer course under Canada-India Institute Industry Linkage Project. She is Life member of "The Indian Society for Technical Education. She is an approved PhD guide in CSVTU Bhilai, Dr. C V Raman University Bilaspur, Periyar University Salem, Vinayaka University Tamil Nadu.