© 2017 IJSRST | Volume 3 | Issue 4 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Use of Supervised Learning Technique in Student Performance Prediction

A. M. Chandrashekhar, Megha M. G.

Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering(SJCE),JSS S&T University Campus, Mysore, Karnataka, India

ABSTRACT

Predicting the performance of student plays an important role in educational environment. The educational database contains a huge amount of data. These database contain hidden information for evaluation and improvement of student's performance. The feasible technique to achieve this prediction is data mining. Personal, social, psychological and other environmental variables are the factors that affects the performance of student. There are many classifier technique that can be applied for predicting the performance of student. This study explores the impact of supervised learning technique for predicting the performance of student.

Keywords : Educational data mining, supervised learning technique C4.5, Multilayer perceptron, Naïve Bayes

I. INTRODUCTION

Data mining can be used to find the existing patterns and relationships. Data mining consists of various technique such as machine learning, statistics and visualization techniques to identify and extract knowledge. This technique can be applied on a large amount of data to identify hidden patterns and relationships which helps in making decision. It can be used to improve the quality of education. Mining in educational environment is called educational data mining. Data mining is the process of discovering the interesting pattern from the huge amount of data stored in database, data warehouse or other repositories.

The various objectives that are used to predict the performance of student includes:

- Generation of data source of predictive variables. As input to the model 12 variable are used whose attribute and coding is shown in the Table 1.
- Identifying the different attributes which effect's the student learning behavior and performance in their academics.
- Build prediction model using data mining technique on the basis of identified predictive attributes.
- Validating the developed model. This prediction improves the quality of education

Br	Variable	Coding	Br	Variable	Coding
1.	Gender (S)	A- Male B-Female	2.	Family (BCD)	Numeric value
3.	Distance (UAS)	Numeric value	4.	High school (VSS)	A – Grammar school B–High school for economics C-Rest
5.	GPA (PO)	Numeric value	6.	Entrance exam (URK)	Numeric values
7.	Scholarship (SS)	A–Not B– sometimes C-yes	8.	Time (VRI)	A–Less than 1 hour B–From 1 to 2 hours C–From 2 to 3 hours D–From 3 to 4 hours E–From 4 to 5 hours
9.	Materials (MAT)	A–book, B–The notes of other students, C – Notebook from the lectures, D-Notes edited or made by student E–All that is available to student	10.	The Internet (INT)	A – Yes B - No
11.	Grade importance (VO)	A–Not important at all, B–Not important C – Somewhat important D – Important E–Very important	12.	Earning (MPD)	Numeric values

Table 1. Student related attribute

II. DATA MINING METHOD

Data mining show one common feature that is identifying new relationships and dependencies of attributes in the observed pattern. The main goal of the analysis is to categories the data by class that is the data on which it belongs to particular class. In this the algorithm is divided into two groups:

- Unsupervised algorithm
- Supervised algorithm

Unsupervised algorithm

The algorithm is to identify the essential pattern in the data without the prior knowledge about which class the data belongs [4]. This technique finds the pattern and structure among all the variables. The model produced by the unsupervised learning algorithm can be used for prediction even though it is not designed for such task. Clustering and association are the example for this group.

Supervised algorithm

The algorithm consists of a target (or dependent variable) which is to be predicted from a given set of predictors (independent variables). With these set of variables, generating a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data Training machine learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training. The algorithm seeks a function from inputs to the respective targets.

There are various classifier technique are present and one cannot choose the best, because they vary from various aspects such as learning rate, classification speed, robustness etc. In this study investigation is made on three algorithm for intelligent data analysis: C4.5, multilayer perceptron, Naïve Bayes [1]. Classification model are built by using this algorithm whose aim is to predict the class to which new sample will belongs. The selection of this three algorithm technique is to identify the most suitable way to predict the student performance.

Naive Bayes algorithm (NB)

It is a method of classifying based on the theory of probability [2]. Bayesian theorem is called as naïve because it solves the problems relying on two important assumptions: it presumes that prognostic attribute are conditionally self-reliant with the familiar classification, and it assumes that there is no hidden attribute that affect the process of prediction. This classifier gives the approach to the probabilistic discovery of knowledge, [11] and it is also an efficient algorithm for classification.

Multilayer perceptron (MLP)

It is widely used and popular neural networks. The network which consists of collection of sensory elements which makes the input layer, one or more invisible layers of processing elements and output layer of processing elements. MLP is especially applicable for approximating a classification function.

C4.5

It is widely used decision tree algorithm. Professor Ross Quinlan introduced decision tree algorithm known as C4.5 in 1993[4]. C4.5 has many feature such as handling missing values, pruning of decision trees, derivation of rule, classification and others. C4.5 uses divided and conquer method.J48 algorithm is an implementation of C4.5 in Weka software tool [5]. Flowchart of decision tree is represented by tree structure. Every internal node is represented has the condition of attribute to be examined, branches of tree represents the result of the study [19]. Leaves represents the class to which sample belongs [20]. Decision tree algorithm is popular because of its ease of implementation and the result can be displayed graphically.

The robustness of the classifier can be determined by performing cross validation on the classifier. In this 3-fold cross validation is used: split the data set into three subsets of equal size [6]. Two subset is used for training one subset is for cross validating, and one for measuring the accuracy of prediction of the final constructed network. This procedure is performed three times so that each subset is tested at least once. The performance metrics can be calculated by using Weka software toolkit after running a specified K-fold cross validation [21].

III. EXPERIMENT RESULTS AND DISCUSSIONS

For the purposes of this study Weka software package was used, was developed at the University of Waikato in New Zealand. This package is developed using Java language. Today it stands as a most efficient and comprehensive package with machine learning algorithm. Test were conducted for the assessment of the input variables: Chi-square test, one R-test, Gain Ratio test, Info Gain Test [7]. The results of every test include the metrics like Attribute, Merit, Merit deviation, Rank, Rank and deviation. The result of all test and their average rank are shown in Table 2. The goal of this study is to determine the importance of each attribute individually. Where the PO (GPA) as more impact than follows URK (entrance exam), MAT (study material), VRI (weekly average hour dedicated for studies). [12] BCD (family size), UAS (distance from residency to institution) are the attribute which as less impact while predicting the performance.

Attribute	Chi-square	One R	Info Gain	Gain Ratio	AVG Range
PO	1,3	1	1,3	1	1,15
URK	1,7	8	1,7	2	3,35
MAT	4,7	6	4,7	4,3	4,93
VRI	3,7	10,3	3,3	4	5,33
SS	7,7	5	7,7	6	6,6
VO	5,7	10,3	5,3	6	6,83
MPD	5,7	9,3	5,7	6,7	6,85
INT	7	7	7,3	6,7	7
VSS	8,7	4	9	9	7,68
S	9	5,7	9	9,3	8,25
UAS	11	5	11	11	9,5
BCD	12	6,3	12	12	10,58

Table 2. The results of all tests and their average rank

Some of the experiment is carried out to evaluate the performance these three algorithm [14]. The performance of the model (NB, MLP, and J48) is evaluated based on three criteria: prediction accuracy, learning time and error rate which is shown in the following Table 3, 4, 5.

Classifier	TP	FP	Precision	Recall	Class
NB	0,500	0,149	0,517	0,500	A
	0,851	0,500	0,843	0,851	B
MLP	0,371	0,179	0,397	0,371	A
	0,821	0,629	0,804	0,821	B
J48	0,290	0,118	0,439	0,290	A
	0,882	0,710	0,796	0,882	B

Table 3. Comparison of evalution measures by class

The above Table shows the evalution measure of three algorithm where Naïve Bayes is better than other two.

Evaluation Criteria	Classifier		
	NB	MLP	J48
Kappa statistic	0,3552	0,1958	0,1949
Mean absolute squared (MAE)	0,263	0,2856	0,3255
Relative absolute Error(RAE)	71,73%	77,68%	88.53%
Root relative squared Error (RRSE)	98,25%	116,14%	103.55%
Root mean squared Error (RMSE)	0,4204	0,4969	0,4431

Table 4. Comparison of estimates

As shown in the Table 4 the estimation of three algorithm Where Naïve Bayes as low error rate compare to MLP and J48 [15]. The Table 5 shows the predictive performance of the algorithm. The Naïve Bayes takes less time to build model. [13] Where as MLP takes more time to bild model, Naïve Bayes Classify correctly compare to MLP and J48.

Evaluation Criteria	Classifier		
	NB	MLP	J48
Timing to build model (in sec)	0	4,13	0
Correctly classified instances	197	183	190
Incorrectly classified	60	74	67

Table 5. Predictive performance of the classifier

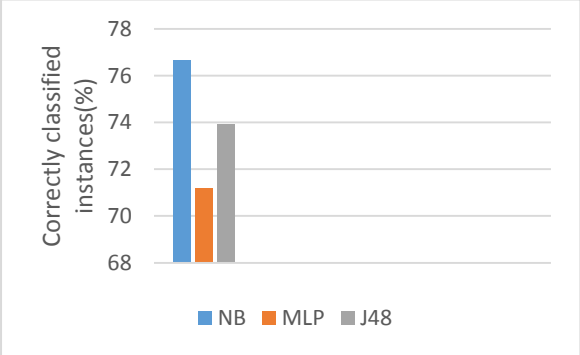

Figure 1. Prediction Accuracy

As shown in the figure 1 Naïve Bayes predict better than J48 and MLP [8]. The MLP as lowest accuracy compare to other two.

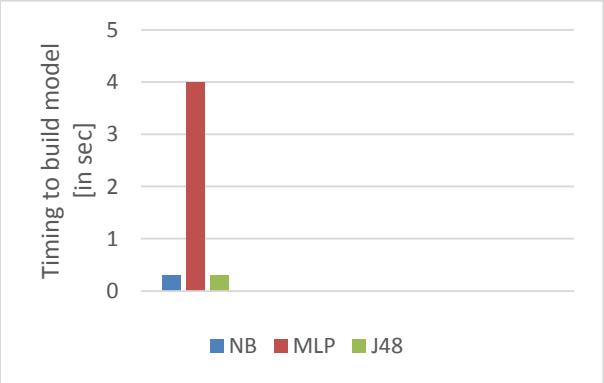

Figure 2. Learning time of three classifiers

As shown in below figure learning time of three classifier, MLP takes more time to build. Naïve Bayes learn rapidly in the time to build model for the given attribute. The Figure 3 shows correctly classified instances versus incorrectly classified instances [16]. The Naïve Bayes classifies more correctly compare to MLP and J48 [17].

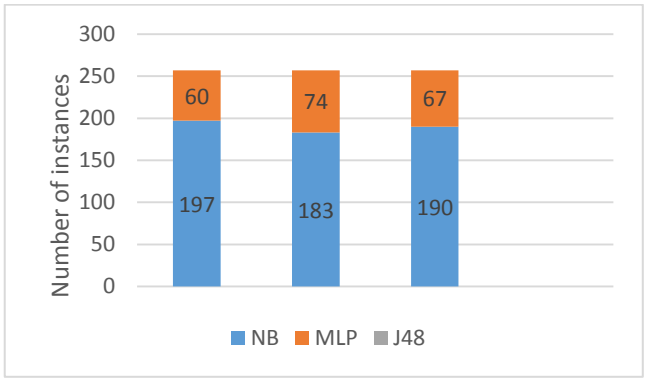

Figure 3. Error Rate

IV. CONCLUSION

In this paper, three supervised algorithm where applied to predict the performance of the student. Performance where evaluated based on the predictive accuracy, Error rate, learning time of the classifier, predictive performance of the classifier [9]. The result indicates that Naïve Bayes predict better than J48 and MLP [10]. Naïve Bayes is the best classifier model which is both accurate and comprehensive [18].

V. REFERENCES

[1]. Wu, X. & Kumar, V. (2009), the Top Ten Algoritms in Data Mining, Chapman and Hall, Boca Raton.

[2]. Witten, I.H. & Frank E. (2000), Data Mining Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco.

[3]. Cios, K.J., Pedrycz W., Swiniarski, R.W & Kurgan, L.A. (2007), Data Mining: A Knowledge Discovery Approach, Springer, New York.

[4]. Quinlan, J.R. (1993), C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco. 5Kumar, V. and Chadha, A. (2011) "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.

[5]. A.M. Chandrashekhar and K. Raghuveer , "Improvising Intrusion detection precision of ANN based NIDS by incorporating various data Normalization Technique – A Performance Appraisal", IJREAT International Journal of

Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014.

[6]. A. M Chandrashekhar and K. Raghuveer, "Diverse and Conglomerate modi-operandi for Anomaly Intrusion Detection Systems", International Journal of Computer Application (IJCA) Special Issue on "Network Security and Cryptography (NSC)", 2011.

[7]. A. M. Chandrashekhar and K. Raghuveer, "Confederation of FCM Clustering, ANN and SVM Techniques of Data mining to Implement Hybrid NIDS Using Corrected KDD Cup Dataset", Communication and Signal Processing (ICCSP) IEEE International Conference,2014, Page 672-676.

[8]. A. M Chandrashekhar and K. Raghuveer, "Hard Clustering vs. Soft Clustering: A Close Contest for Attaining Supremacy in Hybrid NIDS Development", Proceedings of International Conference on Communication and Computing (ICCC - 2014), Elsevier science and Technology Publications.

[9]. A. M. Chandrashekhar and K. Raghuveer, "Amalgamation of K-means clustering algorithem with standard MLP and SVM based neural networks to implement network intrusion detection system", Advanced Computing, Networking, and Informatics –Volume 2(June 2014), Volume 28 of the series Smart Inovation, Systems and Technologies pp 273-283.

[10]. A. M. Chandrashekhar and K. Raghuveer, "Fusion of Multiple Data Mining Techniques for Effective Network Intrusion Detection – A Contemporary Approach", Proceedings of Fifth International Conference on Security of Information and Networks (SIN 2012), 2012, Page 178-182.

[11]. A. M. Chandrashekhar, Jagadish Revapgol, Vinayaka Pattanashetti, "Big Data Security Issues in Networking", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Volume 2, Issue 1, JAN-2016.

[12]. P.Koushik, A.M.Chandrashekhar, Jagadeesh Takkalakaki, "Information security threats, awareness and cognizance" International Journal for Technical research in Engineering (IJTRE), Volume 2, Issue 9, May 2015.

[13]. A.M.Chandrashekhar, Yadunandan Huded, H S Sachin Kumar, "Advances in Information security risk practices" International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[14]. A. M. Chandrashekhar,Muktha G, Anjana D, "Cyberstalking and Cyberbullying: Effects and prevention measures", Imperial Journal of Interdisciplinary Research (IJIR), Volume 2, Issue 2, JAN-2016.

[15]. A.M.Chandrashekhar, Syed Tahseen Ahmed, Rahul N, "Analysis of Security Threats to Database Storage Systems" International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[16]. A.M.Chandrashekhar, K.K. Sowmyashree, RS Sheethal, "Pyramidal aggregation on Communication security" International Journal of Advanced Research in Computer Science and Applications (IJARCSA), Volume 3, Issue 5, May 2015.

[17]. A.M.Chandrashekhar, Rahil kumar Gupta, Shivaraj H. P, "Role of information security awareness in success of an organization" International Journal of Research(IJR), Volume 2, Issue 6, May 2015.

[18]. A.M.Chandrashekhar, Huda Mirza Saifuddin, Spoorthi B.S, "Exploration of the ingredients of original security" International Journal of Advanced Research in Computer Science and Applications(IJARCSA), Volume 3, Issue 5, May 2015.

[19]. A.M.Chandrasekhar, Ngaveni Bhavi, Pushpanjali M K, "Hierarchical Group Communication Security", International journal of Advanced research in Computer science and Applications (IJARCSA), Volume 4, Issue 1,Feb-2016.

[20]. Surjeet Kumar Yadav and Saurabh Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), Vol.2, No. 2, 51-56, 2012.