Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 500 – 508

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015)

Classification and prediction based data mining algorithms to predict slow learners in education sector

Parneet Kaur[a],Manpreet Singh[b],Gurpreet Singh Josan[c]

[a]*Scholar, Department of CSE, Punjab Technical University,Jalandhar 144603,India*
[b]*Assistant Professor, Department CSE&IT, GNDEC, Ludhiana, Punjab, India*
[c]*Assistant Professor, Department of CSE & IT, Punjabi University, Patiala, Punjab, India.*

Abstract

Educational Data Mining field concentrate on Prediction more often as compare to generate exact results for future purpose. In order to keep a check on the changes occurring in curriculum patterns, a regular analysis is must of educational databases. This paper focus on identifying the slow learners among students and displaying it by a predictive data mining model using classification based algorithms. Real World data set from a high school is taken and filtration of desired potential variables is done using WEKA an Open Source Tool. The dataset of student academic records is tested and applied on various classification algorithms such as Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA an Open source tool. As a result, statistics are generated based on all classification algorithms and comparison of all five classifiers is also done in order to predict the accuracy and to find the best performing classification algorithm among all. In this paper, a knowledge flow model is also shown among all five classifiers. This paper showcases the importance of Prediction and Classification based data mining algorithms in the field of education and also presents some promising future lines.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: Educational Data Mining; Knowledge Discovery; Classification; Attribute Evaluator.

1. Introduction

Data mining has attracted lot of attention in the research industry and in society as a whole in recent years, due to enormous availability of large amount of data and the need for turning such data into useful information and knowledge. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering new and potentially useful information from huge databases [12].

* Corresponding author. Tel.: +91-9466688831; fax: 0171-2822002.
E-mail address: kaur.parneet@gmail.com

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.372

Educational Data Mining (EDM) is the application of Data Mining techniques on educational data. The objective of EDM is to analyze such data and to resolve educational research issues. EDM deals with developing new methods to explore the educational data, and using Data Mining methods to better understand student learning environment [1-4]. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. Educational Data Mining researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning ,computer-adaptive testing (and testing more broadly), and the factors that are associated with student failure or non-retention in courses[6,8]. Some other key areas include improvement of student models; application of EDM methods has been in discovering or improving models of a domains knowledge structure and studying pedagogical support (both in learning software, and in other domains, such as collaborative learning behaviours). There are increasing research interests in using data mining in education. This new emerging field, called educational data mining, concerns with developing methods that discover knowledge from data originating from educational environments. Educational data mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K-nearest neighbour and many others. Prediction and analysis of student performance is an important milestone in educational environment. Student's academic performance is a crucial factor in building their future[4,5]. Academic performance of student is not a result of only one deciding factor besides it heavily hinges on various factors like personal, socio-economic, psychological and other environmental variables. This paper identifies the factors associated with students whose academic performance is not good and to improve the quality of education by identifying slow learners so that teachers can assist them individually to improve their performance. Through this paper, the accuracy of some classification techniques for predicting performance of a student is also investigated. The main objectives of this work are: to generate data source of predictive variables, Data mining methodologies to study student performance at high school level, identification of the slow learners performance, identification of the highly influencing predictive variables on the academic performance of high school students and to find the best classification algorithm.

Nomenclature	
EDM	Education Data Mining
SMO	Sequential minimal optimization
J48	Decision Tree Algorithm
REPTree	Reduced Error Pruning Decision Tree
WEKA	Waikato Environment for Knowledge Analysis

2. Background and prior work

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Educational Data Mining (EDM) is still in its infancy [15]. The field of EDM is new and emerging in the field of education sector which can also be applied in other areas like sports, accounts, transportation etc. The international working group in EDM established the Journal of Educational Data Mining (2009) and a yearly international conference that began in 2008. Han and Kamber [17] 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. M.Ramaswami and R.Bhaskaran [12] applied CHAID prediction model to analyze the interrelation between variables that are used to predict the outcome of the performance at higher secondary school education. The CHAID prediction model of student performance was constructed with seven class predictor variable. Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme [13] used machine learning techniques to improve the prediction results of

academic performances in real case studies. Three methods have been used by them to deal with the class imbalance problem and all of them show satisfactory results. They first balanced the datasets and used both cost-insensitive and sensitive learning with SVM for the small datasets with Decision Tree for the larger datasets. Arockiam et al. [14] implemented FP Tree and K-means clustering technique for finding the similarity between urban and rural students programming skills. FP Tree mining was applied to sieve the patterns from the dataset. K-means clustering was used to determine the programming skills of the students. The study clearly indicates that the rural and the urban students differ in their programming skills and found that huge proportions of urban students were good in programming skill compared to rural students. It divulges that academicians provide extra training to urban students in the programming subject. Cortez and Silva [15] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset. Galit [18] gave a case study that uses student's data to analyze their learning behaviour to predict the results and to warn students at risk before their final exams. V.Ramesh et al [16] tries to identify the factors influencing the performance of students in final examination. They adopted survey cum experimental methodology to generate the database. The algorithms which were used by them for implementation were Naïve Bayes, Multi Layer Perception, SMO, J48, and REPTree. The obtained results from hypothesis testing reveals that type of school is not influence student performance but parent's occupation plays a major role in predicting grades.

2.1 Proposed Methodology

A survey cum experimental methodology is used. Through extensive search of the literature and discussion with experts on student performance, a number of factors that are considered to have influence on the performance of a student are identified. These influencing factors are categorized as input variables. For this work, recent real world data is collected from high school. This data is then filtered out using manual techniques. Then data is transformed into a standard format required by the WEKA tool. After that, features and parameters selection is identified.

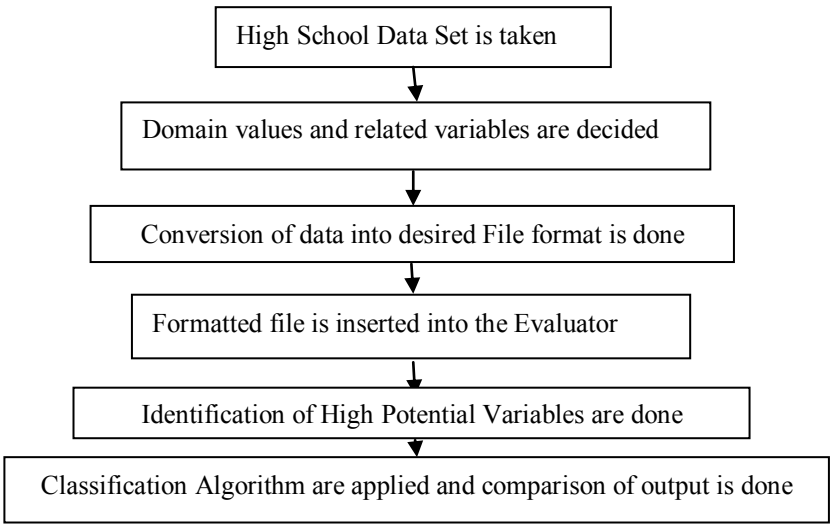

Fig 1. Flowchart of proposed work

Then analysis of identified parameters and implementation is performed on the tool. After implementation results are produced and analyzed. Stepwise description of methodology used is represented with the help of flowchart as shown in Figure 1.A record 152 students of high school is used as dataset and student related variables are defined in the Table 1 along with their domain values.

Table 1. Student related variables

Variable Name	Description	Domain
SEX	Student's Sex	{M,F}
INS-HIGH	Institution at high level	{Private, Government}
TOB	Type of board	{State Board, CBSE}
MOI	Medium of instruction	{Hindi, English}
TOS	Type of school	{Co-ed, Boys, Girls}
PTUI	Private tuition	{Yes, No}
S-AREA	Area at school level	{Urban, Rural}
MOB	Student having mobile	{Yes, No}
COM-HM	Computer at home	{Yes, No}
NETACS	Student having net access	{Yes, No}
ROLL NO.	Student's roll no	Given by school authority
INT-GR	Internal grade of student	{A+, A, B, C}
ATDN	Attendance count	Based on School Attendance count
CLASS(Response Variable)	Whether qualified or not	{NQ, Q}

2.2 Tools and Techniques used

In this paper variety of Data Mining techniques are used for prediction of slow learners in Educational Data Mining. The techniques are Classification, Regression and Density Estimation. During this work, Classification techniques for prediction are used. The output dataset is tested and analyzed with five Classification algorithms which are Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree. For implementation of all these classification tasks we have used WEKA workbench.

3. Simulation Case Study

A total of 152 records are taken for the analysis of this research. In this paper, selected high potential variables using select attributes facility of WEKA is done. For attribute evaluation, Chi Squared attribute, Info Gain attribute, Symmeterical Uncert attribute and ReliefF attribute evaluator are used. To rank variables Ranker Search method technique of WEKA is also applied. Final ranks are generated (using cross validation) manually by taking average. High potential variables are listed below along with their ranks in Table 2.

Table 2. High potential variables

Name of the Variable	Rank Values
INT-GR	1.65
ATDN	2.225
SEX	3.6
PTUI	3.525
MOB	5.375
INS-HIGH	5.925
COM-HM	8.325
NET-ACS	9.2

4. Results

The dataset during this work is tested and analyze with five Classification algorithms those are Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree(using cross validation). All the statistics results are provided in Table 3. Also a comparison of accuracy of all classifiers is done and finally it has been investigated that Multi Layer Perception technique performs best with accuracy 75%. The accuracy level of all the algorithms are given below in Table 4.

Table 3. Statistical Analysis of Classifiers with Cross Validation

Name of Classification Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Multilayer Perception	NQ	0.83	0.44	0.807	0.83	0.822	0.77
	Q	0.55	0.162	0.605	0.553	0.578	0.773
Naive Bayes	NQ	0.76	0.596	0.741	0.762	0.751	0.648
	Q	0.40	0.238	0.432	0.404	0.418	0.648
SMO	NQ	0.88	0.766	0.721	0.886	0.795	0.56
	Q	0.23	0.114	0.478	0.234	0.314	0.56
J48	NQ	0.81	0.574	0.761	0.819	0.789	0.713
	Q	0.42	0.181	0.513	0.426	0.465	0.713
REPTree	NQ	0.83	0.681	0.733	0.838	0.762	0.667
	Q	0.31	0.162	0.469	0.319	0.38	0.667

Table 4. Comparison of Classifiers on the basis of Correctly Classified Instances with Cross Validation

Mining Technique	Accuracy
Multilayer Perception	75%
Naïve Bayes	65.13%
SMO	68.42%
J48	69.73%
REPTree	67.76%

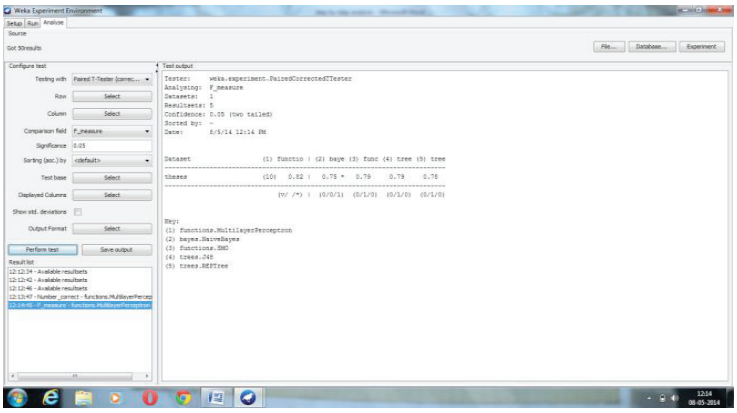

Fig.2. Comparison of Classifiers with use of WEKA Experimenter

Comparison of all classifiers with the help of WEKA Experimenter is shown in fig. 2. In this case also Multi Layer Perception performs best among all classifiers with F-Measure 82%. A model performance chart is also created using GUI of WEKA. The performance comparison on the basis of accuracy among algorithms is shown in the figure 3. Also, knowledge flow model shown in fig.4 which shows membership tree structure.

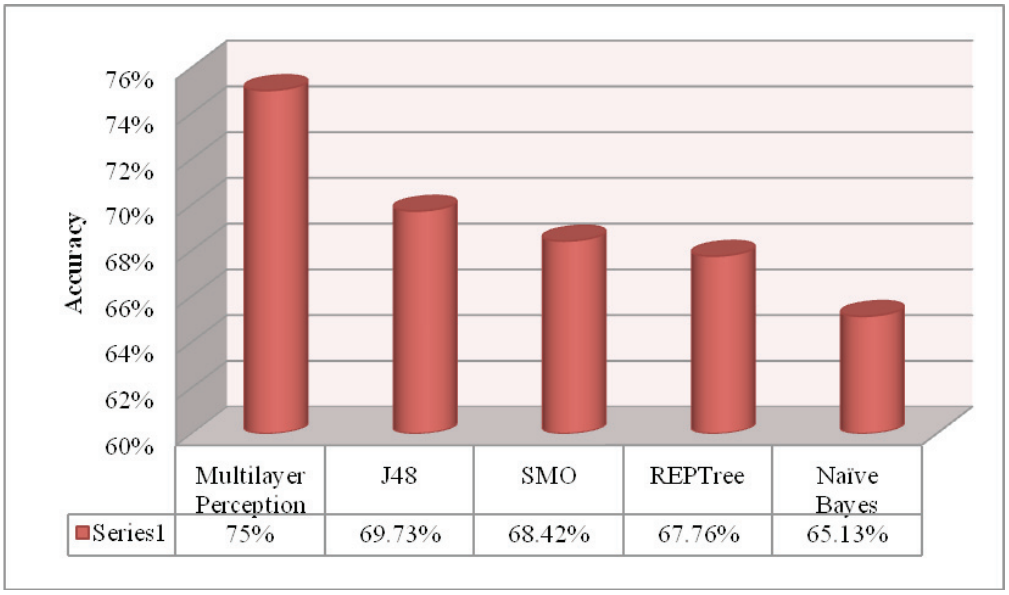

Fig.3. Accuracy Comparison of Classifiers

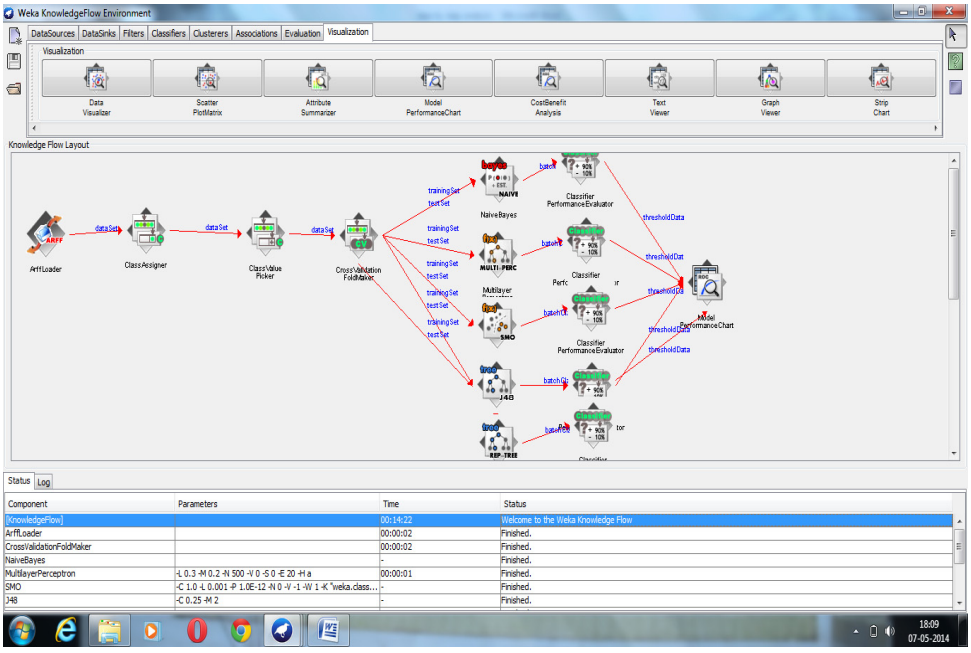

Fig. 4. Knowledge flow model

After loading the data file run the model and a model performance chart as shown in figure 5 for multiple classifiers such as Naive Bayes, Multilayer Perceptron, SMO, J48 and REPTree. Figure 5 shows region of convergence curve (ROC) for each classifier.

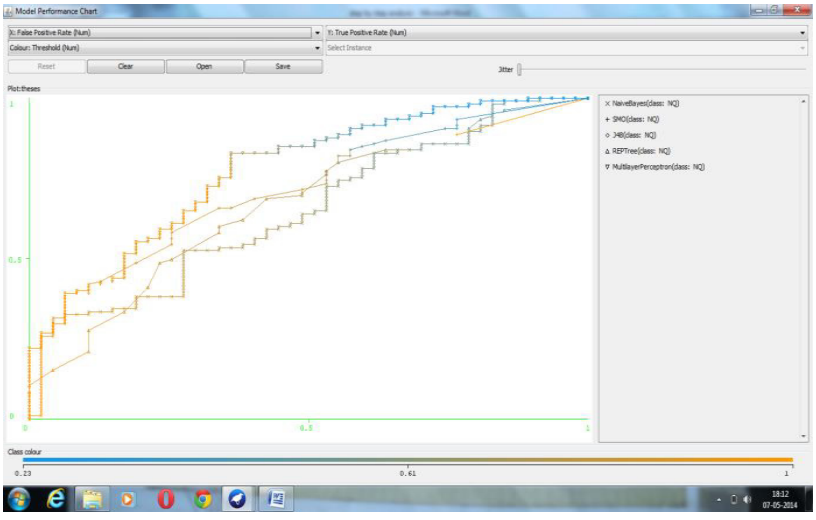

Fig.5. Model performance chart

5. Conclusion and Future Scope

In this paper, classification techniques are used for prediction on the dataset of 152 students, to predict and analyze student's performance as well slow learners among them. In this study, a model was developed based on some selected student related input variables collected from real world (high schools). Among all data mining classifiers Multi Layer Perception performs best with 75% accuracy and therefore MLP proves to be potentially effective and efficient classifier algorithm. Also comparison of all 5 classifiers with the help of WEKA experimenter is also done, in this case also MLP proves to be best with F-measure of 82%. Therefore, performance of MLP is relatively higher than other classifiers. A model performance chart is also plotted. This research help the institutions to identify students who are slow learners which further provide base for deciding special aid to them. EDM is in its infancy and it has lot of potential for education.EDM opens promising and exciting avenues for future research. In future, Integration of data mining techniques with DBMS and E-learning techniques is merged together on different datasets to find accuracy and predictions of desired results. Also, EDM tools are easy to understand and interfaced with various techniques. Educators with no expertise in data mining can also apply their hands in these fields. Also some new factors can be applied to improve the student's performance, learning and retention capabilities among them. Hence the future of EDM is promising for further research and can be applied in other areas like medicine, sports, and share market due to the availability of huge databases.

Acknowledgements

This research paper is truly contribution and guidance who belong to the list of author. The authors pay homage to Punjab Technical University for continuous encouragement, guidance and support during their work.

References

[1] Cristobal Romero (2010), "Educational Data Mining: A Review of the State-of-the-Art", IEEE Transactions on systems , man and cybernetics- Part C: Applications and Reviews vol. 40 issue 6, pp 601 – 618.

[2] Zaïane, O. (2001)," Web usage mining for a better web-based learning environment", Proceedings Of Conference on Advanced Technology For Education, 60-64.

[3] Zaïane, O. (2002), "Building a recommender agent for e-learning systems". Proceedings of the International Conference on Computers in Education,55–59.

[4] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004), "Detecting StudentMisuse of Intelligent Tutoring Systems". Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[5] Tang, T., McCalla, G. (2005)," Smart recommendation for an evolving e-learning system: architecture and experiment", International Journal on E-Learning, vol. 4, issue1, 105–129.

[6] Merceron, A., Yacef, K. (2003)," A web-based tutoring tool with mining facilities to improve learning and teaching". Proceedings of the 11th International Conference on Artificial Intelligence in Education, 201–208

[7] Romero, C., Ventura, S., de Bra, P., & Castro, C. (2003)," Discovering prediction rules in aha! Courses". Proceedings of the International Conference on User Modelling, 25–34.

[8] Beck, J., & Woolf, B. (2000)." High-level student modelling with machine learning". Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 584–593.

[9] Dringus, L.P., Ellis, T. (2005)," Using data mining as a strategy for assessing asynchronous discussion forums", Computer and Education Journal , 45, 141–160.

[10] M.Ramaswami and R.Bhaskaran(2010), "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues Vol. 7, Issue 1, pp 10-18.

[11] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme(2009), "Improving Academic Performance Prediction by Dealing with Class Imbalance", Ninth International Conference on Intelligent Systems Design and Applications,

[12] L.Arockiam, S.Charles,Arulkumar et.al(2010), "Deriving Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, pp 687-690.

[13] P. Cortez, and A. Silva(2008), "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), pp 5-12.

[14] V.Ramesh, P.Parkavi, K.Ramar(2013),"Predicting student performance: A statistical and data mining approach", International journal of computer applications , Volume 63- no. 8, pp 35-39.

[15] Jiawei Han Michelin Kamber(2011), "Data Mining-Concepts and Techniques", Morgan Kaufmann Publishers.