Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.2, 2018, pp.44-47

DOI: https://doi.org/10.51983/ajcst-2018.7.2.1874

Using Predictive Modeling System and Ensemble Method to Ameliorate Classification Accuracy in EDM

Mudasir Ashraf[1], Majid Zaman[2] and Muheet Ahmed[3]
[1&3]Department of Computer Science, [2]Directorate of IT&SS, Department of Computer Science,
University of Kashmir, Jammu and Kashmir, India
E-Mail: mudasir04@gmail.com, zamanmajid@gmail.com, ermuheet@gmail.com
(Received 5 June 2018; Revised 25 June 2018; Accepted 16 July 2018; Available online 24 July 2018)

Abstract **- Educational data mining has illustrated an increasing demand for extracting and maneuvering data from academic backdrop, to generate prolific information which is indispensible for decision making. Therefore in this paper, an attempt has been made to deploy various data mining techniques including base and meta learning classifiers across our pedagogical dataset to foretell the performance of students. Among several contemporary ensemble approaches, researchers have practiced widespread learning classifiers viz. boosting to predict the performance of students. As exploitation of ensemble methods is considered to be significant phenomenon in classification and prediction mechanisms, therefore analogous method (boosting) has been applied across our pedagogical dataset. The entire results have been evaluated with 10-fold cross validation, once pedagogical dataset has been subjected to base classifiers including j48, random tree, naive bayes and knn. In addition, techniques such as oversampling (SMOTE) and undersampling (Spread subsampling) have been employed to further draw a comparison among ensemble classifiers and base classifiers. These methods were exploited with the key objective to observe any improvement in prediction accuracy of students.**
Keywords: **Educational Data Mining, Boosting, Random Tree, Naive Bayes, j48 and K-Nearest Neighbour**

I. INTRODUCTION

EDM is an application of Data mining (DM) that endeavours to estimate the educational data issues by undertaking existing techniques of DM into consideration [1]. Educational data mining (EDM) contemplates to interdisciplinary study that deliberates on the development and improvement of diverse methods to ascertain the academic information produced from heterogeneous sources. The mined data can be suitable to upgrade teaching, learning experiences and accordingly aid in refining the institutional effectiveness. In yester years, it has witnessed impulsive growth in both the fields of software and databases associated with student's information which primarily signify their learning process [2], and this has proved to be a gold mine in the direction of academic research [3].

To determine the student's performance is always considered a tricky task for the reason that his/her performance is based on number of parameters such as character, educational environment, demographics, emotional and other variables. The relationships among these fields are not apparently implicit as they are usually related in intricate nonlinear mode. Furthermore, a variety of data mining techniques in the realm of EDM have been consumed and applied to explore educational data and identify variables responsible for better academic achievements, but there are still deficiencies and it calls for the application of latest data mining tools.

EDM is a subset of DM which encompasses the data that comes from educational background and centers on the development of various techniques and ascertaining various patterns that are exclusive in nature [4]. The identified patterns can be useful for academic stakeholders for decision making, to ameliorate student's performance and to devise healthier teaching and learning strategies. EDM processes raw data coming from educational systems into effectual knowledge that can possibly have a considerable impact on academic strength [5]. EDM is also inexhaustible in model designs, methods and algorithms to investigate academic data [6].

II. LITERATURE REVIEW

Romero et .al (2013) applied Multiple Regression Model (MRM) and support vector machine (SVM) to forecast by and large the individual academic performance of students [7]. Also, Kotsiantis (2012) employed regression method to predict the student's marks in a distance learning system [8]. Noah, Barida and Egertib (2013) studied various parameters associated with the student using regression, k-means and neural networks to identify weak candidates for the purpose of performance enhancement [9]. Moreover, Bichkar (2014) described regression as statistical method of forecasting students' performance based on fields acquired from dataset [10].

Junco et al. (2011) applied ANOVA to calculate the impact on learning outcome and student engagement using twitter [13]. Stafford et al. (2014) examined the wiki activity indicators and the final grades of the students [14]. The results showed that there was significant correlation among the two variables and students who were engaged with wiki activities acquired better overall score. Giovanella et al. (2013) also investigated vigorous participation of students in various social media applications such as wiki, blog,

Delicious and Twitter as a promising learning performance indicator [15].

Suyal and Mohod (2014) investigated students who necessitate special attention by studying the relationship among different attributes using association rule mining [16]. Baderwah and pal (2011) used mining technique namely decision tree on fields such as class test, attendance, assignment and semester marks for early detection of students who are at risk [17]. A number of efforts have been done in this direction and a research team comprising of Jeevalatha, Ananthi, and Saravana (2016) applied decision tree on undergraduate students dataset covering a set of factors such as Communication skills, higher secondary marks and undergraduate marks for performance assessment and placement selection [18]. In addition, Baker and Yacef (2009) conducted a survey on various techniques applied in the field of EDM. They came to the final conclusion that considerable work has been done in the direction of prediction rather than relationship mining [19].

Apart from the statistical measure viz Regression Model employed by a number of researchers, Jothi and venkatalakshmi (2014) made new strides and used clustering technique for analysing and predicting student's performance to improve the student's success rate [11]. Sheik and Gadage (2015) endeavoured to investigate the learning behaviour of various models with the help of various open source tools to get an insight of how these models train, evaluate and predict the performance of students [12].

III. EXPERIMENTAL SETUP

In this subsection, the researchers have applied various base classifiers with boosting approach on academic dataset to corroborate any improvement in classification accuracy of different learning algorithms viz. j48, random tree, naïve bayes and knn. Moreover, the dataset was subjected to techniques such as SMOTE and Spread subsampling to further investigate whether there is any considerable improvement in prediction accuracy of learning classifiers with boosting method.

A. Boosting Approach

Under boosting technique, we introduced the set of classifiers on pedagogical dataset viz. j48, random tree, naïve bayes and knn to predict the performance of students. While exercising boosting method without application of filtering procedures (undersampling and oversampling techniques) on diverse classifiers, we examined that prediction paradigm of j48 over performed in assessment to other models, wherein the accurateness of classifying correct instances was boosted with 95.32% and subsequently observed least misclassification error of 4.67%. The findings of boosting learning classifier are are characterized in table I.

After analyzing results in the below mentioned table I, it is evident that various performance estimates including Tp rate, Fp rate, precision, recall, f-measure, ROC area and relative absolute error associated with the classifiers viz. j48 and naïve bayes have exemplified momentous results when these base classifiers were subjected to ensemble learning via boosting. Furthermore, the performance of random tree has been least significant in contrast to remaining classifiers, and as a matter of consequence, the alternative performance factors such as Tp rate, Fp rate and so on related with the base classifier have also demonstrated low performance while forecasting the correct instances.

TABLE I RESULTS AFTER EMPLOYING BOOSTING APPROACH

Classifier Name	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Rel. Abs. Err.
Boost. with J48	95.32%	4.67%	0.953	0.032	0.955	0.953	0.954	0.995	7.29%
Boost. with Random tree	89.35%	10.64%	0.894	0.073	0.896	0.894	0.895	0.910	16.91%
Boost. with Naïve Bayes	95.04%	4.95%	0.950	0.033	0.951	0.950	0.951	0.988	9.99%
Boost. with KNN	93.59%	6.40%	0.936	0.044	0.937	0.936	0.936	0.948	10.36%

B. Boosting after SMOTE

The oversampling technique namely SMOTE once employed to various individual base classifiers viz. j48, random tree, naïve bayes and knn with boosting approach have exhibited discrepancies in results. The base classifier j48 and naïve bayes with boosting method have demonstrated notable achievements in prediction outcomes when subjected to over sampling technique, in contrast to boosting system without SMOTE (Table II and I depicts results after SMOTE and prior to application of SMOTE respectively). Both classifiers performance amplified from 95.32% to 96.44% (j48 with boosting) and 95.04% to 96.06% (naïve bayes with boosting). However, learning ensembles of each base classifiers viz. random tree and knn also demonstrated substantial improvement in its prediction outcomes from 89.35% to 92.03% (random tree with boosting) and 93.59% to 94.49% (knn with boosting).

TABLE II EXHIBITS RESULTS AFTER OVERSAMPLING METHOD

Classifier Name	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Rel. Abs. Err.
Boost. with J48	96.44%	3.55%	0.964	0.019	0.965	0.964	0.963	0.996	5.35%
Boost. with Random tree	92.03%	7.96%	0.920	0.043	0.921	0.922	0.921	0.938	11.98%
Boost. with Naïve Bayes	96.06%	3.93%	0.961	0.021	0.962	0.961	0.960	0.991	8.26%
Boost. with KNN	94.49%	5.50%	0.945	0.029	0.946	0.946	0.944	0.960	8.41%

C. Boosting after Spread Subsampling

In case of boosting with undersampling technique, the results of diverse base classifiers are furnished in table III. From the findings exemplified in table III and I, it is apparent only single base classifier viz. knn has explained decline in its prediction accuracy from 93.59% to 92.99% using undersampling method. Nevertheless, boosting with other classifiers such as j48 (95.32% to 95.54%), naïve bayes (95.04% to 95.85%) and random tree (89.35%91.61%) have illustrated significant results, and consequently paramount development was experienced across entire performance estimates associated with the classifiers viz. Tp rate, Fp rate, Precision , recall, f- measure and so on.

TABLE III ILLUSTRATES RESULTS AFTER UNDERSAMPLING

Classifier Name	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Rel. Abs. Err.
Boost. with J48	95.54%	4.45%	0.955	0.22	0.956	0.956	0.954	0.996	6.77%
Boost. with Random tree	91.61%	8.38%	0.916	0.042	0.917	0.916	0.915	0.937	12.57%
Boost. with Naïve Bayes	95.85%	4.14%	0.959	0.021	0.960	0.961	0.959	0.992	7.42%
Boost. with KNN	92.99%	7.00%	0.930	0.035	0.932	0.929	0.930	0.985	10.76%

The histograms in figure 1 exposes classification accuracy and relative absolute error of various classifiers using boosting approach. As per the below referenced figure, it publicizes three types of results viz. boosting with each base classifiers prior to filtering procedure, results attained post application of SMOTE and spread subsampling. Primarily, the histograms in figure 6 demonstrate prior application of filtering process in which learning classifier such as j48 with boosting has achieved paramount accuracy of 95.32% among other learning classifiers. Whereas, post deployment of SMOTE and Spread subsampling, j48 and naïve bayes with boosting have attained exceptional prediction accuracy of 96.44% and 95.85% respectively. Moreover, from the figure it is oblivious that the minimum relative error including all phases such as results accomplished ahead of filtering process, post SMOTE and Spread subsampling, has been found with two classifiers namely j48 and naïve bayes.

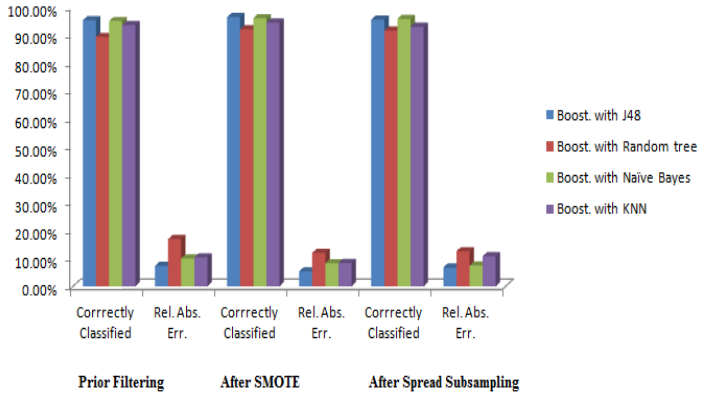

Fig. 1 Accuracy of Classifiers After Pertinence of Three Approaches

IV. CONCLUSION

Ensemble learning classifiers have exceedingly been imperative in predicting the performance of students over the past decade. The ensemble approach has more often fetched significant and accurate results in classifying the correct instances, in contrast to individual base learning algorithms. Moreover in this study, we employed most popular ensemble method viz. boosting on our academic dataset with the aim of predicting the performance of students. Typically, the meta learning algorithm viz boosting was deployed with several base classifiers including j48, random tree, naïve bayes and knn. Among all classifiers with boosting approach, j48 demonstrated compelling performance of 95.32% in predicting the exact instances. Furthermore, after the dataset was subjected to filtering procedures of SMOTE and Spread subsampling, j48 again illustrated substantial prediction accuracy of 96.44% when applied to oversampling method, and naïve bayes exhibited 95.85% accuracy in predicting the precise instances than other learning classifiers when employed with undersampling technique. Moreover, as per the observed statistical figures presented in miscellaneous tables, it was concluded that oversampling approach demonstrated significant results in predicting the outcome of students than other techniques employed.

REFERENCES

[1] T. Barnes, M. Desmarais, C. Romero and S. Ventura, "Educational Data Mining 2009", *Proc. of the 2nd Int. Conf. on Educational Data Mining*, Cordoba, Spain, 2009.

[2] R. S. J. D. Baker, "Data Mining for Education", *Int Encyclopedia Educ*, Vol. 7, pp. 112-118, 2010.

[3] J. Mostow and J. Beck, "Some Useful Tactics to Modify, Map and Mine Data from Intelligent Tutors", *Nat Lang Eng*, 2006; Vol. 12, No. 02):195–208.

[4] C. Romero, S. Ventura, "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, No. 6, November 2010.

[5] C. Romero and S. Ventura, "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, pp. 601-618, 2010.

[6] Ganesh, S. Hari and A. Joy Christy, "Applications of Educational Data Mining: A Survey", *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE*, 2015.

[7] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna and Sebastian Ventura, "Predicting Students' Final Performance From Participation In On-Line Discussion Forums", *Computers & Education*, Vol. 68, pp. 458-472, October 2013.

[8] Sotiris B Kotsiantis, "Use of Machine Learning Techniques for Educational Proposes: A Decision Support System for Forecasting Students Grades", *Artificial Intelligence Review*, Vol. 37, No. 4, pp. 331-344, 2012.

[9] OTOBO Firstman, BAAH Barida Noah and Taylor Onate Egerton, "Evaluation of Student Performance Using Data Mining over A Given Data Space", *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 2, No. 4, pp. 2277-3878, September 2013.

[10] R. S. Bichkar "Predicting Students Academic Perfonnance Using Education Data Mining", *World Journal of Computer Application and Technology*, Vol. 2, No. 2, pp. 43-47, 2014.

[11] J. K. Jothi and K. Venkatalakshmi, "Intellectual Performance Analysis of Students by Using Data Mining Techniques", *International Journal of Innovative Research in Science Engineering and Technology*, Vol. 3, No. 3, March 2014.

[12] Shelke Nikitaben and Gadage Shriniwas, "A Survey of Data Mining Approaches In Performance Analysis and Evaluation", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, 2015.

[13] R. Junco, G. Heiberger and E. Loken, "The Effect of Twitter on College Student Engagement and Grades", *Journal of Computer Assisted Learning*, Blackwell Publishing Ltd, Vol. 27, No. 2, pp. 119-132, 2011.

[14] T. Stafford, H. Elgueta and H. Cameron, "Students' Engagement with A Collaborative Wiki Tool Predicts Enhanced Written Exam Performance", *Research in Learning Technology*, Vol. 22, 2014.

[15] C. Giovannella, E. Popescu and F. Scaccia, "A PCA Study of Student Performance Indicators in a Web 2.0-based Learning Environment", *Proc of the 13th IEEE International Conference on Advanced Learning Technologies, ICALT*, pp. 33-35, 2013.

[16] Sayali Rajesh Suyal and Mohini Mukund Mohod, "Quality Improvisation of Student Performance Using Data Mining Techniques", *International Journal of Scientific and Research Publications*, Vol. 4, No. 4, April 2014.

[17] Baradwaj Brijesh Kumar and Pal Saurabh, "Mining Educational Data To Analyze Students' Performance", *International Journal of Advanced Computer Science and Applications, (IJACSA)*, Vol. 2, No. 6, 2011.

[18] T. Devasia, T. P. Vinushree and V. Hegde, "Prediction of Students Performance Using Educational Data Mining", *International Conference on Data Mining and Advanced Computing (SAPIENCE), IEEE*, pp. 91-95, 2016, March.

[19] J.D.B. Baker Ryan, Yacef Kalina, "The State of Educational Data Mining in 2009: A Review And Future Revisions", *Journal of Educational Data Mining*, Vol. 1, No. 1, February 2009.