An Ensemble-based Decision Tree Approach for Educational Data Mining

Moloud Abdar
School of Computer Science and Engineering
The University of Aizu
Aizu-Wakamatsu, Japan
m.abdar1987@gmail.com

Mariam Zomorodi-Moghadam
Department of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
m_zomorodi@um.ac.ir

Xujuan Zhou
School of Management and Enterprise
University of Southern Queensland
Toowoomba, Australia
Xujuan.Zhou@usq.edu.au

***Abstract*—Nowadays, data mining and machine learning techniques are applied to a variety of different topics (e. g., healthcare and disease, security, decision support, sentiment analysis, education, etc.). Educational data mining investigates the performance of students and gives solutions to enhance the quality of education. The aim of this study is to use different data mining and machine learning algorithms on actual data sets related to students. To this end, we apply two decision tree methods. The methods can create several simple and understandable rules . Moreover, the performance of a decision tree is optimized by using an ensemble technique named Rotation Forest algorithm. Our findings indicate that the Rotation Forest algorithm can enhance the performance of decision trees in terms of different metrics. In addition, we found that the size of tree generated by decision trees ensemble were bigger than simple ones. This means that the proposed methodology can reveal more information concerning simple rules.**

Keywords—Educational data mining; Data mining; Ensemble techninuqe; Rotaion forest algorithm; Decision tree;

I. Introduction

Explosion of information has led to the production of very large amounts of raw data. Data mining and machine learning techniques are the machine-based processes to extract hidden knowledge and information that traditional methods would find either impossible or too time-consuming and costly. Indeed, the algorithms have the role interfacing between human and data. The obtained information extracted from raw data is a turning point in improving services to individuals. Data mining and machine learning algorithms have been widely used in range of research fields such as healthcare [1-4], medicine [5-7], economics [8, 9], decision support systems [10, 11], sentiment analysis [12-14] etc.

Educational data mining (EDM) is a new field that uses various data mining and machine learning methods [15] to improve different aspects of education including e-learning. The EDM can be used to predict students' performance and study outcomes to help increase the efficiency of students' work. The performance of students can be investigated using internal (student-related factors) and external factors. The most important factors are: personal ability, teaching method, superior teaching economics, course type, social and environment. EDM can attempt to uncover data patterns by using available factors. It should be noted that for knowing important factors affecting a student's performance, a proper analysis of the learning process is required. For this reason, EDM can uncover different information regarding a student's performance and the most related factors.

This study concentrates on educational data mining (EDM) using machine learning and data mining algorithms. The study tries to figure out the most important factors which affect the performance of students. The data set is related to the undergraduate students of Department of Electrical Education of Gazi University, Turkey. Firstly, two well-known decision tree (RandomTree and REPTree) algorithms were applied on real-world data set. Then an ensemble technique (Rotation Forest) was applied to improve the performance of these methods. The two new ensemble methods are named RF-Random Tree and RF-REPTree, respectively. The obtained results showed that Rotation Forest can improve the performance of base decision trees significantly. In addition, using proposed approach, several simple and understandable rules were generated. We observed that the Rotation Forest with base decision trees (RandomTree and REPTree) can create deeper trees compared to simple decision trees. The proposed educational data mining methodology not only has good performance (evaluated with different metrics such as FP Rate, Precision, MCC, recall, F_1, ROC Area, and accuracy) but also generated several understandable and simple rules.

The rest of the study is organized as follows. Section II discusses about some related works in the literature. Section III introduces the methods used in current research. The experimental results and discussion will be presented in Section IV. We conclude the research in Section V.

II. Related work

In this section we discuss some studies regarding EDM in the literature. To predict the students' performance, a supervised data mining method - J48 decision tree, was used in [16]. The proposed model was applied to a data set collected from event logs in online learning system to estimate students' final grade. Kumar and Vijayalakshmi [17] proposed a Baker's taxonomy-based application in an educational data set using various single and also multi-instance-based learning methods. According to the achieved outcomes, decision stump tree and the Simple MI methods were much better than other methods among single instance and multi-instance learning algorithms. Moreover, Khan and Ghosh [18] proposed a data mining-based approach to discover the impact of teaching on student performance. It seems that the teaching is an external factor but it is quite impressive on students' performance. Therefore, they applied the association mining technique to 0.2 million academic records which were extracted from one online system of an academic institute of national importance in India. Their results indicated that teaching has a positive effect on students' academic performance. Research [19] concentrated on the students at risk of academic failure using the educational data mining approach. This was because preventing of student failure is a key for schools and universities to improve the quality of education. To this end, extracting most related factors those have impact on students' performance is a prevention solution. The research evaluated the usefulness of a specific hidden class model, the Bayesian Profile Regression, to identify students more likely to drop out. The model was applied on real data extracted from an online questionnaire system which filled by undergraduate students in a university in Italy.

In another research [20], EDM is used to provide a predictive analysis of academic performance of students in public school in Brazil. Therefore, a descriptive statistical analysis performed to profit insight from data. In order to classify the data sets, classification methods based on the Gradient Boosting Machine (GBM) approach were implemented. Their outcomes indicated that although the factors 'grades' and 'absences' were the most related for forecasting at the end of the academic year resulted from student performance, they revealed that factors 'neighborhood', 'school' and 'age' were also potential indicators of a student's academic either success or failure. Study [21] attempted to support the e-learning system using an appropriate educational data mining technique. Their proposed system was developed through utilizing five varying steps of knowledge input. The process was carried out in two modules including the server and client modules.

III. Data set used Method and material

In this section, we discuss about data set used in this paper. This study uses a user knowledge modeling data set from student in Turkey. The data set is related to the knowledge status of students about the field of Electrical DC Machines. The data set has 5 input features with 258 records and one class target as our output. The class feature includes 4 main classes including: VL=Very Low (for beginner level): 50, L=Low (for intermediate level): 129, M=Middle (for expert level): 122, and finally, H=High (or advanced level): 130. This data set can be downloaded in UCI machine learning repository [22]. More information related to attributes is presented in TABLE I:

TABLE I. Attribute information related to student data set

Name	Description of feature	Range	Type
STG	The degree of study time for goal object materails	0-0.99	Input value
SCG	The degree of repetition number of user/student for goal object materails	0-0.90	Input value
STR	The degree of study time of user/student for related objects with goal object	0-0.95	Input value
LPR	The exam performance of user/student for related objects with goal object	0-0.99	Input value
PEG	The exam performance of user/student for goal objects	0-0.93	Input value
UNS	The knowledge level of user/student	VL, L, M, H	Target value

IV. Results ans discussion

We present the obtained results using proposed method in this section. We will then discuss about our findings and try to compare them with previous studies. For this research, K-fold cross validation technique used to evaluate the predictive methods by using partitioning the original data into a training set to train the method, and then a test set to evaluate the proposed model. Hence, we set the value of K equal to 10 (10-fold cross validation). Also, the following metrics (see equations 1-6) are used for evaluating the performance of methods.

$$\text{FP Rate} = \text{FPR} = \text{FP} / (\text{TN}+\text{FP}) \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP}+\text{FP}) \quad (2)$$

$$\text{Recall} = \text{Sensitivity} = \text{TP Rate} = \text{TP} / (\text{TP}+\text{FN}) \quad (3)$$

$$\text{F-measure} = F_1 = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{MCC} = ((\text{TP.TN})-(\text{FP.FN})) / \text{Sqrt}((\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FP})(\text{TN}+\text{FN})) \quad (5)$$

$$\text{Accuracy} = (\text{TP}+\text{TN}) / (\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (6)$$

A. Experimental Result

In the first step, we apply simple RandomTree and REPTree algorithms on the data set, respectively. The results were displayed in Table II and III.

TABLE II. Performance evaluation of simple RandomTree algorithm

Metric	RandomTree				
	Class: VR	*Class: H*	*Class: L*	*Class: M*	*Weighted Average*
FPR	**0.017**	0.036	0.074	0.035	0.046
Precision	0.826	0.897	0.852	**0.924**	0.885
Recall	0.792	**0.968**	0.904	0.830	0.884
F_1	0.809	**0.931**	0.877	0.874	0.883
MCC	0.790	**0.909**	0.817	0.817	0.837
ROC Area	0.887	**0.966**	0.915	0.897	0.919

TABLE III. Performance evaluation of simple REPTree algorithm

Metric	REPTree				
	Class: VR	*Class: H*	*Class: L*	*Class: M*	*Weighted Average*
FPR	0.021	**0.010**	0.063	0.024	0.033
Precision	0.792	**0.969**	0.872	0.952	0.916
Recall	0.792	**0.984**	0.904	0.909	0.915
F_1	0.792	**0.976**	0.888	0.930	0.915
MCC	0.770	**0.969**	0.833	0.896	0.882

ROC Area	0.963	**0.985**	0.961	0.970	0.970

It can be seen from Table II and III that RandomTree algorithm had the accuracy 88.3721% while REPTree algorithm had the accuracy 91.4729%.

B. Optimization using Rotation Forest Algorithm

We applied the ensemble technique (Rotation Forest algorithm) with both decision trees to improve the performance of methods. For this purpose, RandomTree and REPTree algorithms were used as base methods for Rotation Forest. Therefore, the ensemble methods were applied on the data set. More details regarding the obtained results using the ensemble technique (RF-RandomTree and RF-REPTree) are presented in TABLE IV and TABLE V.

TABLE IV. PERFORMANCE EVALUATION OF RF- RANDOMTREE ALGORITHM

Metric	RF-RandomTree				
	Class: VR	*Class: H*	*Class: L*	*Class: M*	*Weighted Average*
FPR	**0.000**	0.010	0.080	0.024	0.036
Precision	**1.000**	0.968	0.854	0.951	0.928
Recall	0.792	0.952	**0.988**	0.875	0.922
F_1	0.884	**0.960**	0.916	0.911	0.922
MCC	0.880	**0.947**	0.877	0.870	0.892
ROC Area	0.990	**0.993**	0.978	0.969	0.980

TABLE V. PERFORMANCE EVALUATION OF RF- REPTREE ALGORITHM

Metric	RF-REPTree				
	Class: VR	*Class: H*	*Class: L*	*Class: M*	*Weighted Average*
FPR	**0.000**	0.005	0.074	0.024	0.033
Precision	**1.000**	0.984	0.863	0.953	0.936
Recall	0.708	0.952	**0.988**	0.920	0.930
F_1	0.829	**0.968**	0.921	0.936	0.929
MCC	0.829	**0.958**	0.885	0.905	0.904
ROC Area	0.991	**0.995**	0.977	0.972	0.981

The achieved results revealed that RF-RandomTree algorithm had the accuracy 92.2481% while RF-REPTree algorithm had the accuracy 93.0233%. This evidence can show that the ensemble-based decision tree have better performance than simple ones.

C. Discussion

According to TABLE II and TABLE III, it can be observed in most of cases both algorithms had the best performance related to *Class H* except FPR and Precision in RandomTree algorithm. The size of the tree generated by REPTree was 15 while the size of the tree generated by RandomTree was 57. Generally, in the first phase REPTree showed better performance than RandomTree. According to TABLE IV and TABLE V, we did not find a specific pattern related to the performance of methods for each class. Moreover, the size of the tree generated by RF-REPTree was 190 while the size of the tree generated by RF-RandomTree was 642. Generally, in the second phase, we observed that RF-REPTree showed better performance than RFRandomTree once again.

According to our findings, the proposed methodology can improve prediction, accuracy and other related metrics. The findings indicated even though Rotation Forest algorithm can enhance the performance of REPTree algorithm, it could improve the performance of RandomTree more significantly. Another significant point is that using Rotation Forest algorithm the size of both REPTree and RandomTree algorithms were increased. Generally, Rotation Forest algorithm could improve the accuracy of RandomTree about 3.876% whereas the accuracy of REPTree using Rotation Forest enhanced about 1.5504%. In other words, the ensemble method can improve the quality of educational data mining not only for its metrics but also about generated trees. Comparison of algorithms before and after applied ensemble technique is presented in TABLE VI.

TABLE VI. COMPARISON OF ALGORITHMS BEFORE AND AFTER APPLIED ENSEMBLE TECHNIQUE

Item	RandomTree	REPTree	RF-RandomTree	RF-REPTree
Accuracy (%)	88.3721	91.4729	92.2481	93.0233
Size of tree	57	15	642	190
Tree sets	1	1	10	10
Running time (second)	0.00 *	0.00 *	0.03	0.02

* Note that the response time was very small and so we consider it as 0.

TABLE VI shows that RF-RandomTree and RF-REPTree algorithms have 10 tree sets while simple RandomTree and RF-REPTree algorithms have only one set. In addition, we can see that the running times (response time) for both decision tree ensemble methods to build the models are very low. This means that the proposed ensemble decision trees have efficiency and effectiveness. The applied methods generate different simple rules, we try to present two rules (selected randomly) for each of simple decision trees which is showed in TABLE VII. It should be noted that each algorithm generated more rules, but however, two of them for each algorithm were chosen due to the limited pages.

TABLE VII. GENERATED RULES USING SIMPLE DECISION TREES AND DECISION TREES ENSEMBLE

Method	Generated rules
RandomTree	IF LPR < 0.32 AND PEG < 0.42 AND PEG < 0.24 AND STR < 0.75 THEN CLASS: very_low
	IF LPR < 0.32 AND PEG < 0.42 AND PEG < 0.24 AND STR >= 0.75 THEN CLASS: Low
REPTree	IF PEG < 0.38 AND PEG < 0.14 AND LPR < 0.62 THEN CLASS: very_low

	IF PEG < 0.38 AND PEG < 0.14 AND LPR >= 0.62 THEN CLASS: Low
RF-RandomTree	IF (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < 0.73) AND (.71 LPR_2+0.702PEG_1-0.057STG_0_0 < 0.68) AND (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < -1.66) AND (0.71 LPR_2+0.702PEG_1-0.057STG_0_0 < 0.07) THEN CLASS: Middle
	IF (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < 0.73) AND (.71 LPR_2+0.702PEG_1-0.057STG_0_0 < 0.68) AND (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < -1.66) AND (0.71 LPR_2+0.702PEG_1-0.057STG_0_0 >= 0.07) THEN CLASS: High
RF-REPTree	IF (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < 0.86) AND (0.71 LPR_2+0.702PEG_1-0.057STG_0_0 < 0.68) AND (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < -2.19) THEN CLASS: High
	IF (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 < 0.86) AND (0.71 LPR_2+0.702PEG_1-0.057STG_0_0 < 0.68) AND (-0.657PEG_2-0.55SCG_0-0.515STR_1_1 >= -2.19) AND (0.753PEG_2-0.51SCG_0-0.416STR_1_1 < -1.78) THEN CLASS: Low

We observed from TABLE VII that the generated rules using ensemble decision tree methods included more details than simple decision trees.

V. Conclusion

Recent development of new technologies in computer science has led to the production of new methods to explore information and knowledge hidden among different raw data sets. One of these data sets is related to students and their performance in various levels of education. Hence, educational data mining can be applied to educational datasets to enhance the quality of education. This study used two well-known decision tree algorithms (RandomTree and REPTree) as our base methods. Our research conducted into two main steps. Firstly, we applied both simple methods on the data set. As a result, we observed that REPTree algorithm had better performance compared to RandomTree algorithm. We then used the ensemble method called Rotation Forest (RF) algorithm. By using ensemble methods, we found that RF can improve the performance of both decision trees significantly. Moreover, our experimental outcomes showed that the ensemble decision tree methods can create more understandable rules than simple ones. In other words, the size of generated trees using RF-RandomTree and RF-REPTree algorithms were greater than simple decision trees (see TABLE VI). As a conclusion, we would argue that the ensemble decision trees can be used for educational data mining due to its good performance.

References

[1] M. Abdar, M. Zomorodi-Moghadam, R. Das, and IH. Ting. Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications. vol. 67, pp. 239-251, January 2017.

[2] M. Abdar. Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease. Carpathian Journal of Electronic and Computer Engineering. vol. 8, no. 2, pp. 31-36, July 2015.

[3] M. Abdar, and M. Zomorodi-Moghadam. Impact of Patients' Gender on Parkinson's disease using Classification Algorithms. Journal of AI and Data Mining, vol. 6, no. 2, pp. 277-285, July 2018.

[4] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar. Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease. In Computer and Applications (ICCA), 2017 International Conference on, IEEE. September 2017, pp. 306-311.

[5] A. G. Dunn, J. Leask, X. Zhou, K. D. Mandl, E. Coiera, Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study, Journal of medical Internet research 17 (6).

[6] X Zhou, E Coiera, G Tsafnat, D Arachi, MS Ong, Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter (2015), Amsterdam: IOS Press.

[7] X. Zhou, Y.Wang, G. Tsafnat, E. Coiera, F. T. Bourgeois, A. G. Dunn, Citations alone were enough to predict favorable conclusions in reviews of neuraminidase inhibitors, Journal of clinical epidemiology 68 (1) (2015) 87–93.

[8] M. Abdar, and NY. Yen. Design of A Universal User Model for Dynamic Crowd Preference Sensing and Decision-Making Behavior Analysis. IEEE Access. August 2017, vol. 5, pp. 24842-24852.

[9] M. Abdar, and NY. Yen. Understanding regional characteristics through crowd preference and confidence mining in P2P accommodation rental service. Library Hi Tech. vol. 35, no. 4, pp. 521-541, 2017.

[10] Y. Li, X. Zhou, P. Bruza, Y. Xu, R. Y. Lau, A two-stage decision model for information filtering, Decision Support Systems 52 (3) (2012) 706–716.

[11] X Zhou, ST Wu, Y Li, Y Xu, RYK Lau, PD Bruza, Utilizing search intent in topic ontology-based user profile for web mining, in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 558-564.

[12] X. Tao, X. Zhou, J. Zhang, J. Yong, Sentiment analysis for depression detection on social networks, in: International Conference on Advanced Data Mining and Applications, Springer, 2016, pp. 807–810.

[13] X Zhou, X Tao, MM Rahman, J Zhang, Coupling topic modelling in opinion mining for social media analysis, in proceedings of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 533-540.

[14] X Zhou, X Tao, J Yong, Z Yang, Sentiment analysis on tweets for social events, in proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 557-562.

[15] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata. Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. Indonesian Journal of Electrical Engineering and Computer Science. vol. 9, no. 2, pp. 447-459, Feburary 2018.

[16] C. C. Kiu. Supervised Educational Data Mining to Discover Students' Learning Process to Improve Students' Performance. In Redesigning Learning for Greater Social Impact Springer, Singapore 2018, pp. 249-258.

[17] S. A. Kumar, and M. N. Vijayalakshmi. Efficiency of Multi-instance Learning in Educational Data Mining. In Knowledge Computing and its Applications. Springer, Singapore 2018, pp. 47-64.

[18] A. Khan, and S. K. Ghosh. Data mining based analysis to explore the effect of teaching on student performance. Education and Information Technologies, pp. 1-21, 2018.

[19] A. Sarra, L. Fontanella, and S. Di Zio. Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. Social Indicators Research. pp. 1-20, 2018.

[20] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research. 2018.

[21] P. Appalla, V. M. Kuthadi, and T. Marwala. An efficient educational data mining approach to support e-learning. Wireless Networks. vol. 23, no. 4, pp. 1011-1024, May 2017.

[22] User Knowledge Modeling Data Set, https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling, [accessed: April 2017].