Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1298~1304
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i3.pp1298-1304

Building student's performance decision tree classifier using boosting algorithm

Farid Jauhari, Ahmad Afif Supianto
Faculty of Computer Science, Brawijaya University East Java, Indonesia

Article Info

Article history:

Received Jun 9, 2018
Revised Aug 20, 2018
Accepted Nov 18, 2018

Keywords:

adaBoost
Boosting algorithm
C5.0
Classification
Prediction
Student's performance

ABSTRACT

Student's performance is the most important value of the educational institutes for their competitiveness. In order to improve the value, they need to predict student's performance, so they can give special treatment to the student that predicted as low performer. In this paper, we propose 3 boosting algorithms (C5.0, adaBoost.M1, and adaBoost.SAMME) to build the classifier for predicting student's performance. This research used 1UCI student performance datasets. There are 3 scenarios of evaluation, the first scenario employs 10-fold cross-validation to compare the performance of boosting algorithms. The result of the first scenario showed that adaBoost.SAMME and adaBoost.M1 outperform baseline method in binary classification. The second scenario was used to evaluate boosting algorithms under the different number of training data. On the second scenario, adaBoost.M1 has outperformed another boosting algorithms and baseline method on the binary classification. The third scenario, we build models from one subject dataset and test using another subject dataset. The third scenario results indicate that it can build prediction model using one subject to predict another subject.

Corresponding Author:

Farid Jauhari,
Faculty of Computer Science,
Brawijaya University East Java, Indonesia.
Email: faridjauhari@ub.ac.id

1. INTRODUCTION

Education is the most important factor in the development of a country, even more in developing country. It always needs information where areas of education which need to improve, and also needs information on how to improve it. In order to get that information, we need data mining to extract the information from the data. Educational data mining is a field of data mining that focuses on the educational data. Romero and Ventura [1] categorized educational data mining by the task, it can be an analysis and visualization of data [2], [3] or recommendation for students [4]. In addition, the task can be in the form of providing feedback for supporting instructors, predicting student's performance, student modeling, detecting undesirable student behaviors, grouping students, social network analysis, developing concept maps, constructing courseware, and planning and scheduling. In this paper, we focus on the student's performance prediction.

Prediction and analysis of student's performance are the most important things to do in the education institutes [5], [6]. Prediction and analysis can improve student's performance more efficient and effective [7], because it can predict the low-performance student, and then we can give special treatment to improve their performance. Predicting student's performance can be done by building accurate classifier from the past student's data, and then we can predict using the classifier with current student's data. In order to build the classifier, we need to understand the parameter or attribute that affect the performance of the students and also the existing method that used.

Shahiri et al. [7] summarized the type of attributes that frequently used in the student's performance prediction in many research, the type of attributes are: cumulative grade point average (CGPA), internal assessments, external assessments, extra-curricular activities, student demographics, and internet and social media interaction. Internal assessments included assignment mark, quizzes, lab work, attendance and other marks that associated with the course or subject. External assessments are final mark of another subject while assessing one specific subject. While student demographics are grouping attribute that includes gender, age, family background, disability, health, want to take higher education or not, does he/she have extra school support and other personal attributes. In most cases, research uses the combination of two or more type of attribute rather than one type of attribute. Kaur et al. [5] were using CGPA, internal assessments, and student demographics. Fernandes et al. [8] and Pandey and Taruna [9] were employing CGPA, internal assessments, and student demographics. Hamsa et al. [10] were utilizing internal and external assessments. Sandoval et al. [11] were CGPA, external assessments, student demographics, internet and social network interaction. While Cortez and Silva [12] were utilizing internal assessments, internet activity, extracurricular dan student demographics. In this paper, we use the same dataset as in [12], that available in 1UCI datasets. There are two subsets in that dataset, the dataset for Mathematics and Portuguese subject. In this paper, we also analyze the possibility of using those datasets as external assessments, by building prediction model with Mathematics dataset and predict using Portuguese dataset, and otherwise.

There are many different methods that have been used to predict student's performance. Naïve Bayes used in [5], [9], [12] as a baseline method. While Pandey and Taruna [9] also employing advance Naïve Bayes that called as Aggregating One-Dependence Estimator (AODE), and it was combined with Decision Tree and K-Nearest Neighbour. As a comparison methods [9] employ Naïve Bayes, ZeroR, OneR, KSTAR, Naïve Bayes Tree (NBTRee), and AODE. Neural network (NN) utilized in [5], [12] as the prediction model, because of its capability to handle non-linear data. While Cortez and Silva [12] also used SVM which accuracy was not much different than NN did. The Fuzzy based method such as Neuro-Fuzzy (ANFIS) [13] and Fuzzy Association Rule Mining [14] also employed by to predict student's performance. Random Forest (RF) were employed by [11], [12] had promising results on the prediction student's performance. Decision Tree (DT) based method had a promising result on [5], [8]-[10], [12]. In Fernandez et al. [8], DT constructed by boosting algorithm was called Gradient Boost Machine. In this research, we compare 3 boosting algorithms, C5.0 (Boosting C4.5), adaBoost.M1, and adaBoost.SAMME to build decision tree as a prediction model.

Boosting is the general method that utilized to reduce the error of learning algorithms [15]. Freund and Schapire [15] introduced two variant of boosting algorithm, adaBoost.M1 and adaBoost.M2 which can combine with other algorithms as the weak learner. The result of [15] shows that adaBoost outperform C4.5, and adaBoost can improve the performance of C4.5 when C4.5 was used as its weak learner. There are many researches on classification and prediction in various fields which employ adaBoost. Boosting algorithm outperforms Bayesian algorithm on spam filtration [16]. Wang et al. [17] hybrid adaBoost and least square based to predict railway turnouts. In Biology, adaBoost was utilized to classify Classification of enzyme and non-enzyme [18] and predict splice site on gene detection [19]. In Medical, adaBoost employ to build the classifier of epilepsy using signals from electroencephalography (EEG) [20]. It also helpful in classification and segmentation of brain tumor [21]. AdaBoost had a good performance on class prediction from gene expression profiles [22], for further this prediction used to enhance cancer diagnostic process. The advantages of adaBoost were stability, accuracy, and it can be used on real-time such as video object tracking [23].As we mentioned before, in this paper, we are using 2 variants of adaBoost, adaBoost.M1 as in [15] and adaBoost.SAMME which introduced in [24].

As we mentioned before, that C4.5 as a weak learner of adaBoost can give the better result than the original C4.5 [15]. There is an improvement of C4.5 which is called as C5.0. The improvement of C4.5 that become base of C5.0 was introduced by Quinlan [25], [26]. The most important improvement is boosting mechanism that added on original C4.5 [27]. As adaBoost, C5.0 was widely used in the research area of classification and prediction. Pang and Gong [27] implemented C5.0 on a classification of individual credit evaluation. C5.0 well performs on liver disease prediction when hybridized with a genetic algorithm which can optimize the rule [28]. Jincheng et al. [29] were implemented C5.0 to predict failure of smart meters. Nia and Khalili [30] build intrusion detector on the computer network using C5.0 as a classifier.

In this paper, we focus on student's performance prediction with C5.0, adaBoost.M1, and adaBoost.SAMME. There are two classification goals in our research: binary classification (pass/fail) and five levels classification (very good, good, satisfactory, sufficient, and fail) as in [12]. We also reconstruct the research of Cortez and Silva [12] to compare the result of DT based classifier with our result. Cortez and Silva [12] build DT using Classification and Regression Trees (CART) or sometimes called RPART (Recursive Partitioning) as in [31]. This paper aims to show the other methods of building DT that better than the method that used in [12] for student's performance prediction.

2. RESEARCH METHODOLOGY

Our research methodology starts with collecting datasets followed by preprocessing step. The next step was building model with 3 boosting algorithms, as we mentioned before. And then applying preprocessed dataset to evaluate boosting methods. As we mentioned before that we are using the same datasets as used in [12].

The datasets consist of 2 subsets, Mathematics subject dataset and Portuguese subject dataset. The mathematics dataset contains 395 student data and the Portuguese contain 649 student data. The datasets consist the same attributes, and there is 33 attribute. The last attribute that named G3 is the final grade of the student in the subject. The range value of this attribute is between 0 and 20. The objective goals of the prediction are at this last attribute.

There are 3 general steps of research methodology in this paper, those are preprocessing, building model, and experiment as a model evaluation. This general methodology is shown in Figure 1. Preprocessing employed to set the class of students based on their final grades (G3). As we mentioned before, there are two classification goals in this paper, that were binary classification (pass/fail) and 5-levels classification (very good, good, satisfactory, sufficient, and fail). Our mapping between grades and classification goals are shown in Table 1 and Table 2.

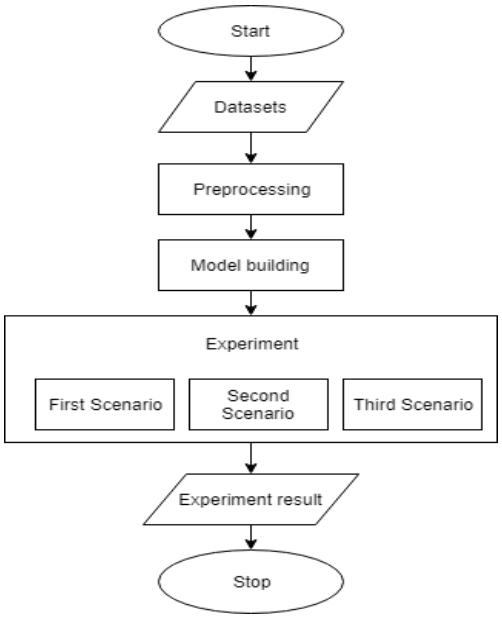

Figure 1. Research methodology

In this research, we build and evaluate prediction model under environment of R language. First prediction model was built under C5.0 algorithm. C5.0 is the improvement of C4.5 as in [26]. As C4.5 does, C5.0 also generate decision tree (DT) as the outcome. AdaBoosts that adopted in this paper are adaBoost.M1 [15] and adaBoost.SAMME [24]. As in [15], we use 100 of boosting iteration. As we mentioned before, we also implement RPART method that used in [12] as a baseline method of student's performance prediction model. We limit the number of recursive partition of trees in 5 level depth. This limitation also used in adaBoost.M1 and adaBoost.SAMME as an attempt to maintain the fairness of comparison.

The accuracy of the prediction models was calculated by using Percentage of Correct Classification (PCC). The PCC value near 100% means that the model has a high accuracy on classification. In this paper, we employ 3 scenarios of evaluation. First, we evaluate using 20 runs of a 10-fold cross-validation as in [12]. The first scenario intends to compare those 3 boosting algorithms and the DT that used in [12].

As the second scenario, we evaluate using 100% datasets as data training (it means data training = data test), 90% dataset as training data (it means randomly sampling of 90% datasets as training data and 10% datasets as test data), 80% - 10% dataset as training data. The evaluation was taking pictures of the average of PCC in 10 runs of each partition scenario. The second scenario aims to show the accuracy of the model on the different number of training data.

As the third scenario, we use Mathematics dataset as training data and Portuguese dataset as testing data, and otherwise (further we call it as 'cross-subject evaluation'). Average of PCC in 10 runs of each of this scenario representing the accuracy of models. The third scenario tends to know "does the characteristics of student's performance differ in different subjects?".

Table 1. Grade in Binary Classification

Classification Goals	Grade
Pass	10-20
Fail	1-9

Table 2. Grade in 5-Levels Classification

Classification Goals	Grade
Very Good	16-20
Good	14-15
Satisfactory	12-13
Sufficient	10-11
Fail	1-9

3. RESULTS AND ANALYSIS

We run our experiment under R language environment [32]. In R language, there are several libraries that used our experiment. As adaBoost.M1 and adaBoost.SAMME implementation, we employ 'adabag' library. And we use 'C50' library of R to build C5.0 models. As in [12], we also install 'rminer' library to reconstruct the DT model of [12].

3.1. First Scenario

As we mentioned before, the first scenario was using 10-fold cross-validation. In order to implement 10-fold cross-validation, we employ 'caret' library of R. Experimental result under the first scenario are summarized in Table 3. Table 3 contains mean of PCC value in different prediction models on two classification goals (binary and 5-levels classification) and two distinct subsets of the dataset (Mathematics and Portuguese subject). Bold and underlines highlighted the best PCC models in different classification goals and distinct dataset subsets. In binary classification, best PCC value achieved by adaBoost.SAMME models, with the mean of PCC is 91.3377 on Mathematics and 93.197 on Portuguese. Although the difference in the accuracy between adaBoost.SAMME and RPART just below 1%. RPART are the best models in 5-levels classification, with 77.203 of PCC on Mathematics and 76.664 on Portuguese. Low accuracy of adaBoost model might be caused by the weak learner that used in 'adabag' library. The 'adabag' library was implemented adaBoost based on Freund and Schapire [15], which was using very simple weak learner that called findAttrTest. Table 3 shows that the lowest performer model is C5.0. It might be because the boosting iteration that used in C5.0 is just 10 times as in [26], while adaBoost using 100 times of boosting iteration.

Table 3. Result of 10-Fold Cross-Validation

	Binary Classification				5-Level Classification			
	RPART	C5.0	M1	SAMME	RPART	C5.0	M1	SAMME
MAT	90.392	88.96	90.893	91.377	77.203	72.804	74.567	72.434
POR	92.512	91.796	92.765	93.197	76.664	69.097	74.536	71.978

3.2. Second Scenario

The result of second experiment scenario on binary classification are summarized in Table 4. As we can see in Table 4, adaBoost.SAMME reach perfect classification when it trains and tests with the same data (100 of PCC). While adaBoost.M1 have very close to perfect accuracy, with PCC 99.747 on Mathematics and 99.661 on Portuguese. It is also noticed that all model of boosting algorithms outperforms the RPART model. This experiment shows us that boosting algorithm can build models that follow the curvature of the data. the 100% or close to 100% accuracy have a bad side, there is a theory of overfitting, which is the condition of the model that over follow the curvature of the data training, so when tested with data that have different curvature the accuracy was low. But based on Freund and Schapire [15], adaBoosts should suffer less from overfitting.

When the model trained with 90% of dataset and test with the remainings (10% of the dataset) the result doesn't have much different from the 10-fold cross-validation, the winner is still adaBoost.SAMME on Portuguese and adaBoost.M1 on Mathematics. When we look into Portuguese results, RPART algorithm outperforms the other models in 80%, 70%, 50%, 40%, 30%, and 10%. We interested to do some simple statistical analysis on this results that we put on Table 5.

Table 4. Binary Classification Result

		100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
MAT	RPART	93.924	90.5	88.987	88.571	89.304	90.909	90.253	89.386	90.696	91.236
	C5.0	96.456	89	88.608	88.487	88.797	89.293	89.198	88.917	87.025	88.68
	M1	99.747	91.25	91.013	89.664	90.57	90.354	90.506	90.397	89.652	88.624
	SAMME	100	89.25	90.759	88.824	89.051	89.545	89.03	89.206	88.291	86.685
POR	RPART	94.915	92.923	93.538	92.821	89.051	92.277	92.513	91.846	91.231	91.556
	C50	95.686	91.231	90.769	91.231	91.538	91.6	91.897	91.692	91.731	90.308
	M1	99.661	93.077	92.462	92.154	92.308	92.123	92.128	91.692	91.481	91.538
	SAMME	100	93.077	91.846	91.795	91.538	92.215	92.231	91.758	92.558	91.368

Due to mean analysis on Table 5, we can conclude that adaBoost.M1 have the best performance for binary classification in the various number of training data, with 91.178 mean of PCC. Although the most robust model is RPART, this conclusion can be inferred by observing the standard deviation among all models. Although the winner in Portuguese datasets is C5.0, in overall, the less value of standard deviation achieved by RPART model. This result also giving information about our statement before about overfitting. Due to small standard deviation (less than 4), it can prove that adaBoost model is far from overfitting.

Table 5. Statistical Analysis on Binary Classificaion Result

		Mean	Max	Min	Stdev
MAT	RPART	90.377	93.924	88.571	1.451
	C5.0	89.446	96.456	87.025	2.412
	M1	91.178	99.747	88.624	2.945
	SAMME	90.064	100	86.685	3.451
POR	RPART	92.267	94.915	89.051	1.467
	C50	91.768	95.686	90.308	1.384
	M1	92.862	99.661	91.481	2.31
	SAMME	92.839	100	91.368	2.435

The result of second experiment scenario on 5-level classification are summarized in Table 6. And the statistical analysis of this result is shown in Table 7. The experiment of using the same data between training data and testing data (in column 100% of Table 6) still have the same conclusion with the conclusion in the binary classification, that is boosting algorithm models outperform RPART model. This reinforces the statement before, that boosting algorithms model can closely follow the curvature of the training data. They can follow the curvature of the training data because of the mechanism of boosting, which can improve the error in line with boosting iterations [15], [26].

According to Table 6 and 7, we can simply conclude, that RPART was overall outperformed the boosting algorithm in 5-level classification as the conclusion on the result of 10-fold cross-validation (first scenario). Although in 10-fold cross-validation just using 90% of the dataset that used as training data and the other 10% of the dataset as testing data. And furthermore, RPART algorithm has produced more robust prediction models for this student performance datasets, than 3 boosting algorithms that implemented in this research.

Table 6. Result of 5-Level Classification

		100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
MAT	RPART	80.506	77.5	78.861	76.891	76.203	75.152	73.755	70.217	67.785	57.865
	C5.0	90.886	73	69.62	69.496	67.595	69.192	68.312	67.004	63.291	56.545
	M1	87.342	75	73.165	72.941	72.975	73.788	71.561	69.856	63.291	57.528
	SAMME	87.57	70.25	70	67.143	68.734	67.323	68.439	65.199	64.241	56.91
POR	Rpart	76.733	76.615	76.231	77.026	75.385	73.6	73.513	73.626	71.115	63.915
	C50	88.598	66.615	68.077	67.692	68.423	67.385	67.718	66.418	67.385	63.675
	M1	82.018	74.615	74	73.128	73.962	73.846	71.359	71.582	68.942	65.248
	SAMME	80.447	66.923	67.462	67.436	67.731	66.554	66.205	67.429	65.019	61.453

3.3. Third Scenario

The third scenario is the cross-subject evaluation. The results of this scenario are shown in Table 8. Bolded and underlined highlight the best PCC, while just bolded showed the lowest PCC in configuration. Mat-Por means that we build the models using Mathematics dataset and predict (test) using Portuguese dataset. In Mat-Por configuration, the highest PCC achieved by RPART with 89.676 and the lowest gained

by adaBoost.SAMME with 89.06. The difference value of PCC on binary classification and Mat-Por among prediction models was very small, which is under 1%. Same result obtained by Por-Mat configuration on binary classification, the best PCC gained by RPART with 89.367 and lowest PCC achieved by 86.228. The difference between highest and lowest PCC was less than 3%. By those result, we can conclude that in binary classification Mathematics dataset or Portuguese dataset build identic prediction models.

Table 7. Statistical Analysis on the Result of 5-Levels Classification

		Mean	Max	Min	Stdev
MAT	RPART	73.473	80.506	57.865	6.351
	C5.0	69.494	90.886	56.545	8.291
	M1	71.745	87.342	57.528	7.35
	SAMME	68.581	87.57	56.91	7.318
POR	RPART	73.776	77.026	63.915	3.747
	C50	69.199	88.598	63.675	6.589
	M1	72.87	82.018	65.248	4.095
	SAMME	67.666	80.447	61.453	4.616

Table 8. Cross-Subject Evaluation Results

	Binary Classification				5-Level Classification			
	RPART	C5.0	M1	SAMME	RPART	C5.0	M1	SAMME
Mat-Por	89.676	89.214	89.183	89.06	68.105	67.951	68.29	63.236
Por-Mat	89.367	87.848	86.81	86.228	74.937	72.152	70.405	64.557

Same as binary classification, RPART had the best PCC 68.105 in Mat-Por configuration and 74.152 in Por-Mat configuration. Lowest PCC gained by adaBoost.SAMME 63.236 and 64.557. The larger difference between 2 configurations gained by RPART in 5-level classification that is 6.83. It's not a big number to tell differences. So, we can conclude that in 5-level classification Mathematics dataset or Portuguese dataset construct identic prediction models.

Due to Table 8, the accuracy of prediction is still high if we compare to result on Table 3, this means that the Mathematics dataset and Portuguese dataset didn't have much different pattern. In other words, Mathematics dataset or Portuguese dataset construct identic prediction models.

4. CONCLUSION

Key to competitiveness value of the educational institution is student's performance. They need to predict low-performance students, in order to improve their performance with some special treatments. We have implemented 3 boosting algorithms (C5.0, adaBoost.M1, adaBoost.SAMME) to build decision tree-based classification models for student's performance prediction. The accuracy of those 3 boosting algorithms compared with RPART as a baseline model of DT based classification.

AdaBoost.M1 and adaBoost.SAMME models outperform RPART on binary classification models. While RPART outperforms boosting models in 5-level classification. The weakness of the adaBoost.M1 and adaBoost.SAMME in 5-level classification might be according to its weak learner, which was too simple. Using more complex weak learner might boost the accuracy of the adaBoost models. C5.0 as boosting of C4.5 failed to outperform RPART because of the limitation of the implementation in 'C50' library of R, which can't change the number of boosting iteration larger than 10 times, while adaBoost.M1 and adaBoost.SAMME can outperform RPART in 100 times of boosting iteration.

As expected, the cross-subject evaluation results show two different subject datasets which constructs identical prediction models. Therefore, on the implementation in the real world, we can just train the model with one subject and use the model to predict all subjects.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[2] A. A. Supianto, Y. Hayashi, and T. Hirashima, "Analysis of steps in posing arithmetic word problem as sentence-integration on interactive learning environment," *ICCE 2016 - 24th Int. Conf. Comput. Educ. Think Glob. Act Local-Main Conf. Proc.*, no. December, pp. 242–251, 2016.

[3] A. A. Supianto, Y. Hayashi, and T. Hirashima, "Visualizations of problem-posing activity sequences toward modeling the thinking process," *Res. Pract. Technol. Enhanc. Learn.*, vol. 11, no. 1, p. 14, 2016.

[4] A. A. Supianto, Y. Hayashi, and T. Hirashima, "Process-based Assignment-Setting Change for Support of Overcoming Bottlenecks in Learning by Problem-Posing in Arithmetic Word Problems," in *Journal of Physics: Conference Series*, 2017, vol. 812, no. 1, p. 012004.
[5] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.
[6] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447–459, 2018.
[7] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015.
[8] 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," *J. Bus. Res.*, no. February, pp. 0–1, 2018.
[9] M. Pandey and S. Taruna, "Towards the integration of multiple classifier pertaining to the Student's performance prediction," *Perspect. Sci.*, vol. 8, pp. 364–366, 2016.
[10] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technol.*, vol. 25, pp. 326–332, 2016.
[11] A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, and M. Montenegro, "Centralized student performance prediction in large courses based on low-cost variables in an institutional context," *Internet High. Educ.*, vol. 37, no. February, pp. 76–89, 2018.
[12] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance," in *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 2008, pp. 5–12.
[13] U. Bin Mat and N. Buniyamin, "Using neuro-fuzzy technique to classify and predict electrical engineering students' achievement upon graduation based on mathematics competency," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 5, no. 3, pp. 684–690, 2017.
[14] S. K. Verma and R. S. Thakur, "Fuzzy Association Rule Mining based Model to Predict Students' Performance," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 4, p. 2223, 2017.
[15] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," *Mach. Learn. Int. Work.*, pp. 148–156, 1996.
[16] W. Gu, "Application of Boosting Algorithm in Spam Filtration," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 12, no. 7, 2014.
[17] G. Wang, T. Xu, H. Wang, and Y. Zou, "AdaBoost and Least Square Based Failure Prediction of Railway Turnouts," *Proc. - 2016 9th Int. Symp. Comput. Intell. Des. Isc. 2016*, vol. 1, pp. 434–437, 2017.
[18] M. M. Sharif, A. Tharwat, A. E. Hassanien, and H. A. Hefny, "Enzyme vs. non-enzyme classification based on principal component analysis and AdaBoost classifier," *Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016*, pp. 288–293, 2017.
[19] E. Pashaei, A. Yilmaz, M. Ozen, and N. Aydin, "Prediction of splice site using AdaBoost with a new sequence encoding approach," *2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc.*, pp. 3853–3858, 2017.
[20] H. Rajaguru and S. K. Prabhakar, "Power spectral density and KNN based adaboost classifier for epilepsy classification from EEG," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 2017, pp. 441–444.
[21] R. Sonavane and P. Sonar, "Classification and segmentation of brain tumor using Adaboost classifier," *Glob. Trends Signal Process.*, pp. 396–403, 2016.
[22] L. Li, Z. Yu, J. Liu, J. You, H. S. Wong, and G. Han, "Multi-view based adaboost classifier ensemble for class prediction from gene expression profiles," *Int. Conf. Pattern Recognit.*, pp. 178–183, 2014.
[23] C. Yi, "Target Tracking Feature Selection Algorithm Based on Adaboost," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 734–740, 2014.
[24] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Stat. Interface*, vol. 2, no. 3, pp. 349–360, 2009.
[25] J. Ross Quinlan, *C4.5: Programs For Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[26] J. R. Quinlan, "Bagging, Boosting, and C4.5," in *Proceedings, Fourteenth National Conference on Artificial Intelligence*, 1996.
[27] S. Pang and J. Gong, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," *Syst. Eng. - Theory Pract.*, vol. 29, no. 12, pp. 94–104, 2009.
[28] M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, "Rule optimization of boosted C5.0 blassification using genetic algorithm for liver disease prediction," *2017 Int. Conf. Comput. Appl. ICCA 2017*, pp. 299–305, 2017.
[29] Y. JINCHENG, J. PING, C. GUANGYU, Y. TIEJIANG, and X. FEI, "Application of C5.0 Algorithm in Failure Prediction of Smart Meters," in *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2016, pp. 328–333.
[30] F. Nia and M. Khalil, "An efficient modeling algorithm for intrusion detection systems using C5.0 and Bayesian Network Structures," *2015 2nd Int. Conf. Knowledge-Based Eng. Innov.*, pp. 1117–1123, 2015.
[31] C. J. Breiman L., Friedman J. H., Olshen R. A., and Stone, *Classification and Regression Trees*. Taylor & Francis, 1984.
[32] "The R Project for Statistical Computing." [Online]. Available: https://www.r-project.org/. [Accessed: 05-Jun-2018].