Genetic Algorithm Based Feature Selection With Ensemble Methods For Student Academic Performance Prediction

Al Farissi[1], Halina Mohamed Dahlan[2], Samsuryadi*[3]

[1,3]Fakultas Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia
[2]Information Systems Department, Azman Hashim International Business School, Universiti Teknologi Malaysia, Johor, Malaysia

alfarissi@unsri.ac.id, halina@utm.my, syamsuryadi@unsri.ac.id

Abstract. Student academic performance is an important factor that affect the achievement of an educational institution. Educational Data Mining (EDM) is a data mining process that is applied to explore educational data that can produce information related to student academic performance. The knowledge produced from the data mining process is used by the educational institutions to improve their teaching processes, which aim to improve student academic performance results. In this paper, a method based on Genetic Algorithm (GA) feature selection technique with classification method is proposed in order to predict student academic performance. Almost all previous feature selection techniques apply local search technique throughout the process, so the optimal solution is quite difficult to achieve. Therefore, GA is apply as a technique of features selection with ensemble classification method in order to improve classification accuracy value of student academic performance prediction, as well as it can be used for datasets with high dimensional and imbalanced class. In this paper, the data used for experiments comes from Kaggle repository datasets which consists of three main categories: behaviour, academic, and demographic. The performances evaluation used to evaluate the proposed method is the Area Under the Curve (AUC). Based on the results obtained from the experiments, shows that the proposed method makes an impressive result in the predictions of student academic performance.

1. Introduction

In recent years, student academic performance prediction has become very important for higher education institutions [1]. Student academic performance is related to fruitfulness in the education process, students who have high academic performance will certainly have a greater chance of completing their studies well [2]. Because of this, prediction of student academic performance is an effective way that aims to prevent and treat academic success. However, to predict student academic performance is not an easy thing to do, because many factors can influence student academic performance, such as demographics, academic background, and behaviour. Therefore, the application of Educational Data Mining (EDM) is a way to overcome this problem [3]. Predicting student academic performance using data mining techniques still has problems related to accuracy for predicting student academic performance. Where, there is no consensus on the comparison of student academic performance predictions using the data mining classification method. This indicates there is no difference in performance that can be detected and there is no specific method that does the best for

student academic performance prediction. In addition, factors that affecting results of classification performance are the high dimensions of the dataset and imbalanced class [4, 5].

In the data mining domain, feature selection is intend to overcome the high dimensional of dataset [6, 7]. The problem of high data dimensions is done by reducing the dimensions of the features in the datasets. The reduction of data dimensions aims to improve machine-learning performance. Application of feature selection for datasets with N and M dimensions (features), feature selection aims to reduce M to M 'and M' ≤ M [7]. Majority of traditional features selection algorithms work by selecting features that range between sub-optimal and almost optimal regions. Therefore, the most optimal solution is quite difficult to obtain using this algorithm[8]. Genetic Algorithms work in a reasonable period by conducting global searches with the aim of finding solutions to the full search space, thus the results obtained significantly can improve performance to find high-quality solutions [9]. This study proposed method by combining Genetic Algorithms with Random Forest (RF) methods to increase the value of accuracy of student academic performance predictions. In the others hand, regarding learning problem from highly imbalanced datasets. Imbalanced class often appear in the real world where data distribution is not balanced. Generally in the case of datasets with two classes, it is assumed that the minority class is a positive class while the majority of the class is a negative class [7, 10]. Often, minority classes have a very low frequency compared to other classes. When a traditional classifier is used in a dataset, the classifier predicts it entirely as negative (majority class). Our proposed method aims to deal with feature selection problems and imbalanced class.

2. Material and methods

This study utilize we use student academic performance dataset from Kaggle repository. Attributes and general description are shown in Table I [11]. This data collection contains 480 student academic performance data, 16 attributes with multi-class labels, to be specific class in three intervals: high, medium and low. Attributes are categorized into three categories: demographic, academic and behavioral categories.

Table 1. Attributes and descriptions of dataset

No	Attributes Category	Attributes	Description
1	Demographic Category	Nationality	Nationality
2		Gender	Male or Female
3		POB	Birthplace
4		Parent responsible for student	Status parent
5	Academic Category	Educational Levels	Levels of school
6		Grade Levels	Student class group
7		Section ID	Register classroom
8		Semester	Academic year
9		Topic	Subject course
10		Student Absence Days	Attendance
11	Behavioral Category	Parent Answering Survey	Parent participation on survey
12		Parent School Satisfaction	Level satisfaction of parent
13		Discussion Activity	Student interaction
14		Active visiting resources	
15		Raised hand on class	
16		Seeing announcements	

The proposed method is a combination of GA feature selection and RF classification for student academic performance prediction. This proposed method compromise with high dimensional dataset and class imbalanced problem in terms to achieve high accuracy value for student academic performance prediction. Optimization features selection based on GA is involves high dimensional datasets and Random Forest classifier is employed to deal with the class imbalance problem. Figure 1 shown the proposed method for student performance prediction.

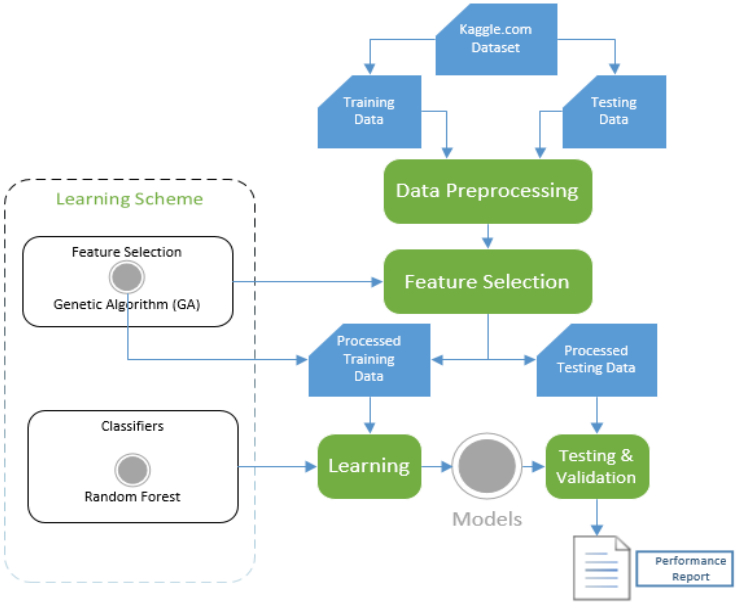

Figure 1. Proposed method of student academic performance prediction

This experiment uses the fifth generation i7 CPU 1.8GHz, 12GB RAM and Windows 10 64 bit operating system. In conducting this experiment, six classifiers are compared, which are: Decision Tree (DT), Artificial Neural Network (ANN), Random Forest, Voting, Bagging and Boosting compare the performance of classification models within the field of student academic performance prediction.

The experiment runs six classifiers where each classifiers are validated using x-fold cross validation to validate training and testing data. As shown in Table 2, this validation method divides the dataset into ten subsets and repeats ten times. In each round, a set of sections is taken to be used as a set of tests and the other sets are combined to become a training set. The final result obtained from average value of all round errors. X-fold cross validation included each instance nine times in the training set and at least once in the test set. Usually, x-fold cross validation is used because it can reduce computing time while maintaining the accuracy of estimates [12].

Table 2. X-fold cross validation

n-validation	Dataset Partition
1	■
2	■
3	■
4	■
5	■
6	■
7	■
8	■
9	■

10

To evaluate the performance of the six classifiers methods used, measurements of this research utilize a six metrics: Geometric Mean (G-Mean), Precision, True Negative Rate (TNR), True Positive Rate (TPR), F1-Score and Area Under Curve (AUC) to compare the performance of classifiers. This evaluation metrics are calculated using confusion matrix as shown in Table 3.

Table 3. Confusion matrix

		Actual Class	
		Positive	Negative
Predicted Condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

G-mean is one of the most comprehensive measurements to evaluate the performance of classification algorithms especially in class imbalance problems in the dataset. Equation 1 can obtain G-Mean values. Precision values refer to the number of positive category data that correctly classify divided by the total data classified as positive. Equation 2 can obtain precision. Furthermore, TPR Equation 3 shows how many percent of the positive category data is correctly classified. TNR is the number of correctly classified class instances that do not belong to the class divided by the total data classified as negative. AUC values refer to how accurate the system can classify data correctly. In other words, the value of AUC is a comparison between data that is correctly classified with the whole data.

$$\mathrm{G-Mean} = \sqrt{TPR.TNR} \quad (1)$$

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

$$\mathrm{TPR} = \frac{TP}{TP+FN} \quad (3)$$

$$TNR = \frac{TN}{TN+FP} \quad (4)$$

$$F1 - Score = 2\frac{Precision\,.\,Recall}{Precision + Recall} \quad (5)$$

$$\mathrm{AUC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (6)$$

3. Results and discussion

In the initial experiment, all classifiers were executed one by one validated using x-cross fold validation with a dataset of 16 attributes and 480 student performance data. The results of experiments with six classification algorithms are show in Table 4. In this table shows the results of the correct classification percentage for multi classes: Geometric Mean (G-Mean), Precision, TPR, TNR, F1 Score and AUC. The results of the experimental values, generally obtained values with high AUC (greater than 73.90%) and TNR greater than 81.80% but TPR was 65.70% lower and G-Mean lower

73.44%. The best algorithm in terms of TPR and Accuracy is RF (78.12% and 83.07% respectively), which are high levels for predicting student performance. However, it is interesting to TNR, where RF gets the best result of 88.03% and G-Mean of 82.93, as well as Bagging and Boosting get the best TNR and G-Mean (86.60%, 80.73 and 86.26, 80.51%, respectively). RF classifiers obtain high TNR with high AUC, as a result they are an accurate but comprehensive classification model to predict student performance with the right trade-off regarding accuracy versus interpretation.

Table 4. Performance classifiers without genetic algorithm feature selection

Classifiers	G-Mean	Precision	TPR	TNR	F1 Score	AUC
DT	74.40	65.68	67.67	81.80	66.66	74.73
ANN	77.53	75.45	71.05	84.60	73.18	77.82
RF	**82.93**	**78.75**	**78.12**	**88.03**	**78.43**	**83.07**
Voting	73.44	75.19	65.70	82.10	70.12	73.90
Bagging	80.73	77.13	75.26	86.60	76.18	80.93
Boosting	80.51	75.26	75.14	86.26	75.20	80.70

In the next experiment, GA employed for features selection that executed for all the classifiers using x-fold cross validation. The results obtained after re-executing the 10 classification algorithms using x-fold cross validation are summarized in Table 5. The improved model for each classifier is highlighted width boldfaced print. To analyse and compare this table with the previous experiment, based on observation for each classifiers has improvement values obtained in all the evaluation measures, and some of them obtain the new best maximum values in almost all measures.

Further, the best overall results are those obtained by GA with RF models, which achieved a TNR 89.33% and a G-Mean 84.93%; these are the best results from all the experiments. It is therefore the classification model, which provides the most accurate and interesting result for prediction student performance. As shown in Table 5, all classifiers that implemented GA outperform the original method. It indicate that the GA based feature selection is effective to improve classification accuracy.

Table 5. Performance classifiers with genetic algorithm feature selection

Classifiers	G-Mean	Precision	TPR	TNR	F1 Score	AUC
GA+DT	81.05	75.37	75.76	86.70	75.56	81.23
GA+ANN	81.98	78.59	76.81	87.50	77.67	82.15
GA+RF	**84.93**	**81.64**	**80.74**	**89.33**	**81.18**	**85.03**
GA+Voting	79.98	79.29	74.03	86.40	76.57	80.22
GA+Bagging	82.25	79.53	77.13	87.70	78.31	82.41
GA+Boosting	83.01	79.37	78.13	88.20	78.74	83.16

Finally, in order to verify whether a significant difference between the proposed method with GA and a method without GA, the results of both methods are compared. This study performed the statistical t-Test for every classifier pair of both on student performance data set. In statistical significance, testing the P value is the probability of obtaining a test statistic at least as significance as the one that was actually observed, assuming that the null hypothesis is true. If P value is less than the predetermined significance level (α), indicating that the observed result would be highly unlikely under the null hypothesis. In this case, value of the statistical significance level (α) to be 0.05. It means that no statistically significant difference if P value > 0.05. The result is shown in Table 6, there are two classifiers (DT, ANN and RF) that have significant difference (P value < 0.05), the results have indicated that those of the rest classifiers (Voting, Bagging and Boosting) have no significant difference (P value > 0.05). The integration between GA and classifier achieved higher classification accuracy for most classifiers. Therefore, based on the results of research conducted shows the proposed method makes an impressive improvement in prediction performance.

Table 6. T test with/without genetic algorithm feature selection

Classifiers	*P* value T-test	Conclusion
DT	0.005	**Significant *P* value < 0.05**
ANN	0.025	**Significant *P* value < 0.05**
RF	0.304	**No Significant *P* value > 0.05**
Voting	0.193	**No Significant *P* value > 0.05**
Bagging	0.323	**No Significant *P* value > 0.05**
Boosting	0.237	**No Significant *P* value > 0.05**

Based on the results of the experiment, showing results methods that integrate genetic algorithms and random forest classification for student academic performance predictions get higher classification accuracy values. Genetic algorithms are applied to deal with high dimensional dataset problems, and RF classification overcomes the problem of class imbalanced. This research performed six classification techniques that were applied to student performance datasets from Kaggle data repository. Based on this, it can be concluded that the application of GA as a feature selection can improve predictive performance for all classifiers.

For future research, it will be related to the comparison of the methods proposed with other metaheuristic optimizations using optimization features selection techniques along with techniques such as PSO or optimization of ant colonies with other ensemble method techniques.

4. Conclusions

In this study, we have conducted experiments using six classifiers with feature selection techniques using genetic algorithms to predict student academic performance. Our experiments use public dataset for student performance from Kaggle repository. The experiments carried out were validated using x-cross fold validation and measured the results of validation through calculations using a confusion matrix. Genetic algorithm is applied to deal with the high dimensional dataset problem, and six classifiers technique is employed to alleviate the class imbalance problem. Therefore, we conclude that GA feature selection with Random Forest classifier method makes an impressive improvement for student academic performance prediction. Further research in our study will conduct experiment using other optimization techniques with different classification algorithms that aim to produce reliable models with high accuracy predictions.

5. References

[1]. S. Kotsiantis, "Educational data mining: a case study for predicting dropout-prone students," *International Journal of Knowledge Engineering and Soft Data Paradigms,* vol. 1, no. 2, pp. 101-111, 2009.

[2]. A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access,* vol. 5, pp. 15991-16005, 2017.

[3]. 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, no. 6, pp. 601-618, 2010.

[4]. X. Zhang, and B.-G. Hu, "A new strategy of cost-free learning in the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 12, pp. 2872-2885, 2014.

[5]. J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data mining and knowledge discovery,* vol. 1, no. 1, pp. 55-77, 1997.

[6]. L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing,* vol. 105, pp. 3-11, 2013.

[7]. C. Sammut, and G. I. Webb, *Encyclopedia of machine learning*: Springer Science & Business Media, 2011.
[8]. M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Systems with Applications,* vol. 39, no. 3, pp. 3747-3763, 2012.
[9]. S. C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," *Pattern Recognition Letters,* vol. 30, no. 5, pp. 525-534, 2009.
[10]. G. Collell, D. Prelec, and K. R. Patil, "A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data," *Neurocomputing,* vol. 275, pp. 330-340, Jan 31, 2018.
[11]. E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict Student's academic performance using ensemble methods," *International Journal of Database Theory and Application,* vol. 9, no. 8, pp. 119-136, 2016.
[12]. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.