Genetic Algorithm Based Feature Selection for Predicting Student's Academic Performance

Al Farissi[1](✉), Halina Mohamed Dahlan[2], and Samsuryadi[1]

[1] Fakultas Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia
{alfarissi, samsuryadi}@unsri.ac.id
[2] Information Systems Department, Azman Hashim International Business School, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
halina@utm.my

Abstract. Recently, student's academic performance prediction has become an increasingly prominent research topic in the field of Educational Data Mining (EDM). The prediction of student's academic performance aims to explore information that is beneficial to the learning process of student. Therefore, accurate prediction of student's academic performance provide benefits for education institutions to improve the quality of their institutions by improving the learning process of students. In predicting the student's academic performance, the problem of high dimensional dataset is often faced in the datasets which significantly impacts the accuracy of student academic performance prediction. This paper proposed Genetic Algorithm based Feature Selection (GAFS) along with selected single classifier for classification in order to improve the accuracy in predicting student academic performance. Kaggle dataset is used in this paper and two phase of experiment have been conducted, single classifier without GAFS, and single classifier with GAFS. Results from the experiments show that, the accuracy of the proposed GAFS for classification makes an impressive performance in predicting student academic performance in terms of accuracy compare to existing techniques.

Keywords: Student academic performance · Feature selection · Genetic Algorithm · Classification · Prediction

1 Introduction

The challenge in predicting student academic performance is increasing due to the greater data in the education database. Prediction and analysis of students academic performance has an important role for students academic development. Identifying factors that influence students academic performance is a complex research task [1]. Predicting student academic performance with high accuracy values will be useful for educational institutions to be able to distinguish the academic performance of the students they have [2]. Educational data mining is able to produce information used by

F. Saeed et al. (Eds.): IRICT 2019, AISC 1073, pp. 110–117, 2020.
https://doi.org/10.1007/978-3-030-33582-3_11

educational institutions in developing educational strategies to improve the quality of education.

Problems that are often encountered in dataset are: high dimensional dataset and noisy attributes that can influence the predictive results. Feature selection is commonly been employed in solving the problem involving high dimensional dataset and noisy attributes [3].

In this paper, four classifiers have been applied, which are Decision Tree (DT), NaïveBayesian (NB), k-NearestNeighbor (k-NN) and Random Forest (RF) with the proposed GAFS. In order to measure the accuracy of the combination, four evaluation measurements are used which are: Accuracy, Precision, Recall and F-Measure.

The rest of this paper is organized into several sections, Sect. 2 presents related work in the field of student academic performance prediction, the student academic performance prediction framework in Sect. 3. Experimental design are presented in Sect. 4 and results and discussions are detailed out in Sect. 5. Final section of this paper, is conclusion, in Sect. 6.

2 Related Work

2.1 Feature Selection

Feature selection (FS) has been widelybeen applied in various applications, except that it has not been so widely applied in predicting student academic performance applications. The main purpose of using feature selection techniques is to minimize redundancy and maximize the subset of relevant features while maintaining high accuracy without losing important information.

Amrieh et al. [4] applied Filter Based Selection (FBS) as a feature selection in their study and combined with the esemble method of the classification: Boosting & Bagging. They concluded that the accuracy produced using FS & ensemble methods increased by 25.8% where the Visited Resource feature was the most effective feature on they student academic performance.

Punlumjeak, W & Rachburee, N [3], in their study applied two feature selection approaches: filter & wrapper approach. They stated that the results of the experiments showed the choice of the most relevant features from the list of features used in the student dataset affecting the prediction model of student academic performance.

Luthfia et al. [5] applying Wrapper and Information Gain approaches as feature selection techniques with a single classification algorithm: NB (Naive Bayes), DT (Decision Tree) and NN (Neural Network). They get better accuracy values after applying feature selection techniques to the models they developed. The combination of Information gain and ANN resulted in an accuracy value of 79.375%.

Zaffar et al. [6] used six feature selection techniques and fifteen single classfiers. From the results of their study, the Principal Components Analysis feature selection technique with the Random Forest classifier gets the best results. Their experimental results illustrate that, a way is needed to set parameters in the feature selection method, in order to produce better performance.

Calderon et al. [7], in their experiments, Genetic Algorithm (GA) feature selection was applied with the ANN algorithm for 1271 data sets with 39 attributes, the data set is the property of the university in southern Peru. The parameters of the Genetic Algorithm adjustment: population size is 25, the crossover level is 0.6, the mutation level is 0.3. For the ANN algorithm parameters are applied three layers (1 input layer, 1 hidden layer, and 1 output layer). The experimental results selected 14 attributes used for predictions. Accuracy value result 88.50% increased by 8% compared to predictions using only ANN classifier.

There have been many study in data mining that have been produced very well by using feature selection techniques. In data mining, feature selection at the preprocessing data has an important role. The feature selection technique is to get a feature subset in order to reduce the occurrence of prediction errors. In addition, the selection of features aims to produce attributes that have strong relevance in order to produce high accuracy predictive values.

However, the majority of feature selection algorithms work in getting feature selection solutions that range between sub-optimal and almost optimal regions, these searches are carried out locally throughout the search process, rather than searching globally. Therefore, it is difficult to get an almost optimal solution to the optimal solution if using this algorithm [8].

The purpose of this paper is to improve the accuracy of prediction in the context of students' academic performance by using GAFS as well as to overcome the problems facing by the current feature selection. GA capable to obtain solutions throughout the search space by using global search capabilities, therefore feature selection using genetic algorithm can produce good solutions in a reasonable period of time [9].

3 Method

3.1 Framework

In this section, the proposed framework as shown in Fig. 1. This framework divided into two steps, which are feature selection and modeling. In the first step which is feature selection, dataset is separated into training data and testing data. Next, GAFS is employed in order to get features subset to improvement accuracy prediction.

In the second step which is modeling, data training with feature subset selected is trained with the selected classifier. Classification accuracy is measured by testing set with selected feature subset. Then a fitness function is constructed using classification accuracy of classifier, the number of selected features and the feature cost.

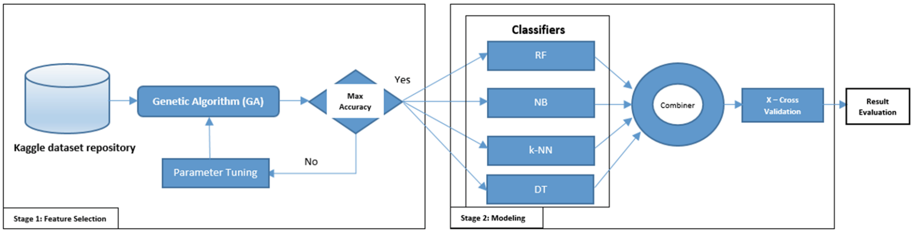

Fig. 1. Framework for student academic performance prediction

Parameters tuning for the Genetic Algorithm Feature Selection (GAFS) and classifiers is shown in the Table 1. These parameters are as follows; initial population values (30), maximum number of generations (30), mutations (0.01), probability of crossover (0.9), k (10), and random number of seeds (1). When the final condition reached, the operation will be stopped.

However, if the conditions are not met, it will go forward with the following generation of operations. The framework searches for better solutions by genetic operations, selection and and mutation, including crossover.

For the value of *k* in the k-NN classifiers, are *k* = 3, 5, 7, and 9. The best accuracy results obtained using the k-NN algorithm are with the value *k* = 5. The performance of base classifiers as single classifier is validated using X-fold cross validation.

Table 1. Parameters tuning for GAFS and selected classifiers

GAFS	RF	NB	k-NN	DT
Initial population: 30 Number of generations: 30 Mutations: 0.01	Number of trees: 1000 Criterion: gain_ratio Maximal depth: 30 Pruning number: 3	Laplace correction: true	Value k training: 5 Weighted voted: true Measure type: Mixed Measure Euclidean Distance	Criteria: gain ratio. Maximal depth: 10. Confidence level: 0,1. Minimal gain:0,01.

4 Experimental Design

4.1 Experimental Setup

Two phase experiments have been done. Phase one is classification without GAFS. In this phase, all classifier which are DT, NB, k-NN and RF are employed without GAFS. In Phase two, GAFS is employed to each classifier.

This experiment conduct used Intel Processor i7 fifth generation with 2.4 GHz CPU, 12 GB Memory RAM and Microsoft Win 10 Edition 64Bit Operating System.

4.2 Dataset Description

This study uses student academic performance dataset from the Kaggle repository. This data collection contains 480 student academic performance data and consists of 16 attributes with one class label that has three intervals: low, medium and high [10]. The features and descriptions of the data set are shown in Table 2.

Table 2. Features and descriptions of student dataset

No	Feature name	Description
1	Nationality	Nationality
2	Gender	Male or female
3	POB	Birthplace
4	Parent responsible for student	Status parent
5	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	Parent answering survey	Parent participation on survey
12	Parent school satisfaction	Level satisfaction of parent
13	Discussion activity	Student interaction with e-learning system
14	Active visiting resources	
15	Raised hand on class	
16	Seeing announcements	

4.3 Model Validation

This paper employed x-fold cross validation to validate training and testing data. Usually, x-fold cross validation is used because it can reduce computing time while maintaining the accuracy of estimates. This method distributes training data into 10 equal parts. The next step, the learning process is carried out ten times as shown in Table 3.

Table 3. X-fold cross validation

n-validation	Dataset Partition									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

4.4 Evaluation Measures

Four general measurements to evaluate quality of the classification are employed: Accuracy, Precision, Recall and F-Measure. Measures calculated using confusion matrix classification based on the Eqs. 1, 2, 3 and 4, respectively.

Accuracy values refer to how accurate the classification can classify data correctly. Equation 1 is used to obtain the accuracy value. In other words, the value of accuracy is a comparison between data that is correctly classified with the whole data. Precision values refer to the number of positive category data that are classified correctly divided by the total data classified as positive. Equation 2 is used to obtain precision. Meanwhile, recall shows how many percent of the positive category data is correctly classified by the classification. Equation 3 is used to obtain the recall value.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F - Measure = \frac{2TP}{TP + FP + FN} \tag{4}$$

5 Result and Discussion

The evaluation result using classifier with all feature is shown in Table 4, and the result of classifier using with GAFS is shown in Table 5. By using GAFS shows a higher increase in accuracy than classifiers without using feature selection. Therefore, the use

Table 4. Classification method result with all attributes

Evaluation measure	Decision tree (DT)	k-nearest neighbour (k-NN)	Naive Bayes (NB)	Random forest (RF)
Accuracy	62.71	61.04	57.50	**79.79**
Recall	62.70	62.04	59.96	**80.34**
Precision	64.73	61.61	58.14	**80.42**
F-measure	63.70	61.79	58.76	**80.28**

of GAFS in classifying techniques is a solution to improve accuracy in predicting student academic performance.

Table 5 shows the classification results using GAFS with classifiers: DT, k-NN, NB and RF. From the results obtained, RF classifier outperform compared to other classifiers. Overall, the most superior performance was shown when GA feature selection was combined with RF classifier. The results of the accuracy obtained were 82.29%, RF successfully classifies student data correctly with 395 data from a total of 480 student data.

Table 5. Classification method result with GA feature selection

Evaluation measure	Genetic algorithm feature selection (GAFS)			
	Decision tree (DT)	k-nearest neighbour (k-NN)	Naive bayes (NB)	Random forest (RF)
Accuracy	74.58	68.54	75.42	**82.29**
Recall	75.15	68.84	76.25	**82.81**
Precision	75.39	69.75	76.26	**82.70**
F-measure	75.26	69.29	76.25	**82.75**

6 Conclusion

In this paper, DT, k-NN, NB and RF classifier have been used with GAFS techniques with the aim of increasing the performance accuracy for student academic prediction. The dataset used containing the performance of students from Kaggle repository. This study succeeded in applying feature selection techniques to reduce the dimensions of data set with the intention of to improve predictive accuracy. From the experimental results show impressive improvements to predict student academic performance. Future research will measure the comparison of several proposed optimization techniques using other feature selection techniques with other ensemble methods.

References

1. Mueen, A., Zafar, B., Manzoor, U.: Modeling and predicting students' academic performance using data mining techniques. Int. J. Mod. Educ. Comput. Sci. **8**, 36–42 (2016)
2. Zimmermann, J., Brodersen, K.H., Heinimann, H.R., Buhmann, J.M.: A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. JEDM-J. Educ. Data Min. **7**, 151–176 (2015)
3. Punlumjeak, W., Rachburee, N.: A comparative study of feature selection techniques for classify student performance. In: 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 425–429. IEEE (2015)
4. Amrieh, E.A., Hamtini, T., Aljarah, I.: Mining educational data to predict student's academic performance using ensemble methods. Int. J. Database Theory Appl. **9**, 119–136 (2016)
5. Rahman, L., Setiawan, A.N., Permanasari, E.A.: Feature selection methods in improving accuracy of classifying students' academic performance. In: 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 267–271. IEEE, Yogyakarta (2017)
6. Zaffar, M., Savita, K.S., Hashmani, M.A., Rizvi, S.S.H.: A study of feature selection algorithms for predicting students academic performance. Int. J. Adv. Comput. Sci. Appl. **9**, 541–549 (2018)
7. Echegaray-Calderon, O.A., Barrios-Aranibar, D.: Optimal selection of factors using genetic algorithms and neural networks for the prediction of students' academic performance. In: 2015 Latin America Congress on Computational Intelligence (LA-CCI). IEEE, Curitiba (2015)
8. Kabir, M.M., Shahjahan, M., Murase, K.: A new hybrid ant colony optimization algorithm for feature selection. Expert Syst. Appl. **39**, 3747–3763 (2012)
9. Yusta, S.C.: Different metaheuristic strategies to solve the feature selection problem. Pattern Recognit. Lett. **30**, 525–534 (2009)
10. Amrieh, E.A., Hamtini, T., Aljarah, I.: Preprocessing and analyzing educational data set using X-API for improving student's performance. In: IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). IEEE, Amman (2015)