Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining

Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining

Hasniza Hassan[1], **Nor Bahiah Ahmad**[1] **and Syahid Anuar**[2]

[1] Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[2] Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia

nieza1212@gmail.com

Abstract. Among the problems raised in the data mining area, the class imbalance is a well-known issue that always occurs. Many researchers studied this issue in several fields using three commonly used techniques: sampling, ensemble, or cost-sensitive learning. However, such studies are still new in education domains. This problem always related to the quality of data that gives the most impact to form an accurate prediction result. Many previous studies focus on binary imbalance classification problems instead of the multi-class imbalance problem in education data. This study used 4413 student instances of two datasets; students' information system and e-learning from the Faculty of Engineering in a Malaysia university for First Semester 2017/2018. Three sampling categories utilized in this study are oversampling techniques, undersampling techniques, and hybrid techniques. The research empirically analyzes five types of ensemble classifiers and seven sampling techniques. The experimental results show a hybrid technique ROS with AdaBoost produces the most excellent performance compared to the other benchmark techniques. SMOTEENN technique with ensembles classifiers consistently produces high results. This technique has great potential in improving the students' performance prediction model.

1. Introduction

Due to education data increased significantly, prediction model development with high accuracy is so important to evaluate and improve students' accomplishment. The academic data is the best predictor in data mining. Besides that, studies by many researchers have proven that demographics and behaviours of students' factors using ensemble learning contributes to building good students' performance prediction model [1-5].

Data preprocessing is the most necessary action to form students' performance prediction models in higher education. The massive number of data might cause data to suffer from noise and other problems like class imbalance. Hence, features selection and noise filtering are essential steps to ensure data is ready for modeling. When trained a machine learning model using imbalanced data, the model will be more towards the majority classes compared to minority class[6]. Consequently, the model will be biased to the majority class.

Class imbalance is one of the ten problems raised in the field of data mining [7]. Many techniques developed to solve the class imbalance problems. However, most of them are to handle binary class imbalance and only a few studies on multi-class imbalance issues for education. Hence, the multi-class imbalance problems need more research for performance prediction model improvement [8].

Multi-class imbalance used class decomposition to convert multi-class to binary. One-versus-all or also known as one-versus-rest (OVA or OVR) and one-versus-one (OVO) types of class decomposition always being used in the literature review [11].

This work compared five ensemble models and have applied sampling techniques to multi-class imbalanced data. The best features based on ranking selected before balanced the data using sampling techniques and learning the data using machine learning classifiers. This paper aims to identify the solution of noise problem and imbalanced in multi-class data to improve students' performance prediction model by evaluating the most suitable imbalanced data classification technique and machine learning algorithm to identify at-risk students.

This study guided by a few research questions: 1. What are the students' data features that might influence the students' performance? 2. What are the suitable ensemble classifiers and imbalanced data classification techniques to improve students' performance prediction models? 3. Do fine-tuning hyperparameters help in improving ensemble classifier performance after applying the imbalance technique?

This paper presents the sequential of sections as follows: Section 2 described the methodology and tools used for the study. This including data collection, integration, preprocessing and modeling. Conversely, this section explains about the classifiers and type of sampling methods. Section 3 shows and discusses the empirical experiment results by using ensemble algorithms and sampling methods. Finally, section 4 concludes the overall study and suggestions for future work.

2. Methodology

This section defined the research problem and described the implementation of the model in detail.

2.1. Datasets Collection

This study combines two real data from the students' information system (SIS) and e-learning (EL) logfile. The data collected from a public research university in Malaysia in the first session of 2017/2018. This study used 4413 numbers of undergraduate students' data from the Faculty of Engineering [9].

2.2. Tools and instruments

This study used Python and Jupyter Notebook to execute the experiment. Scikitlearn python package used for the machine learning algorithm and Imblearn python package used to resample the imbalance class.

2.3. Preprocessing

Preprocessing is an essential step in data mining to increase data quality and assure the modeling process be more efficient [10]. By using Scikitlearn preprocessing package in python, sequences of preprocessing steps done. First, preliminary features filtering executed where the unrequired data removed from the list. Next, the transformation process performed where the categorical data converted to numerical values to ensure they are suitable for modeling. Finally, the normalization process executed to convert all figures to the small range and ready for the modeling process.

2.4. Feature Selection

Features selection reduces data dimensional to improve the result of data mining. This study used feature selection with the filter method that supports by statistical evidence. Steps are including sorting and filtering the features. Finally, the top-ranking features will be selected [8].

There are 19 features selected from the first stage filtering. The features classified into three groups: academic background, demographical with socioeconomic and behavior e-learning. 13 top features out of 19 features used as the predictors (bold features) as shown Table 1.

Table 1. Features Categories and Description [8]

Category	Feature	Description
Academic Background Features	Study_Method	Coursework or research
	Programme	Programme of study
	CGPA	Cumulative Grade Point Average
	Year_Intake	Students' intake in year
	Education_Mode	Part time/Full time
Demographics/ Socioeconomic Features	**Family_Income**	Family income range
	Student_Status	Student status
	Scholarship	Name of Scholarship
	Gender	Student Gender
	Age	Student Age
	Nationality	Student Nationality
	Disability	Disability status
Behaviour Features (E-Learning)	**User_Loggedin**	Count no of login
	Course_Viewed	Count course viewed
	Course_Module_Viewed	Count resources viewed
	Discussion_Viewed	Count forum/discussion viewed
	Submission_Form_Submitted	Count course submitted
	Attempt_Viewed	Count assignment viewed
	Assesable_Submitted	Count assignment submitted

The students' CGPA feature chose as the experiment output or label. This study transformed the CGPA numerical list to 3 classes of categorical formed: low, moderate and excellent according to university standard grading [9]. ClassLabel used as the column name for the label.

2.5. Resampling Technique

Resampling is the method that develops techniques to solve the imbalanced data problem by balancing the imbalanced classes. There are three categories of sampling methods[7] :

- Oversampling Techniques-This technique will be duplicated or generate new minority class instances [9]. It usually used for the small and average size of data. The examples of this technique are Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling (ROS) and Adaptive Synthetic Sampling (ADASYN). In 2002, a reported work of [10] proposed an oversampling technique named SMOTE that able to create a minority synthetic class. It became more popular and the most frequently used in imbalanced classification problems. [16]. However, steps to duplicate or creating new instances in oversampling techniques can cause overfitting [6].
- Undersampling Techniques-This technique will remove data instances from the majority class [9]. Random Undersampling (RUS), Edited Nearest Neighbors (ENN) and Tomek Links (TL) are a few examples of the undersampling techniques. The technique is suitable for vast data. Despite its strength, this technique has a weakness where data reduction will cause information loss.
- Hybrid Techniques-This technique combined oversampling with undersampling or oversampling or undersampling with ensemble techniques. Some examples of this technique are SMOTE-ENN, SMOTE-TL, SMOTEBoost and RUSBoost.

2.6. Classification using Machine Learning Classifiers

This study is using five different techniques of ensemble learning. By default, all ensemble learning classifiers are using the Decision Tree method as a base algorithm. The brief description is as follows:

- Decision Tree (DT) is a partitioning based modelling algorithm that is commonly used [11], [12]. It splits data points into two groups to gain possible answers to the questions.
- Random Forest (RF) ensemble classifier contains many decision trees model and fall to a type of bagging ensemble. Increase number of Decision Tree contributes to the robust Random Forest model. It usually used to perform regression and classification tasks. Its algorithm produces a class that is the mode of overall individual decision trees. Random forest might overfit with a small dataset [3].
- Bootstrapping Aggregating or Bagging is an ensemble classifier that increases the accuracy of weak classifiers. It trained leaners in parallel and learns them differently. It consolidates each learned classifier result using an averaging process [2].
- Boosting is one of the ensemble techniques that manages to convert classifiers that weak to become stronger. It selects wrong predictions data points produced by the previous learner and adjusts their weight. It also can reduce bias and variance in supervised learning.
- AdaBoost is a boosting type algorithm that eliminated the limitation of Boosting. It focused more on the part or pattern that is difficult to classify. Weights are assigned equally to all subsets. It will increase the misclassified instance weight but decrease the truly classified instance weight. Finally, the voting process used to incorporate groups of weak learners to become stronger learners [2].
- Gradient boosting is a type of algorithm to find approximate solutions to the additive modelling problem. Like AdaBoost, Gradient boosting builds weak learners a sequentially. Gradient boosting generates learners during the learning process.
- The extreme Gradient Boosting (XGBoost) technique is a powerful ensemble learning technique. It used the Gradient Boosting designed framework to increase speed and performance.

2.7. Performance Measure

This study used four classification performance metrics to evaluate the performance of machine learning models. The formula is as follows:

$$Accuracy = \frac{TP}{TP+TN+FP+FN} \quad (1)$$

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

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

$$Fmeasure = 2\,X\,\frac{Precision\ X\ Recall}{Precision+Recall} \quad (4)$$

The confusion matrix used to explain the classification model. It classifies instances to four label features: True Positive group, True Negative group, False Positive group, and False Negative group. For the positive group, data will be True Positive when real positive data predicted to be positive. Data will be True Negative when data predicted to be negative. In false groups, data will be False Positive (FP) when it is negative but predicted as positive. Data classified as False Negative (FN) when it is positive but predicted as negative [13].

2.8. Modelling

The data divided into two categories, train and test with ratio 70:30. Then, the 10 folds cross-validation technique applied to every classification model. Grid search used to find the best hyperparameters for every model to get the optimum result.

3. Result and Discussion

3.1. Experiment 1

Table 2 shows the first experiment executed. This experiment used three categories of data: student information system (SIS), e-learning (EL) and a combination of the student information system (SIS) and e-learning (EL). Five ensemble learning models trained: Random Forest (RF), Bagging (BGG), AdaBoost (AB), Gradient boosting (GB) and XGBoost (XGB) to evaluate the performance.

Table 2. Imbalance Results

Algorithm	Features	Accuracy	Precision	Recall	F-measure
Random Forest	SIS	0.695	0.489	0.404	0.409
	EL	0.631	0.415	0.366	0.378
	SIS + EL	0.653	0.517	0.399	0.418
Bagging	SIS	0.692	0.445	0.413	0.413
	EL	0.629	0.385	0.375	0.362
	SIS + EL	0.657	0.481	0.425	**0.425**
AdaBoost	SIS	0.690	0.482	0.407	0.416
	EL	0.643	0.393	0.369	0.365
	SIS + EL	0.654	0.524	0.403	0.395
Gradient boosting	SIS	0.718	0.433	0.404	0.400
	EL	0.689	0.399	0.360	0.330
	SIS + EL	0.706	0.506	0.408	0.407
XGBoost	SIS	0.723	0.442	0.408	0.405
	EL	0.695	0.417	0.360	0.328
	SIS + EL	0.707	0.510	0.402	0.398

The experiment result shows a training model using the XGBoost classifier and SIS data produce the highest accuracy. However, accuracy is not suitable when the classes are imbalance since it does not distinguish correct classified data into multi-class. This study focuses on F-Measure that is mean of precision and recall. Based on experiment results, the highest F-measure gained when trained model using Bagging Classifier and a combination of the SIS and EL data.

According to [2], students' behaviour features contribute to the improvement of accuracy in the model proposed. They discovered a strong relationship between academic and behaviour features. Moreover, [14] and [15] also discovered by having behaviour features, performance accuracy in students' performance prediction models will be improved. A finding by [16] proves the robust relationship between students' behaviour and academic performance. They discovered that by using behaviour features using hybrid classification and clustering technique, prediction result improved. Research by [2], [3], [11], [17]–[22] discovered the ensemble techniques improve students' prediction model. They prove the ensemble technique manages to increase the students' performance prediction model.

3.2. Experiment 2

The second experiment employed sampling methods to overcome the multi-class imbalance problems. Table 3 shows the results of the empirical study using the F-Measure performance metric after applying the sampling techniques in three categories for imbalanced data. The experiment trained models using five ensemble classifiers and a combination of SIS and EL datasets.

Among the classifier trained, all show improvement after balancing the multi-class labels. The highest accuracy result gained after applied AdaBoost ensemble classifiers and ROS. However,

SMOTEENN frequently obtains the highest results among all imbalance techniques employed. These show that SMOTEENN is the most suitable hybrid technique to the education data used in this study.

Table 3. Applied Imbalance Techniques

Type	Technique	RF	BGG	AB	GB	XGB
Oversampling	SMOTE	0.750	0.753	0.757	0.582	0.547
	ROS	**0.870**	0.862	**0.916**	0.662	0.628
	ADASYN	0.716	0.714	0.773	0.630	0.649
Undersampling	RUS	0.548	0.529	0.508	0.536	0.547
	NearMiss	0.792	0.781	0.768	0.771	0.709
Hybrid	SMOTEENN	0.865	**0.872**	0.839	**0.801**	**0.773**
	SMOTETL	0.765	0.769	0.820	0.667	0.650

A study by [6] proved that four classifiers: Logistic Regression, Support Vector Machine, Random Forest and AdaBoost show better performance using two noise filter methods with the class imbalance approach: BST-CF and BST-EF. Research done by [23] proposed the hybrid technique, Bagged NBDT, which are hybrid of bagging and weak classifiers, Naive Bayes and Decision Tree for the multi-class problem that executed on 52 datasets. Experiment results prove the technique outperforms a few ensemble methods used. In another study by [24] proved that a new multi-class imbalance classification algorithm named Diversified Error-Correcting Output Codes (DECOC) produced significant improvement when train using Regression, C4.5, AdaBoost, Random Forests, CART and Multilayer perceptron compare to 17 other benchmark classifiers. A different study by [25] proposed two novel methods, bagging with ADASYN sampling technique and bagging with hybrid technique RSYN. Results show the proposed methods give good results compared with the existing best performing method on 11 imbalanced datasets.

4. Conclusion and Future works

Student achievement and performance are essential criteria that have been monitor frequently by education sectors, particularly by the management of institutions. The institutions are generally taking into account and consistently emphasising the importance of identifying students' performance. Educational Data Mining used many machine learning techniques to identify students' performance. This study examined different ensemble prediction methods to identify method can produce better students' performance prediction result. The experiment executed shows the combination of 13 features as predictors trained using ensemble methods. The experiment result shows the sampling techniques and ensembles classifiers improve the students' performance prediction model. By implement data preprocessing, cross-validation technique with the 10-folds and the algorithms fine-tuning using the grid search method improved the effectiveness of the performance prediction model.

The experiment shows that the highest result gain when training the model using AdaBoost Ensemble Classifier and balanced class using Random Oversampling (ROS). However, SMOTEENN frequently performs well among all sampling techniques used. Results obtained from the model of students' predictions could potentially help educational institutions and educators to monitor students' achievement and identify at-risk students from the earlier stage. The hybrid approach has been proven effective throughout the experiments. It has great potential in facilitating higher education institutions to execute many more accurate prediction models and reduced the at-risk prediction.

In future research, the limitation of this paper will be improved by develops more combinations of hybrid techniques to solve the multi-class classification problem. Then, experiments need to use cross-domain data to have a broad overview of the benchmark. Subsequently, future study might also focus more on the hybrid of ensemble techniques with different base-classifiers, with more hyper-tuning parameters.

Acknowledgement
The authors are grateful to the Ministry of Education and Universiti Teknologi Malaysia for supplied the data.

References

[1] Adejo O and Connolly T 2017 An integrated system framework for predicting students' academic performance in higher educational institutions *Int. J. Comput. Sci. Inf. Technol.* vol. **9** no. 3 pp.149–157
[2] Amrieh E A, Hamtini T and Aljarah I 2016 Mining educational data to predict student's academic performance using ensemble methods *Int. J. Database Theory Appl.* vol. **9** no. 8 pp. 119–136
[3] Salini A and Jeyapriya U 2018 A majority vote based ensemble classifier for predicting students academic performance *Int. J. Pure Appl. Math.* vol. **118** no. 24
[4] Cerezo R, Sánchez-Santillán M, Paule-Ruiz P M and Núñez J C 2016 Students' LMS interaction patterns and their relationship with achievement: a case study in higher education *Comput. Educ*
[5] Hasibur Rahman M and Rabiul Islam M 2018 Predict student's academic performance and evaluate the impact of different attributes on the performance using data mining techniques *2nd Int. Conf. Electr. Electron. Eng. ICEEE 2017*
[6] Radwan A M and Cataltepe Z 2017 Improving performance prediction on education data with noise and class imbalance *Intelligent Automation and Soft Computing*
[7] Yang Q and Wu X 2006 10 challenging problems in data mining research *Int. J. Inf. Technol. Decis. Mak.* vol. **5** no. 4
[8] Wang S and Yao X 2012 Multiclass imbalance problems: Analysis and potential solutions *IEEE Trans. Syst. Man Cybern. Part B Cybern.* vol. **42** no. 4
[9] Hassan H, Anuar S and Ahmad N B 2019 Students' performance prediction model using meta-classifier approach in Higher Education *Communications in Computer and Information Science* pp. 221-231
[10] Costa E B, Fonseca B, Santana M A, de Araújo F F and Rego J 2017 Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses *Comput. Human Behav.* vol. **73** pp. 247–256
[11] Yang X, Kuang Q, Zhang W and Zhang G 2017 AMDO: An Over-Sampling Technique for Multi-Class Imbalanced Problems *IEEE Trans. Knowl. Data Eng.* vol. **30** no. 9 pp. 1672–1685
[12] Blagus R and Lusa L 2013 SMOTE for high-dimensional class-imbalanced data *BMC Bioinformatics*
[13] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res*
[14] Adejo O W and Connolly T 2018 Predicting student academic performance using multi-model heterogeneous ensemble approach *J. Appl. Res. High. Educ.* vol. **10** no. 1 pp. 61–75
[15] Francis B K and Babu S S 2019 Predicting academic performance of students using a hybrid data mining approach *J. Med. Syst*
[16] AL-Malaise A, Malibari A and Alkhozae M 2014 Students performance prediction system using multi agent data mining technique *Int. J. Data Min. Knowl. Manag. Process* vol. **4** no. 5
[17] Nam S J, Frishkoff G and Collins-Thompson K 2017 Predicting students' disengaged behaviors in an online meaning-generation task *IEEE Trans. Learn. Technol*
[18] Zollanvari A, Kizilirmak R C, Kho Y H and Hernandez-Torrano D 2017 Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors *IEEE Access* vol. **5** pp. 23792–23802
[19] Sun Z, Sun L and Strang K 2018 Big data analytics services for enhancing business intelligence *J. Comput. Inf. Syst.* vol. **58** no. 2 pp. 162–169

[20] Pandey M and Taruna S 2014 A comparative study of ensemble methods for students' performance modeling *Int. J. Comput. Appl.* vol. **103** no. 8 pp. 26–32
[21] Satyanarayana A and Ravichandran G 2016 Mining student data by ensemble classification and clustering for profiling and prediction of student academic performance *ASEE Mid-Atlantic Sect. Conf*
[22] Iam-On N and Boongoen T 2017 Improved student dropout prediction in Thai university using ensemble of mixed-type data clusterings *Int. J. Mach. Learn. Cybern*. vol. **8** no. 2 pp. 497–510
[23] Ashraf M, Zaman M and Ahmed M 2018 Using ensemble stackingC method and base classifiers to ameliorate prediction accuracy of pedagogical data *Procedia Comput. Sci.* pp. 1021–1040
[24] Beemer J, Spoon K, He L, Fan J and Levine R A 2018 Ensemble learning for estimating individualized treatment effects in student success studies *Int. J. Artif. Intell. Educ*. vol. **28** no. 3 pp. 315–335
[25] Wanjau S K and Muketha G M 2018 Improving student enrollment prediction using ensemble classifiers *Int. J. Comput. Appl. Technol. Res*. vol. **7** no. 3 pp. 122–128
[26] Singh N Singh P 2019 A novel bagged naıve bayes-decision tree approach for multi-class classification problems *Journal of Intelligent and Fuzzy Systems* p. 2261
[27] Bi J and Zhang C 2018 An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme *Knowledge-Based Syst*. vol. **158** pp. 81–93
[28] Ahmed S, Mahbub A, Rayhan F, Jani R, Shatabda S and Farid D M 2018 Hybrid methods for class imbalance learning employing bagging with sampling techniques *2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut.* CSITSS 2017 pp. 126-131