International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 4, August 2021, pp. 3567~3574
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i4.pp3567-3574

Oversampling technique in student performance classification from engineering course

Nachirat Rachburee, Wattana Punlumjeak
Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathum Thani, Thailand

Article Info

Article history:

Received Jul 31, 2020
Revised Dec 24, 2020
Accepted Jan 18, 2021

Keywords:

Classification
Educational mining
Oversampling

ABSTRACT

The first year of an engineering student was important to take proper academic planning. All subjects in the first year were essential for an engineering basis. Student performance prediction helped academics improve their performance better. Students checked performance by themselves. If they were aware that their performance are low, then they could make some improvement for their better performance. This research focused on combining the oversampling minority class data with various kinds of classifier models. Oversampling techniques were SMOTE, Borderline-SMOTE, SVMSMOTE, and ADASYN and four classifiers were applied using MLP, gradient boosting, AdaBoost and random forest in this research. The results represented that Borderline-SMOTE gave the best result for minority class prediction with several classifiers.

Corresponding Author:

Nachirat Rachburee
Department of Computer Engineering
Faculty of Engineering, Rajamangala University of Technology Thanyaburi
Pathum Thani, Thailand
Email: nachirat.r@en.rmutt.ac.th

1. INTRODUCTION

Nowadays, educational data mining (EDM) is widely used to help student and teacher manage their proper learning environment. Student performance, grade prediction or student dropout prediction are developed in several case studies. Some of the works focus on the classification model and feature selection. However, some studies are the approach to deal with imbalanced data which affects to performance of classification accuracy in minority class. This paper focuses on first year subject of engineering course that all engineering student must take and pass.

One problem of student performance prediction in this research is the imbalanced data from 3 different classes (low, medium and high). We use synthetic minority oversampling techniques (SMOTE), Borderline-SMOTE, SVMSMOTE, and ADASYN to generate synthetic instance data. Then, we apply oversampling data with various classifier techniques to improve performance of classification in minority class. Moreover, student can take proper action if they got a low class predicted.

SMOTE algorithm is the oversampling technique to do resampling and balancing the data set. Resample data was created by interpolation within minority class data. This algorithm relies on the relation of data values. The synthetic data was created by minority class instances. The synthetic data relied on distance and the nearest neighbors which was randomly selected from the same class. Firstly, minority class sample, k of nearest neighbors and the amount of oversampling was setup. Secondly, k nearest neighbors were randomly selected from minority sample class then generated new instance data by interpolation [1].

In this paper, the author reviewed and analyzed various oversampling techniques and many imbalanced datasets. The best result showed 4 classification measures of imbalanced learning. The four first oversampling techniques was polynomial-fitSMOTE, SMOTE-IPF, ProWSyn and Lee.

In [2], the Author reviewed several sampling techniques to deal with the imbalanced dataset. Not only the prediction accuracy rate but also the learning time were appropriated to evaluate the model with balanced dataset. SMOTE and SMOTEBoost were represented as an oversampling technique in this research. The Receiver Operating Characteristic (ROC) curve was used to summaries classifier performance. The area under the curve (AUC) was used to evaluate classifier performance for ROC curve.

This paper used a wrapper approach technique to reduce the dimension of student imbalanced data set. The metric was true positive (TP) rate [3]. In [4, 5], the researcher handled unbalanced data set by a hybrid resampling technique. The combination of SMOTE and DBSCAN technique was used for 12 oversampling data then result represented the highest performance of DBSM by AUC and F-measure.

In [6], the Author used SMOTE and boostrapping to handle unbalanced data. They used several feature selection methods with decision tree, k-NN and Bayes classification model. They found SMOTE and boostrapping increased the accuracy of classification.

Adaptive synthetic (ADASYN) sampling was the oversampling technique by generating minority class instances. ADASYN sampling relied on data density distributions. This technique emphasized on difficult sample set and different weight on different minority instances to compensate for slope distribution [7]. The amount of synthetic instance data depended on density distribution.

The researcher presented ADASYN sampling technique with imbalanced data set and used a decision tree for the classification model. The approach showed the efficiency of ADASYN with five evaluation metrics. ROC was used for based evaluation metric [8].

This paper used extend ADASYN with adaptive synthetic-nominal (ADASYN-N) and ADASYN-KNN (k-nearest neighbor for multiclass imbalance cancer data in Indonesia. ADASYN technique generated instance with nominal data types. ADASYN-kNN generated instance data by voting the attribute of the nearest neighbors. The result of this paper showed that ADASYN-kNN give the highest performance [9].

In [10], ADASYN algorithm and borderline-SMOTE algorithm were used to approach BASMOTE algorithm with stock market data. The result showed that BASMOTE provided higher performance than the traditional oversampling method. This research [11, 12] was an approach to deal with Imbalanced data in customer churn prediction. They used ADASYN to oversampling minority class instances and backpropagation algorithms to classify churn customers. The result represented ADASYN method to increase F1-score performance by threshold correlation 0.01.

SVM was wildly used in regression analysis and classification. This algorithm approach was to create a hyperplane that classifies two or more example sets. It was a good performance algorithm that used with nonlinear classification. This methodology give the highest hyperplane to classify a different class [10]. SVMSMOTE technique used a support vector machine algorithm to synthetic new minority instance observation between minority and majority class's border [13].

This research applied SMOTE-CSVM and OS-SVM as an oversampling technique with 3 human activity open datasets. They used soft margin SVM as a classifier. Their research represented that SMOTE-CSVM and OS-SVM improved prediction accuracy of imbalances human activity datasets [14].

In [15], the researcher applied resampling SVM to various imbalanced datasets. They used 10 imbalanced datasets from UCI machine learning datasets. They applied SVM algorithm to an imbalance dataset with evaluation measurements such as F-measure, AUC. The experiment of this research presented a high performance of SVM algorithms with imbalanced data.

The borderline of minority class examples was used to generate oversampling instances. First, the borderline of minority example sets was found out. Then, the synthetic instance was generated from the borderline data and added to the original training dataset [10, 16].

This paper [17] compared many oversampling algorithms in UCI example data set using SMOTE, borderline-SMOTE, safe-level SMOTE, and ADASYN. They used F-measure value to evaluate the efficiency of experiment with various classifiers such as nearest neighbor, Naïve Bayes, and SVM. The result showed safe level-SMOTE had a higher performance than other algorithms.

Credit data was used to predicted credit risk using K-XGBoost model with border line-SMOTE. The researcher applied SMOTE, safe level SMOTE and borderline-SMOTE with 3 classes unbalanced credit data. Then, they classified credit risk by XGBoost model. The result of the experiment pointed that Borderline-SMOTE had the highest performance algorithm [18].

The researcher explored various oversampling techniques in the imbalanced dataset. SMOTE, ADASYN, SPO, INOS, DataBoost, and Borderline-SMOTE was presented with imbalanced data to improve classification accuracy. The survey showed INOS outperform than other oversampling methods with time series imbalanced data [19].

- Classification method
  a) Multi-layer perceptron (MLP) neural network

     MLP was a classifier technique that arranged in the layer of connecting the compute unit (neuron). Simply the process of MLP, the first layer was an input layer which sent information to calculate in hidden layers with weight. Then, the information in each layer was passed through and sent to the output layer.

     DoS flooding attack was an interesting issue. This research compared an SVM classifier with other methods such as MLP neural network, k-NN, Naïve Bayes, decision tree, random forest and logistic regression for detection system. The result of comparing represented that MLP neural network was the highest accuracy performance classifier [20].

     MLP was used as a classifier without a hidden layer. They used glass dataset and pregnancy dataset applied with mutual information augmentation component. The result showed that the proposed method was improved to generalize performance [21].

  b) AdaBoost

     Adaptive boosting (AdaBoost) was a sequential ensemble method that combined multiple weak learners together, built a strong learner, increasing the final prediction's performance. The main idea was reweighting, which focused on misclassified data points. To be classified in the next round, there will be an increase in the weight of the misclassified data points and decrease the weight of the data points that are correctly classified. Then, take these data points to train and create a new weak learner. These will be sequential, and the weight will be adjusted every round.

     This research proposed BSO-AdaBoost-kNN to deal with imbalance class classification. AUCarea was used as an evaluation metric. They applied the proposed method in oil-bearing of reservoir recognition and provide high precision at 99% [22].

     In [23], the Resampling method in imbalanced dataset was used to improve the performance of classifier. The researcher proposed several ensemble methods with classifiers. The result showed that their proposed method had high performance. AdaBoost was used as classifier method to detect malicious URLs. They showed that AdaBoost algorithm gave more accuracy than other algorithms [24].

  c) Gradient boosting

     Gradient boosting was a model that used decision trees to train several trees together, in which each decision tree learned from previous tree errors, resulting in greater accuracy in prediction. When there was continuous learning of the tree until there was enough depth and the model stopped learning when there was no pattern errors from the previous tree [25].

     Ensemble methods were used in bank direct marketing method [26]. One method was a gradient boosting method. The researcher used ROC curve as an evaluation metric for neural network, logistic regression and gradient boosting classifier.

  d) Random forest

     The principle of the random forest was to create models from multiple decision tree models. Each model received a different data set, which was a subset of all data sets. When making a prediction, each decision tree was given. Make their prediction and calculate prediction by the highest votes chosen by the decision tree or find the mean from the output of each decision tree.

     In [27], k-NN, C4.5, random forest, Naïve Bayes, ANN and AdaBoost were used in customer churn prediction. The researcher used AUC as an evaluation metric. The result represented that random forest was the best classifier in the experiment. This research used random forest with feature selection methods. The result showed that random forest gave the accuracy better than other techniques [28].

- Performance metric

In the unbalanced data issue, the accuracy of classification was not a good indicator for evaluating classification performance. We used various performance measures includes confusion matrix, Precision, Recall, F1-measure and area under curve (AUC).

2. RESEARCH METHOD

In this section, student performance classification with oversampling imbalanced data technique was presented. The procedure consisted of; i) gathering data, ii) Pre-processing data, iii) Oversampling imbalance data, iv) Performance classifier.

2.1. Data gathering

With permission from Faculty of engineering of Rajamagala University of Technology, Thailand, our research picked up 463,956 first year student data records. Then, data were grouped into 6,882 records by

student ID attribute. One record contained 15 attributes including mechanic, material, computer programming, drawing, calculus 1, calculus 2, physic 1, physic 2, English 1, English 2, chemistry, study program, student ID, student name, and class.

2.2. Data cleaning

The example set was managed by detect outlier and remove noise data. Special character such as $ or # and missing value were removed from regrade data. The best grade in the same subject remained in regrade data.

2.3. Data discretization

The example set was separated into 3 classes (High, Medium, Low) by accumulative GPA. High GPA range was 3.00- 4.00 (860 instances) Medium GPA range was 2.00-2.99 (5908 instances) and Low GPA range was 0.00-1.99 (114 instances). The imbalanced data of student performance was showed in Figure 1.

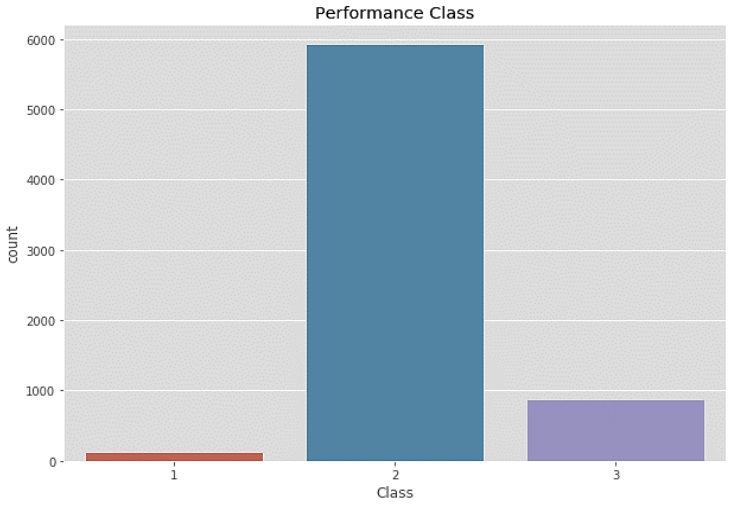

Figure 1. Performance class of original data

2.4. Oversampling imbalanced dataset

We generated synthetic data by SMOTE method. The example instance distribution of oversampling data was shown in Figure 2.

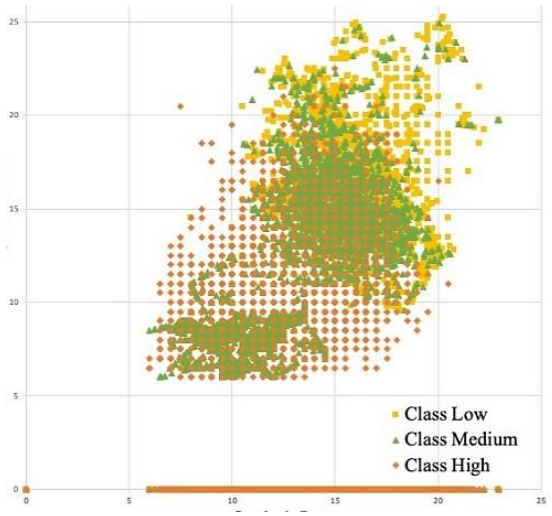

Figure 2. Oversampling instances distribution

2.5. Student performance classifier

We used four classification techniques consisted of neural network, Ada boost, gradient boosting and random forest technique to classify student performance from oversampling imbalanced data set.

3. RESULTS AND DISCUSSION

In oversampling minority class step, we use 4 methods to generate new instances in low and high classes of student performance. Table 1 shows the amount of original data and oversampling data. In performance classification with oversampling data step, we apply 4 classification techniques with 4 oversampling datasets. Figure 3 demonstrates example of confusion matrix of MLP classifier with SMOTE technique.

Table 1. The number of datasets in experiment

Data	Low	Medium	High
Original Data	114	5908	860
SMOTE	5908	5908	5908
ADASYN	5878	5908	5949
Borderline-SMOTE	5908	5908	5908
SVMSMOTE	3754	5908	5908

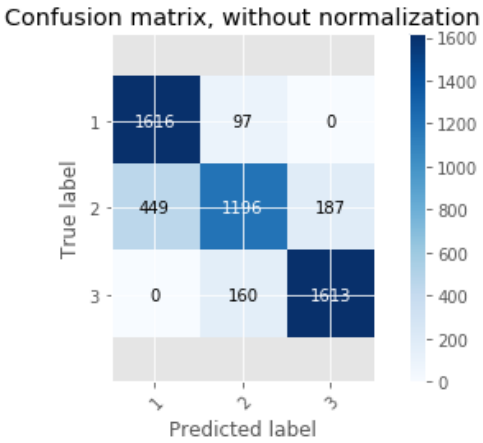

Figure 3. Example Confusion matrix of MLP classifier with SMOTE

We compare the performance of 4 oversampling datasets with 4 classifier methods by 4 evaluation metrics. From Table 2, the result shows the performance improvement of minority class prediction. SVMSMOTE is the best method for minority class (low and high) in recall, f1 and AUC.

Table 2. Performance of oversampling method with neural network

	SMOTE	ADASYN	BorderlineSMOTE	SVMSMOTE
Precision Low	0.78	**0.88**	0.84	0.82
Recall Low	0.94	0.91	**0.96**	**0.96**
F1-measure Low	0.86	0.86	**0.90**	0.88
AUC Low	0.90	0.92	**0.94**	**0.94**
Precision Medium	0.82	0.87	0.89	**0.91**
Recall Medium	0.65	0.60	**0.88**	0.74
F1-measure Medium	0.73	0.70	**0.91**	0.82
AUC Medium	0.79	0.78	0.83	**0.85**
Precision High	**0.90**	0.87	0.88	0.89
Recall High	0.91	0.91	0.94	**0.95**
F1-measure High	0.90	0.89	0.91	**0.92**
AUC High	0.93	0.93	0.94	**0.95**

Gradient boosting provides high-performance precision, recall, f1 and AUC likely as the same value with 4 oversampling methods in Table 3. Table 4, borderline SMOTE and SVMSMOTE are suited in low class. SMOTE is proper with high class. The best method of overall in AdaBoost is SVMSMOTE.

Table 3. Performance of oversampling method with gradient boosting

	SMOTE	ADASYN	BorderlineSMOTE	SVMSMOTE
Precision Low	0.98	0.99	0.99	0.98
Recall Low	0.98	0.98	0.98	0.98
F1-measure Low	0.98	0.98	0.98	0.98
AUC Low	0.99	0.99	0.99	0.99
Precision Medium	0.94	0.94	0.94	0.94
Recall Medium	0.95	0.95	0.95	0.95
F1-measure Medium	0.95	0.95	0.95	0.95
AUC Medium	0.96	0.96	0.96	0.96
Precision High	0.96	0.96	0.96	0.96
Recall High	0.96	0.96	0.96	0.96
F1-measure High	0.96	0.96	0.96	0.96
AUC High	0.97	0.97	0.97	0.97

Table 4. Performance of oversampling method with adaboost

	SMOTE	ADASYN	BorderlineSMOTE	SVMSMOTE
Precision Low	0.76	0.77	**0.80**	0.75
Recall Low	0.96	0.97	0.95	**0.98**
F1-measure Low	0.85	0.86	**0.87**	0.85
AUC Low	0.91	0.92	0.92	**0.93**
Precision Medium	0.90	0.89	0.88	**0.93**
Recall Medium	0.51	0.53	**0.55**	**0.55**
F1-measure Medium	0.65	0.66	0.68	**0.69**
AUC Medium	0.74	0.75	**0.76**	**0.76**
Precision High	**0.82**	0.81	0.81	0.81
Recall High	**0.98**	0.96	0.97	0.97
F1-measure High	**0.89**	0.88	0.88	0.88
AUC High	**0.94**	0.93	0.93	0.92

The result of Table 5 presents the performance of oversampling method with random forest. For low minority class, borderline-SMOTE is the best performance, and high minority class SVMSMOTE is the best performance. The result of the experiment represents performance of the evaluation metric. The performance of minority classes classification from four classifiers shows that borderline-SMOTE is the highest performance. In Table 3. Because of the classifier method, oversampling is not significant in minority class classification.

Table 5. Performance of oversampling method with random forest

	SMOTE	ADASYN	BorderlineSMOTE	SVMSMOTE
Precision Low	0.77	0.77	**0.82**	0.81
Recall Low	0.95	0.95	**0.99**	0.95
F1-measure Low	0.85	0.85	**0.90**	0.88
AUC Low	0.91	0.89	**0.94**	0.91
Precision Medium	0.86	0.87	**0.92**	0.90
Recall Medium	0.58	0.50	0.61	**0.67**
F1-measure Medium	0.69	0.63	0.73	**0.77**
AUC Medium	0.61	0.55	0.66	**0.69**
Precision High	**0.86**	0.80	0.83	0.84
Recall High	0.94	**0.96**	**0.96**	**0.96**
F1-measure High	**0.90**	0.87	0.89	**0.90**
AUC High	0.92	0.91	0.92	**0.93**

4. CONCLUSION

Education mining in several problems such as student drop out, grade prediction and student performance has an imbalance class data problem. The researcher is developing an algorithm to solve the problem at algorithm level and data level to improve classification performance. This research emphasizes in oversampling method to improve imbalance class problem in education mining and overfitting problem also.

From the experiment result, the best performance oversampling method is borderline SMOTE. Students and instructor can use the classification model with oversampling to improve student performance. The future task, we will use more significant feature and more datasets. Moreover, we will extend the algorithm of oversampling or under-sampling methods to improve student performance.

ACKNOWLEDGEMENTS

The authors thanks to Office of Academic Promotion and Registration, Rajamangala University of Technology Thanyaburi for their student data support.

REFERENCES

[1] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Applied Soft Computing*, vol. 83, 2019, Art. no. 105662.
[2] N. V. Chawla, "Data Mining For Imbalanced Datasets: An Overview," *Data Mining And Knowledge Discovery Handbook*, pp. 853-867, 2005.
[3] T.-W. Lim, K.-C. Khor, and K.-H. Ng, "Dimensionality Reduction for Predicting Student Performance in Unbalanced Data Sets," *International Journal of Advances in Soft Computing and its Applications (IJASCA)*, vol. 11, no. 2, pp. 76-86, 2019.
[4] Y. Sanguanmak and A. Hanskunatai, "DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification," *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Khon Kaen, Thailand, 2016, pp. 1-5.
[5] Fernández, A., Garcia, S., Herrera, F., and Chawla, N. V., "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research (JAIR*), vol. 61, pp. 863-905, 2018.
[6] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4331-4339, 2020.
[7] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F., "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484, 2012.
[8] He, H., Bai, Y., Garcia, E. A., and Li, S., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, China, 2008, pp. 1322-1328.
[9] Kurniawati, Y. E., Permanasari, A. E., and Fauziati, S, "Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data," *2018 4th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, 2018, pp. 1-6.
[10] Du, M., Zhang, Z., and Zhang, Y., "Modified Machine Learning Model and Stock Classification Research Based on Unbalanced Data," *2018 7th International Conference on Digital Home (ICDH)*, Guilin, China, 2018, pp. 200-207.
[11] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J. *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access*, vol. 4, pp. 7940-7957, 2016.
[12] A. Aditsania and A. L. Saonard, "Handling Imbalanced Data in Churn Prediction using ADASYN and Backpropagation Algorithm," *2017 3rd International Conference on Science in Information Technology (ICSITech)*, Bandung, Indonesia, 2017, pp. 533-536.
[13] Q. Cao and S. Wang, "Applying Over-sampling Technique Based on Data Density and Cost-sensitive SVM to Imbalanced Learning," *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, Shenzhen, China, 2011, pp. 543-548.
[14] Yala, N., Fergani, B., and Clavier, L., "Soft margin SVM modeling for handling imbalanced human activity datasets in multiple homes," *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, Marrakech, Morocco, 2014, pp. 421-426.
[15] Chen, L., Cai, Z., Chen, L., and Gu, Q., "A Novel Differential Evolution-Clustering Hybrid Resampling Algorithm on Imbalanced Datasets," *2010 Third International Conference on Knowledge Discovery and Data Mining*, Phuket, Thailand, 2010, pp. 81-85.
[16] Han, H., Wang, W. Y., and Mao, B. H., "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International Conference on Intelligent Computing- ICIC 2005*, vol. 3644, pp. 878-887, 2005.
[17] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, 2017, pp. 79-85.
[18] W. Qiu, "Credit Risk Prediction in an Imbalanced Social Lending Environment Based on XGBoost," *2019 5th International Conference on Big Data and Information* Analytics (BigDIA), Kunming, China, 2019, pp. 150-156.
[19] Dhurjad, M. R. K., and Banait, M. S., "A survey on oversampling techniques for imbalanced learning," *International Journal of Application or Innovation in Engineering & Managemen*t, vol 3, no. 10, pp. 279-284. 2014.
[20] M. Latah and L. Toker, "A novel intelligent approach for detecting DoS flooding attacks in software-defined networks*," International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, pp. 11-20, 2018.
[21] R. Kamimura, "Internal and collective interpretation for improving human interpretability of multi-layered neural networks," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 3, pp. 179-192, 2019.
[22] Haixiang, G., Yijing, L., Yanan, L., Xiao, L., and Jinling, L., "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Engineering Applications of Artificial Intelligence*, vol. 49, pp. 176-193, 2016.

[23] Y. Zhang and D. Wang, "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets," *Abstract and Applied Analysis*, vol. 2013, pp. 1-6, 2013.
[24] Khan, F., Ahamed, J., Kadry, S., and Ramasamy, L. K., "Detecting malicious URLs using binary classification through adaboost algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 997-1005, 2020.
[25] Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M. *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing*, vol. 74, pp. 634-642, 2019.
[26] Y. Pan and Z. Tang, "Ensemble methods in bank direct marketing," *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, Beijing, China, 2014, pp. 1-5.
[27] G. Esteves and J. Mendes-Moreira, "Churn perdiction in the telecom business," *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, Porto, Portugal, 2016, pp. 254-259.
[28] R. S. and S. Kumar J., "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 353-359, 2020.

BIOGRAPHIES OF AUTHORS

Nachirat Rachburee is a lecturer at Deprtment of Computer Engineering, Faculty of Engineering, Rajamagala University of Technology, Pathum Thani, Thailand. His research interests include Data Mining, Big data analytics, Deep Learning, Neural Networks and Predictive analytics.

Wattana Punlumjeak is a lecturer at Deprtment of Computer Engineering, Faculty of Engineering, Rajamagala University of Technology, Pathum Thani, Thailand. His research interests include Data Mining, Big data analytics, Deep Learning, Neural Networks and Predictive analytics.