Improving The Effectiveness of Classification Using The Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction

Improving The Effectiveness of Classification Using The Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction

I Kurniawan[1]*, Abdussomad[1], M F Akbar[1], D F Saepudin[2], M S Azis[3], and M Tabrani[4]

[1]Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
[2]Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
[3]Sistem Informasi, Sekolah Tinggi Manajemen Informasi dan Komputer Nusa Mandiri, Indonesia
[4]Teknik Informatika, Sekolah Tinggi Manajemen Informasi dan Komputer Nusa Mandiri, Indonesia

E-mail: ilham.imk@bsi.ac.id

Abstract. Online shopping is a form of trading using electronic devices that allows consumers to buy goods or services from sellers via the internet. Other names for these activities are: e-web-shop, e-shop, e-shop, internet shop, web-shop, web-store, online shop, and virtual shop. An online store generates purchases of products or services at retailers or shopping centers, which are referred to as business-to-consumer (B2C) online shopping. n another process where a business buys from another business, it is called business-to-business (B2B). Nowadays online shopping has become more sophisticated with trading via mobile phones (m-commerce). Cellular phones have been optimized with an application to buy from online sites. In this study, we proposed a data level approach and feature selection techniques as a solution for the classification of imbalanced data. The imbalance class classification is one of the classic problems in the field of artificial intelligence, especially for classification in machine learning. Imbalanced data have been proven to reduce the performance of machine learning algorithms, where imbalance data means that the total data from each class is significantly different. The proposed method is evaluated using a dataset from the UCI repository and area under the curve (AUC) as the main evaluation. The results have shown that the proposed method produces good performance. (AUC¿ 0.8). Overall the second experiment outperformed and was better than the first and third experiments because the main evaluation in the unbalanced class classification is AUC. Therefore, it can be concluded that the proposed method produces optimal performance both for large scale data sets. Overall the second experiment outperformed and better than the first and third experiments, because the main evaluation in the unbalanced class classification was AUC.

1. Introduction

Technology is increasingly changing shopping trends. The consumer purchasing experience has been transformed not only by the Web, but also by new advances in internet-connected smart devices [1][2]. Recent trends that have reshaped international retail, namely the ability of consumers to shop for products from national and international markets at the click of a button, without the need for physical travel. More than 82% of international buyers shop at least once from foreign websites annually [3]. From 2015-2017, the percentage of consumers who still prefer shopping at stores has fallen from 85% to 70% [4]. The imbalance class classification is one of the classic problems in the field of artificial intelligence, especially for classification in machine learning. Imbalance class has been proven to reduce the performance of machine learning algorithms, where imbalance class means the total data from each class is significantly different [5].

Various classification problems to support a prediction, there are a number of different classification errors can have enormous losses, as suggested by [6] and it may be important to control, to some degree, between those mistakes. for example, in the Neyman-Pearson (NP) framework [7]. Tolerable false positive rate (FPR) is set at the specified value, and the aim is to minimize the false negative rate (FNR) provided that the FPR is not greater. it occurs naturally in many situations, especially when the class of interest is the minority.

Various methods have been proposed to solve this problem. These methods can be divided into four types: algorithmic level methods, data level methods, and ensemble classifications [8]. Specifically, data level methods, which focus on preprocessing imbalanced datasets before constructing classification, are widely considered in the literature. Because the tasks of data preprocessing and classifier training can be done independently. In addition, according to [8], which conducts comparative studies of various approaches, the combination of data preprocessing methods with ensemble classifier is better than other methods. The preprocessing data method is based on resampling unbalanced training data sets before the model training stage. To create data balance, the original imbalance dataset can be sampled again by overampling minority classes [9] or undersampling the majority class [10]. Some approaches by combining preprocessing oversampling and undersampling data with ensemble classifier through boosting techniques [11] or bagging [12], for example SMOTEBoost, RUS Boost [13], OverBagging, and UnderBagging [14].

In this study, we propose through data level approach techniques and attribute selection techniques, we show that the type of oversampling strategy, can reduce the risk of removing useful data from the majority class, allowing classifiers built to outperform classifiers developed using SMOTE (synthetic minority over-sampling) strategies and types of random undersampling strategies to balance positive classes and its negative class. In this research we proposes a method for predicting online shoppers purchasing intention prediction by using the integration of particle swarm optimization (PSO) feature selection techniques with data level approaches including Random Under-Sampling (RUS), and SMOTE (Synthetic Minority Over-sampling Technique).

2. Methods

The proposed model includes the application of feature selection using particle swarm optimization (PSO), the data level approach algorithm (SMOTE), the AdaBoost algorithm level approach with several classification algorithms. The final result will be a comparison test and analysis of prediction models that have the highest or best accuracy in predicting online shoppers purchasing intention prediction. The following is a framework for the proposed research model. According to the proposed model consists of two approaches, namely the data level approach and the algorithm level approach. The two approaches will be used interchangeably, and a combination of the two to create various predictive models of online shoppers purchasing intention prediction. The data level approach is intended to balance classes in the dataset which are generally imbalanced. In the data level approach two methods are proposed, namely SMOTE. SMOTE balances minority classes by synthesizing minority class data. The algorithm level approach is intended to improve the performance of classifiers using the ensemble technique using the AdaBoost algorithm. The application of the dataset in the model formed was validated

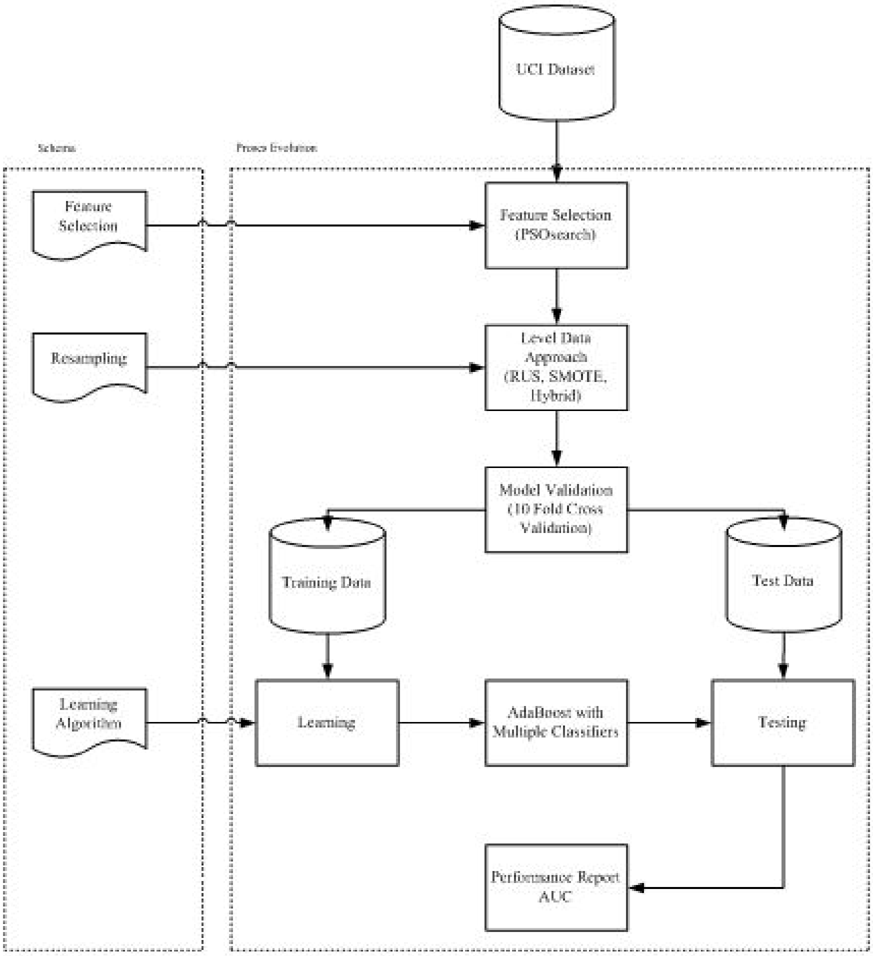
UCI Dataset
Schema
Proses Evaluation
Feature Selection
Feature Selection (PSOsearch)
Resampling
Level Data Approach (RUS, SMOTE, Hybrid)
Model Validation (10 Fold Cross Validation)
Training Data
Test Data
Learning Algorithm
Learning
AdaBoost with Multiple Classifiers
Testing
Performance Report AUC

Figure 1. The Proposed Methods

using 10 fold cross validation. Validation using 10-fold cross validation is done by dividing the dataset into 10 parts, one part as test data, while the other part as training data.

The validation process is repeated, starting from the first part as test data to the tenth section, so that all data in the dataset is tested. The purpose of validation is to produce a performance prediction model for online shoppers purchasing intention prediction. The performance model of online shoppers purchasing intention prediction is measured based on accuracy, sensitivity, F-Measure, and AUC. AUC value can be used as a measure to see the model formed. Area Under ROC Curve (AUC) is used to provide a single numerical metric to be able to compare the performance of the model, the AUC value ranges from 0 to 1 and the model whose prediction is better is close to 1.

In this study, the proposed method was evaluated using classifier effectiveness based on a confusion matrix with the main evaluation being the AUC as used by [15], [16], [17], [18], [19] AUC has the potential to significantly increase convergence across empirical experiments in the prediction of software defects and the use of AUC to improve cross-study comparisons [20]. Evaluation of the proposed method is: f-measure by combining the values of precision, recall and sensitivity (SN) as used by [15], [17], [18], specipies (SP) and precision (PR) as used by [17], [18].

This evaluation is based on a confusion matrix containing true positive (TP), true negative

(TN), false positive (FP) and false negative (FN) values.

3. Result and Discusion

3.1. Single Classifier

The first test was performed using a single classifier model c4.5, multilayer perceptron, support vector machine and random forest on the UCI dataset. The test results are calculated using the Confusion Matrix to look for accuracy, sensitivity / recall / TPrate, specificity / TNrate, FPrate, FNrate, Precision / PPV, F-Measure, and AUC. The calculation results obtained as follows:

Table 1. Results of the Performance of a Single Classification Algorithm on the UCI Repository Dataset.

Algorithms	Recall	Specificity	FPrate	FNrate	Precision	Fmeasure	AUC	Accuracy
C45	0,951	0,589	0,410	0,048	0,926	0,938	0,784	0,895
MLP	0,955	0,555	0,800	0,081	0,921	0,937	0,894	0,893
RF	0,960	0,586	0,413	0,039	0,926	0,942	0,928	0,902
SVM	0,979	0,318	0,643	0,019	0,892	0,933	0,668	0,883

The above table shows that the highest accuracy of the performance of the random forest classification algorithm with an accuracy of up to 90% and the highest AUC value is 0.928. The analysis shows that the average value of the performance of several single classifications includes an accuracy of 89% for the C4.5 algorithm, 89% for the MLP accuracy value, 90% for the RF algorithm accuracy value, and 88% for the SVM algorithm accuracy value. The AUC value for each single classification algorithm is, 0.784 for the C4.5 algorithm, 0.674 for the MLP algorithm, 0.928 for the RF algorithm and 0.668 for the SVM algorithm. These results will be a reference to see how improved the performance of the proposed model in the next trial period.

3.2. Testing Model Level Data Approach, AdaBoost and Classification Algorithms

The first model proposed is to use the SMOTE data level approach technique. The results of the SMOTE (Synthetic Minority Over-sampling Technique) technique for class balance and subsequently performed the AdaBoost ensemble technique with several classification algorithms and validation techniques using 10 fold cross validation. The results obtained for the proposed model are presented in the following table:

Table 2. SMOTE + AdaBoost Performance Results + Classification Algorithm on the UCI Repository Dataset.

Algorithms	Recall	Specificity	FPrate	FNrate	Precision	Fmeasure	AUC	Accuracy
C45	0,929	0,796	0,203	0,070	0,929	0,926	0,945	0,893
MLP	0,930	0,715	0,284	0,069	0,892	0,910	0,900	0,872
RF	0,940	0,814	0,185	0,059	0,932	0,935	0,960	0,907
SVM	0,959	0,531	0,468	0,046	0,847	0,899	0,866	0,840

The table above shows the highest accuracy of the classification of the random forest algorithm with an accuracy reaching 90.7% and the highest AUC value of 0.960. The analysis shows the value of the performance of SMOTE + AdaBoost + classification algorithm includes an accuracy of 89.3% C4.5 algorithm, 87.2% MLP algorithm, 90.7% RF algorithm and 84.0% SVM algorithm. AUC value of 0.945 C4.5 algorithm, 0.900 MLP algorithm, 0.960 random forest algorithm and 0.866 SVM algorithm. These results prove that the second experiment outperformed the first experiment. The second experiment yielded better results in all evaluations than the first experiment. Increased accuracy in all classification algorithms tested. In the case of AUC, the proposed model is classified as an excellent classifier because it is¿ 0.8.

3.3. Testing the PSO, SMOTE, AdaBoost and Classification Algorithm

The second model proposed is to use the PSO (Particle Swarm Optimization) feature selection technique. The results of the feature selection using PSO are then carried out an approach at the data level with the SMOTE resampling technique for class balance and then the Adaboost ensemble technique is performed with a classification algorithm with validation using 10 fold cross validation. The results obtained for the proposed model are presented in the following table.

Table 3. PSO+ SMOTE + AdaBoost Performance Results + Classification Algorithm on the UCI Repository Dataset.

Algorithms	Recall	Specificity	FPrate	FNrate	Precision	Fmeasure	AUC	Accuracy
C45	0,901	0,774	0,225	0,098	0,916	0,908	0,892	0,867
MLP	0,912	0,722	0,277	0,087	0,899	0,905	0,831	0,861
RF	0,930	0,676	0,323	0,069	0,886	0,906	0,805	0,862
SVM	0,961	0,498	0,501	0,038	0,839	0,895	0,726	0,837

The above table shows that the highest accuracy of the classification algorithm performance is the c4.5 algorithm with an accuracy of 86% and the highest AUC value of 0.892. These results prove that the third experiment did not outperform the first and second experiments. The third experiment produced better results on the AUC value compared to the first experiment, for the classification algorithm C4.5 and SVM, the first experiment was worth 0.784 and 0.668, in the third experiment the AUC value increased to 0.892 and 0.726.

3.4. Comparison of Research Results

For a more detailed comparison between the first, second and third experiments, we present a comparison in Figure 4.4. As shown in Figure 4.4, the second experiment (SMOTE + AdaBoost + Classification Algorithm) is better than all experiments on all evaluation models. Of all experiments, the second experiment outperformed all experiments from the first and third experiments. Meanwhile, the second experiment can be said to be a successful experiment. Overall the second experiment outperformed and was better than the first because the main evaluation in the unbalanced class classification was AUC as stated by [16] [20].

4. Conclusion

Comparison of data level approach techniques and PSO feature selection techniques is proposed to see the comparison of the performance results of several classification algorithms, including

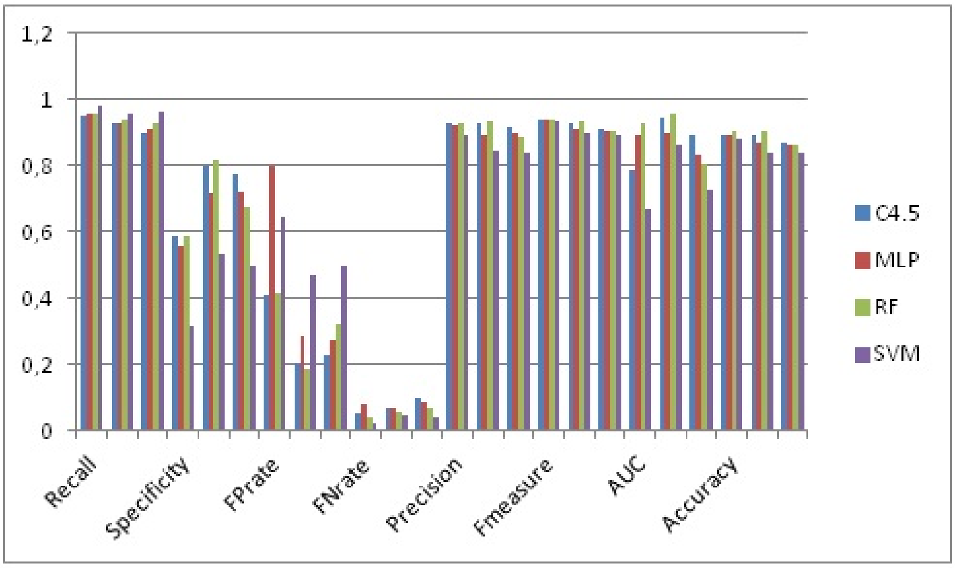

Figure 2. Comparison of Research Results

C4.5, Multilayer Perceptron, Random Forest and Support Vector Machine. Based on the research results, the following conclusions can be drawn. The four classification algorithms proposed by the Random Forest classification algorithm outperforming all the proposed classification algorithms, the random forest algorithm also outperformed the performance results of the three models proposed in this study.The performance of the proposed SMOTE + AdaBoost + Algorithm Algorithm can be seen from the average AUC value higher than other proposed models and in previous studies [21] with an accuracy value of 89.51%. The conclusion is that the proposed model SMOTE + AdaBoost + Classification Algorithm and PSO + SMOTE + AdaBoost + Classification Algorithm can improve the performance of the overall classification model. This result is obtained from the AUC value which can reach 90%. However, when compared from the numerical value of the AUC, the SMOTE + AdaBoost + Classification Algorithm model is said to be better than the two models. The classification criteria of the two models are based on the AUC table, so it can be concluded that the proposed model is in the fair classification criteria with an average AUC value of more than 0.7. in future research it can be done by using other classification algorithms such as Naive Bayes or other classification algorithms, or by changing the selection of attributes such as Genetic Algorithm to increase the accuracy and the AUC.

References

[1] S. J. Kim, R. J. H. Wang, and E. C. Malthouse, 2015, “The Effects of Adopting and Using a Brand’s Mobile Application on Customers’ Subsequent Purchase Behavior,” J. Interact. Mark., vol. 31, no. 2015, p. 28–41.

[2] J. Martins, C. Costa, T. Oliveira, R. Gonçalves, and F. Branco, 2019, “How smartphone advertising influences consumers’ purchase intention,” J. Bus. Res., vol. 94, no. December 2017, p. 378–387.

[3] B. Ramkumar and B. Ellie Jin, 2019, “Examining pre-purchase intention and post-purchase consequences of international online outshopping (IOO): The moderating effect of E-tailer’s country image,” J. Retail. Consum. Serv., vol. 49, no. March, p. 186–197.

[4] TimeTrade, “The State of Retail Report 2017,” 30 March, 2017.

[5] Fanny and T. W. Cenggoro, 2018, “Deep Learning for Imbalance Data Classification using Class Expert Generative Adversarial Network”.

[6] A. Wijaya dan R. S. Wahono, 2017, "Tackling Imbalanced Class In Software Defect Prediction Using Two-Step Cluster Based Random Undersampling And Stacking Technique," J. Teknol., no. November, p. 45–50.

[7] C. Pradhan and A. Gupta, 2017, "Ship detection using Neyman-Pearson criterion in marine environment," Ocean Eng., vol. 143, no. March 2016, p. 106–112.

[8] X. Tao et al., 2019, Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification, vol. 487. Elsevier Inc.

[9] T. Zhu, Y. Lin, Y. Liu, W. Zhang, and J. Zhang, 2018, "Minority oversampling for imbalanced ordinal regression," Knowledge-Based Syst.

[10] C. Tsai, W. Lin, Y. Hu, and G. Yao, 2018, "Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection Chih-Fong," Inf. Sci. (Ny).

[11] D. Z. Li, W. Wang, dan F. Ismail, 2015, "Neurocomputing A selective boosting technique for pattern classi fi cation," vol. 156, hal. 186–192.

[12] W. W. Y. Ng, X. Zhou, X. Tian, X. Wang, and D. S. Yeung, 2017, "Bagging-boosting-based semi-supervised multi-hashing with query-adaptive re-ranking," Neurocomputing, vol. 0, hal. p.

[13] A. R. Hassan and A. Haque, 2016, "An expert system for automated identification of obstructive sleep apnea from single-lead ECG using random under sampling boosting," Neurocomputing.

[14] B. S. Raghuwanshi dan S. Shukla, 2019, "Neurocomputing Class imbalance learning using UnderBagging based kernelized extreme learning machine," Neurocomputing, vol. 329, hal. 172–187.

[15] I. H. Laradji, M. Alshayeb, and L. Ghouti, 2015, "Software defect prediction using ensemble learning on selected features," Inf. Softw. Technol., vol. 58, hal. 388–402.

[16] Z. A. Rana, M. A. Mian, and S. Shamail, 2015, "Improving Recall of software defect prediction models using association mining," Knowledge-Based Syst., vol. 90, p. 1–13.

[17] G. Czibula, Z. Marian, and I. G. Czibula, 2014, "Software defect prediction using relational association rule mining," Inf. Sci. (Ny)., vol. 264, p. 260–278.

[18] R. S. Wahono and N. S. Herman, 2014, "Genetic feature selection for software defect prediction," Adv. Sci. Lett., vol. 20, no. 1, p. 239–244.

[19] O. F. Arar and K. Ayan, 2015, "Software defect prediction using cost-sensitive neural network," Appl. Soft Comput. J., vol. 33, p. 263–277.

[20] F. Cheng, X. Zhang, C. Zhang, J. Qiu, and L. Zhang, 2018, "An Adaptive Mini-Batch Stochastic Gradient Method for AUC Maximization," Neurocomputing, p. 2–25.

[21] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, 2019, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Comput. Appl., vol. 31, no. 10, p. 6893–6908.