Jurnal Mantik, 6 (2) (2022) 2332-2336

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Handling Unbalanced Data With Smote Adaboost

Laila Qadrini

[1]Statistics Department, Mathematics and Natural Science, Universitas Sulawesi Barat, Jalan Prof. Dr. Baharuddin Lopa, S.H, Baurung, Kec. Banggae Timur., Kabupaten Majene, Sulawesi Barat 91412

E-mail: laila.qadrini@unsulbar.ac.id

ARTICLE INFO

Article history:
Received: Jul 25, 2022
Revised: Jul 30, 2022
Accepted: Aug 10, 2022

Keywords:
Unbalanced,
SMOTE,
Adaboost

ABSTRACT

A very important part of data mining is classification techniques, namely how to study a set of data so that rules are generated that can classify or recognize new data that has never been studied. Classification can be defined as a process for declaring a data object as one of the predefined categories (classes). There are many classification techniques global classification techniques have been proposed by experts, which can be grouped into two categories: global classification techniques (taking into account all training data) and local classification techniques (taking into account only part of the training data). each technique has advantages and disadvantages. There are problems that are often encountered in classification, namely the problem of data imbalance. Unbalanced data is a condition where the distribution of data classes is not balanced, the number of one data class is less or more than the number of other data classes. The group of data classes that are less known as the minority group (minority), the other data class group is called the majority group (majority). Classification of data with unbalanced classes is a major problem in the field of data mining, in this research This study resulted in the application of the adaboost smote, the AUC value for the Adaboost SMOTE model was 0.784 while the AUC value for the Adaboost model was 0.664.

Copyright © 2022 Jurnal Mantik.
All rights reserved.

1. Introduction

A significant part of data mining is classification techniques, namely how to study a set of data so that rules are generated to classify or recognize new data that has never been studied. Classification can be defined as a process for declaring a data object as one of the predefined categories (classes) [1].

Classification is widely used in various applications, including fraud detection, customer management, medical diagnosis, sales prediction, etc. How to build a classification model based on the knowledge of an expert but considering an extensive data set [2]. Classification models are more often built using learning techniques in the field of machine learning. The learning process automatically on a data set can produce a classification model (target function) that maps the x data object (input) to one of the previously defined y classes [3]. So the learning process requires feedback in the form of a training set labelled (has a class attribute) and produces an output in the format of a classification model [4].

There are many classification techniques, and global classification techniques have been proposed by experts, which can be grouped into two categories: international classification techniques (taking into account all training data) and local classification techniques (taking into account only part of training data) [5]. Each method has advantages and disadvantages. Some problems are often encountered in classification, namely the data imbalance issue. Unbalanced data is a condition where the distribution of data classes is not balanced. The number of data classes (instances) is one less or more than the number of other data classes [6].

The group of data classes that are less known as the minority group (minority); the other data class group is called the majority group (majority). Classification of data with unbalanced classes is a significant problem in the field of data mining, for example, in medical problems [3], text classification problems [7],

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

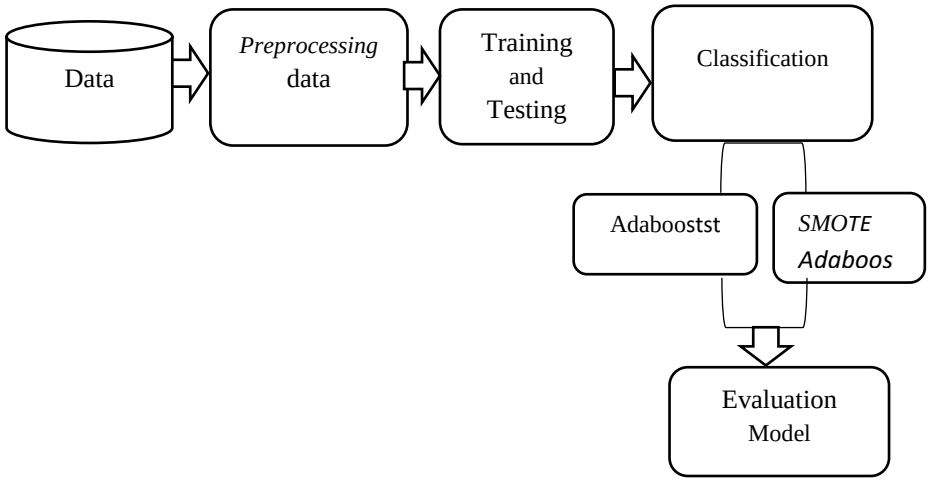
and social media [8]. unbalanced data, Almost all classification algorithms will produce much higher accuracy for the majority class than for the minority class [9]. This difference is an indicator of poor classification performance. In some cases, the minority class is even more critical to identify than the majority class. Applying the classification algorithm without paying attention to class balance results in good predictions for the majority class and the minority class being ignored. If the classification algorithm is implemented directly on imbalanced datasets, performance will decrease [10].

In this study, researchers will handle class imbalances in the case of wine quality data. The data is obtained from the UCI Machine Learning Repository on the https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ page. The resampling technique used is data classification with Adaboost and SMOTE Adaboost. The oversampling algorithm used is Synthetic Minority Over-sampling Technique (SMOTE); this algorithm was chosen from several resampling algorithms because SMOTE produces good accuracy and is effective in dealing with unbalanced classes because it reduces overfitting [11].

The effectiveness of the sampling algorithm is proved by visual experiments. Finally, [18] propose a cost-sensitive algorithm based on Bagging, and combine it with the dynamic sampling algorithm to propose a multi-fusion imbalanced ensemble learning algorithm [19] proposes an ensemble algorithm named of BPSO-Adaboost-KNN to cope with multi-class imbalanced data classification. The main idea of this algorithm is to integrate feature selection and boosting into ensemble. What's more, we utilize a novel evaluation metric called AUC area which is especially for multi-class classification. The goal is to balance the classes in the dataset to improve the performance of the classification algorithm. The results of this study will be compared with the results of the classification without resampling. The evaluation tests are accuracy, Confusion Matrix (CM) and AUC [9].

2. Method

This study applies two classification methods: the Adaboost method and the SMOTE Adaboost method. Quality values that were initially in the range 0-10 were grouped first into two classes; namely, those whose values were less than or equal to 6 were included in class 0 (more petite), and the other was included in class 1 (suitable). Used iteration 5 times. This means that five stumps/small trees are created sequentially and arranged in such a way using the mfinal=3 option. The tree option is maxdepth=1 to ask the program to make a stump as a weak learner of the boosting algorithm. The algorithm is executed, and the obtained model is stored in an object with the name model.AdaBoost. Transform data from numeric to nominal and perform normalization to determine the most appropriate form of data. Then divide the training data and testing data with a ratio of 70%:30%. Data processing is carried out by resampling the SMOTE classifier in the validation process so that higher accuracy is obtained from the processing of each method. The following are the stages of the research carried out:



Picture 1. The Research's Process

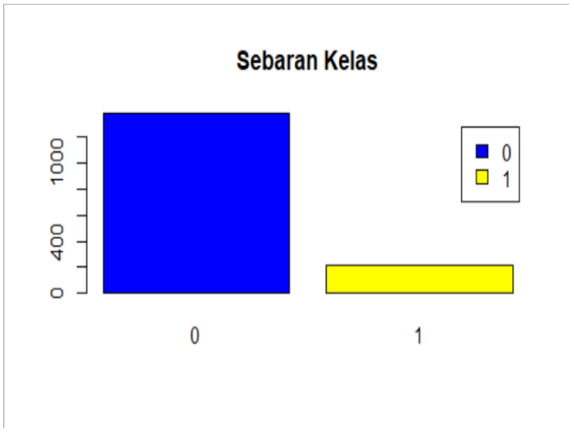
Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

3. Result and Discussion

3.1 Data Overview

The data has 12 attributes, and one of the attributes is a class label. The excellent quality class as the minority class has 217 data, and the poor quality class as the majority class has 1382 data, with the proportion of minority and majority class being 0.86:0.14.



Picture 2. Class Distribution Yes and No

3.2 Preprocessing data

They were preprocessing aims to uniform the data range between 0 to 1. So that the attribute value is between 0 and 1, each attribute value is divided by the maximum attribute value of all data. Quality values initially in the range 0-10 were first grouped into two classes. Namely, those whose values were less than or equal to 6 were included in class 0 (more minor), and the other was included in class 1 (suitable). Then change the numeric type to categorical, i.e. 0 for no and 1 for Yes.

Table 1. Total Training Data and Testing Data.

Type of data	Amount of data Total		Total
	Class 1	Class 0	
Training	152	968	1120
Testing	65	414	479

3.3 Feature Selection

Feature selection is one of the essential activities in building a model, and it is often repeated. Not all features in this study can be helpful in modelling, even if too many irrelevant features in the data can reduce the model's accuracy. We have to choose the features that give the best results; in this study, the feature selection has been done and reduced to 12 elements.

3.4 Determination of training and data testing

The data must first be divided into training and testing data to make predictions. In this study, the overall data is divided into 70%: 30% for 70% of training data and 30% of testing data.

3.5 Adaboost Prediction Results and SMOTE Adaboost

There are several approaches to overcome the problem of imbalanced data, namely the Synthetic Minority Oversampling Technique (SMOTE), selecting data in the minority class randomly and then looking for the nearest k minority class data from the data [10]. From the k minority class data, one data will be chosen randomly, which is then connected to the selected initial data to form a line segment in the feature space [11]. Synthesis data is generated using a convex combination between the two randomly selected data. This approach is considered quite effective because the synthesis data formed is relatively close in the feature space with existing data from the minority class [12]. This procedure can be used to generate as much

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

synthetic data for the minority class as needed. The accuracy of the model with the Adaboost classifier above is 86.85% which means it is very good [13]. In general, in the field of data science, models with accuracy above 70% can be classified as a model that performs quite well [14]. Precision gives an indication of how well the model can predict the positive. The precision value after SMOTE is also 80%. We prefer the occurrence of True Positive and really do not want the occurrence of False Positive. So in this study, we prefer if the quality of the wine is good but is predicted to be poor (so it can still be distributed) rather than the quality of the wine, which is actually not good but is predicted to be good. Sensitivity or recall measures the number of data that were successfully predicted to be positive compared to all data that were in fact positive [15]. The model shows the sensitivity for SMOTE is 76.8%. The AUC value is used as a measure of whether a model is good or bad [16]. An AUC close to 1 indicates a near-perfect model, while an AUC close to 0.5 is a poor model [17]. The AUC value for the SMOTE Adaboost model is 0.784, while the AUC value for the Adaboost model is 0.664.

Table 2.
Prediction Results of Adaboost and SMOTE Adaboost

Metrics	Training data with Adaboost	Training data with SMOTE Adaboost
Accuracy	0,8685	0, 772
Presition	0,385	0,8
Recall	0,944	0,768
AUC	0,664	0,784

4. Conclusion

The prediction results with the Adaboost classifier provide a high accuracy of 86.85%, but actually the low precision value is 0.385, meaning that the model is not good because it is not able to predict the "Yes" class properly. Model evaluation on the adaboost SMOTE classifier is better than adaboost because it produces a higher AUC value of 0.784

References

[1] Zaki, M.J., & Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. USA, NJ : Cambridge University Press.
[2] Ali, S. M. Shamsuddin, & A. L. Ralescu. (2009). Classification with class imbalance problem: a review. Int J Adv. Soft Compu Appl, 7(3).
[3] Kothan. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. Bioinformation, 11(1):6–10.
[4] Wu, Y. Ye, H. Zhang, M. K. Ng, & S.-S. Ho. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. Knowl-Based Syst. 67:105–116.
[5] Li & S. Liu. (2014). A comparative study of the class imbalance problem in Twitter spam Detection. Concurr. Comput. Pract. Exp.,pp. n/a-n/a
[6] Alsawalqah, H., Faris, H., Aljarah, I., Alnemer, L., & Alhindawi, N. (2017, April). Hybrid SMOTE-ensemble approach for software defect prediction. In Computer science on-line conference (pp. 355-366). Springer, Cham.
[7] Wu, Y., Ding, Y., & Feng, J. (2020). SMOTE-Boost-based sparse Bayesian model for flood prediction. EURASIP Journal on Wireless Communications and Networking, 2020(1), 1-12.
[8] Choi, M. J. (2010). A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Iowa: Graduate Theses. Iowa State University.
[9] Yulianto, A., Sukarno, P., & Suwastika, N. A. (2019, March). Improving adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012018). IOP Publishing..
[10] Huang, Q., & Zhang, X. (2016, December). An improved ensemble learning method with SMOTE for protein interaction hot spots prediction. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1584-1589). IEEE.
[11] Rajesh, K. N., & Dhuli, R. (2018). Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier. Biomedical Signal Processing and Control, 41, 242-254.
[12] Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. Information Sciences, 512, 1214-1233.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

[13] Amirruddin, A. D., Muharam, F. M., Ismail, M. H., Tan, N. P., & Ismail, M. F. (2022). Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles. Computers and Electronics in Agriculture, 193, 106646.
[14] Kameswara Rao, B., Prasan, U. D., Rao, J., Pedada, R., & Kumar, P. S. (2022). Identification of Heart Failure in Early Stages Using SMOTE-Integrated AdaBoost Framework. In Computational Intelligence in Data Mining (pp. 537-552). Springer, Singapore.
[15] Hu, S., Liang, Y., Ma, L., & He, Y. (2009, October). MSMOTE: Improving classification performance when training data is imbalanced. In 2009 second international workshop on computer science and engineering (Vol. 2, pp. 13-17). IEEE.
[16] Liu, Q., Wang, X., Huang, X., & Yin, X. (2020). Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. Tunnelling and Underground Space Technology, 106, 103595.
[17] Amirruddin, A. D., Muharam, F. M., Ismail, M. H., Tan, N. P., & Ismail, M. F. (2020). Hyperspectral spectroscopy and imbalance data approaches for classification of oil palm's macronutrients observed from frond 9 and 17. Computers and Electronics in Agriculture, 178, 105768.
[18] Deng, X., Xu, Y., Chen, L., Zhong, W., Jolfaei, A., & Zheng, X. (2020). Dynamic clustering method for imbalanced learning based on AdaBoost. The Journal of Supercomputing, 76(12), 9716-9738.
[19] Haixiang, G., Yijing, L., Yanan, L., Xiao, L., & Jinling, L. (2016). BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. Engineering Applications of Artificial Intelligence, 49, 176-193.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).