Proceeding International Conference on Religion, Science and Education (2026) **5**, 1491-1496

Improving minority class detection in breast cancer dataset using SMOTE with SVM and adaboost

Ulfasari Rafflesia
Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Indonesia,
Jl. WR. Supratman Kandang Limun Bengkulu 38371, Indonesia.

Corresponding author*
ulfasari@unib.ac.id

Abstract: Imbalanced data often hinders the ability of classification models to accurately detect minority classes, leading to biased predictions toward the majority class. This study aims to enhance the detection performance of minority classes in Breast Cancer data by integrating the Synthetic Minority Over-sampling Technique (SMOTE) with two classification algorithms: Support Vector Machine (SVM) and Adaptive Boosting (AdaBoost). The Breast Cancer dataset from the UCI Machine Learning Repository is used due to its imbalanced class distribution. The proposed approach applies SMOTE at the preprocessing stage to generate a more balanced dataset before model training. Both SVM and AdaBoost are then evaluated on the original and SMOTE-balanced data using three performance metrics: accuracy, sensitivity, and F1 score. Experimental results demonstrate that the application of SMOTE significantly improves the classifiers' ability to recognize minority classes. Furthermore, AdaBoost achieves superior performance compared to SVM in terms of both overall accuracy and minority class sensitivity.

Keywords: Classification, imbalanced data, SMOTE, SVM, AdaBoost.

Abbreviations: SMOTE: Synthetic Minority Over-sampling Technique, SVM: Support Vector Machine, AdaBoost: Adaptive Boosting.

Introduction

SMOTE has demonstrated efficacy in enhancing model performance in numerous instances; nevertheless, most prior research has concentrated on binary classification issues or employed a singular classification technique. Classification is a principal machine learning job, extensively applied across diverse domains, including medical diagnosis, pattern identification, and scientific data analysis (Sen, P. C., Hajra, M., & Ghosh, M., 2019., Kumar, Y., Kaur, K., & Singh, G., 2020). The data distribution utilized during model training significantly affects the caliber of classification outcomes. A prevalent issue is class imbalance, characterized by a disproportionate number of samples across classes (Han, J., & Kamber, M., 2001; Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V., 2018). This disparity may lead the model to favor the majority class, diminishing its efficacy in identifying the minority class (Cieslak & Chawla, 2008). This issue arises not only in binary classification but also in multi-class scenarios with more intricate data distributions.

Multiple studies have suggested diverse methods to tackle class imbalance, including adjusting class weights, algorithm modification, and data pretreatment techniques, such as oversampling and undersampling. The Synthetic Minority Over-Sampling Technique (SMOTE) is a widely utilized method that creates synthetic samples for the minority class by using data from nearest neighbors (Blagus, R., & Lusa, L., 2013; Douzas, G., Bacao, F., & Last, F., 2018). While SMOTE has demonstrated efficacy in enhancing model performance in numerous instances, the majority of prior research has concentrated on binary classification issues or employed a singular

ISSN 2829-3355 | EISSN 2828-8467

classification method (Wang, S., Dai, Y., Shen, J., & Xuan, J., 2021). Limited research remains that assesses the efficacy of SMOTE on multi-class data by contrasting the performance of various algorithms, especially with assessment criteria that emphasize the model's capacity to identify minority classes.

This study assesses the effect of implementing SMOTE on the efficacy of three classification algorithms: SVM, decision tree, and Adaptive Boosting (AdaBoost) in multi-class data. It highlights the enhancement of minority class sensitivity while maintaining the model's overall accuracy. Three benchmark datasets from the UCI Machine Learning Repository—Breast Cancer, Ecoli, and Glass—were chosen to exemplify the diversity in multi-class data structures.

This research primarily contributes by: (1) integrating SMOTE with three distinct classification algorithms and assessing their performance in multi-class scenarios, (2) conducting a comparative analysis of performance between imbalanced original data and balanced data, and (3) highlighting the evaluation of sensitivity and F1-score to measure the model's capacity to identify minority classes. This article is structured as follows: Section 2 delineates the experimental methodologies and design, Section 3 elucidates the test results and analysis, and Section 4 encompasses the conclusions and recommendations for future research.

Materials And Methods

Synthetic Minority Oversampling Technique (SMOTE)

Data imbalance occurs when the quantity of observations in one class significantly exceeds that of others. The class with the highest number of observations is referred to as the majority class, and the others are designated as minority classes. This scenario frequently results in biased predictions, as traditional classification methods typically prefer the majority class and regard minority instances as noise, leading to the misclassification of minority observations (Galar et al., 2011). Resampling is frequently conducted before model development to resolve this issue (Boonamnuay et al., 2018).

The Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002), is a commonly employed resampling method. SMOTE produces synthetic instances for the minority class through the k-nearest neighbor method, thereby enhancing class equilibrium. The calculation of the distance between samples is contingent upon the type of attributes. SMOTE utilizes the Euclidean distance for numerical variables.

$$D(x,y) = \sqrt{(x-y)'(x-y)}$$

For categorical attributes, SMOTE employs the Value Distance Metric (VDM):

$$D(x,y) = \sum_{i=1}^{N} \delta(v_{1i}, v_{2i})$$

where the attribute-level distance is defined as:

$$\delta(v_1, v_2) = \sum_{i=1}^{S} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|$$

Through these distance computations, novel minority-class samples are generated between an instance and its closest neighbors, resulting in a more equitable dataset and enhancing classifier efficacy.

Support Vector Machine (SVM)

The SVM method is a supervised machine learning technique grounded in statistical learning theory (Wendong, Y., Zhengzheng, L., & Bo, J., 2017). It extracts a collection of distinctive subsets from the training samples, ensuring that the classification of these subsets corresponds to the segmentation of the complete dataset. The SVM has been effectively employed to address various classification challenges across numerous applications (Tao, P., Sun, Z., & Sun, Z., 2018., Maulud, D., & Abdulazeez, A. M., 2020).

AdaBoost

Adaboost is the preeminent ensemble boosting technique (Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G., 2013). Boosting techniques enhance accuracy in classification and prediction by producing model combinations. Adaboost is extensively utilized across various domains because to its robust theoretical framework, precise predictions, and ease of implementation (Wang, R., 2012; Gamil, S., Zeng, F., Alrifaey, M., Asim, M., & Ahmad, N., 2024; Bian, J., Wang, J., & Yece, Q., 2024). The chosen classification or prediction outcome is the model characterized by attributes comprising weight values (Domingo, C., & Watanabe, O., 2000; Jiang, X., Xu, Y., Ke, W., Zhang, Y., Zhu, Q. X., & He, Y. L., 2022). The weighting in that procedure signifies the significance of an object for accurate classification. The chosen classification outcome is the model with the highest weight value.

The fundamental classifier in the AdaBoost algorithm is referred to as a weak learner, specifically a decision stump. A stump, or little tree, is a classification tree with solely a root node and two leaves. Each stump is fabricated in succession. The prediction error produced by a stump influences the development of succeeding stumps. This indicates that misclassified observations will receive increased emphasis in the subsequent phase. The ensemble method is executed by evaluating the voting weight of each stump.

Data analysis

This study employs a dataset from the UCI Machine Learning Repository, specifically the Breast Cancer dataset. The dataset comprises 699 cases, featuring nine predictive attributes and two target classes. Like numerous medical diagnostic datasets, it demonstrates a significant class imbalance, with the minority class constituting merely 35% of the overall samples. This mismatch can adversely affect the efficacy of classification algorithms, as they often favor the dominant class and may struggle to recognize instances of the minority class reliably. The Synthetic Minority Over-sampling Technique (SMOTE) is utilized before the classification process to resolve this issue.

The Breast Cancer dataset exhibited a notable class imbalance, comprising 458 instances in the benign class (2) and merely 241 instances in the malignant class (4). The imbalanced ratio favored the majority class in the model. After the implementation of SMOTE, the quantity of the malignant class increased to 723. Conversely, the innocuous class stayed at 964, resulting in a more equitable ratio of around 1.33:1. This modification increased the model's sensitivity to malignant cases, hence enhancing predicted performance for previously underrepresented minority instances.

The evaluation of the Breast Cancer dataset concentrated on the efficacy of SVM and AdaBoost, both with and without the implementation of SMOTE. Performance was evaluated through accuracy, sensitivity, and F1-score. The findings demonstrate that the integration of SMOTE significantly enhanced the sensitivity of the minority class for both SVM and AdaBoost, while preserving accuracy at a similar level.

Results And Discussion

Experimental Results

This study assessed two classification algorithms—Support Vector Machine (SVM) and AdaBoost—utilizing the Synthetic Minority Over-sampling Technique (SMOTE). Performance evaluation depends on three often utilized metrics: accuracy, sensitivity, and F1-score. The implementation of SMOTE on the Breast Cancer dataset significantly improved the sensitivity of the minority class while preserving good overall accuracy. This illustrates that pre-training dataset balancing enhances the detection of minority-class occurrences by both SVM and AdaBoost, alleviating the prevalent bias towards the majority class often seen in imbalanced medical datasets. The performance results of both models post-SMOTE application are encapsulated in Table 1, which offers a comprehensive comparison of accuracy, minority-class sensitivity, and F1-score.

Table 1. Evaluation Results of SVM and AdaBoost on the Breast Cancer Dataset with SMOTE

Dataset	Methods	Accuracy (%)	Sensitivity Minority (%)	F1 (%)
Breast Cancer	SVM	97.8	98.9	97.8
	AdaBoost	99.8	99.7	99.8

Note: our own research

According to the evaluation results in Table 1, both SVM and AdaBoost exhibit strong performance on the Breast Cancer dataset after the implementation of SMOTE. The SVM classifier attains an accuracy of 97.8% and a minority class sensitivity of 98.9%. The results demonstrate that SVM effectively identifies nearly all minority-class instances, indicating that SMOTE successfully addresses the imbalance that usually leads SVM to favor the majority class. The elevated sensitivity indicates that SVM is proficient in reducing false negatives—crucial in medical diagnosis, where the failure to detect a positive case might be detrimental.

AdaBoost, nonetheless, surpasses SVM in all assessment metrics. AdaBoost exhibits remarkable resilience and reliability, achieving an accuracy of 99.8% and a minority-class sensitivity of 99.7%. The nearly flawless detection rate of minority instances demonstrates that the boosting process successfully integrates several weak learners to create a highly discriminative model. This robust performance indicates that AdaBoost significantly benefits from the balanced training distribution generated by SMOTE, enhancing its generalization and minimizing both false negatives and false positives.

The results underscore the necessity of rectifying class imbalance before classification. SMOTE consistently augments the capacity of both SVM and AdaBoost to identify minority instances, with AdaBoost exhibiting the most significant enhancement. These findings underscore the need to integrate resampling techniques with ensemble learning methods, especially in critical areas like medical diagnostics, where precise identification of minority cases is vital.

Discussion

This study's results indicate that the implementation of SMOTE significantly enhances the model's capacity to identify minority-class cases in the Breast Cancer dataset. Table 1 demonstrates that both SVM and AdaBoost achieve elevated accuracy and sensitivity following class balancing, signifying that SMOTE effectively alleviates the bias typically present in imbalanced medical datasets. The notable enhancement in minority-class sensitivity and F1-score indicates that the resampling procedure allows the models to prioritize not only overall accuracy but also equitable predictive performance among classes.

The SVM demonstrates robust classification performance after the implementation of SMOTE, attaining elevated sensitivity while preserving competitive accuracy. This suggests that SVM advantages arise from a more equitable distribution of training instances, hence diminishing the likelihood of false negatives—a crucial element in medical diagnosis, as misclassifying positive cases can result in significant clinical repercussions. The enhancement in F1-score underscores the model's improved capacity to equilibrate precision and recall for the minority class.

AdaBoost has the most consistent and exceptional performance among the assessed models. The near-optimal accuracy and sensitivity indicate that the boosting mechanism is exceptionally effective when integrated with SMOTE, enabling the model to discern nuanced patterns within the minority class that the majority class may otherwise obscure. This result underscores the proven efficacy of ensemble learning approaches in addressing intricate classification challenges, especially when complemented by suitable preprocessing techniques like oversampling.

These findings demonstrate that the application of SMOTE before classification markedly improves the efficacy of both SVM and AdaBoost in identifying minority-class cases in the Breast Cancer dataset. The findings highlight the necessity of addressing class imbalance in medical data and illustrate that the integration of SMOTE and ensemble learning provides a dependable and effective method for enhancing minority-class detection in clinical classification tasks.

Conclusions

This study sought to assess the efficacy of the Synthetic Minority Over-sampling Technique (SMOTE) in enhancing the performance of classification models utilized on the unbalanced Breast Cancer dataset. Two classification methods, Support Vector Machine (SVM) and AdaBoost, were analyzed to evaluate the impact of class balancing on predicting performance. The findings indicate that SMOTE effectively augments minority-class representation and significantly enhances model performance, especially regarding sensitivity and F1-score. These enhancements demonstrate that SMOTE facilitates improved pattern recognition within the minority class, therefore diminishing the bias commonly found in imbalanced medical datasets.

The results affirm that the implementation of SMOTE before classification yields more dependable and equitable predictions for both SVM and AdaBoost. AdaBoost, specifically, demonstrated the most significant performance improvements, indicating that ensemble methods can efficiently utilize balanced training data to enhance diagnostic accuracy. Future research may investigate alternate resampling techniques, hybrid oversampling approaches, or the use of cost-sensitive learning to improve further classification efficacy in medical datasets characterized by significant class imbalance. Concluding sentence should be given at after the discussion.

Conflict of Interest: The authors declare that there are no conflicts of interest concerning the publication of this article.

References

Sen, P. C., Hajra, M., & Ghosh, M. (2019). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modelling and graphics: Proceedings of IEM graph 2018 (pp. 99-111). Singapore: Springer Singapore.

Kumar, Y., Kaur, K., & Singh, G. (2020, January). Machine learning aspects and its applications towards different research areas. In 2020 International conference on computation, automation and knowledge management (ICCAKM) (pp. 150-156). IEEE.

Han, J., & Kamber, M. (2001). Data mining concepts and techniques San Francisco Moraga Kaufman.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, 17(2), 225-252.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics, 14(1), 106.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information sciences, 465, 1-20.

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific reports, 11(1), 24039.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). 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), 42(4), 463-484.

Boonamnuay, S., Kerdprasop, N., & Kerdprasop, K. (2018). Classification and regression tree with resampling for classifying imbalanced data. International Journal of Machine Learning and Computing, 8(4), 336-340.

Wendong, Y., Zhengzheng, L., & Bo, J. (2017, June). A multi-factor analysis model of quantitative investment based on GA and SVM. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1152-1155). IEEE.

Tao, P., Sun, Z., & Sun, Z. (2018). An improved intrusion detection algorithm based on GA and SVM. Ieee Access, 6, 13624-13631.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of applied science and technology trends, 1(2), 140-147.

Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.

Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), 745-758.

Wang, R. (2012). AdaBoost for feature selection, classification and its relation with SVM, a review. Physics Procedia, 25, 800-807.).

Gamil, S., Zeng, F., Alrifaey, M., Asim, M., & Ahmad, N. (2024). An efficient AdaBoost algorithm for enhancing skin cancer detection and classification. Algorithms, 17(8), 353.

Bian, J., Wang, J., & Yece, Q. (2024). A novel study on power consumption of an HVAC system using CatBoost and AdaBoost algorithms combined with the metaheuristic algorithms. Energy, 302, 131841.

Domingo, C., & Watanabe, O. (2000, June). MadaBoost: A modification of AdaBoost. In colt (pp. 180-189).

Jiang, X., Xu, Y., Ke, W., Zhang, Y., Zhu, Q. X., & He, Y. L. (2022). An imbalanced multifault diagnosis method based on bias weights AdaBoost. IEEE Transactions on Instrumentation and Measurement, 71, 1-8.