International Journal Of Artificial Intelegence Research
Vol 4, No 2, December 2020, pp. 86-94
ISSN: 2579-7298

Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification

Mardhiya Hayaty [a,1], Siti Muthmainah[b,2], Syed Muhammad Ghufran[c,3*]

[a,b] *Faculty of Computer Science, Universitas Amikom Yogyakarta, Ring Road Utara, Yogyakarta, Indonesia*
Department of Mathematics Abdul Wali Khan University, Mardan Garden Campus, Pakistan
[1] *mardhiya_hayati@amikom.ac.id;* [2] *siti.mutmainah@students.amikom.ac.id;* [3] *smghufran@awkum.edu.pk**
** corresponding author*

ARTICLE INFO

Article history:
Received 07 Feb 2020
Revised 12 April 2020
Accepted 22 Oct 2020

Keywords:
Data Imbalance
Classification
SMOTE
ROS

ABSTRACT

Background: High accuracy value is one of the parameters of the success of classification in predicting classes. The higher the value, the more correct the class prediction. One way to improve accuracy is dataset has a balanced class composition. It is complicated to ensure the dataset has a stable class, especially in rare cases. This study used a blood donor dataset; the classification process predicts donors are feasible and not feasible; in this case, the reward ratio is quite high.
Objective : This work aims to increase the number of minority class data randomly and synthetically so that the amount of data in both classes is balanced.
Method : SMOTE-OverSampling(SOS) and Random Oversampling (ROS) methods make sample data replication in minority class known as synthetic
Results :The application of SOS and ROS succeeded in increasing the accuracy of inappropriate class recognition from 12% to 100% in the KNN algorithm. In contrast, the naïve Bayes algorithm did not experience an increase before and after the balancing process, which was 89%.
Conclusion: The data balancing process succeed to improve the occupation of predicting class, however, the choosing of classification algorithm to be considered.

I. Introduction

High accuracy value is one of the parameters of the success of classification in predicting classes. The higher the value, the more correct the class prediction. Many classification algorithms have been used, such as Naïve Bayes, C.45, KNN, and many more. The algorithm continues to be developed by researchers to produce the best accuracy value, as has been observed by [1] upgrading the Evolutionary Algorithm (EA) in previous studies by automatically selecting the best classifier compared to random forest. Also, in the case of the multiclass dataset and multi-level classification, it has been used for medical purposes, namely the prediction of multiple skin lesions[2].

However, to get the best accuracy value does not only depend on the algorithm used, the character factor dataset used has considerable influence. [3] conducted a comparative empirical study of 11 classification algorithms on results in varying performance. Not surprisingly, algorithms with high accuracy have average or slow training time efficiency. Some algorithms have almost the same performance by only a few percent difference. That means that whatever choice algorithm does not have a significant effect on increasing accuracy.

One way to improve accuracy is that the dataset has a balanced class composition[4]. It is complicated to ensure the dataset has a stable class, especially in rare cases. The case will become minority data and cause prediction accuracy to be minimal [5] compared to other majority data. A substantial prediction error causes it. This event is called a class imbalance, where the number of class members is not balanced compared to other classes [6].

Research on imbalanced data has been widely carried out by researchers with a variety of approaches.[7]–[9] addressing the imbalance data problem at the algorithm level that is an ensemble. Research [8] using RUSboost, LogitBoost, and AdaBoostm1 with the best model results is that RusBoost can classify imbalanced data on high imbalanced ratios. At the same time, [10] proposes Modified Boosted SVM (MBSVM) by making improvements to Wang's Boosted SVM algorithm by updating the imbalance based on distance weights, MBSVM uses 43 datasets.

In the data level approach, using the random sampling method consists of Linear, Shuffled and Stratified to overcome imbalanced data [11] and succeeded in increasing accuracy by which has values of accuracy, precision, recall and AUC> 0.8%

Imbalance of data occurs in rare cases, in the case of health "blood donor," most donors are people who routinely donate blood, so they are categories of people who are eligible to donate blood. Predict that a feasible class has high accuracy compared to an improper class because the number of unfit donors is so small that it has very low predictive accuracy. The amount of majority class data is superior to the minority class. Classes become unbalanced causing prediction errors

This study proposes an approach at the data level by increasing the number of minority class data randomly and synthetically so that the amount of data in both classes is balanced.

II. Material and Method

A. Dataset

The experiment used a primary blood donor dataset at a hospital in the blood transfusion unit. The donor data is 246, with 38 features. Programming tools used python programming and needed some libraries to pre-process, data balancing process, algorithm classification implement.

B. Research Stages

The result of pre-processing divided into training data and testing data. Calculation of imbalance ratio in class to see whether imbalance data occurs. SMOTE-OverSampling and random oversampling processes in minority classes to produce balanced data in all categories.

Implement the Naïve Bayes classification algorithm and K-NN to find the best modeling accuracy. Classification evaluation to see the comparison of efficiency without using a balancing technique by using a balancing method. Work steps, like the chart below.

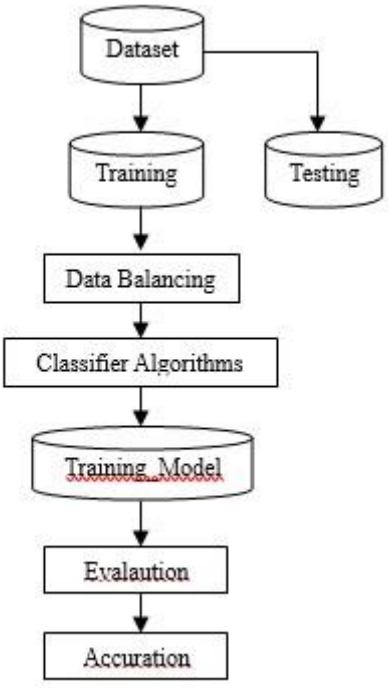

Figure 1. Research Stages

C. Pre-Processing

Pre-processing aims to an understanding of data, improve data quality, and functional data mining results. This process has a comprehensive portion of work and spends almost 70% of the data mining process. Some of the preprocessing work is[12] data cleaning, data integration, normalization, data desiccation

D. Naïve Bayes

Classification methods that are supervised learning or statistics of a class by looking for probabilities[13]. Naïve Bayes classifies data by estimation to determine the probability of P (H | X), where X is the proof and H is the hypothesis and P (H | X) is the probability of posterior H with the condition X to determine the likelihood, or conditional probability[12]

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

E. K-Nearest Neighbour (KNN)

The K-Nearest Neighbor method searches for the closest number of K data objects or training patterns [14] with an input pattern then selects the class of the most models. K value is the number of closest neighbors that will be involved to determine the prediction of class labels in the test data. K was chosen based on class voting from neighboring K.Calculate distances with neighbors using the Euclidean Distance formula [15]

$$dxy = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

F. SMOTE & ROS

The imbalanced data case uses the Synthetic Minority Over-sampling Technique (SMOTE) method[16]. SMOTE will make sample data replication in minority class known as synthetic. How to obtain synthesis data at random the nearest sample data [17] as much as the percentage of data duplication desired

Class imbalance is a class imbalance where the number of majority classes is greater than the minority class, for example, the data has a ratio of 1: 100 where 1 is a minority and 100 is the majority [18]. Whereas Random Over Sampling is one technique that adds data to a minor class randomly without adding variations in class data [19].

Class imbalance results in machine learning incorrectly classifying classes. This approach makes a replica of a minority class, replication known as synthetic data. Each minority data is made of synthetic data as much as the desired duplication percentage.

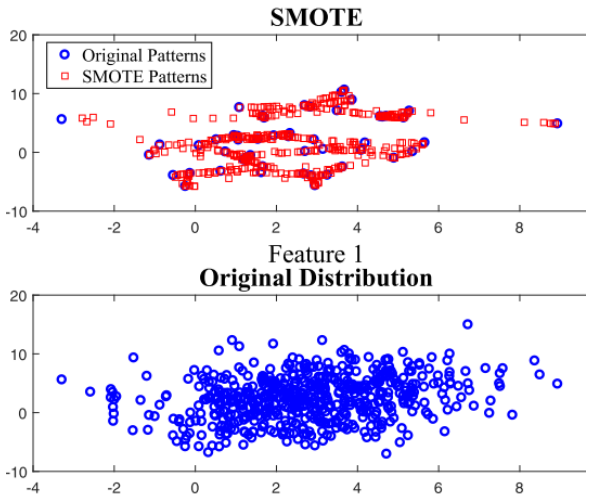

Figure 2. SMOTE Generated vs. Original Example [16]

G. Confusion Matrices

The performance of a system cannot work 100% correctly. Therefore a classification system must measure its performance using confusion matrices. Confusion matrices are tables that record the results of classification performance[12].

		Predicted Class		
		yes	no	Total
Actual class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

Figure 3. Confusion Matrix

Understanding them will make it easy to grasp the meaning of the various measures.

- True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- True negatives(TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- False positives (FP): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer=yes). Let FP be the number of false positives.
- False negatives (FN): These are the positive tuples that were mislabeled as negative (e.g., tuples of class buys computer = yes for which the classifier predicted buys computer=no). Let FN be the number of false negatives.

The accuracy value uses the accuracy formulation to test the correctness [20].

$$Acc = \frac{TP + TN}{P + N} \quad (3)$$

Precision is a metric that measures performance to get relevant data (the amount of right positive data) while recall measures the performance of relevant data reads against the amount of data (true positive + false negative).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

The performance measurement uses a confusion matrix with values of precision, recall, and accuracy. Precision is the level of accuracy of information requested with answers, while recall is the level of success of rediscovering information.

III. Results and Discussion

The blood donor dataset is 247, the qualified donor data is 238, and the unqualified is 9. The dataset consists of training data and testing data.

Table 1. Data Proportion

Data	Qualified	Unqualified
Training	142	9
Testing	96	9

According to the table above, the training data for a class is qualified and unqualified, and there is an imbalance of data and imbalance ratio value is 1:16, Imbalance ratio formula [21]

$$IR = \frac{N_-}{N_+} \quad (6)$$

The SMOTE Oversampling (SOS) and Random Oversampling techniques (ROS) balanced minority data classes (unqualified) and majority classes (qualified) using python library tools such as the following code.

```
From imblearn.over_sampling import SMOTE
From imblearn.over_sampling import RandomOverSampler
#subsequently
x_train, y_train =                          SMOTE()/
RandomOverSampler()
.fit_resample(x_train, y_train)
from collections import Counter
print(sorted(Counter(y_train).items()))
```

Figure 4. Definition: import library imbalance

The ROS method increased the amount of secondary class data (unqualified) randomly taken from the original class date. At the same time, SOS not only increased the amount of data but also adds data variations from the synthetic class originals. The processed data were balancing results the amount of data unqualified equal to the amount of data qualified.

Table 2. Data proportion after balancing

Non/Balancing	Qualified	Unqualified
No-balancing	142	9
SMOTE-OverSampling(SOS)	142	142
Random Over Sampling(ROS)	142	142

Modeling training data used the KNN algorithm and Naïve Bayes. Experiments applied the value of K = 1 to K = 10 on the KNN algorithm. Classification evaluation uses confusion matrix, comparison of accuracy values without data balancing, and with data balancing (SOS and SOS) in the following table.

Table 3.Accuration

Non/Balancing	KNN	Naive Bayes
No-balancing	92%	99.04%
SMOTE-OverSampling(SOS)	87.33%	99.04%
Random Over Sampling(ROS)	97%	99.04%

In table 3 above presents a reasonably high accuracy value using non-balancing data on the two classification algorithms, but the prediction results are biased to the majority class. The accuracy of the Naïve Bayes algorithm shows no difference before or after balancing, as in the graph below.

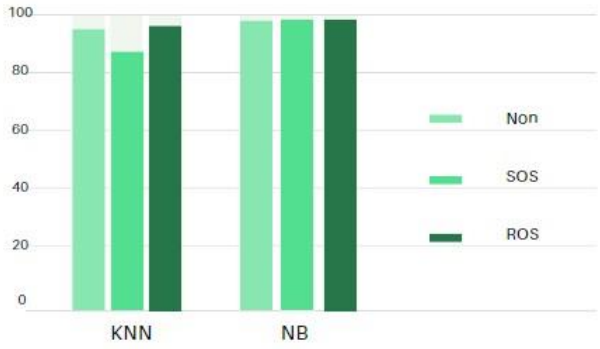

Figure 5. The Accuration

The recall value in table 4 (KNN algorithm) without the balancing process is shallow, meaning that the system recognize qualified class correctly only 12% while the class unqualified is 100%

Table 4. Recall K-Nearest Neighbor

Non/Balancing	Qualified	Unqualified
No-balancing	100%	12%
SMOTE –OverSampling(SOS)	87%	100%
Random Over Sampling(ROS)	97%	100%

The process of balancing data in the minor class succeeded in raising the prediction of the unqualified class significantly to 100% so that the projection of a qualified and unqualified class did not differ

significantly. The SOS and ROS balancing methods can increase the value of recall to an almost balanced on the K-NN classifier, like Figure 6 below.

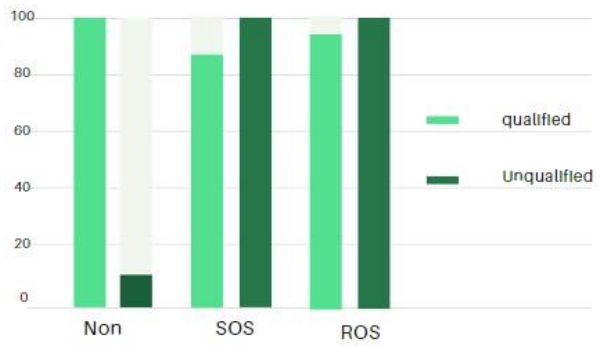

Figure 6. Recall - KNN

Naïve Bayes algorithm produces better predictions even though it does not use the balancing process compared to the KNN implementation. There is no prediction change in both classes, before or after the balancing process. It has a fixed value of 100% (qualified), and 89% (unqualified class).

Table 5. Recall Naïve Bayes

Non/Balancing	Qualified	Unqualified
Non	100%	89%
SMOTE –OverSampling(SOS)	100%	89%
Random Over Sampling(ROS)	100%	89%

Likewise, with research [22], the application of balancing using SMOTE OverSampling (SOS) only slightly increases the value of accuracy.

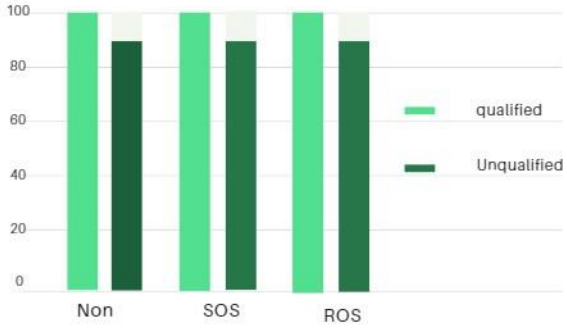

Figure 7. Recall- Naïve Bayes

IV. Conclusion

Data that have a class Imbalance case tend to have the classification results more inclined to the majority class (qualified) than the minority class (unqualified). The blood donor dataset has a reasonably high imbalance ratio between the qualified and unqualified class. In this study, achieved a balanced amount of data in both classes. The number of data in the minority class only 9 data increased to 142 data. The application of SOS and ROS succeeded in increasing the accuracy of inappropriate class recognition from 12% to 100% in the KNN algorithm. In contrast, the naïve Bayes algorithm did not experience an increase before and after the balancing process, which was 89%.

This study used SOS and ROS to overcome imbalance data in binary class cases; it needs to test for multiple class cases

References

[1] J. C. Xavier-Júnior, A. A. Freitas, T. B. Ludermir, A. Feitosa-Neto, and C. A. S. Barreto, "An evolutionary algorithm for automated machine learning focusing on classifier ensembles: An improved algorithm and extended results," *Theor. Comput. Sci.*, vol. 805, pp. 1–18, 2019.

[2] N. Hameed, A. M. Shabut, M. K. Ghosh, and M. A. Hossain, "Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques," *Expert Syst. Appl.*, vol. 141, p. 112961, 2020.

[3] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, 2017.

[4] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Inf. Sci. journal-Elsivier*, no. xxxx, 2019.

[5] W. Lu, Z. Li, and J. Chu, "Adaptive Ensemble Undersampling-Boost: A novel learning framework for imbalanced data," *J. Syst. Softw.*, vol. 132, pp. 272–282, 2017.

[6] M. Palt and M. Palt, "ScienceDirect The Proposal of Undersampling Method for Learning from The Proposal of Undersampling Method for Learning from Imbalanced Datasets Imbalanced Datasets," *Procedia Comput. Sci.*, vol. 159, pp. 125–134, 2019.

[7] H.-J. Xing and W.-T. Liu, "Robust AdaBoost based ensemble of one-class support vector machines," *Inf. Fusion*, vol. 55, no. July 2019, pp. 45–58, 2020.

[8] P. Chujai, K. Chomboon, P. Teerarassamee, N. Kerdprasop, and K. Kerdprasop, "Ensemble Learning For Imbalanced Data Classification Problem," no. January 2015, pp. 449–456, 2015.

[9] B. Krawczyk, A. Cano, and M. Wozniak, "Selecting local ensembles for multi-class imbalanced data classification," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, 2018.

[10] Sundar R and Punniyamoorthy M, "Performance enhanced Boosted SVM for Imbalanced datasets," *Appl. Soft Comput. J.*, vol. 83, p. 105601, 2019.

[11] S. Mutrofin, A. Mu'alif, R. V. H. Ginardi, and C. Fatichah, "Solution of class imbalance of k-nearest neighbor for data of new student admission selection," *Int. J. Artif. Intell. Res.*, vol. 3, no. 2, 2019.

[12] J. Han, Jiawei; Kamber, Micheline; Pei, *Data Mining Concepts and Techniques*. Elsivier, 2012.

[13] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[14] O. Kramer, "Dimensionality Reduction with Unsupervised Nearest Neighbors," *Intell. Syst. Ref. Libr.*, vol. 51, pp. 13–23, 2013.

[15] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 294–298, 2018.

[16] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, 2002.

[18] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018.

[19] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," *Proc. - 2019 IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2019*, pp. 175–183, 2019.

[20] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[21] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large

number of imbalanced datasets," *Appl. Soft Comput.*, vol. 83, p. 105662, 2019.

[22] A. Hanskunatai, "A New Hybrid Sampling Approach for Classification of Imbalanced Datasets," *2018 3rd Int. Conf. Comput. Commun. Syst. ICCCS 2018*, pp. 278–281, 2018.