2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA) | 978-1-6654-2680-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/DATABIA53375.2021.9650084

Improving Machine Learning Prediction of Peatlands Fire Occurrence for Unbalanced Data Using SMOTE Approach

1st Dedi Rosadi *
Department of Mathematics
Gadjah Mada University
Yogyakarta, Indonesia
dedirosadi@gadjahmada.edu*

2nd Deasy Arisanty
Department of Geography Education
Lambung Mangkurat University
Banjarmasin, Indonesia
deasyrsnt@gmail.com

3rd Widyastuti Andriyani
Master of Information Technology
STMIK Akakom
Yogyakarta, Indonesia
Widy.ugm@gmail.com

4th Shelton Peiris
School of Mathematics and Statistics
University of Sydney
Sydney, Australia
shelton.peiris@sydney.edu.au

5th Dina Agustina
Department of Statistics
State University of Padang
Padang, Indonesia
dinagustina31@gmail.com

6th David Dowe,
Dept of Data Science and AI,
Monash University
Clayton, Australia
david.dowe@monash.edu

7th Zheng Fang
PERJ Education
Guangdong, China
mrpercyjardine@gmail.com

***Abstract*—From our previous study, we have known that only a small number of literatures have studied peatlands fire modeling in Indonesia. It is including our recent study on the prediction of the forest fire occurrence in the peatlands area using some machine learning classification techniques. In the previous empirical study using data from South Kalimantan Province in Indonesia, we found that the datasets are unbalanced between the two classes of data, i.e., the occurrence of fire hotspots and the nonoccurrence of fire hotspots areas. In this paper, the performance of the classification method is improved, by balancing the data using what so called Synthetic Minority Over-sampling Technique (SMOTE). In the empirical results, we show the performance of the classification results on the balanced data are mixed. It is found that only using the ensemble AdaBoost with SMOTE balanced data the performance of the methods has always been improved over unbalanced data, either for in-sample or for out-sample cases. The open-source software R is used for implementation of the methods.**

Keywords—peatlands fire, classification methods, balanced data, unbalanced data, SMOTE

I. INTRODUCTION

It has been discussed in various studies that forest fire prediction is an important step for early warning system in forest fire fighting. The accuracy of prediction of the forest fire events relies heavily on the methods of prediction used in the study. In some recent literatures, it is known that various methods can be used to obtain the events prediction. It is including physics-based models, statistical-mathematics models, and data mining/machine learning approaches, see e.g., [1].

Recent literature reviews show machine learning approach have become so popular in the study. Using the meteorological and forest weather index (FWI) variables, [2] study the prediction approach based on the classical classification methods, such as the Support Vector Machine (SVM), Regression based method, Artificial Neural Networks (ANN), Decision Trees (DT) and its extension, i.e., Random Forest (RF). In [3], it was extended the studies in [2] by using proposed hybrid prediction approach based on Fuzzy C-Means clustering and classification based on Back-Propagation Neural Networks (BPNN) model. The studies in [2] and [3] further extended in [4] by applying ensemble classification Adaptive Boosting (AdaBoost) approach ([5]). The study in [4] further improved in [6] by applying bagging (bootstrap aggregating) approach based on multinomial logit method. Literature [7] apply the Decision Tree approach on Bushfire prediction. More recent studies are available. For instance, in [8], it was considered various classical classification methods, such as Naïve Bayes (NB), SVM, DT, k-Nearest Neighbor (kNN) and Logistic Regression (logreg); and, an ensemble approach, namely AdaBoost (DT based) approach. Specifically, in [8], for empirical study using data from South Kalimantan Province, in Indonesia, we found that the datasets are unbalanced between the two classes of data, i.e., the occurrence and the nonoccurrence of fire hotspots area. In this paper, the preprocessing the data by balancing the data using what so called Synthetic Minority Over-sampling Technique (SMOTE). This method is proposed in literature for balancing the categorical data (see e.g., [9]), but has not been used for Indonesia peatlands fire study.

The rest of this paper is organized as follows. In Section 2, we outline short summary of the necessary background for understanding of the study and outlined the improved algorithms. The empirical studies are summarized in Section 3. Last section concludes the studies.

II. METHODS

A. Classification Methods: some classical approaches

In our empirical study, we apply several classification methods which may be considered classical, namely SVM,

NB, logreg, kNN, and Decision Tree (DT) method. As it is considered to be well known, for saving the space, however we do not provide outlines of the methods, see e.g., [8] for the summary. See also, e.g. [10], [11], [12], [13], [14] and [15] for further detail on each method.

B. Adaboost Method

Adaptive Boosting (AdaBoost), is one of the ensemble machine learning approaches. AdaBoost considers to improve the empirical performance using combination of the various "weaker learner" using

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{1}$$

where each *f* denotes a weak learner which uses an input x and gives the outcome the appropriate class of the input *x*. Further detail can be obtained in e.g, [5].

C. SMOTE (Synthetic Minority Oversampling Technique) Method

In solving machine learning problems, especially in classification, an imbalanced dataset is encountered where there is a minority class with a small sample of data. This certainly affects the classification results that are not "optimal". One possible way to handle the imbalanced dataset case, is by resampling data to make it balanced. One of the most common approaches to do oversampling is "the Synthetic Minority Oversampling Technique (SMOTE)" which was introduced in [16]. SMOTE does not replace the data from the minority class but does something called "Synthetic" to generate data for the minority class by implementing a k-nearest neighbor algorithm.
The algorithms of the method work as follows.

SMOTE Algorithm
Input: Number of minority class samples T; Amount of SMOTE N%; Number of *k* nearest neighbors

1. Setting the amount of SMOTE N%
2. The T minority class samples is randomized
3. Choose number of k nearest neighbors
4. One sample is generated using k nearest neighbors as follow
   a. Take the difference between (sample) under consideration and its nearest neighbor.
   b. Multiply a random number between 0 and 1 to the results from part (a)
   c. Add it to the sample under consideration.

   This algorithm will select a random point between two specific two specific features. .

Output: As the output of this algorithm, it will be obtained (N/100) * T synthetic samples of the minority (smaller) class.

See e.g. [15] for further details.

III. RESULTS AND DISCUSSION

A. Data Description

For the empirical study, here we use topographical, satellite, and meteorological data from of peatlands in Kalimantan Selatan Province in Indonesia. For technical reason, however, we only able to obtain the data several days after the occurrence of fire hotspots. In this condition, the variable "area" is labeled as "1". To compare the variables values, we collect the data when the same area is labeled as "0", i.e. the time when there is no peatland fire in the same areas. The variables obtained are the following: the time (of data is collected), the topographical and meteorological data (i.e., district area, LST/Land surface Temperature, Wind Speed, Humidity, Height) and satellite data (NDVI /normalized vegetation index). The time frame of this study is year 2018 and consists of 202 observations, where 160 data are in area "1" and only 40 data in area "0", which can be seen as the imbalanced datasets. This data could be not optimal in the classification and probably can be balanced to obtain the better classification results.

B. Implementation of the methods

For implementation of the methods, we use the following steps. It follows closely the study in [8].

Preprocessing steps

1. We apply the full set of data for the study. Here we do not consider the size of the burned area. First we split the data into the case when the variable area is 0 (denoted as "No Burned Area") and the case of variable area is greater than 0 (denoted as "Burned Area").
2. We further apply normalization step to the data. The normalization step can be done using various approaches, in this study we consider only min-max normalization

$$v_i' = \frac{v_i - \min A}{\max A}(\text{newmax } A - \text{new min } A) + \text{new min } A \tag{2}$$

where
min A is the minimum an attribute A
max A is the maximum an attribute A
v_i is the value in attribute A
Here we use the range [0,1] as the range of [new min A, new max A].

Balancing Step

3. Apply the SMOTE algorithm to balance the data sets. The optimal parameterizations need to be determined and it can be obtained after several testing on the data.

Classification steps

4. For checking the performance of the methods, we randomly split the data into two parts. The first part of the data denotes as the testing data. The other parts is used to check the performance of the methods, and we called as testing data.
5. In the next step, we apply the considered classification approaches to the training data. The testing data is further used to check the performance of the best model obtained using training data.

6. Various measures are computed to check the performance of the considered method in step 5, however, to save the space, only accuracy measure is reported

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where in the categorical classification data,
TP is the true positive case
TN is the true negative case
FP is the false positive case, and
FN is the false negative case

The open-source software R ([17]), is used for implementation of the methods. Various packages and function is used, namely: smote {performanceEstimation} [18], naiveBayes {e1071} [19], svm {e1071} [20] , knn {class} [20], glm {stats} [21], ctree {party} [22] and boosting {adabag} [23] .

IV. Discussion

The sample sizes in the unbalanced (original) and balanced datasets after SMOTE appllication are given in Table 1. It seems that the SMOTE approach is able to improve the ratio between the two class of data, either in the original data before splitting, or in training/testing data. We found that this smote parameterisation are optimal.

The summary of the empirical studies is given in Table 2. The performance of various methods either in the unbalanced (original) data and the SMOTE balanced data either for data training and/or data testing are summarized in Table 2. For the empirical comparison purpose, several training and testing sample sizes are considered in the study. In the empirical study, by balancing the data using the SMOTE approach does not always improve the accuracies of the methods. Either in the unbalanced or balanced datasets, here we can see that the accuracy of AdaBoost method outperforms the other approaches considered in the study in the in-sample data. Here it can be seen that only by using AdaBoost approach the improvement has been obtained in the outsample data. The improvement can be obtained between 2-6% in all cases considered in the study. In general, therefore, this study showed that for peantland fire prediction, the machine learning approaches offers various powerful detection methods.

V. Conclusion

This study showed that the machine learning approaches, either the classical or more recent approaches, could be used for fire occurrence detection of peatlands. These approaches in general can be used in various types of fires, including bush fires, forest fires or peatland fires. When the data is unbalanced between classes, the acccuracy of the prediction can be improved by preprocessing the data using SMOTE approach, to obtain a balanced sample, together with the application of the ensemble classification approach.

Acknowledgement

The authors acknowledge the receive of research funding PD 2021 from Deputy of Reseach and Development, Ministry of Research and Technology of Republic of Indonesia.

TABLE I. SAMPLE SIZES IN THE STUDY

Type of Data		Class	Ratio Data Testing and Training		
			9:1	8:2	7:3
Unbalanced Case (Original Data)	**Sample size before splitting**	0	160	160	160
		1	42	42	42
	Sample Size in Training Data	0	141	124	141
		1	41	38	6
	Sample Size in Testing Data	0	19	36	61
		1	1	4	6
Balanced Case (SMOTE Data)	**Sample size before splitting**	0	168	168	168
		1	126	126	126
	Sample Size in Training Data	0	147	127	114
		1	118	108	92
	Sample Size in Testing Data	0	21	41	54
		1	8	18	34

TABLE II. THE PERFORMANCE OF THE METHODS BEFORE AND AFTER BALANCED THE DATA

Algo-rithms	Ratio Data Testing and Training	Unbalanced Case (Original Data)		Balanced Case (SMOTE Data)	
		Accuracy training	*Accuracy testing*	*Accuracy training*	*Accuracy testing*
SVM	9:1	91,21%	95,00%	93.21%	82.76%
	8:2	90,74%	90,00%	93.62%	88.14%
	7:3	89.36%	91,8%	94.66%	92.05%
kNN	9:1 (k=3)	-	95.00%	-	86,21%
	8:2 (k=3)	-	92,50%	-	88,14%
	7:3 (k=3)	-	86,89%	-	87,50%
Logistic Regress-ion (logreg)	9:1	75.80%	90,00%	68,96%	51,3%
	8:2	74,69%	90,00%	49,78%	62,71%
	7:3	74,46%	86,88%	49,51%	54,54%
Decision Tree (DT)	9:1	91,00%	95,00%	88,00%	90,00%
	8:2	90,00%	92,00%	86,00%	93,00%
	7:3	89,00%	90,00%	90,00%	90,00%
Naïve Bayes (NB)	9:1	83.00%	90,00%	78,90%	82,80%
	8:2	82,1%	87,5%	77,40%	84,7%
	7:3	83%	88,50%	83,50%	84,10%
Adaboost (DT Based)	9:1	100%	95,00%	100%	96.55%
	8:2	100%	92,50%	100%	98.31%
	7:3	100%	91,80%	100%	95.45%

REFERENCES

[1] D. T. Buia, Q.-T. Buib, Q.-P. Nguyenc, B. Pradhand, H. Nampak and P. Trinh, "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area", Agricultural and Forest Meteorology, vol. 233, pp. 32-44, 2017

[2] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data", In: Neves J, Santos M F, Machado J (eds.) Proceding EPIA 2007, pp.512-523, 2007

[3] G. F. Shidik and K. Mustofa, "Predicting size of forest fire using hybrid model", Proceeding ICT EurAsia Conference 2014, pp. 316-327, 2014

[4] D. Rosadi and W. Andriyani, "Prediction of forest fire using ensemble method", Journal of Physics: Conf. Series ICMSE 2020, 1918, 2021

[5] J. Zhu, H. Zou, S. Rosset and T. Hastie, "Multi-class AdaBoost", Statistics and Its Interface, 2, pp.349–360, 2009

[6] D. Rosadi W. Andriyani and D. Arisanty, "Prediction of forest fire using hybrid fuzzy-clustering - bagging method", Presented in ICMSE 2021, Semarang, Indonesia, October 5-6, 2021

[7] D.L. Dowe and N. Krusel, "Decision Tree models of Bushfire Activity". AI Applications, .8, pp 71-72, 1994

[8] D. Rosadi W. Andriyani, D. Arisanty and D. Agustina, "Prediction of Forest Fire Occurrence in Peatlands using Machine Learning Approaches", Proceedings ISRITI 2020, 2021

[9] H. A. Khorshidi, U. Aickelin, "A Synthetic Over-sampling method with Minority and Majority classes for imbalance problems", https://arxiv.org/abs/2011.04170, 2021

[10] T. Hastie, R. Tibshirani and J. Friedman. "The Elements of Statistical Learning", 2nd Eds , New York: Springer, 2009

[11] A. Christmann, I. Steinwart, "Support Vector Machines", New York: Springer-Verlag, 2008

[12] Z. Zhang, "Introduction to machine learning: K-nearest neighbors", Annals of Translational Medicine, vol.4 , pp. 218-218, June 2016

[13] P. Cichosz, "Data Mining Algorithms: Explained Using R", NY: John Wiley & Sons, 2015

[14] G. James, D. Witten, T. Hastie and R. Tibshirani, "An Introduction to Statistical Learning: with Applications in R". NY: Springer, 2013.

[15] L. Rokach and O. Maimoon, "Data Mining with Decision Trees: Theory and Applications, SIngapore: World Scientific, 2008

[16] N. Chawla and K. Bowyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res 2002.

[17] R Core Team, "R: A language and environment for statistical computing", Vienna: R Foundation for Statistical, 2021

[18] L. Torgo, "An Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models in R", CoRR abs/1412.0436 [cs.MS], URL: http://arxiv.org/abs/1412.0436, 2014

[19] D. Meyer, "Support Vector Machines* The Interface tolibsvmin package e1071", Vienna: FH Technikum Wien, Austria, 2020

[19] B. D. Ripley, W. Venables, "Package 'class'", Vienna: CRAN, 2020

[20] R Core Team,"The R Stats Package ", Vienna: CRAN, 2020

[21] T. Hothorn, K. Hornik and A. Zeileis, "ctree: Conditional Inference Trees", Vienna: CRAN, 2020

[22] E. Alfaro," Package 'adabag'", Vienna: CRAN, 2020