Handling Imbalance Data in Classification Model with Nominal Predictors

Kartika Fithriasari, Iswari Hariastuti and Kinanthi Sukma Wening

***Abstract*—Decision tree, one of classification method, can be done to find out the factors that predict something with interpretable result. However, a small and unbalanced percentage will make the classification only lead to the majority class. Therefore, handling imbalance class needs to be done. One method that often used in nominal predictor data is SMOTE-N. For accuracy improving, a hybrid SMOTE-N and ADASYN-N was developed. SMOTE-N-ENN and ADASYN-N were developed for accuracy improvement. In this study, SMOTE-N, SMOTE-N-ENN and ADASYN-N will be compared in handling imbalance class in the classification of premarital sex among adolescent using base class CART. The conclusion obtained regarding the best method for handling class imbalance is ADASYN-N because it provides the highest AUC compared to SMOTE-N and SMOTE-N-ENN. The best decision tree provides information that factors that can predict adolescents having premarital sexual relations are dating style, knowledge of the fertile period, knowledge of the risk of young marriage, gender, recent education, and area of residence.**

***Index Terms*—ADASYN-N, CART, hybrid SMOTE-N, imbalanced data, premarital sex.**

I. Introduction

CLASSIFICATION analysis is one of the supervised methods that is widely used to overcome problems in data mining or industrial and social world. But, one one of the problems in classification is imbalanced data. This case is caused by imbalance class ratio independent or response variable. This imbalanced data will be harmful for researchers in data mining. Machine learning or classification analysis will difficult to classify minority class properly. In imbalanced data, the model that was built tend to lead to majority class, so that the minority class will be predicted to the majority class.

Thus, before doing classification, it is necessary to handle the imbalance case. There are some approaches to deal with imbalanced data, one of them is resampling. There are two kinds of resampling: undersampling and oversampling. Undersampling is a method that reduces the number of majority class until the amount same with minority class. Then oversampling is the reverse oversampling, that is oversample the minority class by replication until the proportion are balanced. Undersampling is rarely used because it can reduce or take important information from the dataset. However, oversampling can increase the possibility of overfitting because it duplicated the instances.

Due to the weakness of undersampling and oversampling, Chawla et al. (2002) proposed SMOTE (Synthetic Minority Oversampling Technique) for a numerical dataset, and SMOTE-N (Synthetic Minority Oversampling Technique Nominal) for nominal dataset [1]. SMOTE and SMOTE-N use k-nearest neighbors to create synthetic data, instead of replicate them. Then, as time goes by, appear the improved of SMOTE, called hybrid SMOTE. That is a combination of SMOTE and undersampling method. There is a lot of combination SMOTE and ENN (Edited Nearest Neighbor), SMOTE and Tomek Link, SMOTE and RUS (Random Undersampling), etc. One of hybrid resampling technique that give good performance is SMOTE-ENN in many dataset and may classifier [2].

Besides SMOTE and hybrid SMOTE, He et al. propose another method method to deal with imbalance, ADASYN (Adaptive Synthetic). Idea of ADASYN is oversampling minority class based on difficulity learning. Minority instances that are more difficult to learn will be given a higher weight and be generated more than minority instances that are easy to learn [3]. ADASYN by He et al. proposed for numerical data, then developed by Kurniawati in 2017 into ADASYN-N and ADASYN-KNN for nominal data. Rahayu et al. (2017) tested SMOTE-N, ADASYN-N, and ADASYN-KNN using Random Forest on several datasets in the multiclass category. The results obtained are ADASYN-N can improve accuracy better than SMOTE-N and ADASYN-KNN [4].

From that problem, this study aims to apply ADASYN-N and hybrid approach SMOTE-N in nominal categorical dataset using base classifier CART. The SMOTE-N hybrid approach used is a combination of SMOTE-N and Edited Nearest Neighbor (SMOTE-N-ENN). From that comparison, this paper contributes to give the best method for handling imbalanced data in classification model with nominal predictor. Also, this paper can help further research if there is an imbalance problem. Dataset that used in this study is highly imbalanced data from Survei Kinerja dan Akuntabilitas KKBPK Badan Kependudukan dan Keluarga Berencana Nasional Indonesia in 2018. The data consist of 15 independent variables with 1 binary dependent variable about premarital sex among adolescent. Proportion of minority class (teenagers who had premarital sex) is 1.6% and majority class (teenagers who had not premarital sex) is 98.4%.

Manuscript received February 23, 2020; accepted May 12, 2020.

K. Fithriasari and K.S. Wening are with the Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. E-mail: kartika_f@statistika.its.ac.id, kinanthi.sukmawening@gmail.com

I. Hariastuti is with the National Family Planning Coordinating Board (BKKBN), East Java, Indonesia. E-mail: iswarihariastuti@yahoo.com

II. Preliminaries

A. Classification and Regression Trees

Some results in the literature say CART is better than other interpretable methods like Logistic Regression and Naïve Bayes. CART gives greater classification accuracy compared with logistic regression [5]. Besides that, CART dan Naïve Bayes provides almost the same value, but CART more suitable for large scale data [6].

CART was first proposed by Breiman et al. (1993). The concept of CART is binary recursive partitioning [7]. The purpose of partitioning is to divide the dataset into sections. The term binary gives the meaning that each group is represented by a node in a decision tree, which can only be divided into two groups. Then, each node is called the parent node and can be divided into two child nodes. The purpose of recursive states that the binary partitioning process can be repeated continuously. Therefore, each parent node will produce two child nodes, and each child node will become a parent node and produce child nodes, and so on.

The process of building a decision tree on CART goes through three main stages: build a classification tree, pruning the tree, and choose the optimal tree. The process of build a classification tree consists of splitting, determining the terminal node, and marking the class label (class assignment). The fundamental idea is to select each split of the subset are purer than the parent subset.

The rules for split parent nodes into two child nodes are based on values derived from one independent variable [7]. If the independent variable is continuous, X_j with n sample space and there are n different sample observation values, then there are $n-1$ different sorting. If the nominal categorical independent variable has an L level, so the number of separation is $2^{L-1}-1$. However, if the independent variable is ordinal categorical, we will get $L-1$ split. Splitting in CART based on the impurity function below.

$$g(t)=\sum_{i\neq j} p(j|t)p(i|t), \tag{1}$$

where $g(t)$ is the gini index in node t, $p(i|t)$ is the proportion of i-th class in node t and $p(j|t)$ is the proportion of j-th class in node t.

After the classification tree formed, the next step is tree pruning. Pruning the classification tree is intended to avoid overfitting because of the smaller number of prediction errors due to a large number of splitting. The method used in the tree pruning is minimal cost complexity [7].

Misclassification cost of the tree T in complexity α can be calculated by the sum of misclassification cost of tree T and multiplication of α and $\dot{T}$ (the number of terminal node in tree T)

$$R_\alpha(T)=R(T)+\alpha|\dot{T}|. \tag{2}$$

Cost complexity pruning determines subtree $T(\alpha)$ that minimizes $R_\alpha(T)$ for every α. Complexity α will increase as the pruning process. Then, search the subtree $T(\alpha)<T_{max}$ that minimizes $R_\alpha(T)$.

After a simple sized classification, the optimal tree can be chosen from subtree that minimizes $R_\alpha(T)$. The other option to build an unbias estimator for misclassification cost is $R^*(T_t)$ [7].

B. Synthetic Minority Oversampling Technique Nominal

Synthetic Minority Oversampling Technique (SMOTE) is one of the methods for handling imbalanced data proposed by Chawla et al. (2002). The idea of SMOTE is oversampling minority instances by making synthetic data rather than doing replication [1]. SMOTE is not focusing on all data classes, but only the minority class. This method adds synthetic data to the original dataset, so the proportion is balanced. SMOTE-N is the development of SMOTE, which can be used for a nominal dataset.

If the distance in numerical data is measured by using Euclidean distance, the distance in categorical data is calculated using a modified version of the Value Difference Metric called MVDM [8]. The distance between the two corresponding feature values is explained by the following equation:

$$\delta(V_1,V_2)=\sum_{i=1}^{h}\left|\frac{C_{1i}}{C_1}-\frac{C_{2i}}{C_2}\right|^k, \tag{3}$$

where C_1 is the number of occurrences of V_1 and C_2 is the number of occurrences of V_2. The number of V_1 that have response i-th class is denoted as C_{1i} and V_2 that have response i-th class as C_{2i}. k is a constant, and some researchers often use 1. Then h is the number of response classes in the dataset.

After that, the distance between two feature vectors is given by:

$$\Delta(X,Y)=w_x w_y\sum_{b=1}^{p}\delta(x_b,y_b)^r, \tag{4}$$

where $r=1$ yields the Manhattan distance, and $r=2$ yields the Euclidean distance [9]. For SMOTE-N, we can ignore these weights in (2), as SMOTE-N is not used for classication purposes directly. After that, generate new minority class feature vectors by creating a new set feature value based on k-nearest neighbors. Result of SMOTE-N, the amount of data between classes will be balanced later.

C. Edited Nearest Neighbor

ENN is an undersampling method that uses the nearest neighbors to choose which instance should be removed. An instance considered to be removed if the response variable different from the majority of its nearest neighbors. Distance between two categories calculated using overlap distance as follows:

$$overlap(V_1,V_2)=\begin{cases}0, & \text{if } V_1=V_2,\\ 1, & \text{otherwise.}\end{cases} \tag{5}$$

Then, the distance between the two instances is calculated below:

$$\Delta(X,Y)=\sum_{i=1}^{p} overlap(V_1,V_2). \tag{6}$$

D. Adaptive Synthetic Nominal

After SMOTE-N, He et al. propose ADASYN for handling imbalance by giving weight in minority class, but only for numerical predictors. Concept of ADASYN is to give weight to minority instances. Synthetic data from minority instances that are difficult to learn will be generated more than minority instances that are easy to learn. Some instances called to be difficult to learn when minority instances located in the majority instances area. As an improvement of ADASYN that only for numerical data, there is ADASYN-N (Adaptive Synthetic Nominal) for nominal data, proposed by Kurniawati (2017). Distance calculation in ADASYN-N using Modified Value Difference Metric (VDM) by Cost and Salzberg, is the same with SMOTE-N. The steps in ADASYN-N are as follows.

Training dataset with n sample $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$, where x_i is data in p dimensional feature space X and $y_s \in Y = \{1, \ldots, C\}$ is class label with biggest amount. Then data is separated into m_s and m_l, number of instances in minority and majority class. Therefore $m_s \leq m_l$ and $m_s + m_l = n$. First of all, we calculate the number of synthetic data that must be generated based on level balance.

$$G = (m_l - m_s) \times \beta \tag{7}$$

where $\beta \in [0, 1]$ is a parameter used in determining the desired level of balance. If $\beta = 1$, that dataset is fully balanced.

For every x_i in minority class, then determine k-nearest neighbors in p dimensional space, and calculate r_i as ratio of majority domination in k-nearest neighbors

$$r_i = \frac{H_i}{k}, \quad i = 1, \ldots, m_s \tag{8}$$

where, H_i is the number of majority instances in nearest neighbor.

When r_i is higher, the more majority instance in nearest neighbors and the more difficult to learn. After that, normalized r_i, so the sum of normalized $\hat{r}_i$ is 1

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{9}$$

where

$$\sum_{i=1}^{m_s} r_i = 1 \tag{10}$$

The final step is calculate number of synthetic data that need to be generated for every minority instance.

$$g_i = \hat{r}_i \times G \tag{11}$$

Data generation is done by replicating x_i (minority instance i) as much as g_i.

III. Results and Discussions

In this section, we present the application of handling imbalance in SKAP data Jawa Timur Province in 2018. SKAP or Survei Kinerja dan Akuntabilitas KKBPK (Kependudukan Keluarga Berencana dan Pembangunan Keluarga) is survei data by BKKBN Indonesia for teenagers. Respond of the data is binary, "0" is no and "1" is yes for have had premarital sex.

TABLE I: Class Distribution

Category	Count
No	1971
Yes	44

There are 15 predictors or independent variable that allegedly effect teenage sexual behaviour. That variable includes social demographic, knowledge, and risk behaviour. Number of instances in data are 695 with weight. But if we unweight the data, the number of instances becomes 2015, and that data will be used for this study. Composition of class responds is presented in Table I.

From Table I, we can say that the amount of teenagers who had and had not premarital sex imbalance. If we continue to classify with imbalanced data, then variables that can predict adolescent having premarital sex are dating style, sex, education of the risk of marrying at young age, and economic status. Average classification accuracy (AUC) produced with imbalanced data only 0.7373, which is not good enough.

For classification, imbalance can cause misleading, because all class predict to majority class. So, it is very important for handling imbalance first before classification. In this study, imbalance data will be handled by SMOTE-N-ENN (hybrid Synthetic Minority Oversampling Technique Nominal and Edited Nearest Neighbor) and Adaptive Synthetic Nominal. Handling imbalance only doing in training dataset, or only use to build model.

A. Hybrid SMOTE-N-ENN

Concept of hybrid SMOTE-N-ENN is oversampling using SMOTE-N and continue with undersampling using ENN. As such, SMOTE-N can oversample minority class (adolescent who had premarital sex):

1) Calulate VDM distance between all instance in minority class.
2) Determine k (number of nearest neighbor), in this study we use 10.
3) Randomly choose one instance in minority class.
4) Determine 10 nearest neighbors by order the distance between choosen instance and all minority instances.
5) Synthetic data are created by determine the value of each independent variable. That value obtained from majority voting from 10 nearest neighbors.
6) Repeat step 3 to 6 until the number of instances in minority class balance with majority class.

After training dataset become balance, the next step is reduce the instance (undersampling) by removing noise data. Instance is called to be noise if the response class different with k-nearest neighbors. So, if one instance has a response that had premarital sex, but majority of k-nearest neighbors never had premarital sex, so that instance is assumed to be noise. Different from SMOTE-N, distance used in ENN is overlap distance.

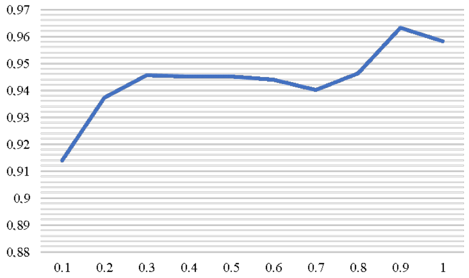

Fig. 1: Classification Accuracy Based on Level of Balance.

B. ADASYN-N

Similar with SMOTE-N, ADASYN-N also oversample training dataset until the proportion between two class balance. Before new training dataset in ADASYN will be build, the amount of synthetic dataset should be determined based on β. But, there is no justification about best β in ADASYN. He et al. compare β from 0 to 1, and say that $\beta = 1$ gave minimum error in their dataset. In this study, we will compare classification accuracy of CART using many level of balance in ADASYN-N.

Figure 1 shows level of balance versus AUC with classifier CART in 5-fold cross validation. There is no specific pattern from classification accuracy in the level of balance. At some point, give high AUC (Area Under ROC-curve), then decrease to low AUC. The graphic shows that AUC increases between level of balance 0.1 until 0.3, then start to decrease until 0.7, and then increase again. Overall, the best balance level due to highest AUC is 0.9, so classification accuracy (AUC) with $\beta = 0.9$ is the best ADASYN-N.

C. Method Comparison

The new dataset produced by ADASYN-N and hybrid SMOTE-N-ENN then tested using Classification and Regression Trees (CART). Implementation with the classifier is done using 5-fold cross validation, which is dividing the dataset into 5 parts, where 4 parts will be training dataset, and one part will be testing dataset.

Training dataset will be modified by handling imbalance. With SMOTE-N, the proportion between two class will balance. Meanwhile in ADASYN-N, not too balance. For example, in fold 1 there are 1577 instances majority class and 36 instances minority class. So, according to (7), the amount of instances should be generated in fold 1 is 1386.9 or 1387.

Using 5-fold cross validation, average AUC of classification using CART will be compared between real dataset with new dataset result of SMOTE-N, hybrid SMOTE-N-ENN, and ADASYN-N. That comparison result is presented in Table II.

From Table II, the best method for dealing with imbalanced data is Adaptive Synthetic Nominal (ADASYN-N). This can be seen from the average AUC on the ADASYN-N method which is higher than the other two methods. The average AUC testing CART with the handling imbalance of ADASYN-N was 0.963266.

TABLE II: Classification Accuracy Testing Dataset

	Average AUC
Imbalance Dataset	0.7373
SMOTE-N	0.916944
SMOTE-N-ENN	0.917808
Best ADASYN-N	0.963266

TABLE III: Classification Accyracy Every Fold in ADASYN-N

Fold	**Average AUC**
1	0.97335
2	0.927947
3	0.965736
4	0.977215
5	0.972081

TABLE IV: Classification Accyracy Every Fold in ADASYN-N

Actual	**Average AUC**	
	No	**Yes**
No	377	18
Yes	0	9

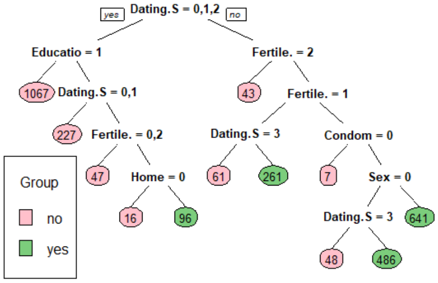

Fig. 2: Best Decision Tree with Best ADASYN-N.

D. Best Method

Because of k-fold cross validation, the number of trees formed with ADASYN-N is 5, so if we want to know which is the best tree weill be interpreted, we should compare all fold's classification accuracy. Here are average of area under roc-curve every fold in ADASYN-N.

Table III says that tree with highest classification accuracy AUC is in fold 4. Confusion matrix in testing data fold 4 ADASYN-N is presented below.

From confusion matrix, we know that there is a missclassification, but all adolescent who had premarital sex classified correctly. It is very important to minimize missclassification, especially in adolescent who had premarital sex. While missclassification in adolescent which had not premarital sex could be a preventive suggestion. Decision tree with best ADASYN-N shown in Fig. 2. Factors that can predict premarital sex among adolescent are dating style, knowledge about fertility, condom usage, sex, latest education, and place of residence.

IV. Conclusions

In this study, we have studied the comparison of ADASYN-N and hybrid SMOTE-N-ENN in highly imbalanced data (SKAP BKKBN, 2018). ADASYN-N with level of balance 0.9 gives the best average AUC compared with SMOTE or hybrid SMOTE-N-ENN. The model is considered to be better because the classification accuracy increases quite high, from 0.7373 to 0.963266. In addition, the factors that appear or can predict premarital sex more reasonable.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[2] S. Vluymans, N. Verbiest, C. Cornelis, and Y. Saeys, “Instance selection for imbalanced data,” in *WorkshopRough Sets: Theory and Applications(RST&A); held at the 2014 Joint Rough Set symposium (JRS 2014)*, 2014.

[3] H. Haibo, B. Yang, G. A. Edwardo, and L. Shutao, “Adaptive synthetic sampling approach for imbalanced learning,” in *IEEE International Joint Conference on Neural Networks, IJCNN*, vol. 8, no. 3, 2016, pp. 1322–1328.

[4] S. Rahayu, T. Adji, and N. Setiawan, “Analisis perbandingan metode oversampling adaptive synthetic-nominal (adasyn-n) dan adaptive synthetic-knn (adsyn-knn) untuk data dengan fitur nominal-multi categories,” 2017.

[5] M. Adiansyah, “Perbandingan metode cart dan analisis regresi logistik serta penerapannya untuk klasifikasi ketertinggalan kabupaten dan kota di indonesia,” Ph.D. dissertation, Institut Pertanian Bogor, 2017.

[6] D. Jeyarani, G. Anushya, R. Rajeswari, and A. Pethalakshmi, “A comparative study of decision tree and naive bayesian classifiers on medical datasets,” *International Journal of Computer Applications*, vol. 975, p. 8887, 2013.

[7] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and regression trees*. CRC press, 1984.

[8] K. Fithriasari, S. Pangastuti, N. Iriawan, and W. Suryaningtyas, “Classification boosting in imbalanced data,” *MJS*, vol. 38, no. Sp2, pp. 36–45, 2019.

[9] S. Cost and S. Salzberg, “A weighted nearest neighbor algorithm for learning with symbolic features,” *Machine learning*, vol. 10, no. 1, pp. 57–78, 1993.