K-Neighbor Over-Sampling with Cleaning Data: A New Approach to Improve Classification Performance in Data Sets with Class Imbalance

Budi Santoso
Department of Statistics, Bogor Agricultural University, Indonesia
Statistics-Indonesia, Jakarta, Indonesia

Hari Wijayanto, Khairil Anwar Notodiputro and Bagus Sartono*
Department of Statistics, Bogor Agricultural University, Indonesia
*Corresponding author

Copyright © 2018 Budi Santoso, Hari Wijayanto, Khairil Anwar Notodiputro and Bagus Sartono. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The problems of class imbalance have attracted concerns of researchers in the last few years. Class imbalance problems occur when the data has unbalanced proportions between groups or levels of the response variable. The class/group with a dominant proportion is called as the majority class, while the small proportion class is known as the minority. These problems relate to the creation of bias in parameter estimation in a parametric model such as logistic regression and high misclassification rate for the minority class. Eventually, it might create risks in the policymaking. To overcome these problems, a new approach called K-Neighbor Over-sampling (KNOS) with cleaning data is proposed in this paper. Unlike the other methods such as SMOTE, Border-line SMOTE (BLS), and Safe Level-SMOTE (SLS), in generating synthetic data, KNOS employs K minority class observations, then KNOS proceeds further by removing some majority class observations. Notice that, similar to BLS and SLS, KNOS only generates synthetic samples from original samples which are safe. In this paper KNOS has been applied to logistic regression classification method. The results showed that KNOS method produced a higher performance in terms of AUC,
G-Mean, and sensitivity compared to BLS and SLS. Moreover, our study has also shown that KNOS produced more consistent result than the other approaches.

**Keywords**: class imbalance, K neighbor over-sampling, logistic regression

1. Introduction

Class imbalance is a kind of data which have unbalanced proportions between two or more groups of data which are usually called as minority and majority classes. The problem is usually found in the real cases such as telephone fraud [6], cases of health diagnosis and pollution detection [5]. In Indonesia, there are many cases of class imbalance which are strategic data development such as data on drop-out rates, illiteracy rates, unemployment and percentage of poor people [1].

The incidence of class imbalance can cause statistical problems such as the occurrence of biases in parameter estimates [9], [10], [11] and inaccuracy prediction in classification especially in minority classes [4], [7], [3], [12]. In various literatures, the problem of class imbalance usually associated with misclassification issues in which the accuracy of classification in minority data will be lower than in the majority class.

There are two approaches that can be used to solve class imbalance problems. They are the solution at data level and solution at algorithm level [12]. The solution at data level is applied by balancing the distribution of the majority and minority class through methods of over-sampling and under-sampling or their combination. The solution at algorithm level is applied by modifying classifier methods or by optimizing the performance of learning algorithm. The advantage of the data level solution is the possibility to select classifier methods independently.

Over-sampling approach is used more frequently than under sampling since the under-sampling method will eliminate data in the majority class which consequently caused loss of important information of the data. According to Batista et.al [2] in general, the synthetic over-sampling method gives better results than under-sampling method. The over-sampling method is more effective than under-sampling, especially on data that has a high imbalance.

The famous over-sampling method is the Synthetic Minority Over-Sampling Technique (SMOTE) [4]. In SMOTE method, the generated synthetic data exists between minority and one of the K nearest neighbors (number of K set to 5). However, the SMOTE method still has some weaknesses that the data synthetic resulted by SMOTE are still possible to spread between minority and majority data. Hence it may reduce the performance of classification. To solve the SMOTE problems, some researchers are conducted to modify SMOTE in order to create a more effective technique for improving the classification performance. The development of SMOTE is mostly done in two ways, the first is to improve the synthetic data results of SMOTE by cleaning data such as SMOTE with Edited Nearest Neighbor (ENN) [2] and SMOTE-Rough Set [12]. The second way is to determine which location will be chosen for data generating as it is expected that the result of data generating will exactly found in the minority data area such as
Borderline-SMOTE (BLS) [7] and Safe level-SMOTE (SLS) [3]. In general, those modification methods have not provided the best result yet.

This paper proposes a different approach to handling class imbalance. This study will use many K nearest neighbors as the basis of synthetic data generating. In addition, this research combines minority data selection and data cleaning approaches in order to produce more optimal classification accuracy.

2. K Neighbor Over-sampling with Cleaning Data (KNOS)

Many over-sampling methods such as SMOTE, BLS, and SLS use one nearest neighbor from K of nearby neighbors as the basis for data generation. Through this process, the result of synthetic data exists between the minority data and its selected nearest neighbor. By selecting only one of K nearest neighbor, it can yield more varied synthetic data and even overgeneralization problems. The use of 1 nearest neighbor is also done by Siriseriwan but with a different approach. Siriseriwan [13] first selects the nearest neighbor based radius so that the number of K nearest neighbor is different. Our method, KNOS is a kind of synthetic data generation which use all of K nearest neighbor from minority class so that the resulting synthetic data is between the K minority data. The result will get more homogeneous synthetic data. Figure 1 show comparison between SMOTE and KNOS using 5 nearest neighbors.

![Figure 1. Comparison SMOTE and KNOS with 5 nearest neighbor](image)

Another difference between the KNOS method and the other method is the use of its nearest neighbor. In the SMOTE, BLS and SLS method, the use of nearest neighbors is only based on minority class so that the minority data being its nearest neighbors may not be the closest possible data. Meanwhile, in the KNOS method, the nearest neighbor concept is based on the distance between the two classes and then the nearest neighbor is chosen which is the minority data. So
the minority data that became its nearest neighbor is the minority data that has the closest distance. In this way, the synthetic data generated is in the safe location. After the data generation, then do the process of cleaning data using adjusted Edited Nearest Neighbor (ENN). The process of cleaning data performed Batista [2] using 3 nearest neighbors, while the KNOS method using the 5 nearest neighbors as the basis of cleaning the data, so it is expected that will provide in-depth data cleaning.

The formula to generate synthetic data by SMOTE with 1 neighbor can be expressed as:

$$D_{new} = D_i + (D_i - D_{\bar{i}}) \times \delta$$

(1)

where:

- $D_i \in S_{\text{min}}$ = minority class example
- $D_{\bar{i}}$ = one of K-nearest neighbor for $D_i$ where $D_{\bar{1}} \in S_{\text{min}}$
- $\delta \in [0,1]$ = random number

the equation (1) can be expressed as:

$$D_{new} = D_i + (D_i \delta - D_i \delta) = D_i - D_i \delta + D_i \delta = (1 - \delta)D_i + \delta D_i$$

(2)

If there are only 2 nearest neighbors then:

$$D_{new:2} = \delta D_i + \delta_1 D_1 + \delta_2 D_2$$

$$D_{new:2} = (1 - \delta_1 - \delta_2) D_i + \delta_1 D_1 + \delta_2 D_2$$

where $\delta = 1 - \delta_1 - \delta_2$

$$D_{new:2} = D_i - \delta_1 D_1 - \delta_2 D_2 + \delta_1 D_1 + \delta_2 D_2$$

$$D_{new:2} = D_i + (D_i - D_1) \delta_1 + (D_2 - D_i) \delta_2$$

(3)

If there are numbers of K nearest neighbors then:

$$D_{new:k} = D_i + (D_i - D_1) \delta_1 + (D_2 - D_i) \delta_2 + \ldots + (D_k - D_i) \delta_k$$

$$D_{new:k} = D_i + \sum_{j=1}^{k} (D_k - D_i) \delta_k$$

(4)

where $\delta, \delta_1, \delta_2, \ldots \delta_k$ is a random number with $\sum_{i=1}^{k} \delta_i = 1$ then:

- $\delta \in [0,1]$
- $\delta_1 \in [0,1 - \delta]$
- $\delta_2 \in [0,1 - \delta - \delta_1]$

...
This model (equation 4) is the general form of over-sampling with K nearest neighbors. In KNOS, we generate a synthetic observation located inside the convex space formed by K minority point. In another hand, SMOTE and related over-sampling with SMOTE generate a synthetic observation at any point along a line connecting two minority points.

Here are the steps of the KNOS method:

a. Selecting the minority observation
   - Set number of K
   - Calculate K nearest neighbors based on the minority and majority observation with Euclidian distance.
   - Selecting minority observation for which the K nearest neighbor is only minority data.

b. Generating a synthetic data by using all of K nearest neighbors correspondings to equation (4) to selected minority observation. If we have data where the majority data is M while the minority data is m. We will generate minority data so that the minority data becomes m*, then we need synthetic data as m*-m.

c. Removing or cleaning noisy majority class observation
   - For each majority class observation, identifying 5 nearest neighbors
   - Remove observation which has all 5 neighbor are minority class

3. Logistic Regression

Logistic regression (LR) is a method used to understand the relation between the categorical response of either binary or multinomial with independent variables in the form of categorical and numerical scales. LR can be used to classify an observation into a positive or negative class. The logistic regression model with a response variable Y and n predictor variable \( x_1, x_2, ..., x_n \) can be formulated by [8]:

\[
E(Y | x) = \pi_i = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} = \frac{e^{(x_\beta)}}{1 + e^{(x_\beta)}}
\]

The link function for binomial data using logit is:

\[
\log it(\pi_i) = g(x) = \ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = X \beta
\]

The method of parameter estimation commonly used in logistic regression is the MLE (Maximum Likelihood Estimation) method. The log of likelihood function can be expressed as:

\[
\ln L(\beta) = \sum_{i=1}^{n} \left( y_i \ln \left( \frac{e^{(x_\beta)}}{1 + e^{(x_\beta)}} \right) + (1 - y_i) \ln \left( 1 - \frac{e^{(x_\beta)}}{1 + e^{(x_\beta)}} \right) \right)
\]
To get parameter coefficient, we compute from the first derivative from \( \ln L(\beta) \) for each \( \beta_0, \beta_1, \beta_2, \ldots, \beta_k \) and then the derivative must be zero.

After the value \( \beta \) obtained, then for the prediction value of \( Y \), we define a cut point value (0.5) to compare each estimated probability to 0.5 as follows:

\[
Y = \begin{cases} 
1; & \pi_i \geq 0.5 \\
0; & \pi_i < 0.5 
\end{cases}
\]

4. Performance Measurements

Evaluation of performance of classification is measured by confusion matrix. Table 1 showed the confusion matrix which contains information about actual and prediction of positive class and negative class where negative indicate majority class and positive indicate minority class.

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Class</td>
<td>Negative</td>
<td>TN: True Negative</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>FP: False Positive</td>
</tr>
</tbody>
</table>

The accuracy, sensitivity, specificity and G-Mean of Table 1 can be calculated as follows:

\[
\text{Accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{8}
\]

\[
\text{Sensitivity} = TP \text{ rate} = \frac{TP}{(TP + FN)} \tag{9}
\]

\[
\text{Specificity} = TN \text{ rate} = \frac{TN}{(TN + FP)} \tag{10}
\]

\[
G - \text{Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{11}
\]

Using accuracy to evaluate the measures of classification on class imbalance is not enough to show the goodness of classification. Accuracy only indicates the precision of classification on the composite of both majority and minority data classes. Performance measures such as sensitivity, specificity, Area under the curve of Receiver Operator Characteristic (AUC) and Geometric Mean (G-Mean) are more appropriate for class imbalance because they are more specific in predicting each class. AUC value is between 0.5 to 1 where the value close to 1 means the accuracy of classification is very good or can distinguish the class very well [8].
5. Result of Study

5.1. Simulation Study

The simulation data used two predictor variables \((X_1 \text{ and } X_2)\) and response variable \((Y)\). The predictor variables are generated using a standard normal distribution \((X_1 \sim N(0,1), X_2 \sim N(0,1))\). While response variables are generated using the Bernoulli distribution where the \(Y\) value follows the distribution:

\[
Y \sim (1, \pi), \quad \pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}
\]  

(12)

The data generation process is performed on different imbalance ratio (IR) with fixed parameter. The simulation data process is applied by \(n=4000\) with \(1000\) replication. Complete can be seen in Table 2 below:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Phi</th>
<th>IR</th>
<th>Linear predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data1</td>
<td>0.2</td>
<td>80:20</td>
<td>(-3 + 2X_1 + 3X_2)</td>
</tr>
<tr>
<td>Data2</td>
<td>0.1</td>
<td>90:10</td>
<td>(-5 + 2X_1 + 3X_2)</td>
</tr>
<tr>
<td>Data3</td>
<td>0.05</td>
<td>95:5</td>
<td>(-7 + 2X_1 + 3X_2)</td>
</tr>
<tr>
<td>Data4</td>
<td>0.01</td>
<td>99:1</td>
<td>(-9 + 2X_1 + 3X_2)</td>
</tr>
</tbody>
</table>

Figure 2. Boxplot of AUC values with various phi
Figure 3. Boxplot of Sensitivity values with various phi

Table 3. The Summary of Statistics for AUC and Sensitivity values by Over-sampling Method on Simulation Dataset

<table>
<thead>
<tr>
<th>Phi</th>
<th>Methods</th>
<th>AUC</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Mean</td>
</tr>
<tr>
<td>0.2</td>
<td>Original</td>
<td>0.784</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.823</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.824</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td><strong>0.825</strong></td>
<td><strong>0.865</strong></td>
</tr>
<tr>
<td>0.1</td>
<td>Original</td>
<td>0.707</td>
<td>0.787</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.809</td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.813</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td><strong>0.833</strong></td>
<td><strong>0.883</strong></td>
</tr>
<tr>
<td>0.05</td>
<td>Original</td>
<td>0.611</td>
<td>0.729</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.711</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td><strong>0.774</strong></td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.767</td>
<td><strong>0.887</strong></td>
</tr>
<tr>
<td>0.01</td>
<td>Original</td>
<td>0.500</td>
<td>0.672</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.643</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.607</td>
<td>0.869</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td><strong>0.661</strong></td>
<td><strong>0.870</strong></td>
</tr>
</tbody>
</table>

*bold print represents the better value
Figures 2 and 3 show the average AUC boxplot value and average sensitivity on different phi in the original data and use the over-sampling method of BLS, SLS, and KNOS. In the original data shows that the value of AUC and sensitivity decreases as the data becomes more unbalanced. Even on data with small phi value (phi≤0.05), the sensitivity value is below 0.5. This shows the accuracy of classification especially minority data on unbalanced data decreases. In addition, the distribution of AUC values and sensitivity has a larger range of unbalanced data. Based on over-sampling method, KNOS method has better performance that is AUC value and sensitivity is higher than another method. This happens at different phi values.

Table 3 present summary statistics of AUC and sensitivity based on the over-sampling method on phi variation. In general, the KNOS method tends to perform better when compared to the BLS and SLS with shorter range values and smaller standard deviation.

5.2. Experimental Study

In application, we use five quantitative data sets. Four data sets from KEEL (http://www.keel.es/datasets.php) and UCI Repository of Machine Learning Databases (http://archive.isc.edu/ml/datasets.html): Ecoli4, Glass2, Haberman’s Survival and Yeast2. The poor households data set based on the results of the National Socio-Economic Survey (Susenas) in Yogyakarta Province in Indonesia. Susenas has been conducted by Statistics-Indonesia (BPS) annually to produce the social-economic indicators. The first to last column of Table 4 represents the data sets name, the number of attributes, the number of examples and imbalance ratio (IR) which ratio of majority class compare to a minority class.

<table>
<thead>
<tr>
<th>Id</th>
<th>Data set</th>
<th># Attributes</th>
<th># Examples</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ecoli4</td>
<td>9</td>
<td>214</td>
<td>11.59</td>
</tr>
<tr>
<td>2</td>
<td>Glass2</td>
<td>7</td>
<td>336</td>
<td>15.8</td>
</tr>
<tr>
<td>3</td>
<td>Haberman’s survival</td>
<td>8</td>
<td>514</td>
<td>9.08</td>
</tr>
<tr>
<td>4</td>
<td>Yeast2</td>
<td>3</td>
<td>306</td>
<td>2.78</td>
</tr>
<tr>
<td>5</td>
<td>Poor Households</td>
<td>2</td>
<td>3662</td>
<td>6.60</td>
</tr>
</tbody>
</table>

The application of the over-sampling method is done on the Ecoli4, Glass2, Haberman's survival, Yeast2 and Poor households data sets which have different IR values. Table 5 shows that all of over-sampling approaches outperforms the results with the original data-sets as expected. KNOS methods have better performance when compared to the BLS and SLS based on sensitivity, AUC values and G-Mean in various data sets.
Tabel 5. The Performance of Over-sampling Method with Application Dataset

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Methods</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>G-Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli4</td>
<td>Original</td>
<td>0.9770</td>
<td>0.8457</td>
<td>0.9851</td>
<td>0.9154</td>
<td>0.9128</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.9760</td>
<td>0.8531</td>
<td>0.9831</td>
<td>0.9181</td>
<td>0.9158</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.9680</td>
<td>0.8601</td>
<td>0.9754</td>
<td>0.9178</td>
<td>0.9160</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.9550</td>
<td><strong>0.8790</strong></td>
<td>0.9588</td>
<td><strong>0.9189</strong></td>
<td><strong>0.9181</strong></td>
</tr>
<tr>
<td>Glass2</td>
<td>Original</td>
<td><strong>0.9313</strong></td>
<td>0.0000</td>
<td><strong>0.9950</strong></td>
<td>0.5025</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.6203</td>
<td>0.4824</td>
<td>0.6285</td>
<td>0.6215</td>
<td>0.5506</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.7203</td>
<td>0.3436</td>
<td>0.7542</td>
<td>0.6131</td>
<td>0.5091</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.6892</td>
<td><strong>0.6884</strong></td>
<td>0.6941</td>
<td><strong>0.6912</strong></td>
<td><strong>0.6912</strong></td>
</tr>
<tr>
<td>Haberman</td>
<td>Original</td>
<td>0.7176</td>
<td>0.1382</td>
<td><strong>0.9612</strong></td>
<td>0.5497</td>
<td>0.3645</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.7330</td>
<td>0.4428</td>
<td>0.8256</td>
<td><strong>0.6342</strong></td>
<td><strong>0.6046</strong></td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td><strong>0.7440</strong></td>
<td>0.3730</td>
<td>0.8900</td>
<td>0.6315</td>
<td>0.5762</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.6132</td>
<td><strong>0.6489</strong></td>
<td>0.6028</td>
<td>0.6258</td>
<td><strong>0.6254</strong></td>
</tr>
<tr>
<td>Yeast2</td>
<td>Original</td>
<td><strong>0.9435</strong></td>
<td>0.6311</td>
<td><strong>0.9783</strong></td>
<td>0.8047</td>
<td>0.7857</td>
</tr>
<tr>
<td></td>
<td>BLS</td>
<td>0.9221</td>
<td>0.7647</td>
<td>0.9416</td>
<td>0.8532</td>
<td>0.8486</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.9325</td>
<td>0.7612</td>
<td>0.9508</td>
<td>0.8560</td>
<td>0.8508</td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.9344</td>
<td><strong>0.8012</strong></td>
<td>0.9474</td>
<td><strong>0.8743</strong></td>
<td><strong>0.8712</strong></td>
</tr>
<tr>
<td>Poor House</td>
<td>Original</td>
<td><strong>0.8652</strong></td>
<td>0.0000</td>
<td><strong>1.0000</strong></td>
<td>0.5000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Household</td>
<td>BLS</td>
<td>0.6533</td>
<td>0.6651</td>
<td>0.6515</td>
<td>0.6580</td>
<td>0.6582</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.6566</td>
<td>0.6622</td>
<td>0.6558</td>
<td>0.6590</td>
<td><strong>0.6590</strong></td>
</tr>
<tr>
<td></td>
<td>KNOS</td>
<td>0.6056</td>
<td><strong>0.7365</strong></td>
<td>0.5853</td>
<td><strong>0.6609</strong></td>
<td>0.6565</td>
</tr>
</tbody>
</table>

6. Conclusion

There are many approaches for handling class imbalance problems but still unsatisfactory. KNOS method can be used as an alternative method to handle class imbalance problem. The results of data simulation and data application show that KNOS method can give a better result from BLS and SLS method based on AUC, G-Mean, and sensitivity performance measures. The results of the KNOS method are able to provide a more consistent based on smaller range and standard deviation values.

Acknowledgements. The authors would like to thank the Statistics-Indonesia for providing funds for this research as a part of the Ph.D. grant research program.

References


K-neighbors over-sampling with cleaning data

https://doi.org/10.1145/1007730.1007735


https://doi.org/10.1002/0471722146

https://doi.org/10.1093/oxfordjournals.pan.a004868


https://doi.org/10.1007/s10115-011-0465-6


**Received: February 26, 2018; Published: April 10, 2018**