A Novel Oversampling Method Based on SMOTE and Local Sets for Imbalanced Classification

Fanyao Dai\textsuperscript{1,2} and Yi Liu\textsuperscript{3}

\textsuperscript{1}Postdoctoral Station of Chongqing Academy of Economics Research, Chongqing 401147, China
\textsuperscript{2}Postdoctoral Station of School of Resources and Safety Engineering, Chongqing University, Chongqing 400044, China
\textsuperscript{3}China Merchants Vehicle Reserch, Chongqing 401329, China
Email: dfy_cq@126.com; cjliuyi@cmhk.com

Abstract. Learning a classifier from imbalanced data continues to be a challenging issue. The oversampling methods can improve the imbalanced classification from the perspective of data preprocessing. Different oversampling methods have been proposed. Nevertheless, most tend to generate unnecessary noise, create redundant synthetic samples in the class center and heavily rely on the parameter $k$. To solve the above issues, this work presents an oversampling method based on local sets and SMOTE (LS-SMOTE). First, the local sets are searched to describe the local characteristic of imbalanced data. Second, a local sets-based noise filter is designed to remove noise and smooth the class boundary. Finally, on each local set, the interpolation of SMOTE between a base sample and a selected sample closest to the majority class is employed to create the synthetic samples. Experimental results with 12 real data sets have proved that LS-SMOTE outperforms representative oversampling methods in training $k$ nearest neighbor classifier.

Keyword. Imbalanced learning; imbalanced classification; oversampling; nearest neighbors; local sets.

1. Introduction

Supervised classification has been widely applied in data mining, artificial intelligence, machine learning, etc. The traditional classifier is trained on data sets with the class-balanced distribution. However, the data distribution is often skewed in many practical situations, such as image processing, gene recognition, natural language processing, etc. In skewed data, the samples’ number of positive cases (minority classes) is much smaller than that of negative cases (majority classes). Hence, minority classes are difficult to be classified correctly but are often the most interesting from the perspective of the application. To improve the imbalanced classification, oversampling methods \cite{1} have been developed and attracted the attention of scholars recently.

The oversampling methods improve imbalanced classification by creating synthetic samples to extend the minority class. Resampling \cite{2} is the classical oversampling method. It amplifies the minority class by copying original samples until the class distribution is balanced. Nevertheless, resampling will cause over-fitting due to many duplicate samples. To improve generalization, the synthetic minority over-sampling technique (SMOTE) \cite{3} creates synthetic examples by employing the interpolation among $k$ nearest minority class neighbors.
So far, SMOTE has received extensive attention. Besides, it has many practical applications, such as recognition of oil and gas, risk evaluation of the pressure for oil and gas, geological hazard evaluation, etc. [4-6]. Because of the advanced nature, many oversampling methods are developed recently and are based on the idea of SMOTE. Examples are Borderline-SMOTE [7], ADASYN [8], Safe-Level-SMOTE [9], k-means SMOTE [10] RSMOTE [11], etc. Despite the effectiveness, most have been proved to have the following limitations:

(a) Most oversampling methods tend to generate unnecessary noise. The interpolation between noise or (and) unsafe borderline samples is used to generate synthetic samples, which amplifies the effect of noise.

(b) Most oversampling methods may create redundant synthetic samples on the class center. Although samples of the class center can maintain the basic distribution of data, they have little contribution to classification performance.

(c) Almost all oversampling methods heavily rely on the neighbor’s parameter $k$. This because they use the interpolation among $k$ nearest neighbors to create synthetic samples. It is difficult to choose an appropriate $k$ value in the different data distribution.

These above challenges have not been overcome simultaneously in oversampling methods so far, to the best of our knowledge.

In order to overcome the above issues, this work presents a method of oversampling based on local sets and SMOTE (LS-SMOTE). In LS-SMOTE, the local sets are searched to describe the local characteristic of imbalanced data. Second, a local sets-based noise filter is designed to remove noise and smooth the class boundary. Finally, on each local set, the interpolation of SMOTE between a base sample and a selected sample closest to the majority class is used to generate the synthetic minority class samples. In experiments, $k$ nearest neighbors classifier (KNN) is employed to validate the LS-SMOTE by comparing 5 oversampling methods on 12 real data sets.

2. Related Work

SMOTE is an important oversampling method proposed by Chawla et al. [3]. SMOTE creates synthetic data by interpolation among $k$ nearest neighbors. However, SMOTE is easy to generate noise due to the interpolation among noise or unsafe borderline samples. Hence, SMOTE-TL and SMOTE-ENN [12] proposed to use the Tomek Link (TL) and the Edited Nearest Neighbor Rule (ENN) to filter out noise in SMOTE. Additionally, the Iterative Partition Filter (IPF) [13] was also used to handle noise in SMOTE-IPF.

Borderline-SMOTE [7] and ADASYN [8] are two improvements of SMOTE. They use $k$ nearest neighbors to creates synthetic samples on the class boundary. However, Borderline-SMOTE and ADASYN are susceptible to noise and heavily rely on the parameter $k$.

Safe-Level-SMOTE [9] and RSMOTE [11] create synthetic minority class samples in the class center, which strengthens the distributional characteristics of original data. Specifically, Safe-Level-SMOTE divides the minority into the noisy region, borderline region and safe region. Then, it just creates synthetic minority class samples in the safe region. RSMOTE introduces relative density and 2-means clustering algorithm to generate synthetic samples in the safe areas. However, Safe-Level-SMOTE and RSMOTE may create redundant synthetic samples in the class center. Synthetic samples in the class center have little contribution to classification performance. Besides, they are also susceptible to noise and heavily rely on the parameter $k$.

Georgios et al. [10] proposed $k$-means SMOTE. $k$-means SMOTE employs the $k$-means clustering algorithm to divide the entire data into $k$ parts. Next, it calculates the sparsity of $k$ parts and creates more synthetic samples in the sparser regions. However, $k$-means SMOTE relies on 5 parameters (including the parameter $k$) and can not detect noise in the original data.

While most existing oversampling methods manage to combat one or two weaknesses, none have been shown to overcome the generation of noise, avoid redundant synthetic samples in the class center and solve the choice of the parameter $k$. To fill this gap, LS-SMOTE is proposed.
3. Local Sets
The idea of local sets (LS) is based on the Nearest Enemy (NE). Specifically, the LS of sample $x_i$ is the set of samples whose distance to $x_i$ is smaller than the distance between $x_i$ and its nearest neighbor from a different class. The nearest enemy $NE(x_i)$ of $x_i$ is defined as follows:

**Definition 1.** (Nearest Enemy): The Nearest Enemy (NE) of sample $x_i$ is its nearest neighbor from a different class. We denote the NE of $x_i$ as $NE(x_i)$.

Based on the definition of NE, the LS and LSC are defined as followed:

**Definition 2.** (Local Set): The Local Set (LS) $LS(x_i)$ of $x_i$ is the set of samples whose distance to $x_i$ is smaller than the distance between $x_i$ and $NE(x_i)$.

$$LS(x_i) = \{ x_j : \text{dist}(x_i, x_j) < \text{dist}(x_i, NE(x_i)) \} \quad (1)$$

In equation (2), $\text{dist}(x_i, x_j)$ or $\text{dist}(x_i, NE(x_i))$ represents the Euclidean distance between $x_i$ and $x_j$ (or $NE(x_i)$).

**Definition 3.** (Local Set cardinality): The Local Set Cardinality (LSC) $LSC(x_i)$ of sample $x_i$ is the number of instances contained in $LS(x_i)$.

$$LSC(x_i) = | \{ x_j : x_j \in LS(x_i) \} | \quad (2)$$

In equation (2), $|$ denotes the number. Figure 1 shows LS on toy data with two classes and different colors. $NE(A) = F$, $NE(C) = G$, $NE(B) = H$, $NE(E) = I$. $LS(A) = \{A, B, D\}$. $LS(B) = \{A, B, C, D\}$. $LS(C) = \{C\}$. $LS(E) = \{E\}$. The LS has the following properties:

(a) Samples have the same class label in a given LS.
(b) If sample $x_i$ is closer to the other classes, the sample $x_i$ will have a smaller value of $LSC(x_i)$, and vice versa. In figure 1, $LSC(A) = 3$, $LSC(B) = 4$ and $LSC(C) = 1$. Sample C is closer to the class boundary.
(c) If more LS contain samples $x_i$, the sample $x_i$ will be safer. In figure 1, noisy sample E is only contained in $LS(E)$. Safer sample C is contained in $LS(C)$ and $LS(B)$. Safer sample D is contained in $LS(A)$ and $LS(B)$.
(d) If more samples regard sample $x_i$ as the NE, sample $x_i$ will be closer to other classes. In figure 1, the most remote minority class sample E has the most number of those samples whose NE is E.

4. Proposed Algorithm
The distance metric of this paper is the Euclidean distance. Let $X = \{x_1, x_2, ..., x_{nmin}, x_{nmin+1}, ..., x_n\}$ be a training set. The number of samples in $X$ is $n$. The number of attributes is $d$. Let $Smin = \{x_1, x_2, ..., x_{nmin}\}$ be the set of minority class samples. The variable $nmin$ represents the instance number of set $Smin$. Likewise, Let $Smaj = \{x_{nmin+1}, x_{nmin+2}, ..., x_n\}$ be the set of majority class samples and $nmaj$ represents the number of $Smaj$. The binary classification is our focus.
The main ideas of LS-SMOTE contain three parts. First, the local sets are searched on imbalanced data. Second, a local sets-based noise filter is designed to remove noise and smooth the class boundary. Third, on each local set, the interpolation of SMOTE between a base sample and a selected sample closest to the majority class is used to generate the synthetic minority class samples. After that, imbalanced data can be improved by using synthetic minority class samples to extend it.

Section 4.1 introduces the noise filter based on the LS. Next, Section 4.2 introduces the oversampling technique with local sets and SMOTE.

4.1. Noise Filter Based on Local Sets

As the analysis in Section 3 shows, if sample \( x_i \) is closer to other classes, more samples regard \( x_i \) as the NE. Based on the above phenomenon, the usefulness of sample \( x_i \) is defined as follows:

**Definition 4.** (Usefulness of samples \( x_i \)): the usefulness of sample \( x_i \) is the number of LS which contains sample \( x_i \).

\[
Usefulness(x_i) = |\{NES(x_j) | x_j \in NES(x_i)\}|
\]  

(3)

**Definition 5.** (Harmfulness of samples \( x_i \)): the harmfulness of sample \( x_i \) is the number of sample \( x_j \) whose \( NE(x_i) \) is \( x_i \).

\[
Harmfulness(x_i) = |\{x_j | x_j \in NE(x_i)\}|
\]  

(4)

According to equations (3)-(4), the usefulness indicates the security of sample \( x_i \), while the harmfulness indicates the abnormality of sample \( x_i \). Hence, the following formula is used to remove noise and smooth the class boundary.

\[
Harmfulness(x_i) \geq Usefulness(x_i)
\]  

(5)

4.2. Oversampling Technique with Local Sets and SMOTE

The pseudo-code for the LS-MOTE is described in Algorithm 1. LS-MOTE requires one parameter, i.e. the parameter \( N \). \( N \) refers to the number of synthetic samples for each minority class sample.

**Algorithm 1: LS-SMOTE**

| Line | Description |
|------|-------------|
| 1    | \( X = Smin \cup Smaj \); |
| 2    | Finding the local set for each sample in \( X \) by formulas (1)-(2); \( O(n^2) \) |
| 3    | Using formula (5) to remove noise and smooth the boundary; \( O(n^2) \) |
| 4    | Finding the local set for each sample in \( X \) by formulas (1)-(2); \( O(n^2) \) |
| 5    | for each \( x_i \in Smin \) \( Base=x_i \); |
| 6    | \( TempN=N \); \( O(nmn) \) |
| 7    | while \( TempN \geq 0 \); |
| 8    | \( SelectedSample \) is the farthest sample in the LS of \( Base \), where \( SelectedSample \) has the smallest LSC; \( O(nmn*N) \) |
| 9    | for \( D=1 \) to \( d \); \( O(nmn*d*N) \) |
| 10   | \( gap=Base[D]-SelectedSample \); \( O(nmn*d*N) \) |
| 11   | \( New[D]=Base[D]+gap*rand(0,1) \); \( O(nmn*d*N) \) |
| 12   | end for |
| 13   | \( TempN=TempN-1 \); \( O(nmn*N) \) |
| 14   | \( Synthetic=Synthetic \cup New \); |
| 15   | end while |
| 16   | end for |
| 17   | return \( Synthetic \); \( O(1) \) |
Local sets are found by Line 2. Then, the noise filter based on local sets (equation (5)) is used to remove noise and smooth the class boundary on Line 3. Next, Local sets are search again on Line 4 since noise can affect the local sets. Lines 5-18 are the process of interpolation. Concretely, a base sample (Base) is selected on Line 6. In the LS of the selected base sample, the farthest sample (SelectedSample) is selected on Line 9. SelectedSample has the smallest LSC. On Lines 10-13, the difference of each attribute between Base and SelectedSample is calculated. Next, the difference is scaled by the function rand(0, 1). The function rand(0, 1) refers to a random number in the range [0, 1]. On Line 12, the difference is employed to synthesize new samples. Finally, the set of synthetic samples (Synthetic) can be got on Line 18. As the analysis in Algorithm 1, the time complexity of LS-SMOTE is \(O(n^2)\).

Figure 2 shows the results of applying LS-SMOTE on an artificial data set. Observing in figure 2, LS-SMOTE can remove noise in the original data and avoid noise generation. Besides, LS-SMOTE creates more synthetic borderline samples, which avoids the generation of redundant samples in the class center and is beneficial to classification performance.

5. Experiments
We select experimental data sets from UCI (University of California Irvine). These experimental data sets are real data sets from various fields. Table 1 introduces the adopted real data sets from 8 aspects, such as the number of samples, the number of attributes, the number of minority class samples, the number of majority class samples, etc.

Multiclass data sets were binarized using the one-versus-rest method. The smallest class is regarded as the minority class, and the remaining data is marked as the majority class. Additionally, we use the tenfold cross-validation to divide data sets into the training and test parts, respectively. Hence, experimental data is obtained by repeating the 10-fold cross-validation 10 times. The KNN classifier with \(k=3\) is adopted to validate the comparison methods. \(F\)-measure with \(\beta=1\) and \(G\)-mean are used as evaluation metrics.

In \(F\)-measure and \(G\)-mean, positive and/or negative cases are implied. In imbalanced classification, the positive case refers to the minority class, while the negative case refers to the majority class. If the value of \(G\)-mean is large, the classifier performs equally well on the samples of positive and negative cases. If the value of \(F\)-measure is larger, the accuracy in the positive case is higher.

5.1. Experiments on Real Data Sets
We validate LS-SMOTE by comparing it with 5 comparative methods, such as SMOTE, Safe-Level-SMOTE, Borderline-SMOTE, ADASYN and \(k\)-means SMOTE. Section 2 has described comparative methods in detail. The parameters of comparative methods are set as their standard versions. The parameter \(N\) is set to 2 in LS-SMOTE. Tables 2-3 shows the results in \(F\)-measure and \(G\)-mean.

In table 2, LS-SMOTE achieves the highest \(F\)-measure for 8 of 12 data sets. In table 3, LS-SMOTE achieves the highest \(G\)-mean for 10 of 12 data sets. Meanwhile, LS-SMOTE outperforms comparative methods in achieving the average \(F\)-measure and \(G\)-mean in the row labeled “Average”. The results prove that LS-SMOTE outperforms comparison methods in improving the minority class. Also, LS-
SMOTE performs slightly worse than the comparative methods on some data sets. The reason is that none of the methods can be applied to all data sets because of the complex distribution for different data sets.

Table 1. Adopted real data sets from UCI.

| ID | Data sets                      | Samples | Features | Minority samples | Majority samples | Imbalance ratios | Classes | Abbreviations |
|----|--------------------------------|---------|----------|------------------|------------------|-----------------|---------|---------------|
| 1  | Heart                          | 270     | 13       | 120              | 150              | 1.25            | 2       | D1            |
| 2  | Haberman’s Survival            | 306     | 3        | 81               | 255              | 2.78            | 2       | D2            |
| 3  | Vertebral Column               | 310     | 6        | 100              | 210              | 2.10            | 2       | D3            |
| 4  | Liver                          | 345     | 6        | 145              | 200              | 1.38            | 2       | D4            |
| 5  | Wholesale Customers            | 440     | 7        | 142              | 298              | 2.10            | 2       | D5            |
| 6  | Tic Tac Toe Endgame            | 470     | 16       | 70               | 400              | 5.71            | 2       | D6            |
| 7  | Wisconsin Diagnostic Breast Cancer | 569   | 30       | 212              | 357              | 1.68            | 2       | D7            |
| 8  | Australian Credit Approval     | 690     | 14       | 307              | 383              | 1.25            | 2       | D8            |
| 9  | Breast Cancer Wisconsin        | 699     | 9        | 241              | 458              | 1.90            | 2       | D9            |
| 10 | Pima Indians Diabetes          | 768     | 8        | 268              | 500              | 1.87            | 2       | D10           |
| 11 | Audio                          | 776     | 26       | 305              | 471              | 1.54            | 2       | D11           |
| 12 | Vehicle                        | 846     | 18       | 199              | 647              | 3.25            | 4       | D12           |

Table 2. Experimental results in terms of F-measure (%).

| Data       | SMOTE Safe-Level-SMOTE | Borderline-SMOTE | ADASYN k-means | SMOTE | LS-SMOTE |
|------------|------------------------|------------------|----------------|-------|----------|
| D1         | 62.76                  | 63.47            | 60.55          | 62.85 | 64.65    | 65.24   |
| D2         | 38.22                  | 38.23            | 43.68          | 37.71 | 38.90    | 44.27   |
| D3         | 73.89                  | 75.63            | 72.91          | 73.57 | 75.86    | 76.72   |
| D4         | 61.90                  | 62.12            | 59.61          | 63.65 | 62.93    | 64.15   |
| D5         | 81.46                  | 84.70            | 81.64          | 83.26 | 84.90    | 84.81   |
| D6         | 59.79                  | 63.23            | 63.50          | 59.19 | 54.04    | 63.78   |
| D7         | 87.53                  | 88.20            | 89.02          | 88.26 | 88.88    | 89.26   |
| D8         | **64.56**              | 61.51            | 62.63          | 63.45 | 62.82    | 63.67   |
| D9         | 92.43                  | 92.54            | **94.35**      | 94.15 | 91.85    | 92.19   |
| D10        | 57.04                  | 57.00            | 55.92          | 57.39 | 58.10    | **58.67**|
| D11        | 94.68                  | 94.84            | 95.17          | 94.79 | 95.10    | **95.21**|
| D12        | 82.04                  | 82.95            | **85.60**      | 82.27 | 83.70    | 84.14   |
| Average    | 71.36                  | 72.04            | 72.05          | 71.71 | 71.81    | **73.51**|
| Wilcoxon + | +                      | +                | +              | +     | N/A      |

We also use the non-parametric two-sided Wilcoxon signed ranks test to assess the significant differences between the comparative method and LS-SMOTE. The significance level is 0.05 in the Wilcoxon test. The symbols +, - and ~, in the row labeled Wilcoxon, indicate that LS-SMOTE is significantly better, worse or equivalent compared with others. Observing the row labeled “Wilcoxon”, LS-SMOTE is significantly better than comparison methods in tables 2-3.

Generally, experiments prove that LS-SMOTE outperforms comparative methods in improving KNN.
Table 3. Experimental results in terms of G-mean (%).

| Data | SMOTE | Safe-Level-SMOTE | Borderline-SMOTE | ADASYN | k-means SMOTE | LS-SMOTE |
|------|-------|------------------|------------------|--------|---------------|----------|
| D1   | 59.64 | 63.28            | 59.60            | 62.89  | 62.93         | 64.12    |
| D2   | 53.82 | 53.71            | 59.23            | 54.01  | 54.41         | 56.24    |
| D3   | 92.87 | 92.48            | 93.98            | 92.75  | 93.20         | 94.12    |
| D4   | 57.01 | 61.88            | 54.24            | 63.15  | 58.46         | 63.41    |
| D5   | 87.81 | 89.51            | 87.86            | 89.45  | 89.80         | 89.88    |
| D6   | 69.09 | 73.22            | 71.28            | 67.63  | 63.95         | 72.14    |
| D7   | 90.11 | 90.46            | 91.22            | 90.70  | 91.04         | 91.36    |
| D8   | 61.79 | 61.92            | 61.69            | 63.93  | 63.06         | 64.12    |
| D9   | 95.73 | 95.89            | 96.43            | 96.43  | 95.28         | 96.51    |
| D10  | 62.75 | 64.12            | 62.47            | 64.64  | 65.51         | 65.67    |
| D11  | 95.51 | 95.52            | 95.63            | 95.49  | 95.62         | 95.79    |
| D12  | 93.63 | 94.77            | 95.45            | 95.35  | 94.41         | 95.92    |
| Average | 76.65 | 78.06            | 77.42            | 78.04  | 77.31         | 79.11    |
| Wilcoxon | +   | +               | +               | +      | N/A           |          |

6. Conclusions
We aim to solve the difficult issues, namely, the generation of unnecessary noise and redundant synthetic samples in the class center, and the choice of parameter k.

Hence, an oversampling method based on local sets and SMOTE (LS-SMOTE) is proposed. First, the local sets are searched to describe the local characteristic of imbalanced data. Second, a local sets-based noise filter is designed to remove noise and smooth the class boundary. Finally, on each local set, the interpolation of SMOTE between a base sample and a sample closest to the majority class is used to generate the synthetic minority class samples.

In experiments, the performance of LS-SMOTE is validated. In detail, we compare LS-SMOTE with 5 popular oversampling methods on 12 real data sets. Experiments have proved that LS-SMOTE outperforms comparative methods in improving the positive case (i.e., minority class) and achieving 2 metrics (i.e., F-measure and G-mean) of KNN classifier.

Acknowledgement
This work is supported by Natural Science Foundation of Chongqing City, China (Grant No. cstc2019jcyj-bshX0097).

References
[1] Hong J, Kang H and Hong T 2020 Oversampling-based prediction of environmental complaints related to construction projects with imbalanced empirical-data learning Renewable and Sustainable Energy Reviews 134 110402 https://doi.org/10.1016/j.rser.2020.110402.
[2] Zhang X, Zhao L, Zhong W and Gu F 2021 A novel hybrid resampling algorithm for parallel/distributed particle filters Journal of Parallel and Distributed Computing 151 24-37.
[3] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic minority oversampling technique J. Artificial Intelligence Research 16 321-357.
[4] Du W Z 2014 The recognition of abnormal gas measurement data and oil and gas show Advanced Materials Research 1010-1012.
[5] Du M, Pierce M, Pivovarova L and Yangarber R 2015 improving supervised classification using information extraction International Conference on Applications of Natural Language to Information Systems pp 3-18.
[6] Aalianvari A, Katibeh H and Sharifzadeh M 2012 Application of fuzzy delphi ahp method for the estimation and classification of ghomrud tunnel from groundwater flow hazard Arabian Journal of Geosciences 5 (2) 275-284.

[7] Han H, Wang W Y and Mao B H 2005 Borderline-SMOTE: A new over-sampling method in imbalanced data sets Learning International Conference on Intelligent Computing pp 878-887.

[8] He H, Bai Y, Garcia E A and Li S 2008 ADASYN: Adaptive synthetic sampling approach for imbalanced learning Proc. Int’l Joint Conf. Neural Networks pp 1322-1328.

[9] Bunkhumpornpat C, Sinapiromsaran K and Lursinsap C 2009 Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem Pacific-Asia Conference on Knowledge Discovery and Data Mining pp 475-482.

[10] Georgios D, Fernando B and Felix L 2018 Improving imbalanced learning through a heuristic oversampling method based on k-means and smote Information Science 465 1-20.

[11] Chen B, Xia S, Chen Z, Wang B and Wang G 2021 RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise Information Sciences 553 397-428.

[12] Batista G E, Prati R C and Monard M C 2004 A study of the behavior of several methods for balancing machine learning training data ACM SIGKDD Explorations Newsletter 6 (1) 20-29.

[13] Sáez J A, Luengó J, Stefanowski J and Herrera F 2015 SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering Information Science 291 (10) 184-203 https://doi.org/10.1016/j.ins.2014.08.051.