ICIC Express Letters
Part B: Applications
Volume **9**, Number **3**, March **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **241–246**

D-SELF-SMOTE: NEW METHOD FOR CUSTOMER CREDIT RISK PREDICTION BASED ON SELF-TRAINING AND SMOTE

Gang Wang

School of Management
Hefei University of Technology
No. 193, Tunxi Road, Hefei 230009, P. R. China
wgedison@gmail.com

Received August 2017; accepted November 2017

Abstract. *In recent years many statistical and machine learning methods have been proposed for customer credit risk prediction. However, the datasets of customer credit risk prediction are often imbalanced and only a small portion of instances are labeled in reality. In this paper, a new customer credit risk prediction method, D-Self-SMOTE, is proposed based on self-training and SMOTE with density based noise filtering strategy to solve the imbalanced data and unlabeled data problems simultaneously. Experimental results on the two public datasets show that among the compared methods, D-Self-SMOTE gets the best result, which could be a potential solution for customer credit risk prediction.*
Keywords: Customer credit risk prediction, SMOTE, Self-training, Density, Noise filtering

1. **Introduction.** The customer credit risk prediction is a critical part of a financial institution's loan approval decision processes. The purpose of customer credit risk prediction is to classify the applicants into two types: applicants with good credit and applicants with bad credit. The accuracy improvement of customer bad credit prediction can retrieve a great loss for the financial institutions [1]. Especially, with the huge growth of the credit industry, building an effective and efficient prediction model has been an important task for saving amount cost [1,2]. Therefore, customer credit risk prediction has raised more and more interests from both academic and industry fields in recent years [3].

For the customer risk prediction problem, the traditional statistical methods have been employed at the earliest, such as Linear Discriminant Analysis (LDA), Logistic Regression Analysis (LRA) and Multivariate Adaptive Regression Splines (MARS) [1,2]. Although these methods are relatively simple and explainable, the problem with applying these statistical methods to the customer credit risk prediction is that some assumptions are frequently violated in reality. In recent years, many studies have demonstrated that machine learning methods, such as Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM) can be used as alternative methods for the customer credit risk prediction [1,3]. In contrast with traditional statistical methods, machine learning methods do not assume certain data distributions. These methods automatically extract knowledge from training instances.

In reality, however, the datasets of customer credit risk prediction are often imbalanced and only a small portion of instances are labeled, which makes above mentioned methods get the unsatisfied prediction performance [3]. For the first problem, i.e., the imbalanced data problem, some researchers have noticed that the imbalanced distribution could greatly degrade the performance of prediction [4,5]. For instance, Brown and Mues gave an experimental comparison for the imbalanced credit scoring [4]. Huang et al. conducted a series of experiments to evaluate neural ANN and data mining methods

for the imbalanced credit assessment task [5]. For the second problem, i.e., the unlabeled data problem, some studies have found that abundant unlabeled instances could improve the prediction accuracy significantly [6,7]. For example, Maldonado and Paredes employed a semi-supervised approach for reject inference in credit scoring using SVM [6]. Kennedy et al., proposed semi-supervised one-class classification method for credit scoring [7]. Although many studies have considered imbalanced data and unlabeled data problems independently, few researches have proposed methods specifically designed to solve these problems simultaneously. However, in order to achieve more effective methods for the customer credit risk prediction, these two important problems should be considered simultaneously.

In this study, a new method, D-Self-SMOTE, is proposed for customer credit risk prediction based on self-training and SMOTE (Synthetic Minority Over-Sampling Technique) with density based noise filtering strategy in this research. D-Self-SMOTE uses SMOTE method to solve the imbalanced data problem, and self-training method to solve the unlabeled data problem. At the same time, in order to avoid introducing new noise from the synthetic instances in SMOTE or new labeled instances in self-training, density based noise filtering strategy is used. For the testing and illustration purpose, two public customer credit risk prediction datasets were selected to verify the effectiveness of the proposed method. Empirical results reveal that D-Self-SMOTE gets the best result among the compared methods. All these results illustrate that D-Self-SMOTE could be used to customer credit risk prediction.

The remainder of the paper is organized as follows. In Section 2, a new method, i.e., D-Self-SMOTE, is proposed for customer credit risk prediction based on self-training and SMOTE. Next, Section 3 presents the experimental design and the experimental results. Finally, Section 4 draws conclusions.

2. A New Method for Customer Credit Risk Prediction Based on Self-training and SMOTE.

2.1. Problem statement.

In this research, customer risk prediction is formulated as a semi-supervised binary classification problem. Let $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ denote the set of labeled instances, and let $U = \{x_{m+1}, x_{m+2}, \ldots, x_N\}$ denote the set of unlabeled instances. x_i is a d-dimensional feature vector. $y_i \in \{-1, +1\}$ is the class label. "+1" is denoted as the minority class, e.g., customer with credit risk. "−1" is denoted as the majority class, e.g., customer with non-credit risk. Both L and U are independently drawn from the same unknown distribution D, whose marginal distributions satisfy $P_D(y_i = +1) \ll P_D(y_i = -1)$.

As SMOTE method synthesizes minority instances and self-training selects high reliable instances into the training datasets, the noise can be potentially introduced. Therefore, density based noise filtering strategy is employed in the research, and some concepts and terms to explain the density based noise filtering method can be defined as follows [8].

Definition 2.1. *(Eps-neighborhood). The Eps-neighborhood of a point x_p, denoted by $N_{Eps}(x_p)$, is defined by $N_{Eps}(x_p) = \{x_q \in X | dist(x_p, x_q) < Eps\}$, $X = \{x_1, x_2, \ldots, x_m\}$.*

Definition 2.2. *(Core point). A core point refers to the point whose neighborhood of a given radius (Eps) has to contain at least a minimum number (MinPts) of the other points.*

Based on above concepts, we can use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method to construct density based clusters.

Definition 2.3. *(Density based cluster). A cluster C is a non-empty subset of X satisfying the following requirements:*

(1) $\forall x_p$, x_q*: if* $x_q \in C$ *and* x_p *is density reachable from* x_q *with respect to Eps and MinPts, then* $P \in C$*.*

(2) $\forall x_p$, $x_q \in C$*:* x_p *is density connected to* x_q *with respect to Eps and MinPts.*

Definition 2.4. *(Noise). Let* $C_1, \ldots, C_k$ *be the clusters of non-empty subset of* X*. Then the noise is the set of points in* X *not belonging to any* C_i*, where* $i = 1, \ldots, k$*, noise* $= \{p \in X | \forall i : p \notin C_i\}$*.*

2.2. **A new customer credit risk prediction method: D-Self-SMOTE.** Based on the above discussion, we can propose a novel customer credit prediction method based on the SMOTE and self-training to solve the imbalanced data problem and unlabeled data problem simultaneously. For the imbalanced data problem, a number of methods were proposed from the perspective of dataset or algorithm [9]. Compared with algorithmic level methods, data level methods are algorithm-independent and often used in practice. Therefore, data level method is considered in this research. For the data level method, random under sampling or over sampling is a straightforward strategy to handle imbalanced data problems [9]. As random under sampling method could discard some useful information, and over sampling method could cause over-fitting, SMOTE, one of the popular advanced over sampling method, is employed in the research [10]. At the same time, SMOTE randomly selects an instance from the line connecting the neighbor and the instance itself, which can potentially introduce noise into the training data set and humble the performance of classifiers. In order to reduce this kind of influence, DSMOTE is proposed, which uses the density based noise filtering strategy. Just like SMOTE, DSMOTE also searches the K nearest neighbors for each minority instance at the very beginning. Unlike SMOTE, DSMOTE uses density based cluster method to judge whether the neighbor belongs to the noise. If the neighbor is the noise, this neighbor can be discarded directly. Then, DSMOTE also employs density based cluster method to judge whether the minority instance and neighbor belong to the same density based cluster. If the minority instance and neighbor do not belong to the same cluster, in order to avoid introducing new noise, DSMOTE uses

$$x_{new} = \dot{x} + rand(0, 1) \times Eps \tag{1}$$

to generate the new instance. $\dot{x}$ denotes the minority instance or the neighbor, which can be randomly selected in the method. If the minority instance and neighbor belong to the same cluster, DSMOTE uses

$$x_{new} = x + rand(0, 1) \times (\tilde{x} - x) \tag{2}$$

to generate the new instance. x denotes the minority instance and $\tilde{x}$ denotes the neighbor. Then DSMOTE judges whether the new instance x_{new} belongs to the noise. If it is the noise, this news instance could also be discarded directly. Compared with the traditional SMOTE method, DSMOTE can generate more accurate minority instances by using density based noise filtering strategy.

For the unlabeled data problem, many approaches have also been proposed in the literature, such as semi-supervised learning and active learning. Among these approaches, disagreement based semi-supervised learning approach explores the unlabeled data automatically, where no human intervention is assumed. Therefore, disagreement based semi-supervised learning approach is employed in this research. Meanwhile, compared with other disagreement based semi-supervised learning approaches, such as co-training, and tri-training, self-training only needs one classifier with no split of features [6,7]. For the customer credit prediction method, two independent feature sets are difficult to be acquired. As a consequence, self-training is employed in the research, which starts with a set of labeled data set and builds a classifier. Then, only these instances with a labeling confidence exceeding a certain threshold are added into the labeled data set. Just like

SMOTE, self-training could also potentially introduce new noise into the labeled data set, especially after DSMOTE is used to generate more minority instances. Therefore, density based noise filtering strategy is also employed into the standard self-training method. For every training iterations, the potential instance, which will be added into the training data set, must be judged whether it belongs to the noise. If the potential instance belongs to the noise, it will be discarded directly.

Based on above analysis, D-Self-SMOTE is proposed for the customer credit risk prediction based on SMOTE and self-training to solve imbalanced data and unlabeled data problems simultaneously. In D-Self-SMOTE, the classifier is generated using the original labeled training dataset with DSMOTE at first. Then, it iterates the following procedures K times. Firstly, the classifier labels the most confident positive and negative unlabeled instances with density based noise filtering strategy. Then, the classifier will be retrained using the updated training dataset. The whole process will repeat until the classifier is unchanged or pre-set number of learning rounds K has been executed. The pseudo code of D-Self-SMOTE is shown in Figure 1.

Input: Labeled instance sets: L;
Unlabeled instance set: U;
Eps-neighborhood: *Eps*;
Mminimum number of points: *MinPts*;
Learning round: K;
Base classifiers: F.
Process:
1. Use density based cluster method with L, *Eps*, *MinPts* to get $C_1^p, \ldots, C_k^p$ and $C_1^n, \ldots, C_k^n$;
2. Use DSMOTE(L) to generate balanced data set L_1 using Equations (1) and (2);
3. Loop for K iterations:
4. Use L_1 to train a classifier F that considers only the x_1 portion of x;
5. Do
6. Allow F to label most confident positive instance from U;
7. If ($x_p \notin$ Noise), add x_p into L_1;
8. While (F labels one positive instance);
9. Do
10. Allow F to label most confident negative instance from U;
11. If ($x_n \notin$ Noise), add x_n into L_1;
12. While (F labels one negative instance);
Output: $F(x)$

FIGURE 1. The pseudo code of D-Self-SMOTE

3. Experiment and Results.

3.1. **Experiment setup.** The effectiveness of D-Self-SMOTE was evaluated on two public customer credit risk prediction benchmark datasets: German and England dataset [1]. It is now well-known that average accuracy is not an appropriate evaluation criterion when there is class imbalance. Thus, AUC (Area Under the ROC Curve) was used as performance evaluation measure in this research. In the experiment, SVM was chosen as base classifier for the D-Self-SMOTE, imbalanced classification methods, i.e., Under Sampling method (US), Over Sampling method (OS), SMOTE, Bagging, Boosting, and self-training related methods, i.e., the standard Self-training, Self-training with Under

Sampling strategy (Self-US), Self-training with Over Sampling strategy (Self-OS), Self-training with SMOTE (Self-SMOTE), Self-training with Bagging (Self-Bagging), Self-training with Boosting (Self-Boosting). In the research OPTICS (Ordering Points To Identify the Clustering Structure) method is used to generate the density based clusters and judge the noise. To minimize the influence of the variability of the training set, ten times 10-fold cross validation is performed on the dataset. For the union dataset of nine subsets, it is partitioned into a labeled training dataset L, and an unlabeled training dataset U under different label rates including 20%, 40%, and 60%.

3.2. **Results and discussion.** Figure 2 summarizes the experiment results of different methods when label rate is 20%.

As shown in Figure 2, for the sampling methods, SMOTE gets the improved results, i.e., 73.31%, and 58.58%. For the Self-SVM it also gets the improved results, i.e., 74.93%, and 60.08%. These results indicate that SMOTE and self-training can solve the imbalanced data and unlabeled data problems individually. Subsequently, D-Self-SMOTE all gets the highest AUC, i.e., 77.62%, and 62.75% at the German and England datasets. These results indicate that D-Self-SMOTE can solve above two problems simultaneously.

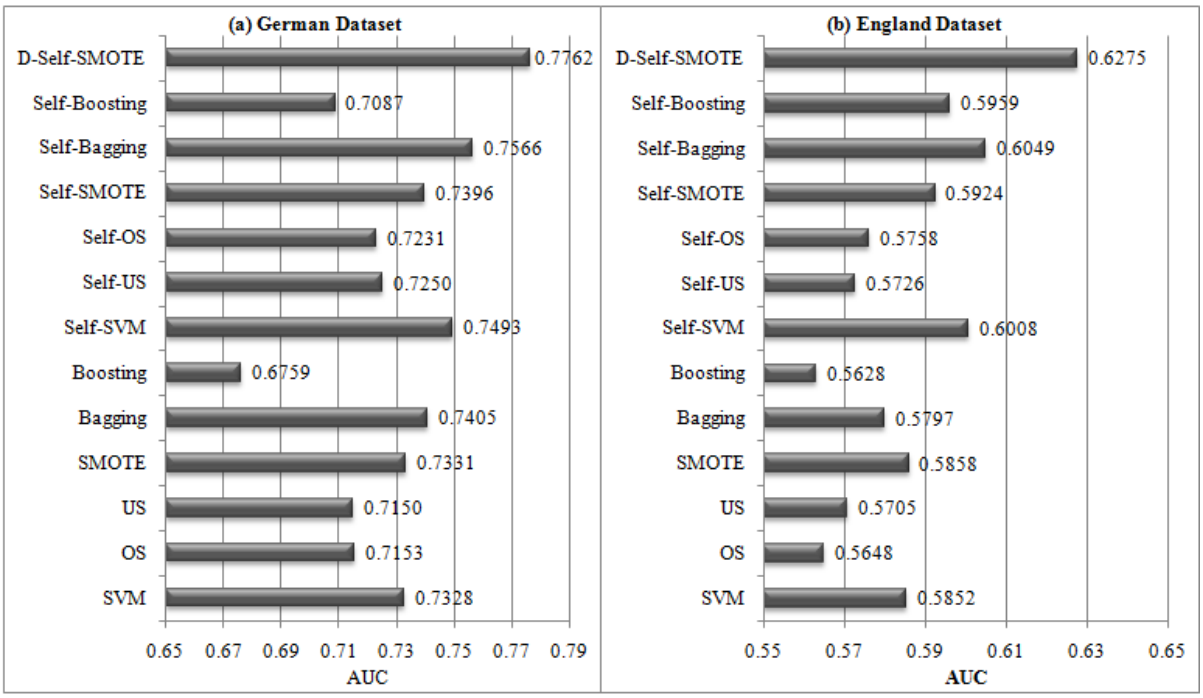

FIGURE 2. Experimental results

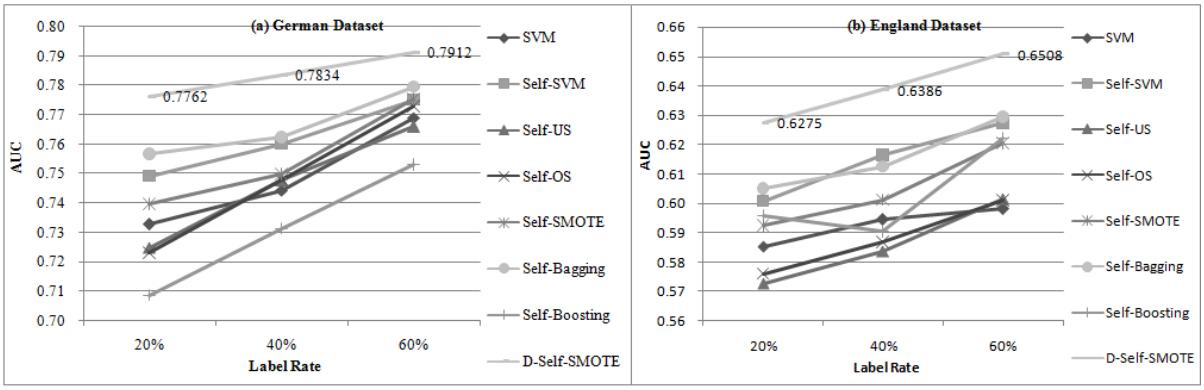

FIGURE 3. Sensitivity analysis of AUC

Next, the average AUCs are compared at the different label rates. The experimental result is shown in Figure 3. D-Self-SMOTE gets the highest average AUCs: 77.62%, 78.34%, 79.12% at the German dataset, and 62.75%, 63.86%, 65.08% at England dataset. It is interesting that with the increase of label rate, the average AUCs of compared methods are also increasing, except for Self-Boosting at England dataset. These results indicate that although the self-training methods can utilize the unlabeled data, the labeled data are very important for customer credit risk prediction in reality.

4. **Conclusions.** In this study, a novel method, D-Self-SMOTE, is proposed for customer credit risk prediction based on SMOTE and self-training with density based noise filtering strategy to solve the imbalanced data and unlabeled data problems at the same time. D-Self-SMOTE employs SMOTE and self-training method to solve the imbalanced data and unlabeled data problems simultaneously. Meanwhile, density based noise filtering strategy is employed to avoid introducing new noise from the synthetic instances in SMOTE or new labeled instances in self-training. Experimental results based on the two public customer credit risk prediction datasets show that D-Self-SMOTE gets the highest average AUC among the compared methods.

Several future research directions also emerge. Firstly, as this research only verifies the proposed method experimentally, more deep theoretical analyses for D-Self-SMOTE are needed in the future research. Secondly, the more diverse semi-supervised methods and imbalanced data classification methods could be explored collaboratively in the future research.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China (71471054, 91646111), Anhui Provincial Natural Science Foundation (1608085MG150), Social Science Knowledge Popularization Foundation of Anhui Province (Y2016016), Special Fund of Political Theory Research Center of Hefei University of Technology (JS2015HGXJ0051), Training Program of Application of Scientific and Technological Achievement of Hefei University of Technology (JZ2017YYPY0235).

REFERENCES

[1] M. R. Sousa, J. Gama and E. Brandão, A new dynamic modeling framework for credit risk assessment, *Expert Systems with Applications*, vol.45, pp.341-351, 2016.

[2] Z.-C. Yang, H. Kuang, J.-S. Xu and H. Sun, Credit evaluation using eigenface method for mobile telephone customers, *Applied Soft Computing*, vol.40, pp.10-16, 2016.

[3] N. Chen, B. Ribeiro and A. Chen, Financial credit risk assessment: A recent review, *Artificial Intelligence Review*, vol.45, pp.1-23, 2016.

[4] I. Brown and C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Systems with Applications*, vol.39, pp.3446-3453, 2012.

[5] Y.-M. Huang, C.-M. Hung and H. C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, *Nonlinear Analysis: Real World Applications*, vol.7, pp.720-747, 2006.

[6] S. Maldonado and G. Paredes, A semi-supervised approach for reject inference in credit scoring using SVMs, *Advances in Data Mining. Applications and Theoretical Aspects*, pp.558-571, 2010.

[7] K. Kennedy, B. M. Namee and S. J. Delany, Using semi-supervised classifiers for credit scoring, *Journal of the Operational Research Society*, vol.64, pp.513-529, 2013.

[8] Y. Lv, T. Ma, M. Tang, J. Cao, Y. Tian, A. Al-Dhelaan et al., An efficient and scalable density-based clustering algorithm for datasets with complex structures, *Neurocomputing*, vol.171, pp.9-22, 2016.

[9] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowledge and Data Engineering*, vol.21, pp.1263-1284, 2009.

[10] 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, no.1, pp.321-357, 2002.