A Re-sampling Method for Class Imbalance Learning with Credit Data

Li Zhang WenXian Wang

School of Information Technology & Management Engineering,
University of International Business and Economics, Beijing, china,100029
zhangli426@hotmail.com

***Abstract*--Credit rating is a typical class imbalance problem. Oversampling methods are commonly used for dealing with this problem. This paper presents an improved oversampling approach based on synthetic minority over-sampling technique(SMOTE). First, use sample distribution of the minority class to estimate whether different types of samples are crossed. Then generate synthetic samples by samples in different class when different classes cross seriously. In addition, increase weights of part positive samples, which may be not on borderline. At last, the proposed method is evaluated on a credit data set. The results indicate that it is more effective than other methods for the class imbalance learning.**

Keywords-- class imbalance; credit rating; SMOTE; sample distribution

I. INTRODUCTION

Credit risk evaluation analysis is a typical classification problem and credit data are predominately composed of ‘‘good’’ samples with only a small percentage of ‘‘bad’’ ones, leading to the so-called class imbalance problem. Currently many data mining techniques have been used to evaluate credit risk, such as logistic regression techniques, linear discriminant analysis, k-nearest neighbor ,neural networks (ANN), genetic algorithm(GA)and support vector machine (SVM)[1-2]etc. Those have been tested and compared being effective classification methods for the credit risk evaluation. In addition, some ensemble models[3] have been adopted to improved the learning performance for the class imbalance problem. But some surveys indicated that the results are far from ideal. For the class imbalance problems, using all the data to build the learning model will usually lead a learning bias to the majority class, while in credit risk evaluation, it is more important to correctly classify the minority class that are the main source of bank losses.

In machine learning and data mining area, a number of solutions to the class imbalance problem were previously proposed both at the data and algorithmic levels. The algorithmic level introduces unequal weights for the minority and majority classes in the training strategy of the classifier or uses cost sensitive learning to modify the classifier. The data level approach is usually based on re-sampling methods, which attempt to balance the class distribution by either adding examples to the minority class (oversampling) or removing examples from the majority class (under-sampling)[4]. The simplest method for re-sampling a data set is random re-sampling. Random oversampling balances a data set by duplicating examples of the minority class until a desired class ratio is achieved. Similarly, random under-sampling (RUS) removes examples (randomly) from the majority class until the desired balance is achieved. There are also “intelligent” methods for oversampling and under-sampling[5-6].

Those re-sampling methods make the problem more tractable and yields good accuracy on the test instances. However, it is worth to note that the imbalanced dataset problem may actually arise from two different sides. One is interclass imbalance where the distribution of class labels varies widely. The other is within-class (intra-class) imbalance, which may occur when the members of a class are not distributed in a uniform distribution. Re-sampling methods often ignore the within-class imbalance problem. In fact, re-sampling techniques may worsen within-class imbalance [7].

978-0-7695-4522-6/11 $26.00 © 2011 IEEE
DOI 10.1109/ICM.2011.34

To remedy this problem, this paper presents an improved oversampling approach based on synthetic minority over-sampling technique(SMOTE)[6] and samples within-class distribution of the minority groups. As multi-class classification can be seen as multiple binary classification problems, only binary classification problem is considered in this paper.

II. RELATED WORK

A. Oversampling Techniques for Class Imbalance

As our work mainly involves oversampling techniques, related literature discussion is limited to existing data oversampling techniques for the class imbalance in this section. One technique which has received much attention in recent literature is SMOTE [6], which creates new minority class examples, rather than simply duplicating them. SMOTE creates attribute values for the new instances by extrapolating values from the *k* nearest neighbors(*kNN*) to each of the original minority class examples. Let $S_{\min}, S_{maj}, S_{syn}$represent the data set of minority class, majority class and synthetic samples set respectively, while $|S_{\min}|, |S_{maj}|, |S_{syn}|$ are numbers of samples in each data set. Each sample has m attributes. Synthetic samples are generated following these steps.

1) Determine *kNN* for each original sample $x_i \in S_{\min}$, and determine the value of $|S_{syn}|$needed.

2) Choose a random sample $x_t (t = 1,2,..., k)$ from k nearest neighbors of sample $x_i \in S_{\min}$.

3) Use (1) to create a synthetic sample x_n .

$$x_{nj} = x_{ij} + gap \times (x_{tj} - x_{ij}) \qquad (1)$$

Where gap is a random number between 0 and 1 and $i = 1,2,..., |S_{\min}|, t = 1,2,..., k, j = 1,2,..., m$.

4) Repeat step 2) and 3) until $|S_{syn}|$ synthetic samples are created.

From the above steps, SMOTE method generates the same number of synthetic data samples for each original minority example and does so without consideration to the distribution of examples, which may cause more overlapping between classes. Borderline-SMOTE [8] is an attempt to improve upon SMOTE by only oversampling minority class examples that are believed to be on the border of the decision regions. It is achieved as follows:

1) Determine *kNN* for each original sample $x_i \in S_{\min}$ and identify the number $m_i^{'}$ of nearest neighbors that belongs to the majority class.

2) If $k/2 < m_i^{'} < k$ is true, $x_i \in S_{\min}$ is considered as borderline sample.

3) Create synthetic sample using borderline samples and their nearest neighbors, which is similar as SMOTE algorithm.

In Borderline-SMOTE method, if all of the nearest neighbors of x_i of minority class are majority examples, x_i is considered as noise and no synthetic examples are generated for it.

Adaptive Synthetic Sampling (ADASYN) algorithm [9] is another representative of oversampling. Its key idea is to change the weights of borderline samples in minority class. So at this point, it is similar to Borderline-SMOTE method.

B. Classifier Performance Metrics

Most of the studies in imbalanced domains mainly concentrate on two-class problem. By convention, the class label of the minority class is positive, and the class label of the majority class is negative. Table 1 illustrates a confusion matrix of a two-class problem[8]. *TP* and *TN* denote the number of positive and negative examples that are classified correctly, while *FN* and *FP* denote the number of misclassified positive and negative examples respectively.

TABLE 1 CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
Positive	TP	FN
Negative	FP	TN

Evaluation metrics are frequently adopted in the research of imbalanced learning problems, namely, *precision; recall;* $F_{measure}$. These metrics are defined as:

$$precision = TP/(TP + FP), recall = TP/(TP + FN)$$

$$F_{measure} = \frac{(1+\beta^2)\, precision \times recall}{\beta^2 (precision + recall)}$$

Where β is a coefficient to adjust the relative importance of precision versus recall (usually $\beta = 1$). The $F_{measure}$ metric combines precision and recall as a measure of the effectiveness of classification, so it provides more insight into the functionality of a classifier than the accuracy metric.

Other evaluation metrics such as *G-mean* and Receiver Operating Characteristics (ROC) Curves are also be used in many literatures [10]. In credit risk evaluation, it is more important to correctly classify the minority class, so in this paper we only use *recall* and $F_{measure}$ metrics.

III. PROPOSED APPROACH

To remedy the problems of SMOTE and Borderline-SMOTE, an improved method called Distribution-SMOTE is proposed in this section, which main idea is oversampling based on the distribution between different classes and SMOTE. Synthetic samples are generated following these steps.

1) Determine the number of synthetic samples $|S_{syn}|$ needed according to the ideal ratio between negative and positive samples.

2) Let $k = \text{int}(|S_{syn}|/|S_{\min}|)$, where int means taking the integer part of expression. For each $x_i \in S_{\min}$, determine its *kNN* in same class and different class set respectively.

3) For each $x_i \in S_{\min}$, calculate the average distance $d_{intra}(i)$ between it and its *kNN* in same class set, and calculate the average distance $d_{exter}(i)$ between it and its *kNN* in different class set. Then calculate the average of $d_{intra}(i)$ and $d_{exter}(i)$ respectively by (2),(3).

$$\overline{d_{intra}} = \frac{1}{M}\sum_{i}^{M} d_{intra}(i) = \frac{1}{M}\sum_{i}^{M}\sum_{j=1}^{k} distance(x_i, x_j) \quad x_j \in S_{\min} \tag{2}$$

$$\overline{d_{exter}} = \frac{1}{M}\sum_{i}^{M} d_{exter}(i) = \frac{1}{M}\sum_{i}^{M}\sum_{j=1}^{k} distance(x_i, x_j) \quad x_j \in S_{maj} \tag{3}$$

4) Let $\alpha = \overline{d_{intra}} / \overline{d_{exter}}$. If $\alpha < T$ (usually $T = 0.5$), generate synthetic samples similar as SMOTE method. Otherwise, it indicates that different types of samples are crossed and the larger of α the more serious crossing, even samples of minority class can be devoted as noise of majority class, synthetic samples is generated as follows:

Determine *kNN* $x_p (p = 1,2,...,k)$ for $x_i \in S_{\min}$ in the whole training data set. Use (4) to create a synthetic sample x_n. If $x_p \in S_{\min}$,use (5) create x_{n+1} at same time, Shown as Fig.1.

$$x_{nj} = x_{ij} + gap_1 \times (x_{pj} - x_{ij}) \quad (j = 1,2,...,m) \tag{4}$$

$$x_{(n+1)j} = x_{pj} + gap_2 \times (x_{ij} - x_{pj}) \quad (j = 1,2,...,m) \tag{5}$$

Where gap_1, gap_2 is a random number between 0 and 0.5.

Compared other three methods mentioned in section Ⅱ, the method proposed in this section has two improvements. Firstly, different data generated formula is adopted based on the distribution. It is fit for data crossing seriously. Secondly, instead of increasing the weights of samples on the class borderline, the weight of samples that may be not on the borderline is increased. The main purpose is not to change samples distribution in minority class.

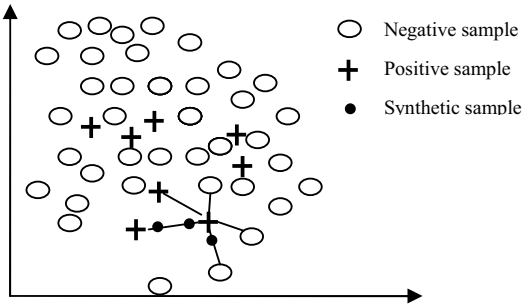

Figure 1 Data generated based on proposed method

IV. EXPERIMENTS

A. Description of the Data set

The proposed methodology is applied to German credit dataset taken from the UCI database. The original dataset is composed by 700 "good" samples and 300 "bad" samples. After changing all qualitative variables into quantitative variables, there are 25 attributes (one is class label) for each sample. Only 600 "good" samples and 100 " bad" samples are selected to build a training set and the rest 100 "good" samples and random selected 100 "bad" samples are used to build the testing set, shown as Table 2. While there is no universally accepted optimal class distribution, a balanced (50:50) distribution is often considered to be near optimal. However, when examples from the minority class are very rare, a ratio closer to 35:65 (minority:majority) may result in better classification performance [10]. So in the experiments we let the ratio in all re-sampling training sets equal or exceed 2:3, which means $k = 3$ in our method.

The experimental design is conducted using 4 classifiers, namely *logistic regression, BP-network,C5.0 and SVM* ,which are used frequently for credit rating in many literatures.

TABLE 2 DATA DESCRIPTION

	"Good"class	"Bad" class	Ratio(min : maj)
Original dataset	700	300	3:7
Training set	600	100	1:6
Testing set	100	100	1:1

B. Experiment Results

First, determine 3-nearest neighbors of each "bad" sample of training set and calculate minority samples distribution $\alpha = 1.185 > 0.5$, which means samples in the training set cross seriously. While there are 50 "bad" samples whose all neighbors are heterogeneous and only 16 "bad" samples have more than two same class neighbors. Those verify that samples in training set cross seriously. In the experiments, Euclidean distance is used to calculate the distance between two samples.

Then to verify our methods is reasonable, a temp training set is built by modifying class labels of 3-nearest neighbors for minority class samples into "bad", which ratio between minority and majority is 338:362(about 1:1). Four different models are built and tested respectively, where *logistic, BP-network*, and *C5.0* models are supported by *SPSS Clementine 11.1* and SVM (RBF kernel) model is built by *LIBSVM* Tools[18]. Experiments results are showed in Table 3 and only the value of *TP* and *TN* are given in this step. The results show that a sample is similar with all its neighbors which can be used to create synthetic samples. On the other hand, a sample in minority class can't be looked as "noise" simply as in literature[8] in crossing seriously data sets.

At last, training set is re-sampled by SMOTE, Borderline-smote, Random oversampling (ROS) and Distribution-SMOTE method, shown in Table 4. For Random oversampling method, all minority samples are copied three times. In SMOTE method, let *gap*=0.5 and 300 synthetic samples are created, while in Borderline-SMOTE method, let *gap*=0.5,0.25,0.125 respectively and create 306 synthetic samples. For Distribution-SMOTE method, let gap_1=0.175 and gap_2=0.25. Testing results of classifiers are shown in Table 5 and Fig.2,Fig.3.

From those results, the performance of classifier varies largely with data sets. Relatively, our method has a better performance compared with others in the credit data set, whatever the *recall* and $F_{measure}$.

From those results, the performance of classifier varies largely with data sets. Relatively, our method has a better performance compared with others in the credit data set.

V. CONCLUSIONS

Credit data sets usually face class imbalance problems, which will decrease the classification performance because the learning algorithms often over-fit the majority class. This paper presents an improved oversampling method based on SMOTE and samples distribution, which is adapted to the case of data crossing seriously. The results of the experiment showed that this approach has better classification performance than other methods. Hence, we conclude that this paper presents a useful approach to deal with the class imbalance problem in credit data set. Future work will consider the re-sampling area and verify our method using other data sets.

TABLE 3 EXPERIMENT RESULTS

Model (200 samples,1:1)	Training set		Training temp set	
	TN	*TP*	*TN*	*TP*
Logistic	83	38	65	66
Bp-Network	99	7	61	60
C5.0	100	0	57	60
SVM	98	7	85	42

TABLE 4 DIFFERENT TRAINING DATA SET

	Good class	Bad class	Synthetic samples
ROS	600	400	300
SMOTE	600	400	300
Borderline-SMOTE	600	406	306
Distribution-SMOTE	600	468	368

TABLE 5 COMPARING OF DIFFERENT METHOD

Classifier	ROS		SMOTE		Borderline-SMOTE		Distribution-SMOTE	
	TN	*TP*	*TN*	*TP*	*TN*	*TP*	*TN*	*TP*
Logistic	78	62	87	51	81	44	75	62
Bp-Network	77	43	90	48	85	32	86	52
C5.0	86	46	76	53	76	53	72	59
SVM	74	27	79	60	72	28	81	63

ACKNOWLEDGEMENT

This work was supported by Program for Innovative Research Team and "211 Program" in UIBE.

REFERENCES

[1] Wei Guo, Meiyan Cao, Ke Gong,The Comparative Study on Credit Risk Evaluation Models of Real-estate for Chinese Commercial Banks,International Management Review Vol. 5,No.2,2009,pp.96-102.

[2] Bellotti, T., & Crook, J. Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, Vol.36,NO.2,2009,pp. 3302-3308.

[3] NATASA SARLIJA, MIRTA BENSIC, MARIJANA ZEKIC-SUSAC. Modeling customer revolving credit scoring using logistic regression,survival analysis and neural networks, Proceedings of the 7th WSEAS International Conference on Neural Networks,2006,pp.164-169

[4] YANMIN SUN,ANDREWK. C. WONG,MOHAMED S. KAMEL. CLASSIFICATION OF IMBALANCED DATA: A REVIEW, International Journal of Pattern Recognition and Artificial Intelligence,Vol. 23, No. 4,2009,pp. 687-719

[5] Francisco Fernandez-Navarro, Cesar Hervas-Martnez, Pedro Antonio Gutierrez. A dynamic over-sampling procedure based on sensitivity for multi-class problems, Pattern Recognition ,No.44 ,2011,pp.1821-1833

[6] N.V.Chawla,K.W.Bowyer,L.O.Hall,W.P.Kegelmeyer. SMOTE:synthetic minority over-sampling technique, Journal of Artificial Intelligence Research,Vol.16,No.3,2002,pp.321-357

[7] Japkowicz N. Class imbalances: are we focusing on the right issue? Notes from the ICML workshop on learning from imbalanced data sets II,2003,pp.17-23

[8] H. Han,W.Y.Wang and B.H. Mao, Borderline-smote: A new over-sampling method in imbalanced data sets learning, International Conference on Intelligent Computing (ICIC'05), 2005,pp. 878-887

[9] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Proc.Int'l J. Conf. Neural Networks,2008,pp.1322-1328

[10] Haibo He, Edwardo A. Garcia. Learning from Imbalanced Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9,2009,pp.1263-1284

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm

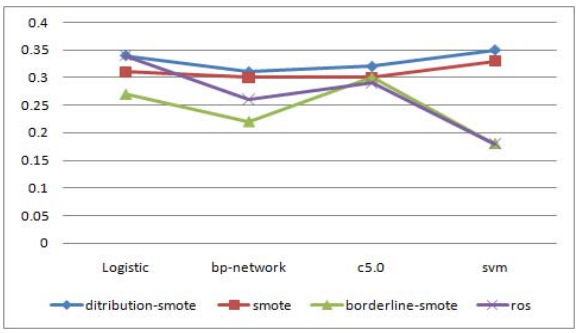

Figure 2 Comparing of $F_{measure}$

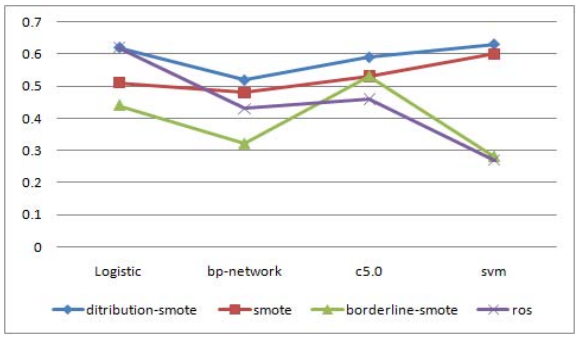

Figure 3 Comparing of *recall*