An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree

Kewen Li
College of Computer & Communication Engineering
China University of petroleum
Qingdao, Shandong Province, China
likw@upc.edu.cn

Wenrong Zhang,Qinghua Lu,Xianghua Fang
College of Computer & Communication Engineering
China University of petroleum
Qingdao, Shandong Province, China
zwr08082103@126.com

***Abstract*—Imbalanced data-set Classification has become a hotspot problem in Data Mining. The essential assumption of the traditional classification algorithms is that the distribution of the classes is balanced, therefore the algorithms used in Imbalanced data-set Classification cannot achieve an ideal effect. In view of imbalance date-set classification, we propose an oversampling method based on support degree in order to guide people to select minority class samples and generate new minority class samples. In the light of support degree, it is now possible to identify minority class boundary samples, then produce a number of new samples between the boundary samples and their neighbors, finally add the synthetic samples to the original data-set to participate in training and testing. Experimental results show that the method has an obvious advantage in dealing with imbalanced data-set.**

Keywords-Imbalanced data-sets; Classification; Boundary sample; Support degree; SMOTE

I. INTRODUCTION

In real-world applications, data-sets are usually imbalanced. So how to classify the imbalance data correctly becomes a hotspot problem. The so-called imbalanced data means that the number of samples from different classes is not nearly the same in a data-set. Focusing on a two-class imbalanced scenario, the class with small sample size is called minority class (hereinafter called positive class), the class with huge sample size is called majority class (hereinafter called negative class). In an imbalanced scenario, the traditional classification algorithms are biased toward the negative class, because it is easier to learn [1]. So the traditional classification algorithms cannot achieve ideal effects because positive class is the more valuable class [2] [3].

People usually start from two aspects depending on how they overcome the imbalance of data-set: data-level treatments and algorithm level modifications. In data-level, people usually change the distribution of data-sets by over-sampling [4] (increasing the number of samples the minority class), under-sampling [5] (reducing the number of samples the majority class) or combination of the both techniques. In algorithm level, people usually deal with imbalanced data classification problems through designing new algorithms or improving the existing algorithms.

On the basis of existing methods and evaluation criteria in imbalanced data classification, we present an improved SMOTE method (improved SMOTE method based on support degree, referred as SDSMOTE).The method chooses minority class as the center, the distance k as the radius to designate an area, then calculate the number of negative class samples as support degree. Whether a sample can be marked as a boundary sample or not depend on its support degree. Using support degree as the guidance to synthesize new sample points has a lot advantages, for example, it can avoid the disadvantages of the current methods which generate new samples blindly, make oversampling in imbalanced data preprocessing more targeted, and improve the ability to enhance the classification of the positive class.

II. IMBALANCED DATA CLASSIFICATION METHOD

A. Methods in data level

In data level, people mainly use sampling techniques to deal with imbalanced data. The basic idea of sampling is that we change the distribution of training samples to overcome the imbalance of data-set. Data sampling techniques include three types: under-sampling, over-sampling, mixed sampling. Under-sampling removes some majority class samples in order to achieve balanced data-set; oversampling increases the number of minority class samples to change the distribution of data-set; mix sampling uses both over-sampling and under-sampling techniques to deal with data-set.

1) Under-sampling technique

Random under-sampling [6] is the most simple and common method in under-sampling technique, it changes the distribution of data-set by removing some negative class samples randomly, but this method also exists shortcomings, such as deleting samples artificially may lose the samples with important information and reduce the performance of classifiers.

NCR(Neighborhood Cleaning Rule, NCR) proposed by J.Laurikkala [7], is an under-sampling method. It uses the nearest neighbor thought to remove negative class samples. Its basic idea is as follows: select a sample X_i from data-set randomly, then find its three nearest neighbors and their categories, compare X_i with the three neighbors: if X_i is a negative class sample, at least two of the three samples are positive class samples, then remove X_i from the data-set; if X_i is a positive class sample, at least two of the three samples are negative class samples, then remove the three neighbors from the data-set. So you can use this method to under-sample negative class samples.

978-1-4799-8003-1/14 $31.00 © 2014 IEEE
DOI 10.1109/IIKI.2014.14

2) Over-sampling technique

Random over-sampling [6] is the most simple and common method in oversampling technique. It increases the number of positive class by copying positive class samples randomly. This method really changes the distribution of data-set, but it has some shortcomings: copying too many positive class samples may cause classifier over-fitting , the time required for building classifiers becomes longer.

SMOTE algorithm [8] is a classic oversampling algorithm. The basic idea of SMOTE is that new positive class samples are synthesized through linear interpolation between two near positive class samples, then add them to the original data-set. The two classes could be balanced by increasing new minority class samples. The specific approach is: for a positive class sample X_i , calculate its distance from other samples of positive class, then select a sample X_j from the k-nearest neighbor samples of positive class randomly, finally generate new samples as the following manner:

$$X_{new} = X_i + rand(0,1) \times (X_j - X_i) \tag{1}$$

According to Eq. 1, X_{new} is added to participate in the training and testing. The method can prevent the occurrence of over-fitting effectively because it is not just copying positive class samples, but it cannot provide a scalar control of the number of new samples, and it cannot select positive class samples and synthesize new samples with guidance, so the quality of the new samples is not very good.

3) Mix sampling technique

Both over-sampling and under-sampling are able to reduce the imbalance of data-set, but they have some drawbacks inevitably. C.Drummond [9] proposed that the performance of classifiers which are built based on under-sampling technology is superior to the performance of classifiers which are built based on over-sampling technology, Chris Seiffert [10] put forward a similar view from the model training complexity and training time, GEBatista [11] thought that over-sampling technique was better than under-sampling techniques when there are overlaps in the data-set. There is not a uniform conclusion about which is better method. Therefore, combination of the two techniques is a common approach to imbalanced data classification.

B. Methods in algorithm level

In algorithm level, people make efforts to enhance the classification of the positive class and the performance of data-set by modifying existing algorithms or proposing new algorithms. The algorithms which are widely used in imbalanced domains include cost-sensitive learning algorithms, integrated learning algorithm, etc. Cost-sensitive learning algorithm assign different misclassification costs for different classes. In particular, it assigns a higher weight for positive class in order to make classifier pay more attention to positive class samples when it is used in imbalanced data classification. C4.5 algorithm is a typical cost-sensitive learning algorithm. Integrated learning algorithm is to solve a problem by combining multiple learners to work together. Although the learners influence each other, the performance is better than a single learner. The typical integrated learning algorithms are AdaBoost algorithm, Bagging algorithm etc. Combining sampling techniques with various classification algorithms is an effective way to solve imbalance data-set classification.

III. IMPROVED SMOTE METHOD BASED ON SUPPORT DEGREE

The SMOTE algorithm cannot provide a scalar control of the number of new samples and cannot guide the selection of positive class samples and the synthesis of new samples, so we proposed SDSMOTE method to solve the shortcomings. The main idea of SDSMOTE is as follows: we select the boundary sample by calculating support degree of each positive class sample, and then use SMOTE algorithm to oversample the selected samples. Boundary samples are difficult to be identified during the classification process, whereas SDSMOTE can achieve the goal of that selecting and synthesizing boundary samples of positive class discriminately, and improving the quality of synthetic samples of minority class.

Support degree could be obtained as follows: Firstly we draw a circle which respectively selects a positive class sample and a certain distance as the center and the radius; secondly we calculate the number of negative class samples involved in a circle area; finally we define the result as the sample's support degree. The larger support degree means the probability that the sample is determined to boundary sample is higher. In this case, the sample should be assigned a high selection probability. The area which is near the sample should produce more samples to strengthen the interface. On the contrary, if the probability is small, the sample should be assigned a small selection probability. So we can effectively avoid the blindness of oversampling.

The fundamentals of SDSMOTE as follows:

S1: Set the number of positive class samples which should be synthesized to Num;

S2: The number of positive class is assumed to be m, the number of negative class is assumed to be n. Select a positive class sample X_i randomly, and calculate the sum S_i of the distance between X_i with every negative class sample y_j according to the formula $S_i = \sum_{j=1}^{n} \sqrt{\|x_i - x_j\|^2}$;

S3: Calculate the sum S of all S_i according to the formula $S = \sum_{i=0}^{m} S_i$;

S4: Calculate the average distance between positive class samples with negative class samples according to the formula $S_{ave} = \frac{s}{m \times n}$;

S5: Set S_{ave} as distance parameter, and respectively select each positive class sample and a certain distance as the center and the radius to draw a circle, and then calculate the number of negative class samples in the region as the support degree k_i . The larger support degree means the sample should be assigned a high selection probability. On the contrary, the sample should be assigned a small selection probability;

S6: According to the selection probability of positive class samples, we can select some positive class samples. Use

SMOTE algorithm on the selected samples, then search its neighbors to synthesize new samples;

S7: Adding the new positive samples to data-set to participate in training and testing.

Our method identifies boundary samples of positive class and selects them discriminatively by support degree. To a certain extent, it avoids the blindness of current method in synthesizing new samples effectively, and makes over-sampling more targeted, and then enhance the classification of the positive class.

IV. EXPERIMENTS AND ANALYSIS

A. Evaluation criteria in imbalanced data classification

In traditional classification problem, people generally use accuracy as the evaluation criteria of classifier performance. But for imbalanced data, measuring the performance of the algorithm by correct ratio is inappropriate, because the classification interface tends to positive class.

In the classification process, multi-classification problem can also be converted into a binary classification problem, so we mainly focus on two-class imbalanced scenario. The evaluation criteria mentioned in this article are based on two-class scenario. The evaluation criteria which are commonly used in imbalanced scenarios are: $F\text{-}value$, $G\text{-}mean$, ROC (Receiver Operating Characteristic) curve, AUC (Area Under the ROC Curve), Etc.

The confusion matrix commonly used in traditional classification algorithm is shown in TABLE I.Usually we call majority classes as negative class (Negative), the minority class as positive class (Positive).

TABLE I. THE CONFUSION MATRIX FOR TWO-CLASS PROBLEM

	Classfied Positive	**Classfied Negative**
True Positive	TP (True Positive)	FN (False Negative)
True Negative	FP (False Positive)	TN (True Negative)

Where, TP denotes the number of true positives, TN denotes the number of true negatives, FN denotes the number of false negatives, FP denotes the number of false positives.

(1) $acc = (TP + TN) / (TP + FN + TN + FP)$

Acc is the ratio between the number of samples correctly classified and the number of the whole samples. The higher acc is, the better the performance of the algorithm is. It is a measure of the overall performance of classifiers, so it does not apply to the imbalanced data classification.

(2) $precision = TP / (TP + FP)$

Precision is the ratio between the number of true positives and the number of the positive samples in classification process.

(3) $recall = TP / (TP + FN)$

Recall is the value between the number of true positives and the number of positive samples, i.e. the accuracy of the minority class.

(4) $\mathrm{F-value} = \dfrac{(1+\beta^2) \times recall \times precision}{\beta^2 \times recall + precision}$

Where, parameter β is adjustable, usually set to 1. F-value is the combination of precision and recall. Only if both of the values are large, the value of $\mathrm{F-value}$ would be large. So it can be used as effective evaluation criteria in imbalanced dataset classification problem.

(5) $\mathrm{G-Mean} = \sqrt{\dfrac{TP}{TP+FN} \times \dfrac{TN}{TN+FP}}$

G-mean reflects the ability of classification algorithms in balancing the two classes. Larger value indicates better classification ability of the two classes.

(6) ROC curve, AUC value

ROC curve compares the performance of different classifiers by two- dimensional curve, but it cannot quantitatively evaluate the performance of the classifiers. AUC (the Area Under the Curve) is the area between ROC curve and the axis. We usually use AUC to evaluate the performance of the classifiers. The larger AUC is, the better classification performance is.

B. Experimental results and analysis

In order to evaluate the effectiveness of SDSMOTE in imbalanced data classification, we choose four public data-sets which have different proportions of positive class to negative class. The glass、blood、wine data-sets are from the UCI Machine Learning Repository, JM1 is from NASA standard data-sets. The information of data-sets are shown in TABLE II.

TABLE II. INFORMATION OF DATA-SETS

Name	**The number of samples**	**The number of attributes**	**Class *ratio* (*positive:* negative)**	**Positive Ratio (%)**
glass	214	10	29:185	13.6
JM1	7782	22	1672:6110	21.5
blood	748	5	178:570	23.8
wine	178	14	48:130	26.9

The experiment uses some public data-sets to compare classification performances of SMOTE + C4.5, SMOTE + AdaBoost, SMOTE + Bagging methods with classification performances of SDSMOTE + C4.5, SDSMOTE + AdaBoost, SDSMOTE + Bagging methods on weka3.6.0 platform. C4.5 decision tree algorithm achieved by J48 classifier. The neighborhood value k of SMOTE algorithm is set to 5.

To increase the objectivity of experimental data, all experiments use ten-fold cross-validation, i.e. the data is divided into ten parts(nine parts for training and one for testing), then use the average of the 10 results as the result of the test ten-fold cross-validation. TABLE III and TABLE IV respectively list $F\text{-}value$ and AUC of the 6 methods on the 4 data-sets. Fig.1, Fig.2 and Fig.3 show the comparison of $F\text{-}value$ values of the six methods on the four data-sets intuitively. Fig.4, Fig.5 and Fig.6 show the comparison of AUC values of the six methods on the four data-sets intuitively.

TABLE III. F-VALUE VALUES OF THE SIX METHODS ON THE FOUR DATA-SETS

Method	glass	JM1	blood	wine
SMOTE +C4.5	0.92	0.649	0.602	0.963
SDSMOTE+ C4.5	0.948	0.651	0.744	0.969
SMOTE+AdaBoost	0.898	0.414	0.577	0.979
SDSMOTE+AdaBoost	0.957	0.638	0.667	0.984
SMOTE+Bagging	0.93	0.679	0.641	0.979
SDSMOTE+Bagging	0.948	0.695	0.737	0.984

TABLE IV. AUC VALUES OF THE SIX METHODS ON THE FOUR DATA-SETS

Method	glass	JM1	blood	wine
SMOTE +C4.5	0.912	0.78	0.751	0.973
SDSMOTE+ C4.5	0.95	0.785	0.839	0.981
SMOTE+AdaBoost	0.967	0.712	0.75	0.999
SDSMOTE+AdaBoost	0.97	0.783	0.834	0.999
SMOTE+Bagging	0.947	0.834	0.805	0.998
SDSMOTE+Bagging	0.95	0.841	0.859	0.999

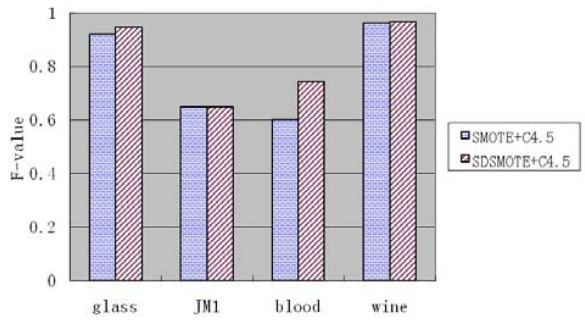

Fig. 1. F-value values of SMOTE +C4.5 and SDSMOTE+ C4.5 on the four data-sets

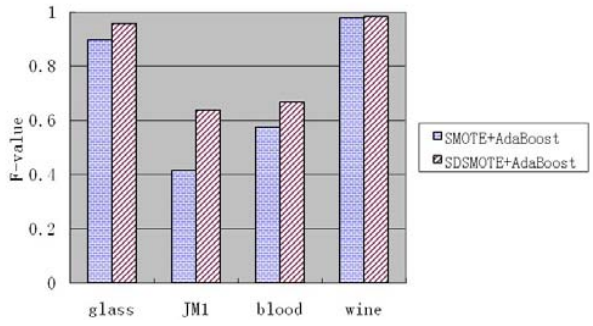

Fig. 2. F-value values of SMOTE+AdaBoost and SDSMOTE+AdaBoost on the four data-sets

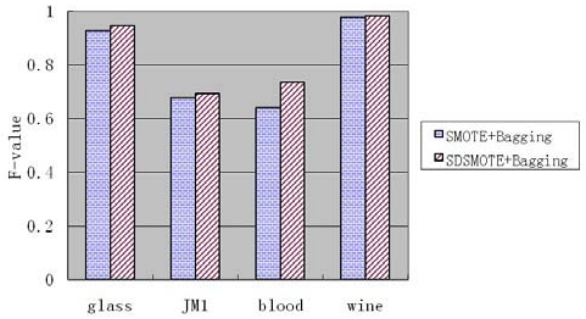

Fig. 3. F-value values of SMOTE+Bagging and SDSMOTE+Bagging on the four data-sets

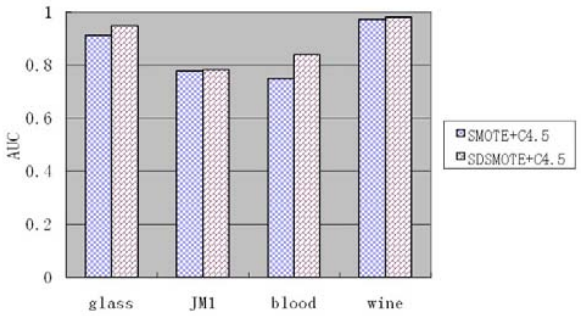

Fig. 4. AUC values of SMOTE +C4.5 and SDSMOTE+ C4.5 on the four data-sets

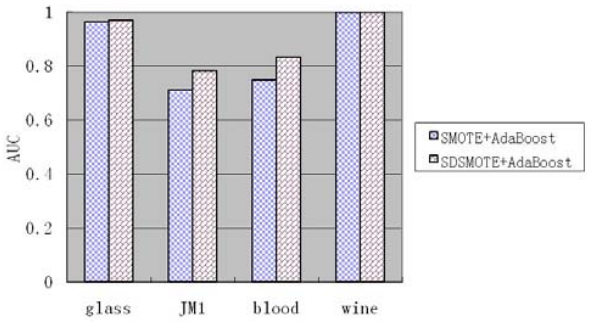

Fig. 5. AUC values of SMOTE+AdaBoost and SDSMOTE+AdaBoost on the four data-sets

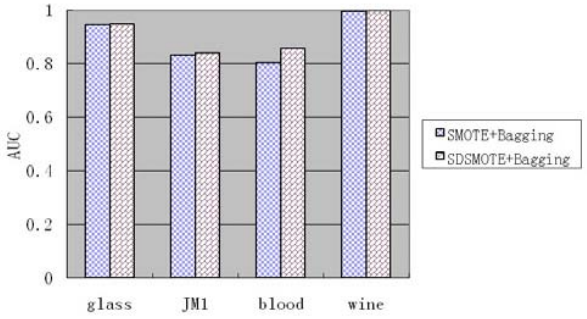

Fig. 6. AUC values of SMOTE+Bagging and SDSMOTE+Bagging on the four data-sets

TABLE III and TABLE IV list the results obtained in the following manner: use SMOTE and SDSMOTE to preprocess the four data-sets, then use C4.5, AdaBoost and Bagging algorithms respectively to deal with the data-sets. The table shows that the ability to identify the positive class and the classification performance of the whole data-sets have been improved by using the two oversampling methods, whereas the classification performance after SDSMOTE treated is better. Preprocessing the data-sets by SDSMOTE, increase boundary samples of positive class and obtain a relatively high F-value and AUC values. From Fig.1, Fig.2, Fig.3, Fig.4, Fig.5 and Fig.6,we can see the classification performance after SD-SMOTE treated is obviously superior to the classification performance after SMOTE treated. Base on the above results, we can draw the conclusion that using SDSMOTE can effectively balance the imbalanced data-set and improve the ability of identifying the positive class during the classifiers deal with data-sets.

V. CONCLUSION

There are a lot of imbalance data-sets in different application domains. In an imbalanced scenario, the traditional classification algorithms are biased toward the negative class because it is easier to learn. So the traditional classification algorithms cannot achieve ideal effects because positive class is the more valuable class. SDSMOTE method which we proposed in the paper, use support degree as the guidance to identify positive class boundary samples and synthesize new samples. It can not only avoid the disadvantages of the current methods which generate new samples blindly, but also make oversampling in imbalanced data preprocessing more targeted and improve the ability to enhance the classification of the positive class. Experimental results show that, SDSMOTE have a high recognition ratio of positive class samples to the whole data-set compared with SMOTE algorithm. Similar to other over-sampling technologies, SDSMOTE method also has the shortcoming that both operation time and storage space have increased due to the new samples addition.

REFERENCES

[1] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Francisco Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," Pattern Recognition , 2013, pp. 3460–3471.

[2] Batista G E A P A, Prati R C, Monard M C, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol.6, 2004, pp. 20-29.

[3] GAO Jia-Wei, LIANG Ji-Ye, "Research and Advancement of Classification Method of Imbalanced Data Sets," Computer Sciense, vol.35, 2008, pp. 10-13.

[4] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers , 2000.

[5] METHA M, AGRAWAL R, RISSANEN J, "S LIQ: A Fast Scalable Classifier for Data Mining," Lecture Notes in Computer Sci. P roc . of the 5th Int. Conf. on Extending Database Tech., 1996, pp. 18 -33

[6] C.Li, "Classifying Imbalanced Data Using A Bagging Ensemble Variation (BEV)," Proceedings of the 45th annual southeast regional conference, March 23-24, 2007, Winston-Salem, North Carolina.

[7] J.Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution," Proceedings of the 8th Conference on AI in Medicine Europe: Artificial. 2001, pp. 63-66.

[8] N.V.Chawla, K.W.Bowyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 2002, Vol. 16, pp. 341-378.

[9] C.Drummond, R.C.Holte, 'C4.5, Class Imbalance and Cost Sensitivity: Why Under-Sampling beats Over-Sampling," Proceedings of the ICML'03 Workshop on Learning from, 2003.

[10] Chris Seiffert, Taghi M.Khoshgoftaar, Jason Van Hulse, Amri Napolitano, "RUSBoost:A Hybrid Approach to Alleviating Class Imbalance," IEEE TRANSACTIONS ON SYSTEM,MAN,AND CYBERNETICS-PART A:SYSTEMS AND HUMANS, vol.40, 2010, pp. 185-197.

[11] G.E.Batista, R.C.Prati, M.C.Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol.6, 2004, pp. 20-29.