Margin-Based Over-Sampling Method for Learning From Imbalanced Datasets

Xiannian Fan, Ke Tang ⋆, and Thomas Weise

Nature Inspired Computational and Applications Laboratory
School of Computer Science and Technology
University of Science and Technology of China
Hefei, China, 230027
xnfan@mail.ustc.edu.cn, {ketang, tweise}@ustc.edu.cn

Abstract. Learning from imbalanced datasets has drawn more and more attentions from both theoretical and practical aspects. Over-sampling is a popular and simple method for imbalanced learning. In this paper, we show that there is an inherently potential risk associated with the over-sampling algorithms in terms of the large margin principle. Then we propose a new synthetic over sampling method, named Margin-guided Synthetic Over-sampling (MSYN), to reduce this risk. The MSYN improves learning with respect to the data distributions guided by the margin-based rule. Empirical study verities the efficacy of MSYN.

Keywords: imbalance learning, over-sampling, over-fitting, large margin theory, generalization

1 Introduction

Learning from imbalanced datasets has got more and more emphases in recent years. A dataset is imbalanced if its class distributions are skewed. The class imbalance problem is of crucial importance since it is encountered by a large number of real world applications, such as fraud detection [2], the detection of oil spills in satellite radar images [3], and text classification [4]. In these scenarios, we are usually more interested in the minority class instead of the majority class. The traditional data mining algorithms have a poor performance due to the fact that they give equal attention to the minority class and the majority class.

One way for solving the imbalance learning problem is to develop "imbalanced data oriented algorithms" that can perform well on the imbalanced datasets. For example, Wu et al. proposed class boundary alignment algorithm

⋆ corresponding author

which modifies the class boundary by changing the kernel function of SVMs [5]. Ensemble methods were used to improve performance on imbalance datasets [6]. In 2010, Liu et al. proposed the Class Confidence Proportion Decision Tree (CCPDT) [7]. Furthermore, there are other effective methods such as cost-based learning [8] and one class learning [9].

Another important way to improve the results of learning from imbalanced data is to modify the class distributions in the training data by over-sampling the minority class or under-sampling the majority class [10]. The simplest sampling methods are Random Over-Sampling (ROS) and Random Under-Sampling (RUS). The former increases the number of the minority class instances by duplicating the instances of the minority, while the latter randomly removes some instances of the majority class. Sampling with replacement has been shown to be ineffective for improving the recognition of minority class significantly. [10][11]. Chawla et al. interpret this phenomenon in terms of decision regions in feature space and proposed the Synthetic Minority Over-Sampling Technique (SMOTE) [12]. There are also many other synthetic over-sampling techniques, such as Borderline-SMOTE [13] and ADASYN [14]. To summarize, under-sampling methods can reduce useful information of the datasets; over-sampling methods may make the decision regions of the learner smaller and more specific, thus may cause the learner to over-fit.

In this paper, we analyze the performance of over-sampling techniques from the perspective of the large margin principle and find that the over-sampling methods are inherently risky from this perspective. Aiming to reduce this risk, we propose a new synthetic over-sampling method, called Margin-guided Synthetic Over-Sampling (MSYN). Our work is largely inspired by the previous works in feature selection using the large margin principle [15] [16] and problems of over-sampling for imbalance learning [17]. The empirical study revealed the effectiveness of our proposed method.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 presents the margin-based analysis for over-sampling. Then in Section 4 we propose the new synthetic over-sampling algorithm. In Section 5, we test the performance of the algorithms on various machine learning benchmarks datasets. Finally, the conclusion and future work are given in Section 6.

2 Related Works

We use A to denote a dataset of n instances $A = \{\boldsymbol{a_1}, ..., \boldsymbol{a_n}\}$, where $\boldsymbol{a_i}$ is a real-valued vector of dimension m. Let $A_P \subset A$ denote the minority class instances, $A_N \subset A$ denote the majority class instances.

Over-sampling techniques augment the minority class to balance between the numbers of the majority and minority class instances. The simplest over-sampling method is ROS. However, it may make the decision regions of the majority smaller and more specific, and thus can cause the learner to over-fit [17].

Chawla et al. over-sampled the minority class with their SMOTE method, which generates new synthetic instances along the line between the minority in-

stances and their selected nearest neighbors [12]. Specifically, for the subset A_P, they consider the k-nearest neighbors for each instances $\boldsymbol{a_i} \in A_p$. For some specified integer number k, the k-nearest neighbors are define as the k elements of A_P, whose Euclidian distance to the element a_i under consideration is the smallest. To create a synthetic instance, one of the k-nearest neighbors is randomly selected and then multiplied by the corresponding feature vector difference with a random number between $[0, 1]$. Take a two-dimensional problem for example:

$$\boldsymbol{a_{new}} = \boldsymbol{a_i} + (\boldsymbol{a_{nn}} - \boldsymbol{a_i}) \times \delta$$

where $\boldsymbol{a_i} \in A_P$ is the minority instance under consideration, $\boldsymbol{a_{nn}}$ is one of the k-nearest neighbors from the minority class, and $\delta \in [0, 1]$. This leads to generating a random instance along the line segment between two specific instances and thus effectively forces the decision region of the minority class to become more general [12]. The advantage of SMOTE is that it makes the decision regions larger and less specific [17].

Borderline-SMOTE focuses the instances on the borderline of each class and the ones nearby. The consideration behind it is: the instances on the borderline (or nearby) are more likely to be misclassified than the ones far from the borderline, and thus more important for classification. Therefore, Borderline-SMOTE only generates synthetic instances for those minority instances closer to the border while SMOTE generates synthetic instances for each minority instance. ADASYN uses a density distribution as a criterion to automatically decide the number of synthetic instances that need to be generated for each minority instance. The density distribution is a measurement of the distribution of the weights for different minority class instances according to their level of difficulty in learning. The consideration is similar to the idea of AdaBoost [18]: one should pay more attention to the difficult instances. In summary, either Borderline-SMOTE or ADASYN improves the performance of over-sampling techniques by paying more attention on some specific instances. They, however, did not touch the essential problem of the over-sampling techniques which causes over-fitting.

Different from the previous work, we resort to margins to analyze the problem of over-sampling, since margins offer a theoretic tool to analyze the generalization ability. Margins play an indispensable role in machine learning research. Roughly speaking, margins measure the level of confidence a classifier has with respect to its decision. There are two natural ways of defining the margin with respect to a classifier [15]. One approach is to define the margin as the distance between an instance and the decision boundary induced by the classification rule. Support Vector Machines are based on this definition of margin, which we refer to as sample margin. An alternative definition of the margin can be the Hypothesis Margin; in this definition the margin is the distance that the classifier can travel without changing the way it labels any of the sample points [15].

3 Large Margin Principle Analysis for Over-Sampling

For prototype-based problems (e. g. the nearest neighbor classifer), the classifier is defined by a set of training points (prototypes) and the decision boundary

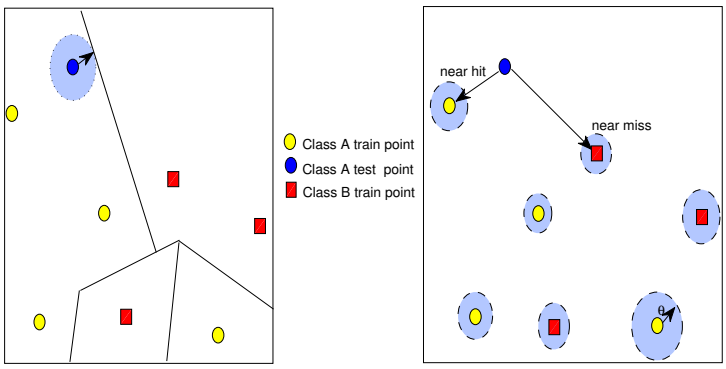

Fig. 1. Two types of margins in terms of the Nearest Neighbor Rule. The toy problem involves class A and class B. Margins of a new instance (the blue circle), which belongs to class A, are shown. The sample margin 1(left) is the distance between the new instance and the decision boundary (the Voronoi tessellation). The hypothesis margin 1(right) is the largest distance the sample points can travel without altering the label of the new instance. In this case it is half the difference between the distance to the nearest miss and the distance to the nearest hit.

is the Voronoi tessellation [19]. The sample margin in this case is the distance between the instance and the Voronoi tessellation. Therefore it measures the sensitivity to small changes of the instance position. The hypothesis margin R for this case is the maximal distance such that the following condition holds: if we draw a sphere with radius R around each prototype, any change of the location of prototypes inside their sphere will not change the assigned labels. Therefore, the hypothesis margin measures the stability to small changes in the prototypes locations. See Figure 1 for illustration.

Throughout this paper we will focus on the margins for the Nearest Neighbor rule (NN). For this special case, it is proved the following results [15]:

1. The hypothesis-margin lower bounds the sample-margin

2. It is easy to compute the hypothesis-margin of an instance $\boldsymbol{x}$ with respect to a set of instances A by the following formula:

$$\theta_A(x) = \frac{1}{2}(||\boldsymbol{x} - nearestmiss_A(\boldsymbol{x})|| - ||\boldsymbol{x} - nearesthit_A(\boldsymbol{x})||) \tag{1}$$

where $nearesthit_A(\boldsymbol{x})$ and $nearestmiss_A(\boldsymbol{x})$ denote the nearest instance to $\boldsymbol{x}$ in dataset A with the same and different label, respectively.

In the case of the NN, we can know that the hypothesis margin is easy to calculate and that a set of prototypes with large hypothesis margin then it has large sample margin as well [15].

Now we consider the over-sampling problem using the large margin principle. When adding a new minority class instance x, we consider the difference of the

overall margins for the minority class:

$$\Delta_P(\boldsymbol{x}) = \sum_{\boldsymbol{a} \in A_P} (\theta_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a}) - \theta_{A\backslash \boldsymbol{a}}(\boldsymbol{a})) \tag{2}$$

where $A\backslash \boldsymbol{a}$ denotes the dataset excluding $\boldsymbol{a}$ from the dataset A, and $A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}$ denotes the union of $A\backslash \boldsymbol{a}$ and $\{\boldsymbol{x}\}$.

For each instance $\boldsymbol{a} \in A_P$, $||\boldsymbol{a} - nearestmiss_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a})|| = ||\boldsymbol{a} - nearestmiss_{A\backslash \boldsymbol{a}}(\boldsymbol{a})||$ and $||\boldsymbol{a} - nearesthit_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a})|| \leq |\boldsymbol{a} - nearesthit_{A\backslash \boldsymbol{a}}(\boldsymbol{a})||$. From Eq. (1), it follows that $\Delta_P(\boldsymbol{x}) \geq 0$. We call $\Delta_P(\boldsymbol{x})$ the margin gain for the minority class.

Further, the difference of the overall margins for majority class is:

$$\Delta_N(\boldsymbol{x}) = \sum_{\boldsymbol{a} \in A_N} (\theta_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a}) - \theta_{A\backslash \boldsymbol{a}}(\boldsymbol{a})) \tag{3}$$

for each instance $\boldsymbol{a} \in A_N$, $||\boldsymbol{a} - nearestmiss_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a})|| \leq ||\boldsymbol{a} - nearestmiss_{A\backslash \boldsymbol{a}}(\boldsymbol{a})||$ and $||\boldsymbol{a} - nearesthit_{A\backslash \boldsymbol{a} \cup \{\boldsymbol{x}\}}(\boldsymbol{a})|| = ||\boldsymbol{a} - nearesthit_{A\backslash \boldsymbol{a}}(\boldsymbol{a})||$. From Eq. (1), it follows that $\Delta_N(\boldsymbol{x}) \leq 0$. We call $-\Delta_N(\boldsymbol{x})$ the margin loss for the majority class.

In summary, it is shown that the over-sampling methods are inherently risky from the perspective of the large margin principle. The over-sampling methods, such as SMOTE, will enlarge the nearest-neighbor based margins for the minority class while may decrease the nearest neighbor based margins for the majority class. Hence, over-sampling will not only bias towards the minority class but also may be detrimental to the majority class. We cannot eliminate these effects when adopting over-sampling for imbalance learning completely, but we can seek methods to optimize the two parts.

In the simplest way, one can maximize the margins for the minority class and ignore the margins loss for the majority class, i.e., the following formula:

$$f_1 = -\Delta_P(\boldsymbol{x}) \tag{4}$$

Alternatively, one may also minimize the margins loss for the majority class, which is

$$f_2 = -\Delta_N(\boldsymbol{x}) \tag{5}$$

One intuitive method is to seek a good balance between maximizing the margins gain for the minority class and minimizing the margins loss for the majority class. This can be conducted by minimizing Eq. (6):

$$f(\boldsymbol{x})_3 = \frac{-\Delta_N(\boldsymbol{x})}{\Delta_P(\boldsymbol{x}) + \varepsilon}, \varepsilon > 0 \tag{6}$$

where ε is a positive constant to ensure that the denominator of Eq. (6) to be non-zero.

4 The Margin-guided Synthetic Over-Sampling Algorithm

In this section we apply the above analysis to the over-sampling techniques. Without loss of generality, our algorithm is designed on the basis of SMOTE.

The general idea behind it, however, can also be applied to any other over-sampling technique

Algorithm 1: MSYN

Input: Training set X with n instances $(\boldsymbol{a_i}, y_i), i = 1, ..., n$ where $\boldsymbol{a_i}$ is an instance in the m dimensional feature space, and y_i belongs to $Y = \{1, -1\}$ is the class identity label associated with $\boldsymbol{a_i}$, Define m_P and m_N as the number of the minority class instances and the number of the majority class instances, respectively. Therefore, $m_P < m_N$. BIN is the set of synthetic instances, which is initialized as empty. Parameter: $Pressure$.

1 Calculate the number of synthetic instances that need to be generated for the minority class: $G = (m_N - m_P) * Pressure$;

2 Calculate the number of synthetic instances that needed to be generated for each minority example $\boldsymbol{a_i}$:

$$g_i = \frac{G}{m_P}$$

3 **for** *each minority class instances* $\boldsymbol{a_i}$ **do**

4 **for** $j \leftarrow 1$ **to** g_i **do**

5 Randomly choose one minority instance, $\boldsymbol{a_{zi}}$, from the k nearest neighbors for the instance $\boldsymbol{a_i}$;

6 Generate the synthetic instances $\boldsymbol{a_s}$ using the technique of SMOTE;

7 Add $\boldsymbol{a_s}$ to BIN

8 sort the synthetic instances in BIN according to the their values of Eq. (6);

9 return $(m_N - m_P)$ instances who have the minimum $(m_N - m_P)$ values of Eq. (6).

Based on the analysis in the previous section, Eq. (6) is employed to decide whether a new synthetic instance is good enough to be added into the training dataset. Our new Margin-guided Synthetic Over-Sampling algorithm, MSYN for short, is given in Algorithm 1. The major focus of MSYN is to use margin-based guideline to select the synthetic instances. $Pressure \in \mathbb{N}$, a natural number, is a parameter for controlling the selection pressure. In order to get $(m_N - m_P)$ new synthetic instances, we first create $(m_N - m_P) * Pressure$ new instances, then we only select top best $(m_N - m_P)$ new instances according to the values of Eq. (6) and discard the rest instances. This selection process implicitly decides whether an original minority instance is used to create a synthetic instances as well as how many synthetic instances will be generated, which is different from SMOTE since SMOTE generates the same number of synthetic instances for each original minority instances. Moreover, it is easy to see that computational complexity of MSYN is $O(n^2)$, which is mainly decided by calculating the distance matrix .

5 Experiment Study.

The Weka's C4.5 implementation [20] is employed in our experiments. We compare our proposed MSYN with SMOTE [12], ADASYN [14], Borderline-SMOTE [13] and ROS. All experiments were carried out using 10 runs of 10-fold cross-validation. For MSYN, the parameter *Pressure* is set to 10 and the ε can be any random positive real number; for other methods, the parameters are set as recommended in the corresponding paper.

To evaluate the performance of our approach, experiments on both artificial and real datasets have been performed. The former is used to show the behavior of the MSYN on known data distributions while the latter is used to verify the utility of our method when dealing with real-world problems.

5.1 Synthetic Datasets

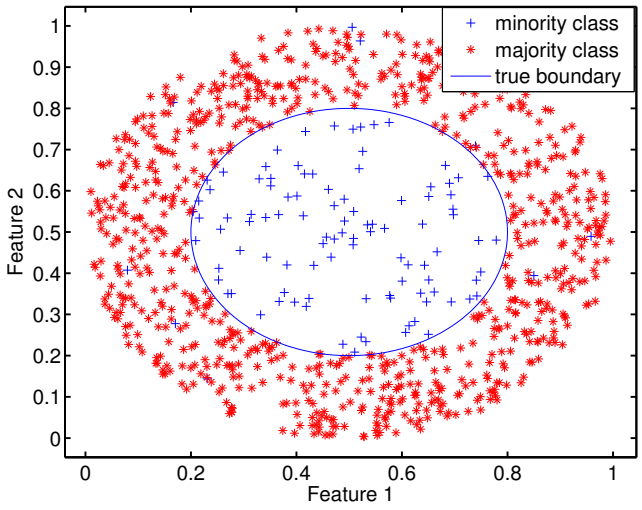

Fig. 2. The distribution of the dataset Concentric with noise.

This part of our experiments focuses on synthetic data to analyze the characteristics of the proposed MSYN. We used the dataset Concentric from the ELENA project [21]. The Concentric dataset is a two-dimensional uniform concentric circular distributions problem with two classes. The instances of minority class uniformly distribute within a circle of radius 0.3 centered on (0.5, 0.5). The points of majority class are uniformly distribute within a ring centered on (0.5, 0.5) with internal and external radius respectively to 0.3 and 0.5.

In order to investigate the problem of over-fitting to noise, we modify the dataset by randomly flipping the labels of 1% instances, as shown in Figure 2.

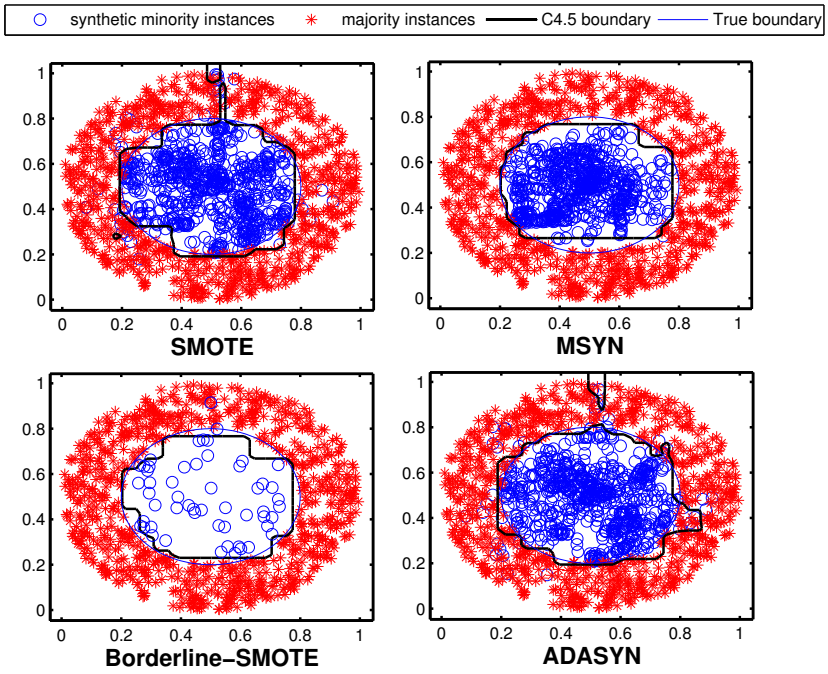

Fig. 3. The synthetic instances and the corresponding C4.5 decision boundary after processing by SMOTE, MSYN, Borderline-SMOTE, ADASYN, respectively.

In order to show the performance of the various synthetic over-sampling techniques, we sketch them in Figure 3. The new synthetic instances created by each over-sampling method, the original majority instances and the corresponding C4.5 decision boundary are drawn. From Figure 3, we can see that MSYN shows good performance in the presence of noise while SMOTE and ADASYN suffer greatly from over-fitting the noise. MSYN generates no noise instances. This can be attributed to the fact that the margin-based Eq. (6) contains the information of the neighboring instances, and this information helps to decrease the influence of noise. Both SMOTE and ADASYN generate a large number of noise instances and their decision boundary is greatly influenced. Borderline-SMOTE generates a small number of noise instances and its decision boundary is slightly influenced. Furthermore, Borderline-SMOTE pays little attention to interior instances and creates only a few of synthetic instances.

5.2 Real World Problems

We test the algorithms on ten datasets from the UCI Machine Learning Repository [22]. Information about these datasets is summarized in Table 1, where *num* is the size of the dataset, *attr* is the number of features, *min%* is the ratio of the number of minority class number to NUM.

Table 1. Summary of the DataSets

Datasets	*num*	*attr*	*min%*
Abalone	4177	8	9.36%
Contraceptive	1473	9	22.61%
Heart	270	9	29.28%
Hypothyroid	3163	8	34.90%
Ionosphere	351	34	35.90%
Parkinsons	195	22	24.24%
Pima	768	8	34.90%
Spect	367	19	20.65%
Tic-tac-toe	958	9	34.66%
Transfusion	748	4	31.23%

Instead of using the overall classification accuracy, we uadopt metrics related to Receiver Operating Characteristics (ROC) curve [23] to evaluate the compared algorithms, because traditional overall classification accuracy may not be able to provide a comprehensive assessment of the observed learning algorithms in case of class imbalanced datasets [4]. Specifically, we use the AUC [23] and F-Measure [24] to evaluate the performance. We apply the Wilcoxon signed rank test with a 95% confidence level on each dataset to see whether the difference between the compared algorithms is statistically significant.

Table 2 and Table 3 show the AUC and F-Measure for the datasets, respectively. The results of Table 2 reveal that MSYN wins against SMOTE on nine out of ten datasets, beats ADASYN on seven out of ten datasets, outperforms ROS on nine out of ten datasets, and wins against Borderline-SMOTE on six out of ten datasets. The results of Table 3 show that MSYN wins against SMOTE on seven out of ten datasets, beats ADASYN on six out of ten datasets, beats ROS on six out of ten datasets, and wins against Borderline-SMOTE on six out of ten datasets. The comparisons reveal that MSYN outperforms the other methods in terms of both AUC and F-measure.

6 Conclusion and Future work

This paper gives an analysis of over-sample techniques from the viewpoint of the large margin principle. It is shown that over-sampling techniques will not only bias towards the minority class but may also bring detrimental effects to the classification of the majority class. This inherent dilemma of over-sampling cannot be entirely eliminated, but only reduced. We propose a new synthetic over-sampling method to strike a balance between the two contradictory objectives. We evaluate our new method on a wide variety of imbalanced datasets using different performance measures and compare it to the established over-sampling methods.

Table 2. Result in terms of AUC in the experiments performs on real datasets. For SMOTE, ADAYSN, ROS and Borderline-SMOTE, if the value is underlined, MSYN has better performance than that method; if the value is starred, MSYN exhibits lower performance compared to that method; if the value is in normal style it means that the corresponding method does not perform significantly different from MSYN according to the Wilcoxon signed rank test. The row W/D/L Sig. shows the number of wins, draws and losses of MSYN from the statistical point of view.

Dataset	MSYN	SMOTE	ADASYN	ROS	Borderline-SMOTE
Abalone	0.7504	0.7402	0.7352	0.6708	0.7967*
Contraceptive	0.6660	0.6587	0.6612	0.6055	0.6775*
Heart	0.7909	0.7862	0.7824	0.7608	0.7796
Hypothyroid	0.9737	0.9652	0.9655	0.9574	0.9653
Ionosphere	0.8903	0.8731	0.8773	0.8970*	0.8715
Parkinsons	0.8248	0.8101	0.8298*	0.7798	0.8157
Pima	0.7517	0.7427	0.7550	0.7236	0.7288
Spect	0.7403	0.7108	0.7157	0.6889	0.7436
Tic-tac-toe	0.9497	0.9406	0.9391	0.9396	0.9456
Transfusion	0.7140	0.6870	0.6897	0.6695	0.6991
W/D/L Sig.	N/A	9/1/0	7/2/1	9/0/1	6/2/2

Table 3. Result in terms of F-measure in the experiments performs on real datasets. For SMOTE, ADAYSN, ROS and Borderline-SMOTE, if the value is underlined, MSYN has better performance than that method; if the value is starred, MSYN exhibits lower performance compared to that method; if the value is in normal style it means that the corresponding method does not perform significantly different from MSYN according to the Wilcoxon singed rank test. The row W/D/L Sig. shows the number of wins, draws and losses of MSYN from the statistical point of view.

Dataset	MSYN	SMOTE	ADASYN	ROS	Borderline-SMOTE
Abalone	0.2507	0.3266*	0.3289*	0.3479*	0.3154*
Contraceptive	0.3745	0.4034*	0.4118*	0.4133*	0.4142*
Heart	0.7373	0.7305	0.7318	0.7151	0.7223
Hypothyroid	0.8875	0.8412	0.8413	0.8771	0.9054*
Ionosphere	0.8559	0.8365	0.8338	0.8668*	0.8226
Parkinsons	0.7308	0.6513	0.6832	0.6519	0.6719
Pima	0.6452	0.6435	0.6499	0.6298	0.6310
Spect	0.4660	0.4367	0.4206	0.4644	0.4524
Tic-tac-toe	0.8619	0.8465	0.8437	0.8556	0.8604
Transfusion	0.4723	0.4601	0.4507	0.4596	0.4664
W/D/L Sig.	N/A	7/1/2	6/2/2	6/1/3	6/1/3

The results support our analysis and indicate that the proposed method, MSYN, is indeed superior.

As a new sampling method, MSYN can be further extended along several directions. First of all, we investigate the performance of MSYN using C4.5. Based on the nearest neighbor margin, MSYN has a bias for the 1-NN. Some strategies, however, can be adopted to approximate the hypothesis margin for the other classification rules. For example, we can use the confidence of the classifiers' output to approximate the hypothesis margin. Thus we expect MSYN can be extended to work well with other learning algorithms, such as k-NN, RIPPER [29]. But solid empirical study is required to justify this expectation. Besides, ensemble learning algorithms can improve the accuracy and robustness of the learning procedure [26]. It is thus worthy of integrating MSYN with ensemble learning algorithms. Such an investigation can be conducted following the methodology employed in the work of SMOTEBoost [6], DataBoost-IM [27], BalanceCascade [28], etc.

Secondly, MSYN can be generalized to multiple-class imbalance learning as well. For each minority class i, a straightforward idea is to extend Eq. (6) to:

$$f_i(\boldsymbol{x}) = \frac{-\sum_{j \neq i} \Delta_{i,j}(\boldsymbol{x})}{\Delta_i(\boldsymbol{x}) + \varepsilon}, \varepsilon > 0 \tag{7}$$

where $\Delta_i(\boldsymbol{x})$ denotes the margin gain of minority class i by adding a new minority instance $\boldsymbol{x}$ ($\boldsymbol{x}$ belongs to class i), and $-\Delta_{i,j}(\boldsymbol{x})$ denotes the margin loss for class j by adding a new minority instance $\boldsymbol{x}$ ($\boldsymbol{x}$ belongs to class i). Then we create the synthetic instances for each minoirty class to make the number of them being equal to the number of the majority class, which has the maximum number of instances. However, this idea is by no means the only one. Extending a technique from binary to multi-class problems is usually non-trivial, and more in-depth investigation is necessary to seek the best strategy.

Bibliography

[1] Xiannian Fan, Ke Tang, and Thomas Weise: Margin-Based Over-Sampling Method for Learning From Imbalanced Datasets. In Joshua (Zhexue) Huang, Longbing Cao, and Jaideep Srivastava, eds., Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Part II (PAKDD'11), May 24–27, 2011, Shenzhen, Guangdong, China, volume 6635 of Lecture Notes in Computer Science (LNCS), Berlin, Germany: Springer-Verlag GmbH, doi:10.1007/978-3-642-20847-8_26

[2] Chan, P.K., Stolfo, S.J.: Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, (2001) 164-168

[3] Kubat, M., Holte R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning. 30(2) (1998) 195-215

[4] Weisis,G.M.: Mining with Rarity: A Unifying Framwork. SiGKDD Explorations. 6(1) (2004) 7-19

[5] Wu, G., Chang, E.Y.: Class-Boundary Alignment for Imbalanced Dataset Learning. Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC (2003)

[6] Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving Prediction of the Minority Class in Boosting. In Proceeding European Conf. Princoples and Practice of Knowledge Discovery in Databases,Dubrovnik, Croatia, (2003) 107-119

[7] Liu, W., Chawla, S., Cieslak, D.A., Chawla, N.V.: A Robust Decision Tree Algorithm for Imbalanced Data Sets. SIAM International Conf. on Data Mining (2010)

[8] Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering. (2006) 63-77

[9] Raskutti, B., Kowalczyk, A.: Extreme re-balancing for SVMs: a case study. SIGKDD Explorations, 6(1) (2004) 60-69

[10] Japkowicz, N.: The Class Imbalance Problem: Significance and Strategies. In Proceeding of the 2000 International Conf. on Artificial Intelligence (ICAI'2000): Special Track on Inductive Learning Las Vegas, Nevada. (2000)

[11] Ling, C., Li, C.: Data Mining for Direct Marketing Problems and Solutions. In Proceeding of the Fourth International Conf. on Knowledge Discovery and Data Mining (KDD-98) New York, NY. (1998)

[12] Chawla, N.V., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research, 16 (2002) 321-357

[13] Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing. (2005) 878-887

[14] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, In Proceeding of International Conf. Neural Networks, (2008) 1322-1328

[15] Crammer, K., Gilad-Bachrach, R., Navot, A., Tishby, N.: Margin analysis of the LVQ algorithm. Advances in Neural Information Processing Systems (2003) 479-486

[16] Gilad-Bachrach, R., Navot, A., Tishby, N.: Margin based feature selection-theory and algorithms. in Proceeding of the twenty-first international conference on Machine learning (2004)

[17] He, H., Garcia, E.A.: Learning from Imbalance Data IEEE Transaction on Knowledge and Data Engineering, 21(9) (2009) 1263-1284

[18] Freund, Y., Schapire, R.: A desicion-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1) (1997) 119-139

[19] Bowyer, A.: Computing dirichlet tessellations. The Computer Journal, 24(2)(1981).

[20] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Record, 31(1) (2002) 76-77

[21] UCL machine learning group, `http://www.dice.ucl.ac.be/mlg/?page=Elena`

[22] Asuncion, A., Newman, D.: UCI machine learning repository (2007)

[23] Bradley A.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7) (1997) 1145-1159

[24] Van Rijsbergen, C.J.: Information Retrieval, Butterworths, London (1979)

[25] Wang, B.X., Japkowicz, N.: Imbalanced Data Set Learning with Synthetic Samples. Proc IRIS Machine Learning Workshop (2004)

[26] Dietterich, T.G.: Ensemble methods in machine learning. Lecture Notes in Computer Secience 1857 (2000) 1-15

[27] Guo, H., Viktor, H.L.: Learning from Imbalanced Data Sets with Boosting and Data Generation: the DataBoost-IM Approach. in SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets, 6(1) (2004) 30-39

[28] Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, 39(2) (2009) 539-550

[29] Cohen, W.: Fast Effective Rule Induction. In Proceeding of 12th International Conf. on Machine Learning, Lake Tahoe, CA. Morgan Kaufmann. (1995) 115-123

```
@inproceedings{FTW2011MBAOOSFIL,
  author    = {Xiannian Fan and Ke Tang and Thomas Weise},
  title     = {{Margin-Based Over-Sampling Method for Learning
               From Imbalanced Datasets}},
  booktitle = {Proceedings of the 15th Pacific-Asia Conference
               on Advances in Knowledge Discovery and Data Mining,
               Part II (PAKDD'11)},
  editor    = {Joshua (Zhexue) Huang and Longbing Cao and
               Jaideep Srivastava},
  publisher = {Berlin, Germany: Springer-Verlag GmbH},
  address   = {Shenzhen, Guangdong, China},
  series    = {{Lecture Notes in Computer Science (LNCS)}},
  volume    = {6635},
  pages     = {309--320},
  year      = {2011},
  month     = may # {~24--27, },
  doi       = {10.1007/978-3-642-20847-8_26},
  eiid      = {20112314036487},
  sciids    = {BCX89},
  sciwos    = {WOS:000311907700026},
},
```