Combination Approach of SMOTE and Biased-SVM for Imbalanced Datasets

WANG He-Yong

***Abstract*—A new approach to construct the classifiers from imbalanced datasets is proposed by combining SMOTE (Synthetic Minority Over-sampling Technique) and Biased-SVM (Biased Support Vector Machine) approaches. A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. The cost of misclassifying an abnormal (interesting) example into a normal example is often much higher than that of the reverse error. It was known as a means of increasing the sensitivity of a classifier to the minority class using SMOTE over-sampling in minority class. But in this paper, it gives a good means of increasing the sensitivity of a classifier to the minority class by using SMOTE approaches within support vectors. As for support vector over-sampling, this paper proposes two different over-sampling algorithms to deal with the support vectors being over-sampled by its neighbors from the k nearest neighbors, not only within the support vectors but also within the entire minority class. Some experimental results confirms that the proposed combination approach of SMOTE and Biased-SVM can achieve better classifier performance.**

I. INTRODUCTION

In the Internet the dissemination of the magnanimous electron text usually to present the data set about the category distribution is imbalanced, namely between the category the sample quantity possibly has the magnitude the disparity, which causes the classified effect not to be ideal. Imbalanced data learning has recently received considerable attention from the research and many real-world datasets. In the imbalanced datasets, most of the cases belong to a majority class and few cases belong to a minority, yet usually more interesting class. Therefore, the imbalanced data learning is problematic as traditional machine learning approaches fail to achieve satisfactory results due to the skewed class distribution.

There are two types of solutions for copping with the imbalanced datasets. The first type tries to increase the number of minority class examples (over-sampling) [1] or decrease the number of majority class examples (under-sampling) [2] in different ways. The second type adjusts the cost of error or decision thresholds in classification for the imbalanced data and tries to control the sensitivity of the classifier. For example, Joshi made an improvement on the boosting algorithm [3], G. Wu enhanced the SVM (Support Vector Machine) algorithm [4], K. Huang proposed the BMPM algorithm [5], X. Peng propsed the SOCP algorithm [6], and E. Seyda proposed the learning on the border method [7], and so on. Those approaches enhanced the accuracy of the classification of the imbalanced data in certain degree.

To improve the classification performance, this paper proposes a combined algorithm from SOMTE sampling on support vector as well as Biased-SVM techniques. First, the original data set is taken into account using Biased-SVM algorithm to determine the boundary. Next, the support vector sampling is carried on to deal with the minority class using SMOTE algorithm. Finally, the Biased-SVM is applied to boost the performance of the overall system. The adopted two different over-sampling algorithms for support vectors are over-sampled by its neighbors from the k nearest neighbors, not only within support vectors but also within its entire minority class. To verify the effort of the proposed algorithm, some simulation data are entered into the combined algorithm to check with their results. The proposed sampling approach presents better performance to the minority class datasets.

Manuscript received November 23, 2007. This work was supported in part by the National Natural Science Foundation of China under Grant 10501014.

WANG HE-YONG is College of E-Business, South China University of Technology Guangzhou 510006, China (corresponding author to provide phone: 86-20-39381128; e-mail: zsuwhy@ hotmail.com).

II. BIASED SVM

SVM was established as a successful approach for various machine learning tasks. The class imbalance issue has also been addressed in the literatures. Through empirical study, Wu et al. [8] reported when the data is highly imbalanced, the decision boundary determined by the training data is largely biased toward the minority class. As a result, the false rate that associates with the minority class might be higher than it really exists. To compensate such skewness, they proposed to enlarge the resolution around the decision boundary by revising their kernel functions. Furthermore, Veropoulos et al. [9] used pre-specified penalty constants on Lagrange multipliers for different classes. Different error costs for the positive C^+ and negative C^- classes are proposed. Thus the objective function becomes:

$$\min_{\omega,\omega_0} \frac{1}{2}\|\omega\|^2 + C^+ \sum_{\{i|y_i=+1\}} \xi_i + C^- \sum_{\{i|y_i=-1\}} \xi_i \qquad (1)$$

Specifically, the primal Lagrangian is:

$$L_p = \frac{1}{2}\omega^T\omega + C^+ \sum_{\{i|y_i=+1\}} \xi_i + C^- \sum_{\{i|y_i=-1\}} \xi_i - \sum_{i=1}^{n} a_i \left(y_i\left(\omega^T x_i + \omega_0\right) - 1 + \xi_i\right) - \sum_{i=1}^{n} r_i \xi_i \qquad (2)$$

978-1-4244-1821-3/08/$25.00©2008 IEEE

Authorized licensed use limited to: Korea University. Downloaded on November 27,2021 at 08:34:05 UTC from IEEE Xplore. Restrictions apply.

where $\alpha_i \geq 0$. It is straightforward to show that the dual formulation gives the same result as the SVM. The constraints on α_i then become:

$$C^+ \geq a_i \geq 0 \qquad if\ y_i = +1 \qquad (3)$$
$$and \quad C^- \geq a_i \geq 0 \qquad if\ y_i = -1$$

Furthermore, it is noted that $\xi_i > 0$, only when $a_i = C$. Therefore non-zero errors on minority support vectors will have larger a_i, while non-zero errors on majority support vectors may have smaller a_i. This net effect is that the boundary is pushed more towards the negative instances.

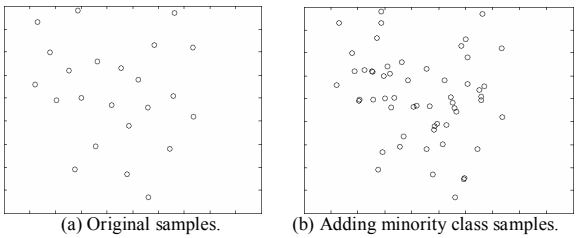
III. SMOTE Sampling Approach for Support Vector

Chawla [1] proposed the SMOTE algorithm in which the minority class is over-sampled by creating the "synthetic" examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing new synthetic examples joining any or all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic examples are generated in the following way: (1) Take the difference between the sample X under consideration and its nearest neighbor $\tilde{X}$ selected randomly from the k minority class nearest neighbors. (2) Take the difference between the feature vector (sample) under consideration of its nearest neighbor. (3) Multiply this difference by a random number between 0 and 1, and add it to the sample under consideration. This causes the selection of a random point between two specific samples. This approach can effectively force the decision region of the minority class to become more general to the dataset. And the new sample is defined as:

$$X_{new} = X + rand(0,1) \times (\tilde{X} - X) \qquad (4)$$

Accordingly, more synthetic examples can be obtained through repeating the above steps.

Figure 1 shows the performance of the adopted SMOTE. The left figure is the distribution of the original samples under consideration; while the right figure is the distribution after adding double synthetic minority class samples through SMOTE. From Figure 1, the synthetic instances can basically keep the distribution of original samples, and the synthetic instances cause the classifier to create larger and less specific decision regions.



(a) Original samples. (b) Adding minority class samples.

Figure 1: The performance of data distribution using SMOTE.

IV. Combination Approach of SOMTE and Biased-SVM

The main idea proposed in this paper is to entirely determine the boundary of datasets by the support vectors. Therefore SMOTE over sampling is only applied in the support vectors. Through this process, the performance of biased SVM can be enhanced in the imbalanced datasets. Moreover, this approach can reduce the processing time because the number of support vectors is bounded and becomes small. According to the proposed idea, the new algorithm can be expressed by flow chart as shown in Figure 2, and is detailed described as follows.

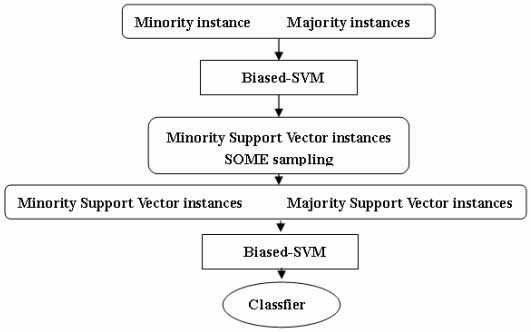


Figure 2: The combination flow chart of SOMTE and Biased-SVM

Step 1: Use the Biased-SVM to deal with the imbalanced training datasets, and record the support vectors.

Step 2: Sample the support vectors to improve the balanced degree between the majority class and the minority class by using SMOTE. In view of SMOTE, there are two different sampling algorithms to adopt, including the support vectors over-sampling by its neighbors from the k nearest neighbors, not only within the support vectors but also within the entire minority class.

Step 3: Use the Biased-SVM to deal with imbalanced datasets, and get an ultimate classifier.

In order to illustrate the effect of the SMOTE on support vector, an artificial imbalanced dataset in two-dimensional space is constructed for simulation. The original data is shown in Figure 3. Note in Figure 3, the boundary as horizontal line is defined in terms of "ideal boundary". The negative (majority) instances can be obtained and shown by ○ , which lie further away from the ideal boundary than the positive (minority) instances as shown by ＋. Some noisy data of the negative instances merely overlap onto the positive instances. The number of negative instances is higher than that of positive instances.

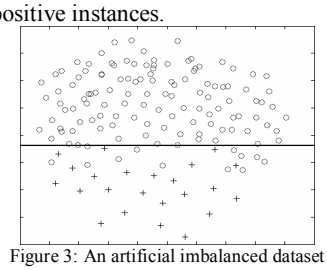


Figure 3: An artificial imbalanced dataset

Authorized licensed use limited to: Korea University. Downloaded on November 27,2021 at 08:34:05 UTC from IEEE Xplore. Restrictions apply.

A. SMOTE on the Support Vector within the Support Vectors

While the positive (minority) samples are carried out using SMOTE approach, this effect can be obtained from Figure 4. The support vectors of the minority (positive) class are shown by □, and the support vectors of the majority (negative) class are shown by ⊕. The minority (positive) class is sampled by creating "synthetic" examples as shown by □. From Figure 4, we know that the synthetic samples can be observed to create larger minority (positive) class samples, which makes the distribution of minority (positive) class much better.

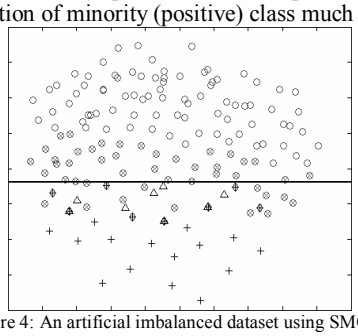


Figure 4: An artificial imbalanced dataset using SMOTE on the support vector within the support vectors

B. SMOTE on the Support Vector within the Minority Class

The support vectors of the minority class are usually quite few, therefore, when the over-sampling rate needs to enhance. It is possible to change the boundary distribution using the SMOTE approach in the interior of minority class. In order to remedy this problem, it is proposed to carry on the over-sampling on the support vector to make improvement. When k is computed around the nearest neighbors for each minority class sample, it is limited not only within the support vectors but also within the entire minority class. As a result, it can reduce the influence of the sample distribution. The performance of this approach is shown in Figure 5. Obviously, as the minority class is over-sampled by increasing amounts, the effect can be significant to add the samples along the boundary, and also reduce the influence on the boundary of the positive class.

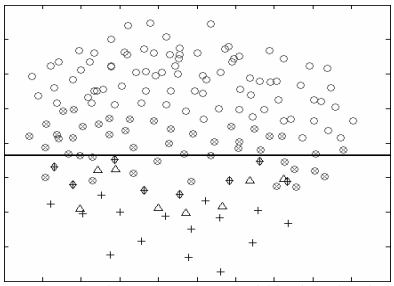


Figure 5: SMOTE on the support vector within the minority class.

V. EXPERIMENT

A. Experimental Data

Four datasets are used with strictly different classes of distributions. All of these datasets come from the UCI repository [10]. Four UCI datasets experimented are Phoneme, Pima, Satimage, and Vehicle. Both Phoneme and Pima are 2-class problem; while both Satimage and Vehicle are more class problem. In order to calculate conveniently, the more class problem is also transform into the 2-class problem. The Satimage dataset has 5 classes originally. The class by labeled 4 is chosen as the minority (positive) class and the rest of the classes are collapsed into one as the majority (negative) class. In the vehicle dataset, the class by labeled 1 is chosen as the minority (positive) class, and the rest of the classes are collapsed into one as the majority (negative) class. The detailed distributions of 4 datasets are shown in Table 1.

TABLE 1: DATA DISTRIBUTION

Datasets	Phoneme	Pima	Satimage	Vehicle
Attribution	5	8	36	18
Sample	5404	768	6435	846
Positive sample	1586	268	626	212
Negative sample	3818	500	5809	634
Imbalanced ratio	2.41	1.87	9.28	2.99
Composing	1:0	1:0	4:other	1:other

B. Evaluation Function

The performance of machine learning algorithms is typically evaluated by a confusion matrix as illustrated in Table 2 for a 2-class problem. The columns are the predicted class and the rows are the actual class. In the confusion matrix , TP represents the true positive samples; FP represents the false positive sample; TN represents the true negative samples; and FN represents the false negative samples.

TABLE 2: CONFUSION MATRIX.

	Predicted Positive	Predicted Negative	$\sum$
Actual Positive	TP	FN	n^+
Actual Negative	FP	TN	n^-

Generally speaking, the accuracy is defined as:

$$\text{Accuracy}=(\text{TP}+\text{TN})/(n^+ + n^-) \qquad (5)$$

But in order to evaluate the classifiers on highly imbalanced datasets, it is virtually unreasonable to aim at accuracy as a metric. This is because with an imbalance of 95 to 5, a classifier that classifies everything negative can be 95% accurate. However, it is completely useless as a classifier. Therefore, other approaches are adopted to evaluate the class of imbalanced datasets. In this paper, the $g-means$ metric is suggested to be defined as:

$$g = \sqrt{acc^+ \cdot acc^-} \qquad (6)$$

Authorized licensed use limited to: Korea University. Downloaded on November 27,2021 at 08:34:05 UTC from IEEE Xplore. Restrictions apply.

Where, $acc^+ = TP/n^+$ is the accuracy in positive (minority) samples and $acc^- = TN/n^-$ is the accuracy on the negative (majority) samples. And the geometric mean of the $g = \sqrt{acc^+ \cdot acc^-}$ reaches high value only if both $acc^+ = TP/n^+$ and $acc^- = TN/n^-$ are high and in equilibrium. The accuracy on positive examples can be increased at the cost of accuracy on negative examples. Moreover, the relation of the two quantities can be captured by curve, and a high acc^+ by a low acc^- will result in poor g.

C. Experimental Results

In this paper，a cross validation method [11] is used. All the samples are separated into 5 subsets randomly. They are the same rate of imbalance as the overall set, then 4 from 5 sets are chosen as the training sets, and accordingly the remaining as the testing set. The value of g is obtained in the testing set five times, that makes the mean of g be calculated as the overall set's value.

In this paper, a dataset processing is experienced by the standard support vector machine (Standard Support Vector Machine, STSVM) [12], Biased-SVM (BSVM) [9], using over-sampling and Biased-SVM (Sos+BVSM) [13] and SMOTE on support vector, not only in the support vector (SVSmt), but also in the entire minority class (SVSmt2).

Table 3 presents the prediction accuracy g for the five representative approachs. Biased-SVM is superior to SVM for almost imbalanced class, especially in processing the Satimage and Vehicle dataset which have a high imbalanced ratio. The combination algorithm of SMOTE and Biased-SVM improves the accuracy over SMOTE by 1%. The proposed SMOTE on support vector based on both in the support vectors and in the minority class performs better than other approaches.

TABLE 3: THE VALUE g OF ALL APPROACHES.

	STSVM	Biased-SVM	Sos+ BSVM	SVSmt	SVSmt2
Phoneme	82.698	84.022	84.429	85.193	85.118
Pima	71.002	73.276	74.343	75.92	75.939
Satimage	75.737	86.764	88.597	89.782	89.604
Vehicle	72.158	81.417	82.086	83.232	83.89
Mean	75.399	81.370	82.364	83.532	83.638

VI. CONCLUSION

In this paper, some algorithms for dealing with the imbalanced datasets are first introduced in all domains. To improve the computation efficiency of the algorithm, it is proposed by combining SMOTE and Biased-SVM approaches, which limit the support vector not only within the support vectors, but also within the entire minority class. To verify the effectiveness of the proposed algorithm, four different UCI datasets are adopted to validate this approach through simulations. The results indicate the proposed approach can receive better performance than the original approaches.

REFERENCES

[1] N. Chawla, K. Bowyer, L. Hall, P. Kegelmeyer. "SMOTE: Synthetic Minority Over-Sampling Technique", Journal of Artificial Intelligence Research, 2002,16:321-357.

[2] Kubat, M., Matwin, S., "Addressing the Course of Imbalanced Training Sets: One-sided Selection", ICML,1997: 179-186.

[3] Joshi, M., Kumar, V., Agarwal, R., "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements", First IEEE International Conference on Data Mining, 2001: 257-264.

[4] Wu, G., Chang, E., "Class-Boundary Alignment for Imbalanced Dataset Learning", The Twentieth International Conference on Machine Learning (ICML) Workshop on Learning from Imbalanced Datasets, Washington DC, 2003,8:49-56.

[5] Huang, K., Yang, H., King, I., Lyu., M. R., "Learning Classifiers from Imbalanced Data Based on Biased Mini-max Probability Machine", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004: 558-563.

[6] Peng, X., Irwin K., "Efficient Training on Biased Minimax Probability Machine for Imbalanced Text Classification Efficient Training on Biased Minimax Probability Machine for Imbalanced Text Classification", In Proceedings of WWW 2007, ACM,2007,1153-1154.

[7] Seyda,E., Jian, H., Léon B., and Lee, C. G., "Learning on the Border: Active Learning in Imbalanced Data Classification", Proceedings of the 16th Conference on Information and Knowledge Management, CIKM2007, ACM Press, 2007.

[8] Wu, G., Chang, E.Y., "Aligning Boundary in Kernel Space for Learning Imbalanced Dataset", ICDM. (2004) 265–272.

[9] Veropoulos, K., Campbell, C., Cristianini, N., "Controlling the Sensitivity of Support Vector Machines", Proceedings of the International Joint Conference on AI, 1999: 55–60.

[10] Blake, C., Merz, C., "UCI Repository of Machine Learning Databases", Department of Information and Computer Sciences, University of California, Irvine,1998, available from http://www.ics.uci.edu/~mlearn/~MLRepository.html.

[11] Webb, A. R., Statistical Pattern Recognition (Second Edition), Beijing: Publishing House of Electronics Industry, 2004,10:33-37.

[12] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, NY, 1995.

[13] Akbani, R., Kwek, S., Japkowicz, N., "Applying Support Vector Machines to Imbalanced Datasets", European Conference on Machine Learning, 2004, 39-50.

Authorized licensed use limited to: Korea University. Downloaded on November 27,2021 at 08:34:05 UTC from IEEE Xplore. Restrictions apply.