Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm

Liumei Zhang, Baoyu Tan, Tianshi Liu, Xiaoqun Sun

School of computer science, Xi’an Shiyou University,710065, China, zhangliumei@xsyu.edu.cn

Abstract. By combining the modified Random-SMOTE oversampling algorithm with the Biased-SVM classification method, this paper has proposed an improved classification approach for the imbalanced data sets. This algorithm is able to cluster the minority samples, and ensures the support vectors as the parent samples according to the distances between the cluster centers in the minority class and the majority class center. It then could generate the new samples for the minority class. The experiment is conducted upon the five imbalanced data sets from UCI data set and the proposed algorithm is compared with other algorithms. The experimental results show that the improved algorithm has significant classification effect for the imbalanced data sets.

1. Introduction

1.1. Research background
Classic classification algorithms show better accuracy for the balanced data sets. On the other hand, the data sets in the real word tend to be imbalanced. An imbalanced data set refers to a data set with a large difference in the number of the samples belong to different classes. For the imbalanced data, the classification accuracy of the minority samples is often more important than the majority samples. However, most classical classification algorithms assume that the prior probability distribution of the samples or the misclassification cost is equal. Therefore, the traditional classification algorithms are no longer suitable for the imbalanced data classification problems. Increasing the recognition rate of the minority classes in the imbalanced data has become an urgent problem to be solved in data mining fields[1-4].

1.2. Research Status

1.2.1. Research Status of the imbalanced Data Classification
In recent years, imbalanced data classification has a pivotal role in the field of data mining. It has been studied by many researchers using different methods. Those methods can be partitioned into two categories based on the data level and algorithms level.

Data level methods mainly focus on the data pre-processing to reduce the imbalanced data. It has two approaches: oversampling refers to manually increase the minority samples, and under-sampling refer to manually reduce the majority samples. On this basis, Chawla et al.[5] proposed an oversampling method, Synthetic Minority Oversampling Technique (SMOTE), which is able to generate samples on the line between minority samples and their neighbors to obtain the balance of the data. To some extent, this method overcomes the over-learning problem of the random oversampling

...
method. But it still has drawbacks, as the samples are prone to be distributed marginalization and blind interpolated for the SMOTE algorithm. A primary concern of improvement of SMOTE algorithm is studied by numeral researchers[6-8]. Moreover, an improved data reconstruction algorithm—the Random-SMOTE algorithm is proposed in [9].

In addition, some improvements for the imbalanced data sets are based on algorithm level which can offset the impact of the imbalance by introducing some mechanisms, such as the improvement to Boosting algorithm[10] proposed by Joshi et al. and a Data Boost-IM method is proposed by Guo H Y et al.[11]. Moreover, the improvement to SVM is proposed by Wu Gang et al.[12], and a BMPM algorithm is proposed by Huang Kaizhu et al.[13].

1.2.2. Research Status of Support Vector Machine

Support Vector Machine (SVM) is classical classification method, proposed by Vapnik and Cortes, which based on the statistical learning theory in the mid-90s[14]. SVM, which has a strong generalization ability, can solve the problems in the traditional machine learning such as local minima, over-learning and dimensionality disasters[15]. Although SVM has many advantages, they also have limitations such as sensitive to the noises[16], the unsatisfied classification effects for the imbalanced data[17] and et al.

In addition, the traditional SVM method has two assumptions: the size of the positive and negative samples for training is balanced, and the cost of misclassification for the samples belong to different classes is basically similar[18]. Besides the processing methods on the data level, in view of the shortcomings of traditional SVM for the classification of the imbalanced data sets, Veropoulos et al. proposed a strategy to solve this problem[19]. Biased-SVM algorithm is dedicated to assign different penalty coefficients C for the positive and negative samples. In this algorithm, the minority samples are given larger penalty factors, and the majority samples are given smaller penalty factors. Therefore, the SVM classifier is able to concentrate on the misclassification rate of the minority class.

2. Existing algorithms

2.1. Random-SMOTE algorithm

With oversampling, SMOTE (Synthetic Minority Oversampling Technique) can generate the samples for the minority class, at the same time, it doesn’t change the sparse distribution of the minority samples [20]. For this purpose, literature [9] proposed an improved oversampling approach—Random-SMOTE algorithm which can generate the new samples for the minority class in a widen field. The algorithm is as follows:

- Linear interpolating randomly between \( y_1 \) and \( y_2 \) to generate \( N \) temporary samples \( t_j \) \((j = 1,2 \cdots N)\)
  \[
  t_j = y_1 + \text{rand}(0,1) \ast (y_2 - y_1), \quad j = 1, 2 \cdots N
  \]

- Linear interpolating randomly between \( t_j \) and \( x \) to construct a new sample \( p_j \) \((j = 1,2 \cdots N)\)
  \[
  p_j = x + \text{rand}(0,1) \ast (t_j - x) \quad j = 1, 2 \cdots N
  \]

2.2. K-means algorithm

The K-means algorithm is a clustering algorithm based on partition, which uses the distance between the samples as the criterion for the similarity measure, that is, the smaller the distance between data samples, the higher their similarity. According to the principle of similarity, the data samples with higher similarity belong to the same cluster, and the data samples with higher dissimilarity are divided into different clusters. The clustering is an unsupervised learning process.
2.3. Biased-SVM algorithm

In the classification for their balanced data, the problem of SVM based on the minimization of structural risk is that the classification weight will be biased towards the majority class, making the classification hyperplane close to the minority class, and thus it is easy to misclassify the minority samples. Veropoulosk et al. proposed a strategy to solve this problem on the algorithm level, and this algorithm gives the different penalty coefficients $C^+$ and $C^−$ for the positive and negative samples[21].

In this way, the objective function becomes:

$$
\min \frac{1}{2} \|w\|^2 + C^+ \sum_{\{i:y_i=+1\}} \xi_i + C^- \sum_{\{i:y_i=-1\}} \xi_i
$$

s.t. $y_i(w \cdot \varphi(x_i) + b) - 1 + \xi_i \geq 0$

$$\xi_i \geq 0, i = 1, ..., l$$

3. The improved classification algorithm for the imbalanced data sets

In the SVM classification, the support vectors play a decisive role for the classification hyperplane. Therefore, in the Random-SMOTE algorithm, sampling all the minority samples will result in a large number of redundancy, which will further increase training time and decrease the qualities of training samples. Therefore, we only consider generating samples that are “close” to the boundary.

Based on this, we deal with the imbalanced data from both the data level and the algorithm level. The steps of this algorithm are follows:

**Algorithm:** for the imbalanced data based on Biased-SVM and the modified over-sampling

**Input:** An imbalanced data set

1. Clustering the minority samples using the K-means algorithm, setting the cluster centers are $L_1, L_2, \ldots, L_H$.

2. Setting the majority class center is:

$$
\bar{M}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_i, (j = 1, 2, \ldots, L)
$$

3. Calculating the distance from each minority class cluster center to the majority class center, setting the distances: $d_1, d_2, \ldots, d_H$.

4. Keeping the $K$ Clusters: $C_1, C_2, \ldots, C_K$ with the first $K$ smallest distances, and take them as the parent samples for oversampling.

5. Oversampling only for clusters $C_1, C_2, \ldots, C_K$, using the Random-SMOTE algorithm in order to balance the data.

6. Using the Biased-SVM algorithm to set the different penalty factors for the minority samples and the majority samples.

$$C^+ = 0.5 \times C \times \frac{p+q}{p}, C^- = 0.5 \times C \times \frac{p+q}{q}$$

In which, $C^+$ is a penalty factor for the majority samples, $C^−$ is a penalty factor for the minority samples, $p$ is the number of the majority samples, $q$ is the number of the minority samples.

4. Experimental results and analysis

In this study, we use five highly imbalanced data sets. These data sets are all from the UCI data set and have different sample sizes and attributes. In addition, they also have the different imbalance ratio (IR). Table 1 summarizes the characteristics of the imbalanced data sets selected in this experiment, including the number of attributes ($#A$), the number of the minority samples ($#M$), the number of the majority samples ($#J$) and the imbalance ratio (IR). In which, $IR = \frac{#M}{#J}$. In this study, the
parameter values are shown also in table 2, including the K value of the K-means algorithm and the number of the support vectors in the minority class.

Table 1 The information of the imbalanced data sets

| Dataset    | #A | #M | #J | IR  | K  | The number of the support vectors |
|------------|----|----|----|-----|----|----------------------------------|
| Banana     | 3  | 75 | 2808 | 0.03 | 5  | 28                               |
| Haberman   | 4  | 81 | 225 | 0.36 | 4  | 35                               |
| Appendicitis | 8 | 21 | 85  | 0.25 | 3  | 8                                |
| Vehicle    | 19 | 199| 647 | 0.31 | 5  | 20                               |
| Wisconsin  | 10 | 241| 458 | 0.53 | 8  | 32                               |

In this experiment, we compare the classification results with the SVM algorithm, the Random-SMOTE algorithm and the improved algorithm in this paper. Table 2 shows the classification results of different data sets by the three different algorithms. We take F-measure and G-mean as the evaluation indexes of classification performance. For each oversampling algorithm, we calculate 10 times to obtain its average value.

Table 2 The experimental performance comparison of the three algorithms

| Dataset   | SVM          | Random-SMOTE | The improved algorithm |
|-----------|--------------|---------------|------------------------|
|           | G-mean | F-measure | G-mean | F-measure | G-mean | F-measure | G-mean | F-measure |
| Banana    | 0.623   | 0.123    | 0.756   | 0.760     | 0.767   | 0.781     |
| Appendicitis | 0.609 | 0.460    | 0.820   | 0.798     | 0.829   | 0.832     |
| Haberman  | 0.607   | 0.495    | 0.743   | 0.731     | 0.764   | 0.741     |
| Vehicle   | 0.924   | 0.892    | 0.949   | 0.947     | 0.956   | 0.955     |
| Wisconsin | 0.928   | 0.910    | 0.938   | 0.933     | 0.956   | 0.947     |

In this experiment, we use the above three algorithms to classify a two-dimensional unbalanced data sets, and we can intuitively compare the classification results between the SVM algorithm and the Random-SMOTE algorithm from Fig.1 and Fig.2.

Fig.1 classification results of SVM algorithm

Fig.2 classification results of improved algorithm

5. Conclusion
This paper has proposed the improved algorithm combined the Random-SMOTE oversampling algorithm with the Biased-SVM. This algorithm extracts the support vectors as the parent samples according to the distances between the cluster centers in the minority class and the majority class center, and generates the new samples for the minority. Although the new algorithm has some improvements on the classification effects, the optimization of the parameters still needs to find the optimal method to solve, so as to improve the generalization ability of the classifier.

Acknowledgments
This work is supported in part by the scholarship from China Scholarship Council(CSC) under the Grant CSC No.201809910007.
References

[1] He H, Garcia E A. Learning from imbalanced data[J]. Knowledge and Data Engineering, IEEE Transactions on, 2009, 21(9): 1263-1284.

[2] Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies[C]//AAAI workshop on learning from imbalanced data sets. 2000, 68:10-15.

[3] Fernández Adel Jesus MJ Herrera F. Hierarchical Fuzzy Rule Based Classification Systems with Genetic Rule Selection for Imbalanced Data-Sets[J]. International Journal of Approximate Reasoning, 2009, 50(3): 561-577.

[4] Han Min, Zhu Xinrong. Mixed algorithm group for unbalanced data classification[J]. Control theory and application, 2011, 28(10): 1485-1489. (in Chinese)

[5] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[6] Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTEBoost: Improving prediction of the minority class in boosting[C]//7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003, 107-119.

[7] Xun H G, Hui H, et al. An over-sampling expert system for learning from imbalanced data sets[C]//International Conference on Neural Networks and Brain. 2005, 537-541.

[8] Hui H, Wang W Y, B H Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]//International Conference on Intelligent Computing. 2005, 878-887.

[9] Dong Yanjie. The Study on Random-SMOTE for the Classification of Imbalanced Data Sets[D]. Dalian university of technology, pp: 20-25. (in Chinese)

[10] Joshi M, Kutnar V, Agarwal R. Evaluating Boosting Algorithms to classify Rare Classes: Comparison and Improvements[C]//First IEEE International Conference on Data Mining. 2001: 257-264.

[11] Guo H Y, Herna L V. Learning from Imbalanced Data Sets with Boosting and Data Generation: The Data Boost-IM Approach[J]. SIGKDD Explorations, 2003, 6: 30-39.

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

[13] Huang Kaizhu, Yang Haiqin, King I, et al. Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004: 558-563.

[14] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995(20): 273-297.

[15] Cristianini, N, Shawe-Taylor. J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. London: Cambridge University Press, 2000.

[16] Cao L J, Lee H P, Chong W K. Modified support vector novelty detector using training data with outliers [J]. Pattern Recognition Letters, 2003, 24 (14): 2479-2487.

[17] Wang B X, Japkowicz N. Boosting support vector machines for imbalanced data sets[J]. Knowledge and Information Systems, 2010, 25(1): 1-20. (in Chinese)

[18] Tang Qin. Unbalanced Data Classification Algorithm and Its Application in Student Loan Risk Management [D]. Huazhong University of Science and Technology, 2011.2: 23. (in Chinese)

[19] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines[C]//Proceedings of the International Joint Conference on AI 1999; 55-60.

[20] Weiss G M, Hirsh H. A Quantitative Study of Small Disjuncts[C]// In Proceedings of the 17th National Conference on Artificial Intelligence. Texas: AAA I Press, 2000: 665-670.

[21] Veropoulosk, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines[C]//Proceedings of the international joint conference on AI, 1999; 55-60.