International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.2 (2014), pp.61-68
http://dx.doi.org/10.14257/ijsip.2014.7.2.06

Imbalanced Data SVM Classification Method Based on Cluster Boundary Sampling and DT-KNN Pruning

Li Peng[1,2], Yu Xiao-yang[1], Bi Ting-ting[2] and Huang Jiu-ling[2]

[1] *Higher Educational Key Laboratory for Measuring and Control Technology, Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, 150080 Harbin, China*
[2] *School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China*
{pli, yuxiaoyang }@hrbust.edu.cn.

Abstract

This paper presents a SVM classification method based on cluster boundary sampling and sample pruning. We actively explore an effective solution to solve the difficult problem of imbalanced data set classification from data re-sampling and algorithm improving. Firstly, we creatively propose the method of cluster boundary sampling, using the clustering density threshold and the boundary density threshold to determine the cluster boundaries, in order to guide the process of re-sampling more scientifically and accurately. Secondly, we put forward a new sample pruning algorithm based on dynamic threshold KNN to deal with the complexity and overlapping problem of imbalanced data set. The phenomenon of data complexity and overlapping will reduce the classification performance and generalization ability of SVM classifier. Experiments show that our method acquires obviously promotion effect in various different imbalanced data sets and it can prove the validity and stability.

Keywords: *Imbalanced Data Sets; Support Vector Machine; Cluster Sampling; Sample pruning; Classification*

1. Introduction

Through the research of many field, people found that the distribution of data categories are often imbalance. Learning with imbalanced class distributions typically addresses the case when, for a two class classification problem, the training data for one class (majority) greatly outnumbers the other class (minority). Imbalanced data set is the real observation data form exists in many fields of natural science. Although it reflects the nature of objective things, but people always only focus on the occurrence of the minority. For example, in credit card fraud detection, the vast majority of users are legitimate, but we hope to predict potential illegal users [1]; In company's bankruptcy risk prediction, the bankrupt companies are in the minority, but companies managers are concerned about the possibility of bankruptcy [2]; In medical diagnosis, healthy people in the real data must be the majority, but people are concerned about is whether the existing data to predict the incidence of the disease [3]; In the biological field, the prediction of protein type also face the imbalance of the data categories[4]. The problem of imbalanced data sets classification is one of the most challenging and difficult task in the field of data mining, have get a great deal of attention from researchers around the world.

ISSN: 2005-4254 IJSIP

Unfortunately, Standard classifier tends to be overwhelmed by the large classes and ignore the small ones [5]. Moreover, the class bias and data overlapping of imbalanced data are main factors which produce the poor classification performance. Therefore, we propose cluster boundary sampling and dynamic threshold KNN to solve above problems respectively.

2. Cluster Boundary Sampling Method based on Density Clustering

Re-sampling method is active to solve imbalanced dataset problem. It focuses on the imbalanced distribution of data label [6]. There are two general strategies to adjust the imbalance of data sets. One named under-sampling is to reduce the number of data samples in majority class; the other named over-sampling is to virtually increase that in minority class [7]. We utilize under-sampling strategy to deal with the problem of data imbalance because this method will remove many data samples of majority class to adjust the distribution but some information has to be lost. In order to reduce its impact on classifier, we will retain more representative data elements in re-sampling as far as possible. Clustering method can cluster similar data elements as several clusters. Samples are as similar as possible in every cluster and the data samples should be as different as possible between different clusters. Clustering method based on density cannot be restricted by data attributes, dimension, sequence and space distribution shape. This method can automatically identify cluster number and has a strong capacity of resisting disturbance [8].

2.1. Density Clustering Algorithm

Clustering is a process that centralizes all objects into a series of collections with similar objects. Divided the data elements into several clusters by calculate their similarity. The elements in each cluster are as similar as possible and elements between the cluster and other cluster are as different as possible. The density-based clustering algorithm can be exempt from the constraints of the data attributes, dimension, order and spatial distribution shape, clustering cluster can automatically identify the number of anti-interference ability. The main idea of density-based clustering method is to select an object as a core object and query neighborhood of the core object, as long as the density of the adjacent area exceeds a certain threshold value, then selecting any object outside the core object within the immediate area as a core objects continue to clustering. Ultimately, high density region is divided by a relatively low density region form a clustering cluster. Density-based clustering algorithm is a kind of widely used clustering method, the algorithm separates the low density regions in the data space from the high density region, it can find clusters of arbitrary shape, and able to identify the noise data. Cluster analysis can handle any data types such as interval scale variables, binary variables, categorical variables, ordinal variables, proportional scale type variables and so on. The metrics of the similarity or dissimilarity of the different data types are also different. Clustering algorithms typically choose the data structure of the data matrix and dissimilarity matrix.

Assume that a data object from the d attribute description, certain data object having a d attribute constitutes a *d*-dimensional data space. In a *d*-dimensional space, the data object is referred to as a *d*-dimensional data points, then the *d*-dimensional data point x can be expressed as $x = (x_1, \cdots, x_d)$, where x_i represents the *i*-th attribute value, d represents the dimension of the space. A collection composed by n *d*-dimensional data

points (also known as *d*-dimensional data set) S can be expressed as $S = (s_1, \cdots, s_n)$, where $s_i = (s_{i1}, \cdots, s_{id})$ and s_{ij} represents the *i*-th data points of the *j*-th attribute values. According to the similarity between the data points, the *d*-dimensional data set V is divided into a process of $\{C_1, C_2, \cdots C_k\}$ known as cluster analysis, where in $k \le n, C_i \ne \varnothing, C_i \subseteq V$ $(i = 1, 2, \cdots, k)$, and $\cup C_i = V$. Here, C_i is generally called a cluster

DBSCN algorithm is a kind of widely applied density clustering algorithm. In this algorithm, cluster is high density area divided by low density area. And DBSCAN can find arbitrary shape cluster and identifies noise data [9]. The main idea of DBSCAN is choosing one object as core and exploring one of its adjacent areas. As long as the density of the adjacent area attains a certain threshold, in the near area, choosing any object as core but pre-cores, and keeping clustering.

2.2. Cluster Boundary Under-sampling Method

Cluster-based sampling method can takes sample for every cluster. Loss of information caused by sampling randomness can be reduced as much as possible. When choosing SVM as classification algorithm, SVM calculates the classification hyper-plane the depended on decision boundary. So centre of cluster will be abandoned when we are sampling. Retain the data of cluster boundary is most likely to help SVM making decision.

The same cluster of data elements in the vector space is relative density and the data contains a high similarity of the content by the distribution of the density-based clustering. Extracting the data elements of the cluster boundaries can be an effective representative of the characteristics of clusters in the data object. For the elements in the data space can be mapped to points in the N-dimensional space. More precisely, any data element x is the vector form of the following characteristics, and using the standard Euclidean distance as the distance between two vectors.

$$\left\langle \alpha_1(x),\ \alpha_2(x),\ \ldots,\alpha_n(x) \right\rangle \tag{1}$$

In that, it means the instance of x of the k -th attribute. Then the Euclidean distance between two instances x_i and x_j is defined as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (\alpha_k(x_i) - \alpha_k(x_j))^2} \tag{2}$$

In the data set D , the neighborhood of the instance x , can be defined as:

$$EPS(x) = \{ y \in D \mid d(x, y) \le EPS \} \tag{3}$$

This method is based on the neighborhood definition to determine the boundary points of clustering elements in the same clusters. If the number of elements contained in the neighborhood of an element is more, indicating that the element located in a region more close to the center of clusters. And if the fewer number of elements contained in the neighborhood of an element, the element located in a region more close to the cluster boundary. We can use the $|EPS(x)|$ represent the number of neighborhood elements in the region of data x .

In order to be able to locate the clustering boundaries more accurately, we use 2 groups density threshold. One group is called clustering density threshold, we can estimate them by using the characteristics and the average distance of the data in order to divide the data set into several clusters. The other group called boundary density threshold estimated by the size of each cluster to be used to find the boundary of the cluster data object. We use the first group clustering density threshold value EPS_1 and the $MINP_1$ to find the similar data elements in data set, the data elements of the data set are divided into a plurality of clusters C . Use the second set of boundary the density the threshold to value EPS_{ci} and $MINP_{ci}$ to find cluster boundary of each cluster C_i the boundary density threshold selection depends on the size of the cluster C_i . We use D to represent the whole set of training data, C_i to represent the *i*-th divided cluster, B_i to represent the boundary of the cluster C_i .

$$D = \{C_1, C_2, C_3, ..., C_n, C_{noise}\} \tag{4}$$

$$C_i = \{x \in D \mid \left|EPS(x)\right| \geq MINP_1\} \tag{5}$$

$$B_i = \{x \in C_i \mid \left|EPS(x)\right| \geq MINP_{ci}\} \tag{6}$$

In imbalanced data set, there is a large gap between the numbers of two classes. Therefore, we ensure the minority class information is complete, and the majority class information is representative as much as possible in the process of cluster sampling. We keep on all minority class information only to cluster the data of majority class and extract the cluster boundary for sampling. At last, make all of these boundaries elements and minority class data as training data for SVM classifier.

The algorithm can be described as follows:

1. Traversing D data elements, and to calculate the distance between the elements and the elements in D ;
2. Estimating clustering density threshold $MINP_1$;
3. Using the first group of density threshold clustering on D ;
4. Mark the data elements in D which belonging to cluster C_i or noise C_{noise} ;
5. For each cluster C_i , compute the number of data elements in cluster N_{ci} ;
6. According N_{ci} to estimate the cluster density threshold $MINP_{ci}$;
7. Calculating each data element in a neighborhood its number of elements belonging to the same cluster;
8. According to density threshold $MINP_{ci}$, the boundary elements B_i extracted from the cluster C_i ;
9. Repeat step 4, until all the elements of the cluster in the D have traversed;
10. Get all B_i .

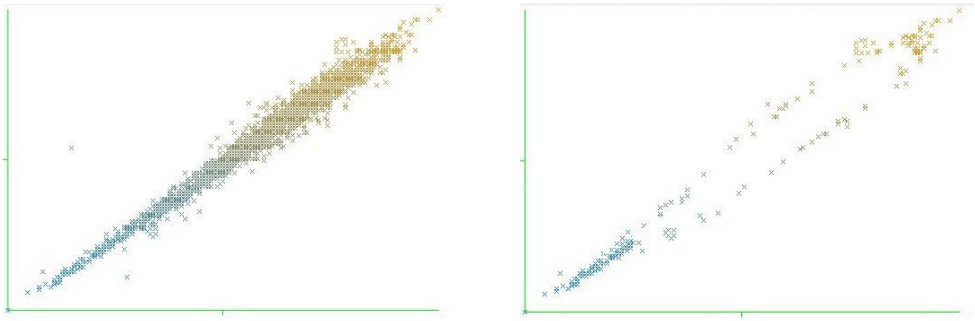

Figure 1. Neighborhood-based Clustering to Determine the Cluster Boundaries

In order to validate the cluster boundary neighborhood-based clustering to determine the calibration algorithm, we use the test data set to verify the algorithm, and visually displayed in the form of two-dimensional scatter plot. Figure 1 (a) shows a cluster of density clustering of the original data; Figure 1 (b) shows a collection of cluster using the boundary points obtained by this method clustering cluster. Figure 1 can be very intuitive proof method can effectively obtain clustering cluster boundaries and the boundary is obtained accurately.

3. Pruning Algorithm based on Dynamic Threshold KNN

The acquisition of data in the actual environment often has the overlapping phenomenon owing to the noise, accidental influence and equipment error etc. Therefore, we usually need to prune the samples in the application of these data sets and pruning algorithm has become an important research hot-spot in recent years. Pruning algorithm is also widely used in the fields of biology [10], economy [11] and medicine [12] *etc*. However, there is little discussion about sample pruning method for imbalanced data set classification. Therefore, we take the lead in exploring this special subject.

3.1. Complexity and Overlapping Analysis of Imbalanced Data Set

Linearly separable data don't exit in reality and most data is complexity with overlapping phenomenon. Especially in the imbalanced data set, this situation is more serious. Figure 2 shows the complexity and overlapping in sample space; it is one of the main reasons led to the decline in the performance of classifier.

When the number of samples is large, complex distribution, and overlapped seriously, which is difficult to judge effective data and remove noise. Therefore, the problem of complexity and overlap is essential to be considered and resolved when imbalanced data classification technology is applied in the practical application. SVM classification performance will decreases obviously because serious data overlapping will rapidly increase the number of support vectors which lead to computational burden, overlearning and weakening generalization ability. Hence, it is significant for SVM classifier to find an effective samples pruning method to improve this issue.

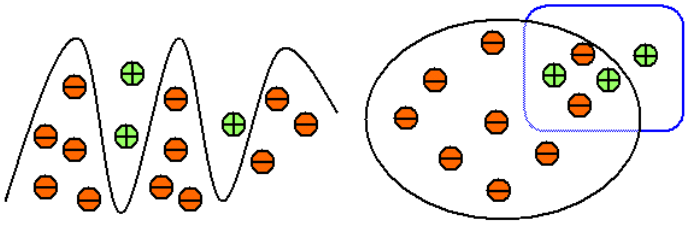

Figure 2. The Phenomenon of Data Complexity and Overlapping

3.2. DT-KNN Pruning Algorithm

We propose a sample pruning algorithm based on dynamic threshold K nearest neighbor (DT-KNN) to solve the complexity and overlap of imbalanced data set. KNN is a mature theory of machine learning and the basic motivation for considering the KNN rule rests on our earlier observation about matching probabilities with nature. The KNN query starts at the test point x and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. We notice first that if k is fixed and the number n of samples is allowed to approach infinity, then all of the k nearest neighbors will converge to x .

KNN algorithm assumes that all samples are mapped into multidimensional space R^n, and to find the k nearest neighbor distance the prediction sample in the multidimensional space, and then determining the prediction samples according to k points' category. More precisely, we represent any instance as the feature vector and use standard Euclidean distance as the value between two vectors.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (\alpha_k(x_i) - \alpha_k(x_j))^2} \tag{7}$$

The KNN algorithm described above is only applicable to general data set, but for imbalanced data set, the effect is not ideal if we directly apply this method. In Imbalanced data set, the positive samples are very scarce, so the positive information is more precious than negative ones. Furthermore, the number of mixed negative samples far exceeds the positive ones because of data class imbalance. Therefore, we use different control thresholds for the prediction of the positive and negative respectively to ensure rare positive information without loss as much as possible. When the positive resource is too extremely scarce, we can even prune the negative and ignore the positive cases. We define the class attribute values is $f(x_i) \in \{1, -1\}$, predictive threshold of query point is calculated by the following formula.

$$\psi(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k} \tag{8}$$

Our algorithm is implemented as follows:

Input: the re-sampled data set D;

Input parameters: K, positive control threshold, negative control threshold;

1. Choosing a not calculated sample as the query point in the data set D;
2. Finding K nearest neighbors of a query point x_q ;

3. Calculating attribute prediction value $\psi(x_q)$ by the formula (3)

   If $f(x_q)$=1 and $\psi(x_q) \le \theta^+$, then delete the current query point x_q;

   If $f(x_q)$=-1 and $\psi(x_q) \ge \theta^-$, then delete the current query point x_q;

4. If not all sample points are calculated then go to step 1 continue;

5. Getting pruned sample set T and output the results, the algorithm finish.

Output: Sample set T after pruning.

We will try our best to ensure the integrity of rare positive information in the process of pruning. We firstly cluster the imbalanced data, and then prune the samples of cluster, finally pruning data as SVM classifier training samples.

4. The Results and Analysis of Experiment

We select four UCI imbalanced data sets to verify the effectiveness of our method and apply SVM to learn the training samples and KNN dynamic pruning samples. Table1 show the number of support vectors before and after pruning and Table 2 express the SVM classification performance before and after our method.

Table 1. Comparison of Support Vectors Number Before and After Pruning

		the number of support vector			
		Shuttle	Abalone	Yeast	Churn
No Sampling	Before pruning	74	40	76	699
	After pruning	43	29	10	296
Boundary Sampling	Before pruning	62	46	47	698
	After pruning	20	8	15	337

Table 2. Comparison of AUC Performance Before and After our Method

		classification performance before and after our method (AUC value)			
		Shuttle	Abalone	Yeast	Churn
No Sampling	Before pruning	0.4792	0.5502	0.6623	0.9013
	After pruning	0.7562	0.7641	0.8591	0.9031
Boundary Sampling	Before pruning	0.7670	0.7004	0.8487	0.9053
	After pruning	0.7948	0.7154	0.9023	0.9143

Table 1 show that DT-KNN sample pruning algorithm can significantly reduce the number of support vectors. It proves that DT-KNN pruning algorithm can lighten the complexity and overlapping of imbalanced data sets, and then improve the generalization ability of SVM classifier. The results of Table 2 prove that cluster boundary sampling is an effective strategy to deal with the problem of imbalanced data classification and pruning algorithm as an auxiliary means can further enhances performance.

5. Conclusion

This paper proposed an effective solving strategy on imbalanced data classification by means of cluster boundary sampling and DT-KNN sample pruning. We creatively apply boundary sampling to reduce the number of negative samples, and then declining

the imbalance ratio. We also propose a new sample pruning algorithm to deal with the complexity and overlapping of imbalanced data sets. Experimental results show that our method is contributed to improving the classification performance of imbalanced data sets.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China (61103149), China Postdoctoral Science Foundation (2011M500682), Postdoctoral Science Foundation of Heilongjiang Province (LBH-Z11106), Technological Innovation Foundation for Youth Scholars of Harbin (2012RFQXG093), Foundation for University Key Teacher of Heilongjiang Province (1252G023), Postgraduate Innovation Research Project of Heilongjiang Province (YJSCX2012-126HLJ), Research Fund for the Doctoral Program of Higher Education (20102303120005), Province Natural Science Foundation of Heilongjiang (QC2013C060) and Science Funds for the Young Innovative Talents of HUST (NO.201304).

References

[1] W. Wei, L. Jin-jiu and C. Long-bing, J. Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web. 31, 6 (**2012**)
[2] Z. Li-gang, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", Knowledge-Based Systems, vol. 41, no. 3, (**2013**).
[3] J. Nahar, T. Imam, K. S. Tickle and C. Yi-ping, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach", Expert Systems with Applications, vol. 40, no. 1, **(2013)**.
[4] Z. Yong-qing, Z. Dan-ling and M. Gang, "Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions", Computational Biology and Chemistry, vol. 36, no. 2, **(2012)**.
[5] J. Burez and D. Vanden, "Handling class imbalance in customer churn prediction. Expert Systems with Applications", vol. 36, no. 3, **(2009)**.
[6] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions", Expert Systems with Applications, vol. 36, no. 3, **(2009)**.
[7] I. Albisua, O. Arbelaitz and I. Gurrutxaga, "The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets", Progress in Artificial Intelligence, vol. 2, no. 1, **(2013)**.
[8] S. Ghosh and B. Lohani, "Mining lidar data with spatial clustering algorithms", International Journal of Remote Sensing, vol. 34, no. 14, **(2013)**.
[9] J. Liu, Z. Huang and J. Luo, "Privacy preserving distributed dbscan clustering", Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, **(2012)** May 16-18.
[10] S. Xiao-feng and W. Ming-hao, "Prediction of pre-miRNA with Multiple Stem-loops Using Pruning Algorithm", Computers in Biology and Medicine, vol. 43, no. 5, **(2013)**.
[11] P. Xingcheng and S. Pengfei, "A New Hybrid Pruning Neural Network Algorithm Based on Sensitivity Analysis for Stock Market Forecast", Journal of Information and Computational Science, vol. 10, no. 3, **(2013)**.
[12] M. Dimitrios, "Genetic Algorithm Pruning of Probabilistic Neural Networks in Medical Disease Estimation", Neural Networks, vol. 24, no. 8, **(2011)**.

