Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 1540628, 9 pages
http://dx.doi.org/10.1155/2016/1540628

Research Article

Imbalanced Data Set CSVM Classification Method Based on Cluster Boundary Sampling

Peng Li,[1,2] Tian-ge Liang,[2] and Kai-hui Zhang[3]

[1]*School of Software, Harbin University of Science and Technology (HUST), Harbin 150080, China*
[2]*School of Computer Science and Technology, Harbin University of Science and Technology (HUST), Harbin 150080, China*
[3]*Academic Journal Center, Heilongjiang University, Harbin 150080, China*

Correspondence should be addressed to Peng Li; pli@hrbust.edu.cn

Received 11 March 2016; Revised 11 June 2016; Accepted 28 June 2016

Academic Editor: Peter Dabnichki

This paper creatively proposes a cluster boundary sampling method based on density clustering to solve the problem of resampling in IDS classification and verify its effectiveness experimentally. We use the clustering density threshold and the boundary density threshold to determine the cluster boundaries, in order to guide the process of resampling more scientifically and accurately. Then, we adopt the penalty factor to regulate the data imbalance effect on SVM classification algorithm. The achievements and scientific significance of this paper do not propose the best classifier or solution of imbalanced data set and just verify the validity and stability of proposed IDS resampling method. Experiments show that our method acquires obvious promotion effect in various imbalanced data sets.

1. Introduction

Imbalanced data set (IDS) is the real observed data form which is prevalent in computer, economics, biology, medicine, and many other natural science fields. That is, there may be magnitude difference in the number of samples between classes. It reflects the nature of the objective things, but, in fact, people tend to only care about the occurrence of small categories. For example, in the detection of financial fraud, the vast majorities are legitimate users, but people do hope through the data to predict those potential illegal users [1, 2]; in the company's bankruptcy risk prediction, the bankrupt company is a small minority, but the enterprise managers are concerned about whether the current business situation has the bankruptcy possibility [3]; in petroleum exploration, there may not be many petroleum areas, but this is exactly what exploration personnel should focus on [4]; in medical diagnosis, the fact is that healthy people must be the majority, but people care about whether they could predict the occurrence of the disease through the current data [5]; in the industrial field, the fault detection is typical of the IDS problem. Most of the equipment must be kept in normal operation in working hours, but it means a lot if a small amount of abnormal equipment or parts can be detected in advance [6]; in the field of biology, the prediction of DNA sequence and protein type also faces the problem of IDS [7, 8]. Therefore, IDS classification not only is one of the most difficult challenges in classification technology research at the technical level, but also has very important practical significance at application level. In recent years it has attracted a great deal of attention from the majority of researchers in various fields.

A large number of studies have shown that satisfactory classification results cannot be achieved if some standard classification models are directly adopted to solve the IDS classification problem [9]. Almost all of the methods were very low in classification accuracy of rare categories and could not improve the recognition level of the rare classes as a whole to an extent of actual acceptance. Researchers are facing a huge challenge; thus the related researching needs to go deeper. Theoretically, we can adopt two strategies to solve the problem of IDS classification. One is the resampling, which can be divided into the down- and upsampling. That is, we can appropriately screen the information content from samples of large class or improve the error costs from samples of small class [10]. The other strategy is to explore more

suitable classification models and mend the IDS classification algorithm based on IDS data characteristics to improve their classification ability.

In this paper, we propose to combine the data resampling and algorithm enhancing strategies for an integrated solution to the problem of IDS classification. We creatively propose the method of cluster boundary sampling and resampling of the IDS, which not only effectively balances the data skew state but also greatly reduces the number of support vectors. It betters the classification results and improves the classification speed significantly. This sampling method overcomes the traditional sampling methods' shortness including lack of theoretical basis, strong randomness, interference of human subjectivity, and serious information loss. At the same time, it is a good solution to the aliasing phenomenon in data, which can greatly improve the generalization performance of the following SVM classifier. In order to adapt to the imbalance state of sample, we have also improved the SVM classification model and use the grid optimization method of cross validation for the training data to determine the punishment factor of the SVM and gamma value of kernel function, which seeks a reasonable theoretical basis for the determination of the penalty factor for SVM.

2. The IDS Classification Algorithm Based on CSVM

At present, the SVM classification algorithm has been able to solve most of the problems with the characteristics of the relatively small data volume, more complete marks, and relatively homogeneous distribution [11]. But when facing the IDS, its performance has dropped significantly. By investigating the reason, we could find that the main reason should be the imbalanced distribution of training data, which makes the positive and negative samples ratio of support vector also obviously imbalanced. Negative samples information occupies a dominant position and thus submerges the positive samples information. Eventually the decision function makes the classification results redundantly lean to negative samples, since only the training samples near the interface could be used as support vector in classification and the samples far from the interface are not likely to be affected. In theory, the SVM should be affected by the IDS classification model with minimal impact compared with other models. For the cases with the smaller imbalance rate, we can get good classification results even though little work has been done to improve the classifying model itself and the SVM learning mechanism could give us a lot of space to improve the classification model. Therefore, this paper still chooses the SVM as the basic algorithm model of IDS classification.

2.1. The Analysis of the IDS's Influence on SVM Classification. In order to make the analysis more visual and intuitive, we analyze the impact of data imbalance and sampling on SVM classification by linear separable data sets. As can be seen from Figure 1(a) of the training process, due to the imbalance of the positive and negative samples, the actual classification hyperplane basically keeps consistent with the ideal hyperplane in the direction, while it is far from the negative samples and close to the positive ones, which is the result of data submerging phenomenon. As shown in Figure 1(b), in the test, this classification hyperplane has a strong tendency towards some negative samples, which makes some positive samples wrongly classified as negative samples.

We randomly select the same number of samples from the positive and negative samples and make the data reach an equilibrium state. Figure 2(a) is the training results after resampling. Although the classification hyperplane we get from learning basically keeps an ideal distance with the positive and negative samples, it gets very big deviation with the direction produced by the ideal super plane, which results from information loss after sampling. As shown in Figure 2(b), such classification hyperplane may also lead to an incorrect classification in the test.

Therefore, how to reduce the information loss in a maximum degree while lowering the deflection rate is an important issue to concern if we want to use the resampling method to solve the IDS.

2.2. SVM Penalty Factor Determination Based on Grid Optimization. Through the previous analysis we can draw such a conclusion. For the IDS, if we directly adopt the traditional SVM model for classification, the actual classification hyperplane basically keeps consistent with the ideal hyperplane in the direction, but we get the deviation in the distance, which is far away from the negative samples and closer to positive samples. This situation makes people naturally think of whether there is a way to pull the actual classification hyperplane near the ideal hyperplane. The most commonly accepted and widely used method is to adjust the penalty factor. Adjusting the penalty factor C is generally considered to be an effective way to improve the IDS classification effect [12].

Give a kernel function K and a set of labeled samples $X_{\text{train}} = \{x_i, y_i\}_{i=1}^{n}$. SVM finds a best α_i for each x_i to make classification interval γ smallest between the classification hyperplane and the sample nearest to it. When a new test sample arrives, it can be predicted by the following formula:

$$\operatorname{sgn}\left(f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b \right). \tag{1}$$

In the formula b is the threshold. First-order soft margin SVM will minimize the initial Lagrange function:

$$L_p = \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i \left[y_i (w \cdot x_i + b) - 1 + \xi_i \right] - \sum_{i=1}^{n}\gamma_i\xi_i. \tag{2}$$

In the formula, $\alpha_i \geq 0$, $\gamma_i \geq 0$. Penalty factor C represents a compromise between the empirical error and the classification interval. To satisfy KKT conditions, the value of α_i should be met:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n}\alpha_i y_i = 0. \tag{3}$$

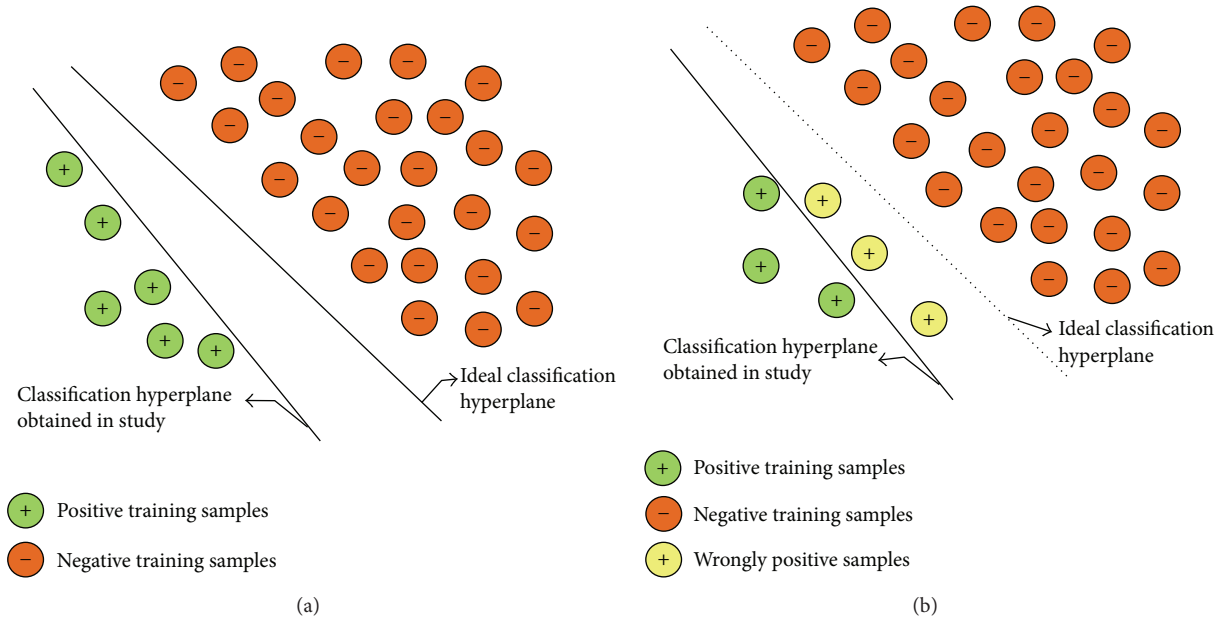

Figure 1: The phenomena of data flow.

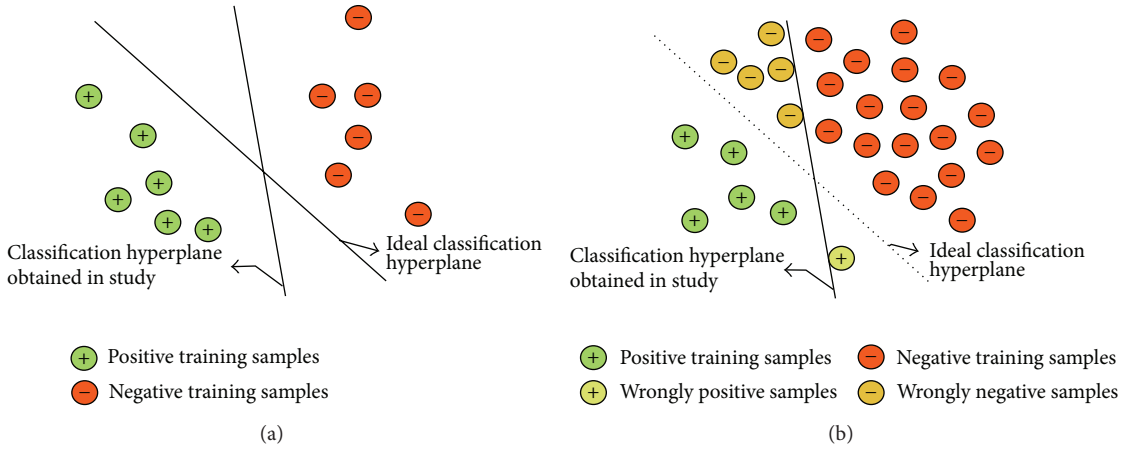

Figure 2: The phenomena of information loss.

The penalty factor produces effect mainly through the restriction of α_i in formula (1). The value of α_i decides how big the vector x_i role is in determining the classification plane. Generally speaking, the bigger its value is the bigger effect it has in classification prediction. According to the optimum conditions (KKT), the value of α_i and its relationship with the penalty factor C are as follows:

$$\begin{aligned} \alpha_i = 0 &\Longrightarrow y_i f(x_i) > 1, \quad \xi_i = 0, \\ 0 < \alpha_i < C &\Longrightarrow y_i f(x_i) = 1, \quad \xi_i = 0, \\ \alpha_i = C &\Longrightarrow y_i f(x_i) < 1, \quad \xi_i \geq 0. \end{aligned} \tag{4}$$

Adjusting the penalty factor can improve the classification effect of SVM on IDS, but this method also has some drawbacks. First, the penalty factor is often determined and adjusted in different application data by virtue of the experience from the researchers and there is no theoretical basis and support. Second, some research shows that the effect is very limited if we improve the IDS classification by adjusting the penalty factor. It is not true that the bigger the penalty factor the better the effect. When the penalty factor exceeds the critical value and if the value becomes bigger, it plays an opposite role [13]. Finally, some studies also put forward a viewpoint that artificially changing the penalty factor is contrary to some basic principles of the kernel method and is lacking in theoretical support.

In this paper, we use the grid optimization method of cross validation on training data, to more reasonably determine penalty factor C of VSM and the parameter gamma value γ of the RBF kernel function. We cannot solve all the problems of the penalty factor. We just want to seek some reasonable theoretical basis and feasible technical means for the determination of the SVM's penalty factor in the

condition of IDS, which could get rid of the serious human experience reliance in the actual situation. We divide positive and negative samples of the training data into 10 parts, respectively, 9 of which are used as the hypothetical training data each time and one is used as hypothetical test data for training and testing. We estimate the scope of C and γ and use (C, γ) to set grid space. We also set the space and seek the optimization in the grid until obtaining the relatively ideal classification results and the correspondent penalty factor and the value of γ to train the VSM classification model.

3. Imbalanced Data Cluster Boundary Sampling Method Based on Density Clustering

For the imbalanced data, resampling is one of the important solutions to the problem of IDS classification, of which the clustering sampling method is widely accepted in recent years and has achieved good results [14]. According to the previous analysis, we get the basic criterion of imbalanced data sampling to minimize the loss of information while trying to reduce the imbalance ratio. It is a paradox to reduce the number of negative samples but also retain the amount of information as far as possible.

At present, using clustering technique as the main method of IDS sampling is one of the main means of downsampling strategy [15]. The vast majority of the idea is to delete the sample containing little information content after the sample clustering so as to achieve the purpose of reducing the imbalance ratio and retaining most information. We have studied and analyzed the sampling method in the process of studying the IDS problem and think that the overall removal of the cluster sample is not necessarily the optimal strategy. We believe that the information content carried by the sample is not evenly distributed, and there should be core information which could represent the core information state of the data and is the key to affecting the classification. Thus we put forward a hypothesis. In the cluster boundary samples, both the majority of the similarity information in the cluster and the difference information among the clusters should be reserved. We assume that the special samples (i.e., cluster boundary samples) which include both the similarity information in the cluster and the difference information among the clusters carry the core information quantity we are looking for. Through this hypothesis, this paper creatively proposes a cluster boundary sampling method based on density clustering to solve the problem of resampling in IDS classification.

3.1. Clustering Method Based on Density. As one of the widely utilized clustering algorithms, density based clustering can get rid of the restriction by data attributes, dimension, arrangement order, and spatial distribution shape and automatically identify the number of clusters with a strong ability to resist interference [16]. The clustering algorithm based on density is widely used and considers the cluster as the high density object region which is separated by the lower one in the data space, which could find the clusters of any shape and can identify the noise data.

For the clustering method based on density, the main idea is to select an object as a kernel object and query the core object's neighborhood. Once the density of the adjacent area exceeds a certain threshold, any object except the previous core would be selected as new core object to continue clustering in the neighborhood. Eventually the relatively high density region is divided into clusters from relatively low density regions and forms clusters.

Assume that a data object is described by d attributes (also called metrics or variables), and several data objects with d attributes constitute the dimensional data space. In d dimension space, data objects are called d dimensional data points, and d dimensional data points x can be expressed as $x = (x_1, \ldots, x_d)$, in which x_d represents the value of i attribute and d represents the dimension of the space (dimensionality). Consider a collection of nd dimensional data points (also known as d dimensional data sets). S could be expressed as $S = (s_1, \ldots, s_n)$, in which $s_i = (s_{i1}, \ldots, s_{id})$ and s_{ij} represents j attribute value of i data point. According to the similarity between the data points, we divide d dimensional data sets V into $\{C_1, C_2, \ldots, C_k\}$. This process is called clustering analysis, in which $k \leq n$, $C_i \neq V$ $(i = 1, 2, \ldots, k)$, and $\bigcup C_i = V$. Here C_i is generally called clusters.

Data matrix, also called object-variable structure: use p variables (also called metrics or attributes) to express n objects. For example, we could use age, height, sex, weight, and other attributes to express people. This data structure is the form of relational tables or $n \times p$ (n objects $\times$ p variables) matrix, as shown in the following:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \tag{5}$$

Dissimilarity matrix, also called object-object structure: store the proximity of n objects of all pairs. It is often shown with $n \times n$ matrix, in which $\langle \alpha_1(x), \alpha_2(x), \ldots, \alpha_n(x) \rangle$ is the measurement difference or dissimilarity between the objects i and j. In general, $d(i, j)$ is a nonnegative number. The more similar or closer the object i is to j, the closer its value is to 0; the more different the two objects are, the bigger their value is. As shown in formula (6), $d(i, j) = d(j, i)$ and $d(i, i) = 0$:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \cdots & \cdots & \cdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}. \tag{6}$$

Similarity is usually defined by the distance between the data points. The shorter the distance is, the greater the similarity will be; the smaller the distance is, the smaller the similarity will be. In an ideal situation, the distance

d_{ij} between data points v_i and v_j must meet the following conditions:

(1) $d_{ij} \geq 0$ (nonnegativeness);

(2) $d_{ij} = 0$ if $v_i = v_j$;

(3) $d_{ij} = d_{ji}$ (symmetry);

(4) $d_{ik} \leq d_{ij} + d_{jk}$, in which $v_i \neq v_j \neq v_k$ (triangle inequality).

The value of d_{ij} should meet the above conditions and range within $0 \sim \infty$. The smaller the value of d_{ij} is, the greater the similarity between v_i and v_j is. On the contrary, the bigger the value of d_{ij} is, the smaller the similarity is.

3.2. Cluster Boundary Determination Method Based on Neighborhood. The data elements obtained in the same cluster by density clustering are relatively densely distributed in the vector space and contain contents of high similarity. We believe that the data elements of the cluster boundary can effectively represent the characteristics of the data objects in the whole cluster. For the elements in the data space, it can correspond to the points in n dimension space. To be more precise, the arbitrary data elements x could be expressed as the vector form with the following feature, and the standard Euclidean distance is taken as the distance between two vectors:

$$\langle \alpha_1(x), \alpha_2(x), \ldots, \alpha_n(x) \rangle, \tag{7}$$

where $\alpha_k(x)$ represents the k attribute of the instance x. Then the Euclidean distance between two instances x_i and x_j is

$$d(x_i, y_j) = \sqrt{\sum_{k}^{n} \left(\alpha_k(x_i) - \alpha_k(x_j)\right)^2}. \tag{8}$$

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

$$\mathrm{EPS}(x) = \{y \in D \mid d(x, y) \leq \mathrm{EPS}\}. \tag{9}$$

The method is based on the definition of the neighborhood to determine the boundary points of the cluster. In the same cluster, for an element, the more this kind of element is contained in the neighborhood, the closer this element is to the center of the cluster; and also the less this kind of element is contained in the neighborhood, the farther this element is to the center of the cluster. We could use $|\mathrm{EPS}(x)|$ to represent the number of data element x in the neighborhood. In order to find the boundary of the cluster more accurately, we chose two groups of density threshold. One group is called clustering density threshold, which is based on the characteristics and the average distance of the whole data set. It is used to divide the whole data into several clusters. The other is called the boundary density threshold, which is estimated by the scale of each cluster. It is used to find the boundary data objects. We use the clustering density thresholds EPS_1 and MINP_1 of the first group to find similar data elements in the data set and then divide these data elements into several clusters C. For each cluster C_i, we use the boundary density thresholds EPS_{C_i} and MINP_{C_i} of the second group to find the cluster boundary ring. The determination of the boundary density thresholds is based on the scale of cluster C_i. In this paper, if we use D to represent the whole training data set, C_i to represent i cluster divided in D, and B_i to represent the boundary ring of cluster C_i, then we have

$$\begin{aligned} D &= \{C_1, C_2, C_3, \ldots, C_n, C_{\mathrm{noise}}\}, \\ C_i &= \{x \in D \mid |\mathrm{EPS}(x)| \geq \mathrm{MINP}_1\}, \\ B_i &= \{x \in C_i \mid |\mathrm{EPS}(x)| \geq \mathrm{MINP}_{C_i}\}. \end{aligned} \tag{10}$$

Details of the implementation of the algorithm are as follows:

(1) Traverse the data elements in D and calculate the distance between the elements in D.

(2) Estimate the clustering density threshold MINP_1.

(3) Use the density threshold of first group to cluster D.

(4) Mark the elements which belong to cluster C_i or noise C_{noise} in D.

(5) Calculate the number N_{C_i} of data elements in one cluster C_i.

(6) Estimate the density threshold MINP_{C_i} of cluster C_i according to N_{C_i}.

(7) Calculate the number of each of the data elements which belong to the same cluster in one certain neighborhood.

(8) Extract the boundary elements B_i from cluster C_i according to the density threshold MINP_{C_i} of the second group.

(9) Repeat the fourth step until all the clusters where the nonnoise elements are have been traversed.

(10) Get all B_i.

The positive and negative samples in IDS are distributed in an imbalanced manner. The data of high imbalance ratio has a centralized distribution. The gap between the numbers of positive and negative samples is always huge. So when extracting the cluster boundary ring of the imbalanced data, we need to ensure that the information of positive samples in the minority is as complete as possible, while the information of negative samples in the majority is as representative as possible. Here we retain all the information of positive samples, only to cluster the negative samples information and extract the cluster boundary. Finally, boundary samples of the entire positive and all negative samples clusters are used as the learning data of the SVM classification.

4. Experimental Verification and Analysis

The IDS has two internal factors: the imbalance ratio and the lack of information. Imbalance ratio refers to the ratio of large and small categories, which represents the degree of data imbalance. The lack of information is the data content

Table 1: The basic information of four UCI data sets.

Data set	Number of negative samples	Number of positive samples	Imbalance ratio	Data description
Shuttle	57829	171	338 : 1	High imbalance ratio High information amount
Abalone	4145	32	130 : 1	High imbalance ratio Low information amount
Yeast	1433	51	28 : 1	Low imbalance ratio Low information amount
Churn	4293	707	6 : 1	Low imbalance ratio High information amount

in a sample of small class, which represents the information content of a small class in the data set. In order to verify the performance of the method in this paper, we select 4 groups of the open data set on UCI public data platform as the experimental data (Table 1 lists the basic information of the 4 data sets), which represents the four possible situations that the imbalanced data has. Using these data sets can reflect the characteristics of IDS from all aspects, which could verify the validity and feasibility of this experimental method.

4.1. Evaluation Metrics of Imbalanced Data Set Classification. The IDS classification is a typical two-classification problem. When doing the classification work, we usually define the smaller categories with fewer data samples as positive samples and the larger categories with more data samples as negative samples. We can describe the results of classification as four situations: the rightly classified positive samples TP, the wrongly classified positive samples FP, the wrongly classified negative samples FN, and the rightly classified negative samples TN. The total number of positive samples P = TP + FN, while the total number of negative samples N = TN + FP. Based on these values, we can get the traditional classification evaluation standards, accuracy, precision, recall, and the calculation formula of F's value, as follows:

$$\begin{aligned} \text{accuracy} &= \frac{(\text{TP} + \text{TN})}{P + N}, \\ \text{precision} &= \frac{\text{TP}}{(\text{TP} + \text{FP})}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}. \end{aligned} \tag{11}$$

We can see that when it comes to the problem of IDS classification, since these commonly used evaluation indicators cannot strongly change the classification distribution, the formula no longer has good performance and even fails (such as for a data set with the proportion of positive and negative samples as 1 : 100, as long as we see the test data as negative samples, the accuracy rate is above 99%). Some studies have also used the accuracy and recall rate as the main evaluation metrics. But the blind pursuit of classification accuracy of positive samples will lead to bad classification results on negative samples, which is not what we would like to see. Thus these traditional evaluation metrics cannot be scientific, reasonable, and accurate to show the performance of the IDS classification effect.

In recent years, the latest research shows that using ROC curve and AUC to evaluate the effect of IDS classification has obvious advantages since the ROC curve and AUC are not affected by the imbalanced distribution of data types. This means that when the number of positive and negative samples of the test data is changed, ROC curve and AUC will not change accordingly, which could evaluate the classification effect in a more scientific and intuitive way [17].

ROC curve is a two-dimensional curve, in which horizontal coordinate represents FPR (false positive rate) and longitudinal coordinate represents TPR (true positive rate). The more the test data we get the more smooth the ROC curve will be. In the ROC curve, the curve X is better than Y if X is always above Y, which means, for all the possible wrong classification costs and class distributions, the expected cost of the classifier corresponding to X is always lower than that of Y.

Although the ROC curve can be intuitive to show whether the classification results are good or bad, in practical application, it is still hoped to use a method of numerical description to evaluate the classification results. As shown in Figure 3, if the two ROC curves X and Y intersect, we can only find that X is better than Y when X is less than 0.23, and FPR is better than Y when FPR is bigger than 0.23. If we only use the ROC curve to measure, it is difficult to explain whether X or Y has better classification effect, not to mention explaining how big the difference is between X and Y. To solve this problem, we can calculate the area (the value of AUC) under the ROC curve, which would be more intuitive and clear to present the good and bad sides of the classification effects:

$$\text{AUC} = \int_0^1 \frac{\text{TP}}{P} d\frac{\text{FP}}{N} = \frac{1}{P \cdot N} \int_0^N \text{TP}\, d\text{FP}. \tag{12}$$

4.2. Comparative Experiment and Analysis. In this paper, four methods are used to carry out experiments on four different data sets to verify the role of the proposed two strategies in classification. In the experiment, 50% of each data set is used for training, and the remaining 50% is used for testing. This paper also uses the method of rotation test and ensures that

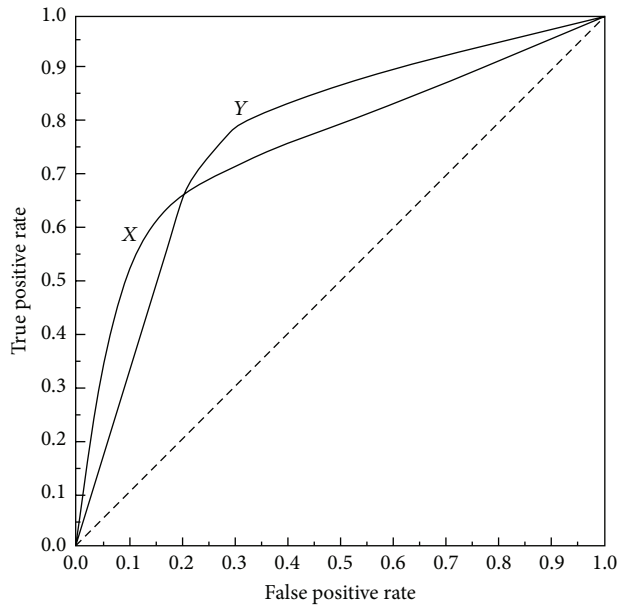

Figure 3: Two ROC curves X and Y.

the training and testing data have the same imbalance ratio. The detailed strategies of the three methods are as follows.

Method 1: classify SVM model using ordinary RBF kernel function.

Method 2: on the basis of method 1, use the grid optimization of 10-fold cross validation to determine the penalty factor and the gamma parameter before training the SVM classification model.

Method 3: on the basis of method 2, use SBC (based cluster SBC) for sampling.

Method 4: on the basis of method 2, use the clustering boundary sampling algorithm proposed in this paper to resample the training set.

By comparing the experimental results of methods 1 and 2, we can verify the effect of changing the penalty factor on the classification of imbalanced data; by comparing the experimental results of methods 2, 3, and 4, we can verify the effect of resampling on the classification of imbalanced data; by comparing the experimental results of methods 3 and 4, we can verify the effect of cluster boundary sampling algorithm on the classification of imbalanced data. The four methods are used to carry out comparison experiments on 4 data sets. The detailed experimental results of ROC curves are shown in Figure 4.

The experimental results of Figure 4 have clearly shown the advantages of this method compared to the traditional SVM classification algorithm. In order to further compare on the quantifiable level, this paper calculates the values of AUC in the four data sets with the four methods, as shown in Table 2.

According to the comparison of the experimental results, we can get some very meaningful observations. (1) For the imbalanced data with high imbalanced ratio, if we directly use

Table 2: The AUC value of three methods in different data sets.

Data set	SVM	C-SVM	SBC-CSVM	BOUND-CSVM
Shuttle	0.5091	0.5792	0.5847	**0.7260**
Abalone	0.5159	0.5332	0.5556	**0.6877**
Yeast	0.5540	0.7776	0.5769	**0.8213**
Churn	0.8745	0.9013	0.8956	**0.9151**

the SVM model for classification, the effect is really poor. In method 1, the AUV values are 0.5091 and 0.5159, respectively, in data sets Shuttle and Abalone with high imbalance ratio, which are almost invalid, while in data sets Yeast and Churn with low imbalance ratio, its performance is accepted. This experimental result validates the existing conclusion. In case of serious IDS, using the SVM method directly has no good effect, but it has good adaptability if the data is not seriously imbalanced. At the same time, it is proved that the data set selected in this paper is representative and scientific. (2) According to the SVM classification of imbalanced data, reasonable adjustment in the value of penalty factor can effectively improve the classification effect, and the higher the data imbalance ratio is, the more obvious the effect is. This conclusion is easy to get by comparing the classification experimental results in the four data sets with methods 1 and 2, which further proves that the penalty factor is a practical method in reality to improve the classification effect. (3) Traditional SBC is not stable. The traditional SBC assumes that the elements in each cluster are highly similar to each other, carrying the same amount of information. But it is proved that the SBC method clustering cannot guarantee the good results every time. Sometimes, the classification effect is reduced due to the serious loss of information. This fully shows that the amount of information carried by the cluster is unequal and some elements should carry more core information which is good for classification. (4) Clustering boundary sampling method has good stability, which is an effective way to solve the imbalanced data classification. By comparing the experimental results in the four data sets with methods 2, 3, and 4, we can see that the method of cluster boundary sampling has a stable effect on improving the classification effect of IDS in different situations. The improving effect is more obvious in the case of serious imbalanced data, which proves that this method is a feasible technique in practical application. (5) Samples on the clustering boundary should carry more information. Through the comparison of methods 3 and 4, there is only a slight difference in the number of removal samples. That is to say, the two methods almost have the same ability to reduce the imbalance ratio, but it is clear that method 4 is more stable and effective. Thus we get a hypothetical conclusion of great significance. That is, the information content carried by samples is imbalanced. There should be more some kind of core information which is more helpful to classification in the information content; and the clustering boundary samples should be the special samples which carry this kind of core information. We will further study and explore the theoretical basis of this hypothesis.

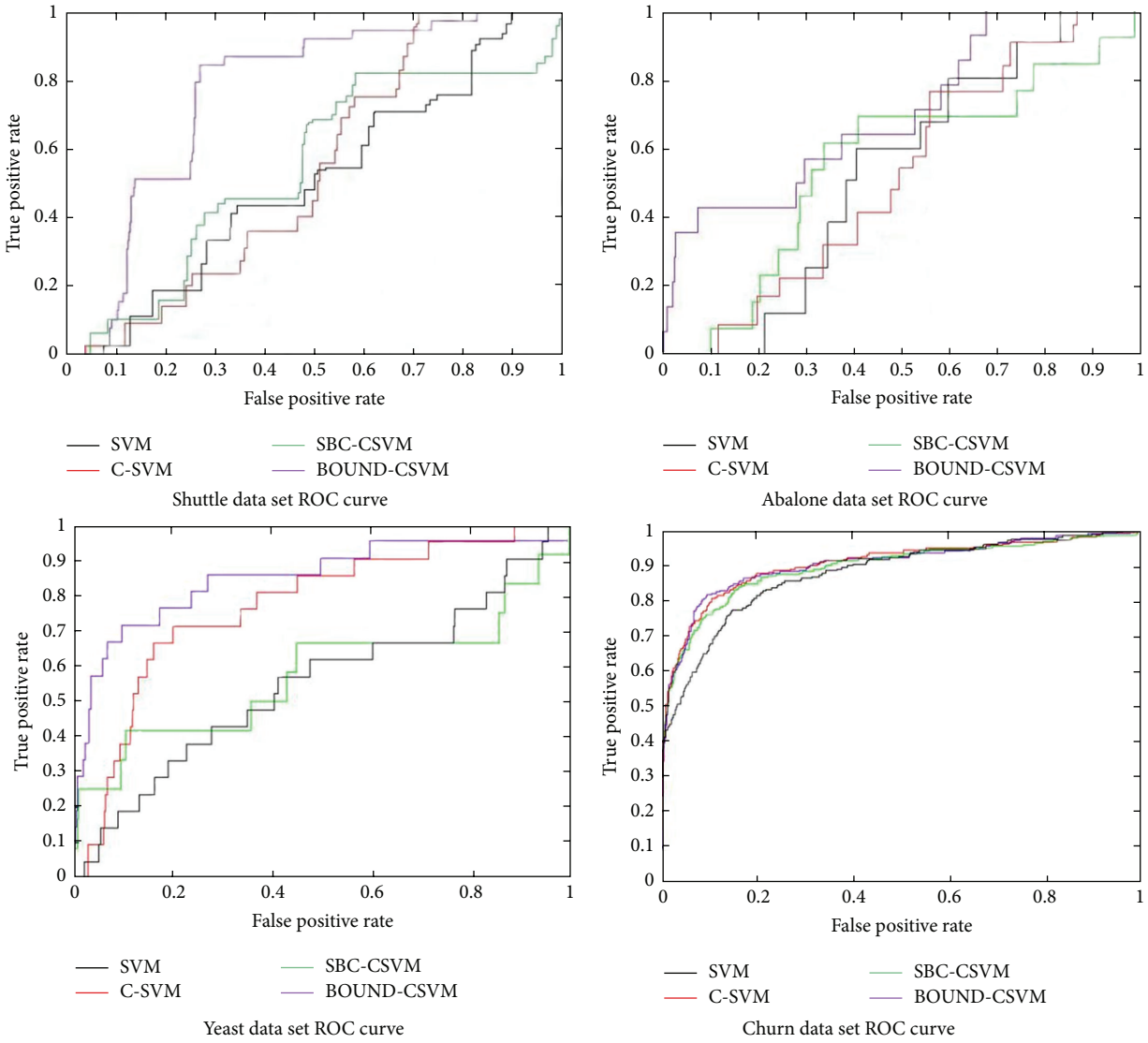

Figure 4: ROC curve of four UCI data sets.

5. Conclusion

In order to address the IDS classification problem, this paper proposes a CSVM classification method based on cluster boundary sampling, which provides an integrated solution to the problem of IDS classification. There are two main contributions of this method. One is the use of grid optimization method of cross validation for the training data to determine the penalty factor of SVM and kernel gamma value. This method does not solve all the drawbacks of the adjustment of the penalty factor, but at least it seeks a more reasonable theoretical basis for the determination of SVM penalty factor. At the technical level, it also provides a practical means of realization. The other is to propose a cluster boundary sampling method based on density clustering. We resample the IDS, which not only effectively balance the data skew state, but also greatly reduce the number of support vectors. It betters the classification effect while significantly improving the classification speed. This sampling method overcomes the shortcomings of the traditional sampling method including the lack of theoretical basis, strong randomness, human subjective interference, and serious information loss. At the same time, it is a good solution to the aliasing phenomenon in data, which can improve the generalization of the following SVM classifier. On a representative public data set, the proposed method is proved to be very stable in improving the classification of IDS. The improvement performance is especially more obvious in the case of high imbalance ratio.

In the future work, the research on the classification of IDS should be further explored in the following aspects. First, resampling is still a major method to solve the problem of imbalance in the world and we need to explore more effective ways to reduce the deviation of data as much as possible while minimizing the information loss. Second, we need to further study the information content and explore whether hypothetical core information content exists. At the same

time, a further study of the theoretical basis for the validity of cluster boundary sampling is also necessary. In the end, we need to study the special kernel function to adapt to the classification of imbalanced data so that the classification algorithm can accommodate to the data imbalance in theory, which could completely solve all the problems of regulating the penalty factor.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This paper is partially supported by Natural Science Foundation of Province (QC2013C060), Science Funds for the Young Innovative Talents of HUST (no. 201304), China Postdoctoral Science Foundation (2011M500682), and Postdoctoral Science Foundation of Heilongjiang Province (LBH-Z11106).

References

[1] W. Wei, L. Jin-jiu, C. Long-bing, O. Yuming, and C. Jiahang, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, no. 4, pp. 449–475, 2013.

[2] L.-F. Zhou and H. Wang, "Loan default prediction on large imbalanced data using random forests," *Telkomnika*, vol. 10, no. 6, pp. 1519–1525, 2012.

[3] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: the effect of sampling methods," *Knowledge-Based Systems*, vol. 41, no. 3, pp. 16–25, 2013.

[4] D. K. Antwi, H. L. Viktor, and N. Japkowicz, "The PerfSim algorithm for concept drift detection in imbalanced data," in *Proceedings of the 12th IEEE International Conference on Data Mining Workshops (ICDMW '12)*, pp. 619–628, Brussels, Belgium, December 2012.

[5] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: a medical knowledge driven approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 96–104, 2013.

[6] H. Yi, Y.-F. Liu, G.-M. Zhang, B. Jiang, and X. Song, "Fault diagnosis in condition of imbalanced samples using candidate set relevance vector machine," *ICIC Express Letters*, vol. 7, no. 2, pp. 505–512, 2013.

[7] H.-L. Yu, J. Ni, and J. Zhao, "ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013.

[8] Y. Zhang, D. Zhang, G. Mi et al., "Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions," *Computational Biology and Chemistry*, vol. 36, pp. 36–41, 2012.

[9] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, 2004.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[11] Y.-Z. Shi, S.-T. Wang, J.-X. Zhang, and T.-G. Ni, "An evolutionary support vector machine for non static data classification," *Journal of Electronic and Information*, vol. 35, no. 6, pp. 1413–1420, 2013.

[12] E.-H. Zheng, C. Zou, J. Sun, L. Chen, and P. Li, "SVM-based cost-sensitive classification algorithm with error cost and class-dependent reject cost," in *Proceedings of the 2nd International Conference on Machine Learning and Computing (ICMLC '10)*, pp. 233–236, Bangalore, India, February 2010.

[13] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2012.

[14] W. Prachuabsupakij and N. Soonthornphisaj, "Clustering and combined sampling approaches for multi-class imbalanced data classification," *Advances in Information Technology and Industry Applications*, vol. 136, no. 10, pp. 717–724, 2012.

[15] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[16] Y.-W. Yu, Q. Wang, J. Kuang, and J. He, "An online clustering algorithm based on density of spatial data stream," *Automation Journal*, vol. 38, no. 6, pp. 1051–1058, 2012.

[17] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.