Research On Classification Method Of High-Dimensional Class-Imbalanced Data Sets Based On SVM

Chunkai Zhang, Jianwei Guo, Ying Zhou, Junru Lu

*School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, People’s Republic of China*

[ckzhang812@gmail.com](mailto:ckzhang812@gmail.com), [guojianwei\_15@163.com](mailto:guojianwei_15@163.com), 1712028923@qq.com

Abstract In recent years, the problem of classification for high dimensional and class-imbalanced data is found in many fields like bioinformatics and so on. High dimensional problems result in bad classification results because some combinations of features have adverse effect on classification; Class-imbalanced problems mean the number of samples of one class is more than another class, which would make the classifier concerns the majority class more but the minority less. The two problems are both exist in high dimensional and class-imbalanced data sets. Many researchers study high dimensional problem and class-imbalanced problem separately and come up with a series of algorithms, but they ignore the new problem arising from the mutual influence of class-imbalanced problem and high dimensional problem, the new problem indicates high-dimensional problem affects sampling process while imbalance problems disturbance feature selection. Firstly, this paper analyses the new problem arising from the mutual influence of the two problems, and then introduces SVM and analyses its advantages on dealing high dimensional problem and class-imbalanced problem. Next, this paper proposes a new algorithm named BRFE-PBKS-SVM aimed at high dimensional and class-imbalanced data sets, which improves SVM-RFE by considering the class-imbalanced problem in the process of feature selection and improves SMOTE so that the procedure of over-sampling could work in the Hilbert space, meanwhile, the over-sampling rates are set adaptively. Finally, the result of some experiments shows the performance of this algorithm.

Keywords high-dimensional· class-imbalanced· feature selection· boundary samples·over-sampling

# **Introduction**

Classification is a common task in pattern recognition and machine learning. So far, researchers have dealt with classification problem and proposed lots of methods such as support vector machine, decision tree, and neural network and so on[1]. However, there are many problems in dealing with imbalanced data and high dimensional data. Imbalanced data refers to such a type of data set: in the sample space, there is a significant difference in the number of samples of one class with other classes of samples. However the minority class of imbalanced data contains more valuable information, so we need to pay more attention to it. But when samples of majority class in a training data set vastly outnumber samples of the minority class, traditional data mining algorithms tend to ignore the minority class because of the global accuracy, which results in bad classification of minority class, so they should be improved such as bagging[2], and improved neural networks[3]. High dimensional data refers to a dataset with a huge number of attributes in the sample space. High dimensional data will result in a time-consuming training process because the excessive number of attributes is related to vector computation, and the computation time increases exponentially when dimensionality increases, and this is the "Curse of Dimensionality"[4]; besides, there are still some high dimensional data redundancy and high correlation feature combination, the existence of these features can not only worsen the performance of classification[5], but also lead to overfitting[6]. For the high-dimensional problem, a series of dimensionality reduction methods[7,8] and feature selection method[9–12] have been developed by scholars.

Aiming at the problems of imbalance and high dimension, many scholars have done some researches and put forward a lot of efficient algorithms separately. But in recent years, data from many fields such as bioinformatics[7,8], image classification[13] show the two characteristics of high dimension and imbalance. Many researchers began to study this new type of high dimensional imbalanced data[14, 15].

This paper mainly studies the binary classification problems of the data set with characteristics of high dimension and imbalance, and proposes a new algorithm named BRFE-PBKS-SVM, which improves SVM-RFE by considering the class-imbalanced problem in the process of feature selection and improves SMOTE so that the procedure of over-sampling could work in the Hilbert space, meanwhile, the over-sampling rates are set adaptively. So high dimensional and imbalanced problems can be solved simultaneously.

# **Related works**

In this section, we analyze the problems of high dimension and imbalance, and then solutions for these problems are briefly introduced, finally we will give a conclusion for high dimensional and imbalanced problems.

## **The high-dimensional and class-imbalanced problem.**

In high dimensional and imbalanced problems, some feature sets show advantages in the recognition of the majority class while they are not conducive to the recognition of the minority class. Under normal circumstances, the classifier pays more attention to majority class than that of the minority class, so high dimensional problems may lead to shaper imbalance problems. For example, in CTR task, the total number of extracted features may be very large, while positive samples and advertising click behaviors are very scarce, and in selecting the features, if the feature sets are not conducive to the recognition of the minority class, it leads to a poor prediction task effect.

There are two main difficulties for classification of high dimensional and class imbalanced data sets. The first problem is that traditional methods cannot restore the relationship between features in high dimensional data after resampling. First of all, feature sets after feature selection without considering the imbalance problem may be not necessarily beneficial to the recognition of the minority class. For example, experiments in [16] show that although the SMOTE method allows the classifier to increase the concern for minority class in low dimensional data, but it doesn’t work effectively in high dimensional data. Second, without considering the effects of imbalanced problems in feature selection, we may easily choose the feature sets which are not conducive to the recognition of minority class [17], and complete feature selection algorithm (LASSO algorithm[18]) can directly eliminate some features which have an important effect on the recognition of minority class, the iterative approach improves the reverse feature elimination method, it selects the feature according to the result of classifier, in each round, each feature is eliminated by a classifier to determine their contribution to the final results, but the iterative approach cannot prevent selecting the feature sets which are conducive to the recognition of majority class either. In conclusion, high-dimensional problem affects sampling process while imbalance problems disturbance feature selection.

## **Solutions for the high dimensional problem**

The processing solutions of high dimensional data mainly include dimensionality reduction and feature selection, LDA (linear discriminant analysis), a classic supervised dimensionality reduction method, has been widely used in high dimensional problems; the classic PCA (principal component analysis) is a kind of method for dimensionality reduction, PCA considers the projection direction according to the distribution of variance in different directions in the original feature space, so that the dimensionality reduction can keep the variance of the distribution.

According to the relationship between feature selection process and classifier training process, the current feature selection methods can be divided into three categories: filtered feature selection, wrapped feature selection and embedded feature selection.

The model training and feature selection process are irrelevant in filter feature selection. The filter feature selection algorithms firstly select some features according to certain rules, and the feature subsets are not limited to the final classifiers.

In contrast to filter feature selection, the wrapper feature selection takes classifier’s feature evaluation as the reference standard of feature selection, and the final feature subset is most beneficial to improve the classification efficiency of certain algorithm model.

Embedded feature selection chooses feature sets according to some parameters of the model after training the model, it is a one-time feature selection process, while the model training is completed, and algorithm completes feature selection.

## **Solutions for the class imbalance problem**

The classification task of imbalanced data mainly consists of two levels: data sampling and classification algorithm. Data sampling is one of the most important methods to solve the problem of imbalanced data from sample space, by reconstructing the numbers of samples in different classes through under sampling, resampling and mixed sampling method, the imbalanced data tend to balance in quantity and reduce the effect of data imbalance on the classification, preventing the classifier to pay too much attention to the majority class to achieve a high global classification accuracy and ignore the minority class which is more important to researchers [19].

A large number of experiments show that the classification results of imbalanced data can be significantly improved by sampling method. Sampling methods have been developed so far, and they have been widely used in imbalanced classification.

At the algorithm level, the classification method mainly focuses on finding more sensitive classification methods for the imbalanced data, which makes the classifier pay more attention to the minority class. In recent years, with the efforts of scholars, some methods, such as cost sensitive learning and ensemble learning, have been proposed. The classification of imbalanced data has achieved satisfactory results.

## **A summary of classification methods for high dimensional imbalance problems**

So far, the approaches to deal with the problems of imbalance and high dimension are initially divided into two groups depending on the order of overcoming the problems of imbalance and high dimension: solving high dimension problems first and solving imbalance problems first. These solutions decompose the high dimensional imbalanced problems into high dimension problems and imbalance problems and resolve them separately. Rok Blagus and Lara Lusa et al. have studied the effect of the SMOTE method on the classification of imbalanced data in high-dimensional and low-dimensional situations[20]. A large number of experiments show that it doesn’t work if we use SMOTE and its derivative methods firstly to solve the imbalance problem and then classify high-dimensional data. There is also a problem in setting the sampling rate when we use the resampling method in solving the problem of imbalance, and there is a risk of losing precious samples when using under sampling method.

\*This work is supported in part by National High Technology Research and Development Program of China (No. 2015AA016008)

By first dealing with the problems of high dimension and then solve the problem of imbalance, we can avoid the problem that the traditional sampling method has no obvious effect on high dimensional data. The main purpose of the feature selection is to find the optimal combination of features, so that the classification efficiency can be improved. In order to classify high dimensional imbalanced data, many researchers solve the problem of imbalance by selecting optimal feature combination firstly and then a series of methods are adopted to solve the problem of imbalance. Richard Weber study the classification results of the combinations of different feature selection methods and SMOTE algorithm[17]. Through the rigorous experimental setup, Richard Weber shows the superiority of SVM in solving the problem of classification of high dimensional imbalanced data. However, since sampling is performed after the feature selection, there is no guarantee that each feature selection is an optimal choice in the case of an imbalanced dataset. And there are limitations of SMOTE algorithm and its derivative algorithms in high dimensional space.

As the problems of high-dimensional and imbalanced problems are intertwined, new problems are formed. We cannot solve the problem of imbalance at one time in the process in dealing with high-dimensional problems. At the same time, we cannot ignore the high-dimensional problem in solving the problem of imbalance, so the classification methods for high dimensional imbalanced data need to be improved.

# **BASIC CONCEPTS**

## **Basic theory of SVM**

SVM's formula is as follows:

 

The formula for separating hyperplane is:

 

The goal of SVM is to solve the following problems:

 

 

The formula (3)(4) can be transformed into a dual problem, such as the formula (5):



The Lagrange function is introduced, and the solution of SVM is as follows:

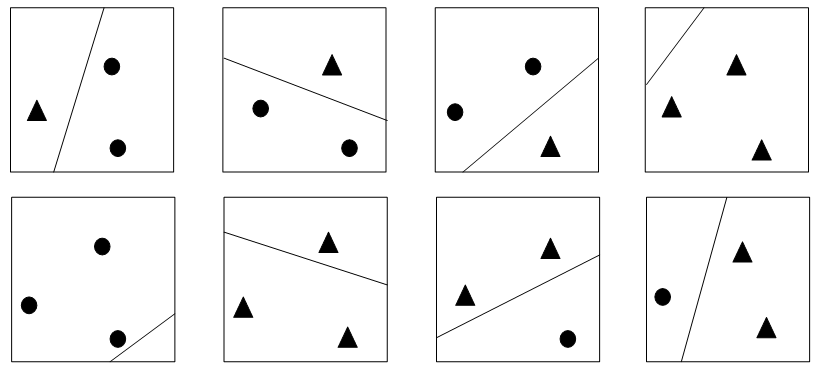
 

 

Where C is the penalty parameter, and is the Lagrange parameter, which can indicate if sample points are support vectors, i.e., the boundary sample [21]:

1. If =0, the sample points are correctly classified but located outside the plane, and they belong to the security sample.
2. If 0<<*C*, the sample points are correctly classified but fall on the edge of the interval, and they belong to the boundary sample.
3. If =*C*, the distance between the sample points and the separation hyperplane are smaller than the geometric margin, the sample points is between the two margins boundaries and may be misclassified.

Support vector machines (SVM), which is based on the VC dimension theory in statistical learning theory, pursues structural risk minimization, which makes it widely used in the field of high dimensional pattern recognition. The VC dimension [22] proposed by Vapnik and Chervonenkis is a description of the expressive power of the SVM hypothesis space. The VC dimension shows the relationship between the total number of samples which can be recognized by SVM correctly and the total number of features in the feature space: for the total N features in sample space, the VC dimension is equal to N+1, when the distribution of sample capacity’s dimension is less than or equal to the VC dimension, SVM will has a correct classification result. For example, the VC dimension in two dimensional Euclidean space is 3, then under this circumstance, for any 3 non collinear samples of 2 dimensional distribution of the portfolio (8 total), there is always a straight line can classify them correctly, as shown in figure 1.



**Fig.1.** combinations of two kinds of samples in dimensional Euclidean space

Therefore, the more features in the feature space, the higher the VC dimension is and the stronger expressive power the SVM has; if the feature number in the feature space is much larger than the total number of samples or an infinite combination, then there is VC (F) =∞. Therefore, SVM has unique advantages in solving the classification problem of high-dimensional data and the classification of small sample data.

Therefore, when we are solving the problem of classification of high dimensional imbalanced data by SVM, the problem of imbalanced data is mainly reflected into the problem of imbalanced boundary samples.

## **Synthetic Minority Oversampling Technique - SMOTE**

SMOTE (Synthetic Minority Oversampling Technique) is a popular oversampling technology[22,23]. By synthetically generating more samples of the minority class, the inductive learners are able to divide their decision regions for the minority class. SMOTE is mainly inspired by an algorithm which has been proposed in a handwriting recognition project. The process of SMOTE is as follows: for each minority class sample, its K nearest neighbors of minority class are selected; liner interpolation method is used between the minority sample and its K nearest neighbors of minority class, and a random interpolation factor is set as well as sampling rate to generate minority class imitation samples. SMOTE interpolation process is shown in the formula (8):



SMOTE algorithm’s interpolation is mainly among similar minority class samples (i.e. adjacent) and thus samples are representatively generated. Therefore, the overfitting problem can be avoided in the SMOTE algorithm, and the decision space of the minority class can also be extended better. Similarly, it can also be applied to the majority of sample space, which can reduce the decision space of the majority class. The pseudo code of SMOTE algorithm is as Table 1.

Modified synthetic minority oversampling technique(MSMOTE)[25]: it is a modified version of SMOTE, this algorithm divides the samples of the minority class into three groups, safe, border and latent noise samples by the calculation of the distances between all samples. When MSMOTE generates new examples, the strategy to select the nearest neighbors is changed with respect to SMOTE that depends on the group preciously assigned to the sample. For safe samples, the algorithm randomly selects a data point from the (same way as SMOTE); for border samples, it only selects the nearest neighbor, finally, for latent noise samples, it does nothing.

**Table 1.** Algorithm 1 SMOTE algorithm pseudo code

|  |
| --- |
| Input: Number of minority class samples T;  Amount of SMOTE N%; over sampling rate K,  Output: Synthesis minority class samples T\* |
| 1. *For i =1 to T*   Compute k nearest neighbors for i, and save the examples in the Karray  (2) *While N != 0*  (3) *SMOTE（N，i，Karray）* # According to formula (1) and sample rate K, for a minority class sample *i*, linear interpolation is performed in Karray.  (4)  *N = N – 1*  (5) *Endwhile*  (6) *Return T\** |

## **SVM-RFE feature selection algorithm**

SVM-RFE uses the weight vector to evaluate the importance of each feature in the feature space. Nested subsets of features are selected in a sequential backward elimination manner, which starts with all the feature variables and removes one feature variable at a time. The pseudo code of SVM-RFE is shown in Table 2.

In the process of feature selection, SVM-RFE evaluates the features according to a certain formula, the order of backward elimination is determined by the evaluated scores. This approach provides a feature evaluation inspiration which considers both high dimensional problem and imbalanced problem: changing the evaluation metric of features to make SVM-RFE pay more attention to the minority class when evaluating features. In each round of the process of backward elimination feature, the remaining features are re-evaluated to evaluate the importance of each feature in the existing feature space.

Because SVM-RFE is a wrapper-type feature selection method, the measure standard of the importance of each feature is the classification effect of SVM after the elimination of a feature, and its order of eliminating features is based on the sorted absolute value of the feature weight W, which provides a way to take imbalanced characteristic into account in the process of SVM-RFE: the hyperplane of SVM can be changed to correctly classify more minority class samples by resampling the boundary minority class samples. So the absolute value of w component and its order will be changed, and thus the score for features also can be changed. The order of eliminating features will be changed, so the process of feature selection will pay more attention to minority class sample.

**Table 2.** SVM-RFE

|  |
| --- |
| Input：Original data set  Out put：Data set after feature selection |
| 1. Solve the classification hyper plane of SVM, W and B are obtained, and the corresponding empirical risk Remp[ƒ] is calculated; 2. Rearrange the weight vector w=(w1,w2,…,wn)according to certain rules: 3. The importance of the features of the original data set: 4. If n=1, then go to (10); otherwise, k=1, go to (4); 5. Remove the feature xik, and set the flag tag=0; 6. Using the data set in which the feature xik is removed to train SVM, get the new w and b, record as wnew, bnew; Calculate the corresponding increment of empirical risk ▽Remp[f new], and set the initial increment of empirical risk▽Remp[f]=0; 7. If ▽Remp[f new]<▽Remp[f], set ▽Remp[f]=▽Remp[f new], tag=k; 8. If ▽Remp[f new]==▽Remp[f] and tag==0, set tag=k; 9. If ▽Remp[f new]>▽Remp[f] and k=n, go to (10); 10. If k==n, remove feature tag and get new\_train\_set and let train\_set=new\_train\_set, n=n-1, go to (3);otherwise go to (3); 11. Return data set after feature selection; |

# **BRFE-PBKS-SVM algorithm for high dimensional imbalanced data classification task**

The traditional method to solve the imbalance problems doesn’t work in high dimension space, so the existing research methods are mainly to solve the high dimensional problems and then solve the imbalanced problems, this idea is very reasonable and targeted, so this paper also use it to solve high dimensional problems, and then solve the imbalanced problem, and we also improve the existing methods according to the hybrid problems’ characteristics.

In this section, we propose a feature selection algorithm based on SVM-RFE to deal with high dimensional problems, then we improved the SMOTE algorithm in Hilbert space by using the characteristics of SVM.

## **SVM-BRFE Feature selection algorithm**

SVM is proposed by Corinna Cortes and Capnik in 1995 (SVM [26],support vector machine) has many advantages in the application of high-dimensional data classification, and has spawned a number of feature extraction method based on improved SVM, such as SVM-BFE, SVM-RFE and algorithm in [27], they all belong to the wrapped feature selection method. Therefore, in this research, we mainly study the classification method of high-dimensional imbalanced data based on SVM.

In SVM-RFE, it mainly uses weight vector of SVM  as the evaluation for the importance of feature as each component of corresponds to a feature, the bigger the absolute value of the component is, the greater the influence of the corresponding feature has on the classification results. The input of SVM is in Euclidean space, while the data is trained in Hilbert space, straight line in Euclidean space may be a higher dimensional curve in Hilbert space, what’s more, it is a usual way to define a kernel function to map the input into Hilbert space rather than a direct mapping function, So it is not feasible to determine the boundary samples by the traditional method. The value of Lagrange parameters  of each sample point indicate the relative position in Hilbert space: samples with are classified correctly and they are safe samples, samples with can be classified correctly but they are at two intervals, and samples with are located in the septal plane, and they are misclassified and boundary samples.

Using the above properties, we can use the Lagrange parameters to determine the location of sample point, we discussed advantages of SVM in imbalance classification: the safe samples have no attributes to determine the final classification hyper plane and interval plane, the original imbalance problem is mainly reflected as the boundary problem of imbalanced samples, as shown in figure 2.

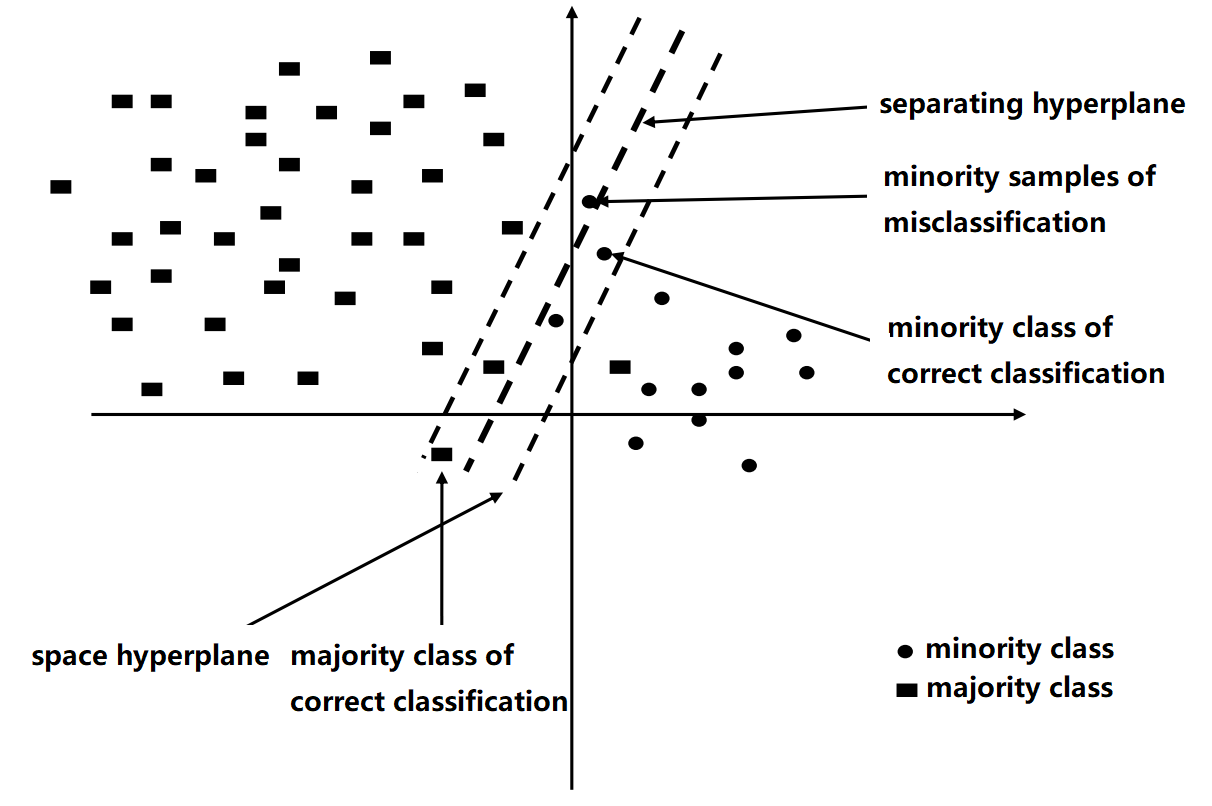


Fig.2 Sketch map of SVM geometry space

Because the sample points which have contributes to the weight vector of SVM are the support vectors between the two interval hyperplanes, we can enhance the degree of concern for the minority class of SVM by resampling the minority class samples located at the interval boundary plane which are misclassified by SVM. Because SMOTE fails in over sampling in high dimensional space, we can only solve the imbalance problem in the feature selection process by resampling. A single rate resampling of minority class samples that are wrongly divided is one such process: copying the misclassified minority class samples, thus we can improve the classifier's attention to the minority class.

Because of the unique advantages of SVM in solving the classification problem of high-dimensional data, the SVM-BRFE algorithm based on SVM is proposed by considering the influence of imbalance on feature selection.

SVM-BRFE feature selection algorithm, i.e., iterative feature selection algorithm based on boundary resampling (Border-Resampling Feature Elimination). For the high dimensional imbalanced data, the selected subsets of features tend to misclassify the minority class sample due to the imbalance problem. So we need to modify the weight vector w of SVM to increase the w’ absolute value of features which tend to increase the minority class recognition rate. The pseudo code of the proposed SVM-BRFE algorithm is as Table 3.

First of all, train SVM to get the original feature weight vector w, Lagrange parameter and values, and these three values are recorded for subsequent use; then, resample the minority class samples when and train SVM using the new resampled data set. The hyperplane of SVM will be changed to increase F1 value by training on new data set. We set sample rate to one because the boundary sample changed with the changing of SVM hyperplane. We need to repeat the process of resampling of minority sample and training SVM until the hyperplane with maximum F1 value is found. The weight vector w corresponding to this founded hyperplane will be used to evaluate features. Finally, sort features in descending according to the importance of features. The iteratively feature elimination will be executed according to the ascending order of importance of features. Each round the feature which increase the F1 value most will be eliminated. Each round SVM separation hyperplane will change after of elimination of a feature, and boundary samples will also change. So the remaining features need to be reevaluated to generate new feature weights to evaluate the importance of each feature in the new feature space.

Note that the resampling samples in the process of feature selection will not be added to the training set. The resampling of boundary minority class sample is just to get a feature weight w that is fair to both majority and minority class. So the weight w can be used to evaluate the importance of features of high dimensional and imbalanced data, rather than improve the classification effect and F1 value by directly making SVM pay more attention to the minority class. In other words, the resampling process before feature selection of each round is only to solve the high-dimensional problem that is affected by imbalanced problem, rather than to solve the problem of imbalance. Therefore, when the maximum F1 value is found, the current round of resampling is over. Record the weight vector w with which SVM achieve maximum F1 value, and then sort the features according to the weight vector w, then remove the resample replication of minority class sample and keep the original minority class samples, then do feature selection. Repeat the above process until the best feature subset is selected. As is shown in Table 3, the process of resampling doesn’t change the train set. Only in the process of feature selection we update the train set after selecting a feature.

**Table 3.** Algorithm 3 SVM-BRFE Feature Selection

|  |
| --- |
| Input：original data set  Out put：Data set after feature selection |
| 1. Solve the classification hyper plane of SVM, W, B, F1 value and lagrange parameters alpha are obtained, set new\_train\_set = train\_set; 2. For new\_train\_set, resample the minority class samples that alpha=c and train SVM. Save F1\_new and wnew, and rearrange the weight vector according to absolute value of w: 3. The importance of the characteristics of the original data set： 4. If there are still misclassification minority class samples go to (2), else go to(4); 5. Set current w as the weight that have maximum F1 value; current F1 value equals to maximum F1 value; 6. Remove the generated samples; 7. If n=1, then turn (3); otherwise, k=1, go to (7); 8. Remove feature xik,and set the flag tag=0; 9. Using the data set in which the feature xik is removed to train SVM, get the new w and b, record as wnew, bnew; Calculate the corresponding increment of F1 value ▽F1\_new, and set the initial increment of empirical risk▽F1\_now =0; 10. If ▽F1\_now <▽F1\_new, set ▽F1\_now =▽F1\_new, tag=k; 11. If ▽F1\_now ==▽F1\_new and tag==0, set tag=k; 12. If ▽F1\_now >▽F1\_new and k==n，go to (13); 13. If k==n, remove the feature ‘tag’ and get new\_train\_set , let train\_set=new\_train\_set, n=n-1, go to(2); otherwise turn to (7); 14. Return data set after feature selection. |

Through the above several steps, we find the best feature weights by resampling boundary samples to evaluate the importance of features and carry out feature selection, then update the training set and repeat the process, finally keep the most conducive feature to enhance the F1 value and other features will be removed. So the following training process can be carried out on the data set which has less redundant feature, irrelevant feature combination and is in low dimension. So for imbalanced problem, the bad effect of high dimensional problem can be reduced and the following oversampling process can be flexible carried out. The above process is beneficial to improve the traditional oversampling algorithm in the follow-up process to solve the imbalance problem and improve the classification effect.

## **PBKS Resampling algorithm**

After solving the high-dimensional problem, the disturbance of the high-dimensional problem to the imbalance problem has been greatly reduced in the high-dimensional imbalanced data, and then the imbalance problem needs to be solved. Although the resampling method can solve the problem of lacking the minority class samples, the sample boundary overlapping, as another problem of imbalanced data, cannot be directly solved by sampling method, as shown in figure 3, it is necessary to cluster the majority and minority classes to determine the boundary overlapping samples, and then use the sampling method to solve the problem [28], while SVM can automatic divide the training set in the feature space of the boundary, so it has a natural advantage to solve the problem of overlapping boundary samples. In this research, we use wrapped feature selection to solve the high dimension imbalance classification in this way we choose the feature sets which is most conducive to improve the efficiency of SVM classification feature subspace, we use the SMOTE method to solve the imbalance problem, and use SVM to deal with the high dimension classification problem.

Fig.3 boundary sample overlap

SVM is a dual problem of solving the Lagrange minimax value in Hilbert space, as shown in formula (5), by introducing the kernel function to transform Euclidean space into Hilbert space so as to solve the nonlinear classification problems in Euclidean space, and a kernel function only corresponds to a Hilbert space. The Hilbert space is used to solve the nonlinear classification problems, therefore, the hyper plane or straight line connecting two points in Euclidean space may be twisted into hyper surface or curve after mapping into Hilbert space by kernel function, and it is same when mapping the super plane or straight line in Hilbert space into Euclidean space. It is necessary to use SMOTE in Hilbert space so that the new samples in minority class are still boundary samples, in which case the separating hyperplane is helpful to improve the recognition rate of the minority class. When we use SVM, the input samples are in Euclidean space while features are computed in Hilbert space, which are different in computation space, so the new SMOTE over sampling the minority class boundary samples needs to find the original sample in Euclidean space to update the training set.

SVM implicitly map Euclidean space to Hilbert space by defining a kernel function, so the exact pre-image of a point in Hilbert space typically does not exist, and we can only settle for an approximate solution. But using the distance relationship mentioned in [29], we can obtain the corresponding input-space distance between the desired pre-images, so we will require the approximate pre-image to satisfy this distance constraints approximately so that we can find the approximate pre-image in Euclidean space.

Before solving this problem, we first propose a distance metric in Hilbert space:

 

 

Implicit mapping of Euclidean space to Hilbert space is shown in (9), and assume that the kernel function is defined as the Gauss kernel function. In this paper, we use to represent which defines inner product in Hilbert space of the two points and in Euclidean space. The square of the distance in the Hilbert space is shown in the formula (10).

When the kernel function is Gauss kernel, the relationship between the square of distance in Euclidean space and Hilbert space is shown as the formula (11) and (12). denotes the square of distance in Euclidean space and denotes the square of distance in Hilbert space.

 

 

SMOTE firstly gets K nearest samples of the sample point , and then selects a sample point in the K samples randomly and linearly interpolates between and . In this paper we only consider oversampling of boundary minority class samples, so for each boundary minority class samples in the Hilbert space, we randomly select another boundary minority class samples to be the input of SMOTE algorithm. The formula of SMOTE in the Hilbert space is shown in formula (13) where is a random number in the open interval.

 

To obtain the approximate pre-image in Hilbert space, the approximate distance constraint between sample points is very important. Suppose that when SMOTE in Hilbert space, the square of the distance between generated oversample points and all points in boundary minority class samples is as shown in formula (14) in which we assume that the number of hypothetical boundary minority class samples is k:



Suppose that is an unknown sample in Euclidean space of the originaltrain set, and the square of the distance between and the k points as shown in (15). In formula (14) and formula (15), subscript and the corresponding k sample points must be consistent.

 

As is shown in formula (16), when the kernel function is the Gauss function, with formula (14) and formula (15), the vector in Euclidean space can be mapped to the corresponding Hilbert space.



The nearer the values of the formula (14) and (15) are, the nearer the point corresponding to Gauss kernel function in Hilbert space and the point generated by SMOTE after transformation of are.

In [11], the constraint of distance between the point generated by SMOTE and its k nearest original minority class samples are considered. In order to find the boundary of the minority class, we consider the minority class samples in the classification boundary of SVM as the distance constraint of instead of the original constraint. By using the grid method we can get the approximate pre-image. Specifically, suppose that is the boundary minority class samples in Hilbert space after training a specific SVM, the upper and lower boundary of the d features of these samples can be obtained by formula (17) and (18). is lower boundary of boundary of minority class samples and is upper boundary of majority class samples.

 

 

Then according to formula (19) we can get the size of each grid. The space of boundary minority class is divided into k \* D grid, and each grid represents one position in Euclidean space. The target is to determine a nearest grid to the oversampled sample after mapping to Hilbert space. Concretely, the size of each grid is equal to the maximum value of this feature minus its minimum value and then it is divided by the total number of samples of the original boundary K. In the process of searching pre-image, the entire grid space will be searched.

 

in formula (13) is a known sample that is generated by SMOTE in the Hilbert space; in formula (15) is a unknown pre-image of . Formula (19) is the grid size of ith feature. In every random grid search, the value of each dimension adds the number of grid size which is optimized by PSO and we get . Substituting this into formula (15) and (16), the results are used to get square of cosine distance, which is shown in formula (20). Iteratively search the maximum value of square of cosine distance. Finally, the point that has largest value of square of cosine distance will be our approximate pre-image .

()

# **Experimental results and analysis**

## **Evaluation criterions**

Taking the special characteristics of the problem of imbalanced sample classification into account and using the traditional evaluation criteria, it will result in the following problems: for a higher global classification accuracy, the traditional classifier directly classifies the sample of minority class into majority class, and obtains a higher global accuracy rate, but the correct classification rate for the sample of minority class is 0, and in this case , the traditional single evaluation system will no longer be applied to the evaluation system of imbalanced sample classification. Therefore, we need some special complex consideration of various indicators to adapt to the special situation of imbalanced sample classification. There are two main categories of these criteria, one called "atomic standard", the other is called "composite standard", it is proposed after a large number of studies, which is compounded from atomic standards and mathematical theory, and they can be well adapted to the evaluation system of imbalanced sample classification problem. In addition, the receiver's curve (ROC) has been widely used in the evaluation of imbalanced sample classification [30].

As shown in Table 4, the confusion matrix for the binary classification problem involved in the classification of imbalanced samples is shown in table 4. By counting the various indicators of the confusion matrix and the compositing all these indicators, we can better count the accuracy of each category respectively, considering the classification of different categories respectively, so as to evaluate the classification algorithm of imbalanced samples, we do not blindly pursue the highest accuracy rate, but consider the classification accuracy of the minority and the majority class.

**Table 4.** Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Classified as positive | Classified as negative |
| positive | Correct positive *TP* | Wrong negative *FN* |
| negative | Wrong positive *FP* | Correct negative *TN* |

Formula (21) to the formula (24) list some of the commonly used atomic evaluation criteria for the classification of imbalanced samples based on the confusion matrix.

(21)

 

 

 

F-Measure is most widely applied to the evaluation of imbalanced sample classification, as shown in the formula (24). It is obtained by combining recall rate, precision rate and balance factor, if and only if both Recall and Precision have a higher value, F-Measure will achieve higher results. In the formula (24) the balance factor is used to adjust the computation rate of recall and precision (usually set to 1).

ROC curve (Receiver Operating Characteristics Curve) was proposed by Swets in 1988 [30] and has been widely used in many fields. FPRate is used for the X axis, and TPRate is used for the Y axis to build the space. By setting a threshold value, we can get a pseudo positive rate and the true positive rate values, and connect these scattered points, we get ROC curve lastly.

ROC curve is not able to directly evaluate the performance of classification in imbalanced samples, so in order to get a quantitative evaluation index, the coverage area AUC (Area under the ROC curve) is proposed. The classification performance of classifier can be evaluated with the area under the right bottom of ROC (that is, AUC), the greater the AUC is, the better performance the classifier has.

## **Experimental data set description**

UCI is a well-known and open machine learning database, in order to get more convincing experiment results, all data sets in this paper are derived from UCI. The used data sets are shown in table 5. Table 5 describes the specific properties of the dataset which are used for all experiments, where No. is the data set index, the Data-Sets is the dataset name, #Attr is the number of attributes contained in the dataset, and %Min represents the proportion of the minority samples. In this paper, there are 6 data sets, in which the imbalanced ratios are not same for each group, and the total sample size are not same either.

**Table 5.** Data sets

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Data-sets | #Attr. | %Min. |
| 1 | Detect Malacious Executable | 513 | 28.57 |
| 2 | SECOM | 591 | 6.64 |
| 3 | Musk | 168 | 15.55 |
| 4 | Heart Disease | 44 | 20.60 |
| 5 | Uerban Land Cover | 147 | 32.20 |
| 6 | Multiple Features | 649 | 12.06 |

There are a lot of problems in UCI database, such as value missing, inconsistent data, ambiguous category labels and therefore cannot be direct input of algorithm, they need some treatments to be clean. Data cleaning and preprocessing are often critical to the effectiveness of the final model. The main preprocessing steps are as follows:

(1) standard class switching

Sometimes, there are more than two categories of labels for a data set. For such a dataset, we take one of the categories as minority class and all the other classes of data are considered as majority class. Therefore, before the experiment is conducted, the class labels of datasets need to be converted based on the characteristics of minority class and majority class.

1. format conversion

The algorithm proposed in this paper is mainly a series of improvements to SVM, which is conducted by libsvm-3.2.1[31], so the format of data needs to be converted into a standard SVM format: for category labels, +1 means minority class sample while -1 means majority class sample, feature of samples is represented sparsely, specifically, when the feature is value, the corresponding column properties should be rewritten as , and when the feature is missing, it is set 0.

(3) padding the missing value

There are some missing value for features in the original dataset, there are two types of missing data: first, for the lack of continuous feature values, fill with average value; second, for discrete feature missing, fill it with the most frequent feature value. If a sample has a missing label, we delete it from the dataset.

(4) Normalization

In the dataset, the range of each attribute will vary, in order to eliminate the influence of numerical value on the final model and make it more convenient to compute between different features, we normalized the data by formula (25)

(25)

In the equation (25), means feature value after normalization, means the original value, means the max value in the specific feature, means the minimum value in the specific feature.

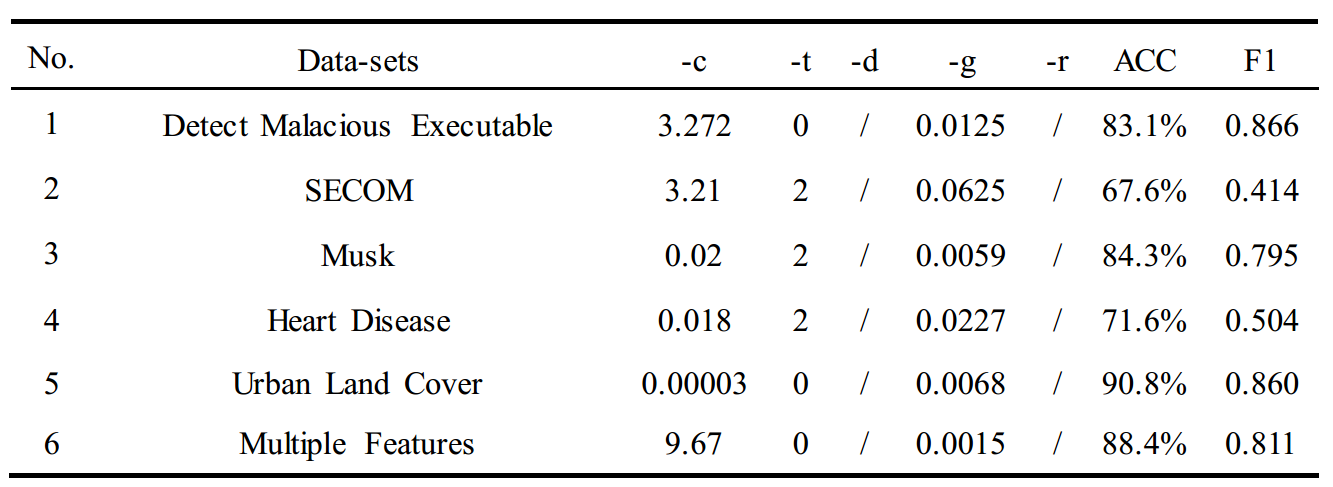
(5) Verification set partition

Some datasets in UCI database have their own training data and validation set, for this kind of data, we can input them into the model after the above 4 steps, and use the training data to train model and then use the validation data set to test the effect of the algorithm, but for other data with no given validation sets, we need to divide it into different parts and use one of them as the validation set, in this paper 20% of the original data is taken as validation set so as to test the effect of the algorithm.

The parameters have a great impact on effectiveness and superiority of an algorithm, parameter adjustment not only relates to the over fitting and under fitting of the training data, but has an influence on the results of the validation sets, and it plays a decisive role in whether it can measure a fair classification algorithm results. In libsvm-3.2.1, the parameters of two classification SVM include the type of kernel function, the penalty parameter size and the constant parameter in kernel function, in the experiments we take 5-fold cross validation and grid search to find the optimal parameters with best average classification performance, and for each dataset, once the optimal parameters are found, they will be fixed in the next experiments, and the grid search results are in table 6 as following:

In table 6, -c represents the penalty parameter, that is, the C in formula 2-10, -t means kernel function types: 0 means a linear kernel, 1 means a polynomial kernel function, 2 represents the Gauss kernel function. –d stands for the power of polynomial kernel function, -g means in the Gauss kernel function, -r is only used in the polynomial kernel, which is set to 1 by default. ACC and F1 respectively represent the result of the whole dataset using the optimal parameters in k-fold cross validation, so that the parameters we set in the training set are not over fitting the training set, and the final model is stable.

**Table 6.** Optimal parameters and results under K fold cross validation



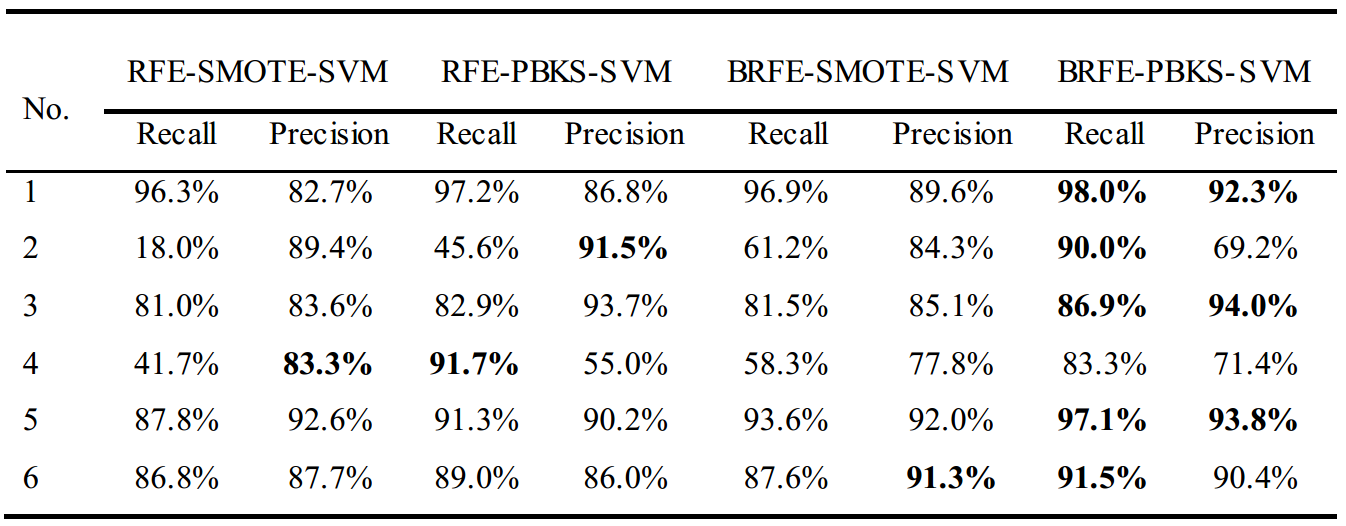
## **Validation of BRFE-PBKS-SVM algorithm**

The BRFE-PBKS-SVM algorithm is divided into two parts. The first part is the feature selection part. The second part is the data resampling part. By combining the two parts, an algorithm is proposed to solve the problem of high-dimensional imbalanced data classification. In the proposed algorithm, what the latter part need to solve is a new problem produced after solving the imbalanced problem in high dimensional imbalanced data classification task using SVM. Next, using the previously mentioned evaluation criteria, the efficiency of the BRFE-PBKS-SVM algorithm will be compared from the following 3 aspects: the improvement of the recognition rate for minority class, the improvement of the overall efficiency and the stability of the algorithm:

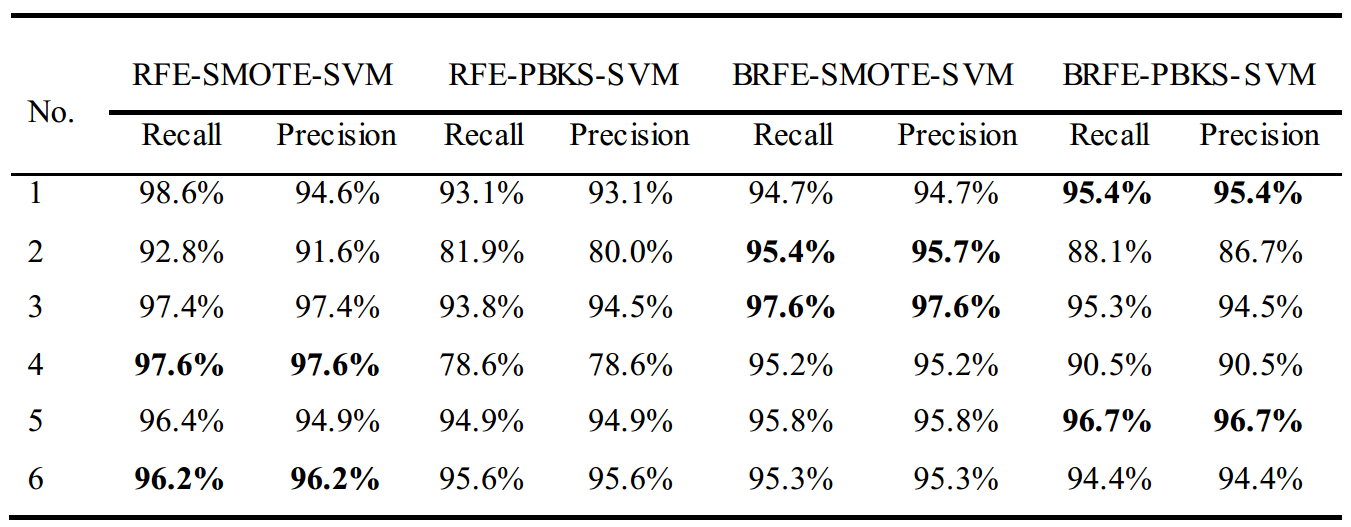
* Changes in minority recall and accuracy
* Global accuracy and changes of F1 value
* The area value of the ROC curve AUC

From table 7 we can see that the BRFE-PBKS-SVM algorithm achieves the highest recall rate of minority class in the 4 algorithms. Compared with the unimproved SMOTE algorithm, PBKS oversampling algorithm has a significant promotion of in the recall rate of minority class, but the accuracy rate has declined. As we can see in table 8, the recall of the majority class by BRFE-PBKS-SVM algorithm is hardly the highest, which shows that the BRFE-PBKS-SVM algorithm has increased the concern for minority class, but has reduced the concern for majority class.

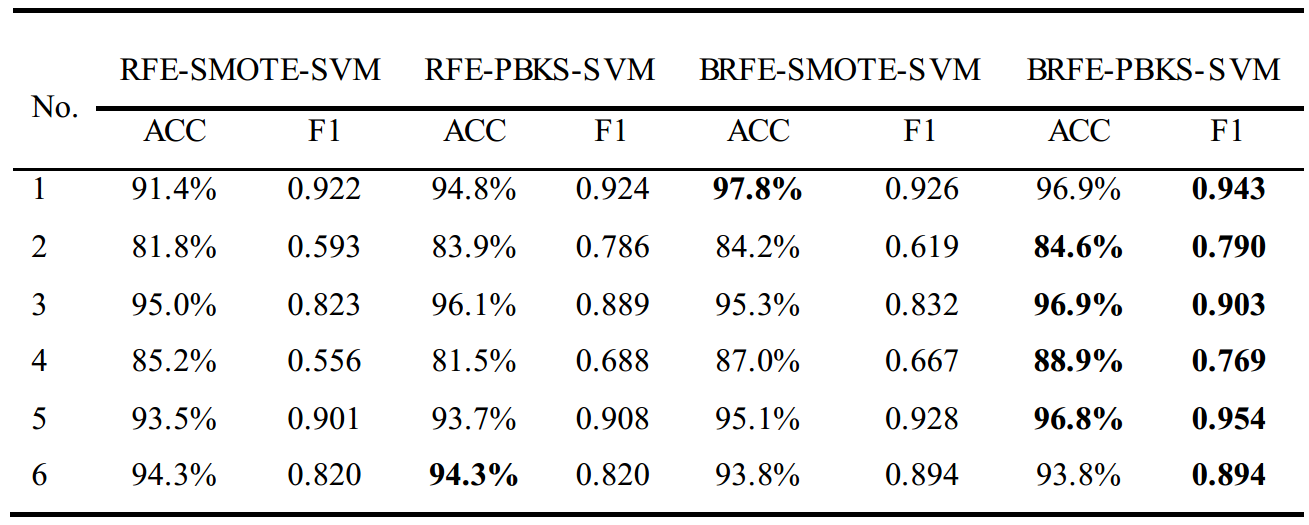
**Table 7.** Comparison of recall and precision rate of minority class

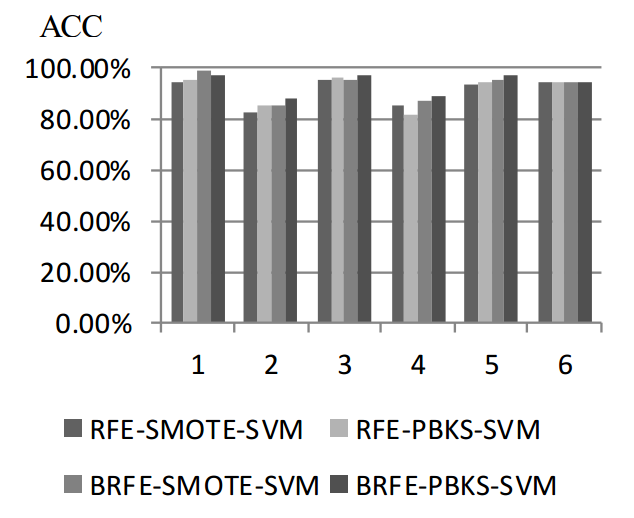


**Table** **8.** Comparison of recall and precision rate of majority class



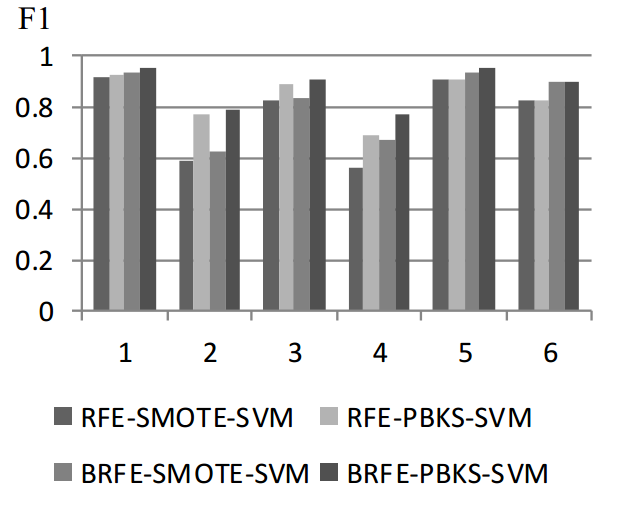
**Table 9.** Comparison of F1 value and ACC value of each algorithm





**Fig. 4.** Histogram of ACC value of each algorithm

By comparing the second columns with the sixth columns and comparing the fifth columns with the eighth columns in figure 4, we can find the effect comparison of feature selection algorithm between the SVM-RFE and the SVM-BRFE. In the second to fifth data sets, the BRFE-PBKS-SVM algorithm is optimal in all algorithms in global accuracy rate ACC; In the case of using the same oversampling algorithm, the improved BRFE feature selection algorithm achieves the best results, because the BRFE feature selection algorithm considers the imbalance problem in the process of feature elimination; In the case of using the same feature selection algorithm, the improved PBKS oversampling algorithm combination achieves the best results, because it trains data under a polynomial kernel function or a Gaussian kernel function corresponding to a Gaussian space, because the sample points generated by the oversampling of the PBKS algorithm can better fill the boundary under the Hilbert space, the spatial distribution is more reasonable, so the classification effect can be improved more.



**Fig. 5.** Histogram of F1 value of each algorithm

The ROC mentioned above is a curve which is commonly used in the evaluation of model performance, the maximum AUC is 1, which means the area surrounded by the ROC curve, a better and more stable classifier has a larger AUC value. In this section, we show ROC curves of 6 datasets with 4 algorithms, and the corresponding AUC curves are drawn with histogram.

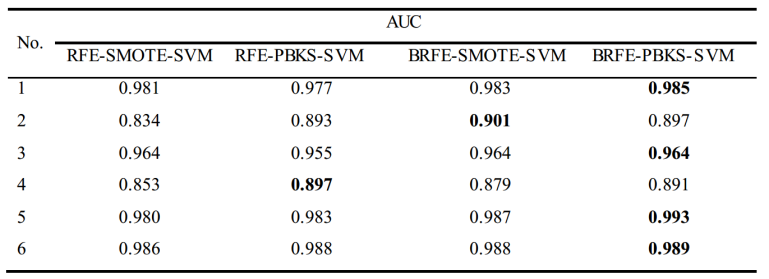
Table 10 and figure 6 show comparison chart of the AUC values of the ROC curves on the six data sets. It can be found in the figure 6 that the BRFE-PBKS-SVM algorithm can obtain the maximum AUC value except for the second and fourth data in the six data sets, in the fourth data set, even if the improved algorithm fails to obtain the optimal AUC value, the difference is only 0.006, generally speaking, the algorithm BRFE-PBKS-SVM has a good stability. Figure 7 shows that the AUC values of the four SVM-based algorithms has less difference in each data set. This also proves that SVM has better stability and superiority in the high-dimensional imbalanced data.

**Fig. 7.** ROC for Detect Malacious Executable

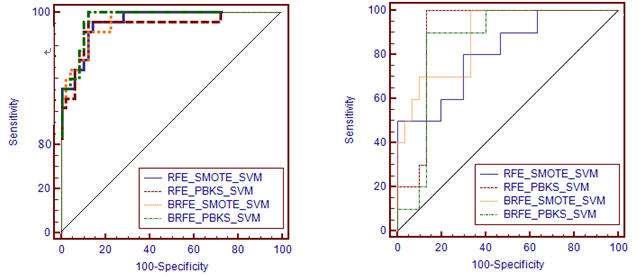
In Figures 7 to 12, the area enclosed by the lines is the AUC value in Figure 6. The diagonal represents a worst classification effect level, and its corresponding AUC value is 0.5, when a classifier in a data set on the ROC curve below this diagonal, its AUC value will be less than 0.5, which means that the classification performance of the classifier on the data set is not as good as a random guessing classifier. The ROC curve tends to the upper left, and the performance of the corresponding algorithm is more significant, AUC value is more close to 1.

According to the six ROC curves obtained from our experiment, except for the second and fourth data sets, for the other data set, the AUC values of these four algorithms are similar, and all algorithms can achieve good results and the proposed algorithm obtains the maximum AUC value for the four data sets. For the second and fourth data sets, these four algorithms have different effects, and their ROC curves are not smooth enough. Though BRFE-PBKS-SVM algorithm cannot achieve the best results for the two data sets, it can achieve compared classification result to best algorithm. Generally, the experiment results show that the BRFE-PBKS-SVM algorithm based on SVM is stable and effective to solve the classification problem for high dimensional imbalanced data.

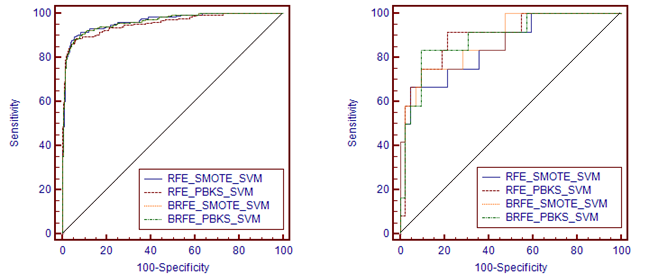
**Table 10.** Various algorithms combine the AUC values on each data set



**Fig. 6.** Histogram of AUC value of each algorithm

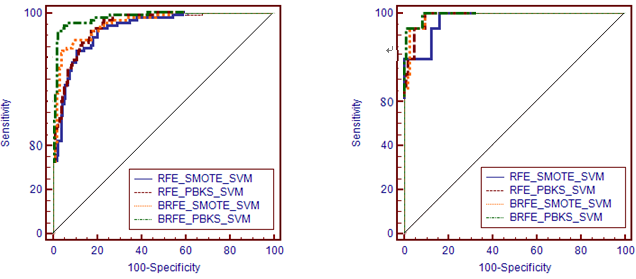


**Fig. 8.** ROC for SECOM



**Fig. 10.** ROC for Heart Disease

**Fig. 9.** ROC for Musk



**Fig. 12.** ROC for Multiple Feature

**Fig. 11.** ROC for Urban Land Cover

# **CONCLUSION**

High dimensional imbalanced data has attracted the attention of many scholars in recent years. Most of the researches solve the two problems separately by firstly carrying out feature selection and then data sampling, which didn’t take the mutual influence of the two problems into consideration, and in the process of solving high dimensional problem by feature selection, the influence of imbalanced problem has not been considered. In the subsequent sampling process, the effects of high dimensional data on the imbalanced problem also has not been considered. In this paper, according to the characteristics of high dimensional imbalanced data, we improve the feature selection and data sampling method, and propose a novel algorithm to solve the high dimensional imbalanced data classification problem:

(1) In the part of feature selection, an improved feature selection algorithm named SVM-BRFE is proposed, which takes the imbalance problem of data into account. The boundary samples in Hilbert space can be decided by SVM, and the misclassified minority class samples in these boundary samples are resampled to adjust the weight value. So we can increase the weight of features that can improve the accuracy of minority class samples. In each round of the process of feature selection, we eliminate the feature that has lowest score, so we can gradually improve the classification effect. The experiments have shown that the performance of our improved SVM-BRFE algorithm is better than the original SVM-RFE algorithm for high dimensional imbalanced data.

(2) In the part of data sampling, we improve the sample process in order to make the training space of SVM and input space to be consistency. In SVM training space, namely Hilbert space, firstly we oversample minority class samples by SMOTE, and then find the approximate pre-image of the synthetic samples in Euclidean space. PSO algorithm is adapted to optimize over sampling rate. Based on these improvements, the PBKS algorithm based on SMOTE is proposed, which is based on PSO. The effectiveness of PBKS algorithm for data sampling is verified by our experiments.

Finally, the two algorithms are combined, and our experiments have shown that the proposed BRFE-PBKS-SVM algorithm is effective for the classification of high dimensional imbalanced data.

##### References

1. Singh A, Purohit A (2015) A Survey on Methods for Solving Data Imbalance Problem for Classification. pp 37-41
2. Provost F (2008) Machine Learning from Imbalanced Data Sets 101 (Extended Abstract). In: Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of. pp 435–439
3. Huang Y-M, Hung C-M, Jiau HC (2006) Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis Real World Applications 7:720–747
4. Bhlmann P, Sara Van De Geer (2013) Statistics for High-Dimensional Data: Methods, Theory and Applications. Journal of the Japan Statistical Society Japanese Issue 44:247–249
5. Guo B, Damper RI, Gunn SR, Nelson JDB (2008) A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. Pattern Recognition 41:1653–1662.
6. Yu L, Liu H (2003) Efficiently handling feature redundancy in high-dimensional data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Dc, Usa, August. pp 685–690.
7. Shen D, Shen H, Marron JS (2013) Consistency of sparse PCA in High Dimension, Low Sample Size contexts. Journal of Multivariate Analysis 115:317–333.
8. Zhuang X-S, Dai D-Q (2007) Improved discriminate analysis for high-dimensional data and its application to face recognition. Pattern Recognition 40:1570–1578.
9. Arif M (2012) Similarity-dissimilarity plot for visualization of high dimensional data in biomedical pattern classification. Journal of Medical Systems 36:1173–1181.
10. Imani M, Ghassemian H (2016) Binary coding based feature extraction in remote sensing high dimensional data. Elsevier Science Inc.
11. Singh B, Kushwaha N, Vyas O-P (2014) A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty. Journal of Data Analysis & Information Processing 2:95–105
12. Eiamkanitchat N, Theera-Umpon N, Auephanwiriyakul S (2015) On Feature Selection and Rule Extraction for High Dimensional Data: A Case of Diffuse Large B-Cell Lymphomas Microarrays Classification. Mathematical Problems in Engineering,2015,(2015-11-16) 2015:1–12
13. García V, Sánchez JS, Mollineda RA (2011) Classification of High Dimensional and Imbalanced Hyperspectral Imagery Data. Springer Berlin Heidelberg
14. Yu H, Ni J (2014) An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. IEEE/ACM Transactions on Computational Biology & Bioinformatics 11:657
15. Dittman DJ, Khoshgoftaar TM, Napolitano A (2014) Selecting the Appropriate Data Sampling Approach for Imbalanced and High-Dimensional Bioinformatics Datasets. In: IEEE International Conference on Bioinformatics and Bioengineering. pp 304–310
16. Blagus R, Lusa L (2012) Evaluation of SMOTE for High-Dimensional Class-Imbalanced Microarray Data. In: International Conference on Machine Learning and Applications. pp 89–94
17. Maldonado S, Weber R, Famili F (2014) Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. Information Sciences 286:228–246
18. Tibshirani R (2011) Regression shrinkage selection via the LASSO. 73:273–282
19. Zhang C, Jia P (2014) DBBoost-Enhancing Imbalanced Classification by a Novel Ensemble Based Technique. In: International Conference on Medical Biometrics. pp 210–215
20. Blagus R, Lusa L (2013) SMOTE for high-dimensional class-imbalanced data. Bmc Bioinformatics 14:106
21. Wu C, Wang X, Bai D, Zhang H (2009) Fast Incremental Learning Algorithm of SVM on KKT Conditions. In: International Conference on Fuzzy Systems and Knowledge Discovery. pp 551–554
22. Vapnik V, Levin E, Le Cun Y (1994) Measuring the VC-dimension of a learning machine. Neural computation 6:851–876
23. Wang J, Yun B, Huang P, Liu YA (2013) Applying Threshold SMOTE Algoritwith Attribute Bagging to Imbalanced Datasets. Rough Sets and Knowledge Technology. pp221-228
24. Han H, Wang W-Y, Mao B-H (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Lecture Notes in Computer Science 3644:878–887
25. Hu S, Liang Y, Ma L, He Y (2009) MSMOTE: Improving Classification Performance When Training Data is Imbalanced. In: Second International Workshop on Computer Science and Engineering. pp 13–17
26. Cortes C, Vapnik V (1995) Support-Vector Networks. Kluwer Academic Publishers
27. Leung Y-Y, Chang C-Q, Hung Y-S, Fung PCW (2006) Gene selection for b cancer classification using support vector machines
28. Lu J, Zhang C, Shi F (2016) A Classification Method of Imbalanced Data Base on PSO Algorithm. In: International Conference of Young Computer Scientists, Engineers and Educators. pp 121–134
29. Kwok JT, Tsang IW (2004) The pre-image problem in kernel methods. IEEE Transactions on Neural Networks 15:1517–25
30. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27:861–874
31. Chang C-C, Lin C-J (2011) Libsvm. 2:1-27