Computer Science and Information Systems 14(3):579–595 https://doi.org/10.2298/CSIS161221017L

Imbalanced Data Classification Based on Hybrid Re-sampling and Twin Support Vector Machine

Lu Cao[1,3] and Hong Shen[1,2]

[1] School of data science and computer science,
Sun Yat-sen University, Guangzhou, China
caolu20001742@163.com
hongsh01@gmail.com
[2] School of Computer Science,
University of Adelaide, Australia
[3] School of Information Engineering,
Wuyi University, Jiangmen, China

Abstract. Imbalanced datasets exist widely in real life. The identification of the minority class in imbalanced datasets tends to be the focus of classification. As a variant of enhanced support vector machine (SVM), the twin support vector machine (TWSVM) provides an effective technique for data classification. TWSVM is based on a relative balance in the training sample dataset and distribution to improve the classification accuracy of the whole dataset, however, it is not effective in dealing with imbalanced data classification problems. In this paper, we propose to combine a re-sampling technique, which utilizes over-sampling and under-sampling to balance the training data, with TWSVM to deal with imbalanced data classification. Experimental results show that our proposed approach outperforms other state-of-art methods.

Keywords: over-sampling, under-sampling, imbalanced dataset, TWSVM, classification.

1. Introduction

Support vector machine (SVM) proposed by V. Vapnik et al. in 1960s is a machine learning technique based on statistical theory. SVM has excellent learning performance in the case of small samples and has been widely used in many fields such as pattern recognition, text classification and regression analysis[1-2]. SVM has a solid theoretical foundation, which is mainly embodied in three aspects: the maximum interval principle, the duality theory and the introduction of kernel function. The theory of maximum interval transforms the original problem of support vector machine to the solution of a convex quadratic programming problem. The kernel function is introduced according to the dual theory, which is used to solve the nonlinear problem. However, the high cost for training data required by SVM makes SVM not applicable for classification tasks on large datasets. A new learning method known as twin support vector machine (TWSVM), which extends a pair of parallel hyper-planes in SVM to the complex non-parallel-plane, was proposed in [3].

Compared with the traditional SVM, TWSVM has two important properties: (1) TWSVM can overcome some of the traditional SVM difficulties in dealing with data distribution, such as cross data. (2) TWSVM solves the quadratic programming problem in the quarter size of the original SVM and the constraint condition of the two programming problem does not contain all the sample points, which makes the training speed of TWSVM is remarkly less than that of traditional SVM. After TWSVM was proposed, researchers have paid close attention to how to further improve the TWSVM, thus a lot of methods have emerged in [4-7]. Although TWSVM has many advantages, it has drawbacks in dealing with imbalanced datasets directly. Imbalanced datasets exist widely in real life, such as cancer diagnosis [8], fraud detection [9] and insurance risk management [10]. The number of instances is much larger than that of the other samples, known as majority and minority class respectively. The recognition of the minority class in imbalanced datasets is greatly important to detect. Such as in the intrusion detection, the number of intrusion events must be far less than the number of normal events, but if an intrusion behavior is judged as a normal event, it may suffer serious losses. TWSVM, as a learning machine designed to optimize the performance of the whole dataset like other traditional classifiers, has low performance for minority class.

In this paper, we propose to combine a hybrid re-sampling technique, which utilizes over-sampling and under-sampling to balance the training data, with TWSVM to improve the recognization rate of the minority class samples in imbalanced datasets. The paper contains three technical components: (1) We present a hybrid re-sampling method to balance the training data by inserting synthetic points into minority classes with the over-sampling technique SMOTE (synthetic minority over-sampling technique) and simultaneously deleting samples carrying little information or noise from majority classes with the under-sampling technique OSS (one side selection). (2) As a new application of TWSVM, we show how to combine the above re-sampling technique with TWSVM to solve the imbalanced datasets classification problem. (3) We conduct extensive experiments to show the effectiveness of the proposed method in comparison with other state-of-art methods in term of *F-measure* and *G-mean*.

The rest of this paper is organized as follows. Section 2 presents the existing imbalanced datasets classification methods. Section 3 describes TWSVM theory and Section 4 introduces our approach to solve the imbalanced datasets classification problem. Section 5 compares the performance of the proposed approach with the existing methods. Finally in section 6, we conclude this paper and indicate our future work.

2. Related Work

A lot of research works have been carried out in the domestic and foreign scholars on the problem of imbalanced classification [11-13]. At present, the existing class imbalance classification methods can be simply categorized into two groups: data level strategy and algorithm level strategy. The data level approaches balance the training dataset of the classifier by re-sampling techniques, while the algorithmic approaches aim

to bias the learning process to enlarge the minority class domination. The two approaches are independent of each other and can be combined.

The data level approach is to resample imbalanced datasets, including under-sampling and over-sampling. The idea of re-sampling is to increase or decrease samples of balance datasets, in order to reduce adverse effects brought by the imbalanced datasets for classifiers. The simplest re-sampling method is to increase or decrease samples randomly, but the effect is not ideal [14]. People are inclined to a heuristic method. Synthetic minority over-sampling technique (SMOTE) is the most common over-sampling method, which adds new synthetic samples to the minority class by randomly interpolating pairs of the closest neighbors in the minority class [15]. SMOTE algorithm generates samples regardless of the majority class and is inclined to increase the minority samples close to the borderline and over-fitting. Han et al. presented an improved strategy of SMOTE, called borderline-SMOTE [16] to solve the problem of over-fitting by generating the minority samples near the classification hyper-plane instead of all minority sample points. Whether the original SMOTE algorithm or its improved algorithm, the generated samples are not consistent with the underlying true distribution of minority class, which could inevitably introduce noise into the training sample set and distort the spatial distribution of data. In [17], Adaptive Synthetic Sampling (ADASYN) algorithm is proposed to overcome the limitation of SMOTE by generate synthetic samples for minority class according to the distribution situation. Gao et al. introduce a novel over-sampling approach, which bases on kernel density method of the minority class to get probability density function estimation to solve two-class imbalanced classification problems [18]. The samples produced by this method can meet the probability density of the minority samples, but this approach is limited by the specific classifier. Zhang et al. presented a RandomWalk Over-Sampling approach (RWO-Sampling) to balance different class samples by creating synthetic samples through randomly walking from the real data. This method keeps the minority data distribution unchanged, but it is stated by the central limit theorem and some conditions must be satisfied [19]. In [20], an over-sampling technique MDO (Mahalanobis Distance-based Over-sampling), which can reduce the risk of overlapping between different class regions, is presented to generate synthetic samples by preserving the covariance structure of the minority class instances according to the probability contours. Two probabilistic over-sampling methods, RACOG (Rapidly Converging Gibbs) and wRACOG (wrapper-based Rapidly Converging Gibbs) are proposed in [21]. Both of these two methods generate new minority samples by using the joint probability distribution of data attributes and Gibbs sampling. RACOG generate new samples based on Markov chain, while wRACOG selects the samples which are most likely to be misclassified in probability.

Under-sampling method reduces the data samples of majority class. Random under-sampling (RUS) is the non-heuristic approach to delete some of the majority samples randomly to rebalance the dataset [22]. This method is simple and easy to implement. Because of the reduction of samples, the under-random sampling technique can reduce the training time. However, the representative information samples are inclined to be lost in this method. Therefore, it is the focus of the future research to retain the samples with large information and eliminate the samples with less information. One Side Selection (OSS) [23] is a typical under-sampling strategy, which divides majority samples into four groups according to Tomek Links technology. And it deletes noise

samples and borderline samples to balance the data samples of minority class. At the same time, researchers begin to try to use clustering method to find the information samples. Yen and Lee [24] propose cluster-based under-sampling approaches, which firstly divide all the training samples into some clusters, then select the representative data as training data in the cluster to improve the classification accuracy for minority class. An adiversified sensitivity-based under-sampling approach is presented in [25] by clustering and sampling iteratively. Majority samples are clustered to obtain the distribution information and improve the diversity of sampling in this method, then a random sensitive strategy is used to select samples from each cluster, finally, a relatively balanced dataset is obtained by iteratively clustering and resampling. In [26], a one-sided dynamic under-sampling (ODU) technique which adopts all samples in the training process, and dynamically determines whether a majority sample should be used for the classifier learning is proposed to solve multi class imbalance problems. For each training sample, ODU algorithm calculates the probability that it may be selected. When the probability is greater than a random number, the sample is considered to be representative. Otherwise the sample is not used in the training process. Lin et al. [27] introduces a dynamic sampling method (DyS) for multilayer perceptrons to solve multi-class imbalance classification. This approach dynamically selects informative samples according to the probability estimated to train the multilayer perceptron. In general, the most important thing in under-sampling is how to select the sample points which are useful for classification.

In addition to data preprocessing techniques, algorithmic level methods are also very popular to handle the imbalanced classification problem. Algorithm level is mainly to improve and enhance the existing algorithm Cost-sensitive learning by setting different misclassification cost to the majority and minority datasets is an effective solution [28]. Castro et al. presents a new cost-sensitive algorithm to improve the discrimination ability of multi-layer perceptrons by learning the Levenberg-Marquadt's rule for class imbalanced problem [29]. With the development of ensemble learning technology, more and more researches introduce the ensemble learning technology to the classification of imbalanced data. People are trying to combine the re-sampling technology and integration technology to come out the imbalanced data classification problem [30-31]. The two algorithms EasyEnsemble and BalanceCascade proposed in [30] are the typical ensemble classification algorithm based on the Boosting and Bagging techniques for undersampling data processing. Chen et al. proposes a Ranked Minority Over-sampling in Boosting (RAMOBoost) algorithm, which adaptively ranks minority class samples at each learning iteration according to a sampling probability distribution [31]. This approach can adaptively shift the decision boundary toward majority and minority samples which are difficult to learn by using a hypothesis assessed procedure. [32] are very comprehensive to summarize the existing boosting and bagging algorithms for imbalanced datasets classification. In addition, Shao et al. firstly introduce an efficient weighted Lagrangian twin support vector machine (WLTSVM) by using different training points to overcome the bias phenomenon in imbalanced classification [33].

3. Comparison of SVM and TWSVM

Consider a binary classification problem in the n dimensional, training dataset $T=\{(x_1,y_1),\cdots,(x_m,y_m)\}$, where m represents the number of samples, x_i is a sample in the input space X, $y_i \in \{-1,1\}$ is the label in the output Y.

The basic idea of support vector machine is to find an optimal hyper-plane, which ensures thd maximizes the area of both sides of the hyper-plane. As for standard linear se accuracy of the classification anupport vector classification (SVC), the separating hyper-plane can be defined as:

$$f(x)=w^{\mathrm{T}}x+b=0 \tag{1}$$

By introducing the regularization term and the slack variable ξ, the optimization problem can be described as follows:

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2+C\sum_{n=1}^{N}\xi_i \\ &\text{s.t.} \quad y_i(w^{\mathrm{T}}x_i+b)\geq 1-\xi_i,\xi_i\geq 0 \end{aligned} \tag{2}$$

The problem of maximizing the interval is transformed into a convex quadratic programming problem by using the dual method in convex optimization. In order to cope with the non-linear problems, SVMs use nonlinear kernels to map low dimensional feature space to high dimensional space.

TWSVM constructs two non-parallel hyper-planes, each of which is close to a class of samples and is far from the other class. Two hyper-planes are denoted as:

$$\begin{aligned} f_+(x)&=w_+^{\mathrm{T}}x+b_+=0 \\ f_-(x)&=w_-^{\mathrm{T}}x+b_-=0 \end{aligned} \tag{3}$$

Two non-parallel hyper-planes can be obtained by solving two optimization problems, and the two optimization problems can be described as:

$$\begin{aligned} &\min_{w_+,b_+,\xi_-} \ \frac{1}{2}\|X_+w_++e_+b_+\|^2+c_1e_-^{\mathrm{T}}\xi_-, \\ &\text{s.t.} \quad -(X_-w_++e_-b_+)+\xi_-\geq e_-, \qquad \xi_-\geq 0 \end{aligned} \tag{4}$$

$$\begin{aligned} &\min_{w_-,b_-,\xi_+} \ \frac{1}{2}\|X_-w_-+e_-b_-\|^2+c_2e_+^{\mathrm{T}}\xi_+, \\ &\text{s.t.} \quad (X_+w_-+e_+b_-)+\xi_+\geq e_+, \qquad \xi_+\geq 0 \end{aligned} \tag{5}$$

where $c_1>0$, $c_2>0$, X_+ and X_- are two types of samples, e_+ and e_- are column vectors, ξ_+ and ξ_- are slack variables. The dual problems of the equation (4) and (5) can be expressed as:

$$\begin{aligned} &\max_{\alpha} \ e_-^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}\overline{X}_-\left(\overline{X}_+^{\mathrm{T}}\overline{X}_+\right)^{-1}\overline{X}_-^{\mathrm{T}}\alpha, \\ &\text{s.t.} \quad 0\leq\alpha\leq c_1e_-, \end{aligned} \tag{6}$$

$$\begin{aligned} &\max_{\gamma} \ e_+^{\mathrm{T}}\gamma-\frac{1}{2}\gamma\overline{X}_+\left(\overline{X}_-^{\mathrm{T}}\overline{X}_-\right)^{-1}\overline{X}_+^{\mathrm{T}}\gamma, \\ &\text{s.t.} \quad 0\leq\gamma\leq c_2e_+, \end{aligned} \tag{7}$$

where $\overline{X}_+=[X_+\ e_+]$, $\overline{X}_-=[X_-\ e_-]$, α and γ are Lagrange multipliers. The inverse of the matrix is solved in (6) and (7). In order to avoid the possible ill-conditioning matrix,

TWSVM introduces factor εI to make the matrix inverse solvable. By solving (6) and (7), the two hyper-planes can be obtained as:

$$\begin{aligned} z_+ &= -\left(\overline{X}_+^{\mathrm{T}}\overline{X}_+ + \varepsilon I\right)^{-1}\overline{X}_-^{\mathrm{T}}\alpha \\ z_- &= -\left(\overline{X}_-^{\mathrm{T}}\overline{X}_- + \varepsilon I\right)^{-1}\overline{X}_+^{\mathrm{T}}\gamma, \end{aligned} \tag{8}$$

where $z_k = \left[\begin{array}{cc} w_k^{\mathrm{T}} & b_k \end{array}\right]^{\mathrm{T}}, (k = +,-)$.

The determination of a new class of samples depends on the distance between the sample points and the two hyper-planes, which can be described as:

$$\text{Class } i = \arg \min_{k=+,-} \frac{\left|w_k^{\mathrm{T}} + b_k\right|}{\left\|w_k\right\|} \tag{9}$$

Like traditional support vector machines, TWSVM maps nonlinear interfaces in the original feature space to high dimensional space through the kernel function to obtain better classification results.

Figure 1 and Figure 2 are two dimensional non-cross data and cross data in the SVM and TWSVM classification effect diagram respectively. Among them, positive sample is represented by the symbol "+", while negative sample is represented by the symbol "。". Two dimensional non-cross data is shown in Figure 1. Figure 1 (a) is the classification effect chart of SVM. The classification hyper-plane can separate the two kinds of data and satisfy the maximal margin. Figure 1 (b) is the classification effect chart of TWSVM. Not the same as the traditional SVM, TWSVM eventually gets two non-parallel classification hyper-planes, in which the solid line is represented for positive class classification plane and the dashed line for negative class classification plane. Two dimensional cross data is shown in Figure 2. Figure 2 (a) is the classification effect chart of SVM. It can be seen that there is only one classification plane in linear SVM, which cannot separate the two classes of samples. In Figure 2 (b), TWSVM using two classification planes can efficiently identify two kinds of samples.

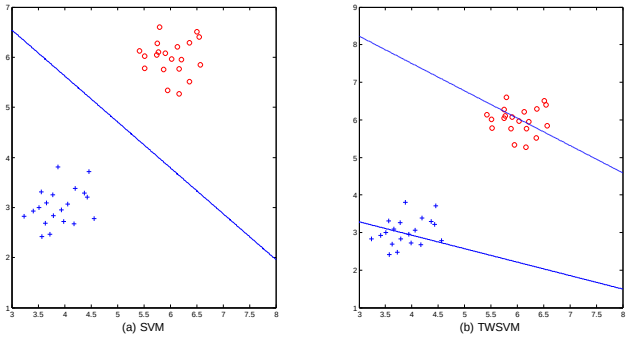

Fig. 1. Classification effect of non-crossing data in SVM and TWSVM. The symbol "+" and the circle "○" represent two kinds of sample points. The solid line in (a) is a support vector machine classification surface to satisfy the maximum interval theory. The dotted lines and solid lines in figure (b) are classification surface of two types of samples in TWSVM

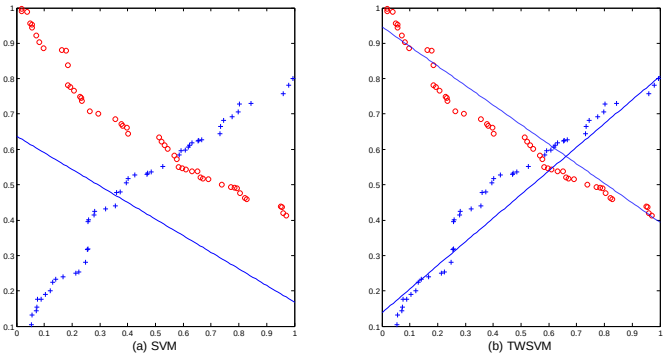

Fig. 2. Classification effect of crossing data in SVM and TWSVM. Figure (a) and (b) are classified as SVM and TWSVM respectively. For cross data, parallel hyper-plane theory of SVM cannot separate the two kinds of data, while non-parallel hyper-plane theory of TWSVM can separate the two kinds of data efficiently

4. The Proposed Approach

4.1. Re-balancing the Data

Over-sampling increases a few samples of the minority, which lead to expand the samples size, increase the training time and easily lead to over-fitting. Under-sampling deletes some samples of the majority. Although the training time is shortened, this method may remove some of the samples which are important for classification in the process of deleting the samples of the majority. In the highly imbalanced dataset, the removal of too many samples leads to serious loss of information, poor sample representation, and a serious departure from the initial data distribution. In this paper, we introduce a hybrid re-sampling technology to balance imbalanced datasets. On the one hand, SMOTE algorithm is used to synthesize new samples for minority class. On the other hand, we use the OSS algorithm to reduce the number of majority samples that have little impact on the classification.

SMOTE is an over-sampling method, the main idea of which is to insert the artificial data in a close distance between the minority samples to increase virtual samples for the minority class. The specific algorithm is as follows: as for every minority class sample of x_i, find k the nearest neighbors, then randomly select one of k neighbor as x_j, finally linear interpolate between x_i and x_j to construct a new minority sample. Figure 3 is an example of SMOTE for single sample. Three nearest neighbors of a given sample point is found. Only an artificial sample made by SMOTE is given in the graph.

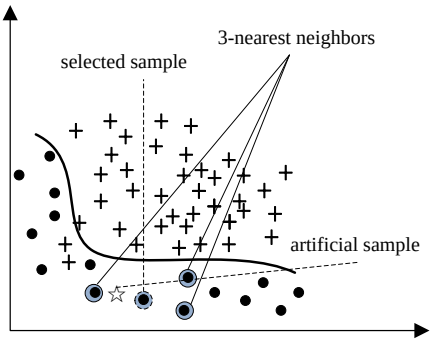

Fig. 3. Example of SMOTE for single sample, in which ● represents minority class, ＋represents majority class, ☆ represents the synthetic sample inserted. Three nearest neighbors of a given sample point is shown and the solid line represents the classification line

OSS algorithm divides majority samples into four groups: noise samples, borderline samples, redundant samples and safe samples. The noise samples are surrounded by the minority class; the borderline samples are close to the boundary; the redundant samples are which can be replaced by other majority class samples and are away from the boundary; the safety samples are which can provide valuable information for classification. Figure 4 shows four groups of samples divided by OSS algorithm specifically. OSS algorithm deletes noise samples and borderline samples to balance the data samples of minority class according to the concept of Tomek Links. Tomek Links algorithm description procedure is as follows. Given a pair of sample points with different sample labels (x_i, x_j) arbitrarily, (x_i, x_j) is called a Tomek Link if no sample x_k exists such that $d(x_i, x_k) < d(x_i, x_j)$ or $d(x_j, x_k) < d(x_i, x_j)$, where $d(x_i, x_j)$ is the euclidean distance between x_i and x_j.

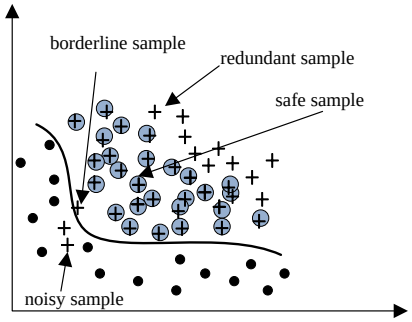

Fig. 4. Four groups of samples divided by OSS algorithm, in which ● represents minority class and ＋represents majority class, the solid line represents the classification line

Figure 5 (a) to (c) shows the effect of SMOTE and OSS. Figure 1 (a) is the distribution of original samples, figure 5 (b) and 5 (c) are distribution after SMOTE and OSS, respectively. From figure 5 we can find that SMOTE method maintains the basic distribution of the original sample, but the virtual sample increased mostly distributed in the original sample with less near the edge. We also can see that OSS method mainly keeps the sample points which are worth to classification.

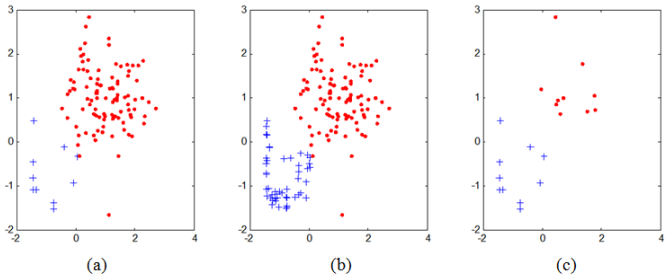

Fig. 5. Distribution graph of different sampling methods. (a) original dataset; (b) dataset distribution after SMOTE; (c) dataset distribution after OSS, in which ● and ＋represent two type of samples respectively

We introduce a hybrid re-sampling technology to balance imbalanced datasets. To be specific, we propose to apply Tomek links to the over-sampled training set as a data cleaning method to decrease over-fitting. Hybrid re-sampling approach can not only get a relatively balanced dataset, but also reduce the overlap between classes, which is beneficial to classification.

4.2. Combining Re-sampling with TWSVM

The scheme of hybrid re-sampling approach is shown as figure 6.

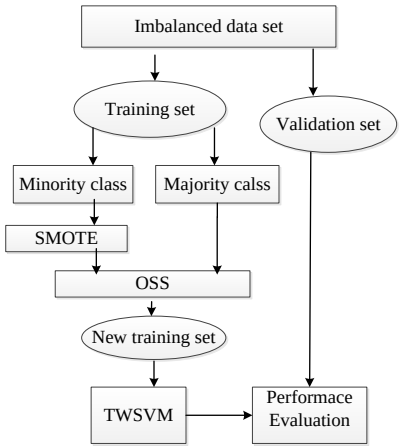

Fig. 6. Scheme of the proposed approach

The proposed approach can be summarized as follows. Firstly, the imbalanced dataset is divided into a training set and a validation set. Among them, the validation set used for cross validation accounts for 20% of the original dataset. Then, SMOTE algorithm is used to increase the minority samples and OSS algorithm is used to decrease the majority samples to get a new relatively balanced training set. Finally, a novel effective classifier TWSVM is used to train the new balanced training set and the validation set which is created initially is used to evaluate the performance of the classifier.

The algorithm for the implementation of our approach is shown in Figure 7.

Algorithm 1.

Input: Original set $\mathbf{S}=\{(\mathbf{x}_i, y_i\}$, $y_i \in \{-1,1\}$ is the label of sample , $\mathbf{x}_i \in \Re^n$, where $i \in [1,n]$

Output: p , is classification performance of TWSVM

1: $\mathbf{T} = 80\% \times \mathbf{S}$, $\mathbf{V} = \mathbf{S} - \mathbf{T}$; /* Randomly selected training set of 80% samples for training set, the rest for validation set*/

2: $\mathbf{M}$= majority($\mathbf{T}$) , $\mathbf{N}$=minority($\mathbf{T}$) /*Get the majority samples and the minority samples from $\mathbf{T}$ */

3:**for each** $\mathbf{x}_i$ in N

4: $\mathbf{x}_j$ = K- nearset- neigbors($\mathbf{x}_i$) /* Find k -th nearest neighbors of $\mathbf{x}_i$ */

5: $\mathbf{x}_{new} = \mathbf{x}_i + r \cdot (\mathbf{x}_j - \mathbf{x}_i)$ /* r is a random number from $[0,1]$ */;

6:end for

7: Noisy set $\mathbf{E} = \varnothing$

8: $\mathbf{T} = \mathbf{T} + \mathbf{x}_{new}$

9: **for each** pair $(\mathbf{x}_i, \mathbf{x}_j)$ in $\mathbf{T}$

10: **if** (class ($\mathbf{x}_i$) !=class($\mathbf{x}_j$)) and

($\notin \mathbf{x}_k$ | $d(\mathbf{x}_i, \mathbf{x}_k) < d(\mathbf{x}_i, \mathbf{x}_j)$ **or** $d(\mathbf{x}_j, \mathbf{x}_k) < d(\mathbf{x}_i, \mathbf{x}_j)$)

11: $E \leftarrow E \cup \{\mathbf{x}_i, \mathbf{x}_j\}$

12: **end if**

13: $\mathbf{T}_{\mathbf{new}} = \mathbf{T} - \mathbf{E}$

14: end for

15: $model$ = TWSVM($\mathbf{T}_{\mathbf{new}}$)

16: p = cross- validation($model$, $\mathbf{V}$)

17: return

Fig. 7. The algorithm for the implementation of our approach

5. Experiments and Analysis

In our experiments, all the classifiers are implemented in MATLAB 12.0. We employ LIBSVM to carry out SVMs. As for the parameters, we set $c_1 = c_2$ in TWSVM. The nearest neighbor number is 5. We focus on the comparison of our approach with SVM, SVM+OSS, SVM+SMOTE and TWSVM.

5.1. Datasets

In the following, we use eight datasets which have different degree of imbalanced from UCI datasets to verify the effectiveness of the proposed hybrid sampling method with TWSVM. UCI datasets can be obtained from http://archive.ics.uci.edu/ml/. In order to construct imbalanced datasets, we reconstruct the UCI datasets. With multiple classes of datasets, we merge some classes or just get two classes. For each dataset, the size of samples, attribution and imbalanced ratio are listed. The specific description about these datasets is summarized in Table 1. N_n and N_p denote the number of samples in the majority and the minority class respectively. Imbalance ratio is defined as N_n / N_p. From table 1, we can see that the datasets are very different in imbalance ratio.

Table 1. Datasets

Datasets	Samples (N_n / N_p)	Attributions	Imbalance ratio
Pima	768 (500/268)	8	1.87
Germen	1000 (700/300)	13	2.33
Haberman	306 (225/81)	3	2.78
Glass7	214 (185/29)	4	6.38
Satimage4	6435 (5809/626)	36	9.28
Vowel	990 (900/90)	9	10.0
Letter	200000(19266/734)	10	26.25
Yeast	1484(1440/44)	11	32.73

5.2. Evaluation Criteria

In general, it usually takes the classification accuracy as the evaluation criteria among the traditional classification methods. However, it is not reasonable to evaluate the performance of the classifier according to the classification accuracy as for the imbalanced data-sets. Because when the proportion of the minority is very low, even all the minority samples are divided into majority, the total accuracy is still very high. But this kind of classifier is not practical. For the issue of the imbalance datasets, there have already been new evaluation criteria such as $F-measure$ and $G-mean$, which are based on the confusion matrix. Confusion matrix is shown in Table 2.

Table 2. Confusion Matrix

	Predicted positive class	***Predicted negative class***
Actual positive class	TP (true positive)	FN (false negative)
Actual negative class	FP (false positive)	TN (true negative)

In this paper, we use $F-measure$ and $G-mean$ as the evaluation measure, defined as follows:

$$precision = TP \big/ (TP + FN) \tag{10}$$

$$recall = TP \big/ (TP + TN) \tag{11}$$

$$F-measure = \frac{(1+\beta^2) \times recall \times precision}{\beta^2 \times recall + precision} \tag{12}$$

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{13}$$

where β is as a parameter and is desirable to 1 in general. The greater the value of $F-measure$, the better classification performance of minority class samples. Because $G-mean$ is based on the accuracy of both classes, it can be used to measure the overall classification performance of the system. In our paper, we utilize *F-measure* and *G-mean* to evaluate the performance of our method by comparing our method with other methods.

5.3. Results and Discussions

In our experiments, we choose Gaussian kernel and use a grid search strategy. The experiment is repeated 10 times for each dataset. Finally, we take the average of 10 experiments for the experimental results.

Table 3 shows the training time classifiers on eight benchmark datasets. From table 3, we can see that OSS algorithm takes the least amount of time, which is consistent with the theoretical analysis. As a kind of under-sampling technique, OSS selects a portion samples from the majority class to balance dataset and decreases the training time. The most time consuming method is the SMOTE algorithm because SMOTE increases the number of minority class samples. For small size datasets such as Glass7 and Haberman, the computing time of SVM and TWSVM is comparable, while for large datasets such as Letter, TWSVM is faster than SVM. Our proposed algorithm is second only to TWSVM in time consuming.

Table 3. The training time of classifiers on benchmark datasets

Datasets	SVM	SVM+OSS	SVM+SMOTE	TWSVM	Our Approach
Pima	3.157	2,154	7.413	1.458	1.947
Germen	28.543	12.422	34.415	7.457	7.498
Haberman	0.457	0.211	0.654	0.138	0.105
Glass7	0.376	0.269	0.557	0.139	0.138
Satimage4	14.675	9.447	18.123	4.116	3.779
Vowel	0.659	0.557	1.214	0.325	0.221
Letter	107.129	97.63	126.258	25.698	30.781
Yeast	0.978	0.615	1.460	0.387	0.526

Experimental results are shown as follows. Figure 8 is performance in $F-measure$ for imbalanced datasets. Figure 9 is performance in $G-mean$ for different datasets. From the experimental results, we can find that the performance of TWSVM is better than SVM in general. For different datasets, the results of SMOTE and OSS are different. On the datasets Haberman and Satimage4, the classification performance of OSS is better than that of SMOTE algorithm. The effect of SMOTE algorithm is better than OSS for other datasets. Compared with TWSVM, SMOTE, or OSS, the hybrid sampling method with TWSVM classification algorithm in this paper is optimal on $F-measure$ and $G-mean$. Specifically, on dataset with low balance rate such as Pima and German, the method proposed in this paper has a different degree of improvement in $F-measure$ and $G-mean$ compared with other algorithms. On the highly imbalanced dataset such as Letter and Yeast, the hybrid re-sampling method with TWSVM also has a good performance. The improvement of $F-measure$ and $G-mean$ shows that this method can not only improve the overall classification performance of the imbalanced data, but also improve the classification performance of the minority class.

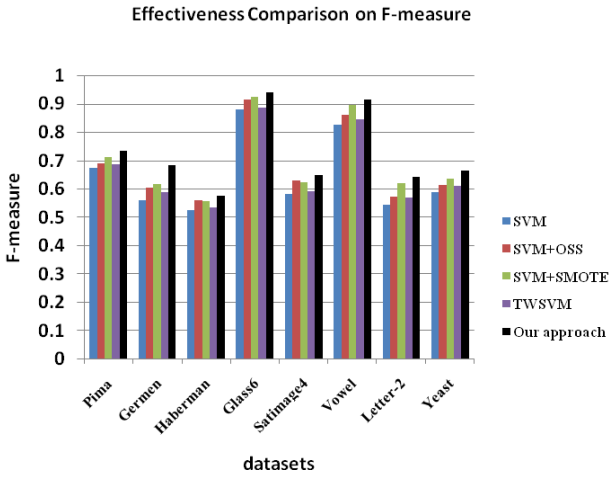

Fig. 8. Effectiveness Comparison on F-measure

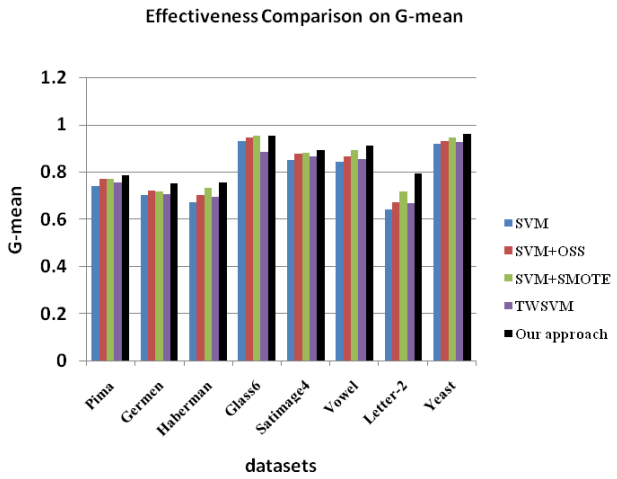

Fig. 9. Effectiveness Comparison on G-mean

6. Conclusion

There are a lot of imbalanced data in practical applications and traditional classification approaches have a low recognition rate for the minority class. An integrated sampling technique, which utilizes SMOTE algorithm and OSS algorithm to balance the training data, combined with the TWSVM classifier is proposed to deal with imbalanced data classification in this paper. As a popular classifier, TWSVM can deal with datasets which SVM is unable to handle, and its computational efficiency is much higher than SVM. Experimental results show that the method of hybrid re-sampling method with TWSVM is feasible and effective. In the experiments, we also find that SMOTE algorithm based on K nearest neighbors is limited to the range of positive samples, which will easily result in over-fitting in practical classification. Therefore, it is the next step to propose a new algorithm with good effect and fast convergence speed. At the same time, we discuss the two classification problem in this paper and multi-class imbalanced data classification is worthy of further study.

Acknowledgments. This work is supported by The 985 Project Funding of Sun Yat-sen University, Australian Research Council Discovery Projects Funding DP150104871, Youth Innovation Talent Project of Guangdong Province (No.2015KQNCX172), Science and Technology Project of Jiangmen City (No.2015[138], No.2016[189]) and Youth Foundation of Wuyi University (No.2015zk11).

References

1. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer, New York, NY, USA. (1995)
2. Deng, N.Y., Tian, Y. J., Zhang, C. H.: Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions. CRC Press. (2012)
3. Khemchandani, R., Chandra, S.: Twin support vector machines for pattern classification. IEEE Transactions on Pattern analysis and machine Intelligence, Vol. 29, No.5, 905-910. (2007)
4. Shao, Y. H., Zhang, C. H., Wang, X. B., Deng, N. Y.: Improvements on twin support vector machines. IEEE Transactions on Neural Networks, Vol.22, No.6,962-968. (2011)
5. Kumar, M. A., Gopal, M.: Least squares twin support vector machines for pattern classification.Expert Systems with Applications, Vol.36, No.4,7535-7543. (2009)
6. Peng, X.: TPMSVM: A novel twin parametric-margin support vector machine for pattern recognition. Pattern Recognition, Vol.44, No.10, 2678-2692. (2011)
7. Shao, Y. H., Chen, W. J., Deng, N.Y.: Nparallel hyperplane support vector machine for binary classification problems. Information Sciences, Vol. 263, No. 3,22–35.(2014)
8. Petrick, N., Chan, H. P., Sahiner, B., Wei, D.: An adaptive density-weightedcontrast enhancementfilter for mammographic breast mass detection. IEEETransactions on Medical Imaging,Vol. 15, No. 1, 59–67. (1996)
9. Fawcett, T., Provost, F.: Adaptive fraud detection. Data Mining Knowledge Discovery,Vol. 1, No. 3, 291–316. (1997)
10. Pednault, E. P. D., Rosen, B. K., Apte, C.: Handling Imbalanced Data Sets inInsurance Risk Modeling. IBM Research Report RC-21731.(2000)
11. Hulse, J. V., Khoshgoftaar, T. M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. The Twenty-fourth International Conference on Machine Learning, DBLP, 935–942. (2007)
12. Chawla, N. V., Japkowicz, N., Kotcz, A.: Editorial: Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations News letter, Vol. 6, No. 1, 1–6. (2004)
13. Wang, S., Yao, X.: MultiClass Imbalance Problems: Analysis and Potential Solutions. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man and Cybernetics Society, Vol. 42, No. 4, 1119-1130. (2012）
14. Wang, Q.: A Hybrid Sampling SVM Approach toImbalanced Data Classification. Abstract and Applied Analysis,Vol. 2014, No.5, 22–35. (2014)
15. Chawla, N. V., Bowyer, K.W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journalof Artificial Intelligence Research, Vol. 16, No. 1, 321–357. (2002)
16. Han, H., Wang, W. Y., Mao, B. H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science, Vol. 3644, Springer-Verlag, Berlin Heidelberg New York, 878–887. (2005)
17. He, H., Bai,Y., Garcia, E. A., Li, S.: Adasyn: Adaptive syntheticsampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks, 1322–1328. (2008)
18. Gao, M., Hong, X., Chen, S., Harris, C. J.: PDFOS: PDF estimation based over-sampling for imbalanced two-class problems. Neurocomputing, Vol. 138, No. 11,7535-7543. (2012)
19. Zhang. H., Li, M.: RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion, Vol. 20, No. 1, 99-116. (2014)
20. Abdi, L., Hashemi,S.: To combat multi-class imbalanced problems by means of over-sampling techniques.IEEE Transactions on Knowledge and Data Engineering, Vol.28, No. 1, 238-251. (2016)

21. Das, B., Krishnan, N. C., Cook, D. J.: RACOG and wRACOG: Two Probabilistic Oversampling Techniques. IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 1, 222-234. (2015)
22. He, H. B., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, 1263–1284. (2009)
23. Kubat, M., Matwin, S., Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the 14th International Conference on Machine Learning. Nashville, Tennessee, USA, 179–186. (2000)
24. Yen, S.J., Lee, Y. S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications An International Journal, Vol. 36, No. 3, 5718 – 5727. (2009)
25. Ng, W. W., Hu, J., Yeung, D. S., Roli, F.: Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems. IEEE Transactions on Cybernetics, Vol.45, No. 11, 2402-2412. (2014)
26. Fan, Q., Wang, Z., Gao, D.: One-sided Dynamic Undersampling No-Propagation Neural Networks for imbalance problem. Engineering Applications of Artificial Intelligence. Vol. 53(C), 62-73. (2016)
27. Lin, M., Tang, K., Yao, X.: Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification. IEEE Transactions on Neural Networks and Learning Systems, Vol. 24, No. 4, 647-660. (2013)
28. Zhou, Z. H., Liu, X. Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, Vol.18, No. 1, 63-77. (2006)
29. Castro, C. L., Braga, A. P.: Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, Vol. 24, No. 6,888-899. (2013)
30. Liu, X. Y., Wu, J., Zhou, Z. H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics, Vol. 39, No.2, 539-550. (2009)
31. Chen, S., He, H., Garcia, E. A.: RAMOBoost: Ranked Minority Oversampling in Boosting. IEEE Transactions on Neural Networks, Vol. 21, No.10, 1624-1642. (2010)
32. Galar, M., Fernández, A., Barrenechea, E., Bustince, H.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews, Vol. 42, No. 4,463-484. (2012)
33. Shao, Y. H., Chen, W. J., Zhang, J. J., Wang, Z., Deng, N. Y.: Anefficient weighted Lagrangian twin support vector machine forimbalanced data classification. Pattern Recognition, Vol. 47, No.9, 3158 – 3167. (2014)
34. University of California-Irvine, [Online]. Available: http://archive.ics.uci.edu/ml/(current April 2013)
35. Cao, L., Shen, H.: Combining Re-sampling with Twin Support Vector Machine for Imbalanced Data Classification. In Proceedings of 17th International Conference on Parallel and Distributed Computing, Applications and Technologies, Guangzhou, 325-329. (2016)
36. Tang, Y., Zhang. Y. Q., Chawla, N. V., et al.: SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics, Vol. 39, No.1, 281 – 288. (2009)
37. Liu, Y., Yu, X., Huang, J. X., et al.: Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing and Management, Vol. 47, No.4, 617-631. (2011)
38. Fu, J. H., Lee, S. L.: Certainty-based active learning for sampling imbalanced datasets. Neurocomputing, Vol. 119, No.16, 350-35. (2013)

39. Chawla, N. V., Lazarevic, A., Hall, L. O., et al.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Lecture Notes in Computer Science, Vol. 2838, Springer-Verlag, Berlin Heidelberg New York,107-119. (2003)
40. Cateni, S., Colla, V., Vannucci, M.: A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing, Vol. 135, No.8, 32-41. (2014)
41. Sun, Y., Kamel, M. S., Wong, A. K. C., et al.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, Vol. 40, No.12, 3358-3378. (2007)
42. Woniak, M., Grana, M,, Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion, Vol. 16, No.1, 3-17. (2014)
43. Akbani, R., Kwek, S., Japkowicz, N.: Applying Support Vector Machines to Imbalanced Datasets. Lecture Notes in Computer Science, Vol. 3201, Springer-Verlag, Berlin Heidelberg New York, 39-50.(2004)

This paper is a rewritten and extended version of an earlier conference paper [35] presented at the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2016)

Lu Cao received the B.S. degrees in Electrical Information Engineering from Changjiang University in 2004 and the M.S degree in Sun Yat-sen University in 2006. She joined Wuyi University since 2006. She is currently working towards her Ph.D. in Sun Yat-sen University since 2014. Her research interests include machine learning, pattern recognition and their applications in imbalanced problems.

Hong Shen is Professor (Chair) of Computer Science in University of Adelaide, Australia, and ”1000 People Plan” Professor and Director of Advanced Computing Institute in Sun Yat-Sen University, China. He received Ph.Lic. and Ph.D. degrees from Abo Akademi University, Finland, M.Eng. degree from University of Science and Technology of China, and B.Eng. degree from Beijing University of Science and Technology, all in Computer Science. He was Professor and Chair of the Computer Networks Laboratory in Japan Advanced Institute of Science and Technology (JAIST) during 2001-2006, and Professor (Chair) of Compute Science at Griffith University, Australia, where he taught 9 years since 1992. With main research interests in parallel and distributed computing, algorithms, data mining, privacy preserving computing and high performance networks, he has published more than 300 papers including over 100 papers in international journals such as a variety of IEEE and ACM transactions. Prof. Shen received many honors/awards including China National Endowed Expert of “1000 People Plan” (2010) and Chinese Academy of Sciences “Hundred Talents” (2005). He served on the editorial board of numerous journals and chaired several conference.

Received: December 21, 2016; Accepted: May 15, 2017.