IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2021, accepted March 30, 2021, date of publication April 20, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074243

Review of Classification Methods on Unbalanced Data Sets

LE WANG, MENG HAN, XIAOJUAN LI, NI ZHANG, AND HAODONG CHENG
School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China

Corresponding authors: Meng Han (2003051@nun.edu.cn) and Le Wang (1004535802@qq.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 62062004, in part by the Ningxia Natural Science Foundation Project under Grant 2020AAC03216, and in part by the Postgraduate Innovation Project of Northern Minzu University under Grant YCX20082.

ABSTRACT This paper studies the classification of unbalanced data sets. First, this kind of data sets is briefly introduced, and then the classification methods of unbalanced data sets are analyzed in detail from different perspectives such as data sampling method, algorithm level, feature level, cost-sensitive function, and deep learning. In addition, the data sampling methods are divided into different technologies for introduction: unbalanced data set classification method based on synthetic minority over-sampling technology (SMOTE), support vector machine (SVM) technology, and k-nearest neighbor (KNN) technology, etc. Then, the advantages and disadvantages of these methods are compared. Finally, the evaluation criteria of the unbalanced data set classifier are summarized, and the future work directions are prospected and summarized.

INDEX TERMS Unbalanced data sets, classification, sampling methods, algorithm level, feature level.

I. INTRODUCTION

Over time, the data tends to change its characteristics, since the number of learning instances in the considered class is not equal, this distribution causes some difficulties in classifying the data sets. The main characteristic of the unbalanced data set is class imbalances, which is caused by the phenomenon that some data streams that do not meet the conditions are ignored when dealing with real problems due to the small number of instances and low priority. Because instances of minority class are usually represented as positive instances, and majority class instances are represented as negative instances [1]. This is a very common challenge faced with the under-representation of minority class instances. Therefore, even if the overall classification model achieves high accuracy, the results of minority class may be poor [2]. When the minority class is particularly important, so it needs to be paid attention to.

Classification is a very useful method for data sets processing, but the data that can be collected are often dirty data without any regularity, it is necessary to do some processing on the data if we want to get meaningful data from a large number of data. Data are divided into many types due to their own characteristics, and they all have different feature attributes. Therefore, the algorithm should be improved for data sets with different features, or the accuracy of classification results will be affected. Among them, the characteristic of unbalanced data sets is that the instances that are concerned when mining data sets are often minority class, but the number of the class is particularly small. In real-life data, these unbalanced data sets are very common, such as solving fraud detection [3], medical diagnosis [4] and spam filtering [5] tasks. Among a large amount of mail data, the amount of spam is relatively small because most mail is normal mail. Therefore, class unbalanced learning always faces the challenge of insufficient instances.

Most of the existing review of classification methods on unbalanced data sets begin with data-level methods, it is mainly divided into the under-sampling method and over-sampling method. Literature [6] introduces unbalanced data classification method from these two aspects. There are also unbalanced data classification algorithms introduced from the perspective of ensemble learning. Literature [7] introduces the classification algorithm of unbalanced data from the ensemble classification algorithm. Literature [8] discusses the ensemble method in terms of a single positive class and multiple classes of data are summarized. However, the data sampling methods in the previous review of classification methods on unbalanced data sets and summarized according to the conventional method of under-sampling and

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

over-sampling, which could not highlight the main techniques used in the algorithm, and the review of the algorithm involved in each chapter was not comprehensive enough. In this paper, the classification algorithm of unbalanced data sets using the sampling method will be summarized according to the types of techniques used, at the end of this chapter, which is more clear than previous reviews.

At present, there are many advanced technologies for the classification of unbalanced data sets, ranging from the most basic under-sampling and over-sampling techniques to real-valued negative selective over-sampling. In the method based on deep learning, more and more researchers combine neural networks with ensemble models and achieve good results. The unsupervised learning method based on the generative adversarial network is also gradually used in the classification method of unbalanced data sets. Classification methods based on feature level and sensitive cost functions are also widely used to deal with problems related to unbalanced data sets, used to deal with problems such as class imbalance. The above related technologies will be introduced in detail in the following sections.

In this paper, several classification methods of unbalanced data sets are summarized in detail. Figure 1 shows the classification methods of several types of unbalanced data sets introduced in this paper. The main contributions of this paper are as follows:

(1) This paper summarizes and analyzes the classification methods for unbalanced data sets in detail from the aspects of data sampling, algorithm level, feature level and, deep learning methods. Compared with the existing review, it is more detailed, comprehensive, and includes updated related methods.

(2) In the sampling methods, this paper summarizes the classification methods for the unbalanced data sets from three aspects, synthetic minority over-sampling technique (SMOTE), support vector machine (SVM), and k-nearest neighbor (KNN) in this review than the previous.

(3) According to the two aspects of single-class learning and ensemble learning, the classification methods of algorithm-level unbalanced data sets are summarized, and the ensemble algorithm part is analyzed according to the Bagging and Boosting methods.

(4)Summarized the cost-sensitive classification methods of the unbalanced data sets, the classification methods at the feature level, deep learning, and so on. Finally summarized the evaluation criteria of the unbalanced data sets classifier, which made this article more comprehensive and complete.

The external method mentioned is the classification method based on the data sampling technique introduced in this chapter, and the internal method for creating or modifying the algorithm is also described in detail in the next chapter.

II. UNBALANCED DATA SETS CLASSIFICATION METHOD BASED ON SAMPLING METHOD

The characteristics of unbalanced data sets have caused the problem of low accuracy for minority class in classification,

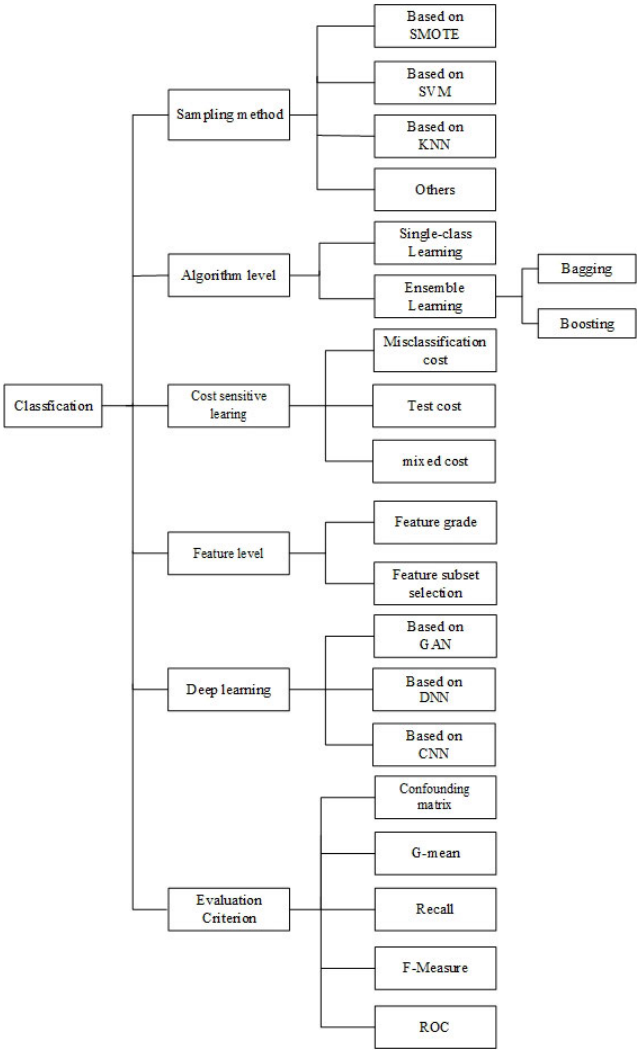


FIGURE 1. The main techniques for unbalanced data sets classification.

which has aroused the interest of many researchers. Researchers have proposed a large number of algorithms to solve the problem of class imbalance. These methods can be divided into two categories: external methods that use existing algorithms without modification, and internal methods that create new algorithms or modify existing algorithms to take into account class imbalances (Pazzani *et al.*, 1994; Riddle *et al.*, 1994; Japkowicz *et al.*, 1995; Kubat *et al.*, 1998.), the two types of methods can be roughly divided into data level and algorithm level. The external method mentioned is the classification method based on the data sampling technique introduced in this chapter, and the internal methods for creating or modifying the algorithm is also be introduced in detail in the next chapter. Data-level methods usually involve data preprocessing. In the process of preprocessing, the data is sampled to rebalance the highly unbalanced class distribution to adjust the class imbalance, further improve the accuracy of classification of unbalanced data sets. This chapter will mainly introduce sampling methods for unbalanced data sets the classification from synthetic minority

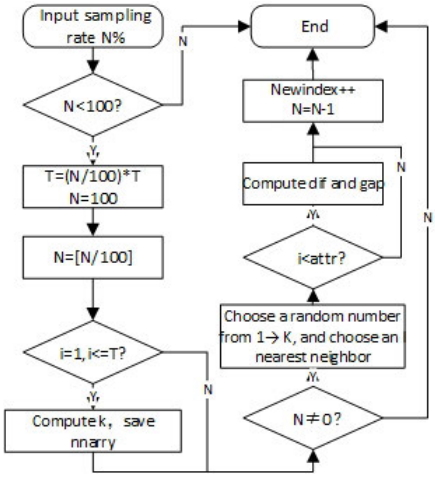


FIGURE 2. Flow chart of SMOTE algorithm.

over-sampling technique (SMOTE), support vector machine (SVM), and k-nearest neighbor (KNN).

A. SAMPLING METHOD BASED ON SMOTE

The sampling methods based on SMOTE is based on an oversampling algorithm proposed by Chawla *et al.* [9], which inherits the excellent parts of the algorithm and makes a series of improvements to it. The SMOTE algorithm increases the number of samples by interpolating between clustered minority samples. It oversamples the minority class by creating ''synthetic'' instances instead of oversampling by replacement. By operating in the ''function space'', a comprehensive instance is generated in a less application-specific way, instead of operating in the ''data space'' as before. By taking samples of each minority class and introducing comprehensive examples along the nearest neighbor line segment connecting any minority class, the minority class is over-sampled. According to the amount of oversampling required, samples from neighbors are randomly selected. Consider the difference between the feature vector (sample) and its nearest neighbor, multiply this difference by a random number between 0 and 1, and add it to the considered feature vector to generate a composite sample. This results in a random point chosen between two specific features along the line segment. This method effectively forces the decision-making area of minority class samples to become more common. The SMOTE algorithm is shown in Figure 2.

Because the SMOTE method currently cannot deal with all characteristics data sets, it has been extended to deal with mixed data sets of continuous and nominal characteristics. Chawla *et al.* [9] called this method synthetic minority oversampling technique-nominal continuous(SMOTE-NC). SMOTE-NC uses the median calculation to calculate the median of the standard deviation of all continuous features in the minority class. If the nominal feature between a sample and its potential nearest neighbor is different, the median value is included in the Euclidean distance calculation. They also used the median to penalize differences in nominal features that are related to typical differences in continuous feature values. And use continuous features to calculate the Euclidean distance space between the k nearest recognized feature vectors (minority class samples) and other feature vectors (minority class samples). For each different nominal feature nearest neighbor between the considered feature vector and its potential value, the method of filling samples is also used. Later they also proposed that SMOTE can also be extended to nominal features (SMOTE-N), using the modified version of the value difference metric (Stanfill and Waltz, 1986) proposed by Cost and Salzberg (1993) to calculate nearest neighbors. And use value difference metric (VDM) to observe the overlap of eigenvalues on all overlap.

Han *et al.* [10] improved SMOTE, called Borderline-SMOTE. The Borderline-SMOTE divides the instances into three zones, using the number of negative instances in the K nearest domains to determine the noise, bounds, and safety. These three regions are defined by specific numerical ranges. Borderline-SMOTE uses the same oversampling technique as SMOTE, but it only oversamples border instances of minority class instead of oversampling all instances of the class like SMOTE. This algorithm considers two positive examples. For the first and second instances, the values of n are k and k-1, respectively. These instances have no obvious difference, but they are divided into different areas, noises, and boundaries. So even if the first instance is rejected, the method will select the second instance for oversampling.

SMOTE is one of the techniques to remedy the problems caused by the classification of unbalanced data sets. It over-samples in the overlapping area to generate minority class instances to solve the problem. However, SMOTE will make minority class instances randomly along a line connecting the minority class instances and their chosen nearest neighbor, ignoring the majority class instances nearby. The technique of Safe-Level-SMOTE is to sample minority class instances with different weight degrees along the same line, use the minority instances in the nearest neighbor to calculate the safety level, make minority class instances near the greater safety level, so the precision property of the Borderline-SMOTE is better than SMOTE. But SMOTE itself randomizes minority class instances along the line connecting the minority class instances and their selected nearest neighbor and ignores the weakness of the majority class instances nearby, Therefore Bunkhumpornpat *et al.* [11] put forward Safe-Level SMOTE, which is based on the SMOTE method, Safe-Level-SMOTE sampling minority class instances with different weights along the same line, and assign each positive instance its security level before generating the synthetic instance. Each synthetic instance is located near the maximum security level, so all instances are generated only in the security area and are defined by the security level. If the instance's security level is close to 0, the instance is almost noisy. If it is close to K, the instance is considered safe. Finally, they give a definition of the safety level, which is used to select a safe location to generate synthetic instances.

Due to the extremely low occurrence rate of minority class, ordinary classifiers usually cannot detect minority class in unbalanced data sets. Bunkhumpornpat *et al.* [12] put forward a new oversampling technique DBSMOTE. This technique relies on the concept of gland-based clusters and is designed to oversample clusters of any shape discovered by DB-SCAN. DBSMOTE generates composite instances along the shortest path from each positive instance to minority class clustering pseudo center of mass. Thus, these synthetic instances are clustered near the center of mass and analyzed far away. The experimental results show that for the unbalance data sets, DBSMOTE better improves the accuracy, F-measure, and AUC than SMOTE, Borderline-SMOTE, and Safe-Level-SMOTE.

Because SMOTE can only solve two classes of problems by adjusting the generated rate to rebalance the class distribution, having more than one of the minority class leads to chaos. In addition, SMOTE is sensitive to the data complexity of the data sets, so in order to solve this phenomenon, Wang and Yao [13] put forward SMOTEBagging algorithm, which involves the generation step of the synthetic instances during subset construction. They extended the SMOTE into the Bagging model to address the multi-class data set. The SMOTE algorithm needs to determine two parameters: the oversampling number of k closest neighbors and minority group −N. The SMOTEBagging algorithm uses a percentage $b\%$ to control the number of instances in each class that is used to generate new instances for a subset. They used two-class data sets and multi-class data sets to conduct experiments on the method. The two-class data sets and multi-class data sets have similar effects on the diversity of each class, eventually proved that its overall performance and diversity have been improved.

When the classifier performs classification in an unbalanced data sets, it usually produces a biased classifier with higher prediction accuracy than the majority class, but worse than the minority class. For this reason, Chawla *et al.* [14] put forward the unbalance data sets learning method SMOTEBoost based on the combination of SMOTE algorithm and Boosting process. By using SMOTE to improve the prediction of minority classes, and using Boosting to improve the performance of the classifier without sacrificing the accuracy of the whole dataset. Boosting gives equal weight to all instances of misclassification. Since the Boosting algorithm mainly samples from a data pool composed of multi-class of data, Therefore, the subsequent sampling of the training set may still be biased towards majority class, Boosting, though, reduces the variance and bias in the final ensemble, but it may be less effective with data sets with the unbalanced class distribution. In order to reduce the inherent bias in the learning process caused by class imbalance, SMOTEBoost introduces SMOTE in each iteration Boosting so that each classifier can learn from more minority class instances, thus learning the broader minority class decision-making area. This method only looks for minority classes of examples in the distribution D_t at iteration t. This implies the effect of increasing the sample weights of minority classes of instances because new instances are created in D_t. The error estimation after Boosting iteration each time is in the original training set. Therefore, before learning a classifier in boosting iteration, they can maximize the boundary of distorted class data sets by adding new minority class samples. This method is used to reduce the bias generated by the classification of unbalanced data sets. The SMOTEBoost algorithm pseudocode is as follows:

TABLE 1. Comparison of classification algorithms for classical unbalanced data sets based on sampling technique.

Typical Algorithm	Compared algorithm	Character
SMOTE[9]	C4.5, Ripper, Navie Bayes	Applying SMOTE to create "synthetic" instances, the methods that could directly handle the class imbalance.
Borderline-SMOTE[10]	SMOTE	Borderline- SMOTE is better suited to two-class problems than SMOTE.
Safe-Level-SMOTE[11]	Borderline-SMOTE, SMOTE	Unlike Borderline-Smote, which has a high accuracy only when it uses Naive Bayes as the base classifier, C4.5 also has high accuracy.
Oversampling & under-sampling		Experiments show that oversampling and under-sampling strategy are not the best methods to use but combining them could be useful, especially if the bias employed by each strategy is of a different nature.

· Given:set $S\{(x_1, y_1),...,(x_m, y_m)\} x_i \epsilon$ X, with labels $y_i \epsilon$ Y $= \{1,..., C\}$,
Where C_P, ($C_P < C$) corresponds to a minority (positive) class.
· Let $B = \{(i, y): i = 1,..., m, y \neq y_i\}$
· Initialize the distribution D_1 over the examples, such that $D_1(i)=1/m$.
· For $t = 1,2,3,4,... T$
1. Modify distribution D_t by creating N synthetic example from minority
Class C_p using the SMOTE algorithm
2. Train a weak learn using distribution D_t
3. Compute weak hypothesis h_t: X × Y → [0, 1]
4. Compute the pseudo-loss of hypothesis h_t:
$\varepsilon_t = \sum_{(i,y)\epsilon B} D_t(i, y)(1 - h_t(x_i, y_i) + h_t(x_i, y))$
5. Set $\beta_t = \varepsilon_t/(1 - \varepsilon_t)$ and $w_t = \left(\frac{1}{2}\right) \cdot (1 - h_t(x_i, y) + h_t(x_i, y_i))$
6. Update D_t: $D_{t+1}(x, y) = \left(\frac{D_t(i,y)}{Z_t}\right) \cdot \beta_t^{w_t}$
where Z_t is a normalization constant chosen such that D_{t+1} is a distribution
· Output the final hypothesis:
$h_{fn} = \text{argmax}_{y\epsilon Y} \sum_{t=1}^{T} \left(log \frac{1}{\beta_t}\right) \cdot h_t(x, y)$

B. SAMPLING METHOD BASED ON SVM

Support vector machine (SVM) is a new two-class learning machine. The machine conceptually implements the following idea: the input vector is non-linearly mapped to a very high-dimensional feature space, in which a linear decision surface is constructed [15]. The special nature of the decision surface ensures the high generalization ability of the learning machine.

Zhang *et al.* [16] proposed an unbalanced data classification method based on the support vector machine. Firstly, the unbalanced data sets were pretreated with PCA whitening and label binarization. PCA set the variance of data feature transformation to 1 to remove the correlation between features and reduce the difficulty of model training. The formula of PCA whitening is,

$$X_{pca\_white,i} = \frac{x_{rot,i}}{\sqrt{\lambda_i + \varepsilon}} \tag{1}$$

After PCA whitening, the data sets were partitioned by leaving one in the cross-validation. The majority of class in the training set is randomly sampled in a certain proportion and combined with minority class in the training set to be relatively balanced data sets. The relative balance data obtained is input to the ensemble model with a support vector machine as the base classifier for weighted training. The ensemble output rule is the sum of the mean value of each group of base classifier decision values as the prediction result of the model. Finally, the test set is input to the ensemble model for model performance evaluation.

Cao and Hong [17] proposed an under-sampling algorithm CUS based on clustering according to the classification characteristics of support vector machines. CUS uses support vectors to play a decisive role in the support vector machine classifier and divides the algorithm, extracts the support vectors of each cluster as information samples, and retains the information of majority class samples to improve the algorithm for effective classification. This algorithm reduces the size of training samples by under-sampling, thus speeding up SVM prediction and making it more effective, and expand on large data sets.

Because the distribution of the data sets of positive and negative class samples has a large deviation, in order to minimize the bias, the data composition of the positive and negative class samples should be reconstructed. Therefore, Huang *et al.* [18] proposed a new over-sampling SVM algorithm NOBDF based on sample characteristics for data set reconstruction. The algorithm firstly to distance with the division of data set, and then take minority class in each distance samples based on the sample characteristics of an improved adaptive variable neighborhood SMOTE algorithm called ANBSC-SMOTE for the synthesis of new samples, and then reconstruct Pima-Indians the original data sets, in the final data sets using the traditional support vector machine(SVM) for classification. The algorithm achieves good results in both positive and negative classification accuracy and overall classification performance, and its robustness is improved, which provides an effective theoretical model for the classification of the unbalanced data sets.

In the classification of support vector machines, support vector plays a decisive role in the classification of hyperplanes. Because in random SMOTE algorithm, sampling of all the minority class samples will lead to a large amount of redundancy, which will further increase the training time, reduce the quality of the training sample, therefore, Zhang *et al.* [19] combine their improved random SMOTE oversampling algorithm and Bias-SVM classification method together, and only consider generating the sample that is ''close'' to the boundary, proposing an improved unbalanced data sets classification method. This algorithm processes unbalanced data from the data level and algorithm level, and can cluster minority class samples, and determine the support vector as the parent sample according to the distance between minority class clustering center and majority class clustering center. It can then generate new samples for the minority class. The improved algorithm has a significant classification effect on unbalanced data sets.

In order to solve the problem of class unbalanced intrusion detection learning algorithm, Ma *et al.* [20] put forward a new mixed method, FSVMs, using the current sampling method, SMOTE in combination with the fuzzy semi-supervised SVM learning method, to classify the unbalanced intrusion detection data. This method uses a support vector machine (SVMw) with weight as the base classifier, because it takes into account the weight of each sample, has better learning performance and high computational efficiency. This method uses a support vector machine (SVMw) with weight as the base classifier, because it takes into account the weight of each sample, has better learning performance and high computational efficiency. On the basis of the segmentation strategy, in order to further solve the problem of multiple class imbalance, the method to make use of unknown label data extension, at the beginning of the training to use mix-SMOTE to deal with the original training samples, using fuzzy partition strategy was carried out on the training data segmentation, based on the hybrid ratio SMOTE on training data sampling, this method can realize high efficient sampling, but also can avoid overfitting phenomenon. Secondly, the classifier SVMw was trained to calculate the fuzziness of unlabeled data and then divided into three groups: low, medium, and high. Finally, the low fuzziness and high fuzziness data are combined with the original marker samples to retrain the classifier model. Finally, experiments show that this method has high precision for unbalanced data sets.

C. SAMPLING METHOD BASED ON KNN

Because of the characteristics of unbalanced data sets, it is difficult to get ideal classification results when dealing with them. According to the advantages of the KNN algorithm and the characteristics of unbalanced data sets, many researchers have proposed some sampling methods to improve the accuracy of classifiers.

Kang *et al.* [21] proposed a new under-sampling method in which a noise filter is added before re-sampling. The existing noise filter is always used in combination with oversampling technology or only deals with noisy instances of most levels. In this paper, a representative under-sampling technology EE(EasyEnsenble) is combined to realize the Noise-filtered Under-sampling Scheme (NUS). Before re-sampling, a noise filtering framework (KF) based on the Clustering method of K nearest neighbor (KNN) is added to solve the imbalance problem. X-KF (UA-KF, EE-KF, etc.) is distinct from other under-sampling methods in that it does not use all minority class examples. Before training the classifier, X-KF first filters out noisy instances from the original minority class data sets, and then trains the classifier using the new minority class data sets. The main difference between X-KF and X is the minority class data sets used to train the classifier, which is also the first time that noise filtering and under-sampling methods have been combined. The pseudocode of the KF algorithm is shown as follows.

1. **Input**: Minority class dataset S_m, Majority class dataset S_M, $|S_m| < |S_M|$, and K that is the number of neighbors we need to calculate.
2. **For** $j = 1$: $|S_m|$
   1) Calculate the K nearest neighbors of sample j in S_m from dataset $(S_m - S_{m,j}) \cup S_M$ through the KNN algorithm.
   2) Count the number of majority samples among K nearest neigh- bors.
   3) If all K nearest neighbors of samples j in S_m are majority samples,
   we can consider j as the noisy sample and mark its label as 2 instead of 1.
3. **End for**
4. Delete all the noisy samples marked as 2.
5. **Output**: A new minority class dataset S'_m.

Li and Hu [22] also proposed a dense-based method combining KNN classifier to reduce the noise of the training data sets. It not only removes the noise samples, but also gives the rules for removing the high-density redundant samples thus reducing the computation of the classifier. When a new sample arrives, KNN finds the k neighborhoods closest to the new sample from the training set according to some appropriate similarity, where the similarity is measured by the cosine between two document vectors. Then, according to the distribution of each sample in the training data, the noise in the training data was removed to make the distribution of the training data more suitable for the K classifier. At the same time, the running time of the KNN method is reduced and the classification accuracy is improved.

Nekooeimehr and Laiyuen [23] developed an under-sampling strategy to deal with noise and class imbalance problems in order to retain useful instances and eliminate noisy ones. First of all, the examples of each type are divided into three categories: useful examples, noise, and potentially useful examples. Then a similarity coefficient is introduced to distinguish the instances of each class. A selection mechanism based on similarity coefficient keeps the useful samples and eliminates the noise samples. They propose a dense-based under-sampling method (DBU), which also uses the KNN clustering method, density to select useful instances and potential instances, uses and introduces a similarity coefficient to distinguish instances in each class, which makes the similarity coefficient of noise and redundant instances lower than that of other instances. The similarity coefficient of these instances is equal to 0. For majority classes, a certain number of samples are selected according to the similarity coefficient of samples as elements of re-sampling of most classes. For the minority classes, the noise is eliminated by deleting the instance with a similarity coefficient of 0 to achieve the purpose of sampling and adjusting the unbalanced phenomenon of data.

D. OTHER SAMPLING METHODS

In addition to the synthetic minority over-sampling technique (SMOTE), the sampling method based on support vector machine (SVM) and the sampling method based on clustering, the data level classification method of unbalanced data sets also involves many under-sampling methods, over-sampling methods, and mixed sampling methods. The simplest under-sampling is the random under-sampling method (RUS) [24], which randomly samples majority class samples, but with minority class samples than normal before, it is a form of data sampling where majority class instances are randomly selected and removed from the data sets until the desired class distribution is achieved [25]. But random under-sampling can lose some useful information, so some researchers have proposed other algorithms to solve the problem, such as Galar *et al.* [26] proposed EUSBoost, Gong *et al.* [27] proposed RHSBoost, Lin *et al.* [28] proposed CBUBoost, etc. In the EUSBoost algorithm, the imbalance ratio is defined as the number of negative class instances divided by the number of positive class instances. The imbalance rate is used to judge whether the data is unbalanced. If the imbalance rate is greater than 9, the data set will be called highly unbalanced [29]. To get a useful subset of the original data sets, the EUSBoost method randomly samples several data subsets and iterates over them until it is unable to further improve the current best re-sampled data sets.

Tao *et al.* [30] proposed a new sampling technology, they use a real-valued negative selection (RNS) to generate the minority class instance to adjust the unbalanced data sets, then the combination of the generated minority class instances and the majority class instances is used as the input of the traditional supervised classification algorithm to determine the optimal decision function. Cui *et al.* [31] proposed an adaptive undersampling method based on density peak clustering. The influence of overlapping region on

classification is considered. According to the density of samples in the subcluster, the weight of samples is calculated and the samples are undersampled. The nearest neighbor search algorithm is used to identify majority class samples in the overlapping regions and delete them. Tao *et al.* [32] proposed a novel adaptive weighted over-sampling for unbalanced data sets based on density peaks clustering with heuristic filtering. This approach can simultaneously accommodate both between-class and within-class imbalances caused by various reasons.

In addition to the simple under-sampling and over-sampling methods, the mixed sampling method is also a feasible data-level method. The mixed sampling method combines the advantages of under-sampling and over-sampling methods and produces better results in the classification of unbalanced data sets. Estabrooks *et al.* [33] conducted a large number of experiments to demonstrate that the over-sampling and under-sampling of C4.5 sampled at different rates produced a new effect, They make different decisions about the various classifiers that make up the combination to make the combination approach work best, because previous research has shown that under-sampling and oversampling will produce classifiers that can make different decisions, based on this decision, instead of having different classifiers vote on a given test point to average the results, a ''good enough'' classifier can decide on that point, and the individual data points need not be the same as the data points selected for the different data points. Their approach allows a single but different classifier to make decisions about each point. Because a single but different classifier makes a decision on each point, the result of the data point is also likely to be unreliable, In order to avoid this possibility, and on this basis, an elimination program is designed to prevent any inappropriate classifier existing at the architectural classification level from participating in the decision process. The elimination program relies on the application of the C4.5 results of ten-fold cross-validation of the original unbalanced training data. The various classifiers of this combination scheme finally show a low error rate, and the overall effect is considerable.

Seiffert *et al.* [34] proposed a hybrid data sampling method combining random over-sampling and random under-sampling techniques, and created a balanced data set for the construction of a decision tree classification model. In order to further improve the classification performance, this method uses two sampling techniques on the same unbalanced training data. One sampling technique is used to partially sample the training data, and the other method is used to complete the whole process. During the experiment, they used five different sampling techniques: random under-sampling (RUS), SMOTE (SM), Random Oversampling (ROS), borderline-SMOTE (BSM), and Wilson's Editing (WE). The experimental results show that the hybrid technique is superior to the single sampling technique in almost all cases. Ng and Dash [35] also use similar techniques to train classifiers on unbalanced data sets. They repeatedly sampled the minority class and the majority class without substitution until no significant improvement to the previous classifier was found. When the supply of minority class samples is exhausted, they will continue to sample from the majority class samples without replacing them but will use random oversampling to keep the distribution of the class balanced, which is essentially a combination of random oversampling and under-sampling.

Estabrooks and Japkowicz [36] combined an under-sampled data-based classifier with an over-sampled data-based classifier. They proposed that the C5.0 classifier performs over-sampled and under-sampled data at ten different rates according to the classifier level. The expert level consists of two experts, one combining the results of an under-sampled classifier and the other combining the results of an over-sampled classifier. Finally, the output level combines the results of the under-sampled expert and the over-sampled expert. In the whole classification process, as the classifier of the combined scheme has a natural tendency to classify the instances as negative, it may be unreliable to assume that a classifier has too many positive decisions. Therefore, they set several combination schemes: (1) A combination plan applicable to each expert at the expert level. (2)The combination scheme is applied at the output level; (3)The elimination scheme is applied to the classifier level. In this way, the classifier is biased to the positive set and the negative bias of the classifier is offset by the higher proportion of positive samples.

When the classical random forest model is faced with unbalanced data sets, Zheng *et al.* [37] proposed an improved random forest unbalanced data classification algorithm based on the mixed sampling strategy, because the sample distribution of each subtree is still consistent with the sample distribution of the original data set and the diversity of base classifiers cannot be guaranteed. By adopting the mixed sampling strategy for each subtree in the random forest, different balanced training subsets are constructed, so as to improve the diversity of base classifiers and enhance the effect of classifiers.

In addition to the above mentioned methods, many researchers have also proposed, with the method of random forests and sampling strategy to deal with the unbalanced data sets, such as Yang *et al.* [38] proposed new random forest algorithm an improved classifying unbalanced data set, the algorithm combines the classification of the decision tree classification effect and the random forest algorithm of weighted voting principle to deal with data. In the process of generating a decision tree, a mixed sampling method is used to generate training subsets, and then the final classification results are selected by weighted voting. In this way, It avoids using the same voting weight on the decision tree to make full use of the high precision decision tree, and the random forest algorithm can have higher accuracy. The algorithm is shown in Figure 3.

TABLE 2. Unbalanced data sets classification method based on sampling method.

Algorithm	Technique	Sampling method	Dataset	Compared algorithm	Advantages	Disadvantages
SMOTE[9]	SMOTE	oversampling	Pima, Phoneme,Adult, E-state, Forest Cover;	Under–C4.5 , Ripper,Naive Bayes, Hull;	Oversampling minority class of samples by creating a "synthetic" instance instead of oversampling by substitution.	Ignore most of the nearby class instances. Sensitive to data complexity of data set.
SMOTE-NC[9]	SMOTE	oversampling	Pima, Phoneme,Adult, E-state, Forest Cover;	Under–C4.5, Ripper,Naive Bayes, Hull;	Use the median to punish differences in nominal features associated with typical differences in continuous eigenvalues.	Sampling on some data sets may be poor.
Borderline-SMOTE[10]	SMOTE	oversampling	Circle, Pima, Satimage, Haberman;	borderline-SMOTE1, borderline-SMOTE2, Random over-sampling;	Only minority class boundary instances are over-sampled to divide different regions, noises, and boundaries, which reduces the operation time.	No different policies are customized for boundary instances.
Safe-Level-SMOTE[11]	SMOTE	oversampling	Satimage, Haberman;	SMOTE, Borderline-SMOTE;	Configuring the level of safety for each instance, this approach gives better precision than SMOTE and the Borderline-SMOTE.	It has not been resolved to automatically determine the number of synthetic instances generated.
DBSMOTE[12]	SMOTE	oversampling	Pima, Haberman, Glass, Segmentatio, Satimage, Ecoli, Yeast;	ORG, SMOTE , BORD, SAFE, DBSMOTE;	DBSMOTE more effectively improves the accuracy, F-measure, and AUC than SMOTE, Borderline-SMOTE, and Safe-level-SMOTE.	Doesn't solve Eps, MinPts automatic measurement problem.
SMOTEBagging[13]	SMOTE	oversampling	Hepatitis, Heart,Liver, Pima,Ionosphere, Breast-w, Glass,Yeast;	Under-sampling, over-sampling, SMOTE.	Using SMOTE algorithm in the Bagging model and extending it to solve multi-class data set, its overall performance and diversity are improved.	Its diversity leads to worse recall rates for majority class, and the differences between two and multiple class cannot be analyzed yet.
SMOTEBoost[14]	SMOTE	oversampling	KDDCup-99 Intrusion, Mammography, Satimage, Phoneme;	First SMOTE, then Boost, AdaBoost.M2, Single RIPPER, SMOTE RIPPER.	In Boosting each iteration, introducing SMOTE will enable each classifier to learn from the minority class instances, thus learning the broader minority decision-making area.	Before learning a classifier in Boosting iteration each time, they can maximize the boundary of distorted class data set by adding new minority class instances.
CUS[16]	SVM	under-sampling	Pima,Haberman, Ecoli, Glass1, Glass16vs2, Shuttle2vs5;	SVM, ROS, SMOTE, RUS, OSS, SBC.	Under-sampling reduces the size of the training sample, thus speeding up SVM prediction and making it more effective and extensible on large data sets.	
NOBDF[17]	SVM	under-sampling	Pima,german, wpbc, Haberman, Yeast, abalone;	SVM, RU, Smote, BSmote, Weight, RUS, AdaSyn, SPU.	The accuracy, overall classification performance, and robustness of positive and negative classes are improved, which provides an effective theoretical model.	The distance band number is not properly divided, which will affect the sample synthesis and data set classification results.
SMOTE and biassed - SVM combination[18]	SVM/SMOTE	under-sampling	Banana, Appendicitis, Haberman, Vehicle, Wisconsin;	SVM, Random-SMOTE.	According to the distance between the cluster center of minority class and the center of majority class, the support vector is determined as the parent sample to generate minority class samples to improve the classification effect.	The optimization of parameters still needs to find the optimal solution, so as to improve the generalization ability of the classifier.

III. UNBALANCED DATA CLASSIFICATION METHOD BASED ON ALGORITHM LEVEL

The data preprocessing method is considered to be the easiest way to deal with unbalanced data classification, but many data sets have some obvious disadvantages in the process of preprocessing. The first is that the distribution of most data sets is often unknown. The second point is that the efficiency of the re-sampling strategy is very low, which will lead to the risk of losing the basic information of most classes when sampling is insufficient, and the risk of over-fitting minority

TABLE 2. *(Continued)* **Unbalanced data sets classification method based on sampling method.**

FSVMs[19]	SVM	oversampling	Balance, Car, Satimage;	Bayes, BP, DT, KNN, RF, SVM, FSVMs.	The fuzzy divide-and-conquer strategy is used to segment the training data. This method can realize efficient sampling and avoid over-fitting.	The detection rate of majority class is lower than that of other classifiers. The effect of this method is limited by the unbalance rate of the data set, and the whole method still has some limitations.
NUS[20]	clustering	under-sampling	Pima,haberman, yeast, balance, letter, cmc, wilt, poker;	UA, UA-KF, RUSB, RUSB-KF, UB, UB-KF, EE, EE-KF.	It is the first time to combine noise filtering and under-sampling, and it is also the first time to discuss the noise problem in minority class.	
Based on density combined with KNN classifier[21]	clustering	under-sampling	a subset of news in People Daily from 1993 to 1997, a collection of news downloaded from the web site sina;		Noise samples are removed, rules for removing high-density redundant samples are given, and the computation of the classifier is reduced.	
DBU[22]	clustering	under-sampling	pimaimb, glass016vs2, glass2, ecoli4, car-good, yeast1289vs7, wine-red3vs5, poker89vs5, abalone19;	RUSBoost, EUSBoost, CBUBoost, RHSBoost, DBSCAN-KNN, EE-KF, SMOTE-IPF, DBUBoost.	The similarity coefficient is used to distinguish each class of instances so that the similarity coefficient of noise and redundant instances is lower than that of other instances to adjust the imbalance phenomenon.	

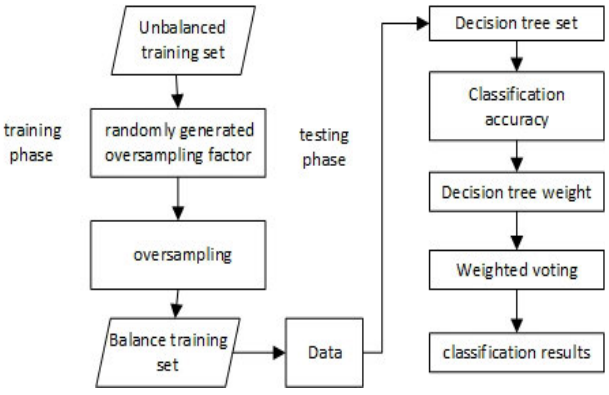


FIGURE 3. Flow chart of improved random forest algorithm.

class when sampling is too much. In view of the problems that will affect the classification accuracy in the pre-processing of unbalanced data sets, people begin to pay attention to the processing of unbalanced data sets at the algorithm level. Many algorithms for processing unbalanced data sets are proposed, and the main methods include single-class learning algorithms and ensemble learning algorithms.

A. SINGLE-CLASS LEARNING ALGORITHM

Single-class learning is to train the data containing only one class of samples. It is a method to learn by using a small number of labeled samples and a large number of unlabeled samples. It is also an effective method to deal with unbalanced data sets. Because most of the raw data we can get from our lives are multi-class data, minority class samples are rare but often critical. Although there is only one class in the sample of single-class learning, it is to find the majority class samples in this sample. When the new samples are classified, the properties of the new samples are judged according to the similarity measure and correlation value.

Yin *et al.* [39] proposed a fault detection method based on robust one-class support vector machines. By changing the weight of samples, the robustness of the algorithm is improved, and the interference of single-class support vector machine to uncorrelated samples is improved. Yin *et al.* designed an adaptive penalty factor according to the Euclidean distance from each normal data point to the data center, so that outliers had little influence on the decision boundary of the first kind of support vector machines. A fault detection method based on a robust single-class support vector machine (SVM) is proposed, in which the optimization model of a robust one-class support vector machine is.

$$\min_{w\in F,\xi\in R^N,\rho\in R} \frac{1}{2}||w|| + \Theta\sum_{i=1}^{N}\hat{d}_i\xi_i - \rho \tag{2}$$

In order to realize the fault detection of SVM, distance measurement and the corresponding threshold value are introduced.

Luca *et al.* [40] used a single-class support vector machine to detect the imbalance between normal data and abnormal data. This paper introduces the use of extreme value theory(EVT) to solve the classification problem of sets (Embrechts *et al.*, 1997). The Poisson point process (PPP) characterization of EVT is used to extract count data that

describes the number of times measurements in S fall into the low-density region defined by X. Compared with existing novelty detection methods, EVT can define a model for exception classes, where data is sparse or even invisible. This enables it to avoid the optimization of hyperparameters, which is usually encountered when using single-class classifiers and often requires data from exception classes. In essence, the use of EVT relies on the extrapolation of normal classes, which provides a class of models for low-density regions. The latter is particularly useful for novelty detection because decision boundaries are expected to be located in areas where data is sparse.

In order to describe the boundary between majority and minority class samples, Han *et al.* [41] used the area under the curve of the sample class robust to replace the empirical error in SVDD, and proposed an unbalanced support vector data algorithm capable of fault diagnosis. In single-class learning, it can be found that it is very important to select a correct threshold because the determination of sample class is affected by the threshold. In addition, the performance of the single-class learning method is also affected by the kernel function. Therefore, choosing an appropriate algorithm can affect the final classification result, and improve the final result by improving the convergence speed and generalization degree of the algorithm.

Scholkopf *et al.* [42] proposed One-Class support vector machines(OCSVM), OCSVM principle similar to SVM, the origin as abnormal points in feature space, and then find a hyperplane as decision boundary to be able to distinguish abnormal points and training samples, and then for each subsequent new samples, the decision boundary can be used to classify it. In the background of information retrieval, Manevitz *et al.* [43] used the OCSVM method to identify and classify unbalanced data, and improved the algorithm by putting forward out-Lier-SVM algorithm. They not only treated the origin as an outlier like the OCSVM but also processed the points very close to the remote point as an outlier. The experimental results show that out-lier-SVM may perform less well on some particular arrays than on a single-class SVM, but it may perform better on some data sets (for example, for a polynomial kernel with 10 features), and out-lier-SVM also produces better results for larger classes. Using macro averaging, that is, taking into account the number of different items in each class, the results of experiments are better than those of single-class SVM.

In the context of the loss of credit card customers, Maldonado *et al.* [44] proposed a method to classify unbalanced data sets through sets. Since this problem is considered to be highly unbalanced, they adopted different classification techniques. For example, support vector data description(SVDD) and two classes of support vector machines, as well as the characteristics of two classes of single-class learning: class imbalance and class overlap, the main idea is to solve the problem of class imbalance and class overlap by superposing different classification methods and consider single-class learning methods to evaluate the diversity of single classifiers. Considering the diversity of individual classifiers, an ensemble approach is used to construct more robust classifiers. In the method of constructing the classifier, the classification strategies of two classes and one class are different. Firstly, the method of binary classification using cost-sensitive SVM is considered. Second, adopt a method that provides a set of object descriptions to detect objects that are in some sense significantly different from the rest of the data set. For the latter approach, SVDD and a method called Parzen density estimation are used as alternatives [45]. The method is based on a mixture of Gaussian cores centered on a single training instance, usually considering diagonal covariance matrices and assuming the same weighted characteristics. Bandwidth is the only parameter that needs to be calibrated during training, usually by maximum likelihood. Finally, experiments with artificial data sets and complex customer churn prediction problems from a financial entity in Chile demonstrate that this method can provide more accurate and robust classification models for different balance and noise levels.

B. ENSEMBLE LEARNING ALGORITHM

More and more attention has been paid to ensemble learning algorithms in the classification of unbalanced data sets. They can also be divided into three subclasses: ensemble based on Bagging, Boosting, and Hybrid, it depends on the ensemble learning algorithm they are based on. Literature [46] makes an extensive empirical analysis on the problem of class imbalance, Boosting [47] and Bagging [48] both achieved the best effect by Boosting [47] and Bagging [48] combined with preprocessing technology. The following will introduce the ensemble classification algorithm of unbalanced data sets from both Bagging and Boosting.

1) UNBALANCED DATA ENSEMBLE CLASSIFICATION ALGORITHM BASED ON BAGGING

In the classic Bagging ensemble method, N classifier sets are constructed from different subsets of the original training data. The individual training data of N classifiers are formed by random sampling and replacement of the original training data. During the classification process, the classifier ensemble uses majority votes from N basic classifiers to determine the class label for each test data. Bagging is an ensemble learning algorithm, which extracts samples from the training set, constructs multiple base learners, and sets their predictions together for final prediction. Diversity can be obtained by re-sampling different data subsets.

UnderBagging [49] randomly undersamples the data set in each Bagging iteration, leaving all the minority class instances in each iteration. To avoid the inherent disadvantages of over-sampling and under-sampling techniques, UnderBagging replaces a separate classification model with an unbalanced training sample (using 1-NN rule), and through the combination of multiple classifiers, each classifier uses a balanced training sample as its learning process. Working in this way, the difficulties of imbalance can be

appropriately addressed. To achieve this, as many training subsamples as are needed to achieve balanced subsets are generated, the number of which will be determined by the difference between the number of prototypes for majority classes and minority classes, for example, if the majority class is about seven times larger than the minority class in a particular data set, the method creates seven subsamples to create a set of seven 1-NN classifiers. Each individual classifier is trained as a learning set consisting of all prototypes in minority classes and an equal number of training instances selected from samples belonging to majority class. The prototypes for most of the classes contained in each training subsample were selected by two different programs. This results in two different types of ensemble: (1) Majority class prototypes are randomly selected and not replaced. (2) Randomly select prototypes from the majority classes and replace them. Experiments show that the ensemble model can improve the accuracy of the classifier.

Another example, SMOTEBagging [13] uses SMOTE in each iteration. The new data sets contains twice the number of instances of majority class. The first half is the guided replication for majority class instances, and the second half is from SMOTE and the random oversampling re-sampling rate. SMOTEBagging's goal is to find the impact of diversity on the unbalanced data sets. First, combine three commonly used re-sampling methods into the ensemble model, give the diversity analysis of under-sampling, over-sampling, and data ensemble algorithm SMOTE, then improve SMOTE to address the multi-class data set in the ensemble model. This algorithm implements three ensemble models, each using Bagging to ensemble each classifier, but with a different re-sampling method. These are called UnderBagging, OverBagging, and SMOTEBagging, respectively.

In order to make the best use of majority class data in the training process without synthesizing minority classes of data, Li *et al.* [50]proposed a method that combines the idea of under-sampling with the classical ''Bagging'' ensemble method. For unbalanced data, the number of target class data is much smaller than that of non-target class data. The most direct approach is to include as much target class data as possible in N training data, and use different non-target data in each training data to introduce variation. In order to equalize each of N training data, the size of target class data and non-target class data in each training data is equal. In the BEV system, the non-target class data is divided into N disjoint sets, where $N = \lfloor M_{nt} / M_t \rfloor$. the M_{nt} is the number of non-target class data, M_t is the number of target class data. Each of the N training data sets is composed of one of the N non-target class data sets combined with the target class data set. This results in N training data sets, each with the same target-class data set and a different set of non-target-class data set. Then learn a classifier from these N training data sets. During the classification process, a class tag is generated for each N classifiers as a vote, and the majority votes of N classifiers are used as the classification of new data. The experimental results show that the BEV system is very effective and generates an ensemble classifier with high sensitivity, specificity, and G-means value in the process of classification.

Hido *et al.* [51] developed a new sampling method to improve Bagging on data sets with the unbalanced class distribution. In the new sampling method ''roughly balanced Bagging'' RB-Bagging, the number of samples in the largest and smallest classes is different, but they are effectively balanced when averaged over all subsets, which supports a more appropriate Bagging method. This approach differs from the existing Bagging method for unbalanced data sets, Bagging method extracts the same number of majority and minority class instances of the sampled subset data. In addition, RB-Bagging takes full advantage of under-sampling in all but minority class instances, which is an effective use of negative binomial distribution. The negative binomial distribution is a probability distribution of the number of failures m in Bernoulli trials given the number of successes n, which is defined by the following probability mass function:

$$p(m|n) = \binom{m+n-1}{n} q^n (1-q)^m \tag{3}$$

Experiments using benchmarks and real data sets show that RB-Bagging packages existing ''balanced'' methods with other commonly used methods. This algorithm can maintain the original Bagging property. Compared with accurate balance models and other algorithms (such as AdaBoost and RIPPER), RB-Bagging generally outperforms them, especially in performance metrics such as AUC, ISE, or G-mean, which are considered appropriate for solving unbalanced problems. RB-Bagging also shows very clear advantages for real-world financial data sets.

2) UNBALANCED DATA ENSEMBLE CLASSIFICATION ALGORITHM BASED ON BOOSTING

Boosting based unbalanced data sets ensemble classification correlation algorithm is to train each classifier serially with the whole data sets, but after each round, it will pay more attention to the wrong instances, the goal is to correctly classify those instances that have been misclassified in the current round in the next iteration. In this work, AdaBoost. M2 [47] is often used, and it is widely used in the field of imbalances in combination with data-level techniques. AdaBoost.M2 is bounded above by

$$\varepsilon \leq (k-1)2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t (1-\varepsilon_t)} \tag{4}$$

Compared with other Boosting methods, AdaBoost.M2 has the main advantage that it makes use of the confidence given by the base classifier when updating the weight.

Seiffert *et al.* [52] proposed a new hybrid sampling algorithm for learning from unbalanced training data, called RUSBoost. RUSBoost uses the random under-sampling technique to remove instances from majority class in each iteration of AdaBoost.M2. The instance weights in the new

under-sampled data sets are normalized to form a distribution. The algorithm provides a simpler and faster alternative to SMOTEBoost, another algorithm that combines boosting and data sampling to improve the performance of the model based on unbalanced data training. RUSBoost uses RUS, a technique for randomly deleting instances from majority classes. The process for RUS to introduce Boosting is simplicity, speed, and performance, and RUS performs well in spite of its simplicity [13].

SMOTEBoost [14] presents minority class examples of the class synthesis using SMOTE. Because a new instance is created, a new weight must be assigned that is proportional to the total number of instances in the new data set. The weights of instances from the original data sets are standardized to form a distribution with the new instances. SMOTE is a more complex and time-consuming data sampling technique than RUS, so SMOTEBoost is more complex and time-consuming to execute than RUSBoost. In addition, SMOTE is an oversampling technique that has the disadvantage of increasing the model training time. SMOTEBoost amplifies this deficiency because the enhanced power requires training a set of models so you have to build many models for a longer training time. SMOTEBoost describes the combination of SMOTE and boosting, which is also a variation of the AdaBoost.M2 program [53], in a series of t rounds. In each round, a weak learning algorithm is called a different distribution D_t, varying the D_t by emphasizing specific training instances. Update the distribution so that the weight of the misclassification is higher than the weight of the correct classification. It is different from the standard boosting, in the standard boosting, the distribution D_t is updated uniformly for instances of both the majority and the secondary classes, but in SMOTEBoost, the distribution D_t is updated in order to oversample the instances of the minority class by creating synthetic minority class instances. The whole weighted training set is given to the weak learner to calculate the weak hypothesis h_t. Finally, combined the different hypotheses into a final hypothesis.

EasyEnsemble [54] performed similar to UnderBagging, but despite training a classifier for each new block, they trained each block using AdaBoost. Thus, although the final classifier is a single ensemble, it looks like an ensemble of the whole. Under-sampling is a common method for dealing with class imbalance, which uses only a subset of most classes and is very effective. The main drawback is that many of the majority class instances are ignored. Liu et al [55] proposed two algorithms to overcome this defect. This algorithm extracts several subsets from majority class, uses each subset to train the classifier, and combines the output of these learners. BalanceCascade trains the classifier sequentially, and most class examples correctly classified by the currently trained learner at each step are excluded from further consideration. They both make better use of majority class than under-sampling because multiple subsets contain more information than a single subset. The main difference is that EasyEnsemble samples independent subsets, while BalanceCascade uses a trained classifier to guide the sampling process of subsequent classifiers. When using the same number of weak classifiers, the training time of the two algorithms is basically the same as that of the under-sampling algorithm. While they provide strong generalization capabilities, they also inherit the weakness of the ensemble approach: a lack of comprehensibility. Even though the basic quantifiers are comprehensible to sign learners, the combine parts are still difficult to understand.

TABLE 3. Comparison of classification algorithms for classical unbalanced data sets based on ensemble learning.

Typical Algorithm	Compared algorithm	Character
UnderBagging[46]	Oversampling, under-sampling	It is more suitable to deal with unbalanced TS problems. Avoiding the inherent drawbacks of over-sampling and under-sampling techniques.
SMOTEBagging[47]	UnderBagging, OverBagging	SMOTEBagging is more suitable for solving problems with multi-class data sets.
SMOTEBoost[14]	AdaBoostAdaCost	SMOTEBoost can construct an ensemble of diverse classifiers and reduce the bias of the classifiers.
RUSBoost[51]	SMOTEBoost	RUSBoost can reduce the model training times.

IV. OTHER UNBALANCED DATA SET CLASSIFICATION METHODS

Unbalanced data sets are very common in studies, and their characteristics lead to poor performance in some classification algorithms. For example, literature [56] points out that support vector machines perform poorly in dealing with learning problems of severely unbalanced data sets. Many studies have used sampling techniques and new classification algorithms to solve the problem of class imbalance in many fields, such as biodata analysis and text classification. However, there are still some problems with these methods in use. The Under-sampling method is likely to eliminate valuable samples when operating on data sets, while over-sampling method will lead to over-fitting of data by classifiers whether copying existing samples or synthesizing new samples [57]. In addition, the problem of class imbalance is usually accompanied by the problem of higher dimensions of the data set. However, research shows that sampling techniques and algorithm methods may not be sufficient to solve the high-dimensional class imbalance problem [58]. In order to solve the above problems, many people choose to classify unbalanced data sets at the feature level. This method mainly combines the data level and the algorithm level and makes full use of their advantages to reduce their disadvantages [59]. Many researches are devoted to the combination of sampling method and cost-sensitive method [60]. At the same time, the hybrid method also gathers the advantages of the data level and the algorithm level. How to make the combination of the two to achieve the best results requires more analysis of

the relationship between them. This chapter will introduce the classification method of unbalanced data sets from the aspect of cost sensitive learning, feature level and deep learning.

A. COST SENSITIVE LEARNING

In addition to using an ensemble method to deal with unbalanced data sets, many researchers also choose a cost sensitive learning method to adjust the situation of unbalanced data in classification. Among them, cost sensitive learning is divided into direct cost sensitive learning and indirect cost sensitive learning. Cost sensitive learning is usually an application-dependent classification algorithm. Cost-sensitive methods can be divided into three categories: the cost of non-uniform misclassification, the cost of testing, and the mixed cost of different types of cost combinations. The non-uniform error classification cost method only considers the misclassification cost of different types of errors and different costs [61]. The test cost-based method considers only the test cost and does not consider any misclassification cost [62]. The hybrid cost method considers various types of costs, including misclassification cost, test cost, and even time cost [63]. Cost sensitive learning is frequently used in neural networks, decision trees and, support vector machines. A cost sensitive support vector machine is a representative method in cost-sensitive learning, which combines different misclassification costs for each class and improves its generalization performance by assigning higher misclassification costs to minority class.

Based on the assumption that the cost of misclassification is known, Li *et al.* [1] proposed a cost-sensitive mixed attribute measure multiple decision tree (CHMDT) method for binary classification of non-equilibrium data sets. Its purpose is to improve the classification performance of minority class in unbalanced data sets. The multi-decision tree is constructed based on different root node information, and then the cost sensitivity theory is embedded into the multi-decision tree of the unbalanced data sets. Based on the Gini index and information gain measure, the concept of mixed attribute measure is proposed as the attribute criterion. The motivation is to punish the two classes for being unbalanced, and punish the positive class more seriously, so as to improve its classification accuracy. They theoretically prove the property of mixed attribute measures and prove the classification performance of the CHMDT method on unbalanced data sets.

Based on similar principles, Veropoulos *et al.* [64] introduced a biased support vector machine by assigning different costs to the majority class and the minority class, which is very useful for making the final hyperplane deviate from the minority class. However, this method does not take into account the different effects of samples of the same class on the formation of the classifier, which may lead to overfitting of noise and outliers. In order to solve this problem, Tao *et al.* [65] proposed a new cost sensitive ensemble method of support vector machines based on adaptive cost weights to deal with unbalanced data sets. Boosting scheme is an improved cost-sensitive one instead of the standard boosting scheme, so as to ensure the consistency between boosting scheme and weak classifier, so as to ensure the consistency between the boosting scheme and the weak classifier. In order to ensure the consistency of the optimization objectives between the weak learner and the boosting scheme, we not only use the cost-sensitive support vector machine as the basic weak learners, but also modify the standard promotion scheme as the cost-sensitive supporter. In order to ensure that more training is provided for minority class instances of subsequent classifiers, especially the minority class instances on the boundary, we also propose an adaptive sequential misclassification cost weight determination method. Based on the classifiers previously obtained in the boosting process, this method can adaptively consider the different contributions of minority class instances to the SVM classifier form in each iteration, which can make it produce a variety of classifiers and thus improve its generalization performance.

Ping *et al.* [66] is proposed based on clustering of weak balance cost sensitive random forest algorithm used for imbalanced data classification algorithm, the first of all, the K-means clustering method is used to send a sample, according to the unbalanced proportion of reduced to set the sampling probability, and then select some samples from each cluster and constitute the new sample to train the cost sensitive decision tree. Through the introduction of misclassification cost splitting attributes, using price decline in value instead of information gain rate, the Gini index, such as indicators, to seek the minimum cost classifications to build a decision tree, in the process of splitting, The misclassification cost of choosing attribute maximization decreases. Since the misclassification substitution value of the minority class is higher than that of the majority class, the attribute selection strategy makes the number of misclassification samples of the minority class as small as possible, so as to improve the attention to the majority class samples. An unbalanced classification method based on cost sensitive learning aims at minimizing the cost of misclassification, which can effectively improve the classification performance of a single decision tree.

Elkan *et al.* [67] proposed a theorem, which explained how to change the negative proportion of the training set by using the classifier learned by the standard non-cost sensitive learning method, so as to make the optimal cost sensitive classification decision. In order to make the target probability threshold $P*$ correspond to the given probability threshold P_0, the negative sample number in the training set is multiplied by

$$\frac{P*}{1-P*}\frac{1-P_0}{P_0} \tag{5}$$

But the theorem does not say in what way the number of counterexamples should change. If a learning algorithm can use the value of the training sample, then the weight of each negative sample can be set to the factor given by the theorem. Otherwise, oversampling or under-sampling must be used. In order to solve the above problems, they proposed another

theorem, which represented probability P' as

$$P' = b' \frac{p - pb}{b - pb + b'p - bb'} \tag{6}$$

which allowed the use of a classifier learned from a training set extracted from a probability distribution and applied it to test sets extracted from different probability distributions.

B. CLASSIFICATION ALGORITHM OF UNBALANCED DATA SETS AT FEATURE LEVEL

Another problem that can occur in a data set is the sample size. Without a large training set, a classifier may not be able to generalize the characteristics of the data. Classifiers can also overfit training data and be misled at test points, especially for high-dimensional data [68]. The main purpose of processing unbalanced data sets at the feature level is to solve a series of problems brought by unbalanced data distribution to classification. The unbalanced sample distribution leads to unbalanced information transmission and expression at the feature level, which brings certain difficulties to minority class in the process of identification. Feature selection will select the optimal data conforming to the evaluation in the sample set as the feature subset according to certain evaluation criteria, and effectively distinguish the data in the sample sets. Feature selection algorithms can be divided into two categories, feature grade and feature subset selection [69]. Feature sequencing evaluates each individual feature against some criteria, and then analyzes and selects some features that are appropriate for a given data set.

Liu *et al.* [70] proposed a feature selection method based on multiple strategies. A strategy that uses information gain, chi-square value, Pearson correlation methods such as correlation analysis as the basic evaluation algorithm, the first to use the basic evaluation method to evaluate software static defect data, and the evaluation results according to certain order is divided into m level, and then according to the characters of each frequency in different levels of secondary sorting, and then selected in the sequence of k characteristics of the new optimal feature subsets is the optimal feature subset, to balance the lack of data in a project. The flowchart of the multi-strategy feature screening method is shown in Figure 3(F is the feature evaluation set.).

Sun *et al.* [71] used the method of feature selection in the process of realizing their unbalanced data classification method. First, the instances in the set are trained to obtain the corresponding class value, and then the characteristic difference degree on different instances is obtained. At this point, the weights of all features are initialized to 0, an instance is randomly selected from the set, an instance with the minimum distance is found from instances with the same or different tag values, the cost of feature testing is taken into account, and the weight of each attribute is finally updated.

According to the idea of feature selection, researchers have proposed many methods about feature selection to improve the accuracy of classification. Research on software defect prediction shows [72] that feature selection and feature extraction can provide great help for its research. VanderPutten *et al.* [73] analyzed the data set of Coil Challenge 2000 and found that feature selection can effectively solve the problem of over-fitting more effectively than the selection of classification algorithm. Secondly, feature selection is a key step in many machine learning algorithms (e.g. [74]–[76]), especially when the data volume is large. The carefully selected features make the mutual variation of the two classes larger and divide the two classes into small clusters [58], [77]. Therefore, greater internal variation can mitigate the impact of decision boundary excursion on classifier performance. A small internal variation will lead to the overlapping of multiple samples of majority classes, thus reducing the skewness of the classes in order to improve the classification accuracy.

Nam *et al.* [78] proposed a TCA (Transfer Component Analysis) method based on feature mapping angle, which maps the features of source project and target project to a potential space closest to them, so as to balance the difference of data distribution among different projects. After that, they retained the advantages of TCA and proposed TCA$^+$ [79] on this basis. This method can not only map two items to a space but also analyze its characteristics and automatically analyze a data normalization method. However, this method is only suitable for the same situation of defect data measurement. When the data measurement differences and distribution differences between projects increase dramatically, this method may be faced with the possibility of failure.

Yu *et al.* [80] proposed a feature selection algorithm for software defect prediction based on a similarity measure (SM). The algorithm firstly updates feature weights according to the similarity of samples of different classes. Secondly, the feature is sorted descending according to the weight to generate the feature sort list, and all feature subsets are selected from the sorted feature list in turn. In order to further improve the feature selection performance for software defect prediction. Okutan and Yildiz [81] use the BN network to sort them according to their correlation degree with defect orientation, so as to select the information content and characteristics. These information measures include software measures and probabilistic influence relationship between defects.

Gao *et al.* [82] proposed a feature mixed attribute selection based on feature subset selection method, sorting priority characteristics, according to the given standard to separate assessment of properties, they put forward a kind of attribute subset selection method called ASH, first, calculate the complete set of properties of CR, CR is to use a consistent count calculation of a kind of evaluation criteria, and then start from any attribute of the size of the 1, choose to have local maximum attribute subsets of the CR. These selected subsets of attributes will be used to generate the superset. This process is repeated until a subset of the attributes with the same CR as the original data set with full characteristics is found or the specified number of attributes is reached. It is then sorted, and the number of related features that are specified in advance is selected to reduce the search space.

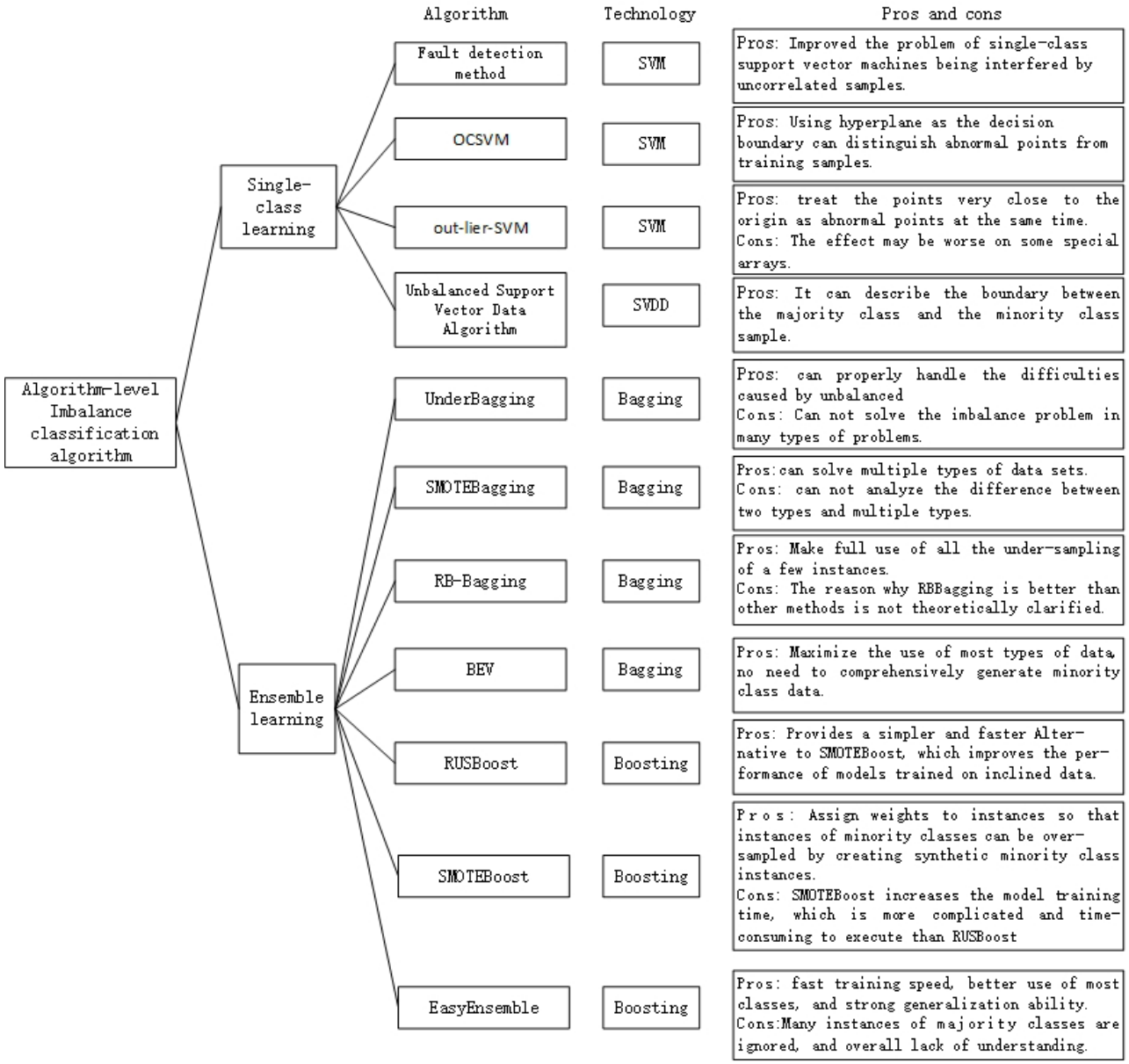


FIGURE 4. Unbalanced data sets classification method based on sampling methods.

This hybrid feature selection algorithm combines feature classification with feature subset selection and is superior to other feature subset selection methods in software defect prediction.

Laradji *et al.* [83] studied the several methods of software defect feature selection technology, found that choose a small number of characteristic can obtain a higher AUC value, they used a forward selection technique, the forward selection is a commonly used technique to select good features. Figure 4 illustrates the process of forwarding selection using a feature set consisting of cyclomatic complexity (CC), weighted methods per class (WMC) and LOC software metrics. Using an initially empty set, forward selection selects the first feature from the full feature set. In the case of Figure 4(a), the feature subset contains only the LOC metric at the first iteration. Then, defect classification is carried out and its performance is evaluated. Afterward, a second feature is selected and the subset contains LOC and CC metrics. Using this augmented subset, the classification performance is evaluated as well. The same process is repeated gradually until all the features are considered. Finally, the feature subset achieving the highest accuracy is retained. However, it is obvious that forward selection becomes time-consuming when dealing with a large set of features. Efficient feature selection is highly desirable. Greedy forward selection (GFS) is an efficient and simple, feature selection technique. Unlike other time-consuming schemes, the GFS selects only those features that contribute positively to the improvement of classification performance. Then, This process is demonstrated in Figure 4(b).

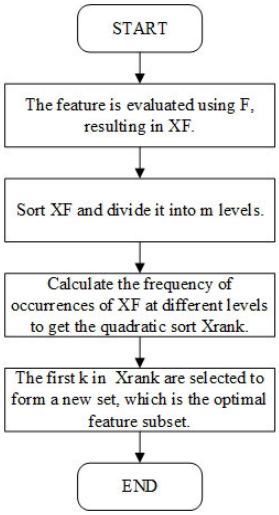


FIGURE 5. Flow chart of a multi-strategy feature screening method.

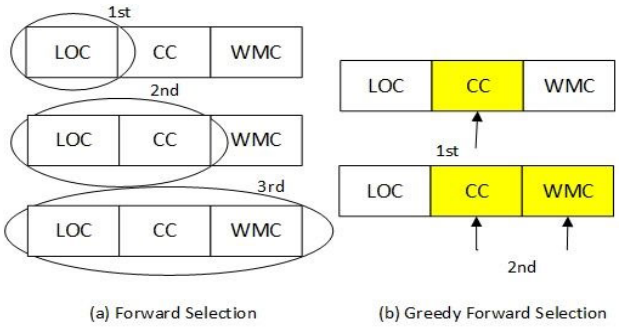


FIGURE 6. Forward selection.

C. CLASSIFICATION ALGORITHM OF UNBALANCED DATA SETS BASED ON DEEP LEARNING

Deep learning is a research hotspot in recent years, its characteristic is to study the inherent law of sample data and presentation level, the researchers use its excellent characteristics of learning data rules which applied to deal with unbalanced data sets, by learning layer by layer feature transformation of neural networks, the characteristics of the samples in the original space said to transform into a new feature space, thus make it easier for classification or prediction.

Zhou *et al.* [84] proposed based on adversarial deep denoising autoencoder, (GAN-DAE), through the generator and the discriminant of adversarial training, access to the unbalanced data the characteristics of positive and negative samples, and training generator to generate a minority class samples, to improve the unbalance in the data sample, optimize the neural network structure and parameters of the improved depth of the neural network performance and classification accuracy. A deep neural network ensemble model optimized by the evolutionary algorithm is proposed. GAN-DAE is taken as the member neural network, and a group of GAN-DAE is ensemble into a comprehensive classification model. The weight of each member neural network in the model is optimized by the evolutionary algorithm. Experimental results also show that this method greatly improves the accuracy of unbalanced data sets classification.

Due to unbalanced data sets of samples of different classes, different importance in the process of training, using the traditional classifier to classify unbalanced data sets is difficult, Xie et al [85] put forward a kind of based on generative adversarial networks (GAN) of unbalanced data sets classification method, using Wasserstein GAN (WGAN) the formation of stable ability to generate minority class samples in order to solve the problem of shortage of one type of class. WGAN modified the loss function and network structure of the original GAN to make the training more stable. It used the stable generation ability of WGAN to synthesize a large number of minority class samples so that the two kinds of samples reached equilibrium and balanced the classes of samples.

Konno *et al.* [86] said such as deep neural networks (DNN) to extract the characteristic of the small sample as the basic characteristics, and then join part of pseudo features to generate new samples to make up for the lack of minority class samples, by using the deep neural networks can extract the characteristics of the samples, from each kind of multivariate probability distribution, the characteristics of the minority class features extracted from multivariate probability, thus increasing the minority class training data, but this method does not generate data, it generates characteristics, then training is carried out according to the extracted features, which can effectively improve the classification results of unbalanced data sets.

When classifying unbalanced data sets, the class unbalance leads to the neural network model being dominated by the majority of classes, leading to a poor classification effect. In view of this phenomenon, Chen *et al.* [87] proposed to introduce the loss function in the process of convolutional neural network training into the class label weight, so as to strengthen the influence of minority classes on model parameters. The text classification experiment also shows that this method can significantly improve the F1 value and improve the ability of the neural network to classify unbalanced data sets.

D. OTHER UNBALANCED DATA SET CLASSIFICATION METHODS

The method described above is not the only way to deal with class unbalanced data sets, and some algorithms designed especially unbalanced data sets run well on raw, unmodified unbalanced data sets. For example, a variant of the association classifier called SPARCCC [88] performs better on balanced data sets than CBA [89] and CMAR [90]. Liu *et al.* [91] designed a decision tree classification algorithm that is robust to inter-class interference in data, and proposed a new measurement method: quasi-confidence ratio (CCP), because the traditional decision tree measurement, such as information gain, is realized through rules of information transmission. Information acquisition, like confidence, is biassed towards the majority class and is therefore sensitive to classification balance. This method uses theoretical and geometric parame-

ters to prove that CCP is insensitive to class distribution, and then embedding CCP into information acquisition and using temporary measures to build a decision tree. This method proves the effectiveness of CCP when the data set is unbalanced.

In addition to the unbalanced data sets ensemble classification method mentioned above, there are also hybrid methods that combine the set method and the data set method. Ren *et al.* [92] proposed a novel classifier, called selection-based resampling ensemble(SRE), whose purpose is to learn non-stationary unbalanced data streams. SRE before the first use of the minority class samples for minority class sets of current sampling, at the same time the SRE also ensembles block-based and online ensemble operators, regularly updated before the classifier, and quickly be equally divided in advance, because the data ensemble members based on the continuous data block, put all the classifier of the weighted results for decision making. It makes full use of data flow knowledge to approximate batch results and is a hybrid ensemble approach.

Although traditional support vector machines are good at classifying unbalanced data sets, the characteristics of support vector machines tend to lead to the final decision boundary biasing toward the majority class especially in the presence of outliers or noises. Tao *et al.* [93] proposed a new Affinity and class probability-based fuzzy support vector machine(FSVM) technique. In ACFSVM, SVDD model training is carried out on all the given majority class samples first to calculate the different affinity of each sample in the majority class. The affinity can effectively identify outliers and some boundary samples in the majority class training samples. At the same time in order to avoid the noise influence on classification, most kernel KNN technique is adopted to define each class of sample classification probability, the samples with lower class probabilities are more likely to be noises, their contribution for learning seems to be reduced by their low memberships constructed by combining the affinities and the class probabilities, so ACFSVM can reduce the affinity and the influence of the lower class probability sample.

The density peak clustering algorithm is a novel density clustering algorithm, but its effect is not ideal when clustering data with unbalanced density distribution, and it lacks the standard when selecting a clustering center. To solve the above two problems, Yang *et al.* [94] proposed an improved density peak clustering algorithm AD-PC-WKNN based on the weighted k-nearest neighbor. This algorithm redefined the local density and predetermined the critical point of the clustering center by combining the idea of k-nearest neighbor, thus improving the classification effect of data sets with the unbalanced distribution.

Bikku method proposed by [95] is also a kind of hybrid method, the efficiency of document analysis in their proposed Hadoop framework is limited by class imbalance and a large set of candidates, they adopted a with text preprocessor and reduce the classification model of two phase mapping framework to deal with this problem, in the first phase is designed based on the mapping of pretreatment method, to eliminate the irrelevant features in the data, missing values, and outliers. In the second stage, a multi-class ensemble decision tree model based on mapping reduction is designed, and the model is implemented on the preprocessed mapper data to improve accuracy and computation time. The whole process adopts the methods of data preprocessing and ensemble learning to deal with some problems caused by unbalanced data sets classification.

LDA [96] deal with unbalanced data sets do not need to adjust any parameters, the basic idea is to make the higher dimensional pattern projection can best recognize the vector space, to extract the classification information and the effect of compression feature space dimension, its and said before the PCA are common dimension reduction techniques, PCA mainly from the characteristics of covariance Angle to find a good way to projection. However, LDA considers the annotation more, that is, the distance between the data points of different classes is expected to be larger after the projection, the data points of the same class can be more compact so that the classification of unbalanced data sets can be handled better.

ELM is an algorithm based on a single hidden layer feed forward neural network. It greatly improves the training speed by randomly generating input weight and the deviation of hidden layer nodes. There is no need to adjust parameters during iteration. ELM has a significant advantage over other traditional neural network algorithms. Using ELM as the basic classifier of an ensemble network can guarantee the accuracy of a single network. However, due to the need to train all samples in the training stage, single ELM still has the disadvantages of overfitting and low generalization ability.

Li *et al.* [97] proposed an ensemble over-limit learning machine based on layered cross-validation, and introduced ensemble learning methods and layered cross-validation strategies into the network training process. The ensemble method and the stratified cross-validation method are ensemble to reduce the over-fitting, improve the generalization ability and enhance the classification ability of unbalanced data. All the samples are divided into K subsets according to the classes, one of which is taken as the verification set, and the rest (K-1) subsets as the training set. In the initial stage, K cross-validation was adopted, and the K classification results are averaged according to the values calculated by the evaluation criteria, and the optimal values are stored, and the corresponding input weights and deviations were adjusted. Then the weight update and ensemble in the process of learning, if a better classification result is to always keep in the iterative optimal input and hidden layer weights deviation decision, after completion of training, use training to get the optimal weights and bias of classifying test sample. In the training process, the distribution of each type of sample of the training set and verification set is consistent with the test set, so the characteristics of samples can be better learned and the testing effect can be guaranteed.

TABLE 4. Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True Positive(TP)	False Negative(FN)
Negative class	False Positive(FP)	True Negative(TN)

Because traditional classifiers tend to have a greater preference for most classes, resulting in lower accuracy for minority class, Sun *et al.* [98] proposed a D-WELM (Data distribution based extreme learning machine). This algorithm uses cost sensitive learning, which takes into account not only the influence of the number of sample classes but also the global loss to design a new weighted scheme. Where the mathematical expression of the weighted scheme is:

$$W_{new_{ii}} = (\frac{1}{r_{i1}} + r_{i2})(1 - \frac{\#(t_i)}{\#(T)}) \tag{7}$$

V. EVALUATION CRITERION OF CLASSFIER

Accuracy is a basic method to judge whether a classifier is excellent. Like the equally common error rate, it is also a common indicator to judge a classifier. However, due to the characteristics of unbalanced data sets, neither of them is applicable in this scenario. Because even if the classifier achieves 90% accuracy in unbalanced data sets,, it is very likely to achieve quite a high accuracy in majority class instances, and its accuracy minority class instances cannot be measured. Therefore, the true positive rate and false positive rate of confounding matrix, ROC curve, G-means, and other methods are usually used in the classification of uneven data to evaluate the performance of classifiers, because they can better measure the effect of classifiers based on the characteristics of unbalanced data sets.

1) CONFUSION MATRIX EVALUATION CRITERION METHODS

It is very important to evaluate its classification performance to guide the modeling. Based on the basic concept of the confusion matrix, the evaluation index of unbalanced data sets classification is proposed. For the problem of two classes, the results of the correct and wrong identification instance of each class can be recorded in the confusion matrix (as shown in Table 4). Among the two classes of questions, minority class of instances are Positive and majority class of instance are Negative. True positive means that the predicted result of the positive sample is still positive, false negative means that the predicted result of the positive sample is still negative, false positive means that the predicted result of the negative sample is still positive, and true negative means that the predicted result of the negative sample is still negative.

Predictive accuracy is a performance measure commonly associated with machine learning algorithms, defined as accuracy $= (TP+TN)/(TP+FP+TN+FN)$. With a balanced data set and equal error costs, it makes sense to use the error rate as a performance metric. The error rate is 1- accuracy. In the case of unbalanced data sets with varying error costs, ROC curves or other similar techniques are more appropriate.

The confusion matrix is used for two classes of problems, algorithms using the confusion matrix as evaluation criterion include Borderline-SMOTE, Safe-level-SMOTE, UnderBagging, etc.

2) G-MEANS EVALUATION CRITERION METHODS OF

Using the product of two classes of prediction coefficients to calculate G-means, even if a model correctly classifies the negative example, the poor prediction performance of the positive example will lead to low G-means. In fact, G-means is very important for measuring the degree to which overfitting avoids negative class and the degree to which positive class is ignored. It is a simple and effective index to measure the unbalanced data sets classification method, which is defined as:

$$G - means = \sqrt{A^+ \times A^-} \tag{8}$$

It can be seen that G-means comprehensively considers the classification accuracy of the two classes of samples. Compared with the accuracy, it can better measure the classification effect of the classification method on the unbalanced data sets. Moreover, it is simple, effective, and easy to understand. It has become one of the commonly used methods in the field of unbalanced data processing.

G-means is suitable for unbalanced data sets where the sample distribution may change over time or where the sample distribution of the training set and the test set differ. Algorithms using G-means as evaluation criterion include SMOTEBagging, RBBagging, A-SUWO, etc.

3) RECALL RATE EVALUATION CRITERION METHODS

The recall rate is for the original sample, and it shows how many positive examples in the sample were correctly predicted. There are two possibilities, one is to predict the original positive class as the positive class (TP), the other is to predict the original positive class as the negative class (FN).

$$R = \frac{TP}{TP + FN} \tag{9}$$

The recall is suitable for more data sets of positive cases because it is an evaluation criterion for the calculation of positive cases. Algorithms using G-means as evaluation criterion include Safe-level-SMOTE, SMOTEBoost, etc.

4) F-MEASURE RATE EVALUATION CRITERION METHODS

F-Measure, also known as F-score, is the weighted harmonic average of recall rate R and precision P. As the name implies, it is to reconcile the contradiction between the increase and decrease of recall rate and precision. This comprehensive evaluation index F introduces a coefficient α to carry out the weighted harmonic of recall rate and precision, and the expression is as follows:

$$F = (\alpha^2 + 1)P.R/\alpha^2(P + R) \tag{10}$$

The most commonly used F1 indicator is the case where the coefficient α in the above formula is 1, that is:

$$F1 = 2P.R/(P + R) \tag{11}$$

F1 has a maximum value of 1 and a minimum value of 0.

F-Measure applies to data sets where there is a contradiction between the accuracy of use and recall rate. Algorithms using F-Measure as evaluation criterion include SMOTEBagging, RUSBoost, etc.

5) ROC CURVE EVALUATION CRITERION METHODS

The quality of the results obtained by a classification algorithm should be evaluated by its performance on both classes, so these individual measures are still insufficient. The receiver operating characteristic(ROC) curve corres-ponds to the probability that the classifier ranks the randomly selected positive instances higher than the randomly selected negative ones. One approach to the ROC curve is by controlling the balance of training samples for each class in the training set. ROC graphics make it intuitive to see the tradeoff between TPrate (benefit) and FPrate (cost), proving that for any classifier, it can't justify increasing the number of true positives rather than false positives. The false positive rate refers to the ratio of the number of positive samples in all negative samples and the true positive rate refers to the ratio of the number of positive samples in all positive samples. The area under the ROC curve (AUC) [99] corresponds to the probability that the classifier sorts the randomly selected positive instances higher than the randomly selected negative ones. AUC provides a scalar measure of classifier performance, which has been widely used in the field of imbalance [29], [100], [101]. The AUC measured value is calculated as the area of the ROC curve:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{12}$$

ROC space [102] represents the FP rate FP/n^- on the X-axis of the graph, and the TP rate TP/n^+ on the Y-axis. Each classifier can be represented by a point in the ROC space corresponding to its FP and TP rates. The point (0,0) corresponds to a strategy that never makes a positive (minority) prediction, and the point (1,1) corresponds to a class that always makes a positive prediction. The point (0,1) represents the perfect classification (all positive patterns are correctly classified, no negative cases are misclassified as positive), and the line $x = y$ represents the random guess classification strategy. In ROC analysis, classifier A is superior to classifier B if it is located in the upper left corner of B in the ROC space (high TP, low FP, or both). A classifier that allows smooth variations in multiple parameters can be represented in the ROC space by an appropriate curve. On the other hand, the geometric mean of accuracy measured separately for each class [103] is defined as $g = (acc^+ \times acc^-)$, where $acc^+ = TP/n^+$ is the accuracy of positive class patterns, and $acc^- = TN/n^-$ represents the accuracy of minority class patterns. This metric is closely related to the distance to perfect classification in the ROC space. The ROC curve can be regarded as the optimal decision boundary class for TP and FP relative costs.

ROC is not affected by sample distribution, so it is suitable for classification evaluation of unbalanced data sets, but it may not be applicable in scenarios sensitive to classification accuracy. Algorithms using the ROC curve as evaluation criterion include SMOTE, RUSBoost, A-SUWO, etc.

VI. FUTURE WORK

Unbalanced data sets classification methods are more and more, and many researchers according to the characteristics of the unbalanced data sets to put forward many methods from a different perspective, but in the face of unbalanced data sets, there are still many aspects need to simplify and overcome, the following will discuss for the further research direction of this kind data sets.

A. IMPROVEMENT OF CLASSIFICATION ALGORITHM BASED ON THE SAMPLING TECHNOLOGY

The sampling technique is a very common data-level classification algorithm for unbalanced data sets, which balances the class imbalance of data sets by under-sampling or over-sampling techniques. Many algorithms have begun to deal with noise in unbalanced data sets, similar to noise filters. However, the techniques used are still limited and other noise filtering and optimization techniques should be tried to deal with some difficult to balance data sets. It is also possible to study different data sets, study the shortage of sampling methods for specific data sets, directly improve the sampling ratio and other factors that can affect the performance of different levels of class imbalance, so as to improve the classification effect of the algorithm on unbalanced data sets.

B. IMPROVED CLASSIFICATION ALGORITHM BASED ON UNBALANCED DATA SETS AT THE FEATURE LEVEL

The existing classification methods for high-dimensional unbalanced data sets mostly adopt the feature level method. When the number of samples increases or the degree of imbalance decreases, the advantage of feature selection in solving unbalanced data sets may gradually weaken. This issue should be studied in detail in future studies. In this study, we cannot only adopt the simplest feature subset search method based on the feature level but also improve it by other methods, such as sequential backtracking float selection and so on. In future work, we can also make a better prediction and estimate of the abrupt changes brought by the feature level method to classification performance, so as to avoid more waste of resources.

1) IMPROVEMENT OF UNBALANCED DATA SETS CLASSIFICATION ALGORITHM BASED ON COST SENSITIVE LEARNING

The existing cost sensitive learning is to set different misclassification costs for the base classifier so that the algorithm can better classify the unbalanced data sets. The cost sensitive function is introduced into a single classifier, and the

minimum feature of misclassification cost is used to divide it, so as to improve the attention to minority class samples and the classification accuracy of unbalanced data sets. In view of the current research situation of using cost sensitive functions to classify unbalanced data sets, more kinds of cost functions such as active learning cost, calculation cost, and storage cost can be used in future research to try to solve the classification problem of unbalanced data sets.

2) IMPROVEMENT OF UNBALANCED DATA SETS CLASSIFICATION ALGORITHM BASED ON DEEP LEARNING

Deep learning based classification algorithms for unbalanced data sets have achieved good results in the past two years, especially in the deep neural network ensemble model. In future research, we can try to improve and optimize the single deep neural network, improve the optimization strategy, and improve the performance of the neural network model. When using GAN to generate sample data, only input minority class samples. In this process, we can also try to use the traditional method of data enhancement to first enhance the unbalanced data sets, and then use GAN to generate data. The effect may be even better.

3) IMPROVEMENT OF THE CLASSIFICATION ALGORITHM OF UNBALANCED DATA SETS BASED ON THE ENSEMBLE METHOD

Current classification problem in view of the unbalanced data sets ensemble research, scholars have proposed many methods to a certain extent, can solve the problem of unbalanced data sets classification, but the ensemble of the existing algorithms are mostly for binary classification of imbalanced data sets is studied, in reality, a lot of problems in the scene are multiple classification problems, such as predict someone's age, or is someone's salary, and so on. Although there is individual research work for many kinds of classification problem are discussed in this paper, but the research on the ensemble classification of multi-class unbalanced data sets is not enough, still need to be further in-depth study. In addition, we can further study the size of the ensemble scale and the weight assigned to the base classifier for a specific algorithm. Choosing the appropriate ensemble size and the weight of the base classifier can make the algorithm get a better effect.

VII. CONCLUSION

The classification of unbalanced data sets is of great significance in data mining, because the unbalanced data sets are very common in real life, and its problems are becoming more and more obvious. As classification study in-depth step by step, more and more researchers are starting to study the extremely unbalanced distribution characteristics of unbalanced data sets, for the classification of unbalanced data sets algorithm is more and more comprehensive. Because unbalanced data sets in real life is very common, so this research has important practical significance. This paper introduces the classification methods of unbalanced data sets from five aspects, including data sampling method, algorithm level classification method, feature level classification method, sensitive cost function, and deep learning classification method. Especially in the data sampling method, a novel perspective based on SMOTE, SVM, and KNN sampling method is proposed to summarize and analyze the classification method of unbalanced data sets. These methods solve the problem of unbalanced data sets in classification. Then the evaluation criteria of unbalanced data sets classifier are introduced. In the end, the problems of unbalanced data sets at the present stage are prospected and the solutions are discussed.

REFERENCES

[1] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu, and Y.-C. Tian, "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets," *Inf. Sci.*, vol. 422, pp. 242–256, Jan. 2018.

[2] N. Seliya, T. M. Khoshgoftaar, and J. V. Hulse, "Predicting faults in high assurance software," in *Proc. IEEE 12th Int. Symp. High Assurance Syst. Eng. (HASE)*, Nov. 2010, pp. 26–34.

[3] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: Classification of skewed data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 50–59, Jun. 2004.

[4] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Appl. Soft Comput.*, vol. 38, pp. 714–726, Jan. 2016.

[5] J. Alqatawna, H. Faris, K. Jaradat, M. Al-Zewairi, and O. Adwan, "Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution," *Int. J. Commun., Netw. Syst. Sci.*, vol. 8, no. 5, pp. 118–129, 2015.

[6] D. X. Liu, S. J. Qiao, and Y. Q. Zhang, "A review of data sampling methods for unbalanced classification," *J. Chongqing Univ. Technol., Natural Sci.*, vol. 33, no. 7, pp. 102–112, 2019.

[7] L. I. Yong, Z. D. Liu, and H. J. Zhang, "Review of ensemble classification algorithms for unbalanced data," *Appl. Res. Comput.*, vol. 31, no. 5, pp. 1287–1291, 2014.

[8] M. Yang, J. M. Yin, and G. L. Ji, "A review of unbalanced data classification methods," *J. Nanjing Normal Univ., Eng. Technol. Ed.*, vol. 8, no. 4, pp. 7–12, 2008.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[10] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Berlin, Germany: Springer, 2005, pp. 878–887.

[11] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2009, pp. 475–482.

[12] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: Density-based synthetic minority over-sampling TEchnique," *Int. J. Speech Technol.*, vol. 36, no. 3, pp. 664–684, Apr. 2012.

[13] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Mar. 2009, pp. 324–331.

[14] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. Knowl. Discovery Databases (PKDD)*, 2003, pp. 107–119.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[16] J. F. Zhang, Z. Wang, W. S. Cui, and M. Liu, "Research on unbalanced data classification method based on SVM," *J. Northeast Normal Univ., Natural Sci. Ed.*, vol. 52, no. 3, pp. 96–104, 2020.

[17] L. Cao and H. Shen, "Imbalanced data classification using improved clustering algorithm and under-sampling method," in *Proc. 20th Int. Conf. Parallel Distrib. Comput., Appl. Technol. (PDCAT)*, Dec. 2019, pp. 358–363.
[18] H. S. Huang, J. A. Wei, and P. D. Kang, "New over-sampling SVM classification algorithm based on unbalanced data sample characteristics," *Control Decis.*, vol. 33, no. 9, pp. 1549–1558, 2018.
[19] L. M. Zhang, L. M. Tan, T. S. Liu, and X. Q. Sun, "Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm," *J. Phys., Conf. Ser.*, vol. 1237, no. 2, pp. 1–5, 2019.
[20] T. Ma, Y. Hou, J. J. Cheng, and X. Y. Chen, "A novel method combining fuzzy SVM and sampling for imbalanced classification," *Int. J. Appl. Systemic Stud.*, vol. 8, no. 1, p. 1, 2018.
[21] Q. Kang, X. Chen, S. Li, and M. Zhou, "A noise-filtered under-sampling scheme for imbalanced classification," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4263–4274, Dec. 2017.
[22] R.-L. Li and Y.-F. Hu, "Noise reduction to text categorization based on density for KNN," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Xi'an, China, vol. 5, 2003, pp. 3119–3124.
[23] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, pp. 405–416, Mar. 2016.
[24] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random undersampling to alleviate class imbalance on tweet sentiment data," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Aug. 2015, pp. 197–202.
[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.
[26] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognit.*, vol. 46, no. 12, pp. 3460–3471, Dec. 2013.
[27] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Comput. Statist. Data Anal.*, vol. 111, pp. 1–13, Jul. 2017.
[28] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci.*, vols. 409–410, pp. 17–26. Oct. 2017.
[29] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, Sep. 2009.
[30] X. Tao, Q. Li, C. Ren, W. Guo, C. Li, Q. He, R. Liu, and J. Zou, "Real-value negative selection over-sampling for imbalanced data set learning," *Expert Syst. Appl.*, vol. 129, pp. 118–134, Sep. 2019.
[31] C. X. Cui, F. Y. Cao, and L. J. Y, "Adaptive undersampling method based on density peak clustering," *Pattern Recognit. Artif. Intell.*, vol. 33, no. 9, pp. 811–819, 2020.
[32] X. Tao, Q. Li, W. Guo, C. Ren, Q. He, R. Liu, and J. Zou, "Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering," *Inf. Sci.*, vol. 519, pp. 43–73, May 2020.
[33] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, Feb. 2004.
[34] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Hybrid sampling for imbalanced data," *Integr. Comput.-Aided Eng.*, vol. 16, no. 3, pp. 193–210, Jun. 2009.
[35] W. Ng and M. Dash, "An evaluation of progressive sampling for imbalanced data sets," in *Proc. 6th IEEE Int. Conf. Data Mining Workshops (ICDMW)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 657–661.
[36] A. Estabrooks and N. Japkowicz, "A mixture-of-experts framework for learning from imbalanced data sets," in *Proc. 4th Int. Conf. Adv. Intell. Data Anal.* London, U.K.: Springer-Verlag, 2001, pp. 34–43.
[37] J. H. Zheng, S. Y. Liu, C. B. He, and Z. Q. Fu, "An improved random forest unbalanced data classification algorithm based on mixed sampling strategy," *J. Chongqing Univ. Technol., Natural Sci.*, vol. 33, no. 7, pp. 113–123, 2019.
[38] X. X. Yang, F. Su, and X. X. Huang, "Research on unbalanced data classification method based on improved random forest algorithm," *Netw. Secur. Technol. Appl.*, vol. 10, pp. 70–71, 2020.
[39] S. Yin, X. Zhu, and C. Jing, "Fault detection based on a robust one class support vector machine," *Neurocomputing*, vol. 145, pp. 263–268, Dec. 2014.
[40] S. Luca, D. A. Clifton, and B. Vanrumste, "One-class classification of point patterns of extremes," *J Mach. Learn. Res.*, vol. 17, no. 1, pp. 6581–6601, 2016.
[41] Z. Y. Han and J. Wang, "Fault diagnosis algorithm based on imbalanced support vector data description," *Comput. Eng.*, vol. 34, no. 5, pp. 156–162, 2017.
[42] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
[43] L. M. Manevitz and M. Yousef, "One–class SVMs for document classification," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 139–154, 2002.
[44] S. Maldonado and C. Montecinos, "Robust classification of imbalanced data using one-class and two-class SVM-based multiclassifiers," *Intell. Data Anal.*, vol. 18, no. 1, pp. 95–112, Jan. 2014.
[45] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[46] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
[47] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
[48] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, Aug. 1996.
[49] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," *Pattern Anal. Appl.*, vol. 6, no. 3, pp. 245–256, Dec. 2003.
[50] C. Li, "Classifying imbalanced data using a bagging ensemble variation (BEV)," in *Proc. 45th Annu. Southeast Regional Conf. (ACM-SE)*, vol. 45, 2007, pp. 203–208.
[51] S. Hido, H. Kashima, and Y. Takahashi, "Roughly balanced bagging for imbalanced data," *Stat. Anal. Data Mining*, vol. 2, nos. 5–6, pp. 412–426, Dec. 2009.
[52] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
[53] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, Mar. 2005.
[54] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 325–332.
[55] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[56] G. Wu and E. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Proc. Workshop Learn. Imbalanced Data Sets II*, Washington, DC, USA, 2003, pp. 49–56.
[57] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 1–6, Jun. 2004.
[58] X.-W. Chen and M. Wasikowski, "FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 124–133.
[59] M. Wozniak, *Hybrid Classifiers: Methods of Data, Knowledge, and Classifier Combination*, vol. 519. Springer, 2013.
[60] Q. Cao and S. Z. Wang, "Applying over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning," *Int. Joint Conf. Neural Netw.*, vol. 2, pp. 543–548, Nov. 2011.
[61] W. Y. Loh, "Classification and regression trees," *Data Mining Knowl.*, vol. 1, no. 1, pp. 14–23, 2011.
[62] P. Melville, F. Provost, M. Saar-Tsechansky, and R. Mooney, "Economical active feature-value acquisition through expected utility estimation," in *Proc. 1st Int. Workshop Utility-Based Data Mining (UBDM)*, 2005, pp. 10–16.

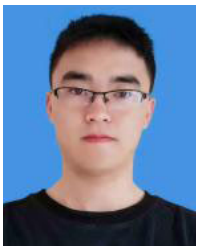
[63] S. Zhang, "Cost-sensitive classification with respect to waiting cost," *Knowl.-Based Syst.*, vol. 23, no. 5, pp. 369–378, Jul. 2010.
[64] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. AI*, 1999, pp. 55–60.
[65] X. Tao, Q. Li, W. Guo, C. Ren, C. Li, R. Liu, and J. Zou, "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification," *Inf. Sci.*, vol. 487, pp. 31–56, Jun. 2019.
[66] R. Ping, S. S. Zhou, and D. Li, "Cost sensitive random forest classification algorithm for highly unbalanced data," *Pattern Recognit. Artif. Intell.*, vol. 33, no. 3, pp. 249–257, 2020.
[67] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.
[68] Y.-J. Cui, S. Davis, C.-K. Cheng, and X. Bai, "A study of sample size with neural network," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2004, pp. 3444–3448.
[69] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[70] S. Y. Liu, Y. Zhai, and D. S. Liu, "Cross-project software defect prediction with multi-strategy feature filtering," *Comput. Eng. Appl.*, vol. 5, no. 8, pp. 53–581, 2019.
[71] Y. G. Sun, "Classification for imbalanced data streams based on cost-sensitive," *J. Xinyang Normal Univ., Natural Sci. Ed.*, vol. 32, no. 4, pp. 670–674, 2019.
[72] J. Chen, S. Liu, W. Liu, X. Chen, Q. Gu, and D. Chen, "A two-stage data preprocessing approach for software fault prediction," in *Proc. 8th Int. Conf. Softw. Secur. Rel.*, San Francisco, CA, USA, Jun. 2014, pp. 20–29.
[73] P. van der Putten and M. van Someren, "A bias-variance analysis of a real world learning problem: The CoIL challenge 2000," *Mach. Learn.*, vol. 57, nos. 1–2, pp. 177–195, Oct. 2004.
[74] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
[75] X.-W. Chen and J. C. Jeong, "Minimum reference set based feature selection for small sample classifications," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 153–160.
[76] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, no. 10, pp. 1205–1224, 2004.
[77] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.
[78] J. Nam, S. J. Pan, and S. Kim, "Transfer defect learning," in *Proc. 35th Int. Conf. Softw. Eng. (ICSE)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2013, pp. 382–391.
[79] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[80] Q. Yu, S.-J. Jiang, R.-C. Wang, and H.-Y. Wang, "A feature selection approach based on a similarity measure for software defect prediction," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 11, pp. 1744–1753, Nov. 2017.
[81] L. Yu and A. Mishra, "Experience in predicting fault-prone software modules using complexity metrics," *Qual. Technol. Quant. Manage.*, vol. 9, no. 4, pp. 421–434, Jan. 2012.
[82] K. Gao, T. M. Khoshgoftaar, H. Wang, and N. Seliya, "Choosing software metrics for defect prediction: An investigation on feature selection techniques," *Softw., Pract. Exper.*, vol. 41, no. 5, pp. 579–606, Apr. 2011.
[83] I. H. Laradji, M. Alshayeb, and L. Ghouti, "Software defect prediction using ensemble learning on selected features," *Inf. Softw. Technol.*, vol. 58, pp. 388–402, Feb. 2015.
[84] X. H. Zhou, "Research on imbalanced data classification based on neural network adversarial and ensemble," Zhejiang Univ. Technol., Hangzhou, China, Tech. Rep., 2019, doi: 10.27463/d.cnki.gzgyu.2019.000386.
[85] X. B. Xie, "Research on imbalanced dataset classification based on generative adversarial networks," Nanjing Univ. Posts Telecommun., Tech. Rep., 2020, doi: 10.27251/d.cnki.gnjdc.2019.000673.
[86] T. Konno and M. Iwazume, "Pseudo-feature generation for imbalanced data analysis in deep learning," *CoRR*, vol. abs/1807.06538, 2018.
[87] Z. Chen and W. Guo, "Text classification based on depth learning on unbalanced data," *J. Chin. Comput. Syst.*, vol. 41, no. 1, pp. 1–5, 2020.
[88] F. Verh and S. Chawla, "Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 679–684.
[89] B. Liu, W. Hsu, Y. Ma, A. A. Freitas, and J. Li, "Integrating classification and association rule mining," in *Proc. 4th KDD*, 1998, pp. 80–86.
[90] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *Proc. IEEE Int. Conf. Data Mining*, Washington, DC, USA, Nov. 2001, pp. 369–376.
[91] W. Liu, S. Chawla, D. A. Cieslak, and N. V. Chawla, "A robust decision tree algorithm for imbalanced data sets," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2010, pp. 766–777.
[92] S. Ren, W. Zhu, B. Liao, Z. Li, P. Wang, K. Li, M. Chen, and Z. Li, "Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning," *Knowl.-Based Syst.*, vol. 163, pp. 705–722, Jan. 2019.
[93] X. Tao, Q. Li, C. Ren, W. Guo, Q. He, R. Liu, and J. Zou, "Affinity and class probability-based fuzzy support vector machine for imbalanced data sets," *Neural Netw.*, vol. 122, pp. 289–307, Feb. 2020.
[94] Z. Yang and H. J. Wang, "Improved density peak clustering algorithm based on weighted K-nearest neighbor," *Appl. Res. Comput.*, vol. 37, no. 3, pp. 667–671, 2020.
[95] T. Bikku, N. S. Rao, and A. R. Akepogu, "A novel multi-class ensemble model based on feature selection using Hadoop framework for classifying imbalanced biomedical data," *Int. J. Bus. Intell. Data Mining*, vol. 14, nos. 1–2, pp. 25–39, 2019.
[96] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007.
[97] H. M. Li, L. D. Wang, and S. K. Duan, "Improved over-limit learning machine and its application in unbalanced data," *J. Southwest Univ., Natural Sci. Ed.*, vol. 42, no. 6, pp. 140–148, 2020.
[98] Q. S. Sun, "Research on weighted extreme learning machine algorithm based on imbalanced data distribution," Xiangtan Univ., Xiangtan, China, Tech. Rep., 2019, doi: 10.27426/d.cnki.gxtdu.2019.001093.
[99] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.
[100] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-NN performance in a challenging scenario of imbalance and overlapping," *Pattern Anal. Appl.*, vol. 11, nos. 3–4, pp. 269–280, Sep. 2008.
[101] J. A. Sáez, J. Luengo, and F. Herrera, "Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification," *Pattern Recognit.*, vol. 46, no. 1, pp. 355–364, Jan. 2013.
[102] J. Swets, R. Dawes, and J. Monahan, "Better decisions through science," *Sci. Amer.*, vol. 283, pp. 82–87, Oct. 2000.
[103] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 179–186.

LE WANG received the bachelor's degree from North Minzu University, in 2018, where she is currently pursuing the master's degree. Her research interests include data mining, machine learning, and ensemble classification of data stream.

MENG HAN received the Ph.D. degree from Beijing Jiaotong University. She is currently an Associate Professor and a Master's Supervisor with North Minzu University. Her research interest includes data mining.

NI ZHANG received the bachelor's degree from Changzhi University, in 2019. She is currently pursuing the master's degree with North Minzu University. Her research interests include data mining, and high utility pattern mining.

XIAOJUAN LI received the bachelor's degree from North Minzu University, in 2018, where she is currently pursuing the master's degree. Her research interests include data mining, and ensemble classification of data streams.



HAODONG CHENG received the bachelor's degree from the Shandong Youth University of Political Science, in 2019. He is currently pursuing the master's degree with North Minzu University. His research interests include data mining and high-utility pattern mining.