Research Article

Jianhua Zhao* and Ning Liu

Semi-supervised Classification Based Mixed Sampling for Imbalanced Data

https://doi.org/10.1515/phys-2019-0103
Received Oct 28, 2019; accepted Nov 15, 2019

Abstract: In practical application, there are a large amount of imbalanced data containing only a small number of labeled data. In order to improve the classification performance of this kind of problem, this paper proposes a semi-supervised learning algorithm based on mixed sampling for imbalanced data classification (S2MAID), which combines semi-supervised learning, over sampling, under sampling and ensemble learning. Firstly, a kind of under sampling algorithm UD-density is provided to select samples with high information content from majority class set for semi-supervised learning. Secondly, a safe supervised-learning method is used to mark unlabeled sample and expand the labeled sample. Thirdly, a kind of over sampling algorithm SMOTE-density is provided to make the imbalanced data set become balance set. Fourthly, an ensemble technology is used to generate a strong classifier. Finally, the experiment is carried out on imbalanced data with containing only a few labeled samples, and semi-supervised learning process is simulated. The proposed S2MAID is verified and the experimental result shows that the proposed S2MAID has a better classification performance.

Keywords: semi-supervised learning; imbalanced data; over sampling; under sampling; ensemble learning

PACS: 89.20.Ff, 89.75.Kd, 89.70.Cf

1 Introduction

The imbalanced data classification is such a problem where the class distribution of training data is not balanced and the number of one class is far less than the other one. It exists widely in real life, such as genetic testing, bad bank loans, fault diagnosis, etc. [1, 2]

In practical application, people are more concerned about the information of the minority class data, and the cost of classification errors of minority classes is much higher than that of majority classes. For example, if a cancer patient is diagnosed as normal, it will delay the optimal timing of treatment, resulting in life-threatening for patients; in network intrusion detection, if the network intrusion behavior is sentenced to normal behavior, it will have the potential danger to cause major network security incidents. Therefore, it is necessary to improve the accuracy of minority classes and it has become an urgent problem in machine learning [3].

Researchers have proposed several different solutions to the classification of imbalanced data. At present, the method for imbalanced data classification can be divided into two levels: one is from the algorithm level [4, 5], the other is from the data processing level [6, 7].

At the algorithm level, the imbalanced data classification method mainly includes ensemble method [4], cost sensitive learning [5], and so on. At the data processing level, it includes over sampling and under sampling, improving the imbalanced data set by some mechanism to obtain a balanced data distribution [6, 7].

At present, there are many imbalanced data classification methods [8–10], most of them are based on supervised learning. However, supervised learning often needs a large amount of labeled samples, and it may take a lot of manpower and material resources to obtain labeled samples. Therefore, it is necessary to use semi-supervised learning for this classification problem. At present, there are also many semi-supervised learning methods [11–13], but most of them assume that the data set is balanced. Therefore, it is very important to design a kinds of semi-supervised algorithm for imbalanced data classification.

For semi-supervised classification of imbalanced data, many researchers have been engaged in the research in this field and have achieved a lot of research results. Li *et al.* [14] proposed a semi-supervised classification method to alleviate the adverse effects of imbalances. This method iteratively selected some unlabeled samples and

***Corresponding Author: Jianhua Zhao:** College of Mathematics and Computer Application, Shangluo University, Shangluo 726000, China; Email: zhaojh2009@aliyun.com
Ning Liu: College of Economics Management, Shangluo University, Shangluo 726000, China

added them to a few classes to form a balanced data set. Pan *et al.* [15] proposed a algorithm based on integrated framework for imbalanced noise graph flow classification. Frasca *et al.* [16] proposed a cost-sensitive neural network algorithm for imbalanced data. Hajizadeh *et al.* [17] proposed a semi-supervised imbalanced image data detection method. Li *et al.* [18] proposed a new label matrix normalization solution to solve the general equilibrium problem. Limin *et al.* [19] proposed a semi-supervised algorithm based on evidence theory and biased-SVM for imbalance data sets which had a number of unlabeled samples.

Although there are some semi-supervised classification methods for imbalanced data, there is no unified opinion about which method is the best to deal with imbalanced data. With the development of society, the social environment becomes more and more complex, and many problems in the real world may become more and more complex, which makes it possible to collect useful labels. The research on classification algorithm for imbalanced data with fewer labeled samples has certain theoretical significance and good practical value.

To solve the classification problem of imbalanced data containing only a few labeled samples, we proposed a semi-supervised learning algorithm based mixed sampling for imbalanced data classification (S2MAID), which combines semi-supervised learning, over sampling, under sampling and ensemble learning. First, the traditional under sampling algorithm is improved to UD-density to select samples with high information content from majority class set, providing data sets for semi-supervised learning. Second, a safe supervised-learning method is used to mark the unlabeled sample and expand the labeled sample. Then, the traditional over sampling algorithm is improved to SMOTE-density to turn the imbalanced data set into balance set. Finally, ensemble technology is used to generate a strong classifier to improve classifier‘s performance. The experiment is carried out on imbalanced data and semi-supervised learning process is simulated to verify the proposed algorithm.

The first part of the paper is the introduction; the second part is the related work, which presents semi-supervised learning, over sampling, under sampling, ensemble technology; the third part is the introduction of our proposed algorithm, including US-density, SMOTE-density and S2MAID; the fourth part is the experiment; the fifth part is acknowledgement.

2 Related works

2.1 Semi-supervised learning

Accompanied by the rapid development of data acquisition and storage technology, it is easier to obtain unlabeled data, but it is relatively difficult to obtain labeled samples because of the need to consume a certain amount of manpower and material resources. Semi-supervised learning [20, 21] is a kind of new method between supervised learning and unsupervised learning, whose purpose is to make full use of large unlabeled samples to make up for the lack of labeled samples. Semi-supervised learning is divided into semi-supervised clustering and classification. The main aim of the latter is to study how to use unlabeled samples to help to train supervised learning classifier, when labeled samples are insufficient. Semi-supervised classification mainly includes disagreement-based method, generative method, discriminative method and graph-based method [20, 21].

The disagreement-based semi-supervised classification realizes the utilization of unlabeled data by using multiple classifiers. In the process of machine learning, the unlabeled data is used as a platform for interaction between multiple classifiers. The original disagreement-based algorithm was developed by Blum and Mitchell [22] in 1998. They assumed that the data set had two views of sufficient redundancy, meeting the following conditions: First, each set of attributes was sufficient to describe the problem; second, each attribute set was conditioned to be independent of another set of attributes when it was marked.

The generative method [23] assumes that the sample and class labels are generated by a set of probability distributions of a certain or certain structural relationship. From these distributions, the sample L with the class label and the sample U without the class label are generated.

The discriminative method [24] uses the maximum interval algorithm to train the learning decision boundary of the labeled sample and unlabeled sample. The purpose of learning is to make the classification hyperplane through the low density data region, and to make the distance maximum between the classification hyperplane and the nearest sample.

Graph-based learning [25] is a very active direction of semi-supervised learning in recent years. The essence of the graph based approach is the label propagation.

2.2 SMOTE algorithm [26]

SMOTE is a kind of over sampling method by changing the balance of the data set. Its purpose is to maintain a balance between the number of majority classes and minority classes, by increasing the number of minority classes.

In SMOTE, it searches for the nearest K adjacent samples in each data sample x of minority class data set and randomly selects N samples in the nearest neighbor data set recorded as $y_1, y_2, y_3, \ldots y_n$. The random linear interpolation operation is carried out between the minority class data x and $y_i (j = 1, 2, N)$ to construct a new sample z_j. The interpolation operation is shown in the Formula (1):

$$z_j = x + randN(0, 1) * (y_j - x), j = 1, 2, \ldots, N \quad (1)$$

Where $randN(0, 1)$ represents a random number, z_j represents new sample, x represents the sample of the minority class, y_j represents the j-th neighbor samples of x, these new synthetic minority class is merged into the original data set to generate new training set.

2.3 Under sampling algorithm [27]

Random under sampling method deletes some samples from the majority class sample set at random, but does not deal with the minority class sample set at all. Because of the randomness and contingency, this method is easy to lose some important information in most classes and affect the classification performance.

Easy Ensemble algorithm randomly extracts multiple subsets from the majority class sample set, then uses each subset and the minority class sample set to form a training set to train a classifier, and finally combines the classification results of multiple classifiers.

Balance Cascade is based on supervised learning combined with Boosting. In the n-round training, the subset sampled from the majority class sample set and the minority class sample set is combined to train a basic learner H. After training, the samples in the majority class sample set that can be correctly classified by H will be eliminated. In the next n+1 round, the classifier is trained by combining a subset from the rejected majority samples with the minority class sample set. Finally, different basic learners are integrated.

KNN-NearMiss is essentially a prototype selection method, which is to select the most representative samples from the majority class sample set for training, mainly to alleviate the problem of information loss in random undersampling.

2.4 Ensemble learning [28]

Ensemble learning gets a number of different based classifiers by using a simple classification algorithm to train data, then it combines the base classifiers into a strong classifier in some way. Ensemble learning plays an important role in the field of machine learning.

With the development of the integrated learning technology, more and more researchers introduce ensemble learning into the classification of imbalanced data, and get a lot of research results.

Galar *et al.* [29] develop a new ensemble construction algorithm, which combines random under sampling with Boosting algorithm. Khoshgoftaar *et al.* [30] compares the performance of boosting and bagging techniques from imbalanced and noisy binary-class data. Ghazikhani *et al.* [31] propose an a two-layer approach online ensemble of neural network classifiers to handle class imbalance and non-stationarity.

The combination of sampling technology and ensemble learning is an effective method to solve the problem of imbalanced data classification [29]. However, the existing algorithms are often unable to combine the advantages of the two methods effectively. For example, the traditional over sampling technique blurred the boundaries of the majority and the minority; the traditional over sampling technology leads to a large scale of data as well as low classification efficiency; It is also possible to cause some valuable data to be lost after processing the imbalanced data sets by using the under sample. In addition, the choice of integration algorithm often affects the classification accuracy of the algorithm.

3 Our Method

In this paper, we proposes a semi-supervised learning algorithm based on mixed sampling for imbalanced data classification (S2MAID), which combines semi-supervised learning, over sampling, under sampling and ensemble learning.

In S2MAID, the data set is divided into two types: the majority class sample set and the minority class sample set. In the majority class sample set there are a lot of labeled samples and a few of unlabeled samples, the same is in the minority class sample set.

In the following content, the main idea of S2MAID is firstly introduced. Then, the improved algorithm UD-density and SMOTE-density is introduced.

3.1 The basic idea of S2MAID

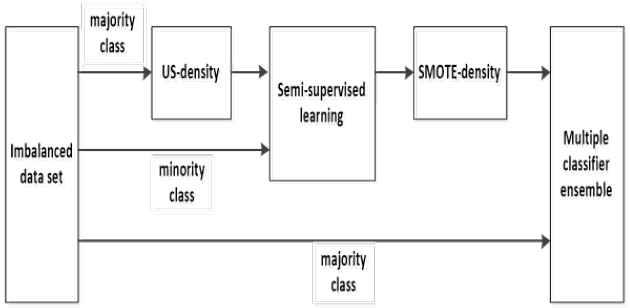

Figure 1: Algorithm framework diagram

In S2MAID, a safe semi-supervised learning algorithm and an improved under sampling method UD-density are used to mark imbalanced unlabeled samples, and expand the number of labeled samples to form a labeled sample set. The improved SMOTE algorithm named SMOTE-density is used to make the imbalanced set become balanced set. Then the ensemble learning method is used to predict the label sample set after sampling. The Framework diagram of S2MAID is shown in Figure 1. The algorithm consists of three parts: data sampling, semi-supervised learning and ensemble learning. First, an improved under sampling method US-density is proposed to sample from majority class set. Second, a safe semi-supervised learning is used to mark the unlabeled data from minority class and the data set sampled by US-density. Next, an over sampling method SMOTE-density is proposed to make the minority class set become balanced. Finally, multiple classifier ensemble method is used to generate the final classifier. The implementation of each step is described in detail below.

3.2 The under sampling process

In the semi-supervised imbalance problem, the number of labeled samples is small, and the classification is imbalanced. In order to improve the accuracy, an under-sampling method US-density is added to the semi-supervised learning algorithm. By deleting the samples with low information content in the training set and retaining a small number of samples with high information content, the balance of the two types of samples is maintained and the accuracy of pre-labeled samples is improved.

For the majority sample set, an under-sampling method based on data density distribution (US-density) is proposed here. Firstly, the labeled samples in the majority sample set are retained. Secondly, for the unlabeled samples in the majority sample set, the data sets are divided into high-density data clusters and low-density data clusters. Different strategies of resampling are applied to different density data clusters to adjust the balance of data sets. Samples in high-density data clusters should be retained as much as possible in the process of under-sampling, and samples in low-density data clusters should be deleted in under-sampling. Finally, the data set is composed of high density unlabeled samples in the majority sample set, the labeled samples in the majority sample set and the minority sample set for semi-supervised learning to mark.

First step, the K-nearest neighbor method is used to classify the majority sample set into noise set, boundary set and safety set. Second step, the point P is selected from the boundary sample set. A circle with P as its center and r as its radius is selected. N denotes the number of samples contained in the circle. If the total sample in the circle is larger than N/2, delete the sample point, otherwise retain the sample point. The description of US-density is shown in Algorithm 1.

Algorithm 1: US-density

```
Input: The majority sample set M; Threshold N.
Output: A new sample set R
    Progress:
1: The k-nearest neighbor method (k=5) is used to
   classify the majority sample set into noise set,
   boundary set and safety set.
2: Delete the noise samples and keep the safe
   samples.
3: A point P is selected from the boundary sample
   set.
4: A circle with P as its center and r as its radius is
   selected, the number of samples in the circle is N
   (here N=10).
5: If the total sample in the circle is larger than N/2,
   delete the sample point, otherwise retain the
   sample point.
6: Repeat step 5 until the set number of samples is
   reached.
7: All retained samples constitute a set R.
```

3.3 Semi-supervised learning process

Here, a safe semi-supervised classification algorithm is used to mark the unlabeled sample and expand the labeled set. The data set for semi-supervised learning is composed of data set R selected from Algorithm 1, the labeled samples in the majority set M and the minority set S. The algorithm description is shown in Algorithm 2 [32]. The main idea of the Algorithm 2 is as follows. First, prelabeling is carried out for unlabeled sample. Then, the sample with pseudo-label is added into the labeled sample set to carry out grouping, training and testing in the labeled set, and the corresponding errors of various pseudo-labels are calculated. At last, the pseudo-label with the lowest classification rate is selected as the candidate label. The loop is executed repeatedly until certain conditions are met.

Algorithm 2: safe semi-supervised classification

```
Input: Labeled set L; Unlabeled set U; Supervised
    classifier.
Output: The candidate label $e_x$ of the unlabeled
    sample $x$ after the prediction
    Progress:
1: While (U not empty)
2: For i=1 to $m$ /* m presents the number of
    categories*/
3: Use pseudo-label $M_i$ to mark x in U, record as $x_i$
4: Add $x_i$ into L to form training set
5: Group L into $k$ groups, record as L(1), L(2)...,L($k$)
6: For $j$=1 to $k$
7: Take L($j$) as validation set, the other $k$-1groups as
    training set
8: Using a training set to train a supervised classifier
    $C_{ij}$
9: The supervised classifier $C_{ij}$ is used to test the
    validation set
10: Record the verifying accuracy at this time as
    acc$_i$ ($j$)
11: EndFor
12: Calculate the average accuracy:
    acc$_i$=(acc$_i$(1)+acc$_i$(2)+...+acc$_i$ ($k$))/$k$
13: Calculate error $e_i$=1- acc$_i$
14: EndFor
15: Get the minimum value $e_x$ in $e_1, e_2, \dots, e_m$
16: $e_x$ is used as the candidate predictive value of the
    sample $x$
17: L=LU$\{e_x\}$
18: Repeat step 1
```

3.4 SMOTE-density

The expanded sample set after semi-supervised learning above is still an imbalanced data set, it is necessary to use oversampling algorithm to make data set balanced [33, 34]. Here an algorithm SMOTE-density is proposed to generate synthetic sample in the minority sample set.

SMOTE-density is improved on the basis of SMOTE algorithm. SMOTE-density algorithm identifies sparse samples based on the density of samples and selects high density samples as seed samples [35]. Then, the method of SMOTE is used to generate synthetic samples between the seed samples and their k-nearest neighbors.

The specific process of the algorithm is as follows:

3.4.1 Determining the neighborhood radius

Suppose x is a sample in the minority sample set, parameter r is the neighborhood radius of sample x, then neighborhood of the sample x is a space, in which x is the circle center and r is the radius.

After experimental comparison, the average distance among all samples in the minority sample set is chose as the neighborhood radius. The calculation formula is as follows.

$$\delta = \frac{m * (m-1)}{2} \sum\nolimits_{i,j=1}^{m} D(x_i, x_j) \qquad (2)$$

Where m represents the number of samples in the minority sample set S, D denotes the Euclidean distance.

3.4.2 Calculate the density of each sample in the minority sample set

The density of δ-neighborhood of sample x is expressed by the number of samples in the δ-neighborhood of sample x. The density threshold of samples is expressed by the mean value of the δ-neighborhood density of all samples. The calculation formula of the density threshold is as follows.

$$DT = \frac{1}{m} \sum_{i=1}^{m} Density_i \qquad (3)$$

Where m denotes the number of samples in the minority sample set S, $Density_i$ denotes the density of δ-neighborhood of sample i.

3.4.3 Generating the seed set

If the density of sample x is greater than DT, x is a dense sample. If the density of sample x is less than DT, x is a sparse object. All sparse samples constitute seed set of this kind.

3.4.4 Generating synthetic samples

According to the idea of SMOTE algorithm, SMOTE-density algorithm synthesizes new samples between seed samples and their neighbors according to Formula (4):

$$p = x + randN(0, 1) * D(x, seed) \quad (4)$$

Where p denotes the new synthetic samples, x denotes the sample of the minority class, $r$and$N$ (0, 1) denotes a random number, D (x,seed) denotes the Euclidean distance between x and seed.

3.4.5 Forming new data set of the minority sample

Here, synthetic samples is added into the minority sample set to form a new sample set. The process of generating seed set mainly includes: calculating the neighborhood radius of the samples in the minority class sample set, calculating the density threshold of the class, and generating the sparse object set as a seed sample set [36, 37]. Synthetic sample segmentation generated by DS-SMOTE algorithm is distributed between sparse objects and their neighbors. The description of SMOTE-density is shown in Algorithm 3.

```
Algorithm 3: SMOTE-density
Input: The majority sample set M; The minority sample set S.
Output: The new data set of the minority sample
  Progress:
  Calculating the neighborhood radius according to formula (2).
  Calculate the density threshold DT according to formula (3).
  for i=1 to m
    if densit $y_i < DT$
        SSET=SSET U{x}
    endif
  endfor
  Generating seed set SSET.
  The samples in SSET are over-sampled using SMOTE method according to Formula (4).
  Forming new data sets of the minority sample.
```

3.5 The integration process of classifiers

In this step, an integrated classification algorithm is used to integrate the balanced data sets. The idea is as follows: different weak classifiers are trained and the weight of each sample in the next iteration is determined according to whether the classification of each sample in each training set is correct and the classification error of the sample [38]. Finally, each training classifier is fused together according to a certain weight, and a strong classifier is formed as the final decision classifier.

4 Experiment and result analysis

4.1 Experimental data set and scheme

The experiment platform is based on Intel Core2 Duo CPU 2.0GHz, memory 2.0GB PC, Windows XP operation system and MATLAB R2009b programming environment.

The experiment is carried out on 5 data sets commonly used in the UCI database (http://archive.ics.uci.edu/ml/), which is shown in Table 1.

Table 1: The data set

data set	num	min/max	ratio	feature
ionosphere	351	126/225	1.79	34
satimage	6435	1358/5077	3.74	36
segment	2310	330/1980	6.00	19
glass	214	29/185	6.38	10
yeast	1484	163/1321	8.10	8

In Table 1, **num** denotes the sample size of the data set; **min** denotes the sample number of minority class and **max** denotes the sample number of majority class; **ratio** denotes the ratio of majority class to minority class, and **feature** denotes the feature number of selected data set. Most of the data sets in Table 1 belong to multi-types data set, so it is necessary to convert multi-types data set into two types. Among them, ionosphere data set itself is a two-types data set; for satimage data set, category 3 is used as a minority class, and the rest category of the samples form a majority class; for segment data set, category 1 is used as a minority class, and the rest of the samples form a majority class; for glass data set, category 7 is used as a minority class, and the rest of the samples form a majority class; for yeast data set, category 4 is used as a minority class, and the rest of the samples constitute a majority class.

In the sample selected in Table 1, the proportion of the sample number of the training set and the test set is set to 1:1, that is, the training set and the test set are 50% of the data set in the list. In order to correctly calculate the classification rate, the training set is divided into labeled sample set and unlabeled sample set. In the first kind of experiment, labeled sample is accounted for 5% of training set; in the second kind of experiment, labeled sample is accounted for 10% of training set; in the third kind of experiments, labeled sample is accounted for 15% of training set.

Because of the randomness in the experiment, in order to fully verify the classification effect of the algorithm, each data set is trained 100 times, and the experimental results are averaged. Before the experiment, the discrete value of all the sample attributes is processed numerically. The Formula (5) is used to normalize the input data.

$$x_k = (x_k - x_{\min})/(x_k - x_{\max}) \tag{5}$$

Where x_{min} denotes the smallest value in data set; x_{max} denotes the maximum value in data set; x_k on the right side of the equal sign denotes the input data; x_k on the left side of the equal sign denotes the normalized data.

4.2 Evaluating indicator

In the classification of balanced data, accuracy is an important index to evaluate the performance of classifiers. However, the evaluation index is not reasonable for imbalanced data, because the cost of error classification for accuracy is the same for all samples.

Here, the confusion matrix shown in Table 2 is used as the evaluation index of imbalanced data. And the description of TP, TN, FN, FP is shown below.

- True Positive (TP): Correct judgement. A positive sample is judged to be a positive one.
- True Negative (TN): Correct judgement. A negative sample is judged to be a positive one.
- False Negative (FN): Erroneous judgement. A positive sample is judged to be a negative one.
- False Positive (FP): Erroneous judgement. A negative sample is judged to be a positive one.

The calculation of Precision rate is shown in Formula (6).

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The calculation of Recall rate is shown in Formula (7).

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The calculation of F-value is shown in Formula (8).

$$F - value = \frac{(1 + \lambda^2) \times Recall \times Precision}{\lambda^2 \times Recall + Precision} \tag{8}$$

Where λ denotes the relative importance of Recall and Precision.

The calculation of G-mean is shown in Formula (9).

$$G - mean = \sqrt{\frac{TN}{TN + FP} \times Recall} \tag{9}$$

Here, F-value and G-mean are used to evaluate the performance of the proposed algorithm.

4.3 Experimental data and results analysis

In order to evaluate the performance of the proposed algorithm S2MAID, it is compared with Algorithm 1 [14], Algorithm 2 [16] and Algorithm 3 [19]. Algorithm 1 is represented by A1, Algorithm 2 is represented by A2 and Algorithm 3 is represented by A3.

The experiment is carried out repeatedly, and the average value is calculated. The results are shown from Table 3 to Table 8. Table 3, Table 4 and Table 5 show the F-value of the 4 algorithms, and Table 6, Table 7 and Table 8 show the G-mean of the 4 algorithms. N presents the proportion of labeled samples and N is 5%, 10% and 15% respectively.

From the experimental results, it can be seen that the semi-supervised classification method S2MAID proposed in this paper has higher recognition rate for minority classes and better stability for the whole data set than the other algorithms.

Table 2: Confusion matrix

Category	Actual positive class	Actual negative class
Experimental positive class	TP	FN
Experimental negative class	FP	TN

Table 3: Comparisons of F-value for minority class data (N=5%)

data set	A1	A2	A3	S2MAID
ionosphere	0.7938	0.7754	0.8034	0.8152
satimage	0.8247	0.8175	0.8421	0.8547
segment	0.9018	0.8745	0.9325	0.9541
glass	0.8217	0.8101	0.8210	0.8224
letter	0.4015	0.4715	0.5184	0.5247

Table 4: Comparisons of F-value for minority class data (N=10%)

data set	A1	A2	A3	S2MAID
ionosphere	0.8117	0.7928	0.8457	0.8517
satimage	0.8280	0.8314	0.8601	0.8684
segment	0.9241	0.8987	0.9424	0.9674
glass	0.8301	0.8178	0.8324	0.8401
letter	0.4587	0.4940	0.5207	0.5598

Table 5: Comparisons of F-value for minority class data (N=15%)

data set	A1	A2	A3	S2MAID
ionosphere	0.8280	0.8024	0.8737	0.8920
satimage	0.8298	0.8274	0.8587	0.8724
segment	0.9455	0.9541	0.9657	0.9742
glass	0.8301	0.8078	0.8418	0.8485
letter	0.4530	0.5235	0.5207	0.5675

Table 6: Comparison of G-mean value for the whole data (N=5%)

data set	A1	A2	A3	S2MAID
ionosphere	0.7987	0.7954	0.8278	0.8314
satimage	0.9024	0.9028	0.9232	0.9374
segment	0.9510	0.9474	0.9540	0.9754
glass	0.8017	0.8217	0.8457	0.8654
letter	0.6274	0.5872	0.7012	0.7512

Table 7: Comparison of G-mean value for the whole data (N=10%)

data set	A1	A2	A3	S2MAID
ionosphere	0.8025	0.8274	0.8421	0.8578
satimage	0.9134	0.9154	0.9417	0.9574
segment	0.9478	0.9325	0.9587	0.9847
glass	0.8357	0.8210	0.8354	0.8721
letter	0.6184	0.6017	0.7254	0.7412

Table 8: Comparison of G-mean value for the whole data (N=15%)

data set	A1	A2	A3	S2MAID
ionosphere	0.8025	0.8341	0.88718	0.8945
satimage	0.9077	0.9201	0.9471	0.9618
segment	0.9065	0.9451	0.9641	0.9898
glass	0.8127	0.8045	0.8421	0.8814
letter	05814	0.5987	0.7468	0.7841

That is to say, this proposed algorithm can improve the F-value value of minority classes without reducing the overall G-mean value of the data set. The main reasons are as follows:

1. The improved under sampling technique UD-density is used to reduce the size of the data set based on keeping the distribution of the whole data set and select samples with high information.
2. The improved under sampling technique SMOTE-density is used to increase the number of minority class and balance the data distribution.
3. The safe semi-supervised classification algorithm is used to increase reliable labeled sample, improving the accuracy of classification.
4. The ensemble algorithm integrates several weak classifiers and constructs a strong classifier, which improves the classification performance. Therefore, the proposed algorithm S2MAID can improve the classification performance better than the other ones.

5 Conclusion

In practical application, there are a large number of imbalanced data set with a small number of labeled samples. Such problems are more important but difficult to deal with. In this paper, a semi-supervised learning algorithm for imbalanced data based on UD-density and SMOTE-density is proposed. The experimental results on UCI data set are compared with the existing semi-supervised classification algorithm for imbalanced set. The results show that the proposed algorithm has a higher G-mean value for the whole data set and a higher F-value value for the minority class. Under similar circumstance, it has a better stability.

Acknowledgement: This work was supported by Shangluo Universities Key Disciplines Project, Discipline name: Mathematics; Natural Science Basic Research Plan in Shaanxi Province of China (No.2015JM6347); Science Research Plan of Shangluo University (No.14SKY026); Horizontal Project of Shangluo University (No.2018HXKY056); Science and technology innovation team construction project of Shangluo University (18scx002).

References

[1] Provost F., Fawcett T., Robust classification for imprecise environments, Mach Learn., 2001, 42(3), 203-231.

[2] He H., Garcia E.A., Learning from Imbalanced Data, IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9), 1263-1284.

[3] Maldonado S., López J., Imbalanced data classification using second-order cone programming support vector machines, Pattern Recogn., 2014, 47(5), 2070–2079.

[4] Sun Z., Song Q., Zhu X., Sun H., Xu B., Zhou Y., A novel ensemble method for classifying imbalanced data, Pattern Recogn., 2015, 48(5), 1623-1637.

[5] Castro C.L., Braga A.P., Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data, IEEE Transactions on Neural Networks & Learning Systems, 2013, 24(6), 888-899.

[6] Barua S., Islam M.A., Yao X., Murase K., MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning, IEEE Transactions on Knowledge & Data Engineering, 2014, 26(2), 405-425.

[7] Ng W.W., Hu J., Yeung D.S., Yin S., Roli F., Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems, IEEE T Cybernetics., 2017, 45(11), 2402-2412.

[8] Hong X., Chen S., Harris C.J., A Kernel-Based Two-Class Classifier for Imbalanced Data Sets, IEEE T Neural Networ., 2007, 18(1), 28-41.

[9] Khan S.H., Hayat M., Bennamoun M., Sohel F., Togneri R., Cost Sensitive Learning of Deep Feature Representations from Imbalanced Data, IEEE Transactions on Neural Networks & Learning Systems, 2015, 29(8), 3573-3587.

[10] Gao M., Hong X., Chen S., Harris C.J., A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, Neurocomputing, 2011, 74(17), 3456-3466.

[11] Zhou Z.H., Li M., Tri-training: exploiting unlabeled data using three classifiers, IEEE Transactions on Knowledge & Data Engineering, 2005, 17(11), 1529-1541.

[12] Yu Z., Lu Y., Zhang J., You J., Wong H.S., Wang Y., et al., Progressive Semisupervised Learning of Multiple Classifiers. IEEE T Cybernetics., 2018, 48(2), 689-702.

[13] Forestier G., Cédric W., Semi-supervised learning using multiple clusterings with limited labeled data, Inform Sciences., 2016, 361-362(C), 48-65.

[14] Li F., Yu C., Yang N., Li G., Kaveh-yazdy F., Iterative Nearest Neighborhood Oversampling in Semisupervised Learning from Imbalanced Data, The Scientific World Journal, 2013, 1, 1903-1912.

[15] Pan S., Wu J., Zhu X., Zhang C., Graph ensemble boosting for imbalanced noisy graph stream classification, IEEE T Cybernetics., 2015, 45(5), 954-968.

[16] Frasca M., Bertoni A., Re M., Valentini G., A neural network algorithm for semi-supervised node label learning from unbalanced data, Neural Networks, 2013, 43, 84-98.

[17] Hajizadeh S., Núñez A., Tax D., Semi-supervised Rail Defect Detection from Imbalanced Image Data, IFAC PapersOnLine, 2016, 49(3), 78-83.

[18] Li F., Li G., Yang N., Xia F., Label matrix normalization for semisupervised learning from imbalanced Data, New Rev Hypermedia M., 2014, 20(1), 5-23.

[19] Du L., Xu Y., Semi-supervised classification method for imbalanced data based on evidence theory, Application Research of Computers, 2018, 35(2), 342-345.

[20] Kawakita M., Kanamori T., Semi-supervised learning with density-ratio estimation. Mach Learn., 2013, 91(2), 189-209.

[21] Belkin M., Niyogi P., Semi-Supervised Learning on Riemannian Manifolds, Mach Learn., 2004, 56(1-3), 209-239.

[22] Blum A., Mitchell T., Combining labeled and unlabeled data with co-training, In Conference on Computational Learning Theory. 1998, 92-100.

[23] Jiang Z., Zhang S., Zeng J., A hybrid generative/discriminative method for semi-supervised classification, Knowl-Based Sys., 2013, 37(2), 137-145.

[24] Appice A., Guccione P., Malerba D., Transductive hyperspectral image classification: toward integrating spectral and relational features via an iterative ensemble system, Mach Learn., 2016, 103(3), 343-375.

[25] Zhuang L., Zhou Z., Yin J., Gao S., Lin Z., Ma Y., et al., Label Information Guided Graph Construction for Semi-Supervised Learning, IEEE T Image Process., 2017, 26 (9), 4182-4192.

[26] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 2011, 16(1), 321-357.

[27] Shalizi C.R., Rinaldo A., Consistency under sampling of exponential random graph models, The Annals of Statistics, 2013, 41(2), 508-535.

[28] Liu Y., Yao X., Ensemble learning via negative correlation. Neural Networks, 1999, 12(10), 1399-1404.

[29] Galar M., Fernández A., Barrenechea E., Herrera F., EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recogn., 2013, 46(12), 3460–3471.

[30] Khoshgoftaar T.M., Van Hulse J., Napolitano A., Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data. IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, 2011, 41(3), 552-568.

[31] Ghazikhani A., Monsefi R., Yazdi H.S., Ensemble of online neural networks for non-stationary and imbalanced data streams. Neurocomputing, 2013, 122, 535-544.

[32] Zhao J., Liu N., Malov A., Safe semi-supervised classification algorithm combined with active learning sampling strategy. J Intell Fuzzy Syst., 2018, 35(4), 4001-4010.

[33] Dewasurendra M., Vajravelu K., On the Method of Inverse Mapping for Solutions of Coupled Systems of Nonlinear Differential Equations Arising in Nanofluid Flow, Heat and Mass Transfer. Applied Mathematics & Nonlinear Sciences, 2018, 3, 1-14.

[34] Fernández-Pousa C.R., Perfect Phase-Coded Pulse Trains Generated by Talbot Effect, Applied Mathematics & Nonlinear Sciences, 2018, 3, 23-32.

[35] Gao W., Wang W., A Tight Neighborhood Union Condition on Fractional (G, F, N', M)-Critical Deleted Graphs, Colloq Math-Warsaw., 2017, 149, 291-298.

[36] Gao W., Wang W., New Isolated Toughness Condition for Fractional (G, F, N) - Critical Graph, Colloq Math-Warsaw., 2017, 147, 55-65.

[37] García-Planas M.I., Klymchuk T., Perturbation Analysis of a Matrix Differential Equation $X\doteq ABx$, Applied Mathematics & Nonlinear Sciences, 2018, 3, 97-104.

[38] Lakshminarayana G., Vajravelu K., Sucharitha G., and Sreenadh S., Peristaltic Slip Flow of a Bingham Fluid in an Inclined Porous Conduit with Joule Heating, Applied Mathematics & Nonlinear Sciences, 2018, 3, 41-54.