Multi-class Unbalanced Data Classification for Sleep Staging

Dan Li[1], Wu Huang[2*], Guobiao Xu[3], Tao Zhang[1], Zhonghui Jiang[1], Xiao Wei[1]

[1] Chengdu Techman Software Co., Ltd, Chengdu, China.
[2] Sichuan University, Chengdu, China.
[3] China Civil Aviation Flight University, Chengdu, China.

* Corresponding author. Tel.:13088058715; email: tmezl@126.com
Manuscript submitted December 18, 2019; accepted March 2, 2020.
doi: 10.17706/ijcce.2020.12.2.58-71

Abstract: Unbalanced data classification is a research focus for many applications, including financial fraud detection, network intrusion detection and cancer classification. However, unbalanced data classification is rarely investigated in the field of EEG-based sleep staging. Herein, considering the idea that old methods can be exploited in new applications, we propose a practical framework aiming to classify sleep stages with unbalanced data. In this framework, the data are balanced by using a SMOTE algorithm, in which the mean sample number is used for data expansion and the nearest neighbour number is set according to the G-mean values. Subsequently, the features are extracted and selected based on the balanced dataset. The effectiveness of the proposed framework is validated by testing eight sets of Sleep-EDF EEG data in the MIT-BIH physiological information database. From the results, the proposed framework can be used to not only improve the F-score value of the minority class but also to improve the G-mean value and the AUC value of the whole data set, which might benefit sleep studies and disorder diagnoses.

Key words: Multi-class unbalanced data, SMOTE algorithm, feature selection, sleep staging, SVM.

1. Introduction

In recent decades, many new data forms have emerged. In particular, the widespread existence of unbalanced data has brought great challenges to traditional machine learning algorithms. Although traditional classification algorithms such as the decision tree [1], [3]-[5], support vector machine (SVM) [4], [6]-[8], naive Bayes [2], and the k-nearest neighbours (KNN) algorithm [3], [9] are successful in dealing with many classification problems, most of these algorithms are based on balanced data. When the data are unbalanced, the performance of the classifier will be reduced; in particular, the minority class cannot be recognized correctly. In some practical applications, the significance of correctly recognizing the minority classes is greater than that of the majority classes, such as in cancer symptoms identification, credit card fraud identification and network hacker intrusion [10], [11].

Unbalanced data refers to the unbalanced distribution of classes [12]; that is, the number of one or more classes in such data accounts for a relatively small proportion of the total sample. For this kind of data distribution, the traditional classification algorithm often fails to achieve good classification effects in practical applications. At present, the research on the classification of unbalanced data sets mainly focuses on two aspects, namely, at the data level and the algorithm level [13], [14]. At the data level, the main idea is to change the distribution of the data sets by random under-sampling or random over-sampling, so that

unbalanced data sets tend to become balanced. For example, Chawla *et al*. [15] proposed an over-sampling method of the minority class, referred to as Synthetic Minority Over-sampling Technology (SMOTE), which can make the minority class have more generalization space and avoid the problem of over-fitting to a certain extent. Xiong Bingyan *et al*. [16] proposed an under-sampling method based on sample weight, K-means AdaCost Bagging (KAcBag). This algorithm uses the Bagging algorithm as the framework, the AdaCost weight updating method as the basis, and the K-means algorithm to cluster the data sets for many times and updates the sample weight according to the clustering results. Experiments on 19 sets of data on UCI and a telecom customer switch show that the KAcBag algorithm solves the problem of class imbalance to some extent, but determining the K value adaptively according to the data distribution is still a problem. Cateni *et al*. [17] proposed a similarity-based under-sampling method (SBU). To reduce information loss and balance the data sets, SBU tends to reduce the majority of class entities in the density space. The dissimilarity values of each pair of majority classes are calculated according to the Euclidean distance, and most classes with small dissimilarity values are removed. At the algorithm level, the problem of class imbalance can be solved by modifying existing classification algorithms or putting forward new ones. For example, Liu Xuying *et al*. [18] used cascade model-based classifiers, which make the data tend to balance by gradually sifting out most of the class samples. A series of classifiers trained in this process will classify the predicted samples through integration. Tang *et al*. [19] solved the problem of SVM classification hyperplane migration caused by unbalanced data by using cost-sensitive learning, over-sampling and under-sampling. Literature [20] proposed a new unbalanced algorithm, PBC4cip, which combines the degree of support of the mode under the unbalanced level to weigh the sum of the degree of support of each class and alleviate the impact of the unbalanced data set on the classifier.

Sleep is an essential physiological activity, and its quality affects physical and psychological health, work status and learning efficiency. Sleep staging is an important way to evaluate sleep quality and diagnose sleep-related diseases [21]. Rechtsehaffen and Kales proposed the R&K sleep staging criteria, which divides sleep into 6 stages: awake (stage 0), non-rapid eye movement (NREM), which was divided into stages 1-4, and rapid eye movement (REM). There are great differences among different classes of data in sleep staging; that is, the duration of each sleep cycle stage is different. For example, for normal people, the duration of the wake stage is usually longer than that of the other stages. Therefore, a problem of class unbalance exists in sleep staging, which may lead to inaccurate results and lower reliability of sleep staging analyses.

In this paper, we propose a practical framework aiming to improve the classification accuracy of the minority class in sleep staging. First, according to the sample mean value of the initial data set, each category is divided into large and small classes. Then the total sample numbers of the large class and small classes are counted, the sample mean μ of the large class is taken as the target of the small class, the SMOTE algorithm is used to over-sample the minority class to achieve the μ value, and the K-mean value is used to cluster the majority class to obtain K multi-class clustering centroids consistent with the number of the minority class, Finally, the K-clustering centroids are combined with all minority class samples to form a balanced training set, and the SVM classification algorithm is used to classify the samples.

2. Materials and Methods

2.1. Materials

The data used in this paper are from 8 groups of Sleep-EDF EEG data in the MIT-BIH physiological information database. The eight subjects included men and women aged 21 to 35 years. All recorded data include EEG, EOG and EMG signals with a sampling rate of 100 Hz. Each recorded datum is accompanied by an annotation which indicates the conclusion of artificial staging by experienced doctors according to the R&K rule using 30 s as the unit. This conclusion serves as a reference standard for studying the

characteristics of each sleep stage. Table 1 presents the sleep expert classification results of the experimental data set in this paper, where W is the awake stage, REM is the rapid eye movement stage, and N1, N2, N3, and N4 are the non-rapid eye movement stages. N1 and N2 are light sleep, and N3 and N4 are deep sleep. The EEG data of Fpz-Cz and Pz-Oz are used in this paper.

Table 1. Sleep Expert Statistical Results

Statistic	Category/R&K Standard						Total
	W	N1	N2	N3	N4	REM	
data/30s	8055	604	4754	672	633	1609	16327
ratio/%	49.34	3.7	29.12	4.12	3.88	9.85	100

As shown in Table 1, the experimental data are unbalanced class data sets according to various criteria; for example, the proportion of the N1 stage is at least 3.7%, showing a serious class unbalance. Therefore, to improve the accuracy for the minority class, the data needs to be balanced. After balancing, the features are then extracted and filtered by the following three steps: a set of optimal feature subsets are obtained, trained and tested for SVM to obtain the classification results.

2.2. Methods

2.2.1. Data balance processing

For unbalanced data sets, if only minority classes are randomly over-sampled or under-sampled, there will be problems such as model over-fitting or data information loss. Therefore, the framework proposed in this paper is a basis for improvement at the data level.

First, the data is divided into a training set and a test set according to 8:2 ratio, and then the data of the training set is processed. The processing method is as follows: according to the sample mean of the data set, each category is divided into a large class and a small class set. For example, the sample mean is N/n, where N is the number of samples and n is the number of classes. If the number of samples belonging to a class is greater than the mean, this class is considered to be a large class, otherwise it is considered to be a small class. Then, the total sample numbers of the large and small sets are counted separately, and the sample mean value of the larger set is taken as the target of the smaller set. The SMOTE algorithm is used to expand the small sample to set A, extract the small class number M of set A, cluster the large class with M as the number of clusters, and obtain the K clustering properties. Then, the K cluster centroids are combined with the small samples to form a new balanced training set. The pseudocode of the algorithm is as follows:

Input: Initial complete data set Q

Output: Balanced training set D_{Train}

1) Q=pre-process();

2) Train [D]= 80% samples were randomly selected from data set Q, and the remaining 20% samples from test [D]= Q.

3) For training data set [D], if the sample number is N and the category number is n, the sample mean value is N/n, the larger class is T1, and the smaller class is T2. The sample numbers of set T1 and T2 are counted separately, and the sample mean value of the larger group, μ , is the target of the smaller group.

4) Balance the data according to T1, T2, μ :

a) According to the μ value, use the SMOTE algorithm to expand the small class samples in T2 to obtain set A;

b) Extract the small class and large class sample set L1 and L2, respectively, in A, and calculate the quantity M of the L1 small class, so that K=M;

c) Perform K-means clustering on samples of the set L2, and obtain the K disjoint subsets and their cluster centroids;

d) Take out the K cluster centroids and record them as set L2';

5) D_{train}=L1∪L2'.

The parameters of the traditional SMOTE algorithm are changed when data balancing is complete. In the past, the SMOTE algorithm was used to expand the data according to the over-sampling rate N; in this paper, the sample mean of the large class was taken as the target of the small class and the sample was then expanded. Additionally, for a parameter K in SMOTE, that is, the number of nearest neighbours, nine numbers have been tested in this paper: 1, 3, 5, 7, 9, 11, 13, 15, and 17. To avoid the dependence on results from data, the final results are averaged five times. By comparing the G-mean values to determine the best K values, the results of different K values of the balanced data set are shown in Fig. 1 below. The horizontal axis represents the KNN number, and the vertical axis identifies the average G-mean value. The graph shows that when k = 11, the G-mean value is the highest. Therefore, in the following experiments, the K value of the SMOTE algorithm is chosen to be 11.

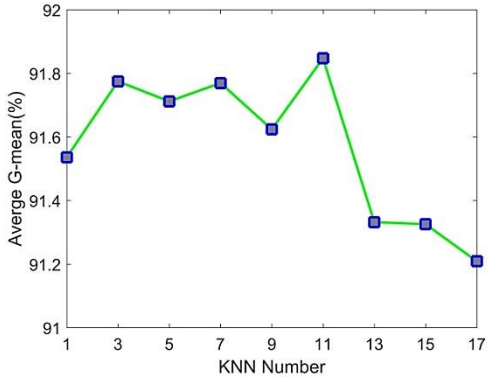


Fig. 1. Classification of the results of different K values on balanced datasets.

Since the original data has noise, invalid data, and non-uniform dimensions between the data, in the initial step, pre-process (), the data needs to be pre-processed. The specific steps are as follows. First, the signal is filtered with six layers of the 'db4' wavelet basis to remove noise. Second, invalid data can be removed by certain rules. For example, a threshold can be set for duplicate data and any data whose similarity is greater than the threshold can be then removed. Incomplete data can be completed by the KNN algorithm. Erroneous data outliers can be eliminated by clustering, regression, sub-boxes and other means, and outliers can also be removed by data distribution characteristics. Finally, the non-uniform dimension is eliminated by normalization. Min-max standardization is commonly used in normalization methods, and the results are normalized to [0,1].

2.2.2. Feature extraction

According to the expert staging results, a segment of data is intercepted every 30 seconds for feature extraction. First, two original EEG signals are superimposed, then 25 features related to sleep staging are extracted from three classes: the time domain, the frequency domain and non-linearity (see Table 2).

Table 2. Selection of Sleep Stage Features

No.	Feature Classes	Features
1	time domain	Minimum, maximum, range, median, standard deviation, coefficient of variation, skewness, kurtosis, first quartile
2	frequency	SEF50 (8~16 Hz), SEF50 (0.5~30 Hz), SEF95 (8~16 Hz), SEF95

	domain	(0.5~30 Hz), SEF95-SEF50 (8~16 Hz), SEF95-SEF50 (0.5~30 Hz), SEF95-SEF50 (0.5-12 Hz), SEF50 (0.5-12 Hz), δ Wave, θ Wave, α Wave, β Wave, spectral peak
3	non-linearity	information entropy, zero-crossing rate, symbol entropy

The above 25 features show different aspects of the features of sleep stages. There are corresponding relationships between the energy ratios of the waves of various features and sleep stages as follows [22]: α and β waves occur in the W phase and REM phase, so α or β waves can distinguish between the W phase and REM phase and NREM phase; the δ wave only occurs in the deep sleep phase, the θ wave occurs in the light sleep phase and REM phase, so the δ wave or θ wave can distinguish between deep sleep and light sleep. Sleep stage entropy is high in the W stage, decreases gradually with the deepening of sleep, and increases again in the REM stage.

2.2.3. Feature selection

In feature selection, a set of optimal feature subsets is selected from the original feature set based on some evaluation criteria. The purpose of feature selection is to select a set of minimal feature subsets according to the given criteria, so that the classification accuracy is not better or worse than the original. Moreover, feature selection can reduce the redundancy between features and remove some irrelevant attributes, making the description of the data sets more accurate and the final model smaller and easier to understand. Therefore, this paper makes feature selection on the basis of balanced data sets using the following three methods.

2.2.3.1. ReliefF algorithm

The Relief algorithm [23] is a single feature optimal strategy algorithm for two classifications. The basic idea is to assign different weights according to the correlation between each feature and category, and features with weights smaller than a certain threshold will be removed. The ReliefF algorithm is an extension of the Relief algorithm that can be used to handle multi-category problems [24]. The ReliefF algorithm randomly extracts one sample x_i from the training sample set at a time, then determines the k nearest neighbour samples NH_i of x_i from the same samples as x_i and finds the k nearest neighbour samples NM_i from each sample of different classes with x_i, and then updates the weight of each feature according to the following rules [24]:

$$w(j) = w(j) + \sum_{c \neq class(x_i)} \frac{\left(\frac{p(c)}{1 - p(class(x_i))} \sum_{j=1}^{k} d(x_i(j), NM_i(j))\right)}{mk} - \sum_{j=1}^{k} \frac{d(x_i(j), NH_i(j))}{mk} \quad (1)$$

In the formula, $x_i(j)$ denotes the value of sample x_i with respect to feature j, $d(\bullet)$ denotes the distance function for calculating the distance between two samples with respect to a feature, m is the number of randomly selected samples, $class(x_i)$ denotes the category to which sample x_i belongs, c denotes a category, and $p(c)$ denotes the prior probability of category c [25], [26].

There are two definitions of the distance function. For numeric feature attributes, the following formula can be used to calculate the distance of different samples for that feature:

$$d(x_i(j), NM_i(j)) = \left| \frac{x_i(j) - NM_i(j)}{\max(j) - \min(j)} \right| \quad (2)$$

For non-numeric feature attributes, the following formula is used to calculate the distance of different samples for that feature:

$$d(x_i(j), NM_i(j)) = \begin{cases} 0 & x_i(j) \neq NM_i(j) \\ 1 & x_i(j) = NM_i(j) \end{cases} \tag{3}$$

The ReliefF algorithm is a well-known feature selection algorithm with comparatively good performance. It is not only simple in principle and efficient in operation but also has no restrictions on data types. Therefore, it has obtained better experimental results in classification problems.

In this paper, the ReliefF algorithm is first used to select 5, 7, 9, 11, 13, and 15 features randomly from 25 extracted features. SVM is then used to classify these features and the F-score average values are calculated for the minority class and all classes. The results are shown in Fig. 2, where the horizontal axis represents the feature sequence number and the vertical axis identifies the average value for the F-score under the number of selected features. The weights of each feature in each group are sorted separately and Table 3 shows the sorted feature sequence. Finally, according to the results in Fig. 2, the best feature subset is determined, that is, the least number of features selected and the highest average F-score value. Considering both Fig. 2 and Table 3, it is found that the greater the number of features, the greater the F-score average of all classes and the minority class. The exception is for 7 features, where the average F-score value decreases. This decrease may be due to the randomness of feature selection, as poor feature selection affects the final result. Considering the time and the average F-score value, the final number of feature subsets is determined to be 11, and the feature sequence number is 25, 24, 23, 20, 21, 15, 16, 1, 3, 4, and 11, that is δ Wave, θ Wave, β Wave, coefficient of variation, SEF95-SEF50 (8-16 Hz), SEF50 (8-16 Hz), SEF95 (8-16 Hz), SEF95-SEF50 (0.5-12 Hz), zero crossing rate, information entropy and symbol entropy.

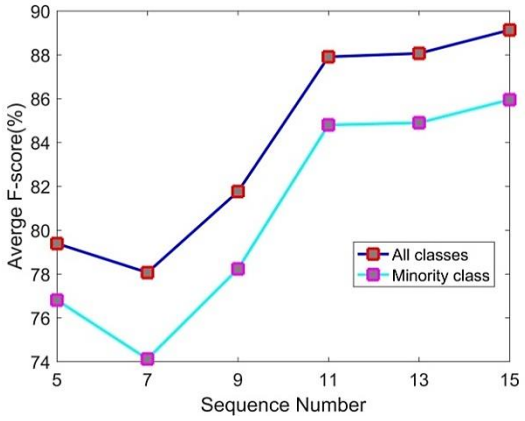


Fig. 2. SVM classification results.

Table 3. Weight Ranking Results

Number of total features	Feature numbers in order of descending weight
5	25, 24, 21, 23, 20
7	25, 24, 23, 21, 15, 20, 1
9	25, 24, 20, 23, 21, 1, 15, 3, 4
11	25, 24, 23, 20, 21, 15, 16, 1 ,3, 4, 11
13	25, 24, 23, 20, 21, 3, 1, 4, 15, 16, 11, 10, 19
15	25, 24, 23, 20, 21, 1, 3, 15, 4, 16, 11, 10, 8, 5, 2

2.2.3.2 Pearson correlation coefficient

A good subset of features should have a high correlation with classification, while the correlation between features should be low. Therefore, the Pearson correlation coefficient is used to calculate the correlation between features according to the following formula:

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\delta_X \delta_Y} \tag{4}$$

In formula (4), μ is the average value, δ is the standard deviation, and E is the mathematical expectation. If the correlation coefficient between the two features is more than 95%, the two features have strong correlation and are redundant; therefore, the feature that ranks behind in ReliefF, that is, the feature with smaller weight, is deleted.

The results of the Pearson correlation coefficient method are shown in Table 4 below. It is found that the correlation coefficients for the pairs of tested features are very low. Therefore, no features are deleted in this step.

Table 4. Pearson Correlation Coefficient Results

No.	1	3	4	11	15	16	20	21	23	24	25
1	1	0.24	-0.71	-0.24	0.07	0.63	-0.15	-0.09	0.48	0.28	0.45
3	0.24	1	-0.67	0.12	0.15	0.54	0.32	0.4	0.49	0.67	0.84
	-0.71	-0.67	1	0.15	-0.07	-0.73	-0.04	-0.08	-0.54	-0.53	-0.76
11	-0.24	0.12	0.15	1	0.06	-0.16	-0.04	0.01	0.05	0.25	-0.01
15	0.07	0.15	-0.07	0.06	1	0.01	-0.28	0.52	0.05	0.06	0.16
16	0.63	0.54	-0.73	-0.16	0.01	1	0	0.01	0.42	0.55	0.72
20	-0.15	0.32	-0.04	-0.04	-0.28	0	1	0.67	0.03	0.23	0.29
21	-0.09	0.4	-0.08	0.01	0.52	0.01	0.67	1	0.06	0.25	0.38
23	0.48	0.49	-0.54	0.05	0.05	0.42	0.03	0.06	1	0.44	0.52
24	0.28	0.67	-0.53	0.25	0.06	0.55	0.23	0.25	0.44	1	0.75
25	0.45	0.84	-0.76	-0.01	0.16	0.72	0.29	0.38	0.52	0.75	1

2.2.3.3 Sequential backward selection

This method is a top-down approach, which first assumes that the whole feature set is the optimal feature set needed at the beginning of the operation. The method then deletes a feature that does not contribute to the criterion function at each step of the algorithm until the number of remaining features meets the requirements of the set function. The advantage of this algorithm is that it takes into full account the statistical correlation between features. In practical applications, it has the characteristics of fast operation and strong computational performance and is a robust algorithm [27]-[29].

Since sequence backward selection is used to determine whether a feature contributes to the criterion function, the feature subset is evaluated by the accuracy of the final classification. Therefore, the feature subset selected in this round is the optimal subset. According to the results of this round of screening, feature numbers 11, 16, and 24 are delete in this step, which shows that the eight remaining features are the optimal subset for this paper, and the feature sequence number is 1, 3, 4, 15, 20, 21, 23, 25. All of the following results will be based on these eight features.

2.2.4. Classification method

The main idea of SVM [30] is to map the training set to a high-dimensional space through a kernel function, which can solve problems such as over-learning, dimension disaster, and local minimum. Some studies show that SVM has a better classification effect on balanced data. Therefore, this paper will give priority to using SVM as the classification method. The objective function of the nonlinear SVM problem is

as follows:

$$\max L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$s.t. \sum_{i=1}^{N} y_j \alpha_i = 0, 0 \le \alpha_i \le C, i = 1, 2, \ldots, N \quad (5)$$

In formula (5), $k(x_i, x_j)$ is a kernel function. The kernel function selected in this paper is the Gauss Radial Basis Kernel Function, shown in Formula (6):

$$K(x, x') = \exp(-\gamma \left\| x - x' \right\|^2) \quad (6)$$

The values of parameters C, γ have a great influence on the accuracy of the model. Therefore, a pair of multi-coding strategies is used to first divide the multi-classification problem into several binary classification problems. Then, a five-fold cross-validation is used, where the data sets are divided into five parts on average, four parts are randomly selected as training sets, and the remaining are used as test sets. Ten results are averaged together and used as the final result to optimize the model and obtain more reliable experimental results.

2.2.5. Model evaluation criteria

In traditional classification methods, accuracy is often used as an evaluation index, but in unbalanced data classification, accuracy is no longer a reasonable index [31], [32]. For unbalanced problems, the commonly used indicators are based on confusion matrices, such as Recall, F-measure, G-mean, and AUC. [33]. The confusion matrices are shown in Table 5.

Table 5. Confusion Matrix

Classification	Predictive Positive	Predictive Negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

The F-measure is the harmonic mean of recall and precision, which is defined as follows:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} \quad (7)$$

where $recall = \frac{TP}{TP + FN}$, $precision = \frac{TP}{TP + FP}$

The G-mean values represents the geometric mean of the classification accuracy of the minority classes and majority classes. This value maximizes the accuracy of both classes while maintaining the balance of the classification accuracy of the majority classes and minority classes. That is, the G-mean values is the largest only when both classes are high. Therefore, the G-mean value can reasonably evaluate the overall classification performance of unbalanced datasets [34]. It is defined as follows:

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{FP + TN}} \quad (8)$$

The Area Under Curve (AUC) is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is not affected by the distribution of the sample. Therefore, the AUC is used as an evaluation

index for the classification model of unbalanced data in many papers. The larger the area under the curve, the larger the AUC value will be, and the model will have a better classification effect. This paper will use the F-measure value to measure the classification performance of minority classes, the G-measure value to measure the overall classification performance of data sets, and the AUC value to measure the classification performance of classifiers.

3. Results

The data set is divided into the training set and test set according to 8:2 ratio. The validation method is a 5-fold cross-validation, and the average of 10 results is used as the experimental results.

Table 6 compares the F-score values of the minority class of different algorithms under the balanced data set and the unbalanced data set, and calculates the average value of the F-scores for each category of the minority class according to formula (7) as the final result.

Table 7 compares the G-mean values of the different sleep stages under different algorithms and whether the data sets are balanced.

Table 8 compares the AUC values of different sleep stages under different algorithms and whether the data sets are balanced.

Table 6. Comparison of Minority Class F-score Values of Different Algorithms in Balanced and Unbalanced Datasets (%)

Algorithm	Balanced	Unbalanced
Decision Tree	76	70
KNN	82	72
Naive Bayes	87	69
SVM	90	76

Table 7. Comparison of G-mean Values of Different Algorithms in Balanced and Unbalanced Datasets (%)

Category	Balanced				Unbalanced			
	Decision Tree	KNN	Naive Bayes	SVM	Decision Tree	KNN	Naive Bayes	SVM
W	89	93	89	99	92	92	83	94
N1	91	90	88	95	66	62	88	82
N2	87	92	90	93	91	88	90	90
N3	93	91	91	96	80	77	88	82
N4	95	93	90	97	90	89	88	90
R	88	92	88	95	87	82	87	87

Table 8. Comparison of AUC Values of Different Algorithms in Balanced and Unbalanced Datasets (%)

Category	Balanced				Unbalanced			
	Decision Tree	KNN	Naive Bayes	SVM	Decision Tree	KNN	Naive Bayes	SVM
W	95	95	95	99	98	98	86	99
N1	95	90	96	98	89	90	97	96
N2	93	94	96	99	96	96	97	98
N3	95	92	97	99	94	97	97	97
N4	97	94	96	99	98	98	97	99
R	94	94	96	99	96	96	96	97

Fig. 3 compares the F-score values for balanced and unbalanced data and shows the evaluation results of different algorithms under balanced and unbalanced conditions. In Fig. 4, the horizontal axis represents four different algorithms, and the vertical axis identifies the values under different evaluation indexes. In the Fig. 4, Un and B represent unbalanced and balanced data sets, respectively. This paper also uses the ROC curve to observe the classifier performance under balanced and unbalanced data sets, as shown in Fig. 5 below.

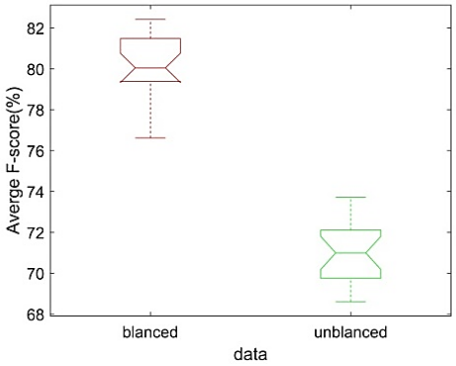


Fig. 3. Comparison of F-score values between balanced and unbalanced datasets.

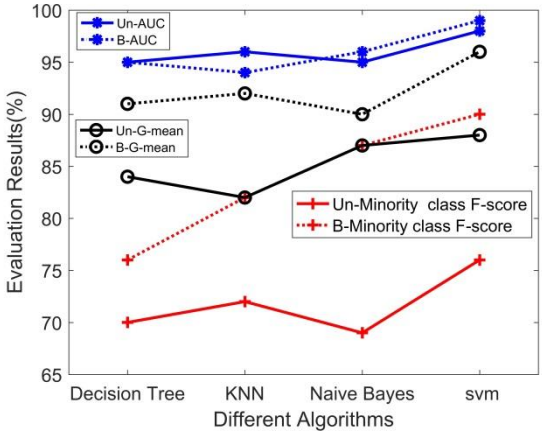
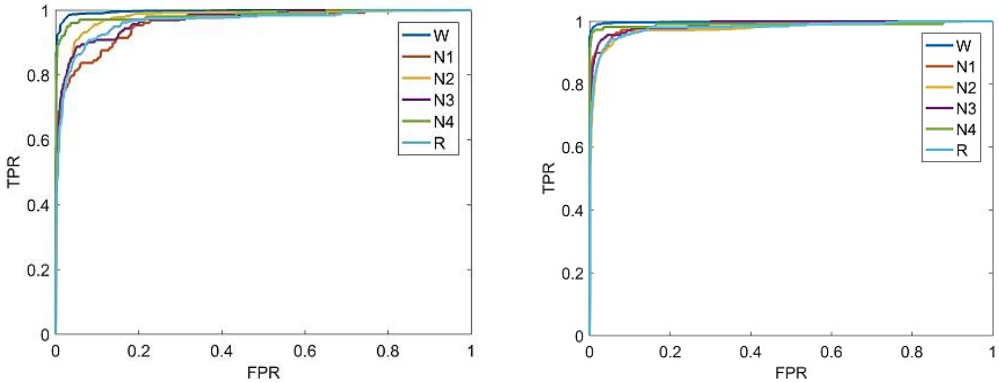


Fig. 4. Comparison of evaluation results between balanced and unbalanced datasets.



(a) ROC Curves for Unbalanced Datasets (b) ROC Curves for Balanced Datasets

Fig. 5. Comparison of ROC curves between balanced and unbalanced datasets.

By considering the results in Table 6, Table 7 and Table 8, it can be seen that after data balancing, the F-score values of the minority class, G-mean values and AUC values classified by different algorithms are higher than those obtained for unbalanced data sets. Additionally, when comparing the four algorithms of decision tree, KNN, Naive Bayes and SVM, the classification result of SVM is the highest in both the balanced and unbalanced data sets. Because SVM is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization, it can achieve global optimal classification with limited sample information [35]. In addition, in Table 7, the G-mean value of the awake stage is the highest under

the SVM classification because the EEG features of waking and sleeping conditions are quite different. From the results shown in Fig. 3 and Fig. 5, the classification effect has been improved after data balancing, and the average F-score value in balanced data sets is 9.2% higher than that in unbalanced data sets. In Fig. 4, the G-mean and F-score values fluctuate greatly before and after balancing, but the AUC values fluctuate very little, indicating that the AUC value is not affected by the sample distribution. AUC values closer to 1 indicate better classification effects. In summary, the proposed framework is suitable for dealing with multi-class unbalanced data models in sleep staging and can more effectively increase the number of minority class samples in unbalanced data sets, thereby improving the classification accuracy of the minority class and the classification performance of classifiers.

4. Conclusion

To address the problem of low classification accuracy of minority classes in sleep staging with unbalanced data, this paper makes improvements on the data level, extracting and filtering the features of the data. Additionally, the proposed framework is validated by using the EEG data of MIT-BIH, which is a public database. The results show that the proposed framework not only improves the recognition effect of the minority class but also improves the overall classification effect. In future work, the following improvements should be considered: (1) reducing the time complexity and improving the real-time performance of data balancing, (2) extracting better and fewer features of classification, and (3) improving the method of synthesizing the minority class samples to make the distribution of new samples more reasonable.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgment

This work is supported by the 13th Five-Year Plan of the National Educational Science Key Issues of the Ministry of Education in 2017 under Grant No. DLA170428.

References

[1] Sahar, S., & Mahdi, E. (2017). A fuzzy decision tree approach for imbalanced data classification. *Proceedings of the International Conference on Computer & Knowledge Engineering* (pp. 292-297).

[2] Annarita, D., Alberto, R., Guido, P., *et al*. (2016). A Bayesian network for flood detection combining SAR imagery and ancillary data. *IEEE Transactions on Geoscience & Remote Sensing, 54(6)*, 3612-3625.

[3] Gustavo, E. A. P. A., Ronaldo, P. C., & Maria, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6(1)*, 20-29.

[4] Liliya, D., & Irina, K. (2017). SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem. *Embedded Computing*, 1-4.

[5] Weiss, G. M. (2004). Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter, 6(1)*, 7-19.

[6] Gururajapathy, S. S., Mokhlis, H., Illias, H., *et al*. (2016). Fault identification in an unbalanced distribution system using support vector machine. *Journal of Electrical Systems, 12(4)*, 786-800.

[7] Lei-Fu, G., Shi-Jie, Z., Dong-Mei, Y. U., *et al*. (2017). Unbalanced support vector machine coupling negative-samples cutting with asymmetric misclassification cost. *Acta. Electronica Sinica., 45(12)*, 2978-2986.

[8] Kaur, R., & Kang, S. S. (2016). An enhancement in classifier support vector machine to improve plant

disease detection. *Proceedings of the IEEE International Conference on Moocs.* (pp. 135-140).

[9] Jinmeng, L. I., Yaping, L, & Tuanfei, Z. (2018). KNN classification algorithm based on hubness and class weighting. *Computer Engineering*.

[10] Chan, P. K., & Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. *Proceedings of the Fourth International Conference on Knowledge Discovery & Data Mining*.

[11] Murph, P. M., & Aha, D. W. (1991). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine.

[12] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21(9)*, 1263-1284.

[13] Japkowicz, N. & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis Journal, 6(5)*, 429-450.

[14] Gustavo, E., Batista, P., & Ronaldo, C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6(1)*, 20-29.

[15] Chawla, N. V., Bowyer, K. W., Hall, L. O., *et al*. (2011). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16(1)*, 321-357.

[16] Bingyan, X., Guoyin, W., & Weibin, D. (2016). Under-sampling method based on sample weight for imbalanced data. *Journal of Computer Research and Development, 53(11)*, 2613-2622.

[17] Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing, 135(8)*, 32-41.

[18] Xuying, L., Jianxin, W., & Zhihua, Z. (2006). A cascade-based classification method for class-imbalanced data. *Journal of Nanjing University (Natural Sciences), 42(2)*, 148-155.

[19] Tang, Y., Zhang, Y. Q., Chawla, N. V., *et al*. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Cybernetics, 39(1)*, 281-288.

[20] Chawla, N. V., Lazarevic, A., Hall, L. O., *et al*. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *Knowledge Discovery in Databases: PKDD, 2838*, 107-119.

[21] Peng, Z., Xiangxin, L., Yi, Z., *et al*. (2013). Research on individual sleep staging based on principal component analysis and support vector machine. *Journal of Biomedical Engineering, 30(6)*, 1176-1179.

[22] Hsu, Y. L., Yang, Y. T., Wang, J. S., *et al*. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing, 104*, 105-114.

[23] Zhang, D., Ma, J., Yi, J., *et al*. (2015). An ensemble method for unbalanced sentiment classification. *Proceedings of the 2015 11th IEEE International Conference on Natural Computation.*

[24] Knonenko, I. (2007). Estimation attributes: Analysis and extensions of relief. *Proceedings of the European Conference on Machine Learning* (pp. 171-182).

[25] Yijun, S. (2007). Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Trans on Pattern Anslysis and Machine Intelligence, 29(6)*, 1035-1051.

[26] Sun, Y., & Wu, D. (2008). A relief based feature extraction algorithm. *Proceedings of the 8th SIAM International Conference on Data Mining* (pp. 188-195).

[27] Furlanello, C., Serafini, M., Merler, S., *et al*. (2003). An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks, 16(5-6)*, 641-648.

[28] Somol, P., Pudil, P., Novovicova, J., *et al*. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters, 20(11-13)*, 1157-1163.

[29] Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15 (11)*, 1119-1125.

[30] Elif, D., Cvetkovic, D., Holland, G., *et al*. (2010). Analysis of sleep EEG activity during hypopnoea

episodes by least squares support vector machine employing AR coefficients. *Expert Systems with Applications*, *37(6)*, 4463-4467.

[31] Yaxiang, G., & Shifei, D. (2011). Advances of support vector machines (SVM). *Computer Science, 38(2)*, 14-17.

[32] Yang, Y., & Shanping, L. (2009). Instance importance based SVM for solving imbalanced data classification. *Pattern Recognition and Artificial Intelligence, 22(6)*, 913-918.

[33] Zhiyong, L., Zhifeng, H., & Xiaowei, Y. (2010). Effects of several evaluation metrics on imbalanced data learning. *Journal of South China University of Technology (Natural Science Edition)*, *38(4)*, 147-155.

[34] Zaixiang, H., Zhong-Mei, Z., Tian-Zhong, H., *et al*. (2015). Improved associative classification algorithm for multiclass imbalanced datasets. *Pattern Recognition and Artificial Intelligence, 28(10)*, 922-929.

[35] Zhaoxia, X., Yiqin, W., Jianjun, Y., *et al*. (2011). Recognition of TCM syndrome types of cardiovascular diseases based on support vector machine and artificial neural networks. *Journal of Beijing University of Traditional Chinese Medicine, 34(8)*, 539-543.

Dan Li received a B.S. degree in electronic information engineering from Xijing University, China, in 2014, and an M.S. degree in computer science from Guizhou Normal University in 2017. She currently works as an algorithm engineer in Chengdu Techman Software Co. LTD. Her current research interests include signal acquisition systems research and the application of machine learning algorithms.

Wu Huang received a B.S. degree in computer science from Sichuan University, China, in 1993, and an M.S. degree in computer science from Sichuan University in 2000. He is currently pursuing a Ph.D. degree in the School of Computer Science at Sichuan University. He is a lecturer in the School of Computer Science at Sichuan University, China. His research interests include embedded real-time systems and intelligent medical instruments.

Guobiao Xu received a B.S. in computer science from Sichuan University, China, in 1990, and he received a M.S. in computer science from Sichuan University, China in 1998. Currently, he is an associate professor in the School of Computer Science at the Civil Aviation Flight University of China. His research interests are computer graphics and CAI.

Tao Zhang received a B.S. degree in automation from Southwest University of Science and Technology, China, in 2009, and M.S. degree in optics from University of Electronic Science and Technology of China, China, in 2017. He currently works as an algorithm engineer in Chengdu Techman Software Co.LTD., China. His professional research interests include machine learning in health care and biological signal processing.

Zhonghui Jiang received a B.S. degree in mechanical engineering from Southwest Jiaotong University, China, in 2015, and M.S. degree in vehicle operation engineering from State Key Laboratory of Traction Power, Southwest Jiaotong University, China, in 2018. He currently works as an algorithm engineer in Chengdu Techman Software Co. LTD., China. His professional research interests are in data processing and analysis and research and application of machine learning algorithm.

Xiao Wei received a B.S. degree in biology science from Xichang College, China, in 2012. He received Ph.D. degree in materials science and engineering through his master's and doctoral programs from Southwest Jiaotong University, China, in 2018. He is currently a senior research fellow in medicine and the manager of research department in Chengdu Techman Software Co. LTD., China. His current research interests mainly focus on the research and application of correlation algorithm for interior space location, and the study of the construction of multi-index system by using the mathematical statistics.