How to balance the bioinformatics data: pseudo-negative sampling

Yongqing Zhang1,2, Shaojie Qiao3,4*, Rongzhao Lu1, Nan Han5, Dingxiang Liu6 and Jiliu Zhou1

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: Imbalanced datasets are commonly encountered in bioinformatics classification problems, that is, the number of negative samples is much larger than that of positive samples. Particularly, the data imbalance phenomena will make us underestimate the performance of the minority class of positive samples. Therefore, how to balance the bioinformatic data becomes a very challenging and difficult problem.

Results: In this study, we propose a new data sampling approach, called pseudo-negative sampling, which can be effectively applied to handle the case that: negative samples greatly dominate positive samples. Specifically, we design a supervised learning method based on a max-relevance min-redundancy criterion beyond Pearson correlation coefficient (MMPCC), which is used to choose pseudo-negative samples from the negative samples and view them as positive samples. In addition, MMPCC uses an incremental searching technique to select optimal pseudo-negative samples to reduce the computation cost. Consequently, the discovered pseudo-negative samples have strong relevance to positive samples and less redundancy to negative ones.

Conclusions: To validate the performance of our method, we conduct experiments base on four UCI datasets and three real bioinformatics datasets. According to the experimental results, we clearly observe the performance of MMPCC is better than other sampling methods in terms of Sensitivity, Specificity, Accuracy and the Mathew’s Correlation Coefficient. This reveals that the pseudo-negative samples are particularly helpful to solve the imbalance dataset problem. Moreover, the gain of Sensitivity from the minority samples with pseudo-negative samples grows with the improvement of prediction accuracy on all dataset.

Keywords: Imbalanced data, Pseudo-negative sampling, Pearson correlation coefficients, Max-relevance, Min-redundancy
**Background**

The work is motivated by the real-world requirement in bioinformatic data processing: it is very common that negative samples greatly dominate positive samples, and this phenomena is called data imbalance problem. In general, we cannot achieve genetic data mining with limited positive samples. So, we think that: whether we could use positive samples by mixing pseudo-negative data (which is classified to be negative data, but they are similar to positive samples with the maximum relevance and they have the minimum redundancy with negative samples) to predict the categories of samples. Because of the lack of enough positive samples, the biologist cannot perform experiments. Consequently, some positive samples cannot be identified or categorised as negative samples which can be viewed defined as pseudo-negative samples. So how to select these pseudo-negative samples will be an alternative method to solve the imbalanced data problem in bioinformatics.

In the post-genome era, with the wide application of various high-throughput technologies, biological data has rapidly increased [1, 2]. Machine learning technology can be applied to discovery important information for understand complex biological processes from large-scale biological data [3–9]. However, imbalanced data is a very common phenomenon in the real dataset (where the positive sample is the minority class). Many bioinformatics applications require class imbalance learning, such as gene expression data [10, 11], protein-DNA binding data [12, 13], N6-methylation sites in mRNAs [14], splice sites prediction [15], prediction of microRNAs [16], prediction of protein interaction [17–21], transcription factor binding sites prediction [22, 23] and so on. In this scenario, the performance of the minority classes can be greatly underestimated [24].

To the best of our knowledge, researchers have proposed some strategies to degrade the influence of imbalance data. These existing methods can be classified into data-level approaches and algorithmic-level approaches [25, 26]. In regard of data-level approaches, re-sampling techniques are employed to balance the sample space w.r.t. an imbalanced dataset in order to alleviate the negative effect of the skewed distribution of samples in the learning process. Resampling methods are very commonly-used approach because they are independent of classifiers. Resampling techniques can be classified into three categories depending on the method used to balance the proportion of positive and negative samples: (1) over-sampling: eliminating the negative effect of skewed distribution by generating new samples of minority class. Two widely-used approaches to generate minority samples are Random Over-Sampling (ROS) which randomly duplicate the minority samples, and SMOTE. (2) Under-sampling: balance the data by discard the samples from the majority class. The simplest yet most effective method is Random Under-Sampling (RUS) which involved the random elimination of majority class examples [27]. RUS deals with the class imbalance problems in an effectively fashion. (3) Hybrid methods: these are a combination of the over-sampling and under-sampling method. The commonly-used algorithmic-level approach is cost-sensitive learning method which assigns higher costs to the minority class [28, 29].

However, RUS often loses some important classification information and ROS is time-consuming and often results in the phenomenon of overfitting. So, it is essential to propose advanced data sampling approaches to maintain the structure of groups and generate new data according to its underlying distribution.

To overcome the problems caused by the imbalanced bioinformatic data, we first propose the pseudo-negative sampling approach based on Max-relevance and Min-redundancy Pearson correlation coefficient (called MMPCC). In the MMPCC approach, Pearson correlation coefficients are used to measure the similarity between positive and negative samples and the coefficients are learned from positive and negative samples based on the max-relevance and min-redundancy criteria. The new algorithm can discover the pseudo-negative samples which may be viewed as positive samples, but their labels are negative. This proposed sampling approach aims at alleviating the imbalanced ratio. The experiments are applied on two UCI data and three real-life bioinformatics data.

**Contribution:** The original contributions of this study can be summarized as follows.

1) We propose a concept of pseudo-negative samples and present a pseudo-negative sampling method which is based on the max-relevance and min-redundancy Pearson correlation coefficient in supervised learning. In particular, both positive and negative samples are taken into full consideration in order to find optimal pseudo-negative samples.

2) We use an incremental searching method for calculating the coefficient of positive and negative samples, which can avoid the high computational cost in selecting the subsets of pseudo-negative samples.

3) We conduct extensive experiments and the results demonstrate the advantage of the MMPCC method for handling the imbalanced bioinformatic data.

**Methods**

**Pseudo-negative sampling method**

Although pseudo-negative samples are viewed to be negative, but they are similar to positive samples with the maximum relevance and they have the minimum redundancy with negative samples. The key idea of pseudo-negative sampling approach is to select a subset from the
negative samples and classify them into positive class by
the method of max-relevance and min-redundancy on
Pearson correlation coefficient in the phase of training.
The formal definition of pseudo-negative samples is given
as follows.

**Definition 1** (Pseudo-negative samples). Given a po-
itive data set \( S^+ = \{ (x_1^+, y_1^+), (x_2^+, y_2^+), ..., (x_m^+, y_m^+) \} \), a
negative data set \( S^- = \{ (x_1^-, y_1^-), (x_2^-, y_2^-), ..., (x_n^-, y_n^-) \} \), then
a pseudo-negative data set is represented by
\( S^* = \{ (x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_l^*, y_l^*) \} \),
where \( m \) is the total number of positive data,
\( n \) is the total number of negative data,
\( m < n \), and \( l \) is the number of pseudo-negative samples.

The purpose of our method is to identify the pseudo-
negative sample set \( S^* \) (which might contain \( l \) samples)
based on \( S^+ \) and \( S^- \), where \( l < m \).

One of the famous sequential search methods is the
incremental sample search algorithm, and we employ it
in the study. To achieve the incremental sample search-
ing, the pseudo-negative sample set starts from \( S_0^* = \emptyset \),
and a quantitative criterion \( Q(S^*_k) \) is used to measure the
similarity of samples in \( S^*_k \).

In each round of searching, a sample \( S^{*'} \) would be added
in the sample set \( S^*_k \).

\[
S^*_k = S^*_{k-1} \cup S^{*'}
\]

(1)

where

\[
S^{*'} = \arg\max_{S^{*'} \subseteq S^* \cap S^{*'} = \emptyset} Q(S^*_k)
\]

(2)

\( Q(S^*_k) \) plays an important role in the sample selection,
which can be defined with different requirements. The
validation accuracy is utilized to evaluate the new sam-
ple subsets. In this study, the metric of Eq. 3 is utilized
to evaluate the similarity of samples in \( S^*_k \) and \( S^{*'} \),
and the corresponding quantitative criterion is given by
the following equation:

\[
Q(S^*_k) = A(S^*_k \cup S^{*'})
\]

(3)

where \( S^{*'} \) is a potential pseudo-negative sample and \( S^*_k \)
is the pseudo-negative sample set, and \( A \) represents the
validation accuracy.

In this study, we employ the Pearson correlation coef-
cient between samples in order to select a new sample.
\( Q(S^*_k) \) can be transformed to be the following equation:

\[
Q(S^*_k) = P(S^*_k \cup S^{*'})
\]

(4)

The details of calculating the Pearson correlation coeffi-
cient are given in the following.

**Max-relevance and min-redundancy on pearson
 correlation coefficient**

Pearson correlation coefficient (PCC) [30] is defined on
the covariance matrix, which is a method to evaluate
the strength of the relationship between two vectors. In
general, the coefficient between two vectors \( \alpha_i \) and \( \alpha_j \) is
defined as follows:

\[
P(\alpha_i, \alpha_j) = \frac{\text{cov}(\alpha_i, \alpha_j)}{\sqrt{\text{var}(\alpha_i) \times \text{var}(\alpha_j)}}
\]

(5)

According to the max-relevance, PCC beyond negative
sample and positive sample are formalized as follows:

\[
D(S^-_i, S^+_j) = P(S^-_i, S^+_j)
\]

(6)

where \( S^-_i \in S^- \), \( i \in N \), \( S^+_j \in S^+ \) and \( j \in M \) agreeing with
the max-relevance criterion. The most relevant feature set
can be obtained by maximizing \( D(S^-_i, S^+_j) \).

\[
S_{\text{max}} = \arg \max \ D(S^-_i, S^+_j)
\]

(7)

Based on the min-redundancy criterion, the samples
could be selected by the following equation:

\[
R = \frac{1}{|S^+|^2} \sum\n P(S^-_i, S^+_k)
\]

(8)

where \( S^-_i \in S^- \) and \( S^+_k \in S^+ \),

\[
S_{\text{min}} = \arg \min \ R
\]

(9)

In terms of incremental search method, an operator
\( \Psi(D, R) \) is defined in Equation 10 in order to optimize
the max-relevance and min-redundancy information. The
best selected sample \( S^{*'} \) is given as follows:

\[
\Psi = D - R
\]

(10)

\[
S^{*'} = \arg \max \ \Psi(D, R)
\]

(11)

Assume we have the sample subsets \( S^*_{k-1} \) which have \( k-1 \)
samples. In the next step of searching, the \( k^{th} \) sample is
obtained from the sample subsets \( \{ S^- - S^*_{k-1} \} \). Then, \( S^*_k \)
can be calculated by Eq. 12 based on \( \Psi(D, R) \).

\[
S^*_k = \arg \max \left\{ P(S^-_i, S^+_j) - \frac{1}{k-1} \sum_{S^*_k \in S^*_{k-1}} P(S^-_i, S^*_k) \right\}
\]

(12)

where \( S^-_i \in \{ S^- - S^*_{k-1} \} \) and \( S^+_j \in S^+ \).

**The proposed pseudo-negative sampling algorithm**

Based on the aforementioned preliminaries, we propose
a pseudo-negative sampling algorithm based on the
max-relevance and min-redundancy on Pearson correlation
coefficient, which is called MMPCC. The detail of the
MMPCC algorithm is presented in Algorithm 1 and the
flow chart is shown in Fig. 1.

As described in Algorithm 1, the selected pseudo-
negative samples can be updated step by step. Firstly,
the max-relevance between the negative sample and the
positive sample is calculated by Equation 7 in order to
choose candidate pseudo-negative samples. Then, the
new selected sample will be identified based on the min-redundancy of samples in the selected pseudo-negative subsets by Equation 9. Lastly, the new sample will be identified to be pseudo-negative sample by Equation 12.

It is worthwhile to note that $l$ is specified by experts in order to determine how many pseudo-negative samples should be inserted into the positive samples set.

The computational complexity of MMPCC, MAXR and MINR includes two parts: the computation of similarity matrices and the computation of sample ranking. The operator $\psi_{\text{MAXR}}$ can be obtained via Equation 7, the operator $\psi_{\text{MINR}}$ can be calculated by Equation 9 and the MMPCC model be figured out by Equation 12.

As for MAXR, the computation of Pearson correlation coefficient of all pairwise negative data and positive data requires the complexity of $O(n \times m \times f)$, where $n$ is the number of negative data, $m$ is the number of positive data and $f$ is the number of attributes of each data. As for MINR, the computation complexity is $O(n \times l \times f)$, where $l$ is the number of pseudo-negative samples. Therefore, the computation complexity of MMPCC is the sum of MAXR and MINR, that is, $O(n \times m \times f + n \times l \times f)$.

**Algorithm 1** Pseudo-negative sampling by Max-relevance and Min-redundancy on Pearson Correlation Coefficient

**Input:** a positive dataset $S^+ = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, a negative dataset $S^- = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

**Output:** a pseudo-negative dataset $S_l^*$, $l$ is the number of pseudo-negative samples.

1. Initialize the target sample subsets $S_0^* = \emptyset$ and the available sample subsets $S_\alpha^- = S^-$
2. for $k = 1$ to $l$
   3. for each $S_j^-$ in $S_\alpha^-$ and each $S_i^+$ in $S_\alpha^+$
   4. search for the new sample $S_k^*$ according to:
   5. $S_k^* = \arg\max [P(S_i^+, S_j^-) - \frac{1}{k-1} \sum P(S_i^-, S_j^-)]$
   6. end for
   7. update $S_k^* = S_i^+ \cup S_k^*$ and $S_\alpha^- = S_\alpha^- - S_k^*$
3. end for
4. return pseudo-negative dataset $S_l^*$

**Classification methods**

**Random forests**

The classifier of Random forests [31, 32] is an ensemble learning method, which works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of overfitting to their training set.

**Neural networks**

A neural network [33] is composed of several simple "neurons", and the output of a neuron will be the input of another. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be -1 and 1.

**AdaBoost**

AdaBoost, short for "Adaptive Boosting", which is a general ensemble method [34]. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. The final equation for classification can be represented by:

$$ F(x) = \text{sign} \left( \sum_{m=1}^{M} \theta_m f_m(x) \right) $$  \hspace{1cm} (13)
where $f_m$ represent the $m^{th}$ weak classifier and $\theta_m$ is the corresponding weight. It is exactly the weighted combination of $M$ weak classifiers.

**Discriminant analysis**

Discriminant analysis (DA) is one of the classification methods. The basic idea is that: two or more clusters or populations are a priori known and one or more observations are classified into one of the known populations according to the measure characteristics [35]. Let $X$ is a $q$-dimensional vector representing an observation from one of several possible classes. If the category is unknown, $X$ can be classified using the discriminant analysis approach. Alternatively, it can be used to characterize the difference between classes via a discriminant function.

**Datasets**

In order to evaluate the prediction performance of MMPCC on pseudo-negative sampling, we compare it with the state-of-the-art prediction methods. In experiments, we use four UCI Repository datasets [36] and three real bioinformatic datasets. Table 1 introduces the detail of the datasets.

From Table 1, we can see that the number of attributes of each dataset is 9, 3, 10, 49, 180, 180 and 25, respectively. We use all attributes of each dataset in MMPCC. In MMPCC, the Pearson correlation coefficient is used to calculate the similarity between negative and positive samples in Equation 7, and is also applied in Equation 10 and Equation 12. Additionally, the coefficient between two vectors $\alpha_i$ and $\alpha_j$ in Equation 5 is obtained by all attributes of each dataset.

In Table 1, Positive represents the number of positive samples, Negative represents the number of negative samples, and Ratio = Negative Numbers / Positive Numbers.

More specifically, the first UCI datasets, Contraceptive Method Choice (CMC) contains 333 minority samples and 1140 majority samples, and the number of attributes is 9. The second UCI datasets, Haberman’s Survival Data, contains 81 minority samples and 225 majority samples, and the number of attribute is 3. The third dataset Solar Flare records the number of solar flares. Each attribute calculates the number of a certain type of Solar Flare within 24 hours. Each instance represents the number of all types of flares in an active region on the sun. The data contains 69 minority classes and 1320 majority classes, with 10 attributes. The fourth datasets Oil contains 41 minority classes and 896 majority classes, including 49 attributes.

The first bioinformatic datasets, SNP data [37], included 183 positive samples and 2891 negative samples, and the number of attributes is 25. The second bioinformatic datasets, PDNA-543 [38], consists of 543 protein sequences, which are all related into the PDB (Protein Data Bank) before October 10, 2014. There are 9,549 DNA-binding residues as positive samples and 134,995 non-binding residues as negative samples in PDNA-543. The third bioinformatic datasets, PDNA-316, is constructed by Si et al [39], which has 316 DNA-binding protein chains and 5,609 binding residues and 67,109 non-binding residues.

**Evaluation metrics**

In this study, four metrics are used to evaluate the performance of different classifiers, including Sensitivity ($Sen$), Specificity ($Spe$), Accuracy ($Acc$) and the Mathew’s Correlation Coefficient ($MCC$). They are calculated according to the following equations:

\[
Sen = \frac{TP}{TP + FN} \quad (14)
\]

\[
Spe = \frac{TN}{TN + FP} \quad (15)
\]

\[
Acc = \frac{TP + TN}{TP + FN + TN + FP} \quad (16)
\]

\[
MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (17)
\]

where $TP$ is the number of true positives, $TN$ is the number of true negatives, $FP$ is the number of false positives, $FN$ is the number of false negatives, $P$ is the number of positives, and $N$ is the number of negatives.

Sensitivity indicates how well the test predicts the true positives, Specificity measures how well the test predicts the true negatives, Accuracy is expected to measure how well the test predicts both true positives and negatives, and MCC considers true and false positives and negatives. So, the higher the values of these evaluation metrics, the better the results.

| Table 1 Description of datasets |
|----------------------------------|
| Dataset   | Positive | Negative | Attributes | Ratio |
|-----------|----------|----------|------------|-------|
| CMC       | 333      | 1140     | 9          | 3.4   |
| Haberman  | 81       | 225      | 3          | 2.7   |
| Solar Flare| 69      | 1320     | 10         | 19.1  |
| Oil       | 41       | 896      | 49         | 21.9  |
| PDNA-543  | 9549     | 134995   | 180        | 14.1  |
| PDNA-316  | 5609     | 67109    | 180        | 11.9  |
| SNP       | 183      | 2891     | 25         | 15.7  |
Results

The purpose of the evaluation is to examine the effectiveness of our proposed MMPCC method on selecting the pseudo-negative samples. Four sets of experiments are conducted. Experiment 1 compares the different percentage of pseudo-negative sampling on two UCI datasets. Experiment 2 compares the different percentage of pseudo-negative sampling on three bioinformatic datasets. Experiment 3 compares MMPCC with the max-relevance and the min-redundancy methods on the PDNA-316 dataset, which aims to evaluate the relation between the relevance and the redundancy. For simplicity, the max-relevance method is represented by MAXR and the min-redundancy method is represented by MINR. Experiment 4 compares MMPCC with other sampling methods on the bioinformatic datasets.

In experiments, five-fold cross-validation is used to train the dataset. In order to give comprehensive results, Discriminant Analysis, AdaBoost, Random Forest and Neural Networks are employed for classification. We use DA, Adaboost, RF and NN to represent these four classifiers in the experiments, respectively.

Experiment 1: experiments on UCI datasets

This set of experiments examines the contribution of different percentage of pseudo-negative sampling on the UCI datasets [36]. The results are shown in Table 2 and Table 3. As mentioned previously, we use the metrics of Sen, Spe, Acc and MCC.

Table 2 presents the performance of different classifiers on the CMC dataset, where the percentage of pseudo-negative samples changes from 0% to around 50%. 0% means the dataset is not used pseudo-negative sampling. We can see that the performance is improved with larger percentage of pseudo-negative samples, where the Random Forest method achieve 28.19%, 39.22%, 43.94%, 50.87%, 56.45% and 62% for Sen when the percentage of pseudo-negative samples is fixed to 0%, 10%, 20%, 30%, 40% and 50%, respectively. In addition, the Acc value is 78.2%, 78.41%, 78.48%, 79.57% and 79.63% and the MCC value is 0.27, 0.369, 0.404, 0.448, 0.505 and 0.532. The performance of different evaluation metrics show a trend of increasing with a higher percentage of pseudo-negative samples, which agrees with the real-world situation that: if we add more positive samples, the classifier will have better performance.

Furthermore, the Neural networks method achieves 27.01%, 40.92%, 47.28%, 53.39%, 54.94% and 61.02% for Sen when the percentage of pseudo-negative samples is fixed to 0%, 10%, 20%, 30%, 40% and 50%, respectively. Moreover, the MCC value is 0.161, 0.302, 0.368, 0.439, 0.439 and 0.505. For Discriminant analysis method, the Sen values are increased by 9.38%, 17.6%, 37.35%, 52.46%, 59.46% and 66.78% and the MCC values are increased by 0.156, 0.198, 0.351, 0.438, 0.485 and 0.530 on different percentage of pseudo-negative samples, respectively. Similarly, the performance of the AdaBoost classifier obtain improvement on Sen and MCC, which demonstrates the effectiveness of the proposed pseudo-negative sampling method.

Table 3 shows similar results on different metrics as Table 2, which verify that pseudo-negative sampling is very useful in classify the imbalance data and can obtain good performance of classification. Furthermore, the results indicates that pseudo-negative samples can be viewed as positive samples and be used to classify objects. For the instability of MMPCC, the results are often not unique in Table 3. There are three reasons about this issue: Firstly, four classification methods were employed, DA, RF, NN and AdaBoost in this study. Different machine learning method has different character, so the experiment results have little instability. Secondly, the value of Sensitivity and Specificity has little instability, but the value of MCC is more stable in most of experiments. As the Sensitivity and Specificity are the singular assessment

| Percentage | Classifier | Sen(%) | Spe(%) | Acc(%) | MCC     |
|------------|------------|--------|--------|--------|---------|
| 0          | DA         | 9.38   | 97.81  | 77.8   | 0.156   |
|            | AdaBoost   | 21.37  | 94.48  | 77.94  | 0.226   |
|            | RF         | 28.19  | 92.8   | 78.22  | 0.27    |
|            | NN         | 27.01  | 87.09  | 73.52  | 0.161   |
|            | DA         | 17.6   | 94.85  | 75.7   | 0.198   |
| 10         | AdaBoost   | 25.76  | 93.58  | 76.78  | 0.266   |
|            | RF         | 39.22  | 91.77  | 78.75  | 0.369   |
|            | NN         | 40.92  | 86.98  | 75.55  | 0.302   |
|            | DA         | 37.35  | 91.71  | 76.94  | 0.351   |
| 20         | AdaBoost   | 40.03  | 91.24  | 77.33  | 0.36    |
|            | RF         | 43.94  | 91.24  | 78.41  | 0.404   |
|            | NN         | 47.28  | 87.22  | 76.38  | 0.368   |
|            | DA         | 52.46  | 88.34  | 77.8   | 0.438   |
| 30         | AdaBoost   | 50.89  | 88.83  | 77.67  | 0.431   |
|            | RF         | 50.87  | 89.98  | 78.48  | 0.448   |
|            | NN         | 53.39  | 87.86  | 77.73  | 0.439   |
|            | DA         | 59.46  | 87.21  | 78.43  | 0.485   |
| 40         | AdaBoost   | 56.01  | 87.61  | 77.61  | 0.461   |
|            | RF         | 56.45  | 90.27  | 79.57  | 0.505   |
|            | NN         | 54.94  | 86.68  | 76.64  | 0.439   |
|            | DA         | 66.78  | 85.42  | 79.08  | 0.530   |
| 50         | AdaBoost   | 64.01  | 87.37  | 79.42  | 0.531   |
|            | RF         | 62     | 88.71  | 79.63  | 0.532   |
|            | NN         | 61.02  | 87.38  | 78.42  | 0.505   |
Table 3 Performance comparison of classifiers under different percentage of pseudo-negative samples on the Haberman data

| Percentage | Classifier | Sen(%) | Spe(%) | Acc(%) | MCC  |
|------------|------------|--------|--------|--------|------|
| 0          | DA         | 17.33  | 95.42  | 74.79  | 0.212|
|            | AdaBoost   | 29.19  | 90.89  | 74.71  | 0.266|
|            | RF         | 34.2   | 82.84  | 70.07  | 0.197|
|            | NN         | 27.98  | 87.28  | 71.68  | 0.202|
|            | DA         | 21.77  | 93.96  | 72.96  | 0.236|
|            | AdaBoost   | 32.72  | 86.12  | 70.58  | 0.214|
| 10         | RF         | 33.38  | 83.91  | 69.38  | 0.197|
|            | NN         | 30.37  | 82.01  | 67.04  | 0.144|
|            | DA         | 30.51  | 94.41  | 74.2   | 0.340|
|            | AdaBoost   | 46.68  | 87.54  | 74.26  | 0.370|
| 20         | RF         | 45.01  | 81.32  | 69.59  | 0.272|
|            | NN         | 37.42  | 82.97  | 68.57  | 0.222|
|            | DA         | 31.73  | 95.1   | 73.32  | 0.36 |
|            | AdaBoost   | 51.81  | 87.15  | 75.65  | 0.422|
| 30         | RF         | 51.06  | 79.6   | 70     | 0.311|
|            | NN         | 42.39  | 84.54  | 70.36  | 0.291|
|            | DA         | 37.13  | 94.38  | 72.93  | 0.404|
|            | AdaBoost   | 50.73  | 86.1   | 72.87  | 0.396|
| 40         | RF         | 56.81  | 78.38  | 69.95  | 0.359|
|            | NN         | 53.63  | 81     | 70.6   | 0.35 |
|            | DA         | 38.61  | 93.83  | 71.74  | 0.405|
|            | AdaBoost   | 61.46  | 82.26  | 73.81  | 0.447|
| 50         | RF         | 60.75  | 78.22  | 70.95  | 0.395|
|            | NN         | 52.41  | 79.81  | 68.56  | 0.339|

metrics. MCC considers true and false positives and negatives and is generally regarded as a balanced measure. MCC can be used even if the class size is very different. Finally, the performance of different evaluation metrics shows a trend of increasing with a higher percentage of pseudo-negative samples.

Experiment 2: experiments on real-Life bioinformatic datasets

In this section, we demonstrate the effectiveness of the proposed method, MMPCC, on the real bioinformatic datasets, including PDNA-543 [38], PDNA-316 [39] and SNP data [37]. The results are given in Fig. 1, Fig. 2 and Fig. 3.

Position Specific Scoring Matrix (PSSM) was used to extract the features from protein sequences of PDNA-543 and PDNA-316. PSSM is a very important type of evolutionary feature, which is obtained by running the PSI-BLAST program to search the SwissProt database via three iteration, with $10^{-3}$ as the E-value cutoff for multiple sequence alignment. In PSSM, there are 20 scores for each sequence position and each score implies the conservation degree of a specific residue type on that position. For each data instance, all the scaled scores in PSSM are used as its evolution features. In this study, we use the window size with 9 residues, and then obtain a vector of normalized PSSM scores whose dimensions of features are $9 \times 20 = 180$.

Figure 2 shows the classification performance on PDNA-543 dataset under different percentage of pseudo-negative samples, where RF-Sen and NN-Sen represent the Sensitivity value of RF and NN classifiers and RF-MCC and NN-MCC represent the MCC value of RF and NN classifiers.

The Sen and MCC metric of NN increase with the percentage of pseudo-negative samples changing from 0% to 50%. When the percentage of pseudo-negative samples changes from 0% to 30%, the Sen and MCC of RF algorithm keep unchanged. However, when the percentage of pseudo-negative samples is above 30%, there is a clear trend that RF has better performance as the percentage of pseudo-negative samples grows.
Figure 3 illustrates the classification performance on the PDNA-316 dataset under different percentage of pseudo-negative samples. The performance of RF is better than NN when the percentage is 0% and 10% in terms of Sen and MCC. When the percentage is above 20%, the performance of NN increases drastically and is better than RF, which shows that adding more pseudo-negative samples could help greatly improve the performance of classification. However, the performance of RF is almost unchanged. This is because the pseudo-negative samples has little effect on the RF algorithm in this dataset.

Figure 4 shows the classification performance for data SNP on different percentage of pseudo-negative samples. The Sen of NN grows rapidly among different percentages of pseudo-negative samples and the MCC of NN gradually increases when the percentage changes from 0% to 30%, and then the fluctuate is small from 40% to 50%. We can find that the Sen and MCC of RF grows as the percentage of pseudo-negative samples gradually increases.

Generally speaking, this set of experiments illustrated that the pseudo-negative samples are very important and can be used to improve the effectiveness of classification.

Experiment 3: comparison of mMPCC, mAXR and mINR on the PDNA-316 datasets

In this section, we employ the five-fold cross-validation to estimate the prediction performance of the proposed mMPCC method on four metrics. We compared mMPCC with other sampling methods including MAXR (max-relevance method based on Equation 7) and MINR (the min-redundancy method based on Equation 9) [30]. In experiments, the PDNA-316 dataset is employed to evaluate the effectiveness of mMPCC. The comparison results are shown in Fig. 5.

According to Fig. 5, it is straightforward to find that mMPCC outperforms the MAXR and MINR method in terms of Sen, Spe, Acc and MCC in the RF and NN classifiers. From Fig. 5(a), the pseudo-negative samples have a big influence on the Sen value. The Sen value of mMPCC is significantly better than MAXR and MINR, when NN is used as a classifier. For the RF classifier, MAXR is the best one when more pseudo-negative samples are added. By Fig. 5(b), with the increases of the percentage of pseudo-negative samples, the Spe value of mMPCC is very stable on RF and NN. This can be explained by the reason that some pseudo-negative samples are still negative ones. In addition, the Sen value can be improved with the cost of degradation of Spe value. Figure 5(c) demonstrates that the mMPCC method is the most stable method on Acc in the RF classifier. Figure 5(d) shows that the MCC value of mMPCC significantly outperforms the MAXR and the MINR methods. The performance of MAXR is better than MINR. The experimental results indicate that mMPCC attempts to utilize more representative samples and find the pseudo-negative samples (which can be viewed as positive samples) from the majority negative samples.

Experiment 4: comparison of mMPCC and classical sampling methods on bioinformatic datasets

In order to verify the advantage of our method, we also compare the prediction performance of mMPCC with other classical over-sampling method, i.e., SMOTE method [40], on the PDNA-316 dataset.

SMOTE is an over-sampling approach in which the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any of the $k$ minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the $k$ nearest neighbors are randomly chosen. In order to compare the performance of the algorithm, we use the default value 5 nearest neighbors the same as the reference [40]. The results of comparison performance are shown in Table 4. Because neural network can learn and model the relationships between inputs and outputs that are nonlinear and complex, and make generalizations and inferences. The runtime performance of random forest is quite good, and they are commonly-used to deal with the unbalanced and missing data.

According to Table 4, we can observe that mMPCC outperforms the SMOTE method in terms of all evaluation metrics. Taking MCC as an example, the mMPCC value in the NN classifier under different percentages of pseudo-negative samples are 0.312, 0.405, 0.464, 0.513 and 0.543, respectively, and the improvements are 0.152, 0.205, 0.248, 0.27 and 0.277, respectively compared to
Fig. 5 Comparison of algorithm MMPCC, MAXR and MINR on RF and NN classifiers for Sen, Spe, Acc and MCC performances

Table 4 Performance comparison between MMPCC and SMOTE under different percentage of pseudo-negative samples

| Percentage(%) | Methods | Classifiers | Sen(%) | Spe(%) | Acc(%) | MCC  |
|--------------|---------|-------------|--------|--------|--------|------|
|              | MMPCC   | RF          | 17.13  | 99.44  | 92.46  | 0.333|
| 10           |         | NN          | 17.05  | 99.23  | 92.25  | 0.312|
|              | SMOTE   | RF          | 16.01  | 98.27  | 91.34  | 0.235|
|              |         | NN          | 5.2    | 99.69  | 91.68  | 0.16 |
|              | MMPCC   | RF          | 17.28  | 99.45  | 91.84  | 0.337|
| 20           |         | NN          | 24.6   | 99.22  | 92.31  | 0.405|
|              | SMOTE   | RF          | 17.07  | 98.14  | 90.75  | 0.246|
|              |         | NN          | 8.05   | 99.49  | 91.16  | 0.2  |
|              | MMPCC   | RF          | 18.18  | 99.44  | 91.29  | 0.351|
| 30           |         | NN          | 30.38  | 99.16  | 92.27  | 0.464|
|              | SMOTE   | RF          | 17.69  | 97.95  | 90.08  | 0.25 |
|              |         | NN          | 10.16  | 99.23  | 90.5   | 0.216|
|              | MMPCC   | RF          | 19.09  | 99.43  | 90.75  | 0.363|
| 40           |         | NN          | 35.94  | 99.08  | 92.26  | 0.513|
|              | SMOTE   | RF          | 18.54  | 97.8   | 89.5   | 0.258|
|              |         | NN          | 12.07  | 99.14  | 90.02  | 0.243|
|              | MMPCC   | RF          | 19.56  | 99.39  | 90.16  | 0.367|
| 50           |         | NN          | 38.82  | 99.13  | 92.15  | 0.543|
|              | SMOTE   | RF          | 18.5   | 97.72  | 88.9   | 0.258|
|              |         | NN          | 14.05  | 99.01  | 89.55  | 0.266|

the SMOTE method. For other three evaluation metrics, the MMPCC method outperforms the SMOTE sampling method as well. As for RF classifier, Table 4 shows that the performance of MMPCC is better than that of the SMOTE method. As shown in Table 4, with the increase of percentage, the MCC value of the MMPCC in the RF classifier are 0.333, 0.337, 0.351, 0.363 and 0.367, respectively, and the improvements are 0.098, 0.091, 0.101, 0.105 and 0.109, respectively over the SMOTE method. This is due to the fact that a number of duplicated or artificial samples were introduced by over-sampling techniques for large-scale imbalanced data. But for MMPCC, there is no man-made duplicated data. In terms of the MMPCC sampling method, the pseudo-negative sampling technique helps identify more useful samples from the negative class which is often neglected, so it performs better than the SMOTE sampling method.

Experiment 5: experiments on highly imbalance ratio datasets
In order to validate the performance of the proposed method on highly imbalance Ratio datasets, the comparative evaluation on two UCI datasets, Solar Flare and Oil, are performed. The dataset Solar Flare contains 69 minority classes and 1320 majority classes; with 10 attributes, and the Ratio is 19.1. The Oil dataset contains 41 minority classes and 196 majority classes; with 10 attributes, and the Ratio is 10.4. The experiment results show that the performance of MMPCC is better than that of the SMOTE sampling method.
Table 5: Classification results of the Solar Flare dataset with highly imbalance Ratio

| Percentage | Classifier | Sen(%) | Spe(%) | Acc(%) | MCC   |
|------------|-----------|--------|--------|--------|-------|
| 0          | RF        | 1.43   | 99.02  | 94.24  | 0.01  |
|            | NN        | 7.25   | 96.90  | 92.51  | 0.05  |
| 10         | RF        | 4.00   | 99.16  | 94.02  | 0.06  |
|            | NN        | 8.01   | 96.73  | 91.94  | 0.06  |
| 20         | RF        | 13.53  | 99.39  | 94.31  | 0.23  |
|            | NN        | 20.88  | 97.63  | 93.09  | 0.24  |
| 30         | RF        | 25.03  | 99.08  | 94.39  | 0.39  |
|            | NN        | 32.03  | 97.08  | 92.95  | 0.33  |
| 40         | RF        | 20.68  | 98.92  | 93.59  | 0.32  |
|            | NN        | 28.33  | 97.64  | 91.79  | 0.28  |
| 50         | RF        | 32.57  | 98.84  | 93.95  | 0.44  |
|            | NN        | 35.48  | 97.05  | 92.51  | 0.38  |

Classes and 896 majority classes, including 49 attributes, and the Ratio is 21.9.

Table 5 demonstrates the classification results of the Solar Flare dataset with highly imbalance Ratio. Overall, the performance is increased with a larger percentage of pseudo-negative samples. For example, the random forest method obtain 1.43%, 4.00%, 13.53%, 25.03%, 20.68% and 32.57% for Sen as the percentage of pseudo-negative samples is fixed to 0%, 10%, 20%, 30%, 40% and 50%, respectively. Moreover, the MCC value is 0.01, 0.06, 0.32 and 0.44. For the neural networks method, the Sen values are increased from 7.25%, 20.88%, 32.03%, 28.33% to 35.48% and the MCC values are increased from 0.05, 0.24 to 0.38 on different percentage of pseudo-negative samples. We can conclude that the performances of different evaluation metrics show a significant improvement with a higher percentage of pseudo-negative samples, even in the situation of highly imbalance Ratio.

Table 6: Classification results of the Oil dataset with highly imbalance Ratio

| Percentage | Classifier | Sen(%) | Spe(%) | Acc(%) | MCC   |
|------------|-----------|--------|--------|--------|-------|
| 0          | RF        | 14.50  | 99.68  | 96.07  | 0.63  |
|            | NN        | 52.18  | 98.90  | 96.83  | 0.58  |
| 10         | RF        | 19.60  | 99.55  | 95.74  | 0.32  |
|            | NN        | 51.95  | 98.54  | 96.37  | 0.54  |
| 20         | RF        | 33.53  | 98.98  | 95.60  | 0.43  |
|            | NN        | 41.26  | 98.65  | 95.72  | 0.48  |
| 30         | RF        | 39.83  | 98.99  | 95.65  | 0.50  |
|            | NN        | 45.83  | 98.29  | 95.32  | 0.51  |
| 40         | RF        | 50.36  | 99.31  | 96.26  | 0.63  |
|            | NN        | 54.96  | 97.83  | 95.08  | 0.55  |
| 50         | RF        | 49.76  | 98.75  | 95.52  | 0.59  |
|            | NN        | 48.09  | 97.84  | 94.58  | 0.51  |

Table 6 demonstrates the classification results of the Oil dataset with highly imbalance Ratio. From the Table 6, the random forest method achieves 14.50%, 19.60%, 33.53%, 39.83%, 50.36% and 49.76% for Sen when the percentage of pseudo-negative samples is fixed to 0%, 10%, 20%, 30%, 40% and 50%, respectively. In addition, the MCC value is 0.27, 0.32, 0.50, 0.63 and 0.59. For the neural networks method, the Sen values are increased from 52.18%, 51.95%, 41.26%, 45.83%, 54.96% to 48.09% and the MCC values are increased from 0.58, 0.54, 0.48, 0.51, 0.55 to 0.51 with different percentage of pseudo-negative samples. It indicates that the proposed method is prone to improve the discrimination of minority class while retains the considerable stability.

Furthermore, Fig. 6 shows the classification performance on the Solar Flare dataset under different percentage of pseudo-negative samples. From Fig. 6(a), the Sen metric of neural network increase with the percentage of pseudo-negative samples changing from 0% to 50%. Even
there is little fluctuation from 40% to 50%. It maybe the distribution of original dataset is unclear. In the future, we will consider how to choose the percentage of pseudo-negative samples automatically. For MCC performance, similar phenomenon can be obtained from Fig. 6(b).

Figure 7 shows the tendency of Oil dataset with highly imbalance Ratio in neural network and random forest classification. We can see that Sen and MCC of random forest gradually increase when the percentage changes from 0% to 50% in Fig. 7(a). However, the value of Sen and MCC of neural network has some fluctuate from 0% to 50%. It indicated that random forest is more stability of the proposed method for this dataset. Similar trends of MCC performance can be obtained from Fig. 7(b).

Discussion
Here we designed a supervised learning method based on max-relevance and min-redundant criterion beyond Pearson correlation coefficient and tested on four UCI datasets and three real bioinformatics datasets. Our results indicated that MMPCC is better than other sampling methods in terms of several evaluation metrics. The performance of different evaluation metrics shows a trend of increasing with a higher percentage of pseudo-negative samples. On the other hand, different machine learning method has different character, so the experiment results have little instability. We also observed that MMPCC method can have good performance even in the situation of highly imbalance Ratio. This reveals that pseudo-negative samples are good at solving the imbalance dataset problem.

Conclusions
In this study, we propose a new sampling method, which is called pseudo-negative sampling, to handle the imbalanced classification problem based on Pearson correlation coefficient which integrates the max-relevant and min-redundant. In addition, an incremental searching method is used to find the target sample with little cost of computation. The experimental results demonstrate the superior performance of our method compared to other algorithms for imbalanced classification problems.

In future, we will apply the proposed MMPCC algorithm in more real-world bioinformatic applications with large-scale imbalanced data. We will investigate the possibility of extending the MMPCC method to handle multiple-classification problem. Furthermore, we will use the state-of-the-art machine learning methods [41–46] to handle the imbalanced classification problem.

Abbreviations
Acc: Accuracy; CMC: Contraceptive method choice; FN: The number of false negative; FP: The number of false positive; MAXR: Max-relevance method; MCC: Mathews correlation coefficient; MINR: Min-redundancy method; MMPCC: Max-relevance and min-redundancy Pearson correlation coefficient; PCC: Pearson correlation coefficient; Pre: Precision; PSSM: Position specific scoring matrix; Sen: Sensitivity; Spe: Specificity; TN: The number of true negatives; TP: The number of true positive

Acknowledgements
Not applicable.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 20 Supplement 25, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-25.

Authors’ contributions
Zhang YQ, Qiao SJ and Zhou JL conceived the study and developed the approach. Lu RZ, Han N and Liu DX carried out and optimized the experiments. All authors contributed to result interpretation. All authors contributed to the drafting and revision of the manuscript. All authors read and approved the final manuscript.

Funding
Publication costs are funded by the National Natural Science Foundation of China under Grants (Nos. 61702058, 61772091, 61802035, 71701026, 61962006), the China Postdoctoral Science Foundation funded project (No. 2017M612948), the Scientific Research Foundation for Education Department of Sichuan Province under Grant (No. 18ZA0098), the Innovative Research Team Construction Plan in Universities of Sichuan Province under Grant (No. 18TD0027), the Natural Science Foundation of Guangxi under Grant (No. 2018GXNSFDA188005), the Sichuan Science and Technology Program under Grant (Nos. 2018JY0448, 2019FG0106, 2019YS0067), the Scientific Research Foundation for Advanced Talents of Chengdu University of Information
Technology under Grant (Nos. KJY201717, KJY201715, KJY201750), the Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology under Grant (Nos. J201706, J201701), Guangdong Province Key Laboratory of Popular High Performance Computers under Grant (No. 2017B030314073).

Availability of data and materials
All relevant data are included in this published article and its additional files. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Author details
1. School of Computer Science, Chengdu University of Information Technology, 610225 Chengdu, China. 2. School of Computer Science and Engineering, University of Electronic Science and Technology of China, 610054 Chengdu, China. 3. School of Software Engineering, Chengdu University of Information Technology, 610225 Chengdu, China. 4. Software Automatic Generation and Intelligent Service Key Laboratory of Sichuan Province, Chengdu University of Information Technology, 610225 Chengdu, China. 5. School of Management, Chengdu University of Information Technology, 610103 Chengdu, China. 6. School of Cybersecurity, Chengdu University of Information Technology, 610225 Chengdu, China.

Published: 24 December 2019

References
1. Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big data bioinformatics. J Cell Physiol. 2014;219(2):1896–900.
2. Greene AC, Giffen KA, Greene CS, Moore JH. Adapting bioinformatics curricula for big data. Brief Bioinforma. 2015;17(1):43–50.
3. Zhang Y, Cao X, Sheng Z. Genemo: a search engine for web-based functional genomic data. Nucleic Acids Res. 2016;44(Web Server issue):122–7.
4. Zhang Y, Pu Y, Zhang H, Cong Y, Zhou J. An extended fractional kalman filter for inferring regulatory networks using time-series data. Chemometrics Intell Lab Syst. 2014;138:57–63.
5. Liu B, Weng F, Huang D-S, Chou K-C. iro-3wpseknc: Identify dna replication origins by three-window-based pseknc. Bioinformatics. 2018;34(18):3086–93.
6. Chuan G, Ma H, Yan J, Chen M, Hong N, Xue D, Zhou C, Chen C, Chen K, Duan B, et al. Deepcrisp: optimized crisp guide ma design by deep learning. Genome Biol. 2018;19(1):80.
7. Liu B, Yang F, Huang D-S, Chou K-C. ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc. Bioinformatics. 2017;34(1):33–40.
8. Yuan L, Zhu L, Guo W-L, Zhou X, Zhang Y, Huang Z, Huang D-S. Nonconvex penalty based low-rank representation and sparse regression for egt training. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2017;14(5):1154–64.
9. Zhu L, Zhang H-B, Huang D-S. Direct auc optimization of regulatory motifs. Bioinformatics. 2017;33(14):243–51.
10. Yu H, Ni J, Zhao J. Acsoocam: An ant colony optimization-based undersampling method for classifying imbalanced dna microarray data. Neurocomputing. 2013;101(2):309–18.
11. Deng S, Yuan J, Huang D, Wang Z. Staps: An r package for structure/function analysis of protein sequences based on informational spectrum method. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine. Washington, DC, IEEE. 2014. p. 29–34.
12. Zhang Y, Qiao S, Ji S, Zhou J. Ensemble-cnn: Predicting dna binding sites in protein sequences by an ensemble deep learning method. In: 14th International Conference on Intelligent Computing. Berlin: Springer-Verlag. 2018. p. 301–6.
13. Hu J, Li Y, Zhang M, Yang X, Shen HB, Yu DJ. Predicting protein-dna binding residues by weightedly combining sequence-based features and boosting multiple svms. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14(6):1389–98.
14. Zhao Z, Peng H, Lan C, Zheng Y, Fang L, Li J. Imbalance learning for the prediction of de-methylation sites in mmr. BMC Genomics. 2018;19(1):574.
15. Du X, Yao Y, Yao D, Zou H, Zhang Y, Li S. Deepic: Exploring splice site motif through convolutional neural network directly from dna sequence. IEEE Access. 2018;6:32958–78.
16. Liu B, Li J, Cairns MJ. Identifying mmr, targets and functions. Brief Bioinforma. 2012;13(1):1–19.
17. Zhang Y, Zhang D, Mi G, Ma D, Li G, Guo Y, Li M, Zhu M. Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. Comput Biol Chem. 2012;36(2):36–41.
18. Zhu L, Deng S-P, You Z-H, Huang D-S. Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2017;14(2):345–52.
19. Huang D-S, Zhang L, Han K, Deng S, Yang K, Zhang H. Prediction of protein-protein interactions based on protein-protein correlation using least squares regression. Curr Protein Peptide Sci. 2014;15(6):553–60.
20. You Z-H, LeY-K, Gao J, Huang D-S, Zhou X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics. 2010;26(21):2744–51.
21. Xia J-F, Zhao X-M, Huang D-S. Predicting protein–protein interactions from protein sequences using meta predictor. Amino Acids. 2010;39(5):1595–9.
22. Shen Z, Bao W, Huang D-S. Recurrent neural network for predicting transcription factor binding sites. Sci Rep. 2018;8(1):15270.
23. Guo W-L, Huang D-S. An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency. Mole BioSystems. 2017;13(9):1827–37.
24. Dan Y, Xu S, Yang W, Sun C, Yu H. A review of class imbalance learning methods in bioinformatics. Curr Bioinforma. 2015;10(4):360–9.
25. Guo H, Li Y, Shang J, Gu M, Huang Y, Gong B. Learning from class-imbalanced data. Review of methods and applications. Expert Syst Appl. 2017;73:220–39.
26. Liu B, Li K, Huang D-S, Chou K-C. enhancer-el: identifying enhancers and their strength with ensemble learning approach. Bioinformatics. 2018;34(22):3835–42.
27. Hassan AR, Haque MA. An expert system for automated identification of obstructive sleep apnea from single-lead ecg using random under sampling boosting. Neurocomputing. 2017;235:122–30.
28. Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans Neural Netw Learn Syst. 2017;28(9):4048–57.
29. Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans Neural Netw Learn Syst. 2017;29(8):3573–87.
30. Jin X, Bo T, He H, Hong M. Semisupervised feature selection based on relevance and redundancy criteria. IEEE Trans Neural Netw Learn Syst. 2016;28(9):1974–84.
31. Pons T, Vazquez M, Mateyherandez ML, Brunak S, Valencia A, Izazouaga JM. Kinmutrf: a random forest classifier of sequence variants in the human protein kinase superfamily. BMC Genomics. 2016;17(2):396.
32. Wang X, Lin P, Ho JW. Discovery of cell-type specific dna motif grammar in cis-regulatory elements using random under sampling boosting. Bioinformatics. 2010;26(21):2744–51.
33. Dutta S, Madan S, Parikh H, Sundar D. An ensemble micro neural network approach for elucidating interactions between zinc finger proteins and their target dna. BMC Genomics. 2016;17(Suppl 13):1033.
34. GutiRrez-Tobal GC, lvarezD, DelCF, HorneroR. Utility of adaboost to detect sleep apnea-hypopnea syndrome from single-channel airflow. IEEE Trans Biomed Engineer. 2015;63(3):636–46.
35. Jin X, Zhao M, Chow TWS, Pecht M. Motor bearing fault diagnosis using trace ratio linear discriminant analysis. IEEE Trans Ind Electron. 2013;60(3):1244–1251.
36. Asuncion A. Ucimachine learningrepository. 2013. https://archive.ics.uci.edu/ml/index.php.
37. Quan Z, Guo M, Yang L, Jun, Wang. A classification method for class-imbalanced data and its application on bioinformatics. J Comput Res Dev. 2010;47(8):1407–14.
38. Hu J, Li Y, Zhang M, Yang X, Shen HB, Yu DJ. Predicting protein-dna binding residues by weightedly combining sequence-based features and boosting multiple svms. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14(6):1389–98.

39. Si J, Zhang Z, Lin B, Schroeder M, Huang B. Metadbite: a meta approach to improve protein dna-binding sites prediction. BMC Syst Biol. 2011;5(1):7.

40. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16(1):321–57.

41. Qiao S, Han N, Wang J, Li R, Gutierrez LA, Wu X. Predicting long-term trajectories of connected vehicles via the prefix-projection technique. IEEE Trans Intell Trans Syst. 2018;19(7):2305–15.

42. Qiao S, Shen D, Wang X, Han N, Zhu W. A self-adaptive parameter selection trajectory prediction approach via hidden Markov models. IEEE Trans Intell Trans Syst. 2015;16(1):284–96.

43. Qiao S, Han N, Zhu W, Gutierrez LA. TraPlan: an effective three-in-one trajectory-prediction model in transportation networks. IEEE Trans Intell Trans Syst. 2015;16(3):1188–98.

44. Qiao S, Han N, Gao Y, Li R, Huang J, Guo J, Gutierrez LA, Wu X. A fast parallel community discovery model on complex networks through approximate optimization. IEEE Trans Knowl Data Engin. 2018;30(9):1638–51.

45. Qiao S, Tang C, Jin H, Long T, Dai S, Ku Y, Chau M. PutMode: prediction of uncertain trajectories in moving objects databases. Appl Intell. 2010;33(3):370–86.

46. Qiao S, Han N, Zhou J, Li R, Jin C, Gutierrez LA. Socialmix: A familiarity-based and preference-aware location suggestion approach. Engineer Appl Artif Intell. 2018;68:192–204.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.