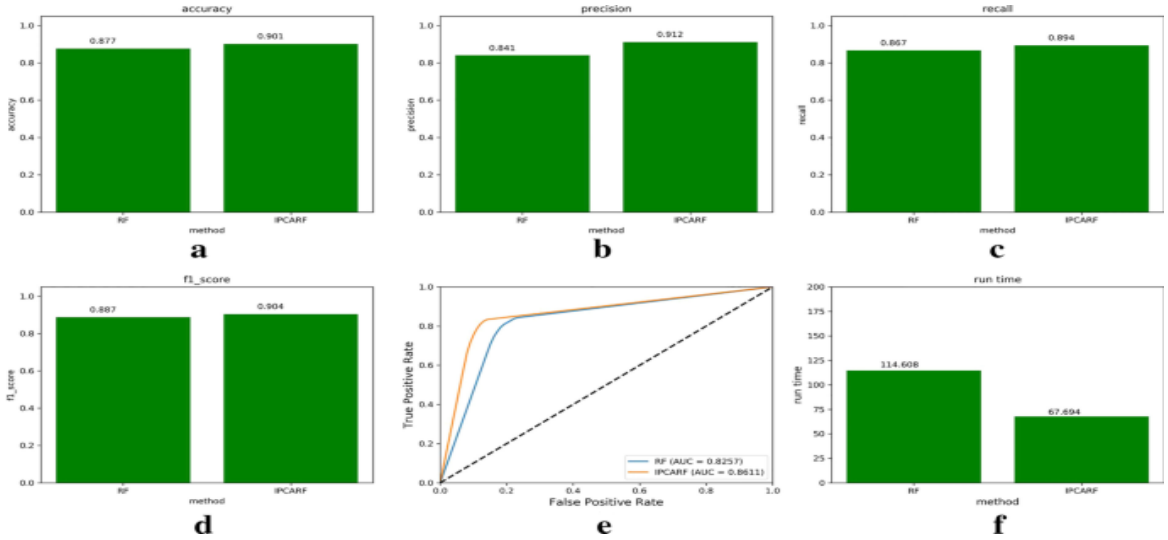
IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier

Discussion

The parameters used in IPCARF also affect its prediction performance. We have performed many experiments and found that, except for the n_estimators parameter, changes in the other parameters have relatively little impact on the prediction results of IPCARF. Therefore, here, we consider only the influence of the n_estimator parameter on the prediction results of the IPCARF algorithm. In this experiment, we set the value range of the n_estimators parameter to [100, 500, 1000, 1500, 2000, and 2500]; then, we selected the optimal n_estimator parameter value using the grid search (GS) method.

The grid search method is a commonly used parameter optimization algorithm. A grid search is a method of finding parameters. Its core principle is to first define the parameter area to be searched and then divide the area into grids. The intersections in the grid form the parameter combinations to be searched. In other words, all the intersections in the grid are parameter combinations (c, g) that should be searched, and each combination (c, g) is retrieved during the grid search process. To obtain the best (c, g) combination, the k-fold method is used to test the classification accuracy of each group (c, g), and the group with the highest accuracy among all selected (c, g) is selected as the parameters for building the model.

Fig. 5



Comparison of prediction results of IPCARF and algorithms (n_estimators = 1500) (this figure is generated in the Python language environment with a 3.7 version)

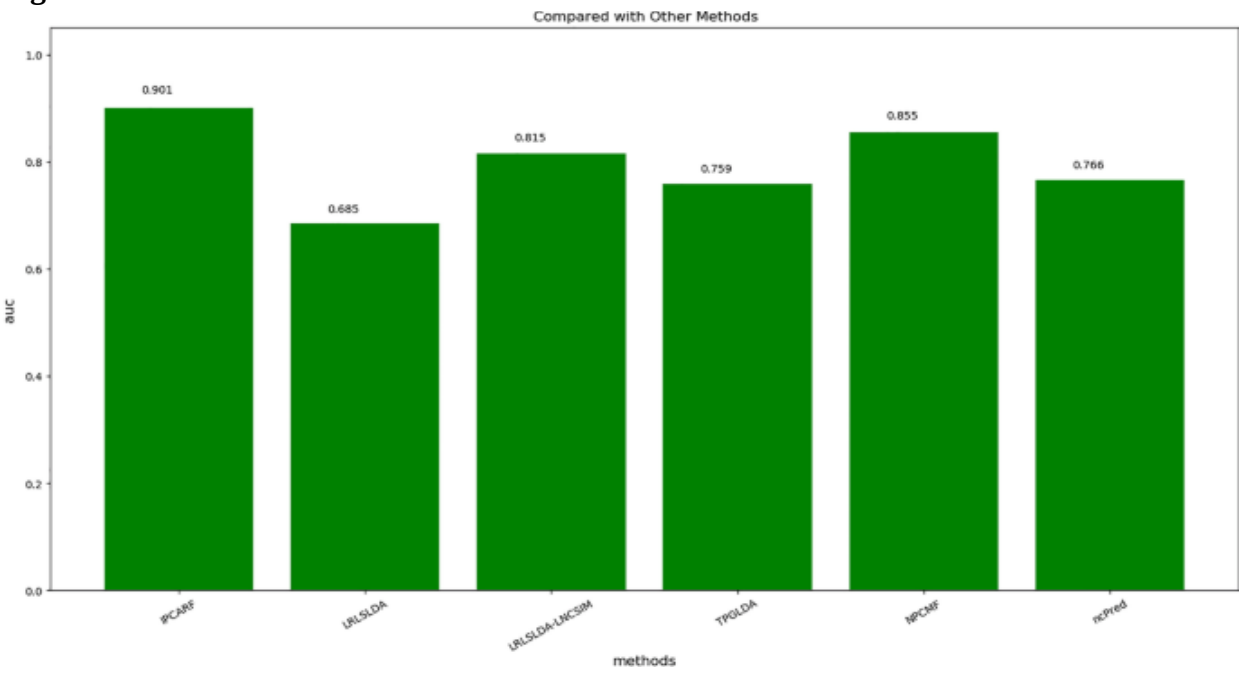
In our experiment, the best n_estimator parameter value found after executing the GS algorithm was 1500. Thus, we adopt n_estimators = 1500 to further compare the execution performances of the IPCARF and RF algorithms. The experimental results are shown in Fig. 5.

Figure 5a–d displays the accuracy, precision, recall, and F1-score results of the two algorithms. Figure 5e shows that the AUC values predicted by the two algorithms are 0.8257 and 0.8611, and Fig. 5f shows that the running time of the IPCAFR algorithm is significantly lower than that of the RF algorithm.

Comparisons with existing works

Previous scholars have developed many effective prediction methods for the prediction of lncRNA-disease associations. However, because the data themselves have problems such as instability and because the evaluation methods used by various methods are inconsistent, the current methods still leave considerable room for improvement. To further verify the effect of IPCARF, we compared it with five other existing works, including LRLSLDA, LRLSLDA-LNCSIM, TPGLDA, NPCMF, and ncPred.The comparison results showing the AUC values of these algorithms are shown in Fig. 6.

Fig. 6



Comparisons with existing work (this figure is generated in the Python language environment with a 3.7 version)

Figure 6 shows that the AUC value obtained when using the IPCARF method to predict lncRNA-disease associations is better than that of the other comparison algorithms. Because of the instability of genetic data, the results of each experimental run differ to some degree. Consequently, we repeated the experiment 10 times and took the average as the final result. In the experiment, the highest value of AUC obtained when running the IPCARF algorithm was 0.906, and the lowest value was 0.861. These experimental results indicate that the prediction performance of the IPCARF method is slightly better than that of the comparative methods.

Case study

Lung cancer is a common malignant lung tumor. The top 5 long non-coding RNAs that use the IPCARF algorithm to predict lung cancer are: GAS5, XIST, CDKN2B-AS1, PVT1 and HOTAIR. Four of the top 5 have the latest literature to verify. Ranked No. 1 is GAS5, and the research in the literature shows that GAS5 may play a role in suppressing cancer. Ranked No. 2 is XIST, and the research in the literature shows that XIST plays an important regulatory role in cancer biology. Ranked No. 4 is PVT1, and the research in the literature shows that PVT1 can inhibit cell proliferation, migration and invasion. Ranked No. 5 is HOTAIR, and the literature found that HOTAIR affects the drug resistance of small cell lung cancer cells by regulating the methylation of HOXA1.

Conclusions

In this study, we proposed a novel model called IPCARF to predict lncRNA-disease associations and compared it with the existing LRLSLDA, LRLSLDA-LNCSIM, TPGLDA, NPCMF, and ncPred prediction methods using 10CV. These methods have achieved excellent performances for predicting lncRNA-disease associations. The comparison results show that the prediction results of the IPCARF method are better than those of the compared methods.

Although the IPCARF method has achieved good prediction results, it still has some limitations that should be improved in future studies. First, the experimental data are still not rich enough, which limits the predicted results. As more data related to lncRNA diseases becomes available, the IPCARF

method will improve. The complexity and inconsistency of biological data also cause certain difficulties in improving and comparing algorithms, especially the inability to obtain completely consistent data sources. In future work, we will consider integrating data from different sources to improve the prediction performance of IPCARF by improving the integrity and quality of the experimental data.

Methods

Data collection

Disease similarity data

The data on disease similarity compiled by different scientific researchers are not the same. Among them, the data compiled by van Driel et al. is the most often cited; it is also the most recognized and is considered to be relatively authoritative disease similarity data. A similarity network of 5080 human genetic diseases is constructed by this database.

lncRNA-disease association data

In 2013, Chen et al. Established the LncRNA disease database, which was the first database of lncRNA-disease association data, and it was manually collected and experimentally verified. Over time and the continuous expansion of lncRNA research, the LncRNA Disease database has also continuously expanded, and the number of entries increases yearly. In this study, we used the v2017 data from the LncRNA disease database. The datasets generated and analyzed during the current study are presented in additional file 1.

Disease semantic similarity

Disease semantic similarity model

Referring to the calculation method, two models are used on the directed acyclic graph (DAG) of diseases to compute a disease semantic similarity score.

First, the contribution of the disease term t in DAG(D) to the semantic value of disease D is defined as follows:

$$sim1(D1,D2)=$$

$$\sum i \in Disease(D1) \cap Disease(D2)(C1D1(i)+C1D2(i)) / C1(D1)+C1(D2)$$

where *sim*1 denotes the disease semantic similarity matrix.

Moreover, the method for calculating disease similarity refers to the calculation method, which provides a detailed description.

Gaussian interaction profile kernel similarity for disease

Similar diseases may have similar related lncRNAs. The similarity of Gaussian interaction kernels can be computed from the known lncRNA-disease association network. The Gaussian interaction kernel similarity between diseases *D*1 and *D*2 is computed as follows:

$$GKS(D_1, D_2) = \exp(-k_{dis}\|D_1 - D_2\|^2)$$

Where $-k_{dis}$ represents the standardized core width, which is calculated by

$$k_{dis} = 1/\frac{1}{m}\sum_{i=1}^{m}\|D(i)\|^2, kdis = \frac{1}{m}\sum_{i=1}^{m}\|D(i)\|2$$

Where m represents the disease number.

Gaussian interaction profile kernel similarity for lncRNA

The Gaussian interaction kernel similarity between lncRNAs L_1L1 and L_2L2 is computed as:

$$GKS(L_1, L_2) = \exp(-k_{lnc}\|L_1 - L_2\|^2), GKS(L1, L2) = \exp(-klnc\|L1 - L2\|2),$$

$$k_{lnc} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\|L(i)\|^2}, klnc = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\|L(i)\|2},$$

Where *n* represents the lncRNA number.

The PCA algorithm

Principal Component Analysis (PCA) is a commonly used data analysis algorithm and an unsupervised linear feature extraction algorithm. PCA has been widely used in applications such as lossy data compression, feature selection, and dimensionality reduction .PCA methods can reduce data from a high-dimensional space to a low-dimensional space because it merges similar features due to the variance. Thus, PCA can reduce both data and the number of data features, which helps to prevent model over fitting.

The main idea underlying the PCA algorithm is to describe things using fewer data features that represent most of the main information. PCA is a statistical method that recombines characteristic variables with linear associations into fewer characteristic variables. The PCA algorithm is essentially a transformation of the variables that introduces a set of new variables that are not related to the original variables; instead, these new variables are linear functions of the original variables. Each new variable is called a principal component. This group of principle is sorted based on variance; the first principal component is the one with the largest

variance in the linear function. The second principal component is the linear function with the second-largest variance, and the first and second principal components are not correlated with each other. The third principal component is also uncorrelated with the first and second principal components and constitutes the linear function with the third-largest variance. By analogy, the original data are transformed using K−LK−L to obtain new data after dimensionality reduction.

The RF classification algorithm

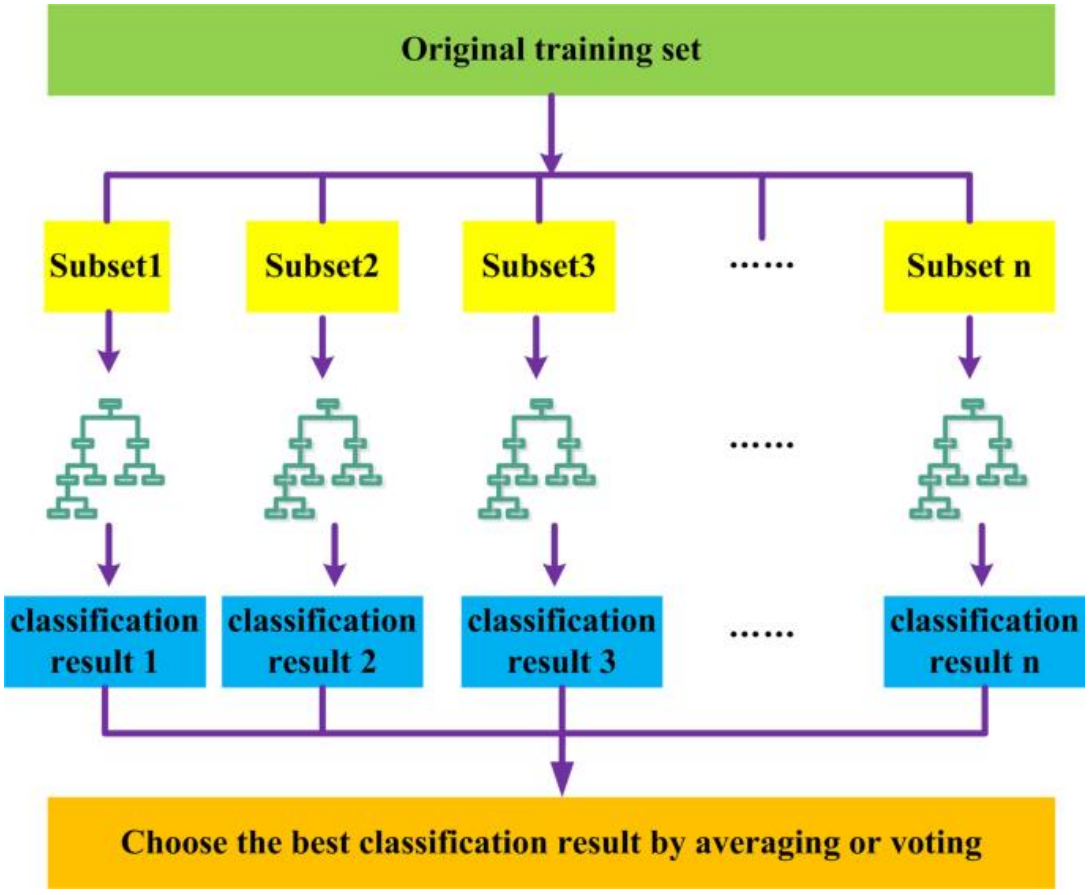
The RF classification algorithm belongs to the supervised learning subfield of the machine learning field. It uses samples from a dataset for training and the trained model is applied to perform predictions on real data to evaluate whether the results meet expectations.

The traditional classification algorithms mainly include k-nearest neighbor (KNN) algorithms, naive Bayes (NB) algorithms, decision tree algorithms and support vector machine (SVM) algorithms. Most of these algorithms are relatively mature, and each has a range of suitable application scenarios, but they also leave space for corresponding algorithm classification performance improvements. The decision tree algorithm is a type of split tree approach based on data attribute characteristics. As research has deepened, improved decision trees such as ID3, C4.5, classification and regression tree (CART), and regression trees have gradually been developed. The decision tree algorithm has advantages such as an easy way to understand the decision results and powerful functions, but it may exhibit problems such as weak fitting. The NB algorithm comes from the field of statistics and predicts the posterior probability based on the prior probability. The advantage of the NB algorithm is its fast calculation speed, while its disadvantage is that there may be dependencies between attributes, which often leads to lower classification accuracy. The SVM algorithm performs high-dimensional and nonlinear classification by constructing a hyper plane. The advantages of the SVM algorithm are that it is highly efficient and provides good classification accuracy. Its disadvantages are the complex structure of its kernel function and a lack of data sensitivity.

The RF method, first proposed by Breiman is a machine learning algorithm consisting of many decision trees. It is a combination of the Bagging and Random Subspaces methods. The RF algorithm is considered to be an ensemble learning and supervised classification method. It first randomly establishes a forest composed of multiple unrelated decision trees; these multiple decision tree classifier models each learn and perform prediction separately. Then, the prediction results of the multiple decision tree classifier models are combined to obtain a final prediction result. There are two typical ways of combining the prediction results from different decision tree classifiers in RF. One is to average the prediction results of all the decision tree classifiers to obtain a prediction result for the entire forest. The other is to conduct a vote on the prediction results from all decision tree

classifiers to select an optimal prediction result as the prediction result of the entire forest. A general flowchart of the RF algorithm is shown in Fig. 7.

Fig. 7



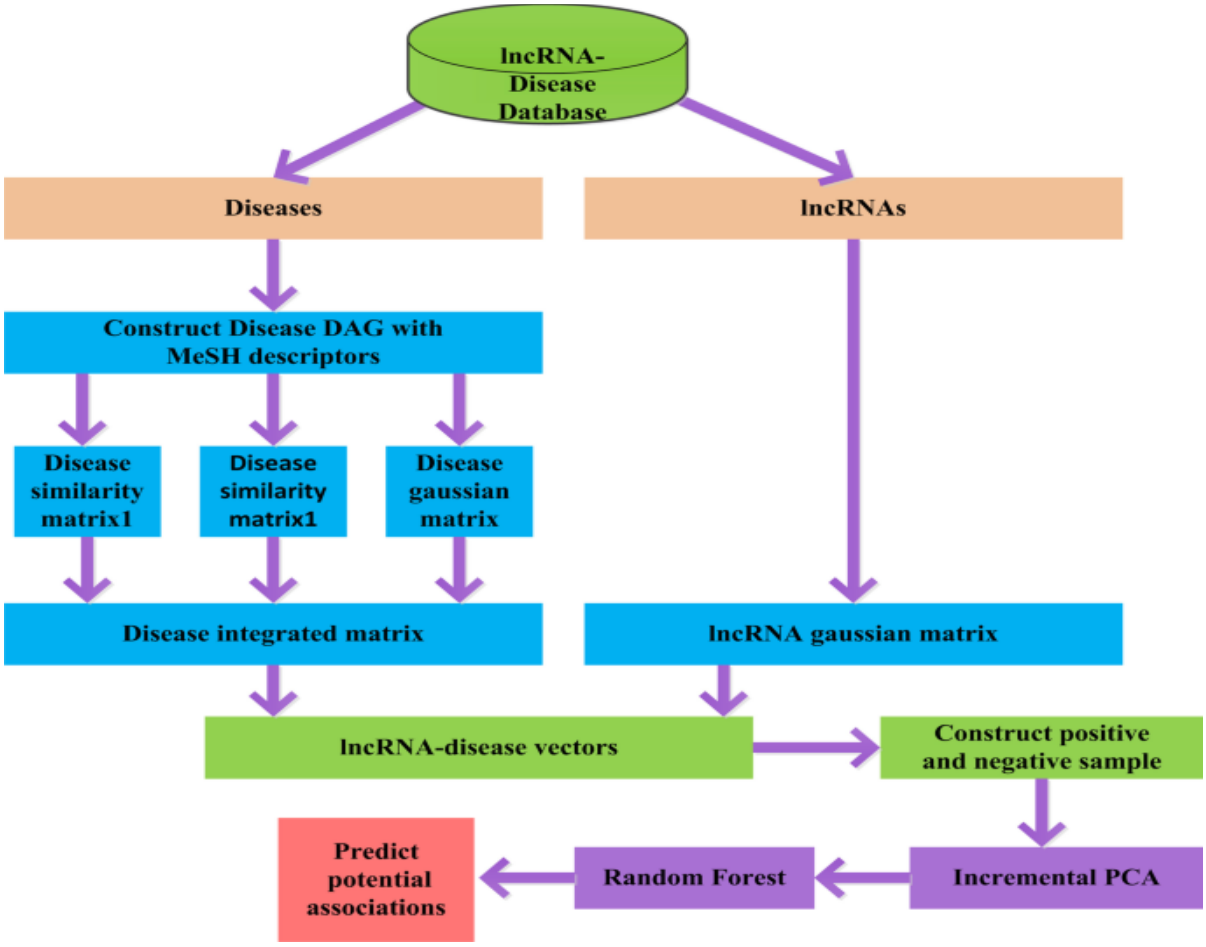
The RF algorithm first selects n samples from the original training set as a training subset and then generates a decision tree for each subset. The above steps are repeated a total of n times to generate n decision trees that form the random forest. Finally, the random forest obtained by training is used to predict test samples, and an optimal classification result is selected using either the mean method or the voting method.

Long noncoding RNA-disease prediction based on IPCA and RF

In this study, we developed an algorithm called IPCARF based on the IPCA and RF methods. First, two semantic similarity matrices, a Gaussian kernel similarity matrix for diseases and a Gaussian kernel similarity matrix for lncRNAs are established. Second, a feature vector is extracted from the similarity matrix to construct an adjacency matrix. Then, the positive samples and negative samples are extracted from the adjacency matrix to construct the dataset for prediction. Next,

the IPCA method is applied to select features and reduce the dataset dimensionality. Finally, the FR classifier is used to make predictions. The IPCARF process is shown in Fig. 8.

Fig. 8



Abbreviations

IPCA: Incremental principal component analysis
RF: Random forests
LncRNA: Long non-coding RNA
DAG: Directed acyclic graph
GS: Grid search
AUC: Area under curve
SVM: Support vector machine
LRLSLDA: Laplacian Regularized Least Squares for LncRNA-Disease Association.
LNCSIM: lncRNA functional similarity calculation models
TPGLDA: lncRNA-disease gene tripartite graph
NPCMF: Nearest profile-based collaborative matrix factorization
ncPred: ncRNA-disease association prediction.