IVYSPRING
INTERNATIONAL PUBLISHER

International Journal of Biological Sciences

2019; 15(13): 2911-2924. doi: 10.7150/ijbs.33806

Research Paper

GBDTCDA: Predicting circRNA-disease Associations Based on Gradient Boosting Decision Tree with Multiple Biological Data Fusion

Xiujuan Lei✉*, Zengqiang Fang

School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

*Member, IEEE

✉ Corresponding author: xjlei@snnu.edu.cn

Received: 2019.02.03; Accepted: 2019.10.15; Published: 2019.11.08

Abstract

Circular RNA (circRNA) is a closed-loop structural non-coding RNA molecule which plays a significant role during the gene regulation processes. There are many previous studies shown that circRNAs can be regarded as the sponges of miRNAs. Thus, circRNA is also a key point for disease diagnosing, treating and inferring. However, traditional experimental approaches to verify the associations between the circRNA and disease are time-consuming and money-consuming. There are few computational models to predict potential circRNA-disease associations, which become our motivation to propose a new computational model. In this study, we propose a machine learning based computational model named Gradient Boosting Decision Tree with multiple biological data to predict circRNA-disease associations (GBDTCDA). The known circRNA-disease associations' data are downloaded from cricR2Disease database (http://bioinfo.snnu.edu.cn/CircR2Disease/). The feature vector of each circRNA-disease association pair is composed of four parts, which are the statistics information of different biological networks, the graph theory information of different biological networks, circRNA-disease associations' network information and circRNA nucleotide sequence information, respectively. Therefore, we use those feature vectors to train the gradient boosting decision tree regression model. Then, the leave one out cross validation (LOOCV) is adopted to evaluate the performance of our computational model. As for predicting some common diseases related circRNAs, our method GBDTCDA also obtains the better results. The Area under the ROC Curve (AUC) values of Basal cell carcinoma, Non-small cell lung cancer and cervical cancer are 95.8%, 88.3% and 93.5%, respectively. For further illustrating the performance of GBDTCDA, a case study of breast cancer is also supplemented in this study. Thus, our proposed method GBDTCDA is a powerful tool to predict potential circRNA-disease associations based on experimental results and analyses.

Key words: circRNA-disease associations, machine learning, Gradient Boosting, multiple biological data

Introduction

Circular RNAs (circRNAs) are one kind of short non-coding RNAs [1] which have no exposed terminating 5'-cap and 3'-polyadenylated tail structures and are closed loops, which are unlike the linear RNAs that have terminated with 5′ caps and 3′ tails [2]. It is this closed loops structure that makes more difficult to detect circRNAs in organisms [3-5]. At the same moment, the closed loops structure makes circRNA more stable and conversed to be regarded as a biomarker to mark some diseases. With the development of basic sequence technologies and high-throughput technologies, more and more circRNAs functions are revealed [6]. Many studies have shown that circRNAs can work as the sponges for competing endogenous RNAs or miRNAs [7-9], which makes circRNAs also can be treated as diseases biomarkers. Secondly, circRNAs also have effects on the alternative splicing and transcription process by isolating the translation start position to modulate protein expression [10, 11]. Thirdly, circRNAs are involved in modulating the expression of parental genes [12]. What's more, circRNAs also function as the retinol-binding protein (RBP) sponges, which can strengthen the interaction between the MBL protein and circMbl [10] or get involve in translating as templates. In addition, increasing numbers of evidences and studies have shown that circRNAs play significant roles in disease diagnosis and treatment [13]. Especially for various cancers [14, 15], cardiovascular disease [16], diabetes [17], etc. Current researches of the associations between circRNAs and diseases are relatively advanced works, which might be based on the pathway circRNA-miRNA-mRNA to explore potential knowledge.

Recently, people pay more and more attention to exploring complicated associations between circRNAs and other biological molecules such as circRNA-miRNA, circRNA-lncRNA, and so on. In order to better promote the researches of circRNA, some useful databases are constructed to collect the information of circRNAs, which include circBase (http://www.circbase.org/) [18], Circ2Traits (http://gyanxet-beta.com/circdb/) [19] and circR2Disease (http://bioinfo.snnu.edu.cn/CircR2Disease/) [20]. CircRNA can obtain highly nuclease-resistant ability because of the particular closed loops structures. A stable closed loop structure of circRNAs also helps circRNAs own longer half-lives than the usual linear RNAs [21], which can be regarded as a unique property to diagnose some circRNA-related diseases. Although traditional RNA-Seq techniques are used widely to detect diseases related circRNAs and high-throughput techniques are adopted to validate, which can help us obtain some accurate experimental results, these techniques are still expensive and time-consuming. There are few computational models to detect potential or promising circRNA-disease associations simultaneously, which is also our motivation to develop this study.

In this study, we adopt the gradient boosting decision tree [22] to predict potential circRNA-disease associations, which is named Gradient Boosting Decision Tree with multiple biological data to predict circRNA-disease associations (GBDTCDA). Multiple biological data such as circRNAs related expression profile data, gene ontology (GO) terms data and base sequence are adopted to construct circRNA similarity network (*CSN*). Diseases related ontology terms and genes are involved in building disease similarity network (*DSN*). Then the statistics information of *CSN*, *DSN* and circRNA-disease associations network, the graph theory information of *CSN* and *DSN*, representational biological indicators of circRNA, such as GC content and K-mer, other information like latent vectors extracted from cicrRNA-disease association network are regarded as feature vector to indicate each circRNA-disease pair. Some of those feature vectors are input to train the model and the rest of data are treated as test data. Here, leave one out cross validation (LOOCV) is adopted to evaluate the performance of GBDTCDA. The area under ROC curve (AUC) value of LOOCV is 0.834, which is a better result than other machine learning methods or network-based methods. In order to further illustrate the performance of GBDTCDA, we also make some case studies. Therefore, GBDTCDA is a powerful method to predict the potential circRNA-disease associations.

Materials & Methods

Human circRNA-disease associations

In this study, human disease-related circRNAs are extracted from the initial dataset which is downloaded from CircR2Disease database [23] (http://bioinfo.snnu.edu.cn/CircR2Disease/). All the collected circRNA-disease associations are validated by biological experiments in CircR2Disease database. There are 739 circRNA-disease associations are collected in the database where includes 661 circRNA entries and 100 disease entries, which are composed of the initial dataset. Then, we pick up the distinct 140 cicrRNA-disease associations involving 132 circRNAs and 40 diseases, which could be regarded as a suitable known numbers of circRNA-disease associations in the circRNA-disease associations' matrix. Thus, matrix A is utilized to describe the circRNA-disease associations. If there is an existing association between the circRNA $c(i)$ and disease $d(j)$, $A(c(i), d(j))$ is equal to 1, otherwise it is equal to 0. The related data used is shown in table 1.

Table 1. The number of experimental data in this experiment.

Experimental data	Number
The number of circRNA-disease associations	140
The number of circRNAs	132
The number of diseases	40

CircRNA similarity

CircRNA sequence similarity

To calculate circRNA sequence similarity, the circRNA related sequence data is downloaded from circBase database [18]. There are 132 circRNA sequence data picked up from circBase database based on the circRNA ID of the circR2Disease database. After we get the circRNA sequence data, a sequence alignment algorithm called the Needleman-Wunsch pairwise alignment algorithm is used to calculate the circRNA sequence similarity. Needleman-Wunsch pairwise algorithm is integrated into a python package called Biopython [24]. The parameter gap-open penalty and gap-open extending penalty are set as 2, -0.5 to -0.1 respectively. Then we describe the matrix *Seq_CS* as the circRNA sequence similarity matrix.

CircRNA functional annotation semantic similarity

Gene ontology (GO) annotation data is downloaded from the human protein reference database (HPRD, http://www.hprd.org/) [25] to calculate the circRNA functional annotation semantic similarity. There are 19701 gene ontology data in the initial dataset. Based on circRNA-disease associations' network, 132 circRNAs related GO terms are extracted, which can be utilized to match circRNA-related gene ontology data. In this study, an information content [26] method is adopted to calculate the annotation semantic similarity of circRNA. Thus, we denote the functional annotation semantic similarity of circRNA network as *Fun_CS* and it can be calculated as follows:

$$Fun\_CS(C_i, C_j) = \frac{2 \times logP(C_{i \cup j})}{logP(C_i) + logP(C_j)} \tag{1}$$

where C_i and C_j represent the GO terms which are related to the target genes of circRNA C_i and C_j, $P(C_i)$ and $P(C_j)$ denote the ratio between C_i and C_j target genes related GO terms and the whole GO terms, respectively. $P(C_{i \cup j})$ describes the proportion between the annotated GO terms on circRNA C_i and C_j and the whole GO terms.

CircRNA expression profile similarity

Expression profile data is extracted from the online database exoRBase (http://www.exorbase.org/)[27], where has collected human circRNA and lncRNA related expression profile data. In this study, we replace the circRNA ID in circBase with the ID in exoRbase manually. Then, "Normal_circRNA_RPM" data, including the names and locations of specific chromosomes and the expression profile of circRNAs at 32 sites in normal human body, is downloaded to calculate the circRNA expression profile similarity which is denoted as *ES*. Furthermore, the Pearson correlation coefficient is adopted to measure the relevance between two circRNAs. The greater correlation score they obtain, the more similar two circRNAs are. Considering that the expression profile of two circRNA C_i and C_j can be expressed as C_{e_i} and C_{e_j}. Thus, the coefficient score can be calculated as follows:

$$ES(C_i, C_j) = \frac{\sum_{t=1}^{N}(c_{it} - \bar{c}_i)(c_{jt} - \bar{c}_j)}{\sqrt{\sum_{t=1}^{N}(c_{it} - \bar{c}_i)\sum_{t=1}^{N}(c_{jt} - \bar{c}_j)}} \tag{2}$$

where N is the number of the circRNA expression profile value.

Fusing multiple circRNA similarity

Based on the previous calculations, three circRNA similarity networks have been constructed, including the circRNA sequence similarity network, the circRNA functional annotation semantic similarity network and the circRNA expression profile similarity network. There will be noise information in the integrated network, if only the linear method is used to integrate multiple data. Thus, a similarity network fusion [28] (SNF) algorithm is adopted to combine multiple biological data resources, which can make each network merge better and keep the most informative information.

In this study, W^M is described as the weighted matrix of each related circRNA similarity matrix and M is set to 3 based on three circRNA similarity networks, where $W^M(i,j)$ is equal to the corresponding circRNA i and j similarity score. Given that $\sum_j P(i,j) = 1$ and we need to avert the scale of self-similarity in the diagonal entries. Thus, a better normalization is defined as follows:

$$P(i, j) = \begin{cases} \dfrac{W^M(i, j)}{2\sum_{k \neq i} W^M(i, k)} \text{ if } & i \neq j \\ \dfrac{1}{2} \text{ if } & i = j \end{cases} \tag{3}$$

What's more, a local kernel similarity matrix of each corresponding normalized matrix can be calculated as follows:

$$S(i, j) = \begin{cases} \dfrac{W^M(i, j)}{\sum_{k \in V(i)} W^M(i, k)} & \text{if} \quad j \in V(i) \\ 0 \quad \text{otherwise} & \end{cases} \tag{4}$$

where $V(i)$ is used to describe the K nearest neighbors of circRNA i in the integrated circRNA similarity matrix W^M. Based on the above operation, the lower value neighbors of circRNA i are picked out and the neighbors that owe high values can be preserved, which can be illustrated as an assumption that the local similarities are more reliable than remote ones. Considering that matrix P takes the whole information of the various circRNA similarity information into account, while matrix S only carries K nearest neighbors' information of the network. To obtain the final circRNA similarity network, we apply the following equation to fuse multiple similarities:

$$P_t^{(i)} = S^{(i)} \times \left(\frac{\sum_{l \neq i} P_{t-1}^{(l)}}{M - 1}\right) \times (S^{(i)})^T \text{ where } \quad i = 1, 2, 3 \tag{5}$$

where $P_t^{(i)}$ denotes the results of t iterations of the ith circRNA similarity network and the $S(i)$ is the KNN (local) similarity matrix of the ith similarity matrix P. M is the number of the multiple different circRNA similarity matrices that need to be fused. In this study, M is equal to 3. When each matrix P is stable after t iterations, we fuse different circRNA matrixes as the following equation:

$$P = \frac{1}{M} \sum_{i=1}^{M} P_t^{(i)} \tag{6}$$

Disease similarity

Disease functional similarity

To calculate disease functional similarity, disease-related genes are extracted from DisGeNet (http://www.disgenet.org/) [29] database where has collected 381,056 gene-disease associations (GDAs) between 16,666 genes and 13,172 diseases database and Online Mendelian Inheritance in Man[30] (OMIM, https://www.ncbi.nlm.nih.gov/omim/) database. Based on the processed circRNA-disease associations dataset, 40 diseases are picked out. Thus, we use those 40 diseases to match their related genes from the above two databases. Here, a statistic method JACCARD is adopted to calculate disease functional similarity as follows:

$$DFS(i, j) = \frac{\left|D_g^i \cap D_g^j\right|}{\left|D_g^i \cup D_g^j\right|} \tag{7}$$

where D_g^i and D_g^j represent collections of genes associated with disease i and disease j, respectively.

Disease semantic similarity

There are 40 individual disease entries according to the pre-processed circRNA-disease associations. The 40 names of diseases are adopted to search their corresponding DOID manually on Disease Ontology website [31] (http://www.disease-ontology.org/). Then we use a R package called DOSE [32] to calculate the disease semantic similarity based on their relevant DOID. Thus, DSS is used to describe the disease semantic similarity matrix. Finally, we adopt the following equation to integrate disease functional and semantic similarities:

$$DS(i, j) = a \cdot DFS(i, j) + (1 - a) \cdot DSS(i, j) \tag{8}$$

where a is a weighted coefficient which is used to adjust the proportion of disease functional similarity and disease semantic similarity in the final disease similarity network and the range of a is between 0 and 1. In this study, a is set to 0.5, which means that those two disease similarities are treated equally.

GBDTCDA

In this study, multiple biological data are adopted to engineer the feature vector of each circRNA-disease pair. CircRNA sequence data, circRNA expression profile and circRNA related GO terms are adopted to build the fusion circRNA similarity matrix. Disease related genes and disease phenotypes data are used to build the combined disease similarity matrix. Furthermore, circRNA nucleic acid sequence data is considered to obtain more biological information. The above data is adopted to calculate the statistic feature, graph theory feature and complex biological feature. Then, the principle component analysis (PCA) algorithm is used to extract the more essential features to reduce the noise of feature vectors. Finally, all the processed features are input into the Gradient Boosting Decision Tree machine to predict the potential circRNA-disease associations. The flowchart of our method is shown in Fig 1.

Engineering the feature vector

There are four different kinds of features extracted from circRNA related data which includes the integrated circRNA similarity network and circRNA nucleic acid sequence data, disease related data which contains the integrated disease similarity network and disease-circRNA associations, respectively. Some of these features are constructed in a way that we refer to previous work [33].

To extract the first type of feature for circRNA $c(i)$ or disease $d(j)$, some statistics information of the circRNA similarity matrix P, the disease similarity DS and the circRNA-disease associations matrix A are taken into our consideration. Based on the matrix A, $F_{1.num.nei}$ can be described that the number of $c(i)/d(j)$'s neighbors can be obtained by calculating the sum of the ith/jth column/row in the matrix A. Then, $F_{1.sim.ave}$ can be presented that the average similarity score of circRNA $c(i)$ and disease $d(j)$ can be calculated based on the matrix P and the matrix DS. What's more, we also take the distribution features of $c(i)$ and $d(j)$ similarity scores into account, which is denoted as $F_{1.dis.num}$. The similarity scores [0, 1] can be divided into different distribution intervals, and then we can calculate the distribution number of similarity scores of $c(i)$ and $d(j)$ in each distribution interval.

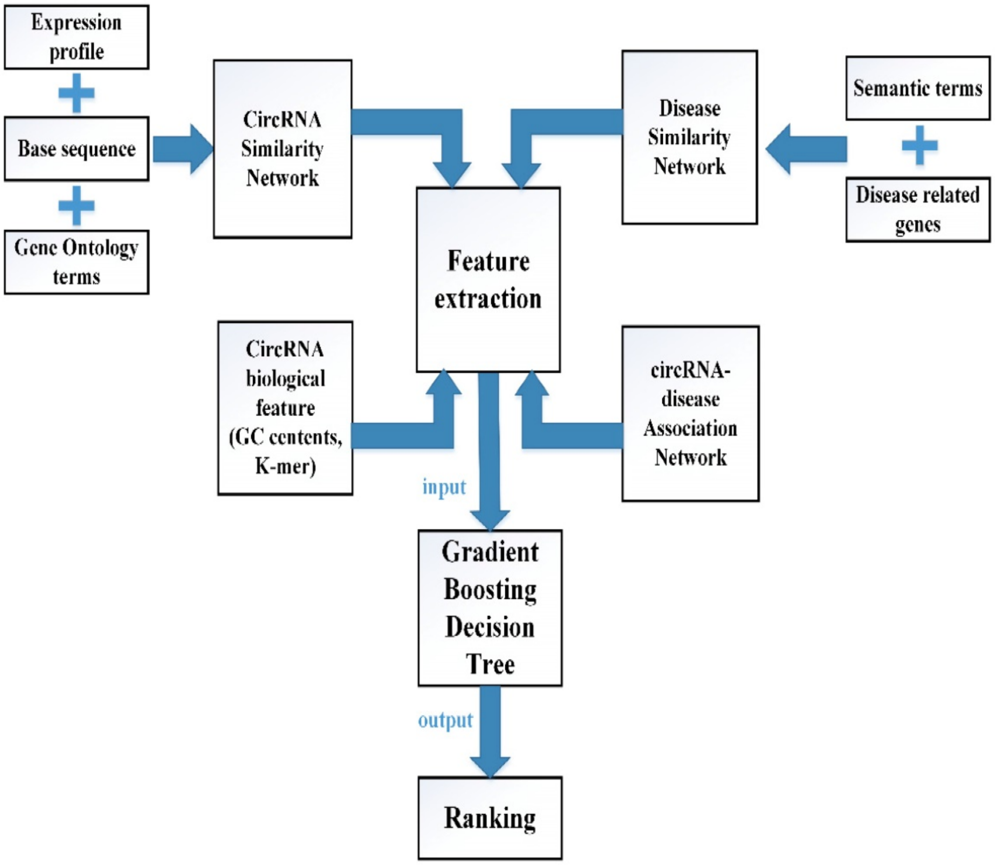
Expression profile
Base sequence
Gene Ontology terms
CircRNA Similarity Network
Feature extraction
Disease Similarity Network
Semantic terms
Disease related genes
CircRNA biological feature (GC centents, K-mer)
circRNA-disease Association Network
input
Gradient Boosting Decision Tree
output
Ranking

Figure 1. The flowchart of computational model GBDTCDA.

To extract the second type of feature for circRNA $c(i)$ or disease $d(j)$, the information of graph theory of the circRNA similarity matrix P, the disease similarity DS are considered. Each similarity score of matrix P and matrix DS is used to calculate the mean similarity score. Thus, the circRNA similarity matrix P and disease similarity matrix DS can be converted into unweighted graph, when the weight of edge exceeds the average value. Then, we can use the reconstructing unweighted graph to obtain the neighbor's number of $c(i)$ and $d(j)$ which can be denoted as $F_{2.num.nei}$. Based on the similarity matrix P and DS, we extract the top 10 similarity scores of $c(i)$ and $d(j)$, which can be denoted as $F_{2.K.sim}$. Given the first type of circRNA and disease feature, we can calculate the average of first type feature by using the top 10 neighbors, which is described as $F_{2.ave.feat1}$. Furthermore, we can also obtain the average of the first type features among the top 10 neighbors weighted by their corresponding similarity values, which are illustrated as $F_{2.W.ave.feat1}$. In order to get more information of the unweighted graph, the betweenness centrality, closeness centrality and eigenvector centrality of each node in the matrix P and DS are calculated, which can be denoted as $F_{2.bc}$, $F_{2.cc}$ and $F_{2.ec}$, respectively.

In order to construct the third feature of circRNA $c(i)$, the nucleotide sequence of $c(i)$ is adopted to calculate the biological feature. The content of GC base in nucleic acid sequence can be regarded as an important indicator of biological characteristics, which is described as $F_{3.GC.Cont}$. Then, the sequence assembly K-mer algorithm is used to count the number of matching base combination patterns, which can be denoted as $F_{3.Base.K\text{-}mer}$. In this study, K is set as 2, 3 and 4, respectively.

For constructing the fourth feature of each circRNA-disease($c(i)$, $d(j)$) pair from the circRNA-disease associations matrix A. The singular value decomposition (SVD) algorithm is adopted to obtain the latent vector of $c(i)$ and $d(j)$, which is denoted as $F_{4.svd}$. What's more, we also calculate the number of $c(i)$'s neighbors and the number of $d(j)$'s neighbors, which are described as $F_{4.c.d.num}$. and $F_{4.d.c.num}$. In addition, the betweenness centrality, closeness centrality and eigenvector centrality of each $c(i)$ and $d(j)$ pair, which can be depicted as $F_{4.c.d.bc}$, $F_{4.c.d.cc}$ and $F_{4.c.d.ec}$, respectively.

After all the information of circRNA similarity network, disease similarity network and circRNA-disease association network are extracted to construct the feature vector of each circRNA-disease pair, which are shown in supplementary Table S1. Four types of characteristics are merged into a feature vector of each circRNA-disease association as follows:

$$F(c(i), d(j)) = \left[F_1, F_2, F_3, F_4\right] \tag{9}$$

where F_1 is the category 1 characteristic which is the statistics information, F_2 is the category 2 characteristic which is the theory information, F_3 is the category 3 characteristic extracted from the circRNA related representational biological indicators, F_4 is the category 4 characteristic extracted from the cicrcRNA-disease associations network.

Gradient Boosting Decision Tree (GBDT) Regression

Gradient boosting[22] is an ensemble machine learning model which combines weak 'learners' into a strong single learner in an iteration fashion. In this study, we adopt a regression tree model to train a training set $\{(x_1, y_1), ..., (x_n, y_n)\}$ of known values of *x* and corresponding values of *y*. The objective is to find an approximation $\hat{F(x)}$ of function *F*(*x*), which minimizes the expected value of a given loss function F_L(*y*, *F*(*x*)). The definition of the approximation functions $\hat{F(x)}$ as follows:

$$\hat{F} = argminEX(F_L(y, F_L(x))) \tag{10}$$

where *y* is a real value. While the gradient boosting decision tree model makes an assumption that a real-valued *y* and seeks an approximation $\hat{F(x)}$ in the form of a weighted sum of functions $h_i(x)$ some class *H*, which can be called weak learners as follows:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + const \tag{11}$$

In accordance with the empirical risk minimization principle, the method tries to find an approximation $\hat{F(x)}$ that minimizes the average value of the loss function on the training test, i.e., minimizes the empirical risk. It does so by starting with a model, consisting of a constant function $F_0(x)$, and incrementally expanding it using a greedy fashion:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} F_L(y_i, \gamma), \tag{12}$$

$$F_m(x) = F_{m-1}(x) + \underset{h_m \in H}{argmin} \left[\sum_{i=1}^{n} F_L(y_i, F_{m-1}(x_i) + h_m(x_i))\right] \tag{13}$$

where $h_m \in H$ is a weak learner function.

Unfortunately, choosing the best function *h* at each step for arbitrary loss function F_L is a computationally infeasible optimization problem generally. Thus, a simplified version is adopted to solve this problem.

This main thought is to apply a steepest descent step to solve this minimization problem. If we considered the continuous case, i.e. where *H* is the set of arbitrary differentiable function on *R*, we would update the model based on the following formulas:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} F_L(y_i, F_{m-1}(x_i)) \tag{14}$$

$$\gamma_m = \underset{\gamma}{argmin} \sum_{i=1}^{n} F_L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} F_L(y_i, F_{m-1}(x_i))) \tag{15}$$

where the derivatives are taken with respect to the functions F_i for $i \in$ {1,..,m}[34]. In the discrete case however, i.e. when the set *H* is finite, we choose the candidate function *h* which is closest to the gradient of F_L for which the coefficient γ may then be calculated with the aid of line search on the above equations. Note that this approach is heuristic and therefore doesn't yield an exact solution to the given problem, but rather an approximation.

Performance Metrics

In this study, two main evaluation metrics are adopted to estimate the performance of our computational method, such as AUC value and *F-measure*. The AUC value is the area of (receiver operating characteristic) ROC curve, which is comprised of true positive rate (*TPR*) and false positive rate (*FPR*). The following equations are adopted to calculate the *TPR* and *FPR*:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{TN + FP} \tag{17}$$

where *TP* are the known circRNA-disease associations, which are distinguished correctly, and *FN* are the unknown circRNA-disease associations, which are identified incorrectly. What's more *TN* are the unknown circRNA-disease associations, which are identified correctly. Finally, *FP* are the known circRNA-disease associations, which identified incorrectly. In addition, in order to further describe the performance of GBDTCDA, *F-measure* is also adopted to integrate *precision* and *recall*, which is a comprehensive evaluation method. *Precision* can be described as the number of true positive samples (known circRNA-disease associations) in a predicted positive sample and *recall* can be depicted as the number of positive examples in the sample predicted correctly. *F-measure* can better evaluate the performance of the model, which is calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{18}$$

$$recall = \frac{TP}{TP + FN} \tag{19}$$

$$F - measure = \frac{2 \times precision \times recall}{(precision + recall)} \tag{20}$$

Results

Leave one out cross validation (LOOCV)

For each given definite disease *i*, there are some circRNAs having associations with the disease *i*. In this study, we pick up all the known circRNA-disease associations as the positive samples and select the same number of positive samples as the negative samples in unknown relationships randomly, which adopt LOOCV to measure the performance of GBDTCDA. During each training session, one circRNA-disease association is left out as the test data. After the score of each circRNA -disease association is obtained and those scores are sorted in descending order. Each score in descending order is regarded as threshold. Along with the changing threshold, we can calculate its corresponding *TPR* and *FPR*, which can be used to draw ROC curve. For the sake of representing our proposed model can compute more accurate results than other existing methods [35-38], which is represented in Fig. 2. In this study, some state-of-the-art methods have been adopted to compare with our computational methods, which include the label propagation algorithm (random walk restart in the heterogeneous network, RWRH and bi-random walk), information flow algorithm (KATZ) and network topology algorithm (Heterogeneous graph inference, HGI). In addition, for obtaining more comprehensive evaluation of our methods, we also use some machine learning algorithm to evaluate the performance of GBDTCDA, which are consisted of ensemble learning algorithm (adapt boost, Adaboost and random forest, RF), regression algorithm (logistic regression), generalized portrait algorithm (liner and poly kernel) and neighbor learning algorithm (*k*-nearest neighbors, KNN).

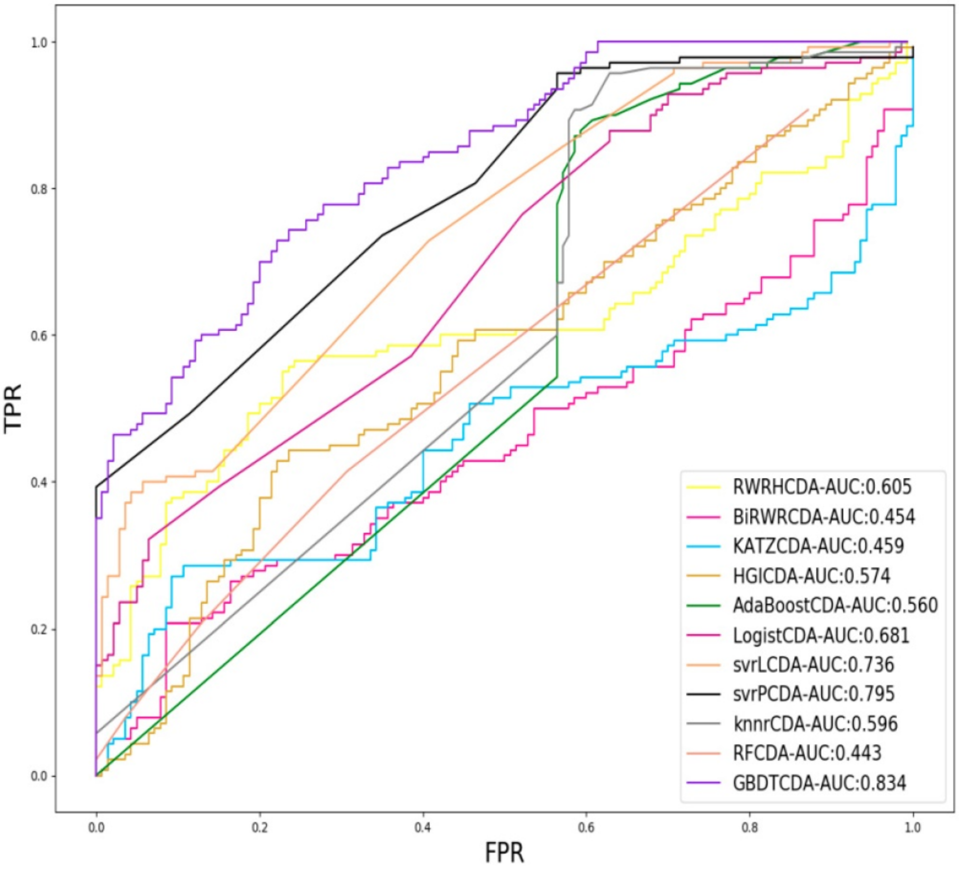

Figure 2. Comparison of the AUC value different methods.

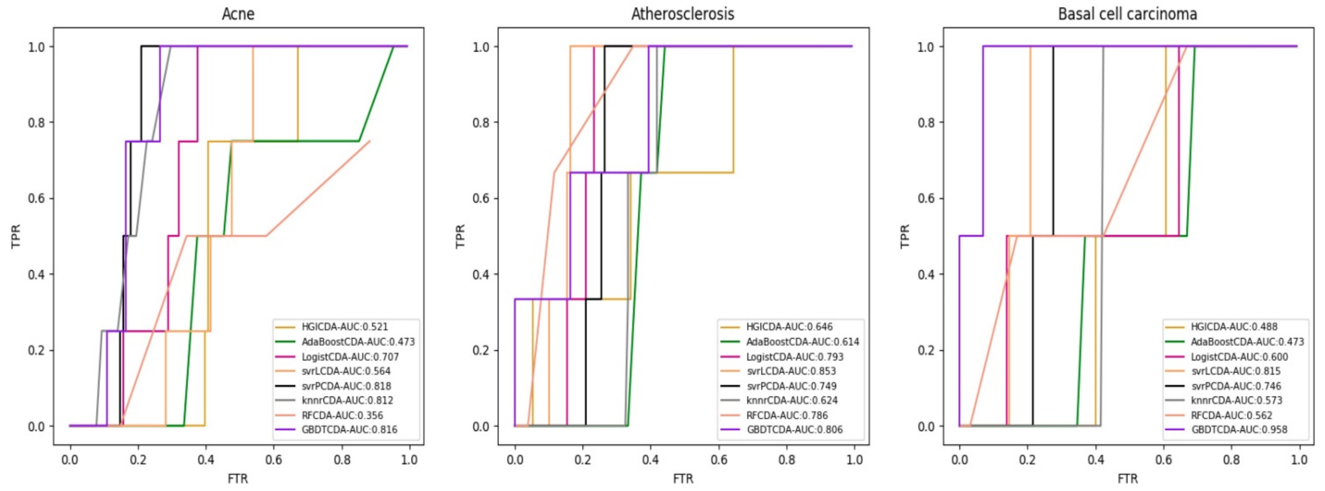

Figure 3. AUC value of Acne, Atherosclerosis and Basal cell carcinoma compared with other methods.

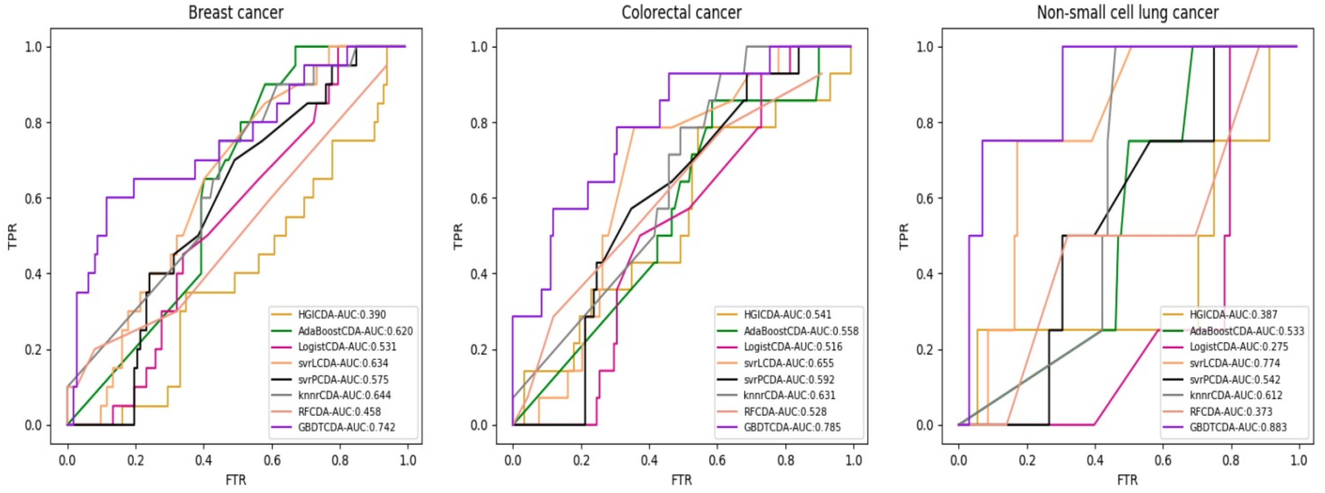

Figure 4. AUC value of Breast cancer, Colorectal cancer and Non-small lung cancer compared with other methods.

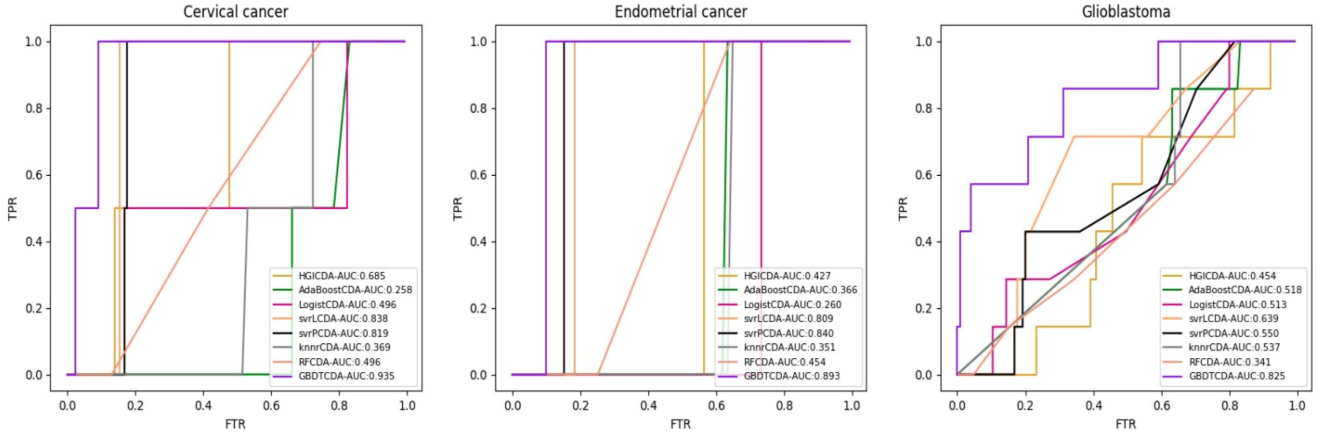

Figure 5. AUC value of Cervical cancer, Endometrial cancer and Glioblastoma compared with other methods.

For some common disease, such as breast cancer, Colorectal cancer, Non-small cell lung cancer, Glioblastoma and other five diseases are implemented separate LOOCV experiments, respectively, which are shown in Fig. 3 to Fig. 5. In almost occasion our proposed method can obtain much better performance than other network-based algorithms and machine learning methods. For obtaining a comprehensive explanation, *F-measure* is also adopted to evaluate the performance of methods which is represented in Fig. 6. *F-measure* value of our proposed method is 0.691, which much better than other methods. What's more, for the predicting scores in top *k* (*k* ∈ [80, 200]) circRNA-disease associations, the number of correct circRNA-disease associations predicted by our proposed method is greater than other methods. The result is shown in Fig. 7.

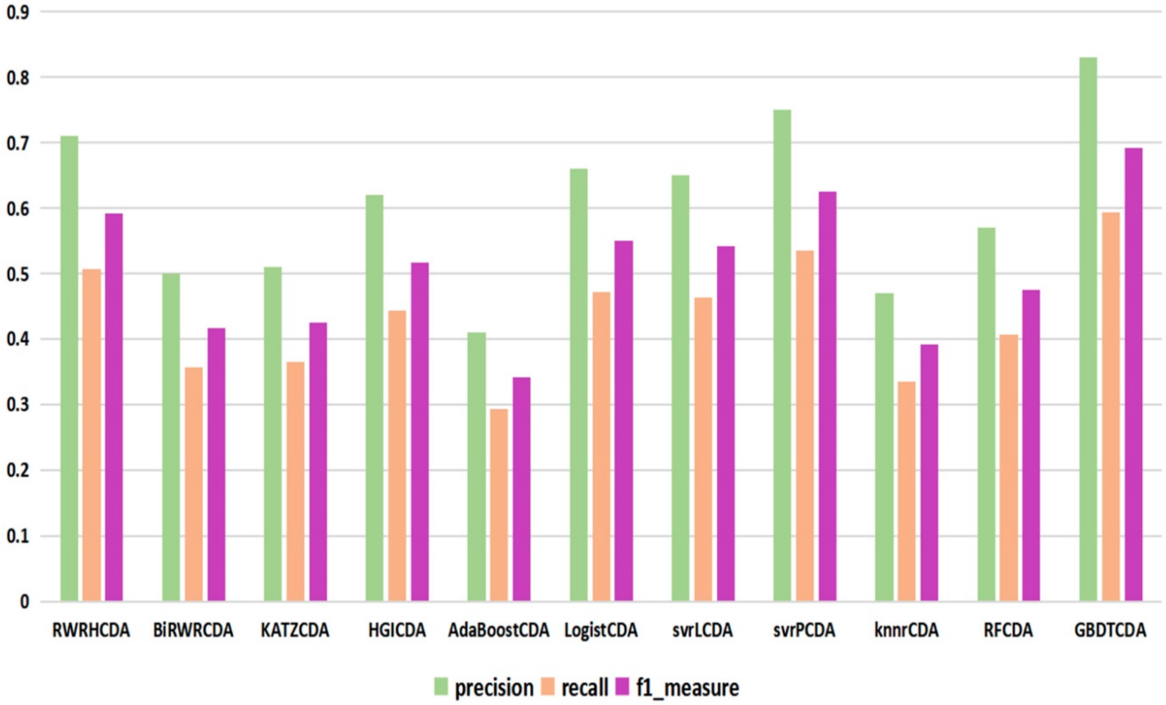

Figure 6. comparison of the precision, recall and f1_measure with different methods.

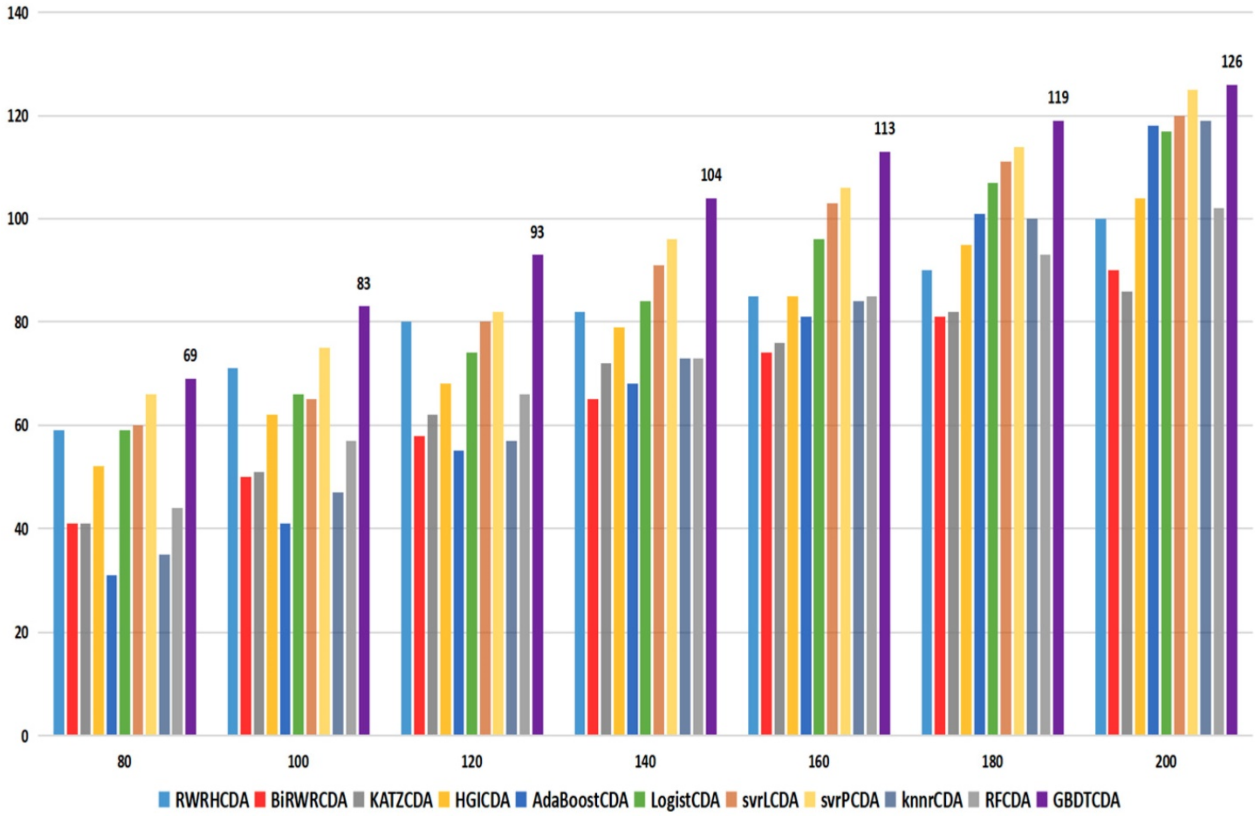

Figure 7. comparison of the top k ranks with different methods.

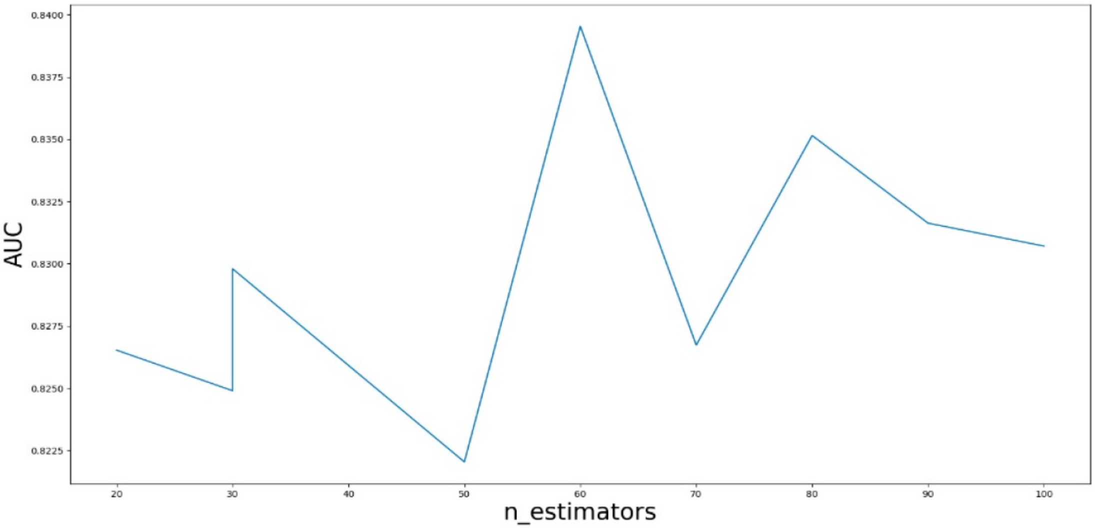

Figure 8. AUC value based on the different parameter n_estimators.

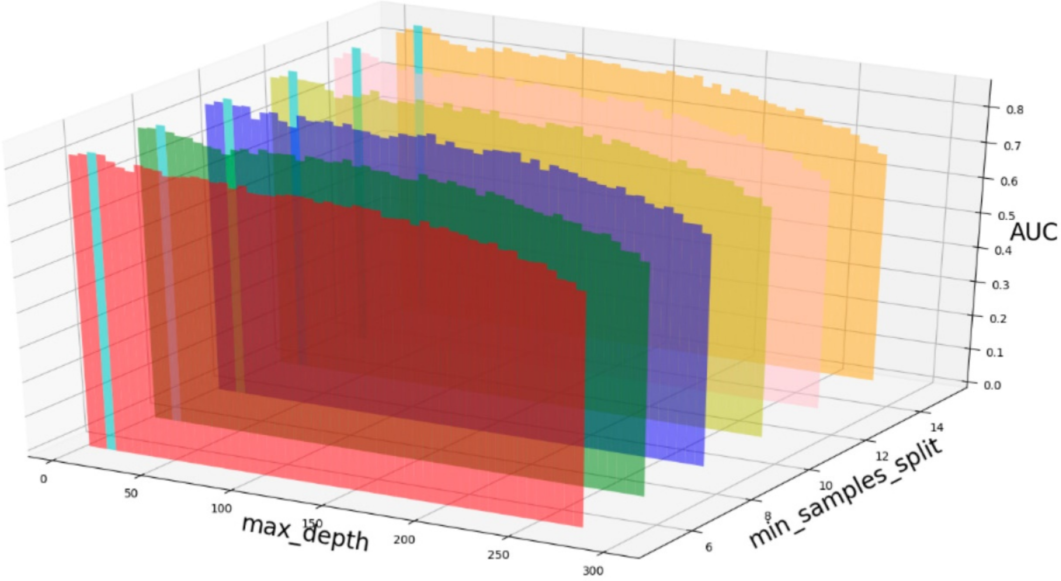

Figure 9. AUC value based on different max_depth and min_samples_split.

Parameter analysis

Based on our experiments, some better parameters are chosen to be set up in our computational model. There are some parameters that need to be adjusted necessarily for GBDT. Firstly, the parameter n_estimators which controls the number of trees of fit sequentially and is set up from 20 to 100 and fix parameter learning_rate as 0.1. The AUC values of different n_estimators based on the fixed learning_rate is represented in Fig. 8. When the parameter n_estimators is set as 60, the AUC value can obtain a better result. Secondly, n_estimators is fixed as 60. The parameters max_depth and min_samples_split which control the maximum depth of each decision tree and the minimum sample number for internal node repartition, respectively and they are set values from 2 to 20 and from 10 to 300, respectively. The results of the parameterization are shown in Fig. 9. Based on the results, max_depth is set as 9 and min_samples_split is set as 24. Thirdly, the parameter min_samples_leaf which control the leaf node minimum sample number is set different step sizes for adjustment, which is described in Fig 10. In according with the results, we can find that min_samples_leaf is set as 8 and 11, the maximum AUC values are obtained. Next, parameter min_samlples_leaf and max_features are adjusted together, which are presented in Fig. 11. In the light of the results, the values of min_samples_leaf and max_features are determined to be 11 and 9, respectively. Finally, the parameter subsample is set up from 0.6 to 0.9, whose results are shown in Fig. 12.

Case study

In order to further validate the capability of predicting potential circRNA-disease associations, some case studies are made to illustrate the performance of our proposed method. The predicting results are proofed by other two circRNA-disease associations databases which are circ2Disease [39] and circRNADisease [40], respectively. In this study, one common disease is adopted to make case studies. Breast cancer [41] is one of the deadly cancers worldwide now, which also becomes a public health issue for people all over the world. Based on previous studies, some factors can increase the risk of breast cancer, such as the age of first birth [42], frequency of regular exercise and some body indices [43], diet styles [44] and environmental factors[45]. While more and more evidences illustrate that circRNAs also can be a biomarker of breast cancer, which is represented in Table 2. Based on our proposed method, the results of those predicting circRNA-disease associations are validated by database circ2Disease (C_1) and circRNADisease (C_2).

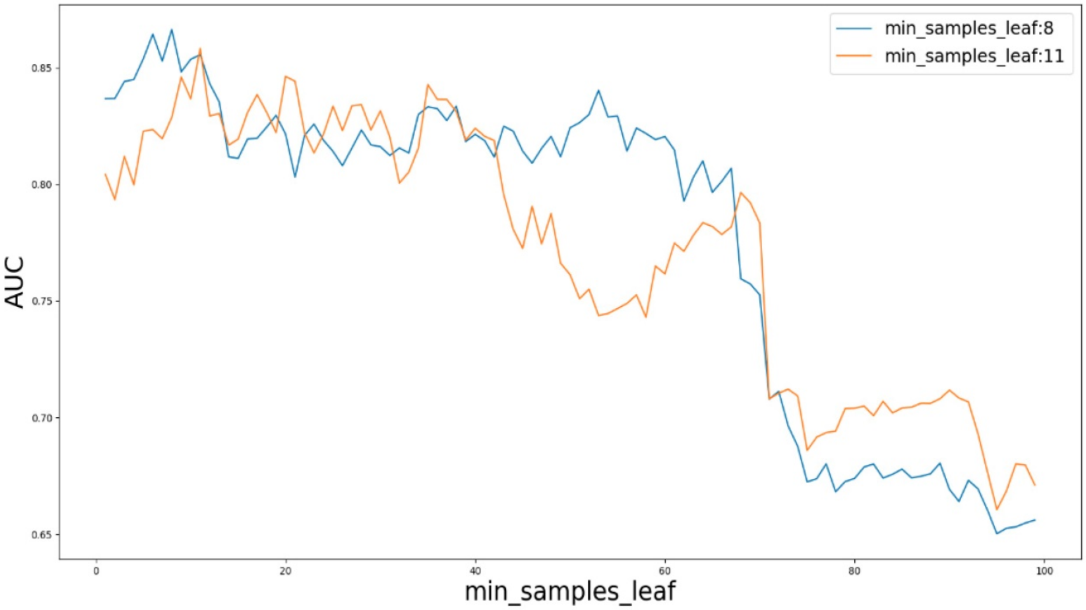

Figure 10. AUC value based on different parameter min_samples_leaf.

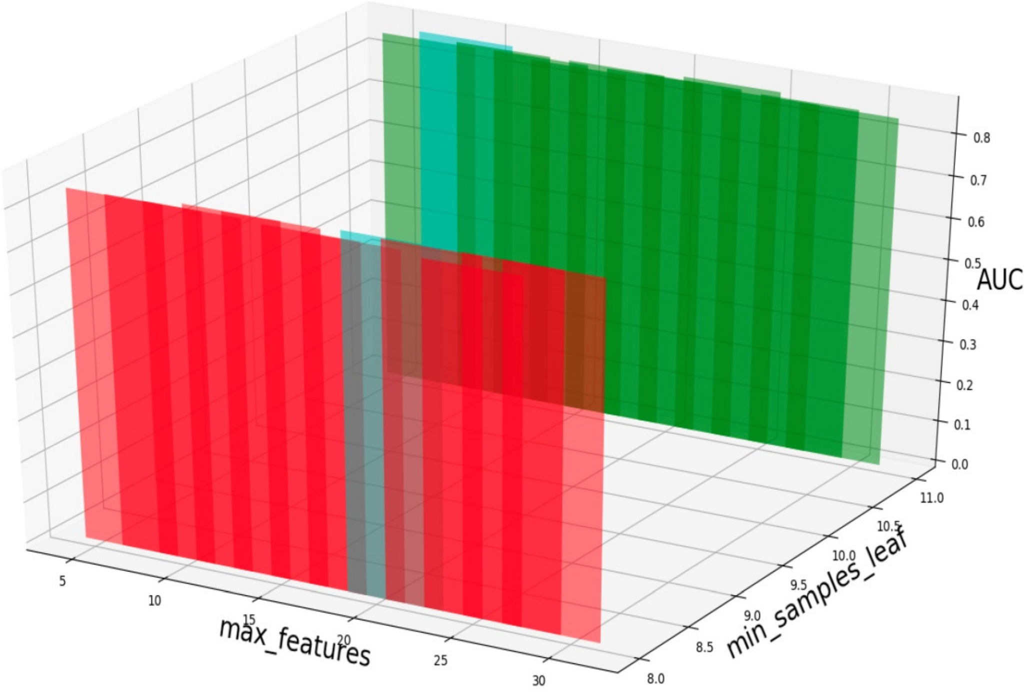

Figure 11. AUC value based on different min_samples_leaf and max_features.

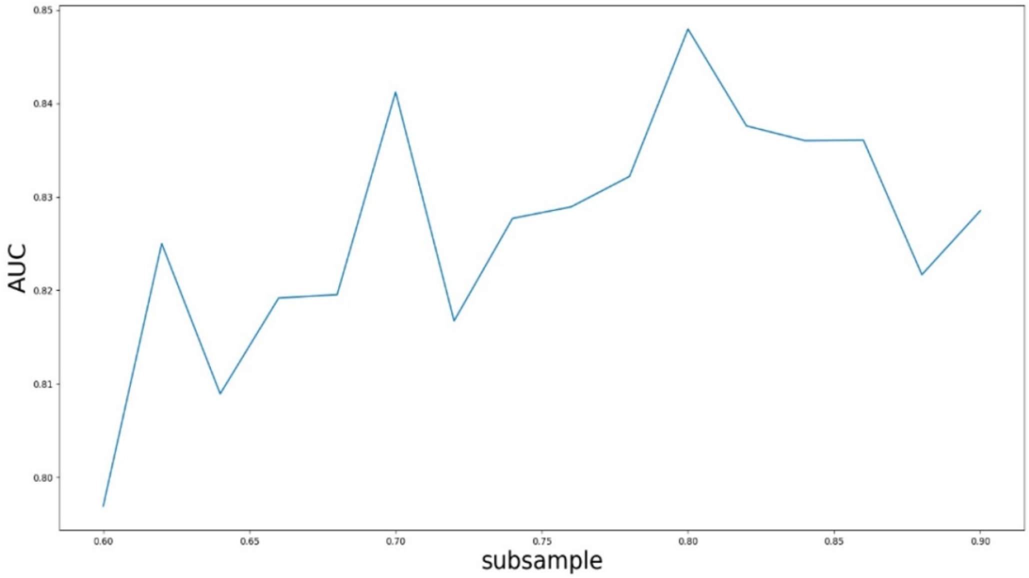

Figure 12. AUC value based on the different parameter subsample.

Table 2. The top 10 breast cancer related candidates circRNAs.

Rank	circRNA name/id	Evidences	Rank	circRNA name/id	Evidences
1	hsa_circRNA_103454/hsa_circ_0067103	unconfirmed	6	hsa_circ_0007534	PMID:29593432
2	hsa_circ_0006411	unconfirmed	7	hsa_circ_0001785	C_2
3	hsa_circ_103110/hsa_circ_0004771	C_1, C_2	8	hsa_circ_0001721	C_1
4	circMYO9B/hsa_circ_0000907	PMID 29702064	9	circAmotl1/hsa_circ_0004214	C_1, C_2
5	circRNA_100984/hsa_circ_0002019	unconfirmed	10	hsa_circ_100219/hsa_circ_0004619	C_1, C_2

Conclusion

With the rapid development of RNA high-throughput technologies, increasing number of diseases related circRNAs are discovered. Therefore, people pay more attention to revealing the intricate relationships between them. While using the traditional biological technologies are expensive and time-consuming. In this study, we propose a new computational method to predict the potential circRNA-disease associations, which is called the GBDTCDA, a machine learning driven method. what's more, gradient boosting model is first used to predict circRNA-disease associations and the LOOCV and *F-measure* evaluation measurements are adopted to illustrate the performance of our proposed method. Compared with other state-of-the-art computational methods, such as network-based methods, propagation methods and machine learning methods, GBDTCDA can get better results than those methods. In order to further describe the performance of GBDTCDA, the case studies of breast cancer are made. Thus, we believe that our proposed method GBDTCDA is a powerful tool to predict potential circRNA-disease associations.

For obtaining the better performance of our proposed computational mothed, some following significant factors cannot be ignored. Firstly, the characteristics, such as statistics information, graph theory information, circRNA sequence information and circRNA-disease latent features are taken into our consideration as comprehensive as possible, which can make the feature of each circRNA-disease pair more allover and promote the gradient boosting machine to be trained well based on those features. Secondly, in order to make our features get more reliable biological significance, multiple biological data, such as circRNA related GO terms, expression profile data and sequence data are adopted to construct the circRNA similarity network. In addition, a multiple data integration algorithm SNF is used to integrate different networks, which makes the integrated network more robust and reliable. Furthermore, an ensemble machine learning method called gradient boosting decision tree is adopted to train our inputting data.

Although our proposed method can obtain the better results compared with other methods, there still are some limits existing in our computational method. On the one hand, many parameters of gradient boosting machine need to be adjusted. In this study, the parameter adjustment is only carried out by some experiments. For our future works, some algorithms might be used to adjust those parameters. On the other hand, more categories of biological data could be taken into account, which can make our computational methods own more biological sense.

Supplementary Material

Supplementary table S1.
http://www.ijbs.com/v15p2911s1.pdf

Acknowledgments

This work was supported by the funding from National Natural Science Foundation of China (61972451, 61672334, 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

Competing Interests

The authors have declared that no competing interest exists.

References

1. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013; 495: 333.
2. Qu S, Yang X, Li X, Wang J, Gao Y, Shang R, et al. Circular RNA: A new star of noncoding RNAs. Cancer Lett. 2015; 365: 141-8.
3. Hsu MT, Coca-Prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature. 1979; 280: 339-40.
4. Kos A, Dijkema R, Arnberg AC, van der Meide PH, Schellekens H. The hepatitis delta (delta) virus possesses a circular RNA. Nature. 1986; 323: 558-60.
5. Pasman Z, Been MD, Garcia-Blanco MA. Exon circularization in mammalian nuclear extracts. RNA (New York, NY). 1996; 2: 603-10.
6. Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic acids research. 2012; 40: 3131-42.
7. Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biol; 2014. p. 409.
8. Shi X, Sun M, Liu H, Yao Y, Song Y. Long non-coding RNAs: a new frontier in the study of human diseases. Cancer Lett. 2013; 339: 159-66.
9. Zhou X, Dai E, Song Q, Ma X, Meng Q, Jiang Y, et al. In silico drug repositioning based on drug-miRNA associations. Briefings in bioinformatics. 2019.
10. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, et al. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014; 56: 55-66.
11. Chao CW, Chan DC, Kuo A, Leder P. The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. Mol Med. 1998; 4: 614-28.
12. Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, et al. Circular Intronic Long Noncoding RNAs. Molecular Cell. 2013; 51: 792-806.
13. Haque S, Harries LW. Circular RNAs (circRNAs) in Health and Disease. Genes (Basel); 2017.
14. Arvinden VR, Arinagiri Kuha DMR, Balaiah M, Ramanathan P, Sundersingh S, Velusami S, et al. RNA sequencing identifies dysregulated circular RNAs in early-stage breast cancer. bioRxiv; 2018.
15. Qin S, Zhao Y, Lim G, Lin H, Zhang X, Zhang X. Circular RNA PVT1 acts as a competing endogenous RNA for miR-497 in promoting non-small cell lung cancer progression. Biomed Pharmacother. 2018; 111: 244-50.
16. Hu M, Wei X, Li M, Tao L, Wei L, Zhang M, et al. Circular RNA expression profiles of persistent atrial fibrillation in patients with rheumatic heart disease. Anatol J Cardiol. 2019; 21: 2-10.
17. Wang L, Luo T, Bao Z, Li Y, Bu W. Intrathecal circHIPK3 shRNA alleviates neuropathic pain in diabetic rats. Biochem Biophys Res Commun. 2018; 505: 644-50.
18. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA (New York, NY). 2014; 20: 1666-70.
19. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Front Genet. 2013; p. 283.
20. Fan C, Lei X, Fang Z, Jiang Q, Wu F-X. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database : the journal of biological databases and curation. 2018; 2018.
21. Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, Yarden Y. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. Nucleic acids research. 2016; 44: 1370-83.
22. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist. 2001; 29: 1189-232.
23. Fan C, Lei X, Fang Z, Jiang Q, Wu F-X. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database : the journal of biological databases and curation. 2018; 2018: bay044.
24. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England). 2009; 25: 1422-3.
25. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database—2009 update. Nucleic Acids Research. 2009; 37: D767-D72.
26. Lin D. An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning: Morgan Kaufmann Publishers Inc. 1998;. p. 296-304.
27. Li S, Li Y, Chen B, Zhao J, Yu S, Tang Y, et al. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. Nucleic acids research. 2018; 46: D106-D12.
28. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11: 333-7.
29. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017; 45: D833-d9.
30. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research. 2002; 30: 52-5.
31. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research. 2015; 43: D1071-8.
32. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics (Oxford, England). 2015; 31: 608-9.
33. Chen J, Li Y, Zheng Q, Bao C, He J, Chen B, et al. Circular RNA profile identifies circPVT1 as a proliferative factor and prognostic marker in gastric cancer. Cancer letters. 2017; 388: 208-19.
34. Xie M, Hwang T, Kuang R. Prioritizing disease genes by bi-random walk. Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II. Kuala Lumpur, Malaysia: Springer-Verlag. 2012; p. 292-303.
35. Liu Y, Zeng X, He Z, Zou Q. Inferring MicroRNA-Disease Associations by Random Walk on a Heterogeneous Network with Multiple Data Sources. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2017; 14: 905-15.
36. Ganegoda G, Wang J, Wu F-X, Li M. Prediction of disease genes using tissue-specified gene-gene network. BMC Syst Biol. 2014; 8 Suppl 3: S3.
37. Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, et al. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. Scientific reports. 2016; 6: 30024.
38. Zhao Z, Wang K, Wu F, Wang W, Zhang K, Hu H, et al. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell death & disease. 2018; 9: 475.
39. Yao D, Zhang L, Zheng M, Sun X, Lu Y, Liu P. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci Rep; 2018. p. 11018.
40. Zhao Z, Wang K, Wu F, Wang W, Zhang K, Hu H, et al. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell Death Dis. 2018; p. 475.
41. Veronesi U, Boyle P, Goldhirsch A, Orecchia R, Viale G. Breast cancer. The Lancet. 2005; 365: 1727-41.
42. MacMahon B. Reproduction and cancer of the breast; 1993.
43. van den Brandt PA, Spiegelman D, Yaun SS, Adami HO, Beeson L, Folsom AR, et al. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. American journal of epidemiology. 2000; 152: 514-27.

44. Smith-Warner SA, Spiegelman D, Adami HO, Beeson WL, van den Brandt PA, Folsom AR, et al. Types of dietary fat and breast cancer: a pooled analysis of cohort studies. International journal of cancer. 2001; 92: 767-74.
45. Modan B, Chetrit A, Alfandary E, Katz L. Increased risk of breast cancer after low-dose irradiation. Lancet (London, England). 1989; 1: 629-31.