BMC Bioinformatics

RESEARCH

Open Access

A novel bi-directional heterogeneous network selection method for disease and microbial association prediction

Jian Guan, Zhao Gong Zhang*, Yong Liu* and Meng Wang

*Correspondence:
2013010@hlju.edu.cn;
2010023@hlju.edu.cn

School of Computer Science and Technology, Heilongjiang University, Harbin, China

Abstract

Microorganisms in the human body have a great impact on human health. Therefore, mastering the potential relationship between microorganisms and diseases is helpful to understand the pathogenesis of diseases and is of great significance to the prevention, diagnosis, and treatment of diseases. In order to predict the potential microbial disease relationship, we propose a new computational model. Firstly, a bi-directional heterogeneous microbial disease network is constructed by integrating multiple similarities, including Gaussian kernel similarity, microbial function similarity, disease semantic similarity, and disease symptom similarity. Secondly, the neighbor information of the network is learned by random walk; Finally, the selection model is used for information aggregation, and the microbial disease node pair is analyzed. Our method is superior to the existing methods in leave-one-out cross-validation and five-fold cross-validation. Moreover, in case studies of different diseases, our method was proven to be effective.

Keywords: Bi-directional heterogeneous network, causal selection model, Potential microorganism disease prediction, Random walk

Introduction

The microbial community is composed of bacteria, fungi, protozoa, and eukaryotes. It has an important impact on human beings in the fields of food, agriculture, environmental governance, and human health [1]. Microorganisms living in different organs of the human body can directly affect human health by regulating human immune system, drug metabolism and pathogen prevention [2]. Therefore, finding more links between microorganisms and diseases can not only help us better under- stand the pathogenesis of diseases, but also promote doctors' diagnosis of diseases. In recent years, many computational methods have been proposed to explore the potential correlation in biological information. The existing calculation methods applied to microbial disease association are mainly divided into three categories. The first is based on fractional functions, the second is based on network algorithms, and the third is based on machine learning.

The first method, the Katz method proposed by Chen et al., is based on scoring function utilizing similarity and known association to predict the potential association between microorganism and disease. For association prediction, this method uses the number and length of paths between nodes in heterogeneous networks to predict [3]. Additionally, Li et al. [4] used Gaussian interaction profle kernel (GIP) similarity to build a weighted heterogeneous network for association prediction.

The second type of prediction method is to use machine learning technology to predict microbial disease association [5, 6]. Wang et al. [7] proposed a semi supervised model based on Laplace regularized least square algorithm. Peng et al. [8] combined multiple weak classifiers to form a strong classifier for prediction.

Lastly, methods based on random walk and matrix decomposition have also been explored to reveal potential microbial disease associations. Several methods use ordinary random walk [9], double random walk of logic function transformation [10] and random walk based on hypergraph [11] to predict potential microbial disease association. Qu et al. [12] used matrix decomposition and label propagation to infer potential associations. In addition, [13] proposed a method based on similarity constraint matrix decomposition to predict potential microRNA and disease associations. Several other methods have been proposed to predict microbial and disease associations based on network consistency projection [14] and multi similarity fusion tag propagation [15]. However, these methods still have some limitations. Because the sparsity of data and the singleness of methods limit the dissemination of information; these methods are difficult to extract the deep-seated Association of microorganisms (or diseases) from the data.

With the development of artificial intelligence, the method based on deep learning has been widely used in various fields. Long et al. proposed a framework to complete prediction based on graph attention network and inductive matrix [16]. Lei et al. [17] used the combination of node2vec algorithm and rule-based reasoning to predict potential associations. Liu et al. [18] have combined non negative matrix decomposition, random walk and capsule neural network to predict the association between microorganisms and diseases. A method based on multi-component graph attention network (GATMDA) was proposed to predict the potential association between microorganisms and diseases [19]. However, many similarities of various microorganisms (diseases) have not been fully utilized. Moreover, most of the previous methods rely on the known microorganism disease association for similarity calculation, thus these methods cannot achieve prediction when involving new diseases (or new microorganisms) due to the lack of training data.

In this paper, we propose a method (BDHNS), to predict microbial disease association based on a bi-directional microbial disease heterogeneous network and selection model. The contribution of our approach is mainly reflected in the following three aspects. First, we constructed a bi-directional heterogeneous microbial disease network based on the similarities between microorganisms and diseases. Different similarities can more comprehensively reflect the relationship between microorganisms and diseases from different angles. Secondly, based on an enhanced bi-directional random walk, we learn the neighbor topology information of microorganism and disease nodes in bi-directional heterogeneous networks from the two directions of microorganism disease and disease microorganism. The multi-feature fusion of microorganism and disease nodes is helpful for the final prediction of microorganism disease association. In

heterogeneous networks, if all the neighbor information of each node is aggregated, the information of different types of nodes may also be aggregated, which will cause information redundancy. Therefore, lastly, we propose a graph convolution-based selection model to selectively aggregate neighbor information of disease (microorganism) nodes. The improved prediction performance is demonstrated by comparison with the most advanced models, ablation experiments, and case studies based on the Loocv and Five-fold cross-validation.

Materials and methods

In order to predict the potential association between microorganisms and diseases, we propose a method based on a bi-directional random walk and selection model (Fig. 1). Firstly, a bi-directional heterogeneous network containing microorganisms and disease nodes is established based on multiple similarities to integrate the Gaussian kernel similarity of microorganisms, the functional similarity of microorganisms, the semantic similarity of diseases, and the similarity of disease symptoms. Secondly, we build an enhanced bi-directional random walk module to learn the neighbor topology information of microorganisms and disease nodes. Then, a graph convolution based selection model is proposed to selectively aggregate the neighbor topology and attribute information of each node in the network and calculate the association probability of node pairs.

Microbial disease association data

We downloaded microbial disease association data from the database HMDAD, including 39 diseases and 292 microorganisms, covering 483 microbial disease associations. After removing duplicate records, we obtained 450 associations involving 39 diseases and 292 microorganisms. Subsequently, we used the data to construct the microbial disease association matrix A, if disease $d(i)$ is associated with microorganism $m(j)$, then $A(i,j) = 1$, otherwise $A(i,j) = 0$.

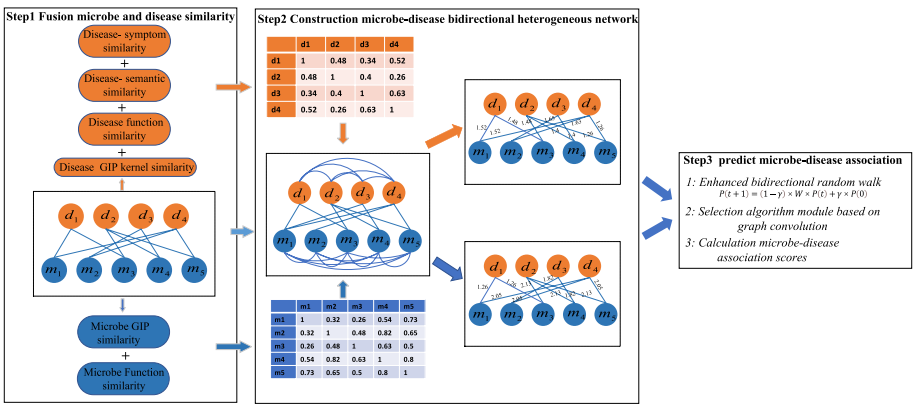

Fig. 1 Overall flowchart of BDHNS. Step1: We fuse the calculated similarity of two microorganisms with the similarity of four diseases. Step2:We first build two one-way heterogeneous network respectively corresponding microorganisms and diseases, and then convert two one-way heterogeneous networks into a two-way heterogeneous network of microorganisms and diseases. Step3:We use the enhanced random walk and selection algorithm to predict the potential association between microorganisms and diseases

Construction of bi-directional heterogeneous network

Based on the similarity of many microorganisms and diseases, we constructed a bi-directional heterogeneous microbial disease network $G = (V, E)$. Among them, node set V is composed of microbial node set V_m and disease node set V_d. Edge $e_{ij} \in E$ represents the edge between a pair of nodes, $V_i, V_j \in V$. Heterogeneous network G contains microbial-microbial similarity, disease-disease similarity and bi-directional association between microorganisms and diseases based on disease similarity and microbial similarity.

Microbial similarity

The more diseases are associated with two microorganisms, the more likely they are to demonstrate the same function. GIP similarity of microorganisms was calculated based on this assumption. The specific calculation process of GIP similarity matrix GM of microorganisms $m(i)$ and $m(j)$ is as follows:

$$GM(m(i,j)) = \exp\left(-r_m \left\|A(m(i)) - A(m(j))\right\|^2\right) \tag{1}$$

$$r_m = r_m{}' / \frac{1}{N}\sum_{i=1}^{N_m} ||A(m(i))||^2 \tag{2}$$

where $A(m(i))$ represents column i of A. $r_m{}'$ is a parameter that affects r_m normalization and $r_m{}' = 1$. N_m indicates that the number of microorganisms is 292.

Information on the resident organs of microorganisms and their effects on disease can be obtained from HMDAD. If two microorganisms live in the same organ and are associated with similar diseases, there is a greater degree of similarity between them. In the same organ, if two microorganisms affect the same disease, the degree of similarity between them is 1, otherwise it is 0. The microbial function similarity matrix FM can be obtained by accumulating the effects of diseases on microorganisms in various organs. Then the FM is normalized to obtain the final FM. The normalization calculation process is as follows:

$$FM(m(i), m(j)) = \frac{FM(m(i), m(j)) - \min(FM)}{\max(FM) - \min(FM)} \tag{3}$$

where $\max(FM)$ and $\min(FM)$ are the maximum and minimum values in the matrix FM.

Given microbial GIP similarity matrix GM and microbial functional similarity matrix FM, the final microbial similarity matrix SM is integrated as follows:

$$SM(i,j) = \begin{cases} GM(i,j) & if\ FM(i,j) = 0 \\ (GM(i,j) + FM(i,j))/2 & else \end{cases} \tag{4}$$

Disease similarity calculation

Similar to the calculation method of microbial GIP similarity matrix, the calculation process is as follows:

$$GD(d(i,j)) = \exp\left(-r_d \left\|A(d(i)) - A(d(j))\right\|^2\right) \tag{5}$$

$$r_d = r_d{}' / \frac{1}{N} \sum_{i=1}^{N_d} ||A(d(i))||^2 \tag{6}$$

where $A(d(i))$ represents row i of adjacency matrix A. $r_d{}'$ set to 1. N_d is 39, representing the number of diseases. Each disease can be represented by constructing a directed acyclic graph (*DAG*), which contains the disease and all its ancestral diseases [20]. Thus, we can calculate the semantic contribution of each disease in *DAG* to each disease in our data. The calculation process is as follows:

$$D_D(d) = \begin{cases} 1 & if\ d = D \\ \max\left\{\Delta \times D_D(d') | d' \in children\ of\ d\right\} & if\ d \neq D \end{cases} \tag{7}$$

where Δ is the semantic attenuation factor. The value of is set to 0.5.

By calculating the semantic contribution values of all diseases in *DAG*, the semantic values of diseases are calculated as follows:

$$D_V(D) = \sum\nolimits_{t \in V_d} D_D(t) \tag{8}$$

where V_d includes disease D and all its ancestral diseases. The more common semantics the two diseases *DAG* contains, the more similar the two diseases will be. The semantic similarity between the two diseases is calculated as follows:

$$DSS(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} D_{(i)}(t) + D_{(j)}(t)}{D_V(D(i)) + D_V(D(j))} \tag{9}$$

Two similar diseases may interact with similar genes [21], and the disease function similarity matrix can be calculated based on the interaction between disease-related genes. Humannet v2.0 database contains gene interactions [22], in which each interaction has a related log likelihood score (LLS) to evaluate the probability of functional linkage between genes. We can get the relevant genomes of diseases $d(i)$ and $d(j)$ that $G_i = \{g_{i1}, g_{i2}, ..., g_{im}\}$ and $G_j = \{g_{j1}, g_{j2}, ..., g_{jn}\}$, respectively. where m is the number of genes in G_i and n is the number of genes in G_j. The association between gene g and gene set $G = \{g_i, g_2, ..., g_k\}$ is as follows:

$$F_G(g) = \max_{g_i \in G}(FSS((g, g_i))) \tag{10}$$

where *FSS* represents the functional similarity score between genes, which is calculated as follows:

$$FSS(g_i, g_j) = \begin{cases} 1 \text{ ifi} = \text{j} \\ LLS'(g_i, g_j) \text{ if i} \neq \text{j} \end{cases} \tag{11}$$

where LLS' is the standardization of gene *LLS*, which is calculated as follows:

$$LLS'(g_i, g_j) = \frac{LLS(g_i, g_j) - LLS_{\min}}{LLS_{\max} - LLS_{\min}} \tag{12}$$

where $LLS_{\max}$ and $LLS_{\min}$ represent the maximum and minimum values in HumanNet respectively.

Finally, we calculate the disease functional similarity as follows:

$$DF(d(i),d(j)) = \frac{\sum_{g_t \in G(d(i))} F_{G(d(j))}(g_t) + \sum_{g_t \in G(d(j))} F_{G(d(i))}(g_t)}{m+n} \tag{13}$$

By integrating disease GIP similarity, disease semantic similarity, disease symptom similarity (TD) and disease function similarity, the final disease similarity matrix SD is expressed as:

$$SD = \frac{GD + DSS + TD + DF}{4} \tag{14}$$

where symptom-based diseases similarity TD was calculated using the associations of diseases and symptoms. The associations between diseases and symptoms is extracted from the human symptomatic disease network [23]. Therefore, we use the vectors of diseases related symptoms and refer to the method in [23] to calculate the similarity between disease-disease based on cosine similarity measurement, and then obtain disease symptom similarity (TD).

Bi-directional correlation calculation between microorganism and disease

For a given disease, the degree of correlation with different microorganisms is different. For example, for a given disease $d(i)$, some microorganisms related to $d(i)$ have a strong similarity relationship, while others have no or low similarity relationship with $d(i)$. Therefore, the similarity topology between microorganism and disease is used to calculate the correlation between disease (microorganism) and microorganism (disease). Specifically, we constructed a bi-directional heterogeneous network containing disease and microbial nodes. When calculating the correlation degree of disease to microorganism, the edge weight transferred from the disease network $d(i)(i = 1, 2, ..., N_d)$ node to the microbial network $m(j)(j = 1, 2, ..., N_m)$ node is defined as the sum of the weights from the disease node associated with microorganism $m(j)$ to the $d(i)$ node. Similarly, we can also calculate the degree of association between microorganisms and disease direction. Therefore, two new adjacent matrices A'_{SD} and A'_{SM} can be obtained based on the similarity matrices SD and SM. The calculation process is as follows:

$$A'_{SD}(i,j) = \sum_{k=1}^{n_d} SD(i,k)a_{kj} \tag{15}$$

$$A'_{SM}(i,j) = \sum_{k=1}^{n_m} a_{ik}SM(k,j) \tag{16}$$

where $a_{ik}(a_{jk})$ is the element on row i and column k (row k and column j) of adjacent matrix A.

Given the bi-directional correlation matrix A'_{SD} and A'_{SM} of microorganisms and diseases,microbial similarity matrix SM and disease similarity matrix SD, a bi-directional heterogeneous microbial disease network can be established. The adjacency matrix of bi-directional heterogeneous networks is A_{all},

$$A_{all} = \begin{pmatrix} SD & A_{SD}^{'} \\ A_{SM}^{'} & SM \end{pmatrix} \tag{17}$$

Learning neighbor topology enhanced bi-directional random walk

The transition probabilities between slave nodes are uniformly distributed in ordinary random walks. In our bi-directional random walk, the transition probability matrix $W_{DM}^{'}$ from disease to microbial node and $W_{MD}^{'}$ from microbial network to disease network are redefined. $W_{DM}^{'}$ and $W_{MD}^{'}$ are as follows:

$$W_{DM}^{'}(i,j) = \varphi \frac{a_{ij} A_{SM}^{'}(i,j)}{\sum_{l=1}^{n_m} a_{il} A_{SM}^{'}(i,l)} \tag{18}$$

$$W_{MD}^{'}(i,j) = \varphi \frac{a_{ij} A_{SD}^{'}(i,j)}{\sum_{l=1}^{n_d} a_{li} A_{SD}^{'}(l,j)} \tag{19}$$

where $\varphi \in (0,1)$ is the jumping probability of walkers between disease network and microbial network. W_d is the transition probability matrix between disease networks, $W_d(i,j)$ represents the jump probability from disease $d(i)$ to disease $d(j)$. $W_d(i,j)$ is expressed as follows:

$$W_d(i,j) = \begin{cases} (1-\varphi) SD(i,j) / \sum_{k=1}^{n_d} SD(i,k) & \text{if } \sum_{k=1}^{n_m} a_{ik} \neq 0 \\ SD(i,j) / \sum_{k=1}^{n_d} SD(i,j) \end{cases} \tag{20}$$

Similarly, the microbial network transition probability matrix W_m is expressed as follows:

$$W_m(i,j) = \begin{cases} (1-\varphi) SM(i,j) / \sum_{k=1}^{n_m} SM(i,k) & \text{if } \sum_{k=1}^{n_d} a_{ki} \neq 0 \\ SM(i,j) / \sum_{k=1}^{n_m} SM(i,k) & \text{otherwise} \end{cases} \tag{21}$$

In general random walk, the microbial disease transfer probability matrix and the disease microbial transfer probability matrix are transposed. Since our heterogeneous network is bi-directional, we propose an enhanced random walk to improve the comprehensive generalization ability of the bi-directional network, which can be described as follows:

$$P(t+1) = (1-r) \times W \times P(t) + r \times P(0) \tag{22}$$

where W is the transition probability matrix of all nodes in the bi-directional heterogeneous network, which is defined as follows:

$$W = \begin{pmatrix} W_d & W_{MD}^{'} \\ W_{DM}^{'} & W_m \end{pmatrix} \tag{23}$$

Selection algorithm module based on graph convolution

Most methods will aggregate all the neighbor information of nodes in heterogeneous graphs (as shown in Fig. 2), which may lead to the aggregation of redundant information to ignore the difference in local information of nodes [24]. Therefore, we adopt a

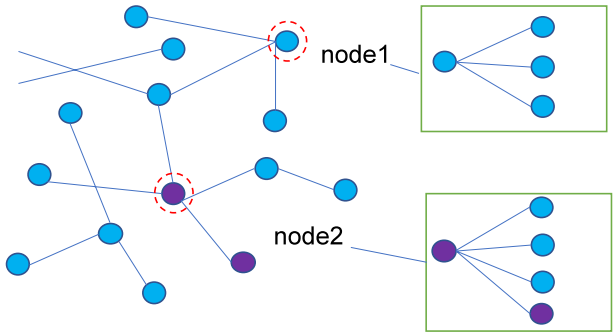

Fig. 2 Heterogeneous map

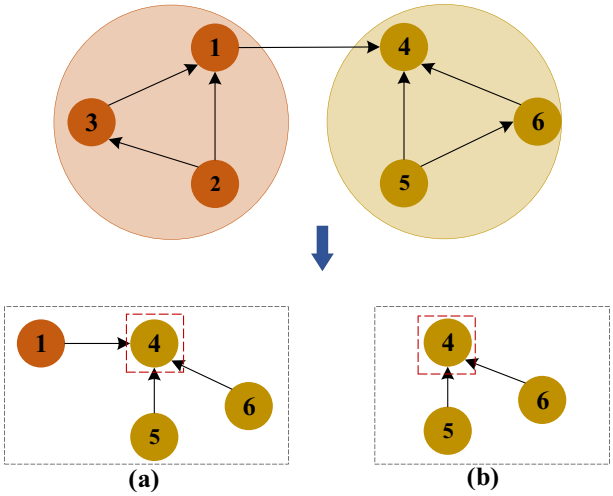

Fig. 3 Select algorithm graph

graph convolution-based selection algorithm to selectively aggregate the neighbor information of each node in the network, so that our model can be applied to more practical scenarios.

We take the transition probability matrix W as the attribute matrix X of microbial and disease nodes, and propose a graph convolution based selection algorithm to selectively aggregate the neighbor topology and attributes of each node in the network, that is, based on graph convolution, we can aggregate them in the following two ways:

Method 1: When the graph convolution aggregates neighbors, all the neighbor information of the node is aggregated, as shown in Fig. 3a.

The adjacency matrix A_{all} and attribute matrix X of the bi-directional heterogeneous network are used as the inputs of the graph convolution based selection algorithm module, and the output of the graph convolution is,

$$\hat{H} = f(x, N(x)|\hat{\theta}) \tag{24}$$

where $\hat{H}$ is the node feature of the convolution output of the original graph, x is the attribute vector of the target node, and $N(x)$ is the attribute vector of the neighbor node of the target node.

We use the node attribute vector of graph convolution output to calculate the association probability $\hat{y}$ between microorganism and disease, as shown in the following formula:

$$y(m(i), d(j)) = \frac{\sum_{k=1}^{m} Sim(h_{m(i)}, h_{m(k)})A(j,k) + \sum_{k=1}^{d} Sim(h_{d(j)}, h_{d(k)})A(k,j)}{\sum_{k=1}^{m} Sim(h_{m(i)}, h_{m(k)}) + \sum_{k=1}^{m} Sim(h_{d(i)}, h_{d(k)})} \tag{25}$$

$$Sim(u, v) = \frac{\sum_{k=1}^{g} u_k v_k}{\sqrt{\sum_{k=1}^{g} u_k^2}\sqrt{\sum_{k=1}^{g} v_k^2}} \tag{26}$$

where $h_{m(i)}$ and $h_{d(j)}$ represent the attribute vectors of disease node $d(j)$ and microorganism node $m(i)$ respectively, and A is the microorganism disease correlation matrix.

Method 2: In the application of graph convolution, only nodes of the same type are aggregated, as shown in Fig. 3b,

$$\hat{H}^s = f(x, \phi|\hat{\theta}^s) \tag{27}$$

where $\hat{H}^s$ is the node attribute of convolution output of graph after intervention. x is the node characteristic, ϕ is the causal intervention. Similarly, the microbial disease prediction association probability $\hat{y}^s$ after the intervention is also calculated based on formulas 24 and 25.

Select algorithm

Two methods of aggregating neighbor information were used to obtain two prediction scores of microbial disease association, i.e. association probability $\hat{y}$ of original prediction and association probability $\hat{y}^s$ of post intervention prediction. Our purpose is that the model can make the best choice between $\hat{y}$ and $\hat{y}^s$ for the final prediction, so as to reduce the impact caused by the local structural difference of nodes [25], which is calculated as follows:

$$\bar{y} = h(\hat{y}, \hat{y}^s|e) \tag{28}$$

where h is the selection function and e is the causal effect factor, which is defined as follows:

$$\begin{aligned} e &= f(x, N(x)|\hat{\theta}) - f(x, do(N = \phi)|\hat{\theta}) \\ &= f(x, N(x)|\hat{\theta}) - f(x, \phi|\hat{\theta}) \\ &= \hat{y} - \hat{y}^s \end{aligned} \tag{29}$$

The final microbial disease association prediction probability is calculated as follows:

$$\bar{y} = \begin{cases} \hat{y}, e >= m \\ \hat{y}^s, e <= m \end{cases} \tag{30}$$

where m is the threshold.

Experimental evaluation and discussion

Parameter setting and evaluation index

Our model BDHNS runs on a GPU(Nvidia GeForceRTX2060). We quantitatively analyze the parameter random walk step number t, restart probability r and characteristic dimension d to determine their values. We set the restart probability r of random walk from 0.1 to 0.9. The embedding dimension d and the number of random walk steps t are set similarly to r, with d and t varying from 8 to 128 and from 5 to 30, respectively. In order to facilitate parameter tuning, one parameter is tested and the other parameters are fixed. When the restart probability r is 0.1, the number of steps t is 20, and the embedding dimension d is 64, our model has the best performance. We use all combinations of parameter d, restart probability r and the range of embedded dimension d to construct our model.

Loocv cross validation and 5-fold cross validation were used to evaluate the performance of our method and other state-of-the-art microbial disease prediction methods. In Loocv, each known association between microorganism and disease is selected as the test sample, while other known associations are training samples. In the 5-fold cross validation, the known association is regarded as a positive sample, and the unobserved association is regarded as a negative sample. All positive samples were randomly divided into five groups, four of which were put into the training set, and the rest were used for testing. In each cross validation, we randomly selected negative samples with the same amount as 4 groups of positive samples for training, and the remaining negative samples are used for testing.

Our evaluation indicators include true positive rate (TPR), false positive rate (FPR), receiver operating characteristic (ROC), and area under curve (AUC). The ROC curve can be drawn and the area under the ROC curve (AUC) can be obtained by sorting the samples with the scores of our method and different thresholds.

Ablation experiment

Under the Five-fold cross-validation and Loocv cross-validation, the ablation experiment was used to verify the contribution of the enhanced random walk module and the graph convolution-based selection algorithm module to the prediction of microbial-disease association (Table 1). In the absence of validation of enhanced random walk, AUC decreased by 6.2% and 2.8% respectively compared with our final model. The main reason is that the enhanced random walk module enhances the neighbor topology representation of microorganisms and disease nodes, which may improve the prediction performance. Compared with the model without the selection module based on graph convolution, the AUC performance of our method is improved by 9.1% and 5.5% under

Table 1 Results of ablation experiments on our method

Enhanced bi-directional random walk	Selection algorithm module based on graph convolution	Five fold cross validation Average AUC	Loocv cross validation Average AUC
×	✓	0.883	0.923
✓	×	0.854	0.896
✓	✓	0.945	0.951

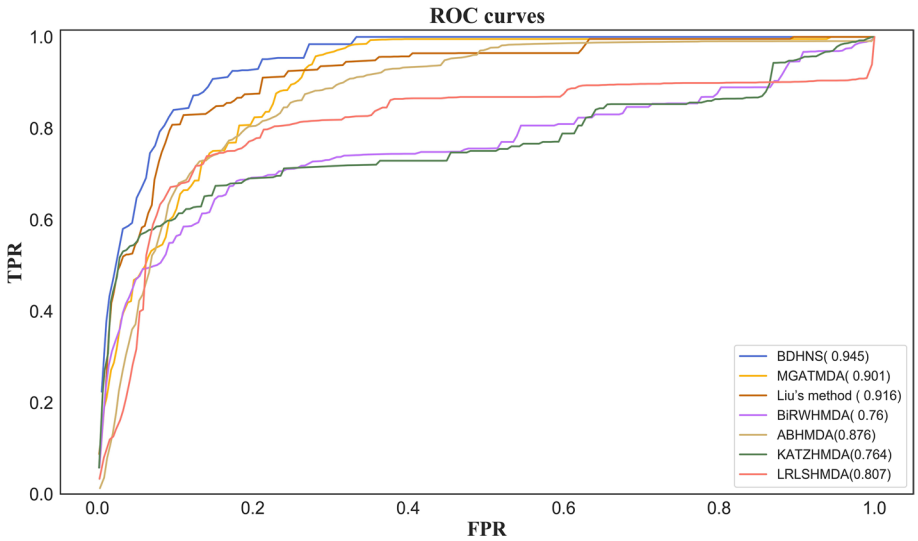

Fig. 4 ROC curve of five methods in 5-fold cross validation

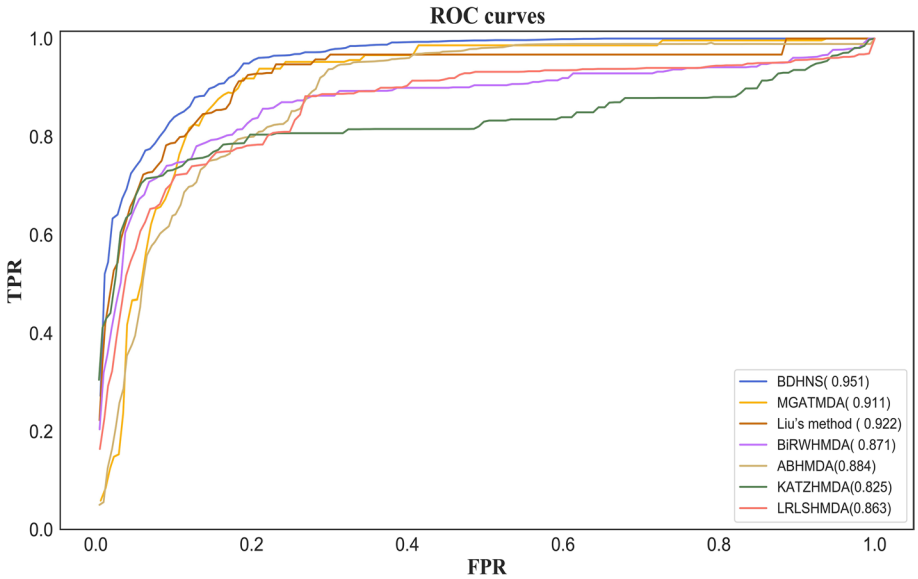

Fig. 5 ROC curve of five methods in Loocv

five-fold cross-validation and Loocv cross-validation, respectively. This indicates that it is necessary to selectively aggregate the attributes of neighbor nodes.

Compare other methods

We compared our method with four microbial disease association prediction methods, including BiRHMDA [4], KATZ [3], LRLSNMDA [7], ABHMDA [8], Liu's method [18], MGATMDA [19]. In five-fold cross-validation and leave-one-out cross-validation, all methods use the same data set for training and testing. Under the five fold cross validation and Loocv cross validation, the average curves of all methods are shown in Figs. 4 and 5, respectively.

Under the five-fold cross validation, the AUC value of our model is the best, which is 0.945, which is 2.9% higher than Liu's method, 4.4% higher than MGATMDA, 18.5% higher than BiRWHMDA method, 6.9% higher than ABHMDA method, 18.1% higher than KATZ method, and 13.8% higher than LRLSNMDA method. One possible reason why Liu's method ranks second is that it integrates the linear features of microorganisms and diseases obtained based on non negative matrix decomposition and random walk and the nonlinear features obtained based on capsule neural network. The third ranked mgatmda utilizes the graph attention network (GAT) to deeply mine the complex association between microorganisms and diseases. BiRWHMDA, ABHMDA, LRLSNMDA and KATZ are shallow prediction methods, which are difficult to deeply integrate the multi-level attributes of microbials and diseases nodes. One possible reason why ABHMDA performs better than BiRWHMDA, LRLSNMDA and KATZ is that it integrates similarity information across multiple microorganisms. The results show that this information is helpful to predict the microbial-disease association.

Under the Loocv cross validation, the AUC value of our method is still higher than that of other methods. The AUC value of our method is 0.951, which is 2.9% , 4%, 8%, 6.7%, 12.6% and 8.8% higher than Liu's method, MGATMDA, BiRWHMDA, ABHMDA, KATZ and LRLSNMDA, respectively.

Case studies

In order to further evaluate the predictive performance of our method for microbial-disease association, we conducted case studies on colon cancer and ulcerative colitis. First, we can obtain the association probability of each microbial candidate disease and rank it in descending order. Then, the first 10 candidate microorganisms of each disease were selected for validation and analysis.

Colon cancer is a life-threatening malignant tumor. The risk of colon cancer is closely related to intestinal flora [25]. In this study, 9 of the top 10 candidate microorganisms related to colon cancer predicted by our method have been confirmed by experiments, as shown in Table 2. Clostridia may promote Colon cancer [26]. It is found that enterobacteriaceae and proteobacteria are very common in colon cancer [27, 28]. Sidhu et al. [29] was found that the bacteria that can produce butyrate in the intestinal microbiota of colon cancer patients were greatly reduced. Clostridium coccoides is one of the bacteria

Table 2 Prediction results of top-10 Colon cancer-associated microbes

Disease name	Rank	Microbe name	Evidence
Colon cancer	1	Clostridia	PMID: 24603888
	2	Enterobacteriaceae	PMID: 26143056
	3	Clostridium coccoides	PMID:21850056
	4	Haemophilus	Unconfirmed
	5	Clostridium	PMID: 30857430
	6	Proteobacteria	PMID: 34650531
	7	Firmicutes	PMID: 34551683
	8	Bacteroidetes	PMID: 34551683
	9	Lactobacillus	PMID: 19647100
	10	Staphylococcus	PMID: 17530358

that can produce butyrate. There is evidence that clostridium is associated with colon cancer [30]. Compared with normal people, fermicutes and bacteroidetes in colon cancer patients decreased and increased significantly respectively [31]. Lactobacillus has immunoregulatory effect on human colon cancer cells [32]. Studies have proved that staphylococcus can prevent cancer [33].

Ulterative colitis is a chronic inflammatory disease of colon and rectum. Its pathogenesis is closely related to intestinal microbial imbalance [34]. Methods 9 of the top 10 candidate microorganisms related to ulcerative colitis predicted by us have been confirmed by experiments, as shown in Table 3. There is evidence that prevotella, clostridium difficile and proteobacteria have a great impact on the pathogenesis of ulcerative colitis [35–37]. Gryaznova et al. [38] reported patients with ulcerative colitis caused by staphylococcus aureus. There is evidence that haemophilus is increased in patients with ulcerative colitis [39]. Helicobacter pylori is also involved in the pathogenesis of ulcerative colitis [40]. The abundance of firmicutes in lymph nodes of patients with ulgenerative colitis is very high [41]. Erysipelotrichaceae was found to be a key bacterium related to ulcerative colitis [42]. It was found that the number of coriobacteriaceae decreased in patients with ulcerative colitis [43].

In conclusion, case studies of two diseases show that our method can indeed find potential microbial disease associations.

Summary

We propose a new microbial disease prediction method to learn and integrate multiple characteristics of diseases and microorganisms. Bi-directional heterogeneous networks are established to help integrate similarities and associations between microorganisms and diseases. An enhanced random walk module is established to learn the neighbor topology information of microorganisms and disease nodes. In order to selectively aggregate node features, a graph convolution-based selection algorithm is further established. In the Loocv and 5-fold cross-validation, the improved performance of our method in AUC was demonstrated by comparing with several microbial disease prediction models. The performance of our method is further demonstrated by the case studies of colon cancer and ulcerative colitis. In the future, we can integrate more types of data, such as gene sequencing and human metabolite data, to help predict the potential

Table 3 Prediction results of top-10 Ulcerative colitis-associated microbes

Disease name	Rank	Microbe name	Evidence
Ulterative colitis	1	Prevotella	PMID: 16585651
	2	Clostridium difficile	PMID: 21272802
	3	Staphylococcus aureus	PMID: 21683308
	4	Haemophilus	PMID: 33748490
	5	Proteobacteria	PMID: 250187840
	6	Helicobacter pylori	PMID: 30430119
	7	Firmicutes	PMID: 30239655
	8	Erysipelotrichaceae	PMID: 32169445
	9	Coriobacteriaceae	PMID: 31812509
	10	Clostridia	Unconfirmed

association between microorganism and disease. In addition, nodes in heterogeneous networks can be connected through different semantic meta paths. Therefore, in the future, we will integrate the information from the meta path in our method.

Acknowledgements
The authors thank the anonymous referees for their careful reading of our manuscript and their extensive comments.

Author contributions
JG, ZGZ designed the study. JG implemented the model. JG, ZGZ, YL performed experiments and analyses. JG drafted the manuscript and JG, ZGZ, YL, MW revised it. All authors have read and approved the final version of this manuscript.

Funding
This work was supported by the Natural Science Foundation of China (61972135); and This work was supported by the Natural Science Foundation of Heilongjiang Province in China (No. LH2020F043); and the Foundation of Graduate Innovative Research Project of Heilongjiang University(YJSCX2022-089HLJU).

Availability of data and materials
The datasets analyzed during the current study are downloaded from the website http://www.cuilab.cn/hmdad. The datasets used during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The data that we used are obtained from the public datasets (http://www.cuilab.cn/hmdad). Therefore, the ethics approval is not applicable for our study.

Consent for publication
Not applicable.

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

Received: 9 July 2022 Accepted: 21 September 2022
Published online: 14 November 2022

References

1. Reiff C, Kelly D. Inflammatory bowel disease, gut bacteria and probiotic therapy. Int J Med Microbiol. 2010;300:25–33.
2. Kreth J, Zhang Y, Herzberg MC. Streptococcal antagonism in oral biofilms: Streptococcus sanguinis and Streptococcus gordonii interference with Streptococcus mutans. Journal of Bacteriology. 2008;190:4632–40.
3. Chen X, Huang YA, You ZH, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017;33:733–9.
4. Li H, Wang YQ, Jiang JW, et al. A novel human microbe-disease association prediction method based on the bi-directional weighted network. Front Microbiol. 2019.
5. Zhao Y, Wang C-C, Chen X. Microbes and complex diseases: from experimental results to computational models. Brief Bioinform. 2020;22:bbaa158.
6. Ying C, Dandan S, Zelong X, Zhaogong Z, et al. Application of convolution neural network in nucleosome localization and recognition. J Harbin Eng Univ. 2021;42(5):8.
7. Wang F, Huang ZA, Chen X, et al. LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction. Sci Rep. 2017;7:7601.
8. Peng LH, Yin J, Zhou LQ, Liu MX, et al. Human microbe-disease association prediction based on adaptive boosting. Front Microbiol. 2018;9:2440.
9. Shen XJ, Chen Y, Jiang XP, et al. Prioritizing disease-causing microbes based on random walking on the heterogeneous network. Methods. 2017;124:120–5.
10. Zou S, Zhang JP, Zhang ZP. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. Plos One. 2017;12:e0184394.
11. Niu YW, Qu CQ, Wang GH, et al. RWHMDA: random walk on hypergraph for microbe-disease association prediction. Front Microbiol. 2017;10:1278.
12. Qu J, Zhao Y, Yin J. identification and analysis of human microbe-disease associations by matrix decomposition and label propagation. Front Microbiol. 2019;10:291.
13. Li L, et al. SCMFMDA: predicting microRNA-disease associations based on similarity constrained matrix factorization. PLOS COMPUT BIOL. 2021;17(7):e1009165.
14. Yin M-M, et al. NCPLP: a novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. IEEE Trans Cybern. 2022;52(6):5079–87.
15. Yin M-M, et al. Multi-similarity fusion-based label propagation for predicting microbes potentially associated with diseases. Futur Gener Comput Syst. 2022;134:247–55.
16. Long Y, Luo J, Zhang Y, et al. Predicting human microbe-disease associations via graph attention networks with inductive matrix completion. Brief Bioinform. 2020;22:146.

17. Lei XJ, Wang YY. Predicting microbe-disease association by learning graph representations and rule-based inference on the heterogeneous network. Fronti Microbiol. 2020;11:579.
18. Liu M, Dai W, Peng W, et al. A multi-view approach for predicting microbedisease associations by fusing the linear and nonlinear features. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM) 2020; p. 323–8.
19. Dayun L, Junyi L, Yi L, et al. MGATMDA: predicting microbe-disease associations via multi-component graph attention network. IEEE/ACM Trans Comput Biol Bioinform 2021; PMID: 34587092.
20. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88:265–6.
21. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 2006;22:2800–5.
22. Hwang S, Kim CY, Yang S, et al. HumanNet v2: human gene networks for disease research. Nucleic Acids Res. 2019;47:D573-80.
23. Zhou XZ, Menche J, Barabasi AL, et al. Human symptoms-disease network. NatCommun. 2014;5:1.
24. Feng F, Huang W. Should graph convolution trust neighbors? A simple causal inference method. 2021, arXiv:2010.11797v2.
25. Moore WE, Moore LH. Intestinal floras of populations that have a high risk of colon cancer. Appl Environ Microbiol. 1995;61:3202–7.
26. Zhu Q, Jin Z, Wu W. Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer. PLoS One. 2014;9(6):e90849.
27. Yurdakul D, Yazgan-Karataş A, Şahin F. Enterobacter strains might promote colon cancer. Curr Microbiol. 2015;71(3):403–11.
28. Wang T, Cai G, Qiu Y, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME. 2012;6(2):320–9.
29. Sidhu JS, Mandal A, Virk J, et al. Early detection of colon cancer following incidental finding of clostridium septicum bacteremia. J Investig Med High Impact Case Rep. 2019;7:2324709619832050.
30. Jin M, Shang F, Wu J, et al. Tumor-associated microbiota in proximal and distal colorectal cancer and their relationships with clinical outcomes. Front Microbiol. 2021;12:727937.
31. He T, Cheng X, Xing C. The gut microbial diversity of colon cancer patients and the clinical significance. Bioengineered. 2021;12(1):7046–60.
32. Paolillo R, Romano Carratelli C, Sorrentino S, et al. Immunomodulatory effects of Lactobacillus plantarum on human colon cancer cells. Int Immunopharmacol. 2009;9(11):1265–71.
33. Noguchi N, Ohashi T, Shiratori T, et al. Association of tannase-producing Staphylococcus lugdunensis with colon cancer and characterization of a novel tannase gene. J Gastroenterol. 2007;42(5):346–51.
34. Yang H, Mirsepasi-Lauridsen HC, Struve C, et al. Ulcerative Colitis-associated E. coli pathobionts potentiate colitis in susceptible hosts. J Gut Microbes. 2020;12(1):1847976.
35. Lucke K, Miehlke S, Jacobs E, et al. Prevalence of bacteroides and prevotella spp.in ulcerative colitis. J Med Microbiol. 2006;55(Pt 5):617–24.
36. Kariv R, Navaneethan U, Venkatesh PG, Lopez R, et al. Impact of Clostridium difficile infection in patients with ulcerative colitis. J Crohns Colitis. 2011;5(1):34–40.
37. Albuquerque A, Magro F, Rodrigues S, et al. Liver abscess of the caudate lobe due to Staphylococcus aureus in an ulcerative colitis patient: first case report. J Crohns Colitis. 2011;5(4):360–3.
38. Gryaznova MV, Solodskikh SA, Panevina AV, et al. Study of microbiome changes in patients with ulcerative colitis in the Central European part of Russia. Heliyon. 2021;7(3):e06432.
39. Walujkar SA, Dhotre DP, Marathe NP, et al. Characterization of bacterial community shift in human Ulcerative Colitis patients revealed by Illumina based 16S rRNA gene amplicon sequencing. Gut Pathog. 2014;6:22.
40. Mansour L, El-Kalla F, Kobtan A, et al. Helicobacter pylori may be an initiating factor in newly diagnosed ulcerative colitis patients: a pilot study. Bioinformatics. 2018;6(13):641–9.
41. Kiernan MG, Coffey JC, McDermott K, et al. The human mesenteric lymph node microbiome differentiates between Crohn's disease and ulcerative colitis. J Crohns Colitis. 2019;13(1):58–66.
42. Sun J, Chen H, Kan J, et al. Anti-inflammatory properties and gut microbiota modulation of an alkali-soluble polysaccharide from purple sweet potato in DSS-induced colitis mice. Int J Biol Macromol. 2020;153:708–22.
43. Pittayanon R, Lau JT, Leontiadis GI, et al. Differences in gut microbiota in patients with vs without inflammatory bowel diseases: a systematic review. Gastroenterology. 2020;158(4):930–46.

Publisher's Note