De-Shuang Huang
Kang-Hyun Jo (Eds.)

LNCS 12464

Intelligent Computing Theories and Application

16th International Conference, ICIC 2020
Bari, Italy, October 2–5, 2020
Proceedings, Part II

2 Part II

Springer

A Graph Convolutional Matrix Completion Method for miRNA-Disease Association Prediction

Wei Wang[1], Jiawei Luo[1](✉), Cong Shen[1], and Nguye Hoang Tu[2]

[1] College of Computer Science and Electronic Engineering,
Hunan University, Changsha 410083, China
luojiawei@hnu.edu.cn

[2] Faculty of Information and Technology, Hanoi University of Industry,
Hanoi 100803, Vietnam

Abstract. MicroRNAs (miRNAs) play a key role in various biological processes associated with human diseases. Identification of miRNA-disease relationships can help to understand disease pathogenesis. Experimentally verifying substantial associations between miRNAs and diseases is the most convincing but time-consuming, while in silico methods can provide efficient alternatives. However, existing computational methods still have room for improvement in considering topology and prior information of network nodes. In this paper, we presented a novel model called GCMCAP, in which we referred to the prediction of potential miRNA-disease associations as a recommendation problem. In our framework, we integrated graph convolution networks as feature extractors into a matrix completion to predict diseases related miRNAs. We tested GCMCAP and other three methods on the same dataset. The results indicate that GCMCAP outperforms other methods with respect to average AUC value. In addition, case studies show that GCMCAP has a great capability to discover novel miRNA-disease associations.

Keywords: MicroRNA · Disease · Association · Graph convolution networks · Matrix completion

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that can regulate gene expression during post-transcription and influence the output of protein-coding genes [1]. Existing researches have shown that miRNAs are involved in many biological processes, such as differentiation [2], development [3], immune reaction [4], apoptosis [5], and pathogenesis. There are already compelling evidences that human miRNAs are associated with complex diseases [6]. For example, studies by Zare et al. have shown that miRNA expression is a regulator of tumorigenesis [7]. In addition, the Human MicroRNA Disease Database (HMDD v2.0) has shown that more than 10,000 associations between miRNA and disease have been identified [8], involving 378 diseases and 572 miRNAs. Therefore, it is necessary to discover other potential associations between miRNA and disease in order to more fully understand pathogenesis of human

D.-S. Huang and K.-H. Jo (Eds.): ICIC 2020, LNCS 12464, pp. 201–215, 2020.
https://doi.org/10.1007/978-3-030-60802-6_18

diseases associated with miRNAs, thereby facilitating disease diagnosis, treatment, and prevention. Identifying the associations between miRNA and disease through biological experiments is the most convincing but time-consuming, while in silico methods can provide efficient alternatives [9–11].

The most of these proposed methods are based on the assumption that similar miRNAs are related to similar diseases [12, 13]. For example, Zhao et al. discovered disease related miRNA candidates using gene expression data and miRNA-gene regulation [14]. Chen et al. proposed a method to identify disease-related miRNAs by random walks with restart on the miRNA similarity network [15]. Xuan et al. evaluated the *k* most functionally similar neighbors by considering the disease terms and phenotype similarity [16]. You et al. constructed a heterogeneous network by integrating different types of heterogeneous biodata sets and proposed a path-based approach to calculate the association score between miRNA and disease [17]. Zou et al. predicted miRNA-disease association based on social network analysis [18]. Xiao et al. used graph regularization non-negative matrix factorization to identify microRNA-disease associations [19]. Luo et al. predicted Small Molecule-microRNA Associations based on Non-negative Matrix Factorization [20]. All these methods make good use of the prior information in the miRNA/disease interaction network.

On the other hand, machine learning-based methods are proposed as many known associations are confirmed by biological experiments. Based on transduction learning, Luo et al. developed a collective prediction method of disease-associated miRNAs [21]. Chen et al. presented a computational model named Laplacian Regularized Sparse Subspace Learning, which projected miRNAs/diseases' graph feature profile to a common subspace [22]. These results show that machine learning based methods improve the performance of association prediction. Based on inductive matrix completion, Ding, X et al. exploited an improved computation method. Li et al. released an effective computational model of Matrix Completion for MiRNA-Disease Association prediction (MCMDA) [23]. Recently, methods combined neural network have been used for association prediction. For example, Hou et al. combined neural network model and induction matrix completion (NIMC) to predict disease-gene association [24]. Xuan, P et al. proposed a method to predict disease related miRNAs based on network representation learning and convolutional neural networks [25]. Han, P et al. predicted disease-gene association by integrating graph convolutional network and matrix factorization [26].

The methods based on similarity measure can effectively integrate heterogeneous data, but it is limited by known associations. On the other side, machine learning-based methods have impressive results for miRNA-disease association predictions, but there is still room for improvement in using miRNAs and disease prior information. In this paper, we referred to the prediction of potential miRNA-disease associations as a recommendation problem and proposed a novel computational framework named GCMCAP. We integrated graph convolutional neural networks into matrix completion to predict the associations between miRNA and disease. With the help of graph convolutional networks, non-linear and high-order neighborhood information can be captured. In addition, the problem of no negative sample in training progress can be circumvented. We tested GCMCAP and other three methods on the same dataset. The results suggest that GCMCAP outperforms other methods in terms of average AUC values. Moreover, case studies show that GCMCAP has a great capability to discover novel miRNA-disease associations.

2 Materials and Methods

2.1 Datasets

We downloaded miRNA-gene interactions from experimentally verified databases, including miRecords v4.0 [27], TarBase v6.0 [28] and miRTarBase v4.5 [29]. After unioning and removing duplicates, we got 38,089 interactions between miRNAs and genes, involving 477 miRNAs and 12,422 genes. We downloaded gene-gene interaction network from HumanNet which contains 476,399 interactions among 16,243 genes [30]. We downloaded validated miRNA-disease associations datasets from the HMDD database v2.0. As done in Xuan et al. [31], we regarded multiple miRNA transcripts as the same mature miRNA. So we acquired 5424 associations involving 378 diseases and 495 miRNAs from HMDD v2.0. We downloaded the disease hierarchical directed acyclic graphs (DAGs) from MeSH (https://www.nlm.nih.gov/mesh/). To ensure consistency of miRNAs, diseases and associations, we removed some irregularly named diseases in MeSH and eliminated miRNAs that are missing from the three miRNA target databases mentioned above. 4887 experimentally validated miRNA-disease associations involving 327 diseases and 351 miRNAs were retained. Consequently, miRNA matrix $S_m \in R^{m \times m}$, disease matrix $S_d \in R^{d \times d}$ and miRNA-disease association matrix $S \in R^{m \times d}$ were formed for prediction task.

2.2 Method Overview

We regarded the prediction of miRNA-disease associations as a recommendation problem with four main steps. In the first step, we calculated miRNA-miRNA and disease-disease pairwise similarities and compiled them into two adjacency matrices. In the second step, we respectively performed a multi-layer graph convolutional networks on the adjacency matrices to assign node feature to each miRNA and disease. In such way, the high-order neighborhood information of each miRNA and disease node is encoded into embeddings. In the third step, we modeled association ratings as the inner product of the embeddings projected onto a latent space. In the last step, we used matrix completion principle to obtain the ratings for each miRNA-disease association. In this stage, the embeddings of miRNA/disease are adjusted by minimizing the difference between the reconstruction matrix and the initial matrix. Figure 1 depicts the whole framework of the proposed method.

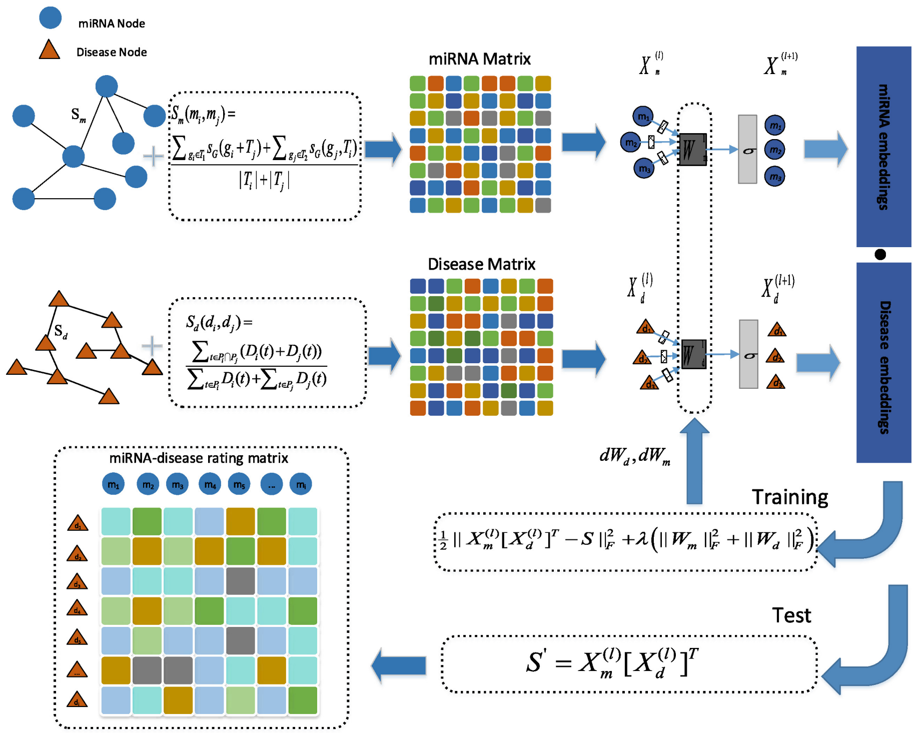

miRNA Node
Disease Node
S_m
$S_m(m_i,m_j)=\frac{\sum_{g_i\in T_1} s_G(g_i,T_j)+\sum_{g_j\in T_2} s_G(g_j,T_i)}{|T_i|+|T_j|}$
miRNA Matrix
$X_m^{(l)}$
$X_m^{(l+1)}$
m_1
m_2
m_3
W
σ
miRNA embeddings
S_d
$S_d(d_i,d_j)=\frac{\sum_{t\in P_i\cap P_j}(D_i(t)+D_j(t))}{\sum_{t\in P_i}D_i(t)+\sum_{t\in P_j}D_j(t)}$
Disease Matrix
$X_d^{(l)}$
$X_d^{(l+1)}$
d_1
d_2
d_3
Disease embeddings
dW_d, dW_m
Training
$\frac{1}{2}\| X_m^{(l)}[X_d^{(l)}]^T - S\|_F^2 + \lambda(\| W_m\|_F^2 + \| W_d\|_F^2)$
Test
$S' = X_m^{(l)}[X_d^{(l)}]^T$
miRNA-disease rating matrix
m_1 m_2 m_3 m_4 m_5 ... m_i
d_1 d_2 d_3 d_4 d_5 ... d_i

Fig. 1. Workflow of GCMCAP

2.3 Diseases Semantic Similarity and MiRNA Functional Similarity

The MeSH database provides a valuable reference system in the form of DAGs (http://www.ncbi.nlm.nih.gov/). Disease similarity can be calculated using DAGs according to Wang's method [32]. In DAGs, nodes and edges represent diseases and the associations between diseases respectively. We calculated the similarity between the two diseases through the disease hierarchical relationship in DAGs. The semantic value of disease d is calculated by the following formula:

$$\mathrm{D(t)} = \begin{cases} \max\{\Delta * \mathrm{D(t')} | \in \mathrm{childrenof(t)}\} & t \neq d \\ D(\mathrm{d}) = 1 & t = d \end{cases} \tag{1}$$

where Δ represents the semantic contribution factor. We set Δ as 0.5 as suggested by Wang et al.. According to the assumption that diseases that share a large part of the DAG tend to have higher semantic similarity, the following calculation methods are available:

$$\mathrm{S_d}\left(\mathrm{d_i},\mathrm{d_j}\right) = \frac{\sum_{\mathrm{t}\in \mathrm{P_i}\cap \mathrm{P_j}} \left(\mathrm{D_i(t)} + \mathrm{D_j(t)}\right)}{\sum_{\mathrm{t}\in \mathrm{P_i}} \mathrm{D_i(t)} + \sum_{\mathrm{t}\in \mathrm{P_j}} \mathrm{D_j(t)}} \tag{2}$$

$D_i(t)$ and $D_j(t)$ are the semantic values related to disease $\mathrm{d_i}$ and disease $\mathrm{d_j}$, respectively. The semantic similarity of each two diseases is calculated based on the position of the two diseases in the DAG and their semantic associations with ancestor diseases.

To avoid reliance on existing associations between miRNA and disease, inspired by Xiao et al., we estimated the similarity between miRNAs using a weighted gene interaction network and an experimentally validated miRNA-target regulatory relationship. Initially the gene function interaction network is downloaded from the HumanNet. Interaction score between two genes indicates the strength of the association between genes. We used the following normalization technique on the gene interaction network to get the pairwise similarity score $\mathrm{S_g}$:

$$\mathrm{S_g}\left(\mathrm{g_i},\mathrm{g_j}\right) = \frac{\mathrm{S_G}\left(\mathrm{g_i},\mathrm{g_j}\right) - \mathrm{S_{min}}}{\mathrm{S_{max}} - \mathrm{S_{min}}} \tag{3}$$

where $S_G\left(g_i, g_j\right)$ denotes similarity score before normalization. We obtained the miRNA-target regulatory relationship from collated miRNA-gene data. Then the similarity between the gene and the target gene set is defined as the maximum similarity between the gene and others. It can be described by the following formula:

$$\mathrm{S}_G(\mathrm{g_t},\mathrm{T}) = \max_{g_{ti}\in T}\left(\mathrm{S_g}(\mathrm{g_t},\mathrm{g_{ti}})\right) \tag{4}$$

Based on the assumption that the greater the number of common target genes, the greater the similarity between miRNAs, the functional similarity $\mathrm{S_m}\left(\mathrm{m_i},\mathrm{m_j}\right)$ between miRNA m_i and m_j can be calculated by the following BMA method [33]:

$$\mathrm{S_m}\left(\mathrm{m_i},\mathrm{m_j}\right) = \frac{\sum_{\mathrm{g_i}\in \mathrm{T_1}} \mathrm{S_G}\left(\mathrm{g_i},\mathrm{T_j}\right) + \sum_{\mathrm{g_j}\in \mathrm{T_2}} \mathrm{S_G}\left(\mathrm{g_j},\mathrm{T_i}\right)}{\left|\mathrm{T_i}\right| + \left|\mathrm{T_j}\right|} \tag{5}$$

where T_i and T_j denote target gene sets of m_i and m_j respectively.

2.4 Graph Convolutional Feature Extractor

The graph convolutional network (GCN) is a kind of multilayer neural network with graph as input. It represents neighbor node features and network information as output vectors. GCNs have been successfully applied in areas such as recommendation systems and drug interactions [34, 35]. In this paper, we resorted to GCNs to convert miRNA and disease networks into embeddings and map the learned embeddings into latent space. The outputs will be applied to the matrix completion in the downstream framework.

Based on the assumption that node features are associated with all of its neighbors, we employed a simple solution to integrate network information. Let miRNA and disease similarity networks (adjacent matrices) be S_m and S_d, respectively. We defined the multiplication of S_m and X_m to obtain the neighborhood information of the current miRNA node. So we obtained the following neural network propagation rule for miRNA nodes.

$$f(X_m^{(l)}, S_m) = \sigma(S_m X_m^{(l)} W_m^{(l)}) \tag{6}$$

where σ denotes the non-linear activation function, $X_m^{(l)}$ denotes the miRNA features output by the l-layer neural network, and $W_m^{(l)}$ represents the weight of the miRNA features. However, this method of multiplication modeling can easily lead to overfitting the local neighborhood structure of a graph with a wide node degree distribution. Therefore, we treated each initial node feature X_m as a graph signal according to spectral graph theory, and use the spectral convolution operation $S_m \bigstar X_m$ on the graph S_m to replace the multiplication. Referring to the convolution theorem that convolution operation is equivalent to the product after Fourier transform, the following equation can be obtained:

$$S_m \bigstar X_m = U_S((U_S^T g) \odot (U_S^T X_m)) \tag{7}$$

where U_S is the eigenvector matrix of S_m, g represents the convolution kernel, and $\odot$ represents the hardmard product. To simplify formula (2), we treat diagonal matrix $g_\theta = diag\left(U_S^T g\right)$ as parameterized convolution kernel. Then the hardmard product can be changed to the form of matrix multiplication. However, Eq. (7) is still limited by 3 problems. First, the information of miRNA nodes is very important for network feature extraction, but the adjacency matrix S_m does not contain node information. Second, S_m is not normalized, which will lead to a larger feature values extracted by nodes with more neighbors. Third, Eq. (7) requires the eigendecomposition of S_m, which has high computational complexity for a large miRNA/disease networks.

To solve the first two limitations, we used $\tilde{S}_m = S_m + I$ to add the identity matrix I to the miRNA matrix to form a self-loop, and used the normalized Laplacian matrix $L_m = I - D^{-1/2} S_m D^{-1/2}$ to represent the network structure. The normalized Laplacian matrix $\tilde{L}_m$ with self-loop is obtained, i.e.

$$\tilde{L}_m = I - \tilde{D}_m^{-\frac{1}{2}} \tilde{S}_m \tilde{D}_m^{-\frac{1}{2}} \tag{8}$$

where $\tilde{D}_m$ denotes the diagonal matrix $\left[\tilde{D}_m\right]_i = \sum_{j=1}^n \left[\tilde{S}_m\right]_{ij}$. In this way, we can perform eigendecomposition on $\tilde{L}_m$, and replace the U_S with the eigenvector matrix U_L obtained from the feature decomposition. Equation (7) can be transformed into the following form.

$$S_m \bigstar X_m = U_L g_\theta U_L^T X_m \tag{9}$$

To circumvent the third limitation, we used Kipf's method to avoid eigendecomposition [36]. Based on Eq. (9), the first-order Chebyshev polynomial is used as the convolution kernel, and the maximum eigenvalue λ of L_m is approximated to 2. The propagation rule of miRNA nodes on the graph S_m can be written as:

$$X_m^{(l+1)} = \sigma(\tilde{D}_m^{-\frac{1}{2}}\tilde{S}_m\tilde{D}_m^{-\frac{1}{2}}X_m^{(l)}W_m^{(l)}) \tag{10}$$

where $X_m^{(l)} \in R^{m \times e}$ denotes the miRNA features output by l-layer network, m is the number of miRNA nodes, and e is the embedding size. The multi-layer networks can be achieved by stacking Eq. (10). Multi-layer networks imply the use of higher-order neighborhood information. Similarly, the propagation rule of disease nodes on the graph S_d can be written as:

$$X_d^{(l+1)} = \sigma(\tilde{D}_d^{-\frac{1}{2}}\tilde{S}_d\tilde{D}_d^{-\frac{1}{2}}X_d^{(l)}W_d^{(l)}) \tag{11}$$

where $X_d^{(l)} \in R^{d \times e}$ denotes the disease features output by l-layer network, d is the number of disease nodes. Disease and miRNA embeddings are trained simultaneously.

2.5 Matrix Completion

Matrix completion has been well used in recommendation system [37]. Similarly, we considered association prediction as a recommendation problem and use known miRNA-disease associations to recover missing entries in association matrix $S \in R^{m \times d}$. Where m is the number of miRNAs. For each entry in matrix S, $\mathrm{S_{ij}} = 1$ if the miRNA is associated with the disease, otherwise, $\mathrm{S_{ij}} = 0$. Let S' be the matrix to be completed. We modeled association ratings as the inner product of the features of miRNAs and diseases projected onto a latent space. i.e. $S' = KH$, where $K \in R^{m \times j}$ and $H \in R^{d \times j}$ denote miRNA and disease space, respectively. According to low-rank assumption, j satisfies $j < < m, d$. The following formula can be defined:

$$\min_{\mathrm{K,H}} \frac{1}{2}\left\|\mathrm{KH^T} - \mathrm{S}\right\|_{\mathrm{F}}^2 + \lambda\left(\|\mathrm{K}\|_{\mathrm{F}}^2 + \|\mathrm{H}\|_{\mathrm{F}}^2\right) \tag{12}$$

where $\| * \|_{\mathrm{F}}$ denotes Frobenius norm and λ denotes the regularization coefficient. However, the classical matrix completion cannot directly take advantage of the embeddings of known miRNA and disease. Inspired by Natarajan et al., we set T_1 and T_2 as projection matrices [38]. The features of miRNA and disease can be mapped into latent spaces with the same dimensions by $K = X_m T_1$ and $H = X_d T_2$, respectively. Therefore, the known associations and the features of miRNA and disease can be simultaneously used by the following improved matrix completion.

$$\min_{\mathrm{T_1,T_2}} \frac{1}{2}\left\|X_m T_1 T_2^T X_d^T - \mathrm{S}\right\|_{\mathrm{F}}^2 + \lambda\left(\|\mathrm{T_1}\|_{\mathrm{F}}^2 + \|\mathrm{T_2}\|_{\mathrm{F}}^2\right) \tag{13}$$

2.6 Training and Evaluation

A classic matrix completion for miRNA-disease association prediction considered the miRNA matrix X_m and disease matrix X_d as inputs. T_1 and T_2 represent the low rank decomposition of the projection matrix T. In proposed method, the similarity matrices X_m and X_d are encoded by graph convolutional networks, respectively. The linear projection $X_m T_1 T_2 X_d$ is replaced by the nonlinear rating $X_m^{(l)}\left[X_d^{(l)}\right]^T$ generated by the graph convolution network. According to the above annotation, the loss function of the graph convolutional matrix completion model can be defined as:

$$loss = \frac{1}{2}\left\| X_m^{(l)}\left[X_d^{(l)}\right]^T - S \right\|_F^2 + \lambda\left(\|W_m\|_F^2 + \|W_d\|_F^2 \right) \tag{14}$$

where $X_m^{(l)} = \sigma\left(\tilde{D}_m^{-\frac{1}{2}}\tilde{S}_m\tilde{D}_m^{-\frac{1}{2}}X_m W_m^{(l)}\right)$ and $X_d^{(l)} = \sigma\left(\tilde{D}_d^{-\frac{1}{2}}\tilde{S}_d\tilde{D}_d^{-\frac{1}{2}}X_d W_d^{(l)}\right)$ denote the outputs of graph convolution networks. Considering the advantage that all operations are differentiable, the proposed model can be train in an end-to-end workflow by gradient descent algorithm. Algorithm 1 shows the overall prediction procedure.

Algorithm 1. GCMCAP Algorithm

```
Input: association matrix $S \in R^{m \times n}$; similarity matrix $S_m \in R^{m \times m}$, $S_d \in R^{d \times d}$; learning rate $\gamma$ ; iterations t; embedding size e; regularization parameters $\lambda$ ;
for p=0 : t do
    Apply the GCNs on $S_d$ to produce embeddings $X_d^{(l)} \in R^{d \times e}$
    Apply the GCNs on $S_m$ to produce embeddings $X_m^{(l)} \in R^{m \times e}$
  for i = 1 : m do
      for j=1:d do
          Apply embeddings $\left[X_m^{(l)}\right]_i$ and $\left[X_d^{(l)}\right]_j$ to generate predicted values $S'_{ij}$
          $loss = \frac{1}{2} \| S' - S \|_F^2 + \lambda\left( \| W_m \|_F^2 + \| W_d \|_F^2 \right)$
      end for
  end for
  Update the GCNs parameters $W_m$ and $W_d$
end for
Apply $S' = X_m^{(l)}\left[X_d^{(l)}\right]^T$ to produce ratings
Output: rating matrix S'
```

3 Results

In this part, we verified the practicality of the proposed method GCMCAP through experiments. Firstly, the evaluation indicators for the performance of all methods are introduced. Then the performance of the GCMCAP and several other common miRNA-disease prediction methods are compared. Next, we carried out parameter analysis to verify the reliability and robustness of the model. Finally, case studies are arranged to explore the capacity of GCMCAP to discover novel disease-associated miRNAs.

3.1 Experiment Settings

To evaluate the performance of the GCMCAP algorithm, we performed 5-fold cross validation (CV) and 10-fold CV. Taking 5-fold CV as an example, in each round, the known miRNA-disease associations are randomly divided into five disjoint subsets. One subset is used as the validation set and the remaining subsets are utilized as the training set. For each fold, embedding size e is set from $\{32, 64, 128, 256\}$, learning rate γ is set from $\{0.001, 0.01, 0.1\}$, the number of GCN layers l is set from $\{2, 3, 4, 5\}$ and probability of dropout is set from $\{0.5, 1\}$. All parameters are considered based on grid search. All experiments are repeated for 5 times, the reports are average of the 5 runs. Refer to Kipf's method, we used the identity matrix as the initial feature matrix to feed GCNs. All methods processes are implemented using Tensorflow framework (v1.9) and trained using the Adam stochastic optimization algorithm [39].

3.2 Evaluation Metrics

After the training progress completed, we obtained all association ratings from the rating matrix. The four values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) are calculated. The true positive rate(TPR),false positive rate (FPR), Precision, and Recall were calculated as following formulas:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{15}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TP}}, \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

we used the TPRs and FPRs to plot the receiver operating characteristic (ROC) curves. Then the area under the ROC curves (AUCs) are used to measure the global performance of models.

To get indicators that reflects global performance in the event of class imbalance, the mean Average Precision (mAP) is used to solve the single-point value limitation of the Precision. We focused more on the set of top K miRNA related to a disease, the sequence of the whole recommendation list may be not important. Therefore, we utilized Precision@N and Recall@N to evaluate the quality of the recommendation list. Where N denotes the percentage of sorted rating results. We set N = $\{10, 20, 30\}$ in all experiments.

3.3 Performance Evaluation

To assess the performance of GCMCAP, we compared it with the other three methods: RWRMDA [15], KATZ [18], MIDPE [31]. RWRMDA is a plain baseline and utilizes random walk algorithm to identify potential unknown association. KATZ combines social network analysis methods with machine learning and predicts unknown miRNA-disease relationships. MIDPE completely integrates various ranges of topologies around the different categories of nodes. We reproduced the three methods based on the formulas in the papers, respectively. We used 5-fold CV and 10-fold CV to compare our method with existing methods based on the same dataset. Table 1 and Table 2 list the results of all methods. GCMCAP outperforms other methods with respect to the most indicators. In 5-fold CV, our method yields an average AUC value of 0.894, which is better than RWRMDA (0.804), MIDPE (0.702) and KATZ (0.864). Table 2 shows that the results of 10-fold CV are similar to the 5-fold CV, and GCMCAP is outperform other methods. Furthermore, P@10 and R@10 of GCMCAP are significantly higher than existing methods, which shows a better performance in top *K* prediction ability. Figure 2 shows the ROC curves of various comparison methods, which suggests the overall performance of the models. σ in the figure denotes the standard deviation. The ROC curve generated by GCMCAP is above all the curves and closer to the upper left corner of the coordinate system. The results suggest that the associations in the similarity matrix maybe not directly reflect the relationship between miRNA or disease. Since GCMCAP integrates graph convolution networks into matrix completion method, non-linear neighborhood information can be abstracted from miRNA network and disease network. Furthermore, GCMCAP does not rely on negative samples for training, which effectively circumvent the problem of no negative sample.

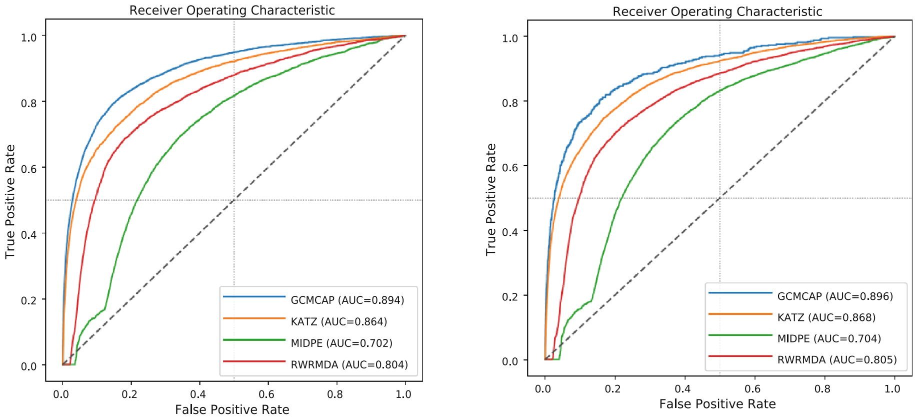


Fig. 2. (a) ROC curves of 5-fold CV; (b) ROC curves of 10-fold CV.

Table 1. Results on 5-fold cross validation

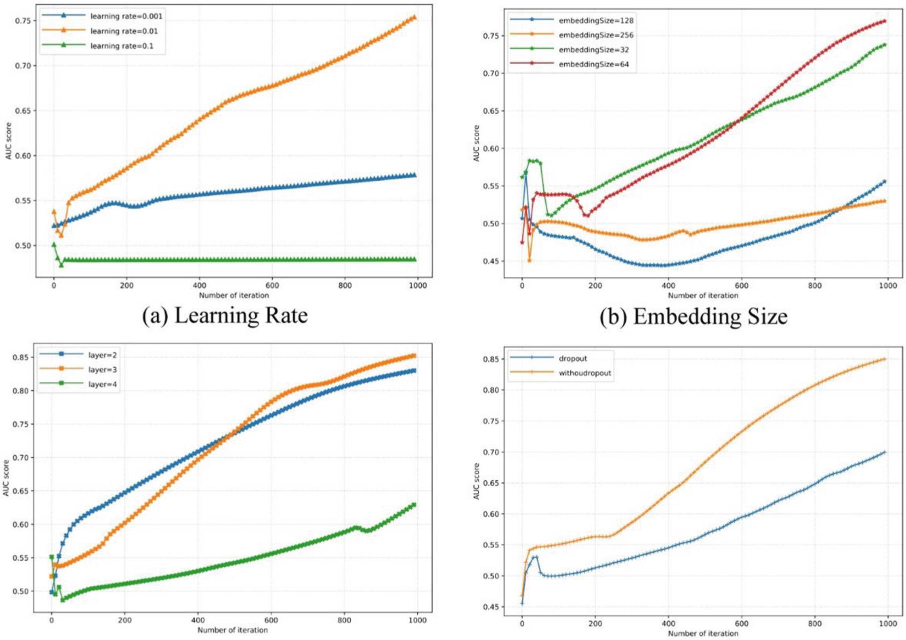
Method	MAP	AUC	P@10	P@20	P@30	R@10	R@20	R@30
RWRMDA	0.032	0.804	0.026	0.039	0.046	0.102	0.207	0.309
MIDPE	0.016	0.702	0.015	0.014	0.018	0.114	0.219	0.325
KATZ	0.14	0.864	0.267	0.206	**0.163**	0.134	0.245	0.346
GCMCAP	**0.197**	**0.894**	**0.347**	**0.223**	0.153	**0.189**	**0.319**	**0.449**

Table 2. Results on 10-fold cross validation

Method	MAP	AUC	P@10	P@20	P@30	R@10	R@20	R@30
RWRMDA	0.016	0.805	0.012	0.018	0.022	0.102	0.204	0.306
MIDPE	0.008	0.704	0.008	0.007	0.009	0.123	0.241	0.358
KATZ	0.083	0.868	0.17	0.12	**0.099**	0.119	0.223	0.327
GCMCAP	**0.153**	**0.896**	**0.238**	**0.142**	0.093	**0.225**	**0.368**	**0.501**

3.4 Parameter Analysis

In this section, we analyzed the effects of parameters through the parameter adjustment. Because AUC is an indicator that can evaluate the comprehensive performance of the model, the influence of parameter changes on the AUC value is analyzed. Learning rate is an important parameter for the optimization model using gradient descent algorithm. We fixed the other parameters and set the learning rate from $\{0.001, 0.01, 0.1\}$. Figure 3(a) shows that there may be an optimal value for the initial learning rate. A small learning rate will cause the model to converge slowly, conversely a large learning rate may make it difficult to converge. Figure 3(b) shows that the embedding size does not affect convergence of GCMCAP within an appropriate size range, which indicates that the proposed method has the ability to obtain prior information steadily. However, when the embedding size is too large, too many parameters make the model difficult to train or even overfitting. This pattern is consistent with other related studies [40].



(a) Learning Rate

(b) Embedding Size

(c) Number of Layer

(d) Dropout

Fig. 3. (a) ROC curves of 5-fold CV; (b) ROC curves of 10-fold CV.

The model with a small number of GCN layers l perform well, and the performance decreases rapidly when $l > = 4$ according to Fig. 3(c). The increase of layer may capture the more global information, but also the more noise is captured. Dropout is a commonly used method to improve performance by avoiding overfitting. However, there is no obvious effect of the dropout on the performance of GCMCAP according to Fig. 3(d). The possible reason is that the effect of dropout is not obvious due to sparse data.

3.5 Case Studies

To further explore the capacity of GCMCAP to discover potential miRNAs associated with disease, 3 given diseases of Glioma, Carcinoma Hepatocellular and Ovarian Neoplasms are analyzed. All known associations in HMDD v2.0 are used to train and unknown associations are used for validation. For each disease, the candidate miRNAs are ranked based on the ratings. The top 10 miRNA candidates are obtained from prediction results.

The latest human microRNA disease database (HMDD v3.2) is used to confirm miRNA candidates for given 3 diseases. HMDD V3.2 provides extensive experimentally supported evidence for human microRNA (miRNA) and disease associations. HMDD v3.2 contains twice as many human miRNA disease associations as previous HMDD v2.0. As shown in Table 3, HMDD v3.2 confirms 6,6 and 5 miRNA candidates are associated with Glioma, Carcinoma Hepatocellular and Ovarian Neoplasms, respectively. In addition, some candidates also rank higher in other methods. For example, 3, 2, and 3 miRNAs are identified by KATZ as the top ten in three diseases, respectively. The results show that miRNA candidates predicted by GCMCAP are very reliable and GCMCAP has stable prediction performance. Furthermore, most confirmed candidates have higher rankings, indicating that top K performance of GCMCAP is outstanding.

Table 3. Evidences of the top 10 associated miRNA candidates for the three given diseases

Disease No. of miRNAs confirmed	By the latest HMDD	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Glioma	6	1	hsa-mir-219	HMDD v3.2	6	hsa-mir-429	HMDD v3.2
		2	hsa-mir-136	HMDD v3.2	7	hsa-mir-642a	Unconfirmed
		3	hsa-mir-135b	HMDD v3.2	8	hsa-mir-103a	Unconfirmed
		4	hsa-mir-216a	Unconfirmed	9	hsa-mir-29c	HMDD v3.2
		5	hsa-mir-195	Unconfirmed	10	hsa-mir-296	HMDD v3.2

(*continued*)

Table 3. (*continued*)

Disease No. of miRNAs confirmed	By the latest HMDD	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Carcinoma, Hepatocellular	6	1	hsa-mir-1296	HMDD v3.2	6	hsa-mir-519a	HMDD v3.2
		2	hsa-mir-632	Unconfirmed	7	hsa-mir-219	Unconfirmed
		3	hsa-mir-17	HMDD v3.2	8	hsa-mir-490	HMDD v3.2
		4	hsa-mir-379	HMDD v3.2	9	hsa-mir-545	Unconfirmed
		5	hsa-mir-518a	HMDD v3.2	10	hsa-mir-659	Unconfirmed
Ovarian Neoplasms	5	1	hsa-mir-134	HMDD v3.2	6	hsa-mir-379	Unconfirmed
		2	hsa-mir-1296	Unconfirmed	7	hsa-mir-518a	Unconfirmed
		3	hsa-mir-632	Unconfirmed	8	hsa-mir-519a	HMDD v3.2
		4	hsa-mir-17	HMDD v3.2	9	hsa-mir-219	HMDD v3.2
		5	hsa-mir-362	Unconfirmed	10	hsa-mir-490	HMDD v3.2

4 Discussion and Conclusion

Identifying miRNA-disease relationships helps to understand disease pathogenesis, diagnosis and treatment. We make observation that existing computational methods still have room for improvement in considering the topology and prior information of network nodes. To adapt to the graph structured disease network and miRNA network, we use graph convolutional networks to model miRNA and disease node embeddings. In addition, GCMCAP also takes advantage of matrix completion theory to circumvent the problem of no negative sample in training progress. Cross validations are implemented to evaluate the performance of GCMCAP. The proposed method GCMCAP obtained an average AUC value of 0.894 and 0.896. In comparison with several other methods, GCMCAP outperforms other baselines in terms of most indictors. The result shows that associations between disease and miRNA can be more accurately identified by GCMCAP. The process of parameter analysis shows that our model is less affected by parameters and has stable prediction ability. Case studies show the ability of discovering new disease associated miRNAs. For future work, we may explore obtaining miRNA and disease node information from other data sources for more prior information. Additionally, graph convolution networks with attention mechanism may further improve the predictive ability for disease-related miRNAs.

Acknowledgements. This work has been supported by the National Natural Science Foundation of China (Grant no. 61873089).

References

1. Ambros, V.: microRNAs: tiny regulators with great potential. Cell **107**(7), 823–826 (2001)
2. Chen, C.-Z., Li, L., Lodish, H.F., Bartel, D.P.: MicroRNAs modulate hematopoietic lineage differentiation. Science **303**(5654), 83–86 (2004)
3. Ambros, V.J.C.: MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell **113**(6), 673–676 (2003)
4. Taganov, K.D., Boldin, M.P., Chang, K.-J., Baltimore, D.J.P.: NF-κB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. Proc. Natl Acad. Sci. **103**(33), 12481–12486 (2006)
5. Petrocca, F., et al.: E2F1-regulated microRNAs impair TGFβ-dependent cell-cycle arrest and apoptosis in gastric cancer. Cancer cell **13**(3), 272–286 (2008)
6. Shi, B., Sepp-Lorenzino, L., Prisco, M., Linsley, P., DeAngelis, T., Baserga, R.: Micro RNA 145 targets the insulin receptor substrate-1 and inhibits the growth of colon cancer cells. J. Biol. Chem. **282**(45), 32582–32590 (2007)
7. Zare, M., Bastami, M., Solali, S., Alivand, M.R.: Aberrant miRNA promoter methylation and EMT-involving miRNAs in breast cancer metastasis: diagnosis and therapeutic implications. J. Cell. Physiol. **233**(5), 3729–3744 (2018)
8. Li, Y., et al.: HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. **42**(D1), D1070–D1074 (2013)
9. Luo, J., Xiao, Q.: A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. J. Biomed. Inform. **66**, 194–203 (2017)
10. Zou, Q., Li, J., Song, L., Zeng, X., Wang, G.: Similarity computation strategies in the microRNA-disease network: a survey. Briefings Funct. Genomics **15**(1), 55–64 (2015)
11. Chen, X.J.M.B.: miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. Mol. BioSyst. **12**(2), 624–633 (2016)
12. Zeng, X., Zhang, X., Zou, Q.: Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Briefings Bioinform. **17**(2), 193–203 (2015)
13. Ding, P., Luo, J., Xiao, Q., Chen, X.: A path-based measurement for human miRNA functional similarities using miRNA-disease associations. Sci. Rep. **6**, 32533 (2016)
14. Zhao, X.-M., et al.: Identifying cancer-related microRNAs based on gene expression data. Bioinformatics **31**(8), 1226–1234 (2014)
15. Chen, X., Liu, M.-X., Yan, G.-Y.: RWRMDA: predicting novel human microRNA–disease associations. Mol. BioSyst. **8**(10), 2792–2798 (2012)
16. Xuan, P., et al.: Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS ONE **8**(8), e70204 (2013)
17. You, Z.-H., et al.: PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. **13**(3), e1005455 (2017)
18. Zou, Q., et al.: Prediction of microRNA-disease associations based on social network analysis methods. BioMed Res. Int. **2015** (2015)
19. Xiao, Q., Luo, J., Liang, C., Cai, J., Ding, P.: A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. Bioinformatics **34**(2), 239–248 (2018)

20. Luo, J., Shen, C., Lai, Z., Cai, J., Ding, P.: Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization. IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (2020)
21. Luo, J., Ding, P., Liang, C., Cao, B., Chen, X.: Collective prediction of disease-associated miRNAs based on transduction learning. IEEE/ACM Trans. Comput. Biol. Bioinf. **14**(6), 1468–1475 (2016)
22. Chen, X., Huang, L.: LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. PLoS Comput. Biol. **13**(12), e1005912 (2017)
23. Li, J.-Q., Rong, Z.-H., Chen, X., Yan, G.-Y., You, Z.-H.: MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget **8**(13), 21187 (2017)
24. Hou, S.: Neural Inductive Matrix Completion for Predicting Disease-Gene Associations (2018)
25. Xuan, P., Sun, H., Wang, X., Zhang, T., Pan, S.: Inferring the disease-associated miRNAs based on network representation learning and convolutional neural networks. Int. J. Mol. Sci. **20**(15), 3648 (2019)
26. Han, P., et al.: GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 705–713 (2019)
27. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T.: miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res. **37**(Suppl. 1), D105–D110 (2008)
28. Vergoulis, T., et al.: TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res. **40**(D1), D222–D229 (2011)
29. Chou, C.-H., et al.: miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. **44**(D1), D239–D247 (2016)
30. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., Marcotte, E.M.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. **21**(7), 1109–1121 (2011)
31. Xuan, P., et al.: Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics **31**(11), 1805–1815 (2015)
32. Wang, D., Wang, J., Lu, M., Song, F., Cui, Q.: Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics **26**(13), 1644–1650 (2010)
33. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.-F.: A new method to measure the semantic similarity of GO terms. Bioinformatics **23**(10), 1274–1281 (2007)
34. Berg, R., Kipf, T.N., Welling, M.: Graph convolutional matrix completion, arXiv preprint arXiv:1706.02263 (2017)
35. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics **34**(13), i457–i466 (2018)
36. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016)
37. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
38. Natarajan, N., Dhillon, I.S.: Inductive matrix completion for predicting gene–disease associations. Bioinformatics **30**(12), i60–i68 (2014)
39. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014)
40. Gui, H., Liu, J., Tao, F., Jiang, M., Norick, B., Han, J.: Large-scale embedding learning in heterogeneous event data. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 907–912. IEEE (2016)