Medicine and Disease Association Prediction via Attention-Based Medical Heterogeneous Information Network Representation Learning

Bin Zhang, Dan Yang, Zhihuang Lin

Abstract—**Knowledge of medication and disease has been rapidly accumulated. Also, an increasing number of researchers have paid more attention to predicting medicine-disease associations by machine learning methods. The associations of entities in the medical heterogeneous information network involve different association types. In this paper, we propose MHIN-MD, a representation learning of medical heterogeneous network model to solve the association between medicines and diseases in medical heterogeneous information network. Specifically, we first construct a medical heterogeneous information network contains diseases, medicines, diagnosis, etc. Next, we design a neural network model to learn the feature vector representation of medical nodes information. We try to find the association between medicines and diseases through the characteristics of medicines structure and the interaction between medicines. Using the Euclidean distance to calculate medicine similarity and disease similarity, generates a set of similar nodes between medicines and diseases to prediction association between medicines and diseases. Finally, our extensive experiments on MIMIC-Ⅲ datasets and DrugBank datasets, the result demonstrates that MHIN-MD can outperform baselines in the association prediction.**

Index Terms—**Medicine and Disease Association; Medical Heterogeneous Information Network; Network Representation Learning; Attention Mechanism.**

I. Introduction

Disease information is an important resource for biomedical researchers to deeply mine and analyze EHRs data. In the medical field, through constructing a complete medical heterogeneous information network, doctors can use the same medicine for patients based on similar symptoms of the disease, which will help to improve the utilization rate of medicines (e.g., Urokinase can treat pulmonary embolism, thromboembolism, and can also improve the anticoagulant activity of acetylsalicylic acid). In the medical field, it is very challenging to solve the association between medicines and diseases. For example, researchers have not yet been found for some of the association between medicines and diseases, make a prediction based on known interactions and paying attention to those possible associations, labor costs can be greatly reduced. Association prediction algorithms involved will play an important role. Since there are always a large number of potential associations between medicines and diseases, the associations [1] and predict whether the association between the two nodes exists. Using network representation learning [2-3] to represent a large number of medical nodes and the association between nodes in the medical heterogeneous information network to vectors in a low-dimensional space, and serve as medicines and diseases associate the sample characteristics of the prediction task. *Li et al.* [4] proposed a thyroid disease knowledge discovery and diagnosis framework AR-ANN, which integrates association rule mining and artificial neural network. At present, most methods of obtaining the association between medicines and diseases by learning the effective vector representation of medical concepts still have many challenges:

(1) **Medicine-Medicine Association Complexity.** There are synergistic effects among various medicines, and even antagonistic effects among some medicines. Therefore, in the prediction results, it is necessary to consider whether the antagonistic medicines are effective in the treatment of diseases. In order to prevent the association between medicines from affecting the results of clinical trials, special attention will be paid to the association between medicines.

(2) **Medicine-Disease Data Sparsity.** The lack of clinical symptom information may lead to the lack of association between medicines. In the process of training data, other data sets are combined to improve medical data, and useful information is extracted from different medical data sets, so as to expand the association between medicines and diseases and better predict the association between medicines and diseases.

(3) **Disease Complexity**. Predicting the association between medicines and diseases involves many aspects such as molecular factors (genes, proteins, metabolism, *etc.*) and environmental factors, phenotype of diseases, therapeutic medicines, *etc.* Although there are differences between diseases, they still have similar characteristics in essence. Considering the complex factors of diseases, the similarity calculation of diseases will be carried out here to find the association between medicines and diseases in similar diseases.

In view of the above challenges, in order to better predicted association between medicines and diseases, we propose MHIN-MD, a representation learning of medical heterogeneous network model, to find the association between medicines and diseases in Electronic Health

Manuscript received May 14, 2021; revised October 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 6167214, in part by General Scientific Research Projects of Liaoning Province under Grant 2019LNJC07, and in part by University of Science and Technology Liaoning Talent Project under Grant 601011507-22

Bin Zhang is a Master of Computer Science and Software Engineering Department, University of Science and Technology Liao Ning, Anshan, China (e-mail: zxinstudios@163. com).

Dan Yang, the corresponding author, is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liao Ning, Anshan, China (e-mail: asyangdan@ 163. com).

Zhihuang Lin is a Master of Computer Science and Software Engineering Department, University of Science and Technology Liao Ning, Anshan, China (e-mail: lzh_ustl@163.com)

Records (EHRs). Specifically, we first extract a large number of medical heterogeneous information (diseases, medicines, diagnosis, *etc.*) from EHRs, and look for the association between medicines and diseases through the characteristics of medicines structure and the interaction between medicines. Next, we design a neural network model to learn the feature vector representation of medical nodes. Finally, we use Euclidean Distance to calculate medicine similarity and disease similarity, generating a set of similar nodes between medicines and diseases to predict the associations between medicines and diseases. To summarize, the main contributions of our work are:

- We formalize the learning problem of heterogeneous information network representation learning involving medical node heterogeneity.
- We propose an innovative heterogeneous information network model, namely MHIN-MD, to predict the associations between medicines and diseases.
- We conduct extensive experiments on MIMIC-III and DrugBank datasets, and the results show that MHIN-MD has superior performance than many baselines.

II. Related Work

In recent years, heterogeneous network representation learning has been studied more and more in the field of medical. This section introduces in detail the heterogeneous network representation learning method that maps various types of medical entities into low-dimensional vectors. Then, the feature vectors of medicines and diseases are fused, and the fused vectors are labeled for semi-supervised dichotomy learning. Finally, the training results are used to predict the association between medicines and diseases, so as to deeply explore the potential performance of medicines.

A. Heterogeneous Information Network Representation Learning

Heterogeneous networks refer to networks with different types of nodes and edges. Great progress has been made in Spoken Language Processing, Natural Language Processing (NLP) and Image Target Recognition. Many researchers utilize representation learning in the medical healthcare domain, for the reason that the sequence of medical codes can be seen as a natural language text. In recent years, many researchers have applied representation learning in the medical healthcare domain because an effective feature representation can simplify the difficulty of dealing with a problem and provide convenience for further applications. *Bengio et al.* [5] first modeled by using neural network language model [6] under the background of statistical language modeling. Med2vec [7] aimed to learn the representation of medical codes to predict future patients access information. However, the method ignored the long-term dependence of doctor's orders during the visit. GRAM [8] was a graph-based representation learning research model for medical health. It used RNN model and accessed patients' medical records according to the robustness of representation learning for medical entities. Both RETAIN and GRAM adopted prediction mechanism to improve prediction performance. *Wang et al.* [9] defined a new consensus clustering method, which uses Euclidean distance and classification distance to automatically cluster numerical data and classify data respectively. *David et al.* [10] proposed an algorithm for anomaly detection and representation based on Euclidean distance between medical laboratory data. *Jiang et al.* [11] proposed a time-aware patient similarity framework from EHRs, in the process of network representation learning, the time decay function is combined with the medical entity representations to obtain the temporal patients' representations. The research level in the homogeneous field has reached maturity. With the construction of knowledge graph in the era of deep learning, the representation learning of heterogeneous networks has been gradually applied. For example, in social networks, attention mechanism is used to separate the association between users and users, users and items and items and item attributes from a heterogeneous information network. In the protein network, amino acid molecules from the association of the network through the structure or distance proximity, characterizing the characteristics of different proteins. In different clinical trials, the kinds of medical entities associated with medicines are different, and the results are different. These medical heterogeneous entities make the medical information network larger and more complex, forming a medical heterogeneous information network containing various types of information. By learning heterogeneous information network, we can deeply explore their potential association and increase the reusability of medicines in the medical field, so that medicines information can be fully applied to the treatment of medical diseases. At present, many fields apply the combination of machine learning and graph embedding technology to predict. Network nodes into are mapped low-dimensional vector representation. The prediction results are more accurate.

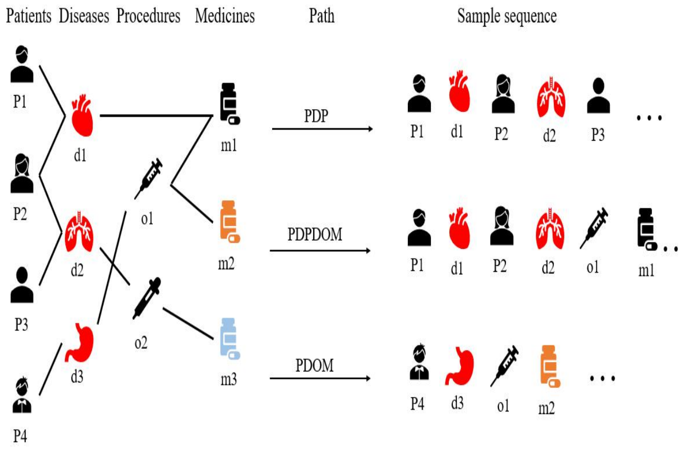


Fig. 1 An example of meta-path-based random walk in a medical heterogeneous information network

B. Association Prediction

So far, many researchers have proposed various association prediction methods. Among them, *Guo Fu et al.* [12] under the background of Medicine mining, Drug Target Interactions (DTIs) can be predicted according to the observed topological features of semantic networks across chemical and biological space. *Lee et al.* [13] uses a cosine-based similarity measure to determine which patient is most consistent with which patient. *Wu et al.* [14] considered the association between the clustering coefficients and similarity of common neighbor nodes, defined the similarity of a node pair to be measured as the

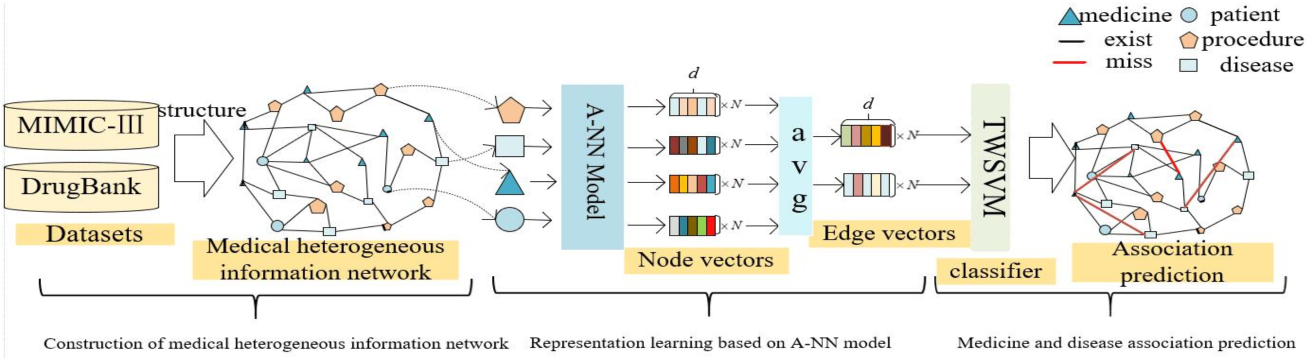

medicine
patient
exist
procedure
miss
disease
MIMIC-III
DrugBank
structure
Datasets
Medical heterogeneous information network
A-NN Model
d
×N
a v g
Node vectors
Edge vectors
TWSVM
classifier
Association prediction
Construction of medical heterogeneous information network
Representation learning based on A-NN model
Medicine and disease association prediction

Fig. 2. Association prediction of medicine and disease base on *MHIN-MD* representation learning

sum of the clustering coefficients of all its common neighbors, and believed that the closeness of the association between the common neighbors of the node pair would affect their similarity. *Liu et al.* [15] proposed a similarity algorithm based on local walk to the network. This method holds that when nodes perform random walk. They do not need to consider the walk within the whole network range, which limit their walk within the limited number of steps. A large number of experiments proved that the operational efficiency of this method was much higher than that of the global random walk algorithm. It applies to large-scale networks.

III. Prepare Knowledge and Problem Definition

Definition 1. **Heterogeneous information network**. A HIN is denoted as *G= (V, E)* consisting of an object set *V* and a link set *E*. A HIN is also associated with an object type mapping function φ: $V \rightarrow A$ and a link type mapping function ψ: $E \rightarrow R$. *A* and *R* denote the sets of predefined object and link types, where $|A|+|R|>2$.

Definition 2. **Medical heterogeneous information network**. A medical heterogeneous information network is denoted as $G=(V_m, V_n, E_o)$ consisting of a medical object set V_m and a medicines object set V_n and a link set E_0. A medical heterogeneous information network is also associated with an object type mapping function φ: $V \rightarrow A_i$,(i ∈ 4) and a link type mapping function ψ: $E \rightarrow R_s$,(s∈5).

Definition 3. **Meta-path.** A meta-path *P* is defined on a medical heterogeneous information network $P=(A_i, R_s)$ and is denoted as a path in the form of $A_1 \xrightarrow{R1} A_2 \xrightarrow{R2} \dots \xrightarrow{Ri} A_{l+1}$ (abbreviated as $A_1 A_2 \dots A_{l+1}$), which describes a composite relation $R= R_1 R_2 \dots R_l$ between object A_1 and A_{l+1}, where denotes the composition operator on relations.

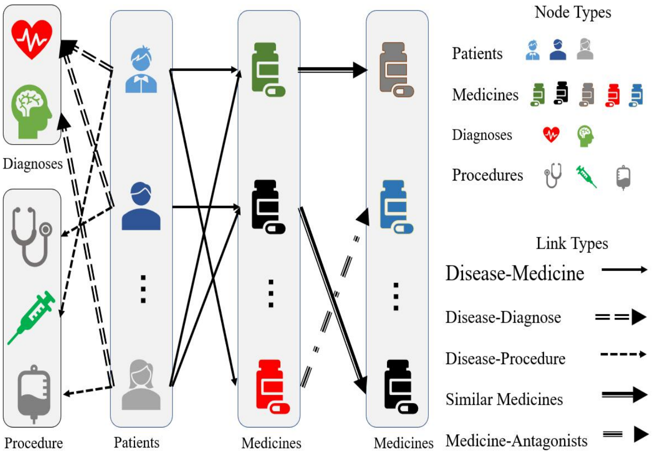


Fig. 3. Medical Heterogeneous Information Network

Example. Two objects can be linked via multiple meta-paths, *e.g.,* "Patient - Patient" (PP) and "Patient - Disease - Patient" (PDP). Different meta-paths usually convey different semantics. We take Fig. 1 as an example, which represents the heterogeneous information network of medical. The PP path indicates similarity between two patients, while PDP path indicates the co-suffer relation between two patients, *i.e.,* they have suffered from the same diseases. It is intuitive to find that these meta-paths can lead to meaningful node sequences corresponding to different semantic relations.

IV. The Proposed Framework

In this section, we proposed a representation learning framework of medical heterogeneous information network based on the attention mechanism. The framework mainly includes three parts as shown in Fig. 2: 1) construction of medical heterogeneous information network; 2) representation learning of medical heterogeneous information network; 3) the association prediction between medicines and diseases.

A. Construction medical Heterogeneous Information Network

We extract the medical entity nodes in Electronic Health Records (EHRs) and the associations between these nodes, a medical heterogeneous information network with node types and including the associations between nodes are constructed. The constructed information network is a directed heterogeneous graph, in which the types of nodes and the associations between nodes. As shown in Fig. 3, there are links between various diseases and various medicines, and there are many different types between medicines, constructing patient-disease, patient-medicine, patient-procedure, medicine-similar medicine, medicine-antagonist.

B. Representation Learning of Medical Heterogeneous Information Network Based on Attention Mechanism

Representation learning of medical heterogeneous information network based on the attention mechanism is to add weight to different meta-paths in medical heterogeneous information network, so as to obtain new vector representation. We proposed a new representation learning model A-NN of medical heterogeneous

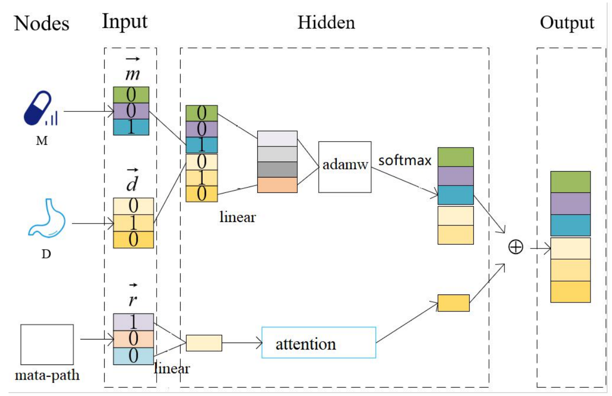


Fig. 4. Medical Node Representation Learning Based on *A-NN*

information networks based on attention mechanism, which generates new feature vectors by learning node vectors and different meta-paths.

1) The A-NN model

Our propose an A-NN model as shown in Fig. 4, the medical node and its meta-path are represented as a one-hot vector at the input layer, and the hidden vector of the medical node and its association is obtained through the transformation of the hidden layer. Because there are different semantics in medical nodes, it is necessary to regularize them so that the value of relation vector is between 0 and 1. This has two purposes: 1) The association between medical nodes takes negative values, which affects the learning of vectors, because when all values are guaranteed to be within the range of positive values, the output layer will always remain positive when passing through the Adam function. 2) If the value of the vector is too large, the effectiveness of learning will be reduced. Finally, the model uses the Hadamard function for the three implicit vectors, which are defined as the formula (1):

$$H = W_m \vec{m} \Theta W_d \vec{d} \Theta f_{01}(W_r \vec{r}) \qquad (1)$$

Then apply a *sigmoid* activation function to the result of Equation (2) as follows:

$$f(x) = \frac{1}{1+e^{-H}} \qquad (2)$$

At the same time, attention mechanism modules are added to different meta-paths in the hidden layer to make the medicine and disease nodes generate different weight, and the newly generated vectors are weighted and summed with the medical node vectors. Output medical entity nodes and their associations are assigned different weights:

$$V_i = tanh\,(w_1 r_i + w_2 b) \qquad (3)$$

Where w_1 and w_2 are the weight matrices in the full association layer, and b is the bias vector. Secondly, weight score of different regions is calculated through Softmax function. The larger the score, the closer the association between the nodes. The function calculation is as following:

$$\alpha = soft\,max\,(uv^T) \qquad (4)$$

The function C_i obtained by vector product calculation of the output obtained by the hidden layer and the weight score is shown in Formula:

$$c_i = \sum_j a_{ij} \cdot h_j \qquad (5)$$

Subject the set C_i obtained by summing to nonlinear transformation by an *Adam* optimizer to obtain a new node vector.

2) Optimization Objective.

In the A-NN model, in order to better learn the vector representation of medical entity nodes, a plurality of objective functions will be set for each meta-path in the association set R. which makes it convenient to iterate and better train data. A training set D is extracted from the medical heterogeneous information network, which includes a set R of medical heterogeneous information network, which includes a set R of medical entity drugs diseases and their associations, and binary tags representing the existence of edges between medicines and diseases. The model A-NN is trained in the back propagation training algorithm combined with random gradient descent, during each training, the weight of the input medical entity node target will be adjusted, and then the medical entity nodes in the training set will be multiplied correspondingly, so as to ensure the maximization of the objective function, facilitate the calculation in the optimization process, prevent the data from being too large, and use log instead of direct calculation. Thus, the formulated as:

$$log\,O = \sum_{M,D,R\in D} log\,O_{M,D,R}\,(M,D,R) \qquad (6)$$

In order to obtain correct prediction, the correct probability $P(R|M,D)$ of prediction training in the training set will be set to the maximum when $L(M,D,R)$ is 1, to the minimum when $L(M,D,R)$ is 0. The association is as follows：

$$O_{M,D,R}(M,D,R) = \begin{cases} P(R|M,D), & if\,L(M,D,R)=1 \\ 1-P(R|M,D), & if\,L(M,D,R)=0 \end{cases} \qquad (7)$$

Different probabilities are assigned to medical entity nodes according to the existence of edges, so as to highlight the association between two medical entity nodes.

$$log\,O_{M,D,R}\,(M,D,R) = L(M,D,R)\,log\,P\,(R|M,D) \\ +[1-L(M,D,R)]\,log\,[1-P(R|M,D)] \qquad (8)$$

In order to prevent the medical entity node after multiplication form being too large to be calculated, the log form will be adopted to avoid errors in calculation results caused by a complicated operation.

$$P(R|M,D) = sigmoid(\sum W_M'\vec{M} \odot W_D'\vec{D} \odot W_R'\vec{R}) \qquad (9)$$

WhereW_M',W_D',W_R' respectively represent the weight of nodes, then the objective function is maximized by applying random gradient descent algorithm, and $\log O_{M,D,R}(M,D,R)$ and weight are added each time the data are trained.

$$W_M'\vec{M}: = W_M'\vec{M} + \frac{d\,log\,O_{M,D,R}\,(M,D,R)}{dW_M'\vec{M}} \qquad (10)$$

$$W_D'\vec{D}: = W_D'\vec{D} + \frac{d\,log\,O_{M,D,R}\,(M,D,R)}{dW_D'\vec{D}} \qquad (11)$$

$$W_R'\vec{R}: = W_R'\vec{R} + \frac{d\,log\,O_{M,D,R}\,(M,D,R)}{dW_R'\vec{R}} \qquad (12)$$

Where respectively represent the weight of medicine nodes M, disease nodes D and their associations R corresponding to medical entities. Each medical entity is assigned different weights for calculation.

3) Algorithm description

Algorithm 1 describes a representation learning algorithm based on A-NN model, input medicine and disease nodes, an empty list is used to store the generated meta-path. The meta-path is generated by random walk, and a label is added to the meta-path of the generated medicines and disease nodes. If it exists, it is set to 0, and if it does not exist, it is set to 1. Then it is put into A-NN model for training to

Algorithm 1 Medical Representation Learning Algorithm Based on A-NN Model

```
Input: Medicine node X={x1， . . . ， xi}； Disease node Y={y1， . . . ， yj}；
    n：Number of nodes；L：label；
Output:  e: node vector set
paths = []
for i=1, . . ., n do
    for j=1, . . ., n do
        paths ← Medicine node X and medicine node Y were selected by a
         random walk
        if paths=null then
            L=1
        else
            L=0
Paths were input into A-NN model for training to generate feature vectors of
    each node.
The medicine node vector x disease node vector y and L are combined into a
    triple<x, y, L>.
Classifier training is used to predict medicine and disease associations.
Get the predicted score AUC.
Return AUC.
```

obtain the feature vector of the node. Finally, it is put into the classifier in the form of triple for prediction. The accuracy of the prediction result is judged according to the value.

C. Prediction of Association between Medicines and Diseases

In this section, by representing each medical node vector obtained after learning. According to the association between nodes, the vector is weighted and summed to obtain the edge vector representing the association between two medical entity nodes. The edge vector and the label representing the association between them are input into binary classifier for semi-supervised learning to predict the association between medicines and diseases, to reduce the prediction time, TWSVM will be used for training in this article, its training time is one quarter of that of SVM, moreover, it can better solve the *exclusive OR* problem of binary classification. In addition, it is very effective for the classification of medical entities with a large number of samples. TWSVM refines the binary classification problem and divides a problem into two smaller quadratic programming problems. By using two non-parallel planes, the data points of the corresponding classes are clustered around the plane. Among them, there are m_1 positive classes and m_2 negative classes, which are represented by matrices $A(m_1\times n)$ and $B(m_2\times n)$ respectively:

$$(TWSVM1)\ \underset{w^{(1)},b^{(1)},q}{Min}\ \frac{1}{2}(Aw^{(1)}+e_1b^{(1)})^T(Aw^{(1)}+e_1b^{(1)}) + c_1e_2^Tq$$

$$S.\,t.\quad -(Bw^{(1)}+e_2b^{(1)})+q\geq e_2, q\geq 0 \tag{13}$$

Table I

Association Types between Medical Entity Nodes

Node1	Node2	Type
P1	40301	Diagnosed
P1218	D5W	Prescribed
P12719	640	Operated
. . .	. . .	. . .
7813	781.3	SameAs
DB03619	DB00605	Interact
158	158.9	Subclass
DB00560	357.2	Indication

$$(TWSVM2)\ \underset{w^{(2)},b^{(2)},q}{Min}\ \frac{1}{2}(Bw^{(2)}+e_2b^{(2)})^T(Bw^{(2)}+e_2b^{(2)}) + c_2e_1^Tq$$

$$S.\,t.\ (Aw^{(2)}+e_1b^{(2)})+q\geq e_1, q\geq 0 \tag{14}$$

Where C_1 and C_2 are two penalty parameters, e_1 is a *m_1-dimensional* unit column vector, and e_2 is a *m_2-dimensional* unit column vector. q is the relaxation vector. TWSVM1 is optimized by Lagrange pair, and the partial derivative solved by the sum is zero. Similarly, TWSVM2 is optimized and the sum is generated. Two hyperplanes are represented as:

$$K(x,C^T)u_1+b_1=0 \tag{15}$$

$$K(x,C^T)u_2+b_2=0 \tag{16}$$

The samples will be classified into categories closer to the hyperplane, which is shown in Equation (17):

$$K(x^T,\text{C}^T)w_r+b_r=\underset{i=1,2}{min}|K(x^T,C^T)w_l+b_l| \tag{17}$$

Where X belongs to class $r(r\in\{1,2\}$), after classification and prediction, by randomly selecting 2000 nodes to form a training set, in order to reduce the number of node pairs to be sorted. Firstly, after a node is selected, a specified range of neighbor nodes around if are taken as candidate nodes. When selecting a certain range of neighbors, appropriate selection will be made according to the size of the data set, so as to ensure that the edges that can be selected can be guaranteed within the range of the data set.

V. Experiments and Evaluation

In this section, we will first introduce the introduction and Pretreatment of medical datasets including disease and medicine datasets, secondly introduce the evaluation indexes applied in this paper and the comparison methods of experiments, and set the parameters in the experiments, and finally explain the proposed model through a case.

A. Experimental Datasets

We use two different datasets MHIN-MD：MIMIC-III and DrugBank. For the MIMIC-III, we extract patients, diseases, programs, *etc.* For the DrugBank, we extract medicines information and its association. Finally, a heterogeneous information network including medicines, diseases and patients is built by combining disease information with medicine information.

1) MIMIC-III datasets

MIMIC-III[16] is a free and open datasets of intensive care patients, which contain the data of ICU patients in Beth

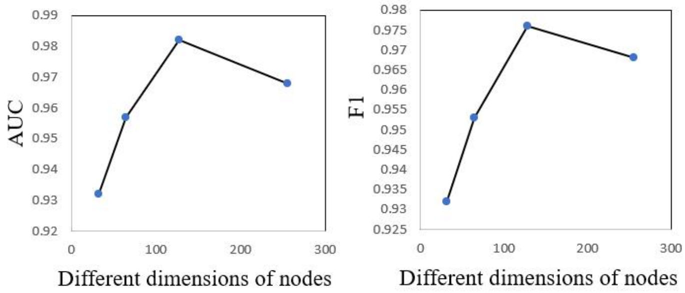


Fig. 5 MHIN-MD sets the performance impact of different node dimensions

Israel Deaconess Medical Center from 2001 to 2012, totaling 46,520 patients. MIMIC-III includes clinical datasets and physiological waveform datasets[17], including patient demographic characteristics, discharge records, clinical diagnostic information and drug prescription information, *etc.* The database records the electrocardiogram, microbiological events and other physiological parameters. In this paper, the datasets are operated as follows: (1) Extract the information of patients' diseases diagnosis, disease characteristics, program operation and medicines used; (2) constructing the association between each medicine and the disease according to the patient information; (3) Add tags to mark the association between medicines and diseases through random walks.

2) DrugBank datasets

DrugBank datasets covers detailed medicine data (*e.g.,* chemical data, pharmacological data) and comprehensive medicine target information (*e.g.,* sequence, structure and pathway of action). The datasets contain a total of 7,685 medicine entities, including more than 6,000 experimental medicines. Using the different characteristics between medicines can discover potential association between medicines and diseases. The medicine similarity network based on chemical structure, the medicine similarity network based on ATC code, By combining with MIMIC-III to form data in triple from as shown in Table I, Where in each row of the list represents the association between medical entity nodes(*e.g.*, patient ID, disease code, program sequence, drug code) and the type of association between nodes(*e.g.* diagnosis, program, prescription, *etc.*), the type between medicines(*e.g.,* similarity, interaction, *etc.*), the type between diseases(*e.g.,* subset), and the type between medicines and diseases(*e.g.,* interaction).

B. Evaluation Metrics and Comparison Methods

In order to evaluate the performance of our proposed framework, we adopt several commonly used evaluation metrics and introduce several comparison methods.

1) Evaluation Metrics

In the verification of association prediction, we consider four representative similarity indexes, among which Common Neighbors (CN) index and Preference Attachment (PA) index are association prediction methods based on node neighbors, while Local Path LP index and Katz index are association prediction methods based on paths. AUC and ROC curves can measure the advantages and disadvantages of classifiers as a whole and judge the accuracy of the proposed experiments.

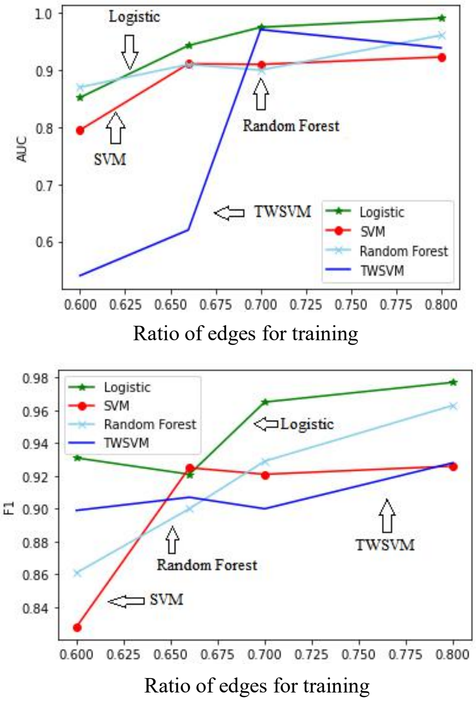


Fig.6 influence of different data sets proportion on accuracy

- **CN** is a commonly used basic graph analysis algorithm, which can get the neighbor nodes shared by the two nodes. More similar neighbor nodes indicate a close association between the two nodes. The similarity is defined as:

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \tag{18}$$

Where S_{xy} is represents the similarity between the two nodes. In fact, the common neighbor of two nodes is the number of two paths between them.

PA is used to generate evolutionary scale-free networks, the index does not require the neighborhood information of each node, so it has the least computational complexity:

$$S_{xy}^{PA} = K_x \times K_y \tag{19}$$

Where the probability of a new link connecting to node x and node y are proportional to $k_x \times k_y$.

LP is a similarity index based on CN, which considers the influence of the third-order path, the influence of direct neighbors and indirect neighbors.

$$S^{LP} = A^2 + \beta A^3 \tag{20}$$

Where β is an adjustable parameter to control the influence of the third-order path; A represents the adjacency matrix of the network.

Katz extends the LP index and further extends the case of all long paths in the network, that is, the weighted sum of paths between two vertices considering sequential paths.

$$\begin{aligned} S^{Katz} &= \sum_{l=1}^{\infty} \beta^l A^l = \beta A \left(I + \sum_{l=1}^{\infty} \beta^l A' \right) \\ &= \beta A(I + S^{Katz}) \\ &= (I - \beta A)^{-1} \beta A \end{aligned} \tag{21}$$

Where β is the weight attenuation factor, in order to ensure the convergence of the sequence, the value must be less than the reciprocal of the maximum eigenvalue of the adjacency matrix A.

ROC is a set of predicted curves, the abscissa is false positive rate (FPR), and the ordinate is the true positive rate (TPR). The closer the ROC curve is to the upper left corner, the better the performance of the classifier.

Table II

Accuracy of different methods

Method	AUC	F1
PA	0.95±0.003	0.93±0.006
Katz	0.96±0.007	0.95±0.005
LP	0.96±0.006	0.97±0.003
CN	0.59±0.002	0.57±0.001
LINE	0.85±0.003	0.90±0.006
DeepWalk	0.87±0.007	0.88±0.005
Node2vec	0.93±0.006	0.91±0.003
MHIN-MD *（Logistic）*	0.99±0.001	0.97±0.007
MHIN-MD *（SVM）*	0.94±0.002	0.92±0.006
MHIN-MD *（Random Forest）*	0.90±0.009	0.91±0.005
MHIN-MD ***（TWSVM）***	**0.97±0.006**	**0.96±0.003**

- **AUC** is the area under the ROC curve. AUC evaluates that a positive sample and a negative sample are randomly given, and the prediction probability of the model for the positive sample is greater than the prediction probability of the model for the negative sample. The calculation formula of association prediction accuracy AUC is as following:

$$AUC = \frac{n'=0.5n''}{n} \tag{22}$$

2) Comparison Methods

Considering the running efficiency on large-scale data sets and the feasibility and expansibility of the algorithm, in this paper, we use three probability-based methods and three different classifiers for performance comparison.

- **DeepWalk** [18] it is learning d-dimensional vectors by acquiring node pairs in the w-hop neighborhood of the network through uniform random walk in the network.

Table III

Accurate Values for Different Parameters of *TWSVM*

Training (%)	C1=0.1	C1=0.2	**C1=0.5**	C1=1	C1=1
	C2=0.05	C2=0.1	**C2=0.25**	C2=0.5	C2=1
66	0.59	0.43	**0.62**	0.601	0.55
70	**0.977**	**0.970**	**0.971**	**0.964**	**0.963**
80	0.928	0.9285	**0.939**	0.945	0.911

- **LINE** [19] learns node vectors by considering the first-order nearest neighbor and the second-order nearest neighbor of nodes in the network respectively. LINE is used to learn *d/2* dimensions by capturing first-order information. Other *d/2* dimensions are learned by capturing second-order information, and then they are used to learn d-dimensional node vectors.
- **Node2vec** [20] it is generalized by DeepWalk. It learns D -dimensional node vectors by capturing node pairs in W -hop neighborhood through parameterized random walks.

C. Parameter determination in MHIN-MD

In the MHIN-MD model of medical heterogeneous information network, in order to obtain better experimental performance, we set MHIN-MD model with four different vector dimensions: 32,64,128 and 256. Next, we use these vectors for training, the results are shown in Fig. 5. The accuracy rate is 0.935,0.954,0.976 and 0.968, respectively. It can be clearly seen that when the vector dimension is 128, the accuracy rate of the experimental accuracy reaches the highest, so in the experiment, we select 128 dimensions to achieve better results. We set the negative sampling rate to 5, and the initial learning rate of random gradient descent is 0.025. After numerous experiments, we found that setting the context window of DeepWalk and Node2vec to 4 can get better performance. Considering that different proportions of

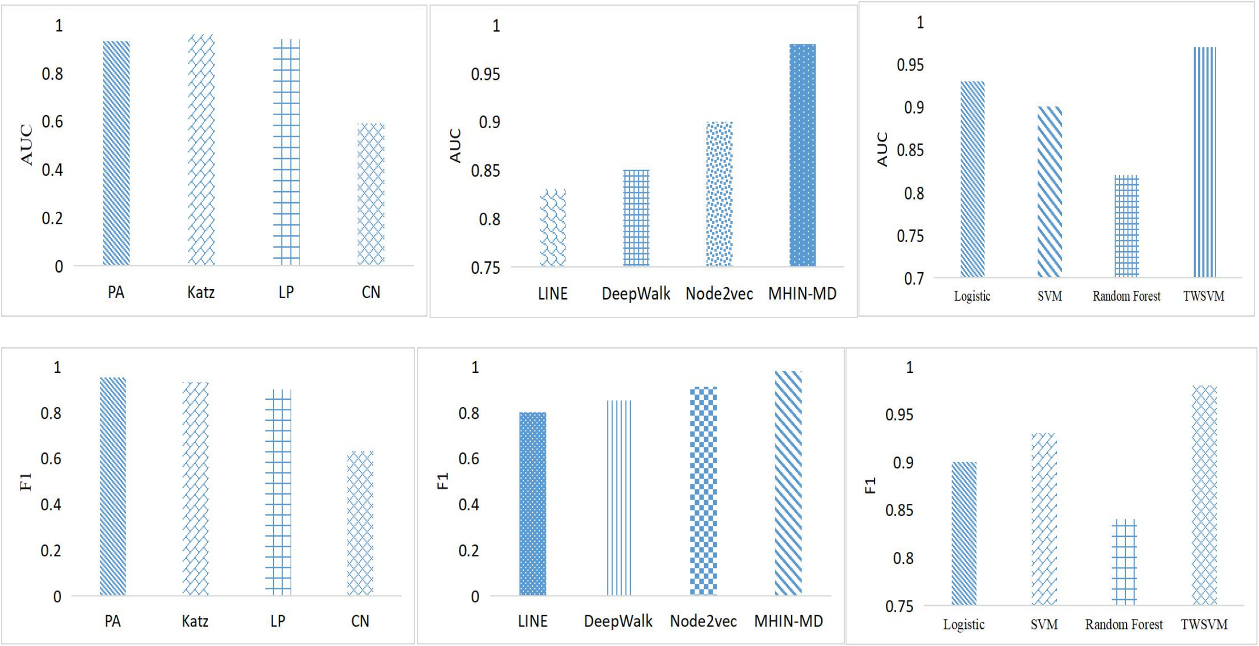


Fig. 7 Performances of variant proposed models

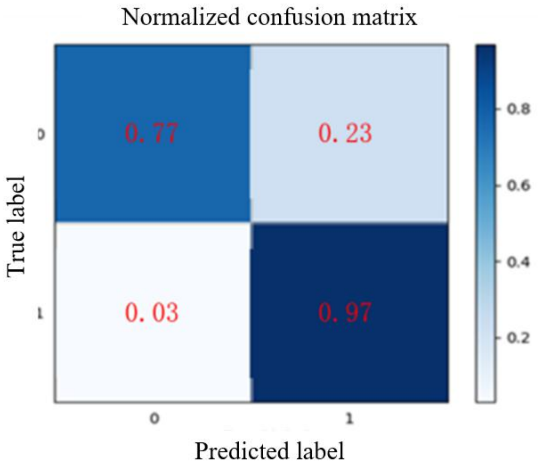


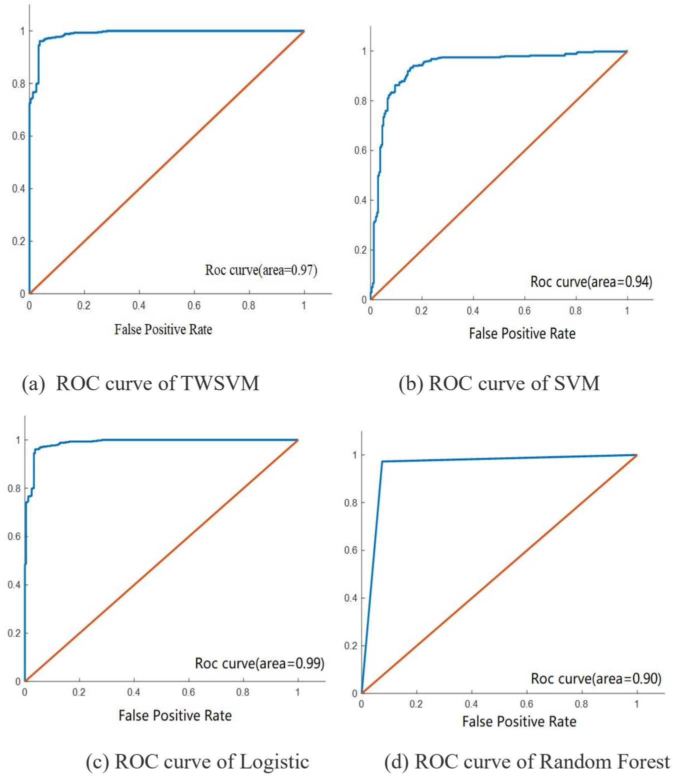
Fig. 8 confusion matrix of association prediction between medicines and diseases

data sets may have diverse effects, we select training sets and test sets with different proportions to carry out experiments, and find that the larger the proportion of training sets, the better the prediction effect. The ratio of Logistic training set is 0.8, which is 0.01~0.05 higher than other ratios. The ratio of SVM training set is 0.8, which is 0.05~0.1 higher than other ratios. The ratio of Random Forest training set is 0.8, which is 0.02~0.05 higher than other ratios. The ratio of TWSVM training set is 0.7, which is 0.05~0.6 higher than other ratios. It can be seen from this that TWSVM needs a lot of data for training, and the results will be better. The results are shown in Fig. 6. When the experimental data training set is set to 0.8, the overall effect is the best. We compared MHIN-MD with different similarity indexes, different isomorphic methods and different classifiers. As shown in table II, we finally found that MHIN-MD framework has obvious effects on logistic regression and TWSVM. Logistic is one of the most commonly used analytical methods to predict the association between medicines and diseases. TWSVM classifier is the latest popular classifier. The datasets are optimized by TWSVM classifier continuously in this paper. We set different parameter values for two necessary parametersC1 and C2 in TWSVM and study their training effects. In order to test the performance of TWSVM classifier more accurately, experiments are carried out on training sets with different proportions when the range of parametersC1 and C2 is selected. From table III, it can be seen that the training performance of 70% training set used in this paper is better than that of 80% training set. Therefore, in this experiment, the training set is set to 70% to learn MHIN-MD.

Table IV

Association between Medicine Testing Prescribed for Certain Diseases and Similar Diseases

Disease Code	Similar Disease Code	Medicine name
10017	10006, 10026,10027, 10033,10040	Magnesium Sulfate, Metronidazole, Senna
10074	10042, 10027, 10044	Aspirin, magnesium oxide
10083	42231,42346	Lisinopril, morphine sulfate, haloperidol
42066	10043	Potassiumchloride, propofol, phenylephrine, fentanyl patch



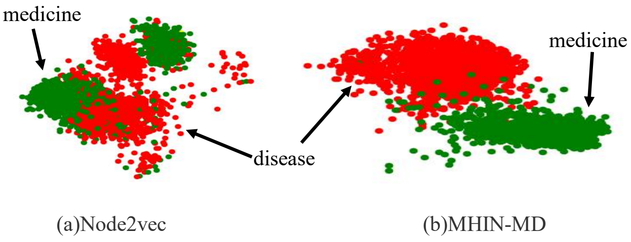
(a) ROC curve of TWSVM (b) ROC curve of SVM

(c) ROC curve of Logistic (d) ROC curve of Random Forest

Fig. 9 ROC curve of association prediction between medicines and diseases

D. Experimental results analysis

In this paper, we proposed model has greatly improved in predicting the association between medicines and diseases. By comparing different methods, the accuracy of MHIN-MD is increased by 2%-12%. As shown in Fig.7, in the performance of different comparison methods in the association between medicines and diseases, we proposed that the accuracy of MHIN-MD has been significantly improved. In order to view the details of the classifier's prediction more clearly, we verified the accuracy of the confusion matrix by experiments. The results are shown in Fig. 8. It is obvious that True Positive reaches 0.97, True Negative reaches 0.77, False Positive reaches0.03, and False Negative reaches 0.23. Experiments prove that the accuracy of our proposed framework MHIN-MD is still very high compared with other methods, and the accuracy of



(a)Node2vec (b)MHIN-MD

Fig.10 Visualization of medicines and diseases

prediction is also relatively high. The ROC curve of the experimental results is shown in Fig. 9, which shows that the framework proposed in this paper has better overall prediction ability in medicine-disease association prediction.

E. Visualization of medicines and diseases

One way of assessing the quality of medicines and diseases embedding representation is through visualization. We conduct visualization experiments to compare the performance of classification between medicines and diseases after Node2vec and MHIN-MD learning. Medicines and diseases embedding representation are fed as features into t-SNE[21],which mapped all points into a 2D space, where two different colors represent medicines and diseases vectors.

From Fig.10, we can see that the visualization of outputs from Node2vec shows I clear boundaries and diffuse clusters. MHIN-MD is much better. In the results we can see the points of different colors are still intermixed in the center of the visualization. However, MHIN-MD can distinguish the two types of nodes better. The MHIN-MD can achieve very good results.

F. Case study

In order to verify the accuracy of medicine-disease association prediction, the unproven records in the current medical records are selected as the test set for training, As shown in Table IV, that is, the diseases with similarity obtained by the similarity measurement of the training diseases are predicted, and the edges without association between the medicines and the similar diseases are predicted(for example, a patient suffers from a fracture, and a blood activating medicines prescribed by a doctor can also be applied to a patient suffering from pulmonary embolism. And the drugs used by patients with septicemia can also be used in pulmonary edema diseases). In order to verify whether the conclusion is reasonable, it will be verified on the CTD (Comparative Toxicogenomic Database) database. By searching the literature, it can be found that the predicted drugs are also related to related diseases, which prove that the method of this framework is reliable. At the same time, it also shows that the accuracy of association prediction in heterogeneous networks is more accurate than that in homogeneous networks.

VI. Conclusions and Future Work

Because MIMIC-III is a clinical diagnostic data set, the information contained in MIMIC-III is very complex. In this paper, a method for predicting the association between medicines and diseases in heterogeneous information networks is proposed. Through the embedding of heterogeneous information, the association between medicines and diseases can be predicted more accurately, which combines the current popular application technologies to mine more potential associations in the association and is more efficient in disease prediction. However, when processing data, only the association between medicines and diseases are considered, and the accuracy of predicting results will be reduced. There are many complexities in medical data, and there are numerous associations between drugs. We found that the associations between drugs should also be paid attention to, and the interactions among drugs sometimes affect the predicted consequences. In the future, we will embed attribute information such as characteristic attributes of patients, disease information, medicine characteristics, and association weight among medicines into the heterogeneous network of medical attributes. Different weights are distributed through different associations among characteristic attributes, which highlight the association between medicines and diseases. This will enrich the heterogeneous medical information network and fully explore the association between medicines and disease.

References

[1] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. USA 106 (2009) 22073.

[2] Peng C, Xiao W, Jian P, et al. A Survey on Network Embedding[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, PP (99):1-1.

[3] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey[J]. Knowledge Structured Systems, 2018, 151(JUL. 1):78-94.

[4] Dongyang Li, Dan Yang, Jing Zhang, and Xuedong Zhang, "AR-ANN: Incorporating Association Rule Mining in Artificial Neural Network for Thyroid Disease Knowledge Discovery and Diagnosis," IAENG International Journal of Computer Science, vol. 47, no.1, pp25-36, 2020

[5] Aditya Grove,Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In KDD '16. ACM，855-864.

[6] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. JMLR, 3, 1137 – 1155.

[7] Multi-layer Representation Learning for Medical Concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). 1495 – 1504.

[8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: Graph-based AŠention Model for Healthcare Representation Learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD17). ACM.

[9] Ding H, Wang C, Huang K, et al. iGPSe: A visual analytic system for integrative genomic based cancer patient stratification[J]. Bmc Bioinformatics, 2014, 15(1):203.

[10] David G, BernsteinL, CoifmanRR. Generating evidence-based interpretation of hematology screens via anomaly characterization. Open Clin Chem J. 2011;4(1):10-6.

[11] Hua Jiang, and Dan Yang, "Learning Graph-based Embedding from EHRs for Time-aware Patient Similarity," Engineering Letters, vol. 28, no.4, pp1254-1262, 2020

[12] Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration[J]. Information ences, 2017:546-560.

[13] Lee J, Maslove DM, Dubin JA.Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS ONE. 2015;10(5): e0127428.

[14] Wu Z, Lin Y, Wang J, et al. Link Prediction with Node Clustering Coefficient[J]. 2015.

[15] Liu W, Lü, Linyuan. association Prediction Based on Local Random Walk[J]. EPL (Europhysics Letters), 2010, 89(5):58007-. Johnson, A, Pollard, T, Shen, L.et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35.

[16] Zhihuang Lin, and Dan Yang, "Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity," Engineering Letters, vol. 28, no.3, pp651-662, 2020

[17] Li G, Luo J, Wang D, et al. Potential circRNA-disease association prediction using DeepWalk and network consistency projection[J]. Journal of Biomedical Informatics, 2020, 112:103624.

[18] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei.2015. LINE: Large-scale information network embedding. In Proceedings of the International Conference on World Wide Web (WWW 2015). ACM, 1067 – 1077.

[19] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. (2016).

[20] S. Arora, W. Hu, and P. Kothari, “An analysis of the t-sne algorithm for data visualization,” in Conference on Learning Theory, 2018, pp. 1455 – 1462.

Bin Zhang is currently pursuing the master's degree in the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China. He received his B.S. in the School of Software Engineering, Jilin Jianzhu University, China, from 2015 to 2019. His current research interests include deep learning, network embedding, and data mining.

Dan Yang is currently a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China. She received her M.S. and Ph.D. degrees in Computer Software and Theory from Northeastern University, China, in 2004 and 2013 respectively. Dr. Yang was a visiting scholar in New Jersey Institute of Technology, U.S.A from June 2015 to May 2016 supported by Chinese Scholarship Council of the Ministry of Education. Dr. Yang is a member of the CCF (China Computer Federation). Her research interests include data integration, big data management and applications in health care.

Zhihuang Lin is currently pursuing the master's degree in the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China. He received his B.S. in the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China, from 2014 to 2018. His current research interests include deep learning, network embedding, and data mining.