Measuring Disease Similarity Based on Multiple Heterogeneous Disease Information Networks

Ling Tian, Jianliang Gao, Jianxin Wang
School of Computer Science and Engineering
Central South University
Changsha, 410083, China

Ying Wang
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100190, China

Bo Song, Xiaohua Hu
College of Computing & Informatics
Drexel University
Philadelphia, PA, 19104, USA

***Abstract*—Quantifying the similarities between diseases is now playing an important role in biology and medicine, which provides reliable reference information in finding similar diseases. Most of the previous methods for similarity calculation between diseases either use a single-source data or do not fully utilize multi-sources data. In this study, we propose an approach to measure disease similarity by utilizing multiple heterogeneous disease information networks. Firstly, multiple disease-related data sources are formulated as heterogeneous disease information networks which include various types of objects such as disease, pathway, and chemicals. Then, the corresponding subgraphs of these heterogeneous disease information networks are obtained by filtering vertices. Topological scores and semantics scores are calculated in these heterogenous subgraphs using Dynamic Time Warping (DTW) algorithm and meta path method respectively. In this way, we transform multiple heterogeneous disease networks to a homogeneous disease network with different weights on the edges. Finally, the disease nodes can be embedded according to the weights and the similarity between diseases can then be calculated using these n-dimensional vectors. Experiments based on benchmark set fully demonstrate the effectiveness of our method in measuring the similarity of diseases through multi-sources data.**

***Index Terms*—Disease Similarity, Disease Prediction, Disease Information Network**

I. Introduction

The relevance of diseases refers to the extended correlation between the diseases. Studying the similarities between diseases could help us obtain deeper understandings of the relationship between different diseases [1] and provide reliable reference information in the development of new drugs [2] [3], the improvement of treatment regimens, in addition to the comprehension of pathogenesis and prevention of major diseases [4] [5].

There exist some challenges to measure the similarity between diseases. Using different data sources for the task is difficult and uncommon, which leads to the lack of comprehensive consideration and evaluation of the similarity measurement between diseases. The measurements between a disease with all other diseases must be done in the same spatial dimension or the same metric to ensure the consistency and accuracy of the outcome of the disease prediction.

In recent years, many methods were proposed to predict similar diseases, which are mainly divided into three categories: semantics-based, function-based and topology-based. The methods based on semantic similarity use the DO term to calculate the similarity between diseases [6]. However, they do not consider the topology between diseases, or disease and other entities, in the disease network. The results obtained by only semantic-based methods will lack comprehensive assessment. For function-based methods, Cheng [7] proposed $Semfunsim$ method combined the semantic similarity and the functional similarity. However, the relationships between disease and genes are rare, and some diseases are even not related to genes. There are also some methods based on topology. For instance, RADAR [8] was proposed as a framework for learning representations of diseases that captures both semantics and structural identities to calculate their similarity. However, it can not achieve real-time update to calculate disease similarity in multiple disease network when a new data source is added. In order to solve the above mentioned problems, we propose a new approach for measuring disease similarity.

II. The Proposed Approach

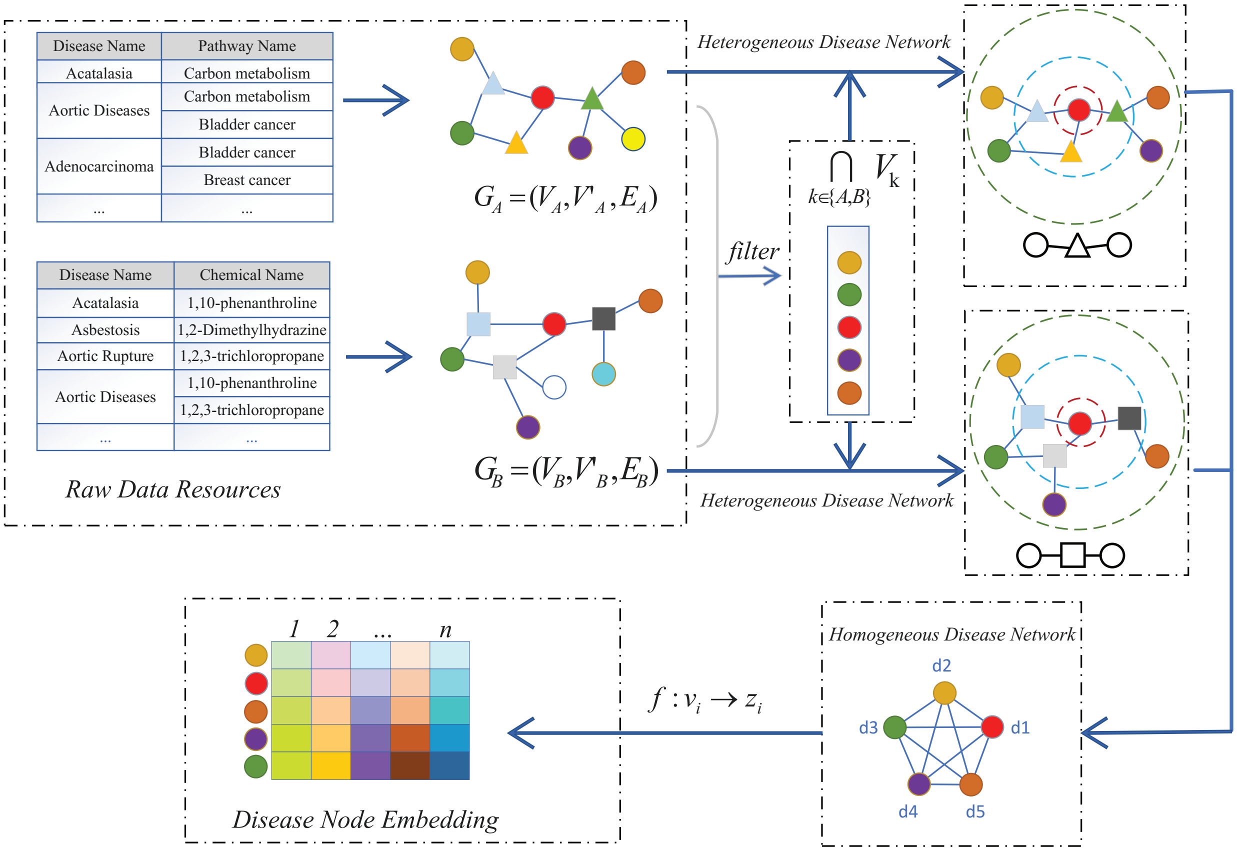
From raw data sources, we construct multiple heterogeneous disease information networks. Disease similarity is then calculated by learning the vector representations on the disease networks. The overall framework is shown in Figure 1.

A. Constructing Disease Information Networks

As shown in Figure 1, the input data sources include multiple types of objects such as disease, pathway, and chemicals. With these objects and their relationships, heterogeneous networks can be constructed accordingly. For example, in Figure 1, the disease-chemical network $\mathcal{G}_B = \{V_B, V'_B, E_B\}$ includes a node set of disease V_B and a node set of chemical V'_B, and E_B represents a set of relationships between the disease node and the chemical node. For multiple heterogeneous networks, we obtain the set of total diseases that appear in all heterogeneous networks by applying the intersect operation on the disease node sets from multiple heterogeneous networks.

In our approach, we transform the heterogeneous networks to homogeneous disease networks by defining two kinds of similarity scores: the semantic score $M(x, y)$ and the topology score $T(x, y)$, where x and y represent any two different disease nodes throughout this paper. For the semantic score $M(x, y)$, where x, y represent two different disease nodes, we calculate it through meta path, which has following definition:

978-1-7281-1867-3/19/$31.00 ©2019 IEEE



Disease Name	Pathway Name
Acatalasia	Carbon metabolism
Aortic Diseases	Carbon metabolism
	Bladder cancer
Adenocarcinoma	Bladder cancer
	Breast cancer
...	...

Disease Name	Chemical Name
Acatalasia	1,10-phenanthroline
Asbestosis	1,2-Dimethylhydrazine
Aortic Rupture	1,2,3-trichloropropane
Aortic Diseases	1,10-phenanthroline
	1,2,3-trichloropropane
...	...

Fig. 1. The similar disease detection framework. Heterogeneous disease information networks (e.g. ○ denotes disease, △ denotes pathway, and □ denotes chemicals) are first constructed from raw data sources. In this example, there are two heterogeneous disease information networks G_A and G_B. The corresponding subgraphs of these heterogeneous disease information networks are obtained by filtering vertices. Then topological scores and semantics scores are calculated in the subgraphs of these heterogenous networks using Dynamic Time Warping (DTW) algorithm and meta path method respectively. In this way, we transform multiple heterogeneous disease networks to a homogeneous disease network with different weights on the edges. Finally, the disease nodes can be embedded according to the weights as n-dimensional vectors and the similarity between diseases can then calculated using these vectors.

*Definition 1: **(Meta Path.)*** A meta path p is a path defined by the path length and the types of nodes and edges in the graph.

For example, $(``disease" \rightarrow ``chemicals" \rightarrow ``disease")$ is a meta path with length three. Its semantic score $M(x,y)$ can then be defined as:

$$M(x,y) = \frac{2 * |\{p_{x \rightarrow y} \mid p_{x \rightarrow y} \in \mathbb{P}\}|}{|\{p_{x \rightarrow x} \mid p_{x \rightarrow x} \in \mathbb{P}\}| + |\{p_{y \rightarrow y} \mid p_{y \rightarrow y} \in \mathbb{P}\}|} \tag{1}$$

where $\mathbb{P}$ represents the set of pre-defined meta paths between disease nodes. For example, $p_{x \rightarrow y}$ is a meta path instance between disease x and disease y, $p_{x \rightarrow x}$ is a meta path instance between disease x and disease x. $M(x,y)$ refers to the similarity score between diseases based on semantic.

For the topology score $T(x,y)$, the definition is:

$$T(x,y) = e^{-(\sum_{i=0}^{\tau} \beta^i * DTW(\mho_i(x), \mho_i(y)))} \tag{2}$$

where $\mho_i(\cdot)$ refers to the degree sequence of the i^{th} hop neighbor node, and $i = 0$ represents the node itself. β is a parameter indicating the weights of the neighbor nodes of different hops. $DTW(\mho_i(x), \mho_i(y))$ represents the distance between the sequence degree of disease nodes using Dynamic Time Warping (DTW) algorithm [9]. $T(x,y)$ refers to the similarity score between diseases based on topology.

Then, we integrate both semantic and topology scores together to be the integrated similarity score, which can be defined as follows:

$$S(x,y) = \alpha * M(x,y) + (1-\alpha) * T(x,y) \tag{3}$$

where $\alpha \in [0,1]$ is a parameter to adjust the contribution of the two similarity scores M and T towards the integrated similarity score S.

Finally, we construct a homogeneous complete graph with different weights on the edges. The weight of the edge between each disease node is generated by similar scores of multiple heterogeneous networks.

$$W(x,y) = \sum_{k=1}^{|\mathbb{G}|} w_k * S_{\mathcal{G}_k}(x,y) \tag{4}$$

where $S_{\mathcal{G}_k}(x,y)$ denotes the integrated similarity score of two disease nodes in the k^{th} $\mathcal{G}$, $\mathcal{G} \in \mathbb{G}$, and w_k represents the proportion of the k^{th} $\mathcal{G}$ network contributing to the weight of the link.

B. Embedding the Diseases for Measuring Similarity

For the constructed networks, we further obtain the vector representation of each node which can be used for various network application tasks such as node similarity measurement.

In our method, we embed each disease node as an n-dimensional vector, where n is the number of nodes in the homogeneous disease network. For any disease node v_i, its vector representation is:

$$v_i = (W(i,1), W(i,2), ..., W(i,n)) \tag{5}$$

where $W(i,j)$ denotes the weight from disease node i to j. The value of each dimension in the vector is generated by the weight of the link edge in the disease similarity network formed by the node and each encoded node.

Since each disease node can be represented by a vector in the same dimensional space, the disease similarity can be calculated by similarity measurement between vectors, such as cosine distance, Euclidean distance, and so on. Our method adopts the Euclidean distance to calculate the similarity of diseases.

$$Sim(x,y) = \frac{1}{1+\sqrt{(v_x - v_y)\cdot(v_x - v_y)^T}} \tag{6}$$

where v_x, v_y are the vector representations of disease nodes x and y respectively. The normalization is to make all the values fall in the scope of [0,1] for fair comparison. The larger $Sim(x,y)$ of two diseases means the higher similarity.

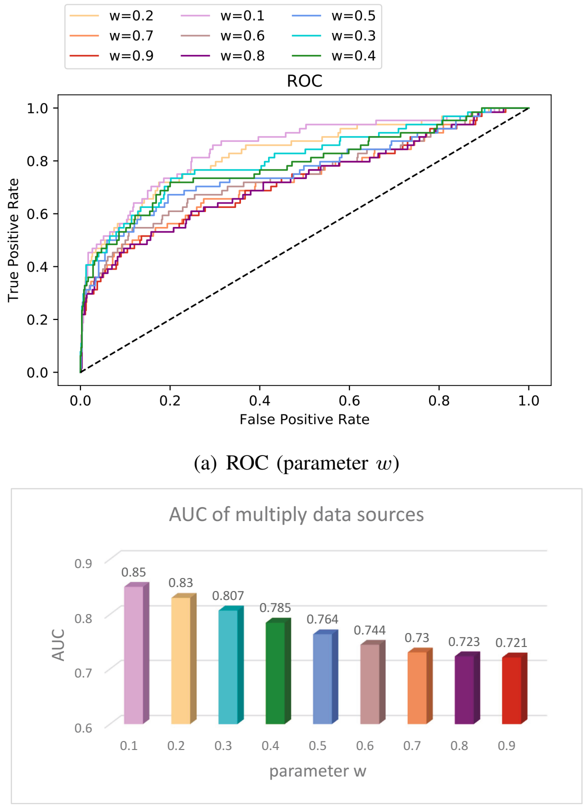
III. Experiment Evaluation

A. Datasets

We use two associations datasets of diseases and other diseases related entities which are extracted from the raw datasets of the ctdbase. The disease-chemical association includes 6,206 disease nodes, 4,180 chemical nodes, and 1,048,547 disease-chemical relationships; the disease-pathway association includes 4,997 disease nodes, 2,338 pathway nodes, and 569,716 disease-pathway relationships. After filtering the raw datasets, the dataset includes 4,986 disease nodes, 4,159 chemical nodes, 2,336 pathway nodes, 1,042,765 disease-chemical relationships, and 569,642 disease-pathway relationships.

B. Experiment Setup

We adjusted the parameters α from 0.1 to 0.9 to explore the influences of different contributions of semantic and topology scores toward the integrated similarity scores on the experimental results. In order to investigate the effects of different disease-related data sources on disease similarity results, we also perform experimental comparisons of multiple values of parameter w from 0.1 to 0.9. To discuss the effectiveness of the proposed method, we utilize the disease pairs in the benchmark and set the range of prediction as 1%, 5%, 10% and 20% to evaluate the correct rate of disease prediction of our method.



(a) ROC (parameter w)

(b) AUC (parameter w)

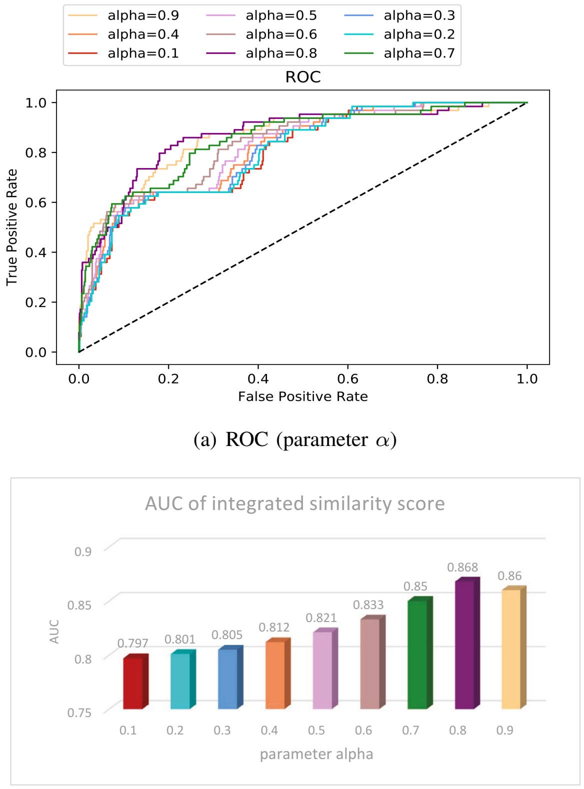
Fig. 2. Parameter w for multiple disease information networks

C. Evaluation Results

a) Parameter w for multiple disease information networks: We define parameter w and 1-w to indicate the effected weights of the chemical and pathway on diseases respectively. We conduct the experiments based on the 65 diseases in the benchmark. Receiver operating characteristic (ROC) curves are then drawn with the benchmark set against 50 random sets. Each random set contains 650 randomly selected pairs. The experimental results are shown in Figure 2.

In the Figure 2(a), each color represents a parameter value of w that varies between [0.1, 0.9]. Each column value in Figure 2(b) is obtained from the area under the ROC Curve (AUC). From Figure 2, we can see that: with the parameter value changing from 0.1 to 0.9, the AUC value tends to decrease gradually, but the overall AUC value can reach 72% or more. When the value of parameter w equals to 0.1, the correct rate of our proposed method of disease similarity can reach 85%. In our experiment of comparing each single data source, the difference of similarity results obtained by single data source of chemical-disease and pathway-disease is relatively large. Therefore, we can conclude from Figure 2 that by combining two or more data sources, the proposed method tend to have more stable results.

b) Parameter α for integrated similarity score: In the research, the contribution of two similarity scores is defined by



(a) ROC (parameter α)

(b) AUC (parameter α)

Fig. 3. Parameter α for integrated similarity score

the parameter α for the integrated similarity score. We set the parameter α to change from 0.1 to 0.9, and the experimental results are shown in Figure 3.

In the Figure 3, each color represents a parameter value that varies between [0.1, 0.9]. Each column value in Figure 3(b) is obtained from the area under the ROC Curve (AUC). From Figure 3, we can see that in the results of combining two similarity scores based on two data sources, our proposed method get the best when α = 0.8, and the accuracy reaches 86.8%. However, the difference in the accuracy obtained by the other parameter values is not obvious, so this result proves that the weight parameters between the two similar score calculation indicators are not sensitive in general.

c) Disease prediction:

The similar disease pair set (A-B) in the benchmark is a set of one-to-one correspondence between nodes in the disease set A and nodes in the disease set B. Our experiments take disease set A as the set of nodes to be predicted, and take disease set B as the reference set. Using our proposed method to calculate the similarity score between any pair of diseases, with reference to the corresponding disease nodes in the reference set B, the accuracy of similar disease predictions in different ranges is counted. The Table I compares the results of similar disease predictions based on a chemical-disease data source, a pathway-disease data source, and a dataset that combines the two data sources.

TABLE I
ACCURACY OF DISEASE PREDICTION BASED ON BENCHMARK

data source	HitΔ1	HitΔ5	HitΔ10	HitΔ20
chemical-disease	9.4%	34.4%	42.2%	57.8%
pathway-disease	12.5%	39.1%	53.1%	78.1%
{chemical,pathway}-disease	14.1%	39.1%	54.7%	78.1%

From Table I, we can see that: As the scope of prediction expands, the accuracy increases. Moreover, our proposed method on multi-source data performs better than single data source. As the range of predictions expands, the proposed method is better than the disease prediction of only chemical-disease data source, and it is better than or equal to the prediction result of the only pathway-disease data source. In summary, the prediction results of disease similarity with combining multiple data sources performs better than a single data source.

IV. CONCLUSION

To measure the similarity between diseases, we propose a new method by utilizing multiple heterogenous disease information networks. Through the validation with the benchmark, the results of our method show high accuracy in predicting similar diseases. The algorithms of our method can also allow update in real- time for the calculation of disease similarity when new data sources are added..

ACKNOWLEDGEMENT

This work was supported in part by NSFC 61873288, NSF IIS 1815256, NSF IIS 1744661 and the Central Universities-Freedom Explore Program of Central South University (No. 2019zzts585).

REFERENCES

[1] Peng Ni, Jianxin Wang, Ping Zhong, Yaohang Li, Fangxiang Wu, and Yi Pan. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP:1–1, 2018.

[2] Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, and Roded Sharan. Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, 18(2):133–145, 2011.

[3] Liang Cheng, Yue Jiang, Zhenzhen Wang, Hongbo Shi, Jie Sun, Haixiu Yang, Shuo Zhang, Yang Hu, and Meng Zhou. Dissim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. *Scientific Reports*, 6:30024–30030, 2016.

[4] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *Plos Computational Biology*, 6(1):e1000641, 2010.

[5] Min Li, Ruiqing Zheng, Qi Li, Jianxin Wang, Fang-Xiang Wu, and Zhuohua Zhang. Prioritizing disease genes by using search engine algorithm. *Current Bioinformatics*, 11(2):195–202, 2016.

[6] Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015.

[7] Liang Cheng, Jie Li, Peng Ju, Jiajie Peng, and Yadong Wang. Semfunsim: a new method for measuring disease similarity by integrating semantic and gene functional association. *Plos One*, 9(6):e99415, 2014.

[8] Ruiqi Qin, Lei Duan, Huiru Zheng, Jesse Li-Ling, Kaiwen Song, and Xuan Lan. Radar: Representation learning across disease information networks for similar disease detection. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 482–487. IEEE, 2018.

[9] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.