Predicting drug-disease associations based on the known association bipartite network

Wen Zhang[1,*], Xiang Yue[2], Yanlin Chen[3], Weiran Lin[1], Bolin Li [2], Feng Liu[2], Xiaohong Li[1,*]

1. School of Computer, Wuhan University, Wuhan 430072, China

2. International School of Software, Wuhan University, Wuhan 430072, China

3. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

zhangwen@whu.edu.cn, tommy96@whu.edu.cn, chenyanlin@whu.edu.cn, waynelin@whu.edu.cn, bolin611@whu.edu.cn, fliuwhu@whu.edu.cn, leexh@whu.edu.cn

*corresponding author

Abstract—*recent studies show that drug-disease associations provide important information for drug discovery and drug repositioning. Wet experimental identification of drug-disease associations is time-consuming and labor-intensive. Therefore, the development of computational methods that predict drug-disease associations is an urgent task. In this paper, we propose a novel computational method named NTSIM, which only uses known drug-disease associations to predict unobserved associations. First of all, known drug-disease associations are represented as a drug-disease bipartite network, and a novel similarity measure named linear neighborhood similarity (LNS) is proposed to calculate drug-drug similarity and disease-disease similarity based on the bipartite network. Then, we predict unobserved drug-disease associations in the similarity-based graph by using label propagation process. In the computational experiments, this proposed method achieves high-accuracy performances, and outperforms representative state-of-the-art methods: PREDICT, TL-HGBI and LRSSL. Our studies reveal that known drug-disease associations can provide enough information to build the high-accuracy prediction models; linear neighbor similarity (LNS) can lead to better performances than other similarity measures such as Jaccard similarity, Gauss similarity and cosine similarity; the bipartite network-derived features outperform the drug biological features and disease semantic features.*

Keywords—*drug-disease associations; association profiles; linear neighborhood similarity*

I. INTRODUCTION

Drugs are chemicals that treat, cure, prevent, or diagnose diseases. Events that drugs exert effects on diseases are referred to as drug-disease associations. Drugs may play a role in the etiology of a disease, e.g. exposure to a drug causes lung cancer; drugs may have a therapeutic role in a disease, e.g. a drug can treat leukemia. Identification of drug-disease associations can provide important information for drug discovery and drug repositioning. Computational methods can guide laborious and costly experiments to identify drug-disease associations, and thus the development of computational methods is an urgent task.

Recently, the machine learning methods were introduced to the drug-disease association prediction, for their capability of dealing with complicated data. Gottlieb et al. [1] constructed a drug-disease association predictor that integrated molecular structures, molecular activities and semantic information. Yang et al. [2] built Naive Bayes models to predict indications of diseases based on drug side effects. Wang et al. [3] trained support vector machine (SVM) models based on drug structures, drug target proteins and drug side effects. Huang et al. [4] built a heterogeneous network for drugs, genomic information and disease phenotypes, and used the random walk to make predictions. Oh et al. [5] characterized drug-disease relationship by using similarity-based features and module distance-based features, and then respectively adopted decision tree, multi-layer perceptron and random forest to build prediction models. Wang et al. [6] proposed a computational framework based on a three-layer heterogeneous network model (TL-HGBI). Martínez et al. [7] built a network of interconnected drugs, proteins and diseases. Wang et al. [8] adopted recommendation systems to make predictions. Moghadam et al.[9] adopted the kernel fusion technique to combine different drug features and disease features, and then built SVM models. Liang et al.[10] proposed a Laplacian regularized sparse subspace learning method (LRSSL) which integrated drug chemical information, drug target domain information and target go information.

Existing machine learning methods usually utilize drug features, disease features and known associations to predict novel drug-disease associations. However, drug features and disease features are not always available, and these methods can't work when information is incomplete. Related studies [11-15] in drug repositioning, drug-target interaction prediction, miRNA-disease association prediction and drug side effect prediction show that known association information is an important information source and can lead to high-accuracy prediction models. In this paper, we propose a computational method named NTSIM, which only uses known drug-disease associations to predict unobserved associations. First of all, known drug-disease associations are represented as a drug-disease association bipartite network, and a novel similarity measure named linear neighborhood similarity (LNS) is proposed to calculate drug-drug similarity and disease-disease similarity based on the bipartite network. Then, we predict unobserved drug-disease associations based on the similarity-based graph by using the label propagation process. In the computational experiments, this proposed method achieves high-accuracy performances, and outperform several state-of-the-art methods: PREDICT, TL-HGBI and LRSSL. The

studies show that known drug-disease associations can provide enough information to build the high-accuracy prediction models; linear neighborhood similarity (LNS) can lead to better performances than other similarity measures, including Jaccard similarity, Gauss similarity and cosine similarity; the bipartite network-derived features outperform drug biological features and disease semantic features. Our proposed method is promising for predicting unobserved drug-disease associations.

II. Materials and Methods

A. Datasets

To the best of our knowledge, there are several databases which describe drug features, disease features and drug-disease associations.

Comparative Toxicogenomics Database (CTD) is a publicly available database for chemical-disease relationship. We collect curated drug-disease associations from the latest CTD database, and these associations were originally extracted from literatures. We collect drug substructures from PubChem Compound database[16], drug targets, drug enzymes, drug-drug interactions from DrugBank database[17], and drug pathways from KEGG DRUG database [18]. Drugs are mapped between different databases according to the ID conversion table from CTD Database. We obtain diseases MeSH descriptors from U.S. National Library of Medicine. Therefore, we can obtain drugs with comprehensive information (substructures, targets, enzymes, drug-drug interactions and pathways) and diseases with MeSH descriptors. We remove drugs associated with less than 10 diseases or disease associated with less than 10 drugs. Finally, we compile a dataset with 18416 associations between 269 drugs and 598 diseases.

Moreover, we consider several benchmark datasets ever used for the drug-disease association prediction. Gottlieb et al. [1] compiled a dataset with 1,933 associations between 593 drugs in DrugBank and 313 diseases in OMIM, and then constructed the "PREDICT" model. The dataset also includes five types of drug-drug similarities and two types of disease-disease similarities. Wang et al. [6] collected 1461 associations between 1409 DrugBank drugs and 5080 OMIM diseases, and calculated the drug-drug structure similarity and the disease-disease semantic similarity. The dataset was used to develop the model "TL-HGBI". Liang et al. [10] adopted 3051 associations between 763 drugs and 681 diseases from [19], and collected drug substructures, protein domains of target proteins, and gene ontology terms of target proteins. Then, three types of drug-drug similarities were calculated, and the disease-disease semantic similarity was obtained from [19]. The dataset was ever used for the model "LRSSL". We obtained PREDICT dataset, TL-HGBI dataset from authors, and LRSSL dataset was available at https://github.com/LiangXujun/LRSSL.

B. Problem Description

Formally, we are given a set of drugs $D = \{D_1, D_2, \cdots, D_n\}$, a set of diseases $S = \{S_1, S_2, \cdots, S_m\}$ and drug-disease associations. Our task is to predict unobserved drug-disease associations based on known associations.

As shown in Fig. 1, drugs, diseases and their associations can be formulated as a bipartite network, which uses drugs, diseases as nodes and uses associations as edges. The bipartite network is characterized by an adjacency matrix M. This is, $M_{ij} = 1$ if drug D_i is associated with S_j; otherwise, $M_{ij} = 0$.

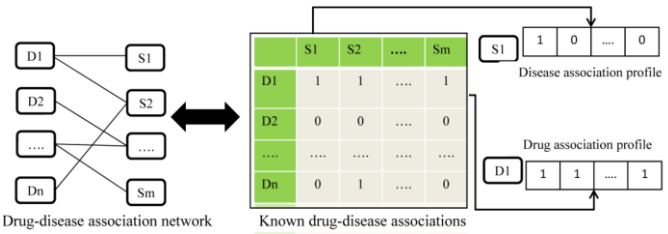

Fig.1. The drug-disease association-based network, disease association profiles and drug association profiles

Based on the bipartite network, we introduce the association profiles of drugs and the association profiles of diseases. The drug association profile X_{D_i} for a drug D_i is a binary vector encoding the presence or absence of associations with every disease in the drug-disease association network; the disease association profile X_{S_i} for a disease S_i is the binary vector specifying the presence or absence of associations with every drug in the drug-disease association network. As shown in Fig.1, drug association profiles and disease association profiles correspond to the row vectors and column vectors of the adjacency matrix M. The association profiles are important information from the drug-disease association network, and we use them to build models and make predictions.

C. Linear Neighborhood Similarity

The assumption that similar drugs (diseases) are likely to have associations with a same disease (drug) is widely used in related works. How to measure drug-drug similarity or disease-disease similarity is a critical issue. Jaccard similarity, cosine similarity and Gauss similarity are usually adopted to measure similarity [15, 20, 21]. However, the study [22] revealed that these similarity measures are not robust to points that connect different classes. According to S.T Roweis's study [22], neighborhood data points in the feature space can be considered to be linear, and Wang [23] stated that each data point can be reconstructed by a linear combination of its neighbors. Therefore, we proposed the linear neighborhood similarity and used it for the drug side effect prediction [20].

Given the association profiles of n drugs $X = \{x_1, x_2, \cdots, x_n\}$, each data point x_i is reconstructed by the linear weighted sum of neighbor data points, and the objective function is to minimize the reconstruction error,

$$\begin{aligned} \varepsilon_i &= \left\| x_i - \sum\nolimits_{i_j, X_{i_j} \in N(X_i)} w_{i,i_j} x_{i_j} \right\|^2 + \lambda_i \|w_i\|^2 \\ &= w_i^T (G^i + \lambda_i I) w_i \\ \text{s.t.} &\sum\nolimits_{i_j : X_{i_j} \in N(x_i)} w_{i,i_j} = 1, w_{i,i_j} \geq 0, j = 1, \cdots, K \end{aligned} \quad (1)$$

where $N(x_i)$ is the set of K nearest neighbors of x_i. $\|\cdot\|$ is the Euclidean norm. x_{i_j} is the jth neighbor of x_i, $w_i = \left(w_{i,i_1}, \cdots, w_{i,i_K}\right)^T$ and $G^i_{i_j,i_k} = (x_i - x_{i_j})(x_i - x_{i_k})^T$ is the entry of the Gram matrix G^i. I is the identity matrix of order n, and λ_i is the regularization parameter. According to (1), weights are calculated for each data point by using the standard quadratic programming, and the weight matrix (similarity matrix) W is obtained. More details are given in [20].

Since $G^i + \lambda_i I$ should be recomputed for each data point in (1), calculating linear neighborhood similarity for large-scale data is extremely time-consuming. Here, we reformulate the

objective function of linear neighborhood similarity in the matrix form,

$$\min_W \frac{1}{2}\|X-(C\odot W)X\|_F^2+\frac{\mu}{2}\|(C\odot W)e\|_F^2 \quad (2)$$
$$\text{s.t.}\,(C\odot W)e=e, W\geq 0$$

where $c_{ij}=1$ if $x_j\in N(x_i)$; otherwise, $c_{ij}=0$. μ is the trade-off parameter. $e=(1,1,\cdots,1)^T$ and $\|\cdot\|_F$ is the Frobenius norm. As analyzed in [20], we can easily set $\mu=1$ in the experiments.

To solve (2), we introduce the Lagrange multipliers $\lambda\in\mathbb{R}^n$ and $\Phi\in\mathbb{R}^{n\times n}$, and obtain Lagrange function,

$$L=\frac{1}{2}\|X-(C\odot W)X\|_F^2+\frac{\mu}{2}\|(C\odot W)e\|_F^2 - \lambda^T\big((C\odot W)e-e\big)-tr(\Phi^T W)$$

We calculate the derivative of L with respect to W,

$$\nabla_W L=C\odot\big((C\odot W)XX^T+\mu(C\odot W)ee^T-XX^T-\lambda e^T\big)-\Phi$$

By KKT complementary slackness condition, we have

$$\big((C\odot W)XX^T+\mu(C\odot W)ee^T-XX^T-\lambda e^T\big)_{ij}W_{ij}c_{ij}=0$$

W_{ij} is updated as follows,

$$W_{ij}=\begin{cases}0 & x_j\notin N(x_i)\\ W_{ij}\dfrac{(XX^T+\lambda e^T)_{ij}}{\big((C\odot W)XX^T+\mu(C\odot W)ee^T\big)_{ij}} & x_j\in N(x_i)\end{cases} \quad (3)$$

However, the update rules in (3) still have the parameter λ. For x_i, (2) is equivalent to the following form,

$$\min_{\omega_i} L^i=\frac{1}{2}\omega_i^T G^i\omega_i+\frac{\mu}{2}\|\omega_i\|_1^2 \quad (4)$$
$$\text{s.t.}\,e^T\omega_i=1,\omega_i\geq 0$$

Then, the Lagrange function of (4) is

$$L^i=\frac{1}{2}\omega_i^T G^i\omega_i+\frac{\mu}{2}\|\omega_i\|_1^2-\lambda_i(e^T\omega_i-1)-\eta^T\omega_i$$

The KKT conditions are

$$\begin{cases}\nabla_{\omega_i}L^i=G^i\omega_i+\mu ee^T\omega_i-\lambda_i e-\eta=0\\ \nabla_{\lambda_i}L^i=e^T\omega_i-1=0\\ \eta\geq 0,\omega_i\geq 0,\eta_j\omega_{i,i_j}=0\end{cases}$$

Then

$$\omega_i^T\nabla_{\omega_i}L^i=\omega_i^T G^i\omega_i+\mu(\omega_i^T e)^2-\lambda_i\omega_i^T e=0$$

We obtain

$$\lambda_i=(\omega_i^T G^i\omega_i+\mu(e^T\omega_i)^2)/e^T\omega_i$$

We notice that the reconstruction error $\omega_i^T G^i\omega_i\approx 0$ and $e^T\omega_i=1$ if ω_i is the optimal solution for (4). Thus $\lambda_i\approx\mu$, and we let $\lambda=\mu e$, and rewrite the update rules (3) as

$$W_{ij}=\begin{cases}0 & x_j\notin N(x_i)\\ W_{ij}\dfrac{(XX^T+\mu ee^T)_{ij}}{\big((C\odot W)XX^T+\mu(C\odot W)ee^T\big)_{ij}} & x_j\in N(x_i)\end{cases} \quad (5)$$

The linear neighborhood similarity for all data points are calculated at once by (5). Compared with our previous approach in (1), the new approach is much more efficient for calculating linear neighborhood similarity.

D. The drug-disease association inference methods

The known drug-disease associations form a bipartite network, and drugs and diseases can be represented by association profiles. Then, we calculate the drug-drug similarity and disease-disease similarity in the association space by using linear neighborhood similarity measure, and propose drug-disease association inference methods, which build prediction models by using the label propagation process.

Given a similarity matrix W for n drugs, we construct a directed graph which considers n drugs as nodes and use the similarity as the weights of edges. Known associations between a specified disease and all drugs are considered as initial labels of nodes.

For the disease S_j, initial labels of nodes are actually the jth column of the association matrix M, which are denoted as $M(:,j)$. The label information is propagated in this directed graph, and the labels of nodes are updated by absorbing labels of neighbors with the probability α and retaining the initial labels with the probability $1-\alpha$. Let P_j^t donate the labels of nodes at the tth iteration, the update from step $t-1$ to step t is,

$$P_j^t=\alpha W P_j^{t-1}+(1-\alpha)M(:,j) \quad (6)$$

If we consider labels for all diseases $S_1,S_2,\cdots,S_m$ at the same time, we can reformulate (6) in the matrix form,

$$P^t=\alpha W P^{t-1}+(1-\alpha)M \quad (7)$$

(7) can be written as,

$$P^t=(\alpha W)^t M+(1-\alpha)\sum_{i=0}^{t-1}(\alpha W)^i M \quad (8)$$

The spectral radius $\rho(W)\leq 1$, $0<\alpha<1$ and $\lim_{t\to\infty}(\alpha W)^t=0$. We can know $\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha W)^i=(I-\alpha W)^{-1}$, and thus the iteration will converge,

$$P=\lim_{t\to\infty}P^t=(1-\alpha)(I-\alpha W)^{-1}M \quad (9)$$

Based on the linear neighborhood similarity and label information, we can develop the drug similarity-based inference method (DSIM), the disease similarity-based inference method (DISIM) and the network topological similarity-based inference method (NTSIM). The methods are briefly introduced as follows.

DSIM predicts unobserved drug-disease associations by label propagation on drug-drug similarity network, and the prediction matrix for drug-disease associations is calculated by,

$$P_{\text{DSIM}}=(1-\alpha)(I-\alpha W_{DD})^{-1}M_{DS} \quad (10)$$

where $M_{DS}=M$, and W_{DD} is the drug-drug similarity matrix.

Similarly, DISIM predicts unobserved disease-drug associations by label propagation on disease-disease similarity network, and the prediction matrix for disease-drug associations is calculated by,

$$P_{\text{DISIM}}=((1-\alpha)(I-\alpha W_{SS})^{-1}M_{SD})^T \quad (11)$$

where $M_{SD}=M'$, and W_{SS} is the disease-disease similarity matrix.

Based on known drug-disease associations, we can predict unobserved associations by using DSIM or DISIM. The method NTSIM is the combination of DSIM and DISIM, and makes use of both drug-drug similarity and disease-disease similarity. The prediction matrix for disease-drug associations is calculated by,

$$P_{\text{NTSIM}}=(P_{\text{DSIM}}+P_{\text{DISIM}})/2 \quad (12)$$

III. Experiments and Results

A. Evaluation Metrics

We implement five-fold cross-validation to evaluate performances of prediction models. The five-fold cross-validation (5-CV) randomly splits known drug-disease associations into five subsets. In each fold, we keep one subset

as the testing set, and use others as the training set. The model is built on the associations in the training set, and then makes predictions for other drug-disease pairs. The predictions and real labels (known associations or not) for these pairs are used to calculate evaluation metrics. The average performances of five folds are adopted.

We use several evaluation metrics: the area under receiver-operating characteristic curve (AUC), the area under precise-recall curve (AUPR), sensitivity (SEN), specificity (SPEC), precision (PRE), accuracy (ACC) and F-measure (F). We adopt AUPR as the primary metric, which takes into account both recall and precision.

B. Performances of prediction models

1) The Influence of Parameters

DSIM, DISIM and NTSIM have two major components: the similarity calculation and similarity-based inference. The linear neighborhood similarity has the parameter: neighbor number K, and the similarity-based inference has the parameter: absorbing probability α. Therefore, DSIM, DISIM and NTSIM have two parameters: the neighbor number K and the absorbing probability α. We consider the combinations of following values: $\{10\%, 30\%, 50\%, 70\%, 90\%\}$ of number of data points for K and $\{0.05, 0.1, 0.15 \ldots 0.95\}$ for α. We use different parameter values to build DSIM models, DISIM models and NTSIM models, and then evaluate the influence of parameters on DSIM, DISIM and NTSIM.

The parameters and AUPR scores of corresponding models are shown in Fig.2. Generally speaking, we can observe similar influence of parameters on AUPR scores of DSIM models, DISIM models and NTSIM models. The neighbor number can directly influence the performances, and greater neighbor number is likely to produce better performance. As the absorbing probability α increases, the performances of three methods will increase, but then decrease after reaching a peak (approximately $\alpha = 0.3$). NTSIM which combines drug-drug similarity and disease-disease similarity produces greater AUPR scores than DSIM which uses drug-drug similarity and DISIM which uses disease-disease similarity.

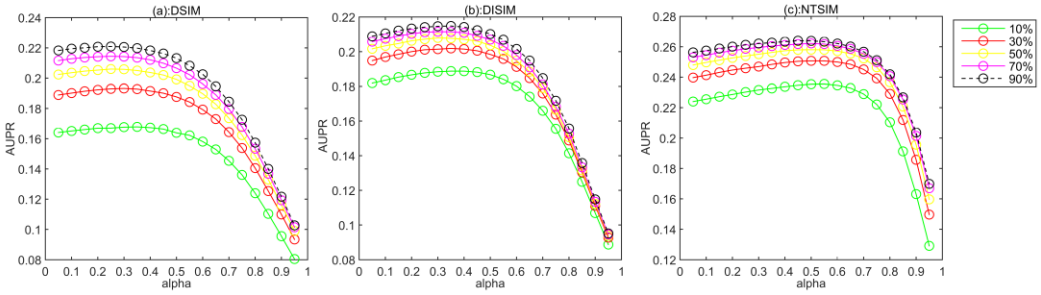

Fig.2. AUPR scores of DSIM models, DISIM models and NTSIM models using different parameter values

2) Comparison of Different Similarity Measures

Based on association profiles of drugs and diseases, DSIM, DISIM and NTSIM calculate drug-drug linear neighborhood similarity and disease-disease linear neighborhood similarity, and build prediction models. For comparison, we also consider other similarity measures: Jaccard similarity, cosine similarity and Gauss similarity. Thus, we calculate drug-drug similarity and disease-disease similarity by different similarity measures, and then build prediction models to compare these similarity measures. In the experiments, the neighbor number K is set as 90% of the data point number for the linear neighborhood similarity; Gauss similarity for feature vectors x_i and x_j are calculated by $\exp(-\sigma\|x_i - x_j\|^2)$, and the bandwidth parameter $\sigma = 1/(\sum_{i=1}^{n}|x_i|/n)$.

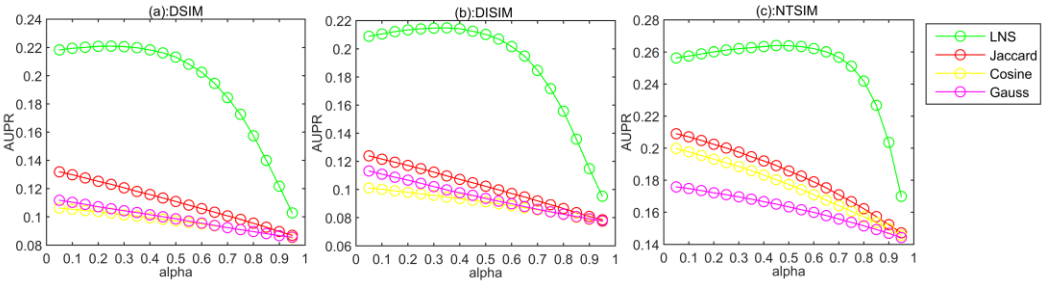

Fig.3. AUPR scores of DSIM models, DISIM models and NTSIM models using different similarity measures

As shown in Fig.3, the linear neighborhood similarity can produce consistently better performances than other similarity measures when used for DSIM, DISIM and NTSIM. We can conclude that the linear neighborhood similarity is superior to Jaccard similarity, cosine similarity and Gauss similarity in our task, and NTSIM outperforms DSIM and DISIM regardless of any similarity measure.

3) Comparison of Different Features

DSIM, DISIM and NTSIM predict unobserved associations based on known drug-disease associations. The first step of DSIM, DISIM and NTSIM is to calculate drug-drug similarity and disease-disease similarity based on association profiles. In order to demonstrate the usefulness of association profiles, we consider five drug features (i.e. substructures, targets, enzymes, pathways and drug-drug interactions) and the disease semantic information. We can calculate five types of drug-drug linear neighborhood

similarities based on drug features and the disease-disease semantic similarity [24-26], and then respectively build different NTSIM models by using one drug feature-derived similarity and the disease semantic similarity. Therefore, we can evaluate the performances of NTSIM models to compare different information sources, i.e. known drug-disease associations and five drug features. As shown in Fig 4, the NTSIM model based on known drug-disease associations can produce better results than the NTSIM model based on drug features. Based on above discussions, known drug-disease associations can form the drug-disease association bipartite network, and the bipartite network-derived association profiles can be used to build high-accuracy models.

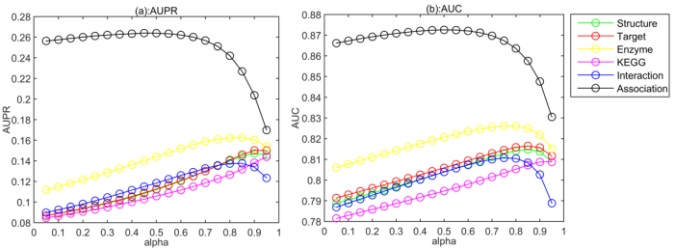

Fig.4. AUPR scores of NTSIM models based on known drug-disease associations and drug features

Since NTSIM can produce better results than DSIM and DISIM, NTSIM is finally adopted to construct models for the drug-disease association prediction. We set the neighbor number K=90% of the data point number and the absorbing probability $\alpha = 0.3$ for NTSIM in the following experiments.

C. Comparison with State-of-the-art Methods

In this section, we compare NTSIM with state-of-the-art methods to demonstrate the superior performance of our method. For comprehensive study, we consider baseline methods and existing drug-disease association prediction methods.

NTSIM predicts unobserved drug-disease associations based on the association bipartite network. Here, we adopt the resource allocation method[27] as the baseline method. The resource allocation method is a popular method for predicting unobserved links in the bipartite network, and has been successfully applied to lots of problems [14, 15, 21, 28]. Therefore, we compare our method NTSIM and resource allocation method by using four datasets, and the results of models evaluated by the cross validation are shown in Table I. Clearly, NTSIM can produce better performances than the resource allocation method in terms of different evaluation metrics, when evaluated by 5-CV on our dataset and three benchmark datasets.

TABLE I. PERFORMANCES OF OUR METHOD AND RESOURCE ALLOCATION METHOD EVALUATED BY 5-CV ON OUR DATASET AND THREE BENCHMARK DATASETS

Methods	Datasets	AUPR	AUC	SEN	SPEC	PREC	ACC	F
Resource allocation	our dataset	0.1895	0.8408	0.2864	0.9738	0.2231	0.9564	0.2494
Our method	our dataset	0.2621	0.8709	0.3250	0.9805	0.3019	0.9640	0.3126
Resource allocation	PREDICT dataset	0.3212	0.8462	0.3580	0.9990	0.4362	0.9977	0.3923
Our method	PREDICT dataset	0.3376	0.9205	0.3678	0.9990	0.4624	0.9977	0.4022
Resource allocation	TL-HGBI dataset	0.0951	0.7747	0.1937	1.0000	0.1718	0.9999	0.1672
Our method	TL-HGBI dataset	0.2631	0.9616	0.4032	0.9999	0.1658	0.9999	0.2349
Resource allocation	LRSSL dataset	0.2094	0.8059	0.2734	0.9994	0.3483	0.9985	0.3025
Our method	LRSSL dataset	0.2693	0.9021	0.3078	0.9994	0.3757	0.9986	0.3384

To the best of our knowledge, a great number of methods have been proposed to predict drug-disease associations. Here, we consider three state-of-the-art methods: PREDICT [1], TL-HGBI [6] and LRSSL [10] as benchmark methods. PREDICT and TL-HGBI are classic methods, and LRSSL is the latest method. According to the publications[1] [6] [10], these methods can produce good performances. PREDICT calculates the score of a given drug-disease association (d_r, d_i) by $max_{d_{r'}, d_{i'} \neq d_r, d_i} \sqrt{S(d_r, d_{r'}) \times S(d_i, d_{i'})}$, where $S(d_r, d_{r'})$ is the drug-drug similarity and $S(d_i, d_{i'})$ is the disease-disease similarity. TL-HGBI is a three-layer heterogeneous network method, which uses existing data about diseases, drugs and drug targets to make predictions. LRSSL is the Laplacian regularized sparse subspace learning method, which integrates drug chemical information, drug target domain information and target annotation information for prediction. We implement PREDICT by following details in [1]; we obtain the source code of TL-HGBI from authors; the source code of LRSSL is publicly available. Therefore, we can fairly compare our method with three methods.

TABLE II. PERFORMANCES OF OUR METHOD AND BENCHMARK METHODS EVALUATED BY 5-CV ON THREE BENCHMARK DATASETS

Methods	Datasets	AUPR	AUC	SEN	SPEC	PREC	ACC	F
PREDICT	PREDICT dataset	0.1507	0.9020	0.3414	0.9929	0.0914	0.9915	0.1437
Our method	PREDICT dataset	0.3376	0.9205	0.3678	0.9990	0.4624	0.9977	0.4022
TL-HGBI	TL-HGBI dataset	0.0492	0.9584	0.1697	0.9999	0.0571	0.9998	0.0840
Our method	TL-HGBI dataset	0.2631	0.9616	0.4032	0.9999	0.1658	0.9999	0.2349
LRSSL	LRSSL dataset	0.1789	0.8250	0.2167	0.9989	0.1988	0.9979	0.2018
Our method	LRSSL dataset	0.2693	0.9021	0.3078	0.9994	0.3757	0.9986	0.3384

As far as we know, PREDICT, TL-HGBI and LRSSL make use of different features to make predictions, and they have to utilize their own datasets to build prediction models. PREDICT dataset, TL-HGBI dataset and LRSSL dataset contain known drug-disease associations, drug features and disease features. Our method "NTSIM" only uses the known drug-disease associations, and builds prediction models based on three benchmark dataset respectively. We conduct 5-fold cross validation to evaluate all models, and results are shown in Table II. We observe that NTSIM can produce better performances than PREDICT, TL-HGBI and LRSSL on the benchmark datasets.

Therefore, our method NTSIM is superior to the baseline method and three drug-disease association prediction methods in the computational experiments.

D. Independent Experiments

In this section, we conduct independent experiments to demonstrate the capability of our method for predicting novel drug-disease associations.

First of all, we respectively build our NTSIM models based on PREDICT dataset, TL-HGBI dataset and LRSSL dataset, and also build PREDICT models, TL-HGBI models and LRSSL models on their own datasets. Then, these models make predictions for other drug-disease pairs without known associations. To the best of our knowledge, CTD database is up-to-date source for the chemical-disease associations, and can help to validate predicted associations. The number of checked predictions and number of confirmed associations are visualized in Fig. 5, and results show that our method can find out more novel associations than benchmark methods. From PREDICT dataset, TL-HGBI dataset and LRSSL dataset, our method can approximately find out 100 novel associations, 60 novel associations and 240 novel associations in top 1000 predictions, respectively.

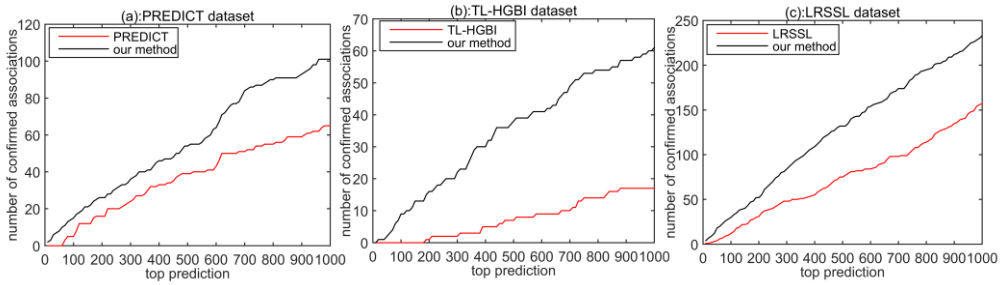

Fig.5. Number of associations confirmed by the latest CTD database in three benchmark datasets

E. Case Study

Here, we test the practical capability of our method for predicting unknown associations. We build our prediction model by using all drug-disease associations in CTD dataset, and then make predictions for other drug-disease pairs. Since all CTD associations have been used to build models, the predicted associations have to be validated by publicly available information sources, such as literatures and websites. The top 10 drug-disease associations predicted by our method are listed in Table III, and 6 novel drug-disease associations can be confirmed.

TABLE III. TOP 10 CHEMICAL-DISEASE ASSOCIATIONS PREDICTED BY OUR METHOD

NO.	Drugs	Diseases	Evidence
1	Methadone	Seizures	https://www.drugs.com/methadone.html
2	Amiodarone	Hypertension	http://factmed.com/drugcover.php?drugname=Amiodarone
3	Clozapine	Headache	https://www.drugs.com/clozapine.html
4	Morphine	Tremor	http://www.medindia.net/doctors/drug_information/morphine.htm
5	Methamphetamine	Hypotension	https://www.drugbank.ca/drugs/DB01577
6	Risperidone	Anxiety Disorders	N.A.
7	Amphetamine	Catalepsy	N.A.
8	Caffeine	Drug-Induced Liver Injury	N.A.
9	Chlorpromazine	Nausea	https://www.drugs.com/mtm/chlorpromazine.html
10	Clozapine	Sleep Initiation and Maintenance Disorders	N.A.

N.A. means that the predicted association cannot be confirmed.

Moreover, we respectively predict drugs that are associated with the disease “Nausea” (MeSH: D009325), and predict diseases that are associated with the drug “Risperidone” (MeSH: D018967). Nausea is a sensation of unease and discomfort in the upper stomach with an involuntary urge to vomit. Risperidone, sold under the trade name Risperdal among others, is an antipsychotic medication. As shown in Table IV, we can find evidences to support 6 drugs for the disease “Nausea” in top 10 predictions, and 4 diseases for the drug “Risperidone” can be confirmed in top 10 predictions.

TABLE IV. TOP 10 PREDICTIONS FOR THE DISEASE “NAUSEA” AND TOP 10 PREDICTIONS FOR THE DRUG “RISPERIDONE”

	Nausea		Risperidone	
NO.	Drug Name	Evidence	Disease Name	Evidence
1	Propranolol	https://www.drugs.com/propranolol.html	Anxiety Disorders	(none)
2	Amitriptyline	https://www.drugs.com/amitriptyline.html	Myoclonus	(none)
3	Chlorpromazine	https://www.drugs.com/mtm/chlorpromazine.html	Hypertension	[29]
4	Celecoxib	(none)	Arrhythmias, Cardiac	(none)
5	Trazodone	http://patientsville.com/trazodone/flatulence.htm	Pain	https://www.drugs.com/risperidone.html
6	Haloperidol	(none)	Chorea	(none)
7	Rifampin	https://www.drugs.com/mtm/rifampin.html	Substance-Related Disorders	(none)
8	Metoprolol	(none)	Inappropriate ADH Syndrome	(none)
9	Simvastatin	https://www.drugs.com/simvastatin.html	Hypothermia	[30]
10	Quinine	https://www.drugs.com/mtm/quinine.html	Xerostomia	https://www.drugs.com/risperidone.html

Therefore, the case study shows that the proposed method can help to identify novel drug-disease associations in the applications.

IV. Conclusion

In our work, we propose a novel method NTSIM to predict unobserved drug-disease associations by only using known drug-disease associations. We present the linear neighborhood similarity measure to calculate the drug-drug similarity and disease-disease similarity based on the association profiles, and combine drug-drug similarity and disease-disease similarity to predict unobserved associations. In the computational experiments, NTSIM can produce high-accuracy performances, and outperform other state-of-the-art methods. The proposed method is promising for predicting unobserved drug-disease associations.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61772381, 61572368) and The Fundamental Research Funds for the Central Universities (2042017kf0219).

References

[1] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Mol Syst Biol,* vol. 7, pp. 496, Jun 07, 2011.

[2] L. Yang, and P. Agarwal, "Systematic drug repositioning based on clinical side-effects," *PLoS One,* vol. 6, no. 12, pp. e28025, 2011.

[3] Y. Wang, S. Chen, N. Deng, and Y. Wang, "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data," *PLoS One,* vol. 8, no. 11, pp. e78518, 2013.

[4] Y. F. Huang, H. Y. Yeh, and V. W. Soo, "Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation," *BMC Med Genomics,* vol. 6 Suppl 3, pp. S4, 2013.

[5] M. Oh, J. Ahn, and Y. Yoon, "A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions," *PLoS One,* vol. 9, no. 10, pp. e111668, 2014.

[6] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics,* vol. 30, no. 20, pp. 2923-30, Oct 15, 2014.

[7] V. Martinez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: network-based drug-disease prioritization by integrating heterogeneous data," *Artif Intell Med,* vol. 63, no. 1, pp. 41-9, Jan, 2015.

[8] H. Wang, Q. Gu, J. Wei, Z. Cao, and Q. Liu, "Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies," *Clin Pharmacol Ther,* vol. 97, no. 5, pp. 451-4, May, 2015.

[9] H. Moghadam, M. Rahgozar, and S. Gharaghani, "Scoring multiple features to predict drug disease associations using information fusion and aggregation," *SAR QSAR Environ Res,* vol. 27, no. 8, pp. 609-28, Aug, 2016.

[10] X. Liang, P. Zhang, L. Yan, Y. Fu, F. Peng, L. Qu, M. Shao, Y. Chen, and Z. Chen, "LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning," *Bioinformatics,* vol. 33, no. 8, pp. 1187-1196, 2017.

[11] H. Chen, H. Zhang, Z. Zhang, Y. Cao, and W. Tang, "Network-based inference methods for drug repositioning," *Comput Math Methods Med,* vol. 2015, pp. 130620, 2015.

[12] D. Sun, A. Li, H. Feng, and M. Wang, "NTSMDA: prediction of miRNA-disease associations by integrating network topological similarity," *Mol Biosyst,* vol. 12, no. 7, pp. 2224-32, Jun 21, 2016.

[13] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics,* vol. 27, no. 21, pp. 3036-43, Nov 1, 2011.

[14] F. Cheng, W. Li, X. Wang, Y. Zhou, Z. Wu, J. Shen, and Y. Tang, "Adverse drug events: database construction and in silico prediction," *J Chem Inf Model,* vol. 53, no. 4, pp. 744-52, Apr 22, 2013.

[15] W. Zhang, H. Zou, L. Luo, Q. Liu, W. Wu, and W. Xiao, "Predicting potential side effects of drugs by recommender methods and ensemble learning," *Neurocomputing,* vol. 173, pp. 979-987, 2016.

[16] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Res,* vol. 37, no. Web Server issue, pp. W623-33, Jul, 2009.

[17] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res,* vol. 36, no. Database issue, pp. D901-6, Jan, 2008.

[18] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res,* vol. 38, no. Database issue, pp. D355-60, Jan, 2010.

[19] F. Wang, P. Zhang, N. Cao, J. Hu, and R. Sorrentino, "Exploring the associations between drug side-effects and therapeutic indications," *Journal of Biomedical Informatics,* vol. 51, pp. 15-23, 2014.

[20] W. Zhang, Y. Chen, S. Tu, F. Liu, and Q. Qu, "Drug side effect prediction through linear neighborhoods and multiple data source integration". Proc. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),IEEE Press, Dec, 2016 pp. 427-434.

[21] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC Bioinformatics,* vol. 18, no. 1, pp. 18, Jan 05, 2017.

[22] S. Roweis, and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, 2000.

[23] F. Wang, and C. Zhang, "Label propagation through linear neighborhoods," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 20, no. 1, pp. 55-67, 2008.

[24] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, and Y. Huang, "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLoS One,* vol. 8, no. 8, pp. e70204, 2013.

[25] X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Sci Rep,* vol. 5, pp. 11338, Jun 10, 2015.

[26] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics,* vol. 26, no. 13, pp. 1644-50, Jul 01, 2010.

[27] T. Zhou, Z. Kuscsik, J. G. Liu, M. Medo, J. R. Wakeling, and Y. C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proc Natl Acad Sci U S A,* vol. 107, no. 10, pp. 4511-5, Mar 09, 2010.

[28] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput Biol,* vol. 8, no. 5, pp. e1002503, 2012.

[29] S. R. Thomson, D. Bhattacharjee, B. C. Magazine, and S. M. Bhat, "Risperidone Induced Hypertension in a Young Female: A Case Report," *Advanced Science Letters,* vol. 23, no. 3, pp. 1980-1982, //, 2017.

[30] M. Razaq, and M. Samma, "A case of risperidone-induced hypothermia," *American Journal of Therapeutics,* vol. 11, no. 3, pp. 229-230, 2004.