Discovering Regularities from Traditional Chinese Medicine Prescriptions via Bipartite Embedding Model

Chunyang Ruan[1,2], **Jiangang Ma**[3], **Ye Wang**[4*], **Yanchun Zhang**[1,2,4], **Yun Yang**[5]
[1]Fudan University, Shanghai,China
[2]Zhejiang Lab, Hanghzou, China
[3]Federation University Australia, Melbourne, Australia
[4]Victoria University, Melbourne, Australia
[5]Longhua Hospital Shanghai University of Traditional Chinese Medicine, Shanghai, China
cyruan16@fudan.edu.cn, j.ma@federation.edu.au, ye.wang10@live.vu.edu.au,
yanchun.zhang@vu.edu.au, 20067225@qq.com

Abstract

Regularities analysis for prescriptions is a significant task for traditional Chinese medicine (TCM), both in inheritance of clinical experience and in improvement of clinical quality. Recently, many methods have been proposed for regularities discovery, but this task is challenging due to the quantity, sparsity and free-style of prescriptions. In this paper, we address the specific problem of regularities discovery and propose a graph embedding based framework for regularities discovery for massive prescriptions. We model this task as a relation-prediction in which the correlation of two herbs or of herb and symptom are incorporated to characterize the different relations. Specifically, we first establish a heterogeneous network with herbs and symptoms as its nodes. We develop a bipartite embedding model termed *HS2Vec* to detect regularities, which explores multiple relations of herb-herb, and herb-symptom based on the heterogeneous network. Experiments on four real-world datasets demonstrate that the proposed framework is very effective for regularities discovery.

1 Introduction

With the continual and rapid development of the distinctive methodology and approach for diagnosing and treating disease, traditional Chinese medicine (TCM) has been playing a vital role in health care for several thousand years, and is well-know because of its unique curative effect and less side effect for complicated diseases (e.g., *SCLC/NSCLC, HIV/AIDS*) [Wen *et al.*, 2018]. As the essence of TCM doctors' clinical experiences, prescriptions plays the most important role in TCM clinic. A TCM prescription mainly consists of patient's symptoms for disease and various kinds of herbs. Based on this, a mainly goal of discovering regularities for prescriptions is to investigate the complex correlations between herbs composition and corresponding symptoms. These regularities are beneficial to inheritance of clinical experience and clinical practices [Yao *et al.*, 2018; Yang *et al.*, 2017]. For example, incompatible herb-pair detection is important to reduce the accidents due to adverse drug reactions unexpectedly[Zhu *et al.*, 2018].

Despite its value and significance, discovering regularities for TCM prescriptions remains in its infancy due to the following challenges: **Massive prescriptions.** With rapid development on the TCM research, massive prescriptions are expressed and collected with handwritten or electronic version, which is a critical task in prescriptions mining. **Sparsity of prescription.** Because of the personalised views of TCM doctors and heavy workload, some herbs/symptoms would not be recorded in the clinical data if they were considered as having no value for diagnosis and treatment. The lack of digitalization and formalization of TCM prescriptions leads to the problems of sparsity. **Free-style.** Prescriptions are represented as natural language and in free-text format. Different from natural language sentences, TCM prescriptions have their own way to organize the herbs and symptoms, which are often put in a weakly ordered form. For example, the herb in the front of the prescription may be connected with the very last herb instead of the surrounding ones. Furthermore, because the various terms used in prescriptions, the co-occurrence of terms contains crucial semantic information. This is in opposition to the current "one target, one drug" approach for relations discovering in biomedicine[Olayan *et al.*, 2018]. In summary, to better detect regularities in prescriptions, it is highly desirable to develop methods that comprehensively consider the heterogeneous data and special semantics and jointly learn the relations of different objects.

Motivated by the aforementioned challenges, we study a heterogeneous graph embedding problem and propose a bipartite framework, which is termed *HS2Vec*. Due to the specific natural language pattern and sparsity, previous works based on natural language processing (NLP) performed bad in discovering regularities in prescriptions [Yao *et al.*, 2018; Wang *et al.*, 2017; Wan *et al.*, 2015]. To avoid this issue, we first establish a TCM heterogeneous network according to the co-occurrence relations among herbs and symptoms. Two objects (herbs/symptoms) are connected if they co-occur in some prescriptions. Considering the existing

*Contact Author

complex relations comprehensively among terms, we propose a bipartite embedding strategy to analyse and predict latent relations. Comparing with existing heterogeneous graph embedding models [Fu *et al.*, 2017; Huang and Mamoulis, 2017], the local-structure and global-semantics are considered simultaneously in our model. Inspired by graph embedding for social relationship extraction [Liu *et al.*, 2018; Tu *et al.*, 2017], we use unsupervised bipartite model with two autoencoders to cope with the node embedding based on structural and semantic relations respectively. Specifically, each autoencoder learns the distance between a input node and its recombination as the output of autoencoder based on local structure. Meanwhile, we learn the distance between two input nodes from two autoencoders based on semantic. By jointly optimizing them in the proposed unsupervised model, we can learn a robust representation of TCM network. Then, We utilize clustering and HCMNed prediction to determine the correlations, which can discover the herbs composition and corresponding symptoms. The proposed method *HS2Vec* is tested on four real-world TCM datasets and compared to state-of-the-art methods and to TCM doctors. The contributions of this paper are listed as follows:

- To our best knowledge, this is the first attempt to discover regularities in TCM prescriptions using heterogeneous graph embedding.
- We develop a bipartite embedding model based on autoencoder termed *HS2Vec* to preserve the structure and semantics of TCM graph, to complete the heterogeneous nodes embedding.
- We demonstrate the effectiveness of our proposed model via heterogeneous network mining tasks on four TCM datasets such as node clustering, linked prediction, and clinical tasks, outperforming the state-of-the-art.

2 Related Works

Knowledge discovering has become a hot topic in healthcare and biomedicine [Wang *et al.*, 2018]. More recent works focus on discovering relations among medical objects. [Luo *et al.*, 2018] presented a low-rank matrix approximation and randomized algorithms to predict drug–disease relations. [Fout *et al.*, 2017] proposed a neighborhood-based graph convolution methods to determine interfaces between proteins. [Zhao *et al.*, 2016] proposed a syntax drug embedding method and used convolutional neural network to detect relations among these drugs. Compared with knowledge discovery research in modern biomedicine, TCM regularities discovery just becomes popular in recent years because of the lack of digitalization and formalization [Yao *et al.*, 2018; Huang *et al.*, 2018].

A number of works have been devoted to detect herbs regularities. [Zhu *et al.*, 2018] incorporated two important herb attributes and their correlation to characterize the incompatible relations among herbs via matrix-factorization. [Yao *et al.*, 2018] proposed a novel topic model to characterize the generative process of prescriptions. [Li *et al.*, 2018] proposed a seq2seq framework enhanced with masking and coverage mechanism to generate prescriptions. [Ji *et al.*, 2017] developed a novel multi-content model based on latent dirichlet allocation to find out the pathogenesis based on the latent semantics analysis of symptoms and herbs. [Wang *et al.*, 2017] proposed a new asymmetric probabilistic model for the joint analysis of symptoms, diseases, and herbs in medical records to discover and extract latent TCM knowledge. [Chen *et al.*, 2018] presented a HIN-based clustering model to find the categories of formulas. [Wan *et al.*, 2015] proposed a novel approach to collectively and globally extract correlations from TCM literature based on HIN. Although these models are actually effective in TCM exploration, these models are limited to the traditional data mining methods or the characteristics of TCM data. Different from these existing NLP based models, our approach deals with prescriptions with multiple heterogeneous nodes and relations, and uses graph embedding to discover the regularities.

G(V,E)	the Graph G, nodes V and edges E for HCMN
τ	the node types of HCMN
ξ	the edge types of HCMN
ρ	a meta-path of HCMN
ϱ	the meta-path set of HCMN
K	the number of layers
C_k	a cluster
T= $\{\mathbf{t}_i\}_{i=1}^n$	the adjacency matrix of HCMN
$\mathbf{x}_i$, $\bar{\mathbf{x}}_i$	the initial data and re-encode data
$\mathbf{y}_i^{(k)}$	the k-th layer hidden representations of node
$\mathbf{W}^{(k)}$, $\bar{\mathbf{W}}^{(k)}$	the k-th layer weight matrix
$\mathbf{b}^{(k)}$, $\bar{\mathbf{b}}^{(k)}$	the k-th layer biases
ω	the set of weighted values of all node pair
d	the number of dimensions
$s(v_i, v_j)$	the proximity of two nodes v_i, v_j

Table 1: Notations and explanations

3 Problem Definitions and Preliminaries

In this section, we provide basic notations and formulate the problem of HCMN embedding. We first give the notation presented in Table 1.

Definition 1 HCMN. *HCMN is a heterogeneous network G=$\{V, E, \omega\}$, where $V = \{v_i\}_{i=1}^n$ represents the all herbs or symptoms of the co-occurrence network, $E = \{e_i\}_{i=1}^m$ denotes the all types edges among these nodes. ω is a set of weight values of all edges, $\omega_{ij} \geqq 0$.*

In HIN, two nodes can be connected via different semantic paths, which are called meta-paths [Fu *et al.*, 2017]. In HCMN, meta-path preserves the semantics and structure of TCM prescriptions.

Definition 2 TCM Meta-path. *A TCM meta-path ρ is defined as the path sequence of node types $\tau_1,\cdots,\tau_n$ connected by edge types $\xi_1,\cdots,\xi_{n-1}$ as follows:*

$$\rho = \tau_1 \xrightarrow{\xi_1} \tau_2 \cdots \tau_{n-1} \xrightarrow{\xi_{n-1}} \tau_n$$

An instance of the meta-path ρ is a path sequence in the HCMN, which conforms to the pattern of ϱ. For example, $herb \xrightarrow{cure} symptom \xrightarrow{cure^{-1}} herb$ is a meta-path in the HCMN, which defines the semantics that two herbs can cure a symptom.

Definition 3 TCM Meta-path based Proximity. *The TCM meta-path based proximity for each pair of nodes $v_i, v_j \in V$ is denoted as:*

$$s(v_i, v_j)=\sum_{\rho}^{|\varrho|} s(v_i, v_j \mid \rho)$$

where ρ is one meta-path of node pair (v_i, v_j); $|\varrho|$ is the number of meta-path between two nodes; and $s(v_i, v_j \mid \rho)$ is the proximity w.r.t. the meta-path ρ. Note that the proximity of two nodes equals the sum of the proximity w.r.t. all meta-paths. It's obvious that this can preserve all kinds of correlations between the two nodes.

Definition 4 TCM graph embedding. *Given the HCMN defined as G=$\{V, E\}$, our goal is to develop a mapping: $v_i \xrightarrow{f} y_i \in \mathbb{R}^d$, where $d \leqq |V|$, y_i is the dimensional latent representations of node v_i. The function f is to get the low-dimensional representations of all nodes. Note that, although there are various types of nodes in HCMN, their representations are mapped into the same latent space. The embedding vectors preserve the semantics in HCMN.*

4 Methodology

4.1 Overview

Given a dataset of graphs, *HS2Vec* considers preserve the whole information of HCMN in a low-dimensional embedding space. Different from general graph, HCMN has multi-types of nodes, and two nodes v_i, v_j may be connected via multi-types of paths. Conceptually, each path represents a specific direction or composite semantics between nodes. So, we use meta-path based proximity to obtain pairwise similarities of nodes. Meanwhile, we attempt to capture and reconstruct its neighborhood information for each node to preserve the local-order proximity. Accordingly, to optimize the proximity in the process of model learning, *HS2Vec* can preserve the highly-nonlinear local-sturcture and global semantics well and is robust to other graph. Getting the low-dimensional embedding space of herbs, symptoms, we can use clustering to detect the groups of herbs and symptoms and use linked prediction to obtain the correlations (i.e. herb-herb, herb-symptom). The remainder of this section, we will present a detailed introduction to *HS2Vec*. The framework of *HS2Vec* is shown in Figure 1.

4.2 PRM Proximity Calculation

According to *Definition 3*, in order to compute the meta-path based proximity that semantics between node v_i and v_j, we need to accumulate the corresponding meta-path based proximity w.r.t. each meta-path ρ. In this paper, we employ transition probability to define the proximity $s(v_i, v_j|\rho_{v_i \to v_j})$. Based on a meta-path ρ:$H \to S \to H$, a herb node v_1 may walk to any node of the next three symptom nodes via weighted edges. Then, one of three symptom nodes transports to a herb node v_2 via the only weighted edge. So, the transition probability from v_1 to v_2 is accumulation of probability $s(v_1, v_2|\rho_{v_1 \to v_2})$= 0.25. This approach is a random walk statistics based on meta-path instance ρ to estimate the transition probability of nodes. Now, we formulate the probability computation as follows:

$$s(v_i, v_j \mid \rho) = \begin{cases} \frac{\mu_{(v_i,v_j)}}{Z} & (v_i, v_j) \in E \\ 0 & otherwise \end{cases} \quad (1)$$

where $\mu_{(v_i,v_j)}$ is the unnormalized transition probability between nodes v_i and v_j, and Z is the normalizing constant. For the case of meta-path based proximity, we have the following property.

$$s_l^{'}(v_i, v_j) = \sum_{(v_i,v^{'}) \in E} \mu_{(v_i,v^{'})}^{\phi(v_i,v^{'})} \times s_{l-1}^{'}(v^{'}, v_j) \quad (2)$$

where $\mu_{(v_i,v^{'})}^{\phi(v_i,v^{'})}$ is the transition probability from v_i to $v^{'}$ w.r.t the edge type $\phi(v_i, v^{'})$. In details, N edges from v_i that belong to the type $\phi(v_i, v^{'})$, the $\phi(v_i, v^{'}) = 1/N$. Next, the **p**roximity based on **r**andom walk path, which is restricted by **m**eta-path instance ρ (PRM), can be defined as follows:

$$s(v_i, v_j) = \sum_{len(\rho) \leq l} s(v_i, v^{'}) \times s(v^{'}, v_j|\rho[2:l]) \quad (3)$$

where l is a length threshold, $\rho[i : j]$ is the subsequence of path instance ρ from the node v_i to the node v_j, and $\rho[2 : l]$ is a length $l-1$ path instance. Summing over all the length-l path instances for Eq.(4), we get the final meta-path based proximity, which is a dynamic programming approach to calculate all proximity. Note that, shorter meta-paths are more informative than longer ones, because longer meta-paths edge more remote objects, which are less related semantically [Huang and Mamoulis, 2017]. Accordingly, we set the length threshold to enhance the estimation of proximity.

4.3 Objective Functions

We introduce the objective functions of *HS2Vec*. We first describe how the *HS2Vec* preserves the second-order proximity that structre. The second-order proximity describes the similarity of neighborhood structures between two nodes. Obviously, the higher second-order proximity denotes that two nodes share more common neighbors for they are more similar. Accordingly, we require our autoencoder to maintain all neighborhoods of each node. For HCMN, the network structures and semantics including global and local information can be described by the adjacency matrix *T*. For example, in *T*, the *i-th* row, $\boldsymbol{t}_i$=$\{x_{ij}\}_{j=1}^{n}$, gives the first-order proximity between nodes v_i and v_j.

The autoencoder attempts to capture the second-order proximity via reconstructing the input data $\boldsymbol{x}_i$. In detail, an autoencoder is composed of two parts, encoder and decoder. The encoder, consisting of multiple non-linear activation functions, maps the $\boldsymbol{x}_i$ into the latent representation space. The decoder is similar to a reversed encoder, reconstructing latent representation to reconstructed input space. Given the initial $\boldsymbol{x}_i$, the hidden representations for each layer are shown as follows:

$$\begin{aligned} \mathbf{y}_i^{(1)} &= \delta(\mathbf{W}^{(1)} \times \mathbf{x}_i + \mathbf{b}^{(1)}) \\ \mathbf{y}_i^{(k)} &= \delta(\mathbf{W}^{(k)} \times \mathbf{y}_i^{k-1} + \mathbf{b}^{(k)}) \\ \bar{\mathbf{x}}_i &= \delta(\bar{\mathbf{W}}^{(k)} \times \bar{\mathbf{y}}_i^{k-1} + \bar{\mathbf{b}}^{(k)}) \end{aligned} \quad (4)$$

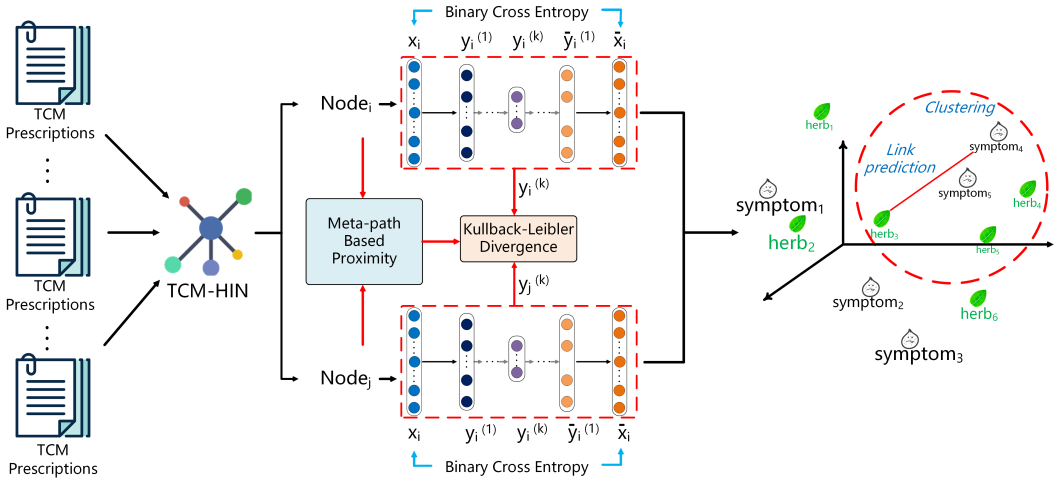


Figure 1: The framework of *HS2Ve* for regularities discovery in prescriptions.

where δ is the sigmoid function. Using the generalized $\boldsymbol{y}_i^{(k)}$, we can obtain the output $\bar{\boldsymbol{x}}_i$ via reversing the calculation process of the encoder. The goal of the autoencoder is to play a *mini-game* that minimizes the reconstruction error of the output and the input. In addition, some nodes contain a small number of edges in HCMN, which results in the sparsity disaster of HCMN. That is because the number of zero elements of the adjacency matrix T is much more than that of non-zeroin elements, which may degrade the performance of reconstruction. The autoencoder component would be inclined to reconstruct more zero elements to output $\bar{\boldsymbol{x}}_i$. Therefore, the weighted *Binary Cross Entropy* as loss function is employed to impose more penalty on the reconstruction error of the non-zero elements than zero elements. For all types of nodes, the second-proximity objective function can be defined as follows:

$$J_1 = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{ij}[x_{ij}\log(\bar{\mathbf{x}}_{ij}) + (1-x_{ij})\log(1-\bar{\mathbf{x}}_{ij})] \tag{5}$$

where λ_{ij} indicates the weight coefficient of the penalty imposed to each elements. If $x_{ij} = 0$, $\lambda_{ij} = 1$, else $\lambda_{ij} = \omega_{ij} > 1$. With the objective function J_1, the latent space can maintain the global structures. It is not only necessary to preserve the global network structure, but also essential to capture the local structure [Huang and Mamoulis, 2017]. The first-order proximity is the general approach to preserve the local structure. However, because the multi-types of nodes and edges have their own special characteristics, we utilize PRM proximity to capture the local information, which can learn unique latent spaces for different node and edge types. To preserve the *PRM* proximity, a distinct goal is to minimize the distance of these two probability distributions between $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$. In our work, we use *Kullback-Leibler* divergence as the distance metric. Then the objective function for this goal is defined as follows:

$$J_2 = -\sum_{v_i,v_j\in V} s(v_i,v_j)\log\delta(\mathbf{y}_i^{(k)}\cdot\mathbf{y}_j^{(k)}) \tag{6}$$

where δ is the sigmoid function. To preserve both global and local proximity of HCMN, we jointly minimize the objective function by training the Eq.(5) and Eq.(6) simultaneously as follows:

$$J = (1-\alpha)J_1 + \alpha J_2 \tag{7}$$

We utilize the asynchronous stochastic gradient descent (*ASGD*) algorithm [Recht *et al.*, 2011] to optimize *HS2Vec*. In detail, we aim to calculate the partial differential function, $\partial J/\partial\bar{\mathbf{W}}^{(k)}$ and $\partial J/\partial\bar{\mathbf{W}}^{(k)}$. In addition, to address the problem that meta-path based algorithm may converge to a trivial solution [Huang and Mamoulis, 2017], we sample multiple negative nodes to enhance the influence of positive nodes. For each pair of nodes with non-zero path based PRM proximity $s(v_i, v_j)$, we redefine object function J_2 as follows:

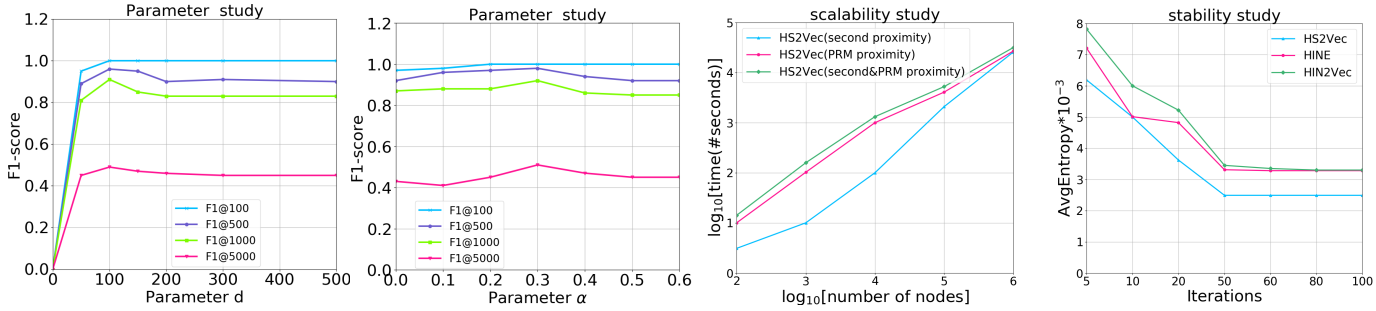
$$J_2 = -\log(1+e^{-\mathbf{y}_i^{(k)}\mathbf{y}_j^{(k)}}) - \sum_{1}^{m}\mathbb{E}_{\hat{v}\in Pr(v_i)}[\log(1+e^{\mathbf{y}_i^{(k)}\hat{v}})] \tag{8}$$

where m is the times of sampling, and $Pr(v)$ is some noise distribution of node v_i.

5 Experiments

5.1 Datasets

- TRE [Wan *et al.*, 2015]: The dataset integrates herbs, symptoms, diseases and their correlations from the Chinese TCM texts.
- TCMSP [Ru *et al.*, 2014]: The dataset describes a pharmacology information TCM, which includes herbs, diseases, chemicals, targets and their correlations.
- TCMGeDIT [Fang *et al.*, 2008]: The dataset provides association information about genes, diseases, TCM effects and TCM ingredients automatically mined from vast amount of biomedical literature.
- The clinical dataset is collected from over 70, 000 lung cancer records, including herbs, symptoms and diseases.



(a) Parameter d. (b) Parameter α. (c) Change of scalability. (d) Change of stability.

Figure 2: Performance evaluation of parameter sensitivity of the embedding dimension d and the penalty coefficient α, scalability and stability.

5.2 Baseline Methods

To evaluate the performance of *HS2Vec* for testing tasks, we compare the following baselines:

- CPM [Wang *et al.*, 2017]: It proposes a novel probabilistic model to capture the asymmetric causal correlations of symptoms, diseases, and herbs.
- TM [Yao *et al.*, 2018]: It designs a novel topic model to characterizes the generative process of prescriptions .
- HIN2Vec [Fu *et al.*, 2017] captures the rich semantics embedded in HIN by exploiting different types of correlation among nodes based on a neural network model.
- HINE [Huang and Mamoulis, 2017] calculates the distance of probability distributions via meta-path and Kullback–Leibler divergence to embed HIN.
- Motif-CNN [Sankar *et al.*, 2018] utilizes convolutional neural network with attention and motif to learn heterogeneous network representations.

5.3 Quantitative Evaluation

For the primary purpose of herb-herb and herb-symptom relations discovery, we first compare the effectiveness of different heterogeneous graph embedding methods at a task of herbs and symptoms clustering, which aims to identify candidate sets for relations discovery. In addition, we use the normalized mutual information (NMI) to evaluate the performance of each method. Finally, in order to ensure that *HS2Vec* can accurately detect the strong herb-herb, herb-symptom correlations, we use the mean average precision (MAP) to evaluate *HS2Vec* for linked prediction. We report the results of NMI and MAP in Table 2.

For clustering, all the embedding methods perform bad on TCMGeDIT, but they have a better performance on the other two datasets. On TCMGeDIT, our model *HS2Vec* gets the best performance than others, and *Motif-CNN* achieves relatively well results. On TCMSP, *HS2Vec* has a better performance than others. On TRE, *HS2Vec* outperforms all other methods. Overall, we can observe that *HS2Vec* outperforms all the other methods in the task of clustering. For the herb-symptom linked prediction, we can observe that the performance of HIN2Vec and HINE is relatively poor in linked prediction. On all the datasets, graph convolutional neural network based embedding methods Motif-CNN obtain better performance than HIN2Vec and HINE, which is very close to that of our proposed model. This is probably because deep neural network can extract the feature of network effectively.

As a result, *HS2Vec* achieves significant improvements over the baselines. It demonstrates that the learned embeddings of *HS2Vec* have strong predictive power for linked prediction.

5.4 Case Evaluation

For the clinical task to generate the effective prescriptions, we first obtain clusters including herbs and symptoms for specific disease conditions or particular patient groups based on the node embedding. In each cluster, we use linked prediction to detect the strong correlation between herbs and symptoms. We now evaluate correlations learned from over 5, 000 lung cancer records via *HS2Vec*. The Micro-F1 and Macro-F1 are employed as the evaluation metrics. CPM, TM, HIN2Vec and HINE were employed as the comparative models. Table 3 shows the results comparison of herb recommendation based on symptoms. We can see that *HS2Vec* achieves high Micro-F1 and Macro-F1 scores, and all other approaches have quite good performance. To show this in detail, recommended prescriptions via *HS2Vec* based on some symptoms of lung cancer were vertified by two experienced TCM doctors. Table 4 gives the comparations of prescription recommendation. We note that *HS2Vec* prescribed sixteen herbs are in common with the herbs prescribed by the TCM doctors, which are marked by bold in this table. *HS2Vec* also recommended five herbs not prescribed by the doctor. A doctor verified that these some of them are all known to be associated with lung tumor. For example, “Platycodon grandiflorus” and “Paris polyphylla” have the similar function.

5.5 Parameter Sensitivity

In this section, we discuss the influence of the parameters dimension d of the latent space and the penalty coefficient α in the Eq.(7). We conduct experiments on a clustering task.

The parameter d measures the appropriate number of embedding dimensions. Figure 2(a) shows the performances of our method w.r.t the varying d. We can observe that the

Model	*NMI*			*MAP*		
	TCMSP	TCMGeDIT	TRE	TCMSP	TCMGeDIT	TRE
HIN2Vec	0.725	0.533	0.757	0.823	0.740	0.796
HINE	0.685	0.530	0.707	0.816	0.742	0.782
Motif-CNN	0.734	0.557	0.706	0.867	0.749	0.846
HS2Vec	**0.735**	**0.569**	**0.710**	**0.898**	**0.753**	**0.849**

Table 2: Performances of node clustering and linked prediction

Training	Metric	Approach				
		HS2Vec	CPM	TM	HIN2Vec	HINE
20%	Micro-F1	**0.362**	0.321 *-12.7%*	0.296 *-22.2%*	0.340 *-6.5%*	0.357 *-1.4%*
	Macro-F1	**0.235**	0.189 *-23.4%*	0.164 *-43.3%*	0.201 *-16.9%*	0.215 *-9.3%*
50%	Micro-F1	**0.511**	0.461 *-10.8%*	0.439 *-16.5%*	0.479 *-6.7%*	0.491 *-4.1%*
	Macro-F1	**0.453**	0.441 *-2.7%*	0.395 *-14.7%*	0.448 *-1.1%*	0.451 *-0.4%*
70%	Micro-F1	**0.693**	0.645 *-7.4%*	0.624 *-11.1%*	0.668 *-3.7%*	0.672 *-3.1%*
	Macro-F1	**0.521**	0.466 *-11.8%*	0.462 *-12.8%*	0.494 *-5.5%*	0.502 *-3.8%*
90%	Micro-F1	**0.813**	0.776 *-3.5%*	0.759 *-7.1%*	0.785 *-3.6%*	0.791 *-2.8%*
	Macro-F1	**0.682**	0.605 *-12.7%*	0.593 *-14.7%*	0.626 *-8.9%*	0.632 *-7.9%*

Table 3: Results of effective herbs recommendation according to symptoms on clinical dataset (Negative percent indicate the ratio decrease, comparing with the highest score of *HS2Vec* in the column)

Symptoms	Prescriptions	
	HS2Vec	***TCM Doctors***
1.lack of strength, 2.insomnia, 3.night sweats, 4.dry mouth, 5. bitter taste, 6.red tongue, 7.white fur, 8.a fine and bowstring pulse	Herba houttuyniae, **Scutellaria baicalensis**, **Salvia chinensis**, **Radix glehniae** **Ligustrum lucidum**, **Dwarf lilyturf**, Pericarpium citrus reticulata, **Herba selaginellae doederleinii**, **Astragalus mongholicus**, **Asparagus**, **Processed rhizoma pinelliae**, Golden thread, Platycodon grandiflorus, **Light wheat**, **Fruit of Chinese wolfberry**, **Chinese yam**, Chicken's gizzard-membrane, **Calcined oyster shell**, **Asarum**, **Dark plum**, **Tuber fleeceflower stem**, **Calcined keel**	**Astragalus mongholicus**, **Salvia chinensis**, **Herba Selaginellae Doederleinii** **Asparagus**, **Dwarf lilyturf**, **Radix glehniae**, Paris polyphylla, Epimedium, **Chinese yam**, **Asarum**, **Dark plum**, **Ligustrum lucidum**, **Scutellaria baicalensis**, **Calcined oyster shell**, **Calcined keel**, **Fruit of Chinese wolfberry**, **Processed rhizoma pinelliae**, **Light wheat**, **Tuber fleeceflower stem**, Raw atractylodes
Precision=0.663	**Recall**=0.893	**F1**=0.692

Table 4: The difference and intersection herbs prescribed by *HS2Vec* and TCM doctors according to clinical symptoms of lung tumor

increasing value of dimension improves the performance of *HS2Vec* in the early. Intuitively, this is because more vectors can encode more useful information in latent space. However, when the value of the dimension over 100, the performance began to drop slowly. The result suggests that excessive dimension values bring noises, which impacts on the performance. Generally, determining the value of the dimension is very important to graph embedding, but *HS2Vec* is less sensitive to d. The parameter α measures a well-balanced point between PRM proximity and Figure 2(b) shows how the value of α affects the performance of *HS2Vec*. When the value of α equals 0, only the second-order proximity is to determine the performance. The performance with α between 0.1 and 0.3 is better than α equaling 0, demonstrating the importance of PRM proximity. When the value is over 0.4, the performance became stable. In summary, both the PRM proximity and the second-order proximity are necessary for *HS2Vec* to maintain network information.

5.6 Scalability and Stability

To testify the scalability, we test node representations with default parameter for TCMSP with increasing number from 100 to 1, 000, 000 nodes. Figure 2(c) shows the test time w.r.t the number of nodes. We can observe that the test time scales linearly with the increased number of nodes. Meanwhile, we examine the stability of *HS2Vec* on the TRE dataset via metric AvgEntropy[Shi *et al.*, 2014]. Figure 2(d) shows the comparison of AvgEntropy w.r.t each iteration on different types of nodes, herbs and symptoms. We can see that *HS2Vec* performs better than HINE and HIN2Vec in all conditions. The reason is that *HS2Vec* effectively preserves more information from the network. Overall, these results show that *HS2Vec* is quite scalable and steady.

6 Conclusion

We proposed a novel bipartite model *HS2Vec* to investigate the relations among herbs and symptoms via graph embedding. *HS2Vec* consists of a local-structure preserving component and a global-semantics preserving component. Using autoencoder with Binary Cross Entropy, we obtain heterogeneous nodes that herbs and symptoms embedding. To better leverage available semantics, we proposed the PRM algorithm to capture semantics among nodes. We then employed Kullback-Leibler divergence to obtain nodes embedding based on semantics. We obtain the node-embedding space from above two embedding approachs, and discover regularities in TCM prescriptions via clustering and linked prediction. Extensive experiments demonstrated the superiority of *HS2Vec*. TCM empirical evaluations also confirmed that *HS2Vec* is helpful for TCM clinical development.

Acknowledgments

We thank the reviewers for their careful consideration and helpful comments. This work was supported by the Nationla Natural Science Foundation of China (No.61672161), Youth Research Fund of Shanghai municipal health and Family Planning Commission (No.2015Y0195).

References

[Chen *et al.*, 2018] Xintian Chen, Chunyang Ruan, Yanchun Zhang, and Huijuan Chen. Heterogeneous information network based clustering for categorizations of traditional chinese medicine formula. In *BIBM*, pages 839–846, 2018.

[Fang *et al.*, 2008] Yu Ching Fang, Hsuan Cheng Huang, Hsin Hsi Chen, and Hsueh Fen Juan. Tcmgenedit: a database for associated traditional chinese medicine, gene and disease information using text mining. *Bmc Complementary & Alternative Medicine*, 8(1):58–58, 2008.

[Fout *et al.*, 2017] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. In *NIPS*, pages 6533–6542, 2017.

[Fu *et al.*, 2017] Tao-Yang Fu, Wang-Chien Lee, and Zhen Lei. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *CIKM*, pages 1797–1806, 2017.

[Huang and Mamoulis, 2017] Zhipeng Huang and Nikos Mamoulis. Heterogeneous information network embedding for meta path based proximity. *arXiv*, abs/1701.05291, 2017.

[Huang *et al.*, 2018] Mengxing Huang, Huirui Han, Hao Wang, Lefei Li, Yu Zhang, and Uzair Aslam Bhatti. A clinical decision support framework for heterogeneous data sources. *IEEE J. Biomedical and Health Informatics*, 22(6):1824–1833, 2018.

[Ji *et al.*, 2017] Wendi Ji, Ying Zhang, Xiaoling Wang, and Yiping Zhou. Latent semantic diagnosis in traditional chinese medicine. *World Wide Web-internet & Web Information Systems*, (3):1–17, 2017.

[Li *et al.*, 2018] Wei Li, Zheng Yang, and Xu Sun. Exploration on generating traditional chinese medicine prescription from symptoms with an end-to-end method. *arXiv*, abs/1801.09030, 2018.

[Liu *et al.*, 2018] Jie Liu, Zhicheng He, and Yalou Huang. Hashtag2vec: Learning hashtag representation with relational hierarchical embedding model. In *IJCAI*, pages 3456–3462, 2018.

[Luo *et al.*, 2018] Huimin Luo, Min Li, Shaokai Wang, Quan Liu, Yaohang Li, and Jianxin Wang. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, 34(11):1904–1912, 2018.

[Olayan *et al.*, 2018] Rawan S. Olayan, Haitham Ashoor, and Vladimir B. Bajic. DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. *Bioinformatics*, 34(7):1164–1173, 2018.

[Recht *et al.*, 2011] Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pages 693–701, 2011.

[Ru *et al.*, 2014] Jinlong Ru, Li Peng, Jinan Wang, Zhou Wei, Bohui Li, Huang Chao, Pidong Li, Zihu Guo, Weiyang Tao, and Yinfeng Yang. Tcmsp: a database of systems pharmacology for drug discovery from herbal medicines. *J Cheminform*, 6(1):13, 2014.

[Sankar *et al.*, 2018] Aravind Sankar, Xinyang Zhang, and Kevin Chen-Chuan Chang. Motif-based convolutional neural network on graphs. *arXiv*, abs/1711.05697, 2018.

[Shi *et al.*, 2014] Chuan Shi, Ran Wang, Yitong Li, Philip S. Yu, and Bin Wu. Ranking-based clustering on general heterogeneous information networks by network projection. In *CIKM*, pages 699–708, 2014.

[Tu *et al.*, 2017] Cunchao Tu, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. Transnet: Translation-based network representation learning for social relation extraction. In *IJCAI*, pages 2864–2870, 2017.

[Wan *et al.*, 2015] Huaiyu Wan, Marie Francine Moens, Walter Luyten, Xuezhong Zhou, Qiaozhu Mei, Lu Liu, and Jie Tang. Extracting relations from traditional chinese medicine literature via heterogeneous entity networks. *Journal of the American Medical Informatics Association Jamia*, 23(2):356, 2015.

[Wang *et al.*, 2017] Sheng Wang, Edward W Huang, Runshun Zhang, Xiaoping Zhang, Baoyan Liu, Xuezhong Zhou, and Cheng Xiang Zhai. A conditional probabilistic model for joint analysis of symptoms, diseases, and herbs in traditional chinese medicine patient records. In *BIBM*, pages 411–418, 2017.

[Wang *et al.*, 2018] Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Personalized prescription for comorbidity. In *DASFAA*, pages 3–19, 2018.

[Wen *et al.*, 2018] Li Wen, Ye Fang Liu, Cen Jiang, Shao Qian Zeng, Yue Su, Wen Jun Wu, Xi Yang Liu, Jian Wang, Ying Liu, and Chen Su. Comparative proteomic profiling and biomarker identification of traditional chinese medicine-based hiv/aids syndromes. *Scientific Reports*, 8(1), 2018.

[Yang *et al.*, 2017] K. Yang, R. Zhang, L. He, Y. Li, W. Liu, C. Yu, Y. Zhang, X. Li, Y. Liu, and W. Xu. Multistage analysis method for detection of effective herb prescription from clinical data. *Frontiers of Medicine*, (7):1–12, 2017.

[Yao *et al.*, 2018] Liang Yao, Yin Zhang, Baogang Wei, Wenjin Zhang, and Zhe Jin. A topic modeling approach for traditional chinese medicine prescriptions. *IEEE Transactions on Knowledge and Data Engineering*, PP(99):1–1, 2018.

[Zhao *et al.*, 2016] Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453, 2016.

[Zhu *et al.*, 2018] Jiajing Zhu, Yongguo Liu, Shangming Yang, Shuangqing Zhai, Zhang Yi, and Chuanbiao Wen. A supervised learning framework for prediction of incompatible herb pair in traditional chinese medicine. In *CIKM*, pages 1799–1802, 2018.