Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

# Heterogeneous graph construction and node representation learning method of *Treatise on Febrile Diseases* based on graph convolutional network

YAN Junfeng[a], WEN Zhihua[a, b], ZOU Beiji[a, c*]

*a. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China*

*b. School of Computer Science and Engineering, Hunan University of Technology, Zhuzhou, Hunan 412008, China*

*c. School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China*

ARTICLE INFO

*Article history*
Received 31 October 2022
Accepted 30 November 2022
Available online 25 December 2022

*Keywords*
Graph convolutional network (GCN)
Heterogeneous graph
*Treatise on Febrile Diseases* (*Shang Han Lun*, 《伤寒论》)
Node representations on heterogeneous graph
Node representation learning

ABSTRACT

**Objective** To construct symptom-formula-herb heterogeneous graphs structured *Treatise on Febrile Diseases* (*Shang Han Lun*, 《伤寒论》) dataset and explore an optimal learning method represented with node attributes based on graph convolutional network (GCN).

**Methods** Clauses that contain symptoms, formulas, and herbs were abstracted from *Treatise on Febrile Diseases* to construct symptom-formula-herb heterogeneous graphs, which were used to propose a node representation learning method based on GCN – the Traditional Chinese Medicine Graph Convolution Network (TCM-GCN). The symptom-formula, symptom-herb, and formula-herb heterogeneous graphs were processed with the TCM-GCN to realize high-order propagating message passing and neighbor aggregation to obtain new node representation attributes, and thus acquiring the nodes' sum-aggregations of symptoms, formulas, and herbs to lay a foundation for the downstream tasks of the prediction models.

**Results** Comparisons among the node representations with multi-hot encoding, non-fusion encoding, and fusion encoding showed that the Precision@10, Recall@10, and F1-score@10 of the fusion encoding were 9.77%, 6.65%, and 8.30%, respectively, higher than those of the non-fusion encoding in the prediction studies of the model.

**Conclusion** Node representations by fusion encoding achieved comparatively ideal results, indicating the TCM-GCN is effective in realizing node-level representations of heterogeneous graph structured *Treatise on Febrile Diseases* dataset and is able to elevate the performance of the downstream tasks of the diagnosis model.

## 1 Introduction

*Treatise on Febrile Diseases* (*Shang Han Lun*, 《伤寒论》) is the first clinical work on medical theories, treatments, prescriptions, and herbs that integrates theories with practices in China, a crowning achievement in the history of traditional Chinese medicine (TCM) development and a classic book with great value. Therefore, a TCM diagnosis-assisting intelligent system based on *Treatise on Febrile Diseases* is also of important significance. The key to the TCM diagnosis-assisting intelligent system is the efficient expressiveness of TCM knowledge, such as *Treatise on Febrile Diseases*, whose attributes that include complicated semantics have created barriers for the

*Corresponding author: ZOU Beiji, Professor, E-mail: bjzou@csu.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2022.12.007

**Citation:** YAN JF, WEN ZH, ZOU BJ. Heterogeneous graph construction and node representation learning method of *Treatise on Febrile Diseases* based on graph convolutional network. Digital Chinese Medicine, 2022, 5(4): 419-428.

knowledge to be understood.

Knowledge representation is widely applied in TCM field. Normally, experts tend to adopt rules-based technology of knowledge representation to systematically study *Treatise on Febrile Diseases* [1, 2]. As machine learning becomes a trending topic, natural language processing (NLP) has been gradually employed in the studies of the expressiveness of *Treatise on Febrile Diseases* and the like TCM knowledge [3-5]. Over the past few years, knowledge graph has also been used in TCM expressiveness studies [6-9]. Graph convolutional network (GCN) is a graph-based deep learning method with powerful capacities fitting for node semantic representation learning, assisted with complicated graph networks. The graph neural network (GNN) or GCN has thus started its journey in herb recommendation because of its powerful graph representation capacity [10-12]. Syndrome-aware multi-graph convolution network (SMGCN) [11] model, after being fused with related symptom embeddings, could generate a generalized TCM syndrome representation system. With this model, the process of physicians revealing TCM syndromes could be simulated to construct a symptom-symptom graph for capturing the cooperative relations among the symptoms, along with the symptom-herb graph, has been taken to build a GCN for symptom embedding learning and a herb recommendation system, with ideal outcomes having being yielded. Inspired by the SMGCN model and how it interacts with symptom and herb node representations, we introduced GCN to address the expressiveness issues of *Treatise on Febrile Diseases*. The clauses in *Treatise on Febrile Diseases* include symptoms, formulas, and herb entities. All selected to construct a complicated GCN in which the symptoms are in the match with formulas, so as to transform the murky expressiveness of *Treatise on Febrile Diseases* into graph representations. A classic heterogeneous graph is made up of nodes with their own attributes, i.e. containing different symptoms, formulas, and herbs messages.

In the study, a TCM-GCN model was proposed based on the original GCN for the representation learning of the heterogeneous graph structured *Treatise on Febrile Diseases* message, including symptom-formula, symptom-herb, and formula-herb graphs, and for integrating corresponding node representation to acquire new node attributes that contain plenty of structured messages. The model was tested using a *Treatise on Febrile Diseases* based dataset, which yielded comparatively ideal results and laid a foundation for the architecture of a TCM diagnosis and treatment intelligent system.

## 2 Related work

### 2.1 Graph representations

A graph representation method is to transform the data in a graph into low-dimensional vectors to get prepared for machine learning [13]. The method at the same time can be used to better analyze the relationships among nodes in a complex network [14]. GCN is primarily used to do graph data convolutions using spectral or space methods. In terms of the spectral method, it is normally used to define the convolutions of the graph structured data in a spectral domain, and then transfer the defined data to a space domain [15]. GCN is a typical example of using the spectral method. For the space method, the convolutions are defined in a space directly. The general idea for the method is to aggregate the message of neighboring nodes to update the central node representation [16]. Graph attention network (GAT) is a classic example of the application of this method.

Currently, only a handful of studies have been carried out on learning representations of *Treatise on Febrile Diseases* based on graphs, but some have been done on knowledge representation in medication [17-20] and TCM herb recommendations [10-12]. In these studies, researchers have represented and learned the symptoms and herbs in a case record using GNN to target tasks on herb or drug recommendation. Such as the SMGCN model, which constructed a symptom-symptom graph for capturing the cooperative relations between symptoms, and was used along with a symptom-herb graph to build a new GCN for learning symptom embeddings, a significant improvement in herb recommendations compared with traditional approaches [11]. Multi-graph convolutional network (MGCN), another herb recommendation model based on multi-graph convolutions, was composed of pooled modules and herb prediction models [12]. MGN applied state elements and the central node of syndromes to achieve treatments based on syndrome. However, the model was limited in its use because the state elements and syndrome data were hard to obtain. Syndrome-aware KG-enhanced attentive multi-graph neural network (KG-ASMGNN) included TCM knowledge schematics to enrich the inputted corpora, which raised the quality of learning representations [21]. Knowledge-driven herb recommendation (KDHR) was an approach driving herb recommendation with the use of multi-layer message fusion based on GCN [22]. This model, assisted by the addition of properties of herbs, ideally represented the characteristics of the herbs. To summarize, graph representation learning can be used to obtain the structured potential information in TCM text as well as its full semantics. Therefore, high-quality vector representations of *Treatise on Febrile Diseases* text could be realized via graph learning, thus significantly improving the performance of the downstream tasks.

### 2.2 GCN techniques

GCN is a multi-layer neural network that is fit for processing the non-euclidean structure data via capturing their attributes at node level, graph level, and through edge prediction. The key to processing graph structured data is how to express the non-euclidean ones. We therefore would like to introduce Graph Embedding, the purpose of which is to transform each node in the given graph into a low-dimensional vector representation, normally realized by approaches such as Deepwalk or Node2vec. Such a vector representation is also called Node Embedding. As neural network technologies grow, GCN, GAT, GraphSAGE, PinSage, and the like methods have become the mainstream for node embedding representation learning [15, 16, 23-25].

For a multi-layer GCN, Equation (1) shows the message passing rule among the middle layers.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

Among which $\tilde{A} = A + I$ represents the adjacency matrix A plus the identity matrix I in graph G; $\tilde{D}$ is the degree matrix of $\tilde{A}$, $H^{(l)}$ is the attribute matrix in layer $l$, $W^{(l)}$ is the weighted matrix also in layer $l$, $\sigma(\cdot)$ is the activation function, $H^{(l+1)}$ is the updated attribute matrix in layer $(l+1)$. The procedure of GCN propagation is as follows.

(i) Message passing among node attributes according to $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}$.

(ii) Aggregating every node using $\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$, i.e. a linear transformation and activation.

(iii) Realizing multi-layer GCN by repeating steps (i) and (ii) $L$ times, and obtaining the final $H^{(L)}$ as the final node representation, which can be sent to the downstream tasks.

GCN has a powerful fitting capacity, which can also rise to gain a filter capacity of a higher-order polynomial frequency response function. GCN has simplified the learning ability of each layer and improved the expressiveness of each element via deep learning, which possesses engineering advantages. Therefore, GCN based models have become researchers' first choices when it comes to performing learning tasks using graphs.

## 3 Methods

### 3.1 Heterogeneous graph architecture of *Treatise on Febrile Diseases*

**3.1.1 Definition of graph representation** Each clause in *Treatise on Febrile Diseases* is a medical case record, such as Clause 12, saying "the pulse of initial Yang febrile disease caused by wind is floating when felts at the surface and weak in depth. Floating at the surface signifies heat. Weak in depth signifies spontaneous perspiration. Prescribe Guizhi Tang (桂枝汤) when the patients feels chill and fears, uneasy because of a fever, nauseous, and with a tendency to snore, etc.". These messages apparently offer the results of the four diagnostic methods, which are a slight chill feeling, aversion to wind, fever, running nose, nausea, and pulse floating at the surface. To treat these symptoms, the author of *Treatise on Febrile Diseases* ZHANG Zhongjing applied the Guizhi Tang, a treatment with significant outcomes (Table 1).

In the intelligent model for syndrome differentiation of the six meridians based on *Treatise on Febrile Diseases*, the clauses in the book were deemed as medical case records, and each clause is one medical case record that includes several symptoms, one formula, and some herbs. So each clause contains several sets of the three elements (symptoms, formulas, and herbs) to form the connectivity among them. Edge was used to link such connectivity and construct symptom-formula-herb heterogeneous graphs. All the clauses in *Treatise on Febrile Diseases* were considered as a set of TCM cases, which were all used to construct the complicated symptom-formula-herb heterogeneous graphs, as shown in Figure 1 (because the edges are numerous, therefore only parts of them that connect symptoms, formulas, and herbs were drawn).

Then the text information was transformed into standard symptom, formula, and herb sets, with each node representing one element: a symptom, a formula or a herb. Edges were used to connect the nodes and construct a graph representation based on these case records for the next learning and prediction, ultimately obtaining a prediction model for TCM diagnosis and treatment. The prediction model was subject to learning from the TCM case records, i.e. inputting all symptoms, formulas, and herbs in the heterogeneous graphs, and thus outputting predicted diagnosis results.

(i) Definition 1

Definition $S = \{s_1, s_2, \cdots, s_M\}$, $S$ denotes for all symptom sets, and $M$ is the set size.

Defining $F = \{f_1, f_2, \cdots, f_N\}$, $F$ stands for all formula sets, and $N$ is the set size.

**Table 1** Clause 12 in *Treatise on Febrile Diseases*

| Symptom | Syndrome | Formula | Herb |
|---|---|---|---|
| The pulse of initial Yang febrile disease caused by wind is floating when felts at the surface and weak in depth. Floating at the surface signifies heat. Weak in depth signifies spontaneous perspiration. When the patients feels chill and fears, uneasy because of a fever, nauseous, and with a tendency to snore | Wind-caused initial Yang febrile syndrome (Syndrome of disharmony between nutrient and defense phases) | Guizhi Tang | Guizhi (Cinnamomi Ramulus), 3 liang (one Eastern Han liang is equivalent to 6.69 grams); Shaoyao (Paeoniae Radix Alba), 3 liang; Zhigancao (Prepared Glycyrrhizae Radix et Rhizoma), 2 liang; Shengjiang (Zingiberis Rhizoma Recens), 2 liang; Dazao (Jujubae Fructus), 12 pcs |

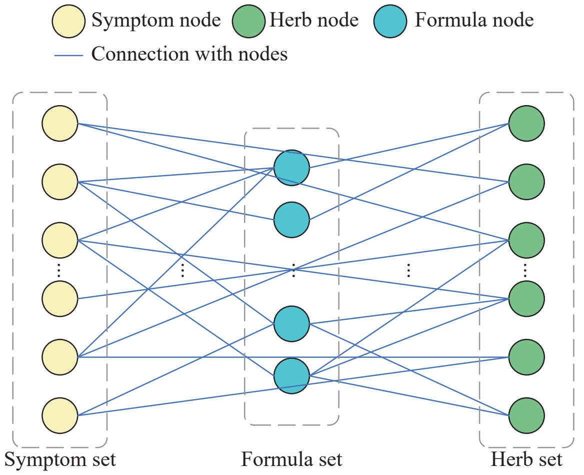

**Figure 1** A symptom-formula-herb heterogeneous graph

Defining $H=\{h_1,h_2,\cdots,h_K\}$, $H$ denotes for all herb sets, and $K$ is the set size.

Defining $C=\{c_1,c_2,\cdots,c_J\}$, $C$ represents all case record sets, and $J$ is the set size. Defining $c=(\{s_1,s_2,\cdots,s_i\},\{f_1,f_2,\cdots,f_j\},\{h_1,h_2,\cdots,h_k\})$, $c$ represents a case record in the dataset, which is composed of subset $S$, subset $F$, and subset $H$. A case record has several symptoms, formulas, and herbs, with $sc=\{s_1,s_2,\cdots,s_i\}$ to represent the aggregations of the symptoms, $fc=\{f_1,f_2,\cdots,f_j\}$ to represent the aggregations of formulas, and $hc=\{h_1,h_2,\cdots,h_k\}$ to represent the aggregations of herbs. So, a case record also represents as $c=(sc,fc,hc)$.

When the symptom subset $sc$ is given, then the task of the model is to calculate the probability $f$ of formulas, and the probability $h$ of herbs for the treatment of $sc$, where $f$ is the $N$-dimensional vector of the probabilities, $N$ is the size of the formula aggregation $F$, and the value of vector $f$ in dimension $i$ denotes for the probability of using the current formula to cure symptom $sc$ subset; $h$ is the $K$-dimensional vector, $K$ is the size of the herb aggregation $H$, the value of vector $h$ in dimension $i$ expresses the probability if using the current herb to cure symptom $sc$ subset.

In order for the symptom, formula, and herb nodes in the heterogeneous graph to be learned efficiently, the node representation learning process in this study is defined as follows according to descriptions in Figure 1 and Definition 1.

(ii) Definition 2

Giving a heterogeneous graph $G=(V,E)$, $V:\{S,F,H\}$, $E:\{sf,sh,fh\}$ to represent the symptom-formula-herb heterogeneous graph is given, where $V$ is made up of nodes $S$, $F$, and $H$, among which $S$ symbols the aggregation of the symptom node $s$, $F$ for the aggregation of formula node $f$, and $H$ for the aggregation of herb node $h$. The edge $sf$ between the symptom and formula is expressed as $(s_i, f_j)$, where $s_i\in S$ and $f_j\in F$. The edge between the symptom and herb is expressed as $(s_i, h_j)$, where $s_i\in S$ and $h_j\in H$. The edge between the formula and herb is expressed as $(f_i, h_j)$, where $f_i\in F$ and $h_j\in H$.

The heterogeneous graph G constructed using *Treatise on Febrile Diseases* text and node attribute sets $X$ are employed here to learn the symptom, formula, and herb node representations, $E_S$, $E_F$, $E_H$, where $E_S\in\mathbb{R}^{|S|\times d_s}$, $|S|$ represents the size of symptom aggregation, $d_s$ is the final value of the learned symptom node vector; $E_F\in\mathbb{R}^{|F|\times d_f}$, $|F|$ denotes for the size of formula aggregation, $d_f$ is the final value of the learned formula node vector; $E_H\in\mathbb{R}^{|H|\times d_h}$, $|H|$ stands for the size of herb aggregation, $d_h$ is the final value of the learned herb node vector. Moreover, $G$ includes all the graph structured information and node attributes. And the representation learning of node $V$ for symptoms, formulas, and herbs in the heterogeneous graph is the main focus of the study.

### 3.1.2 Heterogeneous graph architecture

A total of 195 clauses that have referred to descriptions of symptoms and vital signs as well as the corresponding 112 formulas were selected out of the 398 clauses in *Treatise on Febrile Diseases,* whose descriptions of symptoms, formulas, and herbs were extracted by TCM professionals after proofreading. The extracted information was standardized in accordance with the latest principles from *TCM Syndrome Classification and the Code* (GB/T 15657-2021) published in China.

Each clause has one set of symptom, one of formula, and one of herb. The symptom, formula, and herb sets have appeared in the same clause. Then the nodes representing the symptoms, formulas, and herbs in each clause were connected pair by pair with an edge to construct a symptom-formula-herb heterogeneous network. All of the 195 clauses from *Treatise on Febrile Diseases* were processed this way to ultimately get the whole heterogeneous graph of *Treatise on Febrile Diseases.*

Based on definitions described before, clause (case record) is expressed as $c=(\{s_1, s_2, ..., s_m\}, \{f_1, f_2, ..., f_n\}, \{h_1, h_2, ..., h_k\})$. Those symptoms, formulas, and herbs that have appeared in the same clause were connected by pair. Therefore, we know that in the symptom-formula-herb heterogeneous graph, the aggregation of edges between symptoms and formulas is $\{(s_1, f_1), ..., (s_1, f_n), ..., (s_m, f_1), ..., (s_m, f_n)\}$, that between symptoms and herbs is $\{(s_1, h_1), ..., (s_1, h_k), ..., (s_m, h_1), ..., (s_m, h_k)\}$, and that between formulas and herbs is $\{(f_1, h_1), ..., (f_1, h_k), ..., (f_n, h_1), ..., (f_n, h_k)\}$. The edges are un-directed. Hence, the symptom-formula-herb (S-F-H) heterogeneous graph is composed of three heterogeneous graphs, graph SF, SH, and FH. The edges in the graphs are defined as follows.

$$\mathrm{SF}_{\mathrm{graph}}=\begin{cases}1, & if\ (s_i,f_j)\ co\text{-}occur\ in\ c\\ 0, & otherwise\end{cases} \tag{2}$$

SF represents the symptom-formula graph, if $(s_i,f_j)$ appear in the same clause $c$, so they are connected if their value is 1 and not if their value is 0.

$$\mathrm{SH}_{\mathrm{graph}} = \begin{cases} 1, & if\ (s_i,h_j)\ co\text{-}occur\ in\ c \\ 0, & otherwise \end{cases} \tag{3}$$

SH represents the symptom-herb graph, if $(s_i,h_j)$ appear in the same clause *c*, so they are connected if their value is 1 and not if their value is 0.

$$\mathrm{FH}_{\mathrm{graph}} = \begin{cases} 1, & if\ (f_i,h_j)\ co\text{-}occur\ in\ \ c \\ 0, & otherwise \end{cases} \tag{4}$$

FH represents the formula-herb graph, if $(f_i,h_j)$ appear in the same clause *c*, so they are connected if their value is 1 and not if their value is 0.

Using the clause in Table 1 as an example, the equation for the clause is $c = (\{s_1, s_2, s_3, s_4, s_5, s_6\}, \{f_1\}, \{h_1, h_2, h_3, h_4, h_5\})$ according to the symptoms, formulas, and herbs it contains. Subsequently, all the symptoms, formulas, and herbs are connected with an edge. As a result, edges between symptom *S*, formula *F* are $\{(s_1, f_1), (s_2, f_1), (s_3, f_1), (s_4, f_1), (s_5, f_1), (s_6, f_1)\}$, those between formula *F* and herb *H* are $\{(f_1, h_1), (f_1, h_2), (f_1, h_3), (f_1, h_4), (f_1, h_5)\}$, and those between symptom *S* and herb *H* are $\{(s_1, h_1), (s_2, h_1), (s_3, h_1), (s_4, h_1), (s_5, h_1), (s_6, h_1), (s_1, h_2), (s_2, h_2), (s_3, h_2), (s_4, h_2), (s_5, h_2), (s_6, h_2), (s_1, h_3), (s_2, h_3), (s_3, h_3), (s_4, h_3), (s_5, h_3), (s_6, h_3), (s_1, h_4), (s_2, h_4), (s_3, h_4), (s_4, h_4), (s_5, h_4), (s_6, h_4), (s_1, h_5), (s_2, h_5), (s_3, h_5), (s_4, h_5), (s_5, h_5), (s_6, h_5)\}$. All edges in the graph are undirected. A heterogeneous graph, as shown in Figure 2, is thus constructed.

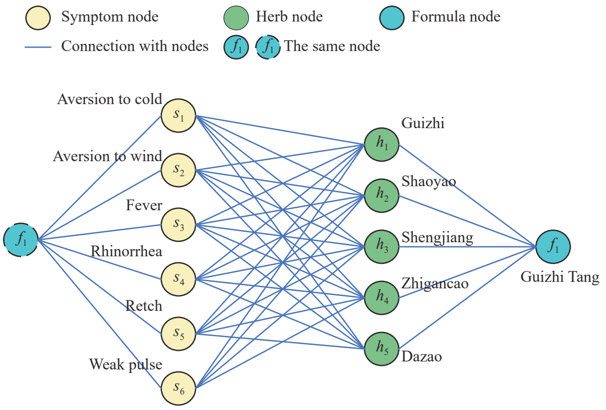

**Figure 2** The heterogeneous graph for Clause 12 in *Treatise on Febrile Diseases*

According to the definitions in this section, the 195 clauses that qualify for the study in *Treatise on Febrile Diseases* were processed to build heterogeneous graphs SF, SH, and FH, which are together made up the whole symptom-formula-herb heterogeneous graph (Figure 3).

## 3.2 Heterogeneous graph representation learning of *Treatise on Febrile Diseases*

**3.2.1 TCM-GCN learning model** The entity vector representations of symptoms, formulas, and herbs are the foundation for predicting the downstream tasks in intelligent TCM diagnosis, whose quality directly determines the performance of the diagnosis prediction model. In the following parts we will introduce how to construct such a model of vector representation for high-order connectivity between TCM entities and their heterogeneous information. We proposed the TCM-GCN model that used message passing and neighbor aggregation to illustrate the learning of nodes in the heterogeneous graph. Message passing and neighbor aggregation are two guides on message passing that aggregate neighboring nodes to update the central node representation, with which the convolutions have been applied for irregular data to realize the connectivity between the graph and neural network.

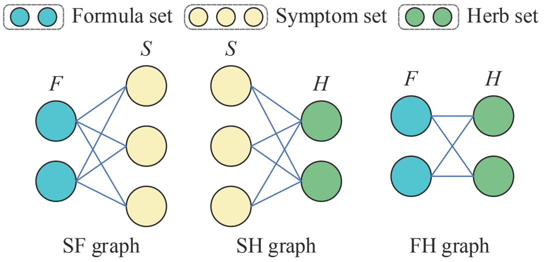

**Figure 3** The symptom-formula-herb heterogeneous graph components

Given that symptoms, formulas, and herbs were represented as three different types of nodes, so they were divided into pairs for representation learning. Symptom nodes were connected with formula and herb nodes, so symptoms were represented by formula nodes in graph SF and herb nodes in graph SH. Next, the two learned vectors from graphs SF and SH were fused. We will now use message passing and neighbor aggregation of the TCM-GCN model to specify this process, as shown in Figure 4.

(i) The graph SF and its attribute matrix were inputted to create the message $\{m_{f_1}, \cdots, m_{f_n}\}$ of the adjacent formula nodes $\{f_1, \cdots, f_n\}$ for each symptom node in $s_i$ the graph. The initial nodes adopted multi-hot encoding.

(ii) The message was then passed and aggregated, and the $s_i$ node representation updated.

(iii) Repeating step (ii) for carrying out *k*-order propagation to obtain the final attribute representation vector $e_{sf_i}$ for node $s_i$ (i.e. node embedding).

(iv) Graph SH underwent the same process as in steps (i), (ii), and (iii) to get the final attribute representation vector $e_{sh_i}$ for node $s_i$.

(v) The attribute vectors $e_{sf_i}$ and $e_{sh_i}$ were fused to get the final attribute representation vector $e_{s_i}$ for node $s_i$.

The same approaches were applied to the formula node $f_i$ and herb node $h_i$, which were used to learn on graphs SF and FH, and graphs SH and FH through the TCM-GCN model to get the final attribute representation vectors for formula nodes and herb nodes $e_{f_i}$ and $e_{h_i}$.

**3.2.2 Message passing and neighbor aggregation** For the representation learning process using the TCM-GCN

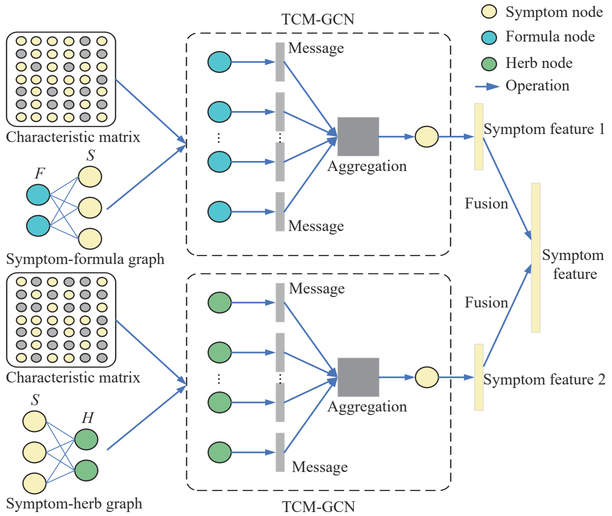

**Figure 4** Symptom representation learning with the TCM-GCN model

model, the message function is defined as follows:

$$m_u^l = MSG^l\left(e_u^{(l-1)}\right) \tag{5}$$

$m_u^l$ refers to the message carried by node $u$ at order $l$ propagation, which was calculated from the attribute vector $e_u^{(l-1)}$ of node $u$ at order $(l-1)$ using the message function $MSG^l$ at order $l$. The message function $MSG^l$ was in line with general linear transformation as follows:

$$m_u^l = W^l e_u^{(l-1)} \tag{6}$$

$e_u^{(l-1)}$ is the attribute vector of node $u$ at order $(l-1)$, the $W^l$ is the weighted matrix. The two times each other can get the transformed message $m_u^l$.

The messages of all neighboring nodes $N(v)$ of the node $v$ at order $(l-1)$ were aggregated, the process of which is defined as:

$$e_v^l = AGG^l(\{m_u^l, u \in N(v)\}) \tag{7}$$

Where $N(v)$ is the sum-aggregation of the neighboring nodes of node $V$, $m_u^l$ is the messages of node $u$ at order $l$, nodes $u$ and $v$ are neighbors, $e_v^l$ is the attribute vectors of node $v$ after aggregation. Aggregation function $AGG$ uesed the sum function $Sum(\cdot)$. Given that messages that the nodes already possess might get lost sometimes during passing, so these messages were transformed following the definitions below:

$$m_v^l = W^l e_v^{(l-1)} \tag{8}$$

$e_v^{(l-1)}$ is the attribute vector of node $v$ at order $(l-1)$, the $W^l$ is the weighted matrix. The two times each other can get the transformed message $m_v^l$. Subsequently, the message of node $v$ and its neighboring nodes were aggregated via concatenation:

$$e_v^l = \sigma(CONCAT(AGG^l(\{m_u^l, u \in N(v)\}), m_v^l)) \tag{9}$$

Where non-linear activation function $\sigma(\cdot)$ used the function $ReLU(\cdot)$, $CONCAT(\cdot)$ represents the concatenating operation.

The Equations (5) – (9) were used for the calculation of message passing and neighboring aggregation in Figure 4. The final attribute vector $e_{sf_i}$ was obtained from learning on the graph SF; the final attribute vector $e_{sh_i}$ was acquired from learning on the graph SH. The two learned vectors were fused via concatenation to get the final attribute vector $e_{s_i}$. Then, all nodes $s$ in $S$ received representation learning to get the matrix $E_S$ of the attribute vector $S$. The exact procedures were repeated to learn on the graph SF and graph FS, and obtain the attribute vector matrix $E_F$ of the formula aggregation $F$, and to learn on the graph SF and graph FS to get the attribute vector matrix $E_H$ of the herb aggregation $H$. Once the attribute vector matrices $E_S$, $E_F$, and $E_H$ of new node representations were gained, the downstream training tasks can be carried out to lay a solid data foundation for the construction of an intelligent diagnosis and treatment model based on *Treatise on Febrile Diseases*.

## 4 Results and discussion

Node representations of symptoms, formulas, and herbs were processed and calculated to get the attribute representation vectors, which were trained using the prediction model. For each clause, the purpose of the training is to minimize the disparity between actual sets and predicted sets of formulas and herbs. Multi-labeling is fit for such a task. The learned node representation vectors got trained using Multilayer Perceptron. The formulas and herbs with the highest probability that rank top $k$ in the results were selected as the predicted formula and herb sets.

### 4.1 Dataset

The clauses from *Treatise on Febrile Diseases* (People's Medical Publishing House, version 2005) were organized, of which 195 eligible clauses were selected. The clauses were then manually processed with references from *TCM Syndrome Classification and the Code* (GB/T 15657-2021) by TCM professionals. A total of 195 sets (symptom sets, formula sets, and herb sets) were gained and used primarily to construct the symptom-formula-herb heterogeneous graph following the steps described above. There were 601 nodes (407 symptom nodes, 112 formula nodes, and 90 herb nodes) and 20177 edges in total on the graph. The dataset was separated as a training set and a test set in a 4 : 1 ratio.

### 4.2 Evaluation index

In order to quantitatively analyze the effects of node representation, three performance evaluation measures, i.e.

precision rate, recall rate, and F1-score, were introduced to evaluate the pros and cons. For clause $c = (sc, fc, hc)$ in the test set, the outcome measures were defined as follows:

$$\text{Precision@}K = \frac{|Top(Set_{pre}, K) \cap Set_{lable}|}{K} \tag{10}$$

$$\text{Recall@}K = \frac{|Top(Set_{pre}, K) \cap Set_{lable}|}{|Set_{lable}|} \tag{11}$$

$$\text{F1-score@}K = \frac{2 \times \text{Precision@}K \times \text{Recall@}K}{\text{Precision@}K + \text{Recall@}K} \tag{12}$$

When the test set $c = (sc, fc, hc)$ was used to recommend herbs, the $Top(Set_{pre}, K)$ represents the top $k$ herbs with the highest scores from prediction, $Set_{lable}$ stands for the actual herb in $hc$, Precision@$K$ stands for the correctly predicted herbs among the top $k$ herbs, Recall@$K$ denotes for the proportion of the correctly predicted top $k$ herbs to the actual herbs in $hc$, and F1-score represents the weighted average value of the precision and recall rates, which offers more objective expressiveness. The $K$ values are 5, 10, and 15.

Similarly, when the test set $c = (sc, fc, hc)$ was used to predict the formula, the $Top(Set_{pre}, K)$ represents the top formulas with the highest scores, $Set_{lable}$ for the actual formula in $fc$, Precision@$K$ represents the precision rate, Recall@$K$ for the recall rate and F1-score have similar connotations as in the prediction of herbs. The $K$ value is 1.

### 4.3 Experiments setup

The study was carried out on the basis of the deep learning library PyTorch Geometric, and the Intel (R) Core i9 10920X processor with 64G memory was used. The learning rate $r$ was set to 0.001, the L2 regularization coefficient was set to 0.001, Dropout value was set to 0.3, and the maximum iteration period was set to 10 000 during training, with the use of Adam optimizer. The optimal depth of TCM-GCN model was 2 layers, the dimension of embedding was 64, and the output dimension of the last layer was 128.

### 4.4 Performance comparision

The MLP was applied to train the node representations calculated from the TCM-GCN. The training results were verified. To measure the pros and cons of different node representations, the multi-hot encoded, non-fusion encoded, and fusion coded node representations were adopted for respective training, and the results were compared.

Table 2 shows that node representation using fusion encoding presented the best performance in herb prediction, those using non-fusion encoding the second best performance, and those with multi-hot encoding yielded the poorest performance among the three. The P@5, R@5, and F1@5 of fusion encoded nodes grew by 7.87%, 1.01%, and 8.97%, respectively, compared with the non-fusion coded ones; however, no significant differences were observed between the results of P@10 and P@15, and of R@10 and R@15, probably because that most formulas *in Treatise on Febrile Diseases* contain only a few herbs. There is only one formula in *Treatise on Febrile Diseases* including over 14 herbs, and a very small proportion of them are made up of more than 14 herbs. So, the changes were not significant because the sample size was comparatively reduced as the herbs increased. The fusion encoding produced better node representation learning in the study than the non-fusion encoding did, suggesting fusion can effectively capture multi-dimensional messages that are related to nodes.

In formula prediction, one clause in *Treatise on Febrile Diseases* corresponds to one formula. No formulas were mixed together, hence the $K$ value was set to 1. Table 3 shows the overall performance of fusion encoding, non-fusion encoding, and multi-hot encoding. Among them, fusion encoding using TCM-GCN produced the best node representation and multi-hot encoding the poorest. Generally speaking, the precision rate in formula prediction was relatively lower than that in herb prediction. The main cause for this might be that symptoms and herbs had more edges between them on the graph, the message the graph carries is more abundant. As opposed to the symptoms and herbs, the relationship between symptoms and formulas was simpler with few messages in their node representations, so the prediction results were poorer by comparison.

In TCM-GCN node representation learning, the layer numbers of GCN also affect the final results. Figure 5 shows the different prediction results of fusion encoding when the convolution layers of TCM-GCN model are 1 to

**Table 2** Results of different node representations for herb prediction

| Encode method | P@5 | P@10 | P@15 | R@5 | R@10 | R@15 | F1@5 | F1@10 | F1@15 |
|---|---|---|---|---|---|---|---|---|---|
| Multi-hot | 0.430 1 | 0.408 1 | 0.425 3 | 0.495 3 | 0.526 1 | 0.501 9 | 0.460 4 | 0.459 7 | 0.460 4 |
| Non-fusion | 0.551 8 | 0.490 2 | 0.510 2 | 0.544 6 | 0.568 6 | 0.540 7 | 0.548 2 | 0.526 5 | 0.525 0 |
| Fusion | 0.595 2 | 0.538 1 | 0.530 8 | 0.599 6 | 0.606 4 | 0.618 9 | 0.597 4 | 0.570 2 | 0.571 5 |
| Gain | 7.87% | 9.77% | 4.04% | 1.01% | 6.65% | 14.46% | 8.97% | 8.30% | 8.86% |

**Table 3** The overall results of different node representations in formula prediction

| Encode method | P@1 | R@1 | F1@1 |
|---|---|---|---|
| Multi-hot | 0.406 6 | 0.383 3 | 0.394 6 |
| Non-fusion | 0.436 0 | 0.460 8 | 0.448 1 |
| Fusion | 0.456 9 | 0.479 8 | 0.468 0 |
| Gain | 4.79% | 4.12% | 4.44% |

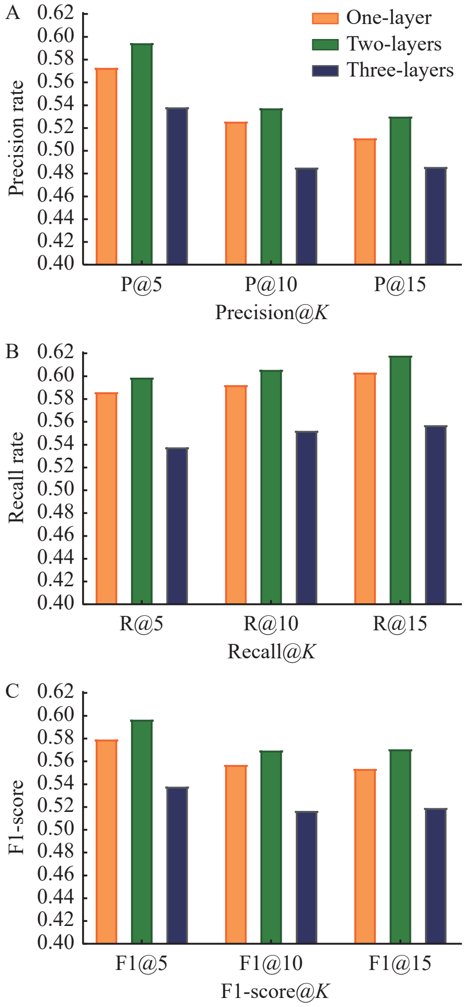

**Figure 5** Effects of TCM-GCN convolutional layer on the prediction results

3 layers deep. It is found that when the number of convolution layers is 1 and 2, the precision rate, recall rate, and F1 value all have better prediction results. However, the improvement of two layers convolution is not obvious compared with that of one layer convolution. When the number of convolution layers reaches 3, the result decreases significantly. The analysis may be caused by the over-fitting of the model when the number of propagation layers becomes larger.

According to the overall results, TCM-GCN can effectively represent the symptom, formula, and herb nodes on the heterogeneous graph of *Treatise on Febrile Diseases*. The convolutions operating at the second layer for message propagation can yield the optimal node representations. Comparisons among the results from the fusion encoding, non-fusion encoding, and multi-hot encoding, fusion encoding is the most efficient in capturing the relationship between nodes, thus improving the effects of node representation.

## 5 Conclusion

In our study, the heterogeneous graph representation learning method was analyzed with the use of *Treatise on Febrile Diseases*, the clauses in which were extracted and processed to build the symptom-formula-herb heterogeneous graph. A node representation learning method for the heterogeneous graph, TCM-GCN, was proposed and applied to learn on symptom-formula, symptom-herb, and formula-herb heterogeneous graphs, during which the high-order propagation was carried out to get new attribute vectors of the node representation via message passing and neighboring aggregation, and in the end to acquire new sets of symptom, formula, and herb node representations. Our study proved the effectiveness of the TCM-GCN model in node representation learning.

Given that the current study has only adopted the simple connectivity among symptoms, formulas, and herbs, without taking into consideration that symptoms in the clauses can be rated as primary or secondary, and herbs in the formulas have different weighted connectivity and relationship such as monarch and minister , assistant and guide. Therefore, an attention mechanism might be taken into future studies to acquire complicated semantics among TCM entities.

## Fundings

New-Generation Artificial Intelligence – Major Program in the Sci-Tech Innovation 2030 Agenda from the Ministry of Science and Technology of China (2018AAA0102100), Hunan Provincial Department of Education key project (21A0250), and The First Class Discipline Open Fund of Hunan University of Traditional Chinese Medicine (2022ZYX08).

## Competing interests

The authors declare no conflict of interest.

## References

[1] ZHANG G. Design of prescription diagnosis and treatment system based on the theory of Sanyin and Sanyang in Treatise on Febrile Diseases. Jinzhong: Shanxi University of Chinese Medicine, 2017.

[2] WU C. A Study on classic prescription clinical decision aid system based on rule engine. Beijing: China Academy of

Chinese Medical Sciences, 2022.

[3] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171–4186.

[4] SHI Q. Research on intelligent prescription model of traditional Chinese medicine based on deep learning. Tianjin: Tianjin University of Traditional Chinese Medicine, 2020.

[5] QU Q. A Study on Treatise on Febrile Diseases based on natural language processing. Heifei: Anhui University of Chinese Medicine, 2021.

[6] WU S, ROBERTS K, DATTA S, et al. Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association: JAMIA, 2020, 27(3): 457–470.

[7] WANG J, XIAO L, YAN J. Construction of knowledge map based on Neo4j's Treatise on Febrile Diseases. Computer and Digital Engineering, 2021, 49(2): 264–267, 396.

[8] KUANG H. Construction of knowledge map of Treatise on Febrile Diseases and its application in Yangming disease. Hangzhou: Zhejiang Chinese Medical University, 2021.

[9] WEI C, YAN J. Study on the construction of syndrome differentiation knowledge graph of diagnostics of traditional Chinese medicine. Basic & Clinical Pharmacology & Toxicology, 2019, 124(S3): 222-223.

[10] RUAN C, MA J, WANG Y, et al. Discovering regularities from traditional Chinese medicine prescriptions via bipartite embedding model. International Joint Conference on Artificial Intelligence (IJCAI), 2019: 3346–3352.

[11] JIN Y, ZHANG W, HE X, et al. Syndrome-aware herb recommendation with multi-graph convolution network. Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020: 145-156.

[12] ZHAO W, LU W, LI Z, et al. TCM herbal prescription recommendation model based on multi-graph convolutional network. Journal of Ethnopharmacology, 2022, 297: 115109.

[13] TU C, YANG C, LIU Z, et al. Network representation learning: an overview. Journal of SCIENTIA SINICA Informationis, 2017, 47(8): 980–996.

[14] CHOI E, XIAO C, STEWART WF, et al. MiME: multilevel medical embedding of electronic health records for predictive healthcare. Proceedings of the 32nd Advances in Neural Information Processing Systems, 2018: 4547–4557.

[15] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks. Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 855–864.

[16] KIPF T, WELLING M. Semi-supervised classification with graph convolutional networks. Proceedings of the 11th International Conference on Learning Representations, 2017: 1–10.

[17] REN Y, SHI Y, ZHANG K, et al. A drug recommendation model based on message propagation and DDI gating mechanism. IEEE Journal of Biomedical and Health Informatics, 2022, 26(7): 3478–3485.

[18] WANG H, LE Z. Expert recommendations based on link prediction during the COVID-19 outbreak. Scientometrics, 2021, 126(6): 4639–4658.

[19] ZHAO T, HU Y, VALSDOTTIR LR, et al. Identifying drug-target interactions based on graph convolutional network and deep neural network. Briefings in Bioinformatics, 2021, 22(2): 2141–2150.

[20] XU X, YUE L, LI B, et al. DSGAT: predicting frequencies of drug side effects by graph attention networks. Briefings in Bioinformatics, 2022, 23(2): bbab586.

[21] JIN Y, JI W, ZHANG W, et al. A KG-Enhanced multi-graph neural network for attentive herb recommendation. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(5): 2560–2571.

[22] YANG Y, RAO Y, YU M, et al. Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation. Neural Networks, 2022, 146: 1–10.

[23] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 1024-1034.

[24] YING R, HE R, CHEN K, et al. Graph convolutional neural networks for web scale recommender systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 974-983.

[25] ZHANG C, SONG D, HUANG C, et al. Heterogeneous graph neural network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 793-803.

# 基于图卷积网络的《伤寒论》异质图构建及节点表示学习方法

晏峻峰[a], 文志华[a,b], 邹北骥[a,c*]

*a. 湖南中医药大学信息科学与工程学院, 湖南 长沙 410208, 中国*
*b. 湖南工业大学计算机学院, 湖南 株洲 412008, 中国*
*c. 中南大学计算机学院, 湖南 长沙 410083, 中国*

**【摘要】目的** 基于图卷积神经网络，构建《伤寒论》“症状-方剂-中药”异质图并探寻节点向量表示的最优学习方法。**方法** 从《伤寒论》含处方的条文中提取出症状、方剂、中药信息，构建“症状-方剂-中药”异质图，基于图卷积网络提出一种“症状-方剂-中药”异质图节点表示学习方法—中医图卷积网络（TCM-GCN），利用 TCM-GCN 分别对症状-方剂、症状-中药、方剂-中药异质图进行学习，基于消息传递和邻居聚合进行高阶传播得到节点的表示特征向量，获得症状、方剂、中药三类节点表示集合，为下游诊断预测模型任务的顺利开展提供基础。**结果** 通过多热编码、非融合编码、融合编码三种节点表示方式在模型预测实验中对比发现，融合编码方式获得了相对较高的精准率、召回率和 F1-score 值，其 Precision@10、Recall@10 和 F1-score@10 值较非融合编码分别提升了 9.77%、6.65% 和 8.30%。**结论** 融合编码方式生成的节点表示在实验中取得了较好效果，表明《伤寒论》异质图节点表示 TCM-GCN 方法的有效性，也将提升其在下游诊断预测任务上的性能。

**【关键词】**图卷积网络；异质图；《伤寒论》；异质图节点表示；节点表示学习