成绩	（采用四级记分制）

西北大学

本科毕业论文（设计）

题目：基于深度学习的中药推荐方法的研究和实现

学生姓名 邹刘文

学　　号 2017111093

指导教师 张海波 副教授

院　　系 信息科学与技术学院

专　　业 软件工程

年　　级 2017 级

教务处制

二○二一年六月

诚信声明

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或在网上发表的论文。

特此声明。

论文作者签名：_____

日　期：　　年　月　日

摘　要

为了更好地推进中医现代化，计算机技术将中医问诊过程建模为草药推荐问题，即根据一组症状推荐一组草药。这既能实现信息共享进而辅助专家医生给出处方建议，又能发掘出症状和药物之间隐匿的联系。

本文主要针对考虑了中医综合症归纳过程的多图卷积模型 SMGCN 进行研究实现，并展开了相关的算法实验。首先将其与电子商务领域经典的图推荐模型 NGCF、HetGCN、PinSage 进行对比，以验证 SMGCN 的独特优势；然后进行消融实验以验证 SMGCN 各级组件的作用；最后展开超参实验以研究不同超参对推荐性能的影响。

实验结果表明，SMGCN 在推荐效果上要优于 NGCF 和 HetGCN，略逊于 PinSage，但是鉴于其训练总时长和单位参数的训练时长都要远低于 PinSage，在需要快速训练部署的生产场景中理应成为首选；SMGCN 模型的 Bipar-GCN、SGE 以及 SI 组件都对推荐效果有提升作用，其中 SGE 的嵌入增强能力最为显著；图卷积层的个数、最终嵌入维度、症状阈值和草药阈值是最影响 SMGCN 性能的四个超参。

关键字：中草药推荐；药症关系；综合症归纳；图神经网络推荐；数据挖掘；

Abstract

For the better promotion of the modernization process of Chinese medicine, computer technology models the consultation process of Chinese medicine as a herbal recommendation problem, that is, recommend a group of herbal medicines based on a group of symptoms. This can not only assist expert doctors in giving prescription advice through information sharing, but also uncover the hidden connection between symptoms and drugs.

This article focuses on the research and implementation of the multi-graph convolution model SMGCN considering the induction process of TCM syndrome, and launches related algorithm experiments. First, we compare it with the classic graph recommendation models NGCF, HetGCN, and PinSage used in the e-commerce field to verify its unique strength; then we conduct ablation experiments to verify the role of SMGCN components at all levels; finally, a super-parameter experiment was carried out to study the influence of different super-parameters on the recommended performance.

Experimental results show that SMGCN is better than NGCF and HetGCN in recommendation effect, and slightly inferior to PinSage, but due to its much shorter total training time and unit parameter training time than PinSage, it should be the first choice in production scenarios that require rapid training and deployment; the Bipar-GCN, SGE and SI components of the SMGCN model all have an effect on the recommendation effect, among which the embedding enhancement ability of SGE is the most significant; the number of graph convolutional layers，the final embedding dimension, and the threshold setting of symptoms and herbs are the four super-parameters that most affect the performance of SMGCN.

Key words： herb recommendation; drug-symptom relationship; syndrome induction; graph neural network recommendation; data mining;

目录

1 绪论

1.1 研究背景与意义

1.1.1 研究背景

在最近经历的新冠疫情中，藿香正气胶囊、连花清瘟颗粒等中药再次发挥了神奇的疗效，作为中国药都的安徽省亳州市仅耗时 43 天就使得全部 108 病患出院，这使得中医和中药在继 2003 年抗击非典、2015 年屠呦呦女士因青蒿素研究获诺贝尔医学奖、2016 年拔火罐风靡里约运动赛场后再次进入人们的视野当中。而在日常生活中，人们经常使用中药来治愈皮肤病、心脏病、头痛、中风等西药难以根治的问题。

然而，中医药由于药理性质类似于黑箱，不如西药透明，在我国现代化的医疗体系中未受到长足的重视。反观同为黑箱方法的机器学习以及深度神经网络，人们并未纠结于其原理的不可解释性，反倒因其极强的泛化能力而大受推崇，我们认为其主要区别在于可重复性。机器学习等黑箱方法只要用户给定相同的数据和算法，就能够得出相对一致的结果；而如图 1-1 所示的传统中药问诊过程包含收集病人症状、推断综合症、确认治疗方案三个核心步骤[12]，这十分依赖于医生的个人经验和主观性，对于同一症状不同医生甚至可能会给出完全不一致的处方建议，对于同一处方患者的治疗效果也有所不同。

为此我们用计算机技术将其建模为数据驱动下的中药推荐问题。我们希望利用机器学习或者深度学习技巧从大量的经典中医处方数据中挖掘出症状和药物间的关联模式，进而根据一组症状来推荐一组中药。

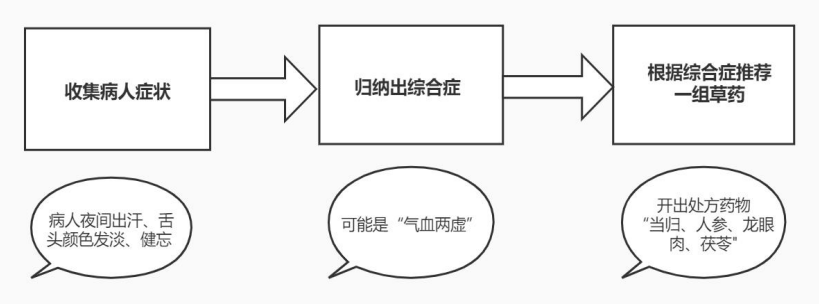

图 1-1　传统中医问诊的过程示意图

1.1.2 研究意义

中医知识体系相对繁杂且经验信息不能很好在医生之间共享。利用推荐机制一方面可以

解决中医处方数据的信息过载[9]问题，为医生用药提供更为全面可靠的辅助意见；另一方面也有益于从遗留的中医处方数据中挖掘出潜藏的知识和原理，进而将杂乱的中医问诊过程标准化、可重复化、信息化，使得中医成为真正意义上屠呦呦女士所说的“中国献给世界的礼物”。

1.2 研究现状

1.2.1 中药推荐问题的研究范畴

第一类是针对某种具体疾病进行中药推荐。既会考虑疾病本身的特点，比如王科文[10]根据舌苔图像来进行中药材推荐的研究；也会考虑到不同病人个体间的差异，例如曹小凤[8]针对不同的高血压患病人群使用了案例推理与贝叶斯推理相结合的混合推理算法。

第二类是先从数据挖掘的角度对症状和药物间的复杂关联进行建模，然后应用关联预测来进行通用的中药推荐。例如王新宇[4]从融合了多源异构中医数据的知识图谱中获取更好的症状和药物实体表示，然后利用内积得分进行推荐；汪浩[2]等人通过推荐模型从海量的疾病和药物的治疗关系、药物和疾病实体内部的相似性关系数据中归纳挖掘出可能的药物-疾病对。

1.2.2 中药推荐算法的研究现状

根据算法策略的不同，目前主流的中药推荐算法可以进行如下划分。

首先是数据挖掘算法。比如于然等人[1]通过在筛选出的高频中药上进行关联规则和聚类分析来分析外治手足综合征的用药规律；唐等人[11]将复杂网络和关联规则结合来捕捉肾系疾病治疗用药的特点；王茹玉[6]则立足于非负矩阵分解（NMF）等数学技巧来探寻症状和药物间的联结方式；庄等人[7]改善了用于聚类的 Biclustering 算法，以此更好地分析药物和症状间的关系。

其次是主题模型，其以包含中药和症状等概念的自然语言文本为数据基础，认为发生同一主题下的中药和症状理论上应该是相似的。Ji 等[13]认为“发病机理”是将症状和中药联系起来的潜在主题。 Ma 等[14]提出了一个“症状综合症”模型来挖掘症状和潜在综合症主题之间的相关性。Lin 等[15]通过主题模型共同对处方中的症状、中药、诊断和治疗进行建模。Wang 等[16]设计了一个不对称概率生成模型，以同时对症状、中药和疾病进行建模。 Yao 等[25]将“综合症”、“治疗”和“中药作用”等中医概念整合到主题建模中，以更好地刻画处方的生成过程。王等[24]将从中药知识图谱上获取的 TransE 嵌入和主题模型进行融合，将处方中的共现信息和知识图谱中中药间的关联都考虑在内。

最后是基于图模型的方法。图推荐在电子商务领域已迅速发展起来，用户和商品构成的结构图相比于商品评分矩阵更能够反映交互信息以及刻画用户画像[2]， 因此得到的嵌入含有更丰富的协同信息。IntentGC[26]通过图卷积网络对用户的行为和商品的信息进行建模来提取其中的偏好和异构关系。MEIRec[27]在异构图卷积网络的基础上引入了元路径的指导，能得到更好的关于被推荐对象的特征表示。而在中药推荐方面，Li 等人[17]借助 SeqSeq 注意力机制构建了一个自动化生成处方的工具，[18],[19]在自动编码器中使用了元路径来挖掘中药异质信息网络， Li[21]等人在中医文献上使用 BRNN 对中药进行文本嵌入学习。特别的是，[20],[12]都将中医问诊过程的综合症归纳或者病人分类的思想纳入到模型的考虑当中，但[20]构建了一个症状-综合症-中药的异构网络并利用拓扑结构来进行学习，而[12]则是借助于 MLP 来将多个症状嵌入融合得到综合症的特征表示然后再与中药嵌入进行交互。

需要指出的是，很多方法起初都是为了解决电子商务中经典的用户-商品推荐或西药与疾病的关系建模，但是基于中药推荐问题的相似性，这些方法是非常容易迁移并且富有成效的。

1.3 本文的研究内容和研究目的

数据挖掘算法更适用于单一疾病的研究；标准主题模型依赖于海量文献中富含的中医药语义知识，在文本本身比较稀疏时，效果便无法得到保证。

由于期望实现通用推荐策略，同时也缺乏海量的中医药文本数据，故本文只对图神经网络方向内的推荐算法展开研究。本文将用户-商品领域的几种经典图推荐算法迁移到症状-中药推荐任务中来，与结合了中医综合症归纳过程的多图卷积算法 SMGCN[12]进行对比，并展开设计了一系列算法实验。此外，本文还基于已有数据集开发了一个小型的中药推荐系统，以期可视化不同算法的实际效果。

1.4 本文的组织结构

本文结构由下述六章组成：

第 1 章首先讲解了本文的研究背景和意义，然后分析了问题研究的具体范畴和研究现状，最后提出了研究目标和研究内容。

第 2 章介绍相关理论和技术，主要概述了图神经网络、推荐策略以及两者的技术结合。

第 3 章详述中药推荐模型（包括 SMGCN 和其他对照模型）的架构和数学原理。

第 4 章详述中医推荐算法实验过程。主要包括实验数据准备、三大核心实验及结果、结

果评估与结论。

第 5 章介绍中药推荐的可视化系统。主要包括其核心功能、系统的关键技术和整体架构。

第 6 章是对课题的相关总结和展望。主要对算法实验的结果进行了总结分析和思路拓展。

2 相关理论和技术

2.1 图神经网络与推荐系统

2.1.1 图神经网络

图神经网络是用图替代传统输入的新的应用模式。

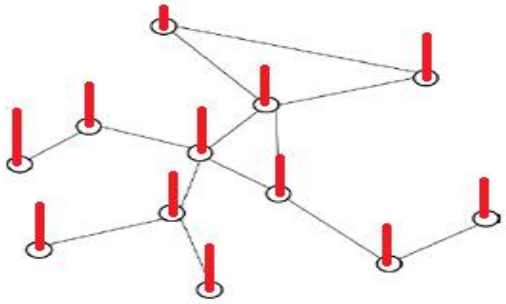

图 2-1 图神经网络示意图（红色方柱表示图节点的嵌入表示）

关于 GNN 的具体工作机制，可以从图和神经网络的不同用途来详细说明。对于 GNN 中的图来说，其没有结构上的明确约束，只需满足 G=（V，E）的通用形式化定义即可，也就是说图可以是同类实体组成的同质图也可以是多种不同类实体构成的异质图，可以是 0-1 关系的硬性连接图也可以是反映关系强弱程度的软性连接图，因此其主要用于准确生动地建模现实世界中各种非欧几里得域的复杂联系。而对于拥有很强非线性能力的深度神经网络来说，GNN 在将图上的每个节点表示为类似于图信号的特征嵌入后，将其输入其中并结合图的结构进行信息传播，最后通过模型训练就能够收敛到既反映节点自身属性又能够反映结构关系的良好嵌入。由此可知，GNN 是图嵌入表示领域中一个全新的强有力工具。相比于基于节点关系矩阵的矩阵分解方法比如[33]等其不会局限于小图，能够在大尺度的图上进行学习；相比于基于随机游走将图转化成序列的方法 DeepWalk[32]等，其不会遗漏掉结构的信息，并能深度融合图的相关属性信息，所以得到了迅速的发展。

关于 GNN 的应用，主要是将获取的图嵌入表示应用于一系列下游任务中。可以用来进行节点分类，比如 text GCN[31]在待分类的文档节点和去重的单词节点构成的大型异质图上进行图卷积以进行文本分类；可以用来进行链接预测，比如根据已有的分子结构去预测新的化合物；可以用来进行节点生成，比如利用计算机视觉领域中 COCO 数据集的所有图片对象来构建一个依赖于图片位置和大小的场景图，然后在图中添加一个新节点后可以利用 GNN 去生成对应图像[23]；可以用来进行网络结构的优化，比如谷歌团队将芯片表征为内存和逻辑单

元构成的图，然后利用 GNN 模型去组合优化硬件电路的布局[22]。

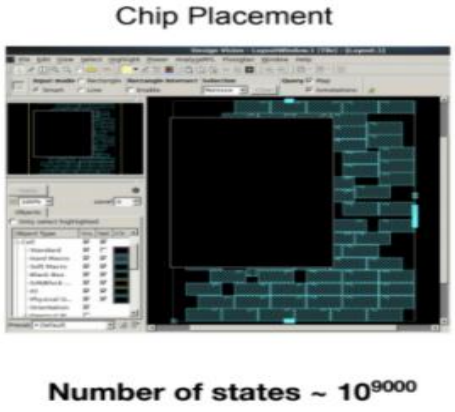

图 2-2　GNN 优化硬件芯片的电路布局[22]

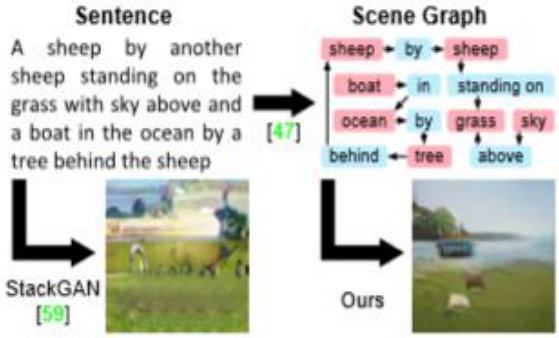

图 2-3 利用 GNN 根据场景描述语句生成图片[23]

2.1.2 推荐策略

相继出现了基于内容的推荐[34]、使用协同过滤策略的推荐[35]以及混合推荐[36]这三类经典策略。

基于内容的推荐从商品的角度出发，认为用户会对具有相似内容特点的商品产生兴趣，是一种从商品到商品的推荐方式。比如说，用户的历史交易记录中含有刘慈欣所著的《三体》小说，那么用户也很有可能购买其他诸如《球状闪电》、《流浪地球》的科幻作品。

使用协同过滤的推荐从用户和用户的关系网络展开，认为相似特征的用户群体会有类似的物品需求和偏好，是一种从用户到商品的推荐方式。比如说，同为软件从业人员的 A 和 B，若 A 经常购买鼠标、键盘等电子产品，则可以向 B 推送 A 的历史购物清单。

而混合推荐则是指通过融合方式来组合不同推荐方法的效果，主要包括前融合、中融合和后融合三类方式。前融合指的是综合所有已知策略的特征输入，利用统一的模型进行推荐；

中融合指的是立足于一种策略，尽量融合进其他策略的算法思想；后融合是对每个策略产生的结果进行一个综合的评估。

需要注意的是，三大策略都只是产生推荐结果的基本思想，而非具体的算法。事实上，针对每一类策略都衍生出众多的机器学习或者深度学习算法。

2.1.3 图神经网络在推荐系统中的应用

图神经网络属于深度学习的分支，其在推荐系统中的应用方法与基于深度学习的推荐系统整体架构是一致的。

（1）基于深度学习的推荐系统架构

如图 2-4 所示，深度学习主要被用于推荐系统中的获取实体嵌入一环中。

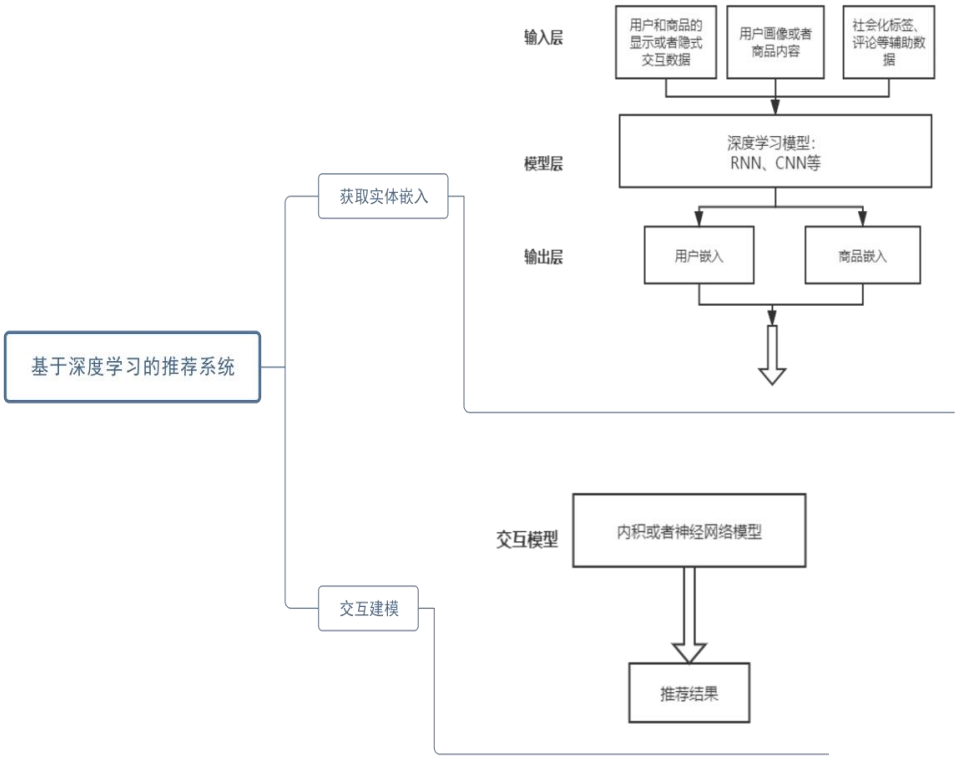

图 2-4 基于深度学习的推荐系统架构图

（2）基于图神经网络的推荐系统

与其他传统深度学习方法的区别仅在与利用图神经网络来获取实体嵌入。

图神经网络按照类别可以划分为图卷积网络、图注意力网络、门图神经网络、图残差网络、图递归网络这 5 类，其中前 3 类经常用于推荐系统。

图卷积网络即 GCN，类比与图像卷积在标准的欧几里得空间聚合局部像素点的细节，图

卷积在非欧几里得空间聚合节点周围的邻居信息以捕获节点间的连接关系，因而可以提取到更好的特征。具体来说，GCN 可以研究社交推荐中社交影响力的扩散情况、预测用户-商品关系从而缓解初始数据的稀疏程度。

图注意力网络主要是区别对待不同的邻居节点信息；门图神经网络增加了一个 GRU 单元，常用于使用会话机制的推荐问题。

2.2 图神经网络领域的深度学习工具

2.2.1 深度学习模型框架

用于深度学习工程开发的集成框架有很多，本文在模型的搭建、训练以及测试的过程主要使用 tensorflow 和 pytorch。

2.2.2 构建图结构的工具

图结构既可以用传统的邻接矩阵、拉普拉斯矩阵等来表示，也可以使用封装了节点创建、边创建、权重设置等图操作的相关 API 辅助创建。本文主要使用了与 pytorch 有良好兼容性的 DGL（deep graph library)，以便从数据集中生成大型图并在图上高效地执行信息传播。

3 中药推荐模型

本章主要介绍针对“根据一组症状推荐一组中药”这一组推荐问题的相关模型，包括它们各自的架构和数学原理。模型包括结合了综合症归纳过程的 SMGCN 模型，以及从传统电子商务领域迁移过来的 NGCF、HetGCN 和 PinSage 模型。

3.1 中药推荐模型的基本思想

所有的模型都是试图参数化症状和中药以重构症状-中药的治疗关系矩阵，并依据参数来进行中药集合的预测，即本质上都是获取症状和中药的有效嵌入。这些模型的通用形式架构如下图 3-1 所示：

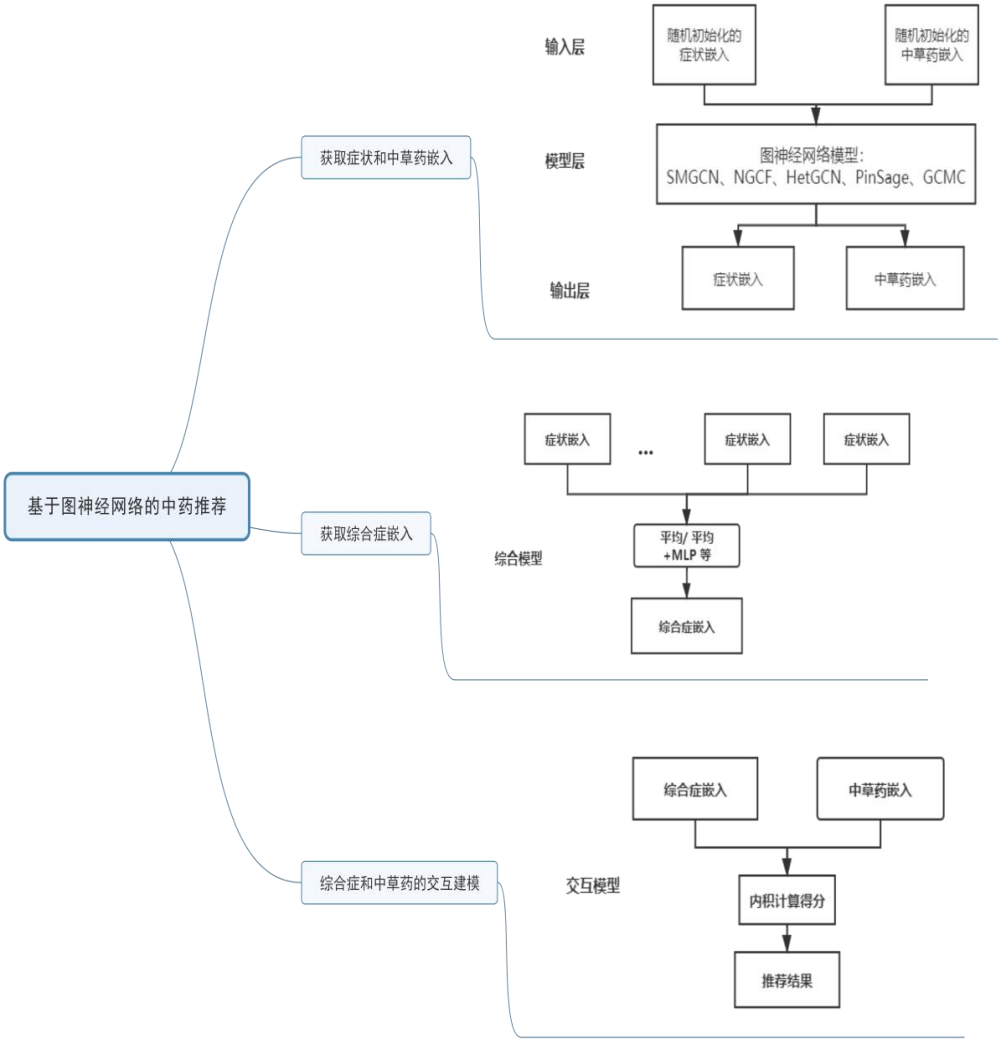

图 3-1　中药推荐模型的架构图

3.2 SMGCN 模型

3.2.1 模型简介

即（Syndrome-aware MultiGraph Convolution Network）基于综合症归纳的多图卷积模型[12]。其使用二部图来表征症状和草药的交互信息，同时使用协同图来建模症状之间、草药之间的相似特性，最后还从综合症推导过程中获得启发，使用万能 MLP 来融合症状组中各个症状的关系。

3.2.2 模型整体架构

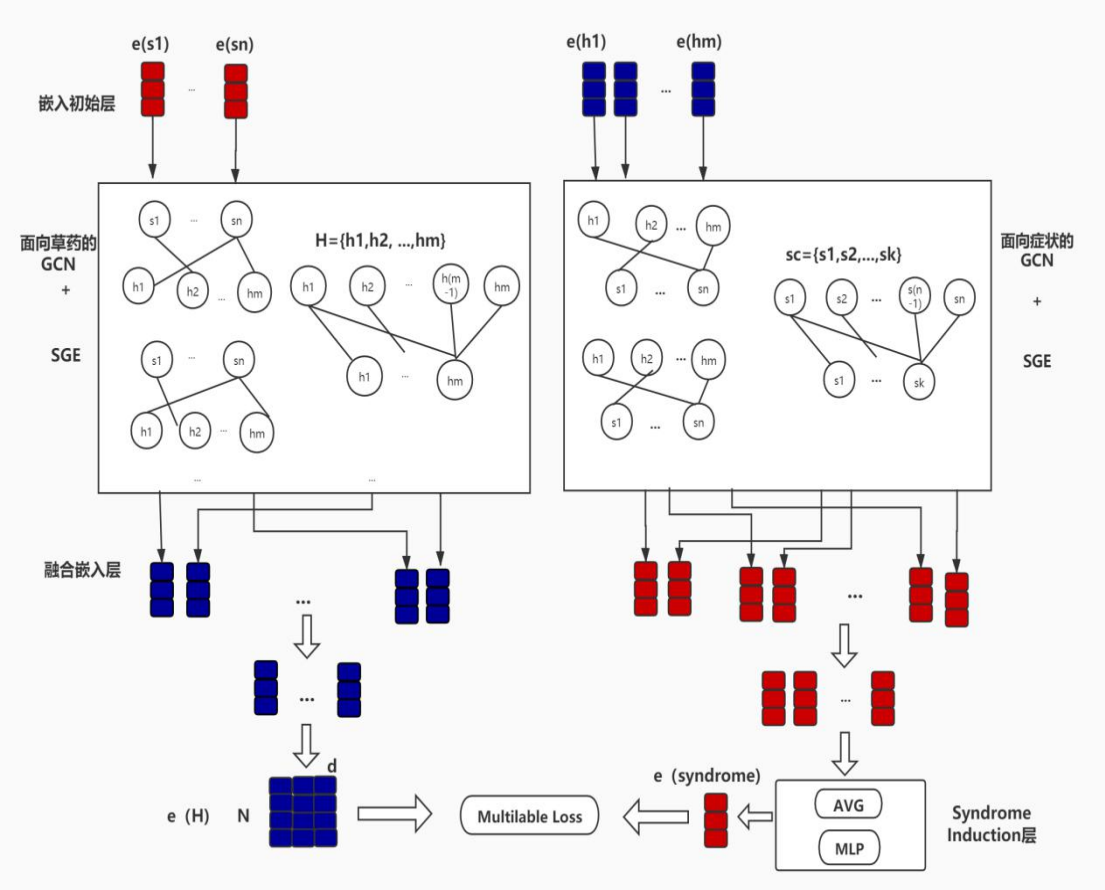

图 3-2 SMGCN 模型的架构图

3.2.3 模型原理

3.2.3.1 嵌入初始化层

假设症状数为 N，草药数为 M，随机初始化所有症状节点和草药节点的嵌入，分别得到其矩阵表示:

$$E_{sympt}^{(0)} = \begin{bmatrix} e_{s_1}^{(0)}, \\ e_{s_2}^{(0)}, \\ \ldots, \\ e_{s_n}^{(0)} \end{bmatrix} \in R^{N \times d_0} \tag{3-1}$$

$$E_{herb}^{(0)} = \begin{bmatrix} e_{h_1}^{(0)}, \\ e_{h_2}^{(0)}, \\ \ldots, \\ e_{h_m}^{(0)} \end{bmatrix} \in R^{M \times d_0} \tag{3-2}$$

然后按照第一维度堆叠得到综合症状和草药节点的嵌入矩阵：

$$E^{(0)} = \begin{bmatrix} E_{sympt}^{(0)} \\ E_{\text{herb}}^{(0)} \end{bmatrix} \in R^{(N+M) \times d_0} \tag{3-3}$$

3.2.3.2 分别面向症状和草药的 Bipar-GCN 层

Bipar-GCN 是 Bipartite-GCN 的缩写，即基于二部图结构的 GCN，其能够很好地建模草药和症状两种实体之间的交互关系。该组件的工作流程包括构建二部图、在二部图上进行卷积操作。

首先我们利用邻接矩阵来构建二部图，即对于同一个处方当中同时出现的症状和草药，我们认为其相互之间具有治疗关系：

$$R_{s_n,h_m} = \begin{cases} 1, \text{ if } (s_n,h_m) \ occurs \ in \ the \ same \ prescription \\ 0, \ otherwise \end{cases} \tag{3-4}$$

然后我们再二部图上构造消息、整合邻居消息、更新节点消息，并不断进行多级传播。我们以图 3-3 中草药节点 h_2 的第一层卷积为例讲述整个过程：

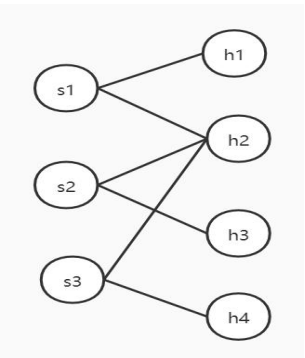

图 3-3　症状-草药二部图结构的简单示意图

1）构造草药节点的所有一阶症状邻居要向草药节点传播的消息：

$$m_{h<-s}{}^{(1)} = e_s{}^{(0)} \bullet T_h{}^{(1)} \quad (s = s_1, s_2, s_3,\ h = h_2) \tag{3-5}$$

2）整合所有症状邻居节点的消息：

$$b_{N_h}{}^{(1)} = tanh\left(\frac{1}{|N_h|}\sum_{s \in N_h} m_{h<-s}{}^{(1)}\right) \tag{3-6}$$

3）更新草药节点信息：

$$b_h{}^{(1)} = tanh\left(W_h{}^{(1)} \bullet (e_h{}^{(0)} || b_{N_h}{}^{(1)})\right) \tag{3-7}$$

4）进行多级传播:

$$b_h{}^{(l)} = tanh\left(W_h{}^{(l)} \bullet (e_h{}^{(l-1)} || tanh\left(\frac{1}{|N_h|}\sum_{s \in N_h} e_s{}^{(l-1)} \bullet T_h{}^{(l)}\right)\right)\right) \tag{3-8}$$

对于症状节点同理可得

$$b_s{}^{(l)} = tanh\left(W_s{}^{(l)} \bullet (e_s{}^{(l-1)} || tanh\left(\frac{1}{|N_s|}\sum_{h \in N_s} e_h{}^{(l-1)} \bullet T_s{}^{(l)}\right)\right)\right) \tag{3-9}$$

这里之所以说是分别面向症状和草药节点的GCN正是因为构造信息和节点更新使用的T和W矩阵在两者之间不进行共享。

最后，我们使用矩阵而非节点采样的方式在整个图上进行操作，如果矩阵过大可以使用稀疏矩阵缓解内存问题，矩阵形式如下：

$$A = \begin{bmatrix} O & R \\ R^T & O \end{bmatrix} \tag{3-10}$$

$$neigh\_info = tanh\left(D^{-1} stack\_mat\_by\_row(AE^{(l-1)}[:N]T_s{}^{(l)}, AE^{(l-1)}[N:]T_h{}^{(l)})\right) \tag{3-11}$$

$$E^{(l)} = tanh\left(stack\_mat\_by\_row((E^{(l-1)}[:N]||neigh\_info[:N])W_s{}^{(l)},\right.$$
$$\left.(E^{(l-1)}[N:]||neigh\_info[N:])W_h{}^{(l)})\right) \tag{3-12}$$

其中D是A对应的度数矩阵，R^T是R的转置矩阵。

3.2.3.3 **协同图编码** SGE(Synery Graph Encoding)

这里的协同图指的是症状和草药的共现图，其通过计算同类实体在所有处方中同时出现

的次数来反映出同类实体间的协同相似性。

以草药-草药同质图为例来说明具体过程。

首先对计算的共现次数矩阵通过设定的阈值进行二值化来创建协同图，大于阈值才认为两类草药间具有相似功效：

$$HH_{h_{m,}\ h_n},\ HH_{h_n,h_m} = \begin{cases} 1, & \text{if } frequency(h_m,h_n) > x_h \\ 0, & otherwise \end{cases} \tag{3-13}$$

然后将其视为草药-草药型的特殊二部图进行一层图卷积，这样可以使得相似的草药具有相似的嵌入表示：

$$r_s^{(1)} = tanh\left(\sum_{k \in N_s^{SS}} e_k^{(0)} \bullet V_s^{(1)}\right) \tag{3-14}$$

$$r_h^{(1)} = tanh\left(\sum_{q \in N_h^{HH}} e_q^{(0)} \bullet V_h^{(1)}\right) \tag{3-15}$$

3.2.3.4 融合嵌入层

使用简单相加来合并从 Bipar-GCN 中获得的交互嵌入与从 SGE 中获得的协同嵌入：

$$e_s^* = b_s + r_s \tag{3-16}$$

$$e_h^* = b_h + r_h \tag{3-17}$$

3.2.3.5 综合症推导层 SI(syndrome induction)

先对所有症状的合并嵌入使用一层平均池化层得到平均嵌入向量，再用一层 MLP 获取综合症嵌入：

$$e_{\text{syndrome}}(\text{sympt\_set}) = ReLU\left(W^{\text{mlp}} \bullet Mean\left(e_{\text{sympt\_set}}\right) + b_{\text{mlp}}\right) \tag{3-18}$$

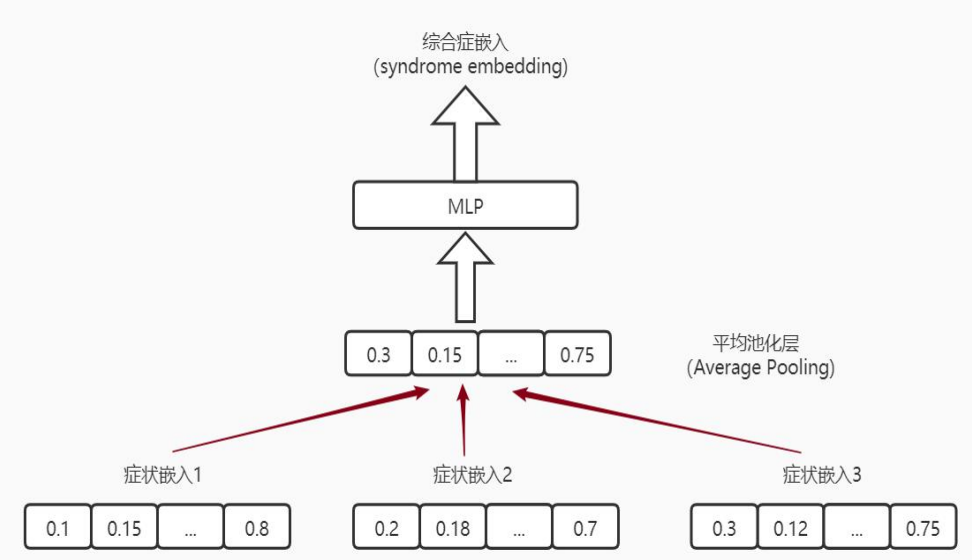

图 3-4 综合症归纳流程图

3.2.3.6 模型预测层

将综合症嵌入e_{syndrome}（$sympt\_set$)与草药堆叠矩阵$e_H \in R^{N\times d}$通过内积进行交互，得到草药推荐的概率向量$y_{\text{sympt_set}}$，其中$y_{\text{sympt_set}}(i)$表示第 i 种草药被推荐的概率：

$$\begin{aligned} y_{sympt\_set} &= g(sympt\_set,h) \\ &= e_{syndrome}（symp\_set) \cdot e_H^T \end{aligned} \tag{3-19}$$

3.2.3.7 模型损失层

草药的组推荐可以视为多标签分类问题，我们利用多标签损失函数进行训练：

$$Loss = \underset{\theta}{argmin} \sum_{(sc,hc')\in P} WMSE(hc',g(sc,H)) + \lambda_\theta ||\theta||^2_2 \tag{3-20}$$

$$WMSE（\text{hc}',g(sc,H)) = \sum_{i=1}^{|H|} w_i(hc'_i - g(sc,H)_i)^2 \tag{3-21}$$

$$w_i = \frac{\max_k \text{freq}(k)}{freq(i)} \tag{3-22}$$

其中 hc'是处方数据集中症状集合对应的经典草药集合，即 ground truth，而 WMSE 意为 hc'和 g（sc，h）之间的（Weighted Mean Square Loss)加权平均平方损失， sc 是 sympt_set 的简写。

3.3 NGCF

3.3.1 模型简介

NGCF[28]是由中科大何向南团队提出的基于图卷积网络的协同过滤推荐模型，其发现传统的协同过滤方法仅利用了用户和商品诸如 ID 编号、属性等描述性的特征，故希望将交互数

据中的协同信号也显示地纳入其中。

NGCF 模型可以很好地迁移到本文问题中来，只需要增加 Average SI 并将正负样本损失函数（BPR Loss）修改为多标签损失函数（MultiLabel　Loss)。

3.3.2 模型整体架构

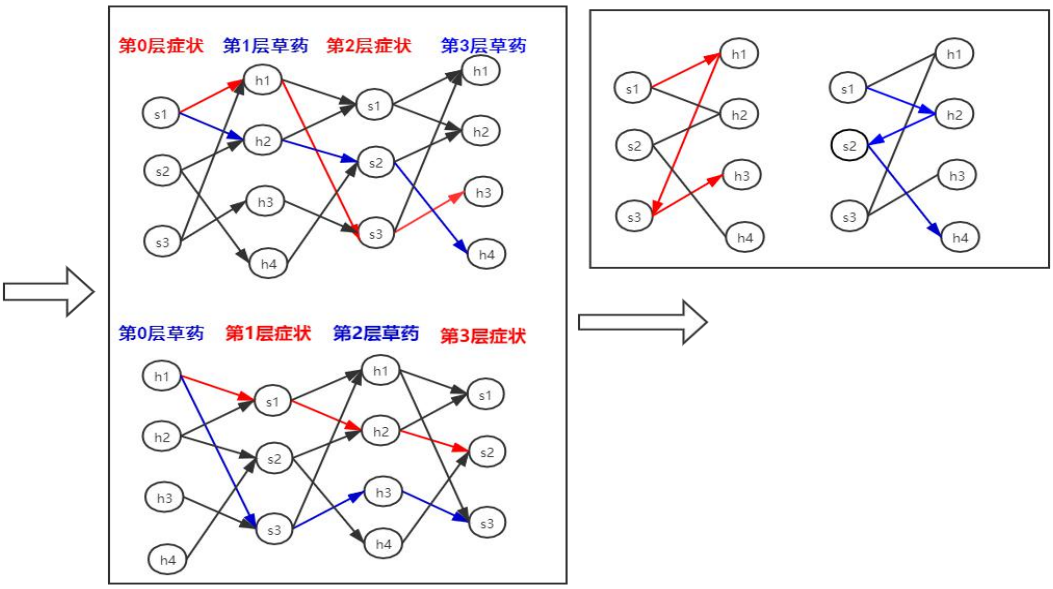

图 3-5 NGCF 利用高阶连通性挖掘协同信号的原理示意图

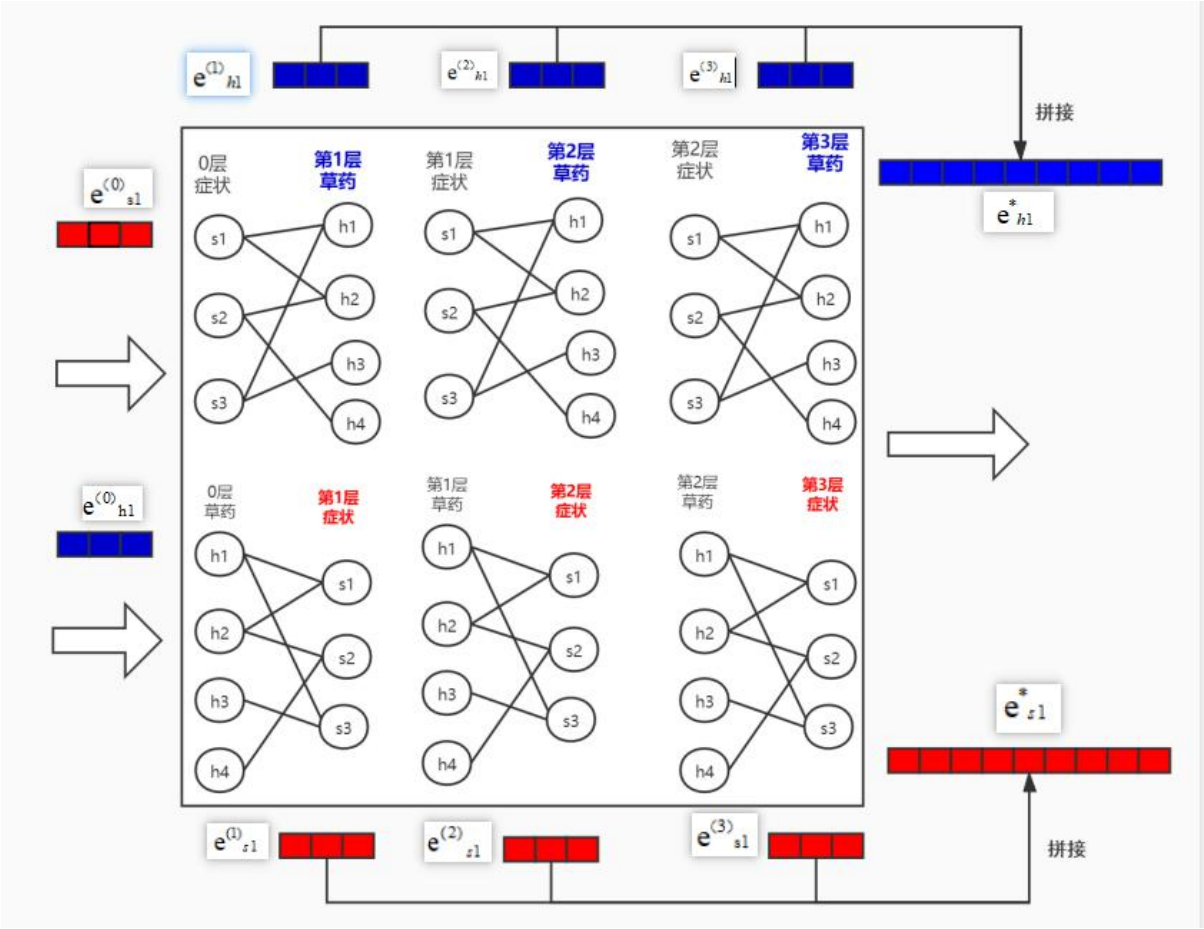

图 3-6　NGCF 模型的整体架构

3.3.3 模型原理

3.3.3.1 嵌入初始化层

NGCF 模型的嵌入初始化层与 SMGCN 相同，即随机初始化所有症状节点和草药节点的嵌入。

3.3.3.2 一阶的嵌入传播

即一层图卷积的过程，分为邻居消息构造、邻居消息聚合两个阶段，我们以第一层草药节点的嵌入传播过程为例来说明。

与 SMGCN 不同的是，这里我们构造其邻居消息时显示地考虑到了反映症状和草药之间治疗关系的交互信息，对于草药节点来说，症状节点要向其传递的消息为：

$$m_{h<-s}{}^{(1)} = f(e_s{}^{(0)}, e_h{}^{(0)}, p_{sh})$$

$$= p_{sh}\ (e_s{}^{(0)}W_1{}^{(1)} + (e_s{}^{(0)} \circ e_h{}^{(0)})W_2{}^{(1)})$$

$$= \frac{1}{\sqrt{|N_s||N_h|}}\ (e_s{}^{(0)}W_1{}^{(1)} + (e_s{}^{(0)} \circ e_h{}^{(0)})W_2{}^{(1)}) \quad (3\text{-}23)$$

其中$W_1{}^{(1)} \in R^{d_0 \times d_1}, W_2{}^{(1)} \in R^{d_0 \times d_1}$, $\circ$代表向量的点积操作，$p_{sh} = \frac{1}{\sqrt{|N_s||N_h|}}$则是代表了当前草药-症状连接重要程度的衰减因子，也就是说当前症状节点和草药节点的邻居数$|N_s|$和$|N_h|$越多，那么当前连接也就越不重要。

然后，我们聚合草药节点和其所有一阶症状邻居节点的信息：

$$e_h{}^{(1)} = Leaky\,Re\,L\,U(m_{h<-h}{}^{(1)} + \sum_{s \in N_h} m^{(1)}{}_{h<-s}) \quad (3\text{-}24)$$

$$m_{h<-h}{}^{(1)} = e_h{}^{(0)}W_1{}^{(1)} \quad (3\text{-}25)$$

对于症状结点整个过程相同，与 SMGCN 不同的是症状和草药节点共享权重参数$W_1{}^{(1)}$和$W_2{}^{(1)}$。

3.3.3.3 高阶嵌入传播

高阶传播即连续堆叠多个一阶嵌入传播层，仍以草药节点为例：

$$e_h{}^{(l)} = Leaky\,Re\,L\,U(m_{h<-h}{}^{(l)} + \sum_{s \in N_h} m_{h<-s}{}^{(l)}) \quad (3\text{-}26)$$

$$m^{(l)}{}_{h<-s} = p_{sh}(e_s{}^{(l-1)}W_1{}^{(l)} + (e^{(l-1)}{}_s \circ e^{(l-1)}{}_h)W_2{}^{(l)}) \quad (3\text{-}27)$$

$$m^{(l)}{}_{h<-h} = e^{(l-1)}{}_h W_1{}^{(l)} \quad (3\text{-}28)$$

此处同 SMGCN，我们仍然使用矩阵形式来执行图上所有节点的多阶嵌入传播：

$$E^{(l)} = Leaky\,Re\,L\,U(m_{self}{}^{(l)} + m_{neigh}{}^{(l)}) \quad (3\text{-}29)$$

$$m_{self}{}^{(l)} = IE^{(l-1)}W_1{}^{(l)} \quad (3\text{-}30)$$

$$m_{neigh}{}^{(l)} = LE^{(l-1)}W_1{}^{(l)} + ((LE^{(l-1)}) \circ E^{(l-1)})W^{(l)}{}_2 \quad (3\text{-}31)$$

$$L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \quad (3\text{-}32)$$

$$A = \begin{bmatrix} O & R \\ R^T & O \end{bmatrix} \tag{3-33}$$

其中 R 为症状-草药治疗关系矩阵（或者说二部图），D 是 A 的对角度矩阵，L 称作二部图的拉普拉斯矩阵。

3.3.3.4 模型预测层

与 SMGCN 不同，NGCF 试图将每个卷积层后获取的嵌入都利用起来以使生成的嵌入包含多阶邻域信息，相比与加权求和、最大池化、LSTM 等复杂手段，我们直接使用维度拼接的简单方式，因为其不会带来额外的待学习参数，即：

$$\begin{cases} e_h{}^* = e_h{}^{(0)} || e_h{}^{(1)} || \ldots || e_h{}^{(l)}, \\ e_s{}^* = e_s{}^{(0)} || e_s{}^{(1)} || \ldots || e_s{}^{(l)} \end{cases} \tag{3-34}$$

然后采用同 SMGCN 的方法来预测得分。

3.3.3.5 模型损失层

NGCF 使用多标签损失的训练过程同 SMGCN，但如果采用 BPR 损失的话，则损失函数中的正样本应对应于症状集合对应的综合症与对应草药集合中的某个草药，负样本应对应于症状集合对应的综合症与对应草药集合之外的某个草药：

$$Loss = \sum_{(sc,hc+,hc-)\in O} - ln\left(e_{syndrome}(sc) \bullet e_{(hc_+)}{}^T - e_{syndrome}(sc) \bullet e_{(hc_-)}{}^T\right) + \lambda \|\theta\|_2{}^2 \tag{3-35}$$

其中$\theta = \{ \quad E^{(0)}, \{W_1{}^{(l)}, W_2{}^{(l)}\}_{l=1}^{k} \quad \}$为模型涉及的所有参数，

$O = \{(sc, hc_+, hc_-) | hc_+ \in hc', hc_- \notin hc', (sc, hc_+) \in R^+, (sc, hc_-) \in R^-\}$为每个处方数据采样的正负样本。

3.4 HetGCN

3.4.1 模型简介

HetGNN[29]模型是为了在 author-paper、paper-paper、paper-venue 组成的异质图上获取 author、paper、venue 三类实体嵌入而提出来的，其主要思想是首先聚合单个节点的异质属性特征、然后聚合同种类型节点的嵌入特征、最后利用注意力机制聚合不同类型节点的嵌入特征。

HetGCN 正是受其启发，在症状-症状、症状-草药、草药-草药图三者拼合起来的异质图

上采用了注意力机制来区分聚合不同类型邻居的信息。

3.4.2 模型整体架构

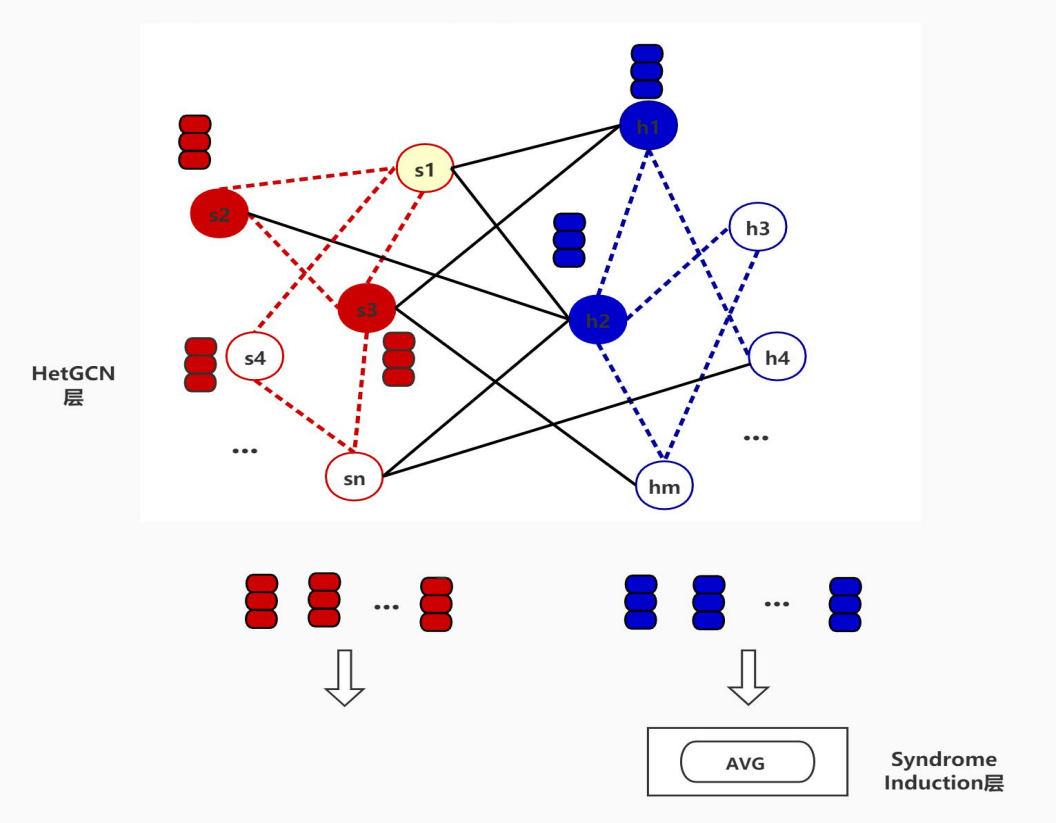

图 3-7 HetGCN 模型的整体架构图

3.4.3 模型原理

与 SMGCN 的 Bipar-GCN 模块原理基本相同，只是引入了 W^{att} 和 z 聚合邻居：

$$b_{N_s}^{(0)} = tanh\left(\sum_{t \in tp} \partial^t \frac{1}{|N_s^t|} \sum_{n \in N^t{}_s} m_n^{(0)}\right) \tag{3-36}$$

$$\partial^t = \frac{exp\left(z^T \, ReLU(W^{att} \bullet (e_s || \frac{1}{|N_s^t|} \sum_{n \in N^t{}_s} m_n^{(0)})\right)}{\sum_{t' \in tp} \exp\left(z^T \, ReLU(W^{att} \bullet (e_s || \frac{1}{|N_s^{t'}|} \sum_{n \in N^{t'}{}_s} m_n^{(0)})\right)} \tag{3-37}$$

此外 SI 部分仅使用平均池化层，同样使用多标签损失损失。

3.5 PinSage

3.5.1 模型简介

PinSage[30]是在图上节点和边的数量达到数亿级别时仍可采用的一种工业量级的方法。其主要提出两点策略，一是对于图上的节点采取包含重启步骤和重要性策略的随机游走采样方

法，能够筛选出真正有意义的边集合;二是对于采样的局部图结构实施连续的多级传播，能够在不同程度上获取到较广邻域内的信息。

本文中批量将处方数据中的症状和草药（元素不重复）分别作为初始的种子节点，然后在双向的症状-草药二部图上通过多跳来获取它们各自的多级同质邻居，最后在这些由“<…<<种子节点-邻居结点>-邻居结点>-...>”组成的多级结构块上进行卷积传播。

本文模型与原始模型的不同之处在于增加 Average SI 并使用多标签损失函数，无需对节点采样正样本、硬性负样本邻居节点来计算 BPR 损失。

3.5.2 模型整体架构

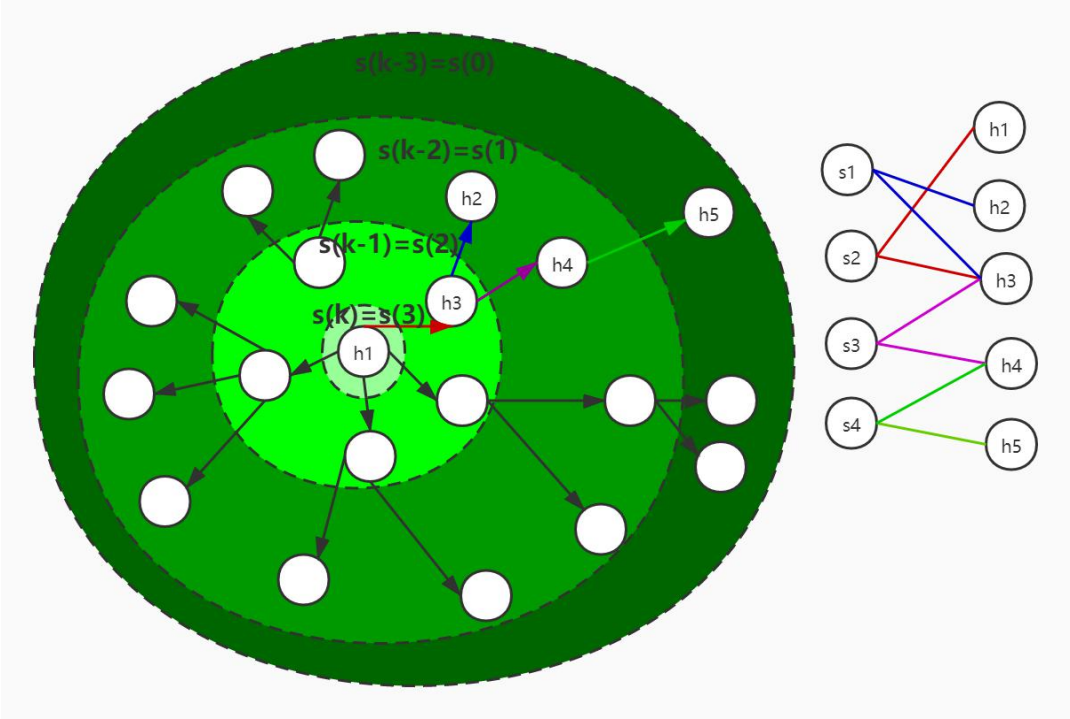

图 3-8 PinSage 在二部图上通过随机游走进行重要性邻居采样的过程示意图

（其中 s（k）代表从外向里的第 k 个子节点集合）

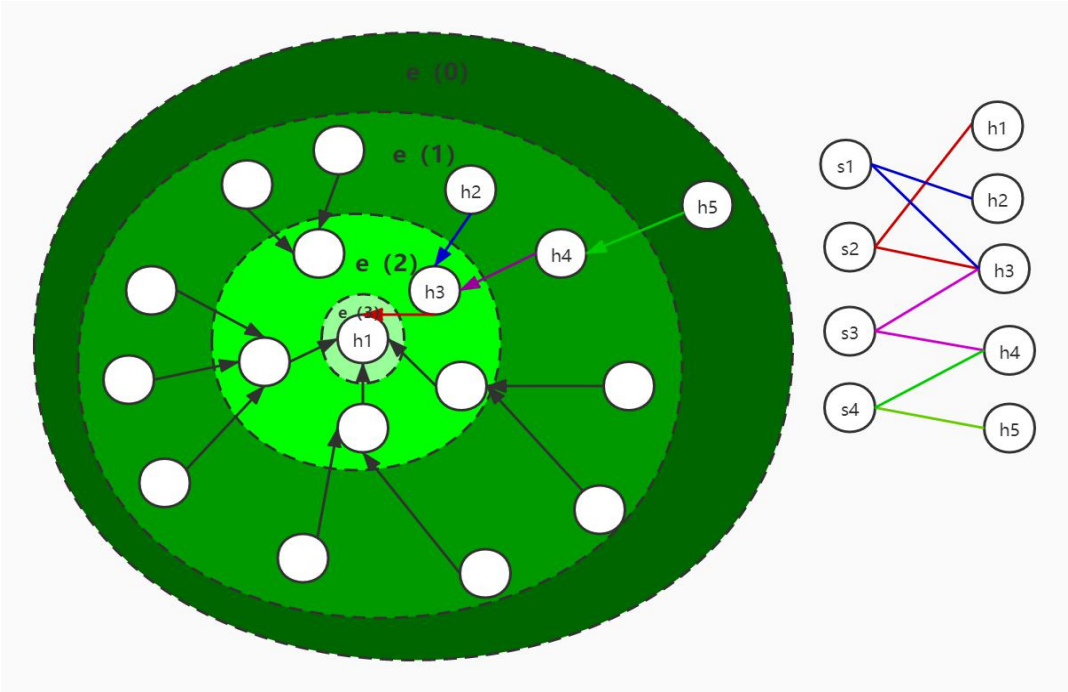

图 3-9 PinSage 模型中信息传播以及节点嵌入的更新过程（最终被更新的节点是 h1）

3.5.3 模型原理

3.5.3.1 嵌入初始化层

PinSage 模型的嵌入初始化层与 SMGCN 相同，即随机初始化所有症状节点和草药节点的嵌入。

3.5.3.2 邻居节点采样层

（1）单层采样过程

以草药为例，种子节点即为初始随机选定的一批节 seeds={h_0,h_1, ...,h_k}。

对于其中的每个节点 h_i (i=1,2,...,k)， 我们执行 num_walk 次长度为 walk_length 的随机游走，每次游走的每一步以一定概率要么回到初始节点、要么沿着边继续走，一共走满 walk_length 步为止。

这样就得到了游走序列 walks={s_0， s_1, s_2, ...， $s_{(walk\_length-1)}$}，然后统计 walks 中每个节点出现的频数并从高到低排序，排序靠前的 num_neigh 个节点即为采样到的邻居节点群。

（2）多层采样

如下图所示，将初始种子节点 seed(0)和采样到的邻居节点 neigh(0) 作为新一轮的种子节点 seed(1)，然后重复上述单层采样结构。

如此从里向外总共重复进行 k 次。从最外层到最里层的节点集依次为 $s^{(0)}$， $s^{(1)}$,...,$s^{(k)}$,且

满足关系

$$s^{(i)} \supset s^{(i+1)},(i = 0,1,\ldots,k-1) \tag{3-38}$$

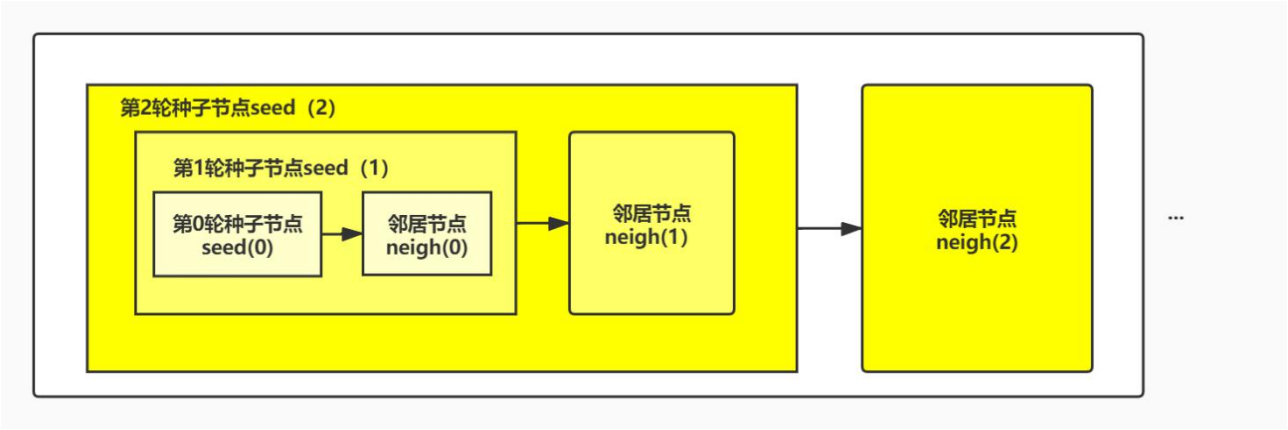

图 3-10 多层采样过程的逻辑示意图

3.5.3.3 多级嵌入传播层

首先，利用嵌入初始化层中的结果获取 s⁽⁰⁾ 中所有节点的嵌入，s⁽⁰⁾ 节点个数记作$N_{s^{(0)}}$：

$$E_h^{(0)} \in R^{N_{s^{(0)}} \times d_0} \longleftarrow e_h^{(0)},\ \forall h \in s^{(0)} \tag{3-39}$$

然后，s⁽⁰⁾ -> s⁽¹⁾ 进行第一层卷积：

1）将 s⁽⁰⁾ 中所有节点的嵌入通过一层神经网络：

$$E_h^{(0)'} = ReLu\left(E_h^{(0)} W_1^{(0)}\right) \in R^{N_{s^{(0)}} \times d_1'},\ \forall h \in s^{(0)}$$

$$其中 W_1^{(0)} \in R^{d_0 \times d_1'} \tag{3-40}$$

2）对于 s⁽¹⁾ 中的每个节点计算加权的邻居信息：

$$E_h^{(1)'} = W_h^{(1)} E_h^{(0)'} \in R^{N_{s^{(1)}} \times d_1'},\ \forall h \in s^{(1)}$$

$$其中 W_h^{(1)} \in R^{N_{s^{(1)}} \times N_{s^{(0)}}} \tag{3-41}$$

3）更新 s⁽¹⁾ 中的每个节点的嵌入：

$$E_h^{(1)} = ReLu\left(E_h^{(1)'} W_2^{(1)}\right) \in R^{N_{s^{(1)}} \times d_1},\ \forall h \in s^{(1)}$$

$$其中 W_2^{(1)} \in R^{d_1' \times d_1} \tag{3-42}$$

其他层同理。

3.5.3.4 模型预测层和损失层

PinSage 的模型预测层和损失层与 SMGCN 完全一致，核心区别只在于获取症状和草药实体嵌入的方式不同。

4 实验

4.1 实验准备

4.1.1 数据准备

（1）数据来源

本文使用的中药处方数据集来自中国工程科技知识中心(CKCEST) [China Knowledge Centre for Engineering Sciences and Technology]，其官网网址是 http://zcy.ckcest.cn/tcm/。其数据内容如表 4-1 所示:

表 4-1 中药数据集内容详情

数据类别	数据文件名	数据说明
实体列表	herbs_contains.txt symptom_contains.txt	均为“一行对应一个草药名或者症状名”的形式，且有行号=编号+1（编号从 0 开始）。草药共计 811 个，症状共计 390 个。
原始处方文本	prescriptions.txt	共包含 98,334 个包含症状集合和草药集合的原始处方。 文本的每一行是一个处方，左边是症状集合，右边是草药集合。
处理过的处方数据	pre_herbs.txt pre_symptoms.txt	共 33,765 经过预处理的处方。pre_herbs 的每一行对应一个处方中草药集合的编号，pre_symptoms 的每一行对应一个处方中症状集合的编号。
处方训练集	pre_herbs_train.txt pre_symptoms_train.txt	同上，但是只包含了 28,746 条用作训练集的处方数据。
处方训练集	pre_herbs_test.txt pre_symptoms_test.txt	同上，但是只包含了 5,019 条用作测试集的处方数据。

（2）数据预处理

这里只需建立“草药编号-草药名”、“症状编号-症状名”、“处方编号-症状集合的症状 id

列表-草药集合的草药 id 列表”三类数据关系(分别由图 4-1、图 4-2、图 4-3 表示，由于数据库中编号只能从 1 开始故图示均比实际 id 大 1)，并用症状-症状（图 4-4）、草药-草药（图 4-5）、症状-草药邻居矩阵（权值为实体在处方中共同出现的次数）构建初步的图。

herb_id	herb_name
1	青木香
2	罗汉松
3	徐长卿
4	荆芥
5	槟目
6	竹叶
7	淡竹叶
8	列当
9	芒硝
10	石决明

图 4-1 草药 id-草药名

symptom_id	symptom_name
1	身热肢寒
2	腹中痞块
3	近视
4	顿咳
5	耳痛
6	气喘
7	咽喉白腐
8	滑胎
9	胎漏
10	健忘

图 4-2 症状 id-症状名

prescrp_id	symptom_sets_id	herb_sets_id	symptom_sets_str	herb_sets_str	train_mask
1	['227']	['55', '71', '169', '275', '295', '395', '649', '657']	['无名肿毒']	['雄黄', '玄参', '乳香', '血竭', '没药', '斑	1
2	['316']	['50']	['大便溏']	['牡蛎']	1
3	['23']	['267', '556', '709', '773']	['小便不利']	['车前子', '罂粟', '罂粟壳', '车前']	1
4	['346']	['554']	['咽喉痛']	['苦参']	1
5	['275']	['13', '201', '703']	['四肢拘急']	['石榴', '天南星', '半夏']	1
6	['358']	['120', '171']	['流泪']	['艾叶', '炉甘石']	1
7	['141', '214']	['138', '175', '182', '548']	['心烦', '烦躁']	['甘草', '细辛', '干姜', '芫花']	1
8	['344']	['124', '138', '175', '360', '426']	['慢惊']	['白矾', '甘草', '细辛', '威灵仙', '僵蚕']	1
9	['275']	['13', '201', '703']	['四肢拘急']	['石榴', '天南星', '半夏']	1
10	['56', '161', '298']	['55', '107', '169', '201', '272', '561', '610', '649',	['腹满', '短气', '恶风']	['雄黄', '白附子', '乳香', '天南星', '天麻	1

图 4-3 处方 id-症状集合 id 列表-草药集合 id 列表

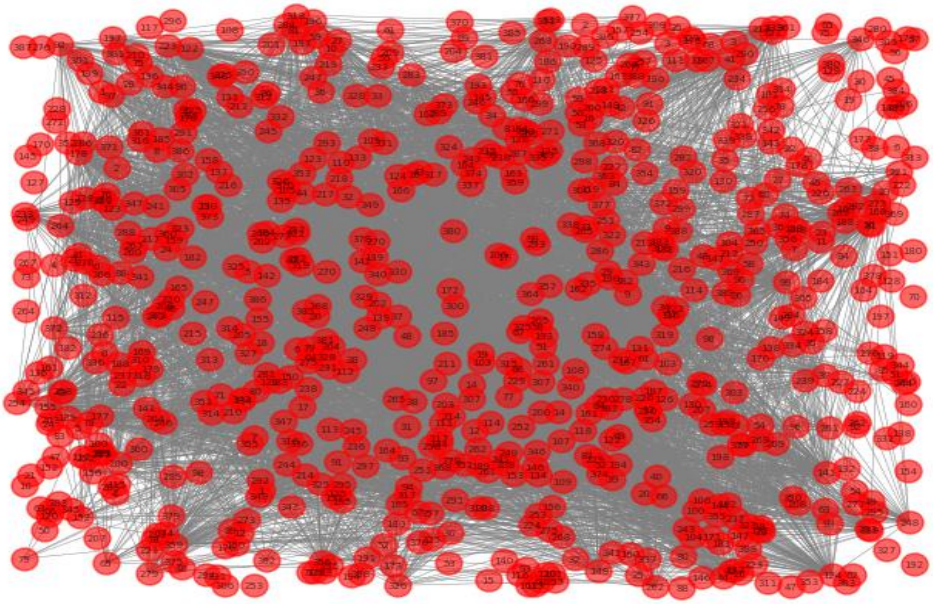

图 4-4 症状-症状初步图的可视化

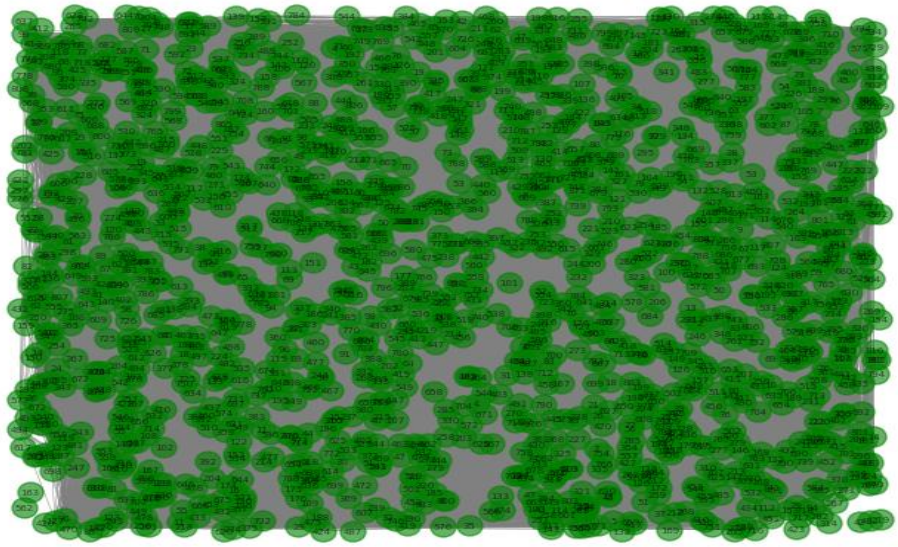

图 4-5 草药-草药初步图的可视化

4.1.2 评价指标

本实验选取草药预测得分中的 top-k（k=5,10,20）个草药作为推荐结果，并将召回率 recall@k、准确率 precision@k、以及归一化的折损累积增益 ndcg@k（其中@是指明 k 值的标识符）作为指标来评估推荐结果的好坏，其计算方式依次为：

$$recall@k = \frac{|top(sympt\_set,k) \cap herb\_set|}{|herb\_set|} \tag{4-1}$$

$$precision@k = \frac{|top(sympt\_set,k) \cap herb\_set|}{k} \tag{4-2}$$

$$ndcg@k = \frac{dcg@k}{idcg@k} \tag{4-3}$$

$$其中\ dcg@k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)} \tag{4-4}$$

$$idcg@k = \sum_{i=1}^{|rel\_num|} \frac{2^{rel_i}-1}{log_2(i+1)} \tag{4-5}$$

其中召回率反映了预测的草药结果的完整性，准确率反映了预测的草药结果的正确性情况，而归一化的折损累计增益则反映了预测的草药结果对不同草药重要性的考量程度。

4.1.3 实验平台

本实验在装有 Nvidia 显卡（cuda11.0）的 unbuntu 服务器上进行，即模型和数据会被同时部署在 GPU 上来加快训练。

实验代码的编写使用“vscode+python 插件+远程服务器连接+jupyter notebook”的方式进行编写，对于 SMGCN、NGCF、HetGCN 模型均使用 tensorflow1.14 进行开发，而对于 PinSage 模型考虑到 dgl API 对游走采样等图上操作的优异支撑性能以及其与 tf1.x 较弱的兼容性，我们转而用 pytorch 1.7.0 进行开发。

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.80.02    Driver Version: 450.80.02    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  GeForce RTX 208...  Off  | 00000000:18:00.0 Off |                  N/A |
| 31%   26C    P8    21W / 250W |      1MiB / 11019MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  GeForce RTX 208...  Off  | 00000000:3B:00.0 Off |                  N/A |
| 31%   26C    P8     1W / 250W |    334MiB / 11018MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  GeForce RTX 208...  Off  | 00000000:5E:00.0 Off |                  N/A |
| 31%   26C    P8     9W / 250W |      1MiB / 11019MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  GeForce RTX 208...  Off  | 00000000:86:00.0 Off |                  N/A |
| 31%   26C    P8    14W / 250W |      1MiB / 11019MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
```

图 4-6 服务器的显卡配置

4.2 对照实验

4.2.1 实验设置

我们设置相同的 sympt_threshold（5）和 herb_threshold（40）、相同的初始嵌入维度（64）、相同的各级隐藏层和输出层维度（PinSage 模型要求输入和输出维度必须一致，其为【64,128】、【128，64】，而其余模型均为【64,128】，【128，256】），然后在预处理过的中药处方训练数据集和测试数据集上进行实验。

不同模型为了收敛到各自的最优结果，在迭代训练次数、学习率、正则强度、节点丢弃率上可以有所区别，根据多次参数调整其最终参数设定分别如下表所示：

表 4-2 各模型的相关参数设定

模型	学习率 lr	正则强度λ	节点丢弃率 dropout
SMGCN	2e-4	7e-3	0.0
NGCF	2e-3	7e-3	0.0
HetGCN	2e-3	7e-3	0.0
PinSage	0.09	1e-3	0.0

4.2.2 实验结果

（1）各模型在测试集上的推荐效果对比

表 4-3 各模型在测试指标上的得分情况

模型	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndgc@20
PinSage	0.22012	0.35498	0.53886	0.30317	0.24754	0.19024	0.38386	0.45925	0.57925
NGCF	0.20385	0.31816	0.46565	0.28193	0.22347	0.16519	0.37875	0.45493	0.55811
HetGCN	0.20238	0.31838	0.46731	0.28037	0.22325	0.16549	0.37794	0.45523	0.55910
SMGCN	0.20760	0.32288	0.47376	0.28703	0.22700	0.16819	0.38669	0.46142	0.56586

表 4-4 SMGCN 模型相比于其他模型在各个测试指标上的提升度（负值表示降低）

模型	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndgc@20

SMGCN	0.20685	0.32451	0.47259	0.28603	0.22734	0.16767	0.38467	0.46118	0.56456
By HetGCN	2.21%	1.93%	1.13%	2.02%	1.83%	1.32%	1.78%	1.31%	0.98%
By NGCF	1.84%	1.48%	1.74%	1.81%	1.58%	1.82%	2.10%	1.43%	1.39%
By PinSage	-5.69%	-9.04%	-12.08%	-5.32%	-8.30%	-11.60%	0.74%	0.47%	-2.31%

（2）各模型的单位参数训练时长对比

表 4-5 各模型的参数训练数据

模型	参数总量	运行时间	单位参数的训练时长
SMGCN	381888	5030.7s	0.01317s
HetGCN	363968	9076.5s	0.02494s
NGCF	159552	5684.6s	0.03563s
PinSage	134656	18493.7s	0.13734s

（2）各模型训练过程中的损失变化曲线对比

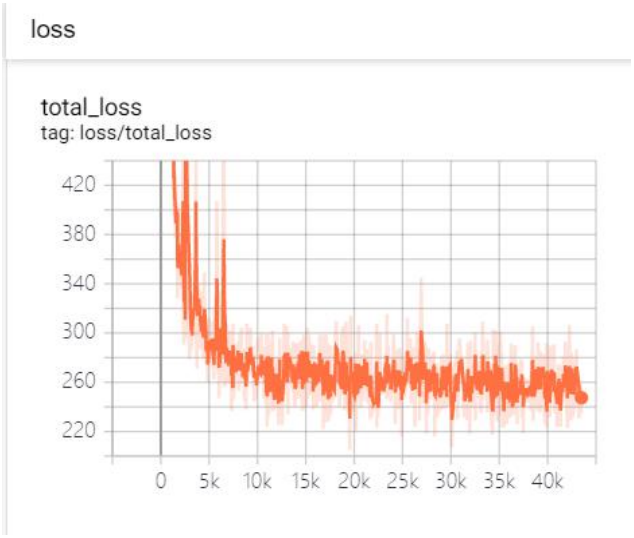

图 4-7 SMGCN 模型的损失变化曲线

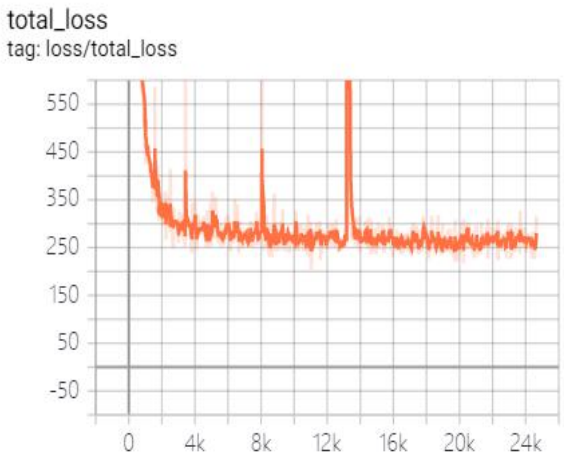

图 4-8 NGCF 模型的损失变化曲线

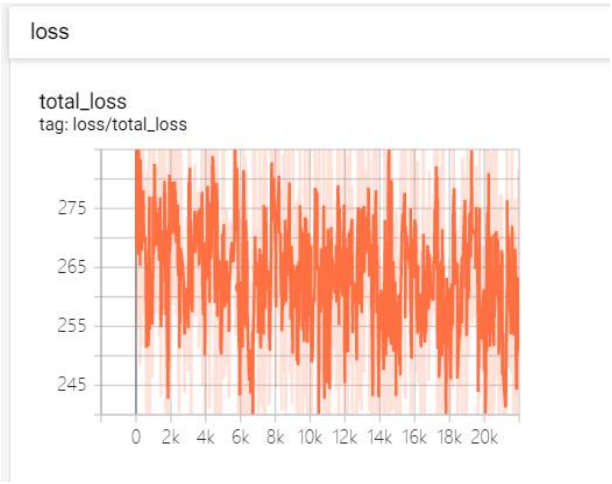

图 4-9 HetGCN 模型的损失变化曲线

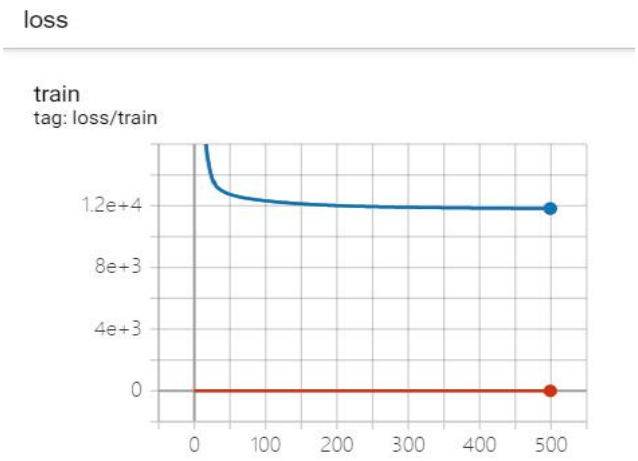

图 4-10 PinSage 模型的损失变化曲线

4.2.3 实验结果分析

（1）表 4-3、4-4 结果分析

可以得知模型的效果有 PinSage > SMGCN > NGCF > HetGCN，总体上各模型效果差异并不明显，但是放在节点数目数以亿计的实际推荐业务即使 0.001%的提升也会影响到 1000 个推荐结果的排序。

加入了额外协同图 SGE 和综合症推导 SI 组件的 SMGCN 模型相比于传统模型 NGCF 和 HetGCN 确有了提升，这间接证明了 SMGCN 组件的合理性。

SMGCN 的效果仍不如传统工业推荐策略 PinSage，这说明基于随机游走的动态邻居方法相比于基于静态邻居的方法能挖掘出更多潜在的协同信号。

（2）表 4-5 结果分析

明显可以发现 SMGCN 模型参数量最大，训练时间反而最短，单位参数的训练时长最小，而 PinSage 模型参数量最小，训练时间反而最长，单位参数的训练时长最大。而其余的两种模型 HetGCN 和 NGCF，单位参数的训练时长略差于 SMGCN，但都大幅好于 PinSage。

在需要快速训练推荐模型的场景中，SMGCN 模型既能够保证速度也能够保证训练结果的质量，理应成为首选。其次依次是 HetGCN、NGCF、PinSage。

（3）图 4-7、4-8、4-9、4-10 分析

HetGCN 模型的损失曲线反复震荡，可以看出在较早迭代次数时就已经收敛了，其余模型的损失曲线都是稳步下降，且 PinSage 模型的损失下降最为平滑。可以观察出 NGCF、HetGCN、PinSage 模型基本都在 10k 次时开始逐渐收敛，而 PinSage 在 100 时开始收敛，但

是由于 PinSage 训练集上并没有采用批处理方式，所以收敛速度此处无法比较。

4.3 消融实验

4.3.1 实验设置

本实验主要验证 Bipar-GCN、SGE 以及 SI 三个核心组件的功能，为此可能的组合包括“SMGCN 完整模型（SMGCN）、SMGCN 模型去掉 Bipar-GCN（SGE w SI）、 SMCGN 模型去掉 SGE（Bipar-GCN w SI）、MGCN 模型去掉 SI（Bipar-GCN w SGE）、SMGCN 模型去掉 Bipar-GCN 和 SGE（不合理）、SMGCN 模型去掉 Bipar-GCN 和 SI（SGE）、 SMGCN 模型去掉 SGE 和 SI（Bipar-GCN）”，在这些模型上进行实验以进行 SMGCN 模型上的消融验证。此外我们还在传统模型添加了 SI 以进一步探究其作用的迁移性情况。

整个消融实验的思路概括如下：

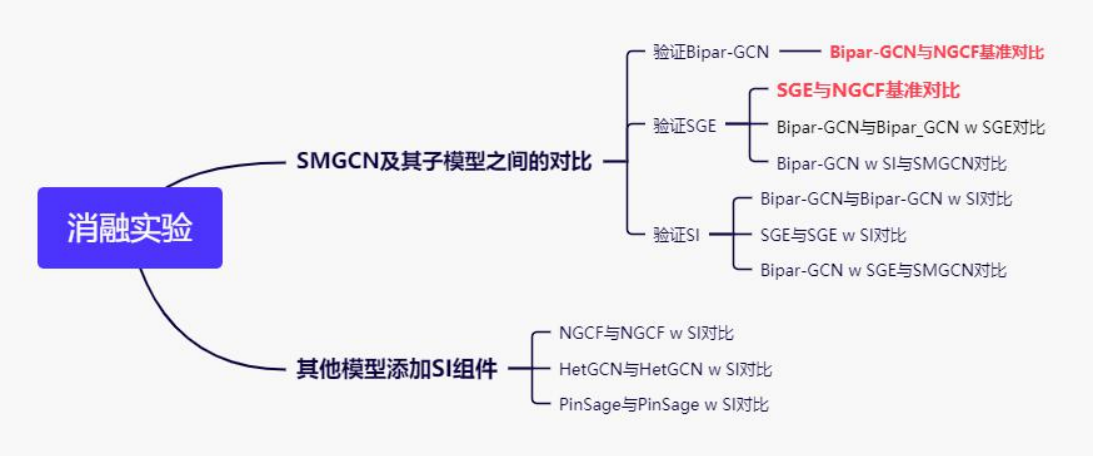

图 4-11 消融实验的整体思路

4.3.2 实验结果

（1）SMGCN 模型与其他各组合子模型在测试集上的推荐效果对比

表 4-6 SMGCN 模型与其他子模型的测试得分情况

submodel	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
NGCF	0.20385	0.31816	0.46565	0.28193	0.22347	0.16519	0.37875	0.45493	0.55811
Bipar-GCN	0.09855	0.15405	0.27177	0.13899	0.10857	0.09831	0.23603	0.29606	0.40302

Bipar-GCN w SGE	0.15921	0.25349	0.37282	0.22837	0.18368	0.13609	0.33657	0.41566	0.50879
Bipar-GCN w SI	0.10655	0.20401	0.29874	0.15409	0.14891	0.11131	0.22023	0.32484	0.40745
SMGCN	0.20760	0.32288	0.47376	0.28703	0.22700	0.16819	0.38669	0.46142	0.56586
SGE	0.18585	0.29468	0.42940	0.26356	0.20972	0.15448	0.36326	0.44150	0.54060
SGE w SI	0.20429	0.31731	0.46245	0.28356	0.22439	0.16561	0.38381	0.45796	0.55942

（2）在其他模型上添加 SI 组件后与原模型的对比

表 4-7 传统模型添加 SI 后的效果对比

模型	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndgc@20
NGCF	0.20385	0.31816	0.46565	0.28193	0.22347	0.16519	0.37875	0.45493	0.55811
NGCF w SI	0.20644	0.32182	0.46900	0.28448	0.22499	0.16646	0.38300	0.45864	0.56180
HetGCN	0.20238	0.31838	0.46731	0.28037	0.22325	0.16549	0.37794	0.45523	0.55910
HetGCN w SI	0.18744	0.29941	0.44438	0.26627	0.21448	0.15967	0.35897	0.43717	0.54170
PinSage	0.22012	0.35498	0.53886	0.30317	0.24754	0.19024	0.38386	0.45925	0.57925
PinSage w SI	0.21693	0.35089	0.53164	0.30201	0.24612	0.18872	0.38334	0.45875	0.57674

4.3.3 实验结果分析

（1）表分析结果

Bipar-GCN 在各项指标上的结果都近似于 NGCF 基准的一半，而 SGE 则与 NGCF 基准的效果都只相差 0.02 左右，这说明 SGE 组件相比 Bipar-GCN 包含了更有价值的协同信息，通过多跳邻居获得的协同信号可能强度不够。

Bipar-GCN　w SGE 在 Bipar-GCN 的基础上添加了 SGE，对应指标得分提升了约 57.15%；

SMGCN 在 Bipar-GCN w SI 的基础上添加了 SGE，对应指标得分提升了 85.56%。这都证明了 SGE 组件有力的嵌入增强能力。

表 4-8 添加 SGE 组件后的模型效果对比

	r@5	p@5	ndcg@5	Avg
Bipar-GCN	0.09855	0.13899	0.23603	
Bipar-GCN w SGE	0.15921	0.22837	0.33657	
Improved.	**64.55%**	**64.31%**	**42.60%**	**57.15%**
Bipar-GCN w SI	0.10655	0.15409	0.22023	
SMGCN	0.20760	0.28703	0.38669	
Improved.	**94.84%**	**86.27%**	**75.58%**	**85.56%**

Bipar-GCN w SI 在 Bipar-GCN 的基础上添加了 SI，对应指标得分提高了约 4.10%；SGE w SI 在 SGE 的基础上添加了 SI，对应指标得分提高了约 7.72%；SMGCN 在 Bipar-GCN w SGE 的基础上添加了 SI，对应指标得分提高了约 23.66%。这些证明了 SI 有相对较弱的嵌入增强能力，远不如 SGE。

表 4-9 添加了 SI 组件后的模型效果对比

	p@5	r@5	ndcg@5	Avg
Bipar-GCN	0.09855	0.13899	0.23603	
Bipar-GCN w SI	0.10655	0.15409	0.22023	
Improved.	**8.12%**	**10.86%**	**-6.70%**	**4.10%**
SGE	0.18585	0.26356	0.36326	
SGE w SI	0.20429	0.28356	0.38381	
Improved.	**9.92%**	**7.59%**	**5.66%**	**7.72%**
Bipar-GCN w SGE	0.15921	0.22837	0.33657	
SMGCN	0.20760	0.28703	0.38669	
Improved.	**30.40%**	**25.69%**	**14.90%**	**23.66%**

（2）表分析结果

可以看出只有 NGCF 模型添加了 SI 后推荐效果得到了提升，HetGCN 模型与 PinSage 模型则反而略有降低。

推测可能的原因是：SMGCN 中包含 Bipar-GCN 和 SGE 两种类型的组件来获取嵌入，且其相比于 Bipar-GCN w SGE 提高了约 23.66%，可以推测出 MLP 对简单的嵌入相加和进行了很好的非线性融合；NGCF 中是把每个卷积层获取的嵌入拼接起来作为最终嵌入，MLP 能很好地混合不同层的信息而使得推荐效果同样得到提高；而 HetGCN 和 PinSage 都是把最后一层卷积层的嵌入作为最终嵌入，没有需要进行非线性融合的其他嵌入对象，所以 MLP 反而可能会扰动原有嵌入中的特征信息。

4.4 超参数性能影响实验

4.4.1 实验设置

我们将可能涉及到的超参数按照 SMGCN 模型组件进行分类，得到如下图所示的实验思路：

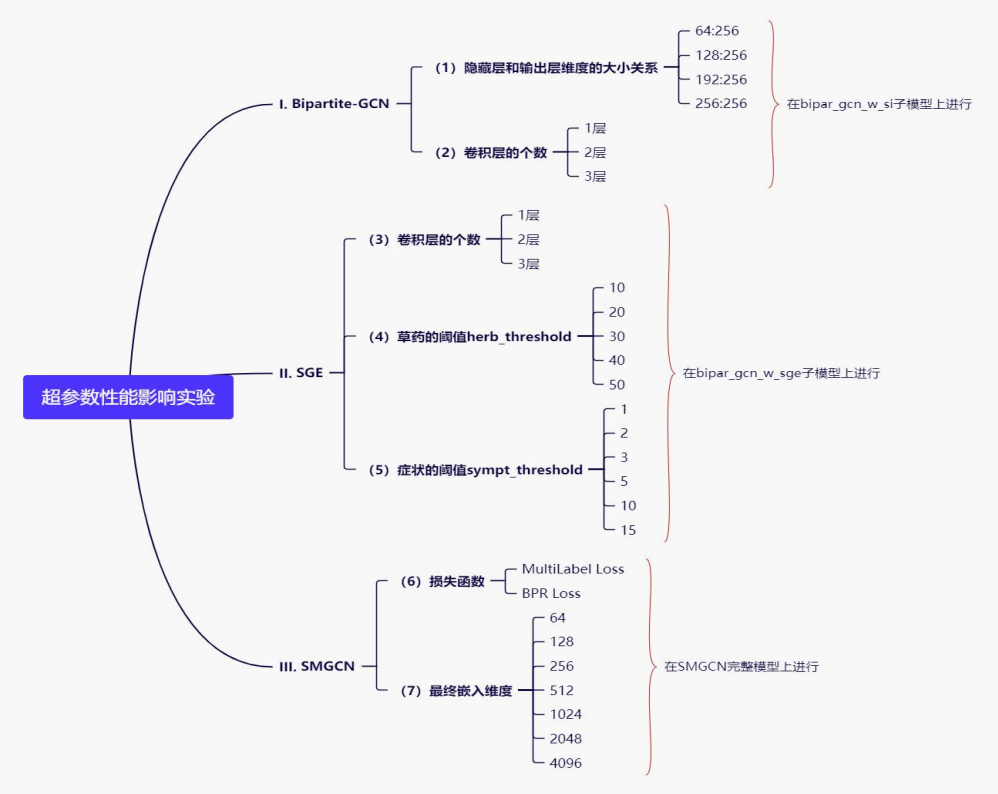

图 4-12 超参数实验的完整思路

其中的编号代表了实验的顺序，每个参数实验都能得到当前所研究的超参数的最佳值。

这里要求下一个参数实验总是将前面所有参数实验中涉及的超参都设定为它们的最佳值，因为这样能将前面实验中所有参数对当前实验的影响程度降到最低。但是同时由于多种参数的共同作用难以相互区分，参数的研究顺序不必严格。

关于最佳超参的选择策略，我们假定 k 的优先级为 5>10>20，指标的优先级为 recall>precision>ndcg。我们首先在 r@5、p@5、ndcg@5 上进行比较，求得 3 个指标上得分最高值的相应参数值，其中出现频率最高的参数值将作为最优超参值；若不能区分则在下一级 r@10、p@10、ndcg@10 上比较，依次类推。

4.4.2 实验结果

按超参名列出实验结果如下，其中表格中的黑色加粗参数为该轮超参实验的最佳超参值，红色数值为 r@5、p@5、ndcg@5 这些指标上各自的最大值：

（1）隐藏层和输出层维度的大小关系

表 4-10 Bipart-GCN 中不同 hid_output_ratio 对应的测试得分

hidden_s：		r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
1:4	64:256	0.14296	0.21013	0.33579	0.20610	0.15597	0.12262	0.32716	0.38770	0.48841
2:4	128:256	0.14546	0.21444	0.33489	0.21427	0.16111	0.12275	0.29746	0.36116	0.45699
3:4	192:256	0.14546	0.21824	0.33450	0.21427	0.16051	0.12247	0.33155	0.39910	0.49189
4:4	**256:256**	0.14581	0.21524	0.33503	0.21423	0.15943	0.12274	0.33240	0.39557	0.49121

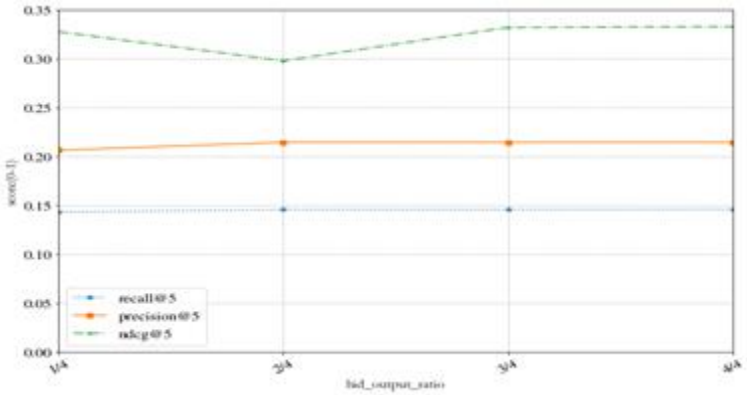

图 4-13 推荐效果关于 hit_output_ratio 因子的变化曲线

（2）Bipar-GCN 中卷积层的个数

表 4-11 Bipart-GCN 中不同 bipartite_gcn_num_layers 对应的测试得分

Depth		r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
1	**[256]** **[256]**	0.14581	0.21524	0.33503	0.21423	0.15943	0.12274	0.33240	0.39557	0.49121
2	[128,256] [128,256]	0.12813	0.17809	0.26954	0.18581	0.12835	0.09792	0.30722	0.36013	0.44572
3	[64,128,256] [64,128,256]	0.05858	0.09651	0.16170	0.08296	0.06760	0.05628	0.14325	0.19794	0.27140

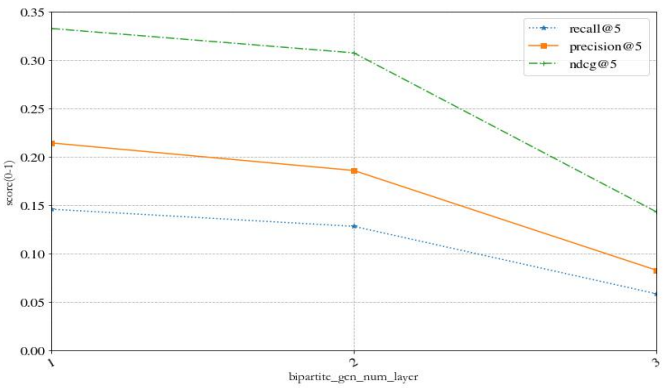

图 4-14 推荐效果关于 bipartite_gcn_num_layer 因子的变化曲线

（3）SGE 中卷积层的个数

表 4-12 SGE 中不同 sge_num_layers 对应的测试得分

Depth	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
1	0.16822	0.2661	0.40210	0.24108	0.19265	0.14512	0.34441	0.42143	0.52270
2	0.12053	0.1794	0.26298	0.17462	0.13140	0.09655	0.31133	0.37396	0.45088
3	0.13293	0.1990	0.30105	0.19362	0.14527	0.10944	0.31662	0.38067	0.46920

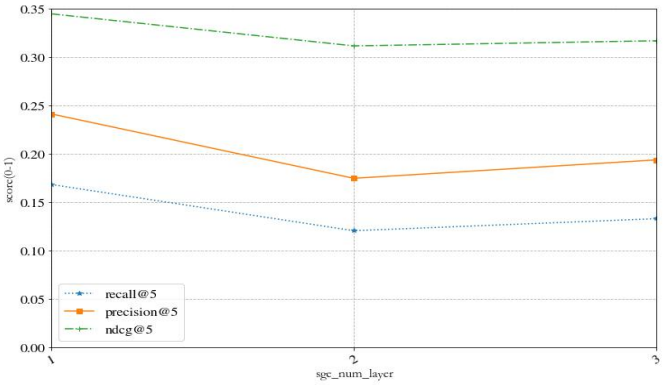

图 4-15 推荐效果关于 sge_num_layer 因子的变化曲线

（4）SGE 中草药阈值的设定值（此处 sympt_threshold 固定为 5）

表 4-13 SGE 中不同 herb_threshold 对应的测试得分

X_h	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@	ndcg@10	ndcg@20
10	0.15803	0.25864	0.38617	0.22861	0.18526	0.13895	0.33650	0.41947	0.51648
20	0.15703	0.25380	0.38260	0.22738	0.18296	0.13769	0.33387	0.41346	0.51164
30	0.16641	0.26610	0.39937	0.23786	0.19099	0.14333	0.34376	0.42315	0.52285
40	0.16822	0.26618	0.40210	0.24108	0.19265	0.14512	0.34441	0.42143	0.52270
50	0.16715	0.27071	0.39749	0.23833	0.19362	0.14432	0.34655	0.42851	0.52570

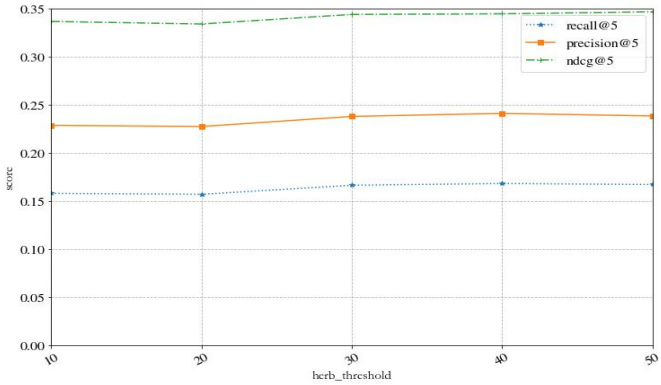

图 4-16 推荐效果关于 herb_threshold 因子的变化曲线

（5）SGE 中症状阈值的设定值（此处 herb_threshold 固定为 40）

表 4-14 SGE 中不同 sympt_threshold 对应的测试得分

X_s	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
1	0.13388	0.20310	0.31516	0.19621	0.15005	0.11576	0.32536	0.39339	0.48741
2	0.16305	0.25786	0.39094	0.23339	0.18619	0.14152	0.34052	0.41621	0.51649
3	0.15778	0.25543	0.38645	0.22650	0.18402	0.13959	0.32857	0.40900	0.50941
5	0.16163	0.26441	0.39827	0.23291	0.19221	0.14428	0.33558	0.41768	0.51851
10	0.16405	0.25900	0.39160	0.23682	0.18838	0.14226	0.34204	0.41837	0.51891
15	0.16187	0.26264	0.40342	0.23606	0.19247	0.14605	0.33947	0.41882	0.52235

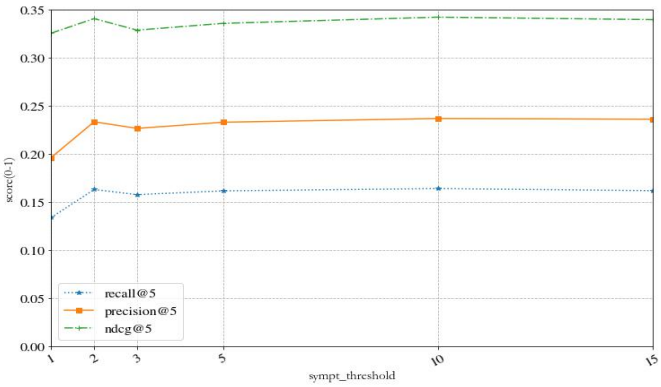

图 4-17 推荐效果关于 sympt_threshold 因子的变化曲线

（6）损失函数（多标签损失和正负样本损失）

表 4-15 SMGCN 和 NGCF 中 BPR Loss 和 MultiLable Loss 对应的测试得分

模型	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
SMGCN BPR	0.20375	0.32109	0.46952	0.27352	0.21978	0.16269	0.37428	0.45658	0.56104
SMGCN MultiLable	0.20728	0.32458	0.47111	0.28675	0.22811	0.16767	0.38590	0.46248	0.56445
NGCF BPR	0.20193	0.31736	0.47163	0.26966	0.21652	0.16229	0.36747	0.44933	0.55692
NGCF MutliLable	0.20499	0.32069	0.46783	0.28380	0.22461	0.16586	0.38221	0.45726	0.56040

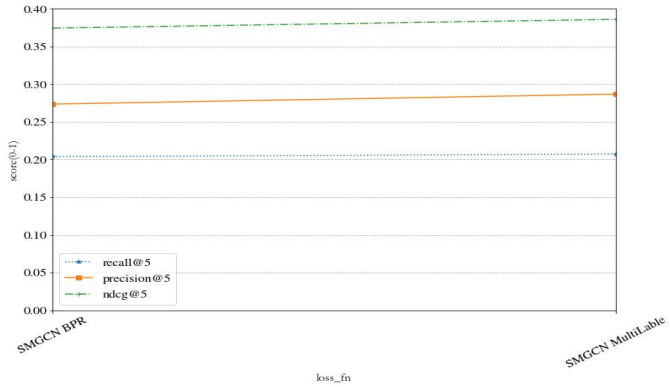

图 4-18 SMGCN 的推荐效果关于损失函数因子的变化曲线

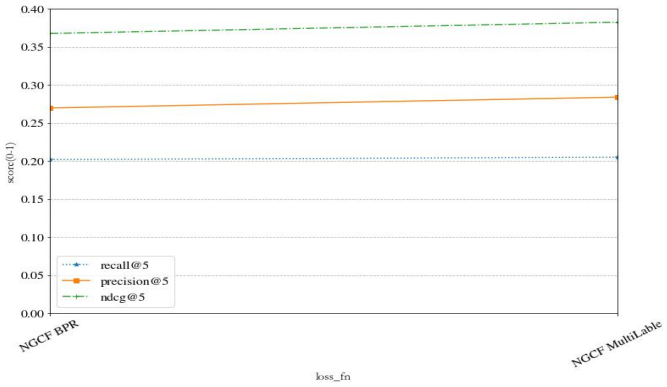

图 4-19 NGCF 的推荐效果关于损失函数因子的变化曲线

（7）最终的嵌入维度

表 4-16 SMGCN 中不同 final_embed_dim 因子对应的测试得分

dimension	r@5	r@10	r@20	p@5	p@10	p@20	ndcg@5	ndcg@10	ndcg@20
64	0.20050	0.31869	0.46121	0.28057	0.22443	0.16508	0.37703	0.45532	0.55639
128	0.20027	0.31674	0.46311	0.27942	0.22359	0.16499	0.37798	0.45623	0.55842
256	0.20617	0.32357	0.47217	0.28504	0.22666	0.16826	0.38344	0.45929	0.56304
512	0.20914	0.32704	0.47364	0.28755	0.22829	0.16910	0.38722	0.46266	0.56548
1024	0.20994	0.32786	0.47387	0.28643	0.22871	0.16809	0.38783	0.46364	0.56516
2048	0.20937	0.32560	0.46855	0.28902	0.22893	0.16740	0.38846	0.46395	0.56310
4096	0.20480	0.32269	0.46741	0.28344	0.22644	0.16668	0.38308	0.45921	0.56128

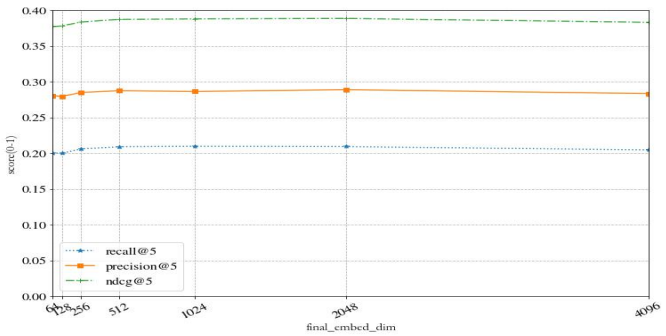

图 4-20 推荐效果关于 final_embed_dim 因子的变化曲线

4.4.3 实验结果分析

将每个超参实验的最佳值和结果分析概括为下表 4-17，从表中可知图卷积层的个数、最终嵌入维度、症状阈值和草药阈值这四个超参最影响 SMGCN 的性能：

表 4-17 超参实验的结果汇总

超参	含义	最佳超参值	分析
hid_output_ratio	Bipar-GCN 中隐藏层和输出层的维度大小关系	4/4	推荐效果总体随 hid_output_ratio 的增大而增强，除了 2/4 时会突然减弱
bipartite_gcn_num_layer	Bipar-GCN 中 GCN 的层数	1	推荐效果总体随着层数增加而不断减弱，且层数越多减弱得越快，在本问题

			中 2 层时已过拟合
sge_num_layer	SGE 中 GCN 的层数	1	推荐效果随着 GCN 层数的增加先减弱后增强，总体趋势还是减弱
herb_threshold	SGE 中的草药阈值	40	推荐效果随着草药阈值先不断增强后减弱，在 40 前模型欠拟合，40 后模型过拟合
sympt_threshold	SGE 中的症状阈值	10	推荐效果总体随着症状阈值先不断增强后减弱，在 10 前模型欠拟合，10 后模型过拟合
loss_fn	SMGCN、NGCF 的损失函数	MultiLabel	SMGCN 和 NGCF 使用 MultiLabel 损失函数均要比 BPR 效果要好
final_embed_dim	SMGCN 的最终嵌入维度	2048	推荐效果随着最终嵌入维度先不断增强后减弱，在维度为 4096 时开始过拟合

5 中药推荐的可视化系统

5.1 系统功能

5.1.1 系统功能模块图

本系统主要由数据集展示、中药推荐两个核心功能模块构成。

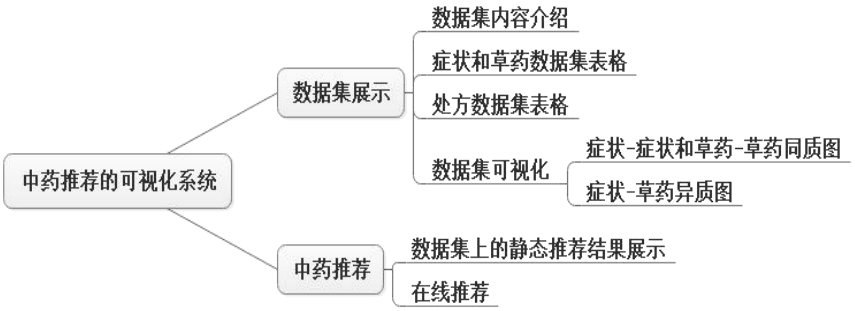

图 5-1 中药推荐系统的功能模块图

5.1.2 模块说明

这两个模块的具体实现都立足于上一章的实验结果。从数据集中构建同质图和异质图有利于理解图模型的概念以及问题的复杂性，而中医推荐模块则以一种更清晰的方式展示了本文中草药推荐问题的含义和各个算法的执行效果。

关于静态中药推荐模块的实现，我们并非直接保存所有的推荐结果，因为只有所有训练迭代结束我们才能得到最优参数，在每一轮测试就保存推荐结果将造成大量数据冗余也可能没有足够的空间保存。我们转而也是将其在线上进行运算。

关于在线推荐，我们也没有采用传统方法，即通过模型检查点文件加载最优参数后再在后台实时计算。一是本问题中的模型实现同时包括 tf1.14 和 pytorch 两种框架，而它们是不便于在系统环境中同时出现的；二是算法代码部署在远程服务器而网页在本地开发，两者难以集成。鉴于此问题只需获取最终症状和草药嵌入以及 MLP 权重参数的特殊性，我们设置了一个全局的 recall 最优值，并在第一轮训练时用测试结果的 recall 更新它并保存相关的嵌入和参数文件，当后续测试结果的 recall 好于全局最优，我们同样更新它并重写同名的嵌入和参数文件。最后线上时直接在后台读取这些矩阵文件进行计算即可，推荐速度更为快速。

5.2 系统运行环境及开发技术

5.2.1 系统开发技术

本文使用“django 后端 + jquery 前端”的方式进行系统开发。

5.2.2 系统运行环境

本文系统配置在 django 框架自带的后台服务器上，页面在 chrome 服务器上本地运行。本机电脑为 win10 系统，内存为 8G。

5.3 系统运行实例

5.3.1 中药数据集页面的功能展示

（1）中药数据集的简介

标注中药数据集 TCM 的版权并叙述其基本内容，方便使用者了解平台的开发背景

图 5-2 TCM 数据集简介页面

（2）数据集中症状和草药数据表

用可翻页的表格直观展示症状列表和草药列表

图 5-3 症状和草药列表页面

（2）数据集中的处方数据表

包含训练集和测试集数据，训练集编号全部在测试集之前

图 5-4 处方列表页面

（3）数据可视化

使用 dgl 自带函数导出的草药-草药、症状-症状的同质图结构，较数据处理部分的图简洁，但能直观展示出孤立的草药和症状节点。

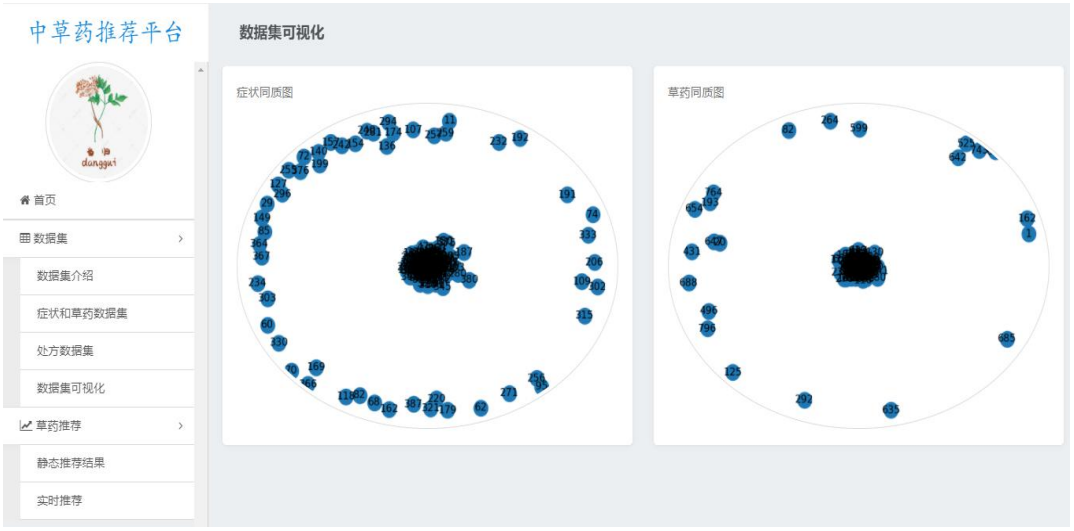

图 5-5 数据集可视化页面

5.3.2 中药推荐页面的功能展示

（1）数据集上的静态推荐结果

下面的图 5-6、图 5-7、图 5-8 分别展示了在处方数据集上采用 top-5、top-10、top-20 策略的推荐结果，其中标记为红色的草药代表其出现在真实草药集合中，即推荐正确。

可以看出在处方上各个算法的预测结果差距不大，部分甚至完全一致只是顺序不同，比如图 5-6 中的 20 号处方；SMGCN 模型能够预测出包括 PinSage 在内的其他模型预测不出来的草药，比如图 5-6 中对 21 号处方预测出了“香附”、图 5-7 中对 33742 号处方预测出了“砂仁”以及对 33744 号处方预测出了“薄荷”、图 5-8 中对 30 号处方预测出了“僵蚕”；但是 SMGCN 模型在部分症状集合上的效果却远不如其他模型，比如图 5-9 中 33747 号处方上 SMGCN 只预测对 1 个，而其他模型均能预测对 4 个，这也解释了为什么其评价指标总体上反而会比 PinSage 差。

图 5-6 top-5 推荐实例

ID	症状集	真实草药集	SMGCN	HetGCN	NGCF	PinSAGE
33742	四肢麻木	菊花 红花 白术 姜黄 五加皮 玉竹 陈皮 牛膝 木瓜 丁香 茯苓 党参 五加 砂仁 檀香 木香 麦冬	白术 当归 杜仲 香附 砂仁 牛膝 川芎 厚朴 红花 苦瓜	防风 独活 羌活 牛膝 木瓜 川芎 草乌 当归 川乌 秦艽	当归 防风 独活 羌活 牛膝 木瓜 川乌 红花 草乌 甘草	当归 防风 甘草 川芎 羌活 牛膝 三棵针 独活 麻黄 白术
33743	白带	鹤草芽	薄荷 黄芩 连翘 虻床子 石膏 杜仲 牛膝 荆芥 木通 肉苁蓉	当归 白芍 白术 川芎 甘草 香附 茯苓 人参 木香 干姜	当归 甘草 白术 茯苓 白芍 人参 川芎 香附 陈皮 干姜	当归 白术 白芍 茯苓 甘草 川芎 香附 木香 人参 陈皮
33744	头痛	薄荷	珍珠母 夜交藤 薄荷 牛膝 佛手 珍珠 山楂 合欢皮 柴胡 香附	甘草 防风 川芎 石膏 黄芩 细辛 麻黄 羌活 白芷 柴胡	甘草 防风 川芎 石膏 黄芩 细辛 麻黄 白芷 当归 茯苓	甘草 石膏 川芎 防风 黄芩 茯苓 柴胡 白芷 细辛 当归

首页1 上一页11247 Page 11248 of 11255. 下一页11249 尾页11255

图 5-7 top-10 推荐实例

ID	症状集	真实草药集	SMGCN	HetGCN	NGCF	PinSAGE
28	鼻塞咳嗽	生姜 甘草 杏仁 阿胶 乌梅 罂粟 半夏 罂粟壳 杏	甘草 杏仁 杏 桔梗 半夏 茯苓 麻黄 防风 陈皮 桑 川芎 人参 荆芥 黄芩 前胡 枳壳 薄荷 白芷 细辛 羌活	甘草 杏 杏仁 桔梗 麻黄 半夏 防风 川芎 细辛 白芷 茯苓 石膏 人参 薄荷 黄芩 荆芥 桑 陈皮 羌活 当归	甘草 杏 杏仁 桔梗 麻黄 细辛 防风 半夏 川芎 茯苓 白芷 薄荷 荆芥 桑 人参 陈皮 黄芩 石膏 羌活 当归	甘草 杏 杏仁 桔梗 川芎 半夏 茯苓 细辛 防风 麻黄 人参 桑 黄芩 薄荷 白芷 陈皮 荆芥 羌活 黄皮 紫菀
29	吞酸胁痛咽干	沙参 当归 北沙参 麦冬	甘草 茯苓 人参 白术 半夏 陈皮 木香 厚朴 香附 砂仁 当归 神曲 青皮 黄芩 白芍 丁香 干姜 豆蔻 柴胡 桔梗	甘草 人参 茯苓 白术 半夏 陈皮 当归 木香 厚朴 黄芩 香附 白芍 黄连 枳壳 砂仁 桔梗 槟榔 柴胡 干姜 大黄	甘草 茯苓 半夏 人参 陈皮 白术 当归 木香 黄芩 厚朴 香附 黄连 桔梗 枳壳 砂仁 白芍 柴胡 槟榔 青皮 大黄	甘草 茯苓 陈皮 半夏 当归 木香 人参 白术 白芍 柴胡 黄芩 枳壳 干姜 香附 桔梗 厚朴 黄连 丁香 青皮 槟榔
30	急惊疳积积滞	全蝎 僵蚕 巴豆 巴豆霜 麝香	麝香 木香 朱砂 黄连 槟榔 陈皮 神曲 巴豆 茯苓 青皮 甘草 全蝎 使君子 三棱 芦荟 麦芽 牛黄 胡黄连 附子 僵蚕	麝香 木香 甘草 朱砂 陈皮 黄连 槟榔 半夏 青皮 神曲 茯苓 大黄 巴豆 白术 三棱 人参 全蝎 使君子 附子 山楂	木香 麝香 槟榔 黄连 甘草 陈皮 朱砂 青皮 神曲 大黄 巴豆 三棱 茯苓 山楂 半夏 芦荟 附子 使君子 麦芽 青黛	木香 麝香 甘草 陈皮 黄连 槟榔 朱砂 大黄 附子 茯苓 半夏 巴豆 青皮 神曲 厚朴 白术 全蝎 山楂 三棱 牛黄

ID	症状集	真实草药集	SMGCN	HetGCN	NGCF	PinSAGE
33745	舌干 头痛 头昏	石决明 赭石 雄黄 玄参 郁金 甘草 黄连 珍珠 大黄 板蓝根 栀子 连翘 天花粉 金银花 黄芩 冰片 蒲公英 朱砂 葛根 磁石 牛黄 麦冬 石膏	大黄 柴胡 香附 薄荷 佛手 珍珠母 珍珠 夜交藤 山楂 郁金	甘草 人参 当归 防风 茯苓 黄芩 川芎 石膏 柴胡 白术	甘草 人参 茯苓 当归 防风 黄芩 石膏 川芎 麦门冬 芍药	甘草 人参 茯苓 川芎 当归 防风 石膏 黄芩 柴胡 麦门冬
33746	咳嗽	甘草 细辛 青黛 人参 党参 炙甘草 五味子	连翘 薄荷 牛膝 杜仲 丹参 红花 滑石 野马追 牛蒡 大黄	杏 杏仁 甘草 桔梗 半夏 人参 桑 茯苓 紫菀 麻黄	甘草 杏 杏仁 桔梗 半夏 茯苓 桑 人参 麻黄 紫菀	甘草 杏 杏仁 桔梗 半夏 茯苓 桑 人参 麻黄 紫菀
33747	恶心 恶寒 呕吐 无汗 腹泻 头痛 恶寒发热 腹痛	生姜 紫苏叶 甘草 陈皮 法半夏 藿香 丁香 白芷 茯苓 苍术 大腹皮 半夏 香薷 广藿香	枳壳 藿香 薄荷 砂仁 麦冬 山楂 佛手 厚朴 香附 泽泻	甘草 白术 半夏 人参 茯苓 陈皮 当归 川芎 厚朴 防风	甘草 白术 茯苓 人参 当归 陈皮 半夏 川芎 防风 厚朴	甘草 白术 茯苓 陈皮 半夏 人参 当归 防风 川芎 木香

首页1 上一页11248 Page 11249 of 11255. 下一页11250 尾页11255

图 5-9 top-10 推荐结果中 SMGCN<其他模型的实例

（3）实时草药推荐

此处用户输入随机的症状名词组合，然后系统会给出对应的推荐结果。以图 5-10、图 5-11 为例，用户输入“胎漏　骨痛\n\n　干呕\n”,系统给出不同算法的推荐结果。

为了避免用户输入重复的或者数据集中不存在的症状，系统预测时会自动过滤掉这些非

法输入，只针对其中的有效输入进行预测。如图 5-12，用户输入中含有重复症状“舌痛”和未知症状“莫名其妙症”，图 5-13 中的推荐会给出过滤的提示信息。

图 5-10 合理症状输入

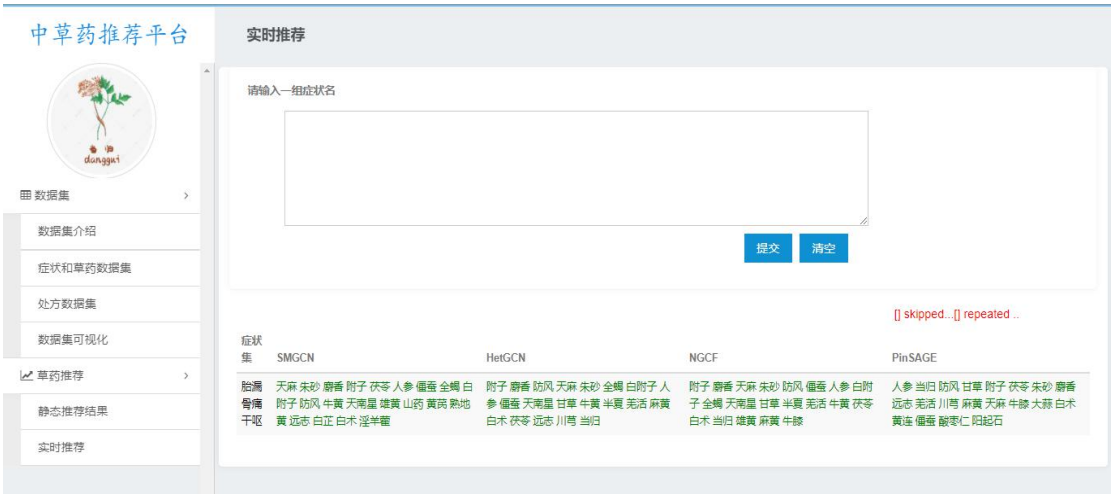

图 5-11 图 5-10 的草药预测结果

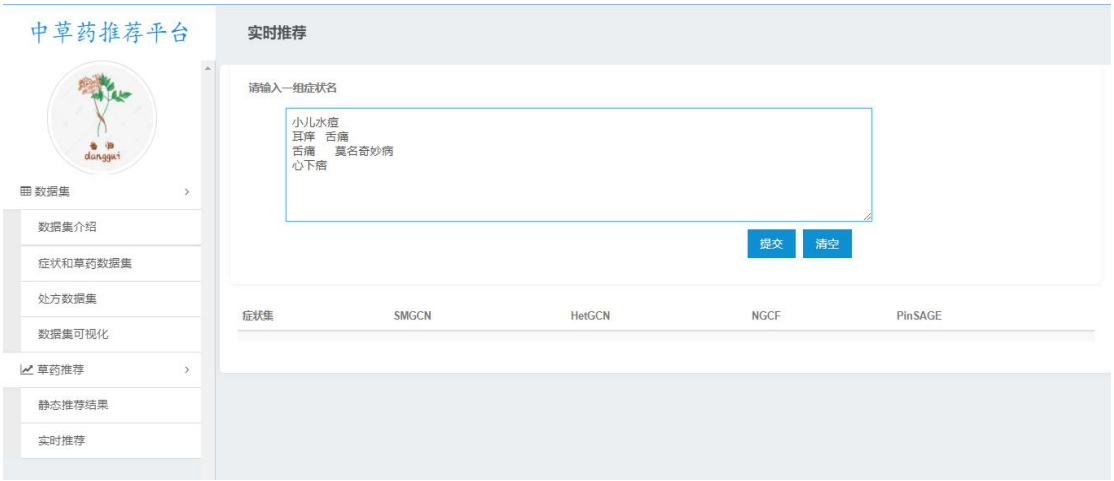

图 5-12 含有重复和不合理内容的症状输入

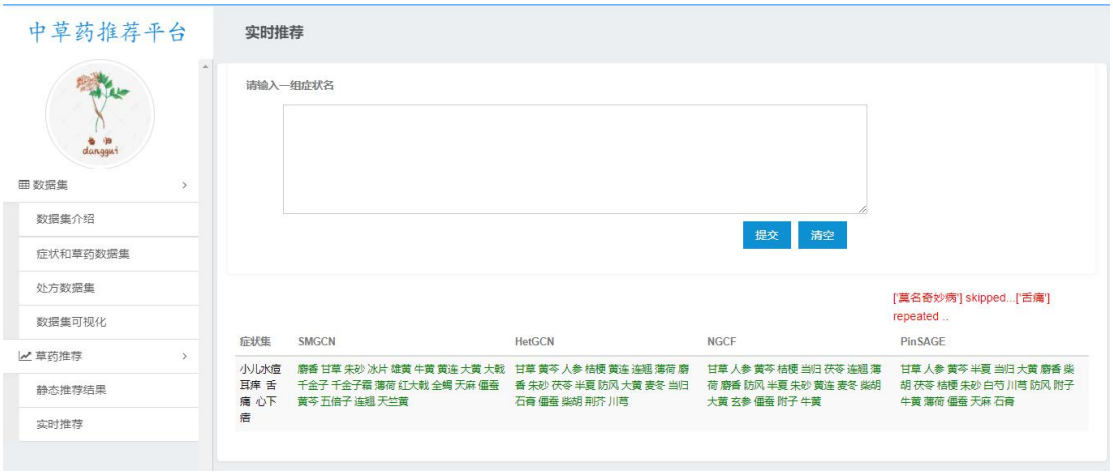

图 5-13 图 5-12 中的推荐结果

6 总结和展望

6.1 总结

本文中的 SMGCN 模型主要引入了 SGE 和 SI 组件。前者直接利用共现图而非在二部图上进行多阶跳跃来获得同类节点的协同信息，而后者使用 MLP 融合多个症状嵌入来模仿中医综合症推导的过程。

其推荐效果相比于从用户-商品领域迁移过来的经典图模型 NGCF、HetGCN 有一定程度的提高，比工业级 PinSage 算法略低，但是 SMGCN 的总训练时长和单位参数训练时长都要远低于 PinSage，所以在需要快速训练部署的生产场景下理应成为首选。

本文还开发了一个中药推荐的可视化平台，一方面展示了中药数据集的来源、数据内容、可视化结果，另一方面实现了各类推荐算法的静态推荐以及基于用户输入的实时推荐，其中对静态结果的观察能发现 SMGCN 能预测出其他模型预测不了的草药。

6.2 展望

虽然 SMGCN 模型有了不少进步，但是仍然存在一些问题和不足，丞待研究和拓展的方面现列举如下：

1）SMGCN 模型利用的数据只有处方中症状集合和草药集合的编号信息，而没有将症状和草药节点自身相关的其他属性用于特征嵌入的构建；

2）SMGCN 模型中模拟综合症推导的过程只有一层简单的 MLP，对于症状集合中随机组合的症状关联可能无法充分挖掘，导致效果提升不多；

3）SMGCN 模型中的交互建模只是简单的内积，可能需要引入非线性神经网络来交互；

4）SMGCN 模型的应用领域可以从中药自然延拓到西药领域，因为西药遗留下来的知识和数据更为丰富，同时西药推荐本身会有更多药理上的支撑。

参考文献

[1] 于然,贾立群,娄彦妮. 基于关联规则与聚类分析探索中药外治手足综合征的用药规律[J]. 中国医药导报,2020,17(15):139-142.

[2] 汪浩,王海平,吴信东,刘琦.药物-疾病关系预测:一种推荐系统模型[J].中国药理学通报,2015,31(12):1770-1774.

[3] 贾香恩,董一鸿,朱锋,钱江波.异构图卷积网络研究进展[J/OL].计算机工程与应用:1-15[2021-04-17].http://kns.cnki.net/kcms/detail/11.2127.TP.20210303.1404.002.html.

[4] 王新宇. 中医实体表示学习与中药推荐技术[D].华东师范大学,2020.

[5] 王越辉. 基于矩阵分解的 lncRNA-疾病关联预测研究[D].西南大学,2020.

[6] 王茹玉. 基于信息推荐的中药适应症发现方法研究[D].北京交通大学,2019.

[7] 庄力,周雪忠,贾彩燕,于剑,张润顺,王映辉.基于 Biclustering 的中医药症关系分析[J].计算机工程,2010,36(11):241-243.

[8] 曹小凤.基于混合推理的高血压药物推荐模型研究[J].软件工程,2018,21(03):41-43.

[9] 高占林.浅谈信息过载的影响及消除[J].天水行政学院学报,2010,(06):63-65.

[10] 王科文. 基于深度神经网络的中医药材推荐[D].华南理工大学,2018:1-52.

[11] 唐丽佳,潘静.运用数据可视化分析肾系疾病用药规律[J].亚太传统医药,2020,16(12):189-192.

[12] Jin Y, Zhang W, He X, et al. Syndrome-aware Herb Recommendation with Multi-Graph Convolution Network[C]//2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020: 145-156.

[13] Ji W， Zhang Y， Wang X , et al. Latent semantic diagnosis in traditional chinese medicine[J]. World Wide Web-internet & Web Information Systems, 2017, 20(5):1-17.

[14] Ma J， Wang Z . Discovering Syndrome Regularities in Traditional Chinese Medicine Clinical by Topic Model[J]. Springer International Publishing, 2016.

[15] Fan L， Xiahou J， Xu Z . TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model[J]. Multimedia Tools & Applications, 2016, 75(22).

[16] Sheng W， Huang E W， Zhang R , et al. A conditional probabilistic model for joint analysis of symptoms, diseases, and herbs in traditional Chinese medicine patient records[C]// 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016.

[17] Li W , Yang Z , Sun X . Exploration on Generating Traditional Chinese Medicine Prescription from Symptoms with an End-to-End method[J]. 2018.

[18] Ruan C, Ma J, Wang Y, et al. Discovering Regularities from Traditional Chinese Medicine Prescriptions via Bipartite Embedding Model[C]//IJCAI. 2019: 3346-3352.

[19] Ruan C, Wang Y, Zhang Y, et al. Exploring regularity in traditional chinese medicine clinical data using heterogeneous weighted networks embedding[C]//International Conference on Database Systems for Advanced Applications. Springer, Cham, 2019: 310-313.

[20] Chen Jinpeng et al. Mining Symptom-Herb Patterns from Patient Records Using Tripartite Graph.[J]. Evidence-based complementary and alternative medicine : eCAM, 2015, 2015 : 435085.

[21] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

[22] Mirhoseini A , Goldie A , Yazgan M , et al. Chip Placement with Deep Reinforcement Learning[J]. 2020.

[23] Johnson J , Gupta A , Fei-Fei L . Image Generation from Scene Graphs[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[24] Wang X , Zhang Y , Wang X , et al. A Knowledge Graph Enhanced Topic Modeling Approach for Herb Recommendation[J]. 2019.

[25] Yao L , Zhang Y , Wei B , et al. A Topic Modeling Approach for Traditional Chinese Medicine Prescriptions[J]. IEEE Transactions on Knowledge & Data Engineering, 2018:1-1.

[26] Zhao J, Zhou Z, Guan Z, et al. Intentgc: a scalable graph convolution framework fusing heterogeneous information for recommendation[C].Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2019: 2347-2357.

[27] Fan S, Zhu J, Han X, et al. Metapath-guided heterogeneous graph neural network for intent recommendation[C].Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2019: 2478-2486.

[28] Wang X, He X, Wang M, et al. Neural graph collaborative filtering[C]//Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval. 2019: 165-174.

[29] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network[C]//Proceedings of the 25th ACM

SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 793-803.

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

[31] Yao L , Mao C , Luo Y . Graph Convolutional Networks for Text Classification[J]. 2018.

[32] erozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 701-710.

[33] 谢雨洋,冯栩,喻文健,唐杰.基于随机化矩阵分解的网络嵌入方法[J].计算机学报,2021,44(03):447-461.

[34] Mooney R J, Roy L. Content-based book recommending using learning for text categorization[C]//Proceedings of the fifth ACM conference on Digital libraries. 2000: 195-204.

[35] Breese J S . Empirical analysis of predictive algorithms for collaborative filtering[C]// Proc.14th Conference on Uncertainty in Artificial Intelligence, Madison, WI. Morgan Kaufmann Publisher, 1998.

[36] Balabanovic M,Shoham Y . Fab: content-based, collaborative recommendation[J]. Communication of the ACM, 1997, 40(3):66-72.

致谢

又是一年春夏之交，伴随着玉兰花开花谢，图书馆前的枫林大道由嫩绿变得葱茏青翠，我在西北大学的四年本科旅程也即将到站。本科阶段的学习体验就如同一只蓝鲸在特定的深海区游动翻滚般充满了未知，我的主动学习能力、探索能力以及沟通表达能力在不知不觉中实现了初步的跃迁。在此，非常感谢老师、同学、朋友和亲人们对我提供的指导和帮助，让我始终能够朝着正确的方向、怀着饱满的热情不断重塑自我。

首先特别感谢我的论文指导老师，西北大学信息科学与技术学院的张海波副教授。在开题阶段为我仔细分析了课题的研究方向，消除了一些认知上的偏差；在项目具体开展阶段及时地解答了我遇到的疑难和困惑；在论文撰写和实验完善阶段，指出并纠正了我的很多错误和不足。最终，我为我的毕业交上了一份满意的答卷。

其次感谢答辩组的张顺利教授和张雨禾副教授给出的针对性指导意见，使得我的毕业作品更加完善。

最后衷心感谢在百忙之中评阅论文的各位专家、教授！