

北京交通大学

硕士专业学位论文

融合领域知识的中医处方推荐方法研究

Study on TCM Prescription Recommendation Method Integrating
Domain Knowledge

作者：董鑫

导师：周雪忠

北京交通大学

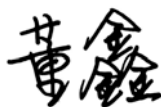
2022 年 6 月

学位论文版权使用授权书

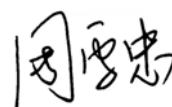
本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：



导师签名：



签字日期：2022年5月21日

签字日期：2022年5月21日

学校代码: 10004

密级: 公开

北京交通大学

硕士专业学位论文

融合领域知识的中医处方推荐方法研究

Study on TCM Prescription Recommendation Method Integrating
Domain Knowledge

作者姓名: 董鑫

学 号: 20125155

导师姓名: 周雪忠

职 称: 教授

专业领域: 计算机技术

学位级别: 硕士

北京交通大学

2022 年 6 月

致谢

时光荏苒，春华秋实，转眼间我在北交的硕士生活即将画上句号。两年的时光，尽管自己依旧是懵懂无知的小白，在大家的帮助提携下，我的硕士阶段亦有不少收获。此刻，硕士的终点无约而至，内心虽感慨万千，但更多的是感激之情。

首先，感谢我最敬爱的导师周雪忠教授，感谢缘分和幸运让我能成为您的学生，感谢两年来您对我的提携和指导，在今后的学习和工作中我会继续以您为榜样，不畏艰难险阻、奋力拼搏，以实际行动回报老师知遇之恩。感谢机器学习与认知计算研究所的贾彩燕教授和桑基韬教授对我的指导和帮助，感谢实验室夏佳楠老师在我学习和生活上的支持和鼓励。

感谢所有师兄师姐、师弟师妹以及同级伙伴们，大家共同营造的融洽氛围为我的硕士生活增添了色彩。在此特别感谢杨扩师兄，感谢两年来师兄对我的指导和鼓励，今后我也会继续以师兄为榜样，在各方面向师兄看齐，不负师兄对我的期望。感谢郑琪光师兄、田昊宇师兄、于泽丛师兄以及已毕业的邹群盛师兄和郑毅师兄对我的帮助。还要感谢湖北中医药大学李君师姐，感谢师姐在合作中给我的帮助和全力支持，更要感谢师姐对我的理解，以及为我开的处方，使我爱上了喝荷叶茶。感谢同届伙伴孙海龙、钟昆禹、董汉阳、李亚茹、王越霄对我的帮助，特别感谢霄霄和汉阳在学习和生活上对我的支持。

感谢所有帮助过我的伙伴们。首先感谢 20 级专硕 1 班的所有同学们，尽管硕士阶段大家各有所志，但是这并不影响咱们集体的团结，班级风采大赛、每次班里精心组织的活动、党支部活动等等，这些都历历在目。特别感谢韩宇臻、霍东昆、王晨宇、董勇在生活上对我的关心和支持。感谢老友小琦、阳阳对我的支持和鼓励，愿我们的友谊地久天长。

更重要的，感谢养育我的家人，感谢妈妈、爸爸、姥姥、姐姐一家对我的养育和鼓励，感谢你们为我撑起了天空，使我能尽情翱翔，我会以实际行动回报家人的养育之恩。

同时感谢各位评审老师，感谢您百忙之中评阅我的论文，也感谢您对我所做工作的认可和指正，本文的撰写过程中也会存在缺陷和不足，感谢您能对我的论文提出宝贵意见，辛苦各位老师。

最后，希望新冠病毒早日远离我们，希望大家都能拥有美好的未来，希望每个人许下的愿望都能如约而至。

2022 年 4 月
于北京丰台

摘要

中医智能处方推荐指利用人工智能技术,根据患者病历信息进行学习,对候选中药进行预测,以模拟医生开具处方的过程。近年来许多学者围绕中医处方推荐开展了相关研究,但目前领域内仍存在亟待解决的问题,如现有临床诊疗数据具有“一多一少”的特点、无法对临床表型中的“未登录词”形成表示、现有处方推荐方法性能较低、现有方法推荐结果的配伍合理性不强等。为解决上述问题,本文围绕中医处方推荐开展了以下三方面研究。

第一,针对现有临床诊疗数据存在的“一多一少”现象,本文提出了基于症状本体库和症状共现关系的处方数据增强方法 SOCO 模型和基于知识图谱采样的处方数据增强方法 SabKG 模型。本文基于症状本体库、临床医案数据和领域知识构建了症状同义关系集、症状共现集和药症知识图谱,在此基础上形成了两种处方数据增强策略。本文将提出的两种策略与基线方法应用于多标签处方推荐任务中,实验结果表明提出的两种策略均能使数据性能得到提升,并且 SabKG 模型达到了相对最优的性能。

第二,针对现有处方推荐方法性能较低以及如何对临床表型中“未登录词”进行表示的问题,本文提出了基于症状术语映射与深度学习的处方推荐方法 TCMPR 模型。该模型以子图抽取方式对患者原始症状进行映射,结合已构建的症状网络对映射后的术语集合进行特征融合,并通过构建的神经网络模型进行学习,得到候选药物的预测概率。本文将提出的 TCMPR 模型与基线方法进行了性能对比,实验结果显示 TCMPR 模型达到了最优的性能,并且能够较好地结合领域知识对临床表型中的“未登录词”进行表示。模型关键模块的相关实验显示,子图筛选策略、症状嵌入表示方法、特征融合方式等模块的选择对模型的性能产生着一定影响。

第三,针对现有方法推荐结果的配伍合理性不强这一问题,本文提出了基于表型相似度与经典名方的处方推荐方法。本文首先形成了结合经典名方的处方推荐策略,实验结果显示出该策略的有效性。考虑到此策略受限于经典名方数据,本文提出了基于表型相似性与经典名方的处方推荐方法,此方法以构建患者表型特征为核心,融合了经典名方和症状网络等领域知识,通过衡量医案数据与经典名方的表型相似度以形成推荐结果。本文对提出的患者表型特征构建及匹配策略进行了实验,结果显示 SSTM-SymJac 和 SSTM-SymCos 两种策略均能有效提升推荐性能,同时使推荐结果的配伍合理性得到了保证。

关键词: 处方推荐; 临床数据增强; 症状术语映射; 深度学习; 表型特征构建

ABSTRACT

TCM intelligent prescription recommendation refers to learn from medical records of patients and predict candidate Chinese herbs by combining artificial intelligent technologies, so as to simulate the diagnose process of doctors. In recent years, many scholars have carried out related research, but there are still some problems need to be solved in the field of prescription recommendation, such as existing clinical data have the characteristics of "One more and one less", the problem of representing "Unrecorded symptom phenotypes", the performance of existing recommendation methods is relatively low, the rationality of collocation among herbs recommended by existing methods need to be improved, etc. In order to solve the above problems, the following three aspects of TCM prescription recommendation were studied in this work.

Firstly, in view of the phenomenon of "One more and one less" existing in clinical case data, a clinical data augmentation method called SOCO was proposed which based on symptom ontology library and symptom co-occurrence relationship, and another data augmentation method called SabKG was proposed based on knowledge graph sampling. First of all, symptom synonymous relation set, symptom co-occurrence set and herb-symptom knowledge graph were constructed by using symptom ontology library, clinical case data and domain knowledge. After that, two clinical data augmentation methods were proposed. The strategies proposed in this work and baseline methods were applied in the multi-label prescription recommendation task, and the experimental results showed that the two proposed methods can improve the performance of clinical case data, and the SabKG model achieves relatively optimal performance.

Secondly, in view of the performance of existing TCM prescription recommendation methods is relatively low, and the problem of how to represent "Unrecorded symptom phenotypes" better, TCMPR model was proposed for prescription recommendation which based on symptom term mapping and deep learning. In this model, the original symptoms of patients were mapped by subnetwork extraction, and the features of the mapped term set were fused with the constructed symptom network, and the prediction probability of candidate herbs was obtained by learning the constructed neural network model. The performance of TCMPR is compared with that of the baseline methods, and the experimental results showed that the proposed TCMPR achieves the optimal performance, and the method could better represent the "Unrecorded symptom phenotypes"

in clinical data by combining domain knowledge. Experiments on key modules of the model showed that the selection of subnetwork screening strategy, symptom representation method, feature fusion method and other modules have impact on the performance of TCMPR model.

Thirdly, in view of the rationality of collocation among herbs recommended by existing methods need to be improved, a prescription recommendation method was proposed which based on phenotypic similarity and classical prescription data. In the first place, a prescription recommendation strategy based on classical prescription data was proposed, and experimental results showed the effectiveness of this framework. Considering that this strategy was limited by classical prescription data, another prescription recommendation method based on phenotypic similarity and classical prescription data was proposed. This strategy focused on building phenotypic characteristics of patients, combined domain knowledge such as classical prescription data and symptom network, and formed recommended result by measuring phenotypic similarity between clinical case data and classical prescription data. Experiments were conducted on the proposed patient phenotypic feature construction methods and similarity matching strategies, the results showed that both the SSTM-SymJac and SSTM-SymCos strategy could effectively improve the recommendation performance, and the compatibility of the recommended results is simultaneously guaranteed.

KEYWORDS: Prescription recommendation, Clinical data augmentation, Symptom term mapping, Deep learning, Phenotype feature construction

目录

摘要	iii
ABSTRACT	iv
1 引言	1
1.1 研究背景和意义	2
1.2 国内外研究现状	3
1.2.1 中医处方推荐研究	3
1.2.2 西药组合推荐研究	6
1.2.3 药物重定向研究	7
1.3 本文主要工作	8
1.4 本文组织结构	9
2 相关研究及方法基础	11
2.1 多分类与多标签分类方法概述	11
2.1.1 多分类问题与多标签分类问题定义	11
2.1.2 多分类经典算法	12
2.1.3 多标签分类经典算法	14
2.2 深度学习方法概述	14
2.2.1 前馈神经网络	15
2.2.2 卷积神经网络	15
2.2.3 注意力机制	16
2.3 文本数据增强方法概述	16
2.3.1 文本数据增强简介	17
2.3.2 EDA 方法	17
2.4 网络表示学习概述	18
2.5 本章小结	19
3 基于数据增强的中医处方推荐方法研究	20
3.1 引言	20
3.2 数据处理与构建	20
3.2.1 临床医案数据预处理	21
3.2.2 基于医案数据的症状共现数据集构建	22

3.2.3	基于症状本体库的症状同义关系集构建	22
3.2.4	药症知识图谱构建	23
3.3	结合领域知识的处方数据增强策略	24
3.3.1	基于症状本体库与症状共现关系的处方数据增强策略	25
3.3.2	基于知识图谱采样的处方数据增强策略	27
3.4	实验设计及验证	28
3.4.1	实验方案设计	28
3.4.2	评价指标与参数设置	30
3.4.3	实验结果及分析	30
3.5	本章小结	33
4	基于症状术语映射的处方推荐方法研究	34
4.1	引言	34
4.2	基于症状术语映射与深度学习的处方推荐框架	34
4.2.1	整体框架介绍	35
4.2.2	症状网络形成及嵌入表示	36
4.2.3	患者症状映射与特征融合	37
4.2.4	基于症状术语映射的处方推荐方法 TCMPR	38
4.3	实验设计与参数设置	39
4.3.1	实验数据与实验设计	39
4.3.2	评价指标与参数设置	40
4.4	实验结果及分析	40
4.4.1	模型与基线方法对比结果及分析	41
4.4.2	TCMPR 方法对未登录词激活效果分析	42
4.4.3	模型关键模块对比实验结果及分析	43
4.4.4	模型其他超参数实验及结果	46
4.5	本章小结	47
5	基于表型相似性的处方推荐方法研究	49
5.1	引言	49
5.2	基于经典名方的处方推荐方法及实验结果	49
5.2.1	基于经典名方的处方推荐策略	50
5.2.2	经典名方数据处理	50
5.2.3	实验设计与评价指标	52

5.2.4	实验结果及分析	53
5.3	基于表型相似性与经典名方的处方推荐方法	56
5.3.1	整体框架	56
5.3.2	患者表型特征构建策略	57
5.3.3	表型相似度匹配方法	58
5.4	基于表型相似性与经典名方的处方推荐实验及结果	59
5.4.1	实验设计与评价指标	59
5.4.2	实验结果及分析	60
5.4.3	案例分析	66
5.5	本章小结	67
6	总结与展望	68
6.1	全文工作总结	68
6.2	未来研究展望	69
	参考文献	70
	作者简历及攻读硕士学位期间取得的研究成果	75
	独创性声明	76
	学位论文数据集	77

1 引言

中医药是我国极具特色的学科领域，是中华文化的瑰宝，几千年来，中医在中国人的健康保障中发挥着不可或缺的作用，并且在当今世界中产生了不可替代的影响，特别是全球新冠肺炎疫情发生后，在人类命运共同体的思想指引下，中医药积极参与到全球疫情防控中，为世界贡献了中国方案，推动了中医药走向世界的进程^[1]。

中医是我国独居优势和特色的传统医学体系，其主要临床模式是以患者就诊时所采集的“四诊”信息为主要根据，对患者进行个性化治疗。中医的治疗过程可称为“理法方药”^[2,3]，分别指理论、治法、方剂和中药，即由中医医生根据病人的症状信息确定病人所患疾病的原因与机理，然后根据疾病机理给出相应的治疗方法，最后形成用于治疗患者症状的诊断方案和对症处方^[4]。在此过程中，中医医生对于患者诊疗则起到了至关重要的作用，目前中医全科人才资源相对稀缺，培养一名合格的中医师是相对漫长的过程，需要较长时间的理论学习和临床实践，基于此现状，中医药的普及化、基层化进程受到了一定影响^[5]。

医学人工智能已成为实现临床诊断决策支持的热点方向。近年来，人工智能产业愈加呈现出爆发式的增长态势，“人工智能+”已深入各行各业，为各领域的信息化发展起到了重要依托作用。在此背景下，国家提出了推进中医药现代化、信息化、智能化的目标，在当今计算机科学技术与人工智能技术蓬勃发展的浪潮下，以信息化驱动中医药现代化，是适应国家信息化发展新形势的重要举措，是推进中医药振兴发展的内在要求，也是实现人人基本享有中医药服务的必然选择^[6,7]。

随着计算机科学技术、人工智能技术等方法与现有医疗诊断技术的不断结合，中医智能处方推荐成为了医学人工智能领域内的一项研究热点，许多专家学者提出了结合计算机技术进行中医开方配伍的智能处方推荐方法。但由于中医诊疗具有个性化的特点，目前领域内已形成的智能处方推荐方法的性能仍不理想，并且推荐结果的配伍合理性、临床可用性还有待考量。

针对目前中医处方推荐领域中存在的问题与难点，本文结合了领域知识，对如何提升中医处方推荐的性能，以及提升结果的配伍合理性等方面进行了相关探索。本文首先结合领域知识，提出了基于数据增强的处方推荐方法；结合了子图抽取和深度学习相关技术，提出了基于症状术语映射的处方推荐方法；并从推荐结果配伍合理性的角度出发，提出了基于表型相似性与经典名方的处方推荐方法。

1.1 研究背景和意义

截止至 21 世纪初,大部分医疗机构的信息化建设水平仍停留在相对较低的阶段,究其原因:一是领域内相关复合型人才储备不够,具体表现为医疗临床人员的计算机信息技术知识储备不够,以及“医疗+计算机技术”复合人才的储备不足;二是资源规划尚不平衡,其表现在东西部资源配比不均、城市与农村资源配比不均等,特别是农村基层地区,缺乏较好的设备支持与良好的网络通信支撑,这使得其信息化进程发展缓慢;三是尚未建立统一化的数据标准,不同平台间缺乏交换接口,导致信息孤岛的现象日益严重,因此同一患者就诊于不同医院时,多家医院形成的不同标识无法得到通用^[8]。

近年来,随着国家将医疗信息化提升到战略层次,《全国医疗卫生服务体系规划纲要(2015-2020)》等政策文件的颁布,医疗信息化发展的进程开启了快进键:“医疗+信息技术”复合型人才储备正不断提升,不同区域资源配比差异逐渐减小,我国医疗行业信息化市场规模逐年递增。经过近几年的发展,大型医疗机构的医院信息系统(Hospital Information System, HIS)已经基本建设完成,中小型医疗机构也在积极配备 HIS 系统,但现阶段仅仅完成了对医疗诊断相关数据的采集工作,采集到的数据质量较低,且采集到的大量临床数据的利用率不高,特别是在临床实践过程中形成的中医处方数据,这些数据至今尚未能被充分利用与挖掘。

随着人工智能技术在医学领域的广泛应用,对医学临床数据的挖掘与应用这一研究方向已逐渐成为医学人工智能领域的重要研究内容与热点,并且经过近几年的相关研究与实践,利用临床医疗数据挖掘出的新知识与规律,对临床决策产生了积极的影响和良好的社会效益^[9]。中医是经验医学,针对患者症状开出的处方数据能够直接体现医生的临床经验与知识水平,而目前较有经验的临床医生的数量与患者的数量是严重不平衡的,如果能利用好已有的中医临床处方数据,结合人工智能相关方法对其挖掘,进行中医智能处方推荐,不仅能对临床医生诊疗起到辅助作用,而且对于探索真实世界临床研究(如临床新药发现、临床机制解释、临床疗效验证等研究热点)将会起到一定的推动作用。中医处方推荐这项领域内热点研究由此孕育而生。

中医智能处方推荐指利用数据挖掘、人工智能等计算机技术,根据病人的电子病历信息及医学领域相关知识进行学习,归纳出病人所属证型,而后根据储备的中医领域知识对候选中药进行预测,从而实现模拟中医医生开具处方的过程,其本质是将患者的就诊信息(如患者基线信息、症状表型信息、理化检验结果等)作为输入特征进行训练,并且结合数据挖掘技术对患者进行诊断(包括归结中医疾病诊断、确定治则治法等),找到对治疗患者症状能够起到直接作用的中药处方。形成符合真实世界开方配伍原则的中医智能处方推荐方法,无论对于辅助医生诊

疗、便利居民日常生活，还是推动计算机科学技术的发展与应用，以及推动中医药现代化、信息化、智能化发展，都具有一定的积极意义和较高的应用价值。

1.2 国内外研究现状

近年来，国内外许多学者围绕中医智能诊断、有效处方发现、中药配伍发现等中医处方数据挖掘的研究热点开展了相关研究。本节将对近年来形成的中医处方推荐方法进行介绍，并对与中医处方推荐密切相关的药物组合推荐、药物重定位两项研究及其进展分别进行阐述。

1.2.1 中医处方推荐研究

根据采用的研究技术和手段，近年来的中医处方推荐相关工作可归结为以下四种类别：基于传统机器学习的中医处方推荐、基于主题模型的中医处方推荐、基于复杂网络的中医处方推荐以及基于深度学习的中医处方推荐。

(1) 基于传统机器学习的中医处方推荐方法及相关研究

基于传统机器学习相关方法的主要思想是将中医处方推荐视为分类问题，即利用患者的临床症状表型作为特征进行训练，或对疾病进行预测（得到与该患者相关的中医疾病诊断），或对药物进行预测（将候选处方或者候选药物作为一个单独存在的类别进行分类，其中使用较多的做法是对候选药物进行预测，最后推荐出一个中药组合作为推荐结果）。Mi 等人^[10]利用逻辑回归（Logistic Regression, LR）、K 近邻算法（K-Nearest Neighbor, KNN）、决策树（Decision Tree, DT）、随机森林（Random Forest, RF）等多种机器学习算法，建立了常用处方的预测模型，阐明了处方预测的可能性和稳健预测所需的数据量，是对处方推荐应用的综合基线模型探索。Xian 等人^[11]通过使用主成分分析等技术，通过海量病历的处方数据进行学习，并结合诊疗指南等知识库数据，实现了基于医生的诊疗习惯及疾病的诊疗原理的处方推荐。Wang 等人^[12]利用基于 K 近邻的多标签分类方法（Multi-Label K-Neighbor Nearest, MLKNN），通过对患者症状句的拆解、形成症状向量，以及与已有向量库进行相似性匹配，实现了对患者疾病诊断的预测，并开发了应用系统。

对于中药处方的预测，目前的相关研究中常见做法是将每味药物视为单独的类别，因此这种做法将中药处方的预测转化为多标签分类问题（Multi-Label Classification, MLC），即为每个患者分配多个中药类别。解决多标签分类问题的最早尝试之一是将其转换为多个单标签分类任务^[13]，但其不足是未考虑标签间存在的相关性。随后，相关学者提出了其他方法来捕获低阶标签相关性^[14-16]或将 MLC 任务转换成二进制分类链^[17]。但是这种做法忽略了中药间的相互作用关系，与实

际中药配伍原理不符。总体上,基于多标签分类的方法常用作各种方法的基线模型。如 Shi 等人^[18]的研究中对帕金森病患者的自动处方推荐进行了讨论,主要做法是通过学习潜伏症状处方 PALAS 模型,利用数据的多模态表示来推荐处方,该文章的实验部分将 PALAS 模型与 MLKNN、Binary Relevance、Rank-SVM 等多种多标签分类算法的性能进行了对比。再如 Wang 等人^[19]的研究中,将提出的 TCM Translator 模型与 MLKNN、Label Powerset 等多标签分类方法进行了性能比较。

整体而言,目前传统机器学习方法经常在深度方法中作为基线模型进行使用,以对比新方法的性能提升能力,特别是多标签分类算法,但这对于中医处方推荐而言并不是一种较为合适的解决方案。

(2) 基于主题模型的中医处方推荐方法及相关研究

主题模型是用来在一系列文档中发现抽象主题的一种统计模型^[20]。近几年主题模型在中医处方推荐领域中得到了广泛的关注^[4,21-25],其主要思想是将每个处方视为一个文档,文档中包含几个潜在的主题,即治疗模式,在该文档中,症状和中药分别被视为一组“词”^[26]。

如 Yao 等人^[4]建立了中医处方的主题模型,该模型描述了中医理论中方剂的生成过程。Zhang 等人^[21]提出了一种症状-中药-诊断主题 SHDT 模型的数据挖掘方法,用以从大规模的中医临床数据中自动提取症状、中药组合和诊断之间的共同关系。Zhang 等人^[22]提出了一种层次化的症状-中药主题模型 HSHT,以自动提取中医临床数据中包含症状及其对应中药的层次化潜在主题结构。Jiang 等人^[23]使用链接潜在狄利克雷分配 LinkLDA 来自动提取包含症状及其相应中药信息的潜在主题结构。Yao 等人^[24]提出了一种利用有监督主题模型和中医领域知识挖掘中医临床病例治疗模式的框架,该框架可以反映中医的基本规律,提高新方的功能预测能力。Wang 等人^[25]在对新方剂功能的预测方法上开展了相关研究,提出了有监督主题模型 Label-Prescription-Herb,该模型将中药配伍规则融入到学习过程,并且提出了基于 TFIDF 特征和中药属性特征构建的多标签分类器。

由于这些优势,主题模型不仅可以推荐合适的中药,还可以探索它们的适应症和配伍^[27]。然而,常用的中药有千余种,总是存在高阶的中药或症状关联,这对于主题模型而言计算起来是相对困难的。此外,主题模型对于语料数据的质量要求较高,目前真实临床处方数据的质量并不能满足主题模型的需要。因此,这使得主题模型在应用于复杂的中药处方预测任务中的应用效果并不理想。

(3) 基于复杂网络的中医处方推荐方法及相关研究

复杂网络已迅速形成了一门贯穿多领域的交叉性学科,其相关理论被应用于诸多领域。人体生命及其疾病系统是典型的复杂系统,因此网络医学的概念随之诞生,旨在采用复杂网络的方法和技术进行临床疾病诊断规律挖掘、新药发现等

方面的研究。

近年来,许多学者从复杂网络的视角对中药药物配伍、核心处方挖掘以及有效处方推荐等热点进行了深入的研究。Zhou 等人^[28]从大量的中医药数据中提取出关键的药物配伍等知识,体现出中药之间并不是相互独立的,而是有着紧密的相互作用关系。Yang 等人^[29]提出了一种将倾向案例匹配、复杂网络分析和中药富集分析相结合的多阶段分析方法,以识别针对特定疾病(例如失眠)的有效中药处方。Jin 等人^[30]在建立单纯性肥胖症耳穴疗法数据库的基础上,基于复杂网络技术对单纯性肥胖症的核心穴位和配伍规律进行分析,总结出了单纯性肥胖症耳穴疗法的特点。Feng 等人^[31]针对不稳定型心绞痛,以部分可观测马尔可夫决策过程模型 POMDP 为基础,选取气虚、血瘀、痰浊 3 个证素的住院患者,对不稳定型心绞痛的诊疗进行深层次的数据挖掘、分析和客观评价。Xu^[32]的工作中,构建了包含症状和处方的异构网络,并在此基础上结合了复杂网络中社团划分的相关方法进行处方推荐,达到了较为理想的性能。Zhang 等人^[33]提出一种基于复杂网络的中医治疗肺癌的处方推荐算法 PRCN,其核心思想是确定治疗肺癌核心药物,根据核心药物来确定最有效的方剂,实验结果表明该算法具有理论正确性,可以为研究中医治疗肺癌提供一定的帮助。Yang 等人^[34]通过 D 指数法筛选潜在获益药物群,并在药物的配伍网络的基础上采用复杂网络的极大团算法,并结合生存分析模型挖掘分析中医药治疗肺癌的核心有效处方。

总体上看,基于复杂网络的处方推荐与药物配伍挖掘的相关方法主要集中在对单病种临床数据的挖掘,而对于多病种数据探索与挖掘的相关工作较少。

(4) 基于深度学习的中医处方推荐方法及相关研究

随着计算能力的增强,深度学习在医疗领域的应用越来越广泛,先前制约深度学习的问题逐渐得到了解决。在医学文本相关任务中,通过构建深度模型训练临床电子病历数据来进行下游预测与挖掘任务,这一方式已经成为广泛使用的研究思路。

近年的研究中,Zhang 等人^[35]的研究根据用户过往病历中的信息,提出了个性化临床处方的框架,将人工神经网络与基于案例的推理相结合,提高了推荐系统的效率,降低了预测误差。Li 等人^[36]提出了一种中药方剂自动生成方法,该任务旨在基于症状文本描述自动地生成中药方剂,并针对其解码器产生重复药物的现象,提出了一种具有覆盖机制和新软损失函数的新型解码器,实验结果证明了该方法的有效性。Wang^[37]的研究中,对中医疾病、方剂、中草药、症状等实体进行提取并构建了知识图谱,然后使用 node2vec 将构建的知识图谱转换为向量空间,最后基于向量之间的相似性进行处方推荐,结果显示该方法具有较好的推荐命中率。Jin 等人^[38]的研究中构建了症状感知多图卷积网络模型 SMGCN,用于中药

推荐任务中证候与中药之间的交互建模，其实验结果表明了该模型融合证候表示与多图卷积的有效性。Yang 等人^[39]通过构造中药知识图谱，引入中药知识作为附加辅助信息，提出了一种多层信息融合的图卷积模型，通过多层图卷积网络的训练得到药物和症状的嵌入表示，并通过训练后的患者表示与药物表示矩阵的乘积得到每味候选药物的预测概率，其实验结果表明该模型在 Top@K 指标上得到了一定的性能提升。Jin 等人^[40]利用了注意力网络来区分症状的重要性，自适应地融合症状嵌入，构建了一种基于知识图谱增强的多图神经网络体系结构，实验结果表明，其提出的模型在性能上产生了较好的效果。Wu 等人^[41]提出了 HsCTRD 模型，该模型结合了构建的 TAHIN 知识图谱信息以形成更高质量的节点表示，实验结果显示，其形成的表示能够提升在中药分类等药物规律挖掘任务上的性能。Li 等人^[42]结合了 seq2seq 框架设计了处方生成模型 Herb-Know，该模型结合了联合注意力机制和指针网络来进行处方生成，提升了生成处方的质量。

尽管深度学习模型相比于传统模型取得了一定的进步，能够达到一个更好的性能，但是由于神经网络“黑盒”的特性，其可解释性不强，即无法对推荐出的结果进行回溯，无法达到“知其所以然”的理想目标，这不是临床医生和就诊患者所希望看到的结果。因此，可解释性与安全性是基于深度学习的处方推荐方法亟待解决的问题。

1.2.2 西药组合推荐研究

药物组合推荐是与中医处方推荐密切相关的一项工作，是诸多国外研究者对医学人工智能领域关注较多的研究热点之一。与中医处方推荐相关研究相比不同之处在于：一是药物组合推荐的研究对象为西药，其组成复杂性远不及中医处方；二是西药组合推荐研究领域已有相对成熟的标准数据集与知识集，如 DDI（反映西药相互作用产生不良反应的知识集）、MIMIC-III 数据集等，而中医处方推荐研究中尚无质量较高的公开数据集，也未形成全面反映中药间相互作用的知识集。虽然两者在研究对象上有着主要区别，但是西药组合推荐的研究思想是值得中医处方推荐研究借鉴的。

近年来，Aujla 等人^[43]提出了一种基于软件定义网络 SDN 和深度学习的疾病分类医疗推荐模型，该系统在收集数据的基础上为医生推荐疾病的治疗方案。Wang 等人^[44]提出了具有循环神经网络与监督强化学习的协同学习框架 SRL-RNN，应用了一个非政策行为者-批评者框架来处理多种药物、疾病和个体特征之间的复杂关系，框架中的“行动者”由指标信号和评估信号调整，以确保有效处方和低死亡率。Wang 等人^[45]提出了 CompNet 模型，将任务转换为无序马尔可夫决策过程问题，并设计了一种深度 Q 学习机制来学习药物之间的相关和不利相互作用。

西药组合推荐的相关工作中，对药物间不良反应的考虑是值得在中医处方推荐中学习的。例如 Gong 等人^[46]的工作中，利用了 MIMIC-III 数据集、ICD-9 本体和 DrugBank 数据库构建了高质量的异构图，然后将疾病、药物、患者及其对应关系共同嵌入较低维度空间，并使用嵌入将药物推荐分解为链接预测过程，同时考虑患者的诊断和药物不良反应。再如 Bohi 等人^[47]的工作中，设计了一个名为 PREMIER 的基于注意力的两阶段个性化药物推荐系统，该系统考虑了药物之间的相互作用，以最大程度地减少对患者的不利影响。

1.2.3 药物重定向研究

一般来说，药物研发从确定思路到投入市场需要 10~17 年的时间成本，资金成本在 20~30 亿美元^[48]。药物重定位，又称旧药新用，是一种用于发现已有药物或正在研发的药物超出原始批准的适应症、扩大其使用范围和用途的方法，具有高效、低成本、无风险的特点^[49]。药物重定位的研究不仅大大减少了新药开发的周期，还降低了经济成本，越来越受到医药界的重视。相关研究以基于临床试验的药物发现与基于计算机技术的药物知识挖掘两种为主，前者的研究基于临床试验，多以探讨药物分子和细胞受体的作用关系为研究方法，临床模型的建立是发现药物潜在作用的主要手段；后者以计算机技术为支撑，研究对象为科学数据间的相关关系，利用计算机构建数据模型是其进行药物知识发现的主要手段^[50]。

下面对基于计算机技术进行药物知识挖掘的药物重定位进行介绍。基于计算机技术进行药物重定向的研究主要包括基于网络药理学的方法、基于机器学习的方法、基于数据挖掘的方法等主要策略：基于网络药理学的方法主要包括基于小分子（配体）、基于药物靶点、基于网络理论三种研究策略^[51]；基于机器学习的方法主要包括基于传统机器学习、基于深度学习、基于矩阵分解等主要策略；基于数据挖掘的方法主要是利用检索及语义推断技术，对现有文献或医学数据库中的信息进行挖掘，从而得到未知实体间的关系。

近年来，许多专家对中药以及中药汤剂的重定向研究进行了探讨。如 Chen 等人^[52]为了发现四物汤的新药理作用，从 GEO 和 CMAP 数据库得到四物汤和 1309 个小分子药物的基因表达谱，利用分子对接技术辨识四物汤的有效成分群，其结果发现四物汤具有抗乳腺癌的作用，并通过文献调研验证了辨识结果的可靠性，将为扩大四物汤的临床应用范围及质量控制奠定基础，为乳腺癌的治疗提供新方法。Yang 等人^[53]的研究中，对丹参三七对在胰岛素抵抗的保护机制进行了探讨，通过网络药理学相关分析发现，人参皂苷 F2、原儿茶酸和丹酚酸 B 通过激活 AMPK 磷酸化和上调胰岛素抵抗细胞模型 HepG2/IR 中 GLUT4 的表达来促进葡萄糖消耗，这揭示了丹参三七对在胰岛素抵抗的治疗中的潜力。Song 等人^[54]的研究结合了

深度学习的方法对中药分子和 SARS-CoV-2 靶标进行重新定位, 在 KIBA 数据集上构建了带残差模块的深卷积神经网络 DCNN-RES, 并对其进行了训练, 其结果对预测药物靶点对结合亲和力的准确率为 85.33%, 并发现甘草和黄芩中的分子与 SARS-CoV-2 的靶标具有很强的结合亲和力, 这也与最新的研究结果一致。

从某种意义上讲, 对中药与中药汤剂的重定位也是一种中医处方推荐的思路。在国民的日常生活中, 到社区诊所、药房买药是普通老百姓相对于去医院就诊而言更加青睐的, 面对常见病与慢性病时采用的治疗策略, 如遇到普通流感、咽喉不适、普通腹泻等常见病, 以及高血压、糖尿病等慢性病时, 人们通常选择就近药店在药房人员的指导下购买对症药物进行治疗, 这更是贴近老百姓日常生活的诊疗方式, 明晰现有药物的适应症, 是满足老百姓“买对药”、“能治病”的需求的重要前提, 只有明确现有药物的适应症, 才能为普通居民买药时为其提供相对合适、治疗有效的药物。同时, 就中成药重定位研究而言, 将中成药二次开发作为中药研发的方向之一, 有助于在中药研发领域创造新价值^[55]。因此, 对现有中成药或中药汤剂的适应症进行完善, 开展中成药重定向研究, 无论是对于满足居民日常生活需要, 还是对于降低药物研发成本, 以及推进中医药现代化、创新化发展进程, 都具有不可替代的积极意义。

1.3 本文主要工作

本文以中医临床医案数据、症状本体库、药症知识图谱等领域知识为基础, 针对目前中医处方推荐领域中存在的问题, 结合数据增强、症状术语映射、表型相似性进行中医处方推荐的方法研究。本文主要工作阐述如下。

(1) 现阶段的研究过程中虽已积累了大量的处方数据, 但是各处方下对应的样本数不足以用于分类任务, 即处方数据的“一多一少”问题。针对这一问题, 本文对临床医案数据进行了预处理和筛选, 并以医案数据为依托构建了症状共现数据集, 并结合症状本体库和相关领域知识形成了症状同义关系集和药症知识图谱。在此基础上, 本文提出了两种临床诊疗数据增强的策略, 将增强后的数据应用于多标签处方推荐任务, 并与数据增强基线方法的性能进行了对比。

(2) 目前已有的处方推荐方法的性能普遍不够理想, 并且对于患者症状中存在的“未登录词”如何表示的问题, 尚无较理想的解决方案。针对这一问题, 本文结合症状术语映射与深度学习技术提出了 TCMPR 模型。首先对原始症状术语通过子图抽取的方式进行术语映射, 然后结合已训练的症状网络的嵌入向量进行患者特征表示, 进而将形成的患者症状表示通过构建的深度模型进行训练, 得到对每味候选药物的预测概率。本文对提出的 TCMPR 模型与多标签分类的基线模型进行了性能对比, 并对 TCMPR 模型中的关键模块进行了实验探究。

(3) 现有处方推荐方法得到的推荐结果多为药物组合, 即忽略了药物之间存在的相互作用, 不符合真实世界开方配伍原则, 结果的配伍合理性不强。针对这一问题, 本文结合经典名方数据与表型相似度匹配方法进行经典名方推荐的方法探索。本文首先将处方推荐任务由多标签分类问题改为多分类问题, 即在推荐结果上推荐中医经典方剂, 来取代原有的对单个药物的预测。本文利用传统机器学习中多分类任务的经典模型进行了实践, 并结合了本文提出的数据增强模型对经典名方数据进行了增强以提升模型性能。但这种方案不是最优的, 因为其性能受限于经典名方样本, 因此本文提出了基于表型相似度与经典名方的处方推荐方法, 并结合医案数据和经典名方数据进行了相关实验。

本文研究工作的创新之处包括以下三方面。

(1) 本文提出的基于症状本体库和症状共现的 **SOCO** 模型和基于药症知识图谱采样的 **SabKG** 模型能够较好的融合领域知识, 以辅助临床诊疗数据增强和提升后续多标签处方推荐任务的性能。

(2) 本文提出的 **TCMPR** 模型充分利用了已形成的药症知识图谱和症状网络等领域知识, 能够较好的形成对症状“未登录词”的表示, 并且构建的模型能够提升处方推荐任务的性能。

(3) 本文形成的基于表型相似度与经典名方的策略融合了症状网络等领域知识形成了患者的症状特征, 并且能够提升推荐结果的配伍合理性和推荐性能。

1.4 本文组织结构

第一章对本文的研究背景和意义进行了介绍, 列举了中医智能处方推荐近年来的相关工作, 对本论文的主要工作进行了阐述, 并介绍了本论文的组织结构。

第二章对本文研究密切相关的研究方法进行了介绍, 首先对目前现有处方推荐中常用的多标签分类和多分类方法进行了介绍, 进而介绍了与本文研究相关的其他相关方法, 包括深度学习方法、文本数据增强方法和图表示学习等内容。

第三章对本文形成的基于数据增强的处方推荐策略进行了阐述。首先详细介绍了对临床医案数据的处理, 以及症状共现集、症状同义关系集以及药症知识图谱的形成过程。之后阐述了本文提出的基于症状本体库和症状共现的处方数据增强策略, 以及基于知识图谱采样的处方数据增强策略。最后进行了相关实验及分析, 以验证本文所提出方法的有效性。

第四章对本文提出的基于症状术语映射和深度学习的处方推荐方法进行了阐述。首先对本文构建的 **TCMPR** 框架及其实现细节(症状网络的形成、症状嵌入表示、症状术语映射以及症状特征融合)进行了详细介绍。本文将提出的方法与相关基线方法进行了对比, 同时对 **TCMPR** 模型中的关键模块进行了相关实验及分析,

探究了影响模型性能的主要因素。

第五章对本文构建的基于表型相似性的处方推荐方法进行了阐述。首先本文对形成的基于经典名方的处方推荐策略进行了阐述，并设计了相关实验对该策略进行实现。而后提出了基于表型相似度与经典名方的处方推荐策略，对提出的症状特征构建策略以及表型相似度匹配策略进行了详细阐述，最后进行了实验探究及结果分析，以对比不同策略下经典名方推荐性能的差异。

第六章对全文的工作进行了总结，并对未来研究工作进行了展望。

2 相关研究及方法基础

本章将对与本文研究密切相关的基本方法进行介绍。首先介绍多分类任务与多标签分类任务及其经典方法，这是本文开展处方推荐工作的基础与关键。然后对本文所使用到的其他相关技术进行简要介绍，包括深度学习技术、文本数据增强方法以及网络表示学习方法等内容。

2.1 多分类与多标签分类方法概述

现有处方推荐策略大多将处方推荐这一任务视为多标签分类问题（每味药物是一个单独的类别，因每位患者所用中药通常为多个，故每位患者可能属于多个类别），若以方剂名称作为患者所属类别，则处方推荐问题可转化为多分类任务。上述两种思路在本文的研究中均已考虑，因此本节对多分类与多标签分类任务及其经典方法进行介绍。

2.1.1 多分类问题与多标签分类问题定义

分类问题是计算机科学领域的主要研究内容之一，其核心流程是在给定输入下预测其所属的类别，即预测建模问题，其中给定输入即为特征、输出的类别也称为标签。对于类标签，传统方法假定输入只属于一个类标签，也就是在多个标签中选择其中一个作为预测结果，这种假定下标签间是互斥关系，称为多分类问题（Multi-class Classification）。

假定 $\mathbf{X} = \mathbb{R}^d$ 代表 d 维输入实例空间， $\mathbf{Y} = \{y_1, y_2, \dots, y_q\}$ 代表标签空间（标签个数为 q ，且 $q > 2$ ），则多分类任务可描述为：利用构建的训练集 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 进行学习，从而建立出从输入空间 \mathbf{X} 到输出标签空间 \mathbf{Y} 的映射，如公式2-1所示。

$$f: \mathbf{X} \rightarrow \mathbf{Y} \quad (2-1)$$

但是现实情况下，有许多学习任务和数据不符合上述的假设，原因在于现实世界的对象可能是复杂的，可能同时隶属于多个类标签，例如在一篇新闻中可以涵盖体育、冬奥会、火炬传递、冰上表演等多个主题，再如一张图片中可能包括河流、高山、小鸟、绿树等多种元素，在这些情境下如果依然采用上述假设是不合理的。一种解决方法是将一组适当的标签分配给该输入对象，以涵盖该输入对象所包含的信息，这时输入所对应的类标签可能不止一个，并且标签间是不互斥的，

称为多标签分类问题^[56]。

假定 $\mathbf{X} = \mathbb{R}^d$ 代表 d 维输入实例空间, $\mathbf{Y} = \{y_1, y_2, \dots, y_q\}$ 代表标签空间 (标签个数为 q), 则多标签分类任务可描述为: 利用构建的多标签训练集 $D = \{(x_i, y_i) | 1 \leq i \leq m\}$ 进行学习, 从而建立出从输入空间 \mathbf{X} 到输出标签空间 \mathbf{Y} 的多标签的映射, 如公式2-2所示,

$$f: \mathbf{X} \rightarrow 2^{\mathbf{Y}} \quad (2-2)$$

其中, 多标签数据集的 $x_i \in \mathbf{X}$ 是 d 维的特征向量 $(x_{i1}, x_{i2}, \dots, x_{id})^T$, $y_i \subseteq \mathbf{Y}$ 是与 x_i 对应的一组类别标签。

2.1.2 多分类经典算法

机器学习领域中已形成了诸多经典的多分类算法, 如支持向量机、决策树、多层感知机等方法。下面对支持向量机和决策树两种经典算法进行简要介绍。

本节以线性可分支持向量机为例进行介绍支持向量机^[57] 算法。考虑一个二分类问题, 假设给定空间上的训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $x_i \in \mathbf{X}$, $y_i \in \mathbf{Y} = \{+1, -1\}$, 即 x_i 为第 i 个特征表示, y_i 为 x_i 的类标签, 当 $y_i = +1$ 时, 称 x_i 为正例, 当 $y_i = -1$ 时, 称 x_i 为负例。学习的目标是在现有特征空间中找到一个超平面, 用以将样本分离成不同的类, 如图2-1所示。

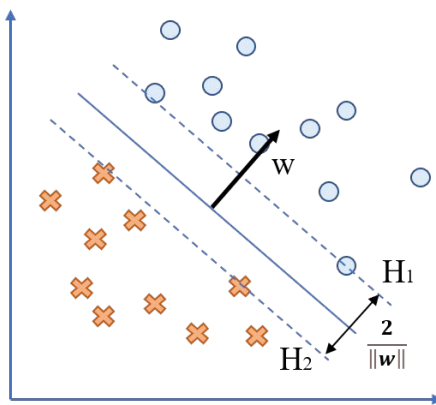


图 2-1 支持向量机算法示例

Fig 2-1 Example of support vector machine method

线性可分支持向量机的策略是尽可能找到能将数据正确分类并且类别间的间隔最大的直线, 其学习过程对应的最优化问题定义为公式2-3所示,

$$\begin{aligned} \min_{\omega, b} \quad & \frac{1}{2} \|\omega\|^2 \\ \text{s.t.} \quad & y_i(\omega \cdot x_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned} \quad (2-3)$$

其中 ω , b 分别是分离超平面 $\omega \cdot x + b = 0$ 的法向量和截距。

对于多分类问题, 领域内学者也提出了将支持向量机用于多分类任务的策略^[58], 思想主要分为两类: 一是直接在目标函数上进行修改, 即把多个分离平面的参数求解合并成一个优化问题中, 通过求解该优化问题的方式来实现多分类, 但其计算复杂度较高, 适用性不强; 另一种思想是通过组合多个二分类器来实现多分类器的构成, 包括“一对多 (One-vs-Rest)”法 (思想是对每个类别进行训练时将该类别样本归为一类、其余类别涉及样本归为一类, 即构建的分类器数目与类别数目相同)、“一对一 (One-vs-One)”法 (思想是任意两类别间建立一个分类器) 等。

决策树算法^[59] 亦是一种经典的分类方法。如图2-2所示, 决策树呈树形结构, 其思想是基于给定特征进行样本的分类, 其本质是“**If-Then**”规则的集合。决策树的优点在于可读性强、分类速度快。

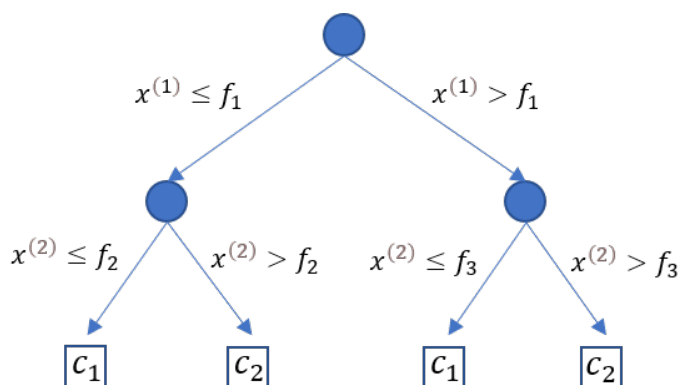


图 2-2 决策树算法示例

Fig 2-2 Example of decision tree method

其通常以递归地选择最优划分特征作为学习方式, 首先构建根节点 (根节点中包含所有训练数据), 然后以计算信息增益等方式选择一个最优特征, 进而根据这个最优特征进行数据集分割, 如果分割后各子集能够被基本正确分类, 则构建叶子节点并分配各子集到叶子节点中, 如果仍有子集不能被正确归类, 则继续在此子集中选择出新的最优特征, 并进行进一步的划分, 直到所有训练子集能够基本分类正确或者无法找到合适的特征时结束构建。决策树对训练数据可能具有较好的分类能力, 但是对于测试数据易产生过拟合现象, 因此在已生成的树结构基础上可以进行剪枝操作以缓解过拟合问题。

2.1.3 多标签分类经典算法

多标签分类问题有两种主要的解决方式^[56]：一是问题转化的方式，即让数据适应算法，经典算法包括 Binary Relevance^[13]、Classifier Chains^[17,60]、Random k-labelsets^[61,62]等；二是算法改编的方式，即让算法适应数据，经典算法包括 MLKNN^[14]、MLDT^[15]、Rank-SVM^[16]等。下面本节对分类器链方法 Classifier Chains 和多标签 K 近邻算法 MLKNN 进行介绍。

分类器链算法（Classifier Chains, CC）的核心思想是将多标签学习转化为二分类器的链式组合，其中分类器链上后续的二分类器是基于前面已形成的分类器链而形成的，如图2-3所示。其优点是实现简单，易于理解，并且是一种高阶的方法，以随机的方式考虑所有类别标签之间的相互关系，能够对标签间的相关性进行考虑。但是这种方法对于各标签的输入顺序较敏感。

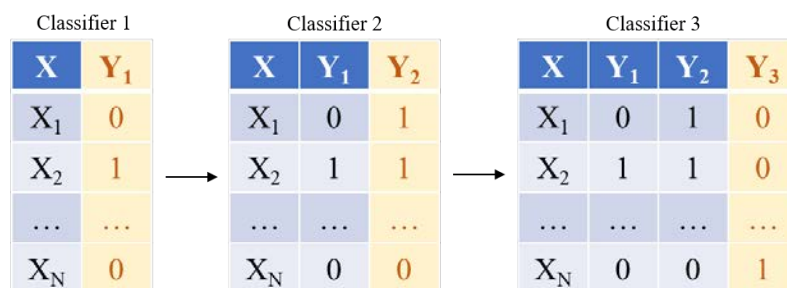


图 2-3 分类器链算法示例

Fig 2-3 Example of classifier chains method

多标签 K 近邻算法 MLKNN 是一种经典的多标签分类算法，其借鉴了 KNN 算法（如图2-4所示）的思想^[63]。首先通过 KNN 算法得到样本最接近的 K 个邻近样本，然后根据 K 个邻近样本的标签，统计属于某一标签的邻近样本个数，最后利用最大后验概率规则对邻居中包含的标注信息进行推理来预测所属标签。MLKNN 算法继承了懒惰学习和贝叶斯推理的优点：由于为每个不可见实例标识的邻居不同，因此可以自适应地调整决策边界；由于每个类标签估计的先验概率，类不平衡问题可以在很大程度上得到缓解。

2.2 深度学习方法概述

深度学习（Deep Learning, DL）是机器学习的一个子问题，其主要目的是学习到输入数据有效的特征表示^[64,65]。目前深度学习以神经网络模型为主要载体，主要原因在于神经网络模型具有超强的表示能力（特别是对于非线性模型的学习和表示），并且其泛化能力和容错能力较强。随着人工智能领域的第三次浪潮的兴起，

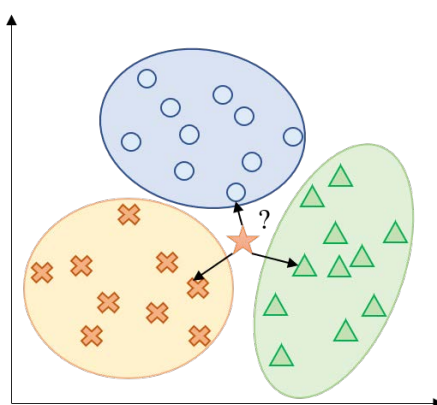


图 2-4 K 近邻算法示例

Fig 2-4 Example of K-nearest neighbor method

深度学习技术在各领域中得到了广泛的应用，也在各领域的相关问题中取得了较好的效果，其中深度学习在医学领域的应用亦是不可胜数。本文的研究中结合了深度学习相关技术辅助处方推荐任务，因此本节对本文使用到的深度学习相关技术进行介绍，包括前馈神经网络、卷积神经网络和注意力机制。

2.2.1 前馈神经网络

前馈神经网络（Feedforward Neural Network, FNN）是早期提出的人工神经网络之一，也经常称为多层感知机，如图2-5所示的前馈神经网络包含了一层输入层、两层隐藏层和一层输出层，各神经元分别隶属于不同层。前馈神经网络具有较强的表示能力，根据万能近似定理可知，对于包含至少一层“挤压性质”的隐藏层和一层线性输出层所组成的前馈神经网络，只要其隐藏层所包含的神经元个数足够多，它就可以近似拟合任意一个有界闭集函数，并且其近似精度也可以是任意的^[66]。但是前馈神经网络也具有不足之处，当神经层数及神经元个数较多时，虽然其表示能力十分强大，但是由于参数量巨大，训练效率会比较低，并且会产生过拟合的现象。

2.2.2 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）是受生物学上感受野机制的启发而提出的，能够弥补前馈神经网络的一些不足（参数量过多、难以提取局部特征等），其结构具有局部连接、平移不变性、参数共享等优点，基本结构如图2-6所示。

一个典型的卷积神经网络包含着卷积层、池化层、全连接层等模块，并且常以

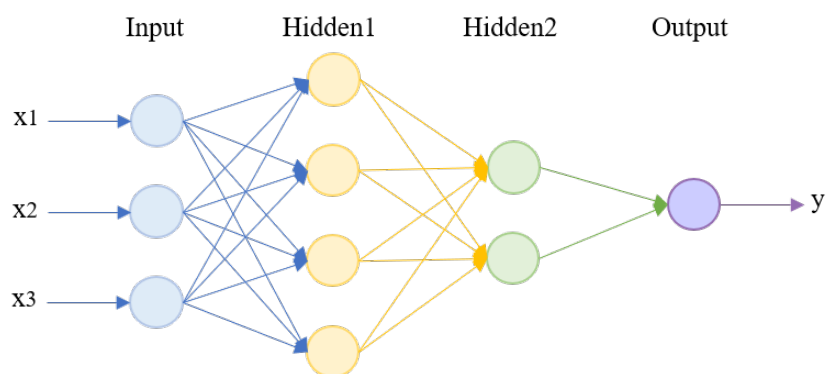


图 2-5 前馈神经网络示例

Fig 2-5 Example of feedforward neural network

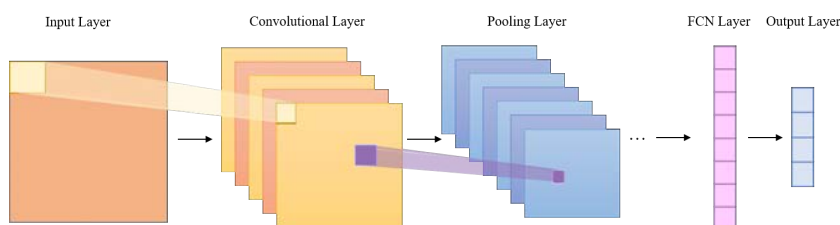


图 2-6 卷积神经网络示例

Fig 2-6 Example of convolutional neural network

交叉堆叠的形式组成，其中，结合了卷积运算的卷积层用于提取特征，特别是局部特征的提取；池化层用于降维，降低卷积层输出的参数量，从而缓解过拟合现象；全连接层对训练结果进行转化形成输出，在卷积神经网络中起到分类器的作用。

2.2.3 注意力机制

神经网络在处理大量输入数据的时候，也可以借鉴人脑的注意力机制进行运算，以提升神经网络的学习效率和性能。注意力的计算可分为两个过程（如图2-7所示）：一是计算所有输入信息的注意力分布 α_n ，二是利用计算得到的注意力分布赋予原始输入，以得到加权平均后的聚合特征^[66]。

图2-7展示了键值对模式下的注意力机制，其中“键”用以计算注意力分布 α_n （即注意力权重），“值”用来计算加权平均后的融合特征。

2.3 文本数据增强方法概述

本节对文本数据增强方法进行概述，首先简要介绍文本数据增强任务，然后对 EDA 文本数据增强方法进行介绍。

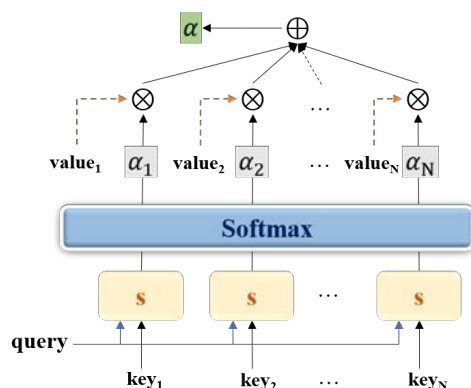


图 2-7 注意力机制示例

Fig 2-7 Example of attention mechanism

2.3.1 文本数据增强简介

基于深度学习的模型学习性能的高低往往取决于训练数据的大小和质量^[67]，而训练数据的收集与整理过程往往较为繁琐，高质量数据的形成需要长时间的收集、整理与加工，特别是对于医学领域，高质量的数据更不易形成。因此可结合文本增强方法对数据进行文本增强，而后运用于下游任务中。近年来相关学者对自然文本领域的文本增强进行了研究，如 2019 年 Wei 等人的研究提出的（Easy Data Augmentation, EDA）方法^[68]，该方法提出了 4 种简易的文本增强策略，取得了理想的效果。下面本节对 EDA 策略进行介绍。

2.3.2 EDA 方法

EDA 方法包括以下 4 种增强方式。

(1) 同义词替换（Synonym Replacement, SR）：从句子中随机选择 n 个非停用词，用随机选择的同义词中的一个替换这些单词。如图2-8中，将原始语句中的“sad”替换为同义词“lamentable”、将“back”替换为“backward”，形成了新的句子。

(2) 随机插入（Random Insertion, RI）：找出句子中不是停用词的词汇，随机选择其同义词插入句子中任意位置，此操作执行 n 次。如图2-8中，在随机位置插入了原句中“comedy”的同义词“funniness”，形成了新的句子。

(3) 随机置换（Random Swap, RS）：随机选择句子中的两个单词并调换它们的位置，此操作执行 n 次。如图2-8中，将原句中的“the”和“roads”两词进行了置换，形成了新的句子。

(4) 随机删除（Random Deletion, RD）：按概率 p 随机删除句子中的每个单词。

如图2-8中，随机删除了三个词汇，分别是“comedy”、“played”和“back”，形成了新的句子。

Operation	Sentence
Original	A sad, superior human comedy played out on the back roads of life.
SR	A lamentable , superior human comedy played out on the backward road of life.
RI	A sad, superior human comedy played out on funniness the back roads of life.
RS	A sad, superior human comedy played out on roads back the of life.
RD	A sad, superior human out on the roads of life.

图 2-8 EDA 文本数据增强方法示例
Fig 2-8 Example of EDA method

这种数据增强的策略采取增强的方式容易实现，并且具有以下优点：第一，在此过程中生成了与原始数据类似的增强数据，能够会引入一定程度的噪声，这样有助于缓解过拟合；第二，使用 EDA 可以通过同义词替换和随机插入操作引入新的词汇，允许模型泛化到未在训练集但存在于测试集中的数据；第三，小数据集在此策略上能够产生相对较好的性能。

2.4 网络表示学习概述

本节对网络表示学习方法进行概述，并介绍本文使用到的 DeepWalk 算法。

网络表示学习（Network Representation Learning）指将网络中的节点表示为低维的、稠密的、实值的向量形式这一学习过程^[69]，亦称图嵌入、网络嵌入，目标是形成能够蕴含原始网络结构和语义信息的节点表示，从而能够进行下游任务（节点分类、推理、链接预测等）。近年来许多专家学者在这一领域中开展了研究，提出了很多经典的网络表示方法，取得了相对较好的性能，如 DeepWalk^[70]、node2vec^[71]、Line^[72] 等方法。下面本节对 DeepWalk 算法进行介绍，其流程如图2-9所示。

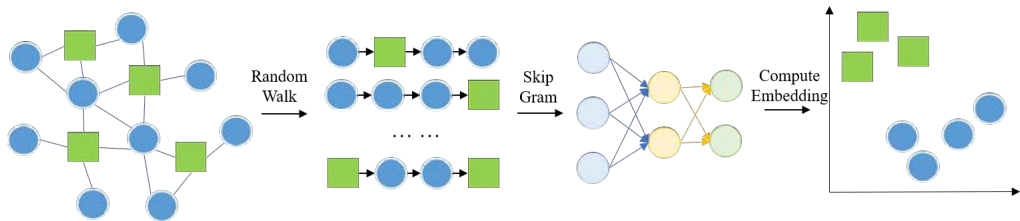


图 2-9 DeepWalk 算法流程
Fig 2-9 Structure of DeepWalk method

DeepWalk 算法的核心流程包括以下过程：首先在给定网络数据中进行随机游

走，产生随机节点序列，而后结合基于 Skip-Gram^[73] 的 Word2vec^[74] 方法对已形成的随机节点序列进行学习，形成网络中节点的表示向量，其核心就是利用截断的随机游走序列表示节点近邻，并将形成的序列看做是句子来作为 Word2vec 的输入来学习节点的向量表示。对于其随机游走策略，该方法首先对图中每个节点产生 γ 个随机游走序列，每个随机游走序列的长度为 t ，节点游走时，每个节点从与之关联的邻居节点中以相同概率任选一节点来进行下一步游走。

2.5 本章小结

本章对与本文研究密切相关的基本方法和相关技术进行了简要介绍，包括多分类与多标签分类任务及经典方法、深度学习技术、文本数据增强方法、网络表示学习等内容，这些方法是本文研究工作的基石，为本文研究工作的开展起到了关键作用。

3 基于数据增强的中医处方推荐方法研究

本章围绕中医临床数据的增强方法及其在处方推荐中的应用开展研究。本文首先对中医临床医案数据进行预处理和样本筛选,进而以此医案数据为依托,构建了症状共现数据集,同时结合了症状本体库和领域知识形成了症状表型同义关系集和药症知识图谱。在此基础上,本文提出了两种对临床诊疗数据进行增强的策略,分别为基于症状本体库与症状共现集的临床数据增强方法 (Clinical Data Augmentation based on Symptom Ontology library and symptom Co-Occurrence set, **SOCO**) 和基于知识图谱采样的临床数据增强方法 (Clinical Data Augmentation based on Sampling by Knowledge Graph, **SabKG**), 将增强后的数据应用于多标签分类任务下的处方推荐,并与数据增强相关基线方法的性能进行了对比。结果显示,本章提出的 **SabKG** 方法在性能上得到了最好的表现, **SOCO** 方法也能够使数据在推荐任务中的性能得到一定提升。

3.1 引言

处方数据质量是提升中医处方推荐方法性能的瓶颈,也是关键。近年来的临床实践中,虽然已经积累了大量的处方数据,但是数据的质量是亟待提升的,如经典名方数据,经典名方经过了几千年临床实践的不断考验,其临床证据体现在数不胜数的古籍文献和现代医学研究中,其自身具有宝贵的临床价值和研究价值。但经典名方等中医临床处方数据具有“一多一少”的特点:如果将中医处方推荐视为分类问题,即把中药处方看作待预测的类别,则尽管临床处方数据的总量可观,但是由于中药处方自身具有的复杂性,使得总量可观的临床处方数据中包含的中药处方数量相对较多,每条中药处方对应的样本数则过少(甚至常出现一条中药处方只有 1~2 条样本与之对应)。这种现象是不利于将分类方法运用于中医处方推荐中的。因此,提出适合中医临床数据的数据增强方法,并将其运用于处方推荐过程中,对于挖掘临床处方蕴含的知识具有积极意义。

3.2 数据处理与构建

本节主要展示本文所使用的临床医案数据的收集与预处理,药症知识图谱的构建等内容,并针对本章的相关研究内容,在症状本体库和医案数据的基础上构建了症状同义关系集和症状共现集。

3.2.1 临床医案数据预处理

首先,本文收集了用于构建处方推荐模型训练与评价的临床诊疗医案数据。临床诊疗医案数据来源于各大名老中医的经典临床诊疗案例。原始临床医案数据中主要包含医案文题、作者、具体日期、医生姓名、患者的脱敏后相关基本信息以及临床诊断和治疗信息等。从这些医案数据中提取了患者的临床诊疗信息,具体包含证候、治法、中西医诊断、处方名称、主诉症状,以及中医用药。原始数据共包含诊次信息 15845 条,数据涉及的期刊来源达 150 多种,包含了 753 位国家级名老中医的经典医案。

为了更好地使用临床医案中病人的临床症状与处方用药构建处方推荐模型,因此本文对原始数据的症状和中药名称进行了人工规范。同时,针对数据中存在“长尾分布”^[75] 的现象(即少数医案数据中症状数量与药物数量过多),本文根据症状数量与用药数量进行了临床病例的筛选。筛选标准为症状数量少于 40 且药物数量少于 20,筛选后获得了 8218 条数据作为实验数据。

此外,本文将原始数据与筛选后的 8218 条数据的症状数目分布与药物数目分布进行了对比。从结果来看,原始数据中包含样本个数 15845 个,涉及症状共 36847 个,涉及药物共 3359 个,病人症状平均数目是 13.40,用药平均数目是 12.06;按照上述阈值进行筛选后保留的 8218 条样本中,涉及症状共有 36145 个,涉及药物共有 2827 个,病人症状的平均数目是 11.45,药物平均数目是 11.31。筛选前后症状个数均近似表现出泊松分布,中药个数均近似表现出泊松分布,分布具体分布如图3-1所示。

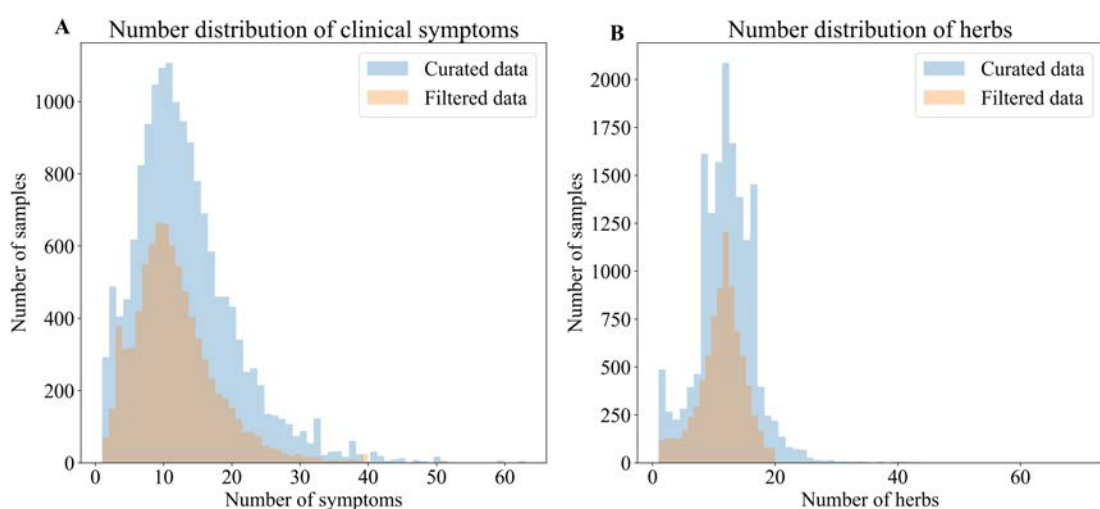


图 3-1 临床医案数据症状药物分布情况

Fig 3-1 Symptom distribution and herb distribution of clinical case data

3.2.2 基于医案数据的症状共现数据集构建

临床诊疗数据中症状表型间的共现关系蕴含着表型间的潜在关联,如表型“反胃”、“胃反酸”等与表型“呕逆”共现频度较高,均反映出人胃部不适、感觉恶心时所产生的的一组关联症状,这种以症状共现形式所存在的症状关联,正可用于症状表型的增强,因此本文基于临床医案数据构建了症状共现关系集,并在此基础上形成了症状的共现评分,以体现症状表型的共现情况和症状间的共现程度,实现流程如图3-2所示。



图 3-2 症状共现集与症状评分集构建

Fig 3-2 The construction of symptom co-occurrence set and symptom score set

首先本文根据医案数据中的患者症状列,形成了症状间的共现关系,需要说明的是,此处的共现指同一患者的症状组中,组内两两症状存在的关联即为症状共现关系。本节对形成的症状共现关系进行了汇总,并进行了不同症状对的共现频次统计,症状对的共现频次越高,说明该对症状共现的频率越高。在此基础上,本文利用症状共现集中的症状共现频度形成了症状评分,症状评分计算过程如 Algorithm 1所示。

从症状评分集的形成过程中可以看出,形成的症状评分代表了该症状在样本中出现的频次情况,以及该症状与其他症状的共现情况,评分越高,代表该症状出现的频次越高。通过上述方法形成的症状共现集和基于症状共现频度的症状评分,将用于后续对表型的数据增强中,根据症状共现关系的频度以及计算得到的症状评分进行症状表型的添加及删除。

3.2.3 基于症状本体库的症状同义关系集构建

症状术语间存在着一定的语义关联,特别是症状的同义关系,同一本体下的症状表型均具有同义关系,正可用以症状的替换来进行症状数据增强。本文利用症状本体库,根据各本体下的症状,形成了 3122 条症状同义关系,形成过程如图3-3所示。

Algorithm 1 基于症状共现频度的症状评分计算框架

Input: 基于共现频度的症状共现集 $Occurrence_{Sym}$;

Output: 症状评分集 $DScore_{Sym}$.

```

1: 初始化字典  $DScore_{Sym}$ ; 初始化总频次  $Sum_{Sym}$ .
2: for  $Sym_i$  in  $Occurrence_{Sym}$  do
3:    $Sum_{Sym} = Sum_{Sym} + Sym_i[freq]$ 
4:   if then  $Sym_i[Source_{Sym}]$  in  $DScore_{Sym}.keys$ 
5:      $DScore_{Sym}[Sym_i] = DScore_{Sym}[Sym_i] + Sym_i[Freq]$ 
6:   else
7:      $DScore_{Sym}[Sym_i] = Sym_i[Freq]$ 
8:   end if
9:   if then  $Sym_i[Target_{Sym}]$  in  $DScore_{Sym}.keys$ 
10:     $DScore_{Sym}[Sym_i] = DScore_{Sym}[Sym_i] + Sym_i[Freq]$ 
11:  else
12:     $DScore_{Sym}[Sym_i] = Sym_i[Freq]$ 
13:  end if
14: end for;
15: for  $Score_{Sym_i}$  in  $DScore_{Sym}$  do
16:    $Score_{Sym_i} = Score_{Sym_i} \cdot Sum_{Sym}$ 
17: end for
18: return  $DScore_{Sym}$ .

```



图 3-3 症状同义关系集构建

Fig 3-3 The construction of symptom synonymous relation set

3.2.4 药症知识图谱构建

为了挖掘和利用中医领域知识进行临床数据增强，本文构建了药症知识图谱，以便利用其形成的药症关系、症状同义关系等信息进行数据增强和下游推荐任务。

构建的药症知识图谱包含以下以下实体和关系。(1) 涉及 5 种实体，分别为中

药、症状、中药功效、中药性味、中药归经。(2) 涉及 5 种关系，分别为：药物-症状关系，源自《中华本草》；症状同义关系，源自症状本体库（即 3.2.3 节构建内容）；中药-中药功效关系、中药-中药性味关系、中药-中药归经关系，源自《2015 版中国药典》和《中华本草》。本文形成的药症知识图谱的实体、关系详细信息如表 3-1 所示，图谱可视化如图 3-4 所示。

表 3-1 药症知识图谱构成

Table 3-1 The construction of herb-symptom knowledge graph			
实体类型	实体数量	关系类型	关系数量
症状	8669	中药-症状关系	37528
中药	8464	症状-症状同义关系	3122
中药功效	1177	中药-中药功效关系	34268
中药性味	45	中药-中药性味关系	22244
中药归经	182	中药-中药归经关系	4858
实体总数	18537	关系总数	102120

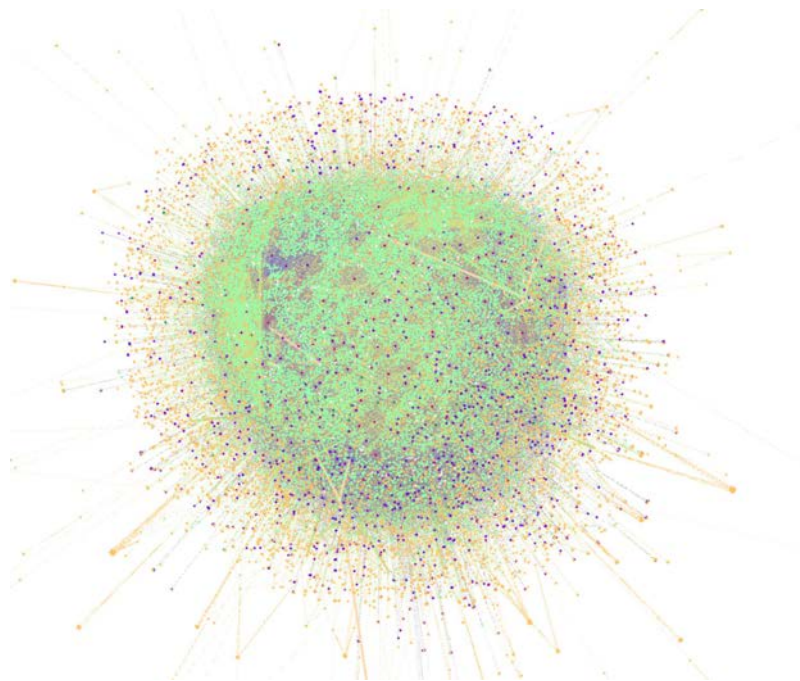


图 3-4 药症知识图谱可视化

Fig 3-4 The visualization of herb-symptom knowledge graph

3.3 结合领域知识的处方数据增强策略

本节将对提出的两种处方数据增强策略进行介绍，首先本文提出了基于症状本体库和症状共现频度的处方数据增强策略，该策略的主要思想是利用上述构建

的症状同义关系集和症状共现频度及得分，对临床诊疗数据中的症状部分进行相应增强。随后本文提出了基于知识图谱采样的处方数据增强策略，主要基于已构建的药症知识图谱，从临床诊疗数据的药物出发对症状数据进行增强。

3.3.1 基于症状本体库与症状共现关系的处方数据增强策略

本文结合症状同义关系集和症状共现频度及评分形成了处方数据增强策略，方法流程如图3-5所示。下面对提出该方法的动机及该方法的增强策略进行阐述。

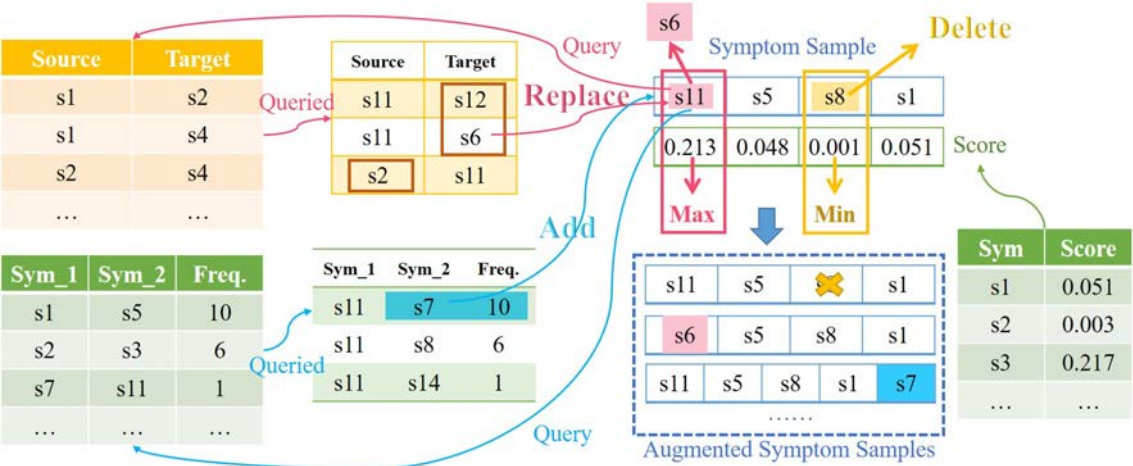


图 3-5 基于症状本体库与症状共现频度的处方数据增强策略

Fig 3-5 Clinical data augmentation based on symptom ontology library and symptom co-occurrence set

在临床实践过程中，对于症状的记录通常由医护人员手动形成，不同人对于同种症状的自然语言描述经常有一定差别。症状本体库对同一本体下不同术语进行了归类，体现了症状表型间的同义关系。如果利用基于症状本体库形成的症状同义关系进行原始症状词替换，这种基于本体的替换不仅符合临床实践，而且能够最大程度上减少由于对原始样本进行增强带来的噪声。

此外，症状间存在着依存关系，如症状表型“咳嗽”与“咳痰”时常共现于临床表现为咳痰、痰多的肺病患者中，但是在临床病历记录过程中，由于目前主要以人为记录为主要形式，对症状表型的记录存在易遗漏的现象。本文形成的症状共现集正可体现症状间的依存关系，用于症状数据的增加（对于共现频度较高的症状对）或去除（共现频度低或者出现频度较低的症状），符合临床实际情况。因此本文结合了症状同义关系集和症状共现频度集及共现评分，形成了对临床诊疗症状数据的增强策略。

本文提出的基于症状本体库与症状共现频度的处方数据增强策略（SOCO 模型）方法框架如 Algorithm 2 所示，具体细节阐述如下。

Algorithm 2 基于症状本体库与症状共现频度的处方数据增强策略

Input: 待增强样本 $Sample$, 症状共现频度集 $G_{Occurrence_{Sym}}$, 症状同义关系集 $G_{Synonymous_{Sym}}$, 症状共现评分集 $DScore_{Sym}$;

Output: 增强后样本集 $Augmented_{Sample}$.

```

1: 初始化样本集  $Augmented_{Sample}$ ; 初始化频度分数集  $Score_{Sym}$ ; 初始化最高分  $Max_{Score} = -1$ , 最低分  $Min_{Score} = 10$ ; 初始化最高分症状词  $Max_{Sym}$ , 最低分症状词  $Min_{Sym}$ .
2: for  $Sym_i$  in  $Sample[Symptom]$  do
3:    $Score_{Sym}[Sym_i] = DScore_{Sym}[Sym_i]$ 
4: end for
5: for  $Sym_i, Score_i$  in  $Score_{Sym}$  do
6:   if  $Score_i > Max_{Score}$  then
7:      $Max_{Score} = Score_i$ 
8:      $Max_{Sym} = Sym_i$ 
9:   end if
10:  if  $Score_i > Min_{Score}$  then
11:     $Min_{Score} = Score_i$ 
12:     $Min_{Sym} = Sym_i$ 
13:  end if
14: end for
15:  $Remove = Sample[Symptom]$ 
16:  $Remove.del[Min_{Sym}]$ 
17:  $Augmented_{Sample}.add = [Remove, Sample[Herb]]$ 
18:  $Replace = Sample[Symptom]$ 
19: Random  $Sym_{Replace}$  from  $G_{Synonymous_{Sym}}[Max_{Sym}].neighbor$ .
20:  $Replace.pop = Max_{Sym}$ 
21:  $Replace.add = Sym_{Replace}$ 
22:  $Augmented_{Sample}.add = [Replace, Sample[Herb]]$ 
23:  $Add = Sample[Symptom]$ 
24: Search  $Freq_{Max}$  in  $G_{Occurrence_{Sym}}[Max_{Sym}].neighbor$  with maximum Frequency
25:  $Add.add = Freq_{Max}$ 
26:  $Augmented_{Sample}.add = [Add, Sample[Herb]]$ 
27: return  $Augmented_{Sample}$ .
```

(1) 症状词添加: 在当前样本的症状组中, 寻找症状共现评分最高的症状词, 然后在症状共现关系集中寻找与之相关的共现对, 并选择与之共现频度最高的症状

(如频度最高的症状有多个, 则随机选取其中一个症状词并返回), 作为新症状加入到原始症状组中, 并与原始样本对应的药物组结合, 形成一条新的患者样本。

(2) 症状词替换: 在当前样本的症状组中, 寻找症状共现评分最高的症状词, 然后在症状本体库中寻找与之相关的同义关系词, 并在找到的同义关系词中随机选取一个症状词, 替换该最高分症状词, 并与原始样本对应的药物组结合, 形成一条新的患者样本。此处如果未找到与最高分症状词相关的同义关系词, 则不做替换(即不形成增强样本)。

(3) 症状词删除: 在当前样本的症状组中, 寻找症状共现评分最低的症状词, 将其删除, 并与原始样本对应的药物组结合, 形成一条新的患者样本。若评分最低的症状不止一个, 则随机选择一个进行删除。

由上述策略不难得知, 如果以上三种增强方式均最多产生一条数据, 则结合后对数据进行增强, 能够生成原始样本数目 2~3 倍左右的新样本。

3.3.2 基于知识图谱采样的处方数据增强策略

本文结合已构建的药症知识图谱形成了基于知识图谱采样的处方数据增强策略, 方法流程如图3-6所示。下面对提出该方法的动机及该方法的增强策略进行阐述。

本文提出的 SOCO 模型虽然能够利用症状相关信息进行增强, 但均基于症状层面进行了相关处理。实际情况中, 患者的症状与对症药物之间存在一定关联, 仅对症状进行增强脱离了中药-症状之间存在的关联。因此, 结合药症知识图谱中存在的药症关系对临床数据进行增强, 不仅能够结合现阶段已形成的以药症关系为主的知识, 而且更贴切临床实际。因此本文提出了基于知识图谱采样的处方数据增强方法。

本文提出的基于知识图谱采样的处方数据增强策略(SabKG 模型)方法框架如 Algorithm 3所示, 具体细节阐述如下。

(1) 对于待增强样本, 首先利用样本的中药组成部分, 以其各味中药为节点, 在已构建的药症知识图谱中查找与各味中药直接相连的症状节点(即查找各中药的一阶症状邻居), 将查找到的症状进行汇总, 同时统计各症状出现的频次;

(2) 使用待增强样本中的症状部分, 与上一步形成的症状频次表中涉及的症状词进行文本匹配(即对前面得到的症状频次表进行筛选, 此处需设置匹配阈值);

(3) 利用匹配后保留的症状频次表进行采样, 首先使用该症状频次表中各症状的频次转为概率, 然后以此概率进行随机采样(这里涉及两个参数, 一是采样后形成的样本个数, 二是采样时每条样本包含的症状个数);

(4) 采样后形成的症状集合, 与原始药物集合拼接后, 形成为增强后的样本。

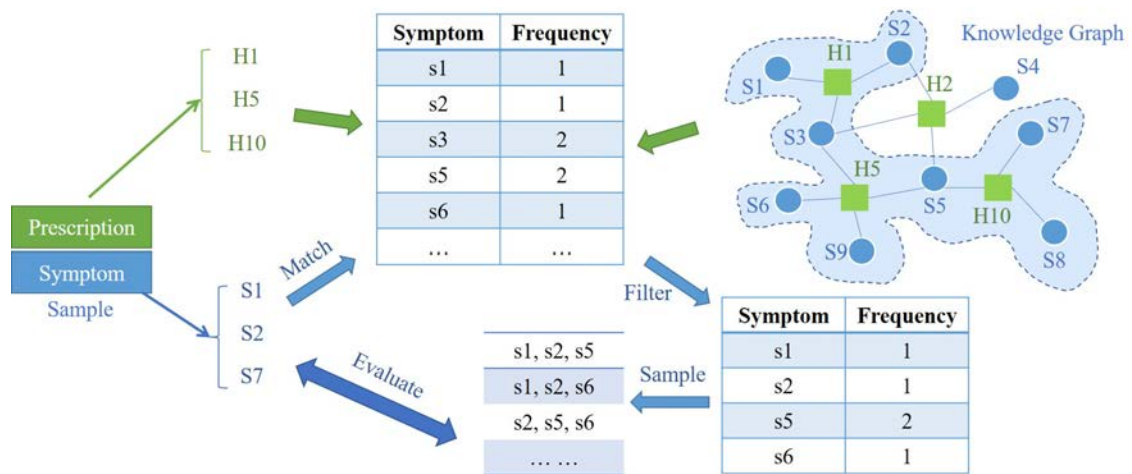


图 3-6 基于知识图谱采样的处方数据增强策略

Fig 3-6 Clinical data augmentation based on sampling by knowledge graph

由上述策略不难得知，以此策略对数据进行增强，生成的增强样本数目由采样后形成的样本个数决定。

3.4 实验设计及验证

为探究本文提出的两种数据增强策略的有效性，本文将提出的数据增强方法与文本数据增强中的相关基线方法进行了对比。使用上述方法对医案数据进行了数据增强，并将增强后的样本和原始样本运用于多标签处方推荐任务中，以对比在多标签处方推荐任务下使用各增强方法产生的增强数据对推荐任务带来的影响。

3.4.1 实验方案设计

本节使用了预处理后的 8218 条临床医案数据进行实验，实验训练数据与测试数据划分为随机划分，训练集：测试集 = 8：2，即形成 6574 条训练样本和 1644 条测试样本。对 6574 条训练集样本，使用各种增强策略进行数据增强，并将增强前后的相关数据运用于多标签处方推荐任务中，并使用 1644 条测试集数据进行统一的测试，以更好地对比不同数据增强方法对推荐任务的影响。

多标签处方推荐任务定义如下。利用患者症状数据作为输入，输出为每味候选药物的预测概率，即将每味中药视为不同的类别，分别进行判断该组症状是否属于当前药物类别，因各症状组下所属的药物可能不止一味，故本问题应归属为多标签分类问题。具体说来，首先利用网络嵌入表示 DeepWalk 算法对已构建的药症知识图谱中各症状形成嵌入表示，之后将训练集的患者症状与药物输入到模型进行训练，训练结束后对测试集的症状样本作为输入进行药物预测，并将得到的

Algorithm 3 基于知识图谱采样的处方数据增强策略

Input: 待增强样本 $Sample$, 已构建的药症知识图谱 G_{HSG} , 症状匹配阈值 Sim_{Sym} , 待形成的增强样本数目 Num_{Sample} , 每条增强样本包含的症状个数 Num_{Sym} ;

Output: 增强后样本集 $DAugment_{Sample}$.

```

1: 初始化样本集  $DAugment_{Sample}$ , 症状频次集  $DFreq_{Sample}$ .
2: for  $herb_i$  in  $Sample[Herb]$  do
3:   for  $Symptom_j$  in  $G_{HSG}.neighbor(herb_i)$  do
4:     if  $Symptom_j$  not in  $DFreq_{Sample}.keys$  then
5:        $DFreq_{Sample}[Symptom_j] = 1$ 
6:     else
7:        $DFreq_{Sample}[Symptom_j] = DFreq_{Sample}[Symptom_j] + 1$ 
8:     end if
9:   end for
10: end for
11: for  $Sym, Freq$  in  $DFreq_{Sample}.items$  do
12:   if  $Freq < Sim_{Sym}$  then
13:      $DFreq_{Sample}.pop[Sym]$ 
14:   end if
15: end for
16: Random  $Num_{Sample}$  Samples with  $Num_{Sym}$  symptoms.
17: for  $Augmented_{Symptom}$  in  $Samples$  do
18:    $DAugment_{Sample}.add([Augmented_{Symptom}, Sample[Herb]])$ 
19: end for
20: return  $DAugment_{Sample}$ .

```

预测结果与原始药物进行评价。本节中使用的训练模型是具有三层的全连接网络，第一层隐藏层的神经元个数是 256，第二层隐藏层的神经元个数是 64，第三层的神经元个数与候选药物总数相同，而后通过一层 SoftMax 激活层进行转化，即得到对每味候选药物的预测概率。

对于数据增强方法，本节将提出的两种策略与文本数据增强基线方法 EDA 进行对比。根据 EDA 方法提及的策略，本节设计了以下 4 个基线实验：症状同义词替换 SR，症状近义字替换（Synonym word Replacement, SR_word），症状随机字删除 RD，症状近邻字置换 RS。对于提出的 SOCO 模型，本节设计了三个消融实验，分别是利用在 SOCO 模型下三种不同的增强方式对应得到的增强样本进行各自实验（分别命名为 SOCO_A, SOCO_D, SOCO_R）。

3.4.2 评价指标与参数设置

由于本章的实验基于多标签分类问题，因此评价指标采用多标签分类相关的 Top@K 系列指标^[76]。下列公式中， i 表示第 i 个测试样本， $T(i)$ 表示第 i 个测试样本中对应的真实药物组， $R(i)$ 表示对第 i 个样本预测到的药物组， N 表示测试集样本数目， K 表示 Top@K 推荐得到药物组的长度。

(1) Precision@K，公式如 3-1 所示。

$$Precision@K = \frac{\sum_{i=1}^N |R(i) \cap T(i)|}{\sum_{i=1}^N |R(i)|} \quad (3-1)$$

(2) Recall@K，公式如 3-2 所示。

$$Recall@K = \frac{\sum_{i=1}^N |R(i) \cap T(i)|}{\sum_{i=1}^N |T(i)|} \quad (3-2)$$

(3) $F_1 - score@K$ ，公式如 3-3 所示。

$$F_1 - score@K = \frac{2 * Precision@K * Recall@K}{Precision@K + Recall@K} \quad (3-3)$$

3.4.3 实验结果及分析

本节对提出的两种数据增强方法以及相关基线方法在临床医案数据集上进行了数据增强，使用增强前后的数据分别进行了多标签处方推荐任务，并采用 Top@K 指标对推荐结果进行了评价，各方法下产生的增强样本数量如表3-2所示，汇总后的实验结果如表3-3所示。

为了尽可能减小因样本数量的差异带来的实验性能的影响，因此对各方法下产生样本的数量进行了控制，使得各方法下产生的数据量都大致在原始样本数量的三倍。具体说来，四种基线方法产生的样本数量的控制条件是：用每条原始样本生成最多 3 条增强样本；对于 SOCO 方法中的删除、替换、添加三种策略均最多产生一条样本，因此三种方法汇集后产生的总样本数量理论上在原始样本的三倍左右；对于 SabKG 方法，控制其产生样本数目的参数为 3，则其产生的样本数目从理论上亦为原始样本的三倍左右。此外，对于 SabKG 方法，设置每条增强样本包含的症状个数均与原始样本数目相同，这种设置使得增强后样本与原始样本在症状数目上差距较小，这也是其他方法较难做到的。

从各方法下的实验性能而言，本文提出的 SabKG 方法在 Top@5、Top@10 和 Top@20 下，无论是准确率、召回率还是 F1 值均取得了相对最优的效果。从表3-3中结果可以看到，本文提出的 SabKG 方法做增强后的数据性能与原始样本的性能相

表 3-2 各增强策略下产生的样本数

Table 3-2 Augmented sample amount under each augment strategy

Method	Sample Amount
Origin	6574
SR	18991
SR_word	19722
RD	19600
RS	19701
SOCO-A	12587
SOCO-D	13060
SOCO-R	8499
SOCO	21051
SabKG	17758

表 3-3 各增强策略下实验性能结果

Table 3-3 Experimental performance results under each augment strategy

Top@K	P@5	P@10	P@20	R@5	R@10	R@20	F1@5	F1@10	F1@20
Origin	0.2869	0.2362	0.1763	0.1296	0.2100	0.3123	0.1785	0.2223	0.2253
SR	0.2880	0.2344	0.1748	0.1302	0.2089	0.3086	0.1794	0.2209	0.2232
SR_word	0.2910	0.2359	0.1778	0.1305	0.2086	0.3145	0.1802	0.2214	0.2272
RD	0.2892	0.2351	0.1787	0.1300	0.2082	0.3174	0.1794	0.2208	0.2287
RS	0.2845	0.2341	0.1772	0.1284	0.2085	0.3143	0.1770	0.2205	0.2266
SOCO-A	0.2805	0.2336	0.1764	0.1263	0.2075	0.3115	0.1742	0.2198	0.2253
SOCO-D	0.2837	0.2321	0.1732	0.1274	0.2067	0.3061	0.1758	0.2186	0.2212
SOCO-R	0.2844	0.2362	0.1764	0.1278	0.2089	0.3106	0.1763	0.2217	0.2250
SOCO	0.2805	0.2336	0.1764	0.1263	0.2075	0.3115	0.1742	0.2198	0.2253
SabKG	0.2923	0.2382	0.1806	0.1320	0.2114	0.3196	0.1818	0.2240	0.2308
Improvement	1.908%	0.850%	2.433%	1.858%	0.628%	2.342%	1.874%	0.732%	2.400%

比，每种评价指标在 Top@20 上均能够提升 2.4% 左右的性能，而其它方法中虽不乏能够使原始样本性能得到提升的，但是提升效果不如本文提出的 SabKG 方法。

图3-7中展示了 SabKG, SOCO 方法与原始样本性能比较情况 (Top@1~Top@20)。从图中可以看出，本文提出的 SabKG 方法取得了最好的性能，尤其在 F1-score@K 上，相比于原始数据能够得到相对明显的性能提升。SOCO 方法在准确率上也能够得到一定提升，并且在 K 较大时 (K>12)，SOCO 方法在 F1-score 上也能够达到与原始数据性能相当的效果，但是其提升效果不如 SabKG 方法明显。总体而言，SabKG 方法不仅能够使原始数据性能得到较好提升，而且这种策略基于已构建的药症知识图谱，能够利用到潜在的药症关系，从而在知识层面上对原始数据的性能进行提升。

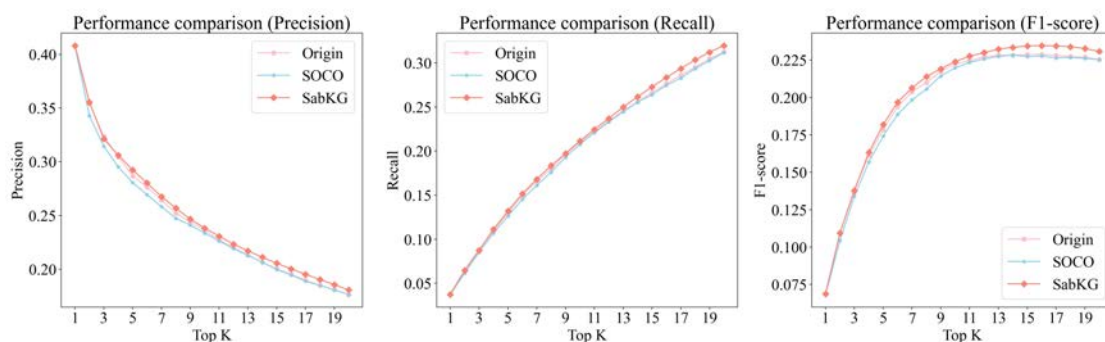


图 3-7 SabKG, SOCO 方法与原始样本性能对比

Fig 3-7 Performance comparison between SabKG, SOCO and the original sample

图3-8展示了原始数据与四种基线方法的性能对比情况。原始数据与基线方法的对比结果显示，基线方法中的近义字替换，随机字删除能够使数据性能得到小幅度的提升，但是这些基线方法的可行性有待考虑，尤其是对于近邻字置换、随机字删除和近义字替换三种方法，这三种方法对于原始症状样本的改变幅度虽然不大（仅仅是单个症状词的删除或者症状词中相应汉字的替换或位置互换），但是这种策略使得原始症状词的实际临床含义可能会发生较大的变化，会破坏原始症状词的自然语言含义，与临床实际应用的贴合性不强。

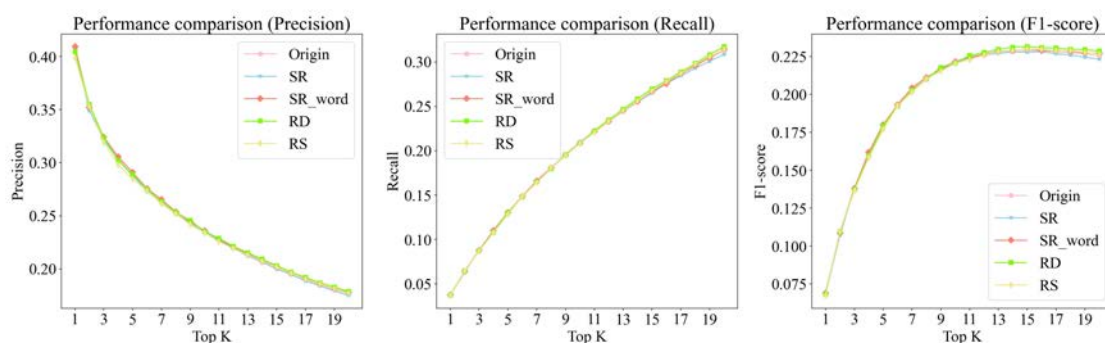


图 3-8 四种基线方法与原始样本性能对比

Fig 3-8 Performance comparison between 4 baseline methods and the original sample

对 SOCO 方法的消融实验显示（如图3-9），SOCO 方法也能够对数据性能进行提升，主要表现在 Precision 上，对于 F1-score 指标也能达到与原始数据相当的性能，但是对于 Recall 指标则未体现出性能的增强。但总体上讲，SOCO 模型也能够使增强后数据性能达到与原始数据性能相当的水平。

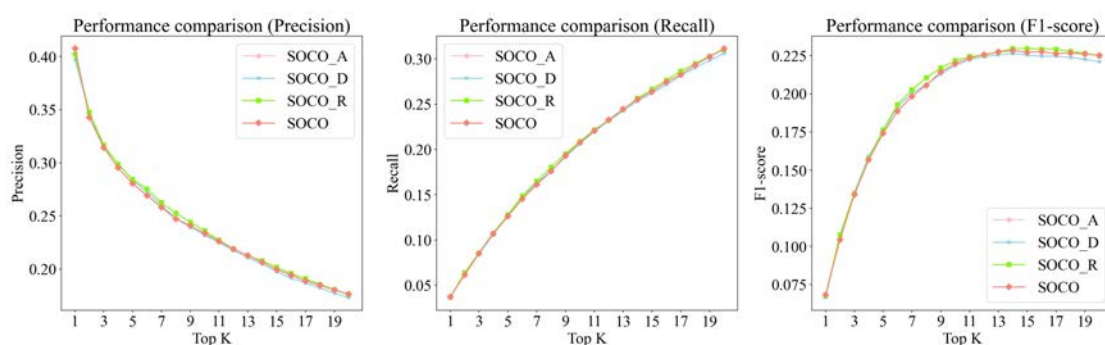


图 3-9 SOCO 模型及其消融实验性能对比

Fig 3-9 Performance comparison between SOCO method and its ablation experiments

3.5 本章小结

在本章的研究中，针对临床诊疗数据存在的“一多一少”现象，提出了基于知识图谱采样的处方数据增强方法 SabKG 模型，以及基于症状本体库与症状共现频度的处方数据增强策略 SOCO 模型。本章将提出的两个模型与四种相关基线方法在多标签处方推荐任务上进行了性能对比。结果显示 SabKG 模型使原始数据得到了相对最好的性能提升，SOCO 模型也能够使增强后数据与原始数据性能相当。提出的两种模型都能够结合领域知识（症状本体库、药症知识图谱、症状共现关系）对原始数据进行症状的增强，特别是 SabKG 模型，其灵活性较好，能够充分地利用现有领域知识中的药症关系，从融入知识的角度对原始数据的性能进行提升。

4 基于症状术语映射的处方推荐方法研究

本章围绕如何提升现有中医处方推荐方法的性能以及如何对症状“未登录词”形成表示的问题进行探讨，提出了基于症状术语映射与深度学习的处方推荐方法（SSTM-based TCM Prescription Recommendation, TCMPR），该方法结合了基于子图抽取的症状术语映射方法（Subnetwork-based symptom term mapping, SSTM）以及深度学习模型来进行中医处方推荐任务，能够有效地表征临床症状术语，特别是形成“未登录词”的嵌入特征。本章将提出的 TCMPR 方法与相关基线方法进行了对比，实验结果显示提出的方法得到了相对最好的性能。此外对 TCMPR 方法中的关键模块进行了相关实验，以更好的探讨本方法的有效性，优化 TCMPR 模型的性能。

4.1 引言

近年来，在临床实践过程中形成的大量的中医临床数据未被充分利用，临床患者的电子病历中通常包含患者主诉、现病史、治疗处方等信息，这类数据中对患者表型信息的记录通常存在较强主观性，如何结合人工智能方法对现有临床电子病历数据进行挖掘，为医生诊疗提供辅助作用是现代信息化临床研究的关键问题之一，领域内许多学者近年来也围绕这一研究热点形成了相关的中医处方推荐方法。

尽管现有处方推荐的相关工作在挖掘和利用中医电子病历方面取得了一定的成效，但整体上看现有处方推荐方法在推荐结果上未能达到一个较好的性能；此外，很多中医临床表型都存在同义关系，一些之前未出现过的术语大部分都可以基于已有表型术语中的字词进行组合得到，例如，术语“脚痛”与术语“脚疼痛”存在同义关系，但两者却被视为两个不同的特征。同时，对于处方推荐中可能出现的“未登录词”，即术语库中未出现的症状词，已有的推荐方法不能有效地解决如何利用现有知识对“未登录词”进行表征这一问题。因此，如何提升现有处方推荐方法的性能，如何利用已有知识形成症状“未登录词”的表示，是两个亟待解决的问题。

4.2 基于症状术语映射与深度学习的处方推荐框架

本节对提出的基于症状术语映射与深度学习的处方推荐方法进行介绍，首先介绍方法整体框架，然后对方法相关细节进行详细阐述。

4.2.1 整体框架介绍

本文提出的基于症状术语映射与深度学习的处方推荐方法框架如图 4-1 所示。方法包括以下部分：首先结合了基于子图抽取的术语映射算法，对患者的所有症状词进行症状术语映射，然后将映射后得到的症状术语集通过已训练的症状网络嵌入表示进行症状特征融合，从而形成原始患者的症状表示，之后通过卷积神经网络学习、全连接网络的训练以及激活层激活后，得到对每种候选中药的预测概率，最后将每味中药的预测概率进行降序排列，并将降序后对应的前 K 味药物作为最终的推荐结果。

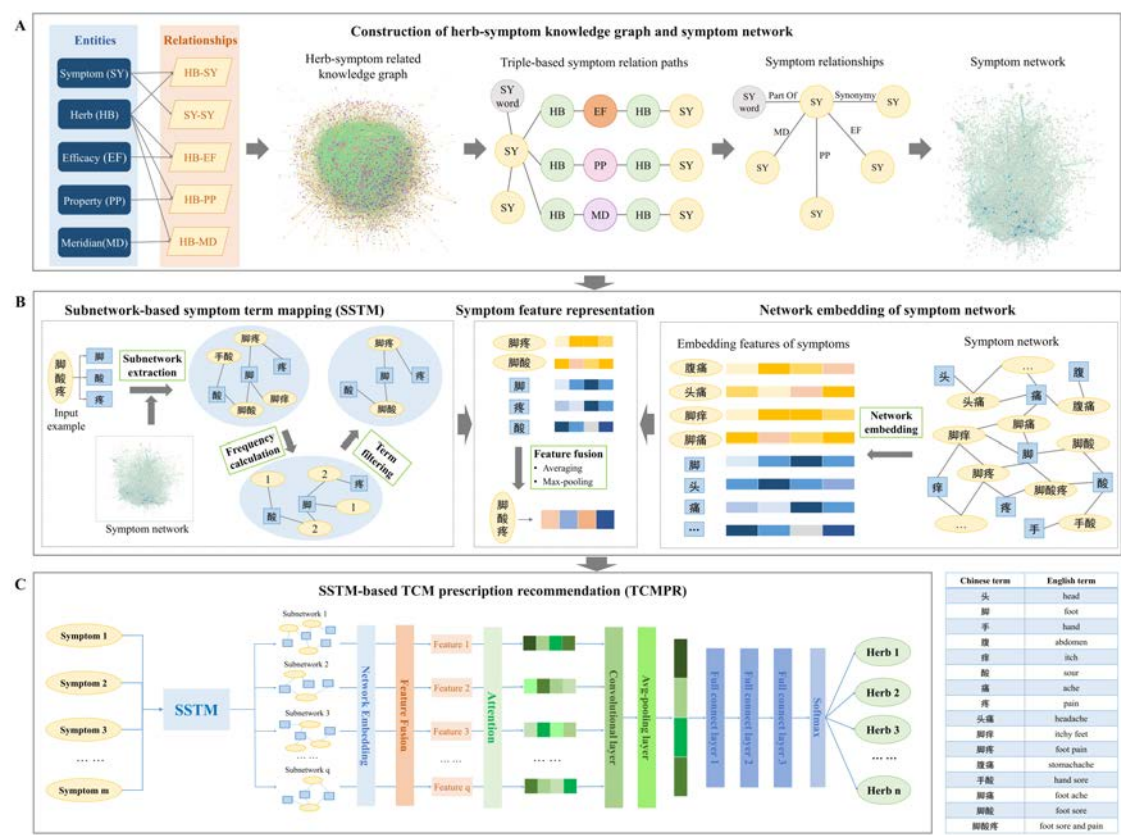


图 4-1 基于症状术语映射与深度学习的处方推荐方法 TCMPR 框架

Fig 4-1 The framework of TCM prescription recommendation method based on subnetwork term mapping and deep learning

在此过程中，涉及到几个关键过程：药症知识图谱构建、症状网络的形成、症状网络嵌入表示、患者症状特征融合以及最终 TCMPR 模型的推荐过程。对于药症知识图谱的构建，本文第3章 3.2.4 节中已对其进行了阐述。下面对症状网络的形成与嵌入表示、症状术语融合以及 TCMPR 模型结构进行详细介绍。

4.2.2 症状网络形成及嵌入表示

在本文 3.2.4 节对构建的药症知识图谱进行了介绍，虽然已构建的药症知识图谱所包含众多实体和关系，其中也覆盖到了许多症状表型实体，但是在实际场景中无法被充分利用，特别是对于描述患者症状间现存或潜在的关系，仅仅靠目前构建的药症知识图谱中症状间的同义关系是远远不够的。本节在已构建的药症知识图谱的基础上，利用中药、中药功效、中药性味和中药归经与症状之间的直接或间接连接关系，构建了三种元路径，并根据这三条元路径寻找症状词间潜在的关联，把形成的症状关联附加到现有的症状关系中，形成更完备的症状网络。下面对构建症状网络的过程进行阐述。

症状术语之间存在着同义关系，中药与症状间也存在着直接关联（如前所述《中华本草》中记录的药症间的关系），并且中药和症状之间还可通过其他相关实体进行间接连接，如两个中药若都能连接相同的中药功效（即两者均具有此功效），那么与这两个中药直接相连的症状词就可通过两个中药和功效形成的实体链而产生连接关系。

基于以上思想，本节构建了三条元路径^[77] 关系，分别是：“症状-中药-中药功效-中药-症状”、“症状-中药-中药性味-中药-症状”和“症状-中药-中药归经-中药-症状”，即分别代表通过中药及其功效、性味和归经属性的直接联系而产生的症状间的间接连接。在此基础上根据已构建的药症知识图谱中的关系进行遍历形成新的症状关系，并将形成的症状关系与现有症状同义关系、症状字词关系进行合并，从而形成了关系相对更完备的症状网络。本节形成症状网络的过程如图4-1A所示，形成的症状网络的规模和各关系的详细信息如表4-1所示。

表 4-1 症状网络构成

Table 4-1 The construction of symptom network

关系来源	关系类型	频度阈值	关系数目
中药-症状	症状-症状字	—	3038
症状-症状	症状-症状（同义关系）	—	3122
症状-症状字	症状-症状字	—	12542
中药-中药功效， 中药-症状	症状-症状（功效）	1000	52729
中药-中药归经， 中药-症状	症状-症状（归经）	1000	51394
中药-中药性味， 中药-症状	症状-症状（性味）	100	45248
关系总数			168073

此处有两点需要进行说明：一是对于症状字词关系的添加，在症状网络的关系中添加“症状词-症状字”关系是为了后面使用基于子图抽取的症状术语映射方法进行症状词的扩散拼接，从而进行子图抽取的过程；二是对于症状网络中新添

加的三种症状关系，基于元路径的方法能够形成数目庞大的关联关系，为了控制症状网络的规模、使新构建的症状关联数目与原始症状关联数目尽可能平衡，本节对新形成的症状关联根据症状关系的频度进行了筛选。

在构建了症状网络后，本节结合了图嵌入相关方法对症状网络中的节点进行表示，以便于后续形成患者特征和下游推荐任务的进行。在本文中，使用到的嵌入方法有 DeepWalk^[70]、node2vec^[71]、Line^[72]、TransE^[78] 和 One-hot^[79]，并在实验中对几种方法的性能进行了对比，嵌入过程如图4-1B 右侧模块所示，实验设计及结果将在后文进行详细介绍。

4.2.3 患者症状映射与特征融合

本节将重点阐述患者症状映射与特征融合的策略。

在处方推荐任务中，一个至为关键的过程便是如何形成对患者的表示（即如何利用输入的患者症状词集合形成患者的特征）。输入的患者症状词集合中，部分症状词是存在于已构建的症状网络中，这些症状词是容易得到其嵌入表示；但是可能存在着现阶段已构建的症状网络中尚未出现的症状术语（即症状“未登录词”），很难使用症状网络直接对这种“未登录词”进行表示。

为解决如何表征症状“未登录词”的问题，本节结合了子图抽取方法对症状“未登录词”进行症状术语映射，方法流程如图4-1 B 左侧模块所示。以术语“脚酸疼”为例做基于子图抽取的症状术语映射：

- (1) 首先对“脚酸疼”进行字的拆解，变为{"脚","酸","疼"}集合；
- (2) 以此集合的每个症状字为出发点，在已构建的症状网络中寻找每个字的一阶邻居，假设找到的一阶邻居有“脚痒”、“手酸”、“脚酸”、“脚疼”4个症状词节点；
- (3) 对找到的邻居节点在该子图中的度进行统计，保留度大于1的节点，最后构建出“酸-脚酸-脚-脚疼-疼”子网络，并将得到的节点集合作为输入症状术语的子图表示。

可以看出，基于子图抽取的术语映射方式能够尽可能地将症状术语进行术语映射，能够充分地利用症状术语词的潜在信息，因为症状术语通常由部位词、症状词、程度词等组成，利用扩散拼接及子图筛选的方式能够将“未登录词”与现有症状术语建立临床语义上的关联，很容易利用已构建的症状网络形成现有症状术语的嵌入表示，因此这种策略不仅能够对症状“未登录词”进行表示，而且形成的映射集合与原始“未登录词”之间在临床含义上相对密切，能够在一定程度上保留住原始症状词的临床含义。

对于输入的患者症状术语集合，无论其是否包含症状“未登录词”，均可用基于子图抽取的症状术语映射方法来对“未登录词”进行表示。在得到映射后的患

者症状术语集合后,即可对该集合中所包含的患者术语所对应的特征进行特征融合,从而形成该患者的最终表示,以便于后面进行模型的训练学习。本章中使用的融合方法包括最大池化和求和平均两种方法,在本章的实验部分中对两种方法的性能进行了对比。

如果症状词 SY_i , 映射后得到的症状术语集合为 $set_i=\{s_1, s_2, ..., s_n\}$, set_i 对应的嵌入表示集合为 $f_i=\{f_1, f_2, ..., f_n\}$ 。则最大池化公式可表示为式4-1,

$$F_{SY_i} = \arg \max_i f_i \quad (4-1)$$

平均池化公式可表示为式4-2。

$$F_{SY_i} = \frac{1}{n} \sum_{i=1}^k f_i \quad (4-2)$$

4.2.4 基于症状术语映射的处方推荐方法 TCMPR

本节介绍本文构建的基于症状术语映射与深度学习的处方推荐方法 TCMPR,其利用患者的所有症状词作为输入,经过症状术语映射、网络嵌入表示、特征融合、卷积神经网络和全连接网络训练等过程后,形成对每味候选中药的预测概率,并以概率排序后的中药序列的前 K 味药物作为最终推荐出的处方。该方法框架的具体流程阐述如下。

输入: 患者的症状词集合,该集合中症状词的数量为 m。

(1) 症状术语映射: 将输入的患者症状词集合通过基于子图抽取的术语映射算法映射成 q 个子图。

(2) 网络嵌入表示: 对症状网络通过网络嵌入表示算法 DeepWalk 进行嵌入表示,得到症状网络中所有症状词和症状字的嵌入表示向量(维度为 d)。DeepWalk 能够对图结构的数据进行节点嵌入表示,并且学习到的节点表示能够保留原始图数据的连接关系,即在图结构上邻近或特征相似的节点训练得到的嵌入表示向量也是近似的。这里的做法是: 对症状网络中的每个节点生成 γ 个随机游走序列,每个游走序列的长度为 t,然后将产生的随机游走序列通过 Word2vec 算法进行向量表示,最终得到每个节点的特征表示,每个表示向量维度均为 d 维。

(3) 症状特征融合: 对于 (1) 得到的 q 个子图,在通过 (2) 得到的症状嵌入表示中寻找其对应症状的嵌入表示,然后进行特征融合,形成对应的 q 个嵌入特征向量,形成的每个向量都是 d 维。此处的特征融合指将每个子图中所有节点的嵌入向量进行求和平均计算,计算后的向量作为特征融合的结果。

(4) 特征权重计算: 将通过 (3) 形成的嵌入向量经过一层注意力层,进行特征重要性计算。本文的做法是,首先对于所有患者症状特征 $\mathbf{X} = \{x_1, x_2, ..., x_q\}$, 根据

注意力打分函数计算注意力分布，然后根据求得的注意力分布来计算输入信息的加权平均，得到对患者特征的重要性评估结果。患者特征的重要性评价向量的计算方式如式4-3所示，

$$att(X, q) = \sum_{i=1}^q \alpha_i x_i \quad (4-3)$$

其中， q 为查询向量（可以是动态生成的，也可以事先指定）， α_i 为注意力分布， α_i 的计算方式如式4-4，

$$\alpha_i = \frac{\exp(s(x_i, q))}{\sum_{j=1}^q \exp(s(x_j, q))} \quad (4-4)$$

在 α_i 中， $s(x, q)$ 为注意力打分函数，其计算方式见下式4-5。

$$s(x, q) = x^T q \quad (4-5)$$

(5) 卷积神经网络学习：将 (4) 学习到的参数输入至包含了 k 个卷积核的一层卷积层中进行训练，再通过一层平均池化层融合，得到患者的最终症状特征。

(6) 全连接神经网络学习：将 (5) 中形成的患者融合特征输入到包含 3 层全连接层的全连接神经网络中，其中第 1 层的神经元个数为 256，第 2 层的神经元个数为 64，第 3 层的神经元个数与待分类类别数目相同（即等于中药总数），进行学习。

(7) Softmax 层激活：将 (6) 中学习到的参数通过一层 Softmax 激活函数^[80]，即将全连接层的输出结果转化为概率，从而得到每种中药被推荐的概率值。Softmax 函数的计算公式如式4-6，

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_{c=1}^n \exp(z_c)} \quad (4-6)$$

其中 z_i 指第 3 层全连接层中第 i 个节点的输出值， n 为该层神经元个数（即中药总数）。

输出：将每味中药的预测概率进行降序排列，并将降序后对应的中药顺序作为输出。

4.3 实验设计与参数设置

本节对本章实验的相关情况进行介绍，包括实验数据、整体实验设计、评价指标及参数设置。

4.3.1 实验数据与实验设计

本节实验所使用的数据是本文第三章中阐述的 8218 条临床医案数据。训练测试集采用随机划分，训练集：测试集 = 8：2，即训练样本为 6574 条，测试样本为 1644 条。

由于提出的方法是对每味候选药物形成预测概率,形成的推荐结果是多味中药的组合(即多标签分类问题),因此将本文提出的处方推荐算法 TCMPR 与两种多标签分类的基线方法进行了对比,其中基线方法包括 MLKNN^[14] 和 MLDT^[15]。MLKNN 算法借鉴了 KNN 算法的思想,通过寻找 k 个近邻样本,并运用贝叶斯条件概率来计算当前标签是否采纳的概率(即判断标签为 1 或 0),概率大的标签则定为样本的预测类别,本节对近邻数 k 分别取 1、5 和 10 分别进行了 3 次实验。MLDT 算法利用了决策树算法对多标签数据进行分类,其主要做法是利用基于多标签熵的信息增益准则递归地构建决策树,在实验中设置叶子结点的最小大小为 40。

此外,实验部分对提出的 TCMPR 方法中的关键模块(子图筛选阈值及登录词处理、嵌入表示方法、患者特征填充策略及患者特征融合方式)和其他超参数(症状嵌入维度、患者症状筛选阈值及全连接层隐藏单元个数)也进行了性能对比实验。

4.3.2 评价指标与参数设置

需要说明的是,为控制变量,以上实验都是基于以下环境和参数下进行的:所有模型基于 Tensorflow2 框架进行实现,神经网络所使用的学习率设置为 $1e-4$,并使用 Adam 梯度下降法进行反向传播,epoch 的默认值为 200,但在训练的过程中若损失 loss 和测试集的命中率在几轮 epoch 后保持不变,则提前停止训练。

在上述实验中:子图筛选阈值分别为 1、2 和 3 分别进行实验,并且同时考虑是否对登录词进行症状映射进行考虑(登录词即存在于症状术语库中的表型术语);对于嵌入表示方法,本章选取了 DeepWalk、node2vec、Line、TransE 和 One-hot 分别进行实验;对于患者特征填充策略,分别采用了最大填充、平均填充和全零填充三种方式进行实验;对于特征融合方式,分别以求和平均以及最大池化两种方式进行。此外对于特征维度分别设置为 100、200、300、400 和 500,将患者症状词最大个数阈值分别设置为 10、20、30、40 和 50,另外对于全连接层隐藏单元个数也进行了几种不同的设置。

由于任务为多标签分类任务,因此仍采用 Top@K 系列评价指标(见 3.4.2 节)。

4.4 实验结果及分析

本节将对各实验及结果进行介绍,包括模型与基线方法对比实验及结果、模型关键模块相关实验及其他超参数相关实验。

4.4.1 模型与基线方法对比结果及分析

本节对提出的 TCMPR 方法与 MLKNN, MLDT 两种多标签分类基线方法进行了对比,其中对于 MLKNN 算法中近邻数 k 分别使用 1、5 和 10 进行三次实验。在此实验中,特征融合方式均为求和平均,特征嵌入表示的特征维度均为 200,子图筛选阈值为默认值 1,特征嵌入方法选择 DeepWalk。在此条件下比较 TCMPR, MLKNN 与 MLDT 三种方法的性能,结果如下图4-2所示, Top@5、Top@10 和 Top@15 的具体结果见表4-2。

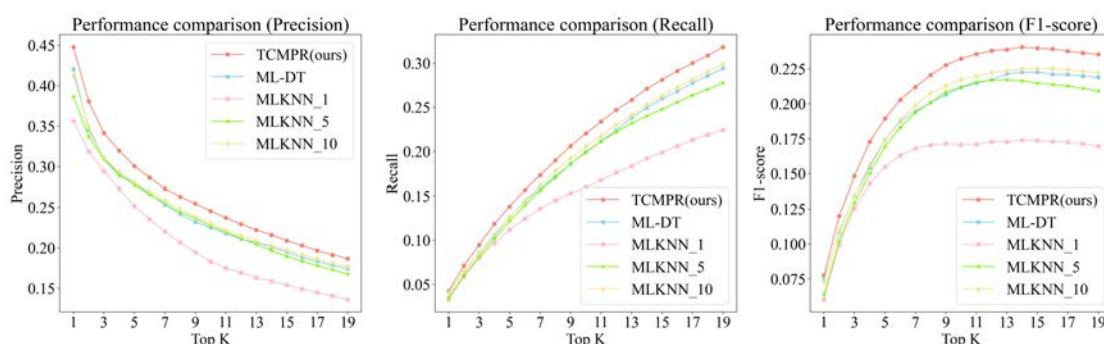


图 4-2 TCMPR 与基线方法性能对比

Fig 4-2 Performance comparison between TCMPR and baselines

从整体上看,本文提出的 TCMPR 比 MLKNN, MLDT 两种基线方法无论是在准确率,召回率和 F1-score 上都有着最好的性能,其中 TCMPR 与最好基线方法 MLKNN-10 相比,在 Top@10 下准确率提升了 9.51%,召回率提升了 7.09%,F1-score 提升了 6.88%。同时,随着 K 值的增加,考虑的药物数量增加了,所有方法的准确率呈下降趋势,召回率呈上升趋势,而 F1-score 呈现上升趋势,说明随着 K 的增加,预测结果是越来越好的。对于 F1-score,从 $K=12$ 开始,所有方法的结果基本呈稳定趋势,并且随着 K 的继续增加,多数算法出现了 F1-score 小幅度下降的现象,这种现象可能是数据本身造成的,如前文所述实验数据的平均药物数目为 11.31 个,因此不难理解此现象产生的原因。

对比 MLKNN 与 MLDT 的结果可以看出,MLDT 算法比 MLKNN-1 性能略好,与 MLKNN-5 性能相当,而 MLKNN-10 比 MLDT 的结果略胜一筹。对于 MLKNN 算法,MLKNN-10 比 MLKNN-5 略胜一筹,而 MLKNN-1 则比 MLKNN-5 和 MLKNN-10 的结果差。从这里可以看出,MLKNN 算法中近邻数 k 的选取对预测结果具有一定的影响。

表 4-2 TCMPR 与基线方法性能对比结果

Table 4-2 Experimental performance between TCMPR and baselines

Top@K	P@5	P@10	P@15	R@5	R@10	R@15	F1@5	F1@10	F1@15
MLKNN-1	0.2519	0.1118	0.1549	0.1832	0.1596	0.1706	0.1541	0.1990	0.1737
MLKNN-5	0.2773	0.1217	0.1692	0.2272	0.1987	0.2120	0.1897	0.2477	0.2148
MLKNN-10	0.2805	0.1262	0.1741	0.2304	0.2058	0.2174	0.1970	0.2630	0.2252
ML-DT	0.2799	0.1264	0.1742	0.2251	0.1996	0.2116	0.1949	0.2597	0.2227
TCMPR(ours)	0.3006	0.1382	0.1894	0.2457	0.2204	0.2324	0.2090	0.2811	0.2397
Improvement	7.17%	9.51%	8.79%	6.64%	7.09%	6.90%	6.09%	6.88%	6.44%

4.4.2 TCMPR 方法对未登录词激活效果分析

为探究 TCMPR 方法对症状“未登录词”的特征形成效果，本节对临床医案数据中“未登录词”的激活情况进行了统计（若“未登录词”通过 TCMPR 方法进行映射后能够形成症状集合及表示，则称该“未登录词”得到了激活），相关统计结果如表 4-3 所示。

表 4-3 临床医案数据中“未登录词”信息及激活情况统计

Table 4-3 Basic statistics and activation effect of unrecorded terms in clinical case data

指标名称	训练集	测试集
样本数	6574	1644
症状平均数	11.41	11.63
登录词平均数	2.81	2.79
登录率	23.76%	23.36%
零登录率样本数	1192	299
映射后症状平均数	44.24	45.53
样本激活率	100.00%	100.00%

其中，“样本数”表示该训练/测试数据包含的样本数目，“症状平均数”表示所有样本包含症状词个数的均值，“登录词平均数”表示所有样本的症状集合中登录词个数的均值，“登录率”表示所有样本的症状集合中登录词所占比率的均值，“零登录率样本数”指包含的所有症状词均为“未登录词”的样本数目，“映射后症状平均数”指原始样本中的症状集合经 TCMPR 方法映射后形成的新症状集合中包含的症状数目的均值，“样本激活率”指得到激活的零登录率样本与原始数据中所有零登录率样本的数目之比。

从结果可以看出，数据的登录率较低（平均登录率不到 24%），这表明原始数据中包含了较多的“未登录词”，而且数据集中包含的零登录率样本的不占少数（零登录率已超过 18%），而 TCMPR 模型中基于子图抽取的症状术语映射方法的使用，能够使临床医案数据中的所有样本形成融合症状网络后的映射集合及嵌入

表示，特别是对零登录率样本的效果更加明显。

4.4.3 模型关键模块对比实验结果及分析

本节对 TCMPR 模型中关键模块进行的相关实验与探讨进行介绍，包括子图筛选阈值及登录词处理策略、嵌入表示方法、患者特征填充策略及患者特征融合方式。

(1) 子图筛选阈值与登录词处理策略

基于子图抽取的症状术语映射是 TCMPR 模型中的核心模块，映射后得到的症状子图（集合）能否反映原始症状“未登录词”的临床含义直接决定着模型的泛化性，其中对于子图的筛选是关键一环。因此本节对子图筛选阈值进行了实验，分别设置子图筛选阈值为 1、2 和 3 进行实验。

此外，对于登录词，虽然其包含于已构建的症状网络中，但是否需要对之进行症状术语映射，这也是需要探讨的：对于症状术语映射而言，其好处在于能够借助症状网络中的知识结构来更好地形成对原始症状词的表示，但同时也存在一个问题，即对原始症状词的症状映射后可能会引入噪声（与原始症状词在临床含义上不太相关的症状字和症状词）。因此本节对是否需要登录词进行术语映射也进行了探讨，分别以对登录词映射和不映射进行实验。

根据上述问题，本节进行了 6 个相关实验。为控制变量，在各实验下嵌入表示方法均为 DeepWalk、特征融合方式均为求和平均、特征嵌入表示的特征维度均为 200。实验结果如图 4-3 所示（其中标签为“All”表示对所有患者原始症状术语均进行映射，标签“UnRecord”则表示不对登录词进行映射）。

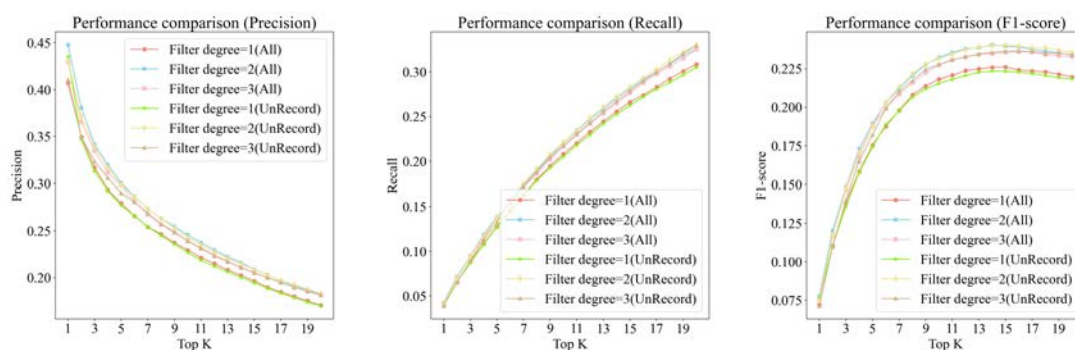


图 4-3 不同子图筛选阈值与“未登录词”处理方式下性能对比

Fig 4-3 Performance comparison between different subnetwork filtering threshold and unrecorded term processing ways

从结果中可以看出，子图筛选阈值是影响模型性能的一个主要因素。在筛选阈值为 2 时相比阈值为 1 的性能得到了较高提升，其原因在于阈值为 1 时形成的

映射后的症状集合中包含的症状数目较多（平均症状术语数目已超过 60 个），其中存在着一些噪声症状词，因此对最后的推荐性能产生了一定影响，而经阈值为 2 的筛选后，保留的症状词则与原始症状术语在临床含义上的关联更为密切，因此能够带来性能上的大幅度提升。在阈值为 3 时虽然与阈值为 1 相比性能仍有提升，但是相比阈值为 2 时则逊色一些，原因在于当阈值为 3 时筛选条件过高，使得保留下来的症状术语多为症状字，并且筛选后症状集合所包含的症状数目骤减（甚至少于患者的原始症状数目），因此性能略逊色于阈值为 2 时的结果。

此外，对登录词是否进行映射这一问题不是影响模型性能的主要因素，但这也可能对模型性能产生一定影响。如图4-3所示，当子图筛选阈值为 1 时，对登录词进行映射要略好于不对登录词进行处理，而在子图筛选阈值为 2 和 3 时，不对登录词进行映射略好于对登录词进行处理。这种现象的出现原因同样在于子图筛选的控制：当子图阈值为 1 时，由于形成的映射后术语集合包含了大量的症状词，同时也引入了大量的噪声，此时对登录词进行处理，确实能够给模型引入知识从而使模型的性能得到一定提升，但是无论是否对登录词进行处理，在此阈值下（阈值为 1）形成的映射后症状集合显然引入了一些噪声；而对于子图阈值为 2 和 3 时，映射后形成的症状集合已经得到了筛选，由于映射带来噪声而引起的影响被减轻了，此时若对登录词再进行映射，对于已形成的映射集合在一定程度上会引入新的噪声，因此这时不宜再对登录词进行映射。综上所述，对登录词不进行处理是相对更优的策略。

(2) 患者症状嵌入表示方法

对于患者嵌入表示策略，本节分别利用 DeepWalk、node2vec、Line、TransE 和 One-hot 五种嵌入表示方法进行实验对比。在此实验中，特征融合方式均为求和平均，嵌入表示的特征维度均为 200。在此条件下比较几种嵌入表示方法的性能。实验结果如图4-4所示。

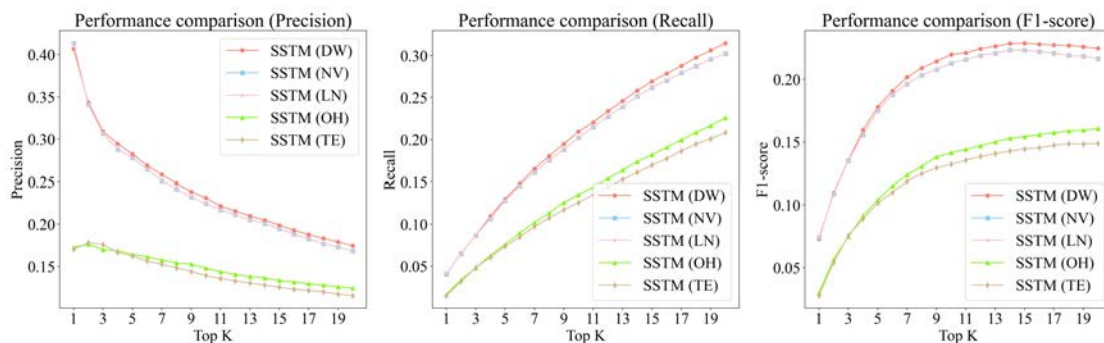


图 4-4 不同症状嵌入方式下性能对比

Fig 4-4 Performance comparison between different embedding methods

图中结果表明,在相同的条件下,DeepWalk、node2vec 和 Line 三种方法在性能上明显优于 TransE 和 One-hot,并且 DeepWalk 展示出了相对最优的性能。DeepWalk 具有 Word2vec 方法所具备的上下文语义的连接,node2vec 方法也可以保留节点的全局和局部特征,Line 方法可以很好的保留一阶和二阶信息,所以这些方法的性能相对较优。TransE 不具备关系特征的支持,这种方法所组合的词的融合可能会有较多的噪声,因此容易产生融合偏倚,而 One-hot 缺乏语义相关性,所以在性能上不占优势。

(3) 患者症状填充策略

由于不同患者症状数目存在着差别,而且不同症状组经映射后形成的症状集合所包含的症状数目更是相差甚异,为便于后续模型训练,需对患者症状特征设定一阈值进行筛选,并且对于未达到该阈值的患者特征需进行特征填充。

本节对患者症状填充策略进行了实验,分别以平均填充、最大填充和全零填充进行实验。此处需说明的是,平均填充指利用现有患者特征向量对各维度取平均值形成的向量来进行缺失填充,最大填充指将患者现有特征向量各维度取最大值形成的向量进行填充,全零填充指以维度相同但特征数值均为零的向量进行缺失填充。结果如图4-5所示。

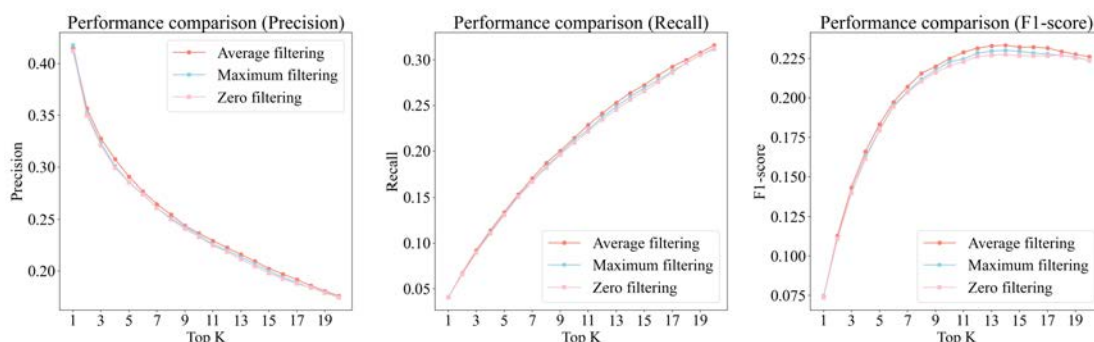


图 4-5 不同患者症状特征填充方式下性能对比

Fig 4-5 Performance comparison between different symptom term filtering ways

从结果中可以看出,平均填充法取得了相对最优的效果,其原因在于平均法对患者已有特征产生的影响相对较小;最大填充的性能略逊于平均填充,原因在于这种“最大”的方式可能会引入偏差;而全零填充则表现出了相对最差的性能,因其对已有症状特征的影响最大。

(4) 患者特征融合方式

对于患者特征的融合方式,本节对求和平均法与最大池化法两种方式进行比较。在此实验中,嵌入表示方法均为 DeepWalk,特征维度均为 200。在此条件下比较两种特征融合方式的性能。结果如图4-6所示。从结果中可以看到,求和平均的

方式表现均优于最大池化的方式。求和的方式对所有的特征部分都是相对公平的，而池化的方式可能导致信息的丢失。可以看出，融合方式的选择对特征的构建以及后续模型的性能能够产生一定影响。

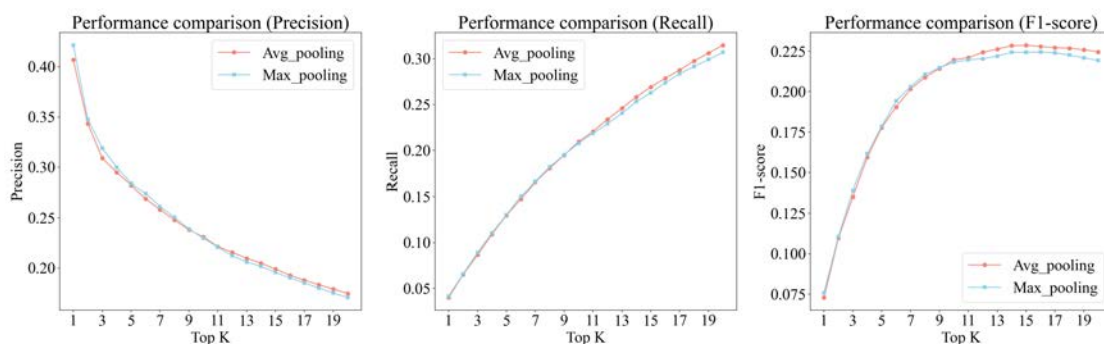


图 4-6 不同症状嵌入融合方式下性能对比

Fig 4-6 Performance comparison under different fusion ways

4.4.4 模型其他超参数实验及结果

本节对模型相关其他超参数进行的相关实验进行介绍，包括症状嵌入维度、患者症状筛选阈值及全连接层隐藏单元个数。

(1) 患者症状特征维度

首先本节对患者症状特征的维度进行了探究，分别设置特征维度为 100、200、300、400 和 500 进行实验。在此实验中，嵌入方式均采用 DeepWalk，特征融合方式均采用求和平均。在此条件下进行实验。结果如图4-7所示。

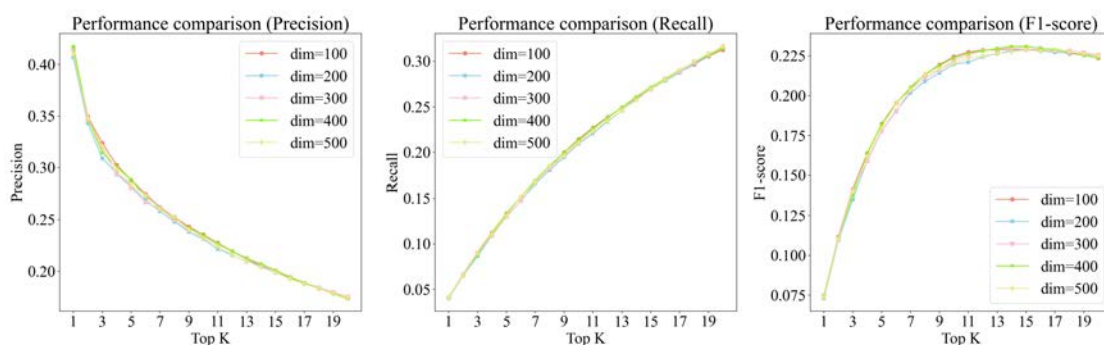


图 4-7 不同症状嵌入维度下性能对比

Fig 4-7 Performance comparison between different embedding dimension

从图中可以看到，当维度从 100 增加至 500 时，模型的性能产生的变化并不明显，而特征维度的增长带来了训练时间的增加。在这五个实验中，当维度在 200

时,模型训练的时间相对较快(平均每轮训练需要约9秒),但是在其他情形下训练时间则有所增长(平均每轮训练时间已超过11秒)。综上所述,症状表示维度为200维是相对较好的选择。

(2) 患者症状数目最大阈值

由于不同患者所带症状数目各异,并且不同症状经子图映射后形成的症状集合所包含的症状数目更是千差万别,因此需设置阈值对患者症状数目进行控制,这一点在上一节中已经提到。因此本节针对患者症状数目最大阈值进行了实验,分别将该阈值设置为10、20、30、40和50进行实验。在此实验中,子图筛选阈值设为1,症状嵌入表示方法为DeepWalk,特征融合方式为求和平均。实验结果如图4-8所示。从图中可以看出,当患者症状数目最大阈值在40时能够取得相对较好的性能。

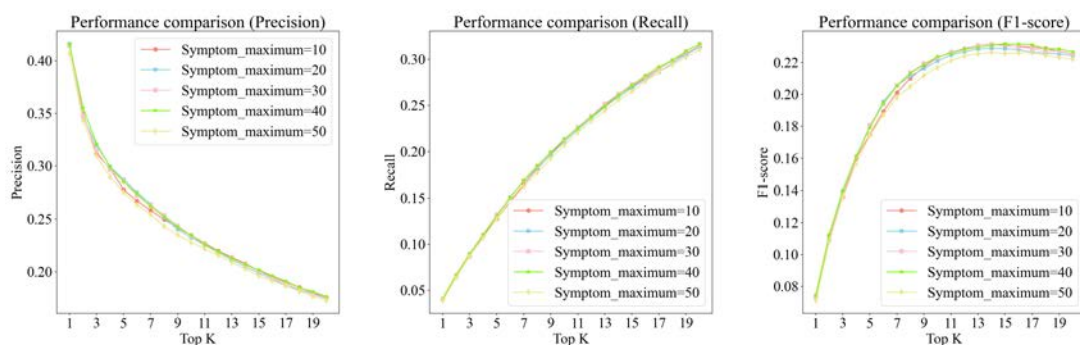


图 4-8 不同症状特征筛选阈值下性能对比

Fig 4-8 Performance comparison under the maximum threshold of symptom terms

(3) 全连接层神经元个数

最后,本节对模型中全连接层的神经元个数进行了相关实验。由于第三层全连接层的神经元个数与中药总数相同,因此只对第一、二层全连接层的神经元个数进行调整。设置方案及实验结果如图4-9所示。从结果可以看出,当第一层神经元个数为256,第二层神经元个数为64时取得了相对最优的结果,其他方案的神经元个数均高于此方案,但在性能上却不如此方案,原因可能是使用的医案数据数量相对较少。

4.5 本章小结

在本章的研究中,针对目前中医处方推荐方法的性能不高、以及现有方法对症状“未登录词”无法形成较好的表示这两个问题,提出了结合症状术语映射与深度学习的中医处方推荐方法TCMPR。本章对提出的TCMPR模型与基线方法进

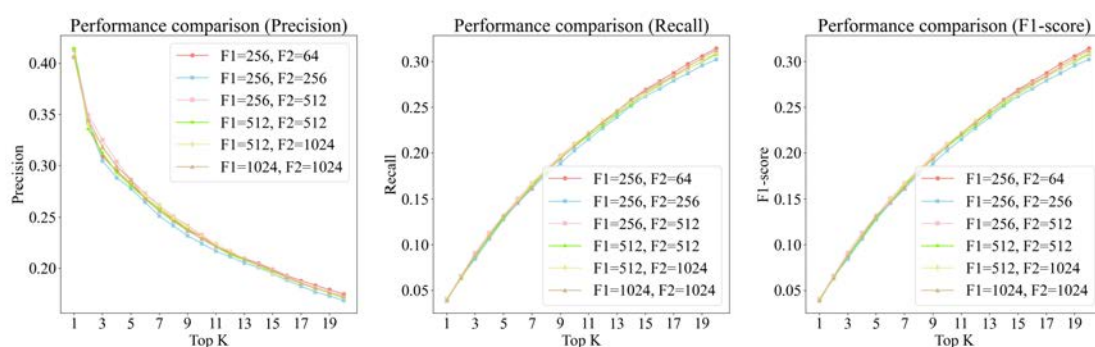


图 4-9 不同网络结构下性能对比

Fig 4-9 Performance comparison between different network structure

行了对比，并针对模型的关键模块和相关超参数进行了相关实验和分析，从结果可以看出 **TCMPR** 方法在性能上优于现有基线方法，并且通过实验发现，子图筛选对于 **TCMPR** 模型而言是主要的影响因素，症状特征嵌入方法、特征填充方式以及特征融合方式都对模型性能产生着一定影响。本章提出的方法融合了领域知识（药症知识图谱、症状网络、症状本体库等），能够更好地对患者“未登录词”进行表示，并且在推荐性能上相比基线方法得到了一定提升。

5 基于表型相似性的处方推荐方法研究

本章围绕如何提升现有处方推荐方法推荐结果的配伍合理性进行开展。首先将待推荐集合由中药变为处方,即将多标签分类问题转换为多分类问题,结合了中医经典名方数据和传统机器学习相关算法进行了处方预测,同时针对经典名方数据存在的处方对应样本数“长尾分布”^[75]的特点,本章结合了第三章提出的 SabKG 模型对经典名方数据进行了数据增强,而后进行多分类任务,结果显示了 SabKG 方法的有效性。但这种策略受限于经典名方样本,因此本章提出了结合表型相似性与经典名方的处方推荐方法,并提出了几种患者症状特征的构建策略,实验结果显示出本章所提出策略的有效性。

5.1 引言

在本文第三章和第四章中,提出了基于数据增强的处方推荐方法和基于症状术语映射与深度学习的处方推荐方法,虽然这些方法解决了现存的相关问题,在性能上达到了相对最优的效果,但是这些方法推荐的结果仅为中药组合,即对每味中药进行单独考虑,分别得到每味药物的预测概率,并未考虑真实世界处方的开方配伍原则,可能会得到配伍不合理的药物组合。如何提升处方推荐结果的配伍合理性是目前中医处方推荐领域内需要亟待解决的问题。

现阶段领域内相关方法中,多数方法将处方推荐视为多标签分类问题,形成的推荐结果均为中药组合(即形成对每味候选药物的概率,然后以排序后的中药列表作为推荐结果,评价指标多为 Top@K 系列指标),这种策略虽然能使问题得以简化,但是与临床实际开方配伍的策略并不贴切,其推荐结果无法直接使用于真实世界的临床患者。较为理想的目标是利用计算机相关方法模拟医生开具真实可靠的处方和处方方解,但是由于中药配伍的复杂性以及传统医学各流派诊疗策略的异质性,因此这也成为利用计算机技术实现这一目标的过程中的瓶颈和挑战。

5.2 基于经典名方的处方推荐方法及实验结果

针对中医处方推荐结果的配伍合理性的问题,本节形成了基于经典名方的处方推荐策略,其基本思想是将处方推荐问题由多标签分类问题转化为多分类问题,利用经典名方数据进行学习和训练,推荐结果转为经典名方(即从中药组合到真实处方的过程)。本节结合了机器学习领域中几种经典的多分类方法,利用筛选后的经典名方数据进行推荐,并针对经典名方具有的“一多一少”的特点,结合了第

三章中提出的 SabKG 模型进行优化。本节围绕这一工作及其实验结果进行详细阐述。

5.2.1 基于经典名方的处方推荐策略

本节首先结合了经典名方数据，以经典处方作为推荐结果实现处方推荐，即以处方名称为推荐对象的多分类任务代替以中药为推荐对象的多标签分类任务，方法流程如图5-1所示。

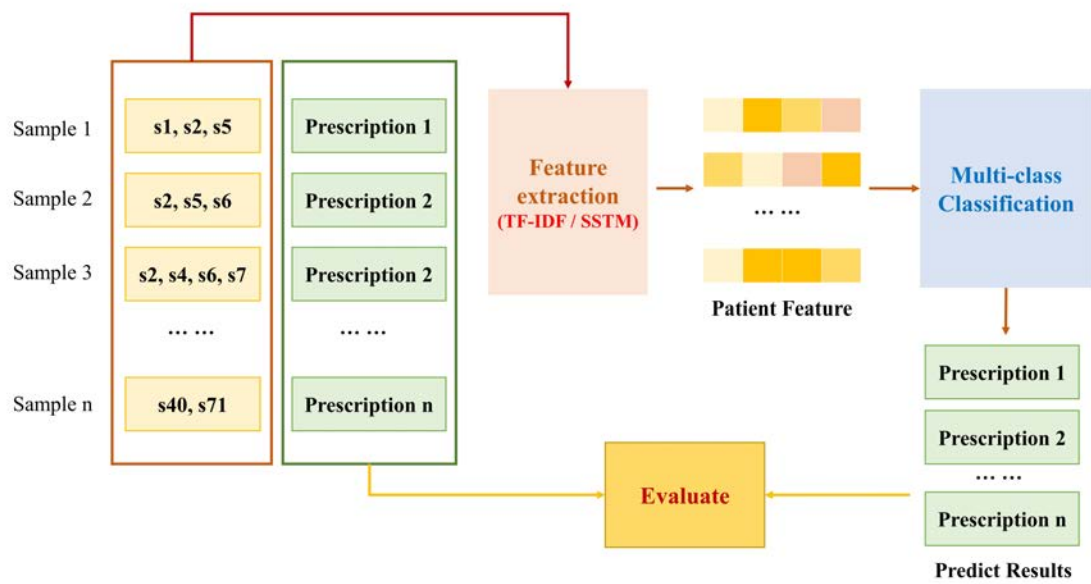


图 5-1 基于经典名方的处方推荐策略

Fig 5-1 Prescription recommendation strategy based on classical prescription data

该框架的具体流程阐述如下。首先，对现有经典名方数据进行预处理，确保数据中经典名方及其适应症的正确关联；使用每条经典名方样本所包含的症状组进行学习，分别形成基于 TF-IDF 的症状特征和基于 SSTM 的症状特征，并形成对原始样本症状组的表示；而后通过多分类算法进行训练，得到预测处方结果；最后对推荐得到的处方及相应样本的原始处方进行对比，评估预测效果。下面对经典名方数据的处理、症状特征的形成、使用的多分类方法及参数设置、实验结果及分析分别进行阐述。

5.2.2 经典名方数据处理

本章所使用的经典名方数据源自中医药经典教材，囊括了《中医内科学》、《中医外科学》、《中医妇科学》、《中医儿科学》等教材中记录的经典名方及相关应用

案例。首先将各教材中所记录的经典名方及其案例进行提取，包括方剂名称、药物组成、症状、疾病名称、证型、治则治法、兼症等信息，整合后的原始案例数据共 2815 条。因原始案例数据中存在症状不规范、药物列缺失等现象，本节对原始案例数据进行了手工预处理，包括对症状进行规范、去除缺失药物组成的数据、合并症状及兼症、去重等步骤，最终保留了 2723 条经典名方数据，其中包含经典方剂 728 种。处理后形成的 2723 条经典名方数据示例如表5-1所示。

表 5-1 经典名方数据示例

Table 5-1 Examples of classical prescription data

序号	症状	方剂名称	药物组成
1	皮肤干燥, 阴伤阳浮, 乏力, 混浊如脂膏, 尿甜, 头晕耳鸣, 腰膝酸软, 瘙痒, 舌红苔少, 尿频量多, 脉细, 口干唇燥, 水竭火烈泄泻腹痛, 小便短黄, 舌质红, 泻下急迫,	六味地黄丸	熟地黄、山萸肉、山药、茯苓、牡丹皮、泽泻
2	肛门灼热, 粪色黄褐臭秽, 泻而不爽, 烦热口渴, 脉滑数或濡数, 发热, 头痛, 脉浮, 苔黄腻	葛根芩连汤	葛根、炙甘草、黄芩、黄连

此外，对整理后经典名方数据的各方剂名对应案例个数的分布进行了统计，分布如图5-2所示，其中图5-2A 展示了所有方剂包含样本数的情况，5-2B 展示了排在前 40 名的方剂对应样本数分布的情况。可以观察到，整理后的经典名方数据存在着“长尾分布”的现象，即少数方剂对应的案例数据个数过多，同时大多数方剂所对应的案例数据数目过少（即本文第三章中提到的“一多一少”问题）。

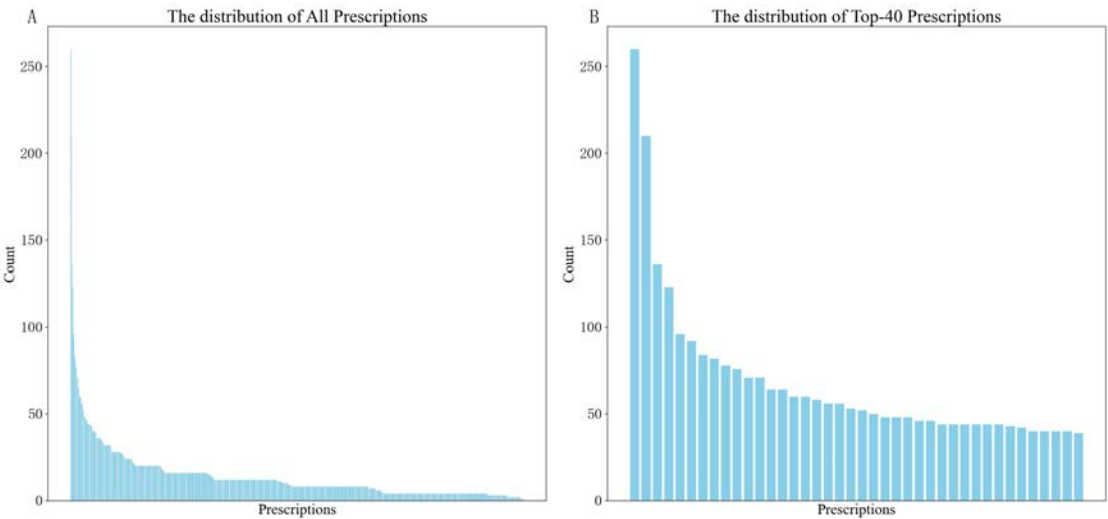


图 5-2 经典名方数据中处方对应样本数分布

Fig 5-2 Sample distribution of prescription corresponding to classical prescription data

由于将方剂名称作为候选类别，因此要保证在各方剂下要具有一定数量的样

本，以便分类任务的进行。因此在 2723 条经典名方数据的基础上，保留了方剂对应样本数目在 5 条以上的相关样本，用于后续的处方推荐。在筛选后保留了 1468 条数据，涉及 158 个经典名方。各筛选阈值下经典名方数据的相关信息如表5-2所示。

表 5-2 各筛选阈值下经典名方数据数目统计

Table 5-2 Statistic result of classical prescription data under different screening threshold

筛选阈值	样本数目	方剂数目	方剂平均样本数
1	2723	728	3.74
2	2519	524	4.81
3	2247	388	5.79
4	1824	247	7.38
5	1468	158	9.29

对于症状特征的形成，本节基于 TF-IDF^[81] 和 SSTM 两种方法进行症状特征的构建：使用 TF-IDF 方法的过程是将每条样本视为一个文档，通过计算该文档中症状术语集合的词频和逆文档频率，并综合得到对每个症状词汇的词向量表示，而后形成对原始样本的症状嵌入表示；使用 SSTM 方法的过程是将原始样本中的症状术语先进行术语映射，将映射后形成的症状集合通过已构建的症状网络的嵌入进行融合，形成对每条原始样本的症状表示。本节利用了这两种策略进行症状特征的形成，并对两种方法下的实验性能进行了对比。

5.2.3 实验设计与评价指标

本节利用筛选后的 1468 条经典名方数据中的症状及方剂名称进行后续的推荐任务，现将实验设计、数据集划分、参数设置、评价指标等内容进行阐述。

本文使用留出法对上述 1468 条经典名方数据进行数据集的划分，训练集：测试集 = 8: 2，即训练数据 1174 条，测试数据 294 条。除留出法外，本文还使用了五折交叉验证划分方法，对数据进行了划分，而后进行了交叉验证。

对于多分类处方推荐任务，本节选用了以下几种机器学习领域中经典的方法进行实验：线性回归 LR^[82]、随机森林 RF^[83]、基于 One-vs-Rest 思想的线性支持向量机 (Linear Support Vector Machine, L-SVM)^[57,58]、高斯朴素贝叶斯 (Gaussian Naive Bayes, G-NB)^[84] 和多层感知机 (Multi-layer Perceptron, MLP)^[85]。参数设置如下：对于随机森林算法，设置其最大深度为 10，叶子节点所包含最小样本数为 3；对于多层感知机，设置其层数为两层，每层的神经元个数都为 100。

本节的评价指标采用多分类任务中的相关指标，各指标的定义如下。

(1) 宏查准率 *macro-Precision* 和加权查准率 *weighted-Precision*。宏查准率公式如 5-1 所示,

$$\text{macro-Precision} = \frac{1}{n} \sum_{i=1}^N \text{Precision}_i \quad (5-1)$$

其中, Precision_i 表示第 i 类下的查准率, 其公式如 5-2 所示,

$$\text{Precision}_i = \frac{P(i) \cap T(i)}{P(i)} \quad (5-2)$$

加权查准率是在宏查准率的基础上, 对于各类别的查准率分别乘其对应权重, 其中权重为各类别的样本数与总样本数的比值。

(2) 宏查全率 *macro-Recall* 和加权查全率 *weighted-Recall*。宏查全率公式如 5-3 所示,

$$\text{macro-Recall} = \frac{1}{n} \sum_{i=1}^N \text{Recall}_i \quad (5-3)$$

其中, R_i 表示第 i 类下的查全率, 其公式如 5-4 所示,

$$\text{Recall}_i = \frac{R(i) \cap T(i)}{R(i)} \quad (5-4)$$

加权查全率是在宏查准率的基础上, 对于各类别的查全率分别乘其对应权重, 其中权重为各类别的样本数与总样本数的比值。

(3) 宏 F_1 值 *macro- F_1* 和加权 F_1 值 *weighted- F_1* 。宏 F_1 值公式如 5-5 所示。

$$\text{macro-}F_1 = \frac{2 * \text{macro-Precision} * \text{macro-Recall}}{\text{macro-Precision} + \text{macro-Recall}} \quad (5-5)$$

加权 F_1 值是在宏 F_1 值的基础上, 对于各类别的 F_1 值分别乘其对应权重, 其中权重为各类别的样本数与总样本数的比值。

(4) 精度 *Accuracy*, 公式如 5-6 所示。

$$\text{Accuracy} = 1 - \frac{a}{m} \quad (5-6)$$

表示在 m 个样本中有 a 个样本分类错误。

5.2.4 实验结果及分析

根据上节中的实验相关设置, 本节利用基于留出法划分的经典名方训练测试数据进行了处方推荐任务, 同时对比了 TF-IDF 和 SSTM 两种特征构建策略下的性能。各方法下测试集的结果见表 5-3 和图 5-3 所示。

其中, 图 5-3 A、B 展示了留出法数据划分下基于 TF-IDF 症状特征的实验结果, 图 5-3 C 展示了五折交叉验证下基于 TF-IDF 症状特征的准确率结果; 图 5-3

表 5-3 留出法下各方法的经典名方数据的实验性能

Table 5-3 Experimental performance of classical prescription data with different classification methods under the leave-out partition strategy

Method	Macro			Weighted			Accuracy
	P	R	F1	P	R	F1	
RF_{tfidf}	0.4162	0.3606	0.3709	0.4231	0.3810	0.3550	0.3810
RF_{sstm}	0.6121	0.5382	0.5494	0.6352	0.5646	0.5469	0.5646
LR_{tfidf}	0.6577	0.5886	0.6053	0.6934	0.6531	0.6321	0.6531
LR_{sstm}	0.8791	0.8843	0.8692	0.8698	0.8673	0.8528	0.8673
NB_{tfidf}	0.9299	0.9409	0.9301	0.9324	0.9422	0.9314	0.9422
NB_{sstm}	0.7284	0.6593	0.6714	0.7133	0.6803	0.6551	0.6803
MLP_{tfidf}	0.9463	0.9378	0.9381	0.9314	0.9388	0.9280	0.9388
MLP_{sstm}	0.8254	0.8215	0.8055	0.8087	0.7993	0.7813	0.7993
SVM_{tfidf}	0.9358	0.9441	0.9342	0.9403	0.9490	0.9373	0.9490
SVM_{sstm}	0.8844	0.8938	0.8764	0.8773	0.8707	0.8570	0.8707

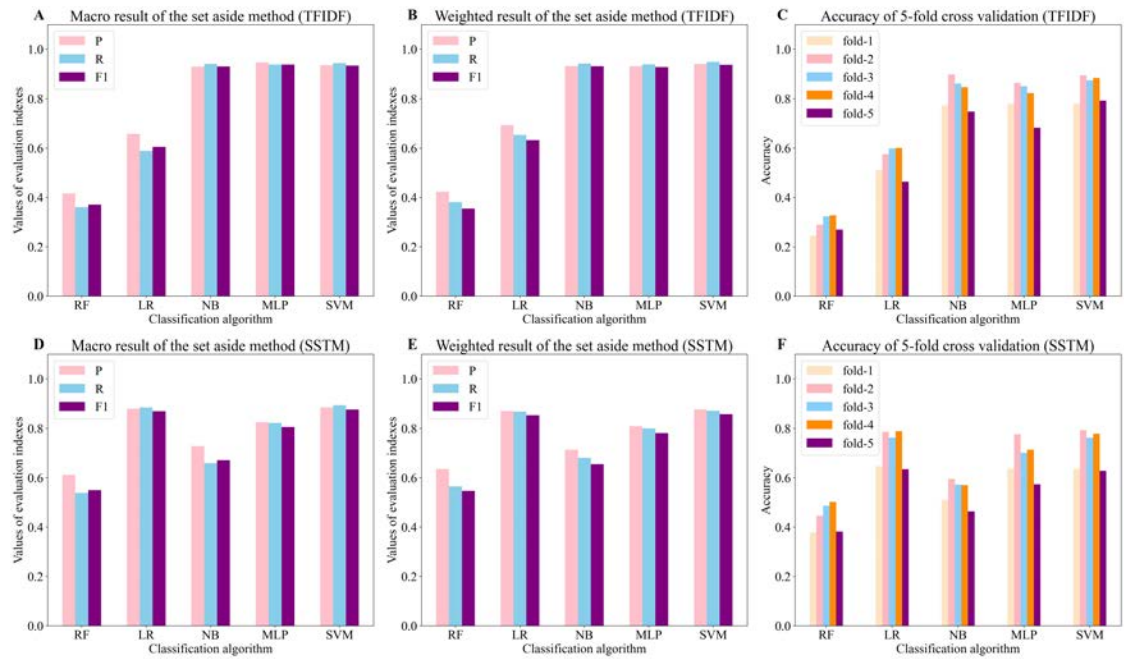


图 5-3 留出法和五折交叉验证下各分类方法的经典名方数据实验性能

Fig 5-3 Experimental performance of classical prescription data with different classification methods under leave-out partition strategy and 5-fold cross validation strategy

D、E 展示了留出法数据划分下基于 SSTM 症状特征的实验结果，图 5-3 F 展示了五折交叉验证下基于 SSTM 症状特征的准确率结果。

从留出法和交叉验证下各分类器的实验性能中可以看出，基于 TF-IDF 特征的相关实验中，高斯朴素贝叶斯、支持向量机和多层感知机三种方法取得了相对较好的效果，而基于 SSTM 的实验中，线性回归、多层感知机和支持向量机三种方

法取得了较好的性能；从构建的症状特征角度观察，基于 SSTM 的特征构建方法在随机森林和线性回归两种策略下的性能超过了基于 TF-IDF 症状特征的结果，而在其它三种方法下则略逊于基于 TF-IDF 症状特征所达到的性能。综上所述，支持向量机和多层感知机在此经典名方数据上均表现出了相对较好的性能，在宏指标和加权指标下，两种方法的性能均为相对较优；同时总体而言，基于 TF-IDF 的症状特征略优于 SSTM 的症状特征。

但是由于本节将此任务视为多分类问题，在经典名方数据筛选过程中去除了大量因其样本数目过少的经典名方数据，这无疑对推荐结果的灵活性造成了不小的影响。因此，本节结合了本文第三章中提出的 SabKG 临床数据增强策略，对 2723 条经典名方数据先进行数据增强，而后基于相对较优的 TF-IDF 特征再进行后续的推荐任务。在进行数据增强的过程中，设置 SabKG 方法的产生样本数目为 3，即 1 条原始数据最多产生 3 条增强后样本，同时将增强后样本与原始样本合并去重，从而形成增强后的经典名方数据。经统计，增强后经典名方数据数目为 10160 条，并在处方对应样本数目阈值为 5 的筛选条件下，保留了 9347 条数据，并进行后续的多分类任务。数据划分与实验参数设置相关内容与增强前数据实验设置保持一致。

基于 SabKG 数据增强与 TF-IDF 症状特征的经典名方推荐结果如表 5-4 和图 5-4 所示，其中表 5-4、图 5-4A 和图 5-4B 展示了留出法下增强后数据的实验结果，图 5-4C 展示了五折交叉验证下增强后数据的准确率结果。从结果中可以看出，高斯朴素贝叶斯、支持向量机和多层感知机三种方法均取得了相对较优的性能，同时增强后数据的性能在这三种方法下均得到了一定的提升，这反映出本文第三章提出的 SabKG 数据增强策略的有效性。

表 5-4 基于增强后经典名方数据的各方法的实验性能

Table 5-4 Experimental performance of augmented classical prescription data with different classification methods

Method	Macro			Weighted			Accuracy
	P	R	F1	P	R	F1	
RF	0.1501	0.1067	0.1129	0.2804	0.2467	0.2126	0.2467
LR	0.4248	0.3423	0.3612	0.5609	0.5411	0.5005	0.5411
NB	0.9407	0.9346	0.9305	0.9476	0.9471	0.9417	0.9471
MLP	0.9437	0.9321	0.9288	0.9472	0.9369	0.9336	0.9369
SVM	0.9765	0.9764	0.9744	0.9760	0.9765	0.9741	0.9765

从上述方法及实验中可以看出，这种基于经典名方数据的多分类处方推荐策略的推荐结果源于经典名方数据，结果必然符合中医开方配伍原则，其配伍合理性可以得到保证。但是由于这种策略视处方推荐为多分类任务，其性能受限于经

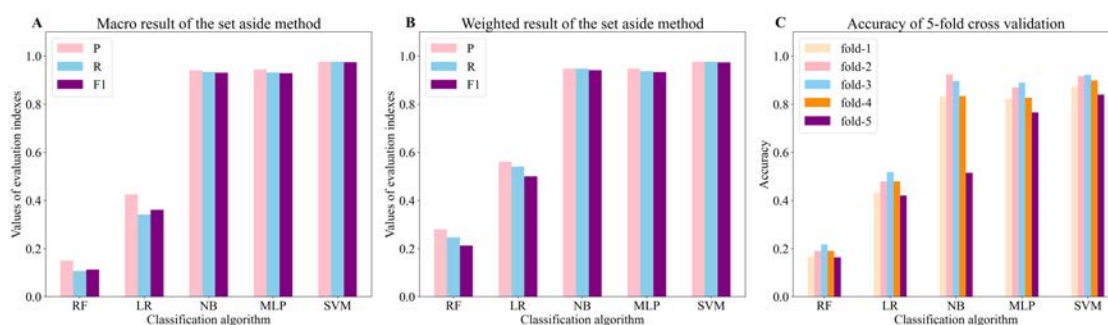


图 5-4 基于增强后经典名方数据的各方法下实验性能对比

Fig 5-4 Performance comparison between different classification methods under augmented clinical prescription data

典名方数据的“一多一少”，即训练过程受限于样本数量和质量。因此，这种方法不是一种相对较优的策略，需要继续探索推荐策略，使得推荐方法和结果不受限于经典名方样本的数量。

5.3 基于表型相似性与经典名方的处方推荐方法

上一节的研究中，本文构建了基于经典名方数据的多分类处方推荐策略，并结合了第三章提出的 SabKG 增强策略以及相关实验说明了这种策略的优势，即推荐结果源于经典名方，符合中医开方配伍原则，但这种策略的性能易受经典名方数据样本量的影响。因此在本节中提出了基于表型相似性与经典名方的处方推荐方法，这种策略推荐结果仍为经典名方，并且该策略不受经典名方的样本量影响。下面对提出的策略进行阐述。

5.3.1 整体框架

本节提出的基于表型相似性与经典名方的处方推荐策略流程如图5-5所示，其主要思想是根据患者间表型的相似性来推荐处方，即如果某一患者能够在某组药物的服用下治疗其症状，那么与这个患者症状高度相似的另一位患者亦可服用这组药物来进行治疗。

本文提出的基于表型相似性与经典名方的处方推荐策略的整体流程阐述如下。首先对临床医案数据和经典名方数据分别形成其症状特征表示，然后使用临床医案数据中的样本作为待推荐对象，利用其样本对应的症状表型特征与经典名方数据中所有样本的症状表型特征进行相似度匹配，以找到与待推荐样本症状相似度最高的经典名方样本，再将找到的经典名方样本所对应的处方及其药物组成作为

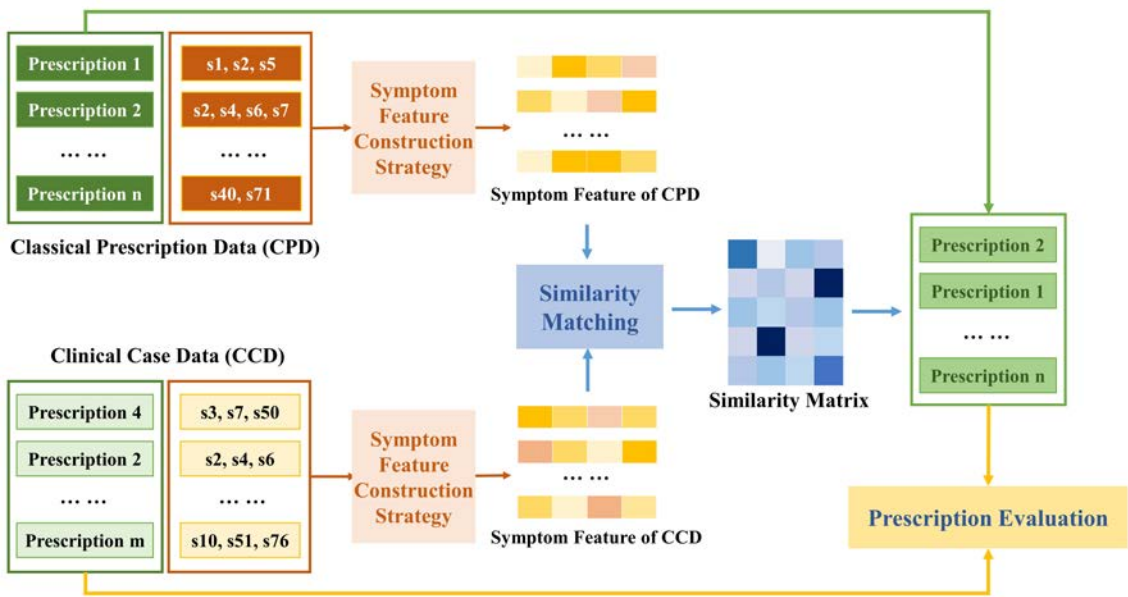


图 5-5 基于表型相似度与经典名方的处方推荐策略

Fig 5-5 Prescription recommendation method based on symptom similarity and the classical prescription data

推荐结果返回给待推荐样本，并根据其药物组成进行评估。在此过程中，对所有经典名方样本和临床医案数据样本的症状特征的形成，以及经典名方与临床医案症状特征相似度的匹配决定了该策略的性能与推荐效果。下面对患者表型特征的构建以及表型相似度匹配方法进行阐述。

5.3.2 患者表型特征构建策略

本节结合了已构建的症状网络和基于子图抽取的症状术语映射方法，形成了患者表型特征的构建策略，构建流程如图 5-6A 所示。构建患者症状特征时所需的输入是患者的症状列表，该策略能够在患者原始症状的基础上形成以下 3 种表示，用于后续表型相似度匹配任务中：

- (1) 融合症状网络嵌入的患者症状表示：在已形成的症状网络表示中寻找患者现有症状术语的嵌入向量，并进行求和平均法融合，形成基于患者原始症状与症状网络的融合嵌入向量；
- (2) 结合症状网络子图抽取的患者症状集合：利用患者原始症状列表，通过基于子图抽取的症状术语映射方法 SSTM 在已构建的症状网络寻找与原始症状关联的症状，并进行一定的筛选，形成结合症状网络的患者症状集合；
- (3) 结合症状网络子图抽取的患者症状表示：从已构建的症状网络表示中寻找

形成的结合症状网络的患者症状集合中涉及的症状术语表示，并进行求和平均法融合，从而形成基于症状网络子图抽取后的患者症状表示向量。

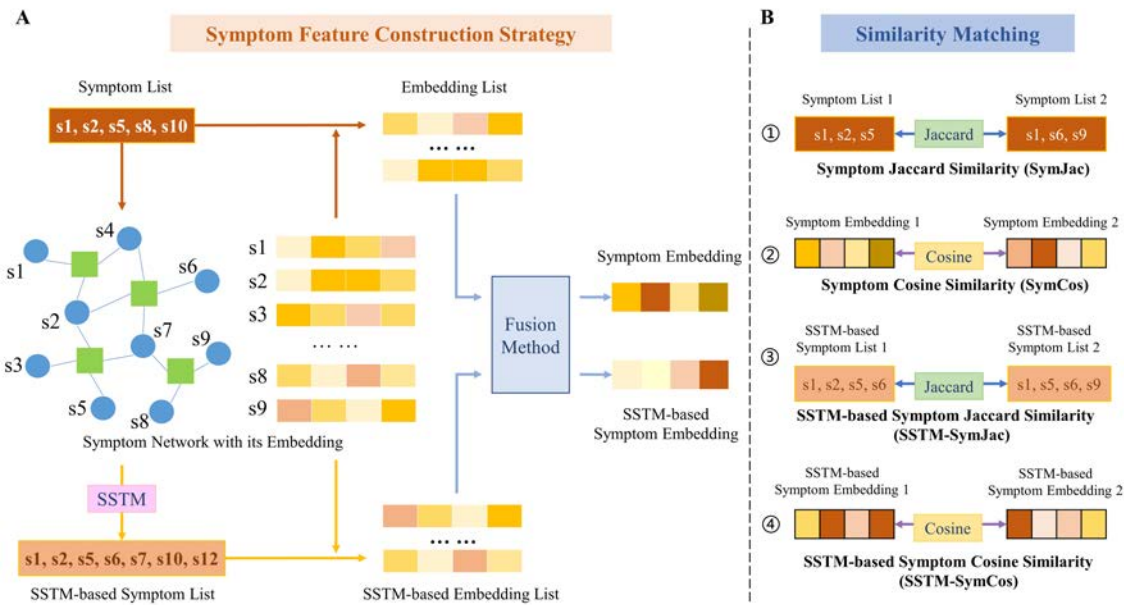


图 5-6 症状表型特征构建策略与表型相似度匹配方法

Fig 5-6 Symptom feature construction strategy and phenotypic similarity matching method

本节对临床医案数据和经典名方数据中的症状，分别使用以上症状特征构建策略形成对每条样本对应症状的表示，而后用于表型相似度匹配任务中，以找到与医案数据相似度最高的经典名方数据。

5.3.3 表型相似度匹配方法

在构建了患者症状特征后，本节将形成的临床医案数据症状特征与经典名方数据的症状特征进行相似度匹配，找到与医案数据最匹配的经典名方，并将其处方和药物组成作为该医案患者的推荐结果。本文使用的表型相似度匹配方法如图 5-6 B 所示，共形成了 4 种匹配策略，方法具体细节阐述如下。

(1) 原始症状集合间 Jaccard 相似度 (Symptom Jaccard Similarity, SymJac): 直接计算临床医案样本包含的症状集合与经典名方样本包含的症状集合的 Jaccard 相似度^[86] (如图 5-6 B ① 所示)。假定临床医案样本所包含的症状集合为 $set_{Sym_{clinical}}$ ，经典名方样本所包含的症状集合为 $set_{Sym_{classical}}$ ，则两者 Jaccard 相似度的计算方式如式 5-7 所示，

$$Jaccard(set_{Sym_{clinical}}, set_{Sym_{classical}}) = \frac{Intersection(set_{Sym_{clinical}}, set_{Sym_{classical}})}{Union(set_{Sym_{clinical}}, set_{Sym_{classical}})} \quad (5-7)$$

其中, $Intersection(x, y)$ 表示 x, y 两个症状集合的交集的元素个数, $Union(x, y)$ 表示 x, y 两个症状集合的并集的元素个数。

(2) 患者症状表示间余弦相似度 (Symptom Cosine Similarity, SymCos)^[87]: 对融合症状网络嵌入后临床医案样本与经典名方样本的患者症状表示计算余弦相似度 (如图5-6 B ②所示)。假定临床医案样本的症状表示为 v_1 , 经典名方样本的症状表示为 v_2 , 则两者的余弦相似度的计算方式如式 5-8 所示。

$$Cosine(v_1, v_2) = \frac{v_1 v_2}{\|v_1\| \|v_2\|} = \frac{\sum_{i=1}^n v_{1i} v_{2i}}{\sqrt{\sum_{i=1}^n (v_{1i})^2} \sqrt{\sum_{i=1}^n (v_{2i})^2}} \quad (5-8)$$

(3) 基于症状网络子图抽取的症状集合间 Jaccard 相似度 (SSTM-based Symptom Jaccard Similarity, SSTM-SymJac): 计算经子图抽取后临床医案样本包含的症状集合与经典名方样本包含的症状集合的 Jaccard 相似度 (如图5-6 B ③所示), 其计算方式同式 5-7。

(4) 基于症状网络子图抽取融合后的症状表示间余弦相似度 (SSTM-based Symptom Cosine Similarity, SSTM-SymCos): 对于临床医案样本与经典名方样本的结合症状网络子图抽取的患者症状表示, 计算两者余弦相似度 (如图5-6 B ④所示), 其计算方式同式 5-8。

5.4 基于表型相似性与经典名方的处方推荐实验及结果

上一节介绍了本文提出的基于表型相似性与经典名方的处方推荐策略, 详细阐述了患者症状表型特征的构建策略和表型相似度匹配策略, 形成了四种根据构建的患者表型特征来衡量患者表型相似度的方法。本节将阐述利用临床医案数据和经典名方数据进行处方推荐任务的过程以及实验结果。

5.4.1 实验设计与评价指标

本节的实验使用了从临床医案数据中随机选取的 1000 条样本, 将其分别与 2723 条经典名方样本进行匹配, 找到与医案样本最相似的经典名方, 并将该处方及药物组成推荐给该患者。对于医案数据患者特征与经典名方症状特征的形成与匹配, 本节使用了在 5.3 节中设计的 4 种表型相似度匹配方法分别进行实验, 以比较不同特征和不同匹配方法的效果。

需要说明的是, 在特征形成过程中使用的症状网络即为本文 4.2.2 节中介绍的内容, 其节点嵌入表示是利用 DeepWalk 算法生成的, 形成的每个节点的嵌入表示维度为 200 维; 对于结合症状网络子图抽取的患者症状集合和表示的形成, 设置子图抽取阈值为 1 和 2 分别进行实验, 以观察使用子图抽取方法提升表征症状的

能力；对于相关的特征融合环节，鉴于本文第四章的相关实验结果，本节统一使用了求和平均法作为症状特征的融合方式。

此外，对于 SymCos 策略，在寻找临床医案和经典名方症状表示的过程中，可能存在因“未登录词”过多而产生无法利用症状网络对患者症状进行表示的现象，因此，在形成医案数据和经典名方的症状特征过程中，如果所有的原始症状均未在现有症状网络表示库中找到对应的表示（即症状组中包含的所有症状词均为“未登录词”），则对其采用子图抽取策略以力求结合症状网络对其形成表示，如出现即使采用了子图抽取策略但是仍未找到与其相关的表示这种极端情况，则设定该症状组的表示为同维度的全零向量。就这一层面而言，本节形成的 SymCos 策略，以及 SSTM-SymJac 和 SSTM-SymCos 策略，都在不同程度上结合了基于子图抽取的症状术语映射策略。

本节对推荐后得到的经典名方的药物组成与医案数据中包含的原始药物信息进行对比，采用 Jaccard 评价指标对推荐性能进行衡量。假定临床医案样本中处方对应的药物集合为 $set_{Herb_{clinical}}$ ，经典名方样本所包含的药物集合为 $set_{Herb_{classical}}$ ，则两者 Jaccard 相似度的计算方式如式 5-9 所示，

$$Jaccard(set_{Herb_{clinical}}, set_{Herb_{classical}}) = \frac{Intersection(set_{Herb_{clinical}}, set_{Herb_{classical}})}{Union(set_{Herb_{clinical}}, set_{Herb_{classical}})} \quad (5-9)$$

其中， $Intersection(x, y)$ 表示 x, y 两个中药集合的交集的元素个数， $Union(x, y)$ 表示 x, y 两个中药集合的并集的元素个数。

5.4.2 实验结果及分析

本节基于 5.3 节中提出的患者表型特征构建和表型相似度匹配方法，根据上一小节设计的实验方案，对各种方法进行了相应实验，并计算得到了各方法下所有待推荐样本的中药的 Jaccard 评价指标的平均值，结果如表 5-5 所示。此外，在得到了所有样本匹配结果后，本节形成了各方法下得到的症状相似度的均值情况（即原始样本的症状与推荐结果对应症状的相似度，如表 5-6 所示）。此外，绘制了药物 Jaccard 指标的分布（见图 5-7）和原始样本与推荐结果的症状处方相似度散点分布情况（见图 5-8）。从结果中可以看出，提出的 SSTM-SymJac 和 SSTM-SymCos 两种方法能够显著提升药物的 Jaccard 指标。下面对各方法下的实验结果进行分析。

首先，从药物 Jaccard 指标均值（表 5-5）和分布上（图 5-7）可以看出，子图抽取度阈值为 2 时的结果均好于度为 1 时的情况，尤其是对于 SSTM-SymJac 和 SSTM-SymCos 两种策略而言，度为 2 时在药物 Jaccard 指标上分别提升了 12.81% 和 9.27%。这进一步显示出基于子图抽取的症状术语映射方法的优势所在：当度为 1 时，无论哪种策略下，基于子图抽取得到的症状集合中所包含的症状个数过多，

表 5-5 各方法下药物 Jaccard 评价指标均值统计

Table 5-5 The mean value of herb jaccard index under each method

Degree Threshold	SymJac	SymCos	SSTM-SymJac	SSTM-SymCos
1	0.0559	0.0485	0.0609	0.0593
2	0.0559	0.0489	0.0687	0.0648
Improvement (Degree 1 Vs. 2)	—	0.82%	12.81%	9.27%

表 5-6 各方法下原始样本与推荐结果的症状相似度均值统计

Table 5-6 The mean value of symptom similarity between the original sample and the recommended result under each method

Degree Threshold	SymJac	SymCos	SSTM-SymJac	SSTM-SymCos
1	0.1081	0.8396	0.3413	0.9135
2	0.1081	0.8352	0.2223	0.8640

并且与原始症状相比变化较大,这一变化虽引进了症状知识,但也引入了一些噪声;而经过了度为 2 时的筛选,保留下来的症状与原始症状组中相关症状的相似度得到了提升,从临床含义上也提升了与原始症状的语义关联,这一结果的出现与本文第四章的相关研究是一致的。

从表 5-6 和图 5-8 可以看到,对于原始样本与推荐样本的症状相似度而言,当度阈值从 1 变为 2 时,二者症状的相似度是普遍下降的,尤其是 SSTM-SymJac 和 SSTM-SymCos 两种策略症状相似度的降幅不小(SSTM-SymJac 方法下症状相似度降幅为 34.87%,SSTM-SymCos 方法下降幅为 5.42%),但是对应的药物 Jaccard 值上升了。这一现象也是融合了子图抽取术语映射方法所产生的:对于利用两条不同的原始症状抽取后形成的两个症状集合,阈值为 2 得到的两个集合元素的关联性比于阈值为 1 而言相对较弱,其原因在于筛选强度的增加能够利用与该症状组密切相关的症状进行引入,这使得抽取后症状集合贴切于原始症状组,不会因症状网络中元素间的复杂相互关联而引入过多噪声。

而对于 SymCos 方法,其性能随度阈值的增加也出现了同样的变化但不明显,其原因这是由于在 SymCos 策略中并未以子图抽取策略作为其形成症状特征的核心,但遇到全为“未登录词”的样本时,子图抽取策略则需要帮助该样本形成特征,因此结果上只产生了较小的变化,但其变化亦和 SSTM-SymJac 和 SSTM-SymCos 的结果变化相同。

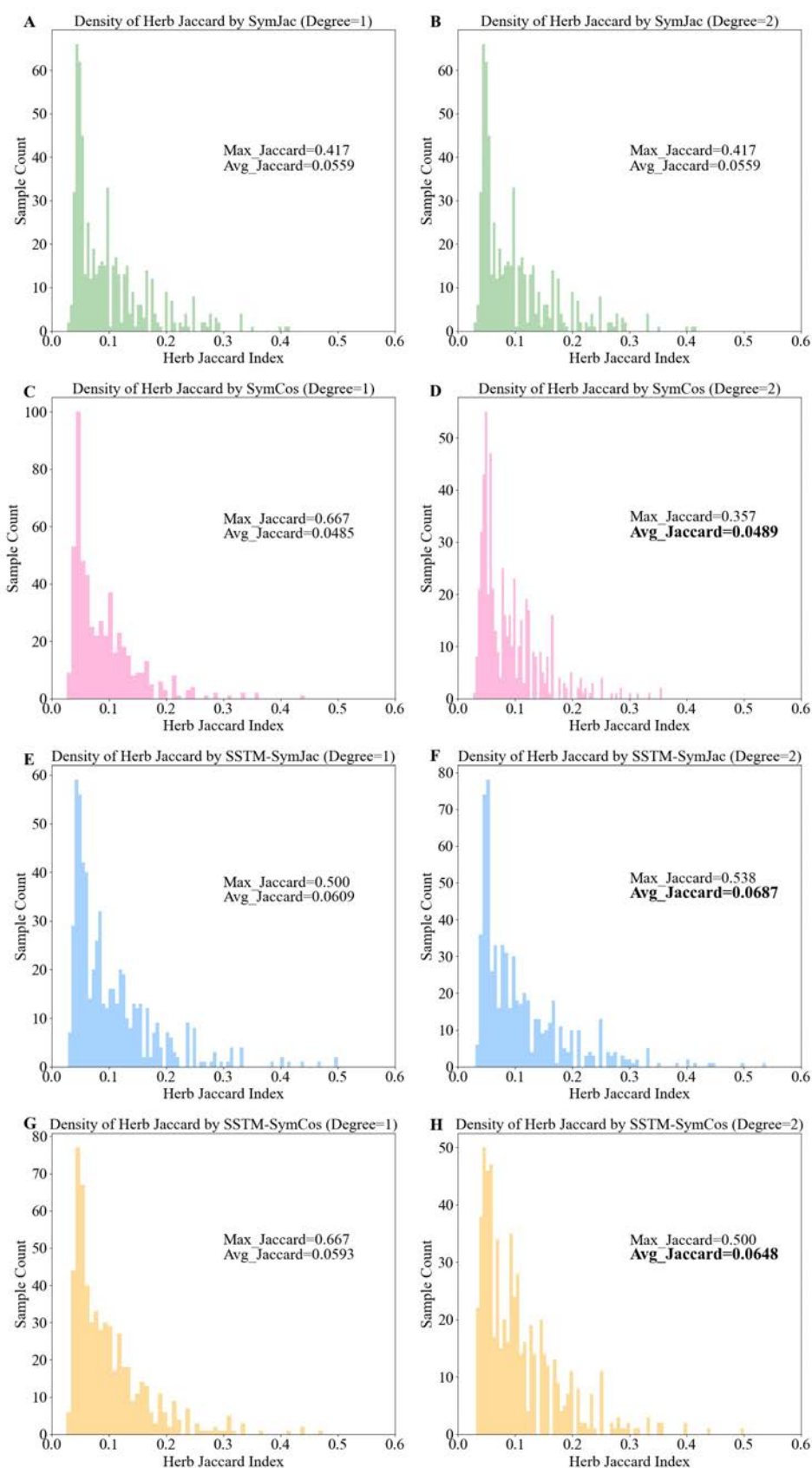


图 5-7 各方法下药物相似度评价结果

Fig 5-7 Experiment results of herb similarity under different methods

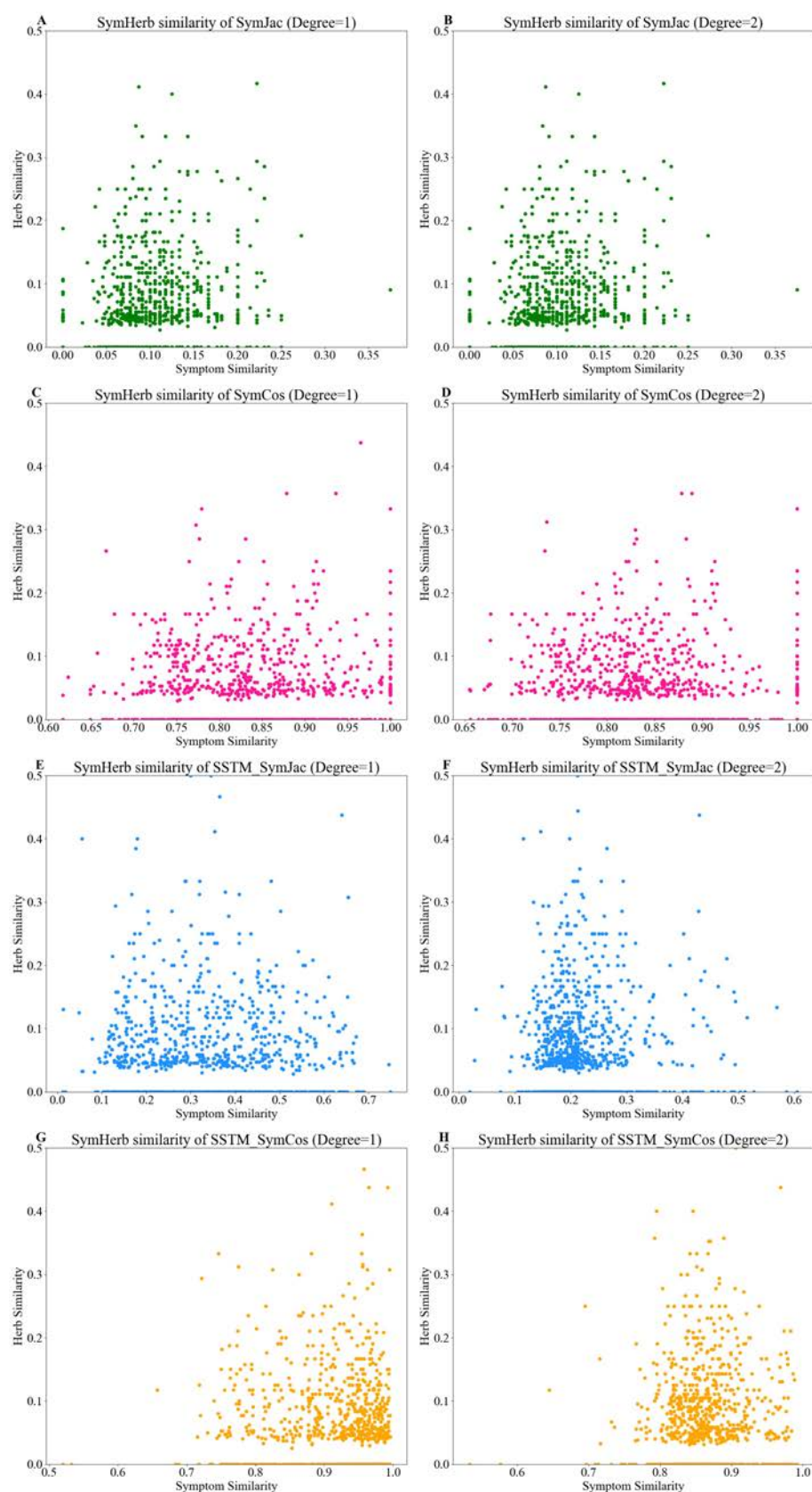


图 5-8 各方法下原始样本与推荐结果的症状药物相似度分布

Fig 5-8 Symptom herb similarity distribution between the original sample and the recommended result under different methods

对于各个方法下症状药物相似度的分布情况，图 5-8 中左侧 A、C、E、G 四张子图是度为 1 时四种方法得到的推荐结果与原始样本的症状处方相似度散点分布，右侧 B、D、F、H 四张子图是度为 2 时四种方法的结果，从图中可以看出 SymCos、SSTM-SymJac 和 SSTM-SymCos 三种策略都能提升原始症状与匹配得到的经典名方的症状相似度和药物相似度，尤其是对于度阈值为 2 时的 SSTM-SymJac 方法（图 5-8F），可以明显看出该方法改善了原始数据症状药物相似度分布值低且发散的现象。但无论哪种方法下，均存在着一些“特异点”，即图中贴近与横轴的一些点，这些点表示无论症状相似度在哪个范围内，其药物相似度均为 0，这一现象显示出数据本身的特点，即存在着症状非常特异或症状“未登录词”较多的数据，使得无法与现有经典名方数据进行匹配，这一点也提示出数据质量的重要性。

从总体上看，SSTM-SymJac 和 SSTM-SymCos 取得了相对较好的效果，并且 SSTM-SymJac 方法的性能相对最优。但是各方法下计算得到的所有待测样本的药物 Jaccard 值均偏低。这一情况出现可能由以下几方面原因导致：

(1) 医案数据中存在的“未登录词”对结果产生着一定影响。本文第四章对“未登录词”激活效果进行了相关探讨，其中原始医案数据集的平均登录率约为 24%，这对症状特征的形成以及后续的推荐任务影响较大。为探讨原始医案数据的登录率对本章相关研究的影响，对在本节实验中所使用的 1000 条医案样本根据登录率进行了筛选，并在筛选后的分组结果下观察子图阈值为 2 时各方法的药物 Jaccard 评价指标以及登录率的变化。经计算，实验中所用的 1000 条测试医案样本的平均登录率为 24.08%。本节以 0.2 作为登录率的划分间隔，对测试样本进行划分，并对几组样本在子图阈值为 2 时各方法的药物 Jaccard 评价指标和登录率情况进行了统计，结果如表 5-7 所示。此处需要说明的是，由于登录率在 0.8~1.0 间的样本数目过少，因此将登录率在 0.6~1.0 的样本进行了合并。

表 5-7 登录率筛选下各方法所得药物 Jaccard 评价指标均值统计（子图阈值为 2 时）

Table 5-7 Average result of herb jaccard evaluation index obtained by various methods under different record term threshold (Subnetwork degree threshold is 2)

登录率区间	样本数	登录率	各方法下药物 Jaccard 指标			
			SymJac	SymCos	SSTM-SymJac	SSTM-SymCos
0.0 ~0.2	451	8.36%	0.0516	0.0469	0.0677	0.0624
0.2 ~0.4	382	30.60%	0.0610	0.0517	0.0718	0.0695
0.4 ~0.6	143	48.39%	0.0591	0.0485	0.0653	0.0629
0.6 ~1.0	24	70.96%	0.0360	0.0446	0.0601	0.0500

从表格中可以看出：登录率区间在 0.2~0.4 的结果优于 0.0 ~0.2 的性能结果，药物的 Jaccard 得到了提升，并且在登录率区间为 0.4~0.6 时部分方法性能得到了再次提升；但由于数据固有的登录率低的特点，登录率高的样本数量过少（登录率

在 0.6 以上的仅有 24 条样本), 因此出现了登录率在 0.6 ~1.0 时药物 Jaccard 指标略有下降的现象; 同时可以看到, 无论在何种筛选条件下, 其结果均展示出与原始测试样本相同的现象, 即 SSTM-SymJac 和 SSTM-SymCos 两种策略均在各组达到了较好性能。这些现象表明随着登录率的提升, 各方法的性能都能得到一定提升。综上所述, 原始数据的登录率对症状特征的形成以及后续的推荐任务有着一定的影响。

(2) 药物 Jaccard 指标不高与数据本身特点直接相关。本节对进行测试的 1000 条医案数据涉及的药物组成与所有经典名方的各药物组成进行了 Jaccard 相似度计算, 统计了各测试样本下与之最相似的经典名方的相似度, 经统计, 这 1000 条样本的最大相似度均值为 0.2928, 从这个数据中可以看出, 即使所有测试医案样本找到的经典名方所包含的药物组成与原医案样本最相似, 其相似度上限的平均水平亦未超过 0.3, 由此可见得到的药物 Jaccard 值较低这一现象也是由数据自身的特点所决定的。

(3) 对于两集合的 Jaccard 评价指标, 由于计算方式是使用两集合交集个数除以并集个数, 并且医案数据和经典名方数据每条样本所包含的平均药物数量较多 (每条医案数据平均包含药物数量为 11.31 个, 每条经典名方数据为 9.16 个), 使得即使推荐的经典名方与医案数据交集相对较多, 但是由于并集数量大, 使得 Jaccard 值偏小。

(4) Jaccard 评价指标只能体现推荐结果与原始样本在药物重叠上的差异, 并未体现出处方功效的内在差异: 对于同一患者, 不同医生可能开出药物组成截然不同的处方, 但是都能够对治疗患者起到帮助作用, 但仅从药物组成的重叠情况来评价似乎不是一种最优策略, 这也是目前领域内工作中存在的另一难点。

同时注意到, 在形成的四种方法中, 以计算症状嵌入余弦相似度的两种方法 (SymCos 和 SSTM-SymCos) 得到的结果均略低于以计算症状集合 Jaccard 的两种方法 (SymJac 和 SSTM-SymJac) 的结果。原因可能是症状“未登录词”带来的影响: 对于 SymCos 方法, 其采用的策略是首先在症状网络中寻找症状词的嵌入表示, 只有当所有症状词均未找到对应的嵌入向量时, 再使用子图抽取策略进行弥补, 这种策略能够尽可能优先于使用原始症状术语进行特征表示, 但是由于“未登录词”的存在, 使得形成的表示向量可能不足以表示原始症状组中的所有症状词, 因此对后续的相似度计算产生着影响, 从结果中也可以看出 SymCos 方法的结果与 SymJac 方法的结果相对弱些; 对于 SSTM-SymCos 方法, 其使用了子图抽取策略作为核心, 能够对“未登录词”进行表示, 因此在性能上与 SSTM-SymCos 略差一点。此外, 已构建的症状网络表示的表征能力可能不够强大, 也可能是这种结果出现的原因。

5.4.3 案例分析

为进一步说明本章所提出的结合表型相似性与经典名方的处方推荐方法的有效性，本节选取部分案例进行展示与分析，选取的案例及其结果如表 5-8 所示，表中展示了医案数据中的两个案例（每个案例包含症状和药物组成信息），并展示了这些案例在不同方法下匹配到的最相似经典名方样本的具体信息（症状、方剂和药物组成）。

表 5-8 处方实例在不同方法下匹配到最相似样本及结果展示

Table 5-8 The most similar result of prescription examples under different matching strategy

实例与匹配策略	症状	处方	药物
实例样本 1	口苦, 口渴不欲饮, 恶心, 腹胀, 纳差, 乏力, 小便短少而黄, 大便粘滞不爽, 发热, 舌质红, 苔黄腻, 脉滑数	—	茵陈, 大黄, 栀子, 车前草, 柴胡, 虎杖, 白花蛇舌草, 金银花, 连翘, 陈皮, 半夏, 黄芩, 甘草
SymJac	纳呆, 舌质红, 足底部红肿热痛, 伴恶寒, 头痛, 发热, 脉滑数, 苔黄腻	五神汤合草薢渗湿汤加减	金银花, 野菊花, 紫背天葵, 紫花地丁, 牛膝, 草薢, 土茯苓, 薏苡仁
SymCos	恶心呕吐, 微恶风寒, 苔薄白或薄黄, 咽红疼痛, 脉浮数, 舌边尖红, 颈部瘰癧, 发热, 瘰癧较大, 纳差	银翘散	金银花, 连翘, 淡豆豉, 山慈菇, 瓜蒌, 牛蒡子, 荆芥, 薄荷, 芦根, 桔梗, 甘草
SSTM-SymJac	小便短黄, 舌质红, 苔黄燥或黄腻, 燥结不甚, 脉滑数, 潮热汗出, 溏滞不爽, 腹痛拒按, 烦渴引饮, 湿热较重, 大便不爽, 大便秘结	大承气汤	大黄, 枳实, 厚朴, 芒硝, 黄芩, 黄连, 神曲, 白术, 茯苓, 泽泻
SSTM-SymCos	小便短黄, 舌质红, 苔黄燥或黄腻, 燥结不甚, 脉滑数, 潮热汗出, 溏滞不爽, 腹痛拒按, 烦渴引饮, 湿热较重, 大便不爽, 大便秘结	大承气汤	大黄, 枳实, 厚朴, 芒硝, 黄芩, 黄连, 神曲, 白术, 茯苓, 泽泻
实例样本 2	胃脘隐痛痞满胀闷, 嘈杂吐酸, 暖气不舒, 便溏, 舌淡红, 舌下络脉淡紫粗长, 苔薄, 脉弦细有抓痕及结痂, 舌淡红, 苔薄, 痒痒剧烈, 脉弦细, 痒痒甚者, 病情反复发作, 皮损肥厚干燥有鳞屑, 病程长, 或呈苔藓样变	—	党参, 白术, 半夏, 陈皮, 降香, 公丁香, 海螵蛸, 瓦楞子, 甘草
SymJac	便溏, 舌淡, 神倦懒言, 素喜热饮, 畏寒肢冷, 面色不华, 甚则色黑, 脉细, 阳虚较甚, 伴脘腹隐痛, 便血紫暗	当归饮子合消风散	当归, 生地黄, 防风, 蝉衣, 牛蒡子, 火麻仁, 僵蚕, 丹参, 甘草
SymCos	食后胃脘不舒, 大便溏薄, 苔薄, 倦怠乏力, 面色萎黄, 舌淡, 饮食减少, 胃脘满闷, 恶心呕吐, 暖气, 脉弱	黄土汤	灶心黄土, 白术, 炮附子, 干地黄, 阿胶, 黄芩, 甘草
SSTM-SymJac	食后胃脘不舒, 大便溏薄, 苔薄, 倦怠乏力, 面色萎黄, 舌淡, 饮食减少, 胃脘满闷, 恶心呕吐, 暖气, 脉弱	加味四君子汤	人参, 黄芪, 白术, 炙甘草, 茯苓, 扁豆
SSTM-SymCos	食后胃脘不舒, 大便溏薄, 苔薄, 倦怠乏力, 面色萎黄, 舌淡, 饮食减少, 胃脘满闷, 恶心呕吐, 暖气, 脉弱	加味四君子汤	人参, 黄芪, 白术, 炙甘草, 茯苓, 扁豆

一方面，案例展示出了本章提出的症状匹配策略的有效性。对于案例 1，患者症状主要表现在小便短黄、大便粘滞不爽、腹胀、恶心等症状，从四种匹配策略可以看出，SSTM-SymJac 与 SSTM-SymCos 两种策略的结果与该患者症状的症状文本在临床含义上覆盖度较高（表中加粗文字含义为各方法的结果与原始样本在临床含义上重叠之处），而 SymJac 与 SymCos 两种策略得到结果的相似性则相对较弱，从案例 2 中亦能得到类似现象，这些现象与前节中的实验结果是相符的，这体现了本章提出的患者表型匹配策略的有效性。

另一方面,对于药物组成,可以观察到无论何种匹配策略,得到的药物组成与原始药物的重叠不高,对于药物的 Jaccard 指标则体现在指标值过低,这亦显示出如果仅从药物组成的重叠情况来评价实验结果,则未必能真实反映匹配到的结果的有效性,此结论与前一节中所得相关结论一致。因此,形成适合智能推荐处方的评价指标是下一步亟待探索的问题。

5.5 本章小结

本章围绕现有处方推荐方法的推荐结果的配伍合理性这一角度出发,结合了经典名方数据,提出了结合经典名方的处方推荐策略和基于表型相似度与经典名方的处方推荐方法。结合经典名方的处方推荐策略将方剂名称作为推荐对象,解决了现有方法中以药物为推荐对象的不足,但这种策略受限于经典名方数据。因此提出了基于表型相似性与经典名方的处方推荐方法,在融合了领域知识的基础上设计了患者特征形成的几种策略,并相应地形成了患者相似度匹配方法,实验结果表明,本章提出的 SSTM-SymCos 和 SSTM-SymCos 策略能够提升推荐结果在药物 Jaccard 评价指标上的表现。两种方法得到的推荐结果都源自经典名方数据,因此结果符合中医开方配伍原则,其药物配伍的合理性得到了保障,达到了本章的研究目标。

6 总结与展望

本章对本文的研究工作进行总结。第一节对全文的研究工作进行了回顾与总结，第二节对本文未来工作进行了研究展望。

6.1 全文工作总结

本文针对处方推荐领域现存的问题和研究难点开展了研究：提出了结合领域知识的处方数据增强策略，包括 SOCO 模型和 SabKG 模型；构建了基于症状术语映射与深度学习的处方推荐方法 TCMPR 模型；形成了基于经典名方的处方推荐方法，并提出了基于表型相似性与经典名方的处方推荐方法。下面对各部分内容进行总结。

(1) 针对临床诊疗数据中存在的“一多一少”现象，提出了结合领域知识的数据增强策略，包括 SOCO 模型和 SabKG 模型。SOCO 模型结合了症状本体库和医案数据中的症状共现关系进行临床诊疗数据增强，SabKG 模型能够结合药症知识图谱中症状和药物间的关联，作为知识进行引入，以进行临床诊疗数据增强。本文将形成的增强数据运用于多标签处方推荐任务，并将提出的方法与基线方法进行了对比实验，实验结果显示出本文提出的两种策略都能够使原始数据的性能在增强后得到提升，并且 SabKG 方法在性能上得到了相对最优的表现。结果展示出本文引入症状本体、症状共现关系、药症知识图谱等领域知识来辅助数据增强这一策略的合理性和有效性。

(2) 针对目前处方推荐现有方法性能不强、以及无法对“未登录词”形成表示的难点，本文提出了基于症状术语映射与深度学习的处方推荐方法 TCMPR。该模型首先结合了基于子图抽取的术语映射方法和已构建的症状网络，将数据中存在的症状“未登录词”进行表示，而后结合了深度学习模型进行处方推荐任务。本文将提出的模型与现有基线方法进行了对比，结果显示出 TCMPR 在性能上优于现有方法，并且能够较好地“未登录词”进行表示。同时对模型中的关键模块进行了相关实验探究，结果显示，子图筛选阈值的设置以及症状嵌入表示方法的选取是影响模型性能的关键因素，对登录词是否进行处理、患者症状填充策略的选择以及患者特征融合方式能够对模型性能产生一定的影响。总体而言，本文提出的模型融合了药症知识图谱和症状网络的信息，这一领域知识的引入能够帮助模型更好地形成对“未登录词”的表示，并且提升了处方推荐方法的性能。

(3) 针对现有方法推荐结果的配伍合理性问题，本文形成了基于经典名方的处

方推荐方法,该策略将推荐结果由药物组合转为真实处方,从实验结果可以看到结合支持向量机和多层感知机的结果达到了较好的性能,并且基于 TF-IDF 的症状特征略优于基于 SSTM 的特征;此外结合了本文第三章提出的 SabKG 方法对经典名方数据进行了增强,其结果显示出 SabKG 方法的有效性。但是这种策略受限于经典名方样本,因此本文提出了基于表型相似性与经典名方的处方推荐方法,该方法融合了症状网络这一领域知识进行症状特征的形成,从结果中可以看到提出的 SSTM-SymJac 和 SSTM-SymJac 方法的有效性。这一章节的工作反映出患者症状特征表示的形成对于后续的任务及性能具有一定的影响。

6.2 未来研究展望

本文针对中医处方推荐领域中现存的一些问题进行了相关工作,研究结果达到了预期目标,但仍存在一些有待进一步解决的问题。因此本节针对本文研究工作中存在的不足,对未来工作提出以下三方面展望。

(1) 构建面向中医处方推荐的标准实验数据集,提高现有经典名方数据集质量。第三章和第五章的相关研究中体现出了数据质量对于研究结果的重要性,近年来的相关工作所各自使用的数据并未统一,并且对于所使用的临床诊疗数据的规范化程度不高,特别是对于症状的规范仍需加强。此外,现有经典名方数据多为从古籍或教材直接摘录,其数据质量仍需提升。

(2) 探索融合中医“理法方药”诊疗策略的智能处方推荐方法,提高推荐过程和结果的可解释性。本文的相关研究仅从融合领域知识的角度进行了一些探索,而未融合真实世界中医“理法方药”的整体诊疗过程。因此,后续的研究中应结合“理法方药”的流程,形成更符合中医真实诊断的处方推荐方法。

(3) 形成针对智能推荐方的评价指标。在本文的第三章和第四章中使用的评价指标多为 Top@K 指标,原因在于形成的结果为药物组合的概率排序。在第五章中主要以药物的 Jaccard 来进行评价,但由于原始处方和推荐方中涉及药物数目较多,这种评价指标下得到的 Jaccard 值相对偏低。上述两种评价指标只是从药物组合角度进行药物重叠的评价,而并未从推荐结果的功效层次进行性能评估,这也是领域中相关研究的一个难点。

综上所述,本文对中医处方推荐领域中现有问题进行了相关探索,取得了初步的研究成果。后续研究将从提升处方推荐实验数据集与经典名方数据集质量、探索融合中医“理法方药”诊疗过程的处方智能推荐策略、形成处方智能推荐结果评价指标等方面开展进一步的研究,以求推动中医处方推荐方法研究的发展。

参考文献

- [1] 马惠聪. 人类命运共同体理念的中国智慧——基于全球共抗时疫的思考 [J]. 攀登, 2020, 39 (05): 135–136.
- [2] Zheng G, Jiang M, Lu C, et al. Prescription analysis and mining [M]. In Data Analytics for Traditional Chinese Medicine Research. Springer, 2014: 97–109.
- [3] Huang Y, Wang S, Wang L, et al. Exploring the rules of li-fa-fang-yao on diabetes mellitus within traditional chinese medicine through text mining [C]. In 2012 7th International Conference on Computing and Convergence Technology (ICCCCT), 2012: 1369–1373.
- [4] Yao L, Zhang Y, Wei B, et al. A topic modeling approach for traditional Chinese medicine prescriptions [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30 (6): 1007–1021.
- [5] 王鸿江, 申俊龙, 徐佩, 等. 中医“医联体+智能化”促进中医基层化的模式研究 [J]. 中国农村卫生事业管理, 2019, 39 (10): 701–704.
- [6] 肖勇, 田双桂, 沈绍武. 我国中医药信息化建设与发展的思考 [J]. 医学信息学杂志, 2019, 40 (7): 12–17.
- [7] 国家中医药管理局. 中医药信息化发展“十三五”规划 [EB/OL]. [2016-12-12]. <http://www.satcm.gov.cn/guicaishi/gongzuodongtai/2018-03-24/2144.html>.
- [8] 叶苏婷, 刘俊彤, 毕迎春, 等. 医疗信息化发展现状及策略实践 [J]. 电脑知识与技术, 2021, 17 (19): 156–158.
- [9] 王昱. 基于电子病历数据的临床决策支持研究 [D]. 杭州: 浙江大学, 2016: 2.
- [10] Mi X, Ikeda H, Nakazawa F, et al. Prescription Prediction towards Computer-Assisted Diagnosis for Kampo Medicine [C]. In 2015 International Conference on Computer Application Technologies, 2015: 126–131.
- [11] 冼向阳, 张志强, 李纪麟. 基于推荐算法和主成分分析的智能处方推荐系统 [J]. 中国数字医学, 2018, 13 (12): 23–24, 67.
- [12] 王斌, 刘涛, 王广志, 等. 支持新型冠状病毒肺炎的中医智能处方推荐和知识库系统 [J]. 中国数字医学, 2020, 15 (5): 25–27.
- [13] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification [J]. Pattern recognition, 2004, 37 (9): 1757–1771.
- [14] Zhang M, Zhou Z. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern recognition, 2007, 40 (7): 2038–2048.
- [15] Clare A, King R D. Knowledge discovery in multi-label phenotype data [C]. In European conference on principles of data mining and knowledge discovery, 2001: 42–53.
- [16] Elisseeff A, Weston J. A kernel method for multi-labelled classification [J]. Advances in neural information processing systems, 2001, 14: 681–687.
- [17] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification [J]. Machine learning, 2011, 85 (3): 333–359.

- [18] Shi Y, Yang W, Thung K-H, et al. Learning-Based Computer-Aided Prescription Model for Parkinson's Disease: A Data-Driven Perspective [J]. IEEE Journal of Biomedical and Health Informatics, 2020, 25 (9): 3258–3269.
- [19] Wang Z, Poon J, Poon S. Tcm translator: A sequence generation approach for prescribing herbal medicines [C]. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019: 2474–2480.
- [20] 韩亚楠, 刘建伟, 罗雄麟. 概率主题模型综述 [J]. 计算机学报, 2021, 44 (6): 1095–1139.
- [21] Zhang X, Zhou X, Huang H, et al. Topic model for chinese medicine diagnosis and prescription regularities analysis: case on diabetes [J]. Chinese journal of integrative medicine, 2011, 17 (4): 307–313.
- [22] Zhang X, Zhou X, Huang H, et al. A hierarchical symptom-herb topic model for analyzing traditional Chinese medicine clinical diabetic data [C]. In 2010 3rd International Conference on Biomedical Engineering and Informatics, 2010: 2246–2249.
- [23] Jiang Z, Zhou X, Zhang X, et al. Using link topic model to analyze traditional Chinese medicine clinical symptom-herb regularities [C]. In 2012 IEEE 14th international conference on e-health networking, applications and services (Healthcom), 2012: 15–18.
- [24] Yao L, Zhang Y, Wei B, et al. Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge [J]. Journal of biomedical informatics, 2015, 58: 260–267.
- [25] Wang L, Zhang Y, Zhang Y, et al. Prescription function prediction using topic model and multi-label classifiers [J]. Evidence-Based Complementary and Alternative Medicine, 2017, 2017.
- [26] Huang Z, Dong W, Bath P, et al. On mining latent treatment patterns from electronic medical records [J]. Data mining and knowledge discovery, 2015, 29 (4): 914–949.
- [27] Wang S, Hu Y, Tan W, et al. Compatibility art of traditional Chinese medicine: from the perspective of herb pairs [J]. Journal of ethnopharmacology, 2012, 143 (2): 412–423.
- [28] Zhou X, Liu B, Wu Z. Text mining for clinical Chinese herbal medical knowledge discovery [C]. In International Conference on Discovery Science, 2005: 396–398.
- [29] Yang K, Zhang R, He L, et al. Multistage analysis method for detection of effective herb prescription from clinical data [J]. Frontiers of medicine, 2018, 12 (2): 206–217.
- [30] Jin Y, Chen X, Huang W, et al. Analysis of the prescription of auricular acupoint therapy for simple obesity based on complex network techniques [J]. World journal of acupuncture-moxibustion, 2018, 28 (1): 38–43.
- [31] Feng Y, Qiu Y, Zhou X, et al. Optimizing prescription of Chinese herbal medicine for unstable angina based on partially observable Markov decision process [J]. Evidence-based Complementary and Alternative Medicine, 2013, 2013.
- [32] 许帆. 基于临床数据的中医处方推荐方法研究 [D]. 北京: 北京交通大学, 2019: 23.
- [33] 章亚东, 胡孔法, 杨涛, 等. 基于复杂网络的中医治疗肺癌的处方推荐算法 [J]. 时珍国医国药, 2019, 30 (5): 1257–1260.
- [34] 杨铭, 李嘉旗, 焦丽静, 等. 基于复杂网络结合生存分析的中医药治疗肺癌的核心有效处方的发现研究 [J]. 中国中药杂志, 2015, 40 (22): 4482–4490.

- [35] Zhang Q, Zhang G, Lu J, et al. A framework of hybrid recommender system for personalized clinical prescription [C]. In 2015 10Th international conference on intelligent systems and knowledge engineering (ISKE), 2015: 189–195.
- [36] Li W, Yang Z. Exploration on Generating Traditional Chinese Medicine Prescriptions from Symptoms with an End-to-End Approach [C]. In CCF International Conference on Natural Language Processing and Chinese Computing, 2019: 486–498.
- [37] Wang Y. A Novel Chinese Traditional Medicine Prescription Recommendation System based on Knowledge Graph [C]. In Journal of Physics: Conference Series, 2020: 012019.
- [38] Jin Y, Zhang W, He X, et al. Syndrome-aware herb recommendation with multi-graph convolution network [C]. In 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020: 145–156.
- [39] Yang Y, Rao Y, Yu M, et al. Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation [J]. Neural Networks, 2022, 146: 1–10.
- [40] Jin Y, Ji W, Zhang W, et al. A KG-enhanced Multi-Graph Neural Network for Attentive Herb Recommendation [J]. IEEE/ACM transactions on computational biology and bioinformatics, 2021.
- [41] Wu Y, Yin Z, Zhou K, et al. A Hybrid-scales Graph Contrastive learning Framework for Discovering Regularities in Traditional Chinese Medicine Formula [C]. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021: 1104–1111.
- [42] Li C, Liu D, Yang K, et al. Herb-know: Knowledge enhanced prescription generation for traditional chinese medicine [C]. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020: 1560–1567.
- [43] Aujla G S, Jindal A, Chaudhary R, et al. Dlrs: deep learning-based recommender system for smart healthcare ecosystem [C]. In ICC 2019-2019 IEEE International Conference on Communications (ICC), 2019: 1–6.
- [44] Wang L, Zhang W, He X, et al. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation [C]. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 2447–2456.
- [45] Wang S, Ren P, Chen Z, et al. Order-free medicine combination prediction with graph convolutional reinforcement learning [C]. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 1623–1632.
- [46] Gong F, Wang M, Wang H, et al. SMR: medical knowledge graph embedding for safe medicine recommendation [J]. Big Data Research, 2021, 23: 100174.
- [47] Bhoi S, Li L M, Hsu W. Premier: Personalized recommendation for medical prescriptions from electronic records [J]. arXiv preprint arXiv:2008.13569, 2020.
- [48] 杨光, 郝逸凡. 基于互信息算法的抗前列腺癌药物重定位分析 [J]. 沈阳师范大学学报 (自然科学版), 2019, 37 (1): 34–37.
- [49] Xue H, Li J, Xie H, et al. Review of drug repositioning approaches and resources [J]. International journal of biological sciences, 2018, 14 (10): 1232.

- [50] 邵杨芳, 王帅, 张芬利, 等. 国内药物重定位研究的主题及发展脉络分析 [J]. 中国新药杂志, 2020, 29 (22): 2541–2551.
- [51] 刘艳飞, 孙明月, 赵莹科, 等. 网络药理学在中药药物重定位研究中的应用现状与思考 [J]. 中国循证医学杂志, 2017, 17 (11): 1344–1349.
- [52] 陈国飞, 沈媛, 宋琦, 等. 基于基因表达谱相似性的四物汤重定位及抗乳腺癌有效成分群辨识 [J]. 世界科学技术-中医药现代化, 2021, 23 (9): 3217–3225.
- [53] Yang X, Wang W, Huang Y, et al. Network Pharmacology-Based Dissection of the Active Ingredients and Protective Mechanism of the *Salvia Miltiorrhiza* and *Panax Notoginseng* Herb Pair against Insulin Resistance [J]. ACS omega, 2021, 6 (27): 17276–17288.
- [54] Song T, Zhong Y, Ding M, et al. Repositioning molecules of Chinese medicine to targets of SARS-Cov-2 by deep learning method [C]. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020: 2306–2312.
- [55] 孙昱, 徐敢, 汪祺. 中药二次开发的研究思路探讨 [J]. 中草药, 2021, 52 (13): 4107–4113.
- [56] Zhang M, Zhou Z. A review on multi-label learning algorithms [J]. IEEE transactions on knowledge and data engineering, 2013, 26 (8): 1819–1837.
- [57] Cortes C, Vapnik V. Support-vector networks [J]. Machine learning, 1995, 20 (3): 273–297.
- [58] Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines [J]. IEEE transactions on Neural Networks, 2002, 13 (2): 415–425.
- [59] De'ath G, Fabricius K E. Classification and regression trees: a powerful yet simple technique for ecological data analysis [J]. Ecology, 2000, 81 (11): 3178–3192.
- [60] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification [C]. In Joint European conference on machine learning and knowledge discovery in databases, 2009: 254–269.
- [61] Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multilabel classification [J]. IEEE transactions on knowledge and data engineering, 2010, 23 (7): 1079–1089.
- [62] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification [C]. In European conference on machine learning, 2007: 406–417.
- [63] Peterson L E. K-nearest neighbor [J]. Scholarpedia, 2009, 4 (2): 1883.
- [64] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521 (7553): 436–444.
- [65] Goodfellow I, Bengio Y, Courville A. Deep learning [M]. MIT press, 2016: 3.
- [66] 邱锡鹏. 神经网络与深度学习 [M]. 北京: 机械工业出版社, 2020: 93.
- [67] Shorten C, Khoshgoftaar T M, Furht B. Text data augmentation for deep learning [J]. Journal of big Data, 2021, 8 (1): 1–34.
- [68] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks [J]. arXiv preprint arXiv:1901.11196, 2019.
- [69] Zhang D, Yin J, Zhu X, et al. Network representation learning: A survey [J]. IEEE transactions on Big Data, 2018, 6 (1): 3–28.

- [70] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C]. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014: 701–710.
- [71] Grover A, Leskovec J. node2vec: Scalable feature learning for networks [C]. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016: 855–864.
- [72] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding [C]. In Proceedings of the 24th international conference on world wide web, 2015: 1067–1077.
- [73] Guthrie D, Allison B, Liu W, et al. A closer look at skip-gram modelling. [C]. In LREC, 2006: 1222–1225.
- [74] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.
- [75] Elberse A. Should you invest in the long tail? [J]. Harvard business review, 2008, 86 (7/8): 88.
- [76] Wang X, Zhang Y, Wang X, et al. A knowledge graph enhanced topic modeling approach for herb recommendation [C]. In International Conference on Database Systems for Advanced Applications, 2019: 709–724.
- [77] Sun Y, Han J, Yan X, et al. Pathsime: Meta path-based top-k similarity search in heterogeneous information networks [J]. Proceedings of the VLDB Endowment, 2011, 4 (11): 992–1003.
- [78] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [J]. Advances in neural information processing systems, 2013, 26.
- [79] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples [C]. In International Conference on Learning Representations, 2018.
- [80] Liu W, Wen Y, Yu Z, et al. Large-margin softmax loss for convolutional neural networks. [C]. In ICML, 2016: 7.
- [81] Ramos J, et al. Using tf-idf to determine word relevance in document queries [C]. In Proceedings of the first instructional conference on machine learning, 2003: 29–48.
- [82] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression [M]. John Wiley & Sons, 2013: 35.
- [83] Breiman L. Random forests [J]. Machine learning, 2001, 45 (1): 5–32.
- [84] Rish I. An empirical study of the naive Bayes classifier [C]. In IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001: 41–46.
- [85] Zhang Z, Lyons M, Schuster M, et al. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron [C]. In Proceedings Third IEEE International Conference on Automatic face and gesture recognition, 1998: 454–459.
- [86] Niwattanakul S, Singthongchai J, Naenudorn E, et al. Using of Jaccard coefficient for keywords similarity [C]. In Proceedings of the international multiconference of engineers and computer scientists, 2013: 380–384.
- [87] Rahutomo F, Kitasuka T, Aritsugi M. Semantic cosine similarity [C]. In The 7th International Student Conference on Advanced Science and Technology ICAST, 2012: 1.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

董鑫，男，1997年12月生，籍贯：河北省秦皇岛市。

2016年9月至2020年6月，就读于江西中医药大学，计算机学院，计算机科学与技术（医药软件工程）专业，获工学学士；

2020年9月至今，就读于北京交通大学，计算机与信息技术学院，计算机技术专业攻读硕士学位，导师周雪忠教授。

二、参与科研项目

[1] 2020年9月至2021年12月，参与国家重点研发计划项目《中医药大数据挖掘研究与创新应用》。

[2] 2021年2月至2021年5月，参与国家重点研发计划项目《真实人脑、人体、人际视/听觉认知、感知神经免疫内分泌系统耦联网络（rBNN+）非线性映射模型构建及其人工智能转化与现实检验》。

[3] 2021年11月至2022年3月，参与国家自然科学基金面上项目《基于真实世界和大数据技术的顽固性高血压痰瘀互结证精准方药方案的方法研究》。

三、发表论文

[1] **X Dong**, Y Zheng, Z Shu, et al. TCMPR: TCM Prescription recommendation based on subnetwork term mapping and deep learning[C]//2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021: 3776-3783. (**EI**, **CCF B** workshop, 第一作者)

[2] **X Dong**, Y Zheng, Z Shu, et al. TCMPR: TCM Prescription Recommendation Based on Subnetwork Term Mapping and Deep Learning[J]. BioMed Research International, 2022, 4845726. (**SCI**, **IF=3.411**, JCR Q2/Q3, 第一作者)

[3] 许宁, **董鑫**, 钟昆禹, 等. 基于复杂网络的特发性肺纤维化病证结合人群分型研究 [J]. 世界科学技术-中医药现代化, 2021, 23(09): 3109-3117. (北大核心期刊, 第二作者)

四、专利

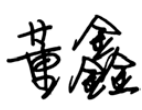
[1] 周雪忠, 杨扩, 郑毅, **董鑫**, 夏佳楠. 药症关系网络构建与概念映射方法及系统 [P]. 北京市: CN113779265A, 2021-12-10.

[2] 杨扩, 周雪忠, 贾彩燕, **董鑫**. 一种基于复杂网络的感知行为与分子网络机制关联方法 [P]. 北京市: CN114121166A, 2022-03-01.

[3] **董鑫**, 周雪忠, 郑毅, 杨扩. 基于症状术语映射与深度学习的中医处方推荐方法 [P]. 北京市: CN114141361A, 2022-03-04.

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：

签字日期：2022年5月21日

学位论文数据集

表 1.1 数据集页

关键词 *	密级 *	中图分类号	UDC	论文资助
处方推荐; 临床 数据增强; 症状 术语映射; 深度 学习; 表型特征 构建	公开			
学位授予单位名称 *		学位授予单位 代码 *	学位类别 *	学位级别 *
北京交通大学		10004	电子信息	硕士
论文题名 *		并列题名		论文语种 *
融合领域知识的中医处方推荐方法研究				中文
作者姓名 *	董鑫		学号 *	20125155
培养单位名称 *		培养单位代码 *	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西 直门外上园村 3 号	100044
专业领域 *		研究方向 *	学制 *	学位授予年 *
计算机技术		医学人工智能	2 年	2022 年
论文提交日期 *	2022 年 6 月			
导师姓名 *	周雪忠		职称 *	教授
评阅人	答辩委员会主席 *		答辩委员会成员	
	刘峰		赵宏智 常冬霞	
电子版论文提交格式 文本 (✓) 图像 () 视频 () 音频 () 多媒体 () 其他 () 推荐格式: application/msword; application/pdf				
电子版论文出版 (发布) 者		电子版论文出版 (发布) 地		权限声明
论文总页数 *	69			
共 33 项, 其中带 * 为必填数据, 为 21 项。				