

基于多源情境协同感知的药品推荐

郑 值¹ 徐 童¹ 秦 川¹ 廖祥文² 郑 毅³ 刘同柱⁴ 童贵显⁴

¹(中国科学技术大学计算机学院 合肥 230027)
²(福州大学数学与计算机科学学院 福州 350116)
³(华为技术有限公司 杭州 310051)
⁴(中国科学技术大学附属第一医院 合肥 230027)
(zhengzhi97@mail.ustc.edu.cn)

Multi-Source Contextual Collaborative Recommendation for Medicine

Zheng Zhi¹, Xu Tong¹, Qin Chuan¹, Liao Xiangwen², Zheng Yi³, Liu Tongzhu⁴, and Tong Guixian⁴

¹(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027)
²(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116)
³(Huawei Technologies., Hangzhou 310051)
⁴(The First Affiliated Hospital of the University of Science and Technology of China, Hefei 230027)

Abstract Recent years have witnessed the accumulation of electronic medical records (EMR), as well as the rapid development of data analytics techniques, which highly support the intelligent medical services, e. g., automatic diagnosis and medicine recommendation. Unfortunately, due to the simplicity of general EMR, the diagnosis model could be easily disturbed by common diseases or symptoms, thus fine-grained prescription with personalized focalization will hardly be achieved. At the same time, we realize that some related context information, e.g., personalized information like age and sexuality, treatment records like examinations, and external information like weather and temperature, could all benefit the diagnosis and medicine recommendation task. However, these information could not be effectively extracted and integrated by current techniques, which constrains the performance of medicine recommendation. To that end, in this paper, we propose a comprehensive framework based on the collaborative awareness of multi-source context information. Specifically, we first utilize the bag-of-words model to process the EMR and related context records. Along this line, a LDA-based contextual collaborative model called Medicine-LDA has been designed to integrate the multi-source information, while at the same time, alleviate the problem of combination explosion of context information. Extensive experiments on the real-world data set from a first-rate hospital demonstrate the effectiveness of our solution.

Key words context-aware; medicine recommendation; topic model; multi-label learning; recommender system

收稿日期:2020-03-10;修回日期:2020-05-18
基金项目:国家重点研发计划项目(2018YFB1004300);国家自然科学基金项目(U1605251,61703386);中央高校基本科研业务费专项资金(WK9110000014);安徽省重点研发计划项目(1804b06020377)
This work was supported by the National Key Research and Development Program of China (2018YFB1004300), the National Natural Science Foundation of China (U1605251, 61703386), the Fundamental Research Funds for the Central Universities (WK9110000014), and the Key Research and Development Program of Anhui Province (1804b06020377).
通信作者:徐童(tongxu@ustc.edu.cn)

摘 要 电子医疗记录的快速积累与数据分析技术的日益成熟,为实现包含智能诊断与药品推荐等功能的智慧医疗服务奠定了基础.然而,电子病历的精简性与患者症状描述的模糊性,导致诊断模型容易受到高发疾病与常见症状的干扰,从而无法支撑细粒度的诊断与处方,在药品推荐上缺乏针对性.与此同时,病情描述以外的许多情境信息,如患者的性别、年龄等个人信息,诊疗过程、检查结果等记录信息,以及所在地的天气、温差等外部信息等,也对于细化对于患者的诊断和处方有着重要的辅助作用.然而,这些多源异构信息往往难以被现有技术所有效提取与整合,从而限制了病情诊断与药品推荐的有效性.针对这一问题,提出了一种基于多源情境协同感知的药品推荐方法,在有效整合多源异构情境信息的基础上,为实现病情诊断与药品推荐提供了具有可解释性的依据.具体而言,首先使用词袋模型对病历和相应的情境数据进行处理,然后设计了一种基于 LDA 模型的情境主题模型 Medicine-LDA,在融合患者病情描述与相应情境信息的同时,有效缓解了情境信息组合爆炸的问题.基于某大型三甲医院的电子病历数据集上的对比实验证明了该方法的有效性与鲁棒性.

关键词 情境感知;药品推荐;主题模型;多标签学习;推荐系统

中图法分类号 TP391

随着人口的增长与老龄化的加剧,人们对于高质量医疗服务的需求不断攀升,医疗系统的接诊压力与日俱增.根据国家卫健委所发布的全国医疗卫生统计数据显示^①,2018 年 1~11 月,全国医院诊疗人次共 32.3 亿人次,同比增长 5.3%.与之相对应的是,截止 2019 年,我国每千人口医生数为 2.59 人^②,与发达国家相比仍有着较大的差距,也导致了医护人员沉重的负担和医疗资源巨大的缺口.因此,若能有效借助人工智能技术,实现患者病情的智能化诊断与药品推荐,将极大地提高医疗系统运行效率,提升我国医疗服务的整体质量.

幸运的是,随着信息技术在医疗行业的普及,电子病历(electronic medical record, EMR)开始被各大医院广泛采用.通过医疗记录的数字化,海量丰富的诊疗信息,如患者的病程记录、检查检验结果、医嘱、手术记录、药品记录等得以记录,并有力支撑下游由大数据驱动的各种智能应用.在此基础上,相关领域的学者们展开了一系列的研究,并在智能导诊^[1]、医疗影像分析^[2]和慢性病随访^[3]等方面取得了不错的成效.但在面向医疗文本的智能化诊断与药品推荐方面,受限于数据、技术等方面的原因,进展尚较为缓慢.首先,我国电子病历推广较晚,导致目前可靠的医疗数据来源较少,且存在数据积累较少、各科室数据无法关联的问题.其次,与传统的推荐系统类似,药品推荐问题对数据质量较为敏感,也极大地影响了其推荐的效果,主要体现在 3 个方面: 1)绝大多数患者仅有一次就诊记录,不可能采用常

规的个性化推荐技术加以建模;2)疾病分布严重不均,常见疾病占据了绝大多数的病历记录,而小人群和罕见疾病则严重稀疏,难以保障其推荐质量;3)疾病的症状之间存在较高比例的重叠.以本次爆发的新冠肺炎为例,其早期症状与其他肺炎乃至普通流感有着颇多相似之处,即使有经验的医生也存在误判的风险.上述 3 个方面最终导致了一个共同的结果,即推荐的药品倾向大众化,缺乏面向特定人群、特定疾病的针对性.因此,我们迫切需要一种更为精细的推荐方法,以提升药品推荐的精确性和针对性.

与此同时,我们注意到电子病历在记录病人病情与诊疗过程的同时,也往往记录一些周边的情境信息,如患者的个人情况、开支、就诊时间等.而这些信息可望为我们实现人群细分的针对性诊断和开药带来新的线索.例如,通过就诊时间追溯天气、温度的变化,有助于我们诊断部分心脑血管或上呼吸道疾病;又如,根据病人是否采用医保支付,可以在确定推荐药品的品种时有所取舍.一个典型的例子是,在空气污染较为严重的时段,“流涕”对应抗过敏药物的概率要高于一般时段,因为此时这一症状更可能由于污染而非常规感冒导致.显而易见,借助情境感知数据挖掘技术,将为我们解决前述的困境提供了新的思路.

然而,这些数据同时也带来了新的难点,即如何综合考虑这些情境信息以实现有效诊疗.一方面,情境信息多属多源异构数据,面临信息不全、结构

① 2018 年 1~10 月全国医疗服务情况. <http://www.nhc.gov.cn/mohwsbwstjxxzx/s7967/201901/296ad9bb4a594a60b8755e7defff0f20.shtml>

② 国家卫健委:我国每千人口医生数达到 2.59 人. <http://news.cctv.com/2019/09/26/ARTIW0F5GSg6C8JBZIBRsFaL190926.shtml>

不统一的问题,除部分结构化数据外,还有大量信息以半结构化或非结构化文本的形式存储,处理起来较为困难;另一方面,情境信息对于药品的影响方式不同,如患者的性别和年龄的不同会导致患者适用药品的不同,而患者发病时的天气、温差等外界因素,则会导致患者易发不同的疾病,从而间接影响医生为患者所开的药品.因此倘若笼统地对所有的情境信息进行不加区分的建模,则可能削弱这些信息所起到的贡献.此外,最为重要的是,情境信息的类别多种多样,每种类别又对应着多种取值,若按照传统推荐系统的做法将每种组合当成一个“用户”,在情境信息数量较多时组合数也会迅速增加,带来“维度灾难”的问题,从而导致对训练数据量的需求过大,在实际应用中难以取得良好的效果.

为了解决上述的问题,本文提出了一个基于多源情境协同感知的药品推荐方法 Medicine-LDA(以下简称 MLDA).事实上,本文通过模拟现实中医生的诊断过程进行建模.在实际诊疗过程中,医生首先获取患者主诉并进行体格检查,然后进行必要的化验,通过化验所得的异常结果以及主诉和体格检查中的信息综合判断患者所患疾病,进而考虑患者的年龄、医保等情境信息,给出一组合适的药品.受此启发,本文通过 4 个步骤对此过程进行建模:

1) 提取数据中患者的主诉文本与检查检验异常,使用词袋模型将其综合为一个文档,用此文档代表此患者,称之为病情文档;

2) 使用主题模型对病情文档进行建模,假设该文档拥有一个潜在主题,该潜在主题代表患者所患疾病,拥有对应的药品分布和词分布;

3) 将患者对应所有情境信息综合为一个文档,称之为情境文档,使用 LDA 模型对其进行建模,获得患者的情境主题分布;

4) 假设情境主题对于患者所患疾病和患者适用药品均会造成影响,使用一个统一的概率模型框架融合药品主题和情境主题,从而对情境信息、病情信息和药品信息综合建模.

本文的主要贡献包括 3 个方面:

1) 提出一种基于 LDA 的情境信息处理方法,将患者的情境信息表示为情境主题分布,解决了不同情境信息组合数过多的问题.

2) 提出一种通用的融合情境信息的主题模型,该模型在融合情境信息时,同时考虑了情境信息对于患者易患疾病的影响和患者适用药品的影响,能够更有效地对情境信息的作用进行建模,并且具有良好的可扩展性与可解释性.

3) 在一个真实的来自大型三甲医院的数据集上进行了充分的实验,实验结果表明本文方法在药品推荐上具有较好的有效性,证实了模型的实际应用价值.

1 相关工作

本节从药品推荐系统,情境感知推荐系统以及主题模型 3 个方面介绍本文的相关工作.

1.1 药品推荐系统

目前,关于药品推荐的研究并不十分广泛,大部分药品推荐主要是利用用户与药品的交互记录,基于协同过滤技术来进行推荐.如文献[4]为了解决协同过滤推荐精度低的问题,提出基于用户相似度和信任度的药品推荐算法.该方法通过对药品聚类来降低时间复杂度,引入共同评分药品阈值和相似度阈值辅助选取相似邻居,并根据用户推荐可信度和评分可信度建立相似计算模型,从而提高系统推荐精度.文献[5]在用户评分相似度的基础上,引入人口属性相似度,通过加权线性融合来得到用户相似度.文献[6]提出利用聚类和基于用户的协同推荐算法对网上购药的用户进行个性化药品推荐.针对协同过滤冷启动以及数据稀疏性等问题,提出使用张量分解对用户、症状、药品三者进行建模.

以上研究均使用传统协同过滤中针对用户和物品的交互关系进行推荐的思想,并且仅使用患者对药品的打分以及患者的症状等简单信息描述一名患者.然而在实际应用中,患者的数量巨大,且绝大多数患者的记录数据较少,使用协同过滤的方法难以扩展至医院等大型医疗场所.另外,由于药品推荐的特殊性,仅仅使用患者口述症状和患者对药品的打分记录对患者进行刻画是十分不准确的.因此,以上方法由于模型本身的缺陷以及使用信息的不全面,难以实现有效的落地应用.

1.2 情境感知推荐系统

传统推荐系统,如基于内容的推荐与协同过滤等方法,忽略了用户在不同情况下会具有不同的偏好行为^[7].因此,研究者们提出了“情境感知”的数据挖掘技术以提升推荐的效果.所谓“情境”,指所有与人机交互相关的,用于区分标定当前特殊场景的信息.由于相同用户在不同情境下的喜好往往不同,基于情境感知的推荐系统可以根据用户所处的特定境况来给出更相关的推荐结果.文献[8]对情境感知推荐系统进行了较为详细的综述,指出情境感知推荐

系统中的推荐流程可以有情境预过滤、情境后过滤、情境建模 3 种形式.其中,情境建模方法直接在推荐中把情境信息作为预测的显式因素来考虑,是研究最为广泛的情境感知方法.例如文献[9]提出将情境信息作为特征维度加入推荐系统的表示向量空间中进行情境建模.该文使用支持向量机(support vector machine, SVM)作为分类器,证实了情境感知的 SVM 推荐效果优于非情境感知 SVM.而文献[10]使用张量分解的方法,对情境信息与基于隐式反馈的协同过滤进行了结合,并针对评价指标 MAP 进行了优化.

在药品推荐系统中,情境信息泛指一切影响患者病情与适用药品的因素.例如患者的性别、年龄、医保信息等内因会影响到患者的适用药品,当地的天气、温差等外因会影响到患者的易患疾病并间接影响患者药品.本文提出了一个能够有效融合各类情境信息的药品推荐系统,能解决药品推荐实际应用中的问题.

1.3 主题模型

主题模型(topic model)是对语料库中的隐含语义结构进行聚类的统计模型,被广泛应用于自然语言处理中的语义分析和文本挖掘问题.隐含狄利克雷分布(latent Dirichlet allocation, LDA)是最常见的主题模型^[11],其假设文档以一定的概率选择某个潜在主题,再从主题中以一定的概率选择一个单词,从而生成整个语料库.LDA 可以通过引入外部标签信息进行各种扩展,如 Labeled-LDA(LLDA)^[12]是对 LDA 模型在多标签语料库中的扩展,其假设语料库中的标签与潜在主题一一对应,从而使得模型直接学习外部标签与文档词之间的关系,在文档多标签学习任务中取得了较好的效果,Author-Topic Model(ATM)^[13]在 LDA 中融合了作者信息,从而获得不同作者对于不同主题的偏好.扩展过的主题模型可以被应用于推荐系统,如文献[14-16]利用主题模型对推特标签进行推荐,文献[17-18]利用主题模型对微博等文档进行推荐.

通过 LDA 模型得到的主题本质是隐式地学习得到词语之间的共现关系,此类方法在文档平均单词数较多的语料中效果较好,而面对如推特、短信息的平均词数较短的文本时,由于短文本中的词语共现不够频繁,此类方法效果往往不佳.有部分工作针对这一问题进行了研究,如文献[19]提出 Twitter-LDA,假设每个文档仅有一个主题,并使用一个伯努利分布判断文档中的词语属于主题词还是背景

词,全部的主题词均由文档的唯一主题对应的词分布生成,在推特数据集上的实验证实 Twitter-LDA 在推特文本数据集中可以比 LDA 取得更好的效果.文献[20]提出 BTM(biterm topic model),直接对整个语料库中的词语共现关系进行显式的建模,从而缓解了短文本中词语共现次数不足的问题.实验表明 BTM 不仅在短文本上的表现超过 LDA,并且在长文本上也取得了优于 LDA 的效果.本文基于文献[19]加以改进,假设每篇文档仅有一个主题,有效地提高了模型的效果.

2 预备知识

本节我们首先介绍本文研究工作中所采用的数据集及预处理情况,然后对本文所探讨的药品推荐问题给出正式的定义.

2.1 数据预处理

首先,我们通过模拟现实中医生对病人的诊断过程对问题进行描述.在实际诊疗过程中,医生往往会在患者的医疗文本中记录患者的主诉病情,以及医生对患者进行的体格检查.医生为患者编写的医疗文本,以及患者进行的化验等检查检验的结果是判断患者病情最为重要的依据.为此,本文选择医疗文本和患者检查检验描述患者病情,并对患者对应的医疗文本与检查检验进行如下处理:首先,对医疗文本进行分词处理,并去除停用词,得到患者对应的文本文档.然后选出患者进行的化验中的全部异常项,并将每一个异常项的与对应的异常状态组合成若干个词,如将“乙肝核心抗体”与其对应的异常状态“偏高”组合成词“乙肝核心抗体偏高”,由此得到患者的化验文档.随后,将患者的文本文档与化验文档组合起来,即得到了基于词袋模型的病情文档,用于描述患者病情.

对于患者对应的情境信息,由于每种情境信息可能对应较多的取值,如“年龄”这一情境信息可以对应上百种取值(0~100 以上),因此需要对情境信息进行分段处理,将同一类别的情境取值范围进行分段,并将每种情境与其对应的分段组合成若干个情境词,如将“年龄”和“老年”组合成“年龄-老年”,由此得到患者的情境文档.在表 1 中,我们对本文用到的全部情境信息与其对应分段方法进行了总结,分段的依据包括:1)生活常识,如性别分为男女,季节分为春夏秋冬;2)数据特征,如麻醉类型分为了全身麻醉、局部麻醉与未麻醉;3)外部知识,如依据常见

年龄分段法对年龄进行分段,依据保险金额与条件对保险情况进行分段;4)对于温差、气温等无参考知识的情境,依据生活常识与数据分布,使得不同分段内的数据分布较为均匀,同时分段方式贴近人们对于生活的认知.

Table 1 Context Information and Segmentation Method

表 1 情境信息与其分段方法

Context Information	Segmentation Method
Gender	male, female
Age	0~3 years old as infant, 4~12 as early youth, 13~45 as youth, 46~59 as middle age, 60+ as agedness
Insurance	high quality, low quality, self-paying, unknown
Operation	yes, no
Anesthesia	general, local, no
Season	spring, summer, autumn, winter
Temperature Difference	less than 10 degree centigrade, more than 10 degree centigrade
Average Temperature	less than 0, 0~10, 10~20, 20~30, more than 30 degree centigrade
Weather	cloudy, rainy, snowy, others

2.2 形式化描述

基于 2.1 节中的定义,我们将本文所研究的问题形式化描述如下:在情境感知药品推荐任务中,训练集中包含每一名患者 i 对应的病情文档 $W_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,p}\}$,其中 $w_{i,\cdot}$ 为医疗文本的分词得到的单词或单个化验异常项.相应的,我们有情境文档 $T_i = \{t_{i,1}, t_{i,2}, \dots, t_{i,q}\}$,其中 $t_{i,\cdot}$ 为单个情境词.此外,我们有药品集合 $D_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,s}\}$,其中 $d_{i,\cdot}$ 为医生为该患者所开的全部药品中的某一种药品.本文所研究问题的目标是通过训练集进行学习,使模型在给定测试集中新病情文档 W' 和新情境文档 T' 时,可以对于任意一种在数据集种出现过的药品 d_i ,输出其对此患者使用的概率 p_i .

由此可见,该问题本质上是一个多标签学习问题,且要求所用方法能够给出全部标签的概率分布.为此,本文使用基于主题模型这一概率模型的方法来融合情境信息进行药品推荐.

3 情境感知药品推荐方法

本节我们将对本文所提出的多源情境协同感知的药品推荐方法,及其所涉及的情境信息建模技术进行详细的介绍.

3.1 模型概述

本文所提出的模型大致步骤为:
1) 模型需要对情境信息进行处理.尽管在数据预处理部分中已经对情境信息的取值进行分段处理,然而,若假设有 K 种情境信息,每种情境有 N 个取值分段,则将有 N^K 种组合方式,导致模型难以处理较多数目的情境信息.为此,本文使用 LDA 模型,通过情境信息主题化的方式对情境信息进行软聚类,从而缓解情境信息组合爆炸的问题.

2) 本文使用一个概率框架对情境信息、病情文档、患者药品进行联合建模.建模思想来自实际医疗场所中的诊疗过程,即医生首先根据患者的症状、体格检查、化验报告来确定患者可能患有的疾病,然后根据患者对应的情境信息给出对此疾病适宜的药品.为此,本文使用基于主题模型的建模方法,假设每一篇病情文档均拥有一个隐变量,表示该患者所患的疾病,因此称之为疾病主题.具体而言,模型通过疾病主题联系情境、病情、药品 3 种信息进行建模,其技术细节将在以下的章节给出详细的介绍.

3.2 情境信息建模方法

为了解决患者情境信息组合爆炸的问题,需要设法对情境信息进行降维处理.一般而言,聚类是解决此类问题的常用手段,其大致可分为硬聚类和软聚类 2 种.其中,硬聚类方法对数据进行确切的分类,规定样本只能完全属于某一个类或完全不属于某一个类,代表算法有 K -Means 算法等.在本文场景中,情境信息代表人群的特征,由于人群特征多种多样,不同人群相似度、不同特征的重要程度均不尽相同,对人群进行硬聚类在应用中缺乏合适的分类标准,且难以给出合理的解释,因此不适用于本文场景.

与硬聚类不同的是,软聚类把数据以一定的概率分到各类中.例如,高斯混合模型(GMM)就是一种典型的软聚类方法.而对于文本数据,利用 LDA 主题模型进行软聚类的方法更为常用.对于患者对应的情境信息而言,若使用词袋模型把每名患者对应的全部情境信息视为一篇情境文档,则可使用 LDA 模型通过主题建模来进行软聚类.图 1 为 LDA 模型的概率图表示,其中 w 为文档单词, z 为隐变量表示文档主题.在训练得到每种主题的词分布后, LDA 模型可以将任意文档表示为一个主题分布,此即为软聚类中各类别的概率分布.

本文采用文献[21]中的 LDA 吉布斯采样法对模型进行求解.首先,对每一篇情境文档中的每一个词赋予一个随机情境主题.在接下来的每一轮迭代

更新中,针对文档中的每一个词,首先根据其他词的主题赋值来求解该词被赋予不同情境主题的概率,然后根据这个概率对该词重新赋予情境主题.在多次迭代收敛后,每个词均被赋予了固定的情境主题,从而可以使用式(1)计算得到每一篇情境文档的主题分布.

$$P(z \mid d) = \frac{n_{d,z} + \alpha}{\sum_{i=1}^k n_{d,z_i} + K\alpha}, \tag{1}$$

其中 $n_{d,z}$ 表示情境文档 d 中词语被赋予主题 z 的数量, α 为狄利克雷先验参数, K 为情境主题数.

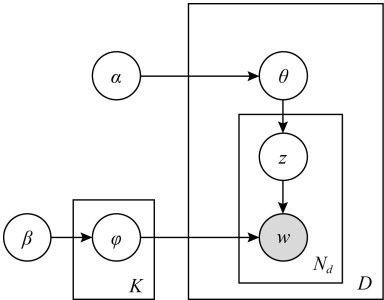


Fig. 1 Graphical representation of LDA
图1 LDA模型的概率图表示

3.3 情境感知的药品推荐方法

完成对情境信息的软聚类后,即可对全部信息进行综合建模.本文提出的建模方法基于对诊疗过程的2点假设:1)情境主题是情境信息的软聚类,每种情境主题既可能影响患者的适用药品,又可能影响患者的易发疾病.如以天气、季节等因素为主的情境主题会更偏向于影响患者的易发疾病,而以手术、医保等信息为主的情境主题更可能影响医生对于患者的药品使用.2)每名病人的数据对应一个隐变量,代表患者所患疾病,称之为疾病主题.疾病主题是连接情境信息、病情文档、药品三者的纽带,情境主题影响疾病主题的分布,而疾病主题影响病情文档中的词分布以及药品集合中的药品分布.根据这2点基于真实诊疗过程的假设,本文提出了一个统一的概率模型 Medicine-LDA(简称 MLDA)进行情境感知的药品推荐.以下给出模型的详细描述.

假设语料库中有 K 个疾病主题, F 个情境主题,每一个疾病主题拥有一个对应的病情文档词分布,每一个疾病主题和情境主题的组合拥有一个对应的药品分布.令 φ 为疾病主题对应的病情文档词分布, φ^B 为背景词分布. θ 为每种人群对应的疾病分布, f 为每种疾病主题和每种情境主题的组合对应

的药品分布, ψ 为患者对应的情境文档的情境主题分布, π 为一个伯努利分布,其作用如文献[19]所述,用于筛选文档中与主题推断关系较小的词语,控制病情文档中的词属于背景词还是主题词.图2给出模型的概率图表示,病情文档与药品集合的生成过程如下所述,其中符号描述如表2所示.

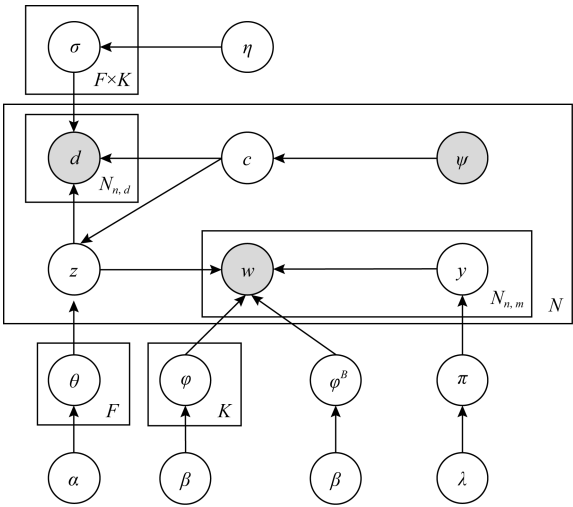


Fig. 2 Graphical representation of MLDA
图2 MLDA模型的概率图表示

Table 2 Basic Symbols Description
表2 基本符号描述

Symbol	Symbol Description
α	Dirichlet prior of θ
β	Dirichlet prior of φ and φ^B
λ	Beta prior of π
η	Dirichlet prior of f
ψ_n	Topic distribution of the n -th context document
c_n	n -th context topic
θ_c	Disease topic distribution of c th context topic
φ_k	Word distribution of k -th disease topic
φ^B	Background word distribution
π	Bernoulli distribution
$f_{z,c}$	Medicine distribution of disease topic z and context topic c
z_n	Disease topic of n -th illness document
$w_{n,m}$	m -th word in n -th illness document
$y_{n,m}$	Sign of m -th word in n -th illness document
$d_{n,m}$	m -th medicine of n -th illness document
K	Number of disease topics
F	Number of context topics
N	Number of patients
$N_{n,m}$	Word number of n -th illness document
$N_{n,d}$	Medicine number of n -th illness document

1) 采样获得 $\varphi^B \sim \text{Dir}(\beta), \pi \sim \text{Beta}(\lambda)$.

2) 对于每一种疾病主题 $z \in \{1, 2, \dots, K\}$ 采样获得 $\varphi_k \sim \text{Dir}(\beta)$, 对于每一种情境主题 $c \in \{1, 2, \dots, F\}$ 采样获得 $\theta_c \sim \text{Dir}(\alpha)$, 对于每一种疾病主题与情境主题的组合 $(z, c) \in \{(1, 1), (1, 2), \dots, (K, F)\}$, 采样获得 $f_{z,c} \sim \text{Dir}(\eta)$.

3) 对于每一名患者 $n \in \{1, 2, \dots, N\}$:

① 采样获得患者情境主题 $c_n \sim \text{Multi}(\phi_n)$;

② 采样获得患者疾病主题 $z_n \sim \text{Multi}(\theta_c)$;

③ 对于病情文档中每一个词 $m \in \{1, 2, \dots, N_{n,m}\}$:

i. 采样获得 $y_{n,m} \sim \text{Ber}(\pi)$;

ii. 若 $y_{n,m} = 0$, 则采样 $w_{n,m} \sim \text{Multi}(\varphi^B)$; 若 $y_{n,m} = 1$, 则采样 $w_{n,m} \sim \text{Multi}(\varphi_{z_n})$;

④ 对于药品集合中的每一种药品 $m \in \{1, 2, \dots, N_{n,d}\}$, 采样获得 $d_{n,m} \sim \text{Multi}(f_{z,c})$.

3.4 采样与推断算法

MLDA 需要对隐变量 z, c, y 进行推断, 由于精确推断算法不可解, 本文使用倒塌的吉布斯采样法 (collapsed gibbs sampling, CGS) 对隐变量进行采样, 并根据统计获得参数 θ, φ, f , 进而使用训练得到的参数对新患者进行药品推荐.

训练时的采样算法如下. 首先对与训练集中所有数据进行随机初始化: 对于第 n 条数据, 随机赋予情境主题隐变量 c_n 和疾病主题隐变量 z_n , 对于第 n 条数据中的第 m 个病情文档词, 随机赋予其对应标识隐变量 $y_{n,m}$. 随后在吉布斯采样的每一轮迭代过程中, 对于任意一篇文档的任一隐变量, 将对给定其余隐变量时该隐变量的条件概率分布进行采样, 从而更新其取值. 隐变量 z 与 c 的条件概率分布为

$$p(z_n = k, c_n = f | \phi_n, w, d, z_{\neg n}, y, \alpha, \beta, \eta) \propto \phi_n^f \times \frac{n_{f, \neg n}^k + \alpha}{\sum_{j=1}^K (n_{f, \neg n}^j + \alpha)} \times \prod_{m=1}^{N_{n,m}, y_{n,m}=1} \frac{n_{k, \neg n}^{w_{n,m}} + \beta}{\sum_{w=1}^V (n_{k, \neg n}^w + \beta)} \times \prod_{d=1}^{N_{n,d}} \frac{n_{f, k, \neg n}^{d_{n,d}} + \eta}{\sum_{d=1}^D (n_{f, k, \neg n}^d + \eta)}, \quad (2)$$

其中, ϕ_n^f 为从 ϕ_n 中采样得到 f 的概率 $p(f | \phi_n)$, $n_{k, \neg n}^w$ 为词 w 被赋予为主题 k 的次数, $n_{f, k, \neg n}^d$ 为药品 d 被赋予为药品主题 k 和情境主题 c 的次数, 以上 2 个计数均排除了文档 n . 值得注意的是, 病情文档中仅有被赋予为主题词, 即 $y = 1$ 的词才对疾病主题的采样具有贡献.

对于病情文档中的每一个词, 其对应的标识变量 y 的条件概率分布为

$$p(y_{n,m} = 0 | w, z, c, d, y_{\neg(n,m)}, \lambda, \beta) \propto \frac{n_{\neg(n,m)}^{y=0} + \lambda}{\sum_{y=0}^1 (n_{\neg(n,m)}^y + \lambda)} \times \frac{n_{y=0, \neg(n,m)}^{w_{n,m}} + \beta}{\sum_{w=1}^V (n_{y=0, \neg(n,m)}^w + \beta)}, \quad (3)$$

$$p(y_{n,m} = 1 | w, z, c, d, y_{\neg(n,m)}, \lambda, \beta) \propto \frac{n_{\neg(n,m)}^{y=1} + \lambda}{\sum_{y=0}^1 (n_{\neg(n,m)}^y + \lambda)} \times \frac{n_{z_{n,m}, \neg(n,m)}^{w_{n,m}} + \beta}{\sum_{w=1}^V (n_{z_{n,m}, \neg(n,m)}^w + \beta)}, \quad (4)$$

其中, $n_{\neg(n,m)}^y$ 为对应标识为 y 的词语个数, $n_{y=0, \neg(n,m)}^w$ 为对应标识为 0 的词语中 w 出现的个数, $n_{z_{n,m}, \neg(n,m)}^w$ 为词语 $w_{n,m}$ 被赋予主题为 $z_{n,m}$ 的次数, 以上计数均排除了第 n 篇文档中的第 m 个词语.

经过多轮迭代至收敛后, 可以根据文档集中全部隐变量的赋值获取参数 $\theta, \varphi, \varphi^B, \sigma, \pi$ 的取值.

$\theta_{f,k}$ 表示情境主题对应的人群患有某种疾病的概率:

$$\theta_{f,k} = \frac{n_f^k + \alpha}{\sum_{j=1}^K (n_f^j + \alpha)}, \quad (5)$$

其中 n_f^j 为采样得到情境主题为 f 的文档对应疾病主题为 j 的次数.

$\varphi_{k,w}$ 表示某种疾病主题生成某个词的概率:

$$\varphi_{k,w} = \frac{n_k^w + \beta}{\sum_{v=1}^V (n_k^v + \beta)}, \quad (6)$$

其中 $\varphi_{k,w}$ 为疾病主题 k 对应的文档中单词 w 出现的次数.

φ_w^B 表示某种背景主题生成某个词的概率:

$$\varphi_w^B = \frac{n_B^w + \beta}{\sum_{v=1}^V (n_B^v + \beta)}, \quad (7)$$

其中 n_B^w 为单词 w 在背景主题中出现的次数.

$\sigma_{f,k,d}$ 表示某种人群在患有某种疾病时使用某种药品的概率:

$$\sigma_{f,k,d} = \frac{n_{f,k}^d + \eta}{\sum_{d=1}^D (n_{f,k}^d + \eta)}, \quad (8)$$

其中 $n_{f,k}^{d_i}$ 为采样得到情境主题为 f 且疾病主题为 k 的文档中药品 d_i 出现的次数.

π_y 表示词语是源自疾病主题还是背景主题的概率:

$$\pi_y = \frac{n^y + \lambda}{\sum_{i=0}^1 (n^i + \lambda)}, \quad (9)$$

其中 n^i 表示标识符 i 出现的次数.

3.5 药品推荐

假设给定一组测试数据集,其中患者对应药品未知.利用式(5)~(9)的结果可以直接计算出该患者对应所有药品的概率,而无需在测试集中进行采样,因此本方法的药品推荐过程可以非常快速的完成.

推荐过程的具体方法为:假设给定一条新病历(W,T),其中 $W=\{w_1,w_2,\cdots,w_p\}$ 为患者对应的病情文档, $T=\{t_1,t_2,\cdots,t_q\}$ 为患者对应的情境文档.首先通过情境文档处理步骤中训练得到的 LDA 模型获得 T 中的情境主题分布 ψ ,随后由式(10)便可获得该患者对应任意一种药品 d 的条件概率.

$$p(d \mid W, \psi) = \sum_{f=1}^F \sum_{k=1}^K \sigma_{f,k,d} \times (\psi_f \theta_{f,k} \prod_w (\pi_0 \phi_w^B + \pi_1 \phi_{k,d})). \quad (10)$$

可以发现,式(10)中括号内的乘积项 $M_{f,k} = \psi_f \theta_{f,k} \times \prod_w (\pi_0 \phi_w^B + \pi_1 \phi_{k,d})$ 与 $\sigma_{f,k,d}$ 互相独立,因此,在对所有药品进行概率计算前,可以预先计算好全部 $M_{f,k}$,计算药品 d 的概率:

$$p(d \mid W, \psi) = \sum_{f=1}^F \sum_{k=1}^K \sigma_{f,k,d} M_{f,k}. \quad (11)$$

获得全部药品概率后,按照概率从大到小进行排序即得到药品推荐的最终结果.

4 实 验

本节详细描述本文实验中所用到的数据集、度量指标、基准方法、参数设置以及实验结果分析.

4.1 实验数据集

本文实验所用数据来源于某大型三甲医院住院部 2015—2018 年的电子病历数据库,且经过脱敏处理.对数据集进行清洗、去重等处理,得到 158 556 条

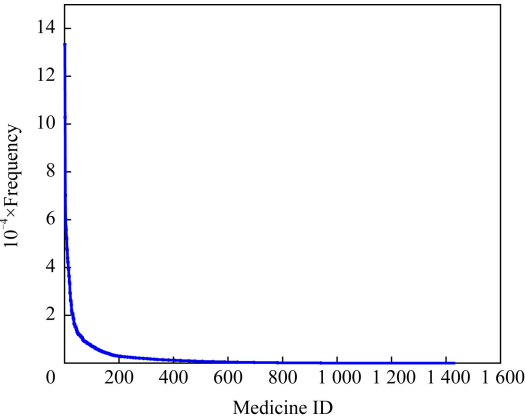


Fig. 3 Medicine Frequency
图 3 药品频率统计

完整电子病历记录,每条数据包括患者从出院到入院所进行的完整诊疗过程记录.通过 3.1 节所述预处理步骤得到最终所用数据集,其中检查异常种类共 1242 种,主诉与体格检查报告分词后得到共 41 309 种词.药品种类共 1 428 种,医生平均为每名患者开出 20 种药品.药品分布存在较为严重的长尾现象,即数据中存在大量低频药品,药品出现频率见图 3,其纵轴表示该药品在多少条记录中出现.图 4 给出了药品出现次数统计的对数形式,纵轴为图 3 对应点取值以 10 为底的对数.

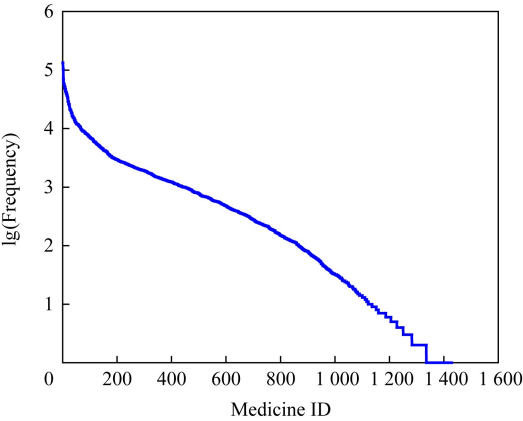


Fig. 4 Medicine Frequency with log
图 4 药品频率统计对数形式

4.2 评价指标

本文所研究的药品推荐问题本质是一个多标签学习(multi-label learning)问题.多标签学习问题的评价指标主要分为 2 类,分别是基于分类的评价指标和基于排序的评价指标^[14].药品推荐的目标是使得真实药品在推荐列表中尽可能排名靠前,为此,本文选用基于分类的评价指标 $Precision@K(P@K)$, $Recall@K(R@K)$, $F@K$ 和基于排序的指标 $NDCG@K$ 两个常用的评价指标评测模型的推荐性能. $Precision@K, Recall@K, F@K$ 的计算公式如下:

$$Precision@K = \frac{1}{N} \sum_{n=1}^N \frac{|\xi_n^T \cap \xi_n^K|}{|\xi_n^K|}, \quad (12)$$

$$Recall@K = \frac{1}{N} \sum_{n=1}^N \frac{|\xi_n^T \cap \xi_n^K|}{|\xi_n^T|}, \quad (13)$$

$$F@K = \frac{1}{N} \sum_{n=1}^N \frac{2 \times Precision_n \times Recall_n}{Precision_n + Recall_n}. \quad (14)$$

其中, N 为测试集总数,为 ξ_n^T 代表第 n 条数据的真实药品集合, ξ_n^K 代表系统对于第 n 条数据的前 K 个推荐药品. $Precision_n$ 和 $Recall_n$ 分别代表针对第 n 条数据计算出的精确率和召回率. $NDCG@K$ 的计算公式为:

$$NDCG@K = \frac{1}{N} \sum_{n=1}^N \frac{\sum_{i=1}^K \frac{rel_{n,i}}{\lg(i+1)}}{\sum_{i=1}^K \frac{1}{\lg(i+1)}}, \quad (15)$$

其中, $rel_{n,i}$ 表示第 n 条数据中的第 i 个药品是否在真实药品集合中, 若在则取 1, 不在则取 0. $|REL_n|$ 表示第 n 条数据真是药品集合大小.

4.3 基准方法

为了验证本文所提出的 MLDA 模型的效果, 本文选择使用以下 12 种基准方法进行对比实验, 包括主题模型、协同过滤、多标签学习、深度学习等方法, 以及本文所提出的模型的变种.

1) Frequency(Freq). 该方法统计训练数据中每种药品出现的次数, 由高到低进行排列. 在进行推荐时, 对于测试集中的每一条数据均给出此列表. 任何效果低于此方法的推荐方法均视为完全无效.

2) LLDA. Labeled-LDA (LLDA)^[12] 是对 LDA 模型在多标签语料库中的扩展, 其假设语料库中的标签与潜在主题一一对应, 在此将每种药品视为一种标签, 将情境文档与病情文档混合作为训练文档, 然后使用文献[12]中的方法进行训练与推荐.

3) TagLDA. TagLDA^[22] 假设每个主题同时拥有文档词分布与标签分布, 同时假设文档的标签是对文档的描述, 因而文档的主题分布应该与其标签的主题分布相同. 在进行推荐时, 通过给定的文档获得文档主题分布 $p(z | doc)$, 然后使用如式(16)计算全部标签 d 得到概率:

$$p(d | doc) = \sum_{z=1}^K p(d | z) p(z | doc). \quad (16)$$

4) KNN. 文献[23]提出了一种文档标签推荐方法, 是 K 近邻算法在多标签数据集中的扩展. 给定一个文档 doc , 此方法根据文档的内容相似度, 如 cosine 距离, 来找出与此文档最相似的 k 个文档. 这 k 个文档中标签出现的概率大小即为模型推荐的标签概率. 在本文的药品推荐任务中, 由于文档单词数过多, 直接计算 cosine 距离会导致计算量过大, 为此本文首先得到患者病情文档, 然后使用 LDA 模型得到病情文档的主题向量, 进而使用 cosine 距离寻找相似文档.

5) MLDT. 多标签决策树 (multi-label decision tree, MLDT)^[24] 是决策树算法 (decision tree) 在多标签学习问题上的扩展. 该方法将决策树算法中熵的计算公式扩展至多标签情形, 形式为

$$MLEnt(\tau) = - \sum_{i=1}^N ((p(c_i) \log p(c_i) + (q(c_i) \log (q(c_i)))), \quad (17)$$

其中, $p(c_i)$ 表示类别 c_i 的概率, $q(c_i) = 1 - p(c_i)$.

6) CTR. 协同过滤 (Collaborative Filtering, CF) 是推荐系统领域应用最为广泛的方法, 其通过相似用户的喜好为用户推荐可能感兴趣的物品. 协同过滤可以分为基于邻居的方法 (neighborhood methods) 和基于隐因子模型的方法 (latent factor models)^[22], 而后者往往可以取得更高的预测精度, 其中基于矩阵分解的隐因子模型^[25-27]是最为成功的一类模型.

然而, 此类方法会遇到冷启动 (cold start) 问题, 对于数据中未曾出现的用户或物品, 算法无法直接通过矩阵分解获得其向量表示. 为此, 文献[28]提出了一种结合主题模型与矩阵分解的协同过滤推荐算法 CTR (collaborative topic regression model). 该方法使用 Topic Model 对文本进行表征, 从而在冷启动环境下依然可以获得向量表示. 根据文献[28], 可以使用 LDA 模型获得每一名患者 j 对应文档的主题分布 θ_j , 进而使用公式 $r_{i,j} = (\mathbf{u}_i)^T \theta_j$ 计算药品 i 推荐给病人 j 的概率大小, 其中 $\mathbf{u}_i$ 为药品的表示向量. 为了得到药品的表示向量, CTR 使用下式对训练集进行建模:

$$r_{i,j} \sim N(\mathbf{u}_i^T \theta_j, c_{i,j}^{-1}). \quad (18)$$

其中, $r_{i,j}$ 取值为 0 或 1 表示药品 i 的在患者 j 的药品集中出现与否. $c_{i,j}$ 用于控制药品出现与否对于向量影响的重要程度, 形式为

$$c_{i,j} = \begin{cases} a, & r_{i,j} = 1, \\ b, & r_{i,j} = 0. \end{cases} \quad (19)$$

7) MLP. 多层感知机 (multilayer perceptron, MLP) 是神经网络的一类经典模型, 包括输入层、输出层和多个隐藏层. MLP 模型可以简单且高效地进行特征表示, 进而进行各类推荐任务. 在此将患者的病情文档和情境文档进行联合作为输入层的输入, 使用 Relu 函数作为激活函数, 训练其输出每种药品推荐的概率.

8) W&D. Wide & Deep (W&D)^[29] 是一种得到广泛应用的深度学习推荐模型, 其模型融合了浅层 (Wide) 和深层 (Deep) 神经网络, 利用浅层模型的记忆能力和深层模型的泛化能力, 实现单模型对精确性和扩展性的兼顾. 在此将患者的病情文档和情境文档进行联合作为模型浅层和深层部分共同的输入, 另外, 由于 W&D 模型是为点击率预测问题设计, 其输出单元维度为 1, 在此对其进行扩展使其输出多个维度.

9) NTM. Neural Topic Model (NTM)^[30-31] 是近年来新兴的一类融合主题模型与深度学习的方法,

在文本建模[30]、推荐系统领域[32]均取得了应用. NTM 采用变分推断对主题模型进行求解, 并使用深度学习技术对变分推断过程进行拟合^[30], 从而提高模型的训练速度与准确率. 本文采用文献[33]中的方法使用变分推断求解 MLDA 模型, 并使用神经网络进行训练. 值得注意的是, MLDA 模型中的情境主题模块无法简单融入 NTM 模型, 因此使用与 MLP 模型相同的联合文档作为输入.

此外, 我们通过对本文方法剔除部分模块以进行消融实验, 所涉及的对比方法包括:

10) MLDA-pure. 修改本文提出模型, 不考虑情境信息的影响, 仅根据病情文档进行推荐.

11) MLDA-z. 修改本文提出的模型, 不考虑情境主题 c 对于药品分布 f 的影响, 仅考虑 c 对于疾病主题 z 的影响.

12) MLDA-c. 修改本文提出的模型对于情境信

息的处理方法, 不对情境信息进行单独处理, 而是将原始的情境文档中的词直接加入患者的病情文档, 直接将其视为与化验异常、医疗文本词作用相同的用于描述患者的文本特征.

4.4 实验结果

表 3 给出了不同方法在数据集上的性能表现, 每个实验结果均为 5 折交叉验证所得. 其中 LLDA 模型中主题-词先验设为 0.01; Tag-LDA 中主题数设为 50; 主题-词先验设为 0.01, CF 中向量维度设为 30, 相似文档数设为 40; CTM 中向量维度设为 30, 参数 a 和参数 b 分别设为 1 和 0.5, 正则化参数设为 2; MLDT 中设置叶节点中最小包含样本数目为 30. 在本文提出的 MLDA 模型中, 设参数 $K = 400, F = 10, \alpha = 0.1, \beta = 0.1, \lambda = 0.1, \eta = 0.01$, 迭代次数为 30 轮. 所有模型参数均通过网格搜索法调至最优.

Table 3 Comparison of Different Methods

表 3 不同方法性能对比

Metric	Freq	LLDA	TagLDA	KNN	MLDT	CF	MLP	W&D	NTM	MLDA	MLDA-pure	MLDA-z	MLDA-c
<i>NDCG@20</i>	0.411 6	0.211 3	0.403 0	0.606 0	0.620 6	0.566 8	0.628 3	0.632 9	0.573 3	0.633 3	0.607 0	0.614 4	0.615 5
<i>NDCG@30</i>	0.449 2	0.243 2	0.441 2	0.667 5	0.679 0	0.624 8	0.689 0	0.695 1	0.633 5	0.696 1	0.667 9	0.675 7	0.677 1
<i>NDCG@40</i>	0.477 9	0.263 1	0.469 6	0.704 6	0.712 5	0.660 9	0.725 0	0.733 3	0.671 3	0.733 4	0.705 7	0.713 5	0.714 9
<i>P@20</i>	0.313 7	0.152 4	0.308 7	0.495 6	0.508 4	0.463 0	0.511 0	0.518 0	0.470 3	0.518 2	0.492 5	0.502 2	0.510 3
<i>P@30</i>	0.248 4	0.125 2	0.239 6	0.404 9	0.412 1	0.374 5	0.414 7	0.420 2	0.385 6	0.421 5	0.400 6	0.407 9	0.415 6
<i>P@40</i>	0.207 5	0.103 2	0.198 2	0.342 6	0.346 9	0.314 4	0.348 1	0.352 5	0.327 3	0.354 5	0.336 9	0.342 6	0.348 6
<i>R@20</i>	0.337 4	0.166 1	0.330 2	0.552 9	0.567 0	0.511 3	0.577 7	0.581 0	0.519 5	0.582 9	0.553 9	0.561 6	0.562 4
<i>R@30</i>	0.396 2	0.221 3	0.394 5	0.656 4	0.663 5	0.608 6	0.678 8	0.687 8	0.621 0	0.687 8	0.656 0	0.664 7	0.665 4
<i>R@40</i>	0.452 1	0.258 7	0.446 1	0.724 7	0.723 9	0.675 7	0.744 4	0.753 7	0.690 5	0.755 9	0.725 9	0.734 1	0.735 4
<i>F@20</i>	0.297 8	0.143 2	0.283 5	0.481 2	0.493 3	0.445 2	0.498 2	0.504 3	0.454 8	0.505 0	0.479 7	0.489 3	0.491 0
<i>F@30</i>	0.283 7	0.139 8	0.278 6	0.463 7	0.471 0	0.432 0	0.476 0	0.482 4	0.441 0	0.484 0	0.459 5	0.468 0	0.472 3
<i>F@40</i>	0.266 3	0.127 6	0.259 6	0.433 4	0.437 4	0.400 5	0.441 2	0.446 8	0.414 1	0.449 4	0.426 8	0.434 1	0.439 8

Note: Our results are in bold and they are better than other methods.

通过对表 3 的观察可以得出 5 个结论:

1) 基于传统主题模型的推荐方法 LLDA 和 Tag-LDA 的效果均未超过 Frequency, 即可视为完全无效果. 这是由于数据的特点导致, 患者的病情文档是病人所患疾病的反映, 其内容往往只围绕某一种或几种疾病. 而传统主题模型假设每篇文档均有较多的主题, 此种假设在文本较长、文本内容较丰富的语料库中效果较好, 而面对文本较短、文本含义较为集中的医疗文本效果不佳.

2) 基于协同过滤方法的 CTR 模型效果较差, 实验效果被传统机器学习模型 ML-KNN 和 ML-DT 所超越. 这是由于药品推荐问题中没有“用户”这

一概念, 每一名患者都需要重新进行诊断, 即文献[25]中所述“out-of-matrix prediction”. 协同过滤方法在此条件下会面临冷启动问题, 从而影响预测的准确率. 文献[25]中的实验部分也证实了这一点.

3) 基于深度学习的推荐系统 MLP、W&D 模型在准确率方面表现较好, 效果与 MLDA 模型十分接近. 然而药品推荐问题对于模型的可解释性要求较高, 医生与患者往往希望得知某种药品得到推荐的理由. 深度学习方法采用的端到端训练模式, 难以提供较好的可解释性. 相较而言, 本文提出的 MLDA 模型采用主题模型进行建模, 在可解释性方面较为优秀, 同时能够取得良好的推荐效果, 因此更适用于

药品推荐领域.

4) 基于深度学习的 NTM 模型相比 MLDA 模型效果有一定差距,这是因为其采用了变分推断的求解方式,相对于 Gibbs 采样在求解精度方面有一定差距.同时,NTM 模型无法直接对情境信息进行充分利用,导致其推荐效果有所下降.

5) 从消融实验的对比来看,由于同时考虑了情境对于疾病和药品的影响,MLDA 在全部评测指标中均取得了最好的效果.相较而言,在 MLDA 模型的 3 类变种中,MLDA- z 模型只考虑情境对于疾病的影响,实验结果表明 MLDA 模型效果优于 MLDA- z .同时,MLDA-pure 模型不考虑任何情境的影响,其效果低于 MLDA 和 MLDA- z .这充分证明了本文所提出模型对于情境信息利用的有效性.

此外,值得注意的是,增加情境对于疾病的影响带来的收益,其效果要低于增加情境对于药品的影响所带来的收益,即情境对于药品的影响要大于对于疾病的影响.这是由于本文所用数据集的特点所导致.本文所用数据来源于大型三甲医院,此类院所接诊的病人中有相当一部分患者身患较严重疾病,如严重的外伤、各类肿瘤、传染病等.此类疾病的

发病与否受患者情境信息的影响相对较小.相对的,在此类大型医疗机构中,医生会更加注意患者的性别、年龄、医保、手术等情境信息对于患者所使用药品的作用效果.因此,情境信息对于药品的作用效果更加显著.

6) MLDA- c 模型相比于 MLDA-pure 模型增加了对情境信息的利用,但其提升效果不是十分明显,相距 MLDA 模型差距较大,这是由于情境信息与患者检查检验与医疗文本词相比而言数量较少,若直接将其与病情文档混合作为描述患者的文本特征,则情境信息对于结果的影响力会被大大减弱,从而导致模型难以取得良好的效果.

下面对本文提出的模型进行参数敏感性分析.图 5(a)展示了情境主题数量对于模型效果的影响,其中 $K=400$,调节不同的 F 值大小.图 5(b)展示了疾病主题数量对于模型效果的影响.其中 $F=10$,调整不同的 K 值大小.观察可见,模型在 $K=400$, $F=10$ 处取得最好效果,并且模型对于参数变化展示出了良好的鲁棒性,参数在较大范围内变化时,模型效果仅在 0.01 量级的范围内波动,这证实了本文提出方法的良好性能.

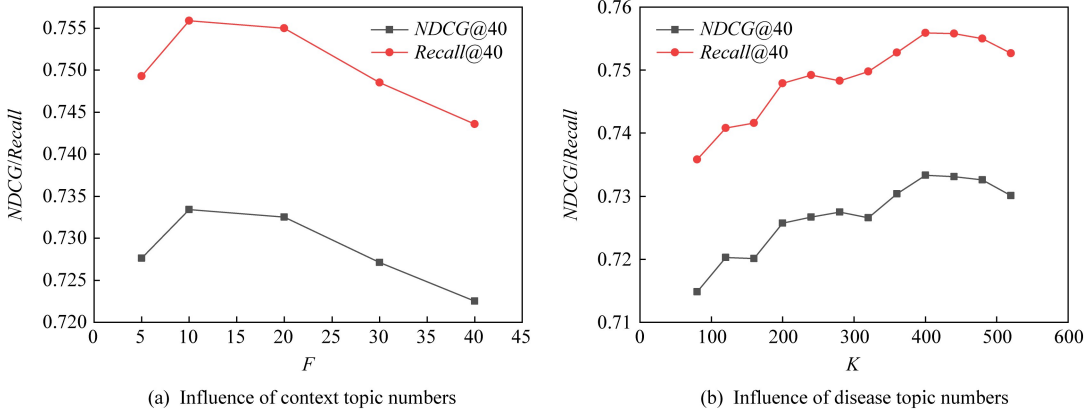


Fig. 5 Influence of context topic numbers and disease topic numbers to results

图 5 情境主题与疾病主题数量对于模型效果的影响

为了进一步研究所提出方法的优越性,我们对模型挖掘所得内容进行分析.具体而言,在本文所提出的模型中,疾病主题表示了一种或一类疾病,而情境主题则表示一类患者,同时疾病主题对应的文档词分布表示了某种疾病对应的化验异常与症状.理想的模型可以较为准确的挖掘得到此类信息,从而为模型的推荐结果提供有力的支持,提高模型的可解释性.

表 4 给出了部分疾病主题的挖掘结果,其中高频词语部分展示了该疾病主题对应的词分布中频率

由高到低排列的前几个词语,并通过对高频词语的观察给出了主题对应的含义.通过对高频词语的观察与分析,可以发现模型挖掘得到的疾病主题具有明显的医学含义,且化验结果和医疗文本均对疾病主题的推断起到了重要的作用.

下面分析不同情境对于疾病和药品的影响程度.定义情境主题 c 对于疾病的影响度 $S_{c,z}$:

$$S_{c,z} = dis(v_{c,z}, \frac{1}{F} \sum_{f=1}^F v_{f,z}). \tag{20}$$

其中, $v_{c,z}$ 代表情境主题 c 对应的疾病主题分布向量,

Table 4 Results of Disease Topic Mining

表 4 疾病主题挖掘结果

Disease Topic	High Frequency Words	Meaning
# 36	creatinine-high, urea nitrogen-high, total number of erythrocytes-low, hematocrit-low, Urine occult bloos-positive	Kidney disease
# 108	hearing loss, tympanic membrane, fibrinogen ca7000-low, external auditory canal, tinnitus, plate volume distribution width-low	Ear disease
# 150	alc of glycosylated hemoglobin-high, D of 25 hydroxyvitamin-low, thyroid, artery, mean blood sugar-high, dorsum of foot	Diabetes mellitus and its complications
# 161	limited activity, lymphocyte percentage-low, endings, blood supply, trauma, sensation, tenderness	Traumatic diseases

dis 代表 2 个向量的欧氏距离,疾病影响度反映了不同情境主题对应的易患疾病相对于平均人群的偏差度,疾病影响度越大,代表该情境主题对于患者易患疾病的影响越大.定义情境主题 c 对于药品的影响度 $S_{c,f}$:

$$S_{c,f} = \frac{1}{K} \sum_{k=1}^K dis\left(v_{c,k}, \frac{1}{F} \sum_{f=1}^F v_{f,k}\right), \quad (21)$$

其中, $v_{c,k}$ 代表情境主题 c 与疾病主题 k 对应的药品分布向量, dis 代表 2 个向量的欧氏距离.药品影响度反映了相同疾病下情境信息对于药品分布的影响.药品影响度越大,代表该情境主题对于患者的药品使用的影响越大.

疾病影响度最大的 2 个情境主题如表 5 所示.可以发现,天气、气温、季节等因素是影响患者易患疾病的主要因素,年龄的不同也起到了一定的影响作用.药品影响度最大的 2 个情境主题如表 6 所示.可以发现,性别、年龄、保险情况、是否手术与麻醉等因素是影响医生对于患者药品使用的主要因素.以上结果有力地证实了本文在 4.3 节中提出的假设.

Table 5 Two Context Topics With the Greatest Disease Impact

表 5 疾病影响度最大的 2 个情境主题

Context Topic	Main Context Word
# 8	weather-cloudy, temperature difference-less than 10 degree centigrade, age-middle age
# 7	average temperature-0~10, season-winter, weather-others

Table 6 Two Context Topics With the Greatest Medicine Impact

表 6 药品影响度最大的 2 个情境主题

Context Topic	Main Context Word
# 6	age-youth, gender-female, insurance-self-paying
# 0	Weather-rainy, anesthesia-no, operation-no

最后对某一情境主题对于药品分布的影响进行

具体分析.定义药品 d 在情境主题 c 下的提升度 $I_{c,d}$:

$$I_{c,d} = rank_{all,d} - rank_{c,d}, \quad (22)$$

其中, $rank_{all,d}$ 代表药品 d 在全部数据中的出现频率排名, $rank_{c,d}$ 代表药品 d 在情境主题 c 对应的数据中出现频率的排名.在此,以情境主题 # 9 为例进行展示,表 7 给出了情境主题 # 9 的主要情境词,表 8 给出了情境主题 # 9 对应的提升度最大的几种药品.可以看到,情境主题 # 9 代表的人群主要为进行了手术的患者,并且进行了全身麻醉,代表患者可能身患重病.表 8 给出了情境主题 # 9 对应的提升度最大的几种药品及其功效,可见各类治疗肿瘤、缓解化疗副作用、消炎药物等提升幅度最大,这与情境主题 # 9 代表的重病类患者的含义一致.

Table 7 Main Context Word of Context Topic # 9

表 7 情境主题 # 9 的主要情境词

Context Topic	Main Context Word
# 9	operation-yes, anesthesia-general, weather-others

Table 8 Medicines With the Greatest Promotion Degree and Their Efficacy

表 8 提升度最大的 5 种药品及其功效

Medicine	Efficacy
Theohydramin	relieves dizziness and nausea
Shenqishiyiwei Granules	promote the growth of leukocytes and alleviate the side effects of chemotherapy
Diyushengbai Tablet	promoting leukocyte growth
Recombinant HumanInterleukin-2 Injection	biotherapy of tumor
Prednisone	antiphlogistic

以上案例分析的内容充分证明了本文提出的模型不仅能够给出准确的推荐结果,还能够从数据中挖掘得到大量有意义的信息和规律,证实了模型的实际应用价值和指导意义.

5 总结与展望

本文提出了一种基于情境感知的药品推荐模型 Medicine-LDA,模型使用一个通用的概率框架来描述真实生活场景中的诊疗过程,具有良好的可解释性.该模型依据患者的化验异常与医疗文本组成的患者文档来推断患者可能患有的疾病,并提出使用疾病主题来描述患者患有不同疾病的概率.同时,该模型考虑到药品推荐的结果会受到情境信息的影响且作用方式不同,进行了有针对性的建模.另外,对于多源情境信息导致组合爆炸的问题,本文提出使用 LDA 模型对情境信息进行建模得到患者的情境主题,并将情境主题融入本文提出的概率模型当中,从而有效地对情境信息的作用进行建模.最后,通过在一个来自某大型三甲医院的数据集上的实验,表明本文提出的模型在准确性上的表现优于常用的传统机器学习模型、主题模型以及协同过滤模型,同时,针对模型学习所得结果进行分析可以得到一系列有价值的信息,证实了模型的良好性能与其实在实际应用中的价值.

本文所提出的方法也有其局限性.由于采用了主题模型,当某些疾病的症状较为相似,且从情境信息难以得到区分时,可能会导致 2 种疾病的混淆.为此,未来我们将对基准方法中的 NTM 模型进行进一步的研究,寻找更佳的情境融合与文本建模方法,从而将主题模型的可解释性与深度学习对于特征的学习能力进行结合,进一步增强模型对于此类问题的处理能力.

参 考 文 献

[1] López-Nores M, Blanco-Fernández Y, Pazos-Arias J J, et al. Property-based collaborative filtering for health-aware recommender systems [J]. *Expert Systems with Applications*, 2012, 39(8): 7451-7457

[2] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis [J]. *Medical Image Analysis*, 2017, 42: 60-88

[3] García-Lizana F, Sarria-Santamera A. New technologies for chronic disease management and control: A systematic review [J]. *Journal of Telemedicine and Telecare*, 2007, 13(2): 62-68

[4] Xiao Xiaoli, Zhou Xiling. Medicine recommendation algorithm based on user similarity and trust [J]. *Computer Engineering and Applications*, 2018, 54(8): 214-219 (in Chinese)

(肖晓丽, 周锡玲. 基于用户相似度和信任度的药品推荐算法 [J]. *计算机工程与应用*, 2018, 54(8): 214-219)

[5] Zhou Xiling, Zhang Limin, Tian Xiaolu, et al. Medicine recommendation algorithm integrating demographic attribute [J]. *Electronic Test*, 2018, 402(21): 70-71 (in Chinese)

(周锡玲, 张莉敏, 田小路, 等. 融合人口统计属性的药品推荐算法 [J]. *电子测试*, 2018, 402(21): 70-71)

[6] Zhang Yin, Zhang Daqiang, Hassan M M, et al. CADRE: Cloud-assisted drug recommendation service for online pharmacies [J]. *Mobile Networks and Applications*, 2015, 20(3): 348-355

[7] Suchman L A. Plans and situated actions: The problem of human-machine communication [M]. Cambridge, UK: Cambridge University Press, 1987

[8] Chen Enhong, Xu Tong, Tian Jilei, et al. Personalized recommendation via mobile context-aware mining [J]. *Communication of China Computer Federation*, 2013, 9(3): 18-24 (in Chinese)

(陈恩红, 徐童, 田继雷, 等. 移动情境感知的个性化推荐技术 [J]. *中国计算机学会通讯*, 2013, 9(3): 18-24)

[9] Adomavicius G, Tuzhilin A. Recommender Systems Handbook [M]. Berlin: Springer, 2011: 217-253

[10] Oku K, Nakajima S, Miyazaki J, et al. Context-aware SVM for context-dependent information recommendation [C] // *Proc of the 7th Int Conf on Mobile Data Management (MDM'06)*. Piscataway, NJ: IEEE, 2006: 109-109

[11] Shi Yue, Karatzoglou A, Baltrunas L, et al. TFMAP: Optimizing MAP for top-*n* context-aware recommendation [C] // *Proc of the 35th Int ACM SIGIR Conf on Research and Development in Information Retrieval*. New York: ACM, 2012: 155-164

[12] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. *Journal of Machine Learning Research*, 2003, 3(1): 993-1022

[13] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora [C] // *Proc of the 2009 Conf on Empirical Methods in Natural Language Proc*. New York: ACM, 2009: 248-256

[14] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents [J]. *arXiv preprint arXiv:1207.4169*, 2012

[15] Wang Yuan, Qu Jishi, Liu Jie, et al. What to tag your microblog: Hashtag recommendation based on topic analysis and collaborative filtering [C] // *Proc of the Asia-Pacific Web Conf*. Berlin: Springer, 2014: 610-618

[16] She Jieying, Chen Ler. Tomoha: Topic model-based hashtag recommendation on twitter [C] // *Proc of the 23rd Int Conf on World Wide Web*. New York: ACM, 2014: 371-372

[17] Zhu Chen, Zhu Hengshu, Ge Yong, et al. Tracking the evolution of social emotions with topic models [J]. *Knowledge and Information Systems*, 2016, 47(3): 517-544

- [18] Xu Tong, Zhu Hengshu, Zhu Chen, et al. Measuring the popularity of job skills in recruitment market: A multi-criteria approach [C] // Proc of the 32nd AAAI Conf on Artificial Intelligence (AAAI'18). Menlo Park, CA: AAAI, 2018; 2572-2579
- [19] Zhao Wanxin, Jiang Jing, Weng Jianshu, et al. Comparing twitter and traditional media using topic models [C] //Proc of the European Conf on Information Retrieval. Berlin: Springer, 2011; 338-349
- [20] Yan Xiaohui, Guo Jiafeng, Lan Yanyan, et al. A biterm topic model for short texts [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013; 1445-1456
- [21] Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008; 569-577
- [22] Si Xiance, Sun Maosong. Tag-LDA for scalable real-time tag recommendation [J]. Journal of Information & Computational Science, 2009, 6(2): 1009-1016
- [23] Kywe S M, Hoang T A, Lim E P, et al. On recommending hashtags in twitter networks [C] //Proc of Int Conf on Social Informatics. Berlin: Springer, 2012; 337-350
- [24] Clare A, King R D. Knowledge discovery in multi-label phenotype data [C] //Proc of European Conf on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 2001; 42-53
- [25] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42(8): 30-37
- [26] Mnih A, Salakhutdinov R R. Probabilistic matrix factorization [C] //Advances in Neural Information Proc Systems. New York: Curran Associates, 2008; 1257-1264
- [27] Hu Yifan, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets [C] //Proc of the 8th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2008; 263-272
- [28] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011; 448-456
- [29] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems [C] //Proc of the 1st Workshop on Deep Learning for Recommender Systems. New York: ACM, 2016; 7-10
- [30] Miao Yishu, Yu L, Blunsom P. Neural variational inference for text processing [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016; 1727-1736
- [31] Miao Yishu, Grefenstette E, Blunsom P. Discovering discrete latent topics with neural variational inference [C] // Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017; 2410-2419

- [32] Bai Haoli, Chen Zhuangbin, Lyu M R, et al. Neural relational topic models for scientific article analysis [C] // Proc of the 27th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2018; 27-36
- [33] Kingma D P, Welling M. Auto-encoding variational bayes [J]. arXiv preprint arXiv:1312.6114, 2013



Zheng Zhi, born in 1997. MSc candidate. Student member of CCF. His main research interests include recommender system and data mining.



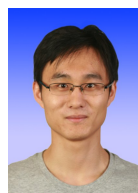
Xu Tong, born in 1988. PhD, associate professor. Member of CCF. His main research interests include data mining and social network.



Qin Chuan, born in 1993. PhD. His main research interests include natural language processing and recommender system.



Liao Xiangwen, born in 1980. PhD, professor. Senior member of CCF. His main research interests include natural language processing, information retrieval and social computing.



Zheng Yi, born in 1987. PhD. His main research interests include natural language processing and machine learning.



Liu Tongzhu, born in 1967. PhD, associate researcher. His main research interests include healthcare big data and hospital management.



Tong Guixian, born in 1991. Master, junior economist. His main research interests include healthcare big data and hospital management.