软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(11):3295−3305 [doi: 10.13328/j.cnki.jos.005481]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于优化主题模型的临床路径挖掘*

徐 啸, 金 涛, 王建民

(清华大学 软件学院,北京 100084)

通讯作者: 金涛, E-mail: jintao05@gmail.com

摘 要: 在健康领域,诊疗过程对于医疗质量至关重要.临床路径集合了各种医疗知识,是对诊疗过程进行标准化的重要途径.然而,当前大多数临床路径由专家研讨制定,往往静态不变,难以部署和实施.在之前的工作中,提出了一种基于主题的临床路径挖掘算法,可以从医疗数据中抽取历史执行路径,客观反映数据中实际存在的医疗模式.算法首先通过主题模型将繁杂的诊疗活动聚合成若干主题,而每个诊疗日就可以表示为一个主题分布,一个病人的诊疗日志也相应的转换为一个主题序列,然后利用过程挖掘方法从这些主题序列中生成基于主题的临床路径模型.但传统主题模型(LDA)的聚类效果往往难以满足医疗数据的特点,导致主题质量不高,影响最终过程模型的可解释性.其中,一个普遍的问题就是LDA无法保证两个相似的诊疗日所得的主题分布也是相似的,这是由于其忽略了诊疗日之间原有的相似性特征.提出了一种优化的主题模型算法,该算法引入了基于本体生成的诊疗日相似性约束,可以有效地提升聚类效果.实验结果表明,提出的方法能够发现更符合医疗领域特点的高质量主题,进而为基于主题的临床路径的挖掘奠定基础.

关键词: 临床路径挖掘;主题模型;过程挖掘

中图法分类号: TP311

中文引用格式: 徐啸,金涛,王建民.基于优化主题模型的临床路径挖掘.软件学报,2018,29(11):3295−3305. http://www.jos.org.cn/1000-9825/5481.htm

英文引用格式: Xu X, Jin T, Wang JM. Optimized topic model for clinical pathway mining. Ruan Jian Xue Bao/Journal of Software, 2018,29(11):3295−3305 (in Chinese). http://www.jos.org.cn/1000-9825/5481.htm

Optimized Topic Model for Clinical Pathway Mining

XU Xiao, JIN Tao, WANG Jian-Min

(School of Software, Tsinghua University, Beijing 100084, China)

Abstract: In healthcare domain, the care process is critical for the care quality. Clinical pathway (CP), which integrates a lot of medical knowledge, is a tool for standardizing the care process. However, most of existing CPs are designed by experts with limited experience and data, and consequently they are always static and non-adaptive for implementation. According to authors' previous work, topic-based CP mining is an effective approach which can discover the process model from clinical data. The various clinical activities are summarized into several topics by latent dirichlet allocation (LDA), and each clinical day in the patient trace is converted to a topic distribution. A CP model can be derived by applying process mining method on the topic-based sequences. However, LDA ignores the similarity between clinical days, which means that in some cases, two similar days may be assigned quite different topic distributions. This paper proposes an optimized topic model for clinical topic discovering by incorporating the similarity constraint, which is based on

* 基金项目: 国家自然科学基金(61325008); 国家科技支撑计划(2015BAH14F02)

Foundation item: National Natural Science Foundation of China (61325008); National Key Technology R&D Program of China (2015BAH14F02)

本文由面向智能制造的业务过程管理与服务技术专题特约编辑王建民教授、刘建勋教授推荐.

收稿时间: 2017-07-20; 修改时间: 2017-09-16; 采用时间: 2017-11-14; jos 在线出版时间: 2017-12-06

CNKI 网络优先出版: 2017-12-06 15:37:41, http://kns.cnki.net/kcms/detail/11.2560.TP.20171206.1537.030.html

the domain knowledge. Experiments on real data demonstrate that this new approach can discover quality topics which are useful for topic-based CP mining.

Key words: clinical pathway mining; topic model; process mining

临床路径(clinical pathway)是针对特定病种的标准化诊疗计划,它将一次诊疗划分为若干阶段,并规定了每个阶段所用项目.临床路径已成为一项重要的医疗管理工具,广泛用于提升预期疗效、控制医疗成本、减少资源浪费.在中国,国家卫计委先后发布并推广了超过1 000种临床路径表单(如图1所示).然而,其中大多数都是由专家研讨制定的,从循证医学角度出发,它们在一定程度上缺乏临床数据的支持.

二、脑出血临床路径表单

适用对象：**第一诊断为**脑出血（ICD-10：I61）

患者姓名：_____ 性别：__ 年龄：__ 门诊号：____ 住院号：_____

住院日期：__年__月__日 出院日期：__年__月__日 标准住院日：8-14天

时间	住院第1天（急诊室到病房或直接到卒中单元）	住院第2天	住院第3天
主要诊疗工作	□ 询问病史与体格检查(包括NIHSS评分、GCS评分及Bathel评分) □ 完善病历 □ 医患沟通，交待病情 □ 监测并管理血压（必要时降压） □ 气道管理：防治误吸，必要时经鼻插管及机械通气 □ 控制体温，可考虑低温治疗、冰帽、冰毯 □ 防治感染、应激性溃疡等并发症 □ 合理使用脱水药物 □ 早期脑疝积极考虑手术治疗 □ 记录会诊意见	□ 主治医师查房，书写上级医师查房记录 □ 评价神经功能状态 □ 评估辅助检查结果 □ 继续防治并发症 □ 必要时多科会诊 □ 开始康复治疗 □ 需手术者转神经外科 □ 记录会诊意见	□ 主任医师查房，书写上级医师查房记录 □ 评价神经功能状态 □ 继续防治并发症 □ 必要时会诊 □ 康复治疗 □ 需手术者转神经外科
重点医嘱	**长期医嘱：** □ 神经内科疾病护理常规 □ 一级护理 □ 低盐低脂饮食 □ 安静卧床 □ 监测生命体征 □ 依据病情下达 **临时医嘱：** □ 血常规、尿常规、大便常规 □ 肝肾功能、电解质、血糖、血脂、心肌酶谱、凝血功能、血气分析、感染性疾病筛查 □ 头颅CT、胸片、心电图	**长期医嘱：** □ 神经内科疾病护理常规 □ 一级护理 □ 低盐低脂饮食 □ 安静卧床 □ 监测生命体征 □ 基础疾病用药 □ 依据病情下达 **临时医嘱：** □ 复查异常化验 □ 复查头CT（必要时）	**长期医嘱：** □ 神经内科疾病护理常规 □ 一级护理 □ 低盐低脂饮食 □ 安静卧床 □ 监测生命体征 □ 基础疾病用药 □ 依据病情下达 **临时医嘱：** □ 异常化验复查 □ 依据病情需要下达

Fig.1 National clinical pathway of intracerebral hemorrhage released by Ministry of Health of China (partial)

图1 国家卫计委发布的关于脑出血的临床路径表单(部分)

随着医疗信息化的迅猛发展,各地积累了大量的医疗数据,因此,由数据驱动的临床路径挖掘方法得到了越来越多的关注.相对于专家制定的临床路径,这类从数据中发现的历史执行路径更加客观、具体,可以有效地辅助临床路径的设计/再设计和异常病历的检测.过程挖掘算法[1]是其中重要的一类,但由于医疗数据的多样性和复杂性,这类方法往往容易产生意大利面状的结果模型,难以被理解.为了提升可解释性,一些研究者尝试使用主题模型(如 LDA[2])从数据中发现常见的诊疗模式,但这类方法忽略了数据中的时序特征,只能作为临床路径模型的参考.

在我们之前的工作中[3],我们将主题模型和过程挖掘相结合,生成了基于主题的临床路径模型,具有良好的简洁性和可解释性.这一工作基于以下两个医疗实践.

(1) 每天的诊疗活动都围绕着若干诊疗目标(主题)制定;

(2) 每个诊疗目标(主题)对应于一组诊疗活动.

基于上述两点,我们将每个诊疗日作为 LDA 中的一篇文档,而诊疗日中每一个诊疗活动作为 LDA 中的一个词语,通过 LDA 的计算得到一组主题(视作诊疗目标).由此,原本由繁杂诊疗活动表示的每个诊疗日可被表示

为一个关于主题的分布,每个病人的日志也相应的转换为一个主题序列.通过过程挖掘,可以得到一个易理解的临床路径模型.

然而,原始的 LDA 假定文档之间是完全独立的,忽略了文档之间可能存在的关联性.在我们的应用中,这就使得两个相似的诊疗日可能会被表示为两组极其不同的主题分布,这与医疗场景的需求是相违背的,进而影响最终临床路径模型的质量.

在本文中,我们提出了一种优化的主题模型算法用于临床路径挖掘,该算法通过引入诊疗日之间的相似性,使得两个相似诊疗日的主题分布也尽可能相似.算法的整体流程如图 2 所示.首先,我们利用医疗本体知识检测出相似的诊疗日,并形成相似性约束;然后,将这些约束转换并融合到 LDA 的目标函数中;最终,通过一个最大化期望(expectation maximization)算法,逐步优化目标函数,得到一个既吻合数据特征,又满足相似性约束条件的聚类结果.完成主题建模后,再利用过程挖掘算法,就可以从所得的主题序列中生成一个易理解的临床路径模型.一系列实验结果表明,该算法相对于原始 LDA,可以更好地对医疗数据进行主题建模,发现高质量主题,用于临床路径的挖掘.

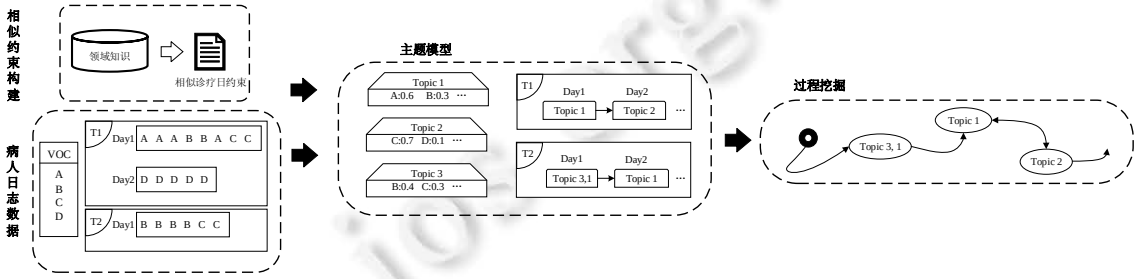

Fig.2 Core process of our approach

图 2 算法核心流程

1 相关工作

早在 2001 年,Lin 等人[4]指出,专家设计的临床路径很难处理医疗领域的个体差异性.为此,他们提出了一套图挖掘算法来抽取脑卒中病人的时序依赖模式,该方法可以用于预测新病人的诊疗路径.过程挖掘可以从日志数据中发现过程模型,因此被广泛用于临床路径的挖掘.Mans 等人[5]采用 ProM[6]中的启发式算法(heuristic miner)[7]对不同医院的中风病人进行挖掘,试图找出普遍的治疗方案.但医疗数据的多样性和复杂性,导致算法最终生成的是意大利面状的、难以理解的过程模型.在改用模糊挖掘(fuzzy miner)[8]方法后,得到了相对较为简洁的结果模型.Poelmans 等人[9]对比了隐马尔可夫模型和形式概念分析在 148 个乳腺癌患者数据的挖掘效果,实验结果表明,前者所得模型更易理解,可用于异常过程检测.然而,正如综述文献[10]中所总结的,由于医疗数据的特殊性,传统的过程挖掘算法很难生成高质量临床路径模型.

针对医疗数据低粒度的数据特点,Huang 等人[11]提出了基于条件概率的诊疗事件打包算法,可以大幅度减少过程挖掘的输入,简化生成的过程模型.近年来,一些研究者尝试使用主题模型对医疗数据进行诊疗模式的抽取,以辅助临床路径的挖掘.在文献[12]中,Huang 等人将一个病人的日志和一项诊疗活动类比于 LDA 中的一篇文档和一个词语,从而发现隐藏主题,作为疾病的诊疗模式.之后,他们在文献[13–15]中分别引入了时间戳、检验检查数据、并发并存症来扩展所发现的诊疗模式的范围和内容.然而,这一方法忽略了临床路径的阶段性特征,无法抓住核心的时序关系,难以作为临床路径的模型进行使用.

在我们之前的工作中,我们提出了基于主题的临床路径挖掘方法[3].本文在此基础上发现了一个制约主题模型效果的关键问题,即 LDA 忽略了诊疗日之间的相似关系,该问题会导致相似的诊疗日有时无法分配到相似的主题分布,进而影响抽取的主题质量,降低挖掘得到的最终临床路径模型的可解释性.为了解决这一问题,我们提出一个优化的主题模型,将基于本体生成的相似诊疗日约束引入其中,提升主题抽取效果.

2 预备知识

2.1 基本定义

本小节给出涉及的相关概念定义.

定义 1(诊疗活动(clinical activity)). 一个诊疗活动是发生在某个时间点的一个事件,是诊疗过程中的最小单元.诊疗活动名称的集合,称为诊疗项目(clinical item).

定义 2(诊疗日(clinical day)). 一个诊疗日指的是一个病人在某一天发生的所有诊疗活动的总和,诊疗日内活动不分时间次序.

定义 3(病人日志(patient trace)). 一个病人日志由若干诊疗日构成.

2.2 基本定义

LDA 是最常用的主题模型之一,可以从大量数据中发现隐含主题.LDA 定义了每篇文档中每个词的生成过程,包含两个重要参数:文档-主题分布、主题-词汇分布.在我们之前的工作中[3],我们将诊疗日和诊疗活动分别对应于 LDA 中的文档和词语概念,从而抽取隐含主题,作为诊疗目的,以吻合前述的两个医疗实践.具体的生成过程如下所述(相关符号注解见表 1).

1. 根据 Dirichlet 先验分布,生成每个主题-诊疗项目分布 $\phi_k \sim Dir(\beta), k=1,2,\ldots,K$;
2. 对于第 d 诊疗日,$d=1,2,\ldots,D$:
   a) 根据 Dirichlet 先验分布,生成诊疗日-主题分布 $\theta_d \sim Dir(\alpha)$;
   b) 对于第 d 诊疗日中第 i 个诊疗活动,$i=1,2,\ldots,N_d$:
      i 根据诊疗日-主题分布生成主题 $z_{d,i} \sim Multi(\theta_d)$;
      ii 根据主题-诊疗项目分布生成诊疗活动 $a_{d,i} \sim Multi(\phi_{z_{d,i}})$.

Table 1 Meanings of the notations

表 1 相关符号注解

符号	含义	符号	含义
$\mathbb{Q}$	所有诊疗日集合	$z_{d,i}$	$a_{d,i}$ 的主题
A	所有诊疗活动集合	$N_d^{(k)}$	第 d 诊疗日中主题为主题为 k 的诊疗活动数量
Z	所有诊疗活动所分配的主题	$N_k^{(v)}$	主题为 k 的诊疗项目 v 的数量
D,K	诊疗日和主题数量	α,β	Dirichlet 先验分布
V	诊疗项目(不同的诊疗活动名称)的数量	ϕ	主题-诊疗活动的多项式分布集合
N_d	第 d 诊疗日的诊疗活动数量	θ	诊疗日-主题的多项式分布集合
N_k	主题为 k 的诊疗活动数量	ϕ_k	主题 k-诊疗日活动多项式分布
$a_{d,i}$	第 d 诊疗日中第 i 个诊疗活动	θ_d	诊疗日 k-主题的多项式分布

通常,两组分布 θ_d 和 ϕ_k 可以利用 Gibbs 采样求解,其依赖于隐含主题与观测到的诊疗活动的联合分布:

$$P(Z,A|\alpha,\beta)=P(Z|\alpha)\cdot P(A|Z,\beta)=\int P(Z|\theta)P(\theta|\alpha)\mathrm{d}\theta\cdot\int P(A|Z,\phi)\cdot P(\phi|\beta)\mathrm{d}\phi \tag{1}$$

给定联合分布,对于每个诊疗活动的主题分配可以由式(2)计算得到.

$$P(z_{d,i}=k|Z_{-z_{d,i}},A)=\frac{P(Z,A)}{P(Z_{-z_{d,i}},A)}\propto\frac{N_{k,-z_{d,i}}^{(v)}+\beta_v}{\sum_{v=1}^{V}N_{k,-z_{d,i}}^{(v)}+\beta_v}\cdot(N_{d,-z_{d,i}}^{(k)}+\alpha_k) \tag{2}$$

采样结束后,ϕ和θ的计算如下:

$$\phi_k^{(v)}=\frac{N_k^{(v)}+\beta_v}{\sum_{v=1}^{V}N_k^{(v)}+\beta_v},\theta_d^{(k)}=\frac{N_d^{(k)}+\alpha_k}{\sum_{k=1}^{K}N_d^{(k)}+\alpha_k} \tag{3}$$

其中,$\phi_k^{(v)}$ 是诊疗项目 v 属于主题 k 的概率,$\theta_d^{(k)}$ 是第 d 诊疗日包含主题 k 的概率.

3 方案流程与算法实现

3.1 相似性约束构建

相似性约束构建可以定义为收集两两间距离小于某个阈值的诊疗日集合.由于一个诊疗日内的诊疗活动是无序的,因此,一个简单的方法就是将每个诊疗日映射到一个V维的TF-IDF向量空间,然后用欧式距离度量它们之间的距离.其中,TF反映的是一个诊疗日中某个诊疗项目的数量,IDF反映的是该项目的信息区分度.

但这一方法存在一个问题,就是医疗领域中有些诊疗项目的功能是相似的,在使用中常常可以互相替代,一个最常见的例子就是有些药品拥有不止一个名称.因此,将诊疗日直接映射到V维空间,没有考虑到医疗领域知识.为了解决此问题,我们引入SNOMED CT(http://www.ihtsdo.org/snomed-ct/)来度量不同诊疗项目之间的相似性,对于相似度高的诊疗项目,作为同一维度计算,从而对向量空间进行降维.

在SNOMED CT中,可以根据诊疗项目的层级位置,计算其信息含量值IC(information content)[16].

$$IC(o)=-\log\frac{\frac{|leaves(o)|}{|ancestors(o)|}+1}{\max\_leaves+1} \tag{4}$$

其中,$leaves(o)$和$ancestors(o)$表示诊疗项目o在SNOMED CT中叶子节点和祖先节点数量,max_*leaves*表示SNOMED CT中叶子节点的总数.给定两个诊疗项目间的最小公共祖先节点MICA(most informative common ancestor),二者间的相似度可由式(5)计算.

$$sim(u,v)=\frac{2\cdot IC(MICA(u,v))}{IC(u)+IC(v)} \tag{5}$$

所有*sim*值小于阈值的诊疗项目对(clinical item pair)均被列入用于降维的候选集合.然而,这些候选集合无法直接用于降维,因为它们之间可能存在着重叠关系(包含相同的诊疗项目).例如,考虑两个候选对(μ,ν)和(ν,ω),将有3种不同的降维策略:(1) 合并μ和ν至同一维度;(2) 合并ν和ω至同一维度;(3) 合并μ,ν和ω三者至同一维度.

为此,我们提出了一种贪心算法来确定具体使用上述哪一策略.算法基于两个原则:(1) 候选对之间无传递性,即已知两个候选对(μ,ν)和(ν,ω),不能推断出(μ,ω)也满足相似关系;(2) 优先选择高*sim*值的候选对.前者意味着只有当一个集合中所有的诊疗项目都两两相似,才能将该集合映射至同一维度.后者意味着对于两个候选对,优先选择*sim*值高的一对用于降维.基于此,算法首先对所有候选对根据*sim*值大小进行降序排序,然后在不违背无传递性原则的前提下,按序将高*sim*值候选对合并为同一维度.值得注意的是,已经被合并降维的诊疗项目不会被再次选中作为降维对象.

完成降维后,我们依据新维度计算两两诊疗日的TF-IDF向量之间的欧式距离,将距离小于阈值τ作为诊疗日相似度约束.相似度约束集合记作$\mathbb{C}=\{\langle s,t\rangle|s,t\in[1,D]\}$,其中,$s$和$t$表示诊疗日的索引,约束数量用$|\mathbb{C}|$表示.

3.2 PS-LDA

我们的目标是构建一个主题模型,可以在吻合数据特征和满足相似性约束之间取得平衡.对于前者,在LDA中通过最大化概率$P(A|\alpha,\beta)=P(A|\theta)\cdot P(\theta|\beta)\cdot P(\phi|\alpha)$实现,而根据LDA中的共轭属性,$\phi$和$\theta$可以由公式(3)计算得到.对于后者,我们希望$\mathbb{C}$中两个相似诊疗日的诊疗日-主题分布$\theta$可以尽量相似.因此,受文献[17]的工作启发,我们使用了一个最大化期望(EM)策略来实现这一目标.

在E阶段,不同于原始LDA,我们固定诊疗日-主题分布θ,然后将其用于构造Gibbs更新规则,即对公式(2)中隐藏变量Z的求解转换为式(6).

$$P(z_{d,i}=k\mid Z_{\neg z_{d,i}},A)=\frac{P(Z,A)}{P(Z_{\neg z_{d,i}},A)}\propto\frac{N^{(v)}_{k,\neg z_{d,i}}+\beta_v}{\sum_{v=1}^{V}N^{(v)}_{k,\neg z_{d,i}}+\beta_v}\cdot\theta^{(k)}_d \tag{6}$$

在M阶段,我们通过求解一个最优化问题,利用E阶段采样得到的主题分配Z来计算诊疗日-主题分布θ.对

应于我们的核心目标,最优化问题的目标函数应包含两部分:(1) 用于数据特征吻合的 $P(A,Z|\alpha,\beta)$;(2) 用于满足相似性约束条件的 $\sum_{c\in\mathbb{C}} dis(\theta_{c.s},\theta_{c.t})$,即两两相似诊疗日的诊疗日-主题分布 $\theta_{c.s}$ 和 $\theta_{c.t}$ 的欧式距离总和.因此,将两部分加权整合,形成可以平衡两者的最优化目标函数:

$$\ell(\theta)=\log(P(A,Z\mid\alpha,\beta))-\gamma\sum_{c\in\mathbb{C}} dis(\theta_{c.s},\theta_{c.t}) \tag{7}$$

然而,在该目标函数中,不仅包含隐变量,且 θ 是有约束的(每个诊疗日,所有主题的诊疗日-主题分布之和 $\sum_{k=1}^{K}\theta_d^{(k)}=1$),因此我们考虑将目标函数转换为非约束形式,以便用 EM 算法进行求解.我们用标准正态分布作为 θ 的先验分布来替代原本的 Dirichlet 先验分布,具体形式如下:

$$\theta_d^{(k)}=\mathrm{e}^{\lambda_d^{(k)}}\Big/\sum_{k=1}^{K}e^{\lambda_d^{(k)}} \tag{8}$$

由此,优化目标函数(7)相应转换为

$$\left.\begin{aligned}\ell'(\lambda)&=\log(P(A,Z\mid\lambda,\beta))+\log(P(\lambda\mid N(0,1)))-\gamma\sum_{c\in\mathbb{C}} dis(\theta_{c.s},\theta_{c.t})\\&=\log(P(Z\mid\lambda)\cdot P(A\mid Z,\phi)\cdot P(\phi\mid\beta))+\log(P(\lambda\mid N(0,1)))-\gamma\sum_{c\in\mathbb{C}} dis(\theta_{c.s},\theta_{c.t})\end{aligned}\right\} \tag{9}$$

由于替换了 θ 的 Dirichlet 先验分布假设,我们采用随机梯度下降法替代原先的 Gibbs Sampling(公式(3))来求解可以最大化目标函数 $\ell'(\lambda)$ 的最优诊疗日-主题分布集合 θ.

3.3 诊疗过程挖掘

根据 PS-LDA 计算得到的诊疗日-主题分布 θ,我们采用文献[3]中的过程挖掘框架进行基于主题的临床路径模型挖掘.主要分为 4 个步骤.

(1) 根据 θ,提取概率值高的主题集合(称为主题标签,例如(1,2))作为诊疗日的表示,用以代替原本繁杂的具体诊疗活动.主题标签中的主题按照概率值降序排列.

(2) 对于低频的主题标签进行合并剪枝,例如,如果(1,2,5,3)低频,则将缩减为(1,2,5),若(1,2,5)仍然低频,则以此类推,直到剩余标签中只包含单一主题.这一步骤的依据就是排序靠前的主题拥有更高的概率值,因此与诊疗日的也更相关.

(3) 病人日志转换为主题序列后,移除序列中的循环结构,包括自循环和包含循环.前者是对连续出现的相同主题标签进行压缩,例如序列{(1,2);(3);(3);(3)}将转为{(1,2);(3)};后者是当某个序列中,中间的所有主题标签都完全包含于首尾标签时,则进行压缩,例如序列{(1,2,5,3);(1);(1,2);(1,2,5,3)}将转为{(1,2,5,3)}.循环结构的压缩,是为了更简洁地展示不同主题标签之间的序列关系.

(4) 对处理后的所有病人的主题序列,用 Fuzzy Miner[8]进行模型挖掘.Fuzzy Miner 是一种基于显著度(significance)和关联度(correlation)的过程挖掘算法,对于高噪声、低结构化的日志数据具有较好的挖掘效果,能够得到易理解的过程模型.

算法 1. PS-LDA(priority similarity latent dirichlet allocation).

输入:*D* 个诊疗日的数据集;主题数目 *K*;超参数 β;权重常量 γ;Gibbs 采样迭代数 *nGS*;梯度下降迭代数 *nGD*;EM 迭代数 *nEM*.

输出:主题-诊疗项目项目 ϕ;诊疗日-主题分布 θ.

1. 随机初始化 *Z* 和 λ;
2. *i*←1;
3. while *i*<*nEM* do
4. *E* 步骤:
5. 根据公式(6)采样得到 *Z*(*nGS* 次迭代);
6. *M* 步骤:
7. *j*←1;
8. while *j*<*nGD* do

9. 设置学习速率ξ;
10. for k=1 to K do
11. for d=1 to D do
12. $\lambda_d^{(k)}(i+1) \leftarrow \lambda_d^{(k)}(i) + \xi \cdot \frac{\partial \ell'(\lambda)}{\partial \lambda_d^{(k)}}$;
13. $j \leftarrow j+1$;
14. $i \leftarrow i+1$;
15. 根据公式(2)和(8)分别计算ϕ和θ并返回;

4 实验评估

本节我们用一系列实验对算法有效性进行评估.

4.1 实验设置

我们使用来自某地级市人民医院的两个病种的收费项数据作为数据集.收费项数据是最普遍使用的医疗信息化数据,各地各级医院都拥有,并且由于其跟医保系统相关,数据可靠度高.但收费项数据非常低层级,涉及的诊疗项目繁多且复杂,因此,收费项数据对于临床路径挖掘既有重要价值,也是重大挑战.我们选择的两个数据集对应的病种分别为脑出血(intra cerebral hemorrhage,简称 ICH)和腹股沟疝(inguinal hernia,简称 IH),是分属于内科和外科的两个常见病种.表 2 罗列了数据集相关的重要统计指标.算法相关参数设置如下:超参数β=0.5,相似项目阈值 sim=0.75,相似诊疗日度量阈值τ=10,两个病种的主题数量 K 分别为 8 和 5.其中,β是一个 V 维的向量;sim 和τ是通过选取一个初始值,由医学专家对筛选出来的(大于阈值)最低值的若干相似对进行评估,根据评估结果对这两个阈值进行调整;主题数量的确定采用的是文献[3]中的权衡策略,该策略在模型的 Perplexity 与模型复杂程度(主题标签数量)之间进行了平衡.

Table 2 Statistics of our datasets

表 2 数据集统计

病种	病人日志数量	诊疗日数量	诊疗项目数量	平均住院日	最小住院日	最大住院日
ICH	240	3 204	752	14	2	34
IH	33	241	447	6	2	10

4.2 主题模型性能

我们从 3 个方面比较了 PS-LDA 和 LDA 的性能:一致度、覆盖度、相似性约束满足度.对于每个主题,我们根据主题-诊疗项目分布ϕ选取了排在前面的诊疗项目(top item)作为该主题的表示,然后邀请医生基于 top item 对每个主题进行专业标注(表 3 罗列了部分主题的 top 10 诊疗项目).

Table 3 Top items of partial topics

表 3 部分主题的 top item

病种	脑出血 ICH		腹股沟疝 IH	
标注**	(6) 药物	(8) 常规护理	(1) 入院检查	(2) 手术
1	泮托拉唑	二级护理	心电图检查(ECG)	高频电刀
2	甘露醇	住院检查费	心电监测	面具
3	氯化钠	静脉输液	静脉采血	血氧饱和度
4	维生素 B6	注射器	采血器	心电监测
5	维生素 C	动静脉置管护理	腹部 X 线摄影	腰椎麻醉
6	门冬氨酸钾镁	静脉穿刺	12 通道动态 ECG	葡萄糖
7	甘油果糖	敷贴	腹部多普勒超声	腹股沟疝修复术
8	钠洛酮	肌肉注射	尿检	手术包
9	吡拉西坦	静脉穿刺置管术	超声计算机图文报告	麻醉监测
10	依替米星	硝苯地平	患者健康教育	敷贴

**主题标注由医生根据 top item 中大多数诊疗项目确定

ICH 的 8 个主题分别被标注为:(1) 入院检查;(2) 医疗影像;(3~5) 生化检查;(6) 药物;(7) 高等级护理; (8) 常规护理.

IH 的 5 个主题分别被标注为:(1) 入院检查;(2) 生化检查;(3) 感染性疾病检查;(4) 手术;(5) 常规护理.

4.2.1 一致度

对于主题模型,一致度主要度量每个主题中的 top words 是否具有一致性,这对于主题的可解释性十分重要.在医疗场景下,我们关注的每个主题下的 top items 是否都围绕同样的诊疗目标展开.

- 定性评估

我们选取了 ICH 的药物主题,对 PS-LDA(表 3 第 2 列)与 LDA(表 4)的结果进行了对比.可以看出,在 LDA 发现的药物主题的 top item 中,有诸如“三通管”“静脉输液”等非药物类诊疗项目;而 PS-LDA 药物主题的 top item 全部都是药物相关诊疗项目.此外,LDA 发现的药物有不少是用于 ICH 并发症的药物,并非治疗 ICH 的常见药物;而 PS-LDA 中的“甘露醇”“甘油果糖”均是 ICH 所需基本药物.

Table 4 Drug-Related topic of ICH discovered by LDA

表 4 ICH 数据集中 LDA 发现的药物主题

药物	乙酰谷酰胺;氯化钠;局部浸润麻醉;依替米星;脑苷肌肽;物理降温;三通管;雾化吸入;氨溴索;静脉输液

- 定量评估

我们通过一个用户调查,对一致度进行定量分析.3 位医生被邀请对每个主题的 top 20 items 进行打分(2 分为非常相关,1 分为相关,0 分为不相关),然后基于投票机制确定最终分数.两个数据集下打分的 Kappa 值分别为 0.73 和 0.77(超过 0.7 可认为具有较高一致认可度).我们采用 *NKQM@n*[18](*NDCG@n* 的扩展)作为一致度的度量指标,其不仅考虑了每个 item 所得分数,还考虑了 item 在主题下的排序信息.计算公式如下:

$$NKQM@n=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{j=1}^{n}\frac{score(M_{k,j})}{\log(j+1)}}{Z_n} \tag{10}$$

其中,K 是主题数量,$M_{k,j}$ 是方法 M(分别为 PS-LDA 和 LDA)生成的主题 k 下的第 j 个诊疗项目,n 是 top item 的数量大小,Z_n 为归一化因子.表 5 展示了两个方法在不同 top item 规模下的效果对比,从表中可以看出,PS-LDA 在各个 n 下均取得了更好的一致度结果.

Table 5 Topic coherence in terms of *NKQM@n*

表 5 主题一致度 *NKQM@n*

病种	方法	*NKQM*@5	*NKQM*@10	*NKQM*@20
ICH	LDA	0.817 9	0.790 7	0.786 1
	PS-LDA	0.832 6	0.853 8	0.819 9
IH	LDA	0.786 0	0.787 4	0.757 4
	PS-LDA	0.821 8	0.819 5	0.800 3

4.2.2 覆盖度

不同于其他主题模型,在医疗场景下,我们还关注的一项重要指标是覆盖度.覆盖度指的是主题模型能否在其 top item 中发现该病种所有重要的诊疗项目.例如,“头部 CT”对于 ICH 是最重要的诊断依据;但在 LDA 的所有主题的 top 10 item 中,没有能够发现该项目.

为此,我们根据国家卫计委发布的临床路径和相关的临床指南,列举了所有必须和推荐的诊疗项目.所有必须项目我们设置为 10 分,而推荐项目设置为 5 分,进而对两种方法所得 top item 进行打分,用以度量算法结果的覆盖度.需要指出的是:我们采用的收费项数据的粒度是较低的,而国家版临床路径和指南中的项目粒度普遍是医嘱级别的,往往收费项中的多个项目才能对应到一个医嘱项目,例如指南中“感染性疾病筛查”在我们的数据集中对应着 11 项收费项目(1 项甲肝测定、7 项乙肝测定、1 项丙肝测定、1 项 HIV 测定和 1 项梅毒螺旋体测定),所以为了验证算法的覆盖度效果,需要在医学专家的帮助下进行项目的映射.表 6 展示了归一化后的总分数

比较,不难发现,在不同的 top item 规模下,PS-LDA 相对于 LDA 都具有更好的覆盖度.

Table 6 Topic coverage in terms of *Score@n*

表 6 主题覆盖度 *Score@n* 统计

病种	方法	*Score*@10***	*Score*@20	*Score*@30
ICH	LDA	0.57	0.83	0.86
	PS-LDA	0.64	0.94	0.98
IH	LDA	0.53	0.85	0.93
	PS-LDA	0.55	0.84	0.95

@n 表示取 topic *n* 个诊疗项目进行覆盖度 *score* 的计算

4.2.3 相似性约束满足度

上面两项指标,主要集中于度量主题-诊疗项目分布ϕ的质量.而本文的一个重要出发点,就是通过引入相似性约束,使得相似诊疗日的θ也尽量相似.因此,这里统计相似诊疗日的θ之间的欧式距离(即公式(7)的后一部分)来度量相似性约束满足度.图 3 展示了在不同相似性约束数量$|\mathcal{C}|$(横轴)下,两种方法的距离统计.从图中可以发现,PS-LDA 在约束满足度方面远超 LDA.

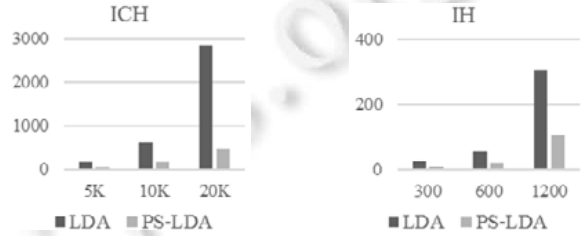

Fig.3 Statistics of similarity constraint conformity in terms of Euclidean distance

图 3 相似性约束满足度欧式距离统计

4.3 临床路径模型展示

经过 PS-LDA 对诊疗项目的聚类,我们可以通过过程挖掘框架得到基于主题的临床路径模型.图 4 和图 5 分别展示了两个病种的挖掘结果.

- 对于 ICH 的临床路径模型,我们可以将其分为 3 个主要阶段:入院检查、药物治疗、复查.这与国家卫计委发布的标准临床路径基本吻合.但有一个较为显著的异常,即 Topic 2 医疗影像并不是对所有病人都必要的.根据医生的分析,这一情况很可能发生在转诊病人上,这类病人有一部分是从其他医院做完了相关影像再来到该医院的(地区核心医院),因此省去了这一重复且较为昂贵的步骤.

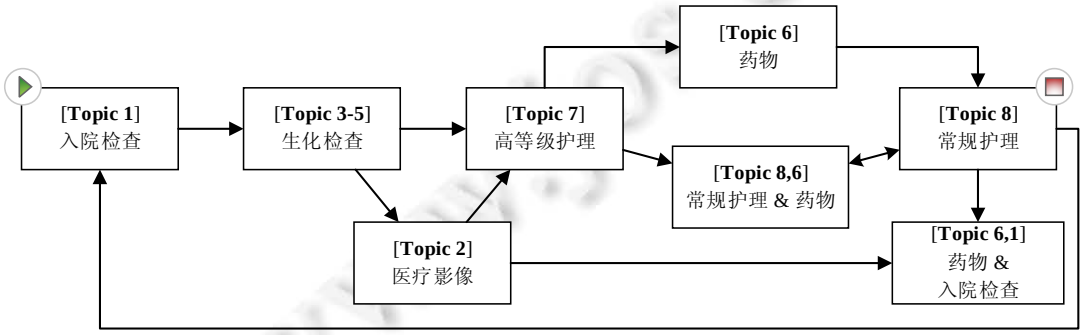

Fig.4 CP model of ICH

图 4 ICH 的临床路径模型

- 对于 IH 的临床路径模型,主要包含两个阶段:入院检查、手术治疗.相对于属于内科的 ICH,外科的 IH 治疗方法更为统一,基本都依赖于手术治疗,这与国家标准临床路径也基本吻合.

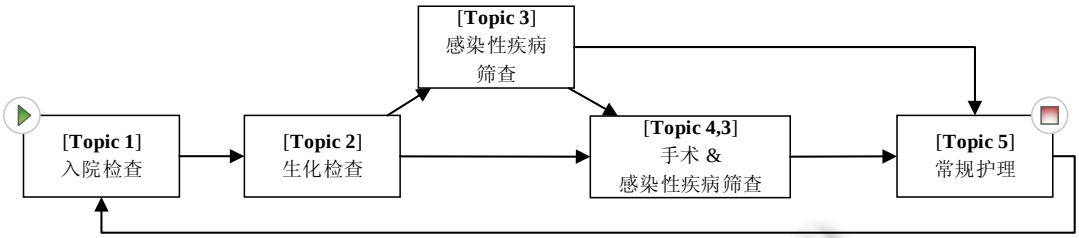

Fig.5 CP model of IH

图 5 IH 的临床路径模型

5 总结与展望

本文我们将诊疗日的相似性特征引入主题模型,用于发现高质量隐含主题,进行临床路径模型挖掘.方法首先基于领域知识检测出相似的诊疗日,构成相似性约束;然后将约束加入主题模型的优化目标函数,通过一个最大化期望框架进行迭代求解;根据所得结果将病人日志转换为主题序列,并通过过程挖掘算法得到最终的临床路径模型.该主题模型算法可以很好地平衡吻合数据特征和满足相似性约束条件的两方面需求.实验结果表明,我们的方法相对于 LDA 在一致度、覆盖度、相似性约束满足度方面都有较大的提升,所生成的临床路径模型简洁、易理解,可以辅助临床路径的设计/再设计及异常检测.

References:

[1] van der Aalst WMP, Desel J, Oberweis A. Business Process Management Models, Techniques and Empirical Studies. Berlin, Heidelberg: Springer-Verlag, 2000.

[2] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3(1):993−1022.

[3] Xu X, Jin T, Wei Z, Lv C, Wang J. TCPM: Topic-Based clinical pathway mining. In: Proc. of the 1st IEEE Int'l Conf. on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE, 2016. 292−301.

[4] Lin F, Chou S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. Int'l Journal of Medical Informatics, 2001, 62(1):11−25.

[5] Mans R, Schonenberg H, Leonardi G, Panzarasa S, Anna C, Quaglini S, van der Aalst WMP. Process mining techniques: An application to stroke care. Studies in Health Technology and Informatics, 2008,136(6):573−578.

[6] Van Dongen BF, de Medeiros AKA, Verbeek HMW, van der Aalst WMP. The ProM framework: A new era in process mining tool support. In: Proc. of the Applications and Theory of Petri Nets 2005. Berlin, Heidelberg: Springer-Verlag, 2005. 444−454.

[7] Weijters A, van Der Aalst WMP, De Medeiros AKA. Process mining with the heuristics miner-algorithm. Technical Report, 166, Technische Universiteit Eindhoven, 2006. 1−34.

[8] Günther CW, Van Der Aalst WMP. Fuzzy mining-adaptive process simplification based on multi-perspective metrics. In: Proc. of the Business Process Management. Berlin, Heidelberg: Springer-Verlag, 2007. 328−343.

[9] Poelmans J, Dedene G, Verheyden G, van der Mussele H, Viaene S, Perters E. Combining business process and data discovery techniques for analyzing and improving integrated care pathways. In: Proc. of the Advances in Data Mining. Applications and Theoretical Aspects, 2010. 505−517.

[10] Yang W, Su Q. Process mining for clinical pathway: Literature review and future directions. In: Proc. of the 2014 11th Int'l Conf. on Service Systems and Service Management (ICSSSM). IEEE, 2014. 1−5.

[11] Huang H, Jin T, Wang J. Clinical-Event packing method based on conditional probability. Computer Integrated Manufacturing Systems, 2017,23(5):1031−1039.

[12] Huang Z, Lu X, Duan H, Fan W. Summarizing clinical pathways from event logs. Journal of Biomedical Informatics, 2013,46(1): 111−127.

[13] Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. Journal of Biomedical Informatics, 2014,47(1):39−57.

[14] Huang Z, Dong W, Bath P, Ji L, Duan H. On mining latent treatment patterns from electronic medical records. Data Mining and Knowledge Discovery, 2015,29(4):914−949.

[15] Huang Z, Dong W, Ji L, He C, Duan H. Incorporating comorbidities into latent treatment pattern mining for clinical pathways. Journal of Biomedical Informatics, 2016,59(1):227−239.

[16] McInnes BT, Pedersen T. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. Journal of Biomedical Informatics, 2015,54(1):329−336.

[17] Du J, Jiang J, Song D, Liao L. Topic modeling with document relative similarities. In: Proc. of the IJCAI. 2015. 3469−3475.

[18] Danilevsky M, Wang C, Desai N, Ren X, Guo J, Han J. Automatic construction and ranking of topical keyphrases on collections of short documents. In: Proc. of the 2014 SIAM Int'l Conf. on Data Mining. Society for Industrial and Applied Mathematics, 2014. 398−406.

徐啸(1990－),男,安徽宁国人,博士,主要研究领域为医疗大数据,数据挖掘.

王建民(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为大数据与知识工程,流程数据管理与挖掘.

金涛(1980－),男,博士,助理研究员,主要研究领域为工作流,业务过程管理,医疗大数据.