计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):220−225 [doi: 10.15888/j.cnki.csa.006073]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于前缀集约束的临床路径挖掘算法①

韩　旭[1], 庄　严[2], 程绍银[2]

[1](中国科学技术大学 计算机科学与技术学院, 合肥 230027)
[2](中国科学技术大学 信息安全测评中心, 合肥 230027)

摘　要: 大量的研究表明, 临床路径在提高医院运行效率上发挥了极大的作用, 但是怎样方便快捷地找到某种疾病的临床路径是一个关键的问题. 随着信息技术的发展, 数据存储能力以及数据收集能力的提高, 各大中型医院都积累了大量的临床诊疗数据, 这为数据挖掘技术应用到临床路径发现提供了基础. 在这篇文章中, 我们把临床路径挖掘问题抽象成频繁序列模式挖掘问题, 我们首次提出了临床路径前缀集的概念, 并在此基础上提出了基于前缀集的临床路径挖掘算法 CPM-PC (Clinical Pathways Mining with Prefix Constraints), 这个算法更适用于临床路径挖掘, 挖掘出的序列模式有更强的医学意义, 这个算法已经被应用到一个真实的数据集上并且取得良好的效果.

关键词: 临床路径; 医疗数据; 序列模式挖掘; 医疗应用; 数据挖掘

引用格式: 韩旭,庄严,程绍银.基于前缀集约束的临床路径挖掘算法.计算机系统应用,2017,26(11):220–225. http://www.c-s-a.org.cn/1003-3254/6073.html

Mining Clinical Pathways Algorithm Based on Prefix Constraints

HAN Xu[1], ZHUANG Yan[2], CHENG Shao-Yin[2]

[1](School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)
[2](Information Technology Security Evaluation Center, University of Science and Technology of China, Hefei 230027, China)

Abstract: As lots of research has discussed, clinical pathways provide an effective way for improving the efficiency of hospitals. However, how to find useful clinical pathways conveniently is a problem. With the rapid development of networking, data storage and the data collections capacity, the hospitals have accumulated a large number of clinical data. In this paper, we characterize the problem of mining clinical pathways as a sequential patterns mining problem. We propose the concept of prefix of CP and integrate the prefix set into our algorithm CPM-PC: Clinical Pathways Mining with Prefix Constraints. The algorithm is more suitable for mining clinical pathways and will not search the sequences that has no medical significance. And the method has been applied to a real world data set to find clinical pathways and performs well.

Key words: clinical pathway; medical data; sequential pattern mining; medical application; data mining

我国是人口大国, 医疗资源相对紧缺, 看病难、看病贵问题是一直困扰政府和人民群众的一个大问题. 低劣质量的医疗服务往往与不规范的临床医疗行为有关[1], 临床路径就是规范临床行为, 在欧美发达国家已经被用于医院管理和疾病的诊断以及治疗, 经过几十年的发展, 临床路径在提高医院运行效率, 尤其在降低疾病治疗费用和提高医患满意度上都有显著的作用[2,3]. 2010 年国务院推动公立医院制定 100 种常见病治疗的临床路径并推进试点管理工作, 可见国家对医院推行临床路径的重视. 现阶段的临床路径制定主要还是通过专家经验来制定, 耗时费力, 如何高效的从海量医疗数据中发现有意义的临床路径是一个有意义的课题.

① 收稿时间: 2017-02-21; 修改时间: 2017-03-23; 采用时间: 2017-03-27

张兆国等[4]指出, 电子病历是实施临床路径的基础, 随着信息技术的发展, 医院等医疗机构都积累了大量的临床数据, 包括患者的基本信息, 诊断数据, 用药信息等, 医疗大数据的分析和应用在提高医疗效率和增强治疗效果将发挥巨大的作用, 同时这为大数据技术在临床路径挖掘领域的应用提供了基础.

1 引言

1.1 临床路径的概念

临床路径起源于工业领域的关键路径, 在 80 年代中期, Zander 首先提出关键路径的概念并把关键路径的概念用于医院临床的管理来提高对病人的治疗和护理的效率以及质量[5]. 临床路径的概念比较多, 表述上也各有其侧重点, 中日友好医院彭明强在借鉴各种定义后给出了一个比较综合和全面的定义, 他认为临床路径是指以循证医学证据和临床诊疗指南为指导, 针对某一疾病或病种制定的一套标准化治疗模式和程序, 是一个有关治疗、护理、康复、检验检测等临床治疗的综合模式[6].

1.2 临床路径的研究进展和应用

中国最早开始探索研究临床路径的是四川大学华西医院[7], 其在 1996 年将膝关节置换术患者纳入临床路径管理, 在这之后山东济宁医院和浙江台州医院也开始探索临床路径的研究应用, 这是国内临床路径研究的探索阶段. 直到 2009 年, 国务院卫生部门发布了《关于开展临床路径管理试点工作的通知》 (卫医政发〔2009〕 116 号), 临床路径才在国内开始较大范围的推广, 也带来了国内临床路径相关研究的高潮, 近几年每年都有数百篇临床路径相关的研究论文发表.

临床路径在医院管理中的应用, 使医院的临床治疗和护理流程的管理更加规范, 明确了各个步骤各个阶段的责任划分, 也让医生的工作更有计划性, 通过规范用药能避免过度医疗的发生, 为国家和社会节约了大量的宝贵医疗资源. 临床路径的采用还给患者带来了更好的治疗和康复体验. 首先, 患者对整个的治疗流程是知晓的, 保证了患者的知情权. 其次, 根据统计, 患者的住院天数和治疗费用也有不同幅度的下降. 表 1 是我们对阜阳市某医院采用临床路径方式治疗和非临床路径方式治疗的阑尾炎患者住院时常和费用的统计结果, 采用临床路径的治疗方案, 住院时间下降了 14.13%, 治疗费用下降了 6.4%.

表 1 采用临床路径前后统计对比

住院时间(天)		治疗费用(元)	
采用临床路径	未采用临床路径	采用临床路径	未采用临床路径
6.5	7.57	1870	1998

本文第一次提出了“前缀集”的概念并在此基础上给出一种基于频繁模式挖掘的临床路径挖掘算法, 并在一个真实的数据集上进行实验, 能在更短的时间内发现有医学意义的频繁序列模式, 从而指导临床路径的制定.

2 基础知识

2.1 序列模式挖掘

序列模式挖掘问题最早是由 R. Agrawal 和 R. Srikant 在 1995 年提出[8], 序列模式挖掘是数据挖掘领域极其重要的一个分支, 所谓频繁序列模式, 就是在一组有序的数据列组成的数据集中, 那些出现次数不小于最小支持度阈值的序列组成的模式. 序列模式同关联规则最大的不同是序列模式是有序的, 而关联规则不关注元素的出现顺序, 只考虑元素的相关性. 在很多有先后顺序约束的应用场景中, 序列模式挖掘能更好的发现一些隐藏的规则, 如在购物篮分析中, 预测用户购买 A 商品后还会购买什么商品, 就可以通过分析购物篮数据找到频繁序列模式从而预测接下来用户的购买习惯. 还有在分析 web 浏览习惯时, 可以通过分析上网用户的点击流数据掌握用户的浏览点击习惯, 为网站的 UI 设计优化提供支持.

下面通过一个例子说明序列模式挖掘的相关基本概念. 表 2 所示是一个有 4 个序列的序列数据集, 记为 S. 一个有序的序列组合被称作序列 Sequence, 如 <a(cd)(abc)de(cdf)>就是一个序列, 序列是由元素组成的, 上面序列中的元素有 a c d e f, 元素记作 Item. 作为同一时间段一起出现的元素被称为元素集, 如 (cd) (abc) (cdf) 都是 10 号序列的元素集, 记为 ItemSet.

表 2 序列数据集示例

序列ID	序列Sequence
10	<a(cd)(abc)de(cdf)>
20	<(ac)c(bcf)(abe)e>
30	<(ef)(ab)(adf)cb>
40	<g(adf)cb(ad)c>

出现频次是指某序列模式出现次数占序列总数的比例, 序列 α 的支持度定义如下:

$$Occur\_Freq(\alpha)=occur\ time(\alpha)/all\ Number(S)$$

在表 2 的数据集中, 假如出现频次阈值是 0.5, 则支持度为 2, 即在表 2 数据集中出现 2 次及以上的序列就是频繁序列模式, 序列 a 在 4 个序列中都有出现, 因此是频繁序列模式, 序列 ac(ab) 在 10 号和 20 号序列都有出现, 因此也是频繁序列. 由此可见, 序列模式是有很强的顺序性的, 而且会有重复出现的情况, 如上面的频繁序列模式 ac(ab) 中 a 出现了 2 次, 这也是序列模式挖掘问题和关联规则挖掘的不同之处.

2.2 PrefixSpan 算法描述

频繁序列模式挖掘算法可以分为两类: 一类是 Apriori 类算法, AprioriAll 算法和 GSP(generalized sequential pattern) 算法[9], 另外一类是模式增长算法, 如 FreeSpan 算法和 PrefixSpan 算法[10]. Apriori 类算法在产生序列模式的过程中会反复多次扫描数据库, 并产生候选序列集, 模式增长算法只需要少量几次扫描数据库, FreeSpan 算法需要扫描 3 次数据库, PrefixSpan 算法只需要扫描 2 次数据库, 并且他们都不会产生候选候选序列集, 在时空效率上, 吴孔玲等通过大量的对比实验证明 PrefixSpan 算法优于 FreeSpan 算法, 并且是上面两类算法中性能最好的[11]. 因此我们的算法是基于 PrefixSpan 算法进行改进, 在这一部分会对 PrefixSpan 算法的相关定义做一个介绍.

前缀. 给定两个序列 $\alpha=(a_1, a_2, a_3, a_4 \dots a_n)$, $\beta=(b_1, b_2, b_3, b_4 \dots b_m)$, $m\le n$, $a_i(1\le i\le n)$ 和 $b_j(1\le j\le m)$ 是包含多个元素的元素集, β 被称作 α 的一个前缀当且仅当:

(1) $b_i=a_i(1\le i\le m-1)$,

(2) $b_m \ a_m$.

例如, (ab)c 是序列 L2=(ab)cd(cef)de 的一个前缀, 我们知道通常一个序列的前缀有多个, 像 a, (ab), (ab)cd(cef) 都是序列 L2 的前缀.

投影. 给定序列 α 和它的子序列 β, α' 也是 α 的一个子序列, α'是 α 关于 β 的投影当且仅当:

(1) β 是 α'的前缀;

(2) α 不存在这样的子序列 α'', 满足 α'是它的子序列, 并且 β 是它的前缀.

例如, 序列 L4=<(bb)cd(def)gd>关于前缀<cd>的投影是<cd(def)gd>.

后缀. 给定两个序列 α'=<$b_1b_2 \dots b_n$>和 β=<$b_1b_2 \dots b_m$>, $m\le n$, α'是 α 关于 β 的投影, 那么 α''=<$b''_m b_{m+1} \dots b_n$>叫作 α 关于 β 的后缀, 当且仅当 $b''_m=b_m-b'_m$, 记作 $\alpha''=\alpha/\beta$.

例如, α'=<(ab)(de)cc(cf)>, β=<(ab)d>, α''=<(e)cc(cf)>是 α 关于前缀 β 的后缀.

投影数据库. 数据集 S 关于前缀 γ 的投影数据库是指这样的序列集合, 它们是序列中以 γ 为前缀的序列的后缀集合.

介绍了 PrefixSpan 算法相关的一些概念之后, 下面是 PrefixSpan 算法的一个基本流程.

输入: 序列数据集 S 和支持度阈值 k

输出: 频繁序列集

1) 找出所有长度为 1 的前缀和对应的投影数据库.

2) 对长度为 1 的前缀进行计数, 将支持度低于阈值 k 的前缀对应的项从数据集 S 删除, 同时得到所有的频繁 1 项序列, i=1.

3) 对于每个长度为 i 满足支持度要求的前缀进行递归挖掘:

a) 找出前缀所对应的投影数据库. 如果投影数据库为空, 则递归返回.

b) 统计对应投影数据库中各项的支持度计数. 如果所有项的支持度计数都低于阈值 k, 则递归返回.

c) 将满足支持度计数的各个单项和当前的前缀进行合并, 得到若干新的前缀

d) 令 $i=i+1$, 前缀为合并单项后的各个前缀, 分别递归执行第 3 步.

3 CPM-PC 算法

医疗数据中临床路径挖掘问题可以被抽象为频繁序列模式挖掘问题, 在医疗信息系统中, 患者的诊断治疗用药以及护理数据都会被采集, 并且发生时间也会被同步记录. 每个患者在医院从住院到出院产生的一系列数据是一个有严格时间先后顺序的序列, Gartner D 等指出, 一天的时间粒度在临床路径挖掘中是足够的[12], 所以可以按照日期对治疗项目排序, 一天的数据对应一个元素集合.

Kaymak 等人分析了把医学知识考虑进医学临床数据分析和发现有用模式中的重要性[13]. 我们观察到, 一个治疗流程往往是从做检查项目或者常规注射开始的, 这意味着临床路径是有特定开端的, 这样的开端可以称为"前缀". 常规的频繁模式挖掘算法获得的临床路径, 有些并无这样的前缀, 从医学意义上这样的临床路径是没有意义的. 因此, 我们提出了"前缀集"的概念, 前缀集是经过医学专家认定的可以作为某种疾病治疗

开端的治疗项目或者检查项目.

在此基础上, 我们提出了一个更加适用于临床路径挖掘的算法 CPM-PC(Clinical Pathways Mining with Prefix Constraints), 即有前缀约束的临床路径挖掘算法. 在频繁 1-项集中通过提前除去不在前缀集中的元素, 达到提前剪枝, 进而缩小搜索空间. 这样挖掘出来的临床路径更有医学意义, 可解释性更强. CPM-PC 算法可粗略地分为两步. 第一步, 扫描数据集 S, 找出所有 1-项集, 如果该项出现频次大于最小支持度且该项属于给定的前缀集, 则保留, 否则删除. 第二步, 对于每个符合条件的频繁 1-项集, 构建投影数据库, 在其中递归挖掘频繁模式.

算法. CPM-PC

输入: 临床数据集 S, 最小支持度 min, 前缀集 P
输出: 临床路径集 CP-set(clinical pathways set)

步骤: 初始化 CP-set=ø, 初始化 S'=ø为临床路径的长度为 1 的前缀集
扫描数据集 S, 如果 1-项集 v 满足 occur Freq(v) ≥ minsupport && v 属于 P, 那么
S'=S'+v
对于 S'中的每一项 α, 递归调用函数 FSP-Miner(α, l, S|α)

函数. FSP-Miner(α, l, S|α)

输入: α 是频繁模式, l 为模式 α 的长度, S|α 为 α 的投影数据库
输出: 频繁序列模式

步骤: 扫描投影数据库 S|α, 对于项 $\boldsymbol{x}$,
1. $\boldsymbol{x}$ 可以添加到 α 后, 形成序列模式 α'
2. <$\boldsymbol{x}$>组合到 α, 形成序列模式 α'
如果 α'是频繁的, 把 α'添加到临床路径集 CP-set
构造 α'的投影数据库 S|α'
递归调用 FSP-Miner(α, l, S|α')

4 数据准备

我们实验数据来自阜阳市某医院的真实临床数据, 数据存储形式为垂直型数据库, 首先要做的是从数据仓库中筛选出某种疾病患者的临床数据, 我们挑选高血压和支气管炎两种疾病的临床数据, 每种疾病单独建表. 在原数据库中, 患者的每条治疗记录有多达十几种属性, 我们对数据属性进行挑选, 选择患者 ID(patient_ID)、治疗项 (treat_Activity)、时间戳 (time) 作为新表的属性, 这三个属性是临床路径挖掘的关键属性.

数据清洗. 由于医院管理系统对数据录入审核机制不完善, 导致数据录入不规范. 数据不规范主要存在两个问题, 一个是汉字拼写错误导致同一种用药被当作不同的治疗项目, 最终挖掘到错误的结果, 例如, "肌酐测定"错误地录入为"肌肝测定", 对于这个问题我们采用模糊匹配的方法首先对具有 90% 相似的项目找出, 再进一步人工识别. 另一个存在的不规范问题是录入的治疗项目中存在空格的情况, 因为我们在算法阶段同一天的治疗项目是用空格作区分, 所以如果治疗项存在空格, 在序列挖掘阶段会被程序解析为多个治疗项目, 例如"5% 葡萄糖注射液", 对于这个问题, 我们在 sql server 里用 like 关键字作模糊查询.

数据格式转换. 数据清理完成后, 还需要把数据转换成易挖掘的格式. 在这部分工作中, 我们用 java 程序把每个病人的就诊治疗的临床数据按照时间顺序写到文件中的一行, 同一天的治疗项目不分先后, 用空格隔开. 一个病人的临床数据经过标准化处理后的例子是[全血细胞分析脑多普勒注射用盐酸川芎嗪 (冻干)][肝功能诊察费 (住院) 静脉输液电解质注射用泮托拉唑钠][加温器 X 线摄影血脂常规心电图检查].

5 实验及结果分析

我们选取两种疾病做实验对比, 高血压和支气管炎, 按照数据准备阶段对数据进行了清洗和标准化的格式转换. 这两个数据集的基本信息如表 3 所示.

表 3 实验数据集基本信息

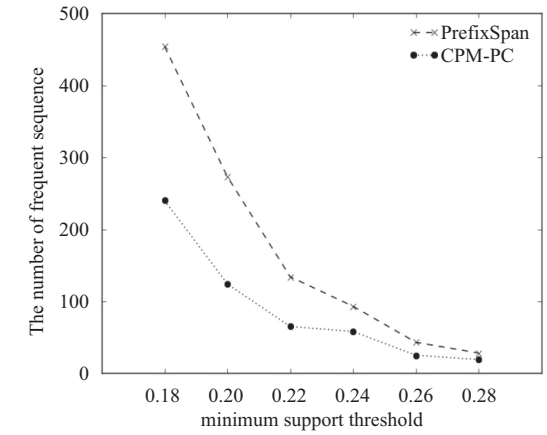
病种	治疗项目数	病人数
高血压	165070	1588
支气管炎	191581	1870

我们的实验对比是比较改进的算法 CPM-PC 算法和经典的频繁序列模式挖掘算法 Prefix-Span 做对比, 在时间性能和挖掘的序列模式两个维度进行比较. 我们的实验是在搭载 Win 10 系统的笔记本上进行, 笔记本配备了 2.30GHz Pentium(R) 处理器, 有 4G 的内存空间, 两个算法都是用 Java 语言实现.

经过向有关医学专家咨询, 我们为高血压和支气管炎选择了临床路径的前缀集. 高血压临床路径前缀

集为 prefix1=<静脉注射, 丹参注射液, 护理, 葡萄糖测定>, 支气管炎临床路径前缀集为 prefix2=<静脉注射, X 射线检查, 血清总胆红素测定, 常规心电图检查, 血常规检查>.

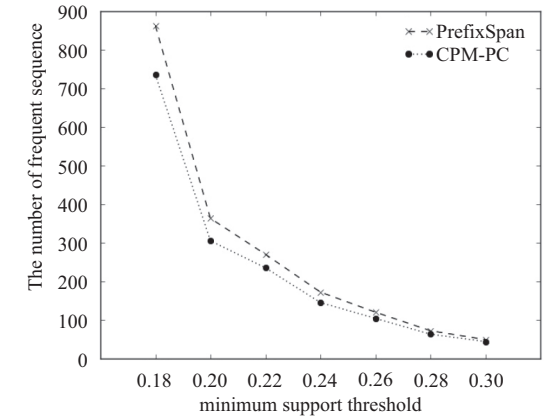
实验结果对比图如图 1 和图 2 所示. 图 1 是高血压数据集上的实验结果对比图, 图 2 为支气管炎数据集上的实验结果对比图. 图 1(a) 和图 2(a) 为挖掘出的频繁序列个数对比图, 图 1(b) 和图 2(b) 为程序执行时间对比图. 从上面的结果图可以看出, 随着支持度的提高, 挖掘出的序列模式个数在减少, 算法耗费的时间也在减少, 而在同一支持度下, CPM-PC 算法比 PrefixSpan 算法挖掘出的频繁序列更少, 这是因为 CPM-PC 算法通过加入临床路径前缀集, 提前去除了不是以前缀集中元素开头的序列模式, 这样挖掘出的临床路径具有更强的医学上的可解释性, 而传统的序列模式挖掘算法会挖掘出更多的模式, 但是这些模式中有些是医学上无法理解的, 不利于临床路径的选择.



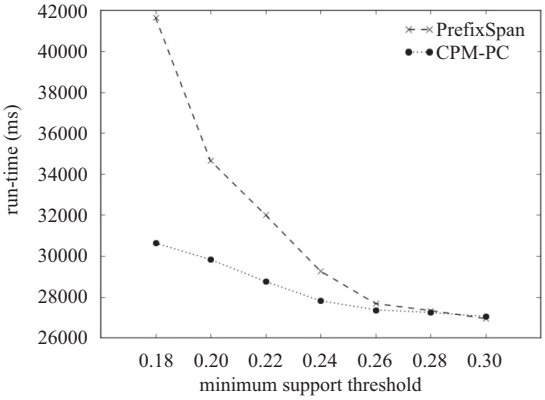
(a) 频繁序列个数对比图

(b) 程序运行时间对比图

图 1 高血压数据集实验结果



(a) 频繁序列个数对比图



(b) 程序运行时间对比图

图 2 支气管炎数据集实验结果

同时, 我们看到在高血压实验结果图 1(b) 中, 在支持度大于 0.24 时, CPM-PC 算法会花费更多的运行时间, 这是因为在通过前缀集对频繁 1 项集进行筛选时花费的时间超过了整个运行时因提前剪枝节省的时间, 当然这是与临床路径前缀集的选取是直接相关的.

6 结语

在这篇文章中, 我们介绍了临床路径的发展和临床路径在提高医院管理效率和患者满意度上的重要意义, 并通过医院的实际统计数据说明了临床路径的有效性. 临床路径的挖掘问题是一个有约束的频繁序列模式挖掘问题, 我们在传统的模式挖掘算法 PrefixSpan 基础上提出了临床路径前缀集的概念并在此基础上提出了一个临床路径挖掘算法 CPM-PC. 在阜阳市某医院的高血压和支气管炎两个病种的临床数据集上进行了对比实验并分析了实验结果, 实验结果表明我们的算法更加适用于临床路径的挖掘.

如何对临床路径进行评价是一个重要的问题, 在

接下来的工作中, 我们将会提出一个评估临床路径的一个数学模型, 用以评价临床路径的实际效果. 这个评估模型应该考虑多种因素, 包括医疗花费, 住院时长, 患者和医护人员满意度等[14]. 此外, 临床路径“前缀集”在本文中是人工选择方式产生的, 效率比较低, 如何通过数据挖掘的方法自动生成前缀集并动态更新是另外一个可以继续深入研究的内容.

参考文献

1 Topal B, Peeter G, Verbert A, *et al*. Outpatient laparoscopic cholecystectomy: Clinical pathway implementation is efficient and cost effective and increases hospital bed capacity. Surgical Endoscopy, 2007, 21(7): 1142–1146. [doi: 10.1007/s00464-006-9083-x]

2 Schuld J, Richter S, Folz J, *et al*. In?uence of IT-supported clinical pathways on patient satisfaction at a surgical department of a university hospital. DMW-Deutsche Medizinische Wochenschrift, 2008, 133(23): 1235–1239. [doi: 10.1055/s-2008-1077245]

3 Wazeka A, Valacer DJ, Cooper M, *et al*. Impact of a pediatric asthma clinical pathway on hospital cost and length of stay. Pediatric Pulmonology, 2001, 32(3): 211–216. [doi: 10.1002/ppul.1110]

4 张兆国, 陈联忠, 范可方, 等. 基于电子病历系统的临床路径管理应用研究. 中国数字医学, 2010, 5(10): 15–17. [doi: 10.3969/j.issn.1673-7571.2010.010.004]

5 Hofmann PA. Critical path method: An important tool for coordinating clinical care. The Joint Commission Journal on Quality Improvement, 1993, 19(7): 235–246. [doi: 10.1016/S1070-3241(16)30004-9]

6 彭明强. 临床路径的国内外研究进展. 中国循证医学杂志, 2012, 12(6): 626–630. [doi: 10.7507/1672-2531.20120102]

7 王思成. 基于循证的中医临床路径研制方法研究[博士学位论文]. 北京: 北京中医药大学, 2010: 65–67.

8 Agrawal R, Srikant R. Mining sequential patterns. Proc. of the 11th International Conference on Data Engineering. Taipei, Taiwan, China. 1995. 3–14.

9 Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. Proc. of the 5th International Conference on Extending Database Technology: Advances in Database Technology. London, UK. 1996. 3–17.

10 Pei J, Han JW, Mortazavi-Asl B, *et al*. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Trans. Knowledge and Data Engineering, 2004, 16(11): 1424–1440. [doi: 10.1109/TKDE.2004.77]

11 吴孔玲, 缪裕青, 苏杰, 等. 序列模式挖掘研究. 计算机系统应用, 2012, 21(6): 263–271.

12 Gartner D, Kolisch R. Scheduling the hospital-wide flow of elective patients. European Journal of Operational Research, 2014, 233(3): 689–699. [doi: 10.1016/j.ejor.2013.08.026]

13 Kaymak U, Mans R, van de Steeg T, *et al*. On process mining in health care. Proc. of 2012 IEEE International Conference on Systems, Man, and Cybernetics. Seoul, South Korea. 2012. 1859–1864.

14 El Baz N, Middel B, Van Dijk JP, *et al*. Are the outcomes of clinical pathways evidence-based? A critical appraisal of clinical pathway evaluation research. Journal of Evaluation in Clinical Practice, 2007, 13(6): 920–929.