文章编号：0253-374X(2018)04-0524-11 DOI:10.11908/j.issn.0253-374x.2018.04.016

基于过程挖掘的临床路径Petri网建模

余建波，郑小云，李传锋，董晨阳
(同济大学 机械与能源工程学院，上海 201804)

摘要：提出基于统计α算法的临床路径Petri网模型，将Petri网和基于统计α算法的过程挖掘算法集成，从事件日志中挖掘重要信息，获得完善的诊疗流程，并在此基础上建立Petri网模型，有效实现诊疗流程的优化和改进. 通过仿真数据试验，验证了本文提出的统计α算法相较于经典α算法在准确度和运行时间上有着较大的优势. 并将模型运用到临床路径真实数据上，证明了模型的有效性和准确性.

关键词：临床路径；统计α算法；Petri网建模；过程挖掘算法
中图分类号：TP18 **文献标志码**：A

Clinical Pathway Modeling by Petri Net Based on Process Mining

YU Jianbo, ZHENG Xiaoyun, LI Chuanfeng, DONG Chenyang
(School of Mechanical Engineering, Tongji University, Shanghai 201804, China)

Abstract: This paper proposed a clinical pathway Petri net model based on statistical α-algorithm, which integrate the basic Petri net with process mining algorithm based on statistical α-algorithm. This model can obtain medical procedure from event log and build clinical pathway Petri net model on the procedure. The medical procedure could be optimized and improved by analysis of the Petri net model. Simulation results shows that the statistical α-algorithm performs better in accuracy and efficiency than classic α-algorithm. The proposed model is verified on the real data of clinical pathway.

Key words: clinical pathway; statistical α-algorithm; petri net modeling; process mining algorithm

临床路径是医生、护士和其他人员共同制定的针对某病种所做的最适当的有顺序性和时间性的整体服务计划，目的是使患者获得最佳的服务，减少康复的延迟和资源的浪费[1]. 通过临床路径模型的建立和分析可以发现诊疗系统中存在的瓶颈问题，同时可以对临床路径的执行实施监控，对临床路径管理具有重大意义. 临床路径执行过程中产生的事件日志中包含大量的信息，是建立临床路径模型的数据来源，因此从事件日志中挖掘相关知识，将信息形成可用的流程，并在此基础上对临床路径进行建模分析是亟待解决的问题.

对于临床路径进行建模，首先需要从事件日志中提取信息，得到完整的诊疗流程，这需要采用过程挖掘(process mining)算法发掘诊疗流程. 过程挖掘是从实际事件日志中，运用过程挖掘算法，发现、监控和改进实际业务流程的思想. 过程挖掘可以深入分析诊疗活动之间可能存在的关系，不遗漏事件日志中任何出现的活动，并且可以自身反复验证结果，从而得到一个完整的流程. 过程挖掘思想最早由Cook等[2]提出，Agrawal等[3]将其引入工作流领域，并正式命名为过程挖掘. Herbst等[4]提出3个可以判断重名任务的过程挖掘算法，在过程挖掘上更加深入了一步. 对于过程挖掘算法的研究，以Aalst等[5]提出的α算法最为全面，目前已经衍生成一系列算法：α+[6]、α++[7]、α#[8]、α*[9]以及Tsinghua-α[10]算法. α系列算法是基于工作流网络(workflow net，WF-net)的行为推理算法，该系列算法不仅可以发掘事件日志中不同活动之间的顺序、并行、因果等基础关系，同时对事件日志中存在的非自由选择、重复活动等特殊关系也有着相当完善的处理. α系列算法都是建立在事件日志中没有噪声且事件日志中活动按照有序排列的前提假设基础上. 当事件日志中存在较多无规律噪声的时候，α系列算法往往会出现过拟合和准确度下降的情况.

临床路径模型是对临床路径整个流程的抽象化

收稿日期：2017-04-13
基金项目：国家自然科学基金(51375290,71777173)；中央高校基本科研业务费专项基金；上海市科技创新行动计划(17511109204)；上海市航天科技创新基金(SAST2015054)
第一作者：余建波(1978—)，男，教授，博士生导师，工学博士，主要研究方向为工作流管理、信号处理等. E-mail：jbyu@tongji.edu.cn

模型，用物理模型的方式将临床路径中诊疗活动、资源、信息等关系表达出来. 临床路径模型是对诊疗流程分析的基础，也是实现临床路径管理的根本，因此模型的准确性和完整性十分重要[1]. 对于临床路径建模的研究主要分为 2 个方向：基于 Petri 网建模和基于 UML 建模. 在实际的运用中，由于 Petri 网更加直观，更加符合使用人员的直观思维，同时 Petri 网是工作流最为常用的建模方法，因此对于临床路径建模研究主要集中在 Petri 网模型上. 文献[11]将保存在数据库中的文本诊疗常规转换为工作流过程描述语言(WPDL)模型和 Petri 网，分析诊疗常规的实施效果，验证了模型的行为正确性. 文献[12]提出利用一种临床路径典型语言 PROforma 对临床路径进行建模，并将临床路径转化成着色 Petri 网络. 文献[13]提出一种基于分层赋时着色 Petri 网对复杂病种建立临床路径模型的方法，实现了对诊疗状态、信息流转及诊疗活动间关系的可视化监控，并基于仿真结果给出了资源配置建议. 文献[14]在[13]的研究基础上做出改进，在对临床路径建模时，修改和新增部分与时间相关的参数和函数，增设费用相关变量及函数，对临床路径的住院时间和诊疗费用进行了定量分析.

综上所述，目前对于临床路径建模研究主要存在以下 2 个问题：第一，建模研究往往立足于已经存在的诊疗流程之上，并不能实现从事件日志中得到临床路径模型. 第二，目前常用的过程挖掘算法对于噪声的控制并不好，而在实际事件日志中的噪声数据总是存在且不可控制的. 因此需要首先给出一个可以消除噪声干扰并且能保证算法准确度的过程挖掘算法，再将其同 Petri 网模型集成，得到一个基于过程挖掘算法的临床路径 Petri 网. 两者集成不仅可以从事件日志中直接提取知识得到完善的工作流程，而且可以将临床路径转换成临床路径 Petri 网络，同时保证了模型准确率和建模效率.

本文提出了一种基于统计 α 算法的临床路径 Petri 网模型，将过程挖掘算法和 Petri 网络进行集成，实现了对于临床路径事件日志知识抽取，得到临床路径完善的诊疗流程，并据此建立 Petri 网模型，进行从事件日志到 Petri 网模型的转换.

1　基于统计 α 算法的临床路径 Petri 网建模方案

提出的基于统计 α 算法的临床路径 Petri 网建模方案如图 1 所示，包括算法挖掘和建模过程两块. 过程挖掘将输入的事件日志通过重名活动判别和统计 α 算法 2 个步骤得到活动关系矩阵及相关临床路径知识. 建模过程将 Petri 网和统计 α 算法以及临床路径的特征集成得到临床路径 Petri 网模型(CP-net). 接着将已经得到的活动关系矩阵和临床路径知识融入已经得到的 CP-net 模型中，得到针对该病种的 CP-net 模型，进一步可以对模型进行可达性、结构完整性和行为完整性的分析，并对该临床路径的完善程度进行考察.

2　基于统计 α 算法的过程挖掘

过程挖掘是从大量的事件日志中挖掘活动之间的关系，得到一个由这些活动关系组成的工作流. 因此，过程挖掘是实现事件日志到工作流模型的重要工具，通过过程挖掘算法对于临床路径事件日志的分析，才能得到完整的临床路径的工作流程，进而建立 CP-net 模型. 在过程挖掘中，对于活动之间关系的定义是整个算法的基础，活动关系的定义如表 1 所示[5].

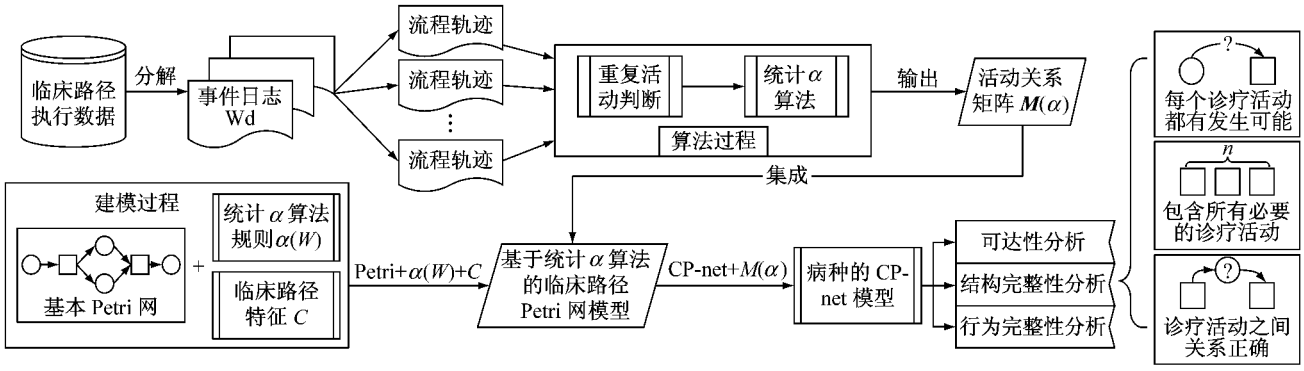

图 1　基于统计 α 算法的临床路径 Petri 网建模方案

Fig.1　Scheme of clinical pathway Petri net modeling based on statistical α-algorithm

表1 活动关系定义

Tab.1 Definition of activity relations

名称	符号	定义
顺序关系	$a>_w b$	$a,b\in\sigma=\{\cdots,a,b,\cdots\}$
因果关系	$a\rightarrow_w b$	$a>_w b\wedge\neg(a>_w b)$
选择关系	$a\#_w b$	$\neg(a>_w b)\wedge\neg(b>_w a)$
并行关系	$a\parallel_w b$	$a>_w b\wedge b>_w a$

2.1 统计 α 算法

统计 α 算法以活动对为识别的噪声最小单位，在创建活动对集合 $Array(AC)$ 时计算每个活动对的频率，在数据量较大时，用频率估计活动对的概率. 在利用统计 α 算法进行活动对关系判断之前，筛选出现概率低于显著性水平的活动对，并将其从 $Array(AC)$ 删除，不在最终的结果中出现. 噪声在事件日志中具体有3种体现：①小概率随机活动的增加. 由于该活动是低概率的，其组成的活动对必然也是低概率的，按照规则将从 $Array(AC)$ 删除. ②小概率活动的替换. 同样地，被替换的活动是低概率的，其组成的活动对也会被删除. ③小概率的活动缺失. 活动对缺失会出现新的活动对，而该活动同样是低概率的，因此也会被删除，从而消除活动缺失带来的噪声影响. 算法具体步骤图如2所示.

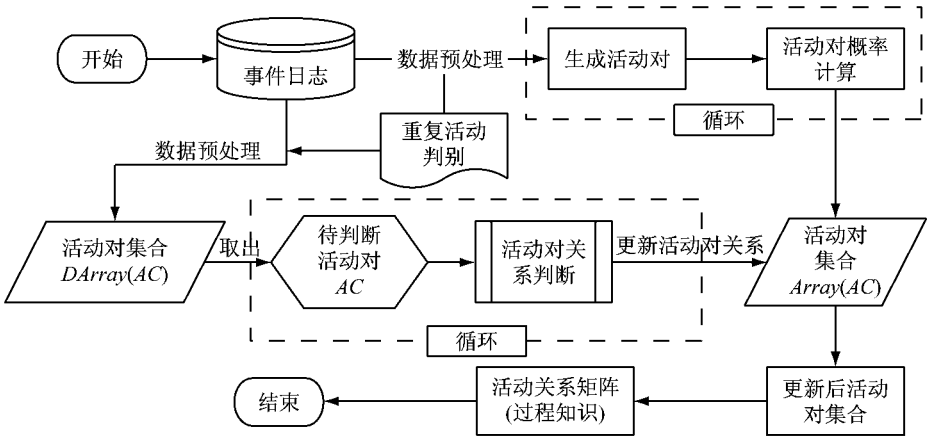

图2 统计 α 算法流程

Fig.2 Procedure of statistical α-algorithm

步骤1：活动对定义. 统计 α 算法以活动对为单位，活动对的定义如下.

定义1 对于任意病种的一个事件日志 W，假设有流程轨迹 $\delta_i=\{\cdots,T_i,T_{i+1},\cdots\}\subset W$，流程轨迹中的元素按照执行时间的先后顺序排列. 其中任意2个相邻的活动 T_i 和 T_{i+1} 及其出现的次数组成一个结构体，称为活动对(activities couple，AC). 具体定义如下.

Structure 活动对{

String FA，第一个活动名称(T_i)；

String SA，第二个活动名称(T_{i+1})；

int F，出现次数=1；

String R，活动关系=顺序关系(默认)；

}

步骤2：活动对概率统计. 遍历事件日志中的所有流程轨迹可得一个由不同活动对组成的活动对集合，记作 $Array(AC)$. 计算活动对集合 $Array(AC)$ 中每个活动对的出现概率. 考虑到重名活动的存在，可得活动对 $AC(T_i,T_{i+1})$ 的概率为

$$P(AC(T_i,T_{i+1}))=\frac{AC(T_i,T_{i+1}).F}{Nn_r}$$

式中：$AC(T_i,T_{i+1}).F$ 为活动对 $AC(T_i,T_{i+1})$ 的出现次数，如果在活动对集合 Array(AC) 中不存在 $AC(T_i,T_{i+1})$，则 $AC(T_i,T_{i+1}).F=0$；N 为流程轨迹数；$n_r=n_r(T_i)+n_r(T_{i+1})$ 等于活动 T_i 和活动 T_{i+1} 重名活动的数目总和.

步骤3：检索需要判断活动关系的活动对. 对于任意病种的流程轨迹来说，顺序关系是出现次数最多的活动关系[10]. 对于可以确定为顺序关系的活动对，不需要对其进行二次判断. 为了简化活动关系的判断流程，减少判断活动对的数目，需要对活动对进行筛选. 以流程轨迹为单位，纵向比较每条预处理后的流程轨迹，排除可以确定为顺序关系的活动对，得到一个活动关系待定的活动对集合，记作 $DArray(AC)$，显然 $DArray(AC)\subseteq Array(AC)$.

步骤4：判断活动对关系. 利用统计 α 算法进行活动对活动关系判断时需要考察与该活动对相关的活动对，因此这里给出与之相关的2个活动对的定义，即同前活动对和转置活动对.

定义2 同前活动对：对于任意活动对 $AC(A,B)$，若存在一个活动对和它有相同的第1个活动 A，

第 2 个活动不同，则称为 $AC(A,B)$ 的同前活动对，记作 $\#AC(A,B)$.

定义 3　转置活动对：对于任意活动对 $AC(A,B)$，若存在一个活动对的第 1 个活动等于它的第 2 个活动，第 2 个活动等于它的第 1 个活动，则称为 $AC(A,B)$ 的转置活动对，记作 $-AC(A,B)$.

结合上述定义，这里给出统计 α 算法，如下.

Input W_d//事件日志，α//显著性水平
Output Matrix[$Array(AC)$]//活动关系矩阵
1. Foreach(σ_i in W_d)
$\sigma_i \leftarrow \{T_1, T_2, \cdots, T_j, T_{j+1}, \cdots\}$//赋值
j : $1 \rightarrow \sigma_i$. Length
If ($AC(T_i, T_{i+1})$ is exist)then
$AC(T_i, T_{i+1}).F$++//活动对频数
Else
$AC(T_i, T_{i+1}).FA \leftarrow T_i$
$AC(T_i, T_{i+1}).SA \leftarrow T_{i+1}$
$P(AC(T_i, T_{i+1})) = \dfrac{AC(T_i, T_{i+1}).F}{N * n_r}$
$Array(AC).Add(AC(T_i, T_{i+1}))$//向活动对集合中添加元素，得到原始活动对集合
2. Foreach(σ in W_d)
$\sigma_i \leftarrow \{A_1, A_2, \cdots, A_j, A_{j+1}, \cdots\}$//构造轨迹
$\sigma_k \leftarrow \{B_1, B_2, \cdots, B_j, B_{j+1}, \cdots\}$//构造轨迹
j : $1 \rightarrow \sigma_i$. Length
If($A_j \neq B_j$)
$DArray(AC).Add(AC(A_j, A_{j+1}))$
$DArray(AC).Add(AC(A_{j-1}, A_j))$
End
Output $DArray(AC)$
3. Foreach (AC in $DArray(AC)$)
$AC \leftarrow AC(A,B)$
$\#AC(A,B) \leftarrow AC(A,C)$//同前活动对
$-AC(A,B) \leftarrow AC(B,A)$//转置活动对
If ($P(AC(A,C)) > \alpha$) then
Dim $AC(B,C)$
If ($P(AC(B,C)) > \alpha$) then
　If ($P(AC(B,C)) > \alpha$) then
　$B||_w C$
　Else
　$B\#_w C$
Else
$B\#_w C$
$A >_w B, A >_w C$
Else
If ($P(AC(B,A)) > \alpha$) then
$A||_w B$
Else
$A \rightarrow_w B$
End.
Output Matrix[$Array(AC)$]//活动关系矩阵

2.2　重名活动判别规则

在事件日志的某一条流程轨迹中，某一活动可能多次出现，但是每次代表的具体含义可能并不相同. 因此在活动关系判断之前，需要对这类活动进行区分，为了方便描述这一类活动，这里给出 2 个定义.

定义 4　在任一条流程轨迹中，若活动 A 出现一次以上，则该活动称为重名活动（cognominal activities）. 为了区分同一条流程轨迹中的重名活动，用记号 $\delta_i(A,n_i)$ 表示在流程轨迹 δ_i 中第 n_i 个活动 A[9].

定义 5　重名活动根据其具体活动内容是否相同分为 2 种：活动内容相同的称为重复性重名活动，简称重复活动（duplicate activities，DA）；活动内容不同的称为非重复性重名活动（homonyms activities，HA）. 提出重名活动判别规则是为了将重名活动区分为 DA 和 HA，并对 2 类重名活动进行不同处理，以消除 HA 对过程挖掘的影响.

为了提高重名活动判断效率，将任意活动 A 的 2 个前驱活动 T_P、T_{PP} 和 2 个后继活动 T_S、T_{SS} 组成的有序集合 $\{T_{PP}, T_P, A, T_S, T_{SS}\}$ 记作一个活动组（activities group），记作 G_A，并以活动组作为重名活动判别的基本单位对重名活动进行判别.

为了进一步确定重复活动的定义，这里引用 Herbst 等[4]对于重名活动和重复活动的定义，采用试验来探索定义. Herbst 等提出的重名活动称为“非独特活动”指的是在一个模型中多次出现的具体活动. 重名活动定义基于文献[9]中的定义，以整个事件日志为对象来寻找重名活动，而 Herbst 更加着眼于模型中的重名活动. Herbst 对于重复活动的定义更加符合实际应用场景中的定义，因此本文在正式提出重复活动定义之前，通过试验将假设与 Herbst 的结果进行对比，确保了本文中重复活动定义的可靠性. 试验采用内蒙古某三甲医院 2010 年 3 月到 10 月之间 7 个月的若干病种的事件日志数据，以活动组为单位，分析 4 个病种，共 510 条流程轨迹. 通过统计发现，当 2 个重名活动对应的 2 个活动组中的元素对应相等，即活动组中的元素和元素顺序都相同时，根据 Herbst 等[4]对于重复活动的定义可知，这 2 个重名活动是重复活动的概率高达 99.7%. 因此可以得到以下重复活动的定义.

定义 6　对于任意流程轨迹 $\delta_i \in W$，如果 δ_i 中存在重名活动 $\delta_i(A,n), \delta_i(A,m) \in \delta_i$，且两者对应的活动组分别为 G_A 和 G'_A，若 G_A 和 G'_A 中除了活动 A 之外的其他活动对应相等，则活动 $\delta_i(A,n)$ 和活动 $\delta_i(A,m)$ 可以定义为重复活动.

定义7 对于任意流程轨迹$\delta_i \in W$，如果δ_i中存在重名活动$\delta_i(A,n),\delta_i(A,m)\in\delta_i$，且两者对应的活动组分别为$G'_A$和$G_A$，若$G_A$和$G'_A$中包含的除了$A$之外的元素相同，但元素排列顺序不同，则称为2个活动组相等，记作$G_A=G'_A$.

对于任意流程轨迹$\delta_i\in W$，如果δ_i中存在重名活动$\delta_i(A,n),\delta_i(A,m)\in\delta_i$，$T_P$和$T_S$分别是活动$\delta_i(A,n)$的前驱任务和后继任务，$T_{PP}$为$T_P$的前驱活动，$T_{SS}$为$T_S$的后继活动；$T'_P$和$T'_S$分别是活动$\delta_i(A,m)$的前驱任务和后继任务，$T'_{PP}$为$T'_P$的前驱活动，$T'_{SS}$为$T'_S$的后继活动. 活动$\delta_i(A,n)$和活动$\delta_i(A,m)$对应的活动组分别为$G_A$和$G'_A$，$U$为重复活动集合，则有以下结论.

(1) 若$G_A=G'_A$，$T_P=T'_P$且$T_S=T'_S$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(2)若$G_A=G'_A$，$T_P=T'_P$，$T_S\neq T'_S$，$T_S=T_{SS}$，且$T_{SS}=T'_S$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(3) 若$G_A=G'_A$，$T_P\neq T'_P$，$T_S=T'_S$，$T_P=T'_{PP}$，且$T_{PP}=T'_P$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(4) 若$G_A=G'_A$，$T_P\neq T'_P$，$T_S\neq T'_S$，$T_P=T'_{PP}$，$T_{PP}=T'_P$，$T_S=T'_{SS}$，且$T_{SS}=T'_S$则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(5) 若$G_A\neq G'_A$，$T_P=T'_P$，$T_S\neq T'_S$且$T_S\#_w T'_S$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(6) 若$G_A\neq G'_A$，$T_P\neq T'_P$，$T_S=T'_S$且$T_P\#_w T'_P$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(7) 若$G_A\neq G'_A$，$T_P\neq T'_P$，$T_P\#_w T'_P$，$T_S\neq T'_S$且$T_S\#_w T'_S$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(8) 若$G_A\neq G'_A$，$T_P\neq T'_P$，$T_P\#_w T'_P$，$T_S\neq T'_S$，$T_S=T'_{SS}$且$T_{SS}=T'_S$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

(9) 若$G_A\neq G'_A$，$T_S\neq T'_S$，$T_S\#_w T'_S$，$T_P\neq T'_P$，$T_P=T'_{PP}$且$T_{PP}=T'_P$，则$\langle\delta_i(A,n),\delta_i(A,m)\rangle\subset U$.

2.3 集成重名活动判别和统计α算法的过程挖掘方法

提出的过程挖掘算法集成了重复活动判别和统计α算法，具体算法实施步骤，如图3所示，可以分为以下步骤：

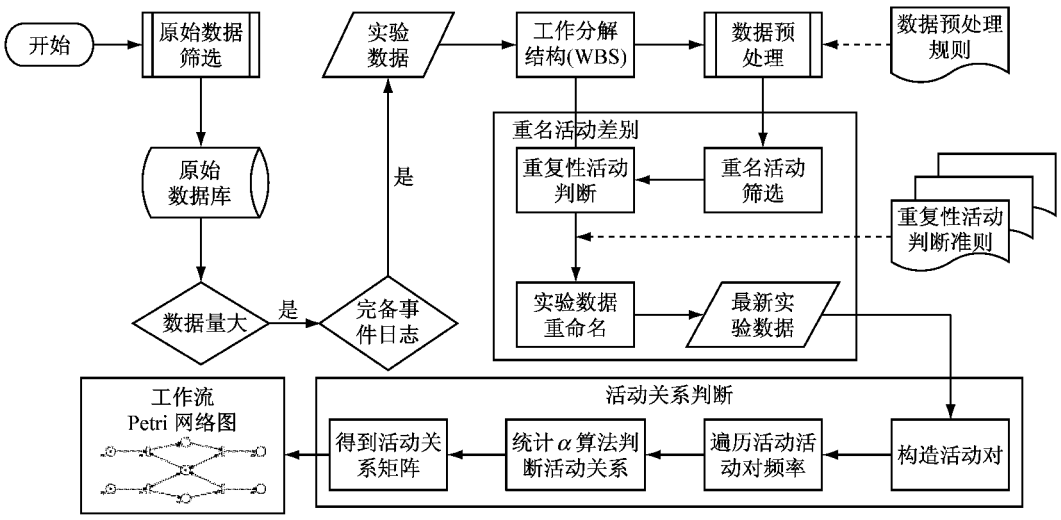

图3 基于重名活动判别和统计α算法的过程挖掘方案

Fig.3 Scheme of process mining that based on statistical α-algorithm and cognominal activity identification

(1) 原始数据的筛选. 获取数据库中结构完整、数据量较大的完备事件日志.

(2)工作分解结构. 将事件日志中的工作流按照其阶段、内容分成若干部分，从而可以减少每一部分包含的活动数目，降低算法运行时间，提高准确度.

(3)数据预处理. 将事件日志中的活动进行重命名排序操作，为重复活动判别做准备.

(4)重复活动判别. 以流程轨迹为单位，对其中出现的重名活动进行分析，根据分析结果重新修正活动命名，为活动关系识别做准备.

(5)活动关系识别. 采用本文提出的统计α算法，提取工作流程知识，分析活动之间的依赖关系，得到活动关系矩阵.

(6)得到结果并修正. 根据上一步骤中得到的活动关系矩阵，并将模型问题还原到实际问题之中.

3 基于过程挖掘算法的临床路径Petri网模型

临床路径是一种特殊的工作流，工作流与Petri

网有着天然的契合关系，因此建立临床路径的 Petri 网模型可以形象地将临床路径流程表现出来. 通过 Petri 网可达性、完整性等属性的分析，可以分析临床路径在执行中存在的问题、可能出现的瓶颈，Petri 网模型使得临床路径管理更加便捷. 临床路径 Petri 网的定义可以在工作流网络(WF-net)[5]的基础上加以修改实现. 在临床路径中，将病人看作唯一实体，病人的工作流是整个临床路径中的唯一工作流，下面给出临床路径 Petri 网(*CP*-net)的定义.

定义 7　令 $CP-\text{net}=(P,T;F,K,W,M_0)$，其中 $N=(P,T;F)$ 是一个 *WF*-net，称为 *CP*-net 的基网，是构成 *CP*-net 的最基本内容. P 为库所集，对应病人状态. T 为变迁集，对应临床诊治操作；F 为流关系，$F=(P\times T)\cup(T\times P)$. K 为 N 上的容量函数，规定了每个位置上的最大令牌数，该容量必须为有限值，可认为是临床路径上医生、护士、器材、药物等资源. W 为流关系上的权函数，对应到临床路径之中，可以认为权函数规定了诊疗活动开展的资源等条件. M_0 为 *CP*-net 的初始标识，标定了病人的初始状态. *WF*-net 是 *CP*-net 的基础，是模型的基本框架，在 *WF*-net 中包含了所有的诊疗活动、病人状态和流关系. *CP*-net 是工作流网的拓展，把临床路径中的相关条件和资源融入其中. 在进行过程挖掘时，主要还是以工作流网为对象，首先从事件日志中挖掘出工作流模型，再从事件日志和相关医疗资料中挖掘诊疗信息，最后在模型形成阶段将诊疗信息融入其中形成 *CP*-net.

CP-net 需要满足以下 5 条约束：

(1)起止唯一性：有且仅有一个 $p_i\in P$ 满足 $\cdot p_i=\varnothing$；有且仅有一个 $p_o\in P$ 满足 $p_o\cdot=\varnothing$. 临床路径的起止分别为病人的最初状态和最终状态，这个状态是唯一的.

(2)无孤性：不存在 $p\in P$，使得 $\cdot p\cap p\cdot=\varnothing$；不存在 $t\in T$，使得 $\cdot t\cap t\cdot=\varnothing$. 病人状态不可能单独存在，同样地，单独存在的诊疗活动也不可能出现在事件日志中.

(3)有界性：对于 $\forall p\in P,\forall M\in\mathbf{R}(M_0)$，存在一个非负整数 k，都有 $k\geqslant M(p)$，即任何状态下，库所的令牌总是有限个的，不存在没有输入库所的变迁. 在临床路径中，每一个诊疗活动的开展都是以病人当前状态为基础的.

(4)无死锁：对于 $\forall t\in T$，都可以通过执行某一变迁序列从而最终使得 t 使能. 即在临床路径中出现的诊疗活动都是有机会实施的，无法实施的诊疗活动不能包含在临床路径中.

(5)无活锁：对于最终库所 p_o，$M(p_o)=W(\cdot p_o,p_o)$，保证最终托肯数量为零. 在临床路径中一旦到达病人最终状态，不再进行该病种的任何诊疗活动. 如果后续其他活动出现，则判定为路径跳转.

基于统计 α 算法的过程挖掘算法是建立 *CP*-net 的基础和前提，通过过程挖掘算法发掘事件日志中活动的相互关系，进而得到整个诊疗流程，形成最终的 *CP*-net 模型. 本文采用的过程挖掘算法是结合了重复活动判别的统计 α 算法，通过将统计 α 算法与 *CP*-net 结合，可以使 *CP*-net 直接使用事件日志中的信息，能消除事件日志中的噪声数据，使得 *CP*-net 模型更加准确. 在本文使用的统计 α 算法中，对于活动关系的判断以活动对为单位，因此需要首先给出活动对在 *CP*-net 中的定义.

定义 8　对于 $CP-\text{net}=(P,T;F,K,W,M_0)$，$\forall a\in\mathbf{N}=(P,T;F)$，用记号 $<a$ 表示节点 a 的前一个节点，记号 $>a$ 表示节点 a 的后一个节点. 显然如果 $a\in P$，则 $<a,>a\in T$.

定义 9　对于 $CP-\text{net}=(P,T;F,K,W,M_0)$，$\exists a,b\in T$，使得 $a\cdot\cap\cdot b=>a=<b$，那么将活动组合 (a,b) 成为活动对，记作 $AC(a,b)$.

表 1 给出了过程挖掘判断的活动基本关系类型：顺序、因果、并行和选择. 这些活动关系是过程挖掘算法的主要结果，因此需要先将这些关系同 Petri 网结合起来. 这里引入工作流 Petri 网(*WF*-net)中的特殊节点 *AND*-split/join 和 *OR*-split/join，描述活动之间关系如图 4 所示. 图 4a 至 4c 分别用图形化的形式表达了顺序、并行和选择关系，对于包括因果关系的 4 个基本活动关系这里给出如下的定义.

定义 10　令 $CP-\text{net}=(P,T;F,K,W,M_0)$，$U$ 是重复活动集合，$\forall a,b\in T$，则

(1)若 $a\cdot\cap\cdot b\neq\varnothing\wedge b\cdot\cap\cdot a=\varnothing\wedge a,b\notin U$，则 $a\rightarrow_w b$(因果关系)

(2)若 $a\cdot\cap\cdot b=>a=<b$，则 $a>_w b$(顺序关系)

(3)若 $OR-\text{split}\in\cdot a\cap\cdot b\wedge OR-\text{join}\in a\cdot\cap b\cdot$，则 $a\#_w b$(选择关系)

(4)若 $AND-\text{split}\in\cdot a\cap\cdot b\wedge AND-\text{join}\in a\cdot\cap b\cdot$ 则 $a||_w b$(并行关系)

本文提出的统计 α 算法是以活动对为单位进行活动关系判断的算法，基于对“同前活动对”、“转置活动对”和活动对概率的计算，这里给出基于统计 α 算法的 Petri 网描述.

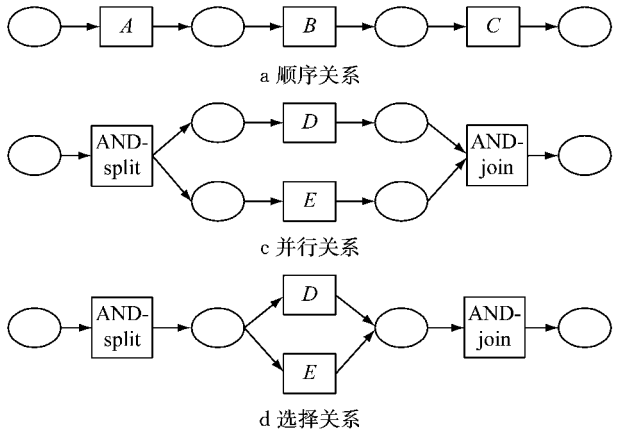

图4 活动关系 Petri 网描述

Fig.4 Description of activity relation in Petri net

令 W 为活动集合 T 上的事件日志，给定显著性系数为 α，统计 α 算法 α 定义如下：

(1) $X_W=\{AC(a,b)\mid \forall a,b\in W, a\cdot\bigcap\cdot b=\succ a=\prec b\}$

(2) $DX_W=\{AC(a,b)\mid \exists_{\sigma,\sigma\in W}AC(a,b)\neq AC'(a,b), AC(A,b)\in\sigma, AC'(a,b)\in\sigma'\}$

(3) $T_W=\{t\in T\mid t\in AC, AC.P\geqslant\alpha\}$

(4) $T_i=\{t\in T\mid \exists_{\sigma\in W}t=\text{first}(\sigma)\}$

(5) $T_o=\{t\in T\mid \exists_{\sigma\in W}t=\text{last}(\sigma)\}$

(6) $P_W=\{p(a,b)\mid (a,b)\in X_W\}\cup\{i_W,o_W\}$，

$$P(a,b)=\begin{cases}1, \text{当 } a\succ_w b \text{ 或 } a\rightarrow_w b\\ 2, \text{当 } a=AND-\text{split}\\ OR-\text{split/join}, \text{当 } a\#_w b\end{cases}$$

(7) $F_W=\{(a,p(a,b))\mid a\in AC(a,b)\in DX_W\}\cup\{b,p(a,b))\mid b\in AC(a,b)\in DX_W\}\cup\{(a,\prec b)\mid a,b\in AC(a,b)\in X_W, AC(a,b)\notin DX_W\}\cup\{(i_W,t)\mid t\in T_I\}\cup\{(t,o_W)\mid t\in T_O\}$

(8) $\alpha(W)=\{P_W,T_W,F_W\}$

在上述定义中，X_W 代表着所有活动对的集合，活动对出现概率需要包含在 X_W 之中. DX_W 则表示存在着差异的活动对集合，相对于上文中的 $DArray(AC)$. T_W 代表所有活动的集合，这些活动出现概率必须高于显著性系数，否则以噪声方式过滤掉，T_W 是 CP-net 中所有变迁的集合. T_i 和 T_o 分别是流程轨迹 $\delta\in W$ 的起始活动和终止活动的集合，根据 T_i 和 T_o 设置起止库所 i_W 和 o_W. P_W 是 CP-net 中所有库所的集合，$P(a,b)$ 表示变迁 a 和变迁 b 之间的库所，库所和变迁直接的连接方式如图 5 所示，因此 $P(a,b)$ 可能不止一个库所，当变迁 a,b 是顺序关系或者是因果关系时，$P(a,b)$ 的数量为一；而当 a,b 是并行关系时，变迁 a,b 的前一个变迁为 AND-split 节点，此时 a,b 变迁与其之间则各有一个库所存在；而当 a 和 b 是选择关系时，a 和 b 的库所则为 OR-split 节点. F_W 是 CP-net 中所有流关系的集合，统计 α 算法得到的 CP-net 主网络则主要由 $\{P_W,T_W,F_W\}$ 三者组成.

根据上述定义，如图 5 给出基于过程挖掘算法的临床路径 Petri 网建模步骤.

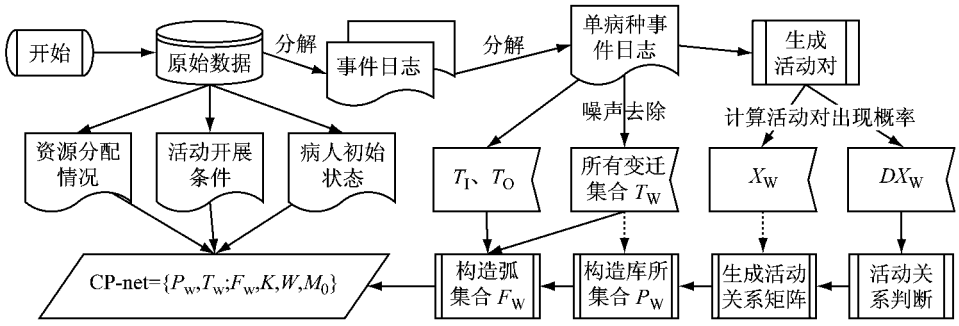

图5 基于过程挖掘算法的临床路径 Petri 网建模过程

Fig.5 Procedure of clinical pathway modeling based on process mining

(1)得到事件日志. 从原始数据开始，首先需要获得构建 CP-net 必需的事件日志，这是过程挖掘的基础，也是 CP-net 主体网络的基本构成.

(2)得到活动关系矩阵. 对于每个单病种事件日志，首先按照定义生成活动对，得到活动对集合 X_W，同时计算每个活动对的出现概率，据此按照规则去除噪声数据得到变迁集合 T_W. 接下来通过比对得到待判断活动对集合 DX_W 以及活动每条流程轨迹的起始和终止变迁集合 T_i 和 T_o. 对于集合 DX_W，需要通过遍历该集合，利用统计 α 算法识别每一个活动对的活动关系，结合 X_W 集合中已知的活动对关系，构造活动关系矩阵.

(3)构造主网络. 在得到的活动关系矩阵以及变迁集合 T_W 基础上按照 $\alpha(W)$ 定义的第 6 条生成相应的库所集合 P_W，并由 P_W、T_W、T_i 和 T_o 得到主网络的流关系集合 F_W.

(4)构造 *CP*-net. 在主网络形成之后，根据原始数据中对资源的分配得到容量函数 K，根据每项诊疗活动的开展条件得到流关系的权函数 W，根据病人的初始状态得到 *CP*-net 的初始标识 M_0，由此便得到最终的 *CP*-net 模型.

4　试验与结果分析

为了验证算法和模型的有效性，试验分为 2 个部分，第一部分采用仿真数据分析和比较统计 α 算法与经典的 α 算法以及 α＋算法在准确度、拟合度和运行时间上的差异[15]. 同时利用其中一组数据进行建模，分析模型的可达性、结构完整性和行为完整性等指标. 第二部分利用从医院采集到的真实临床数据进行建模，评价模型的可达性、结构完整性和行为完整性等指标.

4.1　仿真数据试验

采用文献[16]中给出的仿真数据生成方法，生成了如表 2 中的 4 组数据，每组数据中都包含了顺序、因果、选择和并行 4 种关系，分别从事件日志的轨迹长度、轨迹数目和噪声数目 3 个层面考察 2 种算法的性能表现. 算法准确度和运行时间结果分别如图 6 和图 7 所示.

表 2　仿真数据信息

Tab.2　Details of simulate data

仿真数据代号	轨迹长度	轨迹数目	噪声数目
A	50	100	50
B	60	90	50
C	70	80	50
D	80	70	50
E	90	60	50
N1	50	50	50
N2	50	75	50
N3	50	100	50
N4	50	125	50
N5	50	150	50
Z1	50	50	50
Z2	50	50	60
Z3	50	50	70
Z4	50	50	80
Z5	50	50	90
L1	50	50	50
L2	60	50	50
L3	70	50	50
L4	80	50	50
L5	90	50	50

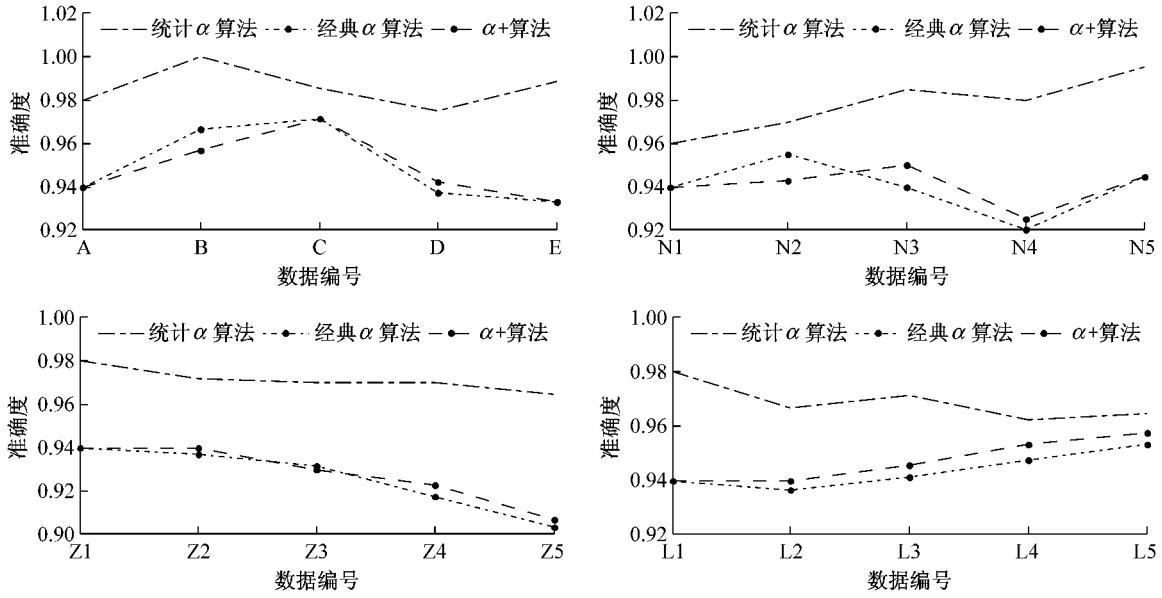

图 6　算法准确度对比

Fig.6　Comparison of accuracy

根据图 6 可以看到，不论仿真数据轨迹长度、轨迹数目和噪声数目如何变化，经典 α 算法和 α＋算法在各种参数上结果相近. 统计 α 算法在准确度上总是好于经典 α 算法和 α＋算法，从多组试验对比来看，统计 α 算法准确度总是比经典 α 算法和 α＋算法高 3%～4%左右，在准确度上有明显的优势. 根据图 8 可以看到，随着轨迹数目的增加，统计 α 算法在运行时间上明显优于经典 α 算法和 α＋算法；而随着噪声数目和轨迹长度的增加，2 个算法在运行时间上不相上下；因此在第 1 组的试验中，大多数情况下统计 α 算法在运行时间上优于经典 α 算法.

为了分析 2 种算法运行时间上的差异，这里给

出算法的时间复杂度分析:假设有一个轨迹数目为 n、轨迹平均长度为 m 的事件日志,可以将该日志看作一个 $m\times n$ 的矩阵.对于统计 α 算法而言,首先需要遍历事件日志得到活动对集合 $Array(AC)$,并计算每个活动对的出现概率,此过程需循环 $(m-1)n$ 次,在得到活动对集合后,计算每个活动对的具体活动关系,判断过程中不再包含循环结构,只有若干次选择结构(假设为 k 次),因此统计 α 算法的总循环次数为 $kn(m-1)^2$.而对于经典 α 算法而言,由于没有以活动对为基本单位,因此共需要对整个矩阵循环 2 次,总循环次数为 $n^2m(m-1)$.因此,当轨迹程度不变时,统计 α 算法运行时间随着轨迹数目的增加线性增加,而经典 α 算法则是幂增加,在 n 较大时,统计 α 算法在运行时间上明显优于经典 α 算法.而在轨迹数目不变时,轨迹程度的变化对于 2 种算法的运行时间影响并不大.

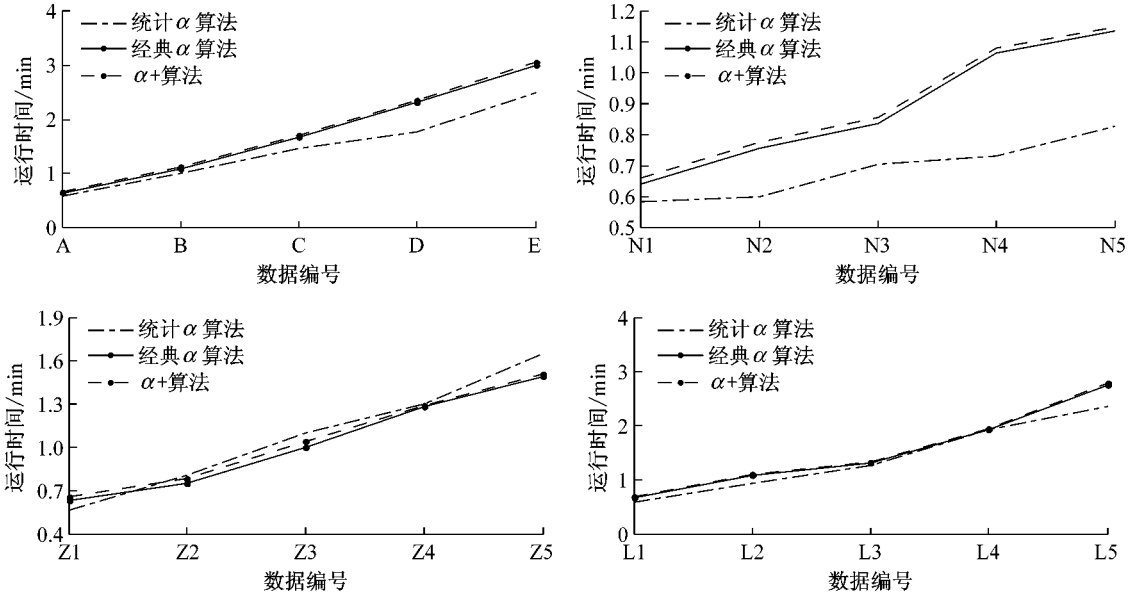

图 7 算法运行时间比较

Fig.7 Comparison of runtime

对于过程挖掘算法评价的另一个常用指标是拟合度.拟合度用来反映过程挖掘结果模型对原始数据的拟合程度.在算法结果中噪声较多时,模型往往会出现拟合程度过高而超过 100%的情况,过拟合同样是不理想的情况.拟合度越接近 100%,结果越好.同样对于上述 4 组仿真数据进行试验,可以得到表 3 中的结果.

α 系列算法由于其对噪声的消除不够,因而常常出现过拟合的问题.而统计 α 算法在进行活动关系判断之前,将其中的噪声进行消除,而这种消除的方式则是通过对于活动概率的统计进行的,因此,统计 α 算法的结果在拟合度上往往接近 100%.

表 3 算法拟合度对比

Tab.3 Comparison of fitness

试验数据	统计α算法				经典α算法				α+算法			
	第1组	第2组	第3组	第4组	第1组	第2组	第3组	第4组	第1组	第2组	第3组	第4组
A,N1,Z1,L1	98.3%	98.0%	98.0%	98.0%	99.2%	过拟合	过拟合	过拟合	99.2%	过拟合	过拟合	过拟合
B,N2,Z2,L2	100.0%	96.7%	98.3%	98.3%	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合
C,N3,Z3,L3	97.7%	98.3%	98.7%	98.7%	过拟合	99.2%	过拟合	过拟合	过拟合	99.6%	过拟合	过拟合
D,N4,Z4,L4	97.3%	97.3%	99.7%	98.7%	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合
E,N5,Z5,L5	99.5%	99.5%	过拟合	99.0%	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合	过拟合

对第 3 组仿真数据进行建模试验,将算法中得到的活动关系矩阵转化为 *CP*-net 模型.对于模型的分析指标主要有 3 个:结构完整性、行为完整性和可达性[14].其中结构完整性衡量了模型对于有意义的活动的包含程度,越高越好;行为完整性则是为了衡量模型中活动关系的准确性和完整性;可达性是 Petri 网模型不发生死锁的衡量标准.图 8 是仿真数据 Z1 的模型部分截图,分析模型可得到如下结果:

(1) 结构完整性.通过对模型和事件日志中活动的对比,该模型并没有包含仿真数据中出现的所有活动,根据计算,该模型的拟合度为 98.2%,若对照对仿真数据 Z1 中噪声的设定,该模型几乎没有包

含任何噪声信息，很好地消除了事件日志中的噪声.

(2) 行为完整性. 这里重点考察仿真数据中添加的特殊结构. 该模型由于其对于因果关系没有直接的体现，因此在因果关系方面略有不足. 但是对并行、选择等关系模型都能精准地表现出来.

(3) 可达性. 通过对于关键结构点的托肯分布以及变迁使能条件的分析，该模型可达率为 100%.

表 4 是对 Z1 至 Z5 这 5 个仿真数据建模分析的结果. 可以看到该模型在结构完整性和可达率上都达到了一个较高的水平，行为完整性由于因果关系在模型中没有直接的展现，所以数据并不算太好，但是也达到了 α 系列算法的平均水平.

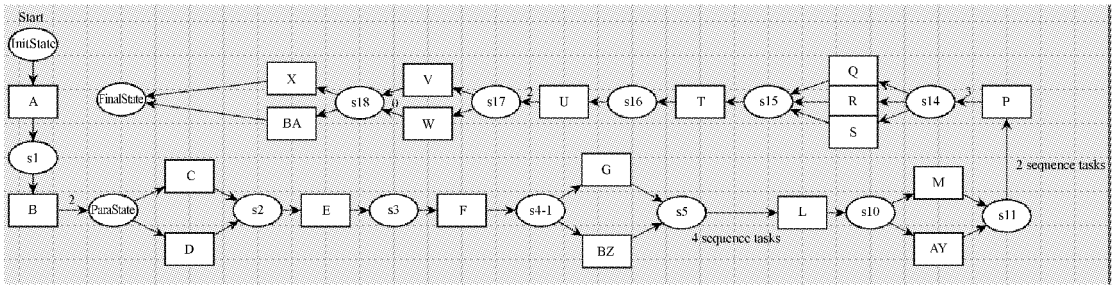

图 8　仿真数据 Z1 的 Petri 网模型(部分)截图

Fig.8　Petri net model of data Z1(part)

表 4　仿真数据 Petri 网建模分析结果

Tab.4　Analysis of Petri net model for data

仿真数据代号	结构完整性	行为完整性	可达率
Z1	98.2%	88.7%	100%
Z2	97.9%	88.7%	100%
Z3	97.5%	87.3%	100%
Z4	97.3%	88.7%	100%
Z5	96.8%	87.3%	99%

4.2　临床路径建模试验

临床路径建模试验的数据来自内蒙古某三甲医院 2010 年 7 月到 12 月锁骨骨折病种的事件日志，为了减少试验的计算时间，这里选取了锁骨骨折手术监护期的数据. 该部分数据流程轨迹平均长度为 60，重复活动有 5 对，事件日志中共有 156 条流程轨迹.

首先，分别使用统计 α 算法、经典 α 算法和 α＋算法对事件日志进行分析，得到算法结果如表 5 所示. 可以看到，结合了重复活动判别的统计 α 算法在准确度上明显优于经典 α 算法和 α＋算法，同时由于统计 α 算法对于噪声的消除，拟合度上也没有出现过拟合的现象，总体优于经典 α 算法和 α＋算法.

表 5　算法结果对比

Tab.5　Comparison of results

比较项目	拟合度	准确度/%	重复活动
统计 α 算法	0.993 3	98	5/5
经典 α 算法	过拟合	80	0/5
α＋算法	过拟合	82	0/5

接下来利用 CPN Tools 软件对算法结果进行 *CP*-net 建模分析，得到图 9 的 *CP*-net 模型图. 分析模型可以得到如下结果：

(1) 结构完整性. 不同于仿真数据中噪声随机的产生，该日志经过人工分析，噪声存在于事件日志中活动偶尔出现的缺失和错位，噪声量并不大，因此最终得到的模型在拟合度方面表现很好，结构完整性高达 99.3%.

(2) 行为完整性. 首先，该病种的事件日志中，存在着 5 对重名活动，模型准确地将重名活动识别出来，并分辨出其中一对为非重复性重名活动，重名活动判别准确率为 100%. 其次，该事件日志中存在的特殊关系较多，模型成功识别并表达出 3 处并行结构和 2 处选择结构，对于并行结构和选择结构的识别准确率为 100%. 对于因果关系该模型同样没有直接表现出来，属于模型欠缺的地方. 通过模型还可以发现，最后一处的选择结构为 2 层选择结构的嵌套，对于该处的选择结构可以进行如下理解：在手术完成后并经过基本护理后，需要对病人的状态进行询问和定义，若病人状态良好让病人正常休息、正常进食即可完成整个临床路径；若病人出现身体疼痛等生理问题，需要进行“疼痛护理”活动，为病人缓解病痛；若病人出现心理不安，则需要进行“心理护理”活动，为病人缓解压力. 总之，该模型对于特殊活动关系的反映率达到了 90%.

(3) 可达性. 该模型不存在死锁之处，短循环也存在着循环次数的限制，可达率为 100%.

总之，基于统计 α 算法的临床路径 Petri 网络模型在结构完整性和可达性上有很好的表现. 尽管在行为完整性上，对于因果关系的表达并不直接，但是因果关系在实际的临床路径执行过程中同顺序关系区别不大，因此整个模型对于实际医院的运作依然具有较好的指导意义. 通过对于具有完整数据的

CP-net模型进行分析,可以得到如下结果:①通过模型建立、实现从事件日志的流程重构至得到实际临床路径与标准路径中的区别,发现诊疗异常;②通过托肯分布和弧权重分布分析,可以发现关键库所,确定关键诊疗活动,从而优化资源分配;③通过*CP*-net模型最短路径搜索和实际可能性分析,优化诊疗流程,形成新的标准化临床路径.

此外,由于模型去除了噪声的干扰,整个模型可

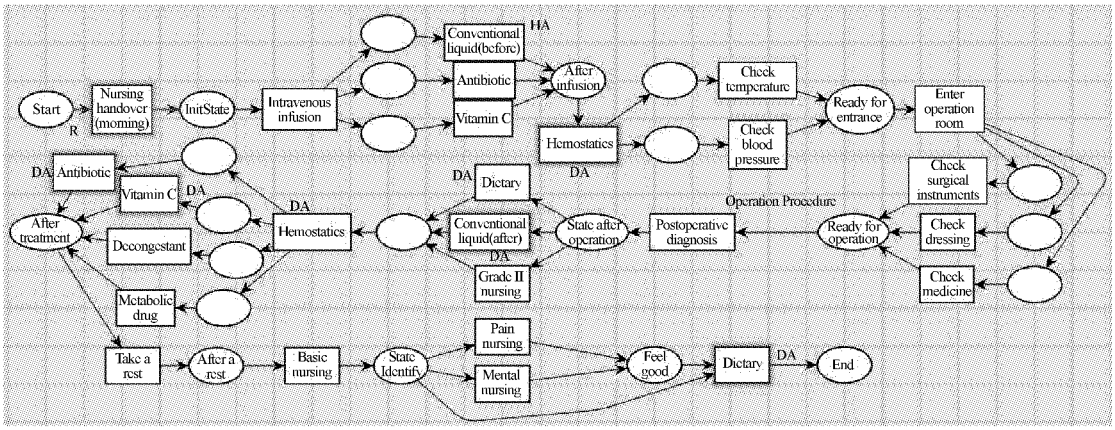

图9 锁骨骨折手术监护期临床路径Petri网模型截屏

Fig.9 The Petri net model of clavicle fracture

看作一个病种最基础的网络,具有极强的拓展性,在此基础上对于模型的衍生可以实现更多的功能,该模型也为临床路径的费用管理、资源调度等进一步的工作奠定基础.

4.3 显著性系数敏感性分析

显著性系数是统计α算法中的重要参数,其数值直接影响最终的结果.为了确定显著性系数α对试验结果的影响,增加显著性系数敏感性分析试验.试验数据采用来自内蒙古某三甲医院2010年7月到12月正常分娩和急性阑尾炎病种的事件日志.将显著性系数分别设定在0.01~0.20之间,共5组数据.试验结果如表6所示.

表6 显著性系数敏感性分析结果

Tab.6 Significant level sensitivity analysis results

显著性系数	正常分娩			急性阑尾炎		
	拟合度	准确度	泛化性	拟合度	准确度	泛化性
0.01	0.96	0.90	0.89	0.96	0.88	0.89
0.05	0.95	0.97	0.98	0.94	0.97	0.98
0.10	0.95	0.98	0.99	0.94	0.98	0.98
0.15	0.93	0.96	0.95	0.92	0.92	0.93
0.20	0.90	0.91	0.90	0.90	0.86	0.90

从试验结果可以看出:当显著性系数较小和较大时,结果的准确度和泛化性都会明显下降;拟合度在显著性系数较大时明显下降.具体来看,当显著性系数较小时,因果关系和选择关系识别率明显降低,出现第一类错误;反之,当显著性系数较大时,结果受到噪声的影响明显增大,保留了较多噪声,出现较多误判的因果关系和选择关系,出现了第二类错误.需要保证敏感性系数在一个合理的范围内,一般可以设定为0.05~0.10.

5 结语

针对临床路径建模提出了一套基于统计α算法的临床路径Petri网建模方法.首先给出了结合重复活动判别的统计α算法,实现了从包含噪声的事件日志中提取知识,形成完整的工作流程;接着提出了基于统计α算法的临床路径Petri网模型,该模型很好地契合了统计α算法,实现了事件日志到Petri网模型的转化.通过仿真数据和真实的临床路径数据试验验证了统计α算法较经典α算法在准确度和运行时间方面的明显优势;试验也证明了基于统计α算法的临床路径Petri网模型的可行性,该模型在可达性和结构完整性上表现优秀,可以用作临床路径管理的辅助工具.

参考文献:

[1] LANG M, BÜRKLE T, LAUMANN S, *et al*. Process mining for clinical workflows: Challenges and current limitations[J]. Studies in Health Technology & Informatics, 2008, 136:229.

[2] COOK J E, WOLF A L. Automating process discovery through event-data analysis[C]//Proceedings of the 17th international conference on Software engineering. [S.l.]: ACM, 1995: 73-82.

[3] AGRAWAL R, GUNOPULOS D, LEYMANN F. Mining process models from workflow logs[C]//International Conference on Extending Database Technology. Berlin Heidelberg: Springer, 1998: 467-483.

(下转第549页)