计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):1−8 [doi: 10.15888/j.cnki.csa.008415]

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

流程挖掘算法综述①

林文祥, 刘德生

(航天工程大学 复杂电子系统仿真重点实验室, 北京 101400)
通信作者: 刘德生, E-mail: liudsnudt@126.com

摘　要: 由于流程挖掘技术的快速发展, 流程挖掘算法种类增加迅速, 已有的算法研究文章介绍已不全面. 针对这一情况对迄今为止的流程挖掘主要算法进行系统性的分析总结. 首先对流程挖掘算法现状进行总体分析, 接着根据算法特性将流程挖掘算法分为传统的流程挖掘算法和基于计算智能和机器学习技术的流程挖掘算法两大类, 简要介绍其中代表性算法的基本思想和相关步骤, 最后比较了各类算法的优势和不足. 其中关于算法的分类和总结有助于初学者梳理流程挖掘领域相关算法知识, 而对发展现状和算法比较的分析则可以启发研究人员有待突破的方面.

关键词: 流程挖掘; 统计分析; 传统算法; 计算智能; 机器学习; 算法比较

引用格式: 林文祥,刘德生.流程挖掘算法综述.计算机系统应用,2022,31(3):1–8. http://www.c-s-a.org.cn/1003-3254/8415.html

Overview on Process Mining Algorithm

LIN Wen-Xiang, LIU De-Sheng

(Key Laboratory of Science and Technology on Complex Electronic System Simulation, Space Engineering University, Beijing 101400, China)

Abstract: Due to the rapid development of process mining technology, the variety of process mining algorithms has increased rapidly, and the introduction of existing algorithm research articles is no longer comprehensive. In view of this, we systematically analyze and summarize process mining algorithms so far. Firstly, we analyze the current situation of process mining algorithms in general and then classify them into two categories according to their characteristics: traditional process mining algorithms and process mining algorithms based on computational intelligence and machine learning technologies. Meanwhile, we briefly introduce the basic ideas and related steps of each subclass of representative algorithms and discuss the current advantages and disadvantages of the algorithms. Finally, suggestions regarding algorithm research and improvement in the next step are proposed. The classification and summary of algorithms can help beginners to sort out relevant algorithm knowledge in the field of process mining, and the analysis of the development status and algorithm comparison can guide researchers in areas that need to be broken through.

Key words: process mining; statistical analysis; traditional algorithm; computational intelligence; machine learning; algorithm comparison

流程挖掘是一门新兴的交叉学科, 起源于软件工程领域. 起初, 软件设计者只能根据业务设计师和业务管理者的相关介绍对工作流程进行建模, 容易导致流程模型具有主观性和片面性, 无法客观反映客户对软件的真实需求; 此外, 由于业务实际运转过程中的不确定性, 容易导致真实的业务流程与事先建好的流程模型存在差异, 难以借助模型分析和优化真实的业务流程. 流程挖掘技术应运而生, 被认为是解决这一问题的可行方案, 即使用系统的实际运行事件日志数据, 通过流程发现、验证等方法, 构建真实的业务流程模型, 用

① 基金项目: 国防科技重点实验室基础研究项目 (DXZT-JC-ZZ-2018-002, DXZT-JC-ZZ-2017-001)
收稿时间: 2021-05-21; 修改时间: 2021-06-21; 采用时间: 2021-07-27; csa 在线出版时间: 2022-01-24

于指导真实业务流程的分析和优化[1].

流程挖掘的概念自 1998 年首次被提出以来, 相关技术发展迅速, 其算法研究更是热点之一. 在 20 余年的时间里, 流程挖掘的新算法和改进算法层出不穷. 本文对流程挖掘算法进行统计学分析, 探讨了算法研究的总体情况和发展趋势; 基于算法特性将流程挖掘算法分为传统算法和基于计算智能和机器学习技术的算法, 对其中主要算法进行简要介绍和对比分析, 并提出下一步改进措施.

1 流程挖掘算法研究现状的统计分析

选取 Web of Science (WoS) 核心合集作为分析流程挖掘总体现状的数据来源, 设置检索的时间跨度为 1998–2020 年, 设置检索条件式为“Topic=‘process mining’” and “Topic=‘Algorithm’”, 共检索出文献 7 085 篇.

图 1 是将检索的 7 085 篇有关流程挖掘算法的文献按照发文年份进行统计分析, 由图可知, 在流程挖掘发展之初, 大约 2010 年之前, 有关算法文章的研究和发表呈现平稳姿态, 数目位于 100–200 之间, 表明这一时间是流程挖掘领域的奠基阶段, 研究发现许多经典算法诞生于此时间段内; 在 2010 年之后, 每年算法的发文数量呈现指数型增长, 在最近两年, 年增长数目均在 200 以上, 表明了当前流程挖掘算法研究正处于爆发期.

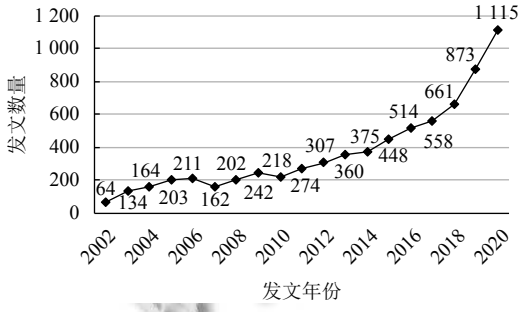

图 1 1998–2020 年外文文献年度趋势

运用 CiteSpace[2] 对 WoS 数据源的文献资料进行分析. 上述 7 085 篇文献经去重得到 6 863 篇独立文献, 分析得到突现关键词 116 个, 其中与算法相关的关键词 58 个. 算法突现词从一定层面上可以反映在某段时间内的研究热点, 甚至可以表明某一算法提出的时间, 比如 evolutionary algorithm, 其突现时间是 2007–2010 年, 这也是遗传挖掘提出并引发相关热点的时候, 再比如与模糊挖掘相关的关键词 Fuzzy 突现时间为 2008–2010 年, 这也与模糊挖掘算法提出时间接近. 此外突现词数量众多反映出流程挖掘相关研究自 1998 年发展至今, 算法提出和改进的相关研究非常多, 尤其是 2010 年以来表现得更加活跃; 突现强度前 30 的关键词如表 1 所示, 其中排名靠前的关键词均与人工智能等前沿领域相关, 相关算法在 2019 年左右开始突现, 表明与 deep learning、machine learning、artificial intelligence 等领域的结合已成为近两年来流程挖掘算法研究的热点.

目前算法的分类方式主要有以下 4 种: (1) 根据流程挖掘结果输出的模型表示方式对算法进行分类; (2) 基于应用领域对算法进行分类; (3) 基于算法特性进行分类; (4) 按照时间顺序介绍算法. 由于第 3 种分类方式具有交叉性小、算法特点明显等优势, 所以本文采取基于算法特性的分类方式, 将流程挖掘领域算法分为两大类, 分别是传统算法和基于计算智能和机器学习技术的算法, 并对其中骨干算法进行简短介绍.

2 流程挖掘传统算法

流程挖掘传统算法的主要特点是具有严密的逻辑推理和因果关系, 提出时间较早且发展较为成熟, 主要包括直接算法、启发式算法和基于概率统计的算法.

2.1 直接算法

直接算法的基本思想是直接对事件日志进行全盘扫描, 从中找出具有特定规律的模式. α 算法和相关改进算法是此类方法的典型代表[3–7].

以 α 算法[3] 为例. α 算法首先基于日志定义了 4 种活动间的次序关系 (以同一日志W中的活动a和b为例), 分别是伴随关系$a >_W b$、因果关系$a \to_W b$、并行关系$a \|_W b$和无关关系$a \#_W b$; 接着根据 4 种次序关系进一步构造 5 种控制流结构, 分别是顺序模式、选择分叉模式、选择合并模式、并行分叉模式和并行合并模式.

图 2 即为 5 种控制流结构 Petri 网的表现形式, 其中, 图 2(a) 是顺序结构, 要求满足的次序关系是$a \to_W b$; 图 2(b) 是选择分叉结构, 要求满足的次序关系是$a \to_W b$, $a \to_W c$并且$b \#_W c$; 图 2(c) 是选择合并结构, 要求满足的次序关系是$a \to_W c$, $b \to_W c$, 并且$a \#_W b$; 图 2(d) 是并行分叉结构, 要求满足的次序关系是$a \to_W b$, $a \to_W c$, 并且$b \|_W c$; 图 2(e) 是并行合并结构, 要求满足的次序关系是$a \to_W c$, $b \to_W c$并且$a \|_W b$. α 算法分类得到事件日志中各活动间关系, 在 5 种控制流结构的基础上, 输出完整的以工作流网为表示形式的流程模型.

表 1 1998–2020 年排名前 30 算法突现词

关键词	突现强度	开始年份	结束年份	1998–2020年
Deep learning	26.854 2	2019	2020	
Machine learning	20.594	2019	2020	
Association rule	15.620 3	2003	2009	
Artificial intelligence	12.227 7	2019	2020	
Random forest	12.120 4	2017	2020	
Sequential pattern	11.426 8	2003	2015	
Regression	11.184 7	2016	2020	
Convolutional neural network	9.302 6	2019	2020	
Gene expression	9.073 9	2004	2014	
Fuzzy set	8.841 3	2003	2012	
Frequent itemset	7.987 7	2003	2012	
Data mining	7.821 4	2009	2009	
Classifier	6.917 8	2011	2016	
Graph mining	6.831 1	2014	2017	
Efficient algorithm	6.780 6	2014	2017	
Logistic regression	6.688 1	2019	2020	
Petri net	6.287 3	2004	2013	
Process model	5.442 3	2018	2018	
K-mean	5.330 7	2008	2013	
Principal component analysis	5.189 1	2019	2020	
Tree	4.977 5	2011	2014	
Extreme learning machine	4.785 5	2018	2018	
Rough set	4.709 3	2016	2016	
Bayesian network	4.682 9	2013	2014	
Particle swarm optimization	4.409 6	2019	2020	
Workflow mining	4.368 2	2004	2006	
Evolutionary algorithm	4.224 2	2007	2010	
Association rule mining	4.135 7	2018	2018	
Social network	4.119 6	2018	2018	
Fuzzy	4.082 2	2008	2010	

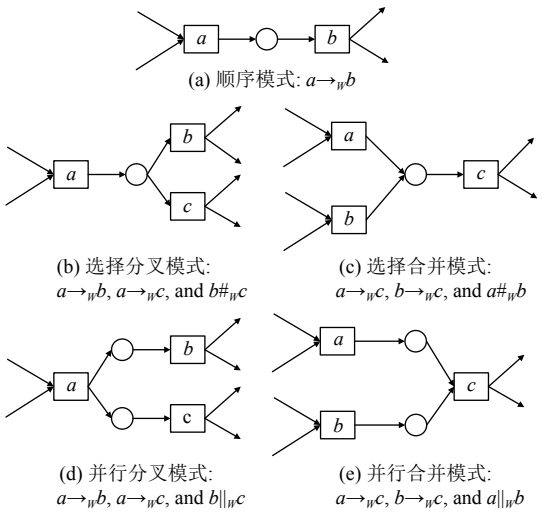

图 2 α 算法中的典型控制流结构

α 算法的提出促进了流程挖掘的快速发展, 目前仍是流程挖掘领域的主流算法之一. 但是, 因为该算法不能处理带有非自由选择结构、短循环结构、重复任务和不可见任务等复杂结构的流程模型, 很多学者基于 α 算法做了深入研究并提出改进措施. 其中, α^+算法[4]可用于处理流程模型中具有长度为 1 或 2 的短循环结构; α^{++}算法[5]则能发现依赖关系大于 1 的隐式依赖关系; $\alpha^{\#}$算法[6]针对的是隐式变迁问题; α^*算法[7]则是可以处理重复活动. 到目前未止还没有一种可以综合上述所有优点的成熟的算法, 文献 [8] 在此方面做了有益的探索和尝试, 使 α 算法初步具备同时处理多种复杂结构的能力.

2.2 启发式算法

α 算法的应用需要一个重要前提, 即日志完备且没有噪声. 噪声指的是记录错误的日志数据或者流程执行出现异常所记录的数据, 这在流程实际运行的时候必然会出现. 为解决流程挖掘中的日志噪声问题, 研究人员提出了启发式算法[9,10].

噪声日志是指出现频次明显低于其他行为的特定行为[11], 这是基于流程运行的稳定性的一种假设, 即噪声日志在正确日志中所占比例很小. 启发式算法的基本思想是基于此基本假设, 在挖掘事件日志时, 考虑流程实例出现的频率, 从概率统计的角度识别噪声, 通过设定阈值, 将低频实例作为噪声过滤掉. 文献 [9] 首次提出了启发式挖掘算法, 其基本思想是在挖掘事件日志时, 使用频率来刻画一对活动之间因果关系的强度, 并以依赖度度量, 通过事件日志中不同活动之间的连接数计算得到依赖度, 依赖度低于某个阈值的日志即视为噪声, 将之过滤后输出相应的控制流结构形成流程模型. 文献 [10] 对启发式挖掘步骤进行了详细介绍, 可总结如下.

1) 挖掘依赖图. 其中包括计算活动间依赖度, 形成活动依赖表, 再设定阈值构造活动间依赖图. 图 3(a) 是一个构造的依赖图的例子[12], 其中活动用字母代替, 活动间关系用有向弧表示, 并在相应有向弧标明活动间的依赖关系, 此环节中低频噪声已通过阈值被过滤.

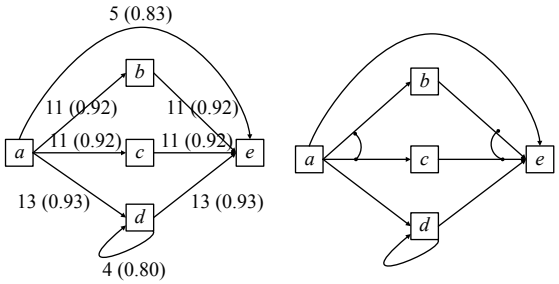

图 3 C-net 构造过程

2) 在依赖图基础上确定并行和选择分支, 通过定义并行选择结构判断的计算公式 $a \Rightarrow_W b \wedge c = \left(\frac{|b>_W c| + |c>_W b|}{|a>_W b| + |a>_W c| + 1}\right)$ 和设定阈值, 当计算值小于某一阈值时被当作选择结构处理, 大于某一阈值时则被当作并行结构处理. 图 3(b) 是构建成功的一个 C-net 例子, 在依赖图的基础上对选择或并行关系进行了标识[12].

启发式算法在应对噪声问题上走出了创新的一步, 但基于频率阈值的对低频事件一刀切也会导致错误删除有效的低频行为, 已有学者考虑这个问题后提出了一些解决方法. 就目前而言, 在针对噪声处理问题上启发式思想仍是主流, 并未出现更多可供选择的噪声处理方法.

2.3 基于概率统计的算法

此类算法是概率统计相关算法在流程挖掘中的应用, 具有明显的或潜在的概率模型, 揭示了特定数据点之间的相关性概率, 与流程挖掘需要得到活动连接概率、现象发生频率等契合度高, 因此在流程挖掘领域发挥重要作用. 常见于流程挖掘领域的概率统计算法有: 马尔可夫 (Markov) 模型[13], 随机活动图 (SAG) 算法[14], 增量分析[15], Apriori 算法[16,17], 随机 Petri 网等.

马尔可夫模型[13] 是一个具有无后效性的随机过程, 表示系统从一个状态 n 转移到另一种状态 n+1 只取决于系统在 n 时刻的状态, 即未来状态的选择仅与系统当前状态相关, 而与之前的状态无关, 并将系统从一个状态转移到另一个状态的概率称之为状态转移概率. 业务流程也可看作一个随机过程, 它描述了系统从一个任务状态到另一个任务状态. 由于流程的任务在时间 t 具有任务状态, 因此业务流程被视为具有有限状态的马尔可夫链, 文献 [13] 设计了一种基于马尔可夫转移矩阵的自动业务流程建模方法, 可推导任务间所有可能的逻辑关系且不受日志质量影响.

随机活动图 (SAG)[14] 是有着明确定义的元组, 包括节点、边、活动、活动分配函数、转移概率、开始节点和结束节点, 由于随机活动图和工作流模型的基本元素都是活动, 可用随机活动图的构建方式发现日志中活动关系. 文献 [14] 通过对 SAG 活动节点和工作流实例的节点进行映射, 并要求 SAG 活动尊重实例活动之间的时间依赖性, 用随机活动图作为工作流模型的中间表现形式, 再进行工作流模型转换.

与 α 算法和启发式算法专注于流程发现不同, 概率统计算法可用在流程挖掘的优化方面, 如利用 Apriori 算法进行关联规则挖掘[18], 从而找出高效的工作组合; 再比如利用 Markov 模型的两个活动先后发生概率得到流程执行的路由概率, 从而检测流程是否异常[19]. 概率统计方法识别流程模型中的缺陷和问题应用较为广泛.

3 基于计算智能和机器学习的流程挖掘算法

此类流程算法主要来自于数据挖掘相关技术在流程挖掘领域的应用, 相比于传统流程挖掘算法, 此类技术的整体成熟性还不够, 大部分还未能达到很好的挖掘效果, 但兴盛的计算智能和机器学习浪潮未必不能给流程挖掘领域带来不一样的前进方向.

3.1 基于智能优化的流程挖掘算法

智能优化算法在流程挖掘领域具有天然的优势, 一是可以处理大多数的控制流结构, 并且挖掘结果具

有很高准确率; 二是算法鲁棒性好, 可以很好地处理噪声; 三是基于局部次序关系的全局搜索, 有利于得到全局优化的结果; 四是适应度函数的自由设置, 便于得到更符合用户倾向的流程模型. 常见的在流程挖掘中使用的搜索算法有: 遗传算法[20]、模拟退火算法[21]、禁忌搜索算法[22]、和粒子群[23] 等.

智能优化算法的基本框架[22] 如图 4 所示.

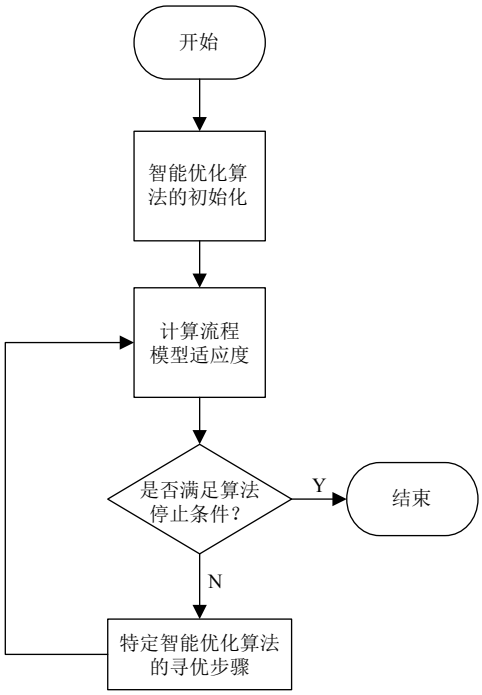

图 4 智能优化算法基本框架

(1) 智能优化算法的初始化. 算法初始化要考虑两个方面因素: 一是算法本身原理, 如遗传算法需要生成初始种群, 而模拟退火算法和禁忌搜索算法仅需生成一个初始个体; 二是要考虑流程挖掘的特点, 主要就是如何表示各流程个体以适应于智能优化技术, 可采用因果矩阵或 Petri 网等表现形式.

(2) 计算流程模型的适应度. 适应度是判断算法停止的主要依据, 流程模型的适应度主要参考 4 个评价指标: 拟合度、泛化度、精确度和简单度, 四者相互制约, 从中寻求平衡是难点.

(3) 智能优化算法停止的条件. 算法停止条件可自行定义, 常用的有: 一是适应度函数若干轮计算后达到预设数值, 二是算法的循环次数达到设定的上限, 三是适应度函数的值经历若干次计算未发生变化或变化量小于设定阈值等.

(4) 智能优化算法的寻优操作. 该步骤是智能优化算法在流程挖掘领域应用的核心步骤, 目的是通过寻优操作提升个体或群体的适应度. 不同算法寻优方式不同, 需要把流程挖掘任务中的概念根据算法进行抽象, 对算法进行适应性修改. 例如基于遗传算法的流程挖掘算法, 在抽象个体表示为因果矩阵后, 需要对针对基因的交叉、变异和选择等策略修改针对因果矩阵的操作, 这些操作是否合理和有效将成为优化算法输出结果优劣的主要影响因素.

智能优化算法随着计算效率的提高而越来越受追捧, 这也反映了智能优化算法的不足, 即算法执行时间过长, 执行效率不高, 利用智能优化算法对业务流程进行离线分析尚还可以, 但流程挖掘的发展趋势是支持在线运行, 目前从效率来看仍有较大差距.

3.2 模糊挖掘

传统流程挖掘技术有很多显现或者隐含的假设, 使其在结构化流程上表现良好, 但应用于结构化程度低的流程, 无法提供具有洞察力的模型. 这不是说传统技术挖掘的模型不正确, 而是挖掘的模型显示了所有细节, 没有从事件日志本身提供任何有意义的抽象, 因此对于流程分析师来说是无用的. 模糊挖掘技术[24] 就是针对此类现象, 实现了流程模型的挖掘和自适应简化, 其基本步骤如下:

(1) 首先创建初始模型

将日志中的所有事件类型都被转换为活动节点, 其重要性由一元重要性来表示. 对于事件类之间的优先关系, 采用相应的有向边进行连接. 这条边由二元重要性和相关性来描述.

(2) 将下列 3 种转换方法应用于流程模型, 依次简化流程模型的特定方面.

1) 二元关系中的冲突解决

若活动A发生后执行B, 定义A对于B的相对重要性为$rel(A,B)$, 同理B对于A的相对重要性为$rel(B,A)$, 根据相对重要性处理 3 种冲突情况.

一是两个冲突关系$rel(A,B)$和$rel(B,A)$的相对重要性均超过先前设定的保留阈值, 则$A \to B$和$B \to A$都将被保留; 二是至少一个冲突关系的相对重要性低于该阈值, 计算确定两个关系的相对重要性之间的偏移量. 如果偏移值超过先前设定的比率阈值, 则移除不重要的边; 三是至少一个关系的相对重要性低于保留阈值, 并且它们的偏移小于比率阈值, 则两个边都从过程模

型中移除.

2) 边的过滤

如图 5 所示, 首先定义可配置的效用比ur, 即带有权重的考虑两活动之间的二元重要性和二元相关性, 计算得到有效选取值Util.后进行归一化得到Normal Util., 再与设定参数删除值Co相比较, 删除小于阈值的边.

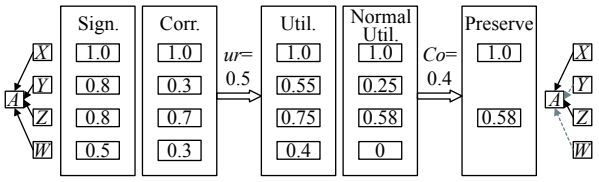

图 5 过滤节点 A 输入边的过程

3) 节点聚合和抽象

借鉴数据挖掘中的模糊聚类方法, 模糊挖掘提出了结点聚集的方法, 对于流程模型中重要性较低但是相关性却很高的多个结点采用聚集的方式, 同时基于结点参数删除值的大小移除孤立且重要性低的结点[25].

模糊挖掘技术对于解决传统流程挖掘技术挖掘的“意大利面式”的流程模型具有显著的作用, 在发现开发人员编码过程[26] 和查找恶意软件[27] 等方面发挥了作用. 模糊挖掘算法针对不同的事件日志可通过配置合理参数得到有效的流程模型, 从适用范围来讲较大. 但优点也是缺点, 针对不同的应用场景, 如何找到正确的参数是一件耗时耗力的事. 模糊挖掘的发展需要得到高适应性的默认参数设置或易操作的合理参数寻找方式, 提高算法在实际应用中的可使用性.

3.3 决策挖掘

事件日志中包含着大量的事件属性信息, 如时间戳、执行者和一些附加数据等, 目前的流程挖掘算法更多的是利用前两个属性来构造反映因果相关性的流程模型, 即基于控制流视角的流程发现, 而对日志中其他数据利用较少, 无法解释活动为何被选择执行. 机器学习算法[28] 已经成为从大量数据中提取知识的广泛采用的手段, 文献 [29] 最早提出了决策挖掘的概念, 通过决策树分析技术分析日志中的各种数据信息如何影响具体的流程实例进行决策, 即得到决策规则, 关注流程挖掘的案例视角.

决策挖掘也称之为决策点分析, 旨在挖掘与业务流程路由相关的数据信息. 现有的决策挖掘算法可分为 3 个步骤.

(1) 使用已有控制流挖掘算法挖掘出业务流程的控制流模型.

(2) 识别过程模型中的决策点.

以 Petri 网为例, 若 Petri 网中某一库所对应了多个输出弧, 则可被判定为决策点.

(3) 使用决策树分析过程模型中的决策点.

在对过程模型中的决策点进行识别之后, 需要判定过程实例的数据属性是否对实例的决策产生影响, 即决策规则挖掘, 其思想就是将每一个决策点转换为一个分类问题, 具体类别则是作出的不同决策[30].

如图 6, 决策树会构建出一个树状模型. 模型存在一个根节点、多个决策结点和多个叶子结点. 其中根结点对应了样本的全部数据集, 而根结点到叶结点的路径则对应一个决策序列. 叶结点则对应的是决策树最后的判断结果, 决策结点设置数据属性, 把每个结点包含的样本数据集合根据属性测试条件分别划分到其子节点中.

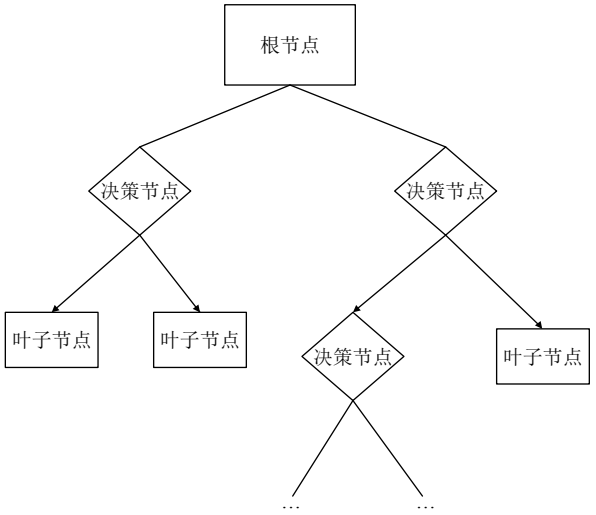

图 6 决策树模型

决策树挖掘提出来之后, 很多学者针对其进行了深入的研究和改进. 文献 [31] 针对流程日志与流程模型不一致会影响挖掘结果的问题, 提出了增加一致性检验步骤; 文献 [32] 则是提出在决策点挖掘时同时考虑外部数据信息和内部各决策点的结构关系, 提升了决策点决策规则挖掘的全面性和准确性. 决策挖掘主要针对案例视角, 运用机器学习相关技术填补了空白, 但在实际应用中还面临许多难题, 包括工作流日志数量好坏直接影响决策挖掘结果、非 Petri 网模型的决

策点挖掘等问题, 还需要更进一步的研究探索.

4 算法比较

前述各类流程挖掘算法的优缺点如表 2 所示.

直接算法以 α 算法及其系列算法[3–7] 代表, 在流程挖掘发展初期诞生, 发展成熟, 根据活动间依赖关系发现流程模型, 原理清晰, 计算简单, 但不能处理许多复杂结构和噪声, 后期虽有学者针对性改进, 但至今仍没有能够处理所有复杂结构的完善算法; 启发式算法[9,10] 解决了噪声处理问题, 基于活动间发生频率来去除噪声, 缺点是易误删低频有效行为, 导致挖掘模型不能反映实际情况; 基于概率统计的算法[13–17] 在处理特定问题时效果显著, 可用于流程优化或者改进算法, 缺点是适用范围不大, 依附性强; 基于智能优化的流程挖掘算法[20–23] 可以处理大多数的控制流结构和噪声, 有利于得到全局的优化结果, 通过设置适应度函数便于得到倾向性模型, 不足在于算法效率太低, 容易产生死锁等问题, 可在质量和效率两方面进行改进; 模糊挖掘[24] 可用于处理现实生活中“意大利面式”的流程模型, 但需要获得合适参数来进行过滤, 可使用性较差; 决策挖掘[29] 关注案例视角, 发现流程走向的原因, 便于研究人员分析, 但是易受事件日志影响.

流程挖掘算法各类算法都有其优越性, 可解决特定领域的特定问题, 但各算法也存在着明显短板, 改进难度大. 从整体上看, 流程挖掘算法发展呈现烟囱式, 目前并无一个算法可以解决绝大多数问题, 探索之路任重而道远.

表 2 流程挖掘算法特点

算法大类	算法小类	代表性算法	优点	缺点
传统算法	直接算法	α算法、α^+算法…	发展成熟, 原理清晰, 计算简单.	不能处理日志噪声, 不能处理多种复杂结构.
	启发式算法	启发式算法	能处理事件日志中的噪声.	容易误删低频有效行为.
	基于概率统计的算法	马尔可夫模、随机活动图算法…	可用于流程优化, 解决特定问题.	各有缺点, 适用性不高, 依附性强.
基于计算智能和机器学习的流程挖掘算法	基于智能优化的流程挖掘算法	遗传算法、模拟退火算法…	可以处理大多数的控制流结构, 可以处理噪声, 有利于得到全局优化结果, 便于得到更倾向性的流程模型.	算法运行效率低, 挖掘模型质量还可提高.
	模糊挖掘	模糊挖掘算法	可处理“意大利面式”的流程模型.	算法参数不易设置, 可使用性有待加强.
	决策挖掘算法	决策挖掘算法	关注流程挖掘的案例视角, 寻找决策原因.	易受事件日志质量影响.

5 总结

流程挖掘是对客观世界中的业务流程、工作流程、信息流程进行发现、验证和优化的有效手段, 目前已经成为系统工程和数据应用工程的研究热点. 流程挖掘算法是流程挖掘的核心, 直接影响流程挖掘效率和挖掘结果质量. 现有的流程挖掘算法中, 以直接算法、启发式算法和概率统计算法等为代表的传统算法具有执行效率高的优点, 但对日志噪声、复杂结构、有效低频行为等问题的处理上难以兼顾, 通常需要组合运用以达到理想效果; 以智能优化、模糊挖掘和决策挖掘等算法为代表的新型算法在处理上述问题方面具有独到的优势, 且更容易获得全局优化结果, 但执行效率会随着挖掘算法的复杂程度而降低, 在应用时需要在挖掘效率和挖掘质量上综合考虑. 随着计算机算力的不断提升, 挖掘算法复杂度对执行效率的制约将逐渐降低, 基于计算智能和机器学习的挖掘算法将以其高质量的挖掘效果成为今后流程挖掘算法的主要发展趋势.

参考文献

1 Aalst W. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Heidelberg: Springer Publishing, 2011.

2 李杰, 陈超美. CiteSpace: 科技文本挖掘及可视化. 北京: 首都经济贸易大学出版社, 2016.

3 van der Aalst W, Weijters T, Maruster L. Workflow mining: Discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1128–1142.

4 de Medeiros AKA, van Dongen BF, van der Aalst WMP, *et al*. Process mining for ubiquitous mobile systems: An overview and a concrete algorithm. Proceedings of the 2nd International Workshop on Ubiquitous Mobile Information and Collaboration Systems. Riga: Springer, 2005. 151–165.

5 Wen LJ, van der Aalst WMP, Wang JM, *et al*. Mining

process models with non-free-choice constructs. Data Mining and Knowledge Discovery, 2007, 15(2): 145–180.
6 Wen LJ, Wang JM, Sun JG. Mining invisible tasks from event logs. Proceedings of the 9th Asia-Pacific Web Conference. Huangshan: Springer, 2007. 358–365.
7 Li JF, Liu DY, Yang B. Process mining: Extending α-algorithm to mine duplicate tasks in process logs. Proceedings of the APWeb/WAIM 2007 International Workshops: DBMAN 2007, WebETrends 2007, PAIS 2007 and ASWAN 2007. Huangshan: Springer, 2007. 396–407.
8 李红. 流程挖掘算法研究 [博士学位论文]. 昆明: 云南大学, 2015.
9 Weijters AJMM, van der Aalst WMP. Rediscovering workflow models from event-based data using little thumb. Integrated Computer-Aided Engineering, 2003, 10(2): 151–162.
10 Weijters AJMM, van der Aalst WMP, de Medeiros AKA. Process mining with the HeuristicsMiner algorithm. Eindhoven: Eindhoven University of Technology, 2006. 1–34.
11 孙笑笑, 张蕾, 俞东进, 等. 基于依赖关联度的业务过程噪声日志过滤方法. 计算机集成制造系统, 2019, 25(4): 969–977.
12 杨丽琴, 康国胜, 蔡伟刚, 等. 业务流程挖掘算法研究. 计算机应用与软件, 2016, 33(4): 44–50.
13 Li Y, Feng YQ. An automatic business process modeling method based on Markov transition matrix in BPM. Proceedings of 2006 International Conference on Management Science and Engineering. Lille: IEEE, 2006. 46–51.
14 Herbst J, Karagiannis D. Workflow mining with InWoLvE. Computers in Industry, 2004, 53(3): 245–264.
15 Sun WX, Li T, Peng W, *et al*. Incremental workflow mining with optional patterns. Proceedings of 2006 IEEE International Conference on Systems, Man and Cybernetics. Taipei: IEEE, 2006. 2764–2771.
16 Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers Inc., 1994. 487–499.
17 Ye YB, Chiang CC. A parallel apriori algorithm for frequent itemsets mining. Proceedings of the 4th International Conference on Software Engineering Research, Management and Applications (SERA'06). Seattle: IEEE, 2006. 87–94.
18 卢盛祺, 李远刚, 管连, 等. 流程挖掘在银行服务管理中的应用. 微型机与应用, 2016, 35(18): 88–92.
19 Lemos AM, Sabino CC, Lima RMF, *et al*. Using process mining in software development process management: A case study. Proceedings of 2011 IEEE International Conference on Systems, Man, and Cybernetics. Anchorage: IEEE, 2011. 1181–1186.
20 de Medeiros AKA, Weijters AJMM, van der Aalst WMP. Genetic process mining: An experimental evaluation. Data Mining and Knowledge Discovery, 2007, 14(2): 245–304.
21 陈希. 一种工作流的挖掘算法研究 [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2016.
22 白雪骢, 朱焱. 一种基于禁忌搜索算法的流程挖掘方法. 计算机科学, 2016, 43(4): 214–218, 240.
23 李洪奇, 李莉, 谢绍龙. 基于粒子群优化算法的过程挖掘. 2008 中国计算机大会论文集. 西安: 中国计算机学会, 2008. 338–345.
24 Günther CW, van der Aalst WMP. Fuzzy mining—Adaptive process simplification based on multi-perspective metrics. Proceedings of the 5th International Conference on Business Process Management. Brisbane: Springer, 2007. 328–343.
25 周逸璇. 基于日志抽象的流程挖掘方法研究 [硕士学位论文]. 昆明: 云南大学, 2012.
26 Ardimento P, Bernardi ML, Cimitile M, *et al*. Learning analytics to improve coding abilities: A fuzzy-based process mining approach. Proceedings of 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). New Orleans: IEEE, 2019. 1–7.
27 Bernardi ML, Cimitile M, Martinelli F, *et al*. A fuzzy-based process mining approach for dynamic malware detection. Proceedings of 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Naples: IEEE, 2017. 1–8.
28 Mitchell T M. Machine Learning. New York: McGraw-Hill, 1997.
29 Rozinat A, van der Aalst WMP. Decision mining in business processes. Russian Mathematical Surveys, 2006, 38(3): 23–95.
30 高昂, 杨扬, 张静乐, 等. 基于工作流日志的决策规则挖掘研究. 计算机应用研究, 2009, 26(11): 4104–4107.
31 de Leoni M, van der Aalst WMP. Data-aware process mining: Discovering decisions in processes using alignments. Proceedings of the 28th Annual ACM Symposium on Applied Computing. Coimbra: ACM, 2013. 1454–1461.
32 瞿华. 一种基于过程挖掘的业务过程决策规则发现算法. 计算机应用研究, 2012, 29(6): 2192–2195.