DOI: 10.16452/j.cnki.sdkjzk.2023.01.013 文章编号:1672-3767(2023)01-0118-11

基于过程挖掘的分层学习行为模型发现方法

李金鹏[1],刘 聪[1],李会玲[1],王 颖[1],曾庆田[2],张 峰[2]
(1.山东理工大学 计算机科学与技术学院,山东 淄博 255030;
2.山东科技大学 计算机科学与工程学院,山东 青岛 266590)

摘 要:教育过程挖掘将过程挖掘技术应用于教育数据分析,是教育数据挖掘的重要分支之一。当前教育数据挖掘主要是用经典的机器学习算法对在线学习数据进行建模分析,难以描述全局的学习过程。凭借解析事件日志发现控制流模型的过程挖掘技术可以解决这一难题,但由于真实数据受各种客观因素的影响,存在大量噪声和无关行为,已有的挖掘方法往往会生成"意大利面模型",不利于分析理解。针对这一问题,本研究提出分层过程挖掘方法发现学生学习过程。具体方法是:通过解析带生命周期事件日志的时间属性,发现活动嵌套关系;然后构造分层事件日志,进而挖掘描述学习行为的分层过程模型;最后用契合度、精确度、F-值三个指标,系统地比较分层过程挖掘方法与已有过程挖掘方法所挖掘模型的区别。

关键词:教育数据挖掘;过程挖掘;分层 Petri 网;学习行为分析
中图分类号:G434,TP181 **文献标志码**:A

Discovery method of hierarchical learning behavior model based on process mining

LI Jinpeng[1], LIU Cong[1], LI Huiling[1], WANG Ying[1], ZENG Qingtian[2], ZHANG Feng[2]
(1. School of Computer Science and Technology, Shandong University of Technology, Zibo 255030, China;
2. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: Educational process mining applies the process mining technology to educational data analysis and it is an important branch of education data mining. Mainly using classic machine learning algorithms to establish models for the analysis of the student online learning data, the current education data mining is difficult to describe the global student learning process. This problem can be solved by the process mining technology used to discover the control flow model by analyzing event logs. But the existence of noise and irrelevant behavior sequences in real data affected by various factors tends to make the existing mining methods generate a "Spaghetti model" and thus are not conducive to analysis and understanding. In view of this problem, this paper proposed the Hierarchical Process Mining (HPM) algorithm to discover the student learning behavior process. Firstly, the time attributes of the life-cycle event logs were analyzed to discover the nesting relationship of activities. Then, the layered event logs were constructed and the hierarchical process model for describing the student learning behavior was mined. Finally, the indicators of fitness, precision, and F-scores were used to make a systematical comparison between the mining models established by the layered process mining method and the existing process mining methods.

Key words: education data mining; process mining; hierarchical Petri net; learning behavior

收稿日期:2021-11-04
基金项目:国家自然科学基金项目(61902222);山东省泰山学者工程专项基金项目(ts20190936,tsqn201909109);山东省自然科学基金优秀青年基金项目(ZR2021YQ45);山东省高等学校青创科技计划创新团队项目(2021KJ031)
作者简介:李金鹏(1996—),男,山东淄博人,硕士研究生,主要从事过程挖掘等方向研究. E-mail:381425318@qq.com
刘 聪(1990—),男,山东淄博人,教授,博士生导师,主要从事业务过程管理、业务流程建模、流程挖掘、人工智能等研究,本文通信作者.E-mail:liucongchina@163.com

过程挖掘(process mining,PM)[1]技术是连接业务过程管理与数据挖掘之间的渠道,旨在发现大规模事件日志中的过程性相关信息,不仅可用于控制流发现与离线分析,亦可用于整合组织挖掘、社会网分析、决策挖掘、时间及结果预测等。自 2011 年电气与电子工程师协会智能计算学会过程挖掘工作组发布初版《过程挖掘宣言》以来,过程挖掘技术逐渐受到各领域的广泛关注。随着业务分工的逐渐标准化、精细化,对业务过程的分析需求与日俱增。此外,机器人流程自动化(robotic process automation, RPA)亟需通用的过程发现技术对接各行业的业务过程,这些客观需求使得过程挖掘技术在未来的普及与广泛应用成为可能。教育数据挖掘(education data mining, EDM)是一个跨学科的研究领域,旨在利用现有的数据挖掘技术从在线学习数据集中发现隐含的模式和信息,例如使用 Apriori 算法发现学生活动序列中存在的关联规则。教育过程挖掘的诞生打破了数据挖掘方法难以发现控制流(control flow)的壁垒,并引入教育过程的发现技术、合规性检查、支持与推荐技术,使得教育数据分析得到进一步扩展。

目前,学生学习过程的发现存在诸多挑战。从事件日志的角度来说,数据本身关键信息缺失会导致事件日志存在"不完备性"[2],且随着时间改变数据流会发生"概念漂移"[3]。从挖掘角度来说,算法本身"偏好"[1]会导致发现的控制流模型不够准确,不能契合学生的真实学习过程,使得下一步研究无法顺利进行。为了解决在挖掘角度上的这一问题,本研究引入能够准确发现学习行为模型的分层过程挖掘方法[4](hierarchical process mining, HPM),且根据学生在线行为事件日志的特点,提出分层过程挖掘算法时间关系(HPM time relation,HPMTR)。此外,本研究所提算法加入时间维度,以解决先前分层过程挖掘算法在发现不同活动之间上下层次关系的不足。

1 相关研究工作

1.1 过程挖掘

过程挖掘是从事件日志数据中发现过程性信息的一种新技术,包含业务过程的发现、合规性检查与改进,是业务过程管理(business process management,BPM)[5]中的一个热点课题,其首要研究是发现过程的控制流结构,即通过解析事件日志中的事件发生信息构造一个能够描述事件之间因果依赖关系的过程模型。过程模型的结构主要由 Petri 网[6]、BPMN(business process model notations)[7]、EPC(event-driven process chain)[8]、YAWL(yet another workflow language)[9]等来描述。过程模型的主要评价指标是契合度(fitness),用以衡量控制流模型解释事件日志的能力。由于契合度与其他指标存在相互竞争关系,因此通常引入多个指标衡量模型,例如精确度(precision),表示模型的行为记录在日志中的比例。目前过程挖掘技术存在诸多挑战,特别是在如何挖掘精确模型与提高挖掘效率等方面[10-13]。

1.2 教育过程挖掘

大规模开放在线课程(massive open online courses,MOOCs)、学习管理系统(learing management system,LMS)、超媒体及其他在线的学习环境为在线社区提供了免费的学习机会。这些系统的日志数据记录了学生如何跟进课程、如何观看视频与讲座、如何处理学习任务。不同的教育信息系统存储信息的方式与类别不同,能够探讨的问题也不同。根据不同在线学习平台的特点,教育过程挖掘在不同教学环境中进行了不同探索。文献[14]讨论过程挖掘在 LMS 中提取信息的方法,特别是学生在考试过程中的轨迹;文献[15]跟踪超媒体环境中的自主管理学习规则(self-regulated learning, SRL);文献[16]监测 SRL 中不同类型事件的发生频率,证明完成课程的学生有更多学习相关的事件;文献[17]跟踪和分析学生在 MOOCs 平台的学习习惯,发现完成课程的学生遵循有规则的序列模式,而其他学生是不可预测的非结构化的过程。

过程挖掘技术面向教育领域的研究目前尚处在初级阶段,研究方法大体可分为三类。第一类是直接将过程挖掘方法引入。例如文献[18]将时序逻辑检查技术引入教育数据分析领域,并在文献[19]中描述了生成案例的方法,用轨迹对齐的方式发现学生学习模式;文献[20]把评论字数、观看时长、成绩等参数相近的学生进行分组,挖掘不同类型学生的学习过程;为了进一步拓展教育过程挖掘的应用范围,文献[21]将在线阅读行为事件日志提升为认知行为事件日志,再用模糊挖掘(fuzzy miner)发现学生在线阅读的认知过程,提出一种验证认知心理学理论模型的手段,使得过程挖掘技术的应用范围得到了极大的扩展。第二类方法是将

过程挖掘技术进行扩展。如文献[22]借鉴 OLAP(on-line anaytical processing)立方体,从扩展散点图的维度,借助"过程立方体"分组事件日志,建立可比较的过程模型(comparative process model),最后分析通过考试和未通过考试的学生在学习过程上的区别。第三类方法是将过程挖掘方法与传统机器学习模型联结。文献[23]引入决策树算法与过程挖掘结合的方式,发现在 Moodle 平台上所需要遵循的隐含学习规则;文献[24]提出将过程相似度与支持向量机(support vector machine, SVM)结合,预测学生在在线判题(online judge,OJ)系统编写程序的表现,并对该方法的有效性进行评估。

1.3 当前存在问题及本研究工作

不同类型的信息系统存储的数据形式不同,因此教育过程挖掘需要解决的首要问题是保证输入日志的正确性,即如何处理数据使其能够达到过程挖掘的输入条件。本研究以学生在线学习《数字电路技术》的数据集为例,为了保证教育过程挖掘输入的正确性,通过四个处理步骤进行数据预处理。针对如何发现准确的控制流模型这一难点,本研究基于文献[4],发现事件日志中不同活动之间存在的上下层关系并进行日志分层,进一步挖掘分层 Petri 网;由于在线学习数据集中多个不同的上层活动嵌套相同的下层活动,导致客观上存在的上层活动嵌套下层活动的发生"频率"过低,很难识别上层活动与其下层活动的对应关系,为了挖掘准确的控制流模型,本研究提出根据事件的时间信息发现上层活动与其下层活动之间"嵌套关系"的方法,并以此重构分层过程挖掘算法。

2 基本概念

事件日志是过程挖掘的输入,在研究型工作中,通常选择 Petri 网为过程挖掘的输出。

2.1 事件日志

定义 1 事件日志[1]。广义的事件日志是一个事件集合,每个事件除了任务属性还可以包括其他属性,如活动标识、名称、时间戳、持续时间、生命周期信息、资源等。令 ε 为事件集,N 为事件的属性集合,对于 $n\in N$,任意一个事件 $e\in\varepsilon$,$\#_n(e)$ 表示属性 n 的值。U_C 为案例集,U_A 为任务集,U_L 为生命周期类型集,U_T 为时间集。假设任意一个事件 $e\in\varepsilon$ 都包含以下属性:$\#_{case}(e)\in U_C$ 为事件 e 所属的案例,每个事件只能属于一个案例;$\#_{act}(e)\in U_A$ 为事件 e 活动名;$\#_{trans}(e)\in U_L$ 为事件 e 的生命周期相关信息;$\#_{time}(e)\in U_T$ 表示事件 e 发生的时间戳。

定义 2 分类器[1]。分类器用于标识事件,定义为 $C:\varepsilon\rightarrow U_A\times U_L$,$\forall e\in\varepsilon$,$C(e)=(\#_{act}(e),\#_{trans}(e))$。若 $\#_{act}(e)=a$,且生命周期的取值范围为 $\#_{trans}(e)\in\{start, complete\}$,则该分类器将$(a, start)$标记为 a_s,将$(a, complete)$标记为 a_c。

定义 3 案例[1]。案例是有限的事件发生序列 $\sigma\in\varepsilon^*$,令 ε^* 是定义在集合 ε 上所有任意长度有限序列的集合,满足:每个事件仅可以出现一次,即 $1\leqslant j<k\leqslant|\sigma|:\sigma(j)\neq\sigma(k)$;并且在每一个案例中的所有事件具有相同的案例属性,即 $\forall e_1, e_2\in\sigma:\#_{case}(e_1)=\#_{case}(e_2)$。

事件日志是一个有限事件发生序列的集合,定义为 $L\subseteq\varepsilon^*$,活动序列相同的多个案例可以用同一条轨迹[1]描述。一个学生在线学习行为的活动序列可以表示为一个案例。

定义 4 多集[1]。多集是元素可以重复出现多次的集合,设 $m=[p^3, q^2]$是定义在集合 $S=\{p, q\}$上的多集,其中 $m\in B(S)$,$m(p)=3$,$m(q)=2$ 表示元素出现次数,$B(S)$表示集合 S 多集的全集。多集不仅可用来表示标识,还用于对事件日志的建模。一条轨迹在不同案例中出现,可以用多集简化表达。

2.2 Petri 网

定义 5 Petri 网[1]。Petri 网用于表示事物的变化过程,其最基本的形式是 $N=(P, T, F)$。其中:T 表示变迁(transition)的有限集合,P 表示库所(place)的有限集合,且 $P\cap T=\varnothing$;$F\subseteq((P\times T)\cup(T\times P))$ 为弧的集合,表示控制流。图 1 所示 Petri 网的简单实例表示为:$F=\{(start, a), (a, p1), (a, p2), (p1, b), (p2, c), (b, p3), (c, p4), (p3, d), (p4, d), (d, end)\}$;$T=\{a, b, c, d\}$;$P=\{start, p1, p2, p3, p4, end\}$。标识 Petri 网$(N, M)$为一个二元对,一个标识(marking)对应由"托肯"(token)构成的多集,token 用黑点表示,$M\in B(P)$表示在库所 P 上定义多集,代表库所中的标识。

定义 6 发生规则[1]。发生规则是指 Petri 网发生变化的条件,对于 $\forall x \in P \cup T, {}^{\cdot}x = \{y \mid y \in P \cup T \wedge (y, x) \in F\}$ 为 x 的前集(pre-set),$x^{\cdot} = \{y \mid y \in P \cup T \wedge (x, y) \in F\}$ 为 x 的后集(post-set)。在包含标识 m 的 Petri 网 N 中,变迁 t 满足任意一个库所 $p \in {}^{\cdot}t$: $m(p) \geqslant 1$,则变迁 t 处于使能状态,可以引发并产生新的标识,记为 $(N, m)[t>(N, m')$。如果一个变迁的输入库所含有足够托肯,消耗该变迁所有输入库所中的一个托肯,所有输出库所中就产生新的托肯。图 1 假设初始标识的多集为 $[start^4]$,a 发生后得到 $[start^3, p1, p2]$,a 仍处于使能状态,再次发生后得到 $[start^2, p1^2, p2^2]$,a 连续发生 4 次后得到 $[p1^4, p2^4]$,同时 b、c 为使能。

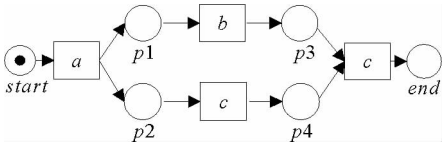

图 1 Petri 网的简单实例

Fig. 1 A simple example of Petri net

2.3 过程发现

过程模型的发现算法主要有四类。第一类算法直接解析事件与事件之间的因果、并发等次序关系并构建过程模型,主要算法有 Alpha Miner[25] 以及通过依赖度量的启发式发现算法 Heuristic Miner[26];第二类是两阶段方法,在低级模型(有向图、马尔可夫模型)的基础上转化为能够描述并发之类机制的高级模型,如 Two-phase Miner[27];第三类是智能计算方法,借助机器学习、神经网络等模拟模型演化过程,如 Genetic Miner[28]、Evolutionary Tree Miner(ETM)[29];第四类是发现规则和频繁模式的局部方法,类似方法可以用于学习基于时序逻辑语言的声明式模型,如 Declare Miner[30]。

3 分层学习行为模型发现方法

分层学习行为模型的发现以带生命周期的学生学习行为事件日志为输入,通过解析事件包含的时间信息,进一步发现活动之间的嵌套关系,并以此构造任务嵌套关系树,用以描述上层活动与其下层活动之间的对应关系,从而根据任务嵌套关系树与嵌套活动发生频率构造分层学习行为事件日志,进而通过分层事件日志分别发现上层过程模型及每个上层变迁所调用的下层过程模型。其中,任务嵌套关系树与学生行为事件日志同步为分层日志构造的输入。文献[4]所提分层过程挖掘方法需要判断嵌套活动关系在两个活动总顺序关系中的发生频率是否达到阈值。由于在学习行为数据中,客观上存在的嵌套活动关系远小于活动顺序关系总数之和,需要不断调试 plugin 中的 nest 值,且易出现将当前活动的开始事件与下一同名活动的结束事件误判为属于同一个活动的情况。基于上述问题,本研究按照活动的开始和结束时间判断嵌套关系,同时根据同一个活动的事件实例(instance)在 XES 文件中"concept:instance"值相同这个特点来判定不同事件是否属于同一个活动,进而实现对学习行为事件日志的分层。本研究提出的分层学习行为模型发现方法架构如图 2 所示。

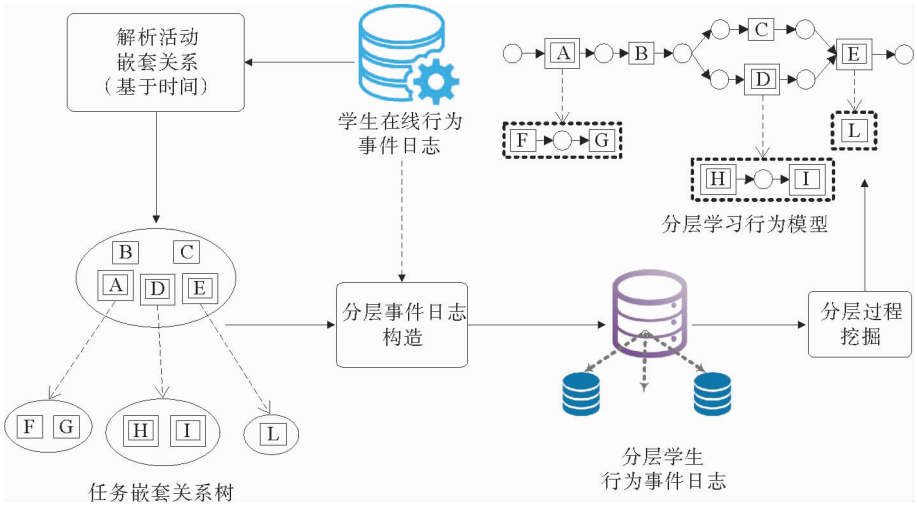

图 2 分层学习行为模型发现架构

Fig. 2 Discovery frame of the hierarchical learning behavior model

3.1 定义活动顺序关系

定义 7 基于时间的活动顺序关系[4]。活动顺序关系如图 3 所示,活动顺序关系用于判断活动之间是否嵌套,令 $L\in \mathrm{B}((U_{\mathrm{A}}\times U_{\mathrm{L}})^{*})$ 为带生命周期的事件日志,对于 $\forall a,b\in U_{\mathrm{A}},\sigma\in L$,$a$ 与 b 为紧邻关系记作:$a\lambda b$;a 与 b 为重合关系记作 $a\Delta b$;a 与 b 为包含关系记作 $a\Omega b$;a 与 b 存在嵌套关系记作 $a\Theta b$。若存在 $i,j,k,l\in\{1,2,\cdots,|\sigma|\}$,$i<j<k<l$,令 $\#_{\mathrm{time}}(\sigma(i))<\#_{\mathrm{time}}(\sigma(j)))<\#_{\mathrm{time}}(\sigma(k))<\#_{\mathrm{time}}(\sigma(l))=TimeCorrect$;$C(e)=(\#_{\mathrm{act}}(e),\#_{\mathrm{trans}}(e))$,$instance(e)=C(e)$;$ConceptInstance(e)=e.\mathrm{concept:instance.id}$。则上述四种关系可形式化定义为:

$$a\lambda b\leftrightarrow instance(\sigma(i))=a_{\mathrm{s}}\wedge instance(\sigma(j))=a_{\mathrm{c}}\wedge instance(\sigma(k))=b_{\mathrm{s}}\wedge instance(\sigma(l))=b_{\mathrm{c}}\wedge TimeCorrect,\tag{1}$$

$$a\Delta b\leftrightarrow instance(\sigma(i))=a_{\mathrm{s}}\wedge instance(\sigma(j))=b_{\mathrm{s}}\wedge instance(\sigma(k))=a_{\mathrm{c}}\wedge instance(\sigma(l))=b_{\mathrm{c}}\wedge TimeCorrect,\tag{2}$$

$$a\Omega b\leftrightarrow instance(\sigma(i))=a_{\mathrm{s}}\wedge instance(\sigma(j))=b_{\mathrm{s}}\wedge instance(\sigma(k))=b_{\mathrm{c}}\wedge instance(\sigma(l))=a_{\mathrm{c}}\wedge TimeCorrect,\tag{3}$$

$$a\Theta b\leftrightarrow\neg(a\Delta b)\wedge\neg(b\Delta a)\wedge(a\Omega b)\wedge\neg(b\Omega a)\wedge(a\lambda b)\wedge\neg(b\lambda a)\wedge ConceptInstance(\sigma(i))=ConceptInstance(\sigma(l))。\tag{4}$$

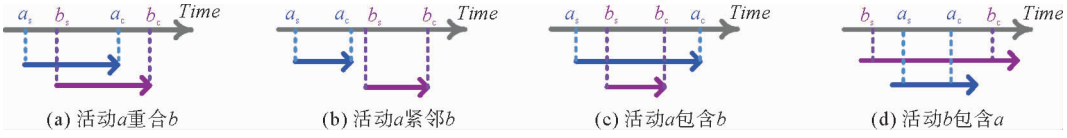

图 3 活动顺序关系

Fig. 3 Relationship of activity order

定义 8 $(A,rootAct,\eta)$ 为 L 的任务嵌套关系树[4],用于描述上层活动与其下层活动的对应关系,其中 A 为活动集合,$rootAct$ 为根节点集。$rootAct\subseteq A$ 满足:

1) $r_1,r_2\in rootAct:\neg(r_1\Theta r_2)\wedge\neg(r_2\Theta r_1)$,意为根节点不存在嵌套关系。

2) $\forall r\in rootAct,\neg\exists a\in A,rootAct:a\Theta r$,意为根节点不能嵌套其他根节点的活动。

$\eta:A\mid\rightarrow\mathrm{P}(A)$,$rootAct$ 为将活动映射到嵌套活动集合中的偏函数,满足:

1) $\forall a,b\in A,\eta(a)\cap\eta(b)=\varnothing$;

2) $\forall b,c\in\eta(a):\neg(b\Theta c)\wedge\neg(c\Theta b)$;

3) $\forall a,b\in A,\eta(a)\cap\eta(b)=\varnothing$。

图 2 所示任务嵌套关系树 A、B、C、D、E 是根节点且相互独立,其中只考虑 H、I 嵌套于 D 这种直接关系,不考虑 D 与 H、I 这种间接关系,间接或者传递嵌套关系被弱化。

3.2 分层事件日志构造

定义 9 分层事件日志[4]。分层事件日志是分层 Petri 网的输入,$(A,rootAct,\eta)$ 为带生命周期事件日志 L 的嵌套关系树,A 为事件日志 L 的活动集合,$HL(L)=(rootLog,HL(rootLog))$ 为 L 的分层事件日志。其中,$rootLog=\bigcup_{\sigma\in L}\sigma\uparrow rootAct$ 为 L 中根节点构成的事件日志,$HL(rootLog)$ 为 $rootLog$ 的分层日志满足:

1) 如果 $NA(rootLog)=\varnothing$,$HL(rootLog)=\varnothing$,其中 $NA(rootLog)$ 为根节点日志中的嵌套活动集合;否则,$HL(rootLog)=\{(na,NLog_{na},HL(rootLog_{na}))\mid na\in NA(rootLog)\}$;

2) $NLog_{na}=\bigcup_{\sigma\in L}\sigma\uparrow\eta$ 为 L 中 na 所映射的下层事件日志。

例如:多集 $L=[\langle a_{\mathrm{s}},b_{\mathrm{s}},b_{\mathrm{c}},c_{\mathrm{s}},c_{\mathrm{c}},a_{\mathrm{c}}\rangle^{10}]$ 为带生命周期的事件日志,若对于日志 L 中任意一个 a_{s} 的“concept:instance”值与 a_{c} 相同,则 a_{s}、a_{c} 属于同一个活动 a 的生命周期;若同时满足任意 a_{s} 的实际发生时间早于 b_{s}、c_{s},任意 a_{c} 的实际发生时间晚于 b_{c}、c_{c},则活动 b、c 嵌套于 a,b、c 看作是 a 的下层活动;若上层日

志为 $rootLog=[\langle a_s, a_c\rangle^{10}]$,则嵌套活动集合 $NA(rootLog)=\{a\}$;a 的下层日志为 $NLog_a=[\langle b_s, b_c, c_s, c_c\rangle^{10}]$。

3.3 分层 Petri 网

定义 10 分层 Petri 网[4]。分层 Petri 网用于描述分层过程,用一个二元组 $HPN=(PN_{n0}, HPN(PN_{n0}))$表示,其中 PN_{n0}是根节点,即顶层包含嵌套变迁的 Petri 网;若 $T_n=\{t\in T \mid N(t)=N\}$是嵌套变迁的集合,且存在 $HPN=HPN(PN_{n0})$ 是 PN_{n0}对应的下层模型,则满足:

1) 如果存在 $T_{n0}=\varnothing$,$HPN(PN_{n0})=\varnothing$;否则 $HPN(PN_{n0})=\{(t_i, PN_{ni}, HPN(PN_{ni})) \mid t_i\in T_{n0}\}$;

2) 其中 PN_{ni}是嵌套任务 t_i 对应的带嵌套变迁的 Petri 网,T_{n0}是 PNn 中所有嵌套变迁的集合。

图 4 给出一个分层 Petri 网的例子,顶层 Petri 网包含三个普通变迁和一个嵌套变迁 t_0,嵌套变迁 t_0 对应下层过程模型 PN_{n1},下层过程模型中包含一个嵌套变迁 t_1,嵌套变迁 t_1 对应下层过程模型 PN_{n2}。

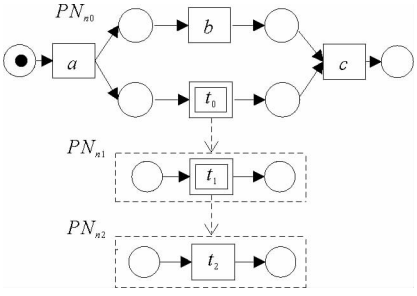

图 4 分层 Petri 网

Fig. 4 Hierarchical Petri net

4 教育过程挖掘及模型评估

将原数据处理为带生命周期事件日志之后,导入开源框架 ProM[30],调用扩展之后的 HPMTR,并在同等条件下与其他算法的性能进行对比。ProM 是一个集成了过程挖掘方法的开源框架,特点是易于扩展,现有过程挖掘技术几乎都可以在 ProM 原有的功能上进行扩展。

4.1 学习行为事件日志处理

实验数据集出自文献[31],记录了 109 名学生学习数据,包含 "Deed""TextEditor""Diagram""Properties""Study_Materials""FSM""Aulaweb""Blank""Other"等活动。其中:"Deed"表示数字电路技术的教学与控件设计(digital electronics education and design suite),"Diagram"表示仿真计时图设计(simulation timing diagram),"FSM"表示有限自动机仿真设计(finite state machine simulator),"Aulaweb"表示登录学习管理系统后并评论,或者下载学习资料,"Other"表示在线访问与学习无关的内容。课程共分为 6 章,每章包含 4~6 节,每周学习一章课程并完成章节测验,课程日期为 2014 年 10 月 2 日~12 月 11 日,每条记录包含活动信息、时间信息、鼠标位置信息等。预处理后的学生在线行为事件日志如表 1 所示。为了构造符合分层过程挖掘算法输入条件的学生在线行为事件日志,通过以下四步对数据集进行预处理。

表 1 预处理后的学生在线行为事件日志

Table 1 Pre-processed student online behavior event log

Session	Student Id	Activity	Start Time	End Time
5	1	Es_5_1	06:02.9	06:36.1
5	1	TextEditor	06:03.0	06:03.0
5	1	Other	06:04.0	06:04.0
5	1	Other	06:05.0	06:07.0
5	1	TextEditor	06:08.0	06:08.0
5	1	TextEditor	06:09.0	06:09.0
5	1	TextEditor	06:10.0	06:17.0
5	1	Other	06:18.0	06:19.0
5	1	TextEditor	06:20.0	06:36.0
5	1	Es_5_2	06:36.9	06:37.1
5	1	Study	06:37.0	06:37.0

续表 1

Session	Student Id	Activity	Start Time	End Time
5	1	Es_5_1	06:37.9	06:39.1
5	1	TextEditor	06:38.0	06:39.0
5	1	Es_5_2	06:39.9	06:43.1
5	1	Study	06:40.0	06:43.0
5	1	Es_5_1	06:43.9	06:51.1
5	1	Deeds	06:44.0	06:51.0
5	1	Es_5_2	06:51.9	06:52.1
5	1	Study	06:52.0	06:52.0

1) 输入实验数据集,将数据集中不相关的鼠标动作与存在空值的记录除去,再用查询语句将成绩表等进行“自然连接”;

2) 定义事件集合、日志集合、寄存事件的栈(stack);

3) 遍历每个案例中的每个事件,当要进栈的事件属于另一个章节时,所有事件出栈并写入新的日志;再把栈底事件的开始时间设为该章节的开始时间并减少 100 ms,栈顶事件结束时间为该小节的结束时间并增加 100 ms,写入新日志;

4) 将连续且重名活动简化为一个活动,设置该活动的开始时间是第一个事件的开始时间,活动的结束时间是最后一个事件的结束时间。

表 1 为学生在线行为事件日志在 MySQL 数据库中的数据保存形式,为了达到分层过程挖掘输入条件嵌入了上层事件。表 2 为在线行为事件日志的整体概况,其中案例数表示学生在线人数。

4.2 实验设计

首先,分别调用 Inductive Miner(IM)[2]、HPMTR 对比两者所挖掘的 Session5 学习过程模型在外观上的区别,并用常用的衡量指标评估挖掘算法的性能。图 5 为用 IM 算法挖掘的描述 Session5 在线学习过程的 Petri 网,其在挖掘效果上更接近于错综复杂且难以观测的“意大利面模型”。图 6 为分层过程挖掘算法发现学生学习行为模型的表现,分层模型能够轻易表达节与节之间的交互过程,且使得在线学习过程的发现与分析不再局限于扁平化的控制流模型。

表 2 学生在线行为事件日志的规模

Table 2 Size of the student online behavior event log

事件日志	案例数	事件数	事件类型数
Session1	77	71 836	28
Session2	82	83 014	38
Session3	87	60 412	48
Session4	99	83 036	40
Session5	91	71 918	42
Session6	84	108 564	42

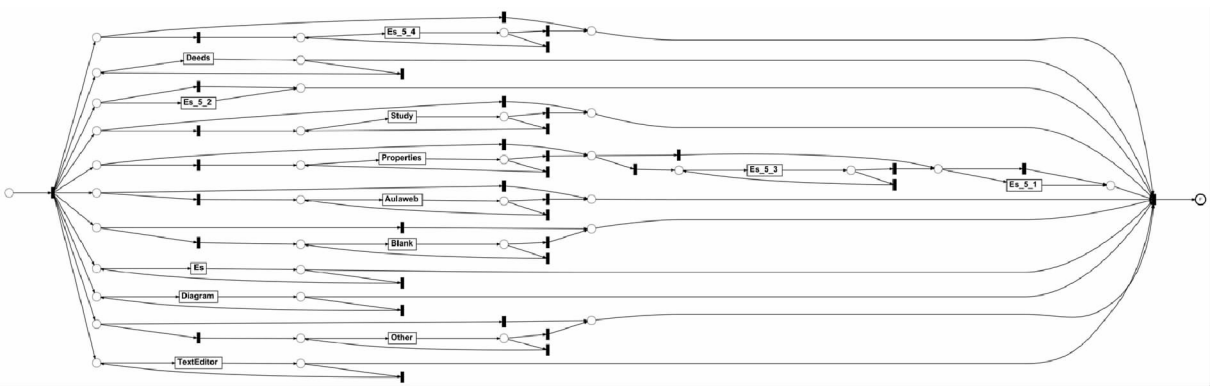

图 5 IM 算法挖掘的描述 Session5 在线学习过程的 Petri 网

Fig. 5 IM algorithm mining Petri net that describes the session5 online learning process

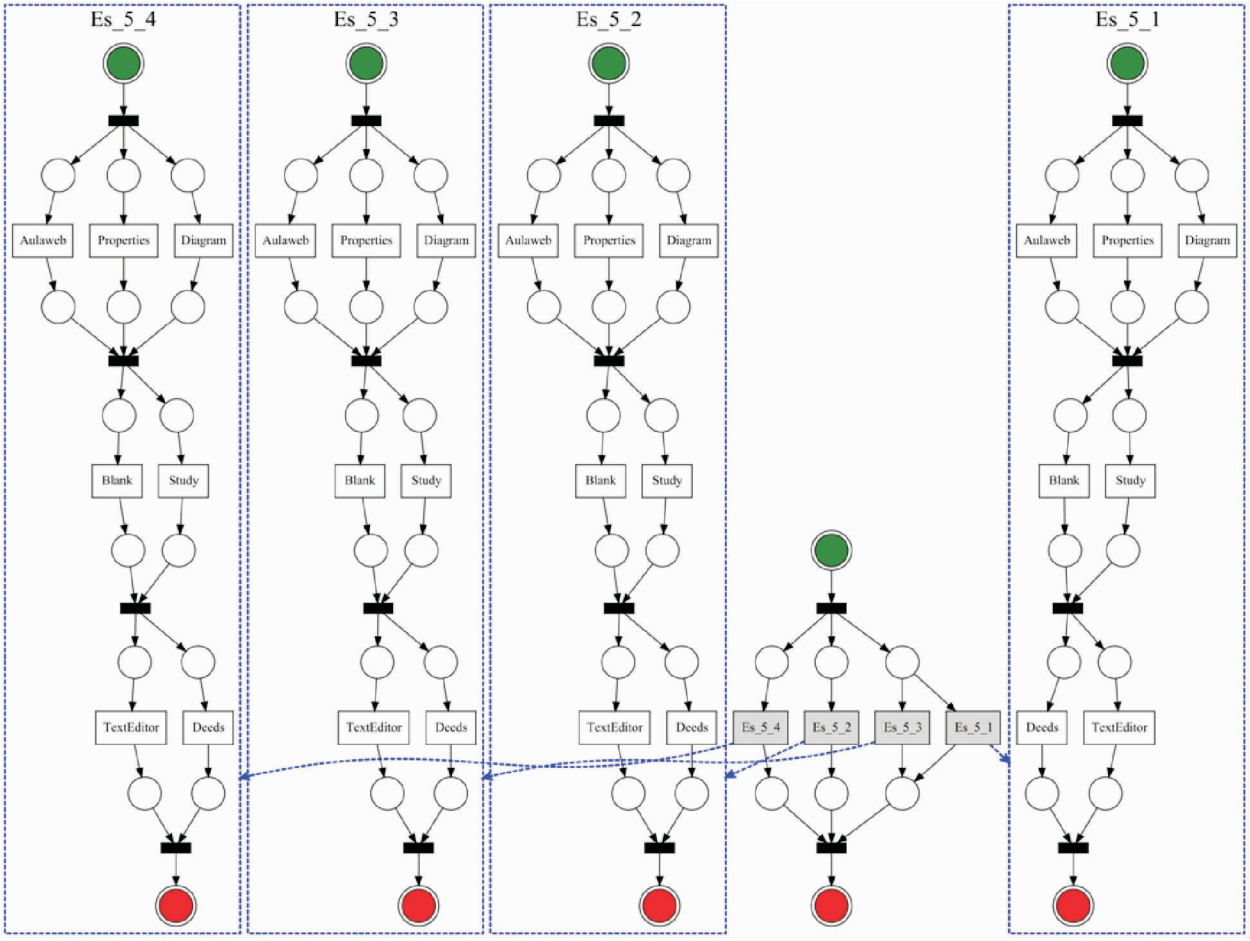

图 6　分层过程挖掘算法发现学生学习行为模型的表现

Fig. 6　Performance of HPM algorithm mining learning behavior model

由于 HPM 的工作原理是将原始日志分解为上层日志和下层日志,再分别挖掘上层过程模型及描述上层活动的下层过程模型,为了通过原始日志计算模型的契合度,将分层过程模型扁平化处理,处理方法如图 7 所示。

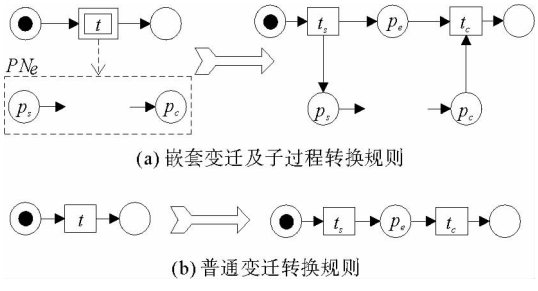

图 7　扁平化处理规则

Fig. 7　Rule of flattening hierarchical Petri Net

然后,调用 IM、IMLC(inductive miner life cycle)挖掘带有开始和结束变迁的 Petri 网并与分层过程掘发现的模型作对比。最后,引用 token 重演技术衡量模型契合度,得到该评估指标的形式化定义为:

$$fitness(\sigma,N)=\frac{1}{2}(1-\frac{m}{c})+\frac{1}{2}(1-\frac{r}{p})\ , \tag{5}$$

$$fitness(L,N)=\frac{1}{2}(1-\frac{\sum_{\sigma\in L}L(\sigma)\times m_{N,\sigma}}{\sum_{\sigma\in L}L(\sigma)\times c_{N,\sigma}})+\frac{1}{2}(1-\frac{\sum_{\sigma\in L}L(\sigma)\times r_{N,\sigma}}{\sum_{\sigma\in L}L(\sigma)\times p_{N,\sigma}})\ 。 \tag{6}$$

式中:p 为重演过程中产生的 token 数,c 为消耗的 token 数,m 为遗失的 token 数,r 为遗留的 token 数,L 代表事件日志,σ 代表轨迹。

精确度的形式化定义为:

$$precision = 1-\frac{|\ pref(M\backslash L)\ |}{|\ pref(M\backslash L)\ |+|\ pref(M\cap L)\ |}。\tag{7}$$

式中:$pref(X)$代表 X 集合所包含的前缀集合,M 代表模型可以描述的轨迹集合,$pref\ (M\backslash L)$ 为模型与日志相悖的轨迹前缀集合。

由于精确度与契合度之间存在相互竞争,为了得到更公正的评价指标,通常计算

$$F\text{-}score=\frac{2\times precision\times fitness}{precision+fitness}。\tag{8}$$

对比结果如表 3 所示,HPM 的契合度略低于 IM,但精确度、*F-score* 却略高于 IM。多数情况下,HPM 的契合度、精确度、*F-score* 优于 IMLC。虽然 IM 的 *fitness* 总体上略高,但分层过程挖掘得到的学生学习行为模型却有着两者在挖掘效果上无法比拟的模型表现能力,因为分层挖掘要将事件日志分层,所以复杂程度更小,更易进行过程梳理,且能够发现层次信息。

表 3 三种算法的挖掘性能对比

Table 3 Performance comparison of three mining algorithms

事件日志	IM			IMLC			HPM		
	fitness	*precision*	*F-score*	*fitness*	*precision*	*F-score*	*fitness*	*precision*	*F-score*
Session1	1.00	0.06	0.11	0.50	0.07	0.12	0.99	0.07	0.13
Session2	1.00	0.04	0.07	0.50	0.06	0.10	0.99	0.07	0.13
Session3	1.00	0.07	0.13	1.00	0.07	0.13	0.99	0.10	0.18
Session4	1.00	0.08	0.14	0.99	0.13	0.22	0.99	0.14	0.24
Session5	1.00	0.07	0.13	0.96	0.07	0.13	0.99	0.09	0.16
Session6	1.00	0.04	0.07	0.50	0.06	0.10	0.99	0.07	0.13

5 总结与展望

本研究根据解析学生在线行为事件日志中上层活动与下层活动的时间信息得到“嵌套关系”,从而提出挖掘分层学习行为模型的分层过程挖掘算法时间关系。将模型扁平化处理之后,引入契合度、精确度、F-值等衡量指标评估所挖掘学生学习行为模型的质量。

本研究的工作只是教育过程挖掘的第一步,教育过程挖掘非仅限于学生行为模型挖掘,教育过程挖掘可以在学生学习行为模型的基础上对学生的实时在线行为进行诊断,检测学生在线行为是否违规。除此之外,教育过程挖掘还可以结合学习行为模型进行学习过程推荐,当实时学生行为诊断到与模型发生“偏移”时,在优秀学生轨迹中搜索到与当前实时轨迹相似度最高的并推荐给该学生。后续的研究工作将着重于教育过程的预测、诊断及推荐。

参考文献:

[1] AALST W M P.Process mining:Making knowledge discovery process centric[J].ACM SIGKDD Explorations Newsletter,2012,13(2):45-49.

[2] YANG H D,WEN L J,WANG J M.An approach to evaluate the local completeness of an event log[C/OL]// 2012 IEEE 12th International Conference on Data Mining.Brussels,Dec. 10-13,2012:1164-1169.DOI:10.1109/ICDM.2012.66.

[3] BOSE R,AALST W M P.Handling concept drift in process mining[C]// International Conference on Advanced Information Systems Engineering.Berlin,Heidelberg:Springer,2011:391-405.

[4] 刘聪,程龙,曾庆田,等.基于 Petri 网的分层业务过程挖掘方法[J].计算机集成制造系统,2020,26(6):1525-1537.
LIU Cong,CHENG Long,ZENG Qingtian,et al.Petri net-based hierarchical business process mining[J].Computer Inte-

grated Manufacturing Systems,2020,26(6):1525-1537.

[5] SCHEER A,NUTTGENS M.ARIS architecture and reference models for business process management[C/OL]//Business Process Management,Models,Techniques,and Empirical Studies.Berlin,Heidelberg:Springer,2000:376-389.DOI:10.1007/3-540-45594-9_24.

[6] 吴哲辉.Petri 网导论[M].北京:机械工业出版社,2006:49-145.

[7] ROSING M V,WHITE S,CUMMINS F,et al.Business process model and notation-BPMN[M].Berlin Heidelberg:Springer,2015:142-159.

[8] SARSHAR K,LOOS P.Comparing the control-flow of EPC and Petri net from the end-user perspective[C]//International Conference on Business Process Management.Berlin Heidelberg:Springer,2005:434-439.

[9] AALST W M P,ALDRED L J,DUMAS M,et al.Design and implementation of the YAWL system[C]//16th International Conference on Advanced Information Systems Engineering.Berlin,Heidelberg:Springer,2004:211-247.

[10] LIU C,PEI Y L,CHENG L,et al.Sampling business process event logs using graph-based ranking model[J/OL].Concurrency and Computation:Practice and Experience,2021,33(5).DOI:10.1002/cpe.5974.

[11] 林蕾蕾,闻立杰,周华,等.基于日志完备性的过程漂移检测方法[J].计算机集成制造系统,2019,25(4):873-881.
LIN Leilei,WEN Lijie,ZHOU Hua,et al.Process concept drift detection approach based on log completeness[J].Computer Integrated Manufacturing Systems,2019,25(4):873-881.

[12] MAGGI F M,CICCIO C D,FRANCESCOMARINO C D,et al.Parallel algorithms for the automated discovery of declarative process models[J].Information Systems,2018,74:136-152.

[13] LIU C,LI H L,ZENG Q T,et al.Cross-organization emergency response process mining:An approach based on Petri Nets[J/OL].Mathematical Problems in Engineering,2020.DOI:10.1155/2020/8836007.

[14] TRCKA N,PECHENIZKIY M,AALST W M P.Process mining from educational data[C/OL]//Handbook of Educational Data Mining,2011:123-142.DOI:10.1201/610274-11.

[15] SAINY J,FAN Y,SINGH S,et al.Using process mining to analyse self-regulated learning:A systematic analysis of four algorithms[C/OL]//11th International Learning Analytics and Knowledge Conference.Irvine,Apr. 12-16,2021:333-343.DOI:10.1145/3448139.3448171.

[16] KIM K,MOON N.A model for collecting and analyzing action data in a learning process based on activity theory[J].Soft Computing,2018,22(20):6671-6681.

[17] ANTOINEV B,JOOS B,AALST W M P.Analysing structured learning behaviour in massive open online courses (MOOCs):An approach based on process mining and clustering[J].International Review of Research in Open and Distributed Learning,2018,19(5):37-60.

[18] MUKALA P P,BUIJS J J,AALST W M P.Uncovering learning patterns in a MOOC through conformance alignments[J].Information Systems,2015,(7):1509-2529.

[19] AALST W M P,GUO S,GORISSEN P.Comparative process mining in education:An approach based on process cubes[C]//3rd International Symposium on Data-Driven Process Discovery and Analysis.Berlin,Heidelberg:Springer,2013:110-134.

[20] BANNERT M,REIMANN P,SONNENBERG C.Process mining techniques for analysing patterns and strategies in students' self-regulated learning[J].Metacognition & Learning,2014,9:161-185.

[21] CAIRNS A H,GUENI B,ASSU J,et al.Analyzing and improving educational process models using process mining techniques[C]//The 5th International Conference on Advances in Information Mining and Management.Brussels,2015:17-22.

[22] BOGARIN A,ROMERO C,CEREZO R,et al.Clustering for improving educational process mining[C/OL]//Proceedings of the 4th International Conference on Learning Analytics and Knowledge:Indianapolis,Mar. 24-28,2014:11-15.DOI:10.1145/2567574.2567604.

[23] BAO Y X,LU F M,WANG Y,et al.Student performance prediction based on behavior process similarity[J].Chinese Journal of Electronics,2020,29(6):1110-1118.

[24] LEEMANS S J J,FAHLAND D,AALST W M P.Discovering block-structured process models from event logs:A constructive approach[C]//Proceedings of 34th International Conference on Application and Theory of Petri Nets and Concurrency.Berlin,Heidelberg:Springer,2013:311-329.

[25] WEIJTERS A J M M,RIBEIRO J T S.Flexible heuristics miner (FHM)[C]// 2011 IEEE Symposium on Computational Intelligence and Data Mining.New York:IEEE,2011:310-317.

[26] AALST W M P,RUBIN V,VERBEEK H M W,et al.Process mining:A two-step approach to balance between underfitting and overfitting[J].Software and Systems Modeling,2010,9:87-111.

[27] DE MEDEIROS A K A,WEIJTERS A J M M,AALST W M P.Genetic process mining:An experimental evaluation[J].Data Mining Knowledge Discovery,2007(14):245-304.

[28] BUIJS J C,VAN DONGEN B F,AALST W M P.Quality dimensions in process discovery:The importance of fitness,precision,generalization and simplicity[J].International Journal of Cooperative Information Systems,2014,23.DOI:10.1142/S0218843014400012.

[29] MAGGI F M,MMOOIJ A J,AALST W M P.User-guided discovery of declarative process models[C]// Proceedings of 2011 IEEE Symposium on Computational Intelligence and Data Mining.New York:IEEE,2011:192-199.

[30] VERBEEK H M V,BUIJS J C A M,DONGEN B F V,et al.ProM 6:The process mining toolkit[C]// The Business Process Management 2010 Demonstration Track.Hoboken,Sep. 14-16,2010,615:34-39.

[31] VAHDAT M,ONETO L,ANGUITA D,et.al.A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator[C]// Proceedings of Design for Teaching and Learning in a Networked World:10th European Conference on Technology Enhanced Learning.Toledo,Sep. 15-18,2015:352-366.

(责任编辑:傅 游)