基于概念漂移检测的大数据交易过程模型优化方法

张 鹏[1],叶 剑[2,3],张 鹏[1]
(1. 山东科技大学,山东青岛 266590;2. 中国科学院计算技术研究所,北京 100190;
3. 移动计算与新型终端北京市重点实验室,北京 100190)

摘 要: 通过大数据交易过程模型优化,实现对大数据交易过程的精确建模,对于构建稳定、鲁棒和精确的交易平台至关重要. 然而,大数据交易流程随时间而变化,传统的静态模型优化方法无法反映现实流程模型的时态变化特征. 为此,本文提出一种基于概念漂移的大数据交易模型优化方法,在概念漂移点检测和定位的基础上,设计大数据交易日志分割算法,演算日志精准分割点,构建具有时变特性的大数据交易分段模型,实现基于日志分割的模型优化. 该方法在天元大数据交易平台的应用实践表明,优化模型在拟合度和精确度方面均优于静态模型,对大数据交易演化过程的适配性更强.

关键词: 大数据交易; 概念漂移; 日志分割; 模型评估
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2019)07-1465-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.07.009

Optimization of Big Data Transaction Process Model Based on Concept Drift Detection

ZHANG Peng[1], YE Jian[2,3], ZHANG Peng[1]
(1. *Shandong University of Science and Technology, Qingdao, Shandong* 266590, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
3. *The Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing* 100190, *China*)

Abstract: Through the optimization of big data transaction process model, the accurate modeling of big data transaction process is realized, which is significant for building a stable, robust and accurate transaction platform. However, the big data transaction process changes over time, and traditional static model optimization methods cannot reflect the characteristics of time-varying changes in real-world process models. For this reason, this paper proposes an optimization approach of big data transaction model. Based on the detection and location of concept drift points, the approach designs a big data transaction log segmentation algorithm and calculates log precise segmentation points to build a large data transaction time-varying segmented model and to realize model optimization. The proposed approach has got used in Tianyuan Big Data Transaction Platform, which shows that the optimization model has an advantage over the static model in fitness, precision and adaptation to the big data transaction process.

Key words: big data transaction; concept drift; log segmentation; model evaluation

1 引言

随着数据经济、数据管理等概念的普及,数据的价值不仅在各个领域得到认可,而且数据本身更成为一种新型交易商品[1]. 大数据交易平台提供了数据出售、数据加工以及数据交易平台服务等多种交易模式[2]. 大数据交易具有业务流程复杂多变等特点,因此面临巨大挑战. 在业务流程不断变化的情况下,大数据交易

收稿日期:2018-04-26;修回日期:2018-12-24;责任编辑:孙瑶
基金项目:国家重点研发计划(No. 2016YFB1001100);工信部 2016 年绿色制造系统集成项目;工信部 2018 年工业互联网创新发展项目和移动计算与新型终端北京市重点实验室研究基金

业务流程管理[3,4]通过对业务流程建模,实现对各个业务环节的管理、监督和优化,并且对优化业务规则、提高交易流程效率具有重要意义.作为业务流程管理的重要环节,过程挖掘[5~9]是服务于数据挖掘和业务过程建模之间的桥梁,它不仅能够发现业务流程的问题和瓶颈,而且能够有效的实现对过程模型的优化.过程挖掘技术通过从系统中提取日志信息,实现对过程模型的发现、监测和优化,从而完成对现实业务流程的改进.

目前,由于大数据交易模式的不成熟以及交易流程的复杂性,为了实现对业务流程的优化,需要对整个流程创建过程模型.然而复杂的交易模式和动态的流程变化都对模型的创建与分析提出了极大的挑战.传统的模型优化方法都假设业务流程是稳定不变的,然后对稳定流程进行建模实现分析优化[10~12].但对于现实的大数据交易平台,由于功能优化升级、业务更新换代等原因,交易流程是一个时变动态演化的过程,那么传统的优化方法就无法解决这种变化业务流程模型的优化.

2 相关工作

传统的模型优化都是以业务流程稳定不变为前提,实现对过程模型的优化.例如文献[10]提出通过通用的分解方法解决模型修复问题,并且可以与现有的不同过程发现和一致性检查技术相结合.文献[11]则通过分析模型校准的结果并分解原日志,得出日志中与模型不拟合的子日志,重演所有子日志,以循环或子过程的方式插入到模型的适当位置.文献[12]在文献[11]的基础上提出修复建议的概念,从时间复杂度和修复资源两个角度,对比不同算法产生的修复建议对模型修复的影响.对于大数据交易流程,不断随时间完成更新,交易流程中部分分支可能会出现新活动的添加或旧活动的删除和替换.因此过程稳定不变条件下的传统优化模型显然不适合业务不断变化的大数据交易流程.针对解决动态模型优化,本文想到使用概念漂移检测解决问题.

模型随时间发生变化的过程被称为概念漂移[13].概念漂移最先在数据挖掘中提出,用于表示目标变量的统计特性随着时间的推移以不可预见的方式变化的现象[14].数据挖掘中概念漂移侧重研究数据流随时间的变化,而过程挖掘则侧重日志活动关系随时间的变化.过程挖掘中的概念漂移的分析分为离线分析[15]和在线分析[16].离线分析包含控制流、数据、资源三个方面的探讨,概念漂移方式也分为突变和渐变[17]两大类.文献[18]对于过程挖掘中概念漂移检测提出了统一解决框架,框架流程从日志文件出发,首先完成特征值的选择与抽取,其次通过合理的滑动窗口生成特征值样本,然后通过对比特征值样本实现对漂移发生的检测,最后通过交互式可视化方式显示漂移,并完成对发生漂移原因的分析.框架的流程提供了检测概念漂移的基本思路.文献[18,19]按照框架流程完成了对突变概念漂移中控制流改变的检测.文献[19]通过自己提出的事件类概念,作为检测概念漂移流程中的特征值,完成对突变概念漂移中控制流改变的检测,并以具体案例验证了特征值的有效性.文献[18]不仅验证轨迹中全活动对的 J-measure 特征值能有效检测突变概念漂移中控制流的改变,而且利用轨迹中单个活动对的 Window Count 特征值检测在控制流中突变概念漂移的大致范围.目前对于控制流中突变概念漂移检测与定位,文献[18]通过观察 J-measure 特征值和 Window Count 特征值 KS 检验的显著性概率图得到大致位置,这种观察确定的突变概念漂移位置和范围的方法,显然无法满足我们日志精准切割和模型优化的需求.

针对上述问题,本文以动态的大数据交易流程为切入点,提取表示活动关系重要性的特征值,通过提出的粗粒度与细粒度优化方法准确定位到轨迹级别上的漂移点,完成对动态模型的日志分割,实现对动态模型的优化.本文将该优化方法用于天元大数据交易平台[20]的流程形式化分析,采用一致性分析标准[21~24],从拟合度、精确度两个方面对优化前后的模型进行对比,充分验证了优化方法的有效性.

3 相关概念

本节简要介绍 Petri 网[5~9]、过程挖掘概念漂移[14~18]、一致性分析[21~24]等基本定义及有关符号表示.Petri 网用于对完整日志和分割后日志进行建模分析,过程挖掘中概念漂移检测为分割日志提供可行性,一致性分析则实现对模型动态优化前后的对比分析.

3.1 多重集、序列

$\mathcal{A}$ 是一个集合,存在映射 $\mathcal{B}$,$\mathcal{B}\rightarrow\mathcal{B}(\mathcal{A})$,其中$\mathcal{B}(\mathcal{A})$代表集合 $\mathcal{A}$ 上所有多重的集合.例如 $\mathcal{A}=\{a,b,c,d\}$,多重集 $B_i\in\mathcal{B}(\mathcal{A})$,$B_1=[\,]$,$B_2=[a]$,$B_3=[a,b,c,c,d,d]$,$B_4=[a^2,b,c^2,d]$.序列 $\sigma=\langle a_1,a_2,\cdots,a_n\rangle\in\mathcal{A}^*$,$a_i\in\mathcal{A}$,$n\in\mathbb{N}_0$,$|\sigma|$表示序列 σ 的长度,$|\sigma|=n$.

3.2 Petri 网

Petri 网是用于描述分布系统的一种模型结构,它既能描述系统结构,又能模拟系统运行.三元组 $N=(P,T,F)$称为一个 Petri 网,其中 P 是一个有限库所集,T 是一个有限变迁集,F 是网 N 的流关系.

设 $N=(P,T,F)$为一个 Petri 网,对于 $x\in P\cup T$,记

$$\cdot x=\{y\mid y\in P\cup T\wedge(y,x)\in F\}$$

$$x\cdot=\{y\mid y\in P\cup T\wedge(x,y)\in F\}$$

称 $\cdot x$ 为 x 的前集或者输入集,$x\cdot$ 为后集或者输出集.

3.3　过程挖掘概念漂移

概念漂移表示目标变量的统计特性随着时间的推移以不可预见的方式变化的现象. 过程挖掘中的概念漂移是指,过程在随着时间变化的同时发生改变的现象. 例如,随着新产品的添加,一个选择结构被插入到一个组织的产品开发过程中.

3.4　控制流漂移

控制流漂移是指过程模型随着时间的改变,模型的行为和结构发生变化. 控制流的结构改变可以分为插入、删除、替换和重新排序过程片段操作. 例如,一个在处理和接受应用程序之后收取费用的组织,现在可以改变其流程,在处理应用程序之前强制支付该费用. 有时,一个过程模型的控制流结构方面保持完整,但是模型行为方面已经发生改变也称为控制流漂移. 例如,考虑一个保险赔偿分类,赔偿金额决定索赔等级为高或者低,保险公司为了提高赔偿额度,将 1000 元索赔金额从高索赔等级改为低索赔等级,过程的模型未发生改变,但是事件的路径发生了改变.

3.5　特征值 Window Count

给定窗口长度 $l\in\mathbb{N}$,Window Count 是关于后继(前继)关系的函数,$f_{WC}^{l,t}:\mathcal{A}\times\mathcal{A}\to\mathbb{N}_0$,定于基于活动对集合. 给定一个轨迹 t 和窗口长度 l,$S^{l,t}(a)$ 表示以活动 a 开始的滑动窗口集合. $\mathcal{F}^{l,t}(a,b)$ 表示以活动 a 开始并且包含活动 b 的窗口集合. 即:

$$\mathcal{F}^{l,t}(a,b)=[s\in S^{l,t}(a)\ \exists\,|1<k\leq|s|,s(k)=b]$$

对于活动 b 跟随活动 a 的 Window Count 值即:

$$f_{WC}^{l,t}(a,b)=|\mathcal{F}^{l,t}(a,b)|$$

3.6　特征值 J-measure

J-measure 是一种表示规则信息量的矩阵. 过程挖掘中概念漂移采用这个指标作为特征值来描述活动之间的重要性,其基础在于可以把活动 b 的跟随关系看作是一个规则:如果活动 a 发生,那么活动 b 将可能发生.

定义活动对集合 $A\times A\to\mathbb{R}^+$,给定窗口长度 $l\in\mathbb{N}$,$p(a)$,$p(b)$是活动 a 和活动 b 在轨迹 t 中发生的概率,$p^{l,t}(a,b)$ 是在窗口长度为 l 下 b 跟随 a 发生的概率,$CE^{l,t}(a,b)$是 b 跟随 a 在窗口长度为 l 的交叉熵.

$$CE^{l,t}(a,b)=p^{l,t}(a,b)\log_2\left(\frac{p^{l,t}(a,b)}{p^t(b)}\right)+(1-p^{l,t}(a,b))\log_2\left(\frac{1-p^{l,t}(a,b)}{1-p^t(b)}\right)$$

关于跟随(先于)关系的 J-measure 函数:

$$F_J^{l,t}(a,b)=p^t(a)CE^{l,t}(a,b)$$

3.7　事件日志

事件日志由案例组成,并且案例由事件组成,案例中的事件以轨迹的形式来表示,即(唯一的)事件的序列. 设集合 $\mathcal{A}$ 代表日志中所有标签的集合,轨迹中的每一个标签 $v\in\mathcal{A}^*$ 代表一个事件. 例如事件日志 $L=[<a,d,f,g,h>^{12},<a,c,d,g,h>^{6},<a,d,l,e,g,h>^{8}]$,轨迹的上标数字代表对应案例的个数.

3.8　一致性分析

过程挖掘一致性分析是对过程模型和实际日志的分析,通过量化的方式给出过程挖掘拟合度、精确度、简洁度和泛化度. 拟合度用以衡量日志记录的系统行为可被过程模型再现的程度,精确度度量过程模型允许发生除日志中的轨迹之外的额外的行为,简洁度表示模型的复杂程度,泛化度表示对事件日志中例子行为的泛化程度.

4　基于概念漂移检测的模型优化

为了清晰地描述大数据交易流程及模型优化过程,本文以大数据交易过程中数据加工简化流程为例,通过对控制流中突变漂移的检测,实现对大数据交易过程模型优化方法的阐述,假设日志 $L=\{<a,b,d,c,e,f>^{18},<a,b,c,d,e,f>^{16},<a,c,b,d,e,f>^{20},<a,b,c,e,d,f>^{16},<a,b,d,c,g,f>^{18},<a,b,c,d,g,f>^{19},<a,c,b,d,g,h,f>^{17},<a,b,c,g,d,h,f>^{16},<a,b,c,g,d,f>^{16},<a,b,c,d,g,h,f>^{19},<a,c,b,d,g,f>^{17},<a,b,d,c,g,h,f>^{18}\}$,其中字母表示活动,上标数字表示轨迹发生次数. 由于概念漂移的检测依赖于时间戳,所以日志中的每个活动都包含时间戳属性. 表 1 为数据加工简化过程中符号含义对照表.

表 1　符号含义对照表

符号	含义
a	进入数据中心
b	数据选择
c	工具选择
d	数据审批
e	工具审批
f	数据分析
g	查看工具使用

在日志 L 上执行挖掘算法,最终得到数据交易流程过程模型. 直接挖掘出的过程模型既无法反映出活动对之间的变化关系,也无法查看交易流程在不同时刻的真实状态. 为了能精确地完成日志分割,显示出业务流程模型在不同时刻的真实状态,我们想到利用概念漂移检测解决问题.

对于控制流中突变概念漂移检测,文献[18]通过轨迹中全活动对的 J-measure 特征值检测突变漂移,利用轨迹中单个活动对的 Window Count 特征值确定漂移大致范围,并且主要通过观察 J-measure 和 Window

Count 的 KS 假设检验概率图得到,但目前缺乏对突变漂移的检测和发生范围的量化方法,这导致得到的突变漂移的检测会存在很大的误差,甚至使得观察出的漂移点中存在假漂移点. 而且当日志中轨迹条数很大时,轨迹范围也很难通过观察得到.

为了适配动态变化的业务流程特点,本文提出基于概念漂移检测的模型优化方法,通过抽取 J-measure 特征值,设计粗粒度优化算法,完成对子日志级漂移的定位,实现对假漂移点的过滤和对漂移活动对的发现;在此基础上,利用抽取 Window Count 特征值,设计细粒度优化算法,完成对轨迹级漂移的定位和模型的分割,最终达到对动态模型的优化.

本文利用概念漂移检测实现对动态模型的优化,不仅在粗细粒度算法中实现对 J-measure 特征值和 Window Count 特征值的量化,而且通过检测出的漂移点实现对日志分割,完成模型优化.

4.1 粗粒度优化方法

对于粗粒度优化方法,本文首先把日志分割成等份子日志,通过 J-measure 特征值检测突变漂移完成日志的分割,进而分割得到粗略分割模型,对比分割模型之间的差异,观察分割模型前后有无活动变化. 粗粒度优化方法识别的漂移点粒度是子日志级的,并且通过粗粒度优化方法能够达到两点目的:(1)去除观察得到的假漂移点;(2)对比模型前后差异,找到真漂移点前后,发生漂移的活动对.

以大数据交易流程中数据加工简化过程为例,将 210 条轨迹分以每组 5 条轨迹进行分割,得到 42 组轨迹. 通过全活动对对 J-measure 特征值检测突变漂移得到图 1,通过观察图 1 将索引为 14 的轨迹组和索引为 28 的轨迹组定为漂移点. 此时的漂移点还比较粗糙,而且漂移点中可能存在观察误差所得到的假漂移点.

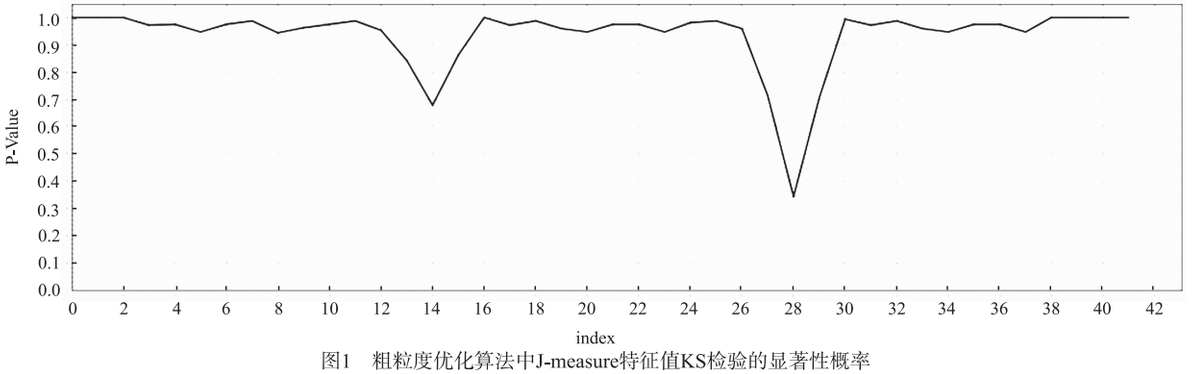


图1 粗粒度优化算法中J-measure特征值KS检验的显著性概率

粗粒度优化方法根据日志大小,选取合适的每组轨迹条数将日志分割成带索引的子日志. 对于 n 条轨迹的日志 L,选取的每组的轨迹条数为 s,日志份数为 m,则 $m=[n/s]$,日志 $L=\{SL_k,k\in[1,\cdots,m]\}$,其中 SL_k 为子日志. 通过观察得到发生漂移的子日志称为漂移子日志. 如图 2 及算法 1 所示.

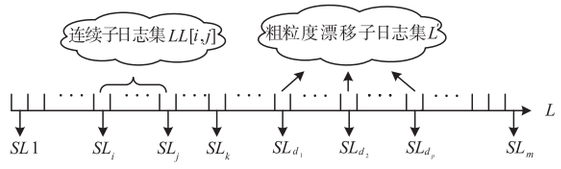


图2 粗粒度优化方法效果图

算法 1 粗粒度优化算法

Input 事件日志 L,粗粒度漂移子日志索引集 RIndex,粗粒度漂移子日志集 L'

Output 真漂移子日志集 L'_t

// Step 1:完成粗粒度模型分割,得到分割后每段连续子日志集所对应的模型定义集合 list 用于存放子日志模型;

for each SL_{d_i} in L' do

if $d_i=d_1$ then

$SL_{\text{beginIndex}}\rightarrow 1$, $SL_{\text{endIndex}}\rightarrow SL_{d_1}-1$;

//从开始子日志索引到结束子日志索引截取连续子日志集

split L from $SL_{\text{beginIndex}}$ to $SL_{\text{endIndex}}\rightarrow LL[1,SL_{d_1}-1]$;

$LL[1,SL_{d_1}-1]$通过挖掘算法得到模型 M_i;

子日志模型 M_i放入集合 list;

if $d_i=d_p$ then

$SL_{\text{beginIndex}}\rightarrow SL_{d_p}$, $SL_{\text{endIndex}}\rightarrow m$;

//从开始子日志索引到结束子日志索引截取连续子日志集

split L from $SL_{\text{beginIndex}}$ to $SL_{\text{endIndex}}\rightarrow LL[SL_{d_p},m]$;

$LL[SL_{d_p},m]$通过挖掘算法得到模型 M_i;

子日志模型 M_i放入集合 list;

else

$SL_{\text{beginIndex}}\rightarrow SL_{d_i}$, $SL_{\text{endIndex}}\rightarrow SL_{d_{i+1}}-1$;

//从开始子日志索引到结束子日志索引截取连续子日志集

split L from $SL_{\text{beginIndex}}$ to $SL_{\text{endIndex}}\rightarrow LL[SL_{d_i},SL_{d_{i+1}}-1]$;

$LL[SL_{d_i},SL_{d_{i+1}}-1]$通过挖掘算法得到模型 M_i;

子日志模型 M_i放入集合 list;

endif;

```
    endfor;
// Step 2:对比相邻模型差异,得到真假漂移子日志集
    for each M_i in list
      M_i与M_{i+1}对比;
      if 模型存在差异 then 输出 L'_t;
       end if;
      if 模型不存在差异 then 输出 L'_f;
       endif;
    endfor;
```

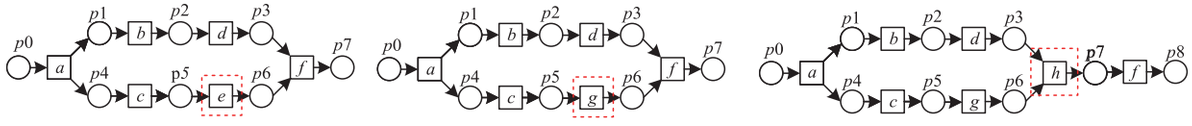
为了描述粗粒度漂移优化,给出定义如下.

定义 1(连续子日志集)　连续子日志集为有序集 $LL[i,j]=\{SL_k \mid i \le k \le j, i,\varphi \in [1,\cdots,m]\}$,其中 SL_k 为子日志,i 为开始子日志索引,j 为结束子日志索引,则 $LL \subseteq L$.

定义 2(粗粒度漂移子日志集)　设 i 为发生漂移时子日志索引,漂移发生 p 次时,有序集 RIndex $=\{d_i\}$, $i \in [1,\cdots,p]$,记粗粒度漂移子日志集为有序集 $L'=\{SL_{d_1},SL_{d_2},\cdots,SL_{d_p}\}$, $d_1<d_2<\cdots<d_p$, $d_i \in$ Rindex,RIndex 为粗粒度漂移子日志索引集,则 $L' \subset L$.

定义 3(真漂移子日志集)　设真漂移子日志集 L'_t 为有序集,真漂移子日志是指真实发生的漂移子日志,$L'_t=\{SL_{d_1},SL_{d_2},\cdots,SL_{d_p}\}$, $d_1<d_2<\cdots<d_p$, $d_i \in$ Rindex,则假漂移子日志集为有序集为 L'_f, $L'_t \subset L'$, $L'_t \cup L'_f = L'$, $L'_t \cap L'_f = \varnothing$.

通过粗粒度优化算法,得到真漂移子日志集 L'_t. 更为重要的是通过真漂移子日志分割子模型后,可以清楚地观察到真漂移子日志前后模型差异,进而得到具体发生漂移的活动对的前后关系. 对于模型前后发生漂移的活动对不止一对的情况,可以任选一对用于漂移范围的确定. 对于粗粒度优化方法,以大数据交易流程中数据加工简化流程为例,粗粒度漂移子日志集 $L'=\{SL_{14},SL_{28}\}$. 通过粗粒度优化算法得到,真漂移子日志集 $L'_t=\{SL_{14},SL_{28}\}$. 真漂移子日志 SL_{14} 和 SL_{28} 将粗粒度日志分割 3 个连续子日志集 $LL[1,13]$, $LL[14,27]$ 和 $LL[28,42]$. 3 个子日志集依次得到 3 个 Petri 网模型 M_1, M_2 和 M_3,如图 3 所示.



(a) 连续子日志集LL[1,13]所对应模型M_1　(b) 连续子日志集LL[14,27]所对应模型M_2　(c) 连续子日志集LL[28,42]所对应模型M_3

图3

4.2　细粒度优化方法

通过分析 3 个模型可以发现,实际模型的控制流发生变化,日志序列随着时间的变化会有新的活动出现或者新的活动替换旧的活动. 在上述事件日志 L 中,活动 g 的出现伴随着活动 e 的消失,活动 h 在日志后期出现. 当 M_1 漂移到 M_2 时选取漂移活动对 c 和 e,且活动 e 跟随活动 c;当 M_2 漂移到 M_3 时选取漂移活动对 g 和 h,且活动 h 跟随活动 g. 选取发生漂移的活动对,通过细粒度优化算法确定漂移发生范围,并在漂移范围中找到漂移点. 细粒度优化算法将突变漂移确定到轨迹级别,实现更为精准的日志的分割以及模型的优化.

以大数据交易流程中数据加工简化流程为例,210 条轨迹不再进行分组,通过全活动对对 J-measure 特征值检测突变漂移得到图 4. 此时的潜在漂移点是一条轨迹,漂移点是轨迹级别的. 对于 n 条轨迹的事件日志,此时的轨迹索引记为 k,则此时的日志 $\overline{L}=\{l_k,k \in [1,\cdots,n]\}$,此时的漂移点也称为漂移轨迹,精确分割日志挖掘出的每个 Petri 模型称为最优子模型,记为 OM_i. 如图 5 及算法 2 所示.

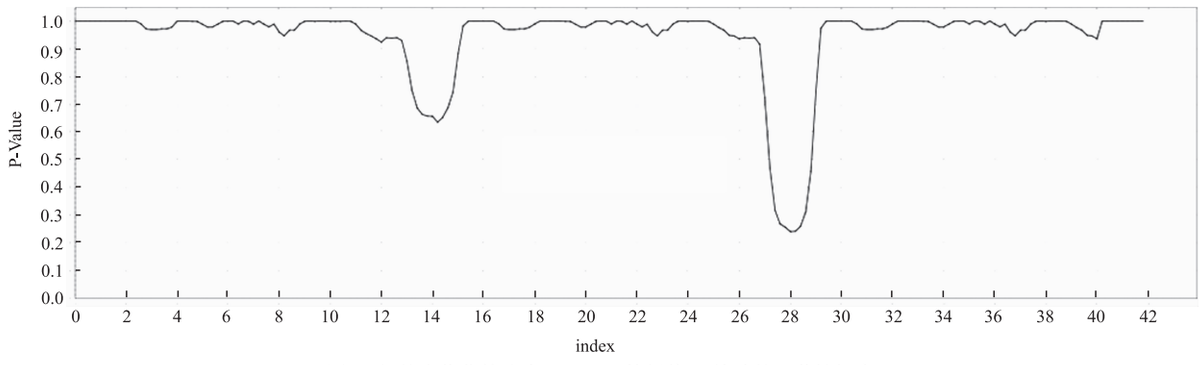


图4　细粒度优化算法中J-measure特征值KS检验的显著性概率

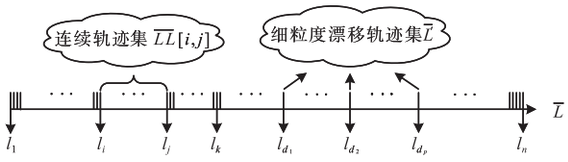


图5 细粒度优化方法效果图

算法 2 细粒度优化算法

```
Input  漂移活动对
Output  最优子模型
// Step1:通过漂移活动对依次求出每次发生漂移的范围
for each 按照时间序列依次输入漂移活动对
  令 p_i = f(l_i^WC) 中 p_i = 0
    第一次 p_i = 0 时,记为漂移范围 ll[i,j] 中的 i;
    最后一次 p_i = 0 时,记为漂移范围 ll[i,j] 中的 j;
    得出漂移活动对 ll[i,j] 并放入 CR;
 endfor;
// Step2:在每次漂移范围中求出一条漂移轨迹并放入细粒度漂移轨迹集
for each ll[i,j] in CR do
  for each l in ll[i,j]
    当 p_min = f(l_i^l). getMin( ) 时,标记轨迹;
  endfor;
  if 标记 p_min 轨迹的个数不止一个
    随机选取一条标记 p_min 轨迹放入 L̄';
  endif;
  if 标记 p_min 轨迹的个数只有一个
    把唯一标记 p_min 轨迹放入 L̄';
  endif;
endfor;
// Step3:完成细粒度模型分割,得到精准分割日志后,每段日志对应的最优子模型
for each l_{d_i} in L̄' do
  if d_i = d_1 then
    l_beginIndex → 1, l_beginIndex → l_{d_1} - 1;
    //从开始轨迹索引到结束轨迹索引截取连续轨迹集
    split L̄ from l_beginIndex to l_beginIndex → LL̄[1, l_{d_1} - 1];
    LL̄[1, l_{d_1} - 1] 通过挖掘算法得到模型 OM_i;
  if d_i = d_p then
    l_beginIndex → l_{d_p}, l_endIndex → n;
    //从开始轨迹索引到结束轨迹索引截取连续轨迹集
    split L̄ from l_beginIndex to l_endIndex → LL̄[l_{d_p}, n];
    LL̄[l_{d_p}, n] 通过挖掘算法得到模型 OM_i;
  else
    l_beginIndex → l_{d_i}, l_endIndex → l_{d_{i+1}} - 1;
    //从开始轨迹索引到结束轨迹索引截取连续轨迹集
    split L̄ from l_beginIndex to l_endIndex → LL̄[l_{d_i}, l_{d_{i+1}} - 1];
    LL̄[l_{d_i}, l_{d_{i+1}} - 1] 通过挖掘算法得到模型 OM_i;
    endif;
  endfor;
```

为了完成细漂移优化,给出定义如下.

定义 4(p-value 轨迹映射) 日志 L 为事件轨迹的集合,日志 L 中事件轨迹总数记为 n,设 l 为事件日志中的一条事件轨迹,记为 $l=\langle e_1,e_2,\cdots,e_n\rangle_i$,$p$ 为在滑动窗口中滑动轨迹 l_i 前后两个特征值的数据集作 KS 假设检验对应的 p-value 值,定义 p-value 与轨迹之间的映射为 f,$p_i=f(l_i^F)$,F 为选取数据集特征值类型,则特征值为 J-measure 时 $p_i=f(l_i^J)$,特征值为 Window Count 时,$p_i=f(l_i^{WC})$.

定义 5(连续轨迹集) 连续轨迹集为有序集 $\overline{LL}[i,j]=\{l_k\mid i\leq k\leq j,i,j\in[1,\cdots,n]\}$,其中 l_k 为轨迹,i 为开始轨迹索引,j 为结束轨迹索引,则 $\overline{LL}\subseteq\overline{L}$.

定义 6(细粒度漂移轨迹集) 设 i 为发生漂移时轨迹索引,漂移发生 p 次时,有序集 $\mathrm{SIndex}=\{d_i\}$,$i\in[1,\cdots,p]$,记细粒度漂移轨迹集为有序集 $\overline{L}'=\{l_{d_1},l_{d_2},\cdots,l_{d_p}\}$,$d_1<d_2<\cdots<d_p$,$d_i\in\mathrm{SIndex}$,SIndex 为细粒度漂移轨迹索引集,则 $\overline{L}'\subset\overline{L}$.

定义 7(漂移范围集) 设 i 为开始出现漂移的轨迹索引,j 为漂移结束的轨迹索引,漂移发生 p 次时,漂移范围为有序集 $ll[i,j]=\{ll_k\}$,$k\in[i,\cdots,j]$,记漂移范围集为有序集 $CR=\{ll_{k_1}[i_1,j_1],ll_{k_2}[i_2,j_2],\cdots,ll_{k_p}[i_p,j_p]\}$,$k_1<k_2<\cdots<k_p$,$k_i\in ll[i,j]$.

通过细粒度优化算法,可以将概念漂移点精准确定到具体的轨迹完成日志分割. 以大数据交易流程中数据加工简化流程为例,使用粗粒度优化方法在两处发生的漂移区域选取漂移活动对 c 和 e,g 和 h,且 e 跟随 c,h 跟随 e. 通过输入漂移范围集 $CR=\{ll[68,72],ll[137,142]\}$,确定了具体漂移轨迹索引为 71 和 140. 通过具体的漂移轨迹索引将模型依次分割为图 6. 通过算法输出的最优子模型和其对应的连续轨迹集,组成一个分段模型 SM,$SM=\begin{cases}OM_1, & TT=\overline{LL}[1,70]\\ OM_2, & TT=\overline{LL}[71,139]\\ OM_3, & TT=\overline{LL}[140,210]\end{cases}$.

5 实验

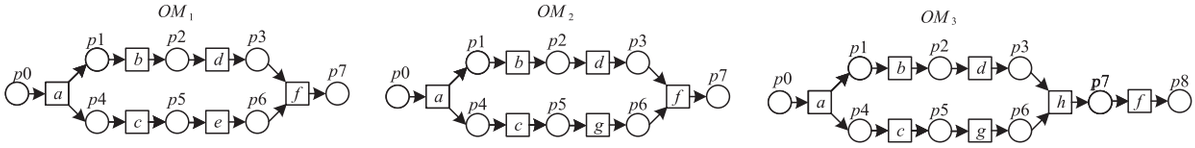
本文实验采用的数据为天元大数据交易平台交易日志,天元大数据交易平台是全国领先的数据资源与交易平台,提供数据交易与在线数据开发等功能. 本文实验在 ProM6.7 平台完成,通过实验仿真优化模型与原模型的比较,进一步验证本文方法的正确性与有效性.

5.1 实验模型与数据

实验参照天元数据网大数据交易平台[20]数据开发的过程模型,表 2 为大数据交易过程中符号含义对照表.

为了验证模型优化效果,实验过程中随机选取了 5 组事件日志 $L_1\sim L_5$,表 2 对 5 组事件日志中的案例总

数、事件总数、轨迹长度以及日志分割后的最优子模型数进行统计.



（a）连续轨迹集$\overline{LL}$［1,70］所对应最优子模型OM_1（b）连续轨迹集$\overline{LL}$［71,139］所对应最优子模型OM_2 （c）连续轨迹集$\overline{LL}$［140,210］所对应最优子模型 OM_3

图6

表 2　符号含义对照表

符号	含义	符号	含义
T_1	用户申请	T_{13}	设置成员权限
T_2	用户申请审批	T_{14}	生成订单
T_3	进入数据中心	T_{15}	查看工具使用说明
T_4	项目申请	T_{16}	项目授权
T_5	选择数据表	T_{17}	订单审批
T_6	工具选择	T_{18}	工具配置
T_7	邀请项目成员	T_{19}	数据加工
T_8	申请数据表	T_{20}	导出数据
T_9	免费试用	T_{21}	部署
T_{10}	按时计费	T_{22}	人群投放
T_{11}	周期计费	T_{23}	应用开发
T_{12}	工具审批	T_{24}	上架售卖

根据天元大数据网的大数据交易平台交易日志,从平台交易日志中随机选取带有时间戳的3000 条轨迹记为日志 L_1,基于 α 算法得如图 7 所示的大数据交易流程数据应用开到发模型作为原始模型.

以日志 L_1为例,按照基于概念漂移检测的优化算法对原始模型进行优化,通过检测出的漂移点完成模型分割优化的目的. 实验针对随机选取的带有时间戳的 3000 条轨迹日志 L,通过基于概念漂移检测的优化算法,确定漂移轨迹索引为 1288 和 2075,并利用具体的漂移轨迹索引,将选取的大数据交易流程应用开发日志 L_1依次分割,最终得到由 3 个最优子模型构成的优化分段模型,如图 8 所示. 通过算法输出的最优子模型和其对应的连续轨迹集组成一个大数据交易流程应用开发分段模型 SM,SM

$$=\begin{cases} OM_1, & TT=\overline{LL}[1,1287] \\ OM_2, & TT=\overline{LL}[1288,2074] \\ OM_3, & TT=\overline{LL}[2075,3000] \end{cases}$$

. 从分段模型可以清晰地看出模型的控制流随时间变化而变化,对比分段模型 OM_1与 OM_2可以看出第一次发生漂移前后,变迁 T_{15}“查看工具使用说明”被变迁 T_{18}“工具配置”所替代;对比分段模型 OM_2与 OM_3可以看出第二次发生漂移前后,变迁 T_{20}“导出数据”加入到模型中,出现在变迁 T_{21}“部署”之前.

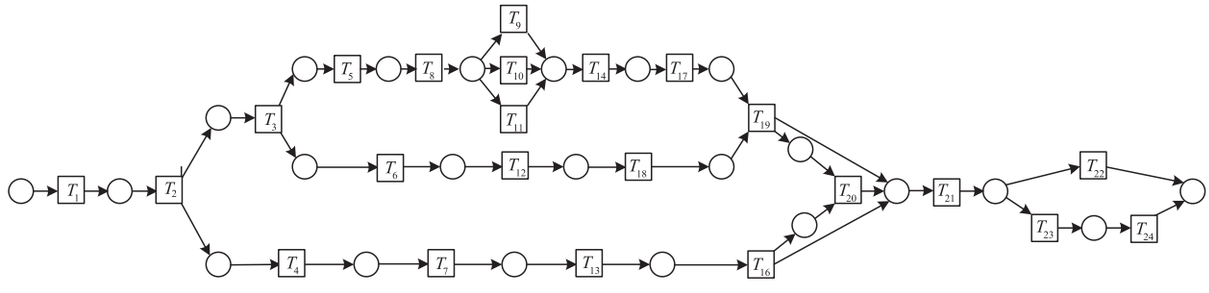


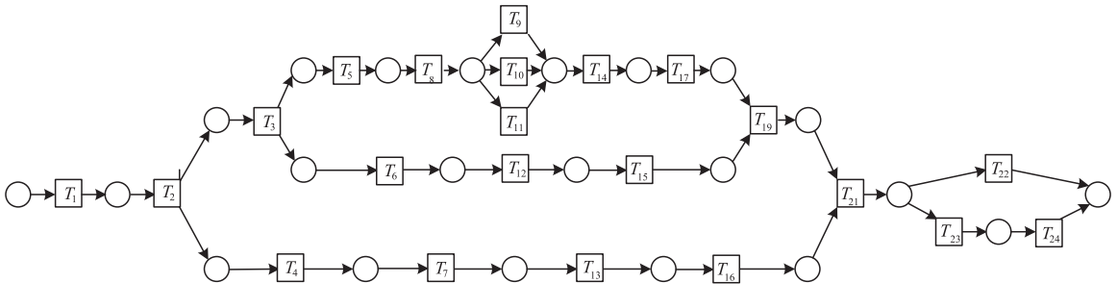
图7　大数据交易流程数据应用开发模型

大数据交易模型在优化前,即使模型发生变化,也只能统一拟合为一整个模型,完全无法观测出日志 L_1所对应的模型的变化. 但大数据交易模型在优化后,不仅确认日志 L_1中模型是否发生变化,而且在模型发生变化的同时,得到由 3 个最优子模型构成的优化分段模型 OM_1、OM_2以及 OM_3. 通过观察对比相邻最优子模型,能够清晰地观察出模型的动态变化.

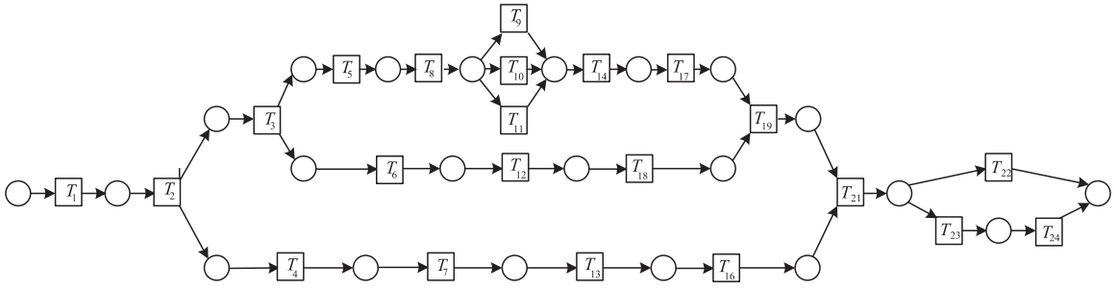
5.2　模型评估

本节从拟合度和精确度两个角度对优化前后模型进行对比分析. 模型评估过程中,优化前后模型的拟合度和精确度均采用文献[24]中拟合度和精确度的计算方法. 对于分段模型的拟合度和精确度,则使用以日志作为权重的加权平均原则,通过各个子模型的拟合度、精确度和轨迹数计算得到.

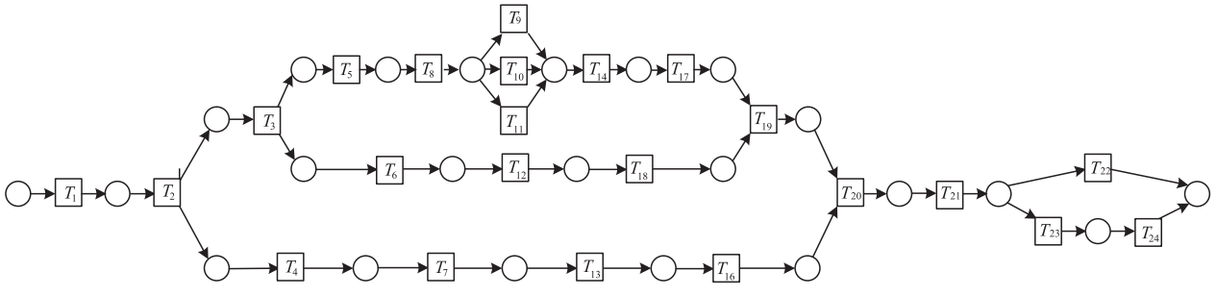
由于模型一直在动态发生变化,通过传统过程挖掘算法直接生成的原始模型并不能准确反映每个时间段模型的真实状态,模型中控制流的改变导致日志中不拟合节点增多和模型精确度下降. 而基于概念漂移的模型优



（*a*）连续轨迹集$\overline{LL}$［1,1287］所对应最优子模型OM_1



（*b*）连续轨迹集$\overline{LL}$［1288,2074］所对应最优子模型OM_2



（*c*）连续轨迹集$\overline{LL}$［2075,3000］所对应最优子模型OM_3

图8

化后的分段模型则是以模型漂移点对动态变化模型进行分割，准确地展现模型在不同阶段对应的不同状态.

对于动态变化模型的分割，本文则采用先粗粒度优化算法后细粒度优化算法的方式，实现对不同阶段模型的展现. 粗粒度算法依赖于对控制流突变概念漂移的漂移点检测，首先把日志分割成等份子日志，通过J-measure 特征值检测突变漂移完成日志的分割，进而分割得到粗略分割模型，对比分割模型之间的差异，观察分割模型前后有无活动变化. 通过观察分割模型前后有无活动变化，找到真实发生漂移的活动对，以备细粒度优化算法所使用. 粗粒度优化算法因为需要遍历每个粗粒度漂移子日志集 L'得到对应子日志模型 M_i以及遍历每个子日志模型 M_i比较前后差异，所以时间复杂度为 $O(n)$. 细粒度优化算法依赖于对控制流突变概念漂移中漂移范围的检测，首先将日志集不在分组，通过轨迹中单个活动对的 Window Count 特征值确定漂移大致范围，并在漂移范围中找到漂移点. 细粒度优化算法因为需要遍历漂移范围集 CR，并在每次漂移范围 $ll[i,j]$中求出一条漂移轨迹并放入细粒度漂移轨迹集，所以时间复杂度为 $O(n^2)$. 粗、细粒度模型优化算法将动态模型精准分割，解决了模型动态变化的难题，极大的减少了模型中因控制流改变而导致的不拟合点数目，实现模型优化前后拟合度和精确度的提高.

以日志L_1为例，从图9 和图10 中可以看出，基于概念漂移的模型优化后的分段模型的拟合度和精确度不仅一直明显高于优化前的模型，而且稳定性也明显高于优化前的模型.

图11 和图12 分别对比日志 $L_1 \sim L_5$模型优化前后的拟合度和精确度. 由图可知，5 组日志的优化后模型拟合度较高于优化前模型，而优化后模型精确度则明

显高于优化前模型. 通过对比日志 $L_1 \sim L_5$ 模型优化前后的拟合度和精确度,进一步反映了基于概念漂移检测模型优化方法的泛化性.

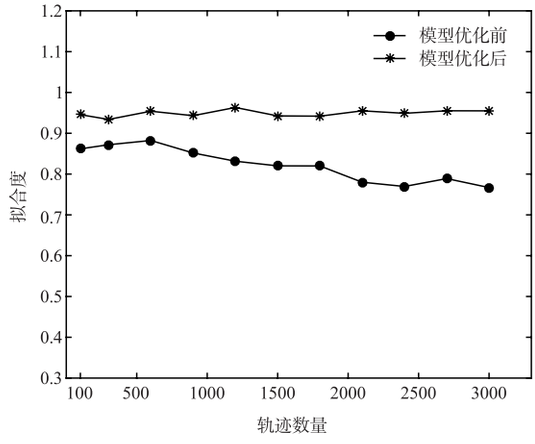


图9　日志–模型间的拟合度

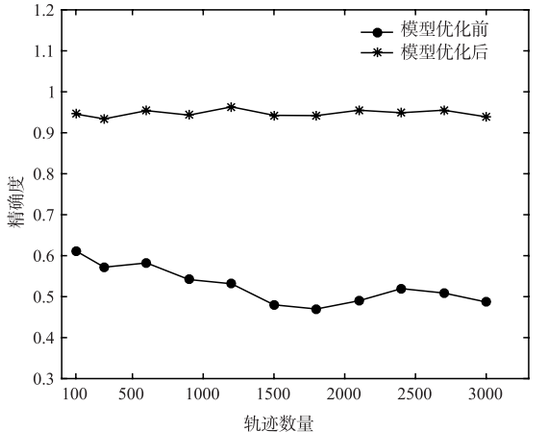


图10　日志–模型间的精确度

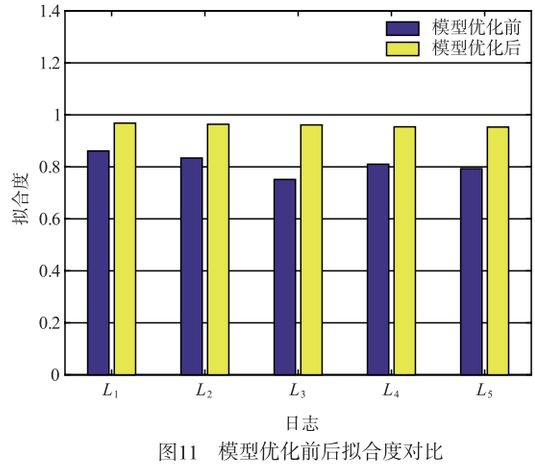


图11　模型优化前后拟合度对比

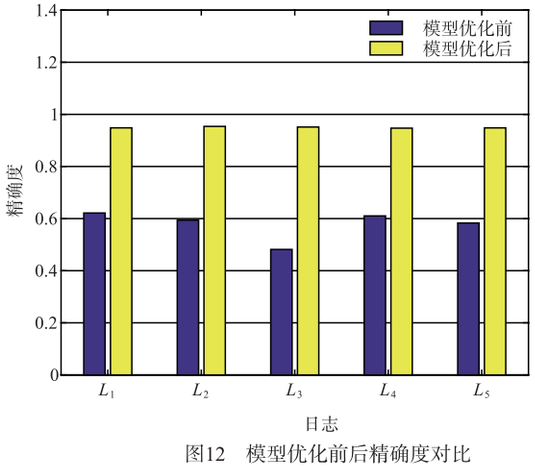


图12　模型优化前后精确度对比

6　结论

本文针对传统模型优化方法在规约业务流程时变演化特性能力不足的缺陷,提出基于概念漂移的模型优化方法,通过发现漂移点将模型依次分割成若干较为稳定的子过程模型,并由子过程模型模型组成分段模型,优化后的分段模型实现了对动态模型的优化. 与此同时,通过实验对比基于概念漂移模型优化方法优化前后的模型,从模型的拟合度以及精确度进行分析,与此同时验证了本论文方法的有效性和正确性.

本文的研究工作主要集中在量化控制流中突变漂移检测,实现对动态模型优化方面,对突变漂移的检测只考虑控制流方面,下一步可以从资源,数据等角度实现对动态模型的优化,同时也可以提出更符合大数据交易特点的特征值,更为准确的定位漂移点.

参考文献

[1] 杨张博,王新雷. 大数据交易中的数据所有权研究[J]. 情报理论与实践,2018,41(6):52.

[2] 宋梅青. 融合数据分析服务的大数据交易平台研究[J]. 图书情报知识,2017(2):13 – 19.

[3] 郭艺,叶剑,张鹏. 基于偏差约减的大数据交易模型分析与修复方法[J]. 电子学报,2018,46(7):1754 – 1761.
GUO Yi, YE Jian, ZHANG Peng. Analysis and repair of big data transaction model based on deviation reduction [J]. Acta Electronica Sinica, 2018, 46(7): 1754 – 1761. (in Chinese)

[4] BAKLIZKY M, FANTINATO M, THOM L H, et al. Business process point analysis: survey experiments[J]. Business Process Management Journal, 2017, 23(2): 399 – 424.

[5] AALST W M P V D, MEDEIROS A K A D, WEIJTERS A J M M. Genetic process mining[J]. Lecture Notes in Computer Science, 2005, 14(2): 76 – 83.

[6] Van der AALST Will. Process Mining: Data Science in Action[M]. Berlin: Springer, 2016.

[7] 杜玉越,朱鸿儒,王路,等. 一种基于逻辑 Petri 网的过程

挖掘方法[J]. 电子学报,2016,44(11):2742 – 2751.
DU Yu-yue,ZHU Hong-ru,WANG Lu,et al. A method of process mining based on logic petri nets[J]. Acta Electronica Sinica,2016,44(11):2742 – 2751. (in Chinese)
[8] KALENKOVA A A,van der AALST W,LOMAZOVA I A,et al. Process mining using BPMN:relating event logs and process models[A]. Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems[C]. US:ACM,2016. 123 – 123.
[9] ZHANG X,DU Y,QI L,et al. Repairing process models containing choice structures via logic petri nets[J]. IEEE Access,2018,PP(99):DOI:10. 1109/ACCESS. 2018. 2870727.
[10] Van der AALST W. Decomposing petri nets for process mining:A generic approach[J]. Distributed & Parallel Databases,2013,31(4):471 – 507.
[11] WYNN M T,et al. Impact-driven process model repair [J]. ACM Transactions on Software Engineering & Methodology,2016,25(4):28 – 80.
[12] FAHLAND D,AALST W. Model repair — aligning process models to reality[J]. Information Systems,2015,47(1):220 – 243.
[13] AALST W,ADRIANSYAH A,MEDEIROS A K A D,et al. Process mining manifesto[A]. International Conference on Business Process Management[C]. Berlin,Heidelberg:Springer,2011. 169 – 194.
[14] 孙雪,李昆仑,韩蕾,等. 基于特征项分布的信息熵及特征动态加权概念漂移检测模型[J]. 电子学报,2015,43(7):1356 – 1361.
SUN Xue,LI Kun-lun,HAN Lei,et al. Construction of the concept drift detection model based on the information entropy of feature distribution and dynamic weighting algorithm[J]. Acta Electronica Sinica,2015,43(7):1356 – 1361. (in Chinese)
[15] BOSE R P J C,AALST W M P V D,ŽLIOBAITĖ I,et al. Handling concept drift in process mining[J]. Lecture Notes in Computer Science,2011,4(1):391 – 405.
[16] BAHNINI A,PLISSONNIER D,KOSKAS F,et al. Online discovery of declarative process models from event streams[J]. IEEE Transactions on Services Computing,2015,8(6):833 – 846.
[17] MAARADJI A,DUMAS M,ROSA M L,et al. Detecting sudden and gradual drifts in business processes from execution traces[J]. IEEE Transactions on Knowledge & Data Engineering,2017,29(10):2140 – 2154.
[18] BOSE R P,WM V D A,ZLIOBAITE I,et al. Dealing with concept drifts in process mining[J]. IEEE Transactions on Neural Networks & Learning Systems,2014,25(1):154.
[19] MANOJ Kumar M V,LIKEWIN Thomas,ANNAPPA B. Capturing the sudden concept drift in process mining[A]. Algorithms & Theories for the Analysis of Event Data (ATAED'15)[C]. Brussels,Belgium,2015. 132.
[20] 天元大数据交易平台[DB/OL]. https://www. tdata. cn/. 2018-10-01.
[21] ROZINAT A,AALST W M P V D. Conformance checking of processes based on monitoring real behavior[J]. Information Systems,2008,33(1):64 – 95.
[22] MANNHARDT F,LEONI M,REIJERS H A,et al. Balanced multi-perspective checking of process conformance [J]. Computing,2016,98(4):407 – 437.
[23] MUNOZ-GAMA J,CARMONA J,AALST W M P V D. Conformance checking in the large:partitioning and topology[A]. International Conference on Business Process Management[C]. Berlin:Springer-Verlag,2013. 130 – 145.
[24] AALST W V D,ADRIANSYAH A,DONGEN B V. Replaying history on process models for conformance checking and performance analysis[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery,2012,2(2):182 – 192.

作者简介

张 鹏 男,1994 年生于安徽阜阳. 研究生. 主要研究方向为过程挖掘、Petri 网、大数据.
E-mail:1192360811@qq. com

叶 剑 男,1974 年生于山东济南. 博士,高级工程师,硕士研究生导师. 主要研究方向为移动互联网挖掘、普适计算.
E-mail:jye@ict. ac. cn

张 鹏(通信作者) 男,1973 年生于山东泰安. 山东科技大学计算机学院副教授. 主要研究方向为 Petri 网、工作流、大数据、高性能计算等.
E-mail:bigbigroc@163. com