文章编号：2095-2163(2022)11-0044-11 中图分类号：TP311 文献标志码：A

面向车载设备数据流的异常检测方法

胡翔宇，陈庆奎
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：针对现有异常检测方法难以通过车辆行驶数据有效地发现车载设备异常的问题，提出一种面向车载设备数据流的异常检测方法。首先，从稳定性、完整性和一致性三种角度分别计算数据的波动、缺失和差异程度，并将其作为检测值放入累计池。然后，采用改进的 Dempster-Shafer 证据理论提取累计池中若干检测值的多个异常特征，并合成特征值，当特征值达到阈值时触发为相应异常事件。最后，结合贝叶斯理论建立概率 Petri 网模型，通过若干异常事件的相互组合推导出设备异常。实验结果表明，在分类模型评价指标下该方法的 F 均值达到近 84%，能够有效地检出可能发生异常的设备。

关键词：车载设备；异常检测；证据理论；Petri 网

Anomaly detection method for on-board equipments data flow

HU Xiangyu, CHEN Qingkui
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Aiming at the problem that the existing anomaly detection methods are difficult to find the anomaly of on-board equipments effectively through the driving data of vehicles, this paper proposes an abnormal detection method for on-board equipments data flow. First, the method calculates the fluctuation, deletion and difference of data from the perspectives of stability, integrity and consistency, and then puts the calculated results into the accumulation pool as the detection value. After extracting multiple abnormal features from several detection values in the cumulative pool, the improved Dempster Shafer evidence theory is used to synthesize multiple features into eigenvalues. When the eigenvalues reach the threshold, the corresponding abnormal events could be triggered. Finally, Bayesian theory and probabilistic Petri nets are combined to model the combination of abnormal events and deduce the fault. Experiments show that the classification model evaluation value F mean of this method reaches nearly 84%, which can effectively detect the equipment that may have abnormalities.

【Key words】 on-board equipments; abnormal detection; evidence theory; Petri nets

0 引 言

在国内外饱受疫情影响的背景下，远程设备的维护工作遭受了巨大冲击。专业维护人员的流动受到了限制，人工检修等服务也受到了影响。以往的设备数量有限、且异常排查难度较低，主要通过定期的人工检修。但随着互联网+与 5G 时代的发展，设备精密化程度提升、设备数量的大幅增长成为了时代发展的趋势，人工检修维护的效率愈加难以满足现实的需要。而大量的设备数据和日志都可以通过网络传输到各个服务商的云平台和数据库中，使设备异常事件分析成为了可能，例如电网状态异常检测[1]、选煤厂设备的远程检测[2]等都是依据设备传感器数据进行异常分析的有效应用场景。

交通网络通过引入车载一体机设备，极大增强了运营数据获取的便利性。该设备通过网络将各类传感器数据传输至云平台，例如车内摄像头的实时监控、GPS 位置信息、行驶速度、报站信息等。管理中心通过以上各种实时数据，可以实现客流量统计[3]、设定调度安排[4]、预测到站时间[5]等。然而针对车载设备的异常排查，现有检修方式还是通过人工。这种方式效率低下，且维护需要定时停运一批车辆，影响正常车辆的工作，因此往往是设备异常已然影响车辆正常运营时才会进行人工检修。针对这种情况，设计一套面向车载设备数据流的异常检测方法显得尤为重要，通过这些运营数据可以发现

基金项目：国家自然科学基金(61572325)；上海重点科技攻关项目(19DZ1208903)；上海智能家居大规模物联共性技术工程中心项目(GCZX14014)。

作者简介：胡翔宇(1997-)，男，硕士研究生，主要研究方向：Petri 网事件分析；陈庆奎(1966-)，男，博士，教授，博士生导师，CCF 会员，主要研究方向：计算机集群、人工智能、并行理论、物联网大规模数据分析等。

通讯作者：陈庆奎 Email：chenqingkui@usst.edu.cn

收稿日期：2022-03-18

车载设备的一些潜在故障，为维修人员提供可能发生异常的设备名单及其优先级，提高检修效率。例如，固定运行路线的车辆通过比对 GPS 信息和运营线路坐标可以判断是否出现 GPS 信号偏移的异常事件，但判断的过程需要考虑诸多因素，短暂的偏移可能是由外部电磁干扰造成，并非车载设备自身的异常，只有出现持久性的或频繁的偏移才能预示着 GPS 模块的损坏。车载设备运营中各类事件纷繁复杂，容易造成误判，如何对车载设备的数据流采取有效的检测，过滤掉一些干扰因素并分析出潜在的异常事件是难点。

目前，采用传感器收集数据并在远程进行设备异常检测的方式已经大量应用于各个领域内，文献[6]引入贝叶斯神经网络建立了卫星遥测数据异常检测模型，通过对系统中不确定性高的样本进行重新评估，提高了异常检测能力。文献[7]通过电力计量装置采集数据、电压互感器的状态量选择和对电压运行状态的在线评估，实现了对异常电压的检测。文献[8]提出了传感器选择策略和数据异常检测的新方案，该方案基于信息论和高斯过程回归实现了对飞机发动机状态的有效监测。

当存在多个传感器或信息源的数据时，由于具备处理不确定性数据的优势，D-S 证据理论成为最常用的多源信息融合技术之一。自该理论提出以来，国内外学者对该理论的冲突悖论和算法改进取得了一定的成果[9-10]，使基于证据理论的异常检测被广泛应用于各个领域[11-13]。但低下的传感器数据质量会极大程度影响检测的效果，针对该问题，文献[14]建立了基于边缘计算的分布式传感数据异常检测模型，提高了检测的效率和准确性。文献[15]提出了基于最近邻的异常检测数据预处理算法，并在实际工业机械的异常检测中得到了验证。

各类异常之间可能具有组合与传递的特性，Petri 网是对事件建模与分析的有力工具，结合 Petri 网进行故障诊断已经在电网[16]和液压器[17]等设备上得到了大量的应用。针对交通设备故障，文献[18]建立了离合器故障树对应 Petri 网，并通过关联矩阵求得最小割集，取得重要度优先级来排序专家系统中的规则，实现了机动车故障的快速定位。但以上方法均未用到实时传感器数据，并且是对故障下的异常模块溯源工作。

针对上述情况，提出一种面向车载设备数据流的异常检测方法，通过发掘车载设备正常行驶数据间的异常关系，实现对车载设备异常事件的检测、累计和组合。为检修人员提供可能发生故障的设备排查名单，提高检测的效率。

本文主要工作有：

（1）底层异常事件生成：针对车载设备实时数据，从 3 种角度判别数据的异常关系提取出检测值，通过累计池收纳检测值并结合证据理论合成特征值，将达到阈值的异常特征触发为底层异常事件，避免因外界因素干扰下数据波动带来的误判。

（2）异常事件的组合推导：通过设备维护日志与历史数据对底层事件次数和故障次数进行统计，采用贝叶斯概率获取各类事件组合的条件概率，使用概率 Petri 网对事件的组合关系建立模型推导故障。

1　异常事件检测方法

车载设备的不同类型数据具有各自协议规定的时间周期、触发条件和数据格式。例如周期位置协议是按规定的时间间隔发送的数据，间隔时间短且较为固定，包含发送时间与经纬度信息等。到站协议是到达目标地点后发送的数据，间隔时间较长且不固定，包含站点信息和发送时间等。

异常是指上述数据出现违背协议规定或无法正常反映车辆状态的情况。由于车辆工作期间发送的数据都是正常的运营数据，单一数据仅能判断格式和缺省情况，无法得知数据内容是否正确，因此本文通过数据间的关系发现异常。异常主要分为 3 种类别：

（1）不稳定：固定时间间隔发送的数据出现了缺失、断连的情况。

（2）不完整：运行触发的数据缺失或数据无法完整反映车辆运行过程。

（3）无效数据：数据的先后逻辑违背、数据间的组合逻辑相互违背。

数据检测是发现上述异常的过程，不同类别的检测需要不同的数据，但不同数据发送频率各不相同，且触发条件也不一致，难以采用统一的方式进行处理。因此本文按照异常检测的时间间隔对协议进行简单分类，主要分为瞬时协议、短周期协议、固定时间协议和长周期协议四种，见表 1。

对不同的协议类别与检测时间间隔，采用不同响应时间的缓冲区积累数据。将其分类存储后，便可进行统一的检测流程。针对上文异常的 3 种类别，本文从稳定性、完整性和一致性三种角度对数据的异常关系进行检测，并提取出检测值。

由于设备异常一旦发生，会导致异常值持久地

或频繁地出现,因此一次程度较轻的异常检测值并不能代表异常事件的真正发生。需要对异常检测值进行累计,达到触发条件才能生成为异常事件。异常事件会相互影响,组合推导出新的事件。例如设备连接性差与整体数据包丢失率高都与网络有关,因此可以相互组合为网络通讯异常。事件组合推导将底层事件提炼为更易于人所感知并理解的组合事件,最终得到包含故障信息的异常事件集。

表 1 协议分类

Tab. 1 Protocols classification

类别名	检测条件描述	检测时间间隔
瞬时协议	单一数据即可判断异常,一旦收到该数据可直接进行检测	瞬时
短周期协议	数据发送频次高,短时间内即可积累足量数据进行检测	较短
固定时间协议	具有固定的检测时间,达到规定时间即可进行检测	固定时间
长周期协议	数据发送频次低,需要较长时间积累足量数据进行检测	较长

本文的异常检测方法主要包含异常检测值提取、底层事件累计生成和事件组合推导三个部分,其工作流程如图 1 所示。

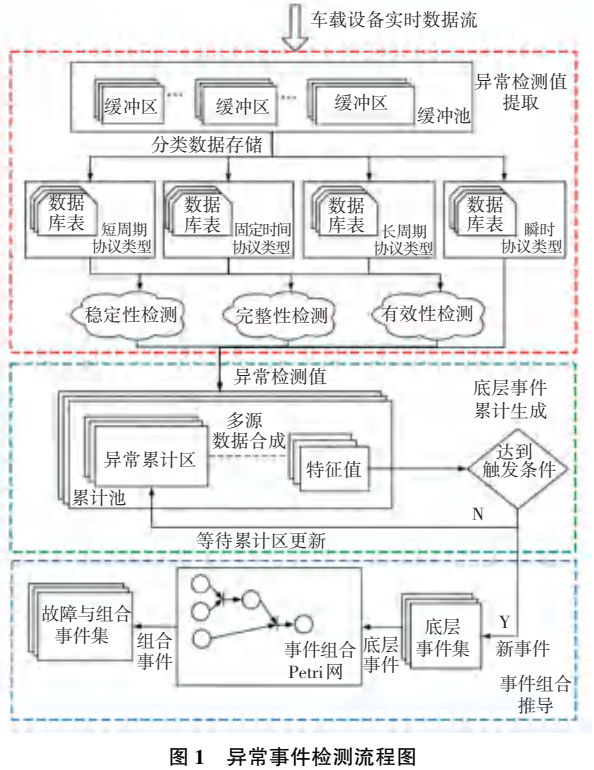


图 1 异常事件检测流程图

Fig. 1 Flow chart of abnormal event detection

车载设备数据流简单分类后被缓冲区接收后持久化到数据库表内。从稳定性、完整性和一致性三种角度提取异常检测值。累计池对这些异常检测值进行收集,通过多源数据合成特征值,并判定是否达到事件生成的条件。产生的新事件在事件组合 Petri 网内进行组合推导,推导出全部组合事件和故障事件。

1.1 异常检测值提取

1.1.1 稳定性检测

稳定性检测是判断固定时间间隔发送的数据是否出现了缺失、断连等一系列不按照规定要求稳定发送数据的情况。划分时间片断示意如图 2 所示。图 2 中,X_s 为按照顺序排布的原始单物联数据,s_{ti} 表示单个数据点,t_i 为数据的发送时间。

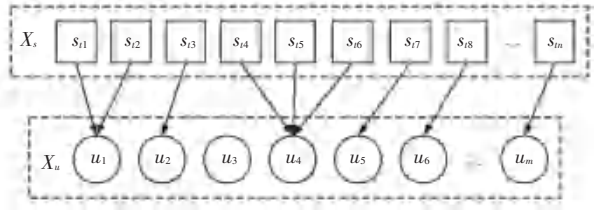


图 2 划分时间片段示意图

Fig. 2 Schematic diagram of time division

单位时间内数据量的大小是度量短时数据稳定性的重要依据,设数据发送时间周期为 k,则该数据序列 X_s 的总运行时长 $t_n - t_1$ 内共可以得到 $m = (t_n - t_1)/k$ 个时间片段。将 X_s 内的数据点按照其所在时间区间放入对应的 u_j 内,得到了一条新序列 $X_u = \{u_1, u_2, \cdots, u_n\}$,其中 u_j 表示第 j 个时间片段内所含数据的数量。在理想情况下,每条数据均按照规定时间间隔发送,则 $u_j = 1, j \in \{1,2,\cdots,m\}$,参见图 3 中的 u_2、u_5 和 u_6。但由于数据发送不稳定的情况存在,实际运行情况下 $u_j \to \{0,1,\cdots,k\}$,参见图 3 中 $u_3, u_j = 0$ 表示该时间片段内的数据缺失。$u_j > 1$ 表示该时间片段收到多条数据,参见图 3 中的 u_1、u_4,这可能是由于缺失的数据在网络通讯恢复后一并发送的结果。

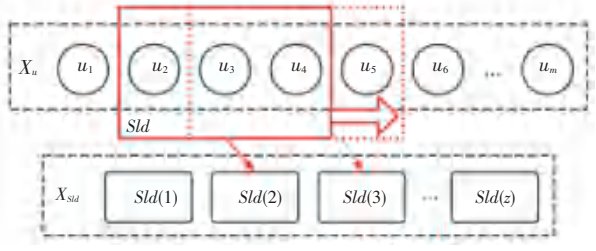


图 3 滑动窗口检测示意图

Fig. 3 Schematic diagram of sliding window detection

针对数据不稳定的情况,滑动时间窗口通过计算一段时间内收到数据量的变化情况并设定阈值,

可以发现短时区间的数据的不稳定，也能过滤数据的正常波动带来的影响。因此本文定义 *Sld* 为检测数据稳定性的一个滑动窗口，其长度为 β 个时间片段的总时长，由图 3 可知，*Sld* 长度为 3 k。通过数据缺失率和时间片段数据量的最大差值来计算该窗口的不稳定率，即：

$$\begin{cases} C_{sld}(j) = \sum_{i=j}^{j+\beta} u_i \\ W(j) = 1 - \dfrac{C_{sld}(j)}{\beta} \\ Sld(j) = \dfrac{u_{\max} - u_{\min} + 1}{C_{sld}(j)} * W(j) \end{cases} \tag{1}$$

其中，$Sld\ (j)$ 表示第 j 个滑动窗口的不稳定率；$W(j)$ 表示窗口内数据缺失率；$C_{sld}(j)$ 表示窗口实际收到的数据量；β 为窗口内时间片段的数量，即应得数据量；$u_{\max}$ 为时间片段收取数据最大量；$u_{\min}$ 为最小量。为避免差值为 0 导致最终不稳率定为 0 的情况出现，因此设定 1 为差值默认最小值。设一组数据序列共检测出异常窗口数 λ_{sld} 个，则不稳定性的异常检测值可由如下公式计算得出：

$$V_{sta} = \frac{\lambda_{sld}}{m - \beta + 1} \tag{2}$$

1.1.2　完整性检测

完整性检测是判断该段数据是否能描述车辆一段完整运行过程，因此数据的缺失率是异常判断的重要依据。车载设备的数据既包含按照固定时间间隔发送的数据，例如周期位置信息、握手连接数据等，也包含运营车辆在工作中随着行进流程触发的事件，例如到达目标地发送的到站信息、驶出站点信息等。

针对运行触发的数据，需要判断其是否与触发条件对应且无缺失。通过比对实际获取到的触发数据与当日行车量调度安排的线路、站点等信息，将不匹配或缺失的数据记为一个异常点。设异常点的发生次数为 *Del*，调度安排的全部触发事件总数为 *Tal*，则异常检测值 V_{int} 的数学定义式具体如下：

$$V_{int} = \frac{Del}{Tal} \tag{3}$$

针对固定时间间隔发送的数据 $X_s = \{s_{t1}, s_{t2}, \cdots, s_{tn}\}$，上文的滑动时间窗口 *Sld* 对每个短时区间内的数据丢失情况进行了检测，但无法排查整体的数据丢失的问题。例如每个滑动窗口都达到了最低数据量要求，但数据总量却缺失较大，这可能预示着车载设备发生了规律性掉线或重启的异常。需要对数据量的整体缺失情况进行检测，设序列实际获取数据量为 n，通过序列中数据的最晚和最早发送时间差值 $t_n - t_1$ 与该种数据规定的发送时间间隔 F_j 可以得到应得数据量，得到异常检测值 V_{int}。其计算公式的数学表述如下：

$$V_{int} = 1 - \frac{n * F_j}{t_n - t_1} \tag{4}$$

1.1.3　一致性检测

上述 2 种检测方式均是对数据外部特征的检测，不涉及数据内容的判断。一致性检测是通过数据具体内容对其先后顺序、数据间的组合逻辑进行异常判断。例如车辆的周期位置信息到达了站点附近，但却缺失相应站点的到/离站数据，这预示着报站模块的异常。通过将实际运营情况、相关协议和检测人员的专业知识相结合，预估出所有可能的异常情况，计算异常点或异常发生时长与全部运行数据的占比得到检测值。异常点的计算与式(3)相同，异常发生时长的检测值 V_{val} 的数学公式为：

$$V_{val} = \sum_{i=1}^{n} \frac{et_i - st_i}{T_e - T_s} \tag{5}$$

其中，et_i 表示第 i 个异常发生的结束时间；st_i 表示第 i 个异常发生的开始时间；数据序列的开始时间为 T_s；结束时间为 T_e；通过计算异常时长占比即可得到异常检测值 V_{val}。

1.2　底层异常事件生成

底层异常事件是对数据检测出的全部异常情况的统称，反映某一时段内检测值的总体情况。同一种检测方式下的异常检测值序列可表示为 $V = [V_1, V_2, \cdots, V_n]$，$V_i \in [0,1]$。车载设备正常运行时，检测值序列的每个值均接近或等于 0，偶尔出现小的波动。车载设备故障真正发生时，一类故障会影响多种数据，使其出现程度严重的、较为频繁的或较为持久的异常波动。

因此一次普通的异常检测值不能直接作为故障的成因，当异常检测值出现以下 3 种情况时可以记录为异常事件：

(1)个别数据的异常程度严重，检测值接近或等于 1。

(2)一段时间内数据频繁地出现异常且检测值较高。

(3)较长的时间区间内稳定地出现异常情况。

针对上述 3 种异常事件的判别条件，本文设置了一个包含三级累计区的异常累计池，并将累计区内序列的特征简单提取后合成为特征值，判断是否

达到底层异常事件生成的触发条件。

1.2.1 三级时长累计池

累计池工作流程如图 4 所示,池内包含 3 个累计区,每种累计区设有不同长度的累计周期时间。一级累计区存放的是最新时间段的异常检测值,用于判断短时区间内是否出现严重异常。二级累计区存放较为近期的异常检测值,判断中等时区内是否频繁地出现较高检测值情况。三级累计区存放的是较为长期的异常检测值,判断是否持续出现异常检测值情况。

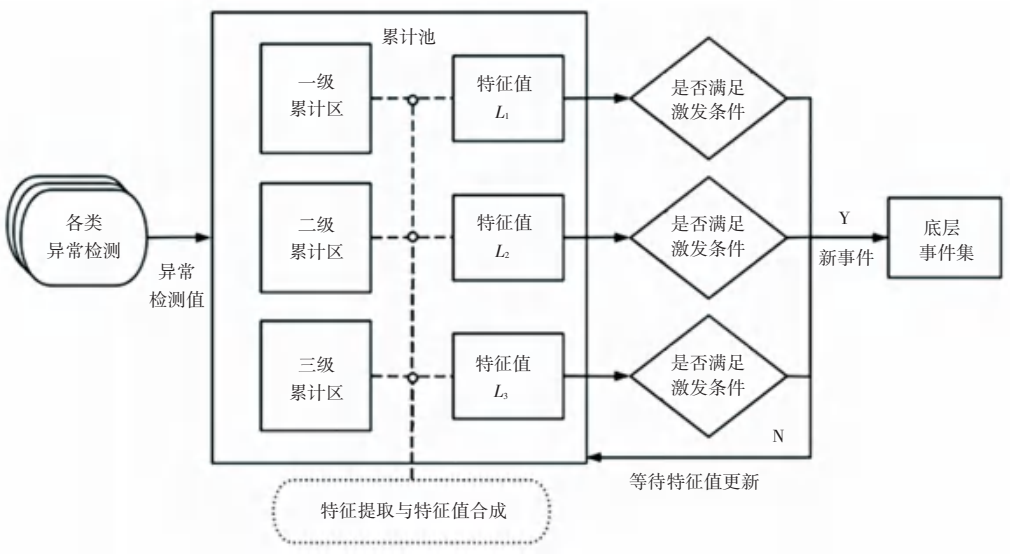


图 4 累计池工作流程图

Fig. 4 Work flow chart of cumulative pool

最新周期的异常检测完成后,各级累计区将最早的数据清除并把新异常检测值放入队尾。待新检测值进入累计区内,重新计算该区的特征值。若 L_i, $i \in \{1,2,3\}$,满足触发条件,则将生成的新事件放入底层事件集,反之则继续等待新检测值的输入。

1.2.2 合成特征值

传统的累计方法采用滑动窗口记录异常数量,由于上文进行了数据异常检测,每个检测值都是对异常程度的推断,无法通过简单地进行数量累计,并且传统的累计无法区分异常严重程度、异常发生频次和稳定地出现异常三种情况,因此本文对区内所有检测值进行特征的简单提取,并进行多源数据合成。

异常并非每次检测都会出现,因此检测值常常出现值为 0 的情况,直接对其使用多源数据融合容易造成冲突的巨大化。为了避免该情况,在提取特征时将所有值为 0 的数据剔除,以非零检测值数量占比作为衡量序列内数据的一个特征。

针对一级累计区判断短时区间内严重异常的目标,选取累计区内前 k 个最大异常检测值 $\lambda_i, i \in \{1,2,\cdots,k\}$ 作为特征,得到特征序列 $FE_1 = \{\lambda_1, \lambda_2,\cdots,\lambda_k\}$;针对二级累计区判断中等时间区间内频繁出现较高异常检测值的目标,选取非零检测值数量占比 γ_1、检测值中位数 γ_2 以及异常检测逐差平均值 γ_3 作为特征,得到特征序列 $FE_2 = \{\gamma_1,\gamma_2,\gamma_3\}$;针对三级累计区判断长时区内稳定出现异常的目标,选取非零检测值数量占比 δ_1、异常检测平均值 δ_2 作为特征,得到特征序列 $FE_3 = \{\delta_1,\delta_2\}$。

D-S 证据理论是一种不确定性推理方法,已大量应用在各类数据融合系统中,其优点是可以在先验知识未知的情况下对多源数据进行融合,即建立在一个非空集合 Θ 上,Θ 由一系列互斥且穷举的对象构成,即 $\Theta = \{\theta_1,\theta_2,\cdots,\theta_n\}$,对于论域中的任意命题 A 均属于 2^Θ,其基本概率函数 $m:2^\Theta \to [0,1]$,且满足:$\sum_{(A\in\Theta)} m(A) = 1$ 且 $m(\varnothing) = 0$。此处需用到的数学公式可写为:

$$\begin{cases} k = \sum\limits_{A_1\cap A_2\cap A_3\cap\cdots=\varnothing} m_1(A_1) * m_2(A_2) * \cdots * m_n(A_m) \\ q(A) = \dfrac{1}{n}\sum\limits_{1\leqslant j\leqslant n} m_j(A) \\ m(A) = \sum\limits_{A_i\cap B_j=A} m_1(A_1) * m_2(A_2) * \cdots * m_n(A_m) + k * q(A) \end{cases} \tag{6}$$

特征序列 $FE_i, i \in \{1,2,3\}$ 内的每个特征均是对同一问题领域的不同证据,可视为多源数据。为保证各命题最终结果之和为 1,弱化冲突带来的误差影响,本文采用文献[9] 中的证据理论合成公式

对累计区的特征序列进行融合，具体参见式(6)。这里，$m(A)$ 为事件融合后的结果值，$k*q(A)$ 为证据冲突情况下的概率分配值，n 为全部证据源的个数，$A_i, i \in \{1,2\cdots,m\}$ 为辨识框架的各个元素，$m_j(A_i)$ 为第 j 个证据源对 A_i 的基本概率赋值。在本节中，n 为特征个数，辨识框架 $\Theta = \{A_1, A_2\}$，此处的 A_1 表明事件判定为异常的情况，A_2 表明事件判定为正常的情况。以 FE_2 为例的基本概率赋值见表 2，通过计算得到各个累计区特征融合的累计结果值 $L_i, i \in \{1,2,3\}$。

表 2　以 FE_2 为证据集的基本概率赋值

Tab. 2　Basic probability assignment with FE_2 as evidence set

	A_1	A_2
m_1	γ_1	$1-\gamma_1$
m_2	γ_2	$1-\gamma_2$
m_3	γ_3	$1-\gamma_3$

1.3　事件组合推导

1.3.1　事件描述与分类

1.1 和 1.2 节分别介绍了异常的检测与底层事件的累计生成过程，但底层事件都是针对数据的某一类具体检测而来的，仅能反映数据间的异常情况，无法反映设备异常的具体现象或故障。车载设备故障会在数据上得以体现，一类故障会影响多种数据，而一类数据的异常也可能是多类故障共同的影响，其间复杂的关系难以通过简单的映射来表示。通过对异常事件建模，剥离其中复杂的相关性，将异常事件进行组合和推导可以发现更为一般性的故障问题，为检修人员提供更为可靠和易于理解的异常信息。因此将累计池内生成的事件与组合推导而来的事件进行区分，事件分类的定义见表 3。

表 3　事件分类

Tab. 3　Events classification

类型	事件分类描述
底层事件	由异常检测与累计池得到，反映数据间异常的事件
组合事件	底层事件与设备故障的中间层，由具有相关性的事件组合而成，可以通俗地反映设备出现的异常现象
故障事件	某一设备模块的故障，是组合事件推导的最终结果

1.3.2　基于概率 Petri 网的事件推导模型

Petri 网是对事件描述与建模分析的有力工具，为适应不同事件的各种组合推导方式，本文引入概率 Petri 网(PPN)。PPN 省去了模糊 Petri 网(FPN)的语言变量和模糊推理逻辑，以阈值控制变迁的触发，无需事前产生模糊推理规则，更加地简便。

PPN 定义为一个八元组，记为 $\sum = (S,T;F, W_t,M,P,f,V)$。其中 $(S,T;F)$ 是一个传统网系统；$W_t: F \to [0,1]$ 是有向弧上的概率权值，默认为 1；$P: S \to [0,1]$，$P(s_i)$ 则是库所 s_i 内标识的概率值。$V: T \to [0,1]$ 是变迁上的阈值集合；M 为各库所的状态标识；t 在 M 上享有发生权的条件为：$\forall s_i \in \cdot t_j: M(s_i) > 0 \wedge f(t_j) > V(t_j)$。$f$ 为定义在变迁的概率计算函数，函数形式见如下：

$$f(t_j) = \frac{\sum_{i=1}^{n} P(s_i) * W_t(F(s_i,t_k))}{n} \tag{7}$$

其中，n 为满足条件的 $s_i \in t_j$ 的元素总数。变迁发生后产生的新标识值 $P(s_i)$ 由前置变迁集的最大值得出：

$$P(s_i) = \max\{f(t_j) \mid t_j \in s_i\} \tag{8}$$

PPN 网的基本型如图 5 所示。图 5 中，圆圈表示库所，带箭头的线段表示有向弧，竖线表示变迁。假设 t_1、t_2 均满足发生条件，$f(t_2) > f(t_1)$，因此输出库所的概率值 $P(s_3)$ 为 $f(t_2)$。

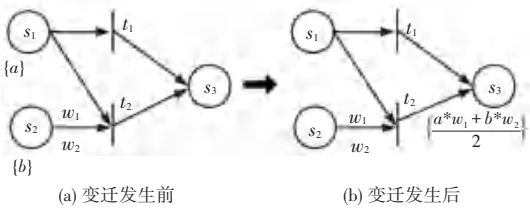


图 5　PPN 变迁示意图

Fig. 5　Transition diagram on PPN

车载设备功能众多，以报站模块为例，其主要工作内容为判别车辆是否到达目标点位，并发送到站数据包和出站数据包。通过 1.1.2 节对异常的分析可以得到 3 种检测方式：一致性下的到达规定位置不报站和到站/离站数据不对应，以及完整性下的报站信息缺失。以这 3 种检测方式为底层事件，车载设备不报站为组合事件，报站模块故障为终点建立事件组合 Petri 网。

报站模块事件组合 Petri 网如图 6 所示。图 6 中，s_1、s_2、s_3 为底层事件库所，分别是到站 / 离站数据不对应事件、报站信息缺失事件和到达规定位置不报站事件。s_4 为组合事件库所，表示车载设备不报站事件。s_5 为报站模块故障事件库所。t_1、t_2 变迁代表事件的组合。有向弧上的概率权值 w_1、w_2、w_3、w_4 代表了事件组合传导的概率，其值通过历史维护信息与维修人员对异常情况和故障的统计得到先验概率，采用贝叶斯定理的后验概率计算公式得出：

$$P(B \mid A_i) = \frac{P(B)P(A_i \mid B)}{P(A_i)} \tag{9}$$

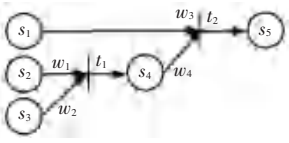


图 6 报站模块事件组合 Petri 网

Fig. 6 Petri nets of station reporting module event combination

当累计池的底层异常事件触发时,将其作为标识放入组合 Petri 网的底层事件库所内,其概率值为事件生成时的特征值,组合事件以及故障的概率值由上层事件概率值与概率权值通过式(8)计算得到,最终推导出所有事件及其发生概率值。

2 实验及分析

2.1 实验数据

本文选用某公交公司 3~11 月期间 46 台车辆的运营数据,包含车载设备的行驶数据和维护报告,由于协议内容众多且包含与异常检测无关的数据,选择其中的 5 种协议作为实验数据,车载设备协议表的内容见表 4。

表 4 车载设备协议表

Tab. 4 Protocol table on on-board equipments

协议编号	协议名称	触发条件	数据内容
DS_1	通讯网络连接	以 5 min 为周期握手	设备序列号、线路名称、线路编号
DS_2	周期位置信息	以 10 s 为周期发送	纬度、经度、瞬时车速、方位角
DS_3	设备自检状态	以 2 min 为周期发送	WiFi、CAN、投币机、DVR、RFID 状态
DS_4	车辆到达站点	到达特定点自动发送	纬度、经度、站点编号、线路编号
DS_5	车辆离开站点	离开特定点自动发送	纬度、经度、站点编号、线路编号

2.2 评价指标

实验目标:在确保尽可能地将异常设备全部检出的前提下,减少误判为异常的设备数量。以查准率和召回率判别方法的准确性。设 TP 为异常设备被正确检出的样本数,FP 为正常设备被误判为异常的样本数,FN 为异常设备被误判为正常的样本数。查准率和召回率的公式分别是:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

查准率和召回率是一对矛盾的度量,通过提高检测标准可以提高查准率、降低召回率,但相应会漏掉许多异常设备。降低检测标准可以提高召回率、降低查准率,带来更多的误判。为了权衡这 2 个指标,取二者调和值 $F-Score$ 作为评判标准,以 β 为加权系数,进而得到:

$$F - Score = (1 + \beta)^2 \frac{Precision * Recall}{\beta^2 * Precision + Recall} \tag{12}$$

β 的取值影响 2 个指标的重要性比例,当 β 为 1 时,二者同样重要;当 $\beta > 1$ 时,召回率更为重要。相反,当 $\beta < 1$ 时,查准率更为重要。由于本方法的目标是为检修人员提供设备排查的优先级和具体的异常信息,提高检测的效率。相比于查准率,召回率更能体现本方法的可行性,因此选择 F_2 分数作为评价指标,将召回率的重要程度设定为查准率的 2 倍。

2.3 实验结果及分析

实验分为 2 个部分。实验一评估数据异常检测结果与底层事件累计生成的情况,并对其结果进行分析。实验二通过事件组合推导出全部异常事件,与实际结果比对验证准确性。

2.3.1 数据检测结果与异常事件生成情况

结合表 4 中车载设备协议类型与 1.1.2 节的异常检测角度,共得到 12 种检测类别,见表 5。该部分实验选择 3~9 月期间所有车辆的运营数据,共计约 11 万趟次。以车辆一趟运行时长作为检测的时间周期,通过表 5 中的各类检测,得到半年间所有趟次的数据检测结果。

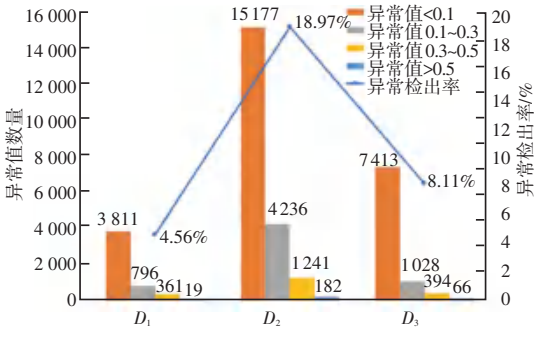
表 5 具体异常检测类别

Tab. 5 Anomaly detection categories

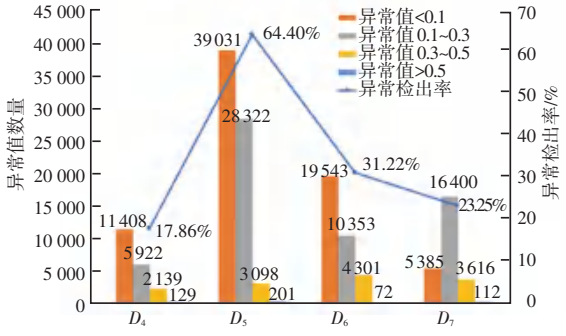
检测类型	检测编号	检测异常名	所用协议编号
稳定性检测	D_1	握手数据稳定性	DS_1
	D_2	周期位置数据稳定性	DS_2
	D_3	设备自检数据稳定性	DS_3
完整性检测	D_4	握手数据完整性	DS_1
	D_5	周期位置数据完整性	DS_2
	D_6	设备自检数据完整性	DS_3
	D_7	报站信息完整性	DS_4、DS_5
一致性检测	D_8	到达规定位置不报站	DS_2、DS_4
	D_9	到站离站信息不对应	DS_4、DS_5
	D_{10}	超出最大可行驶距离	DS_2
	D_{11}	偏移线路次数	DS_2
	D_{12}	偏移线路时间	DS_2

图 7(a)~(c)分别展示了稳定性、完整性和一致性三种性能指标下各种检测方式的异常检测值分

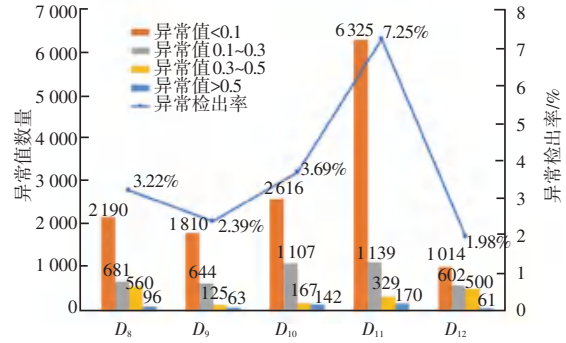
布情况柱状图。横坐标表示的检测编号与表 5 相对应。异常检测值按照严重程度从小到大排序，分为 4 类，分别是：轻（小于 0.1）、较轻（介于 0.1～0.3 之间）、较重（介于 0.3～0.5 之间）和严重（异常值大于 0.5）。图 7 中的折线表示异常检出率，反映该种检测方式下异常的发生率。



（a）稳定性检测结果



（b）完整性检测结果



（c）一致性检测结果

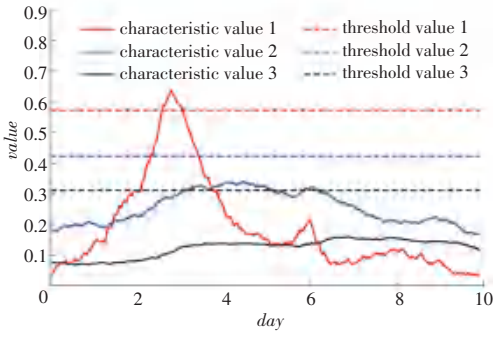
图 7　异常检测值统计结果

Fig. 7　Statistical results of abnormal detection value

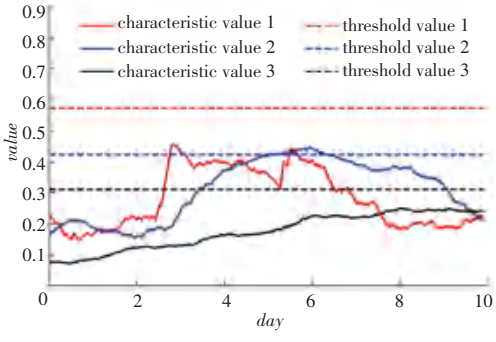
通过对图 7 的分析可以得出：完整性检测的异常检出率远高于其他 2 类，但程度较重的检测值数量占比远小于其他 2 种检测类别。说明车载设备由于信号波动或者网络异常造成的小段数据丢失情况较为普遍。一致性检测异常检出率最低，但程度较重的检测值数量占比高于其他 2 种检测类别，说明该种检测类别对异常更为敏感，具有针对性。D_2、D_5 和 D_{11} 在图 7 中异常检出率远高于其他同类别检测方式，但程度较重的异常值数量占比与其他检测方式并无差异。由于这 3 种检测方式都只用到 DS2 协议编号，说明协议周期时间越短，则数据积攒的数目越多，异常检测效率越高。

针对公交车载设备的运营安排，设定一级、二级和三级累计区的累计时间 $T_i, i \in \{1,2,3\}$ 分别为 1 日、3 日和 7 日。对异常发生下的特征值进行监测得到其变化情况。

图 8 是 3 个累计区的特征值变化情况。图 8（a）是累计池在收到短暂的、且程度严重的异常检测值下特征值的变化情况。



（a）情况一



（b）情况二

图 8　特征值变化趋势图

Fig. 8　Change trend of characteristic value

一级特征值在严重检测值出现当天快速上升到了最高点后、下降至较低水平，变化趋势明显且快速。二级特征值产生了较小幅度的增长、并在随后 3 天均保持稳定，当程度严重的异常值因超出累计周期时间被淘汰后，二级特征值缓慢下降。而三级特征值增长和变化幅度均不明显，说明一级累计区对于短时间内严重异常值判别效果较好。图 8（b）是累计池在第 3～7 日频繁收到较高异常检测值下特征值的积累情况。一级特征值的上下波动较大，但特征值均不超过 0.5，难以作为短时间严重异常事件的生成条件。二级特征值在第 3 日开始呈连续增长的趋势、并且增速较快，至第 6 日达到最高点，随

后缓慢下降。三级特征值呈缓慢增长的趋势,在第 8 日达到最高值、并保持稳定。说明二级累计区对于一段时间内频繁出现较高异常值的情况判别效果较好,三级累计区对异常检测值在更长时间区间频繁出现的情况会有更好的累计结果。

为保证历史数据的故障全部检出,各累计区以历史故障发生情况下的最低特征值作为阈值,最终得到 3~9 月期间全部车辆的底层事件的生成结果,如图 9 所示。图 9 中,每一个柱形反映每种底层事件的生成数量,底层事件库所编号与表 6 中的库所编号一一对应,每个柱形从下至上 3 种颜色分别反映一级、二级和三级累计区的事件生成数目。

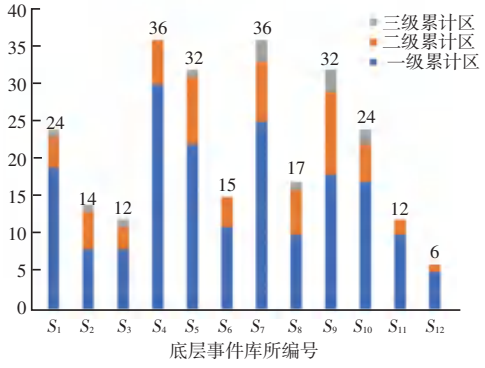


图 9 累计池底层事件生成统计结果

Fig.9 Statistical results for the events generated by the cumulative pool

分析图 9 可以得出:底层事件主要由一级累计区生成,二级与三级累计区生成的占比依次减少,分别为 70.4%、24.6%和 5.0%。说明一日的累计可以判别出大部分的底层异常事件,较长时间周期的累计区可以捕捉少部分遗漏的底层事件。与图 7 内各种检测结果比对可以得出,异常检测值的数量与底层事件数量无线性关系,并且平均各个底层事件生成数占异常检测值数目约 0.12%,证明累计池可以有效地过滤掉大量无法推导故障的冗余异常检测值。

2.3.2 异常事件推导结果

为了使用概率 Petri 网进行故障的推理计算,需要获得组合事件发生下模块故障的条件概率。统计 3~9 月的维护数据,得到故障综合概率为0.072 4%台/天,其中报站模块 0.013%台/天、GPS 模块 0.035%台/天、网络通讯模块 0.024%台/天。

针对组合事件与最终故障推导的不确定性,通过检修人员对故障情况下各类组合事件的发生现象进行判断,得到故障情况下组合事件的发生概率。将上述概率作为先验概率,通过式(10)求出各个有向弧的概率权值,结合 1.3.2 节 Petri 网定义建立事件组合推导 Petri 网。全部事件集见表 6。

表 6 车载设备异常事件集

Tab. 6 Abnormal event set of on-board equipments

库所编号	底层事件名	库所编号	组合与故障事件名
S_1	到达规定位置不报站	S_{13}	车载设备不报站
S_2	报站信息缺失	S_{14}	部分位置信息偏移
S_3	到站离站数据不对应	S_{15}	偏移线路
S_4	相邻点位超出最大可行驶距离	S_{16}	GPS 连接性差
S_5	频繁偏移线路	S_{17}	运行数据丢失
S_6	长时间偏移线路	S_{18}	状态数据丢失
S_7	GPS 总体丢失率高	S_{19}	设备连接性差
S_8	设备自检数据总体丢失率高	S_{20}	报站模块故障
S_9	握手数据总体丢失率高	S_{21}	GPS 信号漂移
S_{10}	GPS 数据不稳定	S_{22}	GPS 模块故障
S_{11}	设备自检数据不稳定	S_{23}	整体数据丢失率高
S_{12}	握手数据不稳定	S_{24}	网络通讯模块故障

车载设备异常事件组合 Petri 网如图 10 所示。图 10 中,S_1 ~ S_{12} 为底层事件的库所,由累计池生成而来,其库所内标识的概率值 P 为事件生成时的特征值。S_{13} ~ S_{24}均由底层事件或其他组合事件推导而来,其库所内标识的概率值 P 由所有指向该库所的变迁通过式(8)计算而来。其中,S_{20}、S_{22}、S_{24} 为最终的故障库所。

该部分实验选用 3~9 月的数据和维护报告对 Petri 网中各个变迁的阈值做调节,9~12 月的真实数据作为数据集Ⅰ,计算方法的最终准确率。由于

3 个月内真实的故障发生次数过少，难以验证本方法的准确性，因此在原数据集的基础上建立仿真数据集Ⅱ和Ⅲ。收集故障发生下的异常检测值和特征值，在周期为 3 个月的运行数据内，选择随机日期、随机车辆的正常数据替换为故障数据，得到数据集Ⅱ。在故障发生下的历史数据最大边界值内对异常数据进行浮动，并作为替换数据插入至正常数据集内，得到数据集Ⅲ。针对每种仿真数据集均进行多次实验，得到平均异常结果见表 7。

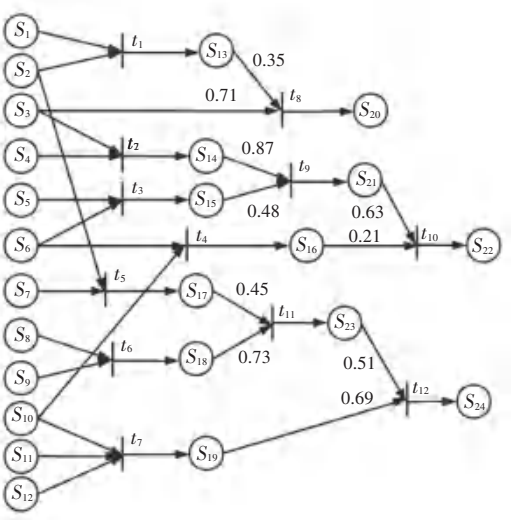


图 10　车载设备异常事件组合 Petri 网

Fig. 10　Petri nets for combination of on－board equipments abnormal events

表 7　车载设备故障检测结果

Tab. 7　Fault detection results of on-board equipments　%

数据集	*Precision*	*Recall*	F_2
Ⅰ	69.8	94.6	88.3
Ⅱ	74.2	96.7	91.2
Ⅲ	61.4	92.4	83.9

由表 7 可知，每行对应不同数据集下的查准率、召回率和 F_2 均值。本文的目标是通过对设备的异常检测为检修人员提供可能发生故障的设备，以提高设备的检修效率。数据结果表明，本方法对于 3 类数据集的召回率均能保持较高的水平，并且在满足高召回率的基础上适当兼顾了查准率，可以有效检测出可能发生故障的设备。

3　结束语

本文提出一种面向车载设备数据流的异常检测方法，从稳定性、完整性、一致性三种角度检测数据间的异常。针对各种异常情况设置不同时间周期的累计区，通过证据理论合成公式对区内数据的特征进行融合，过滤数据波动带来的异常误判情况。分析底层异常事件、设备故障与组合事件的关系，使用概率 Petri 网建立车载设备异常事件组合模型推导设备故障。实验结果表明，该方法可以过滤掉数据波动带来的误判，有效地检测出可能发生异常的车载设备，异常检测 F_2 均值接近 84%。但由于历史故障数据量较少，难以形成有效的数据集调控各个阈值权重。如何在检测中出现新故障数据的情况下，动态地调控累计池以及组合 Petri 网的参数，提高检测的查准率，还有待进一步的研究。

参考文献

[1] 王德文，杨力平. 智能电网大数据流式处理方法与状态监测异常检测[J]. 电力系统自动化，2016，40(14)：122-128.

[2] 庞亮. 马兰矿选煤厂典型设备在线远程智能预测性维护系统的应用[J]. 煤炭加工与综合利用，2019(12)：19-22.

[3] 王璇，李倩丽，宋焕生，等. 基于 AdaBoost 的公交客流量统计算法[J]. 计算机应用研究，2018，35(03)：949-952.

[4] 孙宁 ,吴伟豪，赵风财，等. 基于增强型 Dijkstra 算法的无信号灯交叉路口智能车辆调度研究[J]. 计算机应用研究，2022，39(01)：188-193.

[5] 马书红，张勐豪. 镇村公交下一种新型智能公交到站时间预测算法[J]. 计算机应用研究，2016，33(04)：1044-1046,1061.

[6] CHEN Junfu，PI Dechang，WU Zhiyuan，et al. Imbalanced satellite telemetry data anomaly detection model based on Bayesian LSTM[J]. Acta Astronautica，2021，180：232-242.

[7] 石亚凡. 电力计量装置电压异常在线检测方法研究[J]. 科学技术创新，2021(34)：191-193.

[8] LIU Liansheng，LIU Datong，ZHANG Yujie,et al. Effective sensor selection and data anomaly detection for condition monitoring of aircraft engines[J]. Sensors，2016，16(5)：623-623.

[9] 李弼程，王波，魏俊，等. 一种有效的证据理论合成公式[J]. 数据采集与处理，2002,17(01)：33-36.

[10] 曹洁，孟兴. 一种有效解决 D-S 理论冲突证据合成的方法[J]. 计算机应用研究，2012，29(05)：1815-1817.

[11] 欧志芳，安吉尧，周芳丽. 利用 D-S 证据理论的夜间车辆检测[J]. 计算机应用研究，2012，29(05)：1943-1946.

[12] 李家伟. 基于证据理论和支持向量机的风机故障智能诊断[J]. 吉林大学学报：理学版，2016，54(03)：609-612.

[13] LIN Yun，LI Yuyao，YIN Xuhong，et al. Multisensor fault diagnosis modeling based on the evidence theory[J]. IEEE Transactions on Reliability，2018，67(2)：513-521.

[14] 张琪，胡宇鹏，嵇存，等. 边缘计算应用：传感数据异常实时检测算法[J]. 计算机研究与发展，2018，55(03)：524-536.

[15] LIS A，DWORAKOWSKI Z，CZUBAK P. An anomaly detection method for rotating machinery monitoring based on the most representative data[J]. Journal of Vibroengineering，2021，23(2)：861-876.

[16] QU Liping，LIU Chongjie，LU Zhao，et al. Classified fault diagnosis of power grid based on probabilistic Petri net[C]// Proc of the 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). Piscataway，NJ：IEEE Press，2019：234-237.

(下转第 63 页)