刘尧,邓晓衡,桂劲松,等. 弱链路下基于全对等模式和相关性检测的资源信息共享[J]. 湖南科技大学学报(自然科学版),2021, 36(3):88-96. doi:10.13582/j.cnki.1672-9102.2021.03.013

LIU Y, DENG X H, GUI J S, et al. Resource Information Sharing Based on Full Peer Mode and Correlation Detection for Weak-Link-Oriented Communication Environments [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2021, 36(3):88-96. doi:10.13582/j.cnki.1672-9102.2021.03.013

弱链路下基于全对等模式和相关性检测的资源信息共享

刘尧[1],邓晓衡[1],桂劲松[1*],刘斌[2],付琨[2]

(1.中南大学 计算机学院,湖南 长沙 410083;2. 中国科学院 空天信息创新研究院,北京 100094)

摘 要:针对恶劣网络通信环境导致的数据采集和信息交互异常问题,提出了一种基于多维时序数据子序列的相关性检测方案.该方案针对相同时间段内的多维资源状态时序数据进行相关性计算.在获得其相关性矩阵后,先与已训练好的正常检测模式比对,若结果被判定为异常,再与已训练好的异常检测模式比对,若结果仍被判定为异常,则最终认定结果为异常.通过这种双重检测,尽可能避免对正常数据的误判.仿真结果显示:该方案在确保检测实时性的同时提高了检测的准确率.

关键词:弱链路;对等模式;相关性检测;资源信息;共享

中图分类号:TP393 **文献标志码**:A **文章编号**:1672-9102(2021)03-0088-09

Resource Information Sharing Based on Full Peer Mode and Correlation Detection for Weak-Link-Oriented Communication Environments

LIU Yao[1], DENG Xiaoheng[1], GUI Jinsong[1], LIU Bin[2], FU Kun[2]

(1. School of Computer Science and Engineering, Central South University, Changsha 410083, China;
2. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China)

Abstract: To solve the problem of abnormal data in the process of data collection and information interaction caused by bad network communication environment, a correlation detection scheme based on sub-sequences of multidimensional temporal data was proposed. The scheme calculated the correlation of multi-dimensional resource state temporal data in the same period. After obtaining the correlation matrix, it was firstly compared with the trained normal detection mode. If the result was judged as the abnormal one, it was then compared with the trained abnormal detection mode. If the result was still judged as the abnormal one, it was finally identified as the abnormal one. Through this dual detection, as far as possible to avoid misinterpretation of normal data. Simulation results show that the proposed scheme improves the detection accuracy while ensuring real-time performance.

Keywords: weak-link; peer mode; correlation detection; resource information; sharing

信息系统已与人类活动的一切紧密相连、密不可分.但是,在自然灾害等突发事件发生时或处在战场环境下,支撑信息系统的网络基础设施是没有保障的或者是根本不存在的.然而,现代应急救援决策或战

收稿日期:2020-08-04

基金项目:全军共用信息系统装备预研专用技术项目资助(315105202)

* **通信作者**,E-mail: jsgui06@163.com

场指挥决策急需信息系统发挥人力难以胜任的作用.正确的决策需要大量的、及时的现场信息采集以及强大的信息处理能力.无论针对灾难事件还是战场环境,信息数据采集现场与具备强大信息处理能力的设备通常相距遥远.即使近处存在强大信息处理能力的设备,在重大灾难事件发生时,也很可能遭到破坏,而在战场环境下则更不能有这种假设.

网络通信系统承担将现场采集的数据传输到遥远的信息处理中心的任务.然而,在破坏性极强的环境下,很难保证网络通信系统总是可用的.依靠应急救援人员或战地单兵携带的集感知、通信、处理等功能于一体的信息终端,可以实现现场及时的数据采集、邻近区域自组建网、就近协同信息处理,并提供邻近连通区域内的自主决策.同时,在应急救援现场或战场环境下,也会有应急救援车辆、无人机、战地指挥车、武器装备车辆等.这些装备携带的信息设备相比于应急救援人员或战地单兵携带的信息终端在性能上高出很多,甚至多个装备也可以协同构成信息处理能力更强的集群.这些信息终端和信息装备在文中被统一称为节点.

无线自组网因其良好的自愈性、抗毁性及抗干扰性[1-3],成为将这些节点互联的首选组网模式.两两之间处于通信范围内的节点能快速、便捷、高效地连入网络.同时,当节点彼此不在通信范围内或节点损毁时,也可方便地更新网络拓扑.因此,这些节点无论能力强弱,在地位上都是平等的,而且节点融入网络和撤出网络的便捷性也能够更好地适应应急救援现场或复杂的战场环境[4-5].由于复杂地形阻碍电磁波传播[6]或人为的电磁干扰导致的恶劣通信条件,救灾现场或战地环境无线信号时常出现时断时续的不稳定现象,确保全地域与全空域中每个节点无缝连接几乎不可能.在这种不可预测的条件下,确保集数据、语音、视频于一体的网络应用的可靠性,难度极大,要求极为苛刻.

灾难现场的图像与视频信息对救灾决策有着快捷直观的效果.另外,随着情报、监视与侦察任务变得越来越复杂,战场数据交换量更大,对交换数据的可靠性要求更高、对交换速度要求更快.应急救援决策(例如,抢险现场的图像与视频数据传输与处理)或作战指挥决策(例如,探测跟踪敌方目标、计算指挥控制节点和拦截武器节点的有效协同与整合)需要实时的海量数据传输和超强算力支撑.然而,基于灾难救援现场或战场环境下无信息基础设施支撑的现实,需要节点间的有效协同来形成尽可能强的处理能力.由于节点之间的弱连接特性,互联的网络规模是动态变化的,因而能协同调度的资源总量也是动态变化的.这为基于灾难救援现场或战场环境下的有效资源管理带来了挑战.因此,有必要探讨面向弱链路通信环境的多节点全对等管理技术,为有效管理动态变化的资源,满足网络应用性能需求奠定基础.

1　问题描述

节点可能拥有的资源类型如图 1 所示.节点的无线接口类型决定了其与外界交互的可能性.若一个节点拥有的无线接口类型越少,且相应类型接口的设备量越少,则该节点被孤立的可能性就越大.另一方面,无线接口类型也间接反映了无线通信速率和通信范围,例如,WiFi 的无线通信速率高但通信范围窄,而 GPRS 的无线通信速率低但通信范围宽.

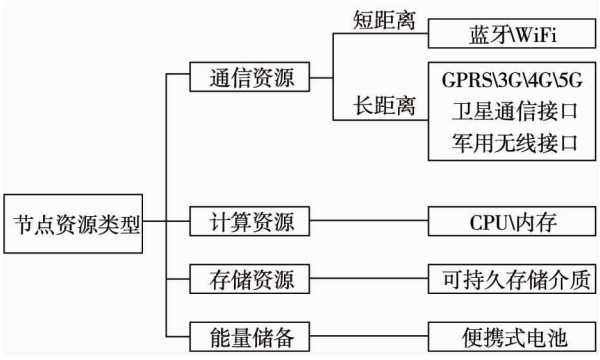

图 1　节点资源类型

个体携带的信息设备无论在计算资源、存储资源、能量储备方面,都无法与大型装备上信息设备相比.前者价格相对低廉、数量更多,是现场数据采集的主要承担者,而后者价格相对昂贵、数量更少,是数据处理、结果分析、应用决策的主要实施者.不同能力的节点的有效协同有可能使有限的资源发挥更大的效用.但前提是每个节点要知道其他节点的资源状况.

既然节点与节点之间的角色是对等的,那么任一节点愿意开放的计算和存储资源,都可以资源状态数据的形式通过广播、发布/订阅、主动推送等方式与其他节点共享.每个节点都可以自主决定是否向其他节点或向哪些节点发出资源状态数据请求.这种自主决定可以是根据自身所知资源状态信息进行决策的结

果.同样,当每个节点得到资源状态数据请求的响应后,也可自行决定是否使用得到的响应状态数据来更新本地缓存信息.例如,发生异常的响应包就应舍弃.为使每个节点总能维护实时有效的资源视图,本文将探讨一种能对多节点全对等关系进行有效管理的方案.

2 解决方案

基于上述问题描述,如何判断响应包是否异常成为本文关注的重点.异常检测研究的目标是要找到数据中不满足常态、约束、规则、给定模型的不寻常数据值或模式[7].异常检测方法通常应用在网络入侵检测、欺诈行为检测、工业损伤检测、文本异常检测、传感器故障检测等过程中.基于统计和机器学习模型的异常检测方法能较为有效地应用于各类检测任务.这些方法大致分为诸如贝叶斯分类和决策树模型等的监督检测[8]、诸如聚类算法的无监督判别以及隐马尔可夫模型的无监督参数估计[9-11]等.随着数据的时态性和时效性越来越引起人们的重视,针对时态数据的异常检测研究也受到了广泛关注.例如,文献[12]介绍了各类时态数据的异常检测问题、研究方法及应用;文献[13]关注了多维时间序列上的异常检测问题,提出的方法特别适宜于对高维时序数据相关性知识的挖掘.

另外,针对时间序列预测研究也引起了广泛关注[14-15],产生了不少研究成果.例如,文献[16]分别从时域、频域、时频域角度研究了时间序列的相关性问题;文献[17]对时序数据的挖掘与分析进行了系统性阐述,并着重讨论了时序数据的预测和有序分类问题;文献[18]提出了一种基于深度神经网络的公交到站时间预测算法;文献[19]关注长时间期预测的误差传播影响问题,提出了一种基于 LSTM 编码解码器结构的长期时间序列预测注意力模型;文献[20]提出了一种新的时间序列预测模型,能够充分学习不同区间长度的时间序列数据的特征;文献[21]提出了基于卡尔曼滤波的改进梯度增强决策树算法;文献[22]提出了一种基于经验模式分解和高阶模糊认知图的时间序列预测方法.

本文关注资源状态数据的异常检测问题,具有典型的多维时序特征,类似于文献[13]关注的问题特征.但与文献[13]相比,数据维度并不属于高维范畴,因此,若文献[13]的异常检测方法直接应用于本文场景,则代价很高且没有必要.另外,基于前述应用场景描述,对资源状态数据异常检测的准确性要求更高,因为基于误判的资源状态的任务调度导致的损失往往远高于资源得不到充分利用所导致的损失.因此,本文借鉴文献[13]对多维时序数据处理的相关性计算思路,但在异常检测算法设计上针对资源状态数据维度不大的特点进行适应性设计以降低检测算法的开销.同时,对被判定为异常的实例进行异常模式训练,以增强异常检测的准确性.

假设每个节点具备周期性检测与采集自身通信资源、计算资源、存储资源、能量储备资源可用水平的能力,其值被归一化为 0~1 的值.若值越接近 1,则可用水平越高,或者说资源空闲率越高.随着时间的推移,检测与采集次数的增多,将形成 4 个维度的时间序列数据.

定义 1 资源状态时序数据:资源状态时序数据是由节点采集的反映自身资源使用状况的一系列有时间顺序的离散数据点.一条长度 L 的资源状态时序数据被表示为 $Z=<d_1,d_2,\cdots,d_i,\cdots,d_L>$,且每个时序数据点可表示为一个二元组 $d_i=<s_i,t_i>$,s_i 表示第 i 个时序数据点的资源空闲率($0\leqslant s_i\leqslant 1, i\in\{1,2,\cdots,L\}$),$t_i$ 是对 s_i 进行采集的时间记录点.对任意 $i\in\{1,2,\cdots,L\}$ 和 $j\in\{1,2,\cdots,L\}$,若 $i<j$,则有 $t_i<t_j$.资源状态时序数据的时间点集合被记作 $\boldsymbol{T}=\{t_i\}_{i=1}^{L}$.

定义 2 多维资源状态时序数据:$\mathbb{Z}$ 是一个包含 K 条具有相同时间点集合 $\boldsymbol{T}$ 的资源状态时序数据集合,被称为 K 维资源状态时序数据,被记作 $\mathbb{Z}=\{Z_1,Z_2,\cdots,Z_K\}$,这里 $K=4$. 进一步,第 $m(m\in\{1,2,\cdots,M\})$ 个节点采集的 $K=4$ 维资源状态时序数据被记作 $\mathbb{Z}_M=\{Z_1^M,Z_2^M,Z_3^M,Z_4^M\}$.

定义 3 资源状态数据采集时间段:时间区间 $t_{[u,v]}$ 是一个节点的一个资源状态数据采集区间,从时间点 t_u 开始到时间点 t_v 结束,并且 $t_{[u,v]}\in\boldsymbol{T}$.

定义 4 资源状态时序数据子序列组:一个节点在时间区间 $t_{[u,v]}$ 上采集的全部资源状态数据,包括 $K=4$ 条具有相同起止时间的资源状态时序数据.

定义 5 资源状态时序数据子序列组之间的相关性:对于给定的具有相同时间点集合的 2 条资源状

态时序数据子序列 Z_i 和 Z_j,若 $p(Z_i,Z_j)$ 是定义在 Z_i 和 Z_j 上的相关性计算函数,则子序列 Z_i 和 Z_j 之间的相关关系被判定为(1)若 $\theta_h < p(Z_i,Z_j) \leqslant 1$,则 Z_i 和 Z_j 强正相关;(2)若 $\theta_l < p(Z_i,Z_j) \leqslant \theta_h$,则 Z_i 和 Z_j 弱正相关;(3)若 $-\theta_l < p(Z_i,Z_j) \leqslant \theta_L$,则 Z_i 和 Z_j 不相关;(4)若 $-\theta_h < p(Z_i,Z_j) \leqslant -\theta_l$,则 Z_i 和 Z_j 弱负相关;(5)若 $-1 < p(Z_i,Z_j) \leqslant -\theta_h$,则 Z_i 和 Z_j 强负相关.这里,θ_h 和 θ_l 是根据经验设置的阈值,满足$0<\theta_l<\theta_h<1$.

将每个节点上的多维资源状态时序数据作为分析单元进行研究.对于一个节点自身采集的所有维度上的序列进行相关性建模计算,不同节点上的多维资源状态时序数据的相关性独立计算.对于一段较长时间内的多维资源状态时序数据,同一个节点在同一种模式的工作周期内,资源状态时序数据的子序列之间通常存在某种较为稳定的相关性关系,而不同的工作模式可能导致这些资源状态时序数据的子序列之间相关关系的差异.由于可变相关关系的子序列可以根据其具体工作模式被分割成若干个更短子序列时间段,因此,本文仅考虑给定的子序列之间相关性阈值无较大变动的情况.

定义 6　基于相关性分析的资源状态时序数据异常检测问题:对于给定的第 m 个节点对应的 $K=4$ 维资源状态时序数据 $\mathbb{Z}_M=\{Z_1^M,Z_2^M,Z_3^M,Z_4^M\}$,需要实现如下任务:(1)设计相关性计算函数 $p(Z_i,Z_j)$ 对 $K=4$ 维资源状态时序数据上任意 2 条资源状态时序数据子序列 Z_i 和 Z_j 进行相关性计算量化,记作 $C_{Z_i,Z_j}=p(Z_i^M,Z_j^M)$. (2)根据任务(1)中的相关性标记,对待检测 $\mathbb{Z}_M=\{Z'^M_1,Z'^M_2,Z'^M_3,Z'^M_4\}$ 上可能存在的异常数据进行检测.若检测资源状态时序数据子序列存在异常,则使用二元组 $<t_{[u,v]},AD(\mathbb{Z}_M)>$ 进行记录.其中,$t_{[u,v]}$ 记录存在异常的资源状态数据采集时间段,起始于时间点 u,结束于 v;$AD(\mathbb{Z}_M)$ 记录该时间段内所有存在异常的资源状态时序数据子序列组的编号.

在本文中,每个节点维护自己采集的关于自身的多维资源状态时序数据.当采集量达到一个时间区间(例如,$t_{[u,v]}$)长度时,对该时间区间内的资源状态时序数据子序列组进行相关性计算,并将计算结果与自身维护的训练模式进行比对.若未发现异常,则保留该时间区间内的资源状态采集结果,并在自身承诺资源共享的前提下向已预订用户推送自身当前的资源状态数据.若发现异常,则舍弃该时间区间内的资源状态采集结果.当资源状态时序数据子序列组积累到一定量后,节点可以采用这些子序列组训练新的检测比对模式.新的检测比对模式与旧的检测比对模式可以采取加权合成的方式得到作为后续检测使用的比对模式.

另一方面,该节点若预定了其他节点的资源状态数据,则在收到响应包(至少包含 1 个子序列组)后,先保存待用.当累积到一定数量后,可以采用这些子序列组训练检测比对模式.再往后,每收到一个响应包,则先对其携带的子序列组进行相关性计算,再与检测比对模式比对以确定是否应该判为异常.若判为异常,则作为异常实例进行保存,当累积到一定量后,可以训练出用于检测比对的异常模式.若判为非异常,则采纳该响应包中数据以更新自身维护的其他节点资源状态数据的缓存信息,以供任务调度模块使用.

当被判定为非异常的响应包积累到一定量后,节点可以采用这些响应包携带过来的子序列组训练新的检测比对模式.新的检测比对模式与旧的检测比对模式可以采取加权合成的方式得到作为后续检测使用的比对模式.针对检测比对的异常模式也可采用类似方式进行更新.通过采用检测比对的异常模式,可以减少对本身实际属于正常的子序列组的误判,但也可能漏掉当前用于比对的异常模式无法检测出的真实异常.因此,若侧重于不误判本身正常的子序列组,则可对正常比对模式检测不通过的子序列组再次使用异常比对模式,若经检测而未被归于异常类别,则仍可接纳.

上述方案描述的核心部分包括训练阶段和检测阶段.在训练阶段,需要积累历史序列数据作为训练数据集,对所有维度上的时间序列进行相关性计算以建立检测时的比对模式.在检测阶段,我们输入待检测的资源状态时序数据子序列组,利用已训练完成的比对模式进行异常模式的识别.我们使用协方差矩阵(Pearson 系数矩阵[23])计算资源状态时序数据子序列组中的子序列间的相关性,其矩阵表示形式为

$$\boldsymbol{M}_{4\times4}=\begin{pmatrix} p(Z_1,Z_1) & \cdots & p(Z_1,Z_4) \\ \vdots & & \vdots \\ p(Z_4,Z_1) & \cdots & p(Z_4,Z_4) \end{pmatrix}. \tag{1}$$

在式(1)中,$\boldsymbol{M}_{4\times4}$ 中的元素 $p(Z_i,Z_j)(i,j\in\{1,2,3,4\})$ 的值由式(2)估算.

$$p(Z_i,Z_j)=\frac{\sum_{l=1}^{L}(d(l)_i-\widehat{d_i})(d(l)_j-\widehat{d_j})}{L}. \tag{2}$$

式中:$d(l)_i$ 和 $d(l)_j$ 为子序列 Z_i 和 Z_j 在第 l 个时间点上的数据值;$\widehat{d_i}$ 和 $\widehat{d_j}$ 为子序列 Z_i 和 Z_j 在整个时间段内全部数据点的均值,其值由式(3)估算.

$$\begin{cases}\widehat{d_i}=\dfrac{\sum_{l=1}^{L}d(l)_i}{L};\\ \widehat{d_j}=\dfrac{\sum_{l=1}^{L}d(l)_j}{L}.\end{cases} \tag{3}$$

节点对自身采集数据的检测与更新过程描述:

算法1: 节点对采集数据的检测与更新过程
输入:当前采集的资源状态与当前维护的训练模式
输出:更新的自身资源状态信息

1:将当前维护的训练模式记作 $\boldsymbol{M}^{t}_{4\times4}$
2:根据式(1)~式(3)计算当前采集的资源状态子序列组的相关性矩阵并记作 $\boldsymbol{M}^{d}_{4\times4}$
3:若 $\boldsymbol{M}^{d}_{4\times4}$ 中元素与 $\boldsymbol{M}^{t}_{4\times4}$ 中元素的相关性关系一致,则使用所有维度的子序列更新自身资源状态信息并结束本次检测与更新过程;否则,继续第4步
4:若当前采集的资源状态子序列组中仅有第 i($i\in\{1,2,3,4\}$)维度子序列与其它维度子序列的相关性未通过训练模式的比对,则使用其他维度子序列更新自身资源状态信息并结束本次检测与更新过程;否则,结束本次检测与更新过程

在恶劣环境下工作的节点,其自身采集的数据并非一定可靠.因此,有必要采用算法1来提升采集数据的可靠性,达到排除采集到的数据中的噪声和异常值的目的.其中,若满足步骤3条件,则采集的数据可被全部采纳;若无法满足步骤3条件,则根据步骤4条件来及时获取某些资源状态数据仍是一种不错的选择.若无法满足步骤4条件,则本次采集的资源状态数据都将被认为不可靠而被舍弃.

节点对接收数据的检测与更新过程描述:

算法2: 节点对接收数据的检测与更新过程
输入:当前接收的资源状态与当前维护的训练模式
输出:更新的关于其他节点资源状态的缓存信息

1:节点维护2种训练模式,分别被记作正常模式 $\boldsymbol{M}^{t'}_{4\times4}$ 和异常模式 $\boldsymbol{M}'^{t'}_{4\times4}$
2:根据式(1)~式(3)计算当前采集的资源状态子序列组的相关性矩阵并记作 $\boldsymbol{M}^{d'}_{4\times4}$
3:若 $\boldsymbol{M}^{d'}_{4\times4}$ 中元素与 $\boldsymbol{M}^{t'}_{4\times4}$ 中元素的相关性关系一致,则使用所有维度子序列更新自身资源状态信息并结束本次检测与更新过程;否则,继续第4步
4:保存存在异常的资源状态子序列组并统计数量是否达到训练集规模;若达到训练集规模,则更新当前维护的异常模式 $\boldsymbol{M}'^{t'}_{4\times4}$ 并继续第5步;否则,继续第5步
5:若 $\boldsymbol{M}^{d'}_{4\times4}$ 中元素与 $\boldsymbol{M}'^{t'}_{4\times4}$ 中元素的相关性关系都不一致,则使用所有维度子序列更新自身资源状态信息并结束本次检测与更新过程;否则,结束本次检测与更新过程

当节点将自身最新更新的资源状态数据推送给订阅用户节点时,因恶劣的通信环境,极易造成资源状态数据发生异常.因此,有必要采用算法2来提升接收数据的可靠性.其中,若满足步骤3的条件,则接收的

数据可被全部采纳.若无法满足步骤 3 的条件,则可根据步骤 5 的条件来判断是否属于异常类别,以便尽可能避免误判.在此之前,可根据步骤 4 的条件决定是否先对异常模式进行更新.

3　仿真分析

仿真所用数据来自作者所在实验室的 1 台笔记本电脑和 9 台配置了 WiFi 通信模块的台式电脑.从笔记本电脑的任务管理器的“性能”菜单可以查看设备用户某一时刻使用该设备资源的详细状况,如图 2 所示.

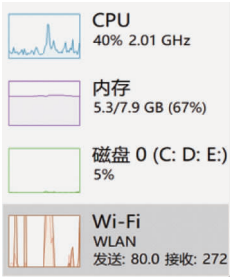

图 2　设备资源使用状况

另外,任务管理器的“用户”菜单可以查看设备用户某一时刻使用该设备资源的百分比,如图 3 所示.

图 3　设备资源使用率

为了对每台电脑做运行情况监测,获取其 CPU、内存、磁盘、网络使用状态数据,使用 Python 编写相关代码来记录连续 3 个月的设备资源状态数据作为数据集,其中每天的记录时间为 9:00—21:00,该程序每 10 s 记录一次数据.针对这些数据,我们将每 1 h 内记录的数据为一个基本处理单元,看作一个包含4 个维度的资源状态子序列组,其中,对能量状态维度在指定范围内做随机填充.所有资源状态子序列组按时间顺序排列,作为方案仿真时模拟设备采集数据和接收其他节点数据之用.

在仿真中,除了实现本文所提出方案(为了方便,称之为 Anomaly Detection Based on Double Correlation Comparison,并简称 ADDC)外,也实现了另外 2 种时间序列异常检测方法作为性能对比方案.一种是基于序列相关性分析的多维时间序列异常检测方法,被称为 CGAD 算法[13].该算法首先对多维时间序列进行分段与标准化计算以得到相关性矩阵,然后建立时序相关图模型以便于划分时间序列团,并进行时间序列团内、团间以及单维的异常检测.另一种是基于机器学习模型的异常检测算法,被称为 LCAD 算法[24].该算法首先对多维序列进行协方差矩阵计算,得到其相关性测量值,然后从矩阵中提出单维的特征向量作为序列的特征值,并用高斯分布模型训练样本集合,对于待检测数据,用最大期望算法得到数据正常或异常的概率分类.

将资源状态数据时间子序列组作为一个分析的基本单位,可将仿真结果分为正确的正例(True Positives, TP)、错误的正例(False Positives, FP)、错误的负例(False Negatives, FN)、正确的负例(True Negatives, TN),其具体含义分别是实际为正常且算法检测为正常的实例数、实际为异常但算法检测为正常的实例数、实际为正常但算法检测为异常的实例数、实际为异常且算法检测为异常的实例数.资源状态数据时间子序列组分析的准确率被定义为正确的正例数与正确的正例数和错误的正例数之和的比值,而召回率被定义为正确的正例数与正确的正例数和错误的负例数之和的比值.通过这 2 个指标,对 3 种方案进行性能的对比分析.

在仿真中,选取 360 个时间点上 4 列数据作为一个时间段,即将时间长度为 360 的数据记为一个数据组.对用于仿真的 3 个月内所采集资源状态数据(近 40 万个时间点上的采集数据),在训练集中最多使用了 2 个月所采集的数据(约 26 万个时间点上的采集数据),而其余数据被用作测试集.异常模式较为均匀

地出现在 4 列数据上,将一条时间序列上的一个长度不少于 18 点的异常模式记为一个异常实例,而任意异常实例的长度值不少于 18 且不大于 180.异常实例较为分散地存在于待检测数据中,且异常实例总数最多为 100 个.分别仿真异常实例总数、测试集规模和训练集规模对上述 3 种算法检测性能的影响,以及本文方案与 CGAD 算法的效率与开销对比.

1)异常实例数量变化的影响:图 4 展示了用 100 组数据做训练集,通过改变异常实例总数的变化仿真对 3 种算法性能的影响.随着数据中异常数据总数的增加,LCAD 算法的准确率和召回率有明显的下降,而 CGAD 算法和本文算法的准确率和召回率稍有下滑,但准确率和召回率都保持在 80%以上,且召回率更为稳定.另外,本文算法的准确率略高于 CGAD 算法,这说明本文设计的正常和异常模式双检测能够避免正常模式的误判.

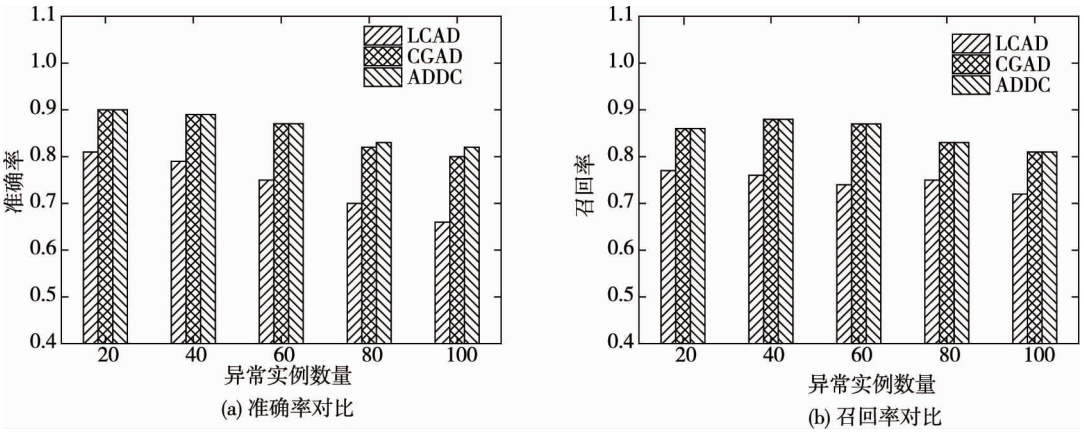

图 4 异常实例数变化对算法性能的影响

2)测试集规模差异的影响:图 5 展示了用 100 组数据做训练集,通过改变测试集规模的大小仿真对 3 种算法性能的影响.

从图 5 看出,相比其他 2 种方案,测试集规模差异对 LCAD 算法的性能影响更大一些不大,但准确率和召回率在 30 组之前变化却不明显,而在 30 组之后均有较明显下降.这说明测试集数量增大到一定程度会降低 LCAD 算法性能.在选定的测试集规模变化范围内,CGAD 算法和本文算法的准确率基本是稳定的,而召回率却略有提高.这证实了 CGAD 算法和本文算法对序列间相关性建模的有效性.由于本文方案采取正常模式与异常模式双检测的思路,因此,在准确率上的表现略好于 CGAD 算法.

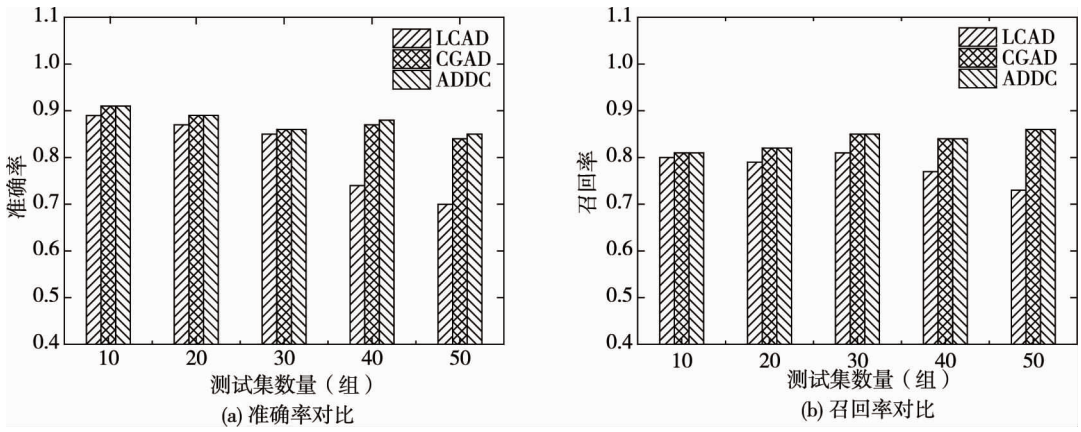

图 5 测试集规模差异对算法性能的影响

3)训练集规模差异的影响:在图 6 中介绍了训练集规模对本文方案识别效果的影响.基于同等测试集规模,分别仿真了规模分别为 50,100 和 300 个序列时间组的训练集情况下的训练过程对识别结果的影响.从图 6a 可以看出,随着测试集规模的增大,这 3 种训练情况的准确率都有微略降低且相同测试集规模

下,训练集数量多时的准确率更高.但总体来说,3 种情况下的准确率比较接近,且当训练集达到 100 及以上时,准确率几乎没有差别.图 6b 显示,这 3 种情况下的召回率都随测试集增多而呈现微略增长.但训练集规模变化对召回率影响更大一些.当训练集规模较小时,召回率较低,而当训练集规模较大时,召回率较高.由仿真结果可见,本文方案识别准确率总体上较为稳定,较小规模的训练集样本也可胜任异常检测任务.

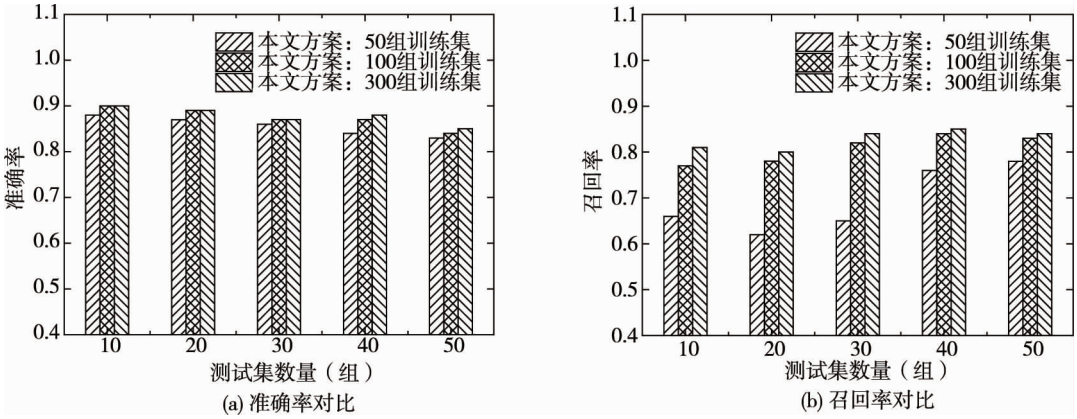

图 6　训练集规模对算法性能的影响

4)方法效率分析:图 7a 和图 7b 分别对比了异常模式总数量和测试集规模对本文方案与 CGAD 算法的异常检测任务所用时间的影响.从图 7a 和图 7b 得知,异常实例数量和测试集数量的差异没有给 2 种算法的训练时间开销带来明显的影响,但是本文方案的训练时间开销明显低于 CGAD 算法.这主要是本文方案不需要时序相关图模型构建等过程,因而节省了训练开销.另外,本文方案检测用时也比 CGAD 算法检测用时要小,且随着数据量的增加,这种优势有扩大趋势.这说明本文算法针对较低维度时序数据分析问题略去时序相关图模型构建的必要性.因此,针对较低维度的时序数据分析问题,本文方案的效率和效果的平衡性更好.

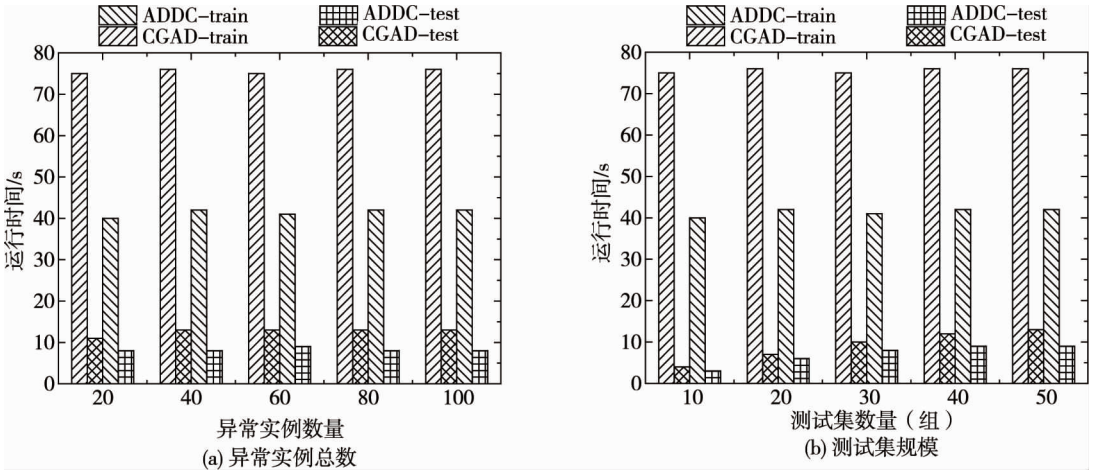

图 7　算法效率与开销对比

4　结论

1) ADDC 算法既能确保资源状态采集与交换过程中异常检测操作的及时性,也能提高检测结果的准确性.

2)ADDC 算法在检测准确性与运行时间上都优于典型的基于序列相关性分析的对比方案 CGAD 算法.

3)ADDC 算法能够满足弱链路环境下资源状态采集与交换过程中异常检测的性能需求,具有实际应用价值.

参考文献:

[1] Mcdonald A B, Znati T F. A mobility-based framework for adaptive clustering in wireless ad hoc networks [J]. IEEE Journal on Selected Areas in Communications, 1999, 17(8): 1466-1487.

[2] Hu Y C, Johnson D B, Perrig A. SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks [J]. Ad Hoc Networks, 2003, 1(1): 175-192.

[3] Hornos C D, Sur A D, Copiapo D H, et al. Wireless ad-hoc network [J]. World Telecommunications, 2003 (6): 1212-1218.

[4] 刘晓鹏,孙际哲,陈西宏.无线自组网在防空导弹网络化作战中的应用研究[J].飞航导弹,2012,4(2):31-34.

[5] 周吉超,顾文珊.无线自组网在野战防空通信系统中的应用[J].数字技术与应用,2016,4(9):41.

[6] 高昊江, 杨晟, 张宜生,等. 恶劣环境下无线网络数据通信策略研究与应用[J]. 计算机工程, 2007, 33(5): 82-83.

[7] Toledano M, Cohen I, Ben Y, et al. Real-time anomaly detection system for time series at scale [C]//In: Proc. of the SIGKDD Workshop, 2017: 56-65.

[8] Wang M, Zhang C, Yu J. Native API based windows anomaly intrusion detection method using SVM [C]//In: Proc. of the IEEE Int'l Conf. on Sensor Networks. IEEE, 2006.

[9] Gao B, Ma H Y, Yang Y H. HMMs (hidden Markov models) based on anomaly intrusion detection method [C]// In: Proc. of the 2002 Int'l Conf. on Machine Learning and Cybernetics, 2002: 381-385.

[10] Qiao Y, Xin X W, Bin Y, et al. Anomaly intrusion detection method based on HMM [J]. Electronics Letters, 2002, 38 (13): 663-664.

[11] Zhang X, Fan P, Zhu Z. A new anomaly detection method based on hierarchical HMM [C]// In: Proc. of the 4th Int'l Conf. on Parallel and Distributed Computing, Applications and Technologies, 2003: 249-252.

[12] Gupta M, Gao J, Aggarwal C, et al. Outlier detection for temporal data [J]. Morgan & Claypool Publishers, 2014, 26(9): 2250-2267.

[13] 丁小欧,于晟健,王沐贤,等. 基于相关性分析的工业时序数据异常检测[J]. 软件学报, 2020, 31 (3): 726-747.

[14] 杨海民,潘志松, 白玮. 时间序列预测方法综述[J]. 计算机科学, 2019, 46 (1): 28-35.

[15] Enders W.应用计量经济学:时间序列分析[M]. 杜江,袁景安,译. 2 版. 北京:高等教育出版社,2006.

[16] 包芬. 一种时序数据相关性分析方法的研究[D]. 西安: 长安大学,2019.

[17] 杨云丽. 大数据背景下的时序数据分析[D]. 合肥: 中国科学技术大学,2019.

[18] 赖永炫,张璐,杨帆,等. 基于时空相关属性模型的公交到站时间预测算法[J]. 软件学报,2020, 31(3): 648-662.

[19] Wang Z M, Zhang L, Ding Z M. Hybrid time-aligned and context attention for time series prediction [J]. Knowledge-Based Systems, 2020, 198: 105937.

[20] Shen Z P, Zhang Y M, Lu J W, et al. A novel time series forecasting model with deep learning [J]. Neurocomputing, 2020, 396: 302-313.

[21] Li L, Dai S D, Cao Z W, et al. Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction [J]. Journal of Supercomputing, 2020,76(9): 6887-6900.

[22] Liu Z D, Liu J. A robust time series prediction method based on empirical mode decomposition and high-order fuzzy cognitive maps [J]. Knowledge-Based Systems, 2020, 203: 106105.

[23] 中国科普·科学百科:皮尔逊相关系数[2020-08-04][EB/OL]. https://baike.baidu.com/item/皮尔逊相关系数/12712835? fr=aladdin.

[24] Ding J, Liu Y, Zhang L, et al. An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model [J]. Applied Intelligence, 2016, 44 (2): 340-361.