

异构复杂信息网络下的异常数据检测算法

穆丽文 彭贤博 黄 岚

(吉林大学计算机科学与技术学院 长春 130012)

摘 要 异构复杂信息网络承载着不同的协议和网络信道,并通过云储存实现资源调度,由此产生的异常数据会给网络信息空间带来安全威胁和存储开销,所以需要进行异常数据准确检测。传统的检测算法采用简化梯度算法进行异常数据检测,不能有效去除多个已知干扰频率成分的异常数据,检测性能不好。提出一种基于自适应陷波级联模型的异常数据检测算法。构建异构复杂信息网络系统模型,采用固有模态分解把异常数据信号解析模型分解为多个窄带信号,设计二阶格型陷波器结构,用多个固定陷波器级联抑制干扰成份,采用匹配投影法寻求优化特征解,找出所有匹配的特征点对,从而实现异常数据检测的改进。仿真实验表明,采用该算法进行异常数据检测时,信号幅值大于干扰噪声数据幅值;该算法提高了检测性能,具有较好的抗干扰性能。

关键词 网络,异常数据,检测,自适应陷波器

中图分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.028

Abnormal Data Detection Algorithm in Heterogeneous Complex Information Network

MU Li-wen PENG Xian-bo HUANG Lan

(College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract Heterogeneous information network carries different protocols and network channels, and realizes the resource scheduling through the cloud storage, resulting in abnormal data which can cause security threat and storage overhead to the network information space, so accurate detection of abnormal data is essential. The traditional detection algorithms use the simplified gradient algorithm for outlier detection, which cannot effectively remove multiple abnormal data of known interference frequency components, so the detection performance is not good. An abnormal data detection algorithm was proposed based on adaptive notch cascade model. The heterogeneous information network system model is constructed, and by using the intrinsic mode decomposition, the abnormal data signal analytical model is decomposed into multiple narrowband signals. Two order lattice notch filter structure is designed, and a plurality of fixing notch filters cascade is used to suppress the interference components. Matching projection method is used to seek the optimized feature solution, and all matching feature point pairs are find out, realizing the improvement of outlier detection. Simulation results show that, when the algorithm is used in data detection, signal amplitude is larger than the amplitude of noise interference data. It can improve the detection performance, and anti-interference performance is better.

Keywords Network, Abnormal data, Detection, Adaptive notch filter

1 引言

异构复杂信息网络是一种由不同制造商生产的计算机、网络设备和系统组成的复杂网络结构,通常情况下,异构复杂信息网络构架在不同的协议上支持不同的功能或应用。异构复杂信息网络在整个互联网体系中体现为一个异构网元,承载着不同的协议和网络信道,并通过云储存实现资源调度,由此产生的异常数据会给网络信息空间带来安全威胁和存储开销,因此需要进行异常数据准确检测。随着信息技术的不断发展和海量的大数据在网络空间的不断增长,研究异构复杂

信息网络中的异常数据检测算法对确保网络安全、净化网络空间具有重要意义,相关算法研究受到了广泛重视。

在异构复杂信息网络中,异常数据受多个级联的干扰成分的影响,自相关特性较弱,特征提取困难,传统的时频特征分析算法和傅里叶频谱检测算法难以对之实现有效的信号检测^[1]。对此,相关学者进行了算法的改进,其中,文献[2]引入信道分配和信息排队论分析机制,构建数据信息链优先链路选择准则,实现对复杂信息网络的异常数据检测,取得了一定的检测效益,但在吞吐量、延时、CPU使用率等指标上性能不佳;文献[3]采用了一种基于博弈论的异构复杂信息网络信道

到稿日期:2014-08-01 返修日期:2014-10-05 本文受国家自然科学基金项目(61472159),吉林省创新团队项目(20122805),吉林省科技发展项目(20140101180JC)资助。

穆丽文(1992—),女,硕士生,主要研究方向为计算智能,E-mail:mlw_mu@163.com;彭贤博(1990—),男,硕士生,主要研究方向为计算机网络,E-mail:peng964286866@163.com;黄 岚(1974—),女,博士,教授,主要研究方向为商务智能理论与应用,E-mail:huanglan@jlu.edu.cn(通信作者)。

分配;文献[4]提出一种基于时间-频率联合分布特征和偏移量递阶控制的网络异常数据监测与参量估计算法,通过递阶控制调整 HHT 频谱偏移,实现对异构网络的复杂信息滤除和异常数据检测,在减小包络线失真引起的边界控制误差上取得了一定的效果,但是该算法存在计算量大、实现困难等问题;文献[5]提出一种基于流数据分类和分形维数分析信号检测算法的异常数据特征提取算法,来提高数据聚类能力,实现数据滤波和数据的异常行为检测,但该算法具有实时性和收敛性不高的缺陷;文献[6]采用基于特征选择和 SVM 参数同步优化的异常数据检测算法,即通过定量递归分析的方法分析网络异常数据的聚类特征,对异常数据的流量进行准确估计和预测,实现数据过滤和监测的目的,但该算法没能有效挖掘流量序列的非线性特征信息,检测概率不高,反而提高了虚警概率,导致对正常网络传输数据的误删;另外,文献[7]采用简化梯度算法进行异常数据检测,结合匹配滤波器的混合无线组网检测,实现异常数据的准确识别,但该算法不能有效应用于异构复杂信息网络中,不能有效去除多个已知干扰频率成分的异常数据,需要进行算法改进设计^[8-10]。

针对上述问题,本文提出一种基于自适应陷波级联算法的异常数据检测算法,首先构建异构复杂信息网络系统模型,并进行信息流模型的构建和异常数据特征的提取,然后设计自适应级联陷波算法进行噪声干扰抑制,最后进行检测算法设计。通过算法的模型构建和仿真测试,得出在实现对异构复杂信息网络的异常数据检测中,本文算法性能优越。

2 异构复杂信息网络系统模型及异常数据特征提取

2.1 异构复杂信息网络信道数据传输模型

为了实现对异构复杂信息网络的异常数据检测,首先构建异构复杂信息网络数据传输模型。异构复杂信息网络中承载着不同的协议和网络信道,并通过云储存实现资源调度。异构复杂信息网络的网络信道模型有轮流协议、随机访问协议和信道划分协议 3 种,信道划分协议主要是 FDMA、CDMA、TDMA 等,它是一种无竞争的协议。网络模型结果分为资源类(resource classes)、资源实例(resource instances)、属性类(attribute classes)、属性实例(attribute instances)等多种类型。在设计异构复杂信息网络的信道数据协议时,采用 SeerSim 分析的算法构建网络数据传输,在各个信道 NAV 任务的有限排队模型的状态空间为:

$$S = \{(k, n), 0 \leq k \leq K, 0 \leq n \leq N\} \quad (1)$$

网络通信过程中,通过云储存实现资源调度,大量数据在不断传输调度过程中产生异常数据,此时局部性交叉项信息链表示为:

$$CUB(i) = \frac{CBT(i)}{T} \times CB \quad (2)$$

其中, CB 表示系统中的数据包的数量, i 表示数据包处理阶段的数量。对于一个通用的接收节点,当接收的功率 $n = N$ 时, CPU 进行数据包处理;当 $n = 1, \dots, N-1$ 时,计算得到信息链状态概率之后,假设 k 节点的通信范围内有 n 个邻居节点,网络的系统平均吞吐量 γ 可以表示为:

$$\gamma = \mu \sum_{k=1}^K p_{k,N} \quad (3)$$

所有信道中异常数据所占的比例为 $CUBP$, 其中路径损

耗因子为:

$$\Delta r_{ij}^k(t) = Q/Z_k \quad (4)$$

式中, Z_k 用于描述第 k 个数据在当前拓扑结构中所需的路线长度。通过上述分析过程实现对异构复杂信息网络信道数据传输模型的构建,如图 1 所示,在此模型中进行数据传输和调度,分析异常数据检测和特征方法。

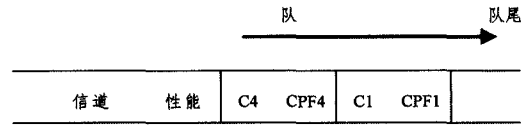


图 1 信道切换队列数据结构

2.2 异常数据的信息流模型构建

在上述构建的异构复杂信息网络信道数据传输模型中,进行异常数据检测算法设计时,需要首先构建异常数据信号模型,引入现代信号检测技术,实现信号检测。在多个已知干扰中,为了使异常数据信号离散数据解析化,构建异构复杂信息网络的异常数据信号解析模型:

$$z(t) = x(t) + jy(t) = a(t)e^{j\theta(t)} \quad (5)$$

式中, $z(t)$ 表示异常数据信号, $x(t)$ 表示异常数据信号解析模型的实部, $y(t)$ 表示异常数据信号的固有模态函数,采用固有模态分解,把异常数据信号解析模型分解为多个窄带信号,得到数据的包络特征为:

$$a(t) = \sqrt{x^2(t) + y^2(t)}, \theta(t) = \arctan \frac{y(t)}{x(t)} \quad (6)$$

其中, $\theta(t)$ 表示高频分量, $a(t)$ 和 $\theta(t)$ 分别是在异构复杂信息网络中的干扰特征幅度和相位信息。异构网络的异常数据干扰频率特征表达式为:

$$f(t) = \frac{1}{2\pi} \times \frac{d\theta(t)}{dt} \quad (7)$$

上式描述了多源进程节点中的异构网络的异常数据生成校验位,用来表征数据干扰频率,其中异常数据出现的概率权重为:

$$w_{ij} = \beta \times w(e_p k_q), \beta > 1 \quad (8)$$

综上分析,可以计算每个信道的信号干扰比 SIR , 得到数据信息采集目标端信息分量为:

$$r_1 = x(t) - c_1 \quad (9)$$

通过上述处理,构建了异常数据的信息流模型,为进行异常数据的有效检测提供信号基础。

3 基于自适应陷波级联模型的异常数据检测算法

3.1 自适应级联陷波算法设计

在异构复杂信息网络中,通过云储存实现资源调度会产生异常数据,给网络信息空间带来安全威胁和存储开销,因此需要进行异常数据检测。根据上述分析,结合式(7)可见,在异构复杂信息网络中的异常数据受到多个干扰频率特征的作用,采用传统的简化梯度算法进行异常数据检测时,不能有效去除多个已知干扰频率成分的异常数据。为了弥补这一缺陷,需要进行算法改进设计,在构建上述异常数据信号模型的基础上,设计二阶格型陷波器结构,进行异常数据的干扰滤波,二阶格型陷波器结构如图 2 所示。图 2 中 $x(k)$ 为观测序列,它是由信号和背景噪声共同构成的, $y(k)$ 为陷波器的输

出,当陷波器的陷波频率点落在 $x(k)$ 中所含正弦信号的频率处时, $x_1(n)$ 是陷波器的一个状态参数。

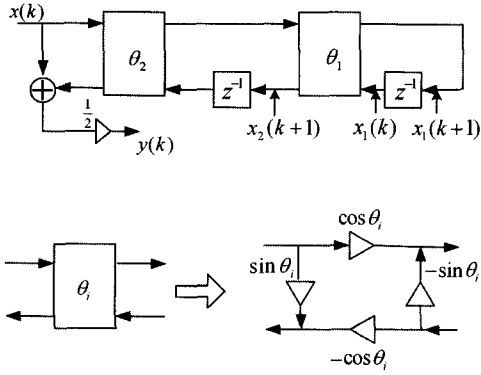


图2 二阶格型陷波器结构

传统梯度算法的滤波器的迭代公式描述为:

$$\theta_1(k+1) = \theta_1(k) - \mu \text{Re}[y(k)\varphi^*(k)] \quad (10)$$

其中, μ 是可以控制收敛速度和精度的参数,称为步长; $\varphi(k)$ 是格型陷波器进行异常数据检测的输出 $y(k)$ 对参数 $\theta_1(k)$ 的差分,由此得到陷波器传输函数为:

$$H_B(z) = \frac{(1 + \sin\theta_2)}{\cos\theta_2} \frac{\cos\theta_1(k) \cos\theta_2 z^{-1}}{1 + \sin\theta_1(k)(1 + \sin\theta_2)z^{-1} + \sin\theta_2 z^{-2}} \quad (11)$$

其中,

$$G(z) = \frac{1 - \sin\theta_2}{2} \frac{1 - z^{-2}}{1 + \sin\theta_1(k)(1 + \sin\theta_2)z^{-1} + \sin\theta_2 z^{-2}} \quad (12)$$

根据滤波器传输函数的幅度和相位响应特性来分析,输入信号是复正弦信号和复噪声的和,可见采用简化梯度算法进行异常数据检测不能有效去除多个已知干扰频率成分的异常数据。陷波器可以抑制干扰中的线谱成份,对于多个已知干扰线谱中的信号检测,为了实现异常数据的准确检测,对于抑制干扰中的线谱成份,本文采用自适应级联陷波设计,用多个固定陷波器级联抑制干扰成份,然后检测信号,在多个已知干扰中检测异常数据的自适应级联陷波设计结构如图3所示。图3中,陷波器级联框图中每一级陷波器都是一个独立的个体,每一级异常数据的梯度信号的产生都只与该级滤波器的参数有关,以此提高对异构复杂网络异常数据的检测性能。

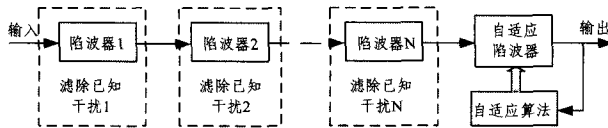


图3 自适应级联陷波设计结构

3.2 检测算法实现

通过上述设计的自适应级联陷波器,对异常数据进行特征提取和滤波处理,去除多个已知干扰频率成分,实现对异常数据的准确检测,假设自适应级联陷波滤波器的输出 $y(k)$ 和梯度信号 $\varphi(k)$ 可以表示为:

$$y(k) = s_1(k) + n_1(k); \varphi(k) = s_2(k) + n_2(k) \quad (13)$$

$$s_1(k) = AA_{H_B} e^{i(\Omega k + \theta_{H_B})}; s_2(k) = AA_{H_B} e^{i(\Omega k + \theta_{H_B})} \quad (14)$$

其中, A_H 、 A_{H_B} 和 θ_H 、 θ_{H_B} 分别是滤波器传输函数 $H(z)$ 和 $H_B(z)$ 的幅度响应和相位响应,用输入 $x(k)$ 减去输出 $y(k)$ 后

就可以得到所期望的异常数据高频分量信号,异构复杂信息网络采集的数据是可分类的。引入了一个管理因子 $\beta \in (0, 0.5)$,在这个管理因子的限定范围内,输入信号中含有 N 个复正弦异常数据信号,选择一定的基函数与异常数据信号进行匹配,信号 f 和基 d_{γ_0} 之间的匹配程度为:

$$\lambda^n(d_{\gamma_0}) = \int_{-\infty}^{+\infty} f(t)d_{\gamma_0}^*(t)dt \quad (15)$$

为了提高对异常数据的检测概率,采用匹配投影法寻求优化特征解,满足:

$$|\langle f, d_{\gamma_0} \rangle| \geq a \sup_{\gamma \in \Gamma} |\langle f, d_{\gamma} \rangle| \quad (16)$$

则可以由若干个向量基函数的线性组合来表示,存在映射 $\Phi: X \rightarrow Y$,满足:

$$\Phi(x \cdot y) = \Phi(x) \cdot \Phi(y) \quad (17)$$

其中, \cdot 是 X 上的运算, \circ 是 Y 上的运算。采集数据集 S , 当 $X \subseteq U, R \subseteq A$ 条件成立时,数据存在异常性的边界条件为:

$$R_{\beta} X = U \{E \in U / R | c(E, X) \leq \beta\} \quad (18)$$

$$R_{\beta} X = U \{E \in U / R | c(E, X) \leq 1 - \beta\} \quad (19)$$

对于任意两个数据块 m_i 和 m_j ,得到基本块文件 $m_{i,j}$ ($1 \leq i \leq n, 1 \leq j \leq k$),异常数据信息预测性能为 $\{\lambda_i, 1 \leq i \leq S\}$,判别准则 $\{R_j, 1 \leq j \leq L\}$,信道分配数据包转换的等待时间为:

$$W_q = W - \bar{X} = \frac{1}{\gamma} \sum_{k=1}^K \sum_{n=1}^N k p_{k,n} - \frac{(N-1)\mu + r}{\mu r} \quad (20)$$

由此找出所有匹配的特征点对,采用异常数据的密度先验信息,通过自适应级联陷波去除数据干扰,实现对异常数据的高效检测。

4 仿真实验与结果分析

为了测试和验证提出的基于自适应陷波级联模型的异常数据检测算法在检测异构复杂信息网络下的异常数据的性能,基于 Matlab 仿真实验平台进行仿真实验。仿真环境为: Intel Core3-530 1GB 内存,操作系统为 Windows 7, VS2008 处理器, X6300 CPU, 2 GDDR 内存。在异构复杂信息网络的构建中,选取了 50000 个存储节点, 200000 个数据对象,节点的存储性能在 1~5 中随机选取,异构网络的管理节点负责系统缓存管理,运算节点负责数据解算、上传等节点端处理。在上述仿真环境中,进行异构复杂信息网络数据通信和传输,数据传输过程中,通过云储存实现资源调度,由此产生的异常数据会给网络信息空间带来安全威胁和存储开销,因此需要进行异常数据检测仿真。网络信息传输处于复杂的含有干扰频率成分的噪声环境中,噪声频谱 $f = 20\text{kHz}$,方位参数 θ 在 $(0, 2\pi]$ 之间均匀分布,异常数据的采样频率选择和网络环境噪声采集的频率相同,为 196.608kHz,信息流的采样频率为 500kHz,采样数据为 TXT 数据,得到的异构复杂信息网络下的两个通道的数据传输信息流信号模型如图4所示。

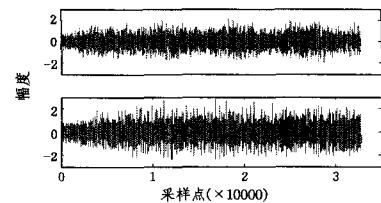


图4 异构复杂信息网络下数据传输信息流信号模型

以上述数据样本为研究对象,将其作为本文设计的自适应级联陷波器的输入向量,对信息流进行特征检测和数据滤波,来实现数据检测。检测实验中,以异常数据的频谱幅度为测试指标,得到了采用本文算法和传统的梯度算法实现数据检测的异常数据频谱幅度结果,如图5所示。

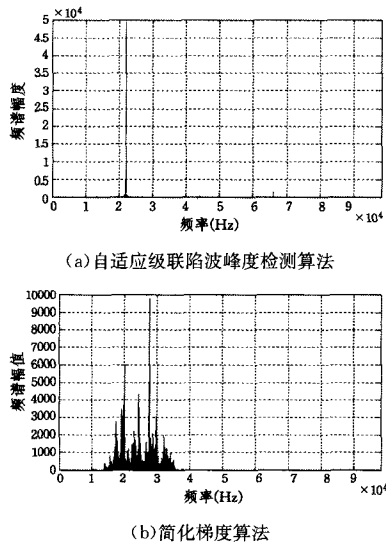


图5 异常数据频谱幅度检测结果对比

从图5可知,采用传统算法进行异常数据检测时,受干扰成分影响较大,对异常数据的检测频谱幅度峰值不明显,检测性能不好。而本文算法用多个固定陷波器级联抑制干扰成份,然后检测信号,可以很好地抑制色噪声的影响,信号幅值大于干扰数据噪声幅值,提高了检测性能。为了定量分析本文算法在检测异常数据时的性能,采用20000次蒙特卡洛实验,在不同信噪比下分析不同算法的检测概率,结果如图6所示。

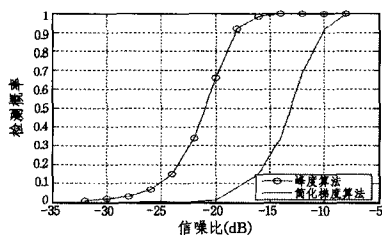


图6 检测性能曲线

分析图6可知,采用本文算法,在信噪比为-16dB时,检测概率为100%,而传统算法的检测概率只有18.7%。可见,采用本文算法进行异常数据检测时检测性能提高较大,抗干扰性能强。

结束语 随着信息技术的不断发展和海量的大数据在网络空间的不断增长,研究异构复杂信息网络中的异常数据检测算法对确保网络安全、提高网络通信质量具有重要意义。本文研究了强干扰低信噪比下的异构复杂网络信息异常数据检测算法,提出了一种基于自适应陷波级联模型的异常数据检测算法。设计自适应级联陷波器,对异常数据检测环境进行干扰噪声抑制,为了提高对异常数据的检测概率,采用匹配投影法寻求次优解,计算所有匹配的特征点对,实现检测算法

的改进。通过理论研究和仿真实验得出,采用本文算法能有效提高异常数据的检测概率,对噪声的抑制性能较好,在网络数据监测等领域具有较好的应用价值。

参考文献

- [1] Zhu Q Y, Yang X F, Yang L X, et al. Optimal control of computer virus under a delayed model [J]. Applied Mathematics and Computation, 2012, 218(23): 11613-11619
- [2] Miorandi D, Sicari S, Pellegrini F D, et al. Internet of things: vision, applications and research challenges [J]. Ad Hoc Networks, 2012, 10(7): 1497-1516
- [3] Chen L, Brian K, Jamie E. Theoretical Characterization of Non-linear Clipping Effects in IM/DD Optical OFDM Systems [J]. IEEE Transactions on Communications, 2012, 60(8): 2304-2312
- [4] 饶雨泰, 杨凡. 网络入侵搅动下的网络失稳控制方法研究 [J]. 科技通报, 2014, 30(1): 185-188
- [5] Rao Yu-tai, Yang fan. Network Intrusion Stir the Network Instability Control Methods of the Research [J]. Bulletin of Science and Technology, 2014, 30(1): 185-188
- [6] 王光浩, 吴越. 一种车载自组织网络路况信息的数据信任模型 [J]. 计算机科学, 2014, 41(6): 89-93
- [7] Wang Guang-hao, Wu Yue. Data Trust Model for Road Information in Vehicular Ad hoc Network [J]. Computer Science, 2014, 41(6): 89-93
- [8] 王瑞, 马艳. 基于分数阶傅里叶变换的线性调频脉冲信号波达方向估计 [J]. 兵工学报, 2014, 35(3): 421-427
- [9] Wang Rui, Ma Yan. DOA Estimation of Wideband Linear Frequency Modulated Pulse Signals Based on Fractional Fourier Transform [J]. Acta Armamentarii, 2014, 35(3): 421-427
- [10] 罗亮, 吴文峻, 张飞. 面向云计算数据中心的能耗建模方法 [J]. 软件学报, 2014, 25(7): 1371-1387
- [11] Luo Liang, Wu Wen-jun, Zhang Fei. Energy Modeling Based on Cloud Data Center [J]. Journal of Software, 2014, 25(7): 1371-1387
- [12] 陈昊, 杨俊安, 庄镇泉. 变精度粗糙集的属性核和最小属性约简算法 [J]. 计算机学报, 2012, 35(5): 1011-1017
- [13] Chen Hao, Yang Jun-an, Zhuang Zhen-quan. The Core of Attributes and Minimal Attributes reduction in Variable Precision Rough Set [J]. Chinese Journal of Computers, 2012, 35(5): 1011-1017
- [14] Zhou Yan, Li Jian-xun, Wang Dong-li. Target tracking in wireless sensor networks using adaptive measurement quantization [J]. Science China Information Sciences, 2012, 55(4): 827-838
- [15] Xu Jian, Li Jian-xun, Xu Sheng. Data fusion for target tracking in wireless sensor networks using quantized innovations and Kalman filtering [J]. Science China Information Sciences, 2012, 55(3): 530-544
- [16] 黄志宏, 巫莉莉, 张波. 基于云计算的网络安全威胁及防范 [J]. 重庆理工大学学报(自然科学版), 2012, 26(8): 85-90
- [17] Huang Zhi-hong, Wu Li-li, Zhang Bo. Network Security Threats and Prevention on Cloud Computing [J]. Journal of Chongqing University of Technology (Natural Science), 2012, 26(8): 85-90