

微服务架构下烟草信息化网络异常数据 层次密度聚类检测算法

叶玉琼^{1†}, 刘杰¹, 张文府¹, 杨虹¹, 钱怡君¹, 夏斌¹, 袁齐风²

(1. 红塔集团楚雄卷烟厂, 云南 楚雄 675000; 2. 北京金蝶云基科技有限公司, 北京 101300)

摘要:当前对于烟草信息化网络异常数据检测主要依托于物联网技术, 由于缺少对异常数据多重特征属性的分类, 使得检测精度较低。为此, 提出了微服务架构下烟草信息化网络异常数据层次密度聚类检测算法。通过计算异常数据样本到聚类中心的局部可达密度确定类簇中心, 并根据聚类决策图对异常数据进行层次密度聚类识别, 结合小波包函数对异常数据进行分解, 根据属性特征将其归类到不同类簇下, 并基于网络传输节点分布模型构建异常数据传输信道模型, 以此为依据, 求取异常数据的熵值, 并将计算结果与固定阈值比较, 将阈值外的数据确定为异常数据, 进而实现异常数据检测。对比实验结果表明, 所提方法对于烟草信息化网络异常数据具有较高的检测精度。

关键词:微服务架构; 烟草信息化网络; 异常数据; 层次密度聚类检测

中图分类号: TP183

文献标识码: A

Hierarchical Density Clustering Detection Algorithm for Abnormal Data in Tobacco Informatization Network under Microservice Architecture

YE Yuqiong^{1†}, LIU Jie¹, ZHANG Wenfu¹, YANG Hong¹, QIAN Yijun¹, XIA Bin¹, YUAN Qifeng²

(1. Chuxiong Cigarette Factory of Hongta Tobacco Group, Chuxiong, Yunnan 675099, China;

2. Beijing Kingdee Yunji Technology Co., Ltd., Beijing 101311, China)

Abstract: Currently, the detection of abnormal data in tobacco information networks mainly relies on internet of things technology. Due to the lack of classification of multiple feature attributes of abnormal data, the detection accuracy is low. To this end, a hierarchical density clustering detection algorithm for abnormal data in tobacco information network under microservice architecture is proposed. By calculating the local reachable density of abnormal data samples to the cluster center, the cluster center is determined, and hierarchical density clustering recognition is performed on the abnormal data based on the clustering decision diagram. The abnormal data is decomposed using wavelet packet function, classified into different clusters based on attribute characteristics, and an abnormal data transmission channel model is constructed based on the network transmission node distribution model. Based on this, the entropy value of the abnormal data is calculated, and compare the calculation results with a fixed threshold, determine the data outside the threshold as abnormal data, and then achieve abnormal data detection. The test results show that the designed method can detect abnormal network operation data with high accuracy, and the detection effect is better.

Key words: microservice architecture; tobacco information network; abnormal data; hierarchical density clustering detection

在现代互联网运营环境下,为了加快烟草生产经营管理一体化进程,实现高质量发展,建设烟草信息网络平台是构建现代烟草经济体系的主要途径之一。数据信息泄露、通信节点攻击、安全管理不规范等严重威胁着烟草信息网络系统的安全运行。因此,应采取合理的技术措施来检测网络异常数据。

文献[1]采用改进 K-means 算法对原始数据集进行划分聚类,根据距离最近原则把数据对象分配给最近的聚类中心,并借助主动迭代方法完成异常数据识别。但该方法对于事件边缘的数据检测具有应用局限性,导致检测精度不高;文献[2]采用遗传算法,通过建立检测模型实现异常数据检测。但此方法存在参数设定困难、计算复杂等缺陷,检测效果不佳。

针对以上分析,本文结合前人研究成果,对微服务架构下烟草信息化网络异常数据层次密度聚类检测算法展开深入研究,以期保证烟草数据的安全性。

1 异常数据层次密度聚类检测算法设计

1.1 异常数据层次密度聚类识别

在检测烟草信息网络中的异常数据时,首先需要对数据信息进行有效的脱敏处理,从而准确识别异常数据^[3]。本文利用聚类算法的树结构来反映数据之间的层次关系,并根据异常数据与正常数据之间特征属性的差异来识别异常数据。

通常情况下,异常数据对烟草信息化系统的影响程度取决于其对应的权值^[4]。因此,假设异常数据的尺度因子为 $C_j, j = 1, 2, \dots, k$, 聚类因子为 D_j , 将 N 个异常数据样本归类到 M 个聚类中心,则由异常数据构成的聚类中心可表示为:

$$W_j = \sum_{j=1}^k C_j + \sum \frac{D_j}{N \times M} \times \frac{u_j}{c_0} \quad (1)$$

式中, u_j 表示给定参数; c_0 表示随机矢量。

假设在烟草信息化系统中有 L 个通信节点,那么节点的汇总信息可表示为:

$$Y_0 = W_j \sum_{l=1}^L f_l(x) \times X_0, f_l \in F_0 \quad (2)$$

式中, f_l 表示第 l 个数据信息; $f_l(x)$ 表示数据信息的粗糙度函数; X_0 表示某一数据样本; F_0 表示信息化网络内部的整个空间。

计算异常数据样本聚类中心的局部密度,即:

$$\rho_i = \sqrt{\frac{Y_0}{d_i}} - d_c \cdot \alpha_0 \quad (3)$$

式中, d_i 表示两个相邻样本之间的距离; d_c 表示样本截断距离; α_0 表示算法参数。

采用高斯核定义数据样本之间的距离因子,即:

$$\delta_i = \sum \exp\left(-\frac{Y_0}{g_0}\right) \quad (4)$$

式中, g_0 表示聚类中心的可调参数。

以 ρ_i 和 δ_i 这两个属性为横、纵指标作出决策图,从而确定类簇中心^[5],如图 1 所示。

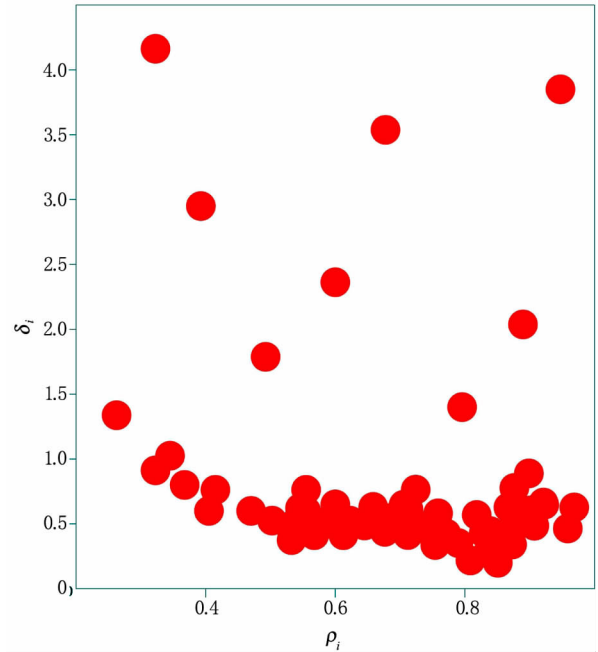


图 1 聚类决策图

为避免人工选取的类簇中心具有主观性,影响后续判断,因此,采用密度指标确定类簇中心^[6],计算公式如下:

$$\gamma_i = \rho_i \times \delta_i \times v_i \quad (5)$$

式中, v_i 表示密度指标。

则类簇中心中异常数据集的每个异常特征对应的密度矢量为:

$$A_i = \sum \exp\left[-\frac{|x_i - x_j|^2}{(\gamma_i + r_j)/2}\right] \quad (6)$$

式中, x_i 、 x_j 分别表示区间内密度的最大值与最小值; r_i 表示衰减常数。

经过局部聚类后,样本 j 对所属该类簇的贡献度可表示为:

$$R_j = \frac{k_c}{|A_i \times h_i|} \quad (7)$$

式中, k_c 表示类簇样本点数量; h_i 表示惩罚因子。

则异常样本数据集所属类簇的确信度为:

$$R' = \sqrt{\frac{R_j}{\varphi_0}} \times \frac{e_0}{d_1} \quad (8)$$

式中, φ_0 表示类簇边界样本数量; e_0 表示偏置向量; d_1 表示密度峰值。

通过样本确信度的变化控制异常数据的波动距离^[7], 则具体的异常数据波动值为:

$$s_j = \frac{1}{u_0} + 3\pi \times \frac{z_0}{R} \quad (9)$$

式中, u_0 表示波动偏差; z_0 表示异常数据识别阈值; π 表示测量距离。

通过上述计算, 比照波动数据的变化情况, 将异常波动值大于波动阈值 ϵ_0 的数据样本识别为异常数据, 为后续异常数据传输信道模型的建立奠定基础。

1.2 异常数据传输模型的构建

基于异常数据识别结果, 在烟草信息化跨平台网络环境下^[8], 首先给出网络传输节点的分布模型, 如图 2 所示。

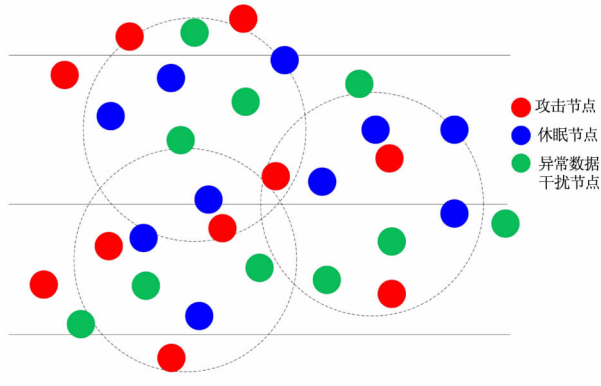


图 2 网络传输节点的分布模型

微服务架构下的烟草信息网络是一个三维连续系统, 参数通道属于时变非平稳通道^[9], 因此, 采用 RSA 算法对网络进行微服务摘要加密, 以保证数据传输的安全性, 计算公式如下:

$$G_t = E_n \times B_t \quad (10)$$

式中, E_n 表示原文数据; B_t 表示私钥。为了防止私钥被攻击暴露, 对微服务网络进行双重加密:

$$D_i = G_t \sqrt{\frac{p_i}{n_x}} \quad (11)$$

式中, p_i 表示原文数据生成的数字签名; n_x 表

示消息摘要。

基于加密后的原文数据, 采用小波包函数对网络信道中的数据进行分解处理^[10], 公式如下:

$$T_t = D_i \sum s_j \varphi_i(2t_0 - l_0) \quad (12)$$

式中, s_j 表示异常数据波动值; t_0 表示监测时间; l_0 表示分解层数; T_t 表示数据分解维度。

结合滑动时窗矩阵, 计算数据分解均值, 即:

$$\bar{T} = \frac{T_t}{E_r \times E_a} \quad (13)$$

式中, E_r 表示小波包系数; E_a 表示小波基函数。

当网络数据传输节点的传输路径存在 χ 条时, 则可认为此时在参数节点分布模型中存在 χ 条通信信道^[11], 再结合数据分解均值对信道中正在传输的数据进行描述, 即:

$$W(t) = W_x \times \bar{T} + \frac{\chi}{\tau(t)} \quad (14)$$

式中, W_x 表示信道数据分布矩阵; $\tau(t)$ 表示双曲正切函数。

异常数据在整个传输过程中的时间序列误差集合可以用下式描述:

$$e_y = \sum_{i=0}^{B_1} f(b) \times W(t) \quad (15)$$

式中, B_1 表示数据时间序列的分段数目; $f(b)$ 表示偏差度函数。

因此, 关于检测目标的函数可以表示为:

$$\begin{cases} H_0 = \sum e_y + \Omega(g_u) \times k_z \\ \Omega(g_u) = h_a + \frac{1}{2}l_a + \delta_0 r_c \\ k_z = t_b + \eta(t) \times \theta_0 \end{cases} \quad (16)$$

式中, H_0 表示检测目标函数; $\Omega(g_u)$ 表示目标函数中不适定因子; k_z 表示输出数据信息的节点序号; h_a 表示节点序号相关值; l_a 表示待测数据; δ_0 表示预测数据; r_c 表示常数; t_b 表示惩罚系数; $\eta(t)$ 表示最优分类函数; θ_0 表示调整系数。

假设异常数据信号的幅值为 A_u , 则传输信道的时域特征可表示为:

$$S(t) = \int_{-\infty}^{+\infty} H_0 \times Z^*(v_0 - t/2) dt \quad (17)$$

式中, Z^* 表示时频特征对; v_0 表示数据传输效率; t 表示异常数据在时频域中的总能量。

由此构建网络异常数据传输信道模型的表达式为:

$$U = a_n(t) \times S(t) [t_a - \vartheta(1 - q_0)] \tag{18}$$

式中， $a_n(t)$ 表示异常数据单分量； t_a 表示传输时延； ϑ 表示单分量传递的数据量； q_0 表示网络带宽。

以识别出的异常数据与网络体现节点分布模型为依据，对微服务系统进行加密处理，结合检测目标函数对网络传输信道的时间序列进行分析，进而构建异常数据传输信道模型，便于后期异常数据检测的实现提供帮助。

1.3 网络异常数据检测

在完成异常数据传输信道模型后，根据簇群异常数据检测目标，计算簇群检测目标的加权平均关联值^[12]，即：

$$P_v = \sqrt{3V_a} - \frac{U \times \sqrt{x_m + 2n_a}}{\zeta_t} \tag{19}$$

式中， V_a 表示综合簇群范围； U 表示异常数据传输信道模型； x_m 表示聚类特征系数； n_a 表示比例系数； ζ_t 表示经验常数。

设定数据信息的相关函数：

$$O_t = \omega_1 - \sqrt{\frac{P_v f_j}{b_d}} \tag{20}$$

式中， ω_1 表示信道传输比； f_j 表示数据平均粒度； b_d 表示异常数据最大维度。

若设异常数据的高阶统计量的聚类度为 λ_u ，则高维数据转换为低维数据的变换函数为：

$$y_0 = \frac{|q_{\max} - q_{\min}|}{O_t} \tag{21}$$

式中， $q_{\max}$ 、 $q_{\min}$ 分别表示数据特征的最大值与最小值。

对异常数据频域空间进行分解^[13]，得到异常数据的传输延迟为：

$$t_x = \frac{\int_{-\infty}^{+\infty} y_0 \times m_n \times J(x) dm_n}{\int_{-\infty}^{+\infty} m_n \times K(x)} \tag{22}$$

式中， m_n 表示异常数据的累积切片量； $J(x)$ 表示有效激活函数； $K(x)$ 表示特征函数。

计算数据样本属性的熵值：

$$H_\beta = - \frac{\sum t_x \times \log_2 Q(t)}{\log_2 n_a} \tag{23}$$

式中， $Q(t)$ 表示样本属性函数； n_a 表示待检测数据样本集合。

结合融入样本方差函数^[14]，设定检测异常数

据的固定阈值，即：

$$\psi = \frac{F(z)}{f_g} \tag{24}$$

式中， $F(z)$ 表示样本方差函数； f_g 表示数据平均适应度值。

将式(23)的计算过程遍历整个自由空间烟草信息化通信网络中的待检测数据，将计算得到的熵值 H_β 与固定阈值 ψ 进行比较，在阈值内的数据确定为正常数据，在阈值以外的数据确定为异常数据^[15]，至此，完成微服务架构下烟草信息化网络异常数据层次密度聚类检测算法的设计。

2 实验论证分析

为了验证本文提出的烟草信息化网络异常数据层次密度聚类检测算法的有效性，需要设计相关实验加以测试，并通过与其他常规方法的对比结果来分析本文方法的实际应用性能。

2.1 实验准备

为了更好地验证实时检测效果，网络异常数据实时检测实验中使用的数据源包括采集的本地数据和 CICID 52017 数据。实验数据集集中的 300 万条数据由正常烟草数据以及异常烟草数据构成，共包括数据堵塞、传输速率异常、路由异常和请求异常 4 种异常类型。

本次选择 Core i3PC 作为主要测试环境，基本内存为 4.5 GB。首先，对以上 4 种异常数据类型进行多属性特征分类，将同一类簇下的数据集合作为测试对象；其次，对网络端口进行连接测试操作，通过滑动窗口将通信节点与主、从节点连接；最后，对信息化网络系统的方差标准进行设定，具体表 1 所示。

表 1 网络端口距离方差标准设定表

网络端口 编号	预测距离	网络通信 系数	方差标准
1280361a	125.36	0.021	12.54
17R73002a	201.96	0.056	17.45
25C73122b	125.89	0.039	13.20
29B75489a	256.70	0.124	16.33
23D0210H	300.56	1.548	25.12
30S451L	426.54	2.060	31.63
15D451123T	152.36	3.012	25.14
18F23457b	230.78	0.257	10.10

将表 1 中的数据信息作为网络端口的通信限制标准,将网络系统的小时参数流量数据作为测试集,在此数据集中引入 2000 个存储节点。通过云存储实现资源调度,并利用管理节点负责数据的解算与上传处理。另外,为了防止实验过程中的外部干扰,采用运算节点对网络系统的存储开销机制进行调整,在验证没有错误后,开始具体的方法测试。

该实验使用 ZigBee 协议栈来实现节点之间的通信,其中路由器节点负责收集网络传输的数据,并定期将数据发送到协调器节点。此外,路由器节点可以自动将信道数据转发到网络中的其他节点。协调器节点接收路由器报告的数据,并通过串口上传到中央基站(PC),在 PC 上完成数据集成与聚类计算。

2.2 实验说明

在此基础上,调用后台数据,设计网络异常数据采样参数,具体如表 2 所示。

表 2 网络异常数据采样参数设置

序号	项目	参数
1	离散数据采样频率/kHz	15
2	异常数据样本长度/bit	1024
3	数据传输信道的频带宽度/kHz	10~30
4	数据传输信道时域宽度/ms	4.5
5	数据网络传输带宽/Hz	1.5×10^3
6	异常信号输出码元宽度/ms	0.15

在掌握了测试样本中携带的异常数据的分布后,使用本文设计的方法来检测异常数据。在检测过程中,建立烟草信息网络中异常通信通道的模型,提取通信过程中异常数据的特征,并通过计算数据熵值实现对异常通信数据的检测。

2.3 异常数据检测结果分析

采用文献[1]基于改进 K-means 算法的异常数据检测方法(方法 1)、文献[2]基于遗传算法的异常数据检测方法(方法 2)与本文方法进行检测性能对比。分别采用不同方法在不同异常数据类型条件下的检测准确度情况,检测准确度越高,表明方法的检测精度越高,结果如图 3 所示。

如上图所示,在不同数据传输量条件下,本文方法对于网络异常数据的检测准确率基本在 70% 以上,主要是由于本文方法能够根据异常数据特征属性进行层次分类,并结合所属类簇的局部可达密

度对不同属性的异常数据进行分段检测,保证了检测精度;方法 1 的检测准确率较低,在 33%~55% 之间,分析原因可知,该方法是以集中式查询与识别方法为基础对异常数据进行检测,对于异常特征变化参量较为复杂的数据,无法准确发掘;方法 2 的检测准确率范围为 32%~62%,该方法的检测准确率较低的原因主要是忽略了通信节点数据辨识度的影响,进而影响了检测精度。通过实验对比结果可知,本文提出的方法对于烟草信息化网络异常数据检测具有更高的检测精度。

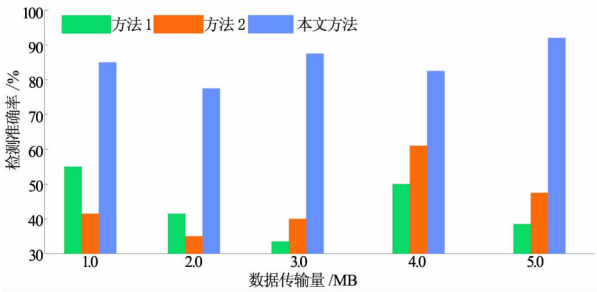


图 3 检测准确率结果对比

2.4 异常数据检测误报率对比实验分析

为进一步验证本文方法在异常数据检测精度方面的优越性,将不同方法应用于不同异常数据维数中,比较各方法的检测误报率,对比结果如图 4 所示。

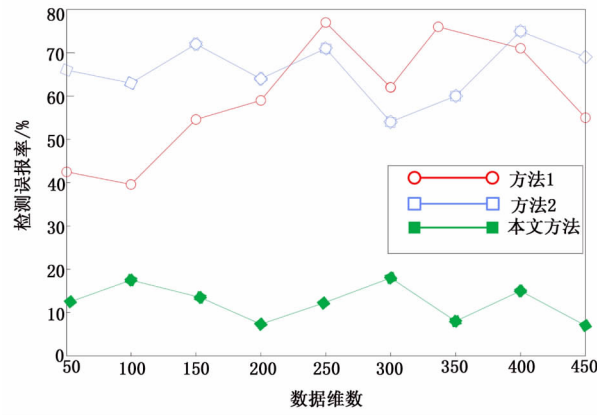


图 4 检测误报率结果对比

根据上图可知,在不同异常数据维数条件下,利用本文方法对网络异常数据进行检测,得到的检测误报率始终控制在 20% 以下,而方法 1 与方法 2 的检测误报率最高分别达到了 77.6% 和 75.3%。由此可以说明本文提出的方法能够以较高的精度检测出烟草信息化网络的异常数据。

3 结 论

为有效弥补当前既有方法对于烟草信息化网络异常数据 检测存在检测精度较低的不足,提出了一种层次密度聚类异常数据检测算法。通过识别异常数据与构建异常数据传输信道模型,并计算数据熵值,进而实现异常数据的检测。经过对比实验验证了本文方法的可靠性,表明提出的方法相较于传统方法具有更大的应用优势。

参考文献

[1] 李东昆,高险峰,张乃平,等. 基于改进 K-means 算法的光通信数据异常检测预警方法[J]. 自动化与仪器仪表,2023(7): 51—54.

[2] 魏丹,杨春琴,李国元,等. 基于遗传算法的网络传输异常数据检测方法[J]. 电子设计工程,2023,31(15): 94—97+106.

[3] 陈怡娜. 基于深度学习算法的无线传感器异常数据检测[J]. 计算技术与自动化,2023,42(2):178—183.

[4] 陈阳. 基于小波神经网络的光纤网络大数据异常负载检测方法[J]. 激光杂志,2023,44(5):123—127.

[5] 周佳航,邢红杰. 基于双自编码器和 Transformer 网络的异常检测方法[J]. 计算机应用,2023,43(1):22—29.

[6] 贾帅帅,田明浩,路红阳. 基于数据增强的组合神经网络异常检测算法[J]. 信息技术与信息化,2023(4):187—190.

[7] 贾旭超,马迅,刘安磊,等. 基于孤立森林法的电能计量数据异常检测方法[J]. 河北电力技术,2023,42(2):41—45.

[8] 拓广忠,葛树峰,李荣让,等. 基于熵值法和 DEA 的电力监测异常数据自动识别算法研究[J]. 微型电脑应用,2023,39(4):160—163+171.

[9] 陈波红,祝金明. 统计向量分析下大规模网络流量异常检测仿真[J]. 计算机仿真,2023,40(4):373—376+401.

[10] 孟献轲,张硕,熊诗,等. 基于时空注意力特征的异常流量检测方法[J]. 计算机应用与软件,2023,40(4):99—106.

[11] 王锐. 基于改进 LOF 的高维数据异常检测方法[J]. 电信工程技术与标准化,2023,36(3):41—45+62.

[12] 徐丽燕,徐康,黄兴挺,等. 基于 Transformer 的时序数据异常检测方法[J]. 计算机技术与发展,2023,33(3):152—160.

[13] 高聚鑫. 基于大数据的网络异常行为检测技术研究[J]. 数字通信世界,2023(2):30—32.

[14] 张仁斌,左艺聪,周泽林,等. 基于多模态生成对抗网络的多元时序数据异常检测[J]. 计算机科学,2023,50(5):355—362.

[15] 刘杰. 基于孤立森林算法的巡检机器人通信异常数据自动检测方法[J]. 信息与电脑(理论版),2023,35(3):86—88.