基于平衡迭代规约层次聚类的无线传感器网络流量异常检测方案

郁 滨 熊 俊*

(战略支援部队信息工程大学 郑州 450000)

摘 要：针对现有网络流量异常检测方法不适用于实时无线传感器网络(WSN)检测环境、缺乏合理异常判决机制的问题，该文提出一种基于平衡迭代规约层次聚类(BIRCH)的WSN流量异常检测方案。该方案在扩充流量特征维度的基础上，利用BIRCH算法对流量特征进行聚类，通过设计动态簇阈值和邻居簇序号优化BIRCH聚类过程，以提高算法的聚类质量和性能鲁棒性。进一步，设计基于拐点的综合判决机制，结合预测、聚类结果对流量进行异常检测，保证方案的检测准确性。实验结果表明，所提方案在检测效果和检测性能稳定性上具有较为明显的优势。

关键词：无线传感器网络；流量异常检测；特征维度扩充；基于平衡迭代规约层次聚类；拐点

中图分类号：TN915; TP391 **文献标识码**：A **文章编号**：1009-5896(2022)01-0305-09

DOI: 10.11999/JEIT201004

A Novel WSN Traffic Anomaly Detection Scheme Based on BIRCH

YU Bin XIONG Jun

(*Information Engineering University, PLA Strategic Support Force, Zhengzhou* 450000, *China*)

Abstract: For the problems that the existing network traffic anomaly detection methods are not suitable for the real-time WSN (Wireless Sensor Networks) and lack reasonable decision mechanisms, a novel Wireless Sensor Networks (WSN) traffic anomaly detection scheme based on BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is proposed. Based on expanding the dimension of traffic characteristics, the scheme uses BIRCH algorithm to cluster traffic characteristics. By introducing the dynamic cluster threshold and neighbor cluster serial numbers, the BIRCH process is optimized to improve the clustering quality and performance robustness. Furthermore, to ensure the detection accuracy of the scheme, a comprehensive decision mechanism based on turning point is designed to detect abnormal traffic, combined with prediction and clustering results. The experimental results show that the proposed scheme has obvious advantages in detection effect and stability of detection performance.

Key words: Wireless Sensor Networks (WSN); Traffic anomaly detection; Characteristic dimension expansion; Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH); Turning point

1 引言

无线传感器网络(Wireless Sensor Networks, WSN)流量异常检测技术运用数据挖掘、机器学习等方法对流量数据进行统计分析，以判断网络是否存在异常运行或入侵行为，对维护网络安全具有重要意义，如何及时、准确地检测出异常流量是当前WSN流量异常检测技术亟需解决的问题[1]。目前，网络流量异常检测技术主要包括基于神经网络模型、基于统计分析方法和基于聚类分析算法[2]。

在WSN流量异常实时检测中，直接通过连续采样获得的流量数据是没有标记的，考虑到通常需要利用大量带标签的数据训练神经网络模型以增强模型的泛化能力[3,4]，因此，基于神经网络模型的检测技术不适用于实时的WSN流量异常检测环境。基于统计分析方法的检测技术在检测实时性、算法计算复杂度等方面具有优势，但现有方案大多缺乏合理的异常判决机制，且忽略了统计误差对判决结果的影响，检测准确度较低[5–7]。基于聚类分析算法的检测技术无需预先处理输入样本，通过分析网络流量数据内部特征的相关性，以聚合具有相似特征的流量，并将稀疏簇内的流量判决为异常数据，逐渐成为网络流量异常检测技术的主要研究方向[8–10]。

现有的聚类分析算法可以被归纳为层次聚类分析、划分聚类分析和智能聚类分析3类[11]。智能聚类分析技术主要基于深度学习、核函数等机器学习

收稿日期：2020-11-30；改回日期：2021-04-21；网络出版：2021-08-18
*通信作者： 熊俊 970121059@qq.com
基金项目：信息保障技术重点实验室开放基金(KJ-15-104)
Foundation Item: The Key Laboratory of Information Assurance Technology Open Fund (KJ-15-104)

方法，该技术在WSN异常流量检测应用中存在与基于神经网络模型的检测技术相同的不适用性。划分聚类分析更倾向处理各个聚类簇大小比较接近的样本，因此该技术往往无法将孤立点数据或异常点数据从样本中分离出来，且聚类中心的选择也会对其聚类结果产生较大影响，同时由于划分聚类一般从总体上评判样本间的相似性，导致其不支持增量式数据源。平衡迭代规约层次聚类(Balanced Iterative Reducing and Clustering using Hierarchies, BIRCH)作为一种经典的层次聚类分析算法，仅通过一次扫描即可有效地组织大规模数据，在聚类质量、效率、稳定性和扩展性方面具有明显优势[12]。BIRCH的基本思想是通过肯定每一个样本点的差异性，先将他们视作一个个单独的聚类簇，再根据簇之间相似性的高低将他们分层合并。通过这种方式，BIRCH聚类分析算法不需要预先设定聚类值和聚类中心，在处理离散点方面表现突出，可以将异常点数据划分到独立的簇中，相较于划分聚类分析技术，更适用于网络流量异常检测。Pitolli等人[13]利用BIRCH聚类算法对样本特征进行分类，用以识别海量软件样本集中的恶意数据，具有较高的检测率和计算效率。Peng等人[14]提出一种基于主成分分析的P-BIRCH聚类算法，时间成本随着聚类数的增加而线性减少，有效地解决移动云环境下的大数据入侵检测问题。

一方面，通过连续采样获得的WSN流量是典型的连续时间序列。根据时间序列相邻值之间具有的时间相关性[15]，针对出现在网络流量平稳阶段或固定周期的突变流量，对比其前后一段时间内的采样流量可以发现，这些流量的急剧变化是反常的，有理由相信他们为异常流量。然而，由于流量值仅仅体现了网络在该采样周期内的流量大小，如果利用BIRCH对单一维度的流量序列进行聚类分析，会因为忽略了流量与其前后临近流量的相关性，从而导致将异常突变流量划分到高峰期或低谷期流量对应的聚类簇中，使得异常检测结果存在较大的偏差。

另一方面，Lorbeer等人[16]指出经典BIRCH基于相同距离划分簇类，存在不能处理自然数据形状集合，对样本输入顺序高度敏感等不足。针对此，Guo等人[17]提出一种基于链接的LBIRCH算法，通过建立邻居表实现对任意形状进行聚类。同时，由于经典BIRCH采用全局静态阈值建立聚类特征树(Clustering Feature Tree, CF-Tree)，只能获得具有相同体积的聚类。因此，尚家泽等人[18]结合朴素贝叶斯算法，提出一种基于自适应阈值的改进BIRCH，一定程度上解决了其不适用于聚类体积差异较大簇类的局限性，但算法执行效率明显下降。

综上所述，针对WSN流量异常实时检测需求，该文提出一种基于BIRCH的WSN流量异常检测方案。首先，扩充WSN流量的特征维度，在此基础上，设计一种优化特征聚类树(Optimized Clustering Feature Tree, OCF-Tree)结构对BIRCH进行优化。然后，提出基于拐点的综合判决机制，进一步弥补聚类偏差对检测结果的影响。最后，根据基于仿真平台获得的WSN流量数据，对比该文方案与其他方案的流量异常检测效果。

2 模型建立

2.1 符号与定义

为方便对方案进行描述，该文相关符号及其含义如表1所示。

定义1 利用WSN的前m个连续时间流量值$s_1, s_2, \cdots, s_m$，预测后p个连续时间流量值$s_{m+1}, s_{m+2}, \cdots, s_{m+p}$，称为$m$-$p$流量预测函数，记作$\Gamma_{m-p}(\cdot)$。

定义2 假设WSN流量序列$\boldsymbol{S}=[s_1, s_2, \cdots, s_n]$，称$\Gamma_{m-p}(\boldsymbol{S})=[\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n]$为$\boldsymbol{S}$的预测序列，记作$\hat{\boldsymbol{S}}$。

定义3 对于WSN流量序列$\boldsymbol{S}$，其预测序列$\hat{\boldsymbol{S}}$，称$[\Delta s_1, \Delta s_2, \cdots, \Delta s_n]$为$\boldsymbol{S}$的预测误差序列，记作$\boldsymbol{S}_\Delta$，其中$\Delta s_i=|s_i-\hat{s}_i|\ (i=1,2,\cdots,n)$。

定义4 假设簇C由n个d维数据$\boldsymbol{X}_i=[x_{1,i}, x_{2,i}, \cdots, x_{d,i}]^{\mathrm{T}}$构成，称$[x_{1,0}, x_{2,0}, \cdots, x_{d,0}]^{\mathrm{T}}$为$C$的质心，记作$\boldsymbol{X}_0$，其中$x_{k,0}=\sum_{=1}^{n} x_{k,i}/n(k=1,2,\cdots,d)$。

表 1 符号定义

符号	含义	符号	含义
$\boldsymbol{S}$	WSN流量序列	$\{C_i\}$	BIRCH聚类簇
s_i	流量值	$\mathrm{tp}(Y)$	序列Y的拐点
$\hat{\boldsymbol{S}}$	预测序列	cluster_T	聚类截断阈值
$\boldsymbol{S}_\Delta$	预测误差序列	predict_T	预测截断阈值
$\boldsymbol{X}$	3维流量特征序列	P	聚类可疑点集合
$\boldsymbol{X_i}$	3维流量特征值	Q	预测可疑点集合

定义5　假设簇C_1和C_2的质心分别为$\boldsymbol{X}_0^1$和$\boldsymbol{X}_0^2$，称$\boldsymbol{X}_0^1$与$\boldsymbol{X}_0^2$的欧氏范数$\left\|\boldsymbol{X}_0^1-\boldsymbol{X}_0^2\right\|$为$C_1$与$C_2$的距离，记作$\mathrm{dist}(C_1,C_2)$。

定义6　假设单调递减序列Y，称斜率$(Y(b)-Y(a))/(b-a)$为Y在区间$[a,b]$的平均变化率，记作k_a^b。

定义7（拐点）[19]　基于定义6，若i_{tp}满足$i_{tp}=\mathrm{argmax}(k_i^{i+1}/k_1^i)$，则称$i_{tp}$为单调递减序列$Y$的拐点，记作$\mathrm{tp}(Y)$。

2.2　流量异常检测模型

基于BIRCH的WSN流量异常检测模型如图1所示，主要包括流量特征聚类分析和流量异常判决两个关键部分。

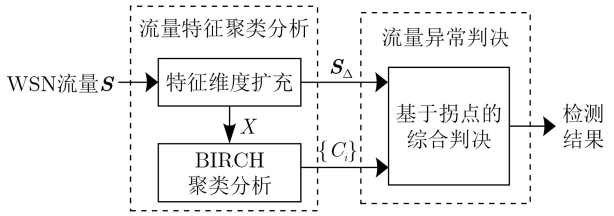


图 1 基于BIRCH的WSN流量异常检测模型

2.2.1　流量特征聚类分析

流量特征聚类分析分为特征维度扩充和BIRCH聚类分析两个环节。

WSN流量预测技术以网络历史流量信息为依据推测未来一段时间内的流量趋势[20]。根据定义2和定义3可知，某时刻流量预测值代表了该时刻流量的变化趋势及基准值，即正常流量的变化可能在预测值附近小范围波动，所以流量预测序列和预测误差序列均隐含了原始流量的时序特征。因此，针对WSN流量的聚类特征维度较低的问题，特征维度扩充环节根据输入的WSN流量序列$\boldsymbol{S}$，利用其预测序列$\hat{\boldsymbol{S}}$和预测误差序列$\boldsymbol{S}_\Delta$作为原始流量的新增特征，以扩充聚类分析输入样本的特征维度。

BIRCH聚类分析环节设计一种特殊的OCT-Tree结构，根据流量特征X将流量划分成簇$\{C_i\}$$(i=1,2,\cdots,K)$，其中$K$为簇个数。一方面，由于WSN流量具有突发性、分布不均匀等特征，且存在稀疏的异常流量点，基于此，通过为每个聚类特征(Clustering Feature, CF) 单独设置增量式的动态阈值T，使其适用于聚类体积存在较大差异的簇。另一方面，对于经典BIRCH算法，每个节点只能容纳固定数目的CF，聚类结果不总对应于自然集群。同时，根据流量的输入顺序，特征差异较大的流量可能会被聚类到同一簇中，针对此，根据每个叶子特征CF^{L}的邻居簇对聚类结果进行全局优化。

2.2.2　流量异常判决

流量异常判决环节根据拐点确定聚类截断阈值cluster_T和预测截断阈值predict_T，将cluster_T作为区分簇体积大小的依据，predict_T作为区分预测误差大小的依据。利用cluster_T从$\{C_i\}$中筛选出体积较小的簇，将小体积簇中的流量构成聚类可疑点集合P，同时利用predict_T从$\boldsymbol{S}_\Delta$中筛选出预测误差较大的流量，构成预测可疑点集合Q，并基于P, Q综合判决WSN流量是否异常。

3　流量特征聚类分析算法

WSN流量的特征维度扩充如图2所示，分别将流量序列作为第1维特征，预测序列作为第2维特征，预测误差序列作为第3维特征，合并$\boldsymbol{S}$, $\hat{\boldsymbol{S}}$和$\boldsymbol{S}_\Delta$，组成3维流量特征序列$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n]$，其中$\boldsymbol{X}_i=[\ s_i\quad \hat{s}_i\quad \Delta s_i\ ]^{\mathrm{T}}, i=1,2,\cdots,n$。特征维度扩充基于流量预测技术，由于目前已有较为成熟的研究，众多流量预测算法具有计算复杂度低、预测速度快、预测精度高等优势[21]，因此m-p流量预测函数$\Gamma_{m-p}(\cdot)$的内部结构不在该文讨论范围内。

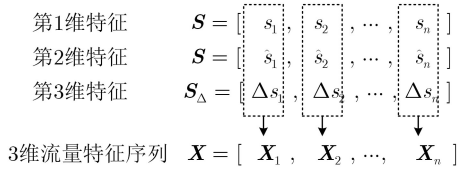


图 2 特征维度扩充示意图

BIRCH聚类分析的实质是构建一棵优化聚类特征树，为不失一般性，假设一棵高度为h+1的OCF-Tree，如图3所示。由下至上，BIRCH将所有样本点划分成若干个聚类簇，第h+1层叶子节点代表一个由若干个聚类簇组成的第h+1层集群，第i层非叶节点代表一个由若干个第i+1层子集群组成第i层集群，则整个OCF-Tree可以被视作一个由所有样本点组成的具有嵌套式结构的集群。其中，叶子特征$\mathrm{CF}_j^L=(N,\mathbf{LS},\mathrm{SS},T,\mathbf{NB})$是对第$j$个聚类簇中样本点总体特征的摘要，非叶特征$\mathrm{CF}_{i,j}^{\mathrm{NL}}=(N,\mathbf{LS},\mathrm{SS})$是对第$j$个第$i$+1层集群中样本点特征的总体描述，$N$为CF含有的样本点个数，$\mathbf{LS}=\sum_{i=1}^{N}\boldsymbol{X}_i$为CF内所有样本点的线性和，$\mathrm{SS}=\sum_{i=1}^{N}\boldsymbol{X}_i^2$为CF内所有样本点的平方和，$T$为动态簇阈值，$\mathbf{NB}$为邻居簇序号。

从层次聚类角度来看，OCF-Tree的构建过程基于欧氏距离最近原则，把相互靠近的样本点聚集成聚类簇，再不断将相近的聚类簇汇聚成更大的集群，最终汇聚成一个包含整个样本的集群。对于每

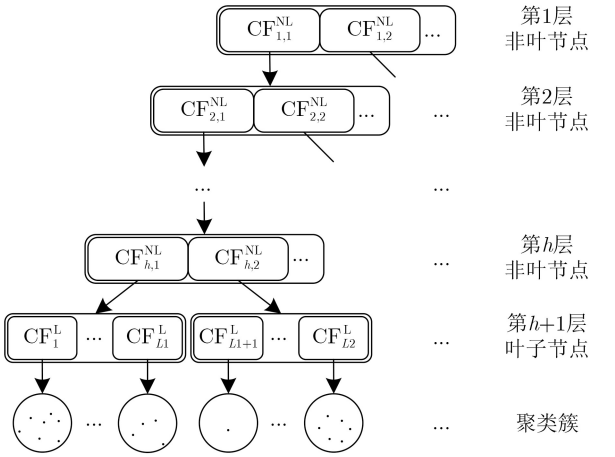


图 3 优化聚类特征树结构

个待聚类的样本点，假设其特征摘要为CF，首先在第1层集群中搜索质心与其欧氏距离最近的$\mathrm{CF}^{\mathrm{NL}}_{1,j_1}$，再在$\mathrm{CF}^{\mathrm{NL}}_{1,j_1}$对应的第2层集群中搜索质心与其欧氏距离最近的$\mathrm{CF}^{\mathrm{NL}}_{2,j_2}$，以此类推由上至下逐层搜索最近的集群，直至搜索到质心与其欧氏距离最近的$\mathrm{CF}^{\mathrm{L}}_{j_{h+1}}$，最后根据$\mathrm{CF}^{\mathrm{L}}_{j_{h+1}}$对应聚类簇的簇内平均距离与其阈值的关系，判断该样本点能否划分到该聚类簇中。通过这种方式，BIRCH在构建OCF-Tree的同时，实现对所有样本点的动态聚类。

当样本点插入最近的聚类簇后，需要依次向上更新搜索路径上叶子节点的CF^{L}和非叶节点的$\mathrm{CF}^{\mathrm{NL}}$。根据定义4和定义5，假设由$n$个$\mathrm{CF}_i(i=1,2,\cdots,n)$合并的CF，对于$\mathrm{CF}^{\mathrm{NL}}$，其更新规则由式(1)—式(3)给出；对于$\mathrm{CF}^{\mathrm{L}}$，其更新规则由式(1)～式(5)给出，其中：$T_C=0.15\cdot {R_C}^2+0.3\cdot \boldsymbol{S}(C)$, $R_C=\sqrt{\sum_{i=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{X}_0\|^2/n}$为CF代表的簇$C$的半径，$S(C)$为CF代表的簇$C$与其质心距离的标准差。

$$N=\sum_{i=1}^{n}N_i \tag{1}$$

$$\mathbf{LS}=\sum_{i=1}^{n}\mathbf{LS}_i \tag{2}$$

$$\mathrm{SS}=\sum_{i=1}^{n}\mathrm{SS}_i \tag{3}$$

$$T=\max\left(\max\left(T_i\right),T_C\right)+0.25\cdot\min\left(\max\left(T_i\right),T_C\right) \tag{4}$$

$$\mathbf{NB}=\mathbf{NB}_0 \tag{5}$$

流量特征聚类分析算法具体如下：

输入：3维流量特征序列$\boldsymbol{X}$。

输出：BIRCH聚类簇$\{C_i\}(i=1,2,\cdots,K)$。

(1) 扩充特征维度

步骤1 利用$\Gamma_{m-p}(\cdot)$预测$\boldsymbol{S}$的$\hat{\boldsymbol{S}}$；

步骤2 根据$\boldsymbol{S}$和$\hat{\boldsymbol{S}}$，计算$\boldsymbol{S}_\Delta$；

步骤3 合并$\boldsymbol{S}$, $\hat{\boldsymbol{S}}$, $\boldsymbol{S}_\Delta$，组成$\boldsymbol{X}$；

(2) 初始化OCF-Tree

步骤4 生成一个根节点$\mathrm{CF}^{\mathrm{NL}}_{1,1}=(0,0,0)$；

步骤5 扫描$\boldsymbol{X}$，为$\boldsymbol{X}$生成一个$\mathrm{CF}_i=(N_i,\mathbf{LS}_i,\mathrm{SS}_i)$，由上至下搜索距离$\mathrm{CF}_i$最近的$\mathrm{CF}^{\mathrm{L}}_j$；

步骤6 根据式(6)计算CF_i与$\mathrm{CF}^{\mathrm{L}}_j$的簇内平均距离$D$。

$$D=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2/n(n-1)} \tag{6}$$

若$D<T$，将CF_i插入$\mathrm{CF}^{\mathrm{L}}_x$，根据式(1)—式(5)更新$\mathrm{CF}^{\mathrm{L}}_j$，转至步骤9；否则，将$\mathrm{CF}_i$转化为$\mathrm{CF}^{\mathrm{L}}_i=(N_i,\mathbf{LS}_i,\mathrm{SS}_i,T_0,\mathbf{NB}_0)$插入该叶子节点中，其中$T_0=0.15\cdot {R_X}^2+0.3\cdot\boldsymbol{S}(\boldsymbol{X})$为初始簇阈值，$\mathbf{NB}_0=[\varnothing]$为初始邻居簇序号，$R_X$为整体样本的半径，$\boldsymbol{S}(\boldsymbol{X})$为整体样本与其质心距离的标准差，转至步骤7；

步骤7 判断该叶子节点中CF^{L}是否超过L，若超过，将节点分裂为两个新节点，将距离最远的两个CF^{L}分别放到两个新节点中，其余的CF^{L}按照距离最近原则划分到两个新节点中，转至步骤8；否则，转至步骤9；

步骤8 针对子节点发生分裂的非叶节点，判断该节点中$\mathrm{CF}^{\mathrm{NL}}$是否超过$B$，若超过，将节点分裂为两个新节点，将距离最远的两个$\mathrm{CF}^{\mathrm{NL}}$分别放到两个新节点中，其余的$\mathrm{CF}^{\mathrm{NL}}$按照距离最近原则划分到两个新节点中，转至步骤8；否则，转至步骤9；

步骤9 从CF_i插入的叶子节点开始由下至上，根据式(1)—式(3)依次更新路径上的每一个$\mathrm{CF}^{\mathrm{NL}}$，转至步骤5重复上述步骤，直至扫描完所有特征值；

(3) 合并邻居簇阶段

步骤10 将OCF-Tree中所有CF^{L}从左至右排序$\mathrm{CF}^{\mathrm{L}}_j(j=1,2,\cdots,K')$，分别对应聚类簇$C_j(j=1,2,\cdots,K')$；

步骤11 对于第k个$\mathrm{CF}^{\mathrm{L}}_k$，计算$\mathrm{dist}(C_k,C_{\mathrm{other}})$ $(\mathrm{other}=k+1,k+2,\cdots,K')$，若$\mathrm{dist}(C_k,C_{\mathrm{other}})<T_k+T_{\mathrm{other}}$，则$C_k$与$C_{\mathrm{other}}$互为邻居簇，分别将对方的簇序号添加到$\mathbf{NB}$中。重复上述步骤，直至更新每一个$\mathrm{CF}^{\mathrm{L}}_j$的$\mathbf{NB}$；

步骤12 根据式(7)在$\mathbf{NB}_k$中寻找C_k的最适邻居簇C_a，其中$a=\mathrm{best}_{\mathrm{NB}}(C_k)\in\mathbf{NB}_k$，$\gamma=0.5$，$\mathrm{intersect}(\mathbf{NB}_k,\mathbf{NB}_a)$为$C_k$与$C_a$的共有邻居簇个数。若$a=\mathrm{best}_{\mathrm{NB}}(C_{\mathrm{best}_{\mathrm{NB}}(C_k)})$，即$C_k$与$C_a$互为最适邻居簇，则将两者合并为一个簇，同时将$k$和$a$从簇

C_b的$\mathbf{NB}_b$中剔除，$b \in (\mathbf{NB}_k \cup \mathbf{NB}_a)$。重复上述步骤，直至合并所有最适邻居簇；

$$\begin{aligned}&\text{best}_{\text{NB}}(C_k)\\&\quad=\text{argmax}\left(\text{intersect}(\mathbf{NB}_k,\mathbf{NB}_a)/(N_k^{\gamma}+N_a^{\gamma})\right)\end{aligned}\tag{7}$$

步骤13　将合并后的C重新排序，输出$\{C_i\}$ $(i=1,2,\cdots,K)$。

值得注意的是，BIRCH聚类算法所需内存与阈值T有关。T越大，构建的CT-Tree规模就越小，则所占用的内存也越小。然而，如果T过大，会导致单个聚类簇的规模较大，则聚类程度十分粗略，从而影响聚类效果。由于本文着重研究如何提高BIRCH的聚类质量，因此，假设流量特征聚类分析算法步骤均在内存足够的前提下进行。

4　流量异常判决方法

相较于正常流量，在异常流量的特征值$\boldsymbol{X}$中，第1维特征值s与第2维特征值$\hat{s}$会存在明显差异，即第3维特征值Δs较大，经BIRCH聚类后，每个异常流量通常会单独构成一个小簇，可以根据BIRCH聚类簇$\{C_i\}$的体积大小来判断流量是否异常。然而，如何选取一个合适的截断阈值以区分簇体积的大小仍需进一步讨论。

根据定义6，对于一个单调递减的离散序列Y，k_i^{i+1}表示Y在区间$\{i,i+1\}$的斜率，k_1^i表示Y在区间$\{1,2,\cdots,i\}$的平均变化率，则将两者的比值k_i^{i+1}/k_1^i称为序列Y在第i个点的趋势变化幅值，可以用来描述序列“总体”变化的急缓趋势。由于趋势变化幅值k_i^{i+1}/k_1^i体现了序列Y在第i个点的“总体”趋势变化的快慢，因此基于定义7，拐点$\text{tp}(Y)$即为趋势变化幅值最大值$\max(k_i^{i+1}/k_1^i)$对应的横坐标点$i_{\text{tp}}=\text{argmax}(k_i^{i+1}/k_1^i)$，其物理意义是序列$Y$中“总体”趋势变化最快的位置，是区分序列变化从骤变到平缓的临界点。

基于此，流量异常判决环节设计一种基于拐点的综合判决机制，利用拐点确定区分聚类簇体积大小的截断阈值，如图4(a) 所示。进一步，利用拐点确定区分流量预测是否失真的截断阈值，如图4(b) 所示，以避免当网络流量出现突发性变化或流量预测误差较大时，正常流量被划分到小体积的簇中的情况。流量异常判决方法具体步骤如下：

输入：BIRCH聚类簇$\{C_i\}(i=1,2,\cdots,K)$，预测误差序列$\boldsymbol{S}_{\Delta}=[\Delta s_1,\Delta s_2,\cdots,\Delta s_n]$。

输出：流量异常检测结果。

(1) 确定截断阈值

步骤1　分别将$\{C_i\}$中的流量个数序列、$\boldsymbol{S}_{\Delta}$由大到小排列并删除重复值，得到关于流量个数的单调递减序列$\boldsymbol{C}^*=[C_1^*,C_2^*,\cdots,C_{K'}^*]$与预测误差单调递减序列$\boldsymbol{S}_{\Delta}^*=[\Delta s_1^*,\Delta s_2^*,\cdots,\Delta s_{n'}^*]$；

步骤2　根据定义7分别求解$\boldsymbol{C}^*$的拐点$\text{tp}(\boldsymbol{C}^*)$与$\boldsymbol{S}_{\Delta}^*$的拐点$\text{tp}(\boldsymbol{S}_{\Delta}^*)$；

步骤3　分别选取拐点$\text{tp}(\boldsymbol{C}^*)$对应的流量点个数$\boldsymbol{C}^*\left(\text{tp}(\boldsymbol{C}^*)\right)$作为聚类截断阈值cluster_T；拐点$\text{tp}(\boldsymbol{S}_{\Delta}^*)$对应的预测误差$\boldsymbol{S}_{\Delta}^*\left(\text{tp}(\boldsymbol{S}_{\Delta}^*)\right)$作为预测截断阈值predict_T；

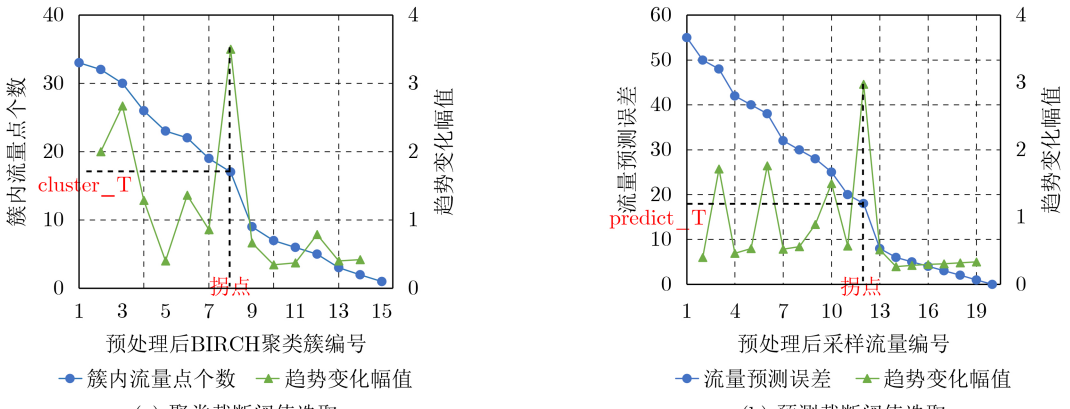
(2) 判定流量正异

步骤4　将$\{C_i\}$中流量个数小于cluster_T的簇判定为小体积簇，簇中的流量构成聚类可疑点集合P；将预测误差大于等于predict_T的流量判定为大误差流量，构成预测可疑点集合Q；

步骤5　取流量值$s_i \in (P \cup Q)$，若$s_i \in (P \cap Q)$，则判决s_i为异常流量；否则，转至步骤6；

步骤6　判决s_i为正常流量。若$s_i \in Q$，则认为s_i对应的预测误差Δs_i在误差允许范围内；否则，认为s_i对应的预测误差Δs_i较大；

步骤7　若完成遍历$(P \cup Q)$，则流量异常判决结束；否则，转至步骤5。



(a) 聚类截断阈值选取　　(b) 预测截断阈值选取

图 4 截断阈值选取

根据上述方法步骤，实现对WSN流量异常的综合检测。

5 实验及结果分析

实验基于Ubuntu系统平台，利用NS-2网络模拟器进行WSN仿真实验，采用Python3.6语言和eclipse环境检验该文方案的有效性。

5.1 实验数据与参数配置

5.1.1 实验数据

实验利用NS-2网络仿真平台搭建WSN仿真环境，配置如表2所示。

实验数据基于上述WSN仿真环境，以采样频率0.25 Hz统计WSN区域内的流量以模拟实时采样环节，得到500个流量样本点。将第1至400个样本点作为$\Gamma_{m-p}(\cdot)$输入，以预测第401至500个流量样本点，作为流量特征的扩充维度。针对本文方案提出的m-p流量预测函数$\Gamma_{m-p}(\cdot)$，采用文献[21]提出的优化FAEMD-OSELM流量预测模型，其中m=400, p=100。

在上述流量采样过程中，分别采用Blackhole, Flooding和Grayhole攻击模型[22]在第451至480个样本时间段模拟网络异常情况，如图5所示。将包含异常流量的第401至500个网络流量分别作为Blackhole, Flooding和Grayhole测试数据集，其中第451至480个为网络异常时的采样流量，其余为网络正常时的采样流量。

5.1.2 实验参数设置

实验选择文献[23]中的ENDTW-O-CFSFDP流量异常检测模型、文献[24]中的BasisEvolution流量异常检测模型和文献[14]中的BIRCH流量异常检测模型与本文方案进行对比，结合Blackhole, Flooding, Grayhole 3种测试数据集，综合比较方案的流量异常检测性能。相较于基于PCA-BIRCH的方案，本文方案设计的启发式阈值T无需人为设定，更具自适应性，其初始值T_0由测试数据集Blackhole, Flooding, Grayhole确定，分别为4.18, 3.89, 3.94。

本文方案需要设定的参数为非叶节点分支因子

表 2 WSN仿真环境配置

环境		配置
操作系统		Ubuntu 18.04.1
仿真平台		NS-2.35网络模拟器
WSN网络设置	区域规模	100×100 (m^2)
	节点个数	终端节点20 (个)
		汇聚节点3 (个)
	路由协议	AODV
	工作模式	周期报告模式

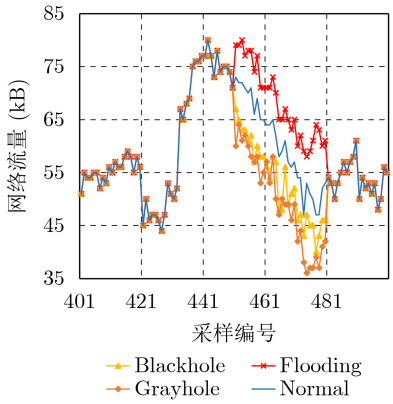


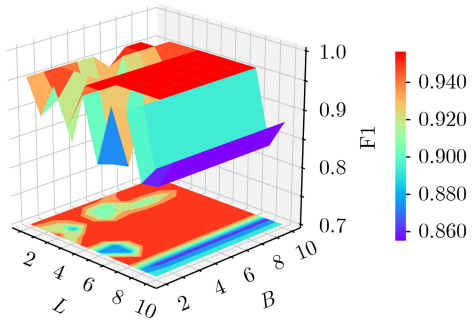
图 5 WSN异常流量数据

B和叶子节点分支因子L，通常B取4, 5, 6，L取4, 5, 6, 7为最佳聚类效果经验值。由于B, L一定程度上决定了OCF-Tree的形状以影响聚类效果，因此实验针对3种测试数据集，分别设置B取2至10，L取1至10，研究B, L取值对本文方案检测结果的F1值和准确率的影响。如图6所示，若选取的B或L过大，距离簇质心较远的正常样本将被划分成独立的聚类簇，并被判决为异常流量，从而增加了正常样本被错误判决的比例；相反地，若选取的B或L过小，受限于OCF-Tree的节点分支数量，异常样本将被强制划分到距离最近的聚类簇中，并增大该聚类簇的规模，导致其被判决为正常流量，从而增加了异常样本被错误判决的比例。在上述两类情况下，本文检测结果的F1值、准确率较低，检测效果较差。因此，实验选择聚类效果最好时的B=4, L=5作为本文方案的实验参数。

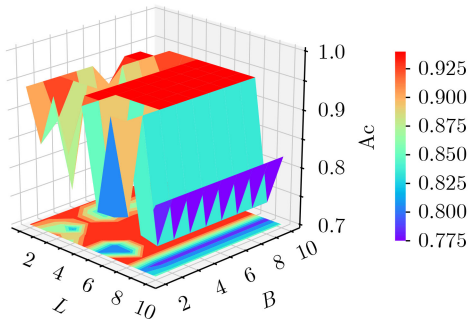
5.2 结果分析

实验选择精确率 (Precise, Pr) 、召回率 (Recall, Re) 、F1值和准确率 (Accuracy, Ac) 作为方案流量异常检测效果的评价指标。Pr为被正确判决的正常样本占所有被判决为正常点的比例，Re为被正确判决的正常样本占所有正常样本的比例，F1为反映方案整体检测效果的综合评价指标，Ac为被正确判决的样本占所有样本的比例，其中Pr, Re, F1$\in$[0,1]，其值越大表明方案检测效果越好。通过20次独立重复实验，检测方案针对3种测试数据集的Pr, Re, F1和Ac如表3所示。

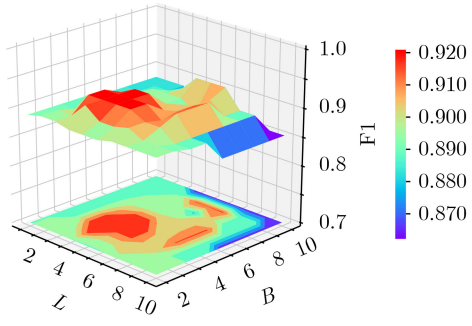
针对3种测试集，基于BasisEvolution方案将网络流量判决为异常样本的数量明显高于其他方案，虽然该方案可以正确检测出大部分的异常流量，在Pr方面具有绝对优势，但同时也将较多的正常流量错误地判决为异常样本，导致其Re较低。相反地，基于ENDTW-O-CFSFDP方案正常流量的误判率低于其他方案，在Re方面具有一定的优势，



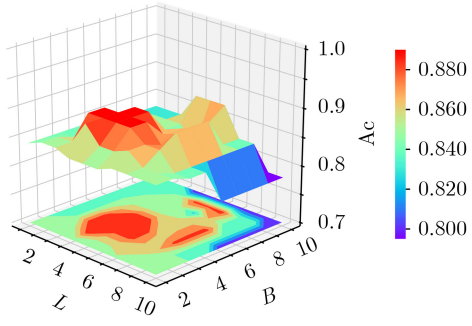
(a) B, L对Blackhole测试集的F1值影响



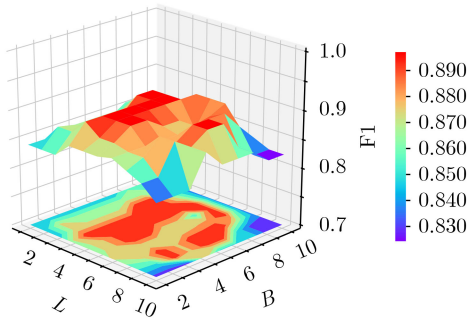
(b) B, L对Blackhole测试集的准确率影响



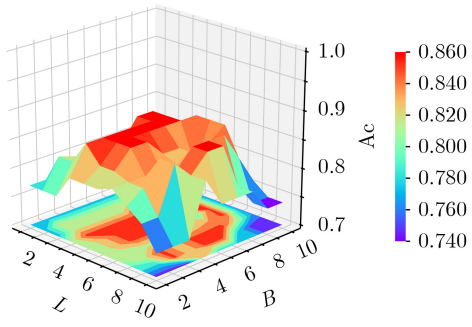
(c) B, L对Flooding测试集的F1值影响



(d) B, L对Flooding测试集的准确率影响



(e) B, L对Grayhole测试集的F1值影响



(f) B, L对Grayhole测试集的准确率影响

图 6 B, L的取值对F1值和准确率的影响

表 3　各方案检测性能对比(%)

方案	测试集	Pr	Re	F1	Ac
ENDTW-O-CFSFDP	Blackhole	82.4	**100.0**	90.3	85.0
	Flooding	82.9	90.0	86.3	80.0
	Grayhole	75.3	**87.1**	80.8	71.0
BasisEvolution	Blackhole	**100.0**	85.7	92.3	90.0
	Flooding	**96.5**	78.5	86.6	83.0
	Grayhole	**92.6**	71.4	87.0	76.0
BIRCH	Blackhole	79.7	90.0	84.6	77.0
	Flooding	78.1	81.4	79.7	71.0
	Grayhole	75.7	75.7	75.7	66.0
本文方案	Blackhole	97.1	94.3	**95.7**	**94.0**
	Flooding	92.8	**91.4**	**92.1**	**89.0**
	Grayhole	92.4	**87.1**	**89.7**	**86.0**

且略高于该文方案，但由于该方案的异常流量检出率较低，其Pr较低。基于BIRCH的方案仅根据BIRCH聚类结果判断流量是否异常，缺乏合理的异常判决机制，其Pr与Re均低于该文方案。

进一步，相较于基于ENDTW-O-CFSFDP, BasisEvolution, BIRCH的方案，在综合指标F1值方面，本文方案对Blackhole测试集分别提高了5.3%, 3.3%, 11.0%，对Flooding测试集分别提高了5.8%, 5.4%, 12.4%，对Grayhole测试集分别提高了8.9%, 9.1%, 14.0%；在Ac方面，本文方案对Blackhole测试集分别提高了9.0%, 4.0%, 17.0%，对Flooding测试集分别提高了9.0%, 6.0%, 18.0%，对Grayhole测试集分别提高了15.0%, 10.0%, 20.0%，表明本文方案在检出异常流量的同时，可以有效避免将正常样本判决为异常样本，且针对不同网络攻击造成的流量异常现象都具有较好的检测能力。

如表4所示，相较于基于ENDTW-O-CFSFDP, BasisEvolution, BIRCH的方案，本文方案F1值的均值分别提高了6.7%, 6.0%, 12.5%，Ac的均值分别提高了11.0%, 6.7%, 18.3%，具有较好的异常流量的检出能力，表明本文方案对整个测试样本的正负类判别准确度较好。同时，本文方案根据网络流量特征之间的差异先筛选出可疑流量点，再结合预测值综合判决流量是否异常，因此，本文方案对应的F1值与Ac的标准差最小，表明本文方案针对不同类型网络攻击造成的异常流量都具有较为稳定的检测性能。

综上所述，本文提出的流量异常检测方案可以有效检测由不同WSN网络攻击导致的异常流量，且检测性能稳定。

6 结束语

本文在深入研究网络流量异常检测技术的基础上，提出了一种基于BIRCH的WSN流量异常检测方案。本文方案设计了基于OCF-Tree的BIRCH聚类分析算法，有效克服样本的特征类型和输入顺序对BIRCH造成的影响，提高了聚类的质量和稳定性。同时，提出一种基于预测的特征维度扩充方法，通过在流量特征中增加隐含流量时序关系的预测序列和误差序列，以适应BIRCH聚类分析算法。进一步，根据流量序列特征分析，定义拐点概念，设计基于拐点的综合判决机制，减少BIRCH聚类分析过程对检测结果的影响，保证了流量异常检测的准确性。实验结果表明，相较于同类方案，本文方案在检测效果和性能稳定性方面具有明显优势。

下一步研究中，将针对OCF-Tree结构中分支因子B, L的选择，设计启发式算法求解最优值，同时，解决如何在保证聚类质量的前提下尽可能减小构建OCF-Tree所需内存的问题。

表 4 各方案F1值、Ac的均值(%) 与标准差对比

方案	F1		Ac	
	均值	标准差	均值	标准差
ENDTW-O-CFSFDP	85.8	0.039	78.7	0.056
BasisEvolution	86.5	0.048	83.0	0.057
BIRCH	80.0	0.036	71.3	0.045
本文方案	**92.5**	**0.030**	**89.7**	**0.040**

参 考 文 献

[1] CHIRAYIL A, MAHARJAN R, and WU C S. Survey on anomaly detection in wireless sensor networks (WSNs)[C]. 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Toyama, Japan, 2019: 150–157. doi: 10.1109/SNPD.2019.8935827.

[2] FERNANDES G Jr, RODRIGUES J J P C, CARVALHO F L F, *et al*. A comprehensive survey on network anomaly detection[J]. *Telecommunication Systems*, 2019, 70(3): 447–489. doi: 10.1007/s11235-018-0475-8.

[3] 闻佳, 王宏君, 邓佳, 等. 基于深度学习的异常事件检测[J]. 电子学报, 2020, 48(2): 308–313. doi: 10.3969/j.issn.0372-2112.2020.02.013.

WEN Jia, WANG Hongjun, DENG Jia, *et al*. Abnormal event detection based on deep learning[J]. *Acta Electronica Sinica*, 2020, 48(2): 308–313. doi: 10.3969/j.issn.0372-2112.2020.02.013.

[4] 董书琴, 张斌. 基于深度特征学习的网络流量异常检测方法[J]. 电子与信息学报, 2020, 42(3): 695–703. doi: 10.11999/jeit190266.

DONG Shuqin and ZHANG Bin. Network traffic anomaly detection method based on deep features learning[J]. *Journal of Electronics & Information Technology*, 2020, 42(3): 695–703. doi: 10.11999/jeit190266.

[5] KADRI F, HARROU F, CHAABANE S, *et al*. Seasonal ARMA-based SPC charts for anomaly detection: application to emergency department systems[J]. *Neurocomputing*, 2016, 173: 2102–2114. doi: 10.1016/j.neucom.2015.10.009.

[6] MATSUDA T, MORITA T, KUDO T, *et al*. Traffic anomaly detection based on robust principal component analysis using periodic traffic behavior[J]. *IEICE Transactions on Communications*, 2017, E100. B(5): 749–761. doi: 10.1587/transcom.2016EBP3239.

[7] DE LA PUERTA J G, FERREIRA I G, BRINGAS P G, *et al*. International Joint Conference SOCO'14-CISIS'14-ICEUTE'14[M]. Bilbao, Spain: Springer, 2014: 545–554.

[8] 王婷, 王娜, 崔运鹏, 等. 基于半监督学习的无线网络攻击行为

检测优化方法[J]. 计算机研究与发展, 2020, 57(4): 791–802. doi: 10.7544/issn1000-1239.2020.20190880.
WANG Ting, WANG Na, CUI Yunpeng, *et al*. The optimization method of wireless network attacks detection based on semi-supervised learning[J]. *Journal of Computer Research and Development*, 2020, 57(4): 791–802. doi: 10.7544/issn1000-1239.2020.20190880.

[9] MAZARBHUIYA F A, ALZAHRANI M Y, and GEORGIEVA L. Anomaly detection using agglomerative hierarchical clustering algorithm[C]. International Conference on Information Science and Applications, Singapore, 2018: 475–484. doi: 10.1007/978-981-13-1056-0_48.

[10] FAROUGHI A and JAVIDAN R. CANF: Clustering and anomaly detection method using nearest and farthest neighbor[J]. *Future Generation Computer Systems*, 2018, 89: 166–177. doi: 10.1016/j.future.2018.06.031.

[11] 章永来, 周耀鉴. 聚类算法综述[J]. 计算机应用, 2019, 39(7): 1869–1882. doi: 10.11772/j.issn.1001-9081.2019010174.
ZHANG Yonglai and ZHOU Yaojian. Review of clustering algorithms[J]. *Journal of Computer Applications*, 2019, 39(7): 1869–1882. doi: 10.11772/j.issn.1001-9081.2019010174.

[12] ZHANG T, RAMAKRISHNAN R, and LIVNY M. BIRCH: a new data clustering algorithm and its applications[J]. *Data Mining and Knowledge Discovery*, 1997, 1(2): 141–182. doi: 10.1023/A:1009783824328.

[13] PITOLLI G, ANIELLO L, LAURENZA G, *et al*. Malware family identification with BIRCH clustering[C]. 2017 International Carnahan Conference on Security Technology, Madrid, Spain, 2017: 1–6. doi: 10.1109/CCST.2017.8167802.

[14] PENG Kai, ZHENG Lixin, XU Xiaolong, *et al*. Balanced iterative reducing and clustering using hierarchies with principal component analysis (PBirch) for intrusion detection over big data in mobile cloud environment[C]. International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, Melbourne, Australia, 2018: 166–177. doi: 10.1007/978-3-030-05345-1_14.

[15] ALKASASSBEH M. A novel hybrid method for network anomaly detection based on traffic prediction and change point detection[J]. *Journal of Computer Science*, 2018, 14(2): 153–162. doi: 10.3844/jcssp.2018.153.162.

[16] LORBEER B, KOSAREVA A, DEVA B, *et al*. Variations on the Clustering Algorithm BIRCH[J]. *Big Data Research*, 2018, 11: 44–53. doi: 10.1016/j.bdr.2017.09.002.

[17] GUO Dongwei, CHEN Jingwen, CHEN Yingjie, *et al*. LBIRCH: an improved BIRCH algorithm based on link[C]. The 2018 10th International Conference on Machine Learning and Computing, New York, USA, 2018: 74–78. doi: 10.1145/3195106.3195158.

[18] 尚家泽, 安葳鹏, 郭耀丹. 基于阈值的BIRCH算法改进与分析[J]. 重庆邮电大学学报: 自然科学版, 2020, 32(3): 487–494. doi: 10.3979/j.issn.1673-825X.2020.03.019.
SHANG Jiaze, AN Weipeng, and GUO Yaodan. BIRCH algorithm improvement and analysis based on threshold value[J]. *Journal of Chongqing University of Posts and Telecommunications*: *Natural Science Edition*, 2020, 32(3): 487–494. doi: 10.3979/j.issn.1673-825X.2020.03.019.

[19] 马春来, 单洪, 马涛. 一种基于簇中心点自动选择策略的密度峰值聚类算法[J]. 计算机科学, 2016, 43(7): 255–258, 280. doi: 10.11896/j.issn.1002-137x.2016.7.046.
MA Chunlai, SHAN Hong, and MA Tao. Improved density peaks based clustering algorithm with strategy choosing cluster center automatically[J]. *Computer Science*, 2016, 43(7): 255–258, 280. doi: 10.11896/j.issn.1002-137x.2016.7.046.

[20] CHEN Yuanfang, GUIZANI M, ZHANG Yan, *et al*. When traffic flow prediction and wireless big data analytics meet[J]. *IEEE Network*, 2019, 33(3): 161–167. doi: 10.1109/MNET.2018.1800134.

[21] 熊俊, 何宽, 李颖川, 等. 基于优化FAEMD-OSELM的WSN流量预测算法研究[J]. 仪器仪表学报, 2020, 41(9): 262–270. doi: 10.19650/j.cnki.cjsi.J2006636.
XIONG Jun, HE Kuan, LI Yingchuan, *et al*. Research on WSN traffic prediction algorithm based on optimized FAEMD-OSELM[J]. *Chinese Journal of Scientific Instrument*, 2020, 41(9): 262–270. doi: 10.19650/j.cnki.cjsi.J2006636.

[22] ALMOMANI I, AL-KASASBEH B, and AL-AKHRAS M. WSN-DS: A dataset for intrusion detection systems in wireless sensor networks[J]. *Journal of Sensors*, 2016, 2016: 4731953. doi: 10.1155/2016/4731953.

[23] 何明, 仇功达, 周波, 等. 基于改进密度聚类与模式信息挖掘的异常轨迹识别方法[J]. 通信学报, 2017, 38(12): 21–33. doi: 10.11959/j.issn.1000-436x.2017287.
HE Ming, QIU Gongda, ZHOU Bo, *et al*. Abnormal trajectory detection method based on enhanced density clustering and abnormal information mining[J]. *Journal on Communications*, 2017, 38(12): 21–33. doi: 10.11959/j.issn.1000-436x.2017287.

[24] XIA Hui, FANG Bin, MATTHEW R, *et al*. A basis evolution framework for network traffic anomaly detection[J]. *Computer Networks*, 2018, 135: 15–31. doi: 10.1016/j.comnet.2018.01.025.

郁　滨：男，1964年生，教授，研究方向为信息安全、无线网络技术、视觉密码等.
熊　俊：男，1996年生，硕士生，研究方向为ZigBee、信息安全技术.

责任编辑：马秀强