DOI:10.19651/j.cnki.emt.2415550

# 基于孤立森林评分扩展的流量异常检测方法

沈 萍 陈俊丽
(上海大学通信与信息工程学院 上海 200444)

**摘 要**:流量异常检测是一种有效识别网络攻击行为的技术。近年来,无监督方法在异常检测领域得到了广泛应用。针对现有流量数据间时序关系挖掘的需求与孤立森林随机选择特征属性进行样本划分的问题,本文提出一种基于孤立森林评分扩展的流量异常检测方法。首先,文章使用滑动窗口机制和信息熵特性,设计了网络流量的熵时序特征提取方法,集成至特征集执行显著特征筛选。然后,文章构建了孤立森林评分扩展模型,在节点样本划分时,利用特征集合迭代方法与特征重要性矩阵,综合集合中孤立树特征,为节点标记综合路径长度代替原路径长度,并计算更能表征样本分布的异常评分。最后,通过设定异常得分阈值判别样本是否异常。在公开数据集上的实验结果表明,文章提出的异常检测模型,相比其他方法有明显优势,具有良好的实时检测性能,误报率更低,可有效用于网络流量的异常检测中,对真实网络环境中攻击事件的识别具有重要意义。

**关键词**:熵时序特征;异常检测;无监督;异常评分
**中图分类号**:TP399 **文献标识码**:A **国家标准学科分类代码**:510.50

# Traffic anomaly detection method based on iForest score extension

Shen Ping Chen Junli
(School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China)

**Abstract**: Traffic anomaly detection is a technique used to identify network attacks effectively. In recent years, unsupervised methods have become prevalent in anomaly detection. Aiming at the demand of mining the temporal relationship between existing traffic data and the problem of randomly selecting feature attributes for sample division in iForest, this paper proposed a traffic anomaly detection method based on iForest score extension. Firstly, the paper used the sliding window mechanism and the information entropy property to design an entropic time-series feature extraction method for network traffic, which was integrated into the feature set to perform significant feature screening. Secondly, the paper constructed an iForest score extension model that utilized the feature set iteration method with the feature importance matrix in the node sample division, integrated the isolated tree features in the set, marked the integrated path length between nodes instead of the original path length, and calculated the anomaly score that better characterized the sample distribution. Finally, by setting the anomaly score threshold, the paper discriminated whether the samples were abnormal. The experimental results on the public dataset show that the anomaly detection model proposed in the paper has obvious advantages over other methods, with good real-time detection performance and lower false alarm rate, which can be effectively used in the anomaly detection of network traffic, and is of great significance for the identification of attack events in real network activities.

**Keywords**: entropy time-series feature; iForest; unsupervised; anomaly score

## 0 引 言

随着计算机网络、新型通信等技术的不断发展与更新,网络流量数据呈现出高维度和大数据量的特征,网络环境中因攻击行为和非正常业务行为引起的异常流量逐渐增多。传统基于标签的流量异常检测方法在一定程度上失效,数据标注变得更加困难,缺少标签的情况也在不断加剧。近年来学者们对无监督异常流量检测的关注逐渐增加,展开了广泛的探索和研究。无监督方法通过分析网络流量的统计特征学习网络行为模式,可以很好地实现流量异常检测。无监督异常检测方法主要分为基于密度、基于聚类和基于重构 3 类。

收稿日期:2024-02-28

基于密度的异常检测方法依赖数据，假设正常样本比异常样本有更高的密度，但面临数据维数灾难和密度变化的问题。基于密度方法的基础算法局部离群因子（local outlier factor，LOF）[1]，通过样本点与其局部邻域样本点疏离程度的局部异常度进行异常检测，测量相对密度代替绝对密度，解决了密度变化的问题，但仍受底层 K 近邻算法对密度估计准确性的限制。

基于聚类的异常检测方法假设异常样本不属于任何聚类集群，远离最近的聚类。常用算法基于密度的聚类算法（density-based spatial clustering of applications with noise，DBSCAN）[2]可以对任意形状的稠密数据集进行聚类，但存在邻域半径和最小点数目参数联调难度大的问题，聚类过程中抛弃噪声会导致不适用在安全性或精密性要求高的领域。Wu 等[3]提出的对称邻域关系密度峰值聚类利用 k 近邻和反向 k 近邻为每个数据点建立对称邻域图用于聚合相似的聚类。

基于重构的异常检测方法假设在低维投影下，异常样本由于数量较少难以进行精准的样本重构，常用方法有主成分分析（principal component analysis，PCA）[4]与自编码器（autoencoder，AE）[5]及相关方法[6-7]。其中，PCA 通过降维后的特征向量衡量线性或非线性数据样本在不同方向上的偏离程度。AE 通过训练使解码器重构的数据最大程度的逼近输入数据，使隐藏层学习到输入数据的良好特征表达，但其压缩和解压缩函数是数据相关的、有损的。Said Elsayed 等[6]提出了一种基于 LSTM 的自动编码器学习流量数据的压缩表示。刘宇啸等[7]利用系数自动编码学习正常流量特征，并引入阈值迭代选取最佳阈值以提高模型检测率。

Liu 等[8]研究的孤立森林算法（isolation forest，iForest）不同于上述基于距离或密度来刻画样本间的疏离程度的异常检测方法，作为少量参数和无监督的算法，同时是机器学习中专为异常检测设计，其采用隔离机制，通过孤立样本点检测异常值。凭借其线性时间复杂度、低常数和低内存要求的优势，该方法在异常检测中应用越来越广泛。为解决应用过程中发现的异常样本聚集形成团簇的检测困难，该团队进一步提出改进算法（isolation forest with split-selection criterion，SCiForest）[9]，随机选择一个超平面以在一个节点中分割数据，不足之处在于具有很高的复杂性。进一步地，Hariri 等[10]提出使用随机斜率超平面分割子空间的扩展孤立森林（extended isolation forest，EIF），解决了 iForest 算法对局部异常的不敏感问题，但对高维数据存在局限性，无法确保低维数据也可以输入到高扩展级别的 EIF 中。杭菲璐等[11]提出融合改进的 iForest 和 LOF 检测结果来确定最终分类，其中 iForest 的改进为，在节点分割时，给定足够的随机生成超平面，通过集成学习器构造最终的分离超平面。Chen 等[12]提出一种箱图采样孤立森林，使用箱图过滤后优化采样来训练和构建树，接着在训练集中选择精度更高的隔离树。Xu 等[13]提出深度孤立森林，利用随意初始化的神经网络映射原始数据、随机轴平行切割数据的协同方式促进异常评分，但随机性问题仍存在。Ahlawat 等[14]通过增量算法形成增量孤立森林以快速更新现有孤立森林，缓解异常可能发生在特征空间新区域和新数据会影响现有数据异常程度的现象。

为了避免对数据标签的依赖，同时基于数据分布不均衡的网络流量符合无监督异常检测算法孤立森林能在高维数据中发现散点异常的应用场景，本文提出一种基于孤立森林评分扩展的流量异常检测方法。在考虑到攻击行为通常出现批量探测呈现聚集的特点，首先提出了熵时序特征提取方法，挖掘流量数据间的熵时序关系特征，用于后续节点分割特征选择使用。然后提出了一种孤立森林评分扩展模型，通过引入特征迭代分割与特征重要性矩阵，在特征集划分形成的 iTree 集合下，为样本点计算并赋值综合路径长度完成异常评分。在保持孤立森林线性检测时间的前提下，根据特征重要性计算综合路径长度代替原路径长度，所得出的异常评分能更准确的表征异常样本分布，提高了对异常事件的识别率。

## 1　相关理论

### 1.1　模型框架

本文提出的基于孤立森林评分扩展的流量异常检测方法总体框架如图 1 所示，包括数据处理、流量异常检测和检测评估 3 个模块。

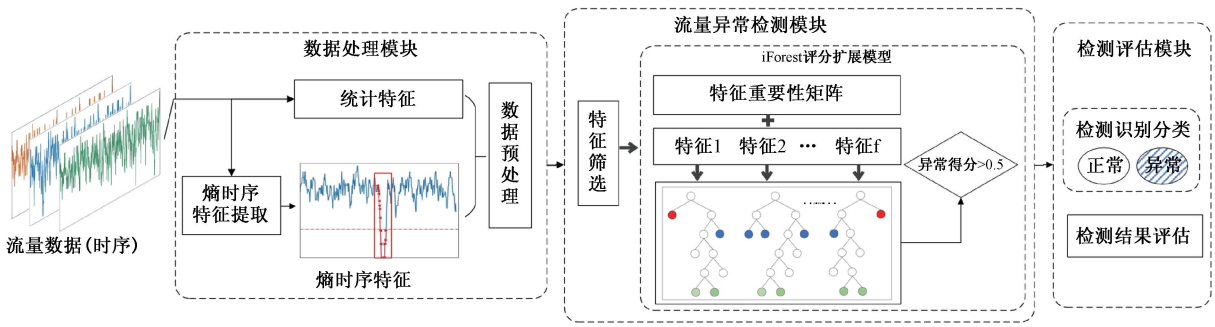


图 1　流量异常检测框架及模块

数据处理模块:挖掘原始流量数据序列之间熵时序特征,提取强关联性的特征以更新原始数据集,再进行数据清洗、转换、归一化的预处理操作,以满足流量异常检测模型的输入要求。

流量异常检测模块:该模块为本文所提方法的核心,由特征筛选和 iForest 评分扩展模型构成,模型在采样集节点分割时使用特征集合迭代方法和特征重要性矩阵,结合初始 iTree 集合下的树结构,获得综合路径长度,用来计算更能表征数据分布特性的异常评分,确定模型最终的无监督异常检测结果。

检测评估模块:利用流量数据集对流量异常检测模块进行测试,得到检测结果的混淆矩阵以计算评估指标值,同时统计模型的检测时间,综合评估模型检测结果与性能。

### 1.2　网络流量熵时序特征提取

网络流量是一个由多个离散信息源构成的随时间变化的信号。在正常且稳定的网络环境下,网络流量呈现出历史相关且平稳的变化趋势,突发和先兆特征不明的异常行为会破坏该状态,网络流量偏离正常行为。异常行为通常表现为在一定时间内对目标网络进行集中攻击形成聚集流量,引发流量激增灾难。

在信息论中,信息熵用于度量随机事件的不确定性,能有效地反映相同属性所对应数据的集中性和分散性,以反映信息源信号发生的可能性。熵具有极值性,当各个信源为状态等概率分布时熵值达到最大值[15]。在大规模网络流量情况下运用该概念,表现为属性数据越聚区域熵值越小,属性数据越散区域熵值越大。

当在正常平稳网络环境中引入异常流量时,流量熵分布将呈现连续下降的波动趋势,与历史正常流量形成明显断层。基于熵值作为时间序列异常样本检测[16]的理论基础,本文提出了一种熵时序特征提取方法,构造过程主要包括 6 个步骤:1)设置滑动窗口宽度 $W$ 和步长 $S$,基于原始时间序列数据完成滑动窗口移动,将数据划分到子序列,生成一组子序列集;2)计算子序列集的流量信息熵;3)构造上异常波动区间:判断流量熵在时间序列上的连续下降区间;4)构造下异常波动区间:利用历史时刻正常流量信息熵最低值作为阈值,判断流量熵由上异常波动区间右边界值反升至阈值的区间;5)建立异常数据波动的完整区间,即上异常区间和下异常波动区间组合;6)将异常波动区间转换为数字序列特征,即熵时序特征。

以一段正常流量时间序列数据为例,得到的信息熵及其分布函数如图 2 所示。图中标注当前环境下正常流量信息熵最小值与子序列 95%分布率时流量熵值。

据图 2 所得,以一段包含异常流量的数据为例,识别异常波动区间,识别结果如图 3 中三角符号标识线所示。

上述熵时序特征提取方法以连续下降与历史正常流量最小信息熵值作为联合条件,识别异常波动区间,避免了单独分析熵下降可能忽略的流量异常区间。把熵时序特征加入原始数据集并使用 SHAP 模型对关键特征进行解释,如图 4 所示。可以发现提取的熵时序特征 entropy 位列顺序靠前。因此认为提取的熵时序特征有效。将该特征加入原始数据集,进行预处理操作后输入流量异常检测模型。

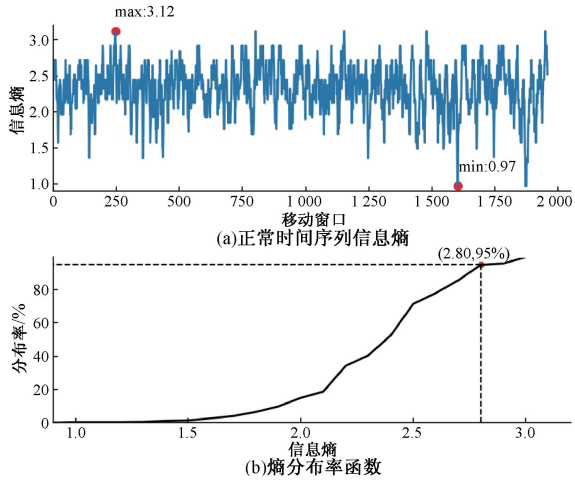


图 2　历史正常流量熵及分布

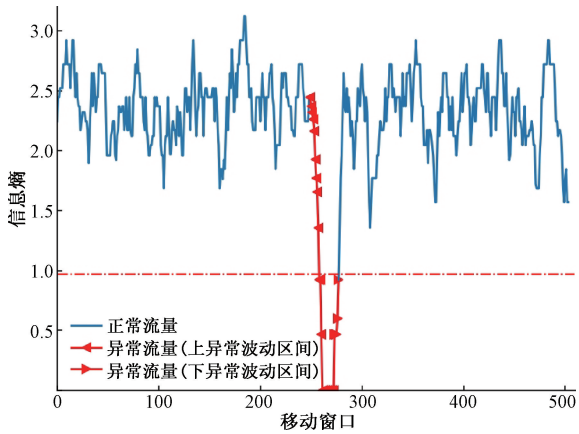


图 3　异常波动区间识别结果

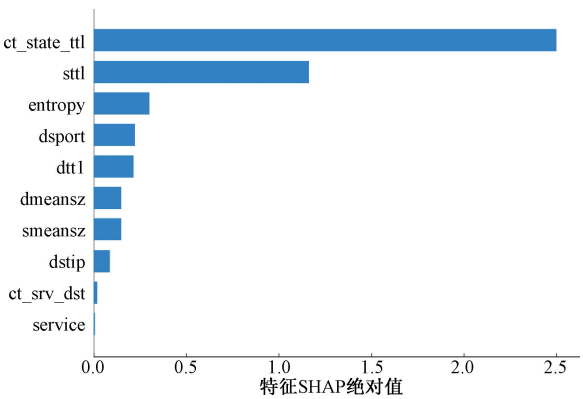


图 4　SHAP 模型特征解释

### 1.3　孤立森林

孤立森林 iForest 是一种大数据背景下的异常检测无监督学习算法，具有不需要计算距离或密度带来的线性执行时间优势，检测性能好。异常流量检测是一个正异常二分类问题，同时也是一个离群点检测[17]的问题。随着网络流量呈海量递增趋势，具备线性时间复杂度和低内存要求的孤立森林算法越来越广泛的运用于流量检测。

iForest 模型由大量的孤立树(iTree)组成，通过 iTree 二叉树搜索结构来孤立样本。二叉树随机选择特征分割，并在该特征的最大值和最小值中随机选择切分值，重复操作不断划分左、右子树，直到满足条件(1)，至此形成树。条件(1)为，下列条件至少满足之一：1)子采样集中只剩一个数据点或者多个相同的数据点，无法进一步划分；2)隔离树的高度达到限定高度。

数据集中的异常数据的样本密度低，在二叉树结构中会较早地被孤立出来，离 iTree 的根节点距离近，正常样本则离 iTree 的根节点较远。$h(x)$是从根节点遍历 iTree 到外部节点的路径长度，即叶节点深度，深度越小，越有可能为异常点。$c(n)$是给定样本数 $n$ 时 $h(x)$的平均值，用来归一化 $h(x)$。异常值分数 $s(x,n)$通过定义为：

$$s(x,n)=2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

其中，

$$c(n)=\begin{cases}2H(n-1)-2(n-1)/n, & n>2\\ 1, & n=2\\ 0, & \text{其他}\end{cases} \tag{2}$$

式中：$H(i)$是通过 ln(i)+0.577 215 664 9(欧拉常数 γ)来估算的调和数。当 $E(h(x))\to c(n)$时，$s\to 0.5$；当 $E(h(x))\to 0$ 时，$s\to 1$；当 $E(h(x))\to n-1$ 时，$s\to 0$。如果 $s\to 1$，那么 $x$ 是绝对异常；如果 $s\to 0$，则 $x$ 被认为是正常数据；如果所有的 $s\to 0.5$，通常整个样本基本正常。因此，设置 0.5 作为异常评分的判断阈值来寻找异常值。

由于 iForest 每次都是随机选取单个特征属性与单个属性值划分样本构建树，但每个样本数据在随机选取的属性上异常程度不同，随机性导致的划分结果不准确。

### 1.4　孤立森林评分扩展模型

根据上述问题，本文在孤立森林算法中引入特征重要性矩阵，构建孤立森林评分扩展模型进行异常检测。与已有改进方法相比，各类方法立足点一致，均为分析随机划分对识别数据集分布特性的影响，提出改进措施，进一步提高模型检测率，但各类方法对孤立树构建机制作用效果存在差异。当前改进方法通常采用新机制划分 iTree 或对 iTree 进行筛选，本文方法为综合特征子集所得到的 iTree 集合，保留 iTree 的独立性，通过特征重要性矩阵确定综合路径长度，重新赋值给样本点，形成全新的孤立森林。

基于 iForest 评分扩展的异常检测方法通过学习特征筛选后的低维深度流量特征，实现异常样本的快速检测。具体实现为：在每个节点切割时，基于预先设定的每棵树训练时参与分裂的树的特征数，选择特征索引来决定在 iTree 上可用于切割的特征集合，特征集合个数定义为 $f$。接着增加算法对不同数据分布的适应性，在每个子采样集上迭代选定的特征，通过迭代样本并尝试不同的切割点，对每一维特征属性的初始路径集合 $set(L)$使用特征重要性矩阵计算样本的综合路径，选择最接近综合路径的初始路径作为划分选择，对这些样本的路径赋值为综合路径。上述步骤重复进行，直到满足 iTree 形成条件。对最终形成的 iTree 计算异常评分。图 5 为本文提出 iForest 评分扩展模型的原理图。

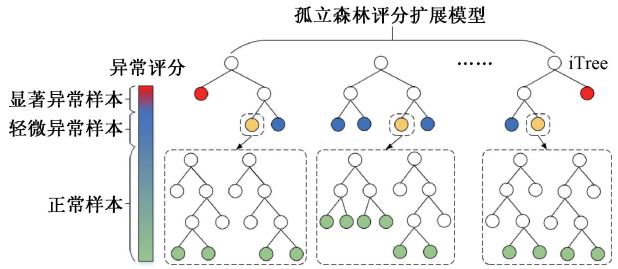


图 5　iForest 评分扩展模型原理图

算法中任一样本点的综合路径计算公式为：

$$L=[L_1\quad L_2\quad\cdots\quad L_f][R_1\quad R_2\quad\cdots\quad R_f]^{\mathrm{T}} \tag{3}$$

式中：$L_i(i=1,2,\cdots,f)$为每个特征集合划分样本得到的路径长度，$R_i(i=1,2,\cdots,f)$为每个特征集合对结果的影响度量参数，满足 $R_1+R_2+\cdots+R_f=1$，共同构成特征重要性矩阵 $\boldsymbol{R}=[\boldsymbol{R}_1\quad\boldsymbol{R}_2\quad\cdots\quad\boldsymbol{R}_f]$。

本文所述特征重要性[18]矩阵使用层次分析法确定。已知系统中各因素之间的关系和层次结构，应用到网络流量数据集中即为特征属性与异常行为的紧密程度与产生异常的严重程度。层次分析法具体实现步骤如下：

1)判断矩阵获取与一致性校验。考虑权威或盲目服从多数的缺陷因素，采用德尔菲法分别得出准则层关于特征属性目标层关于异常评分的判断向量。同时，校验判断矩阵是否满足一致性矩阵的定义。判断矩阵定义为：

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{a}_{11} & \boldsymbol{a}_{12} & \cdots & \boldsymbol{a}_{1f}\\ \boldsymbol{a}_{21} & \boldsymbol{a}_{22} & \cdots & \boldsymbol{a}_{2f}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{a}_{f1} & \boldsymbol{a}_{f2} & \cdots & \boldsymbol{a}_{ff}\end{pmatrix} \tag{4}$$

其中，$a_{ij}$ 表示第 $i$ 个特征属性与第 $j$ 个特征属性的重要程度之比，满足 $a_{ij}=1/a_{ji}$，$a_{ii}=1$。

2)权重矩阵获取。在判断矩阵满足一致性的前提下，使用算术平均法求权重方法，对特征向量进行归一化作为权重向量结果，填入权重矩阵，所得权重向量即构成了特征重要性矩阵 $\boldsymbol{R}=[\boldsymbol{R}_1\quad\boldsymbol{R}_2\quad\cdots\quad\boldsymbol{R}_f]$。

引入特征重要性矩阵计算异常评分的方法在保留孤立森林线性时间复杂度的同时，有效地保护各特征划分下所形成的 iTree 结构特征，消除单一特征划分的不确定性，

虽然在这一过程提高了每一次节点划分的复杂度,但是对每个树结构的度量能更好的捕捉样本分布特点,能更好的适应不同样本的数据分布特性。本文模型中特征用处广泛,所提取的经验证为与网络流量结果显著关联的熵时序特征,有利于模型更准确的正异常样本划分。

## 2　实验与结果分析

### 2.1　数据集

本文在常用的新兴入侵检测数据集 UNSW-NB15[19] 中验证提出方法的有效性。该数据集由正常数据和 9 种攻击数据构成,包含了较新的攻击和入侵手法,生成过程尽量模拟真实网络环境,提出的新特征更能反映当前的真实网络情况。数据集中每条流量由 47 列组成,包含源/目的 $IP$、源/目的端口、协议等 45 个特征列、1 列攻击类型和 1 列标签。本文实验训练数据集抽取 100 000 条数据进行训练。

所使用异常检测算法为无监督方法,故实验中训练集即测试集。同时考虑在无标签情形下评估模型泛化性能,在实验中采用十折交叉验证法对各类模型效果进行比较。

### 2.2　评估指标

实验采用准确率(accuracy)、精确率(precision)、召回率(recall)、误报率(FPR)和 F1 分数(F1-score)对模型的异常检测能力进行评估。评估指标定义如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$F1 - score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{9}$$

### 2.3　实验设计与结果分析

1)模块效果评估

模型效果评估实验旨在评估孤立森林集成各结构模型在数据集上的检测效果,验证算法各模块对整体性能的贡献。有 4 组选项:1)iForest;2)iForest＋熵时序特征;3)iForest 评分扩展;4)本文模型(熵时序特征＋iForest 评分扩展模型)。实验结果表现如图 6 所示。

从图 6 可以看出,相较于 iForest,方法 2 和 3 两种改进方式均在一定程度上提升了流量异常检测精度。引入熵时序特征的模型分割性能更优,说明熵时序特征与异常行为有较高的因果关联性,在孤立森林节点选择特征属性进行分支时提供了更优的组合选择。本文所提模型集成了两种方式的优点,相较于 iForest,准确率由 90.33%提升至 96.32%,精确率由 94.47%提升至 97.69%,误报率由 26.14%降低至 9.11%,验证了本文模型的良好检测性能。

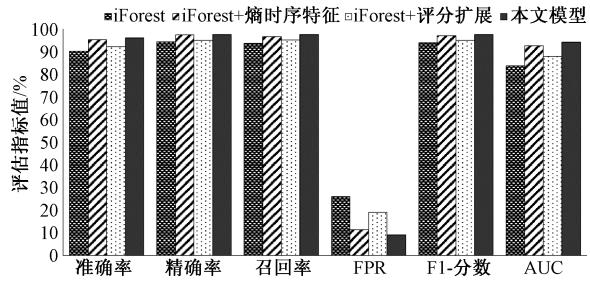


图 6　单一和集成模型检测结果

数据分布存在一定的差异,不同结构能够适用和发现不同分布类型的异常,集成异常检测算法能够更好发挥模型优势以呈现更好的检测效果。进一步地,对不同改进方式下的样本异常得分展开统计,结果如图 7 所示。整体来看,在不同扩展水平下,分布最广泛的样本点基本在异常得分 0.5 左右。在 iForest 模型下,异常分数分布区间最广泛。随着熵时序特征和评分扩展方法的引入,异常样本得分范围更均衡,且正异常样本得分数量峰值分别远离 0.5,表现为识别出的明确正常与异常。

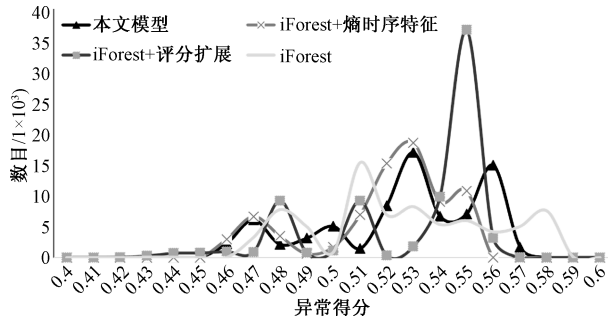


图 7　不同组合方式下异常得分分布密度图

2)模型有效性分析

模型有效性分析为常用无监督异常检测算法与本文模型在数据集上的正异常检测实验。为验证本文方法的有效性,选取了 5 个经典常用无监督检测算法:基于密度的 LOF[1]、基于聚类的 DBSCAN[2]、基于重构的 AE[5]、标准 iForest[8]、iForest 已有改进算法 EIF[10],与本文模型在数据集上进行检测对比。实验结果如表 1 所示。

**表 1　模型有效性分析实验结果**　　%

| 模型 | Accuracy | Precision | Recall | FPR | F1-score |
|---|---|---|---|---|---|
| LOF | 76.84 | 85.87 | 85.91 | 64.57 | 85.89 |
| DBSCAN | 82.12 | 82.26 | 99.72 | 98.46 | 90.15 |
| AE | 72.04 | 82.68 | 83.38 | 79.81 | 83.03 |
| iForest | 90.33 | 94.47 | 93.79 | 26.14 | 94.13 |
| EIF | 95.19 | 95.14 | 99.21 | 23.16 | 97.13 |
| 本文模型 | 96.32 | 97.69 | 97.70 | 9.11 | 97.70 |

从表 1 中可以看出,LOF 和 AE 的异常检测结果较差,DBSCAN 的异常检测结果较这两者,准确率、召回率均

有提升，但异常样本预测正确少导致误报率高，且三者的精确率均不超过 90%。iForest 模型比上述 3 种模型表现优异，准确率、精确率、召回率和 F1-score 均达到 90%。iForest 改进算法 EIF 检测性能较 iForest 有较大提升，准确率与精确率达到 95%以上，本文方法在此基础上各项指标有 1%～2%的提升，表明本文方法对网络攻击具有较好的检测能力，同时误报率为 9.11%，低于其他方法，方法更加可靠。

3)多分类性能分析

在上述是两项实验中，检测方式为正异常的二分类检测，将所有的攻击类型划入异常类，正常数据划入正常类。本节实验为多分类实验。UNSW-NB15 数据集中的攻击类型分为 Analysis、Backdoor、DoS、Exploits、Fuzzers、Generic、Reconnaissance、Shellcode、Worms 9 个类别。实验过程中，分别以某一攻击样本为异常样本，其余作为正常样本，数据训练检测实验结果如图 8 所示。

图 8 中的数据集多分类结果整体取得较好的检测效果，在指定预测值类别下，同类别预测到该类别的数据样本数基本均为占比最大，但检测结果存在较多漏报数据量，异常样本未识别出来预测为正常样本。这是由于部分异常样本占比较小，其中 Analysis、Backdoor、Fuzzers、Reconnaissance、Shellcode、Worms 6 类数据占比均不超过 1%，且该类数据不像其他攻击类别表现为数据集中聚集出现，因此模型对该类数据学习不充分，采样集中划分容易出错，导致有一定的漏报率。数据占比较大的 DoS、Exploits 和 Generic 3 类则表现出良好的检测准确率。

预测值

| 实际值 | Normal | Analysis | Backdoor | DoS | Exploits | Fuzzers | Generic | Reconnaissance | Shellcode | Worms |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | 61 704 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Analysis | 251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Backdoor | 265 | 0 | 6 | 5 | 9 | 1 | 2 | 0 | 0 | 0 |
| DoS | 1 669 | 0 | 0 | 170 | 50 | 9 | 18 | 3 | 0 | 0 |
| Exploits | 2 332 | 0 | 0 | 260 | 460 | 32 | 98 | 50 | 2 | 0 |
| Fuzzers | 652 | 4 | 3 | 26 | 210 | 70 | 27 | 34 | 0 | 0 |
| Generic | 2 398 | 0 | 0 | 24 | 51 | 8 | 5 014 | 6 | 0 | 2 |
| Reconnaissance | 380 | 0 | 1 | 55 | 162 | 1 | 5 | 52 | 0 | 0 |
| Shellcode | 21 | 0 | 0 | 6 | 14 | 1 | 0 | 1 | 0 | 0 |
| Worms | 7 | 0 | 0 | 3 | 7 | 0 | 0 | 0 | 0 | 0 |

图 8　多分类混淆矩阵

4)模型运行时间分析

大数据场景下，异常检测模型在保证了一定异常检测率的同时，同样应确保异常检测速度。因此在实验过程中，对本文方法同其他检测算法的执行耗时进行比较。本文将各无监督模型进行异常检测的总执行时间作为时间指标比较基础。各模型的具体执行时间如表 2 所示。

**表 2　不同模型执行时间**

| 数据比例 | 执行时间/s | | | | | |
|---|---|---|---|---|---|---|
| | LOF | DBSCAN | AE | iForest | EIF | 本文模型 |
| 20% | 6.01 | 18.80 | 6.23 | 1.73 | 1.70 | 1.82 |
| 100% | 119.70 | 131.83 | 76.55 | 9.93 | 11.25 | 10.82 |

在 20%数据集样本下，各模型检测时间相差不大。在 100%数据集样本下，LOF、DBSCAN 和 AE 的时间差异显著，在实际环境下很有可能表现为易受样本数据量影响，无法面对瞬时流量激增场景；EIF 和本文模型同作为 iForest 的改进方法，相比 iForest 模型时间成本增加，但耗时波动小，在可接受范围内，仍表现为线性时间复杂度。同一量级数据下，iForest 和其改进方法在计算时间上没有明显差异。本文模型中 iTree 划分过程中存在一定复杂度增加检测耗时，检测准确率、精确率等更优，说明本文方法在满足实时性的前提下，能够更好的进行流量异常检测。

## 3　结　　论

本文提出了一种基于孤立森林评分扩展的流量异常检测方法，引入了熵时序特征和孤立森林评分扩展概念。该方法首先通过连续下降与历史正常流量最小信息熵值的联合条件判定异常流量上下波动区间，数字化作为熵时序特征。接着，筛选显著强关联性特征作为模型输入，模型在 iForest 基础上扩展，引入迭代特征与特征重要性矩阵的训练，所得出的异常评分对不同数据分布样本有更强的表征能力，更加符合真实网络环境下的流量异常检测。该方法在保持 iForest 线性时间复杂度的前提下，还提高了大规模网络流量检测的精确度，具有更好的异常检测性能。今后工作中考虑动态调整的特征重要性矩阵对不同数据集合进行综合路径计算，使得参数更加合理化。同时，根据相关文献提出未来待检测数据输入后对当前数据异常程度的影响，考虑在稳定的网络环境下，流量数据通常会遵循历史流量数据规范，未来非攻击数据的加入不会对历史数据的异常程度产生显著影响，而未来攻击数据将表现为显著异常，但仍是未来可以思考的方向。

## 参考文献

[1]　BREUNIG M M，KRIEGEL H P，NG R T，et al. LOF：identifying density-based local outliers[C].

Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,2000:93-104.

[2] SCHUBERT E, SANDER J, ESTER M, et al. DBSCAN revisited, revisited: Why and how you should(still) use DBSCAN[J]. ACM Transactions on Database Systems(TODS), 2017, 42(3):1-21.

[3] WU C, LEE J, ISOKAWA T, et al. Efficient clustering method based on density peaks with symmetric neighborhood relationship [J]. IEEE Access,2019,7: 60684-60696.

[4] JOLLIFFE I T, CADIMA J. Principal component analysis: A review and recent developments [J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016,374(2065): 20150202.

[5] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]. Proceedings of the 25th International Conference on Machine Learning, 2008: 1096-1103.

[6] SAID ELSAYED M, LE-KHAC N A, DEV S, et al. Network anomaly detection using LSTM based autoencoder [C]. Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks,2020:37-45.

[7] 刘宇啸,陈伟,张天月,等.基于稀疏自动编码器的可解释性异常流量检测[J].信息网络安全,2023,23(7):74-85.

[8] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]. 2008 Eighth IEEE International Conference on Data Mining,2008: 413-422.

[9] LIU F T, TING K M, ZHOU Z H. On detecting clustered anomalies using SCiForest [C]. Machine Learning and Knowledge Discovery in Databases: European Conference,2010: 274-290.

[10] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest[J]. IEEE Transactions on Knowledge and Data Engineering,2021,33(4): 1479-1489.

[11] 杭菲璐,郭威,陈何雄,等.基于 iForest 和 LOF 的流量异常检测[J].计算机应用研究,2022,39(10): 3119-3123.

[12] CHEN J, ZHANG J, QIAN R, et al. An anomaly detection method for wireless sensor networks based on the improved isolation forest[J]. Applied Sciences, 2023,13(2):702.

[13] XU H, PANG G, WANG Y, et al. Deep isolation forest for anomaly detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12591-12604.

[14] AHLAWAT N, AWEKAR A. Incremental isolation forest to handle concept drift in anomaly detection [C]. Proceedings of the 7th Joint International Conference on Data Science & Management of Data(11th ACM IKDD CODS and 29th COMAD),2024:582-583.

[15] 江魁,丘远东,郑浩城.基于信息熵与 LSTM 的 ICMPv6 DDoS 攻击检测方法[J].计算机工程与应用,2021,57(21):148-154.

[16] 银鹰,周志洪,姚立红.基于 LSTM 的 CAN 入侵检测模型研究[J].信息网络安全,2022,22(12):57-66.

[17] 夏志祥,李准,徐伟.大气电场测量数据的异常检测及校正方法研究[J].电子测量技术,2023,46(1):90-96.

[18] 吴冬,阎卫东,王井利.基于特征重要性加权的随机森林点云分类研究[J].电子测量技术,2023,46(20):120-127.

[19] MOUSTAFA N, SLAY J. UNSW-NB15: A comprehensive data set for network intrusion detection systems(UNSW-NB15 network data set)[C]. 2015 Military Communications and Information Systems Conference(MilCIS),2015: 1-6.

## 作者简介

**沈萍**,硕士,主要研究方向为网络安全、异常流量检测。
E-mail:shen_ping@shu.edu.cn

**陈俊丽**(通信作者),博士,副教授,主要研究方向为网络信息安全、智能信号与信息处理。
E-mail:jlchen@shu.edu.cn