DOI: 10.13382/j.jemi.B2306893

基于特征耦合泛化的流量异常检测方法*

陈万志[1] 张国满[1] 王天元[2]
(1.辽宁工程技术大学软件学院 葫芦岛 125105;2.国网辽宁省电力有限公司营口供电公司 营口 115005)

摘 要:针对现有流量异常检测模型中稀疏特征易被特征选择算法忽略的问题,提出一种基于特征耦合泛化(FCG)的流量异常检测方法。首先,采用 DBSCAN 密度聚类算法去除数据中的离群点,降低异常点对后续 FCG 算法的影响。其次,使用最大相关最小冗余(mRMR)算法对数据特征进行排序,选择对分类最具影响力的特征生成 FCG 算法中的类别区分特征(CDF),以增强分类能力。利用 K 最近邻(KNN)算法填补 CDF 中的缺失值,保持数据完整性。然后,将数据按照攻击类别分组,分别使用 mRMR 算法对特征进行排序,挑选每种攻击类别数据中具有实例区分能力的稀疏特征作为 FCG 算法中的实例区分特征(EDF)。利用两种特征在异常检测数据中的耦合程度和 EDF 的上层概念将 EDF 转化成更泛化的特征。最后,将经过处理的数据输入基于贝叶斯优化(Bayesian optimization, BO)参数的随机森林(RF)模型进行分类识别。通过在 NSL-KDD 数据集上进行仿真实验,准确率达到了 91.79%,验证了所提方法具有较好的检测性能。

关键词:异常检测;离群点检测;特征耦合泛化;特征选择
中图分类号:TP393; TN911.7 **文献标识码**:A **国家标准学科分类代码**:510.40

Traffic anomaly detection method based on feature coupling generalization

Chen Wanzhi[1] Zhang Guoman[1] Wang Tianyuan[2]
(1. College of Software, Liaoning Technical University, Huludao 125105, China; 2. State Grid Yingkou Electric Power Company of Liaoning Electric Power Supply Co.,Ltd., Yingkou 115005, China)

Abstract: Considering the problem that the sparse features in the existing traffic anomaly detection models are easily ignored by the feature selection algorithms, a traffic anomaly detection method based on feature coupling generalization (FCG) was proposed. First, the DBSCAN density clustering algorithm was used to remove outliers in the data to reduce the impact of the anomalies on the subsequent FCG algorithm. Second, the minimal-redundancy-maximal-relevance (mRMR) algorithm was used to sort the data features, and the most influential features for classification were selected to generate the class-distinguishing features (CDF) in the FCG algorithm, in order to enhance the classification ability. The K-nearest neighbors (KNN) algorithm was used to fill in the missing values in CDF to maintain data integrity. Then, the data were grouped according to attack categories, and the features were sorted using the mRMR algorithm respectively, and the sparse features with instance-distinguishing ability in the data of each attack category were selected as the example-distinguishing feature (EDF) in the FCG algorithm. The degree of coupling between the two features in the anomaly detection data and the upper concept of EDF were used to transform EDF into more generalized features. Finally, the processed data were fed into the random forest (RF) model based on Bayesian optimization (BO) parameters for classification and identification. Through simulation experiments on the NSL-KDD dataset, the accuracy reached 91.79%, which verifies the proposed method has a good detection performance.

Keywords: anomaly detection; outlier detection; feature coupling generalization; feature selection

收稿日期:2023-09-15 Received Date: 2023-09-15
*基金项目:国家重点研发计划(2018YFB1403303)、辽宁省教育厅高校科研基金(2021LJKZ0327)项目资助

0　引　言

伴随互联网的不断发展,网络服务已经广泛渗透到各个领域。新用户和新设备不断加入网络,人们享受网络带来的便利,但也随之面临着更多风险[1]。入侵检测已经成为解决网络攻击问题的有效方式之一,不仅可以增强系统的安全管理能力,还能提高信息安全基础结构的完整性[2]。通过收集网络系统中的数据并进行分析,入侵检测系统采用误用检测或异常检测的方式[3],能够及时识别网络中潜在的威胁和入侵迹象,并发出警报或采取相应的防御措施。

随着机器学习和深度学习模型的普及,越来越多基于机器学习和深度学习的流量异常检测模型被提出,有效缓解了传统入侵检测系统的瓶颈问题[4]。机器学习和深度学习可以把异常检测简化为识别和分类问题,智能化的实现网络安全维护。然而,一些异常检测系统过于强调算法模型自身的识别和分类性能,而忽视了前期特征工程对后期模型性能的影响[5-6]。特征生成技术在特征工程中扮演着非常重要的角色,通常情况下,那些在分类过程中起到重要作用的高频特征自然而然会被大多数特征选择方法所保留,而那些在分类过程中影响较小的特征会被特征选择方法忽略,其中就包括稀疏特征。然而,这些稀疏特征很可能蕴含着重要的类别区分信息,因此,研究如何利用这些稀疏特征的有用信息生成新的高质量特征变得至关重要。为了更好的改善异常检测模型的性能,多种技术被应用到异常检测研究中,例如特征生成技术、特征选择技术以及分类算法。

特征生成技术是通过对原始数据进行一系列变换和处理,以创建新的、更有信息量的特征的方法。特征生成技术的多样性为各个领域带来了丰富的研究成果。Li 等[7]提出一种特征耦合泛化(feature coupling generalization, FCG)的特征生成方法,提升了稀疏特征在分类过程中的作用,在多个文本挖掘任务中取得了较好的结果,但其实验数据中特征多为布尔特征,适用性较差。何林娜等[8]在药名实体识别方面应用 FCG 半监督学习方法生成药名字典,得到了 76.73 的 F 值。目前这种方法尚未在异常检测领域得到应用。Niu 等[9]提出一种基于多粒度特征生成方法,在多个异常检测数据集上取得了较好的效果。但随着数据集中特征的增加,可能导致生成更多特征和训练时间的增加,因此引入特征选择技术也是异常检测系统中提高效率的一种解决方案。

特征选择技术旨在从原始特征中选择出少量最具信息量和对模型性能有贡献的特征子集,从而减少维度、降低计算成本、避免过拟合并提高模型的解释性。主要分为 3 种:过滤式、包裹式和嵌入式[10]。其中,Kusumaputri 等[11]在 NSL-KDD 数据集上进行特征选择,分析了多种分类器的性能,为异常检测模型中分类器的选择提供了有价值的参考。Venkatesan[12]和 Leo[13]基于机器学习算法做特征选择,设计了流量异常检测系统,得出了基于 RF 特征选择的异常检测系统效果最佳的结论。杨红浩等[14]采用信息增益率的混合异常检测模型有效解决了仅定性选取特征带来的检测精度较低的问题。

分类算法用于将数据样本分配到预定义的类别或标签中。算法的好坏决定最终的检测结果是否理想,任家东等[15]提出的一种新的混合多层次检测方法结合了 K 最近邻(K-nearest neighbors, KNN)离群点检测和随机森林(random forest, RF),准确率和检测率明显优于其他算法,并且能有效的检测多种攻击类型。然而,该方法未考虑到数据集少数类样本信息不足的问题。胡姣姣等[16]提出一种基于深度卷积神经网络(deep convolutional neural network, DCNN)分类算法的检测方法,弥补了由于忽略数据分布的偏斜性而造成的少数类检测精度低的缺点。梁欣怡等[17]提出一种基于自监督特征增强的卷积神经和双向长短期记忆(convolutional neural network and bidirectional long short term memory, CNN-BiLSTM)网络入侵检测方法,解决了检测中攻击样本和流量特征不足的问题。刘新倩等[18]提出一种基于流量异常分析的 RF 检测方法,在多维度优化下实现了高准确率、低误报率以及有效召回率。

在各种以机器学习为核心的流量异常检测模型中,基于 RF 的实现在检测效果、误报率以及泛化能力上都呈现出一定的优势,并且 RF 自身具备特征选择的能力。因此许多研究都选择使用 RF 模型,或将其与其他模型结合,共同实现流量异常检测。基于上述文献中存在的一些问题,本文提出一种新的流量异常检测模型,基于 FCG 的 RF 流量异常检测方法。其综合考虑基于规则与机器学习两种分类方法的优势,旨在生成具备更好类别区分能力和更强泛化能力的新特征,从而提升模型的检测性能。在实验中,本文提出的异常检测方法相对于基于原始数据特征的检测结果表现出更强的检测能力和更低的误报率。这一成果表明,通过引入 FCG 算法,可以提升异常检测模型的效果,更好的满足实际应用的需求。

本文的贡献如下:

1)提出一种基于 FCG 的流量异常检测方法,利用异常检测数据中的原始特征生成更加有用的新特征。借鉴本体的思想对特征进行组合,以更好地发挥稀疏特征在训练过程中的作用。基于数据特征之间的共现信息,创造新的特征训练分类器,提高了模型的泛化能力。

2)基于原有 FCG 算法选择的类别区分特征(class-distinguishing feature, CDF)存在类别区分能力不足的问题,使用多个强类别区分能力特征依靠规则生成 CDF,并

采用KNN算法填补CDF中的缺失值。通过将两种方法相结合,生成的CDF能够综合两者的优势,具备出色的泛化性能。

1　FCG算法

FCG是一种利用数据中特征之间共现信息转化生成新特征的半监督学习算法,FCG需要两种特征来实现新特征的生成。一种是需要被丰富的能区分实例的低频特征,另一种是用来丰富第1类特征的类别区分能力较强的高频特征。将第1种特征定义为实例区分特征(example-distinguishing feature, EDF),第2种特征定义为CDF。直觉上,这两种特征在数据中的共现信息在一定程度上反映了包含第1种特征的实例的类别倾向。从以上的定义中可以看出EDF的作用为区分不同的实例,即起作用的EDF必然在当前实例中出现,EDF不一定能完全区分出当前实例(即当且仅当该特征出现时才能判断是当前实例),但应该具有较强的区分实例的能力。CDF的作用是区分当前实例的类别,也不一定需要完全区分出当前实例的类别,但应具有较强的区分实例类别的能力。利用这两种特征的共现信息生成的新特征会对判断实例的类别有较大的贡献。由于EDF选取的时候偏向于低频特征,即稀疏特征,导致利用EDF与CDF的共现信息生成的特征也会和EDF一样稀疏。为了解决这一问题,借助本体的思想,利用EDF的上层概念将一组EDF结合起来解决此问题,使用针对同种类别的EDF组来得到更充足的信息。利用组内每个EDF和CDF的共现信息的加和生成新的特征,这样就有效解决了新特征稀疏问题。FCG算法过程如下:

输入:样本集合为 $D=\{d_1,d_2,\cdots,d_n\}$,特征集合为 $F=\{f_1,f_2,\cdots,f_m\}$,基于特征集合下的数据向量表示为 $X=\{x_1,x_2,\cdots,x_m\}\in R^m$,数据集为 U。

输出:新的特征集合 G。

1)从特征集合中选择具有相同上层概念的子集 E 作为EDF集合。

2)确立映射关系 $root(e):E\rightarrow R$,将 E 中的元素根据它们的相似性映射到高级概念集合 R。

3)从特征集合 F 中选择具有较强类别区分能力的子集 C 作为CDF集合。

4)定义特征耦合度(feature coupling degree, FCD)类型 T 来度量每个EDF和CDF的耦合程度。

5)根据数据集合 U 和FCD类型 T,计算每个EDF和CDF之间的FCD值。

6)将集合 $G=R\times C\times T$ 作为新特征集合,使得每个新特征对应一个三元组 (r,c,t),其中 $r\in R,c\in C,t\in T$。

7)生成新的特征列,特征值计算公式为:

$$x_i=\sum_{e\in E\&root(e)=r} BFeature(e,d)*FCD(U,e,c) \tag{1}$$

将数据集中特征定义为一组规则的集合,记作 $L=\{l_1,l_2,\cdots,l_m\}$,将数据集中每个特征作为集合的一个元素 $l_i\in L$。用 $D=\{d_1,d_2,\cdots,d_n\}$ 表示样本的集合,其中每个元素 $d_i\in D$ 表示一个样本。在机器学习中,数据集中的样本元素 d 被标识成一个 m 维向量的形式,记作 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_m\}\in R^m$,向量 $\boldsymbol{x}$ 中的每个元素 $x_i(i=1,2,\cdots,m)$ 对应一个规则 l_i,l_i 和样本 d 一起确定 x_i 的值。如果特征的取值范围仅在两个值之间("是"或"否"),将得到在机器学习领域中常见的布尔特征,通常用0和1表示,取值计算公式为:

$$x_i=BFeature(d,l_i)=\begin{cases}1 & d\in l_i\\ 0 & d\notin l_i\end{cases} \tag{2}$$

2　基于FCG的异常检测方法

本文提出的基于FCG的RF流量异常检测方法,包括对数据的前期预处理、FCG算法的实现以及流量异常检测3个部分。实验过程中使用NSL-KDD数据集进行训练和测试,具体框架如图1所示。

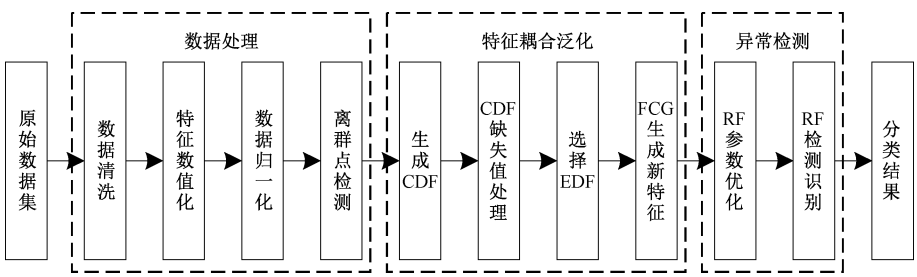

图1　基于FCG的流量异常检测方法框架

Fig. 1　Traffic anomaly detection method framework based on FCG

2.1　数据处理

首先,对 NSL-KDD 数据集中数据特征进行一定的处理,具体包括数据清洗、特征数值化、数据归一化以及离群点检测。

1)数据清洗

在数据集中,经常会出现取值唯一的特征。这些特征对训练分类器影响不大,还会消耗一定的资源,因此需要提前剔除。

2)特征数值化

为了方便处理,需要将数据集中存在的非数值型特征转化为数值型特征。本文采用整数标签编码的方式实现特征数值化。

3)数据归一化

数据的差异可能会对某些算法产生不良影响,数据归一化把所有的数据映射到指定区间内,消除不同特征之间的量纲差异,提高算法性能。本文采用 min-max 方法进行数据归一化,其表达式为:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

其中,x 是原始值,$x_{\min}$ 是该特征的最小值,而 $x_{\max}$ 是特征的最大值。

4)离群点检测

离群点是数据集中与其他数据点明显不同的点,具有与大多数同类别数据点不同的特征或行为。为了降低异常数据点对后期 FCG 方法中 CDF 生成的干扰以及 EDF 与 CDF 耦合程度计算的影响,使用具有噪声的基于密度的聚类算法(density-based spatial clustering of applications with noise, DBSCAN)进行离群点检测和剔除。由于数据集中 root 权限攻击(user-to-root, U2R)和远程用户攻击(remote-to-local, R2L)两种攻击类别中样本数量过少,因此对这两类数据不进行离群点去除,避免训练样本不足导致模型训练不充分的问题。超参数对于离群点检测算法的性能有很大影响,为了得到较优的 DBSCAN 算法超参数,考虑不同参数组合的算法性能,选择性能最优的参数组合作为最终实验参数,在此过程中使用轮廓系数(silhouette coefficient, SC)作为衡量参数组合优劣的指标,其中 DBSCAN 算法通过 Python 中的 scikit-learn 库来实现。

2.2　FCG 算法实现

1)生成 CDF

为了生成 CDF,首先将目标特征定为布尔特征,然后按照以下步骤进行 CDF 生成:

输入:经过处理的数据样本集合$\{x_1, x_2, \cdots, x_n\}$。

输出:生成的 CDF 特征。

步骤(1)将数据中的标签列设置为两种类别取值,其中正常数据为 1,异常数据为 0。

步骤(2)使用特征选择算法[19]最大相关最小冗余(minimal-redundancy-maximal-relevance, mRMR)对数据中的特征进行选择排序。

步骤(3)选择重要性排序第 1 的特征进行分析。

步骤(4)若是离散特征,利用其特征取值将数据集划分成子数据集,执行步骤(6)。

步骤(5)若是连续特征,通过 1R 离散化方法生成离散区间,利用离散区间将数据集划分成子数据集,执行步骤(6)。

步骤(6)计算划分之后每个子数据集的基尼系数(Gini impurity, Gini),若有子数据集 $Gini = 0$,生成子数据集中样本的 CDF 值,选择下一个特征,执行步骤(4);若没有 $Gini = 0$ 的子数据集,执行步骤(7)。Gini 的计算公式为:

$$Gini(t) = 1 - \sum_{i=1}^{K} p_t^{i2} \tag{4}$$

其中,t 代表子数据集节点,K 表示节点 t 上有 K 种不同的类别,p_t^i 表示每个类别在节点 t 上的样本占比,i 取值为 $1 \sim K$。

步骤(7)生成的 CDF 中若存在缺失值,使用 KNN 算法进行填充。

其中,1R 连续特征离散化方法是一种有监督的离散化方法,具体实现如下所示:

输入:数据集中连续特征与其对应的类标签。

输出:连续特征的离散区间。

步骤(1)从第 1 个样本开始,将前 6 个样本放入第 1 个区间。

步骤(2)对于下一个样本,如果其类标签与当前区间中 1/2 以上实例的类标签相同,将其加入当前区间。

步骤(3)如果下一个样本的类标签与当前区间中 1/2 以上样本的类标签不同,形成下一个区间,放入该样本,继续判断下一个样本的加入。

步骤(4)重复上述步骤,直到所有样本都被放入相应的区间。

步骤(5)对每个区间,选择其中样本数量最多的类标签作为该区间的类标签。

步骤(6)最后,检查相邻的区间,如果它们经过上述操作后具有相同的类标签,合并这些相邻区间。

将生成的 CDF 设置为布尔特征有以下优点:首先,通过将正常数据和异常数据进行区分,可以有效降低样本数据的误报率和漏报率。这样划分有助于提高模型的整体准确率,从而更有效地识别和分类异常数据;其次,不对异常数据的类别进行细分,可以增加模型在检测不规范数据样本时的容错性,在某些情况下,两个异常数据样本可能不属于同一种攻击类别,如果在生成 CDF 阶段

过于细致划分,可能会导致类别判断错误,而在生成新特征的过程中,通过 EDF 和 CDF 的共现信息进行更准确的划分,往往可以获得更理想的结果;最后,不细分异常数据类别有助于避免造成稀疏特征问题,当 CDF 取值较多时,EDF 和其对应的 CDF 取值之间的共现信息会减少,会导致生成稀疏特征。不进行细致划分,可以更充分的发挥稀疏特征的类别区分能力,从而提高模型的检测率。

2) CDF 缺失值处理

利用原始特征生成 CDF 的过程中,在有些样本中正常数据和异常数据难以仅通过单个重要特征取值区分开,在这种情况下,这些样本中要生成的 CDF 的值被视为缺失值。为了处理这些缺失值,采用适当的缺失值填充技术是必要的。目前基于机器学习的缺失值处理技术已取得了显著效果[20],因此本文使用一种改进的 KNN 算法实现缺失值填充。KNN 算法在样本不平衡时可能会导致错误的结果,在这种情况下,当某个类别样本数量远大于其他类时,输入一个样本后,其 k 个最近邻中可能有大量来自于数量较大类的样本,导致错误的分类。为了解决这个问题,一种改进方法是引入权重。反比例函数是一种常用的权重函数,其公式为:

$$w(i)=\frac{1}{d(x,x_i)^p} \tag{5}$$

其中,$w(i)$是与第 i 个邻居相关的权重,$d(x,x_i)$是输入样本 x 与邻居 x_i 之间的距离,p 是一个可调节的参数。通过引入这种权重函数,可以减轻大样本类对结果的影响,从而更准确地填充缺失值。这种方法有个潜在的问题,即为近的样本给予了很大权重,而稍远的样本权重会衰减的很快,这会使算法对噪声数据变得更加敏感,高斯函数可以克服这个缺点,其公式为:

$$w(i)=e^{-\frac{d(x,x_i)^2}{2\sigma^2}} \tag{6}$$

其中,σ 是控制高斯分布形状的参数。与之前的权重函数不同,高斯函数在距离较远的样本上也会有一定的权重,能够更好地处理噪声数据,提高算法的鲁棒性,因此选择高斯函数作为 KNN 权重函数。

3) 选择 EDF

针对每种攻击类别的数据进行特征排序,并分析特征的取值情况,在此基础上,挑选一组特征作为同种攻击类别的 EDF,这些特征应具备以下特点:稀疏性、强实例区分能力。通常情况下,高频特征经过分类器训练后已能发挥自身类别区分能力,无需进一步加强学习。因此,EDF 的选择侧重于从数据集中选取一组稀疏特征,这些特征容易被特征选择算法所忽略,但本身很可能包含较强的类别区分信息,将这些特征组合在一起,以同种攻击类别这一概念层次关系为指引,既可以解决数据稀疏问题,又能更好地激发它们的分类能力。

在 NSL-KDD 数据集中,采用 mRMR 算法计算出各个攻击类别的重要性特征。首先,将数据集中数据划分为 4 组,分别为 group_1 = {Normal, Dos}, group_2 = {Normal, Probe}, group_3 = {Normal, U2R}, group_4 = {Normal, R2L}。然后,分别使用 mRMR 算法计算 4 种攻击类型的重要性特征,最后,从每个攻击类型中挑选出一些特征,作为 FCG 方法中每个类别所需的 EDF[21]。例如,针对 U2R 攻击类别,经过特征重要性排序分析之后可以选择特征 root_shell、num_root 以及 num_shells 作为一组 EDF。这 3 个特征都与 U2R 攻击相关,并且几个特征都是稀疏特征,具备良好的实例区分能力,根据 EDF 的上层概念将其设置为一组。通过将它们结合在一起生成特征,可以解决生成特征时出现的数据稀疏问题。

4) FCG 生成新特征

通过选择的 EDF 与生成的 CDF,可以创建每个攻击类别的新特征。首先,利用每个攻击类别的 EDF 组中的每个特征与 CDF 的共现信息,在数据集中计算两种特征之间的 FCD,然后,将 EDF 组中的每个特征与 CDF 的特征耦合度相加,生成对应类别的新特征,生成的这些特征增加了整个数据集合中两类特征之间的相关程度,具备一定的泛化能力。

FCD 类型是 FCG 算法的重要组成部分,它是通过对经典的事件耦合度计算方法进行改进而得到的通用方法。在这个过程中,频率使用了对数运算,b 是平滑因子,设置为 1, $co(x,y)$ 是两个特征 x 和 y 同时出现的次数, $count(x)$ 是特征 x 出现的次数,而 $count(y)$ 是特征 y 出现的次数。

$$FCD=\frac{\log(co(x,y)+b)}{\log(count(x)+b)\log(count(y)+b)} \tag{7}$$

2.3　RF 异常检测

本文选择 RF 模型作为流量异常检测分类器,它是一种集成学习方法,用于解决分类和回归问题。RF 的主要思想是通过组合多个模型的预测,以提高整体模型的稳定性和准确性。它是由多个决策树(Decision Tree, DT)构成的集合,通过对 DT 的预测结果进行整合来产生最终的预测结果,具备良好的泛化性。

RF 算法通过引入两个随机性来增加模型的多样性,从而提高整体性能。自主采样:RF 中的每棵 DT 都是基于一个随机抽样的数据子集进行训练的。特征选择:在构建每棵 DT 的过程中,从所有特征中随机选择一个特征子集。通过这两种随机性,RF 中的每棵 DT 都在不同的数据子集和特征子集上进行训练,这使得每棵 DT 都具有不同的观察和判断,减少了过拟合风险,并在集成预测时提供了多样性。本文使用 Python 中的 scikit-learn 库来实现 RF 算法,并通过使用 BayesianOptimization 库来对 RF 算法的参数进行调优。

3　实验与分析

本文仿真实验基于 Windows 10 操作系统实现;硬件配置:Intel Core i5-7200U 3.40 GHz CPU、16 GB RAM;编程语言:Python3.8。

3.1　训练集与测试集

仿真实验采用 NSL-KDD 数据集,该数据集是基于 KDD Cup 1999 数据集的改进版本,对其进行了一些优化。数据集总共包括 41 个特征、1 个类别标签以及能被正确分类的分类器个数。在这 41 个特征中,有 7 个是离散型特征,分别是第 2、3、4、7、12、14、22 个特征,其余的特征都是连续型特征,在离散特征中,第 2、3 和 4 个特征是字符型特征,而其他的则是数值型特征。这些特征可以被划分为 4 类:TCP 连接基本特征、TCP 连接内容特征、基于时间(2 s 内与当前连接)的网络流量统计特征以及基于主机(前 100 个连接中与当前连接具有相同目标主机)的网络流量统计特征,具体特征描述如表 1~4 所示。数据集中除正常数据外被分为 4 个攻击大类:Dos、Probe、U2R 和 R2L。在该数据集中,训练数据集包含 125 973 条记录,测试数据集包含 22 544 条记录,具体样本数量如表 5 所示。第 43 列为样本能被正确分类的分类器数量,这对于分类结果会有所影响,在实验前直接剔除。

表 1　TCP 连接基本特征

Table 1　Basic characterization of TCP connections

序号	特征名称	描述
1	duration	连接持续时间
2	protocol_type	网络连接协议类型
3	service	目标主机的网络服务类型
4	flag	网络连接正常或错误的状态
5	src_bypes	源主机到目标主机的字节数
6	dst_bytes	目标主机到源主机的字节数
7	land	连接是否同一个主机/端口
8	wrong_fragment	一次连接中错误分段的数量
9	urgent	一次连接中加急包的个数

表 2　TCP 连接内容特征

Table 2　TCP connection Content characterization

序号	特征名称	描述
10	hot	访问系统敏感文件和目录的次数
11	num_failed_logins	一次连接中登录失败的次数
12	logged_in	是否成功登录
13	num_compromised	Compromised 条件出现的次数
14	root_shell	是否获得超级用户权限
15	su_attempted	是否出现“su_root”命令
16	num_root	一次连接中,root 用户操作次数
17	num_file_creations	一次连接中,文件创建操作的次数
18	num_shells	一次连接中,使用 shell 命令的次数
19	num_access_files	一次连接中,访问控制文件的次数
20	num_outbound_cmds	一次连接中,ftp 会话出站命令的次数
21	is_hot_login	登录是否属于热登陆列表
22	is_guest_login	是否是 guest 登录

表 3　基于时间的网络流量统计特征

Table 3　Time-based network traffic statistics characterization

序号	特征名称	描述
23	count	具有相同目标主机的连接数
24	srv_count	具有相同服务的连接数
25	serror_rate	具有相同目标主机出现 SYN 错误的连接的百分比
26	srv_serror_rate	具有相同服务出现 SYN 错误的连接的百分比
27	rerror_rate	具有相同目标主机出现 REJ 错误的连接的百分比
28	srv_rerror_rate	具有相同服务出现 REJ 错误的连接的百分比
29	same_srv_rate	具有相同目标主机相同服务的连接的百分比
30	diff_srv_rate	具有相同目标主机不同服务的连接的百分比
31	srv_diff_host_rate	具有相同服务不同目标主机的连接的百分比

表 4　基于主机的网络流量统计特征

Table 4　Host-based network traffic statistics characterization

序号	特征名称	描述
32	dst_host_count	连接数
33	dst_host_srv_count	具有相同服务的连接数
34	dst_host_same_srv_rate	具有相同服务的连接所占百分比
35	dst_host_diff_srv_rate	具有不同服务的连接所占百分比
36	dst_host_same_src_port_rate	具有相同源端口的连接所占百分比
37	dst_host_srv_diff_host_rate	具有相同服务不同源主机连接所占百分比
38	dst_host_serror_rate	出现 SYN 错误的连接所占百分比
39	dst_host_srv_serror_rate	具有相同服务出现 SYN 错误所占百分比
40	dst_host_rerror_rate	出现 REJ 错误的连接所占百分比
41	dst_host_srv_rerror_rate	具有相同服务出现 REJ 错误所占百分比

表 5　NSL-KDD 实验数据信息

Table 5　NSL-KDD experimental data information

样本类别	训练样本/条	测试样本/条
Normal	67 343	9 711
Dos	45 927	7 458
Probe	11 656	2 421
R2L	995	2 754
U2R	52	200

在整个数据集中,num_outbound_cmds 特征的值都为 0,is_hot_login 特征在整个数据集中只有 1 个样本的值为 1,其余为 0,直接剔除。将特征 protocol_type、service 和

flag 转化为数值型特征,以 protocol_type 为例,该特征包含 3 个不同取值:TCP、UDP 和 ICMP。将这 3 种取值用数值表示,即 1 表示 TCP,2 表示 UDP 以及 3 表示 ICMP。

3.2　试验评估指标

流量异常检测方法涉及许多可利用的指标,为了评估所提方法的有效性,选择其中 4 种主要的评估指标作为实验过程中的衡量指标,分别是准确率(Accuracy)、精确率(Precision)、召回率(Recall)和调和平均 F1 分数。这些指标的意义在于:Accuracy 越高,表示算法的整体性能越好;Precision 能够衡量流量异常检测模型识别攻击的能力,Recall 是模型检测所有攻击的能力,Precision 和 Recall 越高,则表示算法的误报率越低;F1 分数是 Precision 和 Recall 的调和平均值,F1 分数越高,说明算法在 Precision 和 Recall 之间取得的平衡越好,算法越稳定。

$$\begin{cases} Accuracy = \dfrac{TP + TN}{TP + TN + FP + FN} \\ Precision = \dfrac{TP}{TP + FP} \\ Recall = \dfrac{TP}{TP + FN} \\ F1 = 2 \times \dfrac{Precision \times Recall}{Precision + Recall} \end{cases} \tag{8}$$

在这些指标当中,TP(true positive)表示将正常样本分类为正常的数量,TN(true negative)是指将异常样本分类为异常的数量,FP(false positive)是指将异常样本分类为正常的数量,FN(false negative)是指将正常样本分类为异常的数量。

3.3　实验结果与分析

1)实验设置

超参数在机器学习和深度学习中是至关重要的,正确选择合适的超参数可以显著影响模型的性能、训练速度和泛化能力。

(1)DBSCAN 参数

DBSCAN 是一种用于聚类的密度连通算法,不需要预先指定聚类数量。DBSCAN 算法有两个主要超参数需要设置:eps 和 MinPts。eps 定义了一个样本点的邻域范围,即一个样本点的 ε -邻域。一个样本点的 ε -邻域包括其周围距离小于等于 ε 的所有样本点。ε 的选择会影响到最终聚类效果。MinPts 定义了一个样本点的 ε -邻域内最少需要多少个样本点,才能将该样本点视为核心点。为了得到较好的 DBSCAN 算法参数,考虑不同参数组合的得分情况,使用 SC 来评估参数的优劣,SC 越高,说明参数组合效果越佳,其公式为:

$$SC = \frac{b - a}{\max(a,b)} \tag{9}$$

其中,a 表示与相同簇内其他点的平均距离,b 表示与最近不同簇内的点的平均距离。分别将 eps 取值区间设置为[0, 4],步长设为 0.2,MinPts 取值区间设置在[40, 50],步长为 1。通过比较不同参数组合下的 SC 值,选择参数组合 eps = 1.8 与 MinPts = 45。经过离群点去除前后各个样本类别的数据量大小如表 6 所示。

表 6　不同类别数据的大小

Table 6　Size of the different categories of data

样本类别	离群点去除前/条	离群点去除后/条
Normal	67 343	65 626
Dos	45 927	44 728
Probe	11 656	11 346
R2L	995	995
U2R	52	52

(2)KNN 参数

KNN 算法中的核心参数是 k 值,表示在进行预测时需要考虑的最近邻个数。选择适当的 k 值取决于数据集和问题的性质,通常可以通过交叉验证的方式来确定最佳的 k 值。本文使用 5 折交叉验证在区间[2, 20]上寻找最佳 k 值,利用 F1 分数作为衡量指标。通过实验验证,发现在 k = 3 时的 F1 分数最高,因此在实验中选择 k = 3。

(3)RF 参数

RF 分类器中存在许多影响模型分类性能的参数,基于网格搜索的方法虽然可以寻找到合适的参数组合,但其效率较低,使用优化算法能更迅速地找到一个效果较好的参数组合,在众多优化算法中,BO 算法在迭代次数少,运行速度快以及对非凸问题表现稳健等方面具有优势。因此,本文使用 BO 算法对 RF 分类器参数组合进行优化,参数具体优化结果如表 7 所示。

表 7　RF 参数设置

Table 7　RF parameter settings

参数	参数说明	参数值
n_estimators	森林中树的个数	105
max_depth	树的最大深度	12
max_features	节点拆分考虑特征数量	0.8
min_samples_split	叶子节点最小样本数	8

2)消融实验

(1)DBSCAN 实验

为了验证去除离群点对所提方法性能的影响,本文首次进行实验。实验通过对比在所提方法中使用 DBSCAN 离群点检测算法去除离群点后的数据集与使用未去除离群点的数据集在检测效果上的差异,来说明离群点检测算法能提升所提方法的流量异常检测性能,结果如图 2 所示。使用 DBSCAN 离群点检测算法去除数据

集中的离群点后再进行一系列后续操作,在检测性能方面有更出色的表现,各个性能指标的检测结果都有一定的提升。具体来说,在准确率上,提升了 4.56%,在精确率方面,提升了 6.86%,在召回率上,提升了 3.74%,在 F1 分数上,提升了 5.59%。相较于使用原始数据进行流量异常检测,去除离群点后的数据,整体性能指标表现更出色。这得益于离群点检测过程中剔除了每个类别中不符合一般规律的异常数据点,降低了这些异常数据对后续 FCG 算法的实现以及模型训练的影响,使得模型稳定性得到增强,泛化性得到提升。

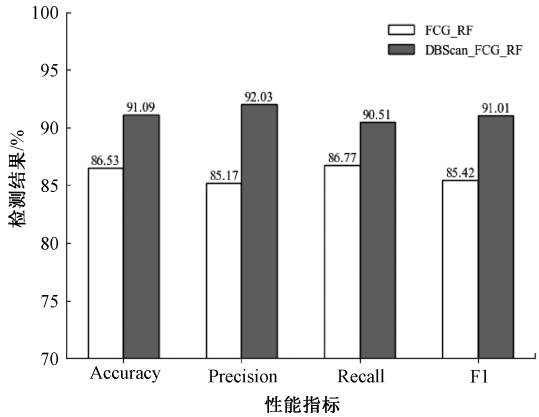

图 2　去除离群点前后的性能比较

Fig. 2　Performance comparison before and after removing outliers

虽然离群点去除之后可以降低异常值对后续算法以及模型的影响,但同时也可能会使得稀疏特征中存在的类别区分信息更加稀少,因为稀疏特征也有可能出现在异常数据点中。为了验证离群点检测与 FCG 算法之间的相互影响,第 2 次进行消融实验。实验旨在通过调整离群点去除的顺序,即将离群点去除分别放在 FCG 算法之前和之后,比较在不同顺序中流量异常检测性能,来分析 FCG 算法与离群点检测之间的相互影响以及确定离群点检测算法在所提方法中的最佳位置,结果如图 3 所示。实验结果表明离群点检测算法去除数据中的异常值能够提升检测性能,在 FCG 算法之前去除离群点虽然会减少稀疏特征中的类别区分信息,但其在降低异常值对后续 FCG 算法中 CDF 生成以及 EDF 与 CDF 之间的共现信息的影响方面作用更大,因此将离群点检测工作放在 FCG 算法之前完成。

(2) CDF 实验

为了验证基于数据集中多个原始特征生成的 CDF 对异常检测分类性能的影响[22],第 3 次进行实验。因为生成的 CDF 为布尔特征,所以进行二分类实验验证 CDF 的性能,即仅区分正常数据和异常数据。实验通过对比分析将 CDF 加入数据集与不加入数据集时的二分类检

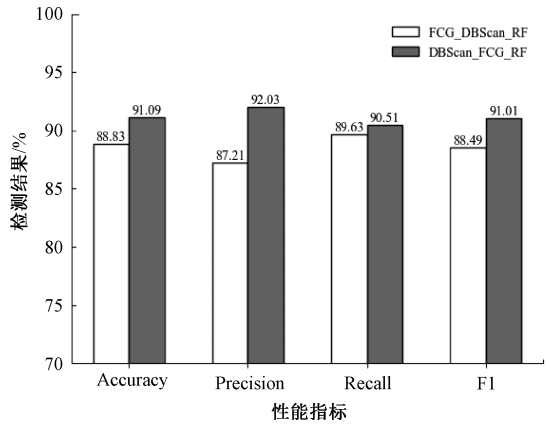

图 3　不同顺序的性能比较

Fig. 3　Performance comparison of different sequences

测结果,验证 CDF 的性能。其中,设置了两组对比实验,第 1 个对比实验,使用具有强类别区分能力的特征生成 CDF,并利用众数填充存在缺失值的样本中的 CDF 值,通过此对比实验,验证缺失值填充方法的有效性,将生成的 CDF 与原始数据一起基于 RF 模型进行二分类实验。第 2 个对比实验,同样使用强类别区分能力的特征生成 CDF,但不同之处在于后者使用 KNN 算法填补存在缺失值的样本中的 CDF 值,同样,将生成的 CDF 与原始特征一起基于 RF 模型进行二分类实验,通过这一对比实验,验证 KNN 算法的高效性。结果如表 8 所示。由实验结果可知,新生成的 CDF 能够更好地区分数据中的正常数据和异常数据。这是因为在生成 CDF 的过程中,融合了基于规则和基于机器学习模型的优点,在无法明确编码规则的情况下,利用机器学习算法填充生成的 CDF 值。对比实验 1 表明缺失值填充技术可以提高 CDF 的分类性能。对比实验 2 表明,与传统的缺失值填充技术相比,利用机器学习算法进行缺失值填充可以取得更好的效果。生成的 CDF 质量越高,区分正常数据和异常数据的能力就越强,最后通过与 EDF 的共现信息生成细分异常数据类别的新特征,在提升流量异常检测性能方面发挥的作用越大。

表 8　二分类检测结果

Table 8　Results of binary classification detection

模型	Accuracy	Precision	Recall	F1-score
RF	0.851 3	0.862 1	0.849 6	0.855 8
RF+CDF+MODE	0.897 9	0.905 4	0.860 4	0.882 9
RF+CDF+KNN	0.935 6	0.940 2	0.957 0	0.949 5

(3) FCG 消融实验

本文 FCG 算法中 EDF 的选取,旨在提升那些具有潜在类别区分能力,但由于其自身稀疏性,容易被特征选择算法忽略的特征在分类任务中的作用。因此,在选择 EDF 时,倾向于选取那些具有实例区分能力的稀疏特征。

通过利用选择的 EDF 与 CDF 之间的共现信息,激发新特征的类别区分能力。在这过程中,使用基于本体的方法,将一组针对同一种攻击类型的 EDF 结合在一起,丰富与 CDF 之间的共现信息,利用这种方式可以有效解决生成的新特征数据稀疏性的问题。每个攻击类别所选 EDF 与 CDF 的具体情况如表 9 所示,其中,new feature 就是上文通过多个强类别区分能力特征生成的 CDF。

表 9 选择的 CDF 和 EDF

Table 9 selected CDF and EDF

攻击类型	EDF	CDF
Dos	7,8,10	new feature
Probe	5,30,31	new feature
R2L	10,12,18	new feature
U2R	14,16,17	new feature

为了验证 FCG 算法的可行性,进行第 4 个消融实验来评估 FCG 算法生成新特征的整体性能。实验通过比较使用未添加 FCG 生成的新特征的数据与添加 FCG 生成的新特征后的数据的检测效果,证明 FCG 算法能提升分类器的性能。为了实现这一目的,利用每种攻击类别所选的 EDF 与 CDF 之间的共现信息计算 FCD,生成新特征,然后将这些新特征添加到数据集中,与原始特征一起基于 RF 模型进行训练和测试,与未使用 FCG 算法的分类性能进行比较。实验结果如图 4 所示。

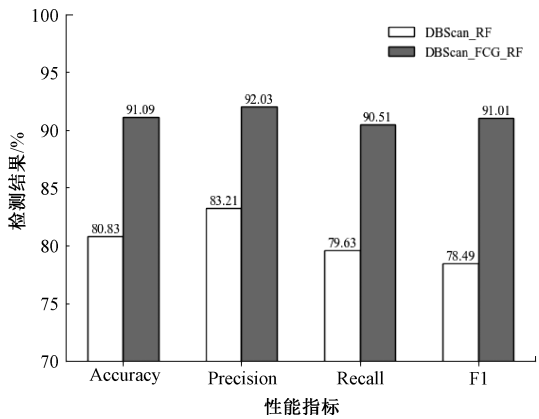

图 4 特征耦合泛化前后的性能比较

Fig. 4 Performance comparison before and after feature coupling generalization

实验结果表明,添加生成的新特征后,在 4 个性能指标上的检测结果相对于基于原始数据特征的 RF 模型均有很大的提升。具体而言,经过 FCG 算法生成新特征之后,在准确率上,提升了 10.26%,在精确率上,提升了 8.82%,在召回率上,提升了 10.88%,在 F1 分数上,提升了 12.52%。FCG 算法之所以对模型检测结果有如此大的提升,首先是因为第 1 步生成的 CDF 在区分正常数据和异常数据上起了关键作用。其次,通过将一组针对同一类别数据的稀疏特征组合起来,生成的新特征成功地保留了原有稀疏特征的类别区分能力,同时过滤掉了对分类影响较小的信息,对异常数据中样本的具体类别区分有重要意义,最后,通过将低频特征利用 FCG 算法组合起来与高频特征一起训练分类器,提升了模型的整体性能。

如图 5 所示,使用 FCG 算法后每个类别的检测性能都有一定的提升。在少数攻击类 R2L 和 U2R 上因为其训练样本不足,模型训练不够充分,导致性能提升效果并不明显。测试样本较少,使得正确检测和错误检测的样本数量的轻微变化会引起模型检测率的较大浮动。

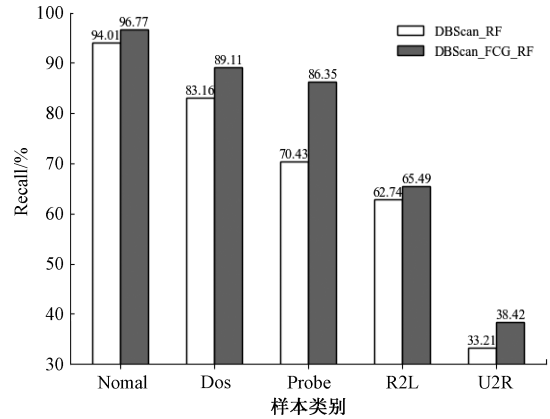

图 5 特征耦合泛化前后的各个类别性能比较

Fig. 5 Performance comparison of each class before and after feature coupling generalization

3)与其他方法对比

为了更加全面地验证所提方法在流量异常检测分类方面的性能,将本文所提方法与经典机器学习方法以及其他流量异常检测方法进行了比较。采用 BO 算法优化 RF 模型的重要参数,进一步提高模型的流量异常检测性能。本文方法与经典机器学习算法以及其他方法的检测结果如表 10 所示,其中使用 4 个评价指标进行对比分析,分别为 Accuracy、Precision、Recall 和 F1 分数。RF 是未使用 BO 优化参数的实验结果。

表 10 与其他方法的比较

Table 10 Comparison with other methods

模型	Accuracy	Precision	Recall	F1-score
NB	0.760 2	0.873 2	0.765 1	0.786 9
DT	0.885 8	0.936 2	0.896 5	0.901 7
KNN	0.870 9	0.891 2	0.870 1	0.872 4
SVM	0.876 6	0.881 2	0.884 3	0.883 9
RF	0.910 9	0.920 3	0.905 1	0.910 1
文献[23]	0.913 6	0.928 1	0.913 6	0.916 7
文献[24]	0.894 0	0.907 5	0.893 5	0.883 6
文献[25]	0.903 1	0.904 3	0.903 2	0.902 3
本文	0.917 9	0.935 9	0.921 4	0.929 1

根据表 10 中的实验结果可知,使用 RF 进行分类识

别,在所提方法中相比其他经典机器学习算法表现更出色,表明采用 RF 集成学习模型具有一定的优势。文献[23]采用双向 LSTM 的网络异常检测方法,各项性能指标都优于传统 LSTM 模型,但双向 LSTM 具有更高的复杂度,比传统 LSTM 模型和其他机器学习模型需要更多的训练时间。文献[24]基于 ADASYN 与改进残差网络的流量异常检测,整体性能相对稳定,但其需要大量的标记数据进行监督学习,在某些场景中可能不易获取足够的标记数据。文献[25]采用双重路由深层胶囊网络,可提取出更高维度的数据特征,使用混合注意力机制使得模型倾向于依赖影响较大的特征,降低了具有潜在影响因素的特征在流量异常检测过程中的作用。相比之下,本文提出的基于 FCG 半监督算法的流量异常检测方法充分挖掘稀疏特征潜在的类别区分能力,结合 RF 模型在检测效果、训练时间以及泛化性等方面的优势,在五分类情况下展现了良好的性能,充分说明该方法的合理性,能有效提高流量异常检测性能。

4　结　论

本文提出一种基于 FCG 的流量异常检测方法。相较于传统的检测方法,基于 FCG 的方法利用原始特征进行组合生成新的特征,降低特征维度的同时充分挖掘存在于稀疏特征中被忽略的有用信息,以提升分类结果的准确性,检测性能更好。生成的新特征与原始特征相比,包含更丰富的信息量和更泛化的表示。针对原始 FCG 算法在选择 CDF 上存在类别区分能力不足的问题,利用多个类别区分能力较强的特征共同生成 CDF。通过在 NSL-KDD 数据集对所提方法进行实验验证其性能,结果表明所提方法能够有效提高流量异常检测分类能力,突出稀疏特征的类别区分能力。由于存在少数类攻击样本不足的问题,模型训练不够充分,导致少数类攻击的检测效果提升不明显,在接下来的工作中需要进行更深入的研究和优化,以进一步增强所提方法在处理多样化异常检测情况下的鲁棒性和准确性。

参考文献

[1] 蹇诗婕, 卢志刚, 杜丹, 等. 网络入侵检测技术综述[J]. 信息安全学报, 2020, 5(4): 96-122.
JIAN SH J, LU ZH G, DU D, et al. Overview of network intrusion detection technology[J]. Journal of Cyber Security, 2020, 5(4): 96-122.

[2] WU W F, LIR F, XIE G Q, et al. A survey of intrusion detection for in-vehicle networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 21(3):919-933.

[3] KWON D, KIM H, KIM J, et al. A survey of deep learning-based network anomaly detection[J]. Cluster Computing, 2019, 22: 949-961.

[4] SUTHISHNI D N P, KUMAR K S S. A review on machine learning based security approaches in intrusion detection system[C]. 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022: 341-348.

[5] RESENDE P A A, DRUMMOND A C. A survey of random forest based methods for intrusion detection systems[J]. ACM Computing Surveys (CSUR), 2018, 51(3): 1-36.

[6] AHMAD I, BASHERI M, IQBAL M J, et al. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection[J]. IEEE Access, 2018, 6: 33789-33795.

[7] LI Y, LIN H, YANG Z. Incorporating rich background knowledge for gene named entity classification and recognition[J]. BMC Bioinformatics, 2009, 10(1): 1-15.

[8] 何林娜, 杨志豪, 林鸿飞, 等. 基于特征耦合泛化的药名实体识别[J]. 中文信息学报, 2014, 28(2): 72-77.
HE L N, YANG ZH H, LIN H F, et al. Drug name entity recognition based on feature coupling generalization[J]. Chinese Journal of Information, 2014, 28(2): 72-77.

[9] NIU Y, CHEN C, ZHANG X, et al. Application of a new feature generation algorithm in intrusion detection system[J]. Wireless Communications and Mobile Computing, 2022, 2022: 1-17.

[10] KHAMMASSI C, KRICHEN S. A GA-LR wrapper approach for feature selection in network intrusion detection[J]. Computers & Security, 2017, 70: 255-277.

[11] KUSUMAPUTRI F H, ARIFIN A S. Anomaly detection based on NSL-KDD using XGBoost with OptunaTuning[C]. 2022 7th International Conference on Business and Industrial Research (ICBIR). IEEE, 2022: 586-591.

[12] VENKATESAN S. Design an intrusion detection system based on feature selection using ML algorithms[J]. Mathematical Statistician and Engineering Applications, 2023, 72(1): 702-710.

[13] LEO B. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[14] 杨红浩, 周治平. 采用信息增益率的混合入侵检测模型设计[J]. 信息与控制, 2019, 48(4): 420-428.
YANG H H, ZHOU ZH P. Design of hybrid intrusion detection model utilizing information gain rate[J]. Information and Control, 2019, 48(4): 420-428.

[15] 任家东，刘新倩，王倩，等. 基于KNN离群点检测和随机森林的多层入侵检测方法[J]. 计算机研究与发展，2019，56(3)：566-575.
REN J D，LIU X Q，WANG Q，et al. An multi-level intrusion method based on KNN outlier detection and random forests[J]. Journal of Computer Research and Development，2019，56(3)：566-575.

[16] 胡姣姣，王晓峰，张萌，等. 基于深度学习的时间序列数据异常检测方法[J]. 信息与控制，2019，48(1)：1-8.
HU J J，WANG X F，ZHANG M，et al. Time-series data anomaly detection method based on deep learning[J]. Information and Control，2019，48(1)：1-8.

[17] 梁欣怡，行鸿彦，侯天浩. 基于自监督特征增强的CNN-BiLSTM网络入侵检测方法[J]. 电子测量与仪器学报，2022，36(10)：65-73.
LIANG X Y，XING H Y，HOU T H. CNN-BiLSTM network intrusion detection method based oneself-supervised feature enhancement[J]. Journal of Electronic Measurement and Instrumentation，2020，36(10)：65-73.

[18] 刘新倩，单纯，任家东，等. 基于流量异常分析多维优化的入侵检测方法[J]. 信息安全学报，2019，4(1)：14-26.
LIU X Q，SHAN CH，REN J D，et al. An intrusion detection method based on multi-dimensional optimization of traffic anomaly analysis[J]. Journal of Cyber Security，2019，4(1)：14-26.

[19] PENG H，LONG F，DING C. Feature selection based on mutual information：Criteria of max-dependency，max-relevance，and min-redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2005，27(8)：1226-1238.

[20] EMMANUEL T，MAUPONG T，MPOELENG D，et al. A survey on missing data in machine learning[J]. Journal of Big Data，2021，8(1)：1-37.

[21] KUMARSHRIVAS A，KUMAR DEWANGAN A. An ensemble model for classification of attacks with feature selection based on KDD99 and NSL-KDD data set[J]. International Journal of Computer Applications，2014，99(15)：8-13.

[22] 吴建胜，张文鹏，马垣. KDDCUP99数据集的数据分析研究[J]. 计算机应用与软件，2014，31(11)：321-325.
WU J SH，ZHANG W P，MA Y. Data analysis and study on KDDCUP99 data set[J]. Computer Applications and Software，2014，31(11)：321-325.

[23] IMRANA Y，XIANG Y，ALI L，et al. A bidirectional LSTM deep learning approach for intrusion detection[J]. Expert Systems with Applications，2021，185：115524.

[24] 唐玺博，张立民，钟兆根. 基于ADASYN与改进残差网络的入侵流量检测识别[J]. 系统工程与电子技术，2022，44(12)：3850-3862.
TANG X B，ZHANG L M，ZHONG ZH G. Intrusion traffic detection and identification based on ADASYN and improved residual network[J]. Systems Engineering and Electronics，2022，44(12)：3850-3862.

[25] 尹晟霖，张兴兰，左利宇. 双重路由深层胶囊网络的入侵检测系统[J]. 计算机研究与发展，2022，59(2)：418-429.
YIN SH L，ZHANG X L，ZUO L Y. Intrusion detection system for dual route deep capsule network[J]. Journal of Computer Research and Development，2022，59(2)：418-429.

作者简介

陈万志(通信作者)，2015年于辽宁工程技术大学(中国测绘科学研究院联合培养)获得博士学位，现为辽宁工程技术大学副教授，硕士生导师，主要研究方向为人工智能与智能信息处理、网络与信息安全和工控软件与数据分析。

E-mail：chenwanzhi@lntu.edu.cn

Chen Wanzhi (Corresponding author) received his Ph. D. degree from Liaoning Technical University (China Academy of Surveying and Mapping Science Joint Cultivation) in 2015, respectively. Now he is an associate professor and master's degree supervisor in Liaoning Technical University. His main research interests include artificial intelligence and intelligent information processing, network and information security and industrial control software and data analytics.

张国满，2021年于辽宁工程技术大学获得学士学位，现为辽宁工程技术大学硕士研究生，主要研究方向为网络安全和入侵检测。

E-mail：481218583@qq.com

Zhang Guoman received his B. Sc. degree from Liaoning Technical University in 2021. Now he is a M. Sc. candidate at Liaoning Technical University. His main research interests include network security and intrusion detection.