**Advances in Applied Mathematics** 应用数学进展, 2023, 12(6), 2736-2748
Published Online June 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.126274

Hans汉斯

# 基于深度学习的入侵检测数据分类研究

**金　颖**[1,2]

[1]成都信息工程大学网络空间安全学院，四川　成都
[2]成都信息工程大学先进密码技术与系统安全四川省重点实验室，四川　成都

收稿日期：2023年5月13日；录用日期：2023年6月7日；发布日期：2023年6月15日

## 摘　要

**针对由于入侵检测数据集中数据类别不平衡，而导致的检测分类准确率低的问题，设计一种基于生成对抗网络(GAN)和深度森林结合的入侵检测模型。首先，基于生成对抗网络独有的对抗思想，通过原数据类的分类结果，筛选出需要生成的类别，生成数据集中缺少的数据，缓解数据集不平衡的问题；然后，针对网络流量特征复杂与深度森林模型数据处理成本高的矛盾，设计了基于主成分分析和线性判别算法结合的特征提取方法，解决了深度森林模型中的数据计算冗余问题，提高了数据传递与处理能力。实验结果证明，所提方法的分类检测精度达到了96%，其中少数类数据的检测精度比没有平衡前提高了30%。**

## 关键词

**生成对抗网络，入侵检测，不均衡数据分类，深度森林，特征提取**

# Research on Intrusion Detection Data Classification Based on Deep Learning

**Ying Jin**[1,2]

[1]School of Cybersecurity, Chengdu University of Information Technology, Chengdu Sichuan
[2]Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chengdu University of Information Technology, Chengdu Sichuan

Received: May 13th, 2023; accepted: Jun. 7th, 2023; published: Jun. 15th, 2023

## Abstract

**Aiming at the problem of low detection and classification accuracy due to the imbalance of data categories in the intrusion detection dataset, an intrusion detection model based on the combination of generative adversarial networks (GAN) and deep forest is designed. .First of all, based on**

文章引用: 金颖. 基于深度学习的入侵检测数据分类研究[J]. 应用数学进展, 2023, 12(6): 2736-2748.
DOI: 10.12677/aam.2023.126274

**the adversarial characteristics of generated adversarial networks, the classes that need to be generated are screened out through the classification results of the original data and the missing data in the dataset is generated to alleviate the problem of dataset imbalance. Then, aiming at the contradiction between the complex network traffic characteristics and the high data processing cost of the deep forest model, a feature extraction method based on the combination of principal component analysis and linear discriminant analysis is designed. It solves the data calculation redundancy problem in the deep forest model and improves the data transmission and processing capabilities. The experimental results show that the classification detection accuracy of the proposed method reaches 96%, and the detection accuracy of the minority class data is 30% higher than that without balance.**

## Keywords

**Generative Adversarial Network (GAN), Intrusion Detection, Imbalanced Data Classification, Deep Forest, Feature Extraction**



## 1. 引言

入侵检测系统可以对外来的数据信息进行分类，通过分析数据的特征，对照数据库进行检测，有助于保护计算机，避免一些潜在的威胁[1]。然而，由于新的攻击方法层出不穷，入侵检测的系统模型需要不断完善来应对新型攻击。入侵检测系统目前存在的主要问题包括 1) 入侵检测数据集存在很大的不平衡性，正常数据的数据量远远大于异常数据。2) 异常数据库更新慢，无法跟上攻击方法的更新，很多未知行为和数据人工无法判断是否异常。3) 要处理的数据信息量大，处理难度高[2]。

其中，数据的不平衡性会严重影响入侵检测系统的检测准确率。对于数量稀少的异常数据类，系统只能提取到部分特征，导致分析结果不全面，后续容易造成判别误差。研究人员为此提出了大量的不平衡数据处理方法[3]，例如随机采样、SMOTE 算法等，但是总会存在一些由于选择不当而造成的数据冗余或者关键数据缺失等问题。生成对抗网络(GAN) [4]是一种被广泛用于数据处理的深度学习算法，由于它对数据独特的生成方式，使得其可以弥补很多数据量过少带来的问题，而且因为生成的数据是根据原数据的本有数据分布特征来生成，避免了增加无用的数据信息。因此，采用生成对抗网络是解决不平衡数据问题的一种有效手段。

深度学习是目前研究的一大热点[5]，它具有学习能力强、性能稳定和提取特征能力强等优点，可以自主学习已有的异常数据，在面对未知的数据时，有更准确的判别能力。但是以往的深度学习算法中所采用的多层神经元结构，使得其对数据量要求过大或者存在过于依赖超参数等问题。为此，周志华等提出了深度森林算法(gcForest) [6]，它是一种由树结构为核心的非神经网络式的深度学习算法。在面对小数据集上，深度森林算法相对于传统深度学习算法具有明显优势，其具有树结构的参数少、调参简单等特点，很好地改善了模型训练效率低的问题。

本论文首先针对入侵检测数据集不平衡的问题，采用生成对抗网络算法，利用生成器和判别器的对抗博弈，生成更加符合原本数据特性的新数据集，改善了异常数据过于稀少而导致分类结果过低的问题。此外，在分类问题上，本论文提出了一个改进的深度森林方法作为分类器，通过高效的特征提取方法提

高了深度森林的训练效率。

## 2. 相关工作

关于入侵检测数据的分类研究，主要分为数据预处理和分类检测两部分。在数据预处理方面，目前大多数的入侵检测数据集都存在数据不平衡的问题，异常的网络攻击数据总是比较稀少，从而导致模型对于异常数据类的学习难度大大增加，影响了检测效果，因此，数据不平衡一直都是阻碍入侵检测效果的一大难题，对此很多研究人员提出了大量的解决方案。2014 年．Yong 等人[7]为了更好地解决不平衡数据的问题，采用几种经典的算法相混合的方法，提出了 SMOTE-NCL 算法，将过采样和欠采样将结合，用 SMOTE 对少数类进行生成，再用欠采样对多数类进行选择，以此来做到最大化的提高。2017 年，Yan 等人[8]试着将算法与机器学习混合使用，提出了先用 SMOTE 算法对数据进行采样，得到的新的数据集后，再与 SVM、RF 等分类算法相结合的办法，实验证明，平衡后的数据比起平衡之前的数据，在训练模型上有了显著地提升。

近几年来，研究人员尝试将 GAN 运用于不平衡数据处理。实验证明，GAN 在处理不平衡数据集上有着很大的优势[9]。2020 年，Lee 等人[10]为了解决数据不平衡问题，提出了 GAN-RF 模型，在 GAN 生成数据后采用 RF 算法进行分类，并对生成前后的数据进行了明显的对比，提供了在合成后和合成前数据集内少数类数据的数量变化表，实验表明，在采用了 GAN 进行处理不平衡数据后，少数类数据的分类精度确实得到了大量的提升。Liao 等人[11]为了解决样本数量不足的问题，同样提出了基于生成对抗网络的入侵检测模型，通过对原数据集的直接处理，不断生成数据，扩充数据集来辅助检测，实验证明分类检测的结果比原数据集要好，但是分类检测提升不是很高，证明直接无筛选的数据生成方法，并没有达到最佳的提升效果。Shahriar 等人[12]提出了 G-IDS 的辅助入侵检测算法，采用了合成数据和原始数据采用两个相同的模型，做两次分类对比分数的方法，分类不结果不理想就删除这部分数据重新生成，这种实验方法更有效的提高了生成的效果，最终 G-IDS 的 F1-score 高达 95%~97%，但是这个数值是针对于 100%的数据集的，若只采用 25%的数据，F1-score 会降低至 86%左右，所以该模型对数据量的要求很大，计算处理更复杂，耗时更多。

在分类检测方面，研究人员也进行了大量的研究。钱铁云等人[13]为了改善传统机器学习算法在处理复杂数据时的不足，提出了一种基于深度神经网络的检测方法，结构采用神经元映射卷积神经网络(NPCNN)，最终实验表明，神经网络结构比起一般机器学习算法，具有更好的特征学习和分类能力。但是深度神经网络也存在自身的局限性，首先，神经网络结构的训练结果非常依赖于大量的数据，一旦数据量不足，学习的效果就会大打折扣，甚至不如一般的算法，而如今在很多领域都没法获取到足够的带标签的数据。其次，深度神经网络属于非常复杂的模型，在模型的设计过程中，需要对模型进行大量的调参，面对不同的问题，不同的领域可能都得大幅度更改参数，一旦参数设置不符合，就非常影响学习能力。鉴于神经网络的这些问题，2019 年，周志华等人提出了深度森林算法，该算法延用了一般算法多层的结构，但是并没有采用神经元作为其中的节点，而是采用了集成的树结构。树结构的参数比起神经网络而言使得该算法在小规模数据上比一般深度学习算法具有更强的训练效果。不仅如此，它对超参数的依赖程度也不高，在不同领域都能直接使用。丁龙斌等[14]为了降低数据处理的复杂程度，提高检测准确率，提出了一种基于集成随机森林的检测方法，将 bagging 集成策略和随机森林将结合，实验结果表明，在分类准确率上有了很大的提高，且其收敛速度也提升至 50%以上。范怡敏等[15]提出一种改进深度森林模型(two boosting deep forest, TBDForest)，采用均等式特征利用方法对原始特征进行变换，将上一层的重要特征传递给下一层，结果表明，改进的算法比起原本的算法分类性能有了提高。颜建军等[16]针对中医问诊复杂性和非线性的特点，采用深度级联森林算法(multi-grained cascade forest, gcForest)构建

分类模型并与多种模型进行比较。实验结果表明，基于 gcForest 算法的模型要明显优于其他算法，在解决慢性胃炎中医问诊证候分类问题上更加有效。2020 年，蒋鹏飞等人[17]提出了基于深度森林与 CWGAN-GP 的关于移动应用网络行为的分类检测，不仅得到了很高的分类准确率，而且解决了样本数量少，数据获取开销大的问题。2021 年，张鹏等人[18]针对无线传感器网络(WSN)节点容易出现故障的问题，提出了基于深度森林的无线传感器网络故障分类算法，与一些经典的神经网络进行对比，实验表明，深度森林可以更准确地识别出不同的故障类型。2022 年，王耀聃等人[19]为了解决目前窃电检测方法提取的特征有效性低的问题，提出一种基于改进深度森林的窃电检测模型。所提模型与 RF 和 ANN 进行对比，实验表明，检测精度分别提高 13.68%和 17.7%，证明深度森林算法有效提高了少数类窃电用户的检出率。

综上所述，不平衡数据集和分类模型难度大是目前入侵检测系统面临的两大难题，不平衡数据的处理从最经典的欠采样和过采样、集成算法、几个算法组合到如今的与机器学习和深度学习结合，面对不同的数据，可能依旧会存在数据处理不当的问题。分类模型从传统机器学习算法到经典的深度学习算法，很好改善了模型的学习能力和面对未知数据的判断能力。因此，本文采用基于生成对抗网络的深度森林模型，通过改进的生成对抗网络进行不平衡数据处理，再选用深度森林作为分类模型，既满足了多层结构的模型复杂性，又结合了树结构，减少了模型训练的难度，提高了效率。

## 3. 方法

入侵检测数据分类的过程由 4 个阶段组成，如图 1 所示。一是数据预处理阶段，为了解决不同特征不同数据类型对数据处理的影响，首先要对数据进行数据编码，使数据类别都转换为数字，然后对数据进行标准化处理，使数据分布在[0,1]区间，便于不同的单位或者不同量级的指标能够进行比较；二是特征提取阶段，先采用 PCA 降维，计算方法简单，便于操作，且能在一定程度上起到降噪的效果，再采用 LDA 降维，LDA 是有监督的方法，可以选择分类性能最好的方向，最终选出最适合的特征作为新的特征。三是不平衡数据处理过程，通过一个控制参数，划定需要生成的类别的范围，然后采用生成对抗网络来生成这部分异常数据，以此增加异常数据的数量，来与正常数据平衡，再采用重采样算法对少部分数据进行优化。四是分类阶段，使用深度森林对处理后的数据进行分类，最终的分类准确率，召回率等指标作为模型的评价指标。

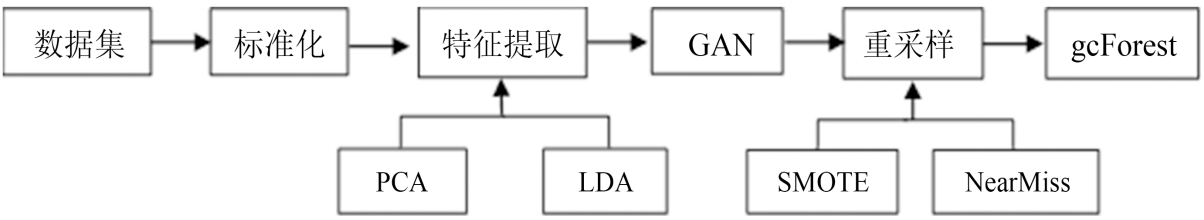


**Figure 1.** Experimental process
**图 1.** 实验过程

### 3.1. 数据集

研究采用的数据集是 KDD99 数据集的修订版本 NSL-KDD (https://www.unb.ca/cic/datasets/nsl.html)，数据集每条记录包含 43 个特征，其中 41 个特征指的是流量输入本身，最后两个是标签(正常或攻击)和分数(流量输入本身的严重性)。数据集中包括一类正常数据(Normal)和四种攻击方法，攻击方法分别是拒绝服务(DoS)、探测(Probe)、用户到根(U2R)和远程到本地(R2L)。数据集的详细说明，如表 1，表 2 所示。

### 3.2. 特征提取

数据集中许多特征在生成模型时没用，为了减少算法复杂性，缩小数据的维度或者尺寸很重要。一个模型是否可以训练地更佳，很大程度上源于数据的优劣，所以，特征提取方法非常重要，如何判断哪些特征更有用，哪些特征是多余的特征更是重中之重。

主成分分析 PCA 使用正交变换将一组可能相关变量的观测值转换为一组称为主成分的线性不相关变量值。每个主成分都是原始变量的线性组合。顾名思义，就是找出数据里最主要的方面，用数据里最主要的方面来代替原始数据。因为所有主成分相互正交，因此没有冗余信息。PCA 计算方法简单，易于实现，且降噪容易，但是因为是无监督的降维技术，降维目的模糊，不能完全区分类别

LDA 是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的，但是与 PCA 不同的是 LDA 的降维维度是有限制的。LDA 方法主要是，对于一个多类别的分类问题，想要把它们映射到一个低维空间，如一维空间从而达到降维的目的，我们希望映射之后的数据间，两个类别之间离得越远，且类别内的数据点之间离得越近，这样两个类别就越好区分。

因此，在本文中选择 PCA 和 LDA 结合，综合两种特征提取方法的优势，既做到操作简单，易于降噪，又可以按照样本标签，更加便捷地选择出让分类性能最佳的特征，在数据处理上有很大优势，减轻了后续分类模型的训练复杂度，降低了消耗，给分类模型训练提供了极大地便捷。

### 3.3. GAN

GAN (Generative Adversarial Networks)是 2014 年由 Goodfellow 提出的一种带有对抗思想的深度学习

**Table 1.** Details on the number of datasets
**表 1.** 数据集数量明细

| 数据集 | 数量明细 | | | | | |
|---|---|---|---|---|---|---|
| | 总数 | Normal | DoS | Probe | U2R | R2L |
| KDDTrain + 20% | 25,192 | 13,449 (53%) | 9234 (37%) | 2289 (9.16%) | 11 (0.04%) | 209 (0.8%) |
| KDDTrain+ | 125,973 | 67,343 (53%) | 45,927 (37%) | 11,656 (9.11%) | 52 (0.04%) | 995 (0.85%) |
| KDDTest+ | 22,544 | 9711 (43%) | 7458 (33%) | 2421 (11%) | 200 (0.9%) | 2654 (21.1%) |

**Table 2.** Subclasses of attack types
**表 2.** 攻击类型的子类细分

| 类型 | Dos | Probe | U2R | R2L |
|---|---|---|---|---|
| 个数 | 11 | 6 | 7 | 15 |
| 子类 | apache2<br>back<br>land<br>neptune<br>mailbomb<br>pod<br>processtable<br>smurf<br>teardrop<br>udpstorm<br>worm | ipsweep<br>mscan<br>nmap<br>portsweep<br>saint<br>satan | buffer_overflow<br>loadmodule<br>perl<br>ps<br>rootkit<br>sqlattack<br>xterm | ftp_write<br>guess_passwd<br>httptunnel<br>imap<br>multihop<br>named<br>phf<br>sendmail<br>snmpgetattack<br>warezclient<br>waremaster<br>xlock<br>xsnoop |

算法，它主要是由一个生成器和一个判别器组成，生成器分析原有的数据特征分布，用一段随机噪音生成与原数据相类似的新数据，判别器来判断接收到的数据是否为原数据，返回一个 0~1 之间的数字。两者相互博弈对抗，直到判别器分辨不出接收到的数据的真假，就代表生成的新数据已经无限接近原本数据了，也就基本可以当作真实数据来使用了。

当原数据集数量过大时，将所有数据都传给 GAN 去生成会花费大量的时间，造成资源的浪费，所以，本研究在 GAN 生成之前增加了一个控制器模块，用于筛选出需要生成的数据类别。首先，将原数据集直接发送给分类检测模型，模型进行第一次训练，给出几项检测结果，分别是准确率、精确率、召回率和 F1 值。

在数据不平衡的情况下，精确率和召回率是比较准确且适合的指标，F1 值是它们的调和平均值，所以综合一下，这里采用 F1 值作为评判标准。

控制器的主要作用是控制需要生成的数据类的范围，全部数据都生成会造成数据量过大，但是生成的类别太少又会导致部分数据类没有优化，影响最终分类性能，所以筛选出适合的范围很重要。因此，我们将控制器的参数设置为 0.6，0.7，0.8，0.9，1.0 进行对比实验，在第一次训练后得到的结果中进行筛选，将 F1 值低于参数的类别标记为弱类，只将这部分数据传给 GAN 进行处理。实验对比证明，当参数在 0.6~0.8 范围时，分类性能随着参数值越高越好，但是最高精度也没满 90%，说明仍然存在一部分数据没有优化，拉低了整体的分类精度。参数为 0.9 和 1.0 时，训练后模型的分类精度达到了比较好的水平，但是在高精度的同时，参数 1.0 时候生成的数据量大，花费的时间也更多，不如参数 0.9 时更好，所以最终选择 0.9 作为控制器的参数，进行后续的实验。

### 3.4. 深度森林

目前的深度学习技术基本上都是通过神经网络模型来实现的，因此普遍意义上，深度学习基本等同于深度神经网络，但是深度神经网络参数太多，学习太过依赖于调参，应用在不同的场合就需要重新调整参数，过于麻烦，鲁棒性低，而且学习特征需要大量的数据。周志华教授等人研究提出了深度森林这个新的完全不同的模型，避免了深度神经网络的缺陷，又继承了深度学习的优点，符合了深度学习中逐层加工、特征变换以及模型复杂度三个特点。深度森林算法训练简单，理论上容易理解，执行效率也更高，还融入进树结构自身的特性，比之深度神经网络有着更大的广泛性。

深度森林是一个非神经网络式的深度模型，主要的结构是级联结构和多粒度扫描。级联结构图如图 2 所示，它是一个多层结构，每一级都包括两个随机森林和两个完全随机森林，多种的森林结构可以增加模型的复杂性和多样性。每个森林将输出一组向量，将其连接起来加上前一级的特征作为下一次的输入。在模型中每一层训练都是独立的，每一层森林训练完都会在测试集上进行检验，得到一组结果，然后与上一层的结果进行对比，假如得到的结果基本没有了提升，那么训练停止。

深度森林主要采用的是多粒度扫描的方法，对于序列数据，确定窗口应该的大小，滑动窗口扫描原始数据，得到输出结果，通过不同的窗口大小，也就可以处理不同的数据，总体类似于卷积神经网络中的卷积核提取局部特征，主要结构如图 3 所示。

由于多粒度扫描会将输入的数据转换为数百甚至数千个新的实例，导致最后输出的数据非常庞大，增加了很大的处理难度，并且还给随后的及联过程产生高维输入。传递数据给及联森林后，导致及联的每一层处理起来时间复杂度增加，内存消耗大。因此，我们采用 PCA 和 LDA 算法来辅助数据的特征提取，减轻后续模型的处理压力和消耗。

本文将 PCA 和 LDA 与级联森林的每一级进行集成，提出了一种基于混合特征提取的深度森林模型，如图 4 所示。PCA 在降维过程中不受其他因素干扰，减少了数据之间的影响。但 PCA 是一种无监督的降维方法，没有类别区分的作用，对于异常检测来说不能划分出最佳的效果，所以加入了 LDA 算法。LDA

与 PCA 不会互相影响，在降维的同时找到分类效果最好的投影方向。

具体的实现过程如下。

1) 首先采用 PCA 算法，分别求出各个特征的平均值，然后对其中每一个样本，都减去对应的均值。

2) 计算出样本的特征协方差矩阵。

3) 计算出特征协方差的特征值和特征向量，并归一化为单位向量。

4) 构建特征向量矩阵。将得到的特征值按照从大到小排序，选择其中最大的 k 个值，然后将其对应的 k 个特征向量分别作为列向量组成特征向量矩阵。

5) 让整个数据集的样本点向向量空间投影，实现了降维的功能。

6) 采用 LDA 算法，计算 PCA 降维后的样本类内散度矩阵 Sw 和类间散度矩阵 Sb。

7) 计算矩阵 Sw-1Sb 以及其特征值和特征向量。

8) 在满足设定条件下构建特征向量矩阵。选择前 k 个特征向量，以列向量的形式组合，构造一个 d*k 的转换矩阵 W。

9) 根据特征向量空间投影，完成分类降维功能。

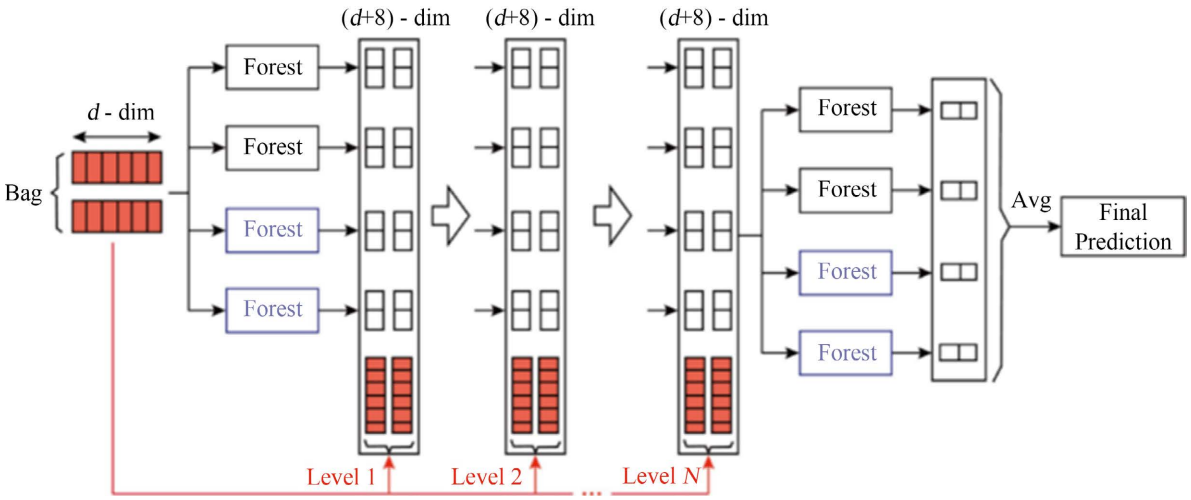


**Figure 2.** Cascade forest structure
**图 2.** 级联森林结构图

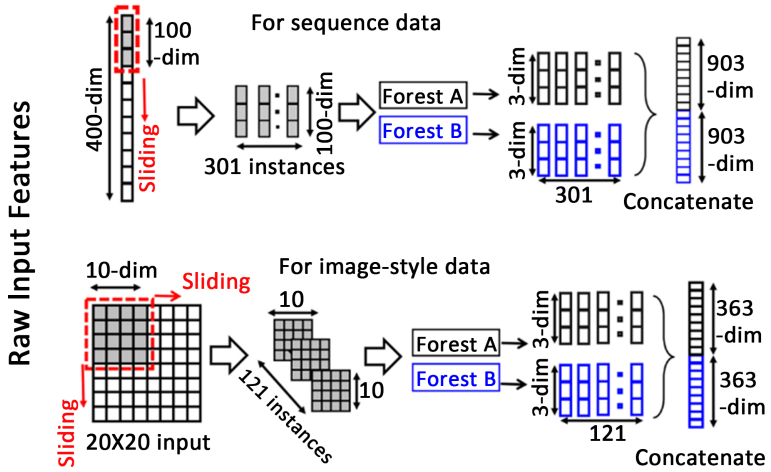


**Figure 3.** Multi-grained scanning
**图 3.** 多粒度扫描结构图

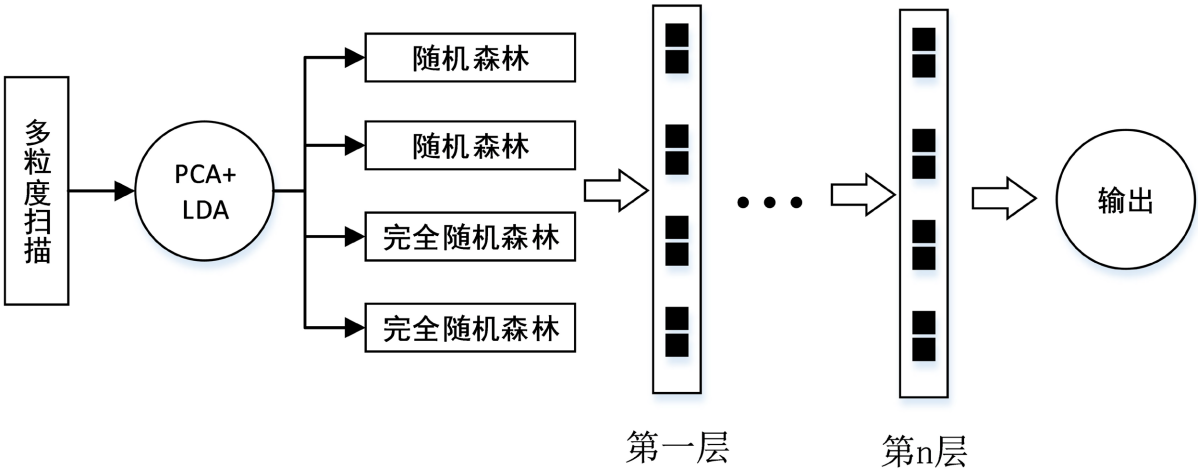


**Figure 4.** Cascaded forest structure map based on mixed feature extraction
**图 4.** 基于混合特征提取的级联森林结构图

## 4. 实验结果与分析

### 4.1. 评估方法

本文采用混淆矩阵来评估分类结果，表 3 为二分类问题的混淆矩阵。常见的分类检测评价指标有准确率、精确率、F1 值、召回率和 ROC 曲线。一般来说，准确率可以反映整体的情况，但是对于不平衡数据这种偏向性比较严重的数据，我们一般会更关注于精确率和召回率，然而在有些特殊情况下，假如产生了两个数值一个过高一个过低的结果，可能难以判别，所以综合这两个数值，参考 F1 值来判断模型的性能可能显得更加准确。

**Table 3.** Binary confusion matrix
**表 3.** 二分类混淆矩阵

| 真实值 | 预测值 | |
|---|---|---|
| | 正类 | 负类 |
| 正类 | TruePositive (TP) | FalseNegative (FN) |
| 负类 | FalsePositive (FP) | TrueNegative (TN) |

准确率——预测正确的样本占总数的比例。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

召回率/敏感性——Sen/Rec 的值越大，说明越好，异常样本会判断为异常的可能性越大，漏检概率越小。

$$\text{Rec} = \text{Sen} = \frac{TP}{TP + FN} \tag{2}$$

精确率——查准率，即正确预测为正的占全部预测为正的比例。

$$\text{Pre} = \frac{TP}{TP + FP} \tag{3}$$

F1 为算术平均数与几何平均数的比值，越大越好。

$$F1 = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{4}$$

ROC 曲线是反应敏感性和特异性连续变量的综合指标，提供不同实验之间在共同标尺下的直观比较，ROC 曲线越凸、越接近左上角表明其诊断价值越大，利于不同指标间的比较，曲线下面积可评价诊断准确性。

## 4.2. 讨论与分析

### 4.2.1. 极少数据类生成前后对比

入侵检测数据集存在着很明显的不平衡性，尤其是有部分数据类非常稀少，甚至只有个位数到两位数。将入侵检测数据集中数量最少的几类数据类结合起来称为为 Attack 类，图 5 显示了训练数据中 Attack 类和 Normal 类的特征关系二维图，从图中可以明显看出数据分布极其不平衡。Normal 类的数据几乎占据了全部，Attack 类只有零星几点。

表 4 为 gcForest、GAN、随机森林(RF)和 Logistic Regression (LR)四种分类模型在部分原数据集上的性能对比，可以看出分类模型对少数类的检测还是很差，甚至有的分类精度只有 20%，在 gcForest 模型上达到了 50%左右和 GAN 模型相当，已经比其他算法高了很多。

为此，本文提出的基于改进后的 GAN 算法来处理数据，并采用 SMOTE 算法和 NearMiss 算法对数据进行调整优化，处理之后的特征关系二维图如图 6 所示。很明显看出代表 Attack 类的小点比之前多了很多，且分布大多数都比较聚集，说明数据比较相似，生成效果较好。

数据优化后再次对比四种模型，结果如表 5 所示，明显看出对于 Attack 这个少数类，每一种模型的检测效果都有了很大的提升，从原本的不到 50%到后续的 70%~80%，GAN 模型的提升反而没有特别大，说明 GAN 模型对于数据的分类没有很擅长，相反，RF 算法在少数类的数据上提升非常大，说明 RF 算法适合用于分类检测，只是对于数据的平衡性要求很高，一旦数据不好，分类的性能就会受到很大的影响。同样作为树结构的 gcForest 模型分类精度也有了比较明显的提升，对比其他算法在多数类和少数类数据上都得到了比较好的检测水平，证明了 gcForest 模型在分类检测方面的巨大优势。

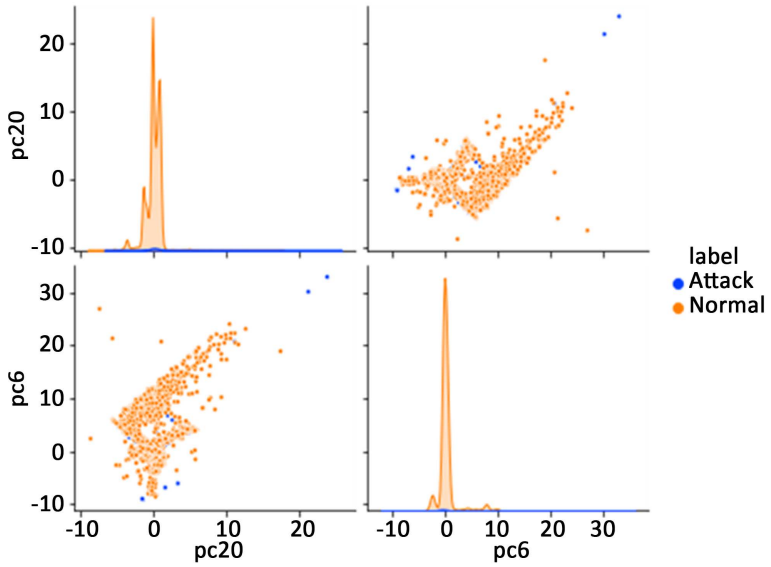


**Figure 5.** Distribution before sampling
**图 5.** 采样前分布图

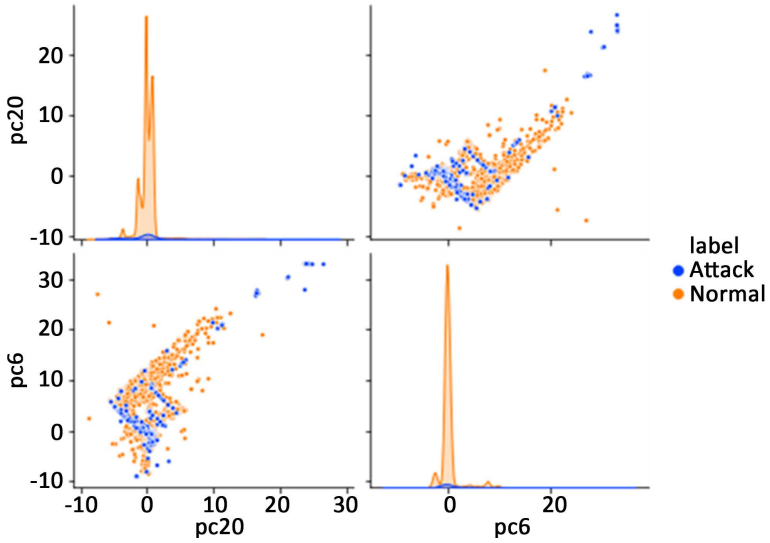


**Figure 6.** Distribution after sampling
**图 6.** 采样后分布图

**Table 4.** Comparison table before sampling
**表 4.** 采样前对比表

| 方法 | Sen/Rec | | Pre | | F1 | |
|---|---|---|---|---|---|---|
| | Normal | Attack | Normal | Attack | Normal | Attack |
| RF | 0.99 | 0.28 | 0.98 | 0.42 | 0.99 | 0.33 |
| LR | 0.95 | 0.21 | 0.98 | 0.38 | 0.97 | 0.27 |
| GAN | 0.99 | 0.52 | 0.99 | 0.58 | 0.99 | 0.56 |
| gcForest | 0.98 | 0.53 | 0.99 | 0.58 | 0.99 | 0.56 |

**Table 5.** Comparison table after sampling
**表 5.** 采样后对比表

| 方法 | Sen/Rec | | Pre | | F1 | |
|---|---|---|---|---|---|---|
| | Normal | Attack | Normal | Attack | Normal | Attack |
| RF | 0.99 | 0.86 | 0.98 | 0.88 | 0.99 | 0.89 |
| LR | 0.96 | 0.75 | 0.98 | 0.80 | 0.97 | 0.77 |
| GAN | 0.99 | 0.79 | 0.99 | 0.80 | 0.99 | 0.79 |
| gcForest | 0.99 | 0.87 | 0.99 | 0.93 | 0.99 | 0.90 |

**Table 6.** Comparison table of classification models
**表 6.** 分类模型对比表

| 方法 | Sen/Rec | Pre | F1 |
|---|---|---|---|
| RF | 0.90 | 0.93 | 0.91 |
| LR | 0.82 | 0.85 | 0.83 |
| GAN | 0.93 | 0.94 | 0.93 |
| gcForest | 0.98 | 0.96 | 0.97 |

#### 4.2.2. 与其他模型性能对比

实验对比选用 gcForest、GAN、随机森林(RF)和 LogisticRegression (LR)四种分类模型，在处理后的完整数据集上进行训练，训练完成后在测试集上进行验证，实验结果如表 6 所示。

#### 4.2.3. 与其他数据集的对比

为了充分评估和验证所提出的方法，本文选用在三个不同的入侵检测数据集上进行实验，分别是 KDD99 数据集，NSL-KDD 数据集，UNSW-NB15 数据集。

KDD99 数据集是一个用来从正常连接中监测非正常连接的数据集。该数据集的标签主要分为五大类，分别是 a. 正常记录，b. 拒绝服务攻击(DOS)，c. 探针攻击(Probing)，d. 来自远程机器的攻击(R2L)，e. 普通用户对于超级管理员用户权限的非法访问(U2R)。实验采用随机抽取的 25%的 KDD99 数据集，数据集中共有 123,530 条网络连接记录，每条网络连接被标记为正常(normal)或异常(attack)，异常类型共细分为 22 种攻击类型。NSL-KDD 数据集是 KDD99 的改进版，数据集每条记录包含 43 个特征，其中 41 个特征指的是流量输入本身，最后两个是标签(正常或攻击)和分数(流量输入本身的严重性)。数据集中存在 4 种不同类型的攻击：拒绝服务(DoS)、探测、用户到根(U2R)和远程到本地(R2L)。UNSW-NB15 数据集中一共有 9 种攻击：Fuzzers，Analysis，Backdoors，DoS，Exploits，Generic，Reconnaissance，Shellcode 和 Worms，每一种攻击有 49 个特征。

实验表明，在三种数据集上对比其他几种算法，gcForest 在分类检测性能上都更具有优势，而且三种数据集对比下来，可以看出 NSL-KDD 数据集在数据质量上比起另外的数据集更加好(表 7)。

**Table 7.** Comparison table of different datasets
**表 7.** 不同数据集的对比表

| 数据集 | 方法 | Sen/Rec | Pre | F1 |
|---|---|---|---|---|
| KDD99 | RF | 0.83 | 0.87 | 0.88 |
| | LR | 0.69 | 0.71 | 0.70 |
| | GAN | 0.91 | 0.92 | 0.91 |
| | gcForest | 0.94 | 0.96 | 0.92 |
| NSL-KDD | RF | 0.90 | 0.93 | 0.92 |
| | LR | 0.82 | 0.85 | 0.83 |
| | GAN | 0.93 | 0.94 | 0.93 |
| | gcForest | 0.98 | 0.96 | 0.97 |
| UNSW-NB15 | RF | 0.88 | 0.89 | 0.88 |
| | LR | 0.75 | 0.75 | 0.75 |
| | GAN | 0.90 | 0.92 | 0.91 |
| | gcForest | 0.92 | 0.94 | 0.93 |

## 5. 结语

本文提出了一种基于改进的深度森林算法的入侵检测数据检测技术，结合了基于 GAN 的不平衡数据处理方法和基于 PCA 和 LDA 的特征提取方法。实验对比了深度森林、GAN、随机森林(RF)和 Logistic Regression (LR)分类模型。实验结果表明，在进行了不平衡数据集的处理后，不仅保证了多数类数据的高精确率，同时对少数类数据的分类检测效果也有着非常显著的提升。此外，深度森林分类模型与其

他几个模型相比，不管是哪一类数据都拥有了更高的检测水平，性能更佳。随着攻击方法的不断发展，入侵检测技术也需要与时俱进，如何将深度学习和入侵检测系统更好地融合将会是未来研究的一大重点。

## 基金项目

国家自然科学基金(No. 62076042, 62102049)，四川省自然科学基金项目(No. 2022NSFSC0535)，四川省科技厅重点研发项目(No. 2021YFSY0012, 2021YFG0332)，成都市科技局技术创新研发项目(No. 2021-YF05-02424-GX)，四川省量子安全通信创新团队(No. 17TD0009)。

## 参考文献

[1] Fernandes Jr., G., Rodrigues, J.J.P.C., Carvalho, L.F., Al-Muhtadi, J.F. and Proença Jr., M.L. (2019) A Comprehensive Survey on Network Anomaly Detection. *Telecommunication Systems*, **70**, 447-489. https://doi.org/10.1007/s11235-018-0475-8

[2] Rahman, M.A., Shahriar, M.H. and Masum, R. (2019) False Data Injection Attacks against Contingency Analysis in Power Grids: Poster. *Proceedings of the* 12*th Conference on Security and Privacy in Wireless and Mobile Networks*, Miami, 15-17 May 2019, 343-344. https://doi.org/10.1145/3317549.3326323

[3] Lee, P.H. (2014) Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets. *International Journal of Environmental Research and Public Health*, **11**, 9776-9789. https://doi.org/10.3390/ijerph110909776

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. MIT Press, Cambridge.

[5] Kocher, G. and Kumar, G. (2021) Machine Learning and Deep Learning Methods for Intrusion Detection Systems: Recent Developments and Challenges. *Soft Computing*, **25**, 9731-9763. https://doi.org/10.1007/s00500-021-05893-0

[6] Zhou, Z.-H. and Feng, J. (2019) Deep Forest. National Science Review, **6**, 74-86. https://doi.org/10.1093/nsr/nwy108

[7] Yong, S. and Feng, L. (2106) SMOTE-NCL: A Re-Sampling Method with Filter for Network Intrusion Detection. 2016 2*nd IEEE International Conference on Computer and Communications* (*ICCC*), Chengdu, 14-17 October 2016, 1157-1161. https://doi.org/10.1109/CompComm.2016.7924886

[8] Yan, B.H., Han, G.D., Sun, M. and Ye, S. (2017) A Novel Region Adaptive SMOTE Algorithm for Intrusion Detection on Imbalanced Problem. 2017 3*rd IEEE International Conference on Computer and Communications* (*ICCC*), Chengdu, 13-16 December 2017, 1281-1286. https://doi.org/10.1109/CompComm.2017.8322749

[9] Belenko, V., Chernenko, V., Kalinin, M. and Krundyshev, V. (2018) Evaluation of GAN Applicability for Intrusion Detection in Self-Organizing Networks of Cyber Physical Systems. 2018 *International Russian Automation Conference* (*RusAutoCon*), Sochi, 9-16 September 2018, 1-7. https://doi.org/10.1109/RUSAUTOCON.2018.8501783

[10] Lee, J. and Park, K. (2021) GAN-Based Imbalanced Data Intrusion Detection System. *Personal and Ubiquitous Computing*, **25**, 121-128. https://doi.org/10.1007/s00779-019-01332-y

[11] Liao, D., Huang, S., Tan, Y. and Bai, G. (2020) Network Intrusion Detection Method Based on GAN Model. 2020 *International Conference on Computer Communication and Network Securit*, Xi'an, 21-23 August 2020, 153-156. https://doi.org/10.1109/CCNS50731.2020.00041

[12] Shahriar, M.H., Haque, N.I., Rahman, M.A. and Alonso, M. (2020) G-IDS: Generative Adversarial Networks Assisted Intrusion Detection System. 2020 *IEEE* 44*th Annual Computers, Software, and Applications Conference* (*COMPSAC*), Madrid, 13-17 July 2020, 376-385. https://doi.org/10.1109/COMPSAC48688.2020.0-218

[13] 钱铁云, 王毅, 张明明, 刘俊恺. 基于深度神经网络的入侵检测方法[J]. 华中科技大学学报(自然科学版), 2018, 46(1): 6-10. https://doi.org/10.13245/j.hust.180102

[14] 丁龙斌, 伍忠东, 苏佳丽. 基于集成深度森林的入侵检测方法[J]. 计算机工程, 2020, 46(3): 144-150. https://doi.org/10.19678/j.issn.1000-3428.0053018

[15] 范怡敏, 齐林, 帖云. 基于基因表达小样本数据的级联森林分类模型[J]. 计算机应用与软件, 2020, 37(11): 165-171.

[16] 颜建军, 刘章鹏, 刘国萍, 等. 基于深度森林算法的慢性胃炎中医证候分类[J]. 华东理工大学学报(自然科学版), 2019, 45(4): 593-599.

[17] 蒋鹏飞, 魏松杰. 基于深度森林与 CWGAN-GP 的移动应用网络行为分类与评估[J]. 计算机科学, 2020, 47(1): 287-292.

[18] 张鹏, 李志, 邸希元. 基于深度森林的无线传感器网络故障分类算法[J]. 计算机测量与控制, 2022, 30(1): 26-33. https://doi.org/10.16526/j.cnki.11-4762/tp.2022.01.005

[19] 王耀聃, 李红娇, 詹清钦. 结合堆叠稀疏自编码器与改进深度森林的窃电检测方法[J]. 计算机应用与软件, 2022, 39(12): 64-72+158.