不平衡数据下基于随机森林的网络入侵检测*

卫昱君 马悦凯 王佳悦 白琳

西安邮电大学计算机学院，西安 710121

摘 要 针对现有入侵检测方法通用性低、数据依赖性强和检测率低等问题，本文提出了一种基于不平衡数据分析的入侵检测模型。通过深入分析网络流量的数据特征，以数据的不平衡特性分析为基础，结合随机森林模型从大量的不平衡网络流量数据中学习并提取出有价值的特征，进而实现不平衡数据的分类。首先，采用 Borderline-SMOTE 过采样策略对原始数据进行均衡化处理以增强少数类样本的分类贡献、提高分类器的分类效率。其次，基于随机森林构建入侵检测分类模型以准确识别出攻击行为。最后，在 NSL-KDD 数据集上进行了对比实验，结果表明 Borderline-SMOTE 结合随机森林的模型在准确率、F1-Score 和攻击检测能力方面均优于对比方法，具备较强的实用性和有效性。该方法为解决小流量攻击检测问题提供了一种行之有效的解决方案。

关键字 入侵检测，Borderline-SMOTE，随机森林，NSL-KDD 数据集

Intrusion Detection Based on Random Forest for Imbalanced Data

WEI Yu-jun MA Yue-kai WANG Jia-yue BAI Lin

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
307043370@qq.com

Abstract—Aiming at the problems of low universality, strong data dependence, and low detection rate in existing intrusion detection methods, an intrusion detection model based on imbalanced data is proposed. By analyzing the data characteristics of network traffic and focusing on the imbalanced nature, combined with the random forest model, valuable features are extracted from large amounts of imbalanced network data to enable effective classification. First, the Borderline-SMOTE oversampling strategy is used to balance the original data, enhancing the contribution of minority class samples and improving classification efficiency. Second, a classification model is built using random forests to accurately identify attack behaviors. Finally, experiments are conducted on the NSL-KDD dataset. Results show that the combination of Borderline-SMOTE and random forest outperforms other methods in accuracy, F1-Score, and attack detection capability, demonstrating strong practicality and effectiveness. This method provides an effective solution for detecting small-traffic attacks.

Keywords—Intrusion detection, Borderline-SMOTE, Random Forest, NSL-KDD data set

1 引 言

随着大数据技术的飞速发展，网络流量每日井喷式增长，网络安全风险与日俱增。在众多网络安全技术中，入侵检测作为一种主动、实时的安全防御技术[1]，越来越受到研究者的关注。其中，异常检测[2]通过学习网络流量的正常模式来判断是否存在与其偏离的数据（异常）。其核心是通过数据分析来识别出异常的计算模型。它可以充分利用机器学习算法建模，自动从大量的网络流量中学习有效的特征，从而构建高效、准确的检测模型，逐渐成为入侵检测研究领域的热点。

Sun 等[3]提出了一种基于 Borderline-SMOTE 和特征选择的方法，有效实现了异常检测。Priyadarsini 等[4]面向入侵检测的不平衡数据问题，提出了一种基于人工蜂群和 Borderline-SMOTE 的随机森林模型。Dinesh 等[5]提出了基于 Borderline-SMOTE 与支持向量机的异常检测模型，在减少误报和及时响应方面具有优势。Bagui 等[6]使用 BSMOTE 和 SVM-SMOTE 两种有效的重采样技术用于识别稀有攻击。已有模型虽在一定程度上实现了入侵检测，但是很多方法在面对复杂的网络流量时，依然存在检测率较低的问题，尤其是在处理海量、高维的数据时容易受到噪声和冗余特征的干扰[7]。另外，异常检测通常对数据集的质量和类型有较强依赖性，尤其在面对数据不平衡问题时，传统分类器容易偏向多数类，而忽视少数类攻击样本，极易导致检测偏差[8]。因此，本文面向网络流量数据的不平衡特征，提出一种基于 Borderline-SMOTE 和随机森林的入侵检测方法，以弥补现有检测技术的不足。

首先，采用 Borderline-SMOTE[9]对少数类样本进行过采样，确保分类模型在训练过程中能够充分学习并识别不同类别的流量特征，提升对少数类攻击的检测能力。其次，针对流量数据的高维特征，通过特征

*基金资助：本文得到西安市重点产业链技术攻关项目（23ZDCYJSGG0010-2023）的资助。

工程[10]和降维[11]技术提取核心特征，去除噪声和无关属性。最后，采用随机森林[12]完成分类，利用其集成学习的优势，增强模型泛化能力，使其能够应对复杂的网络环境。

2 本文方法

2．1 Borderline-SMOTE

Borderline-SMOTE（Synthetic Minority Over-sampling Technique）[13]方法是对SMOTE算法的改进，主要关注少数类中的“边界样本”，即更容易被误分类的少数类样本。算法步骤如下：

（1）距离计算

对于少数类样本，首先通过K近邻方法找到与其距离最近的K个样本。距离计算如式（1）所示。

$$dist(X,Y)=\sum_{i=1}^{n}(x_i-y_i)^2 \quad (1)$$

其中，x_i和y_i是两个样本点在n维空间中的坐标，$dist(X,Y)$表示两个样本点的欧氏距离。

（2）少数类样本分类

通过K近邻找到少数类样本最近的邻居，根据其类别分布，将少数类样本划分为三类，见图1。

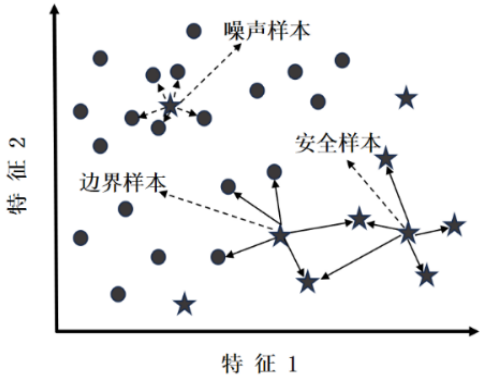

图 1 Borderline-Smote算法原理图

① 安全样本：若K近邻样本中大部分是少数类样本，该样本被认为是安全的。

② 边界样本：若K近邻样本中多数类样本的比例超过一半，则该少数类样本被称为边界样本。是合成新样本的重点对象。

③ 噪声样本：如果该样本的K近邻几乎全部是多数类样本，则认为该样本为噪声，不进行处理。

对边界样本，通过式（2）合成新样本：

$$X_{new}=X_i+\delta\times(X_{\sim i}-X_i) \quad (2)$$

其中，X_i是少数类样本，$X_{\sim i}$是X_i的某个少数类近邻。该公式通过插值合成新的少数类样本。

2．2 随机森林

随机森林（Random Forest, RF）[14]由若干个相互独立的决策树构建产生，使用多数表决的投票机制对测试样本进行投票表决，见图2。

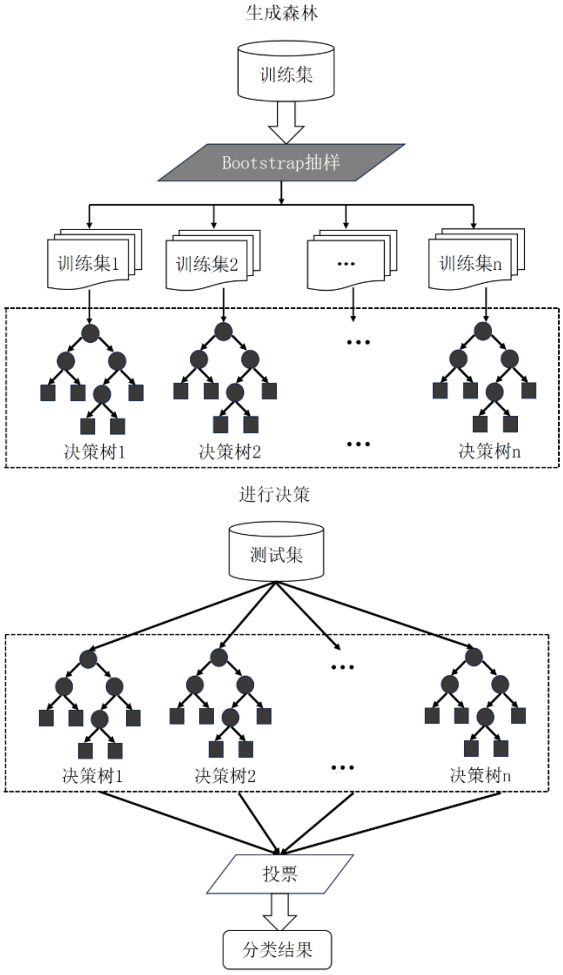

图 2 随机森林投票分类图

算法步骤包括：

（1）计算特征贡献度

随机森林的特征重要性定量为每个特征在每棵树上贡献度的平均值，通过比较不同特征间的贡献度来确定其重要程度。贡献度用平均袋外数据误差表示，具体计算见式（3）和（4）。

$$OobErrorT_i^j=|e_1-e_2| \quad (3)$$

其中，$OobErrorT_i^j$为特征j对RF中树i的袋外数据误差，e_1为用袋外数据样本得到的误差，e_2为随机打乱袋外数据中的第j列后得到的误差。

平均袋外数据误差：

$$MOET^j=\frac{\sum_{i=1}^{T}OobErrorT_i}{T} \quad (4)$$

其中，$MOET^j$为特征j的平均袋外数据误差，T为树的数量。对袋外数据第j列随机打乱的方法为：通过排列的方式将原来所有样本的第j个特征值重新打乱分布。

用平均袋外数据误差来刻画特征j的重要性。依据是：若一个特征很重要，其波动会非常影响测试的误差，若测试误差改变幅度较小，则说明特征j不重要。

（2）计算特征权值

特征权值反映了样本特征的重要程度占比，将每个样本的特征权值表示为该特征的平均袋外误差与所有特征平均袋外误差之和的比，如公式（5）所示：

$$weight_j = \frac{MOET^j}{\sum_{k=1}^{D} MOET^k} \quad (5)$$

其中，$weight_j$为特征j的权值，D为特征总数。

（3）随机森林训练

假设数据的训练集样本量为N，在训练随机森林子树时，有放回地随机从训练集中抽取N个训练样本，用于子树的训练子集；假设每个样本的特征维度为M，随机不重复地抽取m个特征（m<M）组成新的训练子集来生成一棵决策树，每次树进行分裂时，可依据袋外错误率实现最优m的选择；重复以上步骤，得到k棵决策树，组合形成随机森林分类器[15]，最终采用投票的方式来确定样本的最终分类结果。

2.3 基于 Borderline-SMOTE 的随机森林模型

受类别不平衡问题影响，少数类样本会影响分类器的性能。本文通过引入Borderline-SMOTE过采样算法合成更多的边界少数类样本，从而平衡数据集。以便提升随机森林对少数类的识别能力。基于Borderline-SMOTE的随机森林模型见图3。

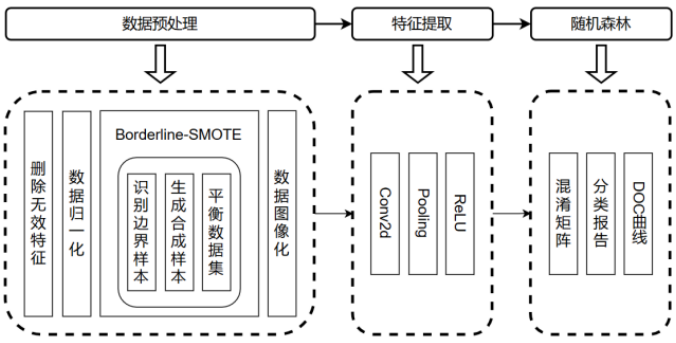

图 3 基于 Borderline-SMOTE 的随机森林分类模型

3 入侵检测模型构建

3.1 数据集

实验采用NSL-KDD标准数据集[16]。其包含43个特征（数值型和离散型），是经典入侵检测数据集KDD99的改进版本。该数据集的类别标签包括“normal”和具体攻击。攻击类型总体分为4大类，具体见表1。

表 1 NSL-KDD 数据集的四类攻击及具体攻击行为

攻击类型	具体攻击行为
拒绝服务攻击（DoS）	如“Neptune”、“smurf”
探测攻击（Probe）	如“satan”、“ipsweep”
用户到根攻击（U2R）	如“buffer_overflow”、“rootkit”
远程到本地攻击（R2L）	如“guess_passwd”、“ftp_write”

实验的NSL-KDD数据集包含两类文件，见表2。

表 2 NSL-KDD 数据集文件和内容

数据集文件	内容
KDDTrain+	训练集文件，包含正常流量和多种攻击类型的样本
KDDTest+	测试集文件，用于评估模型性能，包含和训练集类似的样本

3.2 数据预处理

预处理包括：标签处理、特征工程与编码及归一化与均衡化。

（1）通过读取攻击类型映射文件，将攻击类型按类别进行映射，将“attack_type”列转换为二分类标签，其中正常流量为0，攻击流量为1。

（2）原始数据集包含42个属性字段，包括标签“label”、攻击“attack_type”以及中间预测结果“success_pred”。考虑到“attack_type”作为攻击行为的文本标签可能导致信息泄露，“success_pred”为后验信息，与模型训练无关，这两列在特征工程阶段被移除。再剔除标签列后，剩余39个字段被视为初始的流量特征用于后续建模。这些特征涵盖了基本连接特征（如协议类型、服务类型）、内容特征（如登录尝试次数）以及统计特征，从多个维度刻画了网络流量的行为模式。

（3）为精简特征维度，采取基于主成分分析（PCA）的分类特征降维方法。以便充分考虑数据集中不同类型特征在数据分布和特征表达上的差异，避免在降维过程中出现主成分混合导致的判别信息丢失问题。具体地，将全部数值型输入特征按其语义类别划分为多个子集，包括基本连接子集、内容行为子集和统计特征子集等。随后，对每个子集分别进行Z-score标准化，并单独应用PCA进行降维。根据每个子集的特征维度和信息保留需求，分别设置合适的累计方差贡献率阈值（95%），最终从39个特征中提取出若干组局部主成分，再将其拼接为统一的低维特征向量。最终保留的降维结果包含融合后的16维主成分特征，在保持特征语义结构的基础上，有效提升了数据表达能力和模型

训练效率。

（4）针对分类特征“protocol_type”、“service”和“flag”，分离出离散特征并进行独热编码[17]。为了确保训练集和测试集中的特征列一致，使用“align”函数对齐这些特征，并将编码后的数据与原始数值特征合并。然后通过填充缺失值保证数据完整性，见图4。

（5）为避免不同特征取值范围过大而导致模型训练过程中权重不平衡，使用最小最大缩放（Min-Max Scaling），将数值型特征归一化到[0,1]区间。

（6）采用Borderline-SMOTE算法对少数类样本进行边界过采样，实现数据集的均衡。

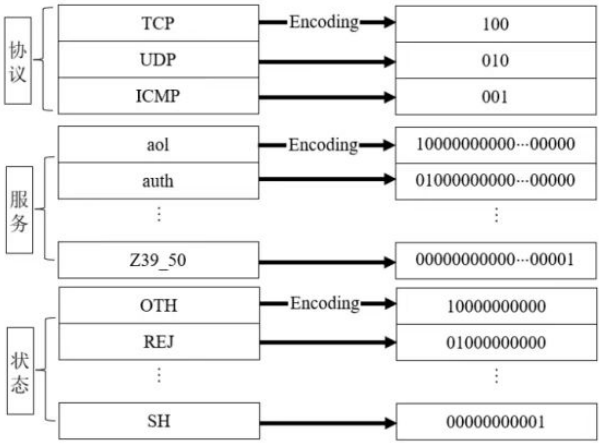

图 4 独热编码-数值化处理

数据预处理后的样本规模见表3。

表 3 数据预处理后的样本规模

类别	样本数（训练集）	样本数（测试集）
normal	67343	9711
neptune	41214	4657
satan	3633	735
ipsweep	3599	141
portsweep	2931	157
smurf	2646	665
nmap	1493	73
back	956	359
teardrop	892	12
warezclient	890	0
pod	201	41
...	...	...
udpstorm	0	2

3．3 实验设置

训练样本总数为125973，测试样本总数为22544。训练集中共有67343条正常流量和58630条攻击；测试集中包含9711条正常流量和12833条攻击。通过Borderline-SMOTE过采样后，训练集的正负类样本达到了平衡，每类均有67343个样本。

通过对训练集的预处理和分类建模，构建出二分类模型以区分正常流量和攻击。模型训练中，随机森林分类器设置101棵决策树[18]并以“entropy”作为划分标准。另外，采用K-fold交叉验证（K=10）来进一步评估模型性能，计算各折交叉验证的准确率，并取均值以获取更稳健的模型性能评估。

3．4 模型评估

采用准确率、精确率、检测率、误判率、F1-score[19]指标对模型进行评估。定义攻击数据预测正确的数目为TP，预测错误数为FN，正常数据预测正确的数目为TN，预测错误数为FP；定义N_P为实际攻击数据总数，则有$N_P = TP + FN$，N_n为实际正常数据总数，则有$N_n = FP + TN$。具体如下：

（1）准确率（Accuracy, ACC）

预测正确的测试数据数占测试数据总量的百分比。

$$ACC = \frac{TP+TN}{N_p+N_n} \quad (6)$$

（2）精确率（Precision, PRE）

被检测为攻击的数据中真正攻击数据数量占比。

$$PRE = \frac{TP}{TP+FP} \quad (7)$$

（3）检测率（召回率）（Recall, REC）

测试集中所有攻击数据被正确检测出来的数量。

$$REC = \frac{TP}{N_p} \quad (8)$$

（4）F1-score

是精确率和召回率的调和均值，可评估模型的综合性能。

$$F1 = 2 \times \frac{PRE \times REC}{PRE+REC} \quad (9)$$

（5）误判率（False Positive Rate, FPR）

为正常数据中被误判为异常的比例。

$$FPR = \frac{FP}{N_n} \quad (10)$$

3．5 实验结果及分析

（1）分类准确率与混淆矩阵

本文模型的总体准确率为81.2%，表明模型能够较好地区分正常流量和攻击。混淆矩阵见图5和表4所示。

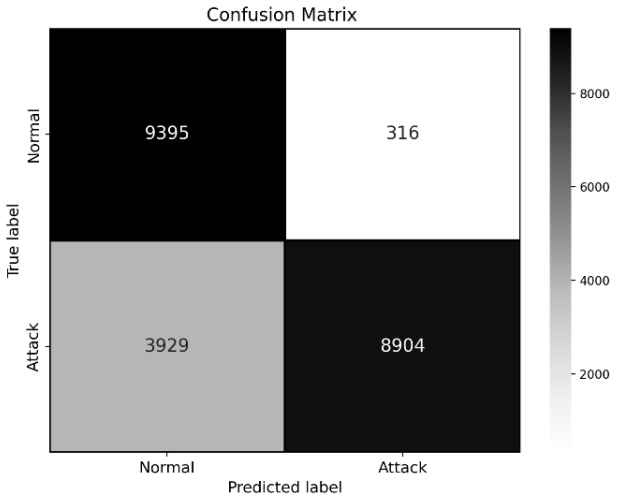

图 5 混淆矩阵

在测试集中，模型正确识别了9395条正常流量和8904条攻击，误将316条正常流量预测为攻击，将3929条攻击预测为正常流量。模型对正常流量的预测更准确，但在攻击流量的识别中仍有一定误差。

表 4 测试集的混淆矩阵

实际类别\预测类别	正常	攻击
正常	9395	316
攻击	3929	8904

（2）ROC曲线与AUC值

ROC曲线见图6。AUC（Area Under Curve）值为0.96，表明模型在不同分类阈值下的表现非常优秀，模型能够很好地区分正常流量与攻击。

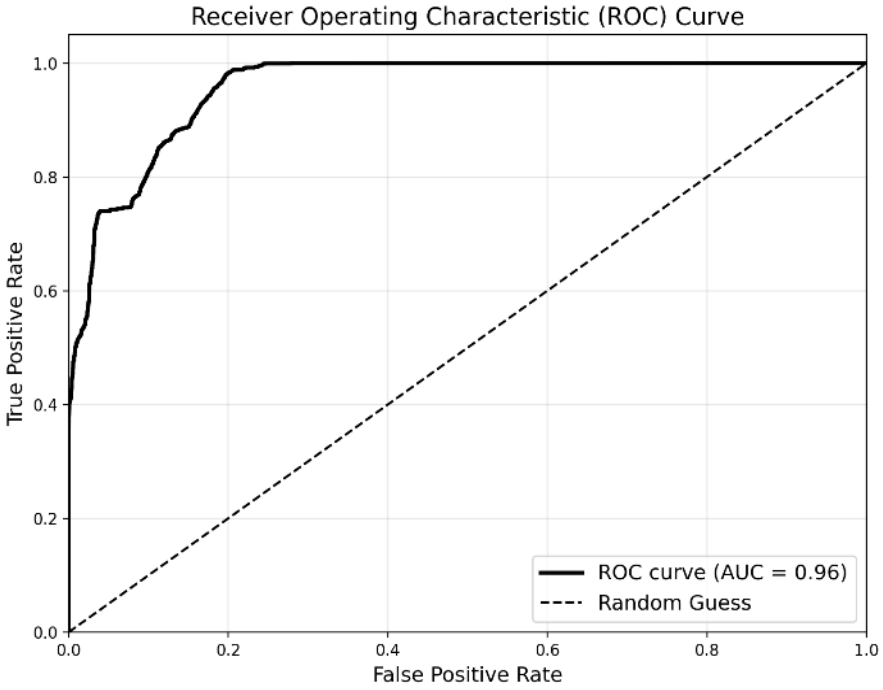

图 6 ROC 曲线图

（3）精确率、召回率及F1-score

表5呈现了3个评估指标，其中，精确率Precision表示模型预测的某类中预测正确的数量。攻击类的Precision高达0.966，说明模型预测攻击的准确率非常高。召回率Recall表示该类别中有多少被正确预测。正常类的Recall为0.968，攻击类的Recall为0.687，表明模型在识别正常流量上表现更好，但攻击流量的召回率相对较低。这是因为对R2L入侵的检测效果不理想，很多R2L入侵是伪装合法用户身份进行攻击的，使得其特征与正常数据包类似，造成了检测的困难。F1-score是Precision和Recall的加权平均值，反映了模型在两类数据上的整体平衡性。

表 5 模型评估指标结果

类别	Precision	Recall	F1-Score	Support
正常	0.705	0.967	0.816	9711
攻击	0.966	0.694	0.808	12833

（4）误判率

FPR是衡量模型性能的重要指标，可以帮助分析模型在分类正负样本时的错误情况。在入侵检测任务中，误判可能带来很大的风险。其表示模型错误地将负样本（正常流量）预测为正样本（攻击）的比例。我们希望FPR越低越好，因为较高的假阳性率可能导致正常操作被误报为攻击，从而影响正常业务流程。本模型FPR为3.3%，说明在正常的流量中，有3.3%被误判为攻击，低FPR有助于减少误报、降低运维负担。

3.6 对比实验及结果

为验证本文方法的有效性，以随机森林（RF）、KNN、决策树（DT）以及卷积神经网络（CNN）作为分类器进行比较。所有分类模型在训练前均进行相同的预处理流程，以确保实验的公平性。数据均衡处理，大多方法采用Borderline-SMOTE算法，KNN方法中使用了SMOTE算法，也在相同条件下进行，以探究不同采样方式的效果。

所有分类器使用相同的训练集和测试集划分（7:3），并在相同的随机种子初始化下进行训练。

各模型超参数设定为：

（1）RF：树数为160、最大深度为8、分裂节点的最小样本数为4、叶子节点的最小样本数为2、分裂准则选择信息增益，以增强特征选择的判别性。

（2）KNN：邻居数为5、距离度量采用欧式距离。

（3）DT：使用Gini系数、最大深度为10。

（4）CNN：采用双卷积块和全连接层结构，输入特征向量重构为一维张量（reshape为[batch_size, 16, 1]），模型包含两层卷积层，每层后接批归一化、最大池化和Dropout（比例0.3），卷积层之后是全连接层、128个神经元，使用Dropout（比例0.5）防止过拟合、输出层为单一节点，使用Sigmoid激活函数进行二分类、优化器为Adam，学习率设为0.001、训练轮次为50。通过上述一致的预处理流程与规范的参数设置，保证了不同方法在相同实验条件下进行公平对比。实验结果对比具体见表6和图7。

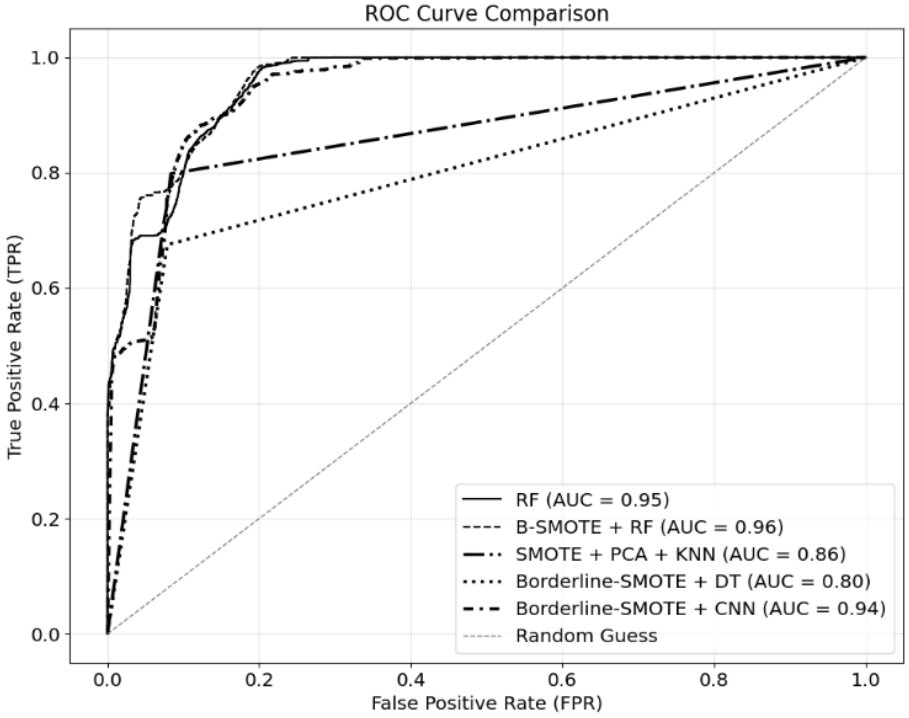

图 7 不同方法的 ROC 曲线比较图

图7给出了ROC曲线对比，表6给出了各算法的准确率及评估指标对比。由结果可见，对于随机森林分类器，经过Borderline-SMOTE均衡处理后，整体预测性能提升理想。Borderline-SMOTE+RF在准确率和AUC值均表现最佳，显示出该方法在入侵检测任务中的明显优势。此外，在所有模型中，本文的Borderline-SMOTE+RF各指标值最高，性能最佳。其中，少数类的精确率和召回率分别为0.966和0.694，F1-Score为0.808。

这表明本文算法不仅能够精准识别正常流量数据，还具备较强的少数类入侵检测能力。

SMOTE+KNN 、 Borderline-SMOTE+DT 以 及 Borderline-SMOTE+CNN在少数类F1-Score上的结果较RF方法的有所提升，说明其对于攻击的检测率有所

提高。Borderline-SMOTE+DT可以以更小的内存占用和更小的计算代价运行[20]，但在处理复杂模式时能力有限。前两种方法只适合相对简单的应用场景。Borderline-SMOTE+CNN的攻击检测率达到了0.672，表现出了良好的检测能力，但在计算资源和训练时间方面的要求较高，不利于实时的入侵检测场景应用。相较于CNN等深度学习模型，Borderline-SMOTE+RF的训练时间更短，对计算资源的需求也更低，展现出了良好的效率和实用性。这使得该方法在保证高性能的同时，兼顾了资源利用效率和可操作性，适合部署在资源有限的实际网络安全系统中，可用性更高。

表 6　本文方法与对比方法各评估指标结果对比

		RF	Borderline-SMOTE +RF	SMOTE+PCA+KNN	Borderline-SMOTE +DT	Borderline-SMOTE +CNN
Accuracy		0.750	**0.812**	0.780	0.779	0.781
Precision	0	0.638	**0.705**	0.680	0.679	0.681
	1	0.966	**0.966**	0.922	0.921	0.922
Recall	0	**0.973**	0.967	0.925	0.924	0.925
	1	0.582	**0.694**	0.670	0.670	0.672
F1-Score	0	0.771	**0.816**	0.784	0.783	0.784
	1	0.726	**0.808**	0.776	0.775	0.778
Macro Avg	Precision	0.802	**0.835**	0.801	0.800	0.801
	Recall	0.777	**0.831**	0.798	0.797	0.799
	F1-Score	0.748	**0.812**	0.780	0.779	0.781
Weighted Avg	Precision	0.825	**0.853**	0.818	0.817	0.818
	Recall	0.750	**0.812**	0.780	0.779	0.781
	F1-Score	0.745	**0.811**	0.779	0.779	0.781

通过实验可以看出，针对不平衡数据问题，Borderline-SMOTE算法在少数类样本生成上的有效性为分类模型性能的提升带来了优势。而在具体分类器的选择上，随机森林因其良好的鲁棒性和泛化能力，在结合Borderline-SMOTE后能够更好地平衡准确率与少数类检测性能。相比之下，KNN和DT虽然在某些评价指标上表现出色，但整体性能不及随机森林；CNN虽具备强大的建模能力，但其训练成本较高，实际应用时需要权衡计算资源与性能提升的需求。

综上所述，Borderline-SMOTE 与随机森林的结合是一种高效且实用的入侵检测方法，不仅能够有效解决数据不平衡问题，还在少数类检测精度和整体性能上表现出色。该方法的研究成果为入侵检测领域的实际应用提供了重要参考，未来可以结合特定场景的需求进一步优化采样策略或随机森林参数设置，以提升模型性能。此外，对于计算资源较为充足的环境，可以探索 Borderline-SMOTE 与深度学习模型的结合，以进一步提升复杂入侵行为的检测能力。

4　结束语

本文提出了一种基于Borderline-SMOTE与随机森林的网络入侵检测方法。该方法以数据为驱动，以不平衡数据处理为突破口，采用Borderline-SMOTE均衡化的数据集来提升随机森林分类模型的性能，使其能够自动从大量网络流量数据中学习并提取有效特征来指导分类过程，以此构建出了有效的入侵检测模型。本文的方法在多项评价指标上均表现良好，尤其在少数类攻击检测方面表现优异。其中，准确率达到0.812，少数类的精确率和召回率分别为0.966和0.694，F1-Score（检测率）为0.808，误判率仅0.033，为复杂网络环境下的入侵检测提供了有价值的解决思路，更为解决小流量攻击检测问题提供了一种有效的解决方案。

参 考 文 献

[1] 邓淼磊, 阚雨培, 孙川川, 等. 基于深度学习的网络入侵检测系统综述[J]. 计算机应用, 2025,45(02): 453-466.

[2] 杨宏宇, 朱丹, 谢丰, 等. 入侵异常检测研究综述[J]. 电子科技大学学报, 2009, 38(05): 587-596.

[3] Sun Y, Que H, Cai Q, et al. Borderline SMOTE algorithm and feature selection-based network anomalies detection strategy[J]. Energies, 2022, 15(13): 4751.

[4] Priyadarsini P I. ABC-BSRF: Artificial Bee colony and borderline-SMOTE RF algorithm for intrusion detection system on data imbalanced problem[C]//Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2020. Springer Singapore, 2021: 15-29.

[5] Dinesh M, Sabarish C S, Yogeshwaran S. Network Anomaly Detection Using Borderline SMOTE Algorithm and Support Vector Machines[C]//2024 5th IEEE Global Conference for Advancement in Technology (GCAT). IEEE, 2024: 1-5.

[6] Bagui S S, Mink D, Bagui S C, et al. Determining resampling ratios using bsmote and svm-smote for identifying rare attacks in imbalanced cybersecurity data[J]. Computers, 2023, 12(10): 204.

[7] 张璐, 胡静, 王旭. 基于大数据分析的网络入侵检测与防御技术研究[J]. 网络安全和信息化, 2024, (08): 132-134.

[8] 王瀚召. 机器学习方法在网络异常检测中的研究与应用[D]. 北京邮电大学, 2023.

[9] 马贺, 宋媚, 祝义. 改进边界分类的 Borderline-SMOTE 过采样方法[J]. 南京大学学报(自然科学), 2023, 59(06): 1003-1012.

[10] 蒋家伟. 机器学习势模型优化研究：自适应数据采样与集成特征选择[D]. 太原理工大学, 2023.

[11] 王晓霞, 孙德才, 唐耀庚. 改进的入侵检测数据降维方法[J]. 计算机工程与应用, 2011, 47(25): 85-88.

[12] 刘琦. 基于机器学习的网络入侵检测算法研究[D]. 大连交通大学, 2022.

[13] 杨毅, 卢诚波, 徐根海. 面向不平衡数据集的一种精化 Borderline-SMOTE 方法[J]. 复旦学报(自然科学版), 2017, 56(05): 537-544.

[14] 王奕森, 夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术, 2018, 12(01): 49-55.

[15] 王立柱, 吴品康, 王秋萍. 随机森林模型参数寻优算法比较分析[J]. 沈阳师范大学学报(自然科学版), 2024, 42(05): 420-426.

[16] 王加梁. 基于属性分类建模的入侵检测方法[J]. 计算机工程与设计, 2022, 43(04): 907-913.

[17] 杨寒雨, 赵晓永, 王磊. 数据归一化方法综述[J]. 计算机工程与应用, 2023, 59(03): 13-22.

[18] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(03): 32-38.

[19] 宋亚统. 机器学习算法评估实战[M]. 人民邮电出版社: 202105. 261.

[20] 周祖灏, 李文鑫, 黄培峰, 等. 基于分类算法的古代玻璃制品成分分析与鉴别模型[J]. 计算机技术与教育学报, 2023, 11(03), P44-49.