计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于集成深度森林的入侵检测方法
作者：丁龙斌，伍忠东，苏佳丽
DOI：10.19678/j.issn.1000-3428.0053018
网络首发日期：2019-04-23
引用格式：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/OL]．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0053018

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2019-04-23 10:13:09
网络首发地址：http://kns.cnki.net/kcms/detail/31.1289.TP.20190422.1420.001.html

基于集成深度森林的入侵检测方法

丁龙斌[1] 伍忠东[1] 苏佳丽[1]
(1兰州交通大学电子与信息工程学院信息安全实验室，兰州 甘肃 730070)

摘要：针对现有基于深度学习的入侵检测算法模型训练时间过长、超参数较多、数据需求量大的缺陷，本文提出一种基于集成深度森林(Ensemble Deep Forests, EDF)的入侵检测算法. 本文首先类比卷积神经网络(CNN)隐藏层结构和集成学习的Bagging集成策略构造随机森林层，对每层中RF输入随机选择的特征训练，然后将输出的类向量和特征向量拼接向下层传递迭代，持续训练直到模型收敛，最后使用NSL-KDD数据集进行实验，比较EDF和CNN入侵检测算法性能. 实验结果显示EDF算法比CNN收敛速度提升了50%以上，而分类准确率几乎相同，表明EDF入侵检测算法是一种高效可行的算法，提升了入侵检测算法的性能.

关键词： 入侵检测；卷积神经网络；深度学习；随机森林；深度森林；

Intrusion detection method based on Ensemble Deep Forests

Ding Longbin[1] *Wu Zhongdong*[1] *Su Jiali*[1]
(1 School of Electronic Information, Lanzhou Jiaotong University, Gansu 730070, China)

Abstract: Aiming at the shortcomings of the existing deep learning algorithm model, such as too long training time, long-consuming with hyper-parameters and large data demand, this paper proposes an Ensemble Deep Forests (EDF) method to solve them. The algorithm first constructs a multi-level cascaded random forest model, Each RF inputs different size of features, then the output class vector and the input feature vector are spliced and transmitted to the lower layer, keeping trains until the model converges. Finally, the simulation is carried out on the NSL-KDD dataset. The results show that the convergence speed of the EDF is 50% higher than that of the CNN, and the accuracy is consistent, which indicates that the EDF algorithm is an effective and feasible method, and improves the performance of the Intrusion Detection Method.

Keywords: Intrusion Detection; CNN; Deep Learning; Random Forest; Deep Forests;
DOI:10.19678/j.issn.1000-3428.0053018

1 概述

入侵检测技术是网络安全的重要组成部分，它通过对网络上的各种信息进行收集和分析来检测各种入侵行为，是维护网络安全的重点. 随着网络的普及和网络速率的提升，网络攻击行为日益增加，攻击方法不断更新，传统的智能化检测技术很难取得期望的成效[1].

近些年来，由于深度学习在分类任务、回归学习等方面的优越性能[2]，基于深度学习的入侵检测算法不断提出. 传统的深度模型多是全连接网络，该类网络较深时，具有参数多、耗时长、易过拟合等缺陷；而基于卷积神经网络(CNN)的入侵检测方法，与传统学习模型如 SVM、决策树、k-近邻等传统机器学习算

基金项目：甘肃省高等学校协同创新团队（编号 2017C-09）、兰州市科技局科技项目（编号：2018-1-51）

作者简介：丁龙斌(1992－)2014年于西安电子科技大学获学士学位，现为兰州交通大学电子与信息工程学院网络安全实验室研究生，研究方向为网络安全、深度学习 **E-mail**：1070648617@qq.com
伍忠东（1968－)教授，硕士生导师，研究方向为网络安全、下一代铁路无线通信 **E-mail**： wuzhd@mail.lzjtu.cn
苏佳丽（1995-）2016年于西北师范大学获学士学位，现为兰州交通大学电子与信息工程学院网络安全实验室研究生，研究方向为下一代铁路移动通信、机器学习 **E-mail:** 605996992@qq.com

法相比在性能上取得了显著的提升. 卷积神经网络拥有更少的链接和参数，易于训练，具有更好的泛化能力能够提取更深层的细微特征[3][4]. 然而，卷积神经网络在实际应用中，首先需要大量带有标签的数据，这大大增加的工作量；其次，深度神经网络处理高维数据时，需要使用高维卷积核进行卷积运算，计算复杂度高，耗时长[5][6]；最后，虽然深度神经网络对比传统深度模型参数较少，但是仍有许多超参数（如节点数、层数、学习效率等）需要花费大量的时间进行调试.

集成学习(Ensemble Learning)通过考虑不同算法的组合，将传统的智能算法或者深度学习算法设计为多个弱分类器，然后通过对分类器群的分类策略统筹来获取更加优良的性能. 集成学习具有稳健性、容错率高等优点. 决策树不依赖于线性或非线性分类，只关心数据样本之间的信息增益，这使得它在分类问题上有着天然的优势. 随机森林是决策树的集成学习方法，继承了决策树分类的优势以及高容错的特点，在大量分类回归问题中有着比 SVM、k-近邻等传统深度学习算法优秀的表现. 然而，依赖传统网络模型的集成学习在获得更高性能的同时，并不能深度挖掘数据中的更深层抽象的信息，这也是限制其性能更好的瓶颈.

本文针对 CNN 入侵检测算法复杂度高、耗时长的缺陷，提出了基于集成深度森林(EDF)入侵检测算法. 深度森林算法综合了 CNN 表征学习的优势和传统集成学习高容错的特点，在与 CNN 类似的级联模型中，加入了收敛条件，使得模型根据复杂度自适应调整模型的深度. 此外，EDF 模型使用森林层带替了卷积层和全连接层，节点模型更为简单，且优化了计算复杂度. 实验表明，EDF 在减少耗时、减少超参数数量等方面对 CNN 有着显著优势.

2 深度森林

2.1 决策树

决策树(Decision Tree)是通过分析对象属性和对象类别的关系而建立的树状图，其分支代表预测方向，叶子节点表示最终预测结果. 它是一种典型的解释型算法，可以根据不同的对象类别将数据分为不同的类，每个类中的数据有着某种同一性[7]. 决策树的分类准确性可以用信息增益或基尼指数衡量. 为获取每个决策的最大信息增益，定义公式(1)为决策树的目标函数:

$$IG\left(D_p, f\right)=I\left(D_P\right)-\sum_{j=1}^{m}\frac{N_j}{N_p}I\left(D_j\right)j \tag{1}$$

其中，f 为数据特征，D_p 和 D_j 分别为父节点和第 j 个子节点，I 为不纯性度量，N_p 为父节点的样本总数，N_j 为第 j 个子节点的样本数目. 为了计算简便并减少搜索空间，决策树一般使用二分树. 此时父节点分出两个子节点 D_{left} 和 D_{right}，带入(1)式即可得二分决策树的目标函数(2):

$$IG\left(D_p,a\right)=I\left(D_P\right)-\frac{N_{left}}{N_p}I\left(D_{left}\right)-\frac{N_{right}}{N_P}I\left(D_{right}\right) \tag{2}$$

其中 $I\left(D_{left}\right)$、$I\left(D_{right}\right)$ 为左右节点的不纯性度量，N_{left}、N_{right} 为左右节点的样本数量，N_P 为样本总量. 不纯性度量是评估分裂节点前后的信息增益的参数，可以由信息熵函数(3)得出:

$$I_H\left(t\right)=-\sum_{i=1}^{c}p\left(i\,|\,t\right)\log_2 p\left(i\,|\,t\right) \tag{3}$$

其中，$p\left(i\,|\,t\right)$ 表示节点 t 中样本属于 c 类的概率. 熵函数定义了树中样本与类别的相关信息增益，增益越大，表示样本属于该类别的概率越高，因此优化熵的最大值即可得到最优决策. 由公式(3)可以得到，在样本类型确定即概率为 1 时，信息熵为 0，而在均匀分布即概率为 50%时，信息熵有最大值 1，即熵策略尽力最大化决策树中的互信息.

在决策树算法中，首先对每个特征属性进行遍历，获取使熵增益最大的属性作为子节点的划分依据，然后对子节点重复迭代，直到分裂出的节点唯一. 决策树算法如表 1 所示:

表 1 决策树算法

算法 1 Decision Trees Algorithm
初始化: 数据样本集 D，属性集 A. While True: TreeGenerate(D，A): 生成节点 node If D 中样本全属于同一类别 C: 将 node 标记为 C 类叶节点 Return If 属性集 A 为空或 D 中所有样本属性值相同: 将 node 标记为最多类 Return 从 A 中选取最佳划分属性 a* For a in a*: 为 node 生成一个分支，令 Dv 表示 D 中属性为 a*的 a 的集合

```
If  Dv 为空:
    Continue
Else:
    TreeGenerate(Dv，A\{a*})
```

2.2 随机森林(RF)

集成学习是通过合并弱分类器增加判决算法来得到分类性能更加优越和稳健的强分类器，能有效减小过拟合，提升分类器准确率. 集成学习分类算法的组合方式有两种：Bagging 和 Boosting. Bagging 是在总样本中随机抽取不同的样本对各分类器进行训练，分类器可以并行训练；Boosting 是将全部的样本送入每个分类器，在每个分类器输出后依据错判率更新样本的权值，然后通过拟合权值残差的方式得到最终模型.

RF 是一种 Bagging 集成学习方式，在分类回归领域有着广泛的应用[8]. RF 通过 Bagging 方式生成多组决策树得到不同的分类策略，然后执行判决算法(如取预测值期望)来达到综合所有分类策略、改善分类器性能的目的. 图 1 阐释了 RF 学习方法，假设回归的输出向量长度为 3，随机森林首先训练 n 组决策树得出每组的预测概率，然后对 n 组输出概率求均值得到随机森林的最终输出[9].

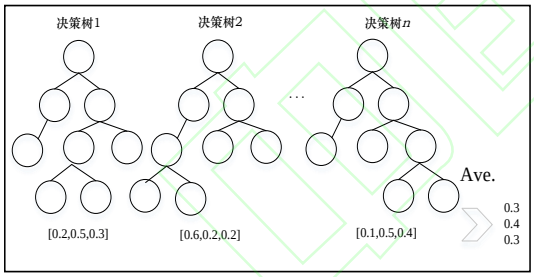


图 1 随机森林决策方法图

2.3 集成深度森林(EDF)

集成深度森林(EDF)中是利用多个 RF 构成一层，然后通过级联(cascade)将构成层间连接. 每一层的输出类向量为一组预测值，使用测试集判定该层模型是否满足收敛条件(如准确率、循环次数等)，若不满足条件则将输出的类向量与初始训练数据连接，作为下一层的输入[10]. DF 模型如图 2 所示，samples 为预处理后的数据向量，输入第一层森林层后，4 个森林分别估计所有样本的类别概率，然后将类别概率作为输出向量和原始样本数据拼接，作为下一层的输入向量，直到达到预设的循环次数或达到收敛条件，最后对输出层的向量求均值，输出概率最大的类别作为预测的样本类别.

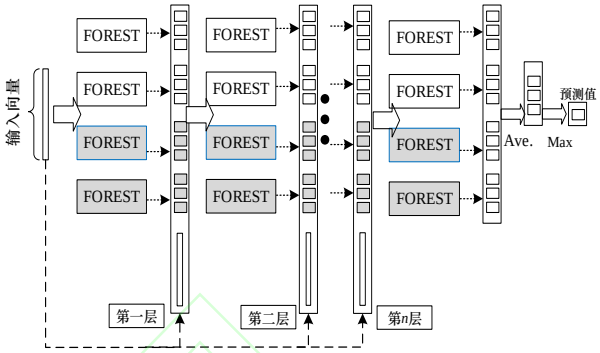


图 2 DF 级联模型示意图

EDF 主要思想是类比 CNN 的网络结构、特征表示方法和表征学习方法，使用 RF 层代替 CNN 的隐藏层和全连接层，构建深度森林网络[11]. 由于 RF 具有优秀的分类特性，超参数少且对参数设置不敏感，对小规模数据和大规模数据都有良好的性能，不需要复杂的卷积运算，在实验中其训练时间、调参难度和测试集识别率与深度神经网络相比表现出了极高的竞争力. 本文首次将 DF 应用到入侵检测中. 本文的 EDF 算法，使用两种特殊的森林构建森林层，使用 bagging 集成方式扩展森林层，使用 ending-to-ending 的方法将上层输入与下层输出合并作为新的输入数据，并分别使用了交叉验证和袋外估计方法预测每一层的输出概率. EDF 学习流程如图 3 所示. 在图 3 中，每一层使用了 4 个森林节点，其中森林模型有两种，分别为 RF 和极限树森林(Extra Trees). 深度森林模型中的深度可以由收敛条件控制，而深度神经网络需要人为设置深度. 深度森林模型深度的自适应，使得该算法适用于不同数据规模.

3 实验设计

实验设计包括数据预处理、分类器算法设计与对照以及实验分析. 实验平台：CPU intel i7-7700Q，python3.6.5.

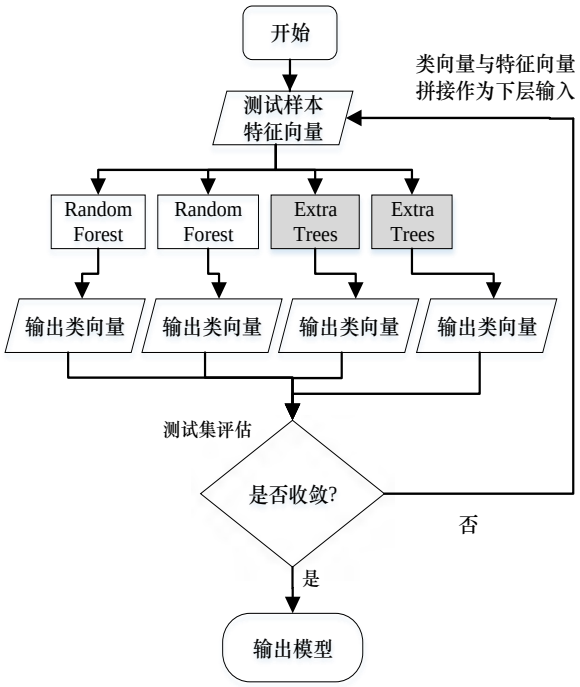
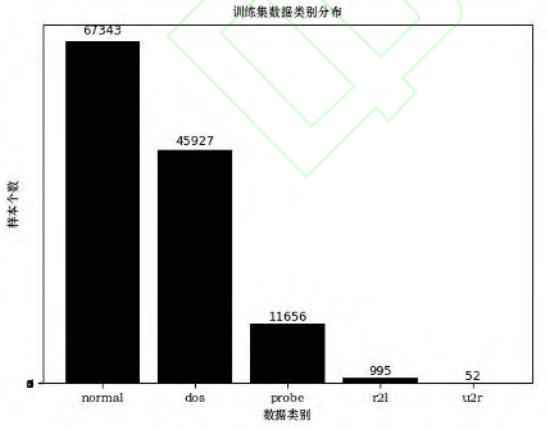


图 3　EDF 算法流程图

3.1　数据预处理

本实验采用 NLP-KDD 入侵检测数据集，此数据集包含 KDDtrain+、KDDtest+、KDDtrain-20percent 和 KDDtest-21 等多组数据，其中 KDDtrain+包含了 125793 条数据，按攻击类型可分为 5 种，按攻击方式可分为 23 种，共 41 种特征. KDDtrian-20percent 为 KDDtrain+的子集，包含 25192 条数据[6][7]. 图 4 展示了 KDDtrain+和 KDDtest+中样本攻击类型的分布情况[12].



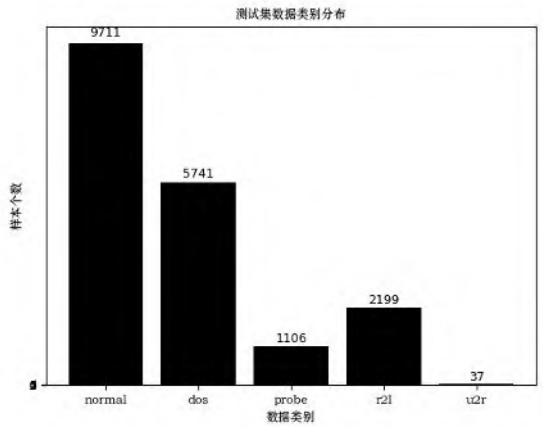


图 4 训练集和测试集数据类分布

数据预处理的主要目的是去除数据噪声、降低数据维度、减少训练时间以及提高运算性能，主要步骤包括畸形样本处理、分类数据和序数数据编码、数据尺度缩放以及特征选择和提取.

在畸形样本处理中，测试集中包含若干条训练集中不存在的攻击方式数据，在使用测试集进行模型验证时，这些数据相当于无用的噪声数据，在实验 c 中将丢弃，其他对照实验不做处理.

数据集中包含三条分类数据特征，分别为 protocol_type、service 和 flag，其中 protocol_type 有三种标识，service 中有 72 种标识，flag 中有 11 种标识. 将这三条特征标识作标签二进制化(LabelBinarizer)处理，共形成 83 条新的特征，将这 83 条特征与其他数据特征连接，形成 122 种特征的无标签数据集.

数据尺度缩放方法包含对数尺度变换、标准化、归一化和中心化等方式，本实验中选择数据标准化，公式（4）描述了标准化的方法：

$$f_{std}^{i}=\frac{f^{i}-\mu_{x}}{\sigma_{f}}\quad i\mu_{f}=\frac{1}{n}\sum_{i=1}^{n}f^{i}$$
$$\sigma_{f}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f^{i}-\mu_{f}\right)^{2}} \tag{4}$$

其中 μ_f、σ_f 为一组特征数据中的均值和标准差，f^i 为特征中第 i 组数据.

特征选择和提取的目的是去除冗余无关特征，从原始特征空间中选取最优特征子集，使其拥有比原始特征空间更好的分类性能并降低数据处理所用的时

间，主要方法包括高维空间映射（如核函数）、主成分分析（PCA）、线性判别分析（LDA）等. 其中 LDA 属于监督学习，需要用到标签数据，与实际情况不符；而客观上入侵方式与协议类型、所耗费时间线性相关，不存在高维度关系，因此不需要高维空间分析；PCA 属于无监督特征选择和提取方式，可以达到降低维度和信息降噪的目的.

本文使用 PCA 将高维数据映射到另一个低维子空间中，子空间的长轴（主成分）代表着数据变化率最高的特征. PCA 还可以反应特征的重要程度和相关程度，进而达到特征选择的目的. PCA 算法如表 2 所示，表中 m 表示选择的特征数量[13].

表 2　主成分分析算法

算法 2　Principal component analysis
1：数据特征 $x_1, x_2, \cdots x_n$ 标准化并中心化
2：计算特征的协方差矩阵 Σ
3：解协方差矩阵得到特征向量 $v_1, v_2, \cdots v_n$ 和特征值 $\lambda_1, \lambda_2 \cdots \lambda_n$
4：选取前 m 个最大特征值和其对应的特征向量
5：由 m 个特征向量构建映射矩阵 W
6：使用映射矩阵 W 将原数据映射至 m 维特征子空间

在数据分类中，如果类别较少而使用的特征较多或模型过于复杂，会发生过拟合现象，具体表现为在训练集上准确率高而在测试集准确率低很多. 为了减少过拟合的影响，本文中为了选取最优特征数量，使用基于相关性的特征选择算法(CFS)估计每个特征对于标签的重要程度，图 5 列出了重要程度最高的 25 种特征及其重要度.

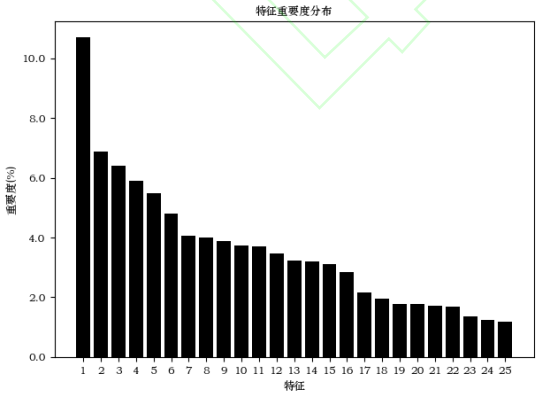


图 5 特征重要度分布图

在图 5 中，重要度最高的 25 组特征重要度之和为 90.27%，而其特征数只占总数的 20.22%，因此选择少量最重要的特征作为训练集，对性能的影响很小，且可大大减少计算量. 本文实验使用 grid-search 方法从(1，56)的范围内选择最优的特征数量(前 56 组重要度最高的特征重要度之和达到了 99. 90%)，最终确定的最优特征数为 15，入侵检测框架如图 6 所示.

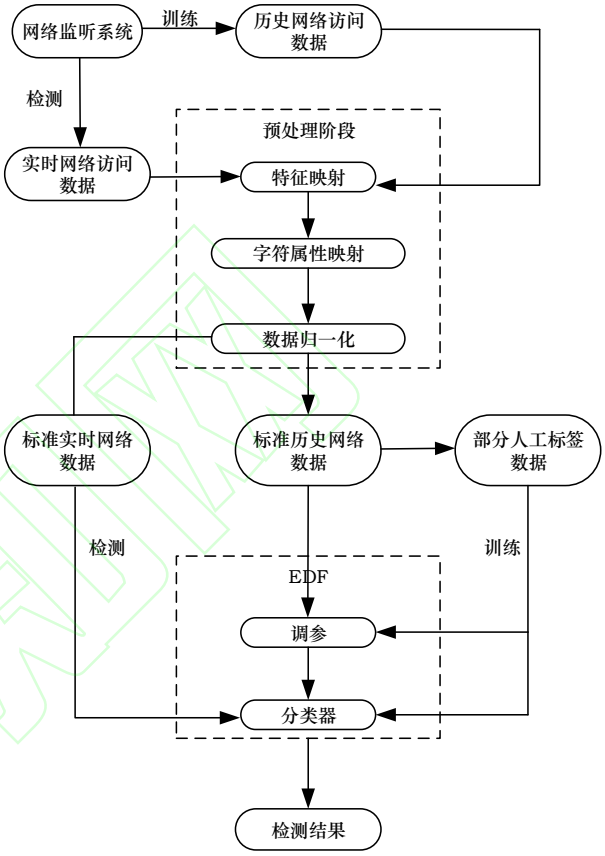


图 6　EDF 入侵检测框架

3.2 分类器算法设计与对照

为进行 EDF 和 CNN 的性能对比，在实验平台、特征数量相同的条件下，首先构建了有三层隐藏层的 CNN 和三层 Forest 层的 EDF 模型，优化两种模型的参数，使其达到最优；然后，设计实验使用两种模型对数据集进行分类预测，评估模型性能. 每种模型在相同条件下实验十次，取平均值作为最终结果.

3.2.1 实验设计与对照:

a. 在 KDDtrain+和 KDDtest+数据集上比较 EDF 和 CNN 的性能，并与其他分类器作比较. 其他分类器性能采用文献[12,14,15]中的结果. 评估策略为在 normal、anomaly 标签下的二分类准确率和模型训练时间.

b. 在 KDDtrain+和 KDDtest+数据集去除畸形样本后比较 EDF 和 CNN 的性能，评估策略为数据集总体的五分类(4 种攻击类和 normal 类)accuracy、precision、recall 和 F1 值以及分别在五种样本类型上的准确率和模型训练时间.

3.2.2 EDF 参数设置

EDF 中每层中的森林使用 Grid Search 方法搜寻最优参数，表 3 给出了最终参数设置:

表 3 EDF 参数配置

RF 模型	ETF	RF
决策树数量	20	20
最小叶子数	20	1
损失函数	gini	gini
最大特征数量	1	12
输入特征数量	15	15

3.2.3 CNN 参数设置

CNN 使用三层卷积网络，每层参数配置如下表所示:

表 4 CNN 卷积层参数设置

卷积层	第一层	第二层	第三层
Filter	16	32	64
Kernel	3	3	3
Padding	Same	Same	Same
Activation	Relu	Relu	Relu

在 CNN 输出层多分类任务激活函数为 Softmax，二分类任务使用 sigmoid 算法，并使用 Adam 算法进行优化，使用交叉熵估计损失. 其中 Adam 算法学习效率设为 0. 0001. CNN 模型迭代轮数(epoch)设置为 40，批量梯度更新数量(batch)设为经典值 100.

3.2.4 实验评价标准

采用分类准确率(accuracy)、精确率（precision)、召回率（recall）和攻击类的 F1 值作为评价指标:

$$\begin{aligned} ACC &= \frac{TP+TN}{FP+FN+TP+T}; \\ PRE &= \frac{TP}{TP+FP}; \\ REC &= \frac{TP}{FN+TP}; \\ F1 &= 2\frac{PRE\times REC}{PRE+REC}; \end{aligned} \tag{5}$$

式(5)中，TP 为模型预测正确的攻击样本数，FP 为预测为攻击类而实际为正常类的样本数，TN 为模型预测正确的正常类样本数；FN 为将攻击类预测为正常类的样本数.

3.3 实验结果和分析

表 5,表 6 和图 7 是实验 a 的结果.

在表 5 中, EDF 在 Normal 类准确率比 CNN 高 5.16%，而在 Anomaly 类上低 2.46%，这是因为两种算法的精确率和召回率不同: EDF 在预测负向类(normal)的样本时有较好表现，即误报率低；而 CNN 在预测正向类(anomaly)时表现优秀，即漏报率低. 但从总体而言，二者的 accuracy 和 F1 值相差不大，而训练时间花销上 EDF 比 CNN 减少了 74.86s，远远优于 CNN.

表 6 为 EDF 算法与其他入侵检测算法的检测率对比，其中 RF 算法使用 weka 的经典参数进行实验，其他算法结果数据来自文献[12,14,15][14][15]. 可以看到，EDF 和 CNN 算法的检测率超过了朴素贝叶斯(NB)、多层感知机(MLP)和支持向量机(SVM)等传统机器学习算法，且与 J48 决策树(J48)，朴素贝叶斯树(NBTree)和随机树(RandomTree)有 2%左右的差距，这是因为这三种树算法都只有单一的特征表征，同种算法每次训练得到的分类数据也具有巨大的差距，即它们的 F1 值较小，并没有获得完全的数据特征，实际中稳定性较差. 三种算法的 F1 值对比如图 7 所示，可以清晰地观察到从决策树到随机森林再到 EDF，随着集成度的提高，F1 值在不断上升，EDF 的 F1 值比决策树高了 2%，比随机森林高了 1%，表明分类器的性能不断提升.

表 5 EDF 和 CNN 入侵检测率比较

评估策略	样本	EDF	CNN
分类	Normal	97.32%	92.16%
准确率	Anomaly	65.77%	68.23%
ACC	总体	79.28%	78.76%
PRE	总体	96.76%	67.74%
REC	总体	64.35%	96.42%
F1	总体	77.30%	77.72%
训练时间	总体	17.77s	92.63s

表 6 KDDtest+上各分类器检测率

分类器	ACC
J48	81.05%
NB	76.56%
NBTree	82.02%
RandomForest	80.67%
ExtraTrees	78.19%
RandomTree	81.59%

MLP	77.41%
SVM	69.52%
CNN	78.76%
EDF	79.28%

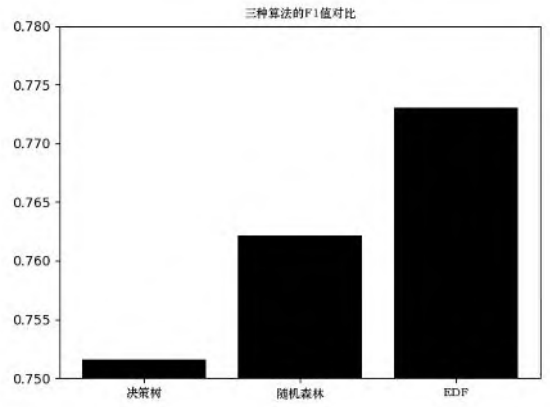


图 7 三种算法的 F1 值对比

表 7 为实验 b 在五分类的条件下，EDF 和 CNN 模型在测试集的预测准确率和模型训练时间. 在 normal、dos 和 probe 类型的预测准确率非常接近，而在 r2l 和 u2r 两种类型下的准确率相差甚远且都准确率不高，这是因为在训练集中 r2l 和 u2r 的样本数量非常少，分别只有 995 和 52，只占总体样本数量 0.79%和 0.04%，因此分类器很难学习得到这两类数据的特征，这两类样本非常容易被误分为具有大量样本的其他类；EDF 比 CNN 的训练时间少了 52 秒，约节省了 55.5%的时长，充分表明了 EDF 算法减少运算复杂度上的优势. 另外，比较实验 a 和 b 的训练时间可知，EDF 训练时间增加较大，而 CNN 训练时间几乎不变，表明 EDF 的运算复杂度受多类分类的影响较大，这是因为在训练时每个 EDF 的网络节点都要受到多分类运算的影响，而 CNN 只有最后的全连接层和输出层进行多分类运算，隐藏层的特征映射计算几乎不变. 但是 CNN 在进行多分类任务时，为保持训练的精确度，需要更多层数来提取更加细微的特征，而且每层的滤波器数保持不变甚至增加，其每层时间复杂度服从 $O\left(F^2 \bullet K^2\right)$，其中 F 为特征数量，K 为卷积核尺寸，而 EDF 时间复杂度与层数线性相关.

表 7 EDF 和 CNN 五分类任务检测率

评估策略	样本	EDF	CNN
分类	Normal	92.72%	93.54%
	DOS	99.20%	98.57%
准确率	Probe	89.15%	91.68%
	R2l	15.42%	27.56%
	U2r	59.60%	36.50%
ACC	总体	87.26%	87.06%
训练时间	总体	41.58s	93.34s

综上，实验结果表明 EDF 入侵检测算法在保持与 CNN 近似的检测率的同时，优化了训练时间和时间复杂度，是一种新的可实用的入侵检测算法.

结束语

本文中的 EDF 入侵检测算法在入侵检测率上与 CNN 相比几乎持平甚至略有优势，而在训练时间上则远远小于 CNN，充分体现了 EDF 算法的运算复杂度小的优越性，能更好的应对当前网络环境中数据量大和处理不及时、不精确的问题. EDF 继承了 CNN 的表征学习能力和集成学习的优秀分类能力，是一种有潜力的新算法模型.

在未来工作中，将致力于中如何搜寻最优森林节点个数，以及如何更进一步优化改进该算法. 此外，将针对样本不平衡数据集的问题对算法做进一步的研究[16].

参 考 文 献

[1] 徐建,薛永隽.机器学习理论在入侵检测技术中的应用研究[J].信息化研究,2014,40(03):6-8+12.

[2] 王晓晖,盛斌,申瑞民. 基于深度学习的深度图超分辨率采样[J]. 计算机工程, 2017, 43(11): 252 - 260.

[3] Ingre B, Yadav A. Performance analysis of NSL-KDD dataset using ANN[C]//Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on. IEEE, 2015: 92-96.

[4] 汪洋,伍忠东,火忠彩.基于 DBN-KELM 的入侵检测算法[J/OL]. 计 算 机 工 程 : 1-6 [2018-11-24]. https://doi.org/10.19678/j.issn.1000-3428.0052314.

[5] Mirza A H, Cosan S. Computer network intrusion detection using sequential LSTM Neural Networks autoencoders[C]//2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018: 1-4.

[6] 胡文君,傅美君,潘文林. 基于 Kaldi 的普米语语音识别[J]. 计算机工程, 2018, 44(1): 199 - 205.

[7] 郭慧,刘忠宝,柳欣.基于云模型和决策树的入侵检测方

法 [J/OL]. 计 算 机 工 程 : 1-9 [2018-11-24]. https://doi.org/10.19678/j.issn.1000-3428.0052276.

[8] Breiman L. Random forests. Machine Learning, 45(1):5–32, 2001.

[9] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[10] Zhou Zhihua, Ji Feng. Deep Forest: Towards an Alternative to Deep Neural Networks[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 3553–3559, 2017.

[11] Xu Yin, Wang Ruilin, Liu Xiaobo, et al. Deep Forest-Based Classification of Hyperspectral Images[C]//Proceedings of the 37th Chinese Control Conference, Wuhan, China: Chinese Control Conference, 2018: 10367-10373.

[12] Tavallaee M, Bagheri E, Lu W, et al . A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications. Ottawa: IEEE, 2009: 1-6.

[13] Raschka S. Python machine learning[M]. Packt Publishing Ltd, 2015.

[14] Revathi S, Malathi A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection[J]. International Journal of Engineering Research & Technology, 2013, 2(12): 1848-1854.

[15] 钱铁云，王毅，张明明，等. 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.

[16] 姚立,张曦煌.MapReduce 环境下处理多类别不平衡数据的改进随机森林算法[J].微电子学与计算机,2018,35(11):139-144.