基于深度学习的入侵检测系统

◆董宁　程晓荣

（华北电力大学（保定）控制与计算机工程学院　河北　071000）

摘要：入侵检测作为一种主动防御系统能够有效阻止来自黑客的多种手段的攻击，随着机器学习和深度学习的发展，相关技术也开始应用到入侵检测中。本文在 KDD-Cup 1999 数据集的基础上，对其进行数据标准化和归一化，然后利用随机森林对数据集处理，计算每个特征的袋外数据误差（OOB），得到每个特征的重要性，提取出 12 个重要性最高的特征，并将数据集按照 7/3 随机划分为训练集与测试集，使用五层深度神经网络训练，三个隐藏层设置为 100 个节点，ReLU Leaky 作为激活函数，使用 Adam 作为优化器，交叉熵作为代价函数，对处理后的数据集训练。以准确率（Accuracy），精确率（Precision），召回率（Recall）为模型的衡量指标，最终得到了精确率为 94.87%，召回率为 94.35%的模型。

关键词：深度学习；神经网络；入侵检测；KDD99

1　引言

入侵检测系统（IDS）是一种能够对网络中的流量和用户行为进行实时跟踪并审计，判断当前操作是否为可疑行为并能主动报警的网络安全设备。区别于其他网络安全设备，IDS 是一种主动防御系统。根据信息的来源可以将入侵检测分为基于主机型与基于网络型。根据检测方法又可分为异常入侵检测型和误用入侵检测型。早期有些入侵检测系统[1]通过建立用户特征表，比较当前特征与已存储定型的以前特征，利用概率统计方法，从而判定是否为攻击行为。然而该方法的缺陷在于用户特征表需要手动更新，且判定是否入侵的参数的确定比较困难，容易造成过高的误报率和漏报率。

随着机器学习和深度学习技术的发展，有人提出了基于 KNN（K 最近邻分类算法）的入侵检测模型，通过将用户行为进行分类，来判定是否为入侵行为[5]。然而该方法随着用户访问量的增加，单次行为分类的时间可长达 60 秒，在实际使用中并不可取，尽管后期有人提出了优化的方案[2]，然而当数量级达到百万时仍无法高速地对用户的行为进行响应和判断。到目前为止基于神经网络的入侵检测系统比较流行，本文则通过对 KDD1999 数据集经过处理后利用深度神经网络对数据进行训练，得到一个精确率和召回率都比较高的模型。

2　基于深度神经网络的 KDD 数据集训练

2.1 数据集数据处理

本文采用 KDD1999 数据集，该数据集每条记录由 41 个特征和一个标签组成。其中，41 个特征由四个类型组成，包含 TCP 基本连接特征（九种，1-9）、TCP 连接的内容特征（13 种，10-22）、基于时间网络的流量统计特征（九种，23-31）和基于主机的流量统计特征（十种，32-41）。在 41 个特征中还存在一些非连续型特征，例如连接的服务和状态等。由于数据集各特征之间数据差异较大，且存在字符型数据，所以需要对数据集进行预处理，即字符型转为数字型，特征标准化，特征归一化。

字符型数据转为数字型可选用 one-hot 编码，然而在本文中为减小后续计算量，将所有可能值编入数组，根据数组中对应值的索引来将字符型数据转换为数字型。数据标准化与数据归一化用于将特征规范在特定区间内，防止差异过大。数据标准化如式 1 所示。

$$X_{ij}^{'} = \frac{(X_{ij} - AVG_j)}{STAD_j} \tag{1}$$

$$AVG_j = \frac{1}{n}(X_{1j} + X_{2j} + \cdots + X_{nj}) \tag{2}$$

$$STAD_j = \frac{1}{n}(|X_{1j} - AVG_j| + |X_{2j} - AVG_j| + \cdots + |X_{nj} - AVG_j|) \tag{3}$$

式 2 中 AVG 为平均值，式 3 中 STAD 为平均绝对偏差，比标准差对于孤立点具有更高的鲁棒性。

数据归一化如式 4 所示：

$$X_{ij}^{'} = \frac{X_{ij}^{'} - X_{\min}}{X_{max} - \mathrm{X}_{\min}} \tag{4}$$

$$X_{min} = \min\{X_{ij}^{'}\} \tag{5}$$

$$X_{max} = \max\{X_{ij}^{'}\} \tag{6}$$

式 5 和 6 分别为提取出所有样本的最小值与最大值，并经过式 4 处理后得到处理后的样本。上述处理后所有特征区间均为[0-1]。在标签中，若一条连接记录正常，则标签为 normal，否则为攻击的类型。KDD 数据集攻击类型有 4 大类共 32 种，本文将其分为两类，即 normal 和 attack，转为数字型即为 01 和 10。

2.2　KDD 数据集特征选择

由于 KDD 数据集共存在 41 个特征，部分特征值在大部分记录中均相同，对最终分类效果无太大影响，且增加了训练时间，所以本文采用随机森林对特征重要性进行排序，使用袋外数据错误率作为衡量指标，选出了 12 个重要性最高的特征，然后利用神经网络对数据访问记录进行分类。最终 12 个特征及权重如表 1 所示。

表 1 特征选择后的十二个特征

特征名及简要概述	权重值
count 两秒内与本记录有相同靶机的记录数	0.219005
dst_bytes 目标机到源主机字节数	0.156562
logged_in 是否登录成功，是为 1，否为 0	0.098179
dst_host_count 前 100 条记录中，与本记录有相同靶机的记录	0.068435
srv_count 两秒内与本记录有相同服务的记录	0.053062
dst_host_srv_diff_host_rate 前 100 条与本记录有同服务同靶机的记录中，与本记录具有不同源主机的记录所占的百分比	0.049701
Service 靶机网络服务类型	0.046566
dst_host_same_src_port_rate 前 100 条记录中与本记录有同靶机同源端口的记录所占的比例	0.039167
flag 连接状态，正常或错误	0.033854
same_srv_rate 两秒内在与本记录有靶机的记录中，与本记录有同服务的记录的比例	0.030765
diff_srv_rate 两秒内在与本记录有同靶机的记录中与本记录有不同服务的记录的比例	0.027173
src_bytes 从源主机到靶机的数据的字节数	0.023462

2.3 深度神经网络设计

本文使用了深度神经网络，其中输入层为特征选择并处理后的 12 个值，中间为三个隐藏层，每个隐藏层有 100 个神经元。每个神经元在经过激活后均进行 dropout 处理，避免过拟合现象。最终输出为两个节点，并随后进行代价函数的计算。其中代价函数使用交叉熵计算，如式 7 所示。

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \quad (7)$$

其中 p，q 为两个独立的概率分布，分别代表网络输出值与实际值。

网络结构如图 1 所示。第一列十二个圆为十二个输入神经元，中间三列每列为 100 个神经元，相邻列每个神经元之间都存在连接，最后一列为两个输出神经元。

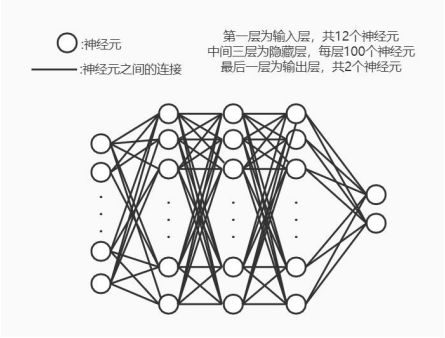

图 1 网络结构

在实验设计初期，神经网络的激活函数为线性整流函数(ReLU)，然而在训练过程中，经常出现损失函数不下降且神经网络中的参数得不到更新等情况。经排查后得知触发了 Dying ReLU 现象。在神经网络中，x 为输入，w 为神经网络的权重，神经元输出为 z，经过 ReLU 后为 a，H 为损失函数，α为学习率。则神经网络的前向传播如式 8，式 9 所示。x 为来自上一层的经过激活后的神经元输出，w 为当前隐藏层与下一层之间连接的权重。

z=w*x　　(8)

a=relu（z）=max（0，z）　　(9)

神经网络的反向传播中，需要求每层权重对应的梯度，并设置学习率不断调整权重。求梯度及权重的更新如式 10，11。当一个巨大的梯度流过神经元且学习率较高时，会导致权重更新过大，造成对于任意输入 x，网络的输出 z 小于 0，经过激活函数后 a 为 0，最终导致梯度恒为 0，后期该权重都得不到更新，形成了神经元死亡的现象。

$$\frac{\partial H}{\partial w} = \frac{\partial H}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial w} \quad (10)$$

$$w = w + \alpha\frac{\partial H}{\partial w} \quad (11)$$

故经过多次失败后最终选用 Leaky ReLU 函数，不同于 ReLU 函数将所有小于 0 的输出设置为 0，Leaky ReLU 是给所有负值赋予一个斜率，从而保证后续的权重更新。

2.4 实验及数据分析

本次训练采用 KDD1999 10%样本集，共 40 余万条，随机划分为 3:7 比例，分别为测试集与训练集。每次随机取 100 条记录为一个 batch 进行训练，并计算为一次迭代。由于样本集中正常访问与攻击的记录比例为 2：8，并不平衡，故仅凭准确率无法有效评估模型的好坏。所以在训练中每达到一百次迭代，在测试集中随机抽取 64 条记录，进行一次准确率，精确率，召回率的计算。三个指标的计算公式如下所示：

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$precision = \frac{TP}{TP+FP} \quad (13)$$

$$recall = \frac{TP}{TP+FN} \quad (14)$$

其中 TP（True Positive）表示将正类预测为正类数，TN（True Negative）表示将负类预测为负类数，FP（False Positive）表示将负类预测为正类数，FN（False Negative）表示将正类预测为负类数。式 12 表示模型预测准确率，式 13 表示在预测为攻击的记录中，真实为攻击的比例，式 14 表示在所有标签为攻击的记录中，模型判定为攻击的记录数。如表 2 所示为训练的结果。

表 2 训练结果

迭代数	准确率	精确率	召回率
2000	91.22%	94.87%	94.35%

如图 2、3、4、5 所示为模型在训练过程中的准确率，精确率，召回率，损失函数曲线，其中横轴为迭代次数，纵轴为对应的指标。为了能更好的体现损失函数值的走势，损失函数进行了平滑处理，其中实线为平滑参数为 0.5 处理后的曲线，而虚线则为未经处理的曲线。从四张图可以看出召回率和精确率在开始时以较大幅度进行增长，在训练 400 次后开始缓慢增长，最终迭代 2000 次时，召回率和精确率均达到 94%以上。而损失函数经过一次大幅度下降后便开始以小幅度变化，且总的趋势为减小最后趋近于不变。

在实际训练过程中，准确率，召回率等参数在开始时并非一直都为 0.2 左右，由于样本分布的不均匀，且训练开始时经过第一次神经网络后的一个 batch 通常表现为全部判定为攻击，或全部判定为正常访问，因此在训练开始时准确率有时会为 20%左右，而有时为 80%左右。然而即使训练开始准确率等参数为 20%，模型的训练时间并没有太大的增加，事实上经多次验证，当训练开始准确率为 20%左右时，在训练初期准确率会经过一段高速增长后会逐渐缓速增长。

（1）准确率曲线

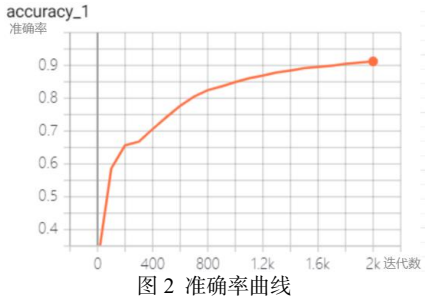

图 2 准确率曲线

（2）精确率曲线

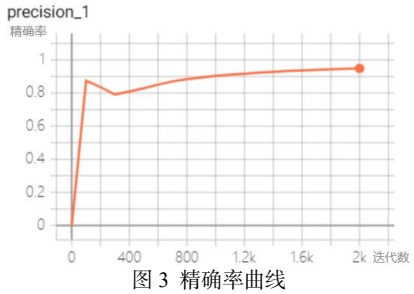

图 3 精确率曲线

（3）召回率曲线

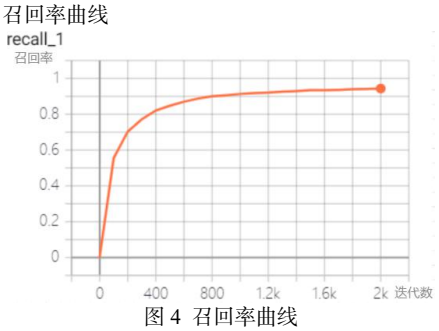

图 4 召回率曲线

（4）损失函数曲线

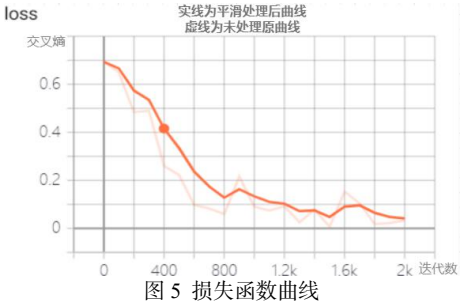

图 5 损失函数曲线

3 结论

本文利用 KDD1999 数据集，对数据集中的数据进行标准化和归一化，通过该方法使数据集中所有数据的值都保持在 0-1 的范围内，避免了在后期训练过程中因特征值之间的差异过大而带来的问题。与此同时采用随机森林选出了 12 个权重最高的特征，大大减少了训练时间和运算量，同时在模型的实际应用中也减少了判断是否为恶意访问所需的时间。然后利用三层神经网络对数据进行了分类和预测，使用 Leaky ReLU 激活函数避免了大规模神经元死亡的结果，并使最终训练结果达到了较高的准确率和召回率。然而在构造随机森林进行特征选择时，由于数据集多，数量大，导致在特征选择时需要花费大量的时间。与此同时，在信息安全的实际应用中，攻击漏报的严重性要远高于攻击误报的严重性。因此在模型的训练过程中，需要刻意提高召回率，在保证高召回率的基础上对精确率进行提高。该问题涉及阈值的选择问题，在本实验中并未体现，是将来需要研究的问题之一。

参考文献：

[1]阮耀平，易江波，赵战生.计算机系统入侵检测模型与方法[J].计算机工程，1999（09）：63-65.

[2]华辉有，陈启买，刘海，张阳，袁沛权.一种融合 Kmeans 和 KNN 的网络入侵检测算法[J].计算机科学，2016，43（03）：158-162.

[3]杨昆朋. 基于深度学习的入侵检测[D].北京交通大学，2015.

[4]C. Xu，J. Shen，X. Du and F. Zhang，"An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units，" in IEEE Access，vol. 6，pp. 48697-48707，2018.

[5]Z.Ma and A.Kaban，"K-Nearest-Neighbours with a novel similarity measure for intrusion detection，" 2013 13th UK Workshop on Computational Intelligence （UKCI），Guildford，2013，pp. 266-271.

基于私有云安全平台的 APT 攻击检测与防御架构研究

◆马琳　房潇　廉新科　张小坤

（91977 部队　北京　100036）

摘要：APT 攻击作为隐蔽性极高、目的性极强、危害性极大的新型网络攻击方式，对具有高战略价值的信息系统的安全稳定运行构成了巨大威胁。本文对 APT 攻击的特征以及攻击流程进行深入分析，总结了传统安全体系架构及防护手段在应对 APT 攻击存在的不足，在此基础上，设计了一种基于私有云安全平台的 APT 攻击检测与防御架构，为重要信息系统抵御 APT 攻击提供了有效解决方案。

关键词：APT 攻击；私有云；蜜罐；沙箱

随着信息技术的高速发展，网络入侵和攻击方式日新月异，逐渐呈现出多样化、复杂化的发展趋势，我们所面临的信息安全形势日趋严峻。近年来，诸如“震网病毒攻击”、“极光行动”等一系列入侵目的性明确、隐蔽性强、危害性大的网络入侵事件被曝光，APT（Advanced Persistent Threats，高级持续性威胁）攻击，这一概念引起了国内外信息安全行业的高度重视。政府、军队的信息系统作为具有高战略价值的关键信息系统，是 APT 攻击的重要对象[1]。传统安全架构、安全技术已经无法有效应对 APT 攻击这类新型综合性网络入侵手段。为此，文章在阐述 APT 攻击内涵、分析 APT 攻击的特征、总结归纳 APT 攻击实施过程的基础上，提出了一种能够有效防御 APT 攻击的基于私有云安全平台的防御与检测方案，为保障信息系统安全运行提供重要技术支撑。

1 对 APT 攻击的理解

1.1 定义

美国国家标准技术研究所（NIST）对 APT 攻击有如下定义：掌握先进的网络入侵知识并拥有资金支持的攻击者，为实现窃取特定大型组织的高战略价值涉密信息或破坏特定组织关键信息系统这一目的，通过技术或“社会工程学”等多种攻击手段实施的具有针对性、持久性的网络攻击[2]。

1.2 APT 攻击特征

与传统网络攻击行为相比，APT 攻击并没有运用新概念的网络攻击技术，攻击的实施者只是更具策略性、计划性的组合使用了多种传统网络攻击技术，提高了达到窃取核心数据或破坏重要计算设施等入侵目的的概率。与传统网络攻击方式比，其具有如下三个显著特征。

高级性（Advanced）。APT 攻击具有的“高级”特征主要体现在两个方面。一是隐蔽性高。攻击者在多年的网络攻防对抗中，积累了丰富的网络入侵经验，“零日”漏洞与未知恶意代码的利用等具有逃逸性质的网络攻击技术在 APT 攻击实施的过程中被作为主要的攻击手段，其帮助攻击者实现了隐藏入侵特征、逃避传统检测与防御安全技术的目的[3]。二是 APT 攻击者呈现出“高级”特征，攻击者目标与目的明确，掌握专业攻击知识，往往拥有巨额资金支持甚至攻击者可能是背负国家意志的网络战士。

持续时间长（persistent）。APT 攻击实施者为成功入侵目标系统会花费较长的时间用于收集系统相关技术与非技术信息。为成功窃取核心资料、对系统造成毁灭性破坏，成功入侵后还通常会通过预留后门等手段保证对目标系统的长期控制以便挖掘高价值信息或适时地对信息系统造成破坏。

威胁大（threats）。APT 攻击与撒网式攻击截然不同，经过精心策划，带有政治色彩抑或是商业目的[4]，政府、军队的关键信息系统是 APT 攻击的主要目标，成功入侵其危害必定较为深远，轻则泄露关键涉密信息，重则威胁国家安全与稳定。

1.3 APT 攻击实施流程

于近年来的众多的 APT 攻击案例，可归纳总结出 APT 攻击的一般流程，如图 1 所示。

图 1　APT 攻击实施流程