

一种基于 LMDR 和 CNN 的混合入侵检测模型

李桥^{1,2}, 龙春^{1,2}, 魏金侠², 赵静²

(1. 中国科学院大学, 北京 101408; 2. 中国科学院计算机网络信息中心, 北京 100080)

摘 要: 随着网络安全技术的飞速发展和大数据技术的广泛应用, 传统的机器学习模型已难以满足大数据环境下高效入侵检测的要求。针对原始数据集特征不够明显的情况, 利用卷积神经网络进行大数据特征提取与数据分析的优势, 文章提出一种基于对数边际密度比 (Logarithm Marginal Density Ratio, LMDR) 和卷积神经网络 (Convolutional Neural Network, CNN) 的混合入侵检测模型。该模型相较于现有传统的机器学习算法和神经网络模型, 能够更充分挖掘数据特征间的联系, 有效提高分类准确率并降低误报率。

关键词: 入侵检测; 对数边际密度比 (LMDR); 卷积神经网络 (CNN); 数据挖掘

中图分类号: TP309 **文献标志码:** A **文章编号:** 1671-1122 (2020) 09-0117-05

中文引用格式: 李桥, 龙春, 魏金侠, 等. 一种基于 LMDR 和 CNN 的混合入侵检测模型 [J]. 信息安全, 2020, 20 (9): 117-121.

英文引用格式: LI Qiao, LONG Chun, WEI Jinxia, et al. A Hybrid Model of Intrusion Detection Based on LMDR and CNN[J]. Netinfo Security, 2020, 20(9): 117-121.

A Hybrid Model of Intrusion Detection Based on LMDR and CNN

LI Qiao^{1,2}, LONG Chun^{1,2}, WEI Jinxia², ZHAO Jing²

(1. University of Chinese Academy of Sciences, Beijing 101408, China; 2. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: With the rapid development of network security technology and the big data technology, the traditional machine learning model has been difficult to meet the requirements of efficient intrusion detection in big data environment. For this reason, considering the advantages of convolutional neural network in feature extraction and data analysis, this paper proposed a mixed intrusion detection model based on logarithm marginal density ratio and convolutional neural network in view of the fact that the characteristics of the original dataset was not obvious enough. Compared with the traditional machine learning algorithm and neural network model, our hybrid model can make full use of the relationship between features for feature enhancement, and effectively improve the classification accuracy and reduce the false alarm rate.

Key words: intrusion detection; logarithm marginal density ratio; convolutional neural network; data mining

收稿日期: 2020-7-16

基金项目: 中国科学院信息化专项 [XXH13507, XXH13513-07]

作者简介: 李桥 (1991—), 男, 天津, 硕士研究生, 主要研究方向为网络空间安全; 龙春 (1979—), 男, 湖北, 高级工程师, 博士, 主要研究方向为网络空间安全; 魏金霞 (1987—), 女, 河北, 高级工程师, 博士, 主要研究方向为网络空间安全; 赵静 (1987—), 女, 甘肃, 高级工程师, 博士, 主要研究方向为网络空间安全。

通信作者: 龙春 anquanip@cnic.cn

0 引言

网络在各个领域都扮演着极其重要的角色。不论是人们的日常生活、娱乐等方面或是军事、科研等领域都要依托于网络^[1]。入侵攻击是一种尝试破坏信息保护、数据完整性和资源获取的行为,保护计算机免受攻击最常用的方法就是利用网络入侵检测系统(Network Intrusion Detection System)^[2]进行数据流量异常检测,从而确保网络环境的安全。

2010年之后,网络入侵检测系统更是受到越来越多人的关注,许多优秀的算法被提出并应用其中,包括朴素贝叶斯算法(Bayesian-based)^[3]、支持向量机算法(SVM)^[4]、决策树算法(DT)^[5]和基于PCA的数据挖掘算法^[6]等。传统的入侵检测算法主要基于模式匹配和数据挖掘,通过对数据本身特性的利用进行数据分类。随着现今网络数据的爆炸式增长,基于神经网络的入侵检测算法应运而生。

HINTON^[7]等人和MATIN^[8]等人于2006年提出了多隐层人工神经网络,可通过逐层预训练的方式实现特征学习,这使得各个领域都开始采用神经网络模型进行数据分析。在入侵检测领域,LIN^[9]等人提出基于字符级的卷积神经网络入侵检测系统,该方法成功地将卷积神经网络和入侵检测结合到了一起。LISEHROODI^[10]等人则利用神经网络和k-means聚类集成算法提高了入侵检测模型的分类准确率。这些方法相较于传统分类模型有了更好的表现。

卷积神经网络(Convolutional Neural Network, CNN)是一种经典的神经网络模型^[11]。它不仅可以进行特征选择,还可以对流量数据进行分类。卷积神经网络在入侵检测领域的优势主要体现在能够通过卷积层和池化层有效提取关键特征信息,从而快速训练模型并做出响应。此外,相较于传统的机器学习算法,卷积神经网络模型能够通过权值共享的方式保持模型效率,并有效减少模型需要训练的参数并防止过拟合的发生。因此可以利用卷积神经网络来构建入侵检测模型。

入侵检测模型的表现与数据集质量密切相关^[12]。

传统的数据预处理算法往往没有充分考虑特征之间的相互依赖性,特征依赖体现在某些单独看起来低信息度的特征,与其他特征一起作为数据考虑时,可能会提高分类性能。FAN^[13]等人提出的利用LMDR进行数据转换的方法是一种优秀的特征转换方法,该方法利用边缘密度对数比对原始特征进行变换,充分利用每个特征所包含的分类信息。2017年REVATHI^[14]等人利用LMDRT-SVM模型对传统SVM分类进行优化的方式也体现出LMDR对于数据集的增强效果。本文提出一种基于对数边缘密度比和卷积神经网络的混合入侵检测模型,LMDR-CNN模型。该模型选择了卷积神经网络作为入侵检测的分类模型,同时考虑到不同信息度的特征之间存在的依赖性以及卷积神经网络输入数据的适配性。

1 LMDR-CNN 模型

1.1 模型概述

本文构建的混合入侵检测模型主要包括数据预处理、LMDR特征增强、特征约减与数据图像化、卷积神经网络。模型首先利用数据预处理模块对数据集进行数据离散值哑编码并完成归一化和标准化;然后利用LMDR模块进行特征增强;再通过特征约减和数据图像化模块使数据集成为符合卷积神经网络输入的数据;最后搭建卷积神经网络模型,通过交叉验证的方式获取最优参数构建最终模型,实现入侵检测功能。算法流程如下所示:

- 1) 将数据集的离散特征哑编码化;
- 2) 将数据集标准化和归一化;
- 3) 将数据集 S 分割为互斥的两个子集 S_1 和 S_2 ;
- 4) 求出 S_1 中类条件密度 g^* 和 f^* ;
- 5) 带入 g^* 和 f^* 求出 S_2 中的对数边缘密度比矩阵 $X^{(2)*}$,构建新数据集 $Z^*=(X^{(2)*}, Y)$;
- 6) 利用随机森林算法对新数据集进行特征约减,使得维度变为 m 维,且满足 $m=a \times a$;
- 7) 重新构建数据集特征并将其转化为 $a \times a$ 维的图像特征数据集;
- 8) 构建CNN神经网络,并将最近的图像特征数据

集数据带入模型,利用10折交叉验证的方式训练模型;

9)重复步骤6)至步骤8),求得最佳的 m 并依次确定最优的模型参数;

10)利用训练好的模型实现入侵检测。

1.2 数据预处理

NSL_KDD数据集^[14]作为KDD CUP99数据集的改进版,是入侵检测领域的开源数据集。本文选择NSL_KDD数据集作为实验用数据集。卷积神经网络需要输入数值型数据,因此需要先将原数据集中的protocol_type、service、flag等3种离散型特征值进行哑编码转换。模型将这3种特征转化成由若干位0和一位1表示而成的序列,即数值哑编码化,共得到122维特征的新数据集。

由于LMDR特征增强模块中的需求以及数据集本身因数据差异过大影响最终分类效果,本文采用标准化和MinMax归一化将整体数据归一化到[0,1]区间范围内。

1.3 LMDR 特征增强

LMDR特征增强能够充分考虑特征之间相互影响的特性,避免因单纯降维而忽略数据之间的相互影响。特征增强变换步骤如下:选取数量为 N 的一个带标签样本,定义 $S=\{(\mathbf{X}_i, \mathbf{Y}_i), i=1,2,\dots,N\}$,其中 \mathbf{X}_i 表示第 i 条样本且该样本属于一个 p 维向量, \mathbf{Y}_i 是对应于 \mathbf{X} 的第 i 条样本观测值的标签。

特征增强变换步骤如下:

1) 数据集分割

将数据集 S 互斥随机分成两个子集 S_1 和 S_2 ,这两个数据集将满足 $S_1 \cap S_2 = \emptyset, S_1 \cup S_2 = S$,定义 $S_1=(\mathbf{X}_1, \mathbf{Y}_1), S_2=(\mathbf{X}_2, \mathbf{Y}_2)$,同时令 N_1 和 N_2 分别为 S_1 和 S_2 的样本数量,则有 $N_1+N_2=N$ 。

2) 核密度估计

将核密度估计用于 S_1 数据集之中求出类条件密度,并分别用 g^* 和 f^* 定义这个估计,其中有 $\mathbf{g}^*=(g_1^*, g_2^*, \dots, g_p^*)^T, \mathbf{f}^*=(f_1^*, f_2^*, \dots, f_p^*)^T$ 。

令 \mathbf{X}^{1+} 定义第一个数据集 S_1 中的标签 $\mathbf{Y}_i^{(1)}=1, i=1,2,\dots,N_1$ 时,与之相关的 $\mathbf{X}_i^{(1)}$ 的数据集合,则有 $\mathbf{X}^{1+}=\{\mathbf{X}_i^{(1)}|\mathbf{Y}_i^{(1)}=$

$1, i=1,2,\dots,N_1\}$;同理,可令 \mathbf{X}^{1-} 定义当 S_1 集合中出现 $\mathbf{Y}_i^{(1)}=0, i=1,2,\dots,N_1$ 时, $\mathbf{X}_i^{(1)}$ 的数据集合,即 $\mathbf{X}^{1-}=\{\mathbf{X}_i^{(1)}|\mathbf{Y}_i^{(1)}=0, i=1,2,\dots,N_1\}$,且满足 $\mathbf{X}^{1+} \cap \mathbf{X}^{1-} = \emptyset, \mathbf{X}^{1+} \cup \mathbf{X}^{1-} = \mathbf{X}^{(1)}$ 。因此本文定义 $g_j^*(\cdot)$ 和 $f_j^*(\cdot)$ 是分别基于如下两个数据子集 $\{\mathbf{X}_1^{1+}, \mathbf{X}_2^{1+}, \dots, \mathbf{X}_{N_1}^{1+}\}$ 与 $\{\mathbf{X}_1^{1-}, \mathbf{X}_2^{1-}, \dots, \mathbf{X}_{N_1}^{1-}\}$ 的,且有 N^+ 和 N^- 分别表示 \mathbf{X}^+ 和 \mathbf{X}^- 这两个数据集的样本数目,且有 $N^++N^-=N$ 。上述概念由公式(1)和(2)表示。

$$g_j^*(x) = \frac{1}{N_1^+ h} \sum_{i=1}^{N_1^+} K\left(\frac{X_{ij}^{1+} - x}{h}\right) \quad (1)$$

$$f_j^*(x) = \frac{1}{N_1^- h} \sum_{i=1}^{N_1^-} K\left(\frac{X_{ij}^{1-} - x}{h}\right) \quad (2)$$

其中, $j=1,2,\dots,p, K(\cdot)$ 代表核函数, h 是带宽。

3) 数据转换

将LMDR用于数据集 S_2 ,从而将 $\mathbf{X}^{(2)}$ 转变为 $\mathbf{X}^{(2)*}$,则有 $\mathbf{X}_i^{(2)*}=\log(f^*(\mathbf{X}_i^{(2)}))-\log(g^*(\mathbf{X}_i^{(2)}))$,构成新的数据集 $Z=(\mathbf{X}^{(2)*}, \mathbf{Y}^{(2)})$ 。其中有 $\mathbf{X}_{ij}^{(2)*}=\log(f_j^*(X_{ij}^{(2)}))-\log(g_j^*(X_{ij}^{(2)}))$ 。

1.4 特征约减与图像化

通过LMDR特征增强构建新的数据集后,为了降低数据特征的维度而提取更有效的特征组合,本文采用随机森林算法计算特征重要度,根据特征重要度进行特征约减。通过比较约减到不同特征维度的数据集,利用相同判别模型的表现确定实验所采用的特征维度。为了使最终模型的输入数据更加适配,本文将原 $n \times 1$ 维的数据转化为 $a \times a$ 维的图像特征数据用于卷积神经网络的模型训练。

图1表示不同情况下正常数据流量和入侵攻击数据流量的灰度图。第1列是将原数据集特征约减到40维的数据灰度图,中间一列是将哑编码后的122维特征约减到81维的数据灰度图,第三列是先将数据集通过LMDR特征增强后再进行特征约减到81维的数据灰度图像。比较每一列的灰度图可以发现,同种类型的灰度图像存在着相同或相似的特征。通过LMDR特征增强后的数据在单位像素点包含的信息更丰富,入侵攻击流量和正常流量能够有更强的区分度。本文采用LMDR特征增强和入侵检测数据图

像化方法相结合的方式生成新的图像数据集，并将其作为卷积神经网络的输入特征进行模型训练，进而提高入侵检测模型的分类表现。











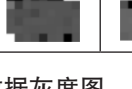

		原数据集 5x8	原数据集 9x9	特征增强 9x9
正常 流量	a			
	b			
入侵攻 击流量	a			
	b			

图 1 入侵检测数据灰度图

1.5 卷积神经网络

神经网络模型架构包括网络输入层、卷积层、池化层、全连接层和输出层等 5 个部分。

1) 网络输入层：模型的第 1 层为网络输入层。通过不同维度的实验最终得出 9x9 维数据作为数据输入效果最好，因此本层网络输入的维度为 9x9 维。

2) 卷积层：模型的第 2、3、5、6 层为网络卷积层。该层利用网络层堆叠的方式逐步提取数据集不同粒度的特征，利用 Relu 激活函数替代传统的 Sigmoid 函数使得网络收敛的速度加快。

3) 池化层：模型的第 4 层和第 7 层为池化层。池化层夹在卷积层之间，用于减少参数量并防止一定程度的过拟合。本文选择最大值池化法，其中池化层采样核的大小为 3x3，步长为 2。

4) 全连接层：全连接层为模型的倒数第 2 层，该层首先利用 dropout 方法随机丢弃一些神经元以防止模型过拟合，然后将数据通过 Flatten 操作进行图像特征打平，最后将数据信息传递给 128 个神经元节点。激活函数为 Relu 函数的全连接层用以进行最终神经元输出。

5) 输出层：该层利用 Softmax 激活函数作为分类器，通过将传递的信息转化为实际的预测值，用以进

行分类预测。

2 实验

2.1 实验设置

实验选用的是 Intel(R)Core-i5-7300HQ CPU@2.5GHz, 8GB 内存的 Windows 10 系统，代码用 python 语言实现。

为了消除因样本特征值数据差异过大所带来的影响，本文对数据样本特征进行标准化和归一化。选择数据集中的 50000 条数据采用 10 折交叉验证的方式对 LMDR-CNN 进行模型训练，并选择 10000 条数据构建测试集进行测试。选择检测率 (Detection Rate, DR), 精确度 (Accuracy), 误报率 (False Alarm Rate, FAR) 3 个指标作为模型评价指标。其中 $DR = TP / (TP + FP)$, 表示预测正确的样本占有所有样本的比例; $Accuracy = (TP + TN) / (TP + TN + FP + FN)$, 表示所有预测为入侵攻击的样本中预测正确的比例, $FAR = FP / (FP + TN)$, 表示实际正常的样本中误报为攻击的比例, 其中 TP 为真正例, TN 为真负例, FP 为假正例, FN 为假负例。

2.2 实验结果与分析

通过将特征约减到不同维度下的数据作为 LMDR-CNN 模型的数据集进行模型训练，并利用测试集进行测试。通过观察可知，LMDR-CNN 模型在数据特征约减到 9x9 维时，模型的表现最好，能够达到 99.49% 的精确度、99.41% 的检测率以及 0.39% 的误报率。图 2 给出了不同特征维度下 LMDR-CNN 的分类效果，可知数据集进行特征约减能够有效减少因差异过大，或不稳定所带来的影响，对数据原有特征进行增强，提高模型的分类表现。而当特征数量进一步减少时，会因约减掉信息量较大的数据特征而导致模型分类效果降低。

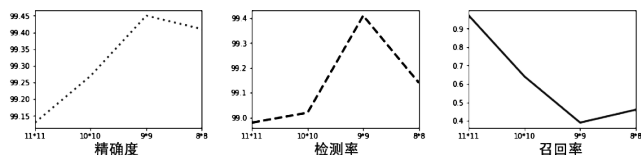


图 2 不同特征维度下 LMDR-CNN 分类效果折线图

在实验结果的基础上，采用多种机器学习算法和 LMDRT-SVM 算法作为混合模型的对照算法，进行多

种算法在9×9维数据集下的模型分类表现。由表1可直观看出，LMDRT-CNN模型相较于传统的机器学习算法表现更加优秀。

表1 不同模型算法的分类结果

	Accuracy/%	DR/%	FAR/%
SVM	96.75	92.97	1.48
随机森林	99.06	98.76	0.45
朴素贝叶斯	93.35	90.29	2.58
LMDRT-SVM	99.31	99.20	0.6
LMDR-CNN(9*9)	99.49	99.41	0.39

此外，本文还进行了CNN(1×41)、CNN(RF,9×9)、LMDR-CNN(1×41)、LMDR-CNN(RF,9×9)4种模型的对比实验，实验结果如表2所示。当模型通过特征约减和图像化，能够使得CNN模型的分类效率进一步提升。LMDR特征增强方法可以提高入侵检测模型分类准确率、精确率并降低误报率。

表2 数据图像化和特征增强前后的模型分类结果比较

	Accuracy/%	DR/%	FAR/%
CNN(1×41)	98.07	97.79	1.70
CNN(RF,9×9)	99.03	98.98	0.73
LMDR-CNN(1×41)	99.09	98.99	0.45
LMDR-CNN(RF,9×9)	99.49	99.41	0.39

3 结束语

入侵检测作为网络安全的关键技术越来越受到广泛关注。分类器的性能与数据集的质量则是构建优秀入侵检测模型的重要考量。本文在传统卷积神经网络模型的基础上，考虑到传统数据处理方法容易忽视部分特征之间的相互依赖关系以及对于卷积神经网络输入的适配性问题，提出一种基于LMDR-CNN的混合入侵检测模型。该模型通过LMDR特征增强、数据图像化方法有效地将原始数据转化为特征更加清晰且与模型输入适配的数据，并通过卷积神经网络进行分类检测。从实验结果可知，本文提出的LMDR-CNN模型更适合当今的大数据网络环境，能够通过数据增强算法和卷积神经网络有效提升入侵检测的准确率、精确率并降低误报率。●（责编 刘洋）

参考文献：

[1] CHEN Lei, YUAN Yuan. Image Recognition of Agricultural Diseases

Based on Deep Transfer Learning[J]. Frontiers of Data & Computing, 2020, 2(2): 111–119.

陈雷, 袁媛. 基于深度迁移学习的农业病害图像识别[J]. 数据与计算发展前沿, 2020, 2(2): 111–119.

[2] RAFFIE Z. A Mohd, MEGAT F. Zuhairi, AKIMI Z. A Shadil, et al. Anomaly-based NIDS: A Review of Machine Learning Methods on Malware Detection[C]// IEEE. 2017 International Conference on Information and Communication Technology Convergence, October 18–20, 2017, Jeju Island, Korea. New York: IEEE, 2017: 266–270.

[3] LEVENT Koc, THOMAS A. Mazzuchi. A Network Intrusion Detection System Based on a Hidden Naive Bayes Multiclass Classifier[J]. Expert Systems with Application, 2012, 39(18): 13492–13500.

[4] AN Wenjuan, LIANG Mangui. A New Intrusion Detection Method based on SVM with Minimum Within-class Scatter[J]. Security and Communication Networks, 2013, 6(9): 1–11.

[5] SIVA S, SIVATHA Sindhu, S. Geetha, et al. Decision Tree Based Light Weight Intrusion Detection Using a Wrapper Approach[J]. Expert Systems with Applications, 2012, 39(1): 129–141.

[6] ARUNA Jamdagni, TAN Zhiyuan, HE Xiangjian, et al. RePIDS: A Multi Tier Real-time Payload-based Intrusion Detection System[J]. Computer Networks, 2013, 57(3): 811–824.

[7] HINTON G, SALAKHUTDINOV R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786): 504–507.

[8] MATIN W, NASIM Z, AHMED N, et al. Artificial Neural Network based System for Intrusion Detection using Clustering on Different Feature Selection[J]. International Journal of Computer Applications, 2015, 126(12): 21–28.

[9] LIN Steven Z, SHI Yong, XUE Zhi. Character-level Intrusion Detection based on Convolutional Neural Networks[C]//IEEE. 2018 International Joint Conference on Neural Networks (IJCNN), July 8–13, 2018, Rio de Janeiro, Brazil. New York: IEEE, 2018: 1–8.

[10] LISEHROODI, MAZYAR Mohammadi, MUDA, et al. A Hybrid Framework based on Neural Network MLP and Means Clustering for Intrusion Detection System[C]//IEEE. 4th International Conference on Computing and Informatics(ICOCI), August 28–30, 2013, Sarawak, Malaysia. New York: IEEE, 2013: 305–311.

[11] YU Yizhou, MA Jiechao, SHI Dejun, et al. Application of Deep Learning in Medical Imaging Analysis: A Survey[J]. Frontiers of Data & Computing, 2019, 1(2): 37–52.

俞益洲, 马杰超, 石德君, 等. 深度学习在医学影像分析中的应用综述[J]. 数据与计算发展前沿, 2019, 1(2): 37–52.

[12] CHEN L S, SYU J S. Feature Extraction based Approaches for Improving the Performance of Intrusion Detection Systems[C]// International Association of Engineers. The International Multiconference of Engineers and Computer Scientists, March 18–20, 2015, Hong Kong, China. United Kingdom: IET INSPEC, 2015: 1–6.

[13] FAN Jianqing, FENG Yan, JIANG Jianchen, et al. Feature Augmentation via Nonparametrics and Selection (FANS) in High-Dimensional Classification[J]. Journal of the American Statistical Association, 2016, 111(513): 275–287.

[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(13): 1848–1853.