doi:10.3772/j.issn.1002-0470.2020.05.013

基于 SDN 的智能入侵检测系统模型与算法①

马 琳②* 张莎莎** 宋姝雨** 王 磊***

(*空军杭州特勤疗养中心信息科 杭州 310012)

(**浙江大学信息与电子工程学院 杭州 310027)

(***国家广播电视总局广播电视科学研究院 北京 100866)

摘 要 基于软件定义网络(SDN)的集中式管理、全局控制等优势,提出了一种智能的入侵检测系统架构模型。基于该模型,可以动态使用不同的机器学习算法对入侵的数据流进行检测,从而提升系统的检测性能。本文针对入侵数据流多特征、不平衡性等特点,提出了一种改进的随机森林算法,通过动态更新决策树的权重来提高分类的准确度。使用 KDD CUP99 数据集对改进的算法进行训练和测试,实验结果表明,改进的随机森林算法在检测精度、代价等指标上都得到了明显的提升,验证了新模型和新算法的有效性。

关键词 软件定义网络(SDN);入侵检测系统;机器学习;随机森林

0 引 言

入侵检测常用的检测方法有异常检测和误用检测[1]。对于异常检测,其假设入侵行为与正常行为明显不同。典型的异常检测方法包括专家系统、统计分析和定量分析。而误用检测假设所有入侵都可以表示为模式或特征,系统基于这些模式或特征来检测入侵行为[2]。异常检测的优势在于对新型攻击敏感,但误报率较高。误用检测的优势在于其对已知攻击具有优越性能,但很难检测到未知的攻击。如何提高入侵检测系统的检测准确率、降低误报率和漏报率一直是研究的重点和热点。

软件定义网络(software defined networking, SDN)作为一种新型网络架构,其控制面与转发面分离、控制器北向接口开放、可实现集中控制等特性为网络技术创新提供了良好的平台。SDN 控制器具有全局网络的信息,可以根据需要监控网络中任意位置的流量,这为集中进行入侵流监测提供了便利。在 SDN 控制器中,可以非常方便地集成各种先进的人工智能算法[3-5],突破了传统特征匹配等方法的局限性,能够显著提升入侵检测的性能。

本文设计了一种基于 SDN 的智能入侵检测系统模型,使用改进的随机森林算法进行智能流分类,以获得更高的检测率和更低的检测代价。本文使用 KDD CUP99 数据集作为训练集和测试集,对改进算法进行了性能仿真和对比分析。

1 相关研究

文献[6,7]研究了传统的流量异常检测方法在 SDN 网络中的应用。然而这些方法采用的流量特征数据较为单一,仅能针对某种特定的异常。文献[8]提出了基于改进非广延熵特征提取的双随机森林实时入侵检测方法,实现对少量异常的有效实时检测,但该方法只适合对骨干链路进行检测。文献[9]融合 OpenFlow 和 sFlow 提出了一种可进行异常检测并可减轻网络异常的机制。Jankowski 等

① 国家重点研发计划(2018YFB0803702)资助项目。

② 女,1975 年生,硕士,高级工程师;研究方向:网络信息安全,智能信息系统;联系人,E-mail: 85413976@qq.com

(收稿日期:2019-06-14)

人[10]提出了一种基于 SDN 的入侵检测架构，该架构从细粒度数据流中提取流量统计信息，并通过分类器对数据进行分类并判断是否有入侵。文献[11]提出了一种称为 AD-ICMP-SDN 的异常检测机制。其中的异常检测组件由因特网控制报文协议（internet control message protocol，ICMP）流量采集、数据训练、支持向量机（support vector machine，SVM）分类 3 个模块组成并对异常检测。但这些研究都缺少统一的系统模型，也缺乏对人工智能算法的灵活集成和应用，无法实现持续的检测性能提升。

2 基于 SDN 的智能入侵检测模型

本文提出一种基于 SDN 的智能入侵检测系统模型，如图 1 所示。该系统模型被分为 3 个平面：转发平面、控制平面、应用和分类平面。

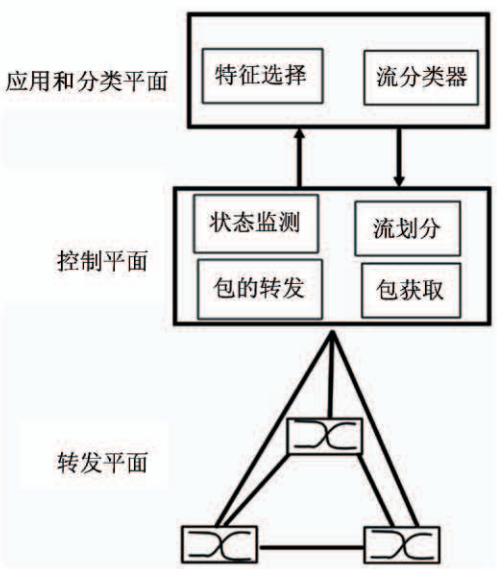

图 1 基于 SDN 的智能入侵检测系统模型

转发平面由网络中的 OpenFlow 交换机构成，它们通过安全通道与控制器进行通信。该平面负责在交换机之间转发数据包。通过收集和上传交换机的流量信息，为控制平面提供实时的网络状态和可疑数据流。另外，交换机也可以根据控制器的指令通过丢弃攻击流的报文来阻止入侵行为。

控制平面对采集的数据流进行初步分析。状态监测模块监视网络状态，并周期性地请求交换机上报流量信息。流划分模块处理和分析流量统计信息，将数据包聚类成流，并生成如下的 6 元组流标识：

$$\{srcip,\ dstip,\ srcport,\ dstport,\ duration,\ protocol\}$$

数据包的收集和检查按照一定的时间间隔执行，合适的时间间隔能够避免降低异常识别的实时性，同时不会大幅增加开销。

应用和分类平面包括特征选择模块和流分类器模块。特征选择模块提取可疑流的有关特征并找到特征的最佳子集，可以高效地处理高维数据。流分类器模块通过标记是否属于特定类型的攻击或正常流量来分类网络流量。基于该系统模型，可以灵活地选择不同的机器学习算法来测量特征的相关性和冗余度，从而找到特征的最佳子集。

3 改进随机森林算法进行流分类

随机森林算法[12-14]具有较高的预测准确率，对异常值和噪声具有很好的容忍度，目前已经在很多领域得到应用。但是，它在噪音较大的分类或回归问题上会出现过拟合，尤其当特征取值划分较多时，对算法性能的影响更大。针对随机森林在对多特征、数据不平衡等特点的网络流数据进行分类时存在的问题，本文提出了改进方法，步骤如下。

（1）初始化样本权重。

对每一个样本赋予权重 $1/N$，其中 N 为样本总数。样本权重的初始值可以根据需要调整，但样本权重之和为 1。为了提高少数类样本的分类精确度，从而提高总体分类性能，改进方法增大了少数类样本的权重。样本权重集为

$$D1 = (\omega_{1,1},\ \omega_{1,2},\ \cdots \omega_{1,i},\ \cdots,\ \omega_{1,N}) \qquad (1)$$

（2）构建决策树，更新权重抽样，并进行多轮迭代。

依据样本分布，对样本进行有权重的随机抽样，生成训练集，并用生成的训练集构造一棵树。用原始数据集对生成的树进行分类，计算分类误差率 e_m，对样本权重进行更新。这里的 e_m 使用袋外数据（即在样本抽样时没有被抽中的数据）作为测试集进行计算，也就是用生成的决策树没有学习到的样

本数据进行计算。更新后的样本权重集为

$$D_{m+1} = (\omega_{m+1,1}, \omega_{m+1,2}, \cdots, \omega_{m+1,i}, \cdots, \omega_{m+1,N}) \tag{2}$$

样本权重更新公式为

$$\omega_{m+1,i} = \frac{\omega_{m,i}}{z_m}\beta \cdot e^{\pm\alpha} \tag{3}$$

$$\alpha = \frac{1}{2}\ln(\frac{1-e_m}{e_m}) \tag{4}$$

其中,Z_m 是规范化因子。

$$Z_m = \sum_{i=1}^{N}\omega_{m,i}\beta \cdot e^{\pm\alpha} \tag{5}$$

$e^{\pm a}$ 取决于分类结果,分类正确为 e^{-a},分类错误则为 e^{a}。

依据上式可知,对于分类正确的样本,其权重更新后会变小,下一次生成分类器的过程中,被抽中的概率也就变小,可以理解为被学到的可能性减小。相反,对于分类错误的样本,若要求之后生成的分类器能够分类正确,就需要加大对这些样本的学习,所以这里更新权重时增大了其权重。式中的 Z_m 为规范化因子,保证每次更新权重后所有样本权重之和恒等于 1。参数 β 用来区别少数类与多数类,因为少数类样本占比较少,这样分类器学到的"知识"也较少。另外,少数类样本分类错误的代价相对更高,所以算法通过 β 提高少数类样本的分类精度,从而提升总体性能。β 的取值由表 1 决定。其中,m 为多数类样本(标签为 0,1,2)的总数,n 为少数类样本(标签为 3,4)总数。

表 1　β 取值

	分类错误	分类正确
少数样本	$\frac{m}{m+n}$	$\frac{n}{m+n}$
多数样本	$\frac{n}{m+n}$	$\frac{m}{m+n}$

样本更新完毕后,对新的样本依据权重抽样,得到新的训练集,生成新的一棵树。

(3)所有决策树生成完成,计算每棵树的分类器权重 w_i。

$$w_i = \frac{2}{\frac{1}{A_i}+\frac{1}{B_i}} \tag{6}$$

其中,A_i 为第 i 棵树正确分类的个数与所有树正确分类个数平均值的比值,B_i 为代价的平均值与第 i 棵树的代价的比值。

w_i 的计算公式权衡了检测精度和代价的值。由公式可知,若某棵决策树分类精度高,则其正确分类个数越多,A_i 也就越大;同时其代价低,则 B_i 也越大,这样 w_i 就较大。训练生成的随机森林中每棵决策树的生成是带有随机性的,所以每个决策树的性能不尽相同。如果一棵决策树的分类精度高,代价低,那么它的性能应该是较好的,所以它的分类器权重应该较大。

4　仿真分析

针对前文提出的模型和算法,利用 KDD CUP99 数据集[15]作为测试数据,在其提供的训练集和测试集的基础上,利用抽样得到训练子集(样本数为 57 114)和测试子集(样本数为 37 464),对改进算法进行训练和测试,模拟入侵检测过程,以对比改进前后算法的性能。

其中,样本类别分布如表 2 所示。采用的评价指标包括精度(P)、召回率(R)、F _ score(F)、准确度(AC)和误报率(FPR)[12]。

表 2　训练子集和测试子集样本分布

标签	类别	训练子集样本数	测试子集样本数
0	Normal	17 129	12 183
1	Probing	3 107	1 880
2	Dos	35 700	21 705
3	U2R	52	228
4	R2L	1 126	1 468
总数		57 114	37 464

由于不同的攻击类型未被系统检测到而产生的后果不尽相同,同时一个正常的数据流被误分类为攻击也会产生不同的代价。定义另一个比较指标(Cost)来度量每个样本被错误分类的成本损失,计算公式如下。

$$Cost = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n}M_{ij} + C_{ij} \tag{7}$$

其中，M_{ij}表示实际为类别 i、分类结果为类别 j 的样本数量。C_{ij}表示从 KDD CUP99 数据集获得的每个样本的代价值，C_{ij}的具体取值如表 3 所示。N 是用于测试的样本总数。从 Cost 的定义可知，它是一个介于 0 ~ 4 之间的数，对于分类算法来说，其 Cost 的值越小越好。

表 3　代价矩阵(C_{ij})

类别	Normal (0)	Probing (1)	Dos (2)	U2R (3)	R2L (4)
Normal (0)	0	1	2	2	2
Probing (1)	1	0	2	2	2
Dos(2)	2	1	0	2	2
U2R(3)	3	2	2	0	2
R2L(4)	4	2	2	2	0

改进前后算法的入侵检测性能如表 4 所示。

表 4　改进前后算法入侵检测性能比较

	AC (%)	*P* (%)	*R* (%)	*F _ score* (%)	*FPR* (%)
原算法	94.66	98.14	93.86	95.96	3.69
改进算法	96.10	99.53	94.66	97.04	0.91

由表 4 可以看出，改进算法相对于原算法在 AC、P、R、F _ score 性能上都有明显的提升，FPR 明显降低。改进算法的代价为 0.1445，而原算法为 0.1735，改进算法明显降低了代价，说明改进随机森林算法优于原算法。

为了进一步分析实验结果，在入侵检测过程中分别计算了各个类别的检测精度，得到结果如表 5 所示。

从表 5 可见，改进算法在 Normal、Probing、R2L 攻击类上检测精度有所提升，而在 U2L 攻击类上检测精度大幅度提高，符合预期设想。因为改进算法的基本思路是增加少数类样本(U2R、R2L)的学习度，从而提高检测精度。

为了验证以上结果的普遍性，进行了如下实验。在不同的测试集样本数下，观察对比检测结果的 Cost 值，结果如图 2 所示。

表 5　算法改进前后各类别检测精度比较

	Normal (%)	Probing (%)	Dos (%)	U2R (%)	R2L (%)
原算法	96.31	96.27	100	1.61	14.33
改进算法	99.09	99.52	100	61.84	14.60

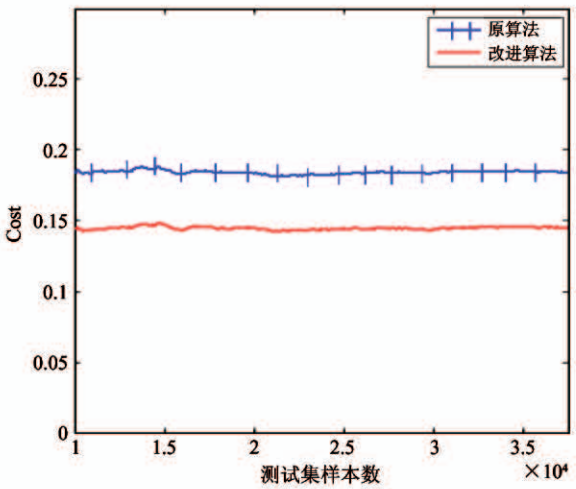

图 2　Cost 随测试集样本数变化曲线图

由图 2 看出，随着测试子集样本数的改变，Cost 的值基本不变，而且改进随机森林算法的 Cost 的值低于原算法的 Cost 值，说明此结果具有普遍性，也进一步验证了本文所提出的改进的随机森林算法在入侵检测中与原算法相比性能更好。

5　结　论

本文提出了一种基于 SDN 的智能入侵检测系统模型，该模型能够充分发挥软件定义网络的集中管理、全局控制等优势。同时，可以灵活使用各种机器学习算法进行流分类，提高检测效率。针对随机森林算法存在的不足，尤其针对需要分类的入侵流具有多特征、数据不平衡等特点，提出了一种改进的随机森林算法，使用 KDD CUP99 数据集进行性能仿真和对比分析。从实验结果来看，相对于原算法，改进的随机森林算法在检测精度、代价等指标上都明显得到了提升，证明了该改进算法在智能入侵检测系统中的有效性。

参考文献

[1] 袁沛沛. 网络安全入侵检测技术[D]. 西安:西安建筑科技大学信息与控制工程学院,2008: 1-58

[2] 蒋建春,马恒太,任党恩,等. 网络安全入侵检测:研究综述[J]. 软件学报, 2000,11(11):1460-1466

[3] Zhang X Y, Wang S P, Yun X C, Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset[J]. *IEEE Transactions on Neural Networks and Learning Systems* (*TNNLS*), 2015, 26(12): 3034-3044

[4] Zhang X Y, Shi H C,Zhu X B,et al. Active semi-supervised learning based on self-expressive correlation with generative adversarial networks [J]. *Neurocomputing*, 2019, 345:103-113

[5] Zhang X Y, Shi H C, Li C S,et al. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Palo Alto, USA, 2019: 1-8

[6] 张勇, 姬生生, 王闵毅. 基于 SDN 架构的 WSN 入侵检测技术[J]. 河南大学学报, 2015,45(2):216-221

[7] Mehdi S A, Khalid J, Khayam S A. Revisiting Traffic Anomaly Detection Using Software Defined Networking [M]. Berlin Heidelberg: Springer, 2011: 161-180

[8] 姚东, 罗军勇, 陈武平, 等. 基于改进非广延熵特征提取的双随机森林实时入侵检测方法[J]. 计算机科学,2013,40(12):192-198

[9] Giotis K, Argyropoulos C, Androulidakis G. Combining OpenFlow and sFlow for an effective and scalable anomaly detection and mitigation mechanism on SDN environments [J]. *Computer Networks*, 2014, 62: 122-136

[10] Jankowski D, Amanowicz M. On efficiency of selected machine learning algorithms for intrusion detection in software defined networks[J]. *International Journal of Electronics and Telecommunications*, 2016,62(3):247-252

[11] 史振华, 刘外喜, 杨家烨. SDN 架构下基于 ICMP 流量的网络异常检测方法[J]. 计算机系统应用, 2016, 25(4):135-142

[12] 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(3):32-37

[13] 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, 2018,1:49-55

[14] 曹正凤. 随机森林算法的优化研究[D]. 北京:首都经济贸易大学统计学院, 2014: 1-135

[15] University of California. UCI KDD archive[EB/OL]. http://kdd. ics. uci. edu/databases/kddcup99/kddcup99: University of California, 2005

SDN based intelligent intrusion detection system model and algorithm

Ma Lin [*], Zhang Shasha [**], Song Shuyu [**], Wang Lei [***]

([*] Information Section, PLA Air Force Hangzhou Special Service Convalescent Center, Hanghzhou 310012)

([**] Institute of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027)

([***] Academy of Broadcasting Science, National Radio and Television Administration, Beijing 100866)

Abstract

An intelligent intrusion detection system model is proposed. The model takes the advantage of central management and global control in software defined networking (SDN). Different machine learning algorithm can be used dynamically for data flow detection. This can improve the performance of intrusion detection system. To solve the problem of diverse characters and unbalance flow data distribution, this work improves current random forest algorithm. The metrics of decision tree will be update dynamically to improve the accuracy of classification. KDD CUP99 dataset is used for algorithm training and testing. Simulation results show that the improved random forest algorithm has good performance on detection accuracy and cost. It proves the efficiency of new model and new algorithm.

Key words: software defined networking (SDN), intrusion detection, machine learning, random forest