文章编号:1671-4598(2020)03-0062-05 DOI:10.16526/j.cnki.11—4762/tp.2020.03.014 中图分类号:TP393.08 文献标识码:A

基于大数据技术的网络异常行为检测模型

刘建兰，覃仁超，何梦乙，熊 健
(西南科技大学 计算机科学与技术学院，四川 绵阳 621010)

摘要：针对传统的网络异常检测受数据存储、处理能力的限制，存在准确率较低、误报率较高以及无法检测未知攻击的问题；在Spark框架下结合改进的支持向量机和随机森林算法，提出了一种基于大数据技术的网络异常行为检测模型；使用NSL-KDD数据集进行了方法验证，表明该方法在准确率和误报率方面明显优于传统的检测算法，整体检测的准确率和误报率分别为96.61%和2.92%，DOS、Probe、R2L和U2R四种攻击类型的准确率分别达到98.01%、88.29%、94.03%和66.67%，验证了方法的有效性。

关键词：大数据技术；网络异常行为；支持向量机；随机森林；模拟退火

Network Abnormal Behavior Detection Model Based on Big Data Technology

Liu Jianlan, Qin Renchao, He Mengyi, Xiong Jian
(School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China)

Abstract: The traditional network anomaly detection is limited by data storage and processing capabilities, which has the problems of low accuracy rate, high false alarm rate and unable to detect unknown attacks. To resolve this, combined with the improved Support Vector Machine and Random Forest algorithm in Spark, a network abnormal behavior detection model based on big data technology is proposed. The method is verified by NSL-KDD data set, which shows that the method is superior to the traditional detection algorithm in accuracy and false positive rate. The accuracy and false positive rate of the overall detection are 96.61% and 2.92%, The accuracy rates of DOS, Probe, R2L and U2R respectively were 98.01%, 88.29%, 94.03% and 66.67%, which verified the effectiveness of the method.

Keywords: big data technology; network abnormal behavior; support vector machine; random forest; simulated annealing

0 引言

随着互联网、云计算、大数据、机器学习等技术的快速发展，目前已经全面进入大数据时代，网络数据迅速增长且类型多样，网络规模越来越大，设备更加多样化[1]。在日益复杂的网络环境中，网络攻击行为越来越复杂，手段也越来越隐蔽，且潜伏周期很长，新型攻击多种多样[2-3]。面对如此复杂严峻的形式，迅速、准确地区分正常和异常网络行为，有效防御非法入侵，成为网络安全领域研究的重点。

自20世纪80年代Denning提出网络入侵检测模型以来[4]，学者们提出了很多网络异常行为检测方法，目前异常行为检测方法主要包含基于统计分析的方法、基于数据挖掘的方法、基于特征规则的方法和基于机器学习的方法[5-6]。其中基于机器学习的网络异常检测是学者们研究的重点，但由于传统的存储空间及计算能力有限，机器学习的检测效果不佳。随着大数据技术的发展，提升了对大量数据的采集、存储及处理能力。将机器学习和大数据技术相结合应用到网络异常行为检测中[7]，通过对数据流实时检测和历史数据的离线分析，可以识别更隐蔽及复杂的网络攻击，而且极大提高了网络异常检测的准确率。因此，已成为近年来安全领域的一大热门。

Son S[8]提出了一种基于apache的数据存储和分析架构，用于处理大量的Hadoop日志数据，并设计实现了三种有效的异常检测方法，进行了对比实验。S. Zhao等[9]提出了一种基于机器学习算法的实时网络流量异常检测的框架，所提出的原型系统结合了机器学习技术和现有的大数据处理框架，并利用校园网数据进行了结果测试。Wang H等人[10]为了解决入侵检测算法耗时大、数据分类占用内存大、单点检测效率低的问题，提出了一种基于Spark平台的并行主成分分析和支持向量机组合算法。B. Senthilnayaki等人[11]在模型预训练过程中首先采用遗传算法进行特征选择，然后应用SVM进行分类，实验表明检测效果较好。S. Sahu等人[12]采用决策树构建了网络入侵检测系统，对网络数据包中的正常、已知攻击和未知攻击进行分类。结果表明，决策树算法对未知攻击具备识别检测能力。王萍[13]提出了一种基于大数据技术的网络异常行为分析的方法，可以识别更隐蔽及复杂手法的攻击行为，并基于该理论方法实现了一个网络异常行为分析监测系统。李若鹏[14]设计与实现了一个基于大数据的网络异常行为检测平台，实现多种海

收稿日期:2019-08-15； **修回日期**:2019-08-29。
基金项目:国防基础科研计划项目(JCKY2017404C004)；四川省教育厅项目(17zd1119,18sxb022)；四川省组织部项目(17sjjg02)。
作者简介:刘建兰(1993-)，女，甘肃陇西县人，在读硕士研究生，主要从事网络安全方向的研究。
通讯作者:覃仁超(1978-)，男，四川武胜县人，博士研究生，副教授，主要从事网络安全、智能计算方向的研究。

量数据的可靠高效的接入、存储以及分析。

综上所述，面对大数据时代网络复杂多样的数据，传统的数据处理平台已无法对海量数据进行高效、全面的处理；传统单一的机器学习算法存在对已知异常检测效果较低，无法检测未知攻击行为的问题。针对以上问题，结合大数据技术和机器学习，提出了一种基于大数据技术的网络异常行为检测方法。详细介绍了该模型的总体架构设计，该模型架构包括数据采集与预处理层、数据分析层、数据存储层和可视化层；结合并行化的模拟退火优化的支持向量机和随机森林算法构建网络异常行为检测模型；利用 Flume、Kafka、Spark streaming 技术以及网络异常行为检测模型实现数据流的实时检测，并将结果可视化展示。利用 NSL－KDD 数据集对本文提出的方法和相关算法进行对比测试，实验结果表明，本文提出的方法在准确率和误报率方面明显优于其他算法，并测试了网络异常行为实时检测流程。

1　相关技术介绍

1.1　支持向量机

支持向量机（support vector machine，SVM）是一种基于 VC 维度理论和结构风险最小化原理的机器学习算法，SVM 是一种二分类模型。基本思路就是在特征空间中找到一个能够满足限制条件的最优超平面，该超平面能够把数据集中的点准确的分类，而且使两侧的点到该超平面的距离最大。其学习策略是求解超平面问题，可以等价为求解相应的凸二次规划问题。SVM 在一定程度上克服了传统机器学习算法中过拟合、维数灾难和局部极小值等不足，具有很强的学习和泛化能力等。

在线性可分的情况下，给出一组训练样本，可将其分类的超平面用一个分类函数表示：

$$f(x)=\omega\cdot x_i+b \tag{1}$$

这里用 1 和－1 分别表示两类样本，当 $f(x)>0$ 时，x 属于 1 的样本点，反之属于－1 的样本点。然而这样的分类超平面在分类问题中并不唯一，需要寻找并确定最佳超平面。因此，将分类问题转化为二次规划问题的求解，公式如下：

$$\begin{cases}\min\left(\dfrac{1}{2}\|\omega\|^2\right)\\ s.t.\ y_i((\omega\cdot x_i)+b)\geqslant 1, i=1,\ldots,l\end{cases} \tag{2}$$

在线性不可分的情况下通过引入非负松弛因子和惩罚因子，允许存在少量分类错误的样本，其约束条件为：

$$\begin{cases}\min\left(\dfrac{1}{2}\|\omega\|^2\right)+C\sum\limits_{i=1}^{l}\xi_i\\ s.t.\ y_i((\omega\cdot x_i)+b)\geqslant 1-\xi_i,\xi_i\geqslant 0\end{cases} \tag{3}$$

基于 spark ML 的 SVM 是具有一个 Hinge 损失函数的线性分类，其中 Hinge loss 用于最大间隔分类，Hinge 函数的标准形式：

$$L(w;x,y):=\max\{0,1-yw^Tx\} \tag{4}$$

默认情况下，线性支持向量机通过 L2 正则化训练，其中正则化系数的设置对训练模型至关重要。支持向量机在高维或无限维空间中构造一个超平面或一组超平面，输出 SVM 模型结构。对于新的数据节点 x，则模型基于 w^Tx 值来预测，如果 $w^Tx>0$，则结果为正常，否则为异常。

1.2　改进的模拟退火优化支持向量机

模拟退火（simulated annealing，SA）算法来源于固体退火原理，即将固体加热到足够高的温度，当温度很高时，内能比较大，内部粒子处于快速无序运动状态，而在温度缓慢冷却时，固体的内能减小，内部粒子逐渐趋于有序，最终趋于平衡状态，内能达到最小，粒子最为稳定。在实际应用中，将优化问题的最优解比作退火过程中能量处于最低的稳定状态，也就是温度达到最低点的状态或概率分布中具有最大概率的状态。其实现主要是在局部搜索的过程中引入了随机扰动的机制，对于产生的新解，如果好于旧解，则无条件接受或更新，如果差于旧解，则以一定概率接受或更新。

模拟退火算法存在两个主要缺点。首先，退火速度问题。对于复杂的目标函数，为了找到全局最优解，设置的初始温度足够高，冷却速度足够慢，导致计算量很大。第二，最优解的丢失和循环问题。由于模拟退火的随机性，可以跳出局部最优，但也可能丢弃当前最优解，以及对某个已经访问过的解进行多次重复的搜索。为了克服以上缺点，本文对传统的模拟退火算法进行改进。改进的 SA 和传统的 SA 主要区别在于：

1）温度冷却函数在传统的 SA 中是恒定的，而在改进的 SA 中采用分段函数。起初温度较高的时候，采用递减较快的函数来降低温度，在退火的后期，采用递减较慢的函数。温度冷却函数如下：

$$T(t)=\begin{cases}\dfrac{T_0}{\ln(k)}, k<m\\ T_0*0.99^k, k\geqslant m\end{cases} \tag{5}$$

其中：m 是预先设定的值，t 是迭代次数，T_0 为起始温度。

2）内循环次数在传统的 SA 中是恒定的，而在改进的 SA 中，如果当前温度的解在足够长的时间内保持不变，超过了算法预设的阈值，则跳出循环。

3）改进的 SA 增加记忆功能来保存最近访问的解和历史最优解。

改进的模拟退火算法的整体步骤大致如下：

1）设定初始温度 T_0，随机产生初始解 ω_0，并计算目标函数值 $f(\omega_0)$，终止温度 T_e，内迭代因子 γ，内循环阈值 θ，设定降温函数分段点 m；

2）通过扰动当前解 ω 产生新解 ω'，若 ω' 内存中存在，则丢弃，否则计算相应的目标函数 $f(\omega')$，得到 $\Delta f=f(\omega')-f(\omega)$；

3）当 $\Delta f<0$ 时，令 $\omega=\omega'$，当 $\Delta f>0$ 时，按概率 $p=\exp\left(\dfrac{-\Delta f}{T_i}\right)$ 接受新解 ω'，并将 ω 保存到内存中；

4）持续在邻近区域内生成新解并重复步骤 2）、3），直到达到内循环阈值 θ；

5）判断温度是否到达终止条件 $T_i < T_e$，如果满足则算法结束，否则根据降温函数降低温度，继续返回步骤 2）。

传统基于 spark ML 的 SVM 正则化系数靠经验设置，存在较强的主观性，不同的值可能会得到不同的结果。然而模拟退火算法非常适合求解优化问题的最优解。因此本文利用改进的 SA 算法优化 SVM 参数的设定，使模型准确率达到全局最优。

1.3 随机森林

随机森林（random forest，RF）算法是一种通过随机的方式建立多棵决策树从而形成一片森林的过程，其输出的类别基于每个决策树输出类别的众数而定。通过 Bagging 方法抽取森林中每棵决策树的训练样本，利用这些子训练集并行训练一套决策树，每棵决策树相互独立。对于测试数据集，在每个决策树做出决策后，随机森林模型将通过投票决定最终结果。随机森林算法是一种集成学习，是一种多个决策树组合成为一个强分类器的过程。在实际应用中，随机森林分类精度高、多次预测结果稳定，且降低过度拟合模型的风险。随机森林的算法流程如下：

1）从原始的数据集中采用 Bagging 思想进行有放回抽样，抽取 K 个和原数据集数量相等的训练子集 $\{D_1, D_2, \ldots, D_k\}$；

2）使用每个训练集构建子决策树，在决策树的每个结点，从全部特征中随机采样有 $m(m \leqslant M)$ 个特征的子空间，并基于这 m 个特征计算所有可能的分裂方式，选择最佳分类方式（如 entropy 度量最大）用于该结点的分裂，继续这个过程直到达到某个预设的停止条件。本文中预设的停止条件为决策树的深度达到设定的值。

3）把 K 个不剪枝的树集成得到一个随机森林，随机森林的分类决策采用投票的方式。当新的数据需要通过随机森林得到分类结果，就可以通过对子决策树的判断结果的投票，得到随机森林的输出结果了。

2 基于大数据技术的网络异常行为检测

2.1 网络异常行为检测模型设计

基于大数据的网络异常行为检测模型如图 1 所示，包括数据采集层、数据预处理层、数据存储层、数据分析层、可视化层。

1）数据采集层。负责从多源异构的数据源采集数据，数据源指用来支撑数据分析和实时监测的各类原始安全数据。通过 Flume 采集数据发布到 Kafka 上。

2）数据预处理层。主要负责对收集到的原始数据进行数值化、归一化的处理工作，方便后面网络异常行为分析使用。

3）数据分析层。负责大批量数据的计算，主要分为离线分析和实时检测两部分。通过基于 Spark ML 的机器学习算法实现离线分析，完成网络行为检测模型的构建。实时检测基于 Spark streaming 框架通过消息订阅读取 Kafka 中的实时数据流，并依据检测模型对实时数据流进行检测。

4）数据存储层：按照数据的类型及使用情况实现各类数据的存储与集中管理，本文将原始数据存储到 HDFS，部分分析处理结果存储在关系型数据库 Mysql 中。

5）可视化层：主要用于结果的展示和查询，方便查看和操作。

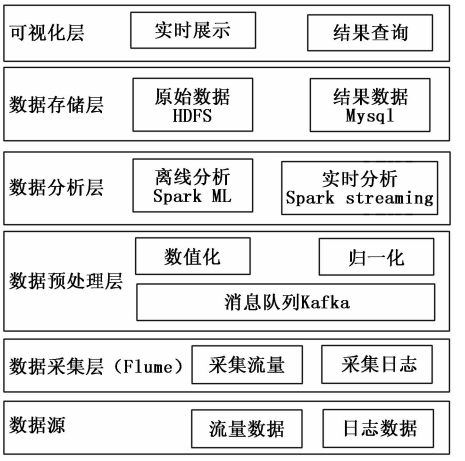

图 1 基于大数据的网络异常行为检测模型

2.2 基于 Spark 的并行 SA _ SVM _ RF 实现

传统串行机器学习算法处理大规模数据时速度会降低。针对此问题，本文将并行框架 Spark 与前面描述的模拟退火、支持向量机和随机森林算法结合，实现了基于 Spark 的并行化 SA _ SVM _ RF 模型。在不降低分类精度的前提下，可以明显提高算法的分类效率。

Spark 有两个内部机器学习库，分别为基于弹性分布式数据集（Resillient Distributed DataSet，RDD）的 MLlib 和基于 DataFrame 的 ML。若数据集结构复杂需经过多次处理或数据需要集成学习的思想进行综合预测时，Spark ML 明显优于 Spark MLlib，因此，本文算法是基于 Spark ML 实现的。

基于 Spark 的并行化 SA _ SVM _ RF 算法的实现步骤如图 2 所示。在模型训练开始之前，将数据集上传到 HDFS 分布式文件存储系统。Spark 集群的任务调度将数据集划分为 K 个部分，每个数据子集都会在 Executor 中创造一项新任务，并分配计算资源。然后在 Spark 集群上先后进行并行 SA 优化的 SVM 和 RF 算法训练以获得 K 个模型，最后，将局部最优模型进行合并操作，输出全局最优模型。

1）数据集切分。利用 textFile（）函数从 HDFS 读取数据集并将其转换为具有列名的 DataFrame，StringIndexer（）对字符串的标签列进行编码，同时按照 partition 类中的分割规则将原数据集随机切分成大小适中的独立数据分区，并分布到各 Executor 上。

2）Map 阶段。每个 Executor 将数据子块里的原始 DataFrame 用 transform（）方法转化成适合 SVM 算法输入格式的 DataFrame，创建一个 Pipeline 对象，同时设置对应

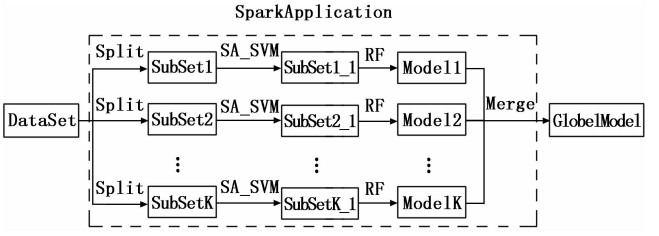

图 2　基于 Spark 的并行化 SA _ SVM _ RF 算法实现流程

的多个顺序执行的 PipelineStage，构建 SA 优化的 SVM 算法对象并设置参数，然后调用 fit（）方法对 SA 优化的 SVM 分类器进行迭代训练产生一个 SA _ SVMModel。接着调用 SA _ SVMModel 的 transform（）方法得到新的 DataFrame，并将其传入下一个 Stage（即 RF 算法），RF 算法用同样的方法得到 Model。

3）Combine 阶段。Combine 阶段位于 Map 和 Reduce 之间，通过 Combiner 对象，将所有数据子块以及 Map 阶段得到的局部模型 Model 洗牌后合并，交予 Reduce 阶段。

4）Reduce 阶段。接收 Combine 阶段的结果并对合并后的分类模型进行测试。按照准确率高误报率低的测试标准输出全局最优分类器 GlobelModel。

2.3　网络异常行为建模与实时检测

本文提出的网络异常行为检测主要分为两个阶段：离线构建检测模型和实时检测数据流，网络异常行为建模与实时检测的流程图如图 3 所示。

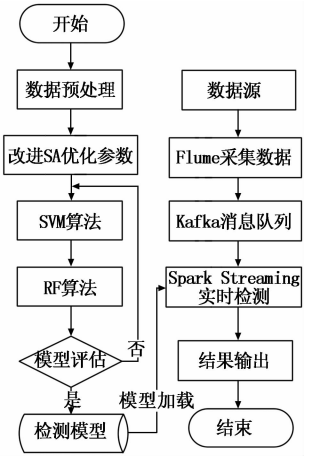

图 3　网络异常行为建模与实时检测流程

基于并行化 SA _ SVM _ RF 的网络异常行为模型构建主要是将模拟退火优化的支持向量机和随机森林有效结合，充分利用各自的优势来提高网络异常检测的效果。模型构建主要包括：数据预处理、模型训练、模型评估。数据预处理主要是对数据集进行数值化、归一化处理。训练阶段主要是先利用模拟退化优化的支持向量机将数据集分成两类，然后利用随机森林再次分类形成五类，得到训练模型。模型评估主要是利用测试数据评估训练的网络异常行为检测模型的好坏，并将测试结果反馈到模型训练阶段来不断调整模型，直至产生满足要求的模型。

基于 Flume _ Kafka _ Spark Streaming 的网络异常实时检测步骤如下：

1）Flume 实时采集数据，并发送到 Kafka 的 topic 中；

2）采用 Spark streaming 实时计算框架通过消息订阅读取 Kafka 的 topic 中实时数据流，利用检测模型进行实时检测，并将结果保存在 mysql 数据库中；

3）以可视化的形式输出最终的异常检测结果。

3　实验结果与分析

3.1　测试环境和测试数据集

测试环境 Spark 集群包括 3 个虚拟机，其中一个主节点，两个从节点，每个节点都具有相同的配置，设置 2 核 CPU，4 GB 内存，20 G 硬盘。测试环境中配置的组件主要包括 Hadoop、Zookeeper、Flume、Kafka、Spark。

本文使用 NSL－KDD 数据集作为模型的训练和测试数据集，同时用多种算法对模型进行比较验证。使用 KDD CUP99 中未标签的数据集模拟实时产生数据流测试网络异常行为实时检测流程。

3.2　算法有效性验证

使用 NSL－KDD 数据集将本文算法与四种传统算法进行比较验证。表 1 表示不同算法分别在 KDDTest^{+} 和 KDDTest^{-21} 上的整体检测效果。表 2 表示不同算法在 KDDTest^{+} 和 KDDTest^{-21} 上对四种攻击类型检测的准确率。

表 1　不同模型在 KDDTest＋和 KDDTest－21 上的整体检测效果

分类模型	KDDTest^{+} 测试集		KDDTest^{-21} 测试集	
	准确率(%)	误报率(%)	准确率(%)	误报率(%)
DT	81.05	1.76	63.97	7.95
SVM	75.24	2.98	50.91	13.24
RF	80.67	2.17	63.26	9.53
Bayes	76.56	6.16	55.77	26.81
本文算法	96.61	2.92	94.41	12.78

从表 1 中可以看出，与传统机器学习算法相比，本文算法的整体准确率都高于以上四种传统机器学习算法，误报率要略微低于决策树和随机森林算法，高于其余两种算法。

表 2　四种攻击类型在 KDDTest^{+} 和 KDDTest^{-21} 上检测的准确率

分类模型	KDDTest^{+} 测试集				KDDTest^{-21} 测试集			
	DOS (%)	Probe (%)	R2L (%)	U2R (%)	DOS (%)	Probe (%)	R2L (%)	U2R (%)
DT	75.64	67.00	1.99	7.00	58.18	66.74	1.99	7.00
SVM	66.47	45.27	0.00	0.00	44.84	45.00	0.00	0.00
RF	79.24	69.15	0.15	3.00	62.00	69.69	0.15	3.00
Bayes	64.16	89.76	19.86	5.00	43.34	89.80	19.86	5.00
本文算法	98.01	88.29	94.03	66.67	95.43	88.99	94.19	66.67

从表 2 中可以看出，与传统机器学习算法相比，本文提出的算法除了对 Probe 的检测准确率略低于 Bayes 算法外，对四种攻击类型的检测准确率明显优于四种传统算法，

尤其是提升了对 R2L 攻击和 U2R 攻击的检测能力。

将本文算法与文献［14］进行比较验证，对比结果如图 4 所示。

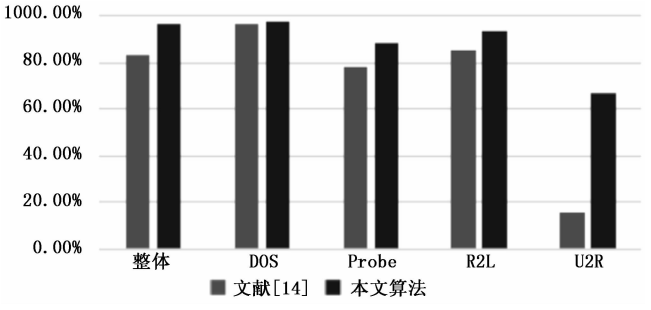

图 4 不同算法检测准确率对比结果

由图可知，本文算法的整体检测准确率比文献［18］高 12.86%。从四种攻击类型的检测准确率来看，本文算法对 DOS、Probe、R2L 和 U2R 的检测准确率比文献［18］分别提高了 1.58%，10.39%，8.27%，51.52%。综合来看，本文算法具有更高的检测准确率，进一步提高了入侵检测分类效果。

3.3 网络异常行为实时检测测试

网络异常实时检测情况如图 5 所示。其中左上角的柱状图展示了一天内网络实时检测到的各类别的总数，右上角的仪表盘展示了一天内网络的总量，下面的散点图可以看出某一时间点出现了某种网络攻击。

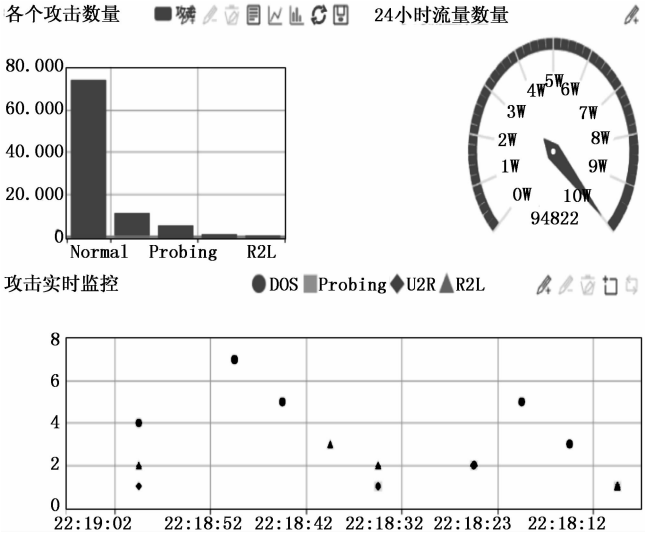

图 5 网络异常实时检测展示

网络异常行为实时检测过程中网络总量实时展示如图 6 所示。从图中可以看出网络总量的实时变化情况。

网络行为历史数据查询部分结果如图 7 所示。可以通过设置网路异常类型、起始时间和结束时间查询网络行为数据记录。

4 总结与展望

针对现有网络异常检测存在的问题，本文将大数据技术和机器学习结合，提出了一种基于大数据技术的网络异常行为检测方法。首先利用基于 Spark 的并行 SA _ SVM _ RF 算法构建网络异常行为检测模型。然后运用 Flume、Kafka、Spark steaming 等技术框架实现了数据流的高效采集、存储和处理，完成网络异常行为实时检测。并且使用 NSL－KDD 数据集进行实验验证，结果表明，本文提出的方法有效提高了网络异常行为检测准确率、降低了误报率，具有较好的检测性能，并验证了大数据环境下实时数据流的处理。下一步工作中，将进一步优化本文提出的算法来完善模型，使之更适用于实际应用。

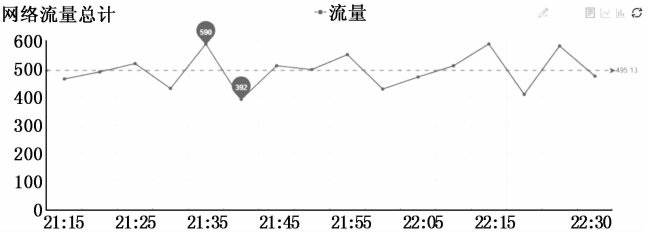

图 6 网络总量实时展示

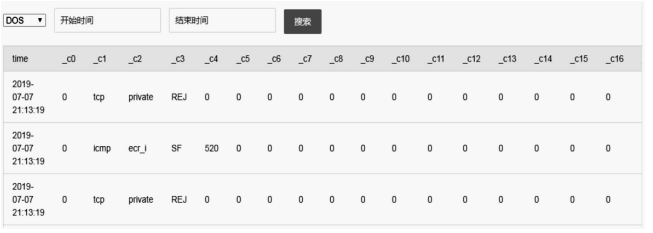

图 7 网络行为历史数据查询展示

参考文献：

［1］陈兴蜀，曾雪梅，王文贤，等. 基于大数据的网络安全与情报分析［J］. 工程科学与技术，2017，49（3）：1－12.

［2］国家计算机网络应急技术处理协调中心，CNCERT/CC2018 年网络安全报告［R］. 2018.

［3］于 洋. 入侵检测系统中特征选择算法与模型构建方法的研究［D］. 兰州：兰州大学，2017.

［4］Denning D E. An intrusion－detection model［J］. IEEE Transactions on software engineering，1987（2）：222－232.

［5］尹传龙. 基于深度学习的网络异常检测技术研究［D］. 郑州：战略支援部队信息工程大学，2018.

［6］Canbay Y，Sagiroglu S. A Hybrid Method for Intrusion Detection［A］. IEEE，International Conference on Machine Learning and Applications［C］. IEEE，2016：156－161.

［7］Belouch M，El Hadaj S，Idhammad M. Performance evaluation of intrusion detection based on machine learning using Apache Spark［J］. Procedia Computer Science，2018，127：1－6.

［8］Son S，Gil M S，Moon Y S. Anomaly detection for big log data using a Hadoop ecosystem［A］. 2017 IEEE International Conference on Big Data and Smart Computing（BigComp）［C］. IEEE，2017：377－380.

（下转第 71 页）