文章编号:1671-4598(2021)08-0168-04 DOI:10.16526/j.cnki.11—4762/tp.2021.08.033 中图分类号:TP316 文献标识码:A

大数据时代网络安全及预测技术研究

梁永坚[1],黄 慷[1],韦 田[1],黎锐杏[2]
(1.国网安徽省电力有限公司 滁州供电公司,安徽 滁州 239000;
2.中能博望(北京)科技有限公司,北京 102488)

摘要:大数据时代信息技术的快速发展,依托于各类硬件防护设备的网络体系架构的异构数据量每天以指数级的量级递增,基于传统的网络安全防护技术无法有效地适用于具有海量数据的特征网络安全和分析预测等工作,因此海量数据的保存、使用以及分析等信息挖掘和数据分析预测逐步成为社会各界重视和当前的研究趋势;以海量的异构数据为研究对象,识别网络安全大数据的典型特征,结合情报预测的主要方法,创新性地提出了大数据特征下的网络安全预测分析技术,提高网络安全风险识别和预测、预警能力,有效地改善网络防护效果。

关键词:大数据;机器学习;网络安全预测

Research on Network Security and Prediction Technology in Big Data Era

LIANG Yongjian[1], HUANG Kang[1], WEI Tian[1], LI Ruixing[2]
(1. State Grid AnHui Electric Power Company Chuzhou Power Supply Company, Chuzhou 239000, China;
2. ZhongNeng Bowang (Beijing) Technology Co., Ltd., Beijing 102488, China)

Abstract: With the rapid development of information technology in the era of big data, the amount of heterogeneous data based on the network architecture of all kinds of hardware protection equipment increases exponentially every day. Based on the traditional network security protection technology, it can not be effectively applied to the work of network security and analysis and prediction with the characteristics of massive data, so it is necessary to save, use and analyze massive data. Data analysis and prediction has gradually become the focus of the community and the current research trend. Taking massive heterogeneous data as the research object, this paper identifies the typical characteristics of network security big data, combined with the main methods of intelligence prediction, innovatively puts forward the network security prediction analysis technology under the characteristics of big data, improves the ability of network security risk identification and prediction, and effectively improves the network protection effect.

Keywords: big data; machine learning; network security prediction

0 引言

我国进入到21世纪后,特别是近10年来,网络科技发展突飞猛进,大数据、云计算、物联网等技术逐步由新兴转为普遍,人类进入了海量信息的时代。各种移动设备的普及应用等带来了新的数据时代[1]。应运而生的是各种网络安全事件频繁出现。其根本原因在于大数据环境下的网络安全预测技术的瓶颈[2],基于传统的网络安全防御技术无法应对海量数据特征下的网络入侵,因而局面较为被动。基于此,研究大数据时代背景下网络安全海量数据的信息分析、提取以及安全问题的预测技术迫在眉睫。

1 分布式模型训练架构

1.1 大数据网络安全预测关键技术

海量数据分析挖掘及预测预警的基础以及核心在于大量异构、多维数据的清洗、降维和同构化等预处理工作[3]。在此核心的基础上,进行数据的分类、学习、训练形成安全预测模型,并结合实际情况,进行网络安全态势的感知和预警。

1)异构、多维数据的清洗:首先结合各类交换机、路由器、网关机、传感器等采集设备的网络安全日志,进行数据的预处理,建立数据关联关系,实现数据融汇,按照固定的规范,将日志的数据进行标准化处理,并统一保存,做好进一步日志分析的准备[4]。

2)多层级网络安全评估。通过建立网络安全多层级的评估模型,结合网络安全威胁评估算法,提炼、获取、形成网络威胁列表,根据目前常见的网络攻击的行为、攻击方式、网络异常的状态、主动攻击的手段等完成建模好训练,从中提取攻击核心代码、异常流程状态数据,并标记、学习、训练异常行为,结合分类计数,进行网络安全的基本评估。

3)网络安全态势预测[5]。将多层级多维度的网络安全评估模型与当前获取的网络安全事件结合,建立网络安全状态图谱,整体分析完成安全态势预测。

后续,结合目前常用的Gognos架构、帆软报表等数据可视化分析常用的工具,建立关系数据模型,以图形化额

收稿日期:2020-11-24; 修回日期:2021-02-04。
作者简介:梁永坚(1964-),男,广西贵港人,研究生,高级工程师,主要从事电力系统规划方向的研究。
引用格式:梁永坚,黄 慷,韦 田,等.大数据时代网络安全及预测技术研究[J].计算机测量与控制,2021,29(8):168-171,195.

形式完成驾驶舱、预警图等可视化图形展示。

目前针对数据的安全态势分析研究主要侧重于数据的预测方面，但是数据处理性能在面对互联网万物感知的海量数据时，性能降低非常大[6-7]，传统的安全态势感知模型已无法适应大数据时代，另外由于科技发展带来的新型的攻击模式层出不穷，如果对各种不同类型的攻击做到精准预测和感知，需要对攻击模型进行不断地学习、训练，并更新攻击库[8-9]。基于上述问题，本文采用分布式的技术对数据进行处理和清洗。在处理数据过程中主要采用有别于传统的机器学习的方法，提出了基于神经网络的采样降维和聚类算法，在此基础上进行网络安全预测。

第一步：使用基于开源平台的 Hadoop 进行分布式数据处理，将通过内存分析处理的数据进行自动划分，将数据随机分布到不同的节点完成基本的处理分析。

第二步：分布式处理完成的数据需要进行降维和聚类，通过改进的聚类算法和基于特征值分解的降维办法进行降维，完成分析预测前的数据清洗。

第三步，清洗后的数据挖掘，针对大数据时代异构数据，采用基于误差反馈的神经网络算法挖掘数据流的深层特质，通过循环、往复、迭代持续进行模型训练，提炼训练模型参数，完成数据的预测，并合理提升预测的准确性。

1.2　分布式数据处理框架

基于传统的神经网络模式主要采取寻找目标函数最小化的方法进行处理模型的参数训练，其不足在于机器学习效率低、标准化能力差，是应用于海量网络安全数据提取的掣肘因素。

考虑到传统算法的不足，设计了改进的前馈神经网络模型，基于 Hadoop 的分部署数据处平台，从算法和算力上解决训练模型的复杂性问题，设计了基于分治策略的分布式模型训练算法。

Hadoop 分布式数据处理平台的核心组件包括 HDFS (hadoop distributed file system) 分布式文件系统以及基于 MapReduce 的并行化处理编程单元。通过分布式文件系统将海量的预处理后的日志文件进行分布式的存储，在这个过程中，通过 MapReduce 完成并行高速运算，其在海量数据环境下的并行计算展示出了强大的能力，尤其适合万物互联状态下的海量网络安全日志数据的处理。因此，本文基于 Hadoop 的优势特点，建立了基于分治策略的分布式模型训练算法。具体算法如图 1 所示。

该算法主要采用的是前反馈式训练神经网络架构，网络数据记录在该架构中包含两种传输路径。路径之一的起始点为输入层，途径隐藏层，然后到达传输层；路径之二为前向反馈型路径，起点为输出层，反向传输到隐藏层。两种路径互相结合、互相补充的模式，使得该架构具有较高的自我训练、自我反馈和协调的能力。通过输入的元数据特征持续的修改框架的训练模型，达到自我调整的目的，该框架尤其适合对于没有经过驯良的特征数据记录的识别，且在海量的网络安全数据集合汇总，该架构对比传统的神经网络算法识别数据的非线性内在规律较高。本文所设计的基于 Hadoop 的分布式数据处理架构其结构相对复杂，具有对原始数据中的异常数据值敏感性不足，对于脏数据、数据噪声的兼容性较好的优势。具体的数据处理流程如表 1 所示。

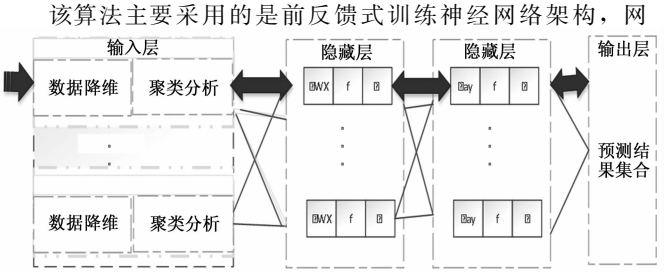

图 1　基于神经网络的网络安全预测算法框架

表 1　基于神经网络的网络安全预测模型的数据处理的流程

步骤	算法过程
MAP 组合阶段	MAP 分类配对阶段
输入	全部网络安全数据集合
输出	数据集合所对应的权重和特征值的阈值，如＜a，Δa＞
第 1 步	根据输入层的数据进行初步计算
第 2 步	根据第一步的计算结果计算隐藏层的输出
第 3 步	由输入层和隐含层计算输出层的结果值
第 4 步	比对输出结果和期望值，计算方差
第 5 步	根据方差的结果反馈调整权重和阈值
规约阶段	Reduce 规约
输入	MAP 组合阶段的输出值
输出	训练集的权重、模型的阈值以及更新后的训练集
第 1 步	定位已经更新的权重和阈值(A)
第 2 步	设置、取得增加值 $\Delta a1$
第 3 步	采取 Hadoop 内置算法计算更改值，并输出到结果训练集

2　数据预处理－PSO－K－Means 聚类算法

数据预处理的第一步为数据清洗，其主要是作用是进行错误数据的识别和纠正，通过两个关键步骤完成数据的清晰；第二步采用分布式聚类算法实现数据聚类，主要对网络安全设备收集的海量的多维数据进行聚类，聚类之前需要做必要的工作为对数据进行统一化处理，也就是降维操作，其作用是提升聚类的效率，提升大数据的处理速度。本文采用的是基于维度特征分析的降维算法，其前提条件是需要收集元数据的协方差矩阵的特征向量和特征值，并结合标准化公式，导出特征向量以及对应的特征值，在此基础上，进行数据的降维操作。处理海量数据的降维算法需要与分布式技术结合，其具体的过程如图 2 所示。

数据预处理过程中的特征分解算法主要采取的是行数与列数保持一致的对角矩阵分解算法，由于原始矩阵的行数与列数不是完全相同，因此该算法无法直接处原始矩阵，通常采取的措施为对矩阵进行初步的降维处理，以得到对

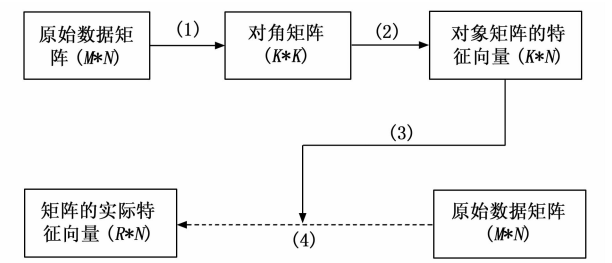

图 2 数据清洗－数据降维处理

称矩阵。具体的步骤如表 2 所示。

表 2 数据清洗－元矩阵降维过程

步骤	算法过程
第 1 步	元矩阵处理，采用变换模型，获得对角矩阵($K*K$)
第 2 步	使用临时矩阵，分解对角矩阵的特征向量及其对应的特征值。
第 3 步	统一处理元矩阵以及特征向量矩阵。
第 4 步	形成预处理的结果，降维结束。

经过降维得到的对称特征向量矩阵后，采取改进后的迭代求解的聚类分析算法——K 均值聚类算法（K－means clustering algorithm），经典的 K－Means 算法目前基本使用在单机的情况下，算法执行效能较低，面对大数据环境下，其可伸缩性不足，且由于其对参数的敏感度非常灵敏，K 值的简单变化都会影响到最终聚类的结果，抗噪声干扰能力很差。改进后的聚类算法可以解决传统的聚类算法的结果不可控的问题，算法的为稳固性更高、弹性更强。具体的做法包括粒子群寻优处理，数据特恒分析，迭代搜索获得最佳聚类中心值。具体的实现过程如图 3 所示。

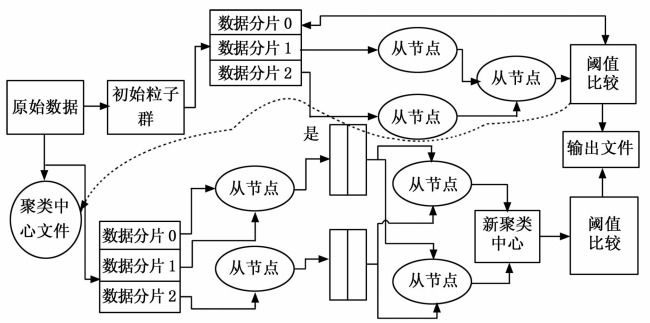

图 3 并行化的粒子群优化 K 一均值算法处理流程

3 数据挖掘－基于 Hadoop 的分布式挖掘算法

常用的挖掘关联规则的频繁项集算法包括 KNN、C4.5、Naive Bayes、CART、SVM、Kmeans、PageRank、AdaBoost、EM、Apriori 等多种，其中性能较高的包括以下 3 种：

1）Apriori 算法。主要应用与 0 对 1 类型的关联规则挖掘，其核心在于建议一个依托于两阶段数据项的递归算法。随着数据规模的扩大，该算法的瓶颈在于 I/O 的吞吐量的指数级增加降低了效率。

2）Eclat 算法。主要应用与关系型数据，其核心在于倒排二分查找思想，建立倒排表，提高频繁项集的产生速度。

3）FP－Growth 算法。其核心在于采用了频繁模式增长策略进行数据挖掘，识别频繁集。其优势在于不需要阐释候选模式，只需要进行两次数据扫描，在处理海量数据时，其性能对比前两种算法，优势非常明显。

本文使用基于 Hadoop 分布式计算框架对 FP－Growth 算法，采用并行分笔试的数据挖掘策略，挖掘数据集的关联规则，其算法的分布式数据处理过程如图 4 所示。

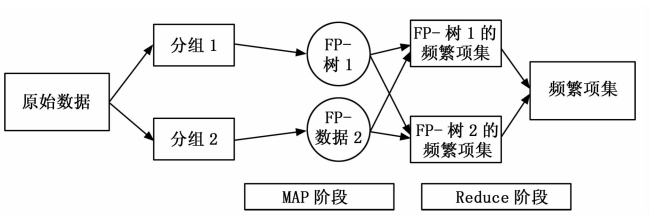

图 4 并行关联规则挖掘流程

在算法中通过上述算法获得最初的网络安全问题预测的结果。数据挖掘的第二步处理是应用基于时间维度的网络安全预测算法对初步的挖掘频繁项集进行处理，进行更精确的网络安全预测。具体的算法如表 3 所示，其执行流程如表 4 所示。

表 3 数据挖掘规则关联核心处理流程

步骤	算法过程
第 1 步	识别元数据中的频繁项集合，并标识类型编码。
第 2 步	根据第一步的频繁项集分组初始的基础数据集
第 3 步	建立各个数据分析结果的 FP 树结构。
第 4 步	为第三步的分组树建立数据挖掘频繁项结果集
第 5 步	汇总各个分组树的结果频繁项集，生成整体频繁项集

表 4 基于时间维度的 Hadoop 预测算法处理流程

步骤	算法过程
第 1 步	识别、汇总各类型攻击数组，生成异常共计库(ADL)
第 2 步	MPA 处理阶段：进行数据集的导入
第 3 步	Reduce 规约阶段：使用 BW(i)标识网络安全威胁行为及其出现的次数
第 4 步	执行异常判断 if(BW(i)＝function(ADL)) ADL＝ADL＋BW(i) Else I＋＋
第 5 步	HBW＝function(ADL)
第 6 步	MAP 处理阶段：输入数据集
第 7 步	Reduce 规约阶段：＜BW(i)，number＞
第 8 步	if(num＞HBW(number)) then 红色预警 else if (num＝HBW(number))黄色预警 esle 正常
第 9 步	生成预测量化结果

步骤一：进行初步网络安全预测初判。输入原始数据

集，进行数据异常情况统计，并拆分汇总生成异常数据库（ADL）。

步骤二和步骤三：通过 Hadoop 的 MapReduce 的映射和规约模型进行分布式计算，输入数据的同时开展异常检测。在算法中使用 BW（i）标识网络危险的类型。

步骤四：设置 BW（i）为已知风险，比对 BW（i）和异常数据库（ADL），并记录 ADL 中的异常类型 i 的出现次数。

步骤五：获取历史威胁数据集 HBW。采用方法 funcation（）在历史库中随机产生初代网络安全威胁记录。

步骤六、步骤七：对新的网络安全威胁记录进行处理。使用数字代表一定时间范围内该威胁出现的频率。

步骤八：进行判断形成结论。比对当前威胁与历史威胁库的数据量。如果大于历史数量，则定义当前网络状态属于高级别风险。如果与历史持平，定义网络安全黄色预警。如果小于历史数据，则定义为安全，同步数据网络安全预测的定性预警和量化数据。

4　实验过程及结果

4.1　实验数据集选取

实验数据集包括：美国空军局域网网络流量数据集经过语出里后的 KDD CUP 99 数据训练集和国家互联网网络安全中心提供的 CAIDA 数据集。

4.2　网络安全预测结果

首先开展的实验为依据数据集合开展网络安全主动性威胁检测率。实验数据结果如表 5 所示。

表 5　4 类入侵检测总体值

类型	Probe	Dos	U2R	R2L
比率/%	95.38	95.63	97.80	95.21

根据表中数据可以看出，基于本文的算法实现入侵检测率均高于 94%，检测平均值为 94.89%。接近 95%。为了从多角度验证本文提出的网络安全预算框架及其算法，研究过程中采用了 KDD CUP 99 数据训练集进行包括不同类型的 5 组实验，第一组实验选取包括 Dos 攻击，Probe，R2L，U2R4 入侵等 4 类异常网络数据以及正常网络流量数据包。

第二组数据选取单一的 Dos 攻击和正常的网络流量访问记录。第三组设置了包括 Probe 主动入侵病毒和正常网络访问数据集。第四组设置 R2L 病毒广播威胁和正常网络访问数据集，第五组数据集为大量的包含 U2R 病毒威胁的数据包和正常网络访问数据集。在全部 5 组数据集中的正常网络访问流量占总体测试数据集的数据比例为四分之三。

基于对实验结果的多维多、多角度分析，以验证算法的可用性和性能，主要从实验结果的误判率、网络安全预测的准确率、网络安全威胁的漏检率这 3 个角度对实验数据进行分析。其中误判率指的是对正常数据表示为威胁数据信息。检测率指的是准确识别异常数据比率，漏检率是指异常数据未被识别，将其标识为了异常威胁数据。各种异常威胁的上述 3 种情况的识别率如图 5 所示。

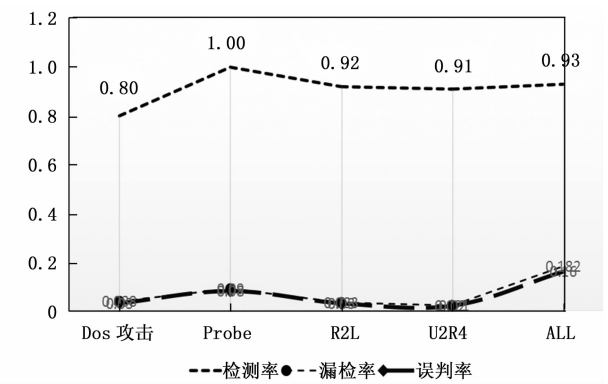

图 5　检测率比对结果图

根据图 5 中的数据可以分析，基于本文的识别算法对于网络信息中的危险预测的误判率非常低，不到 1%，检测率较高接近 94%。其中对于 R2L 的主动入侵式攻击的漏检率对比其他攻击相对较高。由于其入侵行为特征较其他类型相对特殊，算法整体性能较高。

4.3　分布式处理性能实验结果

为了测算分布式平台并行处理的算力和性能，本文在实验过程中选取了更大的数据集 CAIDA 进行实验，具体的实验结果如图 6 所示。

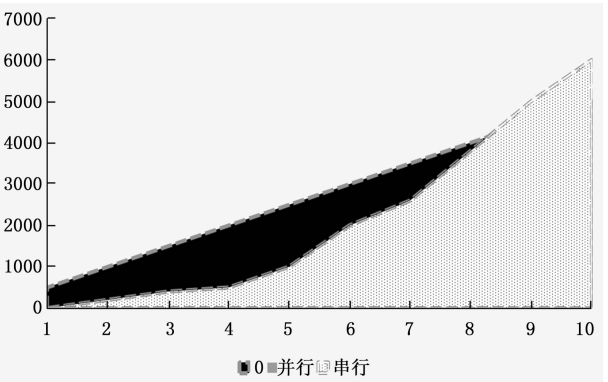

图 6　串行与并行性能对比

根据图 6 可以发现，网络安全预测算法分布式处理平台上的分布式计算时间性能上优势明显，执行性能较高。因此，在进行海量数据处理时，分布式的处理方法是不二之选。实验过程中，设置了 4 种比对性实验用以验证本文的分布式算的处理能力和有效性。第一组实验未进行降维处理只采用分布式的聚类算法进行数据处理实验，实验异常检测率很低，不到 70%。可以看出，多维数据极大地影响算法的准确率。第二组实验时在分布式聚类操作前，增加了降维和特征值的比对处理，异常检测率提升到了平均 80%，效果改善较大；第三组实验，在采用经过同样降维处理的数据集后，并未进行数据清洗和特征值训练，仅采用了 Hadoop 的 MapReduce 算法进行了结果集的关联规则分

（下转第 195 页）