Smart Grid 智能电网, 2018, 8(1), 1-7
Published Online February 2018 in Hans. http://www.hanspub.org/journal/sg
https://doi.org/10.12677/sg.2018.81001

Hans汉斯

Security Analysis and Early Warning Research of Power Big Data

Wei Meng[1], Yudong Wang[1], Jinmei Yang[2], Bin Liu[3], Mao Lin[3]

[1]Information and Communication Company, State Grid Liaoning Electric Power Co., Ltd, Shenyang Liaoning
[2]Zhaoyang Power Supply Company, State Grid Liaoning Electric Power Co., Ltd, Zhaoyang Liaoning
[3]Nanjing NARI Group Corporation, Nanjing Jiangsu
Email: mengmeng2925@sohu.com

Received: Jan. 10th, 2018; accepted: Jan. 23rd, 2018; published: Feb. 1st, 2018

Abstract

With the development of power grid technology, especially the deep integration of power grid and information technology, more and more data have been accumulated by electric power enterprises. These data are the core assets of the power enterprises, which contain a lot of value information, and are also the target of the unlawfully attacked. In this paper, a big data security early warning architecture and core algorithms are proposed based on the safety analysis and early warning research of power big data. The experiment results show that the proposed architecture and core algorithms achieve monitoring and warnings of security events, and judge the security situation intelligently.

Keywords

Big Data Security, Machine Learning, Kernel Density Estimation, Markov Logic Network

电力大数据安全分析和预警研究

孟　威[1]，王玉东[1]，杨金梅[2]，刘　斌[3]，林　茂[3]

[1]国网辽宁省电力有限公司信息通信分公司，辽宁 沈阳
[2]国网辽宁省电力有限公司朝阳供电公司，辽宁 朝阳
[3]南京南瑞集团公司，江苏 南京
Email: mengmeng2925@sohu.com

收稿日期：2018年1月10日；录用日期：2018年1月23日；发布日期：2018年2月1日

摘　要

随着电网技术的发展，特别是电网与信息化深度融合，电力企业积累了越来越多的数据。这些数据蕴含

文章引用: 孟威, 王玉东, 杨金梅, 刘斌, 林茂. 电力大数据安全分析和预警研究[J]. 智能电网, 2018, 8(1): 1-7.
DOI: 10.12677/sg.2018.81001

着大量的价值信息，是电力企业的核心资产，同时也是不法分子攻击的目标。本文基于对电力大数据的安全分析和预警研究提出了电力大数据的安全预警架构以及核心算法。实验证明该框架和算法能够及时发现安全事件并发出警告，智能洞悉电力大数据安全态势。

关键词

大数据安全，机器学习，核密度估计，马尔科夫逻辑网

1. 引言

智能电网是未来电网的发展方向[1] [2]，而信息化是实现智能电网发展的基础和保障，在发电、输电、变电、配电、用电和调度等各个环节，以及电力企业人、财、物集约化管理方面，都离不开信息化的支持[3] [4]。国家电网公司“SG186 工程”和“SG-ERP 工程”的建设和实施实现了国家电网公司总部与各个网省公司之间的横向集成、纵向贯通、数据共享和业务融合，有效促进了电网的迅速发展。

随着电网的不断发展壮大，特别是结构化、非结构化、海量历史/准实时、电网空间等 4 类数据中心平台的深入建设，国家电网公司已积累了丰富的数据资源，形成了电力大数据[5]。电力数据按照数据来源可分为企业内部数据和外部数据，其中内部数据大多来源于关键应用系统，例如客户服务系统、营销应用系统、数据采集与监控系统、协同办公系统、配电管理系统、干部管理系统、生产管理系统等；外部数据则一般来自于互联网、地理信息系统、气象信息系统等。目前为止，国家电网公司在生产、经营、管理、服务等方面积累了海量的各类数据资源，数据总量达到约 5 PB 级别，若用普通 DVD 光盘存储，叠加高度超过 1500 米，其中包括海量的报表、图片、音视频、日志、地理空间数据、时标量测数据等非结构数据[6]。

在一定时间范围内，无法采用常规软件工具捕捉、管理和处理的数据集合，称为大数据(Big Data)。一般情况下，大数据是海量、高增长率和多样化的信息资产，需要新的处理模式才具有更强的决策力和流程优化能力[7]。电网在生产、运行、管理过程中产生庞大的数据信息，数据类型多且增长速度快，需要新的处理模式，符合大数据的特征。这些电力大数据为国家电网公司的决策分析和提高发配电效率提供了良好条件和基础，例如：在电网运行方面，通过精准的负荷预测，推动源网荷协调互动，积极消纳新能源，实现经济调度；在电网规划方面，建立电网规划数据挖掘和诊断评估模型，根据对公司内、外部海量基础数据的挖掘分析，实现电网规划更加精准，提升投资有效性[4] [6]。

大数据在给我们的工作和生活带来便利与好处的同时，其安全问题也逐渐暴露，数据泄露、数据贩卖等事件频发，客户隐私和安全受到极大挑战。在数据驱动环境下，网络攻击也更多地转向存储重要敏感信息的信息化系统，大数据安全防护俨然成为大数据应用发展的一项重要课题[8] [9]。电力大数据存储着电网重要数据和大量用户隐私信息，若被不法分子窃取，易影响电网正常运行，危害用户声誉、财产，甚至造成电网灾难性破坏。如何有效保障电力大数据的安全是亟待解决的课题之一。本文首先分析电力大数据面临的安全威胁，然后重点讨论电力大数据安全预警框架和实现，最后总结全文并展望未来。

2. 电力大数据安全风险

电力大数据的安全风险存在于大数据产生、传输、处理、存储、应用等整个运行周期，例如：大数

据传输中易出现中断、窃听、伪造、篡改等风险，大数据处理及应用中存在用户越权、主机故障、措施不当等风险[10]。电力大数据 80%以上是非结构化数据，通常采用 NoSQL 数据库的形式存储，其它数据采用关系型数据库或者文件服务器存储。在本文，重点讨论电力大数据存储自身存在的安全风险。

2.1. 结构化数据存储安全风险

关系型数据库是结构化数据的主要存储形式。数据库系统漏洞、安全策略不当(例如安全配置、数据库自身的安全防护机制等配置不当)均易引起数据库安全风险。此外，数据库自身的审计系统能记录下数据库某些特定用户的行为信息,但存在以下两个问题：一是无法及时定位不安全行为，即通过数据库的审计系统不能及时识别并定位用户的不安全行为；二是不能对攻击行为实时告警，即无法通过数据库审计系统对攻击行为实时发出警告。此外，数据库开启审计功能后，产生大量审计日志，占用大量的磁盘空间，降低数据库服务性能，更难以在大量审计日志中查找有价值的信息[11]。

2.2. 非结构化数据存储安全风险

NoSQL 是非关系型数据库，可以存储不同类型的数据，包括结构化数据和非结构化数据，数据的多样性特点导致标准 SQL 语句无法访问非结构化数据存储[12]。NoSQL 数据库的每个数据镜像存储于不同位置，保障了数据可用性和无丢失，具有扩展性强、可用性高和灵活性等优点。NoSQL 目前不能沿用 SQL 模式，缺乏严格访问控制和隐私管理技术；作为新代码，难免存在各种漏洞；此外，NoSQL 服务器软件内置安全不够，客户端应用程序需要身份验证、授权管理等安全措施[13] [14]。

2.3. 访问控制安全风险

数据库访问控制的目的是为了防止越权操作，使得只有授权用户在规定范围内使用数据库的某些允许其访问的部分数据[12]。目前，数据库访问控制主要分为三种，分别是自主访问控制、强制访问控制和基于角色的访问控制。

1) 数据库自主访问控制是最简单最灵活的一种访问控制方式，数据合法拥有者可将数据使用权限自主授予其它用户。这种访问控制方式容易受到非法人员攻击，权限管理工作也困难，特别是用户和数据库数据量庞大时，系统开销将剧增[11]。

2) 数据库用户和数据划分不同的安全权限级别,仅有对应安全权限级别的用户才能存取其对应级别的数据信息，这种访问控制方式称为强制访问控制。该方式最大的优点是注重数据库的保密性，缺点是无法应对用户恶意泄露数据，同时授权管理更困难。

3) 基于角色访问控制的核心思想是权限被分配到不同的角色当中，只有拥有某一角色的用户才能访问角色对应范围内的数据[11]。这种访问控制方式最大的安全隐患是存在用户权限过大问题。

此外，电力大数据本身是由海量的各种数据资源构成，恶意软件和病毒代码有可能隐藏其中不易发现，使电力大数据成为可持续攻击的载体。

3. 电力大数据安全预警框架

任何网络攻击都会留下痕迹。针对电力大数据攻击行为的痕迹，往往是以数据形式隐藏于电力大数据中[15]。在本文中，我们从电力大数据的存储方面开展安全分析和预警研究，有效应对大数据的安全威胁。传统的安全分析方法是基于规则和特征的，而这些规则和特征来源于已经被识别的攻击；如果某种攻击是原来未曾识别和未被描述成规则的，规则库和特征库中不存在这种规则和特征，那么传统的安全分析方法将无法识别这种攻击。因此，传统安全分析方法仅对已知攻击和威胁有效，并存在趋势预测难、早期预警能力差等问题。为了更加主动应对新型和未知的威胁和风险，本文借助于机器学习相关算法，

智能化洞悉电力大数据的安全态势，挖掘各类安全事件并及时发出告警。

3.1. 安全预警应用场景

非法入侵或违反安全规则的用户往往会进行大量的数据增、删、查、改及抽取行为。实时监控用户访问的数据文件、数据量大小，并通过统计图(包含柱状图、饼状图、趋势线图等)进行实时展示。根据对正常合法用户的历史信息，通过机器学习统计得到一个安全的阈值范围，针对超出安全阈值范围的用户及其行为进行危险告警，并通过图表进行实时展示。此外，依据用户实时信息学习分析，结合安全阈值范围，实现下一风险预测并及时告警。

3.2. 安全预警异常查询规则

1) 查询量波动阈值

统计最近三年的数据日均查询量，根据核密度估计算法来确定数据年日查询量最大值，将超出阈值的查询量视为异常。

2) 跨地域异常查询

根据访问用户的 IP 地址，判断 IP 访问地址是否在许可权限范围内。如果访问 IP 地址不在许可权限范围内，标记为异常。

3) 休眠用户异常查询

休眠用户是指最近一年内未发生过数据查询行为的用户。休眠用户一旦启动了查询操作称为休眠用户异常查询。

3.3. 安全预警架构设计

电力大数据安全预警架构包括：数据采集-数据接入-流式计算-入侵安全监测检测-数据可视化等五大部分，整个架构如下图 1 所示。其中，数据采集负责从各节点上实时采集数据库日志数据；kafka 是分布式消息系统，用于缓冲和平滑不同步的数据；流式计算对采集到的数据进行实时流式计算，选用 storm；入侵安全监测检测根据数据采集处理和机器自学习两大部分编写监测模型；数据可视化将实时监测的用户行为图表显示并对攻击行为实时告警。

3.4. 机器学习训练

电力大数据安全分析和预警根据用户数据库使用行为习惯来定义行为模式或用户访问能力，无需预先设置固定临界值便可实现用户异常行为的智能化检测。通过机器学习算法来定义用户行为，若用户实时行为模式与其历史行为模式存在较大差异时，则认为用户行为异常。这里采用两种算法检测异常，首先通过核密度估计算法计算用户行为的正常阈值并进行异常监测，其次通过马尔科夫算法进行下一步的预测计算。下图 2 描述了目前分析计算中用户行为的机器学习训练检测框架。在图 2 中，预处理是将采集到的数据转化为相同尺度，保证数据每个特征均值的标准化；白库和黑库是用正常数据和异常数据训练机器学习网络，目标是为了识别用户正常和异常行为。

4. 电力大数据安全预警实现

在实现时，电力大数据安全分析和预警主要包含 4 个步骤：数据准备、数据分析、模型建立、模型验证。

4.1. 数据准备

数据样本选用服务器数据库访问系统中最新的审计日志，包含用户 ID、操作时间、操作行为、操作

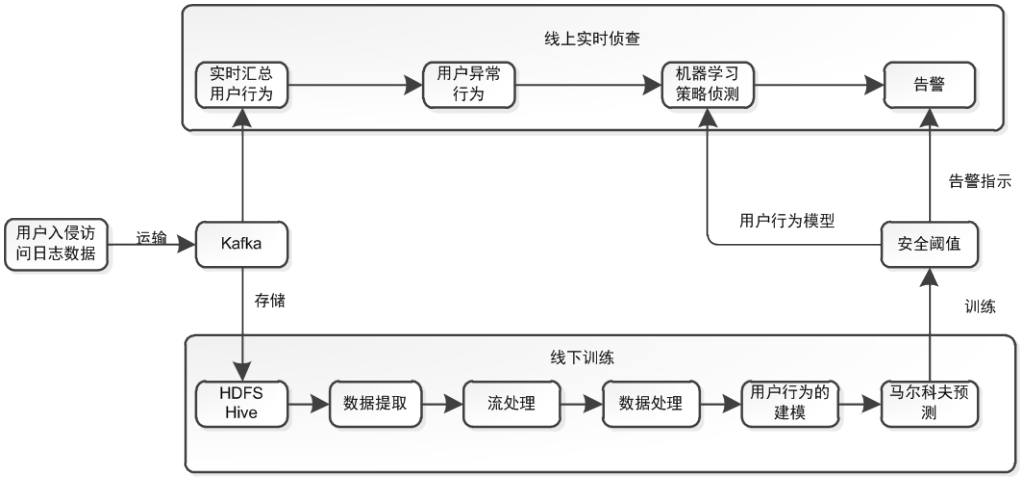

Figure 1. Power big data security analysis and early warning architecture
图 1. 电力大数据安全分析和预警架构

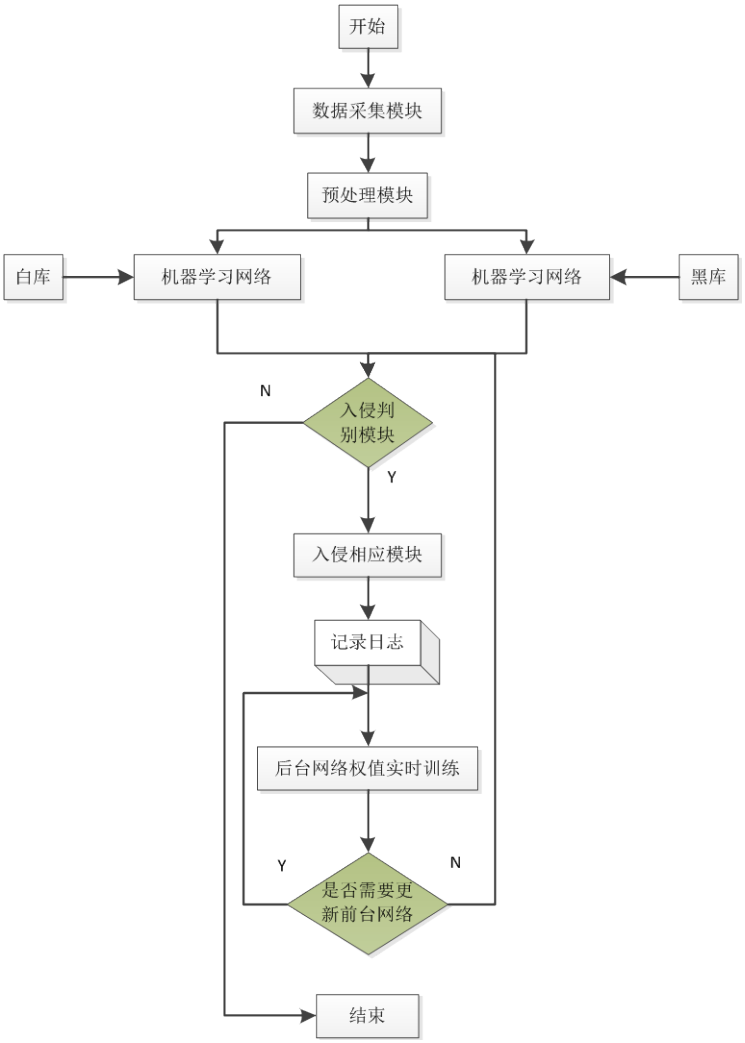

Figure 2. Machine learning monitoring framework
图 2. 机器学习监测架构

终端、返回码、操作对象等，形成一个较大的数据量条数。首先数据异常值剔除，重复值处理，缺失值处理；其次经过多维数据的探索研究，降解数据维度，进行数据归一化处理；最后形成查询数据量矩阵。

4.2. 数据分析

查询量矩阵按照月份来统计。若月查询量矩阵中存储很多为零的值，说明仅有少量用户联系每个月有访问数据库。大量用户的数据库查询操作是断断续续的，仅有少量用户每个月都访问数据库。在这里，我们有必要进一步分析用户查询的连续性。用户查询的连续性可以用两个指标来刻画，一是查询休眠时间，二是查询休眠重启。用户休眠是指用户在自然月份内无查询行为，否则称为用户活跃。用户查询休眠时间是指当前时间减去用户最后一次查询时间得到的天数。用户查询重启是指用户前一个月无数据查询行为，本月有数据查询行为，出于活跃状态。

4.3. 模型建立

本次安全预测模型需要对每个用户构建一个监控与预测模型。用户所属机构不同，查询需求和数据操作行为不同，同一机构用户往往保持相同或相似趋势。

1) 核密度估算法

核密度估计(Density Estimation)是由 Rosenblatt 和 Parzen 提出了非参数估计方法，它从数据样本出发研究数据分布特征，对于数据分布的假定条件和先验知识无任何要求[16]。首先，根据训练样本数据，计算每个用户的概率密度分布函数；然后使用核密度估算法求出访问查询数据库的正态分布，并得到最大值，进而得到其阈值。其核心公式如下：

$$f_h(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x_i-x),x\in R \tag{1}$$

其中，x_1，x_2，…，x_n 是独立同分布的 n 个样本点，其概率密度函数为 f，$K(.)$ 是非负、均值为 0 且积分为 1 的核函数，$h>1$ 是平滑参数，也称为带宽值。

2) 马尔科夫预测算法

为有效解决入侵用户访问数据库过程的复杂性和不确定性，引入马尔科夫逻辑网。马尔科夫逻辑网是一种将传统基于概率图模型的马尔科夫网与不确定推理中的一阶逻辑理论相结合的统计关系学习方法[17]。马尔科夫逻辑网把领域知识带到马尔科夫网中，使一阶逻辑此类简明语言完成描述问题，进而强化了处理复杂性和模糊性问题的功效。通过用户查询访问的实时数据量，在机器学习模块采用马尔科夫预测算法，对应的预测出下一阶段的数据查询量，判断是否超出阈值。

4.4. 模型验证

本文提出的电力大数据安全分析和预警架构框架中，为了验证该模型的有效性，采用平均相对误差和泰勒不等系数来刻画。平均相对误差计算式为：

$$\sum_{i=1}^{n}\left|\frac{\hat{x}_i-x_i}{x_i}\right|\Big/n, \tag{2}$$

其中，n 为预测期数，$\hat{x}_i$ 为预测值，x_i 为实际值。泰勒不等系数计算式为[18]：

$$\frac{\sqrt{\sum_{i=1}^{n}(\hat{x}_i-x_i)^2\Big/n}}{\sqrt{\sum_{i=1}^{n}\hat{x}_i^{\,2}\Big/n}+\sqrt{\sum_{i=1}^{n}x_i^{\,2}\Big/n}}. \tag{3}$$

泰勒不等系数的值应在 0 和 1 之间，当泰勒不等系数等于 0 时，是最优拟合[18]。

本文是基于大数据来研究电力系统安全预警，需要大量的电力信息系统数据，而由于电力系统的特殊性，对数据的安全性要求非常高。本系统的测试数据均来自实际的电力信息系统，担心信息泄密，在相关管理要求下，无法直接给出真实完整的测试数据。利用该模型分析监测营销业务应用系统的测试系统，发现某日某用户数据访问量明显高于其日均预测值，访问异常。经核实，该用户操作确系违规。

5. 总结

大数据在创造价值的同时，其安全问题不容忽视。电网的发展壮大积累了海量的电力大数据，这些数据中蕴含着大量有价值的信息，同时成为不法分子攻击的目标。本文针对电力大数据开展安全分析和预警研究，提出了电力大数据安全预警架构和核心算法，经测试验证该框架可及时发现安全事件并发出警告，智能化洞悉电力大数据的安全态势。在下一步的工作中，我们将继续完善该架构，将数据挖掘、人工智能、深度学习等新技术融合创新并应用到实际生产中。

参考文献 (References)

[1] 黄滨, 安郁滨. 试论中国智能电网的发展[J]. 中外能源, 2014, 10(2): 21-24.

[2] 朱建. 智能电网发展的对策研究[J], 科技风, 2010, 11(1): 104-106.

[3] http://m.book118.com/html/2015/0314/13301098.shtm

[4] 王颖. 信息网络在智能电网中的探索与实践分析[J]. 硅谷, 2014(24): 20-24.

[5] 王继业, 季知祥, 史梦洁, 等. 智能配用电大数据需求分析与应用研究[J]. 中国电机工程学报, 2015, 4(6): 1829-1836.

[6] 黄颖, 张振兴. 国家电网的大数据“云图” [J], 中国电力报, 2016, 11(4): 23-26.

[7] 李昊, 张敏, 冯登国, 惠榛. 大数据访问控制研究[J]. 计算机学报, 2017, 1(2): 72-91.

[8] 大数据安全岌岌可危, 有效防护成当务之急[EB/OL]. http://info.secu.hc360.com/2016/12/080908876891.shtml, 2016-12-08.

[9] 用户隐私面临大考, 大数据安全防护痛点犹存[EB/OL]. http://www.tpy888.cn/news/201612/06/91642.html, 2016-12-06.

[10] 王丹, 赵文兵, 丁治明. 大数据安全保障关键技术分析综述[J]. 北京工业大学学报, 2017, 3(1): 335-349.

[11] 康红丹. 数据库网络服务行为分析与识别技术研究[D]: [硕士学位论文]. 北京: 北京大学, 2012.

[12] 王晶. 非结构化数据结构化存储中的查询语句重写技术研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2013.

[13] 赵宝献, 秦小麟. 数据访问控制研究综述[J]. 计算机科学, 2005, 2(1): 6-11.

[14] 桑运昌. 大数据的安全现状与应对策略研究[J]. 计算机科学, 2015, 11: 372-373.

[15] 曾中良. 大数据时代的企业信息安全保障[J]. 网络安全技术与应用, 2014(8): 137-138.

[16] 方斯顿, 程浩忠, 徐国栋, 等. 基于非参数核密度估计的扩展准蒙特卡洛随机潮流方法[J]. 电力系统自动化, 2015(7): 21-27.

[17] 陈宇, 王亚弟, 王晋东, 等. 马尔科夫逻辑网在信息安全风险管理中的应用[J]. 计算机工程与应用, 2016, 10(1): 104-107.

[18] 姚前, 谢华美, 景志刚, 等. 基于数据挖掘的个人征信系统异常查询实时监测模型及其应用[J]. 大数据, 2016(6): 83-92.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8763，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sg@hanspub.org