计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(3):173−178 [doi: 10.15888/j.cnki.csa.006267]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

电网工控系统流量异常检测的应用与算法改进①

刘亚丽[1,2], 孟令愚[3], 丁云峰[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](国家电网公司东北分部, 沈阳 110180)
通讯作者: 刘亚丽, E-mail: 2352473711@qq.com

摘　要: “两化融合”的工业控制网络的安全问题不断突显. 电力作为国家重要基础设施, 其电网工控系统的安全防护工作极其重要. 本文根据电网工控系统中控制网的内防水平低且其安全监测和防护缺乏内部网络流量异常检测的现状, 分析了电网工控系统的组成结构、网络安全需求及面临的威胁. 提出了将流量异常检测技术应用于针对电网工控系统控制网的安全防护中, 形成针对电网工控系统控制网的两级安全防护. 然后研究了流量异常检测方法的分类和特点以及电网工控系统的网络流量数据特点, 提出了基于熵的动态半监督 K-means 算法并辅以单类支持向量机对半监督 K-means 算法进行改进, 为提升电力系统内防水平奠定基础.

关键词: 电网工控系统; 安全防护; 流量异常检测; 动态半监督 K-means; OCSVM

引用格式: 刘亚丽,孟令愚,丁云峰.电网工控系统流量异常检测的应用与算法改进.计算机系统应用,2018,27(3):173–178. http://www.c-s-a.org.cn/1003-3254/6267.html

Application and Algorithm Improvement of Abnormal Traffic Detection in Smart Grid Industrial Control System

LIU Ya-Li[1,2], MENG Ling-Yu[3], DING Yun-Feng[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](State Grid Corporation Northeast Branch, Shenyang 110180, China)

Abstract: The safety of industrial control network is becoming more prominent. Electric power is an important national infrastructure, so the safety protection of smart grid industrial control system is extremely important. In smart grid industrial control system, according to the status quo of the low internal protection level of the control network and the lack of internal network of anomaly traffic detection, this paper analyzes the composition of the industrial control system, the network security demand, and the threats faced by the smart grid industrial control system. It proposes to apply traffic anomaly detection technology to the security protection of smart grid industrial control system, which forms the two-level security protection. Then, the classification and characteristics of traffic anomaly detection methods and the characteristics of network traffic of smart grid industrial control system are studied. And it proposes a dynamic semi-supervised K-means algorithm based on entropy and OCSVM to improve the semi-supervised K-means algorithm for improving the internal protection level of the smart grid industrial control system.

Key words: smart grid industrial control system; security protection; traffic anomaly detection; dynamic semi-supervised K-means; OCSVM

① 基金项目: 国科控股企业技术创新引导基金 (2015XS0356)
收稿时间: 2017-06-28; 修改时间: 2017-07-17; 采用时间: 2017-07-24; csa 在线出版时间: 2018-02-09

引言

随着“两化融合”的推进, 工控网络已从原有的控制网扩展到了管理网, 这使得工控系统与互联网的关系越来越紧密, 工业控制网络不断智能化, 全球互联互通的大趋势已势不可挡. 然而工业控制网络不断开放的同时也给工控系统带来了越来越多的安全威胁, 电网工控系统也是如此. 著名的攻击事例有 2010 年的“震网”病毒对伊朗核电站工控系统的网络攻击[1].

工控系统涉及很多重要国家基础设施, 比如电力, 所以国家对工控系统安全很重视. 而工控系统安全与传统网络安全又有较多差异, 相当于一个较新的领域. 流量异常检测技术虽已经广泛地应用于互联网信息安全防护中, 但是在工业控制系统 (Industrial Control System, ICS) 中的研究较少[2], 在电网工控系统中研究较之又少.

就电网工控系统来说, 目前企业电力系统内部遭受网络攻击的问题仍无法得到有效的监测、预防与解决[3]. 为了确保电力系统安全运行工作, 重视电力系统内防水平低于外防水平这一问题已成为共识[3]. 对于电力系统内部, 现有的电力监控系统主站与厂站的安全监视主要依赖于内网监视平台, 该平台主要是针对边界处的安全设备和网络设备的日志进行收集, 但没有针对电网工控系统内部网络流量进行异常检测[4].

通过分析网络流量可直观了解网络中的行为, 从而帮助我们监测网络的安全状态[5], 及时发现网络攻击、异常流量、异常行为. 所以将流量异常检测技术应用于电网工控系统的安全防护将利于防范电网安全威胁, 可以在确保电力系统稳定运行的基础上, 提高电力系统自动化、智能化水平, 为社会创造更多经济效益.

本文针对电网工控系统中控制网的内防水平低且其安全监测和防护缺乏内部网络流量异常检测的现状, 提出将流量异常检测技术应用于针对电网工控系统控制网的安全防护中, 并提出一种基于熵的动态半监督 K-means 算法并辅以单类支持向量机 (OCSVM) 对半监督 K-means 算法进行改进, 以提高流量异常检测的效果, 为监测分析网络流量及时感知网络异常发现攻击, 从而提升电力系统内防水平奠定基础.

1 电网工控系统

在“两化融合”的背景下, 电网工控系统已广泛应用信息网络技术且其网络化和智能化也越来越高, 使得电力传输与控制可以自动可靠地进行.

1.1 电网工控系统的组成与结构

通俗来说, 电网工控系统主要指: 电力行业中, 用于监视和控制的, 与计算机和网络技术有关的系统和智能设备, 以及通信与数据网络. 具体来说, 电网工控系统由控制网络和管理网络组成, 管理网络指调度监控管理网, 控制网络分为站控层、间隔层和过程层, 部署在智能变电站[4,6]. 图 1 所示是电网工控系统的结构.

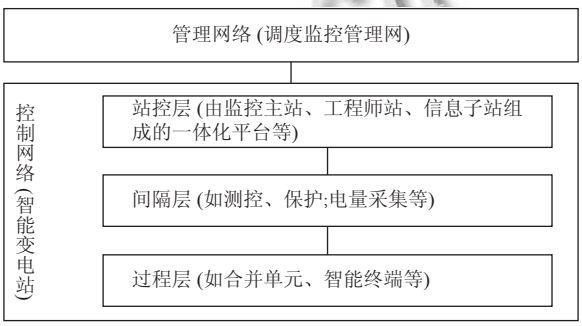

图 1 电网工控系统结构

电网工控系统安全, 总体来说涉及两方面, 分别是电网基础设施安全和电网工控系统网络安全[7]. 随着网络攻击、安全威胁的飞速增长, 针对工控系统的网络攻击和入侵也逐渐变多, 所以目前网络安全已成为电网工控系统安全防护的重中之重.

虽然传统网络安全防护中的流量异常检测技术已相对成熟, 但是由于传统网络与电网工控系统存在许多不同之处, 不能直接照搬使用, 所以需要研究适用于电网工控系统的流量异常检测方法.

1.2 电网工控系统的安全需求和面临的威胁

电网工控系统的安全需求不同于传统网络的安全需求之处[8], 主要是以下的方面:

(1) 电网工控系统侧重可用性, 传统网络侧重机密性, 电网工控系统实时性要求高于传统网络.

(2) 电网工控系统风险管理要求主要是防止监管失控, 而传统网络主要确保数据保密.

(3) 资产保护需求上, 电网工控系统的首要保护目标是控制器及现场设备 (如电力一次设备), 而传统网络首要保护信息资产.

电网工控系统面临多种威胁与攻击. 工控协议在设计之初没有考虑安全性, 所以非常容易遭受网络攻击; “两化融合”的电网工控系统既连接企业内部计算机网络又连接互联网、接入的现场设备多、网络通信

方式多导致入侵途径多, 如现场总线通讯接口、工业以太网通讯接口、U 盘等移动介质、以及误操作等都是可能的入侵途径[9]. 总之“两化融合”的电网工控系统面临的工控攻击的趋势是综合利用系统、网络、终端及管理等多个层面的漏洞进行有针对性的攻击, 因此需要多种技术和多层次的安全防护.

2 电网工控系统两级安全防护

当前我国的电力系统安全措施主要是使用防火墙进行访问控制, 安全防护方法单一. 根据电网工控系统中控制网的安全防护现状, 通过对电网工控系统安全需求和面临的威胁攻击的研究分析, 提出将流量异常检测技术引入电网工控系统控制网的安全防护中, 使访问控制与流量异常检测对电网工控系统的控制网形成两级安全防护, 提高电网工控系统控制网的安全防护水平. 应用流量异常检测技术后形成的两级安全防护的逻辑结构, 具体如图 2 所示.

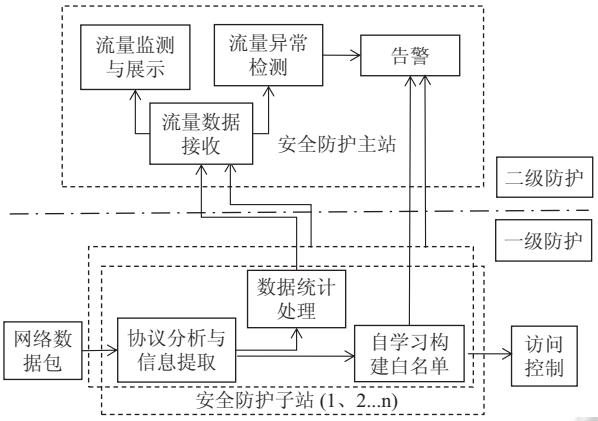

图 2 电网工控系统两级安全防护

安全防护子站主要负责访问控制. 具体是通过协议分析与信息提取阶段对数据采集设备采集来的正常网络数据包进行数据包解析, 特征提取来自动构建白名单, 记录允许通过的源 IP, 目的 IP, 目的端口号[10]. 从而对现场智能终端等设备进行访问控制, 实现一级安全防护. 安全防护子站可部署在智能变电站的站控层(或间隔层).

安全防护主站接收来自于安全防护子站的基于时间窗统计处理获得的流量特征属性, 然后进行流量异常检测, 实现二级安全防护. 安全防护主站可包括流量数据接收、流量监测与展示、流量异常检测和告警等模块. 安全防护主站可以部署在调度监控层 (或智能变电站的站控层).

安全防护子站除了访问控制, 还要为安全防护主站的流量异常检测做准备工作. 流量异常检测技术的引入使得形成了对电网工控系统控制网络的两级安全防护, 下面将详细介绍安全防护中的关键技术流量异常检测及其算法的改进.

3 流量异常检测技术及其算法的改进

对于攻击入口, 最常见的方式还是通过网络, 并且各类攻击通常会引起网络流量异常[11], 所以考虑安全防护时应着重从网络流量入手, 进行网络流量的监测和异常检测, 即电网工控系统安全防护的关键技术是流量异常检测技术.

通过采集网络数据包, 分析得到电网工控系统的网络流量数据的特点[2]如下:

(1) 数据长度小, 频率相对较高.

(2) 网络流量呈周期性.

(3) 响应时间相对小很多.

(4) 数据流向固定.

(5) 时序性强, 如控制信息发生顺序固定.

(6) 工控网络主要以广播多播形式通信, 所以多用 UDP 协议; 而普通 IT 网络常用点对点传输.

总之, 电网工控系统的网络流量与普通 IT 网的网络流量差异较大, 所以有必要针对电网工控系统, 研究网络流量异常检测技术.

3.1 流量异常检测技术的分类

根据相关研究, 将流量异常检测技术常用的方法主要分为统计分析方法和机器学习方法. 统计分析方法是通过选择一些特征属性的统计度量来刻画系统或网络行为状态的边界, 根据测量的统计值是否偏离该边界一定范围来判断是否发生异常[12]. 如基于工控网络中流量周期性的异常检测.

机器学习方法中包括基于数据挖掘的方法、基于生物学原理的方法等[2]. 基于数据挖掘的方法包括使用聚类分析、关联分析、分类等算法的异常检测方法; 基于生物学原理的方法包括基于神经网络、遗传算法或免疫学原理等的异常检测方法, 总之这些方法都属于机器学习方法.

统计分析方法不需要太多的先验知识, 但对事件发生的顺序不敏感、漏报和误报率较高[11], 还需要大量纯净的训练数据, 而在真实的网络环境中这样的数

据很难确保. 所以选择基于机器学习的流量异常检测方法更合适.

3.2 基于机器学习的流量异常检测方法

根据训练数据集有无类标记, 将机器学习方法分为有监督学习和无监督学习; 而训练数据集同时包含有类标记和无类标记的实例时, 则称为半监督学习[13].

有监督学习算法包括决策树、贝叶斯、神经网络、支持向量机等[13]; 无监督学习算法包括 K-means 等聚类分析算法.

有监督学习识别率高, 但需要大量的标记样本代价较大且无法识别未知类别, 无监督可识别未知类别, 但准确率低[14]. 半监督学习用少量的标记样本来辅助大量的无标记样本进行异常检测, 既减少标记工作又提高检测的准确率.

由于工控网络对可用性要求高, 为了避免没能检测出异常而使工控网络不可用, 检测方法的准确性要高. 工控网络的攻击因为多为新型的和被研究时间较短的网络攻击, 规则库覆盖范围较小, 所以需要检测方法的漏检率低, 对未知攻击的检测能力要高[11]. 由于工控网络的实时性要求, 需要检测方法的快速与计算效率高. 所以相对于有监督学习, 无监督学习和半监督学习更合适. 其中的 K-means 算法快速、计算量小、聚类效果好[11]. 与传统的 K-means 算法相比, 半监督 K-means 算法检测率相对较高, 但两者对未知攻击的检测能力低. 为了进一步提高未知攻击检测能力, 降低漏检率, 提高检测率, 提出了基于熵的动态半监督 K-means 算法并辅以单类支持向量机对半监督 K-means 算法进行改进.

3.3 流量异常检测模型

流量异常检测模型如图 3 所示. 网络流量原始数据经过预处理得到训练数据, 对训练数据使用基于熵的动态半监督 K-means 并辅以单类支持向量机的改进半监督 K-means 算法进行训练, 得到检测模型. 经过预处理的实时数据输入检测模型进行实时检测, 获得检测结果, 若发现异常则进行异常告警.

3.3.1 预处理

数据流量上的特点, 使得正常流量和异常流量的各种属性在分布特征上有很大的差异[15], 通过使用特征属性的熵来处理电网工控系统的流量, 将利于进行流量异常检测[4]. 预处理阶段主要进行基于时间窗的特征属性熵量化. 按时间顺序, 每 n 个数据包为一个时间窗口并作为一个单位流量. 对单位流量计算其每个特征属性的熵. 下面具体说明特征属性熵量化. 如单位流量 $F(p_1(T_1, T_2), p_2(T_1, T_2),\cdots, p_n(T_1, T_2))$, 包含 n 个数据包 p, 每个包有 2 个特征属性 $T_i(i=1, 2)$. 属性 T_i 有 m_i 种可能取值, 单位流量中属性 T_i 取值为 j 的数据包数为 n_j, 则 T_i 的熵为:

$$H(T_i) = \sum_{j=1}^{m_i} (\frac{n_j}{n}) \ln(\frac{n_j}{n}) \tag{1}$$

最终获得形如 $F(H(T_1), H(T_2))$ 的单位流量的熵量化特征向量.

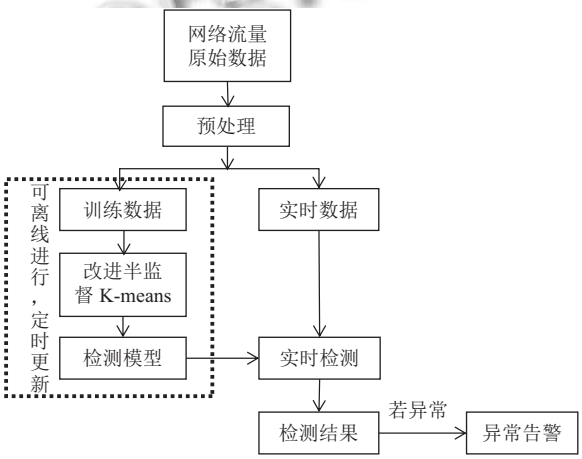

图 3 流量异常检测模型

3.3.2 训练检测模型

传统 K-means 聚类算法是无监督学习, 是 1967 年由 James MacQueen 首次提出的, 其原理是: n 个对象中初始随机选定 k 个对象作为簇中心, 根据欧氏距离, 将 n 个对象就近划分到 k 个簇中, 然后更新簇中心, 即求簇中对象的均值作为新的簇中心, 再重新就近划分 n 个对象到 k 个簇. 重复更新簇中心和就近划分 n 个对象的过程, 直到划分质量最高, 即簇内高相似, 簇间低相似.

Basu[16]等提出的半监督 K-means 算法的基本思想是用覆盖所有类别的少量标记样本来确定总的聚类簇个数 k 和初始簇中心. 其他流程与传统 K-means 聚类算法相同.

SVM 算法是一种建立在统计学习理论基础上的机器学习算法[13]. 其基本思想是用一种非线性映射, 把原训练数据映射到较高的维上, 在新的维上搜索最佳分离超平面[13]. SVM 算法对线性和非线性的数据都可以进行分类. SVM 算法又分为多种, 其中单类支持向量机 (OCSVM) 只需要利用电网工控系统的正常网络

流量数据就可建立正常网络流量的检测模型[17]. 已知只包含正常数据的训练集$D_+ = \{(x_i, y_j)|i \in |D_+|\}$, 其中$y_j$是类别标记, 且$y_j \in \{1, -1\}$, 1 表示正常, –1 表示异常. 在特征空间上, 构造最优分类超平面, 其中分类超平面表达成:

$$(\omega * x) + b = 0 \tag{2}$$

然后对无类标的样本x_l根据判别规则判断是正常还是异常数据:

$$(\omega * x_l) + b \geq 1 \ \rightarrow y_j = 1 \tag{3}$$

$$(\omega * x_l) + b < 1 \ \rightarrow y_j = -1 \tag{4}$$

实际环境中想要使得采集到的标记样本覆盖所有的异常类型难以实现, 且该半监督 K-means 算法对未知攻击的检测能力低. 电网工控系统的正常网络流量数据具有周期性和较小长度性等特性, 且实际环境中较好获取, 适合使用单类支持向量机, 但是若只利用 OCSVM, 不利用已有的异常类型信息, 不利于建立高效的流量异常检测系统. 所以根据以上情况, 将单类支持向量机 (OCSVM) 引入半监督 K-means 聚类算法中以改进异常检测效果, 即提出基于熵的动态优化 k 值的半监督 K-means 辅以单类支持向量机, 简称基于熵的动态半监督 K-means+OCSVM 算法, 其具体内容如下:

1) 从少量已标记样本中, 根据其 k 种类标记, 确定 k 的初始值, 并将每类的样本求均值作为该类的初始簇中心, 即初始共确定 k 个簇中心.

2) 进行一次就近分配, 计算聚类质量 E_1. 其中聚类质量用所有对象到其簇中心的误差的平方和来评估, 即:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dis(p, C_i)^2 \tag{5}$$

且 E 越小, 表示聚类质量越高.

3) 令 k=k+1.

4) 从未标记样本中选取一个点 a 作为新簇的簇中心, 注意 a 要使得那些离其当前簇中心较远的点 b, 划分到以 a 为中心点的簇后, 到中心点的距离缩小的和最大, 即, 使得公式 (5) 取值最大.

$$dsum = \sum_{b=1}^{n} (max\{0, |b - C_b|^2 - |b - a|^2\}) \tag{6}$$

5) 再进行一次就近分配, 计算聚类质量 E_2.

6) 若 k 次的聚类质量高于 k–1 次的聚类质量, 则转至步骤 3); 否则 k=k–1 转至步骤 7).

7) 检查是否有只包含新增簇中心点的孤立簇, 若有则去从簇中心点集和未标记样本集中掉该点, 使 k=k–1 并重复步骤 2)~6);

若无, 则用步骤 6) 确定的新 k 值和 k 个初始簇中心点作为初始输入, 不断进行重分配, 求簇中对象的均值更新簇中心, 直到簇中心收敛, 得到总体样本集的 k 个划分.

8) 用步骤 7) 得到的 k 个最终簇中心点对少量有标记样本集进行一次划分, 找到包含少量有标记样本集中正常样本的簇的簇中心点, 以及不包含任何少量有标记样本集中数据的未知攻击的簇中心.

9) 使用 OCSVM 得到的判别规则, 根据 8) 找到的簇中心点, 找 7) 得到的 k 个划分中的对应簇进行簇内划分修正, 从而区分开距离上较近的正常和异常数据.

3.3.3 实时检测

流量异常检测模型中实时检测部分的具体流程为: 将熵量化特征向量描述的网络流量数据输入动态半监督 K-means 算法得到检测模型, 检测模型根据输入点距离正常簇中心和各异常簇中心的距离来进行划分之后再辅以 OCSVM 的正常数据判别规则修正对网络流量数据的异常检测判别, 从而实现实时检测网络流量的异常与否.

4 测试验证

测试检测算法时, 主要用到三个关键指标来评估, 分别是: 检测率、误报率和未知攻击检测率. 其定义[11]为:

$$\text{检测率} = \frac{\text{识别出的异常样本数}}{\text{所有异常样本数}} \times 100\%$$

$$\text{误报率} = \frac{\text{误判为异常的正常样本数}}{\text{所有正常样本数}} \times 100\%$$

$$\text{未知攻击检测率} = \frac{\text{识别出的未知攻击样本数}}{\text{所有未知攻击样本数}} \times 100\%$$

采用 KDD99 数据集, 根据电网工控系统的网络流量数据长度小、多用 UDP 协议等特点, 选取近似的数据和 41 个特征属性中的 18 个特征属性, 然后从筛选出的训练集中取 2044 个正常样本, 其中异常样本按各类异常的比例共取 2270 个, 组成新的训练集. 取新的训练集的 2.5%, 共 109 个作为有标记样本集. 测试样本集包含正常样本 2619 个, 异常样本 402 个, 其中未知攻击样本 66 个, 共 3021 个样本. 测试半监督 K-means

算法和动态半监督 K-means+OCSVM 算法的检测效果, 结果如表 1 所示.

表 1 两种算法的测试结果 (单位: %)

	检测率	误报率	未知攻击检测率
半监督K-means	78.61	2.67	22.73
动态半监督K-means+OCSVM	95.27	1.10	83.33

表 1 可见, 改进后的半监督 K-means 算法, 即动态半监督 K-means+OCSVM 算法, 提高了检测率、未知攻击检测率, 且降低了误报率, 流量异常检测能力优于半监督 K-means 算法.

5 总结

本文根据电网工控系统中控制网的内防水平低且其安全监测和防护缺乏内部网络流量异常检测的现状, 详细分析了电网工控系统的组成与结构、网络安全需求及其面临的威胁攻击. 提出了将流量异常检测技术应用于电网工控系统控制网的安全防护中, 从而与电力系统当前主要使用的访问控制安全措施共同形成对电网工控系统控制网的两级安全防护. 然后针对安全防护中的关键技术流量异常检测, 研究了流量异常检测方法的分类, 并根据不同方法的特点和电网工控系统的网络流量数据的特点, 选择基于机器学习的流量异常检测方法中的半监督 K-means 聚类算法. 针对半监督 K-means 算法的未知攻击检测率低等问题, 本文提出了基于熵的动态半监督 K-means+OCSVM 算法, 来改进半监督 K-means 算法的性能. 总之, 将流量异常检测技术应用于电网工控系统的安全防护中, 并对流量异常检测算法进行改进, 将有利于提升电网工控系统的内防水平和防范电网安全威胁.

参考文献

1 邹春明, 郑志千, 刘智勇, 等. 电力二次安全防护技术在工业控制系统中的应用. 电网技术, 2013, 37(11): 3227–3232.

2 高春梅. 基于工业控制网络的流量异常检测[硕士学位论文]. 北京: 北京工业大学, 2014.

3 赵延涛. 电力二次系统安全防护策略探究. 电子世界, 2016, (2): 152–154.

4 钟志琛. 基于网络流量异常检测的电网工控系统安全监测技术. 电力信息与通信技术, 2017, 14(1): 98–102.

5 刘松吟. 基于聚类分析的电力通信网络流量分类算法研究[硕士学位论文]. 南京: 东南大学, 2016.

6 陈雪鸿, 叶世超, 石聪聪. 浅谈工业控制系统信息安全等级保护定级工作. 自动化博览, 2015, (5): 66–70.

7 费稼轩, 张涛, 马媛媛, 等. 一种基于 BF-DT-CUSUM 算法的电网工控系统 DDoS 攻击检测方法. 电信科学, 2015, (S1): 106–112.

8 刘威, 李冬, 孙波. 工业控制系统安全分析. 信息网络安全, 2012, (8): 41–43.

9 王孝良, 崔保红, 李思其. 关于工控系统信息安全的思考与建议. 信息网络安全, 2012, (8): 36–37.

10 袁自强. 网络环境下白名单子系统的设计与实现[硕士学位论文]. 成都: 电子科技大学, 2011.

11 王海凤. 工业控制网络的异常检测与防御资源分配研究[硕士学位论文]. 杭州: 浙江大学, 2014.

12 许倩. 基于特征统计分析的异常流量检测技术研究[硕士学位论文]. 郑州: 解放军信息工程大学, 2012.

13 Han JW, Kamber M, Pei J. 数据挖掘-概念与技术 (原书第 3 版). 范明, 孟小峰译. 北京: 机械工业出版社, 2012.

14 张润, 王永滨. 机器学习及其算法和发展研究. 中国传媒大学学报 (自然科学版), 2016, 23(2): 10–18, 24.

15 朱佳佳, 陈佳. 基于熵和 SVM 多分类器的异常流量检测方法. 计算机技术与发展, 2016, 26(3): 31–35.

16 Basu S, Davidson I, Wagstaff K. Constrained Clustering: Advances in Algorithms, Theory, and Applications. London: Chapman & Hall/CRC, 2008.

17 李琳. 基于 OCSVM 的工业控制系统入侵检测算法研究[硕士学位论文]. 沈阳: 沈阳理工大学, 2016.