基于 Canopy-Kmeans 算法的电力企业流量数据分析研究

黄冠杰

(对外经济贸易大学 统计学院,北京 100105)

摘 要:针对电力企业关键信息基础设施大量业务数据易遭受网络攻击的现象,基于各业务信息系统下已有的网络安全设备,通过辅助设备采集流量数据,采用 Canopy-Kmeans 算法进行数据分析研究。首先通过实验证明了 Canopy-Kmeans 算法在处理流量数据时,相比传统 K-means 算法,具有更好的聚类效果,准确率提高约 11%;然后以采集到的电力关键业务系统的流量数据为基础,基于 Canopy-Kmeans 算法进行挖掘分析实验,完成相同类型流量数据的聚类,分析出攻击流量与业务流量的特征项,排除部分误报信息,合理开展网络安全防护工作。

关键词:电力;流量采集;Canopy-Kmeans;聚类;流量分析

中图分类号:TP391.1 文献标识码:A DOI:10.19358/j.issn.2096-5133.2022.01.003

引用格式:黄冠杰.基于 Canopy-Kmeans 算法的电力企业流量数据分析研究[J].信息技术与网络安全,2022,41(1):18-22.

Research on electric power enterprise flow data analysis based on Canopy-Kmeans algorithm

Huang Guanjie

(School of Statistics, University of International Business and Economics, Beijing 100105, China)

Abstract: Aiming at the phenomenon that a large number of business data of the key information infrastructure of electric power enterprises are vulnerable to network attacks, based on the existing network security equipment under each business information system, the flow data is collected through auxiliary equipment, and Canopy-Kmeans algorithm is used for data analysis and research. Firstly, through experiments, it is proved that the Canopy-Kmeans algorithm has a better clustering effect than the traditional K-means algorithm when processing flow data, and the accuracy rate is increased by about 11%. Then, the collected flow data of the power key business system is used, mining and analysis experiments are conducted based on the Canopy-Kmeans algorithm to complete the clustering of the same type of traffic data, analyze the characteristic items of attack traffic and business traffic, eliminate some misreporting information, and carry out network security protection work reasonably.

Key words: electricity; flow collection; Canopy-Kmeans; clustering; flow data analysis

0 引言

随着信息化与工业化的深度融合,各式各样的信息系统得到了广泛应用,信息安全问题不断涌现,关于信息网络的攻防战也愈演愈烈。近几年,国际上不法组织频繁攻击电力企业,层出不穷的网络攻击可能会导致系统出现故障,造成网络瘫痪,严重时造成大范围较长时间的电网故障,产生巨大影响和危害。电力系统作为现代社会的关键信息基础设施之一,其产生的大量业务数据及操作数据,也就自然成为了网络攻击的重点目标[1]。建立健全的电力数据分析体系,助力电力企业识别异常流量,保障电力数据安全迫在眉睫。

目前已有部分企业和专家针对电力数据的网络安全进行了研究,高翔[2]等人采用灰色关联分析和支持向量机算法对电力信息系统网络安全进行态势评估;李群[3]等人提出一种基于“聚类+分类”的恶意攻击检测方法,对流量预处理结果进行聚类,基于 CART 决策树对攻击簇实现分类;高鹏[4]等人

采用国产密码、量子密钥分发和区块链技术对电力终端和数据进行保护；刘川[5]等人基于云计算平台和SDN技术搭建了一体化电力数据安全防护框架，用于身份认证、攻击防范、入侵检测。

但目前大部分电力企业对于收集到的流量数据的挖掘和综合分析明显还不够。若要合理地进行数据分析并分类治理，首先要做到电力企业流量数据的充分采集，通过对采集到的数据进行ETL(Extract-Transform-Load，抽取-转换-加载)、挖掘和分析[6-7]，最终将分析结果应用于实际安全运维中，做到精准治理。本文以某电力企业的实际运行环境为例，首先简述本文所需的流量数据的数据来源及采集方式，得到其各业务系统下已有的网络安全设备中的流量数据，然后分别利用传统K-means算法与Canopy-Kmeans算法进行流量数据聚类分析，挖掘出攻击流量与业务流量的特征项，并排除部分误报信息。本研究对合理开展网络安全防护工作具有指导意义。

1 数据采集

为保证电力关键业务信息系统中的大量数据的安全，应首先明确需检测的业务数据分类。电力数据主要来源于电力企业的生产、运输、营销三层维度，本文主要采集各业务系统下的工控终端、网络设备、安全设备及第三方应用系统上的数据，通过态势感知以及各类辅助数据采集设备进行采集，主要包括日志采集、流量采集和Agent采集三种方式。数据采集的难点是数据体量大、种类多，若采集所有数据，将面临数据过载的问题，因此需要根据安全要素设计相应的特征，进行靶向数据采集[8]。

1.1 日志采集

日志数据的内容主要来自信息系统及涉及的终端、网络安全设备、数据库设备等。日志采集[3]的主要内容有采集时间、IP地址、设备名称、设备类别、操作日志、性能指标、主机原始日志内容等。不仅要对原始日志按照设备类型、日志类型、日志子类型等进行范式化处理，同时要根据事件告警规则(CPU阈值、内存阈值、磁盘阈值等)匹配，将范式化后的事件形成告警，最后记录入数据库中。

1.2 流量采集

流量采集采用旁路方式部署，接收交换机镜像端口提供的数据流，采集的对象主要有服务器、工作站、网络设备、安全防护设备、PLC、数据库等监测对象。采集网络流量数据分析异常网络协议信息，包含工业协议应用畸形报文、IP层畸形报文、协议完整性等。根据采集的网络流量数据建立实时流量模型，在流量分析时应考虑Modbus TCP、S7、IEC104、OPC DA、OPC UA、DNP3.0、MMS、Omron zFins、Ethernet/IP等工业协议深度包监测和分析能力。

1.3 Agent采集

主机探针是应用于收集各种操作系统主机信息的应用软件，软件支持常见的主流操作系统。可对主机产生的各种告警事件进行采集，并对主机进行安全性基线核查，以及对主机数据库运行情况进行采集。

采取综上所述的数据采集方式，能够基本覆盖关键业务信息系统的大量数据，利用网络安全设备掌握业务系统的安全态势，实时发现网络中的异常行为和对重要控制系统的攻击事件，用于事后的调查取证，同时也可为安全防护策略的配置提供参考依据。

2 算法介绍

2.1 K-means算法

K-means是一种较为经典的数据聚类[9-11]算法，该算法快速便捷、容易实现，处理大数据集的时候，该算法具有较好的伸缩性，但是不适合太离散的分类或样本类别不平衡的分类。该算法的基本原理如下：

(1)为每个聚类确定一个初始聚类中心，这样就有k个初始聚类中心；

(2)将样本集中的样本按照最小距离原则分配到最近邻的聚类；

(3)使用每个聚类中的样本均值作为新的聚类中心；

(4)重复步骤(2)~(3)，直到聚类中心不再变化；

(5)结束，得到K个聚类。

设数据集集合$E=\{x_1, x_2, \cdots, x_n\}$，$x_i=(x_{i1}, x_{i2}, \cdots, x_{in})$，$x_j=(x_{j1}, x_{j2}, \cdots, x_{jn})$，则样本$x_i$与样本$x_j$之间的距离为：

$$d(x_i, x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2} \quad (1)$$

用D代表误差，k代表初始聚类中心个数，S_i代表k个类中的一个，u_i是S_i的中心，x_j是S_i中的任意元素，则有效性函数如下：

$$D=\sum_{i=1}^{k}\sum_{x_j \in S_i}(x_j-u_i)^2 \quad (2)$$

针对K-means对初始的聚类中心较为敏感，不

同选取方式会得到不同结果；对噪声点敏感；样本只能归为一类，不适合多分类任务等缺点，可应用以下调优方式进行调优处理：如数据预处理（去除异常点）、合理选择 K 值、高维映射等。本文为提高数据聚类的准确性，提出了 Canopy－Kmeans 算法作为分析企业数据的工具。

2.2 Canopy 算法

Canopy 算法也是一种较为常见的聚类算法，是一种快速聚类技术，不需要事先指定聚类的个数，仅遍历一次样本集，就能得出最优的 canopy 数量。canopy 之间可相互重叠，所以通过 Canopy 算法聚类后样本集中的每一个样本并不能精确地归属于某一个 canopy，因此 Canopy 算法常用于数据预处理。Canopy 算法的执行流程如下：

（1）设定原始数据集阈值 $T1$，$T2$；

（2）随机选取数据集中 P，并计算 P 到数据集中其他点的距离 d；

（3）把所有 d 小于 $T1$ 的点生成 canopy，去除数据集中 d 小于 $T2$ 的点；

（4）不断重复步骤（2）、（3）直到数据集为空；

（5）算法结束，聚类完成。

Canopy 聚类结果示意图如图 1 所示，可以看出通过 Canopy 算法遍历一次样本集后，可以快速得到

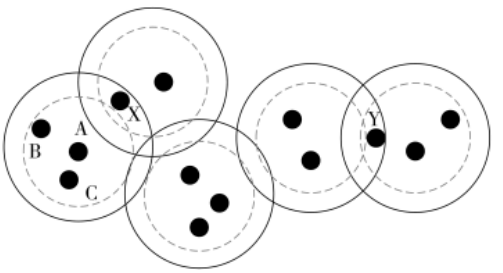


图 1　Canopy 聚类结果示意图

k 个可相互重叠的 canopy，同属于一个 canopy 的样本点最终属于同一个聚类（图中虚线所示），如样本点 A、B、C；但并非每个样本点都只属于一个 canopy，如样本点 X、Y，因此使用 Canopy 算法后仍需进行精确聚类，使得每一个样本点最终只属于一个聚类。

2.3 Canopy－Kmeans 算法

Canopy－Kmeans 算法[12－15]的主要原理为首先通过 Canopy 算法对初始的数据集进行预聚类，得到 k 个 Canopy 聚类中心，然后再将 Canopy 聚类结果作为 K－means 算法的初始聚类中心，再执行 K－means 聚类算法。预处理的优点是去除噪声点对 K－means 算法的干扰，提高算法的准确率，并且通过高效的 Canopy 算法，确定初始聚类中心，可以减少相应的计算量，加快聚类收敛速度。Canopy－Kmeans 算法的整体执行流程如图 2 所示。

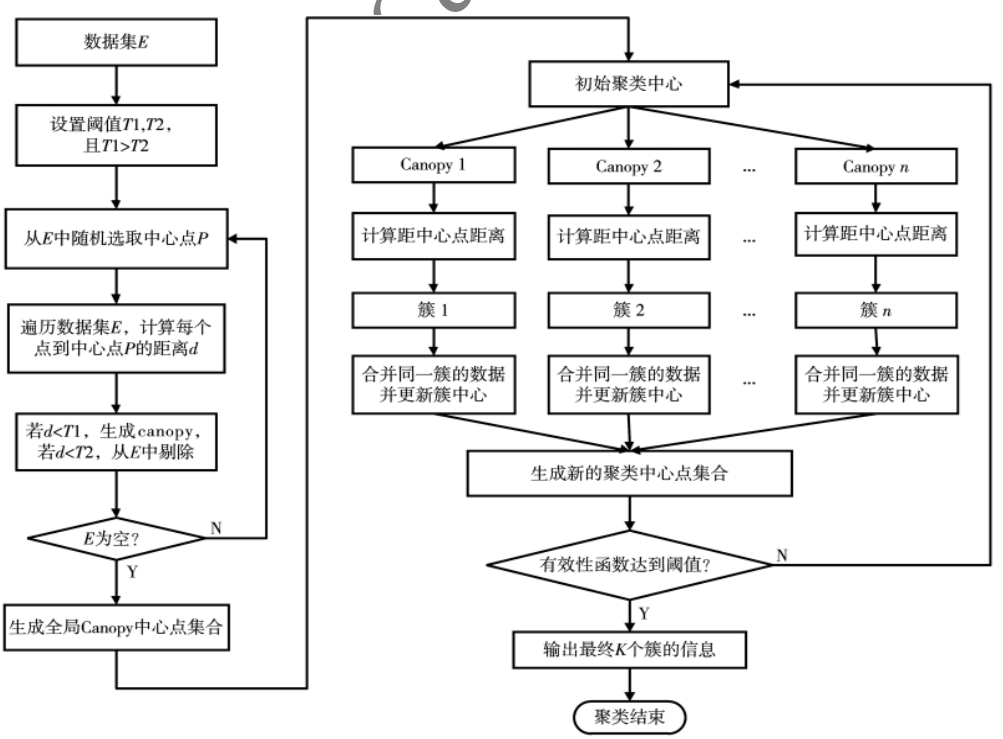


图 2　Canopy－Kmeans 算法执行流程图

3 数据分析

3.1 实验数据

本文的实验数据来源于某电力企业各业务系统下的流量数据，本次实验数据收集三个月的流量数据，分为三个实验数据集，共计10 000条记录，每条记录包含41个特征数，分为三种数据类型，分别为攻击流量、正常业务流量、误报流量，数量分布如表1所示。

表1 流量数据集类别分布 （个）

数据集	攻击流量	业务流量	误报流量	合计
1	1 051	2 258	213	3 522
2	931	2 187	186	3 304
3	857	2 189	128	3 174
合计	2 839	6 634	527	10 000

其中，端口扫描、暴力破解、漏洞利用、DDoS攻击等会被判别为攻击流量，内网主机访问OA系统、综合管理系统或内网终端交互等会被判定为业务流量，下级单位频繁访问业务系统或内网通传递大量数据可能会被判定为误报流量。

3.2 K-means与Canopy-Kmeans聚类分析对比

为了验证Canopy-Kmeans算法的可用性，选用三个数据集作为实验数据，分别利用K-means算法和Canopy-Kmeans算法进行相同的聚类实验，通过最终聚类准确性进行横向对比。

针对采集上来的流量数据，利用K-means算法和Canopy-Kmeans算法进行聚类得到流量分布准确率如表2所示。

表2 K-means和Canopy-Kmeans算法流量分布准确率对比 （%）

实验	算法	攻击流量	业务流量	误报流量
实验一	K-means	72.50	83.62	80.04
	Canopy-Kmeans	85.44	91.29	88.73
实验二	K-means	72.72	84.32	52.15
	Canopy-Kmeans	82.59	89.48	67.54
实验三	K-means	70.95	86.16	57.81
	Canopy-Kmeans	87.51	93.47	71.88

通过三组实验分别对比K-means算法和Canopy-Kmeans算法对实验数据聚类的结果可以发现，采用K-means算法最终得到的攻击流量、业务流量及误报流量和原始数据集分布相比，准确率为73.36%，而采用Canopy-Kmeans算法的准确率可达84.21%。在三组数据集下的K-means算法和Canopy-Kmeans算法的准确率对比如图3所示，相较于K-means算法，Canopy-Kmeans算法对于本次实验数据聚类准确性可提高约10.85%，实验表明Canopy-Kmeans算法拥有更好的聚类效果。

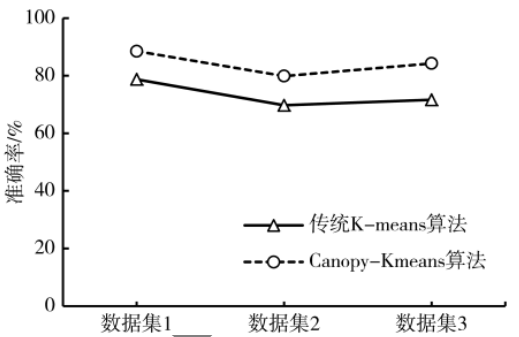


图3 K-means算法和Canopy-Kmeans算法聚类准确性对比图

3.3 基于Canopy-Kmeans算法的流量数据分析

基于本次实验采集的某电力企业流量数据，采用Canopy-Kmeans算法进行聚类，并对41个特征值进行挖掘分析，发现IP地址对结果的表达性更为突出，攻击流量、业务流量以及误报流量中访问频率较高集中分布在某几个IP段内，流量访问频次较高IP分布如表3所示。

表3 流量访问频次较高IP分布表

类别	IP地址段	数量
攻击流量	193.29.107.X 212.102.35.X 212.102.44.X 92.118.188.X	1 728
业务流量	10.81.X.X	5 729
误报流量	10.81.168.X 10.81.172.X	295

通过对本次实验数据进行聚类及挖掘分析，能够发现访问频率较多的IP地址段，对攻击流量及误报流量进行进一步筛查，发现上述攻击频次较高的IP地址段确实存在异常情况，对其采取封禁措施，对访问频次较高的误报流量进行充分核实，确定为某下级单位的正常访问操作，后续将其加入白名单，降低误报率。

4 结论

本文首先对某电力企业关键业务信息系统的数据采集进行了分项分类，包括日志采集、流量采集、Agent 采集三个方面，全面覆盖电力企业生产端的全数据采集，为后续的数据分析奠定基础。

其次，基于获取到的流量数据，分析利用 Canopy-Kmeans 算法较于 K-means 算法进行聚类分析的优势，进而准确地筛选出高危攻击 IP 地址，利用现有的网络安全设备将其封禁，实现主动防御，通过信息化手段在降低设备成本的同时提升准确率，有效提升电力企业信息系统网络安全水平。

后续工作将增加实验数据的多样性，多层次多维度地进行关联分析，并对算法做进一步的优化，通过聚类分析提高数据分类的准确性，以更有针对性地制定网络安全防护的技术方案，降低网络安全风险，提升企业信息系统的稳定性。

参考文献

[1] 吴凯峰，刘万涛，李彦虎，等.基于云计算的电力大数据分析技术与应用[J].中国电力，2015，48(2)：111-116，127.

[2] 高翔，陈贵凤，赵宏雷.基于数据挖掘的电力信息系统网络安全态势评估[J].电测与仪表，2019，56(19)：102-106.

[3] 李群，董佳涵，关志涛，等.一种基于聚类分类的物联网恶意攻击检测方法[J].信息网络安全，2021，21(8)：82-90.

[4] 高鹏，陈智雨，闫龙川，等.面向零信任环境的新一代电力数据安全防护技术[J].电力信息与通信技术，2021，19(2)：7-14.

[5] 刘川，李志伟，沈卫康.基于云计算及 SDN 的电力数据中心安全问题分析和防护策略[J].电子设计工程，2016，24(9)：136-138，143.

[6] 杜研哲.高性能网络流量采集和分析技术的实现[D].北京：北京邮电大学，2017.

[7] 李振国，郑惠中.网络流量采集方法研究综述[J].吉林大学学报(信息科学版)，2014，32(1)：70-75.

[8] 毛杰文.数据中心网络的异常流量检测及探针优化部署研究[D].上海：华东师范大学，2021.

[9] 王千，王成，冯振元，等.K-means 聚类算法研究综述[J].电子设计工程，2012，20(7)：21-24.

[10] 孙吉贵，刘杰，赵连宇.聚类算法研究[J].软件学报，2008，19(1)：48-61.

[11] 杨莉，沈鑫，李英娜，等.基于电力数据聚类分析的算法改进[J].云南电力技术，2017，45(6)：64-68.

[12] 陶莹，杨锋，刘洋，等.K 均值聚类算法的研究与优化[J].计算机技术与发展，2018(6)：90-92.

[13] 毛典辉.基于 MapReduce 的 Canopy-Kmeans 改进算法[J].计算机工程与应用，2012，48(27)：22-26，68.

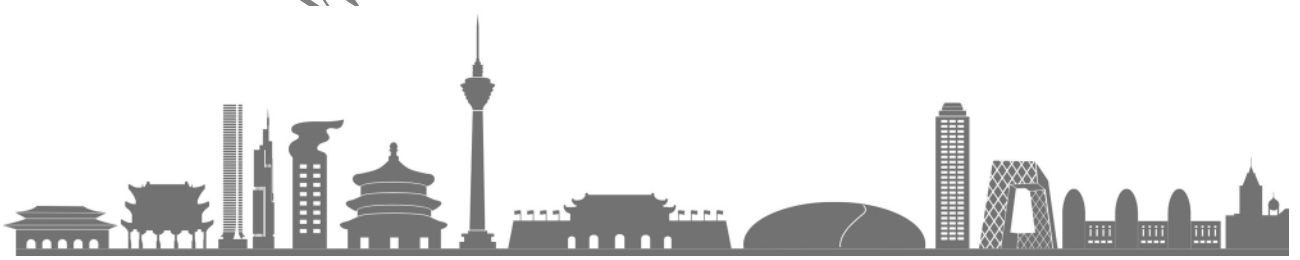
[14] 于连城，张译，张广德，等.基于 canopy-k-means 算法的电网数据挖掘算法的研究[J].国外电子测量技术，2018，37(7)：35-39.

[15] 赵庆.基于 Hadoop 平台下的 Canopy-Kmeans 高效算法[J].电子科技，2014，27(2)：29-31.

(收稿日期：2021-10-10)

作者简介：

黄冠杰(1993-)，男，硕士，助理工程师，主要研究方向：网络安全、工控安全。



版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所