文章编号:1003－6199(2024)01－0072－06 DOI:10.16339/j.cnki.jsjsyzdh.202401012

基于多元增量分析模型的流域电厂大数据采集监视

石发太†,孙卫军

(雅砻江流域水电开发有限公司,四川 成都 610051)

摘 要:提出基于多元增量分析模型的流域电厂大数据采集监视方法,通过数据分析发现异常电厂大数据,保障流域电厂安全运行。数据采集层在PIC18F8722单片机控制下利用多种传感器获取采集对象层的多种流域电厂数据,经数据汇集层的弹性消息总线和转发代理,将采集数据传输至数据处理层数据接收模块,多元增量分析模型调用接收到的数据,通过增量的方式构建电厂数据矩阵正常检测模型,对比其与待检测数据,发现异常电厂大数据,并将异常检测结果族谱写入数据存储层数据库所对应的列簇里,通过前端展示层将监视结果呈现给用户。实验结果表明,该方法可精准、快速采集流域电厂海量数据,可有效发现流域电厂的异常数据,并将其可视化呈现,确保流域电厂安全运行。

关键词:多元增量分析;流域电厂;大数据;采集监视;异常检测;弹性消息总线

中图分类号:TM311 **文献标识码**:A

Big Data Acquisition and Monitoring of Watershed Power Plants Based on Multivariate Incremental Analysis Model

SHI Fatai†, SUN Weijun

(Yalong River Hydropower Development Company, Ltd., Chengdu, Sichuan 610051, China)

Abstract: A big data collection and monitoring method for power plants in the basin based on the multivariate incremental analysis model is proposed to find abnormal big data of power plants through data analysis and ensure the safe operation of power plants in the basin. Under the control of PIC18F8722 single-chip microcomputer, the data acquisition layer uses a variety of sensors to obtain a variety of basin power plant data at the acquisition object layer. Through the elastic message bus and forwarding agent of the data collection layer, the collected data is transmitted to the data receiving module of the data processing layer. The multivariate incremental analysis model calls the received data, builds a normal detection model of the power plant data matrix by incremental means, and compares it with the data to be detected, Discover the big data of abnormal power plants, write the family tree of abnormal detection results into the corresponding column cluster of the data storage layer database, and present the monitoring results to the user through the front-end display layer. The experimental results show that this method can accurately and quickly collect massive data of power plants in the basin, effectively find abnormal data of power plants in the basin, and visually present them to ensure the safe operation of power plants in the basin.

Key words: multivariate incremental analysis; river basin power plant; big data; acquisition and monitoring; anomaly detection; elastic message bus

收稿日期:2022－10－10

作者简介:石发太(1987—),男,甘肃靖远人,本科,高级工程师,研究方向:电力自动化及远程控制,数据管理。

† **通信联系人**,E-mail: bo39639180@163.com

由于社会的进步，电网的发展，我们进入信息时代。各行各业的发展都是以电力为基础的，因此电力系统产生的数据激增，导致电力系统越来越庞大，加大了对流域电厂大数据的采集和监视的难度[1]。一旦流域电厂数据产生异常，就会产生重大影响，所以对流域电厂大数据的采集和监测很重要[2]。

因此寻找流域电厂大数据的采集和监测方法成为研究的目的。例如孙超等人主要采用分布式构架的 Hadoop 技术实现电力大数据的采集，采用 HIVE 分布式的数据库完成数据的存储，然后通过 Hadoop 技术完成电力大数据的处理，但实际应用过程中受 Hadoop 技术影响，存储较小文件受到影响，且延迟低数据访问受限[3]；曾乐等人提出通过 REST 接口和 Flume 框架完成数据的采集，通过 Kafka 实现监视数据，但是该方法实现数据的采集和监视的过程受数据规模影响，导致采集及监视效率低，延时较大[4]。

针对流域电厂数据的复杂性，多元增量分析模型能够将多路径上的流域电厂数据并行处理，并实时在线识别出流域电厂的异常数据，多元增量分析模型简单易懂，检测效率高，可用来监视流域电厂大数据的异常[5,6]。

因此本文提出了基于多元增量分析模型的流域电厂大数据采集监视方法，有效实现流域电厂大数据的采集监视。

1　流域电厂大数据采集监视总体结构

由采集对象层、数据采集层、数据汇集层、数据处理层、数据存储层和前端展示层组成流域电厂大数据采集监视方法的总体结构，用图 1 描述总体结构。

1.1　采集对象层

采集对象层主要是由不同状态与结构的数据组成。主要由五种数据组成：第一种是系统服务和程序的日志文件；第二种是关系数据库中集成数据状态的记录；第三种是为了提高运行状态数据存储和采集统计性能的内存数据库；第四种是实时发生的事件和数据出现异常时的告警；第五种是作为关键的采集监视对象服务网关[7-9]。

1.2　数据采集层

数据采集层主要由 PIC18F8722 单片机、电流传感器、电压传感器、功率传感器、温湿度传感器和 GPRS 无线传输模块构成。利用 GPRS 无线传输模块将 4 个传感器采集的流域电厂数据传输到数据汇集层。数据采集过程如图 2 所示。

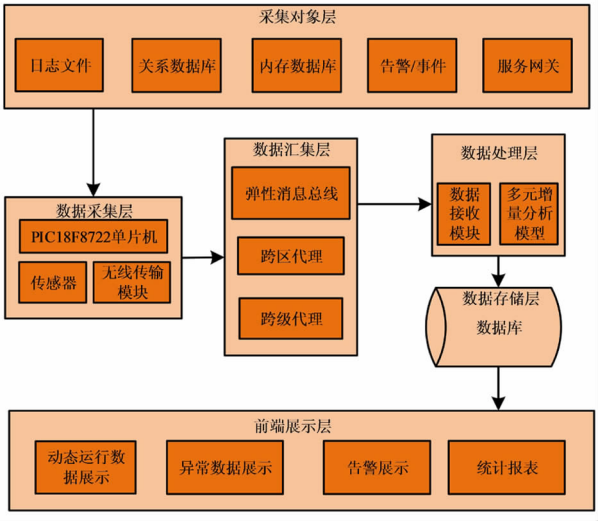

图 1　总体结构图

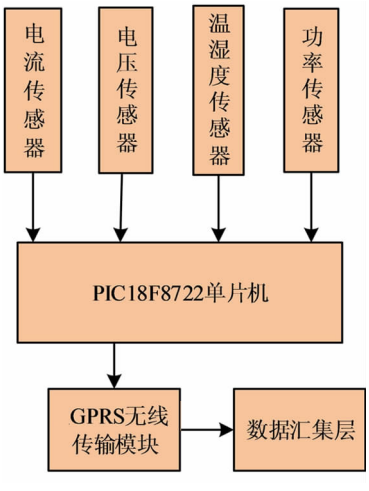

图 2　数据采集过程

数据采集层的核心是智能芯片 PIC18F8722 单片机。PIC 系列单片机具有运行速率快、稳定和持续的性能，性价比高。

通过型号为 WBI412N25 的电流传感器采集流域电厂的电流数据，通过型号为 WBV122S01 的电压传感器采集流域电厂的电压数据，通过型号为 WB9128-1 的功率传感器采集流域电厂的电压数据。电流传感器、电压传感器和功率传感器的参数表如表 1 所示。

表 1 传感器的参数表

参数值	电流传感器	电压传感器	功率传感器
工作相对湿度/%	5～95	5～95	5～95
工作温度/℃	±60	±60	±60
工作电压/V	5	5	—
输入规格	30A～400A	50mV～500V	—
输入方式	穿心输入	端子压接	—
输出类型	直流电压	直流电压	RS485 及两路模拟量输出
相电压 AC/V	—	—	57.7～289
线电压 AC/V	—	—	100～500
电流/A	—	—	1～5

1.3 数据汇集层

流域电厂数据通过弹性消息总线传输到数据处理层，针对流域电厂数据需要跨区跨级传输，通过跨区代理、跨级代理实现电厂数据在线传输。弹性消息总线、跨区代理和跨级代理组成数据汇集层[10,11]。

1.3.1 弹性消息总线

弹性消息总线作为流域电厂数据的传输通道，其具有海量吞吐量、自由扩展、传输机制可靠和支持所有语言接口等特点。

本文方法采用的弹性消息总线为 Kafka 消息总线，通过该消息总线与 Flume(消息采集系统)对接，实现数据直接写入数据库，有效地节省开发和运维的工作时间。

1.3.2 跨区代理

根据电力系统的要求，安全 I/II 的电厂业务数据输入安全 III 区前，需要隔离装置过滤，由于其只能单向通信，然而弹性消息总线通信方式是双向的通信方式，导致其不能跨过隔离装置，因此发送报文时要用到跨区转发代理。负责安全 I/II 区的内网代理和负责安全 III 区的外网代理构成跨区代理。

将安全 I/ II 区的弹性消息总线数据发送给消息消费者模块，然后传输给消息缓存，为减少报文占用隔离装置的带宽，利用压缩模块将报文压缩，然后将报文二次发送给安全 III 区的转发接收模块、解压模块、消息缓存后和消息生产者，安全 III 区的弹性消息总线接收经它们处理后的报文。图 3为跨区转发过程。

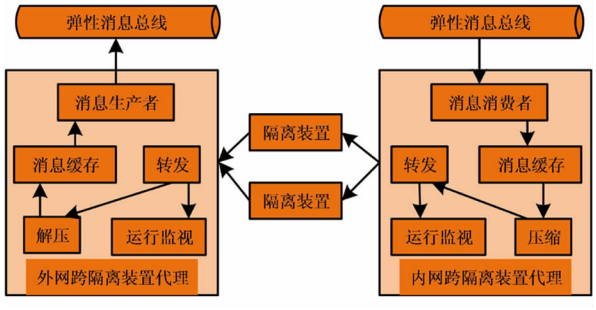

图 3 跨区转发过程

1.3.3 跨级代理

电网调控管理机构通过在各级调控机构设置采集代理，完成分布在省、地、县多级的电厂数据采集。由于跨级代理具有远程传输可靠性高的特点，故利用跨级代理将采集到的电厂数据，通过远程网络向上级单位传递，实现电厂数据的跨级传输。

启动发送代理端的滑动窗口，并编号窗口内的报文，并将其发送给接收代理端，经接收代理端确认后，以编号为依据，接收代理端接收发送代理端的窗口报文，发送至最后一条窗口报文时，在该报文上显示该窗口报文发送结束，经接收代理端确认，发送代理端自动清除已发送报文，完成跨级代理的数据转发[12]。若传输时出现窗口报文丢失，为完成断点续传，确认发送代理端接收到接收代理端最后一条报文编号，将此编号后的报文二次发送给接收代理端，图 4 为跨级转发过程。

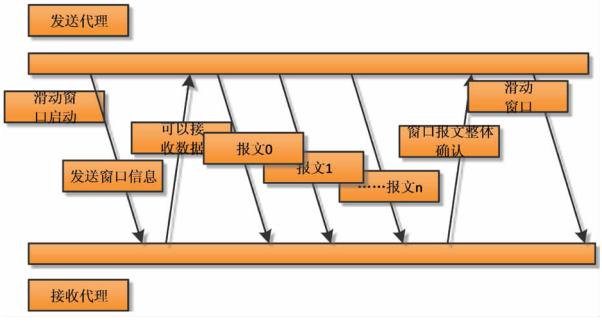

图 4 跨级转发过程

1.4 数据处理层

数据处理层通过数据接收模块得到弹性消息总线传输的采集到的数据，经多元增量分析模型调用完成异常数据检测，并将检测到的异常数据族谱写入数据存储层的数据库所对应的列簇里[13,14]。

1.4.1 数据接收

通过弹性消息队列得到 JSON 格式的电厂大数据报文，报文由报文体和报文头构成，用数组方式表示不同采集对象的族谱，如表 2 所示。

表 2　报文头各属性及 body 部分定义

报文头				boby 消息体
sourcetype	sourcename	msgtype	datapedigree	自定义内容
数据源类型	数据源名	消息类型	数据族谱	由各采集端定义

数据消息的 body 消息体采用 JSON 格式，定义如表 3 所示。

表 3　网关流量消息体定义

Key	Value	说明
id	Chart[32]	对象编号
name	Chart[128]	对象名称
time	longlong	采集时间
flow type	Octet	流量类型
flow rate	Double	数据流量
node_id	Chart[32]	节点 ID

1.4.2　基于多元增量分析模型的流域电厂异常数据检测

流域电厂数据的异常检测是通过与正常数据的对比得出的，通过建立流域电厂数据矩阵的正常检测模型，对比待测流域电厂数据与所构建正常模型，判定待测流域电厂数据是否异常。

为实现流域电厂异常数据的实时、在线检测，通过多元增量分析方法建立流域电厂数据矩阵的正常检测模型，可以通过设定阈值在线实时地发现流域电厂异常数据，对脱离了常态模型的异常数据予以及时报警。并通过新的观测样本不断更新正常检测模型，获取动态更新的多元增量分析模型，更好地实现流域电厂异常数据检测。

(1)建立流域电厂数据矩阵的正常检测模型

利用数据处理层接收到的流域电厂大数据，构建流域电厂数据矩阵 $\boldsymbol{T}$，归一化处理输入向量 $\boldsymbol{T}_i \in R^p(i=1,2,\cdots,N)$ 后，得到其协方差矩阵 $\boldsymbol{R}$，然后以多元分析算法为基础普分解 $\boldsymbol{R}$：

$$\boldsymbol{R}=\frac{\sum_{i=1}^{N}(\boldsymbol{T}_i-\boldsymbol{\mu})(\boldsymbol{T}_i-\boldsymbol{\mu})^N}{N} \tag{1}$$

$$\boldsymbol{RW}=\boldsymbol{W\Lambda} \tag{2}$$

$$\boldsymbol{\mu}=\frac{\sum_{i=1}^{N}\boldsymbol{T}_i}{N} \tag{3}$$

其中：$\boldsymbol{T}_i$ 的均值通过 $\boldsymbol{\mu}$ 来描述，特征值矩阵通过 $\boldsymbol{\Lambda}$ 来表示；特征向量矩阵通过 $\boldsymbol{W}$ 来表示，由特征向量 $\boldsymbol{w}_i$ 构成。

通过向量的方式表示流域电厂数据特征向量矩阵 $\boldsymbol{W}=[w_1 \quad w_2 \quad \cdots \quad w_p]$，把互相正交的所有 $\boldsymbol{w}_i$ 当作坐标轴构建一个新的特征空间，主坐标轴为所有的 $\boldsymbol{w}_i$。从小到大排列特征值，得到 λ_1，$\lambda_2,\cdots,\lambda_p$，进而得到对应的特征向量 $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_p$，其代表在欧氏空间能量(方差)方向对流域电厂数据进行排序，主坐标轴中能量极大值的方向为与特征极值对应的主坐标轴。特征空间的前 k 个坐标轴涵盖流域电厂数据的大多数能量，且该空间中的数据矩阵投影呈现出规律性，因此称该空间为正常子空间 V。除了前 k 个坐标轴之外的特征空间为异常流域电厂数据出现区域，将其定义为非正常子空间 $\widehat{V}$。将向量向空间 V 投影，利用其投影值得到新空间的特征向量矩阵为装载矩阵。仅分析正常子空间 V，其装载矩阵 $\boldsymbol{W}_k$ 对应的 V 投影值 d_i 为：

$$\boldsymbol{d}_i=\boldsymbol{W}_k^N(-\mu+T_i)V \tag{4}$$

其中：投影值 $\boldsymbol{d}_i$ 构成投影矩阵 $\boldsymbol{D}$，以正交矩阵为基础，可得出 $\boldsymbol{W}_k^N=\boldsymbol{W}_k^{-1}$，重构流域电厂数据矩阵输入向量 $\boldsymbol{T}_i$ 为：

$$\boldsymbol{T}_i=\boldsymbol{\mu}+\boldsymbol{W}_k\boldsymbol{d}_i \tag{5}$$

(2)模型的动态更新

通过实时产生的观测样本优化基于多元增量分析的流域电厂异常数据模型参数，建立动态更新的增量模型实现流域电厂数据的在线实时异常检测。假设流域电厂大数据采集监视的时间为 t，观测值用 $\{\boldsymbol{T}_i\}_{i=1}^{t}$ 表示，均值用 $\boldsymbol{\mu}$ 表示，正常子空间 V 下的装载矩阵用 $\boldsymbol{W}_k^{(t)}$ 表示，投影矩阵用 $\boldsymbol{D}^{(t)}=[\boldsymbol{d}_i]_{i=1}^{t}$ 表示。在经过的流域电厂大数据采集监视时间为 $t+1$ 后又产生新增观测值 T_{t+1}，使 T_{t+1} 向正常子空间投影并重构，产生的残余向量表示为：

$$\boldsymbol{s}=\boldsymbol{X}_{t+1}-\boldsymbol{W}_t^{(t)}\boldsymbol{d}_{t+1}-\boldsymbol{u}^{(t)} \tag{6}$$

当前 V 和 s 正交，将 s 变成新的基，使特征向量 $\boldsymbol{W}_k^{(t)}$ 维度加一，与之对应的投影矩阵 $\boldsymbol{D}^{(t)}$ 更新，二者变化表达式为：

$$\boldsymbol{W}'=[\boldsymbol{W}_k^{(t)} \quad \boldsymbol{s}/\|\boldsymbol{s}\|] \tag{7}$$

$$\boldsymbol{D}'=\begin{bmatrix}\boldsymbol{D}^t & \boldsymbol{d}_{t+1}\\ 0 & \|\boldsymbol{s}\|\end{bmatrix} \tag{8}$$

已经存在的特征向量受其空间的改变而旋转，以主元分析法为基础求解 $\boldsymbol{D}'$，以均值 γ 与特征向量为依据，重新构建矩阵 $\boldsymbol{U}$，得到优化参数后的多元增量模型公式为：

$$\boldsymbol{\mu}^{(t+1)}=\boldsymbol{\mu}^{(t)}+W'\gamma \tag{9}$$

$$\boldsymbol{W}_k^{(t+1)}=\boldsymbol{W}'\boldsymbol{U} \tag{10}$$

$$\boldsymbol{D}^{(t+1)} = \boldsymbol{U}^{N}(\boldsymbol{D}^{'} - \gamma 1), 1 \in R^{t+1} \quad (11)$$

(3)数据异常检测

通过对比残余向量与阈值实现异常数据检测。设定阈值 s_t，若阈值 s_t 小于残余向量 $\boldsymbol{s}$ 的 2 范数，则判定观测的流域电厂数据产生异常，然后将异常数据族谱写入数据库中所对应的列簇里。

1.5 数据存储层与前端展示层

数据存储层可将数据处理层通过多元增量分析模型获取的流域电厂异常数据检测结果予以存储，便于前端展示层各功能模块随时调用[15]。

前端展示层等同于人机交互界面，通过调用数据存储层中的列将数据库中的信息展示出来，主要展示 4 项内容，分别为动态运行数据、分类统计数据、告警、统计报表。

2 实验分析

本次实验以某省内的大型流域电厂为实验对象，该流域电厂包括 5 个子电厂，拥有 3 台 32 万千瓦、1 台 64 万千瓦机组，160 万千瓦的总装机容量。该流域电厂每日产生的运行大数据约 308 GB。将本文方法应用至该大型流域电厂，验证本文方法对该流域电厂实时数据的采集监视能力。

采用本文方法采集该流域电厂的电压数据，将采集后的数据与流域电厂数据库中存储的真实电压数据进行对比。并将流域电厂的数据进行极限值处理，只保留范围在 35V 到 60V 之间的电压数据，采集数据与真实数据的对比结果如图 5 所示。

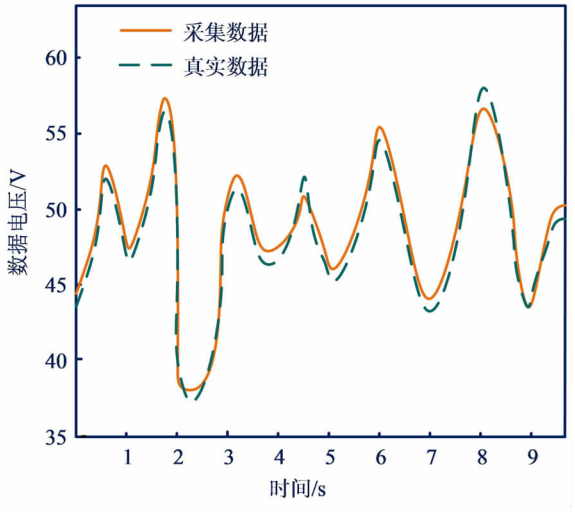

图 5 流域电厂数据采集结果对比

实时采集该流域电厂电压数据时，数据的耗时和数据量如表 4 所示。

表 4 采集数据结果统计

子电厂 ID	开始时间	结束时间	总耗时	数据量/条
30HAH68RT551	10:00:00	10:01:38	0:01:38	421236
30HAH68RT552	10:01:38	10:03:38	0:02:00	407856
30HAH68RT553	10:03:38	10:05:18	0:01:40	412569
30HAH68RT554	10:05:18	10:07:10	0:01:52	425578
30HAH68RT555	10:07:10	10:09:06	0:01:56	415694

通过图 5 可以看出，利用本文方法采集到的数据与该流域电厂的真实数据相近，验证了本文方法可以完成流域电厂数据的精准采集。通过表 7 可以看出，利用本文方法采集该流域电厂电压数据时，耗时仅为 90～120 s，采集数据量为 40 万～43 万条。可以看出采用本文方法对大数据采集效率高采集，可以有效地完成流域电厂大数据的采集。

本次实验验证利用本文方法对该流域电厂的数据进行实时监测的效果，以该流域电厂的数据为数据源，分别在采样周期为 1050 和 2001 时注入异常，异常分别为两节点间不寻常的高速率传输异常(ALPHA/DoS)和非正常服务器数据请求异常(flash crowd)。注入方式如表 5 所示。采用本文方法对该过程进行监视，结果如图 6 所示。

表 5 异常注入方式

注入异常名称	注入周期	持续周期	注入方式
ALPHA/DoS	1050	5	将注入异常数据幅度从 135%迅速增加到 195%
flash crowd	2001	100	将注入异常数据幅度从 120%迅速增加到 160%，再逐步还原

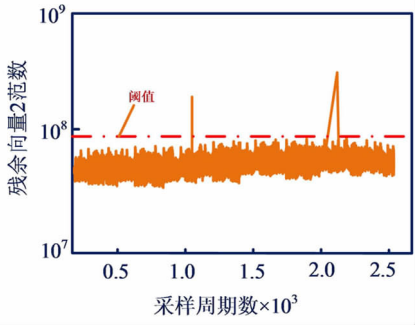

图 6 异常数据检测结果

图 6 为在 2.5×10^3 个采样周期内，本文方法检测到的异常数据情况，图中虚线为设定的阈值。通过图 6 可以看出，本文方法检测到在 1.05×10^3

和 $2.0\times10^{3}-2.1\times10^{3}$ 周期内发现超出设定阈值的残余向量 2 范数值，检测结果与注入攻击时间相同，由此可以看出采用本文方法可以有效地检测流域电厂数据的异常。

采用本文方法在线实时监测该流域电厂大数据，观测该流域电厂一周内出现异常情况，监测界面图如图 7 所示。

图 7　系统监测界面图

通过图 7 可以看出，采用本文方法对该流域电厂进行监视时，流域电厂大数据采集监视界面显示，在 2021 年 10 月 26 日 09：34：23 时，子电厂 1 中主变压器 1 以及子电厂 3 中的变压器 1 存在异常，导致其接地电流出现异常增大。实验结果表明，本文方法可以完成流域电厂大数据的在线实时监测，及时识别流域电厂异常数据，节省维修时间，减小经济损失。

3　结　论

在流域电厂大数据采集监视方法中引入多元增量分析模型来处理数据，提出一种基于多元增量分析模型的流域电厂大数据采集监视方法，本文方法通过数据处理层中的多元增量分析模型处理数据采集层的采集到的流域电厂的数据，将处理后的电厂数据存入数据库，通过前端展示层将数据分类展示出来，达到流域电厂实时数据的接入、跨区和跨级传输，有效地实现了流域电厂大数据采集监视，提高了运维人员工作效率，促进了大数据技术的实用性。

参考文献

[1] 刘长良，许涛，王梓齐，等. 基于智能电厂大数据的关键参数目标值挖掘技术[J]. 热力发电，2019，48(9)：14－21.

[2] 叶康，冷喜武，肖飞，等. 基于大数据标签技术的电网监控智能分析方法[J]. 电测与仪表，2019，56(4)：75－79.

[3] 孙超，常夏勤，王永贵，等. 电力大数据多元数据采集监视技术研究与应用[J]. 计算机技术与发展，2020，30(7)：180－185

[4] 曾乐，孙超，张来恩，等. 基于大数据技术的气象业务监视数据采集处理[J]. 计算机，2021，38(7)：181－188.

[5] 刘晓华. 多元方差分析模型的构建与应用[J]. 统计与决策，2019，35(1)：75－78.

[6] 薛少勃，沈晶，刘海波. 基于持续增量模型的低速端口扫描检测算法[J]. 计算机应用研究，2020，37(4)：1125－1127＋1131.

[7] 刘永辉，张显，孙鸿雁，等. 能源互联网背景下电力市场大数据应用探讨[J]. 电力系统自动化，2021，45(11)：1－10.

[8] 李俊楠，李伟，李会君，等. 基于大数据云平台的电力能源大数据采集与应用研究[J]. 电测与仪表，2019，56(12)：104－109.

[9] 鲍俊如，金莹，熊亮. 基于大数据云平台的电力能源大数据采集方法及应用探讨[J]. 中国新通信，2021，23(14)：101－102.

[10] 黄晓君，陈峥，吴双，等. 基于物联网的电厂通信设备大数据在纺织业供电中的运用[J]. 染整技术，2022，44(5)：44－48.

[11] 黄哲学，何玉林，魏丞昊，等. 大数据随机样本划分模型及相关分析计算技术[J]. 数据采集与处理，2019，34(3)：373－385.

[12] 郭银芳. 大数据环境下网络数据传输及融合优化仿真[J]. 计算机仿真，2019，36(4)：120－123＋183.

[13] 刘凌云，钱辉，邢红杰，等. 一种基于 Q-学习算法的增量分类模型[J]. 计算机科学，2020，47(8)：171－177.

[14] 周利军，邢立勐，白龙雷，等. 基于多元非线性回归模型的 XLPE 电缆终端环形损伤故障特征[J]. 高电压技术，2021，47(9)：3124－3133.

[15] 刘雅婷，王永程，强延飞，等. 基于随机映射与聚类的网络流量异常检测[J]. 计算机仿真，2019，36(3)：289－293.