DOI:10.20033/j.1003-7241.(2026)06-0135-05

一种基于动态融合局部异常因子的时序数据快速清洗方法

郝福忠[1], 杨宇方[1], 姬 哲[1], 张 静[2], 王军义[2]

(1.国网河南省电力公司,郑州 450000;2.国网河南省电力公司信息通信公司(数据中心),郑州 450000)

摘 要:针对现有数据清洗方法存在清洗不彻底、效率较低的问题,提出一种基于动态融合局部异常因子的时序数据快速清洗方法。设置多个不同的 k 值分别计算各数据点的局部异常因子,通过归一化处理消除量纲影响,并结合比值加权计算得到综合局部异常因子,采用动态融合策略对数据点异常程度进行综合评估,实现异常数据的精准识别。利用布谷鸟算法对 K-means 聚类算法的初始聚类中心进行优化,避免传统 K-means 易陷入局部最优的问题。将异常数据删除后形成缺失数据集,利用优化后的 K-means 算法对数据进行聚类,找出包含缺失位置的簇,计算该簇内数据的均值作为原始真实值的估算,并以此替代异常数据,完成数据修复与清洗任务。测试结果表明,该方法清洗全面指数更高、清洗数据均方误差更小、清洗时间开销更少,表明该方法能够快速、全面、准确地完成时序数据中的异常数据清洗任务。

关键词:动态融合局部异常因子;时序数据;改进 K-means 聚类算法;快速清洗方法;异常检测;参数优化

中图分类号: TP132.66 **文献标志码**: A **文章编号**: 1003-7241(2026)06-0135-05

A fast cleaning method of temporal data based on dynamic fusion of local abnormal factors

Hao Fuzhong[1], Yang Yufang[1], Ji Zhe[1], Zhang Jing[2], Wang Junyi[2]

(1. State Grid Henan Electric Power Company, Zhengzhou 450000, China;
2. Grid Henan Information & Telecommunication Company(Data Center), Zhengzhou 450000, China)

Abstract: A research method for rapid cleaning of time-series data based on dynamic fusion of local abnormal factors is proposed to address the problems of incomplete cleaning and low efficiency in existing data cleaning methods. Set multiple different k values to calculate the local anomaly factors of each data point, eliminate the influence of dimensionality through normalization, and combine ratio weighting to calculate the comprehensive local anomaly factors. Use dynamic fusion strategy to comprehensively evaluate the degree of anomaly of data points and achieve accurate identification of anomalous data. Optimize the initial cluster centers of K-means clustering algorithm using cuckoo algorithm to avoid the problem of traditional K-means falling into local optima. After deleting abnormal data, a missing dataset is formed. The optimized K-means algorithm is used to cluster the data, identify clusters containing missing positions, calculate the mean of the data within the cluster as an estimate of the original true value, and use it as a substitute for abnormal data to complete data repair and cleaning tasks. The test results show that this method has a higher comprehensive cleaning index, smaller mean square error in cleaning data, and less cleaning time cost, indicating that the proposed method can quickly, comprehensively, and accurately complete the task of cleaning abnormal data in time-series data.

Keywords: dynamic fusion local anomaly factor; time series data; improved K-means clustering algorithm; quick cleaning method; abnormal detection; parameter optimization

在当今信息化技术飞速发展的时代,数据呈现爆炸式增长,大数据已经充斥了人们的工作、学习和生活。大数据中蕴含了许多有价值的信息,通过挖掘可以实现异常监测、故障检测、分类识别等,因此大数据挖掘已成为许多领域研究的热点[1]。然而,采集到的初始大数据往往质量并不高,其中包含了很多不完整、模糊、重复、异常的数据[2],这些干扰数据的存在使得数据挖掘变得困难,从而导致许多大数据的利用率并不高。面对这种情况,在进行数据挖掘之前,人们通常会进行数据清洗工作,以尽可能降低干扰数据对挖掘结果的影响,提高大数据的质量[3]。其中,异常数据是最难处理的干扰数据之一,因为它们通常表现出与正常数据相似的特征,甚至有时更加难以区分,因此清洗异常数据是非常必要的研究方向。

在上述背景下,很多专家和学者提出了解决方法。例

收稿日期:2025-03-26;**录用日期**:2025-04-17

基金项目:国网河南省电力科技项目,河南公司数据管理应用体系及实施管控研究(5217Q0220005)

作者简介:郝福忠(1971—),男,硕士研究生,正高级工程师,主要研究方向:企业数字化转型。

引用本文:郝福忠, 杨宇方, 姬哲, 等. 一种基于动态融合局部异常因子的时序数据快速清洗方法[J]. 自动化技术与应用, 2026,45(6):135-139.
(Hao Fuzhong, Yang Yufang, Ji Zhe, et al. A fast cleaning method of temporal data based on dynamic fusion of local abnormal factors[J]. Techniques of Automation and Applications, 2026,45(6):135-139.)

如梅玉杰等[4]在其研究中以配电网状态时序数据为研究对象,首先提取了该数据的局部异常因子,然后利用GMM算法对局部异常因子进行聚类并进行聚类排序,得到异常阈值,通过对比异常阈值,检测存在异常的数据,最后通过LSR-RF算法计算异常数据的原始真实数据,以替代异常数据,完成异常数据清洗。李琳等[5]在其研究中以大型风力机工作数据为研究对象,首先将数据划分为不同的区间,然后计算各区间数据差分值,以此为基础估算DBSCAN算法的Eps值,最后利用DBSCAN算法对异常数据进行检测,完成数据清洗。Ding等[6]在其研究中以物联网数据为清洗对象,首先基于物联网数据的时-空相关性特征,通过LRaSMD模型将异常数据清洗问题转换为优化问题,然后利用ISTA算法求解该优化问题,最后将ISTA迭代展开成深度神经网络结构,以此为基础,实现异常数据清洗。

由于有的异常数据特征不明显,导致清洗不彻底,还有遗漏的异常数据,此外待清洗的数据量一般都比较大,因此清洗效率一般都较低,这也是当下异常数据清洗方法常出现的两个问题。针对上述问题,研究一种基于动态融合局部异常因子的时序数据快速清洗方法。

1 时序数据快速清洗方法研究

异常数据的存在对于数据挖掘具有很强的干扰性。为此,研究一种基于动态融合局部异常因子的时序数据快速清洗方法。该研究主要分为两部分,即前一部分进行基于动态融合局部异常因子的异常时序数据检测,后一部分进行基于异常时序数据检测的清洗研究。

1.1 动态融合局部异常检测

局部异常因子是指数据的异常程度,该值越大,越有可能是异常数据。数据清洗的关键是检测出异常数据,而异常数据检测的关键是局部异常因子计算[7]。基于此,本章节首先进行时序数据的局部异常因子提取工作,具体过程如下。

步骤1 输入待清洗的时序数据。

步骤2 对时序数据按照一定的时间长度进行分割,划分为多个时间段。因为一旦数据采集时间较长,采集对象的状态就有可能发生巨大变化,例如从正常状态发展为故障状态。这时后一时间段的数据就与前一时间段的数据有明显的区别,在分布上看,与离群异常数据很相似,很容易将这一部分的正常数据看成异常数据,从而发生清洗错误的问题,所以需要对时序数据进行离散分割[8]。

步骤3 以其中一个离散区间内的数据为例,记为$A=\{a_t,t=1,2,\cdots,T\}$。其中,T代表时序数据分割时间长度;t代表时间点。

步骤4 计算A中每一个数据a_t的k-distance(即距离点a_t第k远的那个距离值),记为k-dis(a_t)。

步骤5 计算a_t的第k距离邻域,即寻找点a_t的第k距离及之内的所有数据点,记为$B_{k-\mathrm{dis}(a_t)}$。

步骤6 计算$B_{k-\mathrm{dis}(a_t)}$中所有点与a_t之间的可达距离,计算公式如下:

$$D_k(a_t,b_t)=\max[k-\mathrm{dis}(b_t),d(a_t,b_t)] \tag{1}$$

式中,$k-\mathrm{dis}(b_t)$代表$B_{k-\mathrm{dis}(a_t)}$中数据点b_t的k-distance;$D_k(a_t,b_t)$代表点b_t到点a_t的第k可达距离;$d(a_t,b_t)$代表a_t与b_t之间的距离。

步骤7 根据$D_k(a_t,b_t)$,按照下述公式计算a_t的局部可达密度。

$$C_k(a_t)=\frac{1}{\left(\dfrac{\sum\limits_{b_t\in B_{k-\mathrm{dis}(a_t)}}D_k(a_t,b_t)}{|B_{k-\mathrm{dis}(a_t)}|}\right)} \tag{2}$$

式中,$C_k(a_t)$代表局部可达密度。

按照上述公式计算a_t的$B_{k-\mathrm{dis}(a_t)}$中各个邻域点的局部可达密度,记为$C_k(b_t)$。

步骤8 按照下述公式计算$C_k(b_t)$与$C_k(a_t)$之间比值的平均值,即a_t的局部异常因子。

$$L(a_t)=\frac{\left(\dfrac{\sum\limits_{b_t\in B_{k-\mathrm{dis}(a_t)}}C_k(b_t)}{|B_{k-\mathrm{dis}(a_t)}|}\right)}{C_k(a_t)} \tag{3}$$

式中,$L(a_t)$代表数据点a_t的局部异常因子。

按照上述流程遍历剩余所有离散区间,计算区间内每个数据的局部异常因子。当$L(a_t)>1$时,说明a_t的密度小于其邻域点密度,认为a_t为异常数据,否则认为a_t为正常数据[9]。

在上述传统局部异常因子计算流程中,参数k的选择至关重要,因为后续计算都是基于该参数的设置来寻找近邻点,然后进行逐步运算的。传统局部异常因子算法中参数k只会设置一个,对于波动较为平缓的小规范时序数据来说是足够的,但是对于波动较为剧烈的大规范时序数据来说,异常数据会更为隐蔽,只设置一个参数k是无法满足准确检测异常数据的要求的[10]。针对这一点,本研究中设置3个不同数值的参数k,重复上述传统局部异常因子计算流程,这样就得到了时序数据中每个数据点的3个局部异常因子。针对一个数据点的3个局部异常因子,为方便后续检测工作,提高检测效率,需要对3个局部异常因子进行动态融合[11]。首先对局部异常因子进行归一化处理,即

$$\dot{L}_i(a_t)=\frac{L_i(a_t)-\bar{L}(a_t)}{\Delta L(a_t)} \tag{4}$$

其中,

$$\bar{L}(a_t)=\frac{\sum\limits_{i=1}^{3}L_i(a_t)}{3} \tag{5}$$

$$\Delta L(a_t)=\sqrt{\frac{\sum\limits_{i=1}^{3}[L_i(a_t)-\bar{L}(a_t)]^2}{3}} \tag{6}$$

式中，$L_i(a_t)$ 代表数据点 a_t 的第 i 个局部异常因子；$\bar{L}(a_t)$ 代表数据点 a_t 的局部异常因子平均值；$\Delta L(a_t)$ 代表数据点 a_t 的局部异常因子标准差；$\dot{L}_i(a_t)$ 代表归一化后数据点 a_t 的第 i 个局部异常因子。

接着计算数据点 a_t 每个局部异常因子与数据点 a_t 局部异常因子总和的比值，即

$$g_i(a_t)=\frac{\dot{L}_i(a_t)}{\sum_{i=1}^{3}\dot{L}_i(a_t)} \tag{7}$$

最后根据 $g_i(a_t)$ 进行动态融合，融合公式如下，即

$$\hat{L}(a_t)=\sum_{i=1}^{3}g_i(a_t)L_i(a_t) \tag{8}$$

式中，$\hat{L}(a_t)$ 代表动态融合后的数据点 a_t 的局部异常因子。

基于融合后的 $\hat{L}(a_t)$，按照 $L(a_t)>1$ 的规则，确定异常数据。

2.2　基于异常数据检测的清洗研究

在利用动态融合局部异常因子完成异常时序数据检测之后，并不意味着清洗工作的完成。本章节基于检测结果，计算异常数据所在位置处的原始的真实的正常数据，以替代异常数据，填补到异常数据位置处，完成对异常数据的覆盖，实现异常数据清洗[12]。在本研究中，利用改进 K-means 聚类算法进行数据清洗。基础 K-means 聚类算法运行效率高且操作简单，但是该算法运行质量受到初始聚类中心选择的直接影响，若是选择不恰当，很容易出现局部最优问题[13]。针对这一点，利用布谷鸟算法进行初始聚类中心选取，以优化 K-means 聚类算法，具体过程如下图 1 所示。

基于图 1 流程，最终得到 M 个 K-means 算法的聚类中心，实现算法的优化和改进。接下来，利用改进后的算法进行数据清洗[14-15]。具体过程如下：

步骤 1　设置算法的初始参数，包括聚类中心个数 M、阈值 λ。

步骤 2　输入不完整的时序数据。将上一章节检测出来的异常数据从时序数据中删除，形成一个具有缺失数据的时序数据，记为 S。

步骤 3　将 S 划分为 n 组。

步骤 4　从 n 组数据中筛选出存在缺失数据的小组，记为 s_j（第 j 个存在缺失数据的小组）。

步骤 5　计算 s_j 中数据的均值，即

$$\bar{x}(s_j)=\frac{\sum_{t=1}^{T}x_t(s_j)}{K} \tag{9}$$

式中，$\bar{x}(s_j)$ 代表 s_j 的数据均值；$x_t(s_j)$ 代表 s_j 中第 t 个数据；T 代表 s_j 中数据长度。

步骤 6　将 $\bar{x}(s_j)$ 填补到被删除的异常数据的位置处，

步骤 7　按照上述过程遍历所有存在缺失数据的小组，形成初始完全时序数据，记为 $\hat{S}$。

步骤 7　利用优化后的 K-means 算法对 $\hat{S}$ 中数据进行聚类。具体过程如下。

1）计算 $\hat{S}$ 中各个数据到聚类中心的距离 $d(x_t,O_m)$，其中 O_m 代表第 m 个聚类中心。

2）将数据划分到距离最近的簇当中；

3）重新计算每个新簇的质心；

4）判断质心是否发生变化。若发生变化，回到 3）；否则数据聚类。

步骤 8　找出包含填补数据的簇。

步骤 9　计算该簇中数据的均值并代替原始填补的数据[16-17]。

步骤 10　计算前后两次填充值的差值，记为 Δx_t。

步骤 11　判断 Δx_t 是否小于设定的阈值？若小于，完成时序数据中异常数据的清洗工作，否则回到步骤 7。

经过上述过程，以估算出来的正常值代替异常值，完成了异常数据清洗工作。

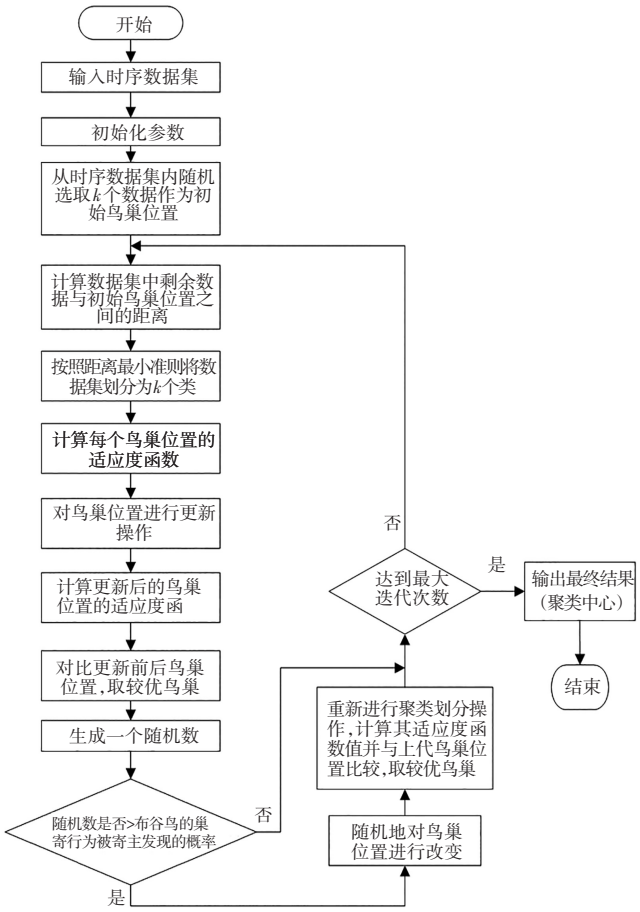


图 1　布谷鸟算法求取 K-means 聚类中心的基本流程

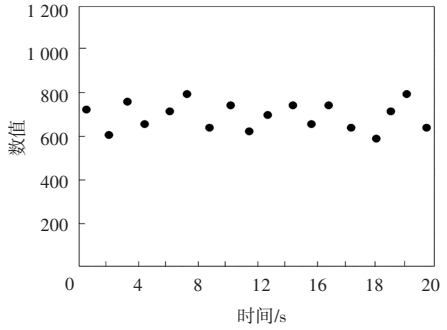
Fig. 1　Basic process of using cuckoo algorithm to obtain K-means clustering centers

3　数据清洗效果验证

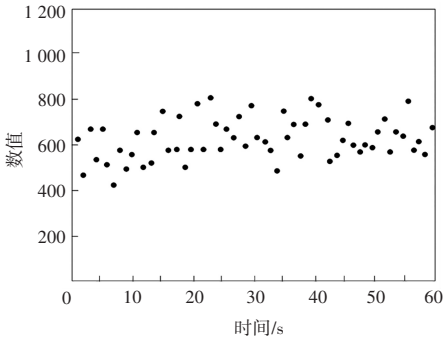
为进一步证明所研究方法的研究效果，与基于 LSR-RF 的清洗方法、基于 DBSCAN 的清洗方法以及基于深度神经网络的清洗方法进行对比。

3.1 数据来源

借助 MATLAB 仿真工具生成 2 个不存在异常数据的仿真时序数据集,时序数据长度分别为短(数据量为 20 个)、长两种(数据量为 60 个),如下图 2 所示。



(a) 短数据集



(b) 长数据集

图 2 仿真时序数据

Fig. 2 Simulation timing data

为模拟异常数据存在的情况,从 2 个仿真时序数据集随机选取一定量的数据并人工设置异常值。设置情况如下表 1 所示。

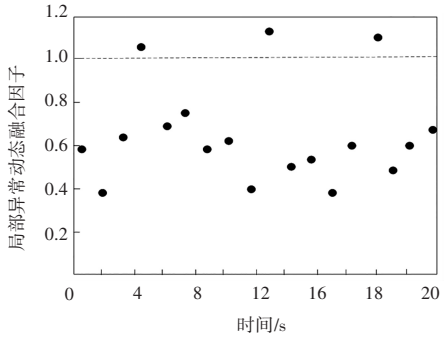
表 1 异常值设置情况

Tab. 1 Setting of Outliers

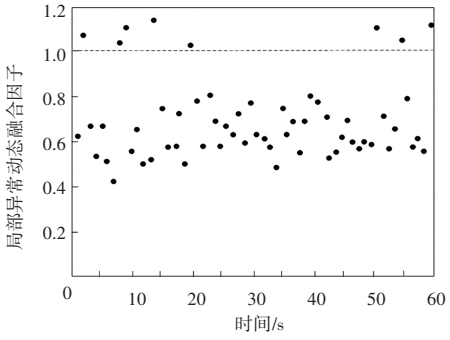
仿真时序数据集	异常值设置数量/个	原始正常数值	异常设置值	异常数据时刻/s
数据集 1	3	787	940	4
		718	425	10
		605	856	17
数据集 2	8	476	636	2
		586	362	8
		486	263	9
		636	421	14
		592	741	20
		668	856	50
		621	320	55
		685	862	60

3.2 局部异常动态融合因子结果

针对这 3 个存在异常数据的仿真时序数据集,计算每个数据动态融合后的局部异常因子并以此为依据,进行异常数据判断,结果如下图 3 所示。



(a) 短数据集



(b) 长数据集

图 3 局部异常动态融合因子计算结果

Fig. 3 Calculation results of dynamic fusion factor for local anomalies

从图 3 中可以看出,所研究方法应用下,短数据集中时刻 4 s、时刻 10 s、时刻 17 s 时的数据局部异常动态融合因子大于 1,说明这 3 个时刻对应的数据为异常数据;长数据集中时刻 2 s、时刻 8 s、时刻 9 s、时刻 14 s、时刻 20 s、时刻 50 s、时刻 55 s、时刻 60 s 时的数据局部异常动态融合因子大于 1,说明这 8 个时刻对应的数据为异常数据。将检测出来的结果对比表 1,得出检测结果与异常值设置情况一致,证明了基于动态融合局部异常因子在异常时序数据检测中的应用效果。

3.3 数据清洗结果

利用改进 K-means 聚类算法估算异常数据所在位置处的原始的真实的正常数据,以替代异常数据,完成数据清洗。估算出来的正常数据值如下表 2 所示。

表 2 正常数据估算值

Tab. 2 Estimated values of normal data

仿真时序数据集	正常数据估算值
数据集 1	784.33
	722.63
	614.68
数据集 2	470.32
	569.87
	483.66
	624.78
	603.98
	664.87
	614.72
	674.06

3.4 数据清洗效果

在相同的实验条件下,利用基于 LSR-RF 的清洗方法、基于 DBSCAN 的清洗方法以及基于深度神经网络的清洗方法进行数据清洗,然后针对清洗结果,选取清洗全面指数、清洗数据均方误差和清洗所产生的时间开销。清洗全面指数用于验证清洗的全面性,判断是否有遗漏的异常数据;清洗数据均方误差用于验证清洗数据的准确性;清洗时间开销用于验证清洗时速度。清洗全面指数计算公式如下,即

$$P = \frac{h}{H} \tag{10}$$

式中, P 代表清洗全面指数; h 代表清洗的正确的异常数据数量; H 代表数据集中异常数据总量。

清洗数据均方误差计算公式如下,即

$$R = \frac{\sum_{i=1}^{n} (r_i - \dot{r}_i)}{n} \tag{11}$$

式中, R 代表清洗数据均方误差; r_i、$\dot{r}_i$ 代表真实数据值与估算出来的替代值; n 代表异常数据数量。

结果如下表 3 所示。从表 3 中可以看出,与其他 3 种方法对比,动态融合清洗方法应用下,清洗全面指数更高、清洗数据均方误差更小、清洗时间开销更少,由此说明本方法能够以更快的速度,全面且准确地完成时序中异常数据的清洗工作。

表 3 数据清洗效果

Tab. 3 Data cleaning effect

方法	仿真时序数据集 1			仿真时序数据集 2		
	全面指数	数据均方误差	时间开销/min	全面指数	数据均方误差	时间开销/min
动态融合清洗方法	1.0	1.58	1.62	1.0	1.37	1.85
基于 LSR-RF 的清洗方法	0.845	3.47	2.32	0.745	6.52	4.52
基于 DBSCAN 的清洗方法	0.862	2.13	2.52	0.824	4.89	4.86
基于深度神经网络的清洗方法	0.921	3.98	1.96	0.836	6.85	3.83

4 结论

为保证数据挖掘信息的全面性和准确性,采集到的大数据一般都是一段时间内时序数据。受到采集设备、采集环境和采集操作等因素影响,某一个时间点的数据就会脱离其他正常数据分布规律,成为离群点,也就是异常数据。该数据的存在对于数据挖掘和利用都产生严重的干扰。针对这一点,研究一种基于动态融合局部异常因子的时序数据快速清洗方法。该研究中通过计算数据点的局部异常因子并动态融合,实现了动态检测,为数据清洗提供依据。最后对清洗效果进行了测试,通过清洗全面指数、清洗数据均方误差和清洗所产生的时间开销 3 个指标的对比,证明了所研究清洗的方法的效果。

参考文献

[1]陈高超, 曾学文, 付名江, 等. 基于条件深度卷积生成对抗的汽轮机数据清洗[J]. 自动化技术与应用, 2025, 44(11):17-21+69.

[2]刘宇璐, 张雅洁, 王罗, 等. 基于分段图像识别的风电场异常运行数据清洗方法[J]. 可再生能源, 2023, 41(4):500-506.

[3]李阳, 沈小军, 张扬帆, 等. 基于速度-关联约束的风电机组风速感知异常数据识别方法[J]. 电工技术学报, 2023, 38(7):1793-1807.

[4]梅玉杰, 李勇, 周王峰, 等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. 电力系统保护与控制, 2023, 51(7):158-169.

[5]李琳, 董博, 郑玉巧. 大型风力机异常功率数据清洗方法[J]. 兰州理工大学学报, 2022, 48(3):65-70.

[6]Ding X, Wang H, Li G, et al. IoT data cleaning techniques: A survey [J]. Intelligent and Converged Networks, 2022, 3(4):325-339.

[7]匡俊搴, 赵畅, 杨柳, 等. 一种基于深度学习的异常数据清洗算法[J]. 电子与信息学报, 2022, 44(2):507-513.

[8]唐艺灵. 改进型加权实时融合算法在提高遥测数据质量中的应用[J]. 探测与控制学报, 2022, 44(4):81-86.

[9]黄光球, 赵羲轩, 陆秋琴. 基于 KPCA-IF-WRF 模型的多源 VOCs 数据清洗方法研究[J]. 安全与环境学报, 2022, 22(6):3412-3423.

[10]武晓冬, 王麟斌, 牛天聪, 等. 基于数据清洗及 LSTM 神经网络的 CVT 故障诊断[J]. 自动化技术与应用, 2025, 44(12):173-176.

[11]汪海宁, 陈昱明, 樊涛, 等. 光伏系统采集数据在线清洗与修复方法研究[J]. 太阳能学报, 2022, 43(6):57-65.

[12]郭蕊, 李奕霏, 高育栋. 基于人工智能技术的电力信息运维数据整合平台[J]. 自动化技术与应用, 2025, 44(06):76-79.

[13]杨军, 刘洋, 杨玉奇. K-means 聚类算法在网络安全检测中的应用研究[J]. 信息与电脑, 2023, 35(10):209-211.

[14]金兰, 陈荆亮. 一种用于异常数据流挖掘的改进 Apriori 算法研究[J]. 计算机仿真, 2025, 42(1):480-484.

[15]匡伟祥. 基于数据挖掘技术的拖拉机发动机故障诊断[J]. 农机化研究, 2025, 47(2): 244-248.

[16]张明泽, 栾文鹏, 艾欣, 等. 基于边缘计算的台区短期负荷预测方法[J]. 电测与仪表, 2024, 61(4): 93-99.

[17]马晓琴, 薛峪峰, 杨媛. 考虑时序数据缺失的配电网线损率动态预测方法[J]. 电子设计工程, 2023, 31(17):132-136.