中国农村水利水电
China Rural Water and Hydropower
ISSN 1007-2284,CN 42-1419/TV

《中国农村水利水电》网络首发论文

题目：水电机组状态监测数据清洗方法
作者：金容鑫，娄岱松，黄华德，毛汉领
收稿日期：2021-09-30
网络首发日期：2022-01-19
引用格式：金容鑫，娄岱松，黄华德，毛汉领. 水电机组状态监测数据清洗方法[J/OL]. 中国农村水利水电.
https://kns.cnki.net/kcms/detail/42.1419.TV.20220119.1056.015.html

网络首发时间：2022-01-19 18:44:06
网络首发地址：https://kns.cnki.net/kcms/detail/42.1419.TV.20220119.1056.015.html

水电机组状态监测数据清洗方法

金容鑫，娄岱松，黄华德，毛汉领

（广西大学机械工程学院，南宁　530004）

摘要：为了提高数据质量，针对水电机组的监测数据中含有高比例的异常错误，从错误数据辨识和缺失数据恢复两方面进行了研究。针对错误数据辨识问题，结合数据中具有的延时性和相似性等特征利用 K-means 聚类方法辨识出异常数据；针对数据缺失等问题，则根据数据中具有的小样本、非线性等特点，利用支持向量回归恢复缺失数据；构建了基于 K-means 聚类和支持向量回归的错误辨识和缺失恢复的数据处理方法。最后利用实际的监测参数验证了此方法的有效性。

关键词：水电机组；监测数据；数据清洗; K-means 聚类；数据恢复

中图分类号： TV734　　**文献标志码：**A

Research on cleaning method of condition monitoring data of hydropower units

JIN Rong-xin, LOU Dai-song, HUANG Hua-de, MAO Han-ling

(School of Mechanical Engineering, Guangxi University, Nanning 530004, China)

Abstract：There have high proportion of abnormal errors in the monitoring data of hydropower units, to improve the data quality, this paper studies the identification of wrong data and the recovery of missing data. Aiming at the problem of error data identification, combined with the characteristics of delay and similarity in the data, the k-means clustering method is used to identify abnormal data; For problems such as missing data, according to the characteristics of small samples and nonlinearity in the data, the support vector regression is used to recover the missing data; A data processing method for error identification and loss recovery based on the k-means clustering and support vector regression is constructed. Finally, the effectiveness of this method is verified by using the actual monitoring parameters.

Key words：hydropower unit; monitoring data; data cleaning; k-means clustering; data recovery

0 引　言

数据的准确性是对水电机组开展运行分析与故障诊断的基础。传感器故障、采集设备故障、电磁信号干扰、通信设备损坏等原因，导致原始数据中存在大量不完整的数据和异常的数据。这些不良的异常数据对水电机组的运行分析与故障诊断将带来严重的负面影响。形成高质量的数据资源，对于提高

收稿日期：2021-09-30
基金项目：广西科技基地和人才专项(桂科 AD19259002)。
作者简介：金容鑫(1997-)，男，硕士研究生，研究领域为水电机组早期故障诊断。E-mail：980499696@qq.com。
通讯作者：毛汉领(1963-)，男，工学博士、教授、博士生导师，主要研究方向为金属结构及零件疲劳损伤的非线性检测、超声非线性检测的理论及方法、机械设备状态监测与故障预示等。E-mail：maohl79@gxu.edu.cn。

水电机组运行分析与故障诊断的准确性具有重要意义。

数据辨识和恢复受到各行业研究者的高度重视，提出了针对各种数据特点的辨识和恢复处理方法。如在数据辨识方面，孟建良[1]提出了基于 Spark 和聚类分析的辨识不良数据的新方法，将抽样技术和最大最小距离法引入到传统 K-means 算法中，克服了收敛速度慢且易陷入局部极小等问题，并用于对输电网状态估计中的不良数据进行检测和辨识。方睿[2]基于 MNMR 状态估计算法提出了一种基于UPU并行加速的量测不良数据辨识方法，有较好的不良数据辨识能力。胡阳[3]提出了一种基于置信等效边界模型的风功率数据清洗方法，用于异常数据识别剔除。Wang[4]提出一种基于时空相关性约束的不良数据检测与识别方法，并用于电力系统功率平衡数据的清理。Yu[5]提出了基于进化对称损失函数的方法，直接识别输出电力系统不良数据。Shuang Hao 等[6]提出了一种基于极大独立集的异常检测方法，由字符串之间距离量化修复数据。钟建伟[7]提出基于蚁群算法的改进新息图法，并用数值仿真结果验证方法的效果。Lin[8]把高效的 LNR 测试应用于识别不同组中的多个坏数据，识别和纠正超大电力系统中的测量误差。在数据恢复方面，洪梓铭[9]提出基于优先级分配策略的电网信息系统数据恢复方法，使物理损坏情况下的数据得以恢复。王方超[10]针对 GPS 坐标序列中的缺失值问题，提出基于数据驱动的 RegEM 插补算法，在大量数据缺失的情况下效果优于传统方法。谢智颖[11]提出了整合缓冲区、四分位数、时间依赖网络等时空处理方法的清洗方法，提高了公交车到达时间的预测精度。张帅[12]建立电力负荷的多尺度时序特征建模，提出周尺度的电力负荷缺失数据恢复方法，并尝试恢复年度等长时段日负荷数据。Fan[13]提出了一种基于卷积神经网络的结构健康监测振动数据恢复方法，具有较好的丢失数据恢复能力。Li[14]提出了一种基于相关隔离森林和注意力的 LSTM(CiF-AL)的数据清理方法，优化了异常数据恢复的定位精度和校正精度。王子馨[15]提出基于长短期记忆网络的缺失数据恢复方法，可用于提高电力系统量测数据质量。针对多源时间序列缺失数据恢复问题，刘歌[16]提出一种基于双重正则矩阵分解的恢复方法，并验证了算法的有效性。

综上所述，不同领域的数据辨识和恢复的研究较多，但鲜有针对水电机组监测数据的研究与应用。本文针对水电机组监测数据的延时性和相似性、小样本和非线性、异常数据和缺失数据并存等特点，利用 K-means 聚类方法辨识异常数据、支持向量回归恢复缺失数据，构建水电机组监测数据的辨识和恢复方法，并利用某水电站的实际监测数据验证方法的有效性。

1 辨识和恢复算法

1.1 K-means 聚类算法

经典的最为广泛使用的 K-means 聚类算法主要以欧氏距离作为相似性衡量指标，表征数据的相似性和延续性，其计算流程如下：

（1）从 N 个数据样本中随机选择 k 个样本并初始化这个 k 聚类中心 $\{C_1, C_2, \cdots, C_k\}$。

（2）计算每一个样本到每一个聚类中心的欧氏距离，依次比较每一个样本到每一个聚类中心的距离，将样本分配到距离最近的聚类中心的类簇，形成 k 簇，并根据以下公式更新 k 簇，计算公式如下：

$$d = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - C_i \right\|_2^2 \tag{1}$$

式中：C_i 为簇 S_i 的中心样本。

（3）对新的 k 簇重新计算该类的聚类中心，计算公式如下：

$$C_i = \frac{1}{S_i} \sum_{x \in S_i} x \tag{2}$$

重复步骤(2)-(3)至满足条件 $\left| C_{n+1} - C_n \right| \le \varepsilon$ 后计算终止。

K-means 聚类算法简单、快速，对大数据集有高效率和可伸缩性，可进行模块化分类。

1.2 支持向量回归方法

应用支持向量回归(Support Vector Regression，SVR)方法，通过非线性映射将样本集从低维空间映射到高维空间。对 n 个训练样本 $D = \{(x_i, y_i)\}_{i=1}^{n} \left(x_i \in X = R^n, y_i \in Y = R \right)$，该非线性映射也就是超平面可以定义为：

$$f(x) = \omega^T x + b \tag{3}$$

式中：x, ω, b 分别是输入向量，权重及截距。

于是，SVR 方法可形式化为：

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \frac{1}{n} \sum_{i=1}^{n} L_\varepsilon [f(x), y, \varepsilon] \tag{4}$$

式中：C 为惩罚因子；L_ε 是不敏感损失函数，将 ε 作为不敏感误差，则不敏感损失函数 L_ε 的表达式为：

$$L_\varepsilon [f(x), y, \varepsilon] = \begin{cases} 0, & |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon, & |y - f(x)| < \varepsilon \end{cases} \tag{5}$$

对于回归错误的数据点，引入松弛变量 ξ_i 和 ξ_i^*，可将 L_ε 代入(4)式可得

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{6}$$

$$\text{st}\quad \begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i, \\ f(x_i) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, \cdots, n \end{cases} \tag{7}$$

引入拉格朗日乘数以及核函数将目标函数转换为对偶形式：

$$\max_{\alpha,\alpha^*}\left[\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)y_i-\sum_{i=1}^{n}\left(\alpha_i+\alpha_i^*\right)\varepsilon -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)K\left(x_i,y_i\right)\right] \tag{8}$$

$$\text{t}\quad \begin{cases} \sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0 \\ 0\le\alpha_i\le C,\ 0\le\alpha_i^*\le C \end{cases} \tag{9}$$

式中：α_i 和 α_i^* 是拉格朗日乘数，$K\left(x_i,x_j\right)$ 为核函数，可以将低维空间的内积运算转换为高维空间的函数运算。最小化拉格朗日函数后，获得 SVR 表达式：

$$f\left(x\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)K\left(x_i,x_j\right)+b \tag{10}$$

SVR 常用的核函数为线性核函数：

$$K\left(x_i,x_j\right)=x_i^T x_j \tag{11}$$

对于样本的分类问题，用基于线性核函数的 Linear SVR 可以快速有效解决。

1.3 算法评价指标

可使用均方根百分比误差（Root Mean Square Percentage Error，RMSPE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）和拟合优度（Goodness of Fit，R^2）等 3 个指标评价 Linear SVR 在回归学习中的性能。RMSPE 表示回归结果的准确性，结果越准确，RMSPE 越小；MAPE 代表回归结果的一致性，结果越一致，MAPE 越小；R^2 代表拟合优度，拟合优度越大，则模型的拟合效果越好。它们计算公式分别为：

$$RMSPE=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t-Y_t}{Y_t}\right|^2} \tag{12}$$

$$MAPE=\frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t-Y_t}{Y_t}\right| \tag{13}$$

$$R^2=1-\frac{\sum_{t=1}^{N}\left(Y_t-Y_t\right)^2}{\sum_{t=1}^{N}\left(Y_t-\overline{Y}\right)^2} \tag{14}$$

式中：N 为样本个数；为 t 时刻的实际值；Y_t 为模型在 t 时刻的预测值；$\overline{Y}$ 为实际值的平均值。

1.4 算法实施步骤

（1）使用 K-means 聚类算法辨识错误数据。从数据集中随机选取 k 个样本点作为初始聚类中心，将剩下的样本分配到欧氏距离最小的聚类中心所对应的类簇，并更新类簇直到满足条件，完成分类。辨识出错误数据，实现数据清洗。

（2）使用清洗后的数据，利用 Linear SVR 求取函数 $f(x)$ 的参数，使其在训练后能够通过

样本缺失的自变量 x 预测对应的因变量，实现对缺失数据的恢复。

（3）计算拟合后 RMSPE、MAPE、R^2 指标，对恢复后的数据进行评价。

2 异常数据辨识

以广西南宁某水电站 2015 年 5 月投入使用的额定功率为 30.77MW 的灯泡贯流式水电机组为研究对象，该水电站计算机监控系统中存储了自运行以来的大量水电机组运行状态监测数据。监控系统采集的监测参数包括电流、电压、功率等电气参数，振动、行程、位移、导叶开度、水位、流量、压力等机械参数，以及瓦温、油温、绕组温度等热量参数，主要测点的部分原始数据见表 1。在实际中，由于传感器异常、机组停机、日常维修等问题会导致存储的数据存在丢失、奇异等问题，在对数据分析之前需要对原始数据进行清洗。

水电机组的运行过程状态是连续的，具有高度重复性和高度的相似性，不易发生突变，可使用 K-means 聚类法完成错误数据的辨识。下面以有功功率、定子线圈温度为例进行数据清洗。每个监测参数有 4998 个数据项，每十分钟记录一次数据。有功功率及定子线圈温度参数的直方图如图 1 所示。

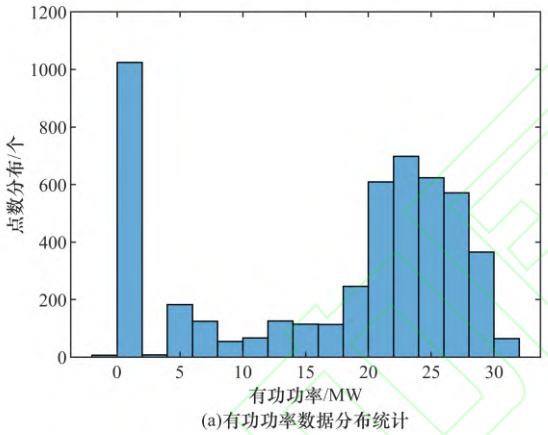
在图 1(a)中，横轴 0-5 范围内出现的频率是 1221 次，很明显该部分数据是存在错点的，需要对这些数据进行预处理。这些接近于 0 的数据大部分是由于机组停机后，由监测系统自动将这些数据补充到当前时刻的数据中。由图 1(b)可知，在机组停机后温度传感器收集到的数据依然存储在监控系统的数据库中，因此需要对该部分数据进行辨识。

使用 K-means 聚类法辨别“功率-定子线圈温度”之间的错误数据，如图 2 所示为功率-定子线圈温度分布图，从其中随机选取 k 个样本点作为初始聚类中心，更新类簇直到满足条件，完成分类，研究不同聚类中心个数对辨识错误数据效果的影响，如图 3 所示。

表 1 2015 年 10 月 9 日部分测点部分原始数据

Tab.1 Part of original data of some measuring points on Oct. 9, 2015

监测参数	时间							
	8:00	8:10	8:20	8:30	8:40	8:50	9:00	9:10
有功功率/(MW)	20.55	15.82	16.04	15.90	15.92	16.09	16.12	10.73
受油器摆度 X/(um)	64.39	120.91	87.60	61.38	86.50	102.61	131.22	97.66
受油器摆度 Y/(um)	62.94	113.73	114.85	69.93	103.24	91.03	112.32	116.25
组合轴承摆度 X/(um)	115.18	112.11	114.35	114.53	112.31	112.40	114.07	115.55
组合轴承摆度 Y/(um)	147.74	146.30	146.12	145.30	145.01	146.80	147.69	146.12
水导摆度 X/(um)	45.09	43.01	40.99	43.62	43.19	41.66	43.62	41.35
水导摆度 Y/(um)	37.58	38.05	37.91	35.41	38.58	39.18	40.82	35.87
受油器振动 X/(um)	102.75	83.42	103.86	92.22	96.54	92.99	72.90	86.63



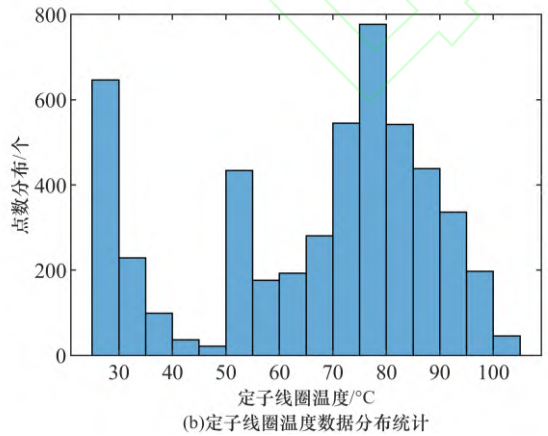


图 1 监测数据分布统计

Fig.1 Distribution statistics of monitoring data

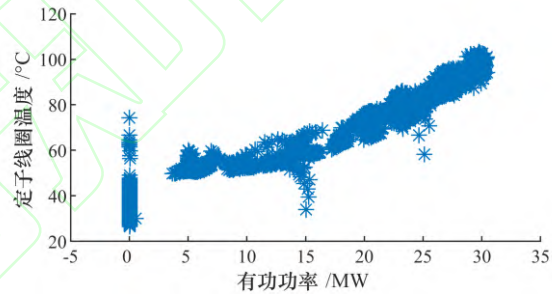
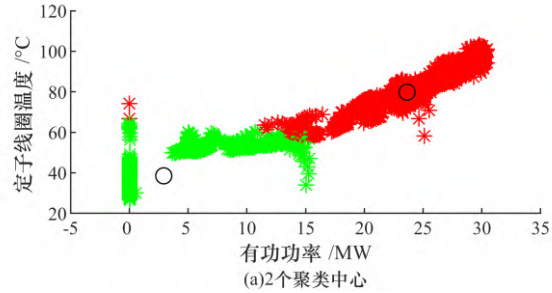
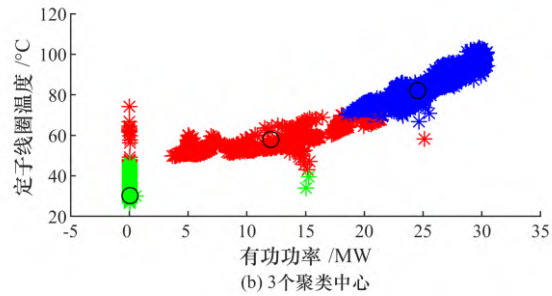


图 2 有功功率-定子线圈温度

Fig.2 Active power-stator coil temperature





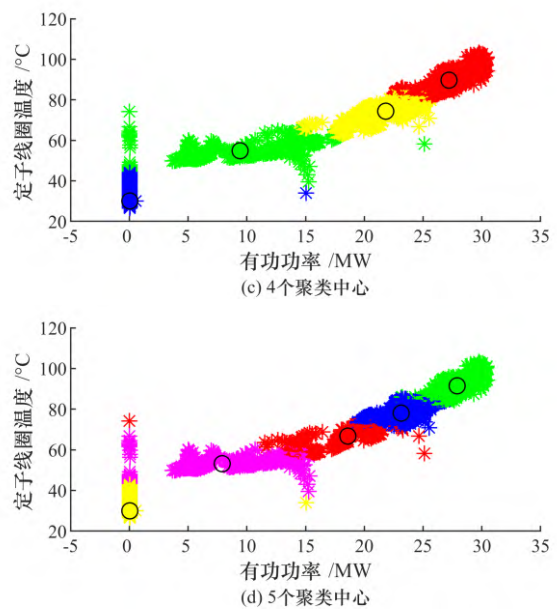


图 3 聚类中心个数对功率-曲线聚类结果的影响

Fig.3 The influence of the number of clustering centers on the power curve clustering results

从图 3 中可以看出聚类中心个数的不同，每个聚类的分布范围存在较大的差异。随着聚类中心个数的变化，零功率点也随之变化，当聚类中心的个数为 3 时，错误数据检测出的概率是 98.5%。因此使用 K-means 聚类方法是可以有效识别出这些异常的错误数据，在实际应用过程中，需要对检出率与误检率进行综合权衡，确定合理的聚类中心个数。在清洗完辨识出的错误数据后，还需要结合以下方式进行数据清洗。

（1）因监控系统自身出现的问题如上位机故障等，无法记录数据，导致一些时间段内出现数据不变，或者数据量全部为 0 的状况。因此，剔除数据中的所有状态量为“0”或者是数据不变的记录。

（2）因机组在停机状态，监控系统在正常运行，此时生成的数据中功率接近于 0，这些数据对机组分析评估没有意义。因此，剔除数据中有功功率接近于 0 且机组转速为 0 的记录。

根据上述方法剔除错误数据后，定子线圈温度的直方图如图 4 所示。

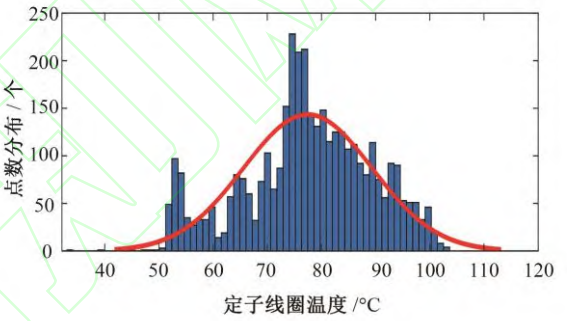


图 4 错误数据处理后定子线圈温度统计分布

Fig.4 Statistical distribution of stator coil temperature after error data processing

从图 4 的频数分布来看，处理后的定子线圈温度数据质量相比于处理前的质量有了显著提升。为了进一步分析该数据，用高斯分布进行曲线拟合并对该统计分布的参数估计，均值为 u=77.4682，方差为 σ=11.8905 均值的 0.95 置信区间为[77.0875, 77.8488]，方差的 0.95 置信区间为[11.6274, 12.1659]，定子线圈温度近似服从于高斯分布，定子线圈温度还受到机组工况等因素影响。

3 缺失数据恢复

以某水电站 1 号机组 2015 年 7 月 15 日

至 2016 年 7 月 30 日的定子线圈温度和有功功率数据为研究对象，共 1982 组数据，前 1487 个数据用于模型训练，后 495 个数据用于模型验证。对这些水电机组监测数据，使用基于线性核函数的 Linear SVR 各个参数拟合模型的性能指标见表 2。

表 2 Linear SVR 模型参数

Tab.2 Parameters of linear SVR model

类型	核函数	内核规模	约束	算法
Linear SVR	Linear	Automatic	Automatic	Automatic

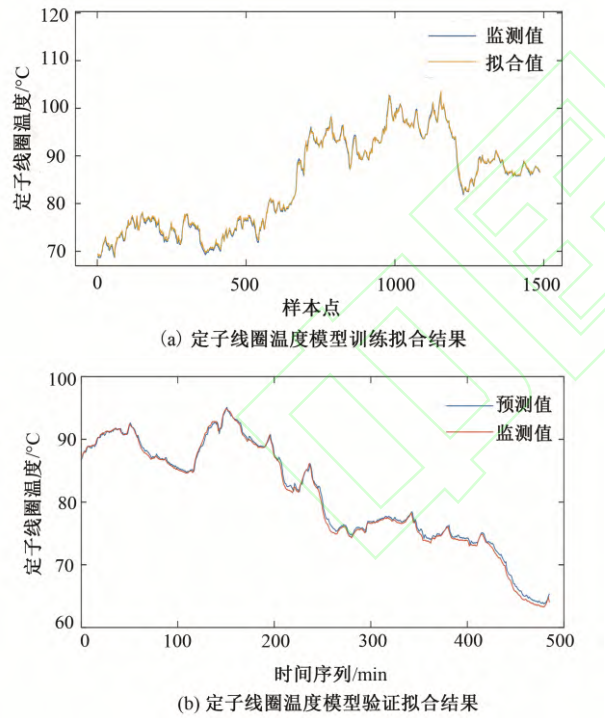


图 5 定子线圈温度线性支持向量回归拟合效果

Fig.5 Fitting effect of stator coil temperature with linear support vector machine

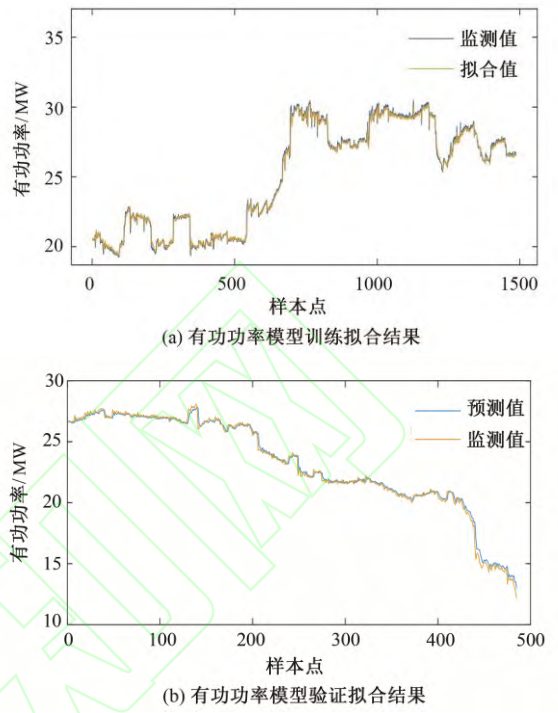


图 6 有功功率线性支持向量回归拟合效果

Fig.6 Fitting effect of active power linear support vector machine

表 3 拟合模型性能指标

Tab.3 Performance index of fitting model

性能指标 / 拟合模型类别	RMSPE/%	MAPE/%	R^2
定子线圈温度训练模型	0.548	0.424	1.00
定子线圈温度验证模型	0.414	0.262	0.99
有功功率训练模型	2.39	1.44	0.99
有功功率验证模型	1.47	0.945	0.99

从表 3、图 5 和图 6 中可以看出算法的 RMSPE 和 MAPE 均不超过 3%，表明该模型训练过程误差小，拟合程度好、具有较高的预测精度，可满足水电机组的预测要求。每个拟合模型类别的拟合优度 R^2 都达到 0.99

以上，很接近 1，说明模型的拟合效果很好。通过分析三个评价性能指标，结果表明，运用 Linear SVR 模型能够高度还原原始数据信息，对水电机组缺失数据进行恢复能达到很好的效果。

4 结论

本文针对水电机组海量监测数据中蕴含的异常数据，提出了一种利用 K-means 聚类方法辨识错误数据，Linear SVR 恢复缺失数据的方法。案例采用了有功功率和定子线圈温度的监测数据验证了所提方法的有效性和可行性，可得以下结论。

（1）K-means 聚类方法只有一个聚类参数可调，计算简单。当聚类中心为 3 个时，使用 K-means 聚类方法辨识错误数据的准确率达 98.5%，对辨识出的异常数据进行清洗，可以获得高质量的数据。

（2）运用 Linear SVR 模型对水电机组缺失的定子线圈温度和有功功率数据进行恢复，得到数据训练模型和验证模型的 RMSPE 和 MAPE 均不超过 3%，其拟合优度 R^2 均在 0.99 以上，说明拟合的准确性高，预测精度高，恢复的数据接近真实数据。

参考文献

[1] 孟建良,刘德超.一种基于 Spark 和聚类分析的辨识电力系统不良数据新方法[J].电力系统保护与控制,2016,44(3):85-91.

[2] 方睿,董树锋,唐坤杰,等.基于最大测点正常率与 GPU 并行加速的不良数据辨识方法[J].电力系统自动化,2019,43(16):86-93+115.

[3] 胡阳,乔依林.基于置信等效边界模型的风功率数据清洗方法[J].电力系统自动化,2018,42(15):18-23+149.

[4] WANG C,MU G,CAO Y. A Method for Cleaning Power Grid Operation Data Based on Spatio-temporal Correlation Constraints[J]. IEEE Access,2020,8:224741-224749.

[5] YU L,SUN W,YANG Z.Bad data identification method of power system based on coevolutionary symmetric loss function[J]. International Transactions on Electrical Energy Systems,2020, 31（11）.

[6] HAO S, TANG N; LI G L ,et al. A Novel Cost-Based Model for Data Repairing[J] . IEEE Transactions on Knowledge and Data Engineering,2017,29(4):727-742.

[7] 钟建伟,刘佳芳,倪俊,等.改进新息图法在不良数据检测与辨识中的应用[J].电力系统及其自动化学报,2018,30(9):83-88.

[8] LIN Y,Abur A. A Highly Efficient Bad Data Identification Approach for Very Large Scale Power Systems[J]. IEEE Transactions on Power Systems,2018,33(6):5979-5989.

[9] 洪梓铭,王忠军.基于优先级分配策略的电网信息系统数据恢复方法[J].电子设计工程,2021,29(9):151-154+159.

[10] 王方超,吕志平,吕浩,等. 基于数据驱动的RegEM 算法在 GPS 坐标时间序列插值中的应用[C]//中国卫星导航年会,2019.

[11] 谢智颖,何原荣,李清泉.基于时空相关性的公交大数据清洗[J/OL].计算机工程与应用,2021[2021-12-07].http://kns.cnki.net/kcms/detail/11.2127.TP.20210430.1401.006.html.

[12] 张帅,杨晶显,刘继春,等.基于多尺度时序建模与估计的电力负荷数据恢复[J].电工技术学报,2020,35(13):2736-2746.

[13] FAN G,Li J,Hao H. Lost data recovery for structural health monitoring based on convolutional neural networks[J]. Structural Control and Health Monitoring. 2019,26（10）.

[14] Li XN,Cai Y,Zhu WH. Power Data Cleaning Method Based on Isolation Forest and LSTM Neural Network[C]//International Conference on Cloud Computing and Security. 2018:539-550.

[15] 王子馨,胡俊杰,刘宝柱.基于长短期记忆网络的电力系统量测缺失数据恢复方法[J].电力建设,2021,42(5):1-8.

[16] 刘歌,芮国胜,田文飚.基于双重正则矩阵分解的缺失数据恢复[J].系统工程与电子技术,2021,43(5):1191-1