引用格式：马良玉，韩立凯，翟亮亮.基于iForest-DBSCAN-RF与优化CATBoost的风电机组齿轮箱油温异常预警[J].电力科学与技术学报,2025,40(4):193-204.

Citation：MA Liangyu, HAN Likai, ZHAI Liangliang. Anomaly warning of gearbox oil temperature in wind power unit based on iForest-DBSCAN-RF and optimized CATBoost[J]. Journal of Electric Power Science and Technology,2025,40(4):193-204.

基于 iForest-DBSCAN-RF 与优化 CATBoost 的风电机组齿轮箱油温异常预警

马良玉[1,2]，韩立凯[1]，翟亮亮[1]

（1.华北电力大学控制与计算机工程学院，河北 保定 071003；2.保定市综合能源系统状态检测与优化调控重点实验室，河北 保定 071003）

摘　要：数据清洗、特征选择和预测模型建立是基于数据采集与监视控制系统（supervisory control and data acquisition，SCADA）数据，实现风电机组异常状态预警不可缺少的重要环节。先结合孤立森林（isolation forest，iForest）和基于密度的空间聚类（density-based spatial clustering of applications with noise，DBSCAN）算法对SCADA数据异常点进行有效清洗，并采用随机森林算法（random forests，RF）与Person相关系数法优选模型输入参数；再进而基于Optuna优化的类别提升树（categorical boosting，CATBoost）算法，建立风电机组正常工况齿轮箱油池温度的预测模型；然后采用滑动窗方法，构建状态评价指标，并使用区间估计理论确定油温异常状态判别的临界阈值；实现油温异常预警；最后利用某风电机组SCADA系统油温异常的真实历史故障数据进行检验，验证了该方法的有效性。

关　键　词：风电机组；故障预警；孤立森林；基于密度的空间聚类算法；随机森林

DOI：10.19781/j.issn.1673-9140.2025.04.018　　**中图分类号**：TM315　　**文章编号**：1673-9140(2025)04-0193-12

Anomaly warning of gearbox oil temperature in wind turbine generator system based on iForest-DBSCAN-RF and optimized CATBoost

MA Liangyu[1,2], HAN Likai[1], ZHAI Liangliang[1]

（1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China;
2. Baoding Key Laboratory of State Detection and Optimization for Integrated Energy System, Baoding 071003, China）

Abstract: Data cleaning, feature selection, and the establishment of a prediction model are indispensable steps to realize anomaly warning of the wind turbine generator system based on data collection and supervisory control and data acquisition (SCADA). Firstly, isolated forest (iForest) and density-based spatial clustering of applications with noise (DBSCAN) algorithm are combined to effectively clean the data outliers of SCADA, and random forest (RF) and Pearson correlation coefficient method are used to optimize the input parameters of the model. Based on the categorical boosting (CATBoost) algorithm optimized by Optuna, a prediction model of gearbox oil pool temperature in the wind turbine generator system under normal operating conditions is established. Then, the state evaluation index is constructed with the sliding window method, and the interval estimation theory is employed to determine its critical threshold for anomaly discrimination of oil temperature. Finally, the anomaly warning of oil temperature is realized. The real historical fault data of oil temperature anomaly in the SCADA system of the wind turbine generator system are used to verify the effectiveness of the method.

Key words: wind turbine generator system; fault warning; iForest; DBSCAN; RF

收稿日期：2024-07-04；**修回日期**：2024-11-26

基金项目：河北省中央引导地方科技发展资金项目（226Z2103G）

通信作者：马良玉（1972—），男，博士，教授，主要从事智能技术在电站建模，优化控制与故障诊断中的应用等方面研究，E-mail：maliangyu@ncepu.edu.cn

当前，作为一种重要的清洁能源利用方式，风力发电的技术日趋成熟，其装机容量逐年攀升[1]。齿轮箱作为风电机组的关键装置，在长期运行过程中，其负载变化、设备异常等会导致油池温度升高。油池温度异常升高是齿轮箱故障的一个重要前兆，油温异常监测对保障风力发电系统的安全稳定运行起了十分关键的作用[2]。

目前，针对风电机组齿轮箱故障的研究主要有2类：一类是利用人工智能算法对数据采集与安全监控系统（supervisory control and data acquisition，SCADA）的正常运行数据进行学习，识别早期异常工况，实现故障的提前预警；另一类则是利用已发生故障的历史数据建立辨识模型，实现故障类型的正确判别。对及时发现早期故障、协助运维及时处理、降低运维成本而言，第一类研究更有价值[3]。

风电机组异常预警一般需要建立正常工况性能参数的预测模型，根据异常或故障发生时参数偏离正常值的程度来及时预警。因此，其对性能预测模型的精度要求较高[4]。而数据清洗、特征选择和模型优化则是基于SCADA数据，实现风电机组异常状态预警不可缺少的重要环节。已有诸多学者针对各环节开展了富有成效的研究。

在SCADA记录的原始数据中，存在停机、限负荷和其他各种原因导致的异常偏离点。有效的数据清洗是保障正常工况预测模型建立的重要前提之一。由于异常点类型较多，单一算法往往无法取得理想效果。现有研究多采用2种以上算法对原始数据进行处理。文献[5]将四分位法与孤立森林算法结合，对异常数据进行清洗，有效去除离群点和限负荷点，但其清洗效果仍有提升空间。文献[6]则提出了Box-Cox变换与Bin算法相结合的方法，实现异常数据清洗，但该方法在一定程度上改变了数据的分布形式，会对清洗效果带来一定影响。

原始数据包含近百个变量，其数据维度高。预测模型建立时，应根据被预测变量选取与之相关性高的关键特征变量作为模型输入，以降低参数冗余，降低建模数据维度，才能保证模型的精度不受影响[7]。常见的特征变量选择法有轻量梯度提升机（light gradient boosting machine，LightGBM）、随机森林（random forests，RF）算法等。文献[8]利用最大互信息系数法，从各变量与模型输出特征的相关性来进行特征选择，但其未考虑输入特征之间的相关性。文献[9]针对风机故障诊断，提出了一种多算法组合的分段式特征选择方法，有效地实现了对高维度数据的降维，其泛化能力也较强，但该方法较为繁琐且未考虑输入特征之间的相关性。

在模型类型选择方面，也涌现了诸多研究成果。他们常采用多模型融合来提高预测精度。文献[10]对长短期记忆（long short-term memory，LSTM）网络进行改进，并将其用于齿轮箱故障预警和电价预测，取得了较好效果，但其模型训练时间较长。文献[11]采用改进的随机森林算法预测齿轮箱油温，模型训练时间较短但精度不够高。文献[12]提出了一种基于Stacking的多模型融合方法，优选各种机器学习算法以提高模型预测精度，但模型结构较复杂，且数据量过少可能导致出现过拟合。文献[13-14]分别利用改进飞蛾火焰（mirage search optimization，MSO）算法优化的LSTM和对输入数据优化的自组织核回归模型进行齿轮箱油温预警，模型预测精度较高，但这些方法均采用样本残差作为状态评价指标，易发生误报警。

本文在现有研究基础上，建立一种基于类别提升树（categorical boosting，CATBoost）算法的齿轮箱油池温度异常预警模型，并采用中国河北省某风电场油温异常故障的真实历史数据来验证该方法的有效性。首先，利用孤立森林算法与基于密度的空间聚类（density-based spatial clustering of applications with noise，DBSCAN）算法，依次清洗原始数据中大量连续的离群点和剩余的离散噪声点，获得高质量的建模数据；其次，将齿轮箱油池温度作为预测模型输出，结合随机森林与Person相关系数法，分别考虑输入特征对输出特征的重要性、输入特征间的相关性，确定模型输入参数；再次，使用Optuna优化的CATBoost算法，建立油温预测模型；最后，以所建预测模型为基础，基于滑动窗口法构建油温状态指标，采用区间估计方法计算异常状态判别阈值，根据实时油温与模型预测偏差，实现齿轮箱油温异常的判别[15]。

1 数据清洗

1.1 手动处理

选取中国河北省某风电场一台风电机组某全年的运行数据进行建模，其采样周期为1 min，共计525 780组数据。该风机的切入、切出风速均为3.25 m/s，其额定功率为1.5 MW，其风速-发电机转速-功率原始数据分布如图1所示。

风电机组正常性能预测模型可预测对机组正常运行中的某些参数，但不包含启动和停机工况。

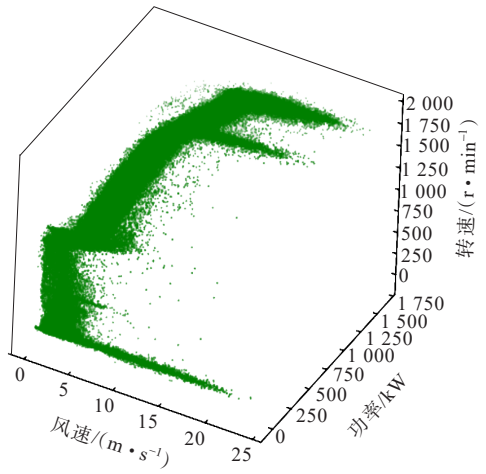


图 1　风速-发电机转速-功率原始数据分布

Figure 1　Wind speed-generator speed-raw power data distribution

通讯中断等意外导致的 SCADA 数据的部分空缺会影响建模。根据风机运行原理，当风速小于等于切入风速时，风电机组为待机或启动状态；当风速大于切出风速时，风电机组处于停机状态。当风电机组停机时，发电机有功功率应不超过 0 kW。因此，应删除风速小于 3 m/s、风速大于 25 m/s、发电机有功功率小于等于 0 kW 的数据及其空缺值。原始数据经处理后，得到 347 804 组有效数据，其对应风速-发电机转速-功率手动处理后数据分布如图 2 所示。

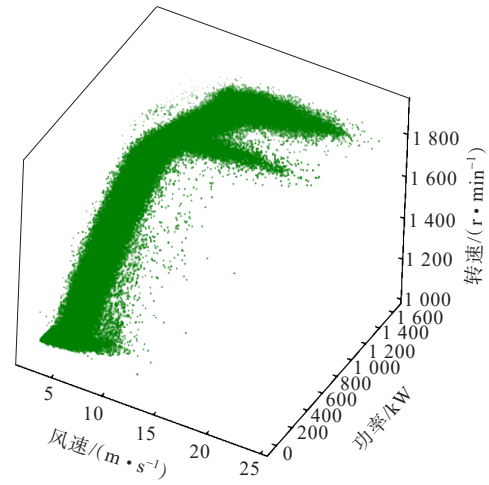


图 2　风速-发电机转速-功率手动处理后数据分布

Figure 2　Wind speed-generator speed-power data distribution after manual treatment

1.2　算法清洗

从图 2 中可以看出，处理后的风电机组数据量大且分布不均，存在大量连续且稀疏的异常点需要剔除。孤立森林（isolated forest，iForest）算法不依赖类似距离、密度等指标来表征一个样本与其他样本的差异，而是直接刻画样本间疏离程度，该算法简单、高效[16]。因此，本文采用 iForest 算法对数据进行清洗，去除规模庞大的连续异常点。

iForest 原理示意如图 3 所示。从图 3 中可以看出，该算法先从待清洗数据集中随机选取 m 个样本作为子集，并将其放入树的根节点，并随机选择一个特征进行分割，生成切割点；再依据这个切割点，将数据空间划分为 2 个子空间，并递归地选择特征和分割数据，构建一棵二叉树，直到满足终止条件；然后，重复此过程，生成 n 棵孤立树，形成孤立森林；最后，将数据点放入孤立森林中，计算其从根节点到叶节点的路径长度，并根据路径长度赋予异常评分。路径越短，图中颜色越深，这些数据点是异常的可能性越大[17]。

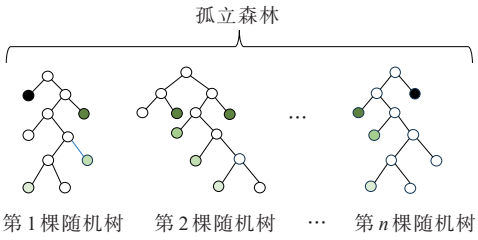


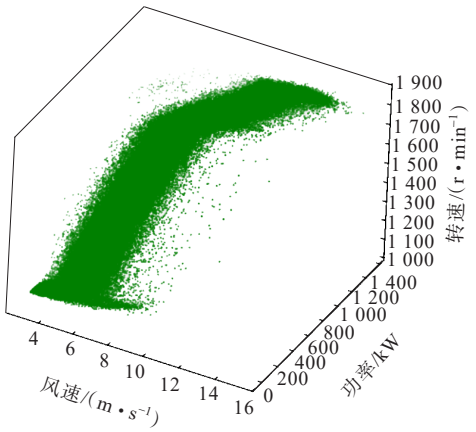
图 3　iForest 原理示意

Figure 3　Schematic diagram of iForest principle

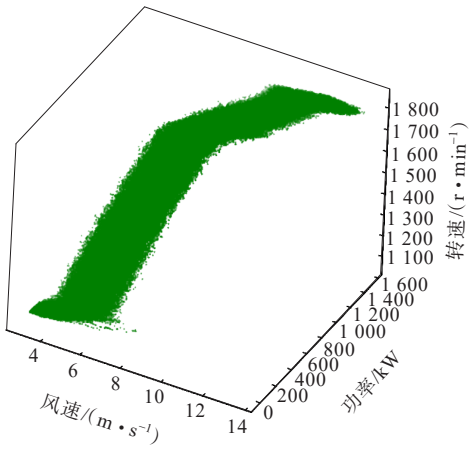
iForest 算法的关键参数是基评估器的数量，该参数值越大，算法准确度越高。但如果该参数值过大，则可能导致计算时间延长，且过大的该参数值对算法准确度提升能力有限[18]。因此，经试验寻优，最终确定该参数值为 200。经过 iForest 一、二次清洗后的数据集如图 4 所示。从图 4(a)中可以看出，iForest 算法能清洗大量的连续异常点，但清洗完成后数据中仍存在少量的噪声数据，应继续进行二次数据清洗。

作为一种基于密度的聚类算法，DBSCAN 的核心思想是，如果一个点在给定半径内邻域内有足够数量的其他点，则该点为核心点，这些点可以一起形成簇；否则，该点为噪声点。对每个数据点重复此检验过程，就可以完成对数据的异常点的筛选[19]。一次清洗后剩余的异常数据大多为离散点，这样的点与周围数据的距离较远，可以通过 DBSCAN 有效地清除剩余的少量噪声点。

DBSCAN 算法有 2 个主要参数：邻域半径和最小点数。邻域半径定义了点的邻域范围，而最小点数则决定了一个点是否为核心点。经多次试验调



(a)iForest算法一次清洗结果



(b) DBSCAN算法二次清洗结果

图4 算法清洗后的数据分布

Figure 4 Data distribution after algorithm cleaning

试,选择邻域半径为0.42,最小点数为5。最终,经DBSCAN算法二次清洗后,得到290 814组正常工况下的数据,其风速-发电机转速-功率数据分布如图4(b)所示。

2 特征选择

SCADA系统包含风电机组的上百种运行参数。在建立正常工况性能预测模型时,过多的输入特征会导致模型结构庞大、训练时间过长或过出现拟合的情况。因此,应对输入变量进行特征选择[20]。特征选择通常利用智能算法计算输入特征的重要度,并对输出特征的重要度排序。但根据重要度选择输入特征的方法仍具有一定主观性,且这种方法没有考虑所选输入特征间的相关性。因此,本文先根据重要度的计算结果,优先选定重要性程度高且稳定的特征量作为输入,再考虑剩余特征间的相关性,据此进行二次特征选择。

齿轮箱油池温度可直观地反应齿轮箱运行是否正常,故将其作为本文的监测对象,确定其为模型输出变量。由于其余特征对齿轮箱油温的重要性不尽相同,故使用RF算法计算其余特征与油池温度的重要性并排序。为克服模型训练结果的随机性,重复4次该试验,列出重要性排名前21的特征,重要性指标试验结果见表1。对这4次试验中的21个特征的重要性指标求均值并排序,结果如图5所示。

表1 重要性指标试验结果

Table 1 Experimental results of importance index

试验	齿轮箱油路入口温度/℃	主轴转速/(r•min⁻¹)	齿轮箱高速轴驱动端轴承温度/℃	轮毂内温度/℃	齿轮箱油路入口油压/MPa	发电机转速/(r·min⁻¹)	齿轮箱高速轴非驱动端轴承温度/℃
1	0.135 7	0.133 8	0.082 3	0.044 8	0.036 7	0.031 8	0.024 3
2	0.137 2	0.134 1	0.088 3	0.045 3	0.035 2	0.032 1	0.023 8
3	0.134 5	0.134 2	0.084 5	0.046 6	0.035 4	0.033 0	0.024 8
4	0.136 4	0.136 1	0.087 5	0.045 5	0.034 7	0.032 5	0.023 9
试验	发电机驱动端轴承温度/℃	发电机非驱动端轴承温度/℃	舱内温度/℃	发电机定子U相线圈温度/℃	发电机定子V相线圈温度/℃	机舱控制柜内温度/℃	2#桨电机温度/℃
1	0.021 8	0.019 2	0.018 8	0.018 3	0.017 8	0.017 1	0.015 8
2	0.019 0	0.021 6	0.019 0	0.018 2	0.018 9	0.017 6	0.015 3
3	0.022 6	0.018 7	0.018 3	0.018 2	0.017 8	0.018 4	0.015 3
4	0.021 9	0.018 4	0.019 3	0.018 6	0.018 3	0.017 9	0.015 4
试验	3#桨电机温度/℃	1#桨电机温度/℃	PLC风机状态(10 min)	网侧频率/Hz	机舱角度/(°)	偏航状态	瞬时设定扭矩/(N·m)
1	0.015 5	0.014 8	0.014 5	0.013 8	0.013 5	0.013 4	0.010 5
2	0.014 6	0.015 0	0.015 2	0.013 1	0.013 6	0.013 8	0.013 7
3	0.014 7	0.015 6	0.014 7	0.012 9	0.013 7	0.013 9	0.013 1
4	0.014 9	0.014 7	0.015 2	0.013 1	0.013 7	0.013 7	0.013 5

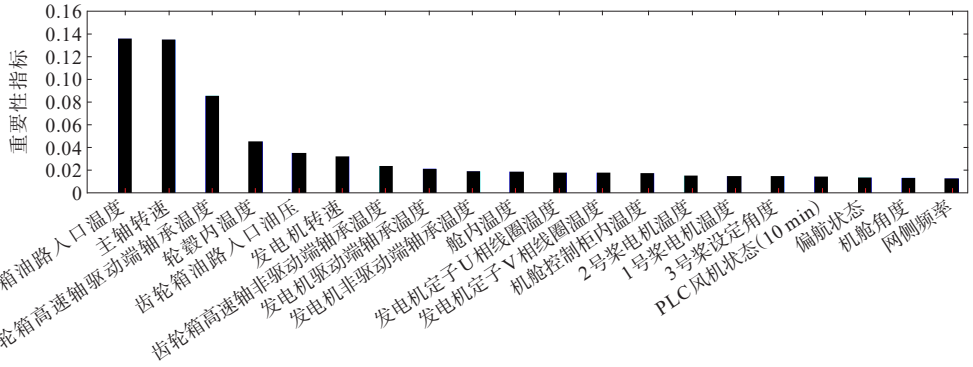


图 5　4 次试验的重要性指标均值

Figure 5　Mean value of importance index of four experiments

由表 1 可知，这 4 次试验得到的特征重要性指标前 7 位排名均稳定不变。从机理上看，齿轮箱油路入口温度与油池温度密切相关；当齿轮箱入口油压减小时，润滑油量减小，冷却效果变差，轴承温度升高，发电机和主轴转速增大均会使齿轮间摩擦生热增加，导致齿轮箱油温升高；齿轮箱油温升高导致润滑失效，此时的齿轮箱高速轴轴承、主轴等机械部件的摩擦力均增大，其温度也会相应升高。同样地，这些机械部件温度升高会将热量传递至齿轮箱油池与轮毂内部，导致油温和轮毂内部温度升高。依据算法所选的这前 7 个特征参数确实与齿轮箱油池温度具有很高的相关性。因此，优先选定这前 7 个特征为模型的输入参数。

利用 Person 相关系数法对一次特征选择后剩余的 14 个特征进行相关性分析，其计算式为

$$\rho_{X,Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-\left(E(X)\right)^2}\sqrt{E(Y^2)-\left(E(Y)\right)^2}} \tag{1}$$

式中，X，Y 均为待计算的随机变量；$E(X)$ 为随机变量 X 的数学期望；$\rho_{X,Y}$ 为 X，Y 的 Person 相关系数。

求出各特征与其他特征之间的相关系数，绘制特征相关性热力图，结果如图 6 所示。将各特征所对应的 14 个相关系数取绝对值并求和，最终得到相关性指标排序，结果见表 2。表 2 中相关性排名前 3 的特征在第一步的 4 次试验中均位于前 14 名的，其对输出特征的重要性影响较高，这些特征与其他的特征相关性最强，对其他参数具有较好的替代性。因此，综合考虑后，补充选取该 3 个特征作为模型的输入参数。

综上所述，最终选择的模型输入参数：齿轮箱油路入口温度、主轴转速、齿轮箱高速轴驱动端轴承温度、轮毂内温度、齿轮箱油路入口油压、发电机转速、齿轮箱高速轴非驱动端轴承温度、发电机非驱动端轴承温度、机舱控制柜内温度、舱内温度，输出特征参数为齿轮箱油池温度。

为验证本文特征二次选择方法的合理性，采用了 3 种单一特征选择方法：RF 特征重要性排序算法、序列前向选择算法（sequence forward selection，SFS）和特征递归消除算法（RFE）与本文方法进行对比[21-22]，并以齿轮箱油池温度为输出，采用这 3 种算法分别选择 7、10、13、16、19 个输入特征，利用与文 3 中相同的数据集和 CATBoost 算法进行建模并进行 100 次参数寻优。最终，比较各模型与本文建立的 10 参数模型针对测试集的预测性能，发电

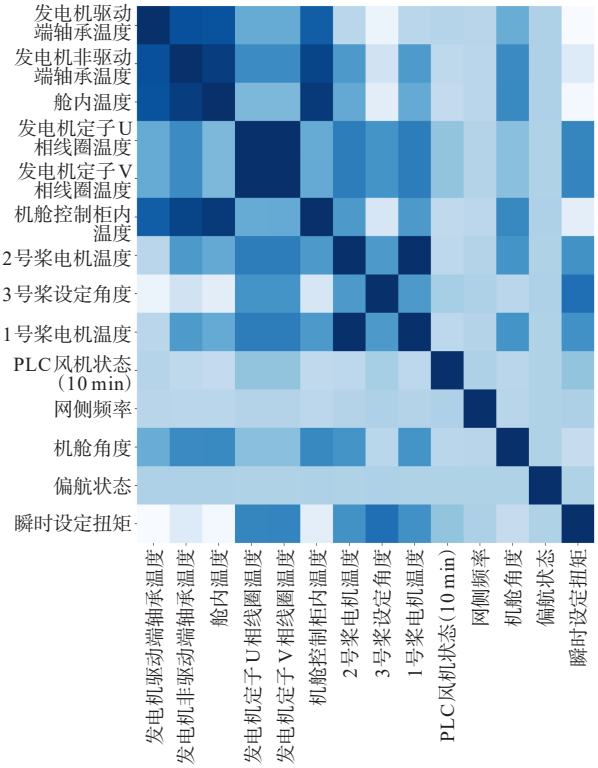


图 6　特征相关性热力图

Figure 6　Feature correlation heat map

表 2 相关性指标排序

Table 2 Correlation index ranking

特征名称	单位	相关性指标
发电机非驱动端轴承温度	℃	6.096 1
机舱控制柜内温度	℃	5.727 7
舱内温度	℃	5.514 1
2号桨电机温度	℃	5.112 9
1号桨电机温度	℃	5.109 2
发电机定子V相线圈温度	℃	4.983 5
发电机定子U相线圈温度	℃	4.981 3
发电机驱动端轴承温度	℃	4.788 8
机舱角度	°	3.995 4
3号桨设定角度	°	3.985 7
PLC风机状态(10 min)		1.759 2
网侧频率	Hz	1.294 1
偏航状态		1.023 6

机非驱动端轴承温度为6.096 1 ℃,机舱控制柜内温度为5.727 7 ℃,舱内温度为5.514 1 ℃,2#桨电机温度为5.112 9 ℃,1#桨电机温度为5.109 2 ℃,发电机定子V相线圈温度为4.983 5 ℃,发电机定子U相线圈温度为4.913 8 ℃,发电机驱动端轴承温度为4.788 8 ℃,机舱角度为3.995 4°,3#桨设定角度为3.985 7°,每隔10 min的PLC风机状态指数为1.759 2,网测频率为1.294 1 Hz,偏航状态指数为1.023 6。

模型性能评价使用拟合优度R^2、均方误差I_{MSE}(mean squared error,MSE)、平均绝对误差绝对值I_{MAE}(mean absolute error,MAE)3种指标[23],其计算公式分别为

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y(i)-\hat{y}(i)\right)^2}{\sum_{i=1}^{n}\left(y(i)-\bar{y}(i)\right)^2} \tag{2}$$

$$I_{\mathrm{MSE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\left(y(i)-\hat{y}(i)\right)\right|^2 \tag{3}$$

$$I_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\left(y(i)-\hat{y}(i)\right)\right| \tag{4}$$

式(2)~(4)中,$y(i)$为数据的实际值;$\hat{y}(i)$为模型预测值;n为数据的个数。

不同特征选择方法对应模型性能指标对比见表3。由表3可知,利用RF算法和SFS算法选择13个特征,RFE算法选择16个特征时,所建各模型预测性能达到最佳。当特征数量减少时,这3种算法构建的模型预测性能均显著下降;而当特征数量进一步增加时,模型预测效果不仅未提升,反而呈现下降趋势。这种现象在一定程度上是特征之间的共线性所导致的。采用本文方法选取的10个特征所建立的模型,其预测性能明显优于采用RF、SFS、RFE 3种算法选择的10个特征对应的模型的。与RF算法选择的13个特征的最佳模型相比,本文10参数模型的R^2降低0.019 1%,MSE提高0.369 1%,但MAE降低0.698 9%,整体预测性能差异较小。与SFS算法选择的13个特征建立的最佳模型相比,本文模型在R^2和MSE上表现更优,但MAE提高4.057 1%。与RFE算法选择的16个特征建立的最佳模型相比,本文模型在所有性能评价指标上均表现最佳。

表 3 不同特征选择方法对应的模型性能指标对比

Table 3 Comparison of model performance indicators corresponding to different feature selection methods

特征选择方法	特征数量	R^2	I_{MSE}	I_{MAE}
RF	7	0.988 308	0.536 108	0.546 804
RF	10	0.989 411	0.485 526	0.492 229
RF	13	0.991 357	0.401 365	0.456 134
RF	16	0.989 824	0.466 576	0.468 504
RF	19	0.988 936	0.507 338	0.471 796
SFS	7	0.987 768	0.560 836	0.401 497
SFS	10	0.988 783	0.514 355	0.431 480
SFS	13	0.990 933	0.415 728	0.408 574
SFS	16	0.989 814	0.467 027	0.406 861
SFS	19	0.988 491	0.527 733	0.432 691
RFE	7	0.986 484	0.619 745	0.572 805
RFE	10	0.986 844	0.590 082	0.575 433
RFE	13	0.988 138	0.543 907	0.517 017
RFE	16	0.989 086	0.500 410	0.478 599
RFE	19	0.987 424	0.576 640	0.495 484
RF+Person	10	0.991 166	0.405 056	0.449 145

综上所述,相比于RF算法通过重要性选取特征、SFS算法逐步增加重要特征,RFE算法根据训练结果逐步剔除特征的3种方法,本文方法在有效降低输入特征维度的同时,较好地保证了模型的预测性能。此外,RF、SFS、RFE这3类算法的最优模型所对应的特征数量均未达到19,尽管19个特征中包含了所有最优的13或16个特征,但由于冗余共线性特征的影响,模型性能反而下降。本文的二次特征选择在降维的同时消除了共线性特征,减少了特征冗余,进一步提升了模型的预测性能。

3　齿轮箱油温预测模型的建立

前文选择的风电机组数据多为温度、转速等变化范围较大的数据。传统的梯度提升算法在工作时，弱学习器模型均在同一完整训练集上训练，再不断提升成强学习器。如果训练集和测试集的数据分布不一致，模型就会在训练集产生过拟合，在测试集上表现不佳，发生预测偏移问题。

CATBoost 基于梯度提升算法，通过迭代地构建决策树来逐步减小预测残差。每个决策树都基于前一个模型的误差进行训练，以此来提高模型的整体性能[24]。梯度提升算法原理如图 7 所示。在图 7 中，$p_k(x)$ 为各决策树预测值，其中 $k=1,2,3,\cdots,N$；y 为训练数据集的实际值。

CATBoost 在训练阶段对数据特征执行处理，采用改进的排序提升方法来避免预测偏移问题，并运用对称的决策树架构来管理叶节点的数量[25]。此外，还引入了优先级或称为先验值的权重系数。设样本 $\sigma=(\sigma_1,\sigma_2,\cdots,\sigma_n)$ 是随机排列顺序，样本 σ_i 的第 j 个特征 $Q_{\sigma_i,j}$ 为数据特征，则有[25]

$$Q_{\sigma_i,j}=\frac{\sum_{k=1}^{j-1}\left[Q_{\sigma_k,j}=Q_{\sigma_i,j}\right]Z_{\sigma_k}+\alpha P}{\sum_{k=1}^{i-1}\left[Q_{\sigma_k,j}=Q_{\sigma_i,j}\right]+\alpha} \tag{5}$$

式中，Q 为样本的数据特征，Z 为样本的特征值；$[\cdot]$ 为指示函数，其在某个元素属于特定子集时取值为 1，否则为 0；P 为先验值，对每一个回归任务，将其特征值的平均值作为先验值；α 为优先级的权重系数，降低低频次特征带来的影响。

由于 CATBoost 模型具有多个超参数，为提高模型预测准确性，本文使用 Optuna 自动寻优框架针对 CATBoost 模型中的树深度、数值变量的分箱个数、正则化系数、学习率、分支数目这 5 个超参数进行寻优[26]。

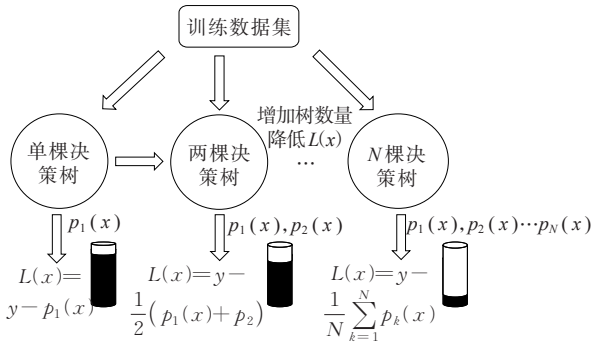


图 7　梯度提升算法原理示意

Figure 7　Schematic diagram of gradient boosting algorithm principle

选择该风电机组正常工况的 50 000 组数据，前 40 000 组数据用于模型训练，后 10 000 组数据用于模型测试和性能评价。

使用 Optuna 框架对模型超参数寻优时，应对 Optuna 框架进行参数选择，选择目标函数迭代次数为 100，损失函数选择均方根误差（mean square error，RMSE）。由于采样器的选择对模型性能有较大影响，分别采用树结构的 Parzen 估计器（tree structured Parzen estimator，TPE）算法、进化策略算法（covariance matrix adaptation evolutionary strategy，CmaEs）、随机搜索（random search，RS）3 类采样器对 CATBoost 模型进行寻优。为克服寻优的随机性，对每个采样器进行 5 次试验，选取效果最好的作为寻优结果，结果见表 4。

表 4　不同采样器模型寻优结果对比

Table 4　Comparison of model optimization results of different samplers

采样器	R^2	I_{MSE}	I_{MAE}
TPE	0.991 166	0.405 056	0.449 145
CmaEs	0.991 032	0.410 342	0.456 527
Random	0.988 605	0.522 494	0.488 442

对比评价指标可知，采用 TPE 采样器寻优的模型均方误差及平均绝对误差均为 3 类采样器中最小值，并且其拟合优度最高。

为展示 Optuna 寻优下的 CATBoost 模型性能的优越性，这里选择集成算法中的 XGBoost 及 LightGBM 算法作对比，采样器均使用 TPE 算法，迭代次数为 100，所建立模型预测性能指标见表 5。

根据这 3 个模型的性能评价指标对比可知，CatBoost 模型的 R^2 和 MSE 均为最优值，XGBoost 模型的 MAE 最佳，LightGBM 的拟合效果各指标均为最差。由于风电机组数据集庞大，拟合优度 R^2 要比平均绝对误差绝对值 MAE 更能反映模型的总体预测效果。因此，本文选择 CATBoost 作为预测模型，模型最优超参数见表 6。

表 5　模型预测性能指标

Table 5　Model prediction performance index

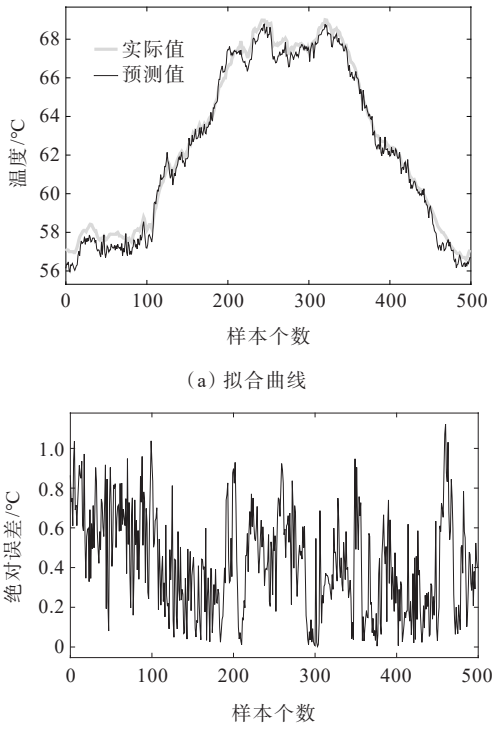
模型	R^2	I_{MSE}	I_{MAE}
CATBoost	0.991 166	0.405 056	0.449 145
XGBoost	0.989 773	0.478 347	0.420 420
LightGBM	0.982 472	0.592 073	0.539 188

表 6 模型最优超参数

Table 6 Optimal hyperparameters of models

树深度	正则化系数	学习率	分支数目	分箱个数
2	3	0.256	1 454	275

利用所得的最优CATBoost预测模型对测试集进行预测，部分数据的拟合结果如图8所示。根据拟合结果可见，绝大多数数据的拟合绝对误差低于0.6，表明模型的拟合精度较高。



(a) 拟合曲线

(b) 绝对误差

图 8 CATBoost模型预测结果

Figure 8 CATBoost model prediction results

4 基于滑动窗口的齿轮箱油温预警

4.1 状态评价指标及预警阈值计算

根据风电机组正常工况数据所建立的预测模型仅在机组正常工作时预测较为精准，而实际由于风电机组工作环境恶劣，当机组发生故障时，模型预测的准确度会呈现不稳定性，若仅根据单一时刻的工况偏离情况进行预测会加大误报警概率[27]。因此，采用基于滑动窗口的故障预警方法，将对于单一工况点的分析转变为对窗口内数据的整体分析。该时间滑动窗口的原理示意如图9所示。在图9中，y为数据实际值；S为窗口内数据的集合，在时间t_k处输入S_k；h为窗口宽度；m为时间增量。

选取单个窗口内齿轮箱油温的实际值与模型

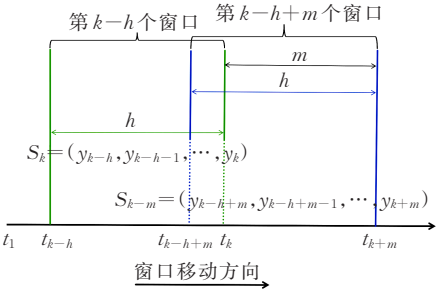


图 9 滑动窗口原理示意

Figure 9 Schematic diagram of sliding window principle

预测值之间的平均绝对误差作为故障预警的状态指标，其计算公式为[28]

$$M_{tk}=\frac{1}{h}\sum_{t=t_k-h}^{t_k}|T_t-T'_t| \tag{6}$$

式中，h为滑动窗口宽度；t为当前时刻；T_t为当前时刻实际温度；T'_t为当前时刻预测温度；M_{tk}为状态指标。

发生故障时，状态指标会不断增加，当其超越某一个值可视为故障的发生。因此，应合理设定状态指标的阈值。本文采用滑动窗口与区间估计相结合的方法设置阈值，即采用滑动窗口进行区间估计，选择置信上限作为阈值，其计算公式如下[28]：

$$N_t=\bar{M}+\frac{\sigma_M}{\sqrt{l}}t_\alpha(l-1)+3\sigma_{\mathrm{M}} \tag{7}$$

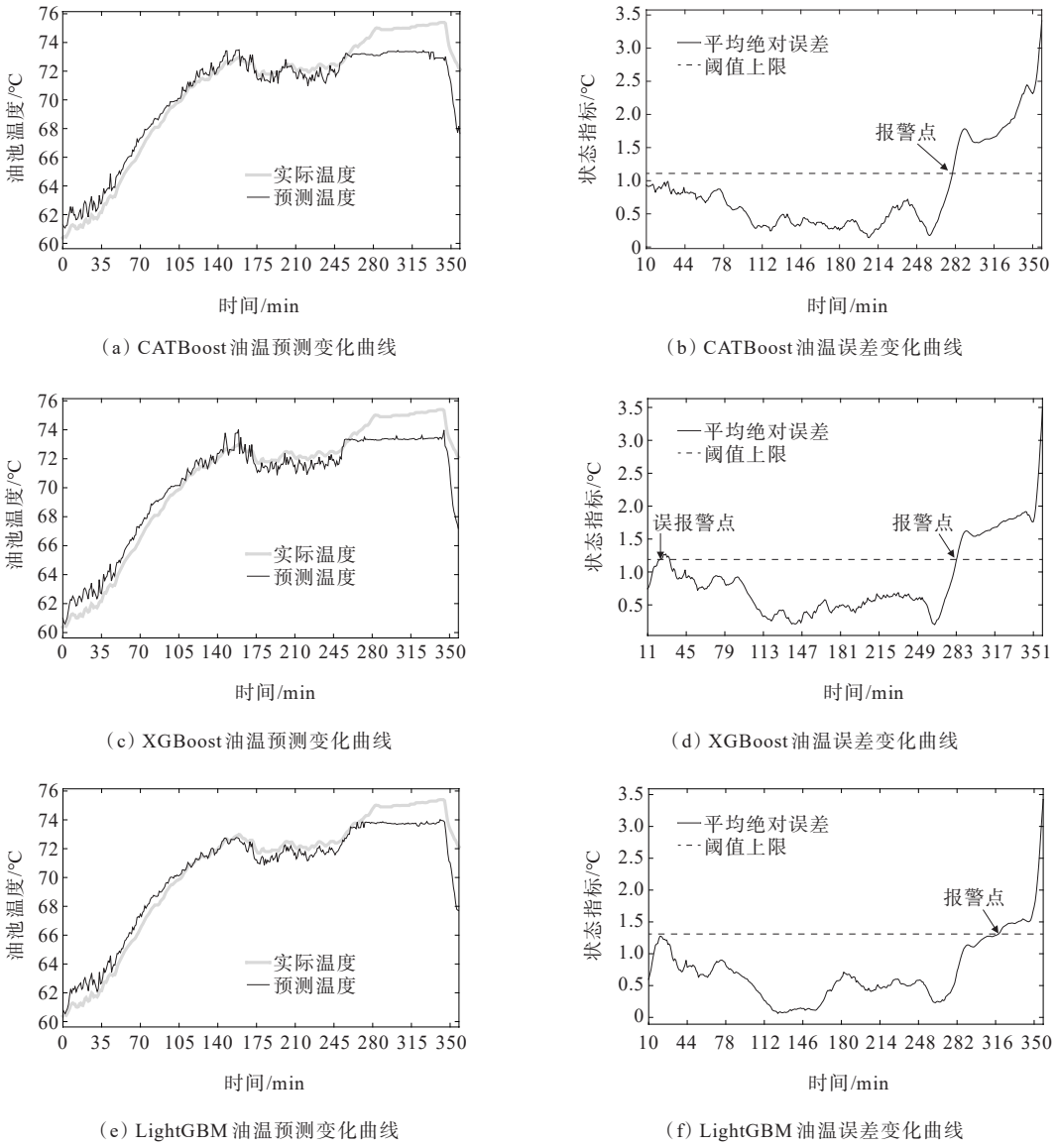
式中，N_t为状态指标的阈值；$\bar{M}$为正常工况下状态指标M_{tk}的平均值；l为滑动窗口个数；σ_{M}为状态指标的标准差；t_α为小概率值。

选取模型建立所使用的机组正常工况数据，经试验，当窗口宽度h选择为12，时间增量m设置为1，α设为0.000 5时，预警效果最好，经过式(7)计算得到状态指标阈值为1.11 ℃。

4.2 实例分析

选取该机组的原始数据，仅剔除数据中的空缺值、风机启动、停机以及风速小于3 m/s的运行数据，以保证故障预警的准确性。为比较本文模型与表6中其他2种模型在异常预警时的性能表现，同时利用所建XGBoost和LightGBM最优模型进行故障预警试验，通过该方法计算2种模型对应的状态指标阈值分别为1.19和1.31 ℃。

1) 实例1。查询该机组故障列表，发现2018年8月31日22时40分SCADA系统记录发生了“齿轮箱油池温度高于上限值”的故障。选取故障发生前6 h至故障记录时刻的数据进行故障预警试验，结果如图10所示。



（a）CATBoost 油温预测变化曲线　（b）CATBoost 油温误差变化曲线

（c）XGBoost 油温预测变化曲线　（d）XGBoost 油温误差变化曲线

（e）LightGBM 油温预测变化曲线　（f）LightGBM 油温误差变化曲线

图 10　故障预警结果（实例 1）

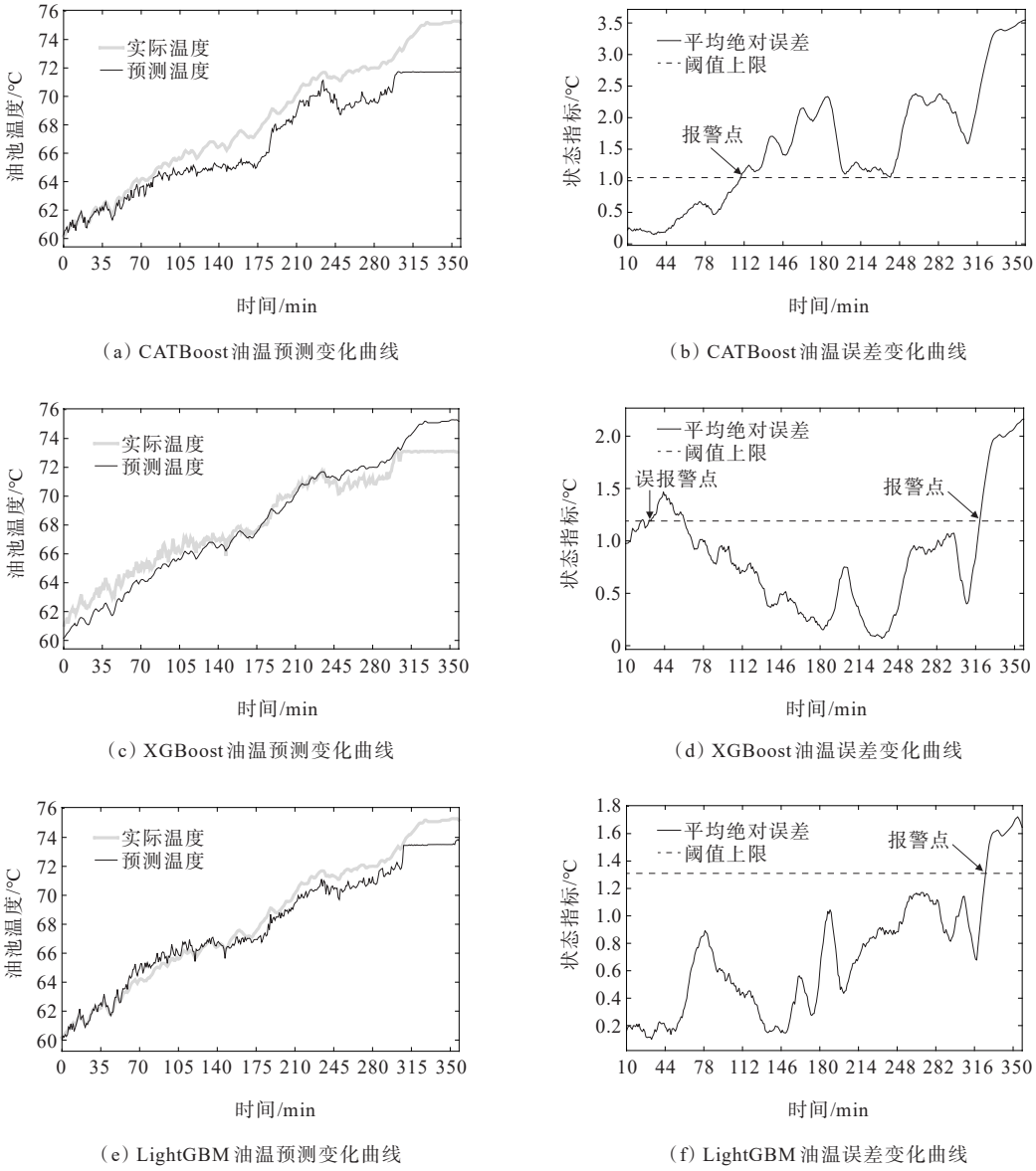
Figure 10　Fault warning result (case 1)

从图 10 中可以看出，CATBoost 模型的状态指标首次大约于第 280 个窗口超出阈值，发出报警提示，其所对应的时间为 2018 年 8 月 31 日 21 时 21 分，比 SCADA 记录的故障时刻提前 79 min 报警。而 XGBoost 及 LightGBM 模型分别比记录时刻提前预警 76 min 和 66 min。此外，XGBoost 大约于第 20 个窗口出现了短暂的超出阈值情况，导致了误报警。相比之下，CATBoost 模型的预警时间最早且没有误报警情况出现。

2）实例 2。查询该机组故障列表，发现同年 9 月 29 日 16 时 38 分 SCADA 系统记录再次发生“齿轮箱油池温度高于上限值”的故障。选取故障发生前 6 h 至故障记录时刻的数据进行试验，结果如图 11 所示。

从图 11 中可以看出，CATBoost 模型预警结果的状态指标首次大约于第 110 个窗口超出阈值，发出报警提示，其所对应的时间为 2018 年 9 月 29 日 12 时 30 分，提前报警了 248 min。而 XGBoost 及 LightGBM 模型分别比记录时刻提前预警 37 min 和 32 min。此外，XGBoost 模型在约第 35 个窗口时出现了短暂的超出阈值情况。相比之下，CATBoost 模型不仅预警时间显著早于其他模型，而且未出现误报警现象。

在实际生产中，故障及时报警可为工作人员提供更多的故障处理时间，而误报警则会干扰工作人员的判断，应尽力避免。由该试验可见，CATBoost 模型比 XGBoost 和 LightGBM 模型的异常预警效果更好。这一方面是由于各模型本身的



(a) CATBoost油温预测变化曲线

(b) CATBoost油温误差变化曲线

(c) XGBoost油温预测变化曲线

(d) XGBoost油温误差变化曲线

(e) LightGBM油温预测变化曲线

(f) LightGBM油温误差变化曲线

图 11 故障预警结果(实例 2)

Figure 11 Fault warning result(case 2)

数据学习机制存在差异,对同一数据集的预测效果存在不一致性。对故障发生前的异常数据,CATBoost模型的敏感度要优于XGBoost和LightGBM模型的,预测值较早偏离实际值,预警更及时;而对于距离故障发生时间较远的正常数据,CATBoost的状态指标波动较小,对正常数据的拟合能力较强,故没有发生误报警。另一方面,本文各预警模型的状态指标阈值均按式(7)计算,该阈值的大小同样会影响各模型预警的及时性和误报警率,而由于不同风电机组的数据差异影响,工程应用中可尝试不同的阈值计算方法或进行阈值微调,以获取最佳阈值,进而提早预警时间,减小误报警概率。

5 结语

针对风电机组齿轮箱异常工况预警,建立一种基于CATBoost算法的油池温度异常预警模型,以实现对风电机组齿轮箱故障的提前预警。本文主要结论如下:

1) 针对风电机组SCADA系统原始数据,依次采用iForest和DBSCAN方法对数据集中大量连续的异常离群点和剩余离散噪声数据进行针对性清洗。试验结果表明,该方法可有效清除异常数据点,获得高质量的机组正常行为建模数据。

2) 使用RF与Person相关系数相结合的方法,分别考虑模型输入特征对输出特征的重要性、输入

特征之间的相关性，实现了模型输入特征的 2 次选择。经试验对比，相比于仅考虑输入特征与输出特征相关性的特征选择方法，本文方法所选特征维度有所降低且保证了所建模型的预测精度。

3）建立了基于 CATBoost 的油温预测模型，利用滑动窗口模型构建状态评价指标，并基于区间估计理论设定阈值。实际故障数据验证，表明该方法所建模型预测准确度较高，能准确及时地对油温异常进行预警，并减少了误报警概率。

应指出，风电机组各类特征值受季节影响显著，采用全年数据训练能提升模型预测精度和预警准确性，但数据量过大会增加模型训练时间并可能导致过拟合。未来研究中可考虑采用迁移学习方法，提升模型的泛化能力与方法不同季节和气候下的故障预警准确性。

参考文献：

[1] 王中行，周元贵，张学广．基于人工智能算法的风电机组状态监测和故障诊断技术研究综述[J]．东北电力大学学报，2024，44(1)：42-51.
WANG Zhongxing，ZHOU Yuangui，ZHANG Xueguang. Review of artificial intelligence algorithms-based wind turbine condition monitoring and fault diagnosis techniques[J]. Journal of Northeast Electric Power University，2024，44(1)：42-51.

[2] 康杰．基于机器学习的风电机组齿轮箱故障预警分析[J]．电气技术与经济，2023(9)：156-158.
KANG Jie. Fault early warning analysis of wind turbine gearbox based on machine learning[J]. Electrical Equipment and Economy，2023(9)：156-158.

[3] 马永光，冯勇升．基于 IICEEMDAN-PCA-GRU 的风电机组齿轮箱故障预警方法研究[J]．太阳能学报，2023，44(4)：67-73.
MA Yongguang，FENG Yongsheng. Research on fault warning method of wind turbine gearbox based on iiceemdan-pca-gru[J]. Acta Energiae Solaris Sinica，2023，44(4)：67-73.

[4] 王旭涛，陈换过，陶瀚宇，等．基于迁移学习的老化风电机组故障预警方法[J]．软件工程，2024，27(5)：68-72.
WANG Xutao，CHEN Huanguo，TAO Hanyu，et al. A fault warning method for aging wind turbine based on transfer learning[J]. Software Engineering，2024，27(5)：68-72.

[5] 魏泰，贺少雄，胡子武，等．基于改进孤立森林算法的风电机组异常数据清洗[J]．科学技术与工程，2024，24(9)：3691-3699.
WEI Tai，HE Shaoxiong，HU Ziwu，et al. Wind turbine abnormal data cleaning based on an improved isolation forest algorithm[J]. Science Technology and Engineering，2024，24(9)：3691-3699.

[6] 韩则胤，王宁，苏宝定，等．基于 Box-Cox 变换结合多种算法的风电机组数据预处理方法研究[J]．计算机测量与控制，2024，32(1)：150-156+164.
HAN Zeyin，WANG Ning，SU Baoding，et al. Research on wind turbine data preprocessing method combined with multiple algorithms based on box-cox transformation[J]. Computer Measurement & Control，2024，32(1)：150-156+164.

[7] 肖湘奇，胡军华，叶志，等．基于多维特征与随机森林的低压用电安全隐患预警方法[J]．电力科学与技术学报，2024，39(2)：143-151.
XIAO Xiangqi，HU Junhua，YE Zhi，et al. A warning method for low-voltage electrical safety hazard based on multi-dimensional features and random forests[J]. Journal of Electric Power Science and Technology，2024，39(2)：143-151.

[8] 王健，李富源，黄鑫，等．基于多参数融合的风电机组齿轮箱油温预警方法[J]．制造业自动化，2023，45(10)：100-103+119.
WANG Jian，LI Fuyuan，HUANG Xin，et al. Fault warning method of gearbox oil temperature of wind turbine based on multi-parameter fusion[J]. Manufacturing Automation，2023，45(10)：100-103+119.

[9] 马良玉，程东炎，梁书源，等．基于 LightGBM-VIF-MIC-SFS 的风电机组故障诊断输入特征选择方法[J]．热力发电，2024，53(1)：154-164.
MA Liangyu，CHENG Dongyan，LIANG Shuyuan，et al. Input feature selection method for wind turbine fault diagnosis based on LightGBM-VIF-MIC-SFS[J]. Thermal Power Generation，2024，53(1)：154-164.

[10] 周伟，魏鑫，李西兴．改进 MFO-LSTM 网络的风电机组齿轮箱故障预警研究[J]．机床与液压，2024，52(4)：185-194.
ZHOU Wei，WEI Xin，LI Xixing. Research on wind turbine gearbox fault early warning based on improved MFO-LSTM network[J]. Machine Tool & Hydraulics，2024，52(4)：185-194.

[11] WANG Z Y，MAE M，YAMANE T，et al. Enhanced day-ahead electricity price forecasting using a convolutional neural network – long short-term memory ensemble learning approach with multimodal data integration[J]. Energies，2024，17(11)：2687.

[12] 王明宇，潘巧波，曹力，等．基于改进随机森林算法的风电机组齿轮箱故障预警方法研究[J]．黑龙江电力，2020，42(6)：508-514.
WANG Mingyu，PAN Qiaobo，CAO Li，et al. Research on fault early-warning method of wind turbine gearbox based on improved random forest algorithm[J]. Heilongjiang Electric Power，2020，42(6)：508-514.

[13] 马良玉，耿妍竹，梁书源，等．基于 Stacking 多模型融合的风电机组齿轮箱油池温度异常预警[J]．中国电机工程学报，2023，43(S1)：242-251.
MA Liangyu，GENG Yanzhu，LIANG Shuyuan，et al.

Early warning of abnormal temperature in gearbox oil pool of wind turbine based on Stacking multi-model fusion[J].Proceedings of the CSEE,2023,43(S1):242-251.

[14] 周伟,魏鑫,李西兴.改进MFO-LSTM网络的风电机组齿轮箱故障预警研究[J].机床与液压,2024,52(4):185-194.
ZHOU Wei, WEI Xin, LI Xixing. Research on wind turbine gearbox fault early warning based on improved MFO-LSTM network[J]. Machine Tool & Hydraulics, 2024,52(4):185-194.

[15] 田雯雯,吕丽霞,刘长良,等.基于动态矩阵与特征相似度的AAKR风电机组状态监测[J].太阳能学报,2024,45(10):536-543.
TIAN Wenwen, LYU Lixia, LIU Changliang, et al. Dynamic matrix and feature similarity-based auto associative kernel regression for condition monitoring of wind turbine[J]. Acta Energiae Solaris Sinica, 2024, 45(10):536-543.

[16] 陈映琼,顾军民,姜胜,等.基于数据清洗和动态神经网络的风电机组齿轮箱油温预警方法研究[J].上海电力大学学报,2020,36(6):547-552.
CHEN Yingqiong, GU Junmin, JIANG Sheng, et al. Research on early warning method of gearbox oil temperature of wind turbine based on data cleaning and dynamic neural network[J]. Journal of Shanghai University of Electric Power,2020,36(6):547-552.

[17] 李加军.基于孤立森林的多离群点数据检测算法设计[J].现代电子技术,2024,47(5):139-142.
LI Jiajun. Design of multi-outlier data detection algorithm based on isolation forest[J]. Modern Electronics Technique,2024,47(5):139-142.

[18] 陈雷,刘林虎,闫川川,等.基于孤立森林和GRU-CNN-Attention的超短期电力负荷预测[J].电子设计工程,2023,31(20):140-144+149.
CHEN Lei, LIU Linhu, YAN Chuanchuan, et al. Ultra-short-term power load forecast based on isolated forests and GRU-CNN-Attention[J]. Electronic Design Engineering,2023,31(20):140-144+149.

[19] YEPMO V, SMITS G, LESOT M J, et al. Leveraging an isolation forest to anomaly detection and data clustering [J].Data & Knowledge Engineering,2024,151:102302.

[20] HOU G L,WANG J J,FAN Y Z.Wind power forecasting method of large-scale wind turbine clusters based on DBSCAN clustering and an enhanced hunter-prey optimization algorithm[J]. Energy Conversion and Management,2024,307:118341.

[21] 罗志宏,刘长良,刘帅.基于近邻元分析的风电机组状态监测特征选择方法[J].华北电力大学学报(自然科学版),2024,51(3):134-142.
LUO Zhihong, LIU Changliang, LIU Shuai. Feature selection method for wind turbine condition monitoring based on neighborhood component analysis[J].Journal of North China Electric Power University (Natural Science Edition),2024,51(3):134-142.

[22] 周鑫,袁泉,张蔷,等.基于信息增益与序列后向搜索的负荷预测方法[J].电气自动化,2020,42(6):73-75.
ZHOU Xin, YUAN Quan, ZHANG Qiang, et al. Load forecasting method based on information gain and sequential backward search[J]. Electrical Automation, 2020,42(6):73-75.

[23] 罗红郊,马晓琴,孙妍,等.基于RFE特征选择的PSO-SVM用电量预测算法[J].电子设计工程,2023,31(20):172-176.
LUO Hongjiao, MA Xiaoqin, SUN Yan, et al. PSO-SVM electricity consumption prediction algorithm based on RFE feature selection[J].Electronic Design Engineering, 2023,31(20):172-176.

[24] 王俊,席芳,周川,等.基于BES-ELM的风电机组故障诊断[J].排灌机械工程学报,2023,41(10):1023-1029.
WANG Jun, XI Fang, ZHOU Chuan, et al. Wind turbine fault diagnostic model based on BES-ELM[J].Journal of Drainage and Irrigation Machinery Engineering,2023,41(10):1023-1029.

[25] 范国庆,李康辉,高捷,等.基于核密度估计和CatBoost算法的光伏功率预测方法[J].上海电力大学学报,2023,39(6):529-535.
FAN Guoqing, LI Kanghui, GAO Jie, et al. A photovoltaic power prediction method based on kernel density estimation and CatBoost method[J]. Journal of Shanghai University of Electric Power,2023,39(6):529-535.

[26] WU Y L, HUANG H. Predicting compressive and flexural strength of high-performance concrete using a dynamic Catboost Regression model combined with individual and ensemble optimization techniques[J]. Materials Today Communications,2024,38:108174.

[27] 徐恒山,莫汝乔,薛飞,等.基于时间戳特征提取和CatBoost-LSTM模型的光伏短期发电功率预测[J].太阳能学报,2024,45(5):565-575.
XU Hengshan, MO Ruqiao, XUE Fei, et al. Short-term photovoltaic power prediction based on timestamp feature extration and CatBoost-lstm model[J]. Acta Energiae Solaris Sinica,2024,45(5):565-575.

[28] 王少科,王培光,张照彦,等.基于机器学习的风电机组齿轮箱故障预警[J].计算机仿真,2023,40(1):99-106.
WANG Shaoke, WANG Peiguang, ZHANG Zhaoyan, et al. Fault early warning of wind turbine gearbox based on machine learning[J].Computer Simulation,2023,40(1):99-106.

[29] 刘杰,曹静,赵昕.基于OOB-GWO-SVR的风电机组齿轮箱故障预警[J].电子测量与仪器学报,2022,36(12):97-105.
LIU Jie, CAO Jing, ZHAO Xin. Wind turbine gearbox fault warning based on OOB-GWO-SVR[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(12):97-105.