DOI:10.19652/j.cnki.femt.2304738

# 基于改进的 K-means 风电机异常数据检测*

陶永辉 王 勇
(上海电力大学计算机科学与技术学院 上海 200090)

**摘 要**:风电机异常数据检测对维护风电设备的稳定运行有着重要意义,为解决 K-means 算法随机指定初始点聚类和风电机数据异常问题,提出一种改进 K-means 算法的风电机数据异常检测方法。改进之后的方法,首先选择数据样本中位数作为第一个初始聚类中心,在选取下一个聚类中心时,距离当前 $n$ 个聚类中心越远的点会有更高的概率被选为第 $n+1$ 个聚类中心,进而达到聚类中心互相距离较远的目的,以此对风电机运行数据进行聚类,检测出离群点及异常点,保障风电设备稳定运行。

**关键词**:K-means 算法;异常检测;初始聚类中心;风电机
**中图分类号**: TP181 **文献标识码**:A **国家标准学科分类代码**: 520.10

# Abnormal data detection of wind turbine based on improved K-means

Tao Yonghui Wang Yong
(College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)

**Abstract**: Abnormal data detection of wind turbine is of great significance for maintaining the stable operation of wind power equipment. In order to solve the problem of clustering randomly assigned initial points of K-means algorithm and abnormal wind turbine data, this paper proposes an improved K-means algorithm to detect abnormal wind turbine data. The improved method first selects the median of the data sample as the first initial clustering center. When selecting the next cluster center, the point farther away from the current $n$ cluster centers will have a higher probability of being selected as the $n+1$ cluster center. And then achieve the goal that the cluster centers are far away from each other. Based on this, the operation data of wind turbines are clustered to detect outliers and outliers, so as to ensure the stable operation of wind power equipment.

**Keywords**: K-means algorithm; anomaly detection; initial cluster center; wind turbine

## 0 引 言

风电行业发展迅速,因此,越来越多的风电机设备被安装,由此也产生了大量的数据,这些数据对风电设备的维护起到至关重要的作用。在这些大量的数据中,无可避免地会夹杂一些风机异常运行产生的数据,而对异常数据的检测则在保障风电设备的正常运行中尤为重要。

异常数据通常可以被认为是其所在数据集内出现频率较少的值,通常这些少数值与其余样本数据的差异较大,包括距离较远、特征不一致等,这就导致异常数据会很清晰地出现在偏离高密度区的地方。数据质量的好坏无疑会和这些异常数据相关,显然,异常数据不仅会影响计算结果,还会导致数据分析错误,甚至还会对下一阶段的发展趋势的预测造成极大的偏差。所以,不管是基于何种原因,对数据进行异常值检测都是非常迫切的。

针对风电设备信息系统数据设计精准高效的异常检测方案,有助于运维人员快速识别影响风电设备的各种异常状况并定位故障,以防故障持续扩大,进而帮助减小损失,提高运维效率。早期对风电设备的异常检测多采用人工巡逻检查方法,而且传统的异常检测方法往往依赖于专家经验设置规则,当数据达到设置好的异常指标时才能检测出异常,不仅效率低,准确率也不高,随着技术不断发展,基于大数据、来源广、异构数据多的情况下,对异常数据进行准确而又高效地自动化检测,成为当前研究

**收稿日期**:2023-02-24
* **基金项目**:上海市自然科学基金(20ZR1455900)、国网上海市电力公司科技项目(SGTYHT/21-JS-223)资助

热点[1]。

针对风电设备领域方面的异常检测工作中，孙滢涛等[2]以多域特征提取为基础，进行电力异常数据检测方案的构建，构造出分类器，然后达到异常检测的目的。张春燕等[3]针对电力数据异常检测效率及精度低的情况，构造了一种电力负荷异常数据检测方案，降低了错误率。余翔等[4]针对电力方面的异常检测，采取孤立森林算法，研究电力消耗数据以及异常数据挖掘过程，结果表明该算法具有优异效果。张昊宇等[5]采用含参数统计的方案，依此进行异常数据检测，史丽萍等[6]则采用了无参统计模型对异常数据检测方案进行构造，两种方法在数据量较小且规律较清晰时能实现较好的检测效果。

胡聪等[7]针对现有异常用电辨识精度低的问题，提出一种基于改进 K-medoids 聚类和支持向量机(support vector machines，SVM)的用电异常行为在线检测方法。吴蕊等[8]改进了传统 K-means 聚类随机选择初始聚类中心的策略，结合数据对象的密集度与最大近邻半径，选择更加接近实际簇中心的数据点作为初始聚类中心，进行电力数据异常检测。黄林等[9]针对传统的异常检测方法难以检测电力信息系统中存在的多个指标综合异常的情况，提出一种基于改进 K-means 算法的异常检测方法，以网格均值点映射该网格内所有样本点来压缩数据。Chahla 等[10]提出了一种新的无监督方法来，通过将 K-means 算法应用于代表当天 24 h 的 24 个不同的 K-means 组来探索用电场景，检测用电数据中的异常。Cui 等[11]提出一种将多项式回归和高斯分布相结合的混合模型，检测和可视化学校用电量数据中的异常情况。Sun 等[12]在对电压数据异常波动检测时，提出了一种改进的 K-means 方法，通过对相关性和可变性分析来进行故障检测。Wu 等[13]提出了一种基于 K-means 参考小区选择方法，以此排查异常数据，有效减少了由于各小区不一致而引起的误报。

在对国内外各大专家学者进行的电力异常检测算法研究来看，近年来，以聚类为底层支撑的异常检测研究依然是重要的课题之一。所以，不管是在理论亦或是在实际应用上来分析，研究聚类算法对电力异常数据的检测依然有其重要的学术和应用价值。

但是，基于多维度指标综合异常的电力异常数据检测的方法较少，且传统 K-means 存在一定缺陷，因此，本文提出了一种经过改进的 K-means 算法，将其应用于风电设备常规运行数据中，并将数据进行降维处理。

## 1 K-means 分析

### 1.1 传统 K-means 缺陷

K-means 算法因其原理简单而成为最受欢迎的聚类算法，同时，在使用过程中，一些缺陷也显现出来了，如算法本身不能自定聚类簇数 $k$；初始聚类中心的确定依靠随机选择的方式；对噪声敏感[14-15]。

针对这些缺点，许多学者对其进行改进。周鑫等[16]提出采取样本分布，以此确定初始聚类中心。黄松等[17]通过改进遗传算法，以并行计算方式弱化 $k$ 值、初始聚类中心对结果的影响。汤深伟[18]基于粒子群优化算法(PSO)，构造出改进 K-means 算法。基于算法对异常值及初始聚类中心敏感，李金涛等[19]进行基于密度的改进。

本文通过对 K-means 算法不能自定 $k$ 值，优化聚类中心选择的方式进行改进，并将其运用到电力数据异常检测中。

### 1.2 传统 K-means 步骤

K-means 聚类算法是一个反复迭代的无监督聚类过程[20]，通过迭代，最终求出最优解，分为如下 4 步。

1)对数据样本随机选取 $k$ 个数据对象，每个数据对象代表一个簇的初始聚类。

2)计算每一个数据样本点与聚类中心的欧氏距离，并将其分配至距离最近的聚类中心所在簇。

3)计算均值向量，将其作为新的聚类中心。

4)判断均值向量是否发生改变，若改变，返回步骤 2)，若不变，则聚类结束。

## 2 异常检测

### 2.1 算法过程

本文首先基于 K-means 无法自定 $k$ 值、初始聚类中心随机选择进行优化。选择手肘法，精确确定聚类初始 $k$ 值。然后使用聚类中心互相远离原则改进后的 K-means 算法进行风电机异常数据检测。算法具体步骤如下。

输入：数据集 $X=\{x_1,x_2,\cdots,x_m\}$；累计解释方差阈值 $\alpha=95\%$，异常值比例 $\beta=1\%$。

输出：异常数据集合 $W$。

风电机异常数据检测算法过程如下。

1)根据手肘法求得初始聚类簇数 $k$。

2)选择数据样本中位数作为首个聚类中心，计算所有样本到各个聚类中心的欧氏距离 $dist(x_j,v)$。

3)对于数据集中的每一个点 $x_i$，计算它与当前已有聚类中心之间的最短距离(即与最近一个聚类中心的距离)，用 $D(x)$ 表示。

4)$D(x)$ 中较大的点作为新的聚类中心，规则为把数据集中每个点到距离其最近的聚类中心的距离相加，其和用 $sum(D(x))$ 表示，在 $0\sim sum(D(x))$ 中取一个随机值 $Random$，执行 $Random=Random-D(x)$，直到 $Random\leqslant 0$，此时的点就是下一个聚类中心。

5)重复步骤 3)、4)，直至选出 $k$ 个聚类中心。

6)如果数据点 $x_j$ 使得 $\mathrm{Min}(dist(x_j,v_i))$ 成立，即 $x_j$ 与第 $i$ 个聚类中心 $v_i$ 最近，那么将 $x_j$ 划分到集合 $c_i$ 中，并重新划分簇。

7)更新聚类中心。

8)聚类中心不再改变，否则返回步骤 2)。

9)根据协方差矩阵、解释方差和累积解释方差、阈值 $\alpha$ 选择主要特征。

10)计算各个样本和最近的聚类中心距离,距离最大就将其视为异常值,异常值比例为 $\beta$。

在实际应用中,方差阈值为95%,这通常被认为是对数据信息的较好保留,可以提供更充分的信息量,同时也可以避免过度拟合的问题,减少了不必要的计算量和噪声。

通常情况下,异常值的比例在0.5%~5%比较合理。如果异常值比例太高,那么可能会影响数据分析和建模的结果;如果异常值比例太低,那么可能会漏掉一些真正的异常值。一般来说,比较常见的设置是0.01或者0.05。

设定一个异常值的比例 $\beta$ 为1%,这样设置是因为在标准正太分布的情况下($N(0,1)$)一般认定3个标准差以外的数据为异常值,3个标准差以内的数据包含了数据集中99%以上的数据,所以剩下的1%的数据可以视为异常值。

## 2.2 算法框架

本文提出的基于改进的K-means算法风电机异常数据检测框架,主要有如下4个部分。

1)使用主成分分析法,以此使数据维度下降和标准化,选取主要特征部分,避免非必要特征对结果产生影响。

2)利用基于初始聚类中心改进的K-means算法对处理过后的风电机数据进行异常检测处理。

3)利用其他算法对该数据集进行异常数据检测。

4)分析实验结果。

整体的算法框架如图1所示。

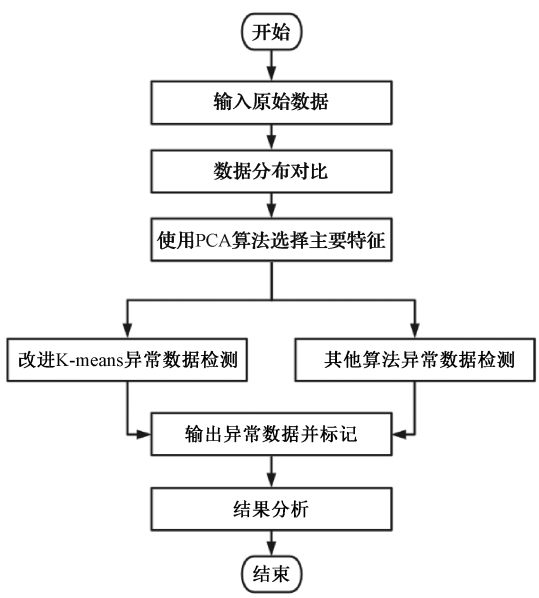



图1 工业异常数据检测框架

## 2.3 主成分分析(PCA)

PCA是数据处理中用到的最常见的一种降维方法[21]。通常情况下,如果数据的变量数减少,那么其精确反映实际信息的程度便会降低,但是,降维的目的却是保留其中较为重要的信息数据,从而不影响整体的精度,因为对于较小的数据集,往往更加容易对其进行处理以及可视化,而且无关紧要的变量也会影响机器学习算法处理数据的速度。总而言之,PCA的目的就是保存较多的信息的同时降低维度。其主要步骤如下。

1)数据标准化。

2)计算协方差矩阵。

3)计算特征向量和特征值。

4)选择主要成分,选择主要特征,并将其作为主轴,后续的数据分析也将围绕该主轴进行。特征向量重要程度,靠其特征值解释方差占比决定。如果有两个不同特征向量 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$,且 $\boldsymbol{X}_1$ 占有39%,$\boldsymbol{X}_2$ 占有11%的方差,因 $\boldsymbol{X}_1$ 具有更多相关信息,则选择其为保留的特征向量。

累积解释方差用于衡量模型与实际数据之间的差异,它是模型总方差的一部分,由存在的因素来解释。通常设定一个贡献率阈值 $\alpha$(一般以80%、85%、90%、95%为主),假定一共 $n$ 个主成分,第1主成分对应的方差即为最大的特征值 $\lambda_1$,第2主成分对应的方差为次大的特征值 $\lambda_2$,计算前 $p$ 个主成分其特征值的累积解释方差率,如果达到 $\alpha$ 时,即可认为这 $p$ 个主成分可以表示原来 $n$ 个变量,公式如下:

$$\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\geqslant\alpha$$

当累计贡献率达到 $\alpha$ 时,拒绝该索引之后的所有特征值和特征向量。

## 2.4 数据集介绍

本文使用数据集来自国家电投某风电场群的一年SCADA真实运行数据,该数据集从风机实际生产过程中收集,是风机在实际工况下运行的典型结果,因此每台风机的原始数据中都包含大量异常数据点。其中,风电场数据中主要信息包括风机运行时的时间戳、风速、功率以及风轮转速,而主要对时间戳以外的3个维度信息进行实验,其具体信息如表1所示。

表1 风机数据集

| 风机编号 | 数据量 | 数据维度 |
|---|---|---|
| 1 | 40 727 | 3 |
| 2 | 38 855 | 3 |
| 3 | 38 995 | 3 |

上述数据都是真实工况下运行的数据,所以不可避免地会夹杂一些异常数据。对风电机数据的异常检测具有重大意义,不仅可以预防事故的发生,还对进一步数据研究以及有效利用起到重要作用。工业数据相较于其他数据来说,具有数据量庞大、数据类型多的特点,在对其进行异常数据检测时,往往需要先对数据进行预处理,然后选

择重要的特征因素进行异常数据检测。

## 3 实验结果及分析

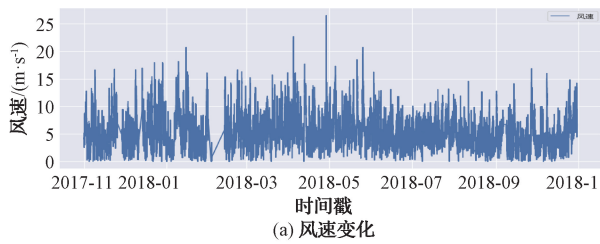
### 3.1 实验环境

操作系统 Windows11；系统类型 64 位；处理器 x64；机带 RAM 8.00 GB；Intel(R) Core(TM) i5-8250U CPU @1.60 GHz 1.80 GHz；实验软件 jupyter notebook；实验环境 Python3。

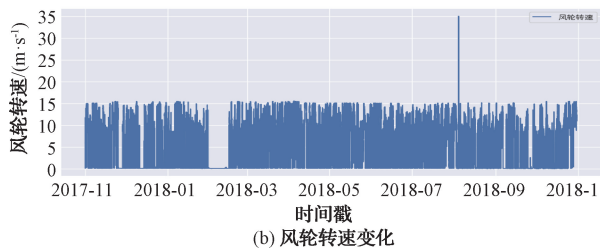
### 3.2 实验过程

1)初始数据分析

首先按照风机编号对风机数据进行统计值计算，以 1 号风机数据为例，按风速、功率、风轮转速变化规律，以时间为序，将其数据变化情况可视化，如图 2 所示。




(a) 风速变化




(b) 风轮转速变化

图 2 1 号风机数据变化

图 2 直观地反映了 1 号风机数据变化情况，便于找到其变化规律，进而得到异常数据点可能出现的情况。求得统计值如表 2 所示。

**表 2 1 号风机风速统计值**

| 统计值 | 数据 | 变量说明 |
|---|---|---|
| count | 40 727.000 000 | 数量统计 |
| mean | 5.734 071 | 均值 |
| std | 2.837 871 | 标准差 |
| min | −0.050 000 | 最小值 |
| 25% | 3.743 079 | 1/4 位数 |
| 50% | 5.264 500 | 1/2 位数 |
| 75% | 7.293 842 | 3/4 位数 |
| max | 26.541 818 | 最大值 |
| Name:风速,dtype:float64 | | |

根据风速大小，对 1 号风机数据进行统计，其可视化如图 3 所示。

统计的目的是为了将 3 个风机风速区间进行对比，找到 3 个风机风速规律，如图 4 所示。

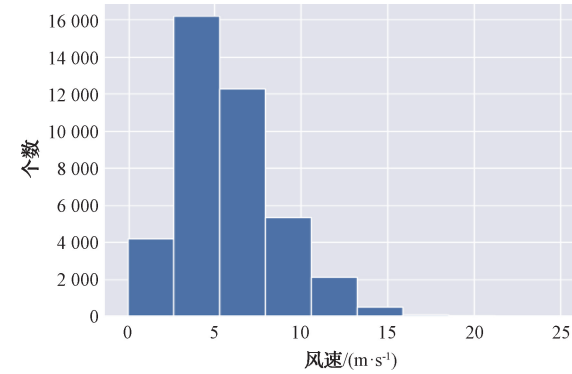



图 3 区间风速统计

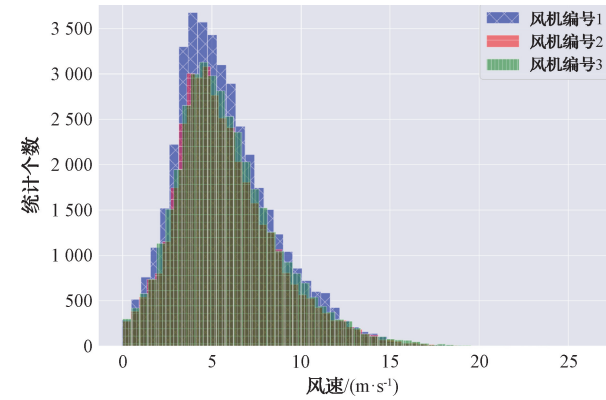



图 4 3 个风机风速对比

从图 4 可以看出，3 个风机风速图像差别不大，即它们有相似的规律，数据的分布区间也都很均匀，便于对其进行数据处理。

2)初聚类

先采取手肘法的方式计算出初始聚类簇数 $k$，手肘法的思想是随着聚类数 $k$ 的增大，每个簇的聚合程度会逐渐提高，误差平方和(sum of the squared errors，SSE)会逐渐变小。当 $k$ 小于真实聚类数时，由于 $k$ 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 $k$ 到达真实聚类数时，再增加 $k$ 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 $k$ 值的继续增大而趋于平缓，也就是说 SSE 和 $k$ 的关系图是一个手肘的形状，而这个肘部对应的拐点 $k$ 值就是数据的真实聚类数。如图 5 所示。

可以清晰地看到，当 $k$ 增加到 5 个以上时，图像开始出现收敛，之后图像趋于平和，因而初始聚类簇数 $k$ 定为 5。选择 2 号风机，所有风机进行聚类可视化如图 6、7 所示。

图 6 和 7 中，数据点分别被分成了 5 种颜色，而这 5 种颜色则代表聚类之后的 5 个簇。对风机数据聚类后，往往需要识别出数据里面的主要成分，通过 PCA 方法确定。首先对数据采取标准化措施，其次，求出特征变量的协方

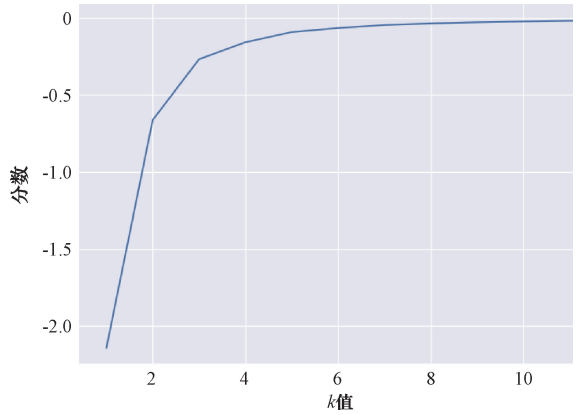



图 5　$k$ 值变化

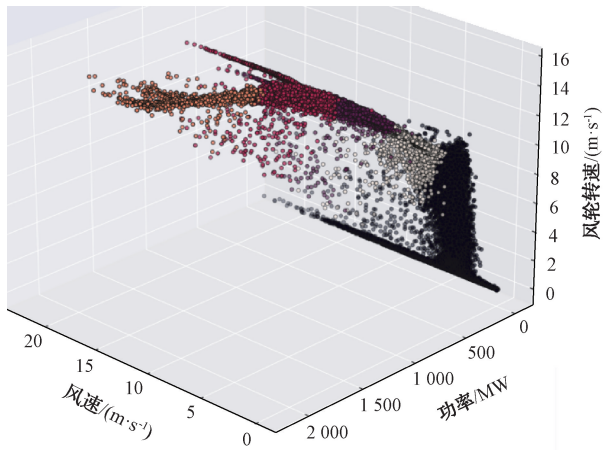



图 6　2 号风机聚类图

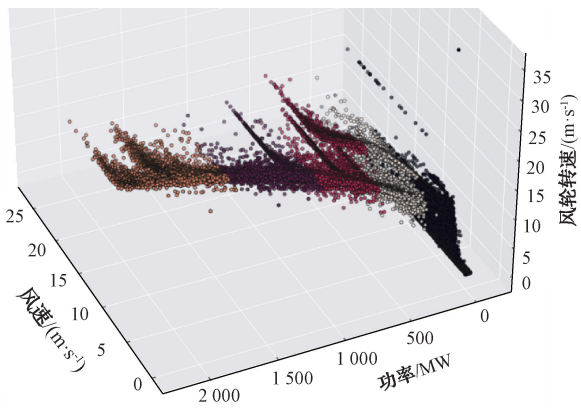



图 7　所有风机聚类图

差矩阵，该项指标可以显示特征变量之间的相关程度。依据特征值，算出各个主特征的解释方差和累计解释方差，将结果可视化如图 8 所示。

从图 8 可以看出，风机数据中有 3 个主要特征。风速概括了约 80%的方差；功率概括了 15%的方差。本文设置累计贡献率阈值 $\alpha$=95%，那么这俩个特征便已经囊括

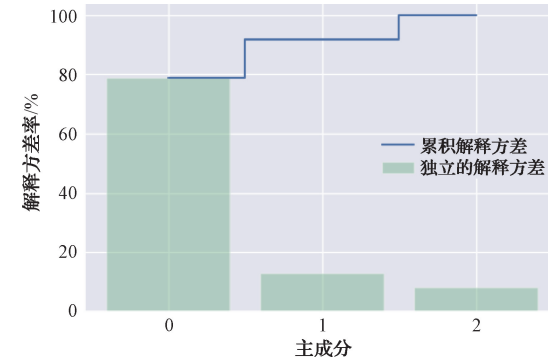



图 8　解释方差

了大约 95%的方差，所以，选取风速以及功率二者作为主成分，然后使用 PCA 算法进行降维。

3）异常检测

通常，假定在对一组数据进行聚类时，认为异常数据不在任何聚类之中，通过以下步骤来判定异常以及将其可视化处理。

（1）计算各个样本和离它最靠近的聚类中心的距离，其中距离值最大，就把它视为异常。

（2）设定异常值的比例为 1%。

（3）依照异常值的比例，求出异常值个数。

（4）对所有数据可视化。

对数据进行计算，得到有效数据量为 117 392，则根据异常值比例 1%设置异常数据，为 1 185，按照主要特征对异常数据进行可视化，如图 9 所示。

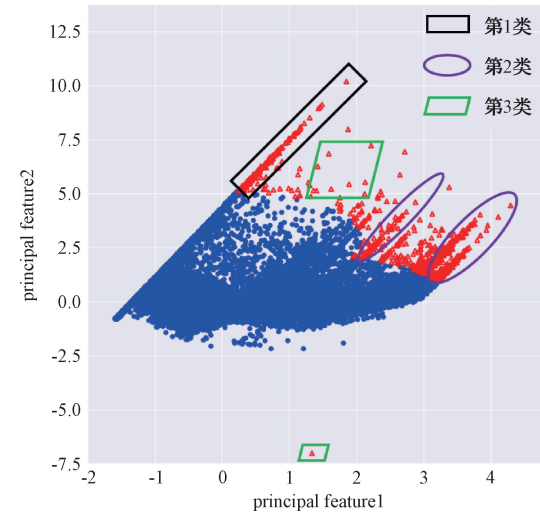



图 9　异常数据－特征图

图 9 中横轴表示数据在第 1 主成分上的值，是通过主成分分析降维方法求得的，是基于多个特征的加权组合，纵轴表示第 2 主成分的数据值。红色点为异常值，蓝色点则为正常值。经分析，其主要故障类型有 3 种[22]，第 1 个为数据底部出现的异常，该种情况的出现可能是机组设备

或者终端等出现故障导致的；第 2 个为中间聚集异常，这一种异常数据可能是由于限电所导致；第 3 个是零散出现的异常，这一种异常数据可能是由于通信出现错误，或者是传感风速的器件故障而导致的。

求出异常数据后，以时间为序，将异常数据显示在风轮转速上，观察异常数据分布规律，如图 10 所示。

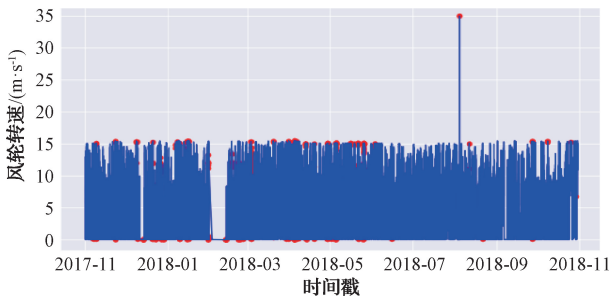



图 10　异常值分布

由图 10 可以得到，经过数据降低维度，以及改进 K-means 算法得到的异常结果，异常点出现在风轮转速最高以及最低处，因为过低或者过高的风轮转速往往违反实际运行规律，极有可能是运行出现了异常，包括数据传输异常、传感器异常等进而造成数据异常。

4)其他异常检测算法

经过改进 K-means 聚类算法对数据进行异常检测后，可以采用其他检测算法对数据进行检测。

(1)孤立森林

使用孤立森立模型对数据进行异常检测，其原理是在数据里面找出规律不太符合其他的，且这种方法并不和以距离为要点的异常检测方法相同，该算法以少数和不同数据为异常的原则，同样选择基于风轮转速可视化异常点，如图 11 所示。

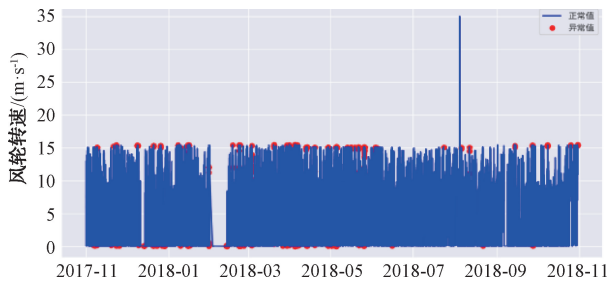



图 11　孤立森立异常数据分布

由图 11 可以得到，使用这种方法算出的结果，和图 10 结果类似，在最低和最高点处均有出现。

(2)支持向量机

SVM 实现异常检测的原理为把高密度数据区域划为正，而低密度的划为负，其可视化结果如图 12 所示。

由图 12 可以得到，使用该算法算出的结果，其异常往往出现在风轮转速最高、最低的地方，但最低点异常数据较多。

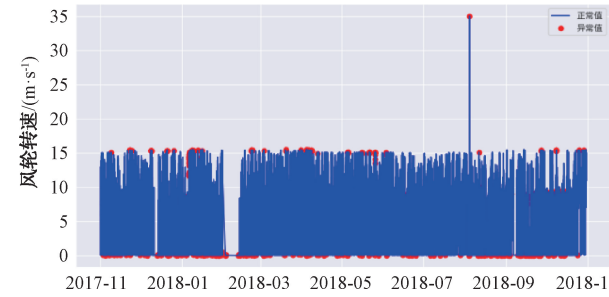



图 12　支持向量机异常数据分布

(3)高斯概率分布

高斯分布通常也可以被叫做正态分布。对于正态分布的数据，其可以实现异常值判定，其可视化结果如图 13 所示。

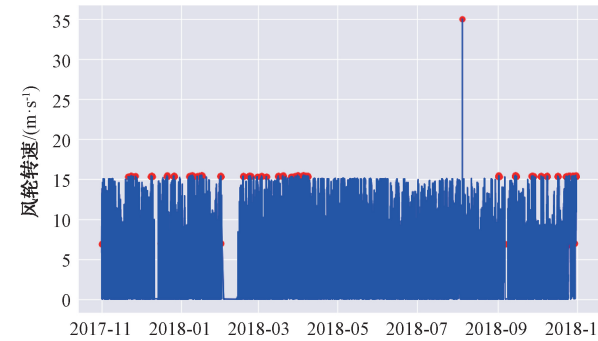



图 13　高斯概率分布异常数据分布

由图 13 可以得到，使用该种方法预测出的结果，其异常经常出现在风轮转速最高的地方，但是，在最低点处却没有。

对 4 种算法求得的异常数据进行汇总，如表 3 所示。

**表 3　各算法异常数据汇总**

| | 1 号风机 | 2 号风机 | 3 号风机 | 汇总 |
|---|---|---|---|---|
| 改进 K-means | 599 | 565 | 21 | 1 185 |
| 孤立森林 | 573 | 584 | 28 | 1 185 |
| 支持向量机 | 490 | 335 | 361 | 1 186 |
| 高斯概率分布 | 412 | 266 | 508 | 1 186 |

通过表 3 可以看出，使用改进 K-means 算法和孤立森林得到的异常数据中，1 号、2 号风机占了绝大多数，但 3 号风机数量较少，而支持向量机和高斯概率分布得到的异常数据较为分散，对该表数据进行可视化，如图 14 所示。

由图 14 可以清晰地看出，改进 K-means 算法和孤立森林异常检测的效果相似，而与其他算法的效果各不相同，结合上文各算法对应的异常数据－风轮转速图来看，孤立森林无疑是最符合现实情况的。因此，本文提出的算法在进行异常检测的时候，具有更贴近实际，以及更准确的效果。

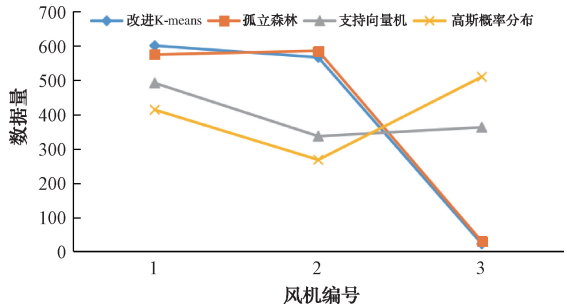



图 14　各算法异常数据对比

到目前为止，本文已经用 4 种不同的方法进行了风机数据异常检测。通常异常检测只有在实际的应用场景中才能测试出效果。由于这些异常检测算法都是无监督学习，所以在构建模型之后，不知道异常检测效果如何，通过对比的方式，找出异常点的共性，无疑是一种最优的评估方案。通过对比可以发现，本文提出的改进后的 K-means 算法可以在实际应用中起到很有效的作用。

因此，基于改进的 K-means 风电机异常数据检测是可行的。

## 4　结　论

本文对国内外关于异常数据检测的现状进行了介绍，并对原始 K-means 的缺点进行剖析，提出了以 K-means 为底层基础的改进算法进行电力异常数据检测，解决了传统 K-means 聚类算法无法自定 $k$ 值，初始聚类中心无法精确预设导致算法出现局部最优或者聚类效果较差的情况，同时针对多维度指标综合异常的电力异常数据，运用了 PCA 降维处理的方法，最后采用多种算法对比的评估方式，对真实 SCADA 数据为基础进行实验，实验结果表明，本文提出的异常检测方案具有优异结果。

### 参 考 文 献

[1] 高书强，李晨. 一种针对电力数据异常检测的改进谱聚类算法[J]. 计算机仿真，2019，36(11)：239-242.
[2] 孙滢涛，张锋明，陈水标，等. 基于多域特征提取的电力数据异常检测方法[J]. 电力系统及其自动化学报，2022，34(6)：105-113.
[3] 张春燕. 基于数据挖掘的电力负荷异常数据检测与修正研究[D]. 兰州：兰州理工大学，2019.
[4] 余翔，陈国洪，李霆，等. 基于孤立森林算法的用电数据异常检测研究[J]. 信息技术，2018，42(12)：88-92.
[5] 张昊宇，姚钢，殷志柱，等. 基于小波神经网络与KNN 机器学习算法的六相永磁同步电机故障态势感知方法[J]. 电测与仪表，2019，56(2)：1-9.
[6] 史丽萍，宋朝鹏，李明泽，等. 基于 SAAFSA 优化加权模糊聚类算法的变压器故障诊断[J]. 电测与仪表，2017，55(11)：12-18.
[7] 胡聪，徐敏，洪德华，等. 基于改进 K-medoids 聚类和 SVM 的异常用电模式在线检测方法[J]. 国外电子测量技术，2022，41(2)：53-59.
[8] 吴蕊，张安勤，田秀霞，等. 基于改进 K-means 的电力数据异常检测算法[J]. 华东师范大学学报(自然科学版)，2020 (4)：79-87.
[9] 黄林，常健，杨帆，等. 基于改进 K-means 的电力信息系统异常检测方法[J]. 深圳大学学报(理工版)，2020，37(2)：214-220.
[10] CHAHLA C，SNOUSSI H，MERGHEM L，et al. A deep learning approach for anomaly detection and prediction in power consumption data[J]. Energy Efficiency，2020(13)：1633-1651.
[11] CUI W，WANG H. A new anomaly detection system for school electricity consumption data[J]. Information，2017，8(4)：151.
[12] SUN Z，HAN Y，WANG Z，et al. Detection of voltage fault in the battery system of electric vehicles using statistical analysis[J]. Applied Energy，2022，307：118172.
[13] WU M，DU W，ZHANG F，et al. Fault diagnosis method for lithium-ion battery packs in real-world electric vehicles based on K-means and the Fréchet algorithm[J]. ACS Omega，2022，7(44)：40145-40162.
[14] 陶永辉，王勇. 基于初始聚类中心选取的改进 K-means 算法[J]. 国外电子测量技术，2022，41(9)：54-59.
[15] 顾洪博，赵万平. 数据挖掘算法性能优化的研究与应用[J]. 长春理工大学学报(自然科学版)，2010 (1)：164-166.
[16] 周鑫，张化祥. K-means 算法的研究与改进[J]. 微计算机信息，2008，24(30)：269-270.
[17] 黄松，邱建林. 改进的遗传 K-means 算法及其应用[J]. 计算机工程与设计，2020，41(6)：1617-1623.
[18] 汤深伟. 基于改进粒子群的 K-means 聚类算法及其在推荐系统中的应用[D]. 合肥：安徽大学，2020.
[19] 李金涛，艾萍，岳兆新，等. 基于 K-means 聚类算法的改进[J]. 国外电子测量技术，2017，36 (6)：9-13.
[20] 颜佩，丁亚军，钱盛友，等. 基于 K 均值聚类的组织损伤等级判定研究[J]. 电子测量与仪器学报，2017，31 (3)：468-473.
[21] 赖思银. 基于 PCA 和改进 SVDD 的异常检测算法[J]. 微型电脑应用，2022，38(8)：129-132.
[22] 饶日晟，叶林，任成，等. 基于实际运行数据的风电场功率曲线优化方法[J]. 中国电力，2016，49(3)：148-153.

作者简介

陶永辉，硕士研究生，主要研究方向为数据挖掘、网络安全、机器学习等。
E-mail：1519212855@qq.com

王勇，博士，教授，主要研究方向为网络安全、恶意代码检测、电力信息安全等。
E-mail：wy616@126.com