# 风电机组数据采集与监控系统异常数据识别方法

李特，王荣喜，高建民
（西安交通大学机械制造系统国家重点实验室，710049，西安）

**摘要**：为了解决原始的风电机组数据采集与监控系统(SCADA)中包含大量异常记录的数据、难以准确反映机组运行状态的问题，提出了一种带噪声基于密度的空间聚类(DBSCAN)模型的风电机组SCADA异常数据识别方法。该方法从分析风速-功率曲线的特点出发，采用预测误差和分类准确度来选取关键聚类参数邻域半径和邻域最小样本点数，避免了人工确定聚类参数的主观性，且参数选择过程可以完全自动化，实现了风电机组SCADA异常数据的有效识别。通过某风场中风电机组的监测数据进行实例验证，结果表明：所提方法能够在保证异常数据被剔除的前提下，保留尽可能多的正常数据，异常识别效果好于现有的$k$-dist图法和基于$k$-平均最近邻算法的改进算法(KANN-DBSCAN)。该研究可为开展风电机组状态分析提供参考。

**关键词**：风电机组；异常数据识别；空间聚类；风速-功率曲线
**中图分类号**：TH17 **文献标志码**：A
**DOI**：10.7652/xjtuxb202403010 **文章编号**：0253-987X(2024)03-0106-11

# A Method for Abnormal Data Recognition of Wind Turbine Supervisory Control and Data Acquisition Systems

LI Te, WANG Rongxi, GAO Jianmin
(State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract**: To address the issue that wind turbines' supervisory control and data acquisition (SCADA) system contains a significant amount of data about abnormal records, which affects the accurate representation of the turbines' operational status, a method for identifying abnormal data based on density-based spatial clustering of applications with noise (DBSCAN) is proposed. Based on the characteristics of the wind speed-power scatter curve, this method involves the use of prediction errors and classification accuracy to select the key clustering parameters: neighborhood radius and minimum number of sample points in the neighborhood. It avoids the subjectivity of manually determining the clustering parameters, allowing for a fully automated parameter selection process. As a result, it achieves effective identification of abnormal data in a wind turbine's SCADA system. The proposed method is validated using monitoring data from wind turbines in a specific wind farm. The results demonstrate that the method helps to retain as much normal data as possible while ensuring the removal of abnormal data. It also shows superior anomaly identification performance compared to $k$-distance graph and KANN-DBSCAN, an improved algorithm based on $k$-nearest neighbors. This study provides valuable insights for the

收稿日期：2023-08-01。 作者简介：李特(1999—)，男，硕士生；王荣喜(通信作者)，男，博士，副研究员。 基金项目：国家自然科学基金资助项目(51905409)。
网络出版时间：2023-12-02 网络出版地址：https://link.cnki.net/urlid/61.1069.t.20231201.1032.002

status analysis of wind turbines.

**Keywords**: wind turbine; abnormal detection; spatial clustering; wind speed-power curve

由于全球气候问题和能源需求的增长，全球在清洁能源领域的投资都在逐渐增加。风力发电是当前可再生能源领域技术最成熟、增长速度最快、商业化发展最好的发电方式之一，其大规模发展对能源结构的调整、应对能源需求增长和环境挑战、实施可持续低碳能源战略具有重要的意义。受风力资源分布限制，风电机组通常建在偏远山区或者海上[1]，工作环境恶劣，长期面临冰冻、台风、潮湿或盐雾腐蚀等问题，机组发生故障的概率显著增加，提高了机组的运维成本[2-3]。

实时监测风电机组的运行状态，可以在机组状态发生异常时甚至在发生异常之前及时地采取针对性的措施，对降低机组的运维成本具有重要意义。数据采集与监控系统(supervisory control and data acquisition，SCADA)记录和存储了大量风电机组的监测数据，蕴含了丰富的状态信息。因此，国内外诸多的研究者都基于 SCADA 数据来对风电机组的状态监测展开研究[4-7]。

由于传感器故障、存储出错和通讯干扰等原因[8]，SCADA 记录的数据中错误地记录了一些监测数据，这些数据也被称为异常数据。异常数据并非风电机组运行时的真实记录，因此，原始的 SCADA 数据难以准确反映风电机组的运行状态。采取有效的措施来对风电机组原始 SCADA 数据中的异常值进行识别并剔除，是后续进行状态监测或状态分析等研究工作的基础。

目前，风电机组 SCADA 数据异常值识别的相关研究通常以风速-功率曲线为依据，大致可以分为基于统计量的方法、基于图像处理的方法和基于聚类分析的方法[9]。常见的基于统计量的方法有 $3\sigma$ 法[10]和四分位法[11]，这些方法的基本思路为计算原始数据的相关统计量，然后将统计量范围以外的数据视为异常数据。虽然基于统计量的方法操作简单，但其有两个明显的缺陷，一方面，这类方法受原始数据质量的影响很大，当数据中异常值的比例较高时，这些方法的表现较差；另一方面，单一的统计量方法难以处理多种不同类型的异常值，通常需要与其他方法组合[12-14]才能达到较好的效果。基于图像处理的方法[15-16]通常将风速-功率曲线转换为二值图像，然后利用图像分割等技术来进行异常值的识别，这类方法通常能够有效识别出堆积型的异常数据，但是识别速度很慢，并且难以实现[9]。基于聚类分析的方法将异常识别视为无监督聚类问题，基于样本之间的距离[17]或簇密度[18-20]来将正常数据和异常数据聚类为不同的类别。这类方法相较于基于图像处理的方法更加简单易行，对原始数据中异常数据的比例不敏感，受到了研究者的广泛关注。

带噪声基于密度的空间聚类(density based spatial clustering of applications with noise，DBSCAN)[21]是一种经典的基于密度的聚类方法，它并不需要事先确定聚类的簇数，聚类速度快，能发现任意形状的空间聚类并且可以有效处理噪声点，对于将 SCADA 中正常数据与异常数据分离开的任务来说尤为适用。然而，邻域半径 $\varepsilon$ 和邻域内最小样本数 $n_p$ 这两个参数的选择对 DBSCAN 的聚类结果影响非常大。经典的 DBSCAN 模型通过 $k$-dist 图法来确定参数，基本过程为：首先计算数据集中每个对象与距该对象第 $k$ 近的对象之间的距离，记作该对象的 $k$-dist 值，然后将所有对象的 $k$-dist 值按照升序或降序进行排列并绘制 $k$-dist 曲线，取 $k$-dist 曲线的第一个“拐点”对应的距离作为 $\varepsilon$，取 $k$ 作为 $n_p$。由于 4-dist 与其他的 $k$-dist 曲线没有显著差异，所以一般默认 $k=4$。显然，$k$-dist 图法在确定参数时需要过多的人工干预，而且 $k$ 的选取具有较大的主观性。

目前，许多研究者针对 DBSCAN 参数确定问题进行了研究，文献[22]中提出的 SA-DBSCAN 使用逆高斯拟合 4-dist 的分布并求分布曲线的峰值点来确定 $\varepsilon$，再通过不同 $n_p$ 下噪声点数与 $n_p$ 的关系来近似求解得到最佳的 $n_p$，但是可能出现无法求得最优解的情况[23]。文献[24]中提出的 I-DBSCAN 方法通过聚类数和噪声点数的下降趋势来确定 $\varepsilon$ 的取值，在此基础上计算每个点邻域内样本点数的平均值作为 $n_p$，但是确定 $\varepsilon$ 的过程中依然需要人的干预。文献[23]中提出的 KANN-DBSCAN 方法通过计算所有对象在不同 $k$ 下的 $k$-dist 值，并对同一个 $k$ 下所有结果进行求平均，得到 $\varepsilon$ 列表，然后对每个 $\varepsilon$ 按文献[24]中的方法确认 $n_p$，生成 $n_p$ 列表，依次取每一对参数组合进行 DBSCAN 聚类，当聚类簇数趋于不变后再次变化时对应的参数组合为最佳聚类参数，但是其容易受噪声数据的影响[25]。

综上所述，现有的改进工作虽然在一定程度上

解决了 DBSCAN 模型参数选择的问题，但是仍存在如下局限：①大多数改进方法并不是真正做到了参数确定的自动化，过程中仍需人为参与来确定其中一个参数的值；②现有的方法都是针对通用数据提出参数选择的判定方法，并没有利用风电机组风速-功率曲线的分布特点。

为此，本文以 DBSCAN 作为基础模型，通过分析风电机组风速-功率曲线的分布特点，设计了预测误差和分类准确度这两个评价指标来实现全自动化选取其邻域半径和最小邻域样本数这两个聚类参数，并以实际生产中的风电机组 SCADA 数据为例验证了本文方法的有效性。

## 1 风电机组 SCADA 异常数据分类

### 1.1 理想功率特性曲线

功率特性曲线直观地描述了风速和发电功率之间的关系，是评价风电机组发电性能的重要指标，也是风电机组 SCADA 数据是否异常的重要依据。设 $v_{in}$ 为切入风速，$v_r$ 为额定风速，$v_{out}$ 为切出风速，$P_r$ 为额定功率。风电机组的理想功率特性曲线如图 1 所示。理论上，当风速很低时，风电机组处于停机状态；当风速达到切入风速时，风电机组开始输出功率，且功率随着风速的增大而增大；当风速达到某一值时，机组的功率达到额定功率，此时的风速即为额定风速；当风速超过额定风速之后，机组的输出功率不会再继续增大，而是被限制在额定功率；当风速过大至超过了风电机组的切出风速时，为了机组的安全，风电机组将进入停机保护状态。功率与风速的关系呈“厂”字型。

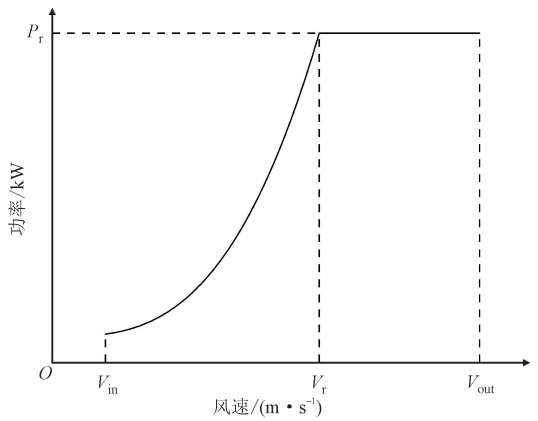


图 1 风电机组理想功率特性曲线

Fig. 1 Ideal power characteristic curve for wind turbines

### 1.2 实际功率-风速曲线

在风电机组实际工作过程中，正常的风电机组的风速-功率曲线应该分布在理想功率曲线周围。然而，由于传感器故障和通讯干扰等原因，SCADA 系统采集到的数据通常包含异常值，这导致实际的风速-功率曲线与理想的功率特性曲线呈现出较大差异。以某风场 32 号风机为例，其在 2021 年 1 月 1 日—2021 年 12 月 31 日一整年的风速-功率曲线如图 2 所示。

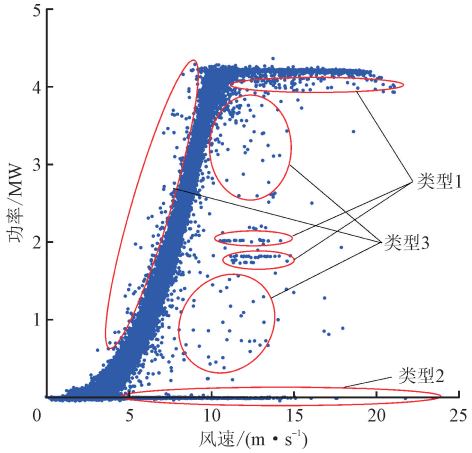


图 2 某风场 32 号风电机组风速-功率曲线

Fig. 2 Scatterplot of wind speed-power of wind turbine No. 32 in a wind farm

从图 2 中可以看出，SCADA 系统中记录的正常数据会较为紧密地分布在功率特性曲线周围，而异常数据看起来是由与“正常”数据完全不同的机制产生的，与风电机组正常的历史数据分布特征明显不符。根据风速-功率曲线的特点，通常可以将风电机组 SCADA 数据中的异常值分为 3 类，下面对每种异常值的分布特点和产生原因进行分析。

#### 1.2.1 中部堆积型

第 1 类异常数据堆积在图形中部，具体表现为 SCADA 系统记录的实时风速接近或大于风电机组的额定风速，但是输出功率小于额定功率，并且输出功率波动非常小，几乎不随着风速的变化而变化。此类异常数据通常是由“弃风限电”造成的，即由于该地区外部输电能力限制和电网消纳能力不足等原因，工作人员强制使风机处于较低的功率输出状态。

#### 1.2.2 底部堆积型

第 2 类异常数据堆积在图形底部，具体表现为 SCADA 系统记录的实时风速大于风电机组的切入风速，但是没有功率输出。此类异常数据通常是由风机操作人员为了检修而强制使风机进行停机而产生的。此外，传感器故障和数据存储出错也可能导致这类异常数据的产生。

#### 1.2.3 离散分布型

第 3 类异常数据离散、稀疏和孤立地分布在功

率特性曲线两侧，通常是由传感器发生故障或者在信号传输和处理过程中受到噪声干扰而产生的。

综上所述，在 SCADA 系统中存在着多种“离群”的异常数据，每种异常数据产生的原因和表现出的数据特点各不相同，这些含有大量异常值的 SCADA 数据对准确反映和识别风机状态带来了影响，针对异常值的数据特性采取适当的预处理措施十分必要。

## 2　基于改进 DBSCAN 的风电机组 SCADA 异常数据识别

### 2.1　DBSCAN 模型

DBSCAN 算法的基本概念如下。

(1)在数据集 $D$ 中，给定一个对象 $p$，以其为中心，$\varepsilon$ 为半径的范围内的区域称为对象 $p$ 的 $\varepsilon$ 邻域；

(2)若一个对象 $p$ 的 $\varepsilon$-邻域中至少包含 $n_p$ 个样本点，记作 $N_\varepsilon(p) \geqslant n_p$，那么称 $p$ 为一个核心对象；

(3)在数据集 $D$ 中，如果 $p$ 是一个核心对象，且 $q$ 为其 $\varepsilon$-邻域内的样本点，那么称 $q$ 是从 $p$ 密度直达的；

(4)在数据集 $D$ 中，如果 $q$ 是从 $p$ 密度直达的，而 $r$ 是从 $q$ 密度直达的，那么称 $r$ 是从 $p$ 密度可达的；

(5)在数据集 $D$ 中，如果 $q$ 和 $r$ 都是从 $p$ 密度可达的，那么称 $q$ 和 $r$ 是密度相连的。

假设邻域半径 $\varepsilon = r$，邻域最小样本数 $n_p = 3$，如图 3 所示。由前文所述可知，对象 $p_2$ 的 $\varepsilon$-邻域内的样本数大于 3，因此该点为核心对象；$p_1$ 在 $p_2$ 的邻域内，因此 $p_1$ 从 $p_2$ 密度直达；同理，$p_5$ 从 $p_1$ 密度直达，因此 $p_5$ 从 $p_2$ 密度可达；同理，$p_4$ 从 $p_2$ 也密度可达，因此 $p_4$ 和 $p_5$ 是密度相连的。

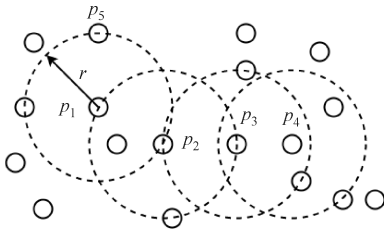


图 3　DBSCAN 算法基本概念示意图

Fig. 3　Schematic diagram of the basic concepts of the DBSCAN algorithm

DBSCAN 聚类的基本思想是将所有互相密度相连的数据点归为一类，其聚类过程如下。

(1)对于给定的数据集 $D$，设定合适的邻域半径 $\varepsilon$ 和邻域最小样本点数 $n_p$；

(2)从数据集 $D$ 中随机选取一个样本点，如果该样本点为核心对象，则找到所有与该点密度相连的样本点，否则暂时将其标记为噪声点，噪声点在后续步骤中仍然可能被考虑到；

(3)重复步骤(2)直到找到一个核心对象，遍历该核心对象 $\varepsilon$-邻域内所有的核心对象，找到所有与这些核心对象密度相连的样本点；

(4)在不属于已确认的任何一类数据的样本点中，重复步骤(2)和(3)，直到没有新的核心对象为止；

(5)最终不属于任何一类数据的样本点被认为是噪声。

尽管 DBSCAN 可以不用事先确定聚类的簇数，也可以找出任何形状的簇群，但是其聚类结果对邻域半径 $\varepsilon$ 和最小点数 $n_p$ 这两个参数很敏感。风电机组 SCADA 异常数据识别任务属于一个没有标签的无监督任务，目前只能通过肉眼观察来选择聚类参数，具有很强的主观性，设计合适的指标和方法来从备选参数中选出最佳参数十分必要。

### 2.2　DBSCAN 最佳聚类参数确定方法

DBSCAN 模型对 SCADA 数据聚类后，正常类别会较为紧密地聚集在理想功率曲线附近，散落在四周的其他类别和噪声点被划分为异常类别。为了评估 SCADA 异常数据识别的有效性，首先需要明确该任务的目标，即保证异常数据被剔除的前提下，保留尽可能多的正常数据。基于此，本文提出预测误差和分类准确度两个指标来进行最佳聚类参数的选取。

(1)预测误差 $e_{pn}$。训练一个回归模型，训练集和测试集均为 DBSCAN 聚类的正常类别的样本点，输入为样本点的风速，输出为对应的功率，$e_{pn}$ 为预测模型的预测误差。风电机组正常的运行监测数据较为紧密地分布在理想功率曲线周围，表明正常数据的风速和功率之间存在特定的映射关系。正常类别中的数据越贴近理想功率特性曲线，那么正常类别中的确为正常数据的比例越高，正常类别中样本的风速-功率映射关系更为“相似”，回归模型相同的情况下，预测误差 $e_{pn}$ 就会越小。

(2)分类准确度 $a_c$。训练一个分类模型，其中输入为风速和功率，输出为数据的类别，训练时假设 DBSCAN 聚类后的类别(正常或异常)为该数据的“真实”标签，$a_c$ 为分类模型的分类准确度。当聚类模型的聚类结果较为准确时，说明数据的“真实”标签也较为准确，分类模型相同的情况下，分类准确度 $a_c$ 也会比较高。

从备选参数范围中选取最合适聚类参数的过程如下：先将每个备选参数组合代入 DBSCAN 模型对原始数据进行聚类，计算出其两个评价指标值，然后

按照 $e_{pn}$ 递增的顺序对各参数组合进行排列，选取 $a_c$ 的第一个极大值点对应的参数组合为最终的聚类参数。

### 2.3 基于改进 DBSCAN 的风电机组 SCADA 异常数据识别流程

本文提出的风电机组 SCADA 异常数据识别流程如图 4 所示。

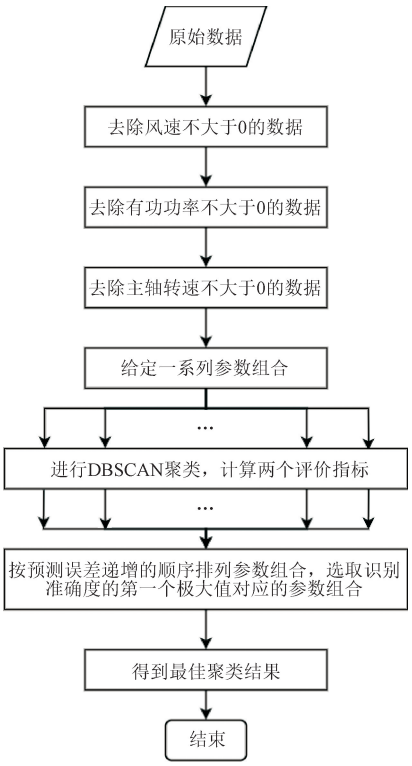


图 4 风电机组 SCADA 异常数据识别流程

Fig. 4 Wind turbine SCADA anomaly data detection process

需要进行基于规则的初步筛选。从理论上来说，当正常监测数据占全部监测数据的比例远大于异常数据所占比例时，基于 DBSCAN 聚类的异常识别方法能够很容易地从数据中发现“正常”的模式，从而剔除异常值。根据前文对异常数据出现原因的分析可知，风电机组监测数据中的各类异常数据占比并不低，特别是底部堆积型，其与正常数据区域相邻，直接针对所有的监测数据使用异常值检测方法难以处理这类堆积型数据[6]。因此，筛选出明显异常的数据从而减小这些数据对后续基于聚类的异常值识别准确性的影响很有必要。本文基于以下规则删除明显异常的数据：①风速不大于 0；②风电机组的有功功率不大于 0；③风电机组的主轴转速不大于 0。

初步筛选之后，进行 DBSCAN 聚类。为了消除风速和功率的不同量纲带来的影响，先按照下式对数据进行 $z$-Score 正则化

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

式中：$x$ 表示需要被标准化的原始数据；$\mu$ 表示原始数据的平均值；$\sigma$ 表示原始数据的标准差；$z$ 表示标准化后的结果。

将备选参数组合分别代入 DBSCAN 模型，得到聚类后的结果，按照 2.2 节中介绍的方法选取最佳聚类参数，从而得到最佳聚类结果。

## 3 实例验证

### 3.1 效果验证

本小节以前文所述某风场 32 号风电机组一年左右的监测数据为例来验证所提出的异常值识别方法的有效性，该监测数据的采样间隔为 10 min，共有 51 101 个样本点，包含环境、工况和状态参数，但在异常识别中只使用风速和功率两个变量。

首先，基于基本规则对原始数据进行初步筛选，经过筛选后剩余 37 519 个数据点，筛选结果如图 5 所示。从图 5 中可以看出，虽然绝大多数底部堆积型异常数据都被剔除，但是仍有部分底部堆积型异常值被保留，同时，中部堆积型和离散分布型两类异常值完全没有被识别出来。因此，需要对筛选出来的数据进行进一步的异常值识别。

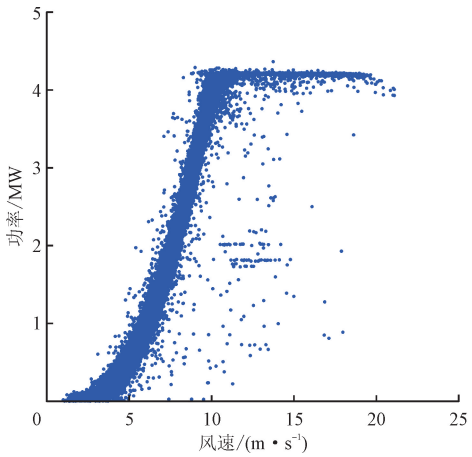


图 5 初步筛选后的结果

Fig. 5 Results after initial screening

在初步筛选的基础上，对剩余的样本使用改进的 DBSCAN 方法来进行分析。在本例中，邻域半径 $\varepsilon$ 从{0.02，0.04，0.06，0.08，0.10}中选取，最小相邻点数 $n_p$ 从{4，6，8，10，12}中选取，构成共 25 种参数组合。因为模型性能评估的结果是比较不同参数组合的聚类结果，而不是为了提高模型的性能，并且模型的输入输出都比较简单，因此使用比较简单的网络结构是合理的。本文中的分类模型和预测模型统一采用简单的多层感知机(MLP)网络结构。

按照图 4 中的流程，依次对 25 种聚类参数组合进行 DBSCAN 聚类，然后计算两个评价指标值。下面首先介绍本文中评价指标的计算方法。

(1)预测误差 $e_{pn}$。预测模型常用的评价指标有均方误差(mean squared error，MSE)、平均绝对误差(mean absolute error，MAE)和平均绝对百分比误差(mean absolute percentage error，MAPE)等，在本文中，定义聚类数据预测误差 $e_{pn}$ 为回归模型在测试集上的 MSE，其计算公式如下

$$e_{pn}=\frac{\sum_{i=1}^{n}(f(s_i)-p_i)^2}{n_t} \tag{2}$$

式中：$s_i$ 表示测试集中第 $i$ 个样本的风速；$f(s_i)$ 表示根据第 $i$ 个样本的风速预测得到的功率；$p_i$ 表示测试集中第 $i$ 个样本的真实功率；$n_t$ 表示测试集样本数。

(2)分类准确度 $a_c$。对于二分类(正常、异常)任务来说，其结果根据分类结果和真实标签的关系可以分为 4 类，如表 1 所示，该表也被称为混淆矩阵。

表 1　混淆矩阵

Table 1 Confusion matrix

| 真实标签 | 模型预测 | |
|---|---|---|
| | 异常 | 正常 |
| 异常 | $T_P$ | $F_N$ |
| 正常 | $F_P$ | $T_N$ |

在本例的混淆矩阵中，$T_P$ 表示标签为“异常”且分类模型也将其分类为“异常”的样本数；$F_N$ 表示标签为“异常”而分类模型将其分类为“正常”的样本数；$F_P$ 表示标签为“正常”而分类模型将其分类为“异常”的样本数；$T_N$ 表示标签为“正常”且分类模型也将其分类为“正常”的样本数。

显然，初步筛选后的风电机组 SCADA 数据中正常数据的样本数应远远多于异常数据的样本数，即两类数据存在严重的不平衡问题。$F_1$ 被定义为精准率和召回率的调和平均数，被广泛用于评估不平衡数据下分类模型的性能，其计算公式如下

$$F_1=2T_P/(2T_P+F_P+F_N) \tag{3}$$

本文中取 $F_1$ 指标作为分类准确度。$F_1$ 的取值范围为 0 到 1，其越接近于 1，说明分类模型的性能越好，也说明该参数组合下聚类得到的数据标签更准确。

异常类别占比 $r_a$ 表示该聚类参数下，聚类后异常类别的数据占总数据的比例

$$r_a=\frac{n_a}{n_o} \tag{4}$$

式中：$n_a$ 表示某一组聚类参数下使用 DBSCAN 聚类方法识别出来的异常样本数；$n_o$ 表示初步筛选后的样本总数。

不同参数组合的 3 个评价指标计算结果如表 2 所示。

表 2　不同参数组合的评价指标计算结果

Table 2 Evaluation indicator results for different parameters

| 参数组合 $(\varepsilon,n_p)$ | 预测误差 $e_{pn}/10^{-3}$ | 分类准确度 $a_c$ | 异常类别占比 $r_a$ |
|---|---|---|---|
| (0.02,4) | 5.956 | 0.434 | 0.034 |
| (0.02,6) | 5.940 | 0.342 | 0.040 |
| (0.02,8) | 5.326 | 0.319 | 0.046 |
| (0.02,10) | 5.217 | 0.296 | 0.053 |
| (0.02,12) | 4.941 | 0.241 | 0.060 |
| (0.04,4) | 6.875 | 0.596 | 0.015 |
| (0.04,6) | 6.866 | 0.611 | 0.016 |
| (0.04,8) | 6.635 | 0.625 | 0.018 |
| (0.04,10) | 6.593 | 0.583 | 0.020 |
| (0.04,12) | 6.686 | 0.602 | 0.023 |
| (0.06,4) | 7.716 | 0.683 | 0.009 |
| (0.06,6) | 7.720 | 0.710 | 0.010 |
| (0.06,8) | 7.491 | 0.697 | 0.011 |
| (0.06,10) | 7.235 | 0.688 | 0.013 |
| (0.06,12) | 7.131 | 0.697 | 0.013 |
| (0.08,4) | 8.466 | 0.748 | 0.007 |
| (0.08,6) | 7.925 | 0.783 | 0.007 |
| (0.08,8) | 8.092 | 0.769 | 0.009 |
| (0.08,10) | 7.679 | 0.744 | 0.009 |
| (0.08,12) | 7.909 | 0.699 | 0.009 |
| (0.10,4) | 8.344 | 0.804 | 0.006 |
| (0.10,6) | 8.265 | 0.754 | 0.006 |
| (0.10,8) | 8.039 | 0.781 | 0.007 |
| (0.10,10) | 7.909 | 0.800 | 0.007 |
| (0.10,12) | 7.630 | 0.792 | 0.007 |

预测误差 $e_{pn}$ 从风速-功率映射关系角度来比较不同的参数组合，该指标最小时的参数组合为：邻域半径等于 0.02 且邻域最小相邻点数等于 12，但是该参数组合下的异常类别占比 $r_a$ 远大于其他的参数组合，说明该聚类参数组合下的 DBSCAN 模型在划分正常数据时更为“严格”，倾向于将更多的样本点聚类到异常类别中，该参数组合下采用 DBSCAN 模型对 32 号风电机组进行异常识别的结果如图 6 所示。这个参数组合下的 DBSCAN 模型将很多实际上是正常的数据也划分到了异常类别，尽管正常类别中正常数据的占比很高，但是浪费了大量的正常数据，不利于后续状态分析等相关研究。

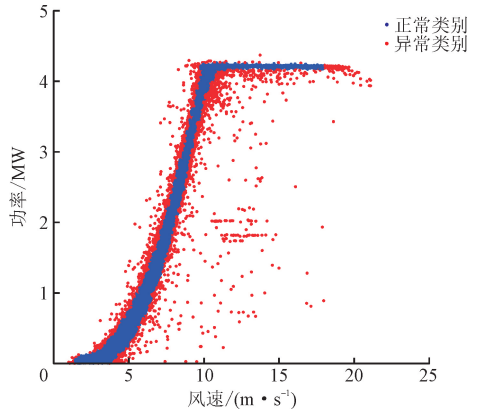


图 6 参数组合为 $\varepsilon=0.2$、$n_p=12$ 时采用 DBSCAN 模型对 32 号风电机组进行异常数据识别的结果

Fig. 6 Result of abnormal data recognition using DBSCAN for the 32nd turbine when $\varepsilon=0.2$ and $n_p=12$

按照 $e_r$ 从小到大的顺序排列备选聚类参数，不同参数组合按递增的排序结果如图 7 所示。

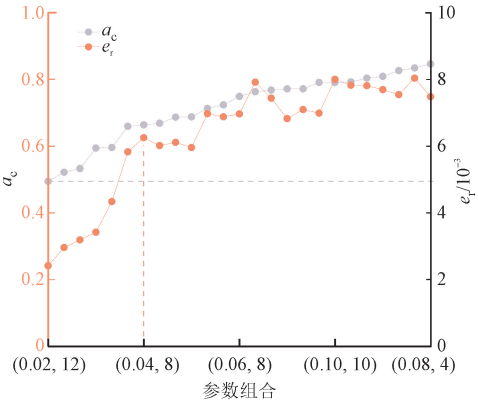


图 7 不同参数组合按 $e_r$ 递增的排序结果

Fig. 7 Sorting results in increasing order of $e_r$

从图 7 中可以看出，随着预测误差开始增大，分类准确度也开始上升，取分类准确度第一个极大值点所对应的参数组合(0.04,8)作为 DBSCAN 的最佳参数组合。使用该聚类参数组合对 32 号风电机组进行异常识别的结果如图 8 所示，从图 8 中可以看出，正常类别数据与图 1 所示的理想功率特性曲线趋势非常相近，同时没有如图 6 中那样浪费大量的正常数据，说明异常数据识别的效果比较好。

为了说明方法具有一定的通用性，选取同一风场 29 号风机一年的 SCADA 监测数据进行验证，原始数据长度为 51 101，采样间隔为 10 min。实验流程和相关设置与前文相同，基于本文提出的改进 DBSCAN 进行异常识别的结果如图 9 所示。从图 9 中可以看出，29 号风电机组的异常数据更加密集，也更贴近于正常数据，本文提出的方法对于该机组的监测数据仍有较好的异常识别效果。

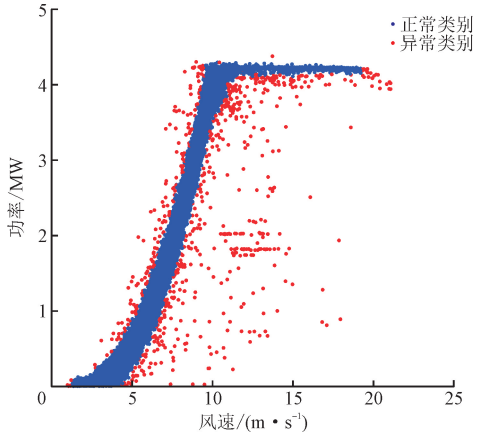


图 8 参数组合为 $\varepsilon=0.04$、$n_p=8$ 时采用 DBSCAN 模型对 32 号风电机组进行异常数据识别的结果

Fig. 8 Result of abnormal data recognition using DBSCAN for the 32nd turbine when $\varepsilon=0.04$ and $n_p=8$

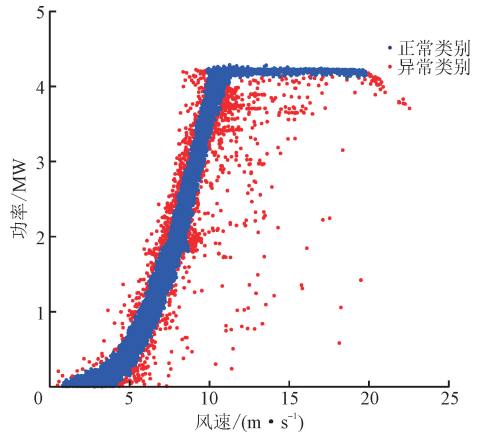


图 9 改进 DBSCAN 模型对 29 号机组异常数据的识别结果

Fig. 9 Recognition result of abnormal data by improved DBSCAN model for 29th turbine

### 3.2 对比验证

为了进一步验证所提方法的优越性，基于 32 号风电机组的监测数据，采用 $k$-dist 图法[21]和 KANN-DBSCAN 法[23]来进行对比实验。

#### 3.2.1 $k$-dist 图法

为了保证对比的一致性，对于 $k$-dist 图法，$k$ 也从{4,6,8,10,12}中选取，由初步筛选后的 SCADA 数据得到的 $k$-dist 图如图 10 所示。因为样本量很大，为了清晰地表达 $k$-dist 曲线的变化趋势，图 10 中只显示了降序排列后的前 200 个值。

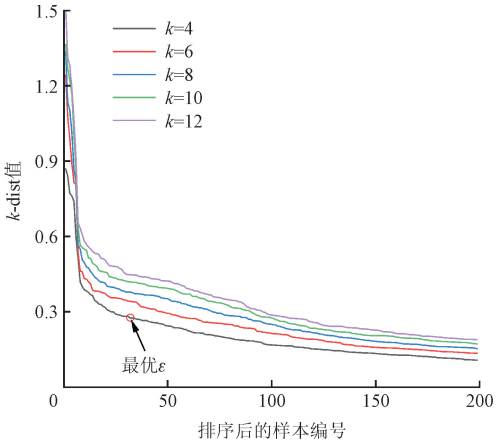


图 10　SCADA 数据的排序 $k$-dist 图

Fig. 10　Sorted $k$-dist graph for SCADA data

从图 10 中可以看出，$k=4$ 时的变化曲线基本可以反映出 $k$ 为其他值时曲线的变化趋势，这与文献[21]中的结论相符。由于更大的 $k$ 意味着聚类时需要更高的计算成本[21]，因此，选择 $k=4$ 时曲线的第一个“拐点”对应的 $k$-dist 值作为最优的 ε。根据 $k$-dist图，最终确定聚类参数为 $\varepsilon=0.27$、$n_p=4$，以该聚类参数对 32 号电机组进行异常数据识别的结果如图 11 所示。

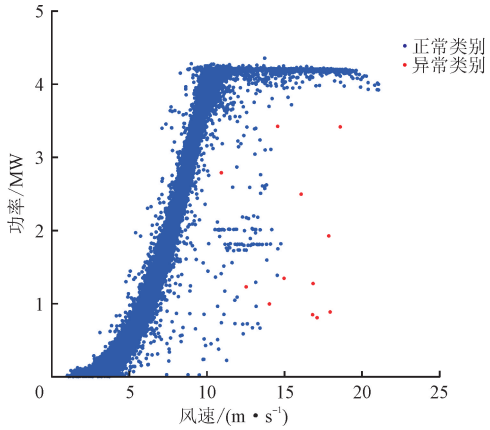


图 11　参数组合为 $\varepsilon=0.27$、$n_p=4$ 时采用 DBSCAN 模型对 32 号风电机组进行异常数据识别的结果

Fig. 11　Result of abnormal data recognition using DBSCAN for the 32nd turbine when $\varepsilon=0.27$ and $n_p=4$

由图 11 可知，只有少部分异常数据被识别出来，大部分离散型异常数据和中部堆积型异常数据并没有被识别出来，识别效果较差，说明基于 $k$-dist 图的 DBSCAN 聚类参数选择不适用于风电机组 SCADA 异常数据识别任务。此外，基于 $k$-dist 图的参数选择过程中需要通过人的介入[25]来选择“拐点”，这一方面引入了人的主观性，当 $k$-dist 曲线变化趋势平缓时这种主观性对参数的选取影响更大，另一方面导致该过程无法完全自动进行，而本文所提出的方法在这两个方面都具有优势。

### 3.2.2　KANN-DBSCAN 法

KANN-DBSCAN 法的基本过程如下[23]。

**步骤 1**　计算数据集 $\boldsymbol{D}$ 的距离分布矩阵

$$\boldsymbol{D}_{n\times n}=\{d_{i,j}\mid 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant n\} \tag{5}$$

式中：$\boldsymbol{D}_{n\times n}$ 为实对称矩阵，$n$ 为数据集中样本的个数；$d_{i,j}$ 为第 $i$ 个对象与第 $j$ 个对象之间的距离；

**步骤 2**　对距离分布矩阵 $\boldsymbol{D}_{n\times n}$ 的每一行按升序排列，排列后第 $k$ 列元素构成所有对象的 $k$-最近邻距离向量 $\boldsymbol{D}_k$；

**步骤 3**　对 $\boldsymbol{D}_{n\times n}$ 每一列的元素 $\boldsymbol{D}_k$ 求平均值 $\bar{\boldsymbol{D}}_k$，将其作为 ε 的候选值。对所有 $k$ 值进行计算，可以得到 ε 候选值列表

$$\boldsymbol{D}_\varepsilon=\{\bar{\boldsymbol{D}}_k\mid 1\leqslant k\leqslant n\} \tag{6}$$

**步骤 4**　对于每个 $k$ 值，选取 $\boldsymbol{D}_\varepsilon$ 中对应的候选 ε，按下式计算该 ε 下的 $n_p$

$$n_p=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{7}$$

式中：$P_i$ 为第 $i$ 个对象在 ε-邻域中的邻域样本数；$n$ 为数据集中的总样本数。将这两个参数代入 DBSCAN 模型对监测数据进行聚类，得到该 $k$ 值下的聚类簇数。当连续 3 次的聚类簇数相同时，认为聚类结果趋于稳定，记录该聚类簇数 $N$ 为最优簇数；

**步骤 5**　继续执行步骤 4，当聚类簇数第一次不再为 $N$ 时，上一个 $k$ 值对应的 ε 和 $n_p$ 为最佳聚类参数。

按照上述步骤对 32 号风电机组的监测数据进行分析，计算得到最优簇数 $N$ 为 7，最佳的聚类参数为 $\varepsilon=0.0216$，$n_p=86$，以该聚类参数对 32 号风电机组进行异常数据识别的结果如图 12 所示。

图 12 中黑色被认为是噪声，其余颜色分别代表某个类别。显然，KANN-DBSCAN 方法选择的参数在进行聚类时，更加关注正常数据中密度不同的区域，而将“外围”的正常数据与异常数据一同划分为了噪声。因此，无论怎样合并不同的类别和噪声，都无法将离散分布型和中部堆积型数据分离出来，即无法实现有效的异常数据识别。

通过上述对比实验可知，相比于 $k$-dist 图法和 KANN-DBSCAN 法，通过本文提出的方法选择出来的聚类参数在对原始数据进行 DBSCAN 聚类时，对异常数据的分离效果更好，且整个过程可以自动进行，说明了本文提出方法的优越性。

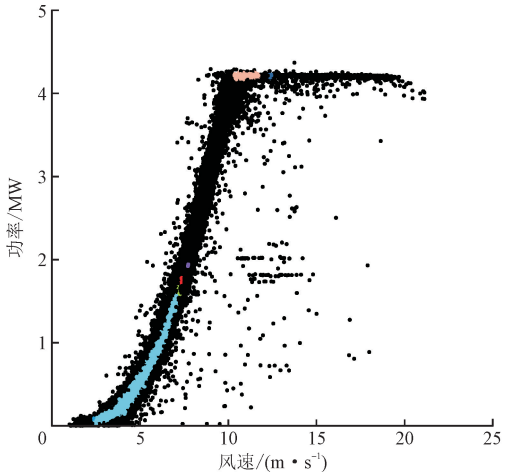


图 12 参数组合为 $\varepsilon=0.021\,6$、$n_p=86$ 时采用 DBSCAN 模型对 32 号风电机组进行异常数据识别的结果

Fig. 12 Result of abnormal data recognition using DBSCAN for the 32nd turbine when $\varepsilon=0.021\,6$ and $n_p=86$

### 3.3 算法复杂度分析

SCADA 异常数据识别实际上只用到了风速和功率这两个变量的数据，即需要聚类的数据为二维数据。设 $n$ 为原始数据的样本点数，则 DBSCAN 模型对二维数据进行聚类时的基本时间复杂度为 $O(n^2)$。KANN-DBSCAN 在 DBSCAN 算法的基础上进行迭代运算，聚类次数由 $k$ 决定，因此其时间复杂度为 $O(kn^2)$，一般情况下 $k\ll n$。

本文提出的算法同样需要多次进行 DBSCAN 模型聚类，聚类次数由备选参数组合的个数 $m$ 决定，聚类过程的时间复杂度为 $O(mn^2)$，通常 $m\ll n$；此外，每次聚类需要计算预测误差和分类准确度两个评价指标，这部分的计算复杂度与模型的选择相关。以本文使用的两层 MLP 为例，其时间复杂度为 $O(nmh^2oi)$，其中 $m$ 为输入维数，$h$ 为每层神经元数，$o$ 为输出维数，$i$ 为每个变量的迭代次数，在本文的实验中 $mh^2oi\approx n$，因此，计算评价指标的时间复杂度约为 $O(n^2)$。

综上所述，本文所提出方法时间复杂度约为 $O(n^2)$，算法复杂度与 KANN-DBSCAN 模型相当，虽然相较于传统的 DBSCAN 算法略高，但是仍然属于同一数量级，且对风电机组 SCADA 异常数据的识别效果更好。

## 4 结 论

本文提出的基于改进 DBSCAN 的风电机组 SCADA 异常数据识别方法，结合风电机组的风速-功率数据分布特点，提出了两个指标，以一种简单而有效的方式来选择合适的聚类参数，能够保证异常数据被剔除的前提下，保留尽可能多的正常数据。以某风场 32 号和 29 号风电机组作为研究实例进行异常数据识别，结果表明本文提出的方法能够有效地将异常数据识别出来。此外，与广泛使用的 $k$-dist 图法以及改进的 KANN-DBSCAN 方法相比，本文提出的参数选择方法受主观性影响小，而且整个过程可以通过程序自动进行，但是算法的时间复杂度较高，如何提高算法的计算效率是后续研究的重点。

**参考文献：**

[1] 符杨，许伟欣，刘璐洁，等. 考虑天气因素的海上风电机组预防性机会维护策略优化方法 [J]. 中国电机工程学报，2018，38(20)：5947-5956.
FU Yang，XU Weixin，LIU Lujie，et al. Optimization of preventive opportunistic maintenance strategy for offshore wind turbine considering weather conditions [J]. Proceedings of the CSEE，2018，38(20)：5947-5956.

[2] 胡姚刚. 大功率风电机组关键部件健康状态监测与评估方法研究 [D]. 重庆：重庆大学，2017.

[3] 尹诗，侯国莲，于晓东，等. 基于 SCADA 数据的风电机组齿轮箱状态监测方法 [J]. 太阳能学报，2021，42(1)：324-332.
YIN Shi，HOU Guolian，YU Xiaodong，et al. Condition monitoring method of wind turbine gear box based on SCADA data [J]. Acta Energiae Solaris Sinica，2021，42(1)：324-332.

[4] DAO P B. Condition monitoring and fault diagnosis of wind turbines based on structural break detection in SCADA data [J]. Renewable Energy，2022，185：641-654.

[5] RAHIMILARKI R，GAO Zhiwei，JIN Nanlin，et al. Convolutional neural network fault classification based on time-series analysis for benchmark wind turbine machine [J]. Renewable Energy，2022，185：916-931.

[6] 江国乾，周俊超，武鑫，等. 基于空洞因果卷积网络的风电机组异常检测 [J]. 太阳能学报，2023，44(5)：368-375.
JIANG Guoqian，ZHOU Junchao，WU Xin，et al. Wind turbine anomaly detection based on dilated causal convolution network [J]. Acta Energiae Solaris Sinica，2023，44(5)：368-375.

[7] 郭怡，王荣喜，高建民. 融合分形特征的风机运行状态辨识方法 [J]. 计算机集成制造系统，2022，28(7)：2139-2148.
GUO Yi，WANG Rongxi，GAO Jianmin. Operation

state recognition method based on fractal features of wind turbines [J]. Computer Integrated Manufacturing Systems, 2022, 28(7): 2139-2148.

[8] MORRISON R, LIU Xiaolei, LIN Zi. Anomaly detection in wind turbine SCADA data for power curve cleaning [J]. Renewable Energy, 2022, 184: 473-486.

[9] 吴永斌，张建忠，袁正舸，等. 风电场风功率异常数据识别与清洗研究综述 [J]. 电网技术，2023，47(6)：2367-2380.
WU Yongbin, ZHANG Jianzhong, YUAN Zhengxi, et al. Review on identification and cleaning of abnormal wind power data for wind farms [J]. Power System Technology, 2023, 47(6): 2367-2380.

[10] WANG Yue, INFIELD D G, STEPHEN B, et al. Copula-based model for wind turbine power curve outlier rejection [J]. Wind Energy, 2014, 17(11): 1677-1688.

[11] HAN Shuang, QIAO Yanhui, YAN Ping, et al. Wind turbine power curve modeling based on interval extreme probability density for the integration of renewable energies and electric vehicles [J]. Renewable Energy, 2020, 157: 190-203.

[12] 沈小军，付雪姣，周冲成，等. 风电机组风速-功率异常运行数据特征及清洗方法 [J]. 电工技术学报，2018，33(14)：3353-3361.
SHEN Xiaojun, FU Xuejiao, ZHOU Chongcheng, et al. Characteristics of outliers in wind speed-power operation data of wind turbines and its cleaning method [J]. Transactions of China Electrotechnical Society, 2018, 33(14): 3353-3361.

[13] SHEN Xiaojun, FU Xuejiao, ZHOU Chongcheng. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm [J]. IEEE Transactions on Sustainable Energy, 2019, 10(1): 46-54.

[14] 梅勇，李霄，胡在春，等. 基于风电机组控制原理的风功率数据识别与清洗方法 [J]. 动力工程学报，2021，41(4)：316-322.
MEI Yong, LI Xiao, HU Zaichun, et al. Identification and cleaning of wind power data methods based on control principle of wind turbine generator system [J]. Journal of Chinese Society of Power Engineering, 2021, 41(4): 316-322.

[15] LONG Huan, SANG Linwei, WU Zaijun, et al. Image-based abnormal data detection and cleaning algorithm via wind power curve [J]. IEEE Transactions on Sustainable Energy, 2020, 11(2): 938-946.

[16] WANG Zhongju, WANG Long, HUANG Chao. A fast abnormal data cleaning algorithm for performance evaluation of wind turbine [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12.

[17] XU Qianyao, HE Dawei, ZHANG Ning, et al. A Short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining [J]. IEEE Transactions on Sustainable Energy, 2015, 6(4): 1283-1291.

[18] ZHAO Yongning, YE Lin, WANG Weisheng, et al. Data-driven correction approach to refine power curve of wind farm under wind curtailment [J]. IEEE Transactions on Sustainable Energy, 2018, 9(1): 95-105.

[19] 王一妹，刘辉，宋鹏，等. 基于多阶段递进识别的风电机组异常运行数据清洗方法 [J]. 可再生能源，2020，38(11)：1470-1476.
WANG Yimei, LIU Hui, SONG Peng, et al. An approach for the cleaning of abnormal wind turbine operation data based on multi-phase progressive recognition [J]. Renewable Energy Resources, 2020, 38(11): 1470-1476.

[20] 雷萌，郭鹏，刘博嵩. 基于自适应 DBSCAN 算法的风电机组异常数据识别研究 [J]. 动力工程学报，2021，41(10)：859-865.
LEI Meng, GUO Peng, LIU Bosong. Study on abnormal data recognition of wind turbines based on adaptive DBSCAN algorithm [J]. Journal of Chinese Society of Power Engineering, 2021, 41(10): 859-865.

[21] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]//Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Palo Alto, CA, USA: AAAI Press, 1996: 226-231.

[22] 夏鲁宁，荆继武. SA-DBSCAN：一种自适应基于密度聚类算法 [J]. 中国科学院研究生院学报，2009，26(4)：530-538.
XIA Luning, JING Jiwu. SA-DBSCAN: a self-adaptive density-based clustering algorithm [J]. Journal of the Graduate School of the Chinese Academy of Sciences, 2009, 26(4): 530-538.

[23] 李文杰，闫世强，蒋莹，等. 自适应确定 DBSCAN 算法参数的算法研究 [J]. 计算机工程与应用，2019，55(5)：1-7.
LI Wenjie, YAN Shiqiang, JIANG Ying, et al. Research on method of self-adaptive determination of DBSCAN algorithm parameters [J]. Computer Engineering and Applications, 2019, 55(5): 1-7.

[24] 周红芳，王鹏. DBSCAN 算法中参数自适应确定方法的研究 [J]. 西安理工大学学报，2012，28(3)：289-292.

ZHOU Hongfang, WANG Peng. Research on adaptive parameters determination in DBSCAN algorithm [J]. Journal of Xi'an University of Technology, 2012, 28(3): 289-292.

[25] 万佳，胡大裟，蒋玉明. 多密度自适应确定 DBSCAN 算法参数的算法研究 [J]. 计算机工程与应用，2022，58(2)：78-85.

WAN Jia, HU Dasha, JIANG Yuming. Research on method of multi-density self-adaptive determination of DBSCAN algorithm parameters [J]. Computer Engineering and Applications, 2022, 58(2): 78-85.

（编辑　武红江）