基于改进 K-means 聚类 k 值选择算法的配网电压数据异常检测

刘明群，何　鑫，覃日升，姜　訸，孟　贤

（云南电网有限责任公司电力科学研究院，云南 昆明 650217）

摘　要：K-means 聚类算法因计算速度快、准确率高等优势被应用于大规模配电网数据异常检测，但当聚类数不合适时，可能导致聚类结果不理想。为此，提出一种基于改进 elbow method 和轮廓系数的聚类数选择算法 IES，首先，该算法利用 elbow method 的聚类评价指标和聚类数上限，确定随数据集不同而自适应变化的阈值，通过自适应阈值求解聚类数下限；其次，在聚类数上下限内计算轮廓系数，并提出“一个极大值”规则避免计算所有轮廓系数，提高算法速度；最后，利用轮廓系数选取合适聚类数，并通过召回率评价异常检测效果，说明为 K-means 聚类算法选取合适聚类数对异常检测的重要性。算例结果表明：IES 算法能在自适应获取最佳聚类数的同时大大削减计算时间，提高 K-means 算法在线监测的准确率和高效性。

关　键　词：配电网电压；在线监测；K-means 聚类算法；最佳聚类数

DOI：10.19781/j.issn.1673-9140.2022.06.010　　**中图分类号**：TM76　　**文章编号**：1673-9140(2022)06-0091-09

Anomaly detection of distribution network voltage data based on improved K-means clustering k-value selection algorithm

LIU Mingqun，HE Xin，QIN Risheng，JIANG He，MENG Xian

（Electric Power Science Research Institute，Yunnan Power Grid Co.，Ltd.，Kunming 650217，China）

Abstract：K-means clustering algorithm has been applied to anomaly detection of large-scale distribution network data due to its advantages of fast computation speed and high accuracy. However，the algorithm may lead to an inaccurate clustering if the assumed clustering number is not appropriate. Therefore，this paper presents a clustering number selection algorithm IES based on the improved elbow method and silhouette coefficient (IES). Firstly，the clustering evaluation index of the elbow method and the upper limit of clustering number are utilized to set a threshold which can adaptively change with data sets. With this threshold，the lower limit of clustering number can be obtained. Secondly，the silhouette coefficient calculated within the upper and lower limit of the clustering number. An “one maximum” rule is proposedin order to improve the algorithm speed and avoid calculating all the silhouette coefficients. In the end，the calculated silhouette coefficients are utilized to select the appropriate clustering number. In addition，the recall rate is employed to evaluate the anomaly detection and illustrate the importance of selecting appropriate clustering number for K-means anomaly detection. Simulation results show that the IES algorithm can obtain the optimal clustering number adaptively，meantime，greatly shorten the calculation time，and improve the accuracy and efficiency of the K-

收稿日期：2021-05-07；**修回日期**：2021-09-11

基金项目：中国南方电网有限责任公司科技项目(YNKJXM20191369)

通信作者：覃日升(1976—)，男，硕士，教授级工程师，主要从事配网无功电压研究；E-mail：qinrisheng2020@126.com

means algorithm in online monitoring.

Key words: distribution voltage; online monitoring; K-means clustering algorithm; optimal number of clusters

近年来,配电网电压监测系统建设加强了对配电网电压、电流等数据的管理与计算分析,实现了电网故障预警、电压准实时在线监测等功能[1]。然而,由于配电网规模和配电网监测数据日益增大,传统在线监测方法已无法满足监测系统对数据挖掘的快速性和准确性要求。因此,为保障配电网安全可靠运行,配电网电压监测系统数据异常检测方法研究具有重要意义。

聚类算法通过对配电网电压数据聚类,挖掘数据特征、区分数据类型、实现实时故障预警。聚类算法可分为划分聚类法、层次聚类法和密度聚类法等,其中划分聚类法计算效率较高,但需事先假定聚类数值[2],包括 K-means、K-medoids 和 CLARA 算法。特别地,K-means 算法原理简单且效率高,非常适合配电网电压数据在线监测;K-means 算法仅在多项式时间内收敛到局部最优解[3],但合理地选择初值将有利于收敛到全局最优解。文献[4]采用 K-means++算法选取 K-means 算法初值,但运用 K-means 算法要事先假定聚类数。聚类数可根据样本情况判断[5],然而,在相关经验不足或数据量过大等情况下无法给出最佳聚类数。对于静态数据库而言,聚类数不会改变,但当在线监测系统数据库是动态时,聚类数会随着配电网故障等问题的产生而动态变化,导致 K-means 算法的聚类效果变差[6],因而需要其他算法辅助选择聚类数。

为解决聚类数选择问题,目前已有学者提出多种聚类数选择算法。文献[7]提出轮廓系数法,算法原理简单且只需给定聚类数上限,对最佳聚类数估计效果很好,但计算速度较慢,不适合在线监测配电网电压;文献[8]运用 DBI 算法为 K-means 算法选取聚类数,DBI 算法对最佳聚类数估计效果较好,但计算速度稍慢;文献[9]运用 Canopy 算法快速自适应选取聚类数,但该算法需根据交叉验证法或先验知识设定松阈值和紧阈值,且紧阈值严重影响聚类数的选取,因此算法自适应能力不够强;文献[10-11]运用 elbow method 选取聚类数,该方法因简单直观且计算速度快而被广泛应用,但常常因“肘部点”不明显而无法估计最佳聚类数。上述常用算法存在自适应能力不强、计算速度慢和准确率不够高等问题,不适合异常数据在线检测。

针对上述常用算法问题,本文提出一种快速选取聚类数的自适应算法。所提算法基于 elbow method 和轮廓系数法,首先利用自适应变化阈值求解聚类数下限,接着在聚类数上、下限内计算轮廓系数。为提高算法速度,提出“一个极大值”规则,避免计算所有轮廓系数。该算法充分考虑 elbow method 的快速性和轮廓系数法的高准确率特点,为 K-means 算法自动选取聚类数,使 K-means 算法在线监测成为可能。最后,为评价所提算法,以 2 个实际配网电压数据为例,通过仿真对比其他聚类数选择算法。结果表明,相比于所对比算法,所提算法能以最高准确率和最快计算速度自适应选取最佳聚类数。

1 K-means 聚类算法

K-means 是一种基于划分的无监督聚类算法,能将数据集分成 k 类,其中 k 是事先假定的。K-means 算法随机产生 k 个聚类中心,根据最近邻原则将数据点归类离其最近的聚类中心,形成 k 个类,并重新计算各类的聚类中心,重复上述步骤直到聚类中心不再改变位置或达到规定的迭代次数。

K-means 算法聚类的目标是使各类误差平方和(sum of the squared errors, SSE)最小,即

$$e_{\mathrm{SSE}}=\sum_{i=1}^{k}\sum_{p\in C_i}||p-m_i||^2 \tag{1}$$

式中 C_i 为第 i 个聚类;p 为 C_i 中的样本点;m_i 为 C_i 的聚类中心,即 C_i 中所有样本的均值;e_{SSE} 为所有样本的聚类误差,代表聚类效果的好坏。

根据拉格朗日定理和最小二乘法原理,确保 SSE 最小的聚类中心[3]应满足:

$$\begin{aligned}\frac{\partial e_{\mathrm{SSE}}}{\partial m_i}&=\frac{\partial}{\partial m_i}\Big[\sum_{i=1}^{k}\sum_{p\in C_i}(p-m_i)^2\Big]=\\&\sum_{i=1}^{k}\sum_{p\in C_i}\frac{\partial}{\partial m_i}(p-m_i)^2=\\&\sum_{p\in C_i}2(p-m_i)=0\end{aligned} \tag{2}$$

由式(2)可得:

$$m_i=\frac{1}{n_i}\sum_{p\in C_i}p \tag{3}$$

其中，n_i 为第 i 聚类的样本总量。因此，聚类中心是该聚类数据的平均值。K-means 算法每一次迭代将聚类中心取为该聚类样本的平均值，确保 SSE 在本次迭代内达到最小，交替采用最近邻原则和以均值计算聚类中心，使 SSE 不断下降，直到平衡收敛。

2　基于 elbow method 和轮廓系数的聚类数自适应确定

结合 elbow method 和轮廓系数，提出改进的 elbow method 和轮廓系数算法（improved elbow method and silhouette coefficient，IES），用于自适应确定聚类数，从而与 K-means 算法结合为自适应 K-means 算法。对于假定的聚类数上限 $k_{\max}$，首先，IES 算法基于 elbow method 的聚类评价指标 SSE 值确定聚类数下限 $k_{\min}$；然后，在聚类数搜索范围 $[k_{\min},k_{\max}]$ 内基于轮廓系数搜寻最佳聚类数 k^*，并利用提出的“一个极大值”规则避免计算 $[k_{\min},k_{\max}]$ 内所有聚类数对应的轮廓系数。当最佳聚类数确定后，可用 K-means 算法进行聚类。

2.1　改进 elbow method

IES 算法的核心是在一定的定义域内利用轮廓系数寻找最佳聚类数。使轮廓系数最大的聚类数为最佳聚类数 k^*。轮廓系数法一般在定义域 $[2,k_{\max}]$ 内计算每一个聚类数 k 对应的轮廓系数，但轮廓系数计算速度慢，不满足配电网在线监测的快速性要求。因此，IES 算法利用计算速度更快的 SSE 确定聚类数下限 $k_{\min}$，将定义域范围缩小为 $[k_{\min},k_{\max}]$。

聚类数为 k 时 elbow method 的误差平方和记为 $e_{\mathrm{SSE}}(k)$。$k=i$ 时的相对 SSE 定义为

$$e_{\mathrm{SSE}}(i)\%=\frac{e_{\mathrm{SSE}}(i)}{e_{\mathrm{SSE}}(1)}\times 100\% \tag{4}$$

式中　$e_{\mathrm{SSE}}(1)$ 为 $k=1$ 时 $e_{\mathrm{SSE}}(k)$ 的值，也是其最大值。

由于 $e_{\mathrm{SSE}}(k)$、相对 SSE 单调递减且离散不可导，无法用极值点和拐点估计最佳聚类数，因此考虑设置阈值来估计聚类数下限。同时，由于不同数据集的手肘曲线不同，因而所设置阈值应随之自适应变化。

定义取值范围为 $(0,1)$ 的常数 K_{elbow}，称为手肘系数。随着 k 的增大，相对 SSE 从 $e_{\mathrm{SSE}}(1)\%$ 开始下降，当 $k=k_{\max}$ 时，相对 SSE 最大下降量为 $e_{\mathrm{SSE}}(1)\%-e_{\mathrm{SSE}}(k_{\max})\%$，而当相对 SSE 下降量达到最大下降量的 K_{elbow} 倍时，对应的相对 SSE 定义为手肘阈值 $e_{\mathrm{SSE,elbow}}\%$，可计算如下：

$$\begin{aligned}e_{\mathrm{SSE,elbow}}\%&=e_{\mathrm{SSE}}(1)\%-K_{\mathrm{elbow}}(e_{\mathrm{SSE}}(1)\%-e_{\mathrm{SSE}}(k_{\max})\%)\\&=100\%-K_{\mathrm{elbow}}\cdot(100\%-e_{\mathrm{SSE}}(k_{\max})\%)\end{aligned} \tag{5}$$

根据大量仿真经验，取 $K_{\mathrm{elbow}}=0.5$，则

$$e_{\mathrm{SSE,elbow}}\%=\frac{100\%+e_{\mathrm{SSE}}(k_{\max})\%}{2} \tag{6}$$

利用式(6)，将聚类数下限 $k_{\min}$ 取为使 $e_{\mathrm{SSE}}(k)\%\leqslant e_{\mathrm{SSE,elbow}}\%$ 成立的最小 k 值。

在聚类数上限为 $k_{\max}$ 以及 $K_{\mathrm{elbow}}=0.5$ 时，$k_{\max}$ 对应人为限定的相对 SSE 最大下降量，所选的 $k_{\min}$ 使相对 SSE 下降量达到最大下降量的一半以上。应指出，IES 算法解出的 $k_{\min}\geqslant 2$。

手肘阈值具有自适应性，因为当 $k>k^*$ 后 $e_{\mathrm{SSE}}(k)\%$ 变化缓慢，即使 $k_{\max}$ 较大，$e_{\mathrm{SSE}}(k_{\max})\%$ 仍接近于 $e_{\mathrm{SSE}}(k^*)\%$，而不同数据集的 $e_{\mathrm{SSE}}(k^*)\%$ 不同，因此，由 $e_{\mathrm{SSE}}(k_{\max})\%$ 计算的手肘阈值也随之自适应变化。对不同数据集，当 $k_{\max}$ 等于样本总数时，均有 $e_{\mathrm{SSE}}(k_{\max})=e_{\mathrm{SSE}}(k_{\max})\%=0$[3]，由式(6)可知，手肘阈值达到其最小值 $e_{\mathrm{SSE,elbow}}\%=50\%$，手肘阈值恒为常数且不随数据集不同而变化，因此 $k_{\max}$ 可取值较大，但不应过大。

不同 K_{elbow} 取值影响 IES 算法计算时间和准确率（能否解出 k^*）。若降低 K_{elbow}，手肘阈值将增大，使得解出的 $k_{\min}$ 变小，有利于 IES 算法解出 k^*（若 $k_{\min}>k^*$，则 IES 算法无法解出 k^*），但 $k_{\min}$ 变小又会使得轮廓系数法的定义域 $[k_{\min},k_{\max}]$ 范围变大，不利于缩短轮廓系数法计算时间。同样分析 K_{elbow} 增大的情况，可知 K_{elbow} 不应过大或过小。由此也能看出，手肘阈值随数据集不同而自适应增大或减小，是在自适应兼顾 IES 算法对不同数据集的准确率和计算速度。实际应用中可根据配网电压历史数据进行测试，适当调整 K_{elbow} 取值。改进 elbow method 确定聚类数下限流程如图 1 所示。

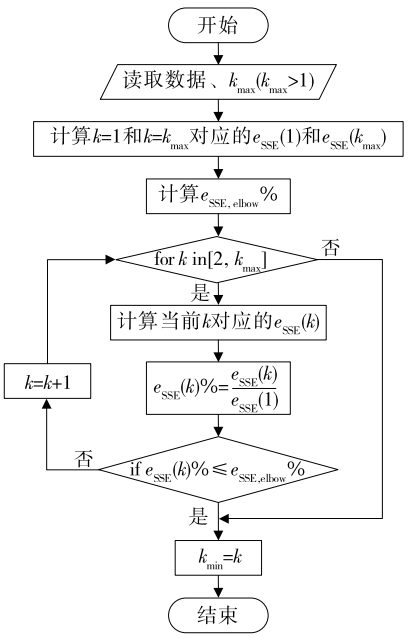

图 1 改进 elbow method 确定聚类数下限流程

Figure 1 The flowchart of determinizing the lower limit of clustering number with the improved elbow method

具体步骤如下：

1)读取人为设定的 k_{max}和数据集；

2)对 $k=1$、k_{max}分别进行 K-means 聚类，对聚类结果应用式(1)计算 SSE，即 $e_{SSE}(1)$、$e_{SSE}(k_{max})$；

3)结合式(4)、(6)计算 $e_{SSE,elbow}$%，令 $k=2$；

4)对当前 k 值进行 K-means 聚类并计算 $e_{SSE}(k)$；

5)采用式(4)计算 $e_{SSE}(k)$%；

6)判断 $e_{SSE}(k)\% \leqslant e_{SSE,elbow}\%$ 是否成立，不成立则令 k 自增 1 并返回步骤 4)，成立则跳出循环进入步骤 7)；

7)记录此时的 k 为 k_{min}，该流程结束。

2.2 “一个极大值”规则

聚类数为 k 时相应的轮廓系数记为 $S(k)$。基于轮廓系数搜寻最佳聚类数，在$[k_{min},k_{max}]$区间内利用轮廓系数算法搜寻最佳聚类数。然而，为确保大于最佳聚类数 k^*，聚类数上限 k_{max}的设置可能会过大，而轮廓系数计算速度慢，对$[k_{min},k_{max}]$区间内每一个聚类数 k 计算轮廓系数将消耗大量时间。

为提高算法速度，IES 算法借鉴 gap statistic 算法中“一个标准错误”(1-standard-error)的规则[12](文献[13]也在其他算法中使用该规则)，提出“一个极大值”规则，即令聚类数 k 在$[k_{min},k_{max}]$区间内每次增加 1，依次计算轮廓系数 $S(k)$，当 $S(k)$首次出现极大值时停止计算 $S(k)$。使用该规则得到多个轮廓系数，选其中最大值对应的 k 为最佳聚类数 k^*。当 $S(k)$在定义域$[k_{min},k_{max}]$内不存在极大值时，“一个极大值”规则失效，需计算定义域内所有 $S(k)$，选最大值对应的 k 为 k^*。本文中“一个极大值”规则在 $k=K$ 生效是指：对于 $K>k_{min}$，当 k 增大到 $K+1$ 时，出现 $S(k)$的极大值 $S(K)$，IES 算法停止计算 $S(k)$。$S(K)$为极大值是指：$S(K)>S(K-1)$且 $S(K)>S(K+1)$。

“一个极大值”规则避免计算所有轮廓系数，相当于降低了实际假定的 k_{max}，从而提高算法速度。

2.3 基于轮廓系数自适应确定聚类数

聚类数下限确定后应用“一个极大值”规则，在定义域$[k_{min},k_{max}]$内利用轮廓系数法求解最佳聚类数。在已计算不同聚类数对应的轮廓系数中自动寻找最大轮廓系数，所对应聚类数为最佳聚类数 k^*，从而实现自适应确定聚类数。基于轮廓系数自适应确定聚类数流程如图 2 所示。

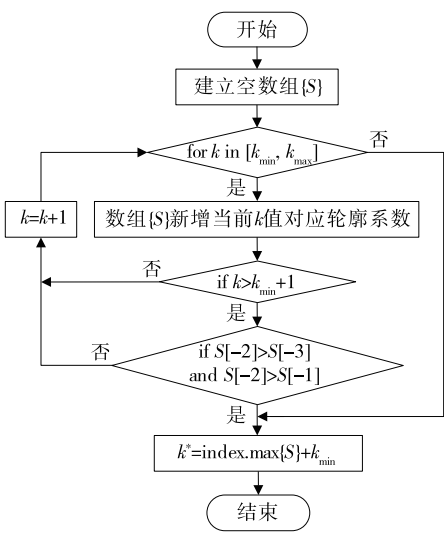

图 2 基于轮廓系数自适应确定聚类数流程

Figure 2 The flowchart of determinizing the clustering number with the improved elbow method based on the silhouette coefficient adaptive determination

具体步骤如下：

1)建立空数组{S}，令 $k=k_{min}$；

2)对当前 k 值进行 K-means 聚类，接着对聚类结果计算轮廓系数并放入数组{S}；

3)若数组{S}中的元素已达 3 个及以上，则说

明可以判断是否出现极大值，进入步骤 4)，否则令 k 自增 1 并回到步骤 2)；

4)判断是否出现极大值，即 $S[-2]>S[-3]$、$S[-2]>S[-1]$是否同时成立，其中 $S[-1]$是数组$\{S\}$倒数第 1 个元素，即本次循环计算得到的轮廓系数；$S[-2]$、$S[-3]$分别是倒数第 2、3 个元素，若不出现极大值，令 k 自增 1 并回到步骤 2)，若出现极大值则跳出循环进入步骤 5)；

5)在数组$\{S\}$中寻找最大轮廓系数，并记录对应的聚类数为最佳聚类数 k^*，IES 算法结束。

2.4　基于自适应 K-means 的实时异常检测模型

IES 算法从图 1 流程开始至图 2 流程结束。自适应确定聚类数的 IES 算法与 K-means 算法结合为自适应 K-means 算法。正常运行时配电网电压数据波动范围较稳定，因此，可利用 K-means 算法对正常运行数据聚类并得到聚类中心，通过判断新输入数据到聚类中心距离是否超过距离阈值 H，从而判断数据是否异常。

$H=(h_1,h_2,\cdots,h_k)$表示各聚类的阈值，其中 k 是聚类数，聚类中数据到聚类中心距离的最大值乘以常数 D 作为 H，综合考虑文献[14]、[15]的实验结果，取 $D=1.04$。若某数据 X 到 k 个聚类中心 C_i 距离均超过相应阈值，则判定为异常数据，即异常数据满足：

$$|X-C_i|>h_i,i=1,2,\cdots,k \tag{7}$$

IES 算法能在异常检测中更新正常数据最佳聚类数，并能在发生异常时帮助挖掘异常数据特征。随着历史正常运行数据的不断增多，正常数据的最佳聚类数可能改变，因此，每隔一段时间需用 IES 算法自适应求解并更新最佳聚类数。当发生异常时，在分析数据异常模式之前，为充分利用当前所有异常数据，可通过 IES 算法对当前所有正常和异常数据的最佳聚类数自适应快速求解，然后利用 K-means 算法将异常与正常数据一起聚类，为挖掘异常数据特征和探测异常来源提供信息。除上述基于自适应 K-means 聚类的方法，文献[14]还利用了其他方法分析数据异常模式。

基于自适应 K-means 的实时异常检测总流程如图 3 所示，具体步骤如下：

1)对配电网电压历史正常运行数据进行 K-means 聚类，并根据聚类得到的最优聚类中心和聚类结果更新距离阈值 H；

2)计算新输入数据到各个聚类中心的距离并与距离阈值比较；

3)若新输入数据属于异常数据则标记为异常，否则将其加入历史正常运行数据；当历史正常运行数据新增数量达到一定量时，利用 IES 算法求解并更新最佳聚类数；

4)将异常数据与历史正常运行数据共同作为新的数据集 D_S；

5)IES 算法根据事先假定聚类数上限计算数据集 D_S 的最佳聚类数 k^*；

6)用 K-means 将数据集 D_S 分为 k^* 个聚类；

7)利用 K-means 聚类结果分析数据异常模式。

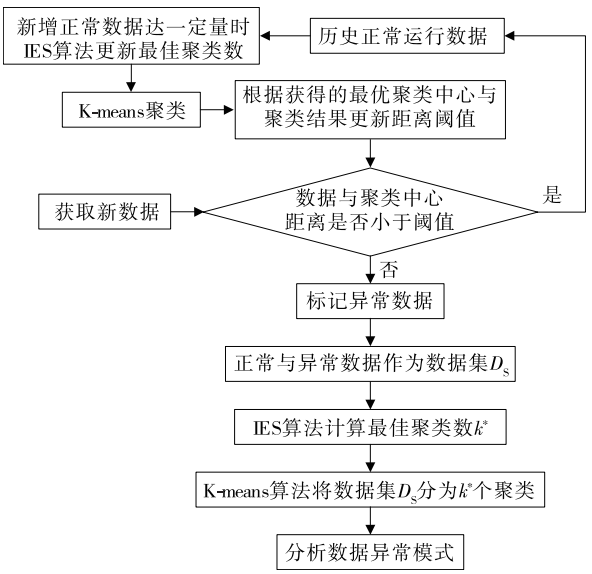

图 3　配电网电压自适应实时异常检测总流程

Figure 3　The general process of adaptive real-time anomaly detection for the distribution network voltage

3　算例分析

3.1　数据集

以 2 个实际配电网电压数据集为例(记为 D_1 和 D_2)，与 DBI 算法和轮廓系数法进行仿真比较，验证所提 IES 算法的有效性。

D_1 有 1 000 个样本点，每个点对应一个时刻三相电压有效值。为体现所提 IES 算法的普适性，对 A、B、C 三相电压分别加入异常数据进行实验。对于 A 相电压，随机抽取 10%，即 100 个正常数据，并加上 4%～10%噪声，生成 1 个数据集，重复进行 50

次，生成 50 个数据集，记为 A 组数据（注意：每个数据集只含 10%异常数据，其中 50 个正常数据加上正噪声 4%～10%，50 个正常数据加上负噪声 −10%～−4%）。用同样方法对 B、C 相电压各生成 50 个数据集，分别记为 B、C 组数据。

D_2 有 4 000 个样本点，每个点对应一个时刻三相电压有效值。为说明选取不同聚类数对聚类效果的影响，对 A、B、C 三相电压均加入异常数据进行实验。对于 A 相，随机抽取 15%，即 600 个正常数据，在±5%处加高斯分布随机函数 G 作为噪声。设正常数据值为 N_{data}，则加噪声后的值为 $N_{\text{data}}\cdot(1\pm0.05)+G$。用同样方法处理 B、C 相电压，最终得到的数据集记为 $D_{2\text{noise}}$。

3.2 评价指标

各算法对最佳聚类数估计的准确率定义为

$$\omega=\frac{N_{\text{T}}}{N} \tag{8}$$

式中 N 为实验次数，对 A、B、C 三相电压的每一组数据实验 50 次，故 $N=50$；N_{T} 为对最佳聚类数估计正确的次数。

对于 A、B、C 三相电压的 3 组数据，最佳聚类数为 3，即分为 1 个正常数据聚类和 2 个异常数据聚类。用 DBI 算法、轮廓系数法和 IES 算法估计所有数据集的最佳聚类数 k^*，若 $k^*=3$ 则估计正确，若 $k^*\neq3$ 则估计错误。

由于出现异常时需用自适应 K-means 算法对正常和异常数据聚类，聚类效果影响下一步分析数据异常模式，因此，需说明 K-means 对正常和异常数据聚类时不同最佳聚类数估计值对聚类效果的影响。为方便说明，简化为分析二分类异常检测效果，再给出异常检测效果评价指标。对于二分类异常检测，可根据真实和检测情况将检测结果分为 4 类，如表 1 所示，TN 为实际正常且被检测为正常的样本，FP 为实际正常但被检测为异常的样本，FN 为实际异常但被检测为正常的样本，TP 为实际异常且被检测为异常的样本。

对于异常检测，相比于不将正常样本判定为异常，更重要的是检测到更多的异常点[16]。因此，采用召回率评估异常检测效果，其值越高意味着检测到越多真实异常点，其最大值为 1。召回率：

$$R_{\text{ecall}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{N}}} \tag{9}$$

表 1 检测结果分类

Table 1 Classification of test results

真实情况	检测情况	
	正常	异常
正常	TN	FP
异常	FN	TP

3.3 实验结果与分析

计算环境如下：计算机 CPU 为 Core i7-10700，内存 16 GB，主频 2.90 GHz，操作系统为 Windows 10（64 bit），数据分析工具为 Python 3、Jupyter NoteBook。

对 A、B、C 三相电压的 3 组数据分别用 3 种算法进行最佳聚类数估计，为满足在线监测的自适应性，应取足够大的聚类数上限 $k_{\max}$，以保证其大于最佳聚类数，因此取 $k_{\max}=20$，实验结果如表 2～4 所示。

表 2 对 A 组数据应用 3 种算法

Table 2 Three algorithms are applied to group A data

最佳聚类数估计值	最佳聚类数估计值取不同值的次数		
	DBI 算法	轮廓系数法	IES 算法
2	9	0	0
3	41	50	50

表 3 对 B 组数据应用 3 种算法

Table 3 Three algorithms are applied to group B data

最佳聚类数估计值	最佳聚类数估计值取不同值的次数		
	DBI 算法	轮廓系数法	IES 算法
2	4	0	0
3	46	50	50

表 4 对 C 组数据应用 3 种算法

Table 4 Three algorithms are applied to group C data

最佳聚类数估计值	最佳聚类数估计值取不同值的次数		
	DBI 算法	轮廓系数法	IES 算法
2	3	0	0
3	47	50	50

由表 2 可以看出，轮廓系数法和 IES 算法对最佳聚类数的估计值稳定为 3；DBI 算法的估计值仅在 2、3 之间波动，其中，对于 A 组 9 个数据集，DBI 算法解出 $k^*=2$，其余 41 个数据集解出 $k^*=3$。对表 3、4 不再赘述。根据表 2～4，计算 3 种算法对 A、B、C 三相电压的 3 组数据最佳聚类数估计的准确率 ω，如表 5 所示。

表 5　3 种算法准确率

Table 5　Accuracy of three algorithms accuracy　%

电压数据集	准确率		
	DBI 算法	轮廓系数法	IES 算法
A	82	100	100
B	92	100	100
C	94	100	100

由于 3 种算法对 k^* 的估计值只有 2 和 3，因此分别取 $k^*=2$、3，对 A、B、C 三相电压的 3 组数据进行 K-means 聚类，其中，$k^*=3$ 时将聚类后的 2 个异常数据聚类合并，从而得到正常和异常数据二分类（注意：实际中不将异常数据聚类合并，因为会损失异常数据特征信息，此处仅是为了计算召回率）。计算不同 k^* 取值下各组数据召回率均值，如表 6 所示。

表 6　不同 k^* 取值下各组数据召回率均值

Table 6　Mean recall rates of data in each group under different k^* values

最佳聚类数取值	召回率均值		
	A	B	C
2	0.496	0.494	0.494
3	0.989	0.986	0.982

由表 6 可见，选取合适的聚类数能大幅提升异常检测效果。对于 A、B、C 三相电压的 3 组数据取 $k^*=2$ 显然不合适，而在表 2～4 中，DBI 算法对 k^* 的估计值多次为 2，结合表 5 可知轮廓系数法对最佳聚类数估计的准确率比 DBI 算法更高，而 IES 算法保持了轮廓系数法的高准确率。

为进一步说明轮廓系数法和 IES 算法在准确率方面比 DBI 算法更适合为 K-means 选择最佳聚类数，对 A 组数据中某一数据集进行 K-means 聚类，取 $k^*=3$，聚类结果如图 4 所示，可明显区分 3 类数据，但对于该数据集，DBI 算法解出 $k^*=2$，而轮廓系数法和 IES 算法解出 $k^*=3$。

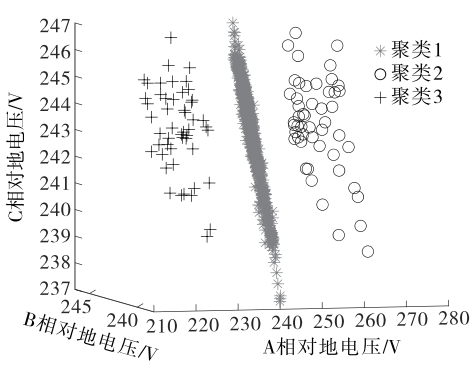

图 4　$k^*=3$ 时 K-means 算法聚类结果

Figure 4　Clustering results of K-means algorithm when $k^*=3$

图 4 中聚类 2、3 为异常数据类，聚类 1 为正常数据类，此时召回率的值为 1，表明 K-means 算法适合对该数据集聚类，而轮廓系数法和 IES 算法比 DBI 算法更适合为 K-means 选择最佳聚类数。

为说明选取不同聚类数对聚类效果的影响，分别取聚类数为 2～9，对数据集 D_{2noise} 进行 K-means 聚类；为方便计算各聚类结果的召回率，人为将异常数据聚类合并，从而得到正常和异常数据二分类。数据集 D_{2noise} 召回率随聚类数变化曲线如图 5 所示，可见召回率随着聚类数增大而增大，说明聚类效果越来越好，当聚类数为 7、8、9 时达到最大值 1。应指出，3 种聚类数选择算法对于 D_{2noise} 的最佳聚类数估计值均为 7。

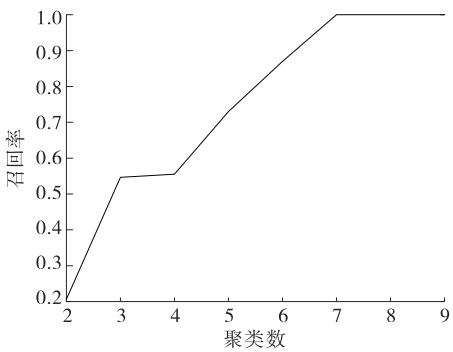

图 5　数据集 D_{2noise} 召回率随聚类数变化曲线

Figure 5　The curve of recall rate of D_{2noise} changing with clustering number

进一步分析发现，聚类数大于 7 时会发生模型过拟合。聚类数分别为 7、8 时数据集 D_{2noise} 的 K-means 聚类结果如图 6、7 所示。对比图 6、7 可知，图 6 为最佳聚类，而图 7 中将正常数据过拟合为 2 个聚类（聚类 4、7），不利于对数据进行分析。

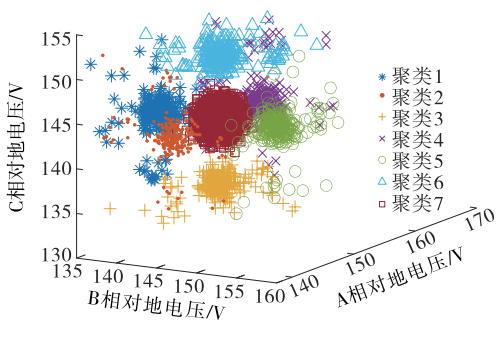

图 6　聚类数为 7 时数据集 D_{2noise} 的 K-means 聚类结果

Figure 6　K-means clustering results of D_{2noise} when the clustering number is 7

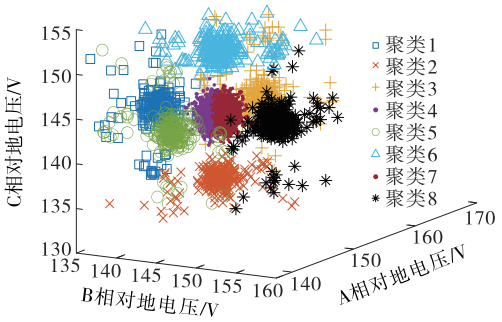

图 7 聚类数为 8 时数据集 D_{2noise} 的 K-means 聚类结果

Figure 7 K-means clustering results of D_{2noise} when the clustering number is 8

用 3 种算法对 A 组 50 个数据集估计最佳聚类数，记录平均运行时间和最小、最大运行时间，如表 7 所示，可见轮廓系数法的最小运行时间大于 DBI 算法最大运行时间，从运行时间均值也能看出 DBI 算法运行速度更快。IES 算法的运行时间均值小于其他 2 个算法，计算速度最快，最符合在线监测的快速性。IES 算法运行时间波动范围大于其他 2 个算法，是因为对于 50 个数据集，IES 算法均解出 $k_{min}=3$，但“一个极大值”规则在不同的聚类数 k 值（$k>3$）处生效，因此，不同数据集的计算量不同，导致运行时间波动。

表 7 3 种算法运行时间对比

Table 7 Running time comparison of three algorithms

算法	运行时间/s		
	平均值	最小值	最大值
DBI 算法	1.072	1.029	1.144
轮廓系数法	1.245	1.209	1.285
IES 算法	0.623	0.348	1.132

综上所述，尽管 DBI 算法计算速度稍快于轮廓系数法，但 DBI 算法准确率是 3 种算法中最低的。与 DBI 算法相比，IES 算法不仅计算速度更快，而且准确率更高；与轮廓系数法相比，IES 算法不仅保持相同的准确率，而且计算速度更快。因此，IES 算法兼顾准确率和计算速度，在保证高准确率的前提下缩短了计算时间，提高了 K-means 算法在线监测的准确率和高效性。

4 结语

K-means 聚类算法计算速度快、准确率高，适合配电网在线监测，但当假定聚类数不合适时，可能导致聚类结果不理想。本文提出了一种快速选取聚类数的自适应 IES 算法，为 K-means 算法自动选取聚类数，使 K-means 算法在线监测配电网成为可能。以召回率评价二分类异常检测效果为例，说明为 K-means 选取合适聚类数对异常检测的重要性。IES 算法首先利用自适应变化阈值求解聚类数下限，接着在聚类数上、下限内计算轮廓系数。为提高算法速度，提出“一个极大值”规则，避免计算所有轮廓系数。所提 IES 算法有如下优点。

1）自适应能力强。IES 算法只需给定聚类数上限这一参数，且该上限允许较大，即使动态数据库的最佳聚类数发生一定的改变，也能保证大于最佳聚类数。所提出用于确定聚类数下限的阈值可随数据集不同而自适应变化，从而自适应兼顾 IES 算法准确率和计算速度。

2）计算速度快。IES 算法利用计算迅速的 SSE 求解聚类数下限，缩小了最佳聚类数的搜寻范围，又利用所提出的“一个极大值”规则减少计算量，提高了计算速度。

3）准确率高。IES 算法充分利用了轮廓系数高准确率的特点。

算例表明，所提 IES 算法能自适应快速选取最佳聚类数，与轮廓系数法相比，IES 算法准确率相同而计算速度更快，与 DBI 算法相比，IES 算法不仅准确率更高，而且计算速度更快。因此，IES 算法兼顾准确率和计算速度，更有利于应用于配电网在线监测。

参考文献：

[1] 邓鹏，刘敏.基于改进聚类和 RBF 神经网络的台区电网线损计算研究[J].智慧电力，2021，49(2)：107-113.
DENG Peng，LIU Min.Power line loss calculation in low voltage region based on improved clustering algorithm and RBF neural network[J].Smart Power，2021，49(2)：107-113.

[2] 刘君，余思伍，陈沛龙，等.基于聚类分析的变压器有载分接开关储能弹簧故障识别[J].高压电器，2020，56(7)：159-165+172.
LIU Jun，YU Siwu，CHEN Peilong，et al.Fault recognition for on-load tap changer storage spring of power transformer by clustering analysis algorithm[J].High Voltage Apparatus，2020，56(7)：159-165+172.

[3] 韩帅，孙乐平，杨艺云，等.基于改进 K-Means 聚类和误差反馈的数据清洗方法[J].电网与清洁能源，2020，36(7)：9-15.
HAN Shuai，SUN Leping，YANG Yiyun，et al. A data cleaning method based on improved K-Means clustering and error feedback[J].Power System and Clean Energy，2020，36(7)：9-15.

[4] 侯庆春，杜尔顺，田旭，等.数据驱动的电力系统运行方式分析[J].中国电机工程学报，2021，41(1)：1-12.
HOU Qingchun，DU Ershun，TIAN Xu，et al.Data-driven power system operation mode analysis[J]. Proceedings of the CSEE，2021，41(1)：1-12.

[5] 秦佳倩，唐海国，张帝，等.加权模糊 C 均值聚类和主客观赋权结合的厂用电关联特征挖掘方法[J].电力科学与技术学报，2020，35(4)：122-127.
QIN Jiaqian，TANG Haiguo，ZHANG Di，et al.Auxiliary power consumption feature mining method weighted fuzzy C-means clustering and subjective and objective weighting combined[J]. Journal of Electric Power Science and Technology，2020，35(4)：122-127.

[6] 尚学军，霍现旭，郑晓冬，等.基于离散小波分析与 K-means 聚类算法的 MMC-HVDC 输电线路保护方案[J].电测与仪表，2020，57(24)：52-57.
SHANG Xuejun，HUO Xianxu，ZHENG Xiaodong，et al. MMC-HVDC transmission line protection scheme based on discrete wavelet analysis and K-means clustering algorithm[J]. Electrical Measurement & Instrumentation，2020，57(24)：52-57.

[7] PETER R J.Silhouettes：a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics，1987，20：53-65.

[8] 刘洋，刘洋，许立雄，等.计及数据类别不平衡的海量用户负荷典型特征高性能提取方法[J].中国电机工程学报，2019，39(14)：4093-4104.
LIU Yang，LIU Yang，XU Lixiong，et al. A high performance extraction method for massive user load typical characteristics considering data class imbalance[J]. Proceedings of the CSEE，2019，39(14)：4093-4104.

[9] 宋军英，崔益伟，李欣然，等.改进分段线性表示与动态时间弯曲相结合的负荷曲线聚类方法[J].电力系统自动化，2021，45(2)：89-96.
SONG Junying，CUI Yiwei，LI Xinran，et al.Load curve clustering method combining improved piecewise linear representation and dynamic time warping[J]. Automation of Electric Power Systems，2021，45(2)：89-96.

[10] 黄雨薇，彭道刚，姚峻，等.基于 SSA 和 K 均值的 TD-BP 神经网络超短期光伏功率预测[J].太阳能学报，2021，42(4)：229-238.
HUANG Yuwei，PENG Daogang，YAO Jun，et al.Ultra-short-term photovoltaic power forecast of TD-BP neural network based on SSA and K-means[J]. Acta Energiae Solaris Sinica，2021，42(4)：229-238.

[11] 赵晶晶，贾然，陈凌汉，等.基于深度学习和改进 K-means 聚类算法的电网无功电压快速分区研究[J].电力系统保护与控制，2021，49(14)：89-95.
ZHAO Jingjing，JIA Ran，CHEN Linghan，et al. Research on fast partition of reactive power and voltage based on deep learning and an improved K-means clustering algorithm[J].Power System Protection and Control，2021，49(14)：89-95.

[12] TIBSHIRANI R，HASTIE W T.Estimating the number of clusters in a data set via the gap statistic[J]. Journal of the Royal Statistical Society，Series B (Methodological)，2001，63(2)：411-423.

[13] BREIMAN L I，FRIEDMAN J H，OLSHEN R A，et al. Classification and regression trees[J]. Biometrics，1984，40(3)：358.

[14] 钱宇骋，甄超，季坤，等.变压器在线监测数据异常值检测与清洗[J].哈尔滨理工大学学报，2020，25(5)：15-22.
QIAN Yucheng，ZHEN Chao，JI Kun，et al.Transformer online monitoring data abnormal value detection and cleaning[J]. Journal of Harbin University of Science and Technology，2020，25(5)：15-22.

[15] 严英杰，盛戈皞，刘亚东，等.基于滑动窗口和聚类算法的变压器状态异常检测[J].高电压技术，2016，42(12)：4020-4025.
YAN Yingjie，SHENG Gehao，LIU Yadong，et al. Anomalous state detection of power transformer based on algorithm sliding windows and clustering[J]. High Voltage Engineering，2016，42(12)：4020-4025.

[16] 侯慧，耿浩，肖祥，等.台风灾害下用户停电区域预测及评估[J].电网技术，2019，43(6)：1948-1954.
HOU Hui，GENG Hao，XIAO Xiang，et al.Research on prediction and evaluation of user power outage area under typhoon disaster[J]. Power System Technology，2019，43(6)：1948-1954.