电测与仪表
Electrical Measurement & Instrumentation
ISSN 1001-1390,CN 23-1202/TH

《电测与仪表》网络首发论文

题目：基于改进式 k-prototypes 聚类的坏数据辨识与修正
作者：王孝慈，董树锋，刘育权，王莉，李俊格
收稿日期：2020-01-23
网络首发日期：2020-11-24
引用格式：王孝慈，董树锋，刘育权，王莉，李俊格．基于改进式 k-prototypes 聚类的坏数据辨识与修正．电测与仪表.
https://kns.cnki.net/kcms/detail/23.1202.TH.20201124.1047.020.html

网络首发时间：2020-11-24 11:10:45
网络首发地址：https://kns.cnki.net/kcms/detail/23.1202.TH.20201124.1047.020.html

基于改进式 k-prototypes 聚类的坏数据辨识与修正

王孝慈[1]，董树锋[1]，刘育权[2]，王莉[2]，李俊格[2]

（1. 浙江大学 电气工程学院，杭州 310027；2. 广州供电局有限公司，广州 510620）

摘要：工业领域很多技术的实现都以准确的负荷数据为基础，而工厂现有的负荷数据测量体系常因为通讯、存储等故障，导致负荷数据中出现大量坏数据。因此，提出基于改进式 k-prototypes 聚类的坏数据辨识与修正方法，通过在聚类中引入非负荷数据特征，削弱负荷坏数据对聚类结果的影响，使坏数据辨识和修复结果更准确。改进式 k-prototypes 算法通过随机初始化，并行计算择优，克服了标准 k-prototypes 容易随初始聚类中心陷入局部最优解的缺陷；并通过聚类数量的自适应处理，解决了主观决定聚类数量的问题。基于聚类结果，根据正态分布原则确定负荷数据可行域，识别坏数据，并利用类中心置换法进行修正。实验表明，该方法较只考虑负荷数据的模糊均值聚类法效果更好，坏数据识别的召回率与修正的准确率显著提高。

关键词：k-prototypes 聚类；混合数据集聚类；坏数据辨识；类中心置换修正法；工业负荷预处理

中图分类号：TM933　　**文献标识码**：A

Industrial load data identification and correction method with improved k-prototypes clustering algorithm

Wang Xiaoci[1], Dong Shufeng[1], Liu Yuquan[2], Wang Li[2], Li Junge[2]

(*1. College of Electrical Engineering Zhejiang University, Hangzhou 310027, China.*
2. Guangzhou Power Supply Bureau Co., Guangzhou 510620, China)

Abstract: The realization of many technologies in the industrial field is based on accurate load data, while the existing measurement system in factories often results in a large number of bad data due to communication and storage failures. Therefore, an industrial load data identification and correction method with improved k-prototypes clustering algorithm was proposed to reduce the impact of bad load data on the clustering results by introducing characteristics of non-load data, so as to make the identification and repair results more accurate. Through random initialization and parallel calculation, the improved k-prototypes algorithm overcomes the defect that standard algorithm tends to fall into the local optimal solution. And the problem of subjectively determining the number of clusters is solved by adaptive processing. Based on the clustering results, the feasible region of load data is obtained according to the principle of normal distribution, and the bad data is identified. The identified bad data is corrected by centroid vector replacing. Experiments show that the method outperforms the fuzzy C-means clustering method which only considers the load data.

Keywords: k-prototypes clustering, mixed dataset clustering, bad data identification, correction with centroid vector replacing, industrial load data preprocessing

0　引 言

随着数据挖掘技术在工业用电领域的逐渐应用，准确的负荷数据变得至关重要。对于工厂来说，准确的负荷数据可以支持其负荷预测、需求响应等多种高级应用，从而提升用能的经济性。另一方面，对于电力企业，准确的用户用电数据可以降低其参与售电市场的风险，并避免用户窃电带来的经济损失[1-2]。然而，由于设备停运、仪表故障、通讯线路异常等原因，导致工厂负荷数据中存在大量坏数据，

基金项目：国家重点研发计划资助项目（2016YFB0901300）

影响工厂和电力企业的正确决策[3]。因此，在对负荷数据挖掘之前，进行坏数据的辨识与修正非常重要。

目前，坏数据辨识与修正方法的研究主要针对系统负荷或母线负荷，主要方法有状态估计法、横向纵向比较法、聚类法。传统的基于加权残差或标准残差的状态估计法，容易出现残差污染和残差淹没现象，造成坏数据的漏检和误检[4-5]。横向纵向对比法根据历史负荷数据值确定正常数据范围，对相邻时刻的负荷数据值非常依赖，因此在一定程度上无法处理连续丢失或突变的坏数据[6-7]。聚类法通过提取用户典型用电模式，确定每种用电模式下负荷数据的合理范围完成负荷辨识，取得了不错的效果[8-10]。文献[8]利用快速爬山法改善模糊 C 均值(fuzzy C-means, FCM)聚类算法，改善了聚类数难以选择，初始聚类中心随机选择等缺点，并根据每个用电模式中历史负荷的最大最小值确定正常数据可行域完成坏数据辨识。文献[9]利用极限学习机提取数据特征，并利用空间核密度聚类分析特征并识别不良数据。但是，上述聚类的特征向量全部为负荷用电数据，在聚类向量中本身包含坏数据的情况下，聚类结果无法准确反映待测日的用电模式特征，会对数据辨识与修正造成影响。文献[10]为了解决上述问题，提出了一种利用灰色关联分析引入非负荷数据信息，改善 FCM 聚类的坏数据辨识与修正模型，实验结果表明在聚类中引入非负荷数据特征值，可以提高模型的准确性和实用性。

文献[11]指出在进行负荷模式提取时，不存在一种聚类方法普遍优于其他聚类方法。并且对于工业负荷模式提取，直接移植现有的聚类方法效果不佳，需要更有针对性的研究。文献[12]采用统计模糊矩阵分类法对工业负荷进行分类，并通过非参数回归分析方法提取中心负荷向量，进而构造异常数据域，完成负荷辨识。但该方法在落地时，需要海量数据，现有大部分工厂无法满足其对数据存储的要求。

针对上述不足，提出了一种基于改进式 k-prototypes 聚类的坏数据辨识与修正方法。主要贡献在于：

（1）构建聚类特征向量时，考虑工厂用电特点，引入非负荷数据，削弱负荷数据中坏数据对聚类结果的影响；

（2）对标准 k-prototypes 算法进行改进，增加了多组初值并行择优，改善了其容易陷入局部最优的缺点，并对聚类数进行自适应处理，解决了主观选择聚类数量的问题；

（3）结合聚类结果，提出了负荷可行域的计算方法，并基于质心曲线置换对坏数据进行修正。

算例分析表明，所提改进式 k-prototypes 聚类算法较 FCM 聚类算法在工厂用电模式提取的效果更好，应用到坏数据辨识与修复中，识别的召回率和修复的准确率都有所提高；较简单置信区间坏数据识别、线性插值坏数据修复，效果提升显著。

1 混合特征聚类

1.1 混合聚类特征选择

利用负荷聚类算法进行坏数据辨识与修正的本质是：提取用户用电的行为模式，将不符合其行为模式的数据找到，并进行修正。然而用于聚类的工厂负荷曲线本身就包含坏数据，如果在聚类时，仅考虑负荷数据，会带来两个问题：

（1）聚类时，结果受负荷坏数据的影响大，无法对用户用电的行为模式进行精确提取，从而影响负荷辨识与修正结果；

（2）修正时，对于某些用电数据严重缺失的时段，无法通过其他信息辅助对其进行填补。

基于上述原因，需要在特征向量中引入工厂的其他用电特征，修正工厂负荷时序值的聚类结果。工厂用电与其生产计划、生产模式强相关。对于有规律性生产模式的大多数工厂，其生产活动一般按周开展，并受节假日影响。因此，需在特征向量中引入“工作日属性”与“节假日属性”。另外，部分工厂除生产用电外，空调用电占比最大，如轮胎工厂的空调用电可达其总用电量的 20%~30%，空调用电量与气温强相。故在考虑温度敏感度大的季度、空调负荷占比高的工厂时，需要增加“气温”特征维度。聚类特征选取结果如表 1 所示。

表 1 聚类特征选取

Tab.1 Feature selection

数据	类型
日负荷数据	数值型
气温	数值型
工作日属性（如：星期一）	非数值型
节假日属性	非数值型

1.2 k-prototypes 算法

对于混合类型数据向量，同时包括数值型与非数值型数据。传统的聚类算法会将非数值型数据数值化，在计算聚类损失函数时仍然使用欧式距离。这样不但使非数值型数据脱离了本身的物理含义，还在聚类中引入了干扰因素。针对上述情况，选择

k-prototypes 算法，在计算聚类损失函数时对数值型、非数值型数据分别进行考虑。

对含有 n 个向量的集合 $\boldsymbol{X}=\{\boldsymbol{x_1},\boldsymbol{x_2},\cdots,\boldsymbol{x_n}\}$，其第 j 个向量由一组特征值组成，可表示为：

$$\begin{aligned}\boldsymbol{x_j}=[&x_{j,1}^r,x_{j,2}^r,\cdots,x_{j,m}^r,\cdots,x_{j,m_r}^r,x_{j,m_r+1}^c,\cdots\\&\cdots,x_{j,m}^c,\cdots,x_{j,m_r+m_c}^c]\end{aligned}\tag{1}$$

式中 $x_{j,m}$ 为 $\boldsymbol{x_j}$ 的第 m 个特征值，上标 r 表示数值型特征，上标 c 表示非数值型特征。m_r 为数值型特征的总数，m_c 为非数值型特征的总数。

通过 k-prototypes 算法将所有向量分为 k 类，则向量集合 $\boldsymbol{X}$ 可表示为：

$$\begin{cases}\boldsymbol{X}=\boldsymbol{X}_1\cup\boldsymbol{X}_2\cup\cdots\cup\boldsymbol{X}_k\\\boldsymbol{X}_i\cap\boldsymbol{X}_j=\varnothing,\ i\neq j\end{cases}\tag{2}$$

式中 $\boldsymbol{X}_i$ (i = 1, 2, ⋯, k)为向量聚类后的第 i 类向量的集合。

向量聚类中，数值属性的相似距离为欧式距离，非数值型属性的相似距离为分类属性距离[13]，则 $\boldsymbol{x_j}$ 到其类心的距离可表示为：

$$\begin{cases}d(\boldsymbol{x}_j,\boldsymbol{v}_i)=\sum\limits_{m=1}^{m_r}\left\|x_{j,m}^r-v_{i,m}^r\right\|^2+\gamma\sum\limits_{m=m_r+1}^{m_r+m_c}\delta(x_{j,m}^c,v_{i,m}^c)\\\delta(x_{j,m}^c,v_{i,m}^c)=\begin{cases}0,\ x_{j,m}^c=v_{i,m}^c\\1,\ x_{j,m}^c\neq v_{i,m}^c\end{cases}\end{cases}\tag{3}$$

式中 $\boldsymbol{x_j}$ 所属类 $\boldsymbol{X}_i$ 的中心向量为：

$$\boldsymbol{v}_i=[v_{i,1}^r,v_{i,2}^r,\cdots,v_{i,m_r}^r,v_{i,m_r+1}^c,\cdots,v_{i,m_r+m_c}^c]\tag{4}$$

式中 γ 为非数值型变量的权重，可在数值数据分布距离标准差的 1/3~2/3 之间进行选择[14]。

在聚类过程中，定义各个向量到所属类中心的总距离为聚类损失函数，聚类的目标为使聚类损失函数最小，可表示为：

$$\min\sum_{i=1}^{k}\sum_{j=1}^{n_i}d(\boldsymbol{x}_j,\boldsymbol{v}_i)\tag{5}$$

式中 n_i 集合 $\boldsymbol{X}_i$ 中向量的数量。

1.3 改进式 k-prototypes 聚类过程

为了克服标准 k-prototypes 容易陷入局部最优，聚类数量难以选择等缺点，对 k-prototypes 算法进行了如下改进：

（1）聚类过程中，随机选取多组聚类中心初值，并行计算，选取代价函数值最小的作为聚类结果，解决陷入局部最优的问题；

（2）提取聚类效果关键指标，设定阈值，对聚类数量进行自适应处理，克服类别数选择的主观性。

（3）将向量数量较少的类拆散，向量合并到距离最小的其他类，避免算法将坏数据单独分类，无法进行识别。

改进后的 k-prototypes 聚类主要过程如图 1 所示。

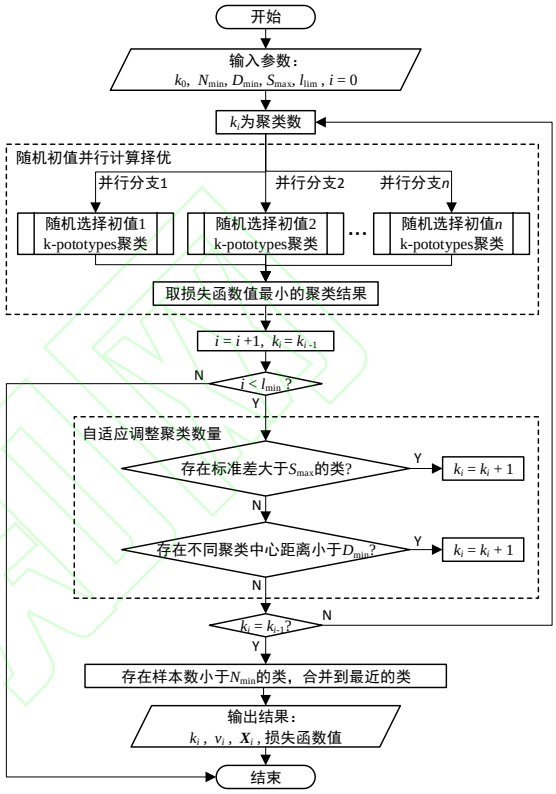


图 1 改进式 k-prototypes 聚类流程图

Fig.1 Flow chart of improved k-prototypes clustering algorithm

对于其输入参数进行如下说明：

k_0 为类数的初始值，由于算法对类数有自适应调整的过程，所以 k_0 的选择不会影响聚类的最终结果。但 k_0 越靠近最终的聚类类数，算法的迭代步骤越少，运算速度越快；

S_{max} 为每一类向量距离分布的最大标准差，若某类的距离分布标准差超过 S_{max}，说明该类内部相似度较低，应进行拆分。S_{max} 的选取可以根据聚类数据的标准差选取其 5%~20%，其选值越小，类内越紧凑；

D_{min} 为不同聚类中心的最小距离，若两类距离小于 D_{min}，则需进行合并。D_{min} 的选取可以根据聚类数据的平均距离选取其 10%~20%，其选值越大，类间分隔越明显；

N_{min} 为每一类最少的向量数目，若少于此数，则不能作为一个独立的类。N_{min} 的选取可以根据聚类数

据集的长度选择其 5% ~ 10%的数量，如果 $N_{\min}=1$ 则不对每类最少向量数目进行约束。$N_{\min}$ 的限制可以避免将包含大量坏数据的向量单独分类，使其无法被辨识与修正；

$l_{\lim}$ 为算法最大迭代次数。

2 坏数据辨识及修正

2.1 坏数据辨识过程

根据聚类结果，提取每类集合中每个向量的负荷数据。对于采样点数量为 s 的负荷数据（若采样间隔为 15min，则 $s=96$），向量 $\boldsymbol{x}_j$ 的提取结果可表示为：

$$\boldsymbol{p}_j=[x^r_{j,1},x^r_{j,1},\cdots,x^r_{j,s}] \tag{6}$$

聚类中心向量 $\boldsymbol{v}_i$ 的提取结果为：

$$\boldsymbol{v}^*_i=[v^r_{i,1},v^r_{i,2},\cdots,v^r_{i,s}] \tag{7}$$

数据提取后，对应分类关系不变，即若 $\boldsymbol{x}_j\in\boldsymbol{X}_i$，则 $\boldsymbol{p}_j\in\boldsymbol{P}_i$。

每类的负荷曲线具有相似性，即曲线形状大致相似，且几个峰谷时刻基本相同，可认为同一类型负荷曲线以 $\boldsymbol{v}_i^*$为中心成正态分布[15]。根据正态分布理论计算每类负荷功率的可行域，具体步骤如下：

步骤 1：针对每一类负荷 $\boldsymbol{P}_j$，计算正态分布参数：

$$\begin{cases}\boldsymbol{\mu}_i=\boldsymbol{v}^*_i=[v^r_{i,1},v^r_{i,2},\cdots,v^r_{i,s}]\\ \boldsymbol{\sigma}^2_i=[\dfrac{1}{n_i}\sum\limits_{j=1}^{n_i}(x^r_{j,1}-v^r_{j,1})^2,\dfrac{1}{n_i}\sum\limits_{j=1}^{n_i}(x^r_{j,2}-v^r_{j,2})^2,\cdots\\ \qquad\cdots,\dfrac{1}{n_i}\sum\limits_{j=1}^{n_i}(x^r_{j,s}-v^r_{j,s})^2]\end{cases} \tag{8}$$

步骤 2：利用步骤 1 获得的参数，计算负荷曲线可行域的上下限：

$$\begin{cases}\boldsymbol{l}^{up}_i=\boldsymbol{\mu}_i+3\boldsymbol{\sigma}^2_i\\ \boldsymbol{l}^{down}_i=\boldsymbol{\mu}_i-3\boldsymbol{\sigma}^2_i\end{cases} \tag{9}$$

步骤 3：形成负荷分类的可行域矩阵，对于第 i 类负荷其可行域矩阵为：

$$\boldsymbol{L}^{feasible}_i=\begin{bmatrix}l^{up}_{i,1} & l^{up}_{i,2} & \cdots & l^{up}_{i,s}\\ l^{down}_{i,1} & l^{down}_{i,2} & \cdots & l^{down}_{i,s}\end{bmatrix} \tag{10}$$

2.2 坏数据修正

基于负荷曲线相似的性质，提出一种基于类心曲线置换的坏数据修正方法，其原理为：用待修正数据曲线所属的聚类中心负荷曲线的相应部分，根据待修正数据部分首尾差值等比伸缩，置换待修正的数据。

如图 2 所示，图 2(a)为待修正的负荷数据，其中 $\left[p^r_{j,m^s},p^r_{j,m^s+1},\cdots,p^r_{j,m^e}\right]$ 为待修正数据。图 2(b)为图 2(a)的聚类中心曲线，需要将图 2(b)中对应图 2(a)待修正部分的数据$\left[v^r_{j,m^s},v^r_{j,m^s+1},\cdots,v^r_{j,m^e}\right]$按比例伸缩变换，填补到图 2(a)中。

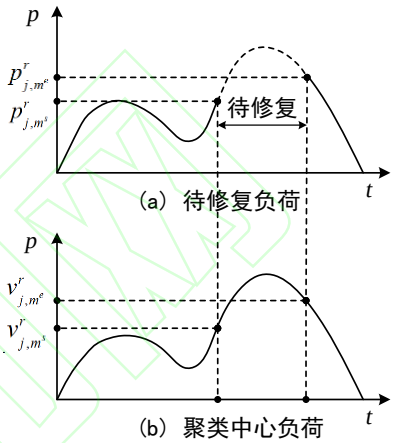


图 2 数据修正示意图

Fig.2 Schematic of bad data correction

计算$\left[v^r_{j,m^s},v^r_{j,m^s+1},\cdots,v^r_{j,m^e}\right]$中每个点需要按比例伸缩的步长，则点 $m\,(m^s<m<m^e)$ 的伸缩步长 $\triangle p^r_{j,m}$ 可表示为：

$$\triangle p^r_{j,m}=\frac{m-m^s}{m^e-m^s}\cdot[(p^r_{j,m^e}-v^r_{i,m^e})-(p^r_{j,m^s}-v^r_{i,m^s})] \tag{11}$$

那么，修复后的数据可表示为：

$$\left[p^r_{j,m^s},p^r_{j,m^s+1},\cdots,p^r_{j,m^e}\right]=\left[v^r_{j,m^s},v^r_{j,m^s+1},\cdots,v^r_{j,m^e}\right]+\cdots+\left[\triangle p^r_{j,m^s},\triangle p^r_{j,m^s+1},\cdots,\triangle p^r_{j,m^e}\right] \tag{12}$$

2.3 方法应用流程

基于改进式 k-prototypes 聚类的坏数据辨识与修正方法如图 3 所示。在进行坏数据的辨识与修复时，含有缺失数据的向量直接标记为待修复数据，不参与聚类，减小坏数据对聚类结果的影响。

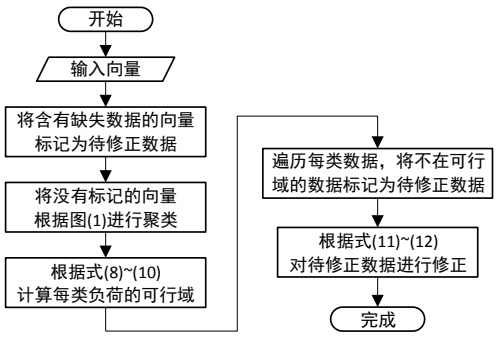


图 3 方法应用流程图

Fig.3 Flow chart of method application

3 算例分析

算例数据集包括负荷用电数据、天气数据、节假日数据。用电数据为广州某工业园现场采集的 3 个工厂从 2018 年 7 月 1 日—2018 年 10 月 24 日的负荷 96 点功率数据（去除光伏）。3 个工厂在数据采集期间，以周为单位从事规律性的生产活动，并根据国家法定节假日调整生产模式。天气数据为广州市同期的平均气温，节假日数据来源于国家法定节假日。对负荷数据进行处理：

（1）制造空白数据：每个工厂随机选择 10 条日负荷曲线，将每条曲线的部分数据删除，删除数据部分连续，长度随机且不超过整条曲线的 40%；

（2）制造坏数据：每个工厂随机选择 10 条日负荷曲线，每条曲线随机选择 3~20 个点，升高或降低 60% ~ 70%。

根据 1.3 章节所述，选取改进式 k-prototypes 的算法参数，并结合具体工厂数据微调，如表 2 所示。

表 2 改进式 k-prototypes 算例参数

Tab.2 Key parameters of improved k-prototypes

	k_0	S_{max}	D_{min}	N_{min}	l_{min}
空调厂	2	0.15	0.15	0.1	10
冷机厂	2	0.12	0.15	0.1	10
漆包线厂	2	0.15	0.15	0.1	10

3.1 改进式 k-prototypes 聚类效果

为测试随机初值，并行择优对 k-prototypes 算法陷入局部最优值的改善效果，对 3 个工厂进行仿真：选取不同的聚类数，从 1 逐渐增加并行分支数，记录代价函数值的变化，并重复多次。

图 4 为对空调厂聚类（k=5），并行分支数从 0 增至 50，重复实验 50 次的效果图。图中每条曲线为一次实验结果，较粗的曲线为多次实验的平均值，数据点在底部形成的平行线为全局最优解。可见，随着并行分支数量的增加，平均代价函数值逐渐趋于全局最优解；并且对于单次运行结果，随着并行分支数量的增加，其代价函数值围绕全局最优解的波动幅度越来越小。

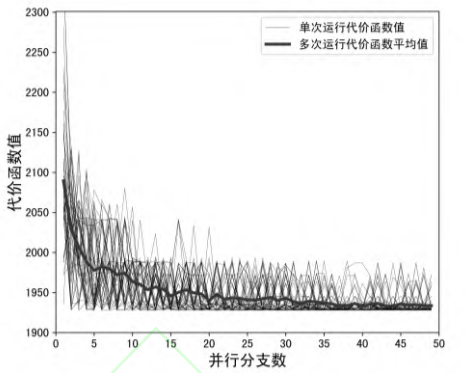


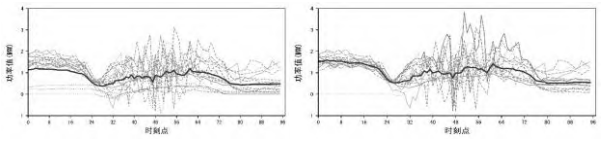
图 4 优化后对陷入局部最优的改善

Fig.4 Improvement of trapped local optima

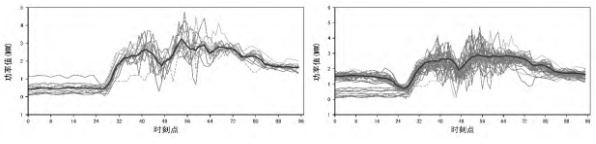
利用改进式 k-prototypes 对工厂数据进行聚类，并选择 FCM 聚类算法进行对比。FCM 聚类算法在坏数据辨识与修正的研究中应用广泛，较传统硬聚类算法效果更好。空调厂的聚类结果如图 5 所示，每条曲线为一条日负荷向量，较粗的曲线为聚类中心向量。由图可见，当聚类数相同时，此聚类算法由于引入非负荷数据削弱坏数据的影响，聚类效果更好：

（1）每类向量数量更均匀，类内更紧致，不受异常数据影响单独分类；

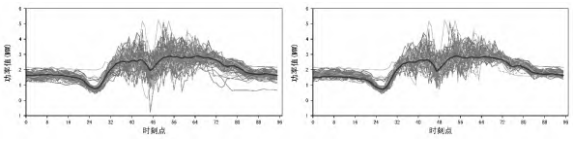
（2）不同类间分隔更明显，聚类的类心向量有明显区分，而 FCM 得类Ⅱ和类Ⅲ的中心向量比较相似。



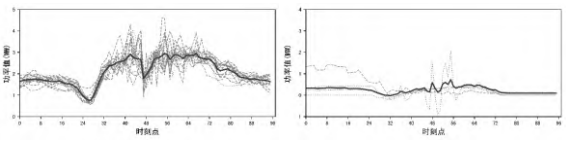
（a）本文聚类：类Ⅰ　　（b）FCM 聚类：类Ⅰ



（c）本文聚类：类Ⅱ　　（d）FCM 聚类：类Ⅱ



（e）本文聚类：类Ⅲ　　（f）FCM 聚类：类Ⅲ



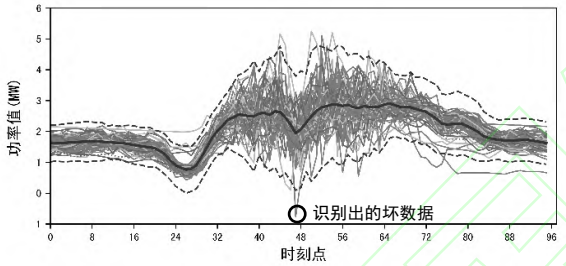
（g）本文聚类：类Ⅳ　　（h）FCM 聚类：类Ⅳ

图 5 改进式 k-prototypes 与 FCM 聚类结果对比

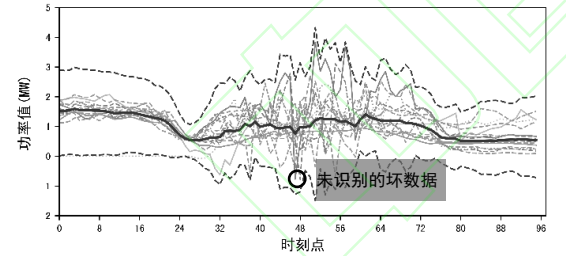
Fig.5 Comparison between the improved k-prototypes and the FCM clustering

3.2 坏数据辨识效果

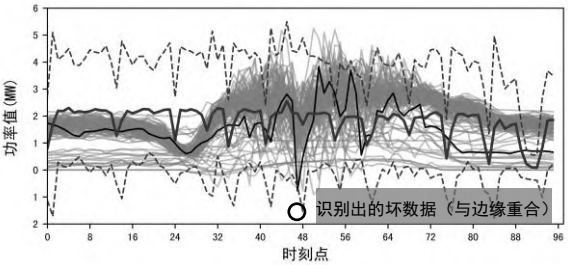
不同的聚类结果会对坏数据的辨识效果产生影响。图 6 为空调厂某个坏数据的辨识结果，图中虚线为计算的可行域。在文中算法中，坏数据所属向量被分到类Ⅲ，由于其越出可行域，被成功识别出来；而在 FCM 聚类算法中，坏数据所属向量被分到类Ⅰ，在该类可行域里，没有被正确识别；如果不进行聚类，虽然坏数据可以识别出来，但是识别结果在置信区间边缘，识别结果不稳定。



（a）本文聚类：类Ⅲ



（b）FCM 聚类：类Ⅰ



（c）置信区间法

图 6 聚类结果对坏数据辨识的影响

Fig 6. The influence of clustering results on bad data identification

对 3 个工厂的坏数据辨识结果进行统计，坏数据的召回率与辨识的准确率如表 3 所示。与 FCM 聚类算法相比，文中算法在准确率保持不变的情况下，能辨识出更多的坏数据，显著提高了坏数据的召回率。相比于置信区间法，坏数据的召回率与准确率都有显著提升。

表 3 坏数据辨识结果

Tab.3 Bad data identification results

	召回率			准确率		
	文中算法	FCM 聚类	置信区间法	文中算法	FCM 聚类	置信区间法
空调厂	91.4%	71.4%	52.5%	74.4%	73.5%	56.2%
冷机厂	87.1%	37.1%	65.4%	65.6%	26.1%	35.4%
漆包线厂	70.1%	65.5%	69.0%	85.0%	85.0%	84.2%

3.3 坏数据修正效果

利用所提的类心置换法对坏数据进行修正，通过与 FCM+类心置换法比较，分析聚类对类心置换法修正准确率的影响；同时，对比线性插值法，分析所提基于聚类算法的类心置换法与直接插值法的修正准确率的区别。

如表 4 所示。对比基于 FCM 聚类的类心置换修正法，文中方法的修正准确率在空调厂、冷机场有少量的提高，在漆包线厂与其持平，可见聚类算法对修正准确率有一定影响。相比于线性插值法，所提类心置换法在修正数据时，由于考虑了数据变化趋势，对坏数据修正的准确率有显著提高。

表 4 坏数据修正结果

Tab.4 Bad data correction results

	文中方法	FCM 聚类 + 类心置换修正	线性插值
空调厂	83.9%	83.0%	78.6%
冷机厂	73.7%	72.5%	60.3%
漆包线厂	72.6%	72.6%	71.1%

4 结束语

基于工业场景中混合数据集的聚类分析，提出了一种有效的坏数据辨识与修正方法。聚类过程中引入随机选择多组初值，并行聚类择优，克服传统 k-prototypes 算法容易陷入局部最优解的缺陷。并通过对聚类数的自适应处理，解决主观选择聚类数的问题。由于引入了非负荷数据，削弱了本身存在的坏数据对聚类结果的影响，使坏数据辨识的召回率和坏数据修正的准确率有所提高。

文中算法适用于大多数存在规律性生产模式的工厂，在实际生产过程中，一些小型工厂可能会根据需求缺口调整灵活的调整生产活动。后续的研究中，可进一步挖掘影响工厂生产活动的因素及其表征方法，应用到坏数据的辨识与修正的研究中。

参 考 文 献

[1] 赵天辉, 王建学, 马龙涛, 等. 基于非参数回归分析的工业负荷异常值识别与修正方法[J]. 电力系统自动化, 2017, 41(18): 53-59.
Zhao Tianhui, Wang Jianxue, Ma Longtao, et al. Outlier Detection and Correction Method for Industrial Loads Based on Nonparametric Regression Analysis[J]. Automation of Electric Power Systems, 2017, 41(18): 53-59.

[2] 李宁, 尹小明, 丁学峰, 等. 一种融合聚类和异常点检测算法的窃电辨识方法[J]. 电测与仪表, 2018, 55(21): 19-24.
Li Ning, Yin Xiaoming, Ding Xuefeng, et al. A method of stealing identification based on fusion of clustering and abnormal-point detection algorithm[J]. Electrical Measurement & Instrumentation, 2018, 55(21): 19-24.

[3] 张昀, 周涼, 任海军, 等. 数据挖掘在电力负荷坏数据智能辨识与修正中的应用[J]. 重庆大学学报, 2013, 36(2): 69-74.
Zhang Yun, Zhou Quan, Ren Haijun, et al. Application of data mining method in power load bad data intelligent identification and correction[J]. Journal of Chongqing University, 2013, 36(2): 69-74.

[4] 康重庆, 夏清. 灰色系统参数估计与不良数据辨识[J]. 清华大学学报：自然科学版, 1997, (4): 72-75.
Kang Chongqing, Xia Qing. Parameter Estimation and Bad Data Identification Of Grey Systems[J]. Journal of Tsinghua University (Science And Technology), 1997, (4): 72-75.

[5] 刘科研, 盛万兴, 张东霞, 等. 智能配电网大数据应用需求和场景分析研究[J]. 中国电机工程学报, 2015, 35(2): 287-293.
Liu Keyan, Sheng Wanxing, Zhang Dongxia, et al. Big Data Application Requirements and Scenario Analysis in Smart Distribution Network[J]. Proceedings of the CSEE, 2015, 35(2): 287-293.

[6] 肖白, 徐潇, 宋坤, 等. 空间电力负荷预测中异常数据的辨识与处理[J]. 东北电力大学学报, 2013, 33(1): 45-50.
Xiao Bai, Xu Xiao, Song Kun, et al. Abnormal Data Identification and Treatment in Spatial Electric Load Forecasting[J]. Journal of Northeast Dianli University (Natural Science Edition), 2013, 33(1): 45-50.

[7] 李光珍, 刘文颖, 云会周, 等. 母线负荷预测中样本数据预处理的新方法[J]. 电网技术, 2010, 34(2): 155-160.
Li Guangzhen, Liu Wenying, Yun Huizhou, et al. A New Data Preprocessing Method for Bus Load Forecasting[J]. Power System Technology, 2010, 34(2): 155-160.

[8] 孔祥玉, 胡启安, 董旭柱, 等. 引入改进模糊 C 均值聚类的负荷数据辨识及修复方法[J]. 电力系统自动化, 2017, 41(9): 96-101.
Kong Xiangyu, Hu Qian, Dong Xuzhu, et al. Load Data Identification and Correction Method with Improved Fuzzy C-means Clustering Algorithm[J]. Automation of Electric Power Systems, 2017, 41(9): 96-101.

[9] 熊晓琪, 黄鹤鸣, 郝亮亮, 等. 基于 ELM 与 DBSCAN 的微电网不良数据检测方法[J]. 电测与仪表, 2018, 55(6): 59-65.
Xiong Xiaoqi, Huang Heming, Hao Liangliang, et al. A micro-grid bad data detection method based on ELM and DBSCAN[J]. Electrical Measurement & Instrumentation, 2018, 55(6): 59-65.

[10] 林顺富, 谢潮, 李东东, 等. 基于灰色关联与模糊聚类分析的负荷预处理方法[J]. 电测与仪表, 2017, 54(11): 36-42.
Lin Shunfu, Xie Chao, Li Dongdong, et al. Load preprocessing method based on grey relational analysis and fuzzy clustering[J]. Electrical Measurement & Instrumentation, 2017, 54(11): 36-42.

[11] 张铁峰, 顾明迪. 电力用户负荷模式提取技术及应用综述[J]. 电网技术, 2016, 40(3): 150-157.
Zhang Tiefeng, Gu Mingdi. Overview of Electricity Customer Load Pattern Extraction Technology and Its Application[J]. Power System Technology, 2016, 40(3): 150-157.

[12] 赵天辉, 王建学, 马龙涛, 等. 基于非参数回归分析的工业负荷异常值识别与修正方法[J]. 电力系统自动化, 2017, 41(18): 53-59.
Zhao Tianhui, Wang Jianxue, Ma Longtao, et al. Outlier Detection and Correction Method for Industrial Loads Based on Nonparametric Regression Analysis[J]. Automation of Electric Power Systems, 2017, 41(18): 53-59.

[13] HUANG Z，MA NG.Fuzzy k-modes algorithm for clustering categorical data [J].IEEE Translations on Fuzzy Systems.1999,7(4).

[14] Huang, Z.: Clustering large data sets with mixed numeric and categorical values, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997.

[15] 刘莉, 王刚, 翟登辉. k-means 聚类算法在负荷曲线分类中的应用[J]. 电力系统保护与控制, 2011, (23): 74-77, 82.
Liu Li, Wang Gang, Zhai Denghui. Application of k- means clustering algorithm in load curve classification[J]. Power System Protection and Control, 2011(23): 74-77, 82.

作者简介：

王孝慈（1994—），女，硕士研究生，从事智能用电与需求响应相关研究。Email: 982128684@qq.com

董树锋（1982—），男，博士，副教授，从事状态估计和有源配电网分析相关研究。Email: dongshufeng@zju.edu.cn

收稿日期：2020-01-23；修回日期：2020-04-11

（田春雨　编发）