文章编号：2096-9317（2022）06-0160-03

基于 k-means 算法的电力负荷数据聚类方法分析

刘姣姣

长江工程职业技术学院，湖北 武汉 430212

摘要：文章针对电力负荷数据的特点，对原始数据进行了异常修正及归一化处理，针对 k-means 算法容易陷入局部最优解的缺陷进行了优化改进。然后通过算例分析，对比了传统 k-means 算法与文章改进的 k-means 算法的聚类结果。实验结果表明，改进算法迭代次数少、收敛速度快，具有更准确的负荷聚类效果。

关键词：k-means 算法；数据挖掘；电力负荷数据；聚类分析

分类号：TM715

0　引言

随着智能电网的推广，电力行业累积了海量的负荷数据。根据用户的相关属性对不同类型的负荷进行划分，挖掘不同时段负荷的构成及其相互之间的关系称为负荷聚类。科学合理的负荷聚类既可以帮助供电企业制定合理电价、实现错峰管理，也可以发现短板，为调整服务策略、降低成本、提升能源利用率提供数据依据。

1　电力负荷数据及其处理

1.1 电力负荷数据的特点

（1）数据量大。电网规模庞大、设备众多，随着智能电表的普及，电网采集、计算和传输的数据量都大幅增长，负荷数据增长速度加快，数据规模变大。

（2）类型复杂。电网数据来源众多、种类繁杂，包括时间序列数据和文本、图像数据等多种类型，且不同行业用户负荷曲线存在相似性，难以提供高效的决策管理支持。

（3）时序性。在电力行业，很多负荷数据具有时序性，根据用户需求，电厂发电、电能转换、电能传输都有时序要求。挖掘负荷数据价值必须以时序特点为前提，否则将失去研究意义 [1]。

1.2 电力负荷的时间特性指标

针对电力负荷的时间特性，对负荷进行分析与计算时，需要使用一些特性指标。

（1）负荷率。日负荷率 = 日用电总量 ÷ 日最大负荷 ×24 h×100%；日最小负荷率 = 日最小负荷 ÷ 日最大负荷 ×100%。

（2）负荷曲线。日负荷曲线随时间变化，受温度、地区、时间因素影响，曲线可以反映当天总用电量与 24 h 最大负荷乘积之间的占比关系，由此得出不同用户的用电特征。根据负荷曲线的峰谷差可以实现错峰管理、电网调度、分时电价、电量分配等。

（3）负荷峰谷差。日峰谷差 = 日最大负荷 – 日最小负荷。

（4）年最大负荷利用小时数。年最大负荷利用小时数 =（年用电量 ÷ 年最大负荷）=8 760 × 年负荷率。

（5）负荷同时率。同时率 = 区域最高负荷 ÷ 各个分区最高负荷之和 ×100%。

1.3 电力负荷数据的预处理

异常数据产生的原因有许多，如终端设备异常或缺陷、传输线路通信错误、数据产生机制存在问题，在某段时间内出现的相同数值、空值或者突变值都可以认为是异常数据。通常采用物理识别、统计识别两种方式判定异常数据，物理识别的根据是以往经验；统计识别需要设定置信区间及概率，超出误差范围即判定为异常 [2]。电力负荷的异常数据一般采用物理识别方法。针对异常数据，可以采用如下方法处理。

（1）个别剔除。方法简单有效，是各类软件最常用的方法。如果数据集的样本数据存在缺失则剔除，适用于样本中空值较少的情况。但这种方法也存在局限性，因为其通过减少数据集的样本量来计算，如果样本量过小，会对结果准确性有比较大的影响。

（2）均值替换。异常数据占比过多且样本变量较重要时，均值替换相对于个别剔除更为适用。可以将样本变量分为数值型和非数值型两种，如果缺失的是

作者简介：刘姣姣，女，硕士，讲师，研究方向为电气工程及其自动化。

数值型，则参照剩余样本值的均值做补充；如果缺失的是非数值型，则根据众数原则，利用其他样本中次数最高的值进行补充。

（3）热卡填充。利用热卡填充，可以在样本集中找到与异常数据最相似的样本，通常采用相关系数矩阵判断样本集之间的相似性。方法的优点是标准差差别小，缺点是用于回归方程时误差会增大。

1.4 电力负荷数据的归一化处理

数据归一是指按照某类算法将数据限制在指定区间内，设负荷数据序列为 $L=(l_1,l_2,\cdots,l_k)$，常用以下归一方法。

（1）极差归一。将负荷数据的值 l_k 映射到指定区间 [m,n] 内的 i_k，保留原始数据之间的关系：

$$\dot{l}_k=\frac{l_k-\min(L)}{\max(L)-\min(L)},k=1,2,\cdots,t \tag{1}$$

式中：max(*L*) 为最大值；min(*L*) 为最小值。

（2）最大值归一。取负荷曲线中的最大值为参考，通过计算归一化到区间 [m,n] 内。

$$\dot{l}_k=\frac{l_k}{\max(L)},k=1,2,\cdots,t \tag{2}$$

（3）Z-score 归一。采用负荷数据的标准差和均值实现归一。

$$\dot{l}_k=\frac{l_k-\mu}{\sigma},k=1,2,\cdots,t \tag{3}$$

式中：μ 为均值，σ 为标准差。

2　k-means 算法及其改进

2.1 k-means 算法步骤

k-means 算法是根据参数 *k* 将 *n* 个数据集划分为 k-means（k 聚类），最终使各个聚类的数据点到聚类中心的距离的平方和达到最小的方法。

k-means 算法的具体步骤如下：（1）任意选 *k* 个点作为初始聚类的中心或者均值；（2）计算其他数据点到聚类中心的距离；（3）按最近距离原则将数据点分配到最近的中心；（4）利用均值算法计算新的聚类中心；（5）若相邻中心无变化或准则函数 *E* 已收敛，算法结束，否则继续迭代；（6）最后产生的 *k* 个聚类中心点和以它为中心的聚类划分是最终结果。

2.2 k-means 算法模型

假设对 *n* 个 *m* 维样本聚类，得到样本集 $X=\{X_1,X_2\cdots,X_n\}$，其中，$X_i=(X_{i1},X_{i2}\cdots,X_{im})$，*k* 个分类记为 $C=\{C_1,C_2\cdots,C_k\}$，质心 $z_j=\frac{1}{n_j}\sum_{x\in c_j}X_i,i=1,2,\cdots,k$（其中，$n_j$ 为 c_j 中数据点数量），聚类的目标就是让 *k* 个类满足以下公式：

$$\sum_{j=1}^{k}\sum_{x\in c_i}d_{ij}\left(x_i,z_j\right)\rightarrow\min \tag{4}$$

式中：$d_{ij}(x_i,z_j)$ 为距离计算函数，文章选用欧式距离来计算；*k* 为聚类数量；z_j 为样本 *j* 的聚类中心。

2.3 改进的 k-means 算法

针对 k-means 算法容易陷入局部最优的缺点，文章提出一种基于对数自适应 GSA 的改进 k-means 算法。

（1）初始化种群规模 *n*，最大迭代次数 *T*，引力系数衰减因子参数 λ、*t*。

（2）计算适应度的值。

（3）更新函数 *best*(*t*)，*worst*(t) 和 $M_i(t)$。

$$best(t)=\min_{i\in\{1,2,\cdots,N\}}fit_i(t) \tag{5}$$

$$worst(t)=\max_{i\in\{1,2,\cdots,N\}}fit_i(t) \tag{6}$$

$$M_i(t)=\frac{m_i(t)}{\sum_{j=1}^{N}m_i(t)} \tag{7}$$

$$m_i(t)=\frac{fit_i(t)-worst(t)}{best(t)-worst(t)} \tag{8}$$

式中：$fit_i(t)$ 和 $M_i(t)$ 分别为粒子 *i* 在第 *t* 次迭代时的适应值和质量；*best*(*t*) 和 *worst*(*t*) 分别为适应值的最小值和最大值。

（4）将参数 α 改进为 *t* 的对数函数。

$$\alpha(t)=\lambda\times\ln\frac{t+T}{T} \tag{9}$$

式中：$\alpha(t)$ 为引力系数的衰减因子；λ 为 α 函数的参数；*t* 为当前迭代次数；*T* 为最大迭代次数。

（5）计算粒子的引力、速度、加速度。

$$F_i^d(t)=\sum_{j\in kbest,j\neq i}^{N}rand_jF_{ij}^d(t) \tag{10}$$

$$v_i^d(t+1)=rand\times v_i^d(t)+a_i^d(t) \tag{11}$$

$$a_i^d(t)=\frac{F_i^d(t)}{M_i(t)} \tag{12}$$

式中：$F_{ij}^d(t)$ 为粒子 *j* 对粒子 *i* 在 *d* 维上第 *t* 次迭代时的作用力；$F_i^d(t)$ 为粒子 *i* 在 *d* 维上第 *t* 次迭代时的总作用力；$v_i^d(t+1)$ 和 $v_i^d(t)$ 为粒子 *i* 在 *d* 维上第 *t* 次和第 *t*+1 次迭代时的速度；$a_i^d(t)$ 为粒子 *i* 在 *d* 维上第 *t* 次迭代时的加速度；*rand* 为 0 ～ 1 之间的随机变量。

（6）更新粒子位置、适应度的值和全局最优解。

（7）重复迭代直至满足终止条件。

（8）得到最优解，作为初始聚类中心进行 k-means 算法聚类。

3　算例分析

为验证文章提出算法的有效性，选取国内某市的日负荷曲线，通过传统 k-means 算法及文章改进的 k-means 算法进行聚类研究。设聚类数目 k=4，对负荷曲线进行聚类分析，研究对象中每一类的实际用户数量占比如表 1 所示。每一类用户根据行业属性又可分为 7 种，具体用电类型如表 2 所示。

表 1　用户数量占比

负荷曲线聚类类别	用户数 / 户	百分比
1	18 478	54%
2	7 186	21%
3	6 501	19%
4	2 053	6%

3.1 传统 k-means 算法电力负荷聚类结果

设聚类数目 k=4，利用传统 k-means 算法计算聚类结果，具体如下。

（1）三峰型曲线。在 8：00—9：00、11：00—12：00、17：00—21：00 这三个时间段内，曲线有上升趋势，每个区间之后的 1 ～ 2 h，曲线呈现下滑趋势，剩余时间曲线基本处于平稳状态。20：00 负荷功率最大，为 10.2 kW。

（2）晚高峰型曲线。在 7：00—9：00、10：00—12：00、17：00—22：00 这三个时间段内，曲线有上升趋势，每个区间之后的 1 ～ 2 h，曲线呈现下滑趋势，剩余时间曲线基本处于平稳状态。20：00 负荷功率最大，为 12 kW[3]。

表 2　用户用电类型统计

负荷曲线聚类类别	不同用电类型的用户数 / 户							总计
	农业生产用电	乡村居民用电	城镇居民用电	大工业用电	非居民照明	商业用电	其他	
1	4 123	162	3 786	3 954	3 122	5 012	1 441	18 478
2	1 155	60	2 438	1 082	1 014	1 160	277	7 186
3	880	48	1 189	1 943	995	1 401	145	6 501
4	611	20	388	306	411	302	15	2 053

（3）平稳型曲线。在 10：00—13：00 和 17：00—24：00，曲线有轻微波动，剩余时间曲线比较平稳。11：00 负荷功率最大，为 13.8 kW。

（4）错峰型曲线。在 10：00—11：00 和 17：00—21：00，曲线有上升趋势，在 1：00—9：00 和 21：00—24：00，曲线呈现下滑趋势。21：00 负荷功率最大，为 8.6 kW。

3.2 改进 k-means 算法电力负荷聚类结果

利用改进 k-means 算法计算聚类结果，具体如下。

（1）平稳型曲线。10：00—14：00、17：00—24：00，曲线有轻微波动，剩余时间曲线比较平稳。11：00 负荷功率最大，为 13.4 kW，用户用电以商业用电为主。

（2）晚高峰型曲线。1：00—9：00，负荷功率较低；11：00—13：00，曲线小幅上升；17：00—20：00，曲线大幅上升。20：00 负荷功率最大，为 11.8 kW，用户用电以城镇居民用电为主。

（3）三峰型曲线。在 6：00—7：00、10：00—11：00、17：00—18：00 这三个时间段内，曲线有上升趋势，11：00 出现最大值；在 7：00—9：00、11：00—15：00、21：00—24：00 这三个时间段，曲线出现下滑趋势，15：00 出现最小值。用户用电以大工业用电为主。

（4）错峰型曲线。0：00—7：00、18：00—24：00 两个区间内功率较大；7：00—18：00 区间功率较小。21：00 负荷功率最大，为 8.5 kW，用户用电与用电高峰期错开，用户以农业用户为主。

3.3 结论

实验表明，改进的 k-means 算法具有很好的搜索能力，初始聚类中心更接近实际，迭代次数较少，收敛速度更快，最终的聚类结果准确性更好。

4　结束语

文章改进了传统的 k-means 算法，并通过实验证明改进后算法的寻优耗时更短、初始聚类与实际偏差较小、对电力负荷数据挖掘的聚类效果更好。但在聚类结果对后续决策影响、负荷预测、管理办法等实际应用方面还需进一步研究。

参考文献

[1] 王凤领，梁海英，张波．一种基于改进差分进化的 K- 均值聚类算法研究 [J]. 计算机与数字工程，2019, 47 (5) : 1042-1048.

[2] 胡阳春．基于改进 k 均值聚类算法的电力负荷模式识别方法研究 [D]. 成都：电子科技大学，2018.

[3] 金之榆，王毛毛，史会磊．基于 DBSCAN 和改进 K-means 聚类算法的电力负荷聚类研究 [J]. 东北电力技术，2019, 40 (6) : 10-14.