Smart Grid 智能电网, 2023, 13(2), 45-51
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sg
https://doi.org/10.12677/sg.2023.132004

Hans汉斯

基于PK-Means算法的用电模式聚类

奚增辉，屈志坚，许唐云

国网上海市电力公司，上海

收稿日期：2023年3月18日；录用日期：2023年4月8日；发布日期：2023年4月19日

摘　要

用电模式聚类是电网需求侧管理，负荷预测、电力系统规划等工作的重要基础，对电力系统的分析、运行、规划都具有重要意义。针对传统的K-Means算法在进行用电模式聚类时没有有效利用时序特征的问题，提出了一种基于K-Means算法改良的时间序列聚类算法PK-Means,并在SSE评价指标基础上进行了改进，提出了一种用于时间序列聚类算法的评价指标累计相似度(CS)，通过皮尔逊相关系数的引入，PK-Means算法在用电模式聚类的场景下相较于传统的K-Means取得了更好的聚类效果。

关键词

用电模式分析，聚类分析，累计相似度，皮尔逊相关系数

Clustering of Power Consumption Patterns Based on PK-Means Algorithm

Zenghui Xi, Zhijian Qu, Tangyun Xu

State Grid Shanghai Municipality Electric Power Company, Shanghai

Received: Mar. 18th, 2023; accepted: Apr. 8th, 2023; published: Apr. 19th, 2023

Abstract

Clustering of power consumption patterns is an important basis for power grid demand side management, load forecasting, and power system planning, and is of great significance to the analysis, operation, and planning of power systems. Aiming at the problem that the traditional K-Means algorithm does not effectively use time series features when clustering electricity consumption patterns, an improved time series clustering algorithm PK-Means based on the K-Means algorithm is proposed, and based on the SSE evaluation index an improvement was made, and an evaluation index cumulative similarity (CS) for time series clustering algorithm was proposed. Through the

文章引用: 奚增辉, 屈志坚, 许唐云. 基于 PK-Means 算法的用电模式聚类[J]. 智能电网, 2023, 13(2): 45-51.
DOI: 10.12677/sg.2023.132004

introduction of Pearson correlation coefficient, PK-Means in the scenario of electricity consumption pattern clustering compared with the traditional K-Means achieves better clustering results.

Keywords

Electricity Consumption Pattern Analysis, Cluster Analysis, Cumulative Similarity, Pearson Correlation Coefficient

1. 引言

电网安全稳定运行是全社会生产作业的重要保障，电力公司的首要任务是确保能够提供稳定、安全的电力[1]。近年来，以云计算、大数据、移动互联网、物联网、智能计算为代表的新兴信息通信技术促进各个行业生产、运营方式和模式发生变化，国家电网公司在数字化潮流中也一直走在时代前列，随着大数据中台、云平台等基础设施的完善，目前已经积累了海量的用户数据，能通过海量电力数据的充分挖掘，识别用户的用电模式特点，可以针对不同用电模式的电力用户制定具有针对性的能效管理策略和负荷模式优化方案，有效缓解电网压力[2]、支撑有序用电安排等工作的开展；可以为负荷预测、需求响应策略制定以及节能减排工作优化等提供数据支撑[3]；也可以提升电网的运行可靠性，促进电网的精细化管理[4]。传统的用电模式分析方法主要是按照电力客户的所属行业进行分析[5]，该方法需要对各个行业下的用户进行细化分类，随后以定性或定量的方法对用户的用电模式进行具体分析：通过分析不同行业用户的负荷特征可以得出各个行业用电负荷伴随时间的变化规律。随着大数据、人工智能的发展，越来越多的方法被应用到用电模式分析领域。李培强等提出基于模糊聚类原理提出了模糊 C 均值算法和模糊等价关系两种算法对变电站的综合负荷进行分类的方法[6]，Chunlin Zhong 等提出使用 K-Means 算法对用户添加标签，通过用户画像为电力公司了解用户的电力消耗习惯、了解用户需求、提高服务质量提供数据，支撑公司业务发展[7]，卜祥国、Wang 等通过 K-Means 算法对家庭月用电量数据进行了家庭用电模式的挖掘与分析，并基于分析结果进行用电量预测应用[8] [9]，王建元等人通过 DPeaks 算法对用户用电数据进行了聚类，并利用聚类结果进行异常用电模式的判定[10]。上述算法在各自领域都取得了良好的效果，但时间序列聚类问题的本质实际上是趋势聚类，而传统的聚类算法评价都是通过距离来评价样本间的一致性，而忽略了样本特征间的时序关系。

综上所述，为了解决传统的聚类算法在用电模式聚类时不能有效利用用电时序特征的问题，本文提出了一种基于 K-Means 改良的 PK-Means 算法，通过皮尔逊相关系数度量样本间的相似性，有效利用了待聚类样本的时序特征，并在用电模式聚类领域取得了良好的效果。

2. PK-Means 算法

2.1. K-Means 算法原理

K-Means 是数据挖掘领域最常用的无监督聚类算法，主要思想是基于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个类，目标是让类内的点尽量紧密的连在一起，而让类间的距离尽量的大。假设 K 个类为 $(C_1, C_2, \cdots, C_k)$，优化目标是最小化平方误差和(SSE, Sum of Squares due to Error)，其数据表达式如下：

$$\mathrm{SSE}=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|_2^2 \tag{1}$$

其中 μ_i 是类 C_i 的均值向量，也称为质心。

K-Means 用于聚类分析时的一般步骤如下：

1) 设置要聚类的数量为 K，程序最大迭代次数为 N。

2) 从给定的数据集中随机选择 k 个样本作为初始的 k 个质心向量：$\{\mu_1,\mu_2,\cdots,\mu_k\}$。

3) 计算其他样本到每个质心的欧式距离：$d_{ij}=\left\|x_i-\mu_j\right\|_2^2$，并将当前样本划分到距离最小的类中。

4) 更新质心数据

$$\mu_j=\frac{1}{|C_i|}\sum_{x\in C_i}x \tag{2}$$

5) 重复上述 3)、4)过程直到质心不在变化或者程序达到最大迭代次数。

2.2. 皮尔逊相关系数

皮尔逊相关系数(Pearson Correlation Coefficient)，又称皮尔逊积矩相关系数(Pearson Product-Moment Correlation Coefficient，简称 PPMCC 或 PCCs)，是用于度量两个变量 X 和 Y 之间的相关性，其值介于−1 与 1 之间[11]。其定义如下：

$$\begin{aligned}\rho_{X,Y}&=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}\\&=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-(E(X))^2}\sqrt{E(Y^2)-(E(Y))^2}}\\&=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}\end{aligned} \tag{3}$$

在进行用电模式分析时，通过皮尔逊相关系数可以判断两个个体在用电模式上的相关性，其值越接近 1，说明两个个体的用电模式越接近；其值越接近−1，说明两个个体的用电模式差异越大。

2.3. PK-Means 算法

基于 K-Means 算法应用在用电模式分析时存在不能度量时序信息的问题，本文提出一种基于相似度的 K-Means 聚类算法，因为所采用的相似度度量为皮尔逊相关系数，所以称之为 PK-Means 算法。PK-Means 通过用皮尔逊相关系数替代传统 K-Means 中的欧式距离作为距离度量，在聚类时充分考虑了用电时序特征，弥补了 K-Means 在用电模式分析应用方面的缺陷，并取得了较好的聚类效果。

如图 1 所示，PK-Means 的算法执行步骤如下：

步骤 1：设置初始化参数，聚类数量为 K，最大迭代次数为 N；

步骤 2：初始化第一个质心，先随机选取一个样本作为第一个质心；

步骤 3：初始化其他质心，遍历不属于质心的样本，计算每个样本与当前所有质心的皮尔逊相关系数之和 $P(x)$，选取 $P(x)$ 最小的样本作为新的质心，重复上述操作直到质心数据等于 K；

步骤 4：样本类别划分，遍历所有样本，计算当前样本到每个质心的皮尔逊相关系数 $P(x_i)$，并将当前样本划分到 $P(x_i)ddssss$ 最大的质心中；

步骤 5：质心更新，遍历所有质心，计算当前质心类别下所有样本的均值向量作为新的质心；

步骤 6：重复步骤 4、步骤 5，直到质心不在发生变化或者达到最大迭代次数。

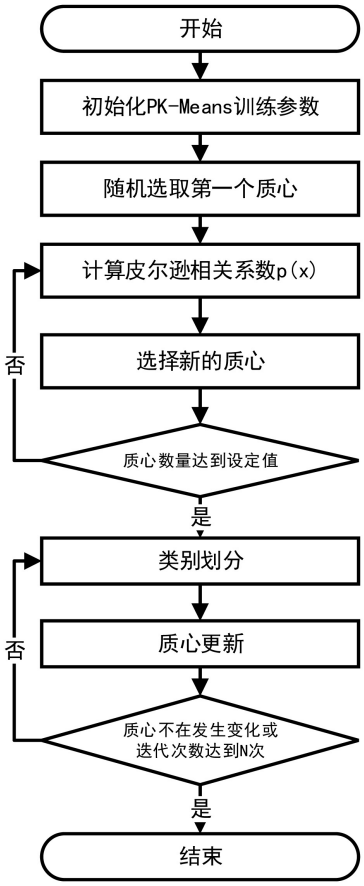

Figure 1. PK-Means algorithm flow chart
图 1. PK-Means 算法流程图

3. 实验分析

3.1. 数据集

本实验所采用的数据集为上海市某地区 146 家企业在 2022 年 1~12 月每个月的用电数据，该数据由 3 列构成，分别是户号、月份、用电量，无空值和异常值。

数据集预处理包括转置和数据标准化。如下所示：通过对月份进行转置得到用电时序矩阵，每一行代表了一个用户在 2022 年 12 个月的用电时序数据。

$$\begin{matrix} x_1^1 & x_2^1 & \cdots & x_{11}^1 & x_{12}^1 \\ x_1^2 & x_2^2 & \cdots & x_{11}^2 & x_{12}^2 \\ x_1^3 & x_2^3 & \cdots & x_{11}^3 & x_{12}^3 \end{matrix} \tag{4}$$

因为每个样本的用电体量都不一样，为了方便用电模式的聚类结果可视化，本实验还对每个用电时序数据做了数据标准化，通过下式将用电量映射到 0~1 之间。

$$y_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{5}$$

3.2. 评价标准

在传统的聚类算法中常用 SSE 作为聚类效果的评价指标，在类别数量一定时，SSE 越小，聚类的效

果越好，但在时间序列聚类中用 SSE 作为评价指标时也存在忽略样本时序关系的问题，基于此本文提出了一种基于 SSE 改良的聚类评价指标 CS (Cumulative Similarity)，即累计相似度：其计算方法如下：

$$\mathrm{CS}=\sum_{i=1}^{k}\sum_{x\in C_i}\rho_{x,\mu_i} \tag{6}$$

其中 μ_i 是类 C_i 的均值向量，在类别数量一定时，CS 越高，聚类的效果越好。

除了 CS 之外，本文还会通过对聚类结果的可视化来评价聚类效果。

3.3. 实验结果分析

本实验中，分别设置了 3、4、5、6、7 个聚类数量，记录了 CS 随迭代次数的变化数据，并对 K-Means 和 PK-Means 算法的聚类效果进行对比分析，结果如表 1 所示，在不同的聚类数量下，PK-Means 算法的 CS 值都要大于 K-Means，说明 PK-Means 在时间序列数据上的聚类表现要优于 K-Means。

Table 1. CS table under different *K* values
表 1. 不同 *K* 值下的 CS 表

算法 \ 聚类数量	3	4	5	6	7
K-Means	116.35	118.04	122.72	124.48	127.01
PK-Means	**118.42**	**121.84**	**123.09**	**125.74**	**128.38**

如图 2 所示，当 $K=4$ 时，CS 的增长相对比较明显，后续随着 K 值增大，CS 值的增长相对平缓，聚类数量带来的 CS 提升收益并不明显，所以选择 $K=4$ 作为最终的聚类数量。

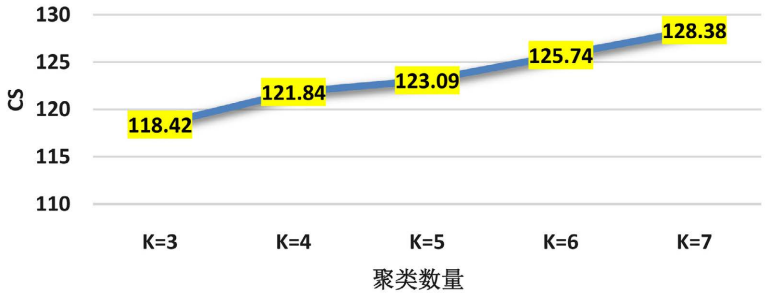

Figure 2. CS curve
图 2. CS 曲线图

如表 2 所示，第 1 次迭代时，K-Means 算法的 CS 值要大于 PK-Means，但等到第三次迭代时，PK-Means 算法的 CS 值实现了反超，说明 PK-Means 可以提取到时序特征信息，而 K-Means 算法的 CS 值几乎是不变的，进一步证明了 K-Means 在时间序列聚类中的缺陷。

Table 2. CS table under different iterations
表 2. 不同迭代次数下的 CS 表

算法 \ 迭代次数	1	3	10	100	1000	10,000	100,000
K-Means	**119.45**	119.61	118.04	118.04	118.04	118.04	118.04
PK-Means	117.30	**121.69**	**120.76**	**122.02**	**119.75**	**121.24**	**121.67**

如图 3 所示，在样本特征不复杂的情况下，K-Means 与 PK-Means 的聚类结果非常接近(如类别 1、类别 2、类别 4)，但如 K-Means 算法聚类结果类别 2 红框内所示，在 K-Means 的聚类结果中，往往会出现与当前类别的趋势不一致的样本，这些样本在趋势上是不应该划分到类别 2 的，属于误分类的样本，PK-Means 的表现整体来说比 K-Means 更稳定，在用电模式分析时，我们需要样本在时间趋势上保持一致的聚类效果，所以从可视化的结果来看 PK-Means 也是优于 K-Means 的。

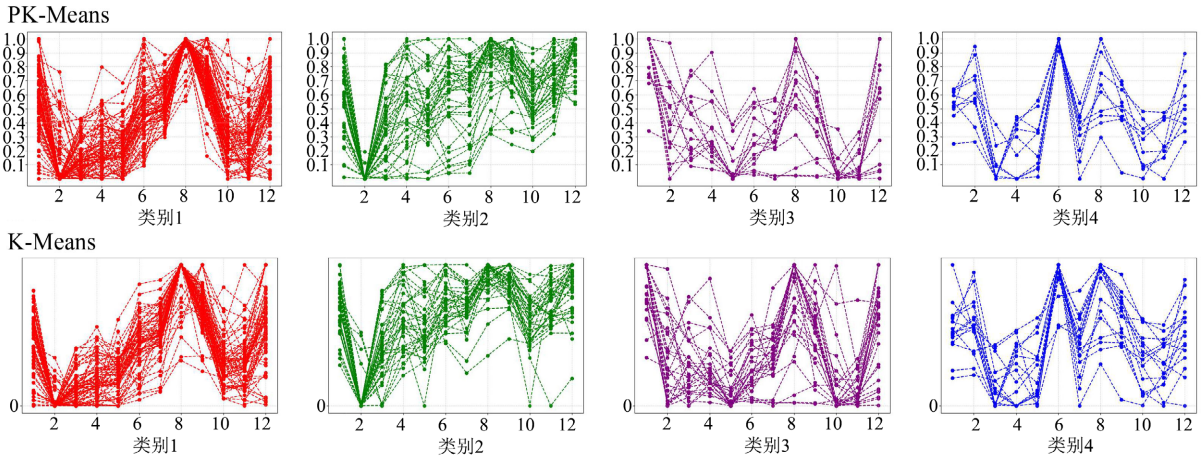

Figure 3. Visualization of clustering results
图 3. 聚类结果可视化

4. 结论

本文提出了一种基于 K-Means 改进的 PK-Means 算法，用于用电模式的聚类分析，通过皮尔逊相关系数来度量个体之间用户模式的一致性，解决了传统 K-Means 算法在时序数据聚类方面不能有效利用时序信息的缺陷，并使用累计相似度(CS)对聚类效果进行度量，通过上海某地区的真实用电数据验证了本文所提方法的有效性，在用电模式聚类场景下取得了更好的聚类效果，为电网精细化管理，用电策略规划等管理政策提供了数据支撑。

参考文献

[1] 冉冉, 陈硕, 刘颖, 李钊. 基于聚类分析的用电模式判别研究[J]. 电力大数据, 2019, 22(4): 43-49.

[2] 钱科军, 沈杰, 刘乙, 徐涛, 张政, 宋杰. 基于负荷聚类的居民需求响应积分精准激励机制[J]. 智慧电力, 2019, 47(7): 29-35.

[3] 刘俊, 罗凡, 刘人境, 徐辉, 严杰. 大数据背景下电力需求侧管理的应用策略研究[J]. 电力需求侧管理, 2016, 18(2): 5-10.

[4] 张昕, 李栋华, 程明. 基于大数据技术的错峰用电管理应用研究[J]. 现代电力, 2015, 32(3): 66-70.

[5] 隋兴嘉. 基于配用电大数据的用电行业分类和用电量需求预测建模分析[D]: [硕士学位论文]. 长春: 长春工业大学, 2018.

[6] 李培强, 李欣然, 陈辉华, 等. 基于模糊聚类的电力负荷特性的分类与综合[J]. 中国电机工程学报, 2005, 25(24): 73-78.

[7] Zhong, C., Shao, J., Zheng, F., *et al.* (2018) Research on Electricity Consumption Behavior of Electric Power Users Based on Tag Technology and Clustering Algorithm. 2018 5*th International Conference on Information Science and Control Engineering* (*ICISCE*), Zhengzhou, 20-22 July 2018, 459-462. https://doi.org/10.1109/ICISCE.2018.00102

[8] 卜祥国. 基于电力大数据的家庭用电模式分析与负荷预测[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2022.

[9] Wang, Y., Yang, Z., Wang, Y., *et al.* (2022) Research on Customer's Electricity Consumption Behavior Pattern. *Jour-*

nal of Physics: *Conference Series*, **2290**, 012042. https://doi.org/10.1088/1742-6596/2290/1/012042

[10] 王建元, 张少锋. 基于线性判别分析和密度峰值聚类的异常用电模式检测[J]. 电力系统自动化, 2022, 46(5): 87-98.

[11] 卜祥国, 纪德洋, 金锋, 冬雷, 等. 基于皮尔逊相关系数的光伏电站数据修复[J]. 中国电机工程学报, 2022, 42(4): 1514-1523.