DOI: 10.13873/J.1000-9787(2020)12-0140-03

基于 DTW 直方图的电力负荷数据聚类算法*

郝晓弘[1]，李亚岚[2]，顾　群[2]，裴婷婷[1]，宋吉祥[2]，周　强[3]
(1.兰州理工大学 电气工程与信息工程学院，甘肃 兰州 730050；
2.兰州理工大学 计算机与通信学院，甘肃 兰州 730050；
3.国网甘肃省电力公司风电技术公司，甘肃 兰州 730000)

摘　要： 针对电力负荷数据聚类过程中 K 均值算法人为指定聚类个数，导致聚类结果陷入局部最小解的问题，提出了基于动态时间归整(DTW)直方图的电力负荷数据聚类方法。利用主成分分析(PCA)法对高维电力负荷数据进行降维；引入直方图法确定负荷数据的初始聚类数目；通过 DTW 将负荷曲线分为 K 个类别；在 MATLAB 仿真平台上验证了该方法的有效性。实验结果表明：本文提出的算法在电力负荷数据聚类分析时减少了运算过程的迭代次数，加快了算法的收敛速度，并且聚类数目达到全局最优解的效果。

关键词： 电力负荷；聚类；动态时间归整；直方图法；K 均值
中图分类号： TP311　　**文献标识码：** A　　**文章编号：** 1000-9787(2020)12-0140-03

Power load data clustering algorithm based on DTW histogram*

HAO Xiaohong[1], LI Yalan[2], GU Qun[2], PEI Tingting[1], SONG Jixiang[2], ZHOU Qiang[3]
(1. School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China;
2. School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China;
3. State Grid Gansu Electric Power Company Wind Power Technology Company, Lanzhou 730000, China)

Abstract: Aiming at the problem that the K-means algorithm needs to artificially specify the number of clustering in the process of power load data clustering, which leads to clustering result falling into the local minimum solution, a clustering method for power load data based on dynamic time warping(DTW) histogram is presented. Principal component analysis(PCA) method is used to reduce the dimension of high dimensional power load data. Histogram method is introduced to determine the initial clustering number of load data, the load curves are divided into K categories by DTW. The effectiveness of the method is verified on MATLAB simulation platform. The experimental results show that the proposed algorithm reduces the number of iterations in the operation process, accelerates the convergence speed of the algorithm, and achieves the effect of global optimal solution in the clustering analysis of power load data.

Keywords: power load; clustering; dynamic time warping(DTW); histogram method; K-means

0　引　言

文献[1]提出电力负荷数据聚类应使得不同类用户的负荷特性差异尽可能大，同一类用户尽可能相似；文献[2]提出并比较了聚类日常电力负荷曲线的特征构造和校准方法，基于时间分辨率特征的条件滤波器为获得能耗模式的洞察提供了新手段；文献[3]通过引入置信半径来得到簇密度，即选取距离最远且簇密度最大的点为初始簇中心；文献[4]提出了基于密度和向异性的初始中性化 K 均值电力负荷模式识别方法的研究；文献[5]选取 K 均值算法进行分析，然后使用降维技术，使用了 DBI 指数(Davies-Bouldin index)来衡量算法的有效性；文献[6]提出一种以牺牲数据结构为代价，来获得聚类效果的大幅提升的算法。现有研究中聚类算法是通过迭代进行求解，运用此类算法对智能电网背景下高维海量的负荷曲线进行聚类分析时，其全局收敛性不能得到保证。

本文提出来了一种可以达到全局最优解、运算时间短、实际操作性强的算法，解决了负荷预测精度低、分时电价不合理等问题，为将来的应用提供理论依据。

收稿日期：2019-08-23
* 基金项目：甘肃省部级资助项目(5227221600KQ)

1　K 均值聚类概述

1.1　K 均值算法

传统算法的步骤[7,8]

输入: 所有数据点 A,聚类个数 K

输出: K 个聚类中心点

1) 随机选取 K 个初始聚类中心

2) Repeat

3) 计算每个点与各中心点之间的距离,将点分配到距离最近的中心点所属的簇中

4) 通过式(1)和式(2),求出 c_j,更新簇的中心点

$$SSE(C) = \sum_{j=1}^{k}\sum_{x_i\in c_j} \| x_i - c_j \|^2 \tag{1}$$

$$c_i = \frac{1}{C_j}\sum_{x_i\in c_j} x_i \tag{2}$$

5) 中心点不发生变化

1.2　聚类算法质量评估

DBI 指标[9] 以类内类样本点到其所属类的中心距离估计类内紧密性,类中心之间的距离表示类间分散性,定义为

$$I_{\mathrm{DBI}} = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq h} D_{k,h} \tag{3}$$

$$D_{k,h} = \frac{\bar{d}_k + \bar{d}_j}{d_{k,h}} \tag{4}$$

式中 $\bar{d}_k$,$\bar{d}_j$ 分别为第 K 类,第 j 类中的数据对象到相应类的类中心的平均距离。$d_{k,h}$ 为第 k 类到第 h 类的类中心的欧氏距离。I_{DBI} 越小表示聚类效果越好。

2　基于动态时间规整的直方图算法

2.1　直方图法

直方图的概念是用条形图表示矩阵的每个元素值。如果直方图是 T 维的,矩阵的列数为 $T-1$。该方法的主要思想是将数据投影到对应的维度上,然后在每个维度上进行区间划分,将每个区间的对象个数与相邻的其他区间进行比较,并选择对象数目超过阈值 ξ 的为峰值。

用直方图对该数据集 $X\{R\times T\}$ 进行统计,X 有 R 个元素,每个元素有 T 维,可以得到用条形图表示的区间集合

$$Group = \{ group_{L_1L_2\cdots L_T} \}\mid_{L_1=1,2\cdots q_1,\cdots L_T=1,2\cdots q_T}$$

每个区间内所含对象个数对应着一个直条高度,定义为

$$Y = \{ y_{L_1L_2\cdots L_T} \}\mid_{L_1=1,2\cdots q_1,\cdots L_T=1,2\cdots q_T}$$

然后,搜寻直方图中的峰值个数,直方图法根据数据样本自身的分布特性将其进行区域划分,可以科学地得到数据的划分结果[10]。

2.2　动态时间规整算法

在时间序列中,相同时间段内的时间序列长度可能并不相等,不同时间轴上的位移导致时间序列之间有差异,两个时间序列想达到相同的状态,必须还原位移。为了更好地求得两个时间序列之间的相似性,动态时间规整(dynamic time warping,DTW)对时间序列进行了扩展和缩短。

如图 1(a)所示,两个时间序列由上下实线表示,虚线表示两个时间序列之间的相似点。DTW 通过对所有相似点的距离求和来度量两个时间序列之间的相似性,称为归整路径距离(warp path distance)。

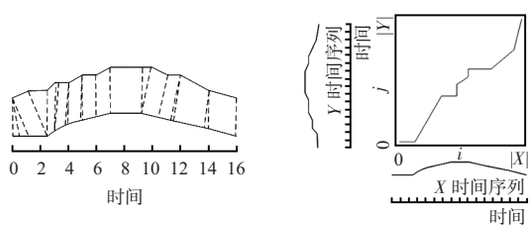
DTW 计算方法: 令要计算相似度的两个时间序列为 X 和 Y,长度分别为 $|X|$ 和 $|Y|$。归整路径的形式为 $S = s_1, s_2, \cdots, s_l$,其中

$$\max(|X|,|Y|) \leqslant l \leqslant |X| + |Y| \tag{5}$$

s_l 的形式为 (i,j),归整路径 S 必须从 $s_1 = (1,1)$ 开始,到 $s_l = (|X|,|Y|)$ 结尾,最后要得到的归整路径是距离最短的一个归整路径且单调递增

$$D(i,j) = Dist(i,j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)] \tag{6}$$

最后求得的归整路径距离为 $D(|X|,|Y|)$,使用动态规划来进行求解。



(a) 时间序列曲线　　(b) 归整路径距离

图 1　两个时间序列曲线及其间的归整路径距离

2.3　基于 DTW 的直方图算法的步骤

1) 负荷数据是多维数据,可设其维度为 m,建立样本空间的 m 维坐标,将每个负荷数据投影在 m 维坐标上。

2) 在 m 维坐标上建立样本直方图,将坐标单位设为 q ($q=0.5$),根据靠左原则(恰好落在刻度线上的投影点,将其归属于左侧的区域),为了直观了解,以二维样本空间为例(PCA 进行降维),将其分别投影在 x 和 y 轴上。

3) 将直方图的统计个数设置一个阈值 ξ,将大于 ξ 值的个数统计起来作为聚类个数。

4) 将时间序列曲线通过 DTW 计算(曲线之间的相似性小于给定阈值,将其归为一类),分为 k 类。

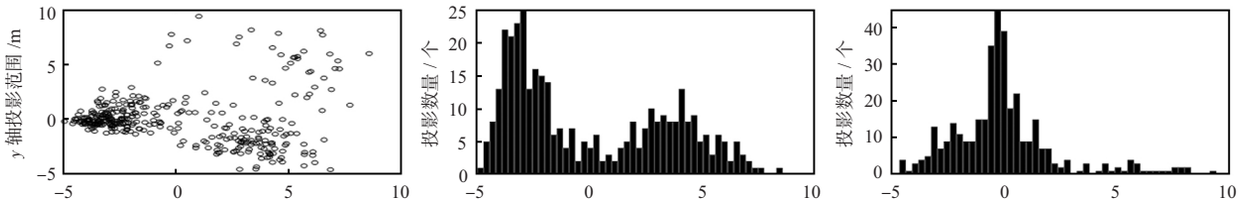
5) 通过 k 类中心曲线的变化来分析负荷曲线变化趋势及所属类型。

3　电力负荷聚类分析

为了验证改进后 K 均值算法的聚类效果,本文在 MATLAB仿真平台上进行实验,将基于 DTW 的直方图算法应用于某市 2017 年日负荷数据中进行应用(每隔 1 h 采集一次,共采集 365 天)。

1) PCA 降维后的负荷数据在二维平面所呈现的状态,如图 2(a)所示; 将横轴称为 X 轴,在 X 轴上的直方图统计如

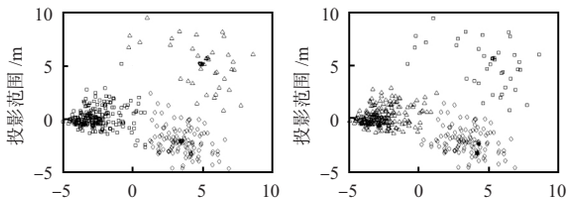
图 2(b) 所示; 将纵轴称为 Y 轴,在 Y 轴上的直方图统计如图 2(c) 所示。本文设定 ξ 值为 22,可以得出聚类个数为 3。



(a) 二维空间投影 (b) X 轴的直方图 (c) Y 轴的直方图

图 2 二维空间投影及直方图统计

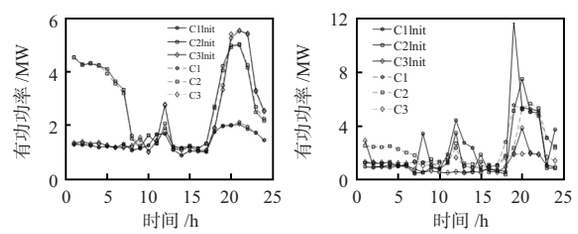
2) 用直方图法确定聚类个数,通过设定阈值计算出聚类个数是 3,如图 3(a) 所示,每类负荷数据的初始聚类中心和最后聚类中心几乎重合。K 均值算法的聚类结果如图 3(b) 所示,初始聚类种类中心和最终聚类中心相差较大。



(a) 直方图算法 (b) K 均值算法

图 3 二种算法聚类中心

3) C3 负荷曲线属于居民用电,波动较小,较为平缓,该曲线与居民的生活习惯密切相关,中午做饭时间稍有起伏,下班后家电器以及健身器材的使用,使得用电量大幅度提升; C2 负荷曲线,夜间的用电量要高于白天用电量,避开用电高峰期,可知这类用户主要在夜间低价时段进行工作,避开用电高峰期,缓解用电压力; C1 属于农业用电,与 C3 的负荷曲线很相似,但晚上用电量要比白天高出许多,晚上进行排水灌溉,如图 4(a) 所示。相同的负荷数据用 K 均值算法聚类时,得到的负荷曲线特征并不明显,如图 4(b) 所示。



(a) 直方图算法 (b) K 均值算法

图 4 二种算法聚类中心时移图

如表 1 所列,从 3 方面对改进前后的算法做了比较,由表可知直方图算法加快了收敛速度,减少了迭代次数,提高了算法有效性。

表 1 算法性能优化对比表

指标	算法	
	K 均值算法	直方图算法
一次耗时/s	0.646	0.117
迭代次数	85	11
DBI	1.932	1.561

4 结束语

随着智能电网的推进与建设,电力负荷数据急剧增加,聚类的准确性降低,在此基础上,本文提出了基于 DTW 的直方图算法,结果证明: 改进后的算法在处理负荷数据聚类时不仅收敛速度快,而且聚类精度高,且 DBI 值也减少了。本文只解决了人为指定初始聚类个数的问题,如何科学确定初始聚类中心将是下一步的研究方向。

参考文献:

[1] 李文江,陈阳. 基于改进 ABC 与 LS-SVM 算法的电力负荷预测的研究 [J]. 传感器与微系统,2013,32(5) : 57 - 59,63.

[2] Al-OTAIBI R,JIN N,WILCOX T,et al. Feature construction and calibration for clustering daily load curves from smart-meter data [J]. IEEE Transactions on Industrial Informatics,2016,12(2) : 645 - 654.

[3] 安计勇,高贵阁,史志强,等. 一种改进的 K 均值文本聚类算法 [J]. 传感器与微系统,2015,34(5) : 130 - 133.

[4] 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.

[5] 程祥. 基于负荷量测数据的电力负荷聚类方法研究 [D]. 杭州: 浙江大学,2017.

[6] 王帅,杜欣慧,姚宏民,等. 面向含多种用户类型的负荷曲线聚类研究 [J]. 电网技术,2018,42(10) : 3401 - 3412.

[7] 郁启麟. K-means 算法初始聚类中心选择的优化 [J]. 计算机系统应用,2017,26(5) : 170 - 174.

[8] 肖琪. 基于优化 K-means 算法的电力负荷分类研究 [D]. 大连: 大连理工大学,2015.

[9] 白雪峰,蒋国栋. 基于改进 K-means 聚类算法的负荷建模及应用 [J]. 电力自动化设备,2010,30(7) : 80 - 83.

[10] 张健沛,杨悦,杨静,等. 基于最优划分的 K-means 初始聚类中心选取算法 [J]. 系统仿真学报,2009,21(9) : 2586 - 2590.

作者简介:

郝晓弘(1960 -) ,男,教授,博士研究生导师,研究领域为分布式能源。

李亚岚(1994 -) ,女,通讯作者,硕士研究生,研究方向为信息信息的获取、融合与处理,E-mail: 2894785172@ qq. com。