

基于改进kmeans++算法的用户分类与 电价政策影响分析

杨 纲,寇 健,严思唯,芦金雨

(上海市电力公司 奉贤供电公司,上海 201400)

Classification of users and analysis of influence of electricity price policy based on improved kmeans++ algorithm

YANG Gang, KOU Jian, YAN Siwei, LU Jinyu

(Fengxian Electric Power Supply Company, Shanghai Electric Power Company, Shanghai 201400, China)

摘要:以上海市张江某工业园大工业用户为研究对象,采用大数据方法分析典型用户用电与电价政策的相关性。针对传统kmeans聚类算法的缺点,提出基于复合距离的kmeans++算法,考虑用电曲线的空间相近性与形态相似性,并优化初值选择,具有更好的聚类效果。采用该算法对大工业用户进行聚类,得到大工业具有5种典型行业用户并分析其原因。基于典型行业用户变压器容量和用电量数据,计算各典型行业用电量标杆值并分析其在新电价政策实施前后的波动情况,明确电力公司后续提升工作的重点方向并对不同用户的用电提出合理意见。

关键词:复合距离;kmeans++;典型行业用户;用电量标杆值;电价政策相关性

Abstract: The large industrial users of an industrial park in Zhangjiang, Shanghai are studied to analyze the relevance between power consumption benchmark value of typical industry users and power price policy using big data technology. To solve problems of traditional kmeans clustering algorithm, a kmeans ++ algorithm based on composite distance is proposed, which takes into account the spatial and morphological similarities of electricity curves, and optimizes the selection of initial values, so clustering effect is better. This algorithm is used to cluster large industrial users. Finally, five typical industrial users are found in large industry and their reasons are analyzed. Based on the transformer capacity and power consumption data of typical industries, the power consumption benchmark values of each typical industry are calculated and their fluctuations before and after the implementation of the new tariff policy are analyzed. These help electricity companies to identify the key direction of follow-up work and put forward reasonable opinions to different users.

Key words: compound distance; kmeans++; typical industry users; power consumption benchmark value; power price policy relevance

0 引言

在用电信息采集系统以及客户服务系统中积累了大量的用户信息,如何从这些数据中分析出具有现实意义的结论已成为智能电网下一个重要课题。而工业负荷作为电网用电侧的重要组成部分,如何有序高效的用电对电力经济与节能减排具有重要意义。因此,依靠大数据的方法,分析不同用户用电与电价政策的相关性,提出相关应对策略,可以帮助电力公司了解用户特点,给不同的用户提供个性化服务与科学用电建议^[1]。

近年来已经开展了大量的研究,文献[2]—文献

[3]采用了不同算法对用电曲线进行聚类,总结了相关用电特征,但缺乏实际应用。文献[4]通过聚类分析了用电特征,以此设计了错峰方案,没有涉及电价政策的相关。文献[5]分析了两部制电价中分时电价的调整对用户用电成本进行了分析,但其分析数据量少且分析方法较为粗糙,并不适合目前精细化智能化的要求。因此,采用大数据手段对用户用电与电价政策之间的关联进行细致分析是十分具有现实意义的。

本文以上海市张江地区大工业用户为研究对象,提出一种基于复合距离的kmeans++算法,充分考虑了用电曲线之间的空间相近性与形状相似性,提高了聚类的准确性。分析典型用户用电特点以及电价政策变化对不同用户的影响关系,提出科学用电方案。

1 基于复合距离的kmeans++聚类算法

聚类算法是一种可以度量样本相似性的方法,

收稿日期:2019-01-22;修回日期:2019-11-12

基金项目:国家电网公司科技项目资助(52090016002M)

This work is supported by Science and Technology Project of State Grid Corporation of China(No. 52090016002M)

其中kmeans算法具有简单、快速、效果好的特点,但是该算法在初值选择、相似性度量等方面依然存在不足,提出基于复合距离的kmeans++算法。

1.1 算法框架

基于复合距离的kmeans++算法包括以下部分:聚类初始中心选定、复合距离计算、聚类主体。图1为算法框架图。

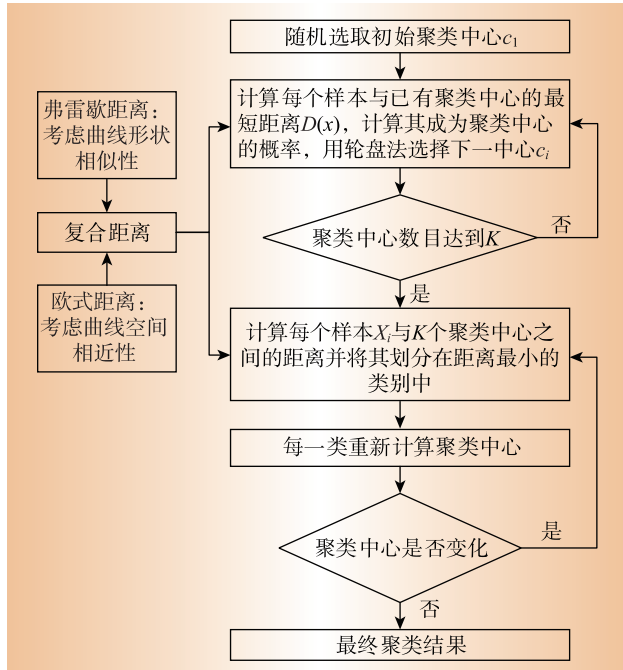


图1 基于复合距离的kmeans++算法框架图

Fig. 1 Framework of kmeans++ algorithm based on compound distance

1.2 考虑形状相似性的复合距离度量

欧式距离虽然可以在一定程度上表征曲线相似关系,但对于具有时间序列特性的电量曲线来说存在一定的局限性^[6-7]。

为了消除欧式距离的不利性,并且将曲线在形态上的相似性也考虑进聚类过程中,引入弗雷歇距离(fréchet distance, FD)。

FD的基本定义为:根据抽象空间理论,定义在单位区间上的映射 $\gamma: [0,1] \rightarrow S$ 是连续映射,则称 γ 是S上的连续曲线。那么就可以假设A和B是S上的两条连续曲线,即 $A: [0,1] \rightarrow S, B: [0,1] \rightarrow S$;又设 α 和 β 是单位区间的两个重参数化函数,即 $A: [0,1] \rightarrow [0,1], B: [0,1] \rightarrow [0,1]$;则曲线A与曲线B的弗雷歇距离 $FD(A,B)$ ^[8]为

$$FD(A,B) = \inf_{\alpha, \beta} \max_{t \in [0,1]} \{d(A(\alpha(t)), B(\beta(t)))\} \quad (1)$$

式中: d 是S上的度量函数。两条曲线之间的弗雷歇距离越小说明两条曲线在形态上越相似。然而在电力系统中采集到的数据并不是连续的,而是离散的,需要对上述公式进行离散化。

给定两条曲线 $A = \{a_1, a_2, \dots, a_m\}$, $B = \{b_1, b_2, \dots, b_n\}$, 1个沿着A和B的组合步是1个沿着A的 k 步 $\{A_i\}_{i=1,2,\dots,k}$ 和1个沿着B的 k 步 $\{B_i\}_{i=1,2,\dots,k}$ 组成,使得对于 $1 \leq i \leq k$, α 和 β 是单位区间的两个重参数化函数^[6]。

$$FD^n(A,B) = \inf_{\alpha, \beta} \left\{ \max_{t \in \{t_k\}_{k=0}^n} \{d(A(\alpha(t)), B(\beta(t)))\} \right\} \quad (2)$$

为了简化相关计算, d, A, α, B, β 均为固定的已知函数,通过 t 的遍历进行计算。

在欧式距离的基础上增加弗雷歇距离构造考虑曲线形态相似性的复合距离,定义如下:假设两曲线 $A = \{a_1, a_2, \dots, a_m\}$, $B = \{b_1, b_2, \dots, b_n\}$, 两条曲线的欧式距离表示为

$$ED(A,B) = \sum_{i=1}^n \|a_i - b_i\|^2 \quad (3)$$

则复合距离 $dis(A,B)$ 可表示为

$$dis(A,B) = \alpha_1 ED(A,B) + \alpha_2 \beta FD^n(A,B) \quad (4)$$

式中: α_1, α_2 为距离权重系数,且满足 $\alpha_1 + \alpha_2 = 1$,可以调整不同距离的重要程度,这里取 $\alpha_1 = \alpha_2 = 0.5$; β 为平衡系数,目的是保证两距离数值数量级相近防止数据不平衡造成结果偏差。

复合距离不仅考虑了空间上的相近性,同时利用弗雷歇距离考虑了曲线形态上的近似性,同时保证精度不会丢失。

1.3 初始聚类中心选定

传统的kmeans算法对初值的选定是采用随机的方法,这使得多次聚类结果稳定性很差^[9]。采用kmeans++方法对初值的选择进行优化。其核心思想是选取第 i 个聚类中心时应是其距离以选出的聚类中心最远^[10]。采用步骤如下:

用户的用电曲线集合为 $S = \{X_1, X_2, \dots, X_n\}$, $X_i = \{x_1, x_2, \dots, x_{96}\}$ 。

(1) 从输入的集合中随机选取一个对象 X_i 作为初始聚类中心 c_1 。

(2) 计算样本中每个对象与已有的聚类中心的最短距离,用 $D(X_i)$ 表示,即

$$D(X_i) = \min \{dis(X_i, c_k)\} \quad (5)$$

(3) 计算样本中每个对象成为下一个聚类中心的概率 P (如式(6)),选取概率最大的作为下一个聚类中心。

$$P = \frac{D(X_i)^2}{\sum_{X \in S} D(X_i)^2} \quad (6)$$

(4) 重复步骤(2)与(3),直到选出的聚类中心数目达到要求 k 个。

1.4 聚类有效性检验

为了评价与衡量聚类的结果的好坏,需要指定

统一的评价指标,这里采用业界比较常用的 *DBI* 指标作为评价标准^[11-12]。

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\overline{X}_i + \overline{X}_j}{\|c_i - c_j\|^2} \right) \tag{7}$$

式中： $\overline{X}_i$ 、 $\overline{X}_j$ 为任意两类别的类内距离平方和，*DBI* 越小意味着类内距离越小，同时类间距离越大。

2 算例分析

2.1 分析流程

典型行业用户用电量标杆值与电价政策相关性分析包括以下内容:数据采集、数据预处理、核心算法、结果可视化与相关性分析。图2为分析框架图。

2.2 分析对象

本文选取上海张江地区某大型工业园区内的 95 用户的电量数据为研究对象。所采集的数据为每 15 min 采集一次,一天 96 个点,夏季属于工业用电的高峰期,对电价政策变化的敏感度更高,为了更好的体现变化相关性,采用夏季的相关数据。园区内典型的典型行业与用户对应表如表 1 所示。

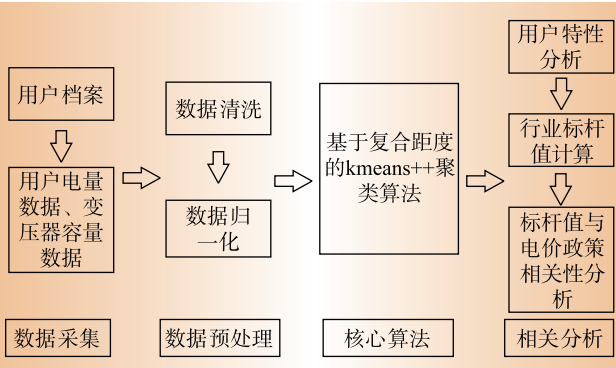


图2 分析框架图
Fig. 2 Analytical framework

表1 行业与用户对照表
Table 1 Contrast table of industries and users

行业	用户编号
纸制品制造	10 21 22 32 33 46 91
金属加工制造业	2 13 23 35 45 54 63 77 81 90 95
肥料制造	1 3 11 18 30 53
电子原件制造	6 17 31 36 43 47 52 61 62 72 76 78 84 92
电子计算机制造	5 16 25 41 42 44 51 60 68 71 75 82 83 85 86
印刷业	4 14 15 24 29 55 65
化学药品制造	8 9 37 40 48 56 64 69 70 73 80 81 89 93
生物生化制品	7 19 26 28 49 50 57 88
涂料、油墨、颜料制造	12 20 27 38 39 58 59 66 67 74 79 87 94

2.3 数据预处理

考虑数据的幅值差异很大可能的影响,对数据进行归一化处理^[10]。

$$X_h = \frac{x_h}{x_{\max}} \tag{8}$$

式中： X_h 为归一化之后 *h* 时刻的电量值； x_h 为归一化之前 *h* 时刻的电量值； $x_{\max}$ 为最大电量值。

2.4 聚类结果

在聚类之前需要优先确定最佳聚类数目。本文采用 *DBI* 指数判别法, *DBI* 指数越小,说明聚类效果越好,绘制 *DBI* 指数与聚类数目的关系图,如图 3 所示。大工业不同行业至少应具有 3 种及以上用电模式,聚类数目应大于等于 3。由图 3 可知,最佳聚类数目为 5。

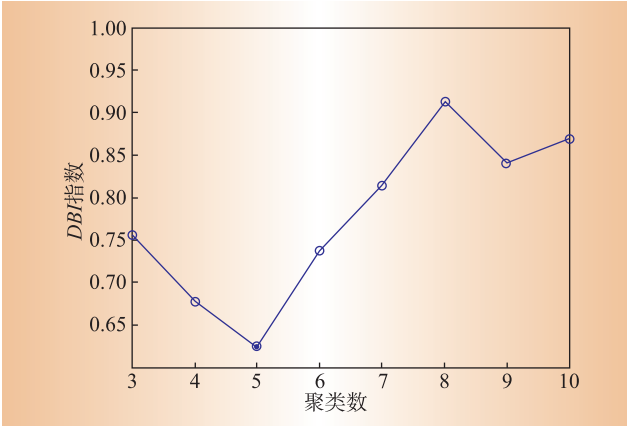


图3 不同聚类数目对应的 *DBI* 指数值
Fig. 3 *DBI* value of different number of clusters

以 5 为聚类数目,采用基于复合距离的 *kmeans++* 算法,对前述 2.2 节中对象进行聚类。聚类的结果如图 4、图 5、表 2 所示。

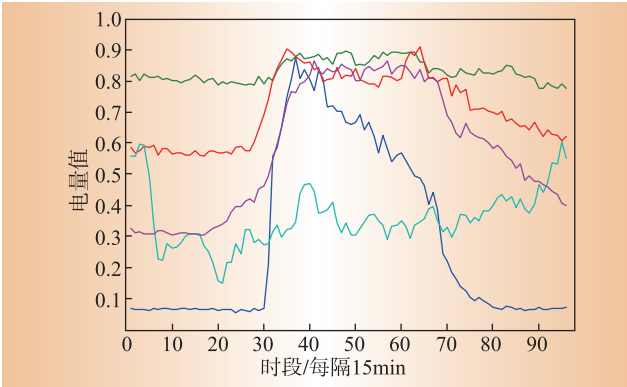


图4 典型行业用户用电模式图
Fig. 4 Power consumption patterns of typical industry users

根据上述结果,并结合每一聚类结果对应的典型行业,可以发现该工业园区的工业型企业具有 5 种基本用电模型,可以总结出 5 个典型行业代表,分别为:日间用电型,全天高峰型,全天次高峰型,随机型,日夜差别型。

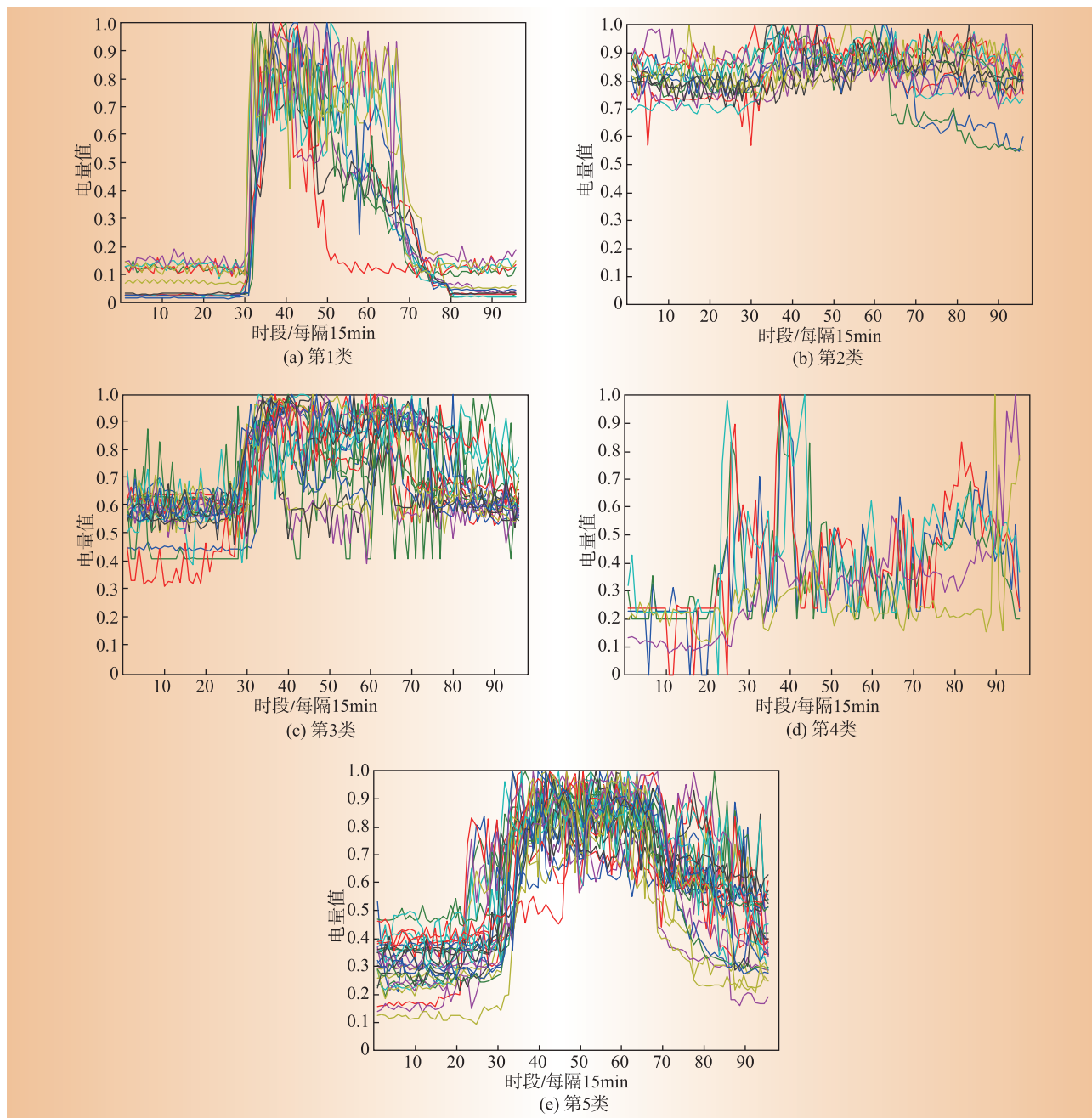


图5 不同聚类结果图

Fig. 5 Different clustering results

第1类为日间用典型,其主要特点是用电时间段主要集中在日间,夜间用电很低。代表行业是印刷业,印刷业是一种半自动化的行业,需要人工进行

表2 不同聚类结果与典型行业对应表

Table. 2 Different clustering results and typical industries contrast table

聚类类别	典型行业
类别1	印刷业
类别2	电子元件、计算机制造业
类别3	金属加工、电机制造
类别4	肥料制造
类别5	生物生化制品、医药

编辑工作,与工作人员的工作时间息息相关,在日间集中办公,用电量高,夜间人员下班,用电量低。

第2类为全天高峰型,其主要特点是全天的用电量都保持在高位,没有明显的峰谷差别。其代表行业为电子元件制造、计算机制造,这些行业属于自动化程度很高的行业,可以实现全天的自动化工作,这使得该类企业的用电量较高且比较平稳。

第3类为全天次高峰型,其主要特点是用电量全天较高,但是日夜略有差异。夜间用电量处于中等水平。其代表行业为金属加工制造业、电机制造业,这些行业均为典型的传统加工制造类行业。其自动化程度要低于第2类行业,其用电量与人员的

工作时间有一定的关系。一般该类行业的企业为了减少机器的关停损耗、增加产量等,在夜间也会保持一定的制造量,使得夜间具有一定水平用电量。

第4类为随机型,其主要特点是全天用电量均在中等水平附近波动,没有固定规律的峰谷变化,与人员工作时间没有明显的相关性。这类结果的代表行业为肥料制造业,所获得该行业用户的数据较少,不同用户之间存在着很大的差异性。

第5类为日夜差别型,其特点是日间与夜间用电量具有较大的差异性,其在夜间用电处于一个较低的水平。其代表行业为生物生化制品、医药研究与制造,都可以归结为生物化学类,该行业是研究型行业,与人员活动关系密切,具有一定的自动化程度。

为了检验聚类的实际效果,分别计算基于复合距离的kmeans++算法与传统算法的DBI指标值,如表3所示。

根据表3的结果可以看出,通过本文基于复合距离的kmeans++算法在DBI指标上要低于传统基于欧式距离的kmeans算法与kmeans++算法,说明本文的方法确实可以获得更好的聚类结果。

表3 聚类效果评价

Table 3 Clustering effect evaluation

方法	DBI 指数值
基于复合距离的kmeans++算法	0.628
基于欧式距离kmeans算法	0.761
基于欧式距离kmeans++算法	0.693

2.5 电价政策对不同用户的影响分析

分析、选择出大工业中的典型行业用户,从而分析典型行业用户用电对电价政策改变的敏感度。选择该工业园区在电价政策改变前2个与改变后7个连续工作日作为分析样本。改变前电价政策为单一制分时电价,改变后为两部制分时电价。

为了更加方便的比较电价政策对不同行业用户用电的影响,设立行业标杆值,其表达式如下

$$IBv = \frac{1}{n} \sum_{i=1}^n \frac{Q_i}{Q_{trani}} \eta \quad (9)$$

式中:IBv为行业标杆值(benchmark value of industry);n为该行业的用户数量; Q_i 为第*i*个用户的日用电量; Q_{trani} 为第*i*个用户对应的变压器容量; η 为单位转换系数。相关结果如图6所示。

根据图6的结果,可以发现规律如下。

(1) 电价政策的变化对大工业内不同行业用电的影响是不一致的,在新的电价政策实施后,印刷业行业标杆值逐步下降;电子元件、计算机制造业的标

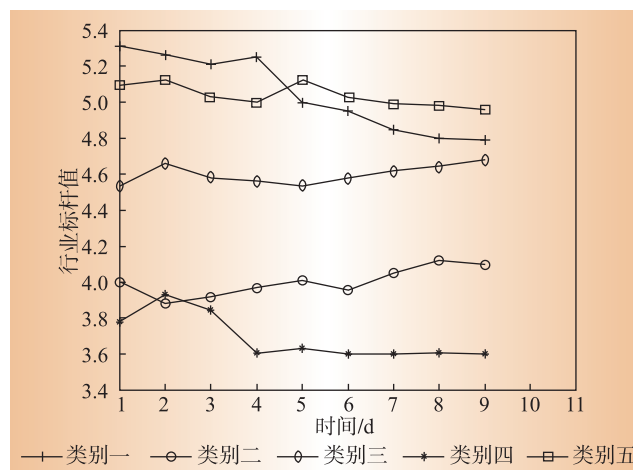


图6 典型行业用电量标杆值波动图

Fig. 6 Benchmark value fluctuation chart of typical industries

杆值逐步上升并稳定;金属加工、电机制造业行业标杆值略有上升;肥料制造业的标杆值快速下降并快速趋于稳定;生物生化类、医药类的标杆值存在波动性,总体水平略高于政策实施前。也就是说新政策的实施,对电子元件、计算机制造、金属加工、电机制造、生物生化制造、医药研究与制造行业的用电具有拉动作用,对印刷业与肥料制造业存在一定的抑制作用。分析其原因在于,两部制分时电价政策相对于单一制分时电价,多要承担一部分基于变压器容量的基础电费,对于自动化程度较高的行业来说,可以通过提高负荷率来降低电价成本,而对于印刷业与肥料制造,难以较大规模的提高负荷率,而通过降低用电量以降低成本具有更好的效果。

(2) 大工业内的不同行业用电对电价政策的敏感度并不一致。印刷业标杆值在新政策实施2日后才发生明显变化;电子元件、计算机制造业的标杆值在新政策实施后就逐步变大;金属加工、电机制造业行业标杆值变化敏感度较低;肥料制造业的标杆值在新政后便有所变化;生物生化类、医药类的标杆值在新政策实施后第3天才发生明显变化。

根据上述结论,实施两部制电价将拉动自动化程度较高产业的用电而抑制印刷与肥料产业的用电。所以电力公司应坚定的实施两部制电价,促进高新技术产业的向上发展而抑制印刷等限制类产业。同时为了保证在结构优化的过程中,部分企业能够逐步适应新的调整,在进行两部制电价实施的时候,应适当放宽一些措施:① 放宽基本电价计费方式变更周期的限制;② 放宽减容期限的限制。在电价政策实施之前应告知所有企业做好相关的准备工作。

对于产业终端的企业来说,根据电价政策,制

定合理的生产计划,将有利于降低电费成本。对于所有产业的企业来说都应该进行技术革新,提高生产效率,优化生产结构。对于印刷类企业,其日夜差别大,需要提高印刷线自动化水平,并提高日间生产效率。电子元件、计算机制造、金属加工、电机制造、生物生化制造等自动化程度较高的企业应提高夜间生产效率。肥料生产企业应制定合理的生产计划,有序的进行生产,并提高制造效率,降低污染排放。

3 结束语

本文围绕着典型行业典型用户做了相关研究。

(1) 为了弥补传统基于欧式距离的kmeans算法的不足,提出一种基于复合距离的kmeans++算法,充分考虑了曲线在空间的相近性与曲线形态的相似性,并对初值选择进行了优化,结果表明,该方法具有更好的聚类效果。

(2) 选取了大工业的典型行业,本文选取的工业区分成5种用电类型:日间用电型,全天高峰型,全天次高峰型,随机型,日夜差别型,分析了其对应的行业与产生原因。

(3) 两部制分时电价对不同行业用电拉动作用不同,对于自动化程度较高的电子元件制造、金属加工与生物生化医药制造的用电具有拉动作用,对于印刷与肥料制造具有一定的抑制作用。D

参考文献:

- [1] 王继业,季知祥,史梦洁,等.智能配用电大数据需求分析与应用研究[J].中国电机工程学报,2015,35(8):1 829-1 836.
WANG Jiye, JI Zhixiang, SHI Mengjie, et al. Scenario analysis and application research on big data in smart powerdistribution and consumption systems[J]. Proceedings of the CSEE, 2015, 35(8):1 829-1 836.
- [2] LI K. Analysis on residential electricity consumption behavior using improved K-means based on simulated annealing algorithm[C]. 2016 IEEE Power and Energy Conference at Illinois (PECI), 1-6.
- [3] 李智勇,吴晶莹,吴为麟,等.基于自组织映射神经网络的电力用户负荷曲线聚类[J].电力系统自动化,2008,32(15):66-70,78.
LI Zhiyong, WU Jingying, WU Weilin, et al. Customers load profile clustering using the som neural network [J]. Automation of Electric Power Systems, 2008, 32(15):66-70, 78.
- [4] 胡长华.基于大用户用电行为分析的错峰管理系统研究与设计[J].现代计算机:上下旬,2014(14):42-47.
HU Changhua. Research and design of peak load shifting management system based on analysis of electrical behavior[J]. Modern Computer, 2014(14):42-47.
- [5] 夏启明.企业合理利用电价政策对成本的影响[J].电力需求侧管理,2001,3(4):41-43.
XIA Qiming. Utilize the electric charge policy's influence on enterprise's cost with reasonableness [J]. Power Demand Side Management, 2001, 3(4):41-43.
- [6] 段青,赵建国,罗珂.基于形状相似的日负荷曲线多重聚类分析及其应用[J].电气应用,2008,27(20):53-56.
DUAN Qing, ZHAO Jianguo, LUO Ke. Application and electrical load data clustering algorithm using multihierarchy analysis based on time series curve contour similarity [J]. Electrotechnical Application, 2008, 27(20):53-56.
- [7] 苑津莎,李中.基于形状相似距离的K-means聚类算法[J].华北电力大学学报(自然科学版),2009,36(6):98-103.
YUAN Jinsha, LI Zhong. K-means clustering algorithm based on the shape similarity distance [J]. Journal of North China Electric Power University (Natural Science Edition), 2009, 36(6):98-103.
- [8] 朱洁,黄樟灿,彭晓琳.基于离散Fréchet距离的判别曲线相似性的算法[J].武汉大学学报(理学版),2009,55(2):227-232.
ZHU Jie, HUANG Zhangcan, PENG Xiaolin. Curve similarity judgment based on the discrete Fréchet distance [J]. Journal of Wuhan University (Natural Science Edition), 2009, 55(2):227-232.
- [9] 徐磊,杨秀,张美霞.基于数据挖掘的工业用户用电行为分析[J].电测与仪表,2017,54(16):68-74.
XU Lei, YANG Xiu, ZHANG Meixia. Industrial users of electricity behavior analysis based on data mining[J]. Electrical Measurement & Instrumentation, 2017, 54(16):68-74.
- [10] 郑丹,王潜平. K-means 初始聚类中心的选择算法[J].计算机应用,2012,32(8):2 186-2 188,2 192.
ZHENG Dan, WANG Qianping. Selection algorithm for K-means initial clustering center[J]. Journal of Computer Applications, 2012, 32(8):2 186-2 188, 2 192.
- [11] CHICCO G. Overview and performance assessment of the clustering methods for electrical load pattern grouping[J]. Energy, 2012, 42(1):68-80.
- [12] 杨燕,靳蕃, KAMEL Mohamed. 聚类有效性评价综述[J].计算机应用研究,2008(6):1 630-1 632,1 638.
YANG Yan, JIN Fan, KAMEL Mohamed. Survey of clustering validity evaluation [J]. Application Research of Computers, 2008(6):1 630-1 632, 1 638.

作者简介:

杨纲(1973),男,上海浦东人,大学本科生,经济师,主要研究领域为电力营销;

寇健(1993),女,辽宁阜新人,硕士研究生,助理工程师,主要研究领域为电力系统运行及控制;

严思唯(1991),男,上海南汇人,大学本科生,助理工程师,主要研究领域为电力营销;

芦金雨(1992),女,江苏南京人,大学本科生,助理工程师,主要研究领域为电气工程。

(责任编辑 柴明哲)