

基于K-Means聚类算法的行业用电行为分析

利用数据挖掘与数理统计的相关技术,分析某省各个地区行业用电行为,进而研究各地区行业用电特征以及客户群体特征,在一定程度上指导电力营销的行业方向和地区重点,同时也能作为用电量运营监测的参考。

周小明¹ 苏安龙¹ 杨宏宇^{2/1}. 国网辽宁省电力有限公司 2. 南瑞集团公司
信息通信技术分公司

0 引言

用电量作为衡量一个地区工业生产活跃度以及企业运行状态的重要指标,与该地区的经济发展有着密切的联系。在经济增长时期,由于投资扩大,需求旺盛,用电量也随之持续上升;而在经济衰退时期,由于工厂开工不足,设备闲置,用电量就会下降。除此之外,行业用电量还能反映不同地区的行业 and 产业发展差异,为供电企业不同地区之间优化资源配置提供依据。因此,大到宏观经济预测,小到输配电设施调整改造,都能体现行业用电行为监测的重要意义。



周小明/高级工程师

关键词/Keywords

用电行为·

K-Means·

聚类·

1 K-Means聚类算法

1.1 算法原理及模型

聚类是一种在没有先验知识的条件下,根据某种相近程度的度量指标,对数据自动进行子集划分的技术。所形成的子集内部数据的结构特征相近,不同子集之间的数据结构特征有较大差异。

K-Means算法,也被称为K-平均或K-均值,是一种得到最广泛应用的经典聚类算法。它是将各个聚类子集内的所有数据样本的均值作为该类的代表点,

算法的主要思想是通过迭代过程把数据集划分为不同的类别,使得评价聚类性能的准则函数达到最优,从而使同一个聚类中的对象相似度较高,而不同聚类中的对象的相似度较小。

K-Means算法的空间需求是适度的,因为只需要存放数据点和质心,所需要的存储量为 $O[(m+k)n]$,其中 m 是点数, n 是属性数。同时时间需求也是适度的-基本上与数据点个数线性相关,所需要的时间为 $O(I, K, m, n)$,其中, I 为收敛所需要的迭代次数。

K-Means模型一般都采用平方误差准则,定义为

$$\min E = \min \sum_{i=1}^k \sum_{p \in C_i} |p - m_i| \quad (1)$$

式中, E 为数据集中所有样本的平方误差的综合; p 为空间中的点,表示给定的样本; m_i 为聚类 C_i 的平均值。这个准则试图使生成的聚类结果尽可能的紧凑和独立。

K-Means算法中判断样本点属于哪一类,是依据该点到类质心的距离来确定的,样本点会被划分到距离最近的一类中,距离函数一般采用欧氏距离。

$$\text{Euclid}(x, y) = \sqrt{\sum (x_i - y_i)^2} \quad (2)$$

式中, x_i 为样本 x 的第 i 个变量值; y_i 为类质心 y 的第 i 个变量值。

1.2 K-Means算法实现步骤

K-Means基本算法步骤为:①选择 K 个点作为初始质心;②计算各个点到 K 个质心的距离;③根据最近原则将所有点分到 K 个类;④计算 K 个类的质心;⑤重复步骤2、步骤3和步骤4,直到质心不发

生变化；⑥完成聚类。图1为K-Means算法流程图。

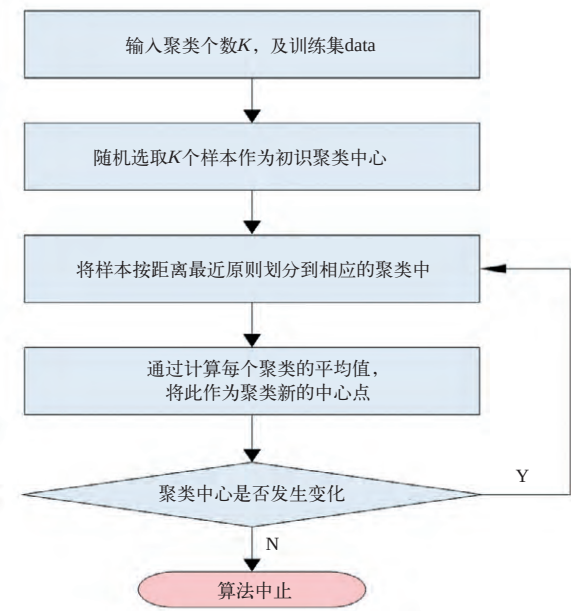


图1 K-Means算法流程图

K-Means算法的优点：解决聚类问题的一种经典算法，简单、快速；对处理大数据集，该算法是相对可伸缩和高效率的；当各类的样本是密集的，而且类与类之间区别明显时，其效果较好。

2 基于K-Means聚类算法的行业用电行为分析

2.1 样本数据

通过某省电力公司用户应收表可以核算出每个客户每个月的用电量，本次研究共选取了2010~2013年共4年的数据，用户数为181 181个。

为了便于深入分析行业用电行为，按照国家行业和产业的划分标准对用户进行分类。在所有行业中，用电量主要集中在工业，因此又将工业再划分为3个子行业，分别为采矿业、制造业以及电力、燃气及水的生产和供应业。用户的行业和产业分类情况如表1所示。

表1 用户行业、产业分类

行业	行业代码	产业	用户数 / 个
农业、林业、牧业和渔业	A000	第一产业	41 780
采矿业	B000	第二产业	10 475
制造业	C000	第二产业	60 306
电力、燃气及水的生产和供应业	D000	第二产业	7 427
建筑业	E000	第二产业	14 346
交通运输、仓储和邮政业	F000	第三产业	494
信息传输、计算机服务和软件业	G000	第三产业	14 275
商业、住宿和餐饮业	H000	第三产业	7 239
金融、房地产、商务和居民服务业	J000	第三产业	7 986
公共事业及管理组织	M000	第三产业	16 853

根据用户应收表的电量信息汇总最终形成如表2所示的用户用电明细数据。

表2 用户用电明细数据

地市	客户编号	行业类别	月份	电量 / kW·h
地市1	0023000588	(二) 制造业	201001	1 824 000
地市1	0023000663	(二) 制造业	201002	890 400
地市1	0023485378	(二) 制造业	201003	1 281 360
地市1	0023485378	(二) 制造业	201004	1 724 640
地市1	0023485378	(二) 制造业	201005	1 399 320
地市1	0023485378	(二) 制造业	201006	1 735 080
地市1	0023485381	(二) 制造业	201007	933 700
地市2	0023485743	(二) 制造业	201001	1 207 400
地市2	0023485743	(二) 制造业	201002	785 350
...	...	...	...	...

本研究主要关注各个行业在不同地区的分布特征情况，因此需要将行业用电量作为列指标，对数据库中的用户电量行数据进行汇总，然后做行转列处理。得到的样本数据为2010年1月~2013年12月的数据，共672条。处理之后的数据如表3所示。

表3 行转列后的样本数据

月份	地区	采矿业	制造业	...
201001	地市1	144 277 496	1 036 857 293	...
201001	地市2	5 241 062	649 040 016	...
201001	地市3	99 610 208	30 085 154	...
201001	地市4	61 564 827	386 945 404	...
201001	地市5	58 126 200	682 427 443	...
...	...	...	...	...

2.2 K-Means聚类结果

本研究聚类所采用的工具为Pluto数据挖掘平台,将10个行业用电量作为属性变量,聚类个数设为3。观察用电量的数据,可以看到用电量大多为 10^6 以上,将初始值设为1000000, K-Means聚类结果如表4所示。

从表4中可以看出,3个聚类的样本数分别为94、210和368,各占样本总数的31.25%、13.99%和54.76%。表5列出了各聚类的中心点,从3个聚类的中心点分布可以看出,农业、林业、牧业和渔业,信息传输、计算机服务和软件业等行业差距较大。

表4 K-Means聚类结果

聚类标号	样本数 / 个	比例 (%)	比例 (%)
1	210	31.25	31.25
2	94	13.99	13.99
3	368	54.76	54.76

表5 K-Means 聚类各行业分布

行业	聚类一	聚类二	聚类三
A000	4 814 170.114 3	24 213 805.925 5	6 628 190.415 8
B000	91 070 792.214 3	36 393 151.968 1	73 772 437.179 3
C000	799 204 298.119	756 782 939.308 5	196 588 727.279 9
D000	14 084 642.738 1	70 558 725.946 8	12 189 123.437 5
E000	6 117 097.157 1	37 119 200.946 8	3 996 196.660 3
F000	15 805 957.152 4	58 474 535.138 3	12 391 499.337
G000	2 840 747.6	21 461 257.212 8	2 047 073.918 5
H000	14 949 639.642 9	103 784 088.074 5	6 755 328.222 8
J000	8 002 697.633 3	55 013 878.053 2	2 677 967.793 5
M000	12 753 067.090 5	88 655 514.468 1	7 988 561.429 3

2.3 聚类结果特征分析

2.3.1 地域特征

3个聚类的地域分布情况如表6所示。

表6 K-Means 聚类各地域分布

地域	聚类一	聚类二	聚类三
地市1	48		
地市2	1	46	1
地市3			48
地市4		48	
地市5	47		1
地市6	47		1
地市7			48
地市8	47		1
地市9			48
地市10			48
地市11			48
地市12			48
地市13	18		30
地市14	2		46

从以上聚类结果的地域分布状况可以看出,聚类结果有明显的地域分布特征,群体在地域分布上特点为:地市8、地市5、地市6和地市1聚到群体1中,地市4和地市2聚到群体2中,地市12、地市9、地市10、地市11、地市3、地市7和地市14聚到群体3中。地市13有60%的样本在群体3中,40%的样本在群体1中,从时间变化趋势上来看,地市13在近期从群体3转向了群体1,因此可以将其归为群体1中。

2.3.2 行业特征

(1) 农业、林业、牧业和渔业

图2为农业、林业、牧业和渔业用电量的分布范围。群体1和群体3为 $[0, 22\ 000\ 000]$,群体2为 $[22\ 000\ 000, 60\ 000\ 000]$,群体2农业发达程度高,相应用电量也大,其他地区用电量相对较少。并且群体1和群体2的农业用电量从2010年开始都成上升趋势,但群体2的增长速度最为明显,群体3整体呈现一个很低的增长趋势,在前两年有几个比较高的点,为朝阳前两年的数据,在2012年后突然大幅度降低。

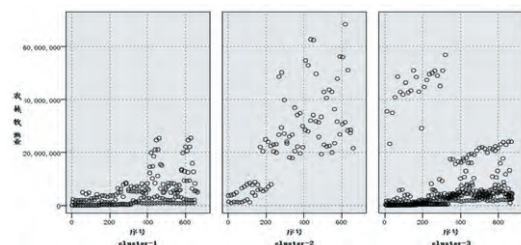


图2 农业、林业、牧业和渔业用电量分布范围

(2) 采矿业

图3为采矿业用电量的分布范围。群体1和群体3为 $[0, 20\ 000\ 0000]$,群体2为 $[0, 60\ 000\ 000]$,与农业用电量相比,采矿业用电量在地区之间的分布正好相反,群体2采矿业用电量较少,而其他地区在采矿业方面用电量都较多,相对来说群体3的采矿业用电增速最大。

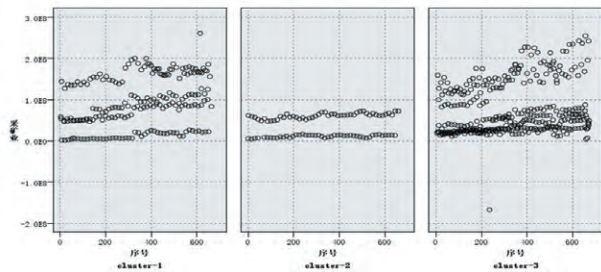


图3 采矿业用电量分布范围

(3) 制造业

图4为制造业用电量的分布范围。群体1和群体2为 $[400\ 000\ 000, 1\ 200\ 000\ 000]$ ，群体3为 $[0, 400\ 000\ 000]$ ，说明群体1和群体2的地市制造业的行业规模都较大，用电量多，并且增速也快，相对群体1来说，群体2增长的速度更为明显，而群体3地市制造业相对落后，增速也较慢。

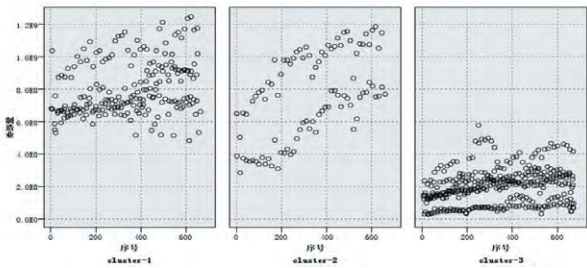


图4 制造业用电量分布范围

(4) 电力、燃气及水的生产和供应业

图5为电力、燃气及水的生产和供应业用电量的分布范围。群体1和群体3为 $[0, 40\ 000\ 000]$ ，群体2为 $[25\ 000\ 000, 150\ 000\ 000]$ ，群体2的用电量比其他地市的用电量都要大很多，说明该行业在群体2发展比较迅速，其他地市基本变化不大。该行业的数据有很强的冬季型用电特征，每年的11月到次年的3月用电量明显很高，剩余月份用电量降低很快。其中群体1体现的最为明显，群体2由于数据较少，不是很明显，群体3在2010年和2011年呈现的冬季特征不是很明显，后两年数据呈现的比较明显。

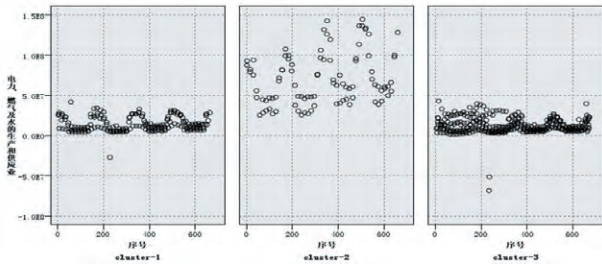


图5 电力、燃气及水的生产和供应业用电量分布范围

(5) 交通运输、仓储和邮政业

图6为交通运输、仓储和邮政业用电量的分布范围。群体1和群体3为 $[0, 50\ 000\ 000]$ ，群体2为 $[40\ 000\ 000, 100\ 000\ 000]$ 。该群体在3个群中都存在比较强的冬季用电特征，即每年的11月到次年的3月用电量明显很高，剩余月份用电量降低很快。群体2中都具有，群体1中的上面

两个城市地市1、地市8，群体3中的上面两个地市11和地市12也具有相同的特征。其余城市每年基本变化不大。

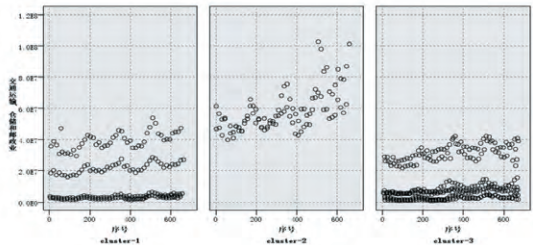


图6 交通运输、仓储和邮政业用电量分布范围

(6) 信息传输、计算机服务和软件业

图7为信息传输、计算机服务和软件业用电量的分布范围。群体1和群体3为 $[0, 5\ 000\ 000]$ ，群体2为 $[10\ 000\ 000, 35\ 000\ 000]$ 。该行业在所有地市整体呈现一个上升趋势，群体1和群体3增速较慢，同时在2011年底有一个明显的增长，但之后变化也很小，群体2增速非常快，说明该行业在地市2和地市4发展迅速。

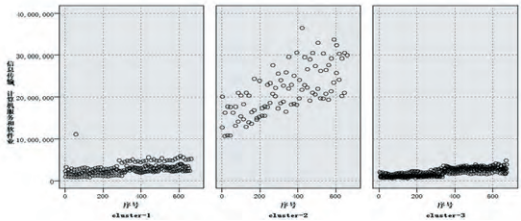


图7 信息传输、计算机服务和软件业用电量分布范围

(7) 金融、房地产、商务和居民服务业

图8为金融、房地产、商务及居民服务业用电量的分布范围。群体1为 $[0, 30\ 000\ 000]$ ，群体3为 $[0, 10\ 000\ 000]$ ，群体2为 $[30\ 000\ 000, 120\ 000\ 000]$ 。群体2的用电量比其他地市的用电量都要大很多，说明该行业在地市2和地市4发展比较迅速，其他地市基本变化不大。该行业的数据有很强的冬季型用电特征，每年的11月到次年的3月用电量明显很高，剩余月份用电量降低很快，和电力、燃气及水的生产和供应业发展很相似。其中群体1和群体3体现的都很明显，群体2由于数据较少，不是很明显。

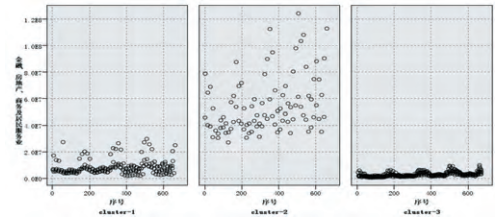


图8 金融、房地产、商务及居民服务业用电量分布范围

(8) 公共事业及管理组织

图9为公共事业及管理组织用电量的分布范围。群体1和群体3为 $[0, 20\ 000\ 000]$ ，群体2为 $[50\ 000\ 000, 150\ 000\ 000]$ 。该行业用电特征与金融、房地产、商务和居民服务业发展很相似。

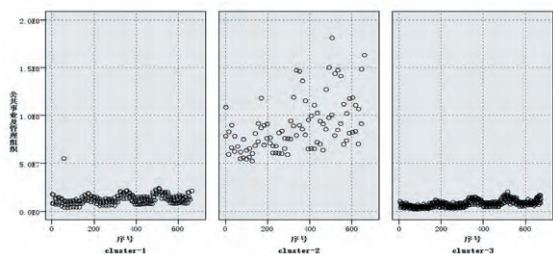


图9 公共事业及管理组织用电量分布范围

3 结束语

通过以上聚类和分析，可以发现辽宁省各行业用电行为具有以下特征。

一是行业用电量具有明显的地域特征。群体1的地市采矿业和制造业用电量都处于比较高的水平，其中制造业发展速度比较快，采矿业速度比较缓慢，其他行业用电量处于比较低的水平。群体2的地市除了在采矿业上的用电量比较少之外，其他行业用电量均比较高，并且发展比较迅速。群体3的地市采矿业处于比较高的用电水平，并且发展速度也比较快，其他行业用电量均处于比较低的水平。

二是行业用电量具有明显的时间特征。农业在2012年之前，用电量以地市14为主，从2012年之后，以群体2城市的用电量为主体，总体来看，农业在全省用电中呈上升趋势，并且增速较快；采矿业的用电量主要集中在除了群体2以外的地区，并且在持续增长；制造业主要集中在群体1和群体2城市，并且增速也较快；电力、燃气及水的生产和供应业和建筑业的用电量都主要集中在群体2，其中电力、燃气及水的生产和供应业用电量变化不大，但是都具有比较明显的冬季用电特征，即每年的11月到次年的3月用电量明显很高，剩余月份用电量降低很快。建筑业在群体2用电量增速明显，在其他城市增速较小。

第三产业的各行业用电量均集中在群体2，且增速明显，其他城市用电量不大并且增速也不明显。其中交通运输、仓储和邮政业（其中部分城市如：群体1的地市1、地市8，群体3的地市11

和地市12，群体2的地市2、地市4）、金融、房地产、商务和居民服务业、公共事业及管理组织3个行业的用电特征非常相似，都具有比较明显的冬季用电特征，即每年的11月到次年的3月用电量明显很高，剩余月份用电量降低很快；商业、住宿和餐饮业也具有比较强的周期波动用电特征，其主要表现为每年的1~3月和7~9月用电量明显升高，剩余月份用电量持续下降。

参考文献

- [1] 叶子菀. 高耗能工业用电行为分析与治理研究, [D]. 武汉: 武汉大学, 2013.
- [2] 何永秀, 王冰, 熊威, 等. 基于模糊综合评价的居民智能用电行为分析与互动机制设计[J]. 电网技术, 2012, 36(10): 248-252.
- [3] 张素香, 刘建明, 赵丙镇, 等. 基于云计算的居民用电行为分析模型研究, [J]. 电网技术, 2013, 37(6): 1542-1546.
- [4] 崔丹丹. K-Means聚类算法的研究与改进[D]. 合肥: 安徽大学, 2012.
- [5] 冯超, K-means聚类算法的研究[D]. 大连: 大连理工大学, 2007.
- [6] 雷小锋, 谢昆青, 林帆, 等. 一种基于K-Means局部最优性的高效聚类算法[J]. 软件学报, 2008, 19(7): 1683-1692.
- [7] 杨善林, 李永森, 胡笑旋, 等. K-Means算法中的K值优化问题研究[J]. 系统工程理论与实践, 2006(2): 97-101.
- [8] 汪中, 刘贵全, 陈恩红. 一种优化初始中心点的K-means算法[J]. 模式识别与人工智能, 2009, 22(2): 299-304.
- [9] 邢留伟. K-Means算法在客户细分中的应用研究[D]. 成都: 西南财经大学, 2007.

(收稿日期: 2015-05-26) EA

作者简介

周小明(1978-), 男, 博士, 高级工程师, 主要研究方向为运营监测模型、信息化技术研究。

苏安龙(1970-), 男, 硕士研究生, 高级工程师, 主要研究方向为运营监测模型、信息化技术研究。