

# 基于目标聚类与偏最小二乘法的电力系统日最大负荷预测

朱金鑫, 田亚生

(扬州供电公司, 江苏 扬州 225009)

**摘 要:** 提出了采用目标聚类与偏最小二乘法的负荷预测方法, 采用模糊 c-均值聚类对负荷采样数据进行预处理, 考虑气温、天气、节假日等影响因素, 采用偏最小二乘法建立负荷预测模型对扬州地区夏季日最大负荷进行预测; 预测结果表明该模型具有较高的拟合与预测精度, 对于制定迎峰度夏期间地区电网日调度方案有一定意义。

**关键词:** 日最大负荷预测; 迎峰度夏; 模糊 c-均值聚类; 偏最小二乘回归

## 0 引言

电力系统的稳定运行要求发电量能够随时紧跟系统负荷的变化, 即电厂发出的电能必须能够平衡用户负荷。准确预测负荷, 对确定日运行方式有重要作用, 有助于确定机组组合方案, 电力系统调度方案。特别在迎峰度夏期间, 准确预测次日最大负荷对于下达用电限额、提前做好限电准备、确保有序用电、电网设备安全稳定运行有着重要的意义。

本文着重讨论运用目标聚类与偏最小二乘法建模对夏季日最大负荷进行预测。

## 1 日负荷数据的预处理

电力系统负荷数据繁多, 如果不经预处理直接建立负荷预测模型, 存在输入数据复杂、预测结果误差较大等问题, 因此考虑将电网的负荷数据进行分类后再建立负荷预测模型。

聚类分析是多元分析的一种, 也是非监督模式识别的一个重要分支。它把一个没有类别标记的样本集按某种准则划分为若干个子集, 使相似的样本尽可能归为一类, 而不相似的样本尽量划分到不同的类中。

### 1.1 聚类分析的数学模型

设  $X = \{X_1, X_2, \dots, X_n\}$  是待聚类分析对象的全体,  $X$  中的每个对象  $X_k (k = 1, 2, \dots, n)$  常用有限个参数值来刻画, 每个参数值刻画  $X_k$  的某个特征。于是对象  $X_k$  就伴随着一个向量  $p_{(X_k)} = (X_{k1}, X_{k2}, \dots, X_{ks})$  其中  $X_{kj} (j = 1, 2, \dots, s)$  是  $X_k$  在第  $j$  个特征值上的

赋值,  $p_{(X_k)}$  称为  $X_k$  的特征向量或者模式向量。聚类分析就是分析论域  $X$  中的  $n$  个样本所对应的模式矢量间的相似性, 并且按照各样本间的亲疏关系把  $X_1, X_2, \dots, X_n$  划分成多个不相交的子集。

在模糊划分中, 样本集  $X$  被划分成为  $c$  个模糊子集  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ , 而且样本的隶属函数从 0, 1 二值扩展到  $[0, 1]$  区间, 满足条件:

$$E_f = \mu_{ik} \quad (1)$$

$$\mu_{ik} = [0, 1]; \sum_{i=1}^c \mu_{ik} = 1, \forall k \quad (2)$$

$$0 < \sum_{i=1}^n \mu_{ik} < n, \forall i \quad (3)$$

### 1.2 模糊 c-均值聚类

初始化: 给定聚类类别数  $c$ ,  $2 \leq c \leq n$ ,  $n$  是数据个数, 设定迭代阈值  $\varepsilon$ , 初始化聚类原型模式  $P^0$ , 设定迭代计数器  $b = 0$ ;

步骤一: 计算或更新划分矩阵  $U^b$

对于  $\forall i, k$  如果  $\exists d_{ik}^{(b)} > 0$ , 则有:

$$\mu_{ik}^{(b)} = 1 / \sum_{i=1}^c [d_{ik} / d_{jk}]^{2/m-1} \quad (5)$$

如果  $\exists i, r$ , 使得  $d_{ir}^{(b)} = 0$ , 则有:

$$\mu_{ik}^{(b)} = 1, \text{ 且对 } j \neq r, \mu_{ik}^{(b)} = 0。$$

步骤二: 更新聚类原型模式矩阵  $P^{(b+1)}$ :

$$P_i^{(b+1)} = \frac{\sum_{k=1}^n (\mu_{ik}^{(b+1)})^m}{\sum_{k=1}^n (\mu_{ik}^{(b)})^m}, i = 1, 2, \dots, c \quad (6)$$

步骤三: 如果  $|P^b - P^{(b+1)}| < \varepsilon$ , 则算法停止并输

入划分矩阵  $U$  和聚类原型  $P$ , 否则令  $b = b + 1$ , 转向步骤一。

### 1.3 电网负荷数据聚类仿真

取扬州电网 2011 年 8 月 1 日负荷数据, 对其进行模糊 c-均值聚类, 得到聚类中心如图 1 所示。

图 1 中日负荷曲线经过聚类分析后得到三个聚类中心, 分别为峰期负荷中心、谷期负荷中心与峰谷期之间负荷中心; 聚类中心数据携带了日负荷基本信息, 同时也解决了负荷预测模型输入数据冗余问题。

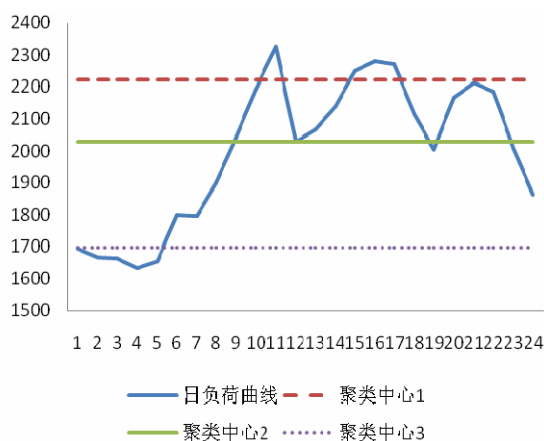


图 1 聚类分析日负荷曲线

## 2 偏最小二乘法的原理和应用

长期以来, 模型式的方法和认识性的方法之间的界限分得十分清楚。而偏最小二乘法则把它们有机的结合, 在一个算法下, 可以同时实现多元线性回归建模、主成份分析以及两组变量之间的相关性分析, 它是一种多因变量对多自变量的回归建模方

法, 可以较好的解决许多以往用普通多元回归无法解决的问题。

### 2.1 偏最小二乘回归的线性模型

设有  $q$  个因变量  $\{y_1, y_2, \dots, y_n\}$  和  $p$  个自变量

$\{x_1, x_2, \dots, x_p\}$ 。为了研究因变量与自变量的统计关系, 观测了  $n$  个样本点, 由此构成了自变量与因变量的数据  $X = (x_1, x_2, \dots, x_p)_{n \times p}$  和  $Y = (y_1, y_2, \dots, y_q)_{n \times q}$ 。

偏最小二乘回归分别在  $X$  与  $Y$  中提取出成分  $t_1$  和

$u_1$  (也就是说  $t_1$  是  $x_1, x_2, \dots, x_p$  的线性组合,  $u_1$  是

$y_1, y_2, \dots, y_q$  的线性组合)。在提取这两个成分时,

为了回归分析的需要, 有下面两个要求:

a)  $t_1$ 、 $u_1$  应尽可能多地携带它们各自数据表中的

变异信息;

b)  $t_1$ 、 $u_1$  的相关程度能够达到最大。

这两个要求表明,  $t_1$ 、 $u_1$  应尽可能好的代表数据表

$X$  与  $Y$ , 同时, 自变量的成分  $t_1$  对因变量的成分  $u_1$

又有很强的解释能力。

在第一个成分  $t_1$ 、 $u_1$  被提取后, 偏最小二乘分

别实施  $X$  对  $t_1$  的回归以及  $Y$  对  $t_1$  的回归。如果回归

方程已经达到满意的精度, 则算法终止; 否则, 将

利用  $X$  被  $t_1$  解释后的残余信息以及  $Y$  被  $t_1$  解释后

的残余信息进行第 2 轮的成分提取。如此反复, 直到能达到一个较满意的精度为止。

### 2.2 偏最小二乘回归的计算方法

首先将数据做标准化处理。  $X$  经标准化处理后的数据矩阵记为:

$$E_0 = (E_{01}, E_{02}, \dots, E_{0p})_{n \times p} \quad (7)$$

$Y$  经标准化处理后的数据矩阵记为:

$$F_0 = (F_{01}, F_{02}, \dots, F_{0p})_{n \times q} \quad (8)$$

记  $t_1$  是  $E_0$  的第一个成分,  $t_1 = E_0 \omega_1$ ,  $\omega_1$  是  $E_0$  的第一个轴, 它是一个单位向量, 即  $\|\omega_1\| = 1$ 。

记  $u_1$  是  $F_0$  的第一个成分,  $u_1 = F_0 c_1$ ,  $c_1$  是  $F_0$  的第一个轴, 并且  $\|c_1\| = 1$ 。

正规的数学表述是求解下列最优化问题

$$\text{Max} = \langle E_0 \omega_1, F_0 c_1 \rangle \quad (9)$$

$$s.t. = \begin{cases} \omega_1^T \omega_1 = 1 \\ c_1^T c_1 = 1 \end{cases} \quad (10)$$

采用拉格朗日算法, 记

$$s = \omega_1^T E_0^T F_0 c_1 - \lambda_1 (\omega_1^T \omega_1 - 1) - \lambda_2 (c_1^T c_1 - 1) \quad (11)$$

对  $s$  分别关于  $\omega_1, c_1, \lambda_1, \lambda_2$  的偏导, 并令为 0,

可以推出:

$$2\lambda_1 = 2\lambda_2 = \omega_1^T E_0^T F_0 c_1 = \langle E_0 \omega_1, F_0 c_1 \rangle \quad (12)$$

记  $\theta_1 = 2\lambda_1$ ,  $\theta_1$  正是优化问题的目标函数值。

由式 (11) 和式 (12) 可得:

$$E_0^T F_0 F_0^T E_0 \omega_1 = \theta_1^2 \omega_1 \quad (13)$$

$$F_0^T E_0 E_0^T F_0 c_1 = \theta_1^2 c_1 \quad (14)$$

求得  $\omega_1$  和  $c_1$  后, 既可得到成分  $t_1 = E_0 \omega_1$ ,

$u_1 = F_0 c_1$ 。

然后分别求  $E_0$  和  $F_0$  对  $t_1, u_1$  的 3 个回归方程:

$$E_0 = t_1 p_1^T + E_1 \quad (15)$$

$$F_0 = u_1 q_1^T + F_1 \quad (16)$$

$$F_0 = t_1 r_1^T + F_1 \quad (17)$$

用残差矩阵  $E_1, F_1$  取代  $E_0, F_0$ , 然后, 求第 2

个轴的  $\omega_2, c_2$  以及第 2 个成分  $t_2, u_2$ , 得出回归系数向量  $p_2, r_2$ , 如此迭代下去, 如果  $X$  的秩是  $A$ , 则:

$$E_0 = t_1 p_1^T + t_2 p_2^T + \dots + t_A p_A^T \quad (17)$$

$$F_0 = t_1 r_1^T + t_2 r_2^T + \dots + t_A r_A^T + F_A \quad (18)$$

## 2.3 交叉有效性判别

在许多情况下, 偏最小二乘回归方程不需要选用所有的成分  $t_1, t_2, \dots, t_A$  进行回归建模, 而是采用截尾的方式选择前  $m$  个成分, 就可以得到一个预测性能较好的模型。

究竟应该选取多少个成分为宜, 可通过考察增加 1 个新的成分后, 能否对模型的预测功能有明显的改进来考虑。

## 3 地区日最大负荷预测模型建立与仿真

### 3.1 目标聚类偏最小二乘回归模型的建立

定义模型自变量为: 日负荷聚类中心数据、每日最高气温、平均气温、最低气温、降雨量、每日积分电量与当日最大负荷; 定义模型因变量为次日最大负荷。考虑节假日信息, 将工作日与节假日分开建立负荷预测模型。见图 2。

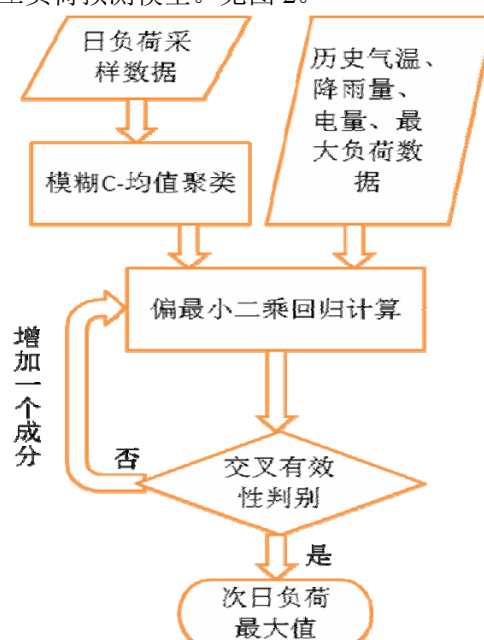


图2 负荷预测模型流程图

3.2 预测结果与误差分析

将 2011 年 7 月 1 日~8 月 30 日的数据进行偏最小二乘回归建模预测次日最大负荷, 得出预测结果如表 1 所示, 表 1 中负荷单位为 MW。

表 1 7 月 2 日~8 月 31 日最大日负荷预计表

7 月	实际	预测	误差%	8 月	实际	预测	误差%
2	2500	2529	1.16	1	2331	2288	1.84
3	2471	2379	3.72	2	2489	2364	5.02
4	2310	2409	4.28	3	2409	2499	3.73
5	2061	2058	0.14	4	2378	2343	1.47
6	2303	2303	0.00	5	2447	2443	0.16
7	2518	2344	6.91	6	2347	2424	3.28
8	2263	2318	2.43	7	2152	2203	2.36
9	2367	2331	1.52	8	2475	2462	0.52
10	2154	2188	1.57	9	2503	2393	4.39
11	2377	2392	0.63	10	2496	2542	1.84
12	2194	2173	0.95	11	2403	2393	0.41
13	2206	2229	1.04	12	2459	2427	1.30
14	2080	2076	0.19	13	2482	2491	0.36
15	2078	2090	0.58	14	2071	1992	3.81
16	2001	1973	1.39	15	2857	2753	3.64
17	1903	1897	0.32	16	2635	2627	0.30
18	2172	2172	0.00	17	2456	2634	7.24
19	2168	2273	4.84	18	2784	2574	7.51
20	2169	2208	1.79	19	2665	2578	3.26
21	2322	2340	0.77	20	2291	2411	5.23
22	2532	2546	0.55	21	2003	1967	1.79
23	2594	2556	1.46	22	2077	2031	2.21
24	2543	2583	1.57	23	2114	2089	1.18
25	2616	2557	2.25	24	2045	2122	3.76
26	2320	2397	3.31	25	2052	2163	5.40
27	2430	2396	1.39	26	2119	2074	2.12
28	2419	2373	1.91	27	2089	2188	4.73
29	2521	2580	2.34	28	1960	2186	11.5
30	2633	2585	1.82	29	2234	2196	1.70
31	2284	2293	0.39	30	2170	2174	0.18
				31	2168	2088	3.69

采用平均绝对百分误差 (MAPE) 和均方根误差 (RMSE) 作为模型的拟合程度与预测精度的指标。

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \tilde{y}_i}{y_i} \right| \times 100 \quad (19)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \frac{y_i - \tilde{y}_i}{y_i} \right)^2} \quad (20)$$

式 (19) 与 (20) 中,  $\tilde{y}_i$  为  $y_i$  的预测值, 根据式 (19) 与 (20) 得出  $MAPE=2.41$ ,  $RMSE=0.032$ , 模型预测精度较高, 预测值与实际值曲线如图 3 所示。

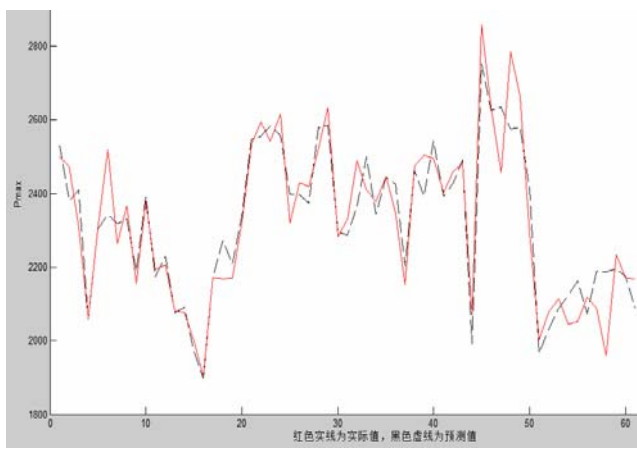


图 3 日最大负荷预测值与实际值曲线

4 结论

电力系统是一个周期性和随机性都很强的系统, 与众多因素有极为复杂的关系。在电力系统负荷预测时, 既要充分分析、掌握并利用其规律性, 又要兼顾各种因素的影响。

考虑迎峰度夏期间日负荷预测对于电网运行的重要性以及扬州电网夏季负荷特性, 本文对 2011 年 7、8 月份扬州地区日最大负荷进行预测。采用模糊 c-均值聚类方法对日负荷数据进行预处理, 对日负荷数据进行分类, 降低了输入数据的冗余度; 考虑影响日负荷的气温、天气、节假日因素, 运用偏最小二乘回归方法建立负荷预测模型, 模型拟合程度良好、具有较高预测精度, 验证了预测模型的合理性。

由于负荷采样数据本身存在一定误差, 而电力系统日最大负荷又受一些随机发生的事件影响, 在考虑众多的影响因素后, 仍存在极少部分预测结果误差偏大的问题; 因此, 剔除与修正负荷数据特异点、降低随机事件对预测模型的影响仍是今后需要进一步研究的问题。

参考文献:

[1] 王文圣, 丁晶, 赵玉龙, 等. 偏最小二乘回归的年用预测研究[J]. 中国电机工程学报, 2003, 23 (10): 17-21.  
[2] 刘耀年, 王卫, 杨冬峰. 基于模糊划分聚类中的中长期用电量预测[J]. 东北电力学院学报, 2004, 24 (4): 39-42.  
[3] 肖春景, 张敏. 基于减法聚类与模糊 c-均值的模糊聚类研究[J]. 计算机工程, 2005(31) (增刊): 135-137.  
[4] 诸克军, 苏顺华, 黎金玲. 模糊 C-均值中最优聚类与最佳聚类数[J]. 系统工程理论与实践, 2005(3): 52-61.

[5] 高新波.模糊聚类分析及其应用[M].西安:西安电子科技大学出版社, 2004.

朱金鑫(1987—),男,江苏扬州人,助理工程师,从事电网调度员工作;

田亚生(1957—),男,江苏扬州人,技师,从事电网调度员工作。

---

作者简介: