基于因果聚类分析理论的维修成本控制体系研究

麻兴斌[1]，孟祥君[2]，刁柏青[2]

（1. 山东科技大学 财经系，济南 250031；2. 国家电网山东电力公司，济南 250001）

摘要：供电网络的维修维护工程比较复杂，影响因素较多，影响方式多种多样，这给成本控制造成了很大困难。为了解决这个问题，采用数据分析，建立预测模型，通过分析预测误差和预测的置信区间，确定维护成本的控制范围。为了提高预测得准确性，利用历史数据对要素变量进行因果聚类。然后将各类型中的变量进行主成份分析，降低变量集的维数。对选出的变量用神经网络算法构建预测模型，并利用其误差的性质确定误差范围，实现对成本的控制。

关键词：成本控制；供电网络；因果聚类；神经网络；置信区间

中图分类号：TM715 **文献标志码**：A **文章编号**：1671－1807(2018)06－0096－06

电网系统是输电系统的主要设备，它的维修业务占电力企业生产任务的70%以上[1]，在电力行业逐步走向市场化的情况下，维修、检修的业务模式也呈现出多样化的现象。这样就相对弱化了在微观层次上对成本构成成分的控制，凸显出维护成本整体控制的重要性。长期以来，电力经营的客观垄断性和环境的复杂性，国内外都没有对其维修费用有一个成熟的成本控制方法。主要的控制方法是标准成本法和预测控制法[2]，常用的预测控制法有多元回归法、人工神经网络法等[3]。本文用区间预测方法，对电网复杂的设备故障维修费进行控制，结合运营中心数据的挖掘实现实时控制的目的。

1 成本分析与变量的确定

1.1 分析成本构成要素

电力企业的线路、设备体系复杂，需要按不同的电压等级、线路/设备、作业等分成不同的等次。在不同的电压等级中，同样的设备维护成本是不同的；不同的设备要求不同，其作业成本也就不同。作业成本由以下四个方面构成：

1）外委成本：从外委成本总额度中区分出材料费、人工费、机械台班费，并以分类总额的方式通过对工单成本进行计划和确认。

2）材料成本：包括装置性材料与消耗性材料。根据要求的管理细度对材料合理归类，并按此类别在系统中建立活动类型，维护定额单价，并对不同的标准作业分别维护其定额数量。装置性材料的成本等于对工单发货时产生的成本；消耗性材料的成本，一部分来源于对工单发货时产生的成本，另一部分来源于对“零购固定资产及卡片式低值易耗品”采购流程所购物料的领用。

3）人工成本：以正式、外聘员工标准定额时薪做为统计的类别来管理人工成本。在系统中建立活动类型，分别维护其定额单价，并对不同的标准作业分别维护正式、外聘员工的工时。业务部门依据实际发生对标准作业的两类工时分别确认。

4）机械台班成本：对不同的机械台班按规范标准归类，按相应类别管理机械台班工时。在系统中建立活动类型，分别维护其定额单价，并对不同的标准作业分别维护机械台班的工作量。业务人员依据实际发生的机械台班工作量分别确认。

由此可见，总的维修成本 M_i 包括两部分，一部分是由装置性材料与消耗性材料等构成，不受天气、地形等因素影响，用 C_i 表示；另一部分受到环境和装置体积等因素的影响，即 $Y_i=\frac{M_i}{C_i}$。

利用要素核算的方式进行成本控制，很难在这些复杂的要素及其变化的规律中预测维修的总成本以及它们的变化范围。电网的成本控制是电网系统预测控制的一个侧面，属于过程控制的类型，虽然没有快速反应的要求，但是具有大系统的复杂性、非线性

收稿日期：2017－05－15

基金项目：国家自然科学基金项目(71071089)。

作者简介：麻兴斌(1965—)，男(回族)，山东长清人，山东科技大学财经系，副教授，管理科学与工程博士，研究方向：复杂系统与组织科学。

的特点[4]、低成本的要求。目前对于这种系统的控制无法确定因变量的变化区间，只能通过多元线性模型[5-6]，依概率确定总成本的变化范围。因此，我们采用因果聚类的方法将数据划分为若干类，每一类进行主成分分析，选择有效主成分作为模型的变量，进行预测。进行控制时，将收集到的相关因素的状态值归入一个最贴近的类别，从这个类别的历史数据中归纳出的特征值进行控制。

1.2 选择变量并建立数据矩阵

建立成本控制体系的目的不但是逐渐降低线路或设备的维修中产生的成本费用，同时也为业务外包的价格提供基础依据。因此在选择变量时，将材料费、人工费、机械台班费等各类直接产生的费用作为检测的变量，将气象因素、人员构成因素、维修材料的重量和体积等作为影响变量。

设有 n 个变量 $X_1, X_2, \cdots, X_n$，有 m 组观察数据，构建 $m \times n$ 矩阵

$$X=(X_1, X_2, \cdots, X_n)=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mm} \end{bmatrix} \quad (1)$$

式中，x_{ij} 是第 j 个变量的第 i 个观察值。

表 1 维护工程指标体系

序号	准则层	指标层	符号
1	因变量	总成本	x_0
2	过程因素	电压级别	x_1
3		维修级别	x_2
4		维修时间	x_3
5	运输因素	维修地点的距离	x_4
6		到达线路的难易程度	x_5
7		所需车辆的级别	x_6
8		设备类型	x_7
9	气象因素	空气温度	x_8
10		空气湿度	x_9
11	现场因素	人员构成	x_{10}
12		固定材料	x_{11}
13		更换材料长度	x_{12}
14		更换材料重量	x_{13}

表 2 维修费用及影响参数

	x_0(元)	x_1(kV)	x_2	x_3(h)	x_4(km)	x_5	x_6	x_7	x_8(级)	x_9(级)	x_{10}(元)	x_{11}	x_{12}(Kg)	x_{13}(m)
1	70 196	35	3	38	77.9	3	2	10	3	4	28 000	4.01	1 122	1.5
2	64 122	220	1	22	56.3	5	1	5	4	4	17 600	2.96	0	2.9
3	70 666	35	5	46	28.2	5	4	9	4	3	30 100	4.64	5 239	5.0
4	75 368	35	5	63	54.5	4	4	9	3	4	28 300	5.95	4 713	8.2
5	72 216	10	2	77	68.3	3	1	6	5	5	29 000	7.93	0	3.0
6	80 464	220	2	84	77.1	7	3	4	6	3	30 800	7.09	0	3.5
7	91 210	500	3	86	58.4	5	4	8	3	5	39 600	6.51	1 769	4.3
8	57 000	10	4	81	69.7	3	4	5	4	4	16 500	3.04	4 889	3.3
9	82 750	10	4	79	61.9	2	7	12	5	2	31 500	5.65	4 970	5.4
10	97 743	220	5	85	8.5	7	7	3	7	5	48 300	7.77	0	2.3
11	110 500	500	3	86	30.2	7	2	9	2	6	47 600	6.75	0	2.2
12	113 768	220	4	81	15.1	6	2	10	1	4	49 100	6.69	4 761	5.1
13	106 580	10	1	79	8.0	5	1	6	4	4	38 700	3.70	0	6.6
14	95 277	35	5	91	3.0	7	7	6	5	6	28 100	9.80	4 150	6.4

2 对要素矩阵进行因果聚类

在变量维数很大的体系中，进行多变量分析，虽然能获得更多的信息，但是可能会降低有效变量的灵敏度，还可能带来噪声干扰、复共线性。在分析中进行降维处理，聚类分析就是常用的一种降维方法[7]。聚类分析是根据变量之间的相似程度或定义距离进

行分类。常用的是根据变量的特征值的相思程度进行分类，也可以根据需要确定变量的相似特征。本文是根据变量与预测变量的相关性确定变量间的相似程度。

2.1　构建相似程度变量

相似程度的定义决定了聚类所根据的变量的性质[8]，这里是研究与预测变量的相关程度和性质，因此首先考虑的是与预测变量的相关性进行聚类。再者变量间的相似性也是重要的聚类特征，因此将变量间的距离也考虑在内，构成判断标准。

首先计算变量间的相似度距离，用 d_{1ij} 表示。

$$d_{1ij}=\left|\sqrt{\sum_{k=1}^{m}(x_{ki}-x_{kl})^2}-\sqrt{\sum_{k=1}^{m}(x_{kj}-x_{kl})^2}\right|$$

其中，x_{kl} 是预测对象，即 Y_i。

其次计算变量间的相关性相似系数，同样分别计算预测变量与预测对象之间的相关性，然后再描述变量间的相似距离与相似系数的共同影响，以此作为聚类分析的相似程度。这里先不考虑变量是否服从正态分布，因此可以用 Spearman's Rank 相关系数计算变量间的相关性。

令 $r_{il}=1-\dfrac{6\sum_{i=1}^{N}D_i^2}{N(N^2-1)}$；$r_{jl}=1-\dfrac{6\sum_{j=1}^{N}D_j^2}{N(N^2-1)}$

式中，$D_i=(x_{ki}-x_{kl})$，$D_j=(x_{kj}-x_{kl})$。

得到相关性相似系数 $d_{2ij}=|r_{il}-r_{jl}|$　　(2)

其中，r_{il}、r_{jl} 分别表示变量 X_i、X_j 同预测变量之间的相关系数；N 为样本数。

构建相似程度变量：

$$d_{ij}=d_{1ij}+\omega d_{2ij}　　(3)$$

根据表 2 中的数据构建相似矩阵 R。

$$R=\begin{bmatrix}
1 & 0.86 & 0.69 & 0.78 & 0.46 & 0.42 & 0.76 & 0.24 & 0.17 & 0.30 & 0.13 & 0.14 & 0.42 & 0.33\\
 & 1 & 0.79 & 0.54 & 0.38 & 0.43 & 0.82 & 0.62 & 0.28 & 0.23 & 0.35 & 0.37 & 0.48 & 0.62\\
 & & 1 & 0.35 & 0.54 & 0.52 & 0.48 & 0.58 & 0.43 & 0.36 & 0.28 & 0.71 & 0.42 & 0.35\\
 & & & 1 & 0.46 & 0.56 & 0.13 & 0.03 & 0.43 & 0.13 & 0.01 & 0.56 & 0.14 & 0.31\\
 & & & & 1 & 0.62 & 0.35 & 0.79 & 0.65 & 0.43 & 0.61 & 0.21 & 0.45 & 0.64\\
 & & & & & 1 & 0.78 & 0.51 & 0.72 & 0.31 & 0.38 & 0.35 & 0.46 & 0.73\\
 & & & & & & 1 & 0.63 & 0.65 & 0.36 & 0.14 & 0.25 & 0.56 & 0.79\\
 & & & & & & & 1 & 0.61 & 0.56 & 0.52 & 0.34 & 0.12 & 0.01\\
 & & & & & & & & 1 & 0.52 & 0.48 & 0.29 & 0.61 & 0.11\\
 & & & & & & & & & 1 & 0.64 & 0.41 & 0.32 & 0.19\\
 & & & & & & & & & & 1 & 0.56 & 0.08 & 0.61\\
 & & & & & & & & & & & 1 & 0.81 & 0.44\\
 & & & & & & & & & & & & 1 & 0.23\\
 & & & & & & & & & & & & & 1
\end{bmatrix}$$

2.2　因果聚类算法

具体聚类方法采用 K-neans 聚类算法。首先确定 k 为聚类分析的簇数，式(3)为相似程度确定变量之间的距离。这样设定的初始凝聚点的集合为：

$$L^{(0)}=(x_1^{(0)},x_2^{(0)},\cdots,x_k^{(0)})$$

据此初始分类为：

$$G^{(0)}=\{G_1^{(0)},G_2^{(0)},\cdots,G_k^{(0)}\}$$

计算初始分类集 $G^{(0)}$ 中每个子集 $G_i^{(0)}$ 的重心，并以此作为下一轮循环的凝聚点集，即

$$x_i^{(1)}=\frac{1}{n_i}\sum_{x_i\in G_i^{(0)}}x_i,i=1,2,\cdots,k$$

式中，n_i 为子集 $G_i^{(0)}$ 的变量数目。得到第二轮循环的凝聚点集：

$$L^{(1)}=\{x_1^{(1)},x_2^{(1)},\cdots,x_k^{(1)}\}$$

同样以式(3)作为相似程度，得到第二轮的分类集合：

$$G^{(1)}=\{G_1^{(1)},G_2^{(1)},\cdots,G_k^{(1)}\}。$$

以此类推重复下去。

当 m 逐渐增大时，子集中的变量趋于稳定。判断式为各子集 $G_i^{(m)}$ 的重心趋于稳定，即 $x_i^{(m-1)}\approx x_i^{(m)}$，$G_i^{(m-1)}\approx G_i^{(m)}$，算法结束，得到最终分类集

$$G^{(m)}=\{G_1^{(m)},G_2^{(m)},\cdots,G_k^{(m)}\}　　(4)$$

计算结果如下：

$$R^{*}=\begin{bmatrix}1 & 0.76 & 0.76 & 0.56 & 0.76 & 0.46 & 0.42 & 0.76 & 0.57 & 0.21 & 0.08 & 0.23 & 0.01 & 0.01\\ & 1 & 0.76 & 0.56 & 0.32 & 0.46 & 0.75 & 0.50 & 0.43 & 0.43 & 0.16 & 0.23 & 0.24 & 0.32\\ & & 1 & 0.35 & 0.32 & 0.43 & 0.31 & 0.51 & 0.51 & 0.48 & 0.43 & 0.73 & 0.44 & 0.32\\ & & & 1 & 0.40 & 0.40 & 0.41 & 0.52 & 0.21 & 0.45 & 0.38 & 0.47 & 0.21 & 0.31\\ & & & & 1 & 0.43 & 0.40 & 0.72 & 0.59 & 0.34 & 0.45 & 0.37 & 0.37 & 0.31\\ & & & & & 1 & 0.71 & 0.43 & 0.72 & 0.35 & 0.35 & 0.36 & 0.28 & 0.23\\ & & & & & & 1 & 0.39 & 0.55 & 0.32 & 0.13 & 0.34 & 0.34 & 0.73\\ & & & & & & & 1 & 0.53 & 0.45 & 0.34 & 0.34 & 0.41 & 0.12\\ & & & & & & & & 1 & 0.54 & 0.35 & 0.32 & 0.38 & 0.46\\ & & & & & & & & & 1 & 0.53 & 0.53 & 0.50 & 0.73\\ & & & & & & & & & & 1 & 0.34 & 0.12 & 0.54\\ & & & & & & & & & & & 1 & 0.72 & 0.57\\ & & & & & & & & & & & & 1 & 0.53\\ & & & & & & & & & & & & & 1\end{bmatrix}$$

对原始矩阵进行聚类分析，令类别 $k=3$ 得出分类矩阵及相关性。

表 3　对原始数据的聚类分析

类别	变量	数据	相关系数
类别Ⅰ	x_1 … x_{10}	3,2,5,…,3 … 23,15,20,…,6	0.78
类别Ⅱ	x_8 … x_{16}	1,3,3,…,1 … 4,5,3,…,7	−0.36
类别Ⅲ	x_7 … x_{14}	7,10,8,…,10 … 8.7,9.5,…,11.3	0.45

注：气象、类别等数据采用分级制，如风力采用气象分级，但是 8 级以上则不利于高空作业，则属于灾害天气，而在一些室内变压器、继电器的维修中则属于正常天气。

3　构建预测模型

3.1　提取各类别中的主成分

影响维修成本的因素变量较多，因此对他们进行聚类分析是一种降维方法，用因素的类别代替众多的因素变量，增加变量的影响显著程度。但是由于各类别中变量数目是不一样的，因此类别作为一个变量对成本的影响也不像一个变量那样的敏感程度。这样，用主成分分析法，提取主要的几个主成分作为变量加入模型中。

对选取的 8 个样本点的数据每个类别进行主成份分析，并估算主成分的方差贡献率。

经过分析，类别Ⅰ中选取 c_1、c_2、c_3 作为变量组成 C_1。

表 4　类别Ⅰ的主成份分析

类别	因子	特征值	方差贡献率%	累计方差贡献率%	备注
类别Ⅰ	c_1	4.51	45.86	45.86	*
	c_2	2.14	24.56	70.42	*
	c_3	1.25	17.17	87.59	*
	c_4	0.60	5.13	92.72	
	c_5	0.31	3.45	96.17	
	c_6	0.08	2.31	96.48	
	c_7	0.02	0.98	99.46	
	c_8	0.01	0.54	100	

同理，将类别Ⅱ、类别Ⅲ进行主成份分析，选取累计方差超过 85%的主成分组成新的变量集 $C=(C_1, C_2, C_3)$，共 8 个主成分变量，分别是从类别Ⅰ中提取 3 个主成分，类别Ⅱ中提取 2 个主成分、类别Ⅲ中提取 3 个主成分。

3.2　构建预测模型

经过对变量的聚类分析和主成分的提取，降低了维数并降低了多重共线性[9]。通过聚类分析，类别和类别之间的距离较大。在每个类别中进行主成份分析，提取出的主成分是相互独立的，但是类别之间的变量虽然相关性变弱，但是仍存在着相关性，因此预测模型还是采用对变量的性质要求较低的神经网络法建立预测模型[10]。

BP 算法在神经网络算法中应用较多，在本方案中采用 BP 算法。将聚类分析和主成份分析提取的 8 个变量，作为神经网络的输入层节点，选择隐含层节点 6 个[11]，输出层就是因变量一个节点。构建预测模型：

$$\hat{Y}_i = f(x_1(i), x_2(i), \cdots, x_8(i)) \quad (5)$$

4　确定成本控制区间

对于理想的预测模型，则

$Y_i = \hat{Y}_i + v(i)$

Y_i 为实测值，$\hat{Y}_i$ 为预测值，$v(i)$为误差。

若 $\hat{Y}_i$ 为理想输出，则 $v(i)$为均值为零的高斯白噪声，即 $E\{v(i)\}=0$。当样本足够大时，观察值呈正态分布，相应预测误差值也呈正态分布，因此可以估计预测误差的分布特征。以最近的 50 个样本做预测误差计算。利用预测模型(5)进行预测，假设预测值为 $\hat{Y}(i)$，与相应的观察值 $Y(i)$进行比较，计算误差均方根

$$\sigma_E = \left(\frac{1}{N-1}\sum_{i-1}^{N}(\hat{Y}(i)-Y(i))^2\right)^{\frac{1}{2}} \quad (6)$$

其中，N 是观察样本的个数，$\hat{Y}(i)$为预测值，$Y(i)$为观测值。

误差均方根能够反映预测值偏离实际观测值的程度，其值一般大于等于零，可以用来观察模型的预测效果。本文用来进行预测值的区间估计，以确定维修成本的范围，进行控制。

本文中，计算给出的一系列样本，得出 $\sigma_E = 0.021$。取置信度 95%，概率度 K＝1.96，则置信区间为$[Y(i)-1.96\sigma_E, Y(i)+1.96\sigma_E]$。成本控制区间$[Y(i)-1.96\sigma_E, Y(i)+1.96\sigma_E]$，即$[Y(i)-0.041, Y(i)+0.041]$。

在上述的维修工程系列中，某工程所消耗的固定费用，包括材料费和设备费用等共计 20.3 万元。将其它变量代入预测模型，得 $Y(i)=2.37$。则总费用预测值为 $M_i = Y_i C_i = Y(i)C_i = 48.11$。成本控制区间为[47.27，48.95]万元。当实际费用超过这个区间时，说明施工过程产生了不合理的费用，或者出现意外的情况，需要进行分析说明。

5　结论

像输电线路的维护维修这样的工程任务，由于影响因素复杂多变，其成本的预算难度很大。在本文中，利用因果聚类的方法，对历史数据进行分析和降维，然后又用神经网络算法构建预测模型，在此基础上以预测的误差范围作为成本检测的控制范围。这种方法的特点是：(1)避免分析电网维护施工过程中各因素的具体影响方式，而是采用统计分析的方法描述各种影响因素与维护成本的关系。(2)采用因果聚类的方法，以因素变量与分析变量之间的相关性的相似程度作为度量，更能符合实际情况。(3)采用神经网络的算法构建模型，能够利用其自我学习的能力实现实施监控。

在本文中，利用本系列的样本得出的是一个固定的区间范围，没有考虑项目成本的大小对控制范围的影响，这是一个需要进一步探讨的问题。

参考文献

[1] 麻兴斌，孟祥君，曹煊，等. 电网运行的可靠性、适应性和经济性研究[M]. 济南：山东大学出版社，2014：48－52.

[2] 陈迎春，宋业新，吴晓平. 基于模糊逻辑的舰船维修经费预测[J]. 海军工程大学学报，2002，14(1)：6－9.

[3] 刘宝平，孙胜祥，徐一帆，等. ANFIS 网络在舰船维修费用预测中的应用[J]. 海军工程大学学报，2004，16(4)：57－60.

[4] CASTILLO O，MELIN P. Hybird intelligent systems for time series prediction using neural networks，fuzzy logic，andfractal theory[J]. IEEE Transaction on Neural Networks，2002，13(6)：1395－1408.

[5] RADAN HUTH，LUCIE POKORN. Multaneous analysis of climatic trends in multiple ariables：an example of application of multivariate statistical methods[J]. International Journal of Climatology，2005，25(4)：469－484.

[6] ARISTITA BUSUIOC，DELIANG CHEN，CECILIA HELLSTR? M. Performance of statistical downscaling models in GCM validation and regional climate change estimates[J]. International Journal of Climatology，2001，21(5)：557－578.

[7] CHRISTIAN HENNIG C. Identifiability of models for clusturwise linear regression[J]. Journal of Classification，2000，17：273－297.

[8] DESARBO WAYNE S，CRON W L. A maximum likelihood methodology for clusterwie linear regression[J]. Journal of Classification，1988，5：249－282.

[9] 朱凯，王正林. 精通 MATLAB 神经网络[M]. 北京：电子工业出版社，2010：132－145.

[10] CHEN S，COWAN C F N，GRANT P M. Orthog onal least squares learning algorithm for radial basis function networks[J]. IEEE Trans on Neural Networks，1991，2(2)：302－309.

[11] 蔡自兴. 神经控制器的典型结构[J]. 控制理论与应用，1998，15(1)：21－24.

Research on the Maintenance Cost Control System of the Power Grid Enterprise

MA Xing-bin[1], MENG Xiang-jun[2], DIAO Bai-qing[2]

(1. Shandong University of Science and Technology, Jinan 250031, China; 2. Shandong Electric Power Group Corporation, Jinan 250001, China)

Abstract: The maintenance of power supply network is more complex, there are so many influence factors that influence the cost of it in various ways, which caused the cost difficult to control. In order to solve this problem, the data analysis is used to establish the method of forecasting model, and the confidence interval of forecast error is made, which determine the range of the maintenance cost control. To improve the accuracy of the forecast, the historical data is used for causal clustering of the factor variables. Then, principal component analysis of the variables is conducted in each type of variable, and the dimension of the variable set is reduced. The neural network algorithm is used to construct the forecasting model and the error range is determined by the error property

Key words: cost control; power supply network; causal clustering; neural network; confidence interval

(上接第 78 页)

The Spatio-temporal Coupling Relationship between Industry Development and Industry Pollution Emission in Henan

NIU Yi-yuan, WU Fan

(Research Institute of Yellow River Civilization and Sustainable Development & Collaborative Innovation Center on Yellow River Civilization of Henan Province, Henan University, Kaifeng Henan 475001, China)

Abstract: The spatial-temporal pattern and coupling relationship between industrial output and pollutant emissions in Henan Province were analyzed using GeoDa spatio-temporal analysis and decoupling index. The factors and contribution of industrial pollution emission in Henan Province were discussed using Kaya equation and LMDI factor decomposition method. The research shows that: ①Various types of industrial pollution are relatively serious in western Henan and northern Henan, and local cities should carry out targeted pollution control and emission reduction; ②Industrial structure optimization and emission intensity reduction can reduce industrial pollutants The increase in the scale of output value and the increase in the proportion of pollutant source industries are positively driving the emission of industrial pollutants.

Key words: industry development; industry pollution emissions; spatio-temporal analysis; decoupling; LMDI factor decomposition