

# 基于 PCA-ANN 的水利工程标高金预测模型

石林林<sup>1,2</sup>

(1. 河海大学商学院, 江苏 南京 211100; 2. 河海大学项目管理研究所, 江苏 南京 211100)

**摘要:**在水利工程投标竞争中, 施工企业投标报价的合理性是赢得该项目的重要前提, 而标高金的准确性直接影响投标报价的合理性。采用基于主成分分析的 BP 神经网络方法, 建立了标高金预测模型。利用主成分分析(PCA)方法, 对标高金的影响因素进行降维, 在保留原始数据信息的条件下, 通过消除样本数据间相关性、减少模型输入变量的数目, 简化网络结构。利用 BP 神经网络(ANN)寻找数据的内在联系和规律, 客观而较为准确地预测出标高金。根据该模型, 以某承包商历年来水利工程投标报价的原始数据为基础, 对标高金进行了预测, 结果表明模型预测结果满足实际需要。

**关键词:**水利工程; 标高金; 主成分; BP 神经网络

**中图分类号:**F284      **文献标识码:**A      **文章编号:**1003-9511(2014)03-0052-04

水利工程项目采购基本上采用招标投标方式, 水利工程领域的竞争达到了空前激烈程度, 施工企业面临着越来越大的竞争压力。水利工程建设投标报价的决策问题是一个非常复杂的过程, 它不仅要考虑企业自身因素, 还要全面考虑招标工程项目相关情况、招投标市场现状、竞争对手历史资料和投标可能性等因素。可见, 投标报价关系到施工企业中中标的可能性以及利润水平。要在激烈的市场竞争中生存下来以及得到进一步发展, 企业必须要提高中标的概率。在水利工程投标竞争中, 施工企业投标报价的合理性是获胜的关键, 它主要取决于成本估算的准确性、其他影响因素以及这些因素和最终报价之间的关系。

水利工程投标报价主要由成本和标高金两部分组成, 成本确定后, 标高金主要由不可预见费和项目的利润构成。目前, 投标报价模型主要有 BP 神经网络<sup>[1]</sup>、模糊自适应粒子群优化<sup>[2]</sup>、支持向量机<sup>[3]</sup>粗糙集与自适应模糊神经网络<sup>[4]</sup>、蒙特卡罗模拟方法<sup>[5-6]</sup>、不完全信息动态博弈分析<sup>[7-8]</sup>、仿真博弈<sup>[9]</sup>等。对比上述方法, BP 神经网络的模拟程度更高, 预测更接近真实值。这些文献选择了主成分 BP 神

经网络的投标报价模型作为预测标高金的工具, 先用主成分分析法(principal components analysis, 缩写 PCA)对影响标高金的多种因素进行数学降维, 减少模型输入变量的数目, 再利用 BP 神经网络(artificial neural network, 缩写 ANN)的自组织、自学习和非线性映射能力来找到大量数据中的内在联系和规律, 客观而较为准确地预测出标高金。相比于其他模型, 基于 PCA-ANN 的水利工程标高金预测结果更加具有信服力, 操作上比较快速、简单、精确, 有助于提高报价的准确性。

## 1 标高金影响因素的识别

在水利工程招投标报价中, 影响标高金的因素很多, 国内外许多学者做了许多研究<sup>[10-11]</sup>, 但目前还没有统一的标准。笔者在阅读大量文献的基础上, 经过实地调查, 结合当前施工企业市场竞争的现状将影响标高金的因素分为 3 类: 项目因素、企业因素、环境因素, 共包括 19 个因素。各承包商可以根据本公司的实际情况选择相应的指标。各影响因素的定义和指标值见表 1。

**基金项目:**“十一五”国家科技支撑计划重大项目(2006BAB04A13); 江苏省教育厅高校哲学社会科学基金(07SJD630006)

**作者简介:**石林林(1989—), 女, 江苏南通人, 硕士研究生, 从事项目管理与工程管理研究。E-mail:827238627@qq.com

表 1 投标报价影响因素的定义和指标值

类别	因素名称	代号	定义	指标值
项目因素	工程性质	$X_1$	工程的具体情况是否在本企业范围	是=100,否=0
	规模	$X_2$	预计的中标合同价	万元
	业主	$X_3$	和投标商关系	好=100,一般=50,差=0
	风险	$X_4$	工程项目不可预测的风险大小	大=100,中=50,小=0
	复杂性	$X_5$	工程的复杂程度是否超出预期	是=100,否=0
	持续时间	$X_6$	期望的项目周期	月
企业因素	目标任务	$X_7$	企业当前项目任务的多少	多=100,一般=50,少=0
	回报率	$X_8$	依据投入多少确定回报率	%
	市场份额	$X_9$	企业当前的实际和预期市场份额比	%
	管理费	$X_{10}$	本年度收回的间接管理费	%
	资金需求量	$X_{11}$	对项目利润的期望值	高=100,中=50,低=0
	盈利情况	$X_{12}$	企业同类项目上获得利润均百分比	%
	索赔经历	$X_{13}$	业主过去承包项目失败状况	多=100,少=50,没有=0
环境因素	市场条件	$X_{14}$	目前是否参与其他项目	多=100,一般=50,少=0
	竞争程度	$X_{15}$	该项目参与竞标其他投标商数量	数量
	未来市场	$X_{16}$	预计会参与的投标项目	多=100,一般=50,少=0
	工人资源	$X_{17}$	当地工人资源是否充足	容易=100,困难=50,没有=0
	工人素质	$X_{18}$	工人技能水平	好=100,一般=50,差=0
	地点	$X_{19}$	项目所在地是否在企业可控的地区内	是=100,否=0

## 2 基于 PCA-ANN 的标高金预测模型

### 2.1 主成分分析

选取标高金的影响因素应尽可能全面地覆盖投标报价的所有信息,使得预测模型结果更准确,但是过多的变量也会造成 BP 神经网络的过度训练,时间延长,反而结果可能不准确。为了提高模型的识别效率,应尽可能选择少的指标来代替原数据从而建立模型。通常可采用粗集理论或灰色关联分析的方法,但这两种方法实质上只是选择部分指标来代替原来信息,因此忽略了部分信息。而主成分分析方法应用了数学降维的方式,生成几个综合变量来代替原有的多个变量,同时又保证这些综合变量可以涵盖原有变量的信息,且相互之间互斥。即在不忽略或很少忽略原有信息量的前提下,把原有的多个相关的参数转换成少数几个互斥的综合参数。这样减少原有的指标数量,既解决了原有数据相关性的问题,又完成了对原始数据的降维工作。其基本算法如下<sup>[12]</sup>:

a. 取  $m$  维随机向量  $X = (x_1, x_2, \dots, x_m)^T$  的  $n$  个样品  $x = (x_{i1}, x_{i2}, \dots, x_{im})^T$ , 则  $X = (x_{ij})_{n \times m}$ 。

b. 标准化原始数据。  $X_{ij} = \frac{x_{ij} - x_i}{\sqrt{\text{var}(x_j)}}$ ,  $i = 1, 2, \dots, n; j = 1, 2, \dots, m$ , 得到矩阵:

$$Z = \begin{bmatrix} z_1^T \\ \vdots \\ z_n^T \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

c. 解得上述矩阵方程,解相关系数矩阵为

$$R = [r_{ij}]_{m \times m} = \frac{Z^T Z}{n-1}$$

d. 计算样本的相关系数矩阵方程,得到  $m$  个特征值  $\lambda_1 \sim \lambda_m$ , 其中  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ 。

e. 解得主成分  $Y = UX$ 。

上述方法所确定的包含原指标信息量最多的成分称为第一主成分,其次是第二主成分。如果有  $K$  个指标,则可以得到  $K$  个主成分,但只有贡献率较大,一般是累计达到 85% 的前几个主成分就能覆盖绝大部分信息量,主成分标准化得分的大小和各个输入变量因子的绝对值大小、符号有关。测试值越大,则第一主成分标准化的得分就越大。

### 2.2 BP 神经网络

神经网络是根据模拟人类大脑神经系统组织结构的原理来进行求解的。其中,目前神经网络技术的主流应用是 BP 神经网络,根据具体问题确定输入变量和目标输出变量,对 BP 神经网络进行设计,为神经网络选择适合的内部结构,主要包括选择网络的层数、每层的神经元数目和每层的激活函数等。

随着施工单位不断发展壮大,不断承接新的项目,企业过去投标的数据就会越来越完整,则建立 ANN 模型所需要的数据就会比较充分和完整,继而有利于标高金的预测。当网络的层数越多时,误差的平方和就会越小,神经网络预测的准确性就会提高。但过多的网络层数,又会使问题复杂化,增加训练的时间,使训练时间延长到几个小时甚至更长时间,不利于得出结果和优化结构。所以为提高速度,在设计中通常要尽可能地减少网络模型的层数,来

减少网络的训练时间。一般来说,隐含层神经元个数的增多,会提高网络训练的精度。但在选择隐含层神经元数目的时候,要考虑到神经元数目的增加对训练时间的影响。训练时间由循环的次数和每一次循环所要耗费的时间所决定。如果神经元的数目过少,训练时间减少了,网络精度也会降低;如果神经元数目过多,既会增加训练的时间,网络训练又会有训练过度导致失常的可能。目前,快速选取隐含层神经元数目的做法是:在能够解决问题的前提下,适当地增加 1~2 个神经元即可。

### 2.3 PCA-ANN 模型的构建

本文使用 BP 神经网络作为预测模型,网络输入端是标高金的影响因素,但由于影响因素较多,避免输入层的层数过多,导致训练过多目标层输出失常,反而结果会不准确,所以先用主成分分析法降维,减少输入端的网络层数,可以有效解决这一问题。输出端为标高金的预测值,模型的基本算法步骤如下:

a. 选取某一施工单位近几年的水利工程中标项目报价标高金原始资料。

b. 提取主成分。用 PCA 对输入变量进行降维处理,减少输入变量个数,确定神经元数目。选取累积方差 85% 的变量因子,降维处理后得到主成分变量因素。

c. 设计 BP 神经网络

①选择输入输出神经元数目。输入样本为影响投标报价的因素,其输出的特征向量为标高金,先确定输入层神经元的个数为降维后的个数和输出层的神经元数目。

②选择神经网络层次。根据 Kolmogrov 理论,对于任何在闭区间内的连续函数都可以用单隐含层的 BP 神经网络来逼近<sup>[13]</sup>。所以 3 层 BP 神经网络可以完成任意  $n$  维到  $m$  维的映射。也有研究表明,单隐含层的神经网络比多隐含层的神经网络的模拟结果有更高的精度。

③选择隐含层神经元数目。隐含层中神经元的个数要根据设计者的经验和反复的实验来确定。隐含层神经元的个数和实际的问题要求、输入和输出神经元数目都有紧密关系,如果隐含层神经元的数目过多,则会把样本中非规律性的内容也当成学习训练的内容,减低系统的泛化能力;如果神经元数目过少,网络从样本中获取信息的能力就会降低,不能够准确完全地反映样本规律。选用以往研究给出的经验公式来确定隐含层的神经元数目:

$$n_h = \sqrt{n + m} + a \quad (1)$$

式中: $n$  为输入层的神经元数目; $m$  为输出层的神经元数目; $a$  为 1~10 之间的常数。

根据以下公式确定的隐含层神经元的精度,精度最高的座位隐含层神经元数目。用总平均精度 OAA 来评价其优劣。定义 OAA 为

$$OAA = \left(1 - \frac{sd}{y_{mean}}\right) \times 100\% \quad (2)$$

式中: $sd = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$ ;  $y_{mean} = \frac{\sum_{i=1}^n y_i}{n}$ ;  $y_i (i=1, 2, \dots, n)$  为实际标高金值; $\hat{y}_i (i=1, 2, \dots, n)$  为 ANN 模型预测的标高金值; $n$  为样本数目。

d. BP 神经网络模型训练。可选用 MATLAB 软件帮助计算,先将原始数据进行归一化处理,将主成分对应的数据作为输入变量,输出变量标高金。网络训练的终止条件设定为网络总误差小于 0.1 或者训练次数大于 10000,得到 BP 神经网络训练误差曲线。

e. 预测数据。选择几组数据作为测试数据,输入网络中得到预测值,得到平均误差,分析是否满足要求。

## 3 案例分析

### 3.1 数据来源

数据来自于某一大型施工企业近五年的水利工程中标项目报价标高金原始资料,见表 2。

$X_1 \sim X_{10}$  代表每次投标的影响指标的取值, $Y$  代表了实际标高金(万元), $N$  表示实例的个数。由于影响标高金的因素差异大,具有一定的代表性,所以经过训练得到的模型可以有广泛的实用性。选取表 2 的前 27 个实例进行 ANN 训练,训练结束后利用没有参与训练的后 3 个实例进行预测。标高金及其影响因素的指标值见表 1。

### 3.2 提取主成分

由于变量 1 的指标值均为 100,其方差等于 0,可直接剔除不作为影响因素。再用 SPSS 数据编辑器将剩余 18 个影响因素指标值进行主成分分析,降维处理后得到 9 个主成分变量因素,见表 3。

### 3.3 设计 BP 神经网络

先确定输入层神经元的个数为降维后的个数 9 个,输出层的神经元数目为 1。根据公式(1),神经元数目为 6~16,使用 ANN 模型进行训练,得到的网络模型总平均精度如图 1 所示。当隐含层的神经元个数选取 11 时,总平均精度达到最高点,为 92.3%。

表2 标高金及其影响因素的指标值

N	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>	X <sub>6</sub>	X <sub>7</sub>	X <sub>8</sub>	X <sub>9</sub>	X <sub>10</sub>	X <sub>11</sub>	X <sub>12</sub>	X <sub>13</sub>	X <sub>14</sub>	X <sub>15</sub>	X <sub>16</sub>	X <sub>17</sub>	X <sub>18</sub>	X <sub>19</sub>	Y/万元
1	100	2300	100	50	0	14	0	8.2	8.9	3.7	230	11.2	50	50	4	50	50	100	100	6.4
2	100	6000	100	50	50	15	50	7.9	8.6	6.3	524	15.6	0	0	9	0	100	100	100	7.6
3	100	4200	50	0	0	21	0	6.2	9.2	4.5	305	14.6	0	100	7	100	50	50	100	8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	100	2900	100	0	50	16	0	8.5	9.3	3.6	325	18.2	50	100	5	100	100	100	0	7.9
26	100	3200	100	50	50	18	50	8.6	9.4	3.2	235	18.5	0	50	6	50	50	100	100	7.1
27	100	2900	50	50	100	16	0	7.9	8.6	3.7	420	18.3	0	0	7	0	100	50	100	6.2
28	100	2500	50	0	100	25	0	6.9	8.9	2.8	218	17.2	0	100	4	100	50	50	100	5.7
29	100	3700	50	0	100	26	0	6.7	8.6	3.1	390	18.5	0	50	5	50	50	100	100	6.4
30	100	4000	100	50	50	30	0	8.3	9	2.8	342	19.6	0	100	6	100	100	100	100	7.6

表3 各主成分的方差贡献率及累积方差贡献率

主成分	特征值	方差贡献率/%	累积方差贡献率/%
1	3.792	21.067	21.067
2	2.738	15.208	36.275
3	2.250	12.498	48.773
4	1.992	11.064	59.838
5	1.468	8.155	67.993
6	1.236	6.867	74.859
7	0.984	5.465	80.324
8	0.821	4.563	84.887
9	0.751	4.173	89.060

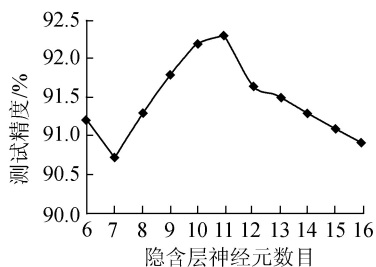


图1 不同数目的隐含层神经元产生的最好测试精度

所设计的BP神经网络投标报价预测模型选定为9-11-1的层次结构。

### 3.4 BP神经网络模型训练和预测

将1~27组的原始数据进行归一化处理,9个主成分对应的数据作为输入变量,输出变量为1~27组的Y。BP神经网络训练误差曲线见图2。

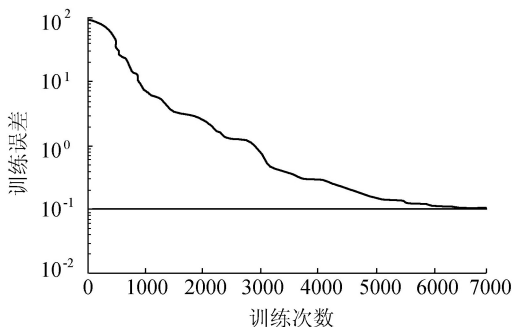


图2 BP神经网络训练误差曲线

图2表明,BP神经网络在训练次数达到6458次后总误差达到既定的要求0.1。用训练结束的神经网络开始进行预测时,将剩余的第28、29、30组数据

作为测试数据,输入网络中得到预测值,结果见表4。

表4 BP神经网络预测结果

序号	实际标高金/万元	计算标高金/万元	相对误差/%
28	5.7	5.43	4.74
29	6.4	6.19	3.28
30	7.6	7.49	1.45

经计算,平均误差率为3.17%,用神经网络模型预测得到的标高金值接近于实际值,可以对投标报价进行合理预测。

## 4 结论

a. PCA方法可重构多元变量输入状态空间,有效地避免了嵌入维数等参数的计算,能建立网络输入输出样本对。标高金的影响因素具有非线性、波动性、不确定性和复杂性的特征,PCA方法能使降维后的新向量较好地表征原始特征向量,有效地解决了原始特征参数维数过高的问题,使输入向量包含充分的预测信息。

b. 基于PCA-ANN相结合的标高金预测模型优化了BP神经网络建模方法的训练速度和迭代次数,可以准确地拟合原始样本。经过对模型的训练、检验,采用算例验证平均误差率为3.17%,说明此模型的预测精度较高,具有实际运用前景。

### 参考文献:

- [1] 韩敏,林云,孙燕楠,等. 基于神经网络的建筑行业投标报价研究[J]. 系统工程学报,2003(4):62-69.
- [2] BAJPAI P, SINGH S N. Strategic bidding in network constrained electricity markets using FAPSO[J]. International Journal of Energy Sector Management, 2008, 2(2): 274-296.
- [3] 叶公伟,袁永博. 基于最小二乘支持向量机的标高金预测方法研究[J]. 建筑管理现代化,2009(3):78-81.
- [4] 喻刚,王雪青,赵辉. 基于RS与ANFIS的投标报价决策研究[J]. 湖南大学学报:自然科学版,2008,35(5):94-97.
- [5] 陈群,孙磊. 蒙特卡罗法在投标报价中的应用[J]. 武汉理工大学学报:信息与管理工程版,2009(3):157-160.

(下转第66页)

0.832。这说明移民对目前最基本的生活条件感到满意,对技能培训带来的效果也予以肯定,同时移民迁入以后对后期扶持政策的落实以及移民干部的工作都感到满意。

根据标准化路径系研究发现,移民对生产条件改善不满意的地方主要集中于耕地数量和质量方面,主要是由于移民最初安置方式主要以后靠安置、集中迁建为主,这导致了移民迁入之后的耕地质量与数量都较之前标准有所降低,同时受库区和安置区自然资源的限制,通过土地整理等手段很难有效增加资源占有量;对生活条件中医疗卫生和子女上学的路径系数只有 0.674 和 0.704,这主要是因为部分村位置相对偏远导致孩子上学交通不便;虽然农村合作医疗的基本全覆盖使得移民的医疗卫生条件有了基本的保障,但是常年零碎的、较高的药费花费不予报销,无疑使得部分移民家庭的负担变重;后期扶持政策实施中路径系数较低的为资金发放,主要是由于移民户家庭收入的增加以及物价的上涨,使得每人每年 600 元的发放方式所起的作用相对变小,部分移民反映集中发放作用更大。

### 3 结 语

移民满意度较低的主要原因在于提高收入方式的转变。因此,在库区和移民安置区资源占有有限制以及固有收入模式发展进入瓶颈的同时,如何进一

步提供移民生活水平是后续工作的重点。笔者建议后期扶持政策实施中,进一步多措并举增加移民群众收入,进一步加快扶持重点转变,在巩固前阶段后期扶持政策实施成效的基础上,加大生产开发力度,努力实现改善移民生活环境、拓宽移民就业渠道、推动库区经济发展等新目标;同时,完善移民社会管理,进一步提高移民部门的服务工作能力,切实保障农民权益。

### 参考文献:

- [1] 谭宇,李忠斌.巴东县三峡库区移民安置满意度调查与分析[J].当代经济,2009(17):109-111.
- [2] 唐贵忠,周恒,云露.小浪底工程征地移民心理满意度的分析[J].人民黄河,2005(04):54-56.
- [3] 马力,夏立忠,李运东,等.三峡库首移民安置区土地资源、移民经济状况及移民满意度的调查与分析[J].长江流域资源与环境,2011(01):21-27.
- [4] 刘成斌,风笑天.三峡移民迁移满意度的转变及其根源[J].人口研究,2007(01):76-85.
- [5] 国务院.国务院关于完善大中型水库移民后期扶持政策的意见[EB/OL].[2006-05-17].[http://www.sdpc.gov.cn/ncjj/nczc/t20060814\\_80111.htm](http://www.sdpc.gov.cn/ncjj/nczc/t20060814_80111.htm).
- [6] 王卫东.结构方程模型原理与应用[M].北京:中国人民大学出版社,2010.

(收稿日期:2013-09-25 编辑:张志琴)

(上接第 55 页)

- [6] 宋维佳,张巍.基于先验信息的无标底投标报价策略研究[J].预测,2007(1):35-39.
- [7] 陈起俊,梁宝栋.BQ 计价模式下工程项目投标报价博弈模型研究[J].科技进步与对策,2012(18):68-71.
- [8] PNGHO S,ASCE A M,LING Y,et al. An analytical model for analyzing[J]. Journal of Construction Engineering and Management,2003,37(7):102-109.
- [9] NASSAR K. Pricing Construction Contracts in a Competitive Market: A Simulation Game[C]//ASC Proceedings of the 38th Annual Conference. Virginia Polytechnic Institute and State University-Blacksburg Virginia,2002:47-54
- [10] LIU S L,WANQ S Y,LAI K K. A general multivariate analysis approach for determining bid mark-up strategy [J]. Construction Management and Economics,2005,23(4):347-353.
- [11] FLORENCE Y Y,LIU M. Factors considered by successful and profitable contractors in mark-up size decision in Singapore[J]. Building and Environment,2005,40(11):1557-1565.

- [12] 张凌,李亚平.基于主成分和模糊聚类的区域自主创新环境评价[J].科技管理研究,2009(9):138-139.
- [13] 罗松江,丘水生,陈旭.一种混沌伪随机序列复杂度分析方法[J].华南理工大学学报:自然科学版,2010(1):22-25.

(收稿日期:2013-11-19 编辑:陈玉国)

