文章编号:1000-582X(2009)09-1104-07

特征提取和小样本学习的电力工程造价预测模型

彭光金,俞集辉,韦俊涛,杨　光
(重庆大学 输配电装备及系统安全与新技术国家重点实验室,重庆 400030)

摘　要:通过特征提取和小样本学习的结合,提出一种全新的基于混合算法的电力工程造价预测模型。利用主成分分析对原始指标进行预处理,消除原始指标之间的相关性,并提取潜在的综合独立指标,将新指标作为输入集构造基于最小二乘支持向量机的预测学习模型,将其预测结果和神经网络模型预测对比分析。并通过不同主成分数目预测结果的比较,确定最优的主成分个数,达到理想的预测效果。实例预测结果表明:该方法可以有效提取原始指标的信息量,在小样本学习方面表现突出,能够达到期望的预测效果。

关键词:电力工程;预测模型;主成分分析;最小二乘支持向量机;小样本学习
中图分类号: TM 743　　文献标志码:A

Cost forecast model for power engineering based on feature extraction and small-sample learning

PENG Guang-jin, *YU Ji-hui*, *WEI Jun-tao*, *YANG Guang*
(State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing University, Chongqing 400030, P. R. China)

Abstract: A novel power engineering cost forecast model was proposed by combining feature extraction and small-sample learning. The initial data was preprocessed with principal component analysis to remove the correlation among the original indexes and get the potential independent indexes. The new indexes acted as the input set to build a new forecast model based on least squares support vector machines. The results of this model were compared with the forecast results getting from artificial neural network. By comparing the forecast results with different principal components number, the optimal number was determined to achieve the desired forecast effect. The prediction results indicate that the method can extract the feature of initial data effectively and is good at small-sample learning . The expected forecasting results can be reached.

Key words: power engineering; forecast model; principal component analysis; least squares support vector machines; small-sample learning

国民经济的快速发展对电网容量提出了更高的要求,随之而来电力建设投入在不断的增加,也给电力建设投资方的投资管理带来了大量的问题。一方面,传统的定额概预算管理制度虽然对于合理确定和控制工程造价起着积极作用,但“量价合一”的管理方式过分限制了投资方自由定价的权利,使得工程造价脱离市场,不利于施工方改进技术条件,提高竞争力。另一方面,在实际电力工程造价管理中除

收稿日期:2009-05-18
基金项目:重庆市自然科学基金资助项目(CSTC2006BA6015);国家电力公司科技项目(04207520070603)
作者简介:彭光金(1970-),男,重庆大学博士,主要从事电力工程优化管理,电力市场等领域研究。
俞集辉(联系人),男,重庆大学教授,博士生导师,(Tel)023-65112230;(E-mail)yujihui@cqu.edu.cn。

了受电网整体规划、总容量、地形特征、设计施工水平等因素影响外还与建设地域的综合经济水平有密切关系。在同一段时期内，收集的相似可比对的历史工程不多，从有限的资料积累中提取更多的知识信息变得困难重重，不能及时有效地对新的工程进行预测和指导。因此，投资方和施工单位迫切需要寻求一种理想的预测方法，能够利用有限的历史工程以及较少的指标，快速估算出新建工程的造价，以便合理指定投资方案，为工程顺利建设争取主动时间，提高工程项目投资的审查效率，指导新建工程的造价管理。

同时，在工程造价预测方面，随着智能算法研究的深入，出现了很多新的预测方法，例如神经网络[1-2]，遗传算法[3]，小波分析[4]，支持向量机[5]等等。但单一的算法在实际工程应用中有着各自的缺陷。神经网络出现了网络结构设计困难，需要大量的数据样本和长时间训练的问题[6]；遗传算法存在交叉率、变异率等复杂参数设置问题，虽然可以保证群体的进化性，但不可避免地出现了个体退化现象[7]；小波分析对小样本学习训练精度很低且学习过程复杂 。虽然支持向量机算法能够较好地解决小样本、非线性、高维数以及局部极小点等实际问题，但针对电力工程预测的特殊性，仅仅利用支持向量机预测，计算精度较差的弱点难以克服。有鉴于此，笔者提出一种基于主成分分析（principal component analysis，PCA）和改进支持向量机——最小二乘支持向量机（least squares support vector machines，LS-SVM）的混合算法，通过特征提取和小样本学习来构建电力工程造价预测模型，以达到较快地预测电力工程造价和指导新建工程造价管理的目标。

1　主成分分析的数学原理

主成分分析法[8-9]最早是由 Pearson 对非随机变量研究时引入的，后来 Hotelling 将此方法推广到随机向量的情形，该方法具有严格的数学理论基础。主成分分析的主要目的是通过特征提取，得到较少的综合独立指标，并且最大限度地保证原有信息量不丢失[10]。从另一个角度来看，主成分分析实质上也是一种数据降维的方法。

设有某个 p 维总体变量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$，$\boldsymbol{R}$ 为 $\boldsymbol{X}$ 的协方差矩阵，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$ 为 $\boldsymbol{R}$ 的特征值，$\boldsymbol{U}=(u_1,u_2,\cdots,u_p)^{\mathrm{T}}$ 为对应的标准正交特征向量，则对应的主成分可表示为：

$$Z_i=u_{i1}X_1+u_{i2}X_2+\cdots+u_{ip}X_p=\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{X}。\quad(1)$$

式(1)中 $Z_i(i=1,2,\cdots,p)$即为新的主成分。从代数观点来看，Z_i 为原始变量 $\boldsymbol{X}$ 的一些特殊线性组合，且主成分间互相正交，没有冗余信息，构成了数据空间的正交基。而特征值 $\lambda_i(i=1,2,\cdots,p)$主要反映了与之对应主成分所包含原信息的比重。各个主成分包含的信息量随着特征值的减小而减少。

定义第 j 个主成分 Z_j 所提出的信息量占全部信息量比例为 $\lambda_j/\sum_{i=1}^{p}\lambda_i$，称为第 j 个主成分的贡献率 q_j。那么，前 m 个主成分的贡献率之和 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$ 称为前 m 个主成分的累积贡献率。在实际应用中，一般要求累积贡献率 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i>T$（某一特定值），就可以认为前 m 个主成分综合了原始变量的绝大部分信息，不再提取新的主成分。

2　最小二乘支持向量机模型

支持向量机理论是 Vapnik 等人[11-12]在1995 年根据统计学习理论提出的一种新的机器学习方法，近年来在模式识别、回归分析和特征提取等方面得到了广泛的应用。它采用结构风险最小化原则代替了传统统计学习中的经验风险最小化原则，通过寻求结构风险最小来提高学习机的泛化能力，实现了经验风险和置信范围均最小。从而达到在小样本情况下，亦能获得良好的统计规律的目的。并且有效地解决了过学习问题，具有良好的推广性能和较好的预测精确性。1999 年，Suykens J. A. K[13]提出了最小二乘支持向量机理论，将最小二乘线性系统引入到支持向量机中，代替了传统采用二次规划方法解决函数估计问题，从而进一步提高了学习的精度。

假设给定数据集 $S=\{x_i,y_i\}$，$i=1,2,\cdots,m$，其中 $x_i,y_i\in\boldsymbol{R}^n$。那么，预测问题的实质也就是，通过样本数据集 $S_{m-l}\cup\{x_{m-l+1},\cdots,x_m\}$ 获得预测值 $\{\tilde{y}_{m-l+1},\cdots,\tilde{y}_m\}$，$l=1,2,\cdots,m$。在线性函数集中即为寻找函数

$$f(x)=(\boldsymbol{w}\cdot\boldsymbol{x})+b,\quad(2)$$

其中：$\boldsymbol{w}\in\boldsymbol{R}^n$（原始空间）为权向量；$b$ 为偏置量。

而利用支持向量机求解非线性函数优化问题，主要是通过一个非线性映射将 S 非线性映射到一个高维特征空间（Hilbert 空间），将非线性函数优化转化为高维特征空间中的线性函数优化问题[14]，关系如式(3)。

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b,\quad(3)$$

其中 $\varphi(x):\boldsymbol{R}^n\rightarrow\boldsymbol{R}^{n_c}$ 为 Hilbert 空间的映射函数。

对于LS-SVM优化问题来说，可以转化成

$$\left.\begin{aligned}&\min_{w,b,\xi} J(\boldsymbol{w},\xi)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\gamma\sum_{k=1}^{N}\xi_k^2,\\&\text{s.t. } y_k-[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]=1-\xi_k。\end{aligned}\right\}\quad(4)$$

LS-SVM与SVM的最大区别就是使用二次损失函数取代SVM中的不敏感损失函数，将不等式约束条件变为等式约束，其中，$k=1,2,\cdots,m$；γ是可调超参数，亦叫惩罚系数；$\xi_k\in\boldsymbol{R}$为一误差变量。

根据式(4)中的目标函数和约束条件建立Lagrange函数[15]，可将原优化求解问题转化成式(5)。

$$L(\boldsymbol{w},b,\xi_k,\alpha_k)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+r\sum_{k=1}^{m}\xi_k^2-\sum_{k=1}^{m}\alpha_k\{y_k-[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]+\xi_k-1\},\quad(5)$$

其中拉格朗日乘子$\alpha_k\in\boldsymbol{R}$。

对式(5)进行优化求解，即L分别对变量$\boldsymbol{w}$，b，ξ_k，α_k求偏导并等于0，结果如式(6)。

$$\left.\begin{aligned}&\frac{\partial L}{\partial \boldsymbol{w}}=\boldsymbol{w}+\sum_{k=1}^{m}\alpha_k\varphi(x_k)=0,\\&\frac{\partial L}{\partial b}=-\sum_{k=1}^{m}\alpha_k y_k=0,\\&\frac{\partial L}{\partial \xi_k}=2r\sum_{k=1}^{m}\xi_k-\sum_{k=1}^{m}\alpha_k=0,\\&\frac{\partial L}{\partial \alpha_k}=y_k-[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]+\xi_k-1=0。\end{aligned}\right\}\quad(6)$$

消除变量$\boldsymbol{w}$，ξ_k，可得以下矩阵方程

$$\begin{bmatrix}0 & -\boldsymbol{Y}^{\mathrm{T}}\\ \boldsymbol{Y} & \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}-\dfrac{1}{2\gamma}I\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ I\end{bmatrix},\quad(7)$$

其中：$\boldsymbol{Z}=[\varphi(x_1)^{\mathrm{T}}y_1\quad\cdots\quad\varphi(x_m)^{\mathrm{T}}y_m]^{\mathrm{T}}$，$\boldsymbol{Y}=[y_1\quad\cdots\quad y_m]^{\mathrm{T}}$，$\boldsymbol{I}=[1\quad\cdots\quad1]^{\mathrm{T}}_{m\times1}$，$\boldsymbol{\alpha}=[\alpha_1\quad\cdots\quad\alpha_m]$，根据mercer条件[16]，一定存在着映射函数φ和核函数$\Psi(\cdot,\cdot)$，使得式

$$\Psi(x_k,x_l)=\varphi(x_k)^{\mathrm{T}}\varphi(x_l)\quad(8)$$

成立。

那么，最小二乘支持向量机的函数估计可表示为

$$y_k=\sum_{k=1}^{N}\alpha_k\psi(x,x_k)+b,\quad(9)$$

其中：α，b由式(7)求解，$\alpha_k\neq0$对应的样本即为支持向量。核函数$\Psi(\cdot,\cdot)$的目的就是从原始空间中抽取特征，将原始空间中的样本映射为高维特征空间中的一个向量，以解决原始空间中的线性不可分的问题。

核函数有很多设置形式，在文中使用了径向基核函数

$$\psi(x,x_i)=\exp\left(-\frac{|x-x_i|^2}{\sigma^2}\right),\quad(10)$$

其中σ^2为核函数的宽度系数，可根据实际情况设置。

3 电力工程造价预测模型

针对电力工程造价中的实际情况，笔者设计了基于特征提取和小样本学习的预测模型，如图1所示。

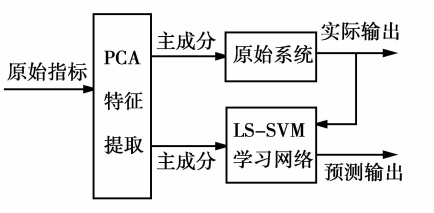

图1 预测模型原理图

电力工程造价管理中受影响因素多的问题，在此将原始指标进行PCA特征提取，以累积贡献率为标准，得到新的主成分，达到对工程属性指标维数压缩的目的。将主成分作为新的输入指标集，同时将历史工程分为两部分，一部分作为训练样本，另一部分作为测试样本。将测试样本输入学习网络，利用LS-SVM在小样本学习方面的优势，构建基于LS-SVM的学习模型，进行网络训练。将测试样本输入训练网络，得到预测输出并与实际值对比分析。这样在大大降低原始空间维数的同时，最大限度地保证原有信息量不丢失，并且可以利用LS-SVM小样本学习的优点提高预测精度。

4 实际工程造价数据预处理

4.1 原始指标数据的简化规约

结合某地区电力工程造价管理实际情况，笔者拟以送电线路工程历史数据为研究对象，对造价情况进行分析。通过资料收集，初步得到某地区220 kV历史送电线路工程44个，但由于不同的导线截面对单位造价影响较大，所以仅保留截面为400 mm^{2}的工程。同时结合电力专家建议：过短的线路长度会大幅提高工程单位造价，将不符合条件的工程剔除后，样本工程剩下37个。

对样本指标进行简化处理，可得到以下原始输入集和输出集。

输入集＝{ 单回比例，双回比例，综合地形，单位转角次数(次/km)，单位跨越(次/km)，单位平均档距，单位耐张段，单位人力运距，单位汽车运距，单位铁塔(基/km)，导线(万元/km)，杆塔钢材(t/km)，基础钢材(t/km)，接地钢材(t/km)，地线(t/km)，挂线金具(t/km)，土石方(m^3/km)，混凝土(m^3/km)}；

输出集＝{单位静态投资(万元/km)}；

说明：输入集中无单位的指标，均进行单位化处理，指标无量纲。

通过输入集和输出集可以看出，输入集共包含 18 个指标，输出集有 1 个指标。因此指标输入构成了 18×37 的矩阵。

4.2 *数据预处理*：

在实际工程中，原始指标表示含义不同，从而数量级之间相差甚远，而主成分分析法在提取主成分时，会“偏爱”数量级偏大的指标，而忽略或者抛弃数量级特别小的指标，从而造成主成分提出主观上的错误，因此在主成分分析前，一般会对指标样本进行标准化处理。

针对原始 p 个指标变量 $X_1, X_2, \cdots, X_P$，用公式(11)进行标准化。

$$\widetilde{X}_i = \frac{X_i - \boldsymbol{E}(X_i)}{\sqrt{\boldsymbol{D}(X_i)}}, \tag{11}$$

其中：$i=1,2,\cdots,p$；$\boldsymbol{E}(X_i)$为变量 X_i 对应的数学期望；$\boldsymbol{D}(X_i)=\frac{1}{n}\sum_{i=1}^{p}(X_i-\overline{X})^2$ 为变量 X_i 对应的方差，反映数据集的离散程度；$\overline{X}$ 为平均值。

针对原始数据集，p 取 18，在 Matlab6.5 下编制标准化程序，并调试运行，结果进行对比分析：原始指标数据分布在 0～1 400 之间，数据分布不均，数量级相差较大。特别是指标 6(单位平均档距)和指标 17(土石方量)表现最为突出。而标准化处理后各指标数据分布变化不大，分布均匀：$\max(\widetilde{X}_i)=5.517$，$\min(\widetilde{X}_i)=-3.692$。大大降低了由于指标数量级不同而对主成分分析造成的影响。从而得到了新的数据集($\widetilde{X}_1,\cdots,\widetilde{X}_p$)。

5　预测模型的建立与仿真分析

5.1 *主成分的提取*

对应经标准化处理得到的新数据集 $\widetilde{X}_1,\cdots,\widetilde{X}_p$ 建立主成分分析模型。在 Matlab6.5 下，将 $\widetilde{X}_1,\cdots,\widetilde{X}_p$ 作为输入数据集编制主成分分析程序，经运行测试得到各主成分的特征值和贡献率如表 1 所示。

表 1　特征值和主成分贡献率

主成分	特征值	贡献率/%	累积贡献率/%
1	8.006 100	44.478 00	44.478
2	2.724 200	15.134 00	59.612
3	1.628 500	9.047 20	68.659
4	1.265 600	7.031 10	75.691
5	1.040 800	5.782 20	81.473
6	0.896 900	4.982 70	86.455
7	0.685 300	3.807 20	90.263
8	0.447 080	2.483 80	92.746
9	0.372 980	2.072 10	94.819
10	0.307 260	1.707 00	96.526
11	0.253 370	1.407 60	97.933
12	0.143 590	0.797 72	98.731
13	0.099 469	0.552 60	99.283
14	0.071 465	0.397 02	99.680
15	0.029 329	0.162 94	99.843
16	0.022 646	0.125 81	99.969
17	0.005 543	0.030 79	100.000
18	$1.128\ 3\times10^{-12}$	$6.277\ 7\times10^{-12}$	100.000

通过表 1 可以看出，18 个主成分对应的特征值均大于 0，满足主成分分析要求，且特征值从主成分 Z_1(8.006 1)开始依次递减，表明了主成分 1 包含了最多的原指标信息量，信息量也随着主成分的增大而依次减少，直到主成分 18，特征值为 $1.128\ 3\times10^{-12}$，所包含原指标的信息量几乎可以忽略不计。通过贡献率的计算，主成分 Z_1 的贡献率已达到了 44.478%，即主成分 Z_1 携带了原指标将近一半的信息量。在实际工程应用中，一般设定有累积贡献率 $Q\geqslant85\%$，可认为新主成分已经可以替代原指标数据。可以看出，前 6 个主成分的累积贡献率为 86.455%，已经达到了主成分提取的要求。但为了更好地反映原指标的情况以及原指标信息量多少对预测模型的影响，分别选取前 6 个主成分($Q=86.455\%$)，7 个主成分($Q=90.263\%$)，10 个主成分($Q=96.526\%$)来分别代替原指标集。

5.2 ***LS-SVM*** *预测模型*

经主成分提取后，利用达到累积贡献率 Q 要求

的主成分作为预测模型的输入集，单位静态投资作为输出集，构建造价预测平台。根据专家经验，一般选取核函数的宽度系数为0.2，而惩罚系数 γ 选取为50。

具体步骤如下：

1)利用PCA提取的主成分形成新的训练样本集和测试集(5个测试样本)；

2)在Matlab6.5添加LS-SVM软件包，编写训练程序；

3)取前6个主成分作为输入变量，分别取15，25，30个学习样本训练网络，得到学习网络后，对5个测试样本进行预测比对；

4)分别取前6个，7个，10个主成分，训练LS-SVM学习网络和ANN学习网络，比对输出结果；确定主成分数目，得到最优预测值。

在步骤3)中，通过不断增添新的训练样本到学习网络中来提高学习网络的泛化能力，得到理想预测效果，不同的学习样本数对学习结果的影响如图2和表2所示。

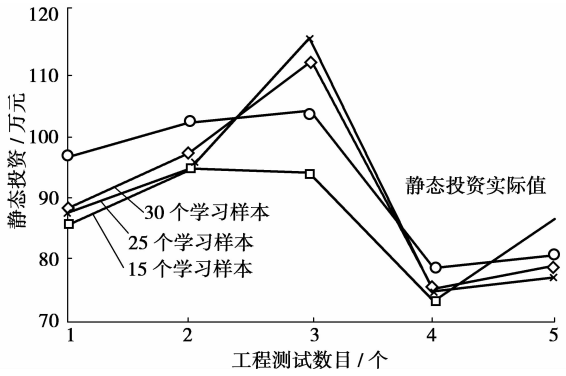

图2 15,25,30个学习样本预测效果对比

表2 相对误差对比

静态投资	相对误差/%		
	15个样本	25个样本	30个样本
96.45	−11.370 0	−9.511 3	−8.715 6
101.63	−7.332 7	−7.185 9	−4.752 3
103.55	−9.198 5	11.144 0	7.841 6
78.31	−6.333 2	−4.466 4	−3.984 5
80.82	6.699 2	−4.153 5	−2.142 0

通过图2可以看出，随着学习样本数的增加，预测模型的泛化能力在不断的提升。15个样本的学习能力较差，与实际值误差较大，当样本数目分别为25个和30个时，学习能力明显提高，误差大幅降低，特别是学习样本增加到30个时，证明了LS-SVM针对小样本学习具有较好的泛化能力。

在表2中，15个学习样本时预测值与实际值误差较大，当样本增加到25个时相对误差 $\max(\sigma)=11.144\%$，$\min(\sigma)=-4.153\,5\%$，到学习样本达到30个时，预测值的相对误差 $\max(\sigma)=-8.715\,6\%$，$\min(\sigma)=-2.142\%$。通过25个测试样本和30个测试样本比较可以看出 $\max(\sigma)=3.302\,4\%$，$\min(\sigma)=0.481\,9\%$。预测结果误差变化趋势减弱，模型趋于稳定。表明该学习网络较稳定。

为了尽可能反映原指标所携带的信息量，将实际工程项目37个分为32个学习样本和5个测试样本，并分别用前6，7，10个主成分来训练LS-SVM学习网络和ANN学习网络，2个学习网络的预测结果对比如表3所示。

表3 LS-SVM和ANN预测结果对比分析

主成分个数	静态投资/万元	LS-SVM			ANN		
		测试值/万元	绝对误差/万元	相对误差/%	测试值/万元	绝对误差/万元	相对误差/%
前6个主成分	96.45	89.682	−6.768	−6.536 0	72.762	−23.688 0	−24.56
	101.63	98.217	−3.413	−3.296 0	93.874	−7.756 3	−7.49
	103.55	109.340	5.790	5.591 5	81.503	−22.047 0	−21.29
	78.31	74.708	−3.602	−3.478 5	55.708	−22.602 0	−21.83
	80.82	79.416	−1.404	−1.355 9	59.626	−21.194 0	−20.47

续表

主成分个数	静态投资/万元	LS-SVM			ANN		
		测试值/万元	绝对误差/万元	相对误差/%	测试值/万元	绝对误差/万元	相对误差/%
前 7 个主成分	96.45	91.308	−5.142	−4.965 7	100.600	4.145 8	4.00
	101.63	98.522	−3.108	−3.001 4	109.760	8.130 0	8.00
	103.55	107.100	3.550	3.428 3	90.080	−13.467 9	−13.01
	78.31	75.501	−2.809	−2.712 7	52.471	−25.839 0	−24.95
	80.82	79.922	−0.898	−0.867 2	52.621	−28.199 0	−27.23
前 10 个主成分	96.45	90.693	−5.757	−5.970 0	90.272	−6.178 4	−5.97
	101.63	95.165	−6.465	−6.360 0	86.900	−14.730 0	−14.23
	103.55	110.360	6.810	6.580 0	52.313	−51.237 0	−49.48
	78.31	84.902	6.592	8.420 0	55.221	−23.089 0	−22.30
	80.82	80.329	−0.491	−0.610 0	34.292	−46.528 0	−44.93

通过表 3 可以看出,除一个预测点外,LS-SVM 网络预测结果的相对误差均小于 ANN 网络。同时,ANN 网络预测结果相对误差变化较大。表明 LS-SVM 学习网络在小样本学习方面无论是预测精度还是网络稳定性均优于 ANN 学习网络。

在 LS-SVM 学习网络模型中,用前 6,7,10 个主成分构建的学习网络对预测值的影响如图 3 所示。

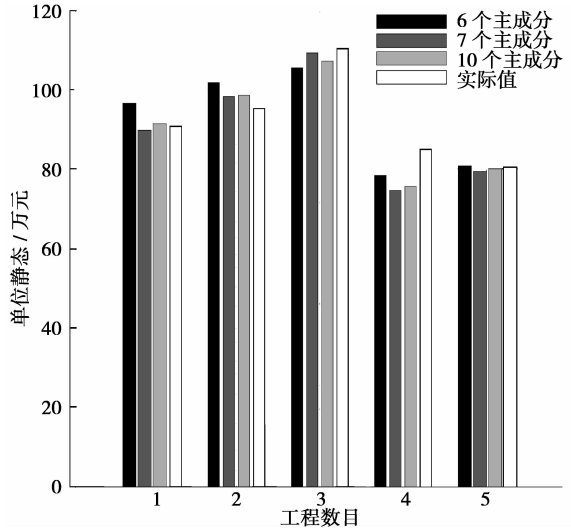

图 3　预测值与实际值对比

图 3 中,每个预测工程从左向右依次为 6,7,10 个主成分以及实际单位静态投资。可以看出当主成分个数为 7 时,学习网络的预测值与实际值误差最小,同时相比其余两组预测效果较好。通过分析可知,前 6 个主成分携带的信息量少于前 7 个主成分是误差增大的关键。而前 10 个主成分时,由于主成分 8,9,10 的特征值过小,虽然累积贡献率在增加,但主成分之间的自相关性也随之增大,反而降低了学习网络的积极性,从而造成了误差的增大。所以在实际预测模型中采用 7 个主成分作为输入集,构建预测模型。结果表明:相对误差在 −4.965 7%～3.428 3%时,基本符合实际工作需要。

6　结　论

将特征提取和小样本学习相结合,形成了一种新的混合型预测算法,并将其应用于电力工程造价预测中。在工程造价受影响因素多,且一段时期内搜集的历史工程有限的前提下,构建了基于 PCA 和 LS-SVM 的电力工程造价预测模型。结合某一地区的实际送电线路工程历史数据,使用 PCA 对原始指标进行特征提取,得到少量综合独立指标。将综合指标作为输入集,构建了基于 LS-SVM 的预测模型,确定了最优的主成分数目,得到预测结果。从结果来看,该模型在小样本预测方面,无论从预测的精度还是泛化能力,均优于 ANN 学习网络。因此,该模型能够在项目建设初期辅助建设方法人合理确定投资方案,为工程顺利建设争取时间和取得主动权,提高了工程造价投资方案的审查效率。实例仿真分析表明:该种混合型工程造价预测方法具有较好的可行性和适应性,便于推广应用到相类似的领域中。

参考文献：

[1] MANDAL P, SENJYU T, UEZATO K, et al. Several-hours-ahead electricity price and load forecasting using neural networks[C] // 2005 IEEE Power Engineering Society General Meeting, June 12-15, 2005, San Francisco, California. [S. l.]: IEEE, 2005: 2146-2153.

[2] DONG J R. A nonlinear combining forecast method based on fuzzy neural network[C]// 2002 International Conference on Machine Learning and Cybernetics, Nov 4-5, 2002, Beijing, China. [S. l.]: IEEE, 2002: 2160-2164.

[3] 周双喜,郑智,鲁宗相. 基于多种群遗传算法的规划[J]. 电力系统及自动化,2007,19(6):66-71.
ZHOU SHUANG-XI, ZHENG ZHI, LU ZHONG-XIANG. New reactive power planning based on the multiple-population genetic algorithm [J]. Proceedings of the CSU EPSA,2007,19(6):66-71.

[4] 姚李孝,刘学琴. 基于小波分析的月度负荷组合预测[J]. 电网技术,2007,31(19):65-68.
YAO LI-XIAO, LIU XUE-QIN. A wavelet analysis based combined model for monthly load forecasting[J]. Power System Technology, 2007,31(19):65-68.

[5] 罗楠,朱业玉,杜彩月. 支持向量机方法在电力负荷预测中的应用[J]. 电网技术,2007,31(2):215-218.
LUO NAN,ZHU YE-YU, DU CAI-YUE. Application of support vector machine method in electric load forecasting [J]. Power System Technology, 2007, 31(2):215-218.

[6] METHAPRAYOON K, LEE W J, RASMIDDATTA S, et al. Multistage artificial neural network short-term load forecasting engine with front-end weather forecast [C] // 2006 IEEE Industrial and Commercial Power Systems Technical Conference. [S. l.]: IEEE, 2006: 1410-1416.

[7] 顾峰,艾芊,凌建峰. 基于小生境免疫算法的月负荷预测组合模型[J]. 电网技术,2007,31(1):1-5.
GU FENG, AI QIAN , LING JIAN-FENG. A comprehensive model of power load forecasting based on niche immune algorithm[J]. Power System Technology, 2007,31(1):1-5.

[8] GAO H B, HONG W X, CUI J X, et al. Optimization of principal component analysis in feature extraction [C]// International Conference on Mechatronics and Automation, Aug 5-8, 2007, Harbin, China. [S. l.]: IEEE, 2007:3128-3132.

[9] LU S, LI M. Bearing fault diagnosis based on PCA and SVM[C] // 2007 International Conference on Mechatronics and Automation, Aug 5-8,2007,Harbin, China. [S. l.]: IEEE,2007: 3503-3507.

[10] MAENAKE T, HONDA K, ICHIHASHI H. Local independent component analysis with fuzzy clustering and regression-principal component analysis[C]// 2006 IEEE International Conference on Fuzzy Systems, Sept 17-20, 2006, Baoding, China. [S. l.]: IEEE, 2006: 857-862.

[11] VAPNIK V N. Statistical learning theory [M]. New York:John Wiley,1998.

[12] 朱家元,杨云,张恒喜. 基于优化最小二乘支持向量机的小样本预测研究[J]. 航空学报,2004,25(6):565-568.
ZHU JIA-YUAN, YANG YUN, ZHANG HENG-XI. Data prediction with few observations based on optimized least squares support vector machines [J]. Acta Aeronautica et Astronautica Sinica, 2004,25(6):565-568.

[13] 阎威武,邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. 控制与决策,2003,18(3):358-360.
YAN WEI-WU, SHAO HUI-HE. Application of support vector machines and least squares support vector machines to heart disease siagnoses [J]. Control and Decision,2003,18(3):358-360.

[14] 田景文,高美娟. 人工神经网络算法研究及应用[M]. 北京:北京理工大学出版社,2006.

[15] 王晶,靳其兵,曹柳林. 面向多输入输出系统的支持向量机回归[J]. 清华大学学报:自然科学版, 2007,47(S2): 1737-1741.
WANG JING, JIN QI-BING, CAO LIU LIN. Support vector regression algorithm for multi-input multi-output systems [J]. Journal of Tsinghua University: Science and Technology, 2007,47(S2): 1737-1741.

[16] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge: Cambridge University Press ,2000:30-34.

（编辑 李胜春）