[文章编号] 1671-9727(2010)02-0206-05

基于主成分分析的BP神经网络及其在需水预测中的应用

龙训建[1,3] 钱 鞠[2] 梁 川[1]
(1.四川大学水电学院,成都 610065; 2.兰州大学资源环境学院,兰州 730000;
重庆水电职业技术学院,重庆 402160)

[摘要] 以甘肃省瓜州县为例,利用1988～2007年的总需水量数据,采用主成分分析法对影响水资源需求量的7个因子进行主要影响因子分析,根据确定的主要影响因子构造BP神经网络的输入样本,从而进行不同水平的年总需水量预测。结果表明:国内生产总值、工业总产值、农业总产值和大牲口数4个因子为影响研究区需水量的主要因子,将此作为主要因子构造BP神经网络的输入样本,确定网络输入节点数,建立瓜州县总需水量预测模型。模拟计算结果表明,基于主成分分析的BP神经网络模型取,预测结果的绝对误差小于$\pm 0.05\times 10^9$ m^3。

[关键词] 需水预测;主成分分析法;BP神经网络
[分类号] TP183;TV214 [文献标识码] A

水资源是人类社会生存与发展中不可替代的重要自然资源及生态环境系统的基本要素[1]。随着人口增长和工农业生产的发展,水资源供需矛盾日益加剧,所面临的水危机日益严重;加上社会经济用水挤占了生态用水,使得天然河湖萎缩、消失,土地荒漠化等,造成生态环境失调,严重阻碍了经济社会的持续发展[2～4]。用水量的高速增长和水资源的短缺使得水资源规划和用水系统的优化调度越来越重要,进行需水量预测则成为了实现水资源规划和管理的有效手段之一[5]。作为全国首批节水型社会试验县之一,近年来,甘肃省瓜州县经济建设快速发展,产业结构战略性调整,对水资源需求、开发和利用提出了新要求,因此,确保水资源的高效利用成为了经济和社会可持续发展的重要保证。

国内外关于需水预测研究历史悠久,但由于近年来水资源与经济社会之间的矛盾愈加突出,对于水资源需求的研究方式也呈现多样化发展趋势[6]。各种计算模型、模拟程序(如定额预测、回归分析预测、指数法预测、灰色模型预测、神经网络模型等)都尝试应用于需水预测方面的研究[7～9]。选择合理的需水预测方法,不仅可以增加水资源配置研究工作的实际操作性,还可得到更符合社会经济发展趋势的结果。无论采用何种需水预测模型,都需要对影响因子进行筛选。因子选择过少,必然会影响预测结果的准确性;因子过多,会使网络训练复杂化,可能陷入局部优化问题,难以得到全局优化解。人工神经网络是当前国际学术界十分活跃的前沿研究领域,具有广泛的应用领域。本文首先应用主成分分析法确定影响总需水量的主要因子,然后以此构造BP神经网络的输入样本,进行BP神经网络的训练与预

[收稿日期] 2009-04-14
[基金项目] 国家科技支撑计划项目(2007BAD88B0804-3)
[作者简介] 龙训建(1982—),女,硕士,讲师,研究方向:水资源规划与管理,E-mail:lmcx402331@foxmail.com。
[通讯作者] 钱鞠,男,副教授,研究方向:水资源与水环境,E-mail:qianju@lzu.edu.cn。

测,以提高模型的学习和泛化能力。

1　研究方法

1.1　主成分分析法

在相关影响因子分析中,主成分分析是将多个指标化为少数相互无关的综合指标的统计方法。其基本思想是通过变量的相关系数矩阵内部结构的研究,找出能控制所有变量的少数几个随机变量去描述多个变量直接的相关关系[10]。从数学角度而言,这属于降维处理技术。对于有 n 个样本的 p 个变量的原始资料矩阵 $X_{(n\times p)}$,进行主成分分析过程为:

a.对原始数据矩阵 $X_{(n\times p)}$ 标准化处理,得到新的数据矩阵

$$Y=(y_{ij})_{n\times p} \tag{1}$$

式中:$y_{ij}=\dfrac{x_{ij}-\overline{x_j}}{\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}}$;$\overline{x_j}=\dfrac{1}{n}\sum_{i=1}^{n}x_{ij}$;$j=1,2,\cdots,p$。

b.建立标准化后的 p 个指标的相关系数矩阵 R

$$R=(r_{ij})_{p\times p} \tag{2}$$

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2\sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}} \tag{3}$$

c.计算相关矩阵 R 的特征值及相应的特征向量 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$,并使其从大到小排列;同时求得对应的特征向量 $u_1,u_2,\cdots,u_p$。

d.计算贡献率 e_m 和累计贡献率 E_m。

$$e_m=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{4}$$

$$E_m=\sum_{j=1}^{m}\lambda_j/\sum_{i=1}^{p}\lambda_i \tag{5}$$

e.计算主成分荷载 z_m。它表示主成分与变量之间的相关系数。

$$z_m=\sum_{j=1}^{n}\sum_{i=1}^{p}u_{ij}y_{ij} \tag{6}$$

1.2　BP 神经网络需水预测

人工神经网络是通过数学方法对人脑若干基本特性进行的抽象和模拟,是一种模仿人脑结构及其功能的非线性信息处理系统[11,12]。经过半个多世纪的发展,由于各种网络结构和算法系统的产生,已逐渐发展成较为完善的人工神经网络理论体系。BP 神经网络是该技术中应用最为广泛的一种,其特点有[11,13~15]:自适应、自组织、自学习的能力、非局域性和非凸性的突出优点。正是这些特点使得该方法已经解决了许多实际问题,其生命力也恰恰在于广泛的实用价值[11]。为了能够更好地泛化全局最优问题,许多学者提出了很多针对性的办法,主要包括以下三方面的改进[11]:一是提高网络的训练速度;二是提高训练精度;三是避免落入局部极小点。

通常情况下,需水预测 BP 神经网络模型包括输入层、隐含层和输出层 3 部分。3 层 BP 神经网络模型的理论计算步骤大致为:

第 1 步,将样本的输入、输出变量归一化处理,即将所有数据转化至[0,1]之间。给每个连接权值 w_{ij},v_{jt},阈值 θ_j 与 γ_t 赋予区间(−1,1)内的随机值。

第 2 步,用输入样本 $x_k=(x_1{}^k,x_2{}^k,\cdots,x_n{}^k)$、连接权值 w_{ij} 和阈值 θ_j 计算隐层各单元的输入 a_j,然后用 a_j 通过传递函数计算隐层各单元的输出 b_j。

$$a_j=\sum_{i=1}^{n}w_{ij}x_i-\theta_j\quad(j=1,2,\cdots,p) \tag{7}$$

$$b_j=f(a_j)\qquad(j=1,2,\cdots,p) \tag{8}$$

第 3 步,利用隐层的输出 b_j、权值 v_j 和阈值 γ_t 计算输出层各单元的输出 L_t,然后通过传递函数计算输出层各单元的实际输出 C_t。

$$L_t=\sum_{j=1}^{p}v_{jt}b_j-\gamma_t\quad(t=1,2,\cdots,q) \tag{9}$$

$$C_t=f(L_t)\qquad(t=1,2,\cdots,q) \tag{10}$$

第 4 步,利用网络目标向量 $T_k=(y_1{}^k,y_2{}^k,\cdots,y_q{}^k)$ 与网络的实际输出 C_t,计算输出层的各单元训练误差 $d_t{}^k$。

$$d_t{}^k=(y_t{}^k-C_t)\cdot C_t(1-C_t)\quad(t=1,2,\cdots,q) \tag{11}$$

第 5 步,利用连接权值 v_{jt}、输出层的训练误差 d_t 和中间层的输出 b_j 计算隐层各单元的训练误差 $e_j{}^k$。

$$e_j^k=\Big[\sum_{t=1}^{q}d_t\cdot v_{jt}\Big]\cdot b_j(1-b_j) \tag{12}$$

第 6 步,利用输出层各单元的训练误差 $d_t{}^k$ 与隐层各单元的输出 b_j 来修正连接权值 v_{jt} 阈值 γ_t。

$$v_{jt}(N+1)=v_{jt}(N)+\alpha\cdot d_t{}^k\cdot b_j \quad (13)$$

$$\gamma_t(N+1)=\gamma_t(N)+\alpha\cdot d_t{}^k \quad (14)$$

$t=1,2,\cdots,q$; $j=1,2,\cdots,p$; $0<\alpha<1$。

第 7 步,利用隐层各单元的训练误差 $e_j{}^k$,输入层各单元的输入 x_k 来修正连接权值 w_{ij} 和阈值 θ_j

$$w_{ij}(N+1)=w_{ij}(N)+\beta\cdot e_j{}^k\cdot x_i{}^k \quad (15)$$

$$\theta_j(N+1)=\theta_j(N)+\beta\cdot e_j{}^k \quad (16)$$

$i=1,2,\cdots,n$; $j=1,2,\cdots,p$; $0<\beta<1$。

第 8 步,通常利用误差测度准则平方误差最小,即能量函数 $E=\frac{1}{2}\sum_{i=1}^{q}(y_i-C_i)$ 来确定研究的网络学习是否满足精度要求。若能量函数 E 小于预先设定的一个极小值,则表明网络收敛,训练达到精度要求;反之,则需要对参数进行调整与选择,或重新分析输入因子与输出因子的相关性。

2 结果与分析

2.1 确定需水量主要影响因子

选取瓜州县 1988～2007 年总需水量序列资料和区域社会经济资料作为基础数据(部分统计结果见表 1),将总需水量作为主成分回归分析的因变量,国内生产总值、工业产值、农业总产值、人口总数、耕地面积、播种面积和大牲口数量 7 个因子作为自变量,应用 SPSS 分析软件的主成分分析功能,求得相关系数矩阵 R,结果见表2。从表 2 可知,7 个因子存在不同程度的相关性。其中,国内生产总值与农业总产值的相关系数为0.981,国内生产总值与人口总数、人口总数与农业生产总值的相关系数均为 0.965。由此可提取出彼此独立的变量,筛选有代表性的因子构造 BP 神经网络的输入样本。

根据表 2 的相关系数矩阵和主成分分析步骤 b～d,得到所筛选 7 个因子的相关系数矩阵 R 的特征值和贡献率计算结果,列于表 3。从表 3 可以看出,第 1 个因子的贡献率为 80.987%,前 2 个因子的累计贡献率达到 93.193%,由此表明这 2 个因子基本上代表了原来 7 个因子 93.193%的信息。由于通常情况下,因子累计贡献率达到 90%以上时就可以反映相关因子的影响,因此,可确定所选 7 个因子中的前 2 个因子代替原变量。

由式(6)计算表 3 中前 2 个因子的荷载矩阵,结果见表 4。由表 4 可看出,国内生产总值、工业产值和农业总产值对第一主成分的相关系数都超过了 0.95,相对其他因子贡献最大;大牲口数对第二主成分贡献最大。因此,选用国内生产总值、工业总产值、农业总产值和大牲口数 4 个因子作为主成分,并以此构造 BP 神经网络输入样本。

2.2 改进的 BP 神经网络

2.2.1 模型的建立

建立 BP 神经网络需水预测模型,首先要确定输入层、隐含层和输出层的节点数。输入层的

表 1 瓜州县典型年需水量及社会经济统计表

Table 1 Water demand & social economy statistics of Guazhou, Gansu

年份	需水量 /10^8 m^3	国内生产总值 /10^4 元	人口	工业产值 /10^4 元	耕地面积 /10^4 亩	大牲口 /头	播种面积 /10^4 亩	农业总产值 /10^4 元
1988	4.12	8452.69	69100	1219.23	22.17	32390.00	24.72	7290.64
1995	4.21	38713.00	82968	8097.00	22.08	18768.00	23.24	25946.24
2000	5.50	66214.00	99615	13908.00	22.76	17870.00	24.53	31264.34
2007	5.44	175165.00	117100	35310.00	35.94	15598.00	36.53	61154.15

表 2 各自变量的相关系数矩阵

Table 2 Correlation coefficients matrix of variables

变量	国内生产总值	人口总数	工业总产值	耕地面积	大牲口数	播种面积	农业总产值
国内生产总值	1.000	0.965	0.949	0.874	−0.521	0.824	0.981
人口总数	0.965	1.000	0.878	0.778	−0.543	0.712	0.965
工业总产值	0.949	0.878	1.000	0.860	−0.629	0.808	0.936
耕地面积	0.874	0.778	0.860	1.000	−0.381	0.975	0.810
大牲口数	−0.521	−0.543	−0.629	−0.381	1.000	−0.275	−0.582
播种面积	0.824	0.712	0.808	0.975	−0.275	1.000	0.743
农业总产值	0.981	0.965	0.936	0.810	−0.582	0.743	1.000

表 3　主成分特征值和贡献率

Table 3　Eigenvalues and contribution rates of principle constituents

因子	1	2	3	4	5	6	7
特征值	5.669	0.854	0.352	0.081	0.021	0.016	0.006
贡献率/%	80.987	12.207	5.029	1.154	0.305	0.230	0.088
累计贡献率/%	80.987	93.193	98.223	99.376	99.682	99.912	100.000

表 4　主要因子的荷载矩阵

Table 4　Load matrix of main indices

变量	国内生产总值	人口总数	工业总产值	耕地面积	大牲口数	播种面积	农业总产值
因子 1	0.985	0.939	0.969	0.916	−0.596	0.865	0.966
因子 2	0.029	−0.078	−0.072	0.309	0.749	0.421	−0.089

节点数为影响水资源需求量的因子数。通过主成分分析法,与瓜州县年总需水量有显著关系的主要影响因子包括:国内生产总值、工业总产值、农业总产值、大牲口数,由此可得出输入层为 4 个节点。各因子 1988～2007 年的统计结果见图 1。预测对象为年总需水量,因此输出层为 1 个节点。隐含层节点数的确定采用试算法,选取训练与测试结果误差最小所对应的隐含层神经元数作为最后确定的隐含层节点数。

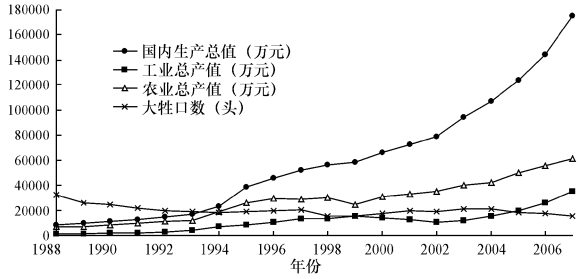

图 1　四个主成分因子年际变化序列

Fig.1　Variation of annual distribution of the four main indices

2.2.2　模型求解

通常情况下,为了加快训练过程中的收敛速度,需对原始数据进行归一化处理,具体计算公式为:

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{17}$$

式中:$x_{\max}$ 和 $x_{\min}$ 分别为系列中的最大值和最小值。根据式(17)可知 x'_i 的范围在[0,1]之内。

从样本数据中选取 1988～2004 年样本进行网络训练,2005～2007 年的已知样本对网络进行检验。同时,选取适当初始学习率 $\eta=0.9$,运算次数 10 000,允许精度 $E=0.05$,训练函数采用 Polak-Ribiere 共轭梯度法的 Purelin 函数。经试算,在隐含层神经元数为 8 时,训练效果最佳,预测值与实际值对比结果见图 2。采用此次训练结果进行样本检验,检验结果及检验误差见表 5。

从图 2 可知,各年需水量预测值与实际值拟合精度较高,曲线基本重合。表 5 中检验结果表明,绝对误差小于 $\pm0.05\times10^9$ m^3,在允许误差范围内。显然,这种基于主成分分析法构造 BP 神经网络的输入矩阵的训练过程误差较小,取得的预测结果也令人满意。

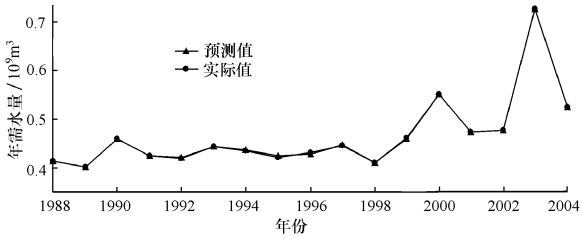

图 2　BP 神经网络预测结果

Fig.2　Prediction outcome of BP neutral networks

表 5　总需水量预测检验结果

Table 5　Testing outcome of total water demand prediction

年份	期望输出 /10^9 m^3	实际输出 /10^9 m^3	绝对误差	允许误差
2005	0.524	0.514	0.010	0.0524
2006	0.531	0.569	−0.038	0.0531
2007	0.544	0.575	−0.031	0.0544

3　结论

a.通过主成分分析,确定了影响瓜州县需水量的主要因子,包括国内生产总值、工业总产值、农业总产值、大牲口数。这 4 个因子的累计贡献率达到 93.193%。

b.由于神经网络模型具有局部逼近的特征

和较强的非线性映射能力,因此它能够较好地模拟具有较强非线性变化特点的需水预测问题。基于主成分分析的 BP 神经网络简化了网络输入样本,消除了网络输入之间的相关性,降低了网络的输入层数,改善了程序执行效率,从整体上提高了网络的性能。最终取得了良好的预测结果。

[参 考 文 献]

[1] 王浩.我国水资源合理配置的现状和未来[J].水利水电技术,2006,37(2):7—14.

[2] 郑度.中国西北干旱区土地退化与生态建设问题[J].自然杂志,2007,29(1):7—12.

[3] 朱丹果,上官智锋.西北地区水资源可持续发展的障碍及解决策略[J].环境科学与管理,2007,32(6):51—53.

[4] 沈福新,耿雷华,曹霞莉,等.中国水资源长期需求展望[J].水科学进展,2005,16(4):522—525.

[5] 吕智,陈文贵,丁宏伟.干旱区内陆盆地水资源的合理配置——以甘肃省高台县为例[J].水资源保护,2005,21(6):45—48.

[6] 王浩,游进军.水资源合理配置研究历程与进展[J].水利学报,2008,39(10):1168—1175.

[7] 和刚,吴泽宁,胡彩虹.基于定额定量分析的工业需水预测模型[J].水资源与水工程学报,2008,19(2):60—63.

[8] 刘俊萍,畅明琦.径向基函数神经网络需水预测研究[J]. 水文,2007,27(5):12—16.

[9] 甘治国,蒋云钟,鲁帆,等.北京市水资源配置模拟模型研究[J]. 水利学报,2008,39(1):91—95.

[10] 张妍,尚金城,于相毅. 主成分-聚类复合模型在水环境管理中的应用——以松花江吉林段为例[J]. 水科学进展,2005,16(4):592—595.

[11] 苑希民,李鸿雁,刘树坤,等.神经网络和遗传算法在水科学领域的应用[M].北京:中国水利水电出版社,2002.

[12] 高隽.人工神经网络原理及仿真实例[M].北京:机械工业出版社,2007.

[13] 凌和良,桂发亮,楼明珠. BP 神经网络算法在需水预测与评价中的应用[J].数学的实践与认识,2007,37(22):42—47.

[14] 张雪飞,郭秀锐,程水源,等. BP 神经网络法预测唐山市需水量[J].安全与环境学报,2005,5(5):95—98.

[15] 董长虹.Matlab 神经网络与应用[M].北京:国防工业出版社,2007.

Water demand forecast model of BP neutral networks based on principle component analysis

LONG Xun-jian[1,3], QIAN Ju[2], LIANG Chuan[1]

1. *College of Water Resource and Hydropower Institute, Sichuan University, Chengdu* 610065, *China*;

2. *College of Resources and Environment, Lanzhou University, Lanzhou* 730000, *China*;

3. *Chongqing Water Resources and Electric Engineering College, Chongqing* 402160, *China*

Abstract: The principle component analysis is based on few extraneous variables that have controlled over all variables and can describe the correlativity among multiple variables. Taking the water demand data from 1988 to 2007 of Guazhou County of Gansu Province for example, this paper analyzes the main factors that influences the water resource quantity based on the principle component analysis method. According to these main factors, the input samples of BP neutral network are definite. Thereby, the BP neutral networks could be trained to predict. The results show that the gross domestic product (GDP), total industrial output value, total output value of agriculture and animals population are the primary indexes that touch to the water resource demand. The corresponding prediction modeling outcome shows that the simulated experiment is quite fit for the practical situation and the absolute error of prediction is lower than $\pm 0.05\times10^9$ m^3.

Key words: water demand prediction; principle component analysis; BP neutral networks