基于主成分分析的 BP 神经网络在延安市需水预测中的应用

舒媛媛，周维博，刘 雷，董起广，李云排
（长安大学 环境科学与工程学院，陕西 西安 710054）

摘 要：利用延安市 1990～2010 年的需水量数据，采用主成分分析法对影响水资源需求量的 10 个因子进行了分析。结果表明：GDP、降雨量、居民生活用水量及生态环境用水量 4 个因子为影响需水量的主要因子，将此作为主要因子构造 BP 神经网络的输入样本，建立延安市需水量预测模型。模拟结果与实际值相吻合，并利用模型对 2015 年需水量进行了预测。

关键词：需水预测；主成分分析法；BP 神经网络；延安市
中图分类号：TP389.1；TU991.31 **文献标识码**：A **文章编号**：1672-643X(2012)06-0172-04

Application of BP neutral networks to water demand forcast of Yan'an City based on principle component analysis

SHU Yuanyuan, ZHOU Weibo, LIU Lei, DONG Qiguang, LI Yunpai
(*School of Environment Science and Technology, Chang'an University, xi'an* 710054, *China*)

Abstract: Taking the water demand data from 1990 to 2010 of Yan'an City of Shaanxi Province, this paper analyzes the main factors that influences the water resource quantity based on the principle component analysis method. The results show that GDP, rainfall, residents living water consumption and ecological environment water consumption are the primary indexes that affact the water demand. Taking the main indexes as input samples, the paper set up forcast model of water demand. The simulation results are consistent with the actual value, and use the model to predict water demand in 2015.

Key words: water demand prediction; method of principle component analysis; BP neutral networks; Yan'an City

0 引 言

为应对社会经济日益增长的用水需求，许多国家的规划部门把需水量预测作为计划工作的手段，以期达到宏观规划、解决水资源供需矛盾的目的[1]。延安市近年来经济持续快速发展，2010 年经济总量位居陕西省第四。然而，地处陕北黄土高原的延安市是一个水资源十分贫乏的地区，长期以来干旱缺水，水是延安市经济社会发展的重要制约因素。现作为全国节水型社会建设试点城市之一，在“十二五”规划中提出经济建设快速发展，产业结构战略调整，并对水资源需求、开发和利用提出了新的要求[2]。因此，为满足用水量的高速增长以及确保水资源的高效利用成为延安市经济、社会可持续发展的重要保证[3]。对延安市需水量进行合理预测可为其经济社会发展规划提供科学依据。

国内外关于需水量预测研究历史悠久[4]。近年来水资源与经济、社会及环境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[5]，常用需水量预测方法有定额法、多元线性回归分析法、灰色模型法以及神经网络法等[6]。神经网络和灰色预测模型是近年来紧密结合计算机技术新发展起来的预测方法，都属系统方法范畴，均是较为准确的方法。其中应用较广泛的是 BP 神经网络模型，其广泛的应用显示出活跃的生命力，具有不可比拟的准确性和优越性[7]。选择合理的预测方法，不仅可以增加水资源配置研究工作的实际操作性，而且更符合社会经济发展趋势的结果。但无论

收稿日期：2012-08-10； **修回日期**：2012-09-08
作者简介：舒媛媛(1986-)，女，湖北襄阳人，硕士研究生，主要从事给排水及水资源开发利用研究。
通讯作者：周维博(1956-)，男，陕西乾县人，教授，博士生导师，从事水资源与水环境及节水灌溉方面的教学与研究工作。

采用何种预测方法,均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;因子选择过多,会使神经网络训练复杂化,可能陷入局部优化问题,难以得到全局优化解[8]。本文根据影响延安市需水量的 10 个因素:人口、GDP、万元 GDP 用水量、降雨量、地表水供水量、地下水供水量、居民生活用水量、工业用水量、农业用水量以及生态环境用水量,首先应用主成分分析法确定影响总需求量的主要因子,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测,以提高模型的学习与泛化能力。

1　研究方法

1.1　主成分分析法

主成分分析法是数学上对数据降维处理的一种方法,将原来众多的具有一定相关性的指标 $X_1, X_2, \cdots, X_p$(比如 p 个指标),重新组合成一组较少个数的互不相关的综合指标来代替原来指标。此种方法以最少的信息丢失为前提,将众多的原有变量综合成较少几个综合指标,这将大大减少分析过程中的计算工作量。

1.2　BP 神经网络

人工神经网络(Artificial Neutral Network)是一种模仿人脑结构及其功能的信息处理系统[9]。BP(Back - Propagation)神经网络,是人工神经网络中最具有代表性和广泛应用的一种。BP 算法是一种简单的最速下降静态寻优算法,用于网络权值和阈值的学习。但是,BP 算法存在以下主要缺点:①采用残差平方和最小作为目标函数,会过分地突出极大值的影响;②难以解决具有局部最小的复杂欺骗性问题,尤其当求取一般非凸目标函数的全局最优解时;③学习算法的收敛速度很慢[10]。针对可能陷入局部优化,难以得到全局优化解这一问题,本文采用通过主成分分析法取得的有显著关系的主要影响因子作为输入层的节点。

BP 神经网络以三层最为常用,也最为成熟。理论上已证明,一个三层网络可以以任意精度逼近任何给定的连续函数。因此,本文采用三层网络模型。三层神经网络技术应用的关键是构造输入层、隐含层和输出层的节点数,以及各层之间连接权值和节点阈值的初始化,其中输入和输出层节点数一般是取输入和输出向量的维数,而连接权值和阈值一般的初始化方法是取一定范围内较小的随机数,对于隐含层节点数的确定目前尚无标准,一般采用试算法。

2　结果与预测

2.1　确定需水量主要影响因子

选取延安市 1990 - 2010 年的需水量、人口(X_1)、GDP(X_2)、万元 GDP 用水量(X_3)、降雨量(X_4)、地表水供水量(X_5)、地下水供水量(X_6)、居民生活用水量(X_7)、工业用水量(X_8)、农业用水量(X_9)以及生态环境用水量(X_{10})的序列年资料作为基础数据,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,$X_1 \sim X_{10}$ 个因子作为自变量,应用 SPSS 分析软件的主成分分析功能,得到相关系数矩阵 R 以及特征根数值衰减折线图,结果见表 2 及图 1。从表 2 可知,10 个因子存在不同程度的相关性,例如 GDP(X_2)与总人口(X_1)的相关系数为 0.898,GDP(X_2)与居民生活用水量(X_7)的相关系数为 0.786。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 神经网络的输人样本。根据表 2 的相关系数矩阵和主成分分析步骤,得到所筛选 10 个因子的相关系数矩阵 R 的特征值 λ 和贡献率计算结果,列于表 3。根据 λ 值决定主成分数目的准则为:

(1)只取 $\lambda > 1$ 的特征根对应的主成分;

(2)累计百分贡献率达到 80% ~85% 以上的 λ 值对应的主成分;

(3)根据特征根变化的突变点决定主成分的数量。

从表 3 可以看出,前两个因子的特征值均大于 1,第 1 个因子的贡献率为 73.07%,前 2 个因子的累计贡献率达到 86.66%,由此表明这 2 个因子基本上代表了原来 10 个因子 86.66% 的信息。从图 1 可以看出,第 2 个 λ 值是一个明显的折点,这表明选取的主成分数目应有 $P \leqslant 2$。从以上信息中可以得出结论,可以确定所选 10 个因子中的前 2 个因子代替原变量。前 2 个因子的荷载矩阵结果见表 4。由表 4 可看出,GDP、居民生活用水量及生态环境用水量对第一主成分的相关系数均不低于 0.987,相对其他因子贡献最大,其它 9 个因子的相关系数均低于 0.5。

另外,从延安市不同水平年的经济社会发展情况来看,延安市生产总值的增长率呈逐年上升趋势,用水速度加快,并且伴随着城镇化率的提高带来居民用水量的增加。降雨量在不同频率年里也在很大程度上影响需水量比如农林牧渔业。目前延安市正在进一步推广棚栽农业滴灌节水技术,在未来水平

年里农业用水将不是影响需水量的主要因素。同时,延安市坚持不懈地抓以退耕还林为主的生态环境建设,退还挤占河湖生态环境用水的水源等措施,努力改善人居环境,实现生态、经济、社会三大效益的有机结合,生态用水在不断增加[12]。

表 1 延安市需水量及影响因子统计表[11] 亿 m^3, 万人, 亿元, m^3/万元, mm, 万 m^3

年份	需水量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1990	1.11	178.69	18.03	616	649.3	0.86	0.23	0.23	0.21	0.65	33
1995	1.37	189.09	48.49	283	336.9	1.04	0.29	0.27	0.33	0.74	58
2000	1.77	196.64	130.6	136	419.6	1.24	0.46	0.29	0.64	0.79	159
2005	1.77	211.85	394.7	48	459.3	1.04	0.67	0.37	0.47	0.82	45
2010	2.42	218.87	885.4	31	506.4	1.49	0.90	0.47	0.75	1.05	428

表 2 各自变量的相关矩阵

年份	$Z(X_1)$	$Z(X_2)$	$Z(X_3)$	$Z(X_4)$	$Z(X_5)$	$Z(X_6)$	$Z(X_7)$	$Z(X_8)$	$Z(X_9)$	$Z(X_{10})$
$Z(X_1)$	1.000	0.898	−0.207	−0.171	0.844	0.869	0.940	0.917	0.467	0.829
$Z(X_2)$	0.898	1.000	−0.481	−0.023	0.946	0.977	0.986	0.976	0.633	0.970
$Z(X_3)$	−0.207	−0.481	1.000	−0.188	−0.374	−0.393	−0.396	−0.341	−0.181	−0.440
$Z(X_4)$	−0.171	−0.023	−0.188	1.000	0.071	0.032	−0.069	−0.131	0.399	0.043
$Z(X_5)$	0.844	0.946	−0.374	0.071	1.000	0.945	0.949	0.932	0.812	0.969
$Z(X_6)$	0.869	0.977	−0.393	0.032	0.945	1.000	0.970	0.968	0.714	0.983
$Z(X_7)$	0.940	0.786	−0.396	−0.069	0.949	0.970	1.000	0.974	0.638	0.956
$Z(X_8)$	0.917	0.976	−0.341	−0.131	0.932	0.968	0.974	1.000	0.591	0.962
$Z(X_9)$	0.467	0.633	−0.181	0.399	0.812	0.714	0.638	0.591	1.000	0.749
$Z(X_{10})$	0.829	0.970	−0.440	0.043	0.969	0.983	0.956	0.962	0.749	1.000

表 3 主成份特征值和贡献率 %

成份	初始特征值			提取平方和载入		
	特征值	贡献率	累计贡献率	特征值	贡献率	累计贡献率
1	7.308	73.077	73.077	7.308	73.077	73.077
2	1.358	13.583	86.661	1.358	13.583	86.661
3	0.882	8.821	95.482			
4	0.308	3.079	98.561			
5	0.093	0.926	99.487			
6	0.032	0.323	99.810			
7	0.011	0.114	99.925			
8	0.005	0.051	99.975			
9	0.002	0.025	100.000			
10	0.000	0.000	100.000			

表 4 主要因子的载荷矩阵

变量	成份		变量	成份	
	1	2		1	2
$Z(X_1)$	0.893	−0.296	$Z(X_6)$	0.986	0.000
$Z(X_2)$	0.989	−0.056	$Z(X_7)$	0.987	−0.116
$Z(X_3)$	−0.429	−0.313	$Z(X_8)$	0.972	−0.184
$Z(X_4)$	0.023	0.944	$Z(X_9)$	0.725	0.472
$Z(X_5)$	0.978	0.073	$Z(X_{10})$	0.987	0.044

综合上述分析,选用 GDP、降雨量、居民生活用水量及生态环境用水量 4 项作为构造 BP 神经网络输入样本的因子是合理的。

2.2 BP 神经网络需水预测

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。输入层的节点数为主要影响水资源需求量的因子数,包括 GDP、降雨量、居民生活用水量及生态环境用水量,由此可得出输入层为 4 个节点。预测对象为年需水量,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。

2.2.2 *模型求解* 为了加快训练过程中的收敛速度,需要对原始数据进行归一化处理,即将所有数据转化至[0,1]之间。具体计算公式为:

$$\chi_i' = \frac{\chi_i - \chi_{\min}}{\chi_{\max} - \chi_{\min}} \tag{1}$$

式中:$\chi_{\max}$ 和 $\chi_{\min}$ 分别为系列中的最大值和最小值[13]。并且程序运行后的预测结果均需进行还原,即:

$$\chi_i = \chi_i'(\chi_{max} - \chi_{min}) + \chi_{min} \quad (2)$$

式中:χ_{max} 和 χ_{min} 分别为训练样本中最大值和最小值。

从样本数据中选取 1990 – 2005 年样本进行网络训练,2006 – 2010 年的已知样本对网络进行检验。同时,选取适当初始学习率 $\eta = 0.1$,运算次数 50000 次,允许精度 $E = 0.001$,训练函数采用 Sigmoid 激发函数。经试算,在隐含层神经元数为 6 时,训练效果最佳,拟合值与实际值对比结果见图 2。采用此训练结果进行样本检验,以 2006 年 – 2010 年作为检验样本,检验结果及检验误差见表 5。

从图 2 可知,各年需水量检验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 ±0.05 亿 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。由于训练过的网络已"模拟"并"记忆"了输入变量和输出变量之间的"函数"关系,因而可以用它进行延安市需水量的预测,延安市 2011 – 2015 年需水量预测结果见表 6。实际值与预测值变化趋势见图 3。

表 5　需水量预测检验结果　　亿 m^3

年份	期望输出	实际输出	绝对误差	绝对误差	允许误差
2006	2.200	2.239	–0.038	0.017	0.05
2007	2.350	2.363	–0.014	0.006	0.05
2008	2.385	2.343	0.042	0.018	0.05
2009	2.403	2.362	0.041	0.017	0.05
2010	2.423	2.408	0.015	0.006	0.05

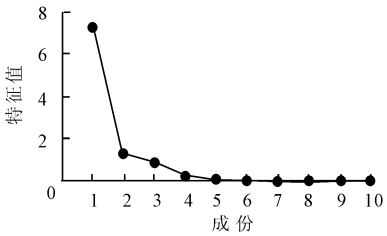

图 1　特征根数值衰减折线图

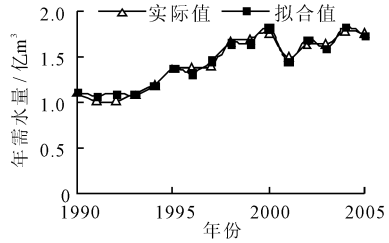

图 2　BP 神经网络拟合结果

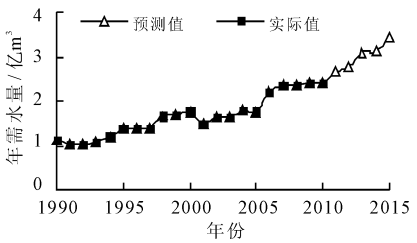

图 3　BP 神经网络预测结果

表 6　需水量预测结果　　亿 m^3

年份	2011	2012	2013	2014	2015
预测值	2.67	2.78	3.11	3.13	3.44

3　结　语

(1)通过主成分分析,确定了影响延安市需水量的主要因子,包括 GDP、降雨量、居民生活用水量及生态环境用水量。这四个因子的累计贡献率达到 86.66%。

(2)应用 BP 神经网络在延安市需水预测模型的实例分析可以看出,BP 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。而基于主成分分析的 BP 神经网络简化了网络输入样本,很大程度上消除了网络输入之间的相关性,改善了程序执行效率,从整体上提高了网络的性能,最终取得良好的预测结果。

参考文献:

[1] Peter H Gleick. Water in crisis paths to sustainable water use[J]. Ecological Applications,1998, 8(3):571 – 579.

[2] 周维博,刘晓宇,田 倩,等. 延安市节水型社会建设"十二五"规划[R]. 2011,8.

[3] 殷 辉. 延安市水资源可持续利用研究[J]. 陕西水利,2008(6):75 – 76.

[4] 姜 琼. 基于可持续利用水量的需求预测方法研究——以延安市为例[D]. 河海大学,2006.

[5] 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.

[6] 张雅君,刘全胜. 需水量预测方法的评估与选择[J]. 给水排水,2003,29(4):26 – 29.

[7] 魏海坤. 神经网络结构设计的理论与方法[M]. 北京:国防工业出版社,2005:25 – 36.

[8] 龙训建,钱 鞠,梁 川. 基于主成分分析的 BP 神经网络及其在需水预测中的应用[J]. 成都理工大学学报,2010,37(2):206 – 210.

[9] 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.

[10] 麻凤海,杨 维,杨 帆,等. 应用改进 BP 神经网络进行用水量预测[J]. 辽宁工程技术大学学报,2004,23(2):191 – 193.

[11] 张 欣,高宏周,张海军,等. 陕西省各市区"十二五"用水总量控制指标及水量分配方案[R]. 2011,9.

[12] 王 浩,于福亮,王 研,等. 延安市水资源综合规划总报告[R]. 2008,5.

[13] 阎立华,吕科峰. 城市日用水量预测的神经网络方法[J]. 沈阳建筑工程学院学报,2004,20(2):136 – 138.