第 16 卷 第 2 期
2005 年 3 月

水 科 学 进 展
ADVANCES IN WATER SCIENCE

Vol. 16 ,No. 2
Mar. , 2005

基于熵值法的 BP 网络输入变量加权分层方法研究

张小峰, 袁 晶

(武汉大学水资源与水电工程科学国家重点实验室, 湖北 武汉 430072)

摘要: 当 BP 网络模型的输入变量包括多个类别时, 如果其中几类变量的个数远多于其它类别的变量, 变量多的这几类会削弱其它类变量对输出变量的影响, 导致模型预报误差增大。提出 BP 网络输入变量加权分层的改进方法。根据熵值法模型对每个类别包含的所有变量按其重要程度加权平均, 得到代表各类的综合影响指标, 将这些综合影响指标作为 BP 网络模型的输入变量得到模型预报结果。改进后的模型更全面合理地考虑了各类输入变量的变化对输出变量的影响, 发展了神经网络的应用理论。实例计算表明, 模型预报精度得到明显提高。

关 键 词: BP 网络; 神经网络; 预报方法; 河床变形; 加权分析; 信息熵

中图分类号: TV147 **文献标识码**: A **文章编号**: 1001-6791(2005)02-0263-05

神经网络的理论和技术已被广泛应用于许多研究领域。多层前向神经网络及其相应的 BP 算法是当今应用最广泛的一种神经网络。该方法对非线性函数有较强的逼近能力, 且学习算法简单, 为未知非线性系统辨识与控制提供了一个新的有效方法。但现行 BP 网络应用于大样本集和复杂模式的问题时, 特别当输入变量包括多个类别且其中几类变量的个数远多于其它类的变量个数时, 个数多的这几类对输出变量构成最主要的影响因素而削弱其它类变量对输出变量的影响, 导致模型不能全面考虑各类输入变量对输出变量的影响, 难以对输出变量进行较为准确的预报。

熵是源于热力学的概念, 熵的本质是系统状态的不确定性。目前熵值法已被广泛应用于确定各指标的客观权重。王[illegible]等[1]引入熵信息的概念, 确定各指标的客观权重, 对诸投标方案进行综合评定, 最终实现投标方案的排序和优选。魏隽等[2]用熵值法对软件产业各专业领域进行发展次序优先排序。本文在 BP 网络的基础上, 提出了一种基于熵值法的输入变量加权分层处理方法。改进后的方法以输入变量的类别作为划分单位, 将各类包含的所有变量按其重要程度加权平均, 自动获取代表各类变量的综合影响指标, 使模型能够更全面合理地考虑各类输入变量的变化对输出变量的影响, 提高模型的预报精度。

1 输入变量加权分层方法的思想

1985 年 Rumelhart 和 McCelland 等提出的误差反向传播(Back Propagation)算法, 是目前人工神经网络理论中最重要的一种学习算法, 其学习过程是由信息正向传播和误差反向传播两个反复交替的过程所组成[3,4]。在信息正向传播过程中, 输入信息经隐含单元逐层处理转向输出层, 可简单地表征为

$$y_j^k = f_j^k\left(\sum_{i=1}^{n_{k-1}} W_{ij}^{k-1} y_i^{(k-1)} - \theta_j^k\right) \qquad j = 1,2,\cdots,n_k; \quad k = 1,2,\cdots,m \tag{1}$$

式中 $W_{ij}^{(k-1)}$ 为第 $(k-1)$ 层中第 i 个神经元到第 j 个神经元的连接权因子; θ_j^k 为该神经元的阈值。

如果输出层不能得到所期望的输出, 则转入反向传播过程。将实际值与网络输出之间的误差沿原来的连接通路返回, 通过修改各层神经元的连接权重使误差减少, 然后再转入正向传播过程。如此反复计算, 直至误差

收稿日期: 2003-11-24; **修订日期**: 2004-03-30

基金项目: 国家重点基础研究发展计划 (973) 资助项目 (2003CB415203); 国家自然科学基金资助项目 (50279035)

作者简介: 张小峰(1962 -), 男, 浙江嵊洲人, 教授, 主要从事水力学及河流动力学研究。

E-mail: yj790902 @sina.com

小于设定值为止。该过程可描述为

$$E = \frac{1}{2}\sum_{p=1}^{p}\sum_{k=1}^{n_3}(T_{pk} - O_{pk})^2 \leq \varepsilon \tag{2}$$

式中 T_{pk}、O_{pk}分别表示输入训练样本为 P 时输出节点 K 的计算输出和期望输出；ε 为允许的最大误差。

为避免输入向量物理意义和单位的不同对 BP 网络模型的影响，需对输入向量作标准化处理：

$$\overline{X_i} = \frac{X_i - X_{i\min}}{X_{i\max} - X_{i\min}} d_1 + d_2 \tag{3}$$

式中 $X_{i\min}$、$X_{i\max}$为输入样本的 i 个节点中的最小值和最大值；d_1、d_2 为网络标准参数；$\overline{X_i}$ 表示标准化的输入变量。由于标准化处理后的输入变量没有了量纲的区别，每个变量对输出变量的影响所占比例均等，当各类变量的个数差别较大，变量多的几类将削弱其它类变量对输出结果的影响，导致模型预报误差增大。

为解决这一问题，按照输入变量的类别，运用熵值法对各类变量进行加权平均，自动获取代表各类变量的综合影响指标：

$$y_{i_1}^{(k-1)} = \sum_{n=1}^{l} w_{i_1 n} y_{i_1 n}^{(k-1)} \tag{4}$$

式中 $w_{i_1 n}$为第 i_1 类影响因子中第 n 个变量的权重，由熵值法计算确定，满足 $\sum_{n=1}^{l} w_{i_1 n} = 1$；$l$ 是第 i_1 类影响因子的个数。

得到代表各类的综合影响指标后，将这些综合影响指标作为 BP 网络模型的输入变量代入式(1)进行训练，得出各类综合影响指标与网络各层及输出变量之间的所有连接权因子和阈值，用于计算和预测。

2 熵值法

运用综合评价方法对众多样本进行权重分析的方法很多，一般有层次分析法、德尔菲法、模糊聚类法、熵值法[5]，评价的过程涉及 4 项关键工作：①影响因素分析及评价指标体系的设计；②样本数据的获取；③各指标重要程度的辨识；④综合评价方法。目前对指标权重的确定存在着不少主观随意性，所列出的 4 种方法都有着各自的实用范围，一般如果样本数据不全，应采用 AHP 法和德尔菲法；若样本中含有大量的模糊数据，且同一层次指标个数较多，应先采用模糊聚类分析方法，再用 AHP 法分别做各个子类的权重；如果样本数据比较完整，则应采用熵值法，由于本文所考虑问题的数据完整，则采用熵值法对数据进行加权平均。

在信息系统中，信息熵是信息无序度的度量，信息熵越大，信息的无序度越高，其信息的效用值越小；反之，信息熵越小，信息的无序度越小，信息的效用值越大[6~9]。在综合评价中，运用信息熵评价所获信息系统的有序程度及信息的效用值是很自然的，统计物理中熵值函数形式对于信息系统是一致的。对于 m 个样本的 n 个评价指标，其初始数据矩阵 $X = \{X_{IJ}\}_{m\times n}$，由于各指标的量纲、数量级及指标优劣的取向均有很大差异，故需对初始数据做标准化处理：

$$y_{ij} = x_{ij} / \sum_{i=1}^{m} x_{ij} \qquad 0 \leq y_{ij} \leq 1 \tag{5}$$

由此得数据的标准化矩阵

$$Y = \{y_{ij}\}_{m\times n} \tag{6}$$

第 j 项指标的信息熵值[5]表示为

$$e_j = -k\sum_{i=1}^{m} y_{ij} \ln y_{ij} \tag{7}$$

式中 常数 k 与系统的样本数 m 有关，对于一个信息完全无序的系统，有序度为零，其熵值最大，$e = 1$，m 个样本处于完全无序分布状态时，$y_{ij} = \frac{1}{m}$，由式(7)可得

$$e = -k\sum_{i=1}^{m} \ln\frac{1}{m} = k\sum_{i=1}^{m}\frac{1}{m}\ln m = k\ln m = 1 \tag{8}$$

于是得到

$$k = (\ln m)^{-1} \qquad 0 < e \leq 1 \tag{9}$$

由于信息熵 e_j 可用来度量 j 项指标信息(指标的数据)的效用价值，当完全无序时，$e_j = 1$，此时，e_j 的信息(也就是 j 指标的数据)对综合评价的效用值为零，因此，某项指标的信息效用价值取决于该指标的信息熵 e_j 与 1 的差值 h_j：

$$h_j = 1 - e_j \tag{10}$$

可见，利用熵值法估算各指标的权重，其本质是利用该指标信息的价值系数来计算，其价值系数越高，对评价的重要性就越大，于是 j 项指标的权重为

$$w_j = h_j / \sum_{j=1}^{n} h_j \tag{11}$$

3　计算实例

本文以长江马家咀水道为例，介绍改进后的方法在该水道断面变形预测中的应用。

3.1　长江马家咀水道概况

马家咀水道位于长江中游上荆江河段，是沙市至陡湖堤河弯之间的过渡段，江面逐渐增宽，上下深槽交错，主流极不稳定，是长江上著名的碍航河段[10]，其河势见图 1。

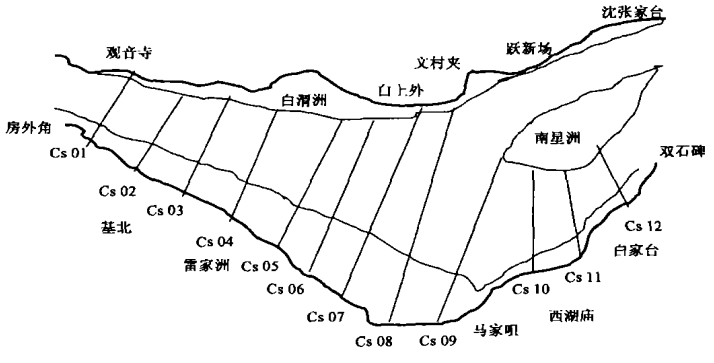

图 1　长江马家咀水道河势图

Fig. 1　Majiazui channel regime of the Yangtze River

3.2　马家咀水道断面变形 BP 网络模型的改进

影响河道断面变形的因素众多，有河段上游的来水、来沙过程以及上、下游河势等。其中反映河段上游来水过程这个类型的变量可以选用年最大流量 Q_{max}，最小流量 Q_{min}、流量出现大于各级大流量的天数等数据，反映河势这个类型的变量可以选用计算断面上、下游河道若干断面地形数据、水流主流位置等。河道上游来水过程和上下游河势均对河道断面变形起着重要作用。但一般情况下，断面数据个数远远多于代表来水过程的数据个数，若按以往方法直接将所有断面数据和其它类型的数据(经标准化处理)作为输入变量代入式(1)建立断面变形的神经网络模型，断面数据因数据量多成为输出变量的主要影响因素，代表上游来水过程的数据因数据个数少对输出变量的影响变得非常微小，模型不能全面考虑断面变形影响因子，导致模型对断面变形预测产生较大误差。针对这一问题，本文利用熵值法原理对模型的输入变量进行加权处理，使模型能够全面合理的考虑各类输入变量对输出变量的影响。

根据马家咀水道的特点，确定该水道断面变形预报模型的影响因素(数据)为：水沙条件包括当年最大流量 Q_{max}，最小流量 Q_{min}、年平均输沙率 $\overline{Q_S}$，最后考虑到流量的大小对河床的冲淤变化有一定影响，模型选取当年流量出现大于 20 000 m^3/s、30 000 m^3/s、40 000 m^3/s 和 50 000 m^3/s 的天数 $T_{Q\geq 20000\ m^3/s}$、$T_{Q\geq 30000\ m^3/s}$、$T_{Q\geq 40000\ m^3/s}$和 $T_{Q\geq 50000\ m^3/s}$。模型的输出变量为计算断面各点的绝对高程。根据各断面不同部位的冲淤变化，断面各点采用不同的间距。如对冲淤变化剧烈的地方，各点间距取得小一些，一般取 50～100 m；对于冲淤变化不大的地方，各点间距可取得大一些，取 200 m。本文以 Cs-04、Cs-05、Cs-06、Cs-07、Cs-08 等 5 个断面作为研究断面，选取 1971 年 12 月、1975 年 5 月、1976 年 1 月、1978 年 12 月、1979 年 1 月、1980 年 12 月、1981 年 1 月、1982 年 12 月、1984 年 5 月、1985 年 2 月、1986 年 11 月、1987 年 4 月、1990 年 1 月、1992 年 4 月、1993 年 4 月、1995 年 4 月、1996 年 11 月、1997 年 4 月、1998 年 4 月和 1999 年 1 月等 20 个测次为模型的训练样本，以 2000 年 5 月为预测样本。

考虑到在输入变量中，断面数据的个数远远多于其它影响因子个数，为了解决因为各类输入变量个数不同给模型预报带来的误差，运用熵值法对各类输入变量按其重要性加权处理。以断面 Cs-05 为例，表 1 列出了熵值法计算后的各变量在其所属类别里所占权重，其中类别 1 表示最小流量 Q_{min}，类别 2 表示最大流量 Q_{max}，类别3 表示当年流量出现大于 20 000 m^3/s、30 000 m^3/s、40 000 m^3/s 和 50 000 m^3/s 的天数 $T_{Q\geq 20\,000\ m^3/s}$、$T_{Q\geq 30\,000\ m^3/s}$、$T_{Q\geq 40\,000\ m^3/s}$和 $T_{Q\geq 50\,000\ m^3/s}$，类别 4 表示年平均输沙率 $\overline{Q_S}$，类别 5、6、7，表示各研究断面上、下游 3 个断面深泓线的位置，类别 8、9、10、11、12、13 表示各研究断面上、下游 6 个断面的地形。模型输入变量的个数经过熵值法加权处理后，对所需预测的断面 Cs-04、Cs-05、Cs-06、Cs-07、Cs-08，模型的输入变量个数分别由 97、96、96、96、97 个减少为 13 个，模型的训练样本和预测样本个数与前述一致，计算过程中对输入变量的读取采用“滚动式”方法，网络隐含层单元数采用“试错法”确定。通过对以上实测资料进行训练和测试，以 Cs-05 断面为例，输入变量数为 13，输出变量数为 12，计算的学习速率 η 取 0.05，动量因子 α 取 0.5，网络误差为 0.033。计算结果如表 2 所示。

为分析神经网络的多输入、多输出结构与多输入、单输出模型结构计算结果之间的差异，本文将 Cs-05 断面的 12 个点分别作为输出变量对模型进行了重新训练和计算。发现在计算过程中，单输出模型的网络误差比多输出的网络误差更小，但由于神经网络模型具有考虑全局最优的功能，两种模型结构的计算结果比较起来基本相同。

表 1 断面 Cs-05 各个影响因子所占权重

Table 1 Weights of each affecting factor of section Cs-05

类别	权重	类别	权重	类别	权重	类别	权重	类别	权重	类别	权重
1	1.000	8	0.041	9	0.170	10	0.049	11	0.025	13	0.022
2	1.000	8	0.079	9	0.070	10	0.036	11	0.013	3	0.049
3	0.191	8	0.087	9	0.002	10	0.035	12	0.000	13	0.002
3	0.143	8	0.065	9	0.025	10	0.030	12	0.174	13	0.064
3	0.255	8	0.050	9	0.002	10	0.024	12	0.153	13	0.036
3	0.412	8	0.036	10	0.017	10	0.022	12	0.112	13	0.055
4	1.000	8	0.020	10	0.123	10	0.002	12	0.060	13	0.033
5	1.000	9	0.004	10	0.134	11	0.000	12	0.051	13	0.111
6	1.000	9	0.051	10	0.143	11	0.244	12	0.095	13	0.066
7	1.000	9	0.019	10	0.111	11	0.234	12	0.124	13	0.143
8	0.001	9	0.060	10	0.066	11	0.116	12	0.158	13	0.035
8	0.231	9	0.050	10	0.028	11	0.037	12	0.050	13	0.134
8	0.128	9	0.112	10	0.029	11	0.056	12	0.019	13	0.029
8	0.131	9	0.158	10	0.033	11	0.093	12	0.004	13	0.123
8	0.088	9	0.153	10	0.064	11	0.075	13	0.030	13	0.028
8	0.044	9	0.124	10	0.055	11	0.107	13	0.024	13	0.017

表 2 Cs-05 断面计算值与实测值比较

Table 2 Comparison between measured and computed value of section Cs-05

断面测点号	实测值/m	改进前计算结果/m	绝对误差/m	相对误差/%	改进后计算结果/m	绝对误差/m	相对误差/%
1	39.940	40.53	0.590	1.48	41.000	1.060	2.65
2	27.849	21.35	-6.499	-23.34	23.300	-4.549	-16.33
3	29.095	22.30	-6.795	-23.35	26.500	-2.595	-8.92
4	26.183	20.62	-5.563	-21.25	20.184	-5.999	-22.91
5	21.460	25.40	3.940	18.36	24.000	2.540	11.84
6	19.540	25.70	6.160	31.53	22.400	2.860	14.64
7	27.740	31.37	3.630	13.09	30.000	2.260	8.15
8	35.631	34.52	-1.111	-3.12	36.794	1.163	3.26
9	38.469	34.50	-3.969	-10.32	39.594	1.125	2.93
10	38.565	40.75	2.185	5.67	40.685	2.120	5.50
11	38.64	40.85	2.210	5.72	40.844	2.204	5.70

比较表 2 的计算结果，可以看出改进后的模型在进行长江马家咀岸线变形的预测时，精度在原模型的基础

上有了很大的提高。例如: 模型对 Cs-05 断面测点 3 的预测误差由 6.795 m 减小为 2.595 m; 测点 6 的预测误差由 6.16 m 减小为 2.86 m。该断面的 12 个测点中, 测点 1、4、8 的预测误差稍有增大, 最大误差在测点 1, 增加幅度为 0.47 m。总体来看, 本文建立的基于熵值法的断面变形神经网络模型, 使 BP 网络模型能够更全面合理地考虑各类输入变量的变化对输出变量的影响, 减小了预测误差, 比较准确地预测河道的断面变形。

4　结　　论

本文以马家咀水道的断面变形 BP 网络模型为例, 通过计算比较, 得出当断面数据的个数远多于其它水沙条件的影响因子个数时, 会给模型的断面预测带来很大误差, 分析其原因是由于当 BP 模型输入变量包括多个类别时, 如果其中几类变量的个数远多于其它类别的变量, 变量多的这几类会削弱其它类变量对输出变量的影响, 导致模型预报误差增大, 基于此, 本文提出了一种基于熵值法的结构自适应神经网络模型, 该模型能够自动将各类影响因子的多个变量按其重要性进行加权处理, 计算结果表明: 改进后的模型能够更全面合理地考虑各类输入变量对输出变量的影响, 提高了模型预报精度, 发展了神经网络的应用理论。

参考文献:

[1] 王　[illegible], 黄文杰. 优选投标项目的决策方法[J]. 现代电力, 2003(2): 86 - 90.

[2] 魏　隽, 吴育华, 秦智辉. 熵权系数法在软件产业发展战略选择中的应用[J]. 河北经贸大学学报, 2002(2): 82 - 87.

[3] 胡铁松, 袁　鹏, 丁　晶. 人工神经网络在水文水资源中的应用[J]. 水科学进展, 1995, 6(3): 76 - 82.

[4] 胡铁松. 神经网络预测与优化[M]. 大连: 大连海事大学出版社, 1994.

[5] 王　靖, 张金锁. 综合评价中确定权重向量的几种方法比较[J]. 河北工业大学学报, 2001(4): 52 - 57.

[6] 时光新, 尹成信. 基于熵的小流域治理效益评价模型及其应用[J]. 水土保持通报, 1999(5): 38 - 40.

[7] 曾　谦, 曾黄麟. 系统参数重要性评价方法[J]. 四川轻化工学院学报, 1999(6): 10 - 13.

[8] 方创琳, 毛汉英. 区域发展规划指标体系建立方法探讨[J]. 地理学报, 1999(5): 411 - 419.

[9] Liang Zongxia, Zheng Mingli. Exponential Integrability, Exponential Decay of Entropy and Logarithmic Sobolev Inequalities of Symmetric Diffusions[J]. MATHEMATIC APPLICATA, 1998, 11(1): 9 - 13.

[10] 张小峰, 谈广鸣, 许全喜, 等. 基于 BP 神经网络的河道断面变形预测模型[J]. 水利学报, 2002(11): 8 - 13.

Weight analysis based on the information entropy research on the inputs of ANN*

ZHANG Xiao-feng, YUAN Jing

(State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China)

Abstract: When the input includes several regimentations and the number of some variables in some regimentations is much more than that of the other regimentations, the former will weak the latter's effect on the output, which leads to the augment about the forecasting error of the model. The entropy based the self-accommodation back propagation neural model is introduced to solve this problem, in which the several variables of each regimentation are weighted according to their importance, so each regimentation is turned into one input respectively in the back-propagation (BP) net work model. The improved model can take the all kinds of inputs into account entirely and reasonably, and boost the forecast accuracy, which develops the applied theory of the neural network.

Key words: BP network; artificial neural network; forecasting method; riverbed-deformation; weight analysis; information entropy

* The study is financially supported by the Basic Research Program of China (2003CB415203) and the National Natural Science Foundation of China (No. 50279035).