基于因子分析的 Hopfield 神经网络在水质评价的应用

卢文喜，初海波，王喜华，龚 磊

（吉林大学 环境与资源学院，吉林 长春 130021）

摘 要：针对 Hopfield 神经网络的过度拟合问题，在因子分析的基础上，结合 Hopfield 神经网络模型提出了因子分析—Hopfield 神经网络模型。以东辽河为例，采用因子分析法确定 7 个水质评价因子，再建立 5×7 的 Hopfield 神经网络进行水质综合评价，并与单一的 Hopfield 网络和传统的内梅罗指数法的结果进行了比较。结果表明，因子分析—Hopfield 神经网络明显优于单一的 Hopfield 神经网络，不仅在一定程度上弥补了因子分析在实际应用中没有实现水质分级的缺陷，而且有效地降低了 Hopfield 神经网络的过度拟和的程度，评价结果更为科学合理，为水质综合评价提供了一种新的方法，具有极好的应用前景。

关键词：因子分析；Hopfield 神经网络；水质评价；东辽河

文献标识码：A **文章编号**：1000-288X(2012)01-0197-04 **中图分类号**：X824

Application of Hopfield Neural Network Based on Factor Analysis to Water Quality Evaluation

LU Wen-xi, CHU Hai-bo, WANG Xi-hua, GONG Lei

(*College of Environment and Resources, Jilin University, Changchun, Jilin* 130021, *China*)

Abstract: To solve the over-fitting problem of Hopfield neural network, Hopfield neural network model with factor analysis was proposed based on factor analysis combined with Hopfield neural network model. Taking Dongliao River for an example, the model determined seven water quality evaluation factors using factor analysis method, created a 5×7 Hopfield neural network to evaluate water quality comprehensively, and compared the results from single Hopfield neural network and traditional Nemero Index method. Results showed that the factor analysis Hopfield neural network is much better than single Hopfield neural network. It not only makes up for the defect that factor analysis does not achieve the classification of water quality in practical applications, but also effectively reduces the extent of over-fitting of Hopfield neural network. The evaluation results are more reasonable and the model provides a new approach to comprehensive water quality evaluation with some excellent prospects.

Keywords: factor analysis; Hopfield neural network; water quality evaluation; Dongliao River

人工神经网络模型是一种基于生物学的智能仿生模型，是由大量处理单元（神经元）互联组成的非线性大规模自适应动力学系统[1-2]。从网络结构角度可分为前向网络与反馈网络，而 Hopfield 神经网络是反馈型神经网络的典型代表[3-4]。基于它的联想记忆功能，利用 Hopfield 神经网络进行水质综合评价，取得了一定的成果。

在水质评价过程中，多种水质指标共同影响水体，同时各指标对水体影响大小又各不相同[5]。对于 Hopfield 神经网络模型而言，并非所有因子都是影响水质评价的必需因子。实际上，如果不加选择地将这些因子作为输入层引入，就会将一些代表噪音的因子加到模型中，会导致模型的预测能力下降，这种情况称之为过度拟合[6-8]。因子分析方法是一种可以从多变量中确定主要影响因子的统计分析方法，将因子分析引入到神经网络模型中即可以在一定程度上解决以上问题。国内外学者将因子分析法和 BP，RBF 等神经网络相结合应用于各领域，以工程实例论证表明该类模型的估算精度比一般的神经网络估算模型要好[9-11]。目前，只有因子分析与 Hopfield 神经网络方法的融合在实际例子的应用少见。

本文首先应用因子分析法确定影响水质综合评

收稿日期：2011-02-19 修回日期：2011-04-19

资助项目：国家水体污染控制与治理科技重大专项“辽河流域控制单元水质目标管理技术”(2009ZX07526-006-04-01)；国家自然科学基金项目“基于数值模拟的表面活性剂强化的 DNAPLs 污染含水层修复过程优化问题研究”(41072171)

作者简介：卢文喜(1956—)，男(汉族)，吉林省德惠市人，博士后，教授，博士生导师，主要从事地下水数值模拟与优化管理及生态水文等方面的研究。E-mail：luwenxi@jlu.edu.cn。

价的主要因子，然后以此构造 Hopfield 神经网络的输入样本，建立 Hopfield 神经网络的水质评价模型，以提高模型的学习和泛化能力，得到更为客观准确的评价结果。

1 方法简介

1.1 因子分析

因子分析是多元分析中处理降维的一种统计方法，它是将具有错综复杂关系的变量（或样品）综合为数量较少的几个因子，以再现原始变量与因子之间的相互关系，同时根据不同因子还可以对变量进行分类[12]。这种方法通过确定 p 个“公因子”正交变量，以便最优地显示 m 个变量之间的差异（$p<m$）。经过标准化处理和数学变换，提取出公共因子，再用这些公共因子来描述和代替原始变量，可以反映和解释原始变量之间的复杂关系[13]，达到数据降维的目的。假如有 n 个水质样本，每个样本共有 p 个水质指标（$x_1, x_2, \cdots, x_p$），构成 $n \times p$ 的数据矩阵：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

用因子分析确定评价因子的具体步骤可参考文献[14]。

1.2 Hopfield 神经网络模型

Hopfield 网络简称 HNN，由 J. J. Hopfield 于 1982 年提出，是由若干基本神经元构成的一个单层全互连神经网络，任意 2 个神经元之间都有连接，是一种对称连接结构[15-16]。Hopfield 网络模型分离散型和连续型 2 种网络，本文采用离散 Hopfield 神经网络对水体进行分类。离散型 Hopfield 网络是一种反馈型单层二值型神经元网络，它的所有结点之间互相连接，每个结点连接权接受其他结点反馈来的信息，以便让任一个神经元输出都受其他神经元的控制，从而各神经元的输出能相互制约。每个神经元设有一个阈值，以反映对输入的噪声控制[17]。在 DHNN 模型中，定义 $X=[x_1, x_2, \cdots, x_n]^T$ 为网络状态的矢量，其分量是 n 个神经元的输出，仅取 -1 或 1 二值。

$$X_i=\begin{cases} 1 & \left(\sum\limits_{j\neq i} W_{ij}x_j-\theta>0\right) \\ -1 & \left(\sum\limits_{j\neq i} W_{ij}x_j-\theta_i\leqslant 0\right) \end{cases} \quad (1)$$

式中：x_j——第 j 个输入分量；X_i——单个神经元的状态变数，其值为 1 或 -1；W_{ij}——连接权值；θ_i——连接阈值。$W_{ij}=W_{ji}$，且 $W_{ij}=0$，即 DHNN 采用对称连接、无自身反馈。

Hopfield 神经网络是一个复杂的非线性动力系统，用“能量函数”分析系统的稳定性，网络从高能状态转移到最小能量状态，最后收敛于系统的稳定点[18]。Hopfield 网络用李亚普洛夫函数很好地解决了联想记忆问题。李亚普洛夫函数形式：

$$E=-\frac{1}{2}\sum_i\sum_j W_{ij}X_iX_j+\sum_j\theta_jX_j \quad (2)$$

式中：X_i, X_j——单个神经元的状态变数，其值为 1 或 -1 两级值；W_{ij}——连接权值；θ_j——连接阈值。

2 实例应用

选取东辽河为研究区域，在因子分析和 Hopfield 神经网络 2 种算法的基础上，建立基于因子分析的 Hopfield 神经网络水质评价模型。东辽河位于吉林省中西部地区，是吉林省 19 条主要江河之一，是水资源供需矛盾比较突出的河流[19]。此次计算利用 2010 年 7 月的水质数据，从影响水质的 12 个水质指标中提取 4 个公共因子，并由因子得分系数和方差贡献率确定各水质指标权重，根据指标权重确定主要影响水质指标并构造 Hopfield 神经网络结构，从而进行水质综合评价。图 1 表示因子分析—Hopfield 神经网络模型的流程图。

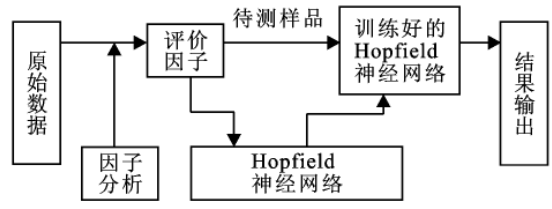

图 1 因子分析－Hopfield 神经网络模型的流程图

2.1 因子分析

按照文献[14]中所述步骤进行计算如下：

（1）将水质数据标准化。

（2）建立变量的相关系数阵 R。

（3）求 R 的特征根及相应的单位特征向量。

由以上步骤可求得主因子有 4 个，其方差累计贡献率达到 86.8%，可以代表原评价指标绝大部分的信息（表 1）。

表 1 因子的贡献率及相应的累计贡献

主因子	特征值	贡献率	累计贡献率
F_1	5.006	0.417	0.417
F_2	3.263	0.272	0.689
F_3	1.247	0.104	0.793
F_4	0.905	0.075	0.868

（4）对因子载荷阵施行方差最大正交旋转，再采用回归分析的方法计算因子得分系数（表 2）。

表 2　前 4 个主因子的权系数

主因子	权系数											
	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
F_1	−0.370	0.175	−1.060	−0.100	0.081	−0.080	0.398	0.362	0.658	−3.690	1.913	2.456
F_2	0.365	−0.050	0.539	−0.320	−0.130	0.594	−0.540	−0.210	−0.790	0.826	−0.780	0.470
F_3	0.080	0.871	−0.240	0.343	0.079	0.046	−0.170	0.060	−0.180	0.220	−0.020	0.009
F_4	0.117	0.095	0.406	0.026	0.822	0.528	0.173	−0.330	−0.240	0.602	−0.160	0.026

（5）利用式(3)得出各评价参数的权重(表 3)。

$$W_j = \sum_{i=1}^{m} \beta_{ij} e \Big/ \sum_{j=1}^{p} \sum_{i=1}^{m} \beta_{ij} e_i \qquad (3)$$

式中：β_{ij}——第 i 个公共因子在第 j 个指标上的因子得分系数；e_i——第 i 个公共因子的贡献率在累计贡献率所占的比例。

表 3　各评价参数的权重

参数权重	TN	TP	氨氮	高锰酸盐指数	COD	DO	BOD_5	Cr	As	Cu	Zn	Pb
W_j	0.105	0.046	0.136	0.108	0.023	0.028	0.088	0.039	0.010	0.156	0.127	0.134

注：W_j 为各评价参数的权重；TN 为总氮；TP 为总磷；COD 为化学需氧量；DO 为溶解氧；BOD_5 为 5 d 生化需氧量。

2.2　建立 Hopfield 神经网络水质综合评价模型

根据因子分析的结果，按照表 3 的各水质指标的权重由大到小排列，前面 7 个水质指标的权重累积达到 85.58%，因此，从影响水质的 12 个指标中确定 7 个评价因子，包括五日生化需氧量、高锰酸盐指数、氨氮、铜、锌、铅、总氮，建立 5×7 Hopfield 神经网络水质综合评价模型，以水质评价标准作为训练样本：

（1）设计网络的记忆模式，即将预存储的模式进行编码，得到取值为 1 和 −1 的记忆模式。记忆模式可以用图 2 来表示。

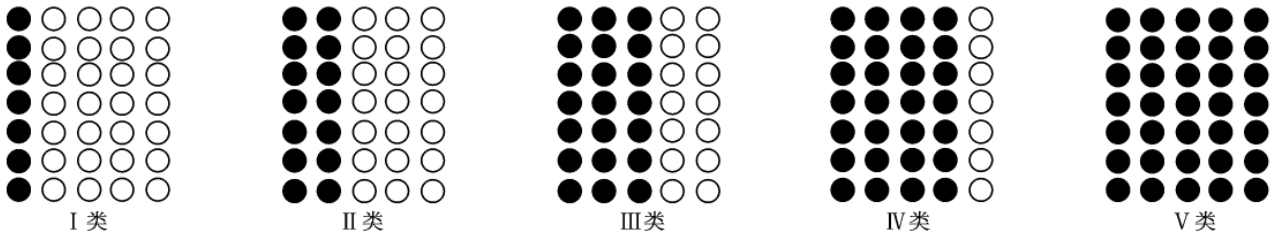

图 2　设计网络的记忆标准模式

（2）调用 MATLAB 的 newhop 函数建立离散型 Hopfield 网络，经训练记忆得到权值矩阵 w 及阈值向量 b。

（3）把原始数据的编码模式设为初始状态，调用 MATLAB 的 sim 函数进行多次迭代使网络收敛。

（4）待到网络稳定后，可得到各测点水质的综合评价结果[20]。

为了说明因子分析—Hopfield 神经网络评价结果客观准确，符合实际，根据地表水环境质量标准的Ⅲ类标准，采用内梅罗指数评价法和 Hopfield 神经网络评价法对东辽河水质进行评价，并将 3 种方法的评价结果进行对比。

吉林省政府发布的最新环境状况公报显示：二龙山水库和辽河源镇水质类别分别Ⅳ类水体与Ⅲ类水体。因子分析—Hopfield 方法中二龙山水库的水质类别为Ⅴ类，辽河源镇水质类别为Ⅲ类水体，而单一的 Hopfield 网络中，两地的水质类别分别为Ⅱ类和Ⅰ类。孔家屯由于受公主岭市工业污水的直接影响，污染严重，因子分析—Hopfield 方法中孔家屯的水质类别为Ⅴ类，符合实际情况，而单一的 Hopfield 网络的水质类别为Ⅱ类，这与实际情况不相符。从表 4 来看，单一的 Hopfield 神经网络评价结果偏小，这是一些代表噪音的因子加到模型中，引起网络的过度训练所造成的，而因子分析—Hopfield 神经网络法和内梅罗指数法的评价结果比较接近，在这 2 种方法中双城的水质类别分别为Ⅲ类和Ⅴ类，根据水质数据可知，双城监测点仅有氨氮和总氮为Ⅴ类标准，其他的几项优于Ⅲ类标准，其中 3 项还达到Ⅰ类标准，可知，因子分析—Hopfield 方法比传统的内梅罗指数法更为合理。

表 4 3 种方法的评价结果比较

名 称	内梅罗指数法	Hopfield	因子分析—Hopfield	名 称	内梅罗指数法	Hopfield	因子分析—Hopfield
东西辽交汇	Ⅲ	Ⅰ	Ⅲ	温家河口	Ⅲ	Ⅰ	Ⅳ
三江口大桥	Ⅳ	Ⅰ	Ⅲ	兴隆河	Ⅴ	Ⅱ	Ⅴ
新开河	Ⅲ	Ⅰ	Ⅲ	七一村大桥	Ⅴ	Ⅲ	Ⅳ
万斤沙场	Ⅲ	Ⅰ	Ⅲ	二龙山水库	Ⅴ	Ⅱ	Ⅴ
温德河	Ⅲ	Ⅰ	Ⅲ	二道河	Ⅴ	Ⅲ	Ⅴ
桑树台	Ⅳ	Ⅰ	Ⅲ	双城	Ⅴ	Ⅲ	Ⅲ
十 屋	Ⅴ	Ⅱ	Ⅳ	西孟东辽河	Ⅴ	Ⅱ	Ⅳ
孔家屯	Ⅳ	Ⅱ	Ⅴ	梨树大桥	Ⅲ	Ⅰ	Ⅳ
卡伦河	Ⅴ	Ⅱ	Ⅳ	寿山镇	Ⅲ	Ⅱ	Ⅳ
二十家子	Ⅲ	Ⅰ	Ⅴ	辽河源镇	Ⅲ	Ⅰ	Ⅲ

单一的 Hopfield 神经网络由于水质指标种类及数量不同，水质评价结果会有很大的差别。因子分析—Hopfield 神经网络模型利用因子分析法确定水质指标种类及数量，因而其水质评价结果更为科学合理。因子分析—Hopfield 神经网络不仅适用定量的水质指标，也适用于定性的水质指标，网络设计简单、过程直观、稳定性好，比一般方法人为主观影响因素少，不须试算，其算法明显优于单一的神经网络。

因子分析法用于水质评价时，可以根据综合因子得分反映各个断面的污染严重程度，也可以根据各主因子得分，识别河流水质中的主要污染物及其来源，但在实际应用中却不能对水质进行分级。因子分析—Hopfield 神经网络模型将因子分析找出主要影响因子的功能和 Hopfield 神经网络非线性计算特点很好结合起来，不仅在一定程度上弥补了因子分析在实际应用中没有实现水质分级的缺陷，而且通过因子分析的特征降维，排除具有相关性数据的影响，有效地降低了 Hopfield 神经网络的过度拟和的程度，进而提高了网络识别精度，同时简化了神经网络结构，加快网络收敛速度，节省运行时间。

然而因子分析—Hopfield 神经网络模型是在 Hopfield 神经网络模型基础上建立起来的，它实际上是黑箱模型，无法说明其内部逻辑关系，很难查找可能出现的问题，同时由于 Hopfield 网络采用编码制输入，在模型实际应用过程中要注意存储容量的问题。

3 结 论

本文针对众多的存在一定相关性的影响水质因子引起的神经网络过度拟合的问题，在因子分析的基础上，结合 Hopfield 神经网络算法提出了因子分析—Hopfield 神经网络算法。因子分析—Hopfield 神经网络明显优于单一的 Hopfield 神经网络，不仅在一定程度上弥补了因子分析在实际应用中没有实现水质分级的缺陷，而且利用因子分析找出主要影响因子的特点有效地降低了 Hopfield 神经网络的过度拟和的程度，评价结果具有客观性，同时水质综合评价网络设计简单、过程直观，减少了人为主观因素的影响，不须试算，加快了网络收敛速度，进而节省了运行时间，为水质综合评价提供了一种新的思维方式，具有极好的应用前景。

[参 考 文 献]

[1] 刘同有. 地下采矿系统分析与综合集成[M]. 北京：地质出版社，1998：131-135.

[2] Hornik K. Approximation Capability of Multi-layer Feedforward Networks[J]. Neural Networks，1991，4(12)：241-257.

[3] 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，2003：23-27.

[4] 霍卫峰. 基于因子分析的神经网络方法的研究及在分子筛合成数据分析中的应用[D]. 长春：吉林大学，2007.

[5] 王里奥，任家宽，刘阳生，等. 基于因子—神经网络预测南渡江海口段水质状况[J]. 环境科学与管理，2008，33(6)：176-179.

[6] 吴玉鸣. 中国农业因子分析神经网络组合预测研究[J]. 农村经济，2003(3)：58-61.

[7] Lee K C，Han I，Kwon Y. Hybrid neural network models for bankruptcy predictions [J]. Decision Support Systems，1996，18(1)：63-72.

[8] Capolo M，Andreussi P，Soldati A. River filled forecasting with a neural network model [J]. Water Resour. Res.，1999，35：1191-1197.

[9] Kevin R J，Simon X Y，Roger R H. Pork farm odour modelling using multiple-component multiple-factor analysis and neural networks[J]. Applied Soft Computing，2005，6(1)：53-61.

（下转第 237 页）

之间河段水环境容量达不到国家 3 级水质标准,到现在为止仍然为劣 5 类水质。

西充河 2 个监测断面长达 7 a 的观测中,COD_{Cr} 超标含量从以前的 16 倍降到现在的 1.37 倍,改善效果明显。但是氨氮含量却从 20 世纪 80 年代的 2 倍上升到 153 倍,氨氮含量的增高会对鱼类等水生生物的酶水解反应和膜稳定性产生明显影响,含量过高可导致水生生物死亡[11]。这也解释了西充河看似污染减轻却鲜有鱼虾的原因。

氨氮排放中有 84.63%来自于华凤街道与晋城镇(包括常林乡)。华凤街道居民依河而居,生活污水大多都未经处理就直接排入河中。常林乡河道两旁有大量农田,种植较多经济作物,对化肥的施用量较大,很多未被植物吸收就随着雨水或灌溉水流入河中。这两方面或成为导致西充河中氨氮严重超标的主要因素。

在今后的治理过程中,减少生活污水的直接排放与控制农业生产中的化肥施用量将成为改善西充河水环境的主要方向。

[参 考 文 献]

[1] 杜德泉.南充西河的污染与防治[J].水资源保护,1992(1):37-42.

[2] 罗明云.南充市西河治理探讨[J].环境科学与管理,2006(4):116-119.

[3] 冯永春,胡本君,巫幸福,等.西河污灌区环境污染对居民健康影响的调查研究[J].四川环境,1992,11(2):27-30.

[4] 刘开第,吴和琴,庞彦军,等.盲数的概念、运算与性质[J].运筹与管理,1998,7(3):14-17.

[5] 李如忠,洪天求.盲数理论在湖泊水环境容量计算中的应用[J].水利学报,2005,36(7):1-9.

[6] 李如忠,钱家忠,汪家权,等.基于盲数理论的河流水质未确知风险分析初探[J].水电能源科学,2003,21(1):18-21.

[7] 高苏蒂,祝荣,景连茴.基于盲数理论计算一般河流水环境容量可能值[J].安徽建筑工业学院学报:自然科学版,2010,18(6):89-91.

[8] 贾瑞娟.盲数的运算律及证明[J].河北建筑科技学院学报,1998,15(2):69-72.

[9] 闫欣荣,史忠科.盲信息下渭河水环境容量及消减量研究[J].西安建筑科技大学学报:自然科学版,2007,39(5):684-689.

[10] 邓柏旺,袁晓宇.水环境容量计算在工程中的应用[J].中国水利,2010(18):48-52.

[11] 蔡继晗,沈奇宇,郑向勇,等.氨氮污染对水产养殖的危害及处理技术研究进展[J].浙江海洋学院学报:自然科学版,2010,29(2):167-172.

(上接第 200 页)

[10] 陈卫东,李俊梅,李跃根,等.基于因子分析和 BP 神经网络的海洋平台拖航风险分析[J].海洋技术,2010,29(2):107-111.

[11] 吴倩,宋永发.基于因子分析和 RBF 神经网络的快速估算模型[J].工程管理学报,2010,24(1):65-68.

[12] 于秀林.多元统计分析[M].北京:中国统计出版社,1999:31-33.

[13] 陈东景,马安青,徐中民,等.因子分析法在水质评价中的应用[J].水文,2002,22(3):31-33.

[14] 李文生,许士国.基于因子分析定权的水质评价模型[J].辽宁工程技术大学学报,2008,27(3):444-446.

[15] 高芳,赵强,赵刚.基于离散型 Hopfield 神经网络的供应商评价模型[J].计算机集成制造系统,2004,10(1):95-98.

[16] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proc. Natt. Acad. Sci. USA, 1982,79(8):2554-2558.

[17] 沈清.模式识别导论[M].长沙:国防科技大学出版社,1991:313-316.

[18] 虞和济,陈长征,张省.基于神经网络的智能诊断[M].北京:冶金工业出版社,2000:120-127.

[19] 韩庆良,郑文卓.影响东辽河上游河道水量减少的因素[J].吉林水利,2005,6(6):18-19.

[20] 崔永华,左其亭.基于 Hopfield 网络的水质综合评价及其 matlab 实现[J].水资源保护,2007,23(3):14-16.