活性污泥系统神经网络建模与仿真研究

楼文高[1,2]　刘遂庆[1]

（1. 同济大学环境科学与工程学院，上海 200092；2. 上海理工大学，上海 200093）

摘要　采用神经网络技术对松江污水厂污水处理活性污泥系统进行建模试验研究，在对实际运行数据按机理准则和范围准则剔除异常数据后，将样本数据随机分成训练样本、检验样本和测试样本。用试凑法确定合理的神经网络隐层节点数，以避免采用过大或过小的网络结构，在训练过程中用检验样本实时监控从而避免“过训练”现象的影响，较好地解决神经网络方法建模的两大难题，从而建立可靠、有效的活性污泥系统神经网络模型。并应用建立的网络模型对活性污泥系统的运行情况进行了仿真研究。建模研究表明，神经网络技术能较好地应用于活性污泥系统的建模，模型具有较好的泛化能力，有很好的实用价值。

关键词　活性污泥系统　神经网络　建模　仿真　泛化能力　样本

Neural network for modeling and simulation of activated sludge system　*Lou Wengao*[1,2], *Liu Suiqing*[1]. (1. *School of Environmental Sciences and Technology, Tongji University, Shanghai* 200092; 2. *University of Shanghai for Science and Technology, Shanghai* 200093)

Abstract: The actual operation data of the activated sludge system of Shanghai Songjiang Sewage Treatment Plant were used to establish a neural network-based (NN-based) model of the activated sludge system. After deleting abnormal data, the remaining data were divided into three sets: training data, verification (validation) data and test data. The reasonable neuron number on hidden layer was determined by trail-and-error. The verification data set was used to monitor the training process and to avoid the over-trained phenomena. A reliable NN-based model for activated sludge system was developed and was successfully employed for performance simulation of the non-linear and complicated activated sludge system. The NN-based activated sludge model is an attractive simulation tool.

Keywords: Activated sludge system　Nneural network　Modeling　Simulation　Generalization　Sample

污水处理系统是一个复杂的非线性系统，尽管对污水处理的机理进行了几十年的研究，提出了一些基于基质降解和微生物生长规律和反应器理论的数学模型（如莫诺方程式、劳伦斯－麦卡蒂方程式等），对污水处理技术发展起到了积极的推动作用[1]，但实际过程要复杂得多，同时上述模型还存在很多难以解决的问题。从根本上说，给排水学术界现阶段对活性污泥系统还知之不多。用机理模型等描述“白箱”系统的方法研究这个“灰箱”或者“黑箱”系统，很难取得满意的效果。

神经网络技术具有好的自适应性、自学习性和容错性的非线性建模技术，已在很多复杂系统的建模中取得了成功，得到了广泛的应用[2~4]，被誉为“21 世纪最有前途的建模技术之一”。自 20 世纪 90 年代初以来[5,6]，神经网络技术已在污（废）水处理领域得到了一定的应用[7]，取得了一些探索性的研究成果。但由于多数作者对神经网络技术缺乏较深入的理解，在建模过程中采用的网络结构不尽合理，也没有采用检验样本实时监控训练过程，即训练过程是否发生“过训练”现象无法判定。

笔者在分析总结的基础上，提出了神经网络建模的基本原则和步骤，用试凑法确定合理的神经网络结构，用训练样本实时监控训练过程以避免出现“过训练”现象的影响，从而确保建立的神经网络模型的泛化能力和实用价值。

1　污水处理活性污泥系统简述

松江污水厂采用普通活性污泥处理系统。对污水厂尾（出）水的设计排放要求见表 1。

表 1　松江污水处理厂污水处理指标[7]　mg/L

指标	COD_{Cr}	BOD_5	SS	NH_3-N
进水水质	450	210	250	—
尾水水质设计标准	100	30	30	5

影响污水处理效果的因素很多，根据报表能提供的数据，取原污水水质指标 COD_{Cr}、BOD_5、SS、NH_3-N、pH、TKN 和运行控制参数的水力平均停留时间（沉砂池或曝气池容积/进水流量，HRT）、曝气池内 MLSS 和 MLVSS、污泥沉降比 SVI 等 11 项指标。

第一作者：楼文高，男，1964 年生，教授，博士。主要从事人工神经网络理论、多指标综合评价等现代数据处理技术在环境科学与工程中的应用研究与教学工作。

* 上海市教委高等学校科学技术发展基金资助项目（No. 01H03）

根据要求，需要建立的是某时出水水质与该部分水进入污水厂时的水质及其运行控制参数之间的数学模型，而不是建立同一时刻出水水质与进水水质、运行控制参数之间的关系。根据该厂设计资料可知污水从进厂到出厂的 HRT 为 21 h，实际运行的 HRT 为 25 h。因此，建立某天的出水水质与前一天进水水质与当天运行控制参数之间的数学模型是合理的。如果，HRT 与 24 h 相差很大，则对测定进水和出水水质的时间应该相差水力平均停留时间，不能测定同时取得的水样水质。这一点尤为重要，否则建立的是伪模型，没有实际意义。

取松江污水厂 2003 年 1 月 1 日～2004 年 10 月 31 日的实际运行数据为研究的原始数据。对原始数据首先剔除不完整的数据，在校核无误的情况下，剔除明显过大或过小的数据。再根据污水本身固有的大小关系或污水处理前后的大小关系（污水处理机理准则）进行判断，校核数据。经上述整理与预处理，得有效的数据 630 组。

2　人工神经网络模型

在现有的人工神经网络应用中，80%～90%采用 BP 网络模型（包括用拟牛顿法、L-M 法等）。它由一个输入层、一个输出层和若干个隐（含）层组成。同层之间的节点没有联系，相邻层的节点两两相连，前一层的输出即为后一层的输入。基本运行机制是：由信息正向传播和误差反向传播两个过程组成，工作原理详见文献[2～4]。

2.1　BP 网络模型存在的不足

BP 网络模型能较好地应用于模式识别、分类、数值逼近、控制等领域，但 BP 网络模型良好的学习性、非线性逼近能力和泛化能力并不是其本身所固有的，而是在满足建模条件的情况下才特有的。大量研究表明，BP 网络建模的关键和核心问题是：①训练过程中如何避免进入局部极小点；②如何得到合理的网络结构。网络结构太大，是出现“过训练”的内因，训练过程往往出现“过训练”现象，结构太小，往往表现为“欠拟合”，不能得到满意的精度。出现“过训练”现象，建立的模型就没有泛化能力，模型就没有实用价值。根据存在性定理，只要隐层节点数足够多（比训练样本数少一个），网络总可以以任意精度逼近期望值（即训练样本误差趋近零）[2～4, 8～10]，但可能没有泛化能力（即非训练样本误差比训练样本误差大得多）。但遗憾的是，合理隐层及其节点数的多少与问题的复杂程度等因素有关，迄今为止，还没有理论计算公式。

2.2　建立 BP 网络模型的基本原则和步骤

为了确保 BP 网络模型的泛化能力，必须围绕和解决好上述关键和核心问题，遵循一定的基本原则和步骤。

2.2.1　样本数据的分组

将收集到的数据随机地分成训练样本、检验样本（10%以上）和测试样本（10%以上）3 部分。训练样本用来根据相应的算法调整网络连接权值而使误差函数趋于极小。检验样本用来实时动态监控训练过程，判定训练过程是否出现了“过训练”现象[2～4, 8～10]，随着训练次数的增加，训练样本误差始终是减小，而检验样本误差从开始时的减小继而出现增大的趋势，表明出现了“过训练”现象。测试样本用来判定建立的模型的泛化能力。如果检验样本、测试样本的误差与训练样本误差基本相同或稍大，表示模型具有较好的泛化能力。目前国内（污）水处理神经网络建模的所有论文均没有检验样本（有的论文把测试样本误称为检验样本）[4,5]，无法判断这些论文的建模（训练）过程是否发生了“过训练”现象。

2.2.2　确定合理的网络结构

针对一定数量的训练样本，总存在一个合理的隐层数和隐层节点数。合理的网络结构是指在满足精度要求的前提下取尽可能紧凑的结构，即取尽可能少的隐层数和隐层节点数。必须注意两点：一般取一个隐层；对于三层网络，输入层和隐层节点数必须至少比训练样本数少一个。一般要求，训练样本数多于网络连接权值数；如果不采用检验样本实时监控训练过程，训练样本数是网络连接权值数 10 倍以上时通常也能取得较好的结果[4, 6, 11]。合理的隐层节点数不仅与输入/输出层节点数有关，更与需解决问题的复杂程度等因素有关。目前各种文献提供的确定隐层节点数的计算公式都是针对训练样本任意多或个别特殊情况，不具有一般性，不宜直接采用。目前最有效的途径是采用试凑法（节点删除法和扩张法）确定合理的隐层节点数。

2.2.3　随机改变网络初始权值

如果网络初始权值相同，则每次训练只能搜索到相同的极值点（局部或全局），很难求得全局极小点邻域。因此，程序必须具有能够随机改变网络初始权值的功能[8,9]。

2.2.4　训练网络模型

训练的目的就是根据算法不断调整网络连接权值，使训练样本的误差平方和达到最小或小于某一期望值。目前，在给定有限个（训练）样本的情况下，如何通过训练设计一个合理的 BP 网络模型能满意地逼

近这些样本所蕴含的规律(不仅仅使训练样本误差达到很小),很大程度上还需要依靠先验知识和设计者的经验。从存在性结论知,即使每个训练样本的误差都很小,非训练样本误差仍可能很大。判断建立的模型是否已有效逼近训练样本所蕴含的规律,应该也必须用随机抽取的非训练样本(本文称为检验样本和测试样本)误差的大小来表示和评价,最直接和客观的指标是非训练样本误差与训练样本误差具有相近的数值[2~4, 9~11],否则,说明建立的模型没有有效逼近训练样本所蕴含的规律,而只是在这些训练样本点上逼近而已。对于同一网络结构,通过选取多组(通常是几十组,由问题的复杂程度而定)不同的网络初始连接权值对网络进行训练,选取没有发生"过训练"时的精度较高的网络连接权值(即可认为是全局极小点)。

2.2.5　确定合理的网络模型

随着网络结构的增大,训练样本误差变小。合理的网络模型是指具有合理隐层及其节点数的、训练时没有发生"过训练"现象、求得全局极小点邻域的模型。

3　建立活性污泥系统神经网络模型

针对 630 组建模数据,取 29 组(2004 年 10 月份有效数据)为预测数据,各随机抽取 60 组数据为检验样本和测试样本,其余 481 组数据为训练样本。

大量实践和示例试算表明:训练多个具有一个输出的网络模型比训练一个具有多个输出的网络模型要方便和简单得多,同时也不会影响后续的预测与控制。因此,根据本研究的原始数据,神经网络模型输入层节点有 11 个节点(变量),输出层有 1 个节点(本文以出水 COD_{Cr} 为例)。根据训练样本数要多于网络连接权值数的要求,隐层节点数不能多于 36 个(连接权值数为 471)。采用节点扩张试凑法:每次增加 3~5 个隐层节点,针对每个结构,随机改变网络初始连接权值试算 50 次以上)。

本研究采用 STATSOFT 公司出品的 STATISTICA Neural Network 软件[9]。经试算,当隐层节点数为 35 个(连接权值数为 456)时,模型具有较好的泛化能力,训练样本、检验样本、测试样本的误差和数据本身特征值如表 2 所示。检验样本、测试样本和预测样本的实测值与模型计算值如图 1~3 所示。

4　活性污泥系统仿真与预测

建立的神经网络模型可以简洁地表示为:

$$y_{COD_{Cr}} = f(X, W_h, W_o) \qquad (1)$$

式中:$y_{COD_{Cr}}$ 为出水 COD_{Cr};X 为输入变量;W_h, W_o 分别为输入层与隐层和隐层与输出层之间的网络连接权值,已通过前述训练得到。

因此,污水厂出水 COD_{Cr} 值主要取决于进水水质和各个运行控制参数(输入变量 X)。因此,可以通过改变任何一个或几个变量,同时保持其他变量(参数)不变,调用神经网络模型,很方便地实现对活性污泥系统出水水质的仿真与预测研究。

以曝气池水力平均停留时间 HRT2 为例加以说明。HRT2 的有效取值范围为 10.5~17.0,其他变量取实际运行数据的均值,如表 3 所示。

表 2　训练样本、检验样本、测试样本和预测数据的各项模型性能指标值和数据特征值

指标	训练样本 模型值(实际数据)	检验样本 模型值(实际数据)	测试样本 模型值(实际数据)	预测数据 模型值(实际数据)
RMSE	5.5526	5.7928	5.9322	4.7764
AAE	3.861	4.397	4.125	4.261
R	0.9300	0.9013	0.8998	0.6843
MAPE	7.13%	8.29%	8.35%	8.32%
σ	14.97(15.11)	12.53(13.48)	12.46(13.71)	7.78(6.67)
平均值	57.24(57.35)	56.66(56.95)	56.67(55.48)	52.06(51.93)
最大值	107.43(110.00)	95.71(97.90)	100.45(103.80)	64.24(60.60)
最小值	30.09(30.30)	31.89(30.90)	37.34(34.70)	38.49(38.20)

注:( )内是实际数据的参数值,如 15.11 表示训练样本实际数据的标准剩余离差值,其余同。RMSE、AAE、R、MAPE 和 σ 分别表示均方根误差、绝对误差均值、相关系数、相对误差均值和标准剩余离差。

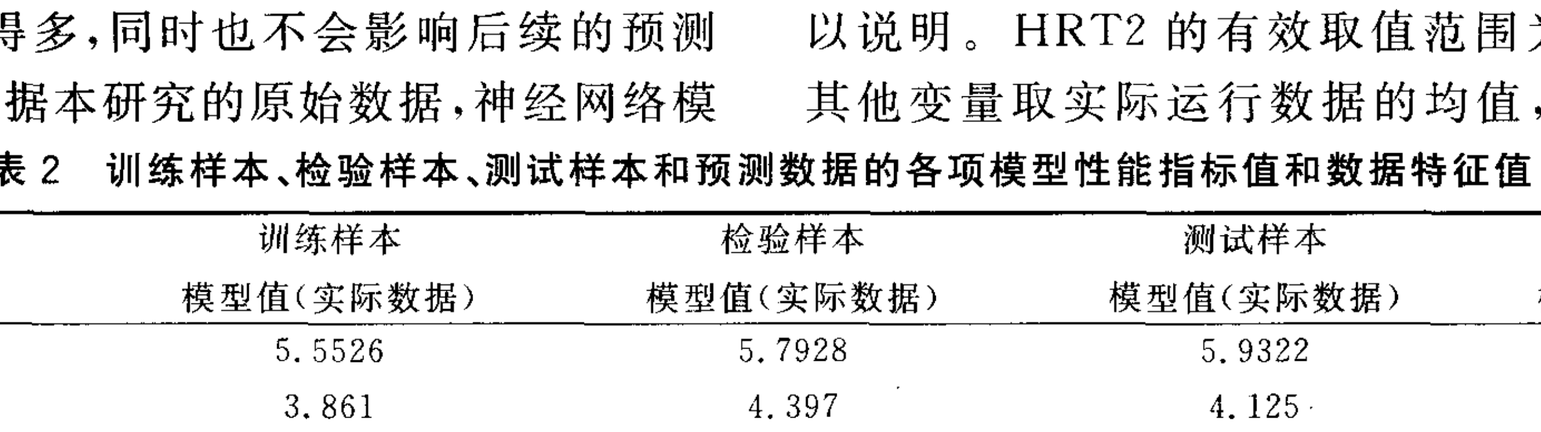

图 1　检验样本的实际数值和模型输出值

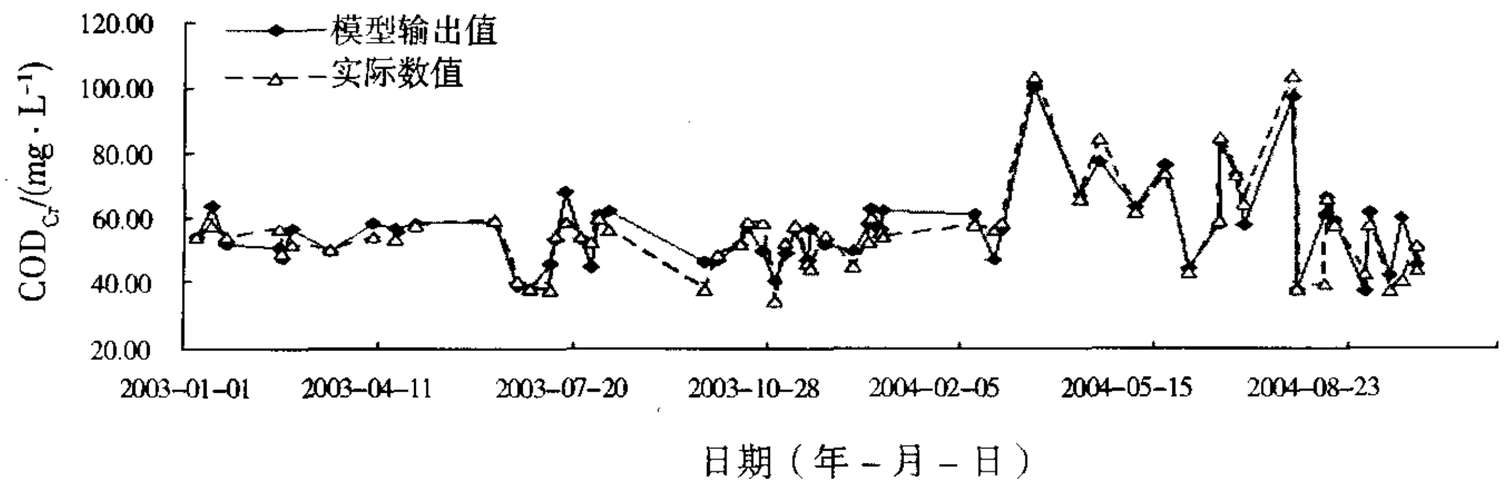

图 2　测试样本的实际数值和模型输出值

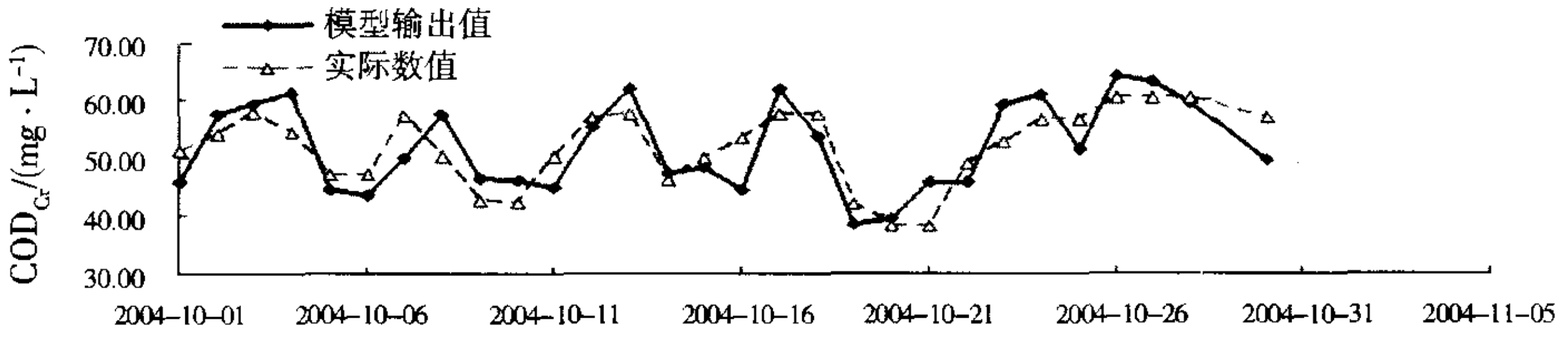

图 3　预测数据的实际数值和模型输出值

表 3　神经网络模型 10 个输入变量(指标)的均值

项目	pH(进)	COD_{Cr}(进)	BOD_5(进)	SS(进)	NH_3-N(进)	TKN(进)	HRT1	SV	MLSS	MLVSS
	/(mg·L⁻¹)	/(mg·L⁻¹)	/(mg·L⁻¹)	/(mg·L⁻¹)	/(mg·L⁻¹)	/(mg·L⁻¹)	/h	/%	/(mg·L⁻¹)	/(mg·L⁻¹)
均值	7.35	879	297	599	31.70	75.24	3.77	58.67	5 094	3 438

调用建立的网络模型，将 HRT2 的最小值设定为 10.5，最大值设定为 17.0，运行程序即仿真活性污泥系统工作情况，预测得到系统出水 COD_{Cr} 与 HRT2 的影响关系，如图 4 所示。图 4 表明，针对上述条件，HRT2 在 15.0 左右时，活性污泥系统处理效果最好。

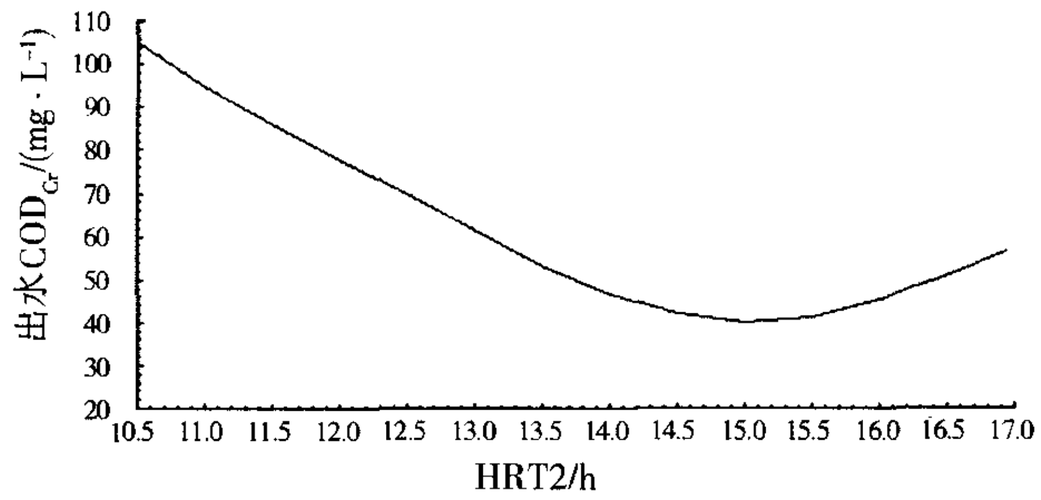

图 4　出水 COD_{Cr} 与 HRT2 的影响关系

5　结　语

(1) 对于高度非线性、工作机理不甚清楚的污水处理活性污泥系统，采用自学习性、自适应性、容错性较好的适用于非线性、黑箱系统建模的神经网络技术，理论上能取得较好的效果，具有较好的理论研究和应用价值。

(2) 针对 BP 网络模型建模的关键和核心问题，本文提出了相应的建模基本原则和步骤：通过检验样本实时动态监控训练过程，如发现进入局部极小点和出现“过训练”现象(表现为训练样本误差减小而检验样本误差出现增大趋势)立即停止训练；采用节点逐步扩张或删除的试凑法确定合理的网络结构，采用几十次随机改变网络初始连接权值以求得全局极小点邻域内的可行解。实例建模表明，遵循上述基本原则和步骤可以建立泛化能力较好的、有效的和可靠的活性污泥系统神经网络模型。

(3) 可以很方便地利用建立的神经网络模型，实现对活性污泥系统运行情况的仿真和出水水质的预测研究，并可验证现有的活性污泥系统机理规律或发现新的机理规律，具有较好的实践意义。

参考文献

1　张自杰. 排水工程(下册). 北京：中国建筑工业出版社，2001
2　王文成. 神经网络及其在汽车工业中的应用. 北京：北京理工大学出版社，1998
3　谢庆生，尹　健，罗延科. 机械工程中的神经网络方法. 北京：机械工业出版社，2003
4　阎平凡，张长水. 人工神经网络与模拟进化计算. 北京：清华大学出版社，2000
5　Krovvidy S, Wee W. A knowledge based neural network approach for waste water treatment system, IJCNN International Joint Conference on Neural Networks, 17-21 June 1990, 1(1):327～332
6　田　禹，王宝贞，周　定. BP 及 RBF 人工神经元网络对臭氧生物活性炭水处理系统建模的比较. 中国环境科学，1998，18(1)：394～397
7　楼文高. 基于神经网络的活性污泥神经网络建模研究.[博士学位论文]，上海：同济大学，2005
8　董　聪. 多层前向网络的全局最优化问题. 大自然探索，1996，15(58)：27～31
9　Statsoft. Statistica Neural Networks. (Manual) Tulsa: Statsoft, Inc., 1999
10　张乃尧，阎平凡. 神经网络与模糊控制. 北京：清华大学出版社，2000
11　张青贵. 人工神经网络导论. 北京：中国水利水电出版社，2004

责任编辑：陈泽军　(修改稿收到日期：2005-05-20)