

DOI :10.3969/j.issn.1004-6933.2012.02.002

等维新息径向基神经网络模型在需水预测中的应用

许银山¹, 吕孙云^{1,2}, 梅亚东¹, 熊莹²

(1. 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 ;

2. 长江水利委员会水文局, 湖北 武汉 430010)

摘要 :针对现有需水预测模型进行多周期预测时存在误差随预测周期延长而累积、抗随机因素干扰能力不足等问题,提出动态等维新息径向基神经网络模型,采用聚类方法进行网络学习,并将其应用于东莞市年需水量预测中。结果表明,动态等维新息径向基神经网络模型相对于基本径向基神经网络模型具有更高的预测精度,并且预测误差不会随着预测周期的延长而累积。

关键词 :城市需水量;预测;动态等维新息径向基神经网络模型

中图分类号 :TV213.4

文献标识码 :A

文章编号 :1004-6933(2012)02-0008-04

Application of equi-dimensional information renewal radial basis function neural network model to urban water demand forecast

XU Yin-shan¹, LV Sun-yun^{1,2}, MEI Ya-dong¹, XIONG Ying²

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

2. Bureau of Hydrology, Changjiang Water Resources Commission, Wuhan 430010, China)

Abstract :To solve the problems of errors increasing with the extension of the forecast cycle and inadequate protection against random factor interference when using the existing models for multi-cycle urban water demand forecast, an equi-dimensional information renewal radial basis function neural network model was proposed, and the clustering approach was used for system learning. This model was applied to the annual water demand forecast in Dongguan City. The results show that the model has higher forecast accuracy than the fundamental radial basis function neural network model, and its forecast errors do not increase with the extension of the forecast cycle.

Key words :urban water demand; forecast; dynamic equi-dimensional information renewal radial basis; neural network model

随着人口的增加和经济社会的发展,人类社会对水资源的需求不仅在数量上日渐增加,而且在质量上渐趋严格。由于天然水资源时空分布不均、自然条件的变迁和人类活动对水资源的破坏及污染,水资源短缺、水灾害加剧、生态环境恶化等问题日趋严重,水资源供需矛盾更加突出,科学配置和严格管理水资源势在必行。需水预测是水资源配置和管理的基础,因此进行合理、有效、精确的需水预测就显得尤为重要。一般采用单周期(或单步)预测与多周期(或多步)预测描述有效预测周期的长短。如果根

据建立的预测模型对即将发生的未来一个时段的需水量进行预测即为单周期预测;同理对未来多个时段的需水量进行预测即为多周期预测。目前国内外应用于需水预测研究的模型和方法很多,但是对于资源规划而言,规划期一般都达到几年甚至几十年,这就需要从各种模型和方法中选择合适的多周期预测模型。然而一般来说,随着预测周期的增加,前几个预测周期的误差会累积并在下一个预测周期放大,因此寻求合适的多周期预测模型有着重要的现实意义。

定额法、系统动力学方法是在分析用水系统、收集多种用水量以及用水量影响因子数据后,分别对影响用水量的驱动因子和单位用水指标进行预测,获得需水量。由于这两种方法对影响用水的各种因素和关系进行了较为详细的描述,可以适用于经济社会和自然环境发生较明显变化下需水量的预测,从而得到较多周期的预测值,属于多周期预测方法,特别是定额法适用于预测周期达 20 a 甚至更长规划期的需水预测。不足的是由于用水量驱动力及用水强度在未来趋势存在很大的不确定性,导致这两种方法的预测结果也具有较大不确定性。时间序列分析本质上属于趋势外推,预测时段越靠近当前,相应的预测结果精度越高,因此其预测周期不宜太长。灰色系统 GM(1,1)模型^[1]同时间序列一样仅考虑了历史用水量序列,而 GM(1,N)模型虽考虑了需水量影响因子,但其实质还是一个指数模型。无论是 GM(1,1)模型还是 GM(1,N)模型,当未来时段需水量零增长或负增长时会产生严重的误差,且误差随预测周期的增加而增加,因此灰色系统模型的预测周期不宜太长。回归分析法、人工神经网络方法^[2-3]、支持向量机法^[4-5]及投影寻踪法^[6]都属于因果关系模型法,立足于现在和过去的用水模式,建立需水量与影响因子之间的因果关系,从而根据未来影响因子来预测未来需水量。当未来用水模式发生改变时,需要根据近期数据动态训练系统,近期数据对系统的影响很大,因此预测周期也不宜太长。

综上所述,对于多周期需水预测而言,大多数预测模型都不能保证其预测精度。因此笔者基于径向机神经网络(RBFANN)模型,引入动态等维新息的概念^[7],提出了动态等维新息径向机神经网络(DEIRBFANN)模型。该模型的核心思想是随着时段的推移,不断引入新信息,剔除旧信息,即每引入一个新信息就剔除掉一个最早的信息,使信息量不断丰富和更新,且维持信息的总量不变。动态等维新息技术通过不断的新陈代谢,更能反映系统在目前的特征,此外还可以避免随着信息的增加,建模运算量不断增大的困难。在多周期需水预测中引入动态等维新息的思想不仅能大大提高预测精度,拓宽适用范围,而且具有较高的理论和应用价值。

1 DEIRBFANN 模型

1.1 模型原理

RBFANN 是一种典型的有导师学习前馈网络,可以根据具体问题确定相应的网络拓扑结构,具有自学习、自组织、自适应功能,它对非线性连续函数具有一致逼近性,学习速度快,可以进行大范围的数

据融合,可以并行高速地处理数据。RBFANN 的优良特性使得其显示出比 BP 神经网络更好的生命力,正在越来越多的领域内替代 BP 神经网络。

DEIRBFANN 模型是指在应用传统的 RBFANN 模型进行多周期需水预测时,每向前预测一个时段,就将预测值作为已知信息加入到原始序列中,同时剔除掉最早的一组数据,保持信息维数不变,重新对模型进行率定预测下一个时段。这样不断地进行新陈代谢,逐个预测,依次递补,直到完成预测目标。

原始数据序列:

$$Q^0 = \{\underbrace{Q_1^0}_{\text{去掉}}, Q_2^0, \dots, Q_N^0, \underbrace{Q_{N+1}^0}_{\text{引入}}\} \quad (1)$$

新数据序列:

$$Q^0 = \{Q_2^0, \dots, Q_N^0, Q_{N+1}^0\} \quad (2)$$

事实上,在任何一个时间序列的发展过程中,随着时间的推移,将会不断地有一些随机扰动或驱动因素进入系统,系统也逐步得到更新,使旧数据的信息意义逐步降低。因此在对时间序列进行预测时,必须不断地考虑那些随着时间推移相继进入系统的扰动或驱动因素,在不断补充新信息的同时,及时去掉老信息,使整个系统一直处于更新和发展的过程中。

1.2 模型的学习过程

对于某一 RBFANN 模型,如果给定了训练样本,那么该网络的学习算法应该解决以下问题:结构设计(即如何确定网络隐节点数 h),确定各 RBF 的数据中心 c_i 及扩展常数 δ_i 、输出权值。一般情况下,如果确定了网络的隐节点数、数据中心和扩展常数, RBFANN 从输入到输出就成了一个线性方程组,此时可以采用最小二乘法求解。而 DEIRBFANN 模型相对于 RBFANN 模型来说,学习方法相同,只是在每进行一个周期的预测后,需要重新进行网络学习,率定模型参数。

聚类方法是最经典的 RBFANN 模型学习算法,由 Moody 和 Darken 在 1989 年提出^[8]。其思路是先用无导师学习方法(K-means 算法)确定 RBFANN 中 h 个隐节点的数据中心,并根据数据中心之间的距离确定隐节点的扩展常数,然后用有导师学习方法训练各隐节点的输出权值。具体步骤如下:

a. 确定预测样本资料。假设预测周期为 N ,如果为第 1 个预测周期,那么以模型率定期资料作为样本,如果为第 2 到 N 个预测周期,需引入前一个周期的预测值作为样本资料,并剔除掉最早一个时段的样本资料。

b. 算法初始化。选择 h 个不同的初始聚类中心,并令迭代次数 $k=1$ 。选择初始聚类中心的方法很多,如从样本输入中随机选取,或者选择前 h 个

样本输入,但这 h 个数据中心必须取不同值。第 k 次迭代时的数据中心为 $c_1^k, c_2^k, \dots, c_h^k$, 相应的聚类域为 $\theta_1^k, \theta_2^k, \dots, \theta_h^k$ 。

c. 计算所有样本输入与聚类中心的距离 $\|X_j - c_i^k\|, i=1, 2, \dots, h, j=1, 2, \dots, n$ 。对样本输入 X_j 按最小距离原则进行分类。即当 $i(X_j) = \min_i \|X_j - c_i^k\|, i=1, 2, \dots, h$ 时, X_j 被归为第 i 类, 即 $X_j \in \theta_i^k$ 。

d. 计算各类的聚类中心。

$$c_i^{k+1} = \frac{1}{n_i} \sum_{X \in \theta_i^k} X, i=1, 2, \dots, h \quad (3)$$

式中: n_i 为第 i 个聚类域 θ_i^k 中包含的样本数。

e. 如果任意 $c_i^{k+1} \neq c_i^k$, 转到步骤 b 进行下一步迭代, 否则聚类结束, 转到步骤 (e) 计算聚类中心。

f. 根据各聚类中心之间的距离确定各隐节点的扩展常数。隐节点的扩展常数取 $\delta_i = \alpha d_i$, 其中 d_i 为第 i 个数据中心与其他最近的数据中心之间的距离, 即 $d_i = \min_j \|c_j - c_i\|, i, j=1, 2, \dots, h$; α 为重叠系数。

g. 当各隐节点的数据中心和扩展常数确定后, 输出权矢量就可以用有导师学习方法训练得到, 但更简洁的方法是使用最小二乘法直接计算。假定输入为 $X(j=1, 2, \dots, n)$ 时, 第 i 个隐节点的输出为:

$$h_{j,i} = f(\|X_j - c_i\|) \quad (4)$$

则隐含层输出矩阵为 $H = [h_{j,i}], H \in R^{n \times h}$, 此时网络输出矢量为:

$$\hat{y} = H\omega \quad (5)$$

式中: ω 为输出权矢量。

令 $\epsilon = \|y - \hat{y}\|$ 为逼近误差, 则如果给定了教师信号 $y = (y_1, y_2, \dots, y_m)^T \in R^m$, 并确定了 H , 便可通过最小化下式求出网络的输出权值:

$$\epsilon = \|y - \hat{y}\| = \|y - H\omega\| \quad (6)$$

通常 ω 可用最小二乘法求得:

$$\omega = H^+ y \quad (7)$$

式中: H^+ 为 H 的伪逆, 即 $H^+ = (H^T H)^{-1} H^T$ 。

h. 通过率定的模型参数进行预测。

i. 判断模型学习停止条件, 即是否到最后一个预测周期, 是, 则停止学习; 否, 则转到第一步, 对下一个预测周期的模型参数进行率定。

2 东莞市需水量预测

2.1 需水量影响因素

东莞市地处南亚热带, 气候温和, 雨量充沛, 水资源丰富, 水环境容量较大, 但降雨径流年内分布不

均, 年际变化较大。随着东莞市产业经济结构的不断优化调整, 工业化程度不断提高, 对水资源的需求不断增加, 同时受水环境恶化及咸水线上溯影响, 沿海片、水乡片水质性缺水更为严重, 对 21 世纪东莞市社会经济的可持续发展构成了严峻挑战, 必须对水资源进行合理的规划和配置。而需水预测是水资源规划管理的基础性研究, 因此分析东莞市需水量的内在驱动力和影响因子具有十分重要的研究意义。笔者结合东莞市的产业结构以及中国气象观测站的实际情况与资料, 选取东莞市年末总人口、GDP、农业总产值、工业总产值、三产增加值、年平均气温、年降雨量、降雨量大于 0.1 mm 的天数和日照时数等共 9 项指标作为东莞市年需水量的影响因子 (其中气象数据为广州站的观测值; 1981—2000 年的社会经济指标和需水量参照文献 [6], 社会经济指标和需水量源自 2001—2007 年《东莞市水资源公报》和《东莞市社会经济》)。

2.2 最佳维数的确定

利用东莞市 1981—2007 年的资料, 以年为预测时段, 分别建立 22 维 (1981—2002 年)、21 维 (1982—2002 年)、...、10 维 (1993—2002 年) 共 13 个模型。按照图 1 所示的流程进行模型的学习率定, 并对未来的需水量进行预测, 同时, 对每个预测周期进行模型率定的详细过程参照 1.2 节。在 RBFANN 模型率定时, 隐含层节点数应该设计满足精度要求最小结构的神经网络, 以保证神经网络的泛化能力 [8]; 另外在应用聚类方法确定扩展长度时, 重叠系数 α 需要通过优选来确定。对比不同维数模型的预测结果,

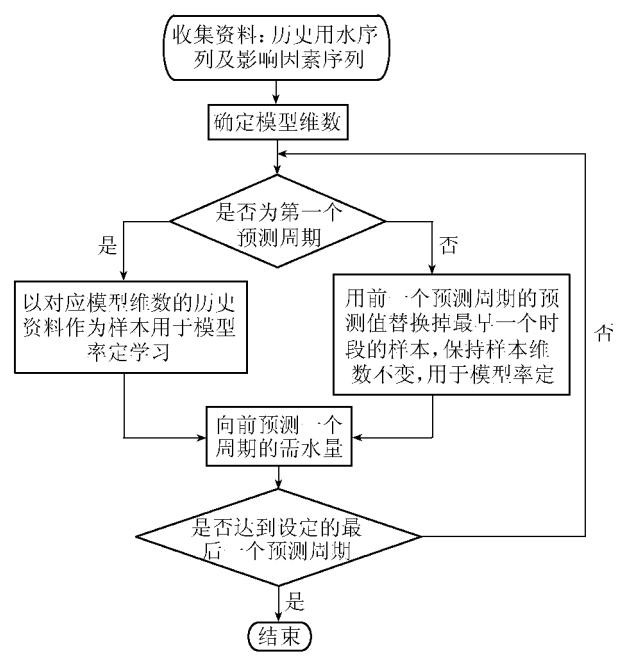


图 1 模型学习及预测流程

表 1 不同维数模型验证期预测结果

模型 维数/ 维	2003 年		2004 年		2005 年		2006 年		2007 年		平均绝对相对 误差/%
	预测值/ 亿 m ³	相对 误差/%	预测值/ 亿 m ³	相对 误差/%	预测值/ 亿 m ³	相对 误差/%	预测值/ 亿 m ³	相对 误差/%	预测值/ 亿 m ³	相对 误差/%	
22	17.61	-0.90	18.85	-2.11	21.31	-4.05	21.96	-1.45	22.36	1.37	1.98
21	17.68	-1.33	18.92	-2.48	21.27	-3.84	22.06	-1.89	22.80	-0.56	2.02
20	17.82	-2.09	18.76	-1.62	21.10	-3.03	21.25	1.86	22.40	1.19	1.96
19	17.79	-1.93	18.47	-0.03	21.74	-6.16	21.45	0.92	22.56	0.50	1.91
18	17.73	-1.61	18.53	-0.34	21.76	-6.27	21.47	0.83	22.63	0.17	1.84
17	17.85	-2.29	18.56	-0.51	21.32	-4.11	21.41	1.10	22.49	0.79	1.76
16	17.87	-2.41	18.55	-0.48	21.21	-3.58	21.52	0.59	22.58	0.40	1.49
15	17.67	-1.26	18.56	-0.55	21.61	-5.51	21.82	-0.79	22.63	0.18	1.66
14	17.82	-2.08	18.62	-0.87	21.74	-6.13	21.89	-1.09	22.72	-0.23	2.08
13	17.90	-2.54	18.63	-0.90	21.88	-6.81	21.65	-0.02	22.62	0.23	2.10
12	17.90	-2.56	18.69	-1.23	21.77	-6.29	21.58	0.34	22.74	-0.32	2.15
11	17.93	-2.74	18.79	-1.77	22.20	-8.41	21.26	1.78	22.65	0.07	2.95
10	17.71	-1.46	18.67	-1.12	21.70	-5.97	21.90	-1.15	22.85	-0.80	2.10

16 维 DEIRBFANN 模型的平均绝对相对误差最小 ,且该模型相对误差最大绝对值出现在 2005 年(3.58%) ,仅比 20 维的模型(3.03%) 大 ,但平均绝对相对误差要小很多。因此可以认为当维数为 16 时 DEIRBFANN 模型能够得到最高精度的预测结果(表 1)。

2.3 年需水量预测

以东莞市 1987—2002 年的资料作为模型率定资料 ,以年为预测时段 ,将 2003—2007 年东莞市年需水量作为未知值对模型进行验证 ,分别建立 16 维的基本 RBFANN 和 DEIRBFANN 模型 ,模型率定期和验证期的输出结果见表 2。

表 2 模型输出结果

年份	实际 用水量/ 亿 m ³	基本 RBFANN 模型		DEIRBFANN 模型	
		输出值/ 亿 m ³	相对 误差/%	输出值/ 亿 m ³	相对 误差/%
1987	8.621	8.552	0.80	8.796	-2.03
1988	8.952	8.975	-0.26	9.016	-0.72
1989	9.253	9.422	-1.82	9.301	-0.52
1990	9.742	9.607	1.38	9.704	0.39
1991	10.374	10.420	-0.45	10.208	1.60
1992	10.685	10.530	1.45	10.601	0.78
1993	11.255	11.049	1.83	11.183	0.64
1994	11.596	11.619	-0.20	11.459	1.18
1995	12.144	12.718	-4.73	12.785	-5.28
1996	13.857	13.900	-0.31	13.737	0.86
1997	14.736	14.777	-0.28	14.662	0.50
1998	14.962	15.085	-0.82	14.836	0.84
1999	15.157	15.095	0.41	15.231	-0.49
2000	16.298	15.617	4.18	15.929	2.26
2001	16.568	16.685	-0.70	16.500	0.41
2002	16.791	16.938	-0.87	17.043	-1.50
2003	17.452	17.091	2.07	17.452	2.07
2004	18.463	17.885	3.13	18.551	-0.48
2005	20.481	19.260	5.96	21.215	-3.58
2006	21.650	20.306	6.21	21.522	0.59
2007	22.670	21.077	7.03	22.579	0.40

在率定期基本 RBFANN 和 DEIRBFANN 模型输

出值绝对相对误差的平均值分别为 1.28% 和 1.23% ,绝对相对误差的最大值均出现在 1995 年 ,分别为 4.73% 和 5.28% ,均小于 10% ,能够满足预测精度要求 ,两种模型在率定期的相对误差差别不大 ,在验证期 ,基本 RBFANN 模型最大绝对相对误差为 7.03% ,平均绝对相对误差为 4.88% ,而 DEIRBFANN 模型的最大绝对相对误差为 3.58% ,平均绝对相对误差为 1.42% ,均远小于基本 RBFANN 模型 ,说明 DEIRBFANN 模型有更高的预测精度 ;基本 RBFANN 模型的绝对相对误差在验证期逐年增大 ,从而可以说明根据最初样本率定得到的模型参数来进行多周期预测时 ,其预测误差会随着周期的延长而累积 ,而且都呈现相同的趋势 ,即预测值比实测值都小且两者之间的差值越来越大。导致这种情况的原因是 :随着预测周期的延长 ,需水影响因素与需水量之间的驱动关系发生了改变 ,而基本 RBFANN 模型不能反映出这种变化 ,只能机械地应用最初的模型率定参数进行预测 ,从而导致预测误差越来越大 ;对于 DEIRBFANN 模型 ,其预测值在实测值上下波动 ,绝对相对误差也没有增大的趋势 ,说明 DEIRBFANN 模型通过不断引入新信息进入模型中对模型参数进行重新率定 ,有效地克服了基本 RBFANN 模型在进行多周期预测时体现出来的预测误差随预测周期延长而累积的缺点 ,即当需水影响因子与需水量之间的驱动发生改变时模型能够快速反应 ,通过在模型率定中纳入新信息而提高预测精度 ,在需水预测领域有较强的实际应用价值。

3 结 语

笔者在分析各种需水预测方法在多周期需水预测中的应用时发现 ,大多数预测方法都不能达到理想的精度 ,因此在应用 RBFANN 模型(下转第 16 页)