DOI:10.16616/j.cnki.10-1326/TV.2017.07.018

基于改进型极限学习机的新疆库尔勒市城市需水量预测

司马义·阿不都热合曼

(新疆塔里木河流域巴音郭楞管理局,新疆 库尔勒 841000)

【摘 要】 城市需水量系统具有大惯性、强耦合、非线性等特性,采用机理分析法,难以建立其准确的数学模型,导致预测效果差。鉴于此,该文将基于正交基函数的改进型极限学习机对城市需水量因子进行辨识,并利用经验模态分解方法确定网络隐含层节点数,建立了库尔勒市城市需水量预测模型。结果表明:模型有效性为0.9714,实测值与预测值的拟合关系比较理想,说明基于正交基函数的改进型极限学习机对城市需水量进行系统辨识是可行的。

【关键词】 城市需水量;预测;极限学习机;经验模态分解;正交基函数

中图分类号:TV211 **文献标志码:A** **文章编号:2096-0131(2017)07-0061-05**

Prediction of Xinjiang Kurle urban water demand based on improved limit learning machine

SIMA Yi·Abudureheman

(*Xinjiang Tarim River Basin Bayinguoleng Administration*, *Kolla* 841000, *China*)

Abstract: Urban water demand system is characterized by large inertia, strong coupling and nonlinearity, etc. Mechanism analysis method is adopted. It is difficult to establish an accurate mathematical model, thereby leading to poor forecast effect. Therefore, urban water demand factors are identified by the improved limit learning machine based on orthogonal basis functions in the paper. The empirical mode decomposition method is used for determining network hidden layer node quantity. An urban water demand prediction model in Korla is established. Results show that the model validity is 0.9714. The fitting relationship between the measured value and the predicted value is more ideal. It is obvious that it is feasible to systematically identify urban water demand by improved limit learning machine based on orthogonal basis functions.

Key words: urban water demand; prediction; limit learning machine; empirical mode decomposition; orthogonal basis functions

城市需水量预测对区域水资源规划、城市供水系统管理和改扩建具有重要指导作用[1-2]。由于城市需水量是一个典型的非线性、大惯性、强耦合和时变的复杂系统,因此,它的预测模型很难通过机理法用简单的数学公式或传递函数来描述[3]。随着人工智能技术的飞速发展,学者们开始利用人工智能技术解决需水量建模的相关问题,其中人工神经网络技术在建模中的应用特别突出[4-5]。人工神经网络可加快计算速度,并提高物理过程参数的精度,能利用有限的参数描述复杂的系统。人工神经网络建模相对于传统建模方法主要优点是不需要用数学表达式,更适合于长期预测。本文以新疆库尔勒市1997—2013年城市需水量及相

关影响因子数据为例,针对城市需水量具有的非线性时变等特性,利用基于正交基函数的极限学习机(extreme learning machine,ELM)对城市需水量因子进行辨识,并利用经验模态分解(empirical mode decomposition,EMD)方法确定网络隐含层节点数,研究结果对于实现区域需水量的精确预测提供了重要参考。

1 极限学习机(ELM)算法

ELM是一种单隐层前馈神经网络,与传统的神经网络的根本区别在于:ELM在训练中可随机产生输入权值和隐含层节点偏移量,且产生后保持不变,只需设置隐含层神经元个数就能获得全局最优解。因ELM学习速度快,泛化性好,故在函数逼近和模式分类方面得到广泛应用。ELM网络拓扑结构由输入层、隐含层和输出层组成。输入层有 n 个神经元,对应 n 个输入变量;隐含层有 L 个神经元;输出层有 m 个神经元,对应 m 个输出变量。ELM算法的主要步骤如下:ⓐ确定隐含层神经元个数,随机产生输入层和隐含层之间的连接权值 w_{ij} 与隐含层节点偏移量 b;ⓑ选取无限可导的函数作为隐含层神经元的激励函数 $g(x)$,计算隐含层输出矩阵 H;ⓒ计算输出权值 $\hat{\beta}$:$\hat{\beta} = H^{+}T$,其中 H^{+} 是矩阵 H 的Moore-Penrose广义逆,T 是系统实际输出。

但传统极限学习机算法存在如下缺陷:ⓐ隐含层激励函数是固定的;ⓑ隐含层节点数一般都是通过试凑法获得;ⓒ所求的输出权值 $\hat{\beta}$ 仅考虑风险最小化,即训练误差最小,但无法保证测试误差也达到最小值。

针对上述缺陷,本文提出基于正交基函数的改进型极限学习机,并利用经验模态分解方法确定网络隐含层节点数。

2 基于正交基函数的改进型极限学习机

2.1 正交基函数

依据最佳平方逼近多项式存在性定理[6,7],任意非线性函数 $y = f(x)$ 都可由一组正交基函数线性表示:

$$y = f(x) = \sum_{i=1}^{L} \omega_i \cdot g_i(x) + R(x) = W^T G(x) + R(x) \tag{1}$$

式中 $G(x)$——正交基函数;

W——相关系数;

余项 $R(x)$——逼近精度误差。

根据式(1),正交基函数神经网络的数学模型可定义为

$$\hat{y} = \hat{f}(x) = \sum_{i=1}^{L} \beta_i \cdot g_i(x) + R(x) = \beta^T G(x) + R(x) \tag{2}$$

式中 x——正交基函数神经网络的输入;

$\hat{y}$——相应神经网络的输出;

$\beta = [\beta_1, \beta_2, \cdots, \beta_L]$——隐含层神经元与网络输出层的连接权值。

基函数 $G(x) = [g_1(x), g_2(x), \cdots, g_L(x)]$ 作为隐含层神经元的激励函数,将网络输入层与隐含层神经元的连接权值设置为1,隐含层神经元与网络输出层的阈值也设置为0。常见的正交基函数有Chebyshev、Hermite和Fourier[8-9],本文选用Fourier正交基函数作为ELM隐含层神经元的激励函数,具体如下:

$$\begin{aligned} g_1(x) &= \cos\left[\frac{\pi}{q}(x-p)\right], g_2(x) \\ &= \cos\left[2\frac{\pi}{q}(x-p)\right], \cdots, g_L(x) \\ &= \cos\left[L\frac{\pi}{q}(x-p)\right] \end{aligned}$$

式中 x——正交基函数神经网络的输入;

p, q——Fourier正交基函数的系数。

2.2 EMD算法

经验模态分解算法[10]是一种基于时域的信号处理方法,它仅仅基于这样的假设:任何信号都是由不同的本征模态函数(intrinsic mode function,IMF)组成,其目的是把复杂的信号分解成有限个本征模态函数之和。

每一个本征模态函数都具有相同的极值点和过零点,在任意两个相邻的过零点之间仅有一个极值点,且上下包络曲线是关于时间轴局部对称,任意的两个本征模态函数之间是相互正交的。本文利用经验模态分解方法确定网络隐含层节点数。

2.3 网络隐含层节点数确定

目标函数 Y 经过 EMD 分解后得 n 个相互正交的分量 $c_1,c_2,\cdots,c_n$，由 EMD 的完备性可知：

$$Y = \sum_{i=1}^{n} c_i \tag{3}$$

式(3)左乘 c_i^T 后得：

$$c_i^T Y = c_i^T \sum_{i=1}^{n} c_i,\ i = 1,2,\cdots,n \tag{4}$$

于是得：

$$Y^T Y = \sum_{i=1}^{n} \| c_i \|^2, i = 1,2,\cdots,n \tag{5}$$

$$1 = \frac{\sum_{i=1}^{n} \| c_i \|^2}{Y^T Y}, i = 1,2,\cdots,n \tag{6}$$

含有 L 个隐含层节点的 ELM 网络的数学模型可表示为

$$Y = \sum_{i=1}^{L} \beta_i \cdot g_i(w_i,b_i,x) \tag{7}$$

式中 $x \in R^n, w_i \in R^n, \beta_i \in R^m$；

w_i——网络输入层与第 i 个节点的连接权值；

b_i——第 i 个隐含层节点的阈值；

β_i——第 i 个隐含层节点与输出层的连接权值；

$g_i(w_i,b_i,x)$——第 i 个隐含层节点的激励函数。

故具有 L 个隐含层节点的能量总贡献率可表示为

$$V = \sum_{i=1}^{L} v_i = \frac{\sum_{i=1}^{L} \beta_i \| g_i(w_i,b_i,x) \|^2}{Y^T Y} \tag{8}$$

式中 V——能量总贡献率，且 $0 \leqslant V \leqslant 1$；

v_i——第 i 个隐含层节点的能量贡献率，且 $0 \leqslant v_i \leqslant 1$。

选择的隐含层节点数量越多，能量总贡献率越高，逼近精度越高，若 $L = n$，则逼近精度 $V = 1$。

本文对训练样本(1997—2007 年)的城市需水量进行 EMD 分解，分解所得的 IMF 分量个数为 9，故网络隐含层节点数为 9。

2.4 网络输出权值确定

根据统计学习理论可知，实际的风险既有经验风险又有结构化风险。如果想获得一个良好的模型，需要同时考虑这两种风险。因此在输出权值最小化和误差最小化之间做出折中，即：

$$\min\{ \| H\beta - T \|^2 + \| \beta \|^2 \}$$

式中 H——隐含层输出矩阵；

$\beta = [\beta_1,\beta_2,\cdots,\beta_L]$——隐含层与输出层的连接权值；

T——系统实际输出。

3 模型应用

3.1 研究区概况及数据来源

库尔勒市位于新疆塔里木盆地北缘，天山支脉库鲁克塔格山和霍拉山山前冲积平原，全市总面积 7268km^2，是新疆巴音郭楞蒙古自治州的州府所在地，也是当地政治、经济与文化中心。库尔勒市因盛产库尔勒香梨而有“梨城”美誉，这里光照充足，昼夜温差大，降水稀少，蒸发强烈，为典型的暖温带大陆性干旱气候，多年平均气温 11.5℃，降水量 55.6mm，蒸发量 2388.2mm(E20 小型蒸发器)，年日照时数 2990h，无霜期 210d。

库尔勒市 1997—2013 年城市需水量及相关影响因子数据均来自《巴音郭楞蒙古自治州 2014 年统计年鉴》，考虑数据的可获取性，城市年需水量及相关影响因素值见表 1。

表 1 新疆库尔勒市 1997—2013 年需水量及相关数据

年 份	城市需水量/万 m^3	生活需水量/万 m^3	人均日生活用水量/L	用水人口/人	GDP/万元	人均 GDP/元	第二产业 GDP/万元	第三产业 GDP/万元
1997	1247	910	151.95	10.24	723928	21389	523511	143125
1998	1312	1012	155.47	10.3	707400	20243	480365	166244
1999	1296	1054	153.68	10.73	733436	20329	510249	161316
2000	1327	1145	154.9	11.5	974299	26016	721456	183429

续表

年　份	城市需水量/万 m^3	生活需水量/万 m^3	人均日生活用水量/L	用水人口/人	GDP/万元	人均 GDP/元	第二产业 GDP/万元	第三产业 GDP/万元
2001	1369	1095	208	11.5	1038348	26686	745744	216451
2002	1595	1225	177.11	18.65	1080649	26614	746167	252554
2003	1612	1095	127.98	23.44	1385413	32552	978908	292016
2004	1901	1143	123.97	25.26	1702132	38414	1242501	336712
2005	3546	1642	168.55	32.16	2428032	52862	1924873	361050
2006	3034	1316	185.74	21.03	3152175	66678	2589573	403793
2007	4395	1403	185.38	26.27	3538604	72074	2877934	474525
2008	4289	1423	188.6	24.8	4464949	88502	3717496	572093
2009	4677	1483	185.52	31.24	3557023	66574	2763699	589222
2010	5220.64	1657.91	180.3	29.28	4392402	80340	3439955	656504
2011	5317.39	1641.44	183.99	31.01	5582944	99094	4497423	767331
2012	5649.35	1802.45	187.05	32.05	5914147	103336	4692684	867467
2013	6531.14	1720.79	197.76	33.91	6534087	110980	5154560	999960

3.2　模型评价指标

模型性能以均方根误差 RMSE(root mean square error)和模型有效性 MV(model validity)为指标,来衡量模型的泛化能力和精度。

均方根误差 RMSE 表示为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{N}(t_i-\hat{y}_i)^2}{N}} \tag{9}$$

模型有效性 MV 表示为

$$\mathrm{MV}=1-\frac{\sum_{i=1}^{N}(t_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(t_i-\bar{y}_i)^2} \tag{10}$$

式中　t_i——ELM 网络模型输出;

$\hat{y}_i$——实际值;

$\bar{y}$——实际值的平均值;

N——样本数。

均方根误差 RMSE 反映模型输出曲线在实际曲线上的波动情况,模型有效性 MV 反映了模型输出与实际值的偏差相对于数据的离散性,性能良好的模型有效性 MV 为 1。

3.3　模型预测结果分析

以库尔勒市 1997—2007 年城市需水量相关数据为训练样本,以 2008—2013 年城市需水量相关数据为预测样本,模型计算结果如下。

3.3.1　传统 ELM 模型

传统 ELM 算法在建立模型前仅需确定网络激励函数和隐含层节点数即可,本文选用 Sigmoidal 函数作为传统 ELM 算法的激励函数,网络隐含层节点数为 9。基于传统 ELM 网络的城市年需水量预测曲线如图 1 所示,误差范围为-558 万~406 万 m^3。

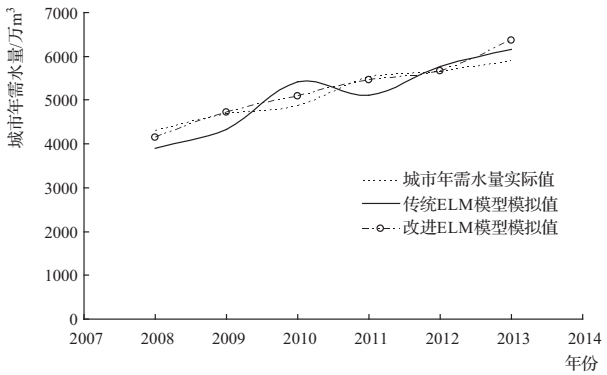

图 1　改进 ELM 模型与传统 ELM 模型的城市年需水量预测

3.3.2 基于正交基函数的改进 ELM 模型

在基于正交基函数的改进 ELM 模型中，其隐含层神经元激励函数系数为：$p=25$，$q=12$，网络隐含层节点数为 9。基于正交基函数的改进 ELM 网络的城市年需水量预测曲线如图 1 所示，其误差范围 -136 万 ~ 171 万 m^3。两种不同的 ELM 模型性能对比见表 2。

表 2 基于正交基函数的改进型 ELM 与传统的 ELM 性能对比

模 型	城市年需水量	
	均方根误差/万 m^3	模型有效性
传统 ELM 模型	408.3	0.7846
改进 ELM 模型	130.2	0.9714

根据表 2 可知，与传统 ELM 模型相比，运用基于正交基函数的改进 ELM 模型的城市需水量均方根误差减小了 278.1 万 m^3，模型有效性相对提高了 0.1868，这说明运用基于正交基函数的改进 ELM 模型对城市需水量进行模拟预测是行之有效的。但是城市需水量误差波动剧烈，离散性大，这是由于需水量受到各种随机因素的影响，并且数据中可能含有噪声，造成 ELM 算法过多的拟合了噪声，最终造成城市需水量的误差波动剧烈，离散性大。

4 结 论

本文提出利用基于正交基函数的改进极限学习机对城市年需水量进行预测。首先对训练样本进行 EMD 分解，根据 IMF 分量的个数确定网络隐含层节点数目；接着，在统计学习理论的基础上，同时考虑经验风险与结构化风险，在输出权值最小化和误差最小化之间做出折中，求解出满足输出权值与误差之和最小化的网络输出权值计算公式。结果表明，与传统 ELM 模型相比，运用基于正交基函数的改进 ELM 模型对城市需水量进行模拟预测是行之有效的。

但是，选用 Fourier 正交基函数作为 ELM 隐含层神经元的激励函数，激励函数中的系数为经验值，需要经过多次试算才能找到最佳值，其他正交基函数有待验证。

参考文献

[1] 杨艳，李靖，马显莹. 基于小波神经网络的城市用水量长期预测研究[J]. 云南农业大学学报：自然科学版，2010，25(2)：272-276.

[2] 梁学玉，张鑫，孙天青. 组合灰色预测模型在城市用水量预测中的应用[J]. 人民黄河，2010，32(4)：79-80.

[3] 李黎武，施周. 基于小波支持向量机的城市用水量非线性组合预测[J]. 中国给水排水，2010，26(1)：54-58.

[4] 王军玺，刘严如，李月娇. 基于区域发展目标的兰州市水资源可持续承载能力预测研究[J]. 水利建设与管理，2013(3)：81-86.

[5] 宓永宁，陈默，张茹. 灰色拓扑法在大伙房水库总氮预测中的应用[J]. 水利建设与管理，2009(3)：72-73.

[6] 曾喆昭. 神经计算原理及其应用技术[M]. 北京：科学出版社，2012：181-185.

[7] 孟广伟，李广博，李锋. 多项式基函数神经网络的结构可靠性分析[J]. 北京航空航天大学学报，2013，39(11)：1460-1463.

[8] 叶军. 基于模拟正交神经网络的电热干燥器温度控制[J]. 农业工程学报，2005，21(10)：105-108.

[9] 杨胡萍，左士伟，涂雨曦. 基于混沌理论和 Legendre 正交基神经网络的短期负荷预测[J]. 电测与仪表，2015，52(13)：63-67.

[10] 张瑜，汪小昂，孙国祥，等. 基于集合经验模态分解与 Elman 神经网络的线椒株高预测[J]. 农业工程学报，2015，31(18)：169-174.