doi:10.11835/j.issn.1674-4764.2016.S2.027

综合法优选局域邻近点的供水量预测模型

刘年东,周明,孙晓婷,杜坤
(昆明理工大学 建筑工程学院, 昆明 650500)

摘　要:基于混沌理论的局域法是目前较先进的预测技术,对于非线性、非平稳特征较强的城市日供水系统的预测,通常混沌局域法更适用。混沌局域法的日供水量预测,最重要的就是对历史信息的利用,即对邻近点的选择和利用,但在运用混沌局域法对城市日供水量预测时却鲜有人考虑邻近点选取问题,或仅考虑邻近点个数与邻近点位置中的一个,很少没有人将 2 种情况综合考虑的研究。据此,运用 BIC 信息准则和衰减系数相结合对相空间重构进行数据挖掘,优选混沌局域法的邻近点,并用该模型对实际水厂日供水量进行了预测。结果表明,运用该方法能显著提高日供水量预测精度,预测平均绝对误差仅为 1.06%,说明该综合法是可行的。

关键词:混沌理论;邻近点;BIC 信息准则;衰减系数;供水量预测
中图分类号:TP183　　**文献标志码:**A　　**文章编号:**1674-4764(2016)S2-0142-05

Study on select mining adjacent point of the local-region method water supply forecasting model

Liu Niandong, Zhou Ming, Sun Xiaoting, Du Kun
(Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, P. R. China)

Abstract: Based on chaos theory local-region method is more advanced forecasting technology, for the prediction of urban daily water supply system with strong nonlinear and non-stationary characteristics, usually chaotic local method is more applicable. for chaos local-region method of daily water supply capacity prediction, most important is the use historical information, namely on adjacent points selection and use, but few people consider the adjacent points selection in the urban daily water supply forecast by using the method of local-region method, or only consider the number of adjacent points, or a position adjacent, no researcher combined the two cases. Accordingly, the combination method of the BIC information criterion and the attenuation coefficient mining data for phase space reconstruction, optimize adjacent points of the local-region method, and the model was used to predict the actual daily water supply of the water plant, the results show that the method can significantly improve the prediction accuracy of daily water supply, the average absolute error of the prediction is only 1.064%, it shows that the combination method is feasible.

Key words: chaos; adjacent points; BIC Information criterion; attenuation coefficient; daily water supplies forecast

收稿日期:2016-10-18
基金项目:昆明理工大学 2016 年学生课外学术科技创新基金(2015YB025);国家自然科学基金(51608242);云南省人才培养计划项目 (14118943)
作者简介:刘年东(1990-),男,硕士生,主要从事市政工程研究,(E-mail) 2896957112@qq.com。
杜坤(通信作者),男,博士,(E-mail) 250977426@qq.com。

供水量预测工作是优化调度环节的重要组成部分,其准确预测与否直接影响给水系统调度决策能否是有效制定[1-2]。供水系统是一个非平稳,非线性的复杂系统,传统预测模型[3]多基于全局建模的角度,缺失局部特征的代表性,例如回归分析法、指数平滑法、趋势外推法等[4],对供水系统趋势把握较差,无法获得较好的预测效果。基于时序特性驱动的混沌理论分析法为供水量预测提供了新的思路,国内外学者对此开展了广泛研究,例如:Jayawardena[5]提出基于广义自由度的新准则,以此来确定混沌预测模型的邻近点个数。Bai 等[6]分析了供水序列的混沌特性,利用自适应混沌粒子群优化 RVM 模型参数,提出一种多尺度的 RVM 用水量预测模型。刘年东等[7]利用 BIC 信息准则优选了混沌局域法邻近点个数,提高了模型预测精度。总之,基于混沌理论的时序预测具有广阔的应用前景。

预测模型大致可概括为 3 个步骤:1)采用何种拟合样本形式能更准确地描述原始数据的时变特性;2)选择多少个样本作为参考样本;3)采用何种模型进行预测运算。笔者发现对于混沌局域法的预测,最重要的就是对历史信息的利用,即对邻近点的选择。但在运用混沌局域法对日供水量进行预测时却鲜有人考虑邻近点的选取问题,多数学者仅凭经验选取 $K(K=m+1)$个邻近点作为参考样本,并未对相空间数据进行进一步的数据挖掘。本文在参考大量文献的基础上,提出 BIC 信息准则和衰减系数法结合的方法来挖掘相空间重构中的数据,优选局域法的邻近点,并利用实测供水量数据验证了该方法提高预测精度有效性。

1 混沌局域法

混沌局域法将相空间运行轨迹的最后一点作为中心点,通过寻找历史相点中最近的若干相点作为运动趋势参考点,预测中心点走向,其步骤可概括为:1)重构相空间,包括嵌入维与时间延迟的确定;2)邻近点选取;3)模型预测。

1)重构相空间[8]是将一维混沌时间序列映射到高维的空间,目的是恢复有规律的吸引子,从而使蕴藏在时间序列中的信息显露出来。

设有混沌时间序列$\{x_1,x_2,\cdots,x_N\}$,则相空间重构为

$$X(t_i)=[x(t_i),x(t_i+2\tau),\cdots,x(t_i+(m-1)\tau)]$$
$$(i=1,2,\cdots N-(m-1)\tau) \quad (1)$$

式中:τ 为时间延迟;m 为嵌入维数;τ 与 m 的取值决定了相空间形态,即拟合样本形式。

2)选取邻近点。选取邻近点的方法通常有两种[9],即固定邻近点个数法和固定邻域半径法。固定邻域半径法是指选取落在以相空间运行轨迹的最后一点 X_M 为中心点,半径取固定值 R 的超球内的点为邻近点,即对邻近点 X_{Mi} 有

$$\| X_M-X_{Mi} \|_2 \leqslant R \quad i=1,2,\cdots q \quad (2)$$

固定邻近点个数法是将中心点 X_M 与所有的相点欧式距离算出,取固定的 K 个最小欧式距离相点为邻近相点,即设中心点 X_M 的到邻近点距离的集合为$\{X_{Mi},i=1,2,\cdots q\}$,$\{X_{Mi}\}$中每个点到中心点 X_M 的欧式距离为 d_i,设 d_i 中的最小值为 d_{min},则

$$d_i=\| X_M-X_{Mi} \|_2 \quad (3)$$

由式(2)、(3)可知 R 和 K 的取值不同,邻近点个数也就不同,本文着重研究固定邻近点个数法。

局域预测模型由 3 个参数决定,即嵌入维数 、延迟时间与邻近点个数。局域法的首要问题就是如何选取中心点 X_M 的邻近点。邻近点个数影响局域模型的预测精度和计算量,对于局域线性预测法,若邻近点的个数 K 取的太小,那么就有可能无法充分利用历史信息,导致大量的有用信息被忽略。但若邻近点的个数 K 取的太大,那么局域线性模型的线性假设条件将不满足;而且邻近点过多时,将引入包括与中心点所在轨道相距较远的轨道上的那些点,由于混沌系统本身所存在的指数发散的性质,那些点加入到模型中以后,将降低模型预测精度。综上所述,对于局域线性预测法,在满足预测精度较高的条件下,邻近点应适中。

3)模型预测。由邻近点距构成的矩阵为$\{X_{Mi}\}$,邻近点的下一步演化点距矩阵为$\{X_{Mi+1}\}$,则基于混沌理论的一阶局域预测模型为

$$X_{Mi+1}=ae+bX_{Mi} \quad (4)$$

式中:$e=[1\ 1\ \cdots 1]^{\mathrm{T}}$,$a$ 和 b 为拟合参数。若要计算中心相点的第 S 步预测值,采用邻近点的第 S 步演化点作为参考点即可,即将上述模型中的 X_{Mi+1} 替换为 X_{Mi+S}。

采用加权一阶局域预测模型时,其权系数可根据邻近点到中心点的欧式距离确定,令 d_i 中的最小值为 d_{min},定义$\{X_{Mi}\}$的权值为 P_i,则

$$P_i=\frac{\mathrm{e}^{-a(d_i-d_{\min})}}{\sum \mathrm{e}^{-a(d_i-d_{\min})}} \quad (5)$$

式中:α 为权重调节系数,一般 $\alpha=1$。根据加权最小

二乘法，a、b 能计算为

$$[a \quad b]^{\mathrm{T}} = (\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{Y} \tag{6}$$

式中：$\boldsymbol{Y}=\boldsymbol{X}_{Mi+1}$；$\boldsymbol{A}=[\mathrm{e}\ X_{Mi}]$；权重矩阵 $\boldsymbol{P}$ 为对角矩阵，对角元素为 P_i。在求解出参数 a、b 后带入式(4)即可得预测值。

2 混沌局域法邻近点的优选

在相空间重构中，邻近的相点具有相似的演化行为，预测点的运动趋势是通过各邻近点的运动趋势推断而来，但是并不是所有的邻近点都能真实反映相点的运动趋势[10]。在确定邻近点个数时，大多数学者往往仅凭经验的确定邻近点个数 K，这样就可能引入一些“弱相关点”，导致预测模型外推误差增大，即出现过拟合的现象[11]。在确定邻近点位置时，常采用欧式法计算距离，但欧式法不能完全反映最邻近点与预测中心点之间的相关性，容易引进“伪邻近点”[12]。鉴于此，笔者使用 BIC 信息准则确定邻近点个数，采用调节衰减系数(即 Lyapunov 指数 λ)计算邻近点位置，对供水量局域预测模型的相空间重构中数据进一步挖掘，优选邻近点，使其能更好地反映供水系统的运动趋势。

2.1 BIC 信息准则选取邻近点个数

信息准则建立在信息熵的概念基础之上，鼓励数据拟合的精度并可避免出现过拟合的情况，为解决过拟合问题提供了途径。日本数学家 Akaike 在贝叶斯原理的基础上提出了 AIC 信息准则的贝叶斯改进形式，简称 BIC 信息准则[13-14]，它所确定的阶数是真阶的相容估计。

$$\mathrm{BIC}(p,q) = \ln\sigma^2 + \frac{(p+q+1)\ln N}{N} \tag{6}$$

式中：σ^2 为拟合方差；N 为拟合数据个数。S 表示预测步距，将时间序列样本均匀地分成 L 段，则得拟合数据个数 $N=L\cdot S$。等式(6)中 $\ln\sigma^2$ 代表模型拟合精度，$(p+q+1)\ \ln N\ /\ N$ 代表模型的复杂度，当 $\mathrm{BIC}(p,q)$取得最小值时，认为模型在精度和复杂度之间取得最佳均衡，不会出现过拟合现象。

2.2 衰减系数法确定邻近点位置

由文献[15]知，相空间重构后，预测值为预测矢量的最后一个分量，它与邻近点的最后分量的相关性最大，与其它分量的相关性依各分量对应的延迟时间呈 Lyapunov 指数衰减。但是欧式法将同一邻近点的所有分量看成是“平等的”进行计算，只是反映邻近点与中心点的距离远近，难以反映运动轨迹的相关性，导致引入“伪邻近点”。孟庆芳等[16]提出减少拟合分量，只用延迟矢量的第一个分量进行一阶拟合，去除其他干扰分量，但在混沌性较弱时可能忽略大量可利用的有效信息。而通过调节衰减系数 Lyapunov 指数的大小，可调节同一邻近点的各分量与中心点各分量的相关程度，使得每个邻近点对应的权值能更好地体现出对预测的贡献，进而提高预测性能。据此，本文利用最大 Lyapunov 指数和邻近点各分量所对应的延迟时间的乘积作为幂，构造一个指数形式的衰减因子 λ，修改了邻近点位置的确定公式，将各分量对预测的贡献通过权值体现出来：

$$d_i = \left[\sum_{j=1}^{m} \mathrm{e}^{-\lambda(m-j)^{\tau}}(x_{MJ}^{j} - x_{M}^{j})^2\right]^{\frac{1}{2}} \tag{7}$$

式中：d_i 为中心点 的第 j 个分量； 为第 i 个邻近点的第 j 分量；λ 为最大 Lyapunov 指数。依据修正后的公式所计算出的矢量间欧式距离，判别了不同邻近点与中心点的相关性，进而提高预测精度。

3 实例分析

某市水厂自 2000 年 1 月至 2006 年 12 月的日供水系统是一种混沌系统。根据文献[17]，为消除年供水量时间序列的季节性和趋势性、减少噪声影响，仅选取 2000—2006 年每年 1 月的日供水量数据作为单独的时间序列进行预测，对于其它月份的预测可按此方法进行处理。本文用互信息法[18]计算得时间延迟 $\tau=7$，用文献[19]方法得嵌入维 $m=15$ 则邻近点法个数为 $K=16$(传统邻近点个数计算方法为 $K=m+1$)。

依据 BIC 信息准则选取邻近点个数：取邻近点个数为 K，设 K 取值范围为 $[K_{\min},K_{\max}]$，依次计算每个 K 值下的 BIC 信息准则值，当 $BIC(K)$取得最小值时即为邻近点个数：

$$\mathrm{BIC}(K) = \ln\sigma^2 + \frac{K\ln N}{N} \tag{8}$$

得本文时序的邻近点个数为 $K=7$，见图 1。

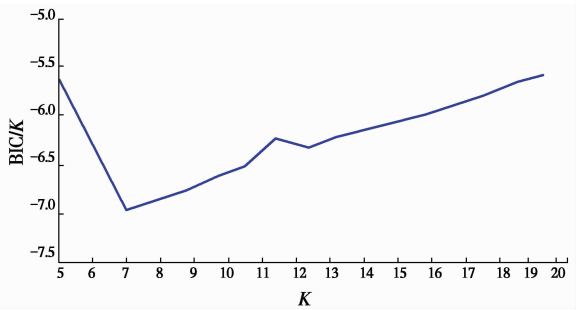

图 1 基于 BIC 信息准则的邻近点个数选择

由 Lyapunov 指数法计算得衰减系数 $\lambda=0.021$，利用公式(7)则可得邻近点位置。

利用混沌局域法预测模型对供水量进行预测，其中 210 个日供水量数据作为参考样本，7 个日供水量数据作为验证样本。总体预测情况见图 2，局部细节见图 3。

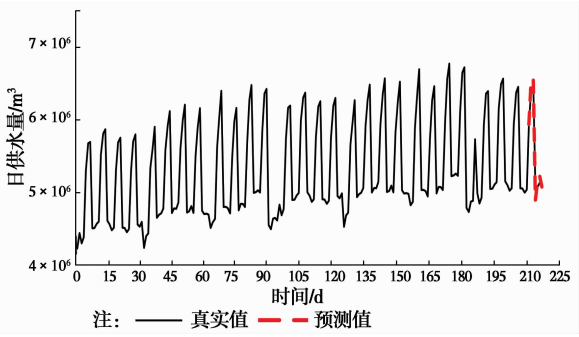

图 2　日供水量总体预测趋势

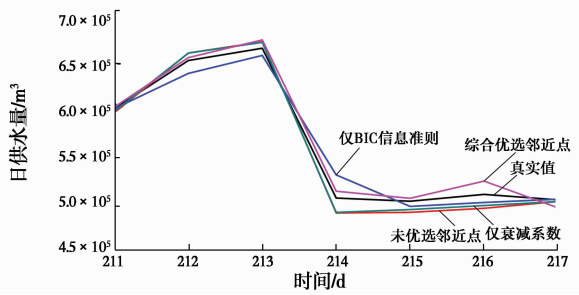

图 3　日供水量局部预测细节

如图 2、3 所示，4 种方法都能较好预测日供水量总体趋势，但对局部细节预测有一定差异。将本文所提“综合优选邻近点”方法与“未优选邻近点”、“仅 BIC 信息准则”及“仅衰减系数”优选邻近点的供水量预测结果进行比较，预测结果的分步相对误差、平均绝对百分比误差(MAPE)如表 1 示。

表 1　预测精度比较　　%

邻近点数	分步相对误差 P_i				MAPE
	211 步	212 步	213 步	214 步	
未优选邻近点	−0.61	1.09	0.92	−2.68	
仅 BIC 信息准则	0.01	−1.86	−1.02	4.17	
仅衰减系数	−0.42	1.08	0.82	−2.56	
综合优选邻近点	0.25	0.43	1.19	1.23	
邻近点数	分步相对误差 P_i				MAPE
	215 步	216 步	217 步		
未优选邻近点	−2.01	−2.49	−0.31		1.45
仅 BIC 信息准则	−0.95	−1.45	0.10		1.37
仅衰减系数	−1.51	−1.99	−0.31		1.24
综合优选邻近点	0.53	2.37	−1.45		1.06

由表 1 知，对比相同预测步数的相对误差，发现传统邻近点个数确定法 $K=16$ 比 BIC 信息准则的 $K=7$ 邻近点个数要多，但 K 值的增加不但没有提高预测精度，反而增加了误差，说明邻近点的个数并不是越多越好，需要考虑过拟合问题；在对比“未优选邻近点”与“仅衰减系数”时发现，后者每一步预测精度都提高了，MAPE 提高了 0.21%，说明“仅衰减系数”剔除了伪邻近点。“仅衰减系数”和“仅 BIC 信息准则”优选邻近点都较“未优选邻近点”的预测精度有所提高，将二者结合后预测精度较单一优选邻近点方法又进一步提高，且预测精度更稳定。

4　结论

提出了将 BIC 信息准则与衰减系数法相结合进一步挖掘相空间重构数据，优选局域法邻近点的新方法，通过应用预测城市日供水量实例证明其可有效提高预测精度。取得结论如下：

1)“仅衰减系数”和“仅 BIC 信息准则”，这种单一优选邻近点法较“未优选邻近点”的预测精度有所提高。

2)实例结果证明，本文所提方法比“未优选邻近点”、“仅 BIC 信息准则”与“仅衰减系数”方法使预测精度的提高更明显，证明了该方法的有效性。

3)衰减系数和 BIC 信息准则的结合是一种有效的数据挖掘方法，能辨别、剔除伪参考样本，获取有效参考样本提高预测精度，在噪声较大的供水系统中会更加明显。

4)值得说明的是，邻近点的优选除利用 BIC 信息准则与衰减系数法结合外，还可用 BIC 信息准则和演化追踪法结合等方法来挖掘出最优邻近点，值得进一步去研究。

参考文献：

[1] YASAR A, BILGILI M, SIMSEK E. Water demand forecasting based on stepwise multiple nonlinear regression analysis [J]. Arabian Journal for Science and Engineering, 2012, 37(8): 2333-2341.

[2] 徐勇鹏，王冬，王媛，等. 供水厂节水优化现状调查与分析[J]. 土木建筑与环境工程，2013，35(Sup2)：45-48.

[3] YUN B, PU W, CHUAN L, et al. Dynamic forecast of daily urban water consumption using variable-structure support vector regression model [J]. Journal of Water Resource Planning and Management, 2014,

140：1943-5452.

[4] 邓宏艳，王成华. 非线性组合模型在库岸边坡地下水位预测中的应用[J]. 土木建筑与环境工程，2010，32(1)：31-35.

[5] JAYAWARDENA A W. Neighbourhood selection for local modelling and prediction of hydrological time series [J]. Journal of Hydrology，2002，258：40-57.

[6] BAI Y，WANG P，LI C，et al. A multi-scale relevance vector regression approach for daily urban water demand forecasting [J]. Journal of Hydrology，2014，517：236-245.

[7] 刘年东，杜坤，周明，等. 局域法邻近点选取对降雨量预测精度影响研究[J]. 给水排水，2016，42：289-292.

[8] TAKENS F. Determining strange attractors in turbulence [J]. Lecture notes in Mathematics，1981，898：361-381.

[9] 韩敏. 混沌时间序列预测理论与方法[M]. 北京：中国水利水电出版社，2007.

[10] 高俊杰. 混沌时间序列预测研究及应用[D]. 上海：上海交通大学，2013.

[11] 孟庆芳，彭玉华. 混沌时间序列改进的加权一阶局域预测法[J]. 计算机工程与应用 2007，43(35)：61-64.

[12] 唐巍，谷子. 基于相关邻近点与峰谷荷修正的短期负荷时间序列预测[J]. 电力系统自动化，2006. 30(14)：25-29.

[13] AKAIKE H. A new look at the statistical model identification[J]. Automatic Control IEEE Transactions on，1974，19(6)：716-723.

[14] AKAIKE H. A Bayesian analysis of the minimum AIC procedure[J]. Annals of the Institute of Statistical Mathematics，1978，30(1)：9-14.

[15] 王振朝，赵晨，张士兵，等. 用于混沌时间序列预测的分维指数加权一阶局域算法[J]. 电测与仪表，2010，47(5)：12-15.

[16] 孟庆芳，彭玉华，曲怀敬，等. 基于信息准则的局域预测法邻近点的选取方法[J]. 物理学报，2008，57(3)：1423-1430.

[17] 张善文，雷英杰，冯有前. MATLAB 在时间序列分析中的应用[M]. 西安：西安电子科技大学出版社，2007.

[18] NA S H，JIN S H，KIM S Y，et al. EEG in schizophrenic patients：mutual information analysis. [J]. Clinical Neurophysiology Official Journal of the International Federation of Clinical Neurophysiology，2002，113(12)：1954-1960.

[19] BOSQ D，GUÉGAN D，LÉORAT G. Statistical estimation of the embedding dimension of a dynamical system [J]. International Journal of Bifurcation & Chaos，2011，9(4)：645-656.

（编辑 郭飞）