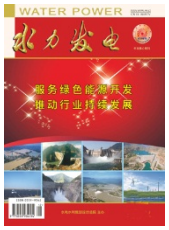

水力发电
Water Power
ISSN 0559-9342,CN 11-1845/TV

《水力发电》网络首发论文

题目：基于高斯过程回归模型的径流短期预测研究
作者：黄亚，易灵，肖伟华，侯贵兵，李媛媛
收稿日期：2020-04-19
网络首发日期：2020-10-15
引用格式：黄亚，易灵，肖伟华，侯贵兵，李媛媛．基于高斯过程回归模型的径流短期预测研究．水力发电.
https://kns.cnki.net/kcms/detail/11.1845.TV.20201015.1022.002.html

网络首发时间：2020-10-15 11:13:51
网络首发地址：https://kns.cnki.net/kcms/detail/11.1845.TV.20201015.1022.002.html

基于高斯过程回归模型的径流短期预测研究

黄　亚[1,2]，易　灵[1]，肖伟华[2]，侯贵兵[1]，李媛媛[1]
（1. 中水珠江规划勘测设计有限公司，广东　广州　510610；
2. 中国水利水电科学研究院流域水循环模拟与调控
国家重点实验室，北京　100038）

摘　要：为寻求更为精确的径流量预测方法，利用传统 BP 神经网络、支持向量机（SVM）以及高斯过程回归（GP）三种模型对径流量进行预测研究，并以广西天峨水文站日入库径流量为例进行预测实践和分析。结果表明，高斯过程回归模型对径流短期预测具有较高精度，预测平均相对误差绝对值为 1.29%，最大相对误差绝对值为 2.71%，预测精度和泛化能力均优于传统 BP 神经网络模型和支持向量机模型，是进一步提高径流预测精度的有效方法。
关键词：径流预测；高斯过程回归；BP 神经网络；支持向量机

Short-term Runoff Prediction Based on Gaussian Process Regression Model

HUANG Ya[1,2], YI Ling[1], XIAO Weihua[2], HOU Guibing[1], LI Yuanyuan[1]
(1. China Water Resources Pearl River Planning, Surveying & Designing Co., Ltd., Guangzhou 510610, Guangzhou, China;
2. State Key Laboratory of Simulation and Regulation of Water Cycle in River Catchment,
China Institute of Water Resources and Hydropower Research, Beijing 100038, China)

Abstract: To seek a more reasonable runoff prediction method, three kinds of models including traditional BP neural network, Support Vector Machine (SVM) and Gaussian process regression (GPR) are used to predict runoff, and the average daily runoff of Guangxi Tiane Hydrological Station is used as study case for runoff prediction and analysis. The results show that the short-term prediction of runoff based on GPR is more accurate, and its absolute value of average relative prediction error is 1.29% and the maximum absolute value relative error is 2.71%. The prediction accuracy and generalization ability of GPR is superior to the traditional BP neural network and SVM. It is one of the effective methods to improve the prediction precision of runoff.
Key Words: runoff prediction; Gaussian process regression; BP neural network; Support Vector Machine

中图分类号：TV12　　文献标识码：A

0　引　言

径流预测一直是人们关注的重大问题。传统方法如时间序列法[1]、回归分析法[2]、模糊分析法[3]、小波分析法[4]、集对分析法[5]以及灰色预测法[6]等均在径流预测预报中取得了一定的成效；但是由于河川径流受众多因素的相互作用、相互影响，具有显著的非线性、高维性、混沌性、模糊性等诸多复杂特征，使上述方法的预测精度受到不同程度的影响[7-8]。近年来，一些学者将神经网络（ANN）、支持向量机(SVM)等机器学习方法引入径流预测研究领域，取得了不少有价值的研究成果[9-10]。但是，这些方法本身仍存在着许多公开的问题，如：神经网络存在着最优网络结构难以确定和过度拟合的问题,且单层神经网络收敛速度慢，容易陷入局部极值[11]；SVM 计算时间受样本数量影响明显，核函数以及惩罚因子的选取经验性较强，对预测结果影响较大等[12]。为此，探讨新的更为精确径流量预测方法显得日益重要。

高斯过程(Gaussian Processes，GP)作为一种新的机器学习技术在多个领域的应用研究引起了相关学者的密切关注[13-16]。它有严格的统计学理论基础，对处理高维数、小样本、非线性等复杂回归问题具有良好的适应性，而且还具有参数自适应获取和预测结果具有概率意义等优点[17]。GP 模型在其他领域已经取得了成功应用，本文以 GP 模型进行径流预测研究，同时采用传统 BP 神经网络模型和 SVM 模型进行参照对比分析，以期寻求更为精确的径流预测方法。

1　GP 基本原理及算法的实现

1.1　高斯过程基本原理[18]

假设 n 个观测数据的训练集为$D=\{(x_i,y_i|i=1,\cdots,n)\}$。其中，$x_i$为 D 中的第 i 个输出变量，y_i为 D 中的第 i 个目标输出，n 为训练集中的样本个数。GP 模型是根据先验知识确定输入向量与目标输出之间的

收稿日期：2020-04-19
基金项目：国家重点研发计划项目（2018YFC1508200、2017YFC0404701）；国家自然科学基金资助项目（51669003）；广西研究生教育创新计划资助项目（YCBZ2018023）
作者简介：黄亚（1990—），男，四川内江人，工程师，博士，从事水文水资源研究.

关系f进行预测，即在给定输入向量时确定目标输出的条件分布。

假定f为一个高斯过程，即$f\sim GP(m,k)$，f是一个以m为均值函数，k为协方差函数的高斯过程。高斯过程是一个随机过程，其与高斯分布类似，高斯过程完全由其均值函数与协方差函数确定。

实际目标输出y往往会包含一些噪声，它与真实输出值$f(x)$相差ε。即

$$y=f(x)+\varepsilon \tag{1}$$

其中，ε为独立的随机变量，符合均值为0，方差为σ_n^2的高斯分布，即

$$\varepsilon\sim N(0,\sigma_n^2) \tag{2}$$

观测值y的先验分布为

$$y\sim N(0,\ \boldsymbol{K}+\sigma_n^2\boldsymbol{I}) \tag{3}$$

式中，$\boldsymbol{K}=\boldsymbol{K}(\boldsymbol{X},\ \boldsymbol{X})$为$n\times n$阶对称正定的协方差矩阵，矩阵中的元素$\boldsymbol{K}_{ij}$度量了$x_i$和$y_j$的相关程度。

此时，训练数据集的n个训练样本输出向量y和测试数据集的预测值f_*构成联合高斯先验分布

$$\begin{bmatrix} y \\ f_* \end{bmatrix}\sim N\left(\begin{bmatrix} \boldsymbol{K}+\sigma_n^2\boldsymbol{I} & \boldsymbol{K}_*^{\boldsymbol{T}} \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{bmatrix}\right) \tag{4}$$

式中，$\boldsymbol{K}_*=\begin{bmatrix} k(x_*,x_1) & k(x_*,x_2) & \cdots & k(x_*,x_n) \end{bmatrix}$；$\boldsymbol{K}_{**}=k(x_*,x_*)$。二者同样反映了其中元素之间的相关程度。

GP 模型可以根据实际预测数据规律选择不同的协方差函数，但都需要满足对任一点集都能够保证产生一个非负正定协方差矩阵[19]。本研究选择各项异性平方指数函数 SE 作为协方差核函数

$$k_{\mathrm{SE}}(\boldsymbol{x}_p,\ \boldsymbol{x}_q)=\sigma_f^2\exp\left(-\frac{1}{2\boldsymbol{l}^2}\left\|\boldsymbol{x}_p-\boldsymbol{x}_q\right\|^2\right) \tag{5}$$

式中，σ_f、$\boldsymbol{l}$为超参数，可以通过极大似然法获得。即，通过建立训练样本条件概率的对数似然函数对超参数求偏导，再采用共轭梯度优化方法搜索出超参数的最优解。对数自然函数为

$$L=\log p(\boldsymbol{y}\mid\boldsymbol{X})=-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K})^{-1}\boldsymbol{y}-\frac{1}{2}\log|\boldsymbol{K}|-\frac{n}{2}\log 2\pi \tag{6}$$

根据式（4）获取最优超参数后，可在训练集（$\boldsymbol{X}$，$\boldsymbol{y}$）的基础上，根据贝叶斯原理预测出与$\boldsymbol{X}_*$对应的最大概率的输出值，其中贝叶斯原理采用观测到真实数据，不断的更新概率预测分布，最后在已有训练样本（$\boldsymbol{X}$，$\boldsymbol{y}$）和预测样本$\boldsymbol{X}_*$条件下推断出$\boldsymbol{y}_*$的最大概率的预测后验分布

$$p(\boldsymbol{y}_*\mid\boldsymbol{y},\ \boldsymbol{X},\ \boldsymbol{x}_*)=N(\boldsymbol{y}_*;\ \mu_*,\ \sigma_*^{\ 2}) \tag{7}$$

其预测值y_*的均值和方差为

$$\mu_*=\boldsymbol{K}^{\mathrm{T}}(\boldsymbol{X}^*)\text{-}\boldsymbol{K}^{-1}\boldsymbol{y} \tag{8}$$

$$\sigma_*^{\ 2}=\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)-\boldsymbol{K}^{\mathrm{T}}(\boldsymbol{X}_*)\,(\boldsymbol{K})^{-1}\boldsymbol{K}(\boldsymbol{X}_*) \tag{9}$$

1.2 算法过程的实现

（1）确定训练样本的输入向量$\boldsymbol{X}_i$中历史点个数p，建立天峨水文站日径流量预测模型的训练样本（X，y）。

（2）数据标准化处理。当天峨水文站径流量实测数据数量级相差较大时或离散性较大时，需对输入、输出数据进行标准化处理。常用的标准化方法是：各训练样本的输入值除以所有样本在各维度上的标准差；对输出进行标准化处理，推荐的标准化方法是对各样本输出减去所有样本输出的平均值。

（3）选择适合的协方差函数，并依据协方差函数对超参数的要求给出初始超参数。

（4）对已建立的天峨水文站径流量训练样本进行训练，并通过对数自然函数式（6）获取最优超参数。

（5）基于在训练过程中获取的最优超参数，通过式（8）、式（9）预测出待预测样本的均值μ_*和方差$\sigma_*^{\ 2}$。

（6）根据上述步骤编制 Matlab 语言，实现基于 GP 模型的径流量数据有效预测。

2 工程实例应用

本文在利用 GP 模型进行径流预测的研究中，截取了广西天峨水文站某一年日平均径流量中的 4 月份前 28 d 数据构成训练样本，预测未来 5 d 该库区径流状况。在建立预测径流量的 GP 模型时，以连续p+1 日径流量为一个训练样本，其中前p日径流量构成训练样本的输入向量，第p+1 日径流量为输出值；p为历史点个数，对结果有较大影响。例如：给定数据x_1，x_2，…，x_6，…，建立 GP 模型过程中以$\boldsymbol{X}_1=[x_1,x_2,\ldots,x_p]$，$y_1=x_{p+1}$作为训练样本（$\boldsymbol{X}$，$\boldsymbol{y}$）的第 1 个输入向量和输出值，其余训练样本输入向量、输出值依次类推。

经比较历史点数和协方差函数对回归、预测数据精度的影响，选择p=4 个历史点数作为输入向量，SE 作为协方差核函数；建立径流的 GP 预测模型，超参数的初始对数设置为：$\log\boldsymbol{l}=[\log(0.1),\log(0.1),$

$\log(0.1),\log(0.01)]$；$\log(sqrt(\sigma_f))=-1$，以日径流量训练样本的极大似然为目标，经计算得到协方差函数最优超参数值为：$\log l=[5.358,-0.8443,0.9470,-0.7709,6.0832]$；$\log(sqrt(\sigma_f))=-6.4190$。

利用上述训练好的 GP 模型对径流量样本进行训练和预测，并以平均相对误差绝对值 *MPE*、均方误差 *MSE* 以及最大相对误差 *MRE* 作为衡量模型精度的评价指标，其值越小，预测模型描述样本数据则越精确。即

$$MPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|\text{实测值}(i)-\text{预测值}(i)|}{\text{实测值}(i)}\times 100\% \quad (10)$$

$$MSE=\frac{1}{N}\sqrt{\sum_{i=1}^{N}(\text{实测值}(i)-\text{预测值}(i))^2} \quad (11)$$

$$MRE=\text{Max}\frac{|\text{实测值}(i)-\text{预测值}(i)|}{\text{实测值}(i)}\times 100\% \quad (12)$$

为了分析 GP 模型在径流量预测精度上较传统机器学习方法的优越性，本文同时利用传统 BP 模型和 SVM 模型对相同的训练样本以及预测样本进行预测，并与 GP 模型的预测结果相比较，结果见表 1、2。从回归角度来说，基于传统 BP、SVM 模型的回归值都较 GP 模型相对误差大很多，反映了 GP 模型在数据回归上的相对优越性；从预测值相对误差来说，BP 模型预测相对误差最小值和最大值分别为 1.41%和 6.18%，SVM 模型分别为 0.01%和 5.24%，而 GP 模型仅为 0.13%和 2.71%，可以看出 GP 模型的预测精度明显优于传统 BP 模型和 SVM 模型；同时，从表 2 的预测结果 *MPE*、*MSE* 以及 *MRE* 来看，GP 模型预测精度均优于传统 BP 模型以及 SVM 模型，反映了传统 BP 模型和 SVM 模型的预测误差偏高,不完全适用于此类小样本问题。

表 1　基于 BP、SVM、GP 模型的训练值及训练误差

序号	实测值/$m^3\ s^{-1}$	BP		SVM		GP	
		拟合值/$m^3\ s^{-1}$	相对误差/%	拟合值/$m^3\ s^{-1}$	相对误差/%	拟合值/$m^3\ s^{-1}$	相对误差/%
1	408	435.2	6.67	435.76	6.80	408	0.17×10^{-4}
2	426	419.55	1.51	417.18	2.07	426	0.00×10^{-4}
3	437	438.81	0.41	436.89	0.02	437	0.24×10^{-4}
4	440	452.28	2.79	447.96	1.81	440	0.30×10^{-4}
5	446	437.28	1.95	436.22	2.19	446	0.11×10^{-4}
6	471	448.28	4.82	449.30	4.61	471	0.35×10^{-4}
7	478	482.58	0.96	481.65	0.76	478	0.30×10^{-4}
8	476	482.56	1.38	483.18	1.51	476	0.06×10^{-4}
9	476	470.46	1.16	476.05	0.01	476	0.03×10^{-4}
10	484	482.76	0.26	488.37	0.90	484	0.23×10^{-4}
11	485	498.61	2.81	501.85	3.47	485	0.06×10^{-4}
12	474	495.94	4.63	499.67	5.41	474	0.12×10^{-4}
13	486	477.49	1.75	484.76	0.25	486	0.19×10^{-4}
14	519	501.71	3.33	505.72	2.56	519	0.05×10^{-4}
15	561	548.89	2.16	543.49	3.12	561	0.27×10^{-4}
16	608	582.67	4.17	572.65	5.81	608	0.05×10^{-4}
17	624	622.54	0.23	609.76	2.28	624	0.32×10^{-4}
18	595	628.6	5.65	616.85	3.67	595	0.09×10^{-4}
19	587	590.95	0.67	587.20	0.03	587	0.06×10^{-4}
20	609	610.49	0.24	609.08	0.01	609	0.12×10^{-4}
21	661	656.37	0.70	653.93	1.07	661	0.10×10^{-4}
22	689	689.82	0.12	689.20	0.03	689	0.10×10^{-4}
23	709	694.26	2.08	687.96	2.97	709	0.15×10^{-4}
24	690	697.48	1.08	690.15	0.02	690	0.13×10^{-4}
25	714	688.79	3.53	679.61	4.82	714	0.30×10^{-4}
26	689	710.42	3.11	717.52	4.14	689	0.04×10^{-4}
27	701	702	0.14	701.20	0.03	701	0.10×10^{-4}
28	705	701.5	0.50	703.64	0.19	705	0.11×10^{-4}
	MPE	2.1		2.16		0.17×10^{-4}	
	MSE	2.77		3.07		0.00	
	MRE	6.67		6.80		0.24×10^{-4}	

表 2　基于 BP、SVM、GP 模型的预测值及预测误差

序列	实测值/$m^3\ s^{-1}$	BP		SVM		GP	
		预测值/$m^3\ s^{-1}$	相对误差/%	预测值/$m^3\ s^{-1}$	相对误差/%	预测值/$m^3\ s^{-1}$	相对误差/%
1	725	714.81	1.41	725.1	0.01	721.8	0.44
2	730	717.17	1.76	725.5	0.62	722.95	0.97
3	767	719.95	6.13	726.8	5.24	783.82	2.19
4	782	733.66	6.18	753.6	3.63	760.84	2.71
5	789	740.35	6.16	759.6	3.73	788.01	0.13
MPE		4.33		2.65		1.29	
MSE		16.95		11.50		5.63	
MRE		6.18		5.24		2.71	

3　结　语

（1）本文以天峨站多年日径流量为基本资料，经 GP 模型、传统 BP 模型以及 SVM 模型径流预测结果分析比较，可以看出：高斯过程回归模型模拟和预测能力要优于传统 BP 模型和 SVM 模型，模拟及预测效果平均相对误差小于 5%。实例应用表明，GP 模型应用于径流预测是可行的，是提高预测精度的有效方法。

（2）三种模型的预测结果都能满足工程实际需要，但相对来说GP模型预测精度较前两者高，它成功克服了BP神经网络小样本推广能力欠缺问题，解决了SVM超参数难以确定的难题，体现出其独特的超参数自适应获取、输出具有概率意义等一系列的优越性以及良好的应用潜力，为更加精确的径流预测提供了一种新的思路，对于地区径流预测以及区域用水具有较大的指导意义。

参考文献:

[1] 张文鸽，黄强，佟春生. 径流混沌时间序列局域多步预测模型及其在黄河上游的应用[J]. 水力发电学报，2007，26(4): 11-15.
[2] 王琪，张亭亭，游海林，等. 基于多元回归分析的大伙房水库径流中长期预报[J]. 水力发电, 2014, 40(5): 17-20.
[3] 彭勇，王国利. 小波、模糊与统计相关结合的径流预测方法研究[J]. 南水北调与水利科技, 2011, 9(4): 47-50.
[4] 李琳琳，岳春芳，张胜江. 基于小波方差分析的BP神经网络年径流预测[J]. 节水灌溉, 2014(6): 44-46.
[5] 汪哲荪，袁潇晨，金菊良，等. 基于集对分析的年径流自组织预测模型[J]. 水利水运工程学报, 2010(4): 33-37.
[6] 叶守泽，夏军. 灰色系统方法在洪水径流预测中的应用研究与展望[J]. 水电能源科学, 1995, 13(3): 197-205.
[7] 曹辉，黄强，白涛，等. 径流预测方法对比分析[J]. 人民黄河, 2009, 31(9): 36-37.
[8] 王鑫，徐淑琴，李洪涛. 河川径流预测方法比较研究[J]. 中国农村水利水电, 2014(4): 98-100.
[9] 陈守煜，王大刚. 基于遗传算法的模糊优选BP网络模型及其应用[J]. 水利学报, 2003(5): 116-121.
[10] 于国荣，夏自强. 混沌时间序列支持向量机模型及其在径流预测中应用[J]. 水科学进展, 2008(1): 116-122.
[11] 刘国东，丁晶. BP网络用于水文预测的几个问题探讨[J]. 水利学报, 1999(1): 66-71.
[12] 汪海燕，黎建辉，杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.
[13] 苏国韶，赵伟，彭立锋，等. 边坡失效概率估计的高斯过程动态响应面法[J]. 岩土力学, 2014, 35(12): 3592-3601.
[14] 孙斌，姚海涛，刘婷. 基于高斯过程回归的短期风速预测[J]. 中国电机工程学报, 2012, 32(29): 104-109.
[15] 苑进，胡敏，Wang Kesheng，等. 基于高斯过程建模的物联网数据不确定性度量与预测[J]. 农业机械学报，2015，46(5): 265-272.
[16] 雷雨，赵丹宁，高玉平，等. 基于高斯过程的日长变化预报[J]. 天文学报, 2015, 56(1): 53-62.
[17] 何志昆，刘光斌，赵曦晶，等. 高斯过程回归方法综述[J]. 控制与决策, 2013, 28(8): 1121-1129.
[18] SEEGER M. Gaussian processes for machine learning[J]. International Journal of Neural Systems, 2004, 14(2): 69-106.
[19] MACKAY D J C. Comparison of approximate methods for handling hyperparameters[J]. Neural Computation, 1999, 11(5): 1035-1068.