基于机器学习的水位流量关系模型参数估计

江 竹，宋文武

（西华大学能源与环境学院，四川 成都 610039）

摘 要：为了克服经典水位流量关系模型在刻画河流动态变化特性时所存在的局限性，提出采用局部加权回归算法估计模型参数；为了提高参数估计精度以及流量的计算效率，提出一种聚类树加权回归方法。首先对训练样本进行聚类，然后使用k-最近邻方法将新的水位样本划分进最恰当的聚类中，最后估计河流日流量。该方法在估计过程中，避免了不相关信息的干扰，从而提高了日流量数据估计的效率和精度。利用某水文站的实测数据对方法进行测试，仿真结果表明，方法估计精度高，为水位流量关系模型参数估计提供了新的有效方法。

关键词：水位流量关系；参数估计；局部加权回归；聚类树；k-最近邻

中图分类号：P641.8 **文献标识码**：A **文章编号**：1000-0852(2013)01-0074-05

水位流量关系曲线是描述测站处基本断面的水位与通过该断面的流量两者之间的关系曲线。在水利水电工程规划、设计与施工过程中，水位和流量的估计与预报一直是一个重要课题[1-3]。当前的河流流量测验技术复杂，耗资昂贵且难以连续进行，而连续观读平均水位则容易办到，因此流量资料通常是根据水位资料，由水位流量关系推算得到。

流量数据的质量直接影响到水文计算和水文预报分析的精度，因此水位流量之间的关系模型一直是研究者们关注的焦点。对于给定的水位和流量数据，由于受变动回水与洪水涨落等因素的影响，因此数据本身并不一定可靠，个别数据的误差可能很大，从一堆看上去杂乱无章的数据中找出一定的规律，能够比较真实地反映天然河道的水位流量关系，常用的方法是构造两者之间的关系式[4]。出于研究的需要，研究人员针对具体的水流条件提出了许多经典的水位流量关系模型，如通过曼宁公式计算的幂指数型、多项式型等[5]。

水流的变化是一个包含大量随机因素的复杂过程，虽然在某一具体条件下可用确定性方程描述，但是受降水天气系统以及人类活动综合作用的影响，河流成为非线性、强相关、高度复杂的动力系统[6]。因此，经典的水位流量关系模型已经不能很好地适用于当前水流特性的研究。近年来，利用新的科学理论探索新的建模方法是人们进行估计与预测课题研究的一个重要方面。机器学习方法是一种非参数学习方法，它不必形成一个明确的假设来定义整个样本空间上的完整目标函数，它可以对每个查询样本形成一个不同的目标函数的局部逼近[7]。因此，本文提出采用一种非参数回归的机器学习方法——局部加权回归建立河道日流量与日平均水位之间的关系，进而进行流量参数的估计；为了进一步提高参数估计的精度，本文提出了一种聚类树加权回归方法。该方法在对训练样本进行聚类后使用k-最近邻方法将新的（用于测试的）水位样本划分进最恰当的聚类中，然后再对日流量进行估计。最后利用实测数据对局部加权回归与本文提出的方法进行了对比测试。

1 水位流量关系模型

水位流量关系是指某基本断面的流量 Q 和所对应的水位 h 之间的经验关系。不同学者对水位流量关系的表达不尽相同，一般的，这种关系有两种形式。其中一种形式为幂律形式的关系：

$$Q=ah^b \tag{1}$$

收稿日期：2011-12-05
基金项目：四川省流体机械重点实验室资助(SBZDPY-11-5)；西华大学高校重点项目(Z1120413)
作者简介：江竹(1979-)，女，四川成都人，讲师，博士，主要从事水文资源教学与研究工作。E-mail：HILL5525@163.com

式中:Q 为通过断面的流量,m^3/s;h 为平均水位,m;a 为系数;b 为指数。

在式(1)的基础上进行自然对数变换可得到流量和平均水位的线性表达式:

$$\ln Q=\ln a+b\ln h \tag{2}$$

另一种水位流量关系式是目前最为常用的多项式型:

$$Q=c_0+c_1h_1+c_2h_2+\cdots+c_mh_m \tag{3}$$

式中:$c_0,c_1,\cdots,c_m$ 为待定系数,可由回归方法计算得到。

多项式型因其图形与大部分观测站的水文特性相符而被得到广泛应用[4]。

上述表达式中的参数 a、b、c_i 一般都是通过最小二乘等回归方法计算得到,它的值受河流流域面积大小、变动回水以及洪水涨落等诸多复杂因素较大的影响,因此采用传统的数学方法所建立的水位流量关系模型并不能较为准确的反应河流中真实的水文情况。

基于此,本文提出采用机器学习方法建立起流量与水位之间的非参数形式的关系模型,然后再进行模型参数的估计,力求为水文计算和水文预报分析提供真实可靠的数据。

2　局部加权回归

局部加权回归(LWR)是一种非参数回归方法,最早由 Cleveland[8]提出,由 Cleveland 和 Develin[9]将其推广到多个自变量的情形。该方法的基本思想是利用加权最小二乘方法在自变量空间的每一点处局部拟合一个多项式函数作为回归函数在该点的估计。局部加权回归在计算变量的关系时采用开放式方法,不套用现成的函数公式,所拟合的曲线可以很好的描述变量之间关系的细微变化[8]。本文在分析流量和水位的变量关系时,实现局部加权回归的具体步骤如下:

(1)首先以由日平均水位构成的变量为中心确定一个区间。区间的宽度取决于:

$$q=fn \tag{4}$$

式中:q 为参加局部回归的观察值的个数;f 为参加局部回归的观察值的个数占观察值个数的比例;n 为观察样本的个数。

本文先选定 f 值,再根据 f 和 n 确定 q 的取值。

(2)定义区间内所有点的权数。权数由权数函数决定,任一点的权数是权数函数曲线的高度。局部加权回归的权数函数形式有很多种类型,大量非参数统计学研究表明,权数函数的选择对加权回归的计算精度并不会带来实质性的影响。本文选用立方权数函数:

$$W(u)=\begin{cases}(1-u^3)^3 & 0\le u\le 1\\ 0 & \text{其他}\end{cases} \tag{5}$$

则观测点 (y_i,x_i) 的权数为:

$$w_i=W(p(x,x_i)/d(x)) \tag{6}$$

式中:x 为日平均水位;y 为流量参数;$p(x,x_i)$ 是 x 与 x_i 之间的距离;$d(x)$ 是第 q 个与 x 最靠近的点 x_i 之间的距离。

(3)利用加权最小二乘方法在自变量空间的每一点处局部拟合一个多项式函数。

(4)拟合流量参数的值。经过以上 4 个步骤便完成局部加权回归过程,最终求得流量的估计值。

3　聚类树加权回归

聚类是将样本集合划分为若干类,使类内的样本之间有较高的相似度,而类间的样本差别尽可能大[10]。聚类过程的本质就是一种最优化过程,即通过一种快速运算使系统的目标函数达到一个极小值。在进行聚类的过程中,主要着眼于将大批样本分割成若干具有某种共性的子集,并不牵涉对具体样本未知域的预测推断,因此不受人的先验知识的约束,可以获得数据集合的最原始信息。

聚类主要包括基于划分的方法、基于层次的方法、基于密度的方法和基于网格的方法[11]。系统聚类法也即层次聚类法[12],它的基本思想是通过建立并逐步更新距离系数矩阵(或相似系数矩阵),找出合并最接近的 2 类,直至全部分析对象合并为 1 类。由于整个过程可表示成一个二叉层次树,故层次聚类又称为聚类树方法,这种方法优点是聚合过程脉络清晰、可视化程度高,且聚类结果不依赖样本的初始排列,聚类结果比较稳定,不易导致类的重构。

为了提高河流流量估计的精度和计算效率,本文提出一种新的学习方法:聚类树加权回归。该方法对水位流量关系进行估计的具体过程为:

首先,对用于训练的水文样本资料进行聚类。

实测的水文数据被作为包含群聚的资料,其中第 k 年的群聚用集合 G_k 来表示,假设 G_k 包含 n 笔资料 $\{x_1,x_2,\cdots,x_n\}$(本例中,x_i 是第 k 群的河流水位和流量数据),方法试图找到一组 m 个代表点 $Y=\{y_1,\cdots,y_m\}$ 使目标函数:

$$J(X;Y,U)=\sum_{i=1}^{n}(|x_i-y_k|^2) \tag{7}$$

最小，其中 y_k 是 G_k 的聚类中心。

在确立了目标函数后，本文进行如下工作：

(1)从 n 个训练对象中任意选择 k 个聚类中心；

(2)对每一个河流水位和流量数据 x_i 寻找与之最接近的聚类中心，并将 x_i 加入该类；

(3)计算目标函数 $J(X;Y,U)$，如果保持不变，代表分群结果已经稳定，跳转至第 5 个步骤，否则进行第 4 步。

(4)由固定的 U 产生最佳的 Y，跳回第 2 个步骤。

(5)建立并逐步更新距离系数矩阵，找出合并有最大相似度的节点，合并这两个节点为一个新的节点，直至只剩一个节点。

接着，本文采用 k-最近邻方法将新的(用于测试的)水位样本划分进最恰当的聚类中。k-最近邻方法的基本思想是在多维空间中找到与未知样本最邻近的 k 个点，并根据这 k 个点的类别来判断未知样本的类[6]。当给定一个新的水位样本时，k-最近邻分类器搜索出 k 个与新样本最为接近的已完成聚类的水位样本。其中，样本 x_i 和 x_j 之间的最近邻是根据标准欧氏距离

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2} \tag{8}$$

进行定义。

其中$<a_1(x),a_2(x),\cdots,a_n(x)>$表示新样本 x 的特征向量。

最后，当每个新的水位样本被分到已完成聚类的群中后，使用上一小节中的局部加权回归对河流流量进行估计。

4 方法测试

4.1 实验数据

本文采用岷江某流域 2007~2010 年 4a 的水文数据对方法进行测试。其中前 3 年的日平均流量和平均水位作为用于训练的历史资料数据，2010 年的数据则作为测试数据用于验证方法的估计精度。其中部分实验数据见表 1。

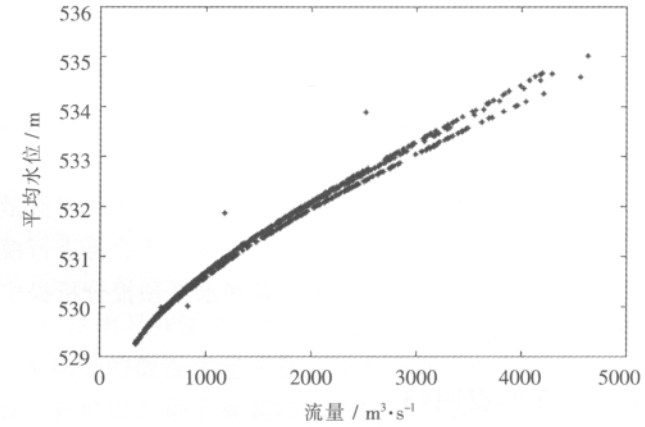
4.2 水位-流量关系模型参数估计

本文在采用聚类树加权回归方法进行实验时，方法的聚类中心从河流资料数据中随机选取。聚类结果如图 1 所示。

表 1 部分实验数据

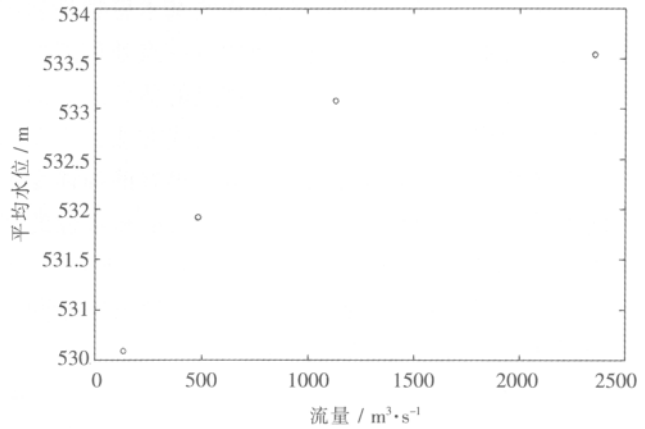
Table 1 Some experimental data

序号	平均水位 / m	平均流量 / $m^3 \cdot s^{-1}$	平均含沙量 / $g \cdot m^{-3}$
1	529.76	534	119
2	529.74	524	106
3	529.73	519	94.8
4	529.73	519	82.9
5	529.74	524	74.4
6	529.74	526	76.4
7	529.73	523	79.2
8	529.74	524	82.1
9	529.71	513	85
10	529.72	514	87.4



(a)观测数据

(a) Observed data



(b) 4 个聚类中心

(b) Four cluster centers

图 1 聚类结果

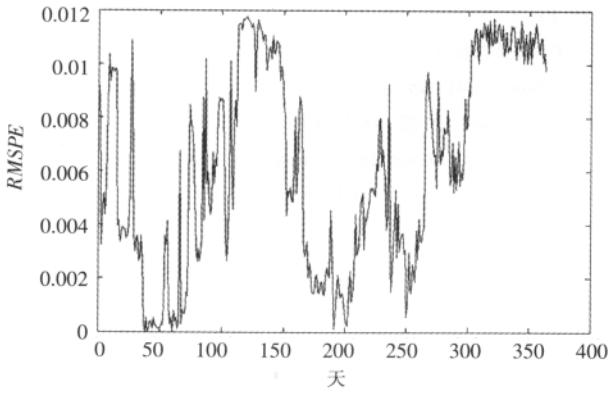
Fig.1 Demonstration of clustering

其中，图 1(a)为观测数据的分布状况，图 1(b)为随机选取的 4 个聚类中心。

本文采用均方根误差:

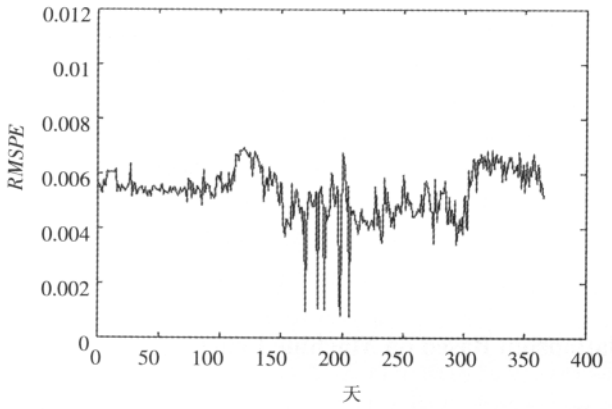
$$RMSPE=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{u_n^e-u_n^o}{u_n^o}\right)^2} \tag{9}$$

作为检验每种方法性能的标准,其中 N 为观测数据的量,角标 e 和 o 分别代表估计值和观测值。部分均方根误差结果如图 2 所示。



(a)局部加权回归的均方根误差

(a) RMSPE of LWR



(b)聚类树加权回归的局方根误差

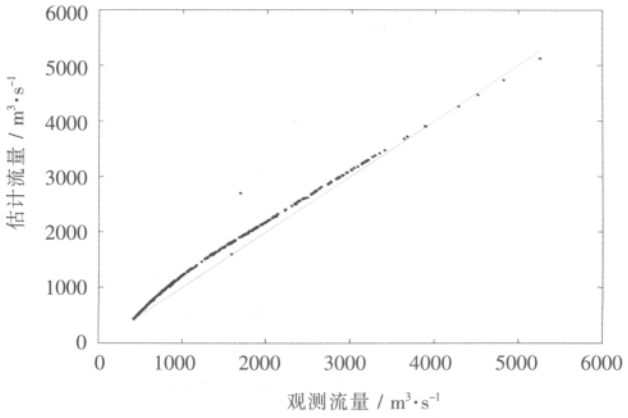
(b) RMSPE of clustering-tree weighted regression

图 2　两种方法的均方根误差

Fig.2 RMSPE of two methods

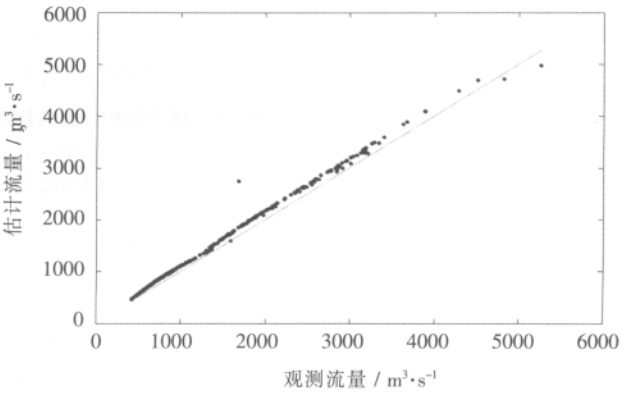
其中图 2(a)为采用局部加权回归方法对日流量估计的均方根误差,图 2(b)为采用聚类树加权回归方法的均方根误差结果。比较二者可知,本文提出方法的估计精度明显优于局部加权回归方法。

图 3 给出了两种方法对河流日流量的估计结果。其中,图 3(a)为采用局部加权回归方法对河流流量的估计结果。由图中的估计结果并结合图 2(a)所示方法的均方根误差值可知,当河流流量较小时,估计的日流量与实际观测值之间存在较大的偏差;当进入大流量区域时,估计结果有所提高。图 3(b)为采用聚类树加权回归方法对河流日流量进行估计的结果。通过分析图 3(b)和图 2(b)可知,本文提出的方法有效改善了正常气候条件下河流日流量的估计值与实测值之间存在较大偏差的现象,这是由于先将河流水文数据作为训练样本进行聚类,再使用 k-最近邻方法将与流量相关的新的河流数据样本划进适当的类中,可以避免其它不相关信息的干扰,从而提高了河流日流量数据估计的效率和精度。在不出现极端气候的情况下,本文提出的方法对更加精确地捕捉河流水文情况变化具有很强的现实意义。



(a) 局部加权回归估计结果

(a) Estimated results by LWR



(b) 聚类树加权回归估计结果

(b)Estimated results by clustering tree weighted regression

图 3　流量估计结果

Fig. 3 Estimated results of discharge

5　结论

水流的运动过程是一个包含大量随机因素的复杂过程,经典水位-流量之间的关系模型建立于经验回

归的基础之上，它们已经不能很好地适用于复杂河流水流特性的研究。

本文提出采用机器学习方法中的局部加权回归以及一种聚类树加权回归方法对河流水位与流量关系模型参数进行估计。利用某水文站的实测数据对方法进行验证，仿真结果表明：本文提出的聚类树加权回归方法在不出现极端气候条件的情况下可以获得比局部加权回归更高的参数估计精度，能更好地捕捉动态水流的变化特性。

参考文献：

[1] 卢敏.径流预报的 SVM 应用研究[J].中国农村水利水电,2006,(2):47–49. (LU Min. Research on the SVM application of runoff forecast [J]. China Rural Water and Hydropower. 2006,(2): 47–49. (in Chinese))

[2] 冯汉中,陈永义.处理非线性分类和回归问题的一种新方法(II)——SVM 方法在天气预报中的应用[J].应用气象学报,2004,(3):355–365. (FENG Hanzhong, CHEN Yongyi. A new method in dealing with nonlinear classification and regression problems (II) – application of SVM method in weather forecast [J]. Journal of Applied Meteorological Science, 2004,(3), 355–365. (in Chinese))

[3] 冯国章,王双银,韦华艳.多元自回归模型在枯水径流预报中的应用[J]. 自然资源学报,1996,(2):84–186. (FENG Guozhang, WANG Shuangyin, WEI Huayan. Application of the multivariate autoregressive model to low flow forecast [J]. Journal of Natural Resources, 1996,(2), 84–186. (in Chinese))

[4] 戴凌全,戴会超. 基于最小二乘法的河流水位流量关系曲线推算[J].人民黄河.2010,(32): 37–39. (DAI Lingquan, DAI Huichao. Calculation of stage–discharge relationship curve based on least square method [J]. Yellow River, 2010,(32), 37–39. (in Chinese))

[5] French M N. Rainfall forecasting in space and time using a neural networks [J]. Journal of Hydrology, 1992,(7):1–13.

[6] Mohsen Behzad, Keyvan Asghari. Generalization performance of support vector machines and neural networks in runoff modeling [J]. Expert Systems with Applications. 2009,(36): 7624–7629.

[7] 朱明. 数据挖掘[M].安徽：中国科技大学出版社.2002.(ZHU Ming. Data Mining [M]. Anhui: Press of University of Science and Technology of China. 2002. (in Chinese))

[8] Cleveland W S. Robust locally weighted regression: an approach to regression analysis by local fitting[J]. Journal of the American Statistical Association,1988.

[9] Cleveland W S, Devlin S J. Locally weighted regression: an approach to regression analysis by local fitting [J]. Journal of the American Statistical Association,1978.

[10] Castro R, Coates M, Nowak R. Likelihood based hierarchical clustering [J]. IEEE Trans Signal Process. 2004,52:2308.

[11] 史坤鹏. 基于经验模式分解的聚类树方法及其在同调机组分群中的应用[J]. 电网技术, 2007, 31(22), 21–25. (SHI Kunpeng. Mode decomposition based clustering–tree method and its application in coherency identification of generating sets[J]. Power System Technology, 2007, 31(22), 21–25. (in Chinese))

[12] Tom Mitchell. 机器学习[M].北京：机械工业出版社,2003,176–187. (Tom Mitchell. Machine Learning [M]. Beijing: China Machine Press, 2003. (in Chinese))

Parameter Estimation of Stage–discharge Relationship Based on Machine Learning

JIANG Zhu, SONG Wenwu

(*School of Energy and Environment, Xihua University, Chengdu* 610039, *China*)

Abstract: To overcome the limitation of the classical stage–discharge relationship model in describing the dynamic characteristics of a river, the locally weighted regression method was used to estimate the model parameters. In order to improve the estimation precision and the calculation efficiency of river discharge, a novel method called clustering–tree weighted regression was proposed. Firstly, the trained samples were clustered in this method. Secondly, k–nearest neighbors method was used to cluster new stage samples into the best fit clustering. Finally, the daily discharge of the river was estimated. During the estimation process, the interference of irrelevant information was avoided, so the estimation precision and efficiency of daily discharge were improved. The data observed at some hydrological stations were used for the test. The simulation results show that the estimation precision of this method is high. This provides a new effective method for the estimation of parameters of stage–discharge relationship model.

Key words: stage–discharge relationship; parameter estimation; locally weighted regression; clustering tree; k–nearest neighbors