文章编号: 1004- 4574(2009) 05- 0174- 04

冰情预报的投影寻踪回归模型

王志兴[1,3], 李成振[2], 陈 刚[1]

(1.西安理工大学 水利水电学院, 陕西 西安 710048; 2.中水东北勘测设计研究有限责任公司, 吉林 长春 130061;
3.黑龙江省水利水电勘测设计研究院, 黑龙江 哈尔滨 150080)

摘 要: 将投影寻踪回归模型应用于黑龙江上游江段开河日期的预报, 并与 GA- BP模型预报的结果进行了对比分析。结果表明, 投影寻踪回归模型预报的精度及稳定性较高, 其性能优于常用的 GA- BP模型。预报采用理论分析与多元逐步回归分析相结合的方法筛选预报因子, 既可确保不遗漏基本影响因子, 又能剔除对目标值影响不显著的因子, 用于确定冰情预报的预报因子较为适宜。

关键词: 防洪工程; 冰情预报; 投影寻踪; 预报因子
中图分类号: P456 文献标识码: A

Projection pursuit regression model for ice situation forecast

WANG Zhi-xing[1,3], LI Cheng-zhen[2], CHEN Gang[1]

(1. The Institute of Water Conservancy and Hydroelectricity, Xi' an University of Technology, Xi' an 710048, China;
2. China Water Northeastern Investigation, Design And Research Co., Ltd, Changchun 130061, China; 3. Heilongjiang Provincial Water Conservancy and Hydroelectric Power Investigation, Design and Research Institute, Harbin 150080, China)

Abstract: This paper applies projection pursuit regression(PPR) model to break-up date of the upstream of Heilong River and gives a comparative analysis with GA-BP model. The result indicates that the PPR model's forecast accuracy and stability are comparatively high and whose working performance is superior to the common GA-BP model. Combining theoretical analysis with multivariate stepwise regression analysis method to select forecasting factors in the forecasting can not only ensure the basic factors completely but also cut out the unessential factors to target value, which is comparatively appropriate to determination of the forecast factors in ice situation forecast.

Key words: flood control works; ice situation forecast; projection pursuit; forecast factor

高寒地区的河流极易发生凌汛灾害, 及时准确预报河段的冰情, 是进行防凌指挥、调度, 采取必要安全措施的重要科学依据。冰情受水力、热力及河势等因素制约, 给传统数学模型的应用造成极大困难。近年来, 人工智能技术的研究与发展为冰情预报开辟了一条崭新的道路, 如 2004年, 陈守煜[1]、冀鸿兰利用模糊优选神经网络 BP模型, 对黄河内蒙古河段封河、开河日期进行预报; 2005年, 王涛[2]、杨开林等应用 Levenberg-Marquart算法改进传统 BP神经网络理论进行黄河宁蒙段冰情预报。

投影寻踪是用来分析和处理高维观测数据, 尤其是非线性、非正态高维数据的一种新兴人工智能技术, 它通过把高维数据投影到低维子空间, 寻找能反映原高维数据结构或特征的投影, 达到研究分析高维数据的目的。一些研究结果表明[3-4], 投影寻踪模型在多个方面要优于 BP网络。与非参数投影寻踪回归模型相比, 基于 Hermite多项式的参数投影寻踪回归模型的优越性已被大量的实践所证实[5], 因此本文尝试将基于 Hermite多项式的参数投影寻踪回归模型引入到冰情预报中来。

收稿日期: 2008- 11- 16; 修订日期: 2009- 07- 16
基金项目: 国家公益性行业专项经费资助项目 (200701006)
作者简介: 王志兴 (1964-), 男, 教授级高级工程师, 博士研究生, 主要从事自然灾害方面研究. E-mail: GHCWZX@163.com

1 基于 Hermite多项式的参数投影寻踪回归模型

1.1 建模过程

基于 Hermite多项式的参数投影寻踪模型 (projection pursuit regression model, PPR)采用可变阶的正交 Hermite多项式拟合其中的一维岭函数, 其数学表达试为

$$h_r(z) = (r!)^{\frac{1}{2}} \pi^{\frac{1}{4}} 2^{\frac{r-1}{2}} H_r(z) \varphi(z), \qquad -\infty < z < +\infty. \tag{1}$$

上式中, $r!$ 代表 r的阶乘; $z = a^{\mathrm{T}} X$ (a为投影方向, $\| a \| = 1$); φ为标准高斯方程, $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$; $H_r(z)$为 Hermite多项式, 采用递推的形式给出, 如 $H_0(z) = 1$, $H_1(z) = 2z$, $H_r(z) = 2(zH_{r-1}(z) - (r-1)H_{r-2}(z))$。

此时参数投影寻踪回归模型的表达式为

$$f(z) = \sum_{i=1}^{m} \sum_{j=1}^{R} c_{ij} h_{ij}(z), \tag{2}$$

式中, m 为岭函数个数, R 为 Hermite多项式的阶数, c是多项式系数, h表示正交 Hermite多项式。用遗传算法优化最佳投影方向 a, 由最小二乘法获得系数 c, 便可确定回归函数 $f(z)$, 进行回归预测。设有因变量 y_i ($i = 1, 2, \cdots, n$)和 p个自变量 $\{x_1, x_2, \cdots, x_p\}$, 观测 n个样本点, 构成自变量与因变量的数据表 $X = [x_1, x_2, \cdots, x_p]_{n \times p}$和 $Y = [y]_{n \times 1}$。以 Hermite多项式为岭函数的参数投影回归模型的建模步骤如下:

步骤 1 随机产生 M 个初始投影方向, 对每个方向计算投影值

$$z_i = \sum_{j=1}^{P} a_j x_{ij} \qquad i = 1, 2, \cdots, n; \ j = 1, 2, \cdots, p, \tag{3}$$

上式 x_{ij}已经进行了归一化处理;

步骤 2 对散步点 (z, y), 按式 (2)计算, 多项式系数 c用最小二乘法获得;

步骤 3 优化投影指标函数。在优化投影方向 a时, 同时考虑多项式系数 c的优化问题, 通过求解投影指标函数最小化问题来估计最佳 a, c值, 即

$$\min Q(a, c) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)}, \tag{4}$$

$$s.t. \sum_{j=1}^{P} a^2(j) = 1, \tag{5}$$

这是一个以 a, c为优化变量的复杂非线性优化问题, 本文采用改进的自适应遗传算法 (AGA)[6]来解决其高维全局寻优问题;

步骤 4 计算第一次的拟合残差 $r_1 = y - \hat{y}$, 如果满足要求则输出模型参数, 否则用 r_1 代替 y, 回到步骤 1开始下一个岭函数的优化, 直到满足一定的要求, 停止增加岭函数个数, 输出最后结果。

1.2 模型控制指标

为提高计算效率及预测精度, 应根据需解决问题的性质为模型设置适当的控制指标, 如表 1所示。

表 1 模型控制指标

Table 1 Controlling indices of model

模型控制指标	说明
拟合误差 E_{PS}	当 $\min Q(a, c) < E_{PS}$时, 样本训练终止, 输出结果。E_{PS}取值不易太小, 否则会出现过度拟合, 导致预报精度下降。
进化指标 D_R	当第 $N+1$代的 $\min Q(a, c)$与第 N代的 $\min Q(a, c)$之差小于 D_R 时, 可认为继续进行遗传进化计算已不能提高训练精度, 结束遗传进化计算。
预测成功与否的判定	因采用归一化处理后的样本进行训练和预测, 故若预报结果在区间 [0, 1], 可以认为预报合理, 将预报结果还原, 否则认为预报失败, 回到步骤 1重新计算。

2 冰情预报

黑龙江上游地处高寒地区, 受地理位置、河流流向、河道特征和水文气象条件等综合因素影响, 黑龙江上游冰坝 (冰塞)出现极为频繁。冰坝多出现在额尔古纳河入汇处至呼玛河入汇处 500 km 的河道上, 大范围冰坝常延续

到结雅河入汇处下游的孙吴县沿江乡,河段长度约 1000 km。及时准确预报河段的冰情,具有十分重要的现实意义。本节将应用投影寻踪回归模型对黑龙江上游河段洛古河、呼玛、上马厂站的开河日期进行预测。

2 1 预报因子的确定

影响凌汛因素很多,可以概括为热力因素、动力因素及河势因素,其中起主要作用的是热力因素和动力因素。合理地选择预报因子是模型预报成败的关键。文献[1]利用相关系数法确定开河日期预报因子为累计正气温、流量、水位及封冻期最大冰厚;文献[2]通过对黄河历史冰情资料分析发现影响开河的因素不仅包括本站开河前期的气温、流量、水位和槽蓄水量,还要受到上游站气温、流量、水位等因素的影响。本文将根据理论分析并结合多元逐步回归分析法[7]筛选预报因子,力求预报因子中只含有对目标值影响显著的因子而不包含对目标值影响不显著的因子。

洛古河、上马厂为水文站,1987年建站,已收集到两站 1987– 2006年的水位、流量、气温及冰情等资料;呼玛站为水位站,收集到该站 1983– 2006年水位、气温及冰情资料。由开河机理可知,开河日期与流量、水位、气温及冰厚有关,这 4个因子直接选出 (呼玛站无流量资料除外),其它因子用逐步回归法筛选,得出各站的预报因子如表 2所示。

表 2 各站开河日期预报因子

Table 2 Forecast factors of break-up date for hydrological Stations

位置	洛古河	呼玛	上马厂
预报因子	预报发布日冰厚;预报发布前 10 d水位、流量均值及气温和;封冻期最大冰厚;最大冰厚发生日期;累积正气温	预报发布日冰厚;预报发布前 10 d水位均值及气温和;封冻期最大冰厚;最大冰厚发生日期累积正气温	预报发布日冰厚;预报发布前 10 d水位、流量均值及气温和;封冻期最大冰厚;最大冰厚发生日期;累积正气温

2 2 黑龙江上游河段开河日期预报

预留后 4a的样本作预测检验,其它样本用于建模。将归一化处理后的样本输入上述模型进行反复训练,训练过程中适当调节 Hermite多项式的阶数及拟合误差 E_{PS}的取值,在达到一定的预测精度的同时还要保障模型的稳健性。以洛古河站开河日期预报为例,根据表 2选定的预报因子,以开河历时 (开河日期距 6月 1日的天数)为目标值,利用 1987– 2002年的资料训练模型,用 2003– 2006年的资料作预测检验,经反复训练最终确定拟合误差 E_{PS}取 0 05,岭函数个数为 1,Hermite多项式的阶数取 7可取得最佳预测结果,结果如表 3所示。与此类似,可得其它两站的预报结果,如表 4、表 5所示。

表 3 洛古河站开河日期预报

Table 3 Forecast of break-up date for Luoguohe Station

年度	实测 (月 – 日)	预测 (月 – 日)	偏差 /d	合格与否
2003	04 – 23	04– 18	5	合格
2004	04 – 28	04– 30	2	合格
2005	05 – 02	05– 05	3	合格
2006	05 – 01	05– 04	3	合格

表 4 呼玛站开河日期预报

Table 4 Forecast of break-up date for Huma Station

年度	实测 (月 – 日)	预测 (月 – 日)	偏差 /d	合格与否
2003	04 – 28	04– 26	2	合格
2004	04 – 26	04– 27	1	合格
2005	05 – 05	05– 02	3	合格
2006	04 – 30	04– 26	4	合格

表 5 上马厂站开河日期预报

Table 5 Forecast of break- up date for Shangmachang Station

年度	实测(月-日)	预测(月-日)	偏差/d	合格与否
2003	04-18	04-20	2	合格
2004	04-17	04-16	1	合格
2005	04-22	04-19	3	合格
2006	05-01	05-03	2	合格

2.3 预报结果分析

黑龙江上游河段冰情预报属中长期预报, 根据《水文情报预报规范》的有关规定[8], 将预见期定为 15 d, 本文封河、开河日期预报合格率为 100%, 精度比较高, 属于甲等预报方案。呼玛站预报因子中缺少流量因子, 但水位与流量有一定的相关性, 故仍能得出较为满意的预测结果, 证明投影寻踪算法对信息含糊、不完整等复杂情况的处理有较强的适应性。

2.4 与 GA-BP算法的比较

从投影寻踪算法及 BP算法与遗传算法的结合上看: 投影寻踪算法利用遗传算法优化单位化投影方向 a, 种群中每个染色体的取值可确定在[-1, 1]区间, 而 GA-BP算法是用遗传算法优化 BP网络的权阈值, 其取值范围较大, 且不能准确确定, 因此前者受初始种群随机性的影响较小, 模型更为稳健, 从这个层面上讲, 投影寻踪算法较 BP算法更适宜与遗传算法相结合。

为进一步考查参数投影寻踪算法的性能, 针对黑龙江上游段开河日期预报, 将参数投影寻踪算法与 GA-BP算法的计算结果做一比较, 如表 6所示。

表 6 PPR 算法与 GA-BP算法的比较

Table 6 Comparison of PPR algorithm with GA-BP method

算法	预测成功率/%	计算用时/s	平均误差/d	预报合格率/%
GA-BP算法	70	13.392	3.76	100
PPR算法	90	12.573	2.58	100

从表 6可以看出, 在相同环境、相同规模下, 投影寻踪算法与 GA-BP算法相比, 两者计算用时相当, 投影寻踪算法的计算精度略高, 稳健性远大于 GA-BP算法。

3 结语

(1)将投影寻踪回归模型应用于冰情预报是切实可行的, 具有较高的精度和稳健性。

(2)采用理论分析与多元逐步回归分析相结合的方法筛选预报因子, 既可确保不遗漏基本影响因子, 又能剔除对目标值影响不显著的因子, 用于确定冰情预报的预报因子较为适宜。

(3)在实际预报中, 受历史资料不全或缺失等影响, 有些影响因子没法考虑或者用其它相关因子替代。实例表明这样处理仍能得出较为满意的结果, 证明投影寻踪回归模型对信息含糊、不完整等复杂情况的处理有较强的适应性。

参考文献:

[1] 陈守煜, 冀鸿兰. 冰凌预报模糊优选神经网络 BP方法[J]. 水利学报, 2004, (6): 114-118

[2] 王涛, 杨开林, 等. 神经网络理论在黄河宁蒙河段冰情预报中的应用[J]. 水利学报, 2005, 36(10): 1205-1208

[3] 杨永生, 何平. 投影寻踪回归与 BP神经网络方法在前汛期降水预测中的比较研究[J]. 气象与环境学报, 2008, (01).

[4] Hwang Teng-Neng Lay Shy-Rong Maechler M. Regression modeling in back-propagation and projection pursuit learing[J]. IEEE Trans Neural Networks, 1991: 342-353.

[5] 付强, 赵小勇. 投影寻踪模型原理及其应用[M]. 北京: 科学出版社, 2006.

[6] 贾嵘, 蔡振华, 罗兴锜. 改进自适应遗传算法及其在水电站最优报价中的应用[J]. 水力发电学报, 2007, 26(1): 11-15.

[7] 丁士晨. 多元回归分析方法及应用[M]. 长春: 吉林人民出版设, 1979.

[8] SL250-2000 水文情报预报规范[S].