doi:10.11988/ckyyb.20170046 2018,35(8):17-21

基于互信息与神经网络的天山西部山区融雪径流中长期水文预报

周育琳,穆振侠,彭 亮,高 瑞,尹梓渊,汤 瑞
(新疆农业大学 水利与土木工程学院,乌鲁木齐 830052)

摘 要:为提高天山西部山区融雪径流的预报精度,更好地指导所在区域的工农业生产发展,针对影响预报精度的关键问题(预报因子的选择),基于互信息法、相关系数法、主成分分析法对研究区的预报因子进行优选,采用RBF神经网络以及组合小波BP神经网络模型进行径流预报研究,并进行不同方案的比较。结果表明:①互信息法优选出的预报因子作为模型输入可以提高预报精度;②采用不同优选预报因子作为RBF神经网络以及组合小波BP神经网络模型的输入变量,结果表明RBF神经网络模型的预测精度要好于组合小波BP神经网络模型;③以相对误差作为评价模型精确度的标准,预测效果最好的是基于互信息方法挑选出的预报因子作为RBF神经模型输入数据的模型预测结果。

关键词:水文中长期预报;互信息法;相关系数;主成分;神经网络
中图分类号:P338 **文献标志码**:A **文章编号**:1001-5485(2018)08-0017-05

1 研究背景

准确的径流预报不仅有助于合理配置水资源,更好地支撑所在区域的工农业生产,而且可以帮助减少气象灾害带来的损失。目前,许多预报模型和组合模型都已应用在中长期径流预报中,如人工神经网络模型[1]、灰色-周期外延模型[2]、组合小波神经网络模型[3]等。人们对各种预报方法及模型进行研究发现,模型的输入变量在一定程度上影响着水文预报的精度[4]。不同研究区其流域径流量受不同相关因子的影响程度也大不相同,因此对研究区相关预报因子的选择成为影响该研究区径流预报水平的关键因素。目前,已有学者对如何选择径流预报因子的问题进行研究:朱永英等[5]借助粗集理论对预报因子进行优化和选择,提高了大伙房水库的中长期径流预报精度;闪丽洁等[6]运用不同方法优选长江流域预报因子建立人工神经网络模型进行径流预报,对比得到精度最高的预报因子挑选方法。然而,目前基于互信息理论进行预报因子选择的研究较少。赵铜铁钢等[7]运用了互信息法确定预报因子对长江各水文站建立神经网络径流预报模型;卢迪等[8]采用互信息量方法筛选预报因子作为BP神经网络的输入数据,对碧流河流的汛期径流进行中长期预报。研究表明互信息量方法可以识别出预报因子与径流的复合相关性,运用互信息挑选出的预报因子作为模型输入可以大大提高模型的预测精度。

新疆天山西部山区河流的补给以融雪为主,区域的水资源相对匮乏,准确地预知该区域的径流量尤其重要,既能够支持当地工农业生产,又对当地社会安定和合理安排水资源起着决定性作用[9]。鉴于不同区域产汇流规律、地貌特征及人类活动等的差异性,目前已有模型并不具有通用性。因此开展天山西部山区水文中长期预报的研究具有重要的意义。本文针对天山西部山区融雪径流预报中气象预报因子的选择问题,基于互信息理论对神经网络预报模型的输入因子选择及衡量因子间复合相关关系的方法进行研究与讨论。首先,通过不同的方法初步选择神经网络模型的输入因子,然后通过不同神经网络模型进行径流预测,并进行不同方案的比较,以期了解不同方法的优劣。

2 研究区概况及数据来源

2.1 研究区概况

喀什河流域位于天山西部,属于伊犁河的支流,

收稿日期:2017-01-01; **修回日期**:2017-02-22
基金项目:国家自然科学基金项目(51469034,51209181);新疆自治区地方公派出国留学成组配套项目(XJDF201307);新疆水文学及水资源重点学科基金项目(xjswszyzdxk20101202)
作者简介:周育琳(1992-),女,福建泉州人,博士研究生,主要从事水文水资源研究。E-mail:597049304@qq.com
通信作者:穆振侠(1980-),男,山东莒县人,副教授,博士,主要从事水文水资源研究。E-mail:muzhenxia@126.com

整个流域全长约为304 km,面积约为9 541 km^2。流域上仅有一个乌拉斯台水文站以及临近的尼勒克、伊宁气象站,基于站点1960—2005年的数据统计:流域的多年平均径流量为102.2 m^3/s;多年平均降雨量为561.7 mm;春季最高气温为8.97 ℃,最低气温为3.03 ℃;夏季最高气温为18.33 ℃,最低气温为14.77 ℃;秋季最高气温为8.40 ℃,最低气温为3.40 ℃;冬季最高气温为-4.83 ℃,最低气温为-13.13 ℃;多年平均气温为5.39 ℃。

2.2 数据来源

借助水文数据、气象数据、探空数据等对研究区融雪为主河流的径流量进行中长期水文预报研究。数据主要来源如表1所示。

表1 数据来源

Table 1 Sources of data

数据类型	站点	统计时段	因子	数据来源
水文数据	乌拉斯台水文站	1960—2005年逐月	降水、气温	新疆水文年鉴
气象数据	尼勒克气象站	1960—2005年逐月	水汽压、相对湿度、蒸发、日照时数、风速	新疆气象年鉴
探空数据	伊宁气象站	1960—2005年逐月	500,700,850 hPa 3个等压面所对应的高度和气温数据	新疆气象年鉴

注:1 hPa= 100 Pa

太阳活动也影响着河川径流变化,因此收集北半球1960—2005年逐月的太阳黑子数作为预报的影响因子。表1中有13个影响因子,加上太阳黑子数合计为14个与研究区径流相关的影响因子。

3 预报因子的确定

3.1 互信息理论

互信息是一种信息度量,可以用来表示2个或多个变量之间的相关性,而且能反映变量间线性相关关系之外的非线性相关关系。如果变量 X,Y 互不相关,则 X,Y 的联合分布密度等于边缘分布密度之积,可表示为

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y) \quad 。 \tag{1}$$

式中:$p_X(x)$ 为 X 的概率密度;$p_Y(y)$ 为 Y 的概率密度;$p_{X,Y}(y)$ 为 X 与 Y 的联合分布密度。

给定 N 个离散观测样本,变量 X,Y 之间互信息计算公式为

$$\mathrm{MI} = \frac{1}{N}\sum_{i=1}^{N} \ln \frac{p_{X,Y}(x,y)}{p_X(x_i)\, p_Y(y_i)} \quad 。 \tag{2}$$

式中MI为互信息量值。

由式(2)可知,当 X 与 Y 互不相关时,MI取值将趋近于0;当 X 与 Y 之间存在函数关系时,MI取值将趋近于正无穷大。

若给定变量 X 的 N 个观测样本,其概率密度 $p_X(x_i)$ 采用核函数(多维高斯分布密度函数),进行估计,即

$$\widehat{p_X}(x_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{(2\pi)^{\frac{d}{2}}\, \lambda^d\, (\det \boldsymbol{S})^{\frac{1}{2}}} \cdot \exp\left\{ \frac{-[x_i - x_j]^{\mathrm{T}}\, \boldsymbol{S}^{-1}\, [x_i - x_j]}{2\lambda^2} \right\} \quad 。 \tag{3}$$

式中:i,j 为观测样本编号;T为向量转置的符号;$\widehat{p_X}(x_i)$ 为 x_i 处的密度函数估计值;d 为向量 $\boldsymbol{X}$ 的维数;$\boldsymbol{S}$ 为向量 $\boldsymbol{X}$ 的协方差矩阵;$\det \boldsymbol{S}$ 为 $\boldsymbol{S}$ 的行列式;λ 为窗口宽度(bandwidth),一般按经验取为

$$\lambda = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} N^{-\frac{1}{d+4}} \quad 。 \tag{4}$$

借助水文数据、气象数据和探空数据共14个相关因子水文序列分别计算与径流序列的互信息量值MI,结果如表2所示。

表2 预报因子与径流序列的互信息量值

Table 2 Coefficient matrix of factor score with MI (mutual information)

相关因子	MI	名次	相关因子	MI	名次
气温	4.25	4	850 hPa高度	0.16	14
太阳黑子	0.68	10	850 hPa气温	4.48	2
降水	4.38	3	水汽压	2.25	9
500 hPa高度	0.30	12	相对湿度	3.68	7
500 hPa气温	4.05	6	蒸发	4.49	1
700 hPa高度	0.43	11	日照时数	2.64	8
700 hPa气温	4.23	5	风速	0.21	13

表2包括14个相关因子与径流序列的互信息MI值以及所占名次(名次是MI值由大到小排序,即名次1为MI值最大所对应的因子)。由表2可见,蒸发、850 hPa气温、降水、气温、700 hPa气温、500 hPa气温这6个相关因子的MI值都>4,是14个相关因子中的前6名,也就是相关性最好的6个因子;相对湿度与径流序列的互信息MI值为3.68,是14个相关因子中的第7名,相关性较好仅次于第6名的MI值0.37。选取互信息MI值最好的7个相关因子(蒸发、850 hPa气温、降水、气温、700 hPa气温、500 hPa气温、相对湿度)作为预报模型的输入变量。

3.2 预报因子的挑选

目前对预报因子的挑选已有大量研究,本文选取以下挑选方法进行预测:

(1)全部预报因子法(全因子法)。借助水文数据、气象数据和探空数据共14个相关因子水文序列直接作为径流预测的预报因子。

(2)相关系数法。对选取的14个相关因子水文序列分别计算与径流序列的相关系数值,选取 R^2 最

大的前 7 个因子作为预报因子。

(3)主成分分析法。运用主成分分析法对 14 组相关因子数据进行主成分提取,提取出 3 组代表 14 个相关因子的主要成分 X_1,X_2,X_3。

(4)互信息法。借助互信息法计算 14 组预报因子与径流之间的互信息量值 MI,选取 MI 最大的 7 个预报因子。

4 种方法的具体因子见表 3。

表 3　乌拉斯台径流预报因子的优选结果

Table 3　Optimized predictors for runoff at Wulasitai Station

预测方法	预报因子
全因子法	气温、降水、每月太阳黑子数、500 hPa 高度、500 hPa 气温、700 hPa 高度、700 hPa 气温、850 hPa 气温、850 hPa 高度、水汽压、相对湿度、蒸发、日照时数、风速
相关系数法	水汽压、500 hPa 气温、气温、500 hPa 高度、降水、日照时数、相对湿度
主成分分析法	X_1,X_2,X_3
互信息法	蒸发、气温、700 hPa 气温、500 hPa 气温、850 hPa 气温、降水、相对湿度

4　实　例

本文通过相关系数法、互信息法及主成分分析方法进行预报因子的优选,确定出 3 组不同的预报因子与不进行因子挑选的全部因子(见表 3),将这 4 组数据作为模型的输入因子。并采用组合小波 BP 神经网络模型与 RBF 神经网络模型对径流序列进行预测。

4.1　不同神经网络预测模型

4.1.1　组合小波 BP 神经网络模型

以 1960 年 1 月—1999 年 12 月逐月预报因子序列和径流序列作训练和测试数据, 2000 年 1 月—2005 年 12 月的径流序列作为检验数据,建立组合小波 BP 神经网络模型进行中长期水文预报研究。4 种挑选方法的预测结果如图 1,相对误差如表 4。

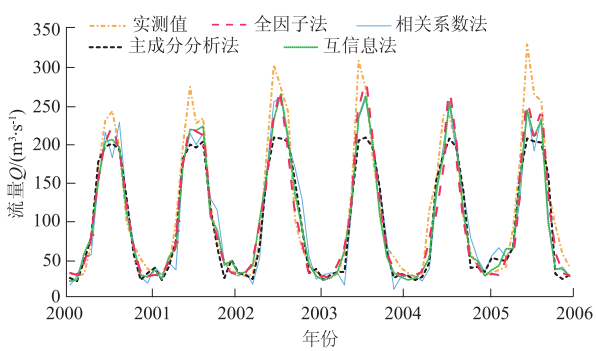

图 1　乌拉斯台站径流小波 BP 神经网络预测结果

Fig.1　Runoff predictions for Wulasitai Station by wavelet BP neural network

表 4　乌拉斯台站径流小波 BP 神经网络预测结果

Table 4　Runoff predictions for Wulasitai Station by wavelet BP neural network

年份	与实测值的平均相对误差/%			
	全因子法	相关系数法	主成分分析法	互信息法
2000	−7.50	−8.42	−5.16	−4.03
2001	−12.90	−3.53	−9.89	−1.42
2002	−1.41	2.93	−3.43	7.50
2003	−7.46	−12.99	−13.83	−10.55
2004	−7.51	−2.41	−9.77	−5.40
2005	−19.51	−6.75	−11.16	−13.48
平均值	−9.38	−5.20	−8.87	−4.56

由图 1 可见:①4 种挑选方法对径流量低值的预测效果都较好,高值预测都不太理想、拟合效果不好;②其中全因子方法的拟合效果最差,对于高值与低值的预测都出现较大的偏差;③互信息法不仅对径流量的高、低值的拟合效果在 4 种挑选方法里最好,而且整个径流过程的拟合程度也较高。因此由互信息法挑选出的预报因子建立的组合小波 BP 神经网络模型预测出的径流量与实测值最为接近。

由表 4 可以看出:①全因子的预测结果的相对误差为−19.51%～−1.41%,平均为−9.38%; ②相关系数法的预测结果的相对误差为−12.99%～2.93%,平均为−5.20%;③主成分分析法的预测结果的相对误差为−13.83%～−3.43%,平均为−8.87%;④互信息法的预测结果的相对误差为−13.48%～7.50%,平均为−4.56%;⑤互信息法在这 6 a 的平均相对误差最小,较全因子法少4.82%,较相关系数法少0.64%,较主成分分析法少4.31%,可作为组合小波 BP 神经网络预报模型的最优预报因子挑选方案。

4.1.2　RBF 神经网络模型

分别采用全因子、相关系数法、主成分分析法和互信息法等方法确定的预报因子,具体见表 3。以 1960 年 1 月—1999 年 12 月逐月的预报因子序列和径流序列作为训练和测试数据, 2000 年 1 月—2005 年 12 月的径流序列作为检验数据,建立 RBF 神经网络模型进行中长期水文预报研究。4 种挑选方法的预测结果如图 2,相对误差如表 5。

由图 2 可见:①4 种挑选方法对径流量低值与高值的预测效果都较好;②其中全因子法和主成分分析法对高值的拟合效果较差;③互信息方法对径流量的高、低值的拟合效果在 4 种挑选方法里最好,整个径流过程的拟合程度也较高。因此由互信息方法挑选出的预报因子建立的 RBF 神经网络模型预测出的径流量与实测值最为接近。

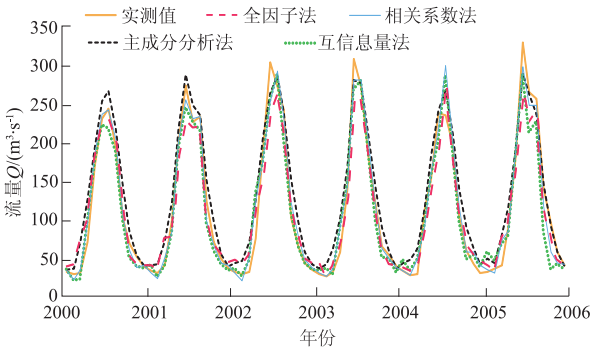

图 2 乌拉斯台站径流 RBF 神经网络预测结果

Fig.2 Runoff predictions for Wulasitai Station by RBF neural network

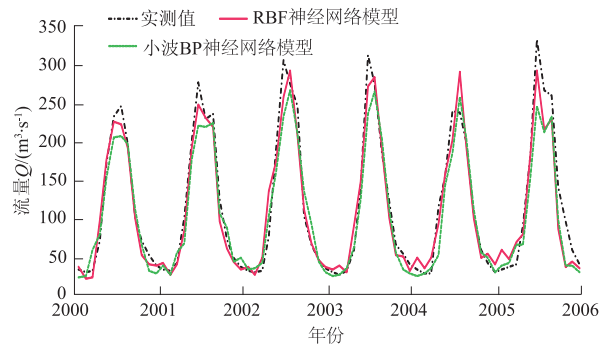

图 3 乌拉斯台站径流预测结果

Fig.3 Result of runoff predictions for Wulasitai Station

表 5 乌拉斯台站径流 RBF 神经网络预测结果

Table 5 Runoff predictions for Wulasitai Station by RBF neural network

年份	与实测值的平均相对误差/%			
	全因子法	相关系数法	主成分分析法	互信息法
2000	−14.99	−1.08	−30.90	5.65
2001	−5.56	4.16	−22.10	7.35
2002	−19.23	−11.07	−37.10	−9.82
2003	−5.33	−5.26	−28.92	−3.03
2004	−11.75	−13.69	−36.21	−13.33
2005	0.01	1.52	−15.57	1.48
平均值	−9.48	−4.24	−28.47	−1.95

计算出 2000—2005 年每年的平均相对误差(表 5)可以看出:①全因子法预测结果的相对误差最大为−19.23%,最小为0.01%,平均为−9.48%;②相关系数法预测结果的相对误差为−13.69%~4.16%,平均为−4.24%;③主成分分析法的预测结果的相对误差为−37.10%~−15.57%,平均相对误差为−28.47%;④互信息法的预测结果的相对误差为−13.33%~7.35%,平均为−1.95%;⑤互信息法在这 6 a 的平均相对误差最小,相对于全因子法减少7.53%,相对于相关系数法减少2.29%,相对于主成分分析法减少26.52%,可作为 RBF 神经网络预报模型的最优预报因子挑选方案。

4.1.3 不同方案的预测结果分析

将各挑选结果分别作为 2 个神经网络模型输入数据,径流序列作为输出数据。得到不同预报结果的相对误差如表 6 所示,将互信息法作为最佳预报因子的挑选方法分别得到 2 个模型的预报结果如图 3所示。

表 6 不同方案下乌拉斯台站径流的预测结果

Table 6 Runoff prediction errors for Wulasitai Station by different methods

预测方法	平均相对误差/%	
	小波 BP 神经网络	RBF 神经网络
全因子法	−9.38	−9.48
相关系数法	−5.20	−4.24
主成分分析法	−8.87	−28.47
互信息法	−4.56	−1.95

从表 6 可以看出,不同的输入数据对 2 种不同模型的预测结果都有影响。其中互信息方法挑选出的预报因子作为输入因子的预测效果要好于其他方法挑选预报因子作为输入数据的预测效果。在小波 BP 神经网络模型预测结果中,互信息法的预测结果平均相对误差相对于全因子法减少4.82%,相对于相关系数法减少0.64%,相对于主成分分析法减少4.31%。从图 3 中可以看出,RBF 神经网络对径流高、低值及径流过程的拟合效果都比小波 BP 神经网络要理想。

5 结 论

基于相关系数法、互信息法和主成分分析法 3 种方法优选预报因子,以及全因子法得到 4 种不同的预报因子挑选结果。将这 4 种预报因子的挑选作为输入因子,径流数据作为输出因子对 RBF 神经网络和组合小波 BP 神经网络进行建模得到 8 种预测结果。可以看出:

(1)不同预报因子在神经网络模型中的预测结果都不同,不同预报因子挑选方法较不进行因子筛选的全因子法相比,其预报精度有着不同程度(0.51%~7.53%)的提高,因此在进行径流中长期预报中重视对预报因子的挑选,可以在一定程度上提高预测精度得到更高的合格率。

(2)互信息量可同时反映出预报因子与径流序列之间的线性关系和非线性关系,较相关系数法、全因子法、主成分分析法的预测结果,平均相对误差大大减少。在小波 BP 神经网络模型预测结果中,互信息法较其他方法的预报精度提高了0.64%~4.82%;在 RBF 神经网络模型预测结果中,互信息法较其他方法的预报精度提高了2.29%~26.52%。因此选用互信息量挑选预报因子作为模型输入可以提高模型的预报精度。

(3)以相对误差作为评价模型精确度的标准得到结果分析,基于互信息方法挑选出的预报因子作为 RBF 神经模型输入数据的模型预测精度最高,较互信息法结合小波 BP 神经网络模型的预测精度提高了2.61%,因此在天山西部山区该方法对径流中长期水文预报研究具有一定参考价值。

从模拟的结果来看,尽管预报模型在天山西部山区具有较好的适用性,但模拟径流值与实测径流值之间仍然存在一定的误差,所有的大误差都是出现在尖峰处,该误差可能来自预报因子的挑选方面,相关系数法只能挑选出与径流序列线性关系较好的因子,主成分分析法也只能挑选出因子的主要成分,即便是互信息方法对预报因子的挑选也存在不足,或者还有影响尖峰变化的因素没有考虑进去,对水文预报精度制约的关键因素是模型输入因子的确定。

在今后的研究中,如果能找到精度更好的挑选因子方法,或加入影响尖峰变化的因素,并充分考虑研究区的实际情况建立带有递归的动态神经网络模型,模型模拟精度可能会进一步提高,以期更好地指导所在区域融雪径流模拟研究与洪水预报方面的工作。

参考文献:

[1] 张　伟.基于人工神经网络的径流预测研究[D].石河子:石河子大学,2008:45-60.

[2] 雷　杰,彭　杨,纪昌明.基于改进灰色-周期外延模型的中长期水文预报[J].人民长江, 2010,41(24):28-31.

[3] 王秀杰,封桂敏,耿庆柱.小波分析组合模型在日径流预测中的应用研究[J].自然资源学报, 2014,29(5):885-893.

[4] 桑燕芳,王　栋,吴吉春,等.基于 WA、ANN 和水文频率分析法相结合的中长期水文预报模型的研究[J].水文, 2009,29(3):12-15.

[5] 朱永英,周惠成,彭　慧.粗集一模糊推理技术在水文中长期预报中的应用研究[J].水力发电学报, 2009, 28(1):45-50.

[6] 闪丽洁,张利平,刘　恋,等.基于多方法优选因子和人工神经网络耦合模型的枯水期径流预报[J].武汉大学学报(工学版), 2015, 48(6): 758-763.

[7] 赵铜铁钢,杨大文.神经网络径流预报模型中基于互信息的预报因子选择方法[J].水力发电学报, 2011, 30(1):24-30.

[8] 卢　迪,周惠成.基于互信息与 BP 神经网络的中长期径流预报方法研究[J].水文, 2014,34(4):8-14.

[9] 穆振侠,姜卉芳.高寒山区降水规律及融雪径流模拟[M].北京:中国水利水电出版社,2015:13.

(编辑:占学军)

Mid-term and Long-term Hydrological Forecasting of Snowmelt Runoff in Western Tianshan Mountains Based on Mutual Information and Neural Network

ZHOU Yu-lin, MU Zhen-xia, PENG Liang, GAO Rui , YIN Zi-yuan, TANG Rui

(College of Water Conservancy and Civil Engineering, Xinjiang Agricultural University, Urumqi　830052, China)

Abstract: The aim of this research is to improve the accuracy of forecasting snowmelt runoff in the mountainous areas of western Tianshan Mountains, and to better support the development of industrial and agricultural production in the study area. The predictor, which is a key issue affecting forecast accuracy, are optimized and selected by using mutual information, correlation coefficient method, and principal component analysis method. The selected predictors are taken as input factors in RBF neural network model and combinatorial wavelet BP neural network model for comparison. Results suggest that: 1) optimized predictors selected by the mutual information method could improve forecast accuracy; 2) according to forecast results under different scenarios, the results of RBF neural network model is superior to those of combinatorial wavelet BP neural network model; 3) with relative error as the standard of accuracy evaluation, RBF neural network model with input factors selected by mutual information method could produce the optimum forecast result.

Key words: mid-term and long-term hydrological forecasting; mutual information; coefficient of correlation; principal component analysis; neural network