中国农业大学学报 2011,16(5):157-162
Journal of China Agricultural University

基于 SVM 分类的淮河流域夏季降水预测模型

吴有训[1] 刘 勇[2] 叶金印[3] 余品忠[1]
(1.安徽省宣城市气象局,安徽 宣城 242000; 2.安徽省气象局,合肥 230061;
3.淮河流域气象中心,安徽 蚌埠 233040)

摘 要 采用 1959—2009 年逐月 74 项大气环流特征量序列、500 hPa 月平均高度场和月平均海温场,计算与预报对象淮河流域夏季降水量相关系数,选取预测因子;用主分量分析方法组合预测因子。用支持向量分类机方法分别建立山东淮河流域、河南淮河流域、江苏淮河流域、安徽淮河流域共 4 个区域夏季降水短期气候预测模型。对 2007—2009 年夏季降水量 SVM 分类预测,4 个区域的训练集回预测正确率为 85%~99%,平均训练集回预测正确率 91%;预测结果误差最大不超过 1 级,绝对值平均为 0.4 级。结果表明,该模型具有较强的预测能力和推广前景,可在气候预测业务中使用。

关键词 支持向量分类机 (SVM); 淮河流域; 夏季降水; 短期气候预测
中图分类号 P 457.6 **文章编号** 1007-4333(2011)05-0157-06 **文献标志码** A

Prediction model for summer precipitation in the huaihe river basin based on support vector machine

WU You-xun[1], LIU Yong[2], YE Jin-yin[3], YU Pin-zhong[1]
(1. Xuancheng Meterorological Bureau, Anhuiprovince, Xuancheng 242000, China;
2. Anhui Meterorological Observatory, Hefei 230061, China;
3. Huaihe River Basin Meteorological Center, Bengbu 233040, China)

Abstract Based on Support Vector Machine (SVM), four short-range summer precipitation prediction models were established for four areas in the Huaihe River basin, respectively. Using the monthly data of 74 circumfluent eigen values, the monthly data of sea surface temperature, the monthly data of 500 hPa height from 1959 to 2009, forecast factors were chosen. Combination of the forcast factors was done by using Principal component analysis. A categorical prediction was performed on summer precipitation data from 2007 to 2009. The results show that the accuracy of four regional training set to predict is 85%-99% and the mean accuracy is 91%, and the lagest grade of error of summer precipitation prediction is no more than 1, and the mean absolute grade is 0.4. These results indicate that the short-range climatic prediction model based on Support Vector machine has a good performance on predictions of summer precipitation.

Key words support vector machine (SVM); Huaihe river basin; summer precipitation; short-range climatic prediction

淮河流域位于我国东部 (111°15′~121°25′ E, 30°55′~36°36′ N),是东亚季风影响区域,夏季降水时空分布不均,年际差异大,降水量多。20 世纪 90 年代以来,淮河流域先后多次发生大洪水,造成巨大损失[1-3]。做好淮河流域夏季降水量短期气候预测具有重要意义。

目前短期气候预测业务技术多用统计预报方法,预测能力受到限制。神经网络方法基于经验风险最小化,易陷入局部最优,训练结果不太稳定,一般需要大样本容量;而支持向量机(support vector

收稿日期: 2011-02-22
基金项目: 中国气象局气象新技术推广项目资助(CMATG2005M34)
第一作者: 吴有训,高级工程师,主要从事天气气候分析和预报研究,E-mail:wuyouxun@yahoo.com.cn

machines,SVM)是 Vapnik 提出的统计学习理论学习方法,它基于结构风险最小化原则,具有严格的理论和数学基础,泛化能力优于前者,算法具有全局最优性,是针对小样本容量统计的理论;应用支持向量机预测夏季降水是一种有效方法[4-6]。

应用 SVM 理论进行建模,用 SVM 回归方法对夏季降水进行预测[7],预测精度受到限制;本研究拟采用主分量分析方法组合预测因子[8],旨在浓缩预测信息,减少计算量,且降维去噪,提高泛化性能;并用 SVM 分类方法建立预测模型,以期提高夏季降水预测能力。

1 支持向量分类机(SVM)夏季降水气候预测模型的建立

1.1 预测对象和预测因子的选取及主分量分析

淮河流域受西风带天气系统与西进北上的西太平洋副热带高压影响;同时受流域西部、西南部及东北部山区、丘陵和广阔的平原的复杂地形干扰,致使降水地理分布差异较大,综合考虑淮河流域各地的降水气候特点和气象服务的需要,将淮河流域分为 4 个区域(图 1):1)山东淮河流域,代表站为日照、临沂、枣庄、济宁、菏泽;2)河南淮河流域,代表站为郑州、驻马店、信阳;3)江苏淮河流域,代表站为连云港、淮阴、盐城、扬州、徐州;4)安徽淮河流域,代表站为宿县、蚌埠、淮南、亳州、阜阳、六安。分别以各区域代表站的夏季(06—08 月,下同)降水量平均值距平百分率作为预测对象(决策属性)。

影响夏季降水的因子很多,包括大气环流因子如西太平洋副热带高压的活动、极涡强度等因子[9-12],也包括一些非环流因子的间接影响[13-14]。74 项大气环流特征量多为具有天气气候学意义的区域天气系统指标,是间接利用高度场网格点的高度

图 1 淮河流域 06—08 月降水分片划分

Fig. 1 Regionalism of the Summer(Jun-Aug) precipitation in the Huaihe River basin

值组成的区域综合因子[15-16],因此用各月 74 项大气环流特征量、北半球 500 hPa 月平均高度场和太平洋月平均海温场①,与预测对象的相关系数计算,分别得到 4 个区域预测因子(特征属性)(表 1)。江苏淮河流域的降水预测因子为 20 个,其他区域因子均为 18 个;相关系数绝对值最大为 0.60;相关系数绝对值最小为 0.31,其中出现在 74 项大气环流特征量因子有 3 个,出现在 500 hPa 高度场因子有 1 个。表 2 给出预报因子 I_i 的具体大气环流特征量。

安徽淮河流域用 1956—2006 年的资料作为训练样本,河南、江苏、山东淮河流域用 1960—2006 年的资料作为训练样本;2007—2009 年资料作为测试样本。考虑到直接用预测因子作为输入,不仅计算量大,而且影响预测模型的泛化性能,因此采用主分量分析方法,浓缩预测因子。设预报因子 p_k 为 m 个,样本容量为 n,其观测值为 p_{ki}($k=1, 2, \cdots, m$;

①74 项大气环流特征量资料、北半球 500 hPa 月平均高度场和太平洋月平均海温场资料取自国家气候中心气候诊断室;环流特征量由国家气候中心气候预测室计算、整理。

表 1　淮河流域 06—08 月降水量与前期预报因子的相关系数

Table 1　Correlation coefficients between previous predictors and the Summer(Jun-Aug) precipitation in the Huaihe River basin

山东淮河流域		河南淮河流域		江苏淮河流域		安徽淮河流域	
预报因子	相关系数	预报因子	相关系数	预报因子	相关系数	预报因子	相关系数
I_1	−0.42	I_1	0.35	I_1	−0.37	I_1	0.34
I_2	0.34	I_2	0.36	I_2	0.36	I_2	−0.31
I_3	−0.34	I_3	−0.49	I_3	0.36	I_3	−0.36
I_4	0.32	I_4	0.40	I_4	0.36	H_4	0.60
I_5	0.41	I_5	0.31	I_5	−0.37	H_5	0.47
H_6	0.39	I_6	0.31	H_6	0.50	H_6	−0.34
H_7	0.39	H_7	0.37	H_7	0.47	H_7	0.35
H_8	0.34	H_8	0.44	H_8	−0.47	H_8	0.47
H_9	0.32	H_9	0.36	H_9	0.41	H_9	0.36
H_{10}	0.34	H_{10}	0.31	H_{10}	−0.33	H_{10}	0.32
H_{11}	0.38	H_{11}	0.56	S_{11}	0.37	S_{11}	0.40
S_{12}	−0.36	H_{12}	0.36	S_{12}	−0.34	S_{12}	−0.41
S_{13}	0.41	H_{13}	0.35	S_{13}	0.34	S_{13}	0.40
S_{14}	−0.41	H_{14}	0.40	S_{14}	0.40	S_{14}	−0.37
S_{15}	0.37	H_{15}	0.51	S_{15}	0.44	S_{15}	0.35
S_{16}	0.47	H_{16}	0.39	S_{16}	−0.39	S_{16}	−0.39
S_{17}	−0.59	S_{17}	−0.39	S_{17}	0.41	S_{17}	0.46
S_{18}	−0.34	S_{18}	0.35	S_{18}	−0.42	S_{18}	−0.42
				S_{19}	0.40		
				S_{20}	−0.47		

注：I_i 为 74 项大气环流特征量资料中表征大气环流的因子；H_i 为 500 hPa 高度场因子；S_i 为海温场因子。

$i=1,2,\cdots,n$)。m 个预报因子线性组合成一新因子 z：

$$z = v_1 p_1 + v_2 p_2 + \cdots + v_m p_m = \boldsymbol{v}'\boldsymbol{p} \quad (1)$$

式中：v_k 为组合系数；$\boldsymbol{v}$ 为组合系数向量；$\boldsymbol{p}$ 为预报因子向量。如果 z 满足方差极大的要求，则称 z 为 m 个变量的主分量，即

$$\boldsymbol{s}_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \rightarrow \text{极大} \quad (2)$$

式中：$\boldsymbol{s}_z^2$ 为方差。满足条件 $\boldsymbol{v}'\boldsymbol{v} = 1$ 的求极值问题，即

$$Q = \boldsymbol{v}'\boldsymbol{S}\boldsymbol{v} - \lambda(\boldsymbol{v}'\boldsymbol{v} - 1) \quad (3)$$

$$\frac{\partial Q}{\partial v} = 0 \quad (4)$$

$\boldsymbol{S}$ 为 $\boldsymbol{p}$ 的 m 个变量协方差阵，问题转化为求矩阵 $\boldsymbol{S}$ 的 m 个特征值 λ 及其对应的特征向量 $\boldsymbol{v}$。取方差占总方差的百分率为 97%的主分量作为输入。

表 2 预报因子 I_i 的各月 74 项大气环流特征量

Table 2 Predictor I_i and monthly data of 74 characteristic quantities of northern hemispheric circulations

区域	因子	月份	环流特征量	区域	因子	月份	环流特征量
山东淮河流域	I_1	1	东亚槽位置(CW)	江苏淮河流域	I_1	1	西太平洋副高脊线(110° E～150° E)
	I_2	2	东亚槽强度(CQ)		I_2	1	西太平洋副高北界(110° E～150° E)
	I_3	2	亚洲经向环流指数(IM,60° E～150° E)		I_3	1	东亚槽强度(CQ)
	I_4	11	南方涛动指数		I_4	1	大西洋欧洲环流型 C
	I_5	12	南方涛动指数		I_5	2	冷空气
河南淮河流域	I_1	2	北美副高面积指数(110° W ～60° W)	安徽淮河流域	I_1	1	东亚槽强度(CQ)
	I_2	2	北美副高强度指数(110° W～60° W)		I_2	1	冷空气
	I_3	10	亚洲区极涡强度指数(1 区,60° E～150° E)		I_3	2	北半球极涡强度指数(5 区,0°～360°)
	I_4	11	东太平洋副高脊线(175° W～115° W)				
	I_5	11	欧亚经向环流指数(IM,0°～150° E)				
	I_6	12	太平洋副高面积指数(110° E～115° W)				

1.2 SVM 分类降水预测模型的设计

1.2.1 对降水量分类建立样本集

对 06—08 月各区域平均降水量的 50(安徽淮河流域 54)个样本,根据标准值(1971—2000 年降水量平均值)计算得各区域平均降水量距平百分率序列并分类。分类标准见表 3。

表 3 淮河流域 06—08 月降水量距平百分率分类标准

Table 3 Category of Summer(Jun-Aug) precipitation in Huaihe River basin

降水量距平百分率/%	类别(级别)
<－60	1
－60～－20	2
－20～20	3
20～60	4
>60	5

1.2.2 确定核函数建立预测模型

选取径向基函数作为 SVM 分类方法的核函数建立预测模型。径向基函数为

$$K(x_i,x)=\exp(-\gamma\|x-x_i\|^2),\gamma>0 \quad (5)$$

式中:x_i 为因子序列;γ 为核参数。采用交叉验证方法[17]计算验证集的分类准确率,优选核函数中参数 γ 及惩罚因子 C;对训练样本进行学习,求出阈值 b^* 及支持向量,确定预测模型;将测试样本代入预测模型,求出预测值。构建 SVM 分类预测模型时采用了 Libsvm-2.89 软件[17]。

2 结果与分析

SVM 分类预测降水量距平百分率类别时,用交叉验证方法优选参数惩罚因子 C 和径向基函数参数 γ。图 2 示出山东淮河流域 2007—2009 年各年 06—08 月降水预测时最优 C、γ 及分类准确率值。可见,验证集的分类准确(预测等级与实况等级完全

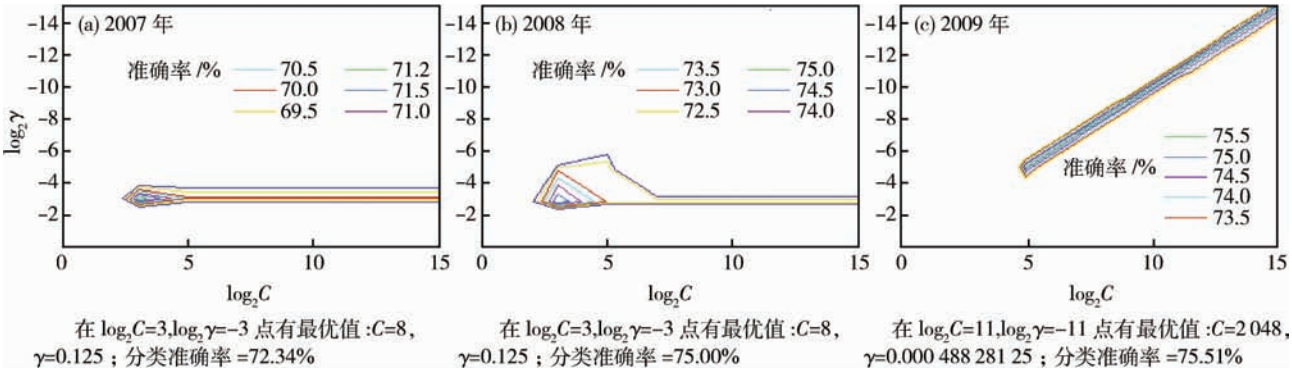

图 2　山东淮河流域夏季降水预测模型优选参数 C、γ 和验证集分类准确率

Fig. 2　Preferred parameters of C and γ and classified forecast accuracy in prediction model of Summer precipitation in Huaihe River basin of Shandong

一致为准确）率分别为 72.34%、75.00% 和 75.51%，表明分类器具有较好的性能指标。

表 4 示出淮河流域 4 个区域 06—08 月降水量距平百分率类别 SVM 分类预测结果。山东淮河流域 2007—2009 年夏季降水量训练集回预测平均准确率为 99%，比其他区域高；对 2007 年降水实况偏多(4 类)，预测正确；2008 年、2009 年实况为偏多和特多，预测结果均比实况偏小 1 级，但定性评价正确；3 年误差绝对值平均为 0.7 级，比其他区域大。河南淮河流域 2007—2009 年夏季降水量训练集回预测平均准确率为 85%，为 4 区域中最低；对 2007 年、2009 年预测正确；2008 年降水实况偏多(4 类)，预测结果比实况偏小 1 级；3 年误差绝对值平均为 0.3 级。江苏淮河流域 2007—2009 年夏季降水量训练集回预测平均平均准确率为 91%；对 2007 年、2009 年预测正确；2008 年降水实况偏多(4 类)，预测结果比实况偏小 1 级；3 年误差绝对值平均为 0.3 级，预测结果和实况同河南淮河流域完全一样。安

表 4　SVM 分类预测模型对淮河流域 2007—2009 年 06—08 月降水量预测结果

Table 4　Forecasting results of Summer(Jun-Aug) precipitation and threshold from 2007 to 2009 in Huaihe River basin by using SVM method

区域	年份	准确率/%($n_{正确}/n_{总}$)①	预测类别	实况类别	绝对误差
山东淮河流域	2007	100(47/47)	4	4	0
	2008	100(48/48)	3	4	−1
	2009	96(47/49)	4	5	−1
河南淮河流域	2007	79(37/47)	4	4	0
	2008	77(37/48)	3	4	−1
	2009	100(49/49)	3	3	0
江苏淮河流域	2007	85(40/47)	4	4	0
	2008	88(42/48)	3	4	−1
	2009	100(49/49)	3	3	0
安徽淮河流域	2007	96(49/51)	5	4	+1
	2008	92(48/52)	3	3	0
	2009	77(41/53)	3	3	0

注：①$n_{正确}$ 和 $n_{总}$ 分别为训练集回预测正确样本数和总样本数。误差绝对值山东淮河流域 3 年平均为 0.7 级，其他区域 3 年平均为 0.3 级；4 区域平均为 0.4 级。

徽淮河流域2007—2009年夏季降水量训练集回预测平均准确率为89%;对2008年、2009年预测正确;2007年降水实况偏多(4类),预测结果比实况偏大1级,定性评价为正确;3年误差绝对值平均为0.3级。

表4数据表明预测模型运行稳定,预测准确率高。预测误差一般为负值,即预测类别比实况类别小,4区域的3年预测误差绝对值平均仅0.4级;对降水正常年份预测正确,偏多、特多年份预测正确或者误差不超过1级,模型具有一定预测能力。

3 结束语

本研究在用主分量分析方法组合预测因子的基础上,构建了基于SVM分类的夏季降水预测模型,主要结论如下:

1)在夏季降水预测模型的建立中,不仅可以从500 hPa月平均高度场和太平洋月平均海温场选取预报因子,74项大气环流特征量多为具有天气气候学意义的区域天气系统指标,这种综合性的大气环流因子也存在淮河流域夏季降水前期重要信号。

2)基于人工神经网络方法夏季降水预测模型,对训练集回预测正确率一般为100%,实际预测结果误差有时较大,显得不太稳定。应用SVM分类的夏季降水预测模型,虽然对训练集回预测正确率不能达到100%,但实际预测结果误差较小,基于SVM分类的夏季降水预测模型的实际应用效果要优于前者。

3)2007—2009年夏季降水量分类预测,淮河流域4个区域的训练集回预测正确率85%~99%,预测结果误差绝对值平均0.4级,表明模型具有较强的预测能力,可在气候预测业务中使用。

参 考 文 献

[1] 刘淑媛,郑永光,王洪庆,等.1998年6月28日—7月2日淮河流域暴雨分析[J].气象学报,2002,60(6):774-779

[2] 矫梅燕,金荣花,齐丹.2007年淮河暴雨洪涝的气象水文特征[J].应用气象学报,2008,19(3):257-264

[3] 毕宝贵,矫梅燕,廖要明,等.2003年淮河流域大洪水的雨情、水情特征分析[J].应用气象学报,2004,15(6):681-687

[4] 杨志民,田英杰.基于模糊系数规划的模糊支持向量分类机[J].中国农业大学学报,2007,12(5):79-85

[5] 刘广利,邓乃扬.基于SVM分类的预警系统[J].中国农业大学学报,2002,7(6):97-100

[6] 邓乃扬,田英杰.数据挖掘的新方法——支持向量机[M].北京:科学出版社,2004:164-223

[7] 张礼平,陈永义,周筱兰.支持向量机(SVM)及其在场预测中的应用[J].热带气象学报,2006,22(3):278-282

[8] 黄嘉佑.气象统计分析与预报方法[M].北京:气象出版社,1990:170-197

[9] 陈兴芳,赵振国.中国汛期降水预测研究及应用[M].北京:气象出版社,2000:60-198

[10] 赵振国.中国夏季旱涝及环境场[M].北京:气象出版社,1999:150-290

[11] 吴有训,程雪生,胡安霞,等.安徽夏季雨型与亚欧500 hPa月平均高度场特征[J].安徽农业科学,2009,38(13):6042-6044,6081

[12] 吴有训,谷经山,张岭,等.典型相关分析方法在梅雨预报中的应用[J].应用数学,2003,16(增刊):40-43

[13] Weng H Y,Lau K M ,Xue Y K. Multi-scale summer rainfall variability over China and its long-term link to global sea surface temperature variability[J]. J Meteor Soc Japan,1999,77(4):845-857

[14] Hu Z Z. Interdecadal variability of summer climate over East A sia and its association with 500hPa heigh t and global sea surface temperature[J]. J Geophy Rea Atmo,1997,102(D16):19403-19412

[15] 魏凤英,黄嘉佑.大气环流降尺度因子在中国东部夏季降水预测中的作用[J].大气科学,2010,34(1):202-212

[16] 吴有训,程雪生,胡安霞.神经网络汛期降水短期气候预测模型[J].数学的实践与认识,2010,40(3):103-106

[17] Chih-Chung Chang, Chih-Jen Lin. LIBSVM-A Library for Support Vector Machines[EB/OL]. (2008-5-16). http://www.csie.ntu.edu.tw/~cjlin/libsvm/

(责任编辑:刘迎春)