

基于主成分分析的 SVM 分类汛期降水预测模型

吴有训¹, 刘 勇², 吴必文²

(1. 安徽省宣城市气象局, 安徽 宣城 242000; 2. 安徽省气象局, 安徽 合肥 230061)

摘要: 采用 1960—2008 年每年 5—9 月安徽宣城 7 县(市)平均降水量作为预报对象, 在降水前期逐月 74 项大气环流特征量、500hPa 月平均高度场和月平均海温场中选取预测因子, 采用主成分分析方法组合预测因子, 建立汛期降水短期气候预测模型。模型预测结果表明 2005—2008 年汛期降水量分类预测误差绝对值小于或等于 1 级, 平均误差绝对值 0.5 级, 训练集回预测准确率达 100%。该模型可在气候预测业务中使用。

关键词: 支持向量分类机(SVM 分类) 汛期降水 短期气候预测 主成分分析
中图分类号: P457.6 文献标识码: A 文章编号: 1006-764X(2011)S1-0030-02

宣城市界于北纬 29°57′~31°19′、东经 117°58′~119°40′之间, 位于长江下游南岸, 汛期降水时空分布不均, 年际差异大, 降水量多, 容易出现暴雨洪涝灾害。如 1999 年的特大洪水, 宣城市 90 个圩口溃破, 全市受灾耕地面积达 10.88 万 hm², 有 30.28 万人被洪水围困, 直接经济损失达 51.95 亿元。因此, 做好汛期降水短期气候预测、为防汛决策提供科学依据意义重大。目前短期气候预测业务技术多采用统计预报方法, 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; 而支持向量分类机(SVM 分类)方法建立在 Vapnik^[1]提出的统计学习理论和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, 应用支持向量分类机建立汛期降水预测方法是一种有效方法^[1-3]。

主成分分析能够把随时间变化的气象要素场分解为空间函数部分和时间函数部分, 主成分的前几个分量占原有变量总方差的很大部分, 因此可用主成分随时间的变化规律代替原始场随时间变化^[4]。用主成分分析方法浓缩预测信息, 减少计算量且降维去噪, 提高泛化性能。本研究用主成分分析方法组合预测因子, 并用 SVM 分类方法建立预测模型, 以期提高汛期降水预测能力^[5-6]。

1 支持向量分类机原理与算法^[7]

给定样本集 $T = \{ (x_1, y_1), \dots, (x_l, y_l) \} \in (x \times$

$y)^l$, 其中 $x_i \in x = R^n, y_i \in y = \{1, -1\}, i = 1, \dots, l$ 。
选取核函数 $K(x, x')$ 和惩罚因子 C , 构造并求解最优化问题:

$$\min \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x'_j) - \sum_{j=1}^l \alpha_j \quad (1)$$

$$\text{s.t.} \sum_{i=1}^l y_i \alpha_i = 0 \quad (2)$$

$$0 \leq \alpha_i \leq C \quad i = 1, \dots, l \quad (3)$$

得最优解 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$ 。选取 α^* 的一个正分量 $0 < \alpha_i^* < C$ 并据此计算阈值 $b_j^* = y_j - \sum_{i=1}^l y_i \alpha_i^* K(x_i, x'_j)$ 。构造决策函数

$$f(x) = \text{sgn} \left[\sum_{i=1}^l \alpha_i^* y_i K(x_i, x'_j) + b_j^* \right] \quad (4)$$

求解多类分类问题的实质就是找到一个把 R^n 上的点分成多部分的规则, 求得问题解。

2 支持向量分类机汛期降水气候预测建模

2.1 预测对象与预测因子选取及主成分分析

用 1960—2008 年每年 5—9 月宣城市 7 县(市)平均降水量作为预测对象。影响汛期降水的因子很多, 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^[8-9], 也包括一些非环流因子的间接影响、太平洋海表温度等^[10-11]。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(资料取自国家气候中心

气候诊断室 通过与预测对象的相关系数计算 选取 11 个预测因子 ($P_1 \sim P_{11}$),其中 $P_1 \sim P_5$ 选自 74 项大气环流特征量,为表征大气环流的因子,分别为 1 月西太平洋副高脊线 ($110^{\circ}\text{E} \sim 150^{\circ}\text{E}$) 1 月西藏高原位势高度 ($25^{\circ}\text{N} \sim 35^{\circ}\text{N}$, $80^{\circ}\text{E} \sim 100^{\circ}\text{E}$) 2 月太平洋副高脊线 ($110^{\circ}\text{E} \sim 115^{\circ}\text{W}$) 2 月北美区极涡强度指数 (3 区, $120^{\circ}\text{W} \sim 30^{\circ}\text{W}$) 2 月亚洲经向环流指数 (IM, $60^{\circ}\text{E} \sim 150^{\circ}\text{E}$); $P_6 \sim P_9$ 为 500 hPa 高度场因子; $P_{10} \sim P_{11}$ 为海温场因子 (500 hPa 高度场因子和海温因子均为不同地理位置格点上的数值)。宣城市 5—9 月降水量与前期预报因子 $P_1 \sim P_{11}$ 的相关系数分别为 0.45, 0.47, -0.36, 0.49, -0.42, 0.42, 0.44, 0.45, 0.34, 0.33, 0.35。

直接用这 11 个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,因此采用主成分分析方法浓缩预测因子。设预报因子为 m 个,样本容量为 n ,其观测值为 p_{ki} ($k=1, 2, 3, \dots, m; i=1, 2, 3, \dots, n$)。 m 个预报因子线性组合成一新因子 z ,如果 z 满足方差极大的要求,则称 z 为原 m 个变量的主成分,问题转化为组合系数向量 v 在满足条件 $v'v=1$ 的求极值问题,设矩阵 S 为 p 的 m 个变量协方差阵,因此问题转化为求矩阵 S 的 m 个特征值 λ_i 及其对应的特征向量 v 。计算得 11 个特征值 3.9813, 2.2948, 1.2692, 1.0081, 0.8001, 0.6895, 0.3999, 0.3137, 0.1475, 0.0616, 0.0343。前 8 个主成分占总方差的百分率为 98%,因此用这 8 个主成分作为输入。用 1960—2004 年资料作为训练样本,2005—2008 年资料作为测试样本。

2.2 支持向量分类机降水预测模型的设计

2.2.1 对降水量分类建立样本集

对 1960—2008 年每年 5—9 月宣城市 7 县(市)平均降水量的 49 个样本根据标准值 (1971—2000 年汛期宣城市 7 县(市)降水量的平均值)分类。表 1 为分类标准。

表 1 5—9 月宣城市 7 县(市)平均降水量分类标准					
降水量距平百分率/%	< -60	-60 ~ -20	-20 ~ 20	20 ~ 60	> 60
类别/级	1	2	3	4	5

2.2.2 确定核函数并建立预测模型

选取径向基函数作为核函数,径向基函数为 $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ $\gamma > 0$ (5) 采用交叉验证的方法^[10]优选径向基函数参数 γ 及惩罚因子 C 。对训练样本进行学习,求出阈值 b_j^* 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。构建支持向量分类机预测模型时采用了 Libsvm-2.89 软件^[12]。

3 结果分析

表 2 给出了 2005—2008 年预测模型的最优参数。用交叉验证方法选择惩罚因子 C 、径向基函数参数 γ 以及宣城市 5—9 月降水量预测模型的计算阈值 b_j^* 。

表 2 2005—2008 年预测模型的最优参数				
年份	γ	C	阈值 b^*	
2005	0.5	32	$\begin{bmatrix} 0.0095 \\ -0.6677 \\ -0.6553 \end{bmatrix}$	$\begin{bmatrix} -0.2107 \\ -0.2951 \\ -0.5285 \end{bmatrix}$
2006	0.0005	2048	$\begin{bmatrix} -6.0693 \\ -0.5108 \\ -0.0110 \end{bmatrix}$	$\begin{bmatrix} -3.0017 \\ -0.2155 \\ -0.0914 \end{bmatrix}$
2007	0.5	8	$\begin{bmatrix} -0.2844 \\ -0.6991 \\ -0.6553 \end{bmatrix}$	$\begin{bmatrix} -0.3824 \\ -0.2951 \\ -0.5285 \end{bmatrix}$
2008	0.5	8	$\begin{bmatrix} -0.3523 \\ -0.7003 \\ -0.6553 \end{bmatrix}$	$\begin{bmatrix} -0.3985 \\ -0.2951 \\ -0.5285 \end{bmatrix}$

表 3 给出了宣城市 5—9 月降水量预测结果。模型预测 2005—2008 年降水量类别时使用的训练集回预测准确率均达到 100%;对降水量偏少的 2005 年、2007 年预测准确,2006 年、2008 年的预测类别均比实况类别偏高 1 级,在实际应用中应注意预测模型的这一特性。表 3 表明预测模型运行稳定,预测准确率高。

表 3 宣城市 5—9 月降水量预测结果				
年份	训练集回预测准确率/%	预测类别/级	实况类别/级	误差级/级
2005	100	2	2	0
2006	100	3	2	1
2007	100	2	2	0
2008	100	4	3	1

4 结 语

利用主成分分析方法组合预测因子,提取主要信息,减少预测模型的计算量,提高泛化性能。连续 4 年的汛期降水量分类预测结果误差小于或等于 1 级,平均误差绝对值 0.5 级,训练集回预测准确率达 100%,模型可在气候预测业务中使用。由于用 SVM 分类方法建立预测模型,预测因子对预测能力的影响很大。广泛收集预测资料选取预测因子,将是下一步要做的工作。

参考文献:

[1] VAPNIK V. 统计学习理论的本质 [M]. 张学工,译. 北京:清华大学出版社,2000.
[2] 王景雷,吴景社,孙景生,等. 支持向量机在地下水位预报中的应用研究 [J]. 水利学报,2003,34(5):122-129.

(下转第 52 页)

Auspice_I3D 钢结构设计软件^[13]主界面如图 7 所示。本软件系统包括 20 个主模块,涵盖了钢结构工程中几乎所有的细节,并集成一体化处理技术,实现设计模型、分析模型、产品模型关联设计,建模采用数据图形技术,并根据用户配置自动生成工程图。



图 7 Auspic_I3D 主界面

4 结 语

通过分析目前三维软件二次开发中存在的问题和对软件二次开发层次的探讨,提出了钢结构三维设计软件二次开发的深层分析方法。该方法通过对软件文件系统的解析找出该软件系统的架构、属性、文件类型及格式,在掌握了数据组织及链接方式的基础上利用 API 提供的接口函数完成界面和交互开发,据此二次开发的软件利用了软件系统多层次的资源,使得开发的软件可以无缝集成到原软件中。该方法为三维设计软件的二次开发提供了一种可行的思路与方法。

参考文献：

[1] 徐国宾,尹伟波. ANSYS 二次开发在水工结构模糊参数反演中的应用[J]. 天津大学学报, 2008, 41(2): 215-220.
[2] 吴余华,童水光,许跃敏,等. 国外 CAD/CAM 系统二次开发关键技术的研究[J]. 计算机辅助设计与制造, 1996(8): 11-12.
[3] 王廷和,孟广耀. 基于 .Net 的 UG 二次开发技术研究[J]. 现代制造技术与装配, 2010(1): 67-68.
[4] 周桂生,陆文龙. CATIA 二次开发技术研究与应用[J]. 机械设计与制造, 2010(1): 81-83.
[5] 刘超,樊菊平,黄克戩,等. 基于 CATIA 的地下厂房三维设计研究[J]. 水电站设计, 2010, 26(1): 22-28.
[6] 邹振宇,刘晶. 基于 CATIA 的建筑构件设计二次开发技术[J]. 吉林建筑工程学院学报, 2010, 27(2): 13-16.
[7] 魏群,张国新,尉军耀,等. 拱坝三维可视化设计软件的开发与应用[J]. 天津大学学报, 2008, 41(9): 1087-1090.
[8] 魏群,张国新. 岩土工程图形计算力学的概念方法及应

用[J]. 岩土工程学报, 2008, 27(10): 2043-2050.
[9] WEIQun. CIS/2 steel structure standards and applications in CAD system[C]//Proceedings of the 2nd China International Steel Construction Congress. Beijing [s. n.]: 2005.
[10] 严秀,李龙澍. 软件逆向工程研究[J]. 计算机技术与发展, 2009, 19(4): 20-24.
[11] SILLS A, AHMED M. XML. NET developer's guide[M]. Beijing: Publishing House of Electronics Industry, 2003.
[12] Autodesk. Autodesk Inventor 2009 API Reference[R]. California: AutoDesk, 2008.
[13] 华北水利水电学院钢结构与工程研究院. Auspic_I3d 研发报告[R]. 郑州: 华北水利水电学院, 2009.

(收稿日期 2010-10-19 编辑 骆超)

+++++

(上接第 31 页)

[3] 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, 17(6): 97-100.
[4] 高尚. 基于主成分分析和支持向量机的作战飞机效能评估[J]. 航空计算技术, 2005, 35(3): 17-20.
[5] 陈永义,余小鼎,高学浩,等. 处理非线性分类和回归问题的一种新方法(I):支持向量机方法简介[J]. 应用气象学报, 2004, 15(3): 345-354.
[6] 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):支持向量机方法在天气预报中的应用[J]. 应用气象学报, 2004, 15(3): 355-365.
[7] 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.
[8] 吴有训,潘荣山,奚和平,等. 皖南冬半年云系和降水与东亚环流结构分析[J]. 安徽师范大学学报:自然科学版, 2007, 30(1): 68-73.
[9] 吴有训,胡雯,袁野,等. 江淮地区夏季环流分型及其云水气候特征[J]. 安徽师范大学学报:自然科学版, 2008, 31(3): 274-278, 287.
[10] 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, 1999, 77(4): 845-857.
[11] HU Z Z. Interdecadal variability of summer climate over East Asia and its association with 500 hPa height and global sea surface temperature[J]. J Geophy Rea Atmo, 1997, 102(16): 19403-19412.
[12] CHANG C C, LIN C J. LIBSVM-A library for support vector machines[EB/OL]. [2008-05-16]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

(收稿日期 2011-03-18 编辑 骆超)