

张礼平,陈正洪,成驰,等.支持向量机在太阳辐射预报中的应用[J].暴雨灾害,2010,29(4):334-336

支持向量机在太阳辐射预报中的应用

张礼平¹,陈正洪²,成驰²,王晓莉³

(1.武汉区域气候中心,武汉 430074;2.湖北省气象科技服务中心,武汉 430074;3.湖北省公众气象服务中心,武汉 430074)

摘要:利用 EOF 能分解数据场和 SVM 回归分析可建立因子与预报量非线性关系的优势,设计预报方案:(1)将多因子和多预报量分别方差标准化,EOF 场展开,提取主分量;(2)用 SVM 回归分析实现多因子主分量对多预报量主分量非线性预测;(3)由预报的多预报量主分量与对应空间函数反演原预报量。选用武汉预报日同一天气类型的上一日逐时(05—18 时)总辐射、日最高温度、温度日较差、日天气类型观测值以及预报日的日最高温度、温度日较差、日天气类型预报值为因子,对预报日逐时辐射量进行预报。独立预报试验表明,预报与实况接近。

关键词:支持向量机;回归分析;太阳辐射

中图分类号:P461+.1 **文献标识码:**A **文章编号:**1004-9045(2010)04-0334-04

1 引言

低碳的生产、生活方式,不仅是解决气候变化问题的根本出路,也将为我国在新一轮全球经济竞争中赢得主动。我国太阳能资源丰富,理论储量大,与同纬度国家相比,资源丰度与美国相近,比欧洲、日本优越得多,是未来最有希望的、可大规模开发利用的可再生能源。太阳能光伏发电被认为是转换效率最高、使用期长、可提供大量电力的一种太阳能利用方式^[1]。国外太阳能光伏发电已经完成了初期开发和示范,现在正向大批量生产和规模应用发展。

由于太阳能利用与太阳辐射密切相关,随着国内太阳能光伏装机容量的迅速扩大,为提高光电转换效率,降低运营成本,研究和开发太阳辐射预报技术显得十分迫切和必要。

支持向量机(Support Vector Machines,简称 SVM)方法是近年国际上开始流行的一种新颖的处理非线性分类和回归的有效方法。它以 V. N. Vapnik 等人提出的统计学习理论^[2-4]为基础,借助 Mercer 核展开定理和近代最优化方法的结果,将样本空间映射到一个更高维以至于无穷维的特征空间,在特征空间中把寻求最优回归超平面问题归结为一个凸约束条件下的二次凸规划问题,从而求得全局最优解。与特征空间中得到的线性解相对应的是样本空间中非线性解。一般升维变换会带来算法的复杂化,但由于核函数的引入,不但没有增加算法的复杂性,

而且在某种意义上避免了“维数灾”。文献[5]分析了 SVM 方法的特点及其在气象业务中的可能应用前景。目前,SVM 方法在太阳辐射预报中的应用较少。本文引进基于 SVM 和 EOF(自然正交分解)的预报方法^[6],设计一种多因子对多预报量非线性预报方案,以实现逐日逐时辐射量预报。

2 基本原理

回归分析又称函数估计。设给定的样本数据集为 $(X_1, y_1), (X_2, y_2), \dots, (X_l, y_l)$

其中 X_i 为预报因子值(N 维向量), y_i 为预报量值, $i=1, 2, \dots, l, l$ 为样本总量。回归分析就是基于样本数据集寻求一个反映预报因子与预报量的最优函数关系 $\hat{y}=f(X)$ 。由于线性函数表述形式最简单,18 世纪 Gauss 提出的回归分析就是在最小二乘法意义下确定线性函数系数 W (N 维向量)和 b ,使 $f(X)=(W \cdot X)+b$ 与实测 y 偏差平方和为最小。当 $\hat{y}=f(X)$ 为线性函数时,通常称 $f(X)$ 为最优回归超平面。

Vapnik 提出一种 ε 不敏感误差函数:

$$L_\varepsilon(y) = \begin{cases} 0 & |f(X) - y| \leq \varepsilon \\ |f(X) - y| - \varepsilon & |f(X) - y| > \varepsilon \end{cases} \quad (1)$$

这里 ε 为非负数。其含义为:当误差小于(或等于) ε 时,认为误差为零忽略不计;误差大于 ε 时,定义误差值为实际误差减去 ε 。这种误差函数给出了一个宽度为 2ε 的不敏感带,称为 ε 管道。若所有样本点均

收稿日期:2010-09-10;定稿日期:2010-11-18

资助项目:科技部公益性行业(气象)科研专项(GYHY201006036)、华中区域气象中心推广项目(QY-T-200902)、华中区域气象中心重点项目(QY-Z-201010)、中国气象局工作任务(气预函[2010]76号)

作者简介:张礼平,男,1956年生,正研级高工,主要从事短期气候预测及方法研究。E-mail: zhangliping_wh@yahoo.com.cn

在 ε 管道中,则总误差为 0;否则如下引入松弛变量 $\xi_i \geq 0$:

$$\begin{cases} \xi_i = |y_i - (W \cdot X_i) - b| - \varepsilon & \text{当 } |y_i - (W \cdot X_i) - b| > \varepsilon \\ \xi_i = 0 & \text{当 } |y_i - (W \cdot X_i) - b| \leq \varepsilon \end{cases}$$

则总误差为所有 ξ_i 的和。图 1 给出了管道和松弛变量的直观图示,这里最优回归超平面为一直线。

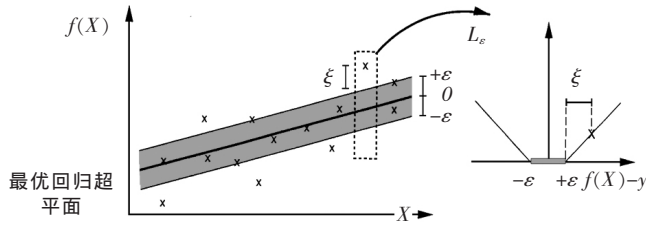


图 1 ε 管道和松弛变量(X 为 1 维向量)示意图

基于 ε 不敏感误差函数,寻求最优回归超平面问题可以归结为如下的凸约束条件下的二次凸规划问题:

$$\min \left\{ \frac{1}{2} \|W\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \right\} \quad (2)$$

$$\text{约束条件: } \begin{cases} y_i - (W \cdot X_i) - b \leq \varepsilon + \xi_i \\ (W \cdot X_i) + b - y_i \leq \varepsilon + \xi_i^* & (i=1, 2, \dots, l) \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

其中 ξ_i 与 ξ_i^* 分别对应于最优回归超平面上方和下方的样本点。 C 为事先给定的惩罚系数,一般由试验确定。

定义相关于(2)Lagrange 函数,对它关于 W, b, ξ_i, ξ_i^* 求偏导数并令偏导数为零,整理后(2)可以转化为如下等价的对偶规划问题:

$$\max \left\{ -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) (X_i \cdot X_j) - \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \right\} \quad (3)$$

$$\text{约束条件: } \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 & (i=1, \dots, l) \\ 0 \leq \alpha_i, \alpha_i^* \leq C \end{cases}$$

α_i, α_i^* 分别为对应 ξ_i 与 ξ_i^* 的 Lagrange 乘子。最后求得最优回归超平面的表达式为:

$$f(X) = (W \cdot X) + b = \sum_{\text{支持向量}} (\alpha_i - \alpha_i^*) (X \cdot X_i) + b \quad (4)$$

$\alpha_i, \alpha_i^* \neq 0$ 对应的 X_i 称为支持向量,非支持向量的 X_i 对应的 $\alpha_i, \alpha_i^* = 0$ 。实际上,最优回归超平面由支持向量对应的样本点完全确定。SVM 通过引入核函数而巧

妙地绕过非线性映射的显式表达,最后得^[2,3,4]:

$$\begin{aligned} f(\varphi(X)) &= (W \cdot \varphi(X)) + b = \sum_{\text{支持向量}} (\alpha_i - \alpha_i^*) (\varphi(X) \cdot \varphi(X_i)) + b \\ &= \sum_{\text{支持向量}} (\alpha_i - \alpha_i^*) K(X, X_i) + b \end{aligned} \quad (5)$$

这就是 SVM 回归得到的非线性回归函数。Mercer 核函数很多,常见的有多项式核、高斯核、拉普拉斯核等。实际计算时, $K(X, X_i)$ 的具体函数形式,通常可由试验确定。尽管是在高维特征空间中解决问题,由于借助了核函数,在实际求解过程中根本不必知道该非线性映射 φ 的显式表达式。特别对于高维数据,由于核函数与向量的维数无关,可避免通常所说的“维数灾”,极大地简化了数值计算,为其业务应用提供了可能。

3 实例

3.1 预报方案设计

为消除量纲不同的影响,分别对每一因子和每一预报量方差标准化,使每一变量的方差和平均值均为 1,0。

由于多因子或多预报量可视为数据场,EOF 能将数据场分解为不随时间变化的空间函数(特征向量)及只依赖时间变化的主分量,利用方差集中在前 N 个主要分量的特征,用前 N 空间函数和主分量的线性组合构成对原场的估计,略去原场中方差较小分量,保留较大分量,正交变换不改变场总方差,因此估计场保留了原场大部方差,反映了原场的主要特征。用 EOF 方法,分别提取多因子和多预报量主分量。考虑到多因子和多预报量主分量可能存在的非线性关系,用 SVM 回归分析实现多因子主分量对多预报量主分量的预测。最后由预测的多预报量主分量与其对应空间函数线性组合还原为预报量的预报。具体步骤如下:

- (1) 将每一因子和每一预报量分别方差标准化;
- (2) 分别对多因子和多预报量进行 EOF 展开,提取主分量;
- (3) 选用不同的核函数,由试验确定较合适的核函数和相关参数,进行 SVM 回归分析,由多因子主分量预测多预报量主分量;
- (4) 由预报的多预报量主分量与对应空间函数线性组合还原为预报量。

3.2 预报实例

为避免不同季节的影响,且考虑到每月气候背景类似,将 1—12 月逐月建立预报模型。预报因子选用武汉站(区站号 57494,位置 30.62°N、114.13°E)2007 年、2008 年、2009 年 3 年逐日与被预报日同一天气类型的上一日逐时(5—18 时)总辐射、同一天气类型的

上一日和预报日的日最高温度、温度日较差、日天气类型数值共 20 个因子(上一日资料均为观测值,预报日数据为气象台预报值), 预报武汉同一天气类型下一天逐时(5—18 时)总辐射(预报量 14 个)。

天气类型是指某日云量、降水概况,日天气类型共分三类:1 类天气类型定义:总云量 ≥ 9 ,低云量 ≥ 3 ,降水量 > 0 ;2 类天气类型定义: $3 <$ 总云量 < 9 ,降水量 $= 0$;3 类天气类型定义:总云量 ≤ 3 ,低云量 ≤ 1 ,降水量 $= 0$ 。这里给出用 2007 年 7 月 1—31 日、2008 年 7 月 1—31 日、2009 年 7 月 1—26 日逐日数据建立模型,预报 2009 年 7 月 27 日逐时总辐射,……,用 2007 年 7 月 1—31 日、2008 年 7 月 1—31 日、2009 年 7 月 1—30 日逐日数据建立模型,预报 2009 年 7 月 31 日逐时总辐射实例。

试验完全模拟实际预报(以预报 2009 年 7 月 27 日辐射为例),其步骤如下:

(1)将 2007 年 7 月 1—31 日、2008 年 7 月 1—31 日、2009 年 7 月 1—27 日逐日多因子数据(样本容量 $31+31+26$)方差标准化,进行 EOF 分析,其中前 88($31+31+26$)逐日数据作为分析训练样本,最后 1 个样本为预报预留。截取前 4 个主分量,累积方差贡献率 75%;

(2)将 2007 年 7 月 1—31 日、2008 年 7 月 1—31 日、2009 年 7 月 1—26 日逐日数据(样本容量 $31+31+26$)多预报量方差标准化,进行 EOF 分析,截取前 3 个主分量,累积方差贡献率 81%;

(3)用 88 个样本多因子前 4 个主分量和多预报量第 1 主分量建立 SVM 非线性回归函数关系式,和多预报量第 2 主分量建立 SVM 非线性回归函数关系式,和多预报量第 3 主分量建立 SVM 非线性回归函数关系式,由多因子主分量预测多预报量主分量,用多因子前 4 个主分量第 89 样本独立预报多预报量第 1、2、3 主分量;

(4)多预报量主分量预报值与对应空间函数线性组合构成 2009 年 7 月 27 日总辐射预报。

与上步骤相同,依次独立制做出 2009 年 7 月 28—31 日逐日逐时总辐射预报。图 2 给出 7 月 27—31 日逐时总辐射预报值与实况的对比情况。由图可见,预报与实况基本接近,特别是 5 d 中,27—28 日总辐射变小,28—30 日总辐射逐日变大、31 日又变小,30 日呈峰值均成功预报。

4 结论与讨论

利用 EOF 能将数据场分解为不随时间变化的空间函数和只依赖时间变化的主分量的特性,以及 SVM 回归分析可建立因子与预报量非线性关系的优势,设

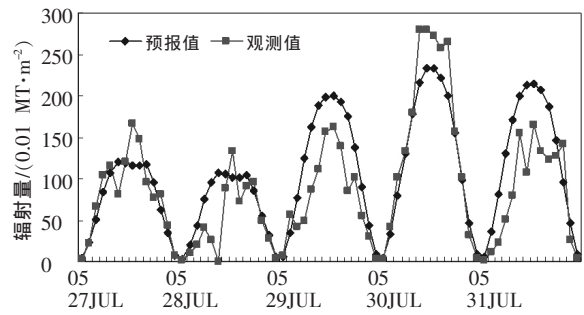


图 2 2009 年 7 月 27—31 日逐日逐时辐射量独立预报结果与实况对比图

计多因子对多预报量非线性预报方案,以实现逐时辐射量预报。对武汉 2009 年 7 月 27—31 日逐日逐时辐射量进行了独立预报试验。试验结果表明,预报与实况接近,5 天中辐射量的两次高低起伏变化的预报均与实况一致。

降维(即将高维样本空间向低维空间投影)是处理复杂问题的传统简化方法,一般认为低维空间数据结构以及内部关系容易研究和认识。与传统思维正好相反,SVM 升维,即将样本空间向更高维空间投影,巧妙地运用 Mercer 核展开定理,通过非线性映射 ϕ ,将原样本空间的非线性关系变为高维特征空间的线性关系,高维特征空间的线性关系也就表述了原样本空间的非线性关系。而在实际求解过程中根本不必知道非线性映射 ϕ 的显式表达式,极大地简化了数值计算。SVM 为我们解决辐射量预报中非线性问题提供了一个新途径。

参考文献:

- [1] 张庆阳. 国外太阳能的开发利用及其借鉴 [J]. 气象科技合作动态, 2009(5):28-32.
- [2] Vapnik V N. Statistical Learning Theory[M]. New York: John Wiley & Sons,Inc. 1998:375-570.
- [3] Vapnik V N. The nature of statistical learning theory [M]. New York:Springer Verlag. 2000: 123-266.
- [4] Courant R, Hilbert D. Method of Mathematical Physics [M]. New York: Springer Verlag.1953:96-110.
- [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 ()——支持向量机方法简介[J]. 应用气象学报,2004,15(3): 345-354.
- [6] 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.

(下转第 355 页)

- 的应用[J].气象,2005,31(10):56-61.
- [19] 刘玉玲. 对流参数在强对流天气潜势预测中的作用[J]. 气象科技, 2003,31(3):147-151.
- [20] 谷秀杰,牛淑贞,介玉娥,等.2007年8月2日郑州大暴雨过程分析 [J].气象与环境科学,2010,33(2):53-58.
- [21] 刘建文,郭虎,李耀东,等.天气分析预报物理量计算基础[M].北京:气象出版社,2005.
- [22] 陆汉城,杨国祥.中尺度天气原理和预报[M].北京:气象出版社,2004.

The Rainstorm Potential Predictability for Rain Area Using the Mesoscale Numerical Model

WANG Jue^{1,2}, SHOU Shao-wen¹, ZHANG Jia-guo², MAO Yi-wei²

(1.Nanjing University of Information Science & Technology,Nanjing 210044;
2.Wuhan Central Weather Office,Wuhan 430074)

Abstract: By studying the relation between the physical quantity output from AREM and rainstorm from 2005 to 2006 and using fuzzy logic method, a potential predictability combination equation with multiple parameters was setup and applied to forecast rainstorm with the average and maximum values of 24hr AREM forecast field. Through contractive analysis of the potential predictability and rainstorm Ts scores from 2007 to 2009, the conclusion is that more strong centers and the false alarms were forecasted by the potential predictability with maximum value. The potential predictability with average value can reflect the whole precipitation condition in 24 hours and its Ts scores are higher, and its whole effect is better though it has deviation to forecast the central rain area of severe precipitation.

Key words: Rainstorm; Mesoscale Model; Potential predictability

(上接第 336 页)

Applications of Support Vector Machines in the Solar Radiation Forecasting

ZHANG Li-ping¹, CHENG Zheng-hong², Cheng Chi², WANG Xiao-li³

(1.Wuhan Central Meteorological Observatory,Wuhan 430074;
2.Hubei Service Center of Meteorological Science and Technology,Wuhan 430074;
3.Public meteorological service center of Hubei, Wuhan 430074)

Abstract: With the superiority of both of EOF (Empirical orthogonal functions) separating fields and nonlinear SVM regression forecasting, a program is projected considering the following: (1) Both of factors and predictors are variance standardized, then EOF, and principal components of both are extracted respectively; (2) With SVM regression, principal components of predictors are estimated by factors; (3) The original predictors are recovered by linear combination of principal components and the eigenvectors.

Solar radiation over Wuhan is forecasted hour by hour with Solar radiation of previous day, the maximum temperature, temperature range and weather type of both previous day and forecasting day. It is indicated that Solar radiation forecasted hour by hour were approximate to the observations.

Key words: Support vector machines; Regression; Solar radiation