基于最小二乘支持向量机的中国粮食产量预测模型研究

李晓东,席升阳,潘 立
(河南科技大学,河南 洛阳 471003)

摘 要:粮食产量预测是制定农业政策的重要依据。针对农业生产系统的特征,在统计学习理论和结构风险最小化原理的基础上,建立了基于最小二乘支持向量机的时间预测模型。预测结果表明该模型具有较高的预测精度,为粮食产量预测提供了一条新的途径。

关键词:粮食预测;支持向量机;最小二乘支持向量机
中图分类号:TP18;S114 **文献标识码**:A **文章编号**:1005-3409(2007)06-0322-03

Forecast of China Grain Production Based on Least Squares Support Vector Machine

LI Xiao-dong, XI Sheng-yang, PAN Li
(He'nan University of Science and Technology, Luoyang, Luoyang, He'nan 471003, China)

Abstract: The forecast of grain production is an important resource for establishing agriculture policy. Aimed at the character of the agriculture system, the least squares support vector machine prediction model is given based on the principle of the statistical learning theory and structural risk minimization. The result is given that the forecasting model is effective and offers a new method to forecast the grain production.

Key words: forecast of grain production; support vector machine; least squares support vector machine

农业是国民经济的基础,农业的健康发展也是国民经济健康发展的根本保证,而农业生产的粮食预测是政府制定和实施农业经济政策的重要依据。做好粮食预测有利于农业经济资源的优化配置,有利于农业结构的合理调整,有利于农业经济的健康发展。目前国际上流行的粮食预测方法有气象产量预测法、遥感技术预测法、统计动力学生长模拟法等。笔者从时间序列的角度,研究我国粮食产量的变化趋势并进行预测。时间序列预测是预测技术体系的重要部分,其基本思想是,时间序列的数据是系统运动规律的外在特征,隐含系统发展变化重要信息,因此可以建立相应的数学模型去反映系统随着时间变化的规律,从系统的表面特征中挖掘出系统演化的信息,进而对系统的未来行为特征进行预测。传统的时间序列预测方法如 ARMA 模型仅反映了系统发展变化的线性特征,而实际的系统发展大多呈现非线性,因此 ARMA 模型具有失真性;最近提出的神经网络预测模型是一种非线形预测方法,具有高度的并行处理和容错能力,但是神经网络有一些难以克服的缺陷,例如神经网络预测模型过分强调对数据的拟合,泛化能力不高,从而预测能力不高。支持向量机是目前提出的新型机器学习方法,由于具有出色的学习性能和预测性能,并且克服了神经网络的过学习和局部极小值问题,在时间序列预测得到广泛的应用。本文根据我国粮食产量的历史数据,建立了基于最小二乘支持向量机的时间序列预测模型,预测结果表明该模型具有较高的预测精度和动态适应性。

1 最小二乘支持向量机回归[1-3]

机器学习理论主要研究从观测数据寻找未知规律,并利用这些规律对未来数据进行预测。现有机器学习方法重要理论基础是统计学,传统统计学研究的是样本数目趋于无穷大的渐进理论,但在实际问题中,样本数目往往是有限的,因此一些理论上很优秀的学习方法在实际中表现却不尽人意。统计学习理论是一种专门研究小样本情况下的机器学习规律的理论,Vapnic 等从 20 世纪六、七十年代开始致力于此方面的研究,到 90 年代中期,形成一个较为完善的理论体系即统计学习理论(Statistical Learning Theory,SLT),统计学习理论被认为是目前针对小样本估计和预测学习的最佳理论。支持向量机理论(Support Vector Machine,SVM)是基于统计学习理论的 VC 维理论和结构风险最小化基础上的新型机器学习方法,从而在统计样本量较少的情况下,也能表现出色的学习性能,并且克服了神经网络出现的泛化能力低和过学习等问题,被认为是神经网络的替代方法。而最小二乘支持向量机(LS-SVM)回归模型是标准支持向量机的改进和扩展,它是支持向量机在二次损失函数下的一种形式,通过构造损失函数将原支持向量机算法的二次优化问题转化为求解线性方程,其求解速度快,因此在各个领域得到广泛的应用和发展。

给定 l 个独立的样本

$$(x_1, y_1), \cdots, (x_1, y_1), x \in R^N, y \in R$$

在函数集 $S = \{f(x, w), w \in \Omega\}$ 中寻求一个最优函数 $f(x, w_0)$ 对 y 与 x 之间的函数关系进行估计,如果所得函数关系是线性函数,则称为线性回归,可以表示为 $f(x) = w^T x + b$ 否则称为非线性回归。对于非线性回归问题,LS-SVM 的基本思想是首先使用一个非线性映射 $\varphi(x)$ 将数据映射到一个高维特

收稿日期:2007-01-03
作者简介:李晓东(1982 -),男,河南驻马店人,硕士生,主要从事控制论研究。

征空间,然后在高维特征空间进行线性回归,最后映射到原空间就完成了输入空间的线性回归。可以表示为 $f(x) = w^T\varphi(x) + b$

并使期望风险最小

$$R(f) = \int [y - f(x,w)]\,p(x,y)\,\mathrm{d}x\mathrm{d}y \tag{1}$$

式中:$L[y - f(x,w)]$ ——用 $f(x,w)$ 对 y 进行预测所造成的损失;$p(x,y)$ ——联合概率密度。由于 $p(x,y)$ 未知,所以期望风险并不可求。传统的学习方法采用了经验风险最小化(ERM)准则,即定义经验风险:

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l} L[y_i, f(x,w)] \tag{2}$$

作为对经验期望风险的估计,并设计学习方法使它最小化。事实上,经验风险最小化代替期望风险最小化并没有可靠的理论依据,只是直观上合理的做法。因此在实际应用中,一般难以取得理想的结果。例如神经网络算法采用经验风险最小化原则,过分地强调训练误差最小并不能得到良好的泛化能力,有时训练误差过小反而导致泛化能力的下降。LS-SVM 是一种以结构风险最小化(Structural Risk Minimization,SRM)为基础的算法,改变了传统的经验风险最小化原则,从而在统计样本量较少的情况下,也能表现出色的学习性能。根据结构风险最小化原理,引入结构风险函数

$$R_{reg}[f] = \frac{1}{2}\|w\|^2 + CR_{emp}[f] \tag{3}$$

式中:$\|w\|^2$ ——控制模型的复杂度参数;$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l} L[y_i, f(x,w)]$ ——经验风险;C ——可调参数,它能够在经验风险和模型复杂度之间进行调节以便使所求的函数具有较好的泛化能力。定义经验风险为

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l}[y_i - f(x_i)]^2 \tag{4}$$

则 LS-SVM 的优化目标函数可以表示为

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{2}r\sum_{i=1}^{l}\xi_i^2 \tag{5}$$

约束条件为

$$y_i = w^T\varphi(x_i) + b + \xi_i,\ i = 1, \cdots, l \tag{6}$$

定义拉格郎日函数为

$$L(w,b,\xi,\alpha) = \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i[w^T\varphi(x_i) + b + \xi_i - y_i] \tag{7}$$

对各参数求偏导数并令其等于零得

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l}\alpha_i\varphi(x_i)\\
&\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l}\alpha_i = 0\\
&\frac{\partial L}{\partial w} = 0 \Rightarrow \alpha_i = \gamma_i\xi_i,\ i = 1, \cdots, l\\
&\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow w^T\varphi(x_i) + b + \xi_i - y_I = 0,\ i = 1, \cdots, l
\end{aligned} \tag{8}$$

最后求解优化问题可转化为求解一线性方程组

$$\begin{bmatrix} 0 & (1_l)^T \\ 1_l & \Phi\gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{9}$$

式中:I ——单位矩阵;Φ ——一矩阵,其定义 $\Phi = (\Phi_{ij})_{l\times l}$,其中 $\Phi_{ij} = \varphi(x_i)^T\varphi(x_j)_I$

$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)^T$;$1_l = (1, 1, \cdots, 1)^T$;

$y = (y_1, y_2, \cdots, y_l)^T$

根据上面推导得出最优解的表达式为

$$f(x) = \sum_{i=1}^{l}\alpha_i K(x_i, x) + b \tag{10}$$

式中:$K(x_i, x_j)$ ——核函数,核函数 $K(x_i, x_j)$ 的值等于 2 个向量 x_i, x_j 在其特征空间 $\varphi(x_i)$ 和 $\varphi(x_j)$ 的内积,即 $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$,因此在实际计算过程中,不必考虑映射函数 $\varphi(x)$ 的具体形式,只需要选择合适的核函数就可以完成其特征空间的内积运算。根据泛函分析理论,只要满足 Mercer 定理条件的函数都可以作为核函数。

目前最常见的核函数有

多项式核函数　　$K(x, x_i) = (1 + x^T x)^d$

高斯径向基核函数　　$K(x, x_i) = \exp(-\|x - x_i\|/\sigma^2)$

神经网络核函数　　$K(x, x_i) = \tanh[k_1(x^T x_i) + k_2]$

2　粮食产量预测模型的建立

LS-SVM 具有较高精度的函数逼近能力,因此可以选择一个合适 LS-SVM 网络拓扑结构向有限个样本的学习来挖掘时间序列内部的运行规律,逼近时间序列样本所隐含的函数关系,可完成对新时间序列的映射关系,从而达到预报的目的。SVM 预测数学关系表达式为

$$Z_{t+1} = F(Z_t, Z_1, \cdots, Z_{t-p}) + \varepsilon_t \tag{11}$$

式中:$Z_t, Z_{t-1}, \cdots, Z_{t-p}$ ——历史时间序列数据;ε_t ——t 时刻的扰动项;Z_{t+1} ——$t+1$ 时刻的预测目标数值。具体的计算步骤如下:

步骤 1　为了提高运算速度和预测精度,对样本进行归一化处理

$$x_i = \frac{z_i - z_{\min}}{z_{\max} - z_{\min}} \tag{12}$$

式中:z_i ——历史数据序列。

步骤 2　确定 LS-SVM 输入向量的个数,建立输入向量 $x_n = [x_{n-1}, x_{n-2}, \cdots, x_{n-m}]$ 到输出向量的 $y_n = [x_n]$ 的映射关系,$R^m \to R$,m 为嵌入维数,可以表示为式(13),本文采取最终误差预报准则确定嵌入维数。

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n-m} & x_{n-m+1} & \cdots & x_{n-1} \end{bmatrix},\ Y = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \cdots \\ x_n \end{bmatrix} \tag{13}$$

步骤 3　选择高斯径向基核函数作为函数。

步骤 4　运用 LS-SVM 对样本进行学习训练,其回归函数可以表示为

$$f(W_j) = \sum_{i=1}^{n-m}\alpha_i K(Z_i, W_j) + b \tag{14}$$

式中:m ——嵌入维数;Z_i, W_j ——矩阵 X 的行向量,$f(W)$ 输出向量与矩阵 Y 相对应。

步骤 5　经过训练后 LS-SVM 就可以对输入向量进行预报,则一步预报方程为

$$\hat{x}(n+1) = f(W) \tag{15}$$

式中:$W = x(n), x(n-1), \cdots, x(n-m+1)$,同理可以得出 l 步长的 LS-SVM 预测模型

$$\begin{cases} \hat{x}(n+l) = f\{\hat{x}(n+l-1), \hat{x}(n+l-2), \cdots, x(n+l-m)\} & l < m \\ \hat{x}(n+l) = f\{\hat{x}(n+l-1), \hat{x}(n+l-2), \cdots, x(n+l-m)\} & l \geq m \end{cases} \quad (16)$$

3 仿真结果

根据我国粮食产量的历史数据[11],如表 1 所示,建立 LS-SVM 模型,经过计算嵌入维数为 3 时最为合理。分别对 2003 - 2005 年的粮食产量进行预测。应用 MATLAB 7.0 编辑程序,并对数据进行计算分析。

表 1 我国粮食产量历年统计表 万 t

年份	1985	1986	1987	1988	1989	1990	1991	1992	1993
产量	37911	39151	40298	39408	40755	44024	43529	44266	45649
年份	1994	1995	1996	1997	1998	1999	2000	2001	2002
产量	44510	46662	50454	49417	51230	50839	46218	45264	45706

仿真结果如表 2 所示

表 2 模型预测值与实际值对比 万 t

年份	原始数据	预测数据	相对误差/ %
2003	43070	43791	1.67
2004	46947	45779	- 2.49
2005	48401	48787	0.80

由表 2 可以看出,模型完全满足预测精度要求。通过以上的实际分析,LS-SVM 模型可以作为我国粮食产量的预测模型。

4 结 论

由于农业生产系统的非线性和不确定性特征,很难在粮食产量与影响因素之间建立确定的数学模型,但是粮食产量的时间数据序列是各种影响因素的综合结果,其中隐含了系统发展的重要信息,因此可以从时间序列特征角度研究其系统的变化规律并对其状态未来特征进行预测。本文应用统计学习理论原理,建立了基于 LS-SVM 的粮食产量预测模型,仿真结果表明,该模型比较全面地反映了系统的变化特征,并对系统的未来状态特征具有较高的预测精度,可以作为我国粮食产量预测的有效工具。

参考文献:

[1] 吴今培,孙德山. 现代数根据分析[M]. 北京:机械工业出版社,2006.

[2] Vapnic V. 统计学习理论的本质[M]. 北京:清华大学出版社,2000.

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32 - 42.

[4] 李栋,王洪礼,杜忠晓,等. 城市生活用水量的支持向量机回归预测[J]. 天津大学学报:社会科学版,2006,8(1):64 - 67.

[5] 卢虎,李彦,宵颖. 支持向量机理论及其应用[J]. 空军工程大学学报:自然科学版,2003,4(4):89 - 91.

[6] 罗雪晖,李霞,张基宏. 支持向量机及其应用研究[J]. 深圳大学学报,2003,20(3):40 - 44.

[7] 杨一文,杨朝军. 基于支持向量机的时间序列预测模型[J]. 系统工程理论方法应用,2005,14(2):176 - 181.

[8] 李智才,马文瑞,李素敏,等. 支持向量机在短期气候预测中的应用[J]. 气象,2006,32(5):57 - 61.

[9] 支持向量机的学习方法的选择与应用[J]. 武汉科技大学学报:自然科学版,2006,29(1):76 - 78.

[10] 中华人民共和国统计年鉴[Z]. 2005.

(上接第 321 页)

[12] 朱守谦,魏鲁明,张丛贵,等. 茂兰喀斯特森林树种生长特点初步研究[C]// 朱守谦. 喀斯特森林生态研究(Ⅱ). 贵阳:贵州科技出版社,1997.

[13] 周政贤,毛志忠,喻理飞,等. 贵州石漠化退化土地及植被恢复模式[J]. 贵州科学,2002,20(1):1 - 6.

[14] 任海. 喀斯特山地生态系统石漠化过程及其恢复研究综述[J]. 热带地理,2005,25(3):195 - 200.

[15] 李阳兵,谢德体,魏朝富. 岩溶生态系统土壤及表生植被某些特性变异与石漠化的相关性[J]. 土壤学报,2004,41(2):196 - 202.

[16] 吴秀芹,蔡运龙,蒙吉军. 喀斯特山区土壤侵蚀与土地利用关系研究:以贵州省关岭县石板桥流域为例[J]. 水土保持研究,2005,12(4):46 - 48.

[17] 姜光辉,郭芳,袁道先. 岩溶石漠化地区土地资源及其开发潜力:以云南木美地下河为例[J]. 水土保持研究,2005,12(5):214 - 217.

[18] 曾土荣. 粤北岩溶石山地区石漠化现状及其对水环境的影响[J]. 水文地质工程地质,2006(3):101 - 105.

[19] 广东省植物研究所编著. 广东植被[M]. 北京:科学出版社,1976.

[20] 宋永昌. 植被生态学[M]. 上海:华东师范大学出版社,2001:47 - 51.

[21] 马克平,黄建辉,于胜利,等. 北京东灵山地区植物群落多样性的研究 Ⅱ. 丰富度、均匀度和生物多样性指数[J]. 生态学报,1995,15(2):268 - 277.

[22] 魏兴琥,杨萍,李森,等. 西藏沙漠化典型分布区沙漠化过程中的生物生产力和物种多样性变化[J]. 中国沙漠,2005,25(5):663 - 667.

[23] 梅再美,王代懿,熊康宁,等. 不同强度等级石漠化土地植被恢复技术初步研究:以贵州花江试验示范区查尔岩试验小区为例[J]. 中国岩溶,2004,23(3):253 - 258.

[24] Toshiya Ohkuro, et al. Vegetation distribution in Naiman, Inner Mongolia, China [C]// Shizuo Shindo and Akihiko Kondoh. The CERes International Symposium on the Role of Remote Sensing for the Environmental Issues in Arid and Semi-Arid Regions. Chiba University, 1997:107 - 110.