基于最小二乘支持向量机的万元工业GDP取水量非线性组合预测

潘国强
(河南省水利科学研究院, 河南 郑州 450003)

摘 要: 根据支持向量机(SVM)和组合预测理论,选择趋势法预测万元工业GDP取水量的3种主要模型,提出基于最小二乘向量机(LS-SVM)的万元工业GDP取水量非线性组合预测方模型。实例表明:与单项预测模型和线性组合预测相比,基于LS-SVM非线性组合预测模型具有更强的泛化能力,能够有效提高区域万元工业GDP取水量预测精度。

关键词: 工业取水量; 最小二乘支持向量机; 非线性组合预测; 定额
中图分类号:TV213.9 **文献标识码**: A **文章编号**: 1672-643X(2013)05-0161-04

Nonlinear combination forecast of water demand quota for ten thousand yuan industry GDP based on LS-SVM

PAN Guoqiang
(*Henan Provincial Water Conservancy Research Institute*, *Zhengzhou* 450003, *China*)

Abstract: Based on the structure risk minimum criterion of support vector machine(SVM) and the nonlinear combination forecast theory, selecting three computation models of trends prediction, the square LS-SVM is applied to establish the combination forecast model for predicting the water demand quota for ten thousand yuan industry GDP. The application example shows that the model has stronger generalization capacity and adaptability than the model of single item and linear combination forecast, and can effectively enhance forecast precision.

Key words: industry water consumption; LS-SVM; nonlinear combination forecast; quota

组合预测理论由Bates和Granger于20世纪60年代提出,它证明了两种无偏的单项预测的组合方法要优于各单项的预测方法。组合预测可以克服单一模型的局限性,有效地综合更多的有效信息,降低预测风险,提高预测精度,在我国社会经济各领域得到广泛的运用。组合预测分为线性和非线性组合预测,其中线性组合预测模型构建相对容易[1],但线性组合预测实际上是不同预测方法之间的值的一种凸组合,当预测对象的实际值曲线位于两种不同方法的预测曲线的上方、下方或相交时,线性组合预测往往就无能为力,此时就需要考虑采用非线性组合预测方法[2]。非线性组合预测尤其适用于信息不完备的复杂经济系统。

在对万元工业GDP取水量进行线性组合预测研究的基础上,笔者基于机器学习的支持向量机(SVM)理论[3],建立基于LS-SVM的万元工业GDP取水量非线性组合预测模型,通过对比分析,证明了其有效性和和更高的预测精度。

1 非线性组合预测

非线性组合预测是将多个单个的预测模型,通过某种非线性函数进行组合得到预测值,公式为:

$$y_t = \phi(\varphi_{1t}, \varphi_{2t}, \cdots, \varphi_{mt}) \quad (1)$$

式中: y_t 为 t 时刻的组合预测结果; φ_{it} 为 t 时刻的方法 i 预测结果; $\phi(x)$ 为非线性函数。

在某种测度下,$\phi(x)$ 的度量要比 $\varphi_i(i = 1,2,\cdots,m)$ 优越。因此非线性组合预测模型构建的核心是构造有效的非线性函数。支持向量机在解决非

收稿日期:2013-05-06; **修回日期**:2013-05-23
基金项目:河南省节水型社会建设"十二五"规划项目
作者简介:潘国强(1969-),男,河北清河县人,高级工程师,主要研究方向为水资源高效利用与调配。

线性问题中具有小样本学习、全局寻优、泛化能力强、解决高维问题等良好特性,在机器学习领域得到广泛应用。最小二乘支持向量机是支持向量机的标准扩展,求解相对简单。本文基于最小二乘向量机提出万元工业 GDP 取水量非线性组合预测的方法。

2 支持向量机与最小二乘支持向量机

支持向量机(Support vector machines,SVM) 由 Vapnik 等人在 1995 年提出,它是在统计学习理论的基础上发展起来的一种通用的学习方法。支持向量机基于统计学习的 VC 维理论和结构风险最小原理,与传统的机器学习方法相比,SVM 具备突出优点:首先它是专门针对有限样本下的最优解,而不是样本数趋于无穷大时的最优解;其次算法最终转化为一个二次型寻优问题,从理论上将得到的是全局最优点,从而避免了传统机器学习方法中易陷于局部极值问题;另外,SVM 的算法是通过非线性变换将实际问题转换到高维的特征空间,在高维的特征空间中构造线性判别函数来实现原空间中的非线性判别函数,解决了高维求解导致的维数灾问题,并且其特殊的性质保证了学习的良好泛化能力。

最小二乘支持向量机 (Least square support vector machine,LS - SVM) 是支持向量机的一种标准扩展[3-4],它采用最小二乘线性系统作为损失函数,用等式约束代替不等式约束,求解过程转化为解一组线性方程组,并且 LS - SVM 不再需要指定不敏感损失函数。相对 SVM,LS - SVM 降低了计算复杂度,加快了求解速度 ,并可获得更高的精度[5]。最小二乘支持向量机(LS - SVM)基本原理如下:

对给定的样本集 $S = \{(x_i,y_i),x_i \in R^n,y_i \in R\},i = 1,2,\cdots,l$, 它的线性回归函数为 $f(x) = w^{\mathrm{T}}\phi(x) + b$。其中,$w_i \in R^n,b \in R$, $\phi(\cdot)$ 为解决非线性问题的核函数。

在 LS - SVM 中,上述回归问题对应的优化问题为:

$$\min_{w,e} Q(w,e) = \frac{1}{2}\| w \|^2 + \frac{c}{2}\sum_{i=1}^{l} e_i^2 \tag{2}$$

约束条件为:

$$y_i = w^{\mathrm{T}}\varphi(x_i) + b + e_i \tag{3}$$

相应的拉格朗日函数为:

$$L(w,b,e,\alpha) = Q(w,b,e) - \sum_{i=1}^{l}\alpha_i[w^{\mathrm{T}}\phi(x_i) + b + e_i - y_i] \tag{4}$$

通过对 w,b,e,α 求偏导数,可以得到该拉格朗日函数的最优化条件为:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l}\alpha_i\phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l}\alpha_i = 0 \\ \dfrac{\partial L}{\partial w} = 0 \Rightarrow Ce_i - \alpha_i = 0 \\ \dfrac{\partial L}{\partial w} = 0 \Rightarrow w^i\phi(x_i) + b + e_i - y_i = 0 \end{cases} \tag{5}$$

上述最优化条件可以转化为矩阵方程形式:

$$\begin{bmatrix} 0 & e^{\mathrm{T}} \\ e & GG^{\mathrm{T}} + \dfrac{1}{C} \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

式中:$y = (y_1,y_2,\cdots,y_N)^{\mathrm{T}}$, e 为 $n \times 1$ 向量,I 为 $n \times n$ 单位向量,$\alpha = (\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$, $G = (\phi(x_1)^{\mathrm{T}},\phi(x_2)^{\mathrm{T}},\cdots,\phi(x_N)^{\mathrm{T}})$,通过解式(6)线性方程组,可以得到 LS - SVM 预测模型:

$$y = \sum_{i=1}^{l}\alpha_i K(x,x_i) + b \tag{7}$$

式中:$K(x,x_i)$ 为满足 Mercei 条件的核函数。

3 用水量非线性组合预测模型的建立

基于最小二乘支持向量机的万元工业 GDP 用水量非线性组合预测是采用其他模型预测的结果作为最小二乘支持向量机的输入向量,将实际万元 GDP 用水量值作为输出,通过训练,建立预测结果与实际值的非线性映射关系,经过学习达到一定精度后,该非线性预测模型成为万元工业 GDP 取水量组合预测的有效工具。

3.1 模型训练样本集的构造

工业用水定额预测的常用方法包括定额法、弹性系数法、相关模型法、重复利用率提高法和趋势法等[6]。由于其涉及的因素较多,难以得出准确的结果。结合当前我国工业用水管理的实际水平,趋势法以其结构简单、不需要率定参数等特点,仍然是我国水资源及水利发展规划当中工业用水预测的推荐方法[7]。根据已有的的文献和实际的应用情况,趋势法中三参数指数模型、阶段乘幂模型和二参数指数模型应用较多。其中三参数指数模型对工业用水预测的精度较高,结果合理可靠,在工业用水管理领域得到广泛的应用[8-9],适合预测长系列的比较平稳的工业取水序列,阶段乘幂模型适合受经济结构调整等经济政策出台等引起用水序列突变的的短期用水预测,多应用于火电等工业行业的预测[10],二参数指数模型应用范围广泛。

趋势法的主要预测模型包括三参数指数模型、二参数指数模型和分段趋势乘幂模型等三种模型,简要介绍如下:

三参数指数模型:

$$W_t = A\exp(B/(t-c)) \tag{8}$$

式中:W_t 为预测的第 t 年万元工业 GDP 取水量;A、B、c 为待定常数;t 为年份。

分段趋势乘幂模型:

$$W_n = An^{-b} \quad (n = 1,2,3,\cdots) \tag{9}$$

式中:W_n 为预测的第 n 年万元工业 GDP 取水量;A、b 为待定常数。

二参数指数模型法:

$$W_n = A\exp^{-nb} \quad (n = 1,2,3,\cdots) \tag{10}$$

式中:W_n 为预测的第 n 年万元工业 GDP 取水量;A、b 为待定常数。

3.2　核函数选择

经常引用的核函数有线性核、多项式核、RBF 核、Sigmoid 核等几种形式。由于 RBF 核函数可以将输入空间以非线性方式影射到特征空间,便于处理现实中以非线性方式存在的问题。其次,与多项式核相比较,RBF 核的参数数目较少,减少了模型的复杂性,所以实践中多使用径向基(RBF)核函数。本文选择径向基函数作为 LS - SVM 的核函数:

$$K(x,x_i) = \exp[(-\|x - x_i\|^2)/\sigma^2] \tag{11}$$

式中:σ 为径向基函数的核宽度。

3.3　模型评价参数

为了定量评价各种模型的预测精度,采用平均绝对误差 *MAE* 以及均方根误差 *RMSE* 作为对预测结果的评估依据,*MAE* 和 *RMSE* 计算公式如下:

$$MAE = \frac{1}{L}\sum_{i=1}^{L}|x_i - \hat{x}_i| \tag{12}$$

$$RMSE = \sqrt{\sum_{i=1}^{L}\frac{(x_i - \hat{x}_i)^2}{L}} \tag{13}$$

式中:i 为样本数,$i = 1,2,\cdots,L$;x_i 为实测值;$\hat{x}_i$ 为预测值。

4　应用实例

采用 2001 - 2011 年河南省万元工业 GDP 取水量数据最为样本,分别应用上述三种趋势预测模型进行预测,同时将其预测结果作为输入序列进行组合预测。

为验证基于 LS - SVM 非线性预测的精度,将线性组合预测作为参照一同分析。线性组合预测采用基于预测方法有效度的优化预测模型[11],并分为一次组合和二次组合两种方式。其中,一次线性组合以两参数指数模型和乘幂模型预测序列为输入样本,二次线性组合以一次组合和三参数指数模型预测序列为输入样本。基于 LS - SVM 的非线性组合预测分为两输入和三输入两种情况,其中两输入采用两参数指数模型和乘幂模型预测序列,三参数还包含三参数指数模型的预测序列。6 种预测模型的具体预测结果见表 1,各个预测模型的精度分析见表 2。

三种单独的趋势法预测模型中,三参数指数模型预测的 *MAE* 为 0.829,*RMSE* 为 1.028;两参数指数模型预测的 *MAE* 为 1.114,*RMSE* 为 1.192;乘幂模型预测的 *MAE* 为 1.929,*RMSE* 为 2.768。可以看出,三参数指数模型预测精度最高,明显优于其他两个模型。

组合预测中,一次线性组合 *MAE* 为 1.114,*RMSE* 为 1.192;二次线性组合 *MAE* 为 0.486,*RMSE* 为 0.773;二输入 LS - SVM 非线性组合预测 *MAE* 为 0.600,*RMSE* 为 0.721;三输入 LS - SVM 非线性组合预测 *MAE* 为 0.578,*RMSE* 为 0.689。结果表明,线性组合的结果优于单项预测,二次线性组合高于一次线性组合,基于 LS - SVM 的非线性组合预测优于线性组合,并且三输入的非线性预测要高于两输入的非线性预测,三输入的非线性预测精度最高。

以上分析计算,说明三参数指数模型确实是一种有效预测模型,但当前三参数指数模型计算多采用基于最小二乘的试算法[8-9]。而在实际应用当中这种方法不仅计算繁杂,并且不太适合基础输入序列数据不平稳即存在强烈突变及连续突变的情况,这在一定程度上制约此模型的应用范围。

表 1　河南省工业万元 GDP 取水量实测值与模型预测值　　m^3

预测模型	2005	2006	2007	2008	2009	2010	2011
实际值	80.6	71.8	63.7	55.4	51.7	46.5	41.7
两参数指数模型	81.6	72.6	64.5	57.4	51.0	45.4	40.3
乘幂模型	80.0	66.5	59.7	55.3	52.1	46.6	44.7
线性组合一次	81.6	72.6	64.5	57.4	51.0	45.4	40.3
三参数指数模型	79.2	70.8	63.5	57.2	51.7	46.9	42.7
线性组合二次	80.6	71.8	64.1	57.3	51.3	46.1	41.4
两输入 LS - SVM	80.6	71.2	63.2	56.8	51.4	45.8	42.4
三输入 LS - SVM	80.7	71.1	63.3	56.8	51.3	45.9	42.3

表 2 6 种模型的计算精度

模型类别	两参数指数模型	乘幂模型	一次组合	三参数指数模型	二次组合	两输入 LS - SVM	三输入 LS - SVM
MAE	1.114	1.929	1.114	0.829	0.486	0.600	0.578
RMSE	1.192	2.768	1.192	1.028	0.773	0.721	0.689

实例表明采用两参数指数模型和乘幂模型的两输入的 LS - SVM 非线性组合预测,其精度已经超过了常用的三参数指数模型预测的精度。所以,建议实际工作中采用多个模型进行组合预测方法,尤其是采用基于 LS - SVM 的非线性组合预测方法,将有效提高预测精度。

5 结 语

提出的基于 LS - SVM 的万元工业 GDP 取水量非线性组合预测模型,综合三参数指数模型、二参数指数模型和分段趋势乘幂模型等万元工业 GDP 取水量趋势预测法的主要模型,实现了三种预测方法优势互补;既考虑长期性的宏观经济调整对工业取水的长期平稳性的影响,又结合近期内受经济结构调整、经济政策出台等对万元工业 GDP 取水量的波动效应,实现短期和长期预测相结合,同时依靠 LS - SVM 强大的非线性处理及寻优能力,提升预测精度,得到满意的预测结果。

参考文献:

[1] 赵永刚,曹红霞,魏新光. 基于组合模型的石羊河流域农业用水量预测[J]. 人民黄河,2012,34(1):99 - 101.

[2] 李黎武,施 周. 基于小波支持向量机的城市用水量非线性组合预测[J]. 中国给水排水,2010,26(1):54 - 56 + 59.

[3] Suykens J A K, Vandewalle J. Least squares support vector machines classifiers[J]. Neural Processing Letters. 1999,9(3):293 - 300.

[4] Suykens J A K. Least squares support vector machines for classification and nonlinear modeling[J]. Neural Network World,2000,10(1):29 - 48.

[5] 张展羽,陈子平,王 斌,等. 基于自由搜索的 LS - SVM 在墒情预测中的应用[J]. 系统工程理论与实践,2010,30(2):201 - 206.

[6] 李 琳,左其亭. 城市用水量预测方法及应用比较研究[J]. 水资源与水工程学报,2005,16(3):6 - 10

[7] 张象明,卢 琼,袁 鹰,等. 松辽流域工业用水定额研究[J]. 中国水利,2006(3):23 - 26.

[8] 秦福兴,耿雷华,陈晓燕. 确定万元 GDP 取水量定额方法的探索[J]. 水利学报,2004,35(8):119 - 122 + 128.

[9] 黄正荣,张振林,贾剑峰,等. 工业用水定额分析与研究[J]. 水资源与水工程学报,2009,20(4):101 - 103.

[10] 宋 轩,耿雷华,杜 霞,等. 我国火电工业取用水量及其定额分析[J]. 水资源与水工程学报,2008,19(6):64 - 66 + 70.

[11] 王明涛. 确定组合预测权系数最优近似解的方法研究[J]. 系统工程理论与实践,2000,20(3):105 - 109.