基于ARIMA和LSSVM的非线性集成预测模型

朱帮助[1*]， 林 健[2]

(1. 五邑大学 管理学院，广东 江门 529020)
(2. 北京航空航天大学 经济管理学院，北京 100083)

摘要： 针对复杂时间序列预测困难的问题，在综合考虑线性与非线性复合特征的基础上，提出一种基于ARIMA和最小二乘支持向量机(LSSVM)的非线性集成预测方法. 首先采用ARIMA模型进行时间序列线性趋势建模，并为LSSVM建模确定输入阶数；接着根据确定的输入阶数进行时间序列样本重构，采用LSSVM模型进行时间序列非线性特征建模；最后采用基于LSSVM的非线性集成技术形成一个综合的预测结果. 将该方法用于中国GDP预测取得的结果，与单独预测方法及流行的其他集成预测方法相比，预测精度有了较大的提高，从而验证了方法的有效性和可行性.

关键词： 时间序列预测；非线性集成；ARIMA；LSSVM

1 引 言

时间序列预测是现代预测领域中最富挑战性的应用之一，一直成为世界各国学者研究的焦点. 传统线性的概率统计模型曾得到广泛的应用，如ARIMA模型，后来还有灰色预测模型、浑沌时间序列预测和人工神经网络等预测方法，其中以人工神经网络最引人注目并且使用最为广泛[1]. 然而，由于人工神经网络学习算法本质上是利用梯度下降法调节权值使目标函数达到极小，导致了神经网络过分强调克服学习错误而泛化性能不强. 同时，人工神经网络还有一些其他难以克服的缺陷，包括易于陷入局部极小、过度拟合、隐层神经元的数目难以确定及网络的最终权值受初始值影响大. 为了解决这些问题，Vapnik于20世纪90年代中期提出了一种新的人工神经网络模型——支持向量机(support vector machine，SVM)[2].

相对于传统人工神经网络基于经验风险最小化的原理，支持向量机是建立在结构风险最小化的原理之上的，前者强调的是训练数据误差的最小化，而后者强调的是推广误差上界的最小化. 因此，支持向量机的解可能是全局最优解，过度拟合也不可能出现[1]. 然而，传统支持向量机的主要缺点在于：支持向量机使用二次规划求解，计算量较大，需要花费较长的训练时间.

针对传统支持向量机存在的问题，1999年Suykens提出了一种改进型支持向量机——最小二乘支持向量机(least square support vector machine，LSSVM)[3]. 与传统支持向量机相比，LSSVM引入最小二乘线性系统到支持向量机中，用等式约束代替不等式约束，求解过程由二次规划方法变为解一组等式方程，求解速度相对加快. 同时，相对于常用的ε-不敏感损失函数，LSSVM不再需要指定逼近精度ε，更容易理解与操作.

鉴于实际问题的复杂性，时间序列数据通常具有线性和非线性的复合特征，单纯的线性或非线性模型都不能够很好地捕捉这种复合性特征[4]. 因此，需要将线性模型和非线性模型

收稿日期：2007-09-18
基金项目：国家自然科学基金(70471074)
*通信作者

集成起来共同开展时间序列预测研究. 最近，一些学者开始注重利用集成预测方法来进行时间序列预测研究，并获得了比单独的线性和非线性模型更好的预测效果. Wedding 提出了一种径向基网络与单变量的Box-Jenkins 模型相结合的集成模型来预测时间序列，获得的预测结果显著优于单独模型的预测结果[5]. Luxho 提出了一个计量经济模型和人工神经网络的集成模型，并用于销售量预测且获得了较好的预测结果[6]. Voort 通过组合Kohonen 自组织映射网络和ARIMA 模型，构建了一个"KARIMA"的集成模型，并用于交通流量预测，获得了较好的预测结果[7]. Tseng 通过集成季节性ARIMA 模型和BP 人工神经网络模型得到一个SARIMABP 模型，并用这个集成模型去预测季节性时间序列，获得了较好的预测结果[8]. Zhang 充分利用ARIMA 模型在线性建模与人工神经网络模型在非线性建模上的优势，建立了一个两者的集成模型，实证结果表明，相对于单独的ARIMA 和人工神经网络模型，该集成模型是一个能有效提高预测精度的模型[9]. Yu 提出了一个基于广义线性回归模型和人工神经网络的集成模型，并用于汇率预测且获得了较好的预测效果[4]. 他们的研究结果都表明了集成预测模型能够有效地提高时间序列预测的精度和可靠性. 然而，在现有文献中，尚未发现基于ARIMA 和LSSVM 的非线性集成预测模型. 本文希望在这方面做些尝试性研究.

2　ARIMA 和LSSVM 的基本原理

2.1　ARIMA 模型

自回归单整移动平均(ARIMA)模型是衡量一个内生变量与其滞后的该变量关系的一个系统模型，即在ARIMA 模型中，一个变量的将来值被假设为该变量的几个滞后项变量的线性函数. ARIMA (p,d,q) 实质上是ARMA (p,q) 的 d 阶单整(即 d 次差分)，它可以将一个非平稳时间序列转化为平稳时间序列[9].

ARIMA (p,d,q) 模型为

$$\Delta y_t = \varphi_1 \Delta y_{t-1} + \varphi_2 \Delta y_{t-2} + \cdots + \varphi_p \Delta y_{t-p} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_q u_{t-q}$$

式中，$\varphi_i (i = 1,2,\cdots,p)$ 为自回归系数，$\theta_i (i = 1,2,\cdots,q)$ 为移动平均系数，是模型的待估参数；Δy_i 为 d 阶单整平稳时间序列，μ_t 为误差项. 显然，Δy_t 可以通过逆向还原为序列 y_t.

ARIMA (p,d,q) 是ARMA (p,q) 模型的一种扩展形式，当 $d = 0$，ARIMA (p,d,q) 就变为ARMA (p,q) 模型. 估计ARIMA (p,d,q) 模型同估计ARMA (p,q) 模型具体的步骤相同，唯一的差异是在估计之前要确定原时间序列的单整(差分)阶数 d.

应用ARIMA (p,d,q) 模型的建模过程包括4 个主要步骤：1)对原时间序列进行平稳性检验，如果序列不满足平稳性条件，可以通过 d 次差分变化使其满足平稳性条件；2)通过计算能够描述序列特征的一些统计量，如自相关系数和偏自相关系数来确定ARIMA 模型的阶数 p 和 q；3)估计ARIMA 模型的未知参数，并检验参数的显著性以及模型本身的合理性；4)进行诊断分析，以证实所得模型确实与所观察到的数据特征相符.

ARIMA (p,d,q) 模型的优势在于其只需要内生变量，不需要任何其他外生变量，因而能够捕捉到内生变量间的影响. 而且，在时间序列预测中，增加一个外生变量可能会增加模型的复杂度，削弱模型的预测能力. 此外，ARIMA (p,d,q) 模型也比较容易理解和操作，因此，本文采用ARIMA (p,d,q) 模型. 当然，ARIMA (p,d,q) 模型的缺点在于它本质上是一类线性模型，因而不能捕捉到时间序列中的非线性特征.

2.2 LSSVM 模型

假定训练样本集：$D=\{(x_k,y_k),k=1,2,\cdots,l\}$，$x_k\in R^n$，$y_k\in R$，$x_k$ 是输入数据，y_k 是输出数据.

LSSVM 定义如下的优化问题[2]：

$$\min \quad J(w,e,b)=\frac{1}{2}w^{\mathrm{T}}w+\frac{1}{2}\gamma\sum_{k=1}^{l}e_k^2$$

$$\text{s.t.} \quad y_k=w^{\mathrm{T}}\varphi(x,k)+b+e_k,\quad k=1,2,\cdots,l$$

其中，$\varphi(\cdot)$ 是非线性空间映射函数，权向量 $w\in R^n$，误差变量 $e_k\in R$ 是偏差量，常数 $\gamma>0$ 是惩罚因子.

定义拉格朗日函数：

$$L(w,e,\alpha,b)=J(w,e)-\sum_{k=1}^{l}\alpha_k\{w^{\mathrm{T}}\varphi(x_k)+b+e_k-y_k\}$$

其中，拉格朗日乘子 $\alpha\in R$.

根据KKT 条件进行求解，令 L 对 w、b、e_k、α_k 的偏导数等于0，即

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\rightarrow w=\sum\limits_{k=1}^{l}\alpha_k\varphi(x_k)\\[2ex]\dfrac{\partial L}{\partial b}=0\rightarrow\sum\limits_{k=1}^{l}\alpha_k=0\\[2ex]\dfrac{\partial L}{\partial e_k}=0\rightarrow\alpha_k=\gamma e_k\\[2ex]\dfrac{\partial L}{\partial\alpha_k}=0\rightarrow w^{\mathrm{T}}\varphi(x_k)+b+e_k-y_k=0\end{cases}$$

消去变量 w、e_k，得

$$\begin{bmatrix}b\\\alpha\end{bmatrix}=\begin{bmatrix}0 & 1_v^{\mathrm{T}}\\1_v & \Omega+\gamma^{-1}I\end{bmatrix}^{-1}\begin{bmatrix}0\\y\end{bmatrix}$$

其中，$y=[y_1,y_2,\cdots,y_l]^{\mathrm{T}}$，$1_v=[1,1,\cdots,1]^{\mathrm{T}}$，$\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_l]^{\mathrm{T}}$，$I$ 为单位矩阵，Ω 为一个方阵，其第 k 行 m 列元素 $\Omega_{km}=\varphi(x_k)^{\mathrm{T}}\varphi(x_m)$，$k,m=1,2,\cdots,l$.

求出 α 和 b，从而得到训练样本集的非线性逼近为

$$y(x)=\sum_{k=1}^{l}\alpha_k\varphi(x)^{\mathrm{T}}\varphi(x_k)+b$$

引入核函数，令 $\varphi(x)^{\mathrm{T}}\varphi(x_k)=K(x,x_k)$，则预测输出为

$$y(x)=\sum_{k=1}^{l}\alpha_kK(x,x_k)+b$$

其中，$K(\cdot)$ 为核函数.

LSSVM 模型的主要优势在于它灵活的非线性建模能力，能够较好地捕捉到时间序列中的非线性特征.

3 基于ARIMA 和LSSVM 的非线性集成模型

现实生活中的时间序列问题通常是非常复杂的，很少是纯粹线性或纯粹非线性的，它们

通常是线性和非线性的综合体，因而在时间序列预测中，没有一个通用的方法能在任何时间序列中都能获得一致好的预测结果. 尽管ARIMA 和LSSVM 在线性和非线性序列中获得了成功，但在时间序列预测中，两种方法中的任何一种都不能充分地建模与预测，因为ARIMA 模型不能很好地处理非线性关系，而LSSVM 模型不能同等地处理时间序列中线性和非线性特征[4]. 因此，ARIMA 和LSSVM 是互补的，把两者集成起来可能会产生一个更加鲁棒的方法，从而可能获得更好的预测结果.

基于ARIMA 和LSSVM 的非线性集成预测建模的基本思路如图1 所示. 首先，利用ARIMA 模型对时间序列的线性趋势建模，为LSSVM 建模确定输入阶数；其次，根据ARIMA 模型确定的输入阶数对时间序列进行样本重构，利用LSSVM 模型对时间序列的非线性特征进行建模；最后，采用基于LSSVM 的非线性集成技术形成一个综合的预测结果.

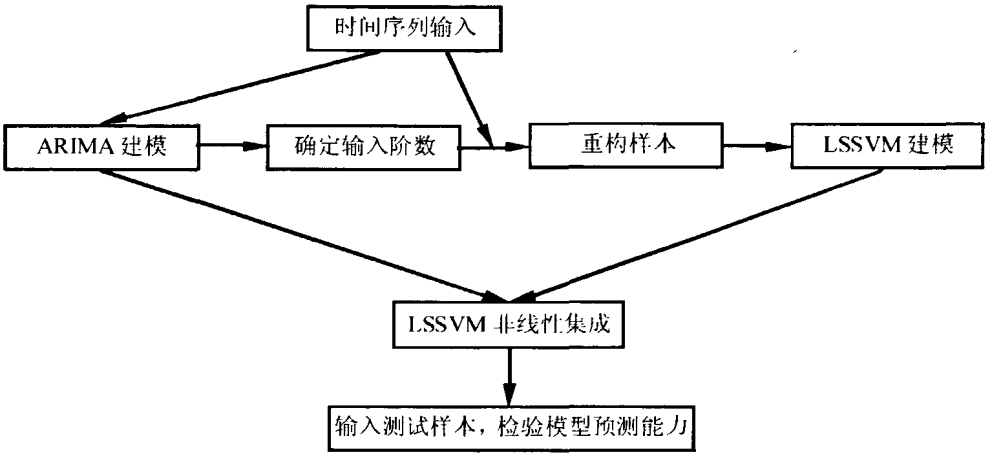

图1　基于ARIMA 和LSSVM 的非线性集成预测建模流程

第一阶段，一个ARIMA 模型被用作拟合时间序列 $\{x_t, t = 1,2,\cdots,n\}$ 的线性部分，记为 $\{\hat{x}_t\}$. 然而，因为时间序列通常还包含一些复杂的非线性特征，因此，仅仅用ARIMA 建模是不充分的. 为了提高预测精度，还需要继续对非线性特征进行建模.

第二阶段，根据第一阶段建立的ARIMA 模型，确定输入阶数（即预测步长），重构时间序列样本，并进行LSSVM 建模，即

$$\hat{x}_{nl} = f(x_{t-1}, x_{t-2}, \cdots, x_{t-r}) + \varepsilon_t$$

式中，$f(\cdot)$ 为由LSSVM 确定的非线性映射函数，r 为预测步长，ε_t 为随机误差. 这样，利用所构建的LSSVM 模型能够预测该时间序列的非线性部分，记为 $\{\hat{x}_{nl}\}$.

第三阶段，为了获得协同效果，将各个体预测模型预测结果集成为最终的预测结果 $\{\hat{y}\}$.

目前，简单平均方法是在集成预测中使用最广泛的一种方法[10]. 即对单独模型的预测结果进行简单平均而获得集成预测结果的方法. 简单平均集成技术简单易行，但将每种预测方法间视为线性关系并不是适用于所有场合，通常它们之间存在着非线性关系[4]. 非线性集成预测方法是近年来兴起的一种新的集成方法，它是利用非线性技术来确定各单独模型的权重[11-12]. 本文采用LSSVM 来确定集成中各单独模型的权重. 具体做法是将各单独模型的预测结果作为LSSVM 模型的输入进行回归，利用核函数将它们转化为支持向量，然后进行学习，进而获得最优权重向量的过程.

一个基于LSSVM 的非线性集成可以看成是一个非线性处理系统，即

$$\hat{y} = f(\hat{x}_l, \hat{x}_{nl})$$

式中，$\hat{x}_l, \hat{x}_{nl}$ 分别是 ARIMA、LSSVM 的预测结果，$\hat{y}$ 是集成预测结果，$f(\cdot)$ 是一个由 LSSVM 确定的非线性映射函数. 利用LSSVM 实现这个非线性映射，实质上是通过LSSVM 的训练来确定集成中的各个体预测模型的权重.

4 模型在中国GDP 预测中的应用

GDP 是衡量一个国家或地区经济发展水平和综合实力大小的重要指标. 改革开放以来，中国的经济发展速度很快，GDP 总量值呈现出很强的非平稳性. GDP 预测一直是各国经济发展预测中一个非常重要的问题，随着经济的发展，对其预测精度的要求也越来越高，这是因为GDP 预测精度的高低对于经济发展目标的制定和调整有显著的影响[13]. 因此需要采用科学有效的方法对其进行预测. 本文从GDP 总量入手，利用1982～2006 年的中国GDP 总量，运用上述构建的非线性集成模型，进行中国GDP 的总量预测，数据来源于中国统计年鉴[14]. 本文将1982～1999 年的18 个GDP 观测值作为训练样本，用来构建模型；2000～2006 年的7 个GDP 观测值作为测试样本，用来检验模型的有效性.

本文借助 EViews 5.0[15]构建 ARIMA 模型，采用静态预测方法，经过比较分析，选择 ARIMA(1,1,0)模型 ($R^2 = 0.7244$, Akaike Info Criterion=18.4099，均方误差：2123.943，平均绝对误差：1652.378，平均绝对相对误差：5.58%). 从中可以发现，ARIMA(1,1,0)模型的 拟合效果较好，可以用来进行GDP 预测. 测试样本的均方误差：12431.97，平均绝对误差：8617.444，平均绝对相对误差：5.54%. 由此可见，ARIMA 模型取得了较好的预测效果.

从模型可以发现，ARIMA 模型选择了1 个 (GDP_{t-1}) 对当期GDP 总量产生重要影响的因素. 这就意味着，可以将前一期的GDP 总量作为LSSVM 的输入来预测当期GDP 总量，即LSSVM 的预测步长为1. 采用MATLAB 7.01 平台和LSSVMlab 1.5 工具箱[16]编程实现 LSSVM 模型，核函数选择高斯径向基核函数. 经过反复实验，当 $\gamma = 1000000, \sigma^2 = 80000$ 时，LSSVM 达到了较好的训练和测试效果.

在获得ARIMA 和LSSVM 的预测结果后，将ARIMA 和LSSVM 的预测结果作为LSSVM 的输入、GDP 观测值作为LSSVM 的输出进行非线性集成建模，核函数仍选择高斯径向基核函数，且 $\gamma = 1000000, \sigma^2 = 400000$. 同时为进一步说明本文提出方法的优越性，选择目前较为流行的基于简单平均的线性集成方法进行对比. 各种方法的预测结果见图2(计量单位：亿元).

为评价预测性能，本文使用均方误差(RMSE)、平均绝对相对误差(MAPE)作为模型的评价准则. RMSE、MAPE 分别定义为

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right| \times 100\%$$

其中，x_t 为GDP 序列第 t 期的实际观测值，$\hat{x}_t$ 为某种预测方法第 t 期的预测值，n 为测试期数，$t = 1, 2, \cdots, n$. 显然，RMSE、MAPE 愈小，则预测精度愈高，误差愈小. 各种预测方法的精度对比结果见表1.

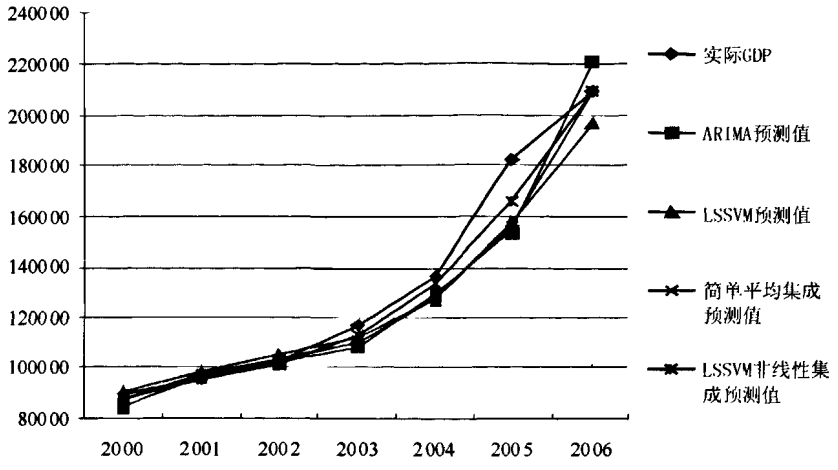

图 2　各种方法的预测结果与实际GDP 对比

从表1 对各种预测方法的预测精度的对比可以发现：1）无论从RMSE 角度还是从MAPE 角度，本文提出的基于LSSVM 的非线性集成预测方法的预测精度与单个预测方法比较有较大的改进，其中相对于ARIMA 方法提高的程度最大；2）与基于简单平均的线性集成预测相比，预测精度也有较大幅度的提高. 因此，本文提出的基于LSSVM 的非线性集成预测方法是最优的，究其原因可能是该方法综合利用了ARIMA 模型和LSSVM 模型各自的优势，充分发挥了它们的协同作用. 3）本文再次证实前人的研究结论，无论是非线性集成预测方法还是线性集成预测方法都能够获得比单独预测方法更好的预测效果.

表 1　各种预测方法预测效果比较

预测方法 / 评价指标	RMSE	MAPE
ARIMA	12431.97	5.54%
LSSVM	11425.24	5.62%
简单平均集成	10917.65	4.84%
LSSVM 非线性集成	6552.25	2.71%

5　总　　结

本文将计量经济模型ARIMA 和最小二乘支持向量机LSSVM 相结合，提出了基于LSSVM 的非线性集成预测方法，为解决复杂时间序列建立LSSVM 模型过程中难于确定输入节点数及非线性集成实现问题提供了较好的方法. 该方法用于中国GDP 总量的预测，所得结果显示，采用该方法进行预测所得到的预测结果与实际情况拟合较好；RMSE、MAPE 与单个预测方法相比均有较大程度的减少，与目前主流的基于简单平均的线性集成方法相比，在预测精度上也有较大幅度的改进. 实证研究结果表明，本文提出的非线性集成预测方法的预测效果最好，不仅能够有效捕捉到变量间的线性关系，而且能够实现变量间的非线性映射关系.

实践中，采用本方法的关键在于时间序列预测步长的确定问题，因为它直接决定着ARIMA 模型的类型与待估参数以及LSSVM 模型的输入变量. 此外，核函数参数的确定和LSSVM 参数的选择优化问题，也是应用本方法必须要解决的基本问题. 此外，从测试样本的预测效果看，本方法对于中长期预测的精度和能力尚有待于进一步提高.

参考文献：

[1] 余乐安，汪寿阳，黎建强. 外汇汇率与国际原油价格波动预测—TEI@I 方法论[M]. 长沙：湖南大学出版社，2006.
[2] Vapnik V. The Nature of Statistics Learning Theory[M]. New York: Springer Verlag, 2000.
[3] Suykenns J A K, Vandewalle J. Least squares support vector machine[J]. Neural Processing Letter, 1999, 9(3): 293-300.
[4] Yu L, Wang S Y, Lai K K. A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates[J]. Computer & Operation Research, 2005, (32): 2523-2541.
[5] Wedding II D K, Cios K J. Time series forecasting by combining RBF networks, certainty factors, and the Box-Jenkins model[J]. Neurocomputing, 1996, (10): 149-168.
[6] Luxhoj J T, Riis J O, Stensballe B. A hybrid econometric neural network modeling approach for sale forecasting[J]. International Journal of Production Economics, 1996, (43): 175-192.
[7] Voort M V D, Dougherty M, Watson S. Combining kohonen maps with ARIMA time series models to forecast traffic flow[J]. Transportation Research Part C: Emerging Technologies, 1996, (4): 307-318.
[8] Tseng F M, Yu H C, Tzeng G H. Combining neural network model with seasonal time series ARIMA model[J]. Technological Forecasting and Social Change, 2002, (69): 71-87.
[9] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003, (50): 159-175.
[10] Lin J, Zhu B Z. Neural network ensemble based on forecasting effective measure and its application[J]. Journal of Computational Information Systems, 2006, (6): 781-787.
[11] Shi S M, Xu L D, Liu B. Improving the accuracy of nonlinear combined forecasting using neural networks[J]. Expert Systems with Applications, 1999, (16): 49-54.
[12] Zhu B Z, Lin J. A novel feature extraction-based selective & nonlinear neural network ensemble model for economic forecasting[J]. International Journal of Computer Science and Network Security, 2007, 7(2): 142-145.
[13] 肖智，吴慰. 基于PSO-PLS 的组合预测方法在GDP 预测中的应用[J]. 管理科学，2008，21(3)：115-121.
[14] 中国统计局. 中国统计年鉴-2007[M]. 北京：中国统计出版社，2008.
[15] 张晓峒. EViews 使用指南与案例[M]. 北京：机械工业出版社，2007.
[16] http://www.esat.kuleuven.ac.be/sista/lssvmlab

A Novel Nonlinear Ensemble Forecasting Model Incorporating ARIMA and LSSVM

ZHU Bang-zhu[1], LIN Jian[2]

(1. School of Management, Wuyi University, Jiangmen 529020, China)
(2. School of Economics and Management, Beihang University, Beijing 100083, China)

Abstract: In order to solve the problem of complex time series forecasting including the linear and nonlinear features, a new ensemble forecasting model incorporating ARIMA and LSSVM is proposed in this paper. This ensemble model uses ARIMA model to capture the linear feature of the time series and LSSVM model to fit the nonlinear component of the time series to obtain the synergetic forecasting results by using LSSVM. The validity of the proposed model has been examined by forecasting GDP of our country. Compared with the traditional forecasting methods and the other popular ensemble forecasting method, the result of the presented method is more accurate.

Keywords: time series forecasting; nonlinear ensemble model; ARIMA; LSSVM