文章编号:1000-6788(2006)04-0097-07

区域经济中长期预测的支持向量回归方法

肖健华[1,2],林　健[3],刘　晋[3]

(1. 五邑大学智能技术与系统研究所,广东 江门 529020;2. 北京航空航天大学经济管理学院,北京 100083;
3. 五邑大学管理学院,广东 江门 529020)

摘要:　分析了区域经济发展特性以及中长期经济预测的特点,对当前经济预测方法存在的不足进行了阐述,指出:由于区域经济系统中存在高度的非线性、耦合性和时变性,使得现有的经济预测方法难以胜任.介绍了支持向量回归算法,并在此基础上,提出了基于支持向量回归的方法对区域经济进行中长期预测的思路,并建立了相应的数学模型.以广东省江门市作为应用对象,说明了该模型的有效性.

关键词:　支持向量回归;统计学习理论;区域经济;中长期预测

中图分类号:　TP181　　**文献标识码**:　A

A SVR-based Model for Regional Economy Medium-term and Long-term Forecast

XIAO Jian-hua[1,2], LIN Jian[3], LIU Jin[3]

(1. Institute of Intelligent Technology and Systems, Wuyi University, Jiangmen 529020, China; 2. School of Economics and Management, Beihang University, Beijing 100083, China; 3. School of Management, Wuyi University, Jiangmen 529020, China)

Abstract:　Based on the analysis of characteristics of regional economy and forecasting methods for medium-term and long-term economic development, drawbacks of current forecasting methods were explained as that those methods are not suitable for forecasting medium-term and long-term economic development due to its nonlinearity, coupling and dynamicity. A support vector regression algorithm was been introduced, and then an idea of forecasting medium-term and long-term regional economy based on the algorithm was explained and a mathematic model based on the idea was proposed. At last, an experiment was conducted to verify the proposed model on the economic dataset of Jiangmen, Guangdong.

Key words:　support vector regression (SVR); statistic learning theory (SLT); regional economy; medium-term; long-term forecast

1　引言

经济发展的中长期预测,一般是指五年以上经济发展趋势的预测,旨在为一个国家或一个区域的中远景规划提供科学的依据.

就经济发展预测而言,与国家宏观经济发展规律相比,区域经济的发展存在自身的特点.首先是波动性大,而且所研究的区域越小波动性越大,有时甚至一个企业的兴衰、降雨量的多少等都可能对一个区域的经济发展产生很大的影响;其次是系统的相对独立性和开放流动性,单个区域的独立性相对较小,各个区域各有侧重,某个区域经济必然与其它的区域经济形成互补.

当然,区域经济的发展预测也具备一般经济系统预测的共同特性[1]:非线性,区域经济系统是众多确定性因素和非确定性因素交互作用下的非线性系统;强耦合性,反映经济发展的各种指标、构成经济系统的各行业、各部门等无不密切关联,每一个指标或行业的变化都会导致其它指标或行业的变化;时变性,经

收稿日期:2004-10-10

资助项目:国家自然科学基金(70471074);中国博士后科学基金(2005038042)

作者简介:肖健华(1970 -),男,汉族,江西永新人,博士后,副教授,主要研究方向:智能信息处理,复杂经济系统建模,E-mail:jianhuaxiao@tom.com;林健(1958 -),男,汉族,福建福州人,博士生导师,五邑大学校长,主要研究方向为复杂系统建模与仿真;刘晋(1956 -),女,湖北孝感人,博士,教授,主要研究方向为管理决策支持系统.

济系统是一个动态的开放系统,反映其运行规律的数学模型始终在不断的变化中.

在对区域经济的发展进行预测时,必须充分考虑上述各种因素,尤其是蕴含在指标数据上的非线性、时变性和不确定性作用关系.

建立在计量经济学理论基础上的各种经济预测模型,大部分属于线性模型.线性模型在发挥巨大作用的同时,也逐渐显露出它的缺点,即很难把握经济系统中的非线性现象,最终必然造成预测结果的较大误差.为弥补这一缺陷,经济领域的一些研究工作者对线性模型进行了修正,如建立分段线性模型和变参数线性模型等,但结果往往不理想[2].以神经网络为代表的非线性建模方法,一度给经济预测带来了希望,理论上也证明了在选择适当的隐层数及相应的神经元数目下,前馈神经网络能以任意精度逼近任意非线性函数.然而在实际应用中,由于神经网络存在的一些缺陷,使得神经网络应用到实际经济系统的发展预测中还存在一定问题.这些缺陷包括:网络结构不能保证最优化;训练算法存在局限,收敛速度得不到保障,且容易陷入局部最优;对训练样本的数量与质量要求较高;基于经验风险最小化的优化目标,不能保证泛化性能[3].

由 Vapnik 提出的统计学习理论(Statistic Learning Theory, SLT)[4]发展而来的核方法(Kernel Method, KM)[5],实现了数据空间与特征空间之间的非线性映射,可以有效地将数据空间中的各种非线性操作演变为特征空间中相应的线性操作,进而大大地提高了非线性处理能力.

作为核方法的一种,支持向量回归(Support Vector Regression, SVR)在非线性回归中具备非常优秀的性能[6].更为难能可贵的是,SVR 建立在结构风险最小化的优化目标上,很大程度上改善了神经网络在非线性拟合上存在的不足.

本文首先介绍 SVR 算法,进而对广东省江门市的经济发展历史数据进行了分析,在此基础上提出基于 SVR 的区域经济中长期预测方法,并建立了相应的数学模型.

2 支持向量回归算法

考虑给定的 n 个学习样本(X_i, y_i), $X_i \in R^d$, $y_i \in R$, $i = 1, 2, \cdots, n$,线性回归的目标就是求回归函数

$$f(X) = (W \cdot X) + b. \tag{1}$$

式中:$W \in R^d$, $b \in R$, $(W \cdot X)$为 W 与 X 的内积.在以往的学习算法中,优化目标是使经验风险即样本损失函数 $L(X_i)$

$$L(X_i) = g(y_i - f(X_i)). \tag{2}$$

的累积 $R_{\text{emp}}(f)$最小化,如最小二乘法,所求的(W, b)应满足

$$\min R_{\text{emp}}(f) = \sum_{i=1}^{n} (y_i - f(X_i))^2. \tag{3}$$

然而,统计学习理论指出,经验风险最小并不能保证期望风险最小[4].在结构风险最小化的优化目标下,线性回归方程式(1)中的参数(W, b)应满足

$$\min Q(W, b) = \frac{1}{2} \| W \|^2 + CR_{\text{emp}}(f). \tag{4}$$

上式中的 $\| W \|^2/2$ 反映了回归函数 $f(X)$的泛化能力,C 为惩罚因子.式(4)表明,结构风险最小化能够折中考虑回归函数的经验风险和泛化能力,因此,回归函数具有更好的性能.

式(2)中常用的损失函数 $L(X_i)$包括二次函数、Huber 函数、Laplace 函数和 ε不敏感函数等.其中 ε不敏感函数能够忽略 ε范围内的回归误差

$$L_\varepsilon(X_i) = \begin{cases} 0, & |y_i - f(X_i)| \le \varepsilon \\ |y_i - f(X_i)| - \varepsilon, & |y_i - f(X_i)| > \varepsilon \end{cases}. \tag{5}$$

比较适合于经济数据处理.

如果,$|y_i - (W \cdot X_i) - b| < \varepsilon (i = 1, 2, \cdots, n)$成立,对应图 1 中外面两条直线所围区域内的样本点,即所有样本的损失函数都为 0,因此有 $R_{\text{emp}}(f) = 0$,式(4)可改写为:

$$\min\ Q(W,b) = \frac{1}{2}\|W\|^2, \tag{6}$$

$$\text{s.t. } y_i - (W\cdot X_i) - b \le \varepsilon, \tag{7}$$

$$(W\cdot X_i) + b - y_i \le \varepsilon. \tag{8}$$

显然,约束条件式(7)和式(8)并不总能得到满足,此时则必须引入松弛因子 $\xi_i \ge 0$ 和 $\xi_i^* \ge 0$,如图 1 所示,在 ε 不敏感损失函数下,式(4)的优化问题变为[6,7]

$$\min\ Q(W,b) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{9}$$

$$\text{s.t. } y_i - (W\cdot X_i) - b \le \varepsilon + \xi_i, \tag{10}$$

$$(W\cdot X_i) + b - y_i \le \varepsilon + \xi_i^*, \tag{11}$$

$$\xi_i, \xi_i^* \ge 0. \tag{12}$$

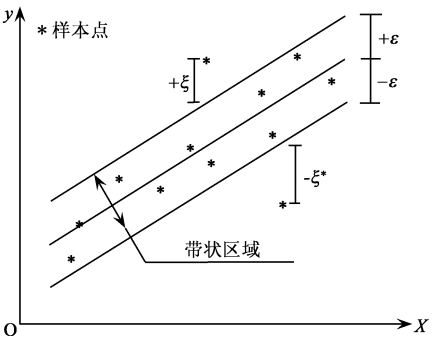


图 1　不敏感函数下的线性回归示意图

引入参数 $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$,构造 Lagrange 函数对上述优化问题进行求解

$$L := \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + (W\cdot X_i) + b) - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - (W\cdot X_i) - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*). \tag{13}$$

考虑到上式关于 W, b, ξ_i, ξ_i^* 取极小,因此对 L 关于 W, b, ξ_i, ξ_i^* 求偏导,并令它们等于 0

$$\frac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0, \tag{14}$$

$$\frac{\partial L}{\partial W} = 0 \to W = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)X_i, \tag{15}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \to C - \alpha_i - \eta_i = 0, \tag{16}$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \to C - \alpha_i^* - \eta_i^* = 0. \tag{17}$$

将式(14)～式(17)代入式(13),得到对偶优化问题

$$\max Q(\alpha_i, \alpha_i^*) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(X_i, X_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*), \tag{18}$$

$$\text{s.t. } \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \tag{19}$$

$$0 \le \alpha_i, \alpha_i^* \le C. \tag{20}$$

由上述优化方程,可求出 α_i 和 α_i^*. 实际上只有一部分 $\alpha_i - \alpha_i^* \ne 0$[4],与之对应的样本 (X_i, y_i) 称为支持向量(Support Vector, SV). 进一步由式(1)和式(15)得到回归方程

$$f(X) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)(X_i\cdot X) + b. \tag{21}$$

根据 KKT 条件,任选一支持向量,上式中 b 依下式计算

$$\begin{cases} b = y_i - (W\cdot X_i) - \varepsilon, & \alpha_i \in (0, C) \\ b = y_i - (W\cdot X_i) + \varepsilon, & \alpha_i^* \in (0, C) \end{cases}. \tag{22}$$

将式(15)代入上式得

$$\begin{cases} b = y_i - \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(X_j\cdot X_i) - \varepsilon, & \alpha_i \in (0, C) \\ b = y_i - \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(X_k\cdot X_i) + \varepsilon, & \alpha_i^* \in (0, C) \end{cases}. \tag{23}$$

对于非线性问题,引入核函数 $K(X_i, X_j)$ 代替样本向量的内积运算,实现数据空间到特征空间的非线性映射,并使低维数据空间的非线性问题转化为高维特征空间的线性问题. 在核函数下,式(18)、式(21)和式(23)变为如下形式

$$\max Q(\alpha_i, \alpha_i^*) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(X_i, X_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*), \quad (24)$$

$$f(X) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(X_i \cdot X) + b, \quad (25)$$

$$\begin{cases} b = y_i - \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)K(X_j \cdot X_i) - \varepsilon \quad \alpha_i \in (0, C) \\ b = y_i - \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)K(X_j \cdot X_i) + \varepsilon \quad \alpha_i^* \in (0, C) \end{cases}. \quad (26)$$

其中式(24)的约束条件依然为式(19)和式(20),对式(24)进行求解可的 α_i 和 α_i^*. 以上各式中常用的核函数包括多项式核函数、高斯径向基核函数、Sigmoid 核函数等.

式(25)即为支持向量回归模型. 从上面的论述中,不难发现与常规的回归方法相比,支持向量回归模型具有两方面的优势:采用结构风险最小化作为优化目标,提高了回归函数的泛化能力;引入了核方法,实现了低维数据空间与高维特征空间的非线性映射,提高了回归函数的非线性数据处理能力.

3 基于支持向量回归的区域经济发展中长期预测:以江门市为例

广东省江门市地处珠三角的西部,人口约 400 万,土地接近 10000 平方公里,人均 GDP 接近 2000 美元,国内外经验表明,该市经济发展正处于高速增长期. 对经济发展进行准确的预测和科学的决策,对于促进江门市经济持续、快速、健康发展具有重要的战略意义.

正是基于上述的背景,江门市政府对开展江门市经济发展的中长期预测十分重视,专门委托五邑大学管理学院进行江门市重大攻关课题"江门市经济发展预测与决策支持系统"的研究工作,本文正是这一项目的部分研究成果.

3.1 区域经济发展中长期预测模型参数的确定

下图(2) (a)～(c)分别是江门市最近 25 年 GDP 净值、增长率、净值的自然对数图.

从上面各图不难看出,GDP 净值的自然对数呈现较强的规律,基本上是线性增长,因而净值的自然对数要容易预测一些. 实际上,如图 2(c)所示,如果将 GDP 对数值中的线性部分去掉,所需预测的非线性部分就控制在一个非常小的区域中,如图 2(d)所示. 显然,以图 2(d)所示的曲线作为预测对象进行预测,并最终将预测结果返回到 GDP 净值中,所得的结果比直接预测 GDP 净值精度要高.

进一步,在 GDP 中扣除物价指数的影响,具体办法是以现有数据库的最后一年的 GDP 数据及其对应的物价指数为标准值,其余数据做相应的调整,这样可方便以后的预测. 在后面的叙述中将经过物价调整的 GDP 数据称为可比 GDP 数据. 调整公式为

$$\text{可比 GDP}_{\text{当前年}} = \frac{\text{物价指数}_{\text{最后一年}}}{\text{物价指数}_{\text{当前年}}} \times \text{实际 GDP}_{\text{当前年}}. \quad (27)$$

调整后的 GDP 净值及其自然对数如下图所示,由于 1982 年前的物价指数不可靠,故将之前的数据舍弃. 图 3 表明,除 1989 年前后外,其它数据的线性特征相当明显,特别是在 1998 年以后.

下面首先讨论如何确定图 3(b)中的线性关系. 考虑到在平稳情况下,GDP 保持一定的比例增长,即

$$\text{GDP}(t) = k' \times \text{GDP}(t-1). \quad (28)$$

从而

$$\log\text{GDP}(t) = \log k' + \log\text{GDP}(t-1) = t\log k' + \log\text{GDP}(0). \quad (29)$$

表明 $\log\text{GDP}(t)$ 基本上呈线性增长. 令 $y(t) = \log\text{GDP}(t)$,设其满足

$$y(t) = kt + b. \quad (30)$$

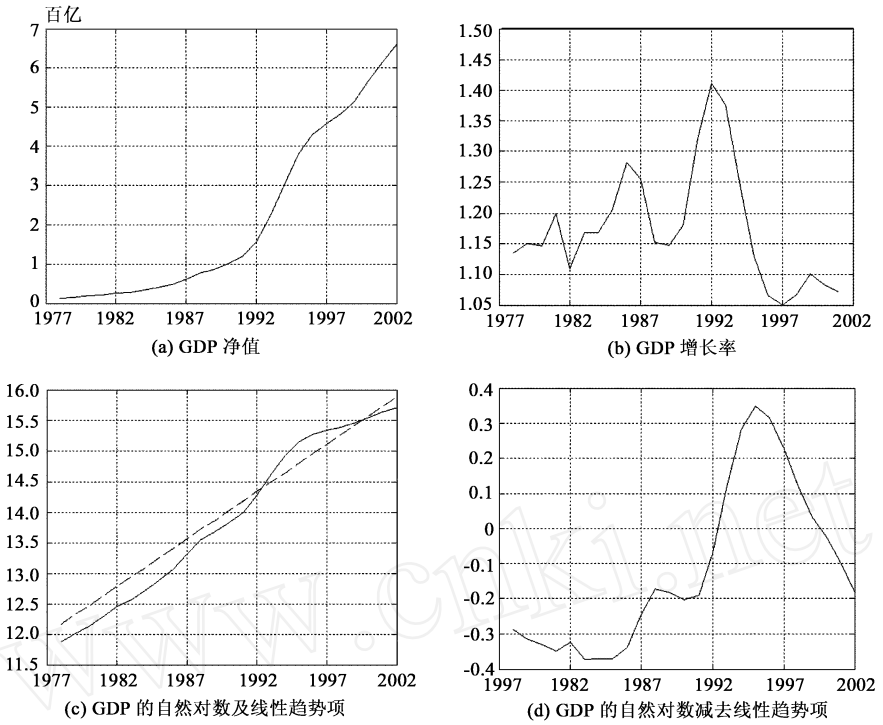


图 2　1978～2002 年间江门市 GDP 相关各量示意图

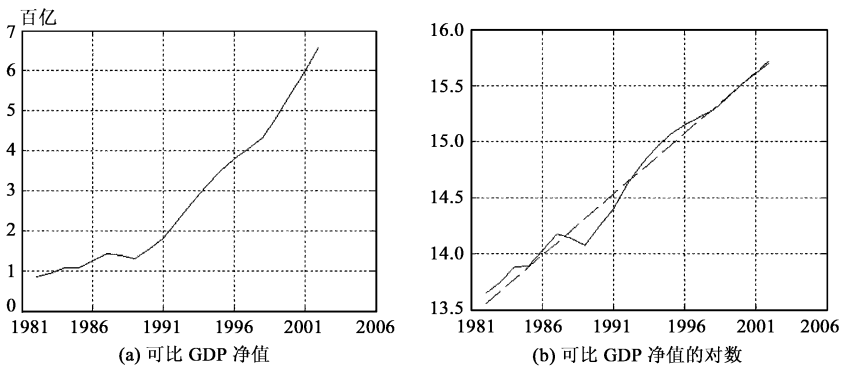


图 3　扣除物价影响后的江门市 1982～2002 年 GDP 相关各量示意图

式中的参数 k, t 可通过最小二乘法求出. 在求解中为了体现对近期数据的重视,可对样本数据进行加权操作,加权比例采用 1.2. 最终所得的趋势项为

$$y(t) = 0.10748 \times t - 199.45. \tag{31}$$

对应的图形如图 3(b) 虚线所示. 最终将可比 GDP 的对数值减去趋势项,所得结果如图 4 所示,为方便叙述,称之为预测变量. 与图 2(d) 相比,预测变量的变化幅度小了一半以上,可见,将预测变量作为研究对象的优势不言而喻.

对预测变量进行自相关分析,如图 5 所示,该图表明,如对预测变量进行回归,只需采用前 8 年的数据即可. 以 $Z(t)$ 表示第 t 年的预测变量值,有

$$Z(t) = f(X_t) = f(Z(t-1), \cdots, Z(t-8)). \tag{32}$$

至此,我们确定了待预测量研究回归模型的阶数.

3.2　区域经济发展中长期预测的 SVR 模型

至今为止,SVR 模型应用研究的最大难点在于模型参数的确定,包括核函数的形式及函数相关参数、折中系数 C 的确定,如果选择 ε 不敏感函数作为损失函数,还存在 ε 具体数值的确定问题. 这些,至今没

有严格的理论作指导,一定程度上依赖于使用者的经验和试凑与比较.

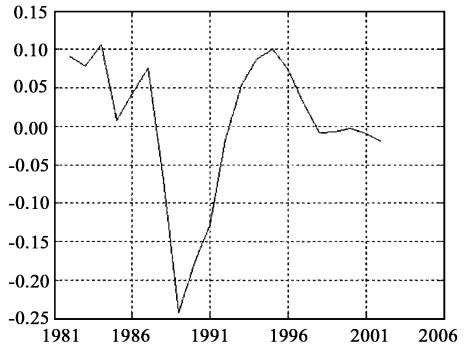


图 4 预测变量

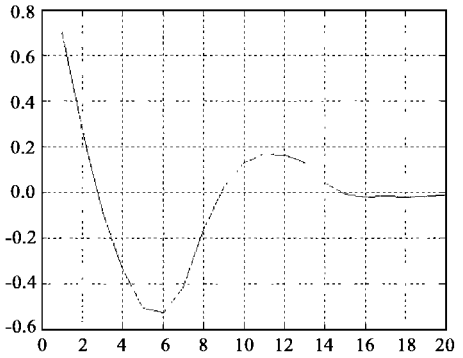


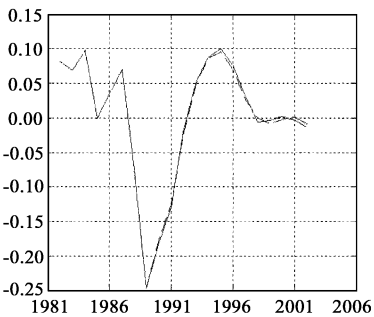
图 5 预测变量自相关图

在本文中,系数 $C = 100$, $\varepsilon = 0.005$,核函数采用高斯径向基函数

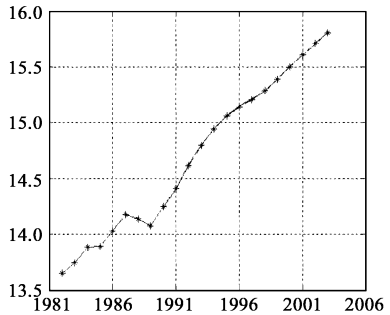
$$K(X_i, X_j) = \exp(-|X_i - X_j|^2/8). \tag{33}$$

3.2.1 对历史数据的拟合分析

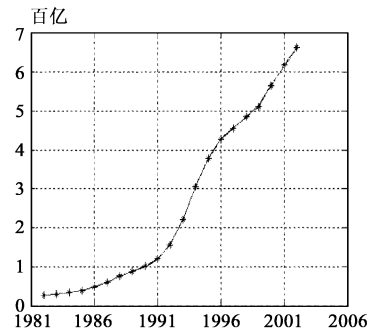
图 6 为采用支持向量回归进行预测的历史数据拟合效果图,可见,效果还是比较理想的. 图 6(a)中,实线为预测数据的真实值,虚线为预测数据的回归值,吻合程度相当高;图 6(b)是在图 6(a)的基础上加回图 3(b)虚线所示的趋势项所得的结果,图中实线为实际可比 GDP 的自然对数值,“ * ”对应预测值. 显然,如果进一步将 GDP 的自然对数值还原成 GDP 当年净值(不考虑物价的影响),会得到更高的拟合性能,如图 6(c)所示,图中实线为 GDP 的当年净值,“ * ”对应预测值.



(a) 预测变量的预测拟合情况



(b) 可比 GDP 对数的预测拟合情况



(c)GDP 的预测拟合情况

图 6 SVR 模型对历史数据的拟合情况

表 1 中长期模型对未来 20 年的预测数据表

年份	2003	2004	2005	2006	2007
GDP	7.36	8.20	9.23	10.40	11.66
增长 %	11.30	11.47	12.57	12.64	12.12
年份	2008	2009	2010	2011	2012
GDP	13.00	14.44	16.02	17.80	19.79
增长 %	11.48	11.10	11.00	11.04	11.20
年份	2013	2014	2015	2016	2017
GDP	22.01	24.47	27.22	30.29	33.74
增长 %	11.21	11.21	11.21	11.29	11.39
年份	2018	2019	2020	2021	2022
GDP	37.60	41.92	46.72	52.04	57.95
增长 %	11.46	11.48	11.44	11.39	11.35

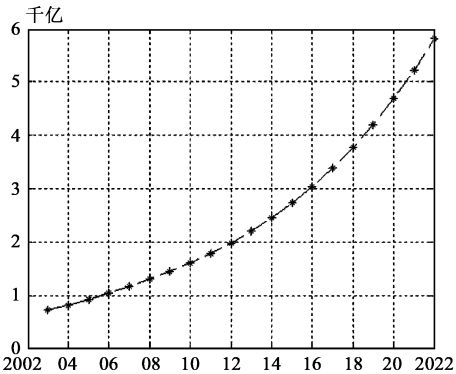


图 7 2003～2022 年间江门市 GDP 预测情况

3.2.2　对未来 20 年的预测

采用支持向量回归的方法对未来 20 年进行预测,所得结果如下表(表中 GDP 单位为百亿元),同时用图形表示为图 7.

预测结果表明,在未来的 20 年内,在现有的环境不发生重大变化的情况下,江门市的经济基本保持 11 %～12.5 %的速度较快增长,并预计 GDP 在 2006 年突破 1000 亿大关.

3.2.3　对 2003 年和 2004 年数据的检验

本论文涉及的研究成果在 2003 年 11 月完成,当时预计 2003 年江门市经济增长 11.3 %,在此后江门市政府公布的数据显示,实际增长 11.1 %.

本成果对 2004 年的预计增长是 11.47 %,前 8 个月的统计数据显示是 12 %,如果扣除物价上涨因素,与系统的预测数据也基本上保持一致.

4　结论

支持向量回归是人工智能领域的最新研究成果,目前已与支持向量分类、基于核的主成分分析等核方法一起成为相关领域研究的前沿和热点.尽管相关的研究还有待于进一步的深入,但它们在非线性数据的处理能力以及对未知样本的泛化能力等方面的优势吸引了众多理论研究者和工程技术人员的广泛关注,其应用面也在迅速的扩展.

将支持向量回归应用到区域经济发展的中长期预测中,属于全新的研究内容,从支持向量回归理论自身具备的特性上讲,这一应用是完全可行的.本文的应用效果也充分说明预测精度是可以接受的.

但正因为支持向量回归本身仍处于发展与完善中,其在经济领域中的应用也刚刚展开,相关的研究还必须深入,尤其是作为一个复杂的社会系统,单纯的定量研究显然不够,如何在相关研究中,融合定性的经济分析知识,并在分析、处理中,更多地进行人机交互,应该是很有意思的一个研究课题.

参考文献:

[1] 郭崇慧.地区中长期发展规划若干定量模型、算法及应用研究[D].大连理工大学,2004.
Guo Chonghui. Study on quantitative models, algorithms and applications for regional medium-and long-term development planning [D]. Dalian University of Technology, 2004.

[2] 王维,贺京同,张建勋,等.人工神经网络在非线性经济预测中的应用[J].系统工程学报,2000,15(2):202 - 207.
Wang Wei, He Jingtong, Zhang Jianxun, et al. Applying artificial neural network to the predicting of nonlinear economy[J]. Journal of Systems Engineering, 2000,15(2):202 - 207.

[3] 邵惠鹤.支持向量机理论及其应用[A].自动化博览二十周年纪念文集,2003,90 - 95.
Shao H. Support vector machines theory and its application[A]. 20th Anniversary Corpus of Automation Panorama[C]. 2003,90 - 95.

[4] Vapnik V. The Nature of Statistical Learning[M]. Theory. New York: Springer-Verlag,1995.

[5] Muller K R, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms[J]. IEEE Trans on Neural Networks, 2001,12(2):181 - 201.

[6] Smola A J, Scholkopf B. A tutorial on support vector regression[R]. NeuroCOLT TR NC-TR-98-030, Royal Holloway College University of London, UK,1998.

[7] Gunn S. Support vector machines for classification and regression[R]. Technical Report. University of Southamptom, 1998.