带约束的最优多元回归模型及其应用*)

肖健华[1] 林 健[2] 刘 晋[2]
(五邑大学智能技术与系统研究所 广东江门 529020)[1]
(五邑大学管理学院 广东江门 529020)[2]

摘 要 针对区域经济发展短期预测建模的特点,结合核方法和支持向量回归的研究进展,提出了一类带约束的最优多元回归模型,该模型综合考虑了多元回归函数的拟合误差、泛化能力以及经济预测的特点,为区域经济的发展预测提供了一种新方案,对广东省江门市最近五年经济发展的预测也验证了该模型的有效性。

关键词 多元回归,支持向量回归,经济预测

Optimal Multiple Regression Model with Constraints and its Application

XIAO Jian-Hua[1] LIN Jian[2] LIU Jin[2]
(Institute of Intelligent Technology and Systems, Wuyi University, Guangdong Jiangmen 529202)[1]
(School of Management, Wuyi University, Guangdong Jiangmen 529202)[2]

Abstract Aimming at the characteristic of short-term forecasting for regional economy, combining with the research achievements of kernel methods and support vector regression, we propose an optimal multiple regression model with constraints. The new model has the specialty of taking the fitting error, generalization ability and the characteristic of economical forecasting into account. In the end of paper, we take Jiangmen City as an example to examine the validity of the new model.

Keywords Multiple regression, Support vector regression, Economic forecasting

1 研究背景

与国家宏观经济预测相比较,区域经济的发展存在自身的特点。首先是经济发展的状况与影响经济发展的指标存在高度的非线性,这一点已在国内外的许多论文和专著中得到关注[1,2],究其原因,主要是因为经济系统自身的复杂性;其次是经济的发展存在更大的波动性,且这种波动性随着区域的缩小而加剧,对于部分小型经济区域,一个大型企业的兴衰、一场暴雨或台风,甚至部分商品价格的变动都可能对该区域经济发展产生决定性的影响。而对于国家而言,这些对经济的影响要小得多,典型的事例是 1998 年的水灾和 2003 年的非典,国家基本上还是以 8%左右的预定速度增长。

区域经济发展的大波动性在图 1 中得到体现,图中实线为 1983 年~2002 年广东省江门市的 GDP 环比数据,虚线为同期国家 GDP 环比数据。位于珠三角的江门市目前人口 382 万,占地接近 10000 平方公里。在图 1 中,不难看出,区域经济发展的波动性远大于国家宏观经济发展的波动性,实际上,如果剔除 1990 年前后由于政治事件造成经济的波动,国家经济整体发展是相当平稳的。

区域经济发展的强非线性和大波动性,使得相关研究人员意识到传统的基于计量经济学的模型难以取得理想的经济预测效果。以神经网络为代表的非线性建模手段,也一度在区域经济发展预测中得到应用,然而深入的理论研究和实际应用表明预测效果同样得不到保证[3,4]。

统计学习理论(Statistic Learning Theory, SLT)[5]的提出,以及由之发展而成的支持向量机(Support Vector Machine, SVM)[6]和支持向量回归(Support Vector Regression, SVR)[7]给数量经济研究者以启示,在为复杂经济系统建模时,有必要考虑降低模型的复杂性程度,提高模型的泛化能力。当然,其前提是必须综合考虑模型的经验误差。

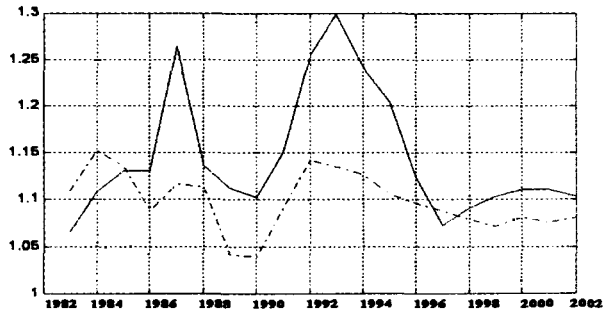

图 1 区域经济与宏观经济的发展波动性比较

事实上,作者在承担江门市科技攻关重点项目《江门市经济预测与决策支持系统》时发现,利用自回归模型对经济发展作中长期预测,非线性模型的准确程度高,而对经济发展作短期预测时,则经过简单变换的线性回归模型往往能取得更好的预测效果。

基于上述的研究背景和区域经济发展具有的特点,作者结合统计学习理论的研究思路,提出了带约束的最优多元回归模型(Optimal Multiple Regression Model with Constraints,简称为 COMR 模型),并将其应用到区域经济发展的短期预测中,取得了一定的效果。

*)资助项目:广东省自然科学基金(032353),国家自然科学基金项目资助(70471074)。肖健华 博士,副教授,主要研究方向为智能信息处理,复杂系统建模;林 健 博士,教授,博士生导师,主要研究方向为复杂系统建模与仿真;刘 晋 博士,教授,主要研究方向为供应链管理。

2 带约束的最优多元回归模型

考虑给定的 n 个已知学习样本(x_i,y_i)，$x_i\in R^d$，$y_i\in R$，$i=1,2,\cdots,n$，在最小二乘法下，线性回归的目标就是求回归函数

$$f(x)=\langle w,x\rangle+b \tag{1}$$

并使参数 w,b 满足

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

式(1)中 $w\in R^d$，$b\in R$，$\langle w,x\rangle$ 为 w 与 x 的内积。式(2)中 $R_{emp}(f)$ 为学习样本集在回归函数 $y=f(x)$ 下的拟合误差，此即所谓的学习误差或经验风险。

机器学习的目的是期望风险最小，即

$$\min R(f)=\int(y-f(x))^2\mathrm{d}F(x,y) \tag{3}$$

式中的联合概率密度分布函数 $F(x,y)$ 为某未知函数。统计学习理论指出[5]：经验风险最小并不能保证期望风险最小。也就是说，建立在经验误差最小化原则上的回归函数的泛化能力得不到保证。结构风险最小化综合考虑了经验风险和置信范围，最终可使期望风险在概率意义下达到最小化。

在 ε 不敏感损失函数下，以结构风险最小化为优化目标，回归函数式(1)应满足[7]

$$\min \Phi(w,\xi,\xi^*)=\frac{1}{2}(w\cdot w)+C\sum_{i=1}^{l}(\xi_i+\xi_i^*) \tag{4}$$

$$s.t.\begin{cases}y_i-\langle w,x_i\rangle-b\leqslant\varepsilon+\xi_i\\-(y_i-\langle w,x_i\rangle-b)\leqslant\varepsilon+\xi_i^*\\\xi_i,\xi_i^*\geqslant 0\end{cases} \tag{5}$$

式中 ξ_i,ξ_i^* 是各学习样本在 ε 不敏感函数下的拟合误差。参数 C 用于折中考虑经验风险和置信范围。

定义 Langrange 函数

$$L=\frac{1}{2}(w\cdot w)+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)-\sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-y_i+\langle w,x_i\rangle+b)-\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+y_i-\langle w,x_i\rangle-b)-\sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*) \tag{6}$$

显然，L 应在 w,b,ξ_i,ξ_i^* 下取最大，分别对这些变量取偏导，有

$$\partial_b L=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0 \tag{7}$$

$$\partial_w L=w-\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i=0 \tag{8}$$

$$\partial_{\xi_i}=C-\alpha_i-\eta_i=0 \tag{9}$$

$$\partial_{\xi_i^*}=C-\alpha_i^*-\eta_i^*=0 \tag{10}$$

进一步考虑到向量 x 中各元素对应区域经济发展中的各个因素，在后面的定性分析中我们将看到，这些因素都对区域经济的发展起积极的推动作用。因此，在区域经济发展预测的特定应用背景下，存在

$$\frac{\partial y}{\partial x}=w'>0 \tag{11}$$

即

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i>0 \tag{12}$$

将式(7)～(10)代入式(6)，并结合式(12)，即得到优化问题

$$\max -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle-\varepsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \tag{13}$$

$$s.t.\ \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i>0$$

$$\alpha_i,\alpha_i^*\in[0,C]$$

此即带约束的最优回归模型(COMR 模型)。

由式(13)可将参数 α_i,α_i^* 解出。实际上，根据 KKT 条件，只有一部分 α_i,α_i^* 不等于 0，与之对应的学习样本称为支持向量。

进一步，由式(8)有

$$w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i \tag{14}$$

最后，由式(1)和式(14)可得到回归方程

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(x_i\cdot x)+b \tag{15}$$

上式中的参数的计算方法如下[6]

$$b=-\frac{1}{2}w\cdot[x_r+x_s] \tag{16}$$

3 基于 COMR 模型的区域经济发展预测：以江门市为例

3.1 影响区域经济发展的因素分析

一般认为，对于区域经济而言，影响其经济增长(通常以 GDP 的变化来衡量)的因素主要包括四个方面：外部环境、投资、消费与净出口，此外，人口受教育程度和金融机构贷款总额也有可能对经济的发展产生较大的影响，如图 2 所示。

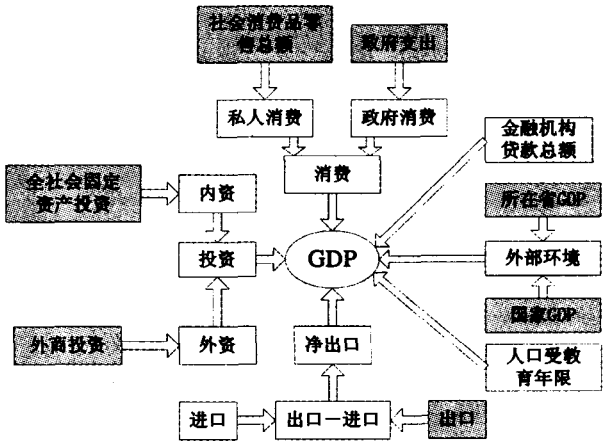

图 2 影响 GDP 增长的因素

其中，人口受教育年限反映了一个城市的人员素质，显然，人员素质的优劣对一个城市的发展是极其重要的，然而，该数据只能在人口普查中获得，因此，没有也不可能有连续的数据，只好舍弃；同样由于其他的原因，进口和金融机构贷款总额的细节数据也无法获得。这样，我们只能考虑用图 2 中标有阴影的 7 个指标来衡量 GDP 的增长：所在省 GDP 增长情况、国家 GDP 增长情况、政府支出、社会消费品零售总额、全社会固定资产投资总额，外商投资总额以及出口。当然，考虑到经济发展的延续性，该区域往年的 GDP 增长数据也是必须考虑的。

将这些指标相对应的增长率与 GDP 的增长率进行相关分析，所得结果如表 1 所示。

表 1　第 t 年 GDP 与第 t−k 年经济指标的相关系数

与第 t 年 GDP 的相关系数		第 $t-k$ 年经济指标数据				
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
经济指标名称	固定资产投资	0.5889	0.5483	0.1775	−0.3132	−0.3850
	实际利用外资	0.7664	0.2519	−0.2602	−0.3822	−0.1750
	外贸出口总额	0.6514	0.2252	0.0703	−0.1630	−0.1807
	所在省 GDP	0.6808	0.4247	−0.0526	−0.4767	−0.2602
	国家 GDP	0.3567	0.0645	−0.2535	−0.5098	−0.4574
	自身 GDP	0.6038	0.1725	−0.2267	−0.4340	−0.1548
	财政支出	0.5596	0.2034	−0.1523	−0.2837	−0.2487
	社会消费品零售总额	0.5010	−0.0317	−0.2032	−0.5375	−0.2395

从表 1 可得出如下结论：固定资产投资对 GDP 的稳定增长综合贡献最大，且可在两年内产生效益，其余的各项指标对 GDP 增长的贡献则主要集中在一年；引进外资和外贸出口对 GDP 的增长作用最直接；就环境而言，GDP 与广东省的 GDP 相关性最大，与国家 GDP 增长关系最小。

以相关系数大于 0.5 为界，对影响 GDP 增长的指标进行预测，最终选取如下 8 个指标作为区域经济发展的预测模型的输入：江门市 GDP($t-1$)、广东省 GDP($t-1$)、外贸出口总额($t-1$)、财政支出($t-1$)、社会消费品零售总额($t-1$)、固定资产投资($t-1$)、固定资产投资($t-2$)以及实际利用外资($t-1$)。

3.2　基于 COMR 模型的江门市经济发展预测

考虑到经济系统的动态性，为了体现对近期数据的重视，首先对样本数据进行等比加权操作，对靠近预测年份的样本赋予较大的权值。

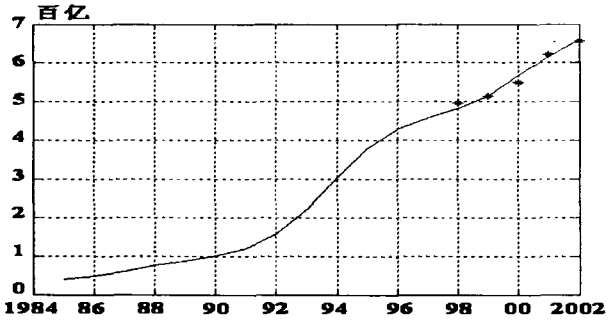

图 3　COMR 预测值与实际值的比较

为了说明方法的有效性，作者采用 1985～1997 年的数据作为学习样本，1998～2002 年的数据为测试样本，采用式(13)所示的 COMR 模型进行检验，所得结果如图 3 和表 2 所示。

图 3 中实线为各年度的 GDP 实际值，“＊”对应相关年份的 COMR 模型预测值。

表 2　预测 GDP 与实际 GDP 的比较

年份	预测 GDP	实际 GDP	误差%
1998	4963300	4818800	2.9990
1999	5139100	5146900	−0.1527
2000	5495800	5675100	−3.1604
2001	6220700	6151600	1.1229
2002	6565300	6608200	−0.6500

图 3 和表 2 表明，考虑到实际经济系统的复杂性，由 CMOR 模型得到的预测结果还是较为理想的。实际上，文献表明，同类研究的预测误差往往在 5%以上。

结论　统计学习理论的提出，为复杂经济数据的处理提供了全新的研究途径。本文作者结合区域经济发展的特点，将统计学习理论与线性回归模型相结合，提出了带约束的多元回归模型。由于该模型是建立在统计学习理论的基础上，因而具有更好的泛化能力。同时，该模型有机地结合了经济发展预测的特点，也就保证了该模型能获得比一般计量经济学模型更好的预测精度。

参 考 文 献

1　Hendry D, Ericsson N. Understanding Economic Forecasts. MIT Press, 2001

2　邓宏钟，迟妍，谭跃进．经济系统中的非线性建模与仿真．计算机工程与应用，2001，37(18)：7～9

3　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．系统工程学报，2000，15(2)：202～207

4　邵惠鹤．支持向量机理论及其应用．见：自动化博览二十周年纪念文集，2003．90～95

5　Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

6　Gunn S. Support Vector Machines for Classification and Regression: [Technical Report]. University of Southamptom, 1998

7　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

(上接第 159 页)

5　Lakshmanan K B, Rosenkrantz DJ, Ravi S S. Alarm placement in systems with fault propagation. Theoretical Computer Science, 2000, 243: 269～288

6　De Bontridder K M J, Halldorsson B V. Halldorsson M M, et al. Approximation algorithm for the minimum test set problem: [Alcom-FT Technical Report Series ALCOMFT-TR-02-80]. Available at: http://www.brics.dk/ALCOM-FT/, 2002.

7　Halldorsson B V. Algorithms for Biological Sequence Problems: [PhD thesis]. Department of Mathematical Sciences. Carnegie Mellon University, USA, 2001

8　Chvatal V. A greedy heuristic for the set covering problem. Mathematics of Operations Research, 1979, 4: 233～235

9　Srivasan A. Improved approximations of packing and covering problems. In: Proc. 27th ACM Symposium on the Theory of Computing, 1995. 268～276

10　Lovasz L. On the ratio of optimal integral and fractional covers. Discrete Mathematics, 1975, 13: 383～390