参数生产前沿面分析的单边支持向量回归模型

彭敏晶[1,2],林 健[3]
PENG Min-jing[1,2],LIN Jian[3]

1.华南理工大学 工商管理学院,广州 510641
2.五邑大学 系统科学与技术研究所,广东 江门 529020
3.北京航空航天大学 经济管理学院,北京 100083
1.School of Business Administration,South China University of Technology,Guangzhou 510641,China
2.Institute of Systems Science and Technology,Wuyi University,Jiangmen,Guangdong 529020,China
3.School of Economics and Management,Beihang University,Beijing 100083,China
E-mail:reggiepeng@163.com

PENG Min-jing,LIN Jian.Single-side support vector regression model for parametric production frontier analysis.Computer Engineering and Applications,2008,44(22):43-45.

Abstract: In order to solve the problem that it is difficult to select the production function in parametric production frontier analysis,a single-side support vector regression model for parametric production frontier is proposed.In the proposed kernel method based model,the input resources data of all decision making units in data space is mapped into feature space,thus the data can be linearly used.In this way,the nonlinear mapping of linear production function is employed to solve the problem of selecting a production function.At last,an experiment of evaluating economic efficiencies of cities in Pearl River Delta is conducted to verify the proposed model.

Key words: production frontier;kernel methods;regression;decision making unit;parametric model

摘 要:为了解决了现有参数生产前沿面分析中先验生产函数难以选择的问题,提出参数生产前沿面分析的单边支持向量回归模型.该模型通过引入核方法,采用非线性映射将各生产决策单元的资源投入原始数据由数据空间映射到特征空间,然后在特征空间进行对应的线性操作。这样,则可以通过线性生产函数的非线性映射来解决生产函数的选择问题。最后,通过对珠三角各城市的经济发展效率进行评价,证明了该模型的有效性。

关键词:生产前沿面;核方法;回归;生产决策单元;参数模型
DOI:10.3778/j.issn.1002-8331.2008.22.012 **文章编号**:1002-8331(2008)22-0043-03 **文献标识码**:A **中图分类号**:TP311

1 引言

对经济生产有效性进行评价具有非常重要的意义。要了解生产行为无效的根源及程度,以提出相应的改进对策和目标,就必须要对生产决策单元(Decision Making Unit,DMU)的生产有效性进行评价。生产前沿面分析是生产有效性评价的重要工具。它的研究思路是:根据已知的一组投入产出观察值,构造出投入产出一切可能组合的外部边界,使得所有投入产出观测点都落在这个边界的下方并且与其尽可能接近[1]。

生产前沿面研究始于1957年经济学家Michael Farrell的生产效率测度研究工作[2],其后产生了丰富的研究方法。其中的一类主要方法是参数模型,它采用的是生产函数估计思想:首先根据需要构造一种具体的生产函数形式,然后通过适当的方法估计位于生产前沿面上的函数参数,从而确定生产前沿面的前沿生产函数。参数方法中的典型模型是Aigner-Chu模型,这种模型的一个最主要的缺点是要求有一定形式的先验生产函数[3]。而事实上,生产函数形式对于不同类型的生产单元可能不相同,这样的前提容易导致所得到生产前沿面并不适合于实际评价的生产单元,进而导致评价不准确。

为了克服以上参数生产前沿面模型中存在的需要确定生产函数形式的问题,本研究基于核方法,提出了参数生产前沿面分析的单边支持向量回归算法。所提出的方法通过采用数据空间到特征空间的映射核函数来解决生产函数形式的确定问题。

2 基于核方法的生产前沿函数构造理论

2.1 参数生产前沿面分析

考虑给定的l个生产决策单元样本(X_i,y_i),$X_i\in R^d$,$y_i\in R$,

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70471074);广东省科技计划项目(the Science and Technology Planning Foundation of Guangdong Province under Grant No.2004B36001051)。

作者简介:彭敏晶(1974-),男,博士生,讲师,主要研究领域为机器学习、管理系统仿真;林健(1958-),男,英国兰卡斯特大学博士,教授,主要研究领域为管理系统仿真。

收稿日期:2008-04-30 **修回日期**:2008-05-29

$i=1,2,\cdots,l$,生产决策单元 i 有 d 个投入 X_i、单产出 y_i。前沿生产函数 y^0 是在所给样本点数据基础上获得的理想产出,因此,$y_i^0-y_i\geqslant 0$,则各个样本点的投入产出效率可用 y_i/y_i^0 表示。在单投入单产出的情况下,生产前沿面 y^0 表示为图 1。

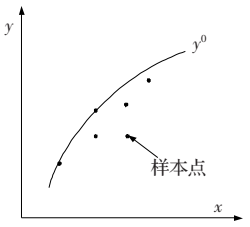

图 1　单投入单产出生产前沿面示意图

在给定数据样本点的情况下,剩下的工作是确定生产前沿函数,具体包括两方面的工作:选择生产函数形式和函数参数估计。不同的函数形式影响效率值。生产前沿面分析中使用的生产前沿函数形式主要有:(1)线性函数;(2)Cobb-Douglas 函数;(3)Log 函数。社会经济系统是一个复杂系统,处于其中的生产决策单元的生产效率受着多种因素的交互影响,采用简单函数来拟合生产前沿面的效果并不理想。对于实际的生产有效性评价问题,由于其生产函数的形式是未知、复杂的,因此,引入非线性处理能力非常强的核方法。

2.2　核方法

目前,核方法(Kernel Methods)在回归、分类、降维等数据处理方面得到了成功的应用[5-9],它是由 1998 年 V. Vapnik 所提出的统计学习理论(Statistical Learning Theory,SLT)发展而成的一类数据处理方法[10],这些方法的共同特征是它们都应用了核映射。

从具体操作过程上看,核方法首先采用非线性映射将原始数据由数据空间映射到特征空间,进而在特征空间进行对应的线性操作,如图 2 所示。由于采用了非线性映射,且这种映射往往是非常复杂的,从而大大增强了非线性数据处理能力。

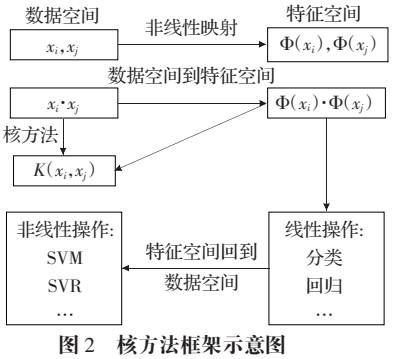

图 2　核方法框架示意图

从本质上讲,核方法实现了数据空间和特征空间之间的非线性变换。设 x_1 和 x_2 为数据空间中的样本点,数据空间到特征空间的映射函数为 Φ,核方法的基础是实现向量的内积变换

$$(x_1,x_2)\rightarrow K(x_1,x_2)=\Phi(x_1)\cdot\Phi(x_2) \tag{1}$$

通常,非线性变换函数 $\Phi(\cdot)$ 相当复杂,而运算过程中实际用到的核函数 $K(\cdot,\cdot)$ 则简单得多,这正是核方法最吸引人的地方。

式(1)中核函数必须满足 Mercer 条件:对于任意给定的对称函数 $K(x_1,x_2)$,它是某个特征空间中的内运算的充分必要条件是对于任意的不恒为 0 的函数 $g(x)$,且 $\int g(x)^2<\infty$,有

$$\int K(x,y)g(x)g(x)\mathrm{d}x\mathrm{d}y\geqslant 0 \tag{2}$$

式(2)给出了一个函数成为核函数的充要条件。考虑到核方法的基础是实现一种由输入空间到特征空间的非线性映射,假设输入空间数据

$$x_i=R^d(i=1,2,\cdots,N) \tag{3}$$

对任意对称、连续且满足 Mercer 条件的函数 $K(x_1,x_2)$,存在一个 Hilbert 空间 H,对映射 $\Phi:R^d\rightarrow H$ 有

$$K(x_i,x_j)=\sum_{n=1}^{d}\Phi_n(x_i)\Phi_n(x_j) \tag{4}$$

式中,d 是 H 空间的维数。

上式进一步说明,输入空间的核函数实际上与特征空间的内积等价。由于在核方法的各种实际应用中,只需要应用特征空间的内积,而不需要了解映射 Φ 的具体形式。换句话说,在使用核方法时,只需要考虑如何选定一个适当的函数,而无需关心与之对应的映射 Φ 可能具有复杂的表达式和很高的维数。

3　基于核方法的前沿生产函数构造

考虑核方法,假定在特征空间下,生产前沿面的线性函数形式如下:

$$y^0=W\cdot X \tag{5}$$

生产前沿面上的理想产出在等于或高过实际产出的同时,与理想的实际产出点有较好的拟合,如图 3 所示。

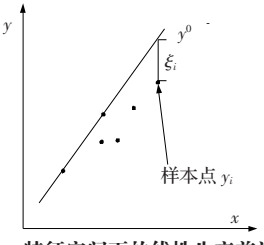

图 3　特征空间下的线性生产前沿面

图中,理想产出与实际产出之差为:

$$\xi_i=y_i^0-y_i=W\cdot X_i-y_i \tag{6}$$

为了使生产前沿面与效率理想化的实际前沿产出有较好的拟合,采用传统的经验风险最小化理论,可得到下面的生产前沿面优化模型:

$$\begin{aligned}&\min\ \sum_{i=1}^{l}\xi_i\\&\text{s.t. }\ y_i-W\cdot X_i=\xi_i\\&\xi_i\geqslant 0\end{aligned} \tag{7}$$

上式中 l 为样本个数。

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[10]。因此,应当限定生产前沿面回归方程的复杂度,所以对 W 作如下限定。

$$\|W\|\leqslant B \tag{8}$$

式(7)和式(8)表明,所提出的生产前沿面分析模型能够折中考虑生产前沿函数经验风险和泛化能力,因此,该生产前沿函数将具有更好的性能。

引入参数 $\beta_i,\lambda_i,\eta_i\geqslant 0$,构造 Lagrange 函数对上述优化模型

求解

$$L:=\sum_{i=1}^{l}\xi_i+\sum_{i=1}^{l}\beta_i(y_i-W\cdot X_i-\xi_i)+\lambda(\|W\|^2-B^2)-\sum_{i=1}^{l}\eta_i\xi_i \quad (9)$$

考虑到上式关于 W,ξ 值取极小,因此分别对 L 关于 W,ξ 求偏导,并令它们等于 0,得到:

$$\frac{\partial L}{\partial W}=0\rightarrow\sum_{i=1}^{l}\beta_i X_i+2\lambda W=0 \quad (10)$$

$$\frac{\partial L}{\partial \xi}=0\rightarrow 1-\sum_{i=1}^{l}(\beta_i+\eta_i)=0 \quad (11)$$

将式(10)和式(11)代入式(9),得到对偶优化问题

$$\min Q=\sum_{i=1}^{l}\beta_i y_i-\frac{1}{4\lambda}\sum_{i,j=1}^{l}\beta_i\beta_j(X_i,X_j)-\lambda B \quad (12)$$

由上述优化方程,可看出,等式右边最后一项 λB 对求极值没有影响。同时,作如下假设。令

$$\alpha_i=\frac{\beta_i}{2\lambda} \quad (13)$$

为对偶系数,则式(12)化为如下优化问题:

$$\min Q=2\lambda\sum_{i=1}^{l}\alpha_i y_i-\lambda\sum_{i,j=1}^{l}\alpha_i\alpha_j(X_i,X_j) \quad (14)$$

把上式中的 λ 消除,则得到

$$\min Q=2\sum_{i=1}^{l}\alpha_i y_i-\sum_{i,j=1}^{l}\alpha_i\alpha_j(X_i,X_j) \quad (15)$$

由上述优化方程,可求出各 $\alpha_i,i=1,2,\cdots,l$。实际上,只有一部分 $\alpha_i\neq 0$,与之对应的样本 (X_i,y_i) 为支持向量(Support Vector,SV)。支持向量即为效率最高的样本点。

进一步由式(13)和式(10),可得到线性特征空间下的最优前沿面方程。

$$y_j^0=\sum_{i=1}^{l}\alpha_i(X_i,X_j) \quad (16)$$

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,并使低维空间的非线性问题转化为高维空间的线性问题。在核函数下,式(14)和式(16)化为如下形式

$$\min Q=2\lambda\sum_{i=1}^{l}\alpha_i y_i-\lambda\sum_{i,j=1}^{l}\alpha_i\alpha_j K(X_i,X_j) \quad (17)$$

$$y_j^0=\sum_{i=1}^{l}\alpha_i K(X_i,X_j) \quad (18)$$

对式(17)进行求解可得各 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。

4　区域经济发展效率评价:以广东省珠三角地级城市为例

广东省江门市地处珠三角的西部,人口约 400 万,土地接近 10 000 平方公里,人均 GDP 接近 2 000 美元,国内外经验表明,该市经济发展正处于高速增长期,对经济发展效率进行客观的评价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,江门市政府对于江门市经济发展的效率评价十分重视,专门委托本项目组进行江门市经济发展的效率评价的研究工作,并把它纳入到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"项目中。本文正是相关研究工作的部分成果。

4.1　投入与产出指标的确定

借鉴参考文献[12]的思路,本文针对当年的各市本地生产总值这个产出,将投入资源分为 3 大类:(1)人力资源;(2)经济资源;(3)环境资源。根据相关性分析,确定对应区域经济系统产出当年生产总值的投入性指标包括:人口[13]、上年 GDP[14]、上年固定资产投资[14]、利用外资[14]、能源消耗[15]等 5 个指标。

为了使经济发展效率具有可比性,选择了整体环境相差不大的珠三角其他地级城市的经济指标来进行比较。各城市的指标如表 1 所示。

表 1　广东省珠三角各城市的指标值

	产出	人力资源	经济资源			环境资源
	当年 GDP/亿元	人口/万人	上年 GDP/亿元	上年固定资产投资/亿元	利用外资/亿元	能源消耗/吨标准煤
广州	5 115.75	949.68	4 450.55	1 312.71	266.68	7 965
深圳	4 926.90	827.75	4 282.14	1 090.12	1 015.22	1 139
珠海	634.58	141.57	331.43	179.85	107.71	901
东莞	2 182.44	656.07	1 806.03	433.90	409.29	2 070
佛山	2 379.80	580.03	1 918.04	568.56	170.80	2 806
中山	817.56	243.46	704.30	291.58	122.55	1 953
惠州	803.94	370.69	686.45	297.62	106.55	1 080
江门	802.16	410.29	695.64	195.69	60.26	1 552
肇庆	453.55	367.60	390.56	145.24	14.16	1 764

4.2　计算步骤

利用 Matlab7.1,通过以下步骤计算得到各城市的经济发展效率:

(1)基于式(16),利用 Matlab 二次优化工具计算出各 α_i 的值;

(2)通过式(17)计算得到各个城市对应的 y_j^0;

(3)把各城市实际 GDP 与所计算得到的 y_j^0 相比较,得到经济发展效率评价值。

4.3　参数确定

核方法应用研究的最大难点在于其参数的确定,其中包括核函数的形式及函数相关参数。

本研究中,核函数采用高斯径向基函数

$$K(X_i,X_j)=\exp\left(-\frac{|X_i-X_j|^2}{8}\right) \quad (19)$$

4.4　计算结果与讨论

计算得到的各城市经济发展效率评价值在表 2 中给出。

表 2　珠三角各地级市经济发展效率评价值

广州	深圳	珠海	东莞	佛山	中山	惠州	江门	肇庆
0.84	0.81	0.61	1.00	1.00	0.88	0.84	0.88	0.52

从表 2 中可以看出,东莞、佛山的经济发展效率为理想状态,而肇庆、珠海的经济发展效率偏低,其原因是因为资源的利用效率不高。江门的经济发展效率为中等,这与实际情况相符。

5　结束语

本研究所提出的参数生产前沿面分析的单边支持向量回归模型通过引入核方法,采用非线性映射将 DMU 的资源投入原始数据由数据空间映射到特征空间,然后在特征空间进行对应的线性操作,解决了现有参数生产前沿面分析需要选定一个先验生产函数的问题。

(下转 80 页)