第32卷 第20期 计 算 机 工 程 2006年10月
Vol.32 № 20 Computer Engineering October 2006
·博士论文· 文章编号：1000—3428(2006)20—0007—03 文献标识码：A 中图分类号：TP18

# 基于支持向量数据描述的预警技术及其应用

林 健[1,2]，彭敏晶[1,2]
(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，江门 529020)

**摘 要：**分析当前的主要预警方法，指出由于缺少非正常数据样本，使得现有的大部分预警方法不适用。为解决该问题，提出了基于核方法的支持向量数据描述预警技术。建立了一个用于检测非正常数据对象的一类分类器，检测数据对象是否在正常值超球体范围内。如果在超球体外，预警专家将最终确认这个数据对象是否为非正常的预警警兆。以广东省江门市的宏观区域经济数据为例，证明了该预警技术的有效性。

**关键词：**预警；支持向量数据描述；核方法；数据样本

# SVDD Based Early Warning Technique and Its Application

LIN Jian[1,2], PENG Minjing[1,2]
(1. School of Business Administration, South China University of Technology, Guangzhou 510641;
2. Institute of Systems Science & Technology, Wuyi University, Jiangmen 529020)

**【Abstract】** After reviewing the current early warning researches, this paper presents that most of current early warning methods are unsuitable because of lacking a historical "ill-represented" dataset. And then the support vector data description early warning technique based on kernel method is proposed to solve the problem. A one-class classifier is fitted to detect the "ill-represented" data objects by enclosing all "good" data objects in a hypersphere. If an object is outside the boundary of the hypersphere, an early warning expert would be prompted to decide whether the object is enough "ill-represented" for issuing a warning. An early warning experiment based on the macro-economic dataset of Jiangmen, Guangdong is conducted to verify the proposed technique.

**【Key words】** Early warning; Support vector data description (SVDD); Kernel method; Data feature

预警是一项重要的工作，它是指围绕特定对象的循环波动这一特定现象所展开的一整套监测和评价的理论和方法体系[1]。目前的研究方法主要有：ARCH预警法[2]、基于概率模式分类法[3]、判别分析法[4]、人工神经网络法[4,5]和基于支持向量机[6]的方法。以上的方法都要求有一定数量的表现预警对象异常波动的预警征兆的非正常数据样本。然而，对动态的、非线性的复杂大系统进行预警工作时，由于预警对象的发展、环境的快速改变及异常出现的概率相对较低等原因[7,8]，收集非正常数据是一件相当困难的工作，这对现有的预警方法来说是一个很大的局限。本文提出的技术通过支持向量数据描述(Support Vector Data Description, SVDD)实现了在没有非正常数据样本集的情况下进行预警工作的目标，同时也解决了收集非正常数据样本的问题。最后，以广东省江门市的宏观区域经济数据为例，证明该预警技术的有效性。

## 1 支持向量数据描述预警技术

本技术把预警工作分为两个步骤，预警步骤见图1。

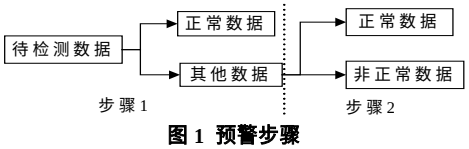



**图1 预警步骤**

(1)区别正常数据和其他数据。建立了一个基于支持向量数据描述的一类分类器[9]，把能确认的所有的正常数据对象包含于一个超球体内，以区别于超球体外的其他数据；(2)确认超球体外其他数据中的非正常数据。球体外的其他数据由经济预警专家来判断是否为非正常数据以发布经济异常警告。

### 1.1 支持向量数据描述

Vapnik的统计学习理论认为[9]，如果样本的数据量比较小时，采用直接寻找封闭区域的方式来分类比估计概率密度的方法更为有效，这种直接寻找封闭区域的一类分类器被称之为数据描述。

对于包含 $n$ 个“正常”数据对象的样本集：

$$\{x_1, x_2, \cdots, x_n\} \tag{1}$$

所形成的超球体由中心点 $a$ 和半径 $R$ 描述。要形成紧凑的球体边界，数据描述的优化问题可以表述为

$$\min \quad L(R) = R^2 \tag{2}$$

$$s.t. \quad R^2 - (x_i - a)(x_i - a)^T \geq 0 \tag{3}$$

根据式(2)和式(3)，定义如下的 Lagrange 函数：

$$L(R, a, \wedge) = R^2 - \sum_{i=1}^{n} \alpha_i \{R^2 - (x_i - a)(x_i - a)\} \tag{4}$$

其中，Lagrange 系数 $\alpha_i \in \wedge$，且 $\alpha_i \geq 0$。把式(4)对 $R$ 求偏微分，并令其等于0，得到

**基金项目：**国家自然科学基金资助项目(70471074)；广东省科技厅计划基金资助项目(2004B36001051)
**作者简介：**林 健(1958－)，男，教授、博导、博士后，主研方向：复杂管理系统仿真；彭敏晶，博士生
**收稿日期：**2006-01-15 **E-mail：**reggiepeng@163.com

$$\sum_{i=1}^{n}\alpha_i = 1 \tag{5}$$

把式(4)对 $a$ 求偏微分，并令其等于 0，结合式(5)，得到

$$\sum_{i=1}^{n}\alpha_i x_i = a \tag{6}$$

把式(5)和式(6)代入式(4)，得到优化方程

$$\max \quad L=\sum_{i=1}^{n}\alpha_i\left(x_i\cdot x_i\right)-\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j\left(x_i\cdot x_j\right) \tag{7}$$

$$s.t. \quad \sum_{i=1}^{n}\alpha_i=1 \qquad \alpha_i\geq 0 \tag{8}$$

根据 KKT 条件，∧中大部分 $\alpha_i=0$，少部分 $\alpha_i\geq 0$。与不为 0 的 $\alpha_i$ 对应的样本 $x_i$ 决定了超球体的边界，这些样本数据被称为支持向量(Support Vector, SV)。

对于已知∧，通过式(6)可求出球心 $a$，任选一个支持向量可由下面的式(9)求出 $R$ ：

$$R^2-\left(x_i-a\right)\left(x_i-a\right)=0 \tag{9}$$

对于待检测的数据 $z$，令：

$$f\left(z\right)=\left(z-a\right)\left(z-a\right)^T \tag{10}$$

式(10)也可变换为

$$f\left(z\right)=\left(z\cdot z\right)-2\sum_{i=1}^{n}\alpha_i\left(z\cdot x_i\right)+\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j\left(x_i\cdot x_j\right) \tag{11}$$

考虑到式(11)中，对于待检验的 $z$，

$$\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j\left(x_i\cdot x_j\right)=const$$

令：

$$g\left(z\right)=\left(z\cdot z\right)-2\sum_{i=1}^{n}\alpha_i\left(z\cdot x_i\right) \tag{12}$$

$$\lambda\left(z\right)=\frac{g\left(z\right)}{g\left(x_s\right)} \tag{13}$$

其中，$x_s$ 为任一支持向量。因此，可以依据式(14)来判定 $z$ 是否在超球体外：

$$\lambda\left(z\right)=\begin{cases}>1,\text{z} & \text{在超球体外}\\ \leq 1,\text{z} & \text{在超球体内}\end{cases} \tag{14}$$

可以看出，采用式(12)～式(14)来确定待判定点 $z$ 的状态，省去了计算半径 $R$ 和中心点 $a$ 的过程，简化了计算。

然而，由于优化方程式(7)所确定的边界形状单一，且边界区域所形成的空间过大，不够紧凑，易于把非正常数据纳入到超球体的范围。通过核方法(Kernel Method, KM)把输入空间的数据对象映射到核空间来解决此问题。把内积运算替换为满足 Mercer 条件的核函数 $\left(x_i\cdot x_j\right)\rightarrow K\left(x_i,x_j\right)$，则属性空间的优化方程式(7)变换为

$$\max \quad L=\sum_{i=1}^{n}\alpha_i K\left(x_i,x_i\right)-\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j K\left(x_i,x_j\right) \tag{15}$$

约束条件不变，式(11)变为

$$f\left(z\right)=K\left(z,z\right)-2\sum_{i=1}^{n}\alpha_i K\left(z,x_i\right)+\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j K\left(x_i,x_j\right) \tag{16}$$

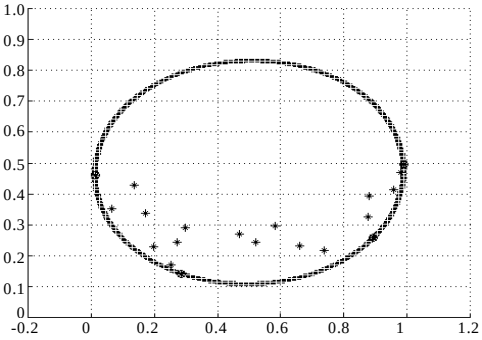
式(12)变为

$$g\left(z\right)=K\left(z,z\right)-2\sum_{i=1}^{n}\alpha_i K\left(z,x_i\right) \tag{17}$$

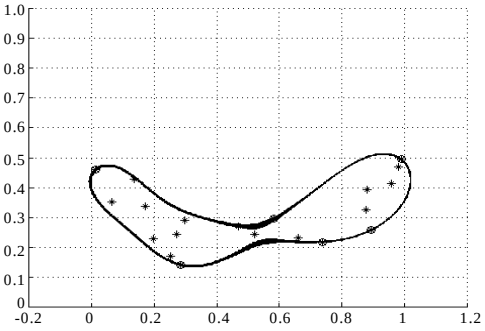
这种基于核方法的直接寻找封闭区域的一类分类方法被称为支持向量数据描述[10]。特别地，选择高斯径向基核函数

$$K\left(x,y\right)=exp\left(-\left\|x-y\right\|^2/\sigma^2\right) \tag{18}$$

其中，$\sigma$ 可事先给出。给定不同的 $\sigma$ 对于同一样本集可能产生不同的超球体边界，如图 2 所示[10]。



(a) 取$\sigma=1$ 时得到的超球体边界



(b) 取 $\sigma=0.3$ 时得到的超球体边界

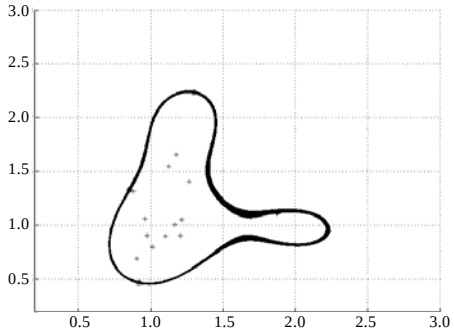
**图 2 $\sigma$ 取值不同时的超球体边界**

引入高斯径向基核函数后，式(15)和式(17)分别简化为

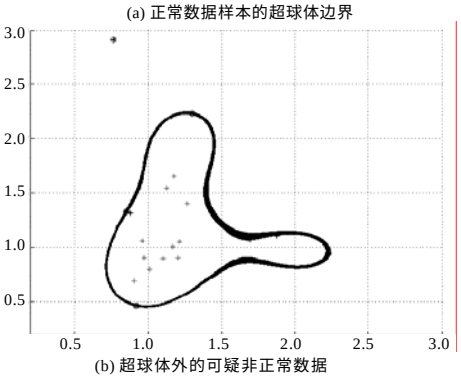
$$\max \quad L=1-\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j K\left(x_i,x_j\right) \tag{19}$$

$$g\left(z\right)=1-2\sum_{i=1}^{n}\alpha_i K\left(z,x_i\right) \tag{20}$$

在引入核方法形成正常数据据集的边界后，如果待确定状态的数据对象 $z$ 的 $g\left(z\right)>g\left(x_s\right)$，即 $\lambda\left(z\right)>1$，则 $z$ 为可疑非正常数据，如图 3。



(a) 正常数据样本的超球体边界



(b) 超球体外的可疑非正常数据

**图 3 正常数据与可疑数据**

### 1.2 技术实现

支持向量数据描述预警技术基于支持向量数据描述，形成包围正常数据样本的超球体边界来判断待确定数据对象是否为其他数据对象，然后再由预警专家通过人机交互确定其他数据对象是否为预警警兆(非正常数据)。具体步骤如图 4。

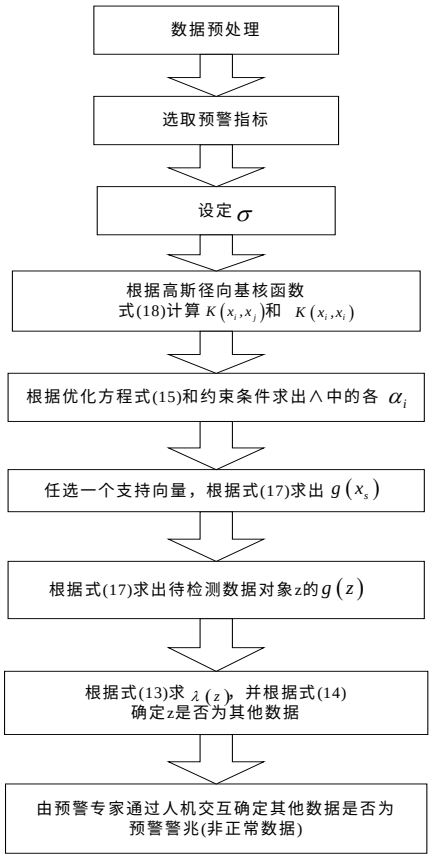



**图 4 预警步骤**

## 2 实例

为了说明支持向量数据描述预警技术的有效性，以广东省江门市的宏观经济数据(1982 年～2002 年)为例进行验证。步骤如下：

(1)考虑到物价因素，同时为使数据具有可比性，根据式(21)对数据进行环比处理。

$$z_t = \frac{y_t / w_t}{y_{t-1} / w_{t-1}} \tag{21}$$

其中，$z_t$ 为 $t$ 年指标的环比；$y_t$ 和 $y_{t-1}$ 分别为 $t$ 年和 $t-1$ 年指标的原始数据；$w_t$ 和 $w_{t-1}$ 分别为 $t$ 年和 $t-1$ 年的累计物价指数。

(2)通过各指标与GDP的互相关性计算，确定出 7 个经济预警相关指标[11]，见表 1。

**表 1 经济预警指标**

| | | | |
|---|---|---|---|
| $x_{1,(t-1)}$ | 上一年度江门 GDP | $x_{2,(t-1)}$ | 上一年度广东省 GDP |
| $x_{3,(t-1)}$ | 上一年外贸出口总额 | $x_{4,(t-1)}$ | 上一年基本建设投资 |
| $x_{4,(t-2)}$ | 上上年基本建设投资 | $x_{5,(t-1)}$ | 上一年实际利用外资 |
| $x_{6,(t-1)}$ | 社会消费品零售总额 | —— | ———— |

(3)选定 $\sigma=1.35$，并事先由经济预警专家确定好对应于 1983 年～1991 年的数据为正常数据对象。然后根据支持向量数据描述预警技术的 7 个步骤计算出对应于 1992 年～2001 年的 $g(z)$ 和 $\lambda(z)$，见表 2。

**表 2 广东省江门市的宏观经济数据**

| 预警年份 (t-1) | $x_{1,(t-1)}$ | $x_{2,(t-1)}$ | $x_{3,(t-1)}$ | $x_{4,(t-1)}$ | $x_{4,(t-2)}$ | $x_{5,(t-1)}$ | $x_{6,(t-1)}$ | $g(z)$ | $\lambda(z)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1983 | 1.099495 | 1.074073 | 1.206206 | 0.903826 | 0.904305 | 1.048000 | 1.003346 | —— | —— |
| 1984 | 1.146769 | 1.219648 | 1.102055 | 0.904305 | 2.910949 | 0.875071 | 1.006119 | —— | —— |
| 1985 | 1.009825 | 1.086892 | 0.759255 | 2.910949 | 0.953558 | 1.266631 | 1.009335 | —— | —— |
| 1986 | 1.159797 | 1.111670 | 2.237508 | 0.953558 | 0.906863 | 2.642502 | 1.008044 | —— | —— |
| 1987 | 1.152515 | 1.144682 | 0.976031 | 0.906863 | 0.689584 | 1.462590 | 1.011566 | —— | —— |
| 1988 | 0.964077 | 1.043557 | 0.901326 | 0.689584 | 0.467928 | 1.024478 | 1.012024 | —— | —— |
| 1989 | 0.937246 | 0.971264 | 0.918675 | 0.467928 | 1.072776 | 0.773341 | 1.011327 | —— | —— |
| 1990 | 1.189334 | 1.169529 | 1.696453 | 1.072776 | 1.187847 | 1.072598 | 1.018243 | —— | —— |
| 1991 | 1.170979 | 1.203577 | 1.527466 | 1.187847 | 1.408381 | 1.790479 | 1.010744 | —— | —— |
| 1992 | 1.243361 | 1.217267 | 1.264464 | 1.408381 | 2.236393 | 1.381315 | 1.011051 | 0.52634 | 1.13710 |
| 1993 | 1.197510 | 1.190295 | 1.296507 | 2.236393 | 1.106619 | 1.703090 | 1.012539 | 0.53101 | 1.14710 |
| 1994 | 1.152661 | 1.102251 | 1.871951 | 1.106619 | 0.797516 | 1.954227 | 1.010137 | 0.43835 | 0.94697 |
| 1995 | 1.128362 | 1.149931 | 1.009599 | 0.797516 | 1.062072 | 1.032458 | 1.008435 | 0.39757 | 0.85888 |
| 1996 | 1.085521 | 1.091107 | 0.959693 | 1.062072 | 1.006108 | 0.962987 | 1.006858 | 0.40325 | 0.87115 |
| 1997 | 1.066950 | 1.122159 | 1.168984 | 1.006108 | 1.549459 | 0.781769 | 1.007239 | 0.43226 | 0.93381 |
| 1998 | 1.071727 | 1.102353 | 1.123718 | 1.549459 | 1.330578 | 1.180231 | 1.004853 | 0.43687 | 0.94377 |
| 1999 | 1.116095 | 1.116870 | 0.85231 | 1.330578 | 1.053136 | 1.214111 | 0.999842 | 0.42898 | 0.92673 |
| 2000 | 1.119409 | 1.15891 | 1.215446 | 1.053136 | 1.317325 | 1.178394 | 1.005306 | 0.39408 | 0.85134 |
| 2001 | 1.107211 | 1.125631 | 0.879345 | 1.317325 | 1.656475 | 1.083229 | 0.999711 | 0.45675 | 0.98671 |

由表 2 中的数据可以看出，1993 年～1994 年将会有较大的经济波动。通过图 5 看实际的 GDP 增长率变化，表明了 1993 年～1994 年的经济波动，从而验证了本技术的有效性。

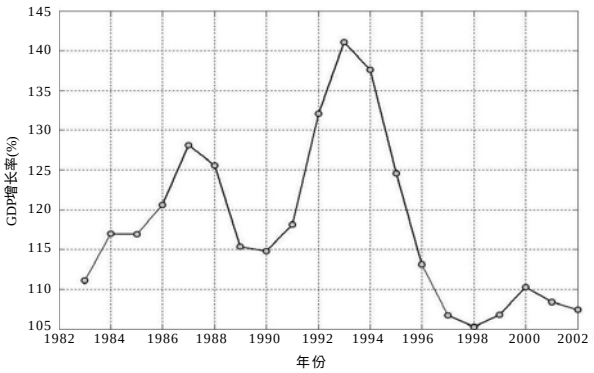



**图 5 GDP 增长率变化**

## 3 结论

支持向量数据描述预警技术的意义在于：(1)解决了没有非正常数据样本集情况下的预警问题；(2)采用此技术预警可以生成分类数据，在一段时间的使用后，将会累积形成非正常样本数据集。

### 参考文献


1 黄继鸿, 雷战波, 凌　超. 经济预警方法研究综述[J]. 系统工程, 2003, 21(2): 64-70.

2 王慧敏. ARCH 预警系统的研究[J]. 预测, 1998, 18(4): 55-56.

3 王建成, 王　静, 胡上序. 基于概率模式分类识别方法的宏观经济预警系统设计[J]. 系统工程理论与实践, 1998, 18(8): 6-10.

4 Altman E. Financial Ratio, Discriminant Analysis and the Predication of Corporate Bankrupt[J]. Journal of Financial, 1958, 123(24)：589-609.

5 王建成, 高大启. 改进的遗传和 BP 杂交算法及神经网络经济预警系统设计[J]. 系统工程理论与实践, 1998, 18(4): 136-141.


(下转第 12 页)