粒子群算法与主成分析法在支持向量机回归预测中的应用研究*

周方军,吕文元

(上海理工大学 管理学院,上海 200093)

摘 要:提出在支持向量机回归预测中采用粒子群算法优化参数和主成分析降维的方法,通过算例分析表明,此法能够显著提高预测的精度。

关键词:支持向量机;粒子群算法;主成分析法;预测

中图分类号:TP181 **文献标识码**:A **文章编号**:1674-7720(2010)23-0080-03

The application research of support vector machine regression based on particle swarm optimization and principal component analysis

ZHOU Fang-Jun, LV Wen-Yuan

(School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: This article proposed a method which apply support vector machine regression based on particle swarm optimization (PSO) and principal component analysis (PCA). The result indicated this method can obviously enhance the precision of prediction.

Key words: SVM; PSO; PCA; prediction

预测是国家、企业等组织制定政策和计划的主要依据,因而预测的准确度是政策与计划制定是否科学的前提。预测的方法有传统的多元回归预测,以及近几年来发展起来的人工神经网络预测[1]、灰色预测[2]。多元回归预测模型简单、易用性强,但难以处理高维、非线性模式;人工神经网络虽然能够较好地解决高维非线性预测的难题,但它需要大量的训练样本,且泛化能力不强,所以当可得到的预测样本是小样本,或者获得大量样本的成本很高时,就难免影响其实用性和经济性;灰色预测虽具有短期预测能力强,可检验等优点,但其长期预测能力较差。Vapnik 等人提出的支持向量机[3-4]是在统计学习理论基础上发展起来的一种新的机器学习算法,是目前针对小样本统计和预测学习的最佳理论,支持向量机具有完美的数学形式、直观的几何解释和良好的泛化性能,解决了模型选择与欠学习、过学习及非线性等问题,克服了收敛速度慢,易陷入局部最优解等缺点,因此支持向量机在分类和回归中均表现出优越的性能。

* 基金项目:国家自然科学基金项目(70301002);国家自然科学基金项目(50875168);高等学校博士学科点专项科研基金资助课题(3309303004)

1 支持向量回归机的基本原理

支持向量回归机[5],主要由 Vapnik 提出的 ε-支持向量回归机(ε-SVR)和 Scholkopf 等提出 ν-的支持向量回归机(ν-SVR)等。本文采用 Vapnik 的 ε-SVR 支持向量回归机。

支持向量机回归实质是要在 R^n 空间寻找一个超平面函数 $y=\boldsymbol{w}^{\mathrm{T}}\cdot x+b$,并使得该超平面与各样本点的偏离最小,其中 $\boldsymbol{w}$ 是超平面 $n-1$ 维法向量。考虑一个样本集 $T=\{(x_1^{\mathrm{T}},y_1),\cdots(x_l^{\mathrm{T}},y_l)\}\in(X\times Y)$,$l$ 为样本数,$\boldsymbol{x}_i$ 是 n 维向量,$y\in R^n$。如果采用 ε 不灵敏函数作为误差函数,当所有的样本点到所求的超平面的距离都不超过 ε 时,如图 1 所示,中间的实线表示 ε 的超平面,超平面两边的 ε 区域为超平面的 ε 带。

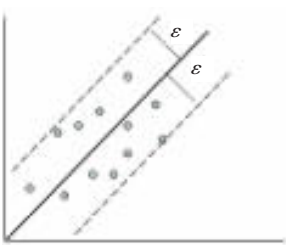

图 1 ε 带超平面

可以想象，一个最优的超平面应该是能够以最小的 ε 带包含训练集中所有样本点的超平面。为求得最优超平面，借鉴支持向量机分类的思想，可将其转化为一个二分类的问题：选择合适的 $\varepsilon(\varepsilon\geqslant\min\varepsilon)$，分别给每个样本点的 y 值加上 ε 或减去 ε，构造新的正负两类样本点：$D^{+}=(\boldsymbol{x}_i^{\mathrm{T}},y_i+\varepsilon;z_i=+1)(i=1,2,\cdots l)$，$D^{-}=(\boldsymbol{x}_i^{\mathrm{T}},y_i-\varepsilon;z_i=-1)(i=1,2,\cdots l)$。考虑会有个别样本点到超平面距离大于 ε 影响求解最优超平面的情况，引入松弛变量 ξ_i,ξ_i^{*} 和惩罚参数 C，构造并求解问题：

$$\min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^{*}) \tag{1}$$

$$s.t.\quad (w\cdot x)+b-y_i\leq\varepsilon+\xi_i,\ i=1,\cdots,l,$$

$$y_i-(w\cdot x)-b\leq\varepsilon+\xi_i^{*},\ i=1,\cdots,l,$$

$$\xi_i,\xi_i\geq 0,\ i=1,2,\cdots,l.$$

引入 Lagrange 乘子，$a^{(*)}=(a_1,a_1^{*}\cdots a_l,a_l^{*})\geq 0$，构造 Lagrange 函数，求式(1)的对偶问题得：

$$\min_{a^{(*)}\in R^{2l}}\ \frac{1}{2}\sum_{i,j=1}^{l}(a_i^{*}+a_i)(a_j^{*}+a_j)K(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(a_i^{*}+a_i)-\sum_{i=1}^{l}y_i(a_i^{*}+a_i) \tag{2}$$

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

$$0\leq a_i,a_i^{*}\leq C,\ i=1,2,\cdots,l.$$

式(2)是一个凸二次规划问题，存在全局最优解 $\bar{a}^{(*)}=(\bar{a}_1,\bar{a}_1^{*}\cdots\bar{a}_l,\bar{a}_l^{*})$，得出最优出超平面函数：

$$f(x)=\sum_{i=1}^{l}(\bar{a}_i^{*}-\bar{a}_i)K(x_i,x)+b^{*} \tag{3}$$

其中 $b^{*}=y_i-\sum_{i=1}^{l}(\bar{a}_j^{*}-\bar{a}_j)K(x_i,x_j)+\varepsilon\quad 0<\bar{a}_i\leq C$ (4)

式(2)~式(4)中 $K(x_i,x_j)$ 是核函数，其值为向量 x_i 和 x_j 在特征空间的 $\varphi(x_i)$ 和 $\varphi(x_j)$ 中的内积，$\varphi(x_i),\varphi(x_j)$ 为映射函数。核函数的作用是当样本点在原空间线性不可分时，可以通过映射函数映射到高维空间，从而达到线性可分的目的，但实际应用中映射函数的显式表达式很难找到，观察式(2)~式(4)中只用到了映射在高维空间的点积，而核函数的特点就是能使变量在低维空间核函数值等于其映射到高维空间的点积值，从而实现不需要知道显式映射函数达到向高维空间映射的目的。任何满足 Mercer 条件的函数均可作为核函数。

2 粒子群算法基本原理

微粒群算法最早是在 1995 年由美国社会心理学家 Kennedy 和 Russell[6]共同提出，其基本思想是受鸟群觅食行为的启发而形成的。PSO 算法把优化问题的解看作是 D 维空间中一个没有体积没有质量的飞行粒子，所有的粒子都有一个被优化目标函数决定的适应度值，而速度决定每个粒子的飞行方向和距离，粒子根据自己先前达到的最优位置和整个群体达到的最优位置来更新自己的位置和速度，从而向全局最优位置聚集。粒子根据以下公式来更新自己的速度和位置：

$$v_{id}(t+1)=wv_{id}(t)+c_1r_1(P_{id}(t)-x_{id}(t))+c_1r_2(P_{gd}(t)-x_{id}(t)) \tag{5}$$

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1)$$

式中，下标 i 代表第 i 个粒子，下标 d 代表速度或位置的第 d 维，t 代表迭代代数，w 代表惯性权重系数，c_1 和 c_2 是学习因子，通常 $c_1,c_2\in[0,4]$，r_1,r_2 是介于 $[0,1]$ 之间的随机数，P_{id} 是粒子 P_i 在第 d 维个体极值坐标，P_{gd} 是粒子群体在第 j 维的全局极值坐标。从式(5)可知，w 越大全局探测能力越强；w 越小则局部探测能力越强。因此可以让 w 随着迭代次数的增加，而动态地减少，以保证算法有较大的机率收敛于全局最优解。但是在算法执行过程中，随着 w 的减少，也在一定程度上导致后期收敛速度降低，从而影响全局收敛性能。为了克服这种缺陷，Clerk 构造了带收缩因子 K 的改进 PSO 模型[7]，试验结果表明收缩因子 K 比惯性权重系数 w 能更有效地控制微粒的飞行速度，同时增强了算法的局部搜索能力，模型如下：

$$v_{id}(t+1)=K(v_{id}(t)+c_1r_1(P_{id}(t)-x_{id}(t))+c_2r_2(P_{gd}(t)-x_{id}(t))) \tag{6}$$

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1)$$

其中，$K=\dfrac{2}{\left|2-C-\sqrt{C^2-4C}\right|}$，$C=c_1+c_2$.

3 主成分析原理

主成分析[8]是利用数学上处理降维的思想，将实际问题中的多个相关性较高的指标设法重新组合成一组新的少数几个互不相关的综合指标来代替原来指标的一种多元统计方法，通常把转化生成的综合指标称为主成份，其中每个主成份都是原始变量的线形组合。主成份要尽可能多地反映原来指标的信息，而且要有较好的解释意义。降维的步骤：(1)将原始数据标准化以消除量纲影响；(2) 计算变量的相关系数矩阵 $\boldsymbol{R}=(r_{ij})_{p'p}$，其中 $r_{ij}(i,j=1,2,\cdots,p)$ 为原来变量 x_i 与 x_j 的相关系数；(3)计算 $\boldsymbol{R}$ 的特征值及相应的特征向量，即 $\lambda_1\geqslant\lambda_2,\cdots\geqslant\lambda_p\geqslant 0$ 然后分别求出对应于特征值 λ_i 的特征向量 $\boldsymbol{a}_i(i=1,2,\cdots,p)$，且 $\boldsymbol{a}_i$ 是正交单位特征向量；(4)写出主成份 $F_i=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_p(i=1,\cdots,p)$。

4 应用实例

试验从 UCI 上选取美国波斯顿地区 1993 年城镇住房数据作为试验数据[9]。试验步骤如下：

(1)应用主成分析法降维

由于统计软件 SPSS 提供了主成份分析功能，而且具有采用交互式、图形化操作界面、结果图形化输出、直观性强等优点，故本文采用 SPSS16.0 作为降维工具，表 1 为最大方差旋转后的因子载荷图，从表中可以看出，7

表 1 主成份载荷旋转

	主成份						
	1	2	3	4	5	6	7
犯罪率(CRIM)	-0.008	0.961	-0.129	0.170	-0.005	0.035	-0.027
常住人口比例(ZN)	0.006	-0.135	0.959	-0.164	0.000	0.112	0.017
无零售商业比例(INDUS)	-0.302	0.471	-0.320	0.228	0.009	0.698	-0.036
有无环城河(CHAS)	0.437	-0.059	-0.008	-0.20	0.891	-0.012	0.047
一氧化碳含量(NOX)	0.946	-0.135	-0.005	0.30	0.192	-0.080	0.064
每方平米人口数(RM)	0.941	-0.170	0.026	-.010	0.187	-0.103	0.109
拥有住宅比例(AGE)	-0.585	0.371	-0.430	0.395	0.047	0.128	0.067
到商业中心的距离权重(DIS)	0.893	-0.269	0.207	-0.068	0.128	-0.174	0.071
高速公路便利性(RAD)	0.953	-0.140	0.019	0.007	0.185	-0.098	0.098
财产税的价值(TAX)	-0.442	0.774	-0.082	0.112	-0.124	0.310	-0.057
教师学生比例(PTRATIO)	0.747	-0.117	0.025	0.111	0.111	-0.048	0.635
每千人中黑人数(BLACK)	-0.919	-0.069	0.017	-0.115	0.115	0.020	0.016
低层次人口比分比(LSTAT)	0.111	0.209	-0.182	-0.022	-0.022	0.113	-0.007

个主成份都有很好的解释意义(载荷绝对值>0.5,说明变量与主成份存在相关性)。主成份 1 为城镇生活环境,主成份 2 为治安环境,主成份 3 为人口密度,主成份 4 为人口层次,主成份 5 为是否有河流,主成份 6 为商业环境,主成份 7 为教育发展水平。

(2)应用粒子群算法优化支持量机参数

支持量机回归待优化的参数有惩罚参数 C 和 ε 带参数 e,采用高斯径向基函数 $K(x_i,x_j)=\exp(-\gamma\|x_i,x_j\|^2)$ 作为 SVM 模型的核函数。选取降维后试验数据前 352 个样本作为训练样本,后 100 个样本作为预测样本。设 $C\in[1,500]$,$\varepsilon\in[0.01,10]$,$V_{id\_max}=X_{id\_max}$,$c_1=2.8$,$c_2=1.3$,种群规模为 30,最大迭代次数为 30,采用 3 折交叉验证模式下的均方误差(MSRE)作为评估粒子的适应度函数。优化后得到最优 C=375.754,ε=0.175,最优目标函数的适应度值 MSRE=7.4801。

(3)应用 e-SVM 进行回归预测

利用优化的 C、ε 对训练样本进行训练得到一个预测模型,然后用所得的预测模型对预测样本进行预测,预测 MSRE=3.987,而采用默认参数 C=1,ε=0.01 所得 MSRE=10.483,采用没有降维的数据并同样利用粒子群优化的方法优化参数,所得 MSRE=5.759,从预测的均方误差可知本文的方法能显著提高预测的精度。图 2 是采用本文方法降维前、后的预测值与真实值的拟合图,从图中可以看出本文方法具有较强的非线性预测能力。

本文把量子群优化算法和主成分析降维的方法应用于支持向量机的回归预测中,试验结果表明此法能显著提高支持向量机的预测精度,同时也表明了支持向量机在非线性、高维模式下的良好预测性能。

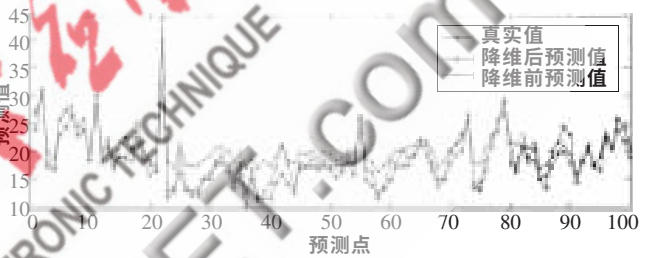

图 2 真实值和预测值曲线拟合图

参考文献

[1] 阎平凡,张长水.人工神经网络与模拟进化计算[M].北京:清华大学出版社,2006.

[2] 韦康南,姚立纲等.基于灰色理论的产品寿命预测研究[J].计算机集成制造系统,2005(10):1491-1495.

[3] VAPNIK V N. The nature of statistic learning theory[M]. New York: Springer, 2005.

[4] VAPNIK V N. Estimation of dependencies based on empiric [M]. Berlin Springer-Verlag, 2003.

[5] 邓乃扬, 田英杰. 数据挖掘中的新方法-支持向量机[M].北京:科学出版社,2004.

[6] KENNEDY J, EBERHART R. Particle swarm optimizat[A]. Proc IEEE Int Conf. on Neural[C]. Perth, 1995. 1942-1948.

[7] CLERK, M. The swarm and the queen: Towards a deterministic and adaptive particle swarm optimization[A]. 1951-1957. 1990. Proc. CEC 1999.

[8] 林海明. 对主成分分析法运用中的十个问题的解析[J].统计与决策(理论版),2007(8):16-18.

[9] http://archive.ics.uci.edu/ml/index.html 1993.07.

(收稿日期:2010-05-04)

作者简介:

周方军,男,1980 年生,硕士研究生,主要研究方向:模式识别与智能计算,设备故障诊断与预测。

吕文元,男,1971 年生,副教授,主要研究方向:模式识别与智能计算,设备故障诊断与预测。