一种基于 LS 拟合判别函数的 SVR 特征选择算法*)

王 浩 王行愚 牛玉刚

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 本文提出一种基于最小二乘(LS)拟合判别函数的 SVR 特征选择算法(简称 LS 特征选择法)。该算法采用了一个适合支持向量回归(SVR)的新目标函数,并在特征子集选择中根据实验数据集冗余特征较少的特点,采用顺序后向选择算法。仿真实验表明,本方法与常用的降维方法 PCA 和 KPCA 相比有更好的效果。

关键词 支持向量回归,特征选择,最小二乘法

SVR Feature Selection Method Based on Least Square Estimate Discriminative Function

WANG Hao WANG Xing-Yu NIU Yu-Gang

(College of Imformation Science and Engineering East China University of Science and Technology, Shanghai 200237)

Abstract This paper develops a SVR feature selection method (LS-FSM) based on least square-based estimate discriminative function. The target function adopted in this method is more suitable for SVR. Moreover, by considering that redundant feature is small in special experiment data, this work uses sequential backward selection method in the feature subset selection. The experiment results show that the proposed LS-FSM has a better performance than the common dimension reduced methods, e. g. PCA and KPCA, in feature subset selection for SVR.

Keywords Support vector regression (SVR), Feature subset selection, Least square method

1 引言

维数约简在支持向量回归(SVR)预测中占有重要的地位。所谓维数约简,通常是寻找包含了原始属性中必要信息的最小特征集。实验表明[1],支持向量机(SVM)在先进行特征选择后的预测效果不仅比不进行特征选择的预测效果好,而且可以提高训练速度。因此,要成功地进行 SVM 预测,第一步应该进行特征提取。但是要严格地确定训练样本的确切数目和特征个数以及它们之间的关系是很困难的。

针对上述问题,人们提出了许多解决方法。其中,文[2]在二类分类问题的支持向量分类(SVC)中首先提出,根据落入分界面附近狭窄空间的样本(即支持向量(SV)),建立预测函数,同时构造出评价判别函数的目标函数,并根据目标函数的结构拟合判别函数,运用顺序前向选择算法进行特征选择。但此方法仅限于二类分类问题。

本文将文[2]的思想拓展到支持向量回归(SVR)的特征选择中,提出一种基于最小二乘(LS)拟合判别函数的 SVR 特征选择算法(简称 LS 特征选择法)。该算法的主要思想是首先训练支持向量机,求出支持向量。采用适合 SVR 的新目标函数,根据 SVR 目标函数的结构构造对应的判别函数,采用顺序后向选择算法进行初步特征选择,并在达到停止条件后,对选择的最优特征子集用 SVR 重新训练。实验表明,在用 SVR 训练前,使用不同特征选择算法对数据集进行维数约简,本特征选择算法具有最小的预测偏差。

2 标准 SVR 方法

设训练样本集$\{(x_i,y_i),i=1,2,\cdots,l\}$,其中 $x_i\in R^N$ 为输入值,$y_i\in R$ 为对应的目标值,l 为样本数。定义 ε 不敏感损失函数为

$$|y-f(x,\omega)|_\varepsilon=\begin{cases}0 & |y-f(x,\omega)|<\varepsilon\\ |\eta|-\varepsilon & |y-f(x,\omega)|\geqslant\varepsilon\end{cases}\tag{1}$$

式中 $f(x,\omega)$是通过对样本集的学习而构造的预测函数,y 为与 x 对应的目标值;$\varepsilon>0$ 为设计参数,它规定了估计函数在样本数据上的误差要求。

在采用支持向量机研究非线性样本集时,通过非线性函数 $\Phi(\cdot)$将训练集数据映射到一个具有高维线性的特征空间,在这个维数可能为无穷大的特征空间中构造估计函数。支持向量机存在对偶表现形式,数据仅作为 Gram 矩阵的项出现,而不需要通过单个属性出现。预测函数 $f(x)$有如下形式[9]:

$$f(x)=\omega\cdot\Phi(x)+b\tag{2}$$

式中:$\omega\cdot\Phi(x)$表示向量 ω 与 $\Phi(x)$的内积,ω 的维数为特征空间维数(可能为无穷维),$b\in R$。根据统计学习理论[3],预测函数 $f(x)$的估计转换成如下的最优化问题:

$$\begin{aligned}&\min_{\omega,b,\xi_i,\xi_i^*}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*),\\&s.t.\ \ y_i-\omega\cdot\phi(x_i)-b\leqslant\varepsilon+\xi_i,\\&\qquad\omega\cdot\phi(x_i)+b-y_i\leqslant\varepsilon+\xi_i^*,\\&\xi_i\geqslant0,\xi_i^*\geqslant0,i=1,2,\cdots,l\end{aligned}\tag{3}$$

对式(3)的最优化问题,一般采用拉格朗日乘子法转换成对偶最优化问题,然后根据 KKT 条件进行最优化计算,得到的预测函数可以由(3)式写为如下具体的形式[3]:

$$f(x)=\sum_{i=1}^{L}(a_i-a_i^*)k(x_i,x)+b\tag{4}$$

训练 SVR 的本质是解决一个二次规划问题,SVR 的训练速度是限制其广泛应用的主要原因。近年来,人们针对该方法本身的特点提出了许多算法(如块算法[4]、分解算法[5]、顺序最小优化算法(SMO)[6])来解决对偶寻优问题,其共同的思想就是采用分而治之的原则,将原问题分解为规模较小的

*)高等学校博士学科点专项科研基金(20040251010)、上海市自然科学基金(04ZR14034)、国家重点基础研究发展规划项目(2002CB312200)。

王 浩 硕士研究生,主要研究方向为人工智能、数据挖掘及其在实际工程中的应用。

子问题。对 SVR 基础理论定理的介绍可见文[9]。

训练 SVR 需要计算和存储核矩阵，其占用的内存是样本数据的平方。在样本稍大的情况下，不可能将整个矩阵保存在内存中，这就增加了虚拟内存页替换的频率，使训练过程非常耗时(这也是本算法中避免用标准 SVR 评价大量候选特征子集，而是采用线性最小二乘法拟合目标函数来间接评价的原因)。在变量维数较大时，训练耗时将更加明显。为了改善训练速度，提高预测精度，应通过特征选择减少向量特征维数。本文中提出一种基于最小二乘(LS)拟合判别函数的 SVR 特征选择算法(简称 LS 特征选择法)，仿真实验结果显示其有良好的效果。

3 LS 特征选择法

给定一个输入数据集，要从中选出在 SVR 预测中效果最好的特征子集，使得用此特征子集训练 SVR 后，对测试集产生的偏差平方和最小。无疑，排列组合出所有可能的特征子集，逐一用 SVR 训练，求出各个偏差平方和，最小偏差和对应的特征子集必为最优特征子集。但是在实际运用中，每评价一个特征子集就要解决一个耗时的二次规划问题，当训练集和维数稍大就将陷入维数灾难，使求解变得不可行。

本文提出的 LS 特征选择法就是通过避免求解大量二次规划问题，使上述方法变得可行并具有较好效果。LS 特征选择法的基本思想是：

首先，对所有训练样本进行训练，求出 SVR 预测函数(6)。该预测函数的结构对于后面的特征子集判别函数的构建和评价具有良好的参考作用。SVR 预测函数的输出可以为任意实数，其预测值可以直接作为判断判别函数好坏的标准，将其作为后面拟合判别函数的目标函数。故在本文中，SVR 预测函数就是拟合判别函数的目标函数，这是不同于文[2]之处。文[2]由于采用二类分类 SVC，预测函数输出值为(1，-1)，因此另构造了一个输出为实数的目标函数，作为拟合判别函数的目标。

其次，当 SVR 目标函数式(6)确立后，待评价特征子集可以通过线性最小二乘算法拟合对应判别函数，逐一进行评价，顺序删除相对冗余属性。它的优点在于利用了 SVR 参数表达的线性关系，可以方便地利用线性最小二乘算法拟合判别函数对各特征值重要性进行估计，避免了求各特征子集对应的 SVR 预测函数导致的大量二次规划问题。另一方面，本文算法利用了 SVR 目标函数的结构，为构建和评价特征子集提供了较好的指导作用。

在 SVR 目标函数公式(6)中，样本保留所有的特征，其中有些特征可能是冗余的。本文算法的目的是为了找到特征子集 v 和一个新的判别函数 g，使下面的损失函数最小：

$$J=\sum_{k=1}^{M}\{f[x(k)]-g[z(k)]\}^2 \tag{5}$$

其中，M 表示总的测试样本数，$z(k)=\Gamma x(k)$，Γ 是降维操作符，表示从 $x(k)$ 中提取一个子集。

$$z(k)=[z_1(k),z_2(k),\cdots,z_m(k)]^T$$

$$z_i(k)\in\{x_1(k),x_2(k),\cdots,x_n(k)\}$$

$i=1,2\cdots,m$，$m<n$。从维数约简的角度，最好 $m\ll n$。SVR 的目标函数由支持向量组成，所以我们更加关注这些特殊的样本点。上式中的 g 和 f 具有相同的结构，因此为了估计 SVR 的目标函数(预测函数)，本算法选择下面的预测函数 $f(x)$ 和判别函数 $g(z)$：

$$f(x)=\sum_{i=1}^{L}(a_i-a_i^*)k(x_i,x)+b$$

可以简化为：

$$f(x)=\sum_{i=1}^{L}w_i k(s(i),x)+b \tag{6}$$

$$g(z)=\sum_{i=1}^{L}v_i k(r(i),z)+a \tag{7}$$

其中，w_i 表示拉格朗日算子，k 是核函数，$s(i)$ 是 L 个支持向量，b 是已确定参数，x 是 M 个 n 维测试向量。$z=\Gamma x$ $r(i)=\Gamma s(i)$，表示 z 和 r 分别是 x 和 s 的特征子集。回归算子 $k(r(i),z)$ 的数目始终等于支持向量的数目 L，特征子集选择因此不同于模式选择问题[7]，模式选择问题的研究中需要首先决定回归算子和它的数目。

在本文算法中，为了评价 LS 特征选择法选出的特征子集，我们首先用线性最小二乘算法估计各个 v_i，构造具有式(6)结构的判别函数(7)式，以使损失函数(5)式最小。虽然由于核函数的关系，判别函数 g 和输入变量 z 具有非线性关系，但是 $g(z)$ 和回归算子 $k(r(i),z)(i=1,2,\cdots,L)$ 却具有线性关系。这种参数线性结构使我们可以避免耗时的二次规划用线性最小二乘法来计算参数 v_i。在确定了判别函数(7)式后，可以用(5)式计算出预测偏差平方和 J，特征子集的优劣就是依靠这些偏差平方和来评价：J 值越小，所选的特征越好。

文[2]在二类分类 SVC 中提出的搜索方法是顺序前向选择算法(sequential forward selection)，循环在整个特征空间中找出冗余度最小的特征，与上一次循环找出的最优特征子集组合成新的特征子集，直到满足停止标准，避免了评价所有可能的特征子集的排列组合。这种搜索方法在冗余特征比较多，保留的最优特征子集比较小时，可以较快收敛。但是考虑到在预测股票指数的应用中，决定股价走势的因素较多，从直观上考虑，显然搜索后保留的最优特征不会太小，且文献中常用于预测股价的候选特征一般不超过 20 个，故不用文[2]的搜索方法，而采用顺序后向选择算法。通过循环对整个特征空间逐一搜索依次删除冗余度最大的特征，直到满足停止标准。两种方法本质上一样，在不同条件下收敛速度却各有优劣。

4 LS 特征选择法的算法步骤

①数据预处理

对 x 进行归一化处理并赋给 z，使 z 的所有特征在[0，1]区间。对后面提到的测试集也要用同样的方法进行归一化处理。设定停止条件，停止条件一般可以设为偏差平方和 J 达到一个预定的阈值，或预设特定的冗余特征个数被删除。

② 用 SVR 对数据集 z 进行训练，得到 L 个支持向量

$s(i),i=1,\cdots,L$

令 $r(i)=s(i),i=1,\cdots,L$，求出 SVR 预测函数 $f(x)$。

③(a)构造 n 个待评价特征子集。依次去掉 z 的第 j 个特征，作为待评价的特征子集 $z^{(j)}$，$j=1,2,\cdots,n$，(为了不混淆，特征序号 j 均指在原始输入空间中的第 j 个特征)。

$z^{(j)}=\Gamma_j z$，Γ_j 表示去掉向量 z 的第 j 维，

$r^{(j)}(i)=\Gamma_j r(i)$，Γ_j 表示去掉向量 $r(i)$ 的第 j 维，

$i=1,2,\cdots,L$，$j=1,2,\cdots,n$。

(b)计算出 L 个回归回子：

$\Phi^{(j)}(i)=K[r^{(j)}(i),z^{(j)}]$，$i=1,2,\cdots,L$

(c)构造 n 个判别函数：

$g(z^{(j)})=v_1\Phi^{(j)}(1)+v_2\Phi^{(j)}(2)+\cdots v_L\Phi^{(j)}(L)+a$

$j=1,2,\cdots,n$

(d)找出本次循环中的较优特征子集。

对每一个判别函数通过最小化(5)式，使用线性最小二乘法估计参数 v_i 和 a，并计算对应的 n 个判别函数对测试数据的偏差平方和 J，导致最小 J 值的待评价特征子集(比如说

$x_k, k_1 \in \{1,2,\cdots,n\}$)，表示去掉 x 的第 k_1 个特征后(显然每次循环中的 k_1 并不相等，因为同一特征不可能被删除多次)，新的特征子集对应的判别函数具有相对最小的偏差。将新的特征子集赋给 z，作为较优特征子集，即

$z=\Gamma_{k_1} z$，同理 $r=\Gamma_{k_1} r$.

$n=n-1$

④判断停止条件

如果不满足停止条件，则跳转第③步，否则转第⑤步。

⑤对选出的特征子集用 SVR 重新训练，生成最终的预测函数。

5 实验仿真

①对公共数据库[8]中的 bodyfat 进行特征选择。共 215 个数据、14 个特征。为了清楚显示算法流程，以及和其它算法进行比较，选择其中第 10～11 步循环，列入表 1。

表 1 LS 特征选择法循环去掉每个特征后对应的偏差平方和

循环次数 \ 特征	1	2	3	4	5	6	7	8	9	10	11	12	13	14	待删特征
10	1.16 e−02	5.59 e−06	#	5.51 e−06	#	#	1.05 e−05	#	#	#	#	#	4.48 e−06	#	13
11	1.17 e−02	7.79 e−06	#	7.25 e−06	#	#	1.54 e−05	#	#	#	#	#	#	#	3

表 1 清晰地表明了，在本文介绍的 LS 拟合判别函数 SVR 特征选择算法中，冗余特征是如何被逐一删除的。例如，在第 10 步循环中，已经删除了第 3、5、6、8、9、10、11、12、14 个特征，在剩下的特征向量中依次去掉每一个特征，构造出 5 个候选特征子集，计算对应的偏差平方和，然后找出最小的偏差 4.48e-06，即表 1 中第 13 个特征。表明本次循环中如果去掉第 13 个特征，对 SVR 预测的影响最小，故删除此特征，进入下一循环，直到达到停止标准。从实验中我们还可以发现，不论在任何一次循环中去掉第一个特征，预测偏差都会明显增大，所以第一个特征应优先保留。本算法中各特征对应的预测偏差类似于 PCA 降维算法中的各特征贡献率。

图 1 是对有 14 个特征的实验数据 bodyfat，进行特征选择后，再用 SVR 预测，所对应的偏差平方和。为了观察预测误差(偏差平方和)的变化趋势，将 LS 特征选择法终止条件中保留的特征数从 1 变化到 13。将不同终止条件选出的最优特征子集用 SVR 训练，分别求出对应偏差平方和，并与常用降维算法 PCA、KPCA 作了比较。

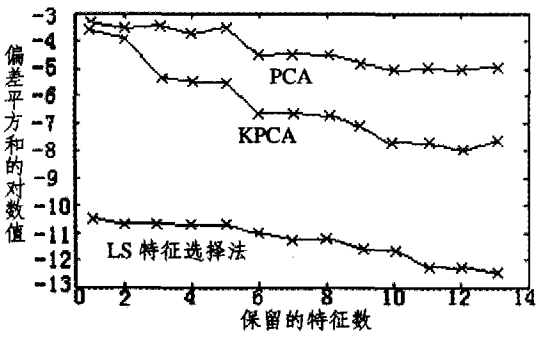

图 1 LS 特征选择法与常用降维算法降维后 SVR 的偏差平方和比较

图 1 表明，随着特征个数的减少，各方法对特征降维后，再用 SVR 训练，对应的预测误差(偏差平方和)均在增大，说明 bodyfat 数据中没有冗余特征。但用三种方法对数据维数约简后，随着特征个数的减少，LS 特征选择法预测误差增大明显较慢，优于 PCA、KPCA 方法。

②对深综指数的 5 周均线进行预测。数据来自 1995 年 12 月 22 日到 2005 年 03 月 18 日的深综指数的周 K 线。从上百个指标中选取下面常用的 16 个指标作为候选特征。

表 2 LS 特征选择法对候选特征进行特征选择后的最优特征

候选特征	成交量 BIAS BRAR. AR VR. VR SAR RSI
	PSY OBV MACD. DIFF KDJ. K MACD. DEA
	KDJ. D DMI. PDI DMI. MDI BOLL. MID KDJ. J
最优特征	成交量 SAR KDJ. K KDJ. D DMI. MDI MACD. DEA

仿真结果如图 2 所示，表明在保留下列 6 个特征时偏差最小，预测精度最好。

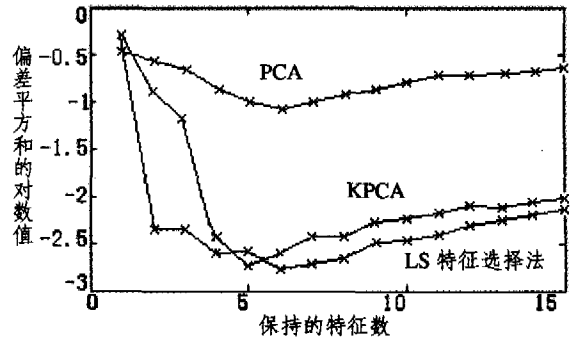

图 2 LS 特征选择法与常用降维算法降维后 SVR 的偏差平方和比较

结论 实验表明，LS 特征选择法在对数据集进行特征选择后，对有冗余特征的原始数据 SVR 预测精度会提高，对无冗余特征的数据如果强行降维，虽然 SVR 预测精度会降低(当然如果很重视精度问题，可以在终止条件中通过设定偏差平方和的阈值，使得对于无冗余特征的数据一个特征也不删除)，但好于常用的 PCA、KPCA 降维方法。

LS 特征选择法实际上是一种特殊的特征选择算法，其设计中充分考虑到了 SVR 预测函数的特点。对数据集做特征选择后，非常适合且仅适合用于 SVR 对数据学习，预测效果比用其它降维算法(PCA，KPCA)进行特征选择再用 SVR 训练的效果好。

参 考 文 献

1 Caoa L J, Chuab K S, Chongc W K, et al. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing, 2003, 55: 321～336

2 Mao K Z. Feature Subset Selection for Support Vector Machines Through Discriminative Function. IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics, 2004, 34(1)

3 Vapnik V N. The nature of statistical learning [M]. Berlin: Springer, 1995

4 Boser B E, Guyon I M, Vapnik V. A Training algorithm for Optimal Margin Classifier. In: Proc. of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 1992. 144～152

5 Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. In: Proc. of the 1997 IEEE Workshop on Neural Networks for Signal Processing, New York: IEEE Press, 1997. 276～285

6 Fast P J C. training of support vector machines using sequential minimal optimization. In: Burges C, Scholkopf B, eds. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999. 185～208

7 Mao K Z, Billings S A. Algorithms for minimal model structure detection for nonlinear dynamic system Identification. Int J Contro, 1997, 68(2): 311～330

8 http://www.csie.ntu.edu.tw/～cjlin/libsvmtools/regression data

9 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Publishing House of Electronics Industry, 2004