基于特征指数加权的最小二乘支持向量机算法①

潘 岚，王仲君

(武汉理工大学 理学院，武汉 430070)

摘 要：根据支持向量回归机原理，针对样本特征对回归预测重要性的差异，采用最小二乘支持向量回归机（LS-SVR）算法，减少参数数量，针对参数对预测效果的影响，并考虑到特征加权的意义，采用特征指数进行加权，其权重系数由灰色关联度确定，提出了基于特征指数加权的最小二乘支持向量回归机算法。为验证该算法的有效性，对实际股票价格进行预测，结果表明该算法较传统最小二乘支持向量回归机算法，其回归估计函数的预测能力明显提高，具有一定的实用价值。

关键词：支持向量回归机；特征加权；灰色关联度；评价指标；核函数

Least Squares Support Vector Machine Based on Exponentially Weighted Feature

PAN Lan, WANG Zhong-Jun

(School of Science, Wuhan University of Technology, Wuhan 430070, China)

Abstract: According to the basic principle of support vector regression algorithm, for the difference of features` correlative degree to the regression, the affect of parameters to the performance of forecast and taking into account the significance of weighted feature after normalization, least squares support vector regression machine (LS-SVR) based on weighted feature is proposed in this paper, in which, least squares support vector regression algorithm is used to reduce the number of parameters and exponentially weighted feature is used to improve prediction accuracy, the weighting coefficients are determined by the grey correlation degree approach. In the meantime, the effectiveness of the algorithm is demonstrated in forecasting the actual stock price. The experimental results show that it is superior to classical support vector machine and can significantly improve the predictive ability.

Key words: least squares support vector regression machine; weighted feature; gray correlation degree; evaluation index; kernel

1 引言

支持向量机是由 Vapnik 等人提出并在统计学习理论（SLT）的基础之上快速发展起来的最开始用于分类的一种新的学习方法。因为其坚实的理论基础以及良好的泛化性能目前被广泛应用于模式识别，文本分类、函数拟合以及回归等诸多领域。在传统支持向量回归机算法中，一般情况下都是认为其使用的样本所具有的所有特征都是具有一样的重要性。但在实际上，每个样本的特征对于问题的相关程度不可能是完全相同的，那么认为所选择的样本中某些冗余的、相互影响的、与回归无关的以及被噪声污染的特征与其它特征拥有相同地位则会对支持向量回归机的实用效果产生影响。文献[1]提出了一种基于特征系数加权的支持向量回归机(WF- SVR) ，它对每一个特征给予一个权重系数来表示该特征与问题的关联程度，并在实验中对一维函数拟合的噪声进行主观加权得到了很好的拟合回归效果。本文提出一种基于特征指数加权的最小二乘支持向量回归机算法。

该算法主要从两个方面进行考虑，一是对于算法中参数的选取。参数的选取对于回归预测的效果有很

① 收稿时间:2011-09-01;收到修改稿时间:2011-10-06

大的影响，传统支持向量机算法中进行回归预测需选取三个参数，但最小二乘支持向量回归机算法中选取的参数只需两个，为减少对参数的选取，本文选用最小二乘支持向量回归机算法；二是对于所使用样本特征对回归预测重要性差异的表示。算法中所使用样本的每个特征对于回归预测效果的影响不可能完全相同，为表示这种差异性，目前比较公认的是利用样本加权来体现不同样本对训练结果的影响，但同时这也是存在问题的，如果对特征直接加权，在理论上的难以体现各特征之间的重要性，如在 SVM 训练过程中，标准方法中还经常要求各个特征值进行“归一化”。很明显，归一化之后，任何特征的“加权”都失去了意义，这也从一个侧面说明了对特征进行“加权”未必能提高该特征的重要性。针对这些存在的问题，本文提出了对其特征指数进行加权的方法，其权重系数由灰色关联度确定。为验证该算法的有效性，本文选取实际股票数据分别用 LS-SVR 和加权 LS-SVR 方法对实际股票价格进行拟合与预测，结果表明加权 LS-SVR 的性能优于传统 LS-SVR，具有一定的实用价值。

2　最小二乘支持向量回归机（LS-SVM）

给定训练集 $\mathrm{T}=\{(\mathrm{x}_1,\mathrm{y}_1),(\mathrm{x}_2,\mathrm{y}_2),\cdots,(x_l,y_l)\}\in(R^n\times R^l)$，其中第 i 个输入数据 $x_i\in R^{n\times 1}$，第 i 个输出数据 $y_i\in R$，$i=1,\cdots,l$。SVM 的目的就是用训练集 T 寻找一个实值函数 $f(x)$，以便用 $y=f(x)$ 来推断任一输入 x 所对应的输出值 y [2]。通过对输入的样本空间利用非线性映射 $\phi:x\to\phi(x)$ 变换到另一个高维特征空间中，在这个特征空间中构造一个回归估计函数, 利用结构风险最小化原则，通过 $\|\omega\|^2$ 最小化来减少模型的复杂度。

对于一个 $\varepsilon-$ 不敏感损失函数作为损失函数[3,4]：$|y-f(x)|_\varepsilon=\max\{0,|y-f(x)|-\varepsilon\}$，引入非负的松驰变量 $\xi^{(*)}=(\xi_1,\xi_1^*,\cdots,\xi_l,\xi_l^*,)^T$ 和惩罚参数 C，得到的标准 $\varepsilon-$ 支持向量回归机($\varepsilon-$ SVR)的原始最优化问题[2]:

$$\begin{aligned}\min_{\omega,b,\xi^{(*)}}\quad & \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\ s.t.\quad & ((\omega\cdot\phi(x_i))+b)-y_i\le\varepsilon+\xi_i\\ & y_i-((\omega\cdot\phi(x_i))+b)\le\varepsilon+\xi_i^*\\ & \xi_i^{(*)}\ge 0,\quad i=1,2,\cdots,l\end{aligned}\tag{1}$$

对于标准 $\varepsilon-$ 支持向量回归机($\varepsilon-$ SVR)，其中需要确定的参数有三个。最小二乘支持向量回归机算法是标准支持向量机的一种扩展，是将原二次规划问题转化为求解线性方程，求解速度快，参数也由传统 $\varepsilon-SVR$ 中的三个变成 LS-SVR 中的两个，减少了对参数的选取。其原始最优化问题为[5]:

$$\begin{aligned}\min_{\omega,e}\quad & \frac{1}{2}\|\omega\|^2+\gamma\sum_{i=1}^{l}e_i^2\\ s.t.\quad & y_i=\omega^T\phi(x_i)+b+e_i,i=1,2,\cdots l\end{aligned}\tag{2}$$

其对偶问题为求解如下方程组：

$$\begin{bmatrix}0 & 1^T\\ 1 & \Omega+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix}\tag{3}$$

其中 $\Omega_{ij}=K(x_i,x_j)=(\phi(x_i)\cdot\phi(x_j))$ 为支持向量机的核函数。则可求得 α、b ,即得回归估计公式：

$$f(x)=\sum_{i=1}^{l}\alpha_iK(x,x_i)+b\tag{4}$$

3　基于加权的最小二乘支持向量回归机

回归模型中所用样本的特征对于所要预测的问题不完全相关或者完全无关时，对于拟合和预测的效果都会产生很大的影响，为了表示每个特征对于回归预测问题重要性的相关程度，可以对不同的特征的指数施加不同的权重。设输入样本的 n 个特征为 $X_1,X_2,\cdots,X_n$，其相应的权重为 $\lambda_1,\lambda_2,\cdots,\lambda_n$，$0\le\lambda_i\le 1$。令特征权重向量 $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)$，则最优化问题(2)转化为：

$$\begin{aligned}\min_{\omega,e}\quad & \frac{1}{2}\|\omega\|^2+\gamma\sum_{i=1}^{l}e_i^2\\ s.t.\quad & y_i=\omega^T\phi(x_i^\lambda)+b+e_i,i=1,2,\cdots,l\end{aligned}\tag{5}$$

其中 $x_i^\lambda=(x_{i1}^{\lambda_1},x_{i2}^{\lambda_2},\cdots,x_{in}^{\lambda_n})$，引入 Lagrange 函数：

$$\begin{aligned}L(\omega,b,e,\alpha)=&\frac{1}{2}\|\omega^2\|+\gamma\sum_{i=1}^{l}e_i^2-\sum_{i=1}^{l}\alpha_i\cdot\\ &(\omega^T\phi(x_i^\lambda)+b+e_i-y_i)\end{aligned}\tag{6}$$

其中 $\alpha\ge 0$ 为 Lagrange 乘子。根据 KKT 条件(Karush-Kuhn-Tucker)，对 ω,b,e,α 求偏导数并令其为 0，则最优解条件为：

$$\begin{cases} \dfrac{\partial L}{\partial \omega}=0 \\ \dfrac{\partial L}{\partial b}=0 \\ \dfrac{\partial L}{\partial e_i}=0 \\ \dfrac{\partial L}{\partial \alpha_i}=0 \end{cases} \text{则} \begin{cases} \omega-\sum_{i=1}^{l}\alpha_i\phi(x_i^{\lambda})=0 \\ \sum_{i=1}^{l}\alpha_i=0 \\ 2\gamma e_i-\alpha_i=0 \\ \omega^T\phi(x_i^{\lambda})+b+e_i-y_i=0 \\ i=1,2,\cdots,l \end{cases} \tag{7}$$

可将其优化求解，转化为求解如下方程组：

$$\begin{bmatrix} 0 & 1^T \\ 1 & \Omega'+\gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{8}$$

其中 $\Omega'_{ij}=K(x_i^{\lambda},x_j^{\lambda})$，求得 α， b，得到回归估计公式：

$$f(x)=\sum_{i=1}^{l}\alpha_i K(x_i^{\lambda},x_j^{\lambda})+b \tag{9}$$

4 实验与分析

为了验证该算法的有效性，本文实验数据选取浦发银行 2011.1.4 到 2011.5.11 之间的 83 个交易日的股票数据，基本特征分别是最高价、最低价、开盘价、成交量、成交金额，来预测股票收盘价，其中 2011.1.4-2011.4.29 之间 76 个交易日数据作为训练集，2011.5.3 - 2011.5.11 这 7 个交易日数据作为测试集。

1) 评价指标。

采用均方误差(mean squared error，MSE)和平均绝对误差百分率(mean absolute percentage error，MAPE)作为评价指标。定义如下：

$$MSE=\frac{1}{l}\sum_{i=1}^{l}(y_i-\hat{y}_i)^2 \tag{10}$$

$$MAPE=\frac{1}{l}\sum_{i=1}^{l}\frac{|y_i-\hat{y}_i|}{y_i} \tag{11}$$

2) 权重。灰色关联度能够反映出各因素相对预测变量的相关程度，实验采用归一化灰色关联度来确定各特征权重，即 $\lambda_i=r_i/\sum_{i=1}^{n}r_i$，为第 i 个特征相对预测变量的关联度。灰色关联度的计算可参见文献[6]。本文权重向量为：

$$\lambda=(0.2222,0.2245,0.2227,0.1646,0.1660)。$$

3) 核函数。常用的核函数有线性核函数，多项式核函数和 Gauss 径向基核函数等。

本文算法中选取 Gauss 径向基核函数进行预测研究。定义如下：

$$K(x_i,x_j)=\exp(-\frac{1}{2\sigma^2}\left\|x_i-x_j\right\|^2) \tag{12}$$

其中核参数 σ 与惩罚参数 γ 由 10-折交叉验证网格法进行选取。

实验采用的是 $LS-SVMlab$ $v1.7$ 工具箱[7]，所有计算均通过 MATLAB 求解。分别采用 LS-SVR 方法与加权 LS-SVR 算法进行预测，其预测结果见表 1，预测效果比较见表 2。对训练样本的拟合及测试样本的预测结果如图 1、图 2 所示。

表 1 两种模型的预测结果

日期	真值	$LS-SVR$	加权 $LS-SVR$
2011.5.3	14.29	14.2360	14.2826
2011.5.4	13.83	13.9377	13.9230
2011.5.5	13.88	13.7756	13.8134
2011.5.6	13.82	13.7830	13.8323
2011.5.9	13.76	13.7682	13.8058
2011.5.10	13.97	13.8075	13.8647
2011.5.11	13.88	13.8375	13.8866

表 2 两种模型的预测效果比较

	训练集		
	$LS-SVR$	加权 $LS-SVR$	提高(%)
MSE	0.0636	0.0628	1.2579
MAPE	0.3591	0.3515	2.1164
	测试集		
	$LS-SVR$	加权 $LS-SVR$	提高(%)
MSE	0.0079	0.0038	51.8375
MAPE	0.0053	0.0035	34.4970

从上述实验结果的图表里都可以看出来不管是对于训练集的拟合还是测试集的预测，加权 LS-SVR 的结果都是优于 LS-SVR 的结果，精度相对都有所提高，特别是在进行预测时，效果更加明显。

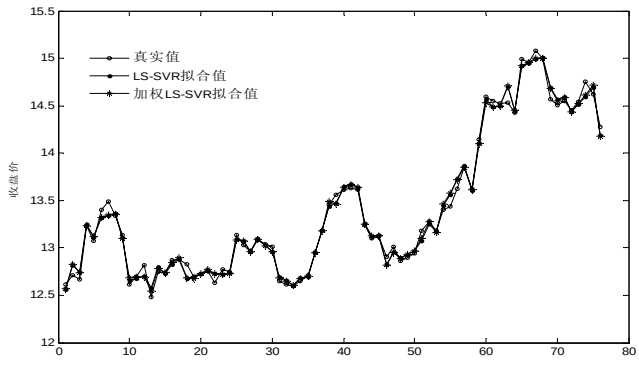

图 1　LS-SVR 与加权 LS-SVR 对训练样本的拟合

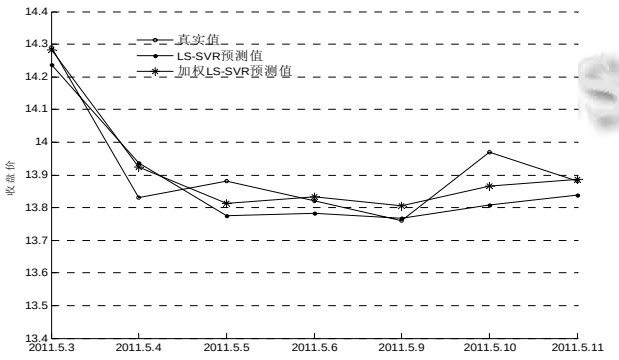

图 2　LS-SVR 与加权 LS-SVR 对测试样本的预测

5　总结

本文考虑到不同特征对于所预测问题的相关程度不同，提出了对于不同的特征指数施加不同的权重，同时根据参数的选取对于预测效果的影响，提出了基于、征指数加权的最小二乘支持向量回归机的模型，解决了样本特征对回归预测结果相关程度的差异性和减少参数选取对预测效果的影响性以及特征加权方法的有效性等问题。同时为验证该算法的有效性，选取实际股票价格数据利用加权 LS-SVR 算法与传统 LS-SVR 算法进行实验验证，并对拟合与预测的结果进行比较，实验结果表明加权 LS-SVR 较 LS-SVR 有更好的回归预测能力，效果基本令人满意，说明该算法具有一定的实用价值。但同时本文还存在一定不足，比如对于特征指数权重的确定，对于测试样本的预测时间长短等，因此下一步的主要任务是进一步的探讨研究如果能更好的确定特征权重，使得特征加权支持向量回归机算法得到更好的改进，以便能有效的应用于实际中。

参考文献

1 金凌宵,张国基.基于特征加权的支持向量回归机研究.计算机工程与应用,2007,43(6):42–44.
2 邓乃杨,田英杰.支持向量机理论、算法与拓展.北京:科学出版社,2009.100–105.
3 邓乃杨,田英杰.数据挖掘中的新方法-支持向量机.北京:科学出版社,2004.226–228.
4 Vapnik V. Statistical leaning theory. New York:Wiley, 1998.
5 王定成.支持向量机建模预测与控制.北京:气象出版社,2009.96–97.
6 罗党.灰色决策问题分析方法.郑州:黄河水利出版社,2005.29–31.
7 http://www.esat.kuleuven.be/sista/lssvml.

(上接第 200 页)

4 侯文静,马永杰,张燕,石玉军.求解 TSP 的改进蚁群算法.计算机应用研究,2010,27(6):2087–2089.
5 黄永.改进蚁群算法及其在公交线网优化中的应用[硕士学位论文].上海:华东师范大学,2009.
6 Qi CM. An ant colony system hybridized with randomized algorithm for TSP. Proc. of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. 2007. 461–465.
7 吴斌,史忠植.一种基于蚁群算法的TSP问题分段求解算法.计算机学报,2001,24(12):1328–1333.
8 于滨,杨忠振,程春田.并行蚁群算法在公交线网优化中应用.大连理工大学学报,2007,47(3):211–214.
9 段海滨.蚁群算法原理及其应用.北京:科学出版社,2005.
10 马良,朱刚,宁爱兵.蚁群优化算法.北京:科学出版社,2008.