ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, Supplement, December 2008, pp.121−130

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向时序预测的支持向量回归参数选择方法*

林树宽+, 徐传飞, 乔建忠, 张少敏, 支力佳, 于 戈

(东北大学 信息科学与工程学院,辽宁 沈阳 110004)

Parameter Choosing Method of Support Vector Regression for Time Series Prediction

LIN Shu-Kuan+, XU Chuan-Fei, QIAO Jian-Zhong, ZHANG Shao-Min, ZHI Li-Jia, YU Ge

(School of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

+ Corresponding author: E-mail: linshukuan@ise.neu.edu.cn

Lin SK, Xu CF, Qiao JZ, Zhang SM, Zhi LJ, Yu G. Parameter choosing method of support vector regression for time series prediction. *Journal of Software*, 2008,19(Suppl.):121−130. http://www.jos.org.cn/1000-9825/19/s121.htm

Abstract: As a new learning method, Support Vector Regression (SVR) has good generalization and prediction performance for time series modeling and predicting. In the course of SVR modeling, parameter choosing is very important to the accuracy of models. Aimed at problems in parameter optimization of SVR models, the paper proposes an SVR parameter choosing method for time series prediction, which improves the traditional cross-validation according to the features of time series prediction and sufficiently mines information included in numbered samples on the basis of maintaining the direction characteristic of time series. Furthermore, it is combined with ε-weighted SVR in order to get good model parameters. Experimental results over typical time series show the validity of the parameter choosing method of SVR. The method gets good effect applied to time series prediction.

Key words: support vector regression; time series prediction; improved cross-validation; ε-weighted; parameter choosing

摘 要: 支持向量回归作为一种新的学习方法,在用于时间序列建模与预测时具有较好的泛化性能和预测能力.在支持向量回归建模的过程中,参数的选择对于模型的准确性至关重要.针对目前支持向量回归模型参数优化中存在的问题,提出一种面向时间序列预测的支持向量回归参数选择方法.根据时间序列及其预测的特点,对传统的交叉验证方法进行了改进,在保证时间序列预测方向性特征的基础上,充分挖掘有限样本所包含的信息,并将之与ε-加权的支持向量回归相结合以选择好的模型参数.典型时间序列上的实验结果表明了所提出的支持向量回归参数选择方法的有效性,该方法在用于时间序列预测时取得了良好的效果.

关键词: 支持向量回归;时间序列预测;改进的交叉验证;ε-加权;参数选择

支持向量机(support vector machine,简称SVM)[1,2]是由Vapnik提出的一种基于统计学习理论的新的学习方

* Supported by the National Natural Science Foundation of China under Grant Nos.60873009, 60773220 (国家自然科学基金); the Natural Science Foundation of Liaoning Province of China under Grant No.20072035 (辽宁省自然科学基金); the Key Project of Liaoning Province of China under Grant No.2007216007 (辽宁省攻关项目); the Key Laboratory for Software System Development of Liaoning Province of China (辽宁省软件系统开发重点实验室)

Received 2008-05-01; Accepted 2008-11-25

法,它遵循结构风险最小化的准则,具有结构简单、全局优化和较好的泛化性能,近年来已成为新的研究热点.它最初主要用于解决模式识别问题[3,4],目前,它的用途已扩展到回归函数估计、非线性系统辨识、预测等方面,显示出较好的学习性能[5].

对于非线性时间序列预测,支持向量回归与其他基于经验风险最小化准则的学习方法相比(如神经网络),具有较好的泛化性能和预测效果.然而,在实际运用支持向量回归进行建模和预测的过程中,能否取得预期的效果,很大程度上依赖于 SVR 模型的参数选择,针对某一特定问题,选择的模型参数不同,模型输出会有较大的差异,因此,模型参数的选择对预测结果有着至关重要的影响.目前,几乎没有有效、实用的SVR模型参数选择方法,一些方法虽有较强的理论基础,但因其参数选择过程的复杂性而并不实用.在实际的SVR建模过程中,通常凭经验手调或采用交叉验证方法选择参数.凭经验选出的参数通常很粗糙且容易陷入局部极小;传统的交叉验证方法在面向时间序列预测时并不是十分有效.本文根据时序数据的特点,提出一种实用的时间序列 SVR 预测模型参数选择方法,将ε-加权的 SVR 与改进的交叉验证有机地结合起来,在时间序列预测中取得了较好的效果.

1 相关工作

关于 SVR 模型的参数选择,一些学者对其进行了研究,并提出了相应的方法.如文献[6]基于迭代技术计算 radius/margin 边界并使其最小化来校准支持向量机的参数;文献[7]在以半径-间距上界最小化为目标的支持向量机参数优化的框架下,提出一种简化算法,利用固定的迭代步长实现核参数的优化;文献[8]采用梯度下降算法,通过最小化泛化误差的估计实现支持向量机的多参数优化.在已有的算法中,有的只对单一核参数进行优化,有的只用来选择模式分类中的 SVM 模型参数,而不是面向时序数据 SVR 模型中的参数选择,有的因计算复杂而并不实用.

一些学者通过遗传算法等优化手段来选择SVM或SVR模型的参数[9],但需要将与预测误差相关的函数定义为适应度函数,使得每一次交叉和变异都要重复建模和计算误差.留一法(leave-one-out)是一种有效的参数选择方法,但由于每次只测试一个样本,使得参数选择过程比较繁琐.交叉验证方法是留一法的扩展,与之不同之处在于:它使用一组样本而非一个样本作为测试数据,因而使每组参数的测试过程得以简化.本文根据时间序列及其预测的特点,改进了传统的交叉验证方法,并将之与一种ε-加权的支持向量回归结合起来选择SVR模型的参数,由于参数优化过程更有针对性,参数优化的结果也更好.

2 时间序列 SVR 模型及其参数

2.1 时间序列支持向量回归建模

设时间序列的样本空间为

$$(x_1,y_1),(x_2,y_2),...,(x_n,y_n) \in R^m \times R \tag{1}$$

当样本数据集为线性时,回归预测函数可表示为

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

其中 $w\cdot x$ 是向量 $w\in R^n$ 与 $x\in R^n$ 的内积, $b\in R$.可通过求解一个凸二次规划问题求解式(2)中的w和b,从而得到预测函数.SVM同时考虑经验风险和置信范围,因此可通过最小化下面的目标函数得到问题的解[1,2]:

$$Q(w)=\frac{1}{2}\|w\|^2+C\cdot R_{emp}(f) \tag{3}$$

其中,第 1 项使得函数更加平坦从而提高推广能力,第 2 项是经验误差,常量C称为惩罚系数,体现对样本预测误差的惩罚程度,也是对置信范围和经验风险的一种折衷[2].$R_{emp}(f)$为损失函数,一般包括Huber函数、Laplace函数、ε不敏感损失函数等,ε不敏感损失函数因其具有较好的性能而被广泛采用.ε不敏感损失函数$L^{\varepsilon}(x,y,f)$定义如下:

$$L^{\varepsilon}(x,y,f)=\left|y-f(x)\right|_{\varepsilon}=\max(0,\left|y-f(x)\right|-\varepsilon) \tag{4}$$

其中,f是域X上的实值函数,$x\in X$且$y\in R$.ε被称为不敏感损失.当$|y-f(x_i)|=|y-(w\cdot x_i)-b|\le\varepsilon(i=1,2,\ldots,l)$时,认为没有误差.这一条件通常不被完全满足,因此引入松弛因子.此时,式(3)的优化目标函数即转换为[1]

$$\begin{aligned}&\text{minimize}\quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\&\text{subject to}\quad (\langle w\cdot x_i\rangle+b)-y_i\le\varepsilon+\xi_i\\&\qquad\quad y_i-(\langle w\cdot x_i\rangle+b)\le\varepsilon+\xi\\&\qquad\quad \xi_i,\xi_i^*\ge 0\quad (i=1,2,...,l)\end{aligned}\tag{5}$$

这里,松弛因子ξ_i是为在目标值之上超出ε所设,ξ_i^*是为在目标值之下超出ε所设.

这是一个凸二次规划问题.通过引入上式的Lagrange函数并求解下面的对偶形式获得乘子α,α^*:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i\cdot x_j\rangle\\&\text{subject to}\quad 0\le\alpha_i,\alpha_i^*\le C\quad (i=1,2,...,l)\\&\qquad\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0\end{aligned}\tag{6}$$

对于非线性时间序列,SVR 引入一个映射函数φ将原始数据映射到一个新的特征空间,从而将非线性问题转化为新特征空间中的线性问题.此时,回归预测函数可表示为

$$f(x)=w\cdot\varphi(x_i)+b\tag{7}$$

式(6)的优化目标函数转化为下面的形式:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle\Phi(x_i)\cdot\Phi(x_j)\rangle\\&\text{subject to}\quad 0\le\alpha_i,\alpha_i^*\le C\quad (i=1,2,...,l)\\&\qquad\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0\end{aligned}\tag{8}$$

通过引入核函数 $K(x,y)=\phi(x)\cdot\phi(y)$,式(8)可写为下面的优化目标函数:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i\cdot x_j)\\&\text{subject to}\quad 0\le\alpha_i,\alpha_i^*\le C\quad (i=1,2,...,l)\\&\qquad\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0\end{aligned}\tag{9}$$

因此,可求得相应的时间序列预测函数为

$$f(x,\alpha_i,\alpha_i^*)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b\tag{10}$$

在式(9)中,核函数的具体形式可为多项式核函数、Sigmoid 核函数、径向基核函数等,本文只讨论径向基核函数,如式(11)所示.

$$K(x,y)=\exp(-gamma\times|x-y|^2)\tag{11}$$

这里,*gamma* 称为核参数.

2.2 SVR模型参数及其影响

由第 2.1 节可知,在建立时间序列的 SVR 模型时,需要选择的参数包括:惩罚系数 C、不敏感损失ε和核参数 *gamma*.

惩罚系数C对于模型结果的影响[10]可以从SVR的目标函数(式(5))中看出,在式(5)中,C表示对经验误差的惩罚,在置信风险与经验误差之间起到折衷的作用,当C取值很小时,对经验误差的惩罚很小,致使不能通过样本的学习获得小的拟合与预测误差;当C取值很大时,对经验误差的惩罚过大,导致过学习,泛化推广能力下降,预测误

差增大,如图 1 所示.因此,为了获得较准确的预测结果,提高泛化能力,C应在一定的取值范围内.同样,太大或太小的ε和$gamma$值都会影响预测性能,对于特定的问题,其取值也应在适当的范围内.如图 2 和图 3 所示*.

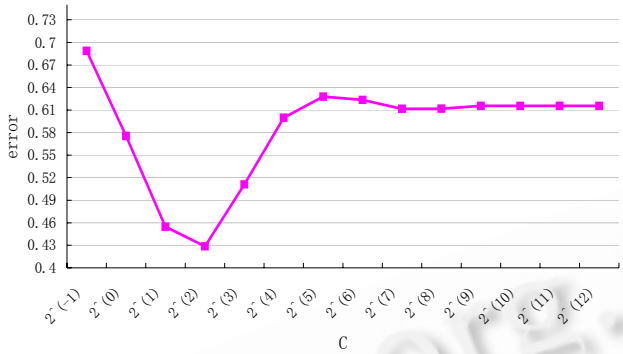


Fig.1 Influence of punishing coefficient C

图 1 惩罚系数 C 对预测误差的影响

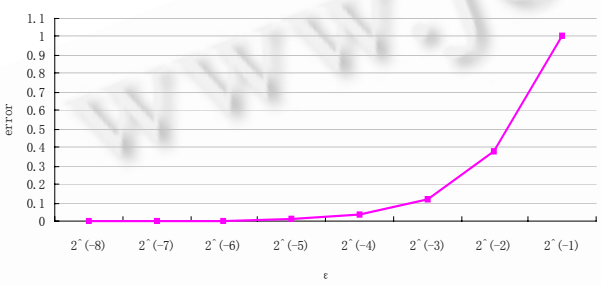


Fig.2 Influence of non-sensitive loss ε

图 2 不敏感损失ε对预测误差的影响

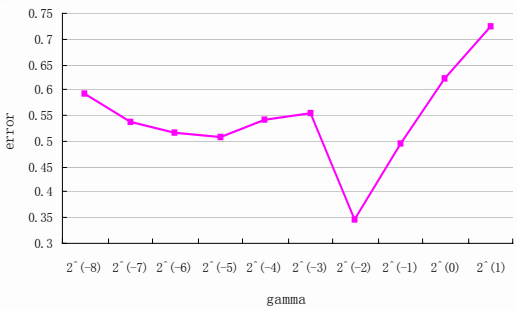


Fig.3 Influence of value *gamma*

图 3 参数 *gamma* 对预测误差的影响

3 时间序列预测及其特点

时间序列预测即是根据时间序列的历史数据$\{y(t),y(t-1),\ldots,y(t-m+1)\}$预测未来 $t+k(k>0)$时刻的值 $y(t+k)$,即寻找 $y(t+k)$与历史数据$\{y(t),y(t-1),\ldots,y(t-m+1)\}$之间的关系.可用下式描述:

$$y(t+k)=F(y(t),y(t-1),...,y(t-m+1)) \tag{12}$$

当k=1 时,称为一步预测;当k>1 时,称为多步预测.参数m称为嵌入维数.这里考虑一步预测.预测模型的训练样本对$(x_i y_i)(i=1,2,\ldots,n-m)$分别来自下面的矩阵$X$和$Y$:

$$X=\begin{bmatrix}x_1\\x_2\\\vdots\\x_{n-m}\end{bmatrix}=\begin{bmatrix}y_1 & y_2 & \cdots & y_m\\y_2 & y_3 & \cdots & y_{m+1}\\\vdots & \vdots & \ddots & \vdots\\y_{n-m} & y_{n-m+1} & \cdots & y_{n-1}\end{bmatrix},Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_{n-m}\end{bmatrix}=\begin{bmatrix}y_{m+1}\\y_{m+2}\\\vdots\\y_n\end{bmatrix} \tag{13}$$

由以上时间序列预测的描述可以看出,时间序列数据及其预测具有以下特点:

(1) 时间序列预测具有方向性,较早的数据会影响它之后的数据,反之则不成立.这就是说,在对时间序列进行 SVR 预测建模时,我们只能使用较早的数据预测较晚的数据,用后边的数据预测前边的数据是没有意义的.

(2) 在时间序列中,预测点的值受它之前的若干历史数据影响,但这些历史数据对它的影响是不均衡的,离

*图 1~图 3 是在典型时间序列数据集D[11]上进行的关于 3 个参数对预测精度影响的实验.

预测点较近的历史数据的影响较大,对于预测数据来说较重要;而离预测点较远的历史数据的影响较小,对于预测数据来说相对不重要.因此,在对时间序列进行 SVR 建模时,应对不同时刻的历史数据给予不同的权重.

4 SVR 模型的参数选择

4.1 传统的交叉验证方法

传统的交叉验证将样本数据分成n个同样大小的子集,依次选出连续的$n-1$ 个子集作为训练子集,测试那个没有参加训练的子集.n次选择以后,就会进行n次训练和测试,n个子集中的每一个都将作为测试子集被预测,也就是说,全部数据中的每个样本点都被预测了一遍.在LIBSVM中使用了一种洗牌方法[12],即先将所有的训练数据顺序全部打乱后重新排列,然后再分成n个同样大小的子集进行交叉验证.

交叉验证通常与网格搜索相结合.所谓网格搜索,就是将待确定的若干参数(限定在一定范围内)按均匀的间隔取出若干个值,然后将这些参数的所有取值加以组合,形成若干组参数组合,每一组参数组合都将参加交叉验证.

传统的交叉验证与网格搜索相结合进行参数选择,首先对C,ε和$gamma$进行网格搜索,然后将搜索的结果执行交叉验证过程,即首先选定C,ε和$gamma$的范围(例如$C=2^{-1},2^{0},\ldots,2^{12}$;$gamma=2^{-8},2^{-7},\ldots,2^{0}$;$\varepsilon=2^{-12},2^{-11},\ldots,2^{-1}$).然后,对应每一组参数组合($C$,$\varepsilon$,$gamma$)进行交叉验证,如图 4 所示,把全部训练数据分成 5 个同样大小的子集(带有“test”字样的子集为测试子集,其他的子集作为训练子集),分别用子集 1,2,3,4 预测子集 5;用子集 2,3,4,5 预测子集 1;用子集 3,4,5,1 预测子集 2;用子集 4,5,1,2 预测子集 3;用子集 5,1,2,3 预测子集 4.这样,共进行了 5 次训练和测试,如果称每次训练和测试为一次交叉验证,则图 4 共执行了 5 次交叉验证过程,每个子集均作为测试子集被预测了一次.对于每一组参数组合(C,ε,$gamma$),都执行这样 5 次交叉验证,相应地,就会得出 5 个测试子集的预测结果,求出 5 次预测误差的平均值,最后选择预测误差平均值最小的那一组参数组合(C,ε,$gamma$)作为支持向量回归模型的参数,即,参数选择的目标函数可表示为

$$\min_{C,\varepsilon,gamma}\frac{1}{n}\sum_{i=1}^{n}E_i \tag{14}$$

其中,n为交叉验证过程中整个训练集所划分的子集数,E_i为每个子集的平均预测误差.

交叉验证与网格搜索相结合的方法将每一组参数组合在每个子集上均预测一遍,取其平均误差最小者,由于对每一组参数的测试均遍历了整个数据集,每个样本点都被预测了一遍,这样选择出的参数对全部数据的适应性较好,准确率是比较稳定的.

1	2	3	4	5 test
1 test	2	3	4	5
1	2 test	3	4	5
1	2	3 test	4	5
1	2	3	4 test	5

Fig.4 Traditional cross-validation

图 4 传统的交叉验证

但是,传统的交叉验证方法没有考虑到时间序列的特点,对于训练子集和测试子集的选择没有考虑时间序列数据的先后顺序,因此,传统的交叉验证比较适合处理不带有时间特征的问题,如分类、函数拟合、概率密度估计等.对于时间序列预测,采用传统的交叉验证进行 SVR 参数选择不仅计算量较大,而且结果也不尽如人意,第 5.2 节的实验结果可以说明这一点.

4.2 面向时间序列预测的SVR参数选择方法

为了获得预测精度较高的模型参数,本文根据时间序列预测的特点,对传统的交叉验证方法和支持向量回

归模型进行了改进,将改进的交叉验证与一种ε-加权的 SVR 模型结合起来,主要体现在两个方面:

(1) 为保证时间序列预测的方向性,同时保证对有限样本的充分学习,对传统的交叉验证进行了改进;

(2) 根据时间序列预测的特点(2),对传统的支持向量回归模型进行改进,使得不敏感参数ε由固定变为随时间序列可调整,以此保证时间序列预测的精度.

4.2.1 面向时间序列预测改进的交叉验证

由图 4 可以看出,在传统的交叉验证中,会出现用后发生的事件预测前边事件的情况,显然,由于时间序列预测具有方向性,这在逻辑上是不合理的.例如,在图 4 中,使用子集 2,3,4,5 预测子集 1 是不恰当的,第 5.2 节的实验结果也证明了这一点.本文提出一种改进的交叉验证方法,改进主要体现在两个方面:

(1) 只用前面的子集测试后面的子集,保证时间序列预测的方向性;

(2) 采用(1)中的方式改进交叉验证似乎只能用子集 1~子集 4 测试子集 5.为了获得更优的模型参数以提高预测精度,本文进一步提出对有限的样本进行充分的学习,因此,不仅要使用子集 1,2,3,4 测试子集 5,还要以子集 1,2,3 测试子集 4,以子集 1,2 测试子集 3,以子集 1 测试子集 2,如图 5 所示,取 4 次测试的平均误差最小的一组参数作为最终的模型参数.这样,既保证了时间序列预测的方向性,又使得样本集中尽可能多的样本均被测试一遍,以此来保证选择的参数对全部数据的有效性,提高基于所选参数进行预测的精度和稳定性.

1	2	3	4	5 test
1	2	3	4 test	
1	2	3 test		
1	2 test			

Fig.5 Improved cross-validation

图 5 改进的交叉验证

4.2.2 面向时间序列预测的ε-加权方法

在传统的 SVR 模型中,不敏感损失ε是固定的.对于时间序列预测,本文提出一种ε-加权方法用于 SVR 参数选择.根据时间序列预测的特点(2),预测点的值受它之前的若干历史数据影响,但这些历史数据的影响是不均衡的,靠近预测点的历史数据的影响较大,而远离预测点的历史数据的影响较小.因此,离预测点较近的数据应该被较多地关注.在公式(5)中,离预测点较近的历史数据具有较大的松弛因子,因而,所对应的不敏感损失ε的值应该较小,反之,离预测点较远数据所对应的ε值应该较大.因此,在式(9)表达的目标函数中,不敏感损失ε应该随时间序列而不断地动态调整,这符合时间序列预测的特点,有助于提高预测的准确性.

按照上边的讨论,式(9)中的二次优化问题可转换成下面的形式:

$$\begin{aligned} \max \quad & \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\varepsilon_i\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j) \\ \text{subject to} \quad & 0\le\alpha_i,\alpha_i^*\le C \quad (i=1,2,...,n) \\ & \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0 \end{aligned} \tag{15}$$

本文借鉴Jordan神经网络和EWMA模型对历史数据的处理方式[13],对不同样本对应的参数ε_i(i=1,2,…,n)按下式进行调整:

$$\varepsilon_i=(1-d)^i\varepsilon_0 \quad (i=1,2,...,n) \tag{16}$$

其中,ε_0为初始选定的参数,d是变化因子(介于 0 与 1 之间),与ε_1相对应的样本x_1是离预测点最远的数据,与ε_n相对应的样本x_n是离预测点最近的数据.

5 面向时间序列预测的 SVR 参数选择方法的有效性

5.1 实验方案设计

本文将第 4 节介绍的两种方法结合起来,根据时间序列的特点,对传统的交叉验证进行改进,同时在 SVR 模型中沿着时间序列对不敏感损失ε进行动态调整.为了验证所提方法的有效性,本文设计了 5 组实验方案.

方案 1. 为传统的交叉验证-网格搜索方法.其中设计了 2 个实验,分别对应加入洗牌和取消洗牌的情况.

方案 2. 在子集 1~4 的基础上预测子集 5,包含 2 个实验,分别对应ε-加权和ε-不加权的情况.这两个实验都体现了时间序列预测的方向性,可显示出ε-加权的 SVR 的效果,如图 6 所示.

Experiment 1 (ε-weighted)	1	2 weighted	3 weighted	4 weighted	5 test
Experiment 2 (ε-not-weighted)	1	2	3	4	5 test

Fig.6 Two experiments of scheme 2

图 6 方案 2 的 2 个实验

方案 3. 在方案 2 的基础上增加了一次训练和预测,即用子集 1~3 预测子集 4,包含ε-加权和ε-不加权两种方式,分别对应 2 个实验,如图 7 所示.参数的选择基于子集 5 和子集 4 的平均预测误差.

Experiment 1 (ε-weighted)	1 weighted	2 weighted	3 weighted	4 weighted	5 test
	1 weighted	2 weighted	3 weighted	4 test	

Experiment 2 (ε-not-weighted)	1	2	3	4	5 test
	1	2	3	4 test	

Fig.7 Two experiments of scheme 3

图 7 方案 3 的 2 个实验

方案 4. 在方案 3 的基础上增加一次训练和预测,即用子集 1、2 预测子集 3,包括ε-加权和ε-不加权两种方式,分别对应 2 个实验,如图 8 所示.参数的选择基于子集 5、子集 4 和子集 3 的平均预测误差.

Experiment 1 (ε-weighted)	1 weighted	2 weighted	3 weighted	4 weighted	5 test
	1 weighted	2 weighted	3 weighted	4 test	
	1 weighted	2 weighted	3 test		

Experiment 2 (ε-not-weighted)	1	2	3	4	5 test
	1	2	3	4 test	
	1	2	3 test		

Fig.8 Two experiments of scheme 4

图 8 方案 4 的 2 个实验

可以看出,除方案 1 体现传统的交叉验证-网格搜索方法外,其余 4 个方案均保证了时间序列的方向性,且从方案 2 到方案 5 逐渐对样本进行更充分的学习,体现了第 4.2.1 节中提出的思想,方案之间实验结果的比较可以验证第 4.2.1 节中所提方法的有效性.在每个方案内部,均对比了ε-加权和ε-不加权两种方式,可以验证第 4.2.2 节中所提方法的有效性.综合比较各个方案以及每个方案内部的实验结果,可以验证本文所提出的面向时间序列预测的 SVR 参数选择方法的有效性.

方案 5. 在方案 4 的基础上增加一次训练和预测,即用子集 1 预测子集 2,同样包括ε-加权和ε-不加权两种方式,对应 2 个实验,如图 9 所示.参数的选择基于子集 5、子集 4、子集 3 和子集 2 的平均预测误差.

Experiment 1 (ε-weighted)	1 weighted	2 weighted	3 weighted	4 weighted	5 test
	1 weighted	2 weighted	3 weighted	4 test	
	1 weighted	2 weighted	3 test		
	1 weighted	2 test			

Experiment 2 (ε-not-weighted)	1	2	3	4	5 test
	1	2	3	4 test	
	1	2	3 test		
	1	2 test			

Fig.9 Two experiments of scheme 5

图 9 方案 5 的 2 个实验

5.2 实验过程、结果及分析

在Sinc函数的仿真数据上比较第 5.1 节所述的 5 组实验方案.Sinc函数是一个典型的时间序列,在很多文献中被广泛地用于回归验证[4].Sinc函数定义为

$$f(t)=\mathrm{sinc}(t+a)+b \tag{17}$$

这里,不失一般性,可设置a=−4,b=0.5.在函数区间$t\in[0,399]$上均匀地取 400 个数据点,构造回归步长为 7 的训练样本 300 个,分别对后边 60 个和 93 个数据进行预测.使用支持向量回归进行预测建模之前,要先确定模型的参数.在网格搜索中,设定C,ε,$gamma$的范围为$C=2^{-1},2^{0},\ldots,2^{12}$;$gamma=2^{-8},2^{-7},\ldots,2^{0}$;$\varepsilon=2^{-12},2^{-11},\ldots,2^{-1}$.

为了评价预测的准确性,采用下面的平均绝对百分比误差 Mape(mean absolute percentage error)作为模型评价指标:

$$\mathrm{Mape}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i-t_i^*}{t_i}\right| \tag{18}$$

其中,n 是被预测的数据数量,t_i 是时间点 i 处的真实值,t_i^* 是该点处的预测值.

第 5.1 节所述的 5 种实验方案的结果见表 1,其中,C,$gamma$ 和ε的取值分别代表每种方案确定的模型参数值,Mape 60 和 Mape 93 分别为 60 个和 93 个预测数据的 Mape 误差,不同方案之间及其内部的预测精度对比如图 10 所示.

从实验结果可以看出:由于时间序列所具有的方向性特征,方案 1 中采用传统的交叉验证方法选择模型参数并不适合(不论是否采取洗牌方式).尤其是洗牌方式,完全打乱了时序,预测误差较大;取消洗牌方式所对应的实验结果虽然较好,但是它选择的参数 C 值过大(是方案 5 加权方式中 C 值的 256 倍).由式(5)可知,这将导致对经验误差的过度惩罚,削弱模型的泛化性能,从而失去 SVR 本身的优势.

由图 10 和表 1 可以明显地看出,方案 2~方案 5 的预测精度逐步提高.

对于方案 2,可以看出预测结果并不好,因为只是简单地用前 4 个子集预测第 5 个子集来确定模型参数,并没有对已有的训练数据进行充分的学习,从而导致训练不足,造成预测结果偏差比较大.

与方案 5 相比,方案 3、4 也存在着和方案 2 一样的问题,但是随着预测子集数量的增加,在训练过程中学习到更多的信息,所以预测精度也随之提高.

方案 5 对样本进行了最充分的学习,从实验结果来看效果也最好.而且惩罚因子 C 没有明显增大,最大也只是 8.因而,避免了方案 1 中取消洗牌方式出现的过学习现象.

从表 1 和图 10 可以看出,在所有带ε-加权方式的 4 组方案中,有 3 组方案ε-加权方式是有效的,只有方案 2 中ε-加权的结果不理想,这是由于决定模型参数的预测子集只有一个(子集 5),训练过程中学习的信息量过少造成的.方案 5 中的实验 1 不仅对样本进行了充分的学习,而且在保证预测方向性的基础上,ε被随时间序列动态调整,因而在具有好的泛化性能的同时,获得了最高的预测精度,从而验证了本文所提出的参数选择方法对于时间

序列预测的有效性.

Table 1 Results of five groups of experimental schemes

表 1 5 组实验方案的结果

		C	*gamma*	ε	$Mape_{60}$	$Mape_{93}$
方案 1	洗牌方式	64	0.125	0.031 25	0.010 338	0.010 47
	取消洗牌方式	2 048	0.007 812 5	0.000 244 141	0.000 559	0.000 561
方案 2	ε-加权	0.5	0.003 906 25	0.062 5	0.035 033	0.035 892
	ε-不加权	2	0.25	0.015 625	0.020 572	0.020 648
方案 3	ε-加权	0.5	0.003 906 25	0.007 812 5	0.011 637	0.011 135
	ε-不加权	2	0.003 906 25	0.031 25	0.014 871	0.015 499
方案 4	ε-加权	0.5	0.015 625	0.000 976 563	0.006 581	0.006 171
	ε-不加权	1	0.007 812 5	0.001 953 125	0.006 948	0.006 693
方案 5	ε-加权	8	0.015 625	0.003 906 25	0.001 189	0.001 144
	ε-不加权	1	0.031 25	0.003 906 25	0.002 341	0.002 238

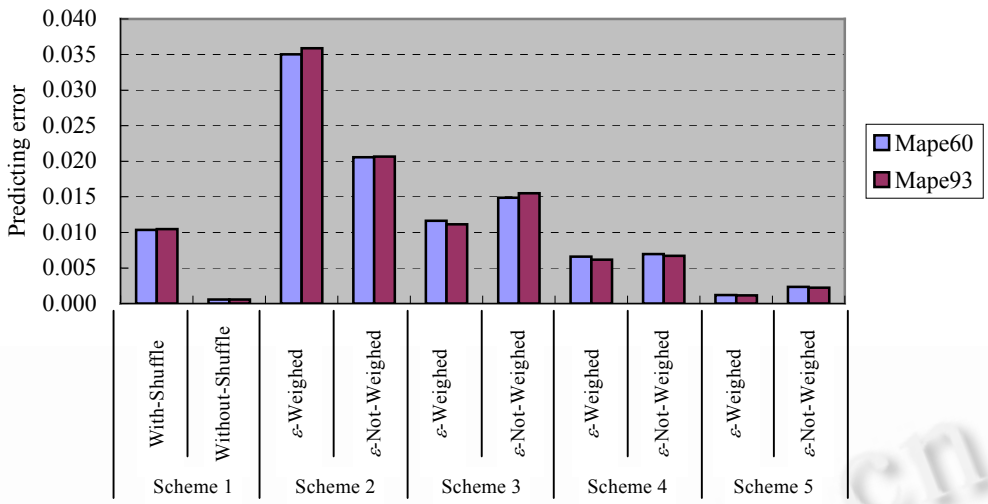


Fig.10 Comparison of predicting errors among five groups of experimental schemes

图 10 5 组实验方案预测误差的比较

6 结 论

本文提出了一种面向时间序列预测的 SVR 参数选择方法,为了获得好的模型参数,根据时间序列及其预测的特点,改进了传统的交叉验证方法,在保证时序数据方向性的同时,充分挖掘有限的训练样本所包含的信息;且在参数选择的过程中,对 SVR 模型中的不敏感损失ε随时间序列进行动态调整,从而在保持模型泛化性的前提下,克服了传统的交叉验证方法不适用于时间序列预测模型参数选取或过拟合的缺陷.通过精心设计实验,验证了本文所提出的 SVR 参数选择方法在处理时间序列预测问题上的有效性.

References:

[1] Vapnic VN. The Nature of Statistical Learning Theory. 2nd ed., New York: Springer-Verlag, 1995. 123−169.

[2] Xu JH, Zhang XG. Statistical Learning Theory. Beijing: Publishing House of Electronics Industry, 2004. 293−363 (in Chinese).

[3] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 121−167.

[4] Du SX, Wu TJ. Support vector machines for regression. Journal of System Simulation, 2003,15(11):1580−1585 (in Chinese with English abstract).

[5] Chuang CC, Su SF, Jeng JT, Hsiao CC. Robust support vector regression networks for function approximation with outliers. IEEE Trans. on Neural Networks, 2002,13(6):1322−1330.

[6] Keerthi SS. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. IEEE Trans. on Neural Networks, 2002,13(5):1225−1229.

[7] Zhang ZS, Li LJ, He ZJ. Research on parameter optimization of fault classifier based on support vector machine. Journal of Xi'an Jiaotong University, 2003,37(11):1101−1109 (in Chinese with English abstract).

[8] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameter for support vector machines. Machine Learning, 2002,46(1):131−159.

[9] Xu L, Li CP. Multi-Objective parameters selection for SVM classification using NSGA-Ⅱ. In: Perer P, ed. ICDM 2006. LNAI 4065, Berlin: Springer-Verlag, 2006. 365−376.

[10] Li GZ, Wang M, Zeng HJ. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Beijing: Publishing House of Electronics Industry, 2004. 82−107 (in Chinese).

[11] Wakuya H, Shida K. Time series prediction by a neural network model based on bi-directional computation style: A study on generalization performance with the computer-generated time series: "Data Set D". System and Computers in Japan, 2003,34(10): 64−74.

[12] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Pham DT, Karaboga D. Training Elman and Jordan networks for system identification using genetic algorithm. Artificial Intelligence in Engineering, 1999,13(2):107−117.

[14] Li LJ, Zhang ZS, He ZJ. Research on condition trend prediction of mechanical equipment based on support vector machine. Journal of Xi'an Jiaotong University, 2004,38(3):230−234 (in Chinese with English abstract).

附中文参考文献

[2] 许建华,张学工.统计学习理论.北京:电子工业出版社,2004.293−363.

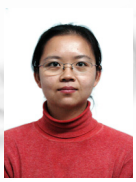
[4] 杜树新,吴铁军.用于回归估计的支持向量机方法.系统仿真学报,2003,15(11):1580−1585.

[7] 张周锁,李凌均,何正嘉.基于支持向量机故障分类器的参数优化研究.西安交通大学学报,2003,37(1):1101−1109.

[10] 李国正,王猛,曾华军.支持向量机导论.北京:电子工业出版社,2004.82−107.

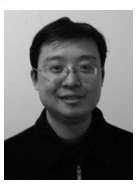
[14] 李凌均,张周锁,何正嘉.基于支持向量机的机械设备状态趋势预测研究.西安交通大学学报,2004,38(3):230−234.

林树宽(1966－),女,吉林长春人,博士,副教授,主要研究领域为时序数据挖掘,人工智能,机器学习.



张少敏(1980－),女,博士生,主要研究领域为模式识别.

徐传飞(1984－),男,硕士生,主要研究领域为时序数据挖掘.



支力佳(1977－),男,博士生,主要研究领域为模式识别.

乔建忠(1964－),男,博士,教授,博士生导师,主要研究领域为并行,分布式计算.

于戈(1962－),男,博士,教授,博士生导师,主要研究领域为数据库.