LS-SVM 的参数优选及铁路客运市场预测

周辉仁,郑丕谔
ZHOU Hui-ren,ZHENG Pi-e

天津大学 系统工程研究所,天津 300072
Institute of Systems Engineering,Tianjin University,Tianjin 300072,China

ZHOU Hui-ren,ZHENG Pi-e.Method for selecting parameters of LS-SVM and forecasting market of railway passenger traffic.Computer Engineering and Applications,2007,43(30):206-208.

Abstract: In Least Squares Support Vector Machines (LS-SVM),a least squares cost function is proposed so as to obtain a linear set of equations in dual space.Through GA,hyper-parameters selection can be solved.The model is then used to forecast the market of railway passenger traffic.It is shown that the hierarchical genetic algorithm proposed is simple and effective.

Key words: Least Squares Support Vector Machines (LS-SVM);Genetic Algorithm (GA);optimization of hyper-parameters;time series prediction

摘 要:提出通过建立验证性能指标用遗传算法优化最小二乘支持向量机的有关参数并进行时间序列预测。将最小二乘支持向量机以铁路客运市场数据进行训练和测试,并与传统的 BP 网络预测模型相比较,结果证明,该模型的预测精确度是令人满意的,提出的方法是可行的。

关键词:最小二乘支持向量机;遗传算法;参数优化;时间序列预测
文章编号:1002-8331(2007)30-0206-03 **文献标识码:**A **中图分类号:**TP181

1 引言

正确预测铁路客运市场，对国家的经济格局和资源配置，以及对铁路内部的投资结构、经营管理等都有重要作用。铁路客运市场受多个因素的影响,但是每个因素对其作用的函数关系又很难界定。因此,铁路客运市场预测属于复杂的非线性系统问题[1]。人工神经网络模拟人的大脑活动,具有极强的非线性逼近、大规模并行处理、自训练学习、自组织和容错能力等优点,用其可进行铁路客运市场预测。但是,目前理论上很难求得网络结构的最佳值,从而其泛化性能很难控制。

Vapnik 在 1995 年提出一种新型统计学习方法。支持向量机(Support Vector Machines,SVM),支持向量机具有完备的统计学习理论基础和出色的学习性能,已成为机器学习界的研究热点,并在很多领域都得到了成功地应用[2,3]。近年,SuyKens J.A.K 提出最小二乘支持向量机方法(Least Squares Support Vector Machines,LS-SVM)[4,5],这种方法采用最小二乘线性系统作为损失函数,求解过程变成了解一组等式方程,求解速度相对加快,并应用到模式识别和非线性函数估计中,取得了较好的效果。

本文首先阐述了最小二乘支持向量机的算法，并提出应用遗传算法优化有关参数，最后采用最小二乘支持向量机对铁路客运市场进行建模,并与神经网络预测结果进行比较,表明最小二乘支持向量机进行的预测精度是比较高，速度是比较快的。

2 最小二乘支持向量机

训练数据的样本可以表示为:$\{x_k,y_k\}$，其中,$\boldsymbol{x}_k$ 是 n 维输入向量,y_k 是一维输出标量。在特征空间中支持向量机模型为:

$$y_k=\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b \tag{1}$$

其中,非线形映射 $\varphi(\cdot)$将输入数据映射到高维特征空间。

用作函数逼近的最小二乘支持向量机,其优化问题为:

$$\min_{w,e} J(w,e)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2 \tag{2}$$

约束条件:

$$y(x)=\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b+e_k,\ \ k=1,\cdots,N \tag{3}$$

最小二乘支持向量机优化问题对应的拉格朗日函数为:

$$L(w,b,e,\alpha)=J(w,e)-\sum_{k=1}^{N}\alpha_k*\{\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b+e_k-y_k\} \tag{4}$$

其中,α_k 为拉格朗日乘子。

拉格朗日函数分别对 w,e_k,b,α_k 求导可得:

$$\frac{\partial L}{\partial W}L=0\rightarrow w=\sum_{k=1}^{N}\alpha_k\varphi(x_k) \tag{5}$$

$$\frac{\partial L}{\partial e_k}=0\rightarrow\alpha_k=\gamma*e_k \tag{6}$$

$$\frac{\partial L}{\partial b}=0\rightarrow\sum_{k=1}^{N}\alpha_k=0 \tag{7}$$

$$\frac{\partial L}{\partial \alpha_k}=0\rightarrow\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b+e_k-y_k=0 \tag{8}$$

其中,$k=1,\cdots,N$。

作者简介:周辉仁(1972-),男,博士研究生,研究方向为神经网络理论及应用、系统管理智能优化与决策等;郑丕谔(1942-),男,教授,博士生导师,研究方向为神经网络理论及其应用、系统管理智能优化与决策等。

由式(5)-式(8)可得如下线形等式:

$$\begin{bmatrix} 0 & \mathbf{1}_N^{\mathrm{T}} \\ \mathbf{1}_N & \Omega+\boldsymbol{I}_N/\gamma \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{9}$$

其中,$\boldsymbol{\Omega}\in R^{N\times N}$,$\boldsymbol{\Omega}_{ij}=K(x_i,x_j)$,$i,j=1,\cdots,N$;$\boldsymbol{I}_N\in R^{N\times N}$,是单位矩阵;$\boldsymbol{I}_N\in R^N$,是元素为1的向量。

因此,对一个新点 x^* 预测函数 $\hat{f}$ 为:

$$\hat{f}(x^*)=\sum_{i=1}^{N}\hat{\alpha}_iK(x_i,x^*)+\hat{b} \tag{10}$$

其中,$\hat{\alpha}$,$\hat{b}$ 可由式(9)唯一求出;核可以取径向基函数核

$$K(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|_2^2}{\sigma^2}\right) \tag{11}$$

σ 为径向基函数的宽度,为待定参数。

3 γ 和 σ 参数的遗传算法选择

对于调整参数 γ 和径向基函数的宽度 σ 的选择,通过建立一种性能指标,设定调整参数 γ 和径向基函数的宽度 σ 的取值范围,利用遗传算法的全局搜索能力进行选择 γ 和 σ。

3.1 验证性能指标

在支持向量机回归情况下,为了优化 γ 和 σ,将初始的含有 N 个样的总训练样本集分为 L 个子样本集,这 L 个子样本集的交集为空集。分别取每个子样本集作为验证样本集,在取每个子样本集作为验证样本时,I 中剩余的其它样本作为训练样本,设验证样本集有 n 个验证样本$\{x_j^v,y_j^v\}$,对于 $l=1,\cdots,L$,每个子样本集验证性能指标可用如下公式:

$$V_l=\frac{1}{n\sum_{j=1}^{n}(y_j^v-\hat{f}_\gamma(x_j^v))^2}=\frac{1}{n\sum_{j=1}^{n}\left(y_j^v-\begin{bmatrix}\frac{1}{\boldsymbol{\Omega}^v}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0 & \mathbf{1}_N^{\mathrm{T}}\\ \mathbf{1}_N & \Omega+\boldsymbol{I}_N/\gamma\end{bmatrix}^{-1}\begin{bmatrix}0\\ y\end{bmatrix}\right)^2} \tag{12}$$

其中,$\boldsymbol{\Omega}^v\in R^{(N-n)\times n}$,$\boldsymbol{\Omega}_{i,j}^v=K(x_i,x_j^v)$,$i=1,\cdots,N-n$,$j=1,\cdots,n$。

最后取总的验证性能指标为:

$$\min_{\gamma,\sigma}V=\frac{V_1+V_2+\cdots+V_L}{L} \tag{13}$$

通过对式(13)的优化即可求得调整参数 γ 和径向基函数的宽度 σ。

3.2 遗传算法设计

3.2.1 群体规模选择

合适的群体规模对遗传算法的收敛具有重要意义。群体太小难以求得满意的结果,群体太大则计算复杂。根据经验,群体规模一般取10~160。

3.2.2 适应度函数的设计

遗传算法中采用适应度函数值来评估个体性能并指导搜索,基本不用搜索空间的知识,因此适应度函数的选取相当重要。根据式(12)所示的验证性能指标,采取如下适应度函数:

$$f(x)=\frac{1}{\sum_{j=1}^{n}\left(y_j^v-\begin{bmatrix}\frac{1}{\boldsymbol{\Omega}^v}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0 & \mathbf{1}_N^{\mathrm{T}}\\ \mathbf{1}_N & \Omega+\boldsymbol{I}_N/\gamma\end{bmatrix}^{-1}\begin{bmatrix}0\\ y\end{bmatrix}\right)^2} \tag{14}$$

其中,$\boldsymbol{\Omega}^v\in R^{(N-n)\times n}$,$\boldsymbol{\Omega}_{i,j}^v=K(x_i,x_j^v)$,$i=1,\cdots,N-n$,$j=1,\cdots,n$。

3.2.3 选择与复制

适值越大的个体被选择的概率也越大。个体 j 的选择概率为:

$$P_i=\frac{f_i}{\sum_{k=1}^{M}f_k}$$

其中,M 表示种群规模,f_i 表示个体 i 的适应度。个体 i 被复制的个数 $R_p(i)=M\times P_i$。从初始化种群中经过选择与复制形成一个子群 $P_1(t)$。

3.2.4 交叉与变异

由于参数基因采用实值编码,为保证交叉后产生新的参数值,并开辟出新的搜索空间,参数基因的交叉操作采用线性组合方式,将两个基因串对应交叉位的值相组合生成新的基因串。交叉在遗传操作中起核心作用,交叉概率较大可增强遗传算法开辟新搜索空间的能力,但性能好的基因串遭到破坏的可能性较大,算法收敛速度降低,且不稳定;若交叉概率较小,则遗传算法搜索可能陷入迟钝状态。

对参数基因,可采用偏置变异,以一定的概率给变异位基因加一个从偏置区域中随机选取的数值。变异在遗传操作中属于辅助性的搜索操作,主要目的是维持群体的多样性,较低的变异概率可以防止群体中重要的单一基因丢失,但降低了遗传算法开辟新搜索空间的能力;较高的变异概率将使遗传操作趋于纯粹的随机搜索,降低了算法的收敛速度和稳定性。一般根据具体问题,变异概率取0.001~0.01之间的值。

4 仿真和比较

在文献[1]提出了BP神经网络在铁路客运市场时间序列预测中的应用,本文对该文献中的数据用遗传算法选择 γ 和 σ,并用训练好的最小二乘支持向量机进行时间序列预测并将结果与BP神经网络结果进行比较。

1985年至2000年铁路客运量的原始时间序列数据(单位:万人)为:

[X]=[11 210 108 579 112 479 122 645 113 807 95 712
95 080 99 693 105 458 108 738 102 745 94 796
93 308 95 085 100 164 105 073]

以连续5年的数据用递阶遗传算法训练过的BP神经网络来预测第6年的数据,按(11 210 108 579 112 479 122 645 113 807;95 712),(108 579 112 479 122 645 113 807 95 712;95 080),(112 479 122 645 113 807 95 712 95 080;99 693),…,依次类推。前10组数据作为学习样本,最后一组数据为测试数据。将前10组数据分为10个子样本集作为验证样本集,即 $L=10$。

本文采用Matlab进行编程构成最小二乘支持向量机,其算法步骤如下:

(1)对输入数据进行归一化处理。

(2)对运行参数进行设置。取种群大小 $N=100$;进化最大代数为500;交叉概率 P_c 为0.55;变异概率 P_m 为0.005。

(3)随机生成 N 个染色体作为初始种群,采用实数编码。

(4)对每个染色体解码按式(14)计算每个染色体的适应度。

(5)按其适应度采用赌轮盘选择法选择和复制个体,生成新的种群。

(6)对种群进行遗传操作。

(7)判断是否满足最大进化代数或停止准则,若满足则转

到步骤(8);若不满足则返回步骤(4)。

(8)对每个染色体解码,构造最小二乘支持向量机,评价最小二乘支持向量机性能。

(9)利用训练好的最小二乘支持向量机进行预测并对预测结果进行反归一化。

用遗传算法迭代 100 次,所得验证误差为 0.017 5,所得 γ=162.620 4,σ^2=1.482 6,所得验证误差函数曲线如图 1 所示。

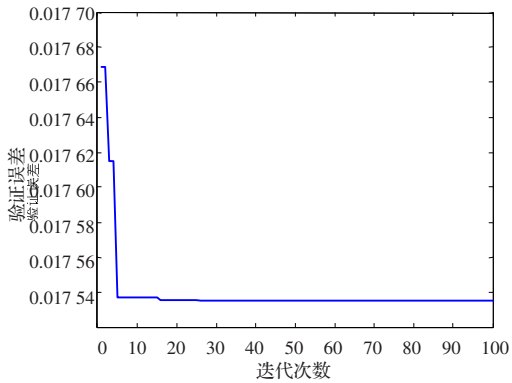

图 1 验证误差函数曲线

将求得 γ=162.620 4,σ^2=1.482 6,代入式(9),求得和,再将和及前 10 组数据代入式(10)便得到预测函数,将前 10 组数据代入预测函数便得训练均方误差为 2.180 4e-4,将第 11 组数据代入预测函数便得到预测结果,将该结果反归一化后就是 2000 年铁路客运市场的预测结果为 105 499.919 0。所有反归一化后的训练及预测结果如表 1 所示;最终预测结果与文献[1]中的 BP 预测结果比较如表 2 所示;反归一化后的输出值和统计值曲线如图 2 所示。

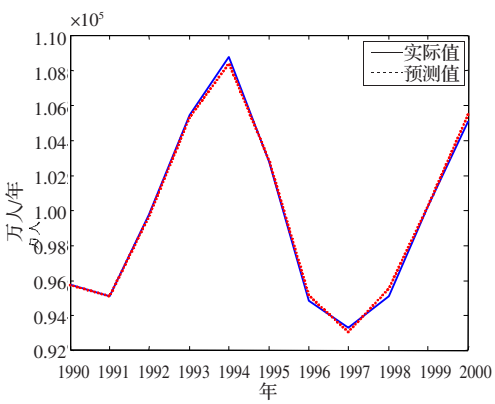

图 2 反归一化后的输出值和统计值曲线

5 结束语

本文尝试了用最小二乘支持向量机回归的方法在铁路客运市场中的预测,首先用遗传算法确定支持向量机的最佳参数,进而建立起基于时间序列的预测模型,从预测结果可以看出,该方法用于铁路客运市场预测具有更高的精度,相对误差比用 BP 网络预测小的多。该方法在训练过程中,所需时间短,用遗传算法优化有关参数能有效的避免过拟合和欠拟合的现象,具有很强的泛化能力。

与神经网络预测相比,支持向量机不需要确定隐节点个数,而对于神经网络这一直是理论上难以解决的问题。最小二乘支持向量机在确定输入输出数据后,经遗传算法确定调整参数后,其连接权重和阈值由解线性等式方程确定,并且存在唯一解。最小二乘支持向量机采用结构风险最小化原则,综合考虑了样本误差和模型复杂度,而大部分未改进的神经网络仅考虑样本误差的最小化,所以最小二乘支持向量机如果有关调整参数选取的恰当具有很好的泛化能力。

(收稿日期:2007 年 3 月)

表 1 LS-SVM 训练及预测结果

时间	统计值(实际值)	输出值(预测值)	相对误差
1990	95 712	95 699.163 6	-0.000 134
1991	95 080	95 097.586 4	0.000 18
1992	99 693	99 638.810 0	-0.000 565
1993	105 458	105 288.728 1	-0.001 672
1994	108 738	108 397.511 5	-0.003 299
1995	102 745	102 809.323 8	0.000 671
1996	94 796	95 138.150 4	0.003 726
1997	93 308	93 013.158 6	-0.003 244
1998	95 085	95 486.020 3	0.004 327
1999	100 164	100 210.547 2	0.000 505
2000	105 073	105 499.919 0	0.004 063

表 2 LS-SVM 与 BP 网络预测结果比较

	学习误差	预测值(万人)	相对误差
LS-SVM	2.180 4e-4	105 499.919 0	0.004 063
BP 网络	0.004 014	101 957.741 2	-0.029 649

参考文献:

[1] 侯福均,吴祈宗.BP 神经网络在铁路客运市场时间序列预测中的应用[J].运筹与管理,2003(4):73-75.

[2] Vapnik V,Levin E,Le C Y.Measuring the VC-dimension of a learning machines[J].Neural Compution,1994(6):851-876.

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

[4] Suykens J A K,Vandewall J.Least squares support vector machine classifiers[J].Neural Processing Letters,1999,9(3):293-300.

[5] Pelckmans K,Suykens J A K,De Moor B.Building sparse representations and structure determination on LS-SVM substrates neurocomputing[J].Special Issue,2005,64:137-159.

(上接 78 页)

计算机模拟结果显示,基于自然梯度的语音盲分离改进算法能够有效地分离随机混合的自然语音信号,但也存在一定的缺陷,在混合矩阵 $\boldsymbol{A}$ 中每列元素的绝对值互相相等时,本文提出的改进算法也无能为力对其进行分离,但发生这种情况的机率较小,因此本文提出的基于自然梯度的语音盲分离改进算法在对语音进行盲分离时还是较为有效的。

(收稿日期:2007 年 3 月)

参考文献:

[1] 孙守宇,郑君里,吴德伟.基于自然梯度算法的盲信源分离研究[J].空军工程大学学报:自然科学版,2003,4(3):50-54.

[2] 孙守宇,郑君里,赵敏,等.不同幅度通信信号的盲源分离[J].通信学报,2004,25(6):132-138.

[3] 张明键,韦刚.一种信号源盲分离的神经网络算法[J].信号处理,2003,19(2):149-152.

[4] 牛奕龙.盲源分离算法研究[D].西安:西北工业大学,2005.