第 39 卷　第 3 期
Vol.39　No.3
山 东 大 学 学 报 (工 学 版)
JOURNAL OF SHANDONG UNIVERSITY (ENGINEERING SCIENCE)
2009 年 6 月
Jun. 2009

文章编号:1672-3961(2009)03-0007-04

一种支持向量机参数选择的改进分布估计算法

王雪松,程玉虎,郝名林
(中国矿业大学信息与电气工程学院, 江苏 徐州 221116)

摘要:支持向量机(support vector machine,SVM)的学习性能和泛化能力在很大程度上取决于参数的合理设置. 将支持向量机的参数选择问题转化为优化问题,以模型预测均方根误差为评价函数,提出一种引入混沌变异操作的改进分布估计算法(estimation of distribution algorithm,EDA),并将其用于优化求解 ε-支持向量机的参数:惩罚因子、不敏感损失系数以及高斯径向基核函数的宽度. 由于改进 EDA 利用混沌运动的随机性和遍历性等特点在解空间内进行优化搜索,能够较好解决传统 EDA 易于陷入局部极小的缺陷. Chebyshev 混沌时间序列预测仿真结果表明:改进 EDA 是选取 SVM 参数的有效方法.

关键词:支持向量机;参数选择;混沌变异;分布估计算法
中图分类号:TP18　　**文献标志码:**A

Parameters selection of a support vector machine using an improved estimation of the distribution algorithm

WANG Xue-song, CHENG Yu-hu, HAO Ming-lin
(School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: The learning performance and the generalization property of support vector machines (SVMs) are greatly influenced by the suitable setting of some parameters. The parameters selection can be transformed into an optimization problem by defining the root mean square error of a SVM prediction model as an evaluation function. A kind of improved estimation of the distribution algorithm (EDA) with a chaotic-mutation operation was proposed and used to optimize parameters of a ε-SVM including a penalty factor, an insensitive loss coefficient and a width of a Gaussian kernel function. The improved EDA could take advantage of the randomness and ergodicity of chaos, which could solve the local minima problem of traditional EDAs. Simulation result of the prediction of a Chebyshev chaotic time series showed that the improved EDA was an effective method of solving the problem for parameters selection of a SVM.

Key words: support vector machine; parameters selection; chaotic-mutation; estimation of distribution algorithm

0 引言

20 世纪 90 年代由 Vapnik 提出的支持向量机是一种典型的核机器学习方法,是借助于最优化方法解决机器学习问题的新工具[1]. SVM 能够较好地解决小样本、非线性和高维数的问题;其训练是一个凸二次规划问题,能够保证找到的极值解就是全局最优解. 一些学者认为,SVM 正成为继神经网络之后新的研究热点,开始成为克服维数灾难、过学习及局部最优等传统困难的有效手段,已在模式识别、信号处理、最优控制、系统建模等领域得到了广泛的应用[2]. SVM 具有优良的学习性能和泛化能力,但是其性能在很大程度上取决于参数的合理选择. 传统的参数选取方法主要有:经验选择法、实验试凑法、梯度下降法、交叉验证法、Bayesian法等. 经验选择

收稿日期:2009-05-10
基金项目:国家自然科学基金资助项目(60804022);教育部新世纪优秀人才支持计划资助项目(NCET-08-0836);高等学校博士学科点专项科研基金资助项目(20070290537,200802901506);江苏省自然科学基金资助项目(BK2008126)
作者简介:王雪松(1974-),女,安徽泗县人,副教授,博士,主要从事智机器学习研究. E-mail: wangxuesongcumt@163.com

要求对所研究问题拥有很好的经验和十足的知识,否则,并不容易获得合适的参数. 实验试凑是通过大量的实验来获得较优的参数,比较费时,而且获得的参数也不一定是最优的. Chapelle 等采用梯度下降法来完成 SVM 参数的选择[3],虽然在计算时间上得到了明显的改善,但是梯度下降法对初始点要求较高,而且是一种线性搜索法,因此极易陷入局部极小. 交叉验证是目前应用较为普遍的一种 SVM 参数选取方法,且易于实现,缺点是计算量大,尤其对于大样本问题. Bayesian 法需要一定的参数空间的先验知识,而且每次确定最优参数均需要多次迭代,与交叉验证法相比,在计算量与计算复杂度上优势不大,实现起来并不容易. 支持向量机的参数选择问题,其实质就是一个优化问题. 因此,随着遗传算法(genetic algorithm, GA)、粒子群优化(particle swarm optimization, PSO)、人工免疫(artificial immune, AI)等智能优化方法在过程控制、经济预测、工程优化等领域取得的成功,陆续有一些学者对如何应用这些方法来优化选择支持向量机的参数进行了研究[4-5]. 最近几年,进化计算领域兴起了一类新型优化算法,即分布估计算法,并迅速成为进化计算领域的研究热点和解决工程问题的有效方法. 分布估计算法提出了一种全新的进化模式,是统计学习理论与随机优化算法的结合[6]. 与 GA、PSO 和 AI 等传统进化算法不同,EDA 是基于对整个群体建立数学模型,直接描述整个群体的进化趋势,是对生物进化“宏观”层面上的数学建模,具有良好的全局搜索能力,但容易陷入局部最优解,出现早熟收敛现象. 通过分析发现,在各种遗传操作算子中,变异算子用新的基因值替换原有基因值,从而可以改变个体编码串的结构,是维持种群多样性,防止出现早熟收敛的重要手段;另外,传统分布估计算法仅仅通过选择算子和基因池重组算子来实施进化,缺少维持种群多样性的变异算子. 为此,在传统 EDA 中引入混沌变异操作,提出一种改进分布估计算法,并将其用于解决支持向量回归机的参数选择问题.

1 支持向量机的参数选择问题描述

支持向量机回归问题可理解为:根据一组独立同分布输入输出数据集 $T=\{(\boldsymbol{x}_q,\boldsymbol{y}_q),\boldsymbol{x}_q\in\mathbf{R}^m,\boldsymbol{y}_q\in\mathbf{R},q=1,2,\cdots,M\}$,其中 $\boldsymbol{x}_q$ 是第 q 个 m 维输入向量,y_q 是第 q 个输出变量,M 为训练样本个数,求取输入与输出之间的关系.

支持向量机回归中,首先要将输入向量映射到高维特征空间,然后在特征空间中构造优化超平面[1]:

$$f(\boldsymbol{x})=\boldsymbol{w}\phi(\boldsymbol{x})+b, \tag{1}$$

式中,$\boldsymbol{w}$ 为权向量,b 为偏置项,$\phi(\boldsymbol{x})$为映射函数,将样本集从输入空间非线性地映射到高维特征空间. 定义 $K(\boldsymbol{x}_q,\boldsymbol{x}_p)=\phi(\boldsymbol{x}_q)\cdot\phi(\boldsymbol{x}_p)$为满足 Mercer 定理的核函数. 目前,应用较多的核函数有线性核、多项式核、高斯径向基核、Sigmoid 核和 Fourier 级数核等. 与其它类型核函数相比,高斯径向基核具有模型选择简单、计算难度小、计算效率高以及算法易于实现的优点. 因此,在实际使用过程中倾向于选择高斯径向基核函数.

$$K(\boldsymbol{x}_q,\boldsymbol{x}_p)=\phi(\boldsymbol{x}_q)\cdot\phi(\boldsymbol{x}_p)=\exp\left(-\frac{\|\boldsymbol{x}_q-\boldsymbol{x}_p\|^2}{2\sigma^2}\right), \tag{2}$$

式中,σ 为高斯径向基核函数的宽度.

根据结构风险最小化原理,综合考虑模型复杂度和分类误差,回归问题可转化为如下优化问题:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{q=1}^{M}(\xi_q+\xi_q^*)$$
$$\text{s.t.}\begin{cases} y_q-\boldsymbol{w}\phi(\boldsymbol{x}_q)-b\leqslant\varepsilon+\xi_q;\\ \boldsymbol{w}\phi(\boldsymbol{x}_q)+b-y_q\leqslant\varepsilon+\xi_q^*;\\ \xi_q=\begin{cases}0, & D(\boldsymbol{x}_q,y_q)-\varepsilon\leqslant 0;\\ D(\boldsymbol{x}_q,y_q)-\varepsilon, & \text{其他};\end{cases}\\ \xi_q^*=\begin{cases}0, & \varepsilon-D(\boldsymbol{x}_q,y_q)\leqslant 0;\\ \varepsilon-D(\boldsymbol{x}_q,y_q), & \text{其他}.\end{cases}\end{cases} \tag{3}$$

式中,C 为惩罚因子,ξ_q 和 ξ_q^* 是不小于零的松弛变量,ε 表示实际输出 y 与 SVM 估计输出 $f(\boldsymbol{x})$之间的残差 $|D(\boldsymbol{x},y)|$ 的上限值,因此,式(3)称为 ε-支持向量机.

此处考虑采用高斯径向基核函数的 ε-支持向量机,需要确定的参数包括惩罚因子 C、不敏感损失系数 ε 以及高斯径向基核函数的宽度 σ.

2 支持向量机参数选择的改进分布估计算法

混沌运动能在一定范围内按其自身的“规律”不重复地遍历所有状态[7]. 改进分布估计算法就是利用混沌变量的随机性和遍历性等特点在解空间内进行优化搜索,能够较好地解决传统 EDA 易于陷入局部极小的缺陷,从而提高进化速度. 具体算法步骤为:

Step 1 根据具体求解问题,确定 ε-支持向量机参数的搜索空间.

Step 2 初始化种群 $\boldsymbol{X}=(C,\sigma,\varepsilon)$. 将待优化

问题解在空间中的坐标做为种群中各个独立的个体,采用十进制编码,如:

$$X_{j,G}=(x_{j,1,G},\cdots x_{j,i,G},\cdots,x_{j,n,G}),\tag{4}$$

式中,$i=1,2,\cdots,n$;$j=1,2,\cdots,\mathrm{NP}$;$G=1,2,\cdots,G_{\max}$. n 为需要优化的参数个数,即为待优化问题的维数,此处 $n=3$,NP 为种群规模,$G_{\max}$为最大进化代数.

为使初始种群不重复地随机遍历整个解空间,采用一维 Logistic 模型,具体步骤如下:

(1) 随机产生一组[0,1]之间的初始值,$\boldsymbol{a}_0=\mathrm{rand}(1,\ n)$;

(2) 将 n 维向量 $\boldsymbol{a}_0$ 代入一维 Logistic 映射模型 $\boldsymbol{a}_{t+1}=\lambda\boldsymbol{a}_t(1-\boldsymbol{a}_t)$,其中 λ 为控制参数. 给定 λ 值,经 NP 次迭代计算就可得到在[0, 1]范围内的混沌时间序列 $\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{\mathrm{NP}}$. 令 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{\mathrm{NP}}]^{\mathrm{T}}$,则 $\boldsymbol{A}$ 为 $\mathrm{NP}\times n$ 维矩阵;

(3) 将混沌时间序列的取值范围从[0, 1]扩展到待优化问题的取值范围,初始种群的第 j 个个体的第 i 个基因位可表示为

$$x_{j,i,1}=\mathrm{low}+(\mathrm{high}-\mathrm{low})\boldsymbol{A}(j,i),\tag{5}$$

式中,low 为解空间的下限值,high 为解空间的上限值.

Step 3　评价初始种群. 以 ε-支持向量机均方根误差 RMSE 作为模型性能的评价函数,如式(6)所示,求函数的极小值:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{q=1}^{M}(y_q-\hat{y}_q)^2}{M}}.\tag{6}$$

Step 4　适应度分配. 将评价函数值按照降序排列,即评价函数值最大(最小适应度)的个体放在列表的第一个位置,评价函数值最小的个体放在位置 NP 上. 那么,种群中第 j 个个体的适应度值 fit(j)可根据它在列表中排序的位置 position(j)来确定,即

$$\mathrm{fit}(j)=2-\mathrm{sp}+2(\mathrm{sp}-1)\frac{\mathrm{position}(j)-1}{\mathrm{NP}-1},\tag{7}$$

式中,position(j)为在区间[1, NP]内取值的整数,选择压差 sp = 2.

Step 5　执行混沌变异,以概率 1 对每个个体进行混沌变异. 为了变异后得到较好的解,在变异时考虑个体的适应度值,对于较好的个体在较小的范围内变异,则第 j 个个体的第 i 个基因的变异半径可表示为

$$r(j,i)=S(i)\frac{1-\mathrm{fit}(j)}{\max\limits_{j'=1,2,\cdots\mathrm{NP}}\mathrm{fit}(j')},\tag{8}$$

式中,$S(i)$为变异步长. 根据进化过程中种群本身的信息来确定步长,则步长可表示为

$$S(i)=\frac{\max\limits_{j'=1,2,\cdots\mathrm{NP}}(x_{j',i,G})-\min\limits_{j'=1,2,\cdots\mathrm{NP}}(x_{j',i,G})}{\mathrm{DIV}},\tag{9}$$

式中,DIV 为一常数. 个体 $x_{j,i,G}$经混沌变异后可表示为

$$x'_{j,i,G}=x_{j,i,G}+r(j,i)*2(1-2\boldsymbol{A}(j,i)).\tag{10}$$

Step 6　评价经混沌变异操作后的种群,如果新个体对应的评价函数值较小,则用其代替旧个体.

Step 7　用分布估计算法产生下一代种群,算法步骤如下:

(1) 从种群 $\boldsymbol{X}$ 中选出 np(np < NP)个较优的个体,对所选样本进行统计分析,建立单变量高斯模型;

(2) 按建立的单变量高斯统计模型产生 $k*\mathrm{NP}$ ($k>1$)个新的个体;

(3) 评价新产生的所有个体,并从新产生的个体中选出 NP 个个体作为下一代种群.

为保持种群的多样性,不仅按个体适应度值的大小选择下一代种群,同时考虑个体的浓度. 因而,个体的选择概率取决于个体的适应度值和浓度,高适应度值和低浓度个体的选择概率大,该个体受到促进. 下一代种群产生策略如下:

① 计算种群中各个体与其他个体的距离之和,则第 s 个个体与其他个体的距离之和为

$$d(s)=\sum_{l=1}^{k*\mathrm{NP}}\sqrt{(\boldsymbol{X}_{s,G}-\boldsymbol{X}_{l,G})^2};\tag{11}$$

② 对求得的 $d(s)$归一化处理得 $d'(s)=\dfrac{d(s)}{\max\limits_{l=1,2,\cdots,k*\mathrm{NP}}d(l)}$;

③ 求得个体的浓度 $\rho(s)=\dfrac{1}{d'(s)}$;

④ 根据个体的浓度调整个体的适应度值 $\mathrm{fit}'(s)=\dfrac{\mathrm{fit}(s)}{\rho(s)}$,按调整后的个体适应度值从大到小对个体排序,并从中选择 NP 个作为下一代种群.

Step 8　判断是否满足收敛条件,如果不满足则转至 Step 4 继续执行. 此处采用的迭代终止条件是设定最大进化代数 $G_{\max}$.

3　仿真研究

考虑一 3 阶 Chebyshev 混沌时间序列:

$$y(t+1)=\cos(3\arccos(y(t))),\ -1\leqslant y_0\leqslant 1.\tag{12}$$

本仿真的目的是用 t 时刻的值预测未来 $t+1$ 时刻的值. 抽取 $t=0$ 到 $t=400$ 共 400 个数据,其中前 200 个数据用于 ε-SVM 的训练数据集,剩下的数据作为 ε-SVM 的检验数据.

分别考查参数 C、ε 和 σ 对 ε-SVM 的均方根误差和训练时间的影响,确定这 3 个参数的选择范围分别为 $C\in[900,1\,200]$、$\varepsilon\in[1.0E\text{-}6,1.0E\text{-}3]$和$\sigma\in[0.01,0.1]$. 应用改进分布估计算法进行支持向量机参数优化,将均方根误差作为优化算法的目标函数,误差越小则参数越好. 改进分布估计算法的参数设置为: $G_{\max}=30$、NP = 30、$\lambda=3.9$、np = 15、$k=1.5$、DIV = 15. 优化后 ε-SVM 的参数值为: $C=1\,093.2$、$\varepsilon=1.0E\text{-}6$、$\sigma=0.043\,318$.

应用优化得到的参数,建立 ε-SVM 模型对 Chebyshev 混沌时间序列在 $t\in[201,400]$区间内的数据进行预测,得到的测试均方根误差为 $6.5854E\text{-}4$,支持向量数为 116. 图 1 给出了模型预测值和真实值,其中"+"表示真实值,"o"表示预测值,这两者之间的误差曲线如图 2 所示. 由图可以看出,基于改进 EDA 优化参数建立的 ε-SVM 模型能够以较高精度预测 Chebyshev 混沌序列,预测误差达到 10^{-5} 数量级,具有良好的泛化性能.

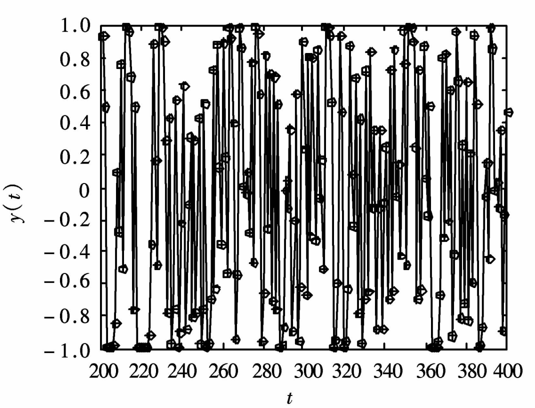

图 1 Chebyshev 混沌时序预测结果

Fig.1 Prediction result of the Chebyshev chaotic times series

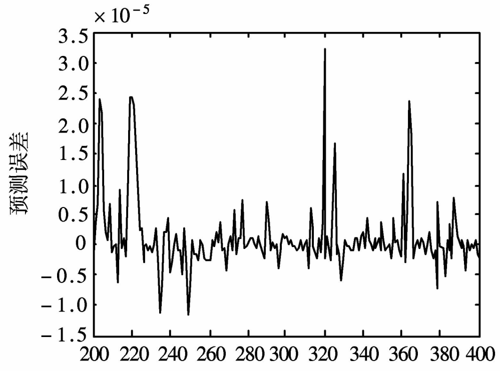

图 2 Chebyshev 混沌时序预测误差

Fig.2 Error prediction of the Chebyshev chaotic times series

4 结论

以采用高斯径向基核函数的 ε-支持向量机为例,定义模型预测均方根误差为评价函数,采用改进分布估计算法优化求解惩罚因子、不敏感损失系数以及高斯径向基核函数的宽度. Chebyshev 混沌时间序列预测仿真结果表明,经改进 EDA 优化构造的 ε-支持向量机的预测精度较高. 另外,值得一提的是,本文虽然是以 ε-支持向量机的参数选择问题为例说明,但是所提改进分布估计算法同样适用于其他类型的支持向量机,为解决支持向量机的参数选择问题提供了新的求解方法.

参考文献:

[1] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer Verlag, 1995.

[2] WANG X S, TIAN X L, CHENG Y H. Value approximation with least squares support vector machine in reinforcement learning system[J]. Journal of Computational and Theoretical Nanoscience, 2007, 4(7/8): 1290-1294.

[3] CHAPELLE O, VAPNIK V, BOUSQUET O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1): 131-160.

[4] WU C H, TZENG G H, LIN R H. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression[J]. Expert System with Applications, 2009, 36(3): 4725-4735.

[5] 邵信光, 杨慧中, 陈刚. 基于粒子群优化算法的支持向量机参数选择及其应用[J]. 控制理论与应用, 2006, 23(5): 740-744.

SHAO Xinguang, YANG Huizhong, CHEN Gang. Parameters selection and application of support vector machines based on particle swarm optimization algorithm[J]. Control Theory & Applications, 2006, 23(5): 740-744.

[6] LARRANAGA P, LOZANO J A. Estimation of distribution algorithms: a new tool for evolutionary computation[M]. Boston: Kluwer Academic Publishers, 2002.

[7] 贾东立, 张家树. 基于混沌变异的小生境粒子群算法[J]. 控制与决策, 2007, 22(1): 708-712.

JIA Dongli, ZHANG Jiashu. Niche particle swarm optimization combined with chaotic mutation[J]. Control and Decision, 2007, 22(1): 708-712.

(编辑:陈燕)