

基于佳点集遗传算法的支持向量机的参数选择

孙浩, 陶亮

(安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039)

摘要:支持向量机(Support Vector Machine, SVM)的参数选择一直缺乏很完善的方法,很大程度上限制了它的应用。为了获得较好的SVM参数,提出了基于佳点集遗传算法的参数选择方法,利用佳点集遗传算法对遗传算法中的交叉操作进行了重新设计,减少了遗传算法的收敛时间并且提高了遗传算法的精度,从而确保了SVM参数选择的准确性。通过数值实验表明由该方法所得的支持向量机可以在一定程度上自动地选择参数,具有一定的推广意义。

关键词:支持向量机;参数选择方法;遗传算法;佳点集遗传算法

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2009)08-0086-03

Parameter Selection Algorithm for Support Vector Machines Based on Good Point Set Based Genetic Algorithm

SUN Hao, TAO Liang

(Ministry of Education Key Laboratory of Intelligent Computing & Signal Processing,
Anhui University, Hefei 230039, China)

Abstract: There have been no perfect algorithms for the selection of the optimal parameters of support vector machines (SVMs), therefore, the applications of SVMs are limited. In order to get optimal SVM parameters, a parameter selection algorithm for support vector machines is given based on good point set based genetic algorithm (GA). The crossover operation in GA is redesigned by using the principle of good point set based genetic algorithm, which can reduce the convergence time, improve the precision of GA, and insure the accuracy of parameter selection. The experiment results show that the proposed algorithm of parameter selection can make the optimal parameter selection automatic and has certain practical application significance.

Key words: support vector machine; parameter selection algorithm; genetic algorithm; good point set based genetic algorithm

0 引言

支持向量机是 Vapnik^[1,2]等人根据统计学习理论提出的一种新的机器学习方法,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,如泛化性能好、无需先验知识等,并能够推广应用到函数拟合等其它机器学习问题中,因此近年来 SVM 的研究受到越来越多的重视。

支持向量机的最大特点就是根据 Vapnik 的结构风险最小化原则,尽量提高学习的泛化能力,即由有限的训练集样本得到的小误差仍能能够保证对独立的测试集小的误差。另外,由于支持向量机算法是一个凸优

化问题,因此局部最优解一定是全局最优解,可防止过学习,这些特点是神经网络等其它学习算法所不及的。

对于分类问题,支持向量机的算法可简述为将输入空间中的样本通过某种非线性函数关系映射到一个特征空间中(维数可能较高),使两类样本(可推广到多类样本)在此特征空间中线性可分,并寻找样本在此特征空间中的最优线性分类超平面。其判别函数为:

$$f(x) = \text{sgn}\left(\sum_{i=1}^l a_i y_i K(x, x_i) + b\right)$$

其中, $K(\cdot, \cdot)$ 是核函数,核函数 $K(\cdot, \cdot)$ 只要满足 Mercer^[3] 条件即可满足上述要求。目前常用的核函数主要有三类:

(a) 多项式核函数: $K(x, x_i) = [(x \cdot x_i) + 1]^q$;

(b) 径向基核函数: $K(x, x_i) = \exp\left\{-\frac{\|x - x_i\|}{2\sigma^2}\right\}$;

(c) 感知机核函数: $K(x, x_i) = \tanh(v(x \cdot x_i) + c)$, 其中 $v > 0, c < 0$ 。

收稿日期:2008-11-19;修回日期:2009-01-13

基金项目:国家自然科学基金项目(60375010);安徽大学人才队伍建设项目和创新团队基金

作者简介:孙浩(1983-),男,硕士研究生,研究方向为图像处理与模式识别;陶亮,教授,博士生导师,研究方向为模式识别、信息处理、多维信号处理等。

对于具体的分类问题,关键是求出 Lagrange 乘数 α_i , 实际上为如下的二次规划问题:

$$\begin{aligned} W(\alpha) &= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s. t. } \sum_{i=1}^N y_i \alpha_i &= 0 \\ 0 \leq \alpha_i &\leq C \\ i &= 1, 2, \dots, N \end{aligned}$$

但是,和其他学习算法一样,SVM 的性能依赖于学习机的参数,而且其参数的选取对于经验的依赖性比较强,到目前为止,还没有 SVM 参数选择的好方法。

1 参数对支持向量机的影响

作为分类器的支持向量机,其泛化性能取决于核函数及其参数的选择,Vapnik 等人在研究中^[2]发现,不同的核函数对 SVM 性能的影响不大,反而误差惩罚参数 C 是影响 SVM 性能的关键因素。

误差惩罚参数 C 的作用是在确定的特征空间中调节学习机器的置信范围和经验风险的比例以调节学习机器的推广能力。它的选取一般是由具体的问题而定,并取决于数据中噪声的数量。在确定的特征子空间中, C 的取值小表示对经验误差的惩罚小,学习机器的复杂度小,分类面较简单而经验风险值较大; C 的取值大则表示对经验误差的惩罚较大,学习机器的复杂度大,分类面较复杂,经验误差相对较小,但这时的学习机器的推广能力则较差;如果 C 取无限大时,则所有的约束条件都必须满足,这意味着训练样本必须要准确地分类。每个特征子空间至少存在一个合适的 C ,使得 SVM 推广能力最好。当 C 超过一定值时,SVM 的复杂度达到了特征子空间允许的最大值,此时经验风险和推广能力几乎不再变化。

2 参数选择方法简介

2.1 传统的参数选择方法

在模型选择以后,首先为常数 C 赋初始值,然后开始实验测试,根据测试精度直到达到要求为止。这种方法是目前比较常用且行之有效的方法,但是基本是凭经验调整,缺乏足够的理论依据,对不同的核函数,不同的样本,其调整方法可能不同,因此在参数调整过程中带有一定的盲目性,且当需要调整幅度较大时,调整次数较多,实验比较复杂^[4]。

2.2 基于遗传算法的参数选择方法

遗传算法^[5](Genetic Algorithm, GA)是一种通过模拟自然进化过程搜索最优解的方法,利用生物遗传

学的观点,结合了适者生存和随机信息交换的思想,通过自然选择、交换、变异等作用机制,实现种群的进化。在寻优过程中,在解空间随机产生多个起始点并同时开始搜索,由适应度函数来指导搜索方向,能够在复杂搜索空间快速寻求全局优化解。

在一些最新研究^[6,7]中,在支持向量机的参数选择中引入了遗传算法,可以在一定范围内快速地得出最佳参数,从而明显提高支持向量机的泛化性能。但是,由于 GA 本身的原因,可能会过早收敛得到的是局部最优解或者是迭代时间过长,从而无法获得最佳性能。

3 基于佳点集遗传算法的支持向量机的参数选择方法

佳点集遗传算法^[8](Good point set based Genetic Algorithm, GGA)是由张铃、张钊提出的,是利用数论中的佳点集理论和方法,对遗传算法中的交叉操作进行了重新设计,不仅具有收敛速度更快,迭代精度更高的优点,而且避免了遗传算法常用的早期收敛的现象。

3.1 佳点集应用于交叉操作

设 $A = a_1 a_2 \cdots a_L$ 与 $B = b_1 b_2 \cdots b_L$ 交叉,得到 $C = c_1 c_2 \cdots c_L$, 其中

$$c_k = \begin{cases} a_k & (a_k = b_k) \\ * & (a_k \neq b_k) \end{cases}$$

设“*”个数为 s ,用于佳点集 G_s 中:

在 s 维空间中含有 n 个点的佳点集中, $P_n(i) = \{(\{r_1 \cdot i\}, \{r_2 \cdot i\}, \dots, \{r_s \cdot i\}), 1 \leq i \leq n\}$, 其中,取 $r_k = \{2 \cos \frac{2\pi k}{p}\}, 1 \leq k \leq s, p$ 是满足 $(p-3)/2 \geq s$ 的最小素数,则 r_k 为佳点;或取 $r_k = \{e^k\}, 1 \leq k \leq s$, 则 r_k 为佳点, $\{a\}$ 表示 a 的小数部分。若 $\{a\} < 0.5$, 取 0, 反之,取 1。

3.2 基于佳点集遗传算法的参数选择算法

设 Ch 为足够大的数; N 为种群规模;支持向量机的样本识别率为适应度函数 $F(t)$ 。

步骤 1: 迭代次数 $t = 0$;

步骤 2: (种群初始化) 在 $[0, Ch]$ 上用“赌轮法”产生 N 个随机数作为初始种群;

步骤 3: 以 $F(t)$ 计算适应度函数值;

步骤 4: 若最优个体所对应的适应度函数值满足要求或达到设定的迭代次数,转到步骤 10;

步骤 5: $t = t + 1$;

步骤 6: (选择算子) 采用锦标赛选择和最优保存策略产生 N 个新个体作为子代;

步骤 7: (交叉算子) 采用 3.1 的佳点集方法对个

体进行两两交叉操作;

步骤 8:(变异算子) 对每个个体按概率进行变异操作;转入步骤 4;

步骤 9:给出最佳的参数,并用其训练数据集以获得最佳模型。

4 数值实验

为了验证文中所提出的算法的有效性,实验数据来源于 UCI 公共数据库^[9]的两个数据集,分别为:ala 数据集和 Wine 数据集。由于 Wine 数据集有三类数据,在本实验中只取对应于标号 1 和标号 2 的两类数据点用于数值试验,从训练集中随机选取 10% 的点作为测试集,其余的点组成训练集;而 WPBC 只有两类数据,也从训练集中随机选取 10% 的点作为测试集;而 ala 数据集也只有两类数据,而且还有另外的测试集,所以直接使用即可。所有数值实验都是在 Intel (R) CPU 3. 0G, 1GM RAM 的兼容机上进行的。

为了测试计算出来的正则参数 C 的性能,将本方法选择出来的 C 和 C 取缺省值 1.0 的结果、还有基于传统遗传算法选择出来的 C 进行了比较,用测试正确率来评价算法的性能。

数值试验结果详见表 1 和表 2。

表 1 三种方法选择的参数 C 的实验结果

数据集	$C = 1.0$	识别率	$C (GA)$	识别率	$C (GGA)$	识别率
ala	1.0	84.05%	3.595	84.37%	3.605	84.39%
WPBC	1.0	80.09%	1.738	81.42%	1.732	81.39%
Wine	1.0	96.84%	0.186	98.46%	0.185	98.45%

表 2 GA 与 GGA 的迭代次数比较

数据集	$C (GA)$	迭代次数	$C (GGA)$	迭代次数
ala	3.595	500	3.605	300
WPBC	1.738	500	1.732	300
Wine	0.186	500	0.185	300

从表 1 可以看出,与一般传统算法相比,采用佳点集遗传算法选择 SVM 参数可以提高数据样本的识别

率,其性能与传统 GA 相当。

从表 2 可以看出,采用佳点集遗传算法选择 SVM 参数比传统 GA 的迭代次数要少很多,因而提高了算法的速度。

5 结束语

文中提出了一种运用佳点集遗传算法对支持向量机的参数进行选择的方法,可以自动地为 SVM 选择合适的参数,省去了人工更改 SVM 参数的麻烦,通过数值实验表明,与传统的 GA 相比,该方法不仅具有收敛速度快、迭代次数少等优点,可以提高算法的速度,而且可以获得比较精确结果,具有较强的泛化能力。这种方法也可以应用于其他类型支持向量机的参数优化,具有一定的推广价值。

参考文献:

- [1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer,1995.
- [2] Vapnik V N. Statistical Learning Theory[M]. New York: Wiely,1998.
- [3] Courant R, Hilbert D. Methods of Mathematical Physics[M]. New York: Wildyinterscience,1953.
- [4] 王 睿. 关于支持向量机的参数选择方法分析[J]. 重庆师范大学学报:自然科学版,2007,24(2):36-38.
- [5] 李敏强,寇纪松,林 丹,等. 遗传算法的基本理论与应用[M]. 北京:科学出版社,2002.
- [6] 刘 胜,李妍妍. 自适应 GA-SVM 参数选择算法研究[J]. 哈尔滨工程大学学报,2007,28(4):398-402.
- [7] 杜京义,侯媛彬. 基于遗传算法的支持向量回归机参数选取[J]. 系统工程与电子技术,2006,28(9):1430-1433.
- [8] 张 铃,张 钺. 佳点集遗传算法[J]. 计算机学报,2001,24(9):1-6.
- [9] Blake C, Keogh E, Merz C J. UCI repository of machine learning databases[EB/OL]. University of California, Irvine, Department of Information and Computer Science, URL. 1998. <http://www.ics.uci.edu/mllearn/MLRepository.html>.

(上接第 85 页)

- [16] 罗永龙,黄刘生,徐维江,等. 一个保护私有信息的多边形相交判定协议[J]. 电子学报,2007,35(4):685-691.
- [17] 罗永龙,黄刘生,荆巍巍,等. 一个保护隐私的布尔关联规则挖掘算法[J]. 电子学报,2005,33(5):900-903.
- [18] 罗永龙,黄刘生,荆巍巍,等. 保护私有信息的叉积协议及其应用[J]. 计算机学报,2007,30(2):248-254.
- [19] Luo Yong-Long, Huang Liu-Sheng, Chen Guo-liang, et

al. Privacy-preserving distance measurement and its applications[J]. Chinese Journal of Electronics, 2006, 15(2):237-241.

- [20] Luo Yong-Long, Huang Liu-Sheng, Zhong Hong. Secure Two-Party Point-Circle Inclusion Problem [J]. Journal of Computer Science and Technology, 2007, 23(1):88-91.