陈 果

基于遗传算法的支持向量机分类器模型参数优化

陈 果

(南京航空航天大学 民航学院,南京 210016)

摘 要:建立在统计学习理论和结构风险最小原则上的支持向量机在理论上保证了模型的最大泛化能力,因此与建立在经验风险最小原则上的神经网络模型相比,理论上更为完善。本文运用支持向量机建立模式识别分类器模型,研究影响模型分类能力的相关参数,在分析参数对分类器识别精度的影响基础上,提出用遗传算法建立支持向量机分类器模型的参数自适应优化算法。最后,用算例表明了本文算法的正确有效性。

关 键 词:支持向量机;模式识别;遗传算法;优化

中图分类号: TH17; TP18 **文献标识码**: A **文章编号**: 1003-8728(2007)03-0347-04

Optimizing the Parameters of Support Vector Machine's Classifier Model Based on Genetic Algorithm

Chen Guo

(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: The support vector machine (SVM), which is based on the statistical learning theory (SLT) and the structural risk minimum principle, guarantees the largest generalization ability of a model. It is, therefore, theoretically more perfect than the neural network model that is based on the empirical risk minimum principle. The paper established the pattern recognition classifier model and studied the parameters that influence the classifier model's classification ability; on the basis of analyzing the parameter's influence on the classifier's recognition accuracy, it proposed the self-adaptive optimization algorithm for the SVM classifier model using genetic algorithm. Finally, calculation instances show the effectiveness of the optimization algorithm.

Key words: support vector machine (SVM); pattern recognition; genetic algorithm; optimization

由 V. Vapnik创立的支持向量机[1]建立在统计学习理论和结构风险最小的基础上,在理论上充分保证了模型的泛化能力,与神经网络相比,具有更坚实的理论基础和完善的理论体系,目前已经广泛运用于模式识别的分类器设计中[2,3]。

但是,用于分类器的支持向量机模型,本身也有许多参数要进行选择,比如惩罚因子 C以及核函数的选取及核函数的相关参数等。这些参数在一定程度上对模型的分类精度具有很大影响,且目前尚无统一选择标准。本文建立支持向量机的模式识别分类器模型,在进行参数影响分析研究的基础上,构造基于遗传算法的模型参数自适应优化算法。

1 支持向量机模式识别原理

支持向量机 SVM是从线性可分情况下的最优分类面提出的。所谓最优分类线就是要求分类线不但能将两类正确分开 (训练错误率为 0),而且使分类间隔最大,前者是保证经验风险最小 (为 0),而使分类间隔最大就是使推广性的界中置信范围最小,从而使真实风险最小,也是对推广能力的控制。推广到高维空间,最优分类线就成为最优分类面。

设线性可分样本集 $(\vec{x}_1, y_1)$, $(\vec{x}_2, y_2)$, …, $(\vec{x}_n, y_n)$, $\vec{x} \in R^d$, $y \in \{-1, +1\}$是类别标号。d维线性空间中线判别函数的一般形式为 $g(\vec{x}) = \vec{w} \cdot \vec{x} + b$,最优分类线方程为

收稿日期: 2005-12-05

作者简介:陈 果 (1972-),男 (汉),四川,副教授, cgzyx@263.net

$$\vec{w} \cdot \vec{x} + b = 0 \quad (1)$$

对样本归一化后,使得对线性可分的样本集满足 $|g(\vec{x})| \geq 1$,即使离最优分类面最近的样本的 $|g(\vec{x})| = 1$,这样分类间隔就等于 $2/\|\vec{w}\|$,因此使间隔最大等价于使 $\|\vec{w}\|$(或 $\|\vec{w}\|^2$)最小;而要求分类线对所有样本正确分类,就是要求它满足

$$y_i[(\vec{w} \cdot \vec{x}) + b] - 1 \geq 0 \quad i = 1, 2, \cdots, n \quad (2)$$

因此满足上述条件且使 $\|\vec{w}\|^2$ 最小的分类面就是最优分类面。显然,过两类样本中离分类面最近的点就是式(2)中使等号成立的那些样本,这些样本被叫做支持向量。因为他们支撑了最优分类面。

最优分类面的求解可以表示成如下约束优化问题,即在式(2)的约束下,求函数

$$\min \frac{1}{2}\|\vec{w}\|^2$$

$$\text{st} \quad y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 \quad (i = 1, 2, \cdots, n) \quad (3)$$

的最小值。在线性不可分的情况下,就是某些训练样本不能满足式(2)的条件,我们可以通过在条件中增加一个松弛项 $\xi_i \geq 0$,变为

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 + \xi_i \geq 0 \quad (4)$$

最小化 $\sum_{i=1}^{n}\xi_i$ 就可以使错分样本最小,则此时优化问题为

$$\begin{aligned} &\min\left(\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right) \\ &\text{s t} \quad y_i(\vec{w} \cdot \vec{x}_i + b) - 1 + \xi_i \geq 0 \\ &(i = 1, 2, \cdots, n), \xi_i \geq 0 \end{aligned} \quad (5)$$

式中:C为某个指定的常数,它实际上起控制对错分样本惩罚的程度的作用,实现在错分样本的比例与算法复杂度之间的折中。将原问题转化成如下对偶问题的最大值

$$\begin{aligned} &Q(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j(\vec{x}_i \cdot \vec{x}_j) \\ &\text{s t} \quad \sum_i \alpha_i y_i = 0 \quad 0 \leq \alpha_i < C \end{aligned} \quad (6)$$

求解上述问题后得到的最优分类函数如下

$$f(x) = \text{sgn}\{(\vec{w}^* \cdot \vec{x}) + b^*\} = \text{sgn}\{\sum_i \alpha_i^* y_i(\vec{x}_i \cdot \vec{x}) + b^*\} \quad (7)$$

核函数 $K(\vec{x}_i, \vec{x}_j)$ 代替最优分类面中的点积,这样就相当于把原特征空间变化换到了某一新的特征空间,此时式(6)的优化问题和式(7)的最优分类函数只需用 $K(\vec{x}_i, \vec{x}_j)$ 代替点积即可。

采用不同的核函数将导致不同的支持向量计算法,目前广泛应用的核函数形式主要有线性核函数、多项式核函数、高斯核函数、Sigmoid核函数。由于高斯核函数可以逼近任意非线性函数,因此,本文选取高斯函数作为和支持向量机的核函数来进行研究。高斯核函数为

$$K(\vec{x}_i, \vec{x}_j) = \exp\left\{-\frac{|\vec{x}_i - \vec{x}_j|^2}{\sigma^2}\right\} \quad (8)$$

2 支持向量机分类器模型参数对分类精度的影响

2.1 分类精度评价函数

由于多类分类问题可以转化为两类问题进行分析,不失一般性,本文仅仅讨论两类问题。设由 n个样本组成的训练样本集为 $(\vec{x}_{\text{Train1}}, y_{\text{Train1}})$, $(\vec{x}_{\text{Train2}}, y_{\text{Train2}})$, …, $(\vec{x}_{\text{Trainn}}, y_{\text{Trainn}})$;由 m 个样本组成的测试样本集为 $(\vec{x}_{\text{Test1}}, y_{\text{Test1}})$, $(\vec{x}_{\text{Test2}}, y_{\text{Test2}})$, …, $(\vec{x}_{\text{Testm}}, y_{\text{Testm}})$, $\vec{x} \in R^d$, $y \in \{+1, -1\}$ 是类别标号。利用训练样本集对分类器进行学习训练,而用测试样本集来对所分类器进行验证,如果测试结果的错误率越少,显然分类器的分类效果越理想,理想情况是对测试样集分类完全正确。通常衡量分类器的分类精度采用识别率 RR(recognition rate),其定义为

$$RR = \frac{\text{测试样本集中分类正确的数目}}{\text{测试样本集中测试样本总数}} \quad (9)$$

判断测试样本集中的第 $j(j = 1, 2, \cdots, m)$ 和样本分类正确与否,通常按照以下两条规则:

(1) if $y_{\text{Test}\,j} = \sum_{i=1}^{n}\alpha_i^* y_{\text{Train}\,i}K(\vec{x}_{\text{Train}\,i} \cdot \vec{x}_{\text{Test}\,j}) + b^* \geq 0$ and $y_{\text{Test}\,j} = -1$ then j样本分类错误

(2) if $y_{\text{Test}\,j} = \sum_{i=1}^{n}\alpha_i^* y_{\text{Train}\,i}K(\vec{x}_{\text{Train}\,i} \cdot \vec{x}_{\text{Test}\,j}) + b^* \leq 0$ and $y_{\text{Test}\,j} = +1$ then j样本分类错误

2.2 评价支持向量机分类精度的数据集

为了分析支持向量机分类器参数对分类精度的影响,需要选取标准的数据集进行实验。

(1) 数据集 1:圆分类问题[4]

在直角平面 xoy上,在圆 $x^2 + y^2 = 16$内定义为一类,标记为“-1”,在圆外定义为一类,标记为“+1”,随机产生 60个训练样本,其中 30个“+1”类,30个“-1”类;随机产生 800个测试样本进行分类精度测试。

(2) 数据集 2:双螺旋问题[5]

训练样本数为 200个,在直角平面 xoy上,设 $i = 0, 1, \cdots, 99$, $\theta(i) = i \times \pi/16$, $r(i) = 6.5 \times (104 - i)/104$, $x(i) = r(i) \times \sin[\theta(i)]$, $y(i) = r(i) \times \cos[\theta(i)]$。由点 $(x(i), y(i))$ $(i = 0, 1, \cdots, 99)$ 组成第一类,标记为“+1”;由点 $(-x(i), -y(i))$ $(i = 0, 1, \cdots, 99)$ 组成第二类,标记为“-1”。

测试样本数为 160个,在直角平面 XOY上,设 $i = 0, 1, \cdots, 80$, $\theta(i) = (i + 0.5) \times \pi/16$, $r(i) = 6.5 \times$

$[104 - i - 0.5]/104$, $x(i) = r(i) \times \sin[\theta(i)]$, $y(i) = r(i) \times \cos[\theta(i)]$。由点 $(x(i), y(i))$ $(i = 0, 1, \cdots, 80)$ 组成第一类,标记为“+1”;由点 $(-x(i), -y(i))$ $(i = 0, 1, \cdots, 80)$ 组成第二类,标记为“-1”。

2.3　支持向量机分类器模型参数的影响分析

支持向量机是建立在统计学习坚实的理论基础之上的,具有理论的完备性,但是在应用上,仍然存在一些问题,典型的问题就是模型参数的选择,目前,也无统一的模型选取标准和理论。在具体使用中,对分类精度有重要影响的参数是:惩罚因子 C,核函数及其参数的选取。

(1) 惩罚因子 C 用于控制模型复杂度和逼近误差的折中,C 越大则对数据的拟合程度越高。但泛化能力将降低。

对不同的类型的核函数,所产生的支持向量的个数变化不大,但是核函数的相关参数,如多项式核函数的多项式次数;对于高斯核函数的 σ 值对模型的分类精度均有重要影响。本论文选定高斯核函数,对其 σ 值进行优化研究。

表 1　惩罚因子 C 对分类精度的影响

惩罚因子 C	高斯核函数的 σ 值	数据集 1 的识别率	数据集 2 的识别率
0.1	1	81.2%	62.5%
1	1	90.2%	50.625%
10	1	87%	42.5%
100	1	87%	42.5%
1000	1	87%	42.5%

表 2　高斯核函数的 σ 值对分类精度的影响

惩罚因子 C	高斯核函数的 σ 值	数据集 1 的识别率	数据集 2 的识别率
1	0.01	22.4%	48.75%
1	0.1	54.2%	64.375%
1	1	90.2%	50.625%
1	10	62%	60%
1	100	48.2%	58.75%

表 1和表 2分别为相同的训练和测试数据集下在不同的模型参数下所得到的分类精度比较。通过比较表 1和表 2,可以看出:惩罚系数 C 和高斯核函数的 σ 值对分类结果的影响均很大。由此可见,支持向量机的优越性能需要通过合适的模型参数才能发挥出来。因此需要寻找支持向量机分类器的最佳模型参数。

3　支持向量机分类模型参数优化的自适应算法

通过上述分析发现,支持向量机模型分类精度与惩罚因子 C 和高斯核函数的 σ 均存在一定的关系,为了获取最佳分类性能的 SVM模型,需要得到最佳的 C 和 σ 值。显然这是一个优化问题,如果采取穷举的方式搜索最优值,计算量将十分巨大以至于无法实现。由于遗传算法[6] 具有隐含的并行性和强大全局搜索能力,可以在很短的时间内搜索到全局最优点。

因此。本文利用遗传算法来进行 SVM 分类模型的参数优化。首先,对 SVM 分类模型惩罚因子 C 和高斯核函数 σ 值进行二进制编码,并随机产生初始化种群;其次,对种群中的各染色体解码,获取 C 及 σ 值,运用一部分训练样本集数据训练 SVM 分类器模型,并用训练好的 SVM 分类器计算测试样本集数据的识别率 RR,根据交叉验证法原理,识别率 RR 在一定程度上反映了 SVM 模型的推广能力和分类能力,因此可以依此构造各基因串的适应度 Fitness = RR;然后判断遗传算法的停止准则是否满足,如果满足则停止计算,输出最优参数,否则,则执行选择、交叉和变异等操作以产生新一代种群,并开始新一代的遗传。

本文遗传算法中:交叉率和变异率分别为 0.50 和 0.05。基因串 (染色体) 中 C 及 σ 值均采用二进制编码,其染色体结构如图 1所示。染色体中的排列顺序为 C 及 σ,设它们的位数分别为 n_1、n_2,则染色体的长度为 $n_1 + n_2$,搜索空间为 $2^{n_1+n_2}$。图 1中 $n_1 = 2$, $n_2 = 10$。

另外,为了与算法适应,规定解码后,对于 C,通过计算 $C = 10^{C-1}$ 得到惩罚因子 C,对于 σ,通过计算 $\sigma = (\sigma + 1)/200$ 得到 σ 值。

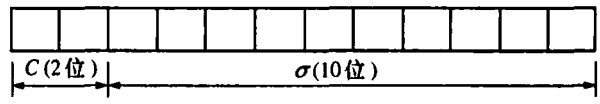

图 1　染色体结构

下面仍然针对数据集 1和 2来对本文算法进行验证。图 2为对数据 1,利用支持向量机双重学习模型得到的最优化分类器对样本进行训练所得到分类曲线;图 3为对数据 2,利用支持向量机双重学习模型得到的最优化分类器对样本进行训练所得到分类曲线。

表 3 基于支持向量机的机器双重学习模型参数及计算结果

分类数据	惩罚系数 C 编码位数	高斯核函数 σ 编码位数	惩罚系数 C 优化结果	高斯核函数 σ优化结果	错误率	识别率
数据集 1	2	10	0.1	0.65	9.2%	90.8%
数据集 2	2	10	100	0.235	0	100%

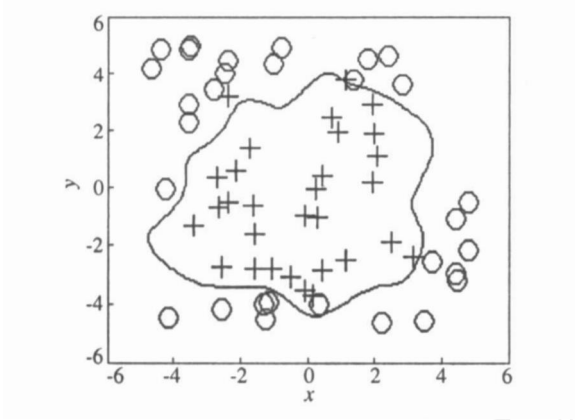

图 2 数据 1的训练样本与分类曲面

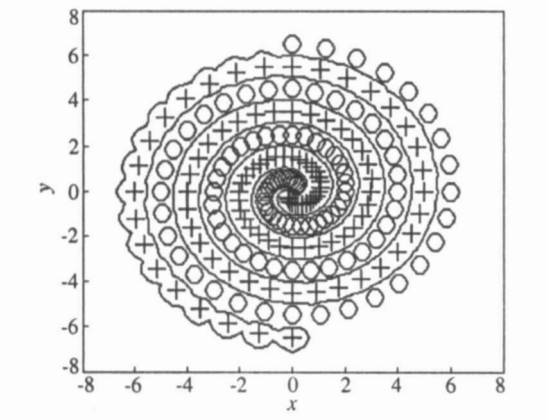

图 3 数据 2的训练样本与分类曲面

从表 3、图 2及图 3可以得出以下结论:

(1) 对数据集 1和 2,本文算法均获得了非常好的计算结果,由遗传算法优化后得到的最优 SVM 分类器模型对数据集 1和 2的识别率均达到了 90%以上,整个学习过程不需要人工干扰,大大提高了模式识别的自动化程度和精度。

(2) 对于双螺旋这样的经典分类难题,许多文献均进行了详细的研究,通常认为用常规的 BP网络很难获得好的泛化特性,Baum 等人试图用一个 2-50-1的 BP网络来解决,未能获得良好的结果[7]; Chen等人提出生成一个收缩算法来训练一个 2-20-20-1的 BP网络,经过 3000次迭代后对测试样本的识别率只有 89.6%[8];文献[9]用优化神经元激活函数的方法使识别率达到了 100%。而通过本文算法,得到了 SVM 分类器的最优参数 (C = 100, σ = 0.235),识别精度达到了 100%。整个学习过程也自动完成。需要指出的是,本文支持向量机的学习算法采用了适用于大样本数据的 SVM 快速算法 SMO (sequential minimal optimization)[4]。由此可见,基于遗传算法的支持向量机分类器具有明显优势。

4 结论

(1) 分析了运用支持向量机进行模式识别分类器的优越性以及存在的问题;

(2) 提出了影响 SVM 分类能力的两个重要参数——惩罚因子 C及高斯核函数 σ值。并对两个参数进行了分类精度影响分析。

(3) 本文用遗传算法构造了同时优化影响 SVM 分类精度的参数 (惩罚因子 C及高斯核函数的 σ值) 的算法,利用遗传算法自动获取最优的 SVM 分类器模型参数。最后用算例表明了本文算法的正确有效性。

[参考文献]

[1] Vapnik V. **The Nature of Statistical Learning** [M]. New York: Springer, 1995

[2] 张周锁, 李凌均, 何正嘉. 基于支持向量机的机械故障诊断方法研究[J]. 西安交通大学学报, 2002, 36(12): 1303～1306

[3] Ge M, Du R, Zhang C C, Xu Y S. Fault diagnosis using support vector machine with an application in metal stamping operations [J]. **Mechanical Systems and Signal Processing**, 2004, 18: 143～159

[4] 褚蕾蕾, 陈绥阳, 周梦编著. 计算智能的数学基础[M]. 北京:科学出版社, 2002

[5] 魏海坤编著. 神经网络结构设计的理论与方法[M]. 北京:国防工业出版社, 2005

[6] Goldberg D. **Genetic Algorithms in Search, Optimization and Machine Learning** [M]. Addison-Wesley, Reading, MA, 1989

[7] Baum E B, Lang K L. Constructing hidden units using examples and queries[A]. In: **Advances in Neural Information Processing System** 3[C], San Maeto. CA: Morgan Kaufmann, 1991: 904～910

[8] Chen Q C. Generating-shrinking algorithm for learning arbitary classfication[J]. **Neural Networks**, 1994, 7: 1477～1489

[9] 吴佑寿, 赵明生, 丁晓青. 一种激励函数可调的新人工神经网络及应用[J]. 中国科学(E辑), 1997, 27(1), 55～60