第28卷第2期
2009年4月

红外与毫米波学报
J. Infrared Millim. Waves

Vol. 28, No. 2
April, 2009

文章编号:1001-9014(2009)02-0141-05

非均衡数据目标识别中SVM模型多参数优化选择方法

郭 雷, 肖怀铁, 付 强
(国防科技大学 电子科学与工程学院 ATR 实验室,湖南 长沙 410073)

摘要:提出了非均衡数据目标识别中SVM模型多参数优化选择方法.首先从理论上分析了SVM模型多参数选择的内涵和必要性,针对非均衡数据的分类识别,基于F测度提出了能全面反映识别性能的多参数优化选择准则.在多参数选择过程中,利用遗传算法进行模型多参数并行优化选择.提出的方法能够寻找模型多参数的全局最优解,避免陷入梯度法常出现的局部最优解情况,同时能够克服传统方法中根据经验选择SVM单参数模型时计算量太大的不足.采用国际通用的标准数据集和雷达目标HRRP数据集进行了仿真实验,实验结果表明,该方法能够得到模型多参数的全局最优值,由此确定的SVM模型分类器性能有较大提高.

关 键 词:目标识别;非均衡数据;支持矢量机;模型优化选择
中图分类号:TP183 **文献标识码**:A

SVM MODEL OPTIMAL MULTI-PARAMETER SELECTION METHOD FOR IMBALANCED DATA TARGET RECOGNITION

GUO Lei, XIAO Huai-Tie, FU Qiang
(ATR Key Laboratory, College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: SVM model optimal multi-parameter selection method for imbalanced data recognition was proposed. First, the connotation and necessity of SVM model multi-parameter selection were theoretically analyzed. Then, a multi-parameter selection criterion based on F-measure was given, which can represent the recognition performance completely. The genetic algorithm was used in search of optimization of multi-parameter in parallel parameter optimal selection. The proposed method can get the global optimal solutions of SVM model multi-parameter, and avoid the local optimal solution caused by gradient method, and can also decrease the computational complexity of experiential selection method. The experimental results of the benchmarks and radar HRRP data sets reveal that the proposed multi-parameter selecting method can get the global optimal solution of SVM model. The recognition performance of the optimal SVM model can achieve much improvement.
Key words: target recognition; imbalanced data; support vector machine(SVM); model optimal selection

引言

支持矢量机[1](Support Vector Machine,简称SVM)是20世纪90年代Vapnik基于统计学习理论中的结构风险最小化原理而提出的一种新的模式识别方法.基于SVM的目标识别一般做法是采用非线性核函数映射的方法,将原始空间的特征数据映射到高维特征空间,使样本在此高维空间线性可分.核函数选择的好坏直接影响到目标识别性能的优劣,如何选择核函数类型以及如何确定核函数最优参数就成为研究基于SVM目标识别的关键问题之一.

很多文献研究了SVM算法及其应用[2,3].对于SVM模型选择,文献[4]研究了核函数单参数和惩罚因子的选择问题;文献[5,6]以半径/边缘为准则给出了基于梯度的模型选择方法,但是该方法敏感于初值的选取.初值不同,优化得到的模型也就不同,比较合理的解释只能是算法陷入了局部最小点.此外,这几篇文献中研究的多参数是指核函数参数

收稿日期:2008-02-27,**修回日期**:2008-09-18 **Received date**: 2008-02-27, **revised date**: 2008-09-18
基金项目:国家自然科学基金(60572138)和重点实验室基金(9140C8001020902)资助项目
作者简介:郭 雷(1980-),男,山东德州人,博士生,主要从事雷达目标识别方面的研究.

σ 与惩罚因子 C.

实际上,不同的核函数及其参数相当于定义了不同映射后的特征空间,目标在不同特征空间中分类识别结果也就不同. 对于每一维特征均有不同核参数 σ_p 的高斯核函数: $K(\mathbf{x},\mathbf{z}) = \exp\left(-\sum_{p=1}^{d}\frac{(x_p - z_p)^2}{\sigma_p^2}\right)$,以前的研究只是简单的认为 $\sigma = \sigma_1 = \cdots = \sigma_p$,然后经验选取最优值 σ. 若令核参数 σ_p 各不相同,则凭经验也无法选取,而实际上这些不同的核参数才是精确的核函数参数. 此外,对于每个特征选用不同的参数也产生了附加信息. 根据最优 σ_p 值,可以估计相应的第 p 个特征的实用性,这也是特征选择的作用. 本文研究的多参数正是指每个不同的 σ_p 和惩罚因子 C.

针对 SVM 模型多参数选择问题,本文提出了多类目标识别中非均衡数据 SVM 模型高斯核函数多参数选择方法. 首先从理论上分析了模型多参数选择的内涵和必要性及非均衡数据参数选择准则,并且基于 F 测度[7,8]出了模型多参数选择准则,在参数优化过程中采用遗传算法多参数并行优化,最后得到核函数的最优参数值,实现 SVM 模型的多参数选择,使得目标识别具有最好正确识别率与最小错误识别率.

1 SVM 模型多参数最优选择问题

1.1 SVM 模型多参数最优选择的内涵

SVM 在特征空间中构造最优分类超平面,这个特征空间是由非线性映射 $\phi(x)$ 决定的,但是一般不需要知道 $\phi(x)$ 的具体形式,因为 SVM 在特征空间中的运算只涉及到 $\phi(x)$ 的内积运算,而在特征空间中 $\phi(x)$ 的内积运算可以用核函数 $K(x,z)$ 来代替,即 $K(x,z) = \langle\phi(x),\phi(z)\rangle$.

核函数 $K(x,z)$ 需要满足如下的 Mercer 定理.

Mercer 定理[1]:令 $x \in R^d$ 和映射 $\phi(x)$,$x \to \phi(x) \in H$,其中 H 是 Hilbert 空间. 内积运算的相应表示是 $\sum_r \phi_r(x)\phi_r(z) = K(x,z)$,其中 $\phi_r(x)$ 是 x 的映射 $\phi(x)$ 的第 r 分量,$K(x,z)$ 是满足如下条件的对称函数 $\int K(x,z)g(x)g(z)\mathrm{d}x\mathrm{d}z \geqslant 0$,其中,对于任意的 $g(x)$ 有: $\int g^2(x)\mathrm{d}x < +\infty$.

由 Mercer 定理可以看出,一旦采用了适当的核函数 $K(x,z)$,就隐含定义了从原始数据空间到高维特征空间的非线性映射 $\phi(x)$. 不同类型的核函数及其参数相当于定义了不同的非线性映射 $\phi(x)$,即定义了不同的特征空间. 原始数据在不同特征空间中可分性不同,构造的分类超平面也不同. 所以 SVM 最优模型选择的实质是选择最优的特征空间(或最优非线性映射 $\phi(x)$),使得目标数据在该特征空间中能够达到最优的分类识别性能.

1.2 进行 SVM 模型多参数最优选择的必要性

本文主要研究具有代表性的高斯核函数,其一般定义是: $K(\mathbf{x},\mathbf{z}) = \exp\left\{-\frac{\|x - z\|^2}{\sigma^2}\right\}$. 之所以选择高斯核函数,主要是因为高斯核函数是 SVM 中被应用最广泛的一种核函数. 高斯核函数是一个普适的核函数,通过参数选择,它可以适用于任意分布的样本.

一般情况下,对于高斯核函数,只是简单的选择某个最优的单个核参数 σ. 实际上,目标特征矢量的各维特征在识别中所起的作用不同,即特征的重要性不同. 例如,对于不同类别目标的某一维(或几维)特征,其差别较大,那么该特征在识别中起比较重要的作用,若各维特征选择了较小的相同高斯核参数 σ,则相应的核矩阵的值比较小,这样降低了目标之间在特征空间的可分性,使得目标识别效果变差.

假设目标特征矢量为 d 维,把按一般定义的高斯核函数改写为

$$K(\mathbf{x},\mathbf{z}) = \exp\left\{-\frac{(x_1 - z_1)^2}{\sigma_1^2} - \frac{(x_2 - z_2)^2}{\sigma_2^2} - \cdots - \frac{(x_d - z_d)^2}{\sigma_d^2}\right\} = \exp\left\{-\sum_{p=1}^{d}\frac{(x_p - z_p)^2}{\sigma_p^2}\right\},$$

在该核函数中为目标每一维特征赋予不同的核参数 σ_p,这些精确的核参数可以反映目标相应特征的性质,使得不同类别目标之间的可分性更优. 此外,为每个特征选用不同的核参数也产生了附加信息,分析这些不同的核参数,也是特征选择的另一思路.

在线性不可分的情况下,SVM 模型参数 C 是对错误识别的惩罚因子. 将核矩阵 K 可以做如下调整[9]: $\mathbf{K} \leftarrow \mathbf{K} + \frac{1}{C}\mathbf{I}$,其中 $\mathbf{I}$ 为单位矩阵. 因此,C 也可以看作核函数的另外一个参数. 对于正类和负类样本数目不均衡的情况,需要选择不同的正类惩罚因子 C_+ 和负类惩罚因子 C_-,常用的选择规则是

$$\frac{C_+}{C_-} = \frac{\text{负类样本数 } n_-}{\text{正类样本数 } n_+} \quad . \tag{1}$$

2 非均衡数据目标识别 SVM 模型多参数优化选择

2.1　非均衡数据目标识别 SVM 模型多参数优化选择准则

目标识别性能的优劣常用识别率来衡量,但是如果各个类别的目标样本数量分布不均衡时,应用识别率这个指标不合适. 例如:两类分类问题,如果正类目标训练样本数量占总训练样本数量的 90%,而负类训练样本数量仅占 10%,那么根据识别率优化得到的 SVM 模型进行识别时可能会把大部分负类目标识别为正类目标. 这时,需要综合考虑两个指标即识别精度 pr 和正确识别精度 re,分别定义如下

$$pr = \frac{tp}{tp + fp}, \quad re = \frac{tp}{p} \quad , \tag{2}$$

其中 p 表示正样本的目标个数,tp 表示正样本中识别为正样本的个数,fp 表示负样本中识别为正样本的个数,tp,fp 更精确定义如下[8]

$$tp = \sum_{i:y_i=1;f(x_i)\geqslant 1} 1 + \sum_{i:y_i=1;0\leqslant f(x_i)\leqslant 1} f(x_i)$$
$$fp = \sum_{i:y_i=-1;f(x_i)\geqslant 1} 1 + \sum_{i:y_i=-1;0\leqslant f(x_i)\leqslant 1} f(x_i) \quad , \tag{3}$$

显然 $re \in [0,1]$,$pr \in [0,1]$.

在这种情况下,采用一种叫做 F 测度的准则来综合衡量目标识别性能是非常合理的. F 测度最初应用于信息恢复与语音信号处理领域,在基于 SVM 的学习算法中也有初步的应用[7],其一般性定义为

$$\frac{1}{F} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \quad . \tag{4}$$

将式(2)定义的 pr 与 re 指标综合考虑,令式(4)中的 $x_1 = pr$,$x_2 = re$,$n = 2$,则有

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{pr} + \frac{1}{re}\right) \quad , \tag{5}$$

简化上式得到

$$F = \frac{2 \cdot pr \cdot re}{pr + re} \quad . \tag{6}$$

当 $pr = re = 0$ 时,定义 $F = 0$;理想情况下,F 近似为 1.

从 F 测度的定义中可以看出,它不仅反映了目标的正确识别率,也约束了目标错误识别率,是目标正确识别与错误识别的折中函数,比单纯正确识别率更能全面反映目标识别性能. 与单纯采用目标正确识别率与 LOO 方法[10]中最小错误率相比,F 测度是更好的模型优化选择的准则,也更能适应不同类别目标样本数量不同的实际情况.

2.2　非均衡数据目标识别的 SVM 模型多参数优化选择方法

传统的基于梯度的优化算法需要目标准则函数是凸函数的或者要求其梯度能准确地估计或者近似,而且求得的往往是局部最优解. 因此,如果想得到最好的模型,必须解一个非凸函数的优化问题. 遗传算法是一种基于自然选择和遗传变异等生物进化机制的全局性概率搜索算法,其搜索的全局性、自身潜在的并行性都是传统的规划方法所不具备的[11].

本文结合遗传算法的优势,提出了非均衡数据目标识别的 SVM 高斯核函数多参数并行优化选择方法,将能全面反映分类器性能的 F 测度函数作为核函数多参数优化准则. 主要思想是利用遗传算法搜索的全局性与并行性,对 SVM 多参数并行优化,采用 F 测度函数作为优化选择的准则函数,得到目标识别的 SVM 模型.

本文方法主要步骤如下:

1. 选定目标识别的训练数据集与测试数据集;

2. 针对高斯核函数 p 个不同的尺度因子 σ_1,σ_2,…,σ_p,以及惩罚因子 C_+(或者 C_-,C_+ 与 C_- 根据式(1)可相互推出)进行初始编码,给出一个具有一定规模染色体的初始群体;

3. 对于每一个染色体(即每一组参数),训练得到其 SVM 模型;

4. 定义染色体的适应度函数为:$fitness = -F = -\frac{2 \cdot pr \cdot re}{pr + re}$,根据步骤③计算得到的 SVM 模型计算群体中每一个染色体的适应度函数;

5. 根据设定的概率,对染色体进行选择、交叉与变异,得到适应度更优的染色体群;

6. 若整个优化过程已得到最优值或者满足迭代结束条件,进入下一步;否则,转入步骤③;

7. 根据得到的最优染色体,解码得到 SVM 模型参数,并且训练得到目标识别的最优 SVM 模型,形成多目标分类器.

3　实验及结果分析

本文在多组数据集上进行识别实验,首先采用国际上目标识别算法常用的基准数据集实验,以验证本文提出的模型参数优化选择方法的有效性;其次采用雷达多类目标 HRRP 数据集实验,验证本文提出算法的有效性和实用性.

3.1　基准数据集实验

基准数据集[12]包括多种不同类别的数据,常用来测试识别算法的性能. 采用本文提出的模型优化选择方法首先在基准数据集上测试,得到的模型选择结果如表1所示. 基准数据集只给出了高斯核函

表1 基准数据集SVM模型多参数优化选择结果

Table 1 Results of SVM model optimal multi-parameter selection on benchmarks

数据	特征维数	训练数据数目(+1/-1)	测试数据数目	基准结果 σ^2	基准结果 C	基准结果 识别率	本文结果 C^+	本文结果 F	本文结果 识别率
Banana	2	400(217/183)	4900	1	316.2	88.47	91.41	0.8713	88.69
Breast cancer	9	200(58/142)	77	50	15.19	73.96	0.121	0.4118	74.03
Diabetis	8	468(170/298)	300	20	(原文未给出)	76.47	0.47074	0.5868	77.00
German	20	700(217/483)	300	55	3.162	76.39	0.0478	0.6028	80.67
Heart	13	170(76/94)	100	120	3.162	84.05	0.6237	0.7901	83.00
Thyroid	5	140(39/101)	75	3	10	95.20	130.22	0.9804	98.67
Waveform	21	400(132/268)	4600	20	1	90.12	0.372	0.8492	89.48
			采用本文方法得到的核函数参数(与特征维数相同)						
Banana			0.256				0.206		
Breast cancer	9.4	155.4	10.7	18.1	11.3	15.4	21.6	13.1	14.7
Diabetis	2265.4	8.3	19.4	127.2	41.7	10.8	8.1	19.0	
German	30.4	20.1	21.5	23.4	595.2	28.9	49.2	59.8	44.6	39.3
	140.9	25.6	2985.1	26.3	33.1	22.0	347.8	81.3	120.3	30.3
Heart	26.1	14.5	79.6	15.8	18.6	35.4	18.9	14.5	123.5	14.3	14.8	13.1	19.9
Thyroid	0.2			0.1		5.9		0.1		1.4	
Waveform	36.5	66.2	22.1	167.9	22.0	88.6	64.8	39.9	41.8	21.5
	23.2	67.3	22.1	36.9	35.7	34.8	94.9	34.7	30.1	52.8	61.9

数的单个参数,即表1中的 σ^2 参数,而本文给出了高斯核函数的多参数即表1中的 σ_p^2 参数,参数个数与特征维数相同.

从表1中可以看出,与文献中给出的模型最优值相比,采用本文提出的模型优化选择方法得到的识别结果与文献给出的结果几乎相等,验证了本文提出方法的有效性,此外本文还给出了各个特征更精确的模型参数,这是以往文献中并未给出的结果.由于目标本身特征固有的可分性的限制,文献[12]中给出的识别率基本已是最好的识别率,所以采用本文提出方法得到的模型参数,使得正确识别率无太大的提高.

3.2 雷达目标HRRP数据实验

采用某电磁特性计算软件计算得到不同弹道点的4种飞机目标HRRP,分别是B737、B747、F16、F18,每一种飞机有855个弹道点,即855幅HRRP,每幅HRRP特征维数为150.实验过程抽取每一类数据的1/3做训练数据,全部的HRRP做测试数据.采用OAA-SVM方法训练测试,此时针对每一类飞机HRRP数据训练时正类样本与负类样本是非均衡的,利用本文提出的多参数并行选择方法进行SVM目标识别的模型选择,得到与HRRP每一维特征对应的精确核函数参数(即150个核函数参数与一个惩罚因子 C),结果如图1所示.与传统的交叉验证方法得到的SVM单参数模型并利用FSVM方法[13]识别得到的结果相比,采用本文得到的多参数模型进行识别的结果非常理想,如表2所列.

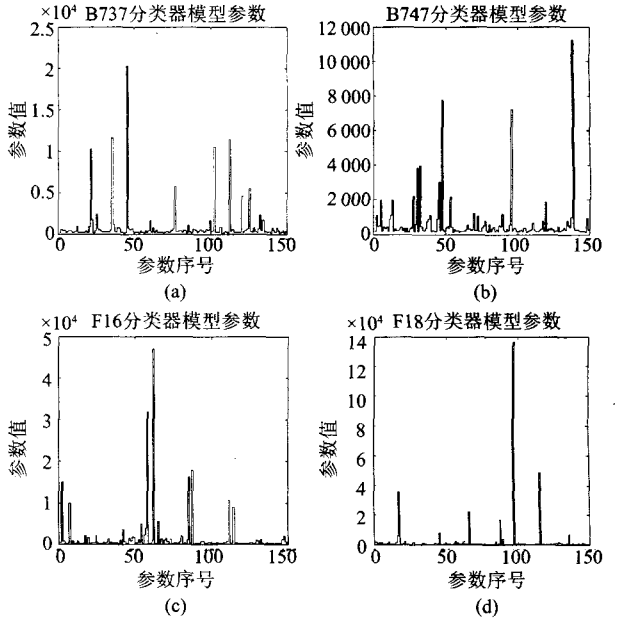

图1 4种飞机目标HRRP分类器SVM模型参数 (a)B737分类器模型参数 (b)B747分类器模型参数 (c)F16分类器模型参数 (d)F18分类器模型参数

Fig. 1 SVM model multi-parameter of 4 planes (a) model multi-parameter of B737 (b) model multi-parameter of B747 (c) model multi-parameter of F16 (d) model multi-parameter of F18

3.3 讨论

采用交叉验证法选择SVM单参数模型是一种常用的纯经验的方法,它通过训练样本来估计模型的错误率,根据实验的中间结果不断调整参数,逐渐寻优以得到一个相对最优的模型.这种纯经验的方法对于训练样本数量巨大时,计算量的代价也是相

表 2　雷达目标 HRRP 识别模型识别结果
Table 2　Classification results of Radar targets' HRRP

目标类型	交叉验证识别参数与结果			本文方法识别结果		
	σ^2	C_+	识别率(%)	F	C_+	识别率(%)
B737	8	300	96.86	0.98893	0.1047	99.50
B747			96.49	0.98891	0.1685	99.30
F16			78.22	0.76705	0.7162	86.52
F18			72.13	0.70366	0.2777	86.73

当大的,而且无法进行 SVM 模型多参数优化选择,因此它不是一种实用的 SVM 模型多参数选择方法.

本文提出的非均衡数据多目标识别 SVM 模型多参数优化选择方法,并且基于 F 测度提出了新的准则函数,更适合大样本、非均衡雷达目标数据集,同时能够智能并行优化模型多参数,而优化得到的多参数正是 SVM 核函数的精确参数.

多参数优化过程中采用了遗传算法进行并行优化,这个过程比交叉验证方法选择和调整参数更具有智能性,而且并没有增加太多的算法复杂性,但本文方法得到的模型结果更精确. 由这种模型优化选择方法确定的 SVM 模型大大提高了雷达多目标正确识别率.

4　结语

基于 SVM 的目标识别性能,不仅取决于给定的训练数据与测试数据,还与其核函数及其参数有密切的关系,精确的核函数参数非常难以选取. 针对这个问题,本文提出了非均衡数据目标识别中 SVM 模型多参数并行选择方法. 实验结果验证了本文提出的多参数智能选择方法的有效性和实用性,能够精确地给出 SVM 模型多参数最优值,并且达到满意的目标正确识别率.

REFERENCES

[1] Vapnik V N. *The Nature of Statistical Learning Theory*[M]. New York: Springer, 2000,123—170.

[2] TAN Kun, DU Pei-Jun. Hyperspectral remote sensing image classification based on support vector machine[J]. *J. Infrared Millim. Waves*(谭琨,杜培军. 基于支持向量机的高光谱遥感图像分类. **红外与毫米波学报**),2008,**27**(2):123—128.

[3] WANG Li, HE Yong, LIU Fei, *et al.* Rapid detection of sugar content and pH in beer by using spectroscopy technique combined with support vector machines[J]. *J. Infrared Millim. Waves*(王莉,何勇,刘飞,等. 应用光谱技术和支持向量机分析方法快速检测啤酒糖度和 pH 值. **红外与毫米波学报**),2008,**27**(1):51—55.

[4] Xu P, Chan A K. An efficient algorithm on multi-class support vector machine model selection[C]. Procee-dings of the International Joint Conference on Neural Net-works 2003. Portland, IEEE, 2003: 3229—3232.

[5] Chapelle O, Vapnik V N, Bousquet O, *et al.* Choosing multiple parameters for support vector machines[J]. *Machine Learning*,2002,**46**(1):131—159.

[6] Keerthi S S. Efficient tuning of SVM hyper parameters using radius/margin bound and iterative algorithms[J]. *IEEE Trans. Neural Networks*,2002,**13**(5):1225—1229.

[7] Musicant D R, Kumar V, Ozgur A. Optimizing F-measure with support vector machines[C]. In the Sixteenth International Florida Artificial Intelligence Research Society Conference, St. Augustine, Florida, USA, AAAI Press,2003: 356—360.

[8] Eitrich T, Lang B. Efficient optimization of support vector machine learning parameters for unbalanced datasets[J]. *Journal of computational and applied mathematics*, 2006, **196**(2):425—436.

[9] Morik K, Brockhausen P, Joachims T. Combining sta-tistical learning with a knowledge-based approach-a case study in intensive care monitoring[C]. In 16th Pro ceedings of the International Conference on Machine Learning. San Mateo, Canada: Morgan Kaufman Publishers, 1999, 268—277.

[10] Bo L F, Wang L, Jiao L C. Multiple parameter selection for LS-SVM using smooth leave-one-out error[J]. Lecture notes in computer science. Berlin: Springer,2005,851—856.

[11] LI Min-Qiang, KOU Ji-Song, LIN Dan. *Theory and application of genetic algorithm*[M]. Beijing Science Press(李敏强,寇纪淞,林丹. **遗传算法的基本理论与应用**. 北京:科学出版社),2002,8—10.

[12] Rätsch G. Benchmarks data sets [Online]. http://ida.first. fraunhofer. de/projects/bench/benchmarks. htm, 1999.

[13] Takuya I, Shigeo A. Fuzzy support vector machine for pattern classification [C]. *Proceeding of International Joint* Conference on Neural Networks, Washington, D. C,2001, 1449—1454.