支持向量分类和多宽度高斯核

常 群[1,2],王晓龙[1,2],林沂蒙[2],王熙照[2,3],Daniel S. Yeung[2,4]
(1. 哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001;
2. 哈尔滨工业大学深圳研究生院媒体与生命科学实验室,广东深圳 518055;
3. 河北大学数学与计算机科学学院,河北保定 071002;4. 香港理工大学计算学系,香港九龙)

摘 要: 支持向量分类中,高斯核不区分样本中各个特征的重要性,显然各个特征对分类的贡献一般是不相同的. 为了体现这种差别从而提高支持向量机的泛化性能,文中提出了多宽度高斯核的概念. 多宽度高斯核增加了支持向量机的超级参数,进一步地,文中提出了多参数模型选择算法. 算法利用误差界自动实现模型选择. 通过实验验证了多宽度高斯核和多参数模型选择算法的有效性.

关键词: 支持向量机; 多宽度高斯核; 多参数模型选择; 误差界
中图分类号: TP181 **文献标识码**: A **文章编号**: 0372-2112 (2007) 03-0484-04

Support Vector Classification and Gaussian Kernel with Multiple Widths

CHANG Qun[1,2], WANG Xiao-long[1,2], LIN Yi-meng[2], WANG Xi-zhao[2,3], Daniel S YEUNG[2,4]
(1. *Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*;
2. *Media and Life Science Lab, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong* 518055, *China*;
3. *Faculty of Mathematics and Computer Science, Hebei University, Baoding, Hebei* 071002, *China*;
4. *Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong, China*)

Abstract: In support vector classification, Gaussian kernel is insensitive to the differences of features. However, generally, different features function differently in classification. To improve the generalization performance of support vector machines, the Gaussian kernel with multiple widths is proposed to emphasize the different contributions of features to classification. With this kernel, the related model selection scheme is designed which can automatically tune multiple parameters for support vector machines by minimizing the error bound. The efficiencies of the proposed kernel and related model selection algorithms are validated via experiments.

Key words: support vector machines; Gaussian kernel with multiple widths; model selection with multiple parameters; error bound

1 引言

高斯核(或称高斯径向基核)是支持向量机(Support Vector Machines, SVM)常采用的核函数[1~4],它的宽度参数定义了核函数的泛化规模,直接影响 SVM 的泛化性能,高斯核形式如下:

$$k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2) \quad (1)$$

很显然,高斯核不区分样本中各个特征的重要性. 但一般来说,各个特征对分类的贡献是不同的,为了体现这种差别,引入多宽度高斯核并命名为 GKMW(Gaussian Kernel with Multiple Widths),GKMW 采用如下形式

$$k(x, y) = \exp\left(-\frac{1}{2}\sum_p \frac{(x_p - y_p)^2}{\sigma_p^2}\right), \ \forall p \quad (2)$$

容易证明 GKMW 是一个正定核,即为 SVM 的可取核. 现在给予每个特征一个权重 fw_p(feature weight)去指示该特征对分类的贡献程度并规定

$$fw_p = 1/\sigma_p, \quad \forall p \quad (3)$$

则有 GKMW 的另一种形式

$$k(x, y) = \exp\left(-\frac{1}{2}\sum_p (fw_p x_p - fw_p y_p)^2\right) \quad (4)$$

确定核参数(又称超级参数)的最优值属于模型选择(Model Selection)的研究领域[5]. 当模型参数较少时,可用穷尽搜索的方法进行模型选择. 当模型参数比较多时搜索空间太大,穷尽搜索变的无能为力,一种可行的办法是利用误差界来选择模型参数. SVM 的归纳学习原则是基于结构风险最小化的,也就是对未来的期望误差最小化,因此利用误差界去实现模型选择是与 SVM 的归纳学习机制一致的.

2 前人及相关的工作

简要介绍 SVM 的形式化,详细的讨论可参考文献[2,3]. 分类时,给定 l 个样本的训练集 $D = \{x_j, y_j\} \in R^n$

收稿日期:2006-03-26;修回日期:2006-10-27
基金项目:国家自然科学基金重点项目(No. 60435020);国家自然科学基金重大研究计划面上项目(No. 90612005)

$\times\{-1,1\}$, $\forall j$,SVM 对未来样本预测如下

$$f(x) = \mathrm{sign}\left(\sum_{j=1}^{l}\alpha_j y_j k(x, x_j) + b\right) \tag{5}$$

其中 $\alpha_j \neq 0$ 的训练样本称为支持向量. 这里 α_j, $\forall j$ 通过优化下面的对偶问题获得

$$\max \quad W(\alpha) = e^{\mathrm{T}}\alpha - \frac{1}{2}\alpha^{\mathrm{T}}Q\alpha$$
$$\text{s.t.} \quad 0 \leq \alpha_j \leq C,\ y^{\mathrm{T}}\alpha = 0 \tag{6}$$

这里 e 是单位向量; Q 是 $l \times l$ 矩阵,其中 $Q_{ij} = y_i y_j K(x_i, x_j)$; C 是惩罚因子. 该对偶问题的原问题是

$$\min \quad J(w, b, \xi) = \frac{1}{2}w^{\mathrm{T}}w + C\sum_{j=1}^{l}\xi_j$$
$$\text{s.t.} \quad y_j(w^{\mathrm{T}}\Phi(x_j) + b) \geq 1 - \xi_j,\ \xi_j \geq 0 \tag{7}$$

这里 ξ_j 称为误差项; Φ 定义了从输入空间到特征空间的映射,映射通过 $K(x, y) = \Phi(x)^{\mathrm{T}}\Phi(y)$ 间接定义. 上述优化问题称为 L1-SVM;当误差项 ξ_j 的指数为 2 时对应的 SVM 称为 L2-SVM. 另外,L1-SVM 在最优解时还有下面的等式

$$w = \sum_{j=1}^{l}\alpha_j y_j \phi(x_j),$$
$$\frac{1}{2}w^{\mathrm{T}}w + C\sum_{j=1}^{l}\xi_j = e^{\mathrm{T}}\alpha - \frac{1}{2}\alpha^{\mathrm{T}}Q\alpha \tag{8}$$

L1-SVM,Chung 等人[6]提出了一个半径-间隔误差界

$$B = \left(R^2 + \frac{1}{C}\right)\left(w^{\mathrm{T}}w + 2C\sum_{j=1}^{l}\xi_j\right) \tag{9}$$

这里 R 是在特征空间中包含所有训练样本的最小球的半径[2,7],进一步考虑到使用 GKMW 时有 $K(x, x) = 1$, R^2 可通过下面的最优化问题得到

$$R^2 = \max_{\beta}\left(1 - \sum_{i,j=1}^{l}\beta_i\beta_j k(x_i, x_j)\right)$$
$$\text{s.t.} \quad \sum_{i}^{l}\beta_i = 1,\ \beta_i \geq 0,\ \forall i \tag{10}$$

Chung 等人利用式(9)去选择高斯核的宽度并且取得了良好的效果. 在利用泛化误差界为 SVM 进行模型选择的研究中,前人的工作主要集中在 2 个参数的选择上,即高斯核宽度 σ 和惩罚因子[5,6,9,10,12] C. Chapelle 等人[8]提出了多参数的 SVM 模型选择方法,其工作是基于 L2-SVM 的. 一般来说,L1-SVM 比 L2-SVM 更常使用,而且 L1-SVM 产生的支持向量的个数比 L2-SVM 的要少[6],普遍认为稀疏性是 SVM 的一个重要特征之一. 本文的工作是基于 L1-SVM 的多参数模型选择.

3　多参数模型选择

采用逐次逼近的迭代程序去实现多参数的模型选择. 用 $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n)^{\mathrm{T}}$ 代表 GKMW 中所有的宽度参数,迭代算法如下

(1) 初始化 σ 和 C.

(2) 用标准的 SVM 算法找出 (α^0, β^0),参考式(6)和(10).

(3) 通过拟牛顿法优化式(9),更新参数 σ 和 C,即

$$(\sigma, C) = \arg\min_{\sigma, C} B(\alpha^0, \beta^0, \sigma, C)$$

(4) 转到步骤(2),或者根据某个停止条件结束迭代程序.

下面我们介绍算法的实现.

3.1　计算误差界的梯度

误差界 $B(\sigma, C)$ 和参数 (σ, C) 之间是一种隐式依赖关系. 为方便,用 ∂v 表示向量 v 中各个元素的梯度. 根据前人在处理隐式梯度问题的研究成果[6,8,9,11],这里给出一个具体的结论:

引理 1　在上面的多参数优化的迭代过程中, $\partial B/\partial\sigma$ 不依赖 $\partial\alpha/\partial\sigma$ 和 $\partial\beta/\partial\sigma$ 的数值; $\partial B/\partial C$ 不依赖于 $\partial\alpha/\partial C$ 的数值.

证明　为方便起见用 θ 代表 (σ, C);用 μ 代表 (α, β); μ_0 代表最优化问题式(6)和式(10)的解 (α^0, β^0);定义 $B(\theta, \mu^0) = \min_{\mu} B(\theta, \mu(\theta))$. 为明晰起见,先看下面的函数依赖关系图:

$$\left.\begin{array}{l}\theta(\sigma, C) \xrightarrow{(6)\ (7)} (\alpha^0, w, \xi)\\ \theta(\sigma) \xrightarrow{(10)} (\beta^0, R)\\ \theta(C)\end{array}\right\} \rightarrow B(w, \xi, R, C)$$

从式(9)可看出:欲求 B 对参数 θ 的梯度,也就是求 B 中的变量 (w, ξ, C, R) 对 θ 的梯度. 表面上看 w 依赖于 α,参见式(8); α 又依赖于 θ,参见式(6). 但实际上变量 (w, ξ, α_0) 的值都是通过同一个优化问题式(6)和式(7)获得,只是它们在名称和形式上不同而已,因此在优化解 $\alpha = \alpha^0$ 处,有 $\partial w^{\mathrm{T}}w/\partial\alpha = 0$; $\partial\xi/\partial\alpha = 0$. 另外 α^0 依赖于 C,但是 C 并不依赖于 α^0,也就是 $\partial C/\partial\alpha^0 = 0$. 总结上面的分析:当 $\alpha = \alpha^0$ 时, $\partial B/\partial\alpha = 0$0,这将会导致

$$\left.\frac{\partial B}{\partial\alpha}\right|_{\alpha=\alpha^0}\frac{\partial\alpha}{\partial\theta} = 0$$

因此迭代计算 $\partial B/\partial\theta$ 时,不需要 $\partial\alpha/\partial\theta$ 的数值. 同样的分析也适合于 β 和 R,参考式(10),即

$$\left.\frac{\partial B}{\partial\beta}\right|_{\beta=\beta^0}\frac{\partial\beta}{\partial\sigma} = 0$$

所以迭代计算 $\partial B/\partial\sigma$ 时也不需要 $\partial\beta/\partial\sigma$ 的数值. 上述分析可形式化为

$$\frac{\partial B}{\partial\theta} = \left.\frac{\partial B}{\partial\theta}\right|_{\mu^0_{\mathrm{fixed}}} + \left.\frac{\partial B}{\partial\mu}\right|_{\mu=\mu^0}\frac{\partial\mu}{\partial\theta} = \left.\frac{\partial B}{\partial\theta}\right|_{\mu^0_{\mathrm{fixed}}}$$

证明完毕.

上面的引理 1 说明了在计算误差期望界 B 的梯度时,不需要考虑 (α, β) 的影响. 下面的一切梯度计算都以引理 1 为基础,不再重复说明.

在模型选择时,前人的参数搜索算法一般都工作

在 $\ln\theta$ 空间,优点是可以消除参数的正条件约束,从而使优化问题变为无约束的. 定义 $B_1 = R^2 + 1/C$ 和 $B_2 = w^T w + 2C\Sigma_i\xi_i$. 在 $\ln\theta$ 空间中,B 的梯度为

$$\frac{\partial B}{\partial \ln\theta_i} = \theta_i\left(\frac{\partial B_1}{\partial\theta_i}B_2 + \frac{\partial B_2}{\partial\theta_i}B_1\right) \tag{11}$$

这里 θ_i 代表 θ 中第 i 个参数,考虑式(6)和(8),有 $\partial B_2/\partial C = \partial(2e^T\alpha - \alpha^T Q\alpha)/\partial C$ 并存在约束条件:$0 \leq \alpha_j \leq C$; $y^T\alpha = 0$, $\forall j$. 消除 C 的约束条件[6]通过变换:$\alpha = C\gamma$. 这样式(6)可变为 $\mathrm{Max}: e^T\gamma C - \gamma^T Q\gamma C^2/2$ 并有约束 $0 < \gamma_j < C$ 和 $y^T\gamma = 0$. 考虑到 $2e^T\alpha - \alpha^T Q\alpha = C^2(2e^T\gamma/C - \gamma^T Q\gamma)$,有

$$\begin{aligned}\frac{\partial B_2}{\partial C} &= \frac{\partial(C^2(2e^T\gamma/C - \gamma^T Q\gamma))}{\partial C}\\ &= 2C\left(\frac{2e^T\gamma}{C} - \gamma^T Q\gamma\right) - C^2\left(\frac{2e^T\gamma}{C^2}\right)\\ &= \frac{2}{C}(2e^T\alpha - \alpha^T Q\alpha) - \frac{2}{C}(e^T\alpha)\\ &= \frac{2}{C}(e^T\alpha - \alpha^T Q\alpha)\end{aligned} \tag{12}$$

其他梯度为

$$\frac{\partial B_1}{\partial C} = \frac{-1}{C^2} \tag{13}$$

根据式(2), $\forall p$

$$\frac{\partial B_1}{\partial\sigma_p^2} = -\sum_{i,j}\beta_i\beta_j\frac{\partial k(x_i, x_j)}{\partial\sigma_p^2} \tag{14}$$

$$\frac{\partial B_2}{\partial\sigma_p^2} = -\sum_{i,j}\alpha_i\alpha_j y_i y_j\frac{\partial k(x_i, x_j)}{\partial\sigma_p^2} \tag{15}$$

$$\frac{\partial k(x_i, x_j)}{\partial\sigma_p^2} = k(x_i, x_j)\frac{((x_i)_p - (x_j)_p)^2}{2\sigma_p^4} \tag{16}$$

这里 $(x)_p$ 代表向量 x 的第 p 个元素. 以上给出了用拟牛顿法优化误差界需要的所有梯度.

3.2 算法实现的一些观点

多参数模型选择算法是这样是实现的:标准的 SVM 算法使用 Chang 和 Lin[4]提供的 LIBSVM 软件,并做了一些修改;拟牛顿下降法使用 Broyden-Fletcher-Goldfarb-Shanno (BFGS) 算法[13]. 因为迭代算法是渐进收敛的,所以需要设立停止条件. 经验显示,较松的停止条件反而会使选择的 SVM 模型有更好的鲁棒性. 对迭代算法的条件设置如下:(1) 最大迭代次数设置为 50 次;(2) $\forall p$, σ_p 和 C 都初始化为 1;(3) 线性搜索时,为防止过大的步长进入错误的搜索区域,其步长 λ 限制在如下范围内:

$$\|\lambda\theta\| \leq \sqrt{2n} \tag{17}$$

n 代表模型参数的个数;θ 中的 t 代表第 t 次迭代.

4 实验部分

实验采用标准数据集,为了防止数值计算时大数据特征吃掉小数据特征的现象,需要对数据做预处理,将各个特征都线性规格化在区间 $[-1, 1]$,这一建议来自于 Chirn-Jen Lin. 所有的模型参数都采用本文介绍的方法选择模型参数,对于使用高斯核的 SVM,模型参数的选取也使用上面的算法,只是将多个宽度(或特征权值)用一个宽度代替即可,涉及的算法改动简单,不再多述. 数据集描述在表 1.

表 1 数据集描述

数据集	类	特征	样本	来源
Heart Disease	2	13	270	[14]
Ionosphere	2	34	351	[15]
Iris Plant	3	4	150	[15]
Wine Recognition	3	13	178	[15]

4.1 实验

实验(a):该实验用于评价使用 GKMW 的 SVM 和使用高斯核的 SVM 性能,将每个数据集随机选取 80 % 的样本作为训练集,剩下的 20 %的样本作为测试集,评价指标采用 5 次实验的平均值,实验结果在表 2.

表 2 高斯核和 GKMW 的性能比较(准确率)

数据集	高斯核	GKMW
Heart Disease	81.4815 %	85.1852 %
Ionosphere	94.0171 %	95.4416 %
Iris Plant	96.6667 %	97.3333 %
Wine Recognition	97.7528 %	98.3176 %

实验(b):该实验用于评价多参数模型选择算法的性能,实验采用 Heart Disease 和 Ionosphere 数据集,训练集/测试集的划分方法同(a). 然后做出误差界(bound, y 轴)和测试错误率(error rate, y 轴)曲线,其中 x 轴为迭代次数(Iterations). 如果 bound 曲线和 error rate 曲线的走向是大体一致的,下降的,最后是稳定收敛的,我们就说这个算法是有效的. 实验结果见图 1 和图 2.

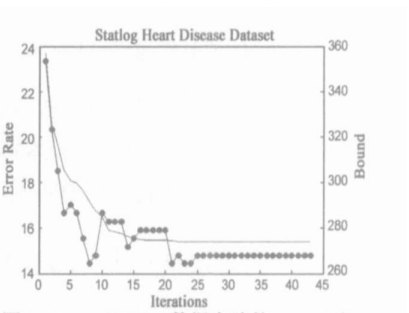

图 1 Heart disease 数据实验的 Bound 和 Error Rate 曲线,点线代表 Error Rate 曲线;另一条为 Bound 曲线

实验(c):该实验用于说明各个特征对分类的贡献是不同的,实验仍采用(b)中的两个数据集,通过条形图显示各个特征权. 实验结果见图 3 和图 4.

4.2 实验分析

实验(a):表 2 显示 SVM 采用 GKMW 以后,其分类性能有所提高. 高斯核是一个性能优越的核函数,因此任何以高斯核为基础所做的改进都是很有意义的. GKMW 的核参数空间覆盖了单宽度高斯核的参数空

间,因此,如果有一个好的模型选择算法的话,理论上应该能够提高泛化能力.

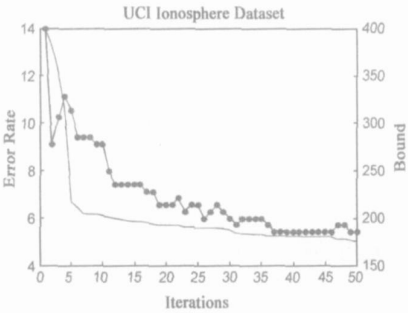

图 2　Ionosphere 数据实验的 Bound 和 Error Rate 曲线,点线代表 Error Rate 曲线;另一条为 Bound 曲线

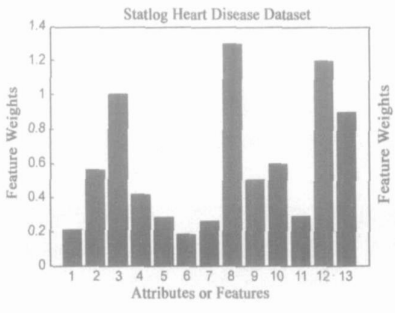

图 3　Heart disease 数据实验的特征权分布图

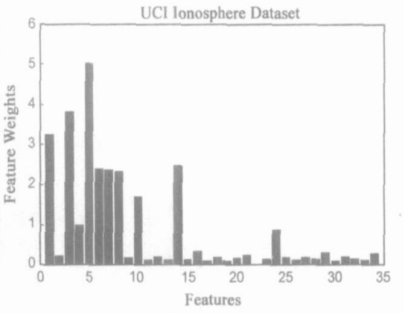

图 4　Ionosphere 数据实验的特征权分布图

实验(*b*):图 1 和图 2 显示随着误差界的下降,其测试误差也随之下降.迭代开始时,由于各个超级参数离最优值相差较大,所以测试错误率较大.随着误差界的下降,测试误差随之降低,迭代优化到 30 次以后,误差界和测试误差的下降已经不明显,并且数值比较稳定,因此我们把迭代次数最大值设为 50 次是比较合理的,两条曲线共同的走向以及最后收敛到比较稳定的数值说明本文的模型选择算法是有效的.

实验(*c*):图 3、图 4 都指示出了不同的特征在分类时具有不同的权值(重要性).

5　结论

本文提出了 GKMW 的概念,通过给样本的各个特征赋予不同的权重,SVM 的分类性能有所提高.采用 GKMW 以后,SVM 的超级参数变多,针对这个问题,文中又进一步提出了多参数的模型选择策略.通过利用误差界去选择模型参数,其底层机制和 SVM 的训练机制是一致的,也就是说模型选择和模型训练都是以最小化 SVM 的泛化误差为标准.实验证实 SVM 采用 GKMW 后,其分类性能比采用单宽度的高斯核有所提高.

参考文献:

[1] V Vapnik. Universal learning technology: support vector machines[J]. NEC Journ of Adv Tech, 2005, 2(2): 137 - 144.

[2] V Vapnik. Statistical Learning Theory[M]. New York: Wiley, 1998.

[3] V Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[4] C C Chang, C J Lin. LIBSVM: a library for support vector machines[DB/OL]. http://www.csie.ntu.edu.tw/～cjlin/libsvm, 2006.

[5] O Chapelle, V Vapnik. Model selection for support vector machines[A]. Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press, 2000. 12: 230 - 236.

[6] K M Chung, W C Kao, C L Sun, L L Wang, C J Lin. Radius margin bounds for support vector machines with the RBF kernel[J]. Neural Computation, 2003, 15(11): 2643 - 2681.

[7] Bernhard Schölkopf, Alex Smola. Learning with Kernels[M]. Cambridge, MA: MIT Press, 2002.

[8] O Chapelle, V Vapnik, O Bousquet, S Mukherjee. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1): 131 - 159.

[9] S S Keerthi. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms[J]. IEEE Trans Neural Networks, 2002, 13(5): 1225 - 1229.

[10] K Duan, S S Keerthi, A N Poo. Evaluation of simple performance measures for tuning SVM hyperparameters[J]. Neurocomputing, 2003, 51: 41 - 59.

[11] J F Bonnans, A Shapiro. Optimization problems with perturbations: A guided tour[J]. SIAM Review, 1998, 40(2): 228 - 264.

[12] D Anguita, et al. Hyperparameter design criteria for support vector classifers[J]. Neurocomputing, 2003, 55: 109 - 134.

[13] D F Shanno, K H Phua. Minimization of unconstrained multivariate functions[J]. ACM Trans Math Software, 1980, 6(4): 618 - 622.

[14] P Brazdil. Statlog Dataset[DB/OL]. http://www.niaad.liacc.up.pt/old/statlog, 2005.

[15] D J Newman, et al. UCI Machine Learning Repository[DB/OL]. http://www.ics.uci.edu/～mlearn/MLSummary.html, 2005.

作者简介:

常　群　男,1969 年生于山东寿光,博士研究生,研究方向为机器学习. E-mail: qchang@insun.hit.edu.cn

王晓龙　男,1955 年生于哈尔滨,博士,博士生导师,研究方向为机器学习和自然语言处理.

林沂蒙　男,1983 年生于山东临沂,硕士,研究方向为机器学习和支持向量机.

王熙照　男,1963 年生于河北曲阳,博士,博士生导师,IEEE 资深会员,研究方向为机器学习.

Daniel S. Yeung　男,博士,讲座教授,IEEE 院士,研究方向为机器学习.