基于支持向量机分类问题的勒让德核函数

张 瑞[1] 王文剑[2,3] 张亚丹[1] 孙芳玲[1]
(山东理工大学理学院 淄博 255049)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]
(山西大学计算机与信息技术学院 太原 030006)[3]

摘 要 基于勒让德正交多项式,提出了一类新的核函数——勒让德核函数。在双螺旋集和标准UCI数据集上的实验表明,在鲁棒性与泛化性能方面,该核函数比常用的核函数(多项式核、高斯径向基核等)具有更好的表现,而且其参数仅在自然数中取值,能大大缩短参数优化时间。

关键词 支持向量机,核函数,模型选择
中图法分类号 TP181 **文献标识码** A

Legendre Kernel Function for Support Vector Classification

ZHANG Rui[1] WANG Wen-jian[2,3] ZHANG Ya-dan[1] SUN Fang-ling[1]
(School of Science, Shandong University of Technology, Zibo 255049, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[3]

Abstract This paper presented a new set of kernel function-Legendre kernel function based on Legendre polynomial. The performance and robustness of the presented kernel were investigated on bi-spiral benchmark data set as well as five data sets from the UCI benchmark repository. The experiment results demonstrate that the presented kernel has competitive robust and generalization performance compared with commonly used kernel functions (polynomial kernel and Radial Basis Function etc.). Moreover, the Legendre kernel has one parameter which is only chosen from natural number, thus parameter optimization is facilitated greatly.

Keywords Support vector machine, Kernel function, Model selection

1 引言

支持向量机由于其优良的泛化性能和强大的模拟非线性关系的能力,近年来在分类和回归等问题中得到了广泛的应用[1-4]。然而,模拟非线性关系的好坏,主要取决于所选择的核函数。一个好的核函数能将非线性的分类(回归)问题映射成线性的分类(回归)问题。然而,可选择的核函数有很多,由于给定数据的分布通常是未知的,因此事先很难从众多的核函数中选择出一个合适的。即使选择了核函数,优化核函数中的参数也是一件非常困难的事情。例如最常用的高斯核含有一个参数,虽然现在已有多种优化其参数的方法[5-7],但寻求最优的参数仍需花费大量的时间。UKF核[8]虽然有良好的泛化性能,但它含有两个参数,要优化这两个参数需要花费更多的时间。为了解决上述问题,基于勒让德正交多项式,提出了勒让德核函数。该核函数最大的优点就是其参数仅在自然数上取值,参数优化时间大大缩短。在双螺旋集和标准的UCI数据集上的实验进一步表明,在泛化性能、鲁棒性方面,该核比常用的核有更好的表现。

2 支持向量机简介

支持向量机(Support Vector Machine)是一种新型的机器学习方法。对于二值分类问题,算法如下:

设已知训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$,其中 $x_i\in R^d$,$y_i\in\{-1,+1\}$。对于非线性可分的分类问题,引入从 R^d 到高维特征空间 H 的映射 Φ,构造优化问题:

$$\min \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t. } y_i(\langle\omega,\Phi(x_i)\rangle+b)\geqslant 1-\xi_i,\ i=1,\cdots,l$$

$$\xi_i\geqslant 0,\ i=1,\cdots,l$$

其对偶问题是

$$\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i,x_j)-\sum_{i=1}^{l}\alpha_i$$

到稿日期:2011-08-17 返修日期:2011-11-12 本文受国家自然科学基金(60975035),教育部博士点基金(20091401110003),山东理工大学博士基金(4041-410002)资助。

张 瑞(1964—),男,博士,副教授,硕士生导师,主要研究方向为机器学习、计算智能等,E-mail:zrlgz@sdut.edu.cn;王文剑(1968—),女,博士,教授,博士生导师,主要研究方向为机器学习、计算智能等。

$$\text{s.t.}\sum_{i=1}^{l}y_i\alpha_i=0$$

$$0\leqslant\alpha_i\leqslant C,i=1,\cdots,l$$

其中 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为对应于映射 Φ 的核函数：

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle$$

其中$\langle,\rangle$为向量的内积。通过求解上述对偶问题，得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*)^{\mathrm{T}}$，选取 α^* 的一个正分量 $0<\alpha_j^*<C$，计算阈值：

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

最后构造决策函数：

$$f(\boldsymbol{x})=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^*y_iK(\boldsymbol{x}_i,\boldsymbol{x})+b^*)$$

最常见的核函数有：

(a)高斯径向基核：$K(\boldsymbol{x},\boldsymbol{z})=\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}}$

(b)多项式核：$K(\boldsymbol{x},\boldsymbol{z})=(\langle\boldsymbol{x},\boldsymbol{z}\rangle+1)^n$

(c)指数核：$K(\boldsymbol{x},\boldsymbol{z})=\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|}{2\sigma^2}}$

3 勒让德核函数的构建

3.1 勒让德多项式简介[9]

勒让德(Legendre)多项式是在区间$[-1,1]$上权函数为常值1的正交多项式，满足正交关系：

$$\int_{-1}^{+1}L_m(x)L_n(x)\mathrm{d}x=\begin{cases}0, & m\neq n\\ \dfrac{2}{2n+1}, & m=n\end{cases}$$

并有递推关系：

$$L_0(x)=1,L_1(x)=x$$

$$(n+1)L_{n+1}(x)=(2n+1)xL_n(x)-nL_{n-1}(x)$$

3.2 勒让德核函数的建立

为了构建勒让德核函数，把上述勒让德多项式推广成向量形式的广义勒让德多项式：

$$L_0(\boldsymbol{x})=1,L_1(\boldsymbol{x})=\boldsymbol{x},(n+1)L_{n+1}(\boldsymbol{x})=(2n+1)\boldsymbol{x}L_n^{\mathrm{T}}(\boldsymbol{x})-nL_{n-1}(\boldsymbol{x})\quad(1)$$

式中，$L_i^{\mathrm{T}}(\boldsymbol{x})$表示 $H_i(\boldsymbol{x})$的转置，$\boldsymbol{x}$ 是行向量。如果 n 是奇数，则 $L_n(\boldsymbol{x})$是一个行向量，否则 $L_n(\boldsymbol{x})$就是实数。根据广义勒让德多项式，定义 n 阶勒让德核函数为

$$K(\boldsymbol{x},\boldsymbol{z})=\sum_{i=1}^{n}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|^2}{d}}\quad(2)$$

式中，d 表示向量 $\boldsymbol{x}$ 的维数。下面证明式(2)定义的函数是核函数。

Mercer 定理[10] 令 X 是 R^d 上的紧集，$K(\boldsymbol{x},\boldsymbol{z})$是 $X\times X$ 上的连续实值对称函数，如果对任意可积函数 $f(\boldsymbol{x})$，都有 $\iint_{X\times X}K(\boldsymbol{x},\boldsymbol{z})f(\boldsymbol{x})f(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\geqslant0$，那么函数 $K(\boldsymbol{x},\boldsymbol{z})$就一定是核函数。先证明$\sum_{i=1}^{n}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})$是核函数。因为

$$\begin{aligned}&\iint_{X\times X}\sum_{i=1}^{n}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})f(\boldsymbol{x})f(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\\&=\sum_{i=1}^{n}\iint_{X\times X}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})f(\boldsymbol{x})f(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\\&=\sum_{i=1}^{n}(\int_X L_i(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x})(\int_X L_i^{\mathrm{T}}(\boldsymbol{z})f(\boldsymbol{z})\mathrm{d}\boldsymbol{z})\\&=\sum_{i=1}^{n}(\int_X L_i(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2\geqslant0\end{aligned}$$

由 Mercer 定理可知，$\sum_{i=1}^{n}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})$是核函数。再由核函数的基本性质[10]知，$\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|^2}{d}}$是核函数，所以 $K(\boldsymbol{x},\boldsymbol{z})=\sum_{i=1}^{n}L_i(\boldsymbol{x})L_i^{\mathrm{T}}(\boldsymbol{z})\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|^2}{d}}$是核函数。

表1列出了从0阶到4阶的勒让德核函数。

表1 0阶到4阶的勒让德核函数

阶数	核函数 K(x,z)
0	L
1	(1+c)L
2	(1.25+c+2.25ab−0.75a−0.75b)L
3	(1.25+c+2.25ab−0.75a−0.75b+6.25abc−3.75ac−3.75bc+2.25c)L
4	[1.25+c+2.25ab−0.75a−0.75b+6.25abc−3.75ac−3.75bc+2.25c+(1225a²b²−1050a²b+105a²−1050ab²+900ab−90a+105b²−90b+9)/64]L

其中，$a=\langle\boldsymbol{x},\boldsymbol{x}\rangle,b=\langle\boldsymbol{z},\boldsymbol{z}\rangle,c=\langle\boldsymbol{x},\boldsymbol{z}\rangle,L=\mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|^2}{d}}$。

图1—图3分别显示了0阶到4阶的勒让德核函数 $K(\boldsymbol{x},\boldsymbol{z})$在 $\boldsymbol{x}\in[-4,4],\boldsymbol{z}=-0.4,0,0.4$ 对应的图形。

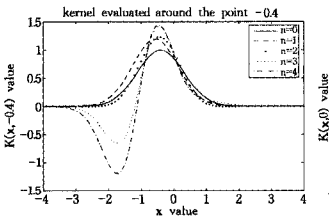

图1 0阶到4阶的勒让德核函数 $K(\boldsymbol{x},-0.4)$对应的图形

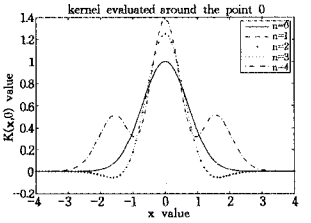

图2 0阶到4阶的勒让德核函数 $K(\boldsymbol{x},0)$对应的图形

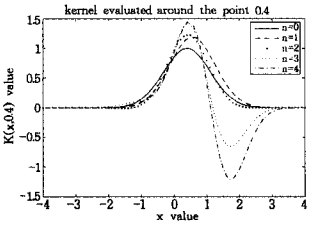

图3 0阶到4阶的勒让德核函数 $K(\boldsymbol{x},0.4)$对应的图形

由上述图形可以看出，0阶到4阶的勒让德核函数 $K(\boldsymbol{x},0)$关于竖轴是对称的，其他的情形非对称。

4 实验结果与分析

为了将勒让德核与常用的核函数(线性核、多项式核和高斯核等)在泛化性与鲁棒性方面做进一步的比较，在双螺旋集和标准数据库 UCI 中的5个数据集上分别做分类实验。以下实验中，支持向量机的参数 C 取值为100。

4.1 在双螺旋集上的分类对比

作为典型的线性不可分问题，双螺旋线问题一直是模式识别领域公认的一个相当有难度的问题，是检验模式识别问题的试金石。试验中我们对有噪声与无噪声的双螺旋集分别进行了试验。

由于多项式核不能将双螺旋集正确分类，因此在实验中选择了高斯核、指数核与勒让德核做对比实验。表2列出了实验核函数参数的取值。表3和表4显示了能将双螺旋集正

确分类的核函数相对应的间隔和支持向量个数，其中 RBF 与 ERBF 所对应的是最大的间隔与相应的支持向量个数。

表 2　试验中所用的核函数及参数的取值

核函数	参数	取值	步长
Legendre	n	0～4	1
RBF	α	0.2～2	0.2
ERBF	α	0.2～2	0.2

表 3　在无噪声条件下几种核对应的间隔、支持向量个数的比较

核函数/参数	间隔	支持向量个数/占总数百分比
Legendre/n=1	0.5621	178/92.7%
Legendre/n=2	3.4353	176/91.7%
Legendre/n=3	11.2909	8/4.2%
RBF	0.156940	190/99%
ERBF	0.150897	192/100%

表 4　在有噪声条件下几种核对应的间隔、支持向量个数的比较

核函数/参数	间隔	支持向量个数/占总数百分比
Legendre/n=1	0.4106	137/71.4%
Legendre/n=2	2.6994	136/70.8%
Legendre/n=3	8.6691	18 /9.4%
RBF	0.149177	189/98.4%
ERBF	0.149277	192/100%

在无噪声条件下，由表 3 可以看出，当 $n=1$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 4 倍；当 $n=2$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 22 倍；当 $n=3$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 72 倍，而支持向量的个数仅为 8 个，占总个数的 4.2%。在有噪声条件下，当 $n=1$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 3 倍；当 $n=2$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 18 倍；当 $n=3$ 时，勒让德核函数的最大间隔大约为 RBF 核和 ERBF 核的 58 倍，而支持向量的个数仅为 18 个，占总个数的 9.4%。

图 4 与图 5 分别给出了二阶 Legendre 核将无噪声双螺旋集和有噪声双螺旋集正确分类的情形。

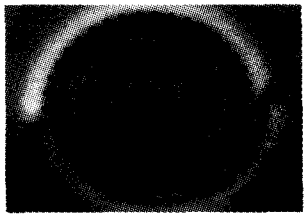

图 4　二阶 Legendre 核将无噪声双螺旋集的正确分类

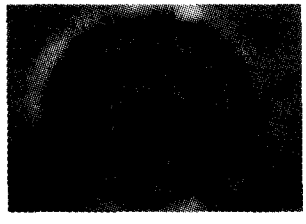

图 5　二阶 Legendre 核将有噪声双螺旋集的正确分类

4.2　在标准 UCI 数据集上的分类对比

实验中，我们在标准的 UCI 数据集上做分类对比试验，用到的数据集如表 5 所列。将勒让德核与最常用的高斯核(RBF)和多项式核(POLY)做实验对比，勒让德核与高斯核的参数取值如表 2 所列，多项式核参数 n 取值从自然数 1 到 10。为了缩短计算时间，banana、splice 与 waveform 3 个数据集从训练集中随机选择了 200 个作为训练集，从测试集中随机选择了 100 个作为测试集。实验结果如表 6 所列。

由表 6 可以看出，勒让德核比常用的高斯核和多项式核具有更好的表现，但勒让德核仅取了参数值为 0 到 4 的自然数。

表 5　试验中用到的 UCI 数据集

数据集	维数	训练点个数	测试点个数
breast-cancer	9	200	77
banana	2	200	100
thyroid	5	140	75
splice	60	200	100
waveform	21	200	100

表 6　在标准 UCI 数据集上分类的最大精度及对应参数

	Legendre	RBF	POLY
breast-cancer	0.7403/n=4	0.7403/σ=0.8	0.7403/n=8
banana	0.9100/n=0	0.9100/σ=1.4	0.8000/n=6
thyroid	0.9867/n=3	0.9867/σ=0.8	0.9867/n=3
splice	0.9000/n=2	0.7200/σ=2.0	0.8800/n=9
waveform	0.9300/n=3	0.9100/σ=1.4	0.9300/n=6

结束语　本文基于勒让德多项式，建立了向量形式的勒让德多项式，并由此提出了勒让德核函数，在双螺旋集及 UCI 数据集上与常用的核函数(多项式核、高斯径向基核，指数核)做了分类实验对比。实验表明，在能将有噪声与无噪声的双螺旋集正确分类的前提下，勒让德核具有最大间隔和最小支持向量个数。在 UCI 数据集上的实验表明，勒让德核比常用的核具有更好的表现。除了上述特点之外，勒让德核最大的优势还在于其参数仅在自然数上取值，这就会大大缩短优化时间。以上仅仅是在双螺旋集与部分 UCI 数据集上所做的分类实验，至于在其他分类问题或在回归及相关应用问题上的表现如何，还有待于进一步研究。

参 考 文 献

[1] Vapnik V. Statictical Learning Theory [M]. New York: Wiley, 1998

[2] Artan Y, Huang X. Combining multiple 2nu-SVM classifiers for tissue segmentation [C]// IEEE ISBI. 2008: 488-491

[3] Chen C H, Ho P G P. Statistical pattern recognition in remote sensing[J]. Pattern Recognition, 2008, 9: 2731-2741

[4] Wu Z L, Li C H, Ng J K Y, et al. Location estimation via support vector regression[J]. IEEE Trans, Mobile Comput, 2007, 6 (3): 311-321

[5] Bi L P, Huang H, Zheng Z Y, et al. New heuristic for determination Gaussian kernels parameter [J]. Mach, Learning Cybernet, 2005, 7: 4299-4304

[6] Liu H J, Wang Y N, Lu X F. A method to choose kernel function and its parameters for support vector machines [J]. Mach, Learning Cybernet, 2005, 7: 4277-4280

[7] Wu K P, Wang S D. Choosing the kernel parameters of support vector machines according to the inter-cluster distance [C]// Neural Networks(IJCNN '06). 2006: 1205-1211

[8] Zhang R, Wang W J. Facilitating the applications of support vector machine by using a new kernel[J]. Expert Systems with Applications, 2011, 38: 14225-14230

[9] 李庆阳，王能超，易大义. 数值分析[M]. 北京：清华大学出版社，2008

[10] 邓乃扬，田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京：科学出版社，2004