一类改进的埃尔米特核函数

田 萌[1,2] 王文剑[1]
(山西大学计算机与信息技术学院 太原 030006)[1] (山东理工大学理学院 淄博 255049)[2]

摘 要 核函数及其参数的选择是决定支持向量机(support vector machine,SVM)分类性能的关键。基于埃尔米特多项式,利用三角核函数构造并证明了一类改进的埃尔米特核函数——三角埃尔米特核函数。该类核函数含两个核参数,其中一个核参数可由样本点到样本均值的距离简单确定,而另一个核参数仅在自然数集中选取,从而简化了该类核函数的参数优化。在双螺线数据集、棋盘格数据集及7个UCI数据集上的实验表明,该类核函数比常见的多项式核函数、高斯核函数及文献[6]提出的埃尔米特核函数有着更好的泛化性能和鲁棒性。

关键词 支持向量机,核选择,埃尔米特多项式,三角核函数
中图法分类号 TP181 **文献标识码** A

A Set of Improved Hermite Kernel Function

TIAN Meng[1,2] WANG Wen-jian[1]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(School of Science, Shandong University of Technology, Zibo 255049, China)[2]

Abstract The selection of kernel function and its parameters plays a significant role in support vector machine (SVM) classification algorithms. Based on Hermite polynomial and the triangular kernel function, a new set of kernel functions—triangular Hermite kernel was proposed. The triangular Hermite kernel has two parameters. One parameter is determined by the distance between sample points and sample mean, and the other parameter is chosen only from non-negative integer. So the parameters of the triangular Hermite kernel can be optimized easily. The experimental results on bi-spiral data, checkerboard data and 7 UCI data sets indicate that the new kernel achieves the competitive classification performance, compared with polynomial kernel, Gaussian kernel, and the previous Hermite kernel proposed in reference [6].

Keywords Support vector machine, Kernel selection, Hermite polynomial, Triangular kernel function

1 引言

由Vapnik等人提出的支持向量机是数据挖掘领域的一种经典算法[1],作为建立在统计学习理论基础上的一种模式识别方法,SVM旨在解决小样本的模式识别问题。SVM利用非线性映射Φ将训练样本X映射到高维特征空间F,$\Phi: X \rightarrow F$,以增强数据的线性可分性。在高维特征空间的学习问题中训练样本总是以成对样本内积的形式出现,核函数的引入能避免显式地描述非线性映射Φ,进而巧妙地解决了非线性求解问题。研究和实验表明核函数的类型及参数直接决定了SVM的学习能力和泛化性能,选择不同的核函数本质上就是选择不同的"相似程度"的度量方式[2]。由于实际问题中样本数据的特征往往是未知的,使得选择合适的"相似程度"的度量方式即核函数变得非常困难。若选择了不合适的核函数或参数,特征空间的线性可分性会变得很差,从而无法实现分类的目的。因此,核函数及其参数选择一直都是SVM理论研究中的核心问题。

核函数选择具体可分为核函数类型的确定和核参数的选择。目前常用的核函数[1-3]有:多项式核函数$K(\boldsymbol{x},\boldsymbol{z})=((\boldsymbol{x},\boldsymbol{z})+1)^n$,高斯核函数$K(\boldsymbol{x},\boldsymbol{z})=\exp(-\|\boldsymbol{x},\boldsymbol{z}\|^2/2\sigma^2)$,Sigmoid核函数$K(\boldsymbol{x},\boldsymbol{z})=S(v(\boldsymbol{x},\boldsymbol{z})+c)$等,其中高斯核因其普适性在实际问题中应用最为广泛。此外,人们还尝试针对具体应用背景来构造核函数[4],但迄今为止还没有得出一个有效的方法。

正交函数系是分析数学中的一个重要组成部分,现已被广泛应用于电子工程、图像处理及数据统计等领域。基于正交函数系的优良性质,许多学者利用正交函数来构造SVM的核函数。Vapnik等人利用埃尔米特正交多项式构造核函数来进行支持向量回归分析[1]。Ozer等[5]将切比雪夫正交多项式推广至向量形式得到了泛化的切比雪夫多项式,从而实现了样本数据向量形式的输入,并用指数函数取代原切比雪夫函数的权函数得到了一类切比雪夫核函数。随后基于泛化多项式的埃尔米特核函数、勒让德核函数等[6,7]相继被应用于支持向量分类问题中。文献[5-7]通过实验说明,构造的

到稿日期:2013-07-19 返修日期:2013-10-27 本文受国家自然科学基金(60975035,61273291),山西省回国留学人员科研基金(2012-008)资助。

田 萌(1979-),女,博士生,讲师,主要研究方向为机器学习、计算智能等,E-mail:luckywalter@163.com;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,E-mail:wjwang@sxu.edu.cn(通信作者)。

这些类核函数相比常见核函数(多项式核函数、高斯核函数)有着更好的鲁棒性和泛化性能。

文献[5-7]中构造的这些核函数本质上都是高斯核和泛化正交多项式 Dirichlet 核的乘积组合核。注意到高斯核的融入在提高正交多项式核函数泛化性能的同时,也给这些类核函数增加了一个参数 σ(高斯核函数的参数)。如果照搬一些高斯核参数的优化方法,比如常规的网格搜索法[8]、核排列[9]和 Fisher 鉴别分析[10]等,势必增加核函数的选择成本。为了规避复杂的 σ 的优化过程,Ozer[5]、张瑞[6,7]等分别令 $\sigma=1$,$\sqrt{d/2}$(d 为样本数据的维数),以节省参数优化时间。但这样的处理太过笼统,忽略了样本数据结构信息对泛化性能的影响。

本文提出利用三角核函数取代高斯核函数来构造出一类新的埃尔米特核函数——三角埃尔米特核函数。该类核函数的核参数能被简单合理地设定,从而在保留了样本数据较多的结构信息的同时,又使其核参数易于优化。实验表明三角埃尔米特核函数相比多项式核函数、高斯核函数及文献[6]提出的埃尔米特核函数有着更好的泛化性能和鲁棒性。

2 三角埃尔米特核函数的构造

2.1 三角核函数

Laplace 核函数 $K(\boldsymbol{x},\boldsymbol{z})=\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma})$是一种近来被广泛关注的核函数[11],与高斯核相比,它对参数 σ 的变化不太敏感。利用所熟知的泰勒展开公式 $e^{-x}\approx 1-x,(x\to 0)$ 可得到一个更为简化的形式:

$$K_{simp}(\boldsymbol{x},\boldsymbol{z})=(1-\frac{\|\boldsymbol{x},\boldsymbol{z}\|}{\sigma})_+ \tag{1}$$

其中,记$()_+$为$(u)_+=\begin{cases}u, & u\geqslant 0\\ 0, & u<0\end{cases}$。文献[12]曾指出 $K_{simp}(\boldsymbol{x},\boldsymbol{z})$有着无穷的 VC 维,它的分类性能与高斯核相当,且有着高斯核所不具备的尺度不变性。

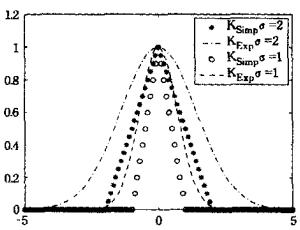


图 1 $\sigma=1,2$ 时,$K_{simp}(\boldsymbol{x},0)$及 $K_{Rbf}(\boldsymbol{x},0)$的曲线图

图 1 给出了当 $\sigma=1,2$ 时,式(1)所示函数及高斯核函数在 $x\in[-5,5]$,$z=0$ 时的函数曲线。从图 1 可看出,这两类函数的图形有着较大区别。$K_{simp}(\boldsymbol{x},\boldsymbol{z})$在函数曲线上呈现一个典型的三角图形,故文献[12]称其为三角核函数,但该文献未给出该函数是核函数的证明。本文为进一步简化函数 $K_{simp}(\boldsymbol{x},\boldsymbol{z})$的表达式,设样本点的全体为 $X=\{\boldsymbol{x}_i\}_{i=1}^{l}$,$\bar{\boldsymbol{x}}$为样本均值,$l$ 为样本个数,记 σ_0 为所有样本点到样本均值$\bar{\boldsymbol{x}}$距离最大值的两倍,则能保证 $P(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma_0}<1)=1$ 始终成立。这样便得到了本文所讨论的函数,其表达式为:

$$K_{Tri}(\boldsymbol{x},\boldsymbol{z})=1-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma_0} \tag{2}$$

其中,$\sigma_0=2\arg\max_{\boldsymbol{x}}\{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|\leqslant r,\boldsymbol{x}\in X,\bar{\boldsymbol{x}}=\frac{1}{l}\sum_{i=1}^{l}\boldsymbol{x}_i\}$。

下面给出式(2)定义的函数 $K_{Tri}(\boldsymbol{x},\boldsymbol{z})$是一个核函数的严格证明。首先列出几个需要的定义及定理。

定义 1[1] 函数 $F(u)$属于 $C^{\infty}(0,\infty)$,并满足条件 $(-1)^kF^{(k)}(u)\geqslant 0,u\in(0,\infty),k=0,1,\cdots$,则称这一函数是$(0,\infty)$上的完全单调函数。

定理 1[1] 若 $X\subset\mathbb{R}^n$,函数 $f:(0,\infty)\to\mathbb{R}$,$K$ 是一个定义在 $X\times X$ 上的函数且 $K(\boldsymbol{x},\boldsymbol{z})=f(\|\boldsymbol{x}-\boldsymbol{z}\|^2)$,则当 f 完全单调时 $K(\boldsymbol{x},\boldsymbol{z})$是正定核。

Mercer 定理[1] 令 X 是 R^n 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 $\iint_{X\times X}K(\boldsymbol{x}-\boldsymbol{z})f(\boldsymbol{x})f(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\geqslant 0$,那么函数 $K(\boldsymbol{x},\boldsymbol{z})$一定是核函数。

由定义 1 及定理 1 知:

定理 2 $K_{Tri}(\boldsymbol{x},\boldsymbol{z})=1-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma_0}$是一个核函数。

证明:设 $f(t)=1-\frac{\sqrt{t}}{\sigma_0}$,计算得 $f'(t)=-\frac{1}{2\sigma_0}t^{-\frac{1}{2}}$,$f^{(k)}(t)=(-1)^k\frac{(2k-3)!!}{2^k\sigma_0}t^{-\frac{(2k-1)}{2}}$,$k\geqslant 2$。自然有结论$(-1)^kf^{(k)}(t)\geqslant 0$成立,由定义 1 知 $f(t)$是完全单调函数。再依据定理 1 可知 $K_{Tri}(\boldsymbol{x},\boldsymbol{z})=1-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma_0}$是正定核。定理 2 得证。

2.2 两类埃尔米特核函数

埃尔米特多项式的一般表达式为:

$$H_n(x)=\sum_{k=0}^{[\frac{n}{2}]}(-1)^k\frac{n!}{k!(n-2k)!}(2x)^{n-2k},n=0,1,2,\cdots$$

它的前两项及递推关系式为:

$H_0(x)=1,H_1(x)=2x$,

$H_n(x)=2xH_{n-1}(x)-2(n-1)H_{n-2}(x),n=2,3,\cdots$

将变量 x 替换为行向量 $\boldsymbol{x}$,并相应地将 x^n 作如下形式的替换:

$$x^n\to\begin{cases}(\boldsymbol{x}\cdot\boldsymbol{x}^{\mathrm{T}})^{n/2}, & n=2k\\ (\boldsymbol{x}\cdot\boldsymbol{x}^{\mathrm{T}})^{(n-1)/2}\boldsymbol{x}, & n=2k+1\end{cases},k=0,1,2,\cdots$$

称所得多项式系为泛化的埃尔米特多项式系$\{H_n(\boldsymbol{x})\}$,其前两项及递推关系式为[6]:

$$\begin{aligned}&H_0(\boldsymbol{x})=1,H_1(\boldsymbol{x})=2\boldsymbol{x}\\ &H_n(\boldsymbol{x})=2\boldsymbol{x}H_{n-1}^{\mathrm{T}}(\boldsymbol{x})-2(n-1)H_{n-2}(\boldsymbol{x}),n=2,3,\cdots\end{aligned} \tag{3}$$

本文中黑体 $\boldsymbol{x}$ 表示行向量,$\boldsymbol{x}^{\mathrm{T}}$ 表示其转置。

基于泛化的埃尔米特多项式,张瑞等构造了一类埃尔米特多项式核,其表达式为[6]:

$$K_{Gauss\_H}(\boldsymbol{x},\boldsymbol{z})=\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{d})\sum_{j=0}^{n}H_j(\boldsymbol{x})H_j^{\mathrm{T}}(\boldsymbol{z}) \tag{4}$$

其中,d 表示向量 $\boldsymbol{x},\boldsymbol{z}$ 的维数。观察式(4)可看出该类核本质上是高斯核与泛化 Dirichlet 核 $\sum_{j=0}^{n}H_j(\boldsymbol{x})H_j^{\mathrm{T}}(\boldsymbol{z})$的乘积组合核,并令高斯核函数参数 $\sigma=\sqrt{d/2}$。基于此类核函数的构造特点,本文称这类埃尔米特核函数为高斯埃尔米特核函数,记为 $K_{Gauss\_H}(\boldsymbol{x},\boldsymbol{z})$。

文献[6]通过实验展示了高斯埃尔米特核相比常见核函数如多项式核、高斯核等有着更好的鲁棒性和泛化性能。但文献[6]用 $\sqrt{d/2}$ 直接赋值给高斯核的参数 σ 的处理方法虽然简化了参数的优化,却丢失了数据大量的结构信息,进而影响了 SVM 的泛化性能。自然直接套用高斯参数 σ 优化方法也是不合适的,因为这样的做法必然将增加核选择成本,给这类核函数在实际中的应用带来麻烦。

考虑到三角核函数相比高斯核函数而言对参数 σ_0 的变化不太敏感，本文利用三角核 $K_{Tri}(x,z)$ 取代高斯核，得到一类新的函数，其表达式为：

$$K_{Tri\_H}(x,z)=(1-\frac{\|x-z\|}{\sigma_0})\sum_{j=0}^{n}H_j(x)H_j^{\mathrm{T}}(z) \tag{5}$$

其中，$\sigma_0=2\arg\max_r\{\|x-\bar{x}\|\leqslant r, x\in X, \bar{x}=\frac{1}{l}\sum_{i=1}^{l}x_i\}$。定理 3 指出式(5)所示函数是一个 Mercer 核。

定理 3 $K_{Tri\_H}(x-z)=(1-\frac{\|x-z\|}{\sigma_0})\sum_{j=0}^{n}H_j(x)H_j^{\mathrm{T}}(z)$，是核函数，其中 $\sigma_0=2\arg\max_r\{\|x-\bar{x}\|\leqslant r, x\in X, \bar{x}=\frac{1}{l}\sum_{i=1}^{l}x_i\}$。

证明：由于

$$\begin{aligned}&\iint_{X\times X}H(x)H^{\mathrm{T}}(z)f(x)f(z)\mathrm{d}x\mathrm{d}z\\&=\int_X H(x)f(x)\mathrm{d}x\int_X H^{\mathrm{T}}(z)f(z)\mathrm{d}z\\&=(\int_X H(x)f(x)\mathrm{d}x)^2\geqslant 0\end{aligned}$$

依据 Mercer 定理知 $H(x)H^{\mathrm{T}}(z)$ 是核函数。又由于核函数满足加法及乘法运算的封闭性，结合定理 2 的结论知 $K_{Tri\_H}(x,z)$ 是一个 Mercer 核。定理 3 得证。

基于该类核函数的构造特点，本文称这类改进的埃尔米特核函数为三角埃尔米特核函数，记为 $K_{Tri\_H}(x,z)$。

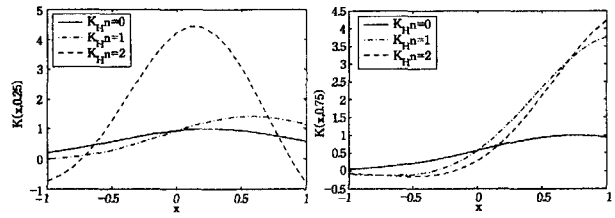
三角埃尔米特核函数含两个核参数，利用样本点到样本均值的距离确定参数 σ_0 所需计算量小，且考虑了样本集的空间信息。在确定参数 σ_0 后，该核函数只剩下一个仅在自然数集中取值的参数 n，这使得三角埃尔米特核函数虽然有两个核参数，但这两个核参数能被较简单快速地确定，避免了网格搜索的麻烦，从而大大节省了参数优化成本。

表 1 给出了 0 阶到 2 阶三角埃尔米特核函数的表达式。

表 1　0 阶到 2 阶三角埃尔米特核函数表达式

n	三角埃尔米特核函数
0	$1\cdot(1-\|\mathbf{x}-\mathbf{z}\|/\sigma_0)$
1	$\{1+4\mathbf{x}^{\mathrm{T}}\mathbf{z}\}\cdot(1-\|\mathbf{x}-\mathbf{z}\|/\sigma_0)$
2	$\{1+4(2\|\mathbf{x}\|^2-1)(2\|\mathbf{z}\|^2-1)+4\mathbf{x}^{\mathrm{T}}\mathbf{z}\}\cdot(1-\|\mathbf{x}-\mathbf{z}\|/\sigma_0)$

图 2 显示了 0 阶到 2 阶的三角埃尔米特核函数在 $x\in[-1,1]$，$z=0.25, 0.75$ 时对应的曲线。



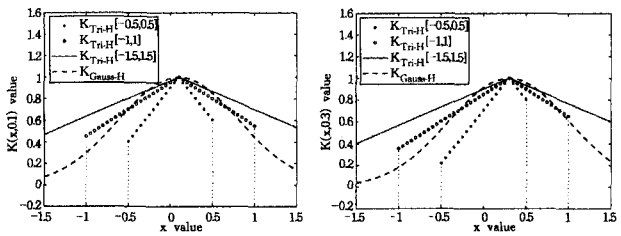
(a) $z=0.25$　　(b) $z=0.75$

图 2　$x\in[-1,1]$时 0 阶到 2 阶 $K_{Tri_H}(x,z)$对应的曲线

2.3　两类埃尔米特核函数间的比较

三角埃尔米特核函数与高斯埃尔米特核函数的表达式虽然都含 $\sum_{j=0}^{n}H_j(x)H_j^{\mathrm{T}}(z)$，但由于分别组合了三角核和高斯核，使得这两类埃尔米特核函数间有着较大区别。为更直观地展示三角埃尔米特核函数与高斯埃尔米特核函数的区别，不失一般性，本文分别取 $x\in[-0.5,0.5]$、$[-1,1]$和$[-1.5,1.5]$，$z=0.1, 0.3$，给出一阶三角埃尔米特核函数与一阶高斯埃尔米特核函数的图形，见图 3。在图 3 中由于 x 是一维向量($d=1$)，故当 z 取定时，高斯埃尔米特核函数的图形是固定不变的，而三角埃尔米特核函数则随着 x 取值区间的变化而改变形状。

为阐述清楚，表 2 列出了两类核函数在图 3 中 4 组点处的具体取值。众所周知选择不同的核函数就是选择不同的标准对相似性和相似性程度进行度量。以表 2 中点对$(x,z)=(-0.4,0.1)$为例，由 $|x-z|=0.5$ 知两点间距离恰为$[-0.5,0.5]$区间长度的 1/2，$[-1,1]$区间长度的 1/4，$[-1.5,1.5]$区间长度的 1/6，自然地这两点在这 3 个区间上的相似程度也应该是不同的。相比高斯埃尔米特核函数在 3 个区间上始终有 $K_{Gauss\_H}(-0.4,0.1)=0.7788$，而本文构造的三角埃尔米特核函数在$[-0.5,0.5]$、$[-1,1]$及$[-1.5,1.5]$区间上的取值分别为 0.5045、0.7506 及 0.8835，因此可以说三角埃尔米特核函数的参数设定考虑了更多的数据集自身的信息，也即三角埃尔米特核函数比高斯埃尔米特核函数保留了样本数据更多的距离“相似性”信息。



(a) $z=0.1$　　(b) $z=0.3$

图 3　$x\in[-0.5,0.5]$,$[-1,1]$,$[-1.5,1.5]$时，$K_{Tri\_H}(x,z)$，$K_{Gauss_H}(x,z)$对应的曲线

表 2　图 3 中核函数在所取点上的值

核函数	K_{Tri_H}			K_{Gauss_H}
样本区间	$[-0.5,0.5]$	$[-1,1]$	$[-1.5,1.5]$	任一区间
$(-0.4,0.1)$	0.5045	0.7506	0.8835	0.7788
$(0.4,0.1)$	0.7027	0.8504	0.9001	0.9139
$(-0.4,0.3)$	0.3186	0.6126	0.6525	0.7674
$(0.4,0.3)$	0.9027	0.9504	0.9668	0.9900

3　实验结果

本文在双螺旋线数据集、棋盘格数据集以及 7 个 UCI 数据集上对比三角埃尔米特核函数 K_{Tri_H}、高斯核函数 K_{Rbf}、多项式核函数 K_{Poly} 及文献[6]中构造的高斯埃尔米特核 K_{Gauss_H} 的泛化性与鲁棒性。表 3 列出了该实验中核函数参数的取值。实验中支持向量机的惩罚因子 $C=100$。

表 3　试验中所用到的核函数及参数

核函数	参数	取值区间	步长
K_{Gauss_H}	n	0—3	1
K_{Tri_H}	n	0—3	1
K_{Rbf}	σ	0.2—2	0.2
K_{Poly}	n	0—3	1

3.1　双螺旋线数据集上的实验结果

双螺旋线问题是一个两类划分问题，该问题的分类要求是把 $x-y$ 坐标平面上两条不同螺旋线上的点正确地分开。取含有 146 个样本点的双螺旋线数据集，其中正负类各 73 个。由于多项式核函数不能将双螺线数据集正确分类，因此在本节试验中选择了高斯核 K_{Rbf}、高斯埃尔米特核 K_{Gauss_H} 与

三角埃尔米特核 K_{Tri_H} 进行对比试验。表 4 列出了 3 类核在双螺旋数据集上当训练精度为 100%时的分类间隔及支持向量个数。图 4 给出的是 2 阶三角埃尔米特核函数在训练集上的分类界面。

表 4　核函数在双螺旋线集上训练结果

核函数	K_{Tri_H}	K_{Gauss_H}	K_{Rbf}
参数	n=2	n=2	σ=0.2
间隔	4.2670	10.5738	0.1658
SV 数	146	112	146

实验结果表明，三角埃尔米特核函数能将双螺旋数据集正确分类，最大间隔约为高斯核的 4 倍，但其支持向量个数及最大间隔稍逊色于高斯埃尔米特核函数。

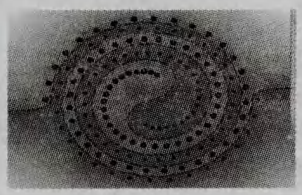


图 4　2 阶 K_{Tri_H} 在双螺旋数据集上的分类边界

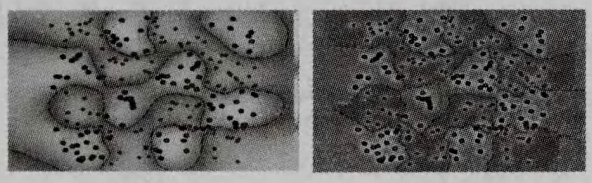
3.2　棋盘格数据集上的实验结果

本节选用了另一个典型的线性不可分问题——棋盘格数据集的分类问题。在[0,200]×[0,200]的二维空间内均匀随机生成 800 个样本点，将其分成 16 个网格，并交替划分为正类和负类。在正负类中各随机取 100 个样本点组成训练集，其余样本作为测试集。由于多项式核函数不能将训练集正确分类，表 5 只列出了三角埃尔米特核函数、高斯埃尔米特核函数以及高斯核函数在取得最高测试精度时的分类间隔、支持向量个数和参数取值。图 5 给出了 $n=1$ 时的核函数 K_{Tri_H} 和 $n=0$ 时的 K_{Gauss_H} 在棋盘格数据集上的分类界面。

实验结果表明，在棋盘格数据集上三角埃尔米特核函数取到了最高测试精度，且支持向量个数也是最少的。从参数优化过程来看，两类埃尔米特核函数都在 n 较小($n<2$)时取得了最高的测试精度，也即能较好地简化核参数的选取。

表 5　核函数在棋盘格数据集上的测试结果

	K_{Tri_H}	K_{Gauss_H}	K_{Rbf}
参数	n=1	n=0	σ=1.8
间隔	2.1307	0.1549	0.1447
SV 数	116	197	132
测试精度(%)	88.33	86.50	87.67



(a) K_{Tri_H}　　(b) K_{Gauss_H}

图 5　$K_{Tri\_H}(x,z)$ 与 $K_{Gauss\_H}(x,z)$ 在棋盘格数据集上的分类边界

3.3　标准 UCI 数据集上的实验结果

本文选取了 7 个 UCI 数据集，相应的数据集信息见表 6。高斯埃尔米特核、高斯核及多项式核与本文提出的三角埃尔米特核函数在这 7 个 UCI 数据集上的实验结果见表 7，其中在各个数据集上所能取到的最高分类精度均用黑体表示。

从表 7 可看出，两类埃尔米特核函数都在参数 n 取值较小(≤2)时取到了最高分类精度，这使得这两类核函数的参数选择相比高斯核函数更加容易，从而更有利于 SVM 在实际问题中的应用。还注意到三角埃尔米特核函数在 7 个数据集上都取到了最高的测试精度，这表明三角埃尔米特核函数比高斯核、多项式核及高斯埃尔米特核函数有着更好的鲁棒性和泛化能力。

表 6　实验中用到的 UCI 数据集

数据集	训练样本数	测试样本数	维数	类别
Diabetis	200	100	8	2
German	600	400	24	2
Wdbc	342	227	30	2
Sonar	125	83	60	2
Splice	200	100	60	2
Letter_IJK	1345	896	16	2
Ionosphere	225	126	34	2

表 7　标准 UCI 数据集上的测试结果(%)

	K_{Tri_H}	K_{Gauss_H}	K_{Rbf}	K_{Poly}
Diabetis	**74.00**	68.00	72.00	71.00
参数	n=0	n=1	σ=1.0	n=1
German	**77.50**	72.50	70.25	72.25
参数	n=1	n=1	σ=1.8	n=0
Wdbc	**92.95**	86.78	84.58	61.23
参数	n=1	n=0	σ=2.0	n=0
Sonar	**98.80**	97.59	97.59	97.59
参数	n=2	n=1	σ=1.8	n=2
Splice	**89.00**	**89.00**	72.00	80.00
参数	n=1	n=1	σ=2.0	n=2
Letter_IJK	**98.10**	97.00	97.66	67.52
参数	n=0	n=0	σ=1.8	n=0
Ionosphere	**93.57**	90.00	92.14	85.71
参数	n=0	n=1	σ=1.6	n=3

结束语　本文基于埃尔米特多项式，利用三角核函数构造了一类新的支持向量机核函数——三角埃尔米特核函数。该类核函数的最大优点在于其核参数易于优化。在双螺旋数据集、棋盘格数据集及 UCI 数据集上的实验表明，三角埃尔米特核函数相比多项式核函数、高斯核函数以及高斯埃尔米特核函数有着更好的鲁棒性和泛化能力，从而有着较广的实际应用前景。上述实验仅仅是在有限数据集上所做的分类实验，今后应进一步研究这类核函数在多分类问题或回归问题上的表现。

参 考 文 献

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

[2] 邓乃扬，田英杰. 支持向量机——理论、算法与拓展[M]. 北京：科学出版社，2009：81-114

[3] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machine[M]. Cambridge: Cambridge University Press, 2000

[4] Tsuda K, Akaho S, Kawanabe M, et al. Asymptotic properties of the Fisher kernel[J]. Neural Computation, 2004, 16(1): 115-137

[5] Sedat O, Chen C H, Cirpan H A. A set of new Chebyshev kernel functions for support vector machine pattern classification[J]. Pattern Recognition, 2011, 44(4): 1435-1447

[6] 张瑞，高红，张立伟. 一类新的支持向量机核函数——埃尔米特核函数[J]. 山西大学学报：自然科学版，2012，35(1)：38-42

[7] 张瑞，王文剑，张亚丹，等. 基于支持向量机分类问题的勒让德核函数[J]. 计算机科学，2012，34(7)：222-224

(下转第 274 页)

规模的增长，字符之间的组合变得更加复杂，造成越来越多垃圾字串通过 HP 过滤，进入到候选重复模式集合中，导致过滤效果有所降低。当语料容量增长到特定规模后，过滤比会保持一个常量，但这个特定规模在实际应用中难以达到。从实验结果来看，语料规模达到特定规模之前，HP 的过滤效果是非常显著的。

4.4 同其它方法比较

目前在处理语料规模大于内存容量的重复模式提取方法中，文献[4]所提出的是比较有代表性的方法，但因实验条件和实验语料不具有可比性，本文没有进行量化比较分析。

根据前述定性分析，因 I/O 读写的速度要远低于内存处理速度，在语料规模超过内存容量后，对于 I/O 的操作次数就成为衡量算法处理性能的关键指标。如在处理中文语料时，文献[4]中方法在语料规模增大到内存容量时其 I/O 操作次数约为汉字字符集规模，不超过 7000，当语料规模进一步增大时，其需要以第二字符进行二次划分，导致 I/O 操作次数呈指数级增长，会严重影响其处理效率；而本文方法的 I/O 操作次数同语料规模是一种线性关系，因此对语料的规模不敏感。前述实验中当处理规模为 32GB 的中文语料时，其 I/O 操作次数约为 16.3+6.3=22.6 次。

当然，文献[4]中方法适用于并行计算，若在并行环境中其处理效率会非常高，而本文算法因需要逐层剪枝和全局垃圾字串过滤，难以用于并行环境中。

结束语 使用逐层剪枝的低频垃圾串过滤，可有效地节省内存消耗，极大地减少 I/O 读写次数，提高重复模式的提取效率。实验表明，本文算法能极大限度地过滤低频字符串，快速地进行大规模语料的重复模式提取，特别适合于在大规模语料中提取所有高频重复模式。

通过大量研究，我们大胆预测：虽然汉字组合数量从理论上讲是无穷的，但当语料规模增大到特定程度后，重复模式的总量基本保持常量。这是因为，汉字字符组合必须要遵循汉语语言习惯，导致重复模式的数量是有限的；当然，一般情况下，我们所寻找的重复模式集合只是这个集合的子集。

尽管使用逐层剪枝过滤方法提高了重复模式提取效率，但低频垃圾模式的过滤效果尚有很大的改进空间；基于语料块的局部重复模式归并算法还需改进，以进一步减少 I/O 读写次数；如有条件，可对更大规模的语料进行研究和处理，以期发现实验中尚未发现的规律和趋势，这些都是本研究下一步需要处理的问题。

参 考 文 献

[1] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007，21(3)：8-19

[2] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测[J]. 中文信息学报，2004，18(6)：1-9

[3] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述[J]. 计算机科学，2010，37(3)：6-10

[4] 龚才春，贺敏，陈海强，等. 大规模语料的频繁模式快速发现算法[J]. 通信学报，2007，28(12)：161-166

[5] Nevill-Manning C G，Witten I H. Identifying Hierarchical Structure in Sequences：A Linear-Time Algorithm[J]. Journal of Artificial Intelligence Research，1997，7(1)：67-82

[6] Yamamoto M，Churcht K W. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus[J]. Computational Linguistics，2001，27(1)：1-30

[7] Larsson N J，Sadakane K. Faster Suffix Sorting[D]. Lund，Sweden：Department of Computer Science，Lund University，1999

[8] Martin F-C，Paolo F，S M. On The Sorting Complexity of Suffix-Tree Construction[J]. Journal of ACM，2000，47(6)：987-1011

[9] Anisa A H，Maxime C，Lucian I，et al. A comparison of index-based lempel-Ziv LZ77 factorization algorithms[J]. ACM COMPUTING SerVey，2012，45(1)：5

[10] 郑家恒，李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报：自然科学版，2002，25(2)：115-119

[11] Clifford R，Sergot M. Distributed and Paged Suffix-Trees for Large Genetic Databases[C]// Proceedings of 14th Annual Symposium on Combinatorial Pattern Matching. 2003：70-82

[12] Schurmann K-B，Stoye J. Suffix Tree Construction and Storage with Limited Main Memory[D]. Bielefeld，Germany：University of Bielefeld，2003

[13] Tian Y，Tata S，Hankins R A，et al. Practical Methods for Constructing Suffix Trees[J]. The VLDB Journal，2005，14(3)：281-299

[14] Cormen T H，Leiserson C E，Rivest R L，et al. In Introduction to Algorithms (2nd Ed)[M]. Cambridge，MA：MIT Press，2001

[15] 张海军，潘伟民，木妮娜，等. 一种自定义顺序的字符串排序算法[J]. 小型微型计算机系统，2012，33(9)：1968-1971

(上接第 242 页)

[8] Schölkopf B，Smola A. Learning with Kernels[M]. Cambridge：MIT Press，2002

[9] Cristianini N，Shawe-Taylor J，Elisseeff A，et al. On kernel-target alignment[C]//Proceedings of Advances in Neural Information Processing Systems. 2001：367-373

[10] Wang Wen-jian J，Xu Zong-ben，Lu Wei-zhen，et al. Determination of the spread parameter in the Gaussian kernel for classification and regression[J]. Neurocomputing，2003，55(10)：643-663

[11] Paclík P，Novovič ová J，Pudil P，et al. Road sign classification using Laplace kernel classifier[J]. Pattern Recognition Letters，2000，21(13/14)：1165-1173

[12] Fleuret F，Sahbi H. Scale-invariance of support vector machines based on the triangular kernel[C]//Proceedings of 3rd International Workshop on Statistical and Computational Theories of Vision. Nice，2003