文章编号:1674-2974(2018)10-0155-06 DOI:10.16339/j.cnki.hdxbzkb.2018.10.021

核函数选择方法研究*

王振武†,何关瑶
(中国矿业大学(北京)机电与信息工程学院,北京 100083)

摘 要:核函数的选择对支持向量机的分类结果有着重要的影响,为了提高核函数选择的客观性,提出了一种以错分实例到支持向量所在界面的距离来表示错分程度,并基于此进行秩和检验的核函数选择方法.通过与 K-折交叉验证、配对 t 测试等参数检验的统计方法进行对比分析,对9种常用核函数的分类能力在15个数据集进行了定量研究.与参数检验方法不同,秩和检验并未假定数据的分布情况(很多情况下数据并不满足假定的分布),而且数据实验证明,秩和检验不但能够对核函数的分类能力进行客观评估,而且在某些数据集上还能产生更好的核函数选择效果.

关键词:核函数;支持向量机;秩和检验;K-折交叉验证;配对 t 测试
中图分类号:TP301.6 **文献标志码**:A

Research on Selection Method of Kernel Function

WANG Zhenwu†, HE Guanyao
(School of Mechanical Electronic & Information Engineering, China University of Mining & Technology, Beijing 100083, China)

Abstract: The selection of kernel functions has an important influence on the classification results of support vector machines. This paper proposed a kernel functions selection method based on rank sum test in order to enhance the selection objectivity, where the error degree adopted in the rank sum test was represented by the distance between the error instance and the interface of support vectors. By comparing with other statistical methods, such as K-folding cross validation and paired t test, the classification abilities of nine common kernel functions were quantitatively studied based on 15 datasets. Different from parameter test methods, the rank sum test does not assume the data distribution (in some cases data cannot satisfy the assumed distribution), the experimental data proves that the rank sum test not only can objectively evaluate the classification abilities of kernel functions, but also can produce better selection results on some data sets.

Key words: kernel function; support vector machines; rank sum test; K- folding cross validation; paired t test

支持向量机(Support Vector Machine, SVM)[1]的使用与核函数的正确选择是密不可分的,核函数技术巧妙地解决了在高维特征空间中计算的"维数灾难"等问题,直接决定了SVM的非线

* **收稿日期**:2017-12-29
基金项目:国家自然科学基金资助项目(61363075),National Natural Science Foundation of China(61363075);国家高技术研究发展计划资助项目(863计划)(2012AA12A308),National High Technology Research and Development Program of China (863 Program) (2012AA12A308);中国矿业大学(北京)越琦青年学者计划项目(800015Z1117),Yue Qi Young Scholars Program of China University of Mining & Technology, Beijing(800015Z1117)
作者简介:王振武(1978—),男,山东青岛人,中国矿业大学(北京)副教授,博士
† 通讯联系人,E-mail:wangzhenwu@126.com

性处理能力[2]. 当前对核函数选择方法的研究主要集中在构造新的核函数[3-7]、核函数参数选择[8-13]以及核函数的评估[1,14-16]上. 由于在使用SVM进行分类的过程中只定义了核函数(并不显式地定义映射函数),所以在同一分类问题上选择不同的核函数对分类效果影响较大,另外映射函数的类型是多变的,在没有先验知识的情况下人们更多地是凭借主观经验进行核函数的选择,具有较大的随意性.

诸多文献从不同的角度给出了构造核函数的新方法. 文献[3]针对多标签数据集的特点构造了新的核函数,文献[4-5]结合切比雪夫多项式构造出新的核函数并以此解决回归问题,而文献[6]对 RBF 核进行极分解,并结合全局多项式核构造混合核函数,文献[7]则针对电力系统的风速概率估计这一具体问题,构造了一种由若干核密度和权重系数组成的混合核函数,消除了传统核密度模型选择最优带宽的问题. 核函数的参数选择方法研究也较多,有些文献[8]针对具体应用问题对核函数参数进行选择,有些文献则致力于研究通用的核函数选择方法. 例如,文献[9]提出了基于代价函数最大化的核函数参数选择方法,文献[10]通过研究边缘正态样本和内部正态样本之间重构误差的差异来寻找满足条件的核函数参数,文献[11]则通过每个样本的最远和最近邻信息来选择核函数参数的方法,文献[12]采用梯度下降法将类内散度矩阵的退化问题转化为迹运算准则,以此来寻找最优参数,而文献[13]则提出了广义核极化准则用来解决分类问题中的高斯核参数优化问题.

一般来说,核函数的评估指标分为四类:一类来自理论分析所给出的界[1],一类是通过考虑数据的分布特征进行核函数的选择[14],第三类是通过研究核函数核矩阵的特征信息来指引核函数的选择[15],第四类则是通过实际数据的验证结果来指导核函数的选择[16]. 遗憾的是,目前还没有成熟的理论来计算推广性的界的范围,只能给出估计值,因此理论分析在实际应用中并不实用;考虑数据的分布特征来选择核函数也有较大的局限性,例如,如果数据的分布特征不符合特定的几何特征(如类圆特征和类球特征)便无法对核函数进行选择;而通过研究核矩阵的特征信息能给出估计的泛化误差界,但算法过于复杂,在实践中很难被应用,因此通过实验结果来评估核函数是最常用的核函数选择方法. 文献[16]采用参数检验的方法对 SVM 分类结果的准确率、召回率等性能评估准则进行分析,通过将其他核函数与径向基核函数(Radial Basis Function,RBF)进行对比,来完成对核函数的综合评估,但文献[16]的方法有两个明显的缺陷:1)由于采用参数检验的方法,需要假定分类结果服从正态分布,而实际上并不是所有数据集都满足此假定;2)对数据集中某一实例的分类结果判断均是非对即错,并没有考虑被错误分类的实例的错分程度,因此对核函数的比较粒度较粗. 针对上述问题,本文提出了一种以错分实例到支持向量所在界面的距离来表示错分程度,并基于此进行秩和检验的核函数选择评估方法.

本文第 1 节对比地分析了三种模型预测性能评估的统计方法,即 K-折交叉验证[17],配对 t 测试[18]与秩和检验[19],并对秩和检验进行预测性能评估的优势进行了讨论;核函数选择的实验结果在第 2 节进行了详细分析和讨论;第 3 节对研究内容进行了总结.

1 模型预测评估方法

文献[16]指出不同评估准则在具体数值上存在差异,但应用统计方法所获得的核函数排序大体上是一致的,这说明传统的性能评估准则(如准确率、召回率和 F-measure 等)对核函数分类性能的影响不大,因此本文主要对模型评估方法进行比较.

1.1 *K*-折交叉验证

K-折交叉验证[16-17]将数据集 $D=\{(x_i,y_i)\}_{i=1}^{m}$(其中 $x_i\in R^n$,$y_i\in R$)随机划分为 K 个不相交的子集 $\{D_1,D_2,\cdots,D_K\}$,且每个子集都有 $\frac{m}{K}$ 个实例. 在分类器训练过程中,K 次迭代均用集合 $D\backslash D_t$($t\in\{1,2,\cdots,K\}$)中的数据进行训练,而用集合 D_t 中的数据进行验证. 此方法估计出的分类器 T 的泛化误差率 $\mathrm{Errcv}(T,D)$ 是 K 次验证误差率 $\mathrm{Err}_i(T,D_i)$ 的平均值. 令 $v_i=<x_i,y_i>$ 表示某一条实例,$D_{(i)}$ 表示包含实例 v_i 的子集,分类器 T 对实例 v_i 进行分类的结果用 $T(D\backslash D_{(i)},v_i)$ 表示,则 K-折交叉验证估计的分类器 T 的泛化误差率 $\mathrm{Errcv}(T,D)$ 为:

$$\mathrm{Errcv}(T,D)=\frac{1}{K}\sum_{i=1}^{K}\mathrm{Err}_i(T,D_i)=\frac{1}{m}\sum_{v_i\in D}\delta(T(D\backslash D_{(i)},v_i),y_i) \tag{1}$$

$$\text{其中,}\delta(i,j)=\begin{cases}1 & i\neq j\\ 0 & i=j\end{cases} \tag{2}$$

容易证明,假设分类器 T 的误差率为 p(p 是真实存在但未知的),则 K-折交叉验证方法估计出的分类器 T 的泛化误差率是 p 的无偏估计. 由于数

据集 D 为有限集,因此很难保证分类器 T 与数据集 D 相互独立,导致这样的无偏估计是有误差的,可以用置信区间的方法对这种误差进行估计.假设 σ^2 表示泛化误差率的方差,当 σ^2 未知时,均值 μ 的置信区间为:

$$\mathrm{Errcv}(T,D)\pm t_{1-\frac{\alpha}{2}}(n-1)s/\sqrt{K} \tag{3}$$

此处 α 表示显著性水平,s^2 是 σ^2 的无偏估计,即:

$$s^2=\frac{1}{K-1}\sum_{i=1}^{n}\left[\mathrm{Err}_i(T,D_i)-\mathrm{Errcv}(T,D)\right]^2 \tag{4}$$

1.2　配对 t 测试

尽管增加实例数 m 可以增加 K-折交叉验证的置信度,但数据集 D 毕竟是有限集,且单纯的增加 m 会导致渐进有偏.另外,应用 K-折交叉验证方法仅仅考察了均值,并未考察模型间的方差差异,其方差差异 σ^2 是用无偏估计量 s^2 来近似计算的,而配对 t 测试[16,18]方法可以准确计算模型之间的均值差异和方差差异.

设第一组样本 $w_1^{H_1},w_2^{H_1},\cdots,w_K^{H_1}$ 是学习模型 H_1 采用某种核函数根据某种性能评估准则在不同的数据集上得到的估计值,第二组样本 $w_1^{H_2},w_2^{H_2},\cdots,w_K^{H_2}$ 是学习模型 H_2 根据同样的性能评估准则,在与模型 H_1 对应的数据集上以另一种核函数得到的估计值,即 $w_1^{H_1}$ 和 $w_1^{H_2}$ 是不同的核函数在相同的数据集上产生的分类结果在同一种评估准则下的估计值,$w_2^{H_1}$ 和 $w_2^{H_2}$ 也是如此,依此类推.假设 μ_1 和 μ_2 分别表示第一组和第二组样本估计值的平均值,因此对学习模型 H_1 和 H_2 的比较便是 μ_1 和 μ_2 之间的比较.由于实验中在每个数据集上有两种学习模型进行训练,得到配对的结果,且使用了 t 检验统计量,因此被称为配对 t 测试.表 1 列出了配对 t 测试的检验方法.例如,$D_0=\mu_1-\mu_2\geqslant 0$ 便是对"μ_1 大于 μ_2"这一零假设的检验方法,即比较模型 H_1 是否比模型 H_2 学习性能差(表 1 中的双侧检验),其他两种情况以此类推.

表 1　配对 t 测试方法

Tab.1　Paired t test method

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\mu_1-\mu_2=0$ $H_1:\mu_1-\mu_2\neq 0$	$H_0:\mu_1-\mu_2\geqslant 0$ $H_1:\mu_1-\mu_2<0$	$H_0:\mu_1-\mu_2\leqslant 0$ $H_1:\mu_1-\mu_2>0$
检验统计量	$t=\dfrac{\bar{d}-D_0}{\delta_d}\sqrt{K}\approx\dfrac{\bar{d}-D_0}{s_d}\sqrt{K}$,自由度:$K-1$		
α 与拒绝域	$\lvert t\rvert>t_{\frac{\alpha}{2}}(K-1)$	$t<-t_\alpha(K-1)$	$t>t_\alpha(K-1)$
假定条件	差值总体的相对频数分布接近正态分布,配对差 d_i 由差值总体随机选出.		

表 1 中的 $\bar{d}$ 表示配对样本数据配对差值的平均值,即 $\bar{d}=\frac{1}{K}\sum_{i=1}^{K}d_i=\mu_1-\mu_2$,$d_i$ 表示第 i 个配对预测结果的差值,即 $d_i=x_i-y_i$,$i=1,2,\cdots,K$;s_d 表示配对预测结果差值的标准差,即 $s_d=\sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(d_i-\bar{d})^2}$;$\delta_d$ 表示配对预测结果差值的方差,即 $\delta_d=\sqrt{\frac{\delta_1^2+\delta_2^2-2\delta_1\delta_2\rho}{K}}$,其中 δ_1 表示第一组样本数据的方差,δ_2 表示第二组样本数据的方差,ρ 表示两组样本的相关系数;α 表示显著性水平.

1.3　秩和检验

K-折交叉验证和配对 t 测试均属于参数检验的方法,而在参数检验中对研究的总体常常要做相关假定,最常见的是假定总体服从正态分布,即 K-折交叉验证和配对 t 测试方法都是在假定泛化误差率服从正态分布的前提下进行的.由于实际的总体与假定的总体往往会有差距,这种假定可能会引起推断结果的错误,而非参数检验方法并未对数据分布进行假定,因此更能在贴近数据的真实情况下对其进行处理,本文采用秩和检验[19]这一非参数检验方法作为核函数中的数据处理方法,它可以用来对两个总体的位置进行比较.

设 $a_1,a_2,\cdots,a_n$ 是一组简单随机样本,$a_{(1)}\leqslant a_{(2)}\leqslant\cdots\leqslant a_{(n)}$ 是其观测值的有序样本,则观测值 a_i 在有序样本中的序号 r 称为 a_i 的秩,记为 $R_i=r$.本文中的简单随机样本用数据集 D 中分类错误的实例和支持向量所在的间隔边界的距离(简称错误距离)来表示.假设 $\{x_1^{H_1},x_2^{H_1},\cdots,x_o^{H_1}\}$($o\leqslant m$)是模型 H_1 在使用某一核函数的分类过程中分类错误的实例集合,$d_{x_1^{H_1}},d_{x_2^{H_1}},\cdots,d_{x_o^{H_1}}$ 表示该集合中各实例分别与支持向量所在超平面的距离,用 θ_1 表示 $d_{x_1^{H_1}},d_{x_2^{H_1}},\cdots,d_{x_o^{H_1}}$ 的中心位置;对应地,$\{x_1^{H_2},x_2^{H_2},\cdots,x_n^{H_2}\}$ 是模型 H_2 在使用另一核函数的分类过程中分类错误的实例集合,$d_{x_1^{H_2}},d_{x_2^{H_2}},\cdots,d_{x_n^{H_2}}$ 表示该集合中各实例分别与支持向量所在超平面的距离,用 θ_2 表示 $d_{x_1^{H_2}},d_{x_2^{H_2}},\cdots,d_{x_n^{H_2}}$ 的中心位置.将这 $o+n$ 个表示距离的样本进行排序产生对应的混合样本,如公式(5)所示:

$$\begin{aligned}&(d_{(x_1^{H_1})},d_{(x_2^{H_1})},\cdots,d_{(x_o^{H_1})},d_{(x_1^{H_2})},d_{(x_2^{H_2})},\cdots,\\&\quad d_{(x_n^{H_2})})\end{aligned} \tag{5}$$

其中 α 表示显著性水平,一般地,检验统计量 W 选择 $\{x_1^{H_1},x_2^{H_1},\cdots,x_o^{H_1}\}$ 和 $\{x_1^{H_2},x_2^{H_2},\cdots,x_n^{H_2}\}$ 中个数较少的进行统计(此处假定 $n\leqslant o$),秩和检验的

方法如表 2 所示,检验统计量为 $W=\sum_{i=1}^{n}R_{d_{x_i}H_2}$,即观测值在混合样本中的秩之和.

表 2 秩和检验方法

Tab. 2 Rank sum test method

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\theta_1-\theta_2=0$ $H_1:\theta_1-\theta_2\neq 0$	$H_0:\theta_1-\theta_2\geqslant 0$ $H_1:\theta_1-\theta_2<0$	$H_0:\theta_1-\theta_2\leqslant 0$ $H_1:\theta_1-\theta_2>0$
检验统计量	$W=\sum_{i=1}^{n}R_{d_{x_i}H_2}$		
α 与拒绝域	$W_I=\left\{\begin{matrix}W\leqslant W_{\alpha/2}(o,n)\\ W\geqslant W_{1-\alpha/2}(o,n)\end{matrix}\right\}$ or	$W_{II}=\{W\geqslant W_{1-\alpha}(o,n)\}$	$W_{III}=\{W\leqslant W_{\alpha}(o,n)\}$

如图 1 所示,SVM 的分离超平面由实线表示,间隔边界由虚线表示,正例点由“O”表示,负例点由“Δ”表示,在间隔边界上的点为支持向量. 当所有分类错误的实例在两个间隔边界之间时所得的距离最短(如图 1 中分类错误的正例点 x_1),其分类效果比错误分类的实例较多地出现在间隔边界之外(如图 1 中分类错误的正例点 x_2)的情况要好.

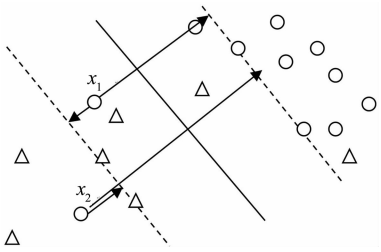

图 1 间隔最大分离超平面实例

Fig. 1 The case of interval maximum separation hyperplane

设有 k 个支持向量 $s_1,s_2,\cdots,s_k$,每个支持向量对应的系数为 $c_1,c_2,\cdots,c_k$,某一错分实例 x_i 在模型 H 下与支持向量所在间隔边界的错误距离表示为:

$$d_i^H=\sum_{j=1}^{k}c_j\parallel x_i-s_j\parallel_2 \tag{6}$$

其中 $\parallel\ \parallel_2$ 表示 L_2 范数.

2 实验结果分析

本文实验所采用的软硬件环境参数如下:CPU E3-1226 v3,内存 4G,操作系统 Windows 7 64bit,开发环境为 MyEclipse6.5,编程语言为 Java.

如表 3 和表 4 所示,本文对 9 个常用核函数在 15 个标准数据集上进行了比较验证,并对结果进行了统计分析. 核函数类型包括平移不变核(RBF、STF、CF、LF、HSF、HTF)和旋转不变核(Linear、PF、SF),数据集包括 UCI 中的常用数据集以及 Statlib 中的部分数据集.

表 3 9 个常见核函数

Tab. 3 Nine common kernel functions

编号	英文名称	缩写	数学表达式
1	Linear Function	Linear	$K(x,y)=(x,y)$
2	Polynomial Function	PF	$K(x,y)=((x,y)+1)^d$
3	Radial Basis Function	RBF	$K(x,y)=\mathrm{e}^{-g\{x-y,x-y\}^2}$
4	Sigmoid Function	SF	$K(x,y)=\tanh(g(x,y)+coef0)$
5	Symmetric Triangle Function	STF	$K(x,y)=\mathrm{Max}(1-g\mid x-y,x-y\mid,0)$
6	Cauchy Function	CF	$K(x,y)=\frac{1}{1+g(x-y,x-y)^2}$
7	Laplace Function	LF	$K(x,y)=\mathrm{e}^{-g\mid\{x-y,x-y\}\mid}$
8	Hyperbolic Secant Function	HSF	$K(x,y)=\frac{2}{\mathrm{e}^{g\{x-y,x-y\}}+\mathrm{e}^{-g\{x-y,x-y\}}}$
9	Hyperbolic Tangent Function	HTF	$K(x,y)=\frac{\mathrm{e}^{g\{x-y,x-y\}}-\mathrm{e}^{-g\{x-y,x-y\}}}{\mathrm{e}^{g\{x-y,x-y\}}+\mathrm{e}^{-g\{x-y,x-y\}}}$

表 4 实验数据集

Tab. 4 Experimental datasets

数据集	数据来源
ionosphere. data	UCI 数据库[20]
pima-indians-diabetes. data	UCI 数据库[20]
sonar. all-data	UCI 数据库[20]
monks-1. train	UCI 数据库[20]
monks-2. train	UCI 数据库[20]
monks-3. train	UCI 数据库[20]
Iris	UCI 数据库[20]
Seeds	UCI 数据库[20]
Wine	UCI 数据库[20]
Yeast	UCI 数据库[20]
Glass	UCI 数据库[20]
bupa-liver-disorders-data	UCI 数据库[20]
Sonar	UCI 数据库[20]
CPS_85_Wages	Statlib 数据库[21]
Veteran	Statlib 数据库[21]

在 3 种模型预测评估方法的实验中,K-折交叉验证采用的 10 折交叉验证,配对 t 测试和秩和检验则是在每个数据集上分别进行核函数的两两对比实验. 另外,所有实验的统计显著性水平均为 5%,实验结果会出现某核函数在某数据集上得不到实验结果的情况,此时判定为“无”.

对 3 种模型预测评估方法实验结果的处理方式为:K-折交叉验证统计 9 个核函数在 15 个数据集上的排名顺序并将其累计求和,排名依据为:置信区间有重叠则判断相等,“无”则被判断为排名最后,否

则按错误率 Errcv(T,D) 大小来排序. 而配对 t 测试和秩和检验则是根据两两对比获胜的次数相加,其中"相等"次数均增加,"无"次数均不增加,统计结果如表 5 所示,括号内的数字是经统计后该核函数在当前检验方法下的排名.

表 5　3 种方法的实验结果

Tab. 5　Experimental results of three methods

核函数	k-折交叉验证	配对 t 测试	秩和检验
Linear	21(1)	72(3)	65(3)
PF	38(4)	45(4)	49(4)
RBF	25(3)	76(1)	76(1)
SF	38(4)	38(5)	38(5)
STF	44(8)	23(8)	36(6)
CF	23(2)	76(1)	72(2)
LF	45(9)	25(7)	28(8)
HSF	40(6)	32(6)	28(8)
HTF	41(7)	23(8)	31(7)

根据表 5 的统计结果可以看出,三种方法对核函数的分类能力进行排序时存在一定差异,但大体是一致的,核函数可以大致分为三级:RBF、Linear、CF 效果最好,PF、SF 其次,STF、LF、HSF、FTF 效果最差.

虽然 3 种方法对 9 种核函数的分类能力在 15 个数据集上得到了大体一致的综合排名结果,但如果针对具体的数据集做仔细分析,会发现 K-折交叉验证和配对 t 测试方法存在较大的局限性. 例如,如图 2 和图 3 所示,在处理数据集 monks-2. train 和 monks-3. train 时,使用 K-折交叉验证在所有的核函数上得出的错误率的置信区间都十分接近,全部存在重合的情况,在统计核函数排名时只能判定它们排名一样,而使用配对 t 测试则得出所有核函数的两两对比结果全为"相等",这说明对于此类数据使用参数检验的方法无法给出比较结果,针对这种情况,秩和检验却能够很好的处理.

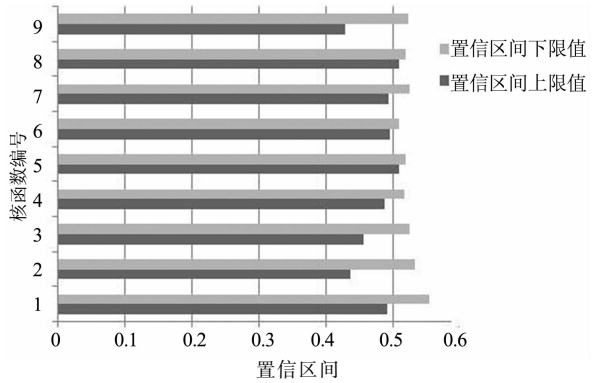

图 2　monks-2. train 的 K-折交叉验证实验结果

Fig. 2　Experimental results of K- folding cross validation on monks-2. train

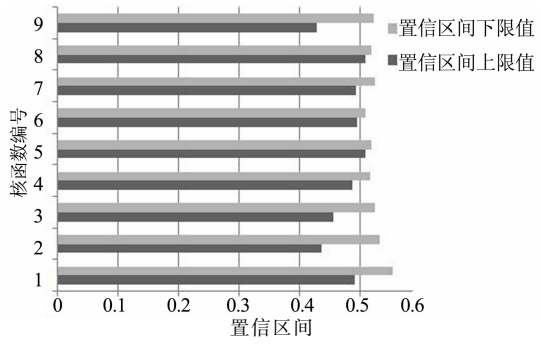

图 3　monks-3. train 的 K-折交叉验证实验结果

Fig. 3　Experimental results of K- folding cross validation on monks-3. train

如图 4 和 5 所示,9 个核函数(用编号表示)被两两对比,顺序为(1,2),(1,2),(1,3),…,(8,9),依次对应横坐标中的 36 个点(1～36). 对于上述括号中的两个核函数,若前者更好则标记为"1",若后者更好则标记为"－1",若两者相等则标记为"0",HTF 核函数参与对比的点的横坐标为 8,15,21,26,30,33,35,36 ,而这些横坐标的值均为"－1",这说明在 monks-2. train 和 monks-3. train 数据集上分类效果最好的为 HTF 核函数,而且基于错误距离的秩和检验在绝大多数的核函数两两对比实验中均能给出明确的判定结果,这是配对 t 测试和 K-折交叉验证方法所无法得到的.

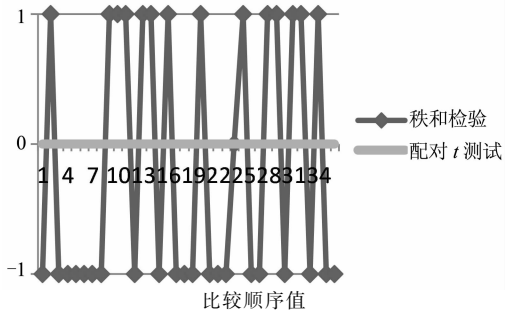

图 4　monks-2. train 实验结果

Fig. 4　Experimental results on monks-2. train

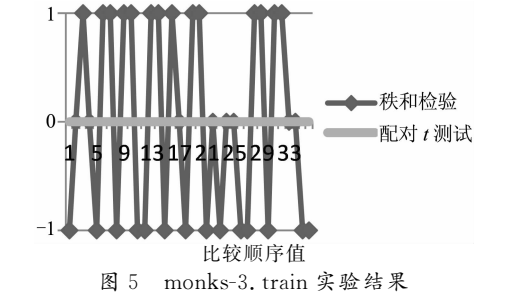

图 5　monks-3. train 实验结果

Fig. 5　Experimental results on monks-3. train

根据上面的分析,由表 5 和图 4～5 可以得出:1)K-折交叉验证、配对 t 测试与秩和检验得到的核函数的综合排序在大体上是一致的,说明秩和检验

可以对核函数的分类能力进行客观评估;2)在数据集的 Errcv(T,D) 不适合使用参数检验方法的情况下,秩和检验却可以对核函数分类能力进行更好的评估.因此,与 K-折交叉验证和配对 t 测试等方法相比,基于错分实例到支持向量所在界面的距离的秩和检验方法具有更高的可行性.

3 结 论

核函数的选择是核方法研究及应用的核心内容,选择的准则和方法目前并没有成型的理论方法,研究人员更多地是凭借主观经验进行选择,因此具有较大的随意性.通过实际数据的验证结果来指导核函数的选择是最常用的方法之一,本文针对参数检验方法的局限性,将秩和检验这一非参数检验方法引入核函数选择中,提出了基于分类错误的实例与支持向量所在的决策界面的距离进行秩和检验的核函数选择方法,实验结果验证了该方法的合理性,并在某些数据集上给出了更好的选择效果.

参数检验方法需要对总体分布进行假定,因此可能会引起推断结果的错误.本文提出的以错分实例到支持向量所在界面的距离来表示错分程度,并基于此进行秩和检验的核函数选择方法,并不需要考虑样本期望和方差,而只需比较其总体位置,因此与参数检验方法相比其适应性更强.另外,错分程度也是参数检验中所没有考虑的因素,在数据集的 Errcv(T,D) 不适合使用参数检验方法的情况下,所提方法能得到较好的结果.另外,本文的方法可以和其他参数检验(如 K-折交叉验证、配对 t 测试等)方法配合使用、相互验证核函数选择的准确性,并且可以在参数检验方法无法分辨核函数优劣的情况下进一步区分核函数的分类性能.

参考文献

[1] VAPNIK V. The nature of statistical learning theory [M]. The second edition. New York: Springer-Verlag, 2000:1-314.

[2] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011, 40(1):2-10.
DING S F, QI B J, TAN H Y. An overview on theory and algorithm of support vector machines [J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1):2-10. (In Chinese)

[3] GHOUTI L. A new kernel-based classification algorithm for multi-label datasets [J]. Arabian Journal for Science and Engineering, 2016, 41(3):759-771.

[4] 赵金伟,冯博琴,闫桂荣. 泛化的统一切比雪夫多项式核函数[J]. 西安交通大学学报, 2012,46(8):43-48.
ZHAO J W, FENG B Q, YAN G R. Generalized uniform Chebyshev polynomial kernel[J]. Journal of Xi'an Jiaotong University, 2012,46(8):43-48. (In Chinese)

[5] ZHAO J W, YAN G R. FENG B Q, *et al*. An adaptive support vector regression based on a new sequence of unified orthogonal polynomials[J]. Pattern Recognition, 2013, 46(3):899-913.

[6] 业巧林,业宁,张训华. 基于极分解下的混合核函数及改进[J]. 模式识别与人工智能, 2009,22(3):366-373.
YE Q L, YE N, ZHANG X H. Extremun decom position based mixteres of kernels and its improvement[J]. Pattern Recognition and Artificial Intelligence, 2009,22(3):366-373. (In Chinese)

[7] MIAO S W, XIE K G, YANG H J, *et al*. A mixture kernel density model for wind speed probability distribution estimation [J]. Energy Conversion and Management, 2016, 126(15): 1066-1083.

[8] TIAN J, YU W Y, XIE S L. On the kernel function selection of nonlocal filtering for image denoising[C]// Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. Kunming, 2008:2964-1969.

[9] ZHU B, CHENG Z D, WANG H. A kernel function optimization and selection algorithm based on cost function maximization[C]// 2013 IEEE International Conference on Imaging Systems and Techniques (IST). 2013:259-263.

[10] WANG S F, NIE B, YUE K, *et al*. Protein subcellular localization with Gaussian kernel discriminant analysis and its kernel parameter selection[J]. International Journal of Molecular Sciences, 2017, 18(12):1-16.

[11] XIAO Y C, WANG H G, ZHANG L, *et al*. Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection[J]. Knowledge-based System, 2014, 59:75-84.

[12] XIONG H L, SWAMY M N S, AHMAD M O. Optimizing the kernel in the empirical feature space[J]. IEEE Transactions on Neural Networks, 2005, 16(2):460-474.

[13] 田萌,王文剑. 高斯核函数选择的广义核极化准则[J]. 计算机研究与发展, 2015,52(8):1722-1734.
TIAN M, WANG W J. Generalized kernel polarization criterion for optimizing Gaussian kernel[J]. Journal of Computer Research and Development, 2015,52(8):1722-1734. (In Chinese)

[14] 梁礼明,冯新刚,陈云嫩,等. 基于样本分布特征的核函数选择方法研究[J]. 计算机仿真, 2013, 30(1):323-328.
LIANG L M, FENG X G, CHEN Y N, *et al*. Method of selection kernel function based on distribution characteristics of samples [J]. Computer Simulation, 2013, 30(1):323-328. (In Chinese)

[15] LIU Y, LIAO S Z. Kernel selection with spectral perturbation stability of kernel matrix [J]. Science China(Information Sciences), 2014,57(11):112103.

[16] 胡包钢,王泳. 应用统计方法综合评估核函数分类能力的研究[J]. 计算机学报,2008,31(6): 942-952.
HU B G, WANG Y. A study on integrated evaluating kernel classification performance using statistical methods [J]. Chinese Journal of Computers, 2008,31(6):942-952. (In Chinese)

[17] BROWNE M W. Cross-validation methods [J]. Journal of Mathematical Psychology, 2000, 4(1):108-132.

[18] SINCICH T. Business statistics by example [M]. The fifth edition. New Jersey: Prentice Hall, 1996:1-1179.

[19] 茆诗松,程依明,濮晓龙. 概率论与数理统计 [M]. 第二版. 北京:高等教育出版社, 2011:1-523.
MAO S S, CHENG Y M, PU X L. Probability theory & mathematical statistics [M]. The second edition. Beijing: Higher Education Press, 2011:1-523. (In Chinese)

[20] NEWMAN D J, HETTICH S, BLAKE C L, *et al*. UCI repository of machine learning databases[D]. Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[21] Statlib—Data, Software and News from the Statistics Community. [http://lib.stat.cmu.edu/datasets/]