软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):2856−2868 [doi: 10.13328/j.cnki.jos.004905]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

普适性核度量标准比较研究*

王裴岩[1,2], 蔡东风[2]

[1](南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

[2](沈阳航空航天大学 人机智能研究中心,辽宁 沈阳 110136)

通讯作者: 王裴岩, E-mail: wangpy@sau.edu.cn, http://www.sau.edu.cn

摘 要: 核方法是一类应用较为广泛的机器学习算法,已被应用于分类、聚类、回归和特征选择等方面.核函数的选择与参数优化一直是影响核方法效果的核心问题,从而推动了核度量标准,特别是普适性核度量标准的研究.对应用最为广泛的5种普适性核度量标准进行了分析与比较研究,包括KTA,EKTA,CKTA,FSM和KCSM.发现上述5种普适性度量标准的度量内容为特征空间中线性假设的平均间隔,与支持向量机最大化最小间隔的优化标准存在偏差.然后,使用模拟数据分析了上述标准的类别分布敏感性、线性平移敏感性、异方差数据敏感性,发现上述标准仅是核度量的充分非必要条件,好的核函数可能获得较低的度量值.最后,在9个UCI数据集和20Newsgroups数据集上比较了上述标准的度量效果,发现CKTA是度量效果最好的普适性核度量标准.

关键词: 核方法;核选择;核参数优化;普适性核度量标准

中图法分类号: TP181

中文引用格式: 王裴岩,蔡东风.普适性核度量标准比较研究.软件学报,2015,26(11):2856−2868. http://www.jos.org.cn/1000-9825/4905.htm

英文引用格式: Wang PY, Cai DF. Comparative study of universal kernel evaluation measures. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):2856−2868 (in Chinese). http://www.jos.org.cn/1000-9825/4905.htm

Comparative Study of Universal Kernel Evaluation Measures

WANG Pei-Yan[1,2], CAI Dong-Feng[2]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

[2](Human Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: Kernel method is a common machine learning algorithm used in classification, clustering, regression and feature selection. Kernel selection and kernel parameter optimization are the crucial problems which impact the effectiveness of kernel method, and therefore motive the research on kernel evaluation measure, especially universal kernel evaluation measure. Five widely used universal kernel evaluation measures, including KTA, EKTA, CKTA, FSM and KCSM, are analyzed and compared. It is found that the evaluation object of five universal kernel evaluation measures mentioned above is average margin of a linear hypothesis in feature space, which has bias against the SVM optimization criterion to maximize minimum margin. Then, this study applies synthetic data to analyze the class distribution sensitivity, linear translation sensitivity, and heteroscedastic data sensitivity. It also concludes that the measures mentioned above are only the unnecessary and sufficient condition of kernel evaluation, and good kernel can achieve low evaluation value. Finally, comparing the evaluation result of the measures mentioned above on 9 UCI data sets and 20 Newsgroups data set suggests that CKTA is the best universal kernel evaluation measure.

Key words: kernel method; kernel selection; kernel parameter optimization; universal kernel evaluation measure

基于核函数的机器学习方法,简称核方法,是机器学习领域的一类重要方法,被广泛地应用于分类、聚类、

* 基金项目: 国家自然科学基金(61402299)
收稿时间: 2015-05-31; 修改时间: 2015-07-14, 2015-08-11; 定稿时间: 2015-08-26

回归和特征选择等方面[1].最具有代表性的方法有:支持向量机、谱聚类、岭回归、核主成分分析等.文献[2]在121个数据集上比较研究了179种分类器的性能,发现基于核函数的支持向量机与极限学习机是5种最优的分类器之一,其效果明显好于其他分类器.然而,核函数的选择与参数优化一直是影响核方法效果的核心问题[3],从而推动了核度量标准,特别是普适性核度量标准(universal kernel evaluation measure)[4]的研究.

普适性核度量标准不直接估计泛化误差界,仅依据给定的问题和样本对核函数质量做出量化评价,与留一法[5]和Span-Bound法[6]等泛化误差界的直接估计法相比具有较高的计算效率,计算代价仅为$O(n^2)$(n为样本容量),与结构风险(structural risk)[7,8]、负对数后验(negative log-posterior)[9]和超核(hyperkernels)[10]等方法相比具有算法无关性,不依赖于具体核学习算法与核函数,具有较好的推广能力.其中,KTA(kernel target alignment)[11]是最早被提出的普适性核度量标准,因此也是应用较为广泛的标准之一.文献[4]综述了KTA的基本思想、理论特性、在多核学习、特征选择与核函数参数优化等方面的应用情况以及与其他普适性度量标准间的关系.

在KTA之后提出的其他普适性度量标准力争沿袭KTA的优点,并对KTA存在的问题进行了改进,其中,EKTA(extension of kernel target alignment)[12]与CKTA(centered kernel target alignment)[13]同样基于KTA的Alignment的基本思想.EKTA采用了每类样本数量对目标矩阵进行了调整,目的是解决KTA的类别分布敏感性问题.CKTA首先将核函数进行了中心化,然后计算与目标矩阵的相似程度.中心化可以消除由于样本远离原点而产生的病态核矩阵的问题[14].

FSM(feature space based kernel matrix evaluation measures)[15]首次指出KTA存在线性变换敏感性问题,并针对该问题进行了改进.FSM与KTA的Alignment基本思想不同,直接度量样本在特征空间中的可分离性,为特征空间中正负类中心的距离与特征空间中同类样本在正负类中心所确定的方向上总偏差的比值.FSM的基本思想与KCSM(kernel class separablity measures)最接近,KCSM是Fisher线性判别准则[16,17]在核函数度量上的应用,为样本在特征空间中类间散布程度和类内散布程度的比值.KCSM被应用于特征选择[18,19]、核函数参数优化[20]等方面.

由此可见,KTA,EKTA,CKTA,FSM与KCSM彼此之间存在一定的相关性,除KCSM外,其他标准都是针对KTA所存在的问题进行的改进,并且EKTA,CKTA与KTA属同族标准,KCSM与FSM属同族标准.因此,对上述5种算法进行比较分析,可进一步揭示其内在相关性以及发现类别分布敏感性与线性变换敏感性等问题的产生原因,为解决上述问题提出新的度量标准提供依据.

对KTA,EKTA,CKTA,FSM与KCSM进行比较研究:首先,发现了上述5种普适性度量标准具有较为相近的形式,可在统一的框架下进行研究与比较,并且发现其度量内容为特征空间中线性假设的平均间隔,与支持向量机最大化最小间隔的优化目标存在偏差;然后,使用模拟数据研究了类别分布敏感性、线性平移敏感性、异方差数据敏感性,指出5种度量标准产生上述问题的原因,并指出该5种普适性度量标准都是核度量的充分非必要条件,好的核函数可能获得较低的度量值;最后,通过在9个UCI数据集和20Newsgroups数据集上的核函数选择实验比较了5种度量标准的度量效果,由于CKTA解决了KTA的线性变换敏感性问题并且受异方差数据影响较弱,是该5种度量标准中度量效果最好的普适性核度量标准.

1 对5种度量标准的分析

考虑二分类问题,$X\subset\mathbb{R}^n$为样本空间,$Y=\{-1,+1\}$为标记集.D代表在$X\times Y$上确定但未知的概率分布,样本集$\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l)\}$依据分布$D$独立同分布抽取.并且有$l_+$个样本属于“+1”类,有$l_-$个样本属于“−1”类,$l=l_++l_-$.每个样本$\boldsymbol{x}_i$通过核函数映射到特征空间$H$中的$\phi(\boldsymbol{x}_i)$:

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle,\phi:X\to H.$$

核函数k的核矩阵为

$$[\boldsymbol{K}]_{i,j}=k(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

1.1 KTA

KTA 由 Cristianini 等人[11]提出,是最早被提出的核度量标准.对于二分类问题,KTA 计算核矩阵与理想目标矩阵的对齐程度.定义理想目标矩阵为

$$[\boldsymbol{Y}]_{i,j}=y_i\cdot y_j=\begin{cases}+1, & y_i=y_j\\ -1, & y_i\neq y_j\end{cases}.$$

定义 KTA 为

$$KTA(\boldsymbol{K},\boldsymbol{Y})=\frac{\langle \boldsymbol{K},\boldsymbol{Y}\rangle_F}{\sqrt{\langle \boldsymbol{K},\boldsymbol{K}\rangle_F}\sqrt{\langle \boldsymbol{Y},\boldsymbol{Y}\rangle_F}},$$

其中,$\langle\cdot,\cdot\rangle_F$ 为 Frobenius 内积.因此,KTA 为归一化后的核矩阵与理想目标矩阵间的 Frobenius 内积.KTA 取值区间为[−1,+1],值越高,表明核矩阵与理想目标矩阵对齐程度越高,核函数越好.

由 $\boldsymbol{Y}$ 的定义可知$\langle \boldsymbol{Y},\boldsymbol{Y}\rangle_F=l^2$,$KTA$ 可转化为如下形式:

$$KTA(\boldsymbol{K},\boldsymbol{Y})=\frac{1}{\sqrt{\frac{1}{l^2}\langle \boldsymbol{K},\boldsymbol{K}\rangle_F}}\hat{A}_u(\boldsymbol{K},\boldsymbol{Y}),$$

其中,$\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})=\frac{1}{l^2}\langle \boldsymbol{K},\boldsymbol{Y}\rangle_F$. 本文将在第 2.1 节讨论 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 与特征空间中线性假设平均间隔间的关系.

1.2 CKTA

Corinna 等人[13]提出了 CKTA,并将其用于多核学习.文献[21]将其用于基于连续基核的多核学习任务.文献[22]将其应用于多核聚类任务.CKTA 使用了中心化核函数代替了 KTA 中的核函数,中心化核函数定义如下:

$$\begin{aligned}k_C(\boldsymbol{x}_1,\boldsymbol{x}_2)&=\langle\boldsymbol{\phi}(\boldsymbol{x}_1)-\bar{\boldsymbol{\phi}},\boldsymbol{\phi}(\boldsymbol{x}_2)-\bar{\boldsymbol{\phi}}\rangle\\&=\left\langle\boldsymbol{\phi}(\boldsymbol{x}_1)-\frac{1}{n}\sum_{i=1}^{l}\boldsymbol{\phi}(\boldsymbol{x}_i),\boldsymbol{\phi}(\boldsymbol{x}_2)-\frac{1}{n}\sum_{i=1}^{l}\boldsymbol{\phi}(\boldsymbol{x}_i)\right\rangle\\&=k(\boldsymbol{x}_1,\boldsymbol{x}_2)-\frac{1}{n}\sum_{i=1}^{n}k(\boldsymbol{x}_i,\boldsymbol{x}_2)-\frac{1}{n}\sum_{j=1}^{n}k(\boldsymbol{x}_1,\boldsymbol{x}_j)+\frac{1}{n^2}\sum_{i,j=1}^{n}k(\boldsymbol{x}_i,\boldsymbol{x}_j).\end{aligned}$$

$\bar{\boldsymbol{\phi}}$ 为特征空间中总体样本的中心向量.中心化的核矩阵 $\boldsymbol{K}_C$ 计算方法如下:

$$\boldsymbol{K}_C=\boldsymbol{K}-\frac{1}{l}\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{K}-\frac{1}{l}\boldsymbol{K}\boldsymbol{1}\boldsymbol{1}^T+\frac{1}{l^2}(\boldsymbol{1}^T\boldsymbol{K}\boldsymbol{1})\boldsymbol{1}\boldsymbol{1}^T=\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]\boldsymbol{K}\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right],$$

其中,$\boldsymbol{1}\in\mathbb{R}^{l\times 1}$ 表示元素全为 1 的向量,$\boldsymbol{I}$ 为单位矩阵.不难得出,$\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]=\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]$. CKTA 的形式如下:

$$CKTA(\boldsymbol{K}_C,\boldsymbol{Y})=\frac{\langle \boldsymbol{K}_C,\boldsymbol{Y}\rangle_F}{\sqrt{\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F}\sqrt{\langle \boldsymbol{Y},\boldsymbol{Y}\rangle_F}}.$$

与 KTA 类似,CKTA 可转换为如下形式:

$$CKTA(\boldsymbol{K},\boldsymbol{Y})=\frac{1}{\sqrt{\frac{1}{l^2}\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F}}\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y}).$$

1.3 EKTA

Kandola[12]提出了 EKTA,使用了样本容量对样本类别标记做如下调整:

$$y_i=\begin{cases}\frac{1}{l_+}, & y_i=1\\ -\frac{1}{l_-}, & y_i=-1\end{cases},$$

则目标矩阵 $\boldsymbol{Y}_E$ 为

$$[\boldsymbol{Y}_E]_{i,j}=y_i\cdot y_j=\begin{cases}\dfrac{1}{l_+^2}, & y_i=y_j=1\\ \dfrac{1}{l_-^2}, & y_i=y_j=-1.\\ -\dfrac{1}{l_+l_-}, & y_i\neq y_j\end{cases}$$

引理 1. 对于任意对称半正定核矩阵 $\boldsymbol{K}\in\mathbb{R}^{l\times l}$ 与目标矩阵 $\boldsymbol{Y}$:

$$\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F=\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F\cdot\left(\frac{l}{2l_+l_-}\right)^2.$$

证明:由矩阵中心化的定义,中心化目标矩阵 $\boldsymbol{Y}_C$ 与 $\boldsymbol{Y}_E$ 具有如下关系:

$$\frac{\boldsymbol{Y}_E}{\sqrt{\langle\boldsymbol{Y}_E,\boldsymbol{Y}_E\rangle_F}}=\frac{\boldsymbol{Y}_C}{\sqrt{\langle\boldsymbol{Y}_C,\boldsymbol{Y}_C\rangle_F}}.$$

因此,

$$\frac{\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F}{\sqrt{\langle\boldsymbol{Y}_E,\boldsymbol{Y}_E\rangle_F}}=\frac{\langle\boldsymbol{K},\boldsymbol{Y}_C\rangle_F}{\sqrt{\langle\boldsymbol{Y}_C,\boldsymbol{Y}_C\rangle_F}}.$$

由于 $\sqrt{\langle\boldsymbol{Y}_C,\boldsymbol{Y}_C\rangle_F}=4\cdot\frac{l_+l_-}{l}$,$\sqrt{\langle\boldsymbol{Y}_E,\boldsymbol{Y}_E\rangle_F}=\frac{1}{l_+l_-}$, 可得出:

$$\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F=\langle\boldsymbol{K},\boldsymbol{Y}_C\rangle_F\cdot\left(\frac{l}{2l_+l_-}\right)^2.$$

由 $\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]=\left[\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^T}{l}\right]$,可得$\langle\boldsymbol{K},\boldsymbol{Y}_C\rangle_F=\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F$,于是有:

$$\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F=\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F\cdot\left(\frac{l}{2l_+l_-}\right)^2.$$ □

依据引理 1 可将 EKTA 转化为

$$EKTA(\boldsymbol{K},\boldsymbol{Y}_E)=\frac{\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F}{\sqrt{\langle\boldsymbol{K},\boldsymbol{K}\rangle_F}\sqrt{\langle\boldsymbol{Y}_E,\boldsymbol{Y}_E\rangle_F}}=\frac{\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F\left(\frac{l}{2l_+l_-}\right)^2}{\sqrt{\langle\boldsymbol{K},\boldsymbol{K}\rangle_F}\sqrt{\langle\boldsymbol{Y}_E,\boldsymbol{Y}_E\rangle_F}}=\frac{l^2}{4l_+l_-}\frac{1}{\sqrt{\frac{1}{l^2}\langle\boldsymbol{K},\boldsymbol{K}\rangle_F}}\widehat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y}).$$

1.4 FSM

FSM 为特征空间中正负类中心的距离与特征空间中同类样本在正负类中心所确定的方向上总偏差的比值:

$$FSM(k)=\frac{std}{\|\phi_+-\phi_-\|},$$

$$std=std_++std_-=\sqrt{\frac{\sum_{i=1}^{n_+}\langle\phi(\boldsymbol{x}_i)-\phi_+,e\rangle^2}{l_+-1}}+\sqrt{\frac{\sum_{i=1}^{n_-}\langle\phi(\boldsymbol{x}_i)-\phi_-,e\rangle^2}{l_--1}},\text{其中},e=\frac{\phi_+-\phi_-}{\|\phi_+-\phi_-\|}.$$

ϕ_+为正类样本在特征空间的中心,ϕ_-为负类样本在特征空间的中心,$\|\phi_+-\phi_-\|$为两类中心的距离,std_+与 std_-分别为特征空间中正类与负类样本在两类中心所确定的方向上的偏差.FSM 值越小,则表明核函数越好.由ϕ_+,ϕ_-与核距离的定义可得:

$$\begin{aligned}\|\phi_+-\phi_-\|^2&=\langle\phi_+,\phi_+\rangle+\langle\phi_-,\phi_-\rangle-2\langle\phi_+,\phi_-\rangle\\&=\left\langle\frac{1}{l_+}\sum_{i=1}^{l_+}\phi_i,\frac{1}{l_+}\sum_{i=1}^{l_+}\phi_i\right\rangle+\left\langle\frac{1}{l_-}\sum_{i=1}^{l_-}\phi_i,\frac{1}{l_-}\sum_{i=1}^{l_-}\phi_i\right\rangle-2\left\langle\frac{1}{l_+}\sum_{i=1}^{l_+}\phi_i,\frac{1}{l_-}\sum_{i=1}^{l_-}\phi_i\right\rangle\\&=\frac{1}{l_+^2}\sum_{i=1,j=1}^{l_+,l_+}\langle\phi_i,\phi_j\rangle+\frac{1}{l_-^2}\sum_{i=1,j=1}^{l_-,l_-}\langle\phi_i,\phi_j\rangle-2\frac{1}{l_+l_-}\sum_{i=1,j=1}^{l_+,l_-}\langle\phi_i,\phi_j\rangle\\&=\frac{1}{l_+^2}\sum_{i=1,j=1}^{l_+,l_+}k(\boldsymbol{x}_i,\boldsymbol{x}_j)+\frac{1}{l_-^2}\sum_{i=1,j=1}^{l_-,l_-}k(\boldsymbol{x}_i,\boldsymbol{x}_j)-2\frac{1}{l_+l_-}\sum_{i=1,j=1}^{l_+,l_-}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&=\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F.\end{aligned}$$

再由引理 1,$\|\phi_+-\phi_-\|^2=\langle\boldsymbol{K},\boldsymbol{Y}_E\rangle_F=\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F\cdot\left(\frac{l}{2l_+l_-}\right)^2$. 定义 std_u 与 e_u 为

$$std_u=\sqrt{\frac{\sum_{i=1}^{n_+}\langle\phi(\boldsymbol{x}_i)-\phi_+,e_u\rangle^2}{l_+-1}}+\sqrt{\frac{\sum_{i=1}^{n_-}\langle\phi(\boldsymbol{x}_i)-\phi_-,e_u\rangle^2}{l_--1}},\ e_u=\phi_+-\phi_-,$$

则 FSM 可转化为

$$\frac{1}{FSM(k)}=\frac{\|\phi_+-\phi_-\|^2}{std_u}=\frac{l^2}{4l_+l_-}\cdot\frac{l^2}{l_+l_-\cdot\left(\sqrt{\frac{1}{l_+-1}\sum_{i=1}^{n_+}\langle\phi(\boldsymbol{x}_i)-\phi_+,\phi_+-\phi_-\rangle^2}+\sqrt{\frac{1}{l_--1}\sum_{i=1}^{n_-}\langle\phi(\boldsymbol{x}_i)-\phi_-,\phi_+-\phi_-\rangle^2}\right)}\widehat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y}).$$

1.5 KCSM

KCSM 的形式如下,其与 Fisher 线性判别分析的形式一致:

$$KCSM(k)=\frac{Tr(\boldsymbol{S}_B)}{Tr(\boldsymbol{S}_W)}.$$

$\boldsymbol{S}_B$ 和 $\boldsymbol{S}_W$ 分别表示类间散布矩阵和类内散布矩阵;Tr 代表矩阵的迹,对于二分类问题:

$$Tr(\boldsymbol{S}_B)=l_+\|\phi_+-\bar{\phi}\|^2+l_-\|\phi_--\bar{\phi}\|^2,$$

$$Tr(\boldsymbol{S}_W)=\sum_{i=1}^{l_+}\|\phi(\boldsymbol{x}_i)-\phi_+\|^2+\sum_{i=1}^{l_-}\|\phi(\boldsymbol{x}_i)-\phi_-\|^2.$$

$\|\phi_+-\bar{\phi}\|$ 与 $\|\phi_--\bar{\phi}\|$ 为类中心到样本中心的距离,$\|\phi(\boldsymbol{x}_i)-\phi_+\|$ 为样本到类中心的距离.

$$\|\phi_+-\bar{\phi}\|^2=\langle\phi_+,\phi_+\rangle+\langle\bar{\phi},\bar{\phi}\rangle-2\langle\phi_+,\bar{\phi}\rangle=\frac{1}{l_+^2}\sum_{i,j=1}^{l_+}k(\boldsymbol{x}_i,\boldsymbol{x}_j)+\frac{1}{l^2}\sum_{i,j=1}^{l}k(\boldsymbol{x}_i,\boldsymbol{x}_j)-2\frac{1}{l_+}\frac{1}{l}\sum_{i=1}^{l_+}\sum_{j=1}^{l}k(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

$$\|\phi_--\bar{\phi}\|^2=\langle\phi_-,\phi_-\rangle+\langle\bar{\phi},\bar{\phi}\rangle-2\langle\phi_-,\bar{\phi}\rangle=\frac{1}{l_-^2}\sum_{i,j=1}^{l_-}k(\boldsymbol{x}_i,\boldsymbol{x}_j)+\frac{1}{l^2}\sum_{i,j=1}^{l}k(\boldsymbol{x}_i,\boldsymbol{x}_j)-2\frac{1}{l_-}\frac{1}{l}\sum_{i=1}^{l_-}\sum_{j=1}^{l}k(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

将 $\|\phi_+-\bar{\phi}\|^2$ 与 $\|\phi_--\bar{\phi}\|^2$ 带入 $Tr(\boldsymbol{S}_B)$:

$$Tr(\boldsymbol{S}_B)=\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F\cdot\frac{l}{4l_+l_-}.$$

由此可得:

$$KCSM(k)=\frac{\frac{l^3}{4l_+l_-}\cdot\frac{1}{l^2}\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F}{\sum_{i=1}^{l_+}\|\phi(\boldsymbol{x}_i)-\phi_+\|^2+\sum_{i=1}^{l_-}\|\phi(\boldsymbol{x}_i)-\phi_-\|^2}=\frac{l^2}{4l_+l_-}\frac{1}{\frac{1}{l}\sum_{i=1}^{l_+}\|\phi(\boldsymbol{x}_i)-\phi_+\|^2+\frac{1}{l}\sum_{i=1}^{l_-}\|\phi(\boldsymbol{x}_i)-\phi_-\|^2}\widehat{A}_u\langle\boldsymbol{K}_C,\boldsymbol{Y}\rangle_F.$$

2 讨 论

本节首先讨论上述 5 种普适性核度量标准与特征空间中线性假设平均间隔的关系,指出上述 5 种度量标准

都是对特征空间中线性假设平均间隔的度量;然后,分别讨论类别分布敏感性、线性平移敏感性和异方差数据敏感性,分析上述问题的产生原因.

2.1 特征空间中线性假设的期望间隔

首先定义特征空间中对偶形式的线性假设,并给出该线性假设的期望间隔.

定义 1(特征空间中线性假设). k 为定义在 $X\times X$ 上的对称半正定核,并且有$(\boldsymbol{x}',y')\in X\times Y$.对于任意 $\boldsymbol{x}\in X$,定义特征空间中线性假设 $h^*(\boldsymbol{x})$为

$$h^*(\boldsymbol{x})=E_{\boldsymbol{x}'}[\alpha' y' k(\boldsymbol{x},\boldsymbol{x}')],\ \alpha'\in\mathbb{R},\ \alpha'\geqslant 0.$$

对于容量为 l 的样本,基于该样本的特征空间中经验线性假设为

$$h^*(\boldsymbol{x})=\frac{1}{l}\sum_{i=1}^{l}\alpha_i y_i k(\boldsymbol{x},\boldsymbol{x}_i),\ \alpha_i\in\mathbb{R},\ \alpha_i\geqslant 0.$$

定义 2(线性假设的期望间隔). 给定任意的$(\boldsymbol{x},y)\in X\times Y$,$h^*(\boldsymbol{x})$的函数间隔为 $yh^*(\boldsymbol{x})$,那么基于分布 D 的期望间隔为

$$E_{\boldsymbol{x}}[yh^*(\boldsymbol{x})]=E_{\boldsymbol{x}}[yE_{\boldsymbol{x}'}[\alpha' y' k(\boldsymbol{x},\boldsymbol{x}')]]=E_{\boldsymbol{x},\boldsymbol{x}'}[\alpha' yy' k(\boldsymbol{x},\boldsymbol{x}')],\ \alpha'\in\mathbb{R},\ \alpha'\geqslant 0.$$

对于容量为 l 的样本,在该样本上的平均间隔为

$$\hat{E}_{\boldsymbol{x}_i}[yh^*(\boldsymbol{x})]=\frac{1}{l}\sum_{i=1}^{l}y_i\left[\frac{1}{l}\sum_{j=1}^{l}\alpha_j y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]=\frac{1}{l^2}\sum_{i=1,j=1}^{l,l}\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j),\ \alpha_j\in\mathbb{R},\ \alpha_j\geqslant 0.$$

定理 1 给出了期望间隔与 $h^*(\boldsymbol{x})$泛化误差间的关系.

定理 1(线性假设的泛化误差上界). 令 $R(h^*)$为 $h^*(\boldsymbol{x})$的泛化误差.对于任意对称半正定核函数 k,$\sup_{\boldsymbol{x},\boldsymbol{x}'\in X}k(\boldsymbol{x},\boldsymbol{x}')\leqslant B$,$\sup_{\boldsymbol{x}\in X}\alpha\leqslant A$,有: $R(h^*)\leqslant 1-\dfrac{E_{\boldsymbol{x}}[yh^*(\boldsymbol{x})]}{AB}$.

证明:依据 Cauchy-Schwarz 不等式:

$$|E_{\boldsymbol{x}}[yh^*(\boldsymbol{x})]|=|E_{\boldsymbol{x},\boldsymbol{x}'}[\alpha' yy' k(\boldsymbol{x},\boldsymbol{x}')]|\leqslant\sqrt{E_{\boldsymbol{x},\boldsymbol{x}'}[(\alpha' yy')^2]E_{\boldsymbol{x},\boldsymbol{x}'}[k^2(\boldsymbol{x},\boldsymbol{x}')]}=\sqrt{E_{\boldsymbol{x},\boldsymbol{x}'}[(\alpha')^2]E_{\boldsymbol{x},\boldsymbol{x}'}[k^2(\boldsymbol{x},\boldsymbol{x}')]}\leqslant AB.$$

基于上式:

$$1-R(h^*)=\Pr[yh^*(\boldsymbol{x})\geqslant 0]=E[\mathbf{1}_{\{yh^*(\boldsymbol{x})\geqslant 0\}}]\geqslant E\left[\frac{yh^*(\boldsymbol{x})}{AB}\mathbf{1}_{\{yh^*(\boldsymbol{x})\geqslant 0\}}\right]\geqslant\frac{E[yh^*(\boldsymbol{x})]}{AB},$$

其中,$\mathbf{1}_e$ 为事件 e 的指示函数,当 e 发生时,$\mathbf{1}_e$ 的值为 1. □

由定理 1 可见,$h^*(\boldsymbol{x})$的泛化误差上界与期望间隔成反比例关系.即,增大期望间隔可降低泛化误差上界.下述定理保证最大化 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 与最大化平均间隔是等价的.

定理 2. 假设α'与 $y'k(\boldsymbol{x},\boldsymbol{x}')$相互独立,对于容量为 l 的样本,最大化 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 与最大化平均间隔 $\hat{E}_{\boldsymbol{x}_i}[yh^*(\boldsymbol{x})]$ 是等价的.

证明:由于α'与 $y'k(\boldsymbol{x},\boldsymbol{x}')$相互独立:

$$E_{\boldsymbol{x}}[yh^*(\boldsymbol{x})]=E_{\boldsymbol{x}}[E_{\boldsymbol{x}'}[\alpha' yy' k(\boldsymbol{x},\boldsymbol{x}')]]=E_{\boldsymbol{x}}[E_{\boldsymbol{x}'}[\alpha']E_{\boldsymbol{x}'}[yy' k(\boldsymbol{x},\boldsymbol{x}')]]=E[\alpha']E[yy' k(\boldsymbol{x},\boldsymbol{x}')].$$

对于容量为 l 的样本:

$$\hat{E}_{\boldsymbol{x}_i}[yh^*(\boldsymbol{x})]=\frac{1}{l}\sum_{j=1}^{l}\alpha_j\frac{1}{l^2}\sum_{i=1,j=1}^{l,l}y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

由 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 的定义以及$\alpha_j\geqslant 0$ 可得:

$$\hat{E}_{\boldsymbol{x}_i}[yh^*(\boldsymbol{x})]=\frac{1}{l}\sum_{j=1}^{l}\alpha_j\cdot\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})=\hat{E}(\alpha_j)\cdot\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})\propto\hat{A}_u(\boldsymbol{K},\boldsymbol{Y}),$$

其中,$\hat{E}(\alpha_j)=\dfrac{1}{l}\sum_{j=1}^{l}\alpha_j$. □

依据定理 2,将 KTA,EKTA,CKTA,FSM 与 KCSM 的 $\hat{E}(\alpha_j)$ 项与 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 项列于表 1 中.可见,5 种度量标准的差异主要体现在了 $\hat{E}(\alpha_j)$ 项上,并可得出如下结论:KTA,EKTA,CKTA,FSM 与 KCSM 是在核函数所构造的特征空间中对线性假设的平均间隔的度量.若某核函数具有较高度量值,那么在由其构造的特征空间中,线性假设将具有较高的平均间隔.需要特别指出的是,这与支持向量机最大化最小间隔的优化标准存在偏差.通常情况下,平均间隔与最小间隔具有较大差异,因此,对于支持向量机更适于直接度量最小间隔而不是平均间隔.然而,平均间隔与最小间隔相比具有较好的统计稳定性,因此需要在统计稳定性与度量偏差间进行折中,可使用 k 间隔[23]或使用靠近边界样本的平均间隔近似最小间隔.

Table 1 $\hat{E}(\alpha_j)$ and $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ of universal kernel evaluation measures

表 1 普适性度量标准的 $\hat{E}(\alpha_j)$ 项与 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 项

核度量标准	$\hat{E}(\alpha_j)$ 项	$\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$ 项
KTA	$\dfrac{1}{\sqrt{\dfrac{1}{l^2}\langle \boldsymbol{K},\boldsymbol{K}\rangle_F}}$	$\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$
EKTA	$\dfrac{l^2}{4l_+l_-}\dfrac{1}{\sqrt{\dfrac{1}{l^2}\langle \boldsymbol{K},\boldsymbol{K}\rangle_F}}$	$\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$
CKTA	$\dfrac{1}{\sqrt{\dfrac{1}{l^2}\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F}}$	$\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$
FSM	$\dfrac{l^2}{4l_+l_-}\cdot\dfrac{l^2}{l_+l_-\cdot\left(\sqrt{\dfrac{1}{l_+-1}\sum_{i=1}^{n_+}\langle\phi(\boldsymbol{x}_i)-\phi_+,\phi_+-\phi_-\rangle^2}+\sqrt{\dfrac{1}{l_--1}\sum_{i=1}^{n_-}\langle\phi(\boldsymbol{x}_i)-\phi_-,\phi_+-\phi_-\rangle^2}\right)}$	$\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$
KCSM	$\dfrac{l^2}{4l_+l_-}\dfrac{1}{\dfrac{1}{l}\sum_{i=1}^{l_+}\|\phi(\boldsymbol{x}_i)-\phi_+\|^2+\dfrac{1}{l}\sum_{i=1}^{l_-}\|\phi(\boldsymbol{x}_i)-\phi_-\|^2}$	$\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$

2.2 类别分布敏感性

类别分布敏感性是指普适性核度量标准的度量值随着两类样本分布变化,即使样本在某核函数构造的特征空间中线性可分,该核函数也可能获得较低的度量值.采用二维空间非平衡数据($X\subset\mathbb{R}^2$)模拟样本经核函数映射后在特征空间中的分布情况,以此展示各普适性度量标准对类别分布的敏感性,如图 1.核函数为 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$,类别分布由$\alpha\in[0,1]$确定,有比例为$\alpha$的样本在(–1,1)处具有类别标记"1",其余 1–α的样本在(1,1)处具有类别标记"–1".显然,对于任意α,样本集都线性可分,各普适性核度量标准的度量值应为其最优值.各普适性度量标准的度量值见表 2.

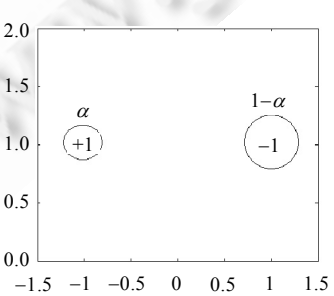

Fig.1 Imbalance data

图 1 非平衡数据

Table 2 Evaluation values on imbalance data

表 2 在非平衡数据上的度量值

核度量标准	度量值
KTA	$\sqrt{\alpha^2+(1-\alpha)^2}$
EKTA	$\frac{2\alpha(1-\alpha)}{\sqrt{\alpha^2+(1-\alpha)^2}}$
CKTA	$4\alpha(1-\alpha)$
FSM	0.0
KCSM	1.0

由表 2 可见:FSM 与 KCSM 达到其最优值;而 KTA,EKTA 和 CKTA 没有达到最优度量值,并且度量值随着类别分布比例α变化.因此,仅有 FSM 与 KCSM 不具有类别分布敏感性.通过表 1 比较 5 种度量标准可发现:FSM 与 KCSM 都具有因子 $\frac{l^2}{4l_+l_-},\frac{l^2}{4l_+l_-}$, 反映了类别的分布比例.乘上 $\frac{l^2}{4l_+l_-}$, 可消除普适性度量标准的类别分布敏感性.通过乘上 $\frac{l^2}{4l_+l_-}$, CKTA 可获得其最优度量值 1.0. $\frac{l^2}{4l_+l_-}$ 仅与数据相关.由于通常假设样本依据分布 D 独立同分布抽取,即,对于不同样本可假定类别分布具有较小波动,当样本给定, $\frac{l^2}{4l_+l_-}$ 为独立于核函数的常数,因此,是否具有 $\frac{l^2}{4l_+l_-}$ 项不影响核函数度量结果的相对关系.EKTA 具有 $\frac{l^2}{4l_+l_-}$ 但其仍然具有类别分布敏感性,其主要原因是:EKTA 与 KTA 具有线性平移敏感性,其度量值与样本的绝对位置相关.

2.3 线性平移敏感性

SVM 等核方法对旋转、平移、尺度变换等线性变换具有不变性,这就要求普适性核度量标准应同样具有该性质[15].半正定核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$具有旋转不变性,因此,普适性核度量标准也具有旋转不变性.5 种普适性核度量标准都具有规范化项($\hat{E}(\alpha_j)$ 项),所以都具有尺度变换不变性.因此,重点讨论线性平移敏感性.线性平移敏感性是指在特征空间中进行平移变换$\phi(\boldsymbol{x})\to\phi(\boldsymbol{x})+\delta$,度量值将是$\delta$的函数.因为平移变换不改变数据的可分性,所以度量值应为恒定值.采用二维空间数据($X\subset\mathbb{R}^2$)模拟样本经核函数映射后在特征空间中的分布情况,展示各普适性度量标准的线性平移敏感性,如图 2 所示.

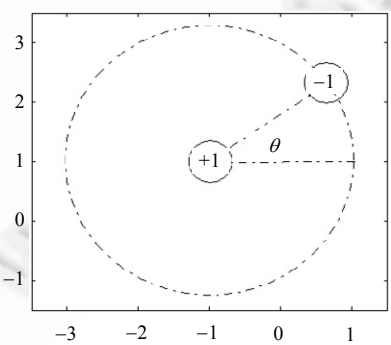

Fig.2 Linear translation data

图 2 线性平移数据

核函数为 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$,在(−1,1)处的样本具有类别标记“1”,在(−1+2cos(θ),1+2sin(θ))处的样本具有类别标记“−1”,两类样本具有相同容量.对于任意θ,该实例皆线性可分,因此,普适性核度量标准应为各自的最优值.各

度量标准的度量值列入表 3 中.通过表 3 可见:CKTA,FSM 与 KCSM 的度量值为各自的最优值;KTA 与 EKTA 具有相同的度量值,并且度量值随着θ而变化.由此可见,KTA 与 EKTA 具有线性平移敏感性,而 CKTA,FSM,KCSM 不具有.

Table 3 Evaluation values under linear translation

表 3 线性平移情况下的度量值

核度量标准	度量值
KTA	$\frac{2}{\sqrt{18-\cos(\theta)\sin(\theta)-16(\cos(\theta)-\sin(\theta))}}$
EKTA	$\frac{2}{\sqrt{18-\cos(\theta)\sin(\theta)-16(\cos(\theta)-\sin(\theta))}}$
CKTA	1.0
FSM	0.0
KCSM	1.0

由表 1,对比 KTA,EKTA,CKTA,FSM 与 KCSM 可发现:除 KTA 外,其他度量标准使用 $\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$ 替代 $\hat{A}_u(\boldsymbol{K},\boldsymbol{Y})$,由于 $\hat{A}_u(\boldsymbol{K}_C,\boldsymbol{Y})$ 使用中心化核函数,因此不具有线性平移敏感性.比较 CKTA 与 EKTA,CKTA 使用$\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F$替代了$\langle \boldsymbol{K},\boldsymbol{K}\rangle_F$,$\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F$不具有线性平移敏感性.FSM 与 KCSM 在计算类内散度时使用的是样本点间的核距离[24],核距离不具有线性平移敏感性,从而使得 FSM 与 KCSM 不具有线性平移敏感性.

2.4 异方差数据敏感性

FSM 被指出具有异方差数据敏感性[25],然而,本文研究发现,5 种度量标准都具有异方差数据敏感性.异方差数据敏感性是指在特征空间中两类样本分布的方差对度量值的影响,即使在线性可分的情况下,也可能获得较低的度量值.各度量标准中 $\hat{E}(\alpha_j)$ 描述了数据的散布程度,其形式与方差相似.KTA,EKTA,CKTA 使用核函数平方的均值($\frac{1}{l^2}\langle \boldsymbol{K},\boldsymbol{K}\rangle_F$)或中心化核函数平方的均值($\frac{1}{l^2}\langle \boldsymbol{K}_C,\boldsymbol{K}_C\rangle_F$)度量数据的散布程度.FSM 与 KCSM 使用数据点到类中心的距离度量数据的散布程度.5 种度量标准具有异方差数据敏感性,是由于 $\hat{E}(\alpha_j)$ 度量数据散布程度具有异方差数据敏感性.

本文采用二维空间数据($X\subset\mathbb{R}^2$)模拟样本经核函数映射后在特征空间中的分布情况,验证普适性核度量标准的异方差数据敏感性是由 $\hat{E}(\alpha_j)$ 产生的.使用两个二元正态分布生成数据,如图 3 所示:类别“1”的中心在(0.4,0),协方差矩阵为 $\begin{bmatrix}0.01 & 0\\ 0 & 1\end{bmatrix}$;类别“−1”的中心在(−0.4,0),协方差矩阵为 $\begin{bmatrix}0.01 & 0\\ 0 & var\end{bmatrix}$.其中,*var* 以 0.1 为步长从 0.1 递增到 10,每类 500 个样本.对于每个 *var*,重复进行 100 次上述过程,其结果取平均值.图 4 展示了 $\hat{E}(\alpha_j)$ 随 *var* 变化的曲线,图中还展示了两类中心距离(DBC)随 *var* 变化的曲线,以反映 $\hat{E}(\alpha_j)$ 的变化是由于异方差数据敏感性产生而不是由于随机数据波动产生的.可见,$\hat{E}(\alpha_j)$ 随 *var* 的变化而大幅度变化,其中,FSM 变化最为剧烈,并且在每个 *var* 的 100 组数据上的方差最大.FSM 更容易受到异方差数据的影响.

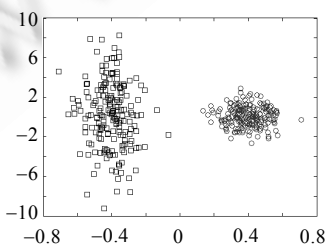

Fig.3 Heteroscedastic data

图 3 异方差数据

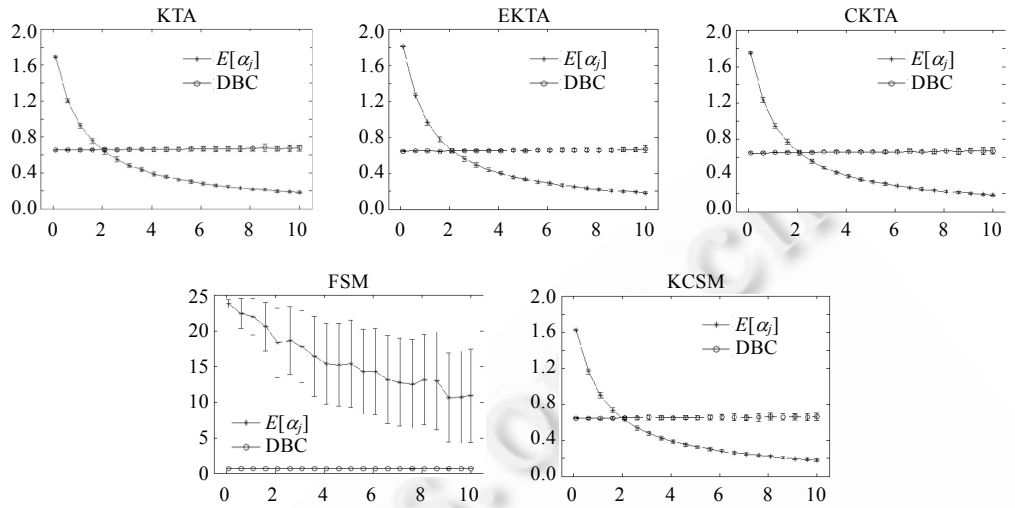

Fig.4 $\hat{E}(\alpha_j)$ on different variances

图 4 不同方差值下的 $\hat{E}(\alpha_j)$

3 实 验

本文使用来自 UCI 的 9 个数据集和 20Newsgroups 数据集进行核函数选择实验.UCI 数据集中的多分类问题使用“one-vs-one”策略转化为多个二分类问题,各数据集信息列于表 4 中,包括特征数量、样本容量.在 UCI 数据集上采用 10 折交叉验证的方法估计错误率.采用文献[26]所提供的 20Newsgroups 数据集(http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html),包含样本 18 846,特征 26 214.该数据集已经划分训练集和测试集,训练集包含样本 11 314(60%),测试集包含样本 7 532(40%).在训练集上采用核函数度量标准选择核函数,在测试集上验证所选择核函数的分类错误率.核方法使用 SVM,训练工具采用 LIBSVM.参数γ为 0.1,1,2,4,8,16,32 的 RBF 核与参数 d 为 1,2,3,4 的多项式核作为被度量的核函数.SVM 的惩罚因子 C 使用 10 折交叉验证,从 0.01,0.1,1,10,100 中选择.

Table 4 Description of UCI data sets

表 4 UCI 数据集描述

数据集	Breast	Diabetes	Ionosphere	Monks 1	Monks 2
特征数量	9	8	33	6	6
样例数量	699	768	351	556	601
类别数	2	2	2	2	2
数据集	Monks 3	Vehicle (bus, opel, saab, van)	Iris (setosa, versicolor, virginica)	Balance (B, L, R)	–
特征数量	6	18	4	4	–
样例数量	554	846	150	652	–
类别数	2	4	3	3	–

KTA,EKTA,CKTA,FSM 与 KCSM 在每个数据集上选出的核函数及 10 折交叉验证错误率(20Newgroups 为在测试集上的分类错误率)列入表 5 中,括号内为错误率方差.其中,BEST 为通过 10 折交叉验证获得的最小错误率(20Newgroups 为在测试集上的最小分类错误率),P 代表多项式核,R 代表 RBF 核,后续数值为核函数的参数,如:P3 代表参数 d 为 3 的多项式核.表中黑体数据表示通过置信度为 95%的 T 检验,与 BEST 无显著差异.标有下划线的数据表示通过置信度 95%的 T 检验,与 BEST 存在显著差异(差于 BEST),但显著优于其他普适性度量标准.Win 1 行代表在 19 组分类问题上与 BEST 无显著性差异的数量,Win 2 行代表在 19 组分类数问题上显著好于其他普适性度量标准的数量.

Table 5 Best kernels selected by each universal kernel evaluation measure and corresponding CV error

表 5 各普适性度量标准选择的最优核及相应的交叉验证错误率

	KTA	EKTA	CKTA	FSM	KCSM	BEST
Breast	**P3 0.0372(0.0263)**	**P1 0.0386(0.0261)**	**P2 0.0386(0.0261)**	**R1 0.0458(0.0307)**	**P1 0.0386(0.0261)**	0.0372 (0.0263)
Diabetes	**R1 0.2342(0.0495)**	**R1 0.2342(0.0495)**	**R1 0.2342(0.0495)**	R0.1 0.3489(0.0419)	**P1 0.2147(0.0447)**	0.2147 (0.0482)
Ionosphere	**R2 0.0513(0.0225)**	**R2 0.0513(0.0225)**	**R2 0.0513(0.0225)**	R0.1 0.3535(0.0932)	**R4 0.0569(0.0230)**	0.0513 (0.0225)
Monks 1	**P4 0.0484(0.0412)**	**P4 0.0484(0.0412)**	**P4 0.0484(0.0412)**	R0.1 0.3130(0.0627)	R32, P1 0.3130(0.0627)	0.0000 (0.0000)
Monks 2	R 32 0.3429(0.0655)	R1 0.0415(0.0352)	R1 0.0415(0.0352)	R0.1 0.1731(0.0505)	P2 0.2080(0.0530)	0.0133 (0.0171)
Monks 3	P1 0.1985(0.0380)	P1 0.1985(0.0380)	R4 0.0307(0.0255)	R0.1 0.3210(0.0831)	P1 0.1985(0.0380)	0.0126 (0.0192)
B_L	R4 0.0326(0.0325)	R0.1 0.1459(0.0615)	R1 0.0298(0.0243)	R0.1 0.1459(0.0615)	R8, R16, R32,P1 0.0680(0.0536)	0.0060 (0.0126)
B_R	R4 0.0414(0.0346)	R0.1 0.1455(0.0437)	R1 0.0296(0.0312)	R0.1 0.1455(0.0437)	R8,R16,R32,P1 0.0768(0.0463)	0.0000 (0.0000)
L_R	P1 0.0417(0.0204)	P1 0.0417(0.0204)	**R4 0.0000(0.0000)**	R0.1 0.4969(0.1653)	P1 0.0417(0.0204)	0.0000 (0.0000)
Setosa_Versicolor	**P1 0.0000(0.0000)**	**P1 0.0000(0.0000)**	**R2 0.0000(0.0000)**	**R1 0.0000(0.0000)**	**P1 0.0000(0.0000)**	0.0000 (0.0000)
Setosa_Virginica	**P1 0.0000(0.0000)**	**P1 0.0000(0.0000)**	**R2 0.0000(0.0000)**	**R2 0.0000(0.0000)**	**P1 0.0000(0.0000)**	0.0000 (0.0000)
Versicolor_Virginica	**P2 0.0600(0.0843)**	**P2 0.0600(0.0843)**	**R2 0.0600(0.0699)**	R0.1 0.3900(0.1197)	**P1 0.0300(0.0675)**	0.0300 (0.0675)
Bus_Opel	R1 0.0163(0.0221)	R1 0.0163(0.0221)	R1 0.0163(0.0221)	R0.1 0.5070(0.0578)	R2 0.0023(0.0074)	0.0000 (0.0000)
Bus_Saab	R1 0.0092(0.0119)	R1 0.0092(0.0119)	R1 0.0092(0.0119)	R0.1 0.4960(0.1804)	R2, R4 0.0069(0.0111)	0.0000 (0.0000)
Bus_Van	**R1 0.0144(0.0258)**	**R1 0.0144(0.0258)**	**R1 0.0144(0.0258)**	R0.1 0.4722(0.0551)	**R32,P1 0.0144(0.0203)**	0.0000 (0.0000)
Opel_Saab	R0.1 0.5528(0.0675)	R0.1 0.5528(0.0675)	R0.1 0.5528(0.0675)	R0.1 0.5528(0.0675)	R1 0.4103(0.0959)	0.2539 (0.0767)
Opel_Van	R1 0.0413(0.0346)	R1 0.0413(0.0346)	R2 0.0243(0.0199)	R0.1 0.4792(0.0465)	P1 0.0195(0.0252)	0.0000 (0.0000)
Saab_Van	R1 0.0289(0.0220)	R1 0.0289(0.0220)	R2 0.0240(0.0226)	R0.1 0.4732(0.0704)	R32 0.0530(0.0297)	0.0000 (0.0000)
20Newsgroups	**P1 0.2403**	**P1 0.2403**	**P1 0.2403**	R0.1 0.9469	**P1 0.2403**	0.2403
Win 1	9/19	9/19	10/19	3/19	8/19	–
Win 2	13/19	12/19	17/19	3/19	12/19	–

由表 5 可见,普适性核度量标准选择出的核函数达到或接近交叉验证方法选择的最优核函数的分类效果,表明了普适性核度量标准在核函数选择问题上的有效性.但是这 5 种普适性度量标准都没有完全与 BEST 一致或无显著性差异,最多仅在 10 个数据集上与 BEST 无显著差异,可见,普适性度量标准还需进一步研究.KTA,EKTA,CKTA 基于 Alignment 思想的度量标准好于 FSM 与 KCSM 的基于可分性度量的标准.在本文的实验中,对 KTA 进行改进的 EKTA 与 FSM 并没有显著好于 KTA 的效果.CKTA 为效果最好的普适性度量标准,在 10 个二分类问题上与 BEST 无显著性差异,在 17 个二分类问题上显著好于其他度量标准.通过比较 KTA,EKTA 与 CKTA,可验证 CKTA 所解决的线性平移敏感性问题在真实问题集上的有效性.FSM 为效果最差的普适性度量标准,并且较倾向于选择具有较小γ值的 RBF 核(R0.1),文献[27]也同样发现了该现象.FSM 更容易受到异方差数据的影响,是导致其效果较差的原因之一.虽然 FSM 同样解决了 KTA 的线性平移敏感性问题,但引入了较为严重的异方差数据敏感性问题,使得其效果较差.KCSM 效果好于 FSM,但由于其基于 Fisher 判别分析,假设数据符合正态分布,因此也限制了其效果[20].

4 结束语

本文对 KTA,EKTA,CKTA,FSM 和 KCSM 这 5 种普适性度量标准进行了比较研究,发现上述 5 种普适性度

量标准具有较为相近的形式;并且发现其度量内容为特征空间中某一线性假设的平均间隔,与支持向量机最大化最小间隔的优化标准存在偏差.使用模拟数据分析了上述标准的类别分布敏感性、线性平移敏感性、异方差数据敏感性,指出了各度量标准的数据敏感性问题的产生原因.最后,在 9 个 UCI 数据集和 20Newsgroups 数据集上比较了上述标准的度量效果,由于 CKTA 解决了 KTA 的线性变换敏感性问题,并且受异方差数据影响较弱,是度量效果相对最好的普适性核度量标准.

References:

[1] Schölkopf B, Smola A. Learning with Kernels. Cambridge: MIT Press, 2002. 1−45.

[2] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifier to solve real world classification problems? Journal of Machine Learning Research, 2014,15(10):3133−3181.

[3] Ramachandram D, Mandava R, Ehsan AM. A survey of the state of the art in learning the kernels. Knowledge and Information Systems, 2012,31(2):193−221. [doi: 10.1007/s10115-011-0404-6]

[4] Wang TH, Zhao DY, Tian SF. An overview of kernel alignment and its application. Artificial Intelligence Review, 2012,43(2): 179−192. [doi: 10.1007/s10462-012-9369-4]

[5] Elisseeff A, Pontil M. Leave-One-Out error and stability of learning algorithms with applications. In: Proc. of the Advances in Learning Theory: Methods, Models and Applications. 1994. 1−15.

[6] Chapelle O, Vapnik V. Model selection for support vector machines. In: Solla SA, Leen TK, Müller K, eds. Proc. of the Advances in Neural Information Processing Systems 12. Cambridge: MIT Press, 1999. 230−236.

[7] Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semi-definite programming. Journal of Machine Learning Research, 2004,5(1):27−72.

[8] Wang CQ, Chen JM, Hu CH, Sun YX. Kernel matrix learning with a general regularized risk functional criterion. Journal of Systems Engineering and Electronics, 2010,21(1):72−80. [doi: 10.3969/j.issn.1004-4132.2010.01.013]

[9] Girolami M, Rogers S. Hierarchic Bayesian models for kernel learning. In: Raedt LD, Wrobel S, eds. Proc. of the 22nd Int'l Conf. on Machine Learning. Bonn, 2005. 241−248. [doi: 10.1145/1102351.1102382]

[10] Ong CS, Smola AJ, Williamson RC. Learning the kernel with hyperkernels. Journal of Machine Learning Research, 2005,6(7): 1043−1071.

[11] Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. In: Dietterich TG, Becker S, Ghahramani Z, eds. Proc. of the Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2001. 367−373.

[12] Kandola J, Shawe-Taylor J, Cristianini N. On the extensions of kernel alignment. In: Proc. of the Neural Networks and Computational Learning Theory. 2002.

[13] Cortes C, Mohri M, Rostamizadeh A. Algorithms for learning kernels based on centered alignment. Journal of Machine Learning Research, 2012,13(2):795−828.

[14] Meilă M. Data centering in feature space. In: Bishop CM, Frey BJ, eds. Proc. of the 9th Int'l Workshop on Artificial Intelligence and Statistics. 2003.

[15] Nguyen CH, Ho TB. An efficient kernel matrix evaluation measure. Pattern Recognition, 2008,41(11):3366−3372. [doi: 10.1016/j.patcog.2008.04.005]

[16] Wang L, Chan KL. Learning kernel parameters by using class separability measure. In: Cristianini N, Jaakkola T, Jordan M, Lanckriet G, eds. Proc. of the 6th Kernel Machines Workshop, in Conjunction with Neural Information Processing Systems. 2002.

[17] Loog M, Heab-Umbach R. Multi-Class linear dimension reduction by generalized fisher criteria. In: Proc. of the 6th Int'l Conf. on Spoken Language Processing. 2000.

[18] Nazarpour A, Adibi P. Two-Stage multiple kernel learning for supervised dimensionality reduction. Pattern Recognition, 2015,48(5): 1854−1862. [doi: 10.1016/j.patcog.2014.12.001]

[19] Ge MM, Fan LY. Learning optimal kernel for pattern classification. WSEAS Trans. on Mathematics, 2013,5(12):491−500.

[20] Wu XY, Mao X, Chen LJ, Xue YL, Rovetta A. Kernel optimization using nonparametric fisher criterion in the subspace. Pattern Recognition Letters, 2015,54:43−49. [doi: 10.1016/j.patrec.2014.11.016]

[21] Afkanpour A, Szepesvári C, Bowling M. Alignment based kernel learning with a continuous set of base kernels. Machine Learning, 2013,91(3):305−324. [doi: 10.1007/s10994-013-5361-8]

[22] Lu YT, Wang LT, Lu JF, Yang JY, Shen CH. Multiple kernel clustering based on centered kernel alignment. Pattern Recognition, 2014,47(11):3656−3664. [doi: 10.1016/j.patcog.2014.05.005]

[23] Gao W, Zhou ZH. On the doubt about margin explanation of boosting. Artificial Intelligence, 2013,203:1−18. [doi: 10.1016/j.artint.2013.07.002]

[24] Schölkopf B. The kernel trick for distance. In: Leen TK, Dietterich TG, Tresp V, eds. Proc. of the Advances in Neural Information Processing Systems 13. Cambridge: MIT Press, 2000. 301−307.

[25] Chudzian P. Evaluation measures for kernel optimization. Pattern Recognition Letters, 2012,33(9):1108−1116. [doi: 10.1016/j.patrec.2012.01.006]

[26] Cai D, Wang XH, He XF. Probabilistic dyadic data analysis with local and global consistency. In: Danyluk AP, Bottou L, Littman ML eds. Proc. of the 26th Annual Int'l Conf. on Machine Learning. ACM Press, 2010. 105−112. [doi: 10.1145/1553374.1553388]

[27] Wang PY, Cai DF. Kernel-Distance target alignment. In: Li ST, Liu CL, Wang YN, eds. Proc. of the Pattern Recognition: Communications in Computer and Information Science. Berlin: Springer-Verlag, 2014. 101−110. [doi: 10.1007/978-3-662-45646-0_11]

王裴岩(1983－),男,辽宁沈阳人,博士生,讲师,CCF 会员,主要研究领域为机器学习,信息抽取.

蔡东风(1958－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为机器学习,人工智能,自然语言处理.