

应用统计方法综合评估核函数分类能力的研究

王 泳^{1),2)} 胡包钢^{1),2)}

¹⁾(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

²⁾(中国科学院研究生院 北京 100049)

摘 要 应用统计方法对支持向量机方法中核函数选择问题进行了研究. 文中将“纠正重复取样 t 测试”引入到核函数选择中, 通过其与 k -折交叉验证、配对 t 测试等多种统计方法的综合应用, 对 9 个常用核函数的分类能力进行了定量研究. 同时, 文中还提出了基于信息增益的评估核函数模式识别能力的定量评估准则, 证明了该准则是传统评估准则的非线性函数. 数值实验表明, 不同模型评估准则之间存在差异, 但应用统计方法可以从这些差异中发现一些规律. 同时, 不同统计方法之间也存在显著差异, 且这种差异对模型评估的影响要大于由于评估准则的不同而产生的影响. 因此, 只有应用综合的评估方法和准则才能对不同核函数的分类能力进行客观评估.

关键词 核函数选择; 模式识别; 纠正重复取样 t 测试; 信息增益; 非线性函数

中图法分类号 TP391

A Study on Integrated Evaluating Kernel Classification Performance Using Statistical Methods

WANG Yong^{1),2)} HU Bao-Gang^{1),2)}

¹⁾(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

²⁾(Graduate University of Chinese Academy of Sciences, Beijing 100049)

Abstract This paper explores the research on evaluating kernel classification performance using statistical methods. By employing the corrected resample t -test and other two statistical methods— k -fold cross-validation and paired t -test, this paper compares their classification abilities on nine normally used kernels. In addition, a new quantitative criterion of evaluating kernel classification performance based on information gain is proposed, which is proved to be the nonlinear function of traditional criteria. Benchmark tests show that there is difference among different criteria, but by using statistical methods some regulations can be turned up among them. Simultaneously, there is great difference among different statistical methods, which affects the evaluating results more than the difference among different criteria does. So only with the integrated methods and criteria the classification performance of different kernels can be evaluated objectively.

Keywords kernel selection; pattern recognition; corrected resample t -test; information gain; nonlinear function

1 引 言

在支持向量机(Support Vector Machines, SVMs)^[1]

方法中,核函数选择十分重要. 研究表明^[1], 针对同一分类问题, 选择不同的核函数, 分类性能可能会相差很大. 这主要是因为构成核函数 $K(x, y)$ 的非线性映射 $\varphi(x)$ 是隐函数, 且这些隐函数的类型是多样

可变的. 所以当人们对特定问题没有任何先验知识的时候, 很难确定应该选择哪一类核函数进行计算. 虽然利用泰勒级数展开和傅立叶级数展开的方法, 已经证明了存在一类最优核, 它所对应的特征映射可以确保任意两个不连接的有界闭集在特征空间中线性可分^[2], 但如何构造这类最优核至今却还缺乏行之有效的办法. 众多学者从不同的角度对核函数选择^[3-4]和构造^[5-8]问题进行了有益的探讨, 但综合性的评估研究仍是缺乏的.

一般说来, 核函数的评估指标可以分为两大类: 一类来自实际数据的实验验证结果; 一类来自理论分析所给出的界. 根据统计学习理论, 核函数推广能力的强弱与由该函数计算得到的分类超平面集合的 VC 维 (Vapnik-Chervonenkis dimension) 相关, VC 维 h , 泛化误差 ϵ 和特征空间中训练样本集与超平面的最短距离 γ 之间存在以下关系^[1]:

$$h \leq \min\left(\left\lceil \frac{R^2}{\gamma^2} \right\rceil, n\right) + 1, \quad \epsilon \leq O\left(\frac{R^2}{m\gamma^2}\right) \tag{1}$$

R 是特征空间中包含所有训练样本的最小超球的半

径, m 是训练样本的个数, n 是特征空间的维数. 因此, VC 维越小, 函数的推广能力越强. 但遗憾的是, 目前尚没有关于如何计算任意函数集的复杂性 (VC 维) 以及推广性界的一般性理论, 能够得到的只是一些估计值^[1,7]. 所以在解决实际问题时, 通常还是以实际数据的实验验证结果作为核函数评估的数量指标.

根据有限数据的实验验证结果进行分类预测性能评估是机器学习领域的一个存在较多争议的研究领域, 这不仅是因为在分类模型预测性能评估体系中存在很多模型评估准则, 而且还存在许多不同的模型评估方法 (图 1). 在实践中, 应用 k -折交叉验证方法 (k -fold cross-validation) 和准确率准则对分类模型进行预测性能评估是最为常规的方法, 但需要注意的是交叉验证技术是一个启发式技术, 未必对各种情况都适用^[9], 尤其是当确定一个学习模型对某个具体问题的解决是否真的优于另一个学习模型, 就需要证明模型之间的这种性能差别不只是评估过程中所产生的偶然结果, 这通常是一项给出置信边界的统计实验工作.

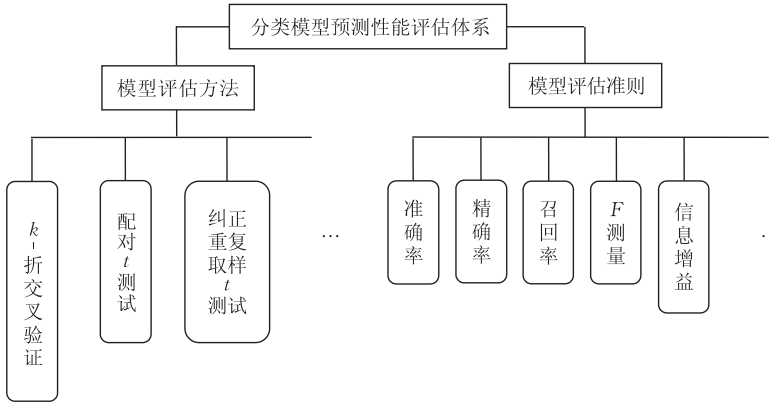


图 1 分类模型预测性能评估体系

本文第 2 节对模型预测性能评估的三种统计方法—— k -折交叉验证、配对 t 测试 (paired t -test)^[10]、纠正重复取样 t 测试 (corrected resample t -test)^[11]进行了对比分析, 引入并讨论了纠正重复取样 t 测试对模型预测性能评估的适用性; 第 3 节提出了基于信息增益^[12]的评估核函数模式识别能力的定量评估准则, 并证明了该准则在一定程度上可以弥补其它评估准则的不足; 第 4 节是实验与分析; 第 5 节对文章内容进行了总结并对进一步研究的方向进行了展望.

2 模型预测性能评估方法

2.1 k -折交叉验证

其基本思想是把样本集 $D = \{(x_i, y_i)\}_{i=1}^m$ (其中

$x_i \in R^n, y_i \in R$) 随机划分为 k 个不相交的子集 $\{D_1, D_2, \dots, D_k\}$, 且每个子集都有 m/k 个样本点. 分类器 T 要迭代训练 k 次, 每次都使用集合 $D \setminus D_i (i \in \{1, 2, \dots, k\})$ 中的数据进行训练, 而用集合 D_i 中的数据进行验证. k -折交叉验证估计出的分类器 T 的泛化误差率 $Err_{CV}(T, D)$ 是 k 次验证误差率 $Err_i(T, D_i)$ 的平均值. 令 $D_{(i)}$ 代表包含样本 $v_i = \langle x_i, y_i \rangle$ 的子集, $T(D \setminus D_{(i)}, v_i)$ 代表分类器 T 对样本 v_i 进行分类的结果, 则 k -折交叉验证估计出的分类器 T 的泛化误差率是

$$Err_{CV}(T, D) = \frac{1}{k} \sum_{i=1}^k Err_i(T, D_i) = \frac{1}{m} \sum_{v_i \in D} \delta(T(D \setminus D_{(i)}, v_i), y_i) \tag{2}$$

$$\delta(i,j)=\begin{cases}1, & i\neq j \\ 0, & i=j\end{cases}\quad (3)$$

定理 1. 给定样本集 D 和分类器 T , 分类器 T 真实但未知的分类误差率是 p , 如果在 k -折交叉验证中删除 D 中任意的样本并不影响估计出的分类器 T 的泛化误差率, 则 k -折交叉验证评估出的泛化误差率是真实误差率的无偏估计.

证明. 因为删除 D 中任意的样本并不影响 $Err_{CV}(T,D)$ 的取值, 所以 D 中样本是从样本空间 \mathcal{D} 中随机选取且与 T 相互独立, 所以 $\delta(T(D\backslash D_{(i)}, v_i), y_i)$ 是满足二项分布的随机变量, 因此

$$E[Err_{CV}(T,D)]-p=\\ \frac{1}{m}E\left[\sum_{v_i\in D}\delta(T(D\backslash D_{(i)}, v_i), y_i)\right]-p=\frac{1}{m}mp-p=0$$

证毕.

但通常情况下, D 中样本有限, 因此很难保证分类器 T 与 D 相互独立, 所以直接将交叉验证估计出的泛化误差率当作模型的真实误差率对模型预测性能进行评估是有误差的, 但可以用置信区间(confidence interval)估计的方法对这种误差进行估计. 根据中心极限定理, 当 $mp\geq 5$ 且 $m(1-p)\geq 5$ 时, 随机变量 $Err_{CV}(T,D)$ 可以用 $\mu=p, \sigma^2=p(1-p)/m$ 的正态分布近似. 所以, 为了得到 μ 的 $100(1-\alpha)\%$ 置信区间, m 的最小取值应满足下式:

$$\frac{4(z_{\alpha/2})^2\sigma^2}{W^2}\leq m\Leftrightarrow \frac{4(z_{\alpha/2})^2p(1-p)}{W^2m}\\ \leq m\Leftrightarrow \frac{2z_{\alpha/2}\sqrt{p(1-p)}}{W}\leq m\quad (4)$$

其中, $z_{\alpha/2}$ 是标准正态分布右尾被分割出 $\alpha/2$ 面积的分割点处的 z 值, W 是置信区间的宽度.

2.2 配对 t 测试

虽然, 增加样本数 m 可以增加 k -折交叉验证的置信度, 但这种数量的增加是有限度的. 研究表明^[13], 单纯增加 m 会导致交叉验证的渐进有偏. 所以, 要想保证 k -折交叉验证方法的有效性, 关键是增加样本集 D 和分类器 T 之间的独立性. 另外, 应用 k -折交叉验证对学习模型进行评估仅考察了模型之间的均值差异, 模型评估的另一个重要考察指标是模型之间的方差差异^[10-11], 这可以用统计学中的配对 t 测试方法实现^[10].

设第一组样本 x_1, x_2, \dots, x_k 是学习模型 X 根据某种性能评估准则在不同的数据集上得到的估计值(所有数据集大小相同, 且来源于同一个领域), 第二组样本 y_1, y_2, \dots, y_k 是学习模型 Y 根据同样的性能评估准则在同样的数据集上得到的估计值, 即 x_1 和

y_1 是使用相同的数据集产生的, x_2 和 y_2 也是如此, 依此类推. 第一组样本的平均值用 μ_1 来表示, 第二组样本的平均值用 μ_2 来表示, 因此学习模型 X 和 Y 的比较就是要判定 μ_1 和 μ_2 是否有显著的差别, 由于实验中两种学习模型在每个数据集上的实验都能获得配对的结果, 因此这种统计测试被称为配对的 t 测试. 表 1 列出了在小样本情况下配对 t 检验方法.

当 $D_0=0$ 时就是对“两个均值相等”这一零假设的检验方法, 即比较模型 X 和模型 Y 学习性能是否一样的检验方法.

表 1 小样本情况下配对 t 检验方法

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\mu_1-\mu_2=D_0$ $H_1:\mu_1-\mu_2\neq D_0$	$H_0:\mu_1-\mu_2\geq D_0$ $H_1:\mu_1-\mu_2<D_0$	$H_0:\mu_1-\mu_2\leq D_0$ $H_1:\mu_1-\mu_2>D_0$
检验统计量	$t=\frac{\bar{d}-D_0}{\sigma_d/\sqrt{k}}\approx\frac{\bar{d}-D_0}{s_d/\sqrt{k}}$, 自由度: $k-1$		
α 与拒绝域	$ t >t_{\alpha/2}(k-1)$	$t<-t_{\alpha}(k-1)$	$t>t_{\alpha}(k-1)$
P 值决策准则	$P<\alpha$, 拒绝 H_0		
假定条件	1. 差值总体的相对频数分布接近正态分布 2. 配对差由差值总体随机选出		

表 1 中:

d_i 表示第 i 个配对样本数据的差值, 即 $d_i=x_i-y_i, i=1, 2, \dots, k$;

\bar{d} 表示配对样本数据差值的平均值, 即 $\bar{d}=\frac{1}{k}\sum_{i=1}^kd_i=\mu_1-\mu_2$;

s_d 表示配对样本数据差值的准则差, 即 $s_d=\sqrt{\frac{1}{k-1}\sum_{i=1}^k(d_i-\bar{d})^2}$;

σ_d 表示配对样本数据差值的总体准则差, 即 $\sigma_d=\sqrt{\frac{\sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2\rho}{k}}$ (其中 σ_1 表示第一组样本数据的总体准则差, σ_2 表示第二组样本数据的总体准则差, ρ 表示两组样本的相关程度);

α 表示置信度(显著性水平);

P 表示观察到的显著性水平.

2.3 改进的配对 t 测试

标准配对 t 检验方法的假定条件 1 要求差值总体的相对频数分布接近正态分布, 而配对数据越多, 其差值总体的相对频数分布越接近正态分布, 因此数据来源越多, 检验所获得的结果越可靠. 但在实践中, 通常只有一个容量有限的数据集可用, 虽然通过增加交叉验证的次数可以增加配对样本的数目, 但重复利用原始数据集得出的交叉验证估计不是独立的, 因此使得配对数据之间具有很强的相关性, 造成

配对 t 检验方法的假定条件 2 无法满足。实际上,通过增加交叉验证次数来增加样本数目,最终将导致产生明显差异,因为 t 统计量在毫无限制的增加着,而这种差异的产生是由于样本的重复使用造成的,并没有真实反映出样本总体的性质。

近年来提出的纠正重复取样 t 测试方法^[11]可以很好地解决这个问题。该方法使用重复旁置法来代替交叉验证法,此时 k -折交叉验证只是一个特例。它将数据集进行不同的随机分割 k 次,每次用 n_1 个样本训练,用 n_2 个样本测试,差值 d_i 则根据在测试数据上的性能计算得出。纠正重复取样 t 测试使用经修改后的统计量:

$$t = \frac{\bar{d} - D_0}{\sigma_d \sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)}} \approx \frac{\bar{d} - D_0}{s_d \sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)}} \tag{5}$$

可以看出,此时 t 统计量不再容易随着 k 值的增加而快速增长了。对于重复的 10 次 10 折交叉验证, $k=100$, $n_2/n_1=1/9$, σ_d 则基于 100 个差值计算得到。

3 模型预测性能评估准则

3.1 信息增益准则

定义 1. 假设数据集 $D = \{(x_i, y_i)\}_{i=1}^w$ (其中 $x_i \in R^n, y_i \in R$) 中包含的 w 个样本属于不同的 K 类,由每类样本构成的集合 $D_i = \{(x_{ij}, y_i)\}_{j=1}^{w_i}$ ($i=1, 2, \dots, K$) 中包含有 w_i 个样本,则数据集 D 的信息量(熵) $entropy(D)$ 为

$$\begin{aligned} entropy(D) &= entropy(D_1, D_2, \dots, D_K) \\ &= entropy\left(\frac{w_1}{w}, \frac{w_2}{w}, \dots, \frac{w_K}{w}\right) \\ &= -\sum_{i=1}^K \left(\frac{w_i}{w} \log_2 \left(\frac{w_i}{w}\right)\right) \end{aligned} \tag{6}$$

$entropy(D)$ 是对数据集 D 的不确定性的度量。当数据集中样本都属于同一类别 k 时, $w_k=w$, 属于其它类别的样本数都为 0, 此时数据集 D 完全确定, $entropy(D) = entropy(D_k) = 0$ (因为 $\lim_{p \rightarrow 0} p \log_2 p = 0$, 所以定义 $0 \log_2 0 = 0$)。当属于不同类别的样本数都相等时, 不确定性最大, 所以对含有 K 类样本的数据集 D 来说:

$$0 \leq entropy(D) \leq \log_2 K \tag{7}$$

定义 2. 假设通过分类模型 f 对数据集 D 进行分类, 由分类结果可以构成新的数据集 $\tilde{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^w$ (其中 \tilde{y}_i 与 y_i 不一定相等)。根据分类结果中每类样本的分布情况, 将数据集 \tilde{D} 划分为 K 个子集

合, 其中第 i 个子集合 $\tilde{D}_i = \{\tilde{D}_{ij}\}_{j=1}^K$ ($i=1, 2, \dots, K$) 由被分为第 i 类的样本组成, \tilde{D}_{ij} 表示由原本是第 j 类但却被分为第 i 类的样本组成的集合, 集合中的样本数是 w_{ij} 个。此时数据集 \tilde{D} 的信息量(熵) $entropy(\tilde{D})$ 为

$$\begin{aligned} entropy(\tilde{D}) &= entropy(\tilde{D}_1, \tilde{D}_2, \dots, \tilde{D}_K) \\ &= entropy(\{\tilde{D}_{1j}\}_{j=1}^K, \{\tilde{D}_{2j}\}_{j=1}^K, \dots, \{\tilde{D}_{Kj}\}_{j=1}^K) \\ &= \sum_{i=1}^K \left[\frac{\sum_{j=1}^K w_{ij}}{w} entropy(\{\tilde{D}_{ij}\}_{j=1}^K) \right] \\ &= -\sum_{i=1}^K \left[\frac{\sum_{j=1}^K w_{ij}}{w} \sum_{j=1}^K \left[\frac{w_{ij}}{\sum_{j=1}^K w_{ij}} \log_2 \left(\frac{w_{ij}}{\sum_{j=1}^K w_{ij}} \right) \right] \right] \\ &= -\sum_{i=1}^K \sum_{j=1}^K \left[\frac{w_{ij}}{w} \log_2 \left(\frac{w_{ij}}{\sum_{j=1}^K w_{ij}} \right) \right] \end{aligned} \tag{8}$$

定义 3. 信息增益(Information Gain) $IG(f)$ 度量的是分类模型 f 从数据集 D 中挖掘出的知识多少, 其定义为

$$IG(f) = entropy(D) - entropy(\tilde{D}) \tag{9}$$

任意分类数据集都可以看成是具有一定不确定性的系统, 一个好的分类器 f 应该表现出最大程度上减少了这个系统的不确定性, 而这种不确定性的减少, 从信息学的角度看就是 f 具有最大的信息增益, 由此可以得出定义 4。

定义 4. 对于分类模型 f_1 和 f_2 , 如果 $IG(f_1) > IG(f_2)$, 则 f_1 比 f_2 更能减少分类数据集的不确定性。

3.2 信息增益准则与其它准则的对比

针对模式识别问题, 实际中常用的模型评估准则有准确率(Accuracy)、精确率(Precision)、召回率(Recall)、 F 测量(F -measure)等^[14]。

对一个 yes 和 no 的二类分类问题, 一个预测可能产生 4 种不同的结果(表 2), 正确的肯定 TP (True Positive)、正确的否定 TN (True Negative)、错误的肯定 FP (False Positive) 和错误的否定 FN (False Negative)。

表 2 二类分类预测的不同结果

预测类			
Yes	No		
正确的肯定 TP	错误的否定 FN	Yes	真实类
错误的肯定 FP	正确的否定 TN	No	

准确率、精确率、召回率和 F 测量分别是根据它们的数值计算获得。

准确率:

$$A=\frac{TP+TN}{TP+TN+FP+FN}$$

(10)

精确率:

$$P=\frac{TP}{TP+FP}$$

(11)

召回率:

$$R=\frac{TP}{TP+FN}$$

(12)

F 测量:

$$F=\frac{2PR}{P+R}=\frac{2\cdot TP}{2\cdot TP+FP+FN}$$

(13)

例 1. 应用不同分类模型解决二类分类问题, 样本总量是 100 个, 其中肯定类是 50 个, 否定类是 50 个. 根据不同模型的分类结果计算信息增益, 并分析信息增益与准确率、精确率、召回率和 F 测量的关系. 结果见表 3.

例 2. 应用不同分类模型解决二类分类问题, 样本总量是 100 个, 其中肯定类是 20 个, 否定类是 80 个. 根据不同模型的分类结果计算信息增益, 并分析信息增益与准确率、精确率、召回率和 F 测量的关系. 结果见表 4.

表 3 根据不同模型的分类结果计算各种评估准则(基于例 1 正负样本比例相等的数据)

模型	TP	FP	TN	FN	准确率	精确率	召回率	F	信息增益
I	25	5	45	25	0.7	0.8333	0.5	0.6250	0.1468
II	30	10	40	20	0.7	0.75	0.6	0.6667	0.1245
III	15	5	45	35	0.6	0.75	0.3	0.4286	0.0468
IV	15	45	5	35	0.2	0.25	0.3	0.2727	0.2958
V	12	26	24	38	0.36	0.3158	0.24	0.2727	0.0611
VI	26	12	38	24	0.64	0.6842	0.52	0.5909	0.0611

表 4 根据不同模型的分类结果计算各种评估准则(基于例 2 正负样本比例不相等的的数据)

模型	TP	FP	TN	FN	准确率	精确率	召回率	F	信息增益
I	5	35	45	15	0.5	0.125	0.25	0.1667	0.0177
II	8	38	42	12	0.5	0.1739	0.4	0.2424	0.0026
III	16	76	4	4	0.2	0.1739	0.8	0.2857	0.0287
IV	16	80	0	4	0.16	0.1667	0.8	0.2759	0.0979
V	8	30	50	12	0.58	0.2105	0.4	0.2759	0.0003
VI	1	5	75	19	0.76	0.1667	0.05	0.0769	0.0003

例 1 和例 2 说明, 不同模型在某个评估准则下的评估结果可能相同, 但同时, 总可以运用其它评估准则来分辨它们的优劣. 同时, 从图 2 和图 3 中还可

以看出信息增益与准确率、精确率、召回率和 F 测量之间存在着复杂的非线性函数簇的关系.

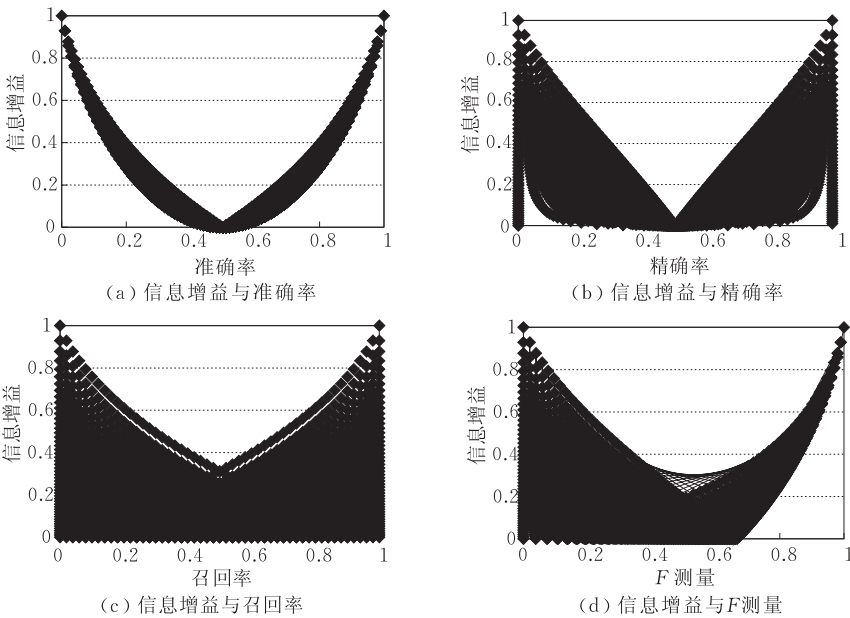


图 2 信息增益准则与其它准则的关系(基于例 1 正负样本比例相等的数据)

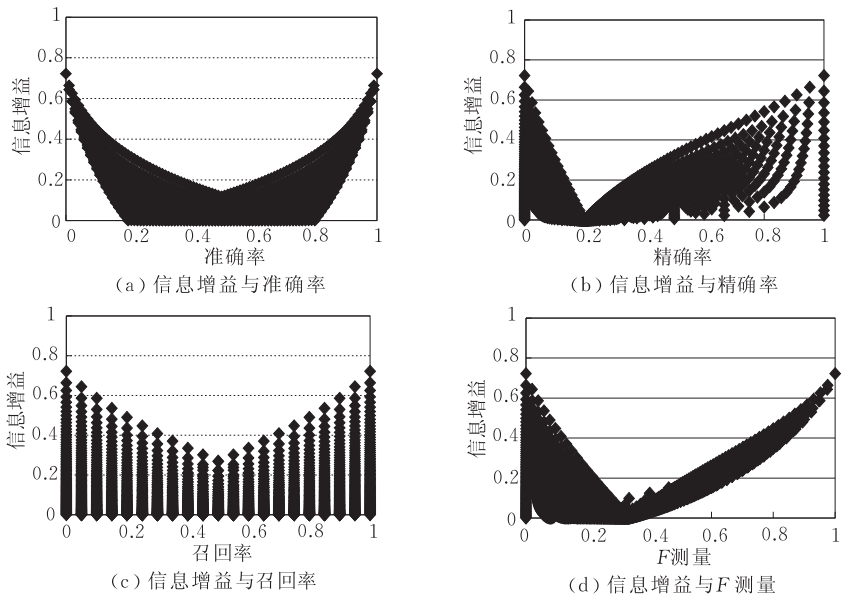


图 3 信息增益准则与其它准则的关系(基于例 2 正负样本比例不相等的的数据)

定理 2. 针对两类分类问题,信息增益是准确率、精确率和召回率的非线性函数.

证明. 根据式(10)~(12)可知

$$FN = \frac{1-R}{R}TP \tag{14}$$

$$FP = \frac{1-P}{P}TP \tag{15}$$

$$TN = \frac{AP+AR-PR-APR}{PR(1-A)}TP \tag{16}$$

根据式(8)和式(9)可知两类分类问题中的信息增益可表示为

$$IG =$$

$$\begin{aligned} & -\frac{TP+FN}{TP+TN+FP+FN}\log_2\left(\frac{TP+FN}{TP+TN+FP+FN}\right)- \\ & \frac{FP+TN}{TP+TN+FP+FN}\log_2\left(\frac{FP+TN}{TP+TN+FP+FN}\right)+ \\ & \frac{TP}{TP+TN+FP+FN}\log_2\left(\frac{TP}{TP+FP}\right)+ \\ & \frac{FP}{TP+TN+FP+FN}\log_2\left(\frac{FP}{TP+FP}\right)+ \\ & \frac{TN}{TP+TN+FP+FN}\log_2\left(\frac{TN}{TN+FN}\right)+ \\ & \frac{FN}{TP+TN+FP+FN}\log_2\left(\frac{FN}{TN+FN}\right) \end{aligned} \tag{17}$$

将式(14)~(16)代入式(17)得

$$\begin{aligned} IG = & \log_2(P+R-2PR) + \\ & \frac{1}{P+R-2PR} [P(1-A)\log_2(1-R)^{(1-R)} + \\ & R(1-A)\log_2(1-P)^{(1-P)} - PR\log_2(1-A)^{(1-A)} + \\ & \log_2(AP+AR-PR-APR)^{(AP+AR-PR-APR)} - \end{aligned}$$

$$\begin{aligned} & \log_2(AP+R-2PR)^{(AP+R-2PR)} - \\ & \log_2(AR+P-2PR)^{(AR+P-2PR)}] \end{aligned} \tag{18}$$

所以,针对两类分类问题,信息增益是准确率、精确率和召回率的非线性函数^①. 证毕.

性质 1. 一般情况下,只有当准确率、精确率和召回率三个变量值全部给定后,信息增益值才可唯一确定.

性质 2. 当准确率为 1 或 0 时,信息增益的值被唯一确定,且是最大值.

不同于应用分类精度为单一准则,信息增益给出了更为综合的模型预测性能评估指标,它相当于以非线性方式平衡了各种准则,因此它可以反映出模型更多的信息,例如:

(1) 信息增益准则可以反映出其它评估准则的评估能力. 精确率和召回率可以反映出模型较多的信息,因此精确率和召回率评估模型的能力较强,而准确率反映出的模型的信息最少,所以将准确率当作模型评估的单一准则是不恰当的;

(2) 信息增益准则的适用范围更广. F 测量准则适用于样本比例严重不平衡的数据集(即针对小概率事件的数据集),而信息增益准则能够反映出它的这种特性,因此也同样适用于这种情况.

所以信息增益在一定程度上可以弥补其它单一评估准则的不足.

^① 需要特别说明的是,定理 2 的证明是在 TP 、 TN 、 FP 和 FN 都不为零的一般情况下进行的. 它们中有为零的情况是特殊情况,可以证明结论依然成立. 文献[15-16]对信息增益与传统模型评估准则之间的关系及其在分类问题中的应用做了进一步的论述.

4 实验与分析

为了阐明不同统计方法和不同模型评估准则之

间的差异,本文在 Weka^[14] 软件平台上,对 9 个常用核函数(表 5,取默认参数)在 21 个标准数据集(表 6)上进行了 10 次 10 折交叉验证,并对实验结果进行了统计分析^①.

表 5 实验中使用的核函数^[8]

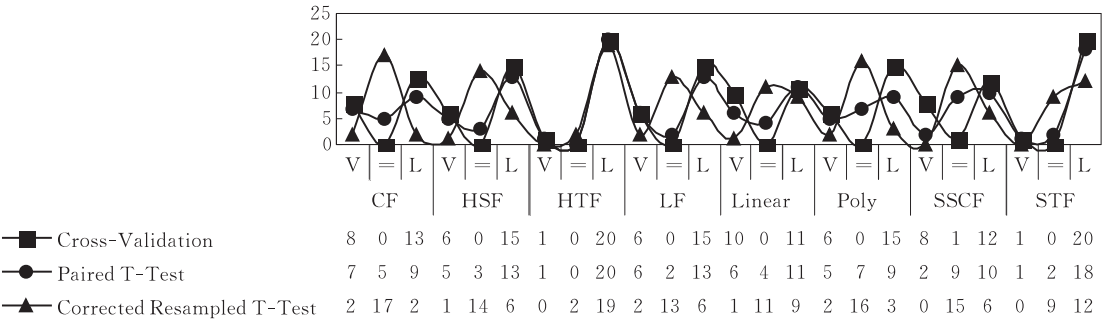
英文名称	缩写	数学表达式	参数定义	参考函数
Linear Function	Linear	$K(x,y)=\langle x,y\rangle$		$\text{Linear}(x)=x$
Polynomial Function	PF	$K(x,y)=(\langle x,y\rangle+1)^d$	$d>0$, default 3	$\text{Poly}(x)=(x+1)^d$
Radial Basis Function or Gaussian Function	RBF	$K(x,y)=e^{-g\langle x-y,x-y\rangle^2}$	$g>0$, default 1	$\text{RBF}(x)=e^{-gx^2}$
Symmetric Triangle Function	STF	$K(x,y)=\text{Max}(1-g \langle x-y,x-y\rangle ,0)$	$g>0$, default 3	$\text{STF}(x)=\text{max}(1-g x ,0)$
Cauchy Function	CF	$K(x,y)=\frac{1}{1+g\langle x-y,x-y\rangle^2}$	$g>0$, default 3	$\text{CF}(x)=\frac{1}{1+gx^2}$
Laplace Function	LF	$K(x,y)=e^{-g \langle x-y,x-y\rangle }$	$g>0$, default 3	$\text{LF}(x)=e^{-g x }$
Hyperbolic Secant Function	HSF	$K(x,y)=\frac{2}{e^{g\langle x-y,x-y\rangle}+e^{-g\langle x-y,x-y\rangle}}$	$g>0$, default 3	$\text{HSF}(x)=\frac{2}{e^{gx}+e^{-gx}}$
Squared Sin Cardinal Function or Squared Sinc Function	SSCF	$K(x,y)=\frac{\sin^2(\langle x-y,x-y\rangle)}{(g\langle x-y,x-y\rangle)^2}$	$g>0$, default 3	$\text{SSCF}(x)=\frac{\sin^2(x)}{(gx)^2}$
Hyperbolic Tangent Function or Sigmoid Function	HTF	$K(x,y)=\frac{e^{g\langle x-y,x-y\rangle}-e^{-g\langle x-y,x-y\rangle}}{e^{gx}+e^{-gx}}$	$g>0$, default 1	$\text{HTF}(x)=\frac{e^{gx}-e^{-gx}}{e^{gx}+e^{-gx}}$

表 6 参与实验的数据集

数据集	特征数	数据量	数据来源	数据集	特征数	数据量	数据来源
Breast-Cancer-Wisconsin	9	699	UCI 数据库 ^[17]	CPS_85_Wages	10	534	Statlib 数据库 ^[18]
DUPA-Liver-Disorders	6	345	UCI 数据库 ^[17]	Plasma_Retinol	13	315	Statlib 数据库 ^[18]
Diabetes_Pima	8	768	UCI 数据库 ^[17]	Prnn-Crabs	7	200	Statlib 数据库 ^[18]
Heart-Statlog	13	270	UCI 数据库 ^[17]	Prnn_Synth_TE	2	1000	Statlib 数据库 ^[18]
Hepatitis	19	155	UCI 数据库 ^[17]	Prnn_Synth_TR	2	250	Statlib 数据库 ^[18]
Ionosphere	34	351	UCI 数据库 ^[17]	Schizo	14	340	Statlib 数据库 ^[18]
Monks-Problems-1	6	432	UCI 数据库 ^[17]	Veteran	7	137	Statlib 数据库 ^[18]
Monks-Problems-2	6	432	UCI 数据库 ^[17]	Nonlineardata100	2	100	基准测试数据 ^[19]
Monks-Problems-3	6	432	UCI 数据库 ^[17]	Nonlineardata1000	2	1000	基准测试数据 ^[19]
Musk-Clean-1	166	476	UCI 数据库 ^[17]	Two-Spirals	2	194	基准测试数据 ^[20]
Sonar	60	208	UCI 数据库 ^[17]				

实验 1. 对模型预测性能评估的 3 种统计方法(交叉验证、配对 t 测试和纠正重复取样 t 测试)进

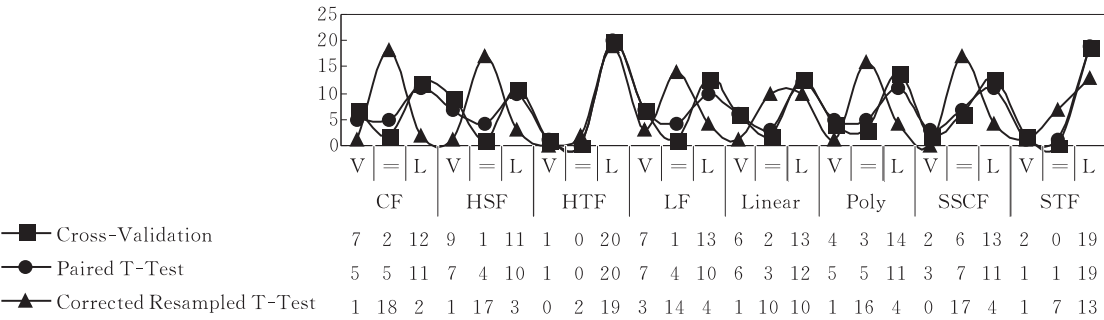
行对比分析. 图 4~图 8 分别显示了在准确率、精确率、召回率、 F 测量和信息增益准则下 RBF 核函数与



(图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF核函数,数据代表相应的次数,统计显著性水平为5%)

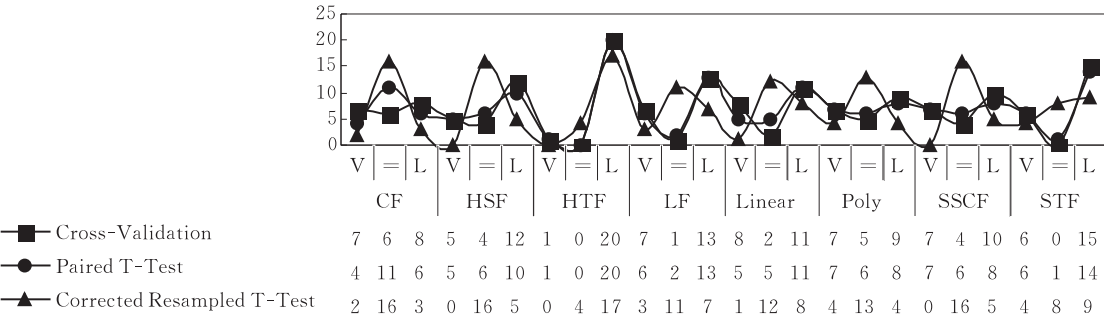
图 4 在准确率准则下对比 3 种统计方法

① 需要特别说明的是,实验中得出的关于核函数分类能力的结论只针对参与实验的数据,并非一般意义下关于核函数分类能力的结论.



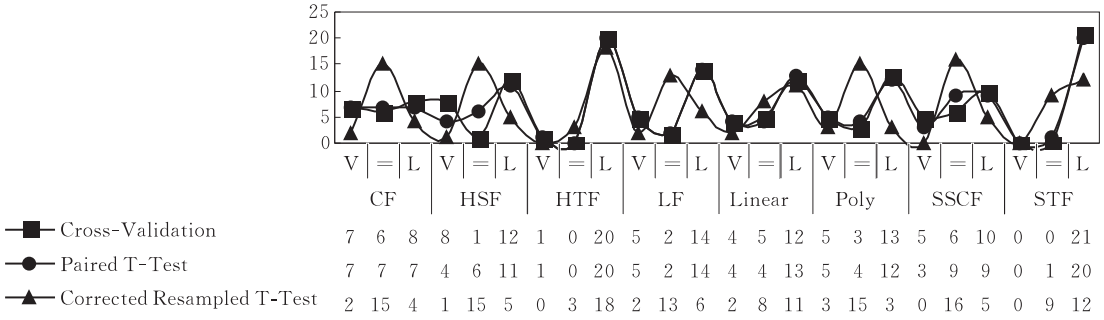
(图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF 核函数,数据代表相应的次数,统计显著性水平为 5%)

图 5 在精确率准则下对比 3 种统计方法



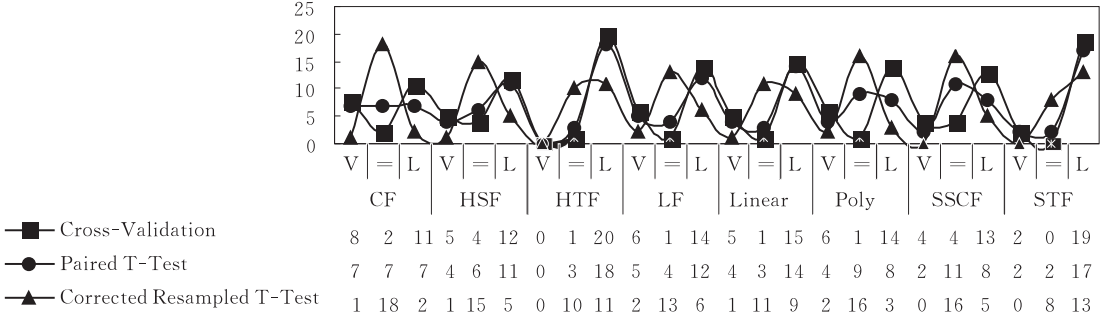
(图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF 核函数,数据代表相应的次数,统计显著性水平为 5%)

图 6 在召回率准则下对比 3 种统计方法



(图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF 核函数,数据代表相应的次数,统计显著性水平为 5%)

图 7 在 F 测量准则下对比 3 种统计方法



(图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF 核函数,数据代表相应的次数,统计显著性水平为 5%)

图 8 在信息增益准则下对比 3 种统计方法

其它核函数进行对比时,不同统计方法所得到的统计结果.图中符号(V、=和L)代表对比核函数的分类结果好于(V)、等于(=)还是差于(L)RBF 核函数,图中数据代表相应的次数,统计显著性水平为 5%.

从图 4~图 8 的统计结果中可以看出,3 种统计方法之间存在较大差异.例如,图 4 中,在准确率准则下,交叉验证的统计结果表明,Linear 的分类能力接近于 RBF,但配对 t 测试方法和纠正重复取样 t

测试的统计结果表明,Linear 的分类能力远远不如 RBF. 又例如,图 5 中,在精确率准则下,交叉验证和配对 t 测试方法的统计结果表明,CF 的分类能力不如 RBF,但纠正重复取样 t 测试的统计结果表明,CF 的分类能力接近 RBF.

当不同统计结果出现矛盾时,需要在应用多种统计方法的基础上,综合考评多种评估准则,然后根据大多数评估结果做出最后评判,真正好的分类模型在所有评估准则下所获得的结果都应当是最好的. 所以,综合图 4~图 8 的统计结果可以得出以下结论:

(1) Linear、HTF 和 STF 的分类能力都远远不如 RBF(取默认参数);

(2) CF 的分类能力最接近 RBF(取默认参数). 而这一结论与图 8 在信息增益准则下的统计结果最为相符,由此看出,信息增益准则最接近综合考评得出的结果.

实验 2. 对模型预测性能评估的多种评估准则(准确率、精确率、召回率、 F 测量和信息增益)进行对比分析. 表 7 显示了应用纠正重复取样 t 测试方法(统计显著性水平为 5%)进行核函数两两对比时,根据不同评估准则所得到的统计结果. 第 1 列是评估准则,第 2 列到第 10 列是参与评估的 9 个核函数,表中数据代表核函数两两对比的获胜次数与失败次数之差,括号内的数字指明该核函数在当前评估准则下的排名.

表 7 应用纠正重复取样 t 测试方法 5 种评估准则的对比结果

准则	RBF	CF	HSF	HTF	LF	Linear	Poly	SSCF	STF
Accuracy	55 (1)	44 (2)	27 (5)	-146 (9)	34 (3)	0 (7)	31 (4)	18 (6)	-63 (8)
Precision	51 (1)	46 (2)	37 (4)	-133 (9)	39 (3)	-21 (7)	31 (5)	22 (6)	-72 (8)
Recall	44 (1)	33 (3)	16 (4)	-123 (9)	16 (4)	-11 (7)	38 (2)	12 (6)	-22 (8)
F Measure	54 (1)	45 (2)	31 (4)	-132 (9)	31 (4)	-16 (7)	39 (3)	22 (6)	-74 (8)
Information Gain	47 (1)	40 (2)	21 (4)	-61 (8)	18 (5)	-15 (7)	31 (3)	6 (6)	-87 (9)

注:表中数据代表核函数两两对比的获胜次数与失败次数之差,括号内的数字指明该核函数在当前评估准则下的排名.

从表 7 的统计结果中可以得出以下结论:

(1) 虽然根据核函数两两对比的获胜次数与失败次数之差对核函数分类能力进行排序时,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是大体一致的;

(2) 对核函数分类能力的评估结论与实验 1 中的结论一致.

5 讨 论

核函数选择的准则和方法作为核方法及其应用的核心内容之一,目前在国际上还没有形成一个统一的模式,在解决实际问题时,人们往往只能是凭借经验,并采用试凑方式,由此产生较大的随意性. 因而有必要对各种常用的核函数进行分类能力的综合评估. 有关评估结论对于在没有先验知识情况下选择核函数具有重要的指导意义.

本文尝试将纠正重复取样 t 测试的统计方法应用到核函数选择中,提出了通过多种评估准则的综合应用来选择核函数的方法. 数值实验表明不同模型评估准则之间存在差异,但应用统计方法可以从这些差异中发现一些规律. 同时,不同统计方法之间也存在差异,且这种差异对模型评估的影响要大于

由于评估准则的不同而产生的影响. 所以,判断核函数分类能力要在应用多种统计方法的基础上,综合考评准确率、精确率、召回率、 F 测量和信息增益等多种评估准则,真正好的核函数分类模型在所有评估准则下所获得的结果都应当是优良的,如针对本文中的实验数据发现 RBF 核函数在各种准则与方法考察中总体最优. 但需要指出的是,统计方法是计算密集型的方法,为了缩短运算时间,可以考虑数据压缩以减少参加运算的数据量.

参 考 文 献

[1] Vapnik V. The Nature of Statistical Learning Theory. 2nd Edition. New York: Springer-Verlag, 2000

[2] Steinwart I. On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research, 2002, 2(2): 67-93

[3] Chalimourda A, Schölkopf B, Smola A. Experimentally optimal v in support vector regression for different noise models and parameter settings. Neural Networks, 2004, 17(1): 127-141

[4] Liu Xiang-Dong, Luo Bin, Chen Zhao-Qian. Optimal model selection for support vector machines. Journal of Computer Research and Development, 2005, 42(4): 576-581 (in Chinese)

(刘向东, 骆斌, 陈兆乾. 支持向量机最优模型选择的研究. 计算机研究与发展, 2005, 42(4): 576-581)

[5] Wang Ling, Bo Lie-Feng, Liu Fang, Jiao Li-Cheng. Least squares hidden space support vector machines. Chinese Journal of Computers, 2005, 28(8): 1302-1307(in Chinese)
(王玲, 薄列峰, 刘芳, 焦李成. 最小二乘隐空间支持向量机. 计算机学报, 2005, 28(8): 1302-1307)

[6] Wu Tao, He Han-Gen, He Ming-Ke. Interpolation based kernel function's construction. Chinese Journal of Computers, 2003, 26(8): 990-996(in Chinese)
(吴涛, 贺汉根, 贺明科. 基于插值的核函数构造. 计算机学报, 2003, 26(8): 990-996)

[7] Tan Y, Wang J. A support vector machine with a hybrid kernel and minimal vapnik-chervonenkis dimension. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(4): 385-395

[8] Chen Y-X, Wang J-Z. Support vector learning for fuzzy rule-based classification systems. IEEE Transactions on Fuzzy System, 2003, 11(6): 716-728

[9] Browne M W. Cross-validation methods. Journal of Mathematical Psychology, 2000, 44(1): 108-132

[10] Sincich T. Business Statistics by Example. 5th Edition. New Jersey: Prentice Hall, 1996

[11] Nadeau C, Bengio Y. Inference for the generalization error. Machine Learning, 2003, 52(3): 239-281

[12] Cover T M. Elements of Information Theory. 2nd Edition. New Jersey: John Wiley & Sons, 2006

[13] Racine J. Consistent cross-validators model-selection for dependent data: hv-block cross-validation. Journal of Econometrics, 2000, 99(1): 39-61

[14] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition. San Francisco: Morgan Kaufmann, 2005(in Chinese)
(董琳等译. 数据挖掘-实用机器学习技术. 第2版. 北京: 机械工业出版社, 2006)

[15] Wang Yong, Hu Bao-Gang. Study of the relationship between normalized information gain and accuracy, precision and recall//Proceedings of the 2007 Chinese Conference on Pattern Recognition (CCPR 2007). Beijing, 2007: 27-34(in Chinese)
(王泳, 胡包钢. 归一化信息增益准则与准确率、精确率、召回率的非线性关系研究//2007年全国模式识别学术会议(CCPR2007). 北京, 2007: 27-34)

[16] Hu Bao-Gang, Wang Yong. Applications of mutual information criteria in classification problems//Proceedings of the 2007 Chinese Conference on Pattern Recognition (CCPR 2007). Beijing, 2007: 35-45(in Chinese)
(胡包钢, 王泳. 关于互信息学习准则在分类问题中的应用//2007年全国模式识别学术会议(CCPR2007). 北京, 2007: 35-45)

[17] Newman D J, Hettich S, Blake C L, Merz C J. UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, 1998

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

[19] Cauwenberghs G, Poggio T. Incremental and decremental support vector machine learning//Proceedings of the Advances in Neural Information Processing Systems (NIPS-13). Cambridge MA: MIT Press, 2000: 409-415

[20] Kevin J L, Michael J W. Learning to tell two spirals apart//Proceedings of the 1988 Connectionist Models Summer School. 1988: 52-59



WANG Yong, born in 1975, Ph. D. candidate. His research interests include pattern recognition, knowledge discovery and data mining.

HU Bao-Gang, Ph. D. , professor, Ph. D. supervisor. His current research interests include pattern recognition and plant growth modeling.

Background

This research is supported by the National Natural Science Foundation of China, No. 60275025 ("Nonlinearity-variation-based Study of Intelligent Systems") and National Natural Science Foundation Outstanding Innovation Group Project, No. 60121302.

The nonlinear-variation ability of functions refers to the ability of functions to approximate a cluster or multi-clusters of nonlinear functions. Though some methods have been proposed to solve the problem from the aspects of "Non-linear domain analysis" and "the application of apriori knowledge",

in view of practical applications, how to measure the "nonlinear-variation ability" quantitatively is still the most difficult and the key part of the research. How to choose the kernel functions is a special case of it, and up to date there is still missing a well-accepted framework to guide kernel selections. Analyzing the nonlinear-variation ability of functions may give us a new choice for kernel selection.

In this paper, the authors applied statistical methods to study the problem of kernel selection quantitatively. K -fold cross-validation is a commonly used statistical method for kernel selection, while it is valid only if the independence between the data and the classifiers is guaranteed. In practical applications this premise condition is usually inaccessible. So

by employing the corrected resample t -test — A newly proposed statistical method, the authors compare it with other two statistical methods — k -fold cross-validation and paired t -test on nine normally used kernels to measure their classification abilities. In addition, a new quantitative criterion of evaluating kernel classification performance based on information gain is proposed, which is proved to be the nonlinear function of traditional criteria and with wider application range. Benchmark tests show that the proposed quantitative statistical method is valid, and the information gain criterion is simple, stable. Furthermore, it can make up other criteria to a certain extent. Similar systematic studies on statistical methods of kernel selection are seldom reported in now.