模糊 K近邻分类器在邻域风险最小化算法中的应用

杞 娴,殷 英,戴 琳
(昆明理工大学 理学院,云南 昆明 650091)

摘要: 在邻域风险最小化原则 (VRM)中运用模糊 K近邻分类器,来提出一种新的定义邻域半径的方法,从而得出一种新的 VRM算法.实例证明这一新算法对解决稀疏小样本的分类和回归有着较好的应用.

关键词: 支持向量机;模糊 K近邻
中图分类号: O235　　**文献标识码**: A　　**文章编号**: 1007 - 855X(2007)06 - 0108 - 05

Application of Fuzzy K Adjacent Classification to Vicinal Risk Minimization

QI Xian, YIN Ying, DAI Lin
(Faculty of Mathematics, Kunming University of Science and Technology, Kunming 650091, China)

Abstract: It is a new fuzzy method to apply K adjacent classification method to defining vicinal radius in vicinal risk minimization (VRM). In this paper, a new fuzzy VRM algorithm is proposed, and three numerical experiments are done to demonstrate the proposed approach's effectiveness in resolving sparse small samples' classification and regression.

Key words: support vector machine; fuzzy K - nearest neighbor

0 引 言

由 Vapnik等[1]提出的支持向量机 (Support Vector Machines, 简记为 SVM)是一种基于统计学习和结构风险最小化原理的新型学习机器.近年来,在 SVM的基础上,一些学者提出了邻域风险最小化 (Vicinal Risk Minimization, 简记为 VRM)原则.目前,VRM已被逐渐应用于模式识别[2]、回归估计[3]和各类金融时间序列预测问题[4]中.

VRM事实上是在不同的样本点上采用不同宽度的核来得到解,对核的宽度的定义,也就是对邻域半径的定义仅仅只考虑了样本点与其最近邻的距离.显然,这样的定义存在一个问题,就是当样本相对较为稀疏时,只考虑离该样本点最近一个点的距离而忽略与其距离较近的 K个点的距离及其距离差别是不恰当的.

在此,论文提出一种新的定义邻域半径的方法,即引入模糊 K近邻分类器[5]在样本点之间建立一个隶属度函数[6]来对邻域半径重新定义,从而在 VRM基础上给出一个新的模糊 K近邻 VRM算法.实例表明,文中提出新算法相比较以前的老算法有一定的改进,在稀疏小样本的分类和回归中展现出了一些优点.

1 VRM 原则

考虑训练集 $D=\{(x_i, y_i)\}_{i=1}^{N}$,其中 $x_i=(x_i^1, \cdots, x_i^n)\in R^n$ 为输入向量, $y_i\in R$ 为输出值, N 为样本数.函数估计问题的标准表示是在函数集 $F=\{f(x, a)\mid a\in\Lambda\}$中最小化泛函

收稿日期: 2007 - 05 - 22 **基金项目**:昆明理工大学青年基金 (项目编号: 2006 - 29).
第一作者简介:杞娴 (1979 -),女,教师.主要研究方向:智能学习,神经网络,支持向量机.
E - mail: qixiancc@sina.com

$$R(a) = \int L(y - f(x, a)) \, \mathrm{d}P(x, y) \tag{1}$$

其中 $L(u)$为一个给定的损失函数, $P(x, y)$为训练集数据所服从的未知概率分布. 由于概率分布 $P(x, y)$未知,通常用经验风险泛函

$$R_{emp}(a) = \frac{1}{n} \sum_{i=1}^{N} L(y_i - f(x_i, a))$$

来替换 (1). 但若密度函数和目标函数都是平滑的,那么经验风险泛函就可能不是对期望风险泛函的最好逼近.

为了寻找对 (1)更好的逼近,假设:未知密度函数任意点 x_i 的一个邻域内是平滑的;使得风险泛函最小的函数也是平滑的,且在任意点 x_i 的邻域内对称. 由此,我们可对所有训练向量 x_i 构造其邻域函数 $v(x_i)$, $i = 1, \cdots, N$,然后再用这些邻域函数构造目标泛函,在此我们将邻域函数区分为硬邻域函数和软邻域函数[1].

硬邻域函数定义了如下一般形式的邻域风险泛函:

$$V(a) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i - \frac{1}{A_i} \int_{v(x_i)} f(x, a) \, \mathrm{d}x\right)$$

而软邻域函数定义了如下一般形式的邻域风险泛函:

$$V(a) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i - \int f(x, a) \, p(x \mid x_i, r_i) \, \mathrm{d}x\right)$$

这里 $p(x \mid x_i, r_i)$为分布函数,例如 $p(x \mid x_i)$, $r_i = N(x_i, \iota_i)$.

在模式识别问题中,我们得到定理 1,在函数回归问题中,可得到定理 2

定理 1:邻域支持向量解的形式为: $f(x) = \sum_{i=1}^{N} \beta_i \phi(x, x_i) + b$,

其中 β为如下问题的最优解: $\max W(\beta) = \mathrm{f}(\mathrm{x}) = \sum_{i=1}^{N} \beta_i - \frac{1}{2} \sum_{i, j=1}^{N} y_i y_j \beta_i \beta_j \psi(x_i, x_j)$

$$\mathrm{s.t.} \quad \sum_{i=1}^{N} y_i \beta_i = 0, \; 0 \leq \beta_i, \beta_i \leq C, \; i, j = 1, \cdots, N$$

其中单邻域核

$$\psi(x, x_i) = E_{v(x_i)} K(x, x') = \int K(x, x') \, p(x' \mid x_i, r_i) \, \mathrm{d}x' \tag{2}$$

与双邻域核 $\psi(x_i, x_j) = E_{v(x_i)} E_{v(x_j)} K(x, x') = \iint K(x, x') \, p(x \mid x_i, r_i) \, p(x' \mid x_j, r_j) \, \mathrm{d}x \mathrm{d}x'$ (3)

这里, E表示期望; $K(x, y)$为核函数且满足 Mercer条件; b值可按 Karush - Kuhn - Tucker(KKT)[7]条件求得; C为一个给定的值. (定理 1的证明可参考文献 [1])

定理 2:邻域支持向量解为如下形式: $f(x) = \sum_{i=1}^{N} (\beta_i - \beta_j^*) \phi(x_i, x) + b$

其中参数 β, β^*为如下最优问题的解:

$$\max W(\beta) = -\varepsilon \sum_{i=1}^{N} (\beta_i + \beta_i^*) + \sum_{i=1}^{N} y_i (\beta_i - \beta_i^*) - \frac{1}{2} \sum_{i, j=1}^{N} (\beta_i - \beta_i^*)(\beta_j - \beta_j^*) \psi(x_i, x_j)$$

$$s.t. \quad \sum_{i=1}^{N} (\beta_i - \beta_i^*) = 0, \; 0 \leq \beta_i, \beta_i^{(*)} \leq C, \; i, j = 1, \cdots, N$$

这里单邻域核 $\phi(x, x_i)$与双邻域核 $\psi(x_i, x_j)$的定义及 b的求法等均与定理 1相同,且该定理的证明也与定理 1的证明类似 (参考文献 [1]).

2 模糊 K近邻 VRM 算法

VRM算法[1]对训练向量 $x_i = (x_i^1, \cdots x_i^n)$的邻域半径的定义只考虑了 x_i 与其最近邻的距离,并取这个距离的 η $(0 < \eta \leq 1/2)$ (通常取 1/2)倍作为 的邻域半径. 显然,当样本相对比较稀疏时,只考虑离该样本点最近一个点的距离而忽略它与其他 K个点的距离及其距离差别是不恰当的.

下面我们在定义邻域半径时,除考虑点之间的距离外,还将考虑点之间的距离差别.我们定义样本 x_j, $j=1\cdots,N$ 对于各个类的隶属度值为

$$H_i(x_j)=\frac{\sum_{s=1}^{k}\mu_i(x_s)\,(1/\|x_j-x_s\|^{2/(p-1)})}{\sum_{j=1}^{k}(1/\|x_j-x_s\|^{2/(p-1)})},\ i=1,\cdots,m \tag{4}$$

其中 $\mu_i(x_s)$为样本 x_s 对第 i类的隶属度值, $s\neq i$; $i=1,\cdots,N$.

要计算 $\mu(x_j)$,我们定义 x_i 与 x_j 的距离为 a_{ij}, $j\neq i$; $i=1,\cdots,N$, $j=1,\cdots,k$,这里 N 是样本总数.将 a_{ij}化为区间 $[0,1]$上的数 a_{ij}';其中

$$a_{ij}'=\frac{a_{ij}-\min\limits_{j\neq i}a_{ij}}{\max\limits_{j\neq i}a_{ij}-\min\limits_{j\neq i}a_{ij}},\ j\neq i;\ j,\ i=1,\cdots,N \tag{5}$$

$$\mu_i(x_j)=1-a_{ij}' \tag{6}$$

若 $H_i(x_j)=\max\limits_{s\neq i}H_i(x_s)$,则把 λa_{ij}定义为 x_i 的邻域半径 $(0<\lambda\leq 1/2)$.如此,就可以得到新的 VRM 算法即模糊 K近邻 VRM 算法.

考虑训练数据 $(x_1,y_1),\cdots,(x_N,y_N)$, $x_i=(x_i^1,\cdots,x_i^n)\in R^n$, $i=1,\cdots,N$.对硬邻域函数用 l_∞度量:

$\|x_j-x_s\|_\infty=\sup\limits_{1\leq k\leq n}|x_i^k-x_j^k|$,则模糊 K近邻 VRM 算法的步骤为:

1)根据训练数据定义向量 x_i 的邻域 $v(x_i)$:

a 计算 x_i 与 x_j 的距离 a_{ij},即 $a_{ij}\|x_j-x_s\|_\infty$, $j\neq i$; j, $i=1,\cdots,N$.

b 按 (5)式将 a_{ij}化为 $[0,1]$区间的数 a_{ij}'

c 选取与 x_i 距离最近的 k个点,并按 (6)式计算这 k个点对第 i类的隶属度 $\mu_i(x_j)$, $i=1,\cdots,N$, $j=1,\cdots,k$, $i\neq j$

d 按 (4)式计算 $H_i(x_j)$, $i=1,\cdots,N$, $j=1,\cdots,k$, $i\neq j$

e 若 $H_i(x_t)=\max\limits_{s\neq i}H_i(x_s)$,则把值 $\ _i=\lambda a_{it}$赋值给 x_i,把值 $\ _t=\lambda a_{it}$赋值给 x_t.这里 $0<\lambda\leq\frac{1}{2}$.若 x_i 曾被赋值,则只需把值赋给 x_t.

f 继续上述过程,直到所有向量都被赋值.

g 定义点 x_i 的邻域: $v(x_i)=\{x: x_i^k-\ _i\leq x^k\leq x_i^k=\ _i,\ k=1,\cdots,n\}$

2)分别根据 (2)和 (3)式计算出单邻域核 $\Phi(x,x_i)$与双邻域核 $\Psi(x,x_i)$.

3)根据定理 1(或定理 2)得出邻域支持向量解 $f(x)$,类似于 SVM 的方法,就可解决模式识别问题(或函数回归问题).

相应地,对软邻域函数,采用以下步骤:

1) 用 l_2 度量: $\|x_j-x_s\|_2=(\sum_{k=1}^{n}|x_i^k-x_j^k|^2)^{\frac{1}{2}}$定义两点之间的距离.

2) 沿用上面 ①~⑤的算法,对 x_i 定义出邻域半径 τ_i.

3) 利用参数 x_i, $\ _i$ 定义分布函数 $p(x|x_i,\ _i)$,如 $p(x|x_i,\ _i)=N(x_i,\ _i)$.

4) 沿用硬邻域函数模糊 VRM 算法的第 2,3步即可.

经过多次实验我们发现:参数 λ对模糊 VRM 算法有一定影响,不过我们只需几次取值就能找到较为理想的 λ值.另外,由于硬邻域函数有一定的"刚性",故我们在实际应用中可优先考虑采用基于软邻域函数的模糊 VRM 算法.

3 实 例

利用模糊 K近邻 VRM 算法和原有 VRM 算法进行比较,参数 C的选取采用 RM界[8].

第 1个例子是对二维平面 $(n=2)$上的 14个稀疏样本点进行分类.

采用损失函数: $L^*(y,f(x,a)=\theta(-yf(x,a)))$,其中 $\theta(u)$是阶跃函数;核函数采用 RBF核函数[9]

$K^*(x, y) = \exp\{-(x-y)^2/(2\sigma^2)\}$并令 $\sigma^2 = 0.0018$;取 $p(x|x_i, \ _i) = N(x_i, \ _i)$可计算出:

单邻域核函数: $\phi(x, x_i) = (1+\frac{a_i^2}{\sigma^2})^{-\frac{n}{2}} \exp\{-\frac{(x-x_i)^2}{2(\sigma^2+a_i^2)}\}$

双邻域核函数: $\psi(x_i, y_j) = (1+\frac{a_i^2}{\sigma^2}+\frac{a_j^2}{\sigma^2})^{-\frac{n}{2}} \exp\{\frac{(x_i-x_j)^2}{2(\sigma^2+a_i^2+a_j^2)}\}$, $i, j = 1, \cdots, 14$

VRM算法参数取为 $\lambda = 1/2$ (对 VRM算法 λ取 1/2最理想);模糊 K近邻 VRM算法参数取为 $\lambda = 1/8$.

经过多次实验,我们发现模糊 K近邻 VRM算法的分类效果明显比原有 VRM算法的好如图 1.另外,模糊 K近邻 VRM算法对螺旋型数据也有很好的分类效果如图 2.

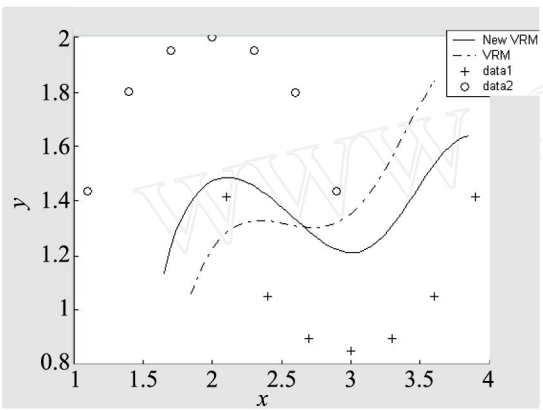

图1　模糊K近邻VRM和VRM对分类的比较

Fig.1　Different results of classification between fuzzy K adjacent VRM and VRM

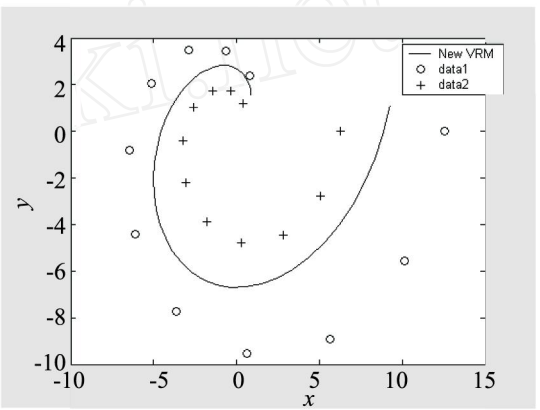

图 2　模糊K近邻VRM在分类中的应用

Fig.2　The application to fuzzy K adjacent VRM in classification

第 2个例子是函数回归估计问题.

这里被用于估计问题为 $z = (\sin\pi y)/(2+\sin\pi x) - 0.5 \le x \le 0.5, 0 \le y \le 1$,对 20个随机样本数据进行训练.取 $p(x)|x_i, r_i = N(x_i, r_i)$.模糊 K近邻 VRM算法参数取为 $\lambda = 1/2$.

从图 3,4可以看出,随样本数的减少,模型的泛化结果和泛化误差变化不大.这说明模糊 K近邻 VRM算法用于小样本逼近的能力很强.

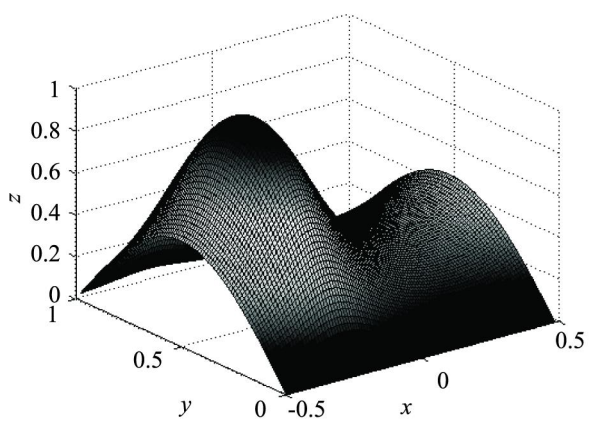

图3　被估计问题的真实图形

Fig.3　True figure of regression

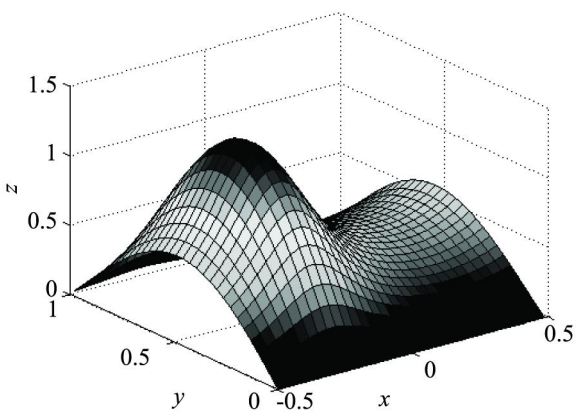

图4　模糊K近邻VRM在回归的应用

Fig.4　The application of fuzzy K adjacent VRM in regression

接下来,对这 20个随机样本数据进行训练,并加入服从均值为零的正态分布 $N(0, 0.02)$的噪声,即:

$z_{ij} = (\sin\pi y_j)/(2+\sin 2\pi x_i) + \zeta_{ij}$, $i, j = 1, \cdots, 30$

这里 ζ_i服从 $N(0, 0.02)$.

VRM算法参数取为 $\lambda = 1/2$,模糊 K近邻 VRM算法参数同上.

实验结果:由图 5,6可看出模糊 K近邻 VRM算法抗噪声能力比 VRM算法强,其逼近能力较 VRM算法稍好.

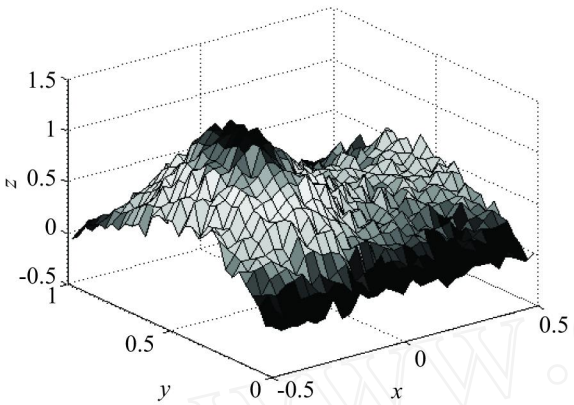

图5 ζ_i=0.02,模糊K近邻VRM在回归中的应用

Fig.5 ζ_i=0.02, the application of fuzzy K adjacent VRM in regression

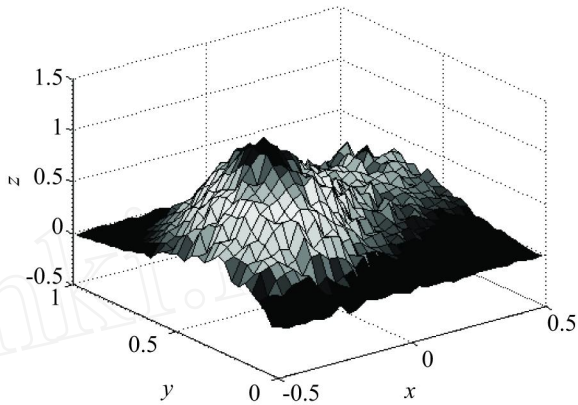

图6 ζ_i=0.02,VRM在回归中的应用

Fig.6 ζ_i=0.02, the application of VRM in regression

4 结束语

在邻域风险最小化原则中,对任意训练样本点,应用模糊 K近邻分类器,来提出一种新的定义邻域半径的方法,从而得出一种新的 VRM算法. 实例表明该方法对稀疏小样本训练集分类及回归是有效的. 当然,邻域风险最小化原则是一种非常新的原则,对这一原则还需做进一步的分析,如怎样增强用于回归的邻域风险最小化原则的抗噪声能力等等.

参考文献:

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. 张学工译. New Yerk: springer,清华大学出版社, 1995.

[2] Opper M, Winther O. Gaussian Processes for Classification[J]. Research Report, Neural Comput 2000(12): 2655 - 2684.

[3] Gao J B, Gunn S R. Mean field Method for the Support Vector Machine Regression[J]. Neurocomputing, 2003(50): 391 - 405.

[4] Francis E H Tay, Cao L J. Modified support vector machines in financial time series forecasting[J]. Neurocomputing, 2002(48): 847 - 861.

[5] 边肇祺,张学工. 模式识别 [M]. 北京:清华大学出版社, 2003.

[6] 张跃. 模糊数学方法及应用 [M]. 北京:煤炭工业出版社, 1992.

[7] Weston J, Gammerman A, Stitson M O, et al. And Kernel Methods Support Vector Learning[C]. MA: M IT Press, 1998. 293 - 306.

[8] Vapnik V N, Chapelleo. Bounds on Error Expectation for Support Vector Machine[J]. Neural Computation, 2000, (1299): 2 013 - 2 036.

[9] Burge C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge discovery, 1998(2): 121 - 167.