第 30 卷 第 1 期
2004 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 30, No. 1
Jan., 2004

最小平方误差算法的正则化核形式[1)]

许建华　张学工　李衍达

(清华大学自动化系智能技术与系统国家重点实验室　北京　100084)

(E-mail: xujianhua99@mail.tsinghua.org.cn)

摘　要　最小平方误差算法是最常用的一种经典模式识别和回归分析方法,其目标是使线性函数输出与期望输出的误差平方和为最小.该文应用满足 Mercer 条件的核函数和正则化技术,改造经典的最小平方误差算法,提出了基于核函数和正则化技术的非线性最小平方误差算法,即最小平方误差算法的正则化核形式,其目标函数包含基于核的非线性函数的输出与期望输出的误差平方和,及一个适当的正则项.正则化技术可以处理病态问题,同时可以减小解空间和控制解的推广性,文中采用了三种平方型的正则项,并且根据正则项的概率解释,详细比较了三种正则项之间的差别.最后,用仿真资料和实际资料进一步分析算法的性能.

关键词　非线性,支持向量机,平方误差,核形式,正则化

中图分类号　TP18

Regularized Kernel Forms of Minimum Squared Error Methods

XU Jian-Hua　ZHANG Xue-Gong　LI Yan-Da

(*State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing*　100084)

(E-mail: xujianhua99@mail.tsinghua.org.cn)

Abstract　Minimum squared error algorithm is one of the classical pattern recognition and regression analysis methods, whose objective is to minimize the squared error summation between the output of linear function and the desired output. In this paper, the minimum squared error algorithm is modified by using kernel functions satisfying Mercer condition and regularization technique, and a nonlinear minimum squared error algorithm based on kernels and regularization technique, *i.e.*, the regularized kernel form of minimum squared error algorithms is proposed. Its objective function includes squared error summation between the output of nonlinear function based on kernels and the desired output, and a proper regularization term. The regularization technique can handle ill-posed problems, reduce the solution space and control the generalization. Three regularization terms of square form are utilized in this paper. According to the probabilistic interpretation of regularization terms, the difference among three regularization terms is given

1) 国家自然科学基金(69885004)资助

Supported by National Natural Science Foundation of P. R. China(69885004)

收稿日期　2002-10-11　　收修改稿日期　2003-01-11

Received October 11, 2002; in revised form January 11, 2003

in detail. The synthetic and real data are used to analyze the algorithm performance.

Key words Nonlinear, support vector machines, squared error, kernel form, regularization

1 引言

在经典模式识别中,最小平方误差分类器(MSE)由于能够处理线性不可分问题,而得到广泛应用[1~3],并且 Fisher 判别分析和 Bayes 分类器是 MSE 的二个特例[1,4,5]. 在线性回归分析中,最小平方误差算法等价于最小二乘线性回归. 岭回归是线性回归的改进,其目标函数中增加一正则项(权向量的模平方),它可以提供一个比最小二乘法更为稳定的估计,且标准差也更小[6].

近几年来,支持向量机(SVM)已成为机器学习领域最有影响力的成果[7~9],它最引人注目的一个特点是利用满足 Mercer 条件的核函数实现非线性变换,而不用知道变换的具体形式[10]. 根据这一思想,其他学者改造一些传统的线性算法,提出了基于核函数的非线性算法,即核形式,例如核主成分分析(KPCA)[11,12]、核 Fisher 判别分析(KFD)[13,14]、最小二乘支持向量机(LS-SVM)[15]、岭回归的核形式(KRR)[16]等. 依据矩阵论,Ruiz 和 Lopez-de-Teruel [17]提出了一种基于广义逆的 MSE 核形式,这一形式的缺点是难于控制其推广性. 许建华等人[18]扼要地提出了二种 MSE 的核形式,并且重点证明它们与 KFD, LS-SVM 和 KRR 的关系.

在核算法中,由于所求参数的个数为样本数加 1,所以问题往往是病态的. 研究人员广泛使用了正则化方法[19,20]. 在 KFD 中[13],为了避免矩阵的奇异性,直接给矩阵的对角线元素加一个常数. 在 SVM[7~9], LS-SVM[15]和 KRR[16]中,特征空间的权向量 2 范数的平方在某种程度上也可以看成一个正则项. Smola 和 Scholkopf [21] 为模式识别、回归估计、函数逼近和多算子反演建立基于核的框架,包括岭回归、SVM 和正则化网络,其中一个重要的数学工具就是正则化技术. Smola 等人[22]研究了正则化算子与支持向量核的关系. 在神经网络领域,正则化技术已得到广泛应用,因为研究人员经验地发现在目标函数中增加一适当的正则项可以有效地改善神经网络的推广性[23]. 除了正则项这一术语,不同的学者使用不同的名称,例如权值衰减[24]、惩罚项[23]、函数光滑或平坦[25,26]、权值修剪[27,28]、先验概率[29~31]和最大间隔[7~9]. 但是,常用的正则项只有三种:平方或 Gauss 型;绝对值或 Laplace 型;归一化或 Cauchy 型[23,31]. Saito 和 Nakano[23]对三种正则项和不同的训练算法进行了详细的比较,其结论是平方正则项和二阶训练算法能够明显地改善神经网络的推广性和收敛性. 特别使我们感兴趣的是, Mackay[29,30]和 Williams [31]给正则项作出了概率解释,即三种正则项对应于不同有关解的先验信息.

本文应用满足 Mercer 条件的核函数推广经典最小平方误差算法,其目标是最小化基于核函数的非线性函数的输出与期望输出之间的误差平方和. 但是其对应的线性方程组却是病态的. 为了处理这一病态问题,同时为了减小模型空间和控制推广性,本文设计三种平方型的正则项,使目标函数包含误差平方和项和正则项. 适当地选取正则化参数可以使对应的线性方程组的系数矩阵为对称正定阵,有效地提高数值计算的稳定性. 我们把这项技术称为

最小平方误差算法的正则化核形式(简称为 RKMSE)，它可以处理非线性分类和回归问题. 根据正则项的概率解释，文中从理论上比较三种算法之间的差别. 仿真和实际数据实验结果进一步证明了 RKMSE 是一类非常有效的非线性算法.

2　经典最小平方误差算法

假设 $\boldsymbol{x}=\{(\boldsymbol{x}_1,y_1),\cdots(\boldsymbol{x}_l,y_l)\}$ 是一样本集，其中 $\boldsymbol{x}_i\in R^n, y_i\in R, i=1,\cdots,l$ 及 l 为样本总数. 对于二分类问题，设 l_1 个样本来自类 ω_1，l_2 个样本来自类 ω_2，且 $l=l_1+l_2$.

在经典 MSE 算法中，用线性函数作为判别函数或回归函数

$$f(\boldsymbol{x})=(\boldsymbol{w}\cdot\boldsymbol{x})+b \tag{1}$$

其中(·)是内积运算，$\boldsymbol{w}\in R^n$ 和 $b\in R$ 分别表示权值向量和阈值. MSE 的目标函数定义为误差平方和[1~3]

$$E_0(\boldsymbol{w},b)=\frac{1}{2}(\boldsymbol{y}-A\boldsymbol{w}-b\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{y}-A\boldsymbol{w}-b\boldsymbol{u}) \tag{2}$$

其中 $A=[\boldsymbol{x}_1\quad\cdots\quad\boldsymbol{x}_l]^{\mathrm{T}}$，$\boldsymbol{y}=[y_1,\cdots,y_l]^{\mathrm{T}}$，$\boldsymbol{u}$ 是所有元素都为 1 的 l 维列向量. 如果计算式(2)关于 $\boldsymbol{w},b$ 的梯度，并令它们为零，则 MSE 算法满足的线性方程组为

$$\begin{bmatrix}A^{\mathrm{T}}A & A^{\mathrm{T}}\boldsymbol{u}\\ \boldsymbol{u}^{\mathrm{T}}A & l\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ b\end{bmatrix}=\begin{bmatrix}A^{\mathrm{T}}\boldsymbol{y}\\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y}\end{bmatrix} \tag{3}$$

对于模式识别问题，MSE 解的性质取决于如何选取样本的期望输出[1]. 一种方式是

$$y_i=\begin{cases}+1, & \text{if } \boldsymbol{x}_i\in\omega_1\\ -1, & \text{if } \boldsymbol{x}_i\in\omega_2\end{cases} \tag{4}$$

这时 MSE 解当样本数趋向无穷时以最小均方误差逼近 Bayes 判别函数；另一种方式是

$$y_i=\begin{cases}+\dfrac{l}{l_1}, & \text{if } \boldsymbol{x}_i\in\omega_1\\ -\dfrac{l}{l_2}, & \text{if } \boldsymbol{x}_i\in\omega_2\end{cases} \tag{5}$$

这时 MSE 解等价于 Fisher 线性判别分析.

对于线性回归分析，期望输出选为样本的输出，这时 MSE 就是最小二乘回归方法. 在岭回归中，目标函数中增加一个正则项($(\boldsymbol{w}\cdot\boldsymbol{w})+b^2$)，其对应的线性方程组相当于在方程(3)的对角元素加一个正常数(正则化参数)，但它可以提供一个比最小二乘法更为稳定的估计，且标准差也更小[6].

3　最小平方误差算法的正则化核形式

假设 Φ 为一将输入空间的向量变换到某一特征空间 F 的向量的非线性变换

$$\Phi:R^n\to F \tag{6}$$

在特征空间 F 中，构造一个权值向量 w^{Φ} 和阈值 β 的线性函数

$$f^{\Phi}(\boldsymbol{x})=(\boldsymbol{w}^{\Phi}\cdot\Phi(\boldsymbol{x}))+\beta \tag{7}$$

依据再生核空间理论，特征空间的解必定位于全体样本所张成的空间中，即解是全体样本的

线性组合[13,32](也可参考 Scholkopf 等人的技术报告[1])

$$\boldsymbol{w}^{\Phi} = \sum_{i=1}^{l} \alpha_i \Phi(\boldsymbol{x}_i) \tag{8}$$

其中 $\alpha_i \in R(i=1,2,\cdots,l)$描述了每一个样本在权向量中所起作用的大小. 把式(8)和核函数的定义

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)) \tag{9}$$

代入式 (7), 则得到空间 F 中的函数

$$f^{\Phi}(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \beta \tag{10}$$

类似地,特征空间 MSE 算法的目标函数可以定义为式(10)的输出与期望输出的平方误差和

$$E_0^{\Phi}(\boldsymbol{\alpha}, \beta) = \frac{1}{2}(\boldsymbol{y} - K^{\mathrm{T}}\boldsymbol{\alpha} - \beta \boldsymbol{u})^{\mathrm{T}}(\boldsymbol{y} - K^{\mathrm{T}}\boldsymbol{\alpha} - \beta \boldsymbol{u}) \tag{11}$$

其中 $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_l]^{\mathrm{T}}$, $(K)_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)(i,j=1,\cdots,l)$为满足 Mercer 条件的半正定核矩阵[8,33]. 最小化式(11),得到对应的线性方程组为

$$\begin{bmatrix} KK^{\mathrm{T}} & K\boldsymbol{u} \\ (K\boldsymbol{u})^{\mathrm{T}} & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} = \begin{bmatrix} K\boldsymbol{y} \\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{12}$$

因为要用 l 样本估计 $l+1$ 个参数,上述方程的系数矩阵一定是病态的,线性方程组(12)存在多解问题.

文献 [24]指出,当两个分类器具有相同的训练误差,容量小的分类器的推广能力更好. 为了从多个解中挑选一个解,可以在目标函数中加入一个适当的正则项. 文献[21] 说明正则项可以有效地减少解空间、控制解的容量和推广性. 因此在目标函数中加入一个正则项,既可以解决解的唯一性问题,又能控制解的推广性. 在基于核的训练算法中,常用的平方型正则项有三个:

1) $(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha}) + \beta^2$ 在类似于 KFD 中使用[13];

2) $(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha})$在 KFD[13]和岭回归[21]中使用;

3) $(\boldsymbol{w}^{\Phi} \cdot \boldsymbol{w}^{\Phi})$在 SVM[7~9,21], LS-SVM[15]和 KRR[16]中使用.

把这三种正则项加入目标函数(11),则得到相应的正则化目标函数

$$E_1(\boldsymbol{\alpha}, \beta) = \frac{1}{2}\mu_1((\boldsymbol{\alpha} \cdot \boldsymbol{\alpha}) + \beta^2) + E_0^{\Phi}(\boldsymbol{\alpha}, \beta) \tag{13}$$

$$E_2(\boldsymbol{\alpha}, \beta) = \frac{1}{2}\mu_2(\boldsymbol{\alpha} \cdot \boldsymbol{\alpha}) + E_0^{\Phi}(\boldsymbol{\alpha}, \beta) \tag{14}$$

$$E_3(\boldsymbol{\alpha}, \beta) = \frac{1}{2}\mu_3(\boldsymbol{w}^{\Phi} \cdot \boldsymbol{w}^{\Phi}) + E_0^{\Phi}(\boldsymbol{\alpha}, \beta) \tag{15}$$

上式中μ_1, μ_2, μ_3为正常数,即正则化参数,可以在解的光滑程度和样本数据的逼近程度之间取折衷[25];

$$(\boldsymbol{w}^{\Phi} \cdot \boldsymbol{w}^{\Phi}) = \boldsymbol{\alpha}^{\mathrm{T}} K \boldsymbol{\alpha} \tag{16}$$

由最小化目标函数(13)~(15),可得到三个线性方程组:

$$\begin{bmatrix} KK^{\mathrm{T}} + \mu_1 I & K\boldsymbol{u} \\ (K\boldsymbol{u})^{\mathrm{T}} & l + \mu_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} = \begin{bmatrix} K\boldsymbol{y} \\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{17}$$

1) Scholkopf B, Herbrich R, Smola A J, Williamson R. A generalized representer theorem. NeuroCOLT2 Technical Report, 2000(http://www.kernel-machines.org/publications)

$$\begin{bmatrix} KK^{\mathrm{T}}+\mu_2 I & K\boldsymbol{u} \\ (K\boldsymbol{u})^{\mathrm{T}} & l \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}=\begin{bmatrix} K\boldsymbol{y} \\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{18}$$

$$\begin{bmatrix} K+\mu_3 I & \boldsymbol{u} \\ \boldsymbol{u}^{\mathrm{T}} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}=\begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \tag{19}$$

在推导方程(19)时，假设核矩阵是非奇异的，并且经过了适当地运算使系数矩阵为对称阵. 从数值计算角度来看，如果正则化参数(μ_1，μ_2，μ_3)足够大，方程(17)～(19) 中系数矩阵就可能成为对称正定阵，使数值计算更加稳定.

本文推导了特征空间的三种线性函数或原始空间中基于核的非线性函数的估计算法. 由于目标函数中包含满足 Mercer 条件的核函数和正则项，我们称这类非线性的训练算法为最小平方误差算法的正则化核形式(regularized kernel form of minimum squared error，RKMSE)，并将式(17)～(19)分别称为 RKMSE-$\alpha\beta$，RKMSE-α 和 RKMSE-w.

对于分类问题，与经典 MSE 相比，RKMSE 的解不仅取决于期望输出，而且取决于正则项. 许建华等人[18]证明了 KFD，LS-SVM 和 KRR 是 RKMSE 的三种特殊形式.

4　三种 RKMSE 算法的差别

三种 RKMSE 算法的差别是它们的目标函数中包含不同的正则项. 对于正则项，不同的学者给出了不同的解释. Mackay[30] 和 Williams[31] 提出了正则项的概率解释. 针对前向神经网络，Williams[31] 指出，根据 Bayes 观点正则项对应于模型参数一种先验分布. 本文依据这种概率解释，比较三种 RKMSE 之间的差别. 这里仍用 Mackay[30] 的概念和符号，正则化目标函数定义为

$$M(\boldsymbol{w})=\theta_D E_D(\boldsymbol{w})+\theta_W E_W(\boldsymbol{w}) \tag{20}$$

上式中 E_D 度量数据拟合的偏离程度；E_W 是正则项；θ_D，$\theta_W>0$ 为正则化参数. Mackay[30] 提出，如果 E_D 是二次误差函数(如函数(11))，则对应于假设期望输出中包含方差为 $\frac{1}{\theta_D}$ 的 Gauss 噪声；如果 E_W 为 $\sum_{i=1}^{m}\frac{1}{2}w_i^2$，则希望权值来自均值为 0、方差为 $\frac{1}{\theta_W}$ 的 Gauss 分布. Williams[31] 进一步说明，如果 E_D 是二次误差函数，则对应于假设期望输出中包含 Gauss 噪声的均值也为 0.

对于 RKMSE，也可以把它看成一个前向神经网络，其拓扑结构如图 1 所示.

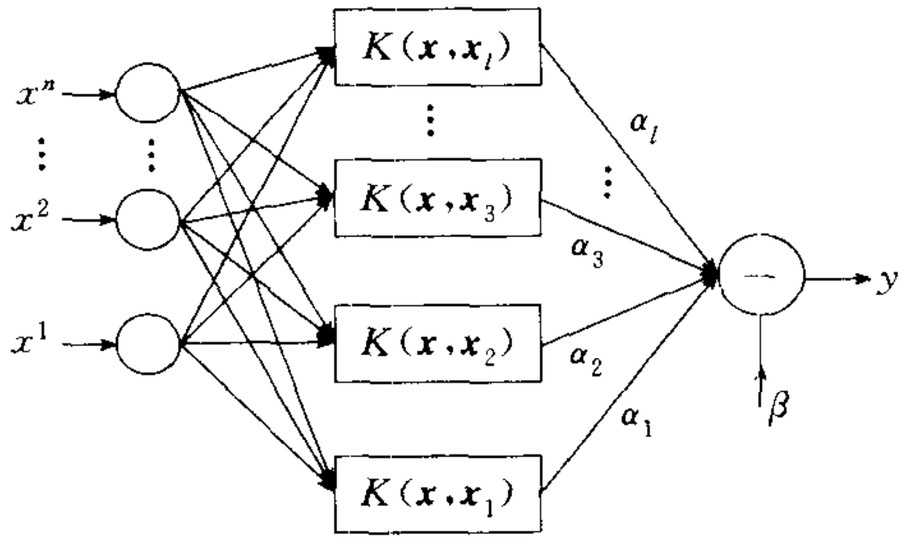


图 1　RKMSE 算法的拓扑结构

Fig. 1　Topological structure of RKMSE algorithm

根据正则项的概率解释及本文的目标函数形式,对目标函数(13),$\{\alpha_1,\cdots,\alpha_l,\beta\}$满足均值为 0、方差与$\mu_1$成反比的 Gauss 分布;对目标函数(14),$\{\alpha_1,\cdots,\alpha_l\}$满足均值为 0、方差与$\mu_2$成反比的 Gauss 分布.

对于正则化目标函数(15),我们给出更详细的解释. 类似于 LS-SVM 和 KRR,可以用二次规划来表示正则化核最小平方误差形式(15)

$$\min_{w^{\Phi},\beta,e} E(w^{\Phi},\beta,e)=\frac{1}{2}(w^{\Phi}\cdot w^{\Phi})+\frac{1}{2\mu_3}\sum_{k=1}^{l}e_k^2 \tag{21}$$

$$\text{s. t. } (w^{\Phi}\cdot\Phi(x_k))+\beta=y_k-e_k,\quad k=1,\cdots,l \tag{22}$$

则对应的 Lagrange 函数为

$$L=E(w^{\Phi},\beta,e)-\sum_{k=1}^{l}\alpha_k((w^{\Phi}\cdot\Phi(x_k))+\beta-y_k+e_k) \tag{23}$$

其中 α_k为 Lagrange 乘子. 问题(23)的最优性条件是

$$\frac{\partial L}{\partial w^{\Phi}}=0\Rightarrow w^{\Phi}=\sum_{k=0}^{l}\alpha_k\Phi(x_k) \tag{24}$$

$$\frac{\partial L}{\partial\beta}=0\Rightarrow\sum_{k=1}^{l}\alpha_k=0 \tag{25}$$

$$\frac{\partial L}{\partial e_k}=0\Rightarrow\alpha_k=\frac{1}{\mu_3}e_k \tag{26}$$

$$\frac{\partial L}{\partial\alpha_k}=0\Rightarrow(w^{\Phi}\cdot\Phi(x_k))+\beta-y_k+e_k=0 \tag{27}$$

从上述方程,可以得到线性方程组(19)和重要的关系式(25),(26). 根据式(25)和(26),有$\{e_1,\cdots,e_l\}$满足 Gauss 分布,其均值为 0;如果假设方差为 σ^2,则 $\{a_1,\cdots,a_l\}$满足均值为 0、方差为$\frac{\sigma^2}{\mu_3^2}$的 Gauss 分布.

当正则化参数μ_1,μ_2,μ_3增大时,解的方差相应减小,这意味着解的大小总体上减小,特别是当正则化参数为无穷大,方差变成 0. 依据正则项的概率解释,有以下结论:

1) 对方程(17),$\{\alpha_1,\cdots,\alpha_l,\beta\}$的均值为 0,方差与正则化参数成反比,无论是回归问题还是分类问题,当正则化参数趋向无穷大时,$\{\alpha_1,\cdots,\alpha_l,\beta\}$趋于 0;

2) 对方程(18),$\{a_1,\cdots,a_l\}$的均值为 0,方差与正则化参数成反比;对方程(19),$\{a_1,\cdots,a_l\}$的均值为 0,方差与正则化参数平方成反比;

3) 对于回归问题,当正则化参数趋向无穷大时,方程(18)和(19)中$\{a_1,\cdots,a_l\}$趋于 0,而 β趋于样本期望输出的平均值;

4) 对于期望输出为式(5)的分类问题,当正则化参数趋向无穷大时,方程(18)和(19)的解$\{\alpha_1,\cdots,\alpha_l,\beta\}$趋向于 0;

5) 对于期望输出为式(4)的分类问题,当正则化参数趋向无穷大时,方程(18)和(19)中的解$\{a_1,\cdots,a_l\}$趋于 0,而 β趋向于$\frac{l_1-l_2}{l}$,这一点意味着随着样本的增加,方程(18)和(19)的解趋向于二类样本的先验概率差.

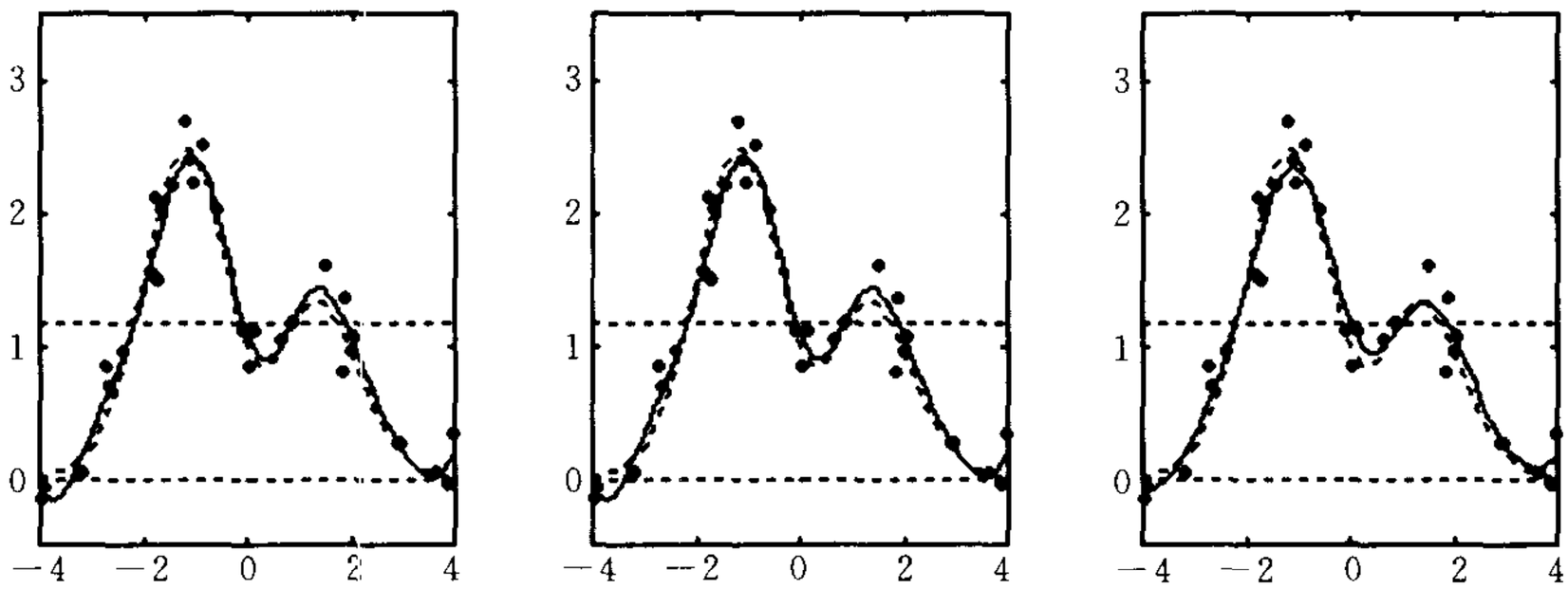
5　实验结果与分析

我们设计三个实验进一步从计算上分析三种最小平方误差算法的正则化核形式的性能：非线性回归实验、双螺旋线分类和图像分割数据.

5.1　非线性回归

对回归分析，我们采用 Satio 等人[23]使用的函数 $f(x)=(1-x+2x^2)e^{-0.5x^2}$. 在实验中，30 个训练样本的 x 值从[−4，+4]范围内随机产生，对应的函数值中加入均值为 0、标准差为 0.2 的 Guass 噪声. 测试样本的 x 值是等间隔的 81 个样本点，文中用均方根误差来评价回归效果. 我们采用 RBF 核函数，实验中 RBF 的参数分别取为 0.25，0.5，1.0，2.0，正则化参数取为 10^{-5}，10^{-4}，10^{-3}，10^{-2}，10^{-1}，1，10，100. 当 RBF 参数固定时，正则化参数增大，回归曲线越来越光滑，均方根误差越来越大；正则化参数减小，回归曲线越来越拟合样本点，均方根误差也越来越大，在一个适当的正则化值，均方根误差达到最小，回归效果最佳；当正则化参数固定，RBF 参数从小变到大，均方根误差也有由大变小再变大的规律. 图 2 是三种算法比较理想的结果，图中已标出对应的参数，其中黑点表示样本，虚线是真实曲线，实线是回归曲线，上面的平行虚线是函数采样值的平均值，下面的平行虚线是 0 线.

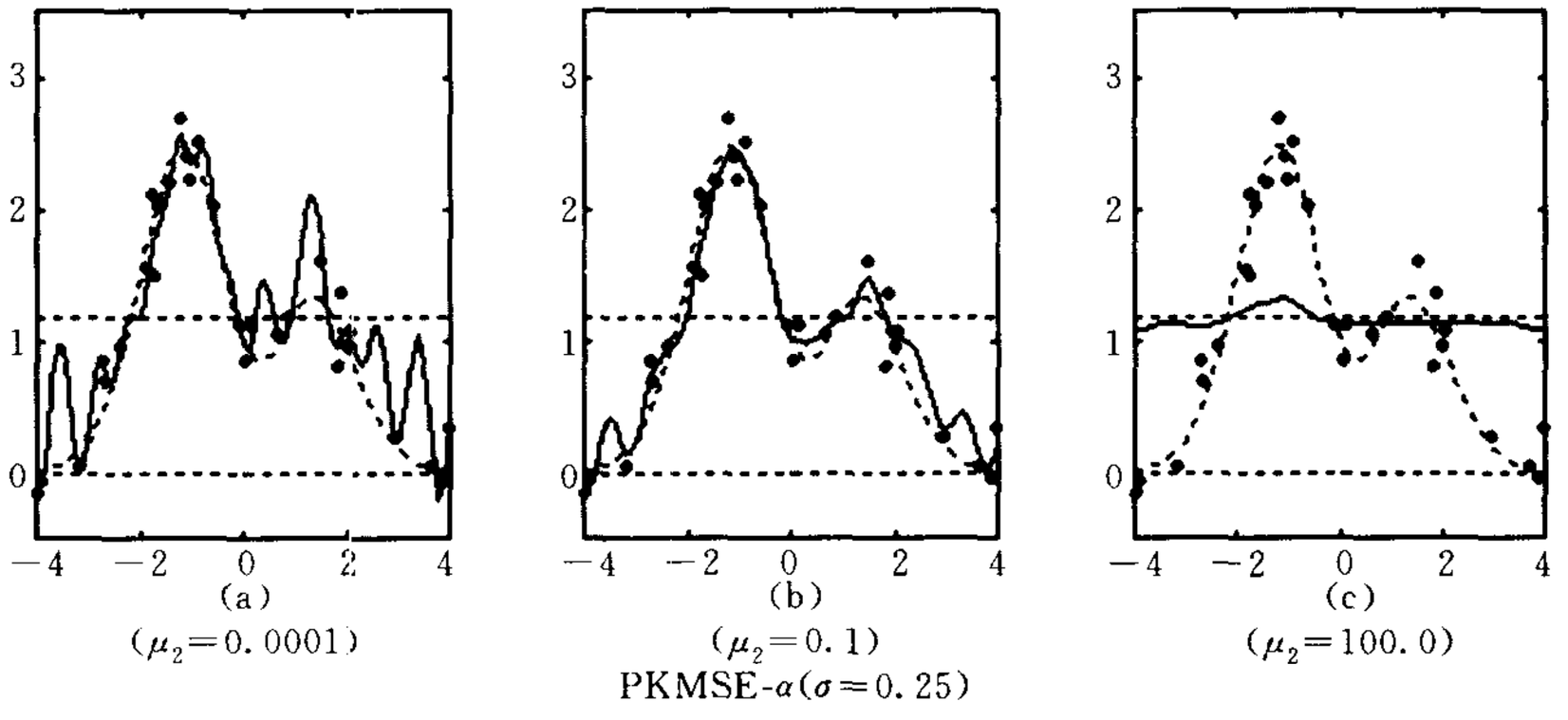
图 3 以 RKMSE-α 算法为例，说明正则化技术是如何改善回归结果. 当选定较小的 RBF



(a) PKMSE-$\alpha\beta$ ($\sigma=1,\mu_1=0.01$)　(b) PKMSE-α ($\sigma=1,\mu_2=0.01$)　(c) PKMSE-w ($\sigma=1,\mu_3=0.1$)

图 2　三种方法比较理想的回归结果(图中标出了 RBF 核参数和正则化参数)

Fig. 2　The best regression results corresponding to three algorithms(The RBF kernel parameter and regularization parameter are listed)



(a) ($\mu_2=0.0001$)　(b) ($\mu_2=0.1$)　(c) ($\mu_2=100.0$)

PKMSE-α($\sigma=0.25$)

图 3　RKMSE－α 算法不同正则化参数的回归结果

Fig. 3　The regression curves corresponding to different regularization parameters with algorithm RKMSE-α

参数时，随着正则化参数的增大，回归曲线越来越光滑，拟合样本点的程度越来越差. 左边的图 3(a)明显过学习，右边的图 3(c)则欠学习. 尽管与图 2 的同一方法相比，图 3(b)在细节上还存在一些差别，但是基本上反映了函数变化规律. 这说明正则化技术可以有效的改善算法的推广性能.

5.2 双螺旋线问题

双螺旋线问题(图 4)要求将位于两条螺旋线上的样本正确地分类. 这是一个高度非线性问题，因为任何一线性分类器都会有一半样本错分. 采用径向基函数核，得到了非常光滑分界面. 图 4(a)是采用第二种方法，其中期望输出为公式(5)；图 4(b)是采用第三种方法，其中期望输出为公式(4).

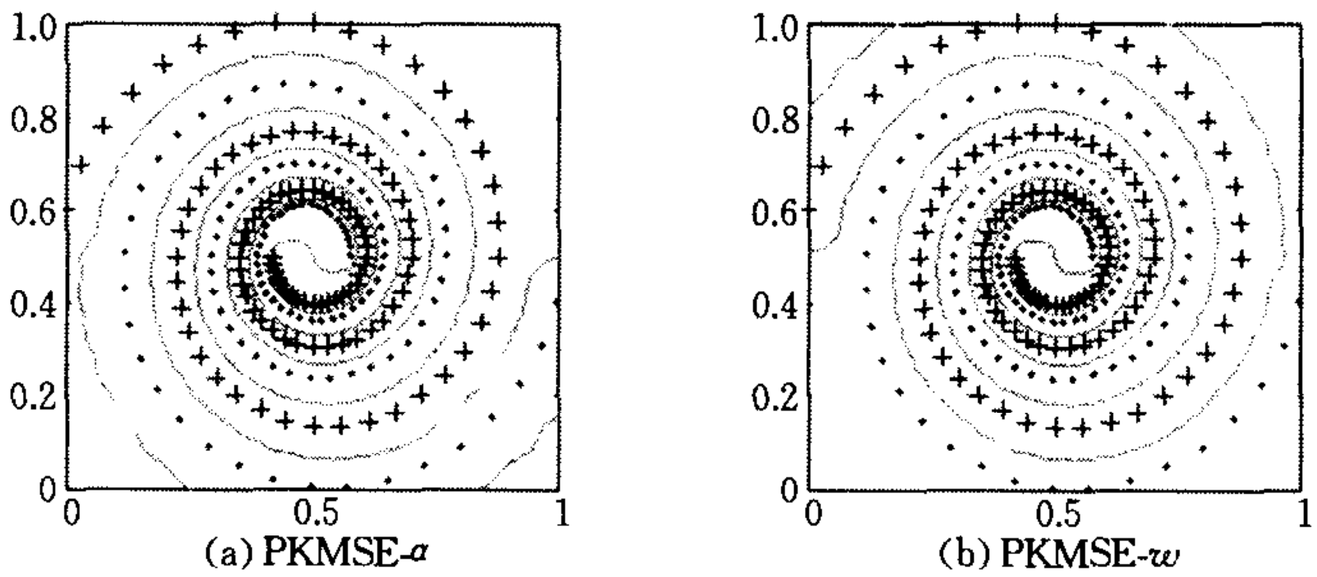


图 4 双螺旋线分类问题

Fig. 4 The classification hyperplanes of two-spiral problem

5.3 图像分割问题

图像分割问题的数据来自 www. cs. toronto. edu/～delve/data/image-seg/desc. html. 该数据集有七类样本：水泥地面，砖面，草地，树叶，天空，小路和窗户，每类有 30 个训练样本和 300 个测试样本，每一样本有 18 个来自原始图像的属性. 为了处理七类问题，我们设计 6 个二类分类器，并且把数据变换到 0 与 1 之间. 仍采用 RBF 核函数，RKMSE-α 的正确率为 85.86%($\sigma=0.25,\mu_2=0.0001$)，RKMSE-w 的正确率为 93.38%($\sigma=1.0,\mu_3=0.0001$). 就这一问题，第三种方法的正确率高于第二种方法.

6 讨论

经典的最小平方误差算法是一种常用的线性分类器设计和回归分析方法，本文应用满足 Mercer 条件的核函数和正则化技术，改造经典的最小平方误差算法，提出最小平方误差算法的正则化核形式(RKMSE). 在 RKMSE 中，目标函数包含两项：基于核函数的非线性函数与期望输出的误差平方和、平方型的正则项. 本文采用了三种正则项，相应获得三种最小平方误差的正则化核形式，其解满足系数矩阵为对称正定的线性方程组. 根据正则项的概率解释，文中给出三种方法之间的关系和差别. 实验中，我们研究了非线性回归和分类问题，进一步说明核函数和正则项在算法中的作用.

References

1 Duda R O, Hart P E. Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973

2 Tou J T, Gonzadez R C. Pattern Recognition Principles. Reading: Addison-Wesly, 1974

3 Bian Zhao-Qi, Zhang Xue-Gong *et al*. Pattern Recognition(Second Edition). Beijing: Tsinghua University Press, 2000(in Chinese)

4 Koford J S, Groner G F. The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE Transactions on Information Theory*, 1966, **12** (1): 42～50

5 Patterson J D, Womack B F. An adaptive pattern classification system. *IEEE Transactions on System Science and Cybernetics*, 1966, **2** (1): 62～67

6 Wang Hui-Wen. Partial Least Sequares Regression Method and Its Applications. Beijing: National Defence Industrial Press, 1999(in Chinese)

7 Cortes C, Vapnik V N. Support vector networks. *Machine Learning*, 1995, **20**(3): 273～297

8 Vapnik V N. Statistical Learning Theory. New York: Wiley, 1998

9 Vapnik V N. The Nature of Statistical Learning Theory (2nd ed). New York: Springer-Verlag, 1999

10 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32～44(in Chinese)

11 Scholkopf B, Smola A, Muller K-R. Kernel principal component analysis. In Proceedings ICANN'97, Springer Lecture Notes in Computer Science, 1997. 583～589

12 Scholkopf B, Smola A, Muller K-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 1998, **10**(5): 1299～1319

13 Mika S, Ratsch G, Weston J, Scholkopf B, Muller K R. Fisher discriminant analysis with kernels. Neural Networks for Signal Processing IX, New York: IEEE Press, 1999. 41～48

14 Muller K-R, Mika S, Ratsch G, Scholkopf B. An introduction to kernel-based learning algorithm. *IEEE Transactions on Neural Networks*, 2001, **12**(2): 181～201

15 Suykens J A K, Vandewalle J. Least squares support vector machines. *Neural Processing Letters*, 1999, **9**(3): 293～300

16 Saunders C, Gammerman A, Vovk V. Ridge regression leraning algorithm in dual variables. In: Shavlik J editor. Machine Learning Proceedings of the 15th International Conference, Morgan Kaufmann, 1998

17 Ruiz A, Lopez-de-Teruel P E. Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 2001, **12**(1): 16～31

18 Xu J, Zhang X, Li Y. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. In: International Joint Conference on Neural Networks 2001. Washington DC, USA: IEEE Press, 2001. 1486～1491

19 Tikhonov A N, Arsenin V Y. Solution of ill-posed problem. Washington DC: Winston and Sons, 1977

20 Tikhonov A N, Goncharsky A V. Ill-posed problems in the natural sciences. Translated from Russian by Bloch M, Moscow: MIR Publishers, 1987

21 Smola A J, Scholkopf B. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica*, 1998, **22**(1～2): 211～231

22 Smola A J, Scholkopf B, Muller K-R. The connection between regularization operators and support vector kernels. *Neural Networks*, 1998, **11**(4): 637～649

23 Saito K, Nakano R. Second order learning algorithm with squared penalty term. *Neural Computation*, 2000, **12** (3): 709～729

24 Guyon I, Stork D G. Linear discriminant and support vector classifiers. In: Smola A J, Bartlett P, Scholkopf B, Schuurmans C, editors (2000) Advances in Large Margin Classifiers, MIT Press, 2000

25 Poggio T, Girosi F. Regularization algorithm for learning that are equivalent to multi-layer networks. *Science*, 1990, **247**(23): 978～982

26 Bishop C M. Curvature driven smoothing: a learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 1993, **4**(5): 882～884

27 Reed R. Pruning algorithms——A survey. *IEEE transactions on Neural Networks*, 1993, **4**(5): 740～747

28 Ishikawa M. Structural learning with forgetting. *Neural Networks*, 1992, **9**(3): 509～521

29 Mackay D J C. Bayesian interpolation. *Neural Computation*, 1992, **4**(3): 415～447

30 Mackay D J C. A practical Bayesian framework for back-propagation networks. *Neural Computation*, 1992, **4**(3): 448～472

31 Williams P M. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 1995, **7**(1): 117～143

32 Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In: Scholkopf B, Burges C J C, Smola A J, editors, Advances in Kernel Methods——Support Vector Learning, Cambridge, MA: MIT Press, 1999. 69～88

33 Cristianini N, Taylor J S. An introduction to support vector machines and other kernel-based learning methods. Cambridge UK: Cambridge University Press, 2000

许建华 1985 年获得成都地质学院应用地球物理系工学学士,1989 年获得中国科技大学地球和空间科学系理学硕士,现为清华大学自动化系博士研究生. 主要从事信号处理、模式识别、神经网络、机器学习等领域的理论和应用研究.

(**XU Jian-Hua** Received his hachelor degree from Chengdu Institute of Geology in 1985 and the master degree from University of Science and Technology of P. R. China in 1987. Currently he is a Ph. D. candidate at Tsinghua University. His research interests include signal processing, pattern recognition, neural network, and machine learning and their applications.)

张学工 1994 年获清华大学自动化系工学博士学位,现为清华大学自动化系教授,信息处理研究所所长. 主要从事信号处理、模式识别、神经网络、统计学习理论研究.

(**ZHANG Xue-Gong** Received his Ph. D. degree in pattern recognition and intelligent system from the Department of Automation, Tsinghua University. He is a professor and Director of Institute of Information Processing in the Department of Automation at Tsinghua University. His research interests include bioinformatics and computational biology, pattern recognition, neural networks, and statistical learning theory.)

李衍达 1959 年毕业于清华大学自动控制系,中国科学院院士,清华大学自动化系教授,清华大学信息科学与技术学院院长. 主要从事信号处理理论、地震勘探数据处理与生物信息学研究.

(**LI Yan-Da** Graduated from the Department of Automatic Control, Tsinghua University in 1959. He is a professor, member of Chinese Academy of Sciences, Dean of School of Science and Technology, Director of Institute of Bioinformatics at Tsinghua University. His research interests include signal processing, intelligent control, and bioinformatics.)