# 基于流形正则化和核方法的最小二乘算法

汪宝彬,彭超权,李学锋
(中南民族大学 数学与统计学学院,武汉 430074)

**摘 要** 研究了再生核希尔伯特空间中流形正则化下的最小二乘算法的学习能力和收敛速度. 该算法能够充分利用输入空间的几何特点以及半监督学习中无标记样本的信息,提高算法的有效性和学习效率. 另外,讨论了该算法中正则参数的选取,这对算法实现具有现实的意义.

**关键词** 流形学习; 正则化; 最小二乘算法; 核方法; 再生核希尔伯特空间
**中图分类号** O211 **文献标识码** A **文章编号** 1672-4321(2017)04-0143-03

# The Least Square Algorithm Based on Manifold Regularization and Kernel Method

*Wang Baobin*, *Peng Chaoquan*, *Li Xuefeng*
(College of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China)

**Abstract** In this paper, we considered the learning ability and convergence rate of the least square algorithm under the manifold regularization in the Reproducing Kernel Hilbert Space (RKHS). This algorithm can make full use of the geometric construction characteristics of the input space and improve the validity and the learning efficiency of the classical least square algorithm by extracting the information from the unlabeled data. Moreover, we discussed the choice of the regularization parameter, which is meaningful to the design of the algorithm.

**Keywords** manifold learning; regularization; least square algorithm; kernel method; reproducing kernel Hilbert space

回归学习是机器学习中非常经典的问题,在很多实际背景中有系统的理论分析和数据模拟,尤其是以核方法为工具的研究引起了工业界、计算机以及统计领域的广泛关注. 回归学习的一般模型如下. 假设 $X$ 是数据输入空间,$Y$ 是输出空间,这里 $X$ 一般假设为 $R^n$ 空间的紧子集,$Y$ 是实数域中的有界子集. 希望找到 $X$ 与 $Y$ 之间的函数关系 $f: X \to Y$. 在回归模型中,有如下关系存在: $Y = f^*(X) + e$,其中 $e$ 是可加噪声,一般满足均值为零的条件. 假设 $\rho$ 是定义在 $X \times Y$ 上的联合 Borel 测度,$\rho_X$ 是关于 $X$ 的边缘分布,$\rho(y \mid x)$ 是任意 $x \in X$ 的条件概率测度,那么可以看到 $f^*(X)$ 是 $Y$ 关于 $\rho(y \mid x)$ 的条件均值,也就是 $f^*(X) = E(y \mid x) = \int_Y y \mathrm{d}\rho(y \mid x)$, $x \in X$. 显然,目标函数 $f^*(X)$ 依赖于样本测度 $\rho$,但在实际应用中不可能事先知道. 因此,借助抽样数据 $z = \{(x_i, y_i)\}_{i=1}^m \subset X \times Y$ ($m$ 是抽样规模) 来估计样本测度 $\rho$,从而推断出 $f^*(X)$. 最小二乘损失函数为 $l(u) = u^2$, $u \in R$,对任意可测函数 $f(x)$ 的泛化误差定义为:

$$\varepsilon(f) = \int_{X \times Y} l(y - f(x)) \mathrm{d}\rho.$$

对应的经验泛化误差为:

$$\varepsilon_z(f) = \frac{1}{m} \sum_{i=1}^m l(y_i - f(x_i)).$$

给定假设函数空间 $H$,最小二乘算法定义为:

$$f_z := \arg \min_{f \in H} \varepsilon_z(f). \quad (1)$$

在本文中,把再生核希尔伯特空间(RKHS) $H_K$[1] 作为假设空间,这里 $K(\cdot, \cdot): X \times X \to R$ 是 Mercer 核,即连续、对称、半正定且:

$$H_K = \overline{span}\{K_x(\cdot) = K(x, \cdot), x \in X\},$$

其中 $\langle \cdot, \cdot \rangle_K$ 是指 $H_K$ 中的内积. 该空间最大特点是再生性质,也就是说对于任意 $f \in H_K$, $x \in X$,有 $f(x) = \langle f, K_x \rangle_K$. 为防止过分拟合,采取正则化方法,对算法 (1) 给定一个正则项 $\lambda \Omega(f)$, $\lambda > 0$ 是正则化参数,

**收稿日期** 2017-09-02
**作者简介** 汪宝彬(1976-),男,副教授,博士,研究方向: 概率统计,E-mail: wbb1818@126.com
**基金项目** 国家自然科学基金资助项目(11671307); 湖北省自然科学基金资助项目(2017CFB523)

$\Omega(f)$ 是正则项,$\Omega(f)$ 根据不同算法的需求设计. 例如,常见 $\Omega(f) = \|f\|_K^2$,我们得到:

$$f_{z,\lambda}: = \arg\min_{f\in H}\varepsilon_z(f) + \lambda\|f\|_K^2.$$

根据表示定理,所学到的函数 $f_{z,\lambda}$ 可以表示为核函数 $K_x$ 的线性组合,即:$f(x) = \sum_{i=1}^{m} C_i K_{X_i}$,其中$\{c_i\}$ 是一个长度为 $m$ 实数序列. 核方法的优势在于:一是将高维空间数据转化为低维空间,保留输入空间的主要特性,同时降低计算复杂度和存储空间;二是在 $H_K$ 空间中的无限维问题能够转化为 $m$ 维空间问题,即为一个求解系数$\{c_i\}$ 问题. 优化问题中的凸函数和凸规划理论能够进行算法(1) 的求解和分析. 对于学到函数 $f_{z,\lambda}$ 与真实目标函数 $f^*$ 之间的关系,在文献[2-7]中已有详细的研究. 在噪声条件和假设空间满足一定条件下,算法(1) 的学习速率最优可以达到 $\|f_{z,\lambda} - f^*\|_{L^2_{\rho_X}} = O(m^{-1})$,这里 $L^2_{\rho_X}$ 是指定义在$(X,\rho_X)$ 上的平方可积空间.

## 1 建立算法

近年来,多罚研究[8] 引起统计学界的广泛关注,其在病态问题中构造稳健解的优势被工业界看重. 它能够结合条件概率中的先验测度信息和罚函数项,进行合适的正则参数选择,达到学习的最优效果. 在学习理论的背景下,Berklin 等人[9] 提出多罚正则化方法,对条件概率中的几何结构进行分析,从理论和实验对该方法进行分析;石等人[10] 通过积分算子理论,在分布式学习背景中对多罚方法的数学理论有更进一步的研究. 本文主要考虑如下算法:

$$f_{z,\mu}: = \arg\min_{f\in H_K}\varepsilon(f) + \mu\|Tf\|_K^2, \qquad (2)$$

这里 $T$ 是定义在 $H_K$ 中的有界线性算子,$\mu > 0$ 是正则参数. 与以往研究不同,本文仅考虑 $\|Tf\|_K^2$ 项,对输入空间 $X$ 结构和半监督数据进行研究.

定义算子 $L_{K,z}: H_K \to H_K$,$L_{K,z}(f) = \frac{1}{m}\sum_{i=1}^{m}\langle f, K_{x_i}\rangle_K K_{x_i}$,$\forall f \in H_K$,则以下性质1 成立.

性质1 对于任意 $\mu > 0$,算法(2) 有唯一解如下:

$$f_{z,\mu} = (L_{K,z} + \mu T^*T)^{-1}\bar{f}_{\rho,z}, \text{ 其中 } \bar{f}_{\rho,z}: = \frac{1}{m}\sum_{i=1}^{m} y_i K_{x_i}, T^* \text{ 是 } T \text{ 的共轭算子.}$$

自然地,定义算法(2) 的积分形式:

$$f_\mu: = \arg\min_{f\in H_K}\varepsilon(f) + \mu\|Tf\|_K^2, \qquad (3)$$

显然,$f_\mu = (L_K + \mu T^*T)^{-1}L_K(f^*)$. 这里,积分算子 $L_K: H_K \to H_K$ 定义为:

$$L_K(f) = \int_X \langle f, K_X\rangle K_X \mathrm{d}\rho_X, \forall f \in H_K.$$

评注:$\|Tf\|_K^2$ 在流形学习中常见形式为 $\|Tf\|_K^2 = \int_{x\in X}\|\nabla_M f\|^2\mathrm{d}\rho_X$,这里 $\nabla_M$ 是 $f$ 在流形 $M \subset R^n$ 上的梯度. $\|Tf\|_K^2$ 能够利用监督数据 $z = \{(x_i,y_i)\}_{i=1}^m$ 和无监督数据 $z' = \{x_j\}_{j=1}^n$ 基于图的拉普拉斯算子逼近. 在流形和半监督学习背景下,得到算法(2) 的明确表达形式为:

$$f_{z,z',\mu}: = \arg\min_{f\in H_K}\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 + \frac{\mu}{(m+n)^2}\sum_{i,j=1}^{m+n}(f(x_i) - f(x_j))^2 W_{i,j} = \arg\min_{f\in H_K}\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 + \frac{\mu}{(m+n)^2}F^{\mathrm{T}}LF,$$

这里 $W_{i,j}$ 是 graph 中各边的权重,$L$ 是图的拉普拉斯算子且 $L = D - W$,这里对角矩阵 $D_{i,j} = \sum_{i,j=1}^{m+n} W_{i,j}$,向量 $F = (f(x_1),\cdots,f(x_{m+n}))^{\mathrm{T}}$.

## 2 主要结果及证明

本文主要研究算法(2) 的学习速度,尤其是在参数 $\mu$ 的选择是否能使算法(2) 的学习能力达到最优. 通常做法是将算法(2) 中 $f_{z,\mu} - f^*$ 分解成抽样误差 $f_{z,\mu} - f_\mu$ 和逼近误差 $f_\mu - f^*$ 两部分. 首先,估计逼近误差 $f_\mu - f^*$. 不失一般性,假设 $B_1 = \inf_{f\in H_K}\frac{\|Tf\|_K}{\|f\|_K}$ 和 $B_2 = \sup_{f\in H_K}\frac{\|Tf\|_K}{\|f\|_K}$. 以下如不做特别声明,$\|\cdot\|$ 是指 $H_K$ 空间中的范数,$\sup_{x\in X}K(x,x): = 1$.

定理1 如果 $\mu > 0$,且存在一个函数 $g \in H_K$ 和 $0 < r < \frac{1}{2}$ 使得 $f^* = L_K^r g$,则:

$$\|f_\mu - f^*\| = O(\mu^r).$$

证明 根据表达式(3),有:

$$f_\mu - f^* = (L_K + \mu T^*T)^{-1}L_K f^* - f^* = -\mu T^*T(L_K + \mu T^*T)^{-1}f^*.$$

因此:

$$\|f_\mu - f^*\| = \|-\mu T^*T(L_K + \mu T^*T)^{-1}f^*\| = \|\mu T^*T(L_K + \mu T^*T)^{-1}L_K^r g\| < \mu B_2^2\|(L_K + \mu T^*T)^{-1}L_K^r g\| < \mu B_2^2\sup_{x>0}(x + \mu B_1^2)^{-1}x^r = c\mu^r,$$

这里常数 $c = B_2^2 B_1^{2r-2}r^r(1-r)^{-r}[(1-r)^{-1} + 1]^{-1}\|g\|$.

现在估计的重点在于抽样误差 $f_{z,\mu} - f_\mu$,首先要用

到以下引理 1.

**引理 1**　假设 $\zeta_i, i = 1, \cdots, m$ 是定义在希尔伯特空间 $(H, \|\cdot\|)$ 上一列独立同分布的随机变量,并且存在常数 $S > 0$,使得 $\|\zeta_i\| \leqslant S, i = 1, \cdots, m$,则对任意 $0 < \delta < 1$,有:

$$P\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\xi_i - E(\zeta)\right\| \geqslant \frac{4S}{\sqrt{m}}\log\left(\frac{2}{\delta}\right)\right\} \leqslant 1 - \delta.$$

于是,有定理 2.

**定理 2**　假设存在 $S > 0$, 有 $|y| \leqslant S$,则对任意 $0 < \delta < 1$,有:

$$P\left\{\|f_{z,\mu} - f_\mu\| \geqslant \frac{4B_1^2(S + \|f_\mu\|)}{\mu\sqrt{m}}\log\left(\frac{4}{\delta}\right)\right\} \leqslant 1 - \delta.$$

**证明**　从算法 (2) 的表达式得到:

$$f_{z,\mu} - f_\mu = (L_{K,z} + \mu T^*T)^{-1}\bar{f}_{\rho,z} - f^* =$$
$$(L_{K,z} + \mu T^*T)^{-1}(\bar{f}_{\rho,z} - (L_{K,z} + \mu T^*T)f^*) =$$
$$(L_{K,z} + \mu T^*T)^{-1}[(\bar{f}_{\rho,z} - L_Kf^*) - (L_K - L_{K,z})f_\mu],$$

从而,

$$\|f_{z,\mu} - f_\mu\| \leqslant$$
$$\|(L_{K,z} + \mu T^*T)^{-1}[(\bar{f}_{\rho,z} - L_Kf^*) - (L_K - L_{K,z})f_\mu]\| \leqslant$$
$$\mu^{-1}B_1^{-2}[\|\bar{f}_{\rho,z} - L_Kf^*\| + \|(L_K - L_{K,z})f_\mu\|] \leqslant$$
$$\mu^{-1}B_1^{-2}[\|\bar{f}_{\rho,z} - L_Kf^*\| + \|L_K - L_{K,z}\|\|f_\mu\|].$$

对于第一项 $\|\bar{f}_{\rho,z} - L_Kf^*\|$,令 $\zeta_i = y_iK_{x_i}$,则 $\|yK_x\| \leqslant S$. 利用引理 1,有:

$$P\left\{\|\bar{f}_{\rho,z} - L_Kf^*\| \geqslant \frac{4S}{\sqrt{m}}\log\left(\frac{2}{\delta}\right)\right\} \leqslant 1 - \delta.$$

对于第二项 $\|L_K - L_{K,z}\|$,令 $\zeta_i = \langle\cdot, K_{x_i}\rangle_K K_{x_i}$,则 $\|\langle\cdot, K_{x_i}\rangle_K K_{x_i}\| \leqslant 1, i = 1, \cdots, m$. 由引理 1 有:

$$P\left\{\|L_K - L_{K,z}\| \geqslant \frac{4}{\sqrt{m}}\log\left(\frac{2}{\delta}\right)\right\} \leqslant 1 - \delta.$$

结合第一项和第二项的估计,并将 $\delta$ 替换成 $\delta/2$,得到在至少 $1 - \delta$ 的概率下,有下式成立:

$$\|f_{z,\mu} - f_\mu\| \leqslant \frac{4B_1^2(S + \|f_\mu\|)}{\mu\sqrt{m}}\log\left(\frac{4}{\delta}\right).$$

定理 2 证毕.

现在已有抽样误差和逼近误差两项估计,由算法 (2) 的学习速率很容易得到定理 3.

**定理 3**　若 $|y| \leqslant S$ 且存在一个函数 $g \in H_K$ 和 $0 < r < \frac{1}{2}$,使得 $f^* = L_K^r g$,在至少 $1 - \delta$ 的概率下,我们有:

$$\|f_{z,\mu} - f^*\| \leqslant c'\left(\frac{1}{\mu\sqrt{m}} + \mu^r\right)\log\left(\frac{4}{\delta}\right). \quad (4)$$

这里 $c'$ 是一个不依赖于 $m, \mu, \delta$ 的常数,它会在证明过程中给出. 进一步,令 $\mu = m^{-\frac{1}{2(1+r)}}$,在至少 $1 - \delta$ 的概率下,有:

$$\|f_{z,\mu} - f^*\| = O(m^{-\frac{r}{2(1+r)}})\log\left(\frac{4}{\delta}\right). \quad (5)$$

**证明**　根据 (3) 式知道, $\|f_\mu\| = \|(L_K + \mu T^*T)^{-1}L_K(f^*)\| \leqslant \|f^*\|$. 进一步,利用定理 1 和定理 2,取常数 $c' = 4B_1^2(S + \|f^*\|) + B_2^2B_1^{2r-2}r^r(1 - r)^{-r}[(1 - r)^{-1} + 1]^{-1}\|g\|$,结论 (4) 即可得证; 结论 (5) 很容易从结论 (4) 推出,证毕.

## 参 考 文 献


[1] Micchelli C A, Xu Y S, Zhang H Z. Universal kernels [J]. Journal of Machine Learning Research, 2006, 7(4): 2651-2667.

[2] Evgeniou T, Pontil M, Poggio T. Regularization networks and support vector machines [J]. Advances in Computational Mathematics, 2000, 13: 1-53.

[3] Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: risk bounds and structural results [J]. Journal of Machine Learning Research, 2002, 3: 463-482.

[4] Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization [J]. Annals of Statistics, 2004, 32: 56-85.

[5] Bartlett P L, Jordan M I, McAuliffe J D. Convexity, classification, and risk bounds [J]. Journal of the American Statistical Association, 2006, 101: 138-156.

[6] Zhou D X. Capacity of reproducing kernel spaces in learning theory [J]. IEEE Transactions on Information Theory, 2003, 49: 1743-1752.

[7] Shi L, Feng Y L, Zhou D X. Concentration estimates for learning with $l^1$-regularizer and data dependent hypothesis spaces [J]. Applied and Computational Harmonic Analysis, 2011, 31: 286-302.

[8] Abhishake, Sivananthan S. Multi-penlty regularization in learning theory [J]. Journal of Complexity, 2016, 36(C): 141-165.

[9] Berklin M, Niyogi P, Sindhwan V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006, 7(1): 2399-2434.

[10] Shi L, Guo Z C, Lin S B. Distributed learning with multi-penalty regularization [J]. Applied and Computational Harmonic Analysis, DOI: 10.1016/j.acha.2017.06.001.