

拟概率空间上学习理论的关键定理 和学习过程一致收敛速度的界

哈明虎¹⁾ 冯志芳²⁾ 宋士吉³⁾ 高林庆¹⁾

¹⁾(河北大学数学与计算机学院 河北 保定 071002)

²⁾(廊坊师范学院数学与信息科学系 河北 廊坊 065000)

³⁾(清华大学自动化系 北京 100084)

摘 要 进一步讨论了拟概率的一些性质,给出了拟概率空间上的拟随机变量及其分布函数、期望和方差的概念及若干性质;证明了拟概率空间上的 Markov 不等式、Chebyshev 不等式和 Khinchine 大数定律;给出并证明了拟概率空间上学习理论的关键定理和学习过程一致收敛速度的界,把概率空间上的学习理论的关键定理和学习过程一致收敛速度的界推广到了拟概率空间,为系统地建立拟概率上的统计学习理论与构建支持向量机奠定了理论基础。

关键词 拟概率;期望风险泛函;经验风险泛函;关键定理;一致收敛速度的界
中图法分类号 TP18

The Key Theorem and the Bounds on the Rate of Uniform Convergence of Statistical Learning Theory on Quasi-Probability Spaces

HA Ming-Hu¹⁾ FENG Zhi-Fang²⁾ SONG Shi-Ji³⁾ GAO Lin-Qing¹⁾

¹⁾(College of Mathematics and Computer Sciences, Hebei University, Baoding, Hebei 071002)

²⁾(Department of Mathematics and Information Sciences, Langfang Normal School, Langfang, Hebei 065000)

³⁾(Department of Automation, Tsinghua University, Beijing 100084)

Abstract Some properties of quasi-probability are further discussed. The definitions and properties of quasi-random variable and its distribution function, expected value and variance are then presented. Markov inequality, Chebyshev's inequality and the Khinchine's law of large numbers on quasi-probability spaces are also proved. Then the key theorem of learning theory on quasi-probability spaces is proved, and the bounds on the rate of uniform convergence of learning process on quasi-probability spaces are constructed. The investigations will help lay essential theoretical foundations for the systematic and comprehensive development of the quasi-statistical learning theory.

Keywords quasi-probability; empirical risk functional; expected risk functional; key theorem; bounds on the rate of uniform convergence

1 引 言

统计学习理论 (Statistical Learning Theory,

SLT) 是 20 世纪 60 年代末由 Vapnik 等人^[1-3]提出并逐步于 20 世纪 90 年代中期建立的一种专门在小样本情况下研究统计学习规律的理论,它为研究有限样本情况下机器学习的理论和方法提供了重要理论

收稿日期:2006-08-06;最终修改稿收到日期:2007-12-07. 本课题得到国家自然科学基金(60573069,60574077,60773062)、教育部科学技术研究重点项目计划(206012)、河北省自然科学基金(F2004000129)和河北省教育厅科研计划重点项目(2005001D)资助. 哈明虎,男,1963 年生,博士,教授,博士生导师,研究领域为不确定信息处理、非可加测度理论与统计学习理论. E-mail: mhha@mail.hbu.edu.cn. 冯志芳,女,1980 年生,硕士,助教,研究方向为统计学习理论. 宋士吉,男,1965 年生,博士,教授,博士生导师,研究领域为不确定系统模型与优化计算、软计算理论及应用. 高林庆,男,1979 年生,硕士研究生,助教,研究方向为统计学习理论.

框架,其核心思想是通过控制学习机器容量的控制来研究学习机器的推广能力.由于SLT为人们系统地研究小样本情况下机器学习问题提供了坚实的理论基础,且从这一理论上发展出的支持向量机(Support Vector Machine, SVM)是一种新的通用学习机器方法,因此,较之以往的机器学习方法,SLT在理论研究和实际应用中均表现出了很大的优势.近40年来,国内外众多学者密切关注和参与这一领域的研究,不仅在理论上取得了一系列丰硕的成果^[1-9],而且在数据分析、生物信息技术、图像图形处理与经济预测等领域中取得了广泛而又成功的应用^[8-17].目前SLT已被学术界公认为机器学习领域一个新的研究热点^[3-7].

SLT主要由4部分内容组成^[1,4]: (1) 学习过程一致性的条件(学习理论的关键定理); (2) 学习过程一致收敛速度的界及基于VC维的推广性的界; (3) 在这些界基础之上的结构风险最小化原则; (4) 实现新的结构风险最小化原则的实际算法——支持向量机.前3部分内容是统计学习理论的基础理论部分,第4部分内容是基于统计学习理论的基础理论发展出的方法及应用部分.本文我们主要讨论学习理论中的关键定理和学习过程一致收敛速度的界.关键定理将SLT中的经验风险最小化的严格一致性问题,转化为均值一致单边收敛于数学期望的存在性问题,换句话说,转化为某一经验过程的收敛问题^[1].因此,对于给定的数据,我们可以通过判断均值是否一致单边收敛于数学期望来判断经验风险最小化原则是否是严格一致的.这样我们就可以用数学分析理论来判断严格一致性的问题.学习过程一致收敛速度的界确定了采用经验风险最小化原则的学习机器的推广能力,它是分析学习机器性能和发展新的学习算法的重要基础.通过对界的估计,可以得到经验风险最小化原则(ERM)中经验风险与实际风险之间的关系,进而可以研究学习机器的推广能力.

尽管统计学习理论被学术界公认为是目前处理小样本学习问题的重要理论,但它仍存在一些需要完善之处^[4,7].如统计学习理论是建立在概率(一类特殊的测度,也可称概率测度)空间上的,且此空间上的概率是一个满足可加性(可列可加性)的非负集函数.由于可加性条件比较苛刻,在实际应用中这个条件往往得不到满足,例如:在现实生活中有时要求人们从价格、品牌和大小相同的若干处理电视机中挑选出自己满意的一台,为此,需要对每台电视

机进行评判.为简单起见,我们只考虑两个主要因素,即图像、音响.设论域 $U = \{\text{图像}(U_1), \text{音响}(U_2)\}$,其中 $\{U_i\} (i=1, 2)$ 为两两不相交的集合, $P(U)$ 表示 U 的幂集.由于人的主观因素的影响,不妨规定主要因素的重要性度量 μ (μ 是从 $P(U)$ 到 $[0, 1]$ 的一个集函数)如下:

$$\mu(\emptyset) = 0, \mu(U_1) = 0.80,$$

$$\mu(U_2) = 0.60, \mu(U_1 \cup U_2) = 1.$$

根据此重要性度量,利用Choquet积分^[18]或Sugeno积分^[19],我们可以给出每台电视机的综合评判值,进而根据综合评判值的大小调选出电视机^[20].但显然有

$$\mu(U_1 \cup U_2) \neq \mu(U_1) + \mu(U_2),$$

故 μ 不是一个概率测度.该简单例子说明了非可加集函数(非可加测度)在现实生活中是客观存在的.因此,对非可加测度的研究就成为很有意义的工作.1954年,Choquet提出的一种称之为容度的理论^[18],可认为是最早的非可加测度研究.到目前为止,比较有代表性的非可加测度有:模糊测度、Sugeno测度、可能性测度、拟概率和可信性测度等^[19-23].经过50余年的完善和发展,非可加测度及积分不仅在理论上取得了系列重要成果,成为数学分析的重要研究方向,而且在信息科学、工程技术、经济管理等领域取得了广泛应用^[18-29],这也意味着把统计学习理论从概率空间推广到非可加测度空间既有理论意义又有广泛应用前景.哈明虎等人^[7]于2006年给出了一类比概率测度更广的有代表性的非可加测度——Sugeno测度空间上的学习理论的关键定理和一致收敛速度的界,这标志着非可加测度空间上统计学习理论研究的开始.

由Wang提出的拟概率^[20,22]是概率和Sugeno测度的重要推广.它也是客观存在的,如进一步讨论上述的例子我们会发现,如果给定一个特殊的函数:

$$\theta_T(x) = -\frac{\ln\left(1 - \frac{5}{6}x\right)}{\ln 6},$$

则

$$\theta \circ \mu(U_1 \cup U_2) = \theta \circ \mu(U_1) + \theta \circ \mu(U_2) = 1,$$

即 $\theta \circ \mu$ 为定义在 U 上的经典概率测度,亦即 μ 为以 θ_T 为其正规 T -函数的拟概率.从而开展拟概率空间上的统计学习理论的研究也是很有意义的,基于此,本文将在拟概率空间上率先讨论统计学习理论.

本文第2节引进了拟概率的定义,进一步给出了拟概率的一些性质,讨论了拟概率空间上的拟随

机变量及其数字特征,并证明了拟概率空间上的 Markov 不等式、Chebyshev 不等式和 Khinchine 大数定律;第 3 节给出并证明了拟概率空间上统计学习理论的关键性定理;第 4 节讨论了学习过程一致收敛速度的界;最后一节给出了本文的结论与展望.

2 预备知识

为了方便和完整起见,本节将给出本文用到的一些基本概念和性质. 本文假定 X 是一个非空集合, \mathcal{F} 是由 X 的子集构成的 σ -代数.

定义 1^[20]. 设 $a \in (0, \infty]$, 广义实值函数 $\theta: [0, a] \rightarrow [0, \infty]$ 称为 T -函数, 当且仅当 θ 是连续严格增加的, 且 $\theta(0) = 0, \theta^{-1}(\{\infty\}) = \emptyset$ 或 $\{\infty\}$ (根据 a 是否有限而定).

显然, 如果 θ 是连续、严格增加的集函数, 那么 θ^{-1} 也是连续、严格增加的集函数.

定义 2^[20]. 集函数 $\mu: \mathcal{F} \rightarrow [0, \infty]$ 称为拟可加的, 当且仅当存在定义域包含 μ 的值域的 T -函数 θ , 使定义在 \mathcal{F} 上的集函数 $\theta \circ \mu$ 是可加的, $\theta \circ \mu$ 通过如下方式运算:

$$(\theta \circ \mu)(E) = \theta(\mu(E)), \text{ 对任意 } E \in \mathcal{F}.$$

μ 称为拟测度, 当且仅当存在 T -函数 θ , 使得 $\theta \circ \mu$ 为一个经典测度. T -函数 θ 称为 μ 的特征 T -函数.

定义 3^[20]. 函数 $\theta: [0, 1] \rightarrow [0, 1]$ 称为正规 T -函数, 当且仅当 θ 是连续、严格增加的, 且 $\theta(0) = 0, \theta(1) = 1$.

若 θ 是正规 T -函数, 则拟测度 μ 称为拟概率, 且三元组 (X, \mathcal{F}, μ) 称为拟概率空间. 本文以下的讨论如无特殊说明均在此拟概率空间上进行.

注 1. 若取 $\theta(x) = x$ 作为其正规 T -函数, 则概率是一种特殊的拟概率; 若取 $\theta(x) = \log_{1+\lambda}(1+\lambda x)$ 作为其正规 T -函数^[20], 则 Sugeno 测度也是一种特殊的拟概率. 相应地, 拟概率空间 (X, \mathcal{F}, μ) 是一种比概率空间和 Sugeno 测度空间更广的一种非可加测度空间.

由拟概率的定义不难得到下面的性质.

性质 1. 设 (X, \mathcal{F}, μ) 为拟概率空间, 则我们有

- (1) $\mu(\emptyset) = 0, \mu(X) = 1$;
- (2) 若 $A, B \in \mathcal{F}, A \subseteq B$, 则 $\mu(A) \leq \mu(B)$;
- (3) 若 $A \in \mathcal{F}$, 则 $\mu(A^c) = \theta^{-1}[1 - \theta \circ \mu(A)]$;
- (4) 若 $A \in \mathcal{F}, B \subset A$, 则 $\mu(A - B) = \theta^{-1}[\theta \circ \mu(A) - \theta \circ \mu(B)]$;
- (5) 若 $A, B \in \mathcal{F}$, 则 $\mu(A \cup B) = \theta^{-1}[\theta \circ \mu(A) +$

$$\theta \circ \mu(B) - \theta \circ \mu(A \cap B)].$$

由于拟概率空间上的随机变量及其分布函数、期望、方差、事件的发生、同分布、 n 维随机变量、 n 维随机变量联合分布函数、边缘分布函数及密度函数等定义分别与概率空间和 Sugeno 测度空间上对应的定义形式(除独立性的定义稍有不同外)是一致的^[1,7], 这里定义形式的一致性是指把定义在概率空间和 Sugeno 测度空间上对应定义的概率和 Sugeno 测度置换为拟概率. 因此我们只列出拟概率空间上的随机变量及其相互独立的定义, 其它定义就不一一列出了.

定义 4. 设 $\xi = \xi(\omega), \omega \in X$ 为定义在 X 上的实值集函数. 对任意给定的实数 x , 若 $\{\omega | \xi(\omega) \leq x\} \in \mathcal{F}$, 则 ξ 称为拟概率空间上的随机变量. 简记为 q -随机变量.

定义 5. 设 $F_\mu(x, y), F_{\mu\xi}(x)$ 和 $F_{\mu\eta}(y)$ 分别是 (ξ, η) 的联合分布函数和 ξ 及 η 的边缘分布函数, 若对任意 x, y 有

$$F_\mu(x, y) = \theta^{-1}[\theta(F_{\mu\xi}(x)) \cdot \theta(F_{\mu\eta}(y))]$$

成立, 则称 ξ 和 η 是相互独立的 q -随机变量.

根据上述定义我们可以给出如下性质.

性质 2. 设 ξ 是一个离散 q -随机变量. 若正规 T -函数为 $\theta, x_1, x_2, \dots, x_n, \dots$ 是 ξ 的所有可能值, 则

$$\theta^{-1}\left(\sum_{i=1}^{\infty} \theta \circ \mu(\{\omega | \xi(\omega) = x_i\})\right) = 1.$$

性质 3. 设 $F_\mu(x)$ 是 q -随机变量 ξ 的分布函数, 则

- (1) 若 $a < b$, 则 $F_\mu(a) < F_\mu(b)$;
- (2) $\lim_{x \rightarrow -\infty} F_\mu(x) = 0, \lim_{x \rightarrow \infty} F_\mu(x) = 1$;
- (3) $F_\mu(x+0) = F_\mu(x)$.

性质 4. 若 $x_1, x_2 \in R, \xi$ 为 q -随机变量, 则 $\mu(\{x_1 < \xi \leq x_2\}) = \theta^{-1}[\theta[F_\mu(x_2)] - \theta[F_\mu(x_1)]]$.

注 2. 当 $\theta(x) = x$ 时, 拟概率 μ 的性质与概率的性质是相同的; 当 $\theta(x) = \log_{1+\lambda}(1+\lambda x)$ 时, 则 $\theta^{-1}(x) = \frac{(1+\lambda)^x - 1}{\lambda}$, 拟概率 μ 的性质与 Sugeno 测度的性质是相同的.

性质 5. 设 $\{\xi_n\}$ 是一个 q -随机变量序列, 且它们的期望存在, 则下面结论成立:

- (1) 对任意 $\lambda, E_\mu[\lambda\xi] = \lambda E_\mu[\xi]$;
- (2) 对任意 $m \geq 1, E_\mu\left[\sum_{i=1}^m \xi_i\right] = \sum_{i=1}^m E_\mu[\xi_i]$;
- (3) 若 $\xi \geq 0$, 则 $E_\mu[\xi] \geq 0$;
- (4) 若 ξ_1, ξ_2 是独立同分布的 q -随机变量, 且

$\xi_1 \leq \xi_2$, 则 $E_\mu[\xi_1] \leq E_\mu[\xi_2]$.

性质 6. 设 $\xi_1, \xi_2, \dots, \xi_n$ 是 n 个相互独立的非负 q -随机变量. 若 θ 和 θ^{-1} 在 $[0, 1]$ 上的 n 阶导数存在且连续, 则下列不等式成立:

$$E_\mu[\xi_1, \xi_2, \dots, \xi_n] \leq C_{n\theta} E_\mu[\xi_1] E_\mu[\xi_2] \cdots E_\mu[\xi_n],$$

其中 $C_{n\theta}$ 为常数.

证明. 下面仅就其 q -随机变量是连续的情形进行证明.

我们先证明二维的情况.

因为

$$\begin{aligned} F_\mu(x, y) &= \theta^{-1} \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \}, \\ f_\mu(x, y) &= \frac{\partial F_\mu(x, y)}{\partial x \partial y} \\ &= \{ \partial \{ (\theta^{-1})' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \} \cdot \\ &\quad \theta'[F_{\mu\xi_1}(x)] F'_{\mu\xi_1}(x) \theta[F_{\mu\xi_2}(y)] \} \} / \partial y \\ &= (\theta^{-1})'' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \cdot \\ &\quad \theta'[F_{\mu\xi_2}(y)] F'_{\mu\xi_2}(y) \theta[F_{\mu\xi_1}(x)] \cdot \\ &\quad \theta'[F_{\mu\xi_1}(x)] F'_{\mu\xi_1}(x) \theta[F_{\mu\xi_2}(y)] + \\ &\quad (\theta^{-1})' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \cdot \\ &\quad \theta'[F_{\mu\xi_1}(x)] F'_{\mu\xi_1}(x) \theta'[F_{\mu\xi_2}(y)] F'_{\mu\xi_2}(y) \\ &= \{ (\theta^{-1})'' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \cdot \\ &\quad \theta'[F_{\mu\xi_2}(y)] \theta[F_{\mu\xi_1}(x)] \theta'[F_{\mu\xi_1}(x)] \cdot \\ &\quad \theta[F_{\mu\xi_2}(y)] + (\theta^{-1})' \{ \theta[F_{\mu\xi_1}(x)] \cdot \\ &\quad \theta[F_{\mu\xi_2}(y)] \} \cdot \theta'[F_{\mu\xi_1}(x)] \theta'[F_{\mu\xi_2}(y)] \} \\ &\quad F'_{\mu\xi_1}(x) F'_{\mu\xi_2}(y). \end{aligned}$$

因为 $F_{\mu\xi_1}(x)$ 和 $F_{\mu\xi_2}(y)$ 的值域是 $[0, 1]$, θ 和 θ^{-1} 是 $[0, 1]$ 上的连续函数, θ^{-1} 在 $[0, 1]$ 上的二阶导数是连续的, 所以

$$\begin{aligned} &(\theta^{-1})'' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \cdot \\ &\quad \theta'[F_{\mu\xi_2}(y)] \theta[F_{\mu\xi_1}(x)] \theta'[F_{\mu\xi_1}(x)] \theta[F_{\mu\xi_2}(y)] + \\ &(\theta^{-1})' \{ \theta[F_{\mu\xi_1}(x)] \cdot \theta[F_{\mu\xi_2}(y)] \} \cdot \\ &\quad \theta'[F_{\mu\xi_1}(x)] \theta'[F_{\mu\xi_2}(y)] \end{aligned}$$

在 $[0, 1]$ 上可以取到它的最大值 $C_{2\theta}$. 从而

$$f_\mu(x, y) = \frac{\partial F_\mu(x, y)}{\partial x \partial y} \leq C_{2\theta} F'_{\mu\xi_1}(x) F'_{\mu\xi_2}(y),$$

因为 q -随机变量 ξ_1, ξ_2 是非负的, 则

$$\begin{aligned} E_\mu(\xi_1 \xi_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_\mu(x, y) dx dy \\ &= \int_0^{\infty} \int_0^{\infty} xy f_\mu(x, y) dx dy \\ &\leq C_{2\theta} \int_0^{\infty} \int_0^{\infty} xy F'_{\mu\xi_1}(x) \cdot F'_{\mu\xi_2}(y) dx dy \\ &= C_{2\theta} \int_0^{\infty} x dF_{\mu\xi_1}(x) \cdot \int_0^{\infty} y dF_{\mu\xi_2}(y) \end{aligned}$$

$$\begin{aligned} &= C_{2\theta} \int_{-\infty}^{\infty} x dF_{\mu\xi_1}(x) \cdot \int_{-\infty}^{\infty} y dF_{\mu\xi_2}(y) \\ &= C_{2\theta} E_\mu(\xi_1) E_\mu(\xi_2), \end{aligned}$$

同理对于 n 维的情况, 存在常数 $C_{n\theta}$, 使得下列不等式成立

$$E_\mu[\xi_1, \xi_2, \dots, \xi_n] \leq C_{n\theta} E_\mu[\xi_1] E_\mu[\xi_2] \cdots E_\mu[\xi_n]$$

证毕.

对于 q -随机变量是离散的情况, 利用微分中值定理可类似给出证明.

性质 7. 假设 ξ 和 η 是两个独立同分布的 q -随机变量且具有有限的期望值, 对任意有限的实数 λ , 有

$$(1) D_\mu[\lambda\xi] = \lambda^2 D_\mu[\xi];$$

$$(2) D_\mu[\xi + \eta] \leq D_\mu[\xi] + D_\mu[\eta].$$

为了在拟概率空间上讨论统计学习理论的关键定理, 下面我们给出并证明拟概率空间上的 Markov 不等式、Chebyshev 不等式和 Khinchine 大数定律.

定理 1 (Markov 不等式). 若 ξ 是一个非负 q -随机变量, 任意 $t > 0$, 则下面不等式成立:

$$\mu\{\xi \geq t\} \leq \theta^{-1} \left[1 - \theta \left(1 - \frac{E_\mu[\xi]}{t} \right) \right].$$

证明. 由定义知 $\theta \circ \mu$ 是概率, 则

$$\begin{aligned} \mu(\{\xi \geq t\}) &= \theta^{-1} [\theta \circ \mu(\{\xi < \infty\}) - \theta \circ \mu(\{\xi \leq t\})] \\ &= \theta^{-1} \left[1 - \theta \left(\int_{-\infty}^t dF_\mu(x) \right) \right] \\ &= \theta^{-1} \left[1 - \theta \left(1 - \int_t^{\infty} dF_\mu(x) \right) \right] \\ &\leq \theta^{-1} \left[1 - \theta \left(1 - \frac{E_\mu[\xi]}{t} \right) \right] \end{aligned}$$

故定理成立.

证毕.

注 3. 在定理 1 中当 θ 取 $\theta(x) = x$ 时, 定理变成

$$\mu\{\xi \geq t\} \leq \frac{E_\mu[\xi]}{t} \text{ 与概率空间上的 Markov 不等式相同; 当 } \theta \text{ 取 } \theta(x) = \log_{1+\lambda}(1+\lambda x), \theta^{-1} \text{ 取 } \theta^{-1}(x) = \frac{(1+\lambda)^x - 1}{\lambda} \text{ 时不等式变为 } \mu\{\xi \geq t\} \leq \frac{E_\mu[\xi]}{t + \lambda t - \frac{\lambda E_\mu[\xi]}{t}},$$

当 $\lambda \leq 0$ 可变为 $\mu\{\xi \geq t\} \leq \frac{E_\mu[\xi]}{t + \lambda t}$, 当 $\lambda > 0$ 变为 $\mu\{\xi \geq$

$t\} \leq \frac{E_\mu[\xi]}{t}$ 与 Sugeno 测度空间上的 Markov 不等式相同, 因此, 此引理是概率空间与 Sugeno 测度空间上 Markov 不等式的推广.

定理 2 (Chebyshev 不等式). 设 ξ 是一个 q -随机变量, 且其期望 $E_\mu(\xi) = a$, 方差 $D_\mu(\xi) = \sigma^2$, 则对任意 $\epsilon > 0$, 下面不等式成立:

$$\mu\{|\xi - a| \geq \epsilon\} \leq \theta^{-1} \left[1 - \theta \left(1 - \frac{D_\mu[\xi]}{\epsilon^2} \right) \right].$$

证明. 对 $|\xi - a|^2$ 直接应用 Markov 不等式即得. 证毕.

定理 3(Khinchine 大数定律). 假设 $\xi_1, \xi_2, \dots, \xi_n$ 是独立同分布的 q -随机变量序列, 若 $\xi_n (n=1, 2, \dots)$ 具有相同的有穷期望值, $E_\mu[\xi_n] = a (n=1, 2, \dots)$, 则对任意 $\epsilon > 0$, 等式

$$\lim_{n \rightarrow \infty} \mu \left\{ \left| \frac{1}{n} \sum_{i=1}^n \xi_i - a \right| < \epsilon \right\} = 1$$

成立.

证明. 由概率空间和 Sugeno 测度空间上的 Khinchine 大数定律(参看文献[7])和定理 2 可知定理成立. 证毕.

3 拟概率空间上学习理论的关键定理

本节我们给出并证明拟概率空间上学习理论的关键定理, 首先我们给出拟概率空间上统计学习理论的一些基本概念.

定义 6. 设 $\xi_1, \xi_2, \dots, \xi_n$ 是一个 q -随机变量序列, a 是一个常数, 对任意 $\epsilon > 0$, 若

$$\lim_{n \rightarrow \infty} \mu(|\xi_n - a| > \epsilon) = 0,$$

则称 $\{\xi_n\}$ 依拟概率 μ 收敛于 a .

定义 7. 设 $F_\mu(x)$ 是 q -随机变量 ξ 的分布函数, z_1, z_2, \dots, z_n 是独立同分布样本. 我们引入集函数 $Q(z, \alpha), \alpha \in \Lambda$, 期望经验风险泛函和经验风险泛函定义如下:

$$R(\alpha) = \int Q(z, \alpha) dF_\mu(z) \quad (1)$$

$$R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \quad (2)$$

经验风险最小化原则. 我们用经验风险泛函

$$R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha), \alpha \in \Lambda$$

来代替期望风险泛函

$$R(\alpha) = \int Q(z, \alpha) dF_\mu(z), \alpha \in \Lambda,$$

假设风险泛函在 $Q(z, \alpha_0)$ 上取得最小值, 经验风险泛函在 $Q(z, \alpha_l)$ 上取得最小值. 我们将函数 $Q(z, \alpha_l)$ 作为函数 $Q(z, \alpha_0)$ 的一个近似. 这一原则称为拟概率空间上经验风险最小化原则(QERM).

定义 8. 对于函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和拟概率分布函数 $F_\mu(z)$, 如果对于集合 Λ 的非空子集 $\Lambda(c) = \{\alpha: \int Q(z, \alpha) dF_\mu(z) \geq c\}, c \in (-\infty, \infty)$ 和任意 $\epsilon > 0$ 使得

$$\lim_{l \rightarrow \infty} \mu \left\{ \left| \inf_{\alpha \in \Lambda(c)} R(\alpha) - \inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) \right| \geq \epsilon \right\} = 0 \quad (3)$$

成立, 则我们称经验风险最小化方法是严格一致的.

定义 9. 对于函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和拟概率分布函数 $F_\mu(z)$ 及任意 $\epsilon > 0$, 如果

$$\lim_{l \rightarrow \infty} \mu \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \epsilon \right\} = 0 \quad (4)$$

成立, 则我们称经验风险一致单边收敛于期望风险.

有了上述定义我们便可给出拟概率空间上学习理论的关键定理.

定理 4(关键定理). 假设存在常数 a 和 A , 使得对于函数集 $\{Q(z, \alpha), \alpha \in \Lambda\}$ 中的所有函数和所有分布函数 $F_\mu(z)$ 有下列不等式成立:

$$a \leq \int Q(z, \alpha) dF_\mu(z) \leq A, \alpha \in \Lambda,$$

则经验风险最小化方法是严格一致的充分必要条件是经验风险一致单边收敛于期望风险.

证明. 必要性. 设经验风险最小化方法在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上是严格一致的. 根据严格一致性的定义, 对于使函数集

$$\Lambda(c) = \{\alpha: \int Q(z, \alpha) dF_\mu(z) \geq c\}, c \in (-\infty, \infty)$$

为非空的任意 c , 下列依拟概率收敛性为真

$$\lim_{l \rightarrow \infty} \mu \left\{ \left| \inf_{\alpha \in \Lambda(c)} R(\alpha) - \inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) \right| > \epsilon \right\} = 0 \quad (5)$$

考虑有限序列 a_1, a_2, \dots, a_n , 满足

$$|a_{i+1} - a_i| < \frac{\epsilon}{2}, a_1 = a, a_n = A.$$

我们用 T_k 表示事件

$$\inf_{\alpha \in \Lambda(a_k)} \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) < \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF_\mu(z) - \frac{\epsilon}{2},$$

则利用式(5), 我们知道

$$\lim_{l \rightarrow \infty} \mu \{T_k\} = 0.$$

令

$$T = \bigcup_{k=1}^n T_k,$$

由性质 1 知

$$\mu \left\{ \bigcup_{k=1}^n T_k \right\} \leq \theta^{-1} \left[\sum_{k=1}^n \theta \circ \mu(T_k) \right].$$

由 θ 和 θ^{-1} 是连续函数且 $\theta(0) = 0$, 得

$$\lim_{l \rightarrow \infty} \theta^{-1} \left[\sum_{k=1}^n \theta \circ \mu(T_k) \right] = \theta^{-1} \left[\sum_{k=1}^n \theta(\lim_{l \rightarrow \infty} \mu(T_k)) \right] = 0,$$

所以

$$\lim_{l \rightarrow \infty} \mu \left\{ \bigcup_{k=1}^n T_k \right\} = 0.$$

我们用 M 表示事件

$$\sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \epsilon,$$

假设 M 出现, 则存在 $\alpha^* \in \Lambda$, 使

$$R(\alpha^*) - \varepsilon > R_{\text{emp}}(\alpha^*)$$

成立. 我们由 α^* 找到一个 k , 使得 $\alpha^* \in \Lambda(\alpha_k)$ 且下面不等式成立:

$$R(\alpha^*) - a_k < \frac{\varepsilon}{2}.$$

由选定的集合 $\Lambda(\alpha_k)$, 我们可知下面不等式

$$0 \leq R(\alpha^*) - \inf_{\alpha \in \Lambda(\alpha_k)} R(\alpha) < \frac{\varepsilon}{2}$$

成立. 事实上, 由 $\inf_{\alpha \in \Lambda(\alpha_k)} R(\alpha) \geq a_k$ 可得

$$0 \leq R(\alpha^*) - \inf_{\alpha \in \Lambda(\alpha_k)} R(\alpha) < R(\alpha^*) - a_k < \frac{\varepsilon}{2}.$$

所以, 对于选定的 α^* 和 $\Lambda(\alpha_k)$, 下面不等式成立:

$$\begin{aligned} \inf_{\alpha \in \Lambda(\alpha_k)} R(\alpha) - \frac{\varepsilon}{2} &> R(\alpha^*) - \varepsilon > R_{\text{emp}}(\alpha^*) \\ &\geq \inf_{\alpha \in \Lambda(\alpha_k)} R_{\text{emp}}(\alpha), \end{aligned}$$

也就是说, 如果事件 M 发生, 则事件 T_k 发生, 那么事件 T 也发生.

由拟概率的单调性可知 $\mu(M) \leq \mu(T)$ 成立, 所以

$$\lim_{l \rightarrow \infty} \mu \{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \} = 0 \quad (6)$$

成立. 即经验风险一致单边收敛于期望风险.

充分性. 现在我们假设式(6)成立. 下面证明严格一致性成立.

用 N 表示事件

$$| \inf_{\alpha \in \Lambda} R(\alpha) - \inf_{\alpha \in \Lambda} R_{\text{emp}}(\alpha) | > \varepsilon.$$

则事件 N 表示两个事件的并集, $N = N_1 \cup N_2$, 其中

$$N_1 = \{ z : \inf_{\alpha \in \Lambda} R(\alpha) + \varepsilon < \inf_{\alpha \in \Lambda} R_{\text{emp}}(\alpha) \},$$

$$N_2 = \{ z : \inf_{\alpha \in \Lambda} R(\alpha) - \varepsilon > \inf_{\alpha \in \Lambda} R_{\text{emp}}(\alpha) \}.$$

由拟概率的拟可加性我们知道

$$\mu(N) \leq \theta^{-1} [\theta \circ \mu(N_1) + \theta \circ \mu(N_2)].$$

假设 N_1 发生, 为了估计 $\mu(N_1)$ 的上界, 我们可找到函数 $Q(z, \alpha^*)$, $\alpha^* \in \Lambda(c)$ 使得

$$R(\alpha^*) < \inf_{\alpha \in \Lambda} R(\alpha) + \frac{\varepsilon}{2}$$

成立, 则下面的不等式成立:

$$\begin{aligned} R(\alpha^*) + \frac{\varepsilon}{2} &< \inf_{\alpha \in \Lambda} R(\alpha) + \varepsilon < \inf_{\alpha \in \Lambda} R_{\text{emp}}(\alpha) \\ &< R_{\text{emp}}(\alpha^*), \end{aligned}$$

即

$$R_{\text{emp}}(\alpha^*) > R(\alpha^*) + \frac{\varepsilon}{2}.$$

由定理 3 可知

$$\mu(N_1) \leq \mu \left\{ R_{\text{emp}}(\alpha^*) - R(\alpha^*) > \frac{\varepsilon}{2} \right\} \xrightarrow{l \rightarrow \infty} 0 \quad (7)$$

另一方面, 假设 N_2 发生, 则存在 α^{**} 使

$$\begin{aligned} R_{\text{emp}}(\alpha^{**}) + \frac{\varepsilon}{2} &< \inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha^{**}) + \varepsilon \\ &< \inf_{\alpha \in \Lambda(c)} R(\alpha^{**}) < R(\alpha^{**}) \end{aligned}$$

成立. 所以

$$\begin{aligned} \mu(N_1) &\leq \mu \left\{ R(\alpha^{**}) - R_{\text{emp}}(\alpha^{**}) > \frac{\varepsilon}{2} \right\} \\ &< \mu \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \frac{\varepsilon}{2} \right\} \xrightarrow{l \rightarrow \infty} 0 \quad (8) \end{aligned}$$

根据式(7), 式(8)和拟概率的拟可加性推断出

$$\begin{aligned} \lim_{l \rightarrow \infty} \mu(N) &< \lim_{l \rightarrow \infty} \theta^{-1} [\theta \circ \mu(N_1) + \theta \circ \mu(N_2)] \\ &= \theta^{-1} [\theta(\lim_{l \rightarrow \infty} \mu(N_1)) + \theta(\lim_{l \rightarrow \infty} \mu(N_2))] \\ &= \theta^{-1}(0) = 0, \end{aligned}$$

所以

$$\lim_{l \rightarrow \infty} \mu \{ | \inf_{\alpha \in \Lambda(c)} R(\alpha) - \inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) | > \varepsilon \} = 0$$

成立. 即经验风险最小化方法在 $Q(z, \alpha)$, $\alpha \in \Lambda$ 上是严格一致的. 至此定理得证. 证毕.

注 4. 由注 1 可知, 概率空间上和 Sugeno 测度空间上学习理论的关键定理^[1,7] 分别是定理 4 的特例.

4 拟概率空间上学习过程一致收敛速度的界

本节我们讨论拟概率空间上经验风险与实际风险之间的关系, 给出学习过程一致收敛速度的界.

首先我们给出拟概率空间上的 Hoeffding 不等式. 本节假定 θ 和 θ^{-1} 的 n 阶导数存在且连续.

定理 5 (Hoeffding's 不等式). 假设 q -随机变

量序列 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立同分布, 令 $S_n = \sum_{i=1}^n \xi_i$,

则 $E_\mu(S_n) = \sum_{i=1}^n E_\mu(\xi_i)$, 对任意 $\lambda > 0$, $t > 0$, 有

$$\begin{aligned} \mu \{ S_n - E_\mu(S_n) \geq t \} &\leq \\ &\theta^{-1} \left[1 - \theta \left(1 - \frac{E_\mu[\exp(\lambda(S_n - E_\mu(S_n)))]}{\exp(\lambda t)} \right) \right] \quad (9) \end{aligned}$$

证明. 由 q -随机变量期望的性质和 Markov 不等式可得. 证毕.

定理 6. 设 q -随机变量 ξ 的期望为零, 且 $\xi \in [a, b]$, 则对任意 $\lambda > 0$, 有

$$E_\mu[\exp(\lambda \xi)] \leq \exp(\lambda^2(b-a)^2/8).$$

证明. 类似于文献[30]中引理 1(或文献[7]中的定理 8). 证毕.

若 q -随机变量序列 $\xi_1, \xi_2, \dots, \xi_n$ 有界, 则拟概率空间上的 Hoeffding 不等式如下.

定理 7. 若有界 q -随机变量序列 $\xi_i \in [a_i, b_i]$, $i=1, 2, \dots, n$ 满足 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立, 则对任意 $t > 0$, 有

$$\mu\{S_n - E_\mu(S_n) \geq t\} \leq \theta^{-1} \left[1 - \theta \left(1 - C_{n\theta} \exp \left(\sum_{i=1}^n \frac{-2t^2}{(b_i - a_i)^2} \right) \right) \right] \quad (10)$$

或

$$\mu\{S_n - E_\mu(S_n) \leq -t\} \leq \theta^{-1} \left[1 - \theta \left(1 - C_{n\theta} \exp \left(\sum_{i=1}^n \frac{-2t^2}{(b_i - a_i)^2} \right) \right) \right] \quad (11)$$

证明. 我们只证明式(10). 由定理 5、性质 6 和定理 6 知

$$\begin{aligned} & \mu\{S_n - E_\mu(S_n) \geq t\} \\ & \leq \theta^{-1} \left[1 - \theta \left(1 - \frac{E_\mu[\exp(\lambda(S_n - E_\mu(S_n)))]}{\exp(\lambda t)} \right) \right] \\ & \leq \theta^{-1} \left[1 - \theta \left(1 - \frac{E_\mu \left[\prod_{i=1}^n \exp(\lambda \xi_i) \right]}{\exp(\lambda t)} \right) \right] \\ & \leq \theta^{-1} \left[1 - \theta \left(1 - \frac{C_{n\theta} \prod_{i=1}^n E_\mu[\exp(\lambda \xi_i)]}{\exp(\lambda t)} \right) \right] \\ & \leq \theta^{-1} \left[1 - \theta \left(1 - C_{n\theta} \exp \left(-\lambda t + \sum_{i=1}^n \frac{\lambda^2 (b_i - a_i)^2}{8} \right) \right) \right], \end{aligned}$$

因此得到

$$\mu\{S_n - E_\mu(S_n) \geq t\} \leq \theta^{-1} \left[1 - \theta \left(1 - C_{n\theta} \exp \left(-\lambda t + \sum_{i=1}^n \frac{\lambda^2 (b_i - a_i)^2}{8} \right) \right) \right],$$

令 $\lambda = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$, 最大化不等式的右端, 得

$$\mu\{S_n - E_\mu(S_n) \geq t\} \leq \theta^{-1} \left[1 - \theta \left(1 - C_{n\theta} \exp \left(\sum_{i=1}^n \frac{-2t^2}{(b_i - a_i)^2} \right) \right) \right].$$

证毕.

下面我们给出 Sugeno 测度空间上学习过程一致收敛速度的界以及这些界与函数集容量之间的关系.

本节只考虑最简单的模型: 函数集仅包含有限个指示函数的情况. 对于实函数集包含有限个实函数的情况, 可以通过指示器将实函数转化为指示函数^[1-2].

进一步, 根据定义 7 和定理 7, 可得如下结果:

$$\mu\{R(\alpha) - R_{\text{emp}}(\alpha) > \epsilon\} = \mu\left\{a - \frac{S_n}{l} > \epsilon\right\} \leq \theta^{-1} [1 - \theta(1 - C_{n\theta} \exp(-2\epsilon^2 l))] \quad (12)$$

和

$$\mu\{R_{\text{emp}}(\alpha) - R(\alpha) > \epsilon\} = \mu\left\{\frac{S_n}{l} - a > \epsilon\right\} \leq \theta^{-1} [1 - \theta(1 - C_{n\theta} \exp(-2\epsilon^2 l))] \quad (13)$$

其中 $b_i - a_i = 1$, l 为样本个数, a 表示事件 $A_a = \{z: Q(z, \alpha) = 1\}$, $\alpha \in \Lambda$ 的拟概率, 即

$$\begin{aligned} a &= R(\alpha) = \int Q(z, \alpha) dF_\mu(z) \\ &= \mu(\{z: \xi(z) = Q(z, \alpha) = 1\}). \end{aligned}$$

此外 $S_n = \sum_{i=1}^n \xi_i$, 从而 $R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) = \frac{1}{l} \sum_{i=1}^l \xi_i = \frac{1}{l} S_n$.

定理 8. 设 $\{Q(z, \alpha_k), k=1, 2, \dots, N\}$ 为包含 N 个元素的指示函数集, $R(\alpha_{k(l)})$ 是使经验风险最小的函数的期望风险值, $R_{\text{emp}}(\alpha_{k(l)})$ 为最小的经验风险值, 则不等式

$$R(\alpha_{k(l)}) \leq R_{\text{emp}}(\alpha_{k(l)}) + \sqrt{-(\ln \xi)/2l}$$

依至少 $\theta^{-1} [1 - \theta(\eta_1)]$ 的拟概率成立, 其中 $\xi = \frac{1 - \theta^{-1} \left(1 - \frac{\theta(\eta_1)}{N} \right)}{C_{n\theta}}$.

证明. 利用不等式(12), 对于任意 $\epsilon_1 > 0$

$$\begin{aligned} & \mu(\{ \sup_{1 \leq k \leq N} (R(\alpha_k) - R_{\text{emp}}(\alpha_k)) > \epsilon_1 \}) \\ & \leq \mu(\{ \bigcup_{k=1}^N (R(\alpha_k) - R_{\text{emp}}(\alpha_k) > \epsilon_1) \}) \\ & \leq \theta^{-1} \left[\sum_{k=1}^N \theta(\mu\{R(\alpha_k) - R_{\text{emp}}(\alpha_k) > \epsilon_1\}) \right] \\ & \leq \theta^{-1} [N[1 - \theta(1 - C_{n\theta} \exp(-2\epsilon_1^2 l))]] \end{aligned} \quad (14)$$

令 $\theta^{-1} [N[1 - \theta(1 - C_{n\theta} \exp(-2\epsilon_1^2 l))]] = \eta_1$, 解 ϵ_1 得

$$\epsilon_1 = \sqrt{-(\ln \xi)/2l},$$

故式(14)可改写成

$$\mu\{ \sup_{1 \leq k \leq N} (R(\alpha_k) - R_{\text{emp}}(\alpha_k)) \leq \epsilon_1 \} \geq \theta^{-1} [1 - \theta(\eta_1)] \quad (15)$$

设 $Q(z, \alpha_{k(l)})$ 为使经验风险最小的函数, 因为上式对函数集 $Q(z, \alpha_k), k=1, 2, \dots, N$ 中的所有不等式成立, 所以不等式

$$R(\alpha_{k(l)}) \leq R_{\text{emp}}(\alpha_{k(l)}) + \sqrt{-(\ln \xi)/2l} \quad (16)$$

依至少 $\theta^{-1} [1 - \theta(\eta_1)]$ 的拟概率成立. 证毕.

注 5. 式(16)估计出所选取函数的风险的上界.

这个界回答了估计经验风险最小化原则的推广能力的第一个问题: 得到最小的经验风险泛函的函数能提供多大的期望风险。

注 6. 当 θ 取 $\theta(x) = x$ 时, 即拟概率退化为概率时, 结果与参考文献[1]中的情况是一样的; 当 θ 取 $\theta(x) = \log_{1+\lambda}(1+\lambda x)$, θ^{-1} 取 $\theta_T^{-1}(x) = \frac{(1+\lambda)^x - 1}{\lambda}$ 时, 即拟概率退化为 Sugeno 测度时, $\theta^{-1}[1 - \theta(\eta_i)]$ 为 $\frac{1 - \eta_i}{1 + \lambda \eta_i}$, 与参考文献[7]中考虑的 $\lambda > 0$ 情况是一样的. 因此定理 8 是概率空间和 Sugeno 测度空间上相应定理的推广.

定理 9. 不等式

$R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \leq \sqrt{-(\ln \xi)/2l} + \sqrt{-(\ln \zeta)/2l}$ 依至少 $\theta^{-1}[1 - 2\theta(\eta)]$ 的拟概率成立, 其中 $\xi = \frac{1 - \theta^{-1}\left(1 - \frac{\theta(\eta_1)}{N}\right)}{C_{n\theta}}$, $\zeta = \frac{1 - \theta^{-1}(1 - \theta(\eta_2))}{C_{n\theta}}$, $\eta = \max\{\eta_1, \eta_2\}$, $R(\alpha_{k(0)})$ 为最小的期望风险值.

证明. 设 $Q(z, \alpha_{k(0)})$ 为使期望风险最小的函数, 根据式(13), 我们有下述不等式成立

$$\mu\{R_{\text{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \epsilon_2\} \leq \theta^{-1}[1 - \theta(1 - C_{n\theta} \exp(-2\epsilon_2^2 l))].$$

令 $\eta_2 = \theta^{-1}[1 - \theta(1 - C_{n\theta} \exp(-2\epsilon_2^2 l))]$, 解 ϵ_2 得

$$\epsilon_2 = \sqrt{-(\ln \zeta)/2l},$$

所以

$$R(\alpha_{k(0)}) \geq R_{\text{emp}}(\alpha_{k(0)}) - \sqrt{-(\ln \zeta)/2l} \quad (17)$$

依至少 $\theta^{-1}[1 - \theta(\eta_2)]$ 的拟概率成立. 因为 $Q(z, \alpha_{k(l)})$ 是使经验风险最小的函数, 所以 $R_{\text{emp}}(\alpha_{k(0)}) - R_{\text{emp}}(\alpha_{k(l)}) \geq 0$ 成立. 令 $\eta = \max\{\eta_1, \eta_2\}$, 有下面不等式成立.

$$\begin{aligned} & \mu\{R(\alpha_{k(l)}) - R(\alpha_{k(0)}) > \epsilon_1 + \epsilon_2\} \\ & \leq \mu\{R(\alpha_{k(l)}) - R_{\text{emp}}(\alpha_{k(l)}) + \\ & \quad R_{\text{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \epsilon_1 + \epsilon_2\} \\ & \leq \mu\{[R(\alpha_{k(l)}) - R_{\text{emp}}(\alpha_{k(l)}) > \epsilon_1] \cup \\ & \quad [R_{\text{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \epsilon_2]\} \\ & = \theta^{-1}[\theta(\mu(R(\alpha_{k(l)}) - R_{\text{emp}}(\alpha_{k(l)}) > \epsilon_1)) + \\ & \quad \theta(\mu(R_{\text{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \epsilon_2))] \\ & \leq \theta^{-1}[\theta(\eta_1) + \theta(\eta_2)] \leq \theta^{-1}[2\theta(\eta)] \quad (18) \end{aligned}$$

所以

$$\begin{cases} \mu\{R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \leq \epsilon_1 + \epsilon_2\} \geq \theta^{-1}[1 - 2\theta(\eta)] \\ R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \leq \sqrt{-(\ln \xi)/2l} + \sqrt{-(\ln \zeta)/2l} \end{cases} \quad (19)$$

依至少 $\theta^{-1}[1 - 2\theta(\eta)]$ 的拟概率成立.

证毕.

注 7. 对于一个给定的函数集, 定理 9 给出的上界估计出所选函数(使经验风险泛函的最小的函数)期望风险值接近最小的期望风险值的程度. 并且定理 8 和 9 所得到的两个不等式都与给定函数集的容量(函数集中函数数目)有关, 这反映了界与函数集的容量的关系.

5 结 论

学习理论的关键定理和学习过程一致收敛速度的界是统计学习理论的重要组成部分. 本文首次给出并证明了拟概率空间上的学习理论的关键定理和学习过程一致收敛速度的界, 为系统地建立拟概率空间上的统计学习理论并以此为基础构建拟支持向量机并应用于实际问题奠定了理论基础. 本文下一步需要研究的主要工作是给出拟概率空间上基于 VC 维的推广性的界, 建立结构风险最小化原则, 构建支持向量机并应用于数据分析、图形图像处理、工程技术、经济管理等领域.

致 谢 作者非常感谢提出对完善本文很有价值的修改意见的两位审稿人, 也十分感谢参与本文讨论和修改的杨兰珍博士生和田大增博士!

参 考 文 献

- [1] Vapnik V N, Xu Jian-Hua, Zhang Xue-Gong translate. Statistical Learning Theory. Beijing: Publishing House of Electronics Industry, 2004(in Chinese)
(Vapnik V N 著. 许建华, 张学工译. 统计学习理论. 北京: 电子工业出版社, 2004)
- [2] Vapnik V N, Zhang Xue-Gong translate. The Nature of Statistical Learning Theory. Beijing: Tsinghua University Press, 2000(in Chinese)
(Vapnik V N 著. 张学工译. 统计学习理论的本质. 北京: 清华大学出版社, 2000)
- [3] Vapnik V N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999
- [4] Bian Zhao-Qi, Zhang Xue-Gong. Pattern Recognition. Beijing: Tsinghua University Press, 1999(in Chinese)
(边肇祺, 张学工. 模式识别. 北京: 清华大学出版社, 1999)
- [5] Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica, 2000, 26(1): 32-42(in Chinese)
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, 26(1): 32-42)

- [6] Liu Hui-Chun, Ma Shu-Yuan. Introduction to support vector machines. *Journal of Image and Graphics*, 2002, 7(6): 618-623(in Chinese)
(柳回春, 马树元. 支持向量机的研究现状. *中国图象图形学报*, 2002, 7(6): 618-623)
- [7] Ha Ming-Hu, Li Yan, Li Jia, Tian Da-Zeng. The key theorem and the bounds on the rate of uniform convergence of learning theory on Sugeno measure spaces. *Science in China (Series E), Information Sciences*, 2006, 36(4): 398-410(in Chinese)
(哈明虎, 李颜, 李嘉, 田大增. Sugeno 测度空间上学习理论的关键定理和一致收敛速度的界. *中国科学(E辑): 信息科学*, 2006, 36(4): 398-410)
- [8] Evgeniou T, Poggio T, Pontil M, Verri A. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 2002, 38: 421-432
- [9] Wechsler H, Duric Z, Li Fa-Yin, V. Cherkassky. Motion estimation using statistical learning theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(4): 466-478
- [10] Raudys S. How good are support vector machines? *Neural Networks*, 2000, 13(1): 17-19
- [11] Tay Francis E H, Cao Lijuan. Application of support vector machines in financial time series forecasting. *Omega*, 2001, 29(4): 309-317
- [12] Tsai C F. Training support vector machines based on stacked generalization for image classification. *Neurocomputing*, 2005, 64: 497-503
- [13] Kim Jong Kyoung, Raghava G P S, Bang Sung Yang, Choi Seungjin. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. *Pattern Recognition Letters*, 2006, 27(9): 996-1001
- [14] Zhan Yi-Qing, Shen Ding-Gang. Design efficient support vector machine for fast classification. *Pattern Recognition*, 2005, 38(1): 157-161
- [15] Jeng Jin-Tsong. Hybrid approach of selecting hyper-parameters of support vector machine for regression. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2006, 36(3): 699-709
- [16] Jin Bo, Tang Y C, Zhang Yan-Qing. Support vector machines with genetic fuzzy feature transformation for biomedical data classification. *Information Sciences*, 2007, 177: 476-489
- [17] Castro J L, Flores-Hidalgo L D, Mantas C J, Puche J M. Extraction of fuzzy rules from support vector machines. *Fuzzy Sets and Systems*, 2007, 158: 2057-2077
- [18] Choquet G. Theory of capacities. *Annales de l'Institut Fourier*, 1954, 5: 131-295
- [19] Sugeno M. Theory of fuzzy integrals and its applications [Ph.D. dissertation]. Tokyo Institute of Technology, Tokyo, 1974
- [20] Wang Zhen-Yuan, George J K. *Fuzzy Measure Theory*. New York: Plenum Press, 1992
- [21] Zadeh L A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1978, 1: 3-28
- [22] Wang Zhen-Yuan. Une classe de mesures floues—les quasi-mesures. *Busefal*, 1981, 6: 28-37
- [23] Liu Bao-Ding. *Theory and Practice of Uncertain Programming*. Heidelberg: Physica-Verlag, 2003
- [24] Ha Ming-Hu, Wu Cong-Xin. *Fuzzy Measure and Fuzzy Integral Theory*. Beijing: Science Press, 1998(in Chinese)
(哈明虎, 吴从妍. 模糊测度与模糊积分理论. 北京: 科学出版社, 1998)
- [25] Grabisch M, Murofushi T, Sugeno M. *Fuzzy Measure and Integrals: Theory and Applications*. New York: Physica-Verlag, 2000
- [26] Sirbiladze G, Gachechiladze T. Restored fuzzy measures in expert decision-making. *Information Sciences*, 2005, 169: 71-95
- [27] Liu Zhi-Qiang, Bruton L T, Bezdek J C et al. Dynamic image sequence analysis using fuzzy measures. *IEEE Transactions on Systems, Man, and Cybernetics-Part. B: Cybernetics*, 2001, 31(4): 557-572
- [28] Liu Bao-Ding, Liu Yan-Kui. Expected value of fuzzy variable and fuzzy expected value models. *IEEE Transactions on Fuzzy Systems*, 2002, 10(4): 445-450
- [29] Zhang Qiang, Gao Long-Chang, Du Wen. Quasi-probabilities and conditional quasi-probabilities. *Journal of Southwest Jiaotong University*, 1998, 33(4): 436-441(in Chinese)
(张强, 高隆昌, 杜文. 拟概率和条件拟概率. *西南交通大学学报*, 1998, 33(4): 436-441)
- [30] Devroye L. Exponential inequalities in nonparametric estimation//Roussas G ed. *Proceedings of the Nonparametric Functional Estimation and Related Topics*. Dordrecht: Kluwer Academic Publishers, 1991: 31-44



HA Ming-Hu, born in 1963, professor, Ph.D. supervisor. His research interests include uncertain information processing, uncertain statistical learning theory, image processing and nonadditive measure theory.

FENG Zhi-Fang, born in 1980, master, teaching assistant. Her research interest is in uncertain statistical learning theory.

SONG Shi-Ji, born in 1965, professor, Ph.D. supervisor. His research interests include uncertain system modeling and optimized computation, soft computing technology and application.

GAO Lin-Qing, born in 1979, master candidate, teaching assistant. His research interest is in uncertain statistical

learning theory.

Background

This work is supported by the National Natural Science Foundation of China "Statistical Learning Theory Based on Uncertain Samples" (grant No. 60573069) and "Study of Uncertain Statistical Learning Theory" (grant No. 60773062).

Statistical Learning Theory or SLT, which deals mainly with the statistical learning principles when samples are limited, is proposed in the 1970s by Vladimir N. Vapnik. It provides important theoretical framework for studying the theory and methods of machine learning when samples are limited, and its kernel idea is to control generalization ability of learning machine by capacity control. SLT provides a solid theoretical foundation for people to study the problem of learning machine when samples are limited, also Support Vector Machine is a new and generic method of learning machine which is developed from this theoretical foundation, and is successfully applied in biological information technology, images graphics processing and economic forecasting. At present, SLT and SVM have become a hot research issue in the field of machine learning. Though SLT is generally acknowledged as a better learning theory handling a small sample in academe, there exist some shortcomings yet. For example, SLT

is provided on the probability (probability measure, a kind of special measure) spaces. As we all know, probability is a set function which satisfies additive or accountable additive on probability spaces. This condition is so strict that sometimes it can't be satisfied in practical application. In other words, there exist a large number of non-additive set functions.

This paper puts forward the concept of Uncertain Statistical Learning Theory. Nowadays, the authors have finished the basic theory research of uncertain Statistical Learning Theory, and have got a series of results on construction and applications of the multistage Support Vector Machine, the Support Vector Machine inverse problem and fuzzy optimization problems.

In the paper, the authors further discuss quasi-probability's proposition, and prove the key theorem of learning theory and the bounds on the rate on uniform convergence of learning process based on quasi-probability space. They expand to scope of theory and application of SLT and build the theoretical foundation of study of SLT on quasi-probability space.