# Sugeno测度空间上学习理论的关键定理和一致收敛速度的界*

哈明虎** 李 颜 李 嘉

(河北大学数学与计算机学院，保定 071002)

田大增

(河北大学物理科学与技术学院，保定 071002)

摘要 讨论了Sugeno测度这类有代表性的非可加测度的性质，给出了Sugeno测度空间上的$g_\lambda$随机变量及其分布函数、期望和方差的定义及性质，证明了Sugeno测度空间上的Markov不等式、Chebyshev不等式和Khinchine大数定律；给出了Sugeno测度空间上的经验风险泛函、期望风险泛函以及ERM原则严格一致收敛的定义，在此基础上给出并证明了Sugeno测度空间上的学习理论的关键定理、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.

关键词 **Sugeno** 测度 经验风险最小化原则 关键定理 一致收敛速度的界

统计学习理论(statistical learning theory, SLT)是在 20 世纪 70 年代由Vapnik[1~3]建立的一种针对小样本情况研究统计学习规律的理论，是传统统计学的重要发展和补充，其核心思想是通过控制学习机器的容量来实现对推广能力的控制[2]，同时也发展了一种新的模式识别方法——支持向量机(support vector machine, SVM). 在短短 30 多年的时间里，国内外涌现出了大量的关于此方面的著作和论文，如文献[1~13]等. 目前，统计学习理论和支持向量机已经成为国际上机器学习领域新的研究热点[4~8].

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. 学习理论的关键定理把学习一致性的问题转化为一致收敛的问题，它

收稿日期: 2005-11-08; 接受日期: 2006-01-15

* 国家自然科学基金项目(批准号: 60573069)、河北省自然科学基金重点项目(批准号: F2004000129)和河北省教育厅科研计划重点项目(批准号: 2005001D)资助

** E-mail: mhha@mail.hbu.edu.cn

指出了经验风险最小化原则(empirical risk minimization, ERM)一致性的充分和必要条件是取决于函数集中“最坏”的函数[1,2]; 学习过程一致收敛速度的界是分析学习机器性能和发展新的学习算法的重要基础[4], 通过对这些界的讨论, 可以得到在ERM原则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.

尽管统计学习理论是学术界公认的较好的处理小样本的学习理论, 但它仍存在一些不完善之处, 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. 众所周知, 概率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. 由于可加性条件非常苛刻, 在实际应用中这个条件往往得不到满足. 换言之, 在实际应用中存在着大量的非可加集函数, 如: 对某商店出售的一种服装进行评判, 为简单起见, 我们只考虑三个主要因素, 即花色式样、耐穿程度及价格, 并由此来组成论域$U$={花色式样($U_1$), 耐穿程度($U_2$), 价格($U_3$)}, 其中$\{U_i\}$ ($i$=1, 2, 3)为两两不相交的集合. 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$\mu$ ($\mu$是从$U$到[0, 1]的一个集函数)如下:

$$\mu(U_1)=0.6,\ \ \mu(U_2)=0.5,\ \ \mu(U_3)=0.8,$$

$$\mu(U_1\cup U_2)=0.7,\ \ \mu(U_1\cup U_3)=0.9,\ \ \mu(U_2\cup U_3)=0.8,\ \ \mu(U_1\cup U_2\cup U_3)=1.$$

显然有$\mu(U_1\cup U_2)\neq\mu(U_1)+\mu(U_2)$, 这从某种意义上说明了非可加集函数的存在. 关于建立在概率测度上的理论与应用的局限性的详细讨论可参见文献[14,15]和脚注 1).

1974 年, 日本学者Sugeno[1)]类非常有代表性的非可加测度, 即Sugeno测度[15,16][1)], 它是概率测度的重要推广. 本文将统计学习理论的研究从概率空间拓展到Sugeno测度空间上, 给出并证明了此空间上学习理论的关键定理, 学习过程一致收敛速度的界及这些界与函数集容量之间的关系.

## 1　Sugeno 测度空间上学习理论的关键定理

为了把学习理论的关键定理从概率(概率测度)空间推广到 Sugeno 测度空间, 我们首先给出本文所需要的一些基本概念、性质与定理.

**定义 1**[15]　设$X$是一非空集合, $\zeta$是由$X$的子集组成的非空类, $\mu$是定义在$\zeta$上的非负实值集函数. 称$\mu$满足$\sigma-\lambda$律(在$\zeta$上)当且仅当存在$\lambda\in\left(-\dfrac{1}{\sup\mu},\infty\right)\cup\{0\}$使得下式成立:

1) Sugeno M. Theory of fuzzy integrals and its applications. Doctoral Thesis, Tokyo Institute of Technology, 1974

$$\mu\left(\bigcup_{i=1}^{\infty}E_i\right)=\begin{cases}\dfrac{1}{\lambda}\left\{\prod_{i=1}^{\infty}[1+\lambda\cdot\mu(E_i)]-1\right\}, & \lambda\neq 0,\\ \sum_{i=1}^{\infty}\mu(E_i) & ,\quad \lambda=0,\end{cases}$$

其中$\{E_i\}$是$\zeta$中一列互不相交集合，且它们的并也在$\zeta$中. 特别地，当$\lambda=0$时，$\sigma-\lambda$律对应$\sigma-$可加.

**定义 2**[15] 设$\mathcal{F}$是$X$的子集组成的$\sigma$代数. $\mu$称为$\mathcal{F}$上的Sugeno测度，当且仅当$\mu$满足$\sigma-\lambda$律且$\mu(X)=1$，简记为$g_\lambda$.

三元组$(X,\mathcal{F},g_\lambda)$称为 Sugeno 测度空间，简记为$g_\lambda$空间. 本文以下的讨论均在$g_\lambda$空间上进行.

**注 1** 从上述定义易知，$g_\lambda$在$\lambda=0$时就是概率(概率测度)，且具有很强的“适应性”，只要选择适当的参数$\lambda$，它就能近似代表相当广泛的一些非可加测度，因而被许多学者认为是很重要的一类有代表性的非可加测度[15,16]1).

直接由定义 2 和数学归纳法，我们不难给出下面的结论:

**定理 1** 对任意的$E_i\in\mathcal{F}$和$E_i\cap E_j=\varnothing\quad(i\neq j),i,j=1,2,\cdots$，下列不等式成立:

$$g_\lambda\{\bigcup_{i=1}^{\infty}E_i\}\leq B\sum_{i=1}^{\infty}g_\lambda(E_i),$$

其中当$\lambda\leq 0$时，$B=1$; 当$\lambda>0$时，$B=1+\lambda$.

由$g_\lambda$的定义和$T$-函数的定义[16]，我们还可证明下面的重要结果:

**定理 2** 对于任意一个$g_\lambda$，总可以找到一个$T$-函数，使得集函数$\theta_T\circ g_\lambda$成为一个概率测度，且此$T$-函数为

$$\theta_T(x)=\begin{cases}\dfrac{\ln(1+\lambda x)}{\ln(1+\lambda)}, & \lambda\neq 0,\\ x, & \lambda\neq 0,\end{cases}\qquad x\in[0,1].$$

利用上述基本定义和定理我们可给出$g_\lambda$随机变量及其分布函数、期望和方差的定义和一些基本性质.

**定义 3** 设$\xi$是从$(X,\mathcal{F},g_\lambda)$到实直线$\Re$上的函数，则称$\xi$是一个$g_\lambda$随机变量.

**定义 4** 设$g_\lambda$随机变量只取有限个数值或可列无穷多个值，则这一类变量称之为离散型$g_\lambda$随机变量.

**定义 5** 设$\xi$是一个离散型$g_\lambda$随机变量，亦即$\xi$的一切可能值为$x_1,x_2,\cdots,x_n,\cdots$，且$g_{\lambda_n}=g_\lambda(\xi=x_n)(n=1,2,\cdots)$，则称$g_{\lambda_1},g_{\lambda_2},\cdots,g_{\lambda_n},\cdots$为$\xi$的分布列.

1) 见 399 脚注 1)

**定义 6**　设 $\xi$ 是一个 $g_\lambda$ 随机变量, 则其分布函数定义为

$$F_{g_\lambda}(x)=g_\lambda\{\xi\le x\},\quad \forall x\in\Re.$$

由定义 3 和 6, 可得如下性质:

**性质 1** 设 $F_{g_\lambda}(x)$ 是 $g_\lambda$ 随机变量 $\xi$ 的分布函数, 则

(1) $F_{g_\lambda}(x)$ 单调不减;

(2) $F_{g_\lambda}(x)$ 是右连续的.

**性质 2**　$g_\lambda\{x_1<x\le x_2\}=\dfrac{F_{g_\lambda}(x_2)-F_{g_\lambda}(x_1)}{1+\lambda\cdot F_{g_\lambda}(x_1)}.$

**注 2**　由上述两性质可以看出 $g_\lambda$ 随机变量的分布函数的性质有的与随机变量的分布函数的性质相同, 有的则不同.

**定义 7**　对于 $g_\lambda$ 随机变量 $\xi$ 的分布函数 $F_{g_\lambda}(x)$, 若存在非负实函数 $f_{g_\lambda}(x)$, 使得对于任意实数 $x$, 有 $F_{g_\lambda}(x)=\int_{-\infty}^{x}f_{g_\lambda}(t)\mathrm{d}t$, 则称 $\xi$ 为连续型 $g_\lambda$ 随机变量, 其中 $f_{g_\lambda}(x)$ 称为 $\xi$ 的密度函数.

**定义 8**　设 $\xi$ 是离散型 $g_\lambda$ 随机变量, 若 $\sum_{i=1}^{\infty}|x_i|g_{\lambda i}<\infty$, 记 $E_{g_\lambda}(\xi)=\sum_{i=1}^{\infty}x_i g_{\lambda i}$, 称 $E_{g_\lambda}(\xi)$ 为离散型 $g_\lambda$ 随机变量 $\xi$ 的数学期望.

**定义 9**　设 $\xi$ 为具有密度函数 $f_{g_\lambda}(x)$ 的 $g_\lambda$ 随机变量, 若 $\int_{-\infty}^{\infty}|x|f_{g_\lambda}(x)\mathrm{d}x<\infty$, 记 $E_{g_\lambda}(\xi)=\int_{-\infty}^{\infty}xf_{g_\lambda}(x)\mathrm{d}x$, 称 $E_{g_\lambda}(\xi)$ 为连续型 $g_\lambda$ 随机变量 $\xi$ 的数学期望.

对于一般的 $g_\lambda$ 随机变量, 需要给出更一般的数学期望定义, 因此, 借助 Riemann-Stieltjes 积分(简称 R-S 积分)给出如下定义:

**定义 10**　设 $g_\lambda$ 随机变量 $\xi$ 的分布函数是 $F_{g_\lambda}(x)$. 若 $\int_{-\infty}^{\infty}|x|\mathrm{d}F_{g_\lambda}(x)<\infty$, 则称积分 $\int_{-\infty}^{\infty}x\mathrm{d}F_{g_\lambda}(x)$ 的值为 $g_\lambda$ 随机变量 $\xi$ 的期望, 记为 $E_{g_\lambda}(\xi)$.

关于 Sugeno 测度空间上事件的发生、二维随机变量、二维随机变量联合分布函数以及密度函数和边缘分布函数等定义也可由类似概率空间上的对应定义给出.

由于 $g_\lambda$ 随机变量分布函数与概率空间上的随机变量的分布函数性质不尽相同, 从而 $g_\lambda$ 随机变量相互独立的重要概念就不能完全按概率空间上的定义给出. 下面我们给出两个 $g_\lambda$ 随机变量相互独立的定义.

**定义 11**　设 $F_{g_\lambda}(x,y)$ 及 $F_{g_\lambda\xi}(x),F_{g_\lambda\eta}(y)$ 分别是二维 $g_\lambda$ 随机变量 $(\xi,\eta)$ 的分布函数及边缘分布函数, 若对于所有的 $x,y$ 有

$$g_\lambda(\xi \le x, \eta \le y) = \theta_T^{-1}\{\theta_T[g_\lambda(\xi \le x)] \cdot \theta_T[g_\lambda(\eta \le y)]\},$$

即

$$F_{g_\lambda}(x, y) = \theta_T^{-1}\{\theta_T[F_{g_\lambda \xi}(x)] \cdot \theta_T[F_{g_\lambda \eta}(y)]\},$$

其中 $\theta_T(x) = \dfrac{\ln(1+\lambda x)}{\ln(1+\lambda)}$, $\theta_T^{-1}(x) = \dfrac{(1+\lambda)^x - 1}{\lambda}$, 则称 $g_\lambda$ 随机变量 $\xi$ 和 $\eta$ 是相互独立的.

由上述定义不难推出 $g_\lambda$ 随机变量的数学期望的如下基本性质.

**性质 3** 设 $\xi, \eta$ 是两个 $g_\lambda$ 随机变量, $C, D$ 是常数, 则有 $E_{g_\lambda}(C\xi + D\eta) = CE_{g_\lambda}(\xi) + DE_{g_\lambda}(\eta)$.

**性质 4** 设 $\xi, \eta$ 是两个相互独立的 $g_\lambda$ 随机变量, 则有 $E_{g_\lambda}(\xi\eta) \le AE_{g_\lambda}(\xi)E_{g_\lambda}(\eta)$, 其中 $A>0$ 为一个常数. 当 $\lambda>0$ 时, $A=[\ln(1+\lambda)+1]\dfrac{\lambda}{\ln(1+\lambda)} \cdot (1+\lambda)$; 当 $\lambda<0$ 时, $A=\dfrac{\lambda}{\ln(1+\lambda)} \cdot \dfrac{1}{(1+\lambda)^2}$; 当 $\lambda=0$ 时, $A=1$ 并且 $E_{g_\lambda}(\xi\eta) = E_{g_\lambda}(\xi)E_{g_\lambda}(\eta)$.

**推论 1** 设 $\xi_1, \xi_2, \cdots, \xi_n$ 是相互独立的 $g_\lambda$ 随机变量, 则有 $E_{g_\lambda}\left(\prod_{i=1}^n \xi_i\right) \le A^{n-1}\prod_{i=1}^n E_{g_\lambda}(\xi_i)$.

**注 3** 由上述性质 3 和 4 可以看出 $g_\lambda$ 随机变量的数学期望性质与随机变量的数学期望的性质有的相同, 有的则不同.

**定义 12** 设 $\xi$ 是一个 $g_\lambda$ 随机变量. 若 $E_{g_\lambda}\{[\xi - E_{g_\lambda}(\xi)]^2\}$ 存在, 则称 $E_{g_\lambda}\{[\xi - E_{g_\lambda}(\xi)]^2\}$ 为 $\xi$ 的方差, 记为 $D_{g_\lambda}(\xi)$.

**性质 5** 设 $\xi$ 是一个 $g\lambda$ 随机变量, $A, B$ 是常数, 则有 $D_{g_\lambda}(A\xi + B) = A^2 D_{g_\lambda}(\xi)$.

**性质 6** 设 $\xi, \eta$ 是两个独立的 $g_\lambda$ 随机变量, 则有 $D_{g_\lambda}(\xi + \eta) \le D_{g_\lambda}(\xi) + D_{g_\lambda}(\eta)$.

**注 4** 由上述性质 5 和 6 可以看出 $g_\lambda$ 随机变量的方差的性质与随机变量的方差的性质有的相同, 有的则不同.

下面我们给出 $g_\lambda$ 空间上的 Markov 不等式、Chebyshev 不等式与 Khinchine 大数定律.

**定理 3** (Markov 不等式) 若 $\xi$ 是非负 $g_\lambda$ 随机变量, 且 $t>0$, 则 $g_\lambda\{\xi \ge t\} \le B\dfrac{E_{g_\lambda}(\xi)}{t}$. 对连续的 $g_\lambda$ 随机变量, 当 $\lambda \le 0$ 时, $B = \dfrac{1}{1+\lambda}$; 当 $\lambda>0$ 时, $B=1$. 对离散

的 $g_\lambda$ 随机变量当 $\lambda>0$ 时，$B=1+\lambda$；当 $\lambda\le 0$ 时, $B=1$.

证　由定义 8~10 以及性质 2 即可得证.

**定理 4** (Chebyshev 不等式)　设 $g_\lambda$ 随机变量 $\xi$ 的期望 $E_{g_\lambda}(\xi)=\mu$，方差 $D_{g_\lambda}(\xi)=\sigma^2$，则对任意正数 $\varepsilon$，不等式

$$g_\lambda\{|\xi-\mu|\ge\varepsilon\}\le B\frac{\sigma^2}{\varepsilon^2}$$

成立.

证　由定理 3 知, $g_\lambda\{|\xi-\mu|^2\ge\varepsilon^2\}\le B\frac{E(|\xi-\mu|^2)}{\varepsilon^2}=B\frac{\sigma^2}{\varepsilon^2}$，因此 $g_\lambda\{|\xi-\mu|\ge\varepsilon\}\le B\frac{\sigma^2}{\varepsilon^2}$.

**定理 5** (Khinchine 大数定律)　设 $g_\lambda$ 随机变量 $\xi_1,\xi_2,\cdots,\xi_n,\cdots$ 相互独立同分布, 且有相同的数学期望 $E_{g_\lambda}(\xi_n)=a(n=1,2,\cdots)$，则对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty} g_\lambda\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_n-a\right|\ge\varepsilon\right\}=0.$$

证　由定理 4 并参照概率空间上的 Khinchine 大数定律的证明即可得证.

**注 5**　定理 5 的证明也可利用定理 2 与概率空间上的 Khinchine 大数定律给出.

为了得到本节的主要结论，下面我们给出 $g_\lambda$ 空间上的经验风险泛函、期望风险泛函以及 ERM 原则严格一致收敛的定义.

**定义 13**　设 $\xi_1,\xi_2,\cdots,\xi_n,\cdots$ 是一个 $g_\lambda$ 随机变量序列, $a$ 是一个常数. 如果对任意 $\varepsilon>0$ 有

$$\lim_{n\to\infty} g_\lambda\{|\xi_n-a|>\varepsilon\}=0,$$

则称 $\{\xi_n\}$ 依 $g_\lambda$ 收敛到 $a$. 简记为 $\xi_n \xrightarrow[n\to\infty]{g_\lambda} a$ .

**定义 14**　设 $F_{g_\lambda}(z)$ 是给定的 $g_\lambda$ 随机变量 $z$ 的分布函数，$z_1,z_2,\cdots,z_l$ 是独立同分布的样本. 考虑函数集 $Q(z,\alpha),\alpha\in\Lambda$，则 $g_\lambda$ 空间上的期望风险泛函和经验风险泛函分别定义为

$$R(\alpha)=\int Q(z,\alpha)\mathrm{d}F_{g_\lambda}(z), \tag{1}$$

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

对于指示函数集，(1)式描述了事件 $A_\alpha=\{z:Q(z,\alpha)=1\},\alpha\in\Lambda$ 的 $g_\lambda$；(2)式描述了事件的频率.

**定义 15** 对函数集 $Q(z,\alpha),\alpha\in\Lambda$ 定义其子集 $\Lambda(c)$ 如下:

$$\Lambda(c)=\left\{\alpha:\int Q(z,\alpha)\mathrm{d}F_{g_\lambda}(z)\geq c,\alpha\in\Lambda\right\}.$$

如果对函数集的任意非空子集 $\Lambda(c)$, $c\in(-\infty,\infty)$ 都有

$$\inf_{\alpha\in\Lambda(c)}R_{\mathrm{emp}}(\alpha)\underset{l\to\infty}{\overset{g_\lambda}{\to}}\inf_{\alpha\in\Lambda(c)}R(\alpha)$$

成立, 则称经验风险最小化(ERM)原则对函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 和分布函数 $F_{g_\lambda}(z)$ 是严格一致的.

本节的最后我们给出 $g_\lambda$ 空间上学习理论的关键定理.

**定理 6** 设对给定的分布函数 $F_{g_\lambda}(z)$ 和函数集 $Q(z,\alpha),\alpha\in\Lambda$ 中所有函数存在常数 $A$ 和 $B$ 使得不等式

$$A\leq R(\alpha)=\int Q(z,\alpha)\mathrm{d}F_{g_\lambda}(z)\leq B$$

成立. 则 ERM 原则严格一致性的充分必要条件是: 经验风险泛函 $R_{\mathrm{emp}}(\alpha)$ 在函数集 $Q(z,\alpha),\alpha\in\Lambda$ 上在如下定义下一致收敛于实际风险 $R(\alpha)$:

$$\lim_{l\to\infty}g_\lambda\{\sup_{\alpha\in\Lambda}(R(\alpha)-R_{\mathrm{emp}}(\alpha))>\varepsilon\}=0,\quad\forall\varepsilon>0$$

时一致收敛于风险 $R(\alpha)$.

证 必要性: 设经验风险最小化原则在函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 上是严格一致的. 根据严格一致的定义可得, $\forall c\in(-\infty,+\infty)$ 使得集合

$$\Lambda(c)=\left\{\alpha:\int Q(z,\alpha)\mathrm{d}F(z)\geq c\right\}$$

非空, 则在如下意义下的收敛是成立的:

$$\inf_{\alpha\in\Lambda(c)}R_{\mathrm{emp}}(\alpha)\underset{l\to\infty}{\overset{g_\lambda}{\to}}\inf_{\alpha\in\Lambda(c)}R(\alpha).\tag{3}$$

考虑有限序列 $a_1,a_2,\cdots,a_n$, 使得

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

以下我们用$T_k$表示事件

$$\inf_{\alpha\in\Lambda(a_k)}R_{\mathrm{emp}}(\alpha)<\inf_{\alpha\in\Lambda(a_k)}R(\alpha)-\frac{\varepsilon}{2}.$$

则由(3)式可得

$$g_\lambda(T_k)\underset{l\to\infty}{\to}0.\tag{4}$$

令 $T=\bigcup_{k=1}^{n}T_k$, 下面我们用归纳法来证明 $g_\lambda\left(\bigcup_{k=1}^{n}T_k\right)\underset{l\to\infty}{\to}0.$

当 $i$=2 时,

$$g_\lambda(T_1\cup T_2)=\frac{g_\lambda(T_1)+g_\lambda(T_2)-g_\lambda(T_1\cap T_2)+\lambda\cdot g_\lambda(T_1)g_\lambda(T_2)}{1+\lambda\cdot g_\lambda(T_1\cap T_2)}\le\frac{g_\lambda(T_1)+g_\lambda(T_2)+\lambda\cdot g_\lambda(T_1)g_\lambda(T_2)}{1+\lambda\cdot g_\lambda(T_1\cap T_2)},$$

由(4)式可知 $g_\lambda(T_1)\underset{l\to\infty}{\to}0$ 和 $g_\lambda(T_2)\underset{l\to\infty}{\to}0$ 成立, 因此我们可得 $g_\lambda(T_1\cup T_2)\underset{l\to\infty}{\to}0$.

设当 $i$=$n$−1 时, $g_\lambda\left(\bigcup_{k=1}^{n-1}T_k\right)\underset{l\to\infty}{\to}0$ 成立. 则 $g_\lambda\left(\bigcup_{k=1}^{n}T_k\right)=g_\lambda\left(\left(\bigcup_{k=1}^{n-1}T_k\right)\cup T_n\right)\underset{l\to\infty}{\to}0$ 也成立, 即

$$g_\lambda(T)\underset{l\to\infty}{\to}0. \tag{5}$$

我们用 $M$ 来表示事件

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

设事件 $M$ 发生. 则存在 $\alpha^*\in\Lambda$, 使得

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

成立. 由 $\alpha^*$ 我们可以找出一个 $k$ 使得 $\alpha^*\in\Lambda(a_k)$ 和

$$0\le R(\alpha^*)-a_k<\frac{\varepsilon}{2}$$

成立. 由选定的集合 $\Lambda(a_k)$, 我们可知下面不等式

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

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

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

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

$$\inf_{\alpha\in\Lambda(a_k)}R(\alpha)-\frac{\varepsilon}{2}>R(\alpha^*)-\varepsilon>R_{\text{emp}}(\alpha^*)\ge\inf_{\alpha\in\Lambda(a_k)}R_{\text{emp}}(\alpha).$$

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 $g_\lambda$ 的单调性, 可知 $g_\lambda(M)\le g_\lambda(T)$ 成立. 由(5)式得 $\lim_{l\to\infty}g_\lambda(T)=0$, 所以

$$\lim_{l\to\infty}g_\lambda\{\sup_{\alpha\in\Lambda}(R(\alpha)-R_{\text{emp}}(\alpha))>\varepsilon\}=0. \tag{6}$$

定理的必要性得证.

充分性: 下面我们用 $N$ 表示事件

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

则事件 $N$ 为下面两个事件的并

$$N = N_1 \cup N_2,$$

其中

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

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

设事件 $N_1$ 发生. 为了估计 $g_\lambda(N_1)$ 的上界, 我们可找到函数 $Q(z,\alpha^*)$, $\alpha^* \in \Lambda(c)$ 使得

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

成立. 则下面不等式

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

成立. 由定理 5 可得

$$g_\lambda(N_1) \le g_\lambda \left\{ R_{\text{emp}}(\alpha^*) - R(\alpha^*) > \frac{\varepsilon}{2} \right\} \underset{l \to \infty}{\to} 0. \tag{7}$$

另一方面, 如果事件$N_2$发生, 则存在函数$Q(z,\alpha^{**})$, $\alpha^{**} \in \Lambda(c)$ 使得下式成立:

$$R_{\text{emp}}(\alpha^{**}) + \frac{\varepsilon}{2} < \inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) + \varepsilon < \inf_{\alpha \in \Lambda(c)} R(\alpha) \le R(\alpha^{**}).$$

再由(6)式成立可得

$$g_\lambda(N_2) \le g_\lambda \left\{ R(\alpha^{**}) - R_{\text{emp}}(\alpha^{**}) < \frac{\varepsilon}{2} \right\} < g_\lambda \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \frac{\varepsilon}{2} \right\} \underset{l \to \infty}{\to} 0. \tag{8}$$

由(7)和(8)式, 我们可知 $g_\lambda(N) = g_\lambda(N_1 \cup N_2) \underset{l \to \infty}{\to} 0$. 定理的充分性也得证.

**注 6** 定理 6 的证明也可利用定理 2 和概率空间上学习理论的关键定理给出, 我们将另文讨论.

## 2 Sugeno 测度空间上学习过程一致收敛速度的界

本节我们首先给出 Sugeno 测度空间上的 Hoeffding 不等式.

**定理 7** (Hoeffding 不等式) 假设 $g_\lambda$ 随机变量 $\xi_1, \xi_2, \cdots, \xi_n$ 相互独立. 设 $S_n = \sum_{i=1}^n \xi_i$ 且 $E_{g_\lambda}(S_n) = \sum_{i=1}^n E_{g_\lambda}(\xi_i)$, 则 $\forall s > 0, t > 0$, 有

$$g_\lambda\{S_n - E_{g_\lambda}(S_n) \ge t\} \le Ce^{-st} \prod_{i=1}^n E_{g_\lambda}\{\exp[s(\xi_i - E_{g_\lambda}(\xi_i))]\},$$

其中 $C = A^{n-1}B$.

证　由 $g_\lambda$ 随机变量的期望定义和性质以及 Markov 不等式即可得证.

然后, 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.

**定理 8**　若 $g_\lambda$ 随机变量 $\xi$ 的均值为 0, 即 $E_{g_\lambda}(\xi)=0$, 且 $\xi\in[a,b]$, 则对于任意 $s>0$, 有

$$E_{g_\lambda}(e^{s\xi})\le e^{s^2(b-a)^2/8}.$$

证　由 $g_\lambda$ 随机变量的期望定义和性质以及指数函数的凸性即可得证.

**定理 9**　若有界 $g_\lambda$ 随机变量 $\xi_i\in[a_i,b_i]$, $i=1,2,\cdots,n$, 满足 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立, 则 $\forall t>0$, 有

$$g_\lambda\{S_n-E_{g_\lambda}(S_n)\ge t\}\le C\exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right\},$$

或

$$g_\lambda\{S_n-E_{g_\lambda}(S_n)\le -t\}\le C\exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right\}.$$

其中$C=A^{n-1}B$.

证　由定理 7 和 8 即可得证.

本文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. 对于实函数的情况可利用一个指示器将实函数转化为指示函数. 将定理 9 应用于两点分布, 结合定义 14 可得到

$$g_\lambda\{R(\alpha)-R_{\mathrm{emp}}(\alpha)>\varepsilon\}=g_\lambda\{g-S_n/l>\varepsilon\}\le C\exp(-2\varepsilon^2 l) \tag{9}$$

和

$$g_\lambda\{R_{\mathrm{emp}}(\alpha)-R(\alpha)>\varepsilon\}\le C\exp(-2\varepsilon^2 l), \tag{10}$$

其中 $b_i-a_i=1$, $l$ 为样本数, $g$ 是事件 $A_\alpha=\{z:Q(z,\alpha)=1\},\alpha\in\Lambda$ 的 $g_\lambda$.

**定理 10**　设 $Q(z,\alpha_k),k=1,2,\cdots,N$ 为函数集, 当 $\lambda\le 0$ 时,

$$R(\alpha_k)-R_{\mathrm{emp}}(\alpha_k)\le\sqrt{\frac{\ln NC-\ln\eta_1}{2l}}$$

依至少 $1-\eta_1$ 的 $g_\lambda$ 成立; 当 $\lambda>0$ 时,

$$R(\alpha_k)-R_{\text{emp}}(\alpha_k)\le\sqrt{\frac{\ln\lambda C-\ln[e^{\ln(1+\lambda\eta_2)-\ln N}-1]}{2l}}$$

依至少 $\frac{1-\eta_2}{1+\lambda\eta_2}$ 的 $g_\lambda$ 成立.

证　利用不等式(9), $\forall\varepsilon_1>0$, 当 $\lambda\le 0$ 时

$$\begin{aligned}&g_\lambda\{\sup_{1\le k\le N}(R(\alpha_k)-R_{\text{emp}}(\alpha_k))>\varepsilon_1\}\\ \le&\sum_{k=1}^{N}g_\lambda\{R(\alpha_k)-R_{\text{emp}}(\alpha_k)>\varepsilon_1\}\\ \le& NC\exp(-2\varepsilon_1^2 l).\end{aligned}$$

$\forall 0<\eta_1\le 1$, 令 $NC\exp(-2\varepsilon_1^2 l)=\eta_1$, 可得 $\varepsilon_1=\sqrt{\frac{\ln NC-\ln\eta_1}{2l}}$. 故

$$g_\lambda\{\sup_{1\le k\le N}(R(\alpha_k)-R_{\text{emp}}(\alpha_k))\le\varepsilon_1\}\ge 1-\eta_1.$$

当 $\lambda>0$ 时, $\forall\varepsilon_2>0$,

$$\begin{aligned}&g_\lambda\{\sup_{1\le k\le N}(R(\alpha_k)-R_{\text{emp}}(\alpha_k))>\varepsilon_2\}\\ \le& g_\lambda\{\bigcup_{i=k}^{N}(R(\alpha_k)-R_{\text{emp}}(\alpha_k))>\varepsilon_2\}\\ \le&\frac{1}{\lambda}\left\{\prod_{i=k}^{N}\left[1+\lambda g_\lambda\left\{(R(\alpha_k)-R_{\text{emp}}(\alpha_k)>\varepsilon_2\right\}\right]-1\right\}\\ \le&\frac{1}{\lambda}\left\{\left[1+\lambda C\exp(-2\varepsilon_2^2 l)\right]^N-1\right\}.\end{aligned}$$

$\forall 0<\eta_2\le 1$, 令 $\frac{1}{\lambda}\left\{\left[1+\lambda C\exp(-2\varepsilon_2^2 l)\right]^N-1\right\}=\eta_2$, 解得

$$\varepsilon_2=\sqrt{\frac{\ln\lambda C-\ln(\sqrt[N]{1+\lambda\eta_2}-1)}{2l}},$$

故

$$g_\lambda\{\sup_{1\le k\le N}(R(\alpha_k)-R_{\text{emp}}(\alpha_k))\le\varepsilon_2\}\ge\frac{1-\eta_2}{1+\lambda\eta_2}.$$

设 $Q(z,\alpha_{k(0)})$ 是函数集中最小化期望风险泛函的函数, $Q(z,\alpha_{k(l)})$ 是最小化经验风险泛函的函数, 因为定理 10 对函数集中所有函数都成立, 所以对 $Q(z,\alpha_{k(l)})$ 同样成立. 当 $\lambda=0$ 时, 上述的界即为概率测度空间上的界.

这些界就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提供多大的期望风险. 下面定理回答了第二个问题:

**定理 11**　当 $\lambda \le 0$ 时, 不等式

$$R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \le \sqrt{\frac{\ln NC - \ln \eta_1}{2l}} + \sqrt{\frac{\ln C - \ln \eta_3}{2l}}$$

依至少 $\frac{1+\lambda\eta-2\eta}{1+\lambda\eta}$ 的 $g_\lambda$ 成立; 当 $\lambda > 0$ 时,

$$R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \le \sqrt{\frac{\ln \lambda C - \ln(\sqrt[N]{1+\lambda\eta_2} - 1)}{2l}} + \sqrt{\frac{\ln C - \ln \eta_3}{2l}}$$

依至少 $\frac{1-2\eta-\lambda\eta^2}{1+2\lambda\eta+\lambda^2\eta^2}$ 的 $g_\lambda$ 成立.

证　对于最小化期望风险泛函的函数 $Q(z,\alpha_{k(0)})$, 由不等式(10)解得 $\varepsilon_3 = \sqrt{\frac{\ln C - \ln \eta_3}{2l}}$, 所以, 当 $\lambda \le 0$ 时, $R(\alpha_{k(0)}) \ge R_{\mathrm{emp}}(\alpha_{k(0)}) - \sqrt{\frac{\ln C - \ln \eta_3}{2l}}$ 依至少 $1-\eta_3$ 的 $g_\lambda$ 成立; 而 $\lambda > 0$ 时, 依至少 $\frac{1-\eta_3}{1+\lambda\eta_3}$ 的 $g_\lambda$ 成立.

因为 $Q(z,\alpha_{k(l)})$ 最小化经验风险泛函, 所以不等式 $R_{\mathrm{emp}}(\alpha_{k(0)}) - R_{\mathrm{emp}}(\alpha_{k(l)}) \ge 0$ 成立. 令 $\eta = \max\{\eta_1, \eta_3\}$ 或 $\max\{\eta_2, \eta_3\}$, 当 $\lambda \le 0$ 时,

$$\begin{aligned}
& g_\lambda(R(\alpha_{k(l)}) - R(\alpha_{k(0)}) > \varepsilon_1 + \varepsilon_3) \\
\le\ & g_\lambda\{R(\alpha_{k(l)}) - R_{\mathrm{emp}}(\alpha_{k(l)}) + R_{\mathrm{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \varepsilon_1 + \varepsilon_3\} \\
\le\ & g_\lambda\{[R(\alpha_{k(l)}) - R_{\mathrm{emp}}(\alpha_{k(l)}) > \varepsilon_1] \cup [R_{\mathrm{emp}}(\alpha_{k(0)}) - R(\alpha_{k(0)}) > \varepsilon_3]\} \\
\le\ & \frac{2\eta}{1+\lambda\eta},
\end{aligned}$$

所以

$$g_\lambda\{R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \le \varepsilon_1 + \varepsilon_3\} \ge \frac{1+\lambda\eta-2\eta}{1+\lambda\eta}.$$

同理, 当 $\lambda > 0$ 时,

$$g_\lambda\{R(\alpha_{k(l)}) - R(\alpha_{k(0)}) > \varepsilon_2 + \varepsilon_3\} \le 2\eta + \lambda\eta^2,$$

故

$$g_\lambda\{R(\alpha_{k(l)}) - R(\alpha_{k(0)}) \le \varepsilon_2 + \varepsilon_3)\} \ge \frac{1-2\eta-\lambda\eta^2}{1+2\lambda\eta+\lambda^2\eta^2}.$$

当 $\lambda = 0$ 时, 上述的界即为概率测度空间上的界.

定理 11 中的界回答了第二个问题: 对于一个给定的函数集, 由最小的经验

风险泛函的函数所提供的期望风险和最小的可能风险有所接近. 并且定理 10 和 11 所得到的两个不等式都与给定函数集的容量(函数集中函数数目的对数)有关, 反映了界与函数集的容量的关系.

**注 7** 由定理 2 和概率空间上学习过程一致收敛速度的界也可得到 Sugeno 测度空间上学习过程一致收敛速度的界, 限于篇幅, 该部分内容我们将另文讨论.

## 3 结论

本文进一步讨论了 Sugeno 测度这类很重要的有代表性的非可加测度的一些性质, 在此 Sugeno 测度空间上给出了 $g_\lambda$ 随机变量及其分布函数等定义, 分析了其数字特征, 并证明了一些重要不等式, 在此基础上把学习理论的关键定理、学习过程一致收敛速度的界以及这些界与给定函数集的容量的关系, 从概率空间实质性地推广到了比它更大的 Sugeno 测度空间上.

致谢 非常感谢香港理工大学计算学系 Yeung 教授给予的指导与帮助.

## 参 考 文 献

1 Vapnik V N. Statistical Learning Theory. New York: A Wiley-Interscience Publication, 1998
2 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
3 Vapnik V N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10(5): 988~999 [DOI]
4 边肇祺, 张学工. 模式识别. 北京: 清华大学出版社, 1999
5 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): 32~44
6 柳回春, 马树元. 支持向量机的研究现状. 中国图像图形学报, 2002, 7(6): 618~623
7 郑红军, 周 旭, 毕笃彦. 统计学习理论与支持向量机概述. 现代电子技术, 2003, 4: 59~61
8 Raudys S. How good are support vector machines? Neural Networks, 2000, 13(1): 17~19 [DOI]
9 Francis E H, Tay F E H, Cao L J. Application of support vector machines in financial time series forcasting. Omega, 2001, 29: 309~317[DOI]
10 Tsai C F. Training support vector machines based on stacked generalization for image classification. Neurocomputing, 2005, 64: 497~503[DOI]
11 Tomonori K, Shigeo A. Comparison between error correcting output codes and fuzzy support vector machines. Pattern Recognition Letters, 2005, 26(12): 1937~1945[DOI]
12 Cawley G C, Talbot N L C. Improved sparse least-squares support vector machines. Neurocomputing, 2002, 48(1-4): 1025~1031[DOI]
13 Zhang Y Q, Shen D G. Design efficient support vector machine for fast classification. Pattern Recognition, 2005, 38(1): 157~161[DOI]
14 哈明虎, 王瑞省, 张 琳. 模糊积分在物流系统工程中的应用. 模糊系统与数学, 2004, 18(4): 72~76
15 Wang Z Y, George J K. Fuzzy Measure Theory. New York: Plenum Press, 1992
16 Weber S. Two integrals and some modified versions critical remarks. Fuzzy Sets and Systems, 1986, 20: 97~105[DOI]
17 哈明虎, 吴从炘. 模糊测度与模糊积分理论. 北京:科学出版社, 1998