Rademacher 复杂度在统计学习理论中的研究: 综述

吴新星 [1,2,3] 张军平 [2,3]

摘 要 假设空间复杂性是统计学习理论中用于分析学习模型泛化能力的关键因素. 与数据无关的复杂度不同, Rademacher 复杂度是与数据分布相关的, 因而通常能得到比传统复杂度更紧致的泛化界表达. 近年来, Rademacher 复杂度在统计学习理论泛化能力分析的应用发展中起到了重要的作用. 鉴于其重要性, 本文梳理了各种形式的 Rademacher 复杂度及其与传统复杂度之间的关联性, 并探讨了基于 Rademacher 复杂度进行学习模型泛化能力分析的基本技巧. 考虑样本数据的独立同分布和非独立同分布两种产生环境, 总结并分析了 Rademacher 复杂度在泛化能力分析方面的研究现状. 展望了当前 Rademacher 复杂度在非监督框架与非序列环境等方面研究的不足, 及其进一步应用与发展.

关键词 机器学习, 统计学习理论, 泛化界, Rademacher 复杂度

引用格式 吴新星, 张军平. Rademacher 复杂度在统计学习理论中的研究: 综述. 自动化学报, 2017, **43**(1): 20−39

DOI 10.16383/j.aas.2017.c160149

Researches on Rademacher Complexities in Statistical Learning Theory: A Survey

WU Xin-Xing[1,2,3] ZHANG Jun-Ping[2,3]

Abstract Measuring the complexity of a hypothesis space plays a crucial role in statistical learning theory. Unlike those data-independent complexities, Rademacher complexity, which is data-dependent, can attain a much more compact generalization representation. In recent years, Rademacher complexity has attracted more attention and found broad applications in the development of statistical learning theory. Because of its importance, in this paper we review several complexity measures of function classes and their relations with Rademacher complexities. Next, we describe the techniques of Rademacher complexities in generalization analysis. Then, we present the recent researches of Rademacher complexity learning bounds for independent and identical distribution (i.i.d.) and non-independent and identical distribution (non-i.i.d.). Finally, we discuss the potential issues and possible directions of Rademacher complexities in statistical learning theory.

Key words Machine learning, statistical learning theory, generalization bounds, Rademacher complexities

Citation Wu Xin-Xing, Zhang Jun-Ping. Researches on Rademacher complexities in statistical learning theory: a survey. *Acta Automatica Sinica*, 2017, **43**(1): 20−39

机器学习是从人工智能中分离出来, 应用驱动的一门学科. 近年来随着大数据时代的来临, 机器学习备受各行各业如计算机视觉、自动控制、生物特征识别、数据分析、互联网、多媒体、社会安全等的广泛关注.

机器学习不仅在应用领域取得了不胜枚举的成功, 在理论方面也在不断完善, 尤其是在统计学习理论的研究方面. 从学习的本质来看, 机器学习旨在基于已知样本数据集, 通过学习来构造逼近真实分布如函数依赖关系或内在规律的学习模型. 其中, 学习模型对未知数据的预测或泛化能力是统计学习理论的主要关注目标.

一般来说, 按训练样本是否有标签可将机器学习大致分为有监督学习、半监督学习和无监督学习三个主要学习框架. 在有监督学习框架下, 训练样本数据集被看作是一组来自某一函数 ($h_{\rm P}$) 依赖关系的输入输出对. 机器学习的目的就是基于该样本数据集, 通过学习的方式, 从给定的函数集 ($\mathcal{H}$, 一般称假设空间) 中找到某一函数 ($\hat{h}$), 使其尽可能地逼近目标函数 ($h_{\rm P}$). 本文讨论的内容主要是在有监督学习框架下展开的.

学习理论的研究最早出现在数学领域中. Tikhonov 和 Arsenin 指出多数数据预测为病态问题 (Ill-posed problem), 需要引入某些限制来保证问题的良态化[1]. 而在机器学习领域, 也存在类似的

收稿日期 2016-02-18 录用日期 2016-07-11
Manuscript received February 18, 2016; accepted July 11, 2016
国家自然科学基金 (61673118, 61273299), 上海浦江人才计划 (16PJD009), 上海市人才发展资金 (201629) 资助
Supported by National Natural Science Foundation of China (61673118, 61273299), Shanghai Pujiang Program (16PJD009), and Shanghai Talents Development Funds (201629)
1. 上海电子信息职业技术学院计算机应用系 上海 201411 2. 复旦大学计算机科学技术学院 上海 200433 3. 上海市智能信息处理重点实验室 上海 200433
1. Department of Computer, Shanghai Technical Institute of Electronics and Information, Shanghai 201411 2. School of Computer Science, Fudan University, Shanghai 200433 3. Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200433

问题. 如果 $\mathcal{H}$ 中函数自由参数过于简单, 那么, 在给定样本数据集上, $\hat{h}$ 将可能不会很好地逼近 h_{P}, 即, 欠拟合; 如果 $\mathcal{H}$ 中函数自由参数过于复杂, 那么, 在给定样本数据集之外的测试数据集或是未知数据集上, $\hat{h}$ 也将可能不会很好地逼近 h_{P}, 即, 过拟合. 因此, 为了权衡 $\mathcal{H}$ 中函数自由参数的复杂性与学习模型的泛化能力, 一些有关参数复杂性的模型准则陆续被一些学者提出, 试图通过控制模型参数的复杂性来获得理想的学习模型, 如 Akaike 提出的用于控制神经网络参数个数的 AIC 准则 (An information criterion)[2], Akaike[3] 和 Schwarz[4] 用于控制多元高斯混合模型参数个数的 BIC 准则 (Bayesian information criterion), Kolmogorov 等从编码角度提出的 Kolmogorov 复杂度 (Kolmogorov complexity, KC)[5−8] 以及基于此发展的针对聚类的最小信息长度 (Minimum message length, MML) 模型[9] 等. 但在实际应用中, 基于这些准则来学习得到模型时将碰到困难, 会出现如 “维数灾难”、实际计算困难等问题. 在 20 世纪 60 至 70 年代, Vapnik 和 Chervonenkis 对复杂性的定义进行了重新思考, 发现学习模型的泛化能力主要取决于学习所基于的假设空间的复杂性, 而假设空间的复杂性本质上不同于模型中自由参数的复杂性或是个数[10−12]. 因此, 如何定义假设空间的复杂性并度量其复杂性, 对学习模型泛化能力的控制和估计显得极为重要. 同时, 在 Popper 哲学思想[13] 的影响下, Vapnik 和 Chervonenkis 提出了 Vapnik-Chervonenkis (VC) 熵、生长函数和 VC 维等一系列著名的复杂性度量, 并将 VC 维用于刻画和度量假设空间的复杂性, 从而来估计和控制学习模型的泛化能力. 同时, 在 1984 年 Valiant 提出了概率近似正确 (Probabilistic approximate correct, PAC) 的概念, 他指出学习模型将以概率 $1-\delta$ $(0<\delta<1)$ 的方式逼近真实分布[14]. 以上这些努力奠定了统计学习理论的基础. 之后, Kearns 等又将 VC 维由示性函数推广到了一般实值函数, 提出了 Fat-shattering 维的概念[15−16]. 但是, 由于 VC 维是在假设空间上引入额外的度量, 并且 VC 维与所给的样本数据集 (分布) 无关或是说数据独立 (Data independent) 等特点, 使其在进行学习模型泛化能力分析方面显得过于保守.

Shawe-Taylor 等在对统计学习理论深入研究中注意到数据相依 (Data dependent) 复杂性度量的重要性[17], Koltchinskii 等将 Rademacher 复杂度引入到了统计学习理论学习模型的泛化能力分析研究中, 发现由于 Rademacher 复杂度在对假设空间进行复杂性度量时不需要引入额外度量, 并且依赖于特定样本数据集 (分布) 或是说数据相依[18−19], 这是明显不同于 VC 维要求的数据独立、一致分布等较强条件. 因此, Rademacher 复杂度能以一种比 VC 维更加紧致的方式关联经验过程极大泛函, 从而, 在学习模型的泛化能力分析方面能得到较好的预测精度及稍快的收敛速度[18−19]. Bartlett 等通过进一步研究发现, 对学习模型泛化能力起关键作用的往往不是整个假设空间中的函数, 而是那些具有较小方差的函数所构成的假设空间的子空间[20−23], 基于此发现给出了局部 Rademacher 复杂度的概念. 此后, Rademacher 复杂度在统计学习理论学习模型泛化能力分析研究方面开始得到高度的关注和广泛的应用.

国内学者在假设空间复杂性与学习模型泛化性能研究方面, 也做了许多相关工作. 如, 西安交通大学的徐宗本院士和西北大学的张海教授在文献 [24] 中主要从学习算法稳定性和泛化性角度综述了现有稳定性框架之间的关系, 湖北汽车工业学院的胡政发硕士在文献 [25] 中基于 VC 维和 Banach 空间 $L-$范数研究了假设空间的复杂性度量, 湖北大学的陈将宏硕士及其导师李落清教授在文献 [26] 中基于覆盖数 (具体定义见附录 A 覆盖数)、VC 维和 Rademacher 复杂度讨论了支持向量机 (Support vector machine, SVM) 的泛化性能, 武汉大学的雷云文博士及其导师丁立新教授等在文献 [27−32] 中基于 Rademacher 复杂度讨论了多核学习 (Multiple kernel learning, MKL)、多模态 (Multi-modal) 学习、神经网络学习、自由节点样条 (Free knot spline, FKS) 学习等几种不同学习算法的泛化性能, 北京大学的许超教授及其合作者在文献 [33] 中运用 Rademacher 复杂度讨论了多标签学习算法的泛化性能, 南京大学的高尉博士及其导师周志华教授在文献 [34] 中基于 Rademacher 复杂度分析了深度神经网络中 Dropout 的泛化界等.

本文主要是从综述的角度, 系统地总结了现有的各种形式 Rademacher 复杂度[18−23, 27, 32, 34−43] 及其与传统复杂度之间的关联性[20−21, 27−29, 32, 36−37, 39−41, 44−50], 并概述了当前 Rademacher 复杂度在统计学习理论泛化界分析方面的应用成果[19, 27−28, 30−33, 36−43, 47−49, 51−56]. 本文的主要内容安排如下: 第 1 节简单介绍了复杂度研究发展的历史背景; 第 2 节给出了本文后续讨论要用的一些定义和符号; 第 3 节总结了各种形式的 Rademacher 复杂度, 并进一步分析了 Rademacher 复杂度的特点及基本的分析处理技巧; 第 4 节讨论了各种形式的 Rademacher 复杂度与传统复杂度之间的关系; 第 5 节在样本数据集为独立同分布的假设下, 讨论 Rademacher 复杂度在泛化界分析方面的研究现状; 第 6 节在样本数据集为非独立同分布的假设下, 讨论 Rademacher 复杂度在泛化界分析方面的研究现状; 第 7 节对本文进行小结, 并展望了 Rademacher 复杂度在泛化界分析方面研究的不足

与进一步发展应用.

1 定义和模型

在本节中, 我们将对本文后续讨论中要用到的一些符号和基本的学习模型分别进行约定与说明.

1) 集合 $\mathcal{X}$ 表示输入空间, 集合 $\mathcal{Y}$ 表示输出空间, 集合 $\mathcal{Z}$ ($\mathcal{Z}=(\mathcal{X}\times\mathcal{Y})$) 表示样本数据集或称样本空间. 记 $\mathcal{P}(\mathcal{Y}^{\mathcal{X}})$ 为 $\mathcal{Y}^{\mathcal{X}}$ 的幂集, 集合 $\mathcal{H}$ 为假设空间且 $\mathcal{H}\in\mathcal{P}(\mathcal{Y}^{\mathcal{X}})$. 记 $Q:(\mathcal{X}\times\mathcal{Y})\times\mathcal{P}(\mathcal{Y}^{\mathcal{X}})\to[0,+\infty]$ 为损失函数, 该函数量化了 $h(\in\mathcal{Y}^{\mathcal{X}})$ 在 $\mathcal{Z}$ 上的学习效果.

2) 集合 $\mathcal{F}:=\{Q(z,h):h\in\mathcal{H},z\in\mathcal{Z}\}$.

3) 集合 $\mathbf{N}$ 为自然数集. 对任意 $n\in\mathbf{N}$, 记 $\mathbf{N}_n=\{1,2,\cdots,n\}$.

4) 集合 $\mathbf{R}$ 为实数集.

5) 记 $f\in\mathcal{F}$, $\|f\|_{L_q(\mu)}:=\left(\int|f|^q\mathrm{d}\mu\right)^{1/q}$.

6) 记 $\|\cdot\|$ 为 $\mathbf{R}^{d\times d}$ ($d\in\mathbf{N}$) 上范数 (如, Frobenius 范数等), 对偶范数定义如下:

$$\|M\|_*:=\sup\{\langle X,M\rangle:X\in\mathbf{R}^{d\times d},\|X\|\le 1\}\quad(1)$$

其中, 内积 $\langle X,M\rangle:=\mathrm{tr}(X^{\mathrm{T}}M)$, tr 表示矩阵的迹.

7) 任意 $d,m\in\mathbf{N}$, $d\times(md)$ 对称矩阵空间为

$$S^{d\times(md)}:=\{(M^1,\cdots,M^m):M^l\in\mathbf{R}^{d\times d},\ (M^l)^{\mathrm{T}}=M^l,l\in\mathbf{N}_m\}\quad(2)$$

定义 1. 对于模型 $h\in\mathcal{H}$, 其泛化误差为[1]

$$\mathrm{E}(h):=\int_{\mathcal{Z}}Q(z(\omega),h)\mathrm{P}(\mathrm{d}\omega)\quad(3)$$

定义 2. 称 h_{P} 为目标函数, 若

$$h_{\mathrm{P}}:=\arg\min_{h\in\Xi}\mathrm{E}(h)\quad(4)$$

其中, $\Xi=\{h|h^{-1}(B)\in\sigma(\mathcal{X})^{\,2},\forall B\subset\mathcal{Y}\}$. h_{P} 表示 Q 意义下预测能力最好的模型.

但在实际应用中, 上面定义 2 中的 P 一般未知, 泛化误差 $\mathrm{E}(h)$ 往往实际无法计算. 所以, 通常采用如下基于给定样本数据集的经验误差

$$\mathrm{E}_n(h):=\frac{1}{n}\sum_{i=1}^{n}Q(z_i,h)\quad(5)$$

并通过取 $\min_{h\in\mathcal{H}}\mathrm{E}_n(h)$ 的方式, 来得到对 h_{P} 的逼近 $\hat{h}$. 这里, z_i, $i=1,2,\cdots,n$, 表示给定样本数据集.

一般地, 将式 (5) 称为经验风险, 基于经验风险最小化来求解得到估计模型的学习策略, 称为经验风险最小化 (Empirical risk minimization, ERM). 由于多数机器学习问题属于病态问题, 常在经验风险后增加与模型复杂度相关的正则项或惩罚项, 并通过最小化引入正则项或惩罚项后的模型来实现良态解. 如下面的结构风险最小化 (Structural risk minimization, SRM) 策略

$$\hat{h}:=\arg\min_{h\in\mathcal{H}_k,k\in\mathbf{N}}(\mathrm{E}_n(h)+\mathrm{pen}(k,n)),\quad\mathcal{H}_k\subset\mathcal{H}_{k+1},\ k=1,2,\cdots\in\mathbf{N}\quad(6)$$

和正则化 (Regularizer) 策略

$$\hat{h}:=\arg\min_{h\in\mathcal{H}}(\mathrm{E}_n(h)+\lambda\Omega(h))\quad(7)$$

注 1[58]**.** 式 (6) 中的 $\mathrm{pen}(k,n)$ 称为惩罚项, 与假设空间 $\mathcal{H}_k$ 的复杂度及样本数据集的容量有关. $\Omega(h)$ 为正则化算子, λ 为正则化参数: λ 较小时, 得到的学习模型拟合能力较强; λ 较大时, 得到的学习模型形式较简单.

在统计学习理论中, 通过式 (5)、(6) 或 (7) 等学习策略获得预测模型 $\hat{h}$ 后, 一般基于如下的偏差–方差分解来研究 $\hat{h}$ 的泛化能力:

$$\mathrm{E}(\hat{h})-\mathrm{E}(h_{\mathrm{P}})=\underbrace{\mathrm{E}(\hat{h})-\mathrm{E}(h^*)}_{\text{估计误差 (方差)}}+\underbrace{\mathrm{E}(h^*)-\mathrm{E}(h_{\mathrm{P}})}_{\text{逼近误差 (偏差)}}\quad(8)$$

其中, $h^*=\arg\min_{h\in\mathcal{H}}\mathrm{E}(h)$.

本文的讨论所关注的估计误差, 主要是通过如下公式

$$\mathrm{E}(\hat{h})-\mathrm{E}(h^*)\le\underbrace{\mathrm{E}(\hat{h})-\mathrm{E}_n(\hat{h})}_{\text{算法稳定性}}+\underbrace{\mathrm{E}_n(h^*)-\mathrm{E}(h^*)}_{\text{一致偏差}}\quad(9)$$

将估计误差分解为算法稳定性和一致偏差等两方面, 本文重点是关于一致偏差的分析研究.

另外, 本文在综述 Rademacher 复杂度在统计学习理论泛化界分析方面的研究时, 考虑到一致性和易读性, 对分散于各文献资料中的相关结果, 按本节的约定和说明作了符号统一化处理.

2 复杂度指标

我们知道假设空间的复杂性对学习模型的泛化能力有重要影响. 在本节中, 将总结分析各种形式的 Rademacher 复杂度, 并讨论应用 Rademacher 复杂度进行假设空间复杂性刻画及学习模型泛化界分析的基本技巧和特点.

Rademacher 复杂度以德国数学家 Hans Adolph Rademacher (1892∼1969) 命名. 早在 Rademacher 复杂度应用于统计学习理论之前, 一些学者就已经对 Rademacher 复杂度的一些特性展

[1] 式 (3) 有时也记为 $\mathrm{E}[Q]$.

[2] 其中, $\sigma(\mathcal{X})$ 表示由 $\mathcal{X}$ 生成的 σ 代数[57].

开过深入的讨论和研究. 如, Rademacher 复杂度的熵积分, Rademacher 复杂度的压缩准则等[59−61].

在统计学习理论的早期研究中, Vapnik 和 Chervonenkis 提出的 VC 维是一种通过度量假设空间复杂性来控制学习模型泛化能力的分析方法. 在 VC 维理论提出之后, 一些学者开始注意到, 基于 VC 维的泛化界是数据独立、分布无关的. 如, $0 < \delta < 1$, 有如下式

$$\sup_{h\in\mathcal{H}} |\mathrm{E}_n(h) - \mathrm{E}(h)| \le \sqrt{\frac{d\log n}{n}} \tag{10}$$

对一切样本数据集分布 P 至少以概率 $1-\delta$ 一致地成立, 即, 对任意的数据集分布 P, 式 (10) 至少以概率 $1-\delta$ 成立. 这里, d 表示 $\mathcal{H}$ 的 VC 维. 从而, 导致通过 VC 维分析得到的学习模型泛化界过于保守. 在式 (10) 中的左边, 如果放弃一致性条件限制, 而是使其依赖于特定样本数据集的分布, 那么, 式 (10) 中右侧的界 (不再数据独立) 将会比原来的界更紧致[62]. 于是, 一些学者开始研究关联实际样本数据集分布的复杂性度量[17]. Koltchinskii 等将 Rademacher 复杂度引入到统计学习理论的泛化界分析研究[18−19], 发现 Rademacher 复杂度作为一种依赖特定数据集分布的复杂性度量, 可以有效地用于学习模型中假设空间复杂性的刻画, 改进学习模型的泛化界. 之后, 经过许多学者的进一步工作, Rademacher 复杂度被广泛应用于统计学习理论的泛化界分析研究.

2.1 Rademacher 复杂度

一般地, Rademacher 复杂度分为两种情形: Rademacher 复杂度和经验 Rademacher 复杂度. Rademacher 复杂度及经验 Rademacher 复杂度的定义分别如下:

定义 3. 假设 X_i, $i = 1, 2, \cdots, n$ 为定义在 $\mathcal{X}$ 上且服从分布 P 的独立同分布 (Independent and identical distribution, i.i.d.) 随机变量序列. 令 $\sigma_i \in \{-1, +1\}$, $i = 1, 2, \cdots, n$ 为 Rademacher 随机变量序列. 记函数 $f : \mathcal{X} \to \mathbf{R}$, $\mathcal{F}$ 是 f 的集合. 记

$$R_n f := \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) \tag{11}$$

和

$$R_n\mathcal{F} := \sup_{f\in\mathcal{F}} R_n f \tag{12}$$

称 $\{R_n f : f \in \mathcal{F}\}$ 为 Rademacher 过程, Rademacher 复杂度记为

$$\mathrm{E}(R_n\mathcal{F}) := \mathrm{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right] \tag{13}$$

而

$$\mathrm{E}_\sigma(R_n\mathcal{F}) := \mathrm{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) | X_1, \cdots, X_n\right] \tag{14}$$

称经验 Rademacher 复杂度. 为便于后面的讨论, 约定将式 (13) 和 (14) 中定义的 Rademacher 复杂度分别称为经典 Rademacher 复杂度和经典经验 Rademacher 复杂度.

经典 Rademacher 复杂度与经典经验 Rademacher 复杂度的关系, 本质上也是期望量与经验量之间的关系, 因而, 可以通过概率集中不等式[63]等数学技术对其中的关系进行刻画. 同时, 在经典与经典经验 Rademacher 复杂度定义中, 涉及的 n 个样本数据集 (分布) 是给定的, 即, 基于 Rademacher 复杂度的泛化界分析依赖于具体学习问题上的数据集分布, 有点类似于为该学习问题 “量身定制” 的[64]. 对于具体的学习模型泛化能力的分析, 如[65], 取 $\mathcal{Y} = \{-1, +1\}$, $0 < \delta < 1$, 任意的 $h \in \mathcal{H}$, 可以得到如下两式

$$\mathrm{E}_n(h) - \mathrm{E}(h) \le 2\mathrm{E}(R_n\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{15}$$

和

$$\mathrm{E}_n(h) - \mathrm{E}(h) \le 2\mathrm{E}_\sigma(R_n\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{16}$$

分别以至少概率 $1-\delta$ 成立. 从式 (13) 和 (14) 可以看出, 上面的不等式 (15) 和 (16) 右侧与具体学习问题的数据分布有关. 同时, 后面也将会看到, 不等式 (15) 和 (16) 的证明采用了比不等式 (10) 的证明更精细的处理技巧. 因而, 基于 Rademacher 复杂度的分析通常会得到比 VC 维更紧致的泛化界, 在实际中也更易于计算.

经典 (经验) Rademacher 复杂度原理上是通过引入 Rademacher 随机变量而得到的一种对函数空间复杂性的度量, 本质上从量化角度揭示了假设空间关联随机噪声的能力. Koltchinskii 在文献 [18] 中指出: 关于经典 Rademacher 复杂度惩罚项的计算与基于随机重新标记 (Randomly relabeled) 样本数据集, 并通过 ERM 策略来获得学习模型之间具有等价性, 即,

$$\sup_{A\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i I_{\{Y_i \ne I_A(X_i)\}} \Leftrightarrow \inf_{A\in\mathcal{H}} \sum_{i=1}^{n} I_{\{\widetilde{Y}_i \ne I_A(X_i)\}} \tag{17}$$

其中, $\sigma_i = -1$, $\widetilde{Y}_i = 0$, 且 $\sigma_i = 1$, $\widetilde{Y}_i = 1$.

证明. 一方面

$$\sum_{i=1}^{n}\sigma_i I_{\{Y_i\neq I_A(X_i)\}}=\sum_{\sigma_i=1,Y_i=1}(1-I_A(X_i))+\sum_{\sigma_i=1,Y_i=0}I_A(X_i)-\sum_{\sigma_i=-1,Y_i=1}(1-I_A(X_i))-\sum_{\sigma_i=-1,Y_i=0}I_A(X_i)=\sum_{i=1,2,\cdots,n,Y_i=1}\sigma_i+\sum_{i=1}^{n}\sigma_i I_A(X_i)\tag{18}$$

另一方面, 注意到

$$\sum_{i=1}^{n}\sigma_i I_A(X_i)=\sum_{\sigma_i=1}I_A(X_i)-\sum_{\sigma_i=-1}I_A(X_i)=-\sum_{\sigma_i=-1}I_A(X_i)-\sum_{\sigma_i=1}(1-I_A(X_i))+num\tag{19}$$

其中, num 表示集合 $\{i\in\mathbf{N}_n:\sigma_i=+1\}$ 中的元素个数.

对式 (19) 取 sup, 实际上就是对下式取 inf

$$\sum_{\sigma_i=-1}I_A(X_i)+\sum_{\sigma_i=1}(1-I_A(X_i))=\sum_{i=1}^{n}I_{\{\widetilde{Y}\neq I_A(X_i)\}}\tag{20}$$

□

注 2[18]**.** 上面的讨论说明: Rademacher 复杂度可以看作是衡量集类 $\mathcal{H}$ (假设空间) 分离能力的一种度量, 即, $\sup_{A\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i I_{\{Y_i\neq I_A(X_i)\}}$ 很大时, 通过 $\mathcal{H}$ 来对样本数据集进行分类, 即使样本数据集中数据标签是随机标记的, 错误也会很小. 另一方面也说明, 过大的 $\sup_{A\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i I_{\{Y_i\neq I_A(X_i)\}}$ 是不恰当的, 因为合理的 $\mathcal{H}$ 是要能对正确标记的样本数据集中数据进行分类, 而不是对任意随机标记的样本数据集中数据进行分类.

2.2 经典 Rademacher 复杂度的推广

在本小节中, 将介绍和讨论在经典 (经验) Rademacher 复杂度基础上推广提出的各种形式 Rademacher 复杂度.

2.2.1 局部 Rademacher 复杂度

一批学者注意到, 一般地, 基于经典 Rademacher 复杂度分析得到的泛化界为 $\mathrm{O}(n^{-1/2})$, 但在实际中这往往是次优的[20]. 同时, 也发现在统计学习理论中对学习模型泛化性能起决定性作用的, 不是全局的假设函数空间, 而是那些具有较小方差的函数所构成的假设空间的子空间. 基于此发现, Bartlett 等提出了局部 Rademacher 复杂度的概念 — 考察原假设空间中具有较小方差的函数子集[20−23], 存在 $r\in\mathbf{R}$ 且 $r>0$,

$$\mathrm{E}R_n\{f\in\mathcal{F}:\mathrm{P}f^2\leq r\}\tag{21}$$

2.2.2 Rademacher chaos 复杂度

以 SVM 为代表的核方法是机器学习中解决非线性模式分析问题的一种有效方法. MKL 是为了解决核方法在核函数选取等方面的不足而被引入到机器学习中[66], 主要从具体样本数据集出发, 在给定的核函数集中寻找最能表达对应样本数据集的核函数. 因此, 与单核学习相比有更好的适应性和灵活性. 但这不能保证基于 MKL 所得到的学习模型一定有更好的泛化能力, 因为还需要同时考察 MKL 对应假设空间的复杂性. 于是, Ying 等[49] 将 $U-$过程 (见附录 A $U-$过程) 中的二阶 Rademacher chaos 复杂度引入到统计学习理论中, 用于刻画 MKL 对应假设空间的复杂性, 分析 MKL 算法的泛化性能. 下面, 给出二阶 (经验) Rademacher chaos 复杂度的相关定义.

定义 4. 假设 X_i, $i=1,2,\cdots,n$ 为定义在 $\mathcal{X}$ 上且服从分布 P 的 i.i.d. 随机变量序列. 令 $\sigma_i\in\{-1,+1\}$, $i=1,2,\cdots,n$ 为 Rademacher 随机变量序列. 记函数 $f:\mathcal{X}\times\mathcal{X}\to\mathbf{R}$, $\mathcal{F}$ 是 f 的集合. 记

$$\mathcal{U}_n(f):=\frac{1}{n}\sum_{i<j\leq n}\sigma_i\sigma_j f(X_i,X_j)\tag{22}$$

和

$$\mathcal{U}_n(\mathcal{F}):=\sup_{f\in\mathcal{F}}\mathcal{U}_n(f)\tag{23}$$

则称

$$\left\{\mathcal{U}_n(f):=\frac{1}{n}\sum_{i<j\leq n}\sigma_i\sigma_j f(X_i,X_j),\quad f\in\mathcal{F}\right\}\tag{24}$$

为二阶齐次 Rademacher chaos 过程.

Rademacher chaos 复杂度记为

$$\mathrm{E}\mathcal{U}_n(\mathcal{F}):=\mathrm{E}\left[\sup_{f\in\mathcal{F}}|\mathcal{U}_n(f)|\right]\tag{25}$$

而

$$\mathrm{E}_\sigma\mathcal{U}_n(\mathcal{F}):=\mathrm{E}_\sigma\left[\sup_{f\in\mathcal{F}}|\mathcal{U}_n(f)|\right]=\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}|\sum_{i<j\leq n}\sigma_i\sigma_j f(X_i,X_j)\|X_1,\cdots,X_n\right]\tag{26}$$

称为经验 Rademacher chaos 复杂度.

注 3. 实际上, 经典的 Rademacher 复杂度可以看作是一阶 Rademacher chaos 复杂度[32].

2.2.3 单模态 Rademacher 复杂度

度量或相似性学习 (Metric or similarity learning) 是计算机视觉和模式识别等领域的热点研究问题之一, 是诸如 K-means 聚类分析、K– 近邻分类等学习算法的基础, 主要基于训练样本数据集学习得到有效的距离度量来衡量目标之间的相似性. Cao 等为了分析度量学习的泛化性能, 构造了用于度量学习的 Rademacher 复杂度[36]. 由于文献 [36] 中考虑的数据描述视角是单一的, 为了区分后面讨论的多模态度量学习情形的 Rademacher 复杂度[27], 我们将其称为单模态度量学习下的 Rademacher 复杂度.

定义 5. 假设 X_i, $i=1,2,\cdots,n$ 为 $\mathcal{X}$ ($\mathcal{X}\subset\mathbf{R}^d$) 上的 i.i.d. 随机变量序列. 令 $\{\sigma_i\}_{i=1}^{\lfloor\frac{n}{2}\rfloor}$ 为 Rademacher 随机变量序列. 则单模态度量学习下 Rademacher 复杂度定义为

$$\mathrm{E}R_n^{\mathrm{single}}:=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathrm{E}\left\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(X_i-X_{\lfloor\frac{n}{2}\rfloor+i})(X_i-X_{\lfloor\frac{n}{2}\rfloor+i})^{\mathrm{T}}\right\|_*\tag{27}$$

而

$$\mathrm{E}_\sigma R_n^{\mathrm{single}}:=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathrm{E}\left[\left\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(X_i-X_{\lfloor\frac{n}{2}\rfloor+i})(X_i-X_{\lfloor\frac{n}{2}\rfloor+i})^{\mathrm{T}}\right\|_*\Bigg|X_1,X_2,\cdots,X_{\lfloor\frac{n}{2}\rfloor+\lfloor\frac{n}{2}\rfloor}\right]\tag{28}$$

称为单模态度量学习下经验 Rademacher 复杂度.

2.2.4 多模态 Rademacher 复杂度

由于实际中涉及的数据信息往往是来自多个异构数据源, 通过多个视角来描述. 如, 音乐网站的歌曲可以通过节奏和音色等声音特征、标签和歌词等语义特征、协同过滤 (Collaborative filtering) 和人物传记等社会特征描述[67−68]. 因此, Lei 等在文献 [36] 中单模态 Rademacher 复杂度基础上, 进一步提出了适用于多模态度量学习研究的 Rademacher 复杂度[27, 32], 用于多模态度量学习的泛化能力分析.

定义 6. 假设 X_i, $i=1,2,\cdots,n$ 为 $\mathcal{X}$ ($\mathcal{X}\subset\mathbf{R}^{md}$) 上的 i.i.d. 随机变量序列. 令 $\{\sigma_i\}_{i=1}^{\lfloor\frac{n}{2}\rfloor}$ 为 Rademacher 随机变量序列. 记 $\mathcal{M}\subset S^{d\times(md)}$, 则多模态度量学习下 Rademacher 复杂度为

$$\mathrm{E}R_n^{\mathrm{multi}}\mathcal{M}:=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathrm{E}\left[\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i d_M(X_i,X_{\lfloor\frac{n}{2}\rfloor+i})\right]\tag{29}$$

而

$$\mathrm{E}_\sigma R_n^{\mathrm{multi}}\mathcal{M}:=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathrm{E}\left[\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i d_M(X_i,X_{\lfloor\frac{n}{2}\rfloor+i})\Bigg|X_1,X_2,\cdots,X_{\lfloor\frac{n}{2}\rfloor+\lfloor\frac{n}{2}\rfloor}\right]\tag{30}$$

称为多模态度量学习下经验 Rademacher 复杂度. 其中, $d_M(X_i,X_{\lfloor\frac{n}{2}\rfloor+i}):=\sum_{k=1}^m(X_i^k-X_{\lfloor\frac{n}{2}\rfloor+i}^k)^{\mathrm{T}}\times M^k(X_i^k-X_{\lfloor\frac{n}{2}\rfloor+i}^k)$, $X_i^k\in\mathbf{R}^d$, $M^k\in S^{d\times d}$, $k=1,2,\cdots,m$.

2.2.5 Dropout Rademacher 复杂度

在训练深度神经网络模型时, 如果训练样本较少, 为防止模型过拟合, Hinton 等提出了 Dropout 技术, 其基本思想是: 训练模型时, 以一定的概率让网络中某些节点不工作[69]. Wan 等利用经典 Rademacher 复杂度对 Dropout 的泛化界进行过研究[70]. Gao 等注意到经典 Rademacher 复杂度仅仅是和训练样本有关, 而没有刻画出 Dropout 技术在训练模型时随机改变网络结构的特性, 于是便提出了 Dropout Rademacher 复杂度[34].

定义 7. 假设 $X_i, i=1,2,\cdots,n$ 为定义在 $\mathcal{X}$ 上且服从分布 P 的 i.i.d. 随机变量序列, 令 $\sigma_i\in\{-1,+1\}$, $i=1,2,\cdots,n$ 为 Rademacher 随机变量序列. 记 $R_s=\{\boldsymbol{r}^s=(r_1,r_2,\cdots,r_s):r_j\sim\mathrm{Bern}(1,p),j=1,2,\cdots,s\}$[3], 其中, s 依赖于具体的 (深度) 神经网络和不同类型的 Dropout (如, 隐藏层节点的 Dropout、权重的 Dropout 等)[4]. 令 $\boldsymbol{r}_i^s=\{r_{i1},r_{i2},\cdots,r_{is}\}\in R_s$, $i=1,2,\cdots,n$ 为对应 n 个样本 Dropout 概率的相关向量序列. 记函数 $f_s:\mathcal{X}\times R_s\to\mathbf{R}$, $\mathcal{F}_s$ 是 f 的集合. Dropout Rademacher 复杂度定义为

$$\mathrm{E}R_n^{\mathrm{Dropout}}\mathcal{F}_s:=\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(X_i,\boldsymbol{r}_i^s)\right]\tag{31}$$

而

[3] $\mathrm{Bern}(1,p)$ 表示参数为 p 的伯努利分布. 如果随机变量 $X\sim\mathrm{Bern}(1,p)$, 则 X 分别以概率 p 取值 1, 以概率 $1-p$ 取值 0.

[4] 如果是隐藏层节点的 Dropout, 那么, $r_j=0$ 表示某节点不工作.

$$\mathrm{E}_\sigma R_n^{\text{Dropout}}\mathcal{F}_s := \mathrm{E}\Big[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i,\boldsymbol{r}_i^s) \Big| X_1,\cdots,X_n;\boldsymbol{r}_1^s,\cdots,\boldsymbol{r}_n^s\Big] \tag{32}$$

即是经验 Dropout Rademacher 复杂度.

上面介绍的 Rademacher 复杂度的几种推广形式都是基于样本数据集产生环境为独立同分布的假定. 但是, 在具体实际应用中存在着大量非独立同分布 (Non-independent and identical distribution, non-i.i.d.) 的数据, 如, 股票市场预测、天气预报、垃圾邮件检测等. 因此, 在 non-i.i.d. 情形下 (如, 平稳 β-mixing、非平稳 β-mixing、独立不同分布、鞅等) 对学习模型的泛化性能进行研究就显得非常必要和有实际意义. 目前, 已经有学者提出了几类特殊 non-i.i.d. 情形的 Rademacher 复杂度 (如, 块 (Block) Rademacher 复杂度、独立不同分布 Rademacher 复杂度、序列 (Sequential) Rademacher 复杂度等), 用于 non-i.i.d. 情形下学习模型的泛化性能分析.

2.2.6 块 Rademacher 复杂度

为了研究样本数据集产生环境为平稳 β-mixing 情形的学习模型泛化能力, Mohri 等在文献 [38] 中将经典 Rademacher 复杂度推广到了平稳 β-mixing 情形, 给出了块 Rademacher 复杂度的定义. Kuznetsov 等进一步地完善文献 [38] 中块 Rademacher 复杂度, 并将其应用到了非平稳时间序列预测[42−43]. Mohri 和 Kuznetsov 等在将块 Rademacher 复杂度用于分析估计学习模型泛化能力过程中, 都使用到了一种独立块 (Independent blocks, IB) 的技巧, 从而将对于平稳 β-mixing 和非平稳 β-mixing 等环境下的分析研究与原有经典 i.i.d. 环境下的研究结果建立关联. 下面给出块 Rademacher 复杂度定义 (具体有关平稳、β-mixing 等概念以及独立块技巧参见附录 A 平稳性、β-mixing、独立块技巧).

定义 8. 记样本 $Z_1^{\mathrm{T}}=\{Z_1,Z_2,\cdots,Z_T\}$. 令 $\sigma_i\in\{-1,+1\}$, $i=1,2,\cdots,m$ 为 Rademacher 随机变量序列. I_1 表示 Z_1^{T} 中被划入奇数编号块的样本数据点对应的指标组成的集合. 记 γ_i 为被划入 $Z(2i-1)$ 块的样本数据点对应的指标组成的集合, $i=1,2,\cdots,m$. $l(f,Z(2i-1))=\sum_{t\in\gamma_i}l(f,Z_t)=\sum_{t\in\gamma_i}Q(Y_t,f(X_t))$, $Z_t=(X_t,Y_t)\in\mathcal{Z}$ [5], 则块 Rademacher 复杂度定义为

$$\mathrm{E}R_{|I_1|}^{\beta\text{-mixing}}\mathcal{F} := \frac{1}{|I_1|}\mathrm{E}_{\tilde{Z}^o,\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i l(f,Z(2i-1))\right] \tag{33}$$

[5] 涉及更进一步的符号参见附录 A 独立块技巧中的说明.

而

$$\mathrm{E}_\sigma R_{|I_1|}^{\beta\text{-mixing}}\mathcal{F} := \frac{1}{|I_1|}\mathrm{E}_{\tilde{Z}^o,\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i l(f,Z(2i-1)) \Big| Z_1,\cdots,Z_{2m-1}\right] \tag{34}$$

即是经验块 Rademacher 复杂度. 对于偶数编号块, 块与经验块 Rademacher 复杂度分别记为

$$\mathrm{E}R_{|I_2|}^{\beta\text{-mixing}}\mathcal{F} := \frac{1}{|I_2|}\mathrm{E}_{\tilde{Z}^e,\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i l(f,Z(2i))\right] \tag{35}$$

和

$$\mathrm{E}_\sigma R_{|I_2|}^{\beta\text{-mixing}}\mathcal{F} := \frac{1}{|I_2|}\mathrm{E}_{\tilde{Z}^e,\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i l(f,Z(2i)) \Big| Z_2,\cdots,Z_{2m}\right] \tag{36}$$

这里, I_2 表示 Z_1^{T} 中被划入偶数编号块的样本数据点对应的指标组成的集合.

2.2.7 独立不同分布 Rademacher 复杂度

为处理独立但不同分布的样本数据, Mohri 等在文献 [37] 中给出了如下的 Rademacher 复杂度定义:

定义 9. 假设 X_i, $i=1,2,\cdots,n$ 为 $\mathcal{X}$ 上的独立不同分布随机变量序列, 其对应的分布分别记为 P_{X_i}, $i=1,2,\cdots,n$. 令 $\sigma_i\in\{-1,+1\}$, $i=1,2,\cdots,n$ 为 Rademacher 随机变量序列. 记函数 f: $\mathcal{Z}\to\mathbf{R}$, $\mathcal{F}$ 是 f 的集合. 有关独立不同分布的 Rademacher 复杂度定义为

$$\mathrm{E}R_n^{\text{non-identical}}\mathcal{F} := \mathrm{E}_{\prod_{i=1}^{n}\mathrm{P}_{X_i},\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right] \tag{37}$$

而经验 Rademacher 复杂度 $\mathrm{E}_\sigma R_n^{\text{non-identical}}$ 类似于定义 3 中的经典经验 Rademacher 复杂度.

注 4. 在上面的定义中, 若 P_{X_i}, $i=1,2,\cdots,n$ 为相同分布时, 则退化为定义 3 中的 Rademacher 复杂度.

2.2.8 序列 Rademacher 复杂度

Rakhlin 等在文献 [39−41] 中考察了更一般的样本数据集产生环境, 将经典 Rademacher 复杂度推广到了鞅的情形, 提出了序列 Rademacher 复杂

度, 用于分析样本数据集产生环境为鞅情形下的学习模型泛化能力. 下面给出序列 Rademacher 复杂度定义 (具体有关鞅的说明参见附录 A).

定义 10. 令 $\sigma_i \in \{-1,+1\}$, $i=1,2,\cdots,n$ 为 Rademacher 随机变量序列. $\mathcal{Z}$ 是可分度量空间[6]. 记函数 $f:\mathcal{Z}\to\mathbf{R}$ 是 $\mathcal{Z}$ 上的有界实值函数, $\mathcal{F}$ 是 f 的集合, Z 为一棵 $\mathcal{Z}$–值树[7]. 序列 Rademacher 复杂度定义为

$$\mathrm{E}(R_n^{\mathrm{mar}}\mathcal{F}) := \sup_{Z}\mathrm{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i(\sigma))\right] \tag{38}$$

其中, $Z_1(\sigma)$ 表示根结点, $Z_t(\sigma)$ 则依赖于 $\sigma_1,\sigma_2,\cdots,\sigma_{t-1}$, $t>1$.

2.3 基于 Rademacher 复杂度的分析技术与特点

在前一小节中, 总结了一批学者针对不同样本数据集产生环境以及不同的假设空间, 进行学习模型泛化能力研究时, 提出的各种形式 Rademacher 复杂度. 在本节中, 将对基于 Rademacher 复杂度进行学习模型泛化能力分析的相关处理技巧和特点进行总结.

在学习模型泛化能力分析研究中, 一种非常有用的数学技术是概率集中不等式, 是关于独立随机变量均值 (或函数) 与其期望之间偏差的概率不等式 (如 Hoeffding 不等式、Bernstein 不等式、McDiarmid 不等式、Talagrand 不等式等[63]). 并且, 让我们感兴趣的是这些不等式的偏差概率界往往是以指数或超级指数 (Super exponentially) 速度衰减, 而学习模型的估计误差则可以通过经验过程与其期望的一致偏差进行估计. 因此, 概率集中不等式为学习模型的泛化能力分析提供了必要的数学技术.

在早期的学习模型泛化能力分析研究中, 一般是采用传统概率集中不等式 (如 Hoeffding 不等式、Bernstein 不等式等) 和 Union 不等式[8] 进行研究, 具体如下:

利用概率集中不等式, 如 Hoeffding 不等式. 一般地, 可以得到, 任意 $\epsilon>0$, 存在 $h_0\in\mathcal{H}$,

$$\mathrm{P}\left(\mathrm{E}(h_0)-\mathrm{E}_n(h_0)\geq\epsilon\right)\leq\exp\left(-\frac{n\epsilon^2}{C^2}\right) \tag{39}$$

这里, C 为常数.

若 $|\mathcal{H}|<\infty$, 则结合式 (39), 并利用 Union 不等式, 便有

$$\begin{aligned}&\mathrm{P}\left(\exists h\in\mathcal{H},\mathrm{E}(h)-\mathrm{E}_n(h)\geq\epsilon\right)=\\&\mathrm{P}\left[\bigcup_{h\in\mathcal{H}}(\mathrm{E}(h)-\mathrm{E}_n(h)\geq\epsilon)\right]\leq\\&\sum_{h\in\mathcal{H}}\mathrm{P}(\mathrm{E}(h)-\mathrm{E}_n(h)\geq\epsilon)\leq\\&|\mathcal{H}|\cdot\exp\left(-\frac{n\epsilon^2}{C^2}\right)\end{aligned} \tag{40}$$

由式 (40), 即有 $0<\delta<1$,

$$\sup_{h\in\mathcal{H}}(\mathrm{E}h-\mathrm{E}_nh)\leq C\sqrt{\frac{\log|\mathcal{H}|+\log\frac{1}{\delta}}{n}} \tag{41}$$

至少以 $1-\delta$ 概率成立. 于是, 就可以用于一致偏差的估计.

若 $|\mathcal{H}|=\infty$, 式 (40) 得到的泛化界就没有意义. 此时, 可以考虑对假设空间 $\mathcal{H}$ 复杂性定义其他类型的度量, 如, 用 VC 维来代替 $|\mathcal{H}|$.

注 5. 从上面的讨论中可以看到, 由于使用了 Union 不等式, 这些分析所得到的界对分布是一致成立的, 与特定的样本数据集 (分布) 无关.

Rademacher 复杂度是与具体样本数据集分布有关的复杂性度量, 基于 Rademacher 复杂度的学习模型泛化能力分析不再使用 Union 不等式, 而是采用了诸如 McDiarmid 不等式、Talagrand 不等式等较新的概率集中不等式成果及链式技巧[9]. 因此, 往往能得到比基于传统复杂度泛化分析相对较紧致的界. 一般地, 基于 Rademacher 复杂度的泛化界分析处理技巧和特点可以总结如下:

1) 在损失函数集 $\mathcal{F}$ 或假设空间 $\mathcal{H}$ 满足一定的假设条件下, 使用概率集中不等式 (如, McDiarmid 不等式), 建立经验过程极大泛函与其期望之间的偏差关系. 如在文献 [52] 中, $0<\delta<1$, 下式至少以概率 $1-\delta$ 成立[10]

$$\sup_{f\in\mathcal{F}}|\mathrm{P}f-\mathrm{P}_nf|\leq\mathrm{E}\sup_{f\in\mathcal{F}}|\mathrm{P}f-\mathrm{P}_nf|+\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \tag{42}$$

2) 不直接刻画函数空间的复杂性规模, 而是关联经验过程的极大泛函. 首先, 针对特定的样本数

[6] 度量空间 (X,ρ) 被称为是可分的, 如果 X 有一可数的稠密子集, 即, 存在 $Y\subset X$, Y 为可数的, 且任意 $x\in X$, 存在 $r>0$, s.t., $B(x,r)(:=\{g\in X:\rho(g,x)<r\})\cap Y\neq\emptyset$[71].

[7] 一棵结点取值于可分度量空间 $\mathcal{Z}$, 深度为 n 的完全二叉树.

[8] $\mathrm{P}(A\cup B)\leq\mathrm{P}(A)+\mathrm{P}(B)$.

[9] 任意 $f\in\mathcal{F}$, 存在 $n_0\in\mathbf{N}$ 及函数 f_i, $i=1,2,\cdots,n_0$, 按链式展开有 $f=f-f_{n_0}+\sum_{i=1}^{n_0}(f_i-f_{i-1})$, 其中, 约定 $f_0=0$[72].

[10] 这里, $\mathrm{P}f=\mathrm{E}(h)$, $\mathrm{P}_nf=\mathrm{E}_n(h)$.

[11] 对称性质的基本思想: 假设 X_i, $i=1,2,\cdots,n$ 为 i.i.d. 随机变量序列. 如果 $\frac{1}{n}\sum_{i=1}^{n}f(X_i)$ 的值接近 $\mathrm{E}f(X_1)$, 那么, $\frac{1}{n}\sum_{i=1}^{n}f(X_i)$ 的值接近 $\frac{1}{n}\sum_{i=1}^{n}f(X_i')$, 其中, X_i' 与 X_i, $i=1,2,\cdots,n$ 独立且同分布.

据集产生环境以及损失函数集 $\mathcal{F}$ (或假设空间 $\mathcal{H}$), 定义具体的 Rademacher 复杂度. 然后, 根据对称性质[11], 得到极大泛函期望与 Rademacher 复杂度的关系不等式. 如, 在文献 [52] 中有, 任意 $\mathcal{F}$,

$$\max\{\mathrm{E}\sup_{f\in\mathcal{F}}(\mathrm{P}f-\mathrm{P}_nf),\mathrm{E}\sup_{f\in\mathcal{F}}(\mathrm{P}_nf-\mathrm{P}f)\}\leq 2\mathrm{E}(R_n\mathcal{F}) \tag{43}$$

最后, 基于得到的关系不等式 (43), 建立经验过程极大泛函与 Rademacher 复杂度之间的关联.

3) 建立 Rademacher 复杂度与经验 Rademacher 复杂度之间的关联. 如, 在文献 [20] 中有, 基于概率集中 (如 Talagrand 不等式), 得到 Rademacher 过程的集中不等式, 记函数 $f:\mathcal{X}\to[a,b]$ $(a,b\in\mathbf{R})$, $\mathcal{F}$ 是 f 的集合, 便有, 任意 $\epsilon>0$,

$$\mathrm{P}\left(\mathrm{E}_\sigma(R_n\mathcal{F})\leq\mathrm{E}(R_n\mathcal{F})-\sqrt{\frac{\epsilon\cdot\mathrm{E}(R_n\mathcal{F})}{(b-a)^{-1}\cdot n}}\right)\leq \exp(-\epsilon) \tag{44}$$

4) 利用数学不等式 (如 Cauchy 不等式等) 与链式技巧[72], 给出经验 Rademacher 复杂度的上界. 如熵积分 (或 Dudley 积分)[61], 任意 $\mathcal{F}$,

$$\mathrm{E}_\sigma(R_n\mathcal{F})\leq 12\int_0^\infty\sqrt{\frac{\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_{L_2(\mathrm{P}_n)})}{n}}\mathrm{d}\epsilon \tag{45}$$

5) 在不同的算法策略下 (如 ERM、SRM、正则化等), 基于经验 Rademacher 复杂度的上界, 计算获得学习模型的泛化界. 一般地, 基于 Rademacher 复杂度的分析可以获得比 VC 维稍快的界.

3 不同复杂度间的关联

在第 2 节中, 对现有文献中各种形式 Rademacher 复杂度进行了总结. 在本节中, 将讨论各种形式 Rademacher 复杂度与传统复杂度[12] 间的相互关系.

传统复杂度之间的关联性已经被许多学者进行过广泛的讨论 (关联性情况见图 1 中的 (12) ∼ (16), 具体关联性讨论见文献 [11, 48, 73−77]). 当前, 关于 Rademacher 复杂度在学习模型泛化能力分析方面的研究, 主要是关注训练样本数据集为 i.i.d. 和 non-i.i.d. 两类产生环境, 相关结果总结至表 1. 为了理解 Rademacher 复杂度在学习模型泛化能力方面的应用与分析, 一批学者在根据具体情形提出各种形式的 Rademacher 复杂度之后, 也对其与传统复杂度之间的关系进行了深入分析. 接下来, 我们对各种形式的 Rademacher 复杂度及其与传统复杂度之间的关联性进行总结, 相关情况见图 1. 下面是按图 1 中关联性 1 ∼ 11 展开的具体讨论.

表 1 Rademacher 复杂度及传统复杂度

Table 1 Rademacher complexities and kinds of complexities of function classes

复杂度类型	样本数据集产生环境	复杂度名称
传统复杂度	i.i.d./non-i.i.d.	VC 熵, 退火 VC 熵, 生长函数, VC 维, 覆盖数, 伪维度, Fat-shattering 维等经典 Rademacher 复杂度, 局部 Rademacher 复杂度
Rademacher 复杂度	i.i.d.	Rademacher chaos 复杂度, 单模态 Rademacher 复杂度, 多模态 Rademacher 复杂度, Dropout Rademacher 复杂度
	non-i.i.d.	独立不同分布 Rademacher 复杂度, 块 Rademacher 复杂度, 序列 Rademacher 复杂度

关联性 1. Rademacher 复杂度与 VC 熵、生长函数、VC 维、退火 VC 熵

Massart 在文献 [50] 中, 尝试建立了经典 Rademacher 复杂度与生长函数之间的关联性. 如,

$$\mathrm{E}(R_n\mathcal{F})\leq\sqrt{\frac{2\cdot S_\mathcal{F}(n)}{n}} \tag{46}$$

其中, $S_\mathcal{F}(n)$ 为生长函数.

Kääriäinen 在文献 [45] 中对 Rademacher 复杂度与 VC 维之间的关系进行了深入的分析讨论. Anguita 等在文献 [44] 中建立了如下经典 Rademacher 复杂度和 VC 的熵关系

$$\mathrm{E}(R_n\mathcal{F})\leq\min_{\lambda\in(0,\infty]}\frac{H_\mathcal{F}(n)}{\lambda n}+\frac{\ln\cosh(\lambda)}{\lambda}$$

且有

$$\lim_{\lambda\to\infty}\frac{H_\mathcal{F}(n)}{\lambda n}+\frac{\ln\cosh(\lambda)}{\lambda}=1 \tag{47}$$

其中, $H_\mathcal{F}(n)$ 为 VC 熵, cosh 为双曲余弦函数. 进一步地, Anguita 等基于组合的分析方法, 讨论分析了经典 Rademacher 复杂度与退火 VC 熵, 生长函数和 VC 维之间的关系, 并给出了经典 Rademacher 复杂度和 VC 熵之间互为可逆的计算关系. 基于这种计算关系, 通过数值模拟的方式表明, 可以将经典

[12] 本文中所讨论的传统复杂度主要限于 VC 熵、退火 VC 熵、生长函数、VC 维、覆盖数、伪维度、Fat-shattering 维等复杂度.

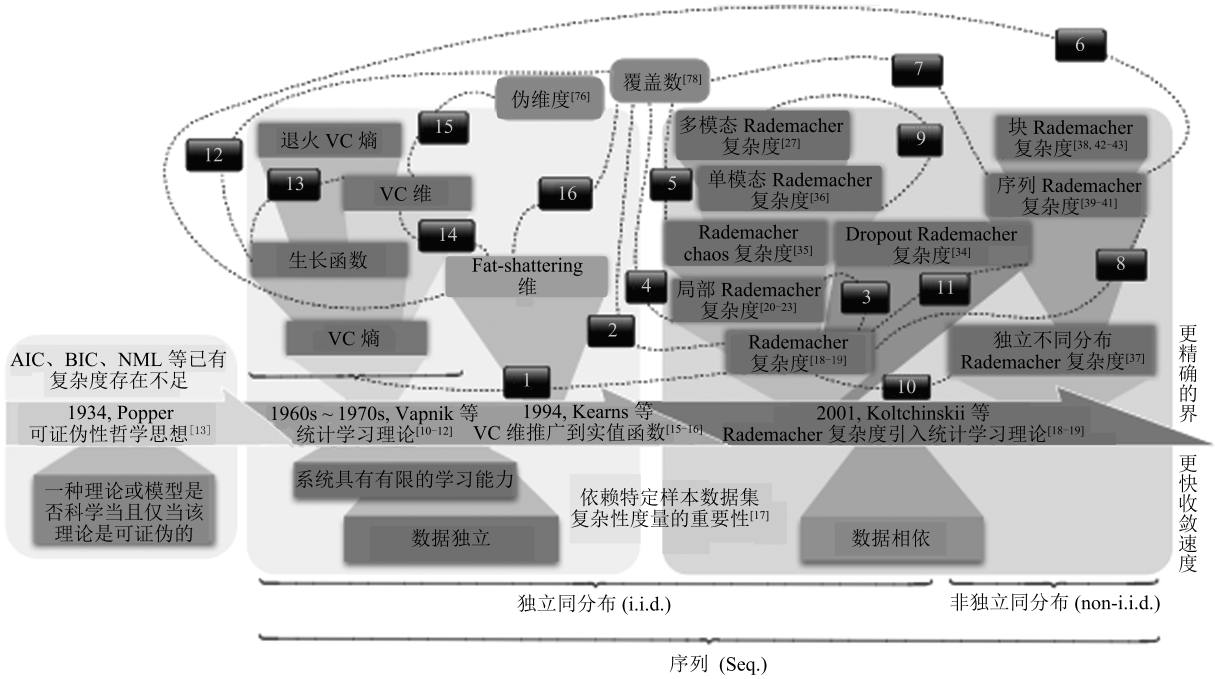

图 1　复杂性度量及其相互关系

Fig. 1　Complexity measures and their relationships

Rademacher 复杂度上界改进为介于 $\mathrm{O}(n^{-1})\sim\mathrm{O}(n^{-1/2})$ 之间.

关联性 2. 经典 Rademacher 复杂度与覆盖数

Srebro 等在文献 [46−47] 中讨论改进了文献 [61] 中经典 Rademacher 复杂度与覆盖数之间的关系, 给出了比式 (45) 更为严格的熵积分, 任意 $\mathcal{F}$,

$$\mathrm{E}_\sigma(R_n\mathcal{F})\leq 4\epsilon+12\int_\epsilon^\infty\sqrt{\frac{\log\mathcal{N}(\epsilon',\mathcal{F},\|\cdot\|_{L_2(\mathrm{P}_n)})}{n}}\mathrm{d}\epsilon' \tag{48}$$

V′Yugin 在文献 [48] 中进一步给出了经典 Rademacher 复杂度−生长函数、经典经验 Rademacher 复杂度−覆盖数与经典 Rademacher 复杂度−覆盖数之间的等价关系.

关联性 3. 局部 Rademacher 复杂度与经典 Rademacher 复杂度

Barlett 等在文献 [20−21] 中研究分析了假设空间中具有较小方差的函数集. 于是, 在 (全局) 经典 Rademacher 复杂度基础上得到了局部 Rademacher 复杂度, 见式 (21).

关联性 4. 局部 Rademacher 复杂度与覆盖数

Lei 等在文献 [72] 基础上得到了局部 Rademacher 复杂度与覆盖数之间的关系[28, 32]

$$\mathrm{E}\mathcal{R}_n\{f\in\mathcal{F}:\mathrm{P}f^2\leq r\}\leq \inf_{\epsilon>0}\left[2\epsilon+\sqrt{\frac{2r\log\mathcal{N}\left(\frac{\epsilon}{2},\mathcal{F},\|\cdot\|_2\right)}{n}}+\frac{8b\log\mathcal{N}\left(\frac{\epsilon}{2},\mathcal{F},\|\cdot\|_2\right)}{n}\right] \tag{49}$$

关联性 5. Rademacher chaos 复杂度与覆盖数

Lei 等在文献 [49] 的基础上进一步给出了 Rademacher chaos 复杂度与覆盖数之间的关系[29, 32]

$$\mathrm{E}_\sigma\mathcal{U}_n(\mathcal{F})\leq \inf_{0<\epsilon<\frac{D}{2}}\left[\sqrt{\frac{n(n-1)}{2}}\epsilon+c\int_{\frac{\epsilon}{2}}^D\log\mathcal{N}(\omega,\mathcal{F}\cup\{0\},d_x)\mathrm{d}\omega\right] \tag{50}$$

关联性 6. Dropout Rademacher 复杂度与经典 Rademacher 复杂度

Gao 等在文献 [34] 中为深入研究深度神经网络中 Dropout 的泛化性能, 提出了用于反映网络结构变化的 Dropout Rademacher 复杂度. 当刻画网络结构变化的随机向量序列 $\boldsymbol{r}_i^s$, $i=1,2,\cdots,n$ 退化为常向量序列时, Dropout Rademacher 复杂度就在一定意义下退化为经典 Rademacher 复杂度.

关联性 7 ∼ 9. 序列 Rademacher 复杂度与 Fat-shattering 维、覆盖数、经典 Rademacher 复杂度

Rakhlin 等在文献 [39−41] 中给出了鞅情形的序列 Rademacher 复杂度, 建立了类似经典 Rademacher 复杂度与覆盖数的熵积分关系 (见式 (48))

$$\mathrm{E}(R_n^{\mathrm{mar}}\mathcal{F}) \leq \inf_{0<\epsilon}\left[4\epsilon + \frac{12}{\sqrt{n}}\int_{\epsilon}^{1}\sqrt{\log \mathcal{N}_2(\omega,\mathcal{F},Z)}\mathrm{d}\omega\right] \tag{51}$$

这里, $\mathcal{N}_p(\epsilon,\mathcal{F},Z)$ 表示树 Z 上 $\mathcal{F}$ 在 p 范数意义下的覆盖数. 并且, 还在一定意义下给出了序列 Rademacher 复杂度和 Fat-shattering 维之间的关联性. 同时, 说明了在 $Z_t(\sigma)$ 不依赖于 $\sigma_1,\sigma_2,\cdots,\sigma_{t-1}$ 时, 则序列 Rademacher 复杂度退化为经典 Rademacher 复杂度. 并且, 还给出了类似经典 Rademacher 复杂度的序列 Rademacher 复杂度结构化结果. 另外, 在鞅条件下, 对统计学习理论中的相关传统复杂性度量性质 (如生长函数与 VC 维关联性等) 进行了推广和研究.

关联性 10. 多模态 Rademacher 复杂度与单模态 Rademacher 复杂度

Lei 和 Cao 等分别在文献 [27, 32, 36] 中研究了单模态 Rademacher 复杂度和多模态 Rademacher 复杂度. 在多模态退化为单模态的情况下, 多模态 Rademacher 复杂度[27] 就退化为了单模态 Rademacher 复杂度[36].

关联性 11. 独立不同分布 Rademacher 复杂度与经典 Rademacher 复杂度

在独立不同分布退化为独立同分布的情况下, Mohri 等在文献 [37] 中关于独立不同分布的 Rademacher 复杂度定义就退化为了经典 (经验) Rademacher 复杂度.

Rademacher 复杂度不同于传统 VC 理论, 是一种依赖特定样本数据集 (分布) 的复杂性度量技术, 能获得比 VC 维更加紧致的界. 基于上面关联性 1 ∼ 11 的讨论, 可以知道, 当前关于 Rademacher 复杂度在泛化界分析研究方面主要集中在训练样本数据集产生环境为 i.i.d. 和 non-i.i.d. 两种情形, 而关于 i.i.d. 的研究已有较长时间, 且研究结果也相对较完善, 而对于 non-i.i.d. 的研究主要是最近几年的工作, 并且研究结果相对尚不完善. 同时也发现, 无论是 i.i.d. 还是 non-i.i.d. 的研究, 都试图将 Rademacher 复杂度转化为有意义且可计算的上界:

1) 将 Rademacher 复杂度转化为 (熵) 积分形式. 如上面讨论中的关联性 2、4、5、7 ∼ 9 等;

2) 将 Rademacher 复杂度 (如单模态 Rademacher 复杂度、多模态 Rademacher 复杂度、独立不同分布 Rademacher 复杂度、块 Rademacher 复杂度等) 通过不同意义下的可计算上界 (如范数界[27, 34, 36]、VC 维界[37]、次根函数上界[42]、核值 Gram 矩阵迹[48] 等) 得到学习模型的泛化界, 从而用于泛化能力的分析研究. 这些复杂度是否存在 (熵) 积分形式仍是一个值得探讨的问题.

4 独立同分布环境

在本节中, 约定样本数据以独立同分布的方式产生, 按照对假设空间 $\mathcal{H}$ 的不同假定, 在相应的策略下来讨论 Rademacher 复杂度在泛化能力方面的应用分析情况. 下面对这些方面的应用成果[19, 27−28, 30−33, 36, 47−49, 51−56] 总结如下:

1) 假定

$$\begin{cases}\mathcal{H} = \{h : h \text{ 为从 } \mathcal{X} \text{ 到 } \{-1,+1\} \text{ 的映射}\}\\ \text{ERM 策略}\end{cases}$$

Boucheron 等在文献 [52] 中总结了 Rademacher 复杂度对泛化界一致偏差的分析情况.

a) 建立经验过程极大泛函与其期望之间的偏差关系, 见式 (42);

b) 建立极大泛函期望与经典 Rademacher 复杂度的关系, 见式 (43);

c) 建立经典经验 Rademacher 复杂度与生长函数的关系, 见式 (46).

一般地, 结合 a) ∼ c) 便可以得到一致偏差为 $\mathrm{O}(1/\sqrt{n})$. Oneto 等在文献 [53−55] 中基于 Talagrand 集中不等式, 将经典 Rademacher 复杂度应用于学习模型泛化误差分析, 将一致偏差改进为介于 $\mathrm{O}(1/n) \sim \mathrm{O}(1/\sqrt{n})$ 之间, 并给出了最优情况为 $\mathrm{O}(1/n)$.

注 6. V′Yugin 在文献 [48] 中, 说明了上述 c) 中关系等价于经典 (经验) Rademacher 复杂度与覆盖数的熵积分关系, 并给出了覆盖数与 Fat-shattering 维之间的关系, 从而, 可以在一定条件下更方便地计算估计经典 Rademacher 复杂度.

2) 假定

$$\begin{cases}\mathcal{H} = \{h : h \text{ 为从 } \mathcal{X} \text{ 到 } \mathbf{R} \text{ 的映射}\}\\ \text{ERM 策略}\end{cases}$$

在经典的统计机器学习理论中, 往往是在一定条件下 (如基于损失函数的 Lipschitz 条件、基于一致有界等[19, 79]), 针对一些具体的 Fat-shattering 维, 覆盖数等的上界, 分析并获得一致偏差或估计误差为 $\mathrm{O}(1/\sqrt{n})$. Srebro 等在文献 [47] 中, 对损失函数进行特殊假定[13]

[13] 满足该条件, 称为 $H-$ 光滑 (Smooth)[47].

$$\left|\frac{\partial^2 Q(z,h)}{\partial h^2}\right| \le H, \quad H\text{为常数} \tag{52}$$

从而, 在一定意义下可以获得更好的估计误差为 $\mathrm{O}(1/n)$, 并将相关结果应用到了在线学习 (On-line learning) 和随机优化 (Stochastic optimization).

Lei 等[28, 32] 在损失函数为一致有界的条件下, 给出了局部 Rademacher 复杂度的熵积分, 见式 (49). 同时, 基于 Bernstein 不等式, 对于特殊情况下可计算的覆盖数上界, 获得诸如 $\mathrm{O}(((\log^p n)/n)^{1/(2-\alpha)})$, $p>0$, $0<\alpha\le 1$ 等估计误差.

3) 假定

$$\begin{cases}\mathcal{H}=\{f: f \text{ 为从 } \mathbf{R}^d \text{ 到 } \mathbf{R} \text{ 的映射}\}\\ \text{ERM 策略}\end{cases}$$

Gnecco 等在文献 [51] 中应用经典 Rademacher 复杂度分析了径向基神经 (Radial basis function, RBF) 网络的逼近误差, 得到一定条件下[14] 的学习率为 $\mathrm{O}(1/\sqrt{n})$.

4) 假定

$$\begin{cases}\mathcal{H}=\{f: f \text{ 为从 } \mathcal{X} \text{ 到 } \{0,1\}^L \text{ 的映射}, L\ge 1\}\\ \text{ERM 策略}\end{cases}$$

Yu 等在文献 [56] 中应用经典 Rademacher 复杂度分析了多标签学习 (Multi-label) 的泛化误差, 得到估计误差为 $\mathrm{O}(1/\sqrt{n})$. Xu 等在文献 [33] 中基于局部 Rademacher 复杂度分析了多标签学习, 改进了文献 [56] 中结果, 在一定条件下得到了 $\mathrm{O}((\log n)/n)$ 的估计误差.

前面的讨论, 对于假设空间的限制较少, 下面将讨论几类特殊假设空间的学习模型泛化能力分析情况.

5) 假定

$$\begin{cases}\mathcal{H}_k=\left\{\pi_b(h): h\in\sum_{k,l}(\mathcal{X})\right\},\ \forall k\in\mathbf{N}\\ \text{ERM 策略}\end{cases}$$

其中, $S_l(T)$ 为关于节点集 T 的 $l-$阶样条集合, $\sum_{k,l}(\mathcal{X}) = \bigcup_{T:=\{a=t_0<t_1<\cdots<t_k=b\}} S_l(T)$, $\pi_b(h)(x) := \max\{-b, \min\{b, h(x)\}\}$, $\forall x\in\mathcal{X}$.

Lei 等在文献 [31−32] 中分析了 FKS, 得到 FKS 的局部 Rademacher 紧致上界, 分析了 FKS 的估计误差, 利用逼近论分析 FKS 的逼近误差, 在对目标函数等做一定的假定下, 得到一定意义下的泛化误差为 $\mathrm{O}((\log n/n)^{2\alpha/(1+2\alpha)})$, 其中, $0<\alpha<l$, l 表示样条的阶.

6) 假定[15]

$$\begin{cases}\mathcal{H}_s=\{h_s: h_s \text{ 为从 } \mathcal{X}\times R_s \text{ 到 } \mathbf{R} \text{ 的映射}\}\\ \text{ERM 策略}\end{cases}$$

Hinton 等为防止深度神经网络中的过拟合问题, 于 2012 年提出了 Dropout 技术[69]. Wan 等基于经典 Rademacher 复杂度对 Dropout 的泛化性进行过分析研究[70]. Gao 等注意到经典 Rademacher 复杂度无法反映出 Dropout 技术所带来的网络结构动态变化, 于是在文献 [34] 中提出了 Dropout Rademacher 复杂度, 并基于此复杂度研究了深度神经模型的泛化性能. 在一定条件下 (如 $Q(\cdot,\cdot)\le M$, M 为常数) 得到了类似于经典情况 (如式 (15) 和 (16) 或是按式 (42)、(43) 和 (44) 得到) 的不等式 (以 $1-\delta$ 概率成立)

$$\mathrm{E}_n(h)-\mathrm{E}(h)\le 2\mathrm{E}(R_n^{\mathrm{Dropout}}\mathcal{H}_s)+M\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{53}$$

和

$$\mathrm{E}_n(h)-\mathrm{E}(h)\le 2\mathrm{E}_\sigma(R_n^{\mathrm{Dropout}}\mathcal{H}_s)+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{54}$$

进一步地, 为分析研究具有 k, $k\in\mathbf{N}_n$ 层隐藏层深度神经网络的误差估计, 讨论得到了深度神经网络中三种不同类型的 Dropout 对应的 Dropout Rademacher 复杂度上界 (刻画了 Dropout Rademacher 复杂度与深度神经网络隐藏层数之间的关系). 具体如下:

a) 如果是节点的 Dropout, Dropout Rademacher 复杂度上界为 $\mathrm{O}(p^{(k+1)/2}/\sqrt{n})$;

b) 如果是权重的 Dropout, Dropout Rademacher 复杂度上界为 $\mathrm{O}(p^{(k+1)/2}/\sqrt{n})$;

c) 如果是节点和权重混合的 Dropout, Dropout Rademacher 复杂度上界为 $\mathrm{O}(p^{(k+1)}/\sqrt{n})$.

7) 假定[16]

$$\begin{cases}\mathcal{H}=\\ \{h: h \text{ 为从 } \mathcal{X} \text{ 到 } \mathbf{R} \text{ 的映射}, \text{且 } h\in\mathcal{H}_K,\ K\in\mathcal{K}\}\\ \text{正则化策略}\end{cases}$$

其中, $\mathcal{H}_K$ 为 RKHS, Mercer 核集合记为 $\mathcal{K}$.

[14] 如对 f 加限制, s.t., $f=\beta_r\cdot\lambda$, β_r 为 r 阶贝塞尔位势 (Bessel potential of order r), λ 满足 L^1 范数条件 [71].
[15] 这里, R_s 参见定义 7.
[16] 再生核希尔伯特空间 (Reproducing kernel Hilbert space, RKHS)[29, 48].

Lei 等[28, 32] 改进了文献 [39] 中 Rademacher chaos 复杂度, 得到了更为一般的结果, 见式 (50). 在损失函数为 Lipschitz 连续, 且假设空间函数满足, 存在 $R > 0$,

$$\|h\|_K \leq R, \quad \forall\, h \in \mathcal{H}_K \tag{55}$$

的条件下, 将文献 [49] 中给出的泛化误差与 Rademacher chaos 复杂度之间量化关系, 应用于 MKL 算法泛化误差的分析[29, 32]. 对于 Hinge 损失函数和 $q-$范数软边缘损失函数, 得到了比文献 [49] 中更一般化的泛化能力或是说额外错分率 (Excess misclassification error) 结果. V′Yugin 则在文献 [48] 中 (这里, $\mathcal{Y}$ 为 $\{-1, +1\}$), 基于样本数据点 $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ 的核值 Gram 矩阵 G, 给出了经典 Rademacher 复杂度的上界

$$R_n\mathcal{F} \leq \frac{1}{n}\sqrt{\mathrm{tr}(G)} \tag{56}$$

然后, 基于式 (42) 和 (43), 在一定意义下给出了泛化误差或是说错分率 (Misclassification error) 为 $\mathrm{O}(\sqrt{\mathrm{tr}(G)}/n + 1/\sqrt{n})$.

8) 假定

$$\begin{cases} \mathcal{H} = \left\{ M \in S^d : \|M\| \leq \dfrac{1}{\sqrt{\lambda}} \right\} \\ \text{正则化策略} \\ \quad (M_z, b_z) := \arg \min\limits_{M \in S^d, b \in \mathbf{R}} \left\{ \mathrm{E}_z(M, b) + \lambda \|M\|^2 \right\} \end{cases}$$

Cao 等在文献 [36] 中研究了单模态度量学习的估计误差 (算法稳定性). 具体如下:

a) 定义 $\mathrm{E}R_n^{\mathrm{single}}$, 给出估计误差与 $\mathrm{E}R_n^{\mathrm{single}}$ 之间的量化关系

$$\begin{aligned} &\mathrm{E}(M_z, b_z) - \mathrm{E}_n(M_z, b_z) \leq \\ &\quad \sup_{(M,b) \in \mathcal{F}} [\mathrm{E}(M, b) - \mathrm{E}_n(M, b)] \leq \\ &\quad \frac{4\mathrm{E}R_n^{\mathrm{single}}}{\sqrt{\lambda}} + \frac{4(3 + 2X_*/\sqrt{\lambda})}{\sqrt{n}} + \\ &\quad 2(1 + X_*/\sqrt{\lambda}) \left(\frac{2\ln(\frac{1}{\delta})}{n} \right)^{\frac{1}{2}} \end{aligned} \tag{57}$$

其中, $X_* = \sup_{x, x' \in \mathcal{X}} \|(x - x')(x - x')^{\mathrm{T}}\|_*$.

b) 在弗罗贝尼乌斯–范数 (Frobenius-norm), 稀疏 L^1-范数 (Sparse L^1-norm) 和混合 $(2,1)-$范数 (Mixed $(2,1)$-norm) 等几种不同 (矩阵) 范数意义下, 估计 $\mathrm{E}R_n^{\mathrm{single}}$ 的上界.

c) 结合 a) 和 b) 得到了诸如 $\mathrm{O}(1/\sqrt{n})$ 等估计误差.

9) 假定

$$\begin{cases} \mathcal{H} = \left\{ M \in S^{d \times (md)} : \|M\| \leq \dfrac{1}{\sqrt{\lambda}} \right\} \\ \text{正则化策略} \\ \quad (M_z, b_z) := \\ \qquad \arg \min\limits_{M \in S^{d \times (md)}, b \in \mathbf{R}} \{ \mathrm{E}_z(M, b) + \lambda \|M\|^2 \} \end{cases}$$

McFee 等在文献 [67−68, 80−81] 对多模态度量学习进行了广泛而深入的研究, 但是这些研究更多的是偏重多模态度量学习的算法, 有关模型学习能力的研究工作则相对缺乏. Lei 等在文献 [27, 32] 中基于文献 [36] 单模态度量学习的工作, 展开了多模态度量模型学习能力的研究.

a) 在一定条件下 (如 $\mathbf{R}^{d \times (md)}$ 上函数满足 $\beta-$强凸[27] 等) 给出了 $\mathrm{E}R_n^{\mathrm{multi}}(\mathcal{M})$ 的估计上界

$$\mathrm{E}(R_n^{\mathrm{multi}}\mathcal{M}) \leq X_* \sqrt{\frac{2f_{\max}}{\beta \lfloor \frac{n}{2} \rfloor}} \tag{58}$$

b) 给出估计误差与多模态 Rademacher 复杂度之间的量化关系

$$\begin{aligned} &\mathrm{E}(M_z, b_z) - \mathrm{E}_n(M_z, b_z) \leq \\ &\quad 2\mathrm{E}(R_n^{\mathrm{multi}}\mathcal{M}) + 2\left(1 + \frac{X_*}{\sqrt{\lambda}}\right) \frac{1 + 2\sqrt{\ln(\frac{1}{\delta})}}{\sqrt{\lfloor \frac{n}{2} \rfloor}} \end{aligned} \tag{59}$$

c) 结合 a) 和 b), 讨论了混合–范数和 Schatten–范数等几种不同 (矩阵) 范数意义下的多模态度量模型学习能力, 得到了诸如 $\mathrm{O}(1/\sqrt{n})$ 等估计误差.

注 7. 当多模态退化为单模态时, 式 (59) 就退化成了单模态情形下 (见式 (57)) 的更紧致的上界.

10) 假定

$$\begin{cases} \mathcal{H}_k = \left\{ \sum\limits_{i=1}^{k} \omega_i K([x - c_i]^{\mathrm{T}} A_i [x - c_i]) + w_0 : \right. \\ \qquad \left. \sum\limits_{i=0}^{k} |\omega_i| \leq b \right\}, \ k \in \mathbf{N} \\ \text{SRM 策略} \end{cases}$$

神经网络的学习能力已经有大批学者从逼近误差和估计误差两方面进行了广泛而深入的研究[51, 76, 82−87]. Lei 等[30, 32] 在文献 [82] 中关于估计误差的分析基础上, 将局部 Rademacher 复杂度引入到神经网络估计误差的分析. 首先, 改进了式 (49) 的结果, 得到

$$\mathrm{E}\mathcal{R}_n\{f\in\mathcal{F}:\mathrm{P}f^2\le r\}\le$$
$$\inf_{\epsilon>0}\left[2\epsilon+(2\sqrt{2b\epsilon}+\sqrt{r})\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}}+\frac{8b\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}\right] \tag{60}$$

然后, 结合式 (60), 并应用 Talagrand 不等式, 在核函数、损失函数等满足一定条件下, 得到估计误差为 $\mathrm{O}((1/n)^{1/(2-\alpha_p)})$, 其中, $\alpha_p=(2/p)\wedge 1$, $p>1$.

注 8. 本节主要讨论了样本数据集为 i.i.d. 的情形, 按照一般的假设空间以及特殊的假设空间 (如 $\mathcal{H}$ 为 RKHS、矩阵空间和样条函数空间等几类情况), 在 ERM、SRM 和正则化等几种不同的学习策略下, 总结讨论了 Rademacher 复杂度在学习模型泛化界方面的应用分析. 在这些应用分析中, 主要采用了概率集中不等式、覆盖数、链式技巧和 Cauchy 不等式等一系列数学技术. 一般地, 基于 Rademacher 复杂度的分析能得到比 VC 维相对紧致的界. 为便于阅读, 相关内容总结至表 2.

5　非独立同分布环境

在前一节中, 主要讨论了在样本数据为 i.i.d. 的假定下, Rademacher 复杂度在泛化能力分析方面的应用. 但是, 现实应用中存在的大量数据, 本质上具有相依性或时间相关性. 所以, 对具体实际应用而言, i.i.d. 是一个非常强的假设. 正是因为如此, 很多学者开始考虑研究 non-i.i.d. 情形下的学习模型泛化能力. 最早的工作可以追溯到 Yu 在文献 [88] 中基于 Bernstein 的独立块技术, 在样本数据分布为平稳分布且满足 β-mixing 的假定下, 对学习模型泛化误差进行的 VC 维界分析研究. 之后, Meir 在文献 [89] 中基于假设空间复杂性的覆盖数度量, 分析研究了学习模型的泛化误差. 针对具体样本数据集分布, Mohri 等在文献 [38] 中开始采用 Rademacher 复杂度来分析假设空间的复杂性, 研究讨论了在样本数据分布为平稳分布且满足 β-mixing 的假定下, 学习模型的一致偏差. 之后, 许多学者陆续将 Rademacher 复杂度的泛化能力研究推广到样本数据集分布为几种特殊 non-i.i.d. 的情形.

本节主要在样本数据分别为独立但不同分布、β-mixing、鞅和非平稳等几种 non-i.i.d. 情形约定下, 按对假设空间 $\mathcal{H}$ 的不同假定, 来讨论 Rademacher 复杂度在学习模型泛化能力方面的分析应用结果 (见文献 [37−43]).

1) 假定

$$\begin{cases}\text{独立不同分布}\\ \mathcal{H}=\{f:f\text{ 为从 }\mathcal{X}\text{ 到 }\mathcal{Y}\text{ 的映射}\}\end{cases}$$

Mohri 等[37] 基于文献 [90−91] 中研究结果, 进一步提出不同分布下的 Rademacher 复杂度, 来分析研究训练样本数据集独立但是不同分布的情况, 在一定条件下 (如 $Q(\cdot,\cdot)\le M$, M 为常数) 得到了

$$\mathrm{E}_{P_{X_{n+1}}}[h]-\mathrm{E}_n[h]\le 2\mathrm{E}(R_n^{\text{non-identical}}\mathcal{H})+M\sqrt{\frac{\log\frac{1}{\delta}}{2n}}+\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}}|\mathrm{E}_{P_{X_i}}[h]-\mathrm{E}_{P_{X_{n+1}}}[h]| \tag{61}$$

基于上一节 i.i.d. 部分的讨论, 可以知道, 在一定意义下一致偏差为 $\mathrm{O}(1/\sqrt{n})$. 同时, Mohri 等还将相关结果应用到了在线学习.

表 2　i.i.d. 情形的泛化界分析

Table 2　Generalization analysis for i.i.d.

样本数据集产生环境	学习策略	假设空间	泛化能力
i.i.d.	ERM	1) $\mathcal{X}\to\{-1,+1\}$	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [52], $\mathrm{O}\left(\frac{1}{n}\right)$ [53−55] 等
		2) $\mathcal{X}\to\mathbf{R}$	$\mathrm{O}\left(\frac{1}{n}\right)$ [47], $\mathrm{O}\left(\frac{\log^p n}{n}\right)^{1/(2-\alpha)}$ [28, 32] 等
		3) $\mathbf{R}^d\to\mathbf{R}$	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [51]
		4) $\mathcal{X}\to\{-1,+1\}^L$	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [56], $\mathrm{O}\left(\frac{\log n}{n}\right)$ [33]
		5) 自由样条函数空间	$\mathrm{O}\left(\left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)}\right)$ [31−32] 等
		6) $\mathcal{X}\times R_s\to\mathbf{R}$	$\mathrm{O}\left(\frac{p^{(k+1)/2}}{\sqrt{n}}+\frac{1}{\sqrt{n}}\right)$, $\mathrm{O}\left(\frac{p^{k+1}}{\sqrt{n}}+\frac{1}{\sqrt{n}}\right)$ [34]
		7) RKHS	$\mathrm{O}\left(\frac{\sqrt{\mathrm{tr}(G)}}{n}+\frac{1}{\sqrt{n}}\right)$ [48]
	正则化	8) S^d	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [36] 等
		9) $S^{d\times(md)}$	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [27, 32] 等
	SRM	10) 神经网络空间	$\mathrm{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{2-\alpha_p}}\right)$ [30, 32]

注 9. 当 P_{X_i}, $i=1,2,\cdots,$ 为相同分布时, 式 (61) 就退化为式 (42) 和 (43) 的情形.

2) 假定

$$\begin{cases}\text{平稳分布且 } \beta\text{-mixing}\\ \mathcal{H}=\{f: f \text{ 为从 } \mathcal{X} \text{ 到 } \mathbf{R} \text{ 的映射}\}\end{cases}$$

Mohri 等[38] 在文献 [88−89] 基础上, 采用独立块技巧, 应用 Rademacher 复杂度分析研究了平稳 β-mixing 序列[17].

a) 建立原始样本数据集中数据块与构造的独立块之间的泛化误差关系

$$|\mathrm{E}_{Z^o}[h]-\mathrm{E}_{\tilde{Z}^o}[h]|\leq(s-1)\cdot M\cdot\beta(a') \tag{62}$$

b) 应用 i.i.d. 情形下的对称性质, 建立极大泛函期望与 Rademacher 复杂度的关系

$$\mathrm{E}_{\tilde{Z}^o}[\Phi(\tilde{Z}^o)]\leq 2\mathrm{E}(R^{\beta\text{-mixing}}_{|I_1+I_2|}\mathcal{H}) \tag{63}$$

其中, $\Phi(\tilde{Z}^o)=\sup_{h\in\mathcal{H}}(\mathrm{E}_{\tilde{Z}^o}\frac{1}{m}\sum_{i=1}^m h(\tilde{Z}(2i-1))-\frac{1}{m}\sum_{i=1}^m h(\tilde{Z}(2i-1)))$.

c) 结合 a) 和 b), 并应用 i.i.d. 情形下的 Mcdiarmid 不等式, 得到一致偏差的 Rademacher 复杂度上界不等式

$$\mathrm{E}\frac{1}{n}\sum_{i=1}^n h(Z_i)-\frac{1}{n}\sum_{i=1}^n h(Z_i)\leq 2\mathrm{E}(R^{\beta\text{-mixing}}_{|I_1+I_2|}\mathcal{H})+M\sqrt{\frac{\log\frac{2}{\delta}}{2s}} \tag{64}$$

其中, $n=2sa$, a 为块的长度, $\mathrm{E}(R^{\beta\text{-mixing}}_{|I_1+I_2|}\mathcal{H})=\frac{1}{|I_1+I_2|}\mathrm{E}_{(\tilde{Z}^o,\tilde{Z}^e),\sigma}[\sup_{f\in\mathcal{F}}\sum_{i=1}^s\sigma_i l(f,Z(i))]$. 其他符号见第 3 节中块 Rademacher 复杂度以及附录 A 独立块技巧的说明.

类似于 i.i.d. 的讨论[48], 在 $\mathcal{Y}=\{-1,+1\}$ 的假定下, 给出了一定意义下泛化误差或是错分率为 $\mathrm{O}(\sqrt{\mathrm{tr}(G)}/s+1/\sqrt{s})$, 其中, $n=2sa$.

注 10. 当假定条件退化为独立同分布时, 式 (64) 就退化为式 (42) 和 (43) 中的特殊情况.

3) 假定

$$\begin{cases}\text{非平稳分布且 } \beta\text{-mixing}\\ \mathcal{H}=\{f: f \text{ 为从 } \mathcal{X} \text{ 到 } \mathcal{Y} \text{ 的映射}\}\end{cases}$$

Kuznetsov 在文献 [42] 中提出了子样选取 (Sub-sample selection) 技巧, 具体如下:

记样本数据集 Z_1^n, 给定常数 $a\geq 1$, 存在 $k\geq 1$, s.t., $n=ka$. 定义子样

$$Z^{(j)}=(Z_{1+j},Z_{2+j},\cdots,Z_{k-1+j}) \tag{65}$$

其中, $j=0,1,\cdots,a-1$.

基于该技巧以及文献 [88] 中的独立块技巧, 考虑分析了平均错误 (Averaged error) 下的一致偏差, 建立了与 Rademacher 复杂度之间的关系不等式

$$\begin{aligned}&\mathrm{E}_{Z_{n+s}}[Q(Z_{n+s},h)]-\frac{1}{n}\sum_{i=1}^n Q(Z_i,h)\leq\\ &2\max\left(\mathrm{E}(R^{\beta\text{-mixing}}_{|I_1|}\mathcal{F}),\mathrm{E}(R^{\beta\text{-mixing}}_{|I_2|}\mathcal{F})\right)+\\ &\max(\Delta_1,\Delta_2)+\\ &M\max\left(\sqrt{\sum_{j=1}^m a_{2j-1}^2},\sqrt{\sum_{j=1}^m a_{2j}^2},\sqrt{\frac{\log\frac{2}{\delta}}{2n^2}}\right)\leq\\ &\frac{2}{a}\sum_{j=1}^{2a}\left(\frac{1}{k}\mathrm{E}[\sup_{h\in\mathcal{H}}\sum_{i=1}^k\sigma_i Q(Z_{a(2i-1)+j},h)]\right)+\\ &\frac{2}{n}\sum_{t=1}^n\sup_{h\in\mathcal{H}}|\mathrm{E}_{Z_{n+s}}[Q(Z_{n+s},h)]-\mathrm{E}_{Z_t}[Q(Z_t,h)]|+\\ &M\sqrt{\frac{\log\frac{2}{\delta}}{8k}}\end{aligned} \tag{66}$$

其中, I_2 表示 Z_1^n 中被划入偶数编号块的样本数据点对应的指标组成的集合, $\Delta_i=\frac{1}{|I_i|}\sum_{t\in I_i}\sup_{h\in\mathcal{H}}|\mathrm{E}_{Z_{n+s}}[Q(Z_{n+s},h)]-\mathrm{E}_{Z_t}[Q(Z_t,h)]|, i=1,2$. 其他符号见第 3 节中块 Rademacher 复杂度以及附录 A 独立块技巧的说明.

同时, 还讨论研究了路径相依错误 (Path-dependent error) 下的一致偏差, 建立了与 Rademacher 复杂度之间的关系不等式

$$\begin{aligned}&\mathrm{E}_{Z_{n+s}}[Q(Z_{n+s},h)]-\frac{1}{n}\sum_{i=1}^n Q(Z_i,h)\leq\\ &\frac{1}{n}\sum_{t=1}^n\sup_{h\in\mathcal{H}}|\mathrm{E}_{Z_{n+s}}[Q(Z_{n+s},h)]-\mathrm{E}_{Z_t}[Q(Z_t,h)]|+\\ &2\mathrm{E}(R^{\mathrm{mar}}_{n-s}\mathcal{F})+\mathrm{O}\left(\sqrt{\frac{(\log n)^3}{n}}\right)\end{aligned} \tag{67}$$

又注意到样本数据的非平稳性, 可能会导致样本分布不收敛到目标分布, 所以定义了比 β-mixing 更强的条件: 随机过程的极限收敛于平稳分布, 并基于次根函数和局部 Rademacher 复杂度, 讨论了该极限平稳分布下学习模型的泛化误差情况. 另外, Kuznetsov 在文献 [43] 中进一步推广了文献 [42] 中有关泛化界分析的结果, 并且基于推广后的结果设计了非平稳时间序列预测的有关算法.

注 11. 若非平稳分布退化为平稳分布, 则上述式 (66) 退化为类似式 (64) 中的情况; 若非平稳分布

[17] 这里, 处理的 $\mathcal{F}$ 是特殊的, 实际上即是假设空间 $\mathcal{H}$.

退化为独立不同分布, 则上述式 (66) 退化为类似式 (61) 中的情况; 若非平稳分布退化为独立同分布, 则上述式 (66) 退化为类似由式 (42) 和 (43) 得到的特殊情况.

另外, 平稳或非平稳的 β-mixing 的 Rademacher 泛化性能分析结果, 目前还未见到用于在线学习, 然而在在线学习中经常会出现 "随时间推移, 未来对过去的依赖充分地小" 的情形, 这正是附录中注 A1 所描述的 β-mixing 的直观意义. 所以, 平稳或非平稳的 β-mixing 学习模型的 Rademacher 泛化性能分析结果, 将会有助于在线学习泛化性能的分析研究.

4) 假定

$$\begin{cases} \text{鞅} \\ \mathcal{H} = \{f : f \text{ 为从 } \mathcal{Z} \text{ 到 } [-1,1] \text{ 的映射}\} \end{cases}$$

其中, $\mathcal{Z}$ 为可分度量空间.

Rakhlin 等在文献 [39−41] 中基于给出的序列 Rademacher 复杂度, 讨论了学习模型的一致收敛性.

a) 建立鞅差序列与序列 Rademacher 复杂度的关系

$$\mathrm{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathrm{E}[f(X_t)|\mathcal{A}_{t-1}]-f(X_t))\right)\leq 2\mathrm{E}(R_n^{\mathrm{mar}}\mathcal{F}) \tag{68}$$

b) 给出序列 Rademacher 复杂度的收敛速度

$$\mathrm{E}(R_n^{\mathrm{mar}}\mathcal{F}) = \mathrm{O}\left(\frac{B}{\sqrt{n}}\right) \tag{69}$$

其中, $B = \inf_{z\in Z}\sup_{f,f'\mathcal{F}}(f(z)-f(z'))\geq 0$.

c) 结合 a) 和 b), 当 $n\to 0$ 时, 保证了鞅意义下学习过程的一致收敛性. Rakhlin 等还进一步讨论了相关结果在在线学习方面的应用.

相比其他 non-i.i.d. 方面的研究, 基于鞅的学习模型泛化界研究工作尚不完善.

注 12. 本节主要讨论了样本数据集分布为 non-i.i.d. 的情形, 从经验过程极大泛函的角度, 基于 Rademacher 复杂度分析讨论学习模型的泛化能力. 在这部分的讨论中, 还引入了独立块与子样选取等一些新的数学处理技巧. 为便于阅读, 相关内容总结至表 3.

6　小结与展望

学习模型的泛化能力分析是当前统计学习理论中的研究热点和难点, 经过很多学者的努力 (如文献 [18−21, 92] 等), Rademacher 复杂性度量已经发展成为一种度量函数空间容量, 进行学习模型泛化能力分析的极其有效的工具. 与已有的 VC 维等函数空间复杂性度量方法相比, Rademacher 复杂度与一致偏差具有更加紧密的关联性, 并且更能客观地定量刻画假设空间的复杂性[45], 因此, 在统计学习理论研究中越来越体现出其重要意义与研究价值.

表 3　non-i.i.d. 情形的泛化界分析

Table 3　Generalization analysis for non-i.i.d.

样本数据集产生环境	假设空间	泛化能力
独立不同分布	1) $\mathcal{X}\to\mathcal{Y}$	$\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$ [37]
平稳分布且 β-mixing	2) $\mathcal{X}\to\mathbf{R}$	$\mathrm{O}\left(\frac{\sqrt{\mathrm{tr}(G)}}{n}+\frac{1}{\sqrt{n}}\right)$ [38]
非平稳分布且 β-mixing	3) $\mathcal{X}\to\mathcal{Y}$	非平稳性可能导致不收敛于 0[42]
鞅	4) $\mathcal{Z}\to[-1,1]$	一致收敛 [39−41]

从当前的研究现状来看, Rademacher 复杂度在统计学习理论的应用主要是: 在学习框架和样本数据空间的假定下, 针对具体的假设空间, 展开对学习模型泛化能力的分析研究. 在学习框架假定方面, 本文关注有监督学习框架, 而关于非监督学习框架下学习模型的泛化能力研究, Rademacher 复杂度也有一些最新的应用, 如, 在直推学习 (Transductive learning) 框架下提出的直推 Rademacher 复杂度和置换 (Permutational) Rademacher 复杂度[93−95] 等, 但相关研究尚不完善和成熟; 在样本数据空间假定方面, 当前关于 Rademacher 复杂度在统计学习理论中的应用主要是关于序列方面, 并且主要集中在序列的样本数据产生环境为 i.i.d. 情形, 在 non-i.i.d 情形, 虽有一些研究但也相对尚不完善. 同时, 对于非序列方面, 如, 有关随机场学习模型泛化能力的研究, 则处于空白阶段. 随着关于随机场的集中不等式研究成果[96−97] 的出现, 将会有助于 Rademacher 复杂度在随机场等非序列学习模型泛化能力方面的应用与研究.

另外, 随着 Rademacher 复杂度在统计学习理论泛化能力方面的进一步深入应用与研究, 一方面, 会有更紧致、更容易指导实践的泛化界被发现, 从而将有助于更好地指导实际算法的设计. 另一方面, 也会促进其他领域如认知心理学等的应用与发展[98−99].

附录 A

在本部分中, 考虑到本文的可读性, 给出前面讨论要用到的相关定义和技巧.

A1 覆盖数

覆盖数作为实变函数理论中的概念, 最早由 Kolmogorov 等给出[78], 本质上是从函数逼近论观点刻画函数空

间的复杂度, 其基本思想是利用有限个半径固定的球来逼近原本无限的函数空间. 覆盖数可以看作是对生长函数的一种推广[76]. 此处引用文献 [32] 中给出的定义形式:

假定 $(\mathcal{G}, d)$ 为度量空间, $\mathcal{F} \subset \mathcal{G}$. 任意 $\epsilon > 0$, 有任意 $f \in \mathcal{F}$, 存在 $g \in \mathcal{F}^{\Delta}$, s.t., $d(f, g) \leq \epsilon$, 则 $\mathcal{F}^{\Delta}$ 为 $\mathcal{F}$ 的一个 $\epsilon-$覆盖. 若 $\epsilon-$覆盖 $\mathcal{F}^{\Delta} \subset \mathcal{F}$, 则称 $\mathcal{F}^{\Delta}$ 为 $\mathcal{F}$ 的一个正则 $\epsilon-$覆盖. 记 $\mathcal{N}(\epsilon, \mathcal{F}, d) := \min\{|\mathcal{F}^{\Delta}| : \mathcal{F}^{\Delta} \subset \mathcal{F}$ 是 $\mathcal{F}$ 的一个 $\epsilon-$覆盖$\}$.

A2 $U-$过程

假设 X_i, $i = 1, 2, \cdots, n$ 为定义在 $\mathcal{X}$ 上且服从分布 P 的 i.i.d. 随机变量序列, 记函数 $f : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$, 且 f 是对称的, $\mathcal{F}$ 是 f 的集合. 则

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} f(X_i, x_j), \ \forall f \in \mathcal{F} \tag{A1}$$

称为 $U-$过程[100].

$U-$过程最初由 Nolan 等引入随机过程理论, 是一簇指标取自对称核函数集合的 $U-$统计量, Peña 等为了研究退化 $U-$过程的渐进行为, 提出了 $m \in \mathbf{N}$ 阶 Rademacher chaos 复杂度的定义[35].

A3 平稳性[101]

随机过程 $\{X_t, t \in T\}$ 称作是平稳的, 若有 $\forall n \geq 1$, t_1, t_2, $\cdots$, $t_n \in T$ 和 $h \in \mathbf{R}$, $t_1 + h$, $t_2 + h$, $\cdots$, $t_n + h \in T$, 有

$$\mathrm{P}(X_{h+t_1}, \cdots, X_{h+t_n}) = \mathrm{P}(X_{t_1}, \cdots, X_{t_n}) \tag{A2}$$

其中, P 为 $\{X_t, t \in T\}$ 对应的联合分布.

A4 β-mixing[42−43]

设 $\{Z_t\}_{t=-\infty}^{+\infty}$ 为 $\mathcal{Z}$ 上对应联合分布为 P 的双无限 (Doubly infinite) 随机变量序列, 则 β-mixing 系数记为: 任意 $a \in \mathbf{N}$ 且 $a > 0$,

$$\beta(a) := \sup_t \mathrm{E}_{Z_{-\infty}^t}[\sup_{A \in \sigma(Z_{t+a}^{\infty})} |\mathrm{P}_{t+a}^{\infty}(A|Z_{-\infty}^t) - \mathrm{P}_{t+a}^{\infty}(A)|] \tag{A3}$$

其中, $\sigma(Z_{t+a}^{\infty})$ 表示由 Z_{t+a}^{∞} 生成的 $\sigma-$代数, 则称 P 是 β-mixing, 若 $a \to \infty$, 有

$$\beta(a) \to 0 \tag{A4}$$

注 A1. β-mixing 的直观意义: 随着时间的推移, 未来对过去的依赖充分地小.

A5 独立块技巧[42, 88]

独立块技巧最早可以追溯至 1927 年 Bernstein 的工作.

记样本 $Z_1^{\mathrm{T}} = \{Z_1, Z_2, \cdots, Z_T\}$, 将其分为大小为 a_i, $i = 1, 2, \cdots, 2m$ 的 $2m$ 块, 且 $T = \sum_{i=1}^{2m} a_i$, 即

$$Z_1^{\mathrm{T}} = \{Z(1), Z(2), \cdots, Z(2m)\} \tag{A5}$$

其中, $Z(1) = Z_{l(i)}^{u(i)}$, $l(i) = 1 + \sum_{j=1}^{i-1} a_j$, $u(i) = 1 + \sum_{j=1}^{i} a_j$, $i = 1, 2, \cdots, 2m$. 于是, 有

奇数编号块

$$Z^o = \{Z(1), Z(3), \cdots, Z(2m-1)\} \tag{A6}$$

偶数编号块

$$Z^e = \{Z(2), Z(4), \cdots, Z(2m)\} \tag{A7}$$

对于表达式 (A6) 和 (A7), 分别构造表达式 (A8) 和 (A9) 如下:

$$\tilde{Z}^o = \{\tilde{Z}(1), \tilde{Z}(3), \cdots, \tilde{Z}(2m-1)\} \tag{A8}$$

和

$$\tilde{Z}^e = \{\tilde{Z}(2), \tilde{Z}(4), \cdots, \tilde{Z}(2m)\} \tag{A9}$$

其中, 在式 (A8) 中 $\tilde{Z}(i)$, $i = 1, 3, \cdots, 2m-1$ 是独立随机变量序列, 且 $\tilde{Z}(i)$ 和 $Z(i)$, $i = 1, 3, \cdots, 2m-1$ 具有相同分布. 在式 (A9) 中 $\tilde{Z}(i)$, $i = 2, 4, \cdots, 2m$ 是独立随机变量序列, 且 $\tilde{Z}(i)$ 和 $Z(i)$, $i = 2, 4, \cdots, 2m$ 具有相同分布.

A6 鞅[71]

随机过程 $\{X_t, t \in T\}$, 其中, 集合 T 为任意指标集, X_t: $\mathcal{X} \to \mathbf{R}$. 若满足: 任意 $t \in T$, X_t 是关于 $\mathcal{A}_t-$可测的随机变量; $\mathrm{E}|X_t| < \infty$, $t \in T$; 并且任意 $s, t \in T$, $s \leq t$, 有

$$\mathrm{E}(X_t|\mathcal{A}_s) = X_s, \ \text{a.e.} \tag{A10}$$

则称 $\{X_t, t \in T\}$ 为鞅. 这里, $\sigma(\mathcal{A}_t)$ 表示由 $\{X_s \leq t\}$ 生成的 $\sigma-$代数.

注 A2. 为简单起见, 假定指标集 T 为离散集. 鞅反映了某种无后效性, 即, 当已知时刻 t 以及之前的值 X_1, X_2, $\cdots$, X_t, 那么, $t+1$ 时刻的值 X_{t+1} 对 $X_1, X_2, \cdots, X_t$ 的条件期望与时刻 t 之前的值 $X_1, X_2, \cdots, X_{t-1}$ 无关, 且等于 X_t.

References

1 Tikhonov A N, Arsenin V Y. *Solution of Ill-posed Problems*. Washington: Winston and Sons, 1977.

2 Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, **19**(6): 716−723

3 Akaike H. A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 1978, **30**(1): 9−14

4 Schwarz G. Estimating the dimension of a model. *Annals of Statistics*, 1978, **6**(2): 461−464

5 Kolmogorov A N. On tables of random numbers. *Sankhya: The Indian Journal of Statistics*, 1963, **25**: 369−376

6 Kolmogorov A N. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1965, **1**(1): 1−7

7 Chaitin G J. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 1966, 13(4): 547−569

8 Solomonoff R J. A formal theory of inductive inference. Part II. *Information and Control*, 1964, **7**(2): 224−254

9 Wallace C S, Boulton D M. An information measure for classification. *The Computer Journal*, 1968, **11**(2): 185−194

10 Vapnik V N, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 1971, **16**(2): 264−280

11 Vapnik V N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

12 Vapnik V N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 1999, **10**(5): 988−999

13 Popper K R. *The Logic of Scientific Discovery*. United Kingdom: Hutchinson, 1959.

14 Valiant L G. A theory of the learnable. *Communications of the ACM*, 1984, **27**(11): 1134−1142

15 Kearns M J, Schapire R E. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 1994, **48**(3): 464−497

16 Bartlett P L, Long P M, Williamson R C. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 1996, **52**(3): 434−452

17 Shawe-Taylor J, Bartlett P L, Williamson R C, Anthony M. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998, **44**(5): 1926−1940

18 Koltchinskii V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 2001, **47**(5): 1902−1914

19 Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: risk bounds and structural results. *The Journal of Machine Learning Research*, 2003, **3**: 463−482

20 Bartlett P L, Bousquet O, Mendelson S. Local Rademacher complexities. *The Annals of Statistics*, 2005, **33**(4): 1497−1537

21 Koltchinskii V. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 2006, **34**(6): 2593−2656

22 Koltchinskii V, Panchenko D. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*. Boston: Birkhäuser, 2004. 443−457

23 Bousquet O, Koltchinskii V, Panchenko D. Some local measures of complexity of convex hulls and generalization bounds. *Computational Learning Theory*. Sydney, Australia: Springer, 2004. 59−73

24 Zhang Hai, Xu Zong-Ben. A survey on learning theory (I): stability and generalization. *Chinese Journal of Engineering Mathematics*, 2008, **25**(1): 1−9
(张海, 徐宗本. 学习理论综述 (I): 稳定性与泛化性. 工程数学学报, 2008, **25**(1): 1−9)

25 Hu Zheng-Fa. The statistic complexity estimates and its application to machine learning. *Acta Automatica Sinica*, 2008, **34**(10): 1332−1336
(胡政发. 统计量复杂性估计及其在机器学习中的应用. 自动化学报, 2008, **34**(10): 1332−1336)

26 Chen Jiang-Hong. Rademacher Complexities and the Generalization Performance of SVM [Master dissertation], Hubei University, China, 2005.
(陈将宏. Rademacher 复杂性与支持向量机的推广性能 [硕士学位论文], 湖北大学, 中国, 2005.)

27 Lei Y W, Ying Y M. Generalization analysis of multi-modal metric learning [Online], available: https://www.researchgate.net/publication/282969709, November 3, 2015

28 Lei Y W, Ding L X, Bi Y Z. Local Rademacher complexity bounds based on covering numbers [Online], available: http://arxiv.org/pdf/1510.01463v1.pdf, October 6, 2015

29 Lei Y W, Ding L X. Refined rademacher chaos complexity bounds with applications to the multikernel learning problem. *Neural Computation*, 2014, **26**(4): 739−760

30 Lei Y W, Ding L X, Zhang W S. Generalization performance of radial basis function networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(3): 551−564

31 Lei Y W, Ding L X, Wu W L. Universal learning using free multivariate splines. *Neurocomputing*, 2013, **119**(16): 253−263

32 Lei Yun-Wen. Rademacher Complexities with Applications to Machine Learning [Ph. D. dissertation], Wuhan University, China, 2015.
(雷云文. Rademacher 复杂度及其在机器学习中的应用 [博士学位论文], 武汉大学, 中国, 2015.)

33 Xu C, Liu T L, Tao D C, Xu C. Local Rademacher complexity for multi-label learning [Online], available: http://arxiv.org/pdf/1410.6990.pdf, October 26, 2014

34 Gao W, Zhou Z H. Dropout Rademacher complexity of deep neural networks. *Science China: Information Sciences*, 2016, **59**(7): 072104: 1−072104: 12

35 de la Peña V, Giné E. *Decoupling: From Dependence to Independence*. New York: Springer-Verlag, 1999.

36 Cao Q, Guo Z C, Ying Y M. Generalization bounds for metric and similarity learning. *Machine Learning*, 2016, **102**(1): 115−132

37 Mohri M, Medina A M. New analysis and algorithm for learning with drifting distributions. In: Proceedings of the 23rd International Conference on Algorithmic Learning Theory. Lyon, France: Springer-Verlag, 2012. 124−138

38 Mohri M, Rostamizadeh A. Rademacher complexity bounds for non-I.I.D. processes. In: Proceedings of the 2008 Advances in Neural Information Processing Systems 21. Vancouver, BC, Canada: Curran Associates, Inc., 2008. 1097−1104

39 Rakhlin A, Sridharan K. On martingale extensions of Vapnik-Chervonenkis theory with applications to online learning. *Measures of Complexity*. Switzerland: Springer International Publishing, 2015. 197−215

40 Rakhlin A, Sridharan K, Tewari A. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 2015, **161**(1−2): 111−153

41 Rakhlin A, Sridharan K, Tewari A. Online learning: random averages, combinatorial parameters, and learnability. In: Proceedings of the 2010 Advances in Neural Information Processing Systems 23. Vancouver, BC, Canada: Curran Associates, Inc., 2010. 1984−1992

42 Kuznetsov V, Mohri M. Generalization bounds for time series prediction with non-stationary processes. In: Proceedings of the 25th International Conference on Algorithmic Learning Theory. Bled, Slovenia: Springer International Publishing, 2014. 260−274

43 Kuznetsov V, Mohri M. Forecasting non-stationary time series: from theory to algorithms [Online], available: http://www.cims.nyu.edu/~vitaly/pub/fts.pdf, July 12, 2016

44 Anguita D, Ghio A, Oneto L, Ridella S. A deep connection between the Vapnik-Chervonenkis entropy and the Rademacher complexity. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(12): 2202−2211

45 Kääriäinen M. Relating the Rademacher and VC Bounds, Technical Report, C-2004-57, Department of Computer Science, University of Helsinki, Finland, 2004.

46 Srebro N, Sridharan K. Note on refined dudley integral covering number bound [Online], available: http://ttic.uchicago.edu/~karthik/dudley.pdf, July 12, 2016

47 Srebro N, Sridharan K, Tewari A. Smoothness, low noise and fast rates. In: Proceedings of the 2010 Advances in Neural Information Processing Systems 23. Vancouver, BC, Canada: Curran Associates, Inc., 2010. 2199−2207

48 V′Yugin V V. VC dimension, fat-shattering dimension, Rademacher averages, and their applications. *Measures of Complexity*. Switzerland: Springer International Publishing, 2015. 57−74

49 Ying Y M, Campbell C. Rademacher chaos complexities for learning the kernel problem. *Neural Computation*, 2010, **22**(11): 2858−2886

50 Massart P. Some applications of concentration inequalities to statistics. *Annales De La Faculté Des Sciences De Toulouse*, 2000, **9**(2): 245−303

51 Gnecco G, Sanguineti M. Approximation error bounds via Rademacher′s complexity. *Applied Mathematical Sciences*, 2008, **2**(4): 153−176

52 Boucheron S, Bousquet O, Lugosi G. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 2005, **9**: 323−375

53 Oneto L, Ghio A, Ridella S, Anguita D. Global Rademacher complexity bounds: from slow to fast convergence rates. *Neural Processing Letters*, 2016, **43**(2): 567−602

54 Oneto L, Ghio A, Ridella S, Anguita D. Fast convergence of extended Rademacher complexity bounds. In: Proceedings of the 2015 International Joint Conference on Neural Networks. Killarney, Ireland: IEEE, 2015. 1−10

55 Oneto L, Ghio A, Anguita D, Ridella S. An improved analysis of the Rademacher data-dependent bound using its self bounding property. *Neural Networks*, 2013, **44**: 107−111

56 Yu H F, Jain P, Kar P, Dhillon I S. Large-scale multi-label learning with missing labels. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014. 593−601

57 Ding Wan-Ding. *Essentials of Measure Theory*. Hefei: Anhui People′s Publishing House, 2005.
(丁万鼎. 测度论概要. 合肥: 安徽人民出版社, 2005.)

58 Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

59 van der Vaart A W, Wellner J A. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.

60 Ledoux M, Talagrand M. *Probability in Banach Spaces: Isoperimetry and Processes*. Berlin: Springer-Verlag, 1991.

61 Dudley R M. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1967, **1**(3): 290−330

62 Lozano F. Model selection using Rademacher penalization. In: Proceedings of the 2nd ICSC Symposium on Neural Networks. London: Academic Press, 2000.

63 Boucheron S, Lugosi G, Massart P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. United Kingdom: Oxford University Press, 2013.

64 Zhou Zhi-Hua. *Machine Learning*. Beijing: Tsinghua University Press, 2016.
(周志华. 机器学习. 北京: 清化大学出版社, 2016.)

65 Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. Cambridge: MIT Press, 2012.

66 Wang Hong-Qiao, Sun Fu-Chun, Cai Yan-Ning, Chen Ning, Ding Lin-Ge. On multiple kernel learning methods. *Acta Automatica Sinica*, 2010, **36**(8): 1037−1050
(汪洪桥, 孙富春, 蔡艳宁, 陈宁, 丁林阁. 多核学习方法. 自动化学报, 2010, **36**(8): 1037−1050)

67 McFee B, Lanckriet G. Learning multi-modal similarity. *Journal of Machine Learning Research*, 2011, **12**: 491−523

68 Xie P T, Xing E P. Multi-modal distance metric learning. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI, 2013. 1806−1812

69 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors [Online], available: http://arxiv.org/pdf/1207.0580v1.pdf, July 3, 2012

70 Wan L, Zeiler M, Zhang S X, LeCun Y, Fergus R. Regularization of neural networks using dropconnect. In: Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA, 2013. 1058−1066

71 Rudin W. *Functional Analysis* (2nd edition). New York: McGraw-Hill, 1991.

72 Mendelson S. A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*. Berlin Heidelberg: Springer, 2003. 1−40

73 Sauer N. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 1972, **13**(1): 145−147

74 Pollard D. *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.

75 Duan H H. Bounding the fat shattering dimension of a composition function class built using a continuous logic connective [Online], available: http://arxiv.org/pdf/1105.4618v1.pdf, May 23, 2011

76 Anthony M, Bartlett P L. *Neural Network Learning-Theoretical Foundations*. Cambridge: Cambridge University Press, 2009.

77 Alon N, Ben-David S, Cesa-Bianchi N, Haussler D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 1997, **44**(4): 615−631

78 Kolmogorov A N, Tihomirov V M. ε-entropy and ε-capacity of sets in functional space. *American Mathematical Society Translations*, 1961, **17**(2): 277−364

79 Bousquet O. New approaches to statistical learning theory. *Journal of the Institute of Statistical Mathematics*, 2003, **55**(2): 371−389

80 Wu P C, Hoi S C H, Xia H, Zhao P L, Wang D Y, Miao C Y. Online multimodal deep similarity learning with application to image retrieval. In: Proceedings of the 21st ACM International Conference on Multimedia. New York, USA: ACM, 2013. 153−162

81 Xia H, Wu P C, Hoi S C H. Online multi-modal distance learning for scalable multimedia retrieval. In: Proceedings of the 6th International Conference on Web Search and Data Mining. New York, USA: ACM, 2013. 455−464

82 Krzyzak A, Linder T. Radial basis function networks and complexity regularization in function learning. *IEEE Transactions on Neural Networks*, 1998, **9**(2): 247−256

83 Niyogi P, Girosi F. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 1996, **8**(4): 819−842

84 Park J, Sandberg I W. Universal approximation using radial-basis-function networks. *Neural Computation*, 1991, **3**(2): 246−257

85 Girosi F. Approximation error bounds that use VC-bounds [Online], available: https://www.researchgate.net/publication/2782224_Approximation_Error_Bounds_That_Use_Vc-Bounds, February 18, 2013

86 Györfi L, Kohler M, Krzyżak A, Walk H. *A Distribution-Free Theory of Nonparametric Regression*. Berlin: Springer-Verlag, 2010.

87 Haussler D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 1992, **100**(1): 78−150

88 Yu B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 1994, **22**(1): 94−116

89 Meir R. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 2000, **39**(1): 5−34

90 Bartlett P L. Learning with a slowly changing distribution. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York, USA: ACM, 1992. 243−252

91 Mansour Y, Mohri M, Rostamizadeh A. Domain adaptation: learning bounds and algorithms. In: Proceedings of the 22nd Annual Conference on Learning Theory. Montreal, QC: Morgan Kaufmann Publishers, 2009.

92 Anguita D, Ghio A, Oneto L, Ridella S. In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(9): 1390−1406

93 El-Yaniv R, Pechyony D. Transductive Rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 2007, **35**: 193−234

94 Tolstikhin I, Blanchard G, Kloft M. Localized complexities for transductive learning [Online], available: http://arxiv.org/pdf/1411.7200.pdf, November 26, 2014

95 Tolstikhin I, Zhivotovskiy N, Blanchard G. Permutational Rademacher complexity. *Algorithmic Learning Theory*. Switzerland: Springer International Publishing, 2015. 209−223

96 Chazottes J R, Collet P, Külske C, Redig F. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 2007, **137**(1): 201−225

97 Belomestny D, Spokoiny V. Concentration inequalities for smooth random fields. *Theory of Probability and Its Applications*, 2014, **58**(2): 314−323

98 Zhu X J, Rogers T T, Gibson B R. Human Rademacher complexity. In: Proceedings of the 2009 Advances in Neural Information Processing Systems 22. Vancouver, BC, Canada: Curran Associates, Inc., 2009. 2322−2330

99 Vahdat M, Oneto L, Ghio A, Anguita D, Funk M, Rauterberg M. Human algorithmic stability and human Rademacher complexity. In: Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: Katholieke Universiteit Leuven, 2015. 313−318

100 Nolan D, Pollard D. U-processes: rates of convergence. *The Annals of Statistics*, 1987, **15**(2): 780−799

101 Karlin S, Taylor H M. *A First Course in Stochastic Processes* (2nd edition). New York: Academic Press, 1975.

吴新星 复旦大学计算机科学技术学院访问学者, 上海电子信息职业技术学院计算机应用系副教授. 主要研究方向为统计学习理论与形式化方法.
E-mail: xinxingwu@yeah.net
(**WU Xin-Xing** Visiting scholar at the School of Computer Science, Fudan University, associate professor in the Department of Computer, Shanghai Technical Institute of Electronics and Information. His research interest covers statistical learning theory and formal methods.)

张军平 复旦大学计算机科学技术学院教授. 主要研究方向为机器学习, 智能交通, 生物认证与图像处理. 本文通信作者. E-mail: jpzhang@fudan.edu.cn
(**ZHANG Jun-Ping** Professor at the School of Computer Science, Fudan University. His research interest covers machine learning, intelligent transportation systems, biometric authentication, and image processing. Corresponding author of this paper.)