



本体学习算法的两类LOO一致稳定性和广义界

朱林立, 华钢, 高伟

引用本文:

朱林立, 华钢, 高伟. 本体学习算法的两类LOO一致稳定性和广义界[J]. *智能系统学报*, 2022, 17(3): 471–479.

ZHU Linli, HUA Gang, GAO Wei. Two classes of LOO uniform stability and generalization bounds of ontology learning algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(3): 471–479.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202101015>

您可能感兴趣的其他文章

仿人机器人步态平衡泛化模型的建立与仿真

Modeling and simulation of humanoid robot gait balance generalization

智能系统学报. 2020, 15(3): 537–545 <https://dx.doi.org/10.11992/tis.201810017>

深度度量学习综述

A brief introduction to deep metric learning

智能系统学报. 2019, 14(6): 1064–1072 <https://dx.doi.org/10.11992/tis.201906045>

结合谱聚类的标记分布学习

Label distribution learning based on spectral clustering

智能系统学报. 2019, 14(5): 966–973 <https://dx.doi.org/10.11992/tis.201809019>

关于深度学习的综述与讨论

Overview on deep learning

智能系统学报. 2019, 14(1): 1–19 <https://dx.doi.org/10.11992/tis.201808019>

基于排序学习的视频摘要

Video summarization based on learning to rank

智能系统学报. 2018, 13(6): 921–927 <https://dx.doi.org/10.11992/tis.201806013>

在线学习的大规模网络流量分类研究

Large-scale network traffic classification based on online learning

智能系统学报. 2016, 11(3): 318–327 <https://dx.doi.org/10.3969/j.issn.1673-4785.201603033>



微信公众平台



期刊网址

DOI: 10.11992/tis.202101015

网络出版地址: <https://kns.cnki.net/kcms/detail/23.1538.TP.20220324.1711.008.html>

本体学习算法的两类 LOO 一致稳定性和广义界

朱林立^{1,2}, 华钢¹, 高炜³

(1. 中国矿业大学 信息与控制工程学院, 江苏 徐州 221116; 2. 江苏理工学院 计算机工程学院, 江苏 常州 213001; 3. 云南师范大学 信息学院, 云南 昆明 650500)

摘要: 近年来, 随着本体研究的深入, 各类机器学习方法被尝试应用于本体相似度计算和本体映射获取。稳定性是本体学习算法的必要条件, 它从本质上体现了算法的可用性, 即要求本体学习算法的最优解不会受到本体样本的小幅度调整而发生大的变化。本文研究了删除一个本体样本点的条件下, 对本体学习算法的期望误差与经验误差的差值产生的影响。分别在本体学习算法一致稳定和假设空间一致稳定两种不同的框架下, 利用统计学习理论的技巧, 得到对应广义界的上界估计。

关键词: 本体; 机器学习; 稳定性; 广义界; 本体数据依赖函数; 本体样本依赖假设集; 拉德马赫复杂度; 经验拉德马赫复杂度

中图分类号: TP391 文献标志码: A 文章编号: 1673-4785(2022)03-0471-09

中文引用格式: 朱林立, 华钢, 高炜. 本体学习算法的两类 LOO 一致稳定性和广义界 [J]. 智能系统学报, 2022, 17(3): 471-479.

英文引用格式: ZHU Linli, HUA Gang, GAO Wei. Two classes of LOO uniform stability and generalization bounds of ontology learning algorithm[J]. CAAI transactions on intelligent systems, 2022, 17(3): 471-479.

Two classes of LOO uniform stability and generalization bounds of ontology learning algorithm

ZHU Linli^{1,2}, HUA Gang¹, GAO Wei³

(1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; 2. School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, China; 3. School of Information, Yunnan Normal University, Kunming 650500, China)

Abstract: Recently, with deepening ontology research, efforts have been made to apply various machine-learning methods to ontology similarity calculations and mapping acquisitions. Stability is a necessary condition for ontology-learning algorithms. It requires that the optimal solution of the algorithm does not undergo major changes due to small adjustments to the ontology samples; thus, it essentially reflects the usability of the algorithm. In this study, we investigated the effect of deleting an ontology sample point on the difference between the expected and empirical errors of the ontology-learning algorithm. In two settings of the ontology-learning algorithm—uniform stability and hypothetical space uniform stability—obtained using statistical learning theory, the corresponding upper bound estimates of generalized bounds are determined.

Keywords: ontology; machine learning; stability; generalized bound; ontology data-dependent function; ontology sample dependent hypothesis set; Rademacher complexity; empirical Rademacher complexity

本体的概念首先属于西方哲学的范畴, 是指客观存在在逻辑层面上的表达和概括。作为形而上学的一个分支, 本体论在哲学领域的主要研究问题是“是否存在非物质事物”。20 世纪 80 年代引入到人工智能领域, 之后对本体有自己的定义。

近年来, 本体作为概念语义表示的工具, 与其他数据库相比, 有着结构化存储、易于在不同本体之间建立映射和本体对齐等优势, 因而被广泛应用于各个学科, 在生物、化学、医学、材料、能源等领域都能看到本体在特定的场景下发挥作用。

Skalle 等^[1] 阐述了如何使用知识建模和本体工程方法开发模型, 并利用本体模型来预测钻井过程中的井下故障。Sobral 等^[2] 提出了一个基于

收稿日期: 2021-01-09. 网络出版日期: 2022-03-25.

基金项目: 国家自然科学基金项目 (51574232).

通信作者: 朱林立. E-mail: zhulinli@jsut.edu.cn.

本体的框架来支持来自智能交通系统的数据的集成和可视化。Al-Sayed 等^[3]设计了一种称为 Cloud-FNF 的综合云本体, 根据该本体结构, 云服务被分类为若干个类别。Tebes 等^[4]给出分析和记录软件测试本体的系统审查结果的方法。Pradeep 等^[5]在基于本体的平衡二叉树中搜索预定的用户查询, 最后使用 Okapi BM25 对相关结果进行排序并交付给用户。Hema 等^[6]提出了一种新的基于信任的隐私保护框架, 通过本体服务排序 TBPFOSR 进行用户身份验证。Messaoudi 等^[7]给出了关于疾病治疗医学本体研究的综述。Mantovani 等^[8]提出了一种本体驱动的地质图知识表示。本体形式语言允许通过语义类别和属性对地球科学家的解释进行机器可读编码, 并支持知识共享和互操作性。Abeyasinghe 等^[9]通过在基因本体 (gene ontology, GO) 本体概念的基于序列的表示之上, 利用新的术语代数以及 3 个条件规则 (单调性、交集和子概念规则) 开发了基于包含的子项推理框架。Kossmann 等^[10]提出了潜艇探测着陆器的本体驱动框架。

随着越来越多的本体被定义和应用, 以及越来越多的本体算法被设计并尝试用于不同的本体工程应用领域, 学习算法也被逐渐应用于本体。本体学习算法就是将机器学习方法与本体自身结构相融合, 通过本体样本的学习得到本体函数。具体地说, 本体用图结构表示, 设 $G = (V, E)$ 为本体图对应某个本体 O , 其顶点对应 O 中概念, 顶点之间的边对应两个顶点某种直接联系。本体可以用有向图来表示, 其中边的权值由边的类型等因素决定。本体学习算法可以看成图算法, 通过本体样本 S 按照一定的学习规则学习得到本体函数 $f: V \rightarrow \mathbf{R}$ 。该本体函数的作用是将整个本体图映射到一维实数轴, 每个本体顶点 (对应本体概念) 映射为一维实数。在最开始的本体数值化表示中, 往往要求每个本体顶点都用一个 p 维向量来表示, 因此本体学习算法本质上就变成一种降维算法, 目标是得到本体函数 $f: \mathbf{R}^p \rightarrow \mathbf{R}$ 。相关本体学习这方面的研究可参考文献 [11-19]。

稳定性是一个本体学习算法在具体工程应用领域可以使用的基本条件, 它要求算法在学习样本集发生轻微改变时输出结果不会发生本质性的改变。常见的变换有: 删除一个样本 (leave one out, LOO), 删除两个样本 (leave two out, LTO), 替换一个样本 (place one, PO)。只有满足稳定性的本体学习算法才具备泛化性, 才有可能对同类数据发挥作用。已有部分文献对本体算法的稳定性进行了讨论, 可参考文献 [15, 20-21]。

具体地说, 通常把本体数据分成两类: 训练集

(样本集) 和测试集。通过本体样本的训练, 按照一定的本体学习算法得到本体函数, 再将本体函数应用于同类本体数据 (测试集)。我们希望通过从样本学习得到的本体函数同样适用于同一个本体的其他本体数据, 即算法具有很好的扩展性能。这必须要求本体学习算法具有稳定性, 即样本的轻微改变对学习得到的结果影响很小。从理论上说, 稳定性是一种学习的平滑性, 即最优函数的性能随着样本容量 n 的增加而缓缓提高, 其特征曲线是平滑过渡的, 不会在任何一点产生剧烈的震荡。而强烈的抖动意味着算法对于特定的本体样本点有着特定的反映, 进而说明本体学习算法本身不适合实验对象数据, 即这样的算法得到的最优本体函数不可能对训练集以外的数据产生很好的效果。

另一方面, 理论研究最关心的是算法的收敛阶和误差界。本文的动机是在稳定性和误差界之间找到必然的联系。因此如果一个本体学习算法有很好的稳定性, 那么它的学习误差就会控制在某个范围内。也就是说, 本文的理论结果暗示了稳定性不仅是本体学习算法具有泛化性的前提, 而且稳定性参数控制了本体学习算法广义界的上界。

本文针对删除一个样本的 LOO 操作, 给出对应的本体学习算法稳定性和本体函数假设空间一致稳定性定义。并利用统计学方法, 针对两类不同的 LOO 一致稳定性定义, 分别给出广义界的估计值, 这些界优于之前同框架 LOO 一致稳定性设定论文中得到的广义界。

1 本体学习算法及稳定性

设 Z 为本体数据集, 在非监督本体学习中, Z 即为本体图顶点数据 V , 对于监督本体学习则 $Z = V \times Y$ 。设本体学习算法 $A: Z^n \rightarrow F$, 其中 F 是本体函数空间, 即假设空间。 $A(S)$ 或者 $A(S, \cdot)$ 表示本体学习算法 A 作用在本体数据 S 的输出函数。 $l: F \times Z \rightarrow \mathbf{R}$ 为本体亏损函数 (也可以将其正则化, 即取值限定在区间 $[0, 1]$ 内, 进而也可以表示为 $l: F \times Z \rightarrow [0, 1]$), 给定本体函数 f 和本体样本点 v , 本体亏损函数表示为 $l(f, z)$ 。设 $S = \{z_1, z_2, \dots, z_n\}$ 为容量为 n 的本体样本。删除样本集 S 中的第 i ($i \in \{1, 2, \dots, n\}$) 个样本 v_i , 对应的样本集变为 $S^{\setminus i} = \{z_1, z_2, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}$, 称此类变换为 LOO (leave one out)。在样本 S 的表示中, 若算法为非监督本体学习算法, 则 $z_i = v_i$; 在监督本体学习下有 $z_i = (v_i, y_i)$ 。如果对于任意位置 i ($i \in \{1, 2, \dots, n\}$), 任意 S 和 $S^{\setminus i}$, 以及任意 $v \in V$, 有

$$|l(A(S), z) - l(A(S^{\setminus i}), z)| \leq \gamma$$

成立, 则称本体算法 A 关于本体亏损函数 l 存在 LOO 一致稳定 γ 。

设 $R_D[l(A(S))] = E_{z \sim D}[l(A(S), z)]$ 为本体算法 A 的期望误差, $\hat{R}_S[l(A(S))] = E_{z \sim S}[l(A(S), z)] = \frac{1}{n} \sum_{i=1}^n l(A(S), z_i)$ 为对应的经验误差。从而本体算法 A 的广义界就定义为期望误差和经验误差的差值, 记为

$$\Delta_{D-S}(l(A)) = R_D[l(A(S))] - \hat{R}_S[l(A(S))]$$

为了方便讨论且说明本文的结果具有更加一般的规律, 定理 1 给出的广义界对任意本体数据依赖函数都成立。具体地说, 设 $M: Z^n \times Z \rightarrow \mathbf{R}$ (也可以限制其值域 $M: Z^n \times Z \rightarrow [0, 1]$) 为本体数据依赖函数 (ontology data-dependent function), 它将给定本体数据集 S 和本体点 z 作为输入, 可以看作计算实值函数 $M(S, \cdot)$ 应用到 z 。这里 $M(S)$ 或者 $M(S, \cdot)$ 表示 M 作用在本体数据 S 上而得到的输出函数。事实上, 在本体学习框架下, 有 $M(S, z) = l(A(S), z)$, 只是前者的表示有另外的数据统计含义, 可以依赖于本体数据。在此设定下, 期望误差和经验误差分别表示为 $R_D[M(S)] = E_{z \sim D}[M(S, z)]$ 和 $\hat{R}_S[M(S)] = E_{z \sim S}[M(S, z)] = \frac{1}{n} \sum_{i=1}^n M(S, z_i)$ 。对应的广义界表示为 $\Delta_{D-S}(M) = R_D[M(S)] - \hat{R}_S[M(S)]$ 。

考虑将本体样本 S 作为输入, 本体数据依赖函数作为输出, 比如在本体监督学习中, 每个本体样本点带有标记信息 y , 因而 $M(S, z) = l_y(f(v), y)$, 即在本体样本 $z = (v, y)$ 上执行本体学习算法 $A(S)$ 。

若对所有 $S \in Z^n$, 任意 $i \in \{1, 2, \dots, n\}$, 任意 $z \in Z$ 有 $|M(S, z) - M(S^{\setminus i}, z)| \leq \gamma$ 成立, 则称本体数据依赖函数 $M: Z^n \times Z \rightarrow \mathbf{R}$ 存在 LOO 一致稳定 γ 。若对所有 $S, S' \in Z^n$, 其中 S 和 S' 是容量相同的两个本体数据集且只有一个数据不相同, 满足 $\forall E \subseteq Y$, 有

$$P[M(S) \in E] \leq e^\epsilon P[M(S') \in E]$$

则称算法 $A: Z^n \rightarrow Y$ 是 ϵ -差分隐私 (leave-one-out- ϵ -differentially private) 的。本体学习算法的假设空间 (hypothesis space) 就是本体函数选取的范围空间, 也称为假设集, 它是本体学习算法设计的核心。假设空间过大意味着函数选取的范围很大, 会导致算法不收敛; 而假设空间过小又会导致得到的最优本体函数性能不佳。理想的本体学习算法都会对假设空间有精巧的设计, 一般在数学上会选取凸函数空间, 常见的本体假设空间有再生核希尔伯特空间、索伯列夫空间等。由于假设空间是一个抽象的函数空间, 一般用抽象的工具来刻画它的大小, 比如覆盖数 (covering number)。

在一般学习框架下, 本体假设空间是由算法本身决定的, 独立于本体样本, 记为 F 。本文将考虑受本体数据集影响的假设空间, 称为本体样本

依赖假设集 (ontology sample dependent hypothesis set), 也称为本体数据依赖假设集 (ontology data dependent hypothesis set), 记为 F_S , 表示根据本体训练样本集 S 而选取的本体函数空间。 F 和 F_S 的关系是: 当 S 给定后, F_S 通常选取为 F 的一个子集, 即 $F_S \subset F$ 。进而可以写成 $F = \bigcup_{S \in Z^n} F_S$ 。这样, F 表示为本体数据依赖假设集的并集, 故也称 $F = (F_S)_{S \in Z^n}$ 为本体数据依赖假设集族。

给定本体样本容量 $n \in \mathbf{N}$ 。如果对任意 $i \in \{1, 2, \dots, n\}$, 任意本体样本集 S 和删除一个元素后的本体样本集 $S^{\setminus i}$, 存在某个 $\beta \geq 0$, 对任意 $f \in F_S$, 都存在某个 $f' \in F_{S^{\setminus i}}$, 使得对任意 $z \in Z$ 都有 $|l(f, z) - l(f', z)| \leq \beta$ 成立。则称本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 存在 LOO 一致稳定 β , 简称 β -稳定。

给定本体样本容量 $n \in \mathbf{N}$ 。若对任意 $S \in Z^n$, 都有 $E_{z \sim S} \left[\sup_{f \in F_S, f' \in F_{S^{\setminus i}}} (l(f, z) - l(f', z)) \right] \leq \chi$ 成立, 则称本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 存在 LOO-CV 稳定 χ , 其中 $\chi \geq 0$ 。更进一步, 若 $E_{S \sim D^n, z \sim S} \left[\sup_{f \in F_S, f' \in F_{S^{\setminus i}}} (l(f, z) - l(f', z)) \right] \leq \bar{\chi}$ 成立, 则称 F 存在 LOO-均值 CV 稳定 $\bar{\chi}$, 同样 $\bar{\chi} \geq 0$ 。

本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 的直径 Δ 和均值直径 $\bar{\Delta}$ 分别定义为

$$\Delta = \sup_{S \in Z^n, z \sim S} E \left[\sup_{f, f' \in F_S} (l(f, z) - l(f', z)) \right]$$

$$\bar{\Delta} = E_{S \in Z^n, z \sim S} \left[\sup_{f, f' \in F_S} (l(f, z) - l(f', z)) \right]$$

对任意两个本体样本集 $S, T \in Z^n$ 以及拉德马赫向量 σ , 集合 $S_{T, \sigma}$ 由 S 通过如下变换得到: 对于 $i \in \{1, 2, \dots, n\}$, 若发现 $\sigma_i = -1$, 则将 S 中第 i 个元素替换成集合 T 中的第 i 个元素。将假设集 $F_{S_{T, \sigma}}$ 记为 $F_{S, T}^\sigma$ 。

固定 $n \in \mathbf{N}$, 本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 关于两个本体样本集 $S = \{z_1^S, z_2^S, \dots, z_n^S\}$ 和 $T = \{z_1^T, z_2^T, \dots, z_n^T\}$ 的拉德马赫复杂度 (Rademacher complexity) 和经验拉德马赫复杂度 (empirical Rademacher complexity) 分别定义为

$$\Xi_n(F) = \frac{1}{n} E_{S, T \sim D^n, \sigma} \left[\sup_{f \in F_{S, T}^\sigma} \sum_{i=1}^n \sigma_i f(z_i^T) \right]$$

$$\hat{\Xi}_{S, T}(F) = \frac{1}{n} E_{\sigma} \left[\sup_{f \in F_{S, T}^\sigma} \sum_{i=1}^n \sigma_i f(z_i^T) \right]$$

2 算法理论分析

定理 1 设 $M: Z^n \times Z \rightarrow [0, 1]$ 为本体数据依赖函数且存在 LOO 一致稳定 γ , 则对于任意 Z 上的概率分布 D 和任意 $\delta \in (0, 1)$, 有

$$\mathbb{E}_{S \sim D^n} [(\Delta_{D-S}(M))^2] \leq 64\gamma^2 + \frac{2}{n} \quad (1)$$

$$P_{S \sim D^n} \left[\Delta_{D-S}(M) \geq 8 \sqrt{(4\gamma + \frac{1}{n}) \ln \frac{8}{\delta}} \right] \leq \delta \quad (2)$$

本文的第一个主要结果给出了本体数据依赖函数框架下的学习算法广义界以及广义界平方的上界, 其中式 (1) 说明本体稳定算法的误差界的平方的期望值可以被一致稳定参数所控制在某个固定的范围内, 而式 (2) 则说明误差界越大, 它发生的概率越小。

证明 考虑 m 个本体数据集, 每个数据集的容量均为 n 。为了防止混淆, 称其为多重本体数据集, 记为 S 。显然 $S \in \mathbb{Z}^{m \times n}$, 将每个本体子数据集分别记为 S_1, S_2, \dots, S_m , 且第 l 个本体子数据集的第 i 个元素记为 S_{li} 。设 $l \in \{1, 2, \dots, m\}$, 定义

$$f_l(S) = \Delta_{D-S_l}(M) = R_D[M(S_l)] - \hat{R}_{S_l}[M(S_l)] \quad (3)$$

设 S 和 S' 是两个多重本体数据集, 且 S' 与 S 相比仅仅是第 k 个本体数据子集的第 i 个元素不同。设 $S^{(k,i)}$ 为多重本体数据集, 它通过删除 S' 中第 k 个本体数据子集的第 i 个元素得到。如果 $l \neq k$, 则 $S_l = S'_l$, 进而有 $f_l(S) = f_l(S')$ 。若 $l = k$, 则

$$\begin{aligned} |R_D[M(S_l)] - R_D[M(S'_l)]| &= |R_D[M(S_l)] - R_D[M(S_l^{(k,i)})] + \\ &R_D[M(S_l^{(k,i)})] - R_D[M(S'_l)]| \leq |R_D[M(S_l)] - R_D[M(S_l^{(k,i)})]| + \\ &|R_D[M(S_l^{(k,i)})] - R_D[M(S'_l)]| = \left| \mathbb{E}_{z \sim D} [M(S_l, z) - M(S_l^{(k,i)}, z)] \right| \end{aligned}$$

$$\left| \mathbb{E}_{z \sim D} [M(S_l^{(k,i)}, z) - M(S'_l, z)] \right| \leq 2\gamma$$

$$|\hat{R}_{S_l}[M(S_l)] - \hat{R}_{S'_l}[M(S'_l)]| =$$

$$\left| \frac{1}{n} \sum_{j=1}^n M(S_l, S_{lj}) - \frac{1}{n} \sum_{j=1}^n M(S'_l, S'_{lj}) \right| \leq$$

$$\left| \frac{1}{n} \sum_{j=1, i \neq j}^n M(S_l, S_{lj}) - \frac{1}{n} \sum_{j=1, i \neq j}^n M(S'_l, S'_{lj}) \right| +$$

$$\left| \frac{1}{n} M(S_l, S_{li}) - \frac{1}{n} M(S'_l, S'_{li}) \right| \leq$$

$$\left| \frac{1}{n} \sum_{j=1, i \neq j}^n M(S_l, S_{lj}) - \frac{1}{n} \sum_{j=1, i \neq j}^n M(S_l^{(k,i)}, S_{lj}^{(k,i)}) \right| +$$

$$\left| \frac{1}{n} \sum_{j=1, i \neq j}^n M(S_l^{(k,i)}, S_{lj}^{(k,i)}) - \frac{1}{n} \sum_{j=1, i \neq j}^n M(S'_l, S'_{lj}) \right| +$$

$$\left| \frac{1}{n} M(S_l, S_{li}) - \frac{1}{n} M(S'_l, S'_{li}) \right| \leq$$

$$2\gamma + \frac{1}{n}$$

这说明 $|f_l(S) - f_l(S'_l)| \leq 4\gamma + \frac{1}{n}$ 。

对任意 $l \in \{1, 2, \dots, m\}$, 设本体数据依赖函数 $M_l: \mathbb{Z}^n \times \mathbb{Z} \rightarrow [0, 1]$ 拥有 LOO 一致稳定 γ , $A: \mathbb{Z}^{n \times m} \rightarrow \{1, 2, \dots, m\}$ 是 ε -差分隐私算法。

设 $X_S = \mathbb{E}_{S \sim D^{nm}, l=A(S)} [\hat{R}_{S_l}[M_l(S_l)]]$, $I(\cdot)$ 为真值函数

(论断为真时取值 1, 否则取值 0), 则

$$X_S = \mathbb{E}_{S \sim D^{nm}, l=A(S)} \left[\frac{1}{n} \sum_{i=1}^n M_l(S_l, S_{li}) \right] =$$

$$\mathbb{E}_{A, S \sim D^{nm}} \left[\frac{1}{n} \sum_{i=1}^n \sum_{l=1}^m I(A(S) = l) M_l(S_l, S_{li}) \right] =$$

$$\frac{1}{n} \sum_{i=1}^n \sum_{l=1}^m \mathbb{E}_{S \sim D^{nm}} [\mathbb{E}_A [I(A(S) = l) \cdot M_l(S_l, S_{li})]] \leq$$

$$\frac{1}{n} \sum_{i=1}^n \sum_{l=1}^m \mathbb{E}_{S \sim D^{nm}} [e^\varepsilon \cdot \mathbb{E}_A [I(A(S^{l(i),z}) = l) \times$$

$$(M_l(S_l^{i,z}, S_{li}) + \gamma)] =$$

$$\frac{1}{n} \sum_{i=1}^n \sum_{l=1}^m \mathbb{E}_{S \sim D^{nm}, z \sim D} [e^\varepsilon \cdot \mathbb{E}_A [I(A(S) = l) \cdot (M_l(S_l, z) + \gamma)] =$$

$$\mathbb{E}_{S \sim D^{nm}, z \sim D, l=A(S)} [e^\varepsilon \cdot (M_l(S_l, z) + \gamma)] =$$

$$e^\varepsilon (\mathbb{E}_{S \sim D^{nm}, z \sim D, l=A(S)} [M_l(S_l, z) + \gamma])$$

式中: $S^{l(i),z}$ 为多重本体数据集, 它通过将第 l 个本体数据子集的第 i 个元素替换成 z 得到; $S_l^{i,z}$ 是通过将 S_l 中第 i 个元素替换成 z 得到的本体数据子集。

由此可以得到

$$e^{-\varepsilon} X_S - \gamma \leq \mathbb{E}_{S \sim D^{nm}, l=A(S)} [R_D[M_l(S_l)]] \leq e^\varepsilon X_S + \gamma \quad (4)$$

式 (4) 的后半部分可以用类似的方法得到。

特别地, 有

$$\left| \mathbb{E}_{S \sim D^{nm}, l=A(S)} [R_D[M_l(S_l)] - \hat{R}_{S_l}[M_l(S_l)]] \right| \leq e^\varepsilon - 1 + \gamma \quad (5)$$

首先证明 (2) 成立。取 $m = \frac{\ln 2}{\delta}$, 设 f_1, f_2, \dots, f_m 为实值函数 (如式 (3) 定义), $f_{m+1}(S) = 0$ 。设 A 是 $f_1, f_2, \dots, f_m, f_{m+1}$ 上的执行算法满足对任意 $l \in \{1, 2, \dots, m\}$ 有 $|f_l(S) - f_l(S'_l)| \leq 4\gamma + \frac{1}{n}$ 。注意到, 对于 $l \in \{1, 2, \dots, m\}$ 有 $M_l = M$ 且 $M_{m+1} = 0$, 因此由式 (5) 可知:

$$\mathbb{E}_{S \sim D^{(m+1)n}} [\mathbb{E}_{l=A(S)} f_l(S)] =$$

$$\mathbb{E}_{a \sim D^{(m+1)n}, -A(a)} [R_D[M_l(Q)] - \hat{R}_{m_l}[M_l(S_l)]] \leq e^\varepsilon - 1 + \gamma$$

运用文献 [22] 中引理 7.1 可知:

$$\mathbb{E}_{S \sim D^{nm}} \left[\max \left\{ \max_{l \in \{1, 2, \dots, m\}} R_D[M_l(S_l)] - \hat{R}_{S_l}[M_l(S_l)], 0 \right\} \right] =$$

$$\mathbb{E}_{S \sim D^{nm}} \left[\max_{l \in \{0, m\}} f_l(S) \right] \leq$$

$$\frac{2 \left(4\gamma + \frac{1}{n} \right)}{\varepsilon} \ln(m+1) \leq$$

$$e^\varepsilon - 1 + \gamma + \frac{8\gamma + \frac{2}{n}}{\varepsilon} \ln(m+1)$$

$$\text{选取 } \varepsilon = \sqrt{\left(4\gamma + \frac{1}{n} \right) \ln(m+1)} = \sqrt{\left(4\gamma + \frac{1}{n} \right) \ln \left(\frac{e \ln 2}{\delta} \right)}$$

当 $\varepsilon \geq \frac{1}{2}$ 时式 (2) 显然成立。当 $\varepsilon < \frac{1}{2}$ 时有 $e^\varepsilon - 1 \leq 2\varepsilon$,

进而根据 $\gamma \leq \sqrt{\gamma}$ 有

$$4 \sqrt{\left(4\gamma + \frac{1}{n} \right) \ln \left(\frac{e \ln 2}{\delta} \right)} + \gamma \leq 4 \sqrt{\left(4\gamma + \frac{1}{n} \right) \ln \left(\frac{8}{\delta} \right)}$$

结合文献 [23] 中引理 1.2 可知:

$$P_{S \sim D^n} \left[R_D[M(S)] - \hat{R}_S[M(S)] \geq 8 \sqrt{\left(4\gamma_n + \frac{1}{n}\right) \ln \frac{8}{\delta}} \right] \leq \frac{\ln 2}{m} \leq \delta$$

从而式 (2) 成立。

式 (1) 证明: 设 $L(S, z) = M(S, z) - R_D[M(S)]$, 则易证 L 是关于分布 D 的无偏估计, 对任意 $S \in Z^n$ 都有 $R_D[L(S)] = 0$ 。由于 M 的取值范围在 $[0, 1]$ 内, 因此 L 取值范围是 $[-1, 1]$ 。由于

$$|R_D[M(S)] - R_D[M(S^V)]| \leq \left| \mathbb{E}_{z \sim D} [M(S, z) - M(S^V, z)] \right| \leq \gamma$$

L 拥有 LOO 一致稳定至多为 2γ 。又因为

$$\Delta_{D-S}(M) = R_D[M(S)] - \hat{R}_S[M(S)] =$$

$$\frac{1}{n} \sum_{i=1}^n (R_D[M(S)] - M(S, z_i)) = -\frac{1}{n} \sum_{i=1}^n L(S, z_i) = -\hat{R}_S[L(S)]$$

得到

$$\mathbb{E}_{S \sim D^n} [(\Delta_{D-S}(M))^2] = \mathbb{E}_{S \sim D^n} [(\hat{R}_S[L(S)])^2]$$

由此, 式 (1) 等价于证明: 设本体数据依赖函数 $L: Z^n \times Z \rightarrow [-1, 1]$ 拥有 LOO 一致稳定 2γ , D 是 Z 上的任意分布。如果 L 是关于 D 的无偏估计, 则有

$$\mathbb{E}_{S \sim D^n} [(\hat{R}_S[L(S)])^2] \leq 64\gamma^2 + \frac{2}{n} \quad (6)$$

记 $S^{i,z}$ 是将 S 中的第 i 个元素替换为 z , 并设

$$R_S^D[L(S)] = \mathbb{E}_{z \sim D} \left[\frac{1}{n} \sum_{i=1}^n L(S^{i,z}, z_i) \right], \text{ 有}$$

$$|\hat{R}_S[L(S)] - R_S^D[L(S)]| =$$

$$\left| \frac{1}{n} \sum_{i=1}^n L(S, z_i) - \mathbb{E}_{z \sim D} \left[\frac{1}{n} \sum_{i=1}^n L(S^{i,z}, z_i) \right] \right| \leq$$

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim D} [L(S, z_i) - L(S^{i,z}, z_i)] \right| =$$

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim D} [L(S, z_i) - L(S^V, z_i) + L(S^V, z_i) - L(S^{i,z}, z_i)] \right| \leq$$

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim D} [L(S, z_i) - L(S^V, z_i)] \right| +$$

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim D} [L(S^V, z_i) - L(S^{i,z}, z_i)] \right| \leq 2\gamma + 2\gamma = 4\gamma$$

$$\mathbb{E}_{S \sim D^n} [(R_S^D[L(S)])^2] \leq \mathbb{E}_{S \sim D^n, z \sim D} \left[\left(\frac{1}{n} \sum_{i=1}^n L(S^{i,z}, z_i) \right)^2 \right] =$$

$$\frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{S \sim D^n, z \sim D} [(L(S^{i,z}, z_i))^2] +$$

$$\frac{1}{n^2} \sum_{i,j \in \{1, 2, \dots, n\}, i \neq j} \mathbb{E}_{S \sim D^n, z \sim D} [L(S^{i,z}, z_i) L(S^{j,z}, z_j)] \leq$$

$$\frac{1}{n} + \frac{1}{n^2} \sum_{i,j \in \{1, 2, \dots, n\}, i \neq j} \mathbb{E}_{S \sim D^n, z \sim D} [L(S^{i,z}, z_i) L(S^{j,z}, z_j)]$$

记 S^{i,j,z_i,z_j} 是将 S 中的第 i 个元素和第 j 个元素分别替换为 z_i 和 z_j 而得到的新本体数据集, 并设

$$h(z) = \min_{z_i, z_j \in Z} L(S^{i,j,z_i,z_j}, z) + 4\gamma。$$

注意到:

$$\max_{z_i, z_j \in Z} L(S^{i,j,z_i,z_j}, z) \leq \min_{z_i, z_j \in Z} L(S^{i,j,z_i,z_j}, z) + 8\gamma$$

可知:

$$|L(S^{i,j,z_i,z_j}, z) - h(z)| \leq 4\gamma$$

因此:

$$\mathbb{E}_{S \sim D^n, z \sim D} [L(S^{i,z}, z_i) L(S^{j,z}, z_j)] =$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [L(S^{i,z}, z_i) L(S^{j,z}, z_j)] =$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [(L(S^{i,z}, z_i) - h(z_i))(L(S^{j,z}, z_j) - h(z_j))] +$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [h(z_i) L(S^{j,z}, z_j)] +$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [h(z_j) L(S^{i,z}, z_i)] - \mathbb{E}_{z_i, z_j \sim D} [h(z_i) h(z_j)] \leq (4\gamma)^2 +$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [h(z_i) L(S^{j,z}, z_j)] +$$

$$\mathbb{E}_{z_i, z_j, z \sim D} [h(z_j) L(S^{i,z}, z_i)] - \left(\mathbb{E}_{z' \sim D} [h(z')] \right)^2$$

考虑到 L 是无偏估计且函数 h 不依赖于 z_i 或 z_j , 因此对应每个给定的 z_i 和 z_j , 有

$$\mathbb{E}_{z_j \sim D} [h(z_i) L(S^{j,z}, z_j)] = h(z_i) R_D [L(S^{j,z}, z_j)] = 0$$

即

$$\mathbb{E}_{z_i, z_j, z \sim D} [h(z_i) L(S^{j,z}, z_j)] = \mathbb{E}_{z_i, z_j, z \sim D} [h(z_j) L(S^{i,z}, z_i)] = 0$$

故有

$$\mathbb{E}_{S \sim D^n} [(R_S^D[L(S)])^2] \leq$$

$$\frac{1}{n} + \frac{1}{n^2} \sum_{i,j \in \{1, 2, \dots, n\}, i \neq j} \mathbb{E}_{S \sim D^n, z \sim D} [L(S^{i,z}, z_i) L(S^{j,z}, z_j)] \leq$$

$$\frac{1}{n} + \frac{1}{n^2} \sum_{i,j \in \{1, 2, \dots, n\}, i \neq j} \left(16\gamma^2 - \left(\mathbb{E}_{z' \sim D} [h(z')] \right)^2 \right) \leq$$

$$\frac{1}{n} + \frac{1}{n^2} \sum_{i,j \in \{1, 2, \dots, n\}, i \neq j} 16\gamma^2 \leq \frac{1}{n} + 16\gamma^2$$

最后,

$$\mathbb{E}_{S \sim D^n} [(\hat{R}_S[L(S)])^2] =$$

$$\mathbb{E}_{S \sim D^n} [(R_S^D[L(S)] + \hat{R}_S[L(S)] - R_S^D[L(S)])^2] \leq$$

$$2\mathbb{E}_{S \sim D^n} [(R_S^D[L(S)])^2] +$$

$$2\mathbb{E}_{S \sim D^n} [(\hat{R}_S[L(S)] - R_S^D[L(S)])^2] \leq$$

$$2\left(\frac{1}{n} + 16\gamma^2\right) + 2(4\gamma)^2 = 64\gamma^2 + \frac{2}{n}$$

由此, 式 (1) 得证。

定义本体学习算法的 LOO 估计为 $\hat{R}_{\text{LOO}}[l(A(S))]$ =

$\frac{1}{n} \sum_{i=1}^n l(A(S^V), z_i)$ 。在监督本体学习下, $Z = V \times Y$, 每个本体样本点为 $z = (v, y)$, 可将 LOO 估计写成

$$\hat{R}_{\text{LOO}}[l(A(S))] = \frac{1}{n} \sum_{i=1}^n l(A_{S^V}(v_i), y_i), \text{ 同时期望误差可以}$$

写成 $R_D[l(A(S))] = E_{(v,y)}[l(A_S(v),y)]$ 。监督本体学习算法 A 关于本体亏损函数 l 存在 LOO 一致稳定 γ 可以重新写为 $l(A_S(v),y) - l(A_{S^v}(v),y) \leq \gamma$, 令 $g_i = g_i(z_1, z_2, \dots, z_n) = E(E_{z \sim D}[l(A_{S^v}(v),y)] - l(A_{S^v}(v_i),y_i))$ 。

本文第二个主要结论刻画了关于 LOO 估计的广义界和 $\left| \sum_{i=1}^n g_i \right|$ 之间的差值可以被 LOO 一致稳定 γ 所决定。

定理 2 设监督本体学习算法 A 关于本体亏损函数 l 存在 LOO 一致稳定 γ , 且本体亏损函数 l 有界: $l(\cdot, \cdot) \leq L$, 有

$$\left| R_D[l(A(S))] - \hat{R}_{\text{LOO}}[l(A(S))] - \sum_{i=1}^n g_i \right| \leq n\gamma$$

证明 通过直接计算, 可得

$$\begin{aligned} & \left| n(R_D[l(A(S))] - \hat{R}_{\text{LOO}}[l(A(S))]) \right| = \\ & \left| \sum_{i=1}^n (E_{(v,y)}[l(A_S(v),y)] - l(A_{S^v}(v_i),y_i)) \right| \leq \\ & n\gamma + \left| \sum_{i=1}^n (E_{(v,y)}[l(A_{S^v}(v),y)] - l(A_{S^v}(v_i),y_i)) \right| \leq \\ & n\gamma + \left| \sum_{i=1}^n g_i \right| \end{aligned}$$

用类似的方法可以得到:

$$\left| n(R_D[l(A(S))] - \hat{R}_{\text{LOO}}[l(A(S))]) \right| \geq \left| \sum_{i=1}^n g_i \right| - n\gamma$$

因此, 定理 2 得证。

定理 3 设本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 的直径和均值直径分别为 Δ 和 $\bar{\Delta}$, 且有 LOO 一致稳定 β , 则 F 有 LOO-CV 稳定 $\Delta + \beta$ 和 LOO-均值 CV 稳定 $\bar{\Delta} + \beta$ 。

证明 设 $S \in Z^n, z \in S$ 。对任意 $f \in F_S, f' \in F_{S^v}$, 由 LOO 一致稳定条件, 存在 $f'' \in F_S$ 使得 $l(f', z) - l(f'', z) \leq \beta$ 。从而

$$\begin{aligned} l(f', z) - l(f, z) &= l(f', z) - l(f'', z) + l(f'', z) - l(f, z) \leq \\ & \beta + \sup_{f, f'' \in F_S} (l(f'', z) - l(f, z)) \end{aligned}$$

进而

$$\sup_{f \in F_S, f' \in F_{S^v}} (l(f', z) - l(f, z)) \leq \beta + \sup_{f, f'' \in F_S} (l(f'', z) - l(f, z))$$

这说明定理 3 得证。

设 φ_S 是与 F_S 关联的本体亏损函数族:

$$\varphi_S = \{z \mapsto l(f, z) | f \in F_S\} \quad (7)$$

设 $\Psi = (\varphi_S)_{S \in Z^n}$ 为 φ_S 的假设集族。为了方便定理 4 讨论, 假设与 F_S 关联的本体亏损函数是 1-有界的, 即对于任意 $f \in F_S$, 任意 $z \sim D$ 有 $0 \leq l(f, z) \leq 1$ 。定理 4 是用本体数据依赖假设集族 LOO 一致稳定和 $\Psi = (\varphi_S)_{S \in Z^n}$ 来刻画拉德马赫复杂度和经验拉德马赫复杂度之间的差值。

定理 4 设 $F = (F_S)_{S \in Z^n}$ 为本体数据依赖假设集族, 且存在 LOO 一致稳定 β 。设 $\Psi = (\varphi_S)_{S \in Z^n}$ 如式 (7) 所定义, 则对任意 $\delta > 0$, 式 (8) 在所有本体样本 $S, T \sim Z^n$ 中以概率至少 $1 - \delta$ 成立:

$$|\hat{\Xi}_{S,T}(\Psi) - \Xi_n(\Psi)| \leq \sqrt{\frac{\left((1 + 2n\beta)^2 + 4n^2\beta^2 \right) \log \frac{2}{\delta}}{2n}} \quad (8)$$

证明 设本体样本 T' 从 T 中得到且和 T 只有一个本体样本点不同, 不妨记第 k 个元素不同, 其中 $k \in \{1, 2, \dots, n\}$ 。给定 $\eta > 0$, 对于任意拉德马赫向量 σ , 根据上确界的定义, 存在 $f' \in F_{S,T}^\sigma$ 使得

$$\sum_{i=1}^n \sigma_i l(f', z_i^T) \geq \sup_{f \in F_{S,T}^\sigma} \sum_{i=1}^n \sigma_i l(f, z_i^T) - \eta$$

根据 F 存在 LOO 一致稳定 β 的假设, 存在 $f \in F_{S,T}^\sigma$, 使得对任意 $z \in Z$, 有

$$\begin{aligned} & |l(f', z) - l(f, z)| = \\ & |l(f', z) - l(f'', z) + l(f'', z) - l(f, z)| \leq \\ & |l(f', z) - l(f'', z)| + |l(f'', z) - l(f, z)| \leq 2\beta \end{aligned}$$

其中 f'' 来自将 $S_{T,\sigma}$ 中第 k 个元素删除后得到的新本体样本集对应的假设集。从而有

$$\begin{aligned} \sup_{f \in F_{S,T}^\sigma} \sum_{i=1}^n \sigma_i l(f, z_i^T) &\leq \sum_{i=1}^n \sigma_i l(f', z_i^T) + \eta \leq \\ & \sum_{i=1}^n (\sigma_i (l(f, z_i^T) + 2\beta)) + \eta \end{aligned}$$

由于 $\eta > 0$ 是任意的, 得到

$$\begin{aligned} \frac{1}{n} \sup_{f \in F_{S,T}^\sigma} \sum_{i=1}^n \sigma_i l(f, z_i^T) &\leq \frac{1}{n} \sum_{i=1}^n (\sigma_i (l(f, z_i^T) + 2\beta)) \leq \\ & \frac{1}{n} \sup_{f \in F_{S,T}^\sigma} \sum_{i=1}^n \sigma_i l(f, z_i^T) + 2\beta + \frac{1}{n} \end{aligned}$$

这意味着, 将 T 替换成 T' , 对 $\hat{\Xi}_{S,T}(\Psi)$ 的影响最多为 $2\beta + \frac{1}{n}$ 。用相同的方法可以得到, 对 S 改变一个本体样本点对 $\hat{\Xi}_{S,T}(\Psi)$ 的影响最多为 2β 。最后, 利用 McDiarmid 不等式^[24] 得到定理 4 的结论。

定理 5 用 $\Xi_n(\Psi)$ 和 $\bar{\chi}$ 来刻画本体数据依赖假设集族 LOO 一致稳定本体学习算法的广义界。

定理 5 设 $F = (F_S)_{S \in Z^n}$ 为本体数据依赖假设集族, 存在 LOO 一致稳定 β 和 LOO-均值 CV 稳定 $\bar{\chi}$ 。设 $\Psi = (\varphi_S)_{S \in Z^n}$ 如式 (7) 所定义, 则对任意 $\delta > 0$, 所有 $f \in F_S$, 式 (9) 在所有本体样本 $S \sim Z^n$ 中以概率至少 $1 - \delta$ 成立:

$$R_D(f) - \hat{R}_S(f) \leq \min\{2\bar{\chi}, 2\Xi_n(\Psi)\} + (1 + 4n\beta) \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (9)$$

这里, $R_D(f) = E_{z \sim D}[l(f, z)]$, $\hat{R}_S(f) = E_{z \sim S}[l(f, z)] = \frac{1}{n} \sum_{i=1}^n l(f, z_i)$ 。

证明 对任意两个本体样本集 S 和 S' (只有对应第 k 个元素不同, 分别记为 z 和 z' , 其他均相同), 令

$$\Theta(S, S') = \sup_{f \in F_S} [R_D(f) - \hat{R}_{S'}(f)]$$

利用简单计算以及本体亏损函数界为 1 的假设, 可以得到:

$$\begin{aligned} & \Theta(S, S) - \Theta(S', S') = \\ & [\Theta(S, S) - \Theta(S, S')] + [\Theta(S, S') - \Theta(S', S')] \\ & \Theta(S, S) - \Theta(S, S') \leq \\ & \sup_{f \in F_S} [R_D(f) - \hat{R}_S(f)] - [R_D(f) - \hat{R}_{S'}(f)] \leq \\ & \sup_{f \in F_S} \frac{1}{n} \sum_{i=1}^n (l(f, z) - l(f, z')) \leq \frac{1}{n} \\ & \Theta(S, S') - \Theta(S', S') = \sup_{f \in F_{S'}} [R_D(f) - \hat{R}_{S'}(f)] - \\ & \sup_{f \in F_{S'}} [R_D(f) - \hat{R}_{S'}(f)] \end{aligned}$$

根据上确界的定义可知, 对任意 $\eta > 0$, 存在 $f \in F_S$, 使得 $\sup_{f \in F_S} [R_D(f) - \hat{R}_{S'}(f)] - \eta \leq [R_D(f) - \hat{R}_{S'}(f)]$ 。

由 $F = (F_S)_{S \in \mathcal{Z}^n}$ 的 LOO 一致稳定 β , 通过类似定理 4 中的讨论, 可知存在 $f' \in F_{S'}$ 使得对任意 z , 有 $l(f', z) - l(f, z) \leq 2\beta$ 。进而

$$\begin{aligned} & \Theta(S, S') - \Theta(S', S') \leq \\ & [R_D(f) - \hat{R}_{S'}(f)] + \eta - \sup_{f \in F_{S'}} [R_D(f) - \hat{R}_{S'}(f)] \leq \\ & [R_D(f) - \hat{R}_{S'}(f)] + \eta - [R_D(f') - \hat{R}_{S'}(f')] = \\ & [R_D(f) - R_D(f')] + \eta + [\hat{R}_{S'}(f') - \hat{R}_{S'}(f)] \leq 4\beta + \eta \end{aligned} \quad (10)$$

由于 $\eta > 0$ 是任意的, 式 (10) 意味着 $\Theta(S, S') - \Theta(S', S') \leq 4\beta$ 。从而有 $\Theta(S, S) - \Theta(S', S') \leq 4\beta + \frac{1}{n}$ 。

通过文献 [25] 中的 McDiarmid 不等式可知, 对任意 $\delta > 0$, 式 (11) 以概率至少 $1 - \delta$ 成立。

$$\Theta(S, S) \leq \mathbb{E}_{S \sim D^n} [\Theta(S, S)] + (1 + 4n\beta) \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (11)$$

下面计算关于拉德马赫复杂度 $\Xi_n(\Psi)$ 和式 (11) 中 $\mathbb{E}_{S \sim D^n} [\Theta(S, S)]$ 项的关系:

$$\begin{aligned} \mathbb{E}_{S \sim D^n} [\Theta(S, S)] &= \mathbb{E}_{S \sim D^n} \left[\sup_{f \in F_S} [R_D(f) - \hat{R}_S(f)] \right] = \\ & \mathbb{E}_{S \sim D^n} \left[\sup_{f \in F_S} \left[\mathbb{E}_{T \sim D^n} [\hat{R}_T(f)] - \hat{R}_S(f) \right] \right] \leq \\ & \mathbb{E}_{S, T \sim D^n} \left[\sup_{f \in F_S} [\hat{R}_T(f) - \hat{R}_S(f)] \right] = \\ & \mathbb{E}_{S, T \sim D^n} \left[\sup_{f \in F_S} \frac{1}{n} \sum_{i=1}^n [l(f, z_i^T) - l(f, z_i^S)] \right] = \\ & \mathbb{E}_{S, T \sim D^n} \left[\mathbb{E}_{\sigma} \left[\sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n \sigma_i [l(f, z_i^T) - l(f, z_i^S)] \right] \right] \leq \\ & \mathbb{E}_{S, T \sim D^n, \sigma} \left[\sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n \sigma_i l(f, z_i^T) + \sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n -\sigma_i l(f, z_i^S) \right] = \end{aligned}$$

$$\begin{aligned} & \mathbb{E}_{S, T \sim D^n, \sigma} \left[\sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n \sigma_i l(f, z_i^T) + \sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n -\sigma_i l(f, z_i^S) \right] = \\ & \mathbb{E}_{S, T \sim D^n, \sigma} \left[\sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n \sigma_i l(f, z_i^T) + \sup_{f \in F_{S, T}^{\sigma}} \frac{1}{n} \sum_{i=1}^n \sigma_i l(f, z_i^S) \right] = \\ & 2\Xi_n(\Psi) \end{aligned}$$

另一方面, 固定 $\eta > 0$, 根据上确界的定义, 对任意 $S \in \mathcal{Z}^n$, 存在 f_S 使得

$$\sup_{f \in F_S} [R_D(f) - \hat{R}_S(f)] - \eta \leq R_D(f_S) - \hat{R}_S(f_S)$$

根据 $R_D(f_S)$ 的定义, 有

$$\mathbb{E}_{S \sim D^n} [R_D(f_S)] = \mathbb{E}_{S \sim D^n} [\mathbb{E}_{z \sim D} [l(f_S, z)]] = \mathbb{E}_{S \sim D^n, z \sim D} [l(f_S, z)]$$

另外, 可知

$$\mathbb{E}_{S \sim D^n} [\hat{R}_S(f_S)] = \mathbb{E}_{S \sim D^n, z \sim S} [l(f_S, z)] = \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{z \leftrightarrow z'}}, z')]$$

其中 $S^{z \leftrightarrow z'}$ 表示将本体数据集 S 中的 z 和 $z' \sim D$ 相交换, 进而有

$$\begin{aligned} \mathbb{E}_{S \sim D^n} [\Theta(S, S)] &\leq \mathbb{E}_{S \sim D^n} [R_D(f_S) - \hat{R}_S(f_S)] + \eta = \\ & \mathbb{E}_{S \sim D^n, z' \sim D} [l(f_S, z')] - \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{z \leftrightarrow z'}}, z')] + \eta = \\ & \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_S, z') - l(f_{S^{z \leftrightarrow z'}}, z')] + \eta = \\ & \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{z \leftrightarrow z'}}, z) - l(f_S, z)] + \eta = \\ & \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{z \leftrightarrow z'}}, z) - l(f_{S^{\setminus z}}, z) + l(f_{S^{\setminus z}}, z) - l(f_S, z)] + \eta \leq \\ & \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{z \leftrightarrow z'}}, z) - l(f_{S^{\setminus z}}, z)] + \\ & \mathbb{E}_{S \sim D^n, z' \sim D, z \sim S} [l(f_{S^{\setminus z}}, z) - l(f_S, z)] + \eta \leq 2\bar{\chi} + \eta \end{aligned} \quad (12)$$

其中, $S^{\setminus z}$ 表示在本体数据集 S 中删除元素 z 得到的集合。由于式 (12) 对所有 $\eta > 0$ 成立, 从而有

$$\mathbb{E}_{S \sim D^n} [\Theta(S, S)] \leq 2\bar{\chi}。$$

结合 $\mathbb{E}_{S \sim D^n} [\Theta(S, S)] \leq 2\bar{\chi}$, $\mathbb{E}_{S \sim D^n} [\Theta(S, S)] \leq 2\Xi_n(\Psi)$ 和式 (11), 可知定理 5 成立。

沿用定理 1 中的标记, $S \in \mathcal{Z}^{m \times n}$ 为多重本体数据集, 它由 m 个本体数据集 S_1, S_2, \dots, S_m 构成, 每个数据集的容量均为 n 。第 l 个本体子数据集的第 i 个元素记为 S_{li} , 其中 $l \in \{1, 2, \dots, m\}$ 。定理 6 刻画了多重本体数据集框架下 LOO-CV 稳定 χ 的作用。

定理 6 设 $F = (F_S)_{S \in \mathcal{Z}^n}$ 为本体数据依赖假设集族, 存在 LOO-CV 稳定 χ , 则

$$\sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}, z' \sim D, z \sim S_l} [P[A(S) = \perp] [l(f_{S_l^{z \leftrightarrow z'}}, z) - l(f_{S_l}, z)]] \leq 2\chi,$$

其中 $S_l^{z \leftrightarrow z'}$ 将本体数据集 S_l 中的 z 和 $z' \sim D$ 相交换。

证明 通过推导可知

$$\begin{aligned} & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}, z' \sim D, z \sim S_l} [P[A(S) = \perp] [l(f_{S_l^{z \leftrightarrow z'}}, z) - l(f_{S_l}, z)]] \leq \\ & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}, z' \sim D, z \sim S_l} [P[A(S) = \perp] \times \end{aligned}$$

$$\begin{aligned}
 & \sup_{f \in F_{S_l}, f' \in F_{S_l^z}} [l(f', z) - l(f, z)] \leq \\
 & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} \left[P[A(S) = l] \times \right. \\
 & \mathbb{E}_{z' \sim D, z \sim S_l} \left[\sup_{f \in F_{S_l}, f' \in F_{S_l^z}} [l(f', z) - l(f, z)] | S \right] \leq \\
 & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} [P[A(S) = l] \times \\
 & \mathbb{E}_{z' \sim D, z \sim S_l} \left[\sup_{f \in F_{S_l}, f' \in F_{S_l^z}, f'' \in F_{S_l^z}} [l(f', z) - l(f'', z) + \right. \\
 & \left. l(f'', z) - l(f, z)] | S \right] \leq \\
 & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} \left[P[A(S) = l] \times \right. \\
 & \mathbb{E}_{z' \sim D, z \sim S_l} \left[\sup_{f' \in F_{S_l^z}, f'' \in F_{S_l^z}} [l(f', z) - l(f'', z)] | S \right] \leq \\
 & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} \left[P[A(S) = l] \times \right. \\
 & \mathbb{E}_{z' \sim D, z \sim S_l} \left[\sup_{f' \in F_{S_l^z}, f'' \in F_{S_l^z}} [l(f', z) - l(f'', z)] | S \right] + \\
 & \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} [P[A(S) = l] \times \\
 & \mathbb{E}_{z' \sim D, z \sim S_l} \left[\sup_{f \in F_{S_l}, f'' \in F_{S_l^z}} [l(f'', z) - l(f, z)] | S \right] \leq \\
 & 2 \sum_{l=1}^m \mathbb{E}_{S \sim D^{mn}} [P[A(S) = l] \chi] = \\
 & 2 \mathbb{E}_{S \sim D^{mn}} \left[\sum_{l=1}^m P[A(S) = l] \right] \chi = 2\chi
 \end{aligned}$$

其中 S_l^z 表示在本地数据集 S_l 中删除元素 z 得到的集合。

利用定理 1 中的本体多重集方法, 结合定理 5 和定理 6, 可得关于用 $\Xi_n(\Psi)$ 和 χ 来刻画本体数据依赖假设集族 LOO 一致稳定本体学习算法的广义界。

定理 7 设 $F = (F_S)_{S \in Z^n}$ 为本体数据依赖假设集族, 存在 LOO 一致稳定 β 和 LOO-CV 稳定 χ 。设 $\Psi = (\varphi_S)_{S \in Z^n}$ 如式 (7) 所定义, 则对任意 $\delta \in (0, 1)$, 所有 $f \in F_S$, 式 (13) 在所有本体样本 $S \sim Z^n$ 中以概率至少 $1 - \delta$ 成立:

$$\begin{aligned}
 & R_D(f) - \hat{R}_S(f) \leq \\
 & \min \left\{ 2\Xi_n(\Psi) + (1 + 4n\beta) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}, \right. \\
 & \left. 2\chi \sqrt{e} + 4 \sqrt{\left(4\beta + \frac{1}{n}\right) \log \frac{6}{\delta}} \right\} \quad (13)
 \end{aligned}$$

由于证明方法与之前定理类似, 我们不再给

出具体证明。同时, 通过替换证明过程中的一些技巧, 可以得到定理 7 的另外一个版本如定理 8。

定理 8 设 $F = (F_S)_{S \in Z^n}$ 为本体数据依赖假设集族, 存在 LOO 一致稳定 β 和 LOO-CV 稳定 χ 。设 $\Psi = (\varphi_S)_{S \in Z^n}$ 如式 (7) 所定义, 则对任意 $\delta \in (0, 1)$, 所有 $f \in F_S$, 式 (14) 在所有本体样本 $S \sim Z^n$ 中以概率至少 $1 - \delta$ 成立:

$$\begin{aligned}
 & R_D(f) - \hat{R}_S(f) \leq \\
 & \min \left\{ 2\Xi_n(\Psi) + 8(1 + 4n\beta) \sqrt{\frac{\log \frac{3}{\delta}}{n}}, \right. \\
 & \left. 2\chi \sqrt{e} + 4 \sqrt{\left(4\beta + \frac{1}{n}\right) \log \frac{3}{\delta}} \right\} \quad (14)
 \end{aligned}$$

3 结束语

本文主要从删除一个本体样本的设定出发, 讨论具有 LOO 一致稳定性的本体学习算法的广义界的统计学特征。主要从两个角度来讨论:

- 1) 本体学习算法关于本体亏损函数具有 LOO 一致稳定性;
- 2) 本体函数假设空间稳定, 即本体数据依赖假设集族 $F = (F_S)_{S \in Z^n}$ 存在 LOO 一致稳定 β 。

利用统计学习理论的方法, 分别给出了两种假设框架下, 本体学习算法的学习率的上界。得到的理论结果表明, 在算法一致稳定和假设空间一致稳定的理论假设下, 本体学习算法的误差界的上界可以被稳定性参数很好地控制, 进而保证了本体算法的收敛性。鉴于各种形式的本体学习算法已被广泛应用于基因学、营养学、化学、地理信息系统等领域, 本文得到的理论结果对本体学习算法的各类实际应用有着潜在的指导意义和参考价值。关于稳定性在具体某个本体学习算法中的表现, 在未来的本体算法执行中还有待进一步实验验证。

参考文献:

- [1] SKALLE P, AAMODT A. Petrol 18 946: downhole failures revealed through ontology engineering[J]. *Journal of petroleum science and engineering*, 2020, 191: 107188.
- [2] SOBRAL T, GALVÃO T, BORGES J. An ontology-based approach to knowledge-assisted integration and visualization of urban mobility data[J]. *Expert systems with applications*, 2020, 150: 113260.
- [3] AL-SAYED M M, HASSAN H A, OMARA F A. Cloud-FNF: an ontology structure for functional and non-functional features of cloud services[J]. *Journal of parallel and distributed computing*, 2020, 141: 143–173.
- [4] TEBES G, PEPPINO D, BECKER P, et al. Analyzing and

- documenting the systematic review results of software testing ontologies[J]. *Information and software technology*, 2020, 123: 106298.
- [5] PRADEEP D, SUNDAR C. QAOC: novel query analysis and ontology-based clustering for data management in Hadoop[J]. *Future generation computer systems*, 2020, 108: 849–860.
- [6] HEMA A M, KUPPUSAMY K. A novel trust-based privacy preservation framework for service handling via ontology service ranking[J]. *Wireless personal communications*, 2020, 112(3): 1339–1354.
- [7] MESSAOUDI R, MTIBAA A, VACAVANT A, et al. Ontologies for liver diseases representation: a systematic literature review[J]. *Journal of digital imaging*, 2020, 33(3): 563–573.
- [8] MANTOVANI A, PIANA F, LOMBARDO V. Ontology-driven representation of knowledge for geological maps [J] *Computers and geosciences*, 2020, 139: 104446.
- [9] ABEYSINGHE R, HINDERER III E W, MOSELEY H N B, et al. SSIF: subsumption-based sub-term inference framework to audit gene ontology[J]. *Bioinformatics*, 2020, 36(10): 3207–3214.
- [10] KOSSMANN M, SAMHAN A, ODEH M, et al. Extending the scope of configuration management for the development and life cycle support of systems of systems—an ontology-driven framework applied to the enceladus submarine exploration lander[J]. *Systems engineering*, 2020, 23(3): 366–391.
- [11] ZHU Linli, HUA Gang, ASLAM A. Ontology learning algorithm using weak functions[J]. *Open physics*, 2018, 16(1): 910–916.
- [12] GAO Wei, ZHANG Yunqing, GUIRAO J L G, et al. A discrete dynamics approach to sparse calculation and applied in ontology science[J]. *Journal of difference equations and applications*, 2019, 25(9/10): 1239–1254.
- [13] ZHU Linli, HUA Gang, ZAFAR S, et al. Fundamental ideas and mathematical basis of ontology learning algorithm[J]. *Journal of intelligent and fuzzy systems*, 2018, 35(4): 4503–4516.
- [14] GAO Wei, GUIRAO J L G, BASAVANAGOUD B, et al. Partial multi-dividing ontology learning algorithm[J]. *Information sciences*, 2018, 467: 35–58.
- [15] 朱林立, 华钢, 高伟. 决定图框架下本体学习算法的稳定性分析 [J]. *计算机科学*, 2020, 47(5): 43–50.
ZHU Linli, HUA Gang, GAO Wei. Stability analysis of ontology learning algorithm in decision graph setting[J]. *Computer science*, 2020, 47(5): 43–50.
- [16] ZHU Linli, YU Pan, FARAHANI M R, et al. Magnitude preserving based ontology regularization algorithm[J]. *Journal of intelligent and fuzzy systems*, 2017, 33(5): 3113–3122.
- [17] ZHU Linli, YU Pan, JAMIL M K, et al. Boosting based ontology sparse vector computation approach[J]. *Engineering letters*, 2017, 25(4): 406–415.
- [18] ZHU Linli, YU Pan, FARAHANI M R, et al. Theoretical characteristics on scoring function in multi-dividing setting[J]. *IAENG international journal of applied mathematics*, 2017, 47(1): 28–36.
- [19] ZHU Linli, TAO Weige, MIN Xiaozhong, et al. Theoretical characteristics of ontology learning algorithm in multi-dividing setting[J]. *IAENG international journal of computer science*, 2016, 43(2): 184–191.
- [20] GAO Wei, XU Tianwei. Stability analysis of learning algorithms for ontology similarity computation[J]. *Abstract and applied analysis*, 2013, 2013: 174802.
- [21] GAO W, ZHANG Y G, XU T W, et al. Analysis for learning a similarity function with ontology applications[J]. *Journal of information and computational science*, 2012, 17(9): 5311–5318.
- [22] BASSILY R, NISSIM K, SMITH A, et al. Algorithmic stability for adaptive data analysis[C]//*Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*. Cambridge: ACM, 2016: 1046–1059.
- [23] STEINKE T, ULLMAN J. Subgaussian tail bounds via stability arguments[EB/OL](2017–04–21)[2021–01–09] <https://arxiv.org/abs/1701.03493v2>.
- [24] MCDIARMID C. On the method of bounded differences[M]//SIEMONS J. *Surveys in Combinatorics*, 1989: Invited Papers at the Twelfth British Combinatorial Conference. Cambridge: Cambridge University Press, 1989: 148–188.
- [25] WALLSTROM T C. On the application of McDiarmid's inequality to complex systems[J]. *SIAM-ASA journal on uncertainty quantification*, 2017, 5(1): 240–245.

作者简介:



朱林立, 高级工程师, 博士, 主要研究方向为人工智能、机器学习。



华钢, 教授, 博士, 主要研究方向为物联网、矿山监控与监管、智能信息处理。



高伟, 教授, 博士, 主要研究方向为图论、人工智能和统计学习理论。