



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS

大规模半监督的核方法模型选择研究

博士学位论文答辩

答 辩 人：李 健
导 师：王 伟 平 研 究 员
指导老师：刘 勇 副研究员

中国科学院信息工程研究所

2020年 5月 14日

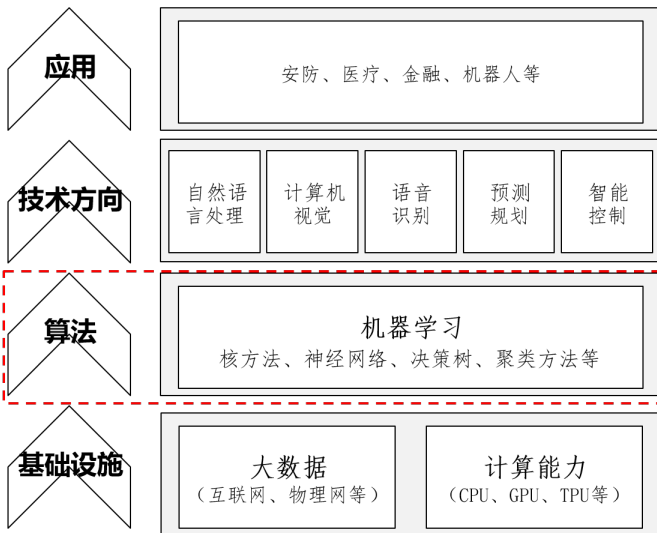
目录

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

目录

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

人工智能技术框架

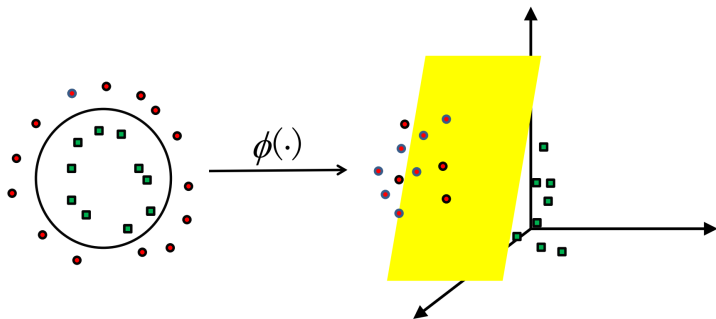


研究背景和意义：核函数

核函数：隐式特征映射的内积形式

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

将数据从低维输入空间隐式地映射到高维特征空间

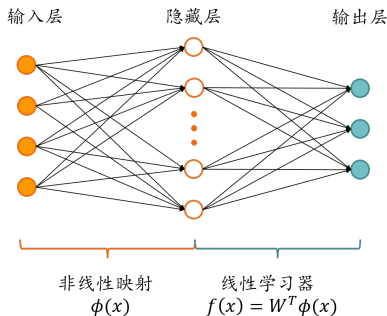


核函数（特征映射） \Rightarrow 特征表达能力（学习性能）

研究背景和意义：核方法

核方法: 基于核函数的方法

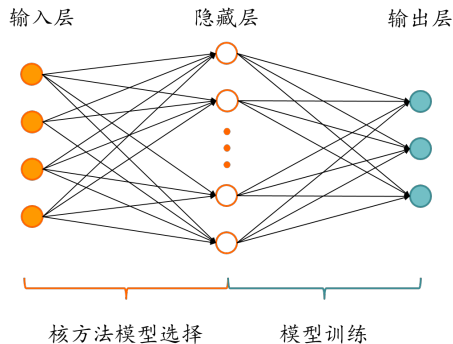
- 非线性映射（核函数） + 线性学习器 \Rightarrow 解决非线性问题



- 学习性能直接由非线性映射（核函数）决定
- 具有坚实的理论基础和完备的非线性学习框架
- 得到广泛研究与应用: 支持向量机 (SVM) [Cortes and Vapnik 1995]、高斯过程 (Gaussian process) [Rasmussen 2003]、核岭回归 (KRR) [Saunders et al. 1998]、核主成分分析(KPCA) [Schölkopf et al. 1998]等

研究背景和意义：核方法模型选择

核方法模型选择：为核方法选择核函数



核方法模型选择 \Rightarrow 核函数（特征映射） \Rightarrow 学习性能

- 核方法模型选择决定了核方法的学习性能
- 是核方法理论研究和实际应用的瓶颈与关键

研究背景和意义：大规模半监督的核方法模型选择

- 当今训练数据的特点

- 数据产生、收集、存储技术的飞速发展 \Rightarrow 大量无标签数据
- 数据标注代价高昂 \Rightarrow 少量有标签数据

- 传统核方法模型选择处理**大规模、半监督数据**时，存在如下不足：

- **缺乏泛化理论保证**：传统核方法模型选择仅使用了有标签数据，摒弃了无标签数据中包含的有用信息，泛化性能较差。
- **缺乏简洁的模型选择准则**：传统核方法模型选择方法大多将核选择、模型训练分开学习，过程繁琐复杂。
- **存在计算瓶颈**：空间复杂度为 $\mathcal{O}(n^2)$ 、时间复杂度至少为 $\mathcal{O}(n^2)$ 。

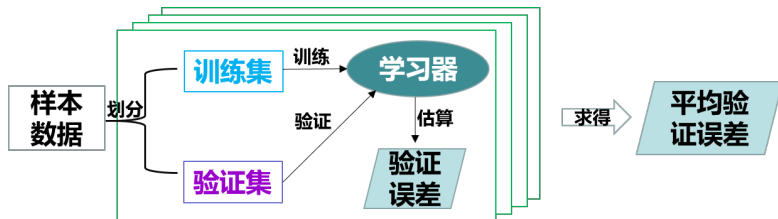
- **Motivation**：当前核方法模型选择方法不适用于大规模、半监督数据。亟需为大规模半监督的核方法模型选择

- 建立大规模半监督核方法的 **近似泛化理论**
- 制定简洁、易求解的 **模型选择准则**
- 设计适用于大规模半监督核方法的 **高效求解算法**

目录

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

研究现状与挑战：交叉验证



常用方法：

- k 折交叉验证 [Stone 1974]
- 自助法 [Efron and Tibshirani 1994]
- 留出法 [Chapelle et al. 2002]

不足之处：

- 需要预先指定候选核函数, 过于依赖人工经验, 缺乏理论保证
- 计算效率低下, 对于每个候选核函数的时间复杂度 $\mathcal{O}(kn^3)$
- 不适用于大规模、半监督数据

研究现状与挑战：近似交叉验证

交叉验证需要在不同划分训练多次，而近似交叉验证只需训练学习器一次，提升了交叉验证进行模型选择的效率。



常用方法：

- 张成界 (span bound) [Chapelle et al. 2002]
- 影响函数 (Influence function) [Debruyne et al. 2008]
- Bouligand 影响函数 [Liu et al. 2014; 2019]

不足之处：

- 需要预先指定候选核函数，过于依赖人工经验，缺乏理论保证
- 计算效率低下，对于每个候选核函数的时间复杂度 $\mathcal{O}(n^3)$
- 不适用于大规模、半监督数据

研究现状与挑战：核对齐方法

核对齐基于“相似样本特征对应的标签也是相似的”思想, 将最大化核矩阵与标签的对齐值作为模型选择准则.

$$\arg \max_{\kappa \in \mathcal{K}} \mathbf{y}^T \mathbf{K} \mathbf{y}$$

主要方法:

- 核对齐 (KTA) [Cristianini et al. 2001, Shen et al. 2016]
- 中心化核对齐 (centered KTA) 准则 [Cortes et al. 2012, Pérez-Ortiz et al. 2016]
- 核极化 (kernel polarization) [Baram 2005]

不足之处:

- 需要预先指定候选核函数, 过于依赖人工经验, 缺乏理论保证
- 独立于具体学习算法, 因此理论上无法保证学习算法的泛化性能
- 不适用于大规模、半监督数据

研究现状与挑战：多核学习

多核学习 (multiple kernel learning, MKL) 的核函数是多个基核的凸组合

$$\kappa(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^m \mu_i \kappa_i(\mathbf{x}, \mathbf{x}'), \mu_i \geq 0, \sum_{i=1}^m \mu_i = 1,$$

其中, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$ 为多核组合系数.

多核学习根据优化目标, 以端到端形式同时学习多个核函数的凸组合系数 (核函数)、模型参数

主要方法:

- 高效求解算法 [Bach et al. 2004, Gönen and Alpaydm 2011]
- ℓ -p 范数自动核选择方法 [Kloft et al. 2011, Han et al. 2017]
- 融入局部 Rademacher 复杂度 [Cortes et al. 2013, Li et al. 2018]
- 无限核学习 [Gehler and Nowozin 2008, Liu et al. 2017]

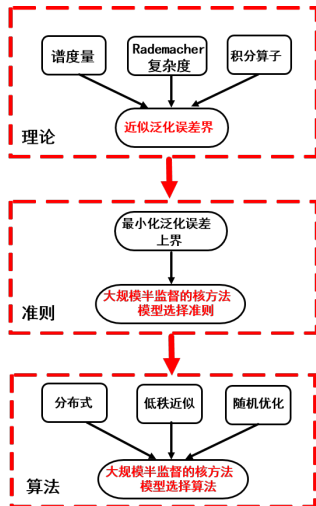
不足之处:

- 求解多核组合问题计算复杂, 求解效率低下
- 不适用于大规模、半监督数据

目录

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

主要研究内容



◆ 大规模半监督的核方法模型选择**理论**研究

- 基于**谱度量**的二分类泛化误差界
- 基于**Rademacher复杂度**的多输出泛化误差界
- 基于**积分算子理论**的最优泛化误差界

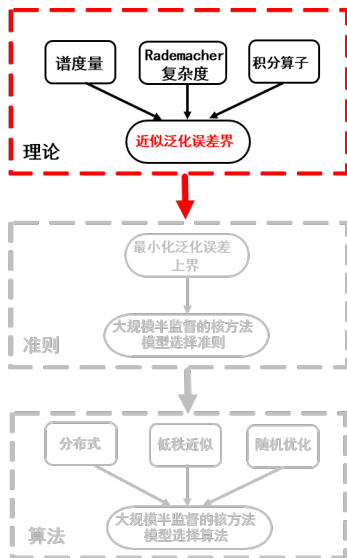
◆ 大规模半监督的核方法模型选择**准则**研究

- 最大化谱度量
- 最小化核矩阵尾部特征值之和
- 反向传播更新核超参数

◆ 大规模半监督的核方法模型选择**算法**研究

- 基于**分布式**核方法的大规模算法
- 基于核方法**低秩近似**的大规模算法
- 基于一阶梯度**随机优化**的大规模算法

大规模半监督的核方法模型选择准则研究



◆ 大规模半监督的核方法模型选择理论研究

- 基于**谱度量**的二分类泛化误差界
- 基于**Rademacher复杂度**的多输出泛化误差界
- 基于**积分算子理论**的最优泛化误差界

◆ 大规模半监督的核方法模型选择准则研究

- 最大化谱度量
- 最小化核矩阵尾部特征值之和
- 反向传播更新核超参数

◆ 大规模半监督的核方法模型选择算法研究

- 基于**分布式核方法**的大规模算法
- 基于核方法**低秩近似**的大规模算法
- 基于一阶梯度**随机优化**的大规模算法

理论研究：泛化误差理论

经验模型、泛化模型：

- $\hat{f}_n = \arg \min_{f \in \mathcal{H}} \hat{\mathcal{E}}(f)$ 经验误差最小化模型, 实际学得模型
- $f^* = \arg \min_{f \in \mathcal{H}} \mathcal{E}(f)$ 期望误差最小化模型, 假设空间中最优模型

泛化误差界：经验学得模型与假设空间中最优模型的期望误差的差距

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*)$$

通常，使用假设空间复杂度度量工具对泛化误差界进行界定

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq \text{Measure}(\mathcal{H})$$

计算学习理论：最小化复杂度度量 \Rightarrow 最小化泛化误差界

$$\arg \min_{\kappa \in \mathcal{K}} [\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*)] \leq \arg \min_{\kappa \in \mathcal{K}} \text{Measure}(\mathcal{H})$$

理论研究：基于谱分析的谱度量泛化误差界

- 核矩阵谱分解(Spectral Decomposition)

$$\mathbf{K}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

其中, $(\lambda_i, \mathbf{v}_i)$ 为核矩阵对应的特征值、特征向量, $i = 1, \dots, n$.

- 将谱度量 (Spectral Measure) 定义为

$$\text{SM}(\mathbf{K}, \varphi) := \frac{1}{n} \sum_{i=1}^n \lambda_i^r \langle \mathbf{y}, \mathbf{v}_i \rangle^2 = \mathbf{y}^T \left(\sum_{i=1}^n \lambda_i^r \mathbf{v}_i \mathbf{v}_i^T \right) \mathbf{y} = \mathbf{y}^T \mathbf{K}^r \mathbf{y}$$

其中, $r \geq 1$.

理论研究：基于谱分析的谱度量泛化误差界

定理 (谱度量泛化误差界)

令 $LSSVM$ 、 SVM 学习器满足 $\forall f \in \mathcal{H}, \|f\|_{\kappa} \leq 1$. 任意训练样本个数满足 $n \geq 5$, 以至少 $1 - \delta$ 的概率存在如下泛化误差界

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq 1 - c_0 \cdot \text{SM}(\mathbf{K}, \varphi) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

其中 c_0 为常数.

- 基于核矩阵谱分析, 为二分类 $LSSVM$ 、 SVM 推导出基于谱度量泛化误差界 [Li et al. 2017, IJCAI 2017].
- 为核对齐 (KTA) 类型的核方法模型选择方法提供泛化理论保证

$$\text{SM}(\mathbf{K}, \varphi) = \mathbf{y}^T \mathbf{K}^r \mathbf{y}$$

理论研究：基于Rademacher复杂度的泛化误差界

定义 (损失空间上的 Rademacher 复杂度)

基于有标签数据集 $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, 定义在损失空间上 \mathcal{L} 的 Rademacher 复杂度定义为:

$$\widehat{\mathcal{R}}(\mathcal{L}) = \frac{1}{n} \mathbb{E}_{\epsilon} \left[\sup_{\ell \in \mathcal{L}} \sum_{i=1}^n \epsilon_i \ell(f(\mathbf{x}_i), \mathbf{y}_i) \right],$$

其中 ϵ_i 为 Rademacher 随机变量 $\Pr(\epsilon_i = +1) = \Pr(\epsilon_i = -1) = 1/2$, ℓ 为损失函数. 期望 Rademacher 复杂度为 $\mathcal{R}(\mathcal{L}) = \mathbb{E} \widehat{\mathcal{R}}(\mathcal{L})$.

- 损失空间复杂度越小, 对应泛化性能越好。
- 局部 Rademacher 复杂度获取更优的子空间, 能得到更好的泛化性能。
- 对应数据依赖的多分类泛化误差界 (K 个类别) 通常为

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq \mathcal{O} \left(\sqrt{\frac{K}{n}} \right).$$

理论研究：基于 Rademacher 复杂度的泛化误差界

首次将局部Rademacher复杂度引入到多分类中, 得到当前最紧的多分类泛化误差界 [Li et al. 2018, NeurIPS]

$$\mathcal{O}\left(\sqrt{\frac{K}{n}}\right) \Rightarrow \mathcal{O}\left(\frac{\log K}{n}\right).$$

- 考虑类别间内在关联, 降低对类别数 K 的依赖

$$\sqrt{K} \Rightarrow \log K$$

- 引入局部 Rademacher 复杂度, 使用更小的假设空间

$$\frac{1}{\sqrt{n}} \Rightarrow \frac{1}{n}$$

理论研究：基于Rademacher复杂度的泛化误差界

使用**半监督数据**构造标签无关 Rademacher 复杂度, 从而对泛化误差界进行提升 [Li et al. 2019d, IJCAI]

$$\mathcal{O}\left(\sqrt{\frac{K}{n}}\right) \Rightarrow \mathcal{O}\left(\sqrt{\frac{K}{n+u}}\right).$$

- 定义在**损失空间**的 Rademacher 复杂度泛化误差界 (**标签相关**), 有监督

$$\widehat{\mathcal{R}}(\mathcal{L}) = \frac{1}{n} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{H}} \sum_{i=1}^n \epsilon_i \ell(f(\mathbf{x}_i), y_i) \right].$$

- 定义在**假设空间**的 Rademacher 复杂度泛化误差界 (**标签无关**), 半监督

$$\widehat{\mathcal{R}}(\mathcal{H}) = \frac{1}{n+u} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{H}} \sum_{i=1}^{n+u} \sum_{k=1}^K \epsilon_{ik} f_k(\mathbf{x}_i) \right].$$

- 基于 L-Lipschitz 连续的假设, 建立标签相关 Rademacher 复杂度与标签无关 Rademacher 复杂度的关联

$$\widehat{\mathcal{R}}(\mathcal{L}) \leq \sqrt{2} L \widehat{\mathcal{R}}(\mathcal{H}).$$

理论研究：基于Rademacher复杂度的泛化误差界

使用局部 Rademacher 复杂度、无标签数据的多输出学习 (Vector-valued learning) 泛化误差界 [Li et al. 2019b, arXiv: 1909.04883]。

对比情况 ($u \gg n$)

泛化误差界结果	最差情况	较好情况
GRC [Cortes et al. 2016]	$\mathcal{O}\left(\sqrt{\frac{\log K}{n}}\right)$	
GRC [Maurer 2016]	$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	
LRC [Li et al. 2019b] †	$\mathcal{O}\left(\frac{1}{\sqrt{n+u}}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$

- 最差情况下：使用 无标签数据 提升泛化误差收敛率

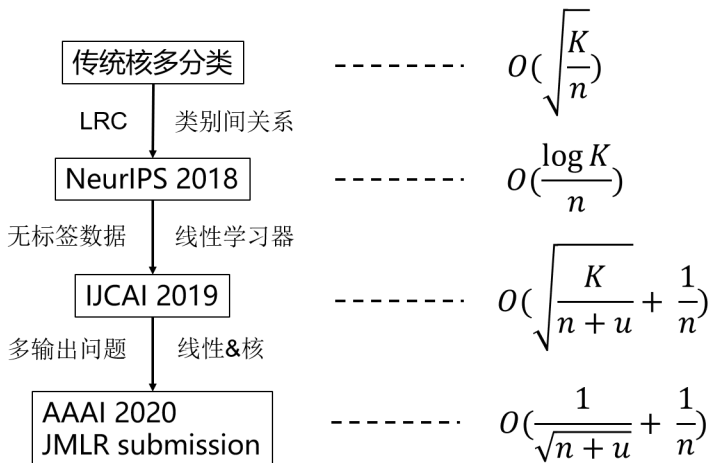
$$\frac{1}{\sqrt{n}} \Rightarrow \frac{1}{\sqrt{n+u}}$$

- 较好情况下：使用 局部 Rademacher 复杂度 提升泛化误差收敛率

$$\frac{1}{\sqrt{n}} \Rightarrow \frac{1}{n}$$

理论研究：基于Rademacher复杂度的泛化误差界

已完成工作针对Rademacher 复杂度泛化误差界研究脉络



理论研究：基于积分算子理论的泛化误差界

基于积分算子的最优泛化理论

- 假设空间的**容量假设**：存在 $Q > 0, \gamma \in [0, 1]$, 对于任意 $1 > \lambda_A > 0$, 满足

$$\mathcal{N}(\lambda_A) \leq Q^2 \lambda_A^{-\gamma}.$$

- 假设空间的**正则化假设**：存在 $r \in [1/2, 1]$ 及 $g \in L^2(\mathcal{X}, \rho_X)$, 满足

$$f^*(\mathbf{x}) = (L^r g)(\mathbf{x}).$$

- 核岭回归 (KRR) 最优泛化误差率

$$\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$$

- γ 越小（假设空间越小）、 r 越大（正则化越强），泛化误差收敛率越快

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \Rightarrow \mathcal{O}\left(\frac{1}{n}\right)$$

任务的难易程度 \Rightarrow 假设的满足情况 \Rightarrow 泛化误差收敛率

理论研究：基于积分算子理论的泛化误差界

将 分布式、随机特征 两个近似算法相结合加速 KRR, 并首次为该组合给出其最优泛化误差收敛界 [Li et al. 2019a, arXiv: 1906.03155].

- 当随机特征个数 M 、分块数 m 分别满足

$$M \gtrsim n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad m \lesssim n^{\frac{2r-1}{2r+\gamma}},$$

则能够达到最优泛化收敛率 $\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$.

- 即使结合了两个近似算法（分布式、随机特征）对原始 KRR 进行了近似，在满足一定条件下，仍能达到原始 KRR 相同的泛化性能。
- 以 $\gamma = 0, r = 1$ 为例
 - KRR

$$\text{空间: } \mathcal{O}(n^2) \quad \text{时间: } \mathcal{O}(n^3) \quad \text{误差收敛率: } \mathcal{O}\left(\frac{1}{n}\right)$$

- KRR-DC-RF

$$\text{空间: } \mathcal{O}(n) \quad \text{时间: } \mathcal{O}(n\sqrt{n}) \quad \text{误差收敛率: } \mathcal{O}\left(\frac{1}{n}\right)$$

计算效率提升极大，仍能达到 $\mathcal{O}(1/n)$ 的收敛率

理论研究：基于积分算子理论的泛化误差界

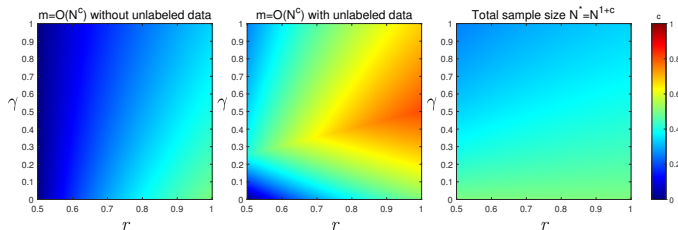
当 $r = 1/2$ 时候, 分块数为常数 $\mathcal{O}(1)$. 使用额外的无标签数据提升分块数 [Li et al. 2019a, arXiv: 1906.03155].

- 当随机特征个数 M , 分块数 m 满足

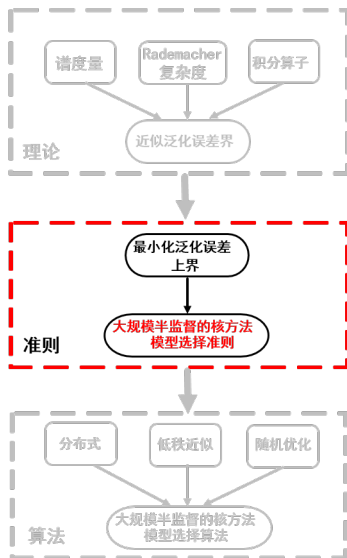
$$M \gtrsim n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad m \lesssim \min \left\{ n^{\frac{2r+2\gamma-1}{2r+\gamma}}, (n+u)n^{\frac{-\gamma-1}{2r+\gamma}} \right\}.$$

则能够达到最优泛化收敛率 $\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$.

- 无标签数据对分块数的提升：避免了常数分块数



大规模半监督的核方法模型选择准则研究



◆ 大规模半监督的核方法模型选择理论研究

- 基于谱度量的二分类泛化误差界
- 基于Rademacher复杂度的多输出泛化误差界
- 基于积分算子理论的最优泛化误差界

◆ 大规模半监督的核方法模型选择准则研究

- 最大化谱度量
- 最小化核矩阵尾部特征值之和
- 反向传播更新核超参数

◆ 大规模半监督的核方法模型选择算法研究

- 基于分布式核方法的大规模算法
- 基于核方法低秩近似的大规模算法
- 基于一阶梯度随机优化的大规模算法

准则研究：最大化谱度量

通过最小化泛化误差界, 将最大化谱度量作为核方法模型选择准则 [Li et al. 2017, IJCAI 2017]

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq 1 - c_0 \cdot \text{SM}(\mathbf{K}, \varphi) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

- 最大化谱度量 \Rightarrow 最小化泛化误差界

$$\arg \max_{\kappa \in \mathcal{K}} \text{SM}(\mathbf{K}, \varphi) = \arg \max_{\kappa \in \mathcal{K}} \mathbf{y}^T \mathbf{K}^r \mathbf{y}$$

- 获得兼具 泛化理论保证、 高效计算效率 的核方法模型选择准则

准则研究：最小化核矩阵尾部特征值之和

通过最小化局部 Rademacher 复杂度的泛化误差界, 将最小化核矩阵的尾部特征值之和作为准则 [Li et al. 2018, NeurIPS]

- 由于单核核矩阵的特征值是固定的, 无法提升学习器。故使用核矩阵特征值可变的多核组合

$$\kappa_{\mu}(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^M \mu_m \kappa_m(\mathbf{x}, \mathbf{x}')$$

- 将最小化多核矩阵的尾部特征值之和放在优化目标中

$$\min_{\mathbf{W}, \mu} \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i) + \frac{\alpha}{2} \|\mathbf{W}\|_{2,p}^2 + \beta \sum_{m=1}^M \mu_m r_m$$

- 以端到端形式, 同时学习多核参数 μ 、核学习模型参数 \mathbf{W}
- 避免了传统核方法模型选择方法存在的不足
 - 需要预先指定候选核函数, 过度依赖于人工经验
 - 割裂了核选择、模型学习的过程, 计算效率低下

准则研究：反向传播更新核超参数 I

多核学习优化复杂, 计算效率较低.

- 构造非平稳谱核的网络, 通过反向传播以端到端形式, 同时学习核超参数、模型参数 [Li et al. 2020a, AAAI].
- 基于随机傅里叶特征, 使用蒙特卡洛采样近似非平稳谱核,

$$\kappa(\mathbf{x}, \mathbf{x}') \approx \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

- 非平稳谱核定义为

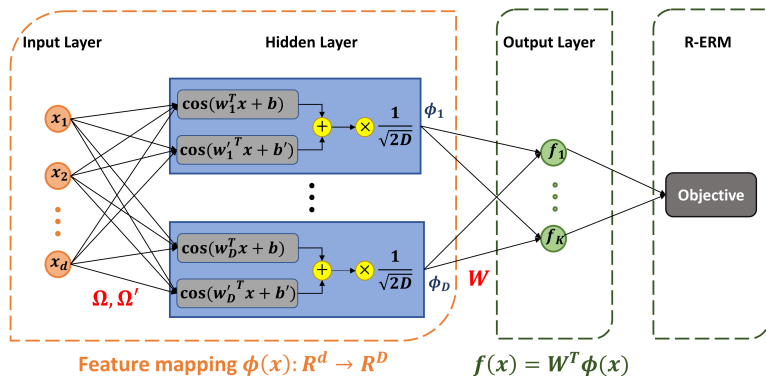
$$f(\mathbf{x}) = \mathbf{W}^T \phi(\mathbf{x}), \quad \phi(\mathbf{x}) = \frac{1}{\sqrt{2D}} [\cos(\mathbf{\Omega}^T \mathbf{x} + \mathbf{b}) + \cos(\mathbf{\Omega}'^T \mathbf{x} + \mathbf{b}')],$$

$\mathbf{\Omega}, \mathbf{\Omega}'$ 为核超参数, \mathbf{W} 为模型参数。

- 优化目标为

$$\arg \min_{\mathbf{W}, \mathbf{\Omega}, \mathbf{\Omega}'} \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda_A \|\mathbf{W}\|_* + \lambda_B \|\phi(\mathbf{X})\|_F^2.$$

准则研究：反向传播更新核超参数 II

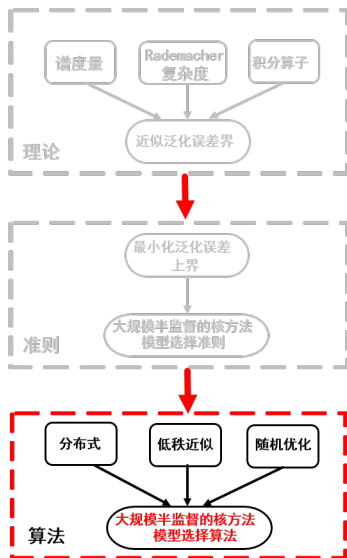


- 提出以端到端的形式同时学习核超参数 Ω, Ω' 、模型参数 W
- 与传统核方法模型选择方法对比
 - 计算高效：与单隐藏层神经网络计算时间复杂度相同。
 - 泛化理论分析保证：使用反向传播能够降低泛化误差上界。

准则研究：核方法模型选择准则内部对比

准则	适用范围	学习方式	人工干预
最大化谱度量	二分类	核选择（枚举） + 模型训练	候选核函数集合，通常较大
最小化核矩阵尾部特征值之和	分类问题	多核学习： 组合系数、模型参数	多个基核
反向传播更新核超参数	有监督问题	单核学习： 核超参数、模型参数	单核
	适用范围变广	计算简化、效率更高	人工依赖更少

大规模半监督的核方法模型选择算法研究



◆ 大规模半监督的核方法模型选择理论研究

- 基于谱度量的二分类泛化误差界
- 基于Rademacher复杂度的多输出泛化误差界
- 基于积分算子理论的最优泛化误差界

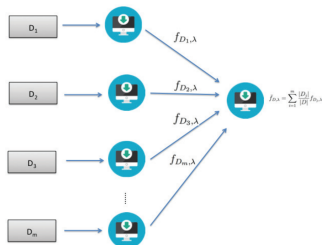
◆ 大规模半监督的核方法模型选择准则研究

- 最大化谱度量
- 最小化核矩阵尾部特征值之和
- 反向传播更新核超参数

◆ 大规模半监督的核方法模型选择**算法**研究

- 基于**分布式**核方法的大规模算法
- 基于核方法**低秩近似**的大规模算法
- 基于一阶梯度**随机优化**的大规模算法

算法研究：基于分布式核方法的大规模算法 I



基于分块数据的局部核岭回归学习器 [Li et al. 2019a, arXiv: 1906.03155]:

$$\hat{f}_{D_j, \lambda}(\mathbf{x}) = \sum_{i=1}^{\bar{n}} \hat{\alpha}_{ij} \kappa(\mathbf{x}_i, \mathbf{x}), \quad \hat{\alpha}_j = (\mathbf{K} + \lambda \bar{n} \mathbf{I})^{-1} \mathbf{y},$$

将局部学习器平均, 构成全局学习器 (KRR-DC)

$$\hat{f}_{D, \lambda} = \frac{1}{m} \sum_{j=1}^m \hat{f}_{D_j, \lambda}.$$

算法研究：基于分布式核方法的大规模算法 II

定理 (KRR-DC 最优泛化误差收敛率保证)

当分治算法的分块数 m 满足

$$m \lesssim \mathcal{O}\left(n^{\frac{2r-1}{2r+\gamma}}\right),$$

KRR-DC 能够达到最优误差收敛率 $\mathcal{O}\left(n^{\frac{-2r}{2r+\gamma}}\right)$

容量假设	正则化假设	分块数 m	空间复杂度	时间复杂度	泛化误差
$\gamma = 1$	$r = 0.5$	$\mathcal{O}(1)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$	$\mathcal{O}\left(\frac{1}{n^{0.5}}\right)$
$\gamma = 0.5$	$r = 0.75$	$\mathcal{O}(n^{0.25})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{2.25})$	$\mathcal{O}\left(\frac{1}{n^{0.75}}\right)$
$\gamma = 0$	$r = 1$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}\left(\frac{1}{n}\right)$

标准 KRR 的空间复杂度为 $\mathcal{O}(n^2)$ 、时间复杂度为 $\mathcal{O}(n^3)$ 。

可以看出除最差情况外, **KRR-DC 能很好地应对大规模核岭回归问题。**

算法研究：基于 Nyström 采样的大规模算法 I

$$\begin{bmatrix} \boxed{K_{ss}} & \boxed{K_{12}} \\ \boxed{K_{21}} & \boxed{K_{22}} \end{bmatrix} \approx \begin{bmatrix} \boxed{K_{ss}} \\ \boxed{K_{21}} \end{bmatrix} \times \boxed{K_{ss}}^{\dagger} \times \begin{bmatrix} \boxed{K_{ss}} & \boxed{K_{12}} \end{bmatrix}$$

将原始核矩阵根据 Nyström 采样点进行分块

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{ss} & \mathbf{K}_{21}^{\top} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}, \quad \mathbf{K}_{ns} = \begin{bmatrix} \mathbf{K}_{ss} \\ \mathbf{K}_{21} \end{bmatrix}.$$

使用 **Nyström 方法** 近似核矩阵, 获得半监督核矩阵的 s -阶低秩近似 [Li et al. 2019c, IJCAI]

$$\mathbf{K} \approx \mathbf{K}_{ns} \mathbf{K}_{ss}^{\dagger} \mathbf{K}_{ns}^T$$

算法研究：基于 Nyström 采样的大规模算法 II

定理 (KRR-Nyström 最优泛化误差收敛率保证)

当 Nyström 采样中心点数满足

$$s \gtrsim \mathcal{O}\left(n^{\frac{1}{2r+\gamma}}\right)$$

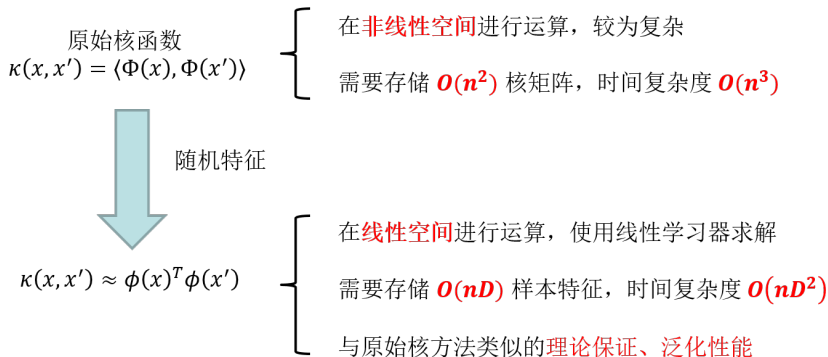
KRR-Nyström 能够达到最优泛化误差收敛率 $\mathcal{O}(n^{\frac{-2r}{2r+\gamma}})$

容量假设	正则化假设	采样数 s	空间复杂度	时间复杂度	泛化误差
$\gamma = 1$	$r = 0.5$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}\left(\frac{1}{n^{0.5}}\right)$
$\gamma = 0.5$	$r = 0.75$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}\left(\frac{1}{n^{0.75}}\right)$
$\gamma = 0$	$r = 1$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}\left(\frac{1}{n}\right)$

标准 KRR 的空间复杂度为 $\mathcal{O}(n^2)$ 、时间复杂度为 $\mathcal{O}(n^3)$ 。

KRR-Nyström 能很好地应对大规模核岭回归问题。

算法研究：基于随机特征的大规模算法 I



构造**随机特征**近似核函数 [Li et al. 2019a, arXiv: 1906.03155]、[Li et al. 2020a, AAAI]、[Li et al. 2020b, arXiv: 2002.12744]

$$\kappa(\mathbf{x}, \mathbf{y}) \approx \phi(\mathbf{x})^T \phi(\mathbf{y})$$

其中，随机特征映射为 $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^D$.

算法研究：基于随机特征的大规模算法 II

定理 (KRR-RF 最优泛化误差收敛率保证)

当随机特征数 M 满足

$$M \geq \mathcal{O}\left(n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\right),$$

KRR-RF 能够达到最优泛化误差收敛率 $\mathcal{O}(n^{\frac{-2r}{2r+\gamma}})$.

容量假设	正则化假设	采样数 s	空间复杂度	时间复杂度	泛化误差
$\gamma = 1$	$r = 0.5$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}\left(\frac{1}{n^{0.5}}\right)$
$\gamma = 0.5$	$r = 0.75$	$\mathcal{O}(n^{0.625})$	$\mathcal{O}(n^{1.625})$	$\mathcal{O}(n^{2.25})$	$\mathcal{O}\left(\frac{1}{n^{0.75}}\right)$
$\gamma = 0$	$r = 1$	$\mathcal{O}(n^{0.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}\left(\frac{1}{n}\right)$

标准 KRR 的空间复杂度为 $\mathcal{O}(n^2)$ 、时间复杂度为 $\mathcal{O}(n^3)$.

KRR-RF 很好地应对大规模核岭回归问题, 但计算效率稍差于 KRR – Nyström.

三种近似核岭回归方法对比

表 1: 最差假设 ($\gamma = 1, r = 0.5$) 下, 各方法空间、时间复杂度对比

	KRR	KRR-DC	KRR-Nyström	KRR-RF
空间复杂度	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$
时间复杂度	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

表 2: 一般假设 ($\gamma = 0.5, r = 0.75$) 下, 各方法空间、时间复杂度对比

	KRR	KRR-DC	KRR-Nyström	KRR-RF
空间复杂度	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.625})$
时间复杂度	$\mathcal{O}(n^3)$	$\mathcal{O}(n^{2.25})$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{2.25})$

表 3: 最优假设 ($\gamma = 0, r = 1$) 下, 各方法空间、时间复杂度对比

	KRR	KRR-DC	KRR-Nyström	KRR-RF
空间复杂度	$\mathcal{O}(n^2)$	$\mathcal{O}(n^1)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$
时间复杂度	$\mathcal{O}(n^3)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

算法研究：对偶梯度下降算法

使用**对偶梯度下降**算法在 **原问题** 上求解核方法 [Li et al. 2018, NeurIPS]. 将优化目标分为两部分, 引入原变量 \mathbf{W} 、对偶变量 $\boldsymbol{\theta}$ 进行求解

$$\min_{\mathbf{W}, \boldsymbol{\mu}} \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(\mathbf{W}, \phi_{\boldsymbol{\mu}}(\mathbf{x}_i), y_i)}_{C(\mathbf{W})} + \underbrace{\frac{\alpha}{2} \|\mathbf{W}\|_{2,p}^2 + \beta \sum_{m=1}^M \mu_m r_m}_{\Omega(\mathbf{W})}$$

- 使用 $C(\mathbf{W})$ 更新对偶变量 $\boldsymbol{\theta}$.
- 根据 $\Omega(\mathbf{W})$ 的费舍尔对偶(Fenchel dual)更新原变量 \mathbf{W} .

使用对偶梯度方法较为**高效地**求解了该多核多分类问题

- 在**原问题**上求解, 避免了对偶问题上使用半正定规划的低效求解方法.
- 使用梯度方法**更新变量的范数**, 避免了高维特征空间的运算.

算法研究：近端梯度下降算法

使用**近端梯度下降(SVT)**算法在求解 **不可微** 的最小化迹范数问题 [Li et al. 2019d, IJCAI], [Li et al. 2019b, arXiv:1909.04883].

$$\arg \min_{f \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i) + \tau_A \|\mathbf{W}\|_F^2 + \tau_I \text{trace}(\mathbf{W}^T \mathbf{x} \mathbf{L} \mathbf{x}^T \mathbf{W})}_{g(\mathbf{W})} + \tau_S \sum_{j > \theta} \lambda_j(\mathbf{W}).$$

- 对可微部分使用梯度下降算法更新 \mathbf{W} , 并求解更新后 \mathbf{W} 的SVD分解.

$$\mathbf{U} \Sigma \mathbf{V}^T = \mathbf{W}^t - \frac{1}{\mu} \nabla g(\mathbf{W}^t, \mathbf{x}_{i_t})$$

- 对首部奇异值矩阵进行更新后, 求解新的 \mathbf{W} .

$$\mathbf{W}^{t+1} = \mathbf{U} \Sigma_{\frac{\tau_S}{\mu}}^{\theta} \mathbf{V}^T$$
$$(\Sigma_{\frac{\tau_S}{\mu}}^{\theta})_{jj} = \begin{cases} \max(0, \Sigma_{jj} - \frac{\tau_S}{\mu}), & i \leq \theta, \\ \Sigma_{jj}, & i > \theta. \end{cases}$$

算法研究：预处理共轭梯度(PCG)

使用**预处理共轭梯度 (PCG)**方法加速闭式解求解 [Li et al. 2019c, IJCAI]

$$\hat{f}_{\lambda}^s(\mathbf{x}) = \sum_{i=1}^s \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}), \quad \text{with}$$
$$\boldsymbol{\alpha} = \underbrace{(\mathbf{K}_{ms}^T \mathbf{K}_{ms} + \lambda_A \mathbf{K}_{ss} + \lambda_I \mathbf{K}_{ns}^T \mathbf{L} \mathbf{K}_{ns})}_{\mathbf{H}}^{\dagger} \underbrace{\mathbf{K}_{ms}^T \mathbf{y}}_{\mathbf{z}},$$

- 梯度方法求解闭式解 $\mathbf{H}\boldsymbol{\alpha} = \mathbf{z}$ 的迭代次数由条件数(Conditon number) $\kappa(\mathbf{H}) = \frac{\lambda_{\max}(\mathbf{H})}{\lambda_{\min}(\mathbf{H})}$ 所决定, 条件数越小(well-conditioned), 迭代次数越少.
- 构造合适的预处理器**减小条件数, 从而降低迭代次数**. 即构造 \mathbf{P} 尽量近似 \mathbf{H} , 同时要易于求解.

$$\mathbf{P}^{-1} \mathbf{H} \boldsymbol{\alpha} = \mathbf{P}^{-1} \mathbf{z}.$$

- 将迭代次数从 $\mathcal{O}(\sqrt{s} \log s)$ 降低到 $\mathcal{O}(\log s)$

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

实验设置

- 对比方法

表 4: 最大化谱度量相关对比方法 [Li et al. 2017, IJCAI]

核方法模型选择准则	时间复杂度	理论
Cross-Validation	至少 $O(n^3)$	有
KTA, CKTA, FSM	$O(n^2)$	无
ER	$O(n^3)$	有
SM (最大化谱度量)	$O(n^2)$	有

- 实验数据集均取自 UCI 和 LIBSVM 中的二分类数据集。

验证准则：最大化谱度量 II

表 5: 最大化谱度量相关方法的平均测试误分类率 (%) [Li et al. 2017, IJCAI]

	SM	CV	ELOO	CKTA	FSM	ER
a1a	16.84±1.39	17.02±1.57	16.88±1.41	18.86±1.49	24.72±1.67	16.97±1.52
a2a	17.78±1.28	17.96±1.25	17.94±1.27	18.52±1.26	25.62±1.47	18.99±1.37
anneal	2.69±3.28	3.81±4.11	2.69±3.28	4.75±4.78	5.13±4.18	5.50±4.95
australian	13.71±2.10	13.84±2.18	13.82±2.04	13.91±1.89	44.71±2.47	13.53±2.06
autos	11.81±11.67	11.81±11.67	12.75±11.06	13.71±12.03	12.71±8.06	12.14±11.51
breast-w	3.27±1.01	3.56±1.16	3.59±1.08	3.51±1.05	3.50±1.05	4.26±1.40
breast-cancer	3.18±1.15	3.63±1.16	3.50±1.23	3.63±1.16	3.60±1.14	4.04±1.12
bupa	30.29±3.48	29.10±4.04	30.31±4.27	35.81±3.45	39.77±3.68	29.13±4.46
colic	15.62±3.00	16.47±2.78	15.73±2.97	19.27±2.58	36.42±3.28	17.35±3.09
diabetes	24.22±2.41	24.69±2.71	23.51±2.75	24.85±2.46	35.30±3.00	23.90±2.48
glass	22.09±5.07	21.82±5.68	20.95±4.82	26.41±7.13	43.00±9.22	22.50±5.08
german.numer	24.09±2.15	25.28±2.38	23.81±2.26	26.02±2.16	29.89±2.41	25.33±2.14
heart	16.53±3.27	16.69±3.36	15.95±3.29	18.67±3.78	44.37±5.50	15.98±3.47
hepatitis	15.57±4.68	17.09±5.74	16.63±4.64	<u>15.74±5.00</u>	21.22±5.41	<u>18.91±6.20</u>
ionosphere	4.88±2.10	5.28±2.11	6.42±2.17	11.70±3.43	35.77±4.00	4.86±1.99
labor	13.65±8.10	14.47±8.08	14.82±8.34	15.41±8.80	34.59±8.70	18.82±8.81
pima	23.80±2.14	<u>22.78±2.36</u>	22.51±2.41	24.38±2.28	34.47±2.42	22.78±2.07
segment	0.01±0.00	0.06±0.24	0.20±0.04	0.32±0.03	0.21±0.01	0.24±0.04
liver-disorders	31.94±3.21	29.00±4.11	30.02±4.76	36.27±3.93	40.90±4.10	29.69±4.97
sonar	15.06±4.80	14.26±4.93	13.68±4.43	15.00±5.51	49.32±6.93	18.84±5.75
vehicle	3.02±1.79	3.33±1.77	3.02±1.79	3.77±1.51	53.32±3.38	5.52±2.44
vote	4.31±1.71	4.78±1.74	4.82±1.73	5.25±1.72	6.37±3.96	7.80±2.33
wdbc	23.10±4.58	22.83±4.32	<u>21.93±4.45</u>	21.87±4.13	<u>22.13±4.19</u>	21.87±4.13
tic-tac-toe	10.10±1.93	10.28±1.66	9.78±1.66	33.62±5.31	34.44±2.04	14.62±2.05
wdbc	2.29±1.15	<u>2.43±1.07</u>	2.73±1.11	2.82±1.20	37.49±3.83	4.75±1.66

验证准则：最大化谱度量 III

表 6: 最大化谱度量相关算法训练时间对比 (秒) [Li et al. 2017, IJCAI]

	SM	CV	ELOO	CKTA	FSM	ER
a1a	63.77	704.50	217.03	87.51	79.71	422.71
a2a	124.53	1995.76	810.87	172.28	156.68	1558.70
anneal	0.28	4.60	0.84	0.43	0.55	1.36
australian	8.30	111.31	30.33	11.34	9.51	52.53
autos	0.10	2.18	0.23	0.17	0.29	0.37
breast-cancer	8.05	104.55	27.13	11.10	9.36	47.68
breast-w	8.59	105.39	29.77	11.80	10.11	49.86
colic	1.55	25.72	6.52	2.15	2.17	11.60
glass	0.33	5.74	1.15	0.53	0.66	2.00
heart	1.02	14.71	3.66	1.12	1.20	6.33
hepatitis	0.38	6.55	1.42	0.58	0.70	2.44
ionosphere	1.59	24.60	6.08	2.03	2.06	10.60
labor	0.16	2.88	0.42	0.28	0.42	0.67
pima	10.99	137.19	36.60	15.11	12.74	62.45
segment	7.50	91.73	23.83	10.49	8.93	42.63
diabetes	10.70	134.71	36.26	14.90	12.51	62.46
german.numer	21.80	249.98	72.63	29.59	26.27	127.51
liver-disorders	1.37	21.04	5.46	1.91	1.94	9.27
sonar	0.62	10.53	2.34	0.87	0.98	4.06
vehicle	2.03	35.64	9.35	2.86	2.85	15.95
vote	2.03	34.43	9.33	2.79	2.81	15.92
wdbc	0.52	9.06	2.07	0.76	0.92	3.33
bupa	1.36	21.09	5.34	1.85	1.88	9.19
tic-tac-toe	18.93	224.13	63.10	26.13	23.04	107.12
wdbc	5.75	61.32	19.18	7.86	6.80	33.65

验证准则：最小化核矩阵尾部特征值之和 I

实验设置

- 对比方法 [Li et al. 2018, NeurIPS]
 - 一对一 (One vs. One) Knerr et al. [1990]
 - 一对多 (One vs. Rest) Bottou et al. [1994]
 - ℓ_1 -范数线性SVM (LMC) Crammer and Singer [2002]
 - 广义极小范数问题求解器 (GMNP) Franc [2005]
 - ℓ_1 -范数多核多分类 (ℓ_1 MC-MKL) Zien and Ong [2007]
 - ℓ_2 -范数多核多分类 (ℓ_2 MC-MKL) Zien and Ong [2007]
 - 混合范数多核学习器 (UFO-MKL) Orabona and Luo [2011].
- 实验数据集均取自 UCI 和 LIBSVM 中的多分类数据集。

验证准则：最小化核矩阵尾部特征值之和 II

表 7: 最小化核矩阵尾部特征值之和和相关算法分类准确率 (%) [Li et al. 2018, NeurIPS]

	Conv-MKL	SMSD-MKL	LMC	One vs. One	One vs. Rest	GMNP	ℓ_1 MC-MKL	ℓ_2 MC-MKL	UFO-MKL
plant	77.14 \pm 2.25	78.01\pm2.17	70.12 \pm 2.96	75.83 \pm 2.69	75.17 \pm 2.68	75.42 \pm 3.64	<u>77.60\pm2.63</u>	75.49 \pm 2.48	76.77 \pm 2.42
psortPos	74.41 \pm 3.35	76.23\pm3.39	63.85 \pm 3.94	73.33 \pm 4.21	71.70 \pm 4.89	73.55 \pm 4.22	71.87 \pm 4.87	70.70 \pm 4.89	74.56 \pm 4.04
psortNeg	74.07 \pm 2.16	74.66\pm1.90	57.85 \pm 2.49	73.74 \pm 2.87	71.94 \pm 2.50	<u>74.27\pm2.51</u>	72.83 \pm 2.20	72.42 \pm 2.65	73.80 \pm 2.26
nonpl	79.15\pm1.51	78.69 \pm 1.58	75.16 \pm 1.48	77.78 \pm 1.52	77.49 \pm 1.53	78.35 \pm 1.46	77.89 \pm 1.79	77.95 \pm 1.64	78.07 \pm 1.56
sector	<u>92.83\pm2.62</u>	93.39\pm0.70	93.16 \pm 0.66	90.61 \pm 0.69	91.34 \pm 0.61	\	\	92.15 \pm 2.57	92.60 \pm 0.47
segment	96.79 \pm 0.91	97.62\pm0.83	95.07 \pm 1.11	97.08 \pm 0.61	97.02 \pm 0.80	96.87 \pm 0.80	96.98 \pm 0.64	<u>97.58\pm0.68</u>	97.20 \pm 0.82
vehicle	79.35\pm2.27	77.28 \pm 2.78	75.61 \pm 3.56	78.72 \pm 1.92	79.11 \pm 1.94	81.57 \pm 2.24	74.96 \pm 2.93	76.27 \pm 3.15	76.92 \pm 2.83
vowel	98.82 \pm 1.19	98.83\pm5.57	62.32 \pm 4.97	98.12 \pm 1.76	98.22 \pm 1.83	97.04 \pm 1.85	98.27 \pm 1.22	97.86 \pm 1.75	98.22 \pm 1.62
wine	99.63\pm0.96	99.63\pm0.96	97.87 \pm 2.80	97.24 \pm 3.05	98.14 \pm 3.04	97.69 \pm 2.43	98.61 \pm 1.75	98.52 \pm 1.89	99.44 \pm 1.13
dna	96.08 \pm 0.83	96.30\pm0.79	92.02 \pm 1.50	95.89 \pm 0.56	95.61 \pm 0.73	94.60 \pm 0.94	<u>96.27\pm0.68</u>	95.06 \pm 0.92	95.84 \pm 0.61
glass	75.19\pm5.05	73.72 \pm 5.80	63.95 \pm 6.04	71.98 \pm 5.75	70.00 \pm 5.75	71.24 \pm 8.14	69.07 \pm 8.08	74.03 \pm 6.41	72.46 \pm 6.12
iris	<u>96.67\pm2.94</u>	97.00\pm2.63	88.00 \pm 7.82	95.93 \pm 3.25	95.87 \pm 3.20	95.40 \pm 7.34	95.40 \pm 6.46	94.00 \pm 7.82	95.93 \pm 2.88
svmguide2	82.69 \pm 5.65	85.17\pm3.83	81.10 \pm 4.15	<u>84.79\pm3.45</u>	<u>84.27\pm3.03</u>	81.77 \pm 3.45	83.16 \pm 3.63	<u>83.84\pm4.21</u>	82.91 \pm 3.09
satimage	91.64 \pm 0.88	<u>91.78\pm0.82</u>	84.95 \pm 1.15	90.67 \pm 0.91	89.29 \pm 0.96	89.97 \pm 0.81	<u>91.86\pm0.62</u>	90.43 \pm 1.27	91.92\pm0.83

实验设置

- 对比方法

表 8: 自动谱核学习 (ASKL) 相关对比方法 [Li et al. 2020a, AAAI]

对比方法	谱核类型	核超参数	正则化项
SK [Rahimi and Recht 2007]	平稳谱核	指定	$\ \mathbf{W}\ _F^2$
NSK [Samo and Roberts 2015]	非平稳谱核	指定	$\ \mathbf{W}\ _F^2$
SKL [Huang et al. 2014]	平稳谱核	自动学习	$\ \mathbf{W}\ _F^2$
NSKL	非平稳谱核	自动学习	$\ \mathbf{W}\ _F^2$
ASKL	非平稳谱核	自动学习	$\ \mathbf{W}\ _*, \ \phi(\mathbf{X})\ _F^2$

- 实验数据集均取自 UCI 和 LIBSVM 中的多分类、多标签数据集。

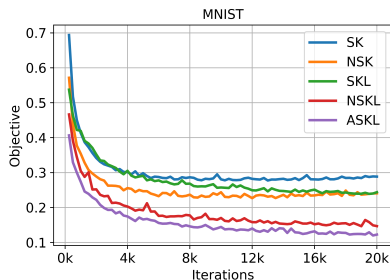
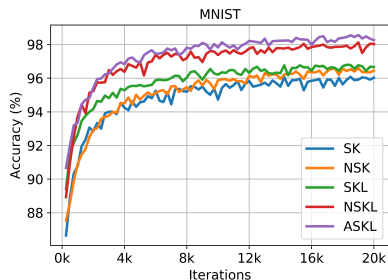
验证准则：反向传播更新核超参数 II

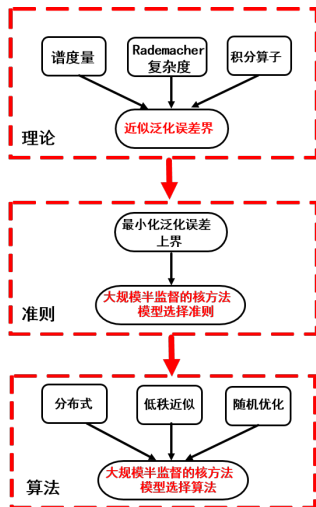
表 9: 自动谱核学习 (ASKL) 相关对比算法的平均测试结果 [Li et al. 2020a, AAAI]

		SK	NSK	SKL	NSKL	ASKL
Accuracy(\uparrow)	segment	89.93 \pm 2.12	90.15 \pm 2.08	94.58 \pm 1.86	94.37 \pm 0.81	95.02\pm1.54
	satimage	74.54 \pm 1.35	75.15 \pm 1.38	83.61 \pm 1.08	83.74 \pm 1.34	85.32\pm1.45
	USPS	93.19 \pm 2.84	93.81 \pm 2.13	95.13 \pm 0.91	95.27 \pm 1.65	97.76\pm1.14
	pendigits	96.93 \pm 1.53	97.39 \pm 1.41	98.19 \pm 2.30	98.28 \pm 1.68	99.06\pm1.26
	letter	76.50 \pm 1.21	78.21 \pm 1.56	93.60 \pm 1.14	94.66 \pm 2.21	95.70\pm1.74
	porker	49.80 \pm 2.11	51.85 \pm 0.97	54.27 \pm 2.72	54.69 \pm 1.68	54.85\pm1.28
	shuttle	98.17 \pm 2.81	98.21 \pm 1.46	98.87 \pm 1.42	98.74 \pm 1.07	98.98\pm0.94
	MNIST	96.03 \pm 2.21	96.45 \pm 2.16	96.67 \pm 1.61	98.03 \pm 1.16	98.26\pm1.78
RMSE(\downarrow)	abalone	10.09 \pm 0.42	9.71 \pm 0.28	8.35 \pm 0.28	7.85\pm0.42	7.88 \pm 0.16
	space_ga	11.86 \pm 0.26	11.58 \pm 0.42	11.40 \pm 0.18	11.39 \pm 0.46	11.34\pm0.27
	cpusmall	2.77 \pm 0.71	2.84 \pm 0.38	2.56 \pm 0.72	2.57 \pm 0.63	2.42\pm0.48
	cadata	50.31 \pm 0.92	51.47 \pm 0.32	47.67 \pm 0.33	47.71 \pm 0.30	46.34\pm0.23

验证准则：反向传播更新核超参数 III

图 1: 自动谱核学习 (ASKL) 相关对比算法的准确率、目标函数的收敛情况 [Li et al. 2020a, AAAI]





◆ 大规模半监督的核方法模型选择**理论**研究

- 基于**谱度量**的二分类泛化误差界
- 基于**Rademacher复杂度**的多输出泛化误差界
- 基于**积分算子理论**的最优泛化误差界

◆ 大规模半监督的核方法模型选择**准则**研究

- 最大化谱度量
- 最小化核矩阵尾部特征值之和
- 反向传播更新核超参数

◆ 大规模半监督的核方法模型选择**算法**研究

- 基于**分布式**核方法的大规模算法
- 基于核方法**低秩近似**的大规模算法
- 基于一阶梯度**随机优化**的大规模算法

1. 研究背景和意义
2. 研究现状与挑战
3. 主要研究成果
4. 实验分析
5. 完成项目及论文发表情况

完成项目及论文发表情况 I

科研成果：以**第一作者**发表(录用)**5篇CCF A类论文**.

- **Jian Li**, Yong Liu, Weiping Wang. [Automated Spectral Kernel Learning](#). In Proceedings of the 34th AAAI Conference on Artificial Intelligence (**AAAI 2020, CCF A**).
- **Jian Li**, Yong Liu, Rong Yin, Weiping Wang. [Multi-Class Learning using Unlabeled Samples: Theory and Algorithm](#). In Proceedings of the 28th International Joint Conference on Artificial Intelligence (**IJCAI 2019, CCF A**).
- **Jian Li**, Yong Liu, Rong Yin, Weiping Wang. [Approximate Manifold Regularization: Scalable Algorithm and Generalization Analysis](#). In Proceedings of the 28th International Joint Conference on Artificial Intelligence (**IJCAI 2019, CCF A**).
- **Jian Li**, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, Weiping Wang. [Multi-Class Learning: From Theory to Algorithm](#). Advances in Neural Information Processing Systems 31 (**NeurIPS 2018, CCF A**).
- **Jian Li**, Yong Liu, Hailun Lin, Yinliang Yue, Weiping Wang. [Efficient Kernel Selection via Spectral Analysis](#). In Proceedings of the 26th International Joint Conference on Artificial Intelligence (**IJCAI 2017, CCF A**).

完成项目及论文发表情况 II

投稿论文

- **Jian Li**, Yong Liu, Weiping Wang. [Convolutional Spectral Kernel Learning](#). arXiv: 2002.12744. (Submission in NeurIPS)
- **Jian Li**, Yong Liu, Weiping Wang. [Distributed Learning with Random Features](#). arXiv: 1906.03155. (Submission in JMLR)
- Yong Liu*, **Jian Li***, Lizhong Ding, Weiping Wang. [Learning Vector-valued Functions with Local Rademacher Complexity and Unlabeled Data](#). arXiv: 1909.04883. (Submission in JMLR)
- Yong Liu, **Jian Li**, Guangjun Wu, Lizhong Ding, Weiping Wang. [Efficient Cross-Validation for Semi-Supervised Learning](#). arXiv:1902.04768.
- Yong Liu, **Jian Li**, Weiping Wang. [Max-Diversity Distributed Learning: Theory and Algorithms](#). arXiv:1812.07738, 2018.

参与科研项目

- 国家自然科学基金项目 (No.61703396), “大规模核方法积分算子谱分析的模型选择方法”.
- 中科院青促会人才项目 (No.2018YFC0823104).
- 信工所引进优秀青年人才项目 (Y7Z0111107).

在学期间获奖情况

- ① 2019年度，博士研究生国家奖学金.
- ② 2019年度，“朱李月华”优秀博士生奖学金.
- ③ 2019年度，中国科学院院长优秀奖.
- ④ 2019年度，国科大博士研究生国际合作培养计划.
- ⑤ 2019年度，中国科学院大学“三好学生”荣誉称号.
- ⑥ 2018年度，博士研究生国家奖学金.
- ⑦ 2018年度，中国科学院大学“三好学生”荣誉称号.
- ⑧ 2018年度，信息工程研究所所长优秀奖.
- ⑨ 2018年度，信息工程研究所第二研究室优秀学生.
- ⑩ 2017年度，信息工程研究所第二研究室优秀学生.

References I

- Francis Bach, Gert Lanckriet, and Michael Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 41–48, 2004.
- Yoram Baram. Learning by kernel polarization. *Neural Computation*, 17(6):1264–1275, 2005.
- Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Lawrence D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pages 77–82, 1994.
- Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.
- Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2760–2768, 2013.
- Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 2514–2522, 2016.
- Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.
- Nello Cristianini, Andre Elisseeff, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. In *International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 367–373, 2001.
- Michiel Debruyne, Mia Hubert, and Johan A.K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9:2377–2400, 2008.
- Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- Vojtech Franc. Optimization algorithms for kernel methods. *Prague: A PhD dissertation. Czech Technical University*, 2005.
- Peter Vincent Gehler and Sebastian Nowozin. Infinite kernel learning. 2008.
- Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011.
- Yina Han, Kunde Yang, Yixin Yang, and Yuanliang Ma. On the impact of regularization variation on localized multiple kernel learning. *IEEE transactions on neural networks and learning systems*, 29(6):2625–2630, 2017.

References II

- Po-Sen Huang, Haim Avron, Tara N Sainath, Vikas Sindhwani, and Bhuvana Ramabhadran. Kernel methods match deep neural networks on timit. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 205–209, 2014.
- Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.
- Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing*, pages 41–50. Springer, 1990.
- Jian Li, Yong Liu, Hailun Lin, Yinliang Yue, and Weiping Wang. Efficient kernel selection via spectral analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2124–2130, 2017.
- Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1591–1600, 2018.
- Jian Li, Yong Liu, and Weiping Wang. Distributed learning with random features. *arXiv preprint arXiv:1906.03155*, 2019a.
- Jian Li, Yong Liu, and Weiping Wang. Learning vector-valued functions with local rademacher complexity. *arXiv preprint arXiv:1909.04883*, 2019b.
- Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Approximate manifold regularization: scalable algorithm and generalization analysis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2887–2893, 2019c.
- Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Multi-class learning using unlabeled samples : Theory and algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019d.
- Jian Li, Yong Liu, and Weiping Wang. Automated spectral kernel learning. In *Thirty-Four AAAI Conference on Artificial Intelligence*, 2020a.
- Jian Li, Yong Liu, and Weiping Wang. Convolutional spectral kernel learning. *arXiv preprint arXiv:2002.12744*, 2020b.
- Yong Liu, Shali Jiang, and Shizhong Liao. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 324–332, 2014.
- Yong Liu, Shizhong Liao, Hailun Lin, Yinliang Yue, and Weiping Wang. Infinite kernel learning: generalization bounds and algorithms. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- Yong Liu, Shizhong Liao, Shali Jiang, Lizhong Ding, Hailun Lin, and Weiping Wang. Fast cross-validation for kernel-based algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.
- Francesco Orabona and Jie Luo. Ultra-fast optimization algorithm for sparse multi kernel learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 249–256, 2011.

References III

- María Pérez-Ortiz, Pedro Antonio Gutiérrez, Javier Sánchez-Monedero, and César Hervás-Martínez. A study on multi-scale kernel optimisation via centered kernel-target alignment. *Neural Processing Letters*, 44(2):491–517, 2016.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1177–1184, 2007.
- Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.
- Yves-Laurent Kom Samo and Stephen Roberts. Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*, 2015.
- Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5): 1299–1319, 1998.
- Huibin Shen, Sandor Szedmak, Céline Brouard, and Juho Rousu. Soft kernel target alignment for two-stage multiple kernel learning. In *International Conference on Discovery Science*, pages 427–441. Springer, 2016.
- Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 111–133, 1974.
- Alexander Zien and Cheng Soon Ong. Multiclass multiple kernel learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 1191–1198, 2007.

恳请各位老师批评指正！