可靠的半监督分类算法研究

答辩编号：19-8

答辩内容

1. 研究问题背景
2. 现有问题与挑战
3. 无标签数据权重学习
4. 半监督数据流扩展算法
5. 总结与展望

研究问题背景

半监督学习

利用大量无标签数据X_U和少量标签数据X_L进行建模

现有模型:

- ✓ Co-training model
- ✓ Graph-based model
- ✓ Micro-cluster model
- ✓ Online manifold
- ✓ …

存在问题:

半监督学习的可靠性问题

$$f_{SL}(X_L) \geq f_{SSL}(X_L, X_U)$$

问题分析

半监督学习的安全性

- **无标签的噪声数据**
  训练数据集中含有噪声数据或者无关数据（Universum）
- **模型假设和数据的不一致**
  模型假设条件过于强烈，使得难以在训练数据上成立
- **非凸优化与局部最小值**
  半监督模型的优化目标往往是非凸优化问题，效果不稳定
- **模型评估的偏执**
  单个评估指标上的优势不代表模型的可靠性
- **其他因素**
  弱监督信息；难分类样本（e.g., self-paced learning）

研究现状

可靠性半监督学习算法

- **基于集成框架的模型**
  维护多个半监督模型和（或）一个监督模型，对数据集上的预测效果进行综合分析 (Max-margin, Worst case min)

- **基于SSL与SL相权衡**
  直接约束与监督模型的差异性 $\|f_{SL}(X_L) - f_{SSL}(X_L, X_U)\|^2$

- **学习无标签数据的权重**
  学习数据权重或数据之间的相似度，降低无关或不利于分类的无标签数据的权重，使得模型更加鲁棒可靠

研究的挑战

现有算法的不足

- **基于集成框架的模型**
  最终模型效果与集成子模型的个数和多样性存在强烈依赖，也忽视了模型的实时性与简洁性

- **基于SSL与SL相权衡**
  直接约束与监督模型的差异性，而忽略监督模型本身训练误差，同时其约束力度也值得考究

- **基于自适应权重模型**
  缺乏深入探索数据固有的结构信息， 而单独依赖特定分类器的划分效果，不能反映真实的数据权重信息

不能处理数据流数据

可靠半监督学习

无标签数据的权重学习策略

ReSSL算法

- 度量聚类假设与数据的一致性
- 簇中的无标签数据共享同一个权重值

RP算法

- 可靠性传播算法
- 选择可靠的无标签数据
- 分布式扩展

离线模型

ReSSL Steam算法

- ReSSL算法的数据流扩展
- 在线动态地维护半监督微簇信息

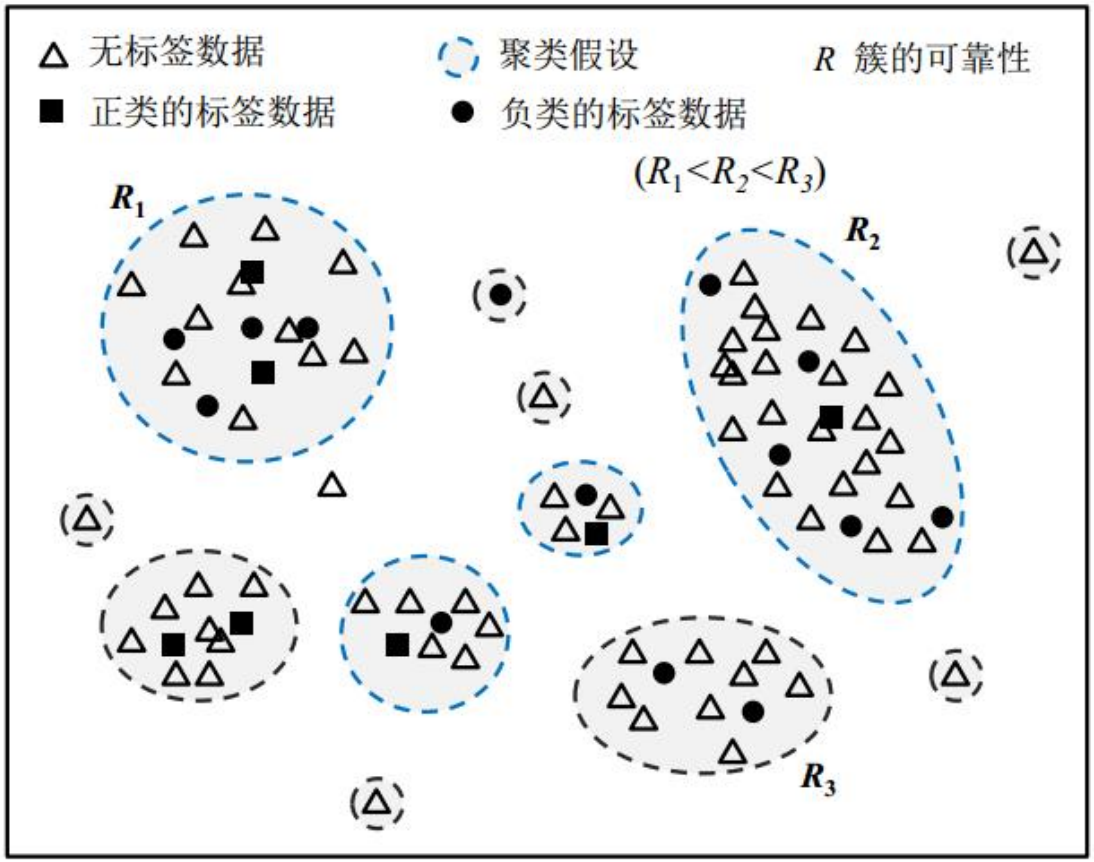
BLS算法

- 在线半监督最小二乘算法
- 松弛的模型假设，让模型更可靠性

在线模型

学习无标签数据权重

模型假设的不一致性



聚类假设：簇内的数据具有相同的标签

学习无标签数据权重

ReSSL算法

- **核心思路：**通过度量假设不一致性，学习每个簇的权重，间接筛选无标签数据
  - **簇规则性：**度量簇中标签的一致性

$$CR(C_i) = \frac{H - H(C_i)}{H}$$

$\boldsymbol{H(C_i)}$**:** 簇C_i 标签分布的熵
$\boldsymbol{H}$: 训练集中标签分布的熵

设 $\boldsymbol{H(C_1) = H}$
$\boldsymbol{CR(C_1) = 0}$

$\boldsymbol{CR(C_1) < 1}$

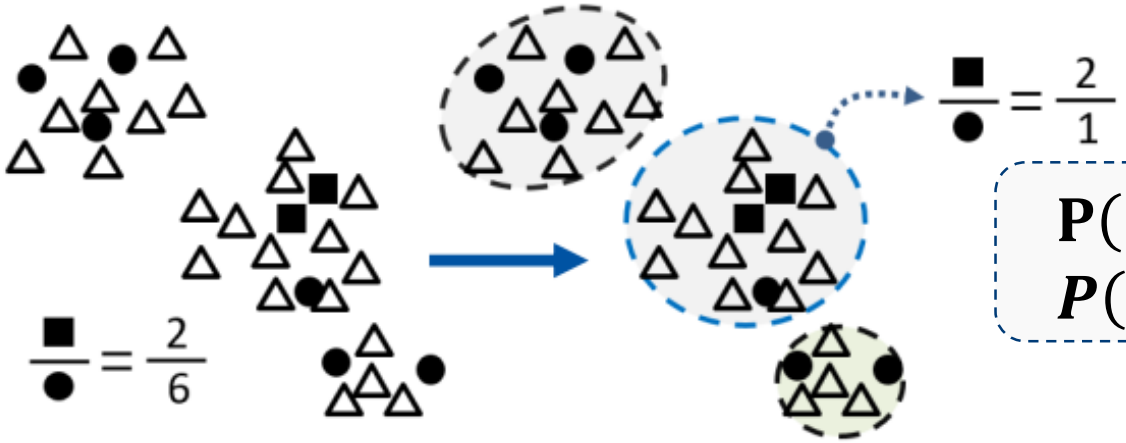
$\boldsymbol{H(C_1) = 0}$
$\boldsymbol{CR(C_1) = 1}$

学习无标签数据权重

ReSSL算法

- **核心思路：**通过度量假设不一致性，学习每个簇的权重，间接筛选无标签数据
  - **簇规则性：**度量簇中标签的一致性
  - **簇优先级：**考虑标签的不平衡问题

$$CP(C_i) = sigmoid\left(\sum_{e=1}^{\kappa} \frac{P(C_i^e) - P(D^e)}{P(D^e)}\right)$$



$\mathbf{P}(\boldsymbol{C_i^e})$: 簇$C_i$中标签$e$占的比例
$\boldsymbol{P}(\boldsymbol{D^e})$: 训练集中标签$e$的比例

学习无标签数据权重

ReSSL算法

- **核心思路：**通过度量假设不一致性，学习每个簇的权重，间接筛选无标签数据
  - **簇规则性：**度量簇中标签的一致性

$$CR(C_i) = \frac{H - H(C_i)}{H}$$

  - **簇优先级：**考虑标签的不平衡问题

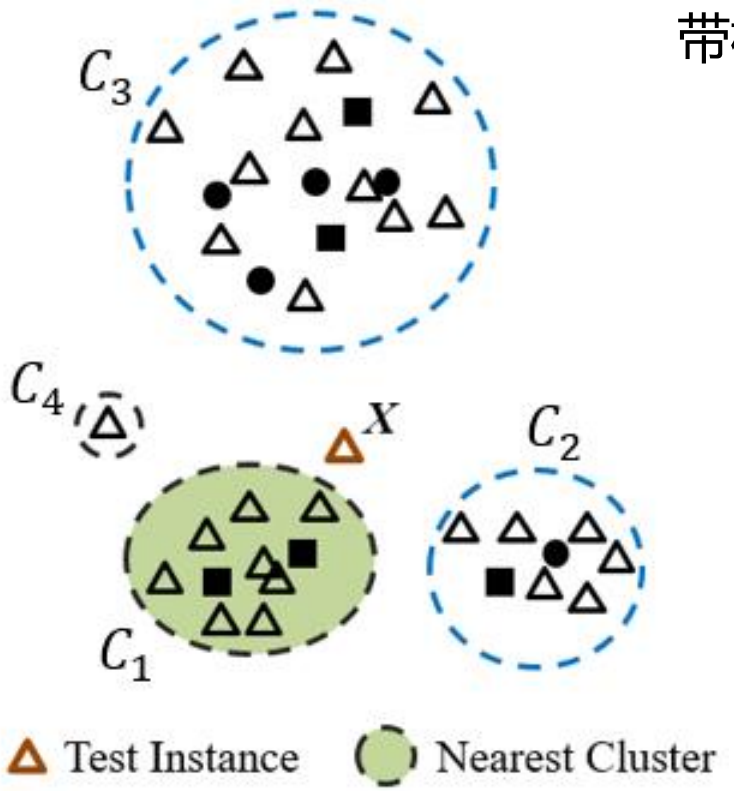
$$CP(C_i) = sigmoid\left(\sum_{e=1}^{\kappa} \frac{P(C_i^e) - P(D^e)}{P(D^e)}\right)$$

可靠性度量： $\boldsymbol{R(C_i) = CR(C_i) * CP(C_i)}$

学习无标签数据权重

ReSSL预测

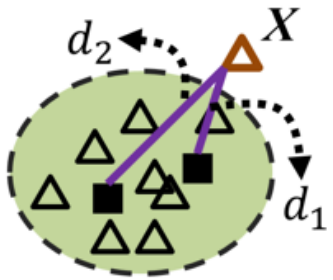
- **如果测试数据的最近簇是可靠的**



带权投票：

$$\mathrm{f}(x) = \max_{l} \frac{P\left(C_1^l\right)}{D^{cs}\left(x, C_1^l\right)}$$

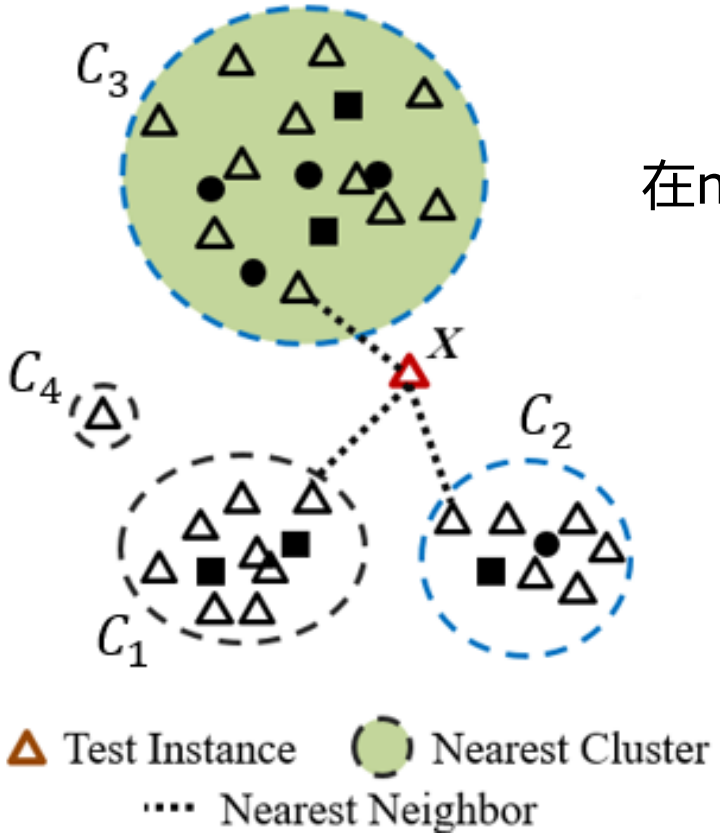
$\boldsymbol{D^{cs}\left(x, C^l\right)}$：点$x$到具有标签$l$的标签数据$C$的平均距离。



学习无标签数据权重

ReSSL预测

– **如果测试数据的最近簇是不可靠的**



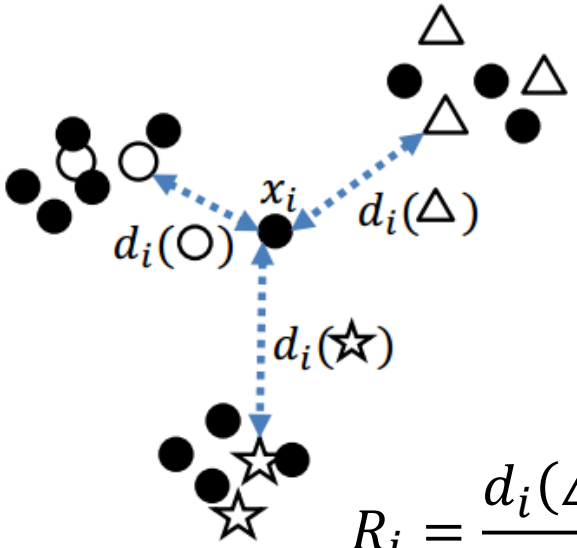
在m个最近簇之间的带权投票：

$$\mathrm{f}(x) = \max_{l} \sum_{i=1}^{m} \frac{R\left(C_i^l\right) P\left(C_i^l\right)}{D^{cs}\left(x, C_i^l\right)}$$

学习无标签数据权重

RP算法

- **核心思路：**邻域标签分布不规整的无标签数据不可信；相似的数据点具有相似的安全性
  - **可靠性先验R：**邻域标签分布规则，数据可靠性高



$$R_i = \frac{d_i(\triangle) - d_i(\bigcirc)}{d_i(\triangle)}$$

$$\mathrm{D}_i = \left(d_i(\bigcirc), d_i(\triangle), d_i(\star)\right)$$

$$R_i = \frac{D_i(k_2) - D_i(k_1)}{D_i(k_2)}$$

$$d_i(k) = \min_{y_j = k} \Phi\left(x_i, x_j\right)$$

学习无标签数据权重

RP算法

- **核心思路：**邻域标签分布不规整的无标签数据不可信；相似的数据点具有相似的安全性
  - **可靠性先验R：**邻域标签分布规则，数据可靠性高
  - **相似性刻画：**拉普拉斯矩阵L

$$\sum_{ij}^{N} W_{ij}\left(C_i - C_j\right)^2 = C^T L C$$

学习无标签数据权重

RP算法

- **核心思路：**邻域标签分布不规整的无标签数据不可信；相似的数据点具有相似的安全性
  - **可靠性先验R：**邻域标签分布规则，数据可靠性高
  - **相似性刻画：**拉普拉斯矩阵L
  - **相似度度量**：边(结构)相似度 & 点(可靠性)相似度

- **可靠性传播算法(Reliability Propagation)**

$$\min_{C_U,N} C^T L_N C + \lambda_1 \|N - W\|_F^2 + \lambda_2 \|C - R\|^2$$
$$s.t. C_L = 1, 0 \le C_U \le 1,$$
$$N_{ij} \ge 0, N = N^T$$

学习无标签数据权重

RP算法

- **核心思路：**邻域标签分布不规整的无标签数据不可信；相似的数据点具有相似的安全性
  - 可靠性先验R：邻域标签分布规则，数据可靠性高
  - 相似性刻画：拉普拉斯矩阵L
  - 相似度度量：边(结构)相似度 & 点(可靠性)相似度
- **可靠性传播算法(Reliability Propagation)**
  - 投影梯度下降

$$\nabla = \mathrm{L}_{LU}^{T} C_L + L_{UU} C_U + \lambda_2 (C_U - R_U)$$
$$C_U \leftarrow \min(1, \max(C_U - \alpha \nabla))$$

  - 开口向上的有界二次函数求解

$$\mathrm{N}_{ij} = \max\left(0, W_{ij} - \frac{(C_i - C_j)^2}{2\lambda_1}\right)$$

学习无标签数据权重

Distributed RP算法

- **核心思路：**基于图的并行计算，每次迭代更新点权C与边权N. *(Bulk Synchronous Parallel (BSP), Spark GraphX)*
  - **先验R的分布式计算：**
    只需标签数据向邻域的无标签数据发送消息(y_i, x_i)
  - **分布式传播算法：**
    - 消息传递

$$m_{ij} = \begin{cases} \left(0, N_{ij}^{(k-1)}, 0\right) & if\ \ x_i\ is\ labeled \\ \left(C_i^{(k-1)} N_{ij}^{(k-1)}, 0, N_{ij}^{(k-1)}\right) & otherwise \end{cases}$$

    - 消息整合计算

$$\nabla_i = M_i(3)C_i - M_i(1) + M_i(2) + \lambda_2(C_i - R_i)$$

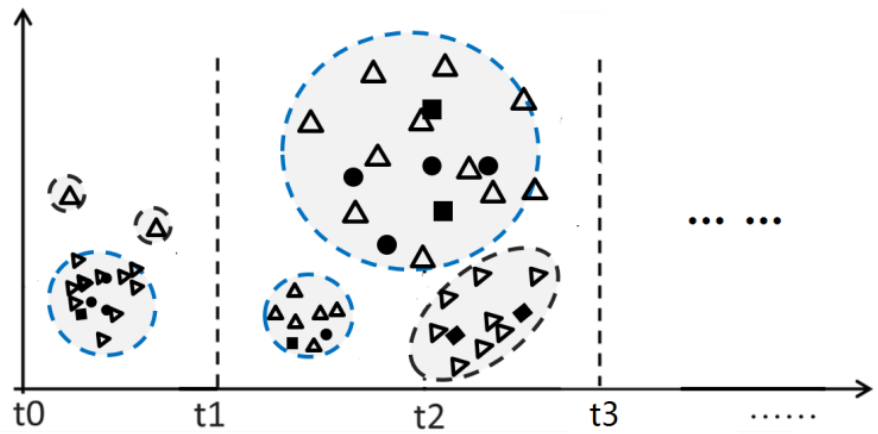
$$M_i = \sum_k m_{ki}$$

在线半监督学习

ReSSL Stream算法

- **核心思想：**在线动态维护微簇
  - **半监督微簇：**记录一个簇的统计信息

$$MC^S = \{LS, SS, N_l, N_u, LD, LC, R\}$$



$\boldsymbol{LS}$: 簇中数据特征的和

$\boldsymbol{SS}$:簇中数据特征的平方和

$\mathbf{N_l}, \mathbf{N_u}$: 簇中标签与无标签数据个数

$\boldsymbol{LD}, \boldsymbol{LC}$：每个类别下，标签数据的个数和标签数据的均值点

$\boldsymbol{R}$: 可靠度量值

在线半监督学习

ReSSL Stream算法

- **核心思想：**在线动态维护微簇
  - **半监督微簇：**记录一个簇的统计信息

$$MC^S = \{LS, SS, N_l, N_u, LD, LC, R\}$$

  - **改进的Den-stream维护微簇策略：**
    - 簇自适应增长
    - 簇的重要性随时间衰减
    - 考虑噪声数据
  - **ReSSL Stream预测：**
    - 距离=测试数据到各类标签数据的中心点

在线半监督学习

BLS算法

- **核心思想：**半监督最小二乘，不施加额外模型假设（如聚类假设，流型假设等等）。通过在线核学习（Online Kernel Learning）扩展到数据流环境。
  - **核半监督最小二乘+表征定理：** $\mathrm{f}(\mathrm{x}) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$

$$\min_{\alpha,z} \left\| \begin{bmatrix} y \\ z \end{bmatrix} - K\alpha \right\|_2^2 + \lambda_1 \alpha^T K \alpha + \lambda_2 \|z\|_2^2$$

  - **新颖的闭式解：**

$$\hat{\mathrm{K}} = I - K(\lambda_1 I + K)^{-1}$$

$$z = -\left(\hat{K}_{uu} + \lambda_2 I\right)^{-1} \hat{K}_{ul} y$$

$$\alpha = (\lambda_1 I + K)^{-1} \begin{bmatrix} y \\ -\left(\hat{K}_{uu} + \lambda_2 I\right)^{-1} \hat{K}_{ul} y \end{bmatrix}$$

在线半监督学习

BLS算法

- **核心思想：**半监督最小二乘模型，不施加额外模型假设（如聚类假设，流型假设等等）。通过在线核学习（Online Kernel Learning）扩展到数据流环境。
  - **核半监督最小二乘+表征定理：**
  - **新颖的闭式解：**

$$z = -\left(\widehat{K}_{uu} + \lambda_2 I\right)^{-1} \widehat{K}_{ul} y$$

标签传播算法： $\min_{\ell} \ell^T L \ell \quad s.t. \ell_i = y_i \ \forall i \in labeled$

Harmonic Solution: $z = \ell_u = -(L_{uu})^{-1} L_{ul} y$

Regularized Harmonic Solution: $z = \ell_u = -(L_{uu} + \lambda I)^{-1} L_{ul} y$

在线半监督学习

BLS算法

- **核心思想：**半监督最小二乘模型，不施加额外模型假设（如聚类假设，流型假设等等）。通过在线核学习（Online Kernel Learning）扩展到数据流环境。
  - **核半监督最小二乘+表征定理：**
  - **新颖的闭式解：**

$$\alpha = (\lambda_1 I + K)^{-1} \begin{bmatrix} y \\ -\left(\widehat{K}_{uu} + \lambda_2 I\right)^{-1} \widehat{K}_{ul} y \end{bmatrix}$$

允许采用基于窗口的在线核学习（Budgeted Kernel Learning）的策略做在线更新

在线半监督学习

BLS算法

- **在线更新（Two Budgets Update）**：限制模型大小的不断增长

$$\mathrm{f(x)} = \sum_{i=1}^{B} \alpha_i k(x_i, x) = \sum_{i=1}^{B_L} \alpha_i k(x_i, x) + \sum_{i=1}^{B_U} \alpha_i k(x_i, x)$$

  - **移除时间最老的数据**

    在窗口中的数据个数超过阈值B_L或B_U的时候被触发。移除对于窗口中时间最老的数据，加入新数据到对应窗口中

  - **基于投影的融合策略**

    数据被投影到前一时刻窗口数据所张成的子空间中

$$f'' = P_b f' = \sum_{i=1}^{b} \alpha'_i k(x_i, \cdot) + \alpha'_{b+1} P_b k(x_{b+1}, \cdot)$$

在线半监督学习

BLS算法

- **增量式/快速求逆** (Chap. 5.2.4)
  - $K_{b+1}^{-1} = f(K_b^{-1})$
  - $K_{b-1}^{-1} = g(K_b^{-1})$

- **Regret Bound** (Chap. 5.2.5)

$$R(T) \leq \frac{\left((2\lambda_1 + 1)U - 2\right)^2}{2}\bar{\eta} + 2U\|\bar{E}\|$$

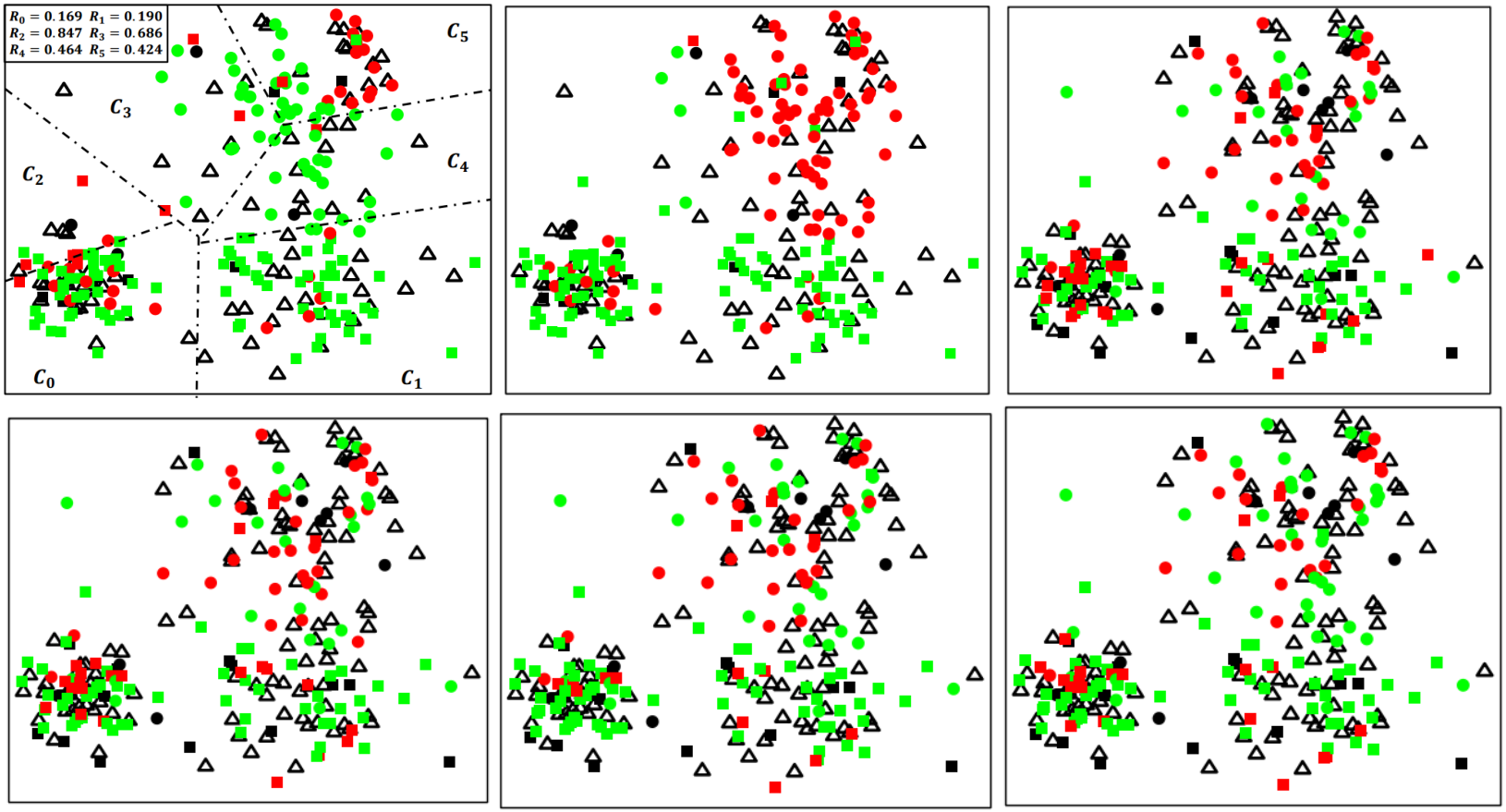
$E_t = \frac{\Delta_t}{\eta_t}$表示为梯度的损失，并且假设$\|E_t\| \leq 1$

$\bar{\mathrm{E}} = \frac{1}{T}\sum_{t=1}^{T}\|E_t\|$表示在所有观测值上的平均梯度损失

$\bar{\eta} = \frac{1}{T}\sum_{t=1}^{T}\|\eta_t\|$表示平均步长。

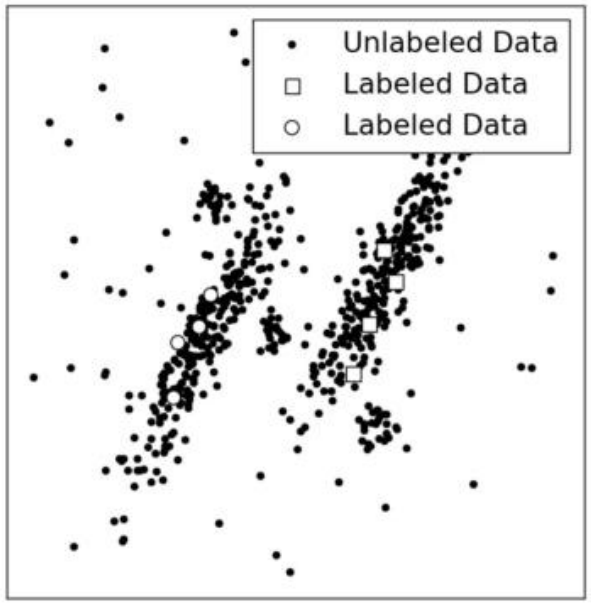
实验结果

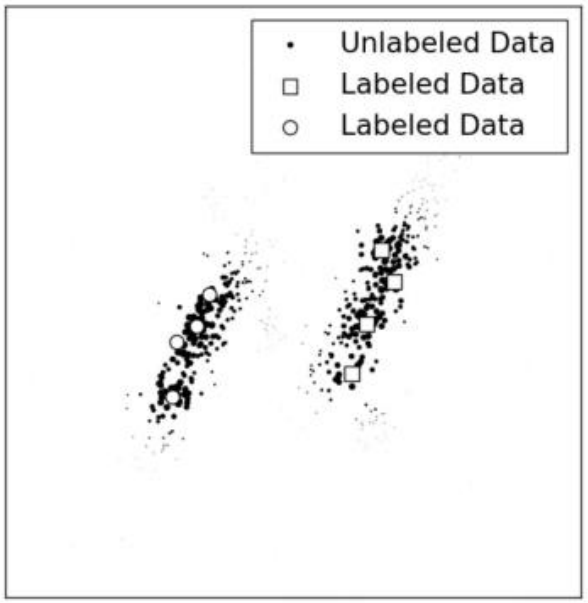
ReSSL算法 部分结果

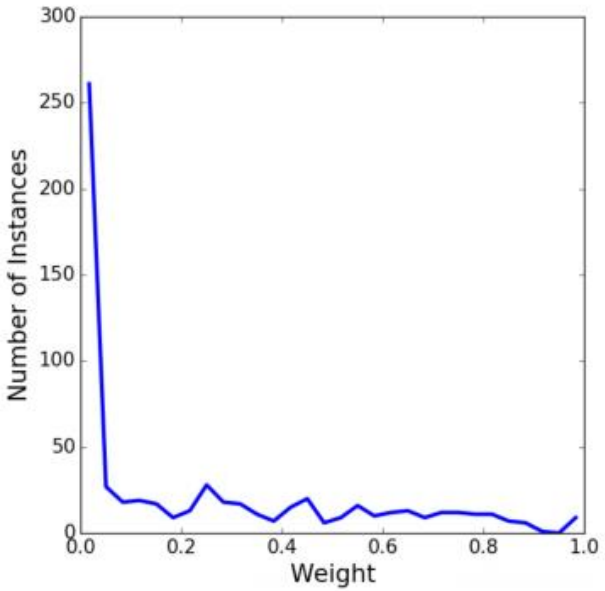


实验结果

RP算法 部分结果

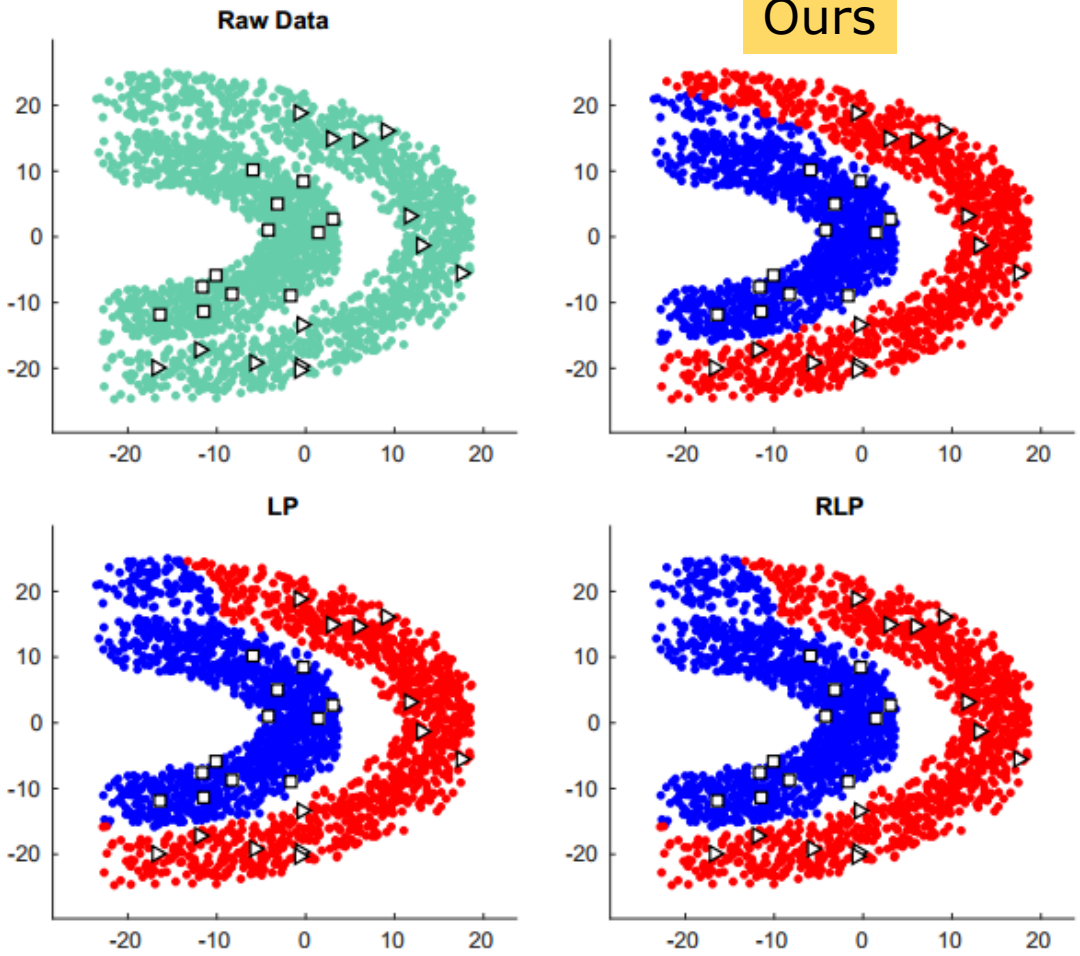






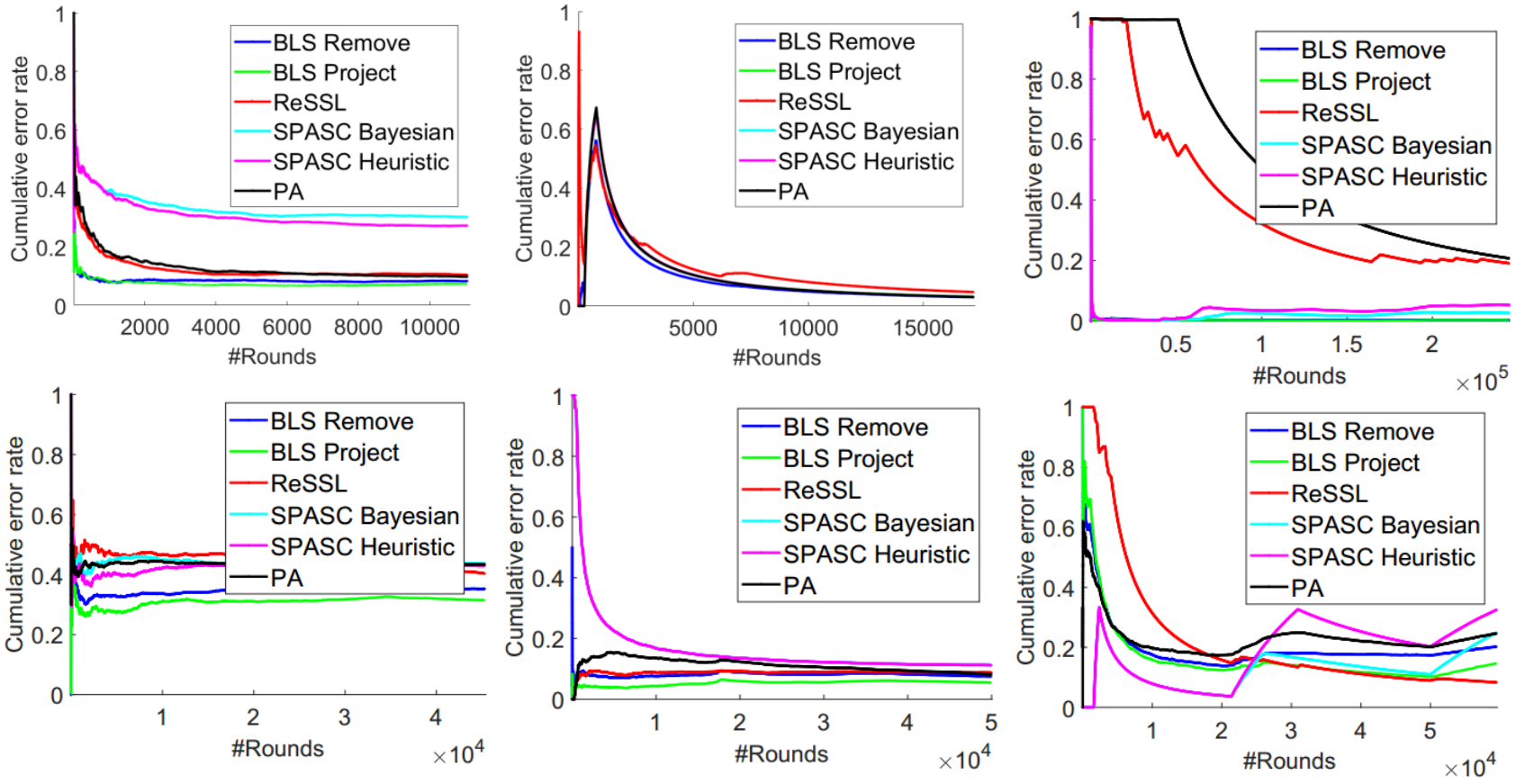
实验结果

BLS算法 部分结果



实验结果

BLS算法 部分结果



总结展望

不足与后续研究

- ReSSL + 结构信息
- BLS算法的时间复杂度
- 可靠性的理论分析
- 可靠半监督回归，聚类的扩展
- 可解释性

感谢您聆听，请指正！