清華大學
Tsinghua University

以数据为中心的机器学习可视分析方法

杨维铠

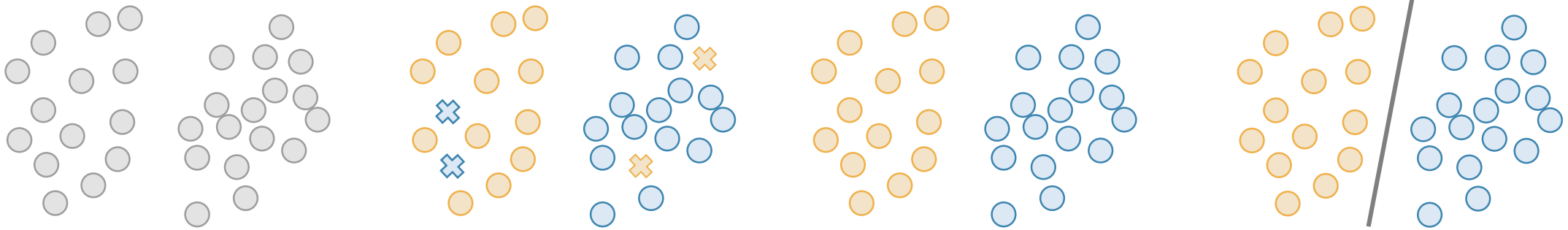
2024. 5. 16

研究背景

机器学习模型构建流程

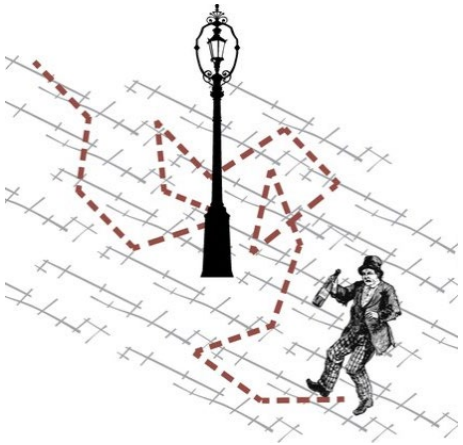
迭代改进

收集数据 → 标注数据 → 清洗数据 → 训练并部署模型



做出调整—训练模型—检查结果

可视分析



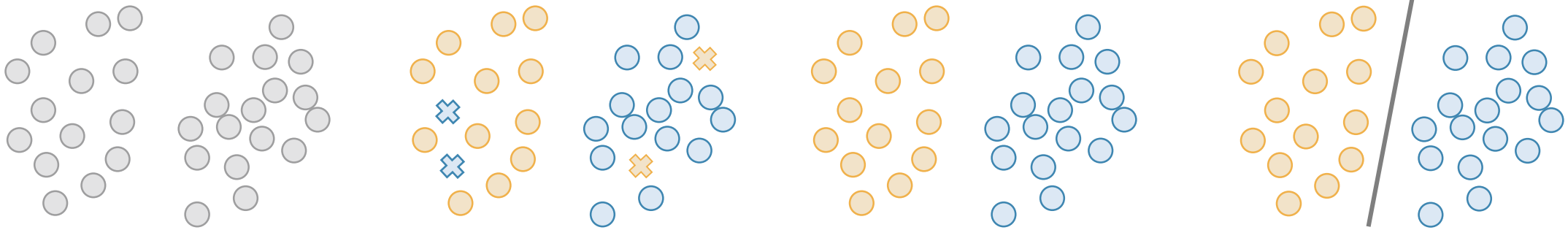
试错过程繁

分析门槛高

机器学习模型构建流程

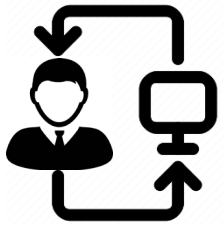
迭代改进

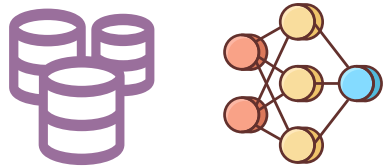
收集数据 → 标注数据 → 清洗数据 → 训练并部署模型

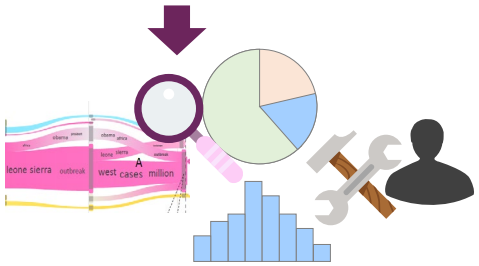


做出调整—训练模型—检查结果

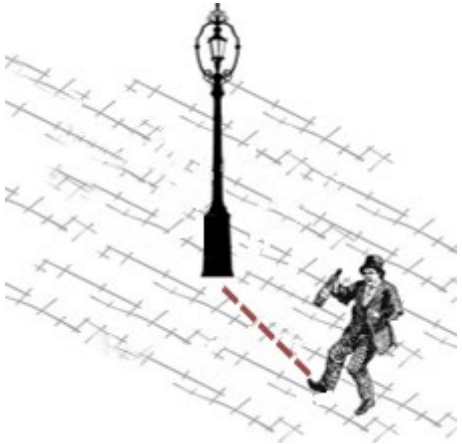
可视分析



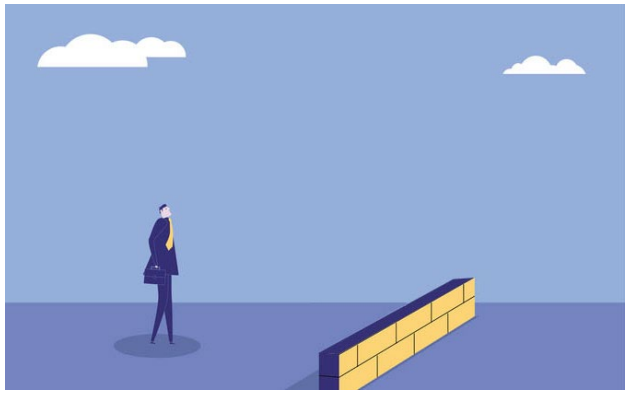




1. 探索数据和模型
2. 分析以理解性能瓶颈
3. 做出有**针对性**的调整



减少试错次数

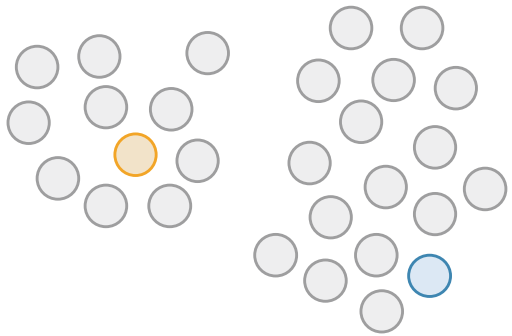


降低分析门槛

核心挑战

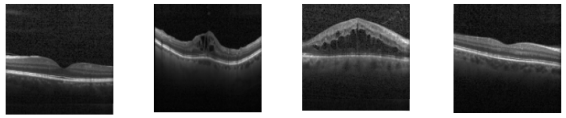
收集数据 → 标注数据 → 清洗数据 → 训练并部署模型

标注数量少



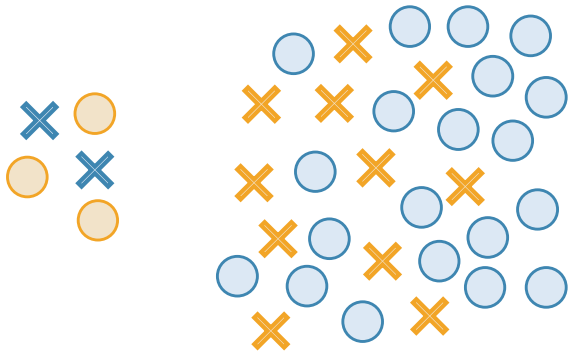
无标注训练数据

有标注训练数据



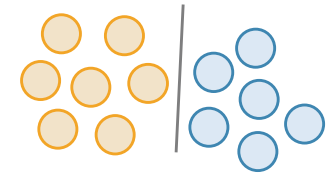
样本偏差大

标注错误多　　类别不平衡

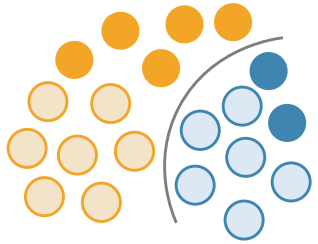


错误标注训练数据

分布变化大



训练数据所得模型的决策边界



加入新数据后模型的决策边界

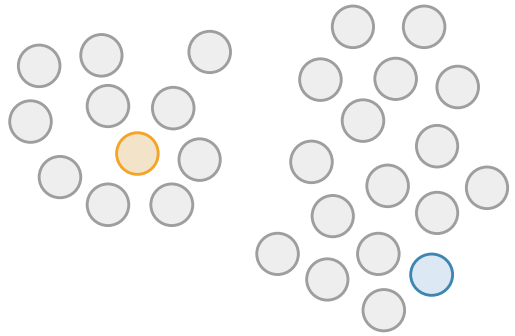
新数据

	每日天气记录				
训练数据	日期	最高温	气压	降雨	正常天气
	22/8/2	36.5℃	101kpa	否	
	22/8/3	32.1℃	102kpa	是	
	⋮	⋮	⋮	⋮	
新数据	23/8/2	38.7℃	100kpa	否	极端天气
	23/8/3	40.1℃	97kpa	否	
	23/8/4	40.6℃	96kpa	否	

研究内容

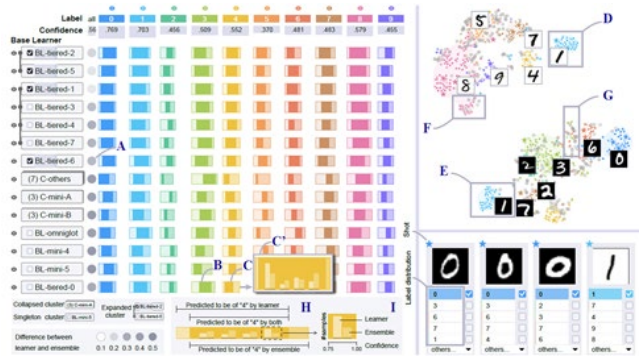
收集数据 → 标注数据 → 清洗数据 → 训练并部署模型

标注数量少

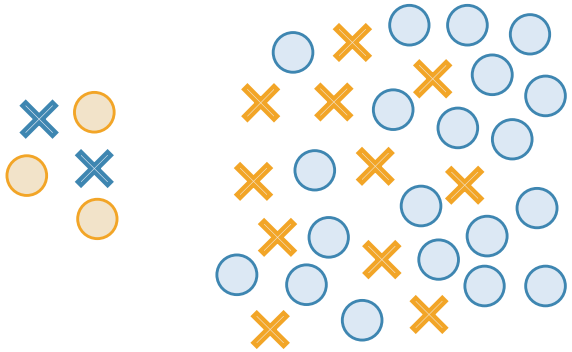


小样本学习**样本选择**方法

[IEEE TVCG 2022]

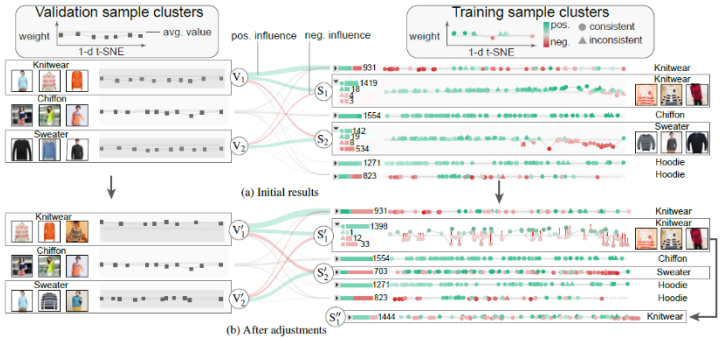


样本偏差大

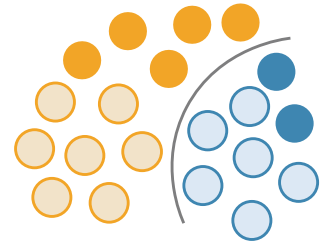


有偏差数据**样本加权**方法

[IEEE TVCG 2024]

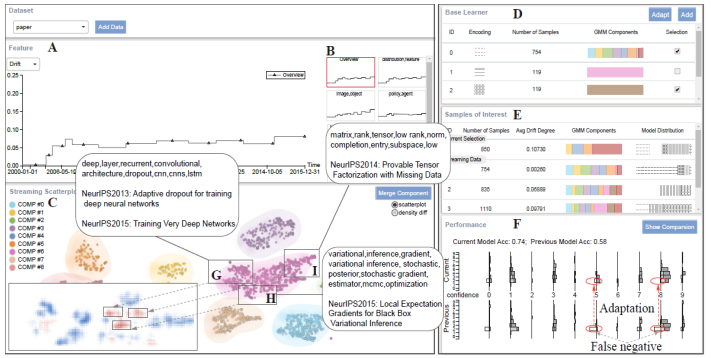


分布变化大

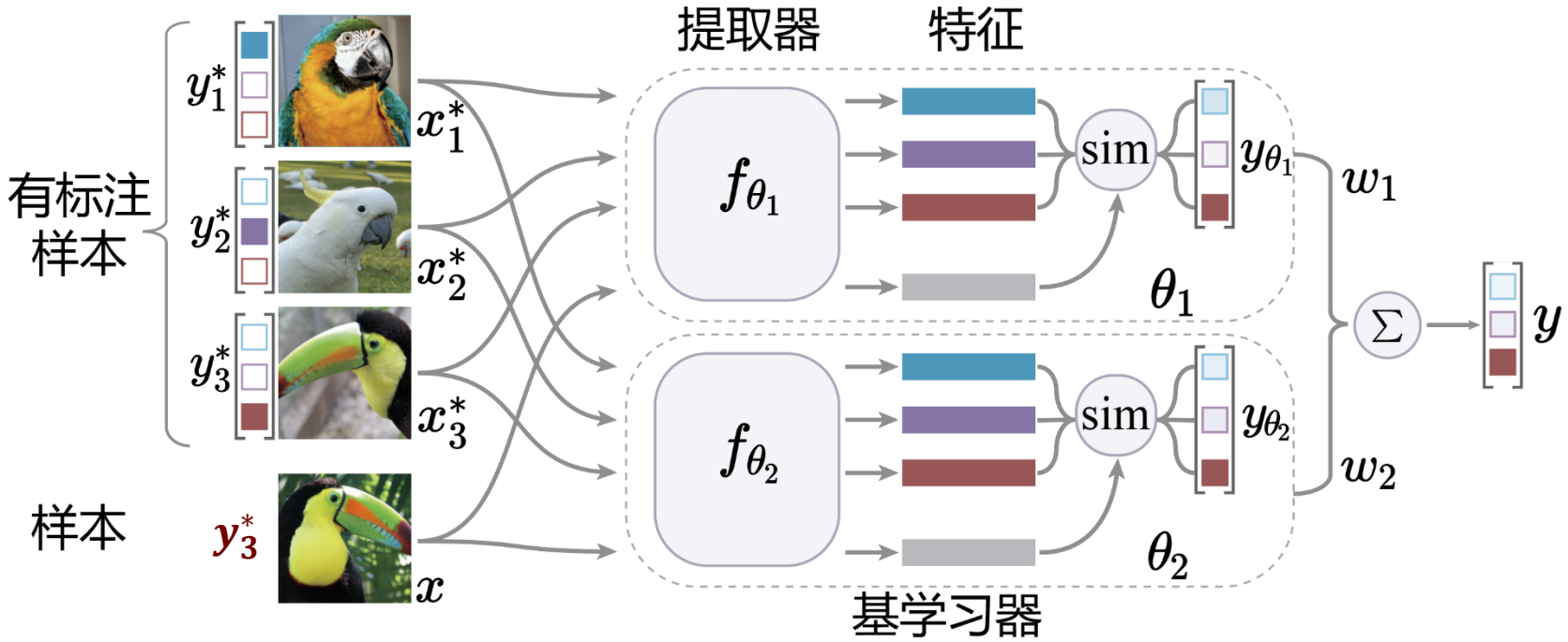


动态环境下**样本更新**方法

[IEEE VIS 2020]



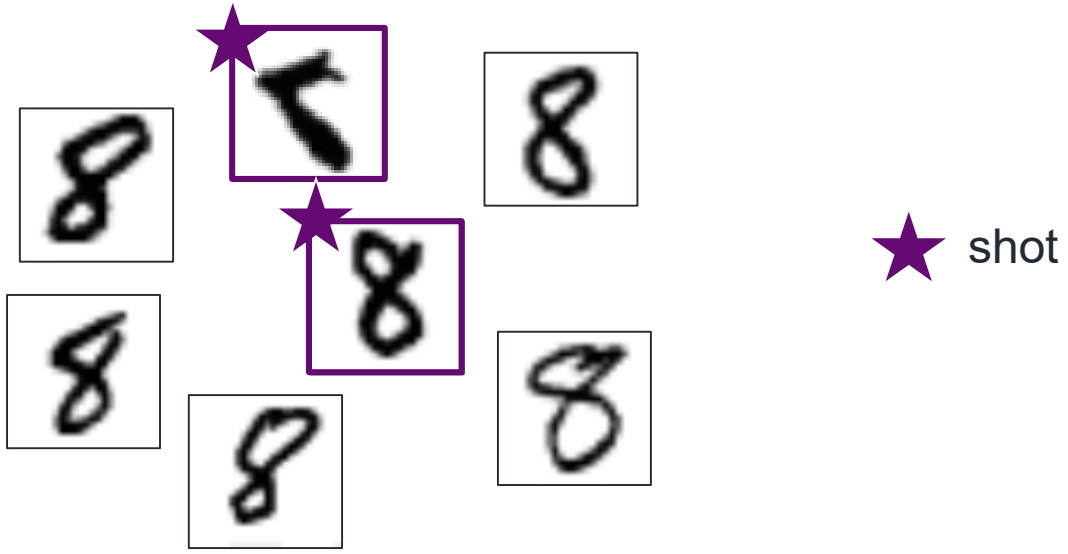
小样本学习可视分析方法



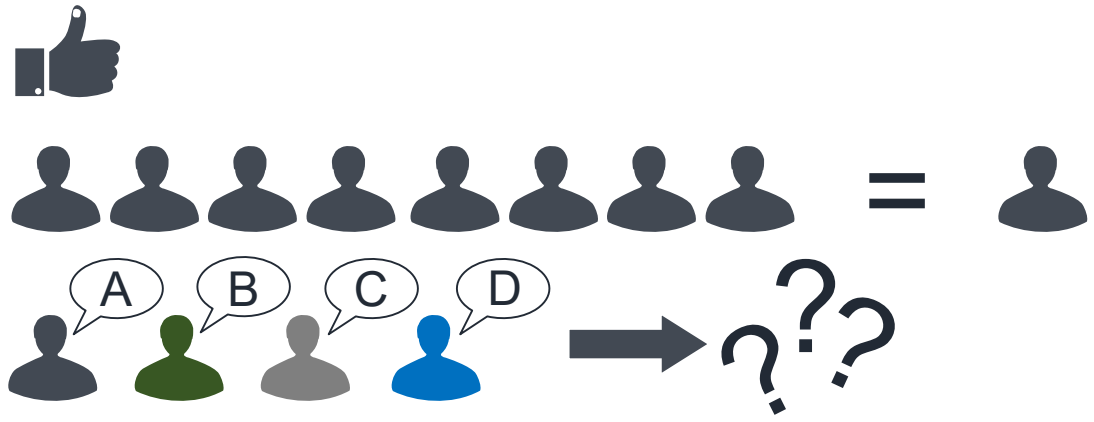
- **有标注样本**和**基学习器**的质量极大影响小样本学习的性能

技术挑战

- 有标注样本选择
  - 去除低质量的有标注样本
  - 加入必要的有标注样本



- 基学习器选择
  - 高性能
  - 多样性
  - 合作性

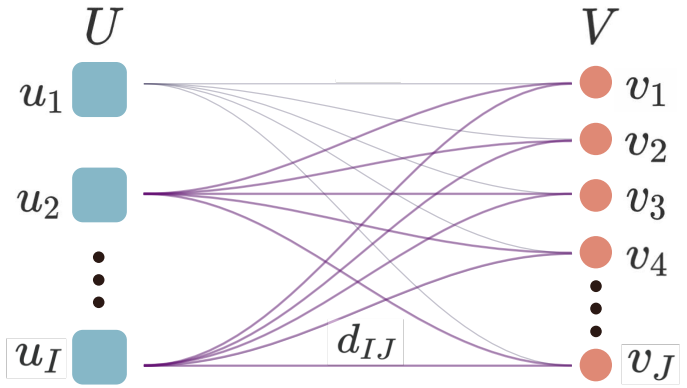


解决方案

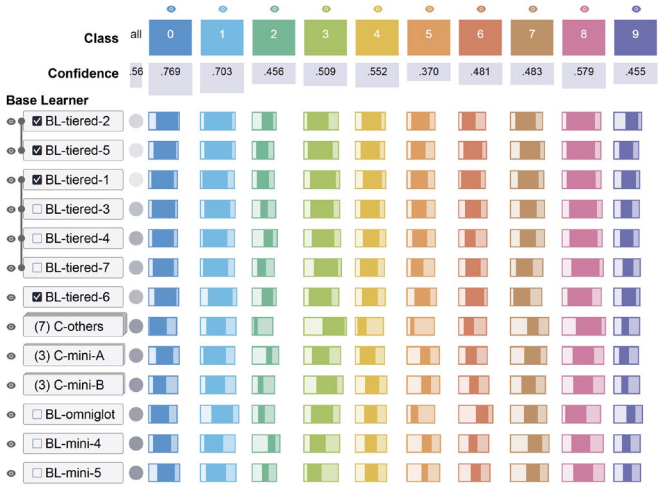
- 有标注样本选择
  - 去除低质量的有标注样本
  - 加入必要的有标注样本

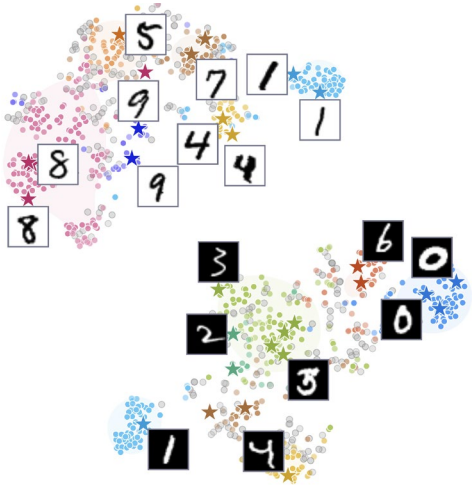
- 基学习器选择
  - 高性能
  - 多样性
  - 合作性

稀疏子集选择问题



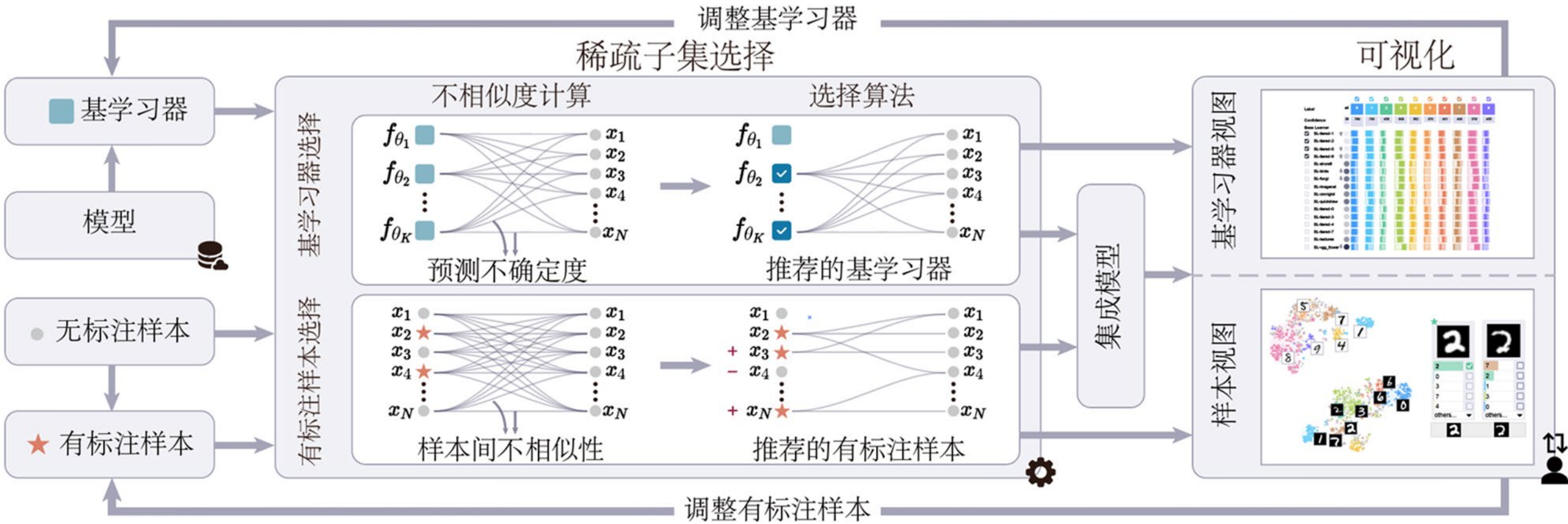
可视化



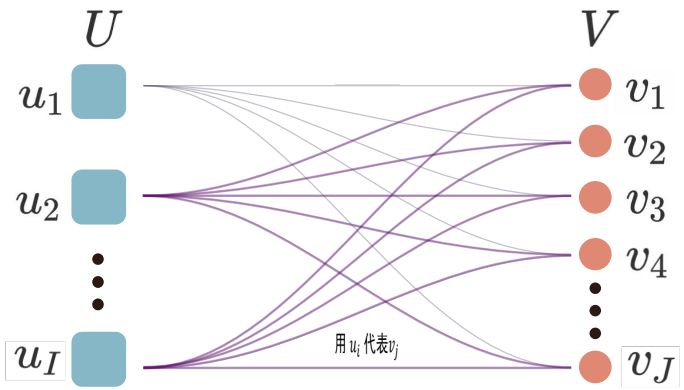


样本选择 → 样本加权 → 样本更新

系统概览


调整基学习器
稀疏子集选择
可视化
基学习器
模型
无标注样本
有标注样本
基学习器选择
有标注样本选择
不相似度计算
选择算法
预测不确定度
推荐的基学习器
样本间不相似性
推荐的有标注样本
集成模型
基学习器视图
样本视图
调整有标注样本

稀疏子集选择



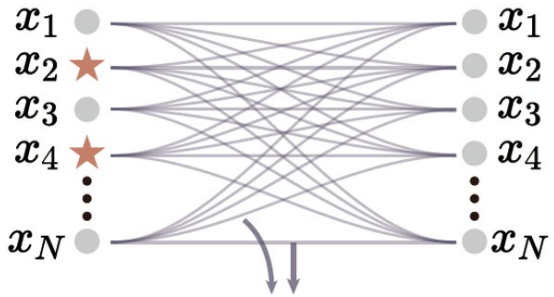
$$\mathcal{L}(A) = \sum_{j=1}^{J} \min_{u_i \in A} d_{ij} + \alpha|A|$$

代表度　　稀疏度

$$\mathcal{L}(Z) = \sum_{j=1}^{J} \sum_{i=1}^{I} z_{ij} d_{ij} + \alpha \sum_{i=1}^{I} \max_{j} z_{ij}$$

用 u_i 代表 v_j　　u_i 被使用

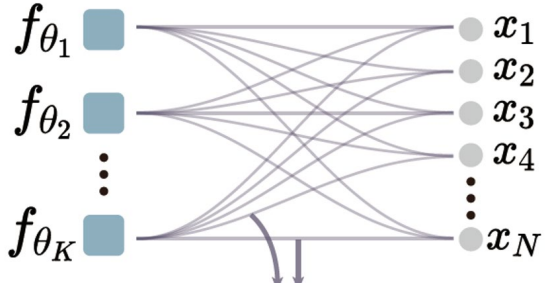
有标注样本选择



图像间不相似度

$$\sum_{j=1}^{N} \sum_{i=1}^{N} z_{ij} d_{ij} + \alpha \sum_{i=1}^{N} \beta_i \gamma_i \max_{j} z_{ij}$$

代表度　　信息量&标注代价



预测不确定度

$$\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ki} d_{ki} + \alpha_1 \sum_{k=1}^{K} \lambda_k \max_{i} z_{ki} + \alpha_2 \sum_{1 \le k < l \le K} \mu_{kl} \max_{i} z_{ki} \cdot \max_{i} z_{li}$$

多样性　　高性能　　合作性

样本选择　样本加权　样本更新

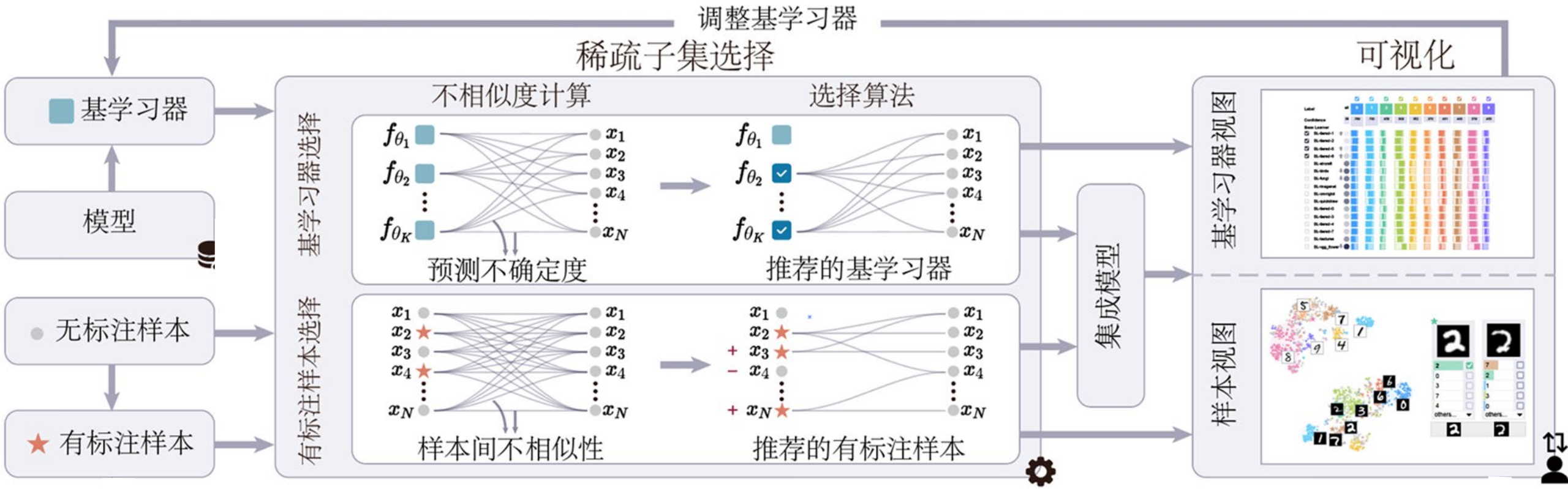
数值实验

- 在四个数据集上验证提出的稀疏子集选择方法的有效性

模型	*mini*	*tiered*	MNIST	CIFAR-FS
随机基学习器/样本	0.873	0.849	0.476	0.447
TIM	0.874	0.898	-	-
推荐样本	0.877	0.862	0.611	0.517
推荐基学习器	0.880 (3.9)	0.868 (3.6)	0.481 (4.3)	0.480 (5.2)
推荐基学习器/样本	**0.896** (3.9)	**0.908** (3.6)	**0.615** (4.3)	**0.541** (5.2)

- 推荐的基学习器和有标注样本都能改善模型性能
- 二者组合可以进一步提高模型性能

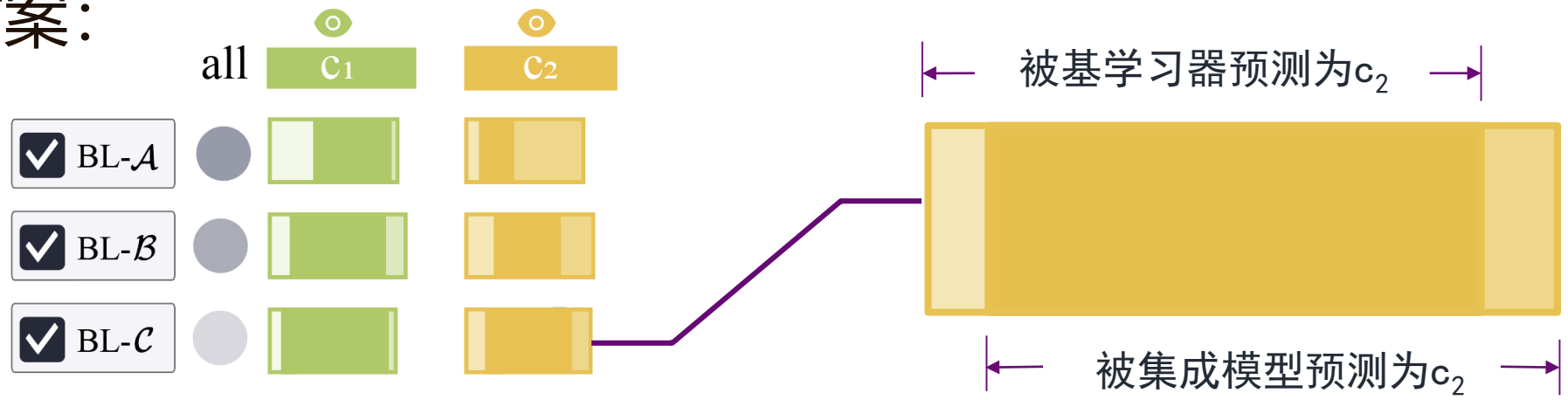
系统概览



样本选择 样本加权 样本更新

基学习器视图

- 挑战：没有真实类标
- 解决方案：

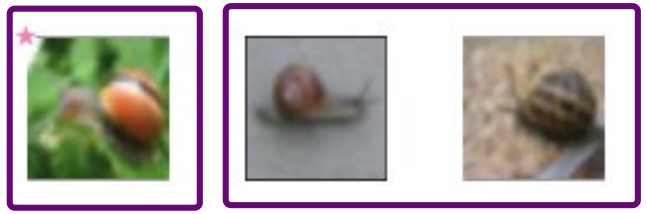


- 挑战：可扩展性
- 解决方案:

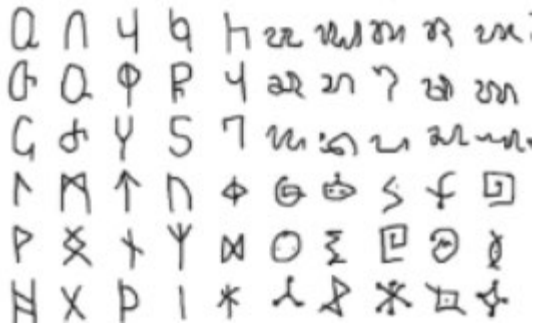
样本选择 样本加权 样本更新

案例分析

- CIFAR-100数据集
  - 使用测试集（20分类问题），初始60个有标注样本
  - 发现了有标注样本覆盖窄、质量低导致性能不佳
  - 加入37个有标注样本，分类准确率 47.4%->59.4%(**+12%**)

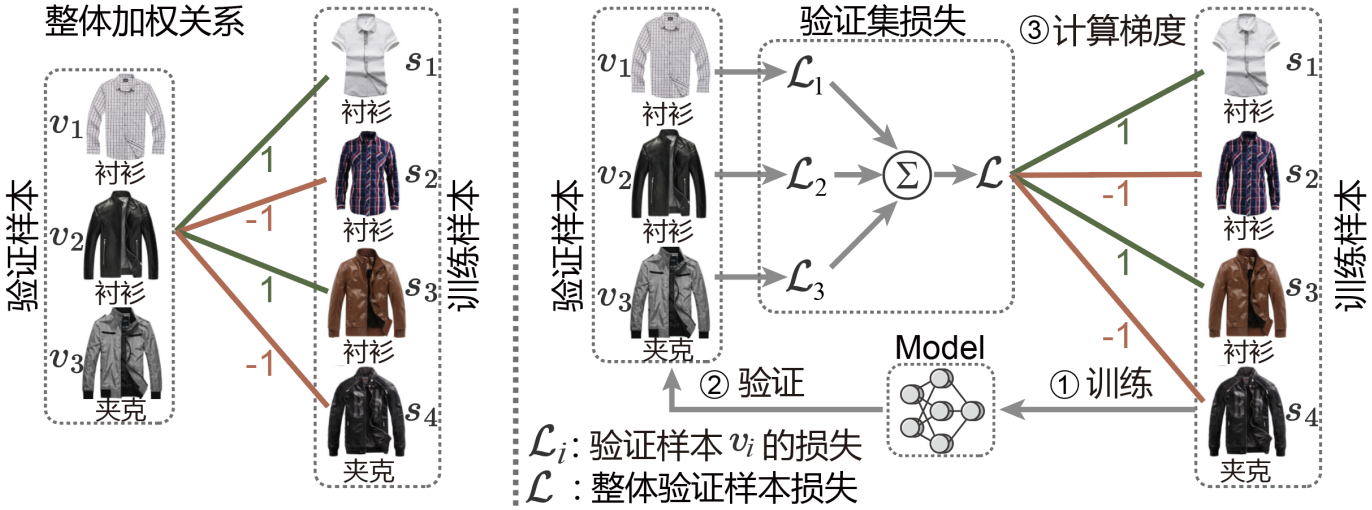


- MNIST手写数字数据集
  - 10分类问题，初始30个有标注样本
  - 发现基学习器组合多样性差，加入Omniglot数据集(多国字符) 上训练的基学习器
  - 也发现了有标注样本覆盖窄、质量低
  - 加入13个有标注样本，分类准确率 49.7%->70.7%(**+21%**)



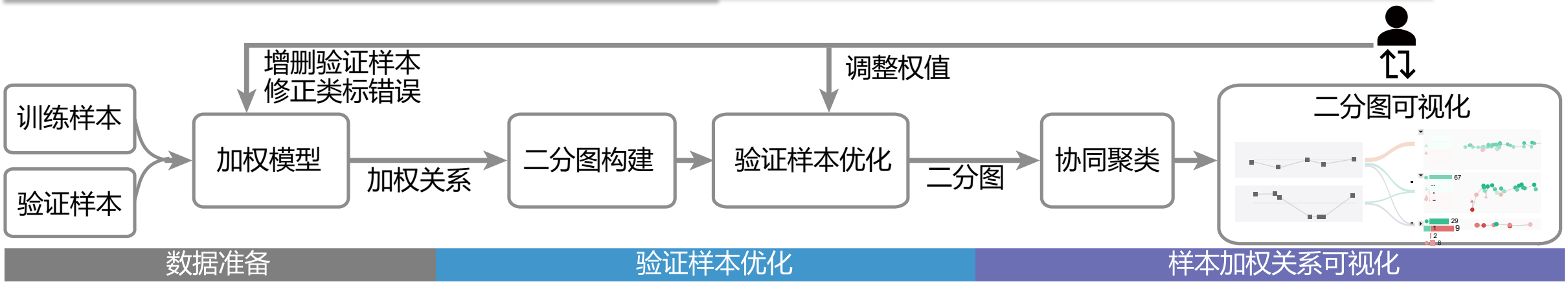
训练样本加权可视分析方法

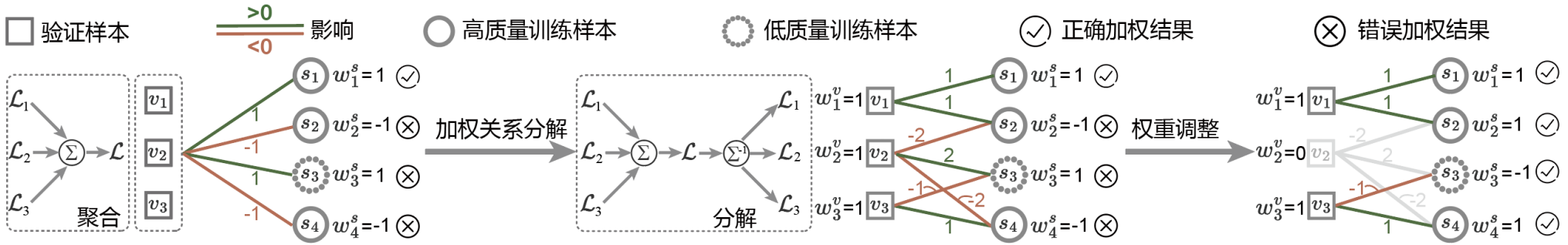
- 修正所有标注错误代价高昂
- 样本加权：降低错误类标样本的影响
- 利用少量验证样本，判断训练样本对模型训练是否有益
- 低质量验证样本给出错误的加权结果

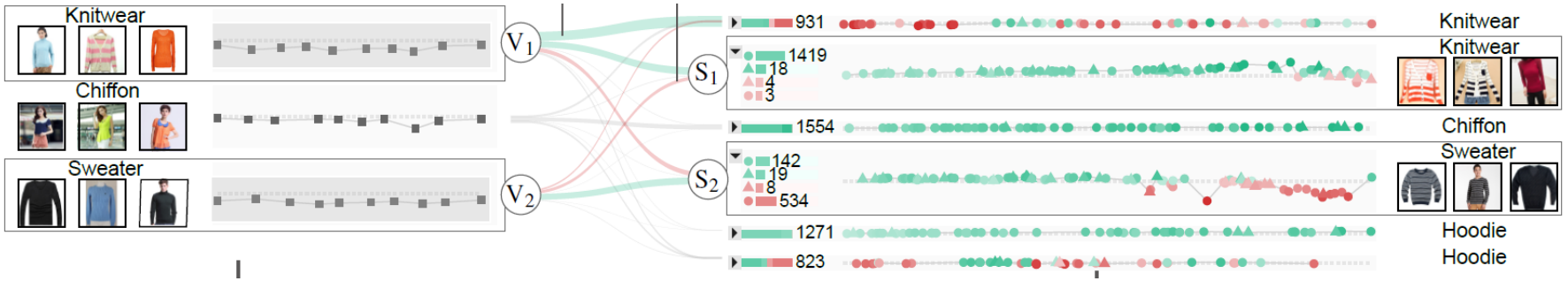


如何找到低质量的验证样本并进行调整？

系统概览



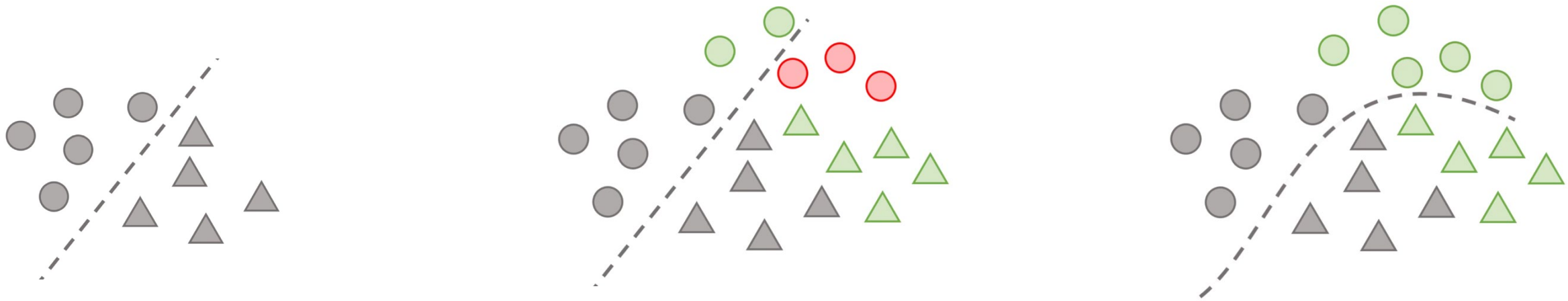




样本选择 | 样本加权 | 样本更新

概念漂移可视分析方法

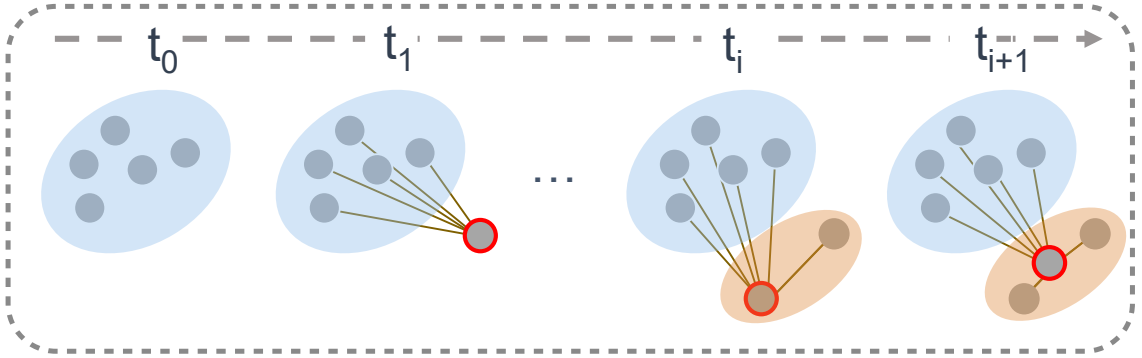
- 概念漂移：数据分布的变化造成数据驱动方法失效



- 需要及时检测、分析、应对概念漂移

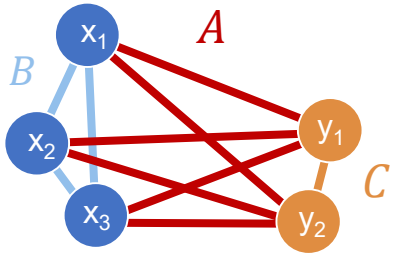
系统概览

概念漂移检测

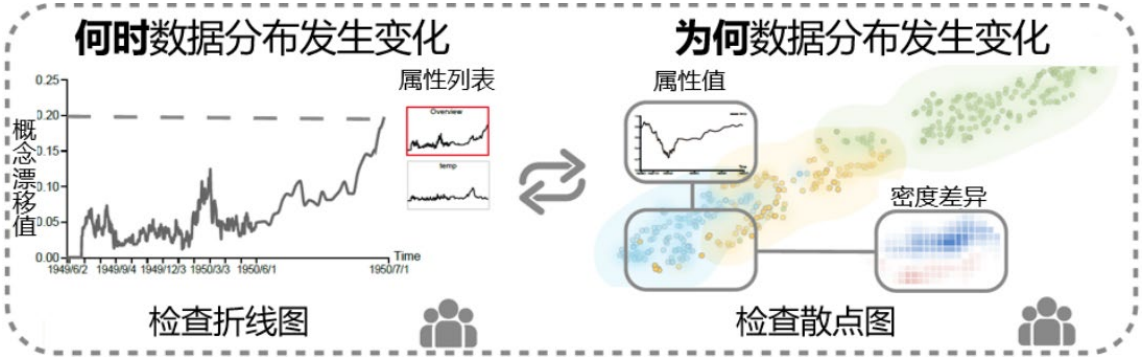


能量距离

$$d(X,Y) = \frac{2A - B - C}{2A}$$

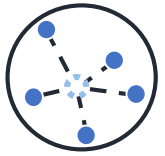


概念漂移分析



有约束 t-SNE

类别可读性



布局稳定性

概念漂移适应

集成模型

样本选择 样本加权 样本更新

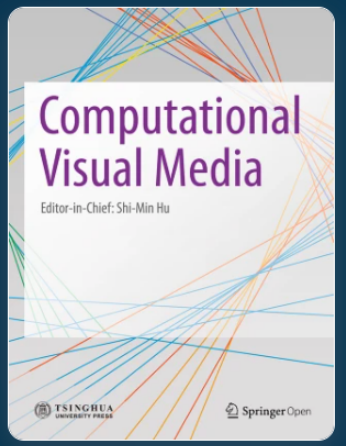
未来工作

Home › Computational Visual Media › Article

Foundation models meet visualizations: Challenges and opportunities

Review Article | Open access | Published: 02 May 2024

(2024) Cite this article



Computational Visual Media

Aims and scope →

Submit manuscript →

Download PDF

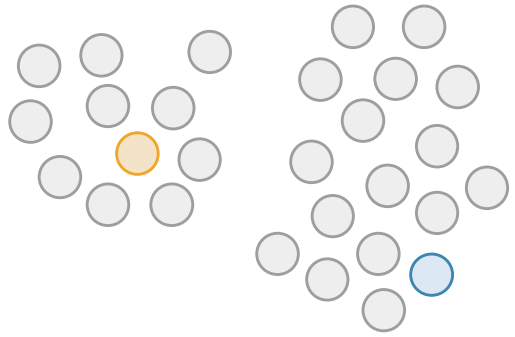
You have full access to this open access article

https://link.springer.com/content/pdf/10.1007/s41095-023-0393-x.pdf

以数据为中心的机器学习可视分析方法

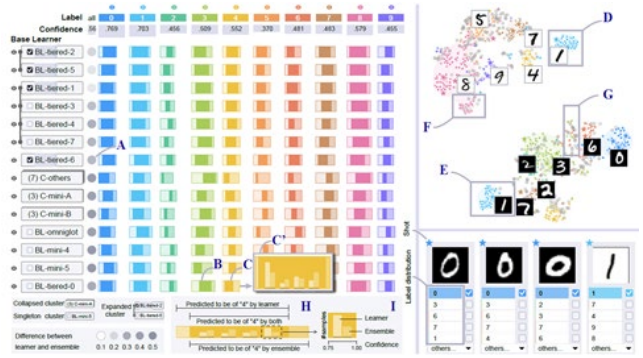
收集数据 → 标注数据 → 清洗数据 → 训练并部署模型

标注数量少

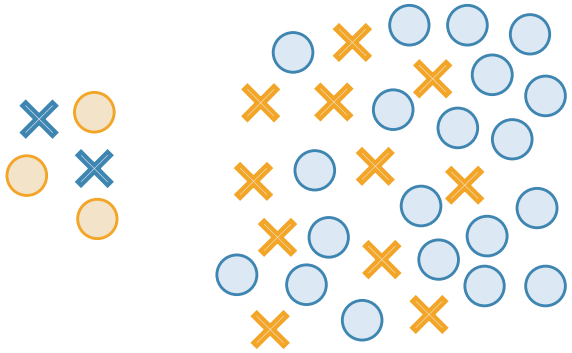


小样本学习**样本选择**方法

[IEEE TVCG 2022]

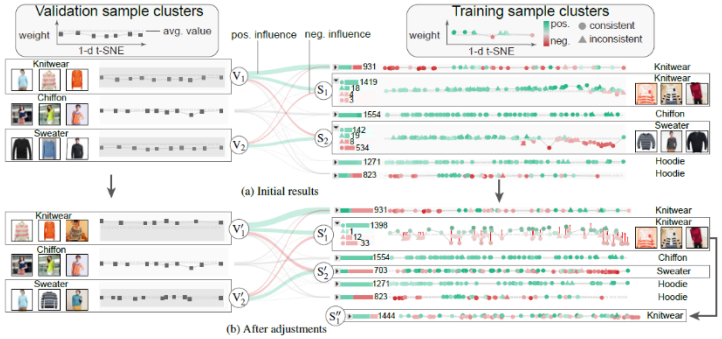


样本偏差大

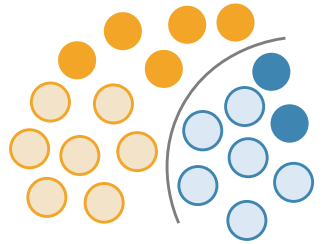


有偏差数据**样本加权**方法

[IEEE TVCG 2024]



分布变化大



动态环境下**样本更新**方法

[IEEE VIS 2020]