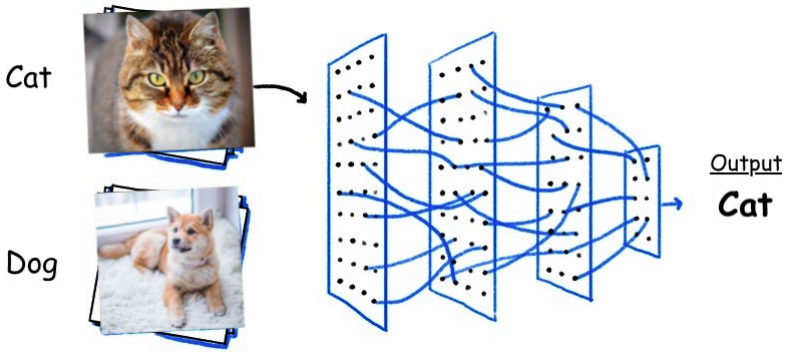
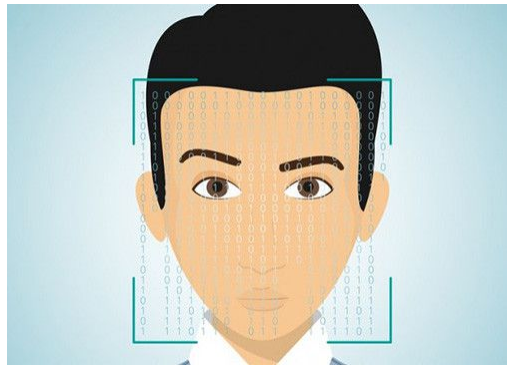
基于样本代表性的
模型性能可视分析及提升方法

清华大学软件学院 杨维铠
2023.11.30

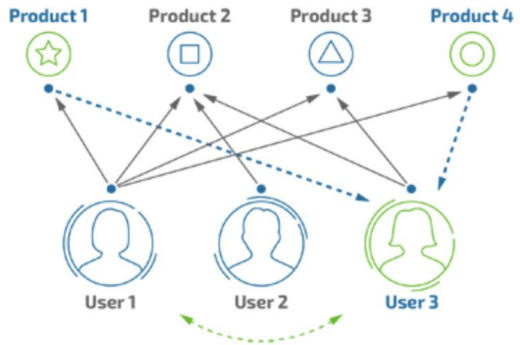
研究背景



图像分类



人脸识别



推荐系统

基于**样本代表性**的**可视分析**技术

数据准备 → 模型训练 → 模型部署

研究思路

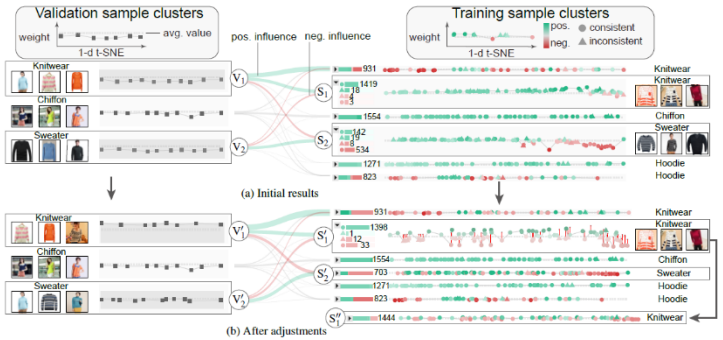
基于**样本代表性**的**可视分析**技术

数据准备 → 模型训练 → 模型部署

更好地代表真实任务　　更好地代表数据分布　　更好地代表当前环境

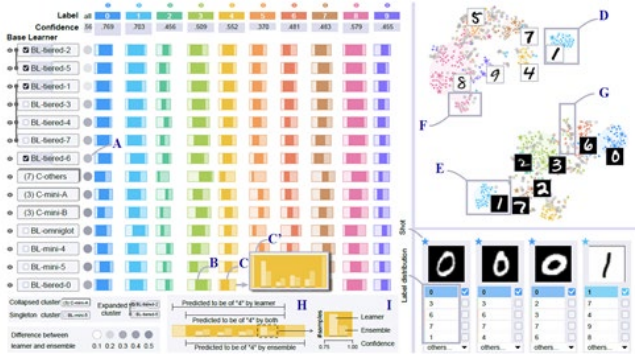
有噪声数据下样本加权

训练样本加权可视分析方法
(IEEE TVCG2024)



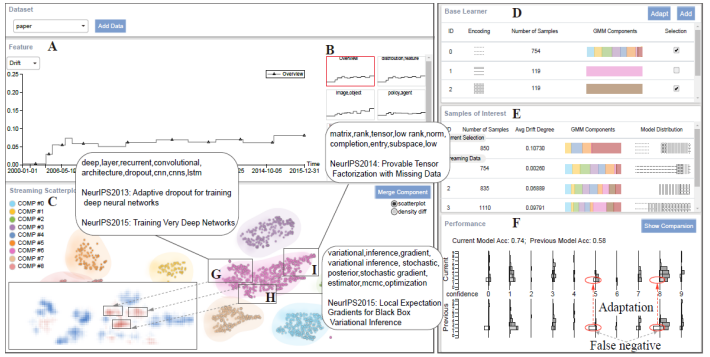
小样本学习中样本选择

小样本学习可视分析方法
(IEEE TVCG2022)



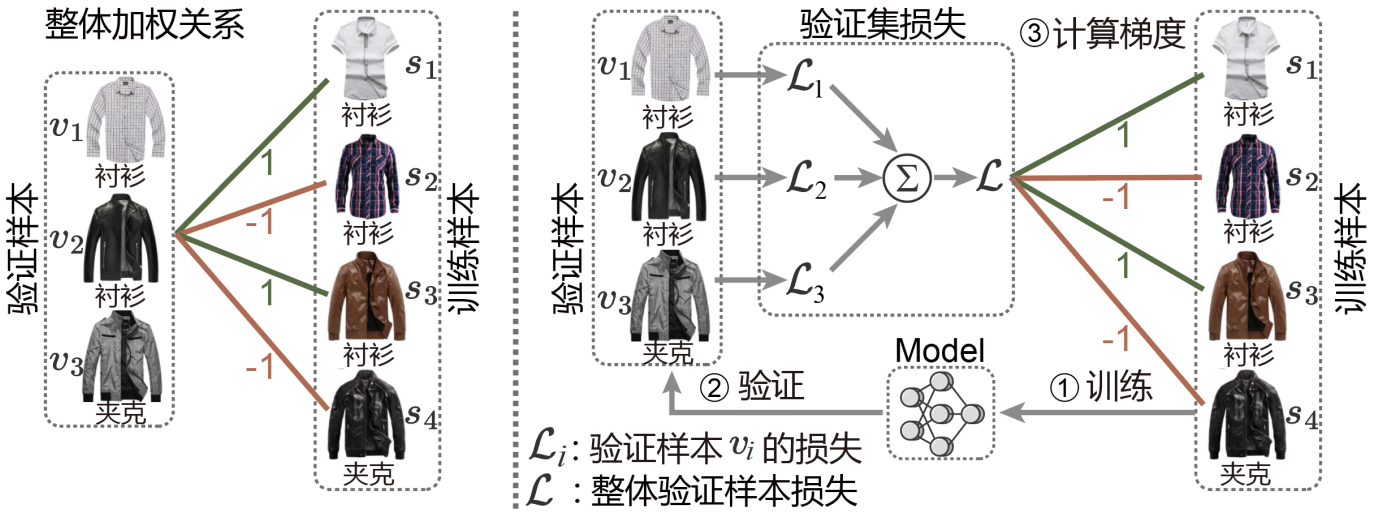
动态环境下样本更新

概念漂移可视分析方法
(IEEE VAST2020)



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?

技术挑战与解决方案

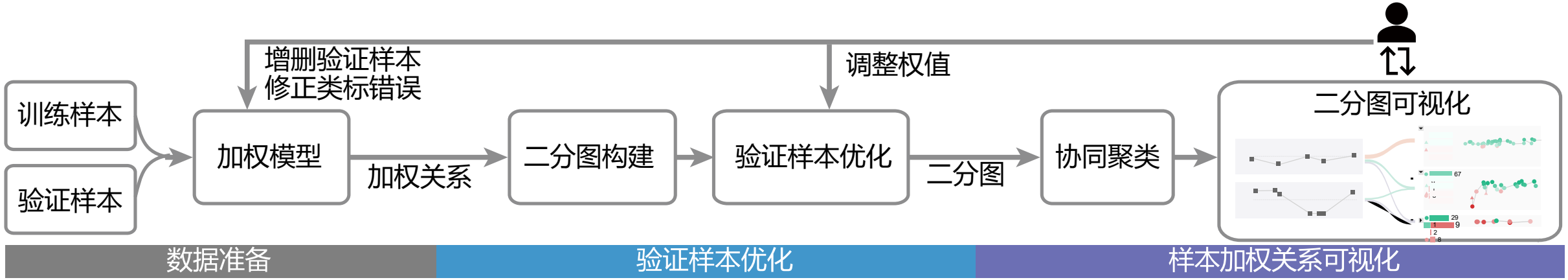
- 构建高质量的验证样本集
  - 自动构建方法往往存在噪声
  - 自动构建方法容易导致样本有偏
  - 逐样本分析过于繁琐
  - 领域专家对机器学习模型不熟悉

需分析验证样本和训练样本间的加权关系

需提供易用的可视分析系统

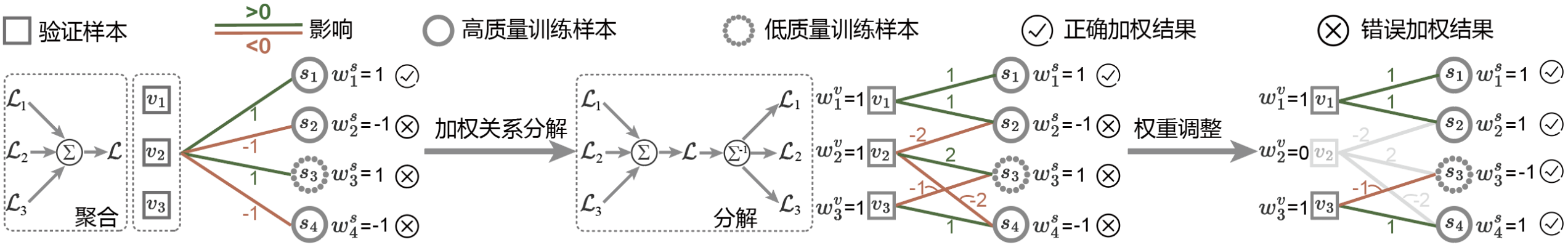
- 解决方案
  - 验证样本优化方法
  - 加权关系可视分析工具

系统概览


增删验证样本
修正类标错误
调整权值
训练样本
验证样本
加权模型
加权关系
二分图构建
验证样本优化
二分图
协同聚类
二分图可视化
数据准备
验证样本优化
样本加权关系可视化

验证样本优化

- 为识别低质量验证样本，首先需要进行加权关系分解
- 如何衡量验证样本的质量
  - 正确性 $$\mathcal{L}_{\mathrm{c}}(w_1^s,\ldots,w_n^s)=\sum_{s_j\in S_+}-\log\phi(w_j^s)+\sum_{s_j\in S_-}-\log(1-\phi(w_j^s)),$$
  - 平衡性 $$\mathcal{L}_{\mathrm{b}}(w_j^s,\ldots,w_j^s)=\sum_{c=1}^{C}p_c\log p_c. \qquad p_c=(\textstyle\sum_{s_j\in \boldsymbol{s}_{c+}}w_j^s)/(\sum_{s_j\in \boldsymbol{s}_{+}}w_j^s)$$
  - 多任务优化 $$\underset{w_i^v,\sigma_c,\sigma_b}{\text{minimize}}\quad \frac{1}{\sigma_c^2}\mathcal{L}_{\mathrm{c}}(w_1^s,\ldots,w_n^s)+\frac{1}{\sigma_b^2}\mathcal{L}_{\mathrm{b}}(w_1^s,\ldots,w_n^s)+\log\sigma_c\sigma_b$$



样本加权 → 样本选择 → 样本更新

数值实验

- 数据集: Cifar10, Cifar100, Clothing (细粒度分类/有噪声), OCT (医疗图像/不平衡)
- 模拟类标噪声：使用伪类标　　模拟类标分布不平衡：帕累托分布
- 方法：Uniform (不加权), FSR (自动选择验证样本加权), Ours (FSR+验证样本优化)

数据集	方法	有标注样本数目 10	20	50	100
CIFAR10	Uniform	56.3%	62.0%	67.1%	70.5%
	FSR	67.6%	74.7%	78.6%	81.2%
	Ours	**68.4%**	**75.4%**	**79.0%**	**81.4%**
CIFAR100	Uniform	30.4%	35.9%	43.9%	50.4%
	FSR	37.0%	43.5%	53.4%	58.8%
	Ours	**38.1%**	**44.2%**	**53.7%**	**59.1%**

(a) 模拟类标噪声

数据集	噪音比例	方法	准确率
Clothing	0.383	Uniform	57.8%
		FSR	70.5%
		Ours	**71.9%**

(b) 真实类标噪声

数据集	方法	不平衡程度 5	10	20	50
CIFAR10	Uniform	92.2%	90.1%	87.1%	81.2%
	FSR	93.4%	90.1%	87.2%	81.7%
	Ours	**93.6%**	**90.3%**	**87.8%**	**82.1%**
CIFAR100	Uniform	63.8%	31.2%	32.6%	34.9%
	FSR	69.8%	63.9%	56.8%	48.9%
	Ours	**70.2%**	**64.5%**	**57.2%**	**49.4%**

(a) 模拟类标分布不平衡

数据集	不平衡程度	方法	准确率
OCT	10	Uniform	78.1%
		FSR	82.8%
		Ours	**83.4%**

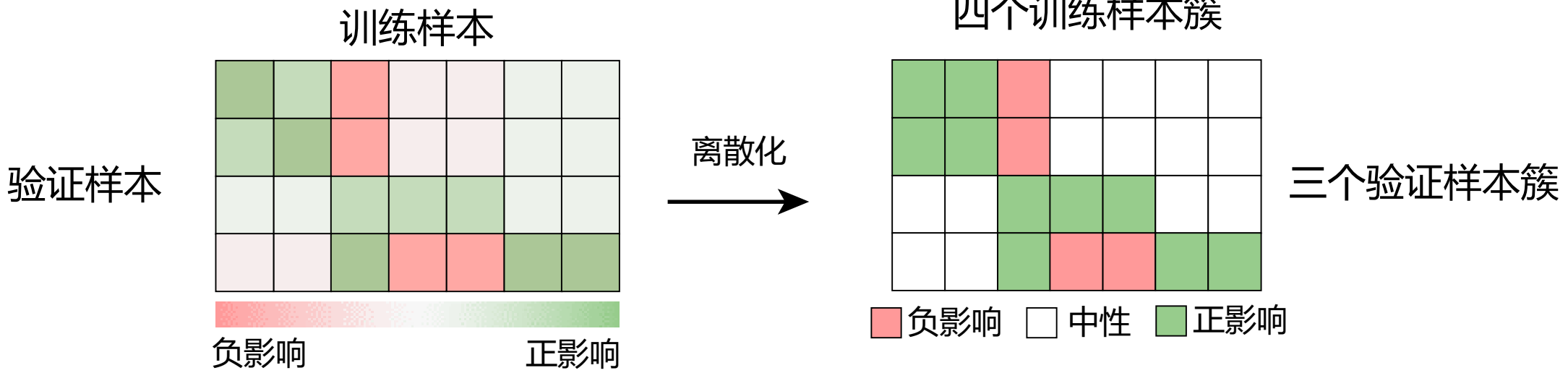
(b) 真实类标分布不平衡

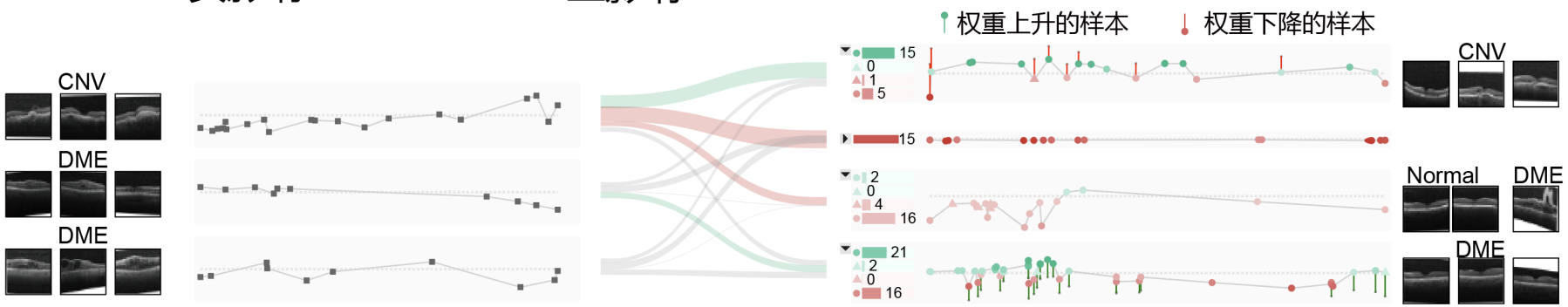
数据集	有标注样本数目	不平衡程度	Uniform	FSR	Ours
CIFAR10	10	5	55.1%	57.4%	**58.0%**
	10	10	54.2%	56.3%	**56.9%**
	20	5	62.4%	62.8%	**63.2%**
	20	10	61.2%	62.1%	**62.7%**
CIFAR100	10	5	29.5%	30.4%	**30.8%**
	10	10	28.2%	29.0%	**29.7%**
	20	5	34.7%	35.0%	**35.2%**
	20	10	32.9%	33.4%	**33.8%**
Clothing	N/A	5	55.9%	59.8%	**62.1%**
	N/A	10	52.6%	57.4%	**59.8%**
OCT	10	N/A	49.5%	57.4%	**59.6%**
	20	N/A	53.0%	63.8%	**65.2%**

类标噪声+类标分布不平衡

样本加权关系可视化

- 挑战：验证样本和训练样本间的加权关系多而复杂
- 解决方案：协同聚类



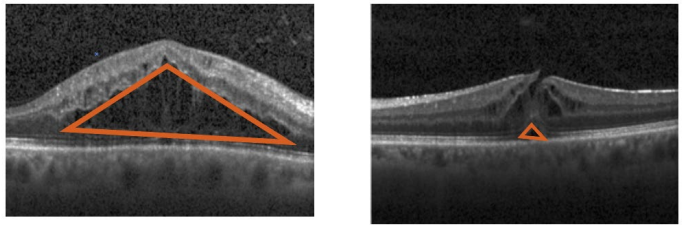


x: 1d-tSNE 投影
y: 样本权值
连边：加权关系

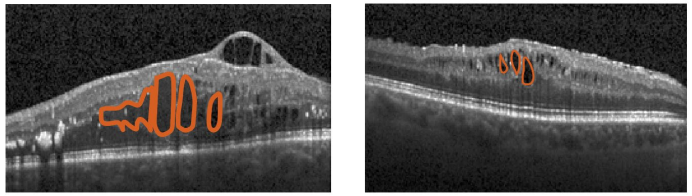
样本加权 → 样本选择 → 样本更新

案例分析

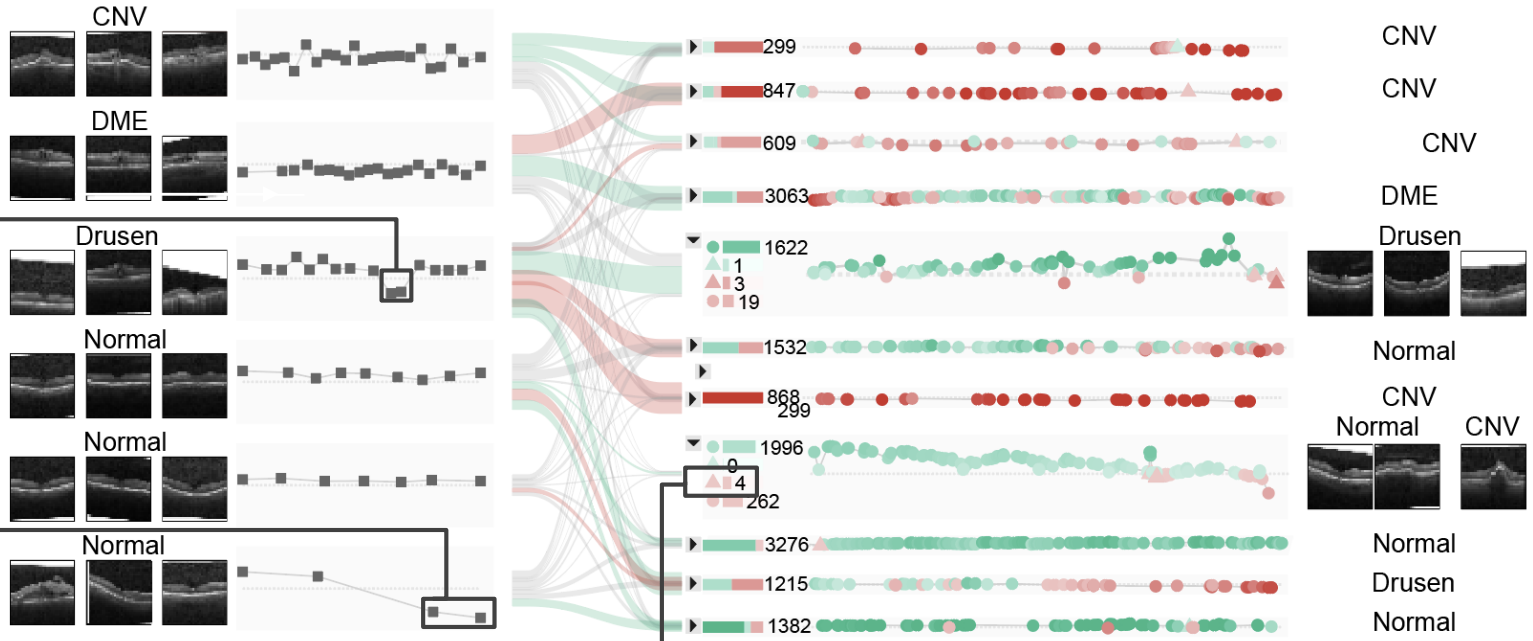
- OCT数据集
  - 检查权重低的验证样本



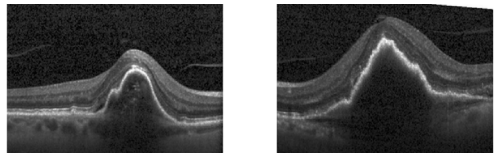
DME 但错标为 Drusen



DME 但错标为 Normal

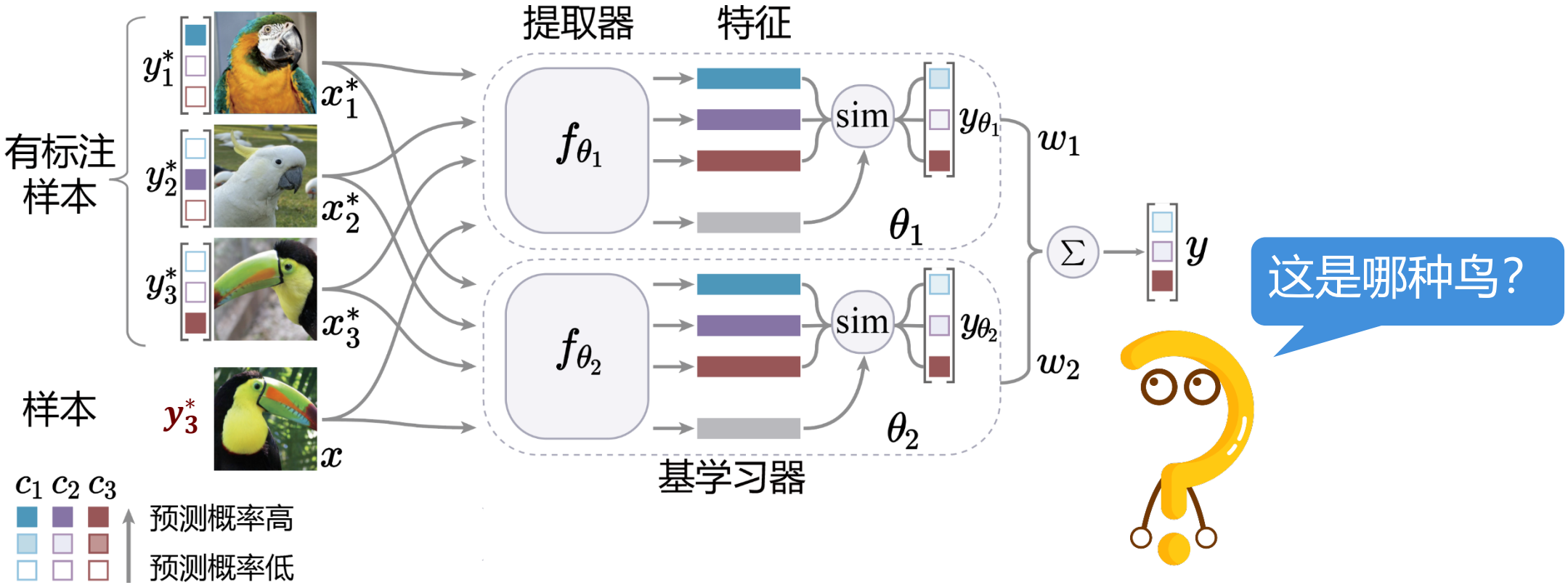


  - 检查训练样本加权结果有问题的样本
    - 验证样本中CNV代表性不足



样本加权 → 样本选择 → 样本更新

小样本学习可视分析方法

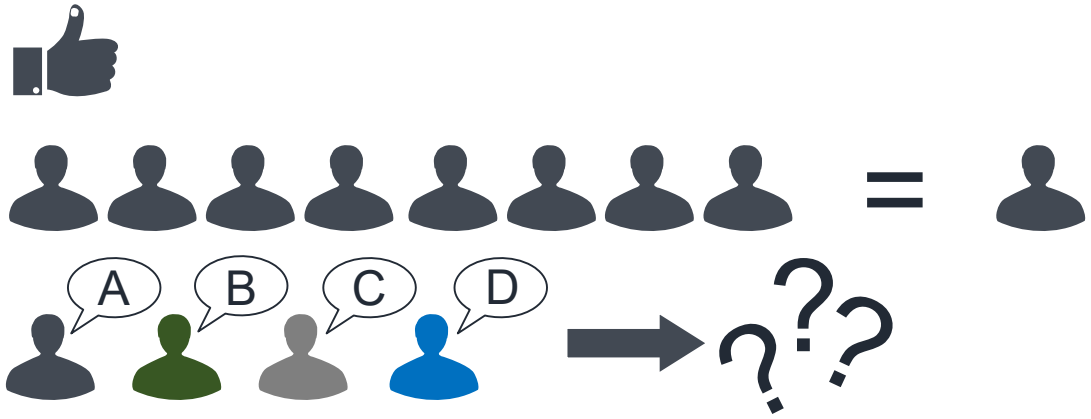


- **基学习器**和**有标注样本**的质量极大影响小样本学习的性能

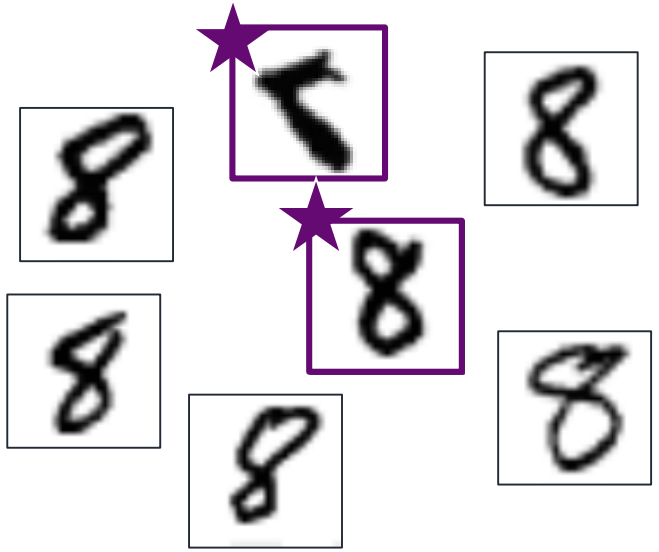
Diagnosing Ensemble Few-shot Classifiers. IEEE TVCG.

技术挑战

- 基学习器选择
  - 高性能
  - 多样性
  - 合作性



- 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



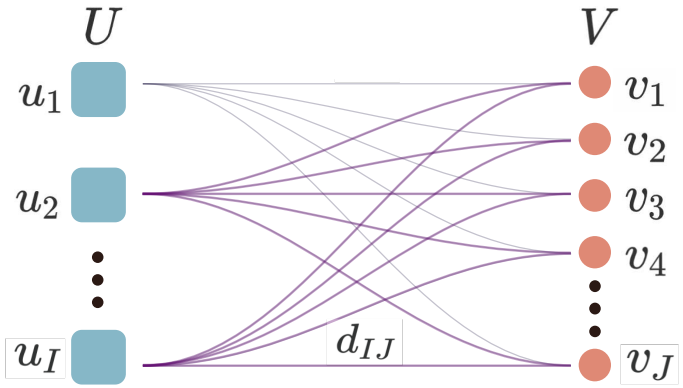
★ shot

样本加权 → **样本选择** → 样本更新

解决方案

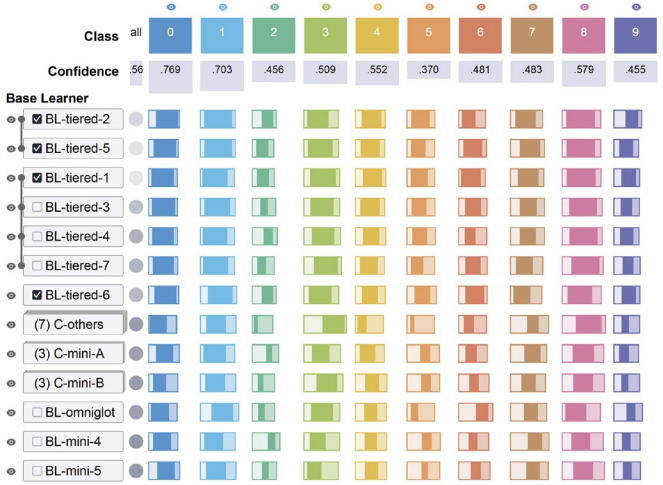
- 基学习器选择
  - 高性能
  - 多样性
  - 合作性

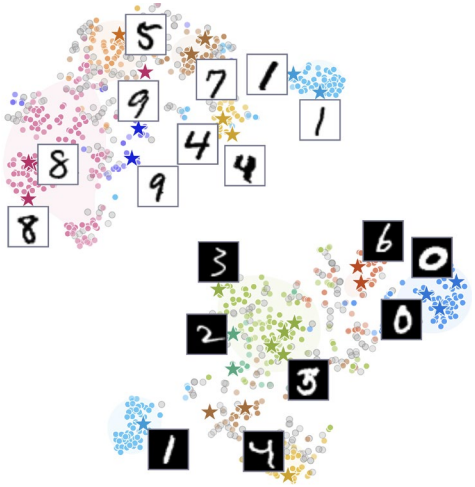
稀疏子集选择问题



- 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

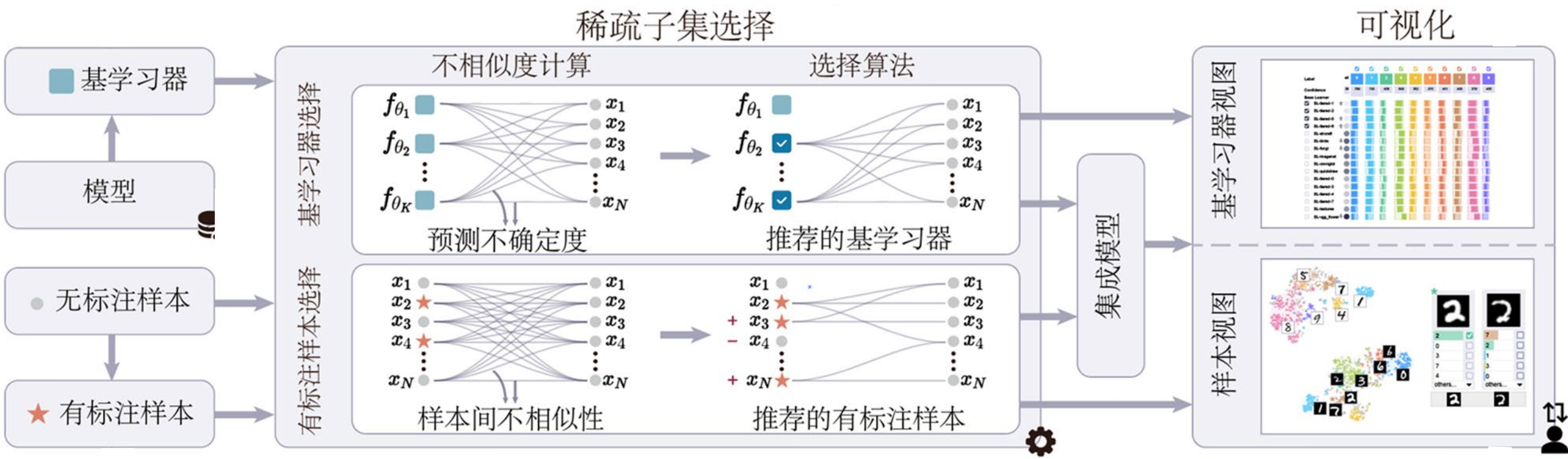
可视化



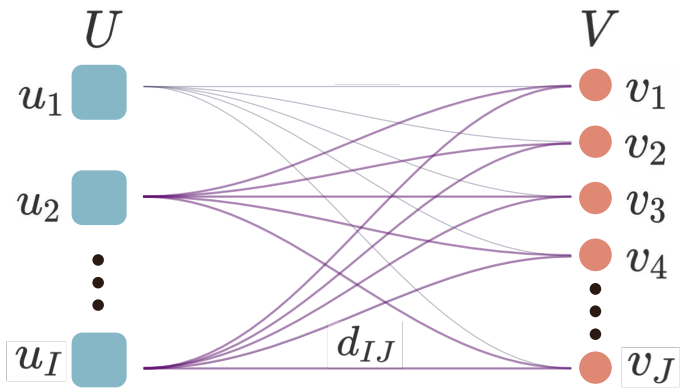


样本加权 → **样本选择** → 样本更新

系统概览


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

稀疏子集选择

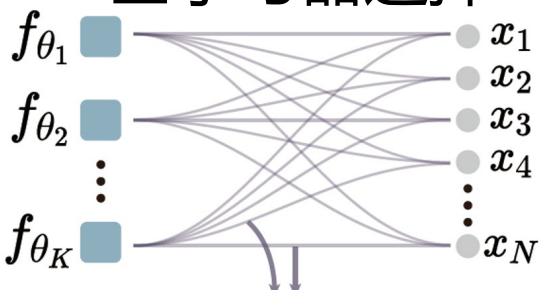


$$\mathcal{L}(A) = \underbrace{\sum_{j=1}^{J} \min_{u_i \in A} d_{ij}}_{\text{代表度}} + \underbrace{\alpha|A|}_{\text{稀疏度}}$$

松弛为连续优化问题
使用ADMM(交替方向乘子法)
进行求解

$$\mathcal{L}(Z) = \sum_{j=1}^{J}\sum_{i=1}^{I} \underbrace{z_{ij}}_{\text{用 } u_i \text{ 代表} v_j} d_{ij} + \alpha \sum_{i=1}^{I} \underbrace{\max_j z_{ij}}_{u_i \text{ 被使用}}$$

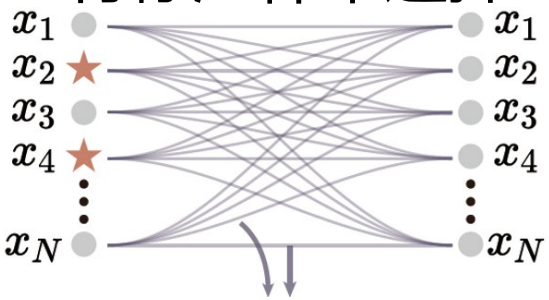
基学习器选择



预测不确定度

$$\underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ki} d_{ki}}_{\text{多样性}} + \alpha_1 \underbrace{\sum_{k=1}^{K} \lambda_k \max_i z_{ki}}_{\text{高性能}} + \alpha_2 \underbrace{\sum_{1 \le k < l \le K} \mu_{kl} \max_i z_{ki} \cdot \max_i z_{li}}_{\text{合作性}}$$

有标注样本选择



图像间不相似度

$$\underbrace{\sum_{j=1}^{N}\sum_{i=1}^{N} z_{ij} d_{ij}}_{\text{代表度}} + \alpha \underbrace{\sum_{i=1}^{N} \beta_i \gamma_i \max_j z_{ij}}_{\text{信息量\&标注代价}}$$

样本加权 → 样本选择 → 样本更新

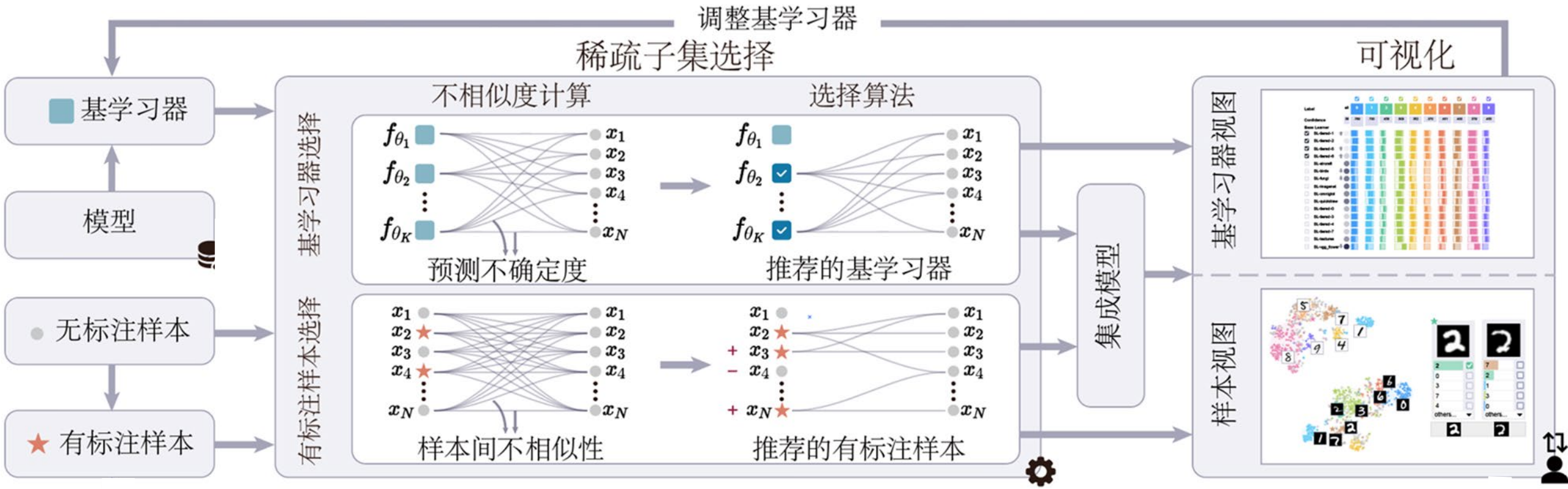
数值实验

- 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)

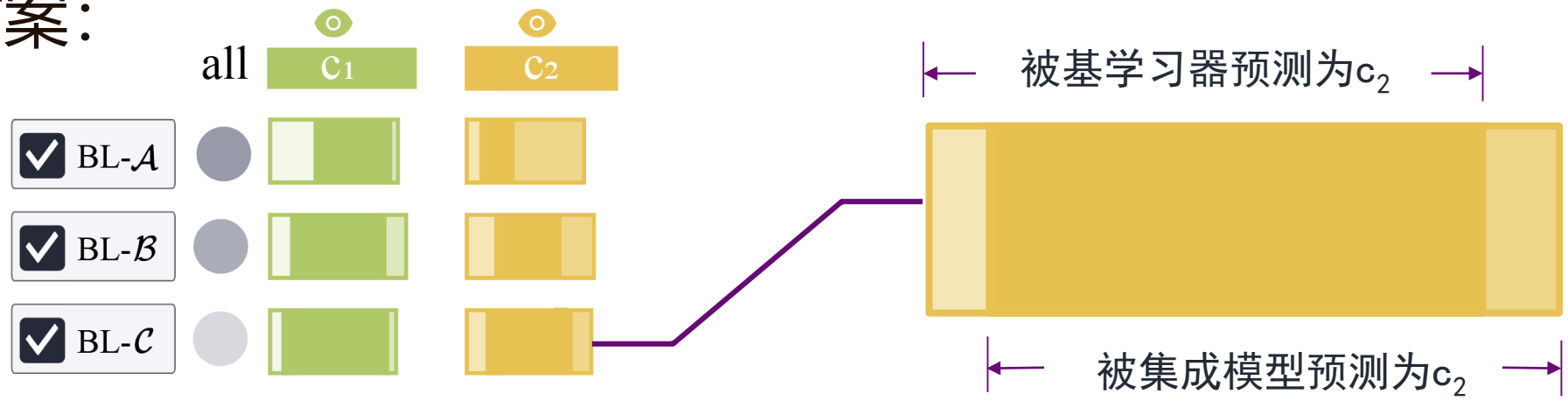
- 推荐的基学习器和有标注样本都能改善模型性能
- 二者组合可以进一步提高模型性能

系统概览


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

基学习器视图

- 挑战：没有真实类标
- 解决方案：



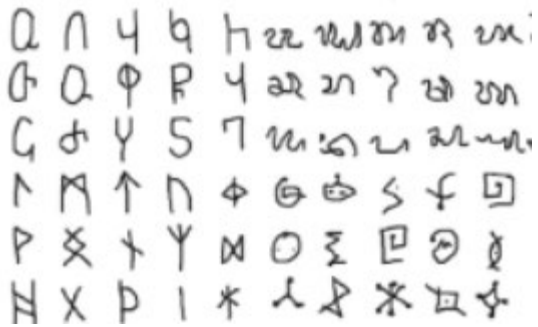
- 挑战：可扩展性
- 解决方案:

样本加权 样本选择 样本更新

案例分析

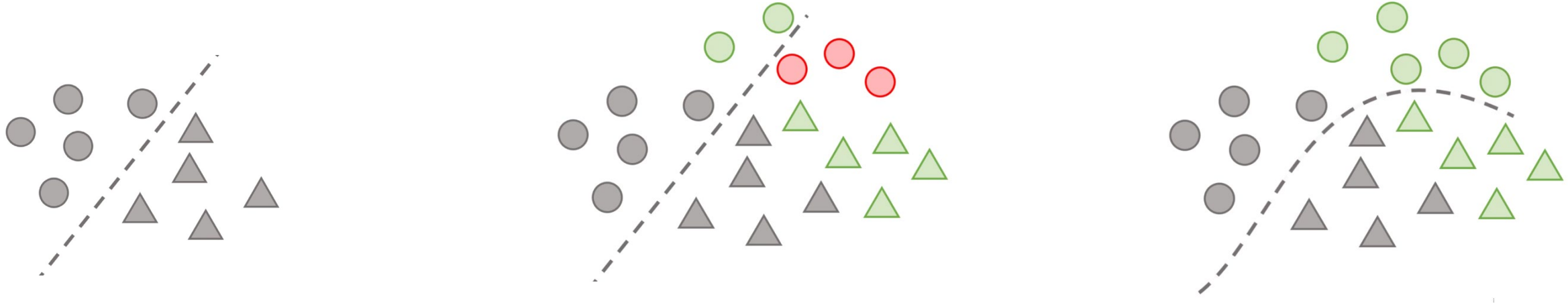
- CIFAR-100数据集
  - 使用测试集（20分类问题），初始60个有标注样本
  - 发现了有标注样本覆盖窄、质量低导致性能不佳
  - 加入37个有标注样本，分类准确率 47.4%->59.4%(**+12%**)

- 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%**)



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



Diagnosing Concept Drift with Visual Analytics. IEEE VIS (VAST).

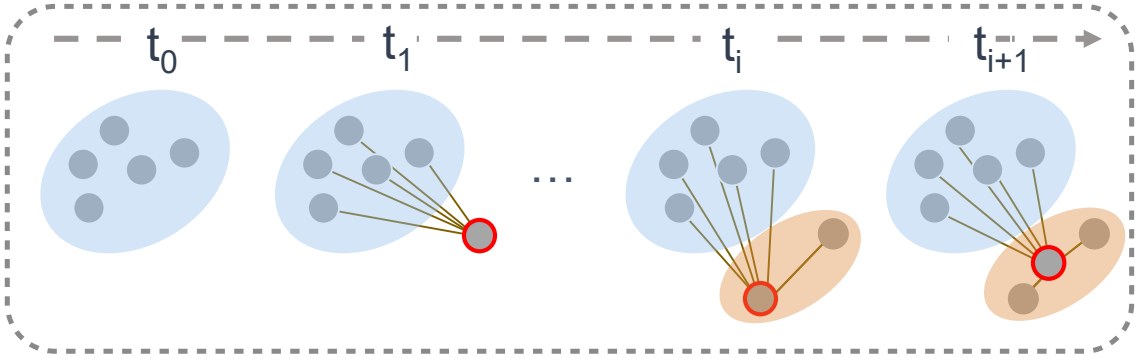
样本加权 样本选择 样本更新

技术挑战

- 类标难以及时获得，不能根据性能下降来检测概念漂移
  - 需要一个基于分布的 **概念漂移检测** 算法
- 概念漂移度量值难以直接理解
  - 需要一个系统性的 **概念漂移分析** 方法
- 对所有数据重新标注来解决概念漂移代价高昂
  - 需要一个轻量级的 **概念漂移适应** 方法

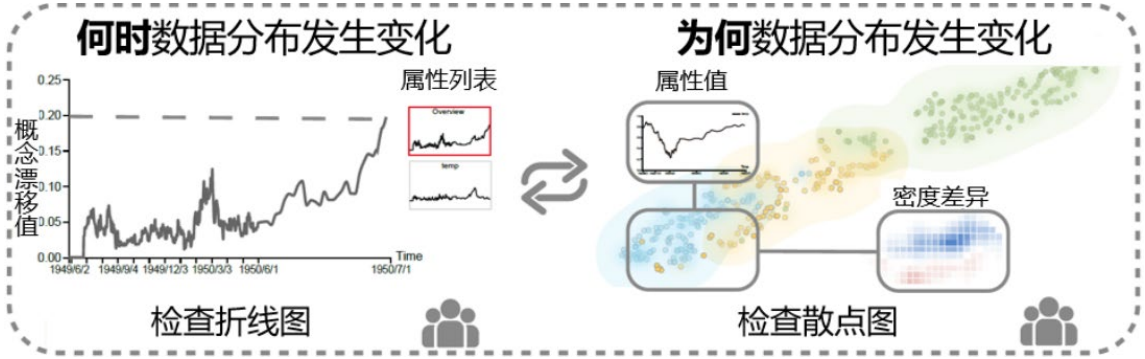
系统概览

概念漂移检测



能量距离

概念漂移分析



有约束t-SNE布局算法

概念漂移适应

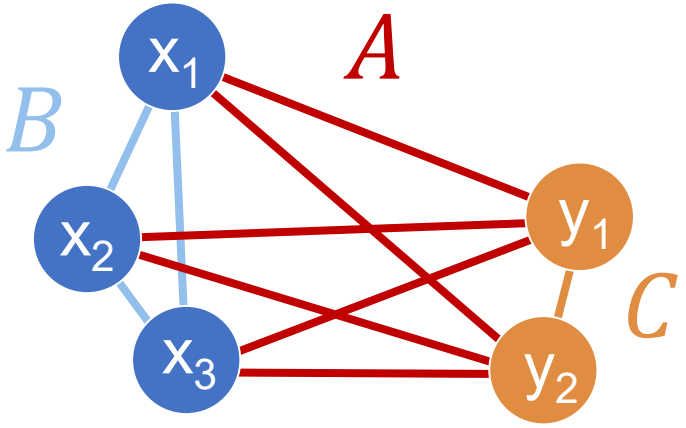
集成模型

样本加权 样本选择 **样本更新**

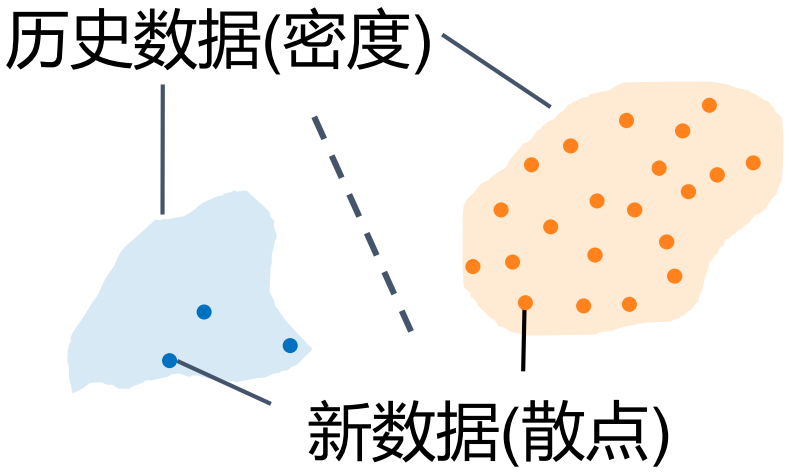
概念漂移检测

- 基于能量距离的度量方法

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



- 高斯混合模型解决过高估计的问题



数值实验

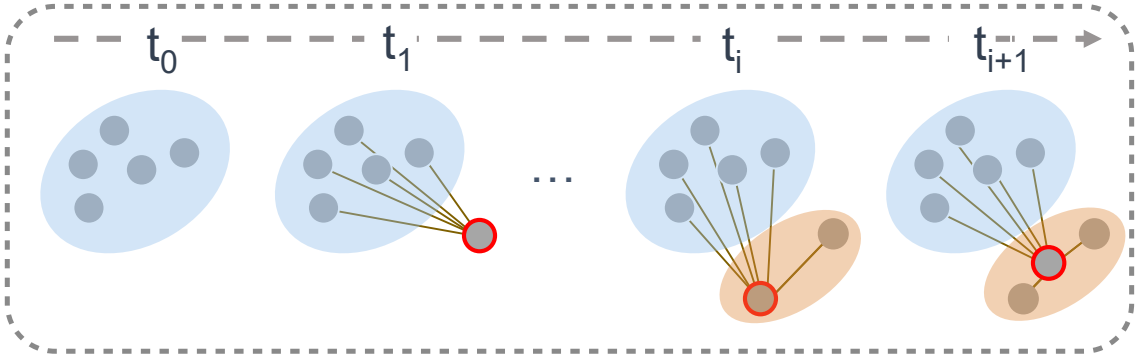
- 数据：模拟了两组含有5,000,000数据点和99次概念漂移的数据流
- 指标：成功检测/延迟检测/未能检测/错误检测
- 与三个基准方法进行对比

数据集	算法	成功检测	延迟检测	未能检测	错误检测
D_1	ITA [13]	75.5	12.3	11.2	4.5
	LDD-DIS [40]	56.5	19.3	23.2	3.9
	TVD [71]	75.5	19.9	3.6	7.0
	Ours	**97.9**	**0.9**	**0.2**	**0.8**
D_2	ITA [13]	75.7	18.0	**5.3**	7.6
	LDD-DIS [40]	44.0	22.2	32.8	12.2
	TVD [71]	71.2	22.3	5.5	**6.0**
	Ours	**77.5**	**14.9**	6.6	6.4

以较低的延迟检测出更多更准的概念漂移

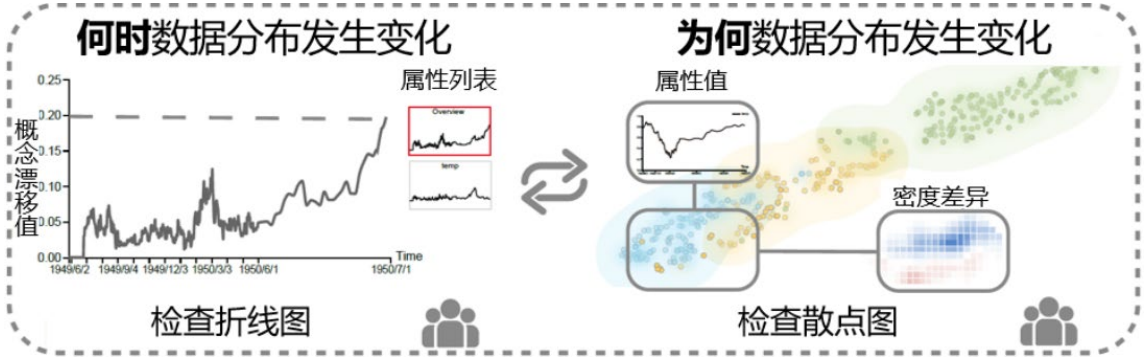
系统概览

概念漂移检测



能量距离

概念漂移分析



有约束t-SNE布局算法

概念漂移适应

集成模型

样本加权 样本选择 **样本更新**

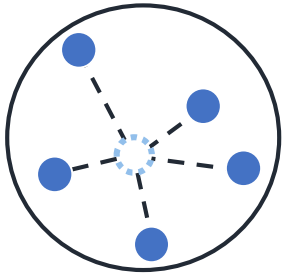
概念漂移分析

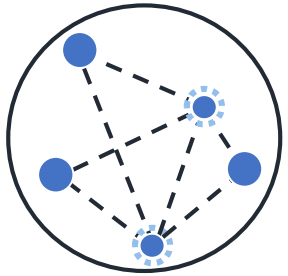
- 基于高斯混合模型的有约束t-SNE布局算法
  - 减少视觉混杂

$$\mathrm{dist}(x_i, x_j) = \begin{cases} \alpha_k \cdot \|x_i - x_j\| & x_i, x_j \in C_k \\ \|x_i - x_j\| & \text{otherwise} \end{cases} \qquad \alpha_k = \alpha \cdot \left(1 - \beta \cdot \frac{H[C_k]}{\max_{k'}\{H[C_{k'}]\}}\right)$$

  - 布局优化：可读性+稳定性

$$\underbrace{\lambda \cdot KL(P\|Q)}_{\text{原始目标函数}} + \underbrace{\varphi \cdot KL(P_c\|Q_c)}_{\text{类别可读性}} + \underbrace{(1-\lambda-\varphi) \cdot KL(P_s\|Q_s)}_{\text{布局稳定性}}$$

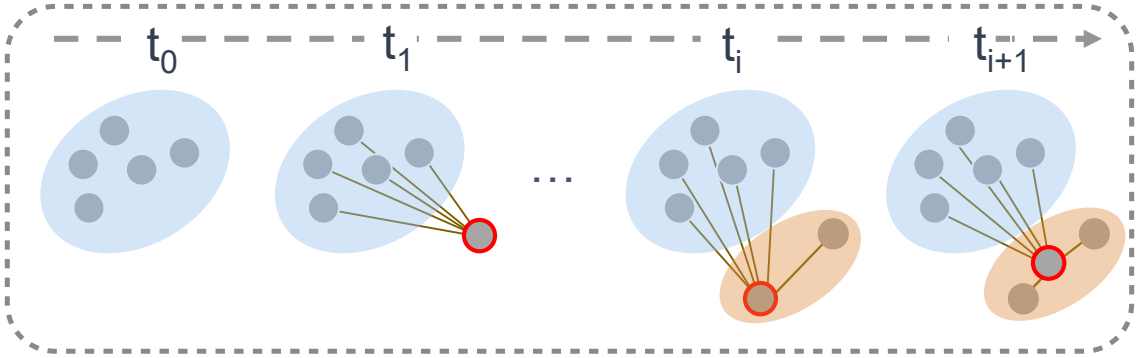




样本加权 → 样本选择 → 样本更新

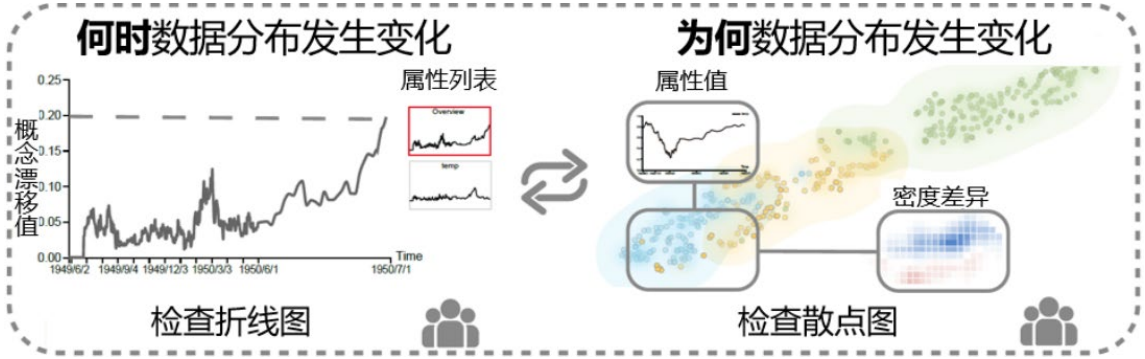
系统概览

概念漂移检测



能量距离

概念漂移分析



有约束t-SNE布局算法

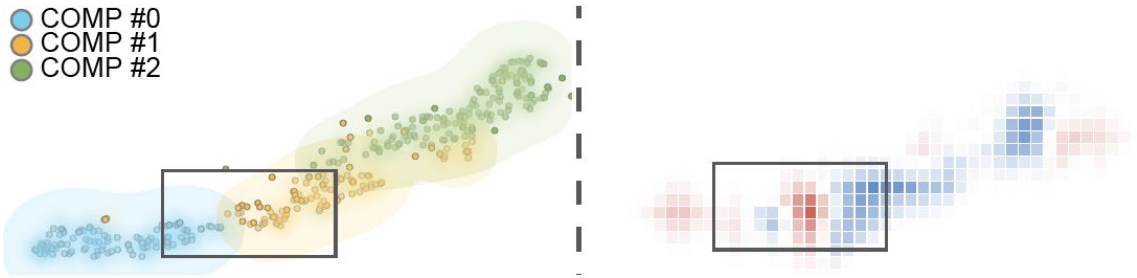
概念漂移适应

集成模型

样本加权 样本选择 样本更新

案例分析

- 天气预测数据集
  - 根据温度、湿度、气压的天气指标，预测第二天是否降雨/雪
  - 发现了由数据不全造成的**季节性**概念漂移与极端天气带来的**偶发性**概念漂移
  - 分类准确率 72%->77%



- 论文数据集
  - 分析NeurIPS 2000-2019间发表论文的变化情况并进行分类
  - 依次发现了SVM/LDA/DNN/VAE/GAN等学术热点带来的概念漂移
  - 分类准确率 58%->74%

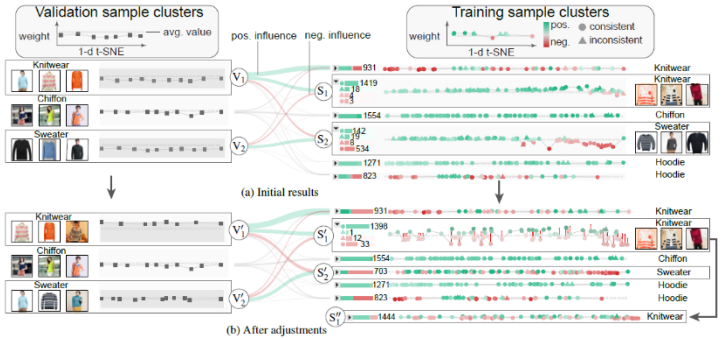
总结

基于**样本代表性**的模型性能**可视分析**及提升方法

数据准备 → 模型训练 → 模型部署

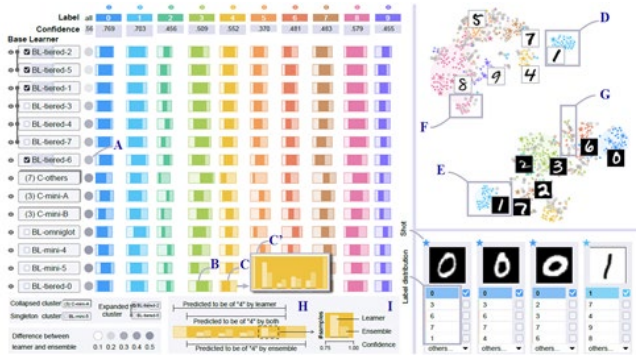
有噪声数据下样本加权

训练样本加权可视分析方法



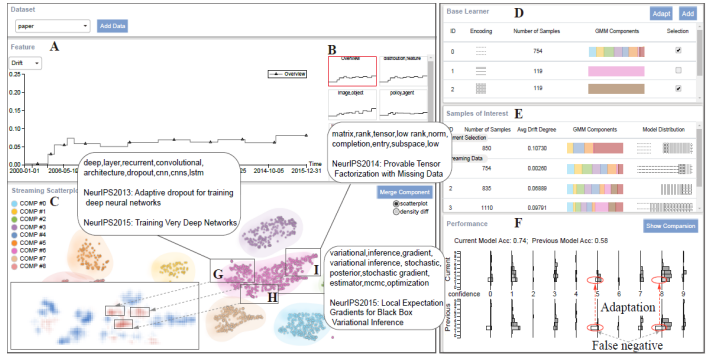
小样本学习中样本选择

小样本学习可视分析方法



动态环境下样本更新

概念漂移可视分析方法



谢谢大家！敬请批评指正！

https://vicayang.cc