# 半监督学习综述

–半监督学习、深度半监督学习、安全深度半监督学习

# 研究背景

## 强泛化能力的机器学习模型需要海量的标记数据

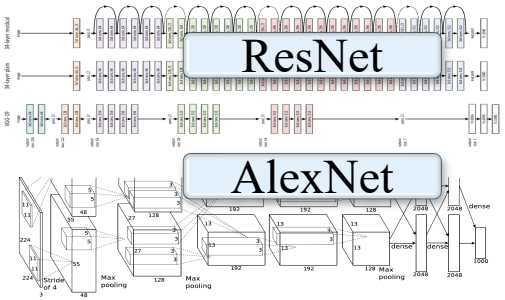
### 应用：

- 业界标杆企业（谷歌、华为等）每年花费千万级成本用于标注数据

Google HUAWEI Alibaba.com

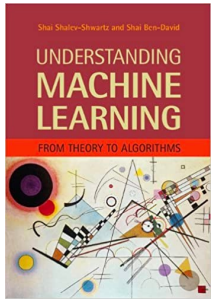
### 算法：

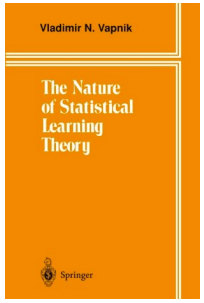
- SOTA深度神经网络模型都依赖海量的标记数据




### 理论：

- 强泛化性能的模型需要海量标记数据（经典统计学习理论）



# 研究背景

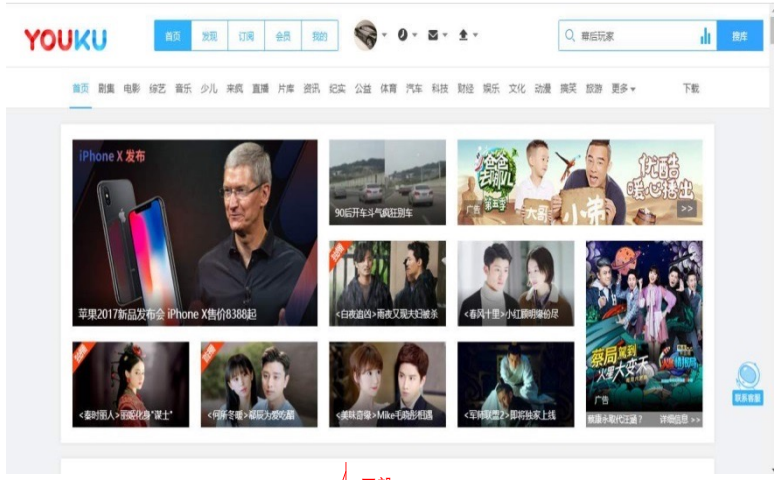
## 标记数据需要付出不菲的代价

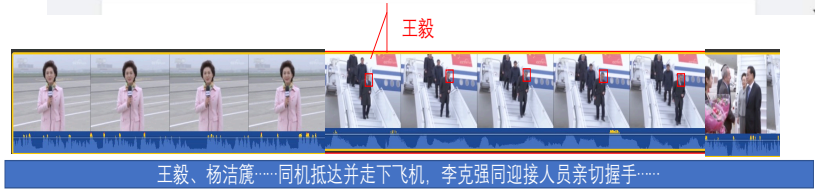
视频理解：最大化使用体验

- 数据：大量业务数据和回流数据
- 系统： “算法+产品+数据” 闭环系统

现行方法：

- ASR/OCR和视觉理解算法存在严重错标、漏标等现象
- 达不到算法迭代优化的标准
- **大量数据资源的浪费**

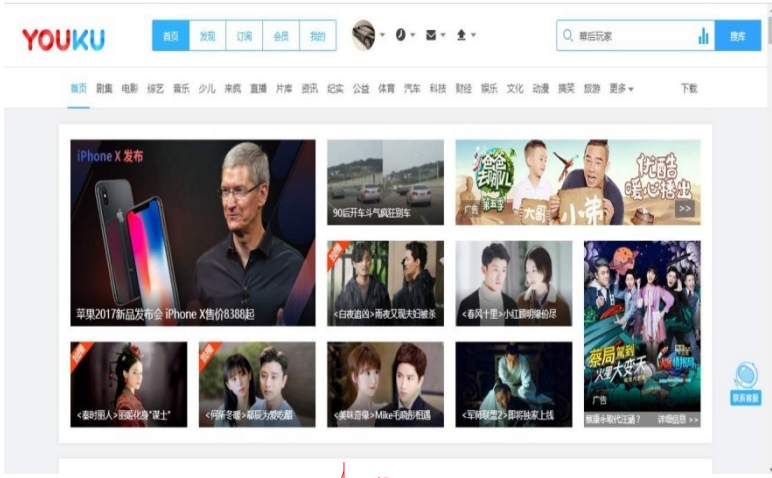



# 研究背景

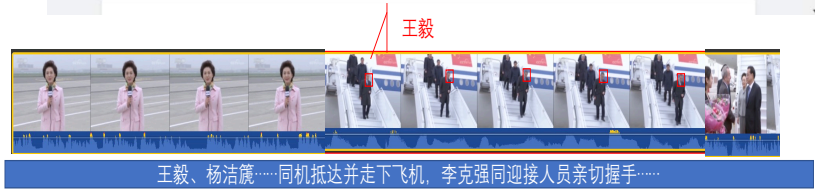
## 标记数据需要付出不菲的代价

自动化清洗视频数据中的标记，
在此基础上自动挖掘潜在的模式
以提升算法的视觉理解能力




**弱监督学习技术**

一种面向大规模弱标注数据自动
提升标注质量的机器学习技术





- ✓ 极具应用价值的研究领域
- ✓ 顶尖研究机构（谷歌、华为等）
  均投入了巨大的研究力量

# 未标记样本的效用

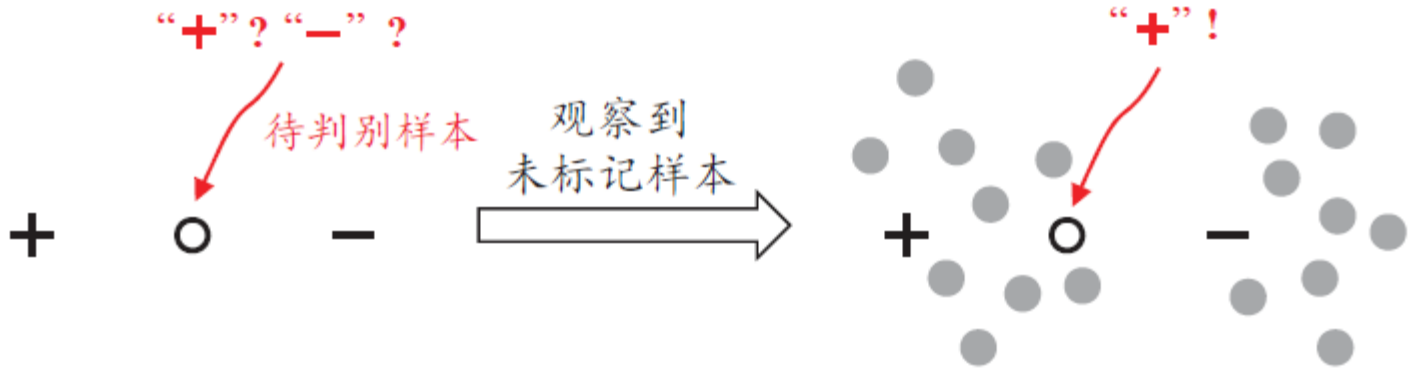



图 13.1　未标记样本效用的例示. 右边的灰色点表示未标记样本

# 未标记样本的假设

要利用未标记样本，必然要做一些将未标记样本所揭示的数据分布信息与类别标记相联系的假设，其中有两种常见的假设。

- 聚类假设（clustering assumption）：

  假设数据存在簇结构，同一簇的样本属于同一类别。

- 流形假设（manifold assumption）：

  假设数据分布在一个流形结构上，邻近的样本具有相似的输出值。

流形假看设可做聚类假设的推广

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 生成式方法

- 假设样本由高斯混合模型生成，且每个类别对应一个高斯混合成分：

$$p(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

其中，$\alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1$

$$p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}$$

# 生成式方法

- 由最大化后验概率可知：

$$
\begin{aligned}
f(\boldsymbol{x}) &= \underset{j \in \mathcal{Y}}{\operatorname{argmax}}\, p(y=j|\boldsymbol{x}) \\
&= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y=j, \Theta=i|\boldsymbol{x}) \qquad p(y=j|\Theta=i) \\
&= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y=j|\Theta=i,\boldsymbol{x}) \cdot p(\Theta=i|\boldsymbol{x})
\end{aligned}
$$

其中 $p(\Theta=i|x) = \dfrac{\alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$

# 生成式方法

- 假设样本独立同分布，且由同一个高斯混合模型生成，则对数似然函数是：

$$\ln p(D_l \cup D_u) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left( \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot p(y_j \mid \Theta = i, \boldsymbol{x}_j) \right)$$

$$+ \sum_{\boldsymbol{x}_j \in D_u} \ln \left( \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) .$$

# 生成式方法

高斯混合的参数估计可以采用EM算法求解，迭代更新式如下：

- E步：根据当前模型参数计算未标记样本$x_j$属于各高斯混合成分的概率：

$$\gamma_{ji} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i}$$

# 生成式方法

- M步：基于$\gamma_{ji}$ 更新模型参数

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left( \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_i, y_i) \in D_l \wedge y_j = i} \boldsymbol{x}_j \right)$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left( \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \right.$$

$$\left. + \sum_{(\boldsymbol{x}_i, y_i) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \right)$$

$$\alpha_i = \frac{1}{m} \left( \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right)$$

# 生成式方法

- 将上述过程中的高斯混合模型换成混合专家模型，朴素贝叶斯模型等即可推导出其他的生成式半监督学习算法。

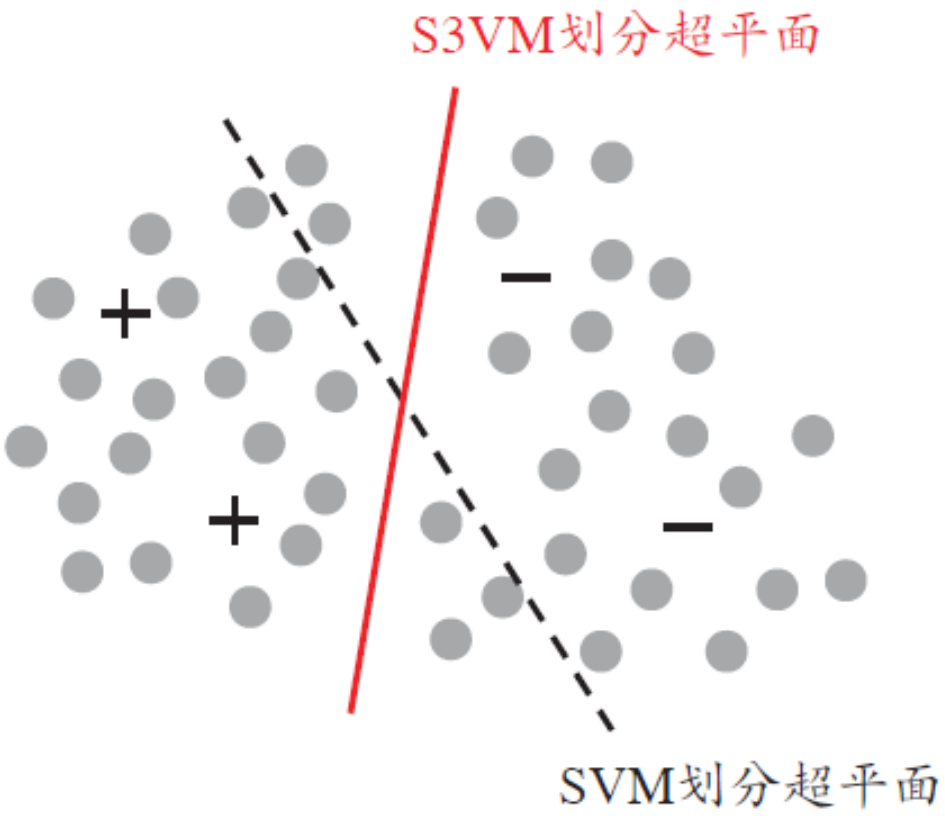
- 此类方法简单、易于实现，在有标记数据极少的情形下往往比其他方法性能更好。

- 然而，此类方法有一个关键：模型假设必须准确，即假设的生成式模型必须与真实数据分布吻合；否则利用未标记数据反而会显著降低泛化性能。

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 半监督SVM



# 半监督SVM

半监督支持向量机中最著名的是TSVM(Transductive Support Vector Machine)

$$\min_{\boldsymbol{w},b,\hat{\boldsymbol{y}},\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i$$

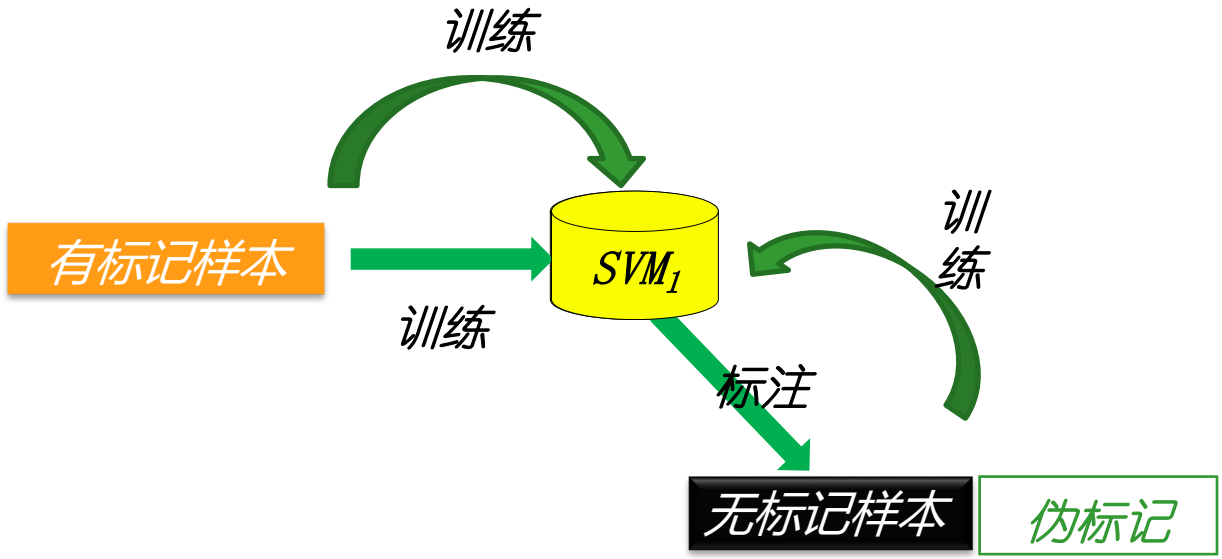
$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l,$$

$$\hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = l+1, \ldots, m,$$

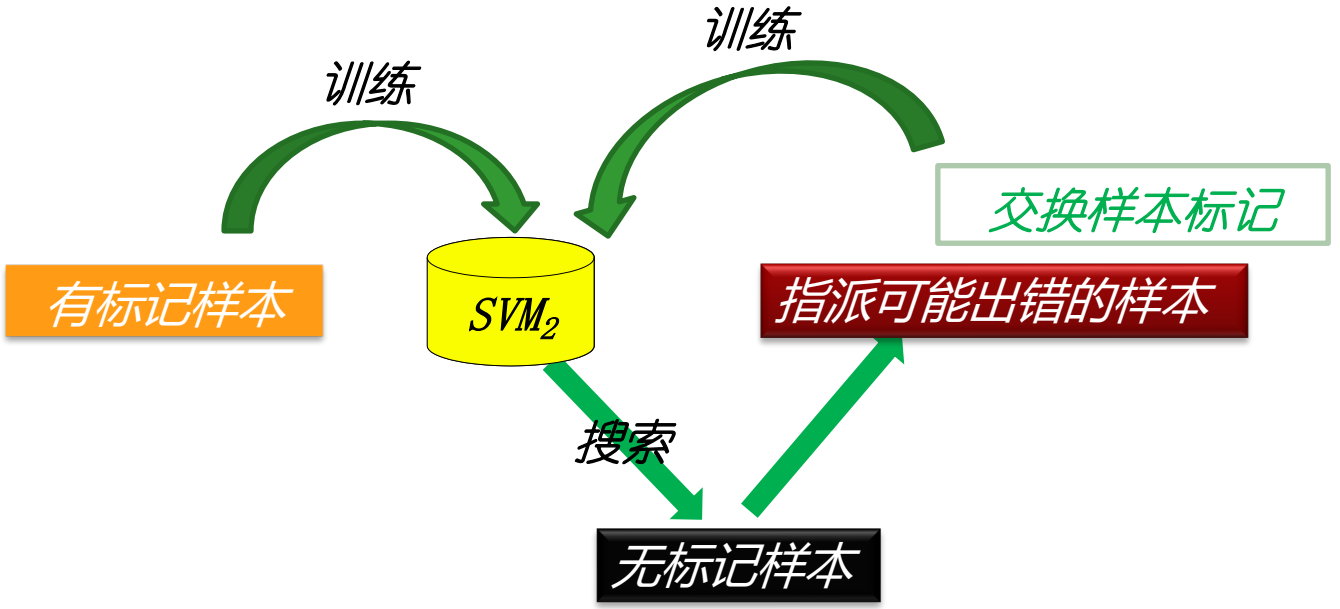
$$\xi_i \geq 0, \quad i = 1, \ldots, m,$$

# 半监督SVM

- TSVM采用局部搜索来迭代地寻找近似解.



# 半监督SVM



# 半监督SVM

*未标记样本的伪标记不准确*

**输入：** 有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$;
未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
折中参数 $C_l$, $C_u$.

**过程：**

用 $D_l$ 训练一个 $\text{SVM}_l$;
用 $\text{SVM}_l$ 对 $D_u$ 中样本进行预测, 得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$;
初始化 $C_u \ll C_l$;
**while** $C_u < C_l$ **do**
  基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$;
  **while** $\exists\{i, j \mid (\hat{y}_i \hat{y}_j < 0) \wedge (\xi_i > 0) \wedge (\xi_j > 0) \wedge (\xi_i + \xi_j > 2)\}$ **do**
    $\hat{y}_i = -\hat{y}_i$;
    $\hat{y}_j = -\hat{y}_j$;
    基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 重新求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$
  **end while**
  $C_u = \min\{2C_u, C_l\}$
**end while**

**输出：** 未标记样本的预测结果: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

图 13.4　TSVM 算法

# 半监督SVM

- 未标记样本进行标记指派及调整的过程中，有可能出现类别不平衡问题，即某类的样本远多于另一类。

- 为了减轻类别不平衡性所造成的不利影响，可对算法稍加改进：

- 将优化目标中的 $C_u$ 项拆分为 $C_u^+$ 与 $C_u^-$ 两项，并在初始化时令：

$$C_u^+ = \frac{u_-}{u_+} C_u^-$$

# 半监督SVM

- 显然，搜寻标记指派可能出错的每一对未标记样本，是一个涉及巨大计算开销的大规模优化问题。

- 因此，半监督SVM研究的一个重点是如何设计出高效的优化求解策略。

- 例如基于图核(graph kernel)函数梯度下降的Laplacian SVM[Chapelle and Zien, 2005]、基于标记均值估计的meanS3VM[Li et al., 2009]等.

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 图半监督学习

- 给定一个数据集，我们可将其映射为一个图，数据集中每个样本对应于图中一个结点，若两个样本之间的相似度很高(或相关性很强)，则对应的结点之间存在一条边，边的“强度”(strength)正比于样本之间的相似度(或相关性)。

- 我们可将有标记样本所对应的结点想象为染过色，而未标记样本所对应的结点则尚未染色．于是，半监督学习就对应于“颜色”在图上扩散或传播的过程。

- 由于一个图对应了一个矩阵，这就使得我们能基于矩阵运算来进行半监督学习算法的推导与分析。

# 图半监督学习

- 我们先基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$，其中结点集

$$V = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, ..., \boldsymbol{x}_{l+u}\}$$

- 边集$E$ 可表示为一个亲和矩阵(affinity matrix)，常基于高斯函数定义为:

$$\mathbf{W}_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise}, \end{cases}$$

# 图半监督学习

- 假定从图 $G = (V, E)$ 将学得一个实值函数 $f : V \to \mathbb{R}$ 。

- 直观上讲相似的样本应具有相似的标记,即得到最优结果于是可定义关于$f$ 的“能量函数”(energy function)[Zhu et al., 2003]:

$$
\begin{aligned}
E(f) &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \boldsymbol{W}_{ij} (f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2 \\
&= \frac{1}{2} \left( \sum_{i=1}^{m} d_i f^2(\boldsymbol{x}_i) + \sum_{j=1}^{m} d_j f^2(\boldsymbol{x}_j) - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \boldsymbol{W}_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j) \right) \\
&= \boldsymbol{f}^T (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{f}
\end{aligned}
$$

# 图半监督学习

- 采用分块矩阵表示方式：

$$E(f) = (\boldsymbol{f}_l^\top \ \boldsymbol{f}_u^\top) \left( \begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix}$$

$$= \boldsymbol{f}_l^\top (\mathbf{D}_{ll} - \mathbf{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^\top \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^\top (\mathbf{D}_{uu} - \mathbf{W}_{uu}) \boldsymbol{f}_u \ .$$

- 由 $\frac{\partial E(f)}{\partial \boldsymbol{f}_u} = \mathbf{0}$ 可得：

$$\boldsymbol{f}_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1} \mathbf{W}_{ul} \boldsymbol{f}_l$$

# 图半监督学习

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W} = \begin{bmatrix} \mathbf{D}_{ll}^{-1} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{ll}^{-1}\mathbf{W}_{ll} & \mathbf{D}_{ll}^{-1}\mathbf{W}_{lu} \\ \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul} & \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu} \end{bmatrix}$$

$$\mathbf{P}_{uu} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu},\ \mathbf{P}_{ul} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}$$

$$\boldsymbol{f}_u = (\boldsymbol{D}_{uu}(\boldsymbol{I} - \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}))^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

$$= (\boldsymbol{I} - \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu})^{-1}\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

$$= (\boldsymbol{I} - \boldsymbol{P}_{uu})^{-1}\boldsymbol{P}_{ul}\boldsymbol{f}_l$$

# 图半监督学习

- 上面描述的是一个针对二分类问题的“单步式”标记传播(label propagation)方法，下面我们来看一个适用于多分类问题的“迭代式”标记传播方法[Zhou et al., 2004].

- 仍基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$

- 定义一个 $(l+u)\times|\mathcal{Y}|$ 的非负标记矩阵 $\mathbf{F} = (\mathbf{F}_1^\top, \ldots, \mathbf{F}_{|\mathcal{Y}|}^\top)^\top$，第i行元素 $\boldsymbol{F}_i = (\boldsymbol{F}_{i1}, \ldots, \boldsymbol{F}_{i|\gamma|})$ 为示例 $\boldsymbol{x}_i$ 的标记向量，相应的分类规则为：

$$y_i = \mathrm{argmax}_{1\le j\le|\gamma|}\, \boldsymbol{F}_{ij}$$

- 将F初始化为：

$$\mathbf{F}(0) = \mathbf{Y}_{ij} = \begin{cases} 1, & \text{if } (1 \le i \le l) \wedge (y_i = j)\ ; \\ 0, & \text{otherwise}\ , \end{cases}$$

# 图半监督学习

- 基于W构造一个标记传播矩阵 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ ，其中 $\boldsymbol{D}^{-\frac{1}{2}} = \left(\frac{1}{\sqrt{d_1}}, \ldots, \frac{1}{\sqrt{d_{l+u}}}\right)$，于是有迭代计算式：

$$\mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t) + (1-\alpha)\mathbf{Y}$$

- 基于迭代至收敛可得：

$$\mathbf{F}^* = \lim_{t\to\infty} \mathbf{F}(t) = (1-\alpha)(\mathbf{I} - \alpha\mathbf{S})^{-1}\mathbf{Y}$$

# 图半监督学习

- 事实上，该算法对应于正则化框架[Zhou et al., 2004]:

$$\min_{\boldsymbol{F}} \frac{1}{2}\left(\sum_{i,j=1}^{l+u} \boldsymbol{W}_{ij}\left\|\frac{1}{\sqrt{d_i}}\boldsymbol{F}_i - \frac{1}{\sqrt{d_j}}\boldsymbol{F}_j\right\|^2\right) + \mu\sum_{i=1}^{l}\|\boldsymbol{F}_i - \boldsymbol{Y}_i\|^2$$

- 当 $\mu = \frac{1-\alpha}{\alpha}$ 时，最优解恰为迭代算法的收敛解 $\mathbf{F}^*$。

# 图半监督学习

**输入:** 有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$;
未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
构图参数 $\sigma$;
折中参数 $\alpha$.

**过程:**

1: 基于式(13.11)和参数 $\sigma$ 得到 $\mathbf{W}$;
2: 基于 $\mathbf{W}$ 构造标记传播矩阵 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$;
3: 根据式(13.18)初始化 $\mathbf{F}(0)$;
4: $t = 0$;
5: **repeat**
6: $\quad \mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t) + (1-\alpha)\mathbf{Y}$;
7: $\quad t = t + 1$
8: **until** 迭代收敛至 $\mathbf{F}^*$
9: **for** $i = l+1, l+2, \ldots, l+u$ **do**
10: $\quad y_i = \arg\max_{1 \le j \le |\mathcal{Y}|} (\mathbf{F}^*)_{ij}$
11: **end for**

**输出:** 未标记样本的预测结果: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

图 13.5 迭代式标记传播算法

# 图半监督学习

- 图半监督学习方法在概念上相当清晰，且易于通过对所涉矩阵运算的分析来探索算法性质。

- 但此类算法的缺陷也相当明显. 首先是在存储开销高。

- 另一方面，由于构图过程仅能考虑训练样本集，难以判知新样本在图中的位置，因此，在接收到新样本时，或是将其加入原数据集对图进行重构并重新进行标记传播，或是需引入额外的预测机制。

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 基于分歧的方法

- 基于分歧的方法(disagreement-based methods)使用多学习器，而学习器之间的“分歧”(disagreement)对未标记数据的利用至关重要。

- 协同训练(co-training)[Blum and Mitchell, 1998]是基于分歧的方法的重要代表，它最初是针对“多视图”(multi-view)数据设计的，因此也被看作“多视图学习”(multi-view learning)的代表.

# 基于分歧的方法

*图片视图*

*文字视图*

周志华 中文简历 Brief CV

**Zhi-Hua Zhou** can be pronounced simply as [Jihua Joe]

**Professor,** Department of Computer Science & Technology, Nanjing University, China

ACM Distinguished Scientist, IEEE Fellow, IAPR Fellow, CCF Fellow

| **Correspondence** | | | 邮政快递地址 |
|---|---|---|---|
| **Mail:** | Zhi-Hua Zhou | **Office:** | Rm 920, Computer Science Building, Nanjing University Xianlin Campus |
| | National Key Laboratory for Novel Software Technology | **Tel:** | +86-25-8968-6268 |
| | Nanjing University, Xianlin Campus Mailbox 603 | **Fax:** | +86-25-8968-6268 |
| | 163 Xianlin Avenue, Qixia District | **URL:** | http://cs.nju.edu.cn/zhouzh/ |
| | Nanjing 210023, China | **Email:** | zhouzh at nju dot edu dot cn or zhouzh at lamda dot nju dot edu dot cn or **zhouzh.gm at gmail.com** *(you may want to contact me using my Gmail account as our university spam setting might be grim)* |

[Interest] [Career] {Education} [Award] [Activity] {Publication} [Course] {Student and Postdoc} {LAMDA Group}

**Research Interest**

I have wide research interests, mainly including *artificial intelligence*, *machine learning*, *data mining*, *pattern recognition*, *evolutionary computation* and *multimedia retrieval*, among which **machine learning** and **data mini**
problem of how to enable computing machines to handle "ambiguity"

Currently I am interested in the following ML/DM topics:

- Multi-label learning
- Multi-instance learning
- Semi-supervised and active learning
- Cost-sensitive and class-imbalance learning
- Metric learning, dimensionality reduction and feature selection
- Ensemble learning
- Structure learning and clustering

For applications, I am mainly interested in the following areas:

- Image retrieval

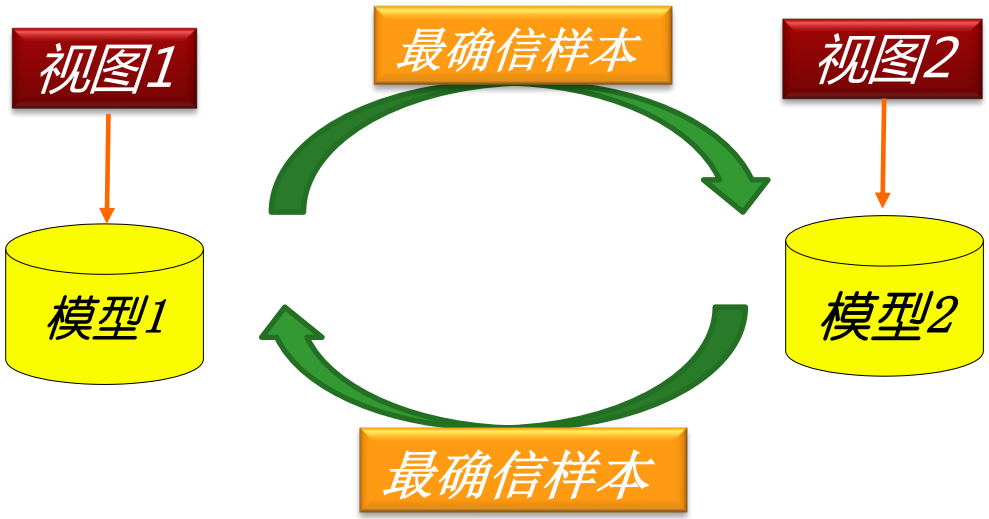


Z.-H. Zhou. Ensemble Methods: Foundations and Algorithms, Boca Raton, FL: Chapman & Hall/CRC, 2012. (ISBN 978-1-439-830031)

网页分类任务中的双视图

# 基于分歧的方法

- 协同训练正是很好地利用了多视图的“相容互补性”. 假设数据拥有两个“充分”(sufficient)且“条件独立”视图。



# 基于分歧的方法

$\boldsymbol{x}_i$ 的上标仅用于指代两个视图，不表示序关系，即 $\langle \boldsymbol{x}_i^1, \boldsymbol{x}_i^2 \rangle$ 与 $\langle \boldsymbol{x}_i^2, \boldsymbol{x}_i^1 \rangle$ 表示的是同一个样本.

令 $p, n \ll s$.

**输入:** 有标记样本集 $D_l = \{(\langle \boldsymbol{x}_1^1, \boldsymbol{x}_1^2 \rangle, y_1), \ldots, (\langle \boldsymbol{x}_l^1, \boldsymbol{x}_l^2 \rangle, y_l)\}$;
未标记样本集 $D_u = \{\langle \boldsymbol{x}_{l+1}^1, \boldsymbol{x}_{l+1}^2 \rangle, \ldots, \langle \boldsymbol{x}_{l+u}^1, \boldsymbol{x}_{l+u}^2 \rangle\}$;
缓冲池大小 $s$;
每轮挑选的正例数 $p$;
每轮挑选的反例数 $n$;
基学习算法 $\mathfrak{L}$;
学习轮数 $T$.

**过程:**

1: 从 $D_u$ 中随机抽取 $s$ 个样本构成缓冲池 $D_s$;
2: $D_u = D_u \setminus D_s$;

初始化每个视图上的有标记训练集.

3: **for** $j = 1, 2$ **do**
4: $\quad D_l^j = \{(\boldsymbol{x}_i^j, y_i) \mid (\langle \boldsymbol{x}_i^j, \boldsymbol{x}_i^{3-j} \rangle, y_i) \in D_l\}$;
5: **end for**

在视图 $j$ 上用有标记样本训练 $h_j$.

6: **for** $t = 1, 2, \ldots, T$ **do**
7: $\quad$ **for** $j = 1, 2$ **do**
8: $\quad\quad h_j \leftarrow \mathfrak{L}(D_l^j)$;
9: $\quad\quad$ 考察 $h_j$ 在 $D_s^j = \{\boldsymbol{x}_i^j \mid \langle \boldsymbol{x}_i^j, \boldsymbol{x}_i^{3-j} \rangle \in D_s\}$ 上的分类置信度，挑选 $p$ 个正例置信度最高的样本 $D_p \subset D_s$、$n$ 个反例置信度最高的样本 $D_n \subset D_s$;
10: $\quad\quad$ 由 $D_p^j$ 生成伪标记正例 $\tilde{D}_p^{3-j} = \{(\boldsymbol{x}_i^{3-j}, +1) \mid \boldsymbol{x}_i^j \in D_p^j\}$;
11: $\quad\quad$ 由 $D_n^j$ 生成伪标记反例 $\tilde{D}_n^{3-j} = \{(\boldsymbol{x}_i^{3-j}, -1) \mid \boldsymbol{x}_i^j \in D_n^j\}$;
12: $\quad\quad D_s = D_s \setminus (D_p \bigcup D_n)$;
13: $\quad$ **end for**
14: $\quad$ **if** $h_1, h_2$ 均未发生改变 **then**
15: $\quad\quad$ **break**
16: $\quad$ **else**

扩充有标记数据集.

17: $\quad\quad$ **for** $j = 1, 2$ **do**
18: $\quad\quad\quad D_l^j = D_l^j \bigcup \left(\tilde{D}_p^j \bigcup \tilde{D}_n^j\right)$;
19: $\quad\quad$ **end for**
20: $\quad\quad$ 从 $D_s$ 中随机抽取 $2p + 2n$ 个样本加入 $D_s$
21: $\quad$ **end if**
22: **end for**

**输出:** 分类器 $h_1, h_2$

图 13.6 协同训练算法

# 基于分歧的方法

- 协同训练过程虽简单，但令人惊讶的是，理论证明显示出，若两个视图充分且条件独立，则可利用未标记样本通过协同训练将弱分类器的泛化性能提升到任意高[Blum and Mitchell, 1998].

- 不过，视图的条件独立性在现实任务中通常很难满足,因此性能提升幅度不会那么大，但研究表明，即使在更弱的条件下，协同训练仍可有效地提升弱分类器的性能[周志华，2013].

# 基于分歧的方法

- 协同训练算法本身是为多视图数据而设计的，但此后出现了一些能在单视图数据上使用的变体算法。

- 它们或是使用不同的学习算法[Goldman and Zhou,2000]、或使用不同的数据采样[Zhou and Li，2005b]、甚至使用不同的参数设置[Zhou and Li，2005a]来产生不同的学习器，也能有效地利用未标记数据来提升性能。

- 后续理论研究发现，此类算法事实上无需数据拥有多视图，仅需弱学习器之间具有显著的分歧(或差异)，即可通过相互提供伪标记样本的方式来提高泛化性能[周志华，2013]。

# 基于分歧的方法

- 基于分歧的方法只需采用合适的基学习器，就能较少受到模型假设、损失函数非凸性和数据规模问题的影响，学习方法简单有效、理论基础相对坚实、适用范围较为广泛。

- 为了使用此类方法，需能生成具有显著分歧、性能尚可的多个学习器，但当有标记样本很少、尤其是数据不具有多视图时，要做到这一点并不容易。

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 半监督聚类

- 聚类是一种典型的无监督学习任务，然而在现实聚类任务中我们往往能获得一些额外的监督信息，于是可通过“半监督聚类”(semi-supervised clustering)来利用监督信息以获得更好的聚类效果。

- 聚类任务中获得的监督信息大致有两种类型：

- 第一种类型是“ 必连”(must-link)与“勿连”(cannot-link)约束，前者是指样本必属于同一个簇，后者则是指样本必不属于同一个簇；
- 第二种类型的监督信息则是少量的有标记样本.

# 半监督聚类

- 约束$k$均值(Constrained $k$-means)算法[Wagstaff et al., 2001]是利用第一类监督信息的代表。

- 该算法是$k$均值算法的扩展,它在聚类过程中要确保“必连”关系集合与“勿连”关系集合中的约束得以满足，否则将返回错误提示。

# 半监督聚类

输入：样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
必连约束集合 $\mathcal{M}$;
勿连约束集合 $\mathcal{C}$;
聚类簇数 $k$.

```
8:    while ¬ is_merged do
9:      基于 K 找出与样本 x_i 距离最近的簇: r = arg min_{j∈K} d_ij ;
10:     检测将 x_i 划入聚类簇 C_r 是否会违背 M 与 C 中的约束;
11:     if ¬ is_voilated then
12:       C_r = C_r ⋃ {x_i};
13:       is_merged=true
14:     else
15:       K = K \ {r};
16:       if K = ∅ then
17:         break并返回错误提示
18:       end if
19:     end if
20:   end while
```

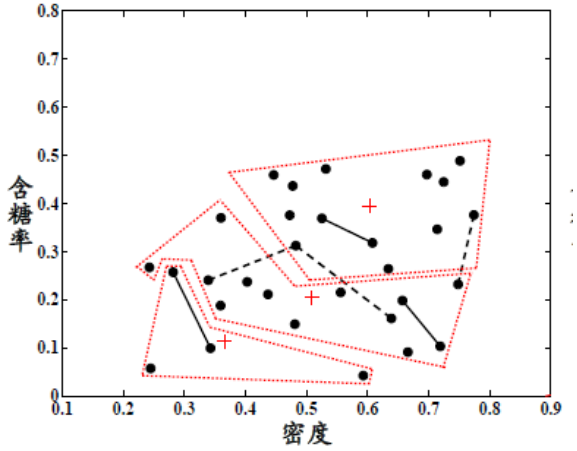
不冲突，选择最近的簇

冲突，尝试次近的簇

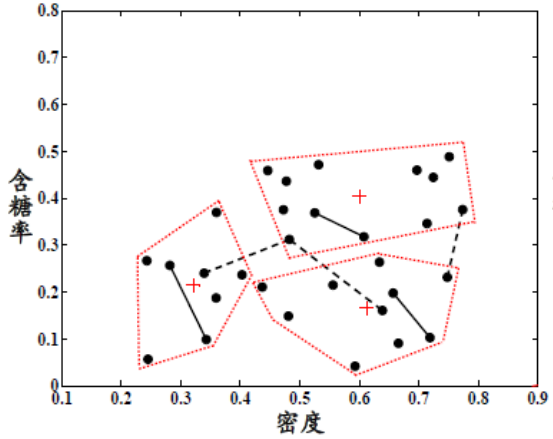
24: end for
25: until 均值向量均未更新
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.7 约束 $k$ 均值算法

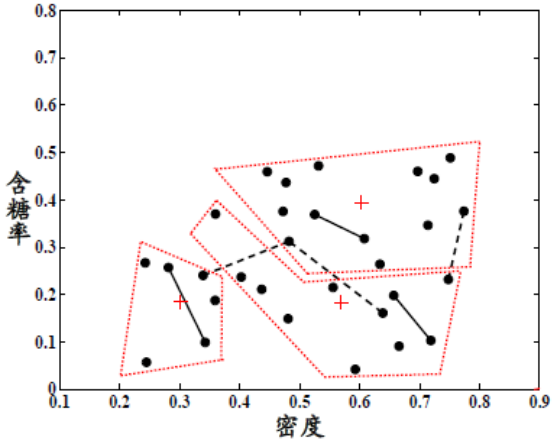
# 半监督聚类



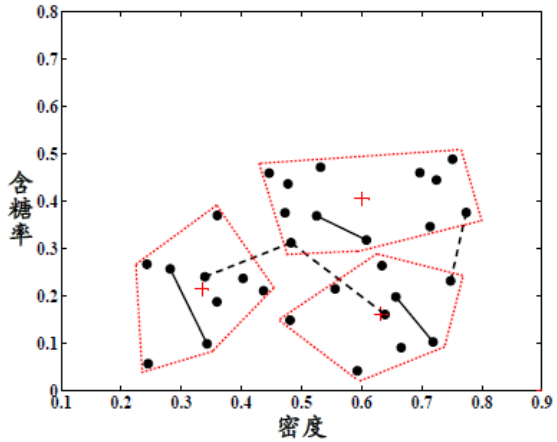

(a) 第 1 轮迭代后




(c) 第 3 轮迭代后




(b) 第 2 轮迭代后




(d) 第 4 轮迭代后

# 半监督聚类

- 第二种监督信息是少量有标记样本。即假设少量有标记样本属于k个聚类簇。

- 这样的监督信息利用起来很容易：直接将它们作为“种子”，用它们初始化$k$均值算法的$k$个聚类中心，并且在聚类簇迭代更新过程中不改变种子样本的簇隶属关系. 这样就得到了约束种子$k$均值(Constrained Seed $k$-means)算法[Basu et al., 2002]。

# 半监督聚类

输入：样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$；

用有标记样本初始化簇中心.

```
for j = 1, 2, ..., k do
    μ_j = (1/|S_j|) Σ_{x∈S_j} x
end for
```

用有标记样本初始化 $k$ 个簇.

```
for j = 1, 2, ..., k do
    for all x ∈ S_j do
        C_j = C_j ⋃ {x}
    end for
```

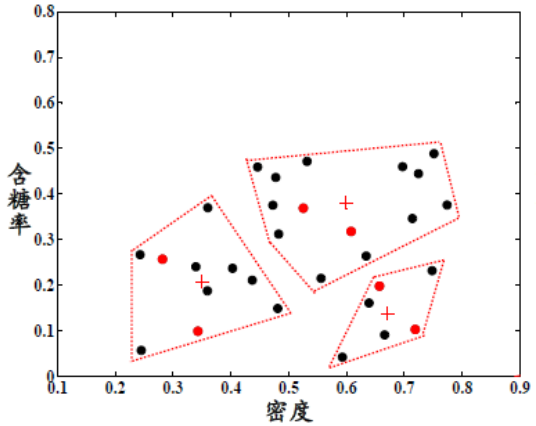
```
    找出与样本 x_i 距离最近的簇: r = arg min_{j∈{1,2,...,k}} d_ij ;
    将样本 x_i 划入相应的簇: C_r = C_r ⋃ {x_i}
end for
for j = 1, 2, ..., k do
    μ_j = (1/|C_j|) Σ_{x∈C_j} x;
end for
until 均值向量均未更新
```

更新均值向量.

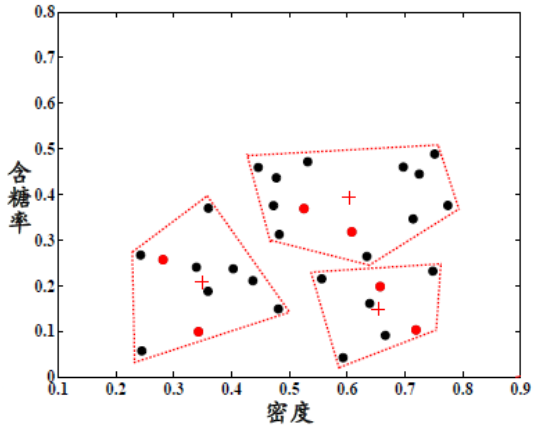
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.9　约束种子 $k$ 均值算法

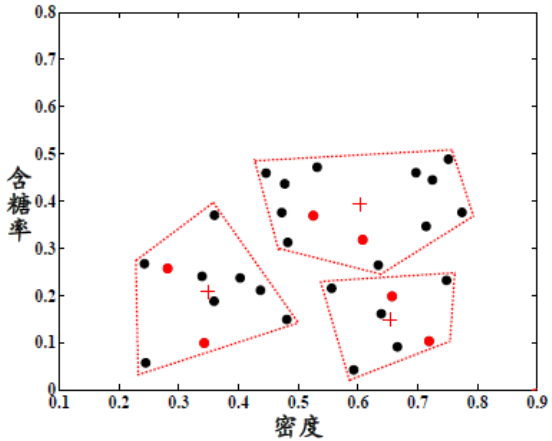
# 半监督聚类



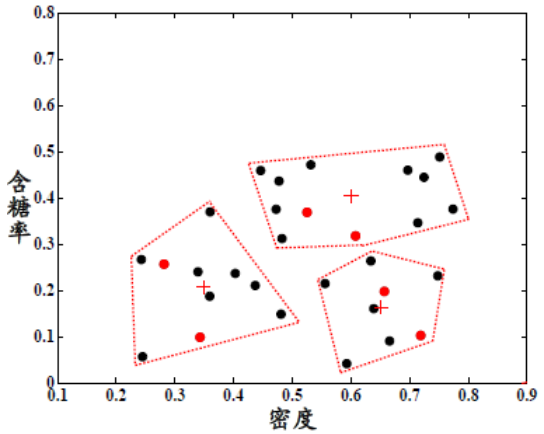

(a) 第 1 轮迭代后




(b) 第 2 轮迭代后




(b) 第 2 轮迭代后




(d) 第 4 轮迭代后

# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 半监督深度学习

深度学习，需要用到大量的有标记数据，即使在大数据时代，干净能用的有标记数据也是不多的

近年来，深度学习与半监督学习思想相结合，产生的半监督深度学习已成为深度学习领域热门的新方向。

包括MIT、Stanford、Google Brain、Facebook等学术界和工业界，在半监督深度学习领域做了大量的工作。

# 半监督深度学习

早期的半监督深度学习算法，主要分为两类：

1）无标签数据预训练网络，然后有标签数据进行微调(fine-tune)

对于神经网络来说，一个好的初始化可以是结果更稳定，迭代次数更少；

这类算法主要关注如何利用无标签数据让网络有一个好的初始化。

代表算法：无监督预训练

# 半监督深度学习

早期的半监督深度学习算法，主要分为两类：

2）利用网络得到的深度特征来做半监督算法：

- 先用有标签数据训练网络（此时网络一般过拟合...）
- 通过隐藏层提取特征，以这些特征来用某种分类算法对无标签数据进行分类
- 挑选认为分类正确的无标签数据加入到训练集
- 重复上述过程

# 半监督深度学习

以上两种方法虽然都用了有标签数据和无标签数据，但就神经网络本身的训练而言，其实还是运行在一种有监督的方式上

而近年来的半监督深度学习研究主要关注，如何让深度学习真正的成为一种半监督算法，

即，端到端的训练半监督深度学习模型

# 半监督深度学习

以上两种方法虽然都用了有标签数据和无标签数据，但就神经网络本身的训练而言，其实还是运行在一种有监督的方式上

而近年来的半监督深度学习研究主要关注，如何让深度学习真正的成为一种半监督算法，

即，端到端的训练半监督深度学习模型

# 两个基本方法

- 一致性正则 (Consistency Regularization)

- 熵最小化(Entropy Minimization)

# 一致性正则

半监督深度学习的一个重要研究方向是利用未标记的数据来强化训练模型，使其符合聚类假设，即学习的决策边界必须位于低密度区域。

这些方法基于一个简单的概念，即如果对一个未标记的数据应用实际的扰动，则预测不应发生显著变化。

具体来说，给定一个未标记的数据样本及其扰动的形式，目标是最小化两个输出之间的距离：

$$d(f_\theta(x) \ , \quad f(\hat{x}))$$

# 一致性正则

流行的距离度量 $d(f_\theta(x), f(\hat{x}))$均方误差(Mean-Squared Error, MSE)，Kullback-Leiber散度 (KL Divergence) 和 Jensen-Shannon 散度(JS Divergence)

$$d_{\mathrm{MSE}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{C}\sum_{k=1}^{C}(f_\theta(x)_k - f_\theta(\hat{x})_k)^2$$

$$d_{\mathrm{KL}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{C}\sum_{k=1}^{C} f_\theta(x)_k \log \frac{f_\theta(x)_k}{f_\theta(\hat{x})_k}$$

$$d_{\mathrm{JS}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{2}d_{\mathrm{KL}}(f_\theta(x), m) + \frac{1}{2}d_{\mathrm{KL}}(f_\theta(\hat{x}), m)$$

# 一致性正则

基于一致性正则(Consistency Regularization)的思想，衍生出一批深度半监督学习算法，如：

Π-Model、Temporal Ensembling、Mean Teacher、VAT、UDA等

具体到每一种算法，核心思想是没有变化的，即最小化未标记数据与其扰动两者之间预测值的距离，主要区别在于：

1）进行数据扰动的方式不同

2）距离的计算方式不同

# Π-Model

## TEMPORAL ENSEMBLING FOR SEMI-SUPERVISED LEARNING

**Samuli Laine**
NVIDIA
slaine@nvidia.com

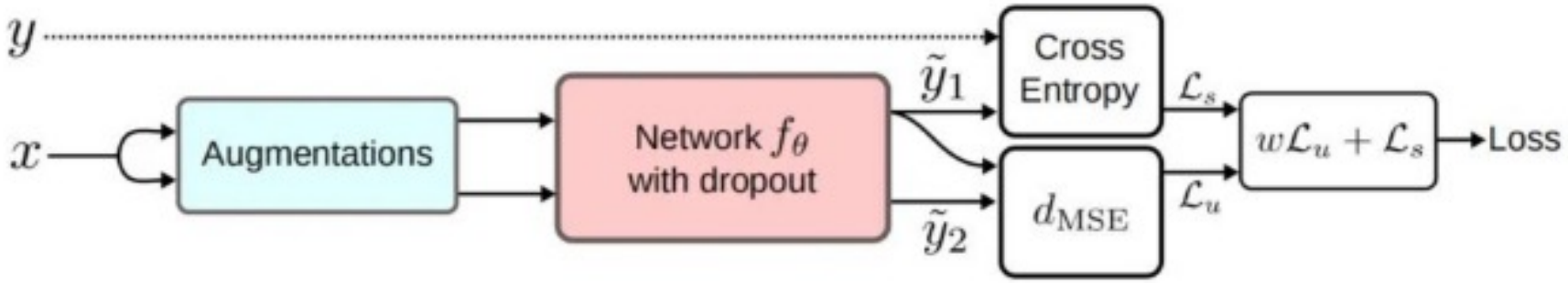
**Timo Aila**
NVIDIA
taila@nvidia.com

论文链接：
https://openreview.net/forum?id=BJ6oOfqge&noteId=BJ6oOfqge

代码链接：
https://github.com/smlaine2/tempens

# Π-Model

对给定的样本x，采用不同的数据增广，得到两次预测结果，目标是减小两次预测之间的距离，提升模型在不同扰动下的一致性



# Π-Model

数据扰动方式：随机反转、平移、剪切等策略

无监督损失：两次前向运算结果的均方误差(MSE)

对每一个参与训练的样本，在训练阶段，Pi-Model 需要进行两次前向计算,时间成本比较高。

# Temporal Ensembling

TEMPORAL ENSEMBLING FOR SEMI-SUPERVISED LEARNING

**Samuli Laine**
NVIDIA
slaine@nvidia.com

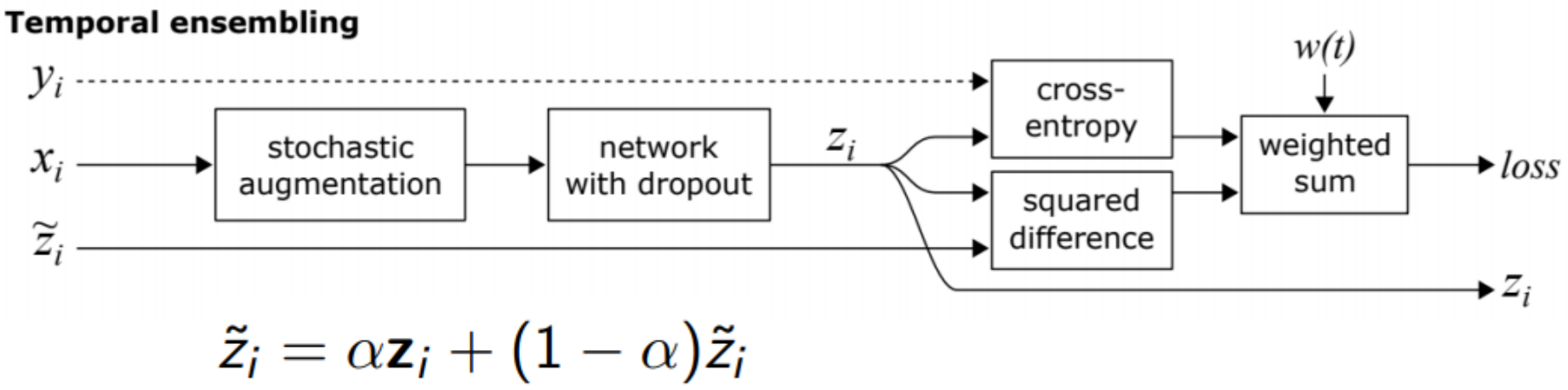
**Timo Aila**
NVIDIA
taila@nvidia.com

论文链接：
https://openreview.net/forum?id=BJ6oOfqge&noteId=BJ6oOfqge

代码链接：
https://github.com/smlaine2/tempens

# Temporal Ensembling

- Π-Model：最小化两次随机增广预测值的均方误差
- Temporal Ensembling：最小化当前模型预测结果与历史模型预测结果的平均值之间的均方误差



# Temporal Ensembling

数据扰动方式：当前轮样本的预测值和该样本在历史轮数上的预测值

无监督损失：两次预测值的均方误差(MSE)

✓ 用空间来换取时间，总的前向计算次数减少了一半
✓ 通过历史预测做平均，有利于减小单次预测中的噪声。

# Mean Teacher

**Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results**

**Antti Tarvainen**
The Curious AI Company
and Aalto University
antti.tarvainen@aalto.fi

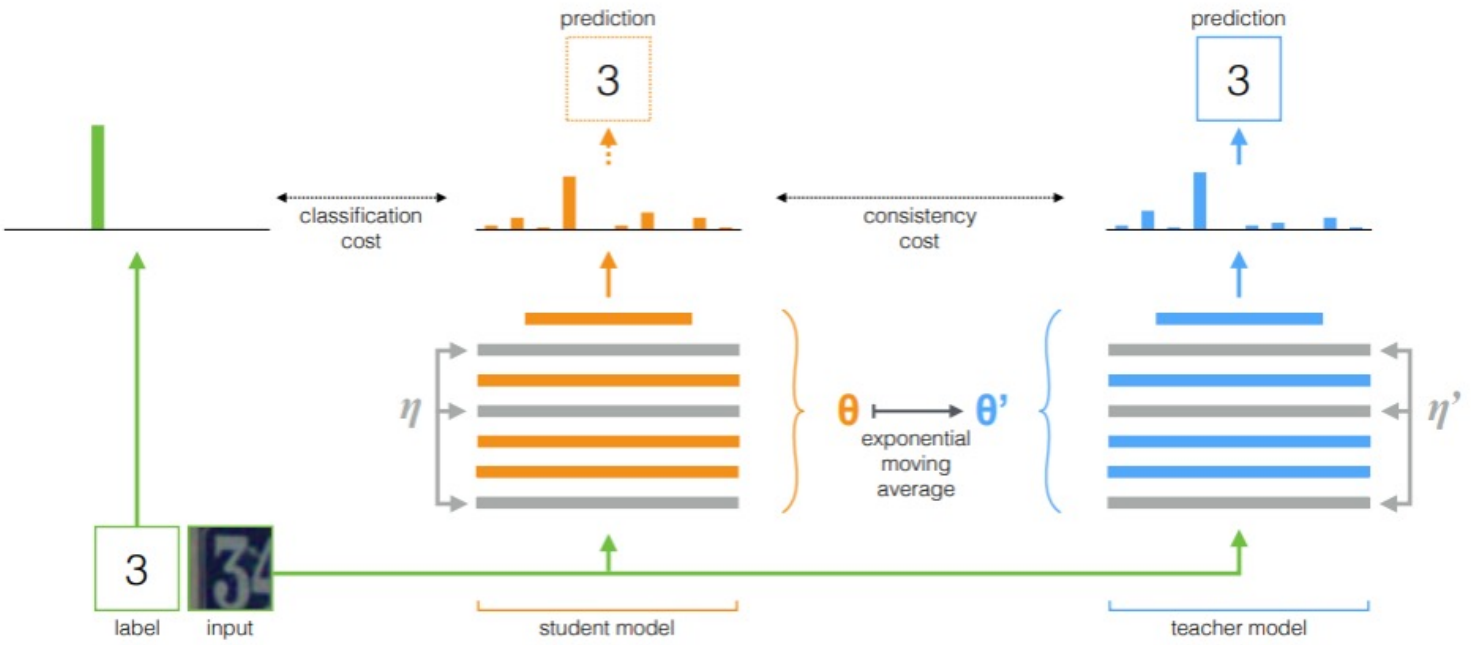
**Harri Valpola**
The Curious AI Company
harri@cai.fi

论文链接：
https://arxiv.org/abs/1703.01780

代码链接：
https://github.com/CuriousAI/mean-teacher

# Mean Teacher

- Temporal Ensembling：保存模型对样本的预测值
- Mean Teacher：**直接保存模型的权重**



# Mean Teacher

数据扰动方式：当前模型对该样本的预测值和历史轮数上模型的集成对该样本的预测值

无监督损失：两次预测值的均方误差(MSE)

保存模型的参数比保存历史预测值更一般化，性能也会更好，但是会带来更大的存储消耗。

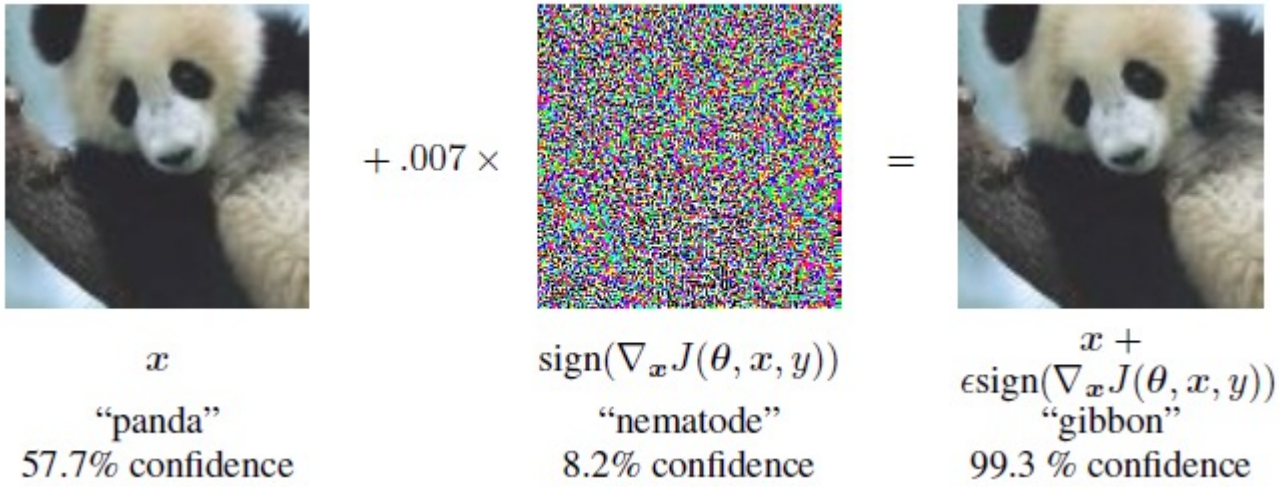
# Virtual Adversarial Training (VAT)

## Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning

Takeru Miyato[*,†,‡], Shin-ichi Maeda[*,†], Masanori Koyama[§,†] and Shin Ishii[†,‡]

论文链接：
https://arxiv.org/abs/1704.03976

# 对抗样本




原有的模型以57.7%的置信度判定图片为熊猫，但添加微小的扰动后，模型以99.3%的置信度认为扰动后的图片是长臂猿。

对抗样本可以让训练优秀的分类网络进行错误的分类，然而人类去看对抗样本的话和真实的样本几乎无异

# Virtual Adversarial Training (VAT)

VAT算法提出对样本添加对抗噪声，然后让模型对原始样本和对抗样本的预测值尽可能的接近

对抗扰动：

$$\mathbf{r}_{adv}(\mathbf{x}) = \arg\max_{||\mathbf{r}||_2 \leq \epsilon} D[p_\theta(\mathbf{y}|\mathbf{x}), p_\theta(\mathbf{y}|\mathbf{x}+\mathbf{r})]$$

# Virtual Adversarial Training (VAT)

无监督损失

最小化原始样本预测值和对抗样本预测值之间的KL Divergence

$$\sum_x D[p_\theta(\mathbf{y}|\mathbf{x}), p_\theta(\mathbf{y}|\mathbf{x}+\mathbf{r}_{adv})]$$

✓ 相比以往的方法，提升了半监督深度学习模型在对抗扰动下的鲁棒性

# Unsupervised Data Augmentation

**Unsupervised Data Augmentation for Consistency Training**

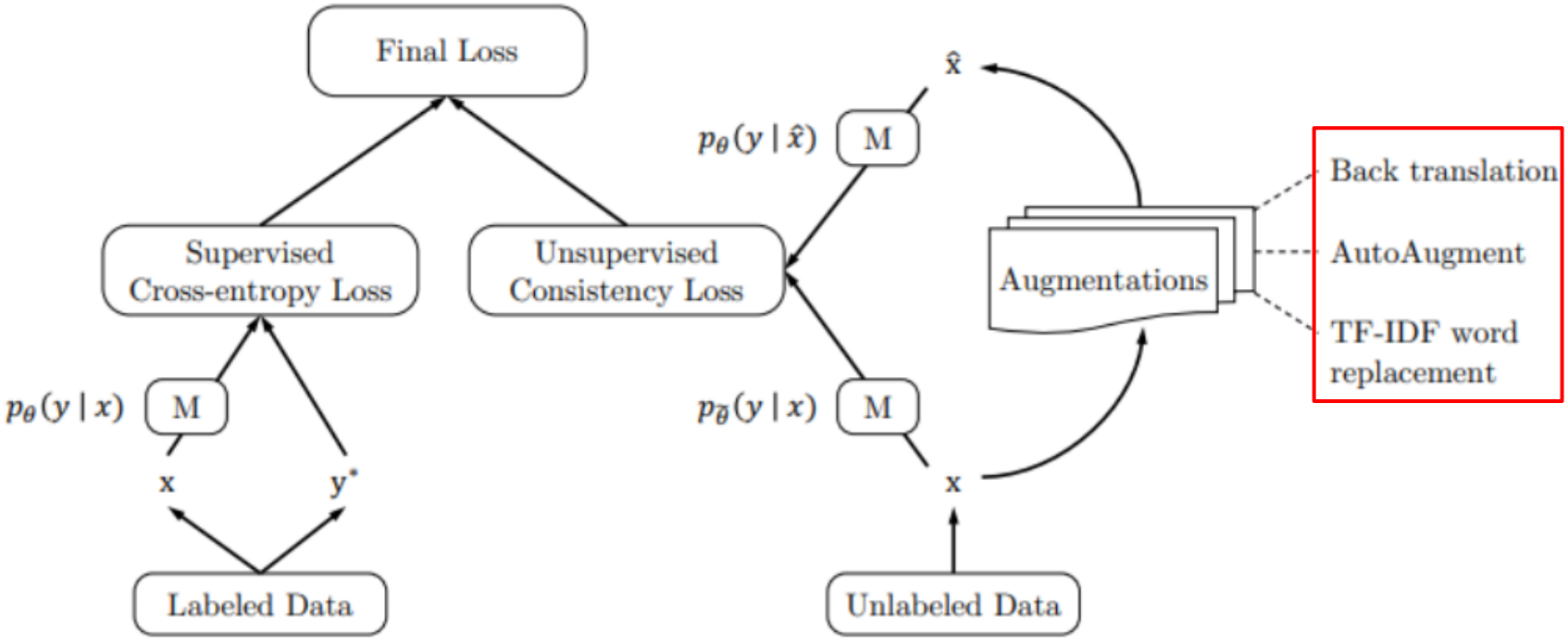
**Qizhe Xie**[1,2], **Zihang Dai**[1,2], **Eduard Hovy**[2], **Minh-Thang Luong**[1], **Quoc V. Le**[1]
[1] Google Brain, [2] Carnegie Mellon University
{qizhex, dzihang, hovy}@cs.cmu.edu, {thangluong, qvl}@google.com

论文链接：
https://arxiv.org/pdf/1904.12848v2.pdf

代码链接：
https://github.com/google-research/uda

# Unsupervised Data Augmentation

该工作提出，对数据进行增广的方式不应该是一成不变的，需要对数据采取更多样化，更领域相关的数据增强方式



# Unsupervised Data Augmentation

无监督损失

最小化未标记数据和增强未标记数据上预测分布之间的 KL Divergence

$$\min_{\theta} \mathcal{J}_{\mathrm{UDA}}(\theta) = \mathop{\mathbb{E}}_{x \in U} \mathop{\mathbb{E}}_{\hat{x} \sim q(\hat{x}|x)} \left[ \mathcal{D}_{\mathrm{KL}} \left( p_{\tilde{\theta}}(y \mid x) \,\|\, p_{\theta}(y \mid \hat{x}) \right) \right]$$

✓ UDA证明了针对性的数据增强效果明显优于无针对性的数据增强

# 基于一致性正则的方法小结

一致性正则(Consistency Regularization)这类方法的主要思想：

- **对于无标记样本，添加噪声后模型的预测值也应该尽可能保持不变**

各方法的主要区别在于如何找到更适合的数据增广

没有哪种数据增广是万能的，数据增广方法不应该是一成不变的，要结合领域知识，不同的任务采用不同的增广方法！

# 两个基本方法

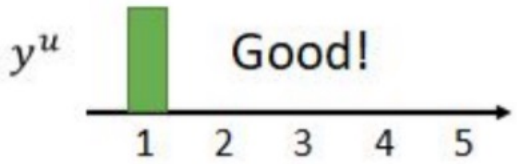
- 一致性正则 (Consistency Regularization)

- 熵最小化(Entropy Minimization)

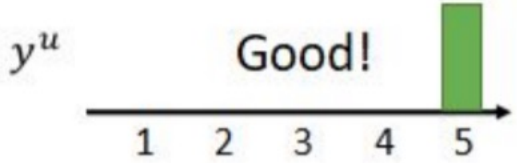
# 熵最小化(Entropy Minimization)

熵最小化：鼓励模型输出的预测值置信度尽可能高


Distribution




$E(y^u) = 0$




$E(y^u) = 0$


$$E(y^u) = -ln\left(\frac{1}{5}\right) = ln5$$

Entropy of $y^u$ :
Evaluate how concentrate the distribution $y^u$ is

$$E(y^u) = -\sum_{m=1}^{5} y_m^u ln(y_m^u)$$

As small as possible

$$L = \sum_{x^r} C(y^r, \hat{y}^r)$$

labelled data

$$+\lambda \sum_{x^u} E(y^u)$$

unlabeled data

# Self-Training

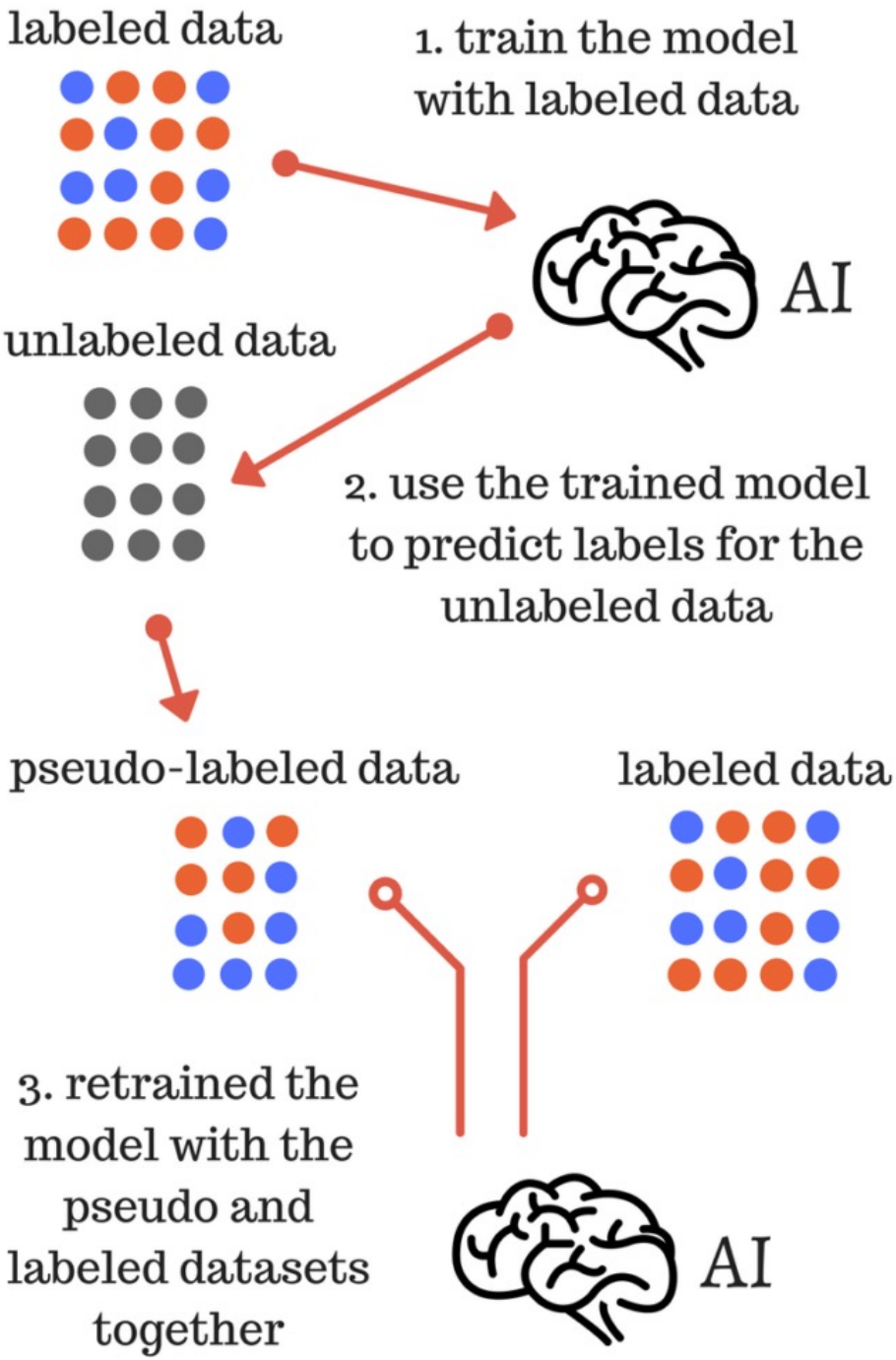
自训练（self-training）

用有标签数据训练一个分类器，然后用这个分类器对无标签数据进行分类，这样就会产生伪标签（pseudo label）或软标签（soft label），挑选你认为分类正确的无标签样本（此处应该有一个挑选准则），把选出来的无标签样本用来训练分类器

# Self-Training

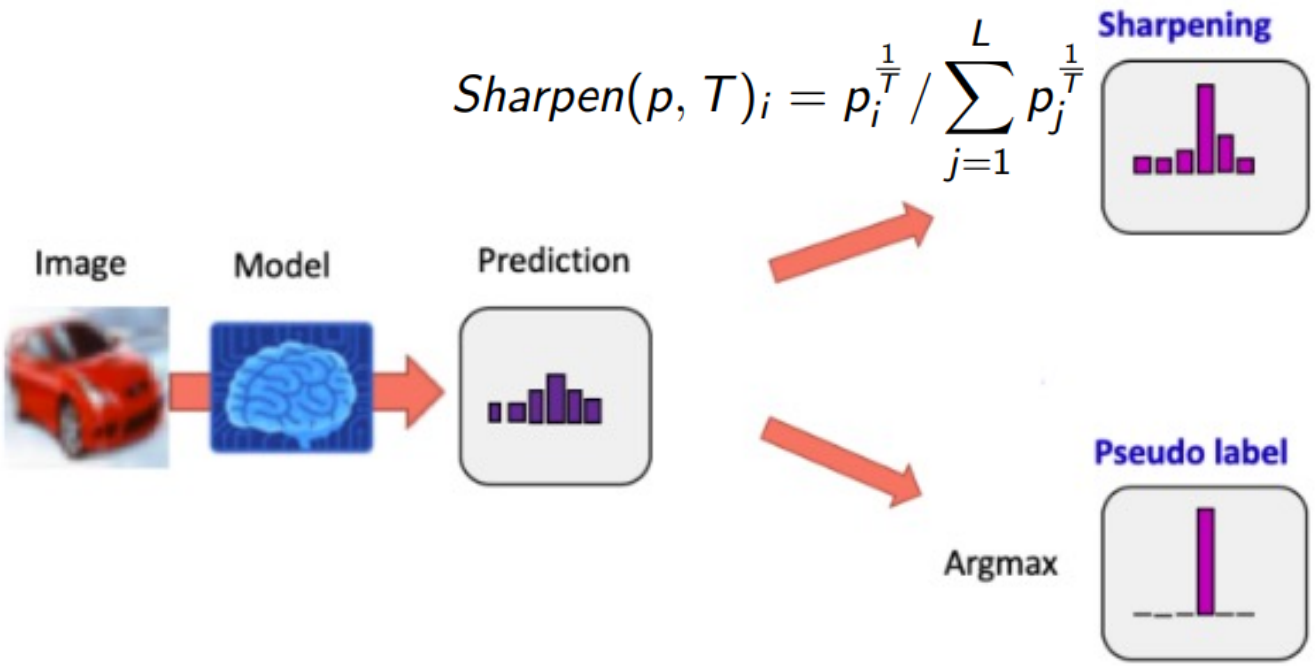
自训练（self-training）

用有标签数据训练一个分 数据
进行分类，这样就会产生
（soft label），挑选你 应该
有一个挑选准则），把选



# Self-Training

如何分配伪标签：最受欢迎的两种是锐化（sharpening）方法和 Argmax 方法



# Self-Training

如何分配伪标签：最受欢迎的两种是锐化（sharpening）方法和 Argmax 方法

前者在保持预测值分布的同时使分布有些极端；
后者仅使用对预测具有最高置信度的预测标签进行标记。

我们还可以对无标签数据进行过滤，如果预测结果大于预定阈值 τ，再将其添加训练中。

# Self-Training

**Step1：** 首先，用少量的标签数据 L 训练模型;
**Step2：** 然后，使用训练后的模型给未标记数据 x∈U 分配 Pseudo-label（伪标签）;
**Setp3：** 通过交叉熵损失计算模型预测和伪标签的损失。
**Step4：** 使用训练好的模型为 U 的其余部分生成标签.

**Algorithm 1** Self-training

```
repeat
    m ← train_model(L)
    for x ∈ U do
        if max m(x) > τ then
            L ← L ∪ {(x, p(x))}
until no more predictions are confident
```

# Holistic Methods

一致性正则和熵最小化各有优劣，

能不能同时考虑两种正则化方法呢？

Google Brain提出MixMatch、FixMatch算法，称为Holistic Method，

即试图在一个框架中整合当前的 SSL 的主要方法，从而获得更好的性能。

# MixMatch

**MixMatch: A Holistic Approach to Semi-Supervised Learning**


**David Berthelot**
Google Research
dberth@google.com

**Nicholas Carlini**
Google Research
ncarlini@google.com

**Ian Goodfellow**
Work done at Google
ian-academic@mailfence.com

**Avital Oliver**
Google Research
avitalo@google.com

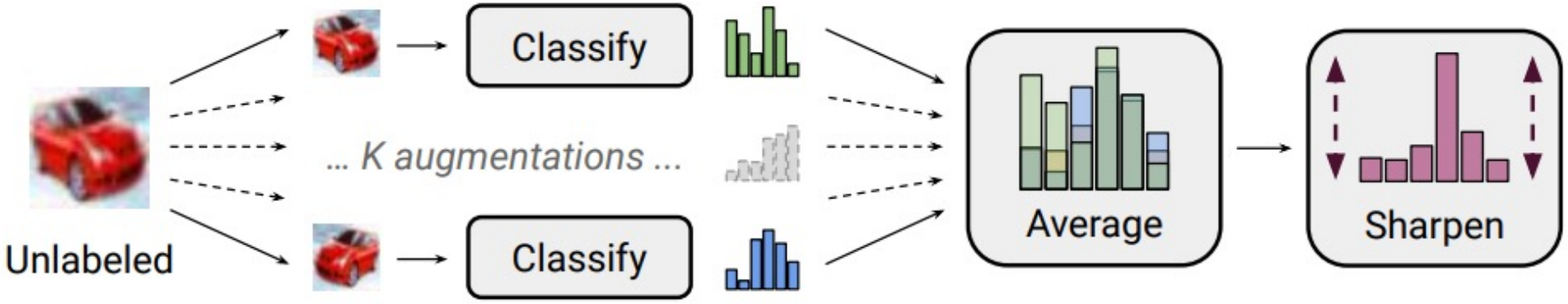
**Nicolas Papernot**
Google Research
papernot@google.com

**Colin Raffel**
Google Research
craffel@google.com


论文链接：
https://arxiv.org/pdf/1905.02249.pdf

代码链接：
https://github.com/google-research/mixmatch

# MixMatch




对样本做k次增广，然后将预测值平均，再*进行锐化操作*

无监督损失：上述操作得到的预测值和模型直接预测值之间的均方误差

# FixMatch

**FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence**

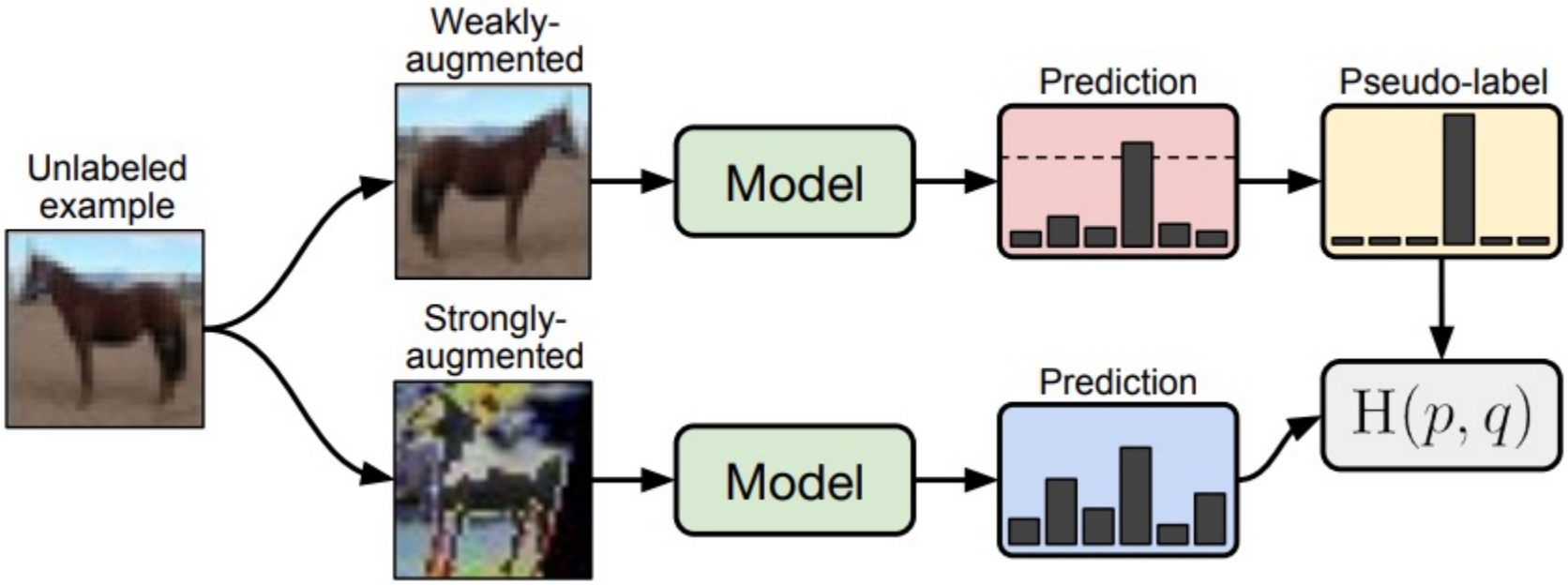
**Kihyuk Sohn*  David Berthelot*  Chun-Liang Li  Zizhao Zhang  Nicholas Carlini**
**Ekin D. Cubuk  Alex Kurakin  Han Zhang  Colin Raffel**
Google Research
{kihyuks,dberth,chunliang,zizhaoz,ncarlini,
cubuk,kurakin,zhanghan,craffel}@google.com

论文链接：
https://arxiv.org/ftp/arxiv/papers/2001/2001.07685.pdf

代码链接：
https://github.com/google-research/fixmatch

# FixMatch

- MixMatch 对预测值进行锐化操作
- FixMatch 利用Argmax的方式得到伪标记，并且只有预测结果大于预定阈值 τ，才将该样本加入训练。



# FixMatch

两种增强

- 弱增强：用标准的翻转和平移策略。
- 强增强：输出严重失真的输入图像。

无监督损失：

利用弱增强的样本得到伪标记后，计算模型预测值和伪标记之间的交叉熵(和监督损失一致)

# FixMatch

**Algorithm 1** FixMatch algorithm.

**Input:** Labeled batch $\mathcal{X} = \{(x_b, p_b) : b \in (1, \ldots, B)\}$, unlabeled batch $\mathcal{U} = \{u_b : b \in (1, \ldots, \mu B)\}$, confidence threshold $\tau$, unlabeled data ratio $\mu$, unlabeled loss weight $\lambda_u$.

$\ell_s = \frac{1}{B} \sum_{b=1}^{B} \mathrm{H}(p_b, \alpha(x_b))$ {*Cross-entropy loss for labeled data*}

**for** $b = 1$ **to** $\mu B$ **do**

  $q_b = p_{\mathrm{m}}(y \mid \alpha(u_b); \theta)$ {*Compute prediction after applying weak data augmentation of* $u_b$}

**end for**

$\ell_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}\{\max(q_b) > \tau\} \mathrm{H}(\arg\max(q_b), p_{\mathrm{m}}(y \mid \mathcal{A}(u_b))$ {*Cross-entropy loss with pseudo-label and confidence for unlabeled data*}

**return** $\ell_s + \lambda_u \ell_u$

# 半监督深度学习小结

一致性正则化方法通过鼓励无标签数据扰动前后的预测相同，增加了模型对数据变化的鲁棒性，减缓了标记数据不足时容易过拟合的问题

熵最小化的方法主要是通过对未标记数据制作满足熵最小化的伪标签然后加入训练，以得到更好的决策边界，

而众多方法中，混合方法表现出了良好的性能，是近来的研究热点。

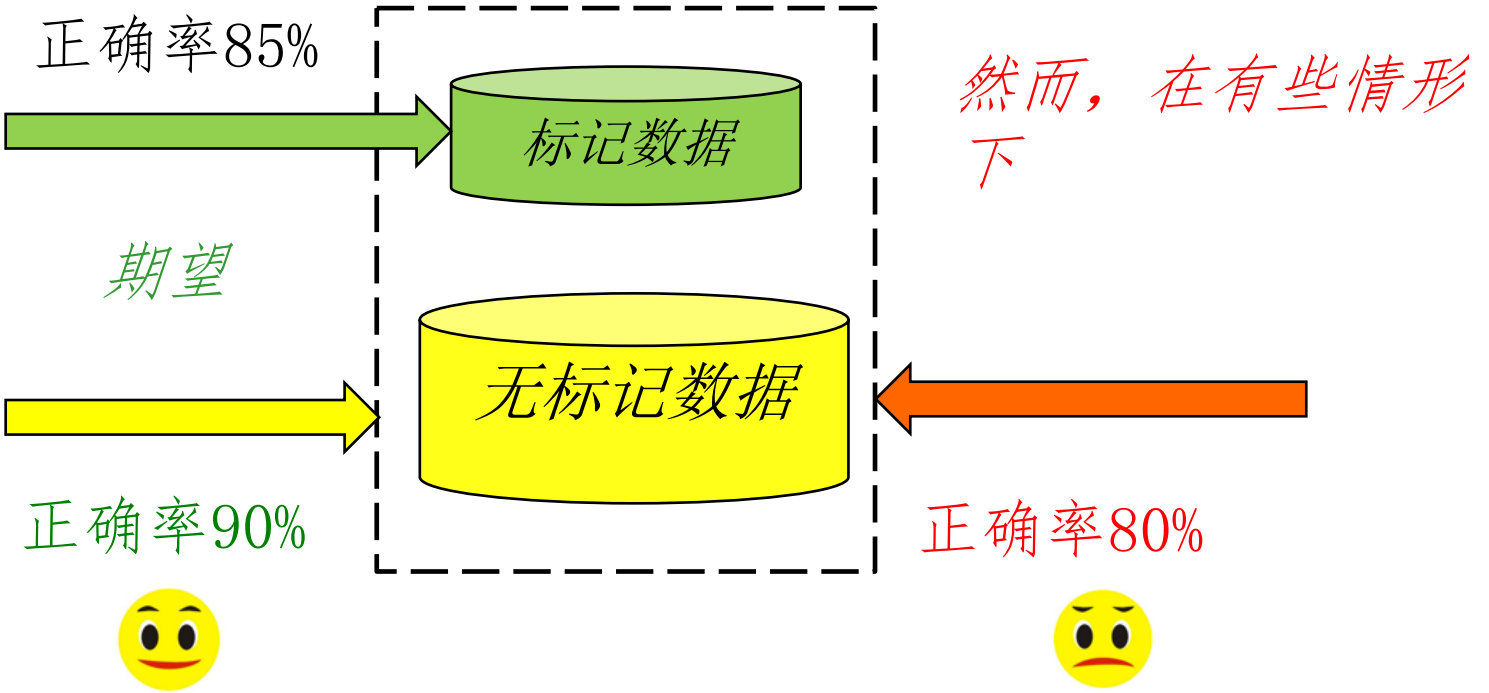
# 大纲

- 未标记样本
- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类
- 扩展-深度半监督学习
- 扩展-安全半监督学习

# 安全半监督学习

半监督学习试图利用更多的未标记数据提升模型性能

然而，半监督学习再利用未标记样本后并非必然提升泛化性能，在有些情形下真实会导致性能下降



# 安全半监督学习

对生成式方法，一个可能的成因是模型假设不准确，因此需要依赖充分可靠的领域知识来设计模型

对半监督SVM，一个可能的成因是训练数据中存在多个“低密度划分”，而学习算法有可能做出不利的选择

安全半监督学习关注的是，如何保证模型在利用了更多的未标记数据后性能相比简单的监督学习不下降

# 安全半监督学习

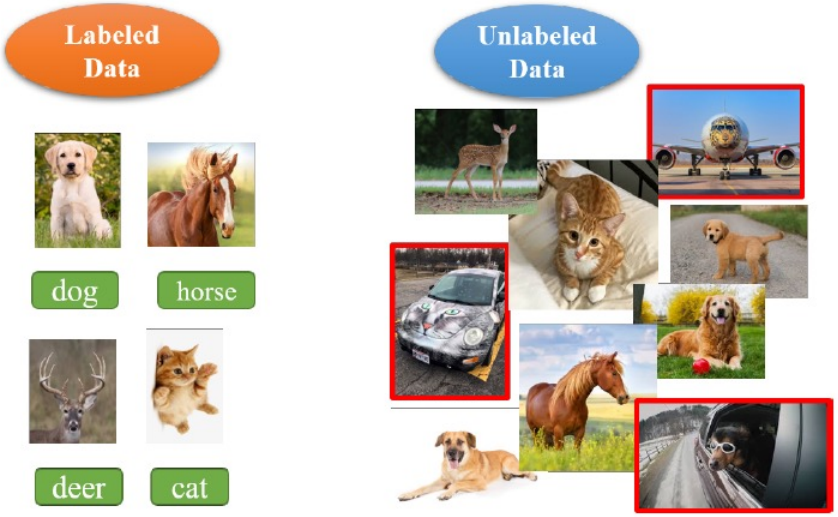
比如，对于半监督SVM，Li and Zhou 提出S4VM算法，通过优化最坏情形下的性能来综合利用多个低密度划分，提升了此类技术的安全性。

但是，更一般的安全(safe)半监督学习仍然是一个悬而未决的问题。

[1] Yu-Feng Li, Zhi-Hua Zhou. Towards making unlabeled data never hurt[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(1): 175-188.

# 类别不匹配的半监督深度学习

Oliver等人指出，当标记数据和无标记数据的类别分布不匹配时，半监督深度学习模型会遇到性能下降的问题




*未标记数据中含有未见类*

[1] Oliver, Avital, et al. "Realistic evaluation of deep semi-supervised learning algorithms." Advances in neural information processing systems 31 (2018): 3235-3246.

# 类别不匹配的半监督深度学习

如何提升半监督深度学习在类别分布不匹配时的安全性，已经成为了半监督深度学习的热点话题

Chen等人提出根据无标记数据预测值的置信度进行无标记数据的筛选

Guo等人提出利用标记数据的性能对未标记样本进行赋权

基本思想都是如何设计准则来实现选择性的利用未标记数据

[1]Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In AAAI, 2020.
[2]Guo, Lan-Zhe, et al. "Safe deep semi-supervised learning for unseen-class unlabeled data." International Conference on Machine Learning. PMLR, 2020.