机器学习

09 半监督学习

李祎

liyi@dlut.edu.cn

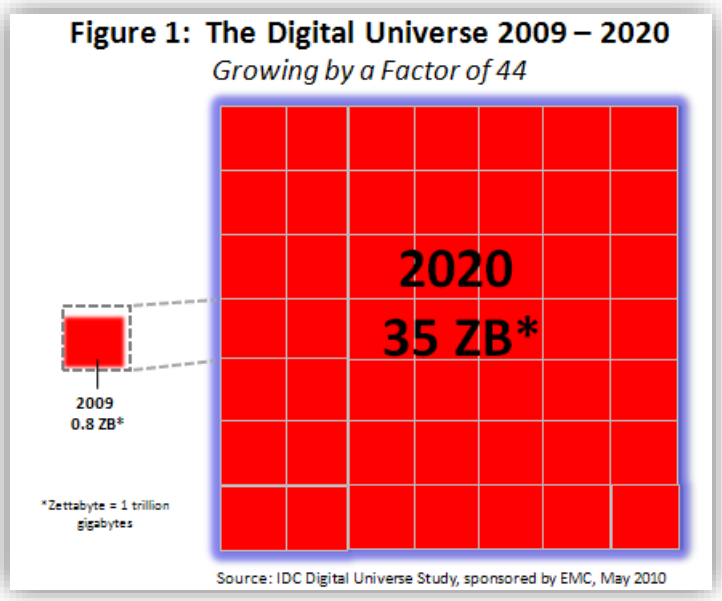
大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

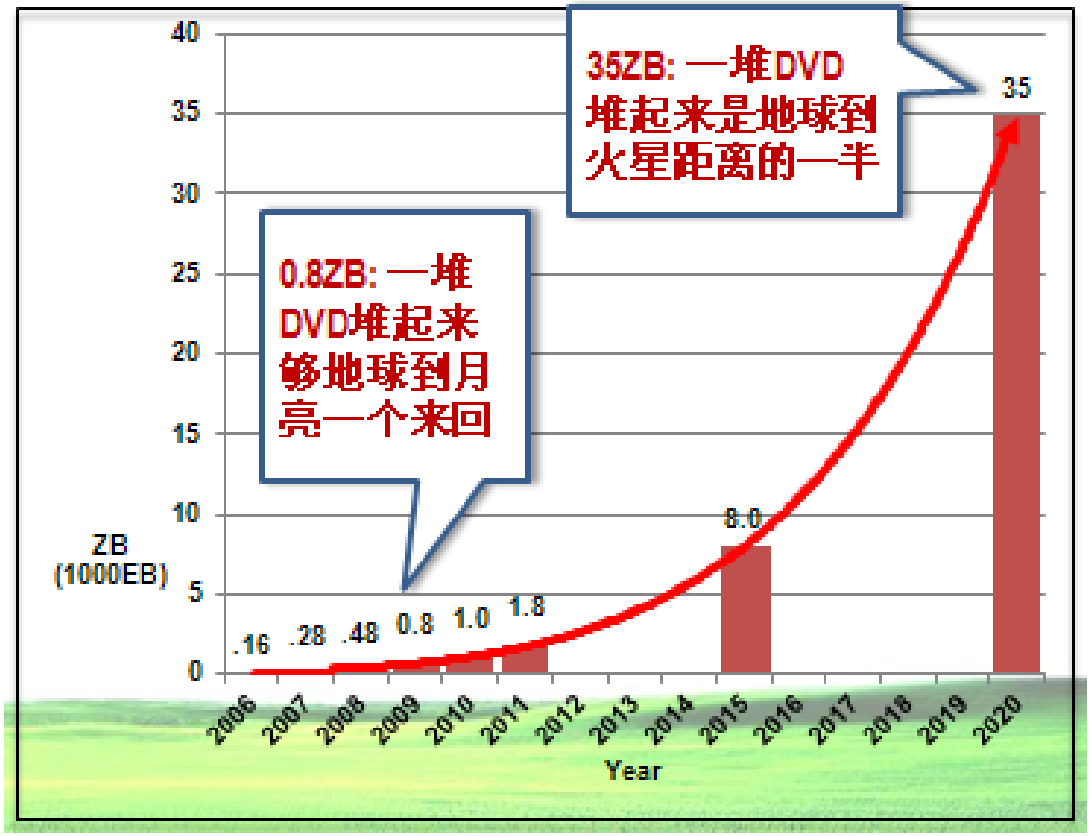
大纲

- 未标记样本与半监督学习
- 基于单学习器的方法：半监督SVM
- 基于分歧的方法：协同训练
- 半监督聚类

大数据时代

- Big data era, obtaining data is getting easier and easier.





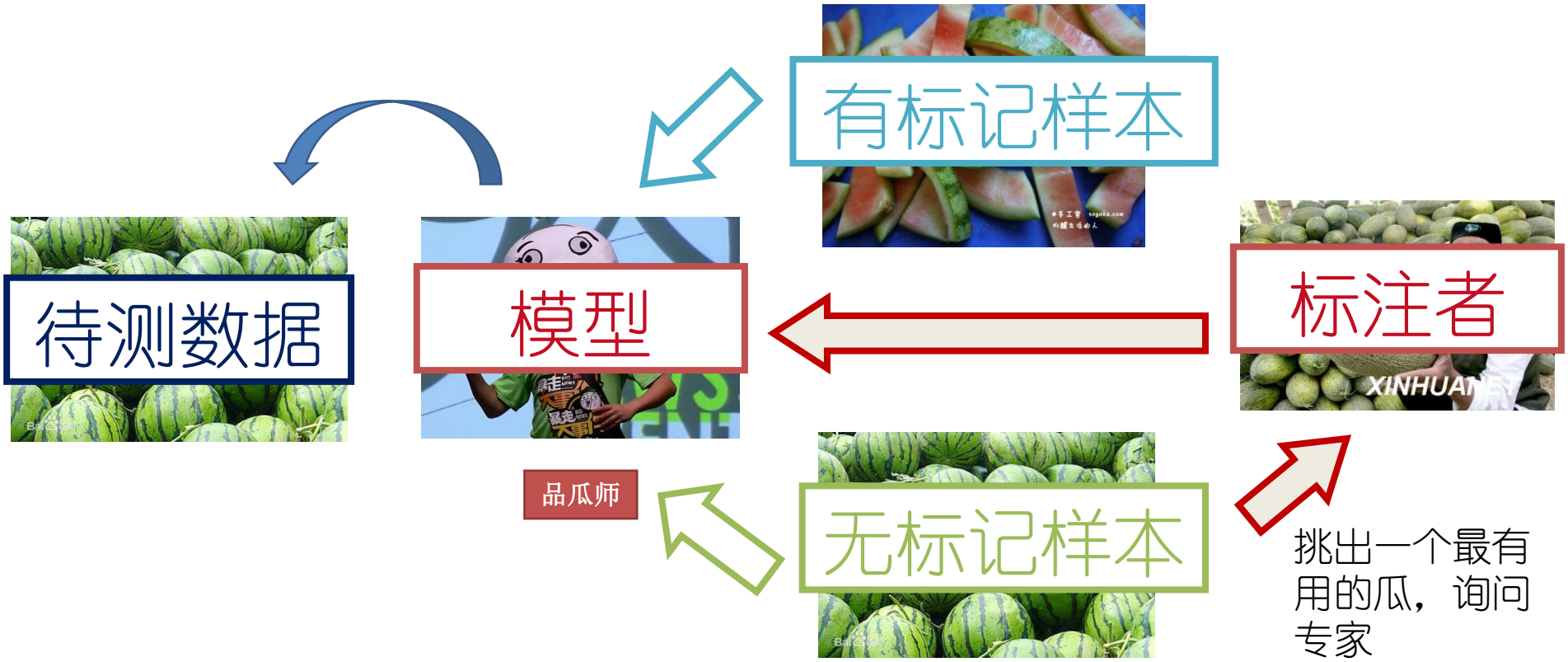
大数据时代

- Labeling the data is difficult! expensive, time consuming, sometimes need experts.

- For example, medical image analysis, webpage recommendation.

- How to use few labeled data and large amount of unlabeled data ?

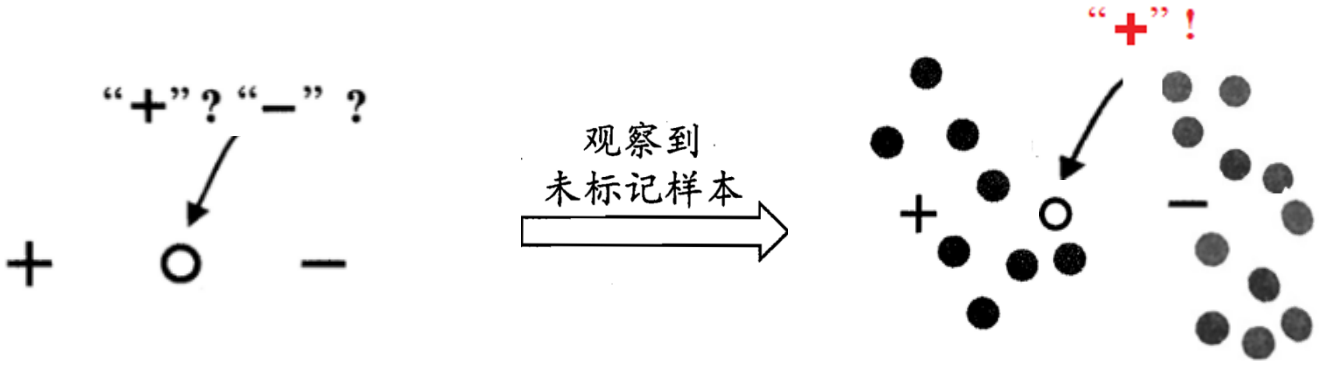
主动学习



- 主动学习引入了额外的专家知识，通过与外界交互将部分未标记样本变为有标记样本。
- 若不与专家交互，还能利用未标记样本吗？

- **Usefulness of unlabeled data**

➢ Can we use unlabeled data to improve the learning ability?



未标记样本效用的例示. 右边的灰色点表示未标记样本

The distribution of the unlabeled data tell us ***something***.

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。

流形假设可看做聚类假设的推广

半监督学习

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

（纯）半监督学习 基于“开放世界”假设

品瓜师

直推学习 基于“封闭世界”假设

Transductive learning

半监督学习

■ Defination

Give a training dataset

$D_l=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$, $y_i(i=1\ldots l)$ is the label ,

and dataset $D_u=\{x_{l+1},x_{l+2},\ldots,x_{l+u}\}$, $l \ll u$.

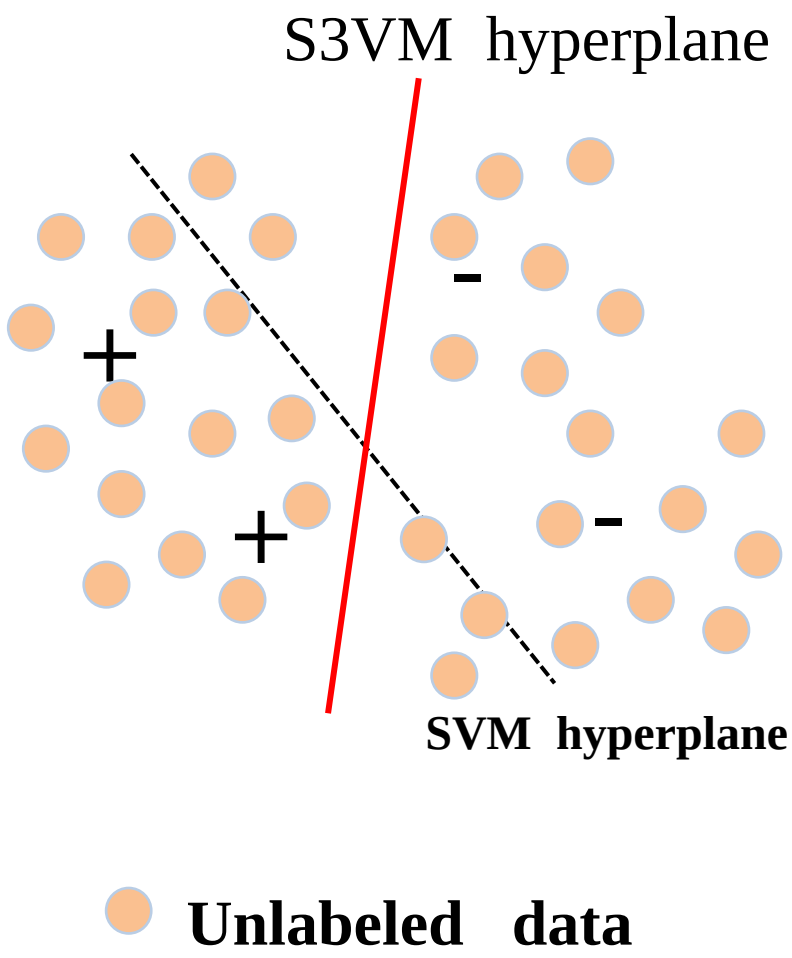
Semi-supervised learning (SSL) is learning a function f(x) by using both D_l and D_u , and then improve (at least not reduce) the classification performance.

Semi-supervised learning is halfway between supervised and unsupervised learning.

半监督SVM

半监督SVM

- Semi-supervised SVM tries to find a hyper-plane that can divide the different class data and goes through the lowest density area.
- Transductive SVM



TSVM

- TSVM(Transductive Support Vector Machine)
- 针对二分类问题，TSVM试图考虑对未标记样本进行各种可能的标记指派，然后在所有结果中，寻求一个在所有样本上间隔最大化的划分超平面。
- 数学模型

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

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

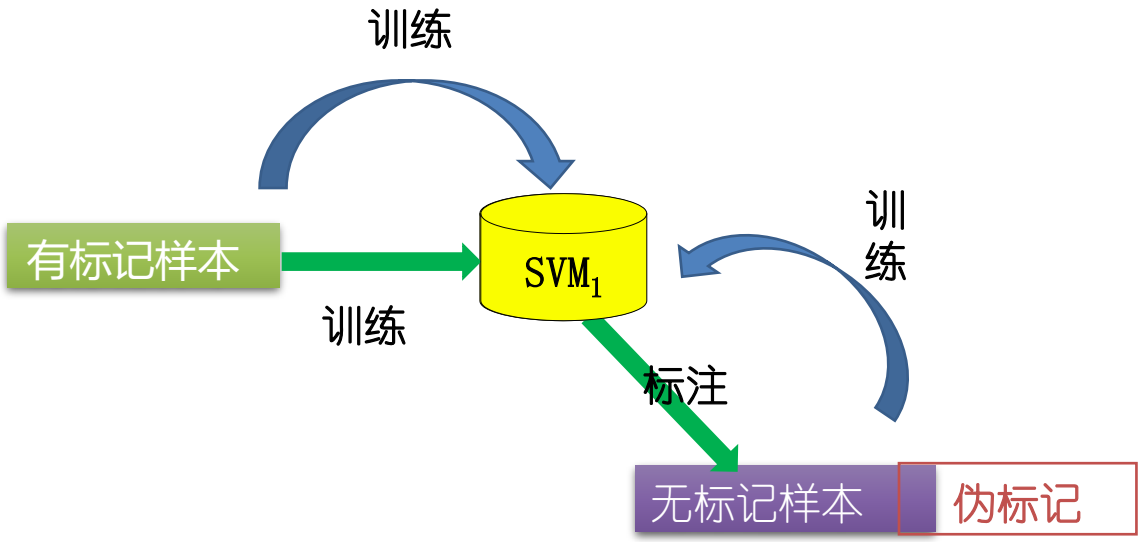
$$\hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = l+1, \ldots, m,$$

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

loss item of Unlabeled data

TSVM

- 显然，尝试未标记样本的各种标记指派是一个穷举过程。
- 在一般情形下，TSVM采用局部搜索来迭代地寻找近似解。



TSVM

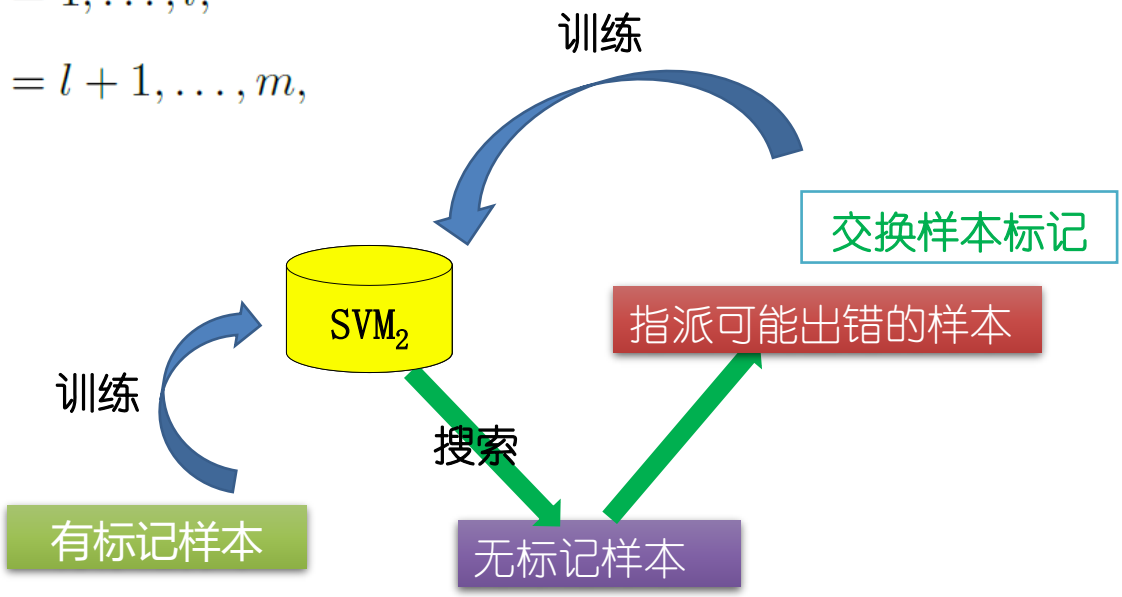
- 注意，开始时未标记样本的伪标记很可能不准确，因此设置 $C_u \ll C_l$ ，迭代过程中，逐渐增大 C_u 以提高未标记样本的作用。

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

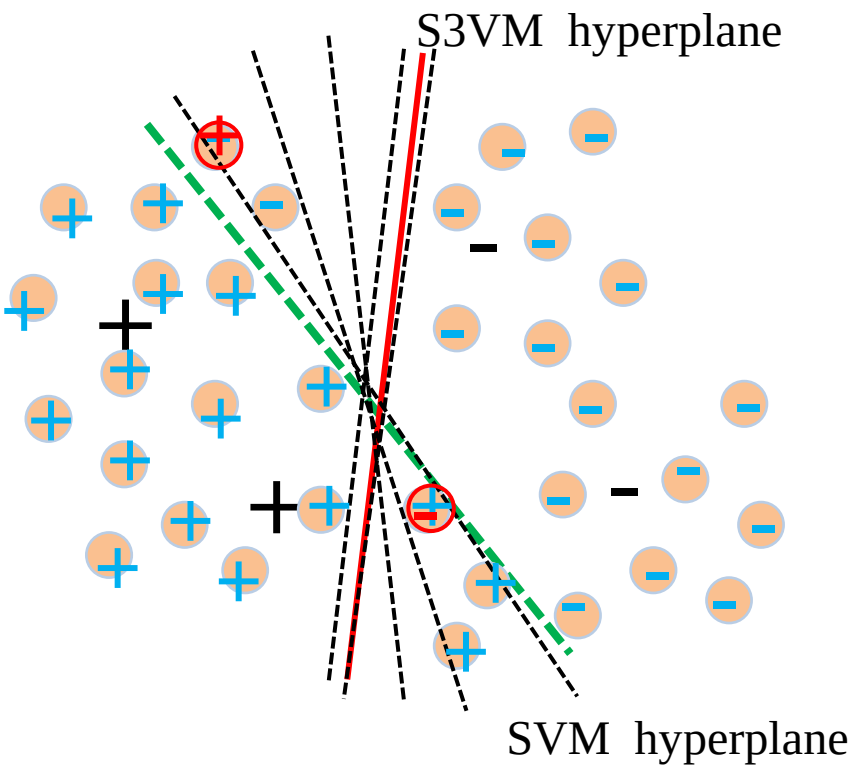
$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l,$$

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

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



TSVM



TSVM algorithm

Step 1 : train SVM on labeled data

→ Step 2 : predict unlabeled data by using SVM

Step 3 : train SVM on label data and psudo-label data

Step4 : find a pair of data which most wrongly predicted and swap their labels

Step 5: goto step2

TSVM

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 SVM_l;
用 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: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

未标记样本的伪标记不准确

图 13.4 TSVM 算法

TSVM

- 显然，搜寻标记指派可能出错的每一对未标记样本进行调整，仍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。

基于分歧的方法

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”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。

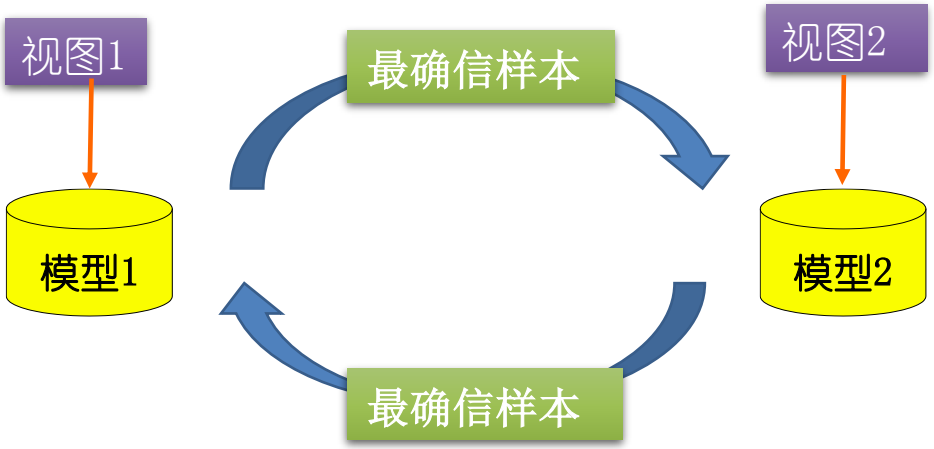
Multi-view

- web pages
- driverless vehicle

Data from different sensors

协同训练

- 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，每个属性集都足以描述该类样本，且相互独立。



Co-training trains two classifiers separately on two *views*, then uses the predictions of each classifier on unlabeled examples to augment the training set of the other.

协同训练

$\boldsymbol{x}_i$ 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**
6: **for** $t = 1, 2, \ldots, T$ **do**

在视图 j 上用有标记样本训练 h_j.

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_u 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

协同训练特点

- 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。

协同训练变体

- 协同训练算法本身是为多视图数据而设计的，性集合的常见数据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]。

半监督聚类

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.

过程：

1: 从 D 中随机选取 k 个样本作为初始均值向量$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k\}$;
2: **repeat**
3: $C_j=\varnothing\ (1\leqslant j\leqslant k)$;　初始化 k 个空簇.
4: **for** $i=1,2,\ldots,m$ **do**
5: 计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j\ (1\leqslant j\leqslant k)$ 的距离: $d_{ij}=\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|_2$;
6: $\mathcal{K}=\{1,2,\ldots,k\}$;
7: is_merged=false;
8: **while** ¬ is_merged **do**
9: 基于 $\mathcal{K}$ 找出与样本 $\boldsymbol{x}_i$ 距离最近的簇: $r=\arg\min_{j\in\mathcal{K}} d_{ij}$;
10: 检测将 $\boldsymbol{x}_i$ 划入聚类簇 C_r 是否会违背 $\mathcal{M}$ 与 $\mathcal{C}$ 中的约束;
11: **if** ¬ is_voilated **then**
12: $C_r=C_r\bigcup\{\boldsymbol{x}_i\}$;
13: is_merged=true
14: **else**
15: $\mathcal{K}=\mathcal{K}\setminus\{r\}$;
16: **if** $\mathcal{K}=\varnothing$ **then**
17: break并返回错误提示
18: **end if**
19: **end if**
20: **end while**
21: **end for**
22: **for** $j=1,2,\ldots,k$ **do**
23: $\boldsymbol{\mu}_j=\frac{1}{|C_j|}\sum_{\boldsymbol{x}\in C_j}\boldsymbol{x}$;　更新均值向量.
24: **end for**
25: **until** 均值向量均未更新

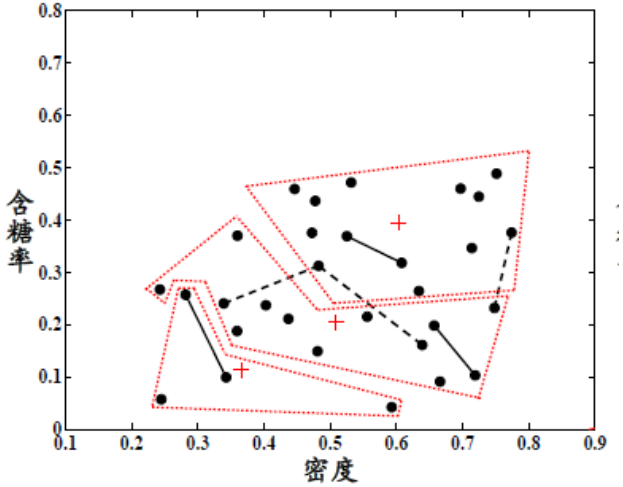
输出：簇划分 $\{C_1,C_2,\ldots,C_k\}$

不冲突，选择最近的簇

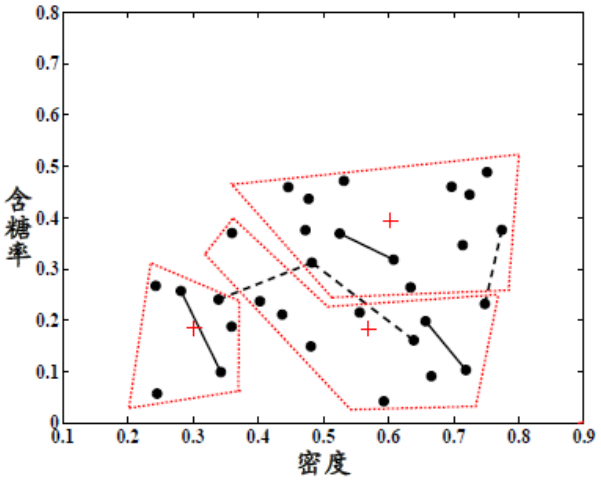
冲突，尝试次近的簇

图 13.7　约束 k 均值算法

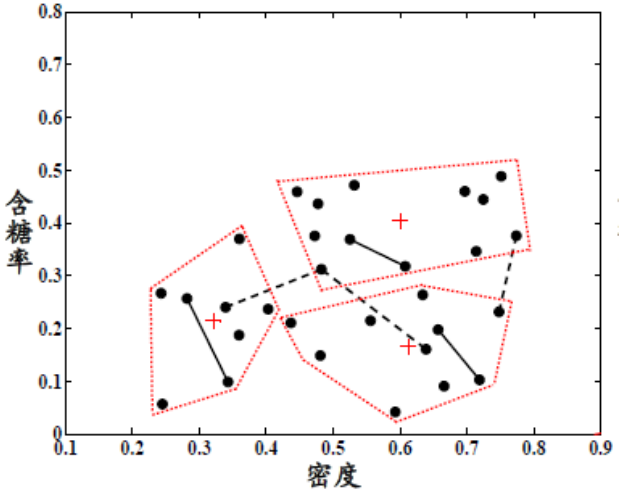
半监督聚类



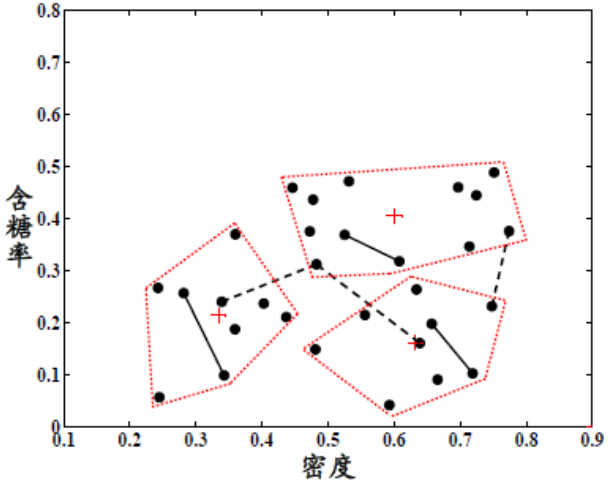
(a) 第 1 轮迭代后



(b) 第 2 轮迭代后



(c) 第 3 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

$S \subset D$, $|S| \ll |D|$.

输入：样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
少量有标记样本 $S = \bigcup_{j=1}^{k} S_j$;
聚类簇数 k .

过程：

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
```

```
repeat
    C_j = ∅ (1 ≤ j ≤ k);
```

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
    end for
    for all x_i ∈ D \ S do
        计算样本 x_i 与各均值向量 μ_j (1 ≤ j ≤ k) 的距离: d_ij = ||x_i − μ_j||_2 ;
        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
```

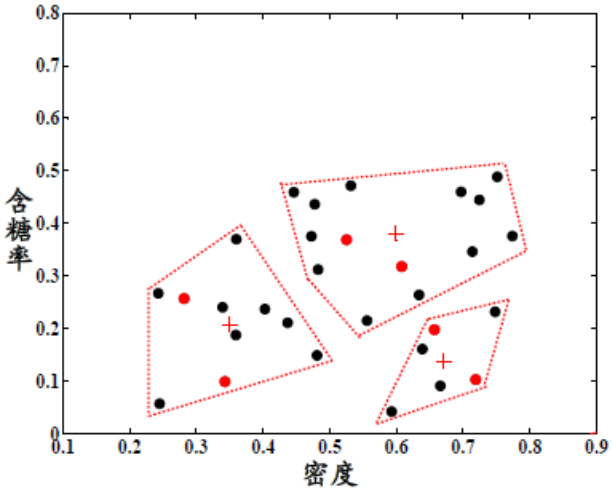
更新均值向量.

```
    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
```

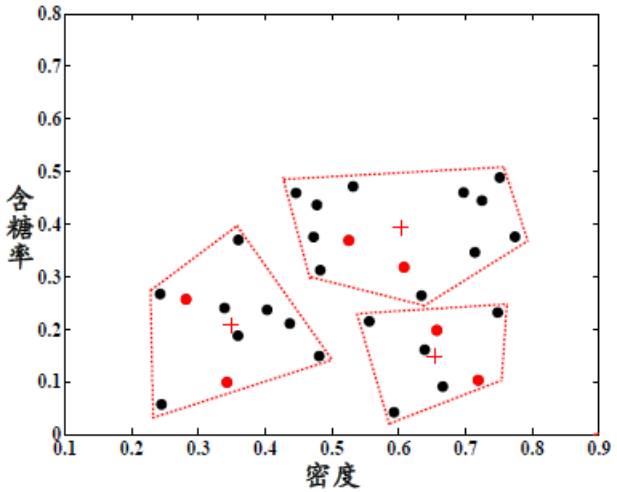
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.9　约束种子 k 均值算法

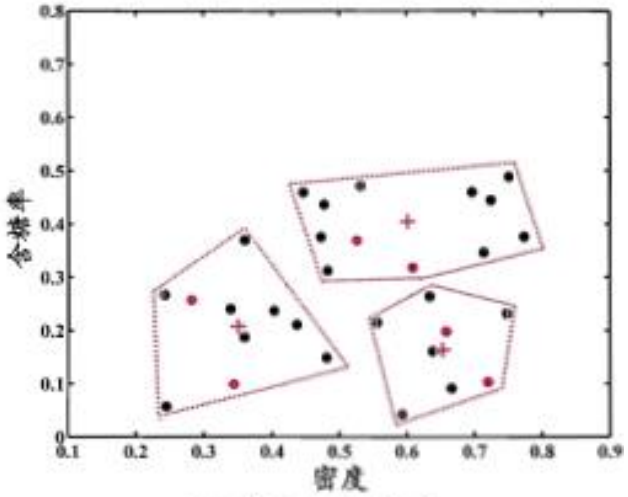
半监督聚类



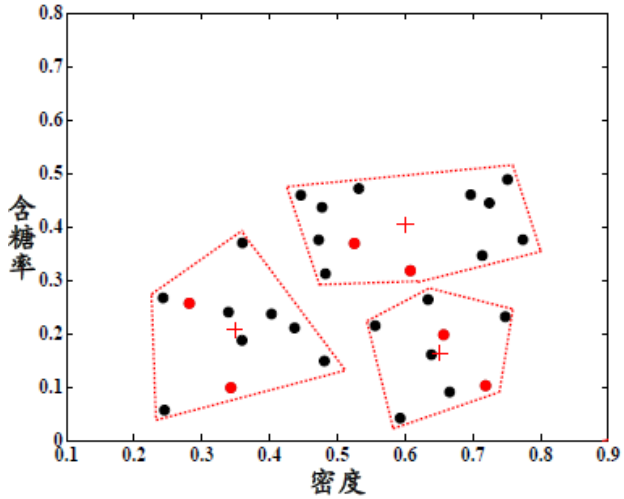
(a) 第 1 轮迭代后



(b) 第 2 轮迭代后



(c) 第 3 轮迭代后



(d) 第 4 轮迭代后

总结

- 未标记样本与半监督学习
- 基于单学习器的方法：半监督SVM
- 基于分歧的方法：协同训练
- 半监督聚类：两种监督信息