半监督学习

机器学习

半监督学习

在实际生活中，常常会出现一部分样本有标记和较多样本无标记的情形，例如：做网页推荐时需要让用户标记出感兴趣的网页，但是少有用户愿意花时间来提供标记。若直接丢弃掉无标记样本集，使用传统的监督学习方法，常常会由于训练样本的不充足，使得其刻画总体分布的能力减弱，从而影响了学习器泛化性能。那如何利用未标记的样本数据呢？

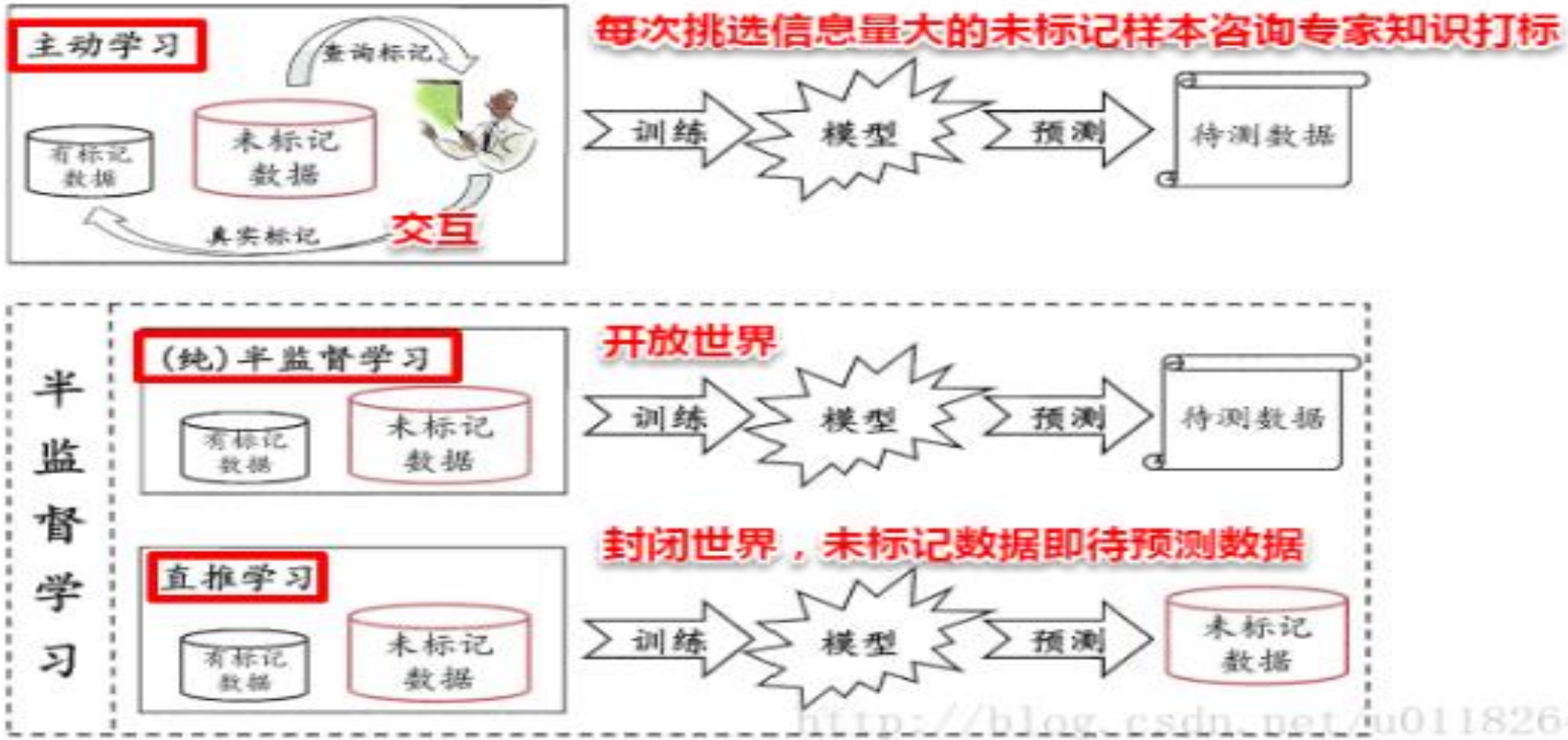
一种简单的做法是通过专家知识对这些未标记的样本进行打标，但随之而来的就是巨大的人力耗费。若我们先使用有标记的样本数据集训练出一个学习器，再基于该学习器对未标记的样本进行预测，从中**挑选出不确定性高或分类置信度低的样本来咨询专家并进行打标**，最后使用扩充后的训练集重新训练学习器，这样便能大幅度降低标记成本，这便是**主动学习**，其目标是**使用尽量少的/有价值的咨询来获得更好的性能**。

半监督学习

显然，**主动学习需要与外界进行交互/查询/打标，其本质上仍然属于一种监督学习**。事实上，无标记样本虽未包含标记信息，但它们与有标记样本一样都是从总体中独立同分布采样得到，因此**它们所包含的数据分布信息对学习器的训练大有裨益**。如何让学习过程不依赖外界的咨询交互，自动利用未标记样本所包含的分布信息的方法便是**半监督学习**.

半监督学习可进一步分为纯半监督学习和直推学习。前者假定训练数据中的未标记样本并非待遇测的数据，而后者则假定学习过程中所考虑的未标记样本恰是待遇测数据。

半监督学习



半监督的分类 根据监督信息的不同

- 半监督分类。半监督分类是作为监督分类问题的延伸，通过对有标签样本和无标签样本训练来进行分类。总的训练样本集包括 l 个有标签样本 $\{(x_i, y_i)\}_{i=1}^{l}$ 和 u 个无标签样本 $\{x_i\}_{i=l+1}^{l+u}$，通常情况下无标签样本的个数往往远大于有标签样本的个数，即 $u >> l$。半监督分类的目标是通过对有标签样本和无标签样本的共同训练产生一个分类器 f，使得它的分类效果优于单一依靠有标签样本进行训练。
- 约束聚类。约束聚类是对无监督学习的拓展，训练样本集 $\{x_i\}_{i=1}^{n}$ 由一些无标签样本组成，但是约束聚类中引入了一些已知的约束信息或称为监督信息。例如，已知 x_i 和 x_j 必定在同一个类簇中，则称之为 must-link，相反如果 x_i 和 x_j 不在同一个类簇中，则称为 cannot-link。同时，还可以对类簇的大小和个数进行限定。约束聚类往往能够比单一的依靠无标签样本聚类取得更好的效果。

半监督的可行性：

半监督学习的可行性建立在以下两个假设下

聚类假设（Cluster Assumption）：位于同一类簇中的样本应该具有相同的类别标签。假设有大量的无标签样本要进行分类，同时给出少量样本的标签信息，这时用所有样本（包括有标签和无标签样本）进行分类就是半监督分类需要解决的问题。根据该假设，标签相同的样本往往分布在同个聚簇中，所以决策边界应该尽量通过数据较为稀疏的地方，这样就不会把分布相对密集的聚簇中的样本点分开。

流形假设（Manifold Assumption）：处于一个很小的局部邻域内的样本应该具有相似的类别信息。假设所有的高维数据都分布在一个低维的子流形上，在寻找最佳特征子空间的过程中，可以通过利用大量的无标签样本来挖掘数据的内在几何结构，进而提高分类器的性能。与聚类假设着眼于数据的全局结构不同，流形假设更关注的是数据的局部特性。

半监督的学习方法分类：

- 生成式方法
- 支持向量机
- 基于图的学习算法
- 基于分歧的方法
- 半监督聚类

生成式方法

生成式方法（generative methods）是基于生成式模型的方法，即先对联合分布P（x, c）建模，从而进一步求解 P（c | x），此类方法假定样本数据服从一个潜在的分布，因此需要充分可靠的先验知识。例如：前面已经接触到的贝叶斯分类器与高斯混合聚类，都属于生成式模型。现假定总体是一个高斯混合分布，即由多个高斯分布组合形成，从而一个子高斯分布就代表一个类簇（类别）。高斯混合分布的概率密度函数如下所示：

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

混合系数　均值向量　协方差矩阵

生成式方法

给定有标记样本集 $D_l=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$ 和未标记样本集 $D_u=\{x_{l+1},x_{l+2},\ldots,x_{l+u}\}$，u 远大于 l，l+u=m。假设所有样本独立同分布，且都是由同一个高斯混合模型生成的。用极大似然法来估计高斯混合模型的参数$\{(\alpha_i,\mu_i,\Sigma_i)|1\le i\le N\}$，$D_l \cup D_u$的对数似然是：

当且仅当i=j时取1

$$LL(D_l\cup D_u)=\sum_{(x_j,y_j)\in D_l}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p(x_j\mid\mu_i,\Sigma_i)\,p(y_j\mid\Theta=i,x_j)\right)+\sum_{x_j\in D_u}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p(x_j\mid\mu_i,\Sigma_i)\right)$$

有类标样本只在特定类簇中出现

无类标样本可能在所有类簇中出现

上式由两项组成，基于有标记数据 D_l 的有监督项和基于未标记数据 D_u 的无监督项。

高斯混合模型参数估计可由 EM 算法求解，迭代更新式如下。

类先验　类条件

$$p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j) = \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\boldsymbol{x}_j \mid z_j = i)}{p_{\mathcal{M}}(\boldsymbol{x}_j)}$$

$$= \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} .$$

《机器学习》P208

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
高斯混合成分个数 k.

过程:

1: 初始化高斯混合分布的模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$
2: **repeat**
3: **for** $j = 1, 2, \ldots, m$ **do**
4: 根据式(9.30)计算 $\boldsymbol{x}_j$ 由各混合成分生成的后验概率, 即
$\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$ $(1 \leqslant i \leqslant k)$ E步
5: **end for**
6: **for** $i = 1, 2, \ldots, k$ **do**
7: 计算新均值向量: $\boldsymbol{\mu}_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} \boldsymbol{x}_j}{\sum_{j=1}^{m} \gamma_{ji}}$;
8: 计算新协方差矩阵: $\boldsymbol{\Sigma}_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i')(\boldsymbol{x}_j - \boldsymbol{\mu}_i')^{\mathrm{T}}}{\sum_{j=1}^{m} \gamma_{ji}}$; M步
9: 计算新混合系数: $\alpha_i' = \frac{\sum_{j=1}^{m} \gamma_{ji}}{m}$;
10: **end for**
11: 将模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ 更新为 $\{(\alpha_i', \boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i') \mid 1 \leqslant i \leqslant k\}$
12: **until** 满足停止条件
13: $C_i = \varnothing$ $(1 \leqslant i \leqslant k)$
14: **for** $j = 1, 2, \ldots, m$ **do**
15: 根据式(9.31)确定 $\boldsymbol{x}_j$ 的簇标记 λ_j;
16: 将 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$
17: **end for**

输出: 簇划分 $C = \{C_1, C_2, \ldots, C_k\}$

半监督的学习方法分类：

- 生成式方法
- 支持向量机
- 基于图的学习算法
- 基于分歧的方法
- 半监督聚类

半监督SVM方法（S3VM）

半监督支持向量机（Semi-Supervised Vector Machin，S3VM）是支持向量机在半监督学习上的推广。不考虑未标记样本的情况，支持向量机试图找到最大间隔划分超平面；在考虑未标记样本的情况下，S3VM试图找到能将两类有标记样本分开，且穿过数据低密度区域的划分超平面。低密度分隔（low-densityseparation）假设是聚类假设在考虑了线性超平面划分后的推广。

TSVM是半监督支持向量机中的著名代表，其核心思想是：尝试为未标记样本找到合适的标记指派，使得超平面划分后的间隔最大化。TSVM采用局部搜索的策略来进行迭代求解，即首先使用有标记样本集训练出一个初始SVM，接着使用该学习器对未标记样本进行打标，这样所有样本都有了标记，并基于这些有标记的样本重新训练SVM，之后再寻找易出错样本不断调整。

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

松弛变量 hinge损失

$$\text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1-\xi_i, \quad i = 1,2,\ldots,l,$$

$$\hat{y}_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1-\xi_i, \quad i = l+1, l+2,\ldots,m$$

$$\xi_i \geqslant 0, \quad i = 1,2,\ldots,m,$$

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; 初始SVM

用 SVM_l 对 D_u 中样本进行预测, 得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$; 伪标记

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

  $C_u = \min\{2C_u, C_l\}$ 逐渐增大Cu

end while

松弛变量越大表示离超平面越近，越容易分错

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

在对未标记样本进行标记指派和调整的过程中，有可能出现类别不平衡问题，即某类的样本远多于另一类，这对SVM的训练造成困扰。为减轻类别不平衡性所造成的不利影响，改进算法如下：将优化目标中的C_u拆分成C_u+和C_u-两项，分别对应基于伪标记而当作正、反例使用的未标记样本，并在初始化时令$(C_u+)=(u-/u+)(\ C_u-)$，其中u-和u+为基于伪标记而当作正、反例使用的未标记样本数。

搜索标记指派可能出错的每一对未标记样本进行调整，涉及巨大的计算开销的大规模优化问题，因此，半监督SVM更多的研究在于如何涉及出更高效的优化求解策略；发展出很多方法，如基于图核（graphkernel）函数梯度下降的LDS、基于标记均值估计的MeanS3VM等。

半监督的学习方法分类：

- 生成式方法
- 支持向量机
- 基于图的学习算法
- 基于分歧的方法
- 半监督聚类

图半监督学习

给定一个数据集，将其映射为一个图，数据集中每个样本对应于图中的一个结点，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）。将有标记样本所对应的结点染色，而未标记样本所对应的结点尚未染色；半监督学习对应于颜色在图上扩散或传播的过程；一个图对应一个矩阵，可基于矩阵运算来进行半监督学习算法的推导与分析。

给定有标记样本集 $D_l=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$ 和未标记样本集 $D_u=\{x_{l+1},x_{l+2},\ldots,x_{l+u}\}$，u 远大于 l，l+u=m。先基于 $D_l \cup D_u$ 构建一个图 G=(V,E)，其中结点集 $V=\{x_1,x_2,\ldots,x_l,x_{l+1},x_{l+2},\ldots,x_{l+u}\}$，边集 E 可表是为一个亲和矩阵（affinity matrix），常基于高斯函数定义：

$$(W_{ij})=\begin{cases}\exp\left(\dfrac{-\|x_i-x_j\|_2^2}{2\sigma^2}\right), & \text{if } i\neq j;\\ 0, & \text{otherwise}.\end{cases}$$

σ =带宽系数【用于控制权值的减缓程度】

w_{ij} 随着欧氏距离 $\|x_i-x_j\|$ 的增加而减少

具有最小能量的函数 f 在有标记样本上满足$f(x_i) = y_i$，i=(1,2,...,l)，在未标记样本上满足$\Delta f = 0$，其中$\Delta = D - W$为拉普拉斯矩阵（Laplacian matrix）。

以第 l 行与第 l 列为界，采用分块矩阵表示方式：$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$，$D = \begin{bmatrix} D_{ll} & D_{lu} \\ D_{ul} & D_{uu} \end{bmatrix}$，则能量函数为：

$$E(f) = f^T(D - W)f = \begin{pmatrix} f_l^T & f_u^T \end{pmatrix}\left(\begin{bmatrix} D_{ll} & D_{lu} \\ D_{ul} & D_{uu} \end{bmatrix} - \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}\right)\begin{bmatrix} f_l \\ f_u \end{bmatrix}$$

$$= f_l^T(D_{ll} - W_{ll})f_l - 2f_u^T W_{ul} f_l + f_u^T(D_{uu} - W_{uu})f_u$$

令$\frac{\partial E(f)}{\partial f_u} = 0$可得：$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$。

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

即：$P_{uu} = D_{uu}^{-1}W_{uu}$，$P_{ul} = D_{uu}^{-1}W_{ul}$，代入$f_u = (D_{uu} - W_{uu})^{-1}W_{ul}f_l$，可得：

$$f_u = \left(D_{uu}(I - D_{uu}^{-1}W_{uu})\right)^{-1}W_{ul}f_l = (I - D_{uu}^{-1}W_{uu})^{-1}D_{uu}^{-1}W_{ul}f_l = (I - P_{uu})P_{ul}f_l$$

推导出的这个式子可以看出，代入$f_l = (y_1, y_2, \ldots, y_l)$，就得到未标记样本的预测值$f_u$。

http://www.docin.com/p-1624081342.html

Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]// International Conference on Neural Information Processing Systems. MIT Press, 2003:321-328.

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。不过该类算法有两个缺点：
1）存储开销上，样本数m下，矩阵规模为$O(m^2)$，很难直接处理大规模数据；
2）构图过程仅能考虑训练样本集，对新样本在图中的位置难以知晓，在接收新样本时或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，如将D_l和经标记传播后得到标记的D_u合并作为训练集，另外训练一个学习器如支持向量机来对新样本进行预测。

半监督的学习方法分类：

- 生成式方法
- 支持向量机
- 基于图的学习算法
- 基于分歧的方法
- 半监督聚类

基于分歧的方法

与生成式方法、半监督SVM、图半监督学习等基于单学习器利用未标记数据不同，基于分歧的方法（disagreement-basedmethods）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）是基于分歧方法的重要代表。它是针对多视图（multi-view）数据设计，也是多视图学习的代表。

多视图数据

多视图数据是指一个数据对象同时拥有多个属性集（attribute set），每个属性集构成一个视图（view）。

如一部电影，拥有图像画面信息所对应的属性集、声音信息所对应的属性集、字幕信息所对应的属性集、网上宣传讨论所对应的属性集等多个属性集。若只考虑电影多视图数据中的图像属性集和声音属性集，一个电影片段样本用$(<x^1,x^2>,y)$表示，其中x^i是样本在视图i中的示例，即基于该视图属性描述而得的属性向量。假定x^1为图像视图中的属性向量，x^2为声音视图中的属性向量；y是标记，如电影类型。

相容性

假设不同视图具有相容性（compatibility），即其所包含的关于输出空间y的信息是一致：令y^1表示从图像画面信息判别的标记空间，y^2表示从声音信息判别的标记空间，则有$y=y^1=y^2$。

在相容性的基础上，不同视图信息是互补的，给学习器的构建带来便利。如某个电影片段，从图像上有两人对视，无法判断电影类型，但若加上声音信息中“我爱你”透露的信息，则可判定为电影类型是爱情片。

协同训练

协同训练正是基于多视图数据的相容互补性。假设数据拥有两个充分（sufficient）且条件独立视图，**充分**是指每个视图都包含足以产生最优学习器的信息；**条件独立**是在给定类别标记条件下每个视图独立。

协同训练如何利用未标记数据呢？

- 首先在每个视图上基于有标记样本分别训练出一个分类器
- 然后让每个分类器去选择自己最有把握的未标记样本赋予伪标记，并将伪标记样本提供给另一个分类器作为新增的有标记样本用于训练更新，这个过程不断迭代进行，直到两个分类器都不再发生变化，或达到预先设定的迭代轮数为止。

总之，每个视图根据有标记样本生成一个学习器，来判别本视图的未标记数据，然后将打上未标记的样本作为其他视图学习器生成的新增有标记样本。

算法分析

若在每轮学习中都考察分类器在所有未标记样本上的分类置信度，会产生很大的计算开销，因此在算法中使用了未标记样本缓冲池。分类置信度的估计因基本学习算法而异，如若使用朴素贝叶斯分类器，可将后验概率转化为分类置信度；若使用支持向量机，则可将间隔大小转化为分类置信度。

协同训练的理论证明显示，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。不过这个前提条件在现实任务中很难满足，但就是视图充分就基本无法做到，不过即便如此，协同训练仍可有效地提升弱分类器的性能。

算法衍生

协同训练算法是为多视图数据而设计的，后面也出现了在单视图上使用的变体算法，或使用不同学习算法，或使用不同数据采样，甚至使用不同的参数设置来产生不同学习器，也能有效地利用未标记数据来提升性能。

实际，原理思想是一致，就是在不同学习器之间产生互补。后续理论研究表明，这类算法并不一定数据是拥有多视图，而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升泛化性能。不同视图、不同算法、不同数据采样、不同参数设置等，都是产生差异的渠道，而非必备条件。

半监督的学习方法分类：

- 生成式方法
- 支持向量机
- 基于图的学习算法
- 基于分歧的方法
- 半监督聚类

半监督聚类

聚类是无监督学习任务，为利用现实任务中获得的监督信息，提出半监督聚类（semi-supervised clustering）来利用监督信息以获得更好的效果。

聚类任务中获得的监督信息分两种：
1）有必连（must-link）和勿连（cannot-link）约束，必连是指样本必属于同一个簇，勿连是指样本必不属于同一个簇；
2）含有少量的有标记样本。

约束k均值

约束k均值算法是在k均值算法基础上扩展，在聚类过程中确保M与C中的约束得以满足；算法描述如下：

利用必连和勿连约束的监督信息。假定样本集$D=\{x_1,x_2,\ldots,x_m\}$以及必连关系集合M和勿连关系集合C；$(x_i,x_j)\in M$表示x_i,与x_j必属于同簇，$(x_i, x_j)\in C$表示x_i,与x_j必不属于同簇。

《机器学习》-K均值算法 P202

输入：

样本集 $D=\{x_1,x_2,...,x_m\}$

必连约束集合 M；

勿连约束集合 C；

聚类簇数 k；

过程：

从 D 中随机选择 k 个样本作为初始均值向量$\{\mu_1,\mu_2,...,\mu_k\}$;

repeat

$C_j = \emptyset(1 \le j \le k)$;

for i=1,2,...,m do

计算样本x_i与各均值向量μ_j的距离$d_{ij} = \|x_i - \mu_j\|_2$;

K={1,2,...,k};

is_merged=true;

while is_merged do

基于 K 找出与样本x_i距离最近的簇：$r = \arg\min_{j\in K} d_{ij}$;

检测将x_i划入聚类簇C_r是否会违背 M 与 C 中的约束;

if is_volilated then //不违背

$C_r = C_r \cup x_i$;

is_merged=false;

else

K=K\{r};

if K=∅ then

break 并返回错误提示;

end if

end if

end while

end for

for j=1,2,...,k do

$\mu_j = \frac{1}{|C_j|}\sum_{x\in C_j} x$;

End for

until 均值向量均未更新

输出：簇划分$\{C_1,C_2,...,C_k\}$

约束种子k均值

给定样本集$D=\{x_1,x_2,\ldots,x_m\}$，假定少量的有标记样本 ，其中S_j为隶属于第j个聚类簇的样本。直接将有标记样本作为种子，初始化k均值算法的k个聚类中心，并且在聚类迭代更新过程中不改变种子样本的簇隶属关系，就是约束种子k均值算法。

输入：

样本集 $D=\{x_1,x_2,\ldots,x_m\}$

少量有标记样本$S = U_{j=1}^{k}S_j$

聚类簇数 k

过程：

for j=1,2,…,k do

$\mu_j = \frac{1}{|S_j|}\sum_{x\in S_j} x$

End for

Repeat

$C_j = \emptyset(1 \le j \le k)$;

for j=1,2,…k do

for all $x \in S_j$ do

$C_j = C_j \cup \{x\}$

End for

End for

For all $x_i \in D \setminus S$ do

计算样本x_i与各均值向量$\mu_j(1 \le j \le k)$的距离：$d_{ij} = \|x_i - \mu_j\|_2$;

找出与样本x_i距离最近的簇：$r = \arg\min_{j\in\{1,2,\ldots,k\}} d_{ij}$；

将样本x_i划入相应的簇：$C_r = C_r \cup \{x_i\}$;

End for

for j=1,2,…,k do

$\mu_j = \frac{1}{|C_j|}\sum_{x\in C_j} x$;

End for

until 均值向量均未更新

输出：簇划分$\{C_1, C_2, \ldots, C_k\}$

谢谢大家