半监督学习综述

klrxbest

摘要：近年来，半监督学习技术成为了机器学习领域的一个热点研究方向。本文首先简单地介绍了半监督学习的发展历史，然后从经典的半监督学习算法、关于半监督学习的理论分析和半监督学习在实际问题中的应用三个方面对近年来半监督学习研究的进展进行了简单介绍，最后指出了目前半监督学习领域存在的一些亟待解决的问题。

关键字：半监督学习

Abstract: In recent years, semi-supervised learning technology has become a hot area of machine learning research. This article first briefly describes the history of development of semi-supervised learning, and then gives a brief introduction to some classic semi-supervised learning algorithms, theoretical analysis on semi-supervised learning and applications of semi-supervised learning techniques, and points out some existing problems to be solved in semi-supervised learning.

Keywords: semi-supervised learning

1. 引言

自从 1946 年第一台电子计算机 ENIAC 诞生至今，计算机技术得到了迅猛的发展，这使得人类采集、存储数据的能力空前的提高，利用计算机对收集到的数据进行分析提取有价值信息的技术（机器学习技术）也随之而生，并得到了很快的发展。传统的机器学习技术一般只利用有标记样本集或者只利用无标记样本集进行学习，而在实际问题中大多是有标记样本与无标记样本并存，为了更好地利用这些数据，半监督学习技术应运而生，近年来，半监督学习技术更成为机器学习领域一个被广泛研究的热点方向。

传统的机器学习技术可以分为两类，一类是无监督学习，另一类是监督学习。

无监督学习一般基于这样的数据设置：

假设数据集 X 包含 n 个样本点 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i \in \mathrm{X}\ (\forall i \in \{1,2,\cdots n\})$，一般假设 x_i 独立同分布(i.i.d., independently and identically distributed)地取样于分布 X 。无监督学习技术的基本目标即是根据这些样本点估计出分布 X 的密度，这类技术的典型代表有聚类、降维等。

监督学习有别于无监督学习的是数据不仅包含样本点本身，而且还包含这些样本点所对应的类别标记(label)，一个样本点的表述形式为 (x_i,y_i)，其中 y_i 为样本点 x_i 的类别标记。监督学习的目标即是从这些数据中学习建立一个从样本点到标记的映射。特别地，监督学习技术当 y_i 为实数值的时候称为回归(regression)技术，否则，称为分类(classification)技术。这类技术的典型代表有支持向量机(SVM, Support Vector Machine)等。

半监督学习则是介于两者之间的学习技术，它同时利用有标记样本与无标记样本进行学习。它所利用的数据集 $X=\{x_1,x_2,\cdots,x_n\}\ (n=l+u)$，可以分为两部分，一部分是有标记的数据集 $X_l=\{x_1,\cdots,x_l\}$，这部分数据集中的样本点 x_i 的类别标记 y_i 已经给出，另一部分是无标记数据集 $X_u=\{x_{l+1},\cdots,x_{l+u}\}$，这部分数据集中样本点的类别标记未知。与实际情况相符一般假设 $u \gg l$，即无标记数据远远多于有标记数据。

在实际问题中，往往无标记数据集的样本点数目会远远多于有标记数据集的样本点数目，这是因为对

某些数据进行标记的代价会很高，比如在生物学中，对某种蛋白质的结构分析或者功能鉴定可能会花上生物学家很多年的工作，而无标记的样本却是随手可得。由于解决实际问题的需要，对半监督学习技术的研究变得尤为重要。

本文对机器学习领域中的半监督学习技术的研究现状进行了简单介绍。其中，第 2 部分对半监督学习技术的发展历史进行了简单的回顾，第 3 部分简单地介绍了半监督学习中的常用假设，第 4 部分从经典算法、理论分析、实际应用三个方面对近年来的半监督学习进行介绍，第 5 部分总结并提出了现有方法存在的一些问题，第 6 部分感谢。

2. 半监督学习技术的历史

最早利用到半监督学习思想的算法应当是自训练(self-training)方法[1]，出现于 20 世纪 60-70 年代的某些文献中（如，1965 年 Scudder 的文献[2]、1967 年 Fralick 的文献[3]，1970 年 Agrawala 的文献[4]）。自训练方法的基本思想[2]是首先利用监督学习技术对有标记的数据进行学习，然后利用学习到的结果对未标记数据进行标记，再将新标记的数据加入到有标记的数据中去再学习，如此迭代。然而这种方法的性能依赖于其中的监督学习技术，而且当利用 0-1 代价函数经验风险最小化进行学习时，未标记样本将失效[1]。

之后出现的则是与半监督学习技术十分相关的直推(transductive)学习技术，直推学习技术由 Vapnik 在 1974 年根据著名的 Vapnik 原理(不要通过解决更复杂的问题来解决问题)提出的学习技术，他认为很多学习任务只需要将当前数据集中的未标记样本进行标记，不需要对样本空间中的所有样本进行标记。直推学习技术与半监督学习技术的本质不同就在于，直推学习技术学习到的结果只需对当前数据中的未标记样本进行标记，半监督学习的学习目标则需是对于整个样本空间中的所有样本都可预测其标记。Z.H. Zhou[29]给出了形象化的描述，直推学习技术是基于封闭世界假设，而半监督学习是基于开放世界假设的。而关于半监督学习技术与直推学习技术的异同点在学术界依然是一个开放式的话题，存在着争论[1]。

半监督学习技术开始发展是由于 20 世纪 70 年代对利用未标记数据学习 Fisher 线性判别规则的研究。这段时间利用 EM 算法结合高斯混合模型或者多项式分布模型的半监督学习技术被广泛提出来。根据 D.J. Miller 和 H.S. Uyar[5]的看法，由于很难利用未标记数据对训练诸如前馈神经网络等当时主流学习技术进行提高，所以半监督学习研究在那个时候没有能够迅速发展开来。

而到了 20 世纪 90 年代，由于自然语言处理的发展，对利用未标记数据帮助提高学习性能的需求越来越强烈，半监督学习才成为了机器学习领域中的研究热点方向。根据 O. Chapelle 等人[1]的看法，半监督学习这个术语最早是在 Merz 等人[6]在 1992 年使用的。

3. 半监督学习中的基本假设

目前的机器学习技术大多基于独立同分布假设，即数据样本独立地采样于同一分布。

除了独立同分布假设，为了学习到泛化的结果，监督学习技术大多基于平滑(smoothness)假设，即相似或相邻的样本点的标记也应当相似。而在半监督学习中这种平滑假设则体现为两个较为常见的假设：聚类(cluster)假设与流型(manifold)假设。

下面对半监督学习中两个常用的假设做简单的介绍。

聚类假设是指同一聚类中的样本点很可能具有同样的类别标记。这个假设可以通过另一种等价的方式进行表达，那就是决策边界所穿过的区域应当是数据点较为稀疏的区域，因为如果决策边界穿过数据点较为密集的区域那就很有可能将一个聚类中的样本点分为不同的类别这与聚类假设矛盾。

流型假设是指高维中的数据存在着低维的特性。Z.H. Zhou 在文献[29]中给了另一种类似的表述，“处于一个很小的局部邻域内的示例具有相似的性质”。关于这两者的等价性没有严格的证明，但是高维数据中

的数据的低维的特性是通过局部邻域相似性体现的，比如一个在三维空间卷曲的二维纸带，高维的数据全局的距离度量由于维度过高而显得没有区分度，但是如果只考虑局部范围的距离度量，那就会有一定意义。

这两种假设一般是一致的，属于监督学习中平滑假设的在半监督学习中的推广。Z.H. Zhou[29]认为流型假设比聚类假设更为一般，因为根据他的描述流型假设是相似的样本点具有相似的性质而不是聚类假设所认为的相同的标记，对于聚类假设无法成立的回归问题上流型假设却可以成立。

4. 半监督学习现状

4. 1 半监督学习算法介绍

监督学习风范(paradigm)可以分为生成式(generative)风范与诊断式(diagnostic)风范。生成式风范是通过估计 $p(x|y)$，得出 x 的产生方式，然后再利用贝叶斯法则估算 $p(y|x)$，从而对未标记样本的标记进行预测，而诊断式风范是直接估计 $p(y|x)$ 而不关心 x 的生成方式。对于半监督学习这两种风范区别就变得模糊了[1]，所以本文不按照这样的分类方式对算法进行介绍，而根据基于假设的不同进行分类介绍。同时，对于机器学习领域中的经典算法如 EM、SVM 等不作赘述。

4. 1. 1 生成式模型

这类算法基于聚类假设。假设数据模型为 $p(x)=p(y)p(x|y)$，其中 $p(x|y)$ 一般表示为“可确认的”(identifiable)混合模型，混合模型的各个组成成分可以通过大量的未标记数据获得，然后再通过少量的标记样本的标记信息即可确定整个混合模型。

离散概率混合模型是假设一个随机变量 X 的概率密度函数由 n 个随机变量的概率密度函数 $\{Y_1,Y_2,\cdots,Y_n\}$ 组合而成，可以表示为这些分量的线性组合（最原始的混合模型的定义只需是凸组合即可）的形式。如下式：

$$f_X(x)=\sum_{i=1}^{n}a_i f_{Y_i}(x)$$

其中，$0\le a_i\le 1$，$\sum_{i=1}^{n}a_i=1$

混合模型较为重要的性质就是“可确认性”(identifiability)，反映的就是给定 $p(x)$ 的分布，并且给定各个混合分量的分布是不是可以唯一确定的给出分量的系数 a_i 的性质，如果可以唯一确定。关于离散概率混合模型的“可确认性”的充要条件在 1963 年 H. Teicher 的文献[7]中已经给出并证明。现仅将定理列出。

定理：对于有穷混合模型 $\kappa=\{\sum_{i=1}^{n}a_iF_i(x):a_i\ge 0,\sum_{i=1}^{n}a_i=1\}$ 为可确认的，当且仅当存在 n 个实数值 $\{x_1,x_2,\cdots,x_n\}$ 使得对于由 $F_i(x_j),1\le i,j\le n$ 所构成的行列式（i 为行数，j 为列数）不为 0。

这个定理结合 H. Teicher 文献[7]中的另一个定理有一个简单的推论，就是所有的有穷正态分布的混合模型是“可确认的”。

鉴于上述良好的数学性质，在半监督学习的生成式模型中，高斯混合模型是被用得最多的混合模型。一般的混合模型可以表示为 $p(x)=\sum_i p(y_i)p(x|y_i)$，其中 y_i 为类别标记。聚类假设即是体现于此。

这类算法一般采用 EM 算法利用未标记样本进行对模型中 $p(x|y_i)$ 组成成分的参数进行估计，再利用少量的未标记样本确定各分量所代表的类别，这类算法比较经典的是 Nigam 等人在 2000 年的工作[8]。

这类算法简单，但是存在着较难解决的问题，如克服 EM 算法的局部最优解的问题，Nigam 在 2001 年的工作[9]在这方面做出了尝试；除了使用混合模型以外，很多算法还需要首先对未标记样本进行聚类，然后再利用已标记样本给这些聚类的结果赋予类别标记，尽管在模型与数据磨合很好的情况下，这类算法可以有很好的性能，但这类算法很难分析[1]。

4.1.2 低密度分割算法

低密度分割就是要尽量让分类边界通过密度较低区域。这类算法即是对聚类假设的第二种描述的很好的应用。下面简单介绍一下比较经典的工作。

4.1.2.1 半监督支持向量机(S3VM, Semi-Supervised SVM)

支持向量机(SVM, Support Vector Machine)是 Vapnik 领导的 Bell 实验室小组提出来的一种分类算法，这种分类方法通过引入核函数，将原本在低维空间线性不可分的数据映射到高维空间（SVM 算法中被称为 RKHS，Reproducing Kernel Hilbert Space），从而变得线性可分，而分类界面即是采用最小经验风险与最大“间隔”(marginal)的标准来确定。

将聚类假设运用到支持向量机中，即是要分类边界绕过数据密集的区域。将 SVM 改进应用到半监督学习中的也是 Vapnik，S3VM 其本质上是直推式的，Vapnik 本人提出时也称之为 TSVM(Transductive SVM)，直推支持向量机。

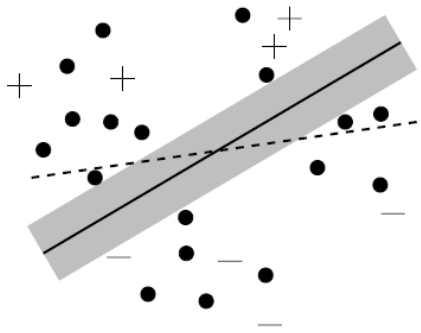

图 1 直推支持向量机的分类结果示意图[1]

就是对样本中的所有的数据包括标记的与未标记的建立一个分类界面，在学习过程中，调整分类界面，使得“间隔”最大，而且尽量避过数据较为密集的区域（通过修改分类边界两侧的未标记样本的类别标记迫使分类界面调整至稀疏区域），如上图所示，实线为调整后分类边界。

形式化定义：

对数据集 $X=\{x_1,x_2,\cdots,x_n\}$（包括标记与未标记数据），需要学到的分类边界为 $\{\mathbf{x}:\mathbf{w}\cdot\mathbf{x}+b=0\}$，最大化到最近的样本点的“间隔”，则得到优化目标：

$$\min_{i\in[1\cdots n]}\left[\frac{\mathbf{y}_i}{\|\mathbf{w}\|}(\mathbf{w}\cdot\mathbf{x}_i+b)\right]$$

Vapnik 同时也给出了 TSVM 的错误率上界这使得 TSVM 在理论上有很好的保证。

4.1.2.2 使用半定规划的半监督学习算法

解 TSVM 的优化目标是一个 NP-难的问题，因为除了分类界面未标记样本的标记也是未知，导致了损

失函数非凸[10]。很多研究人员尝试放松 TSVM 的优化目标以期使得问题可解。其中较为经典的工作就是 De Bie 和 Cristianini[11]提出的将 TSVM 的优化目标放松成为半定规划问题，而半定规划问题是凸优化问题，从而使得问题对于稍大数据集可解。

这里只对放松后的优化问题作简单介绍，对问题如何转化以及转化后与半定规划问题的等价性的证明不作详细介绍，可以参考原文[1]。

定义标记矩阵：

$$\Gamma=\begin{pmatrix}\Gamma_{ll} & \Gamma_{lu}\\ \Gamma_{ul} & \Gamma_{uu}\end{pmatrix}=\begin{pmatrix}Y_lY_l^T & Y_lY_u^T\\ Y_uY_l^T & Y_uY_u^T\end{pmatrix}$$

其中 Y_l 为已标记数据的标记向量， Y_u 为未标记数据的标记向量。

则 TSVM 的优化目标可以放松为：

$$\min_{\Gamma}\max_{\boldsymbol{\alpha}} 2\boldsymbol{\alpha}^T\mathbf{1}-\boldsymbol{\alpha}^T(K\odot\Gamma)\boldsymbol{\alpha}$$

$$s.t. C\geq\alpha_i\geq 0$$

$$diag(\Gamma)=\mathbf{1}$$

$$\Gamma\geq 0$$

其中 α_i 为 $\mathbf{w}=\sum_i\alpha_i c_i\mathbf{x}_i$ 所决定。

De Bie 和 Cristianini 证明了这个优化目标是凸的，而且等价于一个半定规划问题。

但是实验的结果并没有十分的好，因为解半定规划问题所需要的计算开销依然很大。

除此之外，N.D. Lawrence 和 M.I. Jordan[12]通过修改高斯过程的噪音模型来进行半监督学习，对类的标记除了正负类之外，他们还引入了标记 0，并且规定未标记的数据的类别标记不可以标记为 0，这样迫使分类边界避开数据密集区域，这与 TSVM 的想法类似；使用正则化项使学习结果具有某种所需要的特性是机器学习领域中的惯用手法，Grandvalet 和 Bengio[13]通过在优化目标中将未标记样本的熵作为正则化项加入到优化目标中去进行学习，从而使得熵最小化，进而使得分类界面尽量不要切分数据密集区域，因为切分了数据密集区域则使得其不确定性增加，从而使得熵变大。

4.1.3 基于图的半监督算法

这类算法基于流型假设。假设所有的样本点（包括已标记与未标记）以及之间的关系可以表示为一个无向图的形式 $g=<V,E>$。其中图的结点为数据样本点，而边则体现了两个样本点之间的相似度关系。基于图的半监督算法的优化目标就是要保证在已标记点上的结果尽量符合而且要满足流型假设。

很多基于图的算法都使用到图的拉普拉斯(Laplacian)矩阵。这些算法给予图的边赋予权值，然后计算其拉普拉斯矩阵。假设定义图的结点之间的边的权值矩阵为 $\mathbf{W}$，其中 w_{ij} 表示两个结点之间边的权值，当两点之间无边时，$w_{ij}=0$，边的权值可以有多种定义方式，较常见的定义为 k 近邻或者高斯核矩阵的形式。

k 近邻定义：如果结点 x_i 是结点 x_j 的 k 个最邻近结点中的一个，那么 $w_{ij}=1$，否则 $w_{ij}=0$。

高斯核矩阵：$w_{ij}=e^{-\frac{\|x_i-x_j\|_2}{2\sigma^2}}$。特别地，$w_{ii}=0$

令 $\mathbf{D}: d_{ij} = \begin{cases} \sum_j w_{ij} & \text{if } i = j \\ 0 & o.w. \end{cases}$，则图的非规范化拉普拉斯矩阵定义为 $\mathbf{D} - \mathbf{W}$，图的规范化拉普拉斯矩阵定义为 $\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$。

下面以 Zhu 和 Ghahramani[14]在 2002 年提出的一种基于图的算法为例，解释如何使用图的拉普拉斯矩阵将图中已标记点的标记传播到未标记结点。

算法一 标记传播算法（Zhu 和 Ghahramani，2002）

利用高斯核矩阵定义图的权值矩阵 W

计算 $\mathbf{D}: d_{ij} = \begin{cases} \sum_j w_{ij} & \text{if } i = j \\ 0 & o.w. \end{cases}$

初始化样本点的标记向量 $\hat{Y}^{(0)} \leftarrow (y_1, y_2, \cdots y_l, 0, \cdots 0)$

迭代计算 1．2．直至收敛

1. $\hat{Y}^{(t+1)} \leftarrow \mathbf{D}^{-1}\mathbf{W}\hat{Y}^{(t)}$

2. $\hat{Y}_l^{(t+1)} \leftarrow Y_l$

其中步骤 1 是对所有样本点（包括已标记与未标记）的类别标记进行更新，步骤 2 是保证所有已标记点的标记为原始的标记，在整个过程中已标记点的标记始终不会改变。

与该算法类似的是 Zhou 等人[15]与 2004 年提出的标记“扩散”(spreading)算法，与算法一不同点在于，步骤 1 的更新方式为 $\hat{Y}^{(t+1)} \leftarrow \alpha\mathbf{L}\hat{Y}^{(t)} + (1-\alpha)\hat{Y}^{(0)}$，其中 $\mathbf{L}$ 为规范化图的拉普拉斯矩阵。

上述两种算法的收敛性与拉普拉斯矩阵的特征值有关，而计算代价最差情况下为 $O(kn^2)$，其中 k 为平均的样本点的邻居数目，当图接近完全图的时候计算代价高达 $O(n^3)$。

与以上两种算法不同，但同样是标记传播算法，Szummer 和 Jaakkola[17]于 2002 年提出了基于马尔可夫随机游走(Markov Random Walk)策略的给图中未标记点进行标记的算法。其中权值矩阵 $\mathbf{W}$ 同样由高斯核矩阵给出，然后利用权值矩阵定义两点之间标记传播的概率 $p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$。然后假设从一个正类的点出发，经过 t 步的随机游走，到达结点 x_k，结点 x_k 为正类的概率为：$P^{(t)}(y_{start} = 1 | x_k) = \sum_{i=1}^{n} P(y=1|x_i)\, p(x_i | x_k)$，如果该式大于 0.5 则判定为正类，否则为负类，其中 $p(x_i | x_k)$ 的概率可以通过 p_{ij} 计算出，而 $P(y=1|x_i)$ 可以通过 EM 算法计算得出。

以上的这些算法其实都是属于同一个框架下的算法特例，根据流型假设，可以通过计算点之间的类别标记的差异性来定义优化目标的代价函数，尤其在将权值矩阵引入到代价函数中去之后，这种想法变得很自然， 根据高斯核矩阵定义的权值矩阵，邻近点的权值较大，那么就迫使邻近点的类别标记的差异性减小。代价可以写成如下形式：

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(\hat{y}_i-\hat{y}_j)^2=\frac{1}{2}\left(2\sum_{i=1}^{n}\hat{y}_i^{\,2}\sum_{i=1}^{n}w_{ij}-2\sum_{i,j=1}^{n}w_{ij}\hat{y}_i\hat{y}_j\right)$$

$$=\hat{Y}^T(\mathbf{D}-\mathbf{W})\hat{Y}=\hat{Y}^T\mathbf{L}\hat{Y}$$

其中 **L** 为非规范化拉普拉斯矩阵。

同时考虑到为了让已标记点的标记的预测结果要与其真实标记一致，以及处理在图中某一个连通分量没有已标记的情况，可以在这三者之间做一个权衡，写出如下的代价函数：

$$C(\hat{Y})=\|\hat{Y}_l-Y_l\|^2+\mu\hat{Y}^T\mathbf{L}\hat{Y}+\mu\xi\|\hat{Y}\|^2$$

其中第一项是为了让已标记结点的预测结果与真实标记一致，第二项就是第一个代价函数，第三项即是为了处理图中某一个连通分量没有已标记样本的情况。当然求解最小化这个代价的问题等价于求图的最小切割问题，这是一个 NP-难的问题[1]。Zhu 等人[16]在 2003 年的利用高斯随机场与谐波函数进行半监督学习的工作中，通过将代价函数中 $\hat{Y}_l$ 的取值从离散的放松到实值，这使得问题变得简单了许多。

基于图的算法其计算开销都很大，很难应用到较大数据集中去，这使得降低计算开销成为了研究这类算法的重要目标，很多研究者在这方面做了很多工作，鉴于篇幅问题，这里不作介绍。

4.1.4 协同训练

标准协同训练算法是 Blum 和 Mitchell[18]在 1998 年提出的。他们提出了标准协同训练算法的三个基本假设：(1)属性集可以被划分为两个集合；(2)每一个属性集的子集合都足以训练一个分类器；(3)在给定类标签的情况下，这两个属性集是相互独立的。其中每个属性集构成一个“视图”(view)，满足上述假设的“视图”被称为充分冗余“视图”。然后分别对已标记的样本在这两个属性集上训练分类器，这样得到两个分类器，将这两个分类器应用到未标记样本上，然后选择每个分类器对分类结果置信度高的未标记样本以及该样本的预测标记加入到另一个分类器已标记样本集中进行下一轮的训练，如此迭代（如算法二所示）。

协同训练算法其实是通过引入未标记数据缩减假设空间来提高学习算法的性能的[1]。

算法二 标准协同训练算法（Blum 和 Mitchell, 1998）

随机从未标记样本集中选择 u 个样本，建立一个未标记样本池（pool）U

进行 k 次迭代：

1. **利用已标记样本集 L，分别于 x 的两个子属性集 x_1, x_2 上训练两个分类器 h_1, h_2**
2. **用分类器 h_1, h_2 对 U 中的 p 个正类与 n 个负类进行标记**
3. **将这些被标记的样本加入到已标记样本集 L 中**
4. **再从未标记样本集中随机选取 2p+2u 个未标记样本，投入到未标记样本池 U 中**

Blum 和 Mitchell 的那三个假设很强，正如 Z.H. Zhou 在文献[29]中所说，在真实的问题中，满足充分冗余的要求往往很难达到。Z.H. Zhou 举 Blum 和 Mitchell 当年所举的网页分类的例子来说明，“因为网页本身的信息这一视图与超链接上的信息这一视图很难满足条件独立性”，而且“大多数问题不具有充分大的属性集”。很多研究人员就尝试放松这三个假设。Goldman 和 Zhou[19]在 2000 年提出了使用不同的分类器在整

个属性集上训练的方法，训练时，首先利用已标记样本对两个不同的分类器在整个属性集上进行训练，再用这两个分类器互相将自己在未标记样本上置信度较高的标记加入到对方的训练集中去再训练。在这个工作之后他们二人在 2004 年[20]又将集成学习的思想加入到他们的方法中去提高算法性能，基于整个属性集训练一组分类器，利用投票机制对未标记样本进行标记，加入到已标记样本集中再训练，最后的分类结果由加权投票机制的一个变种决定。

但是由于 Goldman 和 Zhou 的算法“在挑选未标记示例进行标记的过程中以及选择分类器对未见示例进行预测的过程中频繁地使用 10 倍交叉验证”，使得其计算开销很大，Z.H. Zhou 和 M. Li[21]在 2005 年提出了 tri-training 的算法，使用三个分类器，如果两个分类器分类结果一致，那么就将该未标记样本加入到已标记样本中去，这样的做法避免了频繁地计算 10 倍交叉验证，节省了计算开销，同时他们的算法不需要基于不同的视图，甚至不需要基于不同的分类器。并且他们基于噪音学习理论给出了“以较高概率确保这一做法有效的条件”，在引入大量未标记样本的情况下，噪声所带来的负面影响可以被抵消。此后，他们还将 tri-training 算法扩展为 Co-Forest 算法。

除此之外 Balcan[22]等人在 2005 年放宽对独立性的假设，并调整了协同训练算法的迭代过程，取得较好的结果。与 Balcan 的工作类似，Johnson 和 Zhang 在 2007 年同样放宽了对独立性的假设，提出了一个二视图模型。

以上的算法讨论的都是半监督分类问题，Z.H. Zhou 和 M. Li[23]在 2005 年最先利用协同训练算法进行半监督回归，他们的算法利用流型假设，放宽了对置信度高的标记的判定准则，使得其可以对连续值（而不是离散的类别标签）进行判定。他们的判定准则如下：

$$\Delta_u = \frac{1}{l}\sum_{x_i \in L}\left(y_i - h(x_i)\right)^2 - \frac{1}{l}\sum_{x_i \in L}\left(y_i - h'(x_i)\right)^2$$

其中 $h(x_i)$ 为对未标记样本的预测标记，$h'(x_i)$ 是将预测标记加入到已标记数据集中进行再训练以后的预测标记。他们利用这样的判定准则提出了 COREG 算法，COREG 利用两个基于不同范式的距离度量的 k 近邻回归模型进行协同训练，最后的预测结果是两个 k 近邻回归模型的预测结果的平均。

4.2 半监督学习理论分析

对于半监督学习的理论分析一般是为了解决这样的问题：(1)半监督学习技术如何奏效：未标记样本是否可以提高学习性能？在什么样的情况下，未标记样本又会损害学习性能？(2)半监督学习的样本复杂度：多少的未标记样本是足够的？还需要多少已标记样本？

半监督学习是近年来才成为机器学习领域的热点方向，关于它的理论分析比较少，关于这点我们认为还有原因就是半监督学习技术种类较多，差异性大，建立一个统一的理论分析的模型较为困难，而且半监督学习从一定意义上讲，其实际应用价值大于其理论价值，它是随着实际应用需求的日益强烈而产生的。

关于半监督学习的理论分析的工作，较为经典的是 Balcan 和 Blum[24]在 2005 年的工作，他们为半监督学习提出了一个扩展的 PAC(probably approximately correct)学习模型，是对监督学习的 PAC 学习模型的扩展，他们引入了“相容性”(compatibility)的概念，他们将相容性定义为一个从假设(hypothesis)以及数据分布到[0,1]区间的一个映射，这个映射反映的是我们所相信的一个假设与当前数据分布的相容程度。

与传统的 PAC 学习模型类似，他们假设数据来自一个分布为 D 的样本空间 X ，数据的类别标记由一个目标函数 c^* 给出，假设空间 ς 则是一组样本空间 X 上的给予类别标记的函数的集合。对于假设空间 ς 中的一个假设 f ，其与分布 D 的相容性定义为函数

$$\chi : \varsigma \times X \to [0,1]$$

$\chi(f,D)=\mathbf{E}_{x\sim D}[\chi(f,x)]$，即 $\chi(f,x)$ 在整个分布 D 上的期望。

其中 $\chi(f,x)$ 的定义，根据不同的半监督学习技术而不同，体现了学习目标所需要满足的性质，即相关的假设。比如在 TSVM 中，我们可以定义一个样本点 x 到 f 所决定的分界面的距离大于某个给定的“间隔”γ，则 $\chi(f,x)$ 为 1，否则为 0，当然也可以定义为关于样本点 x 到 f 所决定的分界面的距离的平滑函数，这样的定义体现了分类见面须经过数据较为稀疏区域的聚类假设；在标准的协同训练算法中，学习的数据样本点的表示是 $<x_1,x_2>$ 代表同一个数据的两个视图，学习的目标是一个函数对 $<f_1,f_2>$，那么这个函数对与 $<x_1,x_2>$ 的 $\chi(f,x)$ 可以定义为 $\Pr_{<x_1,x_2>\sim D}[f_1(x_1)=f_2(x_2)]$，即两者学习的预测结果应当经量相同。其他半监督学习技术的 $\chi(f,x)$ 可以有类似的定义。

通过对各个算法相容性的研究，并加上一些适当的假设 Balcan 和 Blum 得到了一些关于样本复杂度的结果，这些结果反映了这样一个直觉意义上的结论，就是在零训练误差以及高相容性的情况下，只需要很少的已标记样本就可以得到一个很好的假设。

Blum 和 Mitchell[18]也对他们提出的标准协同训练算法进行了分析，其中也用到了相容性的概念，并以一个形象化的二部图的形式展示了他们的研究结果。

除了以上的工作，Leskes[25]和 Kaariainen[26]对半监督学习的泛化错误率上界进行了研究。

最近，Singh 等人[27]在“两个聚类的密度是 Lipschitz 连续的”强假设条件下，对未标记样本对学习性能的影响进行了研究，他们得出的结论是：当两个聚类很远的情况下未标记样本对学习没有帮助；靠近但不是十分近的情况下未标记样本能帮助对分类边界的确定；但是很近重叠的不够多的情况下未标记样本没有帮助；重叠足够多以至于密度不连续的情况下又有帮助。

4.3 半监督学习应用

半监督学习应日益强烈的解决实际问题的需求而产生，在实际问题中，半监督学习有着很广泛的应用，这样的实际问题都存在着本文引言中所阐述的背景，对于样本标记的获得需要花费很昂贵的人力劳动，然而未标记样本却是随手可得，比如在语音识别领域，现在的音频很多，而对这些音频加上标记，需要人去听并辨别这些音频再加上标记，相比于未标记的音频有标记的音频少之又少。

下面就对半监督学习的两个典型的应用做简单介绍。

比较典型的应用就是在自然语言处理领域的应用。更由于互联网的日益发达，指数级增长的网络资源，能进行人工标记的网页等的资源是微乎其微，半监督学习技术在这方面得到了很广泛的应用。前面介绍的 Nigam 等人[8]关于生成式模型方面的工作就是利用 EM 算法进行半监督的文本分类。

半监督学习还有一个典型的应用，就是生物学领域对蛋白质序列的分类问题（蛋白质结构预测）。对一种蛋白质的结构进行预测或者功能鉴定需要耗费生物学家很长时间的工作，知道了一个蛋白质表示序列，如何利用少有的有标记样本以及大量的蛋白质序列来预测蛋白质的结构，而半监督学习技术则是为了解决这类问题而设计的，这使得半监督学习在这个问题上被广泛研究。比如 Weston 等人[28]利用聚类核方法对蛋白质的序列进行半监督分类；Shin 和 Tsuda[1]利用基于图的半监督学习算法对蛋白质的功能进行预测。

5. 结束语

本文从半监督学习的发展历史、经典算法、理论分析和实际应用四个方面对近年来半监督学习方面的研究工作进行了简单的介绍。由于各方面的原因，半监督学习起步较晚，目前的研究还不成熟。

在应用方面，半监督学习技术存在很多需要解决的问题，对半监督学习技术的研究不仅需要关注如何

利用未标记样本提高学习结果的准确率，而且更需要关注的是如何解决“大量”的未标记样本的计算代价问题，目前很多的算法只能对于很小的数据集奏效，而对于具有大量的未标记样本的实际问题计算开销过大。各个算法本身也存在着很多问题，如生成式模型中利用 EM 算法的局部解的问题。如果半监督学习算法的假设与实际问题不一致，那引入未标记样本也许对学习的性能会有降低，如何更好地为实际问题选择适合的半监督学习算法也是很重要的课题，Chapelle 等人[1]做出了一些尝试，邀请了一些半监督学习算法的提出者让他们利用他们的算法对一系列的数据集做实验，并对结果进行了分析。

在理论分析方面，对于未标记样本对学习性能的影响的研究还不够深入，对半监督学习的理论研究还很少，已有的研究也是建立在很强的假设基础上的，然而对于一些实际的问题，可能并不满足这些假设，但是半监督学习算法依然奏效，这需要研究人员给出对半监督学习的更一般的情况进行分析研究。

6. 感谢

感谢 XX 老师对我的指导！

感谢 XX 对我的细心解答！

感谢 XX 其他成员对我的帮助！

References:

[1] O. Chapelle, B. Schölkopf and A. Zien. *Semi-Supervised Learning*. Cambridge, MA:The MIT Press, London, England, 2006

[2] H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11:363–371, 1965

[3] S. C. Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 13:57–64, 1967

[4] A. K. Agrawala. Learning with a probabilistic teacher. *IEEE Transactions on Information Theory*, 16:373–379, 1970

[5] D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In: M. Mozer, M. I. Jordan, T. Petsche, eds. *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, 571-577, 1997

[6] C. J. Merz, D. C. St. Clair, and W. E. Bond. Semi-supervised adaptive resonance theory (smart2).*In Proceedings of: International Joint Conference on Neural Networks*, volume 3, pages 851–856, 1992

[7] H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.* Volume 34, Number 4, 1265-1269, 1963

[8] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000

[9] K. Nigam. Using unlabeled data to improve text classification. *Technical Report doctoral dissertation*, CMU-CS-01-126, Carnegie Mellon University, Pittsburgh, 2001

[10] X. Zhu. Semi-supervised learning literature survey. *Technical Report 1530*, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Apr. 2008

[11] T. De Bie and N. Cristianini. Convex methods for transduction. *In Advances in Neural Information Processing Systems 16*, pages 73–80, 2004

[12] N. D. Lawrence, M. I. Jordan. Semi-supervised learning via Gaussian processes. In: L. K. Saul, Y. Weiss, and L. Bottou, eds. *Advances in Neural Information Processing Systems 17*, Cambridge, MA: MIT Press, 753-760, 2005

[13] Y. Grandvalet, Y. Bengio. Semi-supervised learning by entropy minimization. In: L. K. Saul, Y. Weiss, and L.

Bottou, eds. *Advances in Neural Information Processing Systems 17*, Cambridge, MA: MIT Press, 529-536,2005

[14] X. Zhu and Z. Ghahramani. Towards semi-supervised classification with Markov random fields*Technical Report CMU-CALD-02-106*. Carnegie Mellon University, 2002

[15] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch¨olkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Sch¨olkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 321–328. Cambridge, MA: MIT Press, 2004

[16] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In: *Proceedings of The 20th International conference on Machine Learning* , 2003

[17] M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA:MIT Press, 2002

[18] A. Blum, T. Mitchell. Combining labeled and unlabeled data with co-training. In: *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Wisconsin, MI, 92-100, 1998

[19] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In: *Proceedings of 17th International Conf. on Machine Learning* . pp. 327–334, Morgan Kaufmann, San Francisco, CA, 2000

[20] Y. Zhou and S. Goldman. Democratic co-learing. In: *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, 2004

[21] Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11): 1529–1541, 2005

[22] M.F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In L. K. Saul, Y. Weiss and L. Bottou Eds., *Advances in neural information processing systems 17*. Cambridge, MA: MIT Press, 2005

[23] Z.H. Zhou and M. Li. Semi-supervised learning with co-training. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 908-913, 2005

[24] M.F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In: *Proceedings of In Conference on Computational Learning Theory*, pages 111–126, 2005

[25] B. Leskes. The value of agreement, a new boosting algorithm. In: *Proceedings of the International Conference Of Learning Theory*, 2005

[26] M. Kaariainen. Generalization error bounds using unlabeled data. In: *Proceedings of the International Conference Of Learning Theory*, 2005

[27] Aarti Singh, Robert Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. *In Advances in Neural Information Processing Systems 22*, 2008

[28] E. Ie, J. Weston, W. S. Noble, and C. Leslie. Multi-class protein fold recognition using adaptive codes. In: *Proceedings of the International Conference on Machine Learning*, 2005

[29] Z.H. Zhou. Co-training Paradigm in Semi-supervised Learning. In: *Proceedings of the Chinese Workshop on Machine Learning and Applications*, Nanjing, China, 2007