半监督学习中的协同训练风范*

周 志 华

南京大学计算机软件新技术国家重点实验室，南京 210093

1. 引言

在传统的监督学习中，学习器通过对大量有标记的（labeled）训练例进行学习，从而建立模型用于预测未见示例的标记。这里的“标记”（label）是指示例所对应的输出，在分类问题中标记就是示例的类别，而在回归问题中标记就是示例所对应的实值输出。随着数据收集和存储技术的飞速发展，收集大量未标记的（unlabeled）示例已相当容易，而获取大量有标记的示例则相对较为困难，因为获得这些标记可能需要耗费大量的人力物力。例如在计算机辅助医学图像分析中，可以从医院获得大量的医学图像作为训练例，但如果要求医学专家把这些图像中的病灶都标识出来，则往往是不现实的。事实上，在真实世界问题中通常存在大量的未标记示例，但有标记示例则比较少，尤其是在一些在线应用中这一问题更加突出。例如，在进行 Web 网页推荐时，需要用户标记出哪些网页是他感兴趣的，很少会有用户愿意花大量的时间来提供标记，因此有标记的网页示例比较少，但 Web 上存在着无数的网页，它们都可作为未标记示例来使用。

显然，如果只使用少量的有标记示例，那么利用它们所训练出的学习系统往往很难具有强泛化能力；另一方面，如果仅使用少量“昂贵的”有标记示例而不利用大量“廉价的”未标记示例，则是对数据资源的极大的浪费。因此，在有标记示例较少时，如何利用大量的未标记示例来改善学习性能已成为当前机器学习研究中最受关注的问题之一。

目前，利用未标记示例的主流学习技术主要有三大类[Zhou06]，即半监督学习（semi-supervised learning）、直推学习（transductive learning）和主动学习（active learning）。这三类技术都是试图利用大量的未标记示例来辅助对少量有标记示例的学习，但它们的基本思想却有显著的不同。在半监督学习[ChapelleSZ06][Zhu06]中，学习器试图自行利用未标记示例，即整个学习过程不需人工干预，仅基于学习器自身对未标记示例进行利用。直推学习[Vapnik98][Joachims99]与半监督学习的相似之处是它也是由学习器自行利用未标记示例，但不同的是，直推学习假定未标记示例就是测试例，即学习的目的就是在这些未标记示例上取得最佳泛化能力。换句话说，半监督学习考虑的是一个“开放世界”，即在进行学习时并不知道要预测的示例是什么，而直推学习考虑的则是一个“封闭世界”，在学习时已经知道了需要预测哪些示例。实际上，直推学习这一思路直接来源于统计学习理论

* 本文得到国家自然科学基金(60635030)和全国优秀博士学位论文作者专项基金(200343)资助

[Vapnik98]，并被一些学者认为是统计学习理论对机器学习思想的最重要的贡献[1]。其出发点是不要通过解一个困难的问题来解决一个相对简单的问题。V. Vapnik认为，经典的归纳学习假设期望学得一个在整个示例分布上具有低错误率的决策函数，这实际上把问题复杂化了，因为在很多情况下，人们并不关心决策函数在整个示例分布上性能怎么样，而只是期望在给定的要预测的示例上达到最好的性能。后者比前者简单，因此，在学习过程中可以显式地考虑测试例从而更容易地达到目的。这一思想在机器学习界目前仍有争议，但直推学习作为一种重要的利用未标记示例的技术，则已经受到了众多学者的关注。主动学习[SeungOS92][LewisG94][AbeM98]和前面两类技术不同，它假设学习器对环境有一定的控制能力，可以“主动地”向学习器之外的某个“神谕”(oracle)[2] 进行查询来获得训练例的标记。因此，在主动学习中，学习器自行挑选出一些未标记示例并通过神谕查询获得这些示例的标记，然后再将这些有标记示例作为训练例来进行常规的监督学习，而其技术难点则在于如何使用尽可能少的查询来获得强泛化能力。对比半监督学习、直推学习和主动学习可以看出，后者在利用未标记示例的过程中需要与外界进行交互，而前两者则完全依靠学习器自身，正因为此，也有一些研究者将直推学习作为一种半监督学习技术来进行研究。

本章的主旨是介绍半监督学习中的协同训练（co-training）这一风范（paradigm），因此，对直推学习和主动学习不再做更多的介绍，仅在第 2 节对半监督学习的概况做一简要描述。第 3 至 5 节将从学习算法、理论分析、实际应用等三个方面来介绍协同训练的研究进展，第 6 节则列出几个可能值得进一步研究的问题。

2. 半监督学习

一般认为，半监督学习的研究始于 B. Shahshahani 和 D. Landgrebe 的工作[ShahshahaniL94]，但未标记示例的价值实际上早在上世纪 80 年代末就已经被一些研究者意识到了[Lippman89]。D.J. Miller 和 H.S. Uyar [MillerU97]认为，半监督学习的研究起步相对较晚，可能是因为在当时的主流机器学习技术（例如前馈神经网络）中考虑未标记示例相对比较困难。随着统计学习技术的不断发展，以及利用未标记示例这一需求的日渐强烈，半监督学习才在近年来逐渐成为一个研究热点。

半监督学习的基本设置是给定一个来自某未知分布的有标记示例集$\boldsymbol{L}=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{|\boldsymbol{L}|}, y_{|\boldsymbol{L}|})\}$以及一个未标记示例集$\boldsymbol{U} = \{\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_{|\boldsymbol{U}|}'\}$，期望学得函数$f: X\rightarrow Y$可以准确地对示例$\boldsymbol{x}$ 预测其标记y。这里$\boldsymbol{x}_i, \boldsymbol{x}_j' \in X$ 均为d维向量，$y_i\in Y$为示例$\boldsymbol{x}_i$的标记，$|\boldsymbol{L}|$和$|\boldsymbol{U}|$分别为$\boldsymbol{L}$和$\boldsymbol{U}$的大小，即它们所

[1] 有人认为统计学习理论的最重要贡献是支持向量机，但实际上，支持向量机只是对结构风险最小化原则的一个实现，在处理非线性时用到了核技巧（kernel trick）。结构风险最小化的思想在机器学习中早已有之，只是以往的研究没有适时地总结成一套完整的框架；至于核技巧，则在机器学习和模式识别领域早就在使用了。而直推学习则是和经典的归纳学习很不相同的一个思路。

[2] 这里的“神谕”可以是人，也可以是能够为示例提供真实标记的其他过程。

包含的示例数。

在介绍具体的半监督学习技术之前，有必要先探讨一下为什么可以利用未标记示例来改善学习性能。关于这个问题，有不少研究者给出了解释。例如，D.J. Miller 和 H.S. Uyar [MillerU97] 从数据分布估计的角度给出了一个直观的分析。他们假设所有数据服从于某个由 L 个高斯分布混合而成的分布，即

$$f\left(x|\theta\right)=\sum_{l=1}^{L}\alpha_l f\left(x|\theta_l\right) \tag{1}$$

其中 $\sum_{l=1}^{L}\alpha_l=1$ 为混合系数，$\theta=\{\theta_l\}$ 为参数。这样，标记就可视为一个由选定的混合成分 m_i 和特征向量 $\boldsymbol{x}_i$ 以概率 $\mathrm{P}(c_i \mid \boldsymbol{x}_i, m_i)$ 决定的随机变量。于是，根据最大后验概率假设，最优分类由式 2 给出：

$$h\left(x\right)=\arg\max_{k}\sum_{j}\mathrm{P}\left(c_i=k|m_i=j,x_i\right)\mathrm{P}\left(m_i=j|x_i\right) \tag{2}$$

其中 $\mathrm{P}\left(m_i=j|x_i\right)=\dfrac{\alpha_j f\left(x_i|\theta_j\right)}{\sum_{l=1}^{L}\alpha_l f\left(x_i|\theta_l\right)}$ 。

这样，学习目标就变成了利用训练例来估计 $P(c_i = k \mid m_j = j, \boldsymbol{x}_i)$ 和 $P(m_i = j \mid \boldsymbol{x})$。这两项中的第一项与类别标记有关，而第二项并不依赖于示例的标记，因此，如果有大量的未标记示例可用，则意味着能够用于估计第二项的示例数显著增多，这会使得第二项的估计变得更加准确，从而导致式 2 更加准确，也就是说，分类器的泛化能力得以提高。此后，T. Zhang和F. J. Oles [ZhangO00] 进一步分析了未标记示例在半监督学习中的价值，并指出如果一个参数化模型如果能够分解成 $P(\boldsymbol{x}, y \mid \theta) = P(y \mid \boldsymbol{x}, \theta)\ P(\boldsymbol{x} \mid \theta)$ 的形式，那么未标记示例的价值就体现在它们能够帮助更好地估计模型参数从而导致模型性能的提高。

实际上，只要能够合理建立未标记示例分布和学习目标之间的联系，就可以利用未标记示例来辅助提高学习性能。在[ShahshahaniL94][MillerU97]中，这一联系是通过对生成式模型（generative model）参数的估计来体现的，但在更一般的情况下就需要在某些假设的基础上来建立未标记示例和目标之间的联系。目前，在半监督学习中有两个常用的基本假设，即聚类假设（cluster assumption）和流形假设（manifold assumption）。

聚类假设是指处在相同聚类（cluster）中的示例有较大的可能拥有相同的标记。根据该假设，决策边界就应该尽量通过数据较为稀疏的地方，从而避免把稠密的聚类中的数据点分到决策边界两侧。在这一假设下，大量未标记示例的作用就是帮助探明示例空间中数据分布的稠密和稀疏区域，从而指导学习算法对利用有标记示例学习到的决策边界进行调整，使其尽量通过数据分布的稀疏区域。

聚类假设简单、直观，常以不同的方式直接用于各种半监督学习算法的设计中。例如，T. Joachims

[Joachims99] 提出了TSVM算法[3]，在训练过程中，该算法不断修改SVM的划分超平面并交换超平面两侧某些未标记示例的可能标记，使得SVM在所有训练数据（包括有标记和未标记示例）上最大化间隔（margin），从而得到一个既通过数据相对稀疏的区域又尽可能正确划分有标记示例的超平面；N. D. Lawrence和 M. I. Jordan [LawrenceJ05] 通过修改高斯过程（Gaussian process）中的噪音模型来进行半监督学习，他们在正、反两类之间引入了“零类”，并强制要求所有的未标记示例都不能被分为零类，从而迫使学习到的分类边界避开数据稠密区域；Y. Grandvalet和Y. Bengio [GrandvaletB05] 通过使用最小化熵作为正则化项来进行半监督学习，由于熵仅与模型在未标记示例上的输出有关，因此，最小化熵的直接结果就是降低模型的不确定性，迫使决策边界通过数据稀疏区域。

流形假设是指处于一个很小的局部邻域内的示例具有相似的性质，因此，其标记也应该相似。这一假设反映了决策函数的局部平滑性。和聚类假设着眼整体特性不同，流形假设主要考虑模型的局部特性。在该假设下，大量未标记示例的作用就是让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性，使得决策函数能够更好地进行数据拟合。

流形假设也可以容易地直接用于半监督学习算法的设计中。例如，J. Zhu 等人 [ZhuGL03] 使用高斯随机场以及谐波函数来进行半监督学习，他们首先基于训练例建立一个图，图中每个结点就是一个（有标记或未标记）示例，然后求解根据流形假设定义的能量函数的最优值，从而获得对未标记示例的最优标记；D. Zhou 等人 [ZhouBLWS04] 在根据示例相似性建立图之后，让示例的标记信息不断向图中的邻近示例传播，直到达到全局稳定状态。

值得注意的是，一般情形下，流形假设和聚类假设是一致的。由于聚类通常比较稠密，满足流形假设的模型能够在数据稠密的聚类中得出相似的输出。然而，由于流形假设强调的是相似示例具有相似的输出而不是完全相同的标记，因此流行假设比聚类假设更为一般，这使其在聚类假设难以成立的半监督回归中仍然有效[ZhouL05b][ZhouL07]。

根据半监督学习算法的工作方式，可以大致将现有的很多半监督学习算法分为三大类。第一类算法以生成式模型为分类器，将未标记示例属于每个类别的概率视为一组缺失参数，然后采用 EM 算法来进行标记估计和模型参数估计，其代表包括[ShahshahaniL94][MillerU97] [NigamMTM00]等。此类算法可以看成是在少量有标记示例周围进行聚类，是早期直接采用聚类假设的做法。第二类算法是基于图正则化框架的半监督学习算法，其代表包括[BlumC01][ZhuGL03][BelkinN04][ZhouBLWS04][BelkinNS05]等。此类算法直接或间接地利用了流形假设，它们通常先根据训练例及某种相似度度量建立一个图，图中结点对应了（有标记或未标记）示例，边为示例间的相似度，然后，定义所需优化的目标函数并使用决策函数在图上的光滑性作为正则化项来求取最优模型参数。第三类算法是协同训练（co-training）算法。此类算法隐含地利用了聚类假设或流形假设，它们使用两个或多个学习器，在学习过程中，这些学习器挑选若干个置信度高的未标记示例进行相互标记，从而使得模型得以更新。在 A. Blum 和 T. Mitchell [BlumM98] 提出最早的协同训练算法后，很多研

[3] 这实际上是一个直推算法。

究者对其进行了研究并取得了很多进展，使得协同训练成为半监督学习中最重要的风范（paradigm）之一，而不再只是一个算法。本章接下来的几节就将对协同训练进行进一步的介绍。

3. 协同训练算法

最初的协同训练算法（或称为标准协同训练算法）是A. Blum和T. Mitchell [BlumM98] 在 1998 年提出的。他们假设数据集有两个充分冗余（sufficient and redundant）的视图（view），即两个满足下述条件的属性集：第一，每个属性集都足以描述该问题，也就是说，如果训练例足够，在每个属性集上都足以学得一个强学习器；第二，在给定标记时，每个属性集都条件独立于另一个属性集。A. Blum和T. Mitchell认为，充分冗余视图这一要求在不少任务中是可满足的。例如，在一些网页分类问题上，既可以根据网页本身包含的信息来对网页进行正确分类，也可以利用链接到该网页的超链接所包含的信息来进行正确分类，这样的网页数据就有两个充分冗余视图，刻画网页本身包含的信息的属性集构成第一个视图，而刻画超链接所包含的信息的属性集构成第二个视图。A. Blum和T. Mitchell的算法在两个视图上利用有标记示例分别训练出一个分类器，然后，在协同训练过程中，每个分类器从未标记示例中挑选出若干标记置信度（即对示例赋予正确标记的置信度）较高的示例进行标记，并把标记后的示例加入另一个分类器的有标记训练集中，以便对方利用这些新标记的示例进行更新。协同训练过程不断迭代进行，直到达到某个停止条件。该算法如图 1 所示，其中$\boldsymbol{x}_1$和$\boldsymbol{x}_2$分别指示例$\boldsymbol{x}$在第 1 视图和第 2 视图上对应的示例。A. Blum和T. Mitchell [BlumM98] 对图 1 的算法进行了分析，证明了在充分冗余视图这一条件成立时，图 1 算法可以有效地通过利用未标记示例提升学习器的性能，实验也验证了该算法具有较好的性能。

Input: the labeled training set L

the unlabeled training set U

Process:

Create a pool U' of examples by choosing u examples at random from U

Loop for k iterations:

Use L to train a classifier h_1 that considers only the x_1 portion of x

Use L to train a classifier h_2 that considers only the x_2 portion of x

Allow h_1 to label p positive and n negative examples from U'

Allow h_2 to label p positive and n negative examples from U'

Add these self-labeled examples to L

Randomly choose $2p+2n$ examples from U to replenish U'

图 1　标准协同训练算法 [BlumM98]

然而，在真实问题中充分冗余视图这一要求往往很难得到满足。实际上，即使对 A. Blum 和 T. Mitchell 所举的网页分类的例子来说也是这样，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。K. Nigam 和 R. Ghani [NigamG] 对协同训练算法在不具有充分冗余视图的问题上的性能进行了实验研究，其结果表明，在属性集充分大时，可以随机把属性集划分成两个视图，在此基础上进行协同训练也可能取得较好的效果。遗憾的是，大多数的问题并不具有“充分大”的属性集，而且随机划分视图这一策略并非总能奏效，因此，一些研究者开始试图设计不需要充分冗余视图的协同训练算法。

S. Goldman 和 Y. Zhou [GoldmanZ00] 提出了一种不需要充分冗余视图的协同训练算法。他们使用不同的决策树算法，从同一个属性集上训练出两个不同的分类器，每个分类器都可以把示例空间划分为若干个等价类。在协同训练过程中，每个分类器通过统计技术来估计标记置信度，并且把标记置信度最高的示例进行标记后提交给另一个分类器作为有标记训练例，以便对方进行更新。该过程反复进行，直到达到某个停止条件。在预测阶段，该算法先估计两个分类器对未见示例的标记置信度，然后选择置信度高的分类器进行预测。S. Goldman 和 Y. Zhou 将该算法建立在 A. Angluin 和 P. Laird [AngluinL88] 的噪音学习理论的基础上，并通过实验对算法性能进行了验证。此后，他们 [ZhouG04] 又对该算法进行了扩展，使其能够使用多个不同种类的分类器。

虽然 S. Goldman 和 Y. Zhou 的算法 [GoldmanZ00] 不再要求问题本身具有充分冗余视图，但他们引入了对分类器种类的限制。此外，他们为了估计标记置信度，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，时间开销很大。同时，在少量有标记数据上进行 10 倍交叉验证经常难以得到对置信度的稳定估计。

为了进一步放松协同训练的约束条件，Z.-H. Zhou 和 M. Li [ZhouL05a] 提出了一种既不要求充分冗余视图、也不要求使用不同类型分类器的 tri-training 算法。该算法的一个显著特点是使用了三个分类器，不仅可以简便地处理标记置信度估计问题以及对未见示例的预测问题，还可以利用集成学习（ensemble learning） [Dietterich00] 来提高泛化能力。该算法首先对有标记示例集进行可重复取样（bootstrap sampling）以获得三个有标记训练集，然后从每个训练集产生一个分类器。在协同训练过程中，各分类器所获得的新标记示例都由其余两个分类器协作提供，具体来说，如果两个分类器对同一个未标记示例的预测相同，则该示例就被认为具有较高的标记置信度，并在标记后被加入第三个分类器的有标记训练集。在对未见示例进行预测时，tri-training 算法不再象以往算法那样挑选一个分类器来使用，而是使用集成学习中经常用到的投票法来将三个分类器组成一个集成来实现对未见示例的预测。

与以往协同训练算法需要显式地对标记置信度进行估计不同，tri-training 算法通过判断三个分类器的预测一致性来隐式地对不同未标记示例的标记置信度进行比较，这一做法使得该算法不需要频繁地使用耗时的统计测试技术。但与显式估计标记置信度相比，这一隐式处理往往不够准确，特别是如果初始分类器比较弱，未标记示例可能被错误标记，从而给第三个分类器的训练引入噪音。Z.-H. Zhou 和 M. Li [ZhouL05a] 基于噪音学习理论 [AngluinL88] 推导出了能以较高概率确保这一做法有效的条件，直观地说，如果大多数未标记示例的标记是准确的，那么引入的噪音所带来的负面影响

可以被使用大量未标记示例所带来的好处抵消。为了进一步降低噪音影响，有必要使用一些更可靠的误差估计技术，但这会在一定程度上增大算法的开销。此后，M. Li 和 Z.-H. Zhou [LiZ07] 对 tri-training 进行了扩展，提出了可以更好发挥集成学习作用的 Co-Forest 算法。Tri-training 算法最近被 D. Mavroeidis 等人 [MavroeidisCPCV06] 用来参加欧洲机器学习/数据挖掘竞赛 ECML/PKDD 2006 Discovery Challenge 并获得了较好的名次。

以往的半监督学习研究几乎都是关注分类问题[4]，虽然在监督学习中回归问题的重要性不亚于分类问题，半监督回归却一直缺乏研究。如第二节所述，在半监督回归中由于示例的标记是实值输出，因此聚类假设不再成立，但半监督学习的流形假设仍然是成立的，而且因为回归输出通常具有平滑性，所以流形假设在回归问题中可能比在分类问题中更加有效。因此，如Zhu [Zhu06] 所述，一些基于流形假设的半监督学习技术，例如图正则化算法，在理论上是可以推广到半监督回归中去的。但实际上，此类技术由于要先建立图再进行标记传播，因此若直接推广则只能进行直推回归，要进行半监督回归还需要做一些其他处理。

Z.-H. Zhou和M. Li [ZhouL05b] 最早使用协同训练技术进行半监督回归。在回归问题中，由于示例的属性是连续的实数值，这就使得以往协同训练算法中所使用的标记置信度估计技术难以直接使用。为此，他们提出了一个选择标记置信度最高的未标记示例的准则——标记置信度最高的未标记示例是在标记后与学习器的有标记训练集最一致的示例。更严格的表述是，令h表示当前学习器学得的模型，$\boldsymbol{L}$表示有标记示例集，$\boldsymbol{x}_u \in \boldsymbol{U}$表示一个未标记示例，$h'$表示把$h$标记过的示例$(\boldsymbol{x}_u, h(\boldsymbol{x}_u))$加入训练集后重新训练得到的学习器，则标记置信度最高的未标记示例是在$\boldsymbol{U}$中最大化式 3 的示例。

$$\Delta_u = \frac{1}{|L|}\sum_{x_i \in L}\left(y_i - h(x_i)\right)^2 - \frac{1}{|L|}\sum_{x_i \in L}\left(y_i - h'(x_i)\right)^2 \qquad (3)$$

实际上，式 3 也可以用于半监督分类。基于式 3，Z.-H. Zhou 和 M. Li [ZhouL05b] 提出了 COREG 算法，该算法不要求充分冗余视图，而是通过使用同一学习器的不同参数设置来生成两个初始学习器。具体来说，他们使用了基于不同阶 Minkowski 距离的两个 k 近邻回归模型作为学习器，在协同训练过程中，两个学习器根据式 3 挑选未标记示例进行标记供对方进行更新。最后的回归预测通过对两个 k 近邻回归模型预测值的平均来完成。此后，他们 [ZhouL07] 又将 COREG 推广到使用不同距离度量、不同近邻个数以及其他回归模型的情况。

最近，U. Brefeld 等人 [BrefeldGSW06] 把基于协同训练的半监督回归思想移植到正则化框架下，通过最小化不同视图下回归模型对未标记示例的预测差异来改善各视图的回归模型，也取得了很好的效果。

[4] 半监督聚类已有不少研究，但由于聚类本身是一种非监督学习技术，因此半监督聚类的出发点与半监督分类、回归等期望利用大量未标记示例来辅助对少量有标记示例的学习很不相同，而且其所利用的额外信息也并非未标记示例，而是有标记示例、示例相似性约束等，所以，本章未对半监督聚类进行讨论。

4. 协同训练理论分析

在提出标准协同训练算法时，A. Blum和T. Mitchell [BlumM98] 就对该技术能够奏效的原因进行了探讨。令X_1和X_2分别表示X的两个视图，则一个示例就可以表示为$(\boldsymbol{x}_1, \boldsymbol{x}_2)$，其中$\boldsymbol{x}_1$是$\boldsymbol{x}$在$X_1$视图中的特征向量，$\boldsymbol{x}_2$则是其在$X_2$视图中的特征向量。假设$f$是在示例空间$X$中的目标函数，若$\boldsymbol{x}$的标记为$l$则应有$f(\boldsymbol{x}) = f_1(\boldsymbol{x}_1) = f_2(\boldsymbol{x}_2) = l$。因此，A. Blum和T. Mitchell定义了所谓的“相容性”（compatibility），即对X上的某个分布D，C_1和C_2分别是定义在X_1和X_2上的概念类，如果D对满足$f_1(\boldsymbol{x}) \neq f_2(\boldsymbol{x}_2)$ 的示例 $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ 指派零概率，则称目标函数$f = (f_1, f_2) \in C_1 \times C_2$与$D$“相容”。

基于相容性概念，A. Blum和T. Mitchell揭示了协同训练设置下的一个有趣的现象——即使C_1和C_2是复杂度很高（VC-维很高）的大概念类，与分布D相容的目标概念集相对来说仍然可能会小得多、简单得多。这样，就有可能利用未标记示例来辅助探查哪些目标概念是相容的，而该信息有助于减少学习算法所需的有标记示例数。他们借助于图 2 来直观地展示这一现象。图中二部图左边的每个结点对应了X_1中的一个特征向量，右边的每个结点对应了X_2中的一个特征向量，当且仅当示例$(\boldsymbol{x}_1, \boldsymbol{x}_2)$在分布$D$下以非零概率存在时，结点$x_1$和$x_2$之间才存在边，这些边在图中已经用线条标示出来，其中用实边标示的边对应了已经观察到的未标记示例。在这一表示下，C中与D相容的概念就对应了在图中连通成分之间没有交叉线的划分。显然，属于同一连通成分的示例必然属于同样的类别，而未标记示例可以帮助学习算法了解图中的连通性（实际上也就是了解分布D），因此，通过利用未标记示例，学习算法可以使用较少的有标记示例达到原来需要更多的有标记示例才能达到的效果。

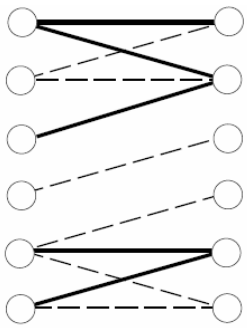

图 2　示例分布的二部图表示 [BlumM98]

进一步，A. Blum和T. Mitchell [BlumM98] 证明了一个定理：如果C_2在有分类噪音时是PAC可学习的，并且两个视图具有条件独立性，那么给定一个初始的弱有效（weakly-useful）学习器$h(\boldsymbol{x}_1)$，协同训练算法只需使用未标记示例就可以学得(C_1, C_2)。这是一个非常强的结论，它意味着只要两个视图的条件独立性成立，那么通过协同训练技术，仅利用未标记示例就可以将一个从有标记示例学得的弱学习器提升到任意精度。

A. Blum和T. Mitchell没有推导出协同训练算法的泛化误差界，为此，Dasgupta等人 [DasguptaLM02] 进行了研究。令S表示一个独立同分布取样；对断言$\Phi[s]$，令$S(\Phi)$表示S中满足该断言的子集$\{s_j: \Phi[s_j]\}$；对两个断言Φ和Ψ，定义经验估计 $\widehat{P}\left(\Phi \middle| \Psi\right) = \left|S\left(\Phi \wedge \Psi\right)\right| / \left|S\left(\Psi\right)\right|$；令$k$表示类

别数；如果学习器h无法判断$\boldsymbol{x}$的类别，则表示为$h(\boldsymbol{x})=\perp$；令$|h|$表示对h的复杂度的一个度量。Dasgupta 等人推导出这样的结论：在S上至少以 1-δ 的概率，对任何一对h_1和h_2来说只要对所有的 $1\le i\le k$都有 $\gamma_i(h_1, h_2, \delta/2)>0$ 以及$b_i(h_1, h_2, \delta/2)\le(k-1)/k$，就有

$$\mathrm{error}(h_1)\le\left(\hat{P}(h_1\neq\perp)-\varepsilon\left(|h_1|,\delta/2\right)\right)\max_j b_j(h_1,h_2,\delta/2)+\frac{k-1}{k}\left(\hat{P}(h_1=\perp)+\varepsilon\left(|h_1|,\delta/2\right)\right) \quad (4)$$

其中 $\varepsilon(k,\delta)=\sqrt{\dfrac{k\ln 2+\ln 2/\delta}{2|S|}}$，

$$b_i(h_1,h_2,\delta)=\frac{1}{\gamma_i(h_1,h_2,\delta)}\left(\hat{P}(h_1\neq i\,|\,h_2=i,h_1\neq\perp)+\varepsilon_i(h_1,h_2,\delta)\right)$$

$$\varepsilon_i(h_1,h_2,\delta)=\sqrt{\frac{(\ln 2)(|h_1|+|h_2|)+\ln\dfrac{2k}{\delta}}{2|S(h_2=i,h_1\neq\perp)|}}$$

$$\gamma_i(h_1,h_2,\delta)=\hat{P}(h_1=i\,|\,h_2=i,h_1\neq\perp)-\hat{P}(h_1\neq i\,|\,h_2=i,h_1\neq\perp)-2\varepsilon_i(h_1,h_2,\delta)$$

值得注意的是，A. Blum 和 T. Mitchell 的工作以及 Dasgupta 等人的工作都假定两个视图间的条件独立性假设成立，但如本章第三节所述，实际上该假设通常是不成立的。这就使得 A. Blum 和 T. Mitchell 的分析结论以及 Dasgupta 等人所推出的协同训练误差上界实际上都只能是理想情况，未必能够适用于实际情况。

M.-F. Balcan等人 [BalcanBY05] 进行了进一步的研究，发现对协同训练技术来说，如果在每个视图上有合适的强学习器，则两个视图的条件独立性假设甚至连弱独立性假设 [Abney02] 都不是必需的，只要数据分布满足比上述假设弱得多的“扩张性”（expansion）假设，迭代式的协同训练算法就可以奏效。“扩张性”是如下定义的：令X^+表示X中的正区域，D^+表示D在X^+上的分布；对$S_1\subseteq X_1$和$S_2\subseteq X_2$，令$\mathbf{S}_i$ (i = 1,2) 表示示例$(\boldsymbol{x}_1, \boldsymbol{x}_2)$有$x_i\in S_i$；令$P(\mathbf{S}_1\wedge\mathbf{S}_2)$表示对$\mathbf{S}_1$和$\mathbf{S}_2$都确信的概率，$P(\mathbf{S}_1\oplus\mathbf{S}_2)$表示对$\mathbf{S}_1$和$\mathbf{S}_2$中至少一个确信的概率；令$H_i\cap X_i^+$表示$\{h\cap X_i^+ : h\in H_i\}$，其中$H_i$（$i$ = 1,2）是假设类。若式 5 对任何$S_1\subseteq X_1{}^+$和$S_2\subseteq X_2{}^+$都成立，则称D^+是ε 扩张的（ε-expanding）；若式 5 对任何$S_1\subseteq H_1\cap X_1{}^+$和$S_2\subseteq H_2\cap X_2{}^+$都成立，则称$D^+$对假设类$H_1\times H_2$来说是$\varepsilon$ 扩张的。

$$P(\boldsymbol{S}_1\oplus\boldsymbol{S}_2)\ge\varepsilon\min\left(P(\boldsymbol{S}_1\wedge\boldsymbol{S}_2),P(\overline{\boldsymbol{S}_1}\wedge\overline{\boldsymbol{S}_2})\right) \quad (5)$$

直观地说，在满足扩张性的数据分布上，对一个与视图j（j =1,2）上的模型所对应的较小的确信集（confidence set）S_j来说，可以利用S_j所导出的另一个视图 3-j上的条件分布对该视图上的正例进行采样，如果利用采样所得的示例学得一个误差小于ε的模型，那么在视图 3-j上的示例出现在该模型所对应的确信集S_{3-j}中的概率将大于出现在S_j中的概率。

值得注意的是，实际使用的协同训练算法（例如第三节中描述的算法）实际上都是迭代式协同

训练算法，而通常在每个视图上使用的都是强学习器[5]，因此，M.-F. Balcan等人的工作在一定程度上解释了为什么两个视图的条件独立性虽然通常不成立，但协同训练算法仍能取得好的效果。

最近，W. Wang 和 Z.-H. Zhou [WangZ07]又做了进一步的分析。一方面，他们证明了只要两个学习器有较大的差异，就可以通过协同训练来利用未标记示例提高学习性能。这不仅解释了为什么在两个视图的条件独立性不成立时协同训练算法可以有好的效果，还解释了那些根本不利用两个视图的算法，例如[GoldmanZ00][ZhouL05b]等奏效的原因。另一方面，从以往的理论分析来看，使用协同训练总可以使得泛化能力提高，甚至可以将弱学习器提升到任意精度；然而，在实际使用协同训练时往往出现这样的情况，即在若干轮协同训练之后如果再继续进行下去，不仅不能改善学习结果，有时甚至会导致性能下降。W. Wang 和 Z.-H. Zhou [WangZ07]对此问题也给出了理论解释。

5. 协同训练的应用

自然语言处理是协同训练技术应用得最为广泛的一个领域。实际上，该领域的研究者在协同训练技术出现之前就已经意识到可以利用问题本身具有的不同属性集来建立模型。例如，D. Yarowsky [Yarowsky95] 在研究词义消歧时，通过同时使用词的局部上下文以及词在文档其他部分出现时的含义这两部分信息，有效减少了对人工标注数据的需求量；E. Riloff 和 R. Jones [RiloffJ99] 在对名词短语进行地理位置分类时，同时考虑了名词短语本身及其出现的上下文；M. Collins 和 Y. Singer [CollinsS99] 进行名实体识别时，也同时使用了名实体的拼写信息及名实体出现的上下文信息。

A. Blum 和 T. Mitchell 提出标准协同训练算法后，协同训练技术很快就在自然语言处理领域受到了重视。D. Pierce 和 C. Cardie [PierceC01] 将协同训练算法用于名词短语识别，他们把当前词及在文档中出现在该词前的 k 个词作为一个视图，把该词及出现在其后的另外 k 个词作为另外一个视图，然后在两个视图上利用协同训练算法进行训练。为了适应多类分类问题，他们还对标准协同训练算法进行了改进。他们的研究结果表明，在使用协同训练技术利用未标记示例后，识别错误率比仅使用有标记示例时下降了 36%。A Sarkar [Sarkar01] 将句法分析器分解为两个相关模型，其中一个负责基于上下文挑选出合适的分析树（parsing tree），另一个则负责计算分析树间的关系并且给出最优的分析结果。在学习过程中，两个模型通过利用未标记示例进行协同训练，每个模型都利用对方提供的信息来帮助自己排除部分句法分析中的不确定因素。其结果表明，通过协同训练学得的句法分析器在精度（precision）和召回率（recall）方面都有显著提高。M. Steedman 等人 [SteedmanOSCHH03] 也提出了一种基于协同训练的统计句法分析方法，与 A. Sarkar 的方法不同，他们使用了两个不同的但功能完整的统计句法分析器进行协同训练。在训练过程中，每个分析器根据自己对未分析句子的

[5] 虽然A. Blum和T. Mitchell [BlumT98] 的理论结果表明弱学习器就够用了，但他们在实验中仍然使用了强学习器。一般来说，在理论分析时为了便于讨论算法的能力通常使用弱学习器；而在实际使用时为了得到更好的性能通常使用强学习器。

分析结果利用某个函数进行打分，作为对该句子分析的置信度，然后把得分最高的若干个示例提交给对方使用。他们的研究结果也证实，使用协同训练技术可以显著提高句法分析器的性能。R. Hwa 等人 [HwaOSS03] 提出了一种基于协同训练的主动半监督句法分析方法，在学习过程中，一个学习器挑选并标记自己最确定的示例给另一个学习器，而另一个学习器则挑选自己最不确定的示例请用户标记后再提交给该学习器用于模型更新。他们的研究结果表明该方法可以减少大约一半的人工标记量。

协同训练技术的另一个重要应用领域是基于内容的图像检索（CBIR）。CBIR 要求检索系统能够根据用户提供的查询图像自动地从图像库中检索出相似图像。在检索过程中通常会利用相关反馈（relevance feedback）来提高性能。具体来说，系统将检索结果提供给用户后，如果用户不满意，就可以从中选择一些图像并标示出其是否是期望的图像，然后系统根据这些信息再重新进行检索。该过程可能会反复进行多轮，直到用户满意或丧失信心为止。值得注意的是，在 CBIR 过程中，即使将用户在相关反馈过程中提供的信息考虑进来，有标记图像的数目仍然是比较少的，因为很少有用户会愿意花大量的时间来提供反馈；但图像库中却通常存在大量的图像，这些图像都是未标记的，因为在查询之前无法事先判断它们是否与查询相关。显然，CBIR 任务是典型的有标记示例很少、未标记示例非常多的任务。因此，基于内容的图像检索是利用未标记示例的学习技术的很好的试验场，另一方面，通过引入这些学习技术可能有助于突破 CBIR 的技术瓶颈 [Zhou06]。

Zhou 等人 [ZhouCJ04][ZhouCD06] 将协同训练引入 CBIR，提出了基于协同训练的主动半监督相关反馈方法。他们在每一轮相关反馈后，利用现有的有标记示例训练两个基于距离度量的简单学习器，然后两个学习器分别对图像库中的图像进行预测从而产生两个排序，排在最前面的是置信度最高的相关图像，排在最后面的是置信度最高的不相关图像，而排在中间的则是置信度比较低的图像。基于这两个排序，两个学习器分别将自己最确定的相关图像和最确定的不相关图像传递给对方，然后两个学习器利用这些新的有标记图像进行更新。更新后的学习器再对图像库中的图像进行预测从而产生两个排序，通过结合这两个排序就得到一个总排序。基于总排序，系统把排在最前面的若干幅图像作为检索结果反馈给用户，而把排在中间的若干幅图像放入反馈池（feedback pool）中，供用户在进行下一轮相关反馈时进行标示。在 COREL 图像库上的实验表明，该技术通过在协同训练设置下结合半监督学习和主动学习，可以有效地利用图像库中的图像来提高检索性能。

6. 结束语

从上世纪 90 年代末标准协同训练算法被提出开始，很多研究者对协同训练技术进行了研究，不仅提出了很多学习方式不同、限制条件强弱各异的算法，对协同训练的理论分析和应用研究也取得了不少进展，使得协同训练成为半监督学习中最重要的风范之一。

但至少在目前，针对协同训练风范仍然存在很多值得进一步研究的问题：

由于协同训练是一种半监督学习技术，因此半监督学习领域存在的主要问题在协同训练风范中都存在。例如，在通过半监督学习利用未标记示例后，有时不仅不能提高泛化能力，反而会使得性能下降。一般认为，在模型假设不符合真实情况 [CohenCSCH04][CozmanC02] 或者未标记示例的分布与有标记示例的分布有较大差异 [TianYXS04] 时，进行半监督学习有可能导致性能下降。另一方面，随着训练不断进行，自动标记的示例中的噪音会不断积累，其负作用会越来越大。利用数据审计（data editing）技术来发现和处理这些噪音数据，也许是一条可能的途径，M. Li 和 Z.-H. Zhou [LiZ04] 对此进行了初步的尝试。总的来说，找到未标记示例导致性能下降的真正原因，有助于更好地发挥半监督学习技术的效用。

目前虽然有了很多协同训练算法，但这些算法都有自身的弱点。如何设计出更强有力的协同训练算法，一直是该领域的重要研究内容。现有对协同训练的理论分析虽然揭示了协同训练的一些内在机理，但是很多分析都建立在一些较强的假设条件上。现在已经知道，在这些较强的假设条件不满足的情况下，协同训练技术仍然能够取得较好的效果。因此，在更一般、更接近真实情况的条件下对协同训练进行理论分析，是一个需要努力的方向。将协同训练技术投入到更多的应用中去，基于协同训练技术研制出实用系统，也是该领域重要的研究内容。

值得注意的是，A. Blum 和 T. Mitchell [BlumM98] 利用数据不同视图的思想受到了机器学习界的很大重视，为“多视图学习”（multi-view learning）这一新的研究领域奠定了基础，这也使得协同训练风范的影响超越了半监督学习领域。例如，I. Muslea 等人 [MusleaMK00][MusleaMK02] 将协同训练的思想引入主动学习，他们在两个视图上分别建立分类器，然后选择两个分类器预测差异最大的示例进行查询。对多视图学习的算法、理论、应用进行研究，是今后的重要研究内容。

到目前为止，对协同训练的研究主要是在机器学习算法这一层面开展的[6]，但笔者认为，协同训练除了在机器学习算法方面具有重要性，深入研究协同训练机制对理解和模仿人类的学习行为也有重要的意义。例如，人类在对外界事物进行学习时，不同感官对同一事物的感知能力通常是不同的，但在学习之后，不同感官对此类事物的感知能力往往都会得到提高。如果把同一事物在不同感官上的反映看作同一示例在不同视图下的特征向量，则学习之后感知能力的提高有可能是因为不同视图下的学习器互相提供了信息。再如，人类在集体环境下进行学习时，如果把每个人看作一个学习器，则即使对同样的事物，不同的人学得的结果也可能是显著不同的，而人们可以通过互相学习来提高自己的能力，这恰恰与协同训练非常相似。笔者认为，通过借鉴人类学习行为，有可能产生更强有力的协同训练技术，而机器学习中对协同训练技术的研究，也许能够为认知科学中对学习的研究提供启示和实验手段。

[6] 虽然本章谈到了学习算法、理论分析、实际应用等方面，但其实机器学习的核心研究内容就是“算法”。这里的“算法”是广义的，不仅包含了学习算法本身，还包含了对算法性质的理论分析或对算法设计的理论讨论，以及对算法的应用等。实际上，计算机科学大多数领域的核心研究内容都是“算法”。

参 考 文 献

[AbeM98] N. Abe, H. Mamitsuka. Query learning strategies using boosting and bagging. In: *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, Madison, WI, 1998, 1-9.

[Abney02] S. Abney. Bootstrapping. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, 2002, 360-367.

[AngluinL88] D. Angluin, P. Laird. Learning from noisy examples. *Machine Learning*, 1988, 2(4): 343-370.

[BalcanBY05] M.-F. Balcan, A. Blum, K. Yang. Co-training and expansion: Towards bridging theory and practice. In: L. K. Saul, Y. Weiss, L. Bottou, eds. *Advances in Neural Information Processing Systems 17*, Cambridge, MA: MIT Press, 2005, 89-96.

[BelkinN04] M. Belkin, P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 2004, 56(1-3): 209-239.

[BelkinNS05] M. Belkin, P. Niyogi, V. Sindwani. On manifold regularization. In: *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS'05)*, Savannah Hotel, Barbados, 2005, 17-24.

[BlumC01] A. Blum, S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In: *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, San Francisco, CA, 2001, 19-26.

[BlumM98] A. Blum, T. Mitchell. Combining labeled and unlabeled data with co-training. In: *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*, Wisconsin, MI, 1998, 92-100.

[BrefeldGSW06] U. Brefeld, T. Gärtner, T. Scheffer, S. Wrobel. Efficient co-regularised least squares regression. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, Pittsburgh, PA, 2006, 137-144.

[ChapelleSZ06] O. Chapelle, B. Schölkopf, A. Zien, eds. *Semi-Supervised Learning*, Cambridge, MA: MIT Press, 2006.

[CohenCSCH04] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, T. S. Huang. Semisupervised learning of classifiers: Theory, algorithm, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(12): 1553-1567.

[CollinsS99] M. Collins, Y. Singer. Unsupervised models for named entity classifications. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99)*, College Park, MD, 1999, 100-110.

[CozmanC02] F. G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In: *Proceedings of the 15th International Conference of the Florida Artificial Intelligence Research Society (FLAIRS'02)*, Pensacola, FL, 2002, 327-331.

[DasguptaLM02] S. Dasgupta, M. Littman, D. McAllester. PAC generalization bounds for co-training. In: T. G. Dietterich, S. Becker, Z. Ghahramani, eds. *Advances in Neural Information Processing Systems 14*, Cambridge, MA: MIT Press, 2002, 375-382.

[Dietterich00] T. G. Dietterich. Ensemble methods in machine learning. In: *Proceedings of the 1st International Workshop on Multiple Classifier Systems (MCS'00)*, Cagliari, Italy, *LNCS 1867*, 2000, 1-15.

[GoldmanZ00] S. Goldman, Y. Zhou. Enhancing supervised learning with unlabeled data. In: *Proceedings of the*

17th International Conference on Machine Learning (ICML'00), San Francisco, CA, 2000, 327-334.

[GrandvaletB05] 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, 2005, 529-536.

[HwaOSS03] R. Hwa, M. Osborne, A. Sarkar, M. Steedman. Corrected co-training for statistical parsers. In: *Working Notes of the ICML'03 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.

[Joachims99] T. Joachims. Transductive inference for text classification using support vector machines. In: *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, Bled, Slovenia, 1999, 200-209.

[LawrenceJ05] 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, 2005, 753-760.

[LewisG94] D. Lewis, W. Gale. A sequential algorithm for training text classifiers. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, Dublin, Ireland, 1994, 3-12.

[LiZ05] M. Li, Z.-H. Zhou. SETRED: Self-training with editing. In: *Proceedings of 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05)*, Hanoi, Vietnam, *LNAI 3518*, 2005, 611-621.

[LiZ07] M. Li, Z.-H. Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man and Cybernetics – Part A*, in press

[Lippmann89] R. P. Lippmann. Pattern classification using neural networks. *IEEE Communications*, 1989, 27(11): 47-64.

[MavroeidisCPCV06] D. Mavroeidis, K. Chaidos, S. Pirillos, D. Christopoulos, M. Vazirgiannis. Using tri-training and support vector machines for addressing the ECML/PKDD 2006 discovery challenge. In: *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, Berlin, Germany, 2006, 39-47.

[MillerU97] D. J. Miller, 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, 1997, 571-577.

[MusleaMK00] I. Muslea, S. Minton, C. A. Knoblock. Selective sampling with redundant views. In: *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'00)*, Austin, TX, 2000, 621-626.

[MusleaMK02] I. Muslea, S. Minton, C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In: *Proceedings of the 19th International Conference on Machine Learning (ICML'02)*, Sydney, Australia, 2002, 435-442.

[NigamG03] K. Nigam, R. Ghani. Analyzing the effectiveness and applicability of co-training. In: *Proceedings of the 9th ACM International Conference on Information and Knowledge Management (CIKM'00)*, McLean, VA, 2000, 86-93.

[NigamMTM00] K. Nigam, A. K. McCallum, S. Thrun, T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000, 39(2-3): 103-134.

[PierceC01] D. Pierce, C. Cardie. Limitations of co-training for natural language learning from large data sets. In: *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP'01)*, Pittsburgh, PA, 2001, 1-9.

[RiloffJ99] E. Riloff, R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In: *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI'99)*, Orlando, FL, 1999, 474-479.

[Sarkar01] A. Sarkar. Applying co-training methods to statistical parsing. In: *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'01)*, Pittsburgh, PA, 2001, 95-102.

[SeungOS92] H. Seung, M. Opper, H. Sompolinsky. Query by committee. In: *Proceedings of the 5th ACM Workshop on Computational Learning Theory (COLT'92)*, Pittsburgh, PA, 1992, 287-294.

[ShahshahaniL94] B. Shahshahani, D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 1994, 32(5): 1087-1095.

[SteedmanOSCHH03] M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, J. Crim. Bootstrapping statistical parsers from small data sets. In: *Proceedings of the 10th Conference on the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, 2003, 331-338.

[TianYXS04] Q. Tian, J. Yu, Q. Xue, N. Sebe. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In: *Proceedings of the IEEE International Conference on Multimedia Expo (ICME'04)*, Taibei, 2004, 1019-1022.

[Vapnik98] V. N. Vapnik. *Statistical Learning Theory*, New York: Wiley, 1998.

[WangZ07] W. Wang, Z.-H. Zhou. Analyzing co-training style algorithms. In: *Proceedings of the 18th European Conference on Machine Learning (ECML'07)*, Warsaw, Poland, 2007.

[Yarowsky95] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, Cambridge, MA, 1995, 189-196.

[ZhangO00] T. Zhang, F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In: *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*, San Francisco, CA, 2000, 1191-1198.

[ZhouBLWS04] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf. Learning with local and global consistency. In: S. Thrun, L. Saul, B. Schölkopf, eds. *Advances in Neural Information Processing Systems 16*, Cambridge, MA: MIT Press, 2004, 321-328.

[ZhouG04] Y. Zhou, S. Goldman. Democratic co-learning. In: *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, Boca Raton, FL, 2004. 594–602.

[Zhou06] Z.-H. Zhou. Learning with unlabeled data and its application to image retrieval. In: *Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06)*, Guilin, China, *LNAI 4099*, 2006, 5-10.

[ZhouCD06] Z.-H. Zhou, K.-J. Chen, H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 2006, 24(2): 219-244.

[ZhouCJ04] Z.-H. Zhou, K.-J. Chen, Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In: *Proceedings of the 15th European Conference on Machine Learning (ECML'04)*, Pisa, Italy, *LNAI 3201*, 2004, 525-536.

[ZhouL05a] Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(11): 1529–1541.

[ZhouL05b] Z.-H. Zhou, M. Li. Semi-supervised learning with co-training. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, Edinburgh, Scotland, 2005, 908-913.

[ZhouL07] Z.-H. Zhou, M. Li. Semi-supervised learning with co-training style algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(11).

[Zhu06] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Apr. 2006.

[ZhuGL03] X. Zhu, Z. Ghahramani, J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In: *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, Washington, DC, 2003, 912-919.