主要周志华老师的综述文章：A brief introduction to weakly supervised learning

强烈建议大家读一下这篇文章，读这一篇文章收获很多，这里总结一下

论文摘要

监督学习通过讯息大量的训练样本来构建预测模型，其中每个训练样本都有一个标签来知名其真值(ground-truth)，尽管现在监督学习技术已经获得了很大的成功，但是值得注意的是，在许多任务中，由于数据标注过程的成本过高，很难获得像全部真值这样的强标注信息。因此采用弱监督的机器学习技术是可取的。本文综述一些弱监督学习的一些研究进展，主要集中关注集中典型的弱监督类型：不完全监督：只有部分训练数据含有标签，不确切监督：训练数据仅仅具有粗粒度的标注信息，不精确监督：标注信息不总是真值。

为何要应用弱监督学习的技术

机器学习，深度学习在很多任务中获得很大的成功，尤其是在监督学习中，深度学习，机器学习取得了突破性的进展。
例如分类和回归的监督任务中，预测模型需要从大量有标注训练样本中学习，训练样本包含两部分：第一部分是对象的特征向量，第二部分是真值标签。在分类任务中，label表示训练样本的类别，在回归任务中，标签是一个与样本相对应的实数值。这些模型依赖强监督信息，例如深度学习模型依赖大量的标注数据，在实际生产中，数据的标注成本很高，获取大量的有标注样本难度大。
因此利用弱监督学习，利用大量无标注或者粗糙标注的样本来进行模型的学习，这样能够有效的利用数据，提升模型的性能

弱监督学习的三种类型

不完全监督(incomplete supervision)： 只有一部分的训练数据含有标注信息，其他数据则没有标签

不确切监督(inexact supervision)： 训练数据仅仅具有粗粒度的标注信息，例如在图像领域中的实例分割任务，仅仅含有bbox标注

不精确监督(inaccurate supervision)： 训练数据的标注信息有误并不完全正确

不完全监督(incomplete supervision)

不完全监督主要应对只有训练集的一个很小的子集含有标签，而大量的样本为无标注的样本。如果仅仅采用有标注的信息训练模型，往往不能得到一个泛化能力强，
非常鲁棒的模型。

应对此不完全监督任务的两种解决方案是：主动学习和半监督学习

主动学习(active learning)

主动学习假设存在一个oracle（我也不知道怎么翻译，神谕？？？），主动学习假设可以从oracle查询选定的未标注实例的真值标签。

简单起见，假设模型的损失依赖于询问的数目，主动学习的目标就是最小化询问的数目，以此来最小化训练模型的损失

对于给定的一部分数目较少的标注样本和大量的无标注样本，主动学习试图寻找最有价值的无标注样本及逆行询问(query),有两种广泛使用的选择策略：信息性与代表性。信息性衡量一个无标注样本降低统计模型不确定性的程度；代表性衡量无标注样本对于表达输入范式的有用程度。

不确定性采样与基于委员会查询的方法是信息性的代表方法，不确定性采样是训练一个学习器，然后选择在这个学习器上有最低置信度的样本，进行查询。后者就是训练多个学习器，然后选择这些学习器最不赞同的无标注样本就行查询。而基于表达的方法主要是利用无标注样本的聚类结果，通常采用聚类的方式实现。

基于信息性的方法的主要缺点就是这种方法严格依赖于已标注的样本来选择query样本，当标注样本非常少的时候，这种方法的效果比较差。而基于表达性的方法则主要依赖于无标注样本的聚类结果，特别是当有标注样本非常少的时候，依赖性更加严重。因此近些年的一些方法试图平衡信息性与表达性。

近些年有很多关于主动学习的理论研究，例如：对于可实现的情况(realizable case)(在假设类中存在数据可完全分离假设)，利用主动学习的方法，样本的复杂度可以获得指数提升。对于不可实现的情况(unrealizable case)(在假设类中不存在任何假设可以将数据完全分离)，已经被证明，没有关于噪声模型的假设，主动学习的上界与下界相匹配，也就是主动学习没有什么用。已经证明，假设Tsybakov噪声模型，对于有界噪声，主动学习可获得指数级的提升。如果能够开发一些特殊的数据特征，例如多视图，主动学习对于无界的噪声依然可用获得指数级的提升。总之，即使对于很困难的情况，通过精妙的设计主动学习依然能获得提升

半监督学习(semi-supervised learning)

半监督学习无需人工的参与，自动开发无标注的数据，来提升模型的性能。

存在一种特殊的半监督学习，称为直推式学习（transductive learning）；直推式学习和（纯）半监督学习的主要区别在于，它们对测试数据，即训练过的模型需要进行预测的数据，假设有所不同。直推式学习持有「封闭世界」假设，即，测试数据是事先给出的、目标是优化测试数据的性能；换言之，未标注数据正是测试数据。纯半监督式学习则持有「开放世界」假设，即，测试数据是未知的，未标注数据不一定是测试数据

在半监督学习中有两个主要的假设就是：聚类假设和流形假设，二者都是关于数据分布的假设。前者假设数据具有连续的聚类结构，因此在相同聚类簇中的结果有相同的类别。后者假设数据依赖于流形，因此相近的实例具有相同的类别。这两种假设都依赖于相似的数据点有相似的输出，因此无标注的数据对相似点的发现有帮助。

半监督学习有四种主要的方法：生成式方法，基于图的方法，低密度分离方法，基于不一致的方法

生成式方法假设有标注与无标注的样本由同一个连续模型生成。因此无标注样本的label作为生成式模型的丢失值，采用EM算法评估。这些模型的不同点在于使用不同的生成式模型来拟合数据，为了能够得到更好的效果通常需要domain的知识来获得充足的生成式模型，也有很多人尝试去混合生成式模型与判别式模型。

基于图的方法构建一张图，节点表示训练实例，边表示节点之间的关系（相似度或者距离），然后利用某种准则在图中传递label信息。这种方法空间以及时间复杂度高，限制扩展性。

低密度分离的方法强迫使边界线穿过密度较低的区域，代表性的方法就是半监督SVM。

基于不一致的方法式使用不同的学习器联合开发无标注数据，在这期间，各个学习器之间的不一致性是训练过程继续的重要因素。以两个学习器为例，在每次迭代中，每个学习器都选择自身置信度最大的label作为这个实例的伪标签去训练前一个分类器。通过ensemble的方法往往可以获得更好的效果。这种基于不一致方法提供了一种将主动学习与半监督学习相互联合的方法，除了联合训练之外，对于几个分类器置信度均较低或者虽然大家的预测置信度都较高，但是label不一样的实例可以选择作为query。

对于无标注样本的使用有时候会让模型的表现力变得更差，由于无标注样本的使用通常会导致多偶遇一个模型选择，因此不充分的选择可能导致更差的表现，为了使半监督学习更加safer，通常联合多个模型来优化。参看文章：Towards making unlabeled data never hurt

基于不一致的方法参考：Theoretical foundation of co-training and disagreement-based algorithms

不确切监督(inexact supervision)

不确切监督关注给定了监督信息，但是监督信息不够准确，仅仅右粗粒度的标签可用，例如在instence segmentation中仅仅含有bbox标注，而没有像素级别的标注。

例如，在药物活性预测的问题中，其目标是建立一个模型学习已知分子的知识，来预测一个新的分子是否适合制造一种特定药物。一个分子可以有很多的低能量形状，而这些分子是否能用于制药取决于这些分子是否具有某些特殊的形状。然而即使对于已知的分子，人类专家也仅知道该分子是否适合制药，而不知道其中决定性的分子形状是什么。

形式化表达为，该任务是从训练数据集中学习，其中每个分子被称为一个包。每个分子的每个形状，是一个示例。

如果一个对应的分子是一个 positive 包，如果存在x的某一个形状是正的，同时是未知的。其目标是为未见过的包预测标签。该方法被称为多示例学习。

已经有许多有效的算法被开发出来并应用于多示例学习。实际上，几乎所有的有监督学习算法都有对等的多示例算法。大多数算法试图调整单示例监督学习算法，使其适配多示例表示，主要是将其关注点从对示例的识别转移到对包的识别；一些其他算法试图通过表示变换，调整多示例表示使其适配单示例算法。还有一种类型，将算法分为三类：一个整合了示例级响应的示例空间范式，一个把 包 视作一个整体的 包 空间范式，以及一个在嵌入特征空间中进行学习的嵌入空间范式中。请注意，这些示例通常被视为 i.i.d. 样本，然而，一些文献表明，多示例学习中的示例不应该被认为是独立的，尽管这些包可以被视为 i.i.d. 样本，并且已经有一些有效的算法是基于此见解进行开发的。

多示例学习已成功应用于各种任务，如图像分类/检索/注释，文本分类，垃圾邮件检测，医学诊断，面部/对象检测，对象类别发现，对象跟踪等。在这些任务中，将真实对象（例如一幅图像或一个文本文档）视为一个包是很自然的。然而，不同于药物活性预测这类包中包含天然示例（分子的各种形状）的例子，需要为每个包生成示例。包生成器制定如何生成示例来构成包。通常情况下，可以从图像中提取许多小的图像块作为其示例，而章节/段落甚至句子可以用作文本文档的示例。尽管包生成器对学习效果有重要影响，但最近才出现关于图像包生成器的全面研究，研究揭示了一些简单的密集取样包生成器比一些复杂的生成器性能更好。

多示例学习的初始目标是为未见过的包预测标签；然而，已有研究尝试识别那些之所以让正包变正的关键示例（key instance）[31,60]。这在诸如没有细粒度标记训练数据的感兴趣区域定位的任务中特别有用。值得注意的是，标准的多示例学习假定每一个正包必须包含一个关键示例，而还有其它研究假定不存在关键示例，每一个示例都对包标签有贡献；甚至假定存在多个概念，而仅当一个包包含满足所有概念的示例时，该包才是正的。

早期的理论结果表明多示例学习对于包中每个示例都由不同的规则分类的异质（heterogeneous）案例来说，是很难的，对于以相同的规则分类所有示例的同质性（homogeneous）案例就是可学习的。幸运的是，几乎所有的实际多示例任务都属于同质性案例。这些分析假定 bag 中的示例是独立的。而不假定示例的独立性的分析更具挑战性，这类研究也出现得较晚，其揭示了在同质性类中，至少存在某些可以用包间的任意分布来学习的案例。尽管如此，与其在算法和应用上的繁荣发展相反，多示例学习的理论研究成果非常少，因为分析的难度太大。

不精确监督(inaccurate supervision)

不精确监督关注的问题是对于给定的监督信息，有一些是错误的，也就是监督信息不总是ground-truth的情况。

一个相对典型的场景就是在有标签噪声的情况下进行学习。目前很多理论研究相关问题，其中大多数均假设存在随机的分类噪声，即标签受随机噪声的影响。基本的思想就是识别潜在的误分类样本，然后尝试对其进行修正。

数据编辑（data-editing）方法构建了相对邻域图（relative neighborhood graph），其中每一个节点对应一个训练样本，而连接两个不同标签的节点的边被称为切边（cut edge）。然后，测量 一个切边的权重统计量，直觉上，如果一个示例连接了太多的切边，则该示例是可疑的。可疑的示例要么被删除，要么被重新标记。值得注意的是，这种方法通常依赖于咨询邻域信息；由于当数据很稀疏时，邻域识别将变得更不可靠，因此，在高维特征空间中该方法的可靠性将变弱。

近期出现的有趣的不准确监督的场景是众包模式(crowdsourcing)，这是一种流行的将工作外包给个人的范式。对于机器学习来说，用众包模式为训练数据收集标签是一种经济的方式。具体来说，未标记的数据被外包给大量的工人来标记。在著名的众包系统 Amazon Mechanical Turk 上，用户可以提交一项任务，例如将图片标注为「树」或「非树」，然后职工完成工作以获取少量报酬。通常这些工人来自世界各地，每个人都可以执行多个任务。这些职工通常互相独立，报酬不高，并通过自己的判断标记数据。这些职工的标记质量参差不齐，但标记质量信息对于用户来说是不可见的，因为工人的身份是保密的。在这些职工中可能存在「垃圾制造者」，几乎用随机的标签来标记数据（例如，用机器替代人类赚取报酬），或「反抗者」，故意给出错误的标签。此外，某些任务可能对一些人来说太难而无法完成。使用众包返回的不准确监督信息来保证学习性能是非常困难的。

很多研究尝试用众包标签推断真值标签。多数人投票策略得到了集成方法的理论支持，在实践中得到了广泛使用并有很好的表现，因此通常作为基线标准。如果预期可以对工人质量和任务难度建模，那么通过为不同的工人在不同的任务上设置权重，则可以获得更好的效果。为此，一些方法尝试构建概率模型然后使用 EM 算法进行评估 [77,78]。人们也使用了极小极大熵方法。概率模型可以用于移除垃圾制造者。近期人们给出了移除低质量工人的一般理论条件。

在机器学习中，众包通常用于收集标签，在实践中，模型的最终性能，而不是这些标签的质量，才是更重要的。目前已有很多关于从低能老师和众包标签学习的研究，这和用带噪声标签学习是很接近的。但其中的区别在于，对于众包设定而言，人们可以方便地、重复地对某个示例提取众包标签。因此，在众包数据学习中，考虑经济性和最小化众包标签的充分数量是很重要的，即有效众包学习的最小代价。很多研究专注于任务分配和预算分配，尝试在准确率和标注开销之间取得平衡。为此，非适应性的任务分配机制（离线分配任务）和适应性机制（在线分配任务）都得到了在理论支持下的研究。需要注意的是，多数研究采用了 Dawid–Skene 模型，其假设不同任务的潜在成本是相同的，而没有探索更复杂的成本设置。

设计一个有效的众包协议也是很重要的。一些文献中提供了「不确定」选项，从而使工人在不确定的时候不被迫使给出确定的标签。该选项可以帮助标记的可靠性获得有理论支持的提升。一些文献中提出了一种「double or nothing」的激励兼容机制，以确保工人能提供基于其自己的信心的标注，诚实地工作。在假定每位工人都希望最大化他们的报酬的前提下，该协议被证实可以避免垃圾制造者的出现。

参看链接：https://academic.oup.com/nsr/article/5/1/44/4093912