IMPROVABILITY THROUGH SEMI-SUPERVISED LEARNING: A SURVEY OF THEORETICAL RESULTS

摘要

半监督学习是一种有标记和无标记数据的学习环境。在这项调查中，我们探讨了在分类和回归任务中使用未标记数据时不同类型的理论结果。大多数使用未标记数据的方法依赖于关于数据分布的某些假设。当这些假设在现实中得不到满足时，包括未标记的数据实际上可能会降低性能。因此，研究这些方法，了解其背后的理论就显得尤为重要。在这篇综述中，我们收集了关于使用半监督学习可能获得的收益的结果，以及关于这些方法的局限性的结果。更准确地说，这篇综述收集了以下问题的答案：在改进监督方法方面，半监督学习的局限性是什么？不同方法的假设是什么？如果假设是真的，我们能得到什么？最后，我们还讨论了半监督学习的最大瓶颈，即他们所做的假设。

**1介绍和范围**

在各种应用中，收集未标记的数据比收集标记的数据更容易、更快和/或更便宜。半监督学习（SSL）的目标是结合未标记和标记的数据来设计分类或回归规则，其性能优于仅基于标记数据的方案。然而，SSL确实带来了固有的风险。众所周知，包含未标记的数据会降低性能[Ben David等人，2008，Cozman和Cohen，2006]。从理论的角度研究和理解SSL可以使我们准确地描述我们需要的假设和我们可以期望的改进，以及上述方法的局限性。有了它，可以制定使用SSL的建议，以尽可能避免性能下降。在本文中，我们收集和介绍了SSL的理论成果，详细研究了相关文献，介绍了它们的主要结果，并指出了与其他工作的联系。

这篇评论针对两组读者。我们针对的第一组是感兴趣的实践者和研究人员，他们致力于实验性SSL。虽然他们可能对我们提供的所有细节都不感兴趣，但我们相信，我们每一节中的介绍都对SSL中的理论结果类型及其引发的主要见解有了很好的高层次理解。第二个目标受众是所有从事SSL理论方面工作的人，我们希望，特别是这一领域的研究人员，能够在我们的综述中找到他们自己工作的灵感和联系。

我们主要用PAC学习框架的语言来描述半监督学习者的表现。我们解释结果，画出它们之间的联系，并指出必须假设它们是有效的。除了某些特定SSL的理论保证之外，我们还介绍了SSL限制的结果。

**1.1大纲**

在下一节中，我们将介绍正式的学习框架，该框架也适用于我们提出的大多数工作。在第3节中，我们给出了SSL限制的结果，这些限制通常是由于对模型或数据生成过程的特定假设而产生的。与SSL的改进可证明是有限的设置相反，我们在同一节中介绍了SSL的改进是无限的三种设置。在这里，我们的意思是SSL可以学习PAC-问题，但是没有监督学习者（SL）可以做到。在第4节中，我们研究了在不进一步假设数据分布的情况下，试图利用未标记数据的方法。在第5节中，我们介绍了半监督学习者，他们对数据分布做出了较弱的假设。这些假设是弱的，因为最终的学习者不能得到比的标准学习速度快的学习速度，其中n是标记样本的数量。这些改进是由一个常数给出的。在第6节中，我们介绍了使用强假设的学习者，在该假设下，学习者可以指数级快速地收敛到给定类中的最佳分类器，即学习率按的顺序排列。在第7节中，我们介绍了在传导设置中的结果，在该设置中，学习者只对未标记数据的标签感兴趣。在同一部分中，我们还提出了一系列的研究，试图构建一个半监督学习者，其学习效果决不会比监督学习者差。在第8节中，我们将讨论总体结果，并指出当前该领域面临的挑战。在第8.4节中，我们进一步详细解释了在SSL中使用假设的正式含义以及由此产生的问题。

**2. 初步设置**

除非另有说明，否则所有结果均在标准统计学习框架中列出。这意味着我们得到了一个特征空间和一个标签空间以及一个未知的XY上的分布P。重载这个符号，我们为X和Y上的边际分布写P（X）和P（Y），类似于条件分布。我们观察到标记的n样本和未标记的M样本，其中，每一个对于和每个对于是根据P相同和独立分布的，然后选择一个假设类H，其中每个是映射、 以及损失函数。除非另有规定，否则我们假设分类为，损失为0-1损失，。在回归任务中，我们假设。基于n个标记的和m个未标记的样本，我们试图找到一个，使得风险很小。最后，每当我们有任何数量A取决于分布P，我们写一个的经验估计版本的A.例如，给定标记的样本Sn，我们写的，用于测量在Sn上测量的的经验风险。如果上下文不清楚，我们将澄清我们测量的样本。在第24页的表1中，我们给出了我们使用的符号的完整列表。

**3 半监督学习的可能性与不可能性**

在SSL中，我们希望使用有关X上的分布的信息来改进学习，但不一定清楚这些信息是否有用。一些作者将这种想法形式化，然后提出无标签数据可以帮助或不能帮助的情况。本节遵循相同的划分。在第3.1小节中，我们提供了不同的设置，作者可以证明未标记的数据没有帮助，而在第3.2小节中，我们提供了三个特定的设置，未标记的数据可以提供无限的改进。我们所说的无限制是指没有一个受监督的学习者可以在这些环境中学习，而一个半受监督的学习者可以。

负的结果往往表明后验概率和边际分布P（X）之间是独立的。然而，这并不直接意味着未标记的数据是无用的，因为我们通常不仅对感兴趣，而且对分类器h，的完全风险感兴趣，后者依赖于P（X）[Peters等人，2017，5.1.2]。例如，在第4.1节和第4.2节中，我们介绍了即使和P（X）是独立的，也显示风险改善的工作。

**3.1不可能情况**

**3.1.1由于数据生成过程而不可能**

Seeger[2000]研究了一个简单的数据生成模型，并研究了当先验信息以贝叶斯方式包含时，关于数据分布的先验信息如何改变我们对模型的后验信念。为了使用贝叶斯方法，假设数据以以下方式生成。我们假设分布P来自一个带参数和的模型类。首先对和进行独立采样，然后通过采集带有相应标签的样本生成数据，如图1所示。在这个设置中的目标是从有限标记样本推断。使用贝叶斯方法，可以很容易地表明独立于任何有限未标记样本和本身。换言之，未标记信息不会改变给定标记数据Sn的后验信念。一个可能的解决方案是假设和之间存在依赖关系，因此在图1中的和之间画一个额外的箭头。

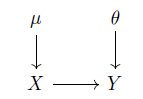


图1:Seeger分析中使用的数据生成过程。

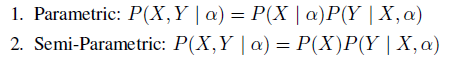
3.1.2由于模型假设而不可能

Hansen[2009]研究了未标记数据何时应该改变我们对模型的后验信念。与Seeger[2000]相比，没有数据生成假设，而是对我们使用的模型进行假设。他研究了由给定模型和真实期望标签输出之间的期望平方损失导出的解。将模型的联合分布分解为:



他得出结论，如果标签和边缘分布之间的共享参数为空，则可以丢弃未标记的数据。

Zhang和Oles[2000]早期的工作区分了相同类型的模型，但不可能性是关于半监督分类器的渐近效率。本文还考虑了两类联合概率模型：



我们可以证明，考虑标记了未标记数据的MLE估计器的Fisher信息可以分解为

。因此，只要有未标记的数据，半监督学习者的Fisher信息就比监督学习者的大，这表明标记了Fisher信息。因此，SSL渐近地更有效，尽管不一定严格。在参数情形下，我们观察到与半监督和监督估计具有相同的渐近性质。在第4.1节中，我们将提出一种方法，即使使用判别模型时，也允许SSL的渐近效率。

**3.1.3由于因果关系而不可能**

Schólkopf等人。[2012，第2节和第3节]分析功能因果模型，如图2所示。他们分析不同的学习场景，假设标签是原因C，特征是效果E，反之亦然。这个模型引入了因果不对称，因为它导致了P（C）和是独立的，而P（E）和不是独立的。假设X是标签Y的原因，我们发现预测与新获得的关于P（X）的信息无关。但是，如果我们假设标签Y是由X引起的，情况就会改变。其中一个问题是，我们不一定知道这个特性是因果关系。但是，例如在医疗环境中，这可能并不太困难，因为我们可以将因果特征识别为真正导致疾病的特征，而效果特征是疾病的症状。Kügelgen等人。[2019]利用这一知识，推导出只考虑效应特征的未标记数据的SSL方法。

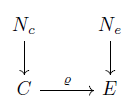


图2:Schólkopf等人使用的简单功能因果模型。[2012年]。效应E是由给定一个确定性映射的C引起的。E和C都受噪声变量NE和NC的影响。

**3.1.4不可能总是胜过有监督的学习者**

Krijthe和Loog[2018]受到Loog生成线性判别模型的成功minimax方法（见第7.2.2节）的启发，研究了一种类似的方法，以找到判别模型的半监督解，这种方法永远不会比监督模型差。它们使用一个设置，其中判别模型是用单调递减损失函数导出的。这个设置也是反导的，所以如果你只对我们的模型在未标记数据Um上的性能感兴趣，请参见第7节。它们本质上表明，在一些温和的条件下，总是有一个未见的数据Um的标签，这样半监督学习者在Um上的表现会比监督解决方案差。从这个意义上说，不可能保证半监督解总是优于监督解。

**3.1.5如果我们只知道流形**

Lafferty和Wasserman[2007，第3节]表明，在没有附加假设的情况下，仅了解流形不足以胜过纯粹的监督学习者。他们在回归环境中工作，并扩展了Bickel和Li[2007]的工作，以表明有一个有监督的学习者能够适应流形的维度，从而能够达到与直接在低维流形上工作的学习者相当的最小最大速率。我们注意到，Lafferty和Wasserman[2007]也表明，如果我们也假设半监督平滑假设，我们基本上可以实现更快的速率。在这一点上，我们不涉及更多细节，但在第6.4节中提供了一个质量上非常相似的分析。

**3.1.6如果我们没有额外的假设则不可能**

Ben David等人。[2008]开始了一系列调查，推测在某种意义上，没有任何假设，SSL通常是不可能的。我们特别假设给定的域分布不限制可能的标记函数，类似于图1中的数据生成过程。他们假设，半监督学习者在没有任何额外假设的情况下，不会有比SL更好的样本复杂度界限（见定义1和2）。这与前面的章节不同，因为对模型或数据生成过程没有进一步的限制。

在下面的两个部分中，我们要说明这些猜想的确切想法，为什么它们一般不成立，在哪些情况下它们是正确的。

我们从本·大卫等人的贡献开始。[2008年]。他们假设，任何半监督学习者的最坏情况样本复杂度最多比监督学习者提高一个常数，这个常数只取决于假设类别。第一个猜想指出，对于可实现的情况。

猜想1（猜想4）。对于任何假设类H，存在一个常数c（H），使得对于X上的任何域分布D



对于足够小的和，其中Dh是上具有边际分布D和条件分布的分布。

第二个猜想对于不可知的情况是一样的，所以我们用Dh代替任意分布的P。

猜想2（猜想5）。对于任何假设类H，存在一个常数c（H），使得对于任何域分布D



对于足够小的和，其中ext（D）是上所有分布P的集合，使得边际分布满足P（X）=D。

换言之，本文猜想，如果给定一个固定的域分布，人们总是可以在它上面找到一个标记函数，因此对于这个标记函数，SL和SSL之间的样本复杂度差距只能是一个常数。

本文证明了在可实现的情况下，实线上的光滑分布和阈值函数，以及在不可知的情况下，阈值函数和区间并的这些猜想。样本复杂度比较是通过构造一个最坏情况分析，在目标假设表现良好的情况下，我们可能仍然会得到非恒定的改进。我们将在第6节探讨这些案例。另一方面，人们也可以问这样一个问题：一个持续的改进本身已经有多好。我们将在讨论中详细阐述这一点。

当假设类具有有限的VC维时，在可实现的情况下，猜想1和2基本上是正确的。Darnstädt等人。【2013】证明了可实现的猜想1是正确的，但有一点小小的改变：被监督的学习者被允许有两倍的不准确，对于有限维的VC情况，我们得到了一个额外的。Mey等人。[2019]在某种程度上，更进一步，证明流形正则化方案遵守猜想所述的限制，即使在这种情况下，域分布携带关于标记函数的信息。Darnstädt等人。[2013]证明猜想1的以下版本。

定理1（定理1）。设H是一个假设类，它包含常数0和常数1函数。然后对于每个域分布D和每个，

1.如果H是有限的，那么



2。如果H有有限的VC维数，那么



首先要注意的是，这个语句适用于所有Dh，特别是如果我们取所有的上确界，如猜想1所示。Golovnev等人。[2019]表明，如果假设类H是由 上的投影给出的，则存在一组域分布，使得任何有监督算法需要与半有监督算法一样多的样本，半有监督算法知道整个域分布。特别是不等式（4）紧跟对数因子。这实际上表明，常数的改进是任意好的，因为我们可以通过增加维数Golovnev等来增加VC维数。[2019年，提案4]。定理1的不可知论版本是一个开放问题。

然而，对于一个具有无穷维VC的假设类，这个猜想也不再适用于稍加修改的公式。这基本上是这样的，因为我们可以从一个具有无限VC维的类开始，因此不能由有监督的学习者学习。然而，一个半监督的学习者可以限制这个类，使其具有有限的VC维。在下一节中，我们将收集三种不同的设置，其中半监督学习者可以学习PAC，而监督学习者不能。

3.1.7如果我们不限制可能的标签功能

Golovnev等人。[2019]表明，如果域X是有限的，并且我们允许其上的所有确定性标记函数，那么即使在一致的有监督学习者之上有一个常数，半监督学习者也无法在可实现的PAC学习框架中改进。一致性意味着学习者获得0个训练错误。然而，被监督的学习者被允许两倍的不准确和两倍的不确定。

定理2（定理8）。设X是一个有限域，设是X上所有确定的二元标记函数的集合。设A是任意一致的有监督学习者，P分布在X和上。那么



虽然更一般的定理1指出，根据假设类别，半监督的常数仍然可以更好，但我们发现，在先前的设置中，半监督甚至失去了这一优势。

对于不可知论的情况也可以找到类似的结果。Gépfert等人的定理2。[2019]基本上指出，假设2（不可知情况）对于有限VC维度情况是正确的，前提是对标记函数没有限制。不同之处在于，它们考虑的是一个期望的而不是一个高概率的框架，并且存在一个关于域分布D的条件，而猜想2被表述为对所有分布D都成立。然而，这个条件非常温和，该定理的基本假设是对标号函数没有限制。

前面两个结果的直觉是相同的：如果我们允许所有的标记函数，就没有关于X支持度的标记信息，我们还没有观察到。在监督和半监督学习者。在下一节中，我们将介绍半监督学习者可以有效学习的假设类。根据前面的结果，这些类被仔细选择并不奇怪。

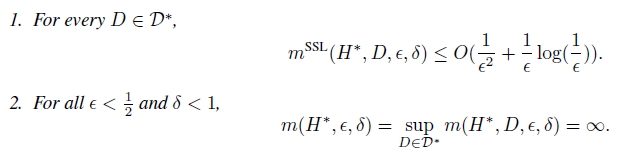
**3.2关于半监督学习可能性的证明**

我们考虑三种特定的设置，在这些设置中可以显示SSL可以学习，而SL不能。我们首先介绍Darnstädt等人的工作。[2013]和Globerson等人。【2017】旨在回答前一小节中的猜想1和2。它们表明存在一个假设类和一组域分布，使得在分布下，没有一个有监督的学习者能够学习。然而，假设任何，一个半监督学习者，能够访问有限的，但依赖于任意大的P，大量未标记的数据可以以相同的收敛速度学习。接下来，我们将介绍Niyogi[2013]的工作，因为它提供了一个最好的例子来说明当从SL和SSL转移时，从不可学习到可学习的转变是如何可能的。

3.2.1用离散集证明可实现情形

Darnstädt等人。[2013]给出第一个例子，说明猜想1通常不成立。这在下面的定理中得到，本节的其他结果将非常相似。

定理3（定理2）。存在一个假设类和一个域分布族，使得



为了使SSL能够为所有学习PAC，它需要了解完整的分布D（尽管对于每个固定的，有限数量的未标记数据就足够了）。由于有监督的学习者只能收集有标签的样本，因此在有限的样本数下永远无法获得这些知识，因此具有无限的样本复杂度。和的构造可以被认为是相当人工的。在接下来的两个小节中，我们将讨论在一个假设类中表现出相似行为的论文，该假设类松散地基于流形假设。尽管如此，我们还是想给出给定例子的直觉，因为它和其他例子一样，使用同样的技巧。

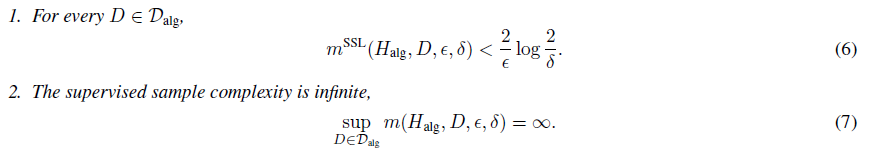
Darnstädt等人。[2013]举例如下。域X由任意长度有限和所有序列组成；上的分布，使得序列其中席是X的长度上的随机置换，假设类包含和常数0假设的所有假设hi。请注意，虽然这个类有无限的VC维，但是它仍然需要一些努力来证明没有一个有监督的学习者能够在中的所有分布中学习它。这是因为在上VC维数可能不是无限的。我们想描述一下SSL如何学习它。在固定和之后，我们绘制足够的未标记样本来识别所有位置，使得具有高概率0。对于所有这些指标i，我们可以从中去掉，因为常数0假设足以准确预测。然后，他们表明，中剩余的假设可以从有限多个样本中学习。请注意，限制可容许域分布是很重要的。如果也包括基本上对所有位置i施加相同权重的分布，则未标记的数据无助于限制。简言之：这个例子，以及下面的例子，基本上是这样设置的：H和D有一定的联系，在这些情况下，关于D的知识实际上可以给出关于H的知识。但是，注意，关于D的知识并没有限制H的可能的标记函数集。相反，D帮助确定了我们完全可以忽略。

3.2.2用代数变种证明不可知论情形

Globerson等人。[2017]使用一个假设类提供了一个不同的例子，该类大致遵循了多种假设。用同样的例子也可以证明猜想2，所以不可知论情形的不可能猜想，一般来说是不成立的。

这个定理与Darnstädt等人的理论非常相似。[2013]的区别在于假设集和分布集的构造。

定理4（定理5）。存在一个假设类Halg和一组分布Dalg。



假设类Halg包含所有在代数集上有类标签1的假设，本质上是流形的一种类型，并且在代数集之外有0。这仍然是一个非常有表现力的集合，具有无限的VC维。但是，如果我们将容许域分布Dalg的集合也限制为（某种类型的）代数集合，则具有知识的半监督学习者可以有效地学习：我们可以将Dalg看作支持可分辨代数集合V1……Vk的有限组合的分布集合。一旦我们知道这个分布支持V1……Vk，我们只需要找出哪些代数集有标签1，哪些有标签0。因此，SSL可以通过只考虑V1……Vk的组合中具有类标签1的假设来减少类暂停。由于所有可能组合的集合都是有限的，SSL可以用不等式（6）限定的样本复杂度来学习它们。注意，尽管真正的标记函数不一定是这个限制集的一部分，但是可以证明，用它的假设来预测总是最优的。这个论点类似于下面对不可知论的解释。

本文还讨论了当真正的目标函数不在Halg中时，这种论证可以扩展到不可知情况。这种扩展起初可能会出现问题，因为半监督算法限制了假设集Halg，为了保证PAC的可学习性，我们需要知道Halg的最佳预测值仍然在这个限制集中。但事实确实如此，因为Dalg的域分布集正是为保持这个特性而创建的。为了证明这一点，假设分布在不可约代数集V0上。我们的SSL现在可以选择将其完全标记为1或0，而这两个选项都可能导致非零错误。但是将它完全标记为1或0已经是理想的，因为使用任何其他代数集将导致这两个标记中的一个。这是因为，通过构造，V1等于V0（这导致将所有内容标记为1），或者具有零质量的交集（这导致将几乎所有内容标记为0）。

这似乎与3.1.5中的研究结果相矛盾，因为Lafferty和Wasserman[2007]表明，有监督的学习者也可以适应潜在的多方面因素。由于Lafferty和Wasserman[2007]在回归设置中工作，而Globerson等人。[2017]分析分类。然而，直觉是Globerson等人。[2017]向受监督的学习者提出一项不可能的任务，即不可学习的任务。另一方面，Lafferty和Wasserman[2007]限制了目标函数的光滑性，因此被监督的学习者面临一个足够容易的问题。Mey等人。[2019]确实表明，如果监督学习者被呈现一个可学习的分类任务，流形正则化只能通过一个常数来提高样本复杂度。

3.2.3使用流形假设使课堂可学习

Niyogi[2013]提供了另一种设置，半监督学习者可以有效学习，而监督学习者不能。然而，动机是独立于本大卫等人的。[2008]是对Belkin等人提出的流形学习框架的一般理论分析。[2006年]。此外，他们的结果是预期的，而先前的论文给出PAC界，这意味着他们持有高概率。尽管本文在预期框架中给出了结果，但我们稍微改变了设置，并将其呈现在PAC学习框架中。我们相信这足以理解这些想法，并使我们能够更好地与以前的论文联系起来。虽然这项工作是基于流形假设的，因此给定的域分布确实限制了可能的标记函数，但我们认为，理解监督学习者无法学习而半监督学习者能够学习的原因是最直观的设置。

示例如下所示。首先，假设容许域分布是由Pc类给出的，Pc类支持圆在欧氏平面上的嵌入，见图3。假设类Hc包含所有可能的半圆二值标记，而圆外的所有内容都标记为1,7，另请参见图3。知道圆的具体嵌入的SSL，只需要在给定的圆上找到两个阈值，一个VC维为2的类，就可以有效地学习SSL。在图4中，我们示意性地展示了为什么Hc有一个无限的VC维度，因此任何有监督的学习者都无法学习。

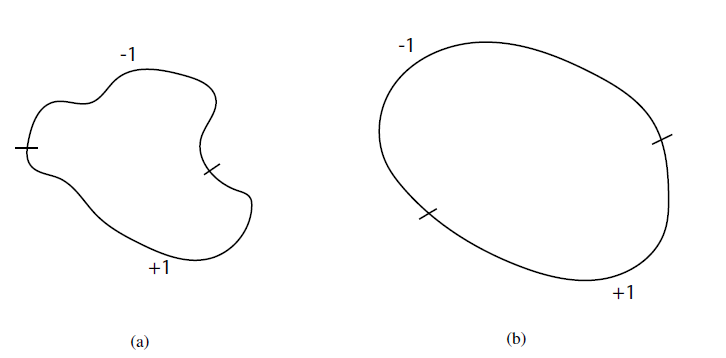
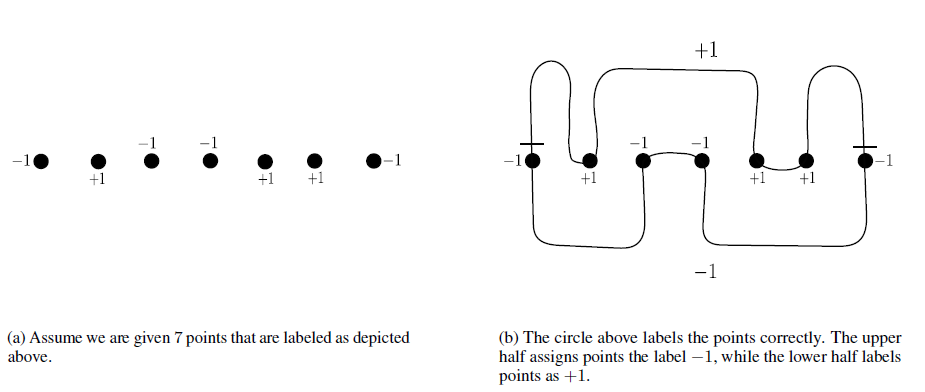


图3：（a）和（b）中所示的形状是圆在欧几里德平面上的两种不同嵌入。圆的一半被标记为1，而另一半被标记为1，而我们假设圆之外的所有东西都被标记为1。



（a） 假设我们得到7个点，上面描述了这些点。（b） 上面的圆圈正确地标记了这些点。上半部分为标签指定点1，下半部分为+1。

图4：假设集Hc具有无限维VC的原理性证明。给定（a）中的点，我们可以用（b）中给定的圆正确地标记它们。

4无假设学习

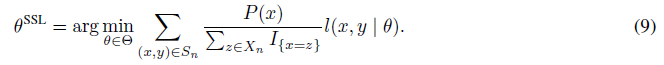
如前一节所述，在没有任何附加假设的情况下使用未标记的数据是很困难的，在某些情况下，可以证明未标记的数据根本没有帮助。如第3节导言中所述，这种不可能性有时源于我们只考虑改进条件概率的估计。我们在这一节中所做的工作着眼于完全风险，这个数量总是受到边际分布P（X）的影响。此外，没有对分布P作额外的假设，因此理论上的保证也很弱。我们首先介绍Sokolovska等人的工作。[2008]使用未标记的数据来重新称重标记点，并显示在渐近效率方面的改进。有趣的是，我们需要错误地指定模型来显示这个结果。其次，我们介绍了Käiriäinen[2005]的工作，他使用未标记的数据来选择版本空间的中心。最好的改进是以2为界的。最后，我们介绍了Leskes[2005]的工作，他使用未标记的数据来组合不同的假设空间，并表明学习率取决于这些假设空间中的最高Rademacher复杂度。

4.1按边际分布重新加权标记数据

Sokolovska等人。[2008]提出了一个半监督学习者，该学习者在重新权衡方案中使用边际分布P（X）的知识。为了避免理论分析的困难，他们将特征空间X限制为只包含有限多个点，并假设SSL可以访问全边际分布P（X）。他们考虑直接估计类概率的模型，而他们通过负对数似然来度量性能。然后他们分析渐近行为，特别是模型估计的渐近方差。他们比较了两个模型，经典的最大对数似然估计仅基于标记数据

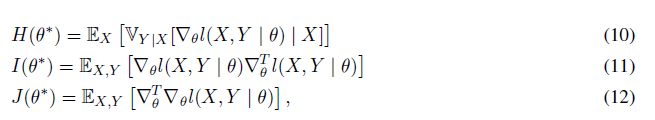


半监督学习者也考虑了边际P（x）

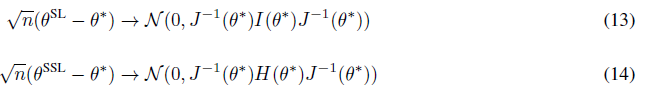


再次注意，半监督学习者用真实的而不是经验的分布来衡量每个特征。让我们先陈述一下关于的结果，然后再讨论它们。

定理5（定理1）。设并定义以下矩阵



式中，是条件随机变量上的方差，则是的一致渐近正态估计



是渐近有效的，这意味着它可以渐近地得到任何无偏估计量的最小方差。

现在问一下什么时候渐近地支配，我们得到了一个有点令人惊讶的答案：我们需要错误地指定模型。从统计学的角度来看，这也许并不奇怪，因为在明确指定的情况下（以及其他一些正则性条件），MLE本身已经是渐近有效的。具体地说，我们得到了，并且我们恢复了经典的结果，即MLE是渐近正态的，且具有逆Fisher信息矩阵的方差。然后用logistic回归模型检验了之间的差异特别大的情况。结果表明，这是离1/2越远的情况，特别是当Bayes误差很小时。这与Tsybakov的低噪声Tsybakov[2004]非常相似，后者用于统计学习以显示快速学习率。在第6.1节和第6.2节中，提出了类似的假设，以证明一些半监督学习者能够指数级快速收敛到Bayes误差。

4.2使用未标记的数据选择版本空间的中心

Kääriäinen[2005]介绍了一种通过使用未标记的数据收集关于两个分类器的一致性的信息来限制风险的方法。然后导出半监督估计量作为最小化该界的假设。不幸的是，这个想法只适用于现实情况。虽然我们还没有得到一个新的不可知情况下的算法，但本文仍然提出了新的边界基于未标记的数据。

4.2.1可实现案例

可实现案例的思想是考虑版本空间，因此包含所有没有训练错误的假设的空间。未标记的数据通过测量该空间上假设的不一致性在该空间上产生伪度量。我们将选择最坏情况下与所有其他情况不一致的假设，在我们假设可实现性时，其中一个必须是真实的。让我们更精确地说。给定两个假设，我们将不一致伪度量定义为



此度量在半监督情况下特别有用，因为它不依赖于标签。我们可以用经验公式来近似它



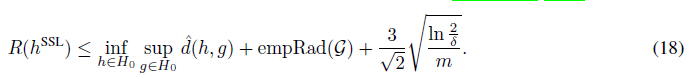
版本空间被定义为。假设是真的，那么我们知道，并且可以证明所有的。这允许我们



当不等式（17）限定了假设h的真实风险时，我们试图通过选择最小化不等式（17）右侧的假设来最小化该风险。更准确地说，我们选择半监督估计作为版本空间的经验中心，所以我们设置



有了这个，我们当然只能控制不等式（17）右边的第一项。然而，我们可以用从空间的Rademacher复杂度导出的浓度不等式来限制第二项，则在概率至少为的情况下是真的[Käriäinen，2005，定理3]



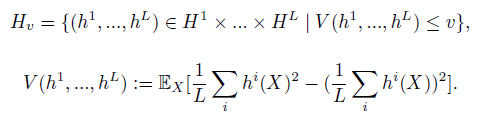
注意不等式（18）右边的两个项变为0表示m的增加，注意在这种情况下也是。所以忽略一分钟，我们只有有限多的未标记数据，我们可以比较SSL（16）纯监督解决方案。注意，在可实现的情况下，纯监督方法也会选择中的假设。由于受监督的学习者没有进一步的信息，我们总能找到一个目标假设，使得。因此，给出了任何监督学习者的最佳界。另一方面，SSL绑定（18）允许我们绑定，至少m是无穷大的。从几何角度来看是的直径，是半径。由于半径与直径之间的差（相对于d）最多为2，我们发现SSL和SL风险边界的差最多为常数因子2。

4.2.2一般情况的界限

在一般情况下，我们不假设目标假设是我们假设类的一部分。为了仍然使用所考虑的度量，在这种情况下，作者提出了以下边界的一般公式。其出发点是观察随机分类器的边界通常比确定性分类器的边界更紧[McAllester，2003c，Langford和Shawe Taylor，2002]。现在的想法是使用这样一个随机分类器frand作为锚，类似于可实现情况下的目标假设。为了得到分类器f的界，我们可以将随机分类器的界与包含的松弛项一起使用。根据我们采用哪种随机分类器，我们得到了不同的界限。这包括例如PAC贝叶斯界以及基于交叉验证和包装方法的界。它们显式地导出了一个交叉验证界，其中随机分类器由在多个交叉验证轮中获得的分类器上的均匀分布给出。

4.3使用未标记的数据组合多个假设空间

Leskes[2005]提出了另一个方案，该方案依赖于测量未标记数据假设之间的分类一致性。这里的想法是使用一个boosting方案，所以我们从不同的假设类开始。我们希望找到所有L假设类中最合适的假设。由于这通常会导致复杂度的过度增加，所以本文只考虑对未标记数据有充分一致意见的人。在这种情况下，充分意味着我们切换到一个新的假设类Hv，对于v>0，定义为



术语本质上测量L个不同假设中的不一致方差，并与未标记数据近似。假设类Hv只保留那些具有足够小的不一致方差的假设集合。然后，本文给出了一个同时适用于所有具有的的推广界，该推广界取决于l基本假设类的最大Rademacher复杂度。

5弱假设下的学习

在前两部分中，我们研究了在没有任何附加假设的情况下，半监督学习者的可能性。现在我们研究在我们称之为弱假设的情况下，SSL可以实现什么。对于弱假设，我们指的是那些本质上不能改变的学习，而是通过一个依赖于假设类的常数来改进的。在第6节中，我们将研究我们必须假设什么才能摆脱领域。我们首先介绍了Balcan和Blum[2010]的工作，因为它是一个通用框架，允许我们分析多个半监督学习者的学习保证。他们表明，在这个框架下的半监督学习者比监督学习者学习常数的速度快，其中常数取决于假设类和我们使用的半监督学习者。

然后，我们将更详细地介绍共同培训的想法。尽管在Balcan和Blum[2010]的框架中也可以看到联合培训，但我们想介绍一些更多的细节。特别是，我们介绍了Sridharan和Kakade[2008]的工作，他们在信息理论框架中提出了共同训练的假设，从而能够精确地量化偏差-方差权衡。

5.1编码弱假设的一般框架

我们从Balcan和Blum[2010]所做的工作开始，因为它提供了一种优雅的方式，可以在一个通用框架中形式化不同的假设。在这个框架中可以投射出许多现有的方法；传导支持向量机（Joachims，1999，Boyd and Vandenberghe，2004）、多视图假设（Blum and Mitchell，1998，Leskes，2005，Sridharan and Kakade，2008）和基于传导图的方法（Blum and Chawla，2001）。其思想是引入一个函数来度量假设h和边际分布P（X）之间的相容性。在这种情况下，兼容性可能意味着许多不同的东西。作为一个简单的例子，如果一个假设h的决策边界经过低密度区域，我们可以称之为与边际分布P（X）相容的假设h。由于我们通常只观察有限的样本量，因此需要为特征空间中的每个点定义函数，因此一个集合

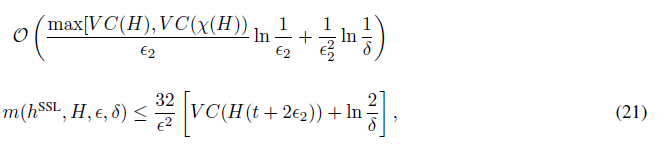


兼容性度量随后产生函数



我们称之为无监督损失。除了标记样品上测量的损耗外，我们还将尝试对其进行优化。本文又提出了几个与本文相同的定理。不同之处主要在于可实现性假设（关于无监督误差和有监督误差）和边界技术。它们给出了一致收敛的界和基于覆盖数的界。下面的定理是双不可知情形（标记的和未标记的损失都不必为零）。

定理6（定理10）。设。然后，假设未标记的样本大小至少为



式中，是最小化受

影响的的假设。这里，^R是用样本Sn测量的经验风险，是用样本Um测量的经验未标记风险。

我们注意到，最初的论文使用了不同的复杂性度量，因此术语是不同的。我们使用标准的VC维度来避免额外的符号，并允许与其他结果进行更简单的比较。他们使用了一种复杂性概念，在Vapnik[1998]中，可以在（指数化）退火熵下找到，并且具有依赖于分布的优点。

我们现在将定理6与前一节的结果进行比较，特别是与猜想1以及定理3和4中对此的答案进行比较。我们知道，在纯监督情况下，我们可以用代替，从而获得与（21）相似的样本复杂度。我们知道（21）给出的样本复杂度接近常数（比较Shalev Shwartz和Ben David[2014]第6章），我们知道，本文定义的纯监督学习者和半监督学习者之间的样本复杂度相差不能超过。因此，学习率的差距实际上是由一个常数给出的，这个常数只取决于猜想2所假设的假设类。然而，如果V C（H）是无限的，而是有限的，则该常数可以是无限的。这正是我们在第3.2节中提出的驳斥这一猜想的例子。

当我们使用假设时，定理6在某种程度上量化了SSL中的基本偏差-方差权衡。

利用半监督相容函数可以有效地限制初始假设空间H，从而减小训练过程的方差。但是，如果相容函数与潜在问题不匹配，则会使训练过程偏离好的解。

5.2假设特征空间可以分割

在多视图学习中，有时也称为联合正则化或联合训练，假设特征空间X可以分解为，并且每个局部特征空间X1,X2已经足够学习了。在Blum和Mitchell[1998]共同训练的早期工作中，在一个网页分类集中使用这个想法。功能的一部分，比如X1，是由网页本身的文本给出的，而另一部分X2，是由指向网页的超链接的锚文本给出的。其思想是，如果两个部分特征空间都有关于正确标签的足够信息，那么我们期望一个正确的分类器预测给定两个部分特征中任何一个的相同标签。因此，我们可以丢弃在这两种观点上有分歧的分类器。

关于这种方法有多种理论成果，可以在前一节的框架内举例分析。Rosenberg和Bartlett[2007]和Farquhar等人。[2006]分析多视图假设下的Rademacher复杂性项。Sindhwani和Rosenberg[2008]定义了一个核，该核直接将假设作为正则化项，从而找到一个自动发生共正则化的RKHS。

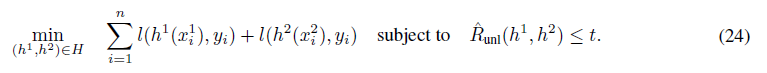
在这里，我们详细介绍了Sridharan和Kakade[2008]的工作，因为这与我们提出的其他结果有着最好的联系。此外，他们的信息论框架还允许分析假设不完全正确时所遭受的惩罚。我们将随机变量X分成X=（X1.X2）中的值。在他们的框架中，多视图假设可以形式化如下。

多视图假设假定是随机变量A和B之间的相互信息，条件是已知随机变量C。然后存在这样的信息



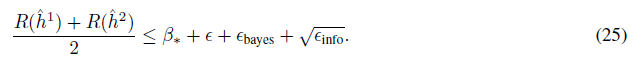
直观地说，一旦我们知道其中一个特性，另一个特性就不会告诉我们更多关于Y的信息。

将此与共同训练相比较，我们可以将其视为一种放松。在共同训练中，一个人假设每个视图已经足以完全学习，这里对应于。然而，如果，我们就无法从一个视图中完美地学习。（但在这个框架中这是可以的）。我们假设，对于每个视图和，都有一个相应的假设集和。我们将使用成对的假设进行预测。本文使用相容函数的概念（19）。特别是，它们定义了一个兼容性函数作为（，其中是一种满足松弛三角形不等式的距离度量，是一个样本。距离d从本质上衡量f1和f2在多大程度上一致一个样本x。对于给定的阈值，我们找到了具有约束经验风险最小化问题的最佳假设对



回想一下方程式（20）中的定义。主要定理保证了由上述过程找到的解，需要以下符号。设为分别从学习时用损失l测量的Bayes误差。我们还设置了，其中是Hi的最佳预测因子。最后，我们设置。

定理7。假设损失l以1为界。存在一个（取决和m），使得在某些进一步的正则性条件下，关于和损失l，并且给出至少标记样本，概率为



我们现在看到，信息论假设允许我们明确地描述当从完全假设集H切换到限制性假设集时引入的偏差。这个偏差是由给出的。

6强假设下的学习

在前一节中，我们分析了一些假设，这些假设只会给我们带来持续的改进，而不允许我们逃避的一般学习速率。现在，我们分析了一些假设，这些假设允许我们逃避这个机制，甚至可以给出指数级的快速收敛。下面的示例说明了其背后的基本思想。假设我们得到一组未标记的数据，并使用它对数据进行聚类。如果我们假设集群是正确的，也就是说每个集群对应于一个类，那么我们实际上只需要足够的标记数据来识别哪个集群属于哪个类。我们在本节中介绍的工作以各种方式扩展了这一思想，并回答了以下问题。如果我们有班级重叠怎么办？如果集群中有噪音怎么办？回归呢？

6.1假设模型是可识别的

Castelli和Cover[1995]完成了半监督学习的经典著作之一，该著作涉及一个与样本复杂度密切相关的主题。该设置非常受限，但可以在标记样本数n中给Bayes风险提供指数级的快速收敛速度。考虑到前面部分的结果通常不能本质上加快的速度（例如，在求解后比较不等式（21））这是非常强大的。

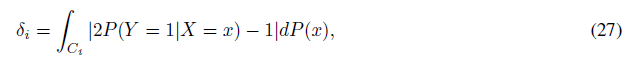
获得这些结果的第一个关键假设是数据生成过程。首先用和绘制标签，然后根据密度绘制特征。因此，未标记的数据是从混合物中提取的。第二个关键假设是混合模型的类是可识别的，即我们可以唯一地推导出混合模型，只给出未标记的数据。在观察到足够多的未标记数据以识别混合物后，我们只需找出如何标记两种混合物成分的每个部分。因此，我们只需在两个备选方案之间做出决定，就可以通过简单的似然比测试找到一个分类器h，该似然比测试在标记样本数n下指数快速收敛到Bayes风险：



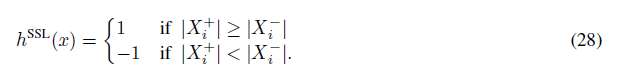
为了进行分析，有必要假设一个人有无限数量的未标记数据。这项工作在Castelli和Cover[1996]中继续进行，作者考虑我们已经了解密度的情况。Sinha和Belkin[2007]将类似的框架扩展到边际分布P（x）未知的情况。相反，他们假设P（x）可以用两个密度函数为和的球形高斯分布的混合来很好地估计。特别是，他们假设和可以用一个小数来限定，其中是Sobolev范数。最后，我们要提到Ratsaby和Venkatesh[1995]的工作，其中超额风险的指数衰减是在井规范的假设下实现的，模型类是两个球面高斯分布的混合。

6.2假设类是集群和分离的

在Rigollet[2007]中，我们使用另一种聚类假设的公式给出了泛化误差的显式界。它与前一节的工作非常相似，在他们的假设下，我们再次获得指数级的快速收敛。他们的第一个简单的设置是，我们得到一组成对不相交的簇我们做一个簇假设，即我们假设标记函数在每个集群Ci上是恒定的。所以这些簇有某种程度的标签纯度，我们可以通过



其中集群Ci是纯iff是1或0。假设我们知道簇，我们让作为每个簇的多数投票分类器。更正式地说，给定一个标记样本Sn，让和类似的。然后给定一个新的数据点，我们设置



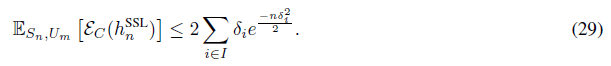
注意，这只定义了集群上的一个函数。然而，该论文认为，未标记的数据在没有观察到未标记数据的情况下是没有帮助的。因此，它只分析集群上未标记数据的可能增益。

因此，超额风险现在被限制在集合，所以我们将超额风险设置为



其中是Bayes分类器。下面的定理描述了一个人可以从预期的集群超额风险中获得的收益。

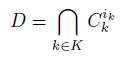
定理8（定理3.1）。设是一个集合，其表示所有，以便此集合满足上述定义的集群假设。则上述多数投票分类器满足



因此，知道了这些聚类，我们就可以在Castelli和Cover[1995]中恢复标记样本量的指数收敛性。然而，本文的最大努力在于对聚类的定义和聚类的有限样本估计。推导过程相当长，在这里我们仅限于描述潜在的直觉。首先我们假设边缘分布P（X）允许密度函数关于Lebesgue测度。有了这一点，就可以定义 w.r.t.的密度水平集参数>0为。对于固定的>0，我们认为聚类本质上是密度水平集的路径连接组件，在这里可以确保排除病理病例。在有限多个未标记样本的情况下，估计集将松弛项添加到不等式（29）中，该不等式在未标记样本大小上多项式下降。因此，为了保证我们仍然能够快速学习，未标记样本的数量必须随着标记样本的数量呈指数增长。

6.3假设类是集群的，但不一定是分开的

辛格等人。[2008]对集群假设提出了一种不同的形式化方法，它允许区分SSL有帮助和没有帮助的情况。这是通过限制分布P的类别，然后研究哪些分布允许成功的半监督学习来实现的。P类的构造使得边缘分布由不同的簇组成，这些簇有时容易区分，有时不易区分。P的边际密度p（x）由K密度pk的混合物给出。所以，，，每个PK在满足一定正则性条件的集上都有支持。我们称之为集合Ck簇，假设每个集合都有自己的光滑标记分布函数。所以在概率ak下，我们从中提取，然后根据标记x。我们进一步只考虑那些导致具有边际的集群（没有重叠）的分布（参见图5），并用表示得到的分布类。在这个公式中，聚类不是主要的兴趣，而是作者所说的决策集。为了定义一个决策集，我们用表示Ck的补码，并定义。如果集合称之为决策集如果可以写成



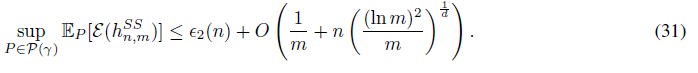
对于，请参见图5（b）中的示例。与聚类相比，决策集的优势在于，在每个聚类上，完全分布p（x，y）不一定是光滑的，因为它们可能在边界处出现跳跃。然而，在决策集上，如果每个是光滑的，则p（x，y）是光滑的。因此，如果我们知道决策集，我们可以使用使用平滑假设的半监督学习者。

主要定理回答了人们是否可以从有限多个未标记点学习决策集的问题。

定理9（推论1）。设是相对于贝叶斯分类器的超额风险。假设E是有界的，并且有一个学习者知道所有决策集D并满足以下超额风险界。

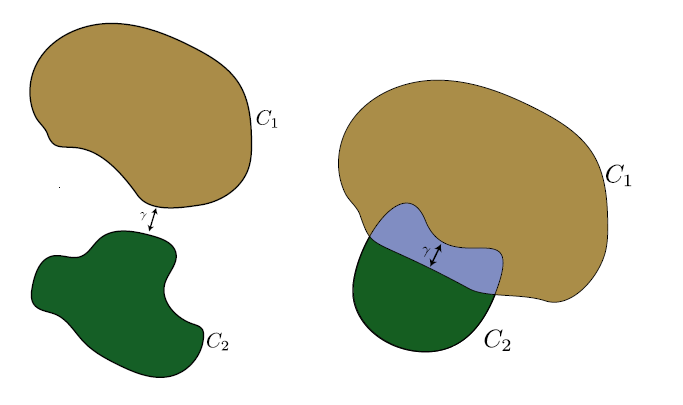


假设，其中是常数，则存在半监督学习者使得



注意以下几点。如果知道决策集的学习者的收敛速度为，则根据不等式（31），未标记数据需要以的速度增加，以确保半监督学习者具有与相同的收敛速度。例如，如果的收敛速度是指数级的，那么我们需要一个指数级的无标记数据，这与上一节中的结果相同。

这里的直觉相当简单。我们需要估计决策集D的未标记样本越大，一旦我们知道这些，我们就可以像一样执行。为了分析首先在经验上学习决策集的半监督学习者是否比所有的监督学习者具有优势，他们首先为所有的完全监督学习者找到最小最大下界。然后，它们给出了特定半监督学习者的上界，并且结论是直观的：为了使SSL有用，参数和未标记样本的数量应使得完全监督学习者不能区分决策集，而半监督学习者可以。所以不应该太大，因为这样有监督的学习者也可以区分决策集。当然，未标记的数据不应该太少，因为半监督学习者无法区分决策集。



（a） 簇C1和C2用边缘隔开。不同的决策区域在这里只是集群。

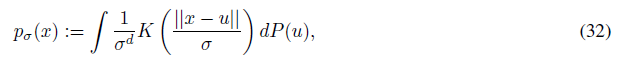
（b） 簇C1和C2有一个带边缘的重叠（浅蓝色）。这三种颜色也构成了三个不同的决策集。

图5：图（a）表示正边距的概念，图（b）表示负边距。

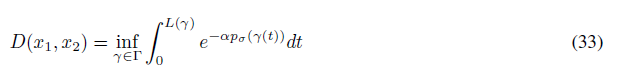
为了说明SSL和SL之间的具体区别，作者假设且条件期望是x中的Hólder-光滑函数。根据本文给出的情况，当SSL可以基本上快于SL时的一个表。在这些情况下，SL的收敛速度为时有一个预期的下限，而SSL的收敛速度为

6.4假设回归函数沿流形平滑

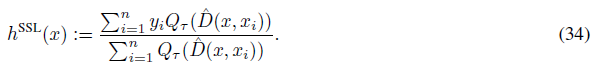
正如我们将在讨论部分进一步阐述的，SSL中的一个问题是，大多数方法都基于对完整发行版的假设。问题是，我们通常无法验证这些假设是否成立。这一点至关重要，因为如果假设不成立，我们很可能希望使用一个有监督的学习者。Azizyan等人的工作。[2012]是少数涉及该主题的论文之一，因为他们介绍了一个依赖于参数的半监督学习者，其中=0恢复一个纯监督学习者。然后给出了交叉验证时半监督学习者的泛化界。由于这项工作使用回归设置，而大多数其他的论文处理分类，并给出了一个干净的形式化的SSL，我们在这里提出的细节。作者使用流形假设的一个版本，因此我们强制我们的估计回归函数在高密度区域平稳地运行。用平滑密度函数测量边缘分布P（X）的密度



其中K是的一个对称内核，具有紧凑的支持，且>0。设为所有连续路径的集合以单位速度从，其中是的长度。利用这一点，我们可以在上定义一个新的度量（所谓的指数度量），它依赖于参数平滑密度。



首先注意=0对应于欧几里德距离。第二，注意两点x1和x2之间的路径上的高值导致新度量中这些点之间的距离较短，这一点用大强调。如果我们假设Q是另一个核，并且我们设置，我们可以将半监督估计定义为



因此，该估计器是最近邻回归器，其中邻域根据其在Dmetric中的距离加权。以下定理给出了假设时平方风险的界。

定理10（定理4.1）。设P是一类概率测度，它根据参数（定理后的更多细节）满足某些正则性。假设对于所有，我们都有，那么

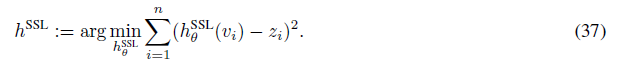


在这个符号中，是D-度量中P的覆盖数：尺寸为的X中的最小闭合球数w.r.t到覆盖P（X）的支撑所需的D-度量，另见Shalev Shwartz和Ben David【2014年，第27章】。在欧几里德情形下，当=0时，我们可以用常数C来约束。当>0和P（X）集中在维数小于d的流形上时，覆盖数可以小得多。关于的正则性条件基本上如下。首先我们假设P（X）具有紧支撑。其次，所有回归函数是L-Lipschitz连续的，其中域具有指数度量D，而协域具有欧氏距离。由于前面的定理可能很难解析，本文在一些进一步的正则性条件下给出了一个简化的推论。

推论1（推论4.2）。假设在一定范围内。此外，假设m足够大并且被很好地选择。那么对于所有的



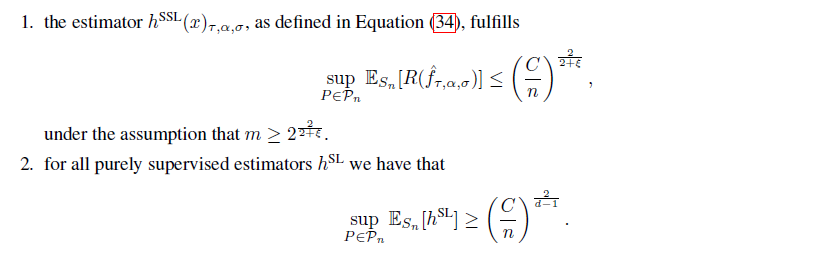
然后，本文分析了我们在寻找最佳时所遭受的额外惩罚。这是通过离散化参数空间，使得和来实现的。假设除了训练样本Sn之外，我们还有一个验证集，这也方便了n的大小。让作为在Sn上训练的半监督假设，参数为。然后，我们通过交叉验证选择来选择最终假设



定理11（定理6.1）。设为超额风险，其中为真回归函数。有一些常数不是普遍的，在某种程度上取决于问题0<a<1和，这样



这特别有趣，因为我们隐式地与有监督的解决方案进行比较，只要我们包含  
。从不等式（38）可以看出，验证过程引入了一个惩罚项。在最坏的情况下，如果我们使用半监督方法，这可以看作是一个附加的错误项，但这种假设实际上是不正确的。最后，作者确定了半监督学习率可以严格优于监督学习率的情况，就像我们在第3.2节中看到的那样。特别是，它们构造了一组分布Pn，这取决于标记样本的数量，这样



为了获得本质上不同的学习率，我们需要，如果P集中在一个维数严格小于d-3的集合上，就是这种情况[Azizyan等人，2012，引理1]。值得注意的是，Pn的构造是通过对大n的分布进行更多的集中来完成的。如果Pn不集中，并且对大n保持平滑，则标记数据已经足够接近边际分布。

这与第6.3节中介绍的工作类似，因为它们还表明，只有当边际分布P（X）不太容易识别时，SSL才能工作。我们还可以与第3.2.3节中的工作进行比较；如果我们限制域分布，以便只允许平滑圆嵌入，则有监督的学习者也可以有效地学习。这是因为有限数量的标记样本足以均匀地学习领域分布，因此半监督学习者会失去其优势。

7导入案例中的学习

虽然许多方法使用未标记的数据来寻找更好的分类规则，但有些方法考虑只关心未标记数据的标签的方案。这些方法通常被称为传导法[Vapnik，1998，第8章]。我们提出了最重要的理论结果。在Pechyony[2008]的第2章中，我们对理论和实践上的跨生产性学习进行了更为详细的调查。在第7.1小节中，我们给出了传导情况下的学习边界。

它们往往是归纳格和相关概念的直接延伸。在第7.2小节中，我们提出了两篇论文，涉及所谓的“安全半监督学习者”这一主题。他们的目标是构建一个半监督学习者，这种学习者永远不会比他们的监督学习者差。

我们可以区分两种传导设置，其中的本质区别是，在一种设置中，我们采样而不替换，因此样本变得依赖。我们在这里介绍的关于跨导学习的工作主要是为了方便而涉及到设置1。但是，我们注意到，可以将设置1的边界转换为设置2的边界[Vapnik，1998，定理8.1]。

设置1

1。我们从一组固定的点开始。

2。我们揭示了随机均匀选择的集的标签Yn。为了便于注释，我们通常假设w.l.o.g，Xn是Xn+m的第一个n，Xm是Xn+m的最后一个m点。

3。基于和Xm，我们试图找到一个在上具有良好性能的分类器h。

设置2

1。我们从上的固定分布P开始。

2。我们根据P抽取n-i.i.d.样本，得到训练集Sn。我们根据P（X）绘制m i.i.d.样本以获得测试集Xm。

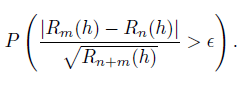
3。基于Sn=（Xn；Yn）和XM，我们试图找到一个在上具有良好性能的分类器h；。

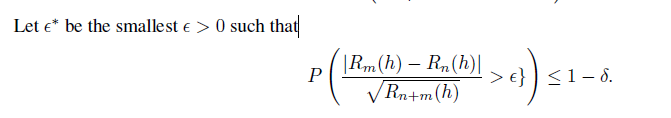
注意，在本节中，我们的测试误差用Rm（h）表示，训练误差用Rn（h）表示。这反映了测试的大小是m，而训练集的大小是n。我们这里不使用hat符号，因为在传导设置中，我们不一定有一个潜在的分布。

7.1传导学习界限

7.1.1 Vapnik的隐式传导界

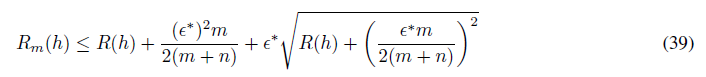
传导推论可追溯到Vapnik[1982]。我们给出了在定理8.2中作为方程（8.15）找到的结果。摘自Vapnik[1998]。假设我们有n+m个样本，我们随机选取n个样本进行训练。然后我们要估计我们在剩余m个样本上的误差。Vapnik表明，超几何分布描述了列车和测试集上观测误差大于的概率





使用一致边界并替换，可以得到以下结果。

定理12。对于所有，以下不等式的概率为



这种不平等的问题在于，术语是n；m；和h的一个隐函数，因此，我们不清楚实际能达到什么样的学习率。下一节将介绍这个问题。

7.1.2边界作为归纳边界的直接扩展

不等式（39）的传导界很难解释，因为它包含一个只能隐式计算的函数。Derbeko等人。[2011]在PAC-Bayes框架中找到显式的传导边界。本文给出了一个界，它实质上是McAllester[2003a]归纳界的直接推广。为了展示结果，他们使用吉布斯分类器。为此，让q是假设集上的任何分布H. The Gibbs分类器Gq将新的实例与相应地绘制为q。在集合席上的Gq的风险是。

定理13（定理17）。设p是H上的任意（先验）分布，它可能依赖于Sn+m，设>0。然后，对于任意随机选择的子集和H上的任意分布q，其概率至少为

这个定理实际上是McAllester[2003a]中等式（6）下归纳监督情形的直接推广，唯一的区别是缺少项。尽管McAllester[2003b]表明，在一定条件下，人们可以在看到Sm之后选择先验p，但归纳PAC贝叶斯理论通常不允许这样做。在传导性设置中，这是允许的，因为我们只关心Sn+m集上的点的性能。在某种程度上，这与使用固定分布学习相同，当我们的固定分布在有限多个点上只有质量时[Benedek和Itai，1991]。

Derbeko等人。[2011]通过使用集群方法选择一个优先p来利用这一点。更准确地说，在观察了数据集Xn+m之后，我们在它上构造了c个不同的聚类。每个聚类通过将一个聚类中的所有点分配给同一个类来产生多个分类器。然后在这些分类器上放置一个一致的先验p，通过最小化不等式（40）选择分类器上的后验分布q，得到Gibbs分类器Gq。

将此方法与完全监督（因此必然是归纳）情况进行比较，我们认识到，可能的性能改进与半监督学习（假设）中的改进具有相同的风格，如第5节和第6节所分析的。使用上述的聚类方法将减少来自KL（qjjp）的不等式（40）中的惩罚。换句话说：我们减少了分类器的方差。另一方面，使用聚类方法会使我们的解产生偏差，并且如果簇具有高杂质，也就是说簇没有明确的多数类，我们将在监督解上降级。

7.1.3基于稳定性的界限

在El Yaniv和Pechyony[2006]中，我们在稳定性的概念下研究了传导性界，假设如果我们稍微扰动输入，分类器的输出不会有太大的变化。传导边界是使用一致稳定性概念的归纳边界的扩展[Bousquet和Elisseeff，2002]和弱稳定性概念[Kutin和Niyogi，2013，Kutin，2002]。给出了基于一致稳定性的更简单的传导界，并解释了弱稳定性的区别。

假设是一个transductive学习者，所以我们（确定地）根据一个标记集Sn和一个未标记集Xm选择一个假设。此外定义和。所以是我们在Sn中用测试集的第j个例子替换训练集的第i个例子时得到的集合。我们认为，如果所有选择，对于所有，和是一致稳定的。

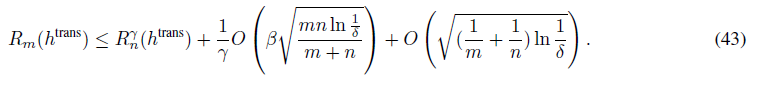


换言之：如果输出变化小于如果我们从训练和测试集中交换两点，则transductive学习者是一致稳定的。边界是用-裕度损失来表示的。对于>0，我们设置



因此，当用损失衡量时，我们为h的风险写。注意损失收敛到0-1损失。

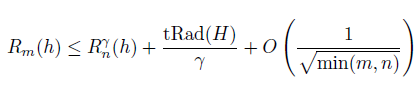
定理14（定理1）。让成为一个-一致稳定的传导学习者，并且；>0。那么，在所有训练和试验分区上，我们至少有1-的概率



注意将依赖于n和m，如果我们交换来自训练集和测试集的两个样本，我们预计我们的训练集越大，我们的算法更改就越少。在基于Rademacher复杂度的反导界中，在下一节中，可以实现的收敛速度。要获得与不等式（43）相同的速率，我们需要表现为。正如Johnson和Zhang[2007]所证明的，正则RKHS方法确实可以达到这种稳定率。

7.1.4基于传导Rademacher复杂性的界限

Rademacher复杂度是一个经过充分研究和确定的归纳案例中风险边界的工具[Bartlett等人，2005]。El Yaniv和Pechyony[2009]引入了这些量的一个传导版本。而在归纳的情况下，我们必须在看到任何数据之前选择我们的假设类，而传导的情况允许我们选择依赖于数据的假设类。假设类H的transductive Rademacher复杂性的定义紧跟归纳情况，将用tRad（H）表示。利用边际损失函数（42）和相应的经验风险，本文表明，对于所有

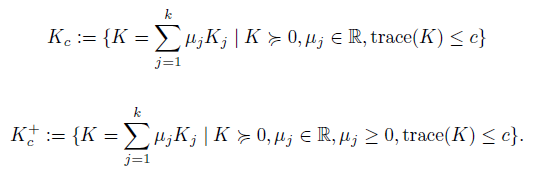


乍看之下，有标签和无标签的数据在收敛性方面扮演着同等的角色，这似乎有些令人惊讶。虽然n<<m的缓慢收敛并不令人惊讶，但人们必须认识到，在m<<n的情况下，传导风险具有非常高的方差，因此我们有很大的高置信估计区间。该界可用于直接估计传导算法的传导风险。

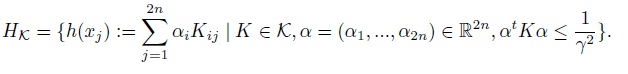
马克西莫夫等人。[2016]使用不同的Rademacher复杂性来推导特定多类算法的风险边界。他们的算法使用基于完整数据的给定聚类来找到一个在某种程度上与所发现的聚类相兼容的假设。然后，transductive多类Rademacher复杂度直接利用这个聚类。利用该算法，作者证明，如果我们有K个初始类，则可以达到的顺序的学习速率[Maximov等人，2016，推论4]。尽管我们注意到这项分析是在设定2的情况下完成的，但毫不奇怪的是，学习率基本上与二元传导的情况相同。

7.1.5基于学习内核的边界

作为归纳案例的直接延伸[Bartlett和Mendelson，2003]，Lanckriet等人。[2004]提出利用未标记的数据来学习适合于传导学习的核。其思想是使用核方法，允许从某一类核中进行选择，以优化目标函数。所提出的PAC界表明，在核类的复杂度和经验误差之间有一个很好的折衷，可以获得良好的（传导）性能。它们的示例性内核类设计如下。给定一组初始的内核，它们定义在有标签和无标签的数据上

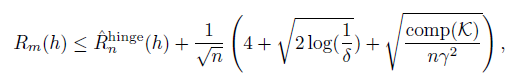


每一类核产生一个假设集



本文发现的误差范围如下。

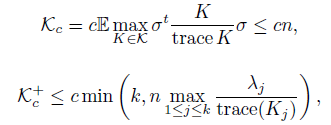
定理15（定理24）。对于每一个>0，且在从（X；Y）中均匀选择的尺寸为n（so m=n）的训练和测试集上，概率至少为1-，则每个函数具有



其中是h的经验铰链损失，comp（）是的复杂性度量，定义为



是2n个Rademacher变量的向量。前面定义的核类的复杂度度量可以由



其中是Kj的最大特征值。

注意，由于m=n，我们发现这个界给出了与第7.1.3节和第7.1.4节同样的的学习率。

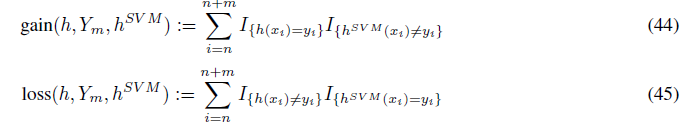
未标记的数据对这个过程的影响取决于最初的内核猜测，但是对本文没有进一步的兴趣。我们可以在Chapelle等人中找到扩展。[2006年]（第282页，底部），其中Ki是以特定方式选择的：如果假设是图拉普拉斯L的第i特征向量，我们可以设置。如Chapelle等人所述。[2006年]（第280页）如果我们在的特征值大的情况下强制很小，那么我们就可以强制这个过程发现的分类器沿着数据流形平滑。Johnson和Zhang[2008]也得到了类似的结果，其中最大的差异是使用的内核。Johnson和Zhang[2008]没有使用核的初始集合，而是使用给定核的谱分解并收缩它，其中收缩依赖于未标记的数据。

7.2安全的导入学习

在半监督学习社区中，众所周知，使用半监督程序通常会带来性能下降的风险[Chapelle等人，2006年，第4章]。这个问题导致一些作者提出这样一个问题：是否可以以安全的方式进行半监督学习，这意味着可以保证SSL不会比受监督的SSL差。到目前为止，我们主要比较了SSL和SL的风险边界。但是，即使风险边界的假设是正确的，较小的边界仍然不能保证改进。我们将特别关注李和周[2011]和卢[2016]的工作。这两项工作的结果都基于minimax公式，并且表明，在某些假设下，通过SSL确实可以保证改进。在导通设置1中也进行了分析。这意味着我们有一个训练集Sn和一个测试集Xm。

7.2.1支持向量机的极大极小方法

李和周[2011]提出的模型的基线是S3VM[Bennett和Demiriz，1999]，它通过寻找低利润率的解决方案来考虑未标记的数据。提出的S4VM模型找到了几种不同的低利润率解决方案，然后在minimax框架内从中选择，以对冲可能出现的最坏情况。假设我们找到了一组建议的解决方案。其思想是将这些解决方案与监督解决方案hSVM进行对比。现在假设我们知道Xm的真正标签。利用此方法，我们可以计算出当监督hSVM与任何其他分类器h进行比较时的性能增益和损失。



如果我们把我们的目标定义为这两者的区别



我们可以将半监督模型定义为该差异的最大化者。由于我们实际上不知道真正的标签，我们假设一个最坏的情况下，导致以下的最大-最小公式。



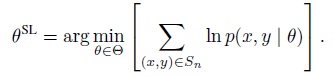
这里是我们可以用Hp实现的所有可能的labeling集合。为了保证SSL不比SL差，必须假设真正的标签Ym是set Yp的一部分，因为只有这样我们才能保证以下内容。

定理16（定理1）。如果，则当在未标记数据Xm上测量性能时，的精度永远不会低于的精度。

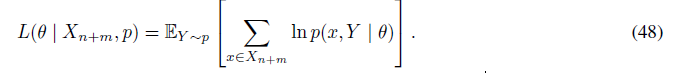
同样，关键的假设是，在这种情况下正好对应于低密度假设。这是因为集合Yp包含来自满足低密度假设的分类器的可能标记。我们可以想象在不同的假设下使用相同的过程，因为我们可以通过Yp对它们进行编码，Yp是我们认为可能的所有labeling的集合。虽然本文仍需要一些假设，但Loog[2016]展示了一个我们获得无假设保证改进的案例。然而，这样做的代价是衡量改进的可能性，而不是准确性。

7.2.2生成模型的极大极小方法

据我们所知，这方面的第二篇论文可能是半监督学习中考虑完全无假设情况的唯一一篇论文。这当然是要付出代价的，以后会有更多。起点是上的概率密度函数族，其中是参数化。首先，我们将设为模型的监督极大似然估计，因此



假设我们知道真实条件概率，对于 ，对于。如果我们知道这一点，我们实际上宁愿优化在完整数据集上计算的模型的预期对数似然性，



为了优于监督模型支持完全（传导）似然（48），我们希望最大化似然增益。所以我们想找到最大化似然增益的



我们不能直接最大化（49），因为我们不知道类真概率分布P。我们反而为所有标记点设置，这给了我们向量，对于未标记点xm，我们考虑最坏的情况，这导致下面的最大最小公式。



注意，向量pm可以是未标记数据Xm的真标签Ym。还要注意 表示所有，因此特别是如果pm=Ym，正如我们总是可以选择的那样。这意味着下面的定理成立。

定理17（引理1）。设为方程（50）中的解，则



对于模型的某些特定选择，前面的不等式几乎肯定是严格的。因此，我们保证了半监督模型的传递似然比监督模型的更大。

这项工作与上一节的一个重要区别是，在本文中，使用了一个生成式

这项工作与前一部分的一个重要区别是，本文采用的是生成模型，而Li和Zhou[2011]使用的支持向量机是一个判别模型，它固有地优化了类概率p（y|x）。Krijthe和Loog[2018]也见第3.1.4小节，表明在某种程度上实际上有必要使用生成模型：方程（50）的半监督估计将与一大类判别模型的监督估计一致。联合模型p（x；y）在这种情况下起作用的原因有多种解释。直观而明显的一点是，该模型的可能性考虑了边际分布P（X），这是一个可以从未标记数据中测量的量。

8讨论

我们讨论了过去四十年来在半监督学习领域提出的主要理论思想和结果。具体地说，我们注重的是能够告诉我们其潜力和缺乏这种潜力的结果。我们讨论了问题的答案：半监督学习的极限是什么？不同方法的假设是什么？如果假设是真的，我们能得到什么？我们喜欢总结我们的调查，并提到一些我们认为是核心的认识。

8.1关于免假设SSL的限制

在第3节中，我们回顾了分析半监督学习的局限性的工作，在没有对分布做出特定假设的情况下，半监督学习者可以利用这些假设。最一般的公式在猜想1和2中得到。它们基本上说，半监督学习者最多可以用一个常数打败所有监督学习者。然后，我们提出的工作表明，这些猜想并不具有完全的普遍性，而是在特定的情况下。它们本质上适用于有限维的可实现情形和假设类，而不适用于无限维的可实现或不可知情形。在有限的VC维数下，不可知PAC学习的情况还有待研究。

8.2持续改进有多好？

第3.1.6节和前一小节研究的问题是，半监督学习者是否能够在样本复杂度方面提供持续的改进。然而，人们也可以问这样一个问题：在实践中不断改进已经有多好。这个问题的答案可以通过一个思维实验得到。假设我们有两个由两个同心的d维球体给出的类。假设我们有足够的未标记数据用于流形正则化方案来识别球体。这样，半监督学习者只需要一个标记样本就可以给出一个完美的分类，而每个监督学习者都需要一个标记样本大小，这个标记样本大小随着维数d的增加而增加。尽管流形正则化分类只需要两个样本，我们从Mey等人那里知道。[2019]流形正则化只能实现持续改进。

这看似矛盾，但当我们研究风险投资的维度时，这种行为很容易理解。如果监督分类器使用一个假设空间H，我们可以将流形正则化解释为切换到一个限制空间。该空间仅包含满足流形假设的假设，其中正则化参数指示该假设的实施程度。Mey等人。[2019]表明流形正则化的改进最多为。如果我们将设得足够高，我们可以保持变，而VC（H）将随着维数d的增加而增加，这表明常数的提高可以任意高。尽管本例使用了流形假设，但Golovnev等人。[2019]举一个半监督学习者的例子，他完全了解领域分布。我们在第3.1.6节中解释了这个特殊的例子。这表明，如果我们有进一步的假设，比如流形假设，或者对边际分布有充分的了解，那么常数的改进是任意高的。如果一个人可以拥有任意高的常数，而不需要假设，并且只有有限的未标记数据，这是一个悬而未决的问题。

8.3我们需要的未标记数据量

在第3.2节中，我们介绍了三种设置，其中半监督学习者可以学习PAC，而没有监督学习者可以。为此，原则上，我们需要无限量的未标记数据，而且我们也不能创建一个不是这样的例子。如果在任何给定的分布P下，固定的有限数量的未标记数据足以学习，那么我们可以使用相同的策略以有监督的方式学习，就像我们总是选择忽略标记一样。这些示例的工作方式是，对于每个固定P，有限数量的未标记数据是足够的，但是这个数量可以任意大。因此，如果我们想了解所有可能的分布，我们需要任意大的未标记数据量（=）。我们在第4节、第5节和第6节中介绍的半监督学习所能提供的改进并不一定需要无限量的未标记数据，尽管它有时是为了方便而假设的。不同的是，在这些设置中，受监督的学习者也能够进行PAC学习，但半受监督的学习者能够使用较少的标记样本来完成这一任务。在第6.2节和第6.3节中，我们看到了两个集群假设的实例，作者表明，未标记数据的数量需要随着标记数据的数量呈指数增长，才能利用这个假设。这是因为发现簇的误差仅在多项式上减少未标记点的数量，如不等式31所示。

8.4在半监督学习中使用假设

在第5节和第6节中，我们将研究半监督学习者在做出假设后可以实现的目标。半监督假设是域分布和标记函数之间的联系。特别地，我们假设在看到特定的域分布之后，我们可以忽略某些标记函数。例如，群集假设将排除不为属于同一群集的点分配相同标签的标记函数。显而易见，但真正的问题是，我们不知道这些假设是否成立。我们推测，如果这样的假设是真的或者不是真的，那么测试所消耗的标记点就相当于直接在有监督的学习者的指导下学习一个好的分类规则。为了使这个陈述更精确，我们定义了一个假设作为上分布P的一个性质。让P a是上的一组分布。我们说P满足假设iff 。例如PA只能包含这样的分布，即边际分布P（X）总是支持簇，每个集群都有一个唯一的标签。然后P满足这个特定的簇假设，即iff，需要注意的是，假设A是P上的一个性质，所以我们需要标记样本来检验它是否正确。因此，为了减少半监督学习者收敛的标签数据消耗假设的不确定性，比较标签数据的消耗是很有兴趣的。当然，我们可能事先知道假设是真的，不需要检验它，但如果不是呢？

第6.4节回顾了少数分析这一问题的工作之一。Azizyan等人。[2012]表明，如果假设是真的，我们可以得到更快的利率，但如果假设不是真的，我们将支付的罚金。Balcan等人。[2011]研究如何以积极的方式测试一个属性，以便我们可以选择要标记的样本。然而，这种测试程序对半监督学习的影响并不清楚。当然，我们可能会声称，甚至没有必要按照Vapnik的原则来测试假设是否正确：当我们最终只关心半监督学习者的表现是否更好时，为什么要测试假设是否正确？我们认为这是半监督学习中一个重要的开放性问题。