小样本学习

小样本学习（FSL），是一种机器学习的方法，它可以通过有限的训练集训练模型。

在机器学习领域，常见的学习策略是从大规模数据集中学习数据分布，从而使模型更好地拟合某一类问题，因此提供足够多的数据是机器学习拟合的关键。小样本学习希望通过少量的样本完成模型的学习构建，在技术上更加具有挑战性。

**一、问题引入**

在理论层面上，人类智能的一个显著特点是能够通过一个或几个示例就快速建立起对新概念的认知能力，例如人们可以在看到几个例子后发现手写字符之间的差异，这是深度学习模型所不具备，而通用人工智能模型需具备的能力。

在技术层面上，受限于成本或客观条件，大规模数据集有时并不容易获取，如对于动物识别问题，珍稀动物的数据会相对较少，这制约了模型最终的学习性能。

在现实层面上，我们有着从少量样本中学习模型的应用需求，如智能手机可以通过学习少量的用户照片，就能具备识别用户人脸的能力。

因此，小样本学习便应运而生，我们希望通过小样本学习使模型具备从少量样本学习的能力，以此解决上述提到的问题。

目前，小样本学习的技术优势在于：

（1）能解决只具有小规模数据集的问题

（2）能减少数据采集的成本

（3）能减少模型的计算成本

**二、概念定义**

给定输入数据x和监督数据y，设X和Y分别为输入数据和监督数据的数据空间。

一个典型的FSL任务被描述为：

其中，

用于任务T的样本来自一个特定的域，由数据空间和边际概率分布组成。

通常中有C个任务类，每个任务类包含K个样本数据（K的值很小），即，将此类任务称为C-way, K-shot任务，目的是生成一个预测函数，该函数用于对中的样本进行预测。受限于样本数据的数量，我们很难用构建高质量的函数模型。

因此，我们引入一个有监督的辅助数据集：

其中，且。

和源于统一数据域，即，但是。

由此，小样本学习可以定义为：给定任务T，以及小规模的有监督数据集和与T不相关的大规模的辅助数据集，小样本学习的任务是利用中少量的监督信息以及中的知识，为任务T构建函数f。

**三、解决方案**

**（一）误差理论构建**

对于某一映射(x, y)，求解它的理论构建如下所示：

假定函数为h，我们希望在数据分布p(x,y) 内，最小化期望风险：

但是由于数据分布 p(x,y) 是未知的，所以在实践中我们通过最小化经验风险来逼近最优结果：

因此，我们可以有三种函数结果表示：

* 表示最小化期望误差的函数结果
* 表示在H假设集内最小化期望误差的函数结果
* 表示在H假设集内最小化经验误差的函数结果

在现实中，我们只能计算得到，因此其理论误差包含两部分：

可以将两个理论误差记为和，前者衡量了函数空间H能否很好地描述，后者衡量了利用数据集拟合的结果与最佳拟合结果的差距。

由此，如果希望减少最终函数与理想结果的误差，可以从以下几个方面入手：

（1）数据，决定了数据集。

（2）模型，决定了函数空间。

（3）算法，决定了在函数空间内搜索适合的最优解的结果。

**（二）解决方案**

在小样本学习中，可用的训练样本I较少，因此最小经验误差与最小期望误差的差距较大，如图1所示，研究小样本学习就是要使在训练样本I较少时，和的差距仍尽可能的小。

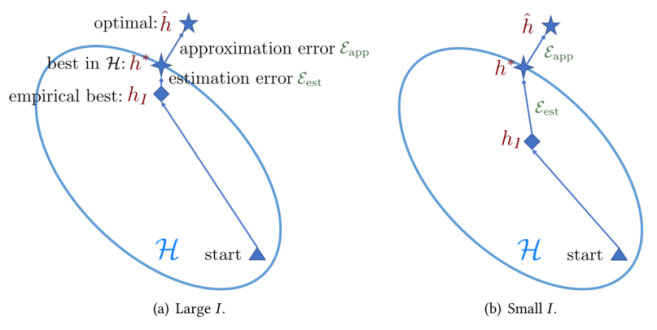


图1 - 大规模数据集与小规模数据集的误差可视图

小样本学习主要利用先验知识辅助模型的训练，这些先验知识可以通过以下三个方面得到应用：

（1）数据

如图2所示，利用数据增强的方式扩充数据集，将样本数量扩充至，确保，以此提升最终的模型训练结果。

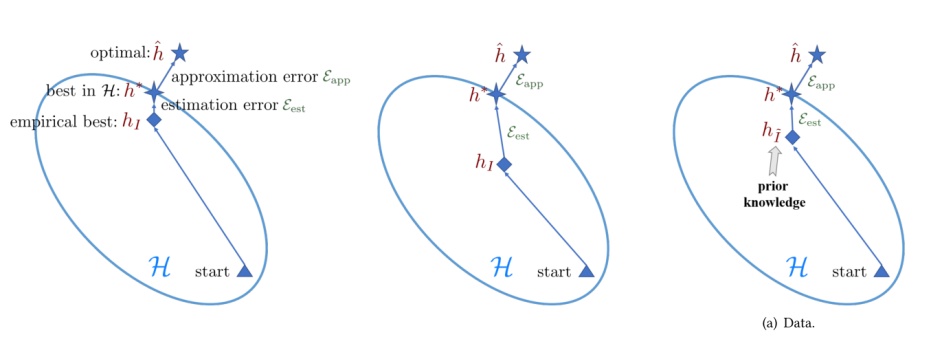


图2 - 利用数据增强解决小样本学习问题的误差可视图

（2）模型

如图3所示，通过先验知识限制函数空间H，使优化搜索的范围缩小至，从而使搜索所需的成本降低，从而提升小样本学习的最终性能。

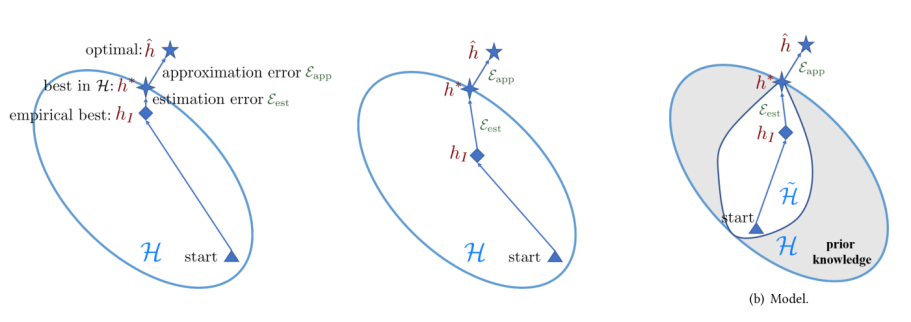


图3 - 通过限制函数空间解决小样本学习问题的误差可视图

（3）算法

如图4所示，通过优化学习策略，如优化搜索的初始起点、提升每次迭代的准确性等，使模型能更快地收敛至最优解，从而提升小样本学习的性能。

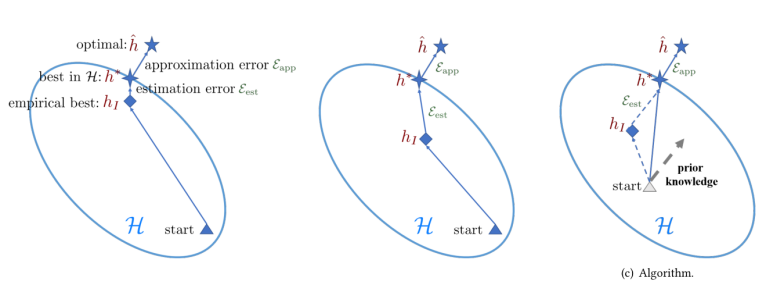


图4 - 通过优化学习策略解决小样本学习问题的误差可视图

**四、总结**

小样本学习是一种非常实用的机器学习方法，在当前的流行认知里，最优的机器学习模型是大模型+大数据集的组合，虽然这种简单粗暴的做法确实行之有效，但是却有着许多隐患。

单从数据层面来讲，大规模数据集的条件其实是一个较为苛刻的条件，受限于客观条件和成本，我们很难获取大量的真实数据。以计算机视觉为例，如果解决简单的RGB图片的猫狗分类问题，我们获取数据是相对较为容易的，但是如果希望解决RGB图片的大熊猫东北虎分类问题，获取数据的难度就上升了，如果希望解决RGB图片的目标分割问题，真实标签的获取难度又上升了，如果希望解决非RGB数据的视觉问题如Depth数据，数据的获取就更加困难了。

如果说不考虑成本，我们将数据集的构建看做复杂的工程问题，通过耗费极大的人力物力去扩充数据，且不说其造成的资源浪费，即使有着付出巨大代价的决心，有时大规模数据的获取是从理论上就不能实现的一件事，例如对手机用户人脸的检测，厂商不可能要求用户提供上千张人脸图片以实现手机对其人脸的面部识别。因此，小样本学习是有着天然的应用土壤的。

目前，基于人工智能的计算机视觉正向多领域进行扩展，面向各种任务、各种传感器的人工智能算法有待进一步研究，当我们面临数据集不足的问题时，考虑使用小样本学习的技术将是一个很好的研究方向。