**机器学习必然需要大量数据吗**

刘逸川

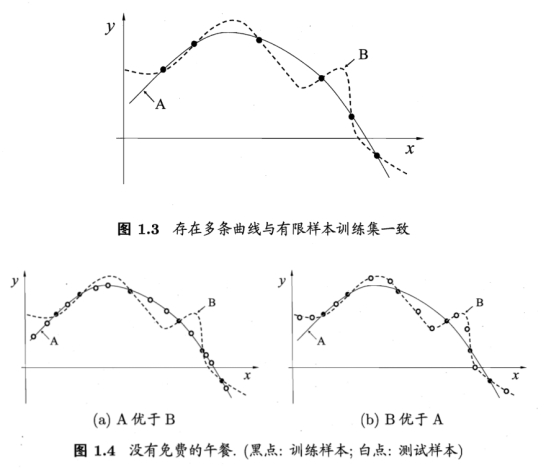
2020-07-18

目前的机器学习大多需要大量数据，相比而言，人类接收少量数据就能达到同样的效果，因此机器常被批评不够聪明。这种批评有一定道理，但是存在两点问题。首先，批评者忽视了**人类在漫长的进化过程中学习过大量数据**，并且以基因的形式将学习结果传递至今。人类出生时并不是一块白板，大脑已经具备了基本结构和学习潜力，所以人类只需要少量数据其实是曾经的大量数据训练的结果。其次，**人类在成长过程中也经历了大量训练**，例如某人从小到大虽然只见过几百只猫，但由于人眼相当于高速摄像机，不间断地拍摄了大量图片，所以其实接收了非常多的训练数据。

当然，不可否认大脑是非常巧妙的。也许有人会说进化过程中的数据和目前的数据差异巨大，不具有足够的价值，但是大脑事实上是通过大量数据学会了如何学习，从而具有很强的迁移能力。此外，大脑也具有很强的想象能力，例如马航MH370失联后，搜索人员希望通过分析南印度洋的卫星图像寻找可能的飞机残骸，但是发现计算机识别的准确率极低。相比而言，人类的分辨能力就要强得多，因为虽然飞机残骸非常不规则，但是人类能够通过飞机的样子想象出飞机的各种残骸的样子。1除了迁移、想象能力，大脑还有很多非常优越的能力。

大脑的这些能力使得它只需要少量数据，如果人类无法直接设计出和大脑同等巧妙的机器，那么只能通过大量训练来弥补，**要想既不需要完善的结构又只需要少量数据是不可能的**，这其实是另一个版本的“没有免费午餐”定理（No Free Lunch Theorem，简称NFL定理）。

NFL定理是机器学习中的一个著名定理，指的是抛开实际情况，任何算法的期望性能都和随机猜测没有区别。例如下面这个例子2：



A、B曲线都能完全拟合训练集，但是真实情况可能是A也可能是B，或者二者都不是，因此不同算法的期望性能是相同的。这是在训练样本数量有限的情况下必然会出现的问题。

**如果希望找到更确切的拟合曲线，要么增加训练样本，要么增加外部信息。**外部信息是指来自实际情况的助于判断的信息，例如大脑的迁移能力就是一种获得外部信息的能力，大脑通过类似的例子来帮助判断。在上面这个例子中，我们倾向于认为平滑的A曲线能够更好地拟合真实情况，这源于我们的经验，经验就是一种外部信息。近些年兴起的小样本学习（Few-shot Learning）的基本思路就是加入外部信息，通过已有知识的迁移来增强泛化能力。

总而言之，机器学习不一定需要大量数据，但是要取得良好效果必须加入外部信息，或者说使机器拥有额外的判断能力。机器需要大量数据来学习是正常的，人类也曾经在进化过程中这么做过，但是人类作为“过来人”应该指导机器——设计更巧妙的硬件、编写更巧妙的程序，使得机器不需要太多的数据，这其实是人类在为机器提供外部信息。最后，值得一提的是，小样本学习和大数据并不矛盾，大数据相当于统计数据，对于社会管理、资源调度、决策等等都是非常重要的，具有少量数据无法替代的作用。

**参考资料**

1 [美]皮埃罗·斯加鲁菲.《智能的本质:人工智能与机器人领域的64个大问题》. 人民邮电出版社. 2017. 140页

2 周志华.《机器学习》. 清华大学出版社. 2016. 7-9页