机器学习面临的挑战

张长水

清华大学自动化系, 智能科学与技术国家重点实验室, 清华信息科学与技术国家实验室 (筹), 北京 100084

E-mail: zcs@mail.tsinghua.edu.cn

收稿日期: 2013-06-20; 接受日期: 2013-11-13

国家重点基础研究发展计划 (批准号: 2013CB329503) 和国家自然科学基金 (批准号: 91120301, 61075004) 资助项目

摘要 该文讨论了机器学习目前面临的几个挑战, 包括: 高维特征空间和数据量问题, 大数据量的计算困难, 寻求最优解的困难和可解释性差等问题. 然后针对当前很多人关心的几个重要问题, 例如大数据问题, 深度学习, 概率图模型等做了分析, 以引起深入思考.

关键词 机器学习 深度学习 概率图模型 大数据 优化问题

1 引言

过去的二十多年的时间里, 机器学习的研究得到了快速发展, 取得了一系列令人瞩目的研究成果, 同时也面临很多的挑战. 我们需要对机器学习的现状和面临的挑战进行探讨和梳理, 以引起更深入的思考.

2 机器学习面临的挑战

虽然机器学习取得了令人瞩目的成果, 但是仍然面临很多困难和问题. 下面列举其中的一些问题.

2.1 高维特征空间与样本数的不足

在很多实际应用问题中, 得到的特征维数是比较高的, 有的甚至是非常高的. 例如, 在图像识别中如果提取 SIFT 特征, 特征维数是 128 维, 如果提取其他特征, 维数也往往是几十维, 或者几百维. 还有, 在文本分类问题中, 如果把每一个单词当作一个特征, 特征的维数可能是几千维, 或者上万维, 这依赖于所使用的字典大小.

下面以概率密度函数的估计为例讨论特征维数和所需要的样本之间的关系. 对于一维的概率密度函数估计来说, 通常在具有几十个以上的样本时可以得到一个比较好的函数估计, 这是因为在每一个点附近应该有一定量的样本 (也就是说, 数据要具有一定的密度) 才能得到好的估计结果. 我们假设至少需要 10 个样本, 这是一个保守的数字. 如果要估计的是一个二维概率密度函数, 就需要把二维空间等分成一些小网格, 需要每一个小网格中有足够多的样本. 因此, 可能需要 $10^2 = 100$ 个样本. 当维数增加的时候, 空间中小的格子的数量随维数的增加而指数上升. 假设维数为 d , 需要的样本数是 10^d .

引用格式: 张长水. 机器学习面临的挑战. 中国科学: 信息科学, 2013, 43: 1612–1623, doi: 10.1360/N112013-00074

按照这种方法计算, 在图像识别问题中, 至少需要 10^{100} 个样本, 而在文本分类中, 需要 10^{1000} 个以上的样本. 可我们知道, 这是不可能的一件事情.

由于样本数不足导致了高维数据学习的困难, 这个现象被称作“维数灾难”. 维数灾难的核心问题是, 高维函数事实上远比低维函数复杂, 而我们对其还没有有效的方法进行分析.

利用具体分类问题的先验知识, 或者利用得到的训练数据, 可能在一定程度上减少维数灾难带来的困难. 例如, 如果知道两组特征 x 和 y 之间是独立的, 那么就有

$$p(x, y) = p(x)p(y). \quad (1)$$

因此, 对于 $p(x, y)$ 的估计就可以通过分别对 $p(x)$ 和 $p(y)$ 估计来完成. 由于单独的 x 或 y 的维数要小于它们联合在一起的 (x, y) 的维数, 因此, 所需要的样本数就会少很多. 概率图模型^[1]研究的就是如何利用随机变量之间的条件独立性对问题建模、学习、推理的理论、算法和应用. 概率图模型在解决高维样本学习方面有很多贡献.

另外, 对实际问题中的数据分析表明, 大量的实际问题的高维数据实际上嵌入在一个低维的流形上^[2,3], 也就是说, 数据并没有充满整个高维空间. 其主要原因就是各个特征之间存在很强的相关性. 因此, 实际上并不需要那么多的数据来估计概率密度函数. 尽管如此, 目前的研究表明, 图像数据的本质维数至少有几十维, 这对样本数的需求还是非常大的.

上面讨论的维数对样本量的需求是以非参数概率密度函数估计为例. 实际上, 维数灾难不仅仅出现在概率密度函数的估计中, 也存在于其他的学习问题中. 上面谈到的特征独立和流形分布的知识同样有助于缓解在其他学习问题中遇到的维数灾难困难.

2.2 寻找最优解的困难

目前很多的机器学习研究沿着这样的思路进行: 把要解决的问题形式化为一个目标函数, 然后通过优化这个目标函数达到对数据学习的目的. 例如: 支持向量机就是把要解决的两类分类问题形式化为最小化如下目标函数:

$$f(w) = \frac{1}{2}w^T w + C \sum_{i=1}^N L(w, x_i, y_i). \quad (2)$$

其中 $x_i, y_i (i = 1, \dots, N)$ 是样本的特征和标签, N 是样本个数, w 是线性分类器的权向量, L 是损失函数, C 是折衷系数. 当 L 是二次, 或者一次函数时, 函数 f 是一个凸函数, 存在一个极值, 可以通过凸优化算法寻优.

我们要解决的实际问题非常复杂, 将其形式化后的目标函数也非常复杂, 往往在目前还不存在一个有效的算法能找到这样复杂目标函数的最优值. 因此, 需要通过一些优化技术寻找次优值. 这样做通常有两个方面的问题存在, 一个是所使用的优化技术可能非常复杂费时, 效率很低; 另一个是得到的局部极值可能距离我们希望的解很远, 不能满足要求.

机器学习领域中有一些优化问题具有一些特殊性. 因此, 有一些研究工作集中在如何求解这些优化问题以及分析所得到的解的性能分析. 但是, 为了解决机器学习领域的问题, 研究针对性的有效优化算法是必要的.

由于求解全局最优是非常困难的, 所以, 通常人们只是采用简单的局部极值求解算法, 例如梯度下降方法. 采用局部极值算法时, 当算法的初值不同, 得到的解也不同. 而要优化的函数往往有非常多

(可能成千上万, 或者更多) 的局部极值, 通过算法得到的解只是其中一个. 我们会关心下面的问题: 这么多的局部极值都是我们所希望的解吗? 如果不是, 其中哪些局部极值是? 如何得到这些解?

另外, 在对要解决的问题建模时, 目标函数有时候只是一种“近似”的建模. 例如: 把要解决的问题形式化成优化下面的函数:

$$f(w) = L(w) + C \times r(w). \quad (3)$$

其中 $L(w)$ 是损失函数, $r(w)$ 是正则项, C 是折衷系数. 目前经常使用的正则项有很多, 例如: 光滑性正则函数, 稀疏性正则函数, 函数复杂性正则函数. 光滑性正则函数假设函数具有光滑性质; 稀疏性正则函数假设要学习的模型具有稀疏性; 函数复杂性正则函数则要求要学习的函数不能太复杂. 这些正则函数的使用基础是假定所要研究的问题满足这样的条件. 但是实际问题是否满足, 在什么程度上满足这样的条件, 我们并不知道. 目标函数的这种“近似”性质, 需要考虑下面这些问题, 一定需要求解目标函数的全局最优值吗? 局部极值 (全部局部极值都) 能满足要求吗?

2.3 可解释性差

从上文的讨论中可知, 机器学习领域里要解决的问题很难用一个容易优化的函数来建模. 对于一个实际问题, 可以构建不同的目标函数来表示要解决的同一个问题. 这样, 也就会得到很多不同的解.

机器学习的另一个问题是得到的模型的可解释性差. 可解释性是和实际应用密切相关的. 机器学习在解决一些具体的问题时, 需要领域的专家能够理解模型, 能够理解“为什么”这个模型能够对未知样本做预测. 例如, 在分子生物学的蛋白质结构分析中, 一个只在输入输出意义下对数据是正确的模型对于分子生物学家还远远不够. 他们需要获得的模型在分子生物学意义下可解释. 同时, 由于所提供的训练数据是不充分的, 机器学习还需要为他们提供进一步工作的线索和可能. 他们需要的是可解释的数据理解工具或工具集.

机器学习应该考虑模型对问题世界的可解释性. 机器学习一直以来所遵循的“输入输出满足”原则对某些问题而言可能是远远不够了. 正如上文所讨论的, 在求解机器学习问题时可能得到多个解, 如果使用“输入输出满足”原则, 可能建立的多个模型获得多个解, 则需要以对问题世界可解释性来分辨其优劣.

大数据时代一书^[4]强调了在大数据时代相关关系的发现和使用更为重要, 而不应该那么追求因果关系. 我认为, 在某些商业领域他们这样做是对的. 但是当我们关注科学问题时, 情况就会不同. 寻根溯源, 或者说追求因果关系是科学研究的一个动力.

关于因果关系和相关关系, 马颂德老师给了意见: “因果关系是一个相对的概念”. 对此的一个解释是: “牛顿看到苹果掉在地上, 发现了万有引力定理, 可以说发现了苹果掉在地上的因果关系. 但也可以说, 这是个相关关系, 因为它没有说明万物之间为什么有引力.”

可以说, 大数据时代一书更强调相关关系, 而我们的科学研究更强调因果性. 就机器学习而言, 因不同的应用问题不同, 对因果关系的需求程度也是不同的. 对于更商业化的应用问题, 即在输入输出意义下对数据是正确的, 可预测性非常重要. 而对于更基础的科学研究问题而言, 可解释性就更为重要.

2.4 大数据量的计算

这里讨论的是数据量这样一个问题, 而不是大数据时代一书中谈到的大数据问题. 下文会讨论大数据问题.

数据量大是机器学习在应用阶段面临的重要问题. 实际上, 在机器学习领域近些年一直关注这个问题, 被称之为“大规模数据的学习” (large scale data learning, 或 big learning).

在过去的十几年中, 人们关注的更多的是好的机器学习算法的设计, 学习算法的性能分析等, 我们统称为学习理论和学习方法. 这是因为当时有太多的问题需要研究和解决, 而机器学习的突飞猛进, 吸引了大部分研究人员的注意力, 很多人沉浸在机器学习的理论方法研究的喜悦中. 而当学习理论和学习方法都已经建立, 几个有代表性的学习算法在实际问题中成功应用后, 大规模数据的学习开始成为了一个受到关注的问题.

大规模数据的学习之所以在近几年才开始受到关注, 主要是因为实际中数据量很大. 而十几年来发展起来的很多学习算法面临的一个尴尬就是: 都很难用于大数据的学习, 主要的问题是时间复杂性和空间复杂性. 例如: 当训练数据超过 10000 时, 支持向量机算法代码 (libsvm) 因为内存不够而无法在一台普通的台式机上运行, 即使扩大内存后, 也需要几个小时才能完成训练. 因此, 不能想象训练数据是十万、百万量级下的支持向量机的学习 (而 libsvm 计算复杂度是样本量的平方). 类似的情况也出现在其他的一些机器学习算法中, 如: EM 算法、主成分分析、线性判别、聚类算法. 因此, 出现了一些工作来解决这个问题. 其基本思路有下面几个:

- 快速算法. 即与原算法等价的快速算法, 类似于为傅里叶变换找到快速傅里叶变换的工作. 这样的工作无疑是重要的, 但是其存在加速极限.

- 近似算法. 不是所有的算法都可以很容易找到和原算法等价的非常快速的算法. 因为要求速度快, 因此就考虑在牺牲很少的精确性的情况下寻找非常快速的 (例如线性的, 或者是亚线性的) 算法. 在这里牺牲精确度是很有必要. 机器学习很多年以来一直在追求学习模型的泛化能力, 但是, 模型的泛化能力取决于很多因素, 而不仅仅是学习算法. 当样本数很大的时候, 其带来的泛化能力的提高往往更有意义. 而对于一些聚类问题而言, 聚类问题通常只是用于所研究问题的一个阶段, 其精度并没有那么重要. 因此, 这里牺牲精度就是一个自然的选择. 当然, 近似算法的设计也是存在加速极限.

- 增量学习和在线学习 (incremental learning, online learning). 这两种学习方式是在一些应用条件具备的情况下进行大数据量学习的一个解决方案. 当需要解决的实际问题的数据是序贯到达的, 就可以考虑增量学习和在线学习. 增量学习和在线学习考虑的是, 当新的样本到达时, 是否可以已经学习到的模型进行小的调整而达到学习的目的. 这里的小意味着少的计算量. 和增量学习一个重要的不同在于, 在线学习要求的“小的调整”能够实时完成. 有些算法本身的性质决定了可以比较容易设计出其增量学习算法, 例如: 感知器学习算法、EM 算法等. 而有的算法则难度更大些, 例如: 支持向量机、主成分分析. 增量学习和在线学习近些年受到重视, 这方面有些优秀的工作发表.

- 并行算法. 利用更多的计算资源以解决大数据的学习问题. 人们考虑过把机器学习算法进行各种并行方案的执行. 例如: 单机多核并行、GPU 方案、多服务器方案等. 一些大的计算机公司更关注这方面的工作. 例如: 有计算机公司曾设法把主成分分析、高斯混合模型的学习算法在 1 万台服务器上并行实现. 这个的工作思路无疑是诱人的. 但是由于通讯开销, 各个计算单元之间的同步等问题的存在, 并行算法这个方向的工作并没有那么理想的进展. 其并行算法的计算能力并不能和所利用的服务器数量呈线性关系, 甚至亚线性关系也达不到. 更多的情况是, 当服务器达到一定数量后, 计算能力几乎不再增长, 见图 1. 当然, 已经有研究者着手解决这个问题.

在大规模数据的计算方面有一些很好的研究工作发表. 有些研究工作解决的问题非常吸引人, 例如: 如果训练数据不能一次放到内存怎么办? 有些算法的加速结果很诱人. 例如: 把最大间隔聚类算法的原算法 (计算复杂度 $O(n^7)$, n 是样本数) 加速到 $O(sn)^{[5]}$, s 是高维数据的稀疏度.

图 1 当前的机器学习算法能够学习的数据量和并行机器数量的关系. 横轴表示并行机器的数据, 纵轴表示可以处理的数据量. 理想情况下, 这是一个线性关系 $x=y$, 但实际上是另一曲线, 即机器数目到达一定量后, 可以处理的数据量几乎不再增长

Figure 1 Relation between data handling capacity and number of parallel unit in a parallel machine learning system recently. In ideal conditions, it follows a linear relationship $x = y$, but actually the data-handling capacity hardly increases when parallel units achieve a certain number

值得说明的是, 并非数据量一定要非常大才叫做大数据的学习问题. 从算法角度看, 只要数据量大, 解空间 (模型空间、参数空间) 就极其大. 因此, 做机器学习研究的人对大数据的体会更深, 压力更大.

3 几个重要问题

除了上面讨论的机器学习面临的挑战外, 下面一些问题也很重要.

3.1 大数据

大数据是当前一个热点问题 [4]. 大数据涉及很多方面的研究, 这包括: 数据的获取、传输、存储、分析等. 这里主要讨论大数据给机器学习的挑战和机遇.

这里的所说的大数据主要是针对由于互联网等技术的发展而出现的大数据问题, 而不仅仅是指数据量比较大 (大数据量的学习已经在前面讨论过了). 这里的“大数据”一词代表了: 数据多, 不够精确, 数据混杂, 自然产生等特点, 这些都在文献 [4] 中做了总结.

大数据给机器学习带来的问题不仅仅是因为数据量大而计算上非常困难, 其带来的更大的困难在于: 数据可能是在不同的服务器上获取的, 这些分布在不同服务器上的数据之间存在某些联系, 但是基本上不满足同分布的假设, 而我們也不可能把所有数据集中起来进行学习和学习. 经典的机器学习理论和算法要求数据是独立同分布的. 当这个条件不满足时, 这时我们的学习模型和学习算法怎么办? 是修改算法从这些数据中学习, 还是整理数据以适应目前的学习算法? 这些服务器上的数据之间的关系如何建模和分析? 另外, 我们已经知道, 在网络上获取的很多数据的分布通常会随着时间的推移发生变化 (称之为演化数据, 在网络的论坛中称之为概念漂移), 这时我们的学习模型和学习算法怎么办?

在数据分布发生变化时, 数据的独立同分布的假设也不再满足, 这时还有什么数学性质可以满足? 如果不满足任何的数据性质或者可以利用的数学性质很少, 其依赖的数学理论是什么? 如何确定给出的模型和算法是可靠的, 而不仅仅是实验室里的算法游戏呢?

大数据除了给机器学习带来了计算上的困难和挑战外, 也带来了一些好处. 其中一个好处体现在数据多了以后, 呈现出小数据情况下没有呈现出的现象, 这被称之为“涌现”(emergence). 实际上, 1990年后曾经有几年人们很集中的研究过这个问题. 人们发现: “微观”地了解每一个个体, 并不能预测“宏观”的整体行为. 例如: 我们知道每个水分子的运动规律和运动方程, 但是你不知道水的沸腾是什么样子, 沸腾的水是“涌现”出来的现象. 在应用领域, 研究人员曾经仅仅使用三条规则来描述一只鸟的飞行. 这样当一群鸟的每一个个体都仅仅遵循这三条规则飞行时, 就“涌现”出看到过的鸟在天空翱翔的景象. “涌现”一词很生动表达了这一含义. 人们也做过很多类似的实验证明了这一点. 因此, 出现了被称之为群体智能 (collective intelligence, wisdom of the crowd) 这样的术语. 当然, 机器学习研究领域对此研究不多.

大数据的另一个好处是: 在某些应用条件下, 数据变得稠密了. 多年以来, 因为很多众所周知的原因机器学习一直在研究小样本的学习问题. 在实际中, 分类器性能不够好的一个原因就是样本太少. 理论上, 我们知道在样本数趋于无穷的时候, 很多算法具有很多良好的性质. 实践中也有这样的体会, 当样本数很多时, 使用简单的模型往往能够取得好的泛化性能. 而在大数据时代, 当样本数量很大的时候, 在样本空间的某些区域会出现稠密的现象, 这些稠密的数据给分类器设计实际上提供了很多的信息. 因此, 在这些局部稠密区域, 分类器的性能有可能接近理论上的极限性能.

大数据的再一个好处是: 大数据使得样本空间原来“空旷”的区域出现了样本, 原来“稀疏”的区域变得不再稀疏, 这在很大程度上为提高分类器性能提供了很好的数据基础. 直观地说, 就是数据本身的多样性能够更多的展现出来. 例如: 在语音识别问题中, 大数据情况下, 人们各种的发音习惯才能更多地体现出来; 在图像识别中, 大数据情况下, 物体在不同情况 (变形、光照、背景等变化) 下的外观表现才更丰富. 而这些数据的缺失很难通过建模和学习算法弥补, 同时, 这些数据也很难 (可以说是不可可能) 通过专家、算法设计人员的设计来获取. 因此数据产生的自发性就很重要.

正是基于上面的原因, 很多从事语音识别的研究人员希望在尽可能多的数据上进行训练: 从几个小时, 到几十个小时, 到几百个小时, 乃至几万个小时的语音数据; 计算机视觉的研究人员也在尽可能多的收集和标注数据: 从几万, 到几十万, 到 Fei-Fei Li 的八千万图像数据^[6], 到几百亿乃至几千亿的语音数据. 八千万图像的 ImageNet 的建立是一个了不起的工作, 然而这些数据对于计算机视觉的任务还远远不够. 而事实上, Hinton 在使用 ImageNet 图像进行物体识别^[7]训练时, 把每张图像进行了很多微小的变换 (旋转、光照变化等) 从而生成了比原图像多几倍, 十几倍的训练数据情况下, 识别率又提高了几个百分点. 这充分说明了我们的训练数据在通常情况下还远远不够.

大数据时代数据的自发性导致了数据本身的不精确性. 不精确意味着数据有错误. 和传统的精确标注的数据相比, 不精确是一个大问题. 而实际上, 对于不精确性的一个补偿就是大量的数据, 由于数据量的巨大, 这一问题变得没有那么严重, 因为其中还有很多高质量的数据.

在机器学习领域几乎没有对这种数据的不精确性做过工作. 可能是因为统计机器学习方法已经对噪声进行了建模, 这噪声也可以包含数据的不精确性. 另外, 针对某些实际应用中的不精确性很难建模, 所以分析算法的性能就太困难.

相比机器学习领域, 数据挖掘领域对此有过一些研究工作. 这些研究工作讨论了当标注数据存在错误时, 是否能够构建好的分类器. 基本结论是: 当大部分标注数据是正确的时候, 少数 (小于 50%)

的错误标注样本对分类器的影响不大, 错误标注数据的比例越小, 分类器的准确性越高^[8]. 当然, 如果大部分标注样本存在错误时怎么办? 对于某些具体的应用问题, 当数据量很大的时候, 可以有针对性地设计算法解决这个问题. 我们曾经考虑一种特殊情况, 在这种情况下, 即使大部分标注数据是错误的, 仍然可以设计出好的分类器^[9].

看起来不精确性对机器学习是个不利的方面, 而实际上, 它并非全是坏处. 例如: 在互联网上搜索时, 百度和 google 会给出一些检索结果. 当用户看到这些结果时, 会点击其中的一条, 然后也许会再点击另外一条. 这些点击信息告诉我们, 用户点击的第一条可能不是他要找的网页, 或者第二条让他更满意. 这些点击信息虽然没有明确做好网页和查询数据之间的标注, 但是告诉我们一些可能的信息. 这样的数据如果足够多, 就有利用价值. 一些研究组报告说, 这些“弱标注”数据给他们的系统提供了很多有用的信息, 而改进了他们的系统.

混杂性是大数据的另一个特性, 是因为数据是自发生成的而带来的一个特性. 混杂性给我们提出的一个课题就是要把这些数据进行分离和整理, 从而为进一步的机器学习做准备. 这个过程是数据挖掘要完成的任务.

3.2 深度学习

在上个世纪八十年代和九十年代, 反向传播算法 (BP 算法) 的出现使得人工神经网络的研究东山再起, 得到了很大的重视和快速发展. 然而经过几年的快速发展后, 又迅速进入研究的低谷.

2006 年 Hilton 发现了深层神经网络的学习算法^[10]. 这之后, 人工神经网络又逐渐得到了重视, 成为近年来的研究热点.

人工神经网络的这次东山又起, 以一个新的面貌出现: 深度学习. 这里的深度指网络的层数多. 二十年前的神经网络研究的大多是三层的神经网络: 一个输入层, 一个隐含层, 一个输出层. 反向传播算法的出现让多层神经网络的学习成为可能. 当时出现了很多令人振奋的研究成果, 其中一个就是关于多层神经网络的表达能力的结果^[11]: 只含有一个隐层的前馈网络是一个通用的函数逼近器, 在一定条件下, 它可以逼近任何一个连续函数. 这个结果说明有一个隐层的前馈网络对于表达非线性函数来说已经足够, 当然这并不说明一个隐层是最好的. 也有研究表明, 要逼近相同输入维数的函数, 两个隐层的网路可能比单个隐层的网络所需隐单元数要少得多. 尽管如此, 因学习算法不令人满意, 很少有人使用两层以上的神经网络结构. 大量的研究表明, 反向传播算法严重过学习 (overfitting). 毫无疑问, 反向传播算法是一个贪婪优化算法, 其收敛到目标函数的一个局部极值. 而目标函数是一个非凸的复杂的目标函数, 存在大量的局部极值. 看起来, 其中很多的局部极值不是我们想要的结果, 而我们又无法提前选择一个好的初值通过反向传播算法得到理想的局部极值. 样本量太少也是导致神经网络严重过学习的一个重要原因. 当时人们没有使用那么多样本有很多原因. 一个是获取足够的样本的代价太高, 另外, 计算能力的不足也限制了人们对于大样本量学习的探索. 因此, 探讨一个以上的隐含层的学习的研究工作非常少, 发表的一些研究工作也不让人乐观.

而这次深度学习的出现是通过逐层学习的方式解决了多个隐含层神经网络的初值选择问题. 图 2 给出的是一个具有三个隐含层的逐层监督学习示意图. 不仅如此, 研究工作表明, 非监督数据在多个隐含层神经网络的初值学习中也起到了很重要的作用. 这是一个很有意思的结果. 在监督学习中, 需要具有样本标号的监督数据. 然而, 获取大量标注的样本的代价过于昂贵. 例如: 语音信号数据库、图像数据库的建立工作都说明了这一点. 但是, 如果不要求数据是监督的, 其数据获取代价就小得多. 例如: 获取大量的语音信号较为容易, 因为每个人每天要说很多话, 也可以从电视、广播中得到这些语音

图 2 具有三个隐含层的逐层监督学习示意图, 每一个方块表示一个隐含层. 逐层的学习过程是: 从输入到临时的输出节点 y_1 构成一个单隐含层网络, 学习该隐含层; 然后以该隐含层样本为输入, 把下一层的隐含层和临时输出 y_2 构成一个单隐含层网络, 学习该第二个隐含层. 这样一直下去直到所有隐含层的学习结束

Figure 2 Sketch map of a layer-by-layer supervised learning process with 3 layers, in which each rectangle represents a hidden layer. The learning process is as follow: the input node and temple output node y_1 form a single hidden layer network. We learn this layer and use its output as input for the next hidden layer and its temple output node y_2 , which can also be seen as a single hidden network. We can go on learn in this way until we complete each hidden layer

信号. 另外也可以在网络上收集大量的图像, 或者通过手机、照相机拍摄大量的图像. 非监督数据也可以用来通过类似监督的逐层学习方式解决多个隐含层神经网络的初值选择问题.

仔细研究可以知道, 深层网络的逐层学习方法是个非常“粗糙”的学习方法, 这其中有太多的近似. 不过, 这个学习过程给我们一些启示: 虽然这是个非常近似的算法, 但是却能得到如此好的泛化能力. 这是什么原因?

另外, 近些年深度学习的研究表明, 深层网络的隐含节点学习到的是所研究问题的特征. 这和机器学习、模式识别之前的研究非常不同. 传统的研究中, 特征提取都是领域专家的工作. 从输入信号中经过怎样的变换来提取特征通常需要很多的领域知识. 而在深度学习中, 学习到的网络的隐含节点对应于从输入信号中经过变换后的特征, 并且这些特征是逐层越来越抽象的. 图 3 是以人脸识别为例, 深层网络隐含节点所学习到的特征的示意图. 在网络的输入层是每一个图像块, 每个像素对应于一个神经元. 在第一个隐含层学习到的是一些非常简单的模式, 即一些带有方向性的各种边缘, 这些模式与计算机视觉多年研究得到的视觉底层特征很类似. 在更高层的隐含节点对应于更高一层, 更抽象的特征. 到最高层, 对应于图像的高级语义特征. 而网络结构逐层学习到更为抽象的特征, 和神经科学中对视神经研究得到的结果也很吻合. 因此, 深度网络学习的不仅仅是分类器, 而且包含了需要的特征.

让深度学习成为热点的另一个重要原因是其在实际应用方面的贡献. 语音识别是一个经过多年研究的课题. 在使用深度学习方法之前的若干年中, 其识别性能没有大的提高. 而在使用了深度学习方法后, 其性能有了一个飞跃 (见表 1). 表 1 中 GMM-HMM 是使用深度网络之前的最好的算法, DNN-HMM 是使用深度网络的算法. 可以看到在这两个测试中, 错误率有了大幅度降低. 这大大推进了语音识别的产业化进程. 到目前为止, 一些最好的语音识别系统都采用了深度学习技术, 这包括微软、谷歌、百度等公司的语音识别系统.

下面再以图像目标识别为例看深度学习的贡献. 在使用 ImageNet 图像数据库进行的图像中的物体识别竞赛中, 在 2010 和 2011 年最好的系统识别率分别为 72% 和 74%. Hinton 带领的研究小组使用了深度学习技术, 2012 年获得了竞赛的冠军, 其识别率为 85%. 由此可以看到深度学习的贡献.

目前, 对深度学习还存在不同的意见. 主要存在下面的批评, 一个是深度学习没有理论. 机器学习领域研究人员非常关心深度学习的理论进展. 人们很想知道深度神经网络为什么可以具有这样好的性能? 其理论依据是什么? 如果从统计学习的角度看, 其泛化能力如何? 这是深度学习目前需要研究和解决的问题.

对深度学习的另一个批评是深度学习中需要太多的经验和人工尝试和技巧. 和传统的神经网络的研究类似, 深度学习中需要确定网络的层数, 每层的节点数, 节点的激发函数等因素. 因此, 不容易掌

图 3 具深层网络隐含节点特征学习示意图 (以人脸识别为例, 修改自文献 [12])

Figure 3 Features learned by hidden nodes in deep network for face recognition, modified from [12]

表 1 两种语音识别系统在两个数据集上的错误率比较

Table 1 Contrast of two speech recognition system on two different datasets

Task	Size of training data (h)	DNN-HMM	GMM-HMM
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6

握和使用.

此外, 大规模数据的学习对于深度学习也是一个不小的挑战. 目前要解决的图像识别和语音识别的学习数据往往几十万、几百万以上的样本量, 利用通常的计算机运行深度学习方法学习一遍数据通常需要几天, 或者是几十天的时间. 这完全不能满足人们的需要. 因此, 一些研究人员做了一些深度学习加速算法的工作.

3.3 概率图模型

概率图模型是对随机变量之间的条件独立性建模的工具. 概率图模型一直受到关注, 只是 Kollar 的书的出版, 让概率图模型变得更受重视.

概率图模型的一个贡献是让很多的模型有了更为简洁的表示, 因此, 我们对于很多模型有了更为直观和深刻的认识. 这包括了很多常见模型: 主成分分析、高斯混合模型、隐马尔科夫模型、独立成分分析等. 特别是在解决实际问题时, 从概率图模型的角度建模, 模型表示简单易理解, 其推理也变得容易理解. 因而, 在计算机视觉、语音识别、文本分析的文章中常常见到用概率图模型建模. 深层神经网络也同样可以用概率图模型表示和分析.

图 4 ImageNet 图像数据库部分图像¹⁾Figure 4 Partial images in ImageNet dataset¹⁾

因为对条件独立性建模, 当所研究的问题中很多变量之间满足条件独立性时, 采用概率图模型建模有助于缓解维数灾难带来的困难. 对此, 上文中已经讨论过.

和基于向量空间数据的建模相比, 概率图模型较好的结合了领域知识, 即随机变量 (特征) 之间的条件独立性. 机器学习通常把研究对象看作一个黑盒子, 这是因为在很多情况下不清楚研究对象的输入输出关系. 当对于这个输入输出关系有所认识时, 充分利用这些知识有利于对于问题的解决. 而概率图模型是充分利用了随机变量之间的条件独立关系这一知识.

概率图模型的一个主要问题在于推理的困难. 在精确推理时, 其算法计算复杂度是指数爆炸的. 因此, 一些近似推理方法被研究和提出. 深度神经网络可以用概率图模型来表示. 深度神经网络采用了逐层学习和推理的方式, 很显然, 这是一种近似, 但却取得了非常好的学习性能. 因此, 是否可以由此获得一些启示而采用类似的近似方法对概率图模型学习和推理?

4 结语

本文主要讨论的是目前机器学习面临的挑战和当前大家关心的几个问题. 其主要目的是希望能够

1) 修改自 <http://www.image-net.org/challenges/LSVRC/2010/index>

对当前的机器学习研究进行一些梳理, 对今后的研究有更多的思考.

参考文献

- 1 Kollar D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge: The MIT Press, 2009
- 2 Balasubramanian M, Schwartz E L. The isomap algorithm and topological stability. Science, 2002, 295: 7
- 3 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290: 2323–2326
- 4 Mayer-Schönberger V, Cukier K. Big Data. Sheng Y Y, Zhou T, translated. Hangzhou: Zhejiang People's Publishing House, 2013 [Mayer-Schönberger V, Cukier K. 大数据时代. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013]
- 5 Wang F, Zhao B, Zhang C. Linear time maximum margin clustering. IEEE Trans Neural Netw, 2010, 21: 319–332
- 6 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009. 248–255
- 7 Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: Proceedings of NIPS Conference, Lake Tahoe, 2012. 1106–1114
- 8 Sheng V S, Provost F, Ipeirotis P G. Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 614–622
- 9 Chen S, Zhang J, Chen G, et al. What if the irresponsible teachers are dominating? a method of training on samples and clustering on teachers. In: Proceedings of 24th AAAI Conference, Atlanta, 2010. 419–424
- 10 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
- 11 Yan P F, Zhang C S. Artificial Neural Networks and Evolutionary Computing. Beijing: Tsinghua University Press, 2000 [阎平凡, 张长水. 人工神经网络与模拟进化计算. 北京: 清华大学出版社, 2000]
- 12 Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of 26th Annual International Conference on Machine Learning. New York: ACM, 2009. 609–616

Challenges in machine learning

ZHANG ChangShui

Department of Automation, Tsinghua University, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China
E-mail: zcs@mail.tsinghua.edu.cn

Abstract In this paper, we describe and analyze the challenge of machine learning research, high dimensional data and the curse of dimensionality, the scalability of the machine learning algorithms, the difficulty of optimization, and the bad explicability. And we discuss a few hot topics: deep learning, big data, and probabilistic graphical models, etc.

Keywords machine learning, deep learning, probabilistic graphical model, big data, optimization

ZHANG ChangShui received his B.S. degree in mathematics from Peking University, Beijing, China, in 1986, and M.S. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, in 1989 and 1992, respectively. He joined the Department of Automation, Tsinghua University, in 1992, and is currently a Professor. He has authored more than 200 papers. His current research interests include pattern recognition and machine learning. He is currently an Associate Editor of the Pattern Recognition Journal. He is a member of the Standing Council of the Chinese Association of Artificial Intelligence.