1 关于机器学习的讨论*

王 珏
中国科学院自动化研究所复杂系统与智能科学实验室,北京 100080

1.1 引言

20 世纪 90 年代初,当时的美国副总统提出了一个重要的研究计划——国家信息基础设施计划(National Information Infrastructure, NII)。在这个计划的推动下,经过大批科学家与工程师的不懈努力,我们的生活与工作方式产生了重要的改变。这个计划的技术含义包含了四个方面的内容:

(1) 不分时间与地域,可以方便地获得信息。

(2) 不分时间与地域,可以有效地利用信息。

(3) 不分时间与地域,可以有效地利用硬软件资源。

(4) 保证信息安全。

经过十年的努力,集计算机科学与技术近四十年的积累,终于实现了以数字网络与浏览器为核心的技术,并做到了“不分时间与地域,可以方便地获得信息”。然而,其他三个目标,特别是“不分时间与地域,可以有效地利用信息”的目标远远未能实现。下面这个感叹,正是我们所面临的现实的写照:

E-mail,汗流浃背找寻信息的日子一去不复返,
无纸办公、轻松工作是我们的憧憬,
然而,我们失望了,
病毒的侵扰,我们提心吊胆,
垃圾的涌现,我们苦不堪言,
我们不安地注视着秘书移动的鼠标,
打印机吐出淹没我们的纸张,

* 本文得到国家重大基础研究项目“数字内容理解的理论与方法(2004CB318103)”的支持。

无纸办公成为嘲讽,轻松工作成为梦想,

我们开始怀念没有 *e-mail* 的时代,

我们开始忧虑进入烦恼的 *e-mail* 时代。

本文不准备讨论“硬软件有效利用”与“信息安全”的问题,而将讨论集中在解决“信息有效利用”的方法之上。“信息有效利用”问题的本质是,如何根据用户的特定需求从海量数据中建立模型或发现有用的知识。对计算机科学来说,这就是机器学习。

2001 年,美国航空航天局 JPL 实验室的科学家 Mjolsness 等人在 *Science* 上撰文指出:“每个科学领域的科学过程都有它自己的特点,但是,观察、创立假设、根据决定性实验或观察的检验、可理解检验的模型或理论,是各学科所共有的。对这个抽象的科学过程的每一个环节,机器学习都有相应的发展,我们相信它将导致科学方法中从假设生成、模型构造到决定性实验这些所有环节的合适的、部分的自动化。当前的机器学习研究在一些基本论题上正取得令人印象深刻的进展,我们预期机器学习研究在今后若干年中将有稳定的进展”[6]。

这个提法是令人惊讶的,大大超过了机器学习研究者的奢望,因为他们从未指望在科学研究的整个过程(观察、假设、实验、检验、模型或理论)中扮演如此重要的角色。

计算机科学,特别是人工智能的研究者一般公认 Simon 对学习的论述:“如果一个系统能够通过执行某个过程改进它的性能,这就是学习”。这是一个相当广泛的说明,其要点是“系统”,它涵盖了计算系统、控制系统以及人系统等,对这些不同系统的学习,显然属于不同的科学领域。即使计算系统,由于目标不同,也分为了“从有限观察概括特定问题世界模型的机器学习”、“发现观测数据中暗含的各种关系的数据分析”,以及“从观测数据挖掘有用知识的数据挖掘”等不同分支。由于这些分支发展的各种方法的共同目标都是“从大量无序的信息到简洁有序的知识”,因此,它们都可以理解为 Simon 意义下的“过程”,也就都是“学习”。本文将讨论限制在“从有限观察概括特定问题世界模型的机器学习”与“从有限观察发现观测数据中暗含的各种关系的数据分析”的方法上,并统称其为机器学习①。显然,相对 Simon 的论述,这只是相当狭义的学习。

我们描述机器学习如下:

令 W 是给定世界的有限或无限的所有观测对象的集合,由于我们观察能力的限制,我们只能获得这个世界的一个有限的子集 $Q \subset W$,称为样本集。机器学习就是根据这个样本集,推算这个世界的模型,使它对这个世界(尽可能地)为真。

---

① 根据 Simon 对“学习”的说明,“机器学习”只是“学习”的一个相当小的子集,本文只讨论计算系统的“学习”问题的一部分。由于讨论的问题比较明确,因此,与计算机科学的普遍说法一样,我们将在本文以后的讨论中,对上述两个概念不作区分。另外,由于本文对数据分析的讨论,只是根据当前应用的需要,建议从机器学习派生新的数据分析方法,因此,我们也将其归为机器学习。

这个描述隐含了三个需要解决的问题:

(1) 一致:假设世界 W 与样本集 Q 有相同的性质。例如,如果学习过程基于统计原理,独立同分布(i. i. d)就是一类一致条件。

(2) 划分:将样本集放到 n 维空间,寻找一个定义在这个空间上的决策分界面(等价关系),使得问题决定的不同对象分在不相交的区域。

(3) 泛化:泛化能力是这个模型对世界为真程度的指标。从有限样本集合,计算一个模型,使得这个指标最大(最小)。

这些问题对观测数据提出了相当严厉的条件,首先需要人们根据一致假设采集数据,由此构成机器学习算法需要的样本集;其次,需要寻找一个空间,表示这个问题;最后,模型的泛化指标需要满足一致假设,并能够指导算法设计。这些条件限制了机器学习的应用范围。

然而,机器学习的动机还是相当诱人的。如果我们可以发现一个通用的机器学习方法,并使用这种方法将观测数据变换为模型,这将一劳永逸地解决信息有效利用问题,这是我们计算机科学家一个不可挥去的理想,然而,这是一个可以无条件实现的理想吗?我们面临的现实是:观测数据性质非常复杂,共享同一个数据集合的用户往往具有不同的需求,换句话说,他们需要从数据集合中寻找不同的解答。这使得我们的理想变成了梦想。

目前,我们面临的观测数据与传统意义下的数据集合并不一样。在过去的日子里,数据一般是通过精心设计的试验,再仔细筛选而获得的,这些数据往往在统计上满足一定的条件(这在试验设计时就已经仔细考虑);而现在我们获得数据是涌现性的,最典型的是网络数据、生物数据以及经济金融数据,人们不能有效地控制这些数据的产生,只能被动地接受它们。这样,这些数据难以满足机器学习建模需要满足的条件,这就是我们的困难所在。事实上,在科学的发展史上,对付这类数据的方法就是数据分析,具体地说,不是使用这些数据直接建立模型,而是缩短数据长度,使得人可以阅读,从而利用人的智慧,洞察数据的内涵。如果需要建立模型,则需要在数据分析的基础上,根据人的洞察,重新设计试验,以从原始观测数据构造新的数据集合,从而,使其满足建模需要的条件 。

数据分析的历史可能远远比机器学习悠长,但是,至今没有一个精确并被人们所公认的定义。从任务上,我们可以粗略地描述如下:

令 Q 是给定世界的有限观测对象的集合,人们需要阅读这个数据集合,以便有所发现,但是,由于我们阅读能力的限制,我们必须将 Q 约简为满足这个限制的描述长度,以便人们可以洞察问题世界的本原。

这个描述暗示以下几个需要解决的问题:

---

② 事实上,在数据挖掘中,这是一类独立的任务,称为数据清洗。

(1) 需求:不同需求,对应解空间中的不同解答。

(2) 解空间:可以从 Q 暗示的解空间中,根据需求,方便地获得其中的任何解答。

(3) 约简:限定解答的长度。

(4) 例外:由于长度的限制,导致例外,例外分析是必要的。

(5) 可阅读:由于阅读解答的对象是人,约简过程和解答必须可解释。

与机器学习相比,数据分析所暗示的需要解决的问题是零散且不系统的,这正表明数据分析研究的复杂性。究其原因,与机器学习中“一致”一样,“需求”的多样性是最困难的问题。

本文主要讨论机器学习,并建议使用某些机器学习原理(例如,符号机器学习与流形机器学习)发展数据分析的新方法。

随着计算机与数字网络的普遍使用,人们方便获得信息的能力已今非昔比,然而,如何有效使用这些信息,以提高社会生产力与改善生活质量,至今还是一个挑战性的问题。机器学习是解决这个问题的重要途径之一。本文是机器学习研究历史与现状的分析与讨论,并借此提出我们对机器学习研究的看法。

1.2　机器学习的发展历史

我们回顾机器学习发展历史[4,5]的动机主要是为了警示我们注意那些本质不能做的事情,以避免重复历史上已经发生过的错误。另外,提示我们关注那些前人研究中的某些动机,在当时,它们可能受到技术条件的限制而不可行,但是,在当前技术条件下,可能获得新生。

机器学习的科学基础之一是神经科学,然而,对机器学习进展产生重要影响的是以下三个发现③,分别是(参见文献[4]):

(1) James 关于神经元是相互连接的发现。

(2) McCulloch 与 Pitts 关于神经元工作方式是“兴奋”和“抑制”的发现。

(3) Hebb 的学习律(神经元相互连接强度的变化)。

其中,McCulloch 与 Pitts 的发现对近代信息科学产生了巨大的影响(参见文献[1])④。对机器学习,这项成果给出了近代机器学习的基本模型,加上指导改变连接神经

---

③ 事实上,机器学习的任何进展主要受到神经科学与认知科学某个原理的启示,但是,这三个来自神经科学的原理可能影响最为深远。

④ 我们没有在这里引用其原始文献,原因是对计算机科学研究者来说,直接阅读这些文献显得生疏,很难抓住其对计算机科学的意义,而 Anderson 编辑的这本书不仅包括了原始文献的全文,而且还在计算的意义下,精炼了一个简洁的对这个贡献的摘要,这对计算机科学家来说,可能比阅读其原始文献更有意义。

⑤ 原始文献出版于 1948 年,本文引用的是 1962 年科学出版社出版的中文翻译版。

元之间权值的 Hebb 学习律,成为目前大多数流行的机器学习算法的基础。

1954 年,Barlow 与 Hebb 在研究视觉感知学习时,分别提出了不同假设:Barlow 倡导单细胞学说,假设从初级阶段而来的输入集中到具有专一性响应特点的单细胞,并使用这个神经单细胞来表象视觉客体。这个考虑暗示,神经细胞可能具有较复杂的结构;而 Hebb 主张视觉客体是由相互关联的神经细胞集合体来表象,并称其为 ensemble。在神经科学的研究中,尽管这两个假设均有生物学证据的支持,但是,这个争论至今没有生物学的定论。这个生物学的现实,为我们计算机科学家留下了想象的空间,由于在机器学习中一直存在着两种相互补充的不同研究路线,这两个假设对机器学习研究有重要的启示作用 。

在 McCulloch 与 Pitts 模型的基础上,1957 年,Rosenblatt 首先提出了感知机算法[7],这是第一个具有重要学术意义的机器学习算法。这个思想发展的坎坷历程,正是机器学习研究发展历史的真实写照。感知机算法主要贡献是:首先,借用最简单的 McCulloch 与 Pitts 模型作为神经细胞模型;然后,根据 Hebb 集群的考虑,将多个这样的神经细胞模型根据特定规则集群起来,形成神经网络,并将其转变为下述机器学习问题:计算一个超平面,将在空间上不同类别标号的点划分到不同区域。在优化理论的基础上,Rosenblatt 说明,如果一个样本集合是线性可分,则这个算法一定可以以任何精度收敛。由此导致的问题是,对线性不可分问题如何处理。

1969 年,Minsky 与 Papert 出版了对机器学习研究具有深远影响的著作 *Perceptron*(《感知机》)[8]。目前,人们一般的认识是,由于这本著作中提出了 XOR 问题,从而扼杀了感知机的研究方向。然而,在这本著作中对机器学习研究提出的基本思想,至今还是正确的,其思想的核心是两条 :

(1) 算法能力:只能解决线性问题的算法是不够的,需要能够解决非线性问题的算法。

(2) 计算复杂性:只能解决玩具世界问题的算法是没有意义的,需要能够解决实际世

---

⑥ 事实上,机器学习的文献并没有引用这些研究。激发我们这样做的原因是,在机器学习划分的研究中,基于这两个假设,可以清晰地将机器学习发展历程总结为:以感知机、BP 与 SVM 等为一类,以样条理论、*k*-近邻、Madaline、符号机器学习、集群机器学习与流形机器学习等为另一类(见本文其他小节)。当然,我们意识到,这存在被计算机科学家批评的风险,因为在人工神经网络的研究中,神经细胞是作为人工神经网络的一个节点,本文则是将模型(可以是人工神经网络本身)考虑为神经细胞。事实上,这个考虑并未违反生物学原理,与 Hebb 学习律也是一致的。然而,我们可以更清晰地对机器学习方法分类,如果这是一种学术“风险”,我们认为这个“风险”是值得冒的(参见本文以后的讨论)。

⑦ 应该指出的是,对计算而言,这是两个相互矛盾的原则,解决这个矛盾的出路,应该是领域知识,这正是 Minsky 想说的要点。既考虑非线性,又考虑计算复杂性,正是我们机器学习研究的核心,因此,研究满足这两个要求的算法,一直是我们机器学习所要解决的问题,然而,它恰恰是在理论上无解,领域知识是将其变换为可解问题的出路。

界问题的算法。

另外,这个著作中对算法设计的方法上,特别强调几何表示的重要性,尽管目前这个观点还不能为所有人接受,但是,我们的经验认为,这是对机器学习算法设计方法论的重要贡献。

在 1986 年,Rumelhart 等人的 BP 算法解决了 XOR 问题,沉寂近二十年的感知机研究方向重新获得认可[9],人们自此重新开始关注这个研究方向,这是 Rumelhart 等人的重要贡献。然而,1988 年,Minsky 与 Papert 在不作修改的情况下,重版他们的这部著作,仅仅增加了一个长篇重版序言与一个跋,并在重版此书的第一页上,他们手写下了“以此纪念 Rosenblatt”。这两个举动表明:其一,重申他们的上述核心思想,并对 BP 算法不以为然;其二,他们预言,最终兴起的将是感知机,而不是 BP。尽管 BP 算法的研究在以后的十年中成为主流,不幸的是,最近几年,Minsky 与 Papert 的预言成为了现实,人们的研究又回到了感知机。

在 20 世纪 60 年代的另一个重要研究成果来自 Widrow。1960 年,Widrow 推出了 Madaline 模型[10],在算法上,对线性不可分问题,其本质是放弃划分样本集的决策分界面连续且光滑的条件,代之分段的平面 。从近代的观点来看,这项研究与感知机的神经科学假设的主要区别是:它是确认 Barlow 假设中神经细胞具有较复杂结构的思想,由此,将线性模型(例如,感知机)考虑为神经细胞模型(而不是简单的 McCulloch 与 Pitts 模型),然后,再基于 Hebb 神经元集合体假设,将这些局部模型集群为对问题世界的表征,由此解决线性不可分问题。但是,这项研究远不如感知机著名,其原因是:其一,尽管 Madaline 可以解决线性不可分问题,但是,其解答可能是平凡的 ;其二,Widrow 没有给出其理论基础,事实上,其理论基础远比感知机复杂,直到 1990 年,Schapire 根据 Valiant 的“概率近似正确(PAC)”理论证明了“弱可学习定理”之后,才真正引起人们的重视。

进一步比较机器学习中两个不同路线的神经科学启示是有趣的:对机器学习来说,它们最显著的差别是对神经细胞模型的假设,例如,感知机是以最简单的 McCulloch 与 Pitts 模型作为神经细胞模型,而 Madaline 是以问题世界的局部模型作为神经细胞模型,两种方法都需要根据 Hebb 思想集群。因此,对机器学习研究,两个神经科学的启示是互补的。但是,两者还有区别:前者强调模型的整体性,这与 Barlow“表征客体的单一细胞论”一致,因此,我们称其为 Barlow 路线;而后者则强调对世界的表征需要多个神经细胞集群,这与 Hebb“表征客体的多细胞论”一致,我们称其为 Hebb 路线。鉴于整体模型与局部模型之间在计算上有本质差别,尽管根据 Barlow 与 Hebb 假设区分机器学习的方法

---

⑧ 事实上,目前更为普遍被我们使用的 k-近邻方法同样是 Hebb 路线的产物。

⑨ 假设在平面上有 n 个点,我们总可以使用 $n-1$ 条直线将其划分(线性分类器),但是,这个解答是平凡的,因为,信息长度没有被缩短。

并不严格,但是,在区分机器学习研究理念上有启示作用[①]。

机器学习早期研究的特点是以划分为主要研究课题,这个考虑一直延续到 Vapnik 在 20 世纪 70 年代发展的关于有限样本统计理论,并于 20 世纪 80 年代末流传到西方之后,在泛化能力意义下指导算法设计才成为人们关注的主要问题,这是本文需要进一步讨论的问题。

尽管以 Open 问题驱动的 BP 算法研究大大推动了感知机研究方向的发展,然而,近十年计算机科学与技术的快速发展,使得人们获得数据的能力大大提高,BP 这类算法已完全不能适应这种需求,退出人们的研究视野成为必然,同时,Minsky 的算法设计原则愈显重要。

然而,沿着 Barlow 路线的机器学习研究并没有终止,自 1992 年开始,Vapnik 将有限样本统计理论介绍给全世界,并出版了统计机器学习理论的著作[2]。尽管这部著作更多地是从科学、哲学上讨论了机器学习的诸多问题,但是,其暗示的算法设计思想对以后机器学习算法研究产生了重要的影响。

Vapnik 的研究主要涉及机器学习中两个相互关联的问题,泛化问题与表示问题。前者包含两个方面的内容:其一,有限样本集合的统计理论;其二,概率近似正确的泛化描述。而后者则主要集中在核函数,由此,将算法设计建立在线性优化理论之上。

Valiant 的"概率近似正确"学习的考虑在机器学习的发展中扮演了一个重要的角色。1984 年,Valiant 提出了机器学习的一个重要考虑,他建议评价机器学习算法应该以"概率近似正确(PAC)"为基础[12],而不是以传统模式识别理论中以概率为 1 成立为基础[13],由此,他引入了类似在数学分析中的 ε-δ 语言来描述 PAC,这个考虑对近代机器学习研究产生了重要的影响。首先,统计机器学习理论中泛化不等式的推导均以这个假设为基础;其次,基于这个考虑的"弱可学习理论",为研究基于 Hebb 路线的学习算法设计奠定了理论基础,并产生被广泛应用的集群机器学习理念(ensemble)。

1990 年,Schapire 证明了一个有趣的定理:如果一个概念是弱可学习的,充要条件是它是强可学习的。这个定理的证明是构造性的,证明过程暗示了弱分类器的思想[14]。所谓弱分类器就是比随机猜想稍好的分类器,这意味着,如果我们可以设计这样一组弱分类器,并将它们集群起来,就可以成为一个强分类器,这就是集群机器学习。由于弱分类器包含"比随机猜想稍好"的条件,从而,避免了对 Madaline 平凡解的批评。另外,由于 Schapire 定理的证明基于 PAC 的弱可学习理论,因此,这种方法又具有泛化理论的支持。

---

① 这似乎存在着矛盾:Barlow 强调整体性,这暗示神经细胞具有复杂结构,而由简单结构形成的整体模型又需要集群,问题是 Barlow 暗示的神经细胞的复杂结构组成部分是否可以使用简单神经模型表述,这是神经科学没有回答的问题。这里,我们将神经细胞结构中的各组成部分假设为一些已知模型,只能是计算机科学家在没有神经科学依据下的无奈的想象而已。因此,这仅仅是对我们研究机器学习的启示罢了。

这样,自 Widrow 提出 Madaline 近 30 年之后,人们终于获得了基于 Hebb 路线下的机器学习算法设计的理论基础。这个学习理念立即获得人们的广泛关注,其原因不言自明,弱分类器的设计总比强分类器设计容易,特别是对线性不可分问题更是如此。由此,Madaline 与感知机一样,成为机器学习最重要的经典。

自 1969 年 Minsky 出版 *Perceptron*(《感知机》)一书以后,感知机的研究方向被终止,到 1986 年 Rumelhart 等发表 BP 算法,近 20 年间,机器学习研究者在做什么事情呢?这段时间正是基于符号处理的人工智能的黄金时期,由于专家系统研究的推动,符号机器学习得到发展,事实上,这类研究方法除了建立在符号的基础上之外,从学习的机理来看,如果将学习结果考虑为规则,每个规则将是一个分类器,尽管这些分类器中有些不一定满足弱分类器的条件,但是,它应该是 Hebb 路线的延续。

符号机器学习的最大优点是归纳的解答与归纳的过程是可解释的,换句话说,数据集合中的每个观测(样本或对象)对用户都是透明的,它在解答以及计算过程中所扮演的角色,用户都是可以显现了解的。然而,它的缺陷同样突出,就是泛化能力。由于学习结果是符号表述,因此,只可能取“真”与“假”,这样大大减低了对具有一定噪音数据的分析能力,需要其他技术来补充:其一,观测世界的数据到符号域的映射,其二,不确定推理机制。但是,这两种方法与符号机器学习方法本身并没有必然的关系。

近几年,由于数据挖掘的提出,符号机器学习原理有了新的用途,这就是符号数据分析,在数据挖掘中称为数据描述,以便与数据预测类型的任务相区别(从任务来说,这类任务与机器学习是一致的)。

与机器学习的目标不同,数据分析不是以所有用户具有相同需求为假设,相反,强调不同用户具有不同的需求。另外,数据分析强调,分析结果是为用户提供可阅读的参考文本,决策将依赖人的洞察。如何根据用户的特定需求将观测数据集合变换为简洁的、可为用户理解的表示成为关键。这是符号机器学习的另一个可以考虑的应用领域。由于符号机器学习在泛化能力上的欠缺,这也是它在与基于统计的机器学习方法竞争中避免遭到淘汰的出路。

在这一节的最后,将 1989 年 Carbonell 对机器学习以后十年的展望[15] 与十年后 Dietterich 的展望[16] 作一个对比,可能是有趣的,我们希望以此说明机器学习研究由于面临问题的改变所发生的变迁(表 1)。

近几年,由于数据性质的复杂性与用户需求的多样性,使得各种机器学习方法层出不穷(参见 1.7 节)。这些问题驱动的方法借用了大量数学工具,以表示不同的需求与数据性质,同时根据数学工具来解决这些问题。精确地预测这些方法的生命力还为时过早,因为这些方法的提出尽管有充分的应用背景,但是,大多数方法至今其理论基础还比较粗糙,甚至没有自己独特的理论基础,同时,也还没有令人信服地解决了已有方法不能解决的问题,尝试与经验积累是当前的研究主流。

表　1

文献[15] (Carbonell, 1989)	文献[16] (Dietterich, 1999)	注　　解
符号机器学习	符号机器学习	保留:发生本质变化,符号数据分析
连接机器学习	统计机器学习	分为:基于 Barlow 路线的方法
	集群机器学习	分为:基于 Hebb 路线的方法
遗传机器学习	强化机器学习	扩展:强调反馈的作用,以及动态规划的解决方案
分析机器学习	—	放弃:由于问题过于复杂

注:需要指出的是,Dietterich 将连接机器学习中感知机类分离出来,并将其分为两个子类,分别对应 Barlow 路线与 Hebb 路线的统计机器学习与集群机器学习。尽管连接机器学习中的其他类型的学习方法可能不再是热点,但是,不等于已无研究的意义。

1.3　统计机器学习

统计机器学习是近几年被广泛应用的机器学习方法,事实上,这是一类相当广泛的方法。更为广义地说,这是一类方法学。当我们获得一组对问题世界的观测数据,如果我们不能或者没有必要对其建立严格物理模型,我们可以使用数学的方法,从这组数据推算问题世界的数学模型,这类模型一般没有对问题世界的物理解释,但是,在输入输出之间的关系上反映了问题世界的实际,这就是"黑箱"原理。一般来说,"黑箱"原理是基于统计方法的(假设问题世界满足一种统计分布),统计机器学习本质上就是"黑箱"原理的延续。与感知机时代不同,由于这类机器学习科学基础是感知机的延续,因此,神经科学基础不是近代统计机器学习关注的主要问题,数学方法成为研究的焦点。

本文所说的统计机器学习,是指 20 世纪 90 年代以 Vapnik 的著作 *The Nature of Statistical Learning Theory* 为标志的机器学习研究。Vapnik 对机器学习的认识不同于神经网络时代的要点是:其一,强调泛化能力,并将学习算法设计建立在泛化指标的基础上;其二,强调线性划分,在学习算法设计上,指出"回归感知机"的重要性[11]。这两个特点可以简单地总结为:泛化和表示两个基本问题。

1.3.1　泛化问题

在 30 年后的今天,我们重温 20 世纪 70 年代的两项研究是有趣的,其一,对机器学习有重要影响的统计模式识别[13],其二,有限样本的统计理论(参见文献[17])[1]。尽管前者

1　很多中文文献将其翻译为小样本统计理论,作者不赞成这个译法。事实上,在泛化不等式的推导过程中,样本个数必须满足一个依赖泛化精度指标的条件。因此,"小样本集合"仅仅是"有限",以便与 Duda 泛化理论依赖无限样本相区别。

的研究是建立在模式识别任务的基础上,然而,这个任务的统计基础却是与机器学习一致的,当然,其方法是建立在经典统计理论的基础上。它对泛化能力的刻画是经典统计理论中的大数定理,具体地说,泛化能力需要以样本数量趋近无穷大来描述。Duda 的贡献主要是提出了以经典统计理论为工具刻画模式识别与机器学习的各类任务,同时暗示了对所建模型的评价方法。而后者则试图建立一种新的统计理论——有限样本的统计理论,具体地说,将学习的样本集合理解为从问题世界随机选取的子集,由于不同的样本集合对应不同的模型(也称为假设),而不同模型对问题世界为真的程度不同(泛化或误差),如何计算对问题世界“最真”的模型就是主要任务。这样,样本集合成为泛化指标的随机变量,由此,建立了结构风险理论。

本文不准备详细讨论 Duda 理论,而将重点强调 Vapnik 的贡献,并指出两者的区别。

从 Duda 开始,泛化问题的理论就是使用“风险”来刻画数学模型与问题世界模型之间的差别,由于问题世界模型不可获得,由此建立的任何数学模型都将有损失,称为风险。我们希望风险最小。理想情况下的风险称为期望风险,记为 R_{exp},由于其不能直接计算,只能估计,这个估计的风险称为经验风险,记为 R_{emp}。由此,泛化问题就是计算使得 $|R_{emp} - R_{exp}|$ 最小的模型。经典方法认为:当样本个数趋于无穷大时,如果所建立的数学模型是成功的(以概率 1 成立),$|R_{emp} - R_{exp}|$ 应该趋近零。这样,由于需要样本集合趋近无穷大,样本集合成为唯一的,因此,经典泛化理论将不包括样本集合这个因素。

Vapnik 的考虑不尽相同:其一,样本集合是风险描述的重要因素,换句话说,样本集合将是风险公式中的一个变量(在统计上,是一个关于风险的随机变量);其二,根据 PAC,模型以概率 $1 - \delta$ 成立,即,模型泛化能力以概率近似正确描述。由此,这个统计理论不能简单地仅仅考虑经验风险与期望风险之间的关系,同时需要考虑划分样本集合函数族的划分能力,称为置信范围 Φ。这样,构成风险不等式:$R_{exp}(\alpha) \leq R_{emp}(\alpha) + \Phi(k/d)$,其中,$d$ 是函数族的 VC 维。这就是所谓的结构风险。我们期望 $R_{emp}(\alpha) + \Phi(k/d)$ 最小,即,$R_{emp}(\alpha)$ 与 $\Phi(k/d)$ 同时最小,以获得对期望风险估计的下界。由于 $R_{emp}(\alpha)$ 与 $\Phi(k/d)$ 同时最小在理论上过于复杂,甚至不可能,在目前的泛化理论(泛化不等式)研究中,在假设 $R_{emp}(\alpha)$ 已经最小的条件下,估计 $\Phi(k/d)$ 最小,这样,泛化理论可以建立在基于参数 α 的函数 $f(\alpha)$ 的 VC 维上。

泛化不等式的研究首先将样本集合考虑为从问题世界中随机选取的一个子集,而每个样本集合对应一个模型,称为假设,这样,泛化不等式经历了三个重要的阶段:①“假设”(模型)个数有限,根据 Valiant 的 PAC 理论,推出泛化不等式,我们称其为 PAC 泛化不等式[12];②“假设”个数无限,根据 VC 维推出泛化不等式,称为 VC 维泛化不等式,这个不等式对以后泛化理论的研究有重要的意义[18];③由于 VC 维泛化不等式不够直观,因此,它还不能指导算法设计,1998 年,具有几何直观的最大边缘不等式被发现[19],这个不等式无论在直观几何解释上还是在指导算法设计上均具有重要意义。由于其直观的几