Boosting 学习理论的探索
——一个跨越 30 年的故事

周志华
南京大学

关键词：机器学习　AdaBoost 间隔理论　Boosting 学习理论

这篇文章尝试用通俗故事的方式讲述一个机器学习理论中重要问题的探索历程。读者或能从中感受到机器学习理论探索的曲折艰辛，体会到理论进展对算法设计的指引意义。

溯源

1989 年，哈佛大学的莱斯利·维利昂特（Leslie Valiant，计算学习理论奠基人、2010 年 ACM 图灵奖得主）和他的学生迈克尔·肯斯（Michael Kearns，后来担任贝尔实验室人工智能研究部主任）提出了一个公开问题："弱可学习性是否等价于强可学习性？"

这个问题大致上是说：如果一个机器学习任务存在着比"随机猜测"略好一点的"弱学习算法"，那么是否就必然存在着准确率任意高（与该问题的理论上限任意接近）的"强学习算法"？

直觉上这个问题的答案大概是"否定"的，因为我们在现实任务中通常很容易找到比随机猜测稍好一点的算法（比方说准确率达到 51%）、却很难找到准确率很高的算法（比方说达到 95%）。

出人意料的是，1990 年，麻省理工学院的罗伯特·夏柏尔 (Robert Schapire) 在著名期刊 *Machine Learning* 上发表论文，证明这个问题的答案是"YES"！更令人惊讶的是，他的证明是构造性的！

也就是说，夏柏尔给出了一个过程，直接按这个过程进行操作就能将弱学习算法提升成强学习算法。过程的要点是考虑一系列"基学习器"，让"后来者"重点关注"先行者"容易出错的部分，然后再将这些基学习器结合起来。

遗憾的是，这个过程仅具备理论意义，并非一个能付诸实践的实用算法，因为它要求知道一些实践中难以事先得知的信息，比方说在解决一个问题之前，先要知道这个问题的最优解有多好。

后来夏柏尔到了新泽西的贝尔实验室工作，在这里遇见加州大学圣塔克鲁兹分校毕业的约夫·弗洛恩德 (Yoav Freund)。凑巧的是，弗洛恩德曾经研究过多学习器的结合。两人开始合作。终于，他们在 1995 年欧洲计算学习理论会议（注：该会议

图 1　时任 ACM 主席戴维·帕特森（David Patterson，右一）和首席运营官约翰·怀特（John White，右四）向夏柏尔（右三）和弗洛恩德（右二）颁发"ACM 帕里斯·卡内拉基斯理论与实践奖"

已经并入 COLT）发表了一个实用算法，完整版于 1997 年在 *Journal of Computer and System Sciences* 发表。这就是著名的 AdaBoost。夏柏尔和弗洛恩德因这项工作获 2003 年“哥德尔奖”、2004 年“ACM 帕里斯·卡内拉基斯 (Paris Kanellakis) 理论与实践奖”。

AdaBoost 的算法流程非常简单，用夏柏尔自己的话说，它仅需“十来行代码 (*just 10 lines of code*)”。但这个算法非常有效，并且经修改推广能应用于诸多类型的任务。例如，在人脸识别领域被誉为“第一个实时人脸检测器”、获得 2011 年龙格 - 希金斯 (Longuet-Higgins) 奖的维奥拉 - 琼斯 (Viola-Jones) 检测器就是基于 AdaBoost 研制的。AdaBoost 后来衍生出一个庞大的算法家族，统称 Boosting，是机器学习中“集成学习”的主流代表性技术。即便在当今这个“深度学习时代”，Boosting 仍发挥着重要作用，例如经常在许多数据分析竞赛中打败深度神经网络而夺魁的 XGBoost，就是 Boosting 家族中 GradientBoost 算法的一种高效实现。

我们今天的故事就是关于 Boosting 学习理论的探索。

异象

机器学习界早就知道，没有任何一个算法能“包打天下”（注：著名的“没有免费的午餐”定理，参见周志华《机器学习》，清华大学出版社 2016，1.4 节）。因此不仅要对算法进行实验测试，更要进行理论分析，因为有了理论上的清楚刻画，才能明白某个机器学习算法何时能奏效、为什么能奏效，而不是纯粹“碰运气”。

1997 年，夏柏尔和弗洛恩德给出了 AdaBoost 的第一个理论分析。他们证明，AdaBoost 的训练误差随训练轮数增加而指数级下降；这意味着算法很快收敛。对于大家最关心的泛化性能，即算法在处理新的、没见过的数据时的性能，他们的结论是：AdaBoost 的泛化误差 ≤ 训练误差 + $\tilde{O}\left(\sqrt{\frac{\ln|\mathcal{H}|T}{m}}\right)$（注：$\tilde{O}$ 隐去了相对不太重要的其他项），这里的 m 是训练样本数，T 是训练的轮数，$\ln|\mathcal{H}|$ 可以大致理解为基学习器的复杂度；因为 AdaBoost 每训练一轮就增加一个基学习器，所以 $\ln|\mathcal{H}|T$ 大致相当于最终集成学习器的复杂度。于是，这个理论结果告诉我们：训练样本多些好、模型复杂度小些好。

希望训练样本多，这容易理解。为什么希望模型复杂度小呢？这是由于机器学习中存在“过拟合”，简单地说，如果对训练数据学得“太好了”，反而可能会犯错误。例如图 2，在学习“树叶”时，如果学习器错误地认为没有锯齿就不是树叶，这就过拟合了。一般认为，产生过拟合的重要原因之一，就是由于模型过于复杂，导致学得“过度”了、学到了本不该学的训练样本的“特性”而非样本总体的“共性”。

图 2　“过拟合”的直观解释

显然，夏柏尔和弗洛恩德在 1997 年的理论蕴义与机器学习领域的常识一致，因此很容易得到大家认可。

然而，AdaBoost 在实践中却呈现出一个奇异的现象：它似乎没有发生过拟合！

如图 3 所示，在训练误差到达 0 之后继续训练，虽然模型复杂度在增大，但泛化误差却仍会继续下降。

科学发现中有一个基本方法论：若有多个理论假设符合实验观察，则选取最简洁的。这就是所谓“奥卡姆剃刀 (Ocama's razor) 准则”。这个准则在众多学科史上都发挥了重要作用。然而如果审视 AdaBoost 的行为，却可以发现它是如此与众不同。

如图 3 中，训练到第 10 轮和第 1000 轮时形成

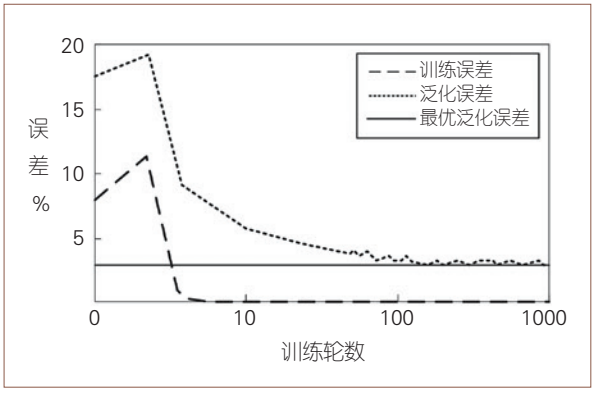

图 3　AdaBoost 的典型表现

的假设（集成学习器）都与“实验观察”（训练数据）一致，前者仅包含 10 个基学习器、后者包含 1000 个基学习器。显然，根据奥卡姆剃刀应该选取前者，但实际上后者却更好。

也就是说，AdaBoost 的行为表现不仅违背了机器学习领域的常识，从更广大的视角看，甚至违背了科学基本准则！

因此，弄清 AdaBoost 奇异现象背后的道理，不仅能满足我们的好奇心，还可能揭开机器学习中以前不知道的某种秘密，进而为算法设计打开一扇新门。“AdaBoost 为何未发生过拟合？”作为 Boosting 最关键、最引人入胜的基础理论问题，吸引了诸多知名学者投入其中。

惊蛰

夏柏尔和弗洛恩德很快意识到 1997 理论中的问题。1998 年，他们与后来领导伯克利著名的西蒙斯计算理论研究所的彼得·巴特莱特 (Peter Bartlett) 和李伟上 (Wee Sun Lee) 合作发表了一个新的基于“间隔 (margin)”的理论。

“间隔”是机器学习中一个非常重要的概念。大致来说，如图 4 所示，假定我们用一个划分超平面把不同类别的样本分开，那么某个样本点与超平面的“距离”就是这个样本点相对该超平面的“间隔”。考虑所有样本点相对这个超平面的“最小间隔”，就定义出了“超平面的间隔”。机器学习中著名的支持向量机 SVM 就是通过优化技术来求解出间隔最大的划分超平面，换一个角度看，就是试图使样本点相对超平面的“最小间隔”尽可能大。

在夏柏尔等人的新理论中，AdaBoost 的泛化误差界包含一个关于间隔的阈值项 θ，并且 θ 出现在分母上；这意味着间隔越大，泛化误差就可能会越小。这个理论漂亮地解释了 AdaBoost 为什么没有发生过拟合：这是因为即便训练误差达到 0，间隔仍有可能增大。如图 4，超平面 B 已经把两类训练样本点完全分开，其训练误差为 0；继续训练可能找到超平面 A，训练误差仍为 0，但是 A 的间隔比 B 更大，所以泛化误差可以进一步减小。

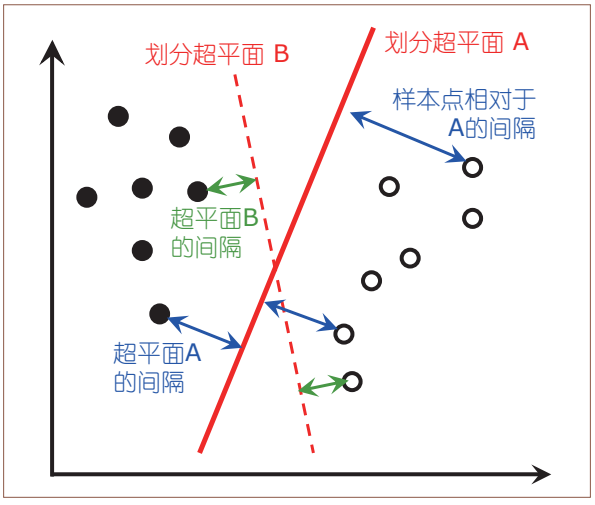

图 4　机器学习中“间隔”的直观含义

这项理论在 1998 年发表，刚好在那一年，多特蒙德大学的托斯腾·约阿希姆斯 (Thorsten Joachims) 在欧洲机器学习大会上报道了支持向量机在文本分类任务上展现出卓越性能，机器学习领域正式进入“统计学习时代”，而“间隔”正是支持向量机的核心概念。从间隔的角度来解释 AdaBoost 的行为，无形中使机器学习的“集成学习”与“统计学习”这两大流派走到了一起。

有趣的是，统计学习奠基人、支持向量机发明人弗拉基米尔·N. 瓦普尼克 (Vladimir N. Vapnik) 在 1990 年（苏联解体前一年）离开苏联来到新泽西的贝尔实验室工作，夏柏尔和弗洛恩德合作时与瓦普

尼克是同事，或许受到了他的影响。

冰封

夏柏尔等人在 1998 论文结尾说，他们与李奥·布瑞曼 (Leo Breiman) 进行了“沉痛 (*poignant*)”的交流。用这个词形容学术交流很少见。给他们带来沉痛感的正是被誉为“二十世纪伟大的统计学家”的加州大学伯克利分校的布瑞曼教授，他不仅是统计学领域巨擘，还是机器学习中著名的集成学习方法 Bagging 和“随机森林”的发明人。

1999 年，布瑞曼的独立署名论文提出了一个新理论。这个理论也是基于“间隔”来对 AdaBoost 进行刻画，但是与夏柏尔等人 1998 理论不同的是，布瑞曼理论的着眼点是“最小间隔”，基于这个间隔物理量，布瑞曼得到了比夏柏尔等人“更紧”的泛化误差界，是 Boosting 间隔理论体系中一个新的里程碑。

在学习理论中，“更紧的界”通常意味着“更本质的物理量”。上一节谈到过，“最小间隔”取决于最靠近划分超平面的样本点，而著名的支持向量机就是在努力寻找使得“最小间隔”最大的划分超平面。所以，布瑞曼理论揭示出“最小间隔”竟然也是 AdaBoost 的关键，这使人们感到“醍醐灌顶”。

到此一切都很美好，只需承认：Boosting 间隔理论体系的关键在于“最小间隔”就 OK 了。

然而，布瑞曼不仅理论造诣深厚，还是一位热忱的机器学习算法发明者。既然证明了“最小间隔”是关键，而 AdaBoost 跟最小间隔的关联是通过复杂的理论分析才发现的，也就是说，AdaBoost 并非“直接”优化最小间隔，那么何不发明一个直接优化最小间隔的新型 Boosting 算法呢？理论上来说这算法应该会比 AdaBoost 更强大。

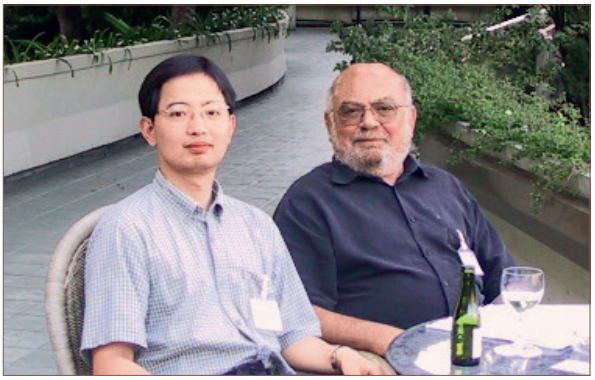

图 5　2003 年笔者与布瑞曼在欧洲机器学习大会期间探讨 Boosting 理论问题

于是，布瑞曼在 1999 年发明了一个新算法 Arc-gv。可以从理论上证明，这个算法能够使最小间隔达到最优。同时，布瑞曼的大量实验也验证了该算法对应的“最小间隔”总是大于 AdaBoost 的“最小间隔”。然而，实验却显示出 Arc-gv 的泛化误差大于 AdaBoost!

这就奇怪了。已经证明了 Boosting 间隔理论体系的关键是“最小间隔”，现在 Arc-gv 的“最小间隔”无论在理论上还是在实验上都优于 AdaBoost，那么 Arc-gv 算法的泛化性能应该更好啊，为什么反倒不如 AdaBoost 呢？！

证明过程无误。这就意味着，把“间隔”作为基石是错误的。因此，布瑞曼的工作不啻于宣告了 Boosting 间隔理论体系的“死刑”。

Boosting 间隔理论体系由此进入“冰封年代”。研究者纷纷转向其他理论体系，其中最重要的是名著 *The Elements of Statistical Learning* 作者、斯坦福大学三位大师杰罗姆·弗里德曼 (Jerome Friedman)、特莱沃尔·哈斯蒂 (Trevor Hastie)、罗伯特·提比希阿尼 (Robert Tibshirani) 提出的“统计视角 (Statistical View)”理论。这一派理论也有不少争议，尤其是未能清楚解释 AdaBoost 为何未发生过拟合。但人们认为至少它“活着”、还可以被改善，而间隔理论体系虽有优美的几何直观解释，但它已经“死了”。

续命

七年后，在 2006 年国际机器学习大会 (ICML) 上，在普林斯顿大学任教的夏柏尔与耶鲁大学的学生列夫·芮因 (Lev Reyzin) 合作获得了“杰出论文奖”。获奖论文做了一件事：把布瑞曼 1999 年的实验重新做一遍。

我们知道，Boosting 间隔理论体系除了间隔，必然还会涉及到训练样本数和模型复杂度。要讨论间隔对泛化误差的影响，就必须把训练样本数和模型复杂度“固定”住。前者容易：指定训练样本的个数即可；后者则必须专门处理。

一般认为，决策树模型复杂度由叶结点的数目决定，因此在 1999 年实验中布瑞曼使用了叶结点数目相同的决策树。芮因和夏柏尔重复了布瑞曼的实验，发现 AdaBoost 决策树虽然与 Arc-gv 决策树的叶结点数目相同，但树的层数却更多。因此芮因和夏柏尔推测，这些决策树的复杂度或许不同？

这个发现并不容易，因为这些决策树的高度差并不大。例如，在各自使用 300 棵决策树时，AdaBoost 决策树平均高度约为 10.5，仅比 Arc-gv 决策树高了约 1 层。从大量实验数据中观察到这么小的差别，需要相当敏锐的观察力。

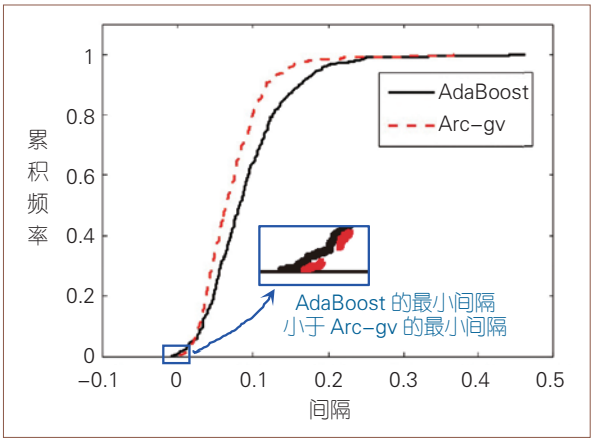

图 6　AdaBoost 的最小间隔小于 Arc-gv，但间隔总体上更大一些

芮因和夏柏尔用仅包含 2 个叶结点的单层决策树（亦称“决策树桩”）作为基学习器重新做了实验。他们发现，虽然 Arc-gv 的最小间隔始终大于 AdaBoost，但是若考虑样本总体，则 AdaBoost 的间隔比 Arc-gv 更大一些。例如从图 6 中可以看到 AdaBoost 的曲线更“靠右”，这意味着有更多的样本点取得较大的间隔值。于是芮因和夏柏尔断言，“最小间隔”并非 Boosting 间隔理论体系的关键，重要的是间隔的总体分布。他们猜测，或许“间隔均值”或“间隔中位数”是关键物理量，但并未给出理论证明。

一般读者今天看这篇论文或许会感到很诧异：它既没有提出新理论，也没有提出新算法，也没有新应用，这个“三无”论文居然获得了 ICML 杰出论文奖？

这项工作的意义必须放到 Boosting 理论探索的整体图景中去考察：它显示出布瑞曼对 Boosting 间隔理论体系的致命打击，至少其中的实验部分并未“实锤”！

那么，芮因和夏柏尔的工作使 Boosting 间隔理论体系“活过来了”吗？还没有。因为布瑞曼的主要攻击来自于他的理论结果，实验起的是“验证”作用。因此，Boosting 间隔理论体系想逃生，就必须有新理论，基于“最小间隔”之外的间隔物理量证明出比布瑞曼更紧的泛化误差界。

磨砺

2008 年，北京大学的王立威博士、封举富教授、杨成同学和后来担任日本 RIKEN 人工智能研究中心主任的杉山将 (Masashi Sugiyama) 与笔者合作，在 COLT 会议发表了一篇论文。这篇论文提出了“均衡间隔”的概念，并且基于它证明出一个比布瑞曼更紧的泛化误差界。由于“均衡间隔”不同于最小间隔，因此不少学者以为 Boosting 理论问题解决了。

然而，笔者心有不安。一方面，在“均衡间隔”理论证明过程中用到的信息与夏柏尔和布瑞曼有所不同，相应的结果未必能直接比较；另一方面，更重要的是，“均衡间隔”是从数学上“强行”定义出来的一个概念，我们说不清它的直观物理含义是什么。

回顾笔者研究 Boosting 理论问题的初心，是希望通过弄清楚“AdaBoost 为何未发生过拟合”来为新的机器学习算法设计获得启发。如果新理论的关键概念缺乏直观物理含义，那么就很难启发算法设计了。因此，笔者继续思考着。

2008 年，南开大学组合数学中心的高尉同学（现南京大学人工智能学院副教授）跟笔者联系攻读博

士学位，被笔者拽进了 AdaBoost 问题。当时估计一年应能解决。然而困难超过预期，三年下来仅基于一个走偏的小结果发表了一篇会议论文。博士生毕竟有毕业文章压力，感觉“走投无路”了。笔者内心也很惶惑，但还是必须斗志高昂地鼓劲。幸好笔者手上还有另一个重要理论问题，高尉在这个问题上发表了一篇颇有影响的 COLT 文章，稍缓解压力，能够继续前行。

几经磨砺，终于在 2013 年我们在 *Artificial Intelligence* 发表了一个新理论，相应的泛化误差界比夏柏尔和布瑞曼的更紧，这确证了“最小间隔”并非 Boosting 间隔理论体系的关键物理量。有意思的是，以往认为应该存在“某个”关键的间隔物理量，而我们的新理论揭示出：应该使得“平均间隔”最大化、同时使“间隔方差”最小化，也就是说，关键物理量并非一个，而是两个！

AdaBoost 为何未发生过拟合？为何在训练误差达到 0 之后继续训练仍能获得更好的泛化性能？新理论给出了答案：因为 AdaBoost 在训练过程中随着轮数的增加，不仅使平均间隔增大，还同时使间隔方差变小。同时，这也意味着 AdaBoost 最终仍有可能发生过拟合，只不过很迟——当平均间隔已无法再增大、间隔方差也无法进一步减小时。

众所周知，以支持向量机为代表的一大类统计学习方法都在试图最大化“最小间隔”，而这个新理论揭示：若能最大化“平均间隔”同时最小化“间隔方差”，得到的学习器会更好！于是，笔者的博士生张腾同学（现华中科技大学计算机学院教师）等开始了这方面的探索。2014 年开始的 5 年里，我们建立起“最优间隔分布学习机 (Optimal margin Distribution Machine, ODM)”这个新的算法族，包括二分类、多分类、聚类、半监督等学习算法，这些受新理论启发的算法工作不属于本文重点，就不赘述了。

定论

2013 年的工作引起了很多反响，如在 2014 年国际人工智能大会 (AAAI) 上，国际人工智能学会主席、卡内基梅隆大学机器学习系主任曼纽拉·维罗索 (Manuela Veloso) 教授的 Keynote 报告将它作为人工智能领域的重要进展介绍，称其“使间隔理论复兴 (*renascence*)”“为学习算法设计带来了新洞察 (*new insight*)”。

然而，笔者仍有隐忧。虽然 2013 理论相应的泛化误差界在当时是最紧致的，但今后会不会有人基于其他的间隔物理量获得更紧的界，导致我们关于“AdaBoost 为何未发生过拟合”的答案和“最大化平均间隔同时最小化间隔方差”的算法指导思想被推翻呢？

六年后，在 2019 年底的 NeurIPS 会议上，丹麦奥胡斯大学的阿兰·格洛隆德 (Allan Grønlund)、卡斯柏·拉森 (Kasper G. Larsen)、莱尔·卡玛 (Lior Kamma)、亚历山大·马塞厄森 (Alexander Mathiasen) 与加州大学伯克利分校的杰拉尼·纳尔逊 (Jelani Nelson) 合作发表了一篇论文（见图 7）。纳尔逊是美国总统奖和斯隆研究奖得主，拉森在 STOC 和 FOCS 曾两获最佳学生论文奖，是理论计算机科学界的新星，卡玛则毕业于以色列魏兹曼研究所这个计算机科学重镇。理论计算机科学家出手机器学习理论问题，是近年来的一个重要趋势。这篇论文最终证明了 2013 年我们给出的已经几乎是最紧的泛化误差上界，至多再改进一个 log 因子，并且这个上界已经与下界匹配，理论上不可能有更好的结果！

终于，心安了。

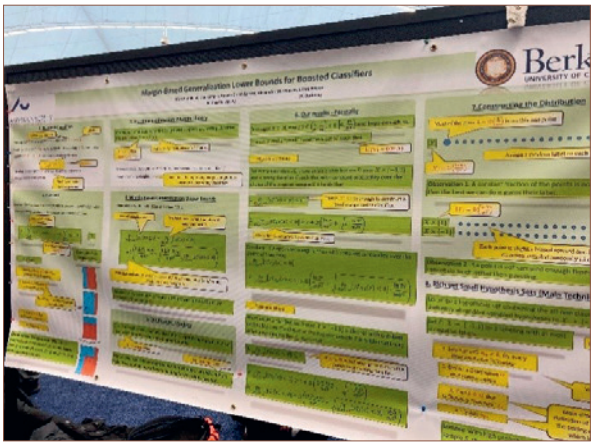

图 7　NeurIPS 2019 会议上这篇论文的 poster

Theorem 1. *(Schapire et al., 1998) For any $\delta > 0$ and $\theta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq \Pr_S[yf(\boldsymbol{x}) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m \ln|\mathcal{H}|}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right).$$

Theorem 2. *(Breiman, 1999) For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq R\Big(\ln(2m) + \ln\frac{1}{R} + 1\Big) + \frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta},$$

where $\theta = \hat{y}_1 f(\hat{\boldsymbol{x}}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}}$, $R = \frac{32\ln 2|\mathcal{H}|}{m\theta^2} \leq 2m$.

Theorem 3. *(Gao and Zhou, 2013) For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geq 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq \frac{2}{m} + \inf_{\theta\in(0,1]}\left[\Pr_S[yf(\boldsymbol{x}) < \theta] + \frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_S[yf(\boldsymbol{x}) < \theta]}\right],$$

where $\mu = \frac{8}{\theta^2}\ln m \ln(2|\mathcal{H}|) + \ln\frac{2|\mathcal{H}|}{\delta}$.

图 8　本文提到的最主要的 3 个理论结果

剧终

从 1998 年 AdaBoost 间隔理论体系萌生，到几经论争跌宕得到 2013 年结果，经过了 15 年。再经 6 年得到该结果的定论。如果从故事开头的 1989 年算起，整整经历了 30 年。故事中的一些主要人物如李奥·布瑞曼已经作古，而当年的研究生已成为教授。最后，本文不加解释地列出故事中最主要的三个理论结果以志纪念（见图 8）。 ■

周志华

CCF 会士、常务理事。南京大学教授、计算机系主任、人工智能学院院长、计算机软件新技术国家重点实验室常务副主任。ACM/AAAS/AAAI/IEEE/IAPR Fellow，欧洲科学院外籍院士。主要研究方向为人工智能、机器学习、数据挖掘。
zhouzh@nju.edu.cn

参考文献

[1] Zhou Z H. Large margin distribution learning[C]// *ANNPR 2014*.(keynote article)

[2] Zhang T, Zhou Z H. Optimal margin distribution machine[J]. *IEEE Transactions on Knowledge and Data Engineering*, DOI:10.1109/TKDE.2019.2897662.

[3] Grønlund A, Kamma L, Larsen K G, et al. Margin-based generalization lower bounds for boosted classifiers[C]// *NeurIPS 2019*.

注：对理论内容感兴趣的读者可以从 [1] 中找到主要文献；对 ODM 算法感兴趣的读者可参阅 [2]；[3] 是“定论”一节谈到的最新工作。