数据挖掘中的集成方法
——通过集成预测来提升精度

Ensemble Methods in Data Mining:Improving Accuracy Through Combining Predictions

〔美〕 Giovanni Seni　John Elder　著
王　攀　张　健　杨　洋　彭宇阳　卢其兵　译

科学出版社

数据挖掘中的集成方法

——通过集成预测来提升精度

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions

〔美〕Giovanni Seni　John Elder　著

王攀　张健　杨洋　彭宇阳　卢其兵　译

科学出版社

北京

图字：01-2014-5707 号

内 容 简 介

本书讨论基于决策树的集成，分析被视为当前现代集成算法高级性能的主要原因之一的正则化问题，描述集成方法领域近年来的两个发展——重要性采样（IS）和规则集成（RE），论述新数据信息集成在复杂性和更高精度方面的悖论等重要命题。全书面向前沿、文字简练、论述充分、可读性好。

本书可供计算机科学技术、控制科学与工程、信息科学与技术、机电工程、管理科学与工程等专业的教师、研究生、高年级本科生参考。

图书在版编目（CIP）数据

数据挖掘中的集成方法：通过集成预测来提升精度/（美）赛尼（Seni，G.），（美）艾德（John，E.）著；王攀等译. —北京：科学出版社，2015.5

书名原文：Ensemble Methods in Data Mining：Improving Accuracy Through Combining Predictions

ISBN 978-7-03-044327-4

Ⅰ. ①数… Ⅱ. ①赛… ②艾… ③王… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2015）第 105540 号

责任编辑：张艳芬 余 丁 赵晓廷 / 责任校对：桂伟利

责任印制：徐晓晨 / 封面设计：陈 敬

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 6 月第 一 版 开本：720 × 1000 1/16

2015 年 6 月第一次印刷 印张：7 插页：2

字数：121 000

定价：80.00 元

（如有印装质量问题，我社负责调换）

作 者 简 介

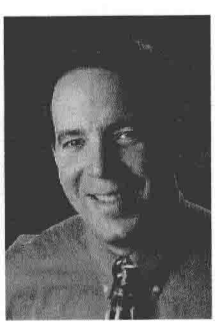
Giovanni Seni

Seni 是 Elder Research 公司的资深科学家，负责该公司的西部办公室。作为硅谷活跃的数据挖掘领域的践行者，他在统计模式识别、数据挖掘及人机交互应用领域具有逾 15 年的研发经历。他既是大企业的技术人员，又是一些小规模企业的贡献者。他拥有五项美国专利，发表了 20 余篇学术论文。

Seni 是圣塔克拉拉大学计算机工程系兼职教师，教授“模式识别与数据挖掘导论”课程。

他于 1989 年在洛斯安第斯大学（波哥大，哥伦比亚）获得计算机工程学士学位，1995 年在纽约州立大学布法罗分校获得计算机博士学位，是该校富布莱特学者。他还在斯坦福大学统计系获得数据挖掘及应用证书。

John Elder



Elder 博士就职于 Charlottesville，Virginia，Washington DC 和 Mountain View，California（www.datainglab.com）的数据挖掘咨询组。艾德研究公司成立于 1995 年，聚焦于联邦、商业、投资、安全领域的高级分析，包括文本挖掘、股票选择、图像识别、生物信息学、过程优化、交叉销售、药物功效、信用评分、风险管理和赝伪检测。ERI 公司已成为最大、最有经验的数据挖掘咨询企业。

Elder 于莱斯大学获得电气工程学士学位和电子电气硕士学位，在弗吉尼亚大学获得系统工程专业博士学位，他是该校数据挖掘课程的兼职教授。在赴 ERI 的前 15 年，他在航空国防咨询界工作了 5 年，在一家投资管理公司任职 4 年，在莱斯大学计算与应用数学系任职 2 年。

Elder 博士开发了创新性的数据挖掘工具，他本人是一位著名的特约报告人，也是在巴黎召开的“2009 Knowledge Discovery and Data Mining”会议的共同主席。他在许多大学、公司和政府实验室教授的分析技术课程以明晰性和有效性著称。Elder 在一个由总统任命的委员会服务了 5 年——为国家安全作技术指导。他与 Bob Nisbet、Gary Miner 合著的面向实际工作者的获奖书籍——《统计分析与数据控据应用手册》于 2009 年 5 月出版。

致　谢

感谢对该计划的构想和完成有贡献的人士。Seni 有幸能经常性地会晤 Jerry Friedman 以讨论集成背后的统计学概念。Friedman 教授的影响深远。Bart Goethels 和 ACM-KDD07 的组织者率先对我们关于该主题的讲座提议表示了欢迎。Tin Kam Ho 仔细审阅了本书的架构，Keith Bettinger 对本书手稿提出了许多有益的建议，Matt Strampe 提供了 R 编码方面的帮助。Morgan & Claypool 的工作人员——尤其是执行编辑 Diane Cerra 在将手稿转化为书籍方面勤勉而耐心。最后，感谢家人对我们的关怀与帮助。

特将本书献给我们的父亲——Tito 和 Flecher。

Giovanni Seni 和 John Elder

2010 年 1 月

译 者 序

江流天地外，山色有无中。

——王维

一句诗，为什么带来无尽的遐想？因为它诗中有画、诗画集成。

在广阔的科学研究领域中，集结人类在解决各类问题中的优秀思想理论方法是行之有效的问题解决之道，并不断在实践中得到印证。在信息大类的诸多学科中，集成理论与方法、集成学习等也被寄予厚望并已大显身手。

本书是两位美国学者于2010年出版的《Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions》的中译本。书中讨论了基于决策树的集成，对被认为是当前现代集成算法高级性能的关键原因之一的正则化问题进行了研究，就集成方法领域两个新进的发展——重要性采样（IS）和规则集成（RE）做了比较深入的分析，论述了新数据信息的集成在复杂性和更高精度方面的悖论等重要命题。

毋庸讳言，相对于人类面临的层出不穷、纷繁复杂的问题世界而言，本书所包含的内容仅为冰山一角、沧海一粟。其问题解决途径可能也只是挂一漏万。然而不应否认：集成方法的火花，往往触发问题求解领域中方法突破方面的燎原之势！

本书的出版得到国家自然科学基金（71371148）和广西科学研究与技术开发计划2011年度项目的资助与支持；武汉理工大学系统科学与工程研究中心的笪东璇、余弘道、刘聪然、舒新卿、陶崇园等同学对本书的出版做出了不可或缺的贡献，谨此诚致谢意！

限于译者水平，书中疏漏之处在所难免，欢迎广大读者批评指正。

王 攀

2014年8月

原书序一

Elder是统计预测领域的著名专家。他也是指导我从许多复杂数据中寻求有用信息的一位好友。我极其有幸与Elder在多个项目中合作，而集成向来都厥功至伟必定是一条好的理由。

下面描述我们是如何相遇的以及集成是如何可信赖的。我在弗吉尼亚大学花了四年时间研究市场。我的计划是毕业后成为一名投资经理。适合本人技能和个性的有益技术形式是我的全部所需（仅此而已）。1991年毕业后，我度过了一个数据引导下的特殊的、以咖啡因作为能量来源的不分昼夜的时期。在疯狂的“试凑”头脑风暴的磨合中，我误打误撞地产生了这样的概念：从广而分散的基预测模型组中创建一个“超级模型”。

在对投资管理进行10年组合模型的研究以后，我决定探究在一般的学术工作中，我的理念何处可行。我在华尔街做自营交易员一段时期后回到Charlottesville，并且寻求该领域的本地专家。

我在网上发现了Elder的公司——艾德研究公司，并且希望他们能有时间与一个数据挖掘方面的新手交流。我随即发现Elder不仅是统计学习方面的著名专家，还是一名非常成功的使这些方法得以流行的演说家。幸运的是，他有兴致讨论预测问题并聆听了我的想法。早先，他就向我提出对于投资的多模型方法用统计预测术语表述，即“集成”。

Elder和我在过去的10年一直共同从事我们感兴趣的项目。2001年，我和艾德研究公司组队一起竞争KDD杯。为了政府立项资助“基于集成的研究和软件”的创意，我们写了一份详细的计划书。2007年，在Netflix大奖赛上我们联合组队与数千个其他团队竞争，并获得了（一个项目的）第三名（这要部分地感谢简单集成）。我们甚至彻夜冥思苦想编出用户评估模型，这带来了多年前原始突破的迷人记忆。

集成的应用实践是数量可观的。它们的大部分实践十分初级，本书将明确地提升这一现实技术水平。Seni 对前沿研究与透彻掌握以及 Elder 的实践经验已联合创造出一部极富可读性的有用专著。

之前，我认为让用户无缝构建集成的软件必将诞生，这与专业建筑师使用 CAD 软件创建设计蓝图的方式一样。我期待 Seni 和 Elder 将（长）立于该领域的发展前沿。如果幸运，我也将涉身其中。

Jaffray Woodriff

Quantitative Investment Management 公司首席执行官

Charlottesville，Virginia

2010 年 1 月

（英文版）编辑注：Woodriff 先生的投资公司已持续正面增长，也已发展成为美国东南部最大的对冲基金管理者。

原书序二

人们发现对于挑战性任务卓有成效的解决方案往往来源于专家的集成。然而，对于一个复杂分类任务的算法求解，集成方法的效用第一次被见证却迟至20世纪80年代后期，这时的计算能力开始支持对丰富的分类方法集合的同步探索和应用。接下来的20年见证了越来越多应用于研究领域的此类方法，以及几种面向集成生成和组合的持久成功策略的发展。如今，虽然对所有要素的完美解释仍然难以捉摸，但是集成方法已成为统计学不可或缺的工具。涉及预测分类问题的每一个研究者和实践者，在很好地理解此方法学所展示的内涵后都能获益良多。

该书作者Seni和Elder对该主题进行了及时、简明的介绍。在对预测学习中的核心观点进行直观阐述之后，通过一条捷径将读者带入流行的基于树的集成创造策略的核心，然后以精炼而清晰的语言介绍统计学前沿的发展，通过传统的统计理论和方法，作出积极尝试来解释和采掘集成的奥秘。贯穿全书，这种方法是通过不同的现实生活中的实例来进行阐释的，并且面向读者用R代码实现，从而获得第一手经验以扩充（方法的阐释效果）。对于实际工作者，这个方便的文献开启了一扇门，（使人们）对极可望解决其面临挑战任务的这一丰富的工具集有很好的理解。对于研究人员和学生，它提供了一个（包含）大量密切相关文献的简明框架，并且充当了这一重要主题的优秀总结。

集成方法的发展绝没有止步。在有趣的开放性挑战中，有如何更透彻地理解数学结构、可应用性的具体情况的映射，寻找可扩展性和解释性的实现、处理不完整或不平衡的训练样本，以及适应环境的变化演化模型。看到该书在未来几十年里将激励智者着手解决这些问题，是何等令人兴奋啊！

Tin Kam Ho

贝尔实验室，Alcatel-Lucent

2010年1月

摘　要

在过去的十年中，集成方法被称为数据挖掘和机器学习领域最具影响力的发展方向。集多个模型于一体往往比最好的单个组件更精确。从投资时机把握到药物发现，从伪造检测到推荐系统等工业挑战（在这些领域，预测精度比模型可解释性更重要），集成都能给以重要的提升。

应用全部建模算法进行集成有成效，但为了给出最直观的解释，本书聚焦于决策树。在描述树的优缺点之后，作者对正则化（当前被认为是现代集成算法性能优良的关键因素）进行概述。随后本书清晰地描述近年来的两个发展方向：重要性采样（IS）和规则集成（RE）。IS 展现了经典集成方法——bagging、随机森林和 boosting，它们是单个算法的特例，揭示如何改善其精度和速度。RE 是源于规则树集成的线性规则模型。它们是集成的最可解释版本，对于如信誉评分和故障诊断等应用是基本要素。最后，解释集成对于新数据在复杂性（显然复杂得多）和更高精度方面的悖论。

本书面向初学者、科研人员和工程技术人员（尤其在工程、统计和计算机科学领域）。那些对集成了解甚少者将掌握为何及如何使用这一突破性方法，工程技术人员将收获构建甚至更强大模型的洞见。书中提供了用 R 完成的代码片段以举例说明算法及鼓励读者尝试使用该技术[①]。

本书作者是数据挖掘和机器学习领域的资深专家，也是兼职教授和颇受欢迎的演说家。虽然他们是早期发现并应用集成的先驱，但是在这里也精炼和阐释了学术带头人（如 Jerome Friedman）新近的突破性工作，以使得实践者受裨益。

---

① R 是通过 Comprehensive R Archive Network（CRAN）用于数据分析和统计建模的开源语言和环境。R 系统库包为许多计算平台提供了大量的功能，下载地址 http://cran.r-project.org/。CRAN 网站既有教程和综合性文档，也有不少优秀的介绍性读本；在此推荐 Peter Dalgaard 的 Introductory Statistics with R，W.N.Venables 和 B.D.Ripley 合著的 *Modern Applied Statistics*。

作者欢迎读者针对本书提出改进意见，可发邮件至 seni@datamininglab.com 和 elder@datamininglab.com。勘误和更新见 www.morganclaypool.com。

目　录

译者序
原书序一
原书序二
摘要
第 1 章　集成发现 …… 1
1.1　建立集成 …… 5
1.2　正则化 …… 6
1.3　现实世界中的实例：信用评分+网飞挑战 …… 7
1.4　本书的组织架构 …… 8
第 2 章　预测学习和决策树 …… 10
2.1　决策树归纳纵览 …… 14
2.2　决策树的性能 …… 16
2.3　决策树的缺陷 …… 17
第 3 章　模型复杂度、模型选择和正则化 …… 19
3.1　什么是树的“合适”规模 …… 19
3.2　偏差-方差分解 …… 20
3.3　正则化 …… 23
3.3.1　正则化与成本-复杂度树修剪 …… 23
3.3.2　交叉验证 …… 24
3.3.3　运用收缩的正则化 …… 26
3.3.4　通过构建增量模型的正则化 …… 30
3.3.5　实例 …… 31
3.3.6　正则化综述 …… 34
第 4 章　重要性采样和经典集成方法 …… 36
4.1　重要性采样 …… 39
4.1.1　参数重要性测度 …… 40
4.1.2　扰动采样 …… 42
4.2　泛化集成生成 …… 42
4.3　Bagging …… 44

4.3.1 实例 …… 47
4.3.2 为什么 Bagging 有用 …… 51
4.4 随机森林 …… 51
4.5 AdaBoost …… 53
4.5.1 实例 …… 54
4.5.2 为什么使用指数损失 …… 56
4.5.3 AdaBoost 的总体最小值 …… 57
4.6 梯度 Boosting …… 58
4.7 MART …… 59
4.8 并行集成与顺序集成的比较 …… 59
第 5 章 规则集成和解释统计 …… 61
5.1 规则集成 …… 61
5.2 解释 …… 63
5.2.1 仿真数据实例 …… 64
5.2.2 变量重要性 …… 68
5.2.3 偏相关 …… 69
5.2.4 交互统计 …… 70
5.3 制造业数据实例 …… 70
5.4 总结 …… 74
第 6 章 集成复杂性 …… 75
6.1 复杂性 …… 75
6.2 广义自由度 …… 77
6.3 实例：带有噪声的决策树表面 …… 78
6.4 广义自由度的 R 代码和实例 …… 82
6.5 总结与讨论 …… 83
参考文献 …… 85
附录 A AdaBoost 与 FSF 程序的等价性 …… 90
附录 B 梯度 Boosting 和鲁棒损失函数 …… 93

第1章 集成发现

...and in a multitude of counselors there is safety*.

Proverbs 24：6b

从数据中归纳模型，可找到大量经典方法，而且其处理能力各具特色。流行算法的精度依赖所处理问题的细节，如图 1.1 所示（Elder 和 Lee（1997）），该图揭示了五种算法用于六个公共领域问题的样本外相对误差。总体来说，神经网络模型对这些问题表现最优，但需指出，每种算法都在六个数据集中的至少两个上表现最优或次优。

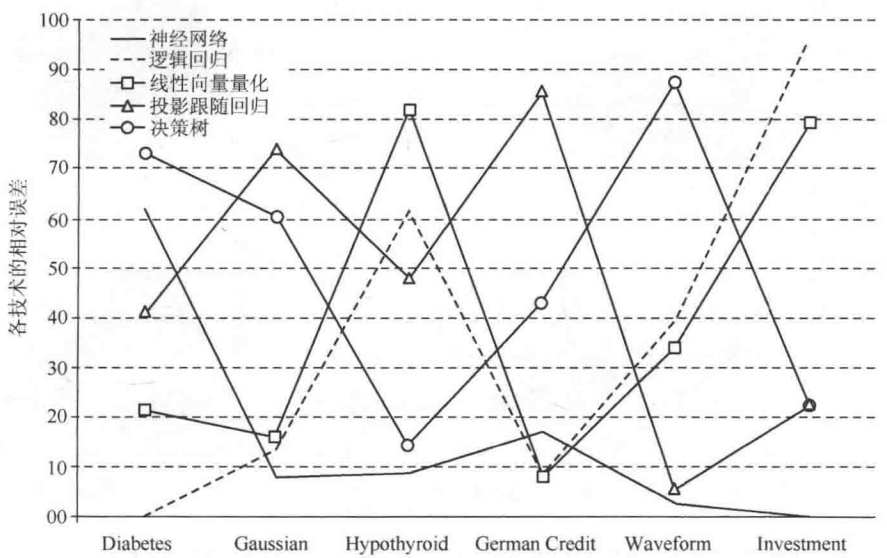


图 1.1 五种算法用于六个公共领域问题的非样本外相对误差（基于 Elder 和 Lee（1997））

* 原文为（还有别的版本）：For by wise counsel thou shalt make thy war：And in multitude of counselors there is safety. 译为：战中取胜凭智谋，群策群力保安宁。

——译者注

对于给定问题哪种算法表现优异？Michie 等（1994）对该问题展开了研究，他们开展了与前述问题相似但更多的工作——23 种算法用于 22 个数据集，并且基于给定的数据集性能构建决策树来预测最优算法[①]。虽然该研究侧重于树——在 23 个算法中占 9 个，而且几个数据集对树易产生不可控阈值但仍然为算法选择提供了有益指导。

然而，还有一种提高模型精度的方式比选择单一模型更容易且效果更显著：将诸模型集成。图 1.2 展示了图 1.1 中模型以四种不同方式集成的非样本精度，这些集成方式包括平均法、投票法和顾问感知器（Elder 和 Lee，1997）。对于每个问题，顾问感知器集成技术都优于简单的平均法，而与集成和单一模型相比，其差别很小。这里每种集成方法都比单一算法更有效。

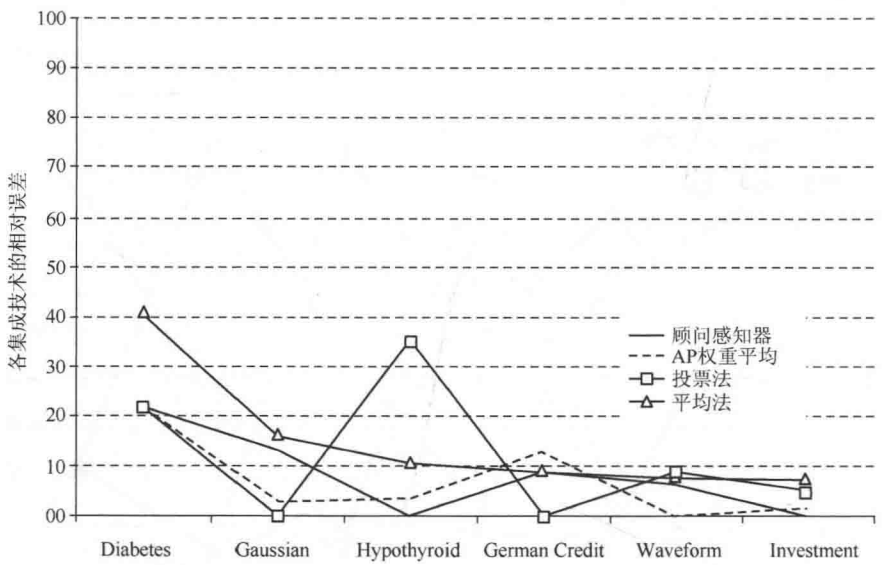


图 1.2　四种集成方法在图 1.1 问题上的非样本相对误差（基于 Elder 和 Lee（1997））

① 研究者（Michie 等，1994，10.6 节）曾检验一种算法的结果并且构建了一种 C4.5 决策树（Quilan，1992）以分离那些算法可应用的（算法在最佳算法的容限内）数据集到其所不能用的数据集中。他们还从树模型中提取了规则并应用专家系统来判定冲突规则以最大化“信息分数”网。这本书的网址为 http：//www.amsta.leeds.ac.uk/charles/statlog/whole.pdf。

这一现象被少数人分别同时发现，他们应用决策树（Ho，Hull 和 Srihari，1990）、神经网络（Hansen 和 Salamon，1990）或数学理论（Kleinberg，1990）来改善分类。最具影响的早期发展是 Breiman（1996）的 Bagging，Freund 和 Shapire（1996）的 AdaBoost，这些将在第 4 章予以描述。

在努力从回声定位信号特征中预测蝙蝠分类时我们偶然见识了集成（当时称为“模型融合”或“捆拢”）的威力（Elder，1996b）[①]。用几个非常不同的算法，如决策树、神经网络、多项式网络和最近邻（见 Nisbet 等（2009）对算法的描述）中的每一个都建立了最优模型。这些方法使用不同的基函数和训练程序，使它们具有不同的表现形式（图 1.3），还常产生意想不到的不同预测向量（即使在集成性能非常相似时）。

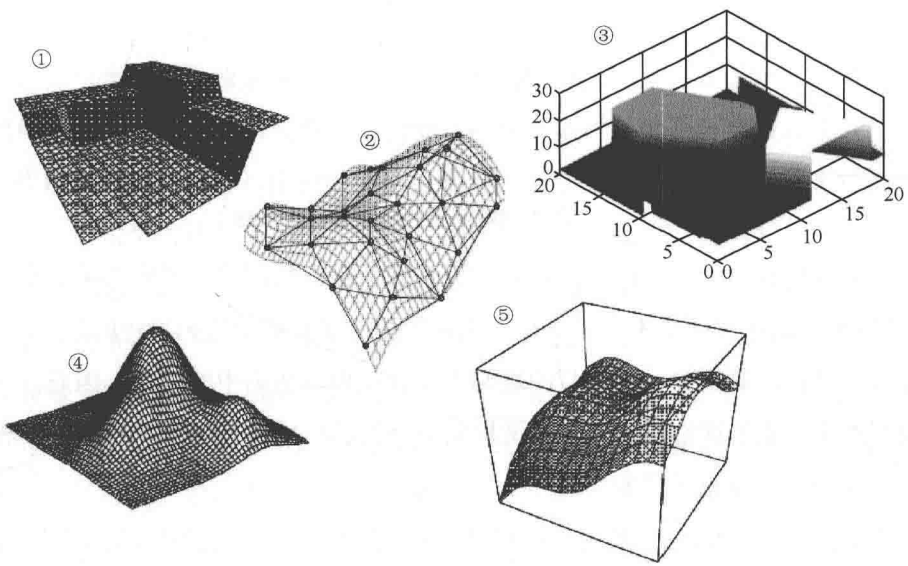


图 1.3 五种建模算法的估计表面实例

①为决策树，②为 Delaunay 规划（基于 Elder（1993）），③为最近邻，④为多项式网络，⑤为神经网络

该计划是只用蝙蝠的叫声来非侵害性地对其物种分类。伊利诺伊大学香槟分校（UIUC）的生物学家捕捉了 19 只蝙蝠，将其分别标注为 6 个物种之一，然后

① 感谢伊利诺伊大学香槟分校的 Doug Jones 及其电子电气专业学生的合作。