中国计算机学会学术著作丛书
——知识科学系列 4

机器学习及其应用

王珏　周志华　周傲英　主编

清华大学出版社

中国计算机学会学术著作丛书
——知识科学系列 4

机器学习及其应用

王珏　周志华　周傲英　主编

清华大学出版社
北京

内 容 简 介

机器学习是计算机科学和人工智能中非常重要的一个研究领域，近年来，机器学习不但在计算机科学的众多领域中大显身手，而且成为一些交叉学科的重要支撑技术。本书邀请国内外相关领域的专家撰文，以综述的形式介绍机器学习中不同领域的研究进展。全书共分 13 章。第 1 章是关于机器学习的一个全局性综述。第 2 至第 6 章分别对统计学习、非监督学习、符号学习、强化学习和流形学习进行了综述，并穿插了作者的一些精彩工作。第 7 和第 8 章分别介绍了作者在集成学习和进化学习中某一具体话题上的研究成果。第 9 和第 10 章对数据挖掘中的一些问题进行了介绍和讨论。第 11 至第 13 章则对机器学习在模式识别、视频信息处理等领域的应用做了介绍。

本书可供计算机、自动化及相关专业的学生、教师、研究生和工程技术人员参考。

图书在版编目(CIP)数据

机器学习及其应用/王珏，周志华，周傲英主编. —北京：清华大学出版社，2006.3
(中国计算机学会学术著作丛书. 知识科学系列)
ISBN 7-302-12038-2

Ⅰ. 机…　Ⅱ. ①王…②周…③周…　Ⅲ. 机器学习　Ⅳ. TP181

中国版本图书馆 CIP 数据核字(2005)第 124726 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
责任编辑：薛　慧
印 装 者：清华大学印刷厂
发 行 者：新华书店总店北京发行所
开　　本：185×230　**印张**：21.5　**字数**：482 千字
版　　次：2006 年 3 月第 1 版　2006 年 3 月第1 次印刷
书　　号：ISBN 7-302-12038-2/TP·7792
印　　数：1～3000
定　　价：42.00 元

评审委员会

序

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想像。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步做出更大的贡献。

中国计算机学会

2002年6月14日

序　言

2002年秋天，由王珏教授策划和组织，复旦大学智能信息处理开放实验室（即现在的上海市智能信息处理重点实验室）举办了一次“机器学习及其应用”研讨会。该研讨会属于实验室的“智能信息处理系列研讨会”之一。十余位学者在综述机器学习各个分支的发展的同时报告了他们自己的成果。鉴于研讨会取得了非常好的效果，而机器学习领域又是如此之广阔，有那么多重要的问题还没有涉及或还没有深入，2004年秋天王珏教授又和周志华教授联合发起并组织第二届“机器学习及其应用”研讨会，仍由复旦实验室举办。这次研讨会又取得了非常好的效果，并且参加的学者比上次更多，报告的内容也更丰富。根据与会者的意见，决定把报告及相关内容编成一本书出版，以便与广大的国内学者共享研讨会的成果。

机器学习是人工智能研究的核心课题之一，不但有深刻的理论内蕴，也是现代社会中人们获取和处理知识的重要技术来源。它的活力久盛不衰，并且日呈燎原之势。对此，国内已经有多种定期和不定期的学术活动。本书的出版反映了机器学习界一种新型的华山论剑：小范围、全视角、更专业、更深入，可与大、中型机器学习会议互相补充。值得赞扬的是，它没有任何学派和门户之见，无论是强调基础的“气宗”，还是注重技术的“剑宗”，都能在这里畅所欲言，自由交流。我很高兴地获悉：第三届“机器学习及其应用”研讨会已经在2005年11月由周志华教授和王珏教授主持在南京大学成功举行。并且以后还将有第四届、第五届……。作为一直跟踪这项活动并从中获得许多教益的一个学习者，我真希望它发展成这个领域的一个品牌，希望机器学习的优秀成果不断地由这里飞出，飞向全世界。

值得一提的是王珏教授有一篇颇具特色的综述文章为本书开道。长期以来，许多有识之士为国内学术界缺少热烈的争鸣风气而不安。因为没有争鸣就没有学术繁荣。细心的读者可以看出，这篇综述的观点并非都是传统观点的翻版，并且很可能不是所有的同行都认同的。作者深刻反思了机器学习这门学科诞生以来走过的道路，对一些被行内人士几乎认作定论的观点摆出了自己的不同看法。其目的不是想推出一段惊世骇俗的宏论，而是为了寻求真理、辨明是非。在这个意义上，王珏教授也可算是一位“独孤求败”。如果有人能用充分的论据指出其中可能存在的瑕疵，他也许会比听到一片鼓掌之声更感到宽慰。

随着本书的出版，中国计算机学会丛书知识科学系列也正式挂牌了。在衷心庆贺这个系列诞生的同时，我想重复过去说过的一段话："二十多年来，知识工程主要是一门实验性科学。知识处理的大量理论性问题尚待解决。我们认为对知识的研究应该是一门具有坚实理论基础的科学，应该把知识工程的概念上升为知识科学。知识科学的进步将从根本上回答在知识工程中遇到过，但是没有很好解决的一系列重大问题"。本系列为有关领域的学者提供了一个宽松的论坛。衷心感谢王珏、周志华、周傲英三位编者把这本精彩的文集贡献给知识科学系列的首发式。我相信今后机器学习著作仍将是这个系列的一个常客。据悉，第四届机器学习研讨会将于今秋在南大举行，届时各种观点又将有进一步的发展和碰撞。欲知争鸣烽火如何再燃，独孤如何锐意求败，且看本系列下回分解。

陆汝钤

2006 年 1 月

前　言

2002年，复旦大学智能信息处理开放实验室（即现在的复旦大学上海市智能信息处理重点实验室）成立之时，陆汝钤老师建议实验室组织“智能信息处理系列研讨会”作为实验室的一项重要学术活动，并将“机器学习及其应用”列为当年支持的研讨会之一。在2002年11月的第一个周末，研讨会成功举行，并确定了会议不征文、不收费、报告人由组织者邀请，以及“学术至上，其他从简”的办会宗旨。

在2004年的研讨会上，两天半的会议保持100余人的旁听者，这令与会专家深受鼓舞，并商定从这次会议开始，将“机器学习及其应用”发展成为一个系列研讨会，在每年11月第一个周末举行。本书就是清华大学出版社为这次会议出版的文集。

随着各行各业大量数据的涌现，如何使得这些数据变为提高管理水平、发展产业效益与保障社会与信息安全的重要资源，成为当前重要且不得不解决的重要问题，这就需要分析或阅读这些数据。数据分析与机器学习是完成上述任务的重要途径，就“分析与阅读数据”而言，它们的目标是一致的，其区别仅仅是学者为了区分研究方法而使用了不同的术语。因此，一般地说，从解决“分析与阅读数据”的角度，我们可以对此不加区分，事实上，“机器学习及其应用”系列研讨会，同样欢迎数据分析的研究报告。

正是由于各行各业需求的推动，近几年，机器学习得到了学术界的充分重视，例如，2005年国际人工智能联合会议（IJCAI'05）收录的文章中将近一半或多或少地与机器学习研究有关，这与以往的国际人工智能联合会议上“机器学习”只有一两个分组会议有天壤之别。

目前，机器学习研究大致可以分为三种不同的途径：其一，将以往机器学习研究整理并上升为理论，例如，统计机器学习理论整理了感知机、Duda的统计模式识别理论等，Reduct理论整理了符号机器学习的各种方法，集成机器学习（ensemble）整理了各种局部模型的方法等。这类研究非常重要，假设空间、线性描述以及边缘与复杂性等均派生于此。这类研究至今还十分活跃，其中重要的结果将机器学习研究提升到一个新的高度。这个论文集并没有包含所有重要的研究结果，我们相信，在以后这个系列会议上，将有更多的研究报道。其二，近几年，各类机器学习范式层出不穷，几乎一两年就有一种范式流行起来，例如，多示例学习、Ranking学习、数据流学习等等。这类研究的特点是应用需求

驱动的，大多数范式的理论基础尚在发展之中，其中包括首先将其他已有的理论基础加以改造，使之适应面临问题的需要。这是一类重要的研究途径，也许其中某些范式将会发展出自己独特的理论，并成为独立的研究课题。应该指出的是，这类研究如此重要，因为它是理论提出与发展所必需的观察，因此，不过分地说，它是机器学习新理论、新技术产生之母。其三，如果将上述两类研究理解为机器学习研究的两个极端，则还存在介于两者之间的一类研究。说其介于两者之间，一方面，其起源完全来自实际问题的需要，而又不能完全纳入某个已有机器学习的理论框架，例如，关系学习，其来源是对关系数据库数据的学习，由于关系数据库无法表示为命题逻辑形式，人们不得不发展新的理论与方法。流形学习、强化学习以及半监督学习等均属于这类研究。有趣的是半监督学习，在前几年的研究中，这类范式的学习还应该属于第二类，近两年，人们发现这类范式的学习可以建立在谱流形上，并与“转导”问题联系在一起，这似乎建立了自己一套理论基础。对这类学习范式还有一个特点，就是尽管已经有了自己的理论，但是，还远远不够完善，在科学意义上，还需要雕琢，还需要进一步证明其价值。换句话说，目前发展的理论，还远远不及机器学习第一类的研究扎实，今天被人们热捧的理论，也许明天就被证明是行不通甚至是不重要的。当然，一旦被实践证明这类研究中的某个理论是有意义的，它将自然被归类于第一类研究。

机器学习的另一个重要趋势，是考虑给定数据集合自身的性质，1995 年出现的“没有免费午餐定理”，近几年得到机器学习研究者的重视，因为任何一般性的方法在面临非线性问题时，如果处理不当，不得不面对“维数灾难”问题，这个问题只有在理论框架下嵌入特定数据集合的特定性质才有可能解决。对这个问题，本论文集并没有仔细讨论，但是，在本系列 2005 年南京大学会议上，已有探讨。

本论文集所收集的每篇文章将讨论一个问题，并使用综述的形式，将报告人自己的研究合理地嵌入在之中，这是本系列会议的一个特殊要求，其目的是尽可能地全面反映机器学习的研究现状，并为同行提供一种观点与索引（请注意，本论文集的每篇文章绝不是一篇研究情况的报道，它们均反映了作者对所研究问题的观点，当然，这些观点并不能完全代表清华大学出版社、研讨会组织和主持者，以及本书（形式上和事实上的）编者及其他各章作者的学术观点）。但是，应该指出，这还不是一个对上述所有机器学习问题均涉及的论文集，一方面，有些机器学习研究范式还不够成熟，另一方面，则是我们的能力还有局限。

我们正在面临如此重大且困难的问题（网络信息、生物信息与金融经济信息），它们要求我们必须认真对待并有效解决。目前，包括科学、技术、安全、军事与金融经济等众多的领域均在关心机器学习的研究进展，这为机器学习研究者提供了大量的机会，机器学习的研究者正在进入激动人心的时代，因为他们的任何有意义的成果，就可能为社会与科学带来进步。有一利必有一弊，在面临如此困难问题的面前，机器学习研究者的危机同时出

现，指望通过写程序或改进已有结果的方式获得成功，已十分困难。不同领域的研究者正在悄然侵入我们的领地，以接替我们。应用者正在关注着这种新陈代谢，他们已经等待太久了，急不可耐了。

本书共分13章。第1章是关于机器学习的一个全局性综述。第2至第6章分别对统计学习、非监督学习、符号学习、强化学习和流形学习进行了综述，并穿插了作者的一些精彩工作。第7和第8章分别介绍了作者在集成学习和进化学习中某一具体话题上的研究成果。第9和第10章对数据挖掘中的一些问题进行了介绍和讨论。第11至第13章则对机器学习在模式识别、视频信息处理等领域的应用做了介绍。

需要说明的是，陈松灿教授、封举富教授和吴高巍博士在研讨会上曾做了精彩的报告，但遗憾的是，由于时间紧迫，他们的文章没有来得及收入本书。

最后，我们衷心感谢陆汝钤老师对这个系列会议一贯的指导与支持，没有陆老师的指导与支持，我们是不可能将这个系列会议办下去。我们也感谢复旦大学上海智能信息处理重点实验室对“机器学习及其应用'04”的支持，他们为组织这次会议作了大量卓有成效的工作。参加本论文集编写的作者感谢不同国家项目对他们研究的支持，没有这些项目的资助，这些研究者也无法完成这些研究。

编者

2006年1月

目　录