

中国计算机学会学术著作丛书
——知识科学系列 9

机器学习及其应用2011

周志华 杨 强 主编

清华大学出版社

需要全本请访问 www.ertongbook.com

中国计算机学会学术著作丛书——知识科学系列 9

机器学习及其应用 2011

周志华　杨　强　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，机器学习不仅在计算机科学的众多领域中大显身手，还成为一些交叉学科的重要支持技术。本书邀请国内外相关领域的专家撰文，以综述的形式分别介绍机器学习不同分支及相关领域的研究进展。全书共分 14 章，内容分别涉及因果推断、流形学习与降维、迁移学习、类别不平衡学习、演化聚类、多标记学习、排序学习、半监督学习等技术和协同过滤、社区推荐、机器翻译等应用，以及互联网应用对机器学习技术需求的探讨。

本书可供计算机、自动化及相关专业的研究人员、教师、研究生和工程技术人员参考。

版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

机器学习及其应用 2011/周志华，杨强主编. --北京：清华大学出版社，2011.11
(中国计算机学会学术著作丛书. 知识科学系列)
ISBN 978-7-302-26853-6

Ⅰ. ①机…　Ⅱ. ①周… ②杨…　Ⅲ. ①机器学习　Ⅳ. TP181

中国版本图书馆 CIP 数据核字(2011)第 187136 号

责任编辑：薛　慧
责任校对：赵丽敏
责任印制：

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：185×230　　**印　张**：17.25　　**字　　数**：371 千字
版　　次：2011 年 11 月第 1 版　　**印　　次**：2011 年 11 月第1次印刷
印　　数：1～　000
定　　价：　.00 元

产品编号：

评 审 委 员 会

名誉主任委员：张效祥

主 任 委 员：唐泽圣

副主任委员：陆汝钤

委　　员：（以姓氏笔画为序）

王　珊　　吕　建　　李晓明

林惠民　　罗军舟　　郑纬民

施伯乐　　焦金生　　谭铁牛

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象。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作出更大的贡献。

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”研讨会已经于 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京大学举行,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

机器学习致力于“利用经验来改善系统自身的性能”。在计算机系统中,“经验”通常是以数据的形式存在的,要利用经验就不可避免地要对数据进行分析,因此,机器学习已逐渐成为计算机数据分析技术的源泉之一。随着人类收集和存储数据能力的不断增长以及计算机运算能力的飞速发展,利用计算机来分析数据的要求越来越广泛、越来越迫切,从而使得机器学习的重要性越来越显著。机器学习不仅是人工智能的核心研究领域之一,目前还成为计算机科学中最活跃、最受关注的领域之一。

2002 年,陆汝钤院士在复旦大学智能信息处理实验室发起组织了“智能信息处理系列研讨会”,并将“机器学习及其应用”列为当年支持的研讨会之一。2002 年 11 月,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。2004 年 11 月,在复旦大学举行了第二次“机器学习及其应用”研讨会,两天半的会议一直有 100 余人旁听。2005 年起研讨会由南京大学软件新技术国家重点实验室举办。2005 年 11 月举办的第三届研讨会吸引了来自全国近十个省市的 250 余人旁听;2006 年 11 月、2007 年 11 月分别由南京航空航天大学信息科学与技术学院、南京师范大学数学与计算机学院协办了第四届和第五届研讨会,两次均吸引了来自全国十余个省市的约 300 人旁听;2008 年 11 月举行的第六届研讨会,适逢南京大学计算机学科建立五十周年,吸引了来自全国十余个省市的 380 余人旁听;此后在 2009 年 11 月和 2010 年 11 月在南京大学分别举行了第七、八届研讨会,均有约 400 人旁听。值得一提的是,为了促进研究生之间以及研究生与资深学者之间的交流,从 2006 年开始,在研讨会期间举行“机器学习及其应用学生研讨会”,由研究生通过墙展方式介绍自己的工作,到目前为止共举行了五次,先后吸引了 100～300 余人参加。

清华大学出版社对推介信息科学技术领域的研究进展一直抱有极大的热情。早在“第二届机器学习及其应用研讨会”举行时清华大学出版社就参与其中,并为该研讨会专门出版了文集,即 2006 年发行的《机器学习及其应用》一书。2005 年第三届研讨会期间,清华大学出版社和与会专家商定,以后每两届研讨会的部分内容将编成一书,以“机器学习及其应用:出版年”的形式冠名。第三至六届研讨会的部分内容已在《机器学习及其应用 2007》以及《机器学习及其应用 2009》中出版发行。

本书是清华大学出版社邀请第七届和第八届“机器学习及其应用研讨会”的部分专家将其报告内容总结成文而得的文集。书中每一章将讨论一个论题，以综述的形式对该方面的研究进展加以介绍，并将报告人自己的一些研究工作嵌入其中。书中章节不仅涉及因果推断、聚类分析、维数削减等传统研究领域，还涉及流形学习、半监督学习、多标记学习等新领域，以及计算语言学、协同过滤、互联网应用等。需要注意的是，书中各章的内容仅表达该章作者本人的见解，并不代表清华大学出版社、编者及其他各章作者的学术观点。本书的出版得到了陆汝钤院士的支持和指导，并得到清华大学出版社计算机专著出版基金的资助，在此谨表示衷心的感谢。

编　者

2011 年 6 月

目　录

因果推断的可分解性和可传递性问题 …… 耿　直　1
1　引言 …… 1
2　图模型结构学习的可分解条件 …… 2
3　直接作用和间接作用 …… 3
3.1　基于关联模型的直接作用与间接作用 …… 4
3.2　基于因果模型的主分层直接作用 …… 4
3.3　控制的和自然的直接作用 …… 6
4　因果作用的可传递性问题 …… 7
5　讨论 …… 11
参考文献 …… 11
机器学习的几何观点 …… 何晓飞　14
1　引言 …… 14
2　监督学习、半监督学习与无监督学习 …… 15
3　基于几何拓扑的降维算法 …… 17
3.1　流形降维 …… 17
3.2　几何和拓扑 …… 18
3.3　保局投影 …… 20
4　主动学习和半监督学习:基于几何的观点 …… 23
5　结束语和展望 …… 29
参考文献 …… 30
协同过滤与链接预测的迁移学习问题 …… 李　斌　朱兴全　杨　强　33
1　引言 …… 33
1.1　问题背景 …… 33
1.2　相关研究工作综述 …… 35
2　基于矩阵分解的潜在特征空间共享 …… 36
2.1　组级评分矩阵共享 …… 37
2.2　项目潜在特征共享 …… 40
3　协同过滤的迁移学习 …… 41
3.1　评分矩阵生成模型 …… 41

3.2 实验结果 …… 43
4 链接预测的迁移学习 …… 45
4.1 集体链接预测模型 …… 45
4.2 实验结果 …… 47
5 结语 …… 49
参考文献 …… 49
LDA 的并行化运算及其应用 …… 李友林 51
1 引言 …… 51
2 LDA 算法介绍 …… 52
3 LDA 算法的并行化——PLDA …… 54
4 LDA 算法的进一步并行化——PLDA+ …… 57
5 AdHeat 算法——PLDA 在社区推荐中的应用 …… 61
6 结束语 …… 65
参考文献 …… 66
关于二类模式分类问题的分解 …… 吕宝粮 赵 海 67
1 引言 …… 67
2 最小最大模块化网络 …… 68
2.1 问题分解 …… 69
2.2 模块集成 …… 70
3 高斯零交叉函数最小最大模块化网络 …… 71
3.1 高斯零交叉函数 …… 71
3.2 高斯零交叉函数最小最大模块化网络的特点 …… 73
3.3 与其他分类器的关系 …… 75
4 大规模二类问题的分解策略 …… 76
4.1 随机分解 …… 76
4.2 超平面分解 …… 77
4.3 聚类分解 …… 78
4.4 基于先验知识的分解 …… 79
5 大规模不平衡专利数据分类 …… 80
5.1 实验数据 …… 80
5.2 最小最大模块化 Liblinear …… 82
5.3 性能评价指标 …… 82
5.4 Section 层上 A 类为正类的二类问题实验 …… 83
5.5 Section 层上的全部二类问题实验 …… 84
6 结论 …… 85

参考文献 …… 86
面向降维的图构建技术 …… 乔立山　张丽梅　陈松灿　89
1　引言 …… 89
2　降维与图构建 …… 90
2.1　降维技术 …… 90
2.2　图及其构建技术 …… 92
3　稀疏表示建图与稀疏保持投影 …… 93
3.1　稀疏表示建图 …… 94
3.2　稀疏保持投影 …… 97
4　面向降维的图优化 …… 98
4.1　图优化的局部保持投影 …… 99
4.2　降维与建图的联合学习框架 …… 101
5　结论 …… 103
6　参考文献 …… 104
统计词对齐 …… 王海峰　刘占一　吴　华　108
1　引言 …… 108
2　机器翻译简介 …… 109
2.1　基于规则的机器翻译 …… 109
2.2　基于实例的机器翻译 …… 109
2.3　统计机器翻译 …… 110
3　双语词对齐 …… 110
3.1　IBM 模型 …… 111
3.2　EM 算法在词对齐中的应用 …… 114
3.3　解码算法 …… 115
3.4　基于 HMM 的统计词对齐模型 …… 116
3.5　其他机器学习方法 …… 118
3.6　双语词对齐评价方法 …… 119
4　单语词对齐与搭配抽取 …… 119
4.1　搭配简介 …… 119
4.2　单语词对齐与双语词对齐的类比 …… 120
4.3　基于单语词对齐的搭配抽取 …… 120
4.4　实验 …… 123
5　利用搭配提高双语词对齐质量 …… 126
5.1　搭配概率 …… 127
5.2　提高 IBM 模型 …… 127

5.3 提高双向词对齐 …… 128
5.4 实验 …… 129
6 讨论与总结 …… 133
参考文献 …… 133
概念、相似性与聚类分析 …… 于 剑 136
1 引言 …… 136
2 相似性与概念 …… 137
3 相似性计算模型 …… 139
3.1 样例相似性计算模型 …… 139
3.2 原型理论下的相似性计算公式 …… 142
3.3 相似性的融合 …… 143
4 结束语 …… 143
参考文献 …… 144
互联网行业对机器学习和其他计算技术的需求 …… 岳亚丁 147
1 引言 …… 147
2 互联网行业现状 …… 147
2.1 互联网企业收入模型 …… 147
2.2 数据计算任务 …… 149
2.3 典型做法 …… 151
3 对计算技术的需求 …… 156
3.1 解决方案框架 …… 156
3.2 并行化算法 …… 157
3.3 其他难题 …… 157
4 小结 …… 161
参考文献 …… 161
基于指数族混合模型的在线式演化聚类算法 …… 张见闻 张长水 162
1 引言 …… 162
1.1 问题背景 …… 162
1.2 相关研究工作综述 …… 165
2 指数族混合模型 …… 167
3 从密度估计的观点看聚类问题 …… 168
4 基于指数族混合模型的演化聚类算法 …… 169
4.1 历史数据相关的途径 …… 170
4.2 历史模型相关的途径 …… 171
5 实验 …… 174

6　结语 …… 176
参考文献 …… 177
多标记学习 …… 张敏灵　周志华　179
1　引言 …… 179
2　学习框架 …… 181
2.1　问题定义 …… 181
2.2　评价指标 …… 182
3　学习算法 …… 185
3.1　算法分类 …… 185
3.2　“问题转换”算法 …… 185
3.3　“算法适应”算法 …… 189
4　结束语 …… 193
参考文献 …… 195
Ranking on Large-scale Graphs with Rich Metadata …… Bin Gao　Tie-Yan Liu　200
1　Introduction …… 200
2　Unsupervised Ranking on Large-scale Graphs with Metadata …… 201
2.1　General Framework for Unsupervised Graph Ranking …… 203
2.2　Unsupervised Graph Ranking Algorithms …… 207
2.3　Further Discussions on the Unsupervised General Framework …… 211
3　Supervised Ranking on Large-scale Graphs with Metadata …… 213
3.1　General Framework for Supervised Graph Ranking …… 214
3.2　Supervised Graph Ranking Algorithms …… 216
4　Summary …… 219
References …… 220
Semi-supervised Learning with Mixed Unlabeled Data
…… Haiqin Yang　Kaizhu Huang　Zenglin Xu　Irwin King　Michael R. Lyu　222
1　Introduction …… 222
2　Literature Review …… 224
3　Model …… 228
3.1　Notation and Problem Definition …… 228
3.2　Framework …… 229
3.3　Properties …… 230
4　Solution and Algorithms …… 231
4.1　Semi-definite Programming Transformation …… 232
4.2　Concave Convex Procedure (CCCP) …… 232

5 Experiment …… 235
5.1 Setup …… 235
5.2 Results on Real World Datasets …… 235
5.3 Efficiency …… 236
6 Conclusion …… 237
References …… 238
Learning with Local Consistency …… Chiyuan Zhang Deng Cai 243
1 Motivation …… 243
2 Problem Formulation …… 244
2.1 Local Consistency Assumption …… 244
2.2 Regularization Framework …… 246
2.3 Graph Representation and Construction …… 247
3 Algorithms …… 248
3.1 LapRLS and LapSVM …… 248
3.2 Graph Regularized Non-negative Matrix Factorization …… 249
3.3 Locally-consistent Topic Modeling …… 251
3.4 Gaussian Mixture Model with Local Consistency …… 253
4 Conclusion …… 255
References …… 256

1 因果推断的可分解性和可传递性问题

耿　直
北京大学 数学科学学院，北京　100871

理性的最高成就是引起了人们对其有效性的怀疑。
（乌纳穆诺（Miguel de Unamuno），1864—1936）

自提出替代指标悖论以来，它成了我心中一片飘逝不去的阴云，使我对因果作用的统计结论的可传递性产生了怀疑。我想借此机会，向大家介绍一下这个困扰我的问题。

1　引言

基于还原论的方法将一个大规模的系统分解为若干个小规模的系统，再对每个小系统进行局部研究，然后，将局部研究的结论汇总后综合得出大系统的结论。但是，一个大系统的研究问题是否能分解为小系统的研究问题？如果分解的话，需要什么条件？本文针对因果推断问题，探讨因果路径上变量之间的因果作用问题。例如，我们希望研究一个系统中三个变量 T，S 和 Y 之间的因果机制。我们是否能将其分解为两个子系统，一个包含 T 和 S，另一个包含 S 和 Y，分别研究各子系统的因果机制，然后组建 T，S 和 Y 之间的因果机制？这种分解需要什么条件？

采用变量选择的手段将一个高维空间的统计问题降到低维空间解决，需要考虑两个问题：可压缩性（collapsibility）和可分解性（decomposability）。可压缩性的意思是将高维数据经变量筛选降到低维空间后能保证得到与高维空间相同的统计结论[1,3,20,23,27]。我们曾经探讨过关于离散数据、连续数据、混合数据和一般分布情况进行统计推断的可压缩问题[8,10,11,15]。可分解性的意思是将一个高维的问题分解到若干变量子集的低维空间分别进行分析，然后将这些低维的结论汇总，最终解决高维的问题[5,7,13,24]。可分解性与可压缩性相互联系，高维问题分解为若干低维问题，希望在分解后的每个低维空间能得到无混杂偏倚的推断结论，即可压缩性。然后将所有低维空间得到的无混杂偏倚的结论进行汇总，得

到高维空间的结论，即可分解性。本文将探讨因果推断的可分解性。在 $X_1 \to X_2 \to \cdots \to X_p$ 的一条因果路径上，是否可以通过分解，得到各个局部 X_i 对 X_{i+1} 的因果作用，然后综合这些局部因果作用得到 X_1 对 X_p 的因果作用？Pearl 定义了因果路径的作用[16]。VanderWeele 和 Robins 提出了利用带正负号的因果网络判断因果作用正负号的方法[22]。他们的方法需要已知一个完整的因果网络。但是在实际应用中，也许很难获得一个完整网络的知识。

在医学临床试验中，称研究的目标变量为结局指标(endpoint)。当这个结局指标在实际研究中难以获得时，常常用一个容易获得的指标替换这个结局指标，称为替代指标(surrogate endpoint)。例如，一个评价预防更年期妇女骨折药物的临床试验，因为观测妇女是否今后会出现骨折需要等待很长时间，因此，将骨密度作为结局指标“骨折”的一个替代指标，通过该药物是否能增加骨密度来评价对于骨折的预防效果。替代指标的方法不仅应用于医学，也广泛应用于自然科学、经济学和人文社会科学的研究。很多学者提出了确定替代指标的准则[2,4,13,14,18,25]。这些准则都有一个共同的特点，就是要求一个替代指标应该是从原因通向结局的因果路径上的变量，并且希望这个替代指标能够在各种意义上切断原因到结局的路径。我们曾指出可能会出现治疗对替代指标有正作用，并且替代指标对结局指标也有正作用，但是治疗对结局指标可能有负作用的替代指标悖论现象[2]。这里的正作用是指总体分布意义上的因果作用。替代指标悖论意味着在总体分布意义上统计推断的结论不具有可传递性。本文将进一步指出，即使总体中每一个个体的替代指标对结局都有严格正的个体作用，由治疗对替代指标的统计结论也不能预测治疗对结局的作用。关于如何综合分析因果路径上变量之间的因果作用，文献[2,12,22]提出了利用变量之间有单调性关系的先验知识进行定性的评价方法。但是，即使观测到因果路径上所有变量的联合数据，这些先验知识也是不可检验的。本文的目的不是探讨如何进行因果路径上的统计推断方法，而是试图论述统计结论不具有可传递性。

本文第 2 节介绍因果网络结构学习的分解方法，在条件独立的情况下，可以将整体的结构学习分解为局部的结构学习。接下来的几节探讨评价因果作用的分解问题。首先，在第 3 节介绍了目前已经提出的几种直接作用的概念。然后，第 4 节探讨在任何意义上都没有直接作用的情况下，也很难采用分解方法进行因果作用的评价，由分解后的局部因果作用难以综合得出整体的因果作用。最后，我们提出如何综合统计推断结论是一个具有挑战性的问题，希望这个问题能得到统计学者和各科学领域研究者的关注。

2 图模型结构学习的可分解条件

令 V 表示包含所有变量的集合，将其划分为三个互不相交的子集 A,B,C。很多人研究过无向图模型的估计和模型选择的分解算法[5,6,9]。如果给定 C 的条件下 A 与 B 相互独立并且 C 上是一个完全无向子图(称为强分解条件)，那么，我们可以分别计算两个子图模型