

机器学习理论及应用

Machine Learning Theory and Its Applications



李凡长 钱旭培 著
谢琳 何书萍

中国科学技术大学出版社

当代科学技术基础理论与前沿问题研究丛书

中国科学技术大学
校友文库

机器学习理论及应用

Machine Learning Theory and Its Applications

李凡长 钱旭培 著
谢琳 何书萍

中国科学技术大学出版社

内 容 简 介

机器学习新方法研究是实现机器学服务各行各业的历史使命. 根据这样的宗旨, 本书系统地介绍了李群机器学习、动态模糊机器学习、Agent 普适机器学习和贝叶斯量子随机学习, 共三篇十二章内容.

本书可供计算机科学技术、认知科学、机器学习、人工智能、控制技术领域的高年级本科生和研究生作为教科书或参考书, 也可供高校教师、科研院所的研究人员使用.

图书在版编目(CIP)数据

机器学习理论及应用/李凡长等著. —合肥: 中国科学技术大学出版社, 2009. 10

(当代科学技术基础理论与前沿问题研究丛书: 中国科学技术大学校友文库)

“十一五”国家重点图书

ISBN 978-7-312-02636-2

I. 机… II. 李… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字(2009)第 182037 号

出版发行 中国科学技术大学出版社

地址 安徽省合肥市金寨路 96 号, 230026

网址 <http://press.ustc.edu.cn>

印 刷 合肥晓星印刷有限责任公司

经 销 全国新华书店

开 本 710mm×1000mm 1/16

印 张 30.75

字 数 533 千

版 次 2009 年 10 月第 1 版

印 次 2009 年 10 月第 1 次印刷

印 数 1—2000 册

定 价 88.00 元

总 序

侯建国

(中国科学技术大学校长、中国科学院院士、第三世界科学院院士)

大学最重要的功能是向社会输送人才,大学对于一个国家、民族乃至世界的重要性和贡献度,很大程度上是通过毕业生在社会各领域所取得的成就来体现的。

中国科学技术大学建校只有短短的五十年,之所以迅速成为享有较高国际声誉的著名大学之一,主要就是因为她培养出了一大批德才兼备的优秀毕业生。他们志向高远、基础扎实、综合素质高、创新能力强,在国内外科、经济、教育等领域做出了杰出的贡献,为中国科大赢得了“科技英才的摇篮”的美誉。

2008年9月,胡锦涛总书记为中国科大建校五十周年发来贺信,信中称赞说:半个世纪以来,中国科学技术大学依托中国科学院,按照全院办校、所系结合的方针,弘扬红专并进、理实交融的校风,努力推进教学和科研工作的改革创新,为党和国家培养了一大批科技人才,取得了一系列具有世界先进水平的原创性科技成果,为推动我国科教事业发展和社会主义现代化建设做出了重要贡献。

据统计,中国科大迄今已毕业的5万人中,已有42人当选中国科学院和中国工程院院士,是同期(自1963年以来)毕业生中当选院士数最多的高校之一。其中,本科毕业生中平均每1000人就产生1名院士和七百多名硕士、博士,比例位居全国高校之首。还有众多的中青年才俊成为我国科技、企业、教育等领域的领军人物和骨干。在历年评选的“中国青年五四奖章”获得者中,作为科技界、科技创新型企业界青年才俊代表,科大毕业生已连续多年榜上有名,获奖总人数位居全国高校前列。鲜为人知的是,有数千名优秀

毕业生踏上国防战线,为科技强军做出了重要贡献,涌现出二十多名科技将军和一大批国防科技中坚。

为反映中国科大五十年来人才培养成果,展示毕业生在科学研究中的最新进展,学校决定在建校五十周年之际,编辑出版《中国科学技术大学校友文库》,于2008年9月起陆续出书,校庆年内集中出版50种。该《文库》选题经过多轮严格的评审和论证,入选书稿学术水平高,已列为“十一五”国家重点图书出版规划。

入选作者中,有北京初创时期的毕业生,也有意气风发的少年班毕业生;有“两院”院士,也有IEEE Fellow;有海内外科研院所、大专院校的教授,也有金融、IT行业的英才;有默默奉献、矢志报国的科技将军,也有在国际前沿奋力拼搏的科研将才;有“文革”后留美学者中第一位担任美国大学系主任的青年教授,也有首批获得新中国博士学位的中年学者……在母校五十周年华诞之际,他们通过著书立说的独特方式,向母校献礼,其深情厚意,令人感佩!

近年来,学校组织了一系列关于中国科大办学成就、经验、理念和优良传统的总结与讨论。通过总结与讨论,我们更清醒地认识到,中国科大这所新中国亲手创办的新型理工科大学所肩负的历史使命和责任。我想,中国科大的创办与发展,首要的目标就是围绕国家战略需求,培养造就世界一流科学家和科技领军人才。五十年来,我们一直遵循这一目标定位,有效地探索了科教紧密结合、培养创新人才的成功之路,取得了令人瞩目的成就,也受到社会各界的广泛赞誉。

成绩属于过去,辉煌须待开创。在未来的发展中,我们依然要牢牢把握“育人是大学第一要务”的宗旨,在坚守优良传统的基础上,不断改革创新,提高教育教学质量,早日实现胡锦涛总书记对中国科大的期待:瞄准世界科技前沿,服务国家发展战略,创造性地做好教学和科研工作,努力办成世界一流的研究型大学,培养造就更多更好的创新人才,为夺取全面建设小康社会新胜利、开创中国特色社会主义事业新局面贡献更大力量。

是为序。

2008年9月

前 言

机器学习是人工智能、计算机科学等领域的重要研究方向之一,近年来深受人们的关注和喜爱.尤其是自 2006 年美国卡内基梅隆大学(CMU)专门设立机器学习系以来,更是在全球产生了史无前例的推动作用.人们意识到在现代科学研究方法中,如果还停留在计算的技术层面上,那整个科学的研究进程可能会受到限制.如对物质特性的分析、分子结构的物理性质的分析等,这些仅凭计算是很难实现的.这些数据结构的深层关系和语义的不确定性使得仅凭计算技术来完成这些任务就显得远远不够了.

针对数据的深层结构关系,我们研究小组从 2004 年开始引入李群(Lie Group),从机器学习的角度进行研究,提出了李群机器学习(Lie Group Machine Learning, LML)的基本理论框架.

针对数据语义的不确定性问题,我们研究小组从 1994 年开始研究,给出了面向不确定数据建模新的集合论、逻辑和模型等系统理论.为了进一步使该理论成为数据分析的有效方法,将动态模糊集合(Dynamic Fuzzy Sets, DFS)、动态模糊逻辑(Dynamic Fuzzy Logic, DFL)引入机器学习领域,提出了动态模糊机器学习(Dynamic Fuzzy Machine Learning, DFML)理论框架.

这些工作已在国际期刊、国际会议上发表,相关内容由作者在全国性的学术会议上做大会报告多次.为了能够快速地在更大范围内进行交流,特由作者整理成《机器学习理论及应用》一书.

该书分三篇共十二章.第 1 章引言,第 2 章李群机器学习模型,第 3 章李群机器学习(LML)子空间轨道生成算法,第 4 章李群机器学习的辛群学习算法,第 5 章李群机器学习的量子群学习算法,第 6 章李群机器学习的纤维丛学习算

法,第7章动态模糊机器学习模型,第8章动态模糊自主学习子空间学习算法,第9章动态模糊决策树学习,第10章基于DFL的多Agent学习模型,第11章Agent普适机器学习,第12章贝叶斯量子随机学习算法.

其中第2,3,7~9章由许欢、陈凤、张静、李明伦、谢琳提供初稿;第4~6章由付会欣、何书萍、张炯提供初稿;第10~12章由谢丽萍、王德鹏、茅伟强、钱旭培提供初稿;其余章由李凡长完成.文字输入由谢琳、钱旭培、何书萍完成,最后由李凡长统一定稿.

在本书的撰写过程中引用了大量的参考文献,在此对所有参考文献的作者表示感谢.

除此之外,南京大学的周志华教授,中科院自动化所的王珏研究员,中国科学院院士、中国科技大学的陈国良教授,中国科技大学的蔡庆生教授、陈恩红教授、陈小平教授、刘贵全博士,北京交通大学的黄厚宽教授、于剑教授,清华大学的石纯一教授、张长水教授,中科院计算所的史忠植研究员,复旦大学的张军平博士,北京大学的封举富教授,美国伯克利大学的 T. Y. Lin 教授,南京航空航天大学陈松灿教授等对本项工作的研究也曾直接或间接地提出过许多中肯的建议,在此一并表示感谢.最后,感谢国家自然科学基金委的支持(60775045),感谢中国科技大学出版社的领导.

本书除第1章外,其余各章自成体系,读者可选择自己感兴趣的章节阅读.若作为教材,最好以篇(第3篇除外)为单元进行讲授,这样可使学生接受系统的知识训练.

书中论及的机器学习新方法是机器学习的前沿领域,许多概念和理论尚待探讨,加之作者水平有限,书中若有不确切的地方在所难免,恳请读者指正,联系方式 Email:lfzh@suda.edu.cn.

李凡长

2009年3月于苏州

目 次

总序	i
前言	iii
第 1 章 引言	1

第 1 篇 李群机器学习

第 2 章 李群机器学习模型	13
2.1 引言	13
2.2 李群机器学习的概念	13
2.3 李群机器学习的代数模型	24
2.4 李群机器学习的几何模型	27
2.5 李群机器学习公理假设	30
2.6 李群机器学习 Dynkin 图的几何学习算法	42
2.7 李群机器学习的线形分类器设计	49
2.8 本章小结	53
参考文献	53
第 3 章 李群机器学习(LML)子空间轨道生成算法	55
3.1 LML 中偏序集及格的基本概念	55
3.2 LML 子空间轨道生成格学习算法	72
3.3 LML 中一般线性群 $GL_n(\mathbb{F}_n)$ 作用下学习子空间轨道生成格学习	

算法	88
3.4 本章小结	97
参考文献	97
第 4 章 李群机器学习的辛群学习算法	99
4.1 问题提出	99
4.2 李群机器学习中的辛群分类器设计	105
4.3 李群机器学习中的辛群分类器算法	123
4.4 应用实例	129
4.5 本章小结	136
参考文献	136
第 5 章 李群机器学习的量子群学习算法	140
5.1 问题提出	140
5.2 李群机器学习中的量子群分类器构造方法	141
5.3 量子群学习算法在分子对接中的应用	149
5.4 本章小结	162
参考文献	162
第 6 章 李群机器学习的纤维丛学习算法	165
6.1 问题提出	165
6.2 纤维丛学习模型	166
6.3 纤维丛学习算法	171
6.4 本章小结	180
参考文献	180

第 2 篇 动态模糊机器学习

第 7 章 动态模糊机器学习模型	185
7.1 问题提出	185
7.2 动态模糊机器学习模型	186
7.3 动态模糊机器学习系统的相关算法	201

7.4 动态模糊机器学习系统的过程控制模型	212
7.5 动态模糊关系学习算法	223
7.6 本章小结	230
参考文献	231
第 8 章 动态模糊自主学习子空间学习算法	235
8.1 自主学习研究现状分析	235
8.2 基于 DFL 的自主学习子空间的理论体系	238
8.3 基于 DFL 的自主学习子空间学习算法	242
8.4 本章小结	251
参考文献	251
第 9 章 动态模糊决策树学习	253
9.1 决策树学习的研究现状	253
9.2 动态模糊格的决策树方法	256
9.3 动态模糊决策树特殊属性处理技术	272
9.4 动态模糊决策树的剪枝策略	284
9.5 应用	291
9.6 本章小结	296
参考文献	297
第 10 章 基于 DFL 的多 Agent 学习模型	301
10.1 引言	301
10.2 基于 DFL 的 Agent 心智模型	303
10.3 基于 DFL 的单 Agent 学习算法	312
10.4 基于 DFL 的多 Agent 学习模型	317
10.5 本章小结	337
参考文献	337

第 3 篇 其他学习方法

第 11 章 Agent 普适机器学习	343
11.1 引言	343

11.2 Agent 普适机器学习	346
11.3 一种 Agent 普适机器学习分类器设计	358
11.4 本章小结	367
参考文献	367
第 12 章 贝叶斯量子随机学习算法	370
12.1 问题提出	370
12.2 相关基本理论	371
12.3 贝叶斯量子随机学习模型	387
12.4 网络结构的贝叶斯量子随机学习算法设计	395
12.5 网络参数的贝叶斯量子随机学习算法设计	403
12.6 面向缺失数据的贝叶斯量子随机学习算法设计	412
12.7 本章小结	420
参考文献	421
附录	
附录 1 拓扑群	427
附录 2 微分几何概念	430
附录 3 流形学习算法	431
附录 4 辛群的基本概念和性质	435
附录 5 量子群的基本概念	439
附录 6 纤维丛	446
附录 7 动态模糊集(DFS)	449
附录 8 动态模糊(DF)关系	455
附录 9 动态模糊逻辑	459
附录 10 动态模糊格及其性质	462
中英文名词对照	467

第 1 章 引 言

科学通过理论、实践形成了人们解释自然界、人类社会等复杂系统的工具. 在上世纪, 整个科学逐步接受了一种新的科学研究方法, 那就是机器学习. 机器学习方法可以对研究由于受经济或者其他条件限制不能进行实验的数据进行有效分析. 比如物质性质的分析、生命的演化等基本问题的回答都需要机器学习方法的支持. 2001 年美国航空航天局 JPL 实验室的科学家在《Science》上撰文指出^[1]: “机器学习对科学研究的整个过程正起到越来越大的作用, 并预计该领域将取得稳定而快速的发展”; 2003 年美国国防部高级研究计划局 DARPA 开始启动了以机器学习为核心的 PAL (Perception Assistant that Learns) 计划, 将机器学习技术的重要性上升到国家安全的高度来考虑; 2004 年, MIT Technology Review 列出的 “10 Emerging Technology That Will Change Your World” 将机器学习方面的研究列在第 4 位; 2006 年美国卡内基梅隆大学 (CMU) 专门成立了机器学习系; 2008 年, 江泽民同志在上海交通大学学报上撰文《新时期我国信息技术产业的发展》指出: “……人工智能等前沿技术研究, 推动信息学、物理学、生物学、认知学等学科的交叉融合, 力争在多进制计算、语义分析、人脑结构模拟、机器学习等关键技术方面有新作为, 实现原始性创新”^[2]. 这充分说明机器学习方法在科学研究中扮演着重要的角色, 许多重大科技活动越来越需要机器学习方法的参与.

因此, 开展机器学习新方法研究是每位机器学习研究者责无旁贷的事情. 为了进一步和同仁们进行交流, 我们将近年来在机器学习方面的一些

成果整理成此书. 全书分三篇介绍: 第 1 篇: 李群机器学习 (Lie Group Machine Learning, LML); 第 2 篇: 动态模糊机器学习 (Dynamic Fuzzy Machine Learning, DFML); 第 3 篇: 其他学习方法, 主要包括两种机器学习方法, 即: Agent 普适机器学习和贝叶斯量子随机学习算法. 下面就各篇的情况简单介绍.

第 1 篇 李群机器学习

近年来, 随着机器学习技术的普及应用, 面对物理、化学、生物信息学等领域的复杂数据, 在物理学家和化学家等广泛使用李群理论来研究相关领域的数据规律的启发下, 李群机器学习 (Lie Group Machine Learning, LML) 这个新的概念于 2004 年被提出, 目前已在加拿大、爱尔兰、芬兰、意大利、美国、中国等国展开了此方面的研究并取得了初步成绩.

李群结构是目前学术界公认的对学习问题研究很有用的一套理论工具. 李群之所以受人们关注, 一方面是李群有好的数学结构, 另一方面受到物理学家和化学家广泛使用李群方法来处理物理学中的晶体数据、有机物和无机物的数据、药物分子结构数据等这些复杂数据的启发. 从数据分析的角度看, 用机器学习方法进行数据分析, 其目的就是揭示这些数据具有的规律, 从而帮助用户提供解释的依据. 引进李群理论研究机器学习问题自然是一种可以探索的新思路.

(1) 从国外的研究方面看

1991 年 Tien-Ben Tsao^[3]等人提出了用李群的方法解决神经系统对视觉运动的三维解释. 它采用李群方法将刚性约束用于噪声抑制和三维运动的计算, 把三维图像解释成一个线性信号变换, 而这个线性变换的基本信号是三维欧几里德的六个极小生成子的二维向量场并由神经网络执行, 其计算机模拟结果表明, 极小随机噪声会导致传统的代数方法失败, 而基于李群的神经系统依然能保证其可靠性. 1999 年 Rajesh P. H. Rao 和 Daniel L. Ruderman^[4]提出了基于李群理论的一个贝叶斯方法来学习不变的视觉感知, 运用基于矩阵指数的图像生成模型, 从包含极小变换的输入数据到一个学习李群算子的非监督学习算法, 并证明了以前基于一阶泰勒级数的方法只是他们新提出的李群方法的一个特例. 2000 年 Gerge A. Kantor 和

P. S. Krishnaprasad^[5]引入一个能够在一个分布式控制网络上执行的线性算子类,即回归正交变换(ROTS),在那些可以表示 $SO(n)$ ($n \geq 4$) 的 ROTs 中获取条件,提出了一个基于平面矩阵对角化的分布式反馈控制的范式,并在合适的内在的李群中提取一个梯度流以找到一个适合这项工作的 ROT. D. Gregory Arnold^[6]使用李群分析作为热物理学模型的一种方法,提出了支配子空间不变量的概念和拟不变量的一种特殊类型,并给出了支配子空间不变量(DSI)算法. David McG. Squire^[7]等人利用李变换群提出了一种新的二维轮廓的不变特征方法. 2002年 Simone Fiori^[8]针对神经学习常常是发生在参数空间中,并且这些参数空间又常常被赋予一个特殊的几何结构,他们给权空间赋予一个特殊的李群结构,用李群的微分几何的结构来学习这些学习属性,通过此方法揭示了不同学习方法间内在的属性的联系,从而提出了基于李群的非监督神经学习. 2004年 Abdol-Reza Mausolur^[9]等人提出了图像分析中的约束主轮廓进化的李群方法,通过使用平面作用的李群及其极小生成子的李代数的对应关系适当地调整欧拉—拉格朗日下降方程,在此保证当保持原函数的下降方程不变时,曲线进化发生在被选变换群的轨道上,此方法的优点在于无须任何原函数的信息,也无须对其进行任何修改就能直接实现. Venu Madhav Govindu^[10]使用运动解释的内在的李群结构来求平均,用特殊正交群 $SO(3)$ 和特殊的欧几里德群 $SE(3)$ 的李代数来定义李群上的平均值,提出了全局相容运动估计的李代数平均值法. 此方法能够线性计算所有可能相对运动及对运动优化快速求平均. 2005年, Yasunori Nishimari, Shotaro Akaho^[11]提出了在 Stiefel 流形上基于测地线和接近于测地线的曲线族(拟测地线)的新的学习算法,强调了李群的非正规化约束在这个流形上的作用. 2006年, Guy Lebanon^[12]提出了一种新的度量学习问题的框架结构,它能够(健壮)学习高维离散数据的局部度量,它是将候选量集限制在一个参数族中,此候选度量是在一个变换李群下的 Fisher 信息量的拉回度量. 2007年,哥伦比亚大学的 Risi Kondor 在 ICML 2007 上组织了“Group Theoretical Methods in Machine Learning”专题讨论. 2008年完成了《Group theoretical methods in machine learning》博士论文^[13],给出了群作用下核不变量的对称化方法及置换群学习的 PerMceptran 算法和在傅里叶谱下群作用核的转换不变量的特性等内容. 由此可见,李群理论正在计算机科学、人工智能、机器学习等领域发挥着重要

作用.

(2) 从国内的研究方面看

2004 年李凡长等人发表了《基于李群的机器学习理论框架》第一篇论文^[14], 经过几年努力, 他领导的李群机器学习 (Lie Group Machine Learning, LML) 研究小组, 取得了一些初步的成果^[15-36], 其内容包括: (1) 提出了李群机器学习的公理系统: 即一致性假设公理、独立性假设公理、对偶性假设公理和泛化假设公理. (2) 提出了李群机器学习的基本模型, 内容包括: ① 李群机器学习的代数模型方面, 根据李群机器学习的定义, 将李代数和单参数子群, 左不变向量场和左不变流形这四种不同的数学结构定义成自然的一一对应, 即 $\{\text{李代数}\} \cong \{\text{左不变向量场}\} \cong \{\text{左不变流形}\} \cong \{\text{单参数子群}\}$, 基于这些数学结构的关系, 给出了李群机器学习的李代数根系计算模型、分解模型及分类模型等, 并将这些基本结构应用于晶体分类, 完成了基于 $SO(3)$ 分类器的晶体晶系分类系统, 形成帮助物理学家对晶体进行分类的计算机辅助系统. ② 在李群机器学习的几何模型方面: 主要利用李群的一些几何性质, 如仿射性、对偶性、平移性、测地线性性质、曲率等几何性质来研究学习问题. 将此讨论的观测集假定成紧致连通的, 这样就可以在观测集上的单位点上取一个自同构作用下不变的内积, 对于这样取定的内积, 再取一组标准正交基, 然后用左平移把它们分别扩张成观测集的左不变向量场, 就可以唯一地在观测集上的每一个点的切空间取定内积, 使得该内积恰好为左不变向量点的切空间的标准正交基, 这样就构成了一个黎曼空间. 所以, 给出了李群机器学习的对偶空间计算模型、素根系的几何结构分类模型及共轭类计算模型等. (3) 提出了李群机器学习子空间轨道生成格理论及相关算法. 根据独立划分假设公理及一致性假设公理和观测空间集合中的每一个元素都可以用一组独立实参数来描写的基本事实, 结合李群机器学习的几何模型特点, 在李群机器学习空间的基础上, 建立李群机器学习子空间轨道生成格的概念, 给出了在观测集上的各种典型群作用下, 由多个相同维数或轨道和秩子空间生成的格理论及相关学习算法, 如一般线性群的轨道生成格学习算法、轨道生成格宽度优先学习算法、深度优先学习算法和启发式学习算法等. (4) 提出了李群机器学习的量子群分类器设计. (5) 提出了李群机器学习的辛群分类器设计. (6) 提出了上同调边缘学习算法、数据分解算法、扩展算法、对偶学习算法及几何学习算法

等.(7) 完成了基于量子群学习算法的三维分子检索系统、晶体辅助分类系统等.

该篇主要包括第2~6等五章.

第2章主要介绍李群机器学习的基本概念及相关学习模型.

第3章主要介绍李群机器学习子空间学习理论及相关学习算法.

第4章主要介绍李群机器学习的辛群学习算法及相关内容.

第5章主要介绍李群机器学习的量子群学习算法及相关内容.

第6章主要介绍李群机器学习的纤维丛学习算法及相关理论.

第2篇 动态模糊机器学习

本篇包括第7~10等四章的内容.

第7章主要介绍动态模糊机器学习模型,面对不确定数据的动态模糊性及Simon对机器学习的描述,提出了动态模糊机器学习的学习模型、参数学习算法、极大似然估计学习算法、动态模糊关系学习算法等相关内容.

第8章主要介绍动态模糊自主学习子空间学习算法,包括:自主学习的基本概念及相关学习算法.

第9章主要介绍动态模糊决策树学习,针对不确定数据将动态模糊集及动态模糊逻辑的基本理论应用于机器学习领域,提出了动态模糊决策树学习的基本概念及相关系统理论.

第10章主要介绍基于DFL的多Agent学习模型,将动态模糊逻辑(DFL)和Agent技术相结合,提出了多Agent学习的模型及相关学习算法.

第3篇 其他学习方法

本篇主要包括两种机器学习方法,即Agent普适机器学习方法和贝叶斯量子随机学习算法.

第11章主要介绍Agent普适机器学习,将Agent技术、普适计算技术及机器学习相结合,形成了Agent普适机器学习方法,该章系统地介绍了Agent普适机器学习的相关内容及相关算法.

第12章主要介绍贝叶斯量子随机学习算法,给出了贝叶斯量子随机学习

模型、网络结构的贝叶斯量子随机算法设计、网络参数的贝叶斯量子随机学习算法设计等相关内容。

现用图将相关内容联系起来,如图 1.1.

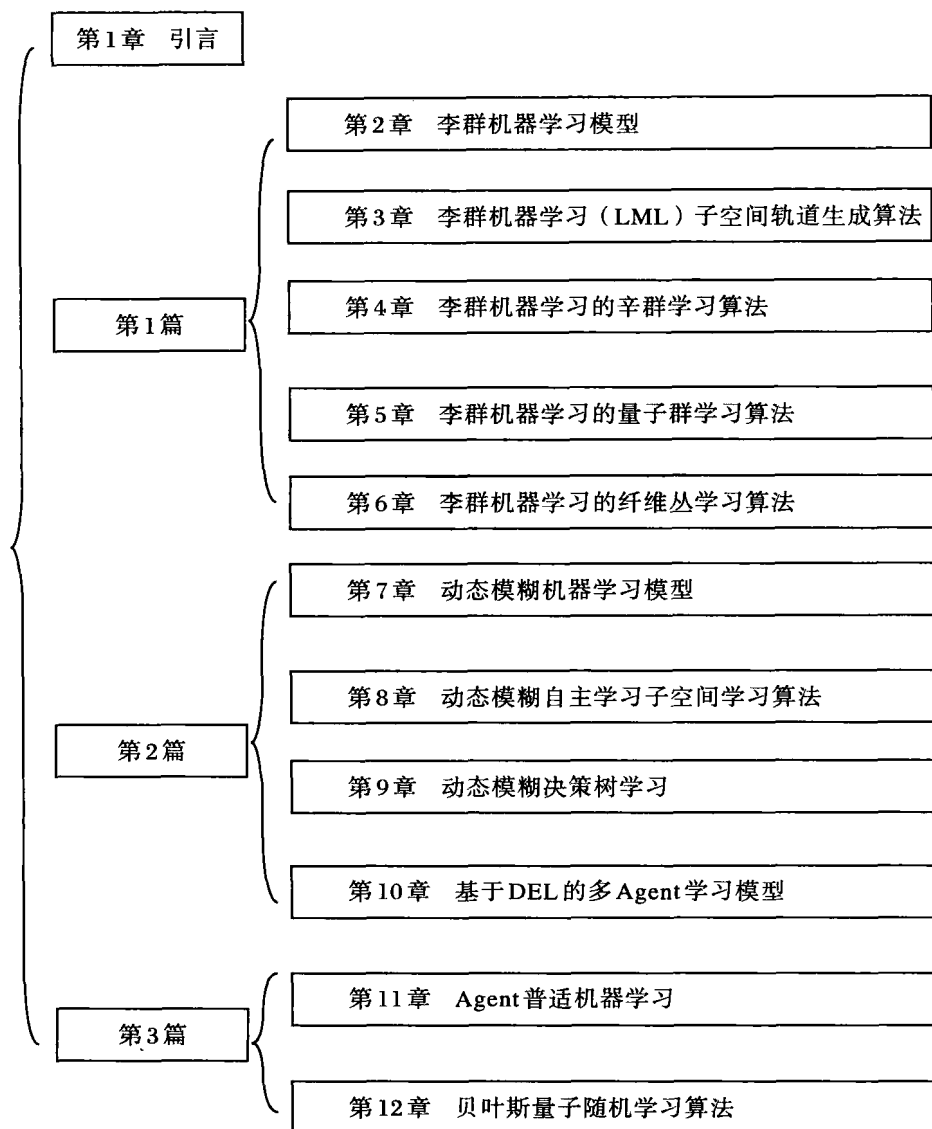


图 1.1 各章联系图