

| 刘颖 著 |

基于机器学习的 遥感影像分类方法研究

清华大学出版社



刘颖 著

基于机器学习的 遥感影像分类方法研究



清华大学出版社
北 京

内 容 简 介

机器学习是人工智能的一个重要领域,源自于统计模型拟合。机器学习通过采用推理及样本学习等方式从数据中获得相应的理论,尤其适合解决“噪声”模式及大规模数据集等问题。本书是作者几年来科研成果的总结。全书共7章,围绕遥感图像分类这一主线,深入研究监督学习、半监督学习、集成学习三大主流机器学习算法,构建完整的遥感图像分类体系。在理论研究的基础上,结合实例,详细介绍了改进机器学习算法及其在遥感分类处理中的应用情况。

本书内容充实、结构清晰、实例丰富,适合从事计算机及相关学科的师生,以及相关科研院所的科研人员阅读。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

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

基于机器学习的遥感影像分类方法研究 / 刘颖 著. —北京:清华大学出版社, 2014
ISBN 978-7-302-35991-3

I. ①基… II. ①刘… III. ①机器学习—应用—遥感图像—分类—研究 IV. ①TP75②TP181

中国版本图书馆 CIP 数据核字(2014)第 065957 号

责任编辑:王粲娉 易银荣

封面设计:牛艳敏

版式设计:方加青

责任校对:曹 阳

责任印制:刘海龙

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 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 装 者:三河市金元印装有限公司

经 销:全国新华书店

开 本:169mm×230mm 印 张:10.25 字 数:114 千字

版 次:2014 年 5 月第 1 版 印 次:2014 年 5 月第 1 次印刷

定 价:58.00 元

产品编号:058438-01

前言

长期以来，土地覆盖变化的研究一直是全球环境研究的热点，无论从社会经济角度还是从生态环境角度均具有重要的意义。为了全面掌握土地覆盖变化信息，迫切需要使用切实有效的方法实现土地覆盖宏观、动态、大尺度的制图与监测，遥感技术的迅猛发展为这一需求提供了可能。然而，目前遥感信息处理和分类的水平大大滞后于遥感影像获取技术的发展。因此，研究新理论、新方法以提高遥感信息的处理能力具有十分重要的意义和应用前景。

支持向量机(Support Vector Machines, SVM)是近年来机器学习与模式识别领域新的研究焦点，它具有结构简单、适应性强、全局最优等特点，能较好地解决高维特征、非线性、过学习与不确定性等问题，广泛地应用于土地覆盖遥感分类。尽管SVM在遥感信息获取中取得了很好的效果，但仍存在有待改进和完善之处，主要表现在以下两方面：①参数选择的问题，即不准确的分类参数常常影响分类器的分类精度；②样本不足且代表性不好的问题，即当训练样本集远远小于测试样本集时，即便SVM具有较强的泛化性，也难以给出令人满意的结果。围绕这些问题，本书开展了如下工作：

1. 选择图们江下游, 中、朝、俄交界处作为研究对象。以行列号115-30一景、近20年的6幅不同时相的Landsat ETM/TM影像作为研究材料。分别讨论本书所采用的影像合成方式、特征采集方法、土地覆盖分类依据, 以及特征选取方法, 为进一步研究分类方法提供必要的材料。

2. 针对SVM分类过程中核函数选择及参数设置不准确的缺点, 提出一种基于自适应变异粒子群优化SVM参数模型(Adaptive Mutation Particle Swarm Optimization SVM, AMPSO-SVM)。AMPSO在运行过程中根据群体适应度方差以及最优解的大小来确定当前最佳粒子的变异概率。与传统粒子群(Particle Swarm Optimization, PSO)优化SVM参数模型(PSO-SVM)相比, AMPSO-SVM能够快速摆脱局部搜索的束缚, 提高全局搜索的性能, 克服早熟收敛造成分类参数寻找不准确的缺点, 同时保持了种群的多样性。最后应用该模型进行多光谱遥感影像的土地覆盖分类实验, 并与SVM分类方法、PSO-SVM分类方法进行对比。分类精度从传统PSO-SVM的91.50%提高到93.59%, Kappa系数由0.890 3提高为0.917 5。 c 和 γ 的取值得到的分类结果明显优于SVM的手工设置值100和0.143所得到的结果(分类精度87.07%, Kappa系数0.837 2)。结果表明, AMPSO-SVM模型有效地提高了遥感影像的分类精度。

3. 提出了一个新的自训练半监督支持向量机方法(PS3VM)。自训练半监督算法最大的弊端在于“错误累积”现象, 即在学习过程

中,一旦某个分类出错,将导致这个错误被继续学习与加强。为了克服这一现象,本书在自训练半监督SVM(S3VM)的基础上引入两个算法:①从分类器的构造角度,利用自适应变异粒子群算法对SVM参数优化,以提高单个分类器的分类精度;②在未标记样本的标注阶段,采用Gustafson-Kessel模糊聚类算法(GKclust)将最接近样本的有效无标签样本作为标注对象,以控制错误信息的输入。为了测试所提模型的有效性,分别针对遥感的数字化集合和影像集合进行分类实验,并与AMPSO-SVM(简称PSVM)监督分类方法、未改进自训练S3VM方法进行对比实验,由PS3VM产生的分类精度(95.10%)分别比S3VM(93.06%)高出2.04%;比PSVM(90.81%)高出4.29%。实验结果一方面说明了已标记样本和未标记样本的用量比例必须满足一定的阈值要求(1:3),才能产生最小的泛化误差;另一方面证实了利用所提出学习框架能够获得较好的分类精度。

4. 对于样本不足且代表不好而造成的小样本问题,学者们普遍采用半监督学习和集成学习两种范式对SVM进行改进。然而,集成学习与半监督学习之间存在许多互补性,且二者的混合范式(即半监督集成)可以更大程度地改进学习系统的泛化能力。因此,本文设计了一种新的半监督集成方案(EPS3VM),PS3VM半监督方法利用未标记数据有效地应对训练样本不足的同时也产生若干性能差异的个体分类器,将这些个体分类器采用加权集成策略进一步提高分类模型的泛化能力。为了测试其性能,应用该模型进行多光谱

遥感影像的土地覆盖分类实验，并与其相关算法进行对比。分类精度从92.16%(PS3VM)提高到96.88%，Kappa系数由0.901 0提高为0.960 6。结果表明，EPS3VM克服传统SVM参数选择不准确的同时有效地应对了小样本问题，分类性能更优。

本书是在吉林财经大学资助下，国家自然科学基金项目(61202306)、吉林省科技厅项目(20130522177JH, 201215119, 20100507)、吉林省教育厅十二五重点规划项目(2012185)、吉林省高校新世纪优秀人才支持计划、吉林财经大学青年学俊等项目支持下完成的。值此专著完成之际，诚挚地感谢吉林财经大学的资金支持，感谢中国科学院东北地理与农业生态研究所张柏教授、吉林财经大学管理科学与信息工程学院王丽敏教授及长春工业大学韩旭明副教授的热情帮助和指点。

由于作者水平有限，加之机器学习领域研究领域纵深宽广，书中难免有考虑不周之处，诚请广大读者批评指正。

刘 颖

2014年1月于长春

目 录

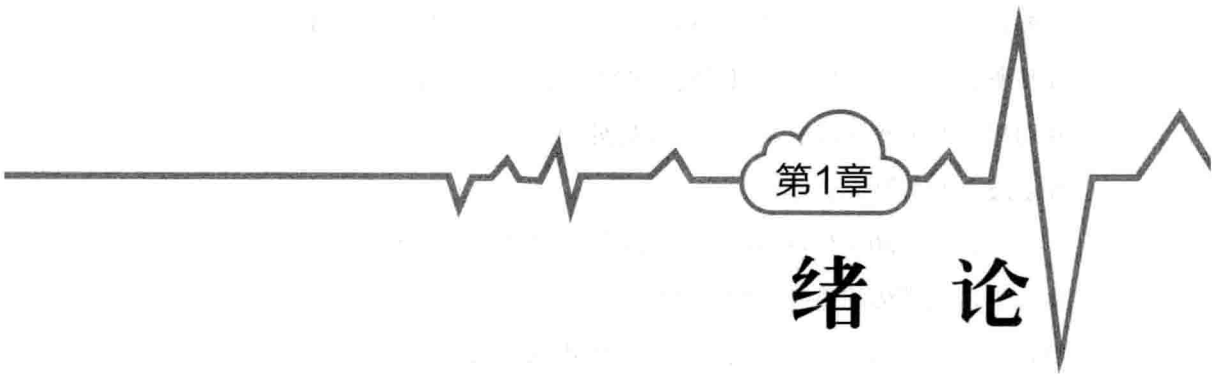
第1章 绪论	1
1.1 基本概念	2
1.1.1 土地覆盖	2
1.1.2 遥感技术	3
1.1.3 机器学习	4
1.2 研究意义	5
1.2.1 丰富土地覆盖遥感分类的理论与方法	6
1.2.2 为土地利用/覆盖的动态监测、保护和管理提供 技术支持	6
1.2.3 一种新的自适应半监督支持向量机遥感分类模型的 提出	7
1.2.4 半监督学习思想和集成学习思想的融合	7
1.3 本书研究方法 & 结构安排	7
1.3.1 研究方法	7
1.3.2 结构安排	10
参考文献	12

第2章	关键技术国内外研究现状	19
2.1	遥感影像信息提取方法	20
2.2	SVM遥感分类研究进展	24
2.2.1	SVM在遥感分类中的优点	24
2.2.2	SVM在遥感影像分类中的不足	26
2.2.3	SVM在遥感影像分类中的应用领域	27
2.3	半监督学习理论及研究进展	29
2.4	半监督分类中的聚类算法	32
2.5	集成学习理论及研究进展	32
	参考文献	36
第3章	遥感图像数字化	49
3.1	研究区位置及遥感影像集	50
3.1.1	研究区位置	50
3.1.2	研究区影像集	52
3.1.3	分类体系的建立	52
3.2	遥感影像数字集	53
3.2.1	样本采集	53
3.2.2	特征选取	56
3.3	本章小结	62
	参考文献	63

第4章 SVM参数优化方法研究	67
4.1 SVM理论及参数优化算法研究进展	68
4.1.1 SVM的核心思想	68
4.1.2 SVM理论	68
4.1.3 SVM参数优化方法研究进展	72
4.2 基于自适应变异粒子群参数优化的土地覆盖分类 模型	75
4.2.1 传统粒子群算法(PSO)	75
4.2.2 自适应变异粒子群优化算法(AMPSO)	76
4.2.3 土地覆盖分类模型构建	79
4.3 实验结果与分析	82
4.3.1 实验影像选取	82
4.3.2 特征选取及样本集表示	83
4.3.3 核函数的选取	83
4.3.4 实验参数及精度评价指标	84
4.3.5 实验结果与比较	85
4.4 本章小结	90
参考文献	91
第5章 基于模糊聚类的半监督支持向量机土地覆盖分类方法 研究	95
5.1 概述	96

5.2	自训练半监督学习	96
5.2.1	无标签样本的重要性	96
5.2.2	自训练半监督算法	97
5.3	模糊聚类理论	99
5.3.1	聚类的概念	99
5.3.2	常用聚类算法	100
5.3.3	聚类有效性验证	105
5.4	一种新的自训练半监督支持向量机分类模型构建	106
5.4.1	未标记样本的选择依据	107
5.4.2	基于GKclust的自训练半监督支持向量机设计 流程	107
5.4.3	基于GKclust的自训练半监督支持向量机算法	109
5.5	实验结果与分析	109
5.5.1	遥感影像数字化	110
5.5.2	参数设置	111
5.5.3	模糊聚类算法的比较	112
5.5.4	无标签样本的参与比例	115
5.5.5	土地覆盖遥感图像分类	121
5.6	本章小结	123
	参考文献	124

第6章 基于半监督集成支持向量机的土地覆盖分类研究	129
6.1 概述	130
6.2 集成学习框架	130
6.2.1 个体生成方法	131
6.2.2 结论生成方法	133
6.3 半监督集成支持向量机的土地覆盖分类模型构建	134
6.3.1 个体生成算法	135
6.3.2 结论生成算法	136
6.4 实验结果与分析	136
6.4.1 实验数据	137
6.4.2 结果与精度分析	137
6.5 本章小结	140
参考文献	141
第7章 总结与展望	145
7.1 研究结论	146
7.2 本书不足之处	148
7.3 研究展望	148



第1章

绪 论

1.1 基本概念

1.1.1 土地覆盖

土地覆盖变化是全球环境变化研究的热点和前沿，是国际地圈—生物圈计划(IGBP)的核心研究内容之一，无论从社会经济角度还是从生态环境角度均具有重要的意义。土地覆盖真实地反映了地表覆盖情况，它并不是单一的植被及土地类型，而是土地利用类型及其所具有的一系列自然属性特征的综合体，包括与土地覆盖类型密切相关的生态环境要素，如植被所处的生态区域、地形与气候条件以及土地利用状况等^[1]。

自20世纪80年代以来，随着臭氧层的破坏、全球变暖、土地退化、土地荒漠化的蔓延、酸雨频繁发生等全球性环境问题的不断涌现，人们逐渐认识到土地利用与土地覆盖变化是气候、生态过程、生物多样性、生物化学循环，乃至全球变化的主要原因^{[2]~[3]}。具体表现在以下几个方面。

(1) 土地覆盖变化对人类资源利用的影响。由于土地覆盖变化使得水资源、耕地资源、草场资源、森林资源的存量发生变化，从而对人类社会产生影响，并对人类环境的安全造成威胁^{[4]~[5]}。

(2) 土地覆盖变化对土壤的影响。土地覆盖的变化使得土壤养分失衡、营养元素衰竭、有机质含量变化、水分循环改变^{[6]~[8]}。

(3) 土地覆盖变化对全球气候变化的影响。地表反照率、粗糙度、植被叶面积指数及植被覆盖度等变化造成降水量减少和温度增加^{[9]~[10]}，同样也引起局地区域气候变化，如城市热岛现象，城市酸

雨等^{[11]~[12]}。

(4) 土地覆盖变化对生态系统的影响,集中于生态系统功能方面的物质循环和能量流动。如在温度、CO₂浓度及降水量等自然环境要素发生变化的情况下,生态系统的初级生产量会受到影响^{[13]~[14]}。

(5) 土地覆盖变化对水文和地表、地下径流的影响。表现在其改变所造成的气候变化,汛期流量增加,枯水期流量减少,水中泥沙含量变化,径流极值变化,进而影响地下水的补给,产生多种水文效应^{[15]~[17]}。

(6) 土地覆盖变化还可能加剧自然灾害。如森林的砍伐、植被覆盖度的降低使得土壤侵蚀面积增大,洪涝灾害加剧^{[18]~[19]}。

因此,有效地管理和保护日益稀缺的土地资源,探索土地覆盖动态变化过程,解析变化的驱动因素,抑制全球环境恶劣变化等诸多问题一直是地学领域学者关注的热点,而这些问题的解决依赖于准确、实时地获取土地覆盖动态变化信息。

1.1.2 遥感技术

遥感RS(Remote Sensing)是20世纪60年代兴起并迅速发展起来的一门包括航天技术、计算机技术、图像处理等综合性的探测技术^[20]。它是一种远离目标,在不与目标直接接触的情况下,通过某种平台上装载的传感器获取其特征信息,然后对其所获得的信息进行提取、判定、加工处理及应用分析^[21]。遥感技术能大面积、重复地获取区域多波段、多时相信息,为大面积、实时、动态监测土地覆盖动态变化提供了可能,是调查、监测和分析最好的手段之一^[22]。从此,人类对地球表层的理解进入一个崭新的阶段。然而,由于地球

系统的复杂性和开放性, 遥感信息在进行地学空间分析和反演过程中具有模糊性和多解性的特点, 即复杂性和不确定性^{[23]~[25]}。表现为不同土地覆盖类型经常有相似的光谱特性(异物同谱现象), 以及相同土地覆盖类型由于地形、光照条件等的影响会具有不同的光谱特性(同物异谱现象), 这些增大了土地覆盖信息的提取难度。因此, 在研究不同土地覆盖类型的光谱特征和空间分布特征的同时, 进一步研究土地覆盖分类策略与方法, 将有助于土地覆盖信息的有效获取。

1.1.3 机器学习

机器学习是人工智能的一个重要领域, 源自于统计模型拟合。机器学习通过采用推理及样本学习等方式从数据中获得相应的理论, 尤其适合解决“噪声”模式及大规模数据集等问题。在大样本、多向量及不确定数据分析工作中发挥着日益重要的作用^{[26]~[27]}。

根据学习方式的不同, 机器学习方法可分为以下几类: ①监督学习(Supervised Learning), 利用一组已知类别的样本调整分类器的参数, 使其达到所要求性能的过程; ②无监督学习(Unsupervised Learning), 是指所有样本的类别都是未知的, 算法通过数据特征自动产生类别属性; ③半监督学习(Semi-supervised Learning), 介于监督学习和无监督学习之间, 所需样本既包括已知类别样本又包括未知类别样本, 通过挖掘未知类别样本中所蕴涵的固有结构信息, 来对已知类别样本可能因代表性不好而造成的拟合分类器有偏差的情况进行校正; ④集成学习(Ensemble Learning), 综合多个同构或异构学习机对同一个问题进行学习进而提高分类器的泛化能力。

当前,许多机器学习技术被广泛地应用于遥感影像的分类,例如最大似然法、K-means算法、神经网络、决策树、支持向量机等^{[28]~[29]}。其中,Vapnik提出的支持向量机(Support Vector Machines, SVM)监督分类技术是近年来模式识别与机器学习领域一个新的研究热点^{[30]~[31]},非常适合处理高维、复杂的小样本多维数据分类^{[32]~[33]},广泛应用于土地覆盖分类^{[34]~[36]}、森林类型识别^{[37]~[38]}、农业作物监测^[39]、道路信息提取^[40]、图像分割^[41]等领域。尽管SVM在遥感信息提取中取得了很好的效果,但是仍然存在需要改进和完善之处。主要表现在以下两方面。

(1) 参数选择的问题:分类参数的选择没有特别好的办法,应用时不容易找到最优分类参数。

(2) 样本不足且代表不好的问题:当训练样本集远远小于测试样本集,即便SVM具有较强的泛化性,也难以给出令人满意的结果。

1.2 研究意义

本研究充分考虑遥感影像特点及SVM土地覆盖分类技术不足的问题,一方面利用智能优化算法,克服传统SVM参数选择不准确弊端;另一方面,将模糊聚类技术、半监督学习和集成学习理论引入SVM土地覆盖分类方法中,有效地解决小样本问题。其目的在于丰富土地覆盖遥感信息提取理论与方法,提高土地覆盖分类精度,为土地资源的合理保护、利用与监测提供更为有效的数据信息。