



中国计算机学会学术著作丛书
——知识科学系列 12

机器学习及其应用 2017

高阳 陈松灿 主编

清华大学出版社

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 12

机器学习及其应用 2017

高阳 陈松灿 主编



清华大学出版社
北京

内 容 简 介

本书是对第十三届和第十四届中国机器学习及其应用研讨会的一个总结,共邀请了与会的7位专家就其研究领域撰文,以综述的形式探讨了机器学习不同分支及相关领域的研究进展。这些内容涉及稳健的矩阵回归模型、低秩模型的闭解、面向大规模学习的随机优化、张量分解、基于递归神经网络的图像描述算法、标记分布学习及其应用,以及软件缺陷挖掘。

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阅读参考。

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 2017/高阳,陈松灿主编. —北京:清华大学出版社,2017
(中国计算机学会学术著作丛书. 知识科学系列)
ISBN 978-7-302-48781-4

I. ①机… II. ①高… ②陈… III. ①机器学习 IV. ①TP181

中国版本图书馆 CIP 数据核字(2017)第 263453 号

责任编辑:薛 慧

封面设计:傅瑞学

责任校对:王淑云

责任印制:沈 露

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:小森印刷(北京)有限公司

经 销:全国新华书店

开 本:185mm×230mm

印 张:9.75

字 数:198 千字

版 次:2017 年 11 月第 1 版

印 次:2017 年 11 月第 1 次印刷

印 数:1~2000

定 价:58.00 元

产品编号:074540-01

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：陆汝铃

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。

陆汝铃

2006 年 1 月

前言

为促进国内人工智能领域学者的交流,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。是年11月,第一届研讨会邀请了10余位专家闭门研讨,在复旦大学成功举行。2004年11月周志华教授与王珏研究员在复旦大学主持举办了第二届“机器学习及其应用”研讨会。此次研讨会除邀请报告人外,还吸引了闻讯自发与会旁听的学者和研究生100余人。自此,研讨会的模式从闭门的学术讨论,走向了开放的学术交流。

由于当时机器学习受到的关注和支持较少,组织者和主办单位需承担较多压力,2005年研讨会移师南京大学软件新技术国家重点实验室,周志华教授和王珏研究员主持举办了第三届研讨会,吸引了来自全国近10个省市的250余人旁听。此次研讨会确定了会议不征文、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,同时确定了研讨会举办的时间,如无特殊,则为当年11月份的第一个周末。此后,组织者争取到兄弟单位给予协助,2006年、2007年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10余个省市的约300人旁听;2008年举行的第六届研讨会,适逢南京大学计算机学科建立50周年,与会人数达到了380余人;2009年和2010年在南京大学分别举行了第七、八届研讨会,均有约400余人旁听。这一时期为国内机器学习界的“垦荒”阶段,众多学者为研讨会做出了重要贡献。研讨会在国内机器学习领域乃至整个计算机领域逐渐产生了重要的影响,学界美名为“南京会议”。

随着国内科研条件的改善,机器学习逐渐获得更多关注和支持。为进一步推动机器学习在国内的发展,从2011年起,研讨会的举办地开始走出南京。2011年和2012年由张长水教授和杨强教授主持,清华大学自动化系、智能科学与系统国家重点实验室、清华大学信息科学与技术国家实验室(筹)举办第九届和第十届研讨会,两次会议均有500多人参加。2013年、2014年由张军平教授和高新波教授主持,分别在复旦大学计算机科学技术学院和上海市智能信息处理实验室举行了第十一届研讨会,在西安电子科技大学举办了第十二届研讨会,这两次会议分别达到600多人和800多人参加。2015年和2016年,研讨会再次回到南京大学举办,此时国内机器学习研究和应用已经发展到一个新的高峰,这两次会议均约有1400人参加。至此,研讨会已经成为备受国内机器学习及其相关

领域研究人员关注的盛会。

研讨会自发起之日起,清华大学出版社敏锐地了解到这一学术动态。研讨会主要组织者与清华大学出版社达成共识,每两年对研讨会上专家交流的部分技术内容,经过整理,结集成书,以飨读者。十余年来,陆续出版了《机器学习及其应用》《机器学习及其应用 2007》《机器学习及其应用 2009》《机器学习及其应用 2011》《机器学习及其应用 2013》《机器学习及其应用 2015》等系列书籍。本书是以上序列书籍的延续。

本书是对第十三届和第十四届中国机器学习及其应用研讨会交流内容的部分总结,共邀请了会议中的 7 位专家就其研究领域撰文,以综述的形式探讨了机器学习不同分支及相关领域的研究进展。全书共分 7 章,内容分别涉及矩阵回归模型、低秩子空间模型、随机优化等经典模型和优化方法,介绍了新型的标记分布机器学习范型,以及机器学习在图像描述、软件缺陷挖掘等方面的应用。

杨健教授介绍了几种全新的稳健回归模型,包括基于核范数的稳健矩阵回归,基于推广幂指数分布的稳健矩阵回归,基于核- L_1 范数的联合矩阵回归,以及基于树结构核范数的稳健矩阵回归。

林宙辰教授讨论了若干具有闭解的低秩模型,以及其在图像处理领域的诸多成功应用。

在大规模数据的机器学习算法中,随机近似是一种新兴的技术手段。张利军教授介绍了阶段混合梯度下降、随机临近梯度下降,能够有效降低学习算法的空间和时间复杂度。

张量结构往往用于刻画数据的多源关系。徐增林教授介绍了非参非线性张量分解,能够提高分解模型的精度。

图像描述是近年来有挑战性的任务。张长水教授介绍了基于注意力的递归神经网络编码器,刻画了图像和句子之间的编码关系。

和图像描述类似,大量的机器学习样本涉及标记多义性的问题。耿新教授详细介绍了一种新的机器学习范型——标记分布学习,为解决标记多义性开辟了一条新途径。

最后,黎铭副教授探讨了机器学习技术在软件缺陷挖掘中的创新应用,以及其中面临的若干挑战性问题。

本书概括了国内机器学习及其应用的最新研究进展,可供计算机、自动化、信息处理及相关专业的研究人员、教师、研究生和工程技术人员参考,也可作为人工智能、机器学习课程的辅助内容,希望对有志于从事机器学习研究的人员有所帮助。

高 阳 陈松灿

2017 年 9 月 4 日

目 录

稳健的矩阵回归模型与方法	杨 健 罗 雷	1
1 引言		1
2 基于核范数的稳健矩阵回归		3
3 基于推广幂指数分布的稳健矩阵回归		6
4 基于核- L_1 范数的联合矩阵回归		8
5 基于树结构核范数的稳健矩阵回归		11
6 结束语		14
参考文献		14
若干低秩子空间恢复模型的闭解及其应用	林宙辰	19
1 引言		19
2 无噪低秩表示模型的闭解		21
2.1 应用：潜在低秩表示模型的推导		23
3 无噪潜在低秩模型的闭解		24
3.1 应用：提升潜在低秩表示模型的聚类性能		25
4 若干主要低秩模型的解之间的关系		26
4.1 解的质量的比较		28
4.2 求解速度的比较		30
5 闭解应用的另一个例子：仿射变换下特征点误匹配检测		30
6 结束语		33
参考文献		33
面向大规模机器学习的随机优化	张利军	35
1 引言		35
2 相关工作		36
3 降低时间复杂度		38
3.1 研究背景		38

3.2 阶段混合梯度下降	39
3.3 实验	42
4 降低空间复杂度	43
4.1 研究背景	43
4.2 随机临近梯度下降	44
4.3 实验	46
5 总结与展望	48
附录	48
参考文献	49
非参贝叶斯张量分解研究	徐增林 贺丽荣 严丰 漆远 53
1 引言	53
2 Tucker 分解	54
3 CP 分解	55
4 非参贝叶斯分解	57
4.1 InfTucker 分解	57
4.2 InfTucker 算法	59
4.3 实验结果	63
5 结束语	65
参考文献	66
基于递归神经网络的图像描述算法	靳骏奇 闫子昂 张长水 69
1 引言	69
2 相关工作	70
3 模型与方法	71
3.1 基于多尺度局部色块的图像表示	71
3.2 基于注意力的递归神经网络解码器	73
4 实验与分析	76
4.1 实验设定	76
4.2 定量评估结果	77
4.3 微软比赛结果	79
4.4 定性评估结果	79
5 结束语	84
参考文献	85

标记分布学习及其应用	耿 新 徐 宁	87
1 引言		87
2 学习框架		88
2.1 符号及形式化定义		88
2.2 评价指标		90
3 标记分布学习算法		91
3.1 “问题转化”算法		91
3.2 “算法改造”算法		92
3.3 专用算法		93
4 标记分布学习应用		95
4.1 原始数据中的标记分布		95
4.2 基于先验知识的标记分布		104
4.3 从数据集中学到的标记分布		109
5 结束语		120
参考文献		120
软件缺陷挖掘	黎 铭	125
1 引言		125
2 基于编程模式挖掘软件缺陷		126
3 基于缺陷标注挖掘软件缺陷		127
4 基于缺陷报告挖掘软件缺陷		131
5 结束语		135
参考文献		136



稳健的矩阵回归模型与方法^{*}

杨 健 罗 雷

南京理工大学模式计算与应用实验室, 南京 210094

1 引言

随着网络和信息技术的不断发展,全球数据量呈现爆炸式增长。特别地,人类在最近两年产生的数据量相当于之前产生的全部数据量。这些海量、复杂的数据已经对社会经济、政治、文化以及生活等方面产生了深远的影响。如何有效地分析它们使其更好地为人类服务是一个亟待解决的问题。

近年来,机器学习方法已经渗透到了数据分析的各个方面,成为了其中的重要组成部分。它利用已知数据来学习和推理其中未知的、潜在的概率分布等重要信息,解释数据样本中变量(或特征)之间的关系。也就是说,它能从庞大的数据中揭示出数据的内在规律或本质结构。这种本质结构可方便人类对数据的理解,提高人类对数据的驾驭能力。为实现这一任务,我们首先要熟知数据分析中可能面临的问题。一方面,在实际的信号和图像采集与处理中,数据的维数越高,给数据的采集和处理带来越多的限制。大规模的数据不仅存在着较多的冗余、无关的属性,还会带来维数的灾难。另一方面,实际中数据往往结构复杂,例如采集的数据本身不完整或者存在大量的噪声,这导致了一些传统的数据处理技巧也许失效。

基于线性回归(linear regression, LR)分析的方法一直是机器学习领域的研究热点。它的目的是估计测试数据与训练数据之间的关系。值得注意的是,这种关系可理解成数据的一种内在结构。为避免过拟合,一个正则项需要强加到 LR 模型中。基于 L_1 范数的正则项和基于 L_2 范数的正则项是目前使用最广泛的两种正则项。 L_2 范数正则化的线性回归一般称为脊回归(ridge regression)。该方法使用 L_2 范数的平方来约束表示系数,并使用同样的方式来刻画表示残差。最近, I. Naseem 等人^[1]将此方法应用于人脸识别并由

^{*} 本文得到国家自然科学基金(91420201 和 61472187)资助。

此提出了线性回归分类器(LRC)。在这个方法中,他们先通过脊回归模型获得测试样本在每类训练样本中的表示系数,然后将测试样本分在离其最近的那一类。事实上,一些以往的工作,诸如最近邻特征线^[2]、最近邻特征面以及最近邻特征空间^[3]都可看作 LRC 的不同变体。

L_1 范数正则化的线性回归称为 Lasso,它已被广泛地应用于稀疏表示中。例如, J. Wright 等人^[4]引入了稀疏表示分类器(SRC)。SRC 使用所有训练样本作为一个字典来表示一个测试样本,并且假设表示系数是稀疏的,即这些稀疏的非零表示系数应该集中在与测试样本具有相同类型标签的训练样本上。为了得到一个稳健的模型,他们进一步假设噪声也是稀疏的,由此建立了推广的 SRC 模型。这个模型对随机像素损坏与块状遮挡的噪声具有一定的稳健性。A. Wagner 等人^[5]进一步推广了 SRC 模型,并将人脸对齐和识别统一到一个框架中。另一方面,一些近期的工作主要集中在探究稀疏在图像分类中的角色^[6~9]上。J. Yang 等人^[8]对 SRC 的合理性提供了一些理论性的解释并得出一个结论:SRC 的有效性是由 L_1 约束带来的,而不是 L_0 约束(原本的稀疏限制)的作用。L. Zhang 等人^[9]分析了 SRC 的运行原理且发现:合作表示策略比基于 L_1 范数的系数约束扮演了更为重要的角色。于是,基于脊回归的思想,他们提出了合作表示分类器(collaborative representation classifier, CRC)。可是,CRC 虽然具有速度上的优势,但不能提供一种噪声移除的机制,所以它不是图像分类的一种稳健的回归方法。

在 LRC、CRC 以及 SRC 模型中,表示余差使用误差向量的 L_2 范数或者 L_1 范数来约束。从统计学来讲, L_2 范数为服从高斯分布的数据提供了一种最优的刻画。但是,在现实的人脸识别中,表示误差的分布是很复杂的,高斯分布无法完美地拟合它们^[10,11]。所以,以上方法对现实中的噪声并不稳健。为解决这类问题,稳健回归分析^[12]应运而生。该方法是对传统的最小均方法在异常点存在情况下的一种改进,它提供给我们这样一个信息:什么是一个有效的观测,并且哪一种观测应被剔除。稳健回归的最初目的是用来拟合一个能够表示出大多数数据的模型。

M-估计是稳健回归中一种常用的技巧,在 19 世纪 60 年代由 P. J. Huber 等人^[13]提出。选择权重函数对稳健回归的性能具有重要影响,由此衍生了许多稳健回归模型,例如,推广的稀疏表示模型^[4]、稳健稀疏编码模型(RSC)^[10]、基于相关熵的稀疏表示模型(CESR)^[14]以及稳健的线性回归分类器(RLRC)^[15]。特别地,RLRC 本质上是一种 Huber 估计方法。此外,R. He 等人^[16]统一了两种稀疏稳健回归模型,即在误差纠正中以 SRC 为代表的相加模型和在误差探测中以 CESR 为代表的相乘模型。R. He 等人^[16]还通过定义不同的半平方函数建立了一种半平方的人脸识别框架。该框架可同时执行误差纠正与探测任务。M-估计的稳健性体现在:异常像素值会被赋予一个小的权值来减小它们在编码过程中的影响,而对其他元素则赋予较大的权值。所以,与传统的 minimum 方法相比,稳健回归更适合处理实际中的噪声。但是,以上所提到的稳健回归方法都使用了

一维的基于像素误差模型,即误差图像的每个像素都被独立地刻画。它们并没有考虑图像的二维结构信息。

不同于以上方法,本文将介绍几种全新的稳健回归模型。它们将图像的结构信息融入到建模中。这些方法包括:基于核范数的稳健矩阵回归,基于推广幂指数分布的稳健矩阵回归,基于核- L_1 范数的联合矩阵回归,以及基于树结构核范数的稳健矩阵回归。

2 基于核范数的稳健矩阵回归

基于像素的一维模型(例如, SRC, RSC 和 CESR)存在两个问题:首先,这类模型假设误差图像的每一个像素是独立同分布的。对于随机的像素损坏,由于噪声被独立地添加在每个像素上,所以这个假设是合理的。但是,对于许多实际的人脸变化,诸如遮挡、伪装或者由光照改变产生的阴影,这个假设不再成立。例如,由黑色围巾造成的遮挡,其中的像素值是 0,那么,遮挡部分中理想的表示误差像素是相关的,因为现实图像中局部区域的像素一般是高度相关的。所以,使用基于像素的一维误差模型(诸如 SRC^[4], RSC^[11], RLRC^[15]等)来解决带有遮挡的图像分类任务在理论上是令人质疑的。

其次,单独地刻画误差的每个像素忽视了误差图像的整体结构,因为误差图像中的所有像素误差也许包含了有意义的结构信息(例如误差图像的秩)。在基于回归分析的人脸识别方法中,可以使用训练图像来表示一张测试图像。理想情况下,误差图像应该是一个零矩阵。因此,它自然是低秩的。在一般情形下,测试图像中存在光照改变和遮挡。事实上,光照和遮挡是影响人脸识别性能的关键因素。实际的光照改变,特别是诸如阴影的部分光照改变一般会导致一个低秩(或者近似低秩)的误差图像,这与满秩的原始图像形成了鲜明的对比。诸如眼镜与围巾遮挡也会导致一个低秩的误差图像。以上所提及的回归方法可独立地处理每一个误差像素,因此无法反映这类结构信息。

为了充分利用误差图像的低秩结构信息,这里介绍我们所提出的一种基于二维图像矩阵的误差模型。相反地,以前的方法,诸如脊回归、Lasso 或者稳健回归,都是基于向量的途径。也就是说,为了处理矩阵形式的二维图像,我们在使用诸如回归的方法时,必须提前将图像转化成向量。在转化阶段,一些结构信息(例如,误差图像的秩)也许会丢失。本节的矩阵回归方法不需要矩阵到向量的转化过程。它通过最小化余差图像的秩来挖掘图像的结构信息,并以此来决定回归系数。为便于优化,秩函数最小化问题一般可转化成核范数最小化问题。根据这一想法,表示余差图像的最小核范数可作为一个准则。因此,此方法也称为基于核范数的矩阵回归(nuclear norm based matrix regression, NMR),其具体形式如下:

给定 n 个图像矩阵 $\mathbf{A}_1, \dots, \mathbf{A}_n \in R^{p \times q}$ 和一个图像矩阵 $\mathbf{B} \in R^{p \times q}$, 则 $\mathbf{B}$ 可被 $\mathbf{A}_1, \dots, \mathbf{A}_n$ 线性地表示,即

$$\mathbf{B} = x_1 \mathbf{A}_1 + x_2 \mathbf{A}_2 + \cdots + x_n \mathbf{A}_n + \mathbf{E} \quad (1)$$

其中, $x_1, x_2, \dots, x_n$ 表示稀疏的集合, $\mathbf{E}$ 表示余差。

定义从 R^n 到 $R^{p \times q}$ 的映射如下:

$$\mathbf{A}(\mathbf{x}) = x_1 \mathbf{A}_1 + x_2 \mathbf{A}_2 + \cdots + x_n \mathbf{A}_n. \quad (2)$$

那么, 公式(1)变为

$$\mathbf{B} = \mathbf{A}(\mathbf{x}) + \mathbf{E} \quad (3)$$

公式(3)或者公式(1)给出了线性矩阵回归模型的一般形式, 它们与传统的线性向量模型形成了鲜明的对比。

在许多应用中, 余差图像 $\mathbf{B} - \mathbf{A}(\mathbf{x})$ 在最优解处是典型低秩的(或近似低秩的)。考虑到核范数是秩函数在单位球上的凸包, 本节将解决以下核范数最小化问题来评估回归系数:

$$\min_{\mathbf{x}} \|\mathbf{B} - \mathbf{A}(\mathbf{x})\|_* \quad (4)$$

其中, $\|\mathbf{X}\|_*$ 表示矩阵 $\mathbf{X}$ 的核范数, 即 $\mathbf{X}$ 的所有非零奇异值之和。此外, 我们可在公式(4)上增加一个类似脊回归的正则项, 由此获得了正则化的矩阵回归模型:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{E}} \|\mathbf{E}\|_* + \frac{1}{2} \lambda \|\mathbf{x}\|_2^2 \\ \text{s. t. } \mathbf{B} = \mathbf{A}(\mathbf{x}) + \mathbf{E} \end{aligned} \quad (5)$$

其中, $\lambda > 0$ 是一个平衡参数。公式(5)可使用交替方向乘子法(ADMM)来解决, 其主要迭代过程被总结在算法 1 中。这里, $D_\zeta(\mathbf{Q})$ ($\zeta > 0$) 表示矩阵 $\mathbf{Q}$ ($\in R^{p \times q}$) 的奇异值收缩算子(详细定义见文献[17]), $\text{Vec}(\cdot)$ 是一个将矩阵转化为向量的操作符, $\mathbf{I}$ 是一个单位矩阵。此外, 我们可以使用文献[18]中类似的途径来设置算法 1 的终止条件。

算法 1 使用 ADMM 解决 NMR

输入: 一组图像矩阵 $\mathbf{A}_1, \dots, \mathbf{A}_n$ 和一个图像矩阵 $\mathbf{B} \in R^{p \times q}$, 模型参数 λ 以及 μ 。

1. 令 $\mathbf{A} = [\text{Vec}(\mathbf{A}_1), \dots, \text{Vec}(\mathbf{A}_n)]$ 。计算 $\mathbf{M} = \left(\mathbf{A}^\top \mathbf{A} + \frac{\lambda}{\mu} \mathbf{I} \right)^{-1} \mathbf{A}^\top$;
 2. $\mathbf{E}^0 = -\mathbf{B}$, $\mathbf{Z}^k = \mathbf{0}$, $k = 0$;
 3. 更新 $\mathbf{x}$: $\mathbf{x}^{k+1} = \mathbf{M} \mathbf{g}$, 其中 $\mathbf{g} = \text{Vec} \left(\mathbf{B} - \mathbf{E}^k - \frac{1}{\mu} \mathbf{Z}^k \right)$;
 4. 更新 $\mathbf{E}$: $\mathbf{E}^{k+1} = D_{1/\mu} \left(\mathbf{B} - \mathbf{A}(\mathbf{x}^{k+1}) - \frac{1}{\mu} \mathbf{Z}^k \right)$;
 5. 更新 $\mathbf{Z}$: $\mathbf{Z}^{k+1} = \mathbf{Z}^k + \mu (\mathbf{A}(\mathbf{x}^{k+1}) + \mathbf{E}^{k+1} - \mathbf{B})$;
-

输出: 最优回归系数向量 $\mathbf{x}^{k+1}$ 。

但是,传统的 ADMM 方法仅能获得一个 $O(1/k)$ 的收敛率^[19],其中 k 代表迭代次数。所以,我们将使用一种加速的 ADMM^[20] 来获得模型(5)的最优解,此算法能达到平方收敛。考虑到加速的(Fast)ADMM 的收敛条件要求目标函数是强凸的,我们在模型(5)上添加一个平方项,从而得到一个近似的 NMR 模型如下:

$$\begin{aligned} \min_{\mathbf{E}, \mathbf{x}} \quad & \|\mathbf{E}\|_* + \gamma \|\mathbf{E}\|_F^2 + \frac{1}{2}\theta \|\mathbf{x}\|_2^2 \\ \text{s. t.} \quad & \mathbf{B} = \mathbf{A}(\mathbf{x}) + \mathbf{E} \end{aligned} \quad (6)$$

其中, $\gamma, \theta > 0$ 。J. Yang 等人^[17]已经说明,在一定条件下,模型(5)与模型(6)的解是一致的。算法 2 列出了解决公式(6)的详细过程,其终止条件类似于算法 1。

算法 2 使用 Fast ADMM 解决 NMR

输入: 一系列图像矩阵 $\mathbf{A}_1, \dots, \mathbf{A}_n$ 和一张图像矩阵 $\mathbf{B} \in \mathbb{R}^{p \times q}$, 模型参数 λ, μ 和 γ 。

1. 令 $\mathbf{A} = [\text{Vec}(\mathbf{A}_1), \dots, \text{Vec}(\mathbf{A}_n)]$ 。计算 $\mathbf{M} = \left(\mathbf{A}^T \mathbf{A} + \frac{\theta}{\mu} \mathbf{I} \right)^{-1} \mathbf{A}^T$, 其中 $\theta = \lambda(1 + \gamma)$;
 2. $\mathbf{x}^0 = \hat{\mathbf{x}}^0 = \mathbf{0}, \mathbf{Z}^0 = \hat{\mathbf{Z}}^0 = \mathbf{0}, \alpha^0 = 1, k = 0$;
 3. 更新 $\mathbf{E}$: $\mathbf{E}^{k+1} = \frac{\mu}{\mu + 2\gamma} D_{1/\mu} \left(\mathbf{B} - \mathbf{A}(\hat{\mathbf{x}}^k) - \frac{1}{\mu} \hat{\mathbf{Z}}^k \right)$;
 4. 更新 $\mathbf{x}$: $\mathbf{x}^{k+1} = \mathbf{M} \mathbf{g}$, 其中 $\mathbf{g} = \text{Vec} \left(\mathbf{B} - \mathbf{E}^{k+1} - \frac{1}{\mu} \hat{\mathbf{Z}}^k \right)$;
 5. 更新 $\mathbf{Z}$: $\mathbf{Z}^{k+1} = \hat{\mathbf{Z}}^k + \mu (\mathbf{A}(\mathbf{x}^{k+1}) + \mathbf{E}^{k+1} - \mathbf{B})$;
 6. 更新 α : $\alpha^{k+1} = \frac{1 + \sqrt{1 + 4(\alpha^k)^2}}{2}$;
 7. 更新 $\hat{\mathbf{x}}$: $\hat{\mathbf{x}}^{k+1} = \mathbf{x}^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}} (\mathbf{x}^{k+1} - \mathbf{x}^k)$;
 8. 更新 $\hat{\mathbf{Z}}$: $\hat{\mathbf{Z}}^{k+1} = \mathbf{Z}^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}} (\mathbf{Z}^{k+1} - \mathbf{Z}^k)$;
-

输出: 最优的回归系数向量 $\hat{\mathbf{x}}^{k+1}$ 。

在 NMR 的建模动机中, J. Yang 等人^[17]使用了“低秩”的概念,这是源于对由遮挡造成的误差图像的一个直观的描述。实际上, NMR 并不局限于“低秩”的误差图像,因为它不直接优化秩函数而是优化误差的核范数。比起秩函数,核范数能对误差图像提供一种更加可行的刻画,因为它在刻画近似低秩的误差图像的时候仍然是有效的。这些误差图像也许是代数满秩的,但它的奇异值接近于 0。例如,在光照改变的情形下,假如人脸是几何光滑的,那么误差图像中的像素则是高度相关的。所以,由光照改变造成的误差图像一般是近似低秩的。