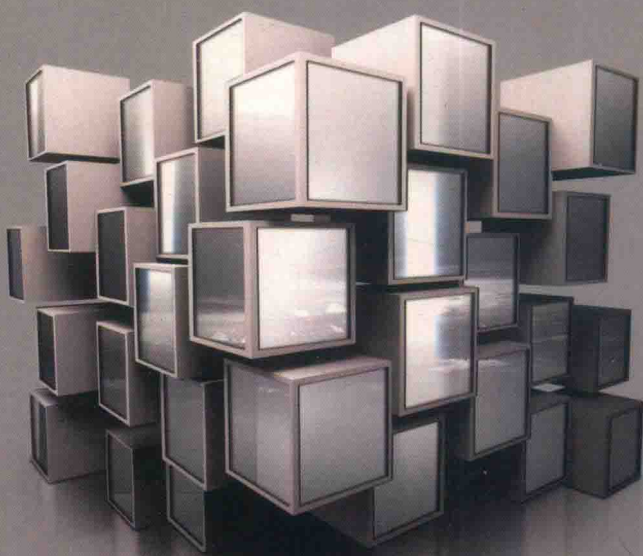


张量数据的特征提取与分类

ZHANGLIANG SHUJU DE TEZHENG TIQU YU FENLEI

张帆 许丽 孙帅 著



中国水利水电出版社
www.waterpub.com.cn

张量数据的特征提取与分类

张帆 许丽 孙帅 著



中国水利水电出版社
www.waterpub.com.cn

· 北京 ·

内 容 提 要

本书的主要研究内容是在模式识别应用领域中,提出新的基于张量数据的特征提取和分类算法,并且对这些张量型算法进行详细的理论推导和性能分析,在实验中验证所提出算法的优越性。主要内容来自于作者多年的研究成果,使读者能够比较全面地了解张量分析的基本知识以及张量型算法在模式识别领域的研究、发展和应用。本书理论联系实际,集知识性、专业性、操作性、技能性为一体。本书的读者对象主要为模式识别、人工智能领域的研究人员以及高等院校高年级的学生和研究生。

图书在版编目(CIP)数据

张量数据的特征提取与分类 / 张帆, 许丽, 孙帅著

—北京: 中国水利水电出版社, 2017.8

ISBN 978-7-5170-5755-0

I. ①张… II. ①张… ②许… ③孙… III. ①张量分析—数据采集 IV. ①O183.2

中国版本图书馆 CIP 数据核字(2017)第 193620 号

责任编辑: 陈 洁

封面设计: 王 斌

书 名	张量数据的特征提取与分类 ZHANGLIANG SHUJU DE TEZHENG TIQU YU FENLEI
作 者	张 帆 许 丽 孙 帅 著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路 1 号 D 座 100038) 网址: www.waterpub.com.cn E-mail: mchannel@263.net (万水) sales@waterpub.com.cn 电话: (010) 68367658 (营销中心)、82562819 (万水)
经 售	全国各地新华书店和相关出版物销售网点
排 版	北京万水电子信息有限公司
印 刷	三河市天润建兴印务有限公司
规 格	170mm×230mm 16 开本 12.25 印张 215 千字
版 次	2018 年 1 月第 1 版 2018 年 1 月第 1 次印刷
印 数	0001—2000 册
定 价	48.00 元

凡购买我社图书,如有缺页、倒页、脱页的,本社营销中心负责调换

版权所有·侵权必究

前 言

在许多应用领域，特别是在云计算、移动互联网、大数据应用等方面，会产生大量的高维高阶的数据，采用张量的数学形式能够恰当地表示这些具有多维结构的数据。这些数据往往含有大量的冗余信息，需要对其进行有效的降维。在模式识别中，特征提取（降维）和分类是两个关键步骤。大多经典的特征提取和分类的算法都是基于向量数据的，处理张量数据时需要将其向量化。张量数据向量化的过程会破坏数据的内部结构，维数会显著增加，使算法的计算量和复杂度也显著增加。因此，对于张量型特征提取和分类算法的研究具有重要的应用价值。

本书以张量分析的数学理论为工具，针对基于张量数据的特征提取和分类算法现有的一些问题进行了研究，并提出新的张量型特征提取和分类算法。

本书内容自成体系，共分为6章，第1章绪论。本章主要论述了本书的研究背景和意义，介绍了张量数据的表示和

定义，详细分析了当前基于张量数据的特征提取和分类方法的研究发展现状，并指出目前存在的一些问题，以及向量型和张量型算法的比较，同时介绍了在本书实验中所用到的数据库。

第2章相关理论基础。本章主要是对在后续章节中所用到的数学和经典算法的相关理论知识介绍。尤其对主成分分析、线性判别分析、最大散度差、极限学习机和岭回归等这些特征提取和分类算法进行了分析和说明，为后续章节张量型的特征提取和分类算法的研究提供了数学分析工具，奠定了理论基础。

第3章基于 MPCA 和 GTDA 的张量型特征提取算法。本章首先介绍了 MPCA 和 GTDA 两种算法的基本原理，然后比较 MDA 和 GTDA 的优缺点，表明了 MPCA + GTDA 二者联合的优越性，并给出了算法的详细流程。同时，对 GTDA 算法的一些先决条件和性质，如收敛性、投影后空间的维度、初始化条件等问题进行了研究。最后，通过在人脸和步态数据库中的实验，表明了 MPCA + GTDA 作为一种新的张量型特征提取算法的优越性。

第4章张量型极限学习机分类算法。本章阐述了如何把一维分类算法极限学习机，分别扩展到二维和张量领域，即二维极限学习机和张量极限学习机分类器的推导过程，从而使其能够直接对二阶和更高阶张量数据进行分类。最后，通过实验结果表明该算法的分类效果。

第5章多线性多秩回归分类算法 (MMRR)。本章首先简要阐述了采用秩1分解得到的张量型分类算法存在的一些不足的前提下,提出了基于多秩分解的多线性多秩回归算法,并且对该算法进行了详细的理论推导,列出了具体的计算流程。同时,对该算法性能有重要影响的一些先决条件,如参数选择、初始化条件、收敛条件等,结合实验结果进行了深入的探讨和分析。最后在多个张量型数据库上验证了该算法的分类效果。

第6章总结与展望。即对全文的内容进行总结,并对张量型特征提取和分类算法进一步的工作进行了展望。

本书由张帆、许丽、孙帅著,由于作者水平有限,编写时间仓促,书中错误和不妥之处在所难免,请读者和专家批评指正。

作者

2017年6月

目 录

前言

第 1 章 绪论	1
1.1 研究背景与意义	2
1.2 传统特征提取和分类方法	7
1.3 基于张量数据的特征提取和分类方法研究现状	9
1.3.1 数据的张量表示	10
1.3.2 基于张量数据的特征提取算法	12
1.3.3 基于张量数据的分类器	17
1.3.4 张量型算法与向量型算法比较	19
1.4 实验数据库	21
1.5 本章小结	24
第 2 章 相关理论基础	25
2.1 张量理论	26
2.1.1 张量的符号表示与展开	27

2.1.2 张量的基本运算	28
2.1.3 张量的秩和分解	30
2.1.4 张量 - 张量投影与张量 - 矢量投影	32
2.2 相关向量型特征提取算法	36
2.2.1 PCA 算法	36
2.2.2 LDA 和 MSD 算法	38
2.3 极限学习机分类算法	40
2.4 岭回归分类算法	44
2.5 本章小结	47
第 3 章 基于 MPCA 和 GTDA 的张量型特征提取算法	48
3.1 引言	49
3.2 融合 MPCA 和 GTDA 的特征提取算法	51
3.2.1 多线性主成分分析	52
3.2.2 广义张量判别分析	53
3.2.3 MPCA 与 GTDA 算法融合	56
3.3 GTDA 算法的先决条件分析	60
3.3.1 投影后特征子空间的维数确定	61
3.3.2 迭代初始化的条件	62
3.3.3 迭代终止条件	65
3.3.4 收敛性	66
3.4 实验结果分析	67
3.4.1 GTDA 初始化条件的选择	67

3.4.2	GTDA 与 MDA 的收敛性分析	73
3.4.3	FERET 人脸库的实验和结果分析	75
3.4.4	步态数据库的实验和结果分析	79
3.4.5	AR 彩色人脸数据库的实验和结果分析	81
3.5	本章小结	84
第 4 章	张量型极限学习机分类算法	86
4.1	引言	87
4.2	二维极限学习机分类器	88
4.2.1	2D-ELM 算法原理	89
4.2.2	2D-ELM 实验结果分析	92
4.3	张量极限学习机分类器	97
4.3.1	TELM 算法原理	97
4.3.2	TELM 的实验结果分析	100
4.4	本章小结	104
第 5 章	多线性多秩回归分类算法(MMRR)	106
5.1	引言	107
5.2	MMRR 算法原理	113
5.2.1	算法推导过程	113
5.2.2	迭代优化过程	114
5.3	MMRR 算法先决条件分析	121
5.3.1	MMRR 算法的收敛性	121
5.3.2	MMRR 算法的初始化条件和计算量比较	123

5.3.3	MMRR 算法中各参数的确定	125
5.4	实验结果分析	128
5.4.1	收敛性分析	128
5.4.2	不同初始化方法对算法分类性能的影响	133
5.4.3	参数变化对算法分类性能的影响	134
5.4.4	MMRR 分类器的分类性能	137
5.5	本章小结	140
第 6 章	总结与展望	142
6.1	全文总结	143
6.2	工作展望	146
	缩写符号对照表	149
	数学符号对照表	159
	参考文献	165

1

绪论

第 1 章 绪论

本书将从一些传统的特征提取和分类算法出发,如主成分分析、极限学习机分类和线性回归分类等,提出新的基于张量数据的特征提取和分类的算法,并将其应用在实际问题中。由于特征提取和分类是模式识别系统中的两个关键步骤,本章首先介绍模式识别系统及其研究背景和意义,以及传统的特征提取和分类算法,基于张量的特征提取和分类算法的研究现状,后续研究中所用到的数据库,最后为本书的研究内容与框架。

1.1 研究背景与意义

当今社会的信息化进程不断发展,科技工作面临的研究对象也与日俱增,如何能准确高效地寻找到具有价值的信息并对其准确定位和分类,成为当下关注和研究的热点问题。目

前已有许多学科从不同途径和角度对此进行了深入的研究,这其中就包括模式识别技术,而特征提取和分类则是模式识别系统中的两个主要步骤,是决定模式识别系统优劣性的重要因素之一。

模式识别也可称为模式分类,是根据研究对象的特性和已知类别进行匹配的过程。这些对象与应用领域有关,可以是图像、信号波形或任何可测量并需要分类的对象。随着模式识别技术的日益丰富和完善,在工业、农业、国防、医学、气象、天文等领域都得到了成功应用^[1-4]。其中,近年来发展最快的应属于计算机视觉领域,如文本识别、生物特征识别(指纹、掌纹、人脸、虹膜、视网膜、人体行为、表情等)、医学图像识别、遥感图像识别等。

典型的模式识别系统框图如图 1.1 所示,主要由信息获取、预处理、特征提取和分类这四个既相互联系又明显区分的步骤组成。根据处理对象的不同,可选用各种传感器、测量装置或输入装置进行信息的采集和获取。在信息获取完成后,常需要对采集样本进行预处理操作,以便抽取出受干扰因素影响比较小的待识别样本。

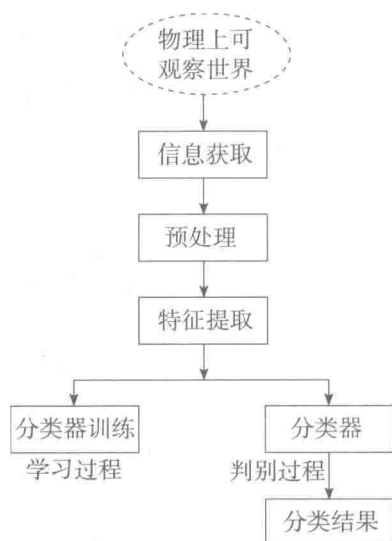


图 1.1 模式识别系统框图

在对客观世界进行信息采集的过程中,总是尽可能多地采集各项数据以保证样本模式的真实性和完整性,从而导致了采集到的样本具有过大维数,使处理数据的过程复杂而耗时。考虑到不同的特征对于不同的分类要求具有不同的意义,可通过特征提取过程将对识别有明显作用的特征提取出来,从而压缩模式的维数,提高识别的效率和准确度。特征提取后的特征空间用于分类,分类即是将未知类别属性的样本判定归属为某一类型。为了使分类器有效地进行分类判决,首先应对分类器进行训练,也就是分类器的学习过程。分类器的学习过程需经过多次重复,不断纠正错误,才能取得满足要求的类别识别率。经过学习过程的分类器,对待识别特征样本进行判

别,确定其类别归属,这一过程称为判别过程或者测试过程。如果训练样本的所属类别是已知的,可在机器学习中起监督引导的作用,称为有监督的模式分类;如果训练样本的所属类别未知,称为无监督的模式分类;如果已知部分样本的类别属性,称为半监督的模式分类。

在新兴的应用领域,特别是在云计算、移动互联网、大数据应用等方面,产生大量高维高阶的数据。这些数据无论大小,大多都具有多维的结构,可以用张量的数学形式来表示。譬如,计算机视觉和图像处理中的灰度图像,神经科学和生物医学领域中的多通道心电图信号,生物信息学中的基因图,都可归类于二阶张量;人机交互中用于动作和姿态识别的灰度视频,人脸识别中的彩色人脸图像,则可用三阶张量来表示。在传感器网络、医学图像分析等领域中,还存在着更高阶次的张量数据。这些张量数据具有很大的维数,并含有大量的冗余信息。而传统的特征提取和分类方法大多都以向量数据为处理对象,面对这些高维高阶的张量数据时,需要将张量数据向量化,而这一做法会带来“维数灾难”。主要有以下两个缺点。

(1) 对张量的向量化即是将张量中各列或行向量依次首尾相连,这一操作将导致高维向量的产生,增加算法的计算量和复杂度,占用更多存储空间,以及容易出现过拟合现象。

(2) 对张量的直接向量化,会改变数据原有的数据结构,丢失大量有用的数据结构信息,譬如原有数据中特征和特征之间的空间位置信息。而内部信息的丢失,会影响最终的识别

效果。

因此,为了更好地在模式识别应用中处理张量数据,避免对其进行向量化操作,基于张量数据的特征提取和分类算法在近年来得到不断的研究和发展。研究成果表明,采用张量型的算法直接处理张量数据,不仅可以保留数据的内部结构信息,还可以在张量化过程中有效控制优化问题中变量的个数,从而减少在学习过程中出现过拟合现象。基于张量数据的特征提取和分类算法优点众多,已得到科学界广泛地研究和应用,理论研究和实际应用都得到了快速的发展。

目前关于张量型特征提取和分类算法的研究,主要集中在由向量型算法向张量型算法扩展的方向上,取得的一些成果和进展将在 1.3 节的研究现状中进行介绍。在现有的张量型特征提取和分类算法中,仍然存在有一些问题。譬如,有些算法在迭代优化的计算过程中不收敛,从而造成算法的计算结果具有波动性,不能获得稳定的识别结果;有些采用张量秩 1 分解的张量型算法由于自由变量太少,存在欠拟合的问题,也会影响算法最终识别效果的情况。这些都是在张量型学习算法的研究中可能遇到的问题。针对这些问题的研究和改进,对张量型特征提取和分类算法的发展和推广也具有重要的意义。

1.2 传统特征提取和分类方法

用于特征提取和分类的算法众多,在计算机视觉应用领域,主要有基于子空间分析的方法、基于神经网络的方法、基于支持向量机的方法和基于 Gabor 小波特征的方法等^[5]。其中,基于子空间分析的特征提取算法通过将样本投影到某最优解子空间,以便达到降低样本维度和提取更具有鉴别度特征的目的。由于映射方法有线性和非线性两种,所以子空间分析法分为线性和非线性两类。用于降维的线性子空间方法主要包括主成分分析(Principal Component Analysis, PCA)^[6]、独立成分分析(Independent Component Analysis, ICA)^{[7][8]}、线性判别分析(Linear Discriminant Analysis, LDA)^[9]、典型相关分析(Canonical Correlation Analysis, CCA)^[10]、偏最小二乘分析(Partial Least Squares Analysis, PLSA)^[11]和最大散度差(Maximum Scatter Difference, MSD)^[12]等方法。非线性子空间方法主要有基于核映射的非线性子空间方法,流形学习(Manifold Learning)等方法。核映射的非线性算法主要是在原有线性特征提取算法基础上引入核方法,其中被广泛应用的有核主成分分析(Kernel Principal Component Analysis, KPCA)^[13]、核典型相关分析(Kernel Canonical Correlation