首都经济贸易大学出版基金资助

流形学习若干关键问题与算法研究

LIUXING XUEXI RUOGAN
GUANJIAN WENTI YU SUANFA YANJIU

古楠楠　樊明宇
王　迪　韩　志　◎ 著

首都经济贸易大学出版社
Capital University of Economics and Business Press

流形学习若干关键问题与算法研究

古楠楠　樊明宇　王　迪　韩　志　著

首都经济贸易大学出版社

·北京·

图书在版编目(CIP)数据

流形学习若干关键问题与算法研究/古楠楠等著. —北京:首都经济贸易大学出版社,2015.7

ISBN 978-7-5638-2374-1

Ⅰ. ①流… Ⅱ. ①古… Ⅲ. ①数据处理 ②数据模型—建立模型 Ⅳ. ①TP274②TP311.13

中国版本图书馆 CIP 数据核字(2015)第 152235 号

流形学习若干关键问题与算法研究

古楠楠 樊明宇 王勉 舒志 著

出版发行	首都经济贸易大学出版社
地　　址	北京市朝阳区红庙(邮编 100026)
电　　话	(010)65976483　65065761　65071505(传真)
网　　址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经　　销	全国新华书店
照　　排	首都经济贸易大学出版社激光照排服务部
印　　刷	北京京华虎彩印刷有限公司
开　　本	880 毫米×1230 毫米　1/32
字　　数	163 千字
印　　张	6.375
版　　次	2015 年 7 月第 1 版　2015 年 7 月第 1 次印刷
书　　号	ISBN 978-7-5638-2374-1/TP·39
定　　价	20.00 元

前　言

数据降维(或称为维数约简)是数据挖掘与数据建模的基本问题,而流形学习是近年来发展起来的数据降维最引人注目的方法之一。流形学习是缓解高维数据的“维数灾难”等问题的有效方法,具有十分重要的理论研究价值和实际应用价值。首先,它涉及数学、计算机科学、信息科学、生物认知等多个领域,是新兴的前沿性的交叉科学研究。其次,流形学习促进了数学中多个领域的交叉。最后,流形学习在机器学习、数据挖掘以及模式识别等领域都有重要的应用。尽管流形学习在理论和应用上都取得了成功,但仍面临很多挑战性的问题,本文针对其中的一些关键问题进行了研究,主要包括:

1. 针对流形学习方法的统一理解与综述问题,我们提出使用流形正则化框架。为了获得从高维表示空间到低维本质空间的降维映射,该框架力图拟合先验的低维表示指导信息,同时考虑降维映射的函数复杂度及其保持数据结构化信息的程度。依据此框架,我们将线性的与非线性的、无监督的与有监督的、单类的与多类的各种流形学习算法联系起来,从一个统一的角度理解它们,同时从此角度对它们进行了综述,并进一步探讨了它们之间的共性与差异。

2. 针对流形学习的降维维数选择问题,我们将近年来出现的本质维数估计方法进行了分类和综合比较研究。按其基本构造

原理的不同，我们将已有的本质维数估计方法分为五类：基于熵图的方法，基于分形维数的方法，基于 k - 近邻距离的方法，基于特征值的方法及其他方法。本文依此对这些方法进行了综述，并通过在一系列典型数据集上的实验对它们的应用性能进行了分析比较，利用这些比较结果，本文得出了针对不同应用问题的本质维数估计方法选择策略。

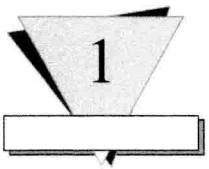
3. 针对传统流形学习方法对于分布在非连通流形（或多流形）上的数据经常失效的问题，我们以分解 - 整合方法这一非连通流形学习算法为切入点，在其基础上提出了过渡曲线降维算法。该算法通过构建非连通类边缘最近点间的平滑过渡曲线来建立类间更客观有效的流形连接关系，进而保证良好的非连通数据全局降维形态。该方法明显地改进了分解 - 整合算法的表现，特别地，分解 - 整合算法所遭遇的边缘问题得到了有效的解决。这一工作扩展了分解 - 整合算法的应用范围。

4. 针对半监督流形学习问题，即面对半监督分类任务时传统流形学习算法无法处理多流形数据、很难引入类别标签、缺乏显性映射的问题，我们采取“根据有标签数据的类别标签随机生成先验低维表示并对其进行拟合，同时保持数据稀疏结构”的策略，提出一种判别性保稀疏投影算法（Discriminative Sparsity Preserving Projection，DSPP）。DSPP 是针对多流形数据设计的，具有显性的降维映射，还从数据稀疏表示中继承了较高的判别性能。实验证明，DSPP 在半监督分类方面有明显优势。

5. 针对流形学习在半监督分类中的应用问题，我们提出了基于稀疏化假设的核稀疏正则化（Kernel - based Sparse Regularization，KSR）半监督分类方法。数据的稀疏化表示，也称之为稀疏编码，本质上是数据在少数非欧坐标轴上面的线性表示

系数,因此可以理解为数据降维的一种特殊形式。KSR 的主要思想是在核特征空间中计算数据的稀疏表示,从而能够避免对非图像数据直接应用稀疏表示时可能会遭遇的“l^2 - 范数问题”,然后通过在流形正则化框架下保持数据的稀疏表示系数来得到分类函数。KSR 能够自适应进行邻域选择,具有较高判别性能,且具有显性的分类函数,能够处理在线分类任务。标准数据集上的实验结果证明了 KSR 的有效性。

目 录



绪论

1.1 研究背景与意义

1.1.1 研究背景

随着信息获取技术与互联网技术的发展，在自然科学和社会科学的各个研究领域，每时每刻都在快速地产生海量的数据，而隐藏在它们背后的内在关系与规律却难以被直观发现，因此人们正逐步陷入“数据丰富，知识匮乏”的尴尬境地[93]。如何从海量数据中有效地挖掘出我们所需要的知识，进而指导相关科学研究、系统优化等，是信息科学的基本问题，也是当前知识经济社会面临的巨大挑战。

在很多实际应用中，如医学数据检测[4][95]、基于基因的疾病诊断[141][223]（经常需要对成千上万基因进行并行处理）、人脸识别[253]、图像分类[61][154]、语音识别[188]、网页分类[169]等，人们所采集到的数据通常是繁杂的、高维的以及非结构化的，数据的这种高维性往往掩盖了其本质特征。且面对高维数据时，传统的数据分析方法往往会遭遇“维数灾难”（Curse of Dimensionality）[18]，即在缺乏简化数据的前提下，要在给定的精度下准确地对某些变量的函数进行估计，我们所需要的样本数量会随着样本维数的增加而呈指数式增长。这会导致多种问题，如采样问题：设数据维数为 1 时为保证采样密度 ρ 所需的采样数为 N，则在数据维数为 D 时为保证同样采样密度所需的采样数为 N^D，当 D 很大时，采样数是一个天文数字。与“维数灾难”相关的另外一个几何现象是“空空间现象”（Empty Space Phenomenon）[185]，即高维空间本质上是稀疏空间，这使得数据在低维数据空间所具备的许多特性在高维

空间中将不再成立，从而会降低相关机器学习算法的性能。同时，数据的高维性增加了信息处理算法的计算复杂度，使得很多处理实时任务的算法效率降低，无法满足如在线分类、视觉跟踪等实时任务的要求。这些问题给高维数据的分析处理和模式识别带来了极大的挑战。因此，在处理高维数据时，降维显得尤为重要，这一任务吸引了众多科研人员的关注，成为机器学习与数据挖掘中的热门研究问题之一。

1.1.2 研究意义

数据降维（Dimensionality Reduction，或称为维数约简）是数据建模与数据挖掘的基本问题，流形学习（Manifold Learning）是数据降维中近年来所兴起的热点方法之一。降维的目的是将数据从高维观测空间通过线性或非线性方法投影到低维特征空间，从而发现隐藏在高维数据中的有意义的低维结构。

降维对信息处理有多方面的益处[110]：①获取本质特征。对数据进行降维以提取蕴含在高维数据中的本质特征，其实也就是从数据中获取有用的知识，这也是数据挖掘和机器学习的主要目标之一。比如，在分类问题中，并非每维特征都会对分类起作用，甚至由于维数过高，有些特征还起误导作用，而对数据进行降维可得到本质有用的特征，从而可以提高分类准确率。②克服维数灾难。维数灾难问题是对高维数据进行分析所遭遇的主要障碍，通过降维可将高维数据转化为本质低维表示，同时尽可能地保留处理任务所需的信息，所以可用维数非常低的低维表示代替原始高维数据，算法的输出精度（如分类准确率）等能够得到保证（甚至会提高），同时算法的效率还会大大提高。于是降维在相当大程度上可减轻甚至避免维数分析方法所面临的维数灾难问题。

③节省存储空间。通过降维获得了数据的本质低维表示，可节约存储数据所需的空间，且能进一步有效降低后续处理的计算代价。④去除无用噪声。高维数据中可能包含很多冗余的噪声信息，通过降维可删除无用或冗余的噪声维数，从而在一定程度上消除高维数据中存在的噪声。⑤实现数据可视化。通过寻找高维数据的2维或3维主特征表示，可将高维数据中蕴涵的规律转化为直观的视觉形式。即使数据的本质特征表示不仅仅只有2维或3维，也可通过观察数据的特征对两个或三个特征的组合进行可视化。

降维问题涉及模式识别、统计学、机器学习、数据挖掘等多个领域，是一个富有挑战性的课题。流形学习是解决此问题的有效途径之一，它自提出以来就吸引了信息科学领域研究人员的广泛关注并成为相关研究热点。在理论和应用上，流形学习都具有重要的研究意义。首先，它涉及数学、信息科学、生物认知、计算机科学等多个领域，是新兴的前沿性的交叉科学研究。其次，它也促进了数学中多个分支的交叉，如微分几何、图论、概率统计等，推动了新算法与新理论的产生。最后，流形学习方法在模式识别、机器学习与数据挖掘的许多高维实际问题中都有重要应用价值，如计算机视觉（人脸识别、行人检测）、自然语言处理（文档分类、机器翻译）、生物计算（蛋白质功能预测、基因筛选）等。因此，对流形学习进行相关研究具有重要的理论意义与应用价值。

1.2 相关研究基础

数据降维是人工智能、信息恢复和数据挖掘等领域最基本的问题之一。数据特征是指通过直接观测或间接计算而获得的用

以描述数据的可测特性的不变量，描述数据的特征个数（或自由度）称为数据的维数。现实中存在两种基本的数据维数：一种是用以表征（例如在计算机上表达）数据的数据特征个数，如用 64×64 维的数据表示一幅灰度图像、用关键词出现的频率表示一个文本等，这种维数我们称之为数据的表示维数；另一种是本质描述数据所需的最少特征数（或自由度），如一组 DNA 数据中决定人的某种特性的基因数目、一组图像特有特征的数目等，这种维数我们称之为数据的本质维数[21][205]。对于应用数据，如 DNA 序列、图像、同主题文本等，数据的表示维数与本质维数之间常存在极大差距。数据降维的根本任务是将数据从高维表示空间通过线性或非线性方法映射到低维本质特征空间，从而得到原高维数据的本质低维表示。

线性降维方法是传统的数据降维方法，也是发展得比较成熟的降维方法，这类方法通过线性映射将高维数据映射到低维空间从而得到低维特征表示。线性降维方法主要包括主成分分析（Principle Component Analysis，PCA）[115]、高维尺度分析（Multidimensional Scaling，MDS）[59]、独立成分分析（Independent Component Analysis，ICA）[108] 以及线性判别分析（Linear Discriminant Analysis，LDA）[167] 等。PCA 用方差刻画数据的信息量，并通过投影的方法，使所得低维数据尽可能保持原高维数据的方差。MDS 则基于数据点间的某种距离或相似性度量，通过使低维数据间的距离与原高维数据间的相似性相一致来实现降维。ICA 基于消除或减少各分量间的相关性，通过在低维特征空间中寻找最能使数据相互统计独立的方向进行降维。LDA 是一种有监督的学习算法，它基于 Fisher 准则，通过选择使类内散度最小而类间散度最大的变换矩阵来实现降维。线性降维方法通过特

征的线性组合来进行降维，具有实现简单、容易计算、解释性强等优点，对很多具备线性结构的数据取得了较好的降维效果。然而对于现实世界的真实数据而言，其具有线性特性往往只是人们所假设的一种理想情形。实际中的很多高维数据是高度非线性或强属性相关的，例如图像数据、视频数据、文本数据、金融数据及基因表达数据等。由于线性假设的不成立，使得传统的线性降维方法无法提取这些数据的内在本质特征。为了解决这一问题，非线性降维方法应运而生。

基于核的降维方法是早期的比较有代表性的非线性降维方法，该类方法运用核技巧(Kernel Trick)将原始数据映射到更高维的特征空间，并期望在原始空间中呈非线性结构的数据集可以在核空间中呈现线性结构，从而可以利用线性降维方法发现蕴含在数据中的低维结构。由于这类方法无需创建复杂的假设空间，通过定义核函数即可隐性地定义出特征空间，故而很多线性降维方法均有其对应的基于核的非线性降维方法。代表性的方法有：核主成分分析(Kernel Principle Component Analysis，KPCA)[183]、核判别分析(Kernel Discriminant Analysis，KDA)[11]以及核独立成分分析(Kernel Independent Component Analysis，KICA)[5]等。虽然基于核的方法在处理非线性数据时比线性方法具有一定的优势，但也存在一定的局限性。首先，由于在算法中引入了核技巧，如何选择合适的核函数并设置函数中的参数，就成为了一个难点。一个合适的核函数能够使数据在特征空间中近似线性可分或者线性可分，但并不存在某个核函数能够适用于所有的数据集。目前，在实际应用中，核函数的选择主要依靠经验或领域知识选取，缺乏一个有效的理论指导。其次，由于在算法中采用的核映射是隐性的，人们并不知道数据集在核特征空间中的具体表达，这使