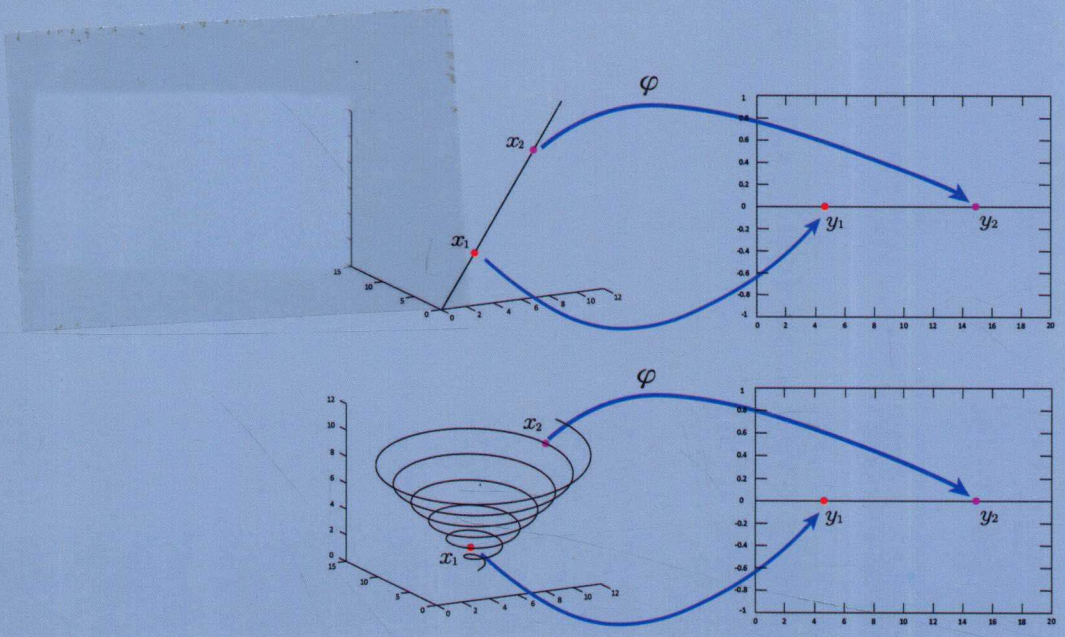

快速与增量式数据降维算法研究

申富饶 竺 涛 赵 健 著

科学出版社

快速与增量式数据降维算法研究

申富饶　竺　涛　赵　健　著

科学出版社

北京

内 容 简 介

本书围绕数据降维技术，分别针对线性降维和非线性降维两种降维手段进行广泛而深入的讨论。对于线性降维技术，本书介绍了常用的降维算法，并对线性降维与矩阵分解的等价性进行了分析，在此基础上提出了 semi-NMF、OCA、IOCA、EOCA 等改进算法并进行详细的理论分析和实验验证。对于非线性降维算法，本书介绍了常用算法，包括流形学习算法、基于核方法和基于神经网络的数据降维，并提出了改进的基准点选取 SL-Isomap 算法及基于拓扑学习的流形学习算法 TLE。

本书可供从事数据分析、模式识别和机器学习相关工作的技术人员以及高等院校计算机相关专业师生参考使用。

图书在版编目(CIP)数据

快速与增量式数据降维算法研究/申富饶，竺涛，赵健著. —北京：科学出版社，2018.11

ISBN 978-7-03-059237-8

Ⅰ. ①快… Ⅱ. ①申… ②竺… ③赵… Ⅲ. ①机器学习-算法-研究 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字 (2018) 第 251011 号

责任编辑：惠 雪 曾佳佳 / 责任校对：赵桂芬
责任印制：张克忠 / 封面设计：许 瑞

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

三河市荣展印务有限公司 印刷

科学出版社发行 各地新华书店经销
*
2018 年 11 月第 一 版 开本：720×1000 1/16
2018 年 11 月第一次印刷 印张：13 1/2
字数：234 000

定价：99.00 元

(如有印装质量问题，我社负责调换)

前　言

随着科技的发展，在互联网、金融、交通运输、生物信息、天文等几乎每个领域中，都在不断生成海量数据。在许多应用中，不但收集到的数据的数量在增加，这些数据的维数也在不断提高。高维数据不仅带来了机遇，更带来了巨大的挑战：一方面，对高维数据直接进行处理计算代价巨大；另一方面，“维数灾难” 会伴随着高维数据而发生，在数据的分析和应用中引起严重的问题。因此，如何从高维数据中高效地提取有用信息是一个值得研究的开放性问题。作为解决上述问题的重要工具，数据降维技术在过去数十年来吸引了研究者们大量的关注。

数据降维算法将数据从高维空间映射到低维空间并保留所感兴趣的信息。针对传统数据降维算法在快速学习与增量学习方面的不足，本书进行了一系列研究：从线性降维到流形学习，从快速降维到增量式降维，提出了数个新的算法。本书的主要贡献包括：①对线性降维进行研究，将线性降维问题转换为基底提取问题。提出了一种简单而高效的阈值机制以自动估计目标维数的方法来实现高速降维。②提出了一种基于数据选择的半非负矩阵分解 (semi-nonnegative matrix factorization，semi-NMF) 算法，它允许分解后系数矩阵中元素出现负值，利用自适应阈值机制，改善了矩阵分解的效率，实现了特征空间维数的自动估计并保证了所获取基底的质量。③ 提出了正交成分分析 (orthogonal component analysis，OCA) 这一快速线性降维算法，能够在不进行矩阵特征值求解运算和矩阵求逆运算情况下实现高效的正交成分 (基底) 提取，自动估计特征空间维数，并保证了数值上的稳定性。④提出了增量式正交成分分析 (incremental orthogonal component analysis，IOCA) 算法，其能够于在线环境中自动且快速地提取所需的正交基底，通过增加特征子空间的维数实现增量学习。⑤提出了进化式正交成分分析 (evolutionary orthogonal component analysis，EOCA) 算法，通过对正交基底的更新，方便地合并两个子空间为一个新的子空间，进而在低计算复杂度的情况下成功实现在线子空间学习。⑥提出了拓扑学习嵌入 (topology learning embedding，TLE) 算法以实现快速和增量式的非线性降维。TLE 通过在线输入数据流生成少量代表节点，构造了拓扑保持网络以实现对数据内在结构的近似，简化了所提取的结构，显著减少了非线性降维的计算量与

存储消耗，实现了增量式的流形结构学习与外样本数据嵌入。本书提出了一系列算法用于实现快速和增量式数据降维，并通过实验结果证明了这些算法的高效性。

本书是作者以及南京大学机器人智能与神经计算研究组 (RINC Group) 多年来在降维问题方面研究成果的总结，赵金熙教授为本书多个算法的形成提供了宝贵思想和理论指导，课题组成员徐烨、干强、梁雨等为本书做出了重要贡献，张天玥、徐百乐多次校对全书，科学出版社的惠雪编辑及其他同仁们也对本书的出版付出了大量心血，在此一并表示感谢。

本书适合人工智能相关领域的科技工作者阅读使用，也可以作为高等院校计算机科学各相关专业研究生的参考书。降维问题的研究仍然是当前的热点，各种方法都在不断发展中，且作者的学术水平有限，书中不足和疏漏之处在所难免，欢迎读者和同行专家批评指正。

作 者

2018 年 6 月于南京

目　　录

第1章 绪　论

随着科学技术特别是信息技术的发展，人类社会进入了大数据时代。在许多应用中，不但收集到的数据的数量在增加，这些数据的维数也在不断提高。如何处理大规模高维数据成为研究者与技术专家面临的一大挑战。一方面，对这些数据的分析和处理往往意味着要付出巨大的存储和计算代价；另一方面，高维数据常常会引发严重的“维数灾难”(curse of dimensionality，维数诅咒)。

“维数灾难”最早由 Bellman [1] 在研究动态优化问题时提出，即当问题中每个参数都可以通过搜索某个离散的搜索空间来确定时，随着参数的增加，通过枚举方式寻找最优解会越来越难。Bellman 指出如果分别对 10 维和 20 维空间中单位立方体在相邻采样点距离规定为 0.1 单位的条件下进行采样，后者所需的采样点数目是前者的 10^{10} 倍。这一现象说明：在缺乏简化假设的条件下，对某个含有多个参数的函数进行优化时，随着参数数量的增长，想要达到恒定的表现水准，需要的数据样本数量必须相应地呈指数级增长 [1]。在现在的大数据时代，“维数灾难”则用于描述在对高维空间中数据进行分析与组织时遇到的在低维空间中不会发生的各种问题 [2]。这些问题的发生是因为随着维数的增加，空间体积急速扩大而令可用数据变得稀疏 —— 这种现象被称为空空间现象 (empty space phenomenon) [3]。“维数灾难”的存在使得对高维数据的处理困难重重。

事实上，我们在描述某一事物时，往往只会选择性地对它的少量关键特征进行描述，而非事无巨细地将它的所有属性一一罗列，因为后者既不划算也不现实，更会把我们真正想要了解的关键信息淹没在大量的无用信息中。这反映了我们对这个世界的一种朴素认识：当利用事物完成某一任务时，并不需要了解该事物的所有属性，少量经过提取和选择的特征才是我们所关心的。

我们会尝试将含有大量属性的集合精练成仅含少量关键特征的较小集合，通过掌握这个较小特征集合以实现对事物的了解和掌握。在上述过程中，所精练出的特征越少越好，而这些特征对事物本质的反映越精确越好。这就提醒我们，当面对高维数据时，我们同样可以从中选择或提取少量的特征，它们能够反映事物的本质却不含冗余或仅含少量的冗余。这一想法便是对数据进行降维。数据降维技术能够显著减轻系统处理数据的花费，是分析高维数据的核心工具之一。

数据降维的基本问题是发现高维数据的紧凑表示形式 [4]。而“内在维数”假设是进行数据降维操作的基本假设。所谓内在维数，简单地说，可以理解为隐藏变量的个数，它的值经常小于被观察到的变量的数目。降维操作便是要从高维数

据中提取出这些隐藏变量。基于内在维数假设，在对高维数据进行降维时，我们相信这些数据中感兴趣的部分并没有填满整个原始高维空间，而我们的工作就是找出它们在所处的低维隐藏空间中的表示。具体地说，就是通过寻找一个映射函数 $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^k (k < d)$，将高维原始数据 $\boldsymbol{x}$ 映射到它的低维表示 $\boldsymbol{y} = \varphi(\boldsymbol{x})$。在这个过程中，$\varphi$ 可能是显式的，但也可能是未知的，即给定某个 $\boldsymbol{x}$，我们可能仅能得到对应的 $\boldsymbol{y}$，却无法知道具体的映射方式。

需要注意的是，数据降维只是手段，对数据的处理才是最终目的。如图 1 所示，数据降维可以看作整个数据处理流程中的预处理步骤：某个给定的数据处理系统无法有效地处理维数过高的数据，因此这些数据需要经过降维操作后才能输入系统。数据降维技术的目标是在减少冗余信息的同时保留或增强有意义的信息，以满足设备的处理要求或人的感知需求。作为一种重要的表示学习 [5] 技术，数据降维的作用主要包括：

(1) 减少数据的存储和运算代价；

(2) 去除数据中的无关信息，提升所提取特征的质量；

(3) 当降维后的维数小于等于 3 时，可实现数据的可视化。

数据降维技术广泛应用于科研、工程、商业等领域，在机器学习、数据挖掘、计算机视觉、信号处理等方面扮演着重要角色。

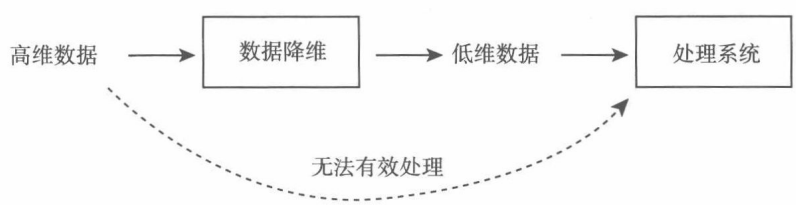

图 1 处理系统无法处理过高维数的数据时，需要进行数据降维

作为高维数据分析的重要手段，数据降维技术扎根于实践，同时在理论研究方面也具有重大意义。在 2000 年举办的 “21 世纪的数学挑战” 论坛上，著名统计学家 Donoho 在名为《高维数据分析：维数的诅咒与祝福》的演讲报告中指出 21 世纪必将是数据的世纪，同时预言在 21 世纪高维数据分析将成为非常活跃的研究领域，必将会出现全新的技术 [6]。而在 2008 年，图灵奖得主 Hopcroft 更是指出：高维数据降维理论是支撑未来计算机科学领域发展的主要理论之一 [7]。

1.1 数据降维算法分类

作为高维数据分析和处理的重要工具，数据降维技术向来受到研究者们的广泛关注。在过去的百余年来，各种降维算法层出不穷，构成了一个庞大的家族。

想要对已有的降维算法进行分类总结，能够采用的分类方式有多种多样 [8]。本书采用最常见的一种分类方式，即根据降维时所依据的模型，将数据降维算法分为线性降维算法和非线性降维算法两大类。

线性降维算法通过建立线性模型对数据进行降维，对于数据集 $\mathcal{X}$ 中的任意数据 $\boldsymbol{x}_i \in \mathbb{R}^d$，线性降维算法通过如下两种基础方式获取其低维表示 [9]：

(1) 寻找投影矩阵$\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k] \in \mathbb{R}^{d\times k}$，通过对 $\boldsymbol{x}_i$ 进行线性变换，获取其对应的低维表示 $\boldsymbol{y}_i \in \mathbb{R}^k$：

$$\boldsymbol{y}_i = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}_i \tag{1}$$

(2) 寻找一组基 $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$，利用这组基对 $\boldsymbol{x}_i$ 进行线性重构：

$$\boldsymbol{x}_i \approx \sum_{j=1}^{k} y_{i,j}\boldsymbol{v}_j \tag{2}$$

所获得的系数向量 $\boldsymbol{y}_i = [y_{i,1}, \cdots, y_{i,k}] \in \mathbb{R}^k$ 即可作为 $\boldsymbol{x}_i$ 的低维表示。

线性降维模型的优势在于简单直观，早期的数据降维算法大多采用线性模型，而后来的许多非线性数据降维算法也都借鉴了线性算法的思想。需要注意的是，式 (1) 与式 (2) 只是线性降维模型的基本框架。对于具体的线性降维算法，可以根据具体需要，在执行时对数据进行平移操作，或在该基本框架基础上添加正则化项。

如图 2 所示，线性降维算法实现的是对呈全局线性分布的数据的维数约简。但是，如果要实现对如图 3 中左侧类型的数据的降维 (这些数据呈非线性结构)，就需要用到非线性的降维算法。

所有未采用前述线性模型的降维算法都可被归于非线性降维算法。显然，与线性降维模型相比，非线性降维模型更加复杂多样，并且深受各种新技术发展的影响。非线性降维算法往往能够对结构更为复杂的数据进行处理，但另一方面，进行非线性降维运算的代价往往也更高。

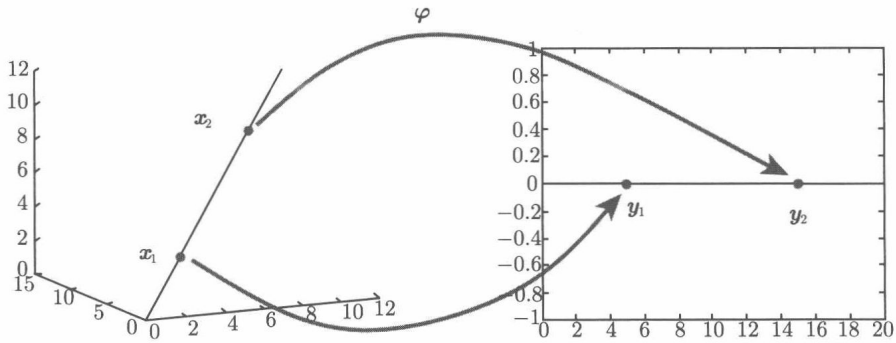

图 2 线性降维示意图

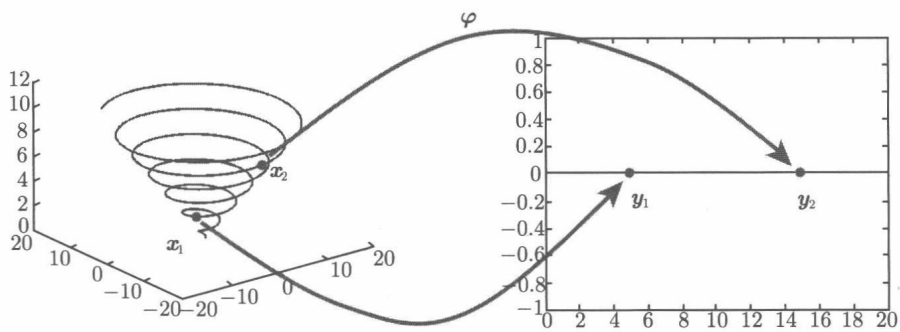

图 3 非线性降维示意图

数据降维中最核心的问题在于发现高维观测值中所隐含的有意义的潜在结构 [10]。无论线性或非线性降维，其本质都是为了解决上述问题 —— 线性降维算法直接通过提取一组基底对数据所处的特征空间进行表示；非线性降维算法则相对复杂，会以各种方式学习数据间的关系，进而确定它们的分布结构。数据降维领域存在很多具有挑战性的课题，而如何快速与增量式地对数据的结构进行学习便是一项重要的研究内容。

在评价一个算法时，低复杂度向来是一个重要的加分项。算法复杂度的高低往往关系到其能否成功地在大规模数据集上进行应用。然而，许多降维算法的空间与时间复杂度并不尽如人意。许多线性降维算法在计算过程中需要进行矩阵特征值运算或反复的迭代更新操作，如果以简单的运算代替或减少上述复杂运算，就能让算法更加简便快捷。另外，快速地对数据的内在维数 (基底维数) 进行估计是绝大多数线性降维算法所不具备却在实际应用中又十分重要的一项功能，实现此项功

能具有重要意义。对于许多非线性降维算法，当数据集规模很大时，巨大的存储和计算消耗成为限制它们应用的瓶颈，因此实现对复杂数据的快速结构学习在非线性降维研究中是一项非常值得进行的工作。

增量学习是指一个学习系统能不断地从新样本中学习新的知识，并能保存大部分以前已经学习到的知识。显然，以人类智能为代表的真正智能对信息的提取和知识的学习是增量式的，而近年来对增量学习的研究在人工智能领域成为一股重要的趋势。多数经典的数据降维算法都要求在进行降维运算前，所有的数据都已获得。然而在实际应用中，数据常常是不断生成的，因此实现对数据的增量式甚至在线式降维更能满足实际要求。

综上所述，在目前这个信息膨胀的“大数据”时代，快速与增量式降维算法在实际应用中具有明显的优势，本书针对这一课题进行了一系列研究。

1.2 本书主要内容与组织结构

本书主要展开关于快速与增量式数据降维算法的研究，按如下基本顺序进行：从快速数据降维算法研究到增量式数据降维算法研究，从线性降维算法研究到非线性降维 (流形学习) 算法研究。

在线性降维中，数据中感兴趣部分被假设为分布在一个低维的线性特征空间中，数据降维的任务可以通过求取特征空间基底而完成。本书关于线性降维研究的具体思路为设计一个自适应阈值机制对数据进行自动选择，随后根据所选择的数据确定特征空间维数并快速地提取特征空间基底。具体来说，在线性降维方面，本书的主要工作包含如下几点：

(1) 将线性降维任务转化为矩阵分解任务，即把原始数据矩阵近似表示为低秩的基底矩阵与系数矩阵相乘的形式，由此对基底的线性独立性等问题进行分析研究，提出一种自适应阈值机制用于进行基底的选择，并同时解决基底数目 (目标维数) 的自动确定问题。

(2) 受非负矩阵分解方法中“部分组成全部”的思想的启发，针对已有非负矩阵分解和半非负矩阵分解算法无法自动确定目标基底的问题以及计算消耗大的问题，提出一种基于数据选择的半非负矩阵分解算法，在取消系数的非负性限制的条件下，实现快速的半非负矩阵分解。

(3) 由于正交基底的性质更加优良，提出一种提取正交成分的线性降维算法 —— 正交成分分析 (orthogonal component analysis，OCA)。通过将施密特正交化与自适应阈值机制进行组合，OCA 具有低计算复杂度以及目标维数自动确定两大优势。

(4) 在 OCA 的基础上，提出了一种快速在线降维算法 —— 增量式正交成分分析 (incremental orthogonal component analysis，IOCA)。IOCA 通过对在线输入数据进行增量式的正交基底提取，实现了对特征子空间的动态更新。由于 IOCA 计算复杂度极低，且具有自动确定目标维数以及保证所提取基底的质量的能力，其成为一种有竞争力的在线降维算法。

(5) 提出一种新的子空间基底在线调整算法，实现对以正交基底表示的线性子空间进行合并的功能。在此基础上，提出名为进化式正交成分分析 (evolutionary orthogonal component analysis，EOCA) 的在线算法。

在非线性降维方面，针对目前已有非线性降维算法计算复杂度高、存储空间需求大的问题，主要进行了如下工作：

(1) 针对现有的基准点 Isomap 算法 (L-Isomap) 难以同时确定基准点数量与位置的问题，本书提出了 SoinnLandmark-Isomap (SL-Isomap) 算法，利用自组织神经网络 SOINN 学习高维数据的拓扑结构，自动确定基准点的数目与位置，同时实现了数据压缩与非线性降维。

(2) 针对 Isomap 的三大主要缺陷 (计算效率低、邻域难确定、无法在线降维)，本书提出了新的流形学习算法：拓扑学习与在线映射算法 (topology learning and out-of-sample embedding, TLOE)。TLOE 利用自组织神经网络学习高维数据的拓扑结构，自适应地确定基准点及其邻域关系，通过估算新样本点与基准点的相似度实现了在线降维映射。解决了流形学习的三大普遍问题：计算效率问题、监督学习问题以及外样本处理问题，并且实现了增量式流形学习。

(3) 在 TLOE 工作的基础上，提出了拓扑学习嵌入 (topology learning embedding，TLE) 算法，从拓扑学习与数据嵌入两方面分别进行了改进。

本书章节的设置如下：

第 1 章为绪论，介绍数据降维研究的意义，简要分析经典降维算法，随后对本书的主要内容及组织结构进行简要叙述。

第 2 章介绍部分现有的经典线性降维算法。

第 3 章中，通过对线性降维中基本问题的研究和分析，基于线性独立性理论，提出一种简单有效的自适应阈值机制，以用于随后的研究。

第 4 章中，从可解释性角度出发，对限定基底矩阵必须非负，系数矩阵可出现负值的半非负矩阵分解算法进行了研究，提出了一种较为快速的基于数据选择的半非负矩阵分解算法。

第 5 章中，提出了一种快速算法 OCA，在不进行矩阵求逆或特征值求解运算的情况下，以简单的运算实现正交成分 (基底) 的提取。

第 6 章中，针对在线学习问题，提出了 OCA 的增量式版本 IOCA，实现了数据的在线降维。

第 7 章中，提出了一种子空间标准正交基底在线调整算法，以此为基础，提出了在线算法 EOCA。

从第 8 章开始展开了关于流形学习的研究，本章对部分现有的经典非线性降维算法进行了介绍。

第 9 章中，提出了基于自组织神经网络的 SL-Isomap 算法，通过改进流形学习中基准点的选取，实现较为高效的非线性降维。

第 10 章中，提出了基于自组织神经网络的 TLOE 算法，以达到快速流形学习与外样本数据嵌入的目的。

第 11 章，继续以实现快速与增量式非线性流形学习为目标，提出了一种能够应用于大规模数据的非线性降维算法 TLE。

第 12 章为总结与展望，对本书所介绍的内容进行了概括性的描述，针对当前研究提出改进的方向，为进一步研究做出展望。