DOI:10.16353/j.cnki.1000-7490.2013.02.017

● 毕达天[1]，邱长波[1]，张　晗[2]
（1. 吉林大学　管理学院，吉林　长春　130022；2. 长春师范学院　经济管理学院，吉林　长春　130032）

数据降维技术研究现状及其进展*

摘　要： 数据挖掘主要用于从原始数据资料中挖掘有用的信息，而这些数据资料的维数已经对目前大多数数据挖掘算法的效率造成了严重的阻碍，这种阻碍被称之为“维数灾难”。数据降维技术可以有效地解决这一问题。文章以数据降维方法为主线，对数据降维问题的分类进行了描述，对数据降维方法的研究现状及主要算法进行了详细的阐述，对数据降维算法最新研究进展进行了简要介绍，并指出其优缺点，最后提出了数据降维技术今后的研究方向。

关键词： 数据挖掘；高维数据；降维；研究现状；研究进展

Abstract: Data mining is mainly used for the mining of useful information from raw data, however the dimensions of the raw data have become a serious obstacle to the efficiency of the most data mining algorithms. The obstacle is called as the “dimensional disaster”. The data dimension reduction technology can be used to effectively solve this problem. Taking the data dimension reduction method as the main clue, this article describes the classification of data dimension reduction, expatiates on the research status and main algorithms of the data dimension reduction method, gives a brief description of the latest research progress on data dimension reduction algorithms, and points out the advantages and disadvantages. Finally, the article presents the future research direction of data dimension reduction technology.

Keywords: data mining; high dimension data; dimension reduction; research status; research progress

数据挖掘技术自 1989 年 8 月的第一届 KDD（Knowledge Discovery in Database）国际学术会议被提出以来，经过 20 多年的发展，现如今已经在金融、电信、零售、医疗、信息处理等各个领域发挥着巨大的作用。目前，许多企业和政府部门等机构都积累了海量的、不同维度的数据资料，数据挖掘主要用于从原始数据资料中挖掘有用的信息，而这些数据资料的维数已经对目前大多数数据挖掘算法的效率造成了严重的阻碍，这种阻碍被称之为“维数灾难”[1]。也就是说，现有的数据挖掘算法对低维度的数据集非常有效，而对高维度的数据集则很难得出有意义的结果。

近几年，人们对数据挖掘技术的研究热情持续升温，在理论研究和技术应用等各个方面都取得了长足的进步，伴随着数据库技术和数据挖掘技术的不断发展，高维数据集的降维这一现实问题已经引起了各界学者越来越多的关注。

本文通过对近几年来数据降维技术成果的研究和归纳，从探求数据集降维技术的发展趋势的角度出发，对目前数据降维技术的研究进展进行了分析和总结。

1　数据降维问题的分类

数据降维的问题可以按照其原因分为四大类，分别为降低学习（建模）成本、提高学习（建模）性能、不相关维度约简和冗余维度约简，如图 1 所示。

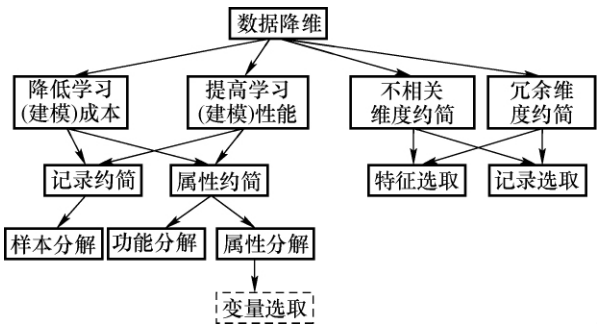

图 1　降维问题的分类

降低学习（建模）成本和提高学习（建模）性能可以进一步分为两个子问题：记录约简和属性约简，其中记录约简在更多的情况下被具体化为样本（或元组）分解；属性约简则包含功能分解和属性分解两个方面，这两个分解问题是分解方法论中的一部分。属性分解有一个子问题是变量选取，它的解决方法是在数据预处理过程中从给定

* 本文为吉林大学“985 工程”项目资助的研究成果。

的数据集中删除某属性，它的基本原理是减少数据挖掘算法运行所需时间，因为运行时间是由记录的数量和每个记录（维度）中属性的数量决定的，但是变量选取可能会降低数据挖掘的精确度。

不相关维度约简和冗余维度约简可以进一步分为两个子问题：特征选取和记录选取。特征选取的目的是确定哪些特征是重要的，哪些特征是不相干的或是冗余的并可以放弃的。特征选取的过程减少了数据集的维数，使数据挖掘算法更快和更有效的运行。在某些情况下，未来的分类精度可以改善，而在其他情况下，其结果将更简洁，更容易用模型来解释[2]。记录选取过程相对简单，但正如数据集中有一些属性比其他属性重要一样，正确的选取可以对后续的数据挖掘结果更有帮助[3]。

2 数据降维方法的分类及研究现状

目前，数据降维方法可以分为两大类，线性方法和非线性方法。

2.1 数据降维的线性方法

当数据集中各个变量间是独立无关的，或者数据为非线性时可在一定程度上用线性结构近似表达的时候，可以运用线性方法来对数据进行降维。

关于数据降维线性方法最初的研究是1958年Togerson提出了多维尺度分析（Classical Multidimensional Scaling，MDS）的方法[4]，多维尺度法是一种将多维空间的研究对象（样本或变量）简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。

主成分分析法[5]（Principal Component Analysis，PCA）与多维尺度法相类似，是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。

线性判别分析[6]（Linear Discriminant Analysis，LDA）与主成分分析法类似，是将高维的模式样本投影到最佳鉴别矢量空间，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，以达到抽取分类信息和压缩特征空间维数的效果。

独立成分分析[7]（Independent Component Analysis，ICA）通常被用来提取独立统计成分，同时最大化非高斯特性的测量，如峰度和偏度，或者将互信息减到最少。

随机投影[8]（Random Projection）是通过构建Lipschitz映射来实现降维，当数据维数和基数很大时，它不是引入了一个显著的畸变，而是随机地将原始的高位数据投射到一个低维子空间，由于投影消耗的是线性计算时间，因此这种方法的计算效率很高。

2.2 数据降维的非线性方法

当数据为高度非线性或强属性相关时，运用线性的方法对数据集进行降维处理的效果则不是很理想，因此这种情况下，需要用非线性的方法对数据集进行降维。目前数据降维的非线性方法有很多，人们也把更多的研究热情投入到其中。数据降维的非线性方法有以下几种。

基于核的主成分分析方法[9]（Kernel Principal Component Analysis），是目前国际上比较流行的一种特征提取方法，它是利用核技巧对主成分分析法进行的一种非线性推广。

2000年，Tenenbaum等人提出了等距离映射算法[10]（Isometrie Mapping，Isomap）。等距离映射算法用于计算一组高维数据点的准等距离低维嵌入，在粗略估计每个数据点的相邻流形的基础上，为计算一个数据流形的内在几何关系提供了一种简单方法。

2000年，Roweis等人提出了局部线性嵌入算法[11]（Local Linear Embedding，LLE），该方法是基于简单的几何直观：如果一个数据集是从光滑流形中取样、每个点相邻的点依然相邻并且同样坐落于低维空间。LLE算法是将数据集的每个点都线性嵌入到流形的一个局部线性贴片中来构造低维空间，例如保存原始数据的局部线性关系。

2003年，Donoho和Grimes提出了Hessian局部线性嵌入算法[12]（Hessian Locally Linear Embedding，HLLE），LLE通过最小化平方误差来实现线性嵌入，而HLLE则是在数据流行中最小化Hessian泛函来实现线性嵌入，HLLE的概念框架可以被看成是拉普拉斯特征映射框架的修改。

拉普拉斯特征映射法（Laplacian Eigenmaps，LE）是由Belkin和Niyogi于2003年提出的[13-14]，它是基于无监督学习流形的想法，首先在数据图形中构建Laplace-Beltrsmi算子，然后从与Laplace-Beltrsmi算子相一致的几个最小特征值的特征向量中得出降维后的低维数据集。

2004年，Lafon提出了扩散映射的方法[15]（Diffusion Maps），是将每个数据集嵌入欧几里得空间使得空间的欧式距离等于数据的扩散距离。由扩散映射引起的各种各样的扩散距离都可以描述为数据的多尺度几何结构。

3 数据降维技术最新研究进展

正如上文中所提到的21世纪初是数据降维技术最高速发展的阶段，随着其理论研究和实践应用的不断深入，关于数据降维技术的理论框架已逐步形成，许多已有的线性降维算法和非线性降维算法在其特定的领域都发挥着重要的作用。反观近些年对于数据降维技术的研究，则大多数集中在原有算法的改进和延伸。

3.1 自适应线性数据降维算法

K. Luebke和C. Weihs针对数据分类的问题，基于各种数据的类型和特点，提出了一种两步骤适应过程的线性

数据降维算法[16]，首先对不同的数据特征进行选择，然后再针对手中的数据进行应用研究。该算法所提到的自适应过程，是采用了“元学习”（Meta-learning）的方法[17]作为其试图覆盖各种数据空间的模拟数据构造规则，引入选择统计的方法对数据进行充足的描述[18]，以达到分类的目的，并极大地减少了分类的错误率。该算法的优点是对数据进行降维分类处理的过程是自适应的，并且适用于各种数据类型，缺点则在处理含有较多和较强异常值的数据集时的性能还需要进一步提高。

3.2 基于统计相关的数据降维方法

K. Lee，A. Gray 和 H. Kim 基于统计相关的方法，提出了一个在变量之间新的概念——“内在”距离。“内在”距离与其他扩散距离的概念相类似，可以作为变量间相关统计方法的拓展，应用在最邻近分类等不同的领域。

在此基础之上，他们又提出了基于“内在”距离这一新概念的降维方法——“相关映射”，并将其应用于真实的数据集中[19]。这种方法可以替代之前的方法，与之前的方法相比具有更强的稳定性，分类效果更好。这种方法的缺点是引入了一个新的参数 t，用来表示邻里演变的连接程度，t 的选取和理论解释还不是很明确，同时，这种方法还需在较大的数据集中进行验证。

3.3 基于基准点的数据降维方法

P. Magdalinos，C. Doulkeridis 和 M. Vazirgiannis 提出了一种快速有效的高维数据集降维算法[20]（FEDRA），这种算法属于基于基准点（Landmark-based）的降维算法的范畴。它的基本思路是通过保留数据对象两两之间的精确距离，并将其嵌入到低维的投影空间。通过理论分析和实验证明这种算法在处理高维和高基数的数据集时所得结果的质量要比同类算法更有效。在此基础之上，他们又将该算法应用到典型的聚类分析中，成功再生初始聚类结构并减少 10% 的初始维度，同时显著加快 k 均值聚类算法的收敛速率，在实验中所使用的数据集也比同类其他算法研究中所使用的数据集要更大。

FEDRA 算法的优点是满足了一个理想的降维算法的几点需求，即较低的时间和空间需求以及最小变形值，等等。在聚类分析应用中还可以改善原有的聚类结构和质量。这种算法目前仅仅在聚类分析应用中得到了验证，还需在其他算法应用以及处理更大的高维数据集中加以验证和改进。

3.4 几何局部嵌入的数据降维方法

Ge Shuzhi Sam，He Hongsheng 和 Shen Chengyao 提出了一种几何局部嵌入（Geometrically Local Embedding，GLE）的降维方法[21]，用于内部特征发现，分类和聚类等方面。

该方法提出了一种用几何距离来测量数据内部相邻节点的距离以达到高维数据的清晰可视化。相比以往的局部嵌入方法，这种几何距离强调了流形由中心向量所生成的局部几何结构，而不用再计算数据的成对距离。这种方法在提取数据集的内部结构方面非常有效，有助于进一步重建权重值以降低噪声数据和奇异值造成的影响，并且通过将特征数据投影到低维的可分离区域，在高维数据的可视化、分类和聚类应用中都取得了不错的结果。这种方法的优点是计算复杂度不高，可以应用在数据集很大的情况下；而缺点则是和其他降维算法相比，在输入数据较小的情况下计算速度较慢。

3.5 几种算法的比较

对上述 4 种算法从属性、应用数据类型、优点、缺点和用途 5 个方面进行横向比较，如表 1 所示。

表 1 几种算法比较

算法名称	属性	应用数据类型	优点	缺点	用途
自适应线性数据降维算法	线性	人工合成数据、宏观经济数据等	降维和分类过程自适应；适用于各种数据类型	处理含有较多较强的异常值的数据集时性能较差	分类
基于统计相关的数据降维方法	非线性	古典流形、手写体数字、挥发性有机化合物等	稳定性强；分类效果好；某些场合可以代替其他方法	引入新参数 t；较大数据集未得到验证	分类
基于基准点的数据降维方法	线性	雷达观测数据、图像分割数据、分子数据、人工合成数据等	较低时间和空间需求、改善原有聚类结构和质量	在聚类中得以验证，较大数据集性能未验证	聚类
几何局部嵌入的数据降维方法	非线性	人工流形数据、手写体数字等	计算复杂度低，可应用于大数据集	应用于小数据集时，计算速度慢	分类、聚类、数据可视化等

比较表 1 可知，4 种方法的应用数据类型都比较广泛，用途也都集中在数据挖掘技术应用最为广泛的两个方向：分类和聚类。算法的性能优缺点差异较大，人们可根据需要进行合理的选择，并针对某种算法的不足进行性能改进和提高。

4 结束语

本文以数据降维方法为主线，对数据降维问题的分类

进行了描述，对数据降维方法的研究历程及主要算法进行了详细的阐述，最后对数据降维算法最新研究进展进行了简要介绍，并指出了其优点和缺点。随着社会的发展，数据将以更多样的形式展现在大家面前，因此给人们处理这些数据所带来的挑战也越来越多。现有的数据降维算法的研究虽然已经取得了长足的进步，可是在以下两个方面依然存在着一些问题。

1）目前的降维方法对于普通高维数据（5 < 维度 < 50）已经非常有效，但是对于超高维数据（维度 > 50）的情况下，降维的效果还有待于进一步提高。

2）目前的线性降维算法受噪声影响较小，但是更多的情况下需要使用非线性降维方法，而大多数非线性降维方法的缺陷是受噪声影响大，因此提高算法的鲁棒性，减少噪声和奇异值对降维结果的影响是今后重点需要突破的研究方向。

可以看出，随着信息技术的不断发展，人类社会的方方面面几乎都要用数据“说话”。尽管数据降维原本是伴随着统计和数据挖掘中数据预处理这一步骤出现的，但由于数据降维技术本身独有的理念给人们处理解决各类问题都提供了一个新的思路和方法，在这一点上数据降维技术一定程度上等同于一种方法论，在未来的一段时期里必将对人类生产生活产生重大影响。□

参考文献

[1] ELDER J F，PREGIBON D. A Statistical perspective on knowledge discovery in databases [M] // FAYYAD U，PIATETSKY-SHAPIRO G，SMYTH P，et al. Advances in Knowledge Discovery and Data Mining. Cambridge，Massachusetts：AAAI/MIT Press，1996.

[2] HALL M A. Correlation-based feature selection for machine learning [D]. Hamiltion，New Zealand：University of Waikato，1999.

[3] BLUM P，LANGLEY P. Selection Of relevant features and examples in machine learning [J]. Artificial Intelligence，1997，97（1/2）：245-271.

[4] TOGERSON W S. Theory and methods of scaling [M]. [s. l.]：Wiley，1958.

[5] JOLLIFFE I T. Principal component analysis [M]. Berlin：Springer Series in Statistics. Springer-Verlag，1986.

[6] DUDA R O，HART P E，STORK D H. Pattern classification [M]. 2nd ed. [s. l.]：Wiley Interscience，2000.

[7] COMON P. Independent component analysis，a new concept? [J]. Signal Processing，1994，36（3）：287-314.

[8] WANG Jianzhong. Geometric structure of high-dimensional data and dimensionality reduction [M]. New York：Springer Heidelberg Dordrecht London，2011：131-147.

[9] SCHOLKOPF B，SMOLA A，MULLER K. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation，1998，10（5）：1299-1319.

[10] TENENBAUM J B，SILVA V，UNGFORD J C. A global geometrie framework for nonlinear dimensionality reduction [J]. Science，2000，290（12）：2319-2323.

[11] ROWEIS S T，SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science，2000，290（5500）：2323-2326.

[12] DONOHO D L，GRIMES C. Hessian eigenmaps：new locally linear embedding techniques for high-dimensional data [D]. Stanford：Department of Statistics，Stanford University，2003：5591-5596.

[13] BELKIN M. Problems of learning on manifolds [D]. Chicago：The University of Chicago，2003.

[14] BELKIN M，NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation，2003，15（6）：1373-1396.

[15] LAFON S. Diffusion maps and geometric harmonics [D]. U. S.：Yale University，2004.

[16] LUEBKE K，WEIHS C. Linear dimension reduction in classification：adaptive procedure for optimum results [J]. Advances in Data Analysis and Classification，2011，5（3）：201-213.

[17] BRAZDIL P，GIRAUD-CARRIER C，SOARES C，et al. Meta learning [M]. Berlin：Springer Heidelberg，2009.

[18] O'GORMAN T W. Applied adaptive statistical methods：tests of significance and confidence intervals [J]. Technometrics，2004，64（4）：484-485.

[19] LEE K，GRAY A，KIM H. Dependence maps，a dimensionality reduction with dependence distance for high-dimensional data [J]. Data Mining and Knowledge Discovery，2012（4）（online first）.

[20] MAGDALINOS P，DOULKERIDIS C，ATHENS M V. Enhancing clustering quality through landmark-based dimensionality reduction [J]. ACM Transactions on Knowledge Discovery from Data，2011，5（2）：1-44.

[21] GE Shuzhi Sam，HE Hongsheng，SHEN Chengyao. Geometrically local embedding in manifolds for dimension reduction [J]. Pattern Recognition，2012，45（4）：1455-1470.

作者简介： 毕达天，男，博士生。研究方向：管理信息系统，数据挖掘。

邱长波，男，教授，博士，博士生导师。研究方向：数据挖掘。

张晗，女。研究方向：管理学，数据挖掘。

收稿日期： 2012－09－03