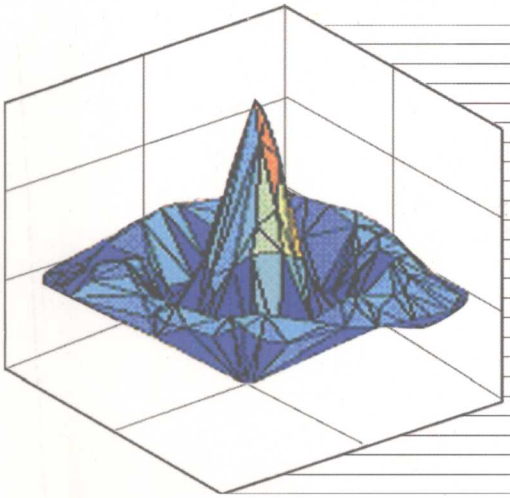


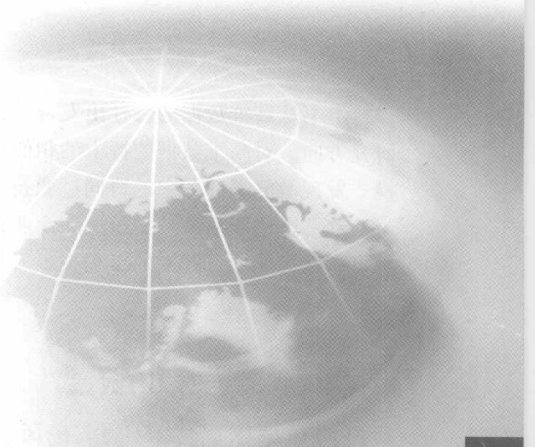
RESEARCH ON TECHNOLOGIES
FOR HIGH DIMENSIONAL DATA MINING

高维数据挖掘技术研究

杨风召 著

东南大学出版社

本书由南京财经大学出版基金资助出版



RESEARCH ON TECHNOLOGIES
FOR HIGH DIMENSIONAL
DATA MINING

高维数据挖掘技术研究

杨风召 著

東南大學出版社
·南京·

内容提要

本书从高维数据的特性出发，指出了高维数据给数据挖掘带来的影响以及高维数据挖掘的研究方向。对高维数据挖掘中的相似性搜索、高维数据聚类、高维数据异常检测、高维数据频繁模式发现及电子商务中的协同过滤技术进行了研究，提出了相关的解决方案和相应算法。

本书适用于从事数据挖掘和商业智能研究的高校教师、研究生、科研院所的科研人员以及从事商业智能项目开发的工程技术人员。

图书在版编目(CIP)数据

高维数据挖掘技术研究/杨风召著. —南京：东南大学出版社，2007.12

ISBN 978-7-5641-0977-6

Ⅰ.高… Ⅱ.杨… Ⅲ.数学采集-研究 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2007)第 165995 号

高维数据挖掘技术研究

出版发行	东南大学出版社
社　　址	南京四牌楼 2 号(邮编 210096)
出 版 人	江　汉
网　　址	http://press.seu.edu.cn
电子邮件	press@seu.edu.cn
经　　销	全国各地新华书店
印　　刷	江苏兴化印刷有限责任公司
开　　本	700 mm×1000 mm　1/16
印　　张	8.75
字　　数	158 千字
版　　次	2007 年 12 月第 1 版
印　　次	2007 年 12 月第 1 次印刷
书　　号	ISBN 978-7-5641-0977-6/TP·162
定　　价	24.00 元

前　言

数据挖掘指的是从大量的数据中提取隐含的、事先未知的、并且潜在有用的知识的技术，是目前国际上数据库和信息决策领域最前沿的研究方向之一。在数据挖掘中有两个基本问题需要解决，一个是如何能挖掘出数据中蕴藏的知识；另一个是面对海量的数据如何在有限的时间内完成。两者缺一不可。因此在数据挖掘中不只是包含统计学、人工智能的研究成果，也需要数据库系统提供有效的存储、索引和查询处理支持，源于高性能计算的技术在处理海量数据集方面常常也是重要的。因此在数据挖掘的研究中常常要涉及多种学科多个领域的知识。本书围绕高维数据挖掘这个主题，对高维数据挖掘中的相似性搜索、高维数据聚类、高维数据异常检测、高维数据频繁模式挖掘及电子商务中的协同过滤等相关技术进行了研究。

高维数据是指维数比较高的数据，如交易数据、文档词频数据、用户评分数据、Web 使用数据及多媒体数据等。由于这种数据存在的普遍性，使得对高维数据挖掘的研究有着非常重要的意义。但由于“维灾”的影响，也使得高维数据挖掘变得异常地困难，必须采用一些特殊的手段进行处理。

随着数据维数的升高，高维索引结构的性能迅速下降，在低维空间中，我们经常采用 L_k-范数作为数据之间的相似性度量，在高维空间中很多情况下这种相似性的概念不复存在，这就给高维数据挖掘带来了很严峻的考验：一方面引起基于索引结构的数据挖掘算法的性能下降；另一方面很多基于全空间距离函数的挖掘方法也会失效。解决的方法可以有以下几种：可以通过降维将数据从高维降到低维，然后用低维数据的处理办法进行处理；对算法效率下降问题可以通过设计更为有效的索引结构、采用增量算法及并行算法等来提高算法的性能；对失效的问题通过重新定义使其获得新生。

到底多少维的数据才算高维数据，是一个首先必须弄清楚的问题。是十维、一百维、一千维还是一万维？我们认为与所解决的问题有关，同时与维之间的相关程度也有关系，不能一概而论。维数的增加对数据挖掘算法的影响也是渐变的

过程，不存在一个明显的界限，最关键的是要看维数的增加能否对原来研究问题的方法造成根本性的影响，比如说由于维数高的原因使原来的方法失效，或者使原来的方法效率降低到难以容忍的程度等。

本书是作者近年来从事高维数据挖掘研究成果的总结，该研究得到江苏省教育自然科学基金(06KJB520031)和江苏政府留学奖学金的支持。

在该书编写的过程中，得到了复旦大学施伯乐教授、朱扬勇教授，澳大利亚悉尼科技大学的 Chengqi Zhang 教授、Longbing Cao 博士的指导和帮助，在此表示感谢。同时感谢数据挖掘讨论组(www. dmgroup. org. cn)的成员们，与他们就某些专业问题的讨论使我受益匪浅。

由于高维数据挖掘涉及的内容较多，该书不可能包含这个研究领域的全部内容，同时由于作者水平所限，不当与错误之处在所难免，敬请广大读者及相关专家批评指正。

杨风召

2007 年 9 月 10 日

目　录

1 绪 论

1.1 研究背景

1.1.1 数据挖掘技术的产生和发展

在过去的三十年，随着计算机硬件技术、数据收集技术和数据存储技术的快速发展，各行各业都逐步建立起各自的数据库体系。在这些数据库中存放着大量的数据，如何能有效地利用这些信息，使之能为生产实践所利用，成为人们所关注的问题。但相对于堆积成山的丰富的数据而言，人们缺乏强有力的分析手段和分析工具，因而造成了“数据丰富而信息缺乏”的状况。显然，数据库的检索和查询难以满足人们的需要，虽然伴随着数据仓库出现的联机分析处理(On-Line Analytical Processing，OLAP)技术具有总结、概化和聚集的功能，可以从不同角度来观察数据，支持多维分析和决策支持，但它不能进行更深层次的分析，挖掘出大量数据背后所蕴藏的知识。在这种情况下，数据挖掘技术便应运而生。

数据挖掘指的是从大量的数据中提取人们感兴趣的知识，这些知识是隐含的、事先未知的、并且是潜在有用的信息[FPSU96]。它是计算机技术研究中的一个很有应用价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术，目前已成为国际上数据库和信息决策领域中最前沿的研究方向之一，引起了学术界和工业界的广泛关注。一些国际上高级别的工业研究实验室，例如 IBM Almaden 和 GTE，众多的学术单位，例如 UC Berkeley，都在这个领域开展了各种各样的研究计划。其研究的主要目标是发展有关的方法论、理论和工具，以支持从大量数据中提取有用的和让人感兴趣的知识和模式。

数据挖掘，也叫数据库中发现知识(Knowledge Discovery in Databases，KDD)。KDD 一词首次出现在 1989 年 8 月举行的第 11 届国际联合人工智能学术会议上。随着 KDD 在学术界和工业界的影响越来越大，国际 KDD 组委会于 1995 年把专题讨论会更名为国际会议，在加拿大蒙特利尔市召开了第 1 届 KDD 国际学术会议，以后每年召开一次。迄今为止，由美国人工智能协会主办的 KDD 国际研讨会已经召开了 13 次，规模由原来的专题讨论会发展成为国际学术大会。

除了美国人工智能协会主办的KDD年会外，还有许多的数据挖掘年会，包括PAKDD、PKDD、SIAM-Data Mining等。PAKDD（Pacific-Asia Conference on Knowledge Discovery and Data Mining）是亚洲-太平洋地区数据挖掘会议，从1997年到现在已经召开了11届。PKDD（European Symposium on Principles of Data Mining and Knowledge Discovery）是欧洲数据挖掘讨论会，也是从1997年开始的，到目前也已经召开了11届。SIAM-Data Mining（Society for Industrial and Applied Mathematics）是SIAM组织召开的数据挖掘讨论会，于2001年4月召开了第1届讨论会，专注于科学数据的数据挖掘。此外，数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了KDD专题或专刊。

按照挖掘结果的模式，数据挖掘任务可以分为两大类：描述性和预测性。描述性数据挖掘是对数据的一般特性进行描述。预测性数据挖掘是通过对现有数据进行分析推理，对未来的行为进行预测。具体来说可以分为概念描述，关联规则分析，分类、聚类分析、异常分析和演化分析等。

1.1.2 高维数据挖掘的概念

数据挖掘技术的应用范围非常广泛，包括政治、经济、资源、环境、安全、科学、管理等诸多方面。在这些应用中，我们经常会碰到一些对象，它们可能有几十、几百甚至成千上万个属性。可以将这些对象表示成高维属性空间中的点或向量，这样就把客观世界中的对象集用高维数据的集合来表示。高维数据挖掘就是指对这种数据进行挖掘，从中发现潜在有用的知识。下面是一些常见的高维数据的例子。

1）*购物篮数据*

在零售商业中，某客户的一次购物行为可以看作是一个交易。该客户所购商品的集合构成了一个交易记录，大家形象地称之为“购物篮”。交易数据库中记录着许多客户的交易记录。我们可以将所有的商品种类看作一个表的列，而客户的一次购物看作表的行，如果在该次购物中有某种商品，就在对应的列上记录为“1”（表示有该商品）或一个其他有意义的数值（如商品的件数或价值等），否则就计为“0”或计为空。这样就可以将购物篮数据看作以商品种类为维，以交易为记录的多维数据。由于商品的种类一般都非常多（几千至几万种），因此购物篮数据实际上是一种特高维的数据。

2）*文档数据*

在信息检索（IR）领域，经常用向量空间模型来表示文档。在这个模型中，每个文档被看作是在词空间（term-space）中的一个向量。最简单的形式，即每个文

档用词频 TF(term-frequency)向量 $\boldsymbol{d}_{tf}=(tf_1,tf_2,\cdots,tf_n)$来表示,这里 ft_i 是指在这篇文档中第 i 个词出现的频率。由于通常词空间是一个高维的空间,因此在空间向量模型中,文档数据是一种高维的数据。

3) 用户评分数据

由于电子商务的兴起和商业竞争的需要,许多企业和组织开始为客户提供个性化的服务,如根据每个用户的特点推荐用户可能需要的产品和服务。目前的推荐系统主要依据客户的消费行为(购物篮等)或用户的评分数据。用户评分是指用户根据自己的感受对使用过的商品或享用过的服务打分,通过这种数据可以使我们了解用户的消费习惯,划分出不同的用户群体,根据相似的群体行为预测某个个体的行为等等。同购物篮数据相似,用户评分数据一般也是高维数据。

4) 多媒体数据

多媒体数据库在许多领域,如地理学、分子生物学、CAD 等学科得到了越来越重要的应用。在多媒体数据挖掘中,基于内容的相似性搜索是一个最基本的操作。解决这一问题最有效的方法是通过特征抽取将多媒体对象映射为高维特征空间中的点(特征向量)。这样就可以将多媒体对象的相似性搜索问题转换成为高维空间中的最近邻查找问题。

5) 时间序列数据

时间序列数据是一种特殊的序列数据,时间序列数据由一系列随时间变化的值组成,测量这些值的时间间隔是等间距的。时间序列数据库在证券期货、科学实验和药物研究等方面都有重要的应用。如一个事件在一个等间隔的时间序列 $t_1,t_2,\cdots,t_n$ 上的值分别为 $x_1,x_2,\cdots,x_n$,那么我们可以把这个事件表示为 $X=x_1x_2\cdots x_n$,这样就将时间序列数据看作一个 n 维的向量,由于在实际中时间序列数据的长度很长,因此时间序列数据是一种高维数据。

6) Web 使用数据

在基于 Web 的应用中,Web 日志记录用户对 Web 页的使用模式,通过分析和探索 Web 日志中的用户规律,可以为电子商务识别潜在客户,提高终端用户的服务质量以及改善 Web 服务器的性能。用户使用 Web 的模式也可以表示成高维数据的形式,Web 服务器中的每一个 Web 页都可以看作高维空间中的一维,在每一维上的数值可以表示多种意义,如用户是否访问该网页或在该网页停留的时间等。

7) 基因表达数据

基因表达数据是由基因芯片实验产生的。如酵母基因芯片实验可产生6 223种基因在 79 种条件下的表达数据。这种数据可以表示成一个矩阵,每个行对应

一个基因，每一列表示一个条件。每个数值表示一个基因在特定条件下的 mRNA 的相对丰度。生物学家意在通过这种数据找到在特定的条件组合下表现出强烈相似基因的集合。

由于在现实世界中存在着大量的高维数据，而这些高维数据与低维数据相比，在许多方面又表现出不同的特征，如果将用于低维数据的挖掘方法直接应用于高维数据，则可能会产生完全不同的结果，因此必须研究适合高维数据挖掘的理论和方法，这对于完善数据挖掘理论以及拓展数据挖掘的应用都有重要的意义。本书主要针对高维数据挖掘中的一些关键问题，如相似性搜索、聚类和异常检测、频繁封闭模式挖掘等进行研究，讨论高维对它们的影响及所采取的手段和方法。

需要说明的是，高维数的数据并没有统一的标准，其与所解决的问题有关，同时与维之间的相关程度也有关系，不能一概而论。比如在研究对象的相似性搜索问题时，如果数据的独立维数超过 10 维就属于维数非常高的数据了，但在频繁模式挖掘问题中数据的维数为几百维都是很正常的。我们认为最关键的是要看维数的增加能否对原来研究问题的方法造成根本性的影响，比如说由于维数高的原因使原来的方法失效，或者使原来的方法效率降低到“难以容忍”的程度等。

1.2 高维数据挖掘所遇到的困难

1.2.1 高维数据的特点

1）稀疏性

假设一个 d 维的数据集 D 存在于一个超级立方体单元 $\Omega=[0,1]^d$ 中，数据在空间中均匀分布，并且各个维之间是独立的。对于一个边长为 s 的超级立方体范围来说，一个点在这个范围内的概率为 s^d，由于 $s<1$，因此随着维数 d 的增大，这个概率的值会越来越小。也就是说，即使在一个很大的范围内很可能连一个点也不包含。例如，当 $d=100$ 时，一个边长为 0.95 的超级立方体范围只包含 0.59% 的数据点。由于这个超级立方体范围可以位于数据空间 Ω 的任何地方，因此我们可以得出结论，在高维空间中，数据点是非常稀疏的。

我们还可以通过基于网格的直方图来反映高维空间中数据点的分布特点。设 $d=50$，数据集中的数据点的数目 $N=10^{12}$。假如我们对每个维只从中点划分成两个部分，这样原来的空间 Ω 就被划分成 2^{50} 个单元，这时会得到 10^{12} 个包含数据的单元，其中每个单元包含一个数据，包含数据的单元数仅占总单元数的大约

1/1 000。实际上这种划分还是非常粗略的，如果每个维划分更多的段数，这时空的单元数还会增加很多。

2）空空间现象

下面我们来分析一下正态分布的数据，一个正态分布可以用中心点（期望值）和标准差（σ）来表示。数据点与期望值点之间的距离符合高斯分布，但与期望点的相对方位是随机选取的。应该注意的是，相对于一个点的可能方向的数目，也是随着维数的增大而呈指数级的增长，其结果是，虽然与中心点之间的距离仍然符合同样的分布，但数据点之间的距离也会随着维数的增大而增加。如果我们考虑数据集的密度函数，就会发现，虽然可能没有一个点离中心点的距离很近，但在中心点还是会出现一个最大值。这种高维空间中在空区域中点的密度可能会很高的现象称为“空空间现象”（empty space phenomenon）[Sco92]。

1.2.2 维灾（the curse of dimensionality）

Bellman 第一次提出了“维灾”①这一术语[Bel61]。它最初的含义是不可能在一个离散的多维网格上用蛮力搜索去优化一个有着很多变量的函数。这是因为网格的数目会随着维数也就是变量的数目呈指数级的增长。随着时间的推移，“维灾”这一术语也用来泛指在数据分析中遇到的由于变量（属性）过多而引起的所有问题。这些问题在信息搜索领域主要表现在两个方面：一方面，随着维数的升高，索引结构的修剪效率迅速下降，当维数增加到一定数量时，采用索引结构还不如顺序扫描[WSB98]；另一方面，在高维空间中由于查询点到它的最近邻和最远邻在很多情况下几乎是等距离的，最近邻的概念常常会失去意义[BGRS99]。

1.2.3 高维对数据挖掘的影响

1）高维对最近邻查询的影响

数据挖掘面对的是海量的数据，为了提高最近邻查询的效率，往往需要索引结构的支持，在进行高维数据的最近邻查询时，由于“维灾”会使索引结构失效或使其性能下降，从而使得算法的时间复杂度增加，导致查询效率的下降。到目前为止，对高维最近邻查询的研究主要集中在一般数值型数据的选维、降维和高维索引结构等方面，高维数据间的相似性度量还主要采用 L_k-范数。这些技术目前都还存在很多局限性，选维和降维技术只有在数据集的维之间存在有较强的相关

① “the curse of dimensionality”的中文翻译，curse 在词典中是咒骂、祸因的意思，在其他文献中也有不同的译法，这里译为“维灾”，意为由于维数过多而引起的各种问题。

关系时，效果才较好，而许多数据集是内在高维的，即通过选维或降维后剩下的维数仍然很高。所有的高维索引结构都有一个性能上限，超过这个界限后，就会失去对数据的修剪作用。我们在实际运用中遇到的高维数据的维数往往超过高维索引结构的性能上限，并且在高维数据中往往有许多空值的存在，如带数量的交易数据、文档数据、用户评分数据等。这种数据既不是二值型数据，也不是一般的数值型数据。显然前面谈到的索引结构都不适合这种数据，同时由于有大量空值的存在，这种数据的相似性度量函数也不适合采用 L_k-范数。更进一步，“维灾”还可能会使高维空间中最近邻概念失去意义。

2）高维对聚类和异常检测的影响

目前在数据挖掘中聚类和异常的概念大多是基于距离或密度的，快速的聚类或异常检测算法往往依赖于索引结构或网格划分。高维对聚类和异常检测的影响也主要表现在两个方面：一方面由于在高维空间中索引结构的失效或网格划分的数目随维数呈指数级增长，使得聚类和异常检测算法的性能下降；另一方面由于在高维空间中的很多情况下，数据点之间几乎是等距离的，使得聚类的概念失去意义，同样由于高维空间中数据的高度稀疏性，每个数据点在距离或密度的意义上都可以看作是一个异常点，这时异常的概念也会变得毫无意义。

3）高维对频繁模式挖掘的影响

在数据挖掘中，最初绝大多数的频繁模式挖掘算法都是基于特征（列或项）计数的，他们将特征（列或项）的组合作为算法的搜索空间。由于特征组合的数目与特征数呈指数关系，当特征数（数据维数）较高时，算法的搜索空间会“爆炸性”的增长，算法的效率会大幅度的下降或者根本得不到结果，所以这些算法通常不适用于维数非常高的数据。

4）高维对分类模式挖掘的影响

传统的分类方法有决策树、贝叶斯法、神经网络等，不同的分类方法有不同的特点。有三种评价尺度：(1)预测准确度；(2)计算复杂度；(3)模型描述的简洁度。分类的效果一般和数据对象的特点有关。有的数据噪声大，有的数据分布稀疏，有的数据属性间相关性强，有的数据属性是离散的而有的数据是连续值，目前普遍认为不存在某种方法能适合于各种特点的数据。一般情况下，传统的分类方法如决策树方法，对低维数据对象即具有较少特征属性分类时可以取得较高的预测精度，分类模型也较为简洁。但对于高维数据对象，传统的分类方法将产生较复杂的分类模型，并会出现分类模型过度拟合数据集的情况，而且由于决策树方法是一个属性一个属性地考虑，所以算法效率难以提高。近年来，为解决高维数据对象的分类问题，学者们提出了一些新的方法，如 SVM(Supporting Vector Ma-

chine,支持向量机)方法和基于规则的分类方法。该部分不是本书的主要研究内容,感兴趣的读者可以参阅有关文献。

1.3 高维数据挖掘的主要研究方向

针对上述高维对数据挖掘的影响,在高维数据挖掘领域主要有以下几个研究方向。

1.3.1 高维空间中的距离函数或相似性度量函数

距离函数和相似性度量函数在很多数据挖掘算法中扮演着非常重要的角色。它常常用来衡量对象之间的差异程度和相似程度。由于"维灾"与传统上采用 L_k-范数作为距离函数有关,因此通过重新定义合适的距离函数或相似性度量函数可以避开"维灾"的影响。

1.3.2 高效的高维数据相似性搜索算法

目前绝大多数的高维索引结构和相似性搜索算法都是基于数值型数据,并且这些索引结构在应用于数据挖掘时都不同程度存在着一定的局限。因此需要设计更为高效的相似性搜索算法,包括两部分内容:一部分是对现未涉及或研究较少的其他类型高维数据相似性搜索方法的研究,另一部分是对现有高维索引结构或搜索算法性能的改进。

1.3.3 高效的高维数据挖掘算法

针对在高维空间中多数数据挖掘算法效率下降的问题,需要设计更为高效的高维数据挖掘算法。如在高维索引结构失效的情况下,在聚类算法或异常检测算法中采用并行算法、增量算法以及采样技术等提高算法的效率。根据高维数据的特点,设计新颖的频繁模式挖掘算法,提高算法的执行效率。

1.3.4 在高维空间中对失效的问题的处理

如前所述,在高维情况下,最近邻的概念失去了意义,从而也会导致基于距离的聚类问题和异常检测问题失去意义。这些问题在高维情况下需要重新进行定义,并设计出相应的挖掘算法。

1.3.5 选维和降维

通过选维和降维,可以将高维数据转换为低维数据,然后采用低维数据的方