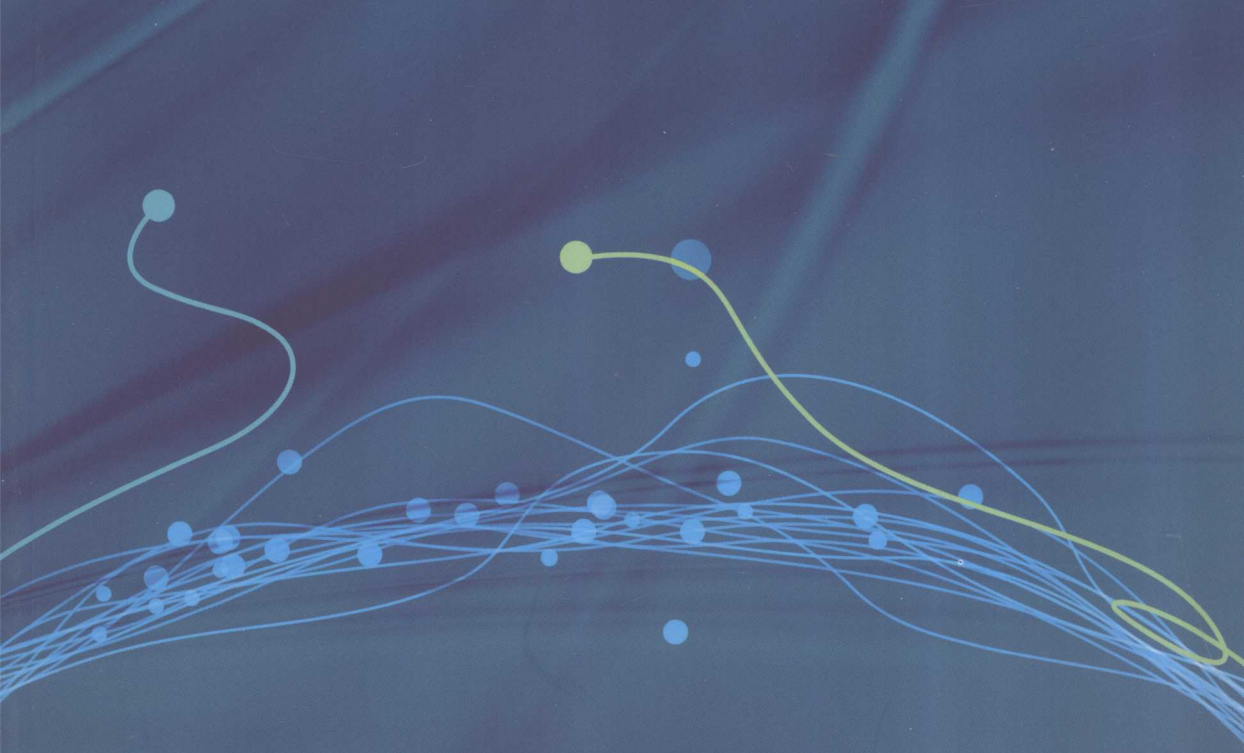

数据挖掘中的特征约简

陈黎飞　吴　涛　著

科学出版社

数据挖掘中的特征约简

陈黎飞　吴　涛　著

科学出版社

北　京

内 容 简 介

特征约简是数据挖掘的一项基础性技术，其目的在于降低数据的维度和提取数据中的重要特征或特征组合。本书系统地阐述了特征变换、特征选择的基本原理、基本过程，介绍了针对连续型、类属型等不同类型数据的过滤型、封装型及嵌入型特征约简方法。着重讨论了近年兴起的软特征选择技术，以及嵌入自动特征约简的子空间聚类、子空间分类技术，并以实例的方式给出了不同方法在文档挖掘、信息安全以及生物信息学等领域的应用。

本书可以作为数据挖掘、机器学习、模式识别理论与技术的教学、实践和应用的教科书或参考书，适合高等院校高年级本科生、研究生以及学习数据挖掘课程的学生使用，也适合相关企事业的技术人员使用。

图书在版编目(CIP)数据

数据挖掘中的特征约简/陈黎飞，吴涛著. —北京：科学出版社，2016

ISBN 978-7-03-049657-7

Ⅰ. ①数… Ⅱ. ①陈… ②吴… Ⅲ. ①数据处理－研究 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字(2016)第 201396 号

责任编辑：王 哲 邢宝钦 / 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：迷底书装

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

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2016 年 8 月第 一 版 开本：720×1 000 1/16

2016 年 8 月第一次印刷 印张：13 1/2

字数：250 000

定价：81.00 元

（如有印装质量问题，我社负责调换）

前　言

随着“大数据”的兴起，数据挖掘研究和应用已深入人心。通过数据挖掘，人们可以在大量数据中提取出概念、规则、变化模式或规律等感兴趣的知识。近年来，信息化技术快速发展，数据挖掘需要处理和分析的数据益显复杂，表现为各式应用中描述事物属性的特征越来越繁杂，特征量也越来越庞大。作为数据归约的一项主要技术，特征约简在数据挖掘任务中的重要性也随之凸显，现已成为许多实用数据挖掘系统的一个基础构件。

特征约简也称为属性约简，在结构化数据的数据挖掘等领域还称为维度约简等，其目的是通过移除冗余特征或对原始特征的重新表示，以减少描述事物的特征数目，提取描述事物最合适的特征空间，降低各种数据挖掘方法计算代价的同时，提升数据质量，进而提高数据挖掘系统的性能。实际上，人们对特征约简的研究要早于数据挖掘。例如，早在 1970 年 10 月，美国阿贡国家实验室举行的专题研讨会（Symposium on Feature Extraction and Selection in Pattern Recognition）就已经关注图像处理、语音处理、光学字符识别及高能物理等模式识别领域中特征抽取和特征选择技术。

在早期的数据挖掘系统中，特征约简主要以一种数据预处理技术出现，它通过特征变换或特征选择将数据归约到低维空间，使得各种数据挖掘算法可以在归约后的特征空间完成挖掘任务，是应对高维数据挖掘等难题的一种有效手段。如今，特征约简已经渗透到数据挖掘过程的其他环节，事实上，一些新型数据挖掘模型和算法已将特征约简作为一个内在的组成部分；在机器学习领域，近年风生水起的表征学习（representation learning）的主要研究内容即为如何为分类等预测性挖掘任务抽取有用信息以有效表达数据。本书重点考察应用于数据挖掘系统的特征约简技术，在讨论理论知识的基础上，介绍过滤型、封装型及嵌入型特征约简的主流方法及其应用，以及聚类、分类挖掘领域近年开发的嵌入型软特征约简技术。

全书由 3 部分构成，共 6 章：第 1 部分包括第 1 章和第 2 章，分别介绍数据挖掘、特征约简的基础知识和主要特征约简技术；第 2 部分包括第 3 章和第 4 章，分别介绍主要的特征变换方法和特征选择方法；第 3 部分包括第 5 章和第 6 章，结合作者近年在相关领域的研究成果，分别介绍基于聚类和分类挖掘的嵌入型特征约简方法及其应用。除第 1 章和第 2 章外，后续各章具有相对独立性，方便读者根据自己的需要和时间、精力的不同情况选择使用。

福建师范大学数学与计算机科学学院人工智能实验室的研究生范宇杰编写了本书部分章节，并为内容整理等做了大量工作；杨天鹏、田健、兰天、林品乐、乔小双和李海超等同学参与了部分章节的材料整理工作；张健飞等同学提供了部分研究材料。

符 号 定 义

$\mathrm{Tr}=\{(\boldsymbol{x}_1,z_1),(\boldsymbol{x}_2,z_2),\cdots,(\boldsymbol{x}_i,z_i),\cdots,(\boldsymbol{x}_N,z_N)\}$	训练数据集
$\mathrm{DB}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_N\}$	无类别标号数据集
N	样本数目
$\boldsymbol{x}_i=(x_{i1},\cdots,x_{ij},\cdots,x_{iD})^{\mathrm{T}}$	第 $i(i=1,2,\cdots,N)$ 个样本（列向量）
D	样本的属性/特征数目
D'	约简属性/特征数目
$\boldsymbol{x},\boldsymbol{y}$	数据集中的任意样本
$\boldsymbol{x}',\boldsymbol{y}'$	样本在约简特征空间的投影
z	类别标号或预测属性
A_j	第 j 个属性
$\mathcal{A}$	属性的集合
x_j 或 x_d	样本 $\boldsymbol{x}$ 的第 j 或第 d 维属性 $(j,d=1,2,\cdots,D)$
O_d	第 d 个离散型属性取值的集合
$\lvert O_d\rvert$	O_d 包含的符号数目
$o\in O_d$	O_d 中的任一符号（离散值）
K	类（簇）数目
c_k	第 k 个类 $(k=1,2,\cdots,K)$
$\lvert c_k\rvert$	c_k 包含的样本数目
$\boldsymbol{v}_k=(v_{k1},\cdots,v_{kj},\cdots,v_{kD})^{\mathrm{T}}$	第 k 个类的中心向量
$\boldsymbol{w}_k=(w_{k1},\cdots,w_{kj},\cdots,w_{kD})^{\mathrm{T}}$	类 k 的特征权重向量
u_{ki}	样本 $\boldsymbol{x}_i$ 相对于 c_k 的隶属度
X,Y	随机变量（自变量）
Z	随机变量（因变量）
$\boldsymbol{A},\boldsymbol{B},\boldsymbol{E},\boldsymbol{H},\boldsymbol{S},\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$	矩阵
κ	核函数（数值型数据）
ℓ	核函数（类属型数据）

目　录

第1章 概 论

1.1 数据挖掘基础

Internet 的出现和通信技术的迅猛发展使得各种信息以指数级增长，受益于数据库等技术的进步，收集和保存各种类型的数据变得越来越容易。面对海量的数据，人们已不再满足于简单数据查询或统计，更为需要的是从大量数据资源中挖掘出对各类决策有指导意义的概念、规则、模式或规律等知识。面对数据极其丰富而信息或知识相对贫乏的现象，数据挖掘（Data Mining，DM）这一现代数据分析和处理技术应运而生，以从大量数据中提取隐含的、事先未知的，并且潜在有用的知识，已引起学术界和工业界的广泛关注[1-4]，也是当前数据库和信息决策领域的一个前沿研究方向。

一般认为，数据挖掘的概念最早是由美国计算机协会（Association for Computing Machinery，ACM）在 1995 年年会上提出。此前，1989 年第 11 届国际联合人工智能学术会议提出了一个更为广泛的概念——数据库的知识发现（Knowledge Discovery in Database，KDD）[1-3]。KDD 过程包括目标数据选择、数据预处理、数据转换，进行数据挖掘和解释并评价发现的知识等阶段。可以认为，数据挖掘是数据库中知识发现过程的一个基本的、最重要的步骤，因为它是从存放在数据库、数据仓库或其他数据源中的大量数据中挖掘隐藏的模式，包括数学方程、规则、聚类、图、树结构以及用时间序列表示的循环模式等。历史上，数据挖掘也被称为“数据打捞”、“数据探查”和“数据垂钓”等。

一个典型的数据挖掘系统结构如图 1.1 所示。其中的数据挖掘引擎根据指定的挖掘任务结合模式评估模块从数据中挖掘有趣的模式，图形用户界面则提供用户和数据挖掘系统间的通信与交互；知识库中存放领域知识，用于指导数据挖掘引擎或评估结果模式的有趣度等。数据挖掘引擎是系统的基本组成部分，按照挖掘结果模式的不同，可完成描述性和预测性两大数据挖掘任务。描述性数据挖掘提供数据内在特性的描述，预测性挖掘通过历史数据的分析和推理，对未来的行为进行预测。具体来说，数据挖掘任务可分为概念描述、关联分析、聚类分析、分类、异常分析和演化分析等。

数据挖掘是一门有较强应用背景的学科，从某种意义上说，正是实际应用的需要推动了数据挖掘的产生和发展。商业智能（Business Intellegence，BI）就是其中一个典型的应用。商业智能由 Gartner Group 于 1996 年提出，定义为一类由数据仓库或数

据集市、报表查询、数据分析、数据挖掘等部分组成的，以帮助企业决策为目的的技术及其应用。数据挖掘是实现商业智能的一项关键技术，它把先进的信息技术应用到整个企业，不仅为企业提供信息获取能力，而且通过对信息的开发将其转变为企业的竞争优势。

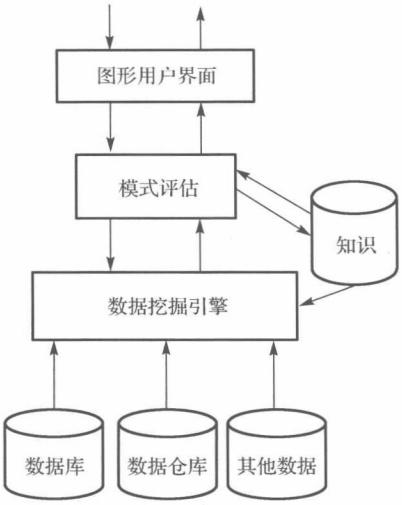

图 1.1　数据挖掘系统结构

1.2　数据挖掘模型

数据挖掘涉及许多学科，涵盖统计学、机器学习、人工智能、数据库技术、数据可视化技术和高性能计算等，是数据库系统、数据库应用以及高级数据处理和分析领域中一项欣欣向荣的科技前沿。对数据挖掘过程模型的研究很多，也已出现多种过程模型，如 SEMMA（Sample Explore Modify Model Assess）过程模型、CRISP-DM（Cross-Industry Process for Data Mining）过程模型、KDD Process 模型等。

图 1.2 显示了 CRISP-DM 过程模型[5]，即跨行业数据挖掘过程标准模型，是数据挖掘业界通用的标准之一，获得了人们的广泛认同。它为一个数据挖掘项目提供了一个完整的过程描述，并从数据挖掘技术的应用角度将一个数据挖掘项目分为 6 个不同的、但顺序并非完全不变的阶段，包括商业理解（business understanding）、数据理解（data understanding）、数据准备（data preparation）、模型构建（modeling）、模型评估（evaluation）和模型发布（deployment）。CRISP-DM 过程模型实施一个数据挖掘项目时，各阶段需开展的工作汇总在表 1.1 中，分述如下。

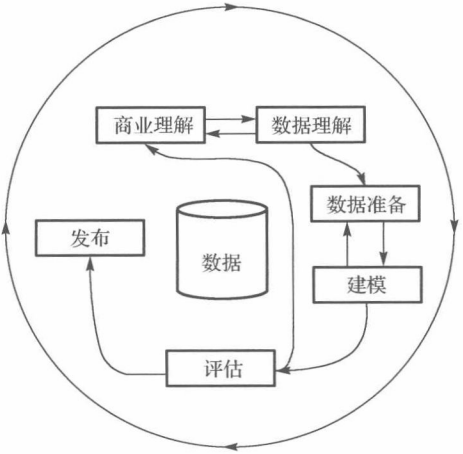

图 1.2　CRISP-DM 过程模型

表 1.1　CRISP-DM 各阶段任务

商业理解	数据理解	数据准备	模型构建	模型评估	模型发布
确定商业目标	收集初始数据	数据选择	选择模型技术	结果评估	模型发布
评估形势	数据描述	数据清洗	生成、测试设计	过程回顾	监控和维护设计
确定挖掘任务	观察数据	数据构造	模型创建	下一步计划	生成最终报告
项目实施计划	确认数据质量	集成数据	评估模型		任务回顾
		数据格式化			

（1）商业理解：这一初始阶段主要从商业角度理解项目目标及需求，并转化为数据挖掘的问题定义，制定达成目标的初步计划。

（2）数据理解：此阶段从数据采集工作开始，并进行诸如熟悉数据、探测数据（以发现其中有趣的数据子集等）、描述数据、检验数据质量等工作。

（3）数据准备：数据准备阶段涵盖了从原始数据中构建最终数据集的全部工作。根据实际情况，此阶段工作有可能反复进行。

（4）模型构建：在实际应用中，对同一个数据挖掘问题，可能存在多种方法供选择。因此，在这个阶段，需要选择和使用不同类型的数据挖掘模型、方法，同时校准模型参数为最优的值。另外，一些模型方法可能对数据格式、数据类型、数据质量有特别的要求，为构建这些模型，有时需要重新回到数据准备阶段执行某些任务。

（5）模型评估：从数据分析的角度看，高质量的模型已经在这一阶段之前构建完成。但为确保所建立的模型确实达到了预期目标，有必要在模型发布前回顾构建模型过程所执行的每一个环节并对模型进行评估。

（6）模型发布：将模型发现的结果以及数据挖掘过程以某种方式组织成可视化的形式（如文本形式），便于用户查看。

在模型构建阶段，当前已有许多模型方法可供选择，比较有代表性的方法包括分类分析、聚类分析、关联分析和回归分析等，分别需要建立分类模型、聚类模型、关联分析模型和回归分析模型，再使用相应的机器学习算法进行分析。根据机器学习算法是否利用和如何利用数据集中样本的类别标号信息，又可以分为监督学习（supervised learning）、半监督学习（semi-supervised learning）和无监督学习（unsupervised learning）等类型。其中，监督学习算法利用样本的类别信息来指导学习过程，无监督学习算法使用的数据均是未标记类别的，而半监督学习主要考虑如何同时利用大量没有类别标记的数据和少量已标记类别的数据进行学习。下面简要介绍这些数据挖掘模型以及它们的基本功能。

1.2.1　分类分析

分类（classification）是一种典型的有监督学习方法，它依据数据特征将数据对象分配到不同的类别，在实际应用中具有重要意义。例如，在银行客户信用评价中，要将客户分为“可信”和“不可信”等类别时，通过分析各类别客户的特征以及特征与类别之间的关系，以发现决定它们分类的关键特征和分类规则，从而预测新客户属于哪种类别，对是否授予客户信用额度等提供辅助决策支持；电商平台将用户在某一段时期内的购买或网页浏览记录划分为不同的类别，在此基础上根据这些分类情况向用户推荐他们可能感兴趣的商品，从而增加平台的销售量等。分类分析主要用于预测数据对象的（离散型）类别。

给定训练数据集 $\mathrm{Tr}=\{(\boldsymbol{x}_1,z_1),(\boldsymbol{x}_2,z_2),\cdots,(\boldsymbol{x}_i,z_i),\cdots,(\boldsymbol{x}_N,z_N)\}$，其中记号 $\boldsymbol{x}_i$ 表示第 i 个训练样本，$z_i\in\{1,2,\cdots,K\}$ 表示 $\boldsymbol{x}_i$ 的类别标号，$K(K>1)$ 是类别数目。设 $\boldsymbol{x}$ 表示任意一个样本，z 是 $\boldsymbol{x}$ 的类别标号，分类问题就是从 Tr 构造一种映射关系 $f:\boldsymbol{x}\to z$。给定一个未知类别的测试样本 $\boldsymbol{x}_t$ 时，可以使用 f 确定的映射关系赋予 $\boldsymbol{x}_t$ 的类别标号 z_t。这样的 f 就称为分类模型或分类函数。

机器学习、专家系统、统计学、生物计算等领域的专家提出了为数众多的分类模型和相应的分类器（classifier），总体而言，可以分为两种类型：“懒（lazy）”型分类器及与之对应的“急切（eager）”型分类器。前者以 k-近邻（k-NN）分类器为代表，它们在训练阶段并未显式地建立分类模型，而在预测阶段使用训练样本实施分类；后者则在训练阶段使用训练样本构造各式分类模型，在预测阶段使用分类模型对新样本进行分类，典型的方法包括贝叶斯分类、决策树、基于规则的方法、神经网络、遗传算法分类等，也有若干介于二者之间的分类方法，通常称为“半懒（semi-lazy）”型方法。一些常用的分类方法参见 6.1 节。

需要注意的是，各式分类器的性能与数据的特性密切相关。例如，有些数据包含较多的噪声，有些数据存在缺失值（missing values）问题，有些数据分布较为稀疏，有些数据的特征间存在显著相关性，而有些数据混合了数值型、类属型等属性。因此，

一般认为，不存在适合于所有特性数据的分类方法。在实际应用中，需要根据数据的不同特点并结合应用背景来选择合适的分类方法。

1.2.2　聚类分析

“物以类聚”，聚类是现实世界中普遍存在的现象。数据聚类（data clustering）是根据数据内在的统计特性，发现其中未知的对象类。直观地说，聚类是一种对具有共同趋势和模式的数据对象进行分组的方法，它将数据对象分组成为多个类或簇（cluster），在一个簇中的对象之间具有较高的以某种度量为标准的相似度，而不同簇中的对象差异较大。通过聚类，人们能够识别密集的和稀疏的区域，发现全局的分布模式，以及数据属性之间有趣的相互关系。

给定数据集 $\mathrm{DB}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{iD})^{\mathrm{T}}$ 表示 D 维空间的一个数据对象，N 为对象数目。给定任意两个数据点 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 间的相似性度量 $\mathrm{sim}(\boldsymbol{x}_1,\boldsymbol{x}_2)$，所谓聚类，就是依据相似性度量将 DB 划分为若干个非空数据子集的集合 $C=\{c_1,c_2,\cdots,c_K\}$ 的过程，称 $c_1,c_2,\cdots,c_K$ 为 DB 的簇或簇类，使得同一个簇中的数据对象彼此相似，而隶属于不同簇的数据对象尽可能彼此不相似。这里 K 表示簇数目，通常 $K>1$。

作为统计学的一个重要分支，聚类分析已经被广泛研究了许多年，早期研究可以追溯到 20 世纪 40 年代，在发展过程中，一些研究曾归入统计学和机器学习范畴。在统计学中，聚类一般称为“聚类分析”（cluster analysis），主要集中于基于距离的聚类分析；在机器学习中，聚类称为无监督的学习，主要体现在聚类学习的数据对象没有类别标记，需要由聚类算法自动计算。在很多应用场合，数据集中蕴涵的聚类数目也是未知的，此时要求聚类方法能够评估聚类结果的质量，进而确定数据集的最佳聚类数目，这是聚类有效性（cluster validity）研究的主要内容。

由于这种无监督特性，聚类得以在许多领域广泛应用，包括模式识别、数据分析、图像处理、市场研究等。在商业上，聚类能帮助市场分析人员从他们的消费者数据库中区分出不同的消费群体，用购买模式刻画不同的客户群体特征，以更好地销售商品和拓展市场；在生物信息学中，聚类可以用于辅助研究动、植物的类别，识别具有相似功能的基因，获得对某种群体固有结构的认识。聚类还可以用来从地理数据库中识别出具有相似土地用途的区域；用于对 Web 文档的分类，以对相似主题的文档进行自动归类等。同时，聚类作为数据挖掘中的一个模块，既可以作为一个单独的工具以发现数据分布的一些深入信息，并且概括出每一类的特点；又可以作为数据挖掘其他分析算法的预处理步骤。

近些年来，随着 KDD 技术的兴起以及应用领域的扩展和深化，聚类研究进入蓬勃发展的阶段。在这个富有挑战性的研究领域中，聚类研究工作已集中于为大型数据库的有效和实际的聚类分析寻找适当的方法，活跃的研究主题集中在聚类方法的可伸缩性、对各种类型数据的有效性以及高维数据聚类分析等[1-4]。若干新型聚类方法参见第 5 章。

1.2.3　关联分析

关联（association）分析用于发现隐藏在交易数据、关系数据或其他信息载体中有价值的数据项之间的关系。所发现的关系通常以关联规则或频繁模式的形式来表示。典型的应用案例包括购物篮分析（market basket analysis）：在商业应用中，通过对顾客购买记录的关联分析，以发现顾客的购买习惯。下面是一个众所周知的例子：某超市通过关联分析，从其交易数据库中惊奇地发现，顾客在购买啤酒的同时经常也购买尿布（这种独特的销售现象出现在年轻的父亲身上）；于是，超市调整货架布局，将啤酒和尿布放在一起以增进销量。由此可见，从大型数据集中发现关联规则，对于改进部分商业活动的决策具有非常重要的作用。

关联规则可以使用表达式 $IS_L \rightarrow IS_R$ 来形式化，其中 IS_L 和 IS_R 是两个不相交的项集，即 $IS_L \cap IS_R = \varnothing$。关联规则的强度一般使用支持度（support）和置信度（confidence）来度量。支持度确定了规则在给定数据集中的频繁程度，而置信度则确定了 IS_R 在包含 IS_L 的事务中出现的频繁程度。支持度和置信度的计算方式分别为

$$\text{Support}(IS_L \rightarrow IS_R) = \Pr[IS_L \cup IS_R] \tag{1.1}$$

$$\text{Confidence}(IS_L \rightarrow IS_R) = \frac{\text{Support}(IS_L \cup IS_R)}{\text{Support}(IS_L)} \tag{1.2}$$

式（1.1）和式（1.2）表明规则 $IS_L \rightarrow IS_R$ 的置信度容易从 IS_L 和 $IS_L \cup IS_R$ 的支持度计算出来，即当我们知道 IS_L、IS_R 和 $IS_L \cup IS_R$ 的计数，就很容易导出对应的关联规则 $IS_L \rightarrow IS_R$ 或 $IS_R \rightarrow IS_L$。这种规则在数据库中是常见的，而人们通常只对满足一定条件（如具有较大的支持度和置信度）的关联规则感兴趣。因此，为发现有价值的关联规则，需事先给定最小支持度和最小置信度两个阈值，前者表示一组商品（项）集在统计意义上需满足的最低程度，后者反映了关联规则的最低可靠度。满足上述条件的规则被称为强关联规则。在算法层面，挖掘强关联规则的问题可以首先归结为频繁项集的挖掘问题，通常包括以下两个步骤。

（1）找出所有的频繁项集：根据定义，这些项集在交易数据出现的频率至少与预定义的最小支持度阈值一样。

（2）由频繁项集产生强关联规则：根据定义，这些规则必须同时满足最小置信度条件。

上述关联分析思想还可用于序列模式挖掘。例如，顾客在购买商品时，除了具有上述关联规律外，还可能存在时间上或序列上的规律。一个典型的应用场景如下：顾客购买了某些商品之后，在下次采购时会购买与这些商品有关的另一些商品。

1.2.4　回归分析

回归分析是一种古老且影响深远的数量分析方法，最早由高尔顿在生物统计研究

中提出。其主要目的是研究目标变量（因变量）与影响它的若干相关变量（自变量）之间的关系，通过拟合类似 $Z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_DX_D$ 的关系式来揭示变量之间的关系。关系式中的待定系数 $\beta_0,\beta_1,\beta_2,\cdots,\beta_D$ 通过最小二乘法等从训练数据中学习得到，一旦确定了这些系数，对于一组新的 $X_1,X_2,\cdots,X_D$ 的值，基于关系式就可以预测未知的 Z 值。因此，回归分析通常用于预测分析、时间序列模型以及发现变量之间的联系。

根据涉及自变量数目的多少，回归分析可以分为一元回归分析和多元回归分析。仅考虑两个变量（一个自变量、一个因变量）间相关关系的分析通常称为一元回归分析，而含有两个或两个以上自变量的回归分析称为多元回归分析。根据因变量和自变量之间内在关系的不同，又可分为线性回归和非线性回归分析。上述的关系式 $Z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_DX_D$ 用于线性回归分析。

回归和 1.2.1 节所述的分类一样，都是用于预测未来数据的模型，区别之处在于，分类用于预测对象的离散型类别，回归则用于预测对象的连续或者有序取值。换句话说，分类主要用于定性预测，例如，基于历史股票交易数据建立分类模型，预测未来股市是“涨”还是“跌”（这里，“涨”和“跌”是两个离散型类别标号）；而回归主要用于定量预测，在上例中，对应于建立回归模型预测未来股市涨跌的幅度（这里的幅度是量化的数值）。

1.3 维 灾 问 题

1.3.1 数据挖掘中的特征

现有的数据挖掘模型大多面向结构化数据，即可以用矩阵表示的关系数据。一些例子见 1.5 节。矩阵的每个行表示一个数据对象（data object），根据上下文有时也称为数据样本（sample）、实例（instance）等；矩阵的每个列为描述数据对象的属性（attribute），也称为特征（feature）、变量（variable）等。数据属性的作用是有差异的，特别在分类分析或回归分析中，需要至少包含一个“类”属性，它是分类或回归方法预测的目标属性。具有 D 个属性的数据对象可以看作 D 维空间的数据点（data point），这里，每个属性就是组成空间的一个特征，每个特征即为空间的一个维（dimension），因此我们也可以将数据对象看作该空间中的 D 维向量（vector）。

包括 1.2 节所述四大模型在内的多数数据挖掘模型都是以这样的矩阵数据为基础。例如，在决策树分类中，从一个数据集（子集）构造一颗决策树（子树）的关键步骤便是从数据矩阵中选择一个重要的列为分割属性；在聚类中，常基于欧几里得距离定义样本间的相似性，而距离函数的计算依据是两个对象属性间的差异；关联分析涉及的事务数据可以看作矩阵数据的一种紧凑表示，其中矩阵的每个行表示一笔交易，

每个列对应一个项，交易包含的项对应的属性取值 1，未包含的取值 0，如此构成了一个稀疏矩阵。在文本挖掘中，这种表示称为向量空间模型（Vector Space Model，VSM）；在回归分析中，主要任务就是分析因变量（目标属性）与其他属性之间的关系，因此其处理的数据也可以用矩阵来表示。

当前，在数据挖掘的许多应用领域，数据正变得越来越复杂。实际上，各种类型的交易数据、文档数据、基因表达数据、网络通信数据等的特征数目（维数）可以达到成千上万。若将这些对象表示成高维属性空间的点，则客观世界中的对象可以用高维数据的集合来表示，对这种数据进行挖掘就是高维挖掘问题[1,6]。典型的高维数据如下。

1）购物篮数据

购物篮数据记录顾客的购买行为，主要用于客户关系管理[7]。考虑用一张表来记录顾客的购买行为，表中的每一行表示一次完整的交易记录，除客户的一些基本信息外，表中的列需要记录交易的时间、地点、交易方式、付款信息以及促销信息等。此外，还可以将每一种商品或服务品种看作一个列，若顾客购买了某种商品或服务，则对应的列做上标志或记录购买的数量或金额，这样购物篮数据就是一种高维的数据。

2）文档数据

各种类型的文档通常使用 VSM 表示。在文本挖掘中，一个典型的文本挖掘系统需要考虑几千个词[8]，每篇文档被表示成一个高维词向量，这里，向量的维数是词条的数目，向量元素为该篇文档中某个词条出现的频度或表示是否出现的 1/0 二元值。在实际应用中，一个文档集合涵盖的词条数量为数众多，有时甚至超过文档的数目。针对文档数据的挖掘有文本分类、文本聚类等方法[9-11]。

3）程序文件数据

程序文件数据是一种反映程序行为特征的数据，它是基于数据挖掘的计算机病毒或恶意软件检测系统的基础。例如，在 Microsoft Windows 操作系统下，可以使用程序调用的 API 函数名[12]或程序文件包含的特征字符串来描述一个程序，这样的数据通常具有几千个维度。

4）基因表达数据

基因表达数据由基因芯片实验产生，数据可以用一个矩阵来表示，矩阵的行代表一个样本，每个列对应一个基因，其数值表示该基因在样本上的表达水平。对基因表达数据的挖掘可以对未知样本进行分类、找到对某种生命现象具有相同表达水平的基因组合等[13]。基因表达数据具有很高的维度，如 ALL-AML 白血病数据[14]的数据维度高达 7129。

5）网络通信数据

网络入侵检测系统为识别网络攻击行为，需要使用许多特征来描述一次网络通信行为。例如，MIT Lincoln Labs 提供的 DARPA 通信数据[15,16]，每条数据由 41 个特征

组成，若对其中的类属型属性展开成二元型属性（展开方法参见1.5.2节），则数据维度达108维。与上述的几种数据相比，通信数据的维度不算太高，但是通信数据具有海量、数据流的特点，这样的数据对基于数据挖掘的入侵检测系统也是一大挑战。

1.3.2 什么是维灾

维灾即维数灾难（the curse of dimensionality），最早由Bellman考虑动态优化问题时提出，指满足一定统计指标（期望与方差）的模型（精度），所需样本的数量将随着维数的增加呈指数增长（或模型复杂程度、表示长度呈指数增长）[17,18]。在数据挖掘领域，维数灾难泛指数据分析中遇到由于属性过多所引发的问题，主要表现在以下几个方面[19]。

1）稀疏性

稀疏性是随着维度增长数据对象在空间分布固有的一个特点。让我们用以下基于网格的直方图方法理解这种稀疏性：考虑一个40维的空间，将空间的每个维度沿中点划分成两个部分，这样可以得到2^{40}个单元；设样本数量为10^6=100万，那么有样本落入的单元数最多只有10^6个。注意到$10^6/2^{40}<10^{-6}$，这意味着在一个单元中发现样本的概率不超过0.000001（假设样本是均匀分布的）。实际上，这是一种非常粗略的空间网格划分方法，即便如此，我们还是可以获得这样的观察：在一个40维的空间中，即使有100万个样本，其分布依然是极其稀疏的。

稀疏性特点提示我们需要谨慎地将一些概念迁移到高维空间，如全局密度的概念。通常，人们将密度定义为一个单元（单位邻域）内的样本数量，而根据上述粗略的空间单元划分方法，每个单元内的样本数目几乎为0。因此，在高维空间中，人们更关心样本的局部密度（如某个子空间中的样本密度）而非全局密度。

2）空空间现象（empty space phenomenon）

空空间现象是维度增长到一定数量时产生的一种“奇怪”现象。下面以球体积随空间维度变化来说明这个现象。用D（$D>1$）表示空间维度，半径为r的“球”（注：$D=2$时为二维平面的圆，$D=3$对应于三维空间的球，$D>3$时通常称为超球，这里统称为“球”）的体积用式（1.3）计算：

$$\text{Volume}(D)=\frac{\pi^{D/2}}{\Gamma(D/2+1)}r^D \tag{1.3}$$

根据式（1.3）计算若干单位球（即$r=1$）体积如下：$\text{Volume}(3)\approx 4.18879$，$\text{Volume}(5)\approx 5.26379$，$\text{Volume}(6)\approx 5.16771$，$\text{Volume}(20)\approx 0.02581$和$\text{Volume}(30)\approx 0.00002$。简单分析可知，当$D<6$时，“球”体积随着维度增加而增大，但是，$D>6$时开始下降，$D>30$时，其数值已经接近0了。这个奇怪的现象告诉我们，用“球形”来描述高维空间（如$D>30$）中的模式时，要谨慎使用“内部”的概念，因为此时球的“内部”可能是空的（体积接近于0）。