第 3 章　数据预处理

3.1　引　　言

数据是对客观世界及对象的一种符号化或数量化的描述与表示。从客观物理世界中获得数据的目的是从中获得能够进行挖掘与分析工作所需要的知识。通过对数据的采集和处理，可以达到获取信息与挖掘知识的目的。例如，气象局采集气象数据以预测天气，海洋生物学家采集海水样品以监测海洋生态的变化等。

随着科学技术的发展，目前可以采用的数据测量手段越来越多，同时可以获得的数据也越来越多。对于通过一定的测量和测试手段获取的数据，在进行挖掘与分析工作之前，数据本身会存在一定的问题。这些问题中有些是因数据自身的不良特性而产生，有些则因受限于获取数据的客观条件而产生。因此在对数据进行挖掘和分析工作之前，通常需要首先对数据进行一定的处理工作，以保证后续挖掘和分析的数据质量，即数据预处理。一个典型的数据预处理过程包括数据清洗、数据集成、数据转换和数据归约等步骤。通常情况下，为了检测和分析数据中所面临的问题，常常会借助数据的描述性汇总方法来观测数据的趋中趋势和散布性，以便分析和发现原始数据中可能存在的问题。

本章的内容安排如下：3.2 节介绍数据预处理相关的基本概念；3.3 节介绍用于检测和分析数据质量问题的数据汇总性描述方法；3.4 节介绍用于消除数据噪声的几类典型的数据清洗方法；3.5 节介绍数据集成相关的概念与方法；3.6 节介绍数据归约和转换相关的方法；3.7 节介绍在数据离散化中所用到的主要方法与技术。

3.2　数据预处理的基本概念

本节将介绍与数据预处理相关的基本概念，包括数据的基本概念、数据的属性，以及实际数据预处理工作中所面对的问题，并介绍数据预处理工作的主要内容。

3.2.1　数据的基本概念

数据是数据对象(data objects)及其属性(attributes)的集合。一个数据对象是对一个事物或者物理对象的描述。一个典型的数据对象可以是一条记录、一个实体、一个案例、一个样本等。而数据对象的属性则是这个对象的性质或特征，例如一个人的肤色、眼球颜色是这个人的属性，而某地某天的气温则是该地该天气象记录的属性特征。

表 3.1 给出了一个关于银行信用卡数据的例子。银行为控制信用卡欺诈风险，对信用卡用户提交的资料都会有记录，表中所示为其中一部分记录的示例。其中，每一行为一条记录，每条记录即一个数据对象，代表一个用户的资料。而每一行的序号、婚姻状态、计税收入、是否欺诈均为数据对象的属性。而每一条记录的某一列即该对象属性的属性值，如序号

为 1 的对象“婚姻状态”属性的值为“单身”。

表 3.1 数据的一个例子：信用卡用户的资料

序号	婚姻状态	计税收入(元)	是否欺诈
1	单身	130000	否
2	已婚	105000	否
3	单身	60000	是

属性值是对一个属性所赋予的数值或符号，是属性的具体化。一个属性可以映射为不同数值类型的属性值，如某人的身高可以是 1.73m，也可以是 173cm，或 1730mm。不同的属性可以映射到相同的属性值空间中，如年龄和序号都可以映射为自然数。受属性的性质影响，不同属性值性质也可能不同，如序号可以不断增长，但人的年龄却有最大值的限制。

属性具有不同的类别，可以按照属性值的类型将属性类别分为以下 4 种。

(1) 名称型属性(nominal)。如身份证号码、眼球颜色和邮政编码等。

(2) 顺序型属性(ordinal)。如比赛排名、学分成绩和身高等。

(3) 间隔型属性(interval)。如日期间隔、摄氏和华氏温度等。

(4) 比率型属性(ratio)。如百分比和人口比例等。

一个属性属于以上 4 种属性的哪一种，取决于属性的属性值是否满足下列 4 种性质：区别性、有序性、可加性和乘除性。名称型属性的属性值只满足区别性性质，即两个名称型属性的属性值可以判断相等或不等，但没有判断大小、加减乘除的意义。顺序型属性的属性值除了满足区别性属性之外，也满足有序性。间隔型属性的属性值满足区别性、有序性和可加性 3 种性质。比率型属性的属性值满足以上全部 4 种性质。

属性除了以上分类之外，还有离散属性和连续属性之分。离散属性只能从有限或可数的属性值集合中取值，通常可以用整数变量表示，如邮政编码、文档中的词数和身份证号码等。二进制属性是离散属性的一个特例。连续属性与离散属性相对，可以从不可数无穷多个属性值中取值，通常取值范围为实数。实际中，通常只用有限多位来表示一个数，因此连续属性在计算机中通常表示为浮点数。

以上介绍了数据对象属性和属性值的概念。与属性和属性值相同，数据也是多种多样的，根据数据的来源、用途和组织方式等可以将数据分成许多类型。这里根据数据的组织方式和相对关系将数据呈现为以下形式。

(1) 记录数据。这种数据由一条条的记录组成，如记录数据、数据矩阵、文档数据和事务数据等。

(2) 图数据。这种数据由记录(点)和记录之间的联系(边)组成，如万维网数据、化学分子结构数据等。

(3) 有序数据。这种数据的记录之间存在时间和空间上的序关系，如序列数据、时间序列数据和空间数据等。

记录数据是数据集由一条条记录组成的数据，每条记录具有相同的属性集合。记录数据是 SQL 数据库所使用的数据类型。表 3.1 所示的数据就是记录数据的一个例子，表中每一行代表一条记录，每条记录都有 4 个属性：序号、婚姻状态、计税收入和是否欺诈。

数据矩阵是记录数据的一种特例。当每个属性都是数值型属性的时候，这些数据对象就可以被看成空间中的点，每一个维度对应一个属性。这样的数据集可以用 $m \times n$ 的矩阵来表示，其中矩阵的行数 m 为记录的条数，矩阵的列数 n 为记录的属性个数。

文档数据是文档集合构成的数据集。在自然语言处理中，在“词袋模型”的假设下将一个文档中词出现的次数作为文档的属性是常见的做法。表 3.2 展示了一个文档集合的数据矩阵表示，其中每一行为一个文档，每一列代表文档中出现某个词的次数。

表 3.2 文档数据表示为数据矩阵：文档中词出现次数

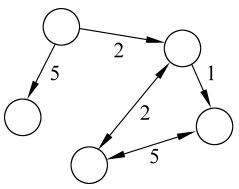
文档序号	team	coach	play	ball	score	game	win	lost	timeout	season
文档 1	3	0	5	0	2	6	0	2	0	2
文档 2	0	7	0	2	1	0	0	3	0	0
文档 3	0	1	0	0	1	2	2	0	3	0

交易数据是记录数据的一种特例，在交易数据中，每一条记录（交易）中包含若干个物品。例如在超市的销售记录中，一笔销售记录包括一个序号和一个物品清单。表 3.3 展示了一个超市销售记录的例子，其中每一条记录是一笔销售。

表 3.3 交易数据的例子：超市销售记录

序　　号	物　　品
1	Bread，Coke，Milk
2	Beer，Bread
3	Beer，Coke，Diaper，Milk
4	Beer，Bread，Diaper，Milk
5	Coke，Diaper，Milk

图数据由点与点之间的连线构成，通常用来表示具有某种关系的数据，如家谱图、分类体系图和互联网链接关系等。在万维网中，网页通常表示为 HTML（超文本标记语言）格式，其中包含可以指向其他网页或站点的链接，如果把这些网页视为点，将链接视为有向边，则万维网数据可以看作一个有向图，如图 3.1 所示。化学分子结构可以视为无向图模型，其中每个点为原子，而其中的线为化学键，如图 3.2 所示。



Data Mining

Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

图 3.1 图数据的例子：万维网数据

有序数据是一种数据记录之间存在序关系的数据，这种序关系体现在前后、时间或者空间上。交易序列数据是一种特殊的有序数据，其中每一个数据都是一个交易序列。表 3.4 所示的超市销售记录序列数据中，每一行为一位顾客的购买记录序列，括号内是一次购买的物品清单，不同括号的先后顺序表示时间上的先后顺序。交易序列数据有助于挖掘在时间上具有先后的一些交易的性质，如重复购买（购买啤酒后常常会购买小孩的纸尿布），或关联商品（购买单反照相机后都会购买镜头）。

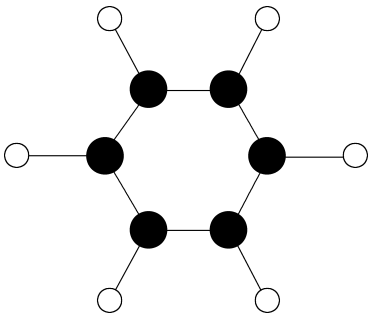


图 3.2　图数据的例子：苯环分子结构

表 3.4　交易序列的例子：超市销售记录

序　　号	购买记录序列
1	(Bread, Coke) (Milk, Coke) (Bread)
2	(Beer, Bread) (Beer, Diaper)
3	(Beer, Coke, Diaper) (Milk) (Beer)
4	(Beer, Bread, Diaper) (Milk)
5	(Coke, Diaper, Milk) (Beer, Diaper)

有序数据还应用在许多其他领域，如生物学中的基因序列、气象学中的气象指数的时空数据等都属于有序数据的范畴。

图数据和有序数据在孤立数据的基础上增加了数据之间的关联性，因此具有比孤立数据更加丰富的信息。由于图数据和有序数据的组织形式的特殊性，通常称对图数据进行的数据挖掘为图挖掘（graph mining），称对序列数据进行的数据挖掘为序列挖掘（sequence mining）。

3.2.2　为什么要进行数据预处理

为什么要进行数据预处理？最主要的原因是数据质量无法满足数据挖掘的要求，如数据可能具有某些不良特性，或者不符合后续挖掘的需要。高质量的数据挖掘结果离不开高质量的数据来源，为了让后续的数据挖掘可以更好地进行，可以对数据进行一些处理和变换，使得预处理后的数据能够满足数据挖掘的需要。

一般来说，高质量的数据应该满足准确性、完整性和一致性的特点。即数据应该准确反映所描述的事实，数据的属性应该是完整的，数据的每个属性应当以一致的原则来表示。遗憾的是，在现实世界中的数据往往不能够满足这些要求。现实世界中的数据有可能是不准确的，如温度、语音和图像等数据经常含有噪声，人工采集和录入的数据可能含有错误等。现实世界中的数据有可能是不完整的，数据的某些属性值可能是缺失的，数据可能没有所关心的某个属性。现实世界中的数据可能是不一致的，同一条数据的不同属性之间可能有着冲突的关系，如某个客户资料中显示他在 1998 年出生却在 1997 年获得博士学位；不同数据记录的同一属性可能具有不同的格式，如一部分学生的成绩可能是百分制分数，而另一部分

学生的成绩却是以优秀、良好、通过、不通过来表示。

数据质量的低劣甚至有着来自现实的原因。由于采集数据时的想法与分析数据时的想法不一定相同，数据的某些属性可能会缺失；采集数据的软硬件可能出现漏洞，致使某些属性值丢失；人工采集数据时可能出现人为错误，被调查用户可能不愿意透露某些数据，这些原因都有可能导致数据的不完整性。同样，在数据的传输过程中可能没有采用无损的传输方式，在数据的采集中有可能没有很好地保存原始信息，这些都可能导致数据中含有噪声。当数据是从不同的源头进行采集时，不同时刻采集过程受到环境的影响，都可能产生数据不一致的现象。

除了以上所提及的数据一致性、完整性、准确性之外，人们通常还会关心其他一些数据质量问题，如时效性、可信性、有价值、可解释性和可访问性等，这里就不一一赘述了。

3.2.3 数据预处理的任务

数据预处理的主要任务包括数据清洗、数据集成、数据转换、数据归约和数据离散化等。

(1) 数据清洗。顾名思义，是对脏数据进行处理并去除这些不良特性的过程。脏数据是指包含噪声、存在缺失值、存在错误和不一致性的数据。通常来说，数据清洗的过程会填补缺失值、对有噪声的数据进行平滑处理、识别并移除数据中的离群点并解决数据的不一致性问题。

(2) 数据集成。是将不同来源的数据集成到一起的过程，这些数据可能来自不同的数据库、数据报表和数据文件。数据集成需要解决数据在不同数据源中的格式和表示的不同的问题并整理为形式统一的数据。

(3) 数据转换。是对数据的值进行转换的过程。在使用某些数据处理方法之前，如 k-均值聚类和贝叶斯分类，对数值进行转换非常必要。因为当数据的不同维度之间的数量级差别很大的时候，分类和聚类的结果会变得非常不稳定，这时通常会对数据进行规范化，对数据值进行统一的放缩。

(4) 数据归约。是对数据的表示进行简化的技术。数据归约使得表示非常复杂的数据可以以更加简化的方式来表示。数据归约可以使得数据处理在计算效率、存储效率上获得较大的提升，而不至于在挖掘分析性能上做出大的牺牲。

(5) 数据离散化。是对连续数据值进行离散化的过程。数据的传输、存储和处理过程都只能对有限位的数据值进行，所以数据离散化是计算机处理数据所必经的一个步骤。数据离散化有时也称为量化，数据在离散化过程中可能会损失部分信息，信息论中的率失真理论给出了量化过程中的信息损失与量化的位数的关系。

数据预处理相关的这些任务都服务于一个目的，即将不完整、不一致、不准确的数据造成的不利影响尽可能地消除，使得后续的数据挖掘工作能够得到高质量的结果。

3.3 数据的描述

当获得大量数据时，比起直接查看这些数据，人们通常更加关心这些数据在整体上具有什么样的特性。为了得到对于数据的整体认识，需要将数据以一定的方式描述出来。

本节将介绍描述数据的方法，包括描述数据中心趋势的方法如均值、中位数，描述数据

的分散程度的方法如方差、标准差，以及数据的其他描述方法如散点图和参数化方法等。

3.3.1 描述数据的中心趋势

假设你是一门课程的教师，拿到了这门课程中学生的百分制成绩列表，你如何评价学生在这门课上的成绩水平？通常会想到首先看一看整体水平是什么样的，如何用一个数来评价学生的整体水平呢？一般会想到的是平均值（mean），又称为均值或算术平均值（arithmetic mean），其计算方式如下：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

例如，对于下列学生成绩列表，其算术平均值为 81.6 分，即平均分是 81.6 分。可以看出，学生的成绩分布大体在平均值附近。

67 98 76 78 70 82 91 85 84 85

如果其中一个同学没有好好复习，在考前打了一晚上游戏，结果考了 12 分，使得成绩列表变成了这样：

12 98 76 78 70 82 91 85 84 85

如果按照平均值计算，就会得出平均值为 76.1 分，但是仔细看一看这个成绩列表就会发现，比平均值低的只有 3 个同学，而有 7 个同学比平均值高，这个平均值并不能很好地反映整体水平，是因为这个考了 12 分的同学造成了这种现象。为了处理这种情况，可以使用截断均值（trimmed mean），即不考虑离群值，用其他值计算平均值。使用截断均值来进行计算：去除第一个同学的分数，余下 9 个同学的分数平均值为 83.2，这比较符合直观印象。

在诸如歌唱比赛、评标等打分环节中，为了避免评委个人的偏好与偏向对整体评分造成影响，通常使用去掉一个最低分，去掉一个最高分，用其他分数计算平均分的手段来进行打分，这就是一种形式的截断均值。

有时在计算平均值时并不希望将所有的数据等同看待，而是希望让一些数据比另一些数据更有代表性。例如在歌唱比赛中，有 5 名评委和 10 名观众（真正的选秀节目的观众数一般很多），他们都可以对歌手进行打分，但是评委一般由专业歌唱家与著名艺人组成，其艺术鉴赏能力和权威性要明显高于一般观众，如果计算算术平均值显然是不适当的。这时可以使用加权算术平均值（weighted arithmetic mean），其计算方式如下：

$$\bar{x}=\frac{\sum_{i=1}^{n}w_i x_i}{\sum_{i=1}^{n}w_i}$$

现在假设一名歌手演唱了一首欣赏难度较大的歌曲，评委对其的评价较好，但是曲高和寡，观众并不买账，打分状况可能就会出现下面的情况。

评委：90 85 80 75 95

观众：80 95 60 40 65 85 70 50 20 40

如果对这样的打分计算简单的算术平均值，就会得到 68.7 的平均分，这个分数并没有显示出该歌手的实际水平，这个打分方式显然并不合理。为了突出评委评分的权威性，同时不打击观众参与的积极性，主办方决定使用加权算术平均值来对歌手进行打分，评委评分的

权重 w_i 是观众评分的权重的 10 倍。这样计算出来的均值就是：

$$\bar{x}=\frac{\sum_{i=1}^{5}10x_i+\sum_{i=6}^{15}x_i}{\sum_{i=1}^{5}10+\sum_{i=6}^{15}1}=\frac{4855}{60}=80.9$$

这个分数既反映了评委在专业层面上对歌手的评分，又在一定程度上反映了歌手缺乏观众人气的现象，是更加合理的打分方式。

有些情况下，均值并不能给出正确的对数据的整体印象。如例 3-1 所示。

例 3-1 某公司的员工收入情况如下：

- 1 名 CEO 兼总裁，年薪 500 万；
- 5 名副总裁，人均年薪 100 万；
- 10 名总监，人均年薪 50 万；
- 40 名中层管理人员，人均年薪 25 万；
- 100 名普通员工，人均年薪 15 万。

该公司的全员平均年薪是 25.6 万，这样的统计数据给人们一种误导，大多数员工也会认为这个统计数据明显与身边的情况不同，他们看到的情况是几乎身边的所有人都会觉得自己的薪资比平均薪资低。这样一个薪资水平能够反映一个公司的一般情况吗？

答案显然是不能，平均薪资不足以反映这个公司的一般薪资状况。因此，可以使用中位数(median)和众数(mode) 两种描述方式。

中位数是将数据排序后处于中间的数，如果数据值是奇数个，则中位数就等于中间的数；如果数据值是偶数个，则中位数等于中间两个数的平均值。使用中位数来描述例 3-1 中的薪资，可以得出该公司的薪资中位数为 15 万的结论。

众数是在数据中出现次数最多的数。使用众数来描述例 3-1 中的薪资，可以得出该公司的薪资众数为 15 万的结论。

如果数据不是离散型数据，而是连续型数据，那么中位数的意义就是累积概率分布函数值为 0.5 的点，该点前的概率密度函数的积分等于 0.5；而众数的意义则是使概率密度函数值最大的点，即最大的峰值对应的数据点。

众数、中位数和均值的关系如图 3.3 所示，对于仅有一个峰值的分布来说，三者之间的关系可以用一个经验公式来描述：

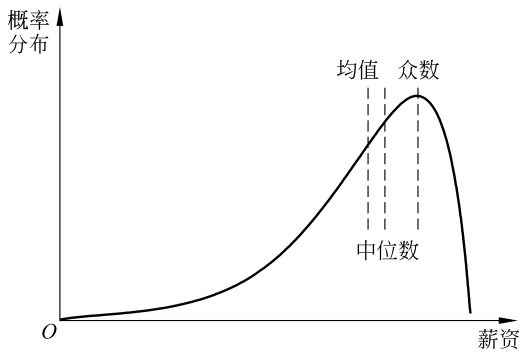


图 3.3 众数、中位数和均值

均值－众数＝3×(均值－中位数)

该公式并不一定总是成立，但是可以在一定程度上反映三者之间的关系。

3.3.2 描述数据的分散程度

除了需要了解数据的中心大体分布在哪里之外，通常还要关注数据的分散程度。回到本节开头的例子，假设你是一门课程的教师，你不仅仅希望了解全班同学成绩的一般水平，可能还希望知道全班同学的成绩之间相差很大还是相差较小，也就是数据的分散程度。

衡量数据的分散程度的一个很好的指标是分位数，α 分位数是从负无穷到某一点概率密度函数的积分(分布列求和)为 α 时那一点的值。比较常用的分位数为最小值(可以认为是 0 分位数)、0.25 分位数(Q_1)、中位数(0.5 分位数)、0.75 分位数(Q_3)和最大值(可以认为是 1 分位数)。

通过这些分位数可以定义一些描述数据分散度的指标。范围是最大值与最小值之差，它描述了数据分布在多大的范围中；中间四分位数极差(IQR)是 Q_3-Q_1，它反映了数据中心部分的分散程度；五数概要是上述 5 个分位数的整体，通常被用在箱线图中，用于形象表示数据的范围。

例如，对于下列学生成绩列表，其范围为 98－67＝31 分，其中间四分位数极差为 85－76＝9 分，其五数概要的箱线图表示如图 3.4 所示。从箱线图中可以看出，前 25％的学生成绩相差较大，而中间 50％的成绩分布相对比较集中。

67 98 76 78 70 82 91 85 84 85

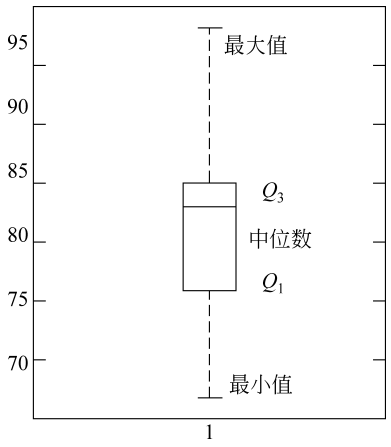


图 3.4 箱线图：五数概要的可视化

在箱线图中，有些数据点由于过于脱离整体，通常希望把它们单独表示出来，这些点称为离群点(outlier)。通常使用点与最近的中间四分位数的差来判断是否属于离群点，通常使用一个常数 k(经验值为 1.5)与中间四分位数极差的成绩来定义这个临界差值。即当数据不属于以下区间时，认为数据为离群点：

$$[Q_1-k(Q_3-Q_1),Q_3+k(Q_3-Q_1)]$$

衡量数据分散程度的另外两个常用的指标是方差和标准差。方差通常用 S^2 表示，是数据的平方误差的期望，样本的(无偏)方差的计算公式为：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

标准差通常用 s 表示，标准差是方差的均方根值。正态分布是一种典型的概率分布，其概率密度函数可以使用均值 μ 和标准差 σ 两个参数来表示：

$$N(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正态分布是分布比较集中的单峰分布，其主要的概率集中在均值附近，其中，$[\mu-\sigma,\mu+\sigma]$集中了 68％的概率，$[\mu-2\sigma,\mu+2\sigma]$集中了 95％的概率，$[\mu-3\sigma,\mu+3\sigma]$集中了 99.7％的概率。正态分布的概率分布如图 3.5 所示。

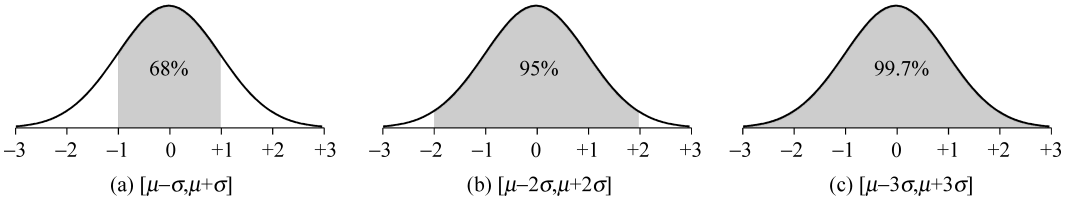


图 3.5 正态分布的概率分布

3.3.3 描述数据的其他方式

除了以上介绍的描述数据中心趋势和描述数据分散程度的描述方式外，还有一些其他数据描述方式能够揭示数据中的更多信息。下面介绍直方图、分位数图、Q-Q 图和散点图这几种数据描述方式。

直方图是一种用长方形表示数据分布的统计图形。直方图将数据的分布范围分为几个区间，并用面积表示每个区间的数量或频率分布。如在本节开始时的成绩分布的例子中，可以画出如图 3.6 这样的直方图。

从直方图可以看出数据在各个区间的分布情况，它比数据的均值和方差更直观地反映了数据的分布情况，通常比较数据值在不同区间上的差异时会使用直方图。

分位数图是一种反映在[0,1]区间内的分位数统计图形。其横轴为概率，通常为[0,1]区间；而纵轴为对应横轴的分位数。分位数图可以直观地看出中位数、上下四分位数等统计指标，也可以通过斜率看出数据的分布情况。分位数图上，斜率越低的地方分布越集中。如本节开始时的例子中，学生成绩的分位数图如图 3.7 所示。

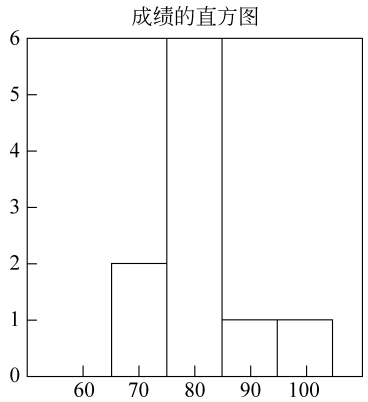


图 3.6 学生成绩的直方图

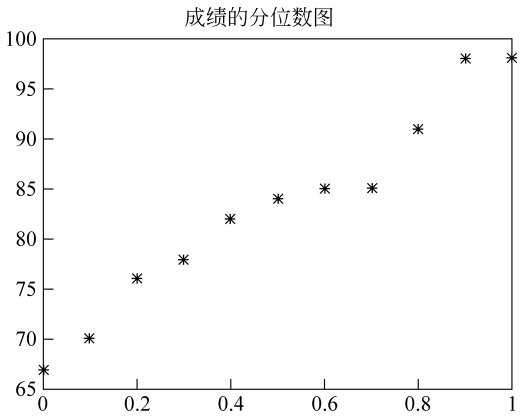


图 3.7 学生成绩的分位数图

与前两个统计图形不同，Q-Q 图和散点图不是描述一个概率分布的统计图形，它们是描述两个分布之间的统计关联的图形。

Q-Q 图（分位数-分位数图）是描述两个单变量分布的分位数的图，从 Q-Q 图上比较容易读出两个分布之间的偏移。Q-Q 图通常用在两个分布比较类似的情况，如一个行业不同品牌的售价分布的比较这样的场合。

对如下两个班的成绩列表，可以得到 Q-Q 图如图 3.8 所示。

A 班：67，98，76，78，70，82，91，85，84，85

B 班：61，68，67，78，82，84，85，88，97，98

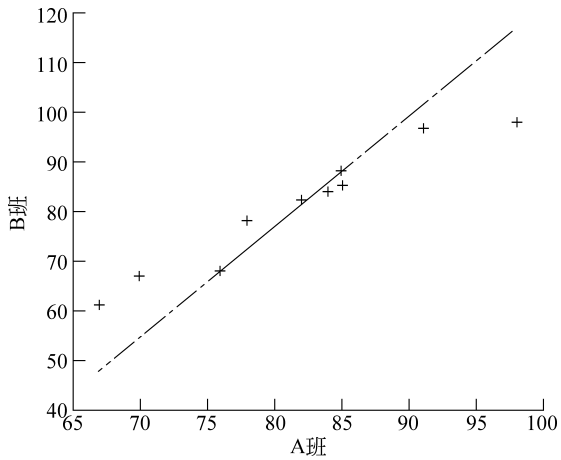


图 3.8　两个班学生成绩分布的 Q-Q 图

从图 3.8 中可以看出，在高分段和低分段中，B 班比 A 班分布都更集中一些。

散点图是一种用散点方式来描述数据在多个维度上分布的统计图形。通常用来通过在两个维度上将数据可视化表示，用以揭示数据在这两个维度上存在的相关关系。散点图中，每一个点对应一条数据记录，点对应的横纵坐标即对应数据在两个维度上的属性。图 3.9 展示了在两个维度有相关性时的散点图的例子，图 3.10 展示了两个维度不相关的散点图。

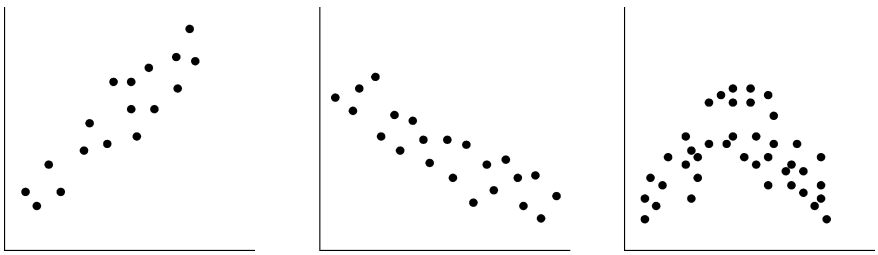


图 3.9　散点图：相关数据

图 3.10　散点图：不相关数据

散点图可以直观地反映两个变量之间是否存在相关与依赖关系，以及一个变量是否可以表示为另一个变量的函数。当一个变量可以近似表示成另一个变量的函数时，散点图可以帮助得出对两个变量依赖关系的直观判断。根据数据拟合两个变量之间依赖关系的过程

称为回归分析，回归分析有参数化方法和非参数化方法两种，其中参数化方法需要对模型具有先验知识，这种先验知识可以通过对散点图的观察获得。

3.4 数据清洗

数据清洗是数据预处理中非常重要的一个环节，在这个环节中进行的任务包括填补数据中的缺失值，识别数据中的离群点，对有噪声数据进行平滑等。由于数据挖掘的质量很大程度上依赖数据本身的质量，而数据清洗在提升数据质量方面具有相当大的作用，因此数据清洗是数据挖掘中的重要步骤。

3.4.1 数据缺失的处理

由于很少能从现实世界中获得完美的数据，在数据预处理中，经常能够见到数据缺少某些属性值的情况。比如在一项调研中，受访者有时会拒绝填写个人收入这样比较敏感的数据，这就造成了数据缺失的情况。

数据缺失可能由各种原因导致，采集设备的故障可能会造成空白数据，一个属性可能与其他属性产生冲突而造成它被删除，数据在录入阶段可能出现误解而未能录入，在数据录入的时刻可能某个属性并不受重视而未被采集，采集数据的需求可能发生了变化造成数据属性集合的变化。这些林林总总的理由都可能导致数据值的缺失，以至于需要对这些数据进行分析时，部分数据很可能缺少某些重要的属性。

怎样处理缺失数据？最简单的处理方法是当数据的某个属性缺失时，丢弃掉整条数据记录。这种处理方式在许多时候是不得已而为之的策略，当缺失的属性值至关重要而且填补属性值的意义不大时，通常采取这个策略。例如，类别标签在有监督分类中是不可或缺的，当拿到一批数据进行有监督分类时，如果数据没有类别标签，这些数据就无法在训练或测试中使用，这时能够应对此类情况的最好策略就是丢弃这部分不完整的数据。

处理缺失数据的另外一种方式是人工填补缺失值，即对于某些缺失的属性，用人工的方式进行填补。人工填补的前提是数据存在一定的冗余，其缺失属性可以通过其他属性进行推断。人工填补的方式存在一些弱点，首先是数据填补的影响难以预计，受人的主观因素和知识背景的影响；其次是人工处理数据的规模受到人工成本的限制，难以处理较大规模的数据；以及当参与处理的人员过多时，填补的标准性难以得到保证。因此，人工填补仅仅应用在有限的情境下。

对于缺失数据采用较多的处理方式是自动对缺失值进行填补。自动填补数据最简单的办法对所有缺失某一属性的数据填补统一的值。例如，在客户调查数据中，收入属性的类型是整数，工作单位属性的类型是字符串，则可对未填写收入的所有客户的收入属性填补为0，对未填写工作单位的所有客户的工作单位属性填补为空字符串。许多数据库如 MySQL 数据库都提供了默认值功能，可以自动对缺失属性填充统一的默认值。

使用统一的值进行填充有时会带来问题，例如，在统计消费者的收入水平，分析消费者的消费能力时，将缺失的收入属性统一填充为 0，就可能对统计结果造成偏差，对相关性的分析造成干扰。这时可以使用属性的均值对该属性的所有缺失值进行填补，这样可以减少对数据的干扰。图 3.11 展示了一个使用属性均值对属性的缺失值进行填补的例子。

名字	公司	工资(元/月)
Amy	A公司	13000
Alice	A公司	?
Mike	A公司	9000
Joey	B公司	5000
Tom	B公司	?
Zelda	B公司	3000

名字	公司	工资(元/月)
Amy	A公司	13000
Alice	A公司	7500
Mike	A公司	9000
Joey	B公司	5000
Tom	B公司	7500
Zelda	B公司	3000

图 3.11　使用平均值填补缺失数据

在图 3.11 所示的例子中，可能会提出一个疑问，从数据来看，A 公司的工资水平似乎比 B 公司要高一些，那么使用属性均值来估计，很可能 Alice 的工资被低估了，而 Tom 的工资被高估了。可以提出，为什么不用相同工作单位的平均工资水平来填补缺失值呢？

图 3.12 所示的例子中，使用 Amy 和 Mike 的工资平均值来填补 Alice 的工资属性，使用 Joey 和 Zelda 的工资平均值来填补 Tom 的工资属性。经过这样的填补，数据变得更加一致，也更加符合常识。还可以利用同类别属性的其他统计特性，如中位数、众数等。

名字	公司	工资(元/月)
Amy	A公司	13000
Alice	A公司	?
Mike	A公司	9000
Joey	B公司	5000
Tom	B公司	?
Zelda	B公司	3000

名字	公司	工资(元/月)
Amy	A公司	13000
Alice	A公司	11000
Mike	A公司	9000
Joey	B公司	5000
Tom	B公司	4000
Zelda	B公司	3000

图 3.12　使用同类别数据的属性平均值填补缺失数据

更进一步地，可以通过更加智能的方式利用更加丰富的信息来处理缺失值，可以将缺失值本身作为预测的对象，通过一些存在的属性来对缺失值进行预测。例如，可以通过客户的工作单位、学历水平、存款的多少、不动产的状况来对客户的收入进行预测。可以采用的预测方法有线性回归、决策树模型和最大似然估计等。

3.4.2　数据清洗

数据噪声是指数据中存在的随机性错误和偏差，许多原因可能导致这些错误与偏差。其中，数据采集中一些客观因素的制约带来了数据噪声。数据采集设备可能具有缺陷和技术限制。例如，在数码相机中使用的 CCD(电荷耦合设备)图像传感器本身可能具有暗电流和热噪声，不可避免地造成图像中的噪声，这种情况在暗光条件下表现得尤为突出。数据传输中信道一般是有失真的。例如，在模拟电视信号的传输中，像素的亮度等信息都是采用模拟方式调制的，这使得模拟电视信号容易出现色度损失、变形、抖动和串扰等情况。同时，数据采集过程中的人为错误也会引起噪声，如数据录入中的读数误差、对于数据的命名约定不一致等情况都会造成错误的数据值。

在数据挖掘领域中，为了保证数据预处理工作的高效，为了处理噪声数据，通常用到的

方法是分箱、聚类分析和回归分析等，有时也会将计算机决策与人的主观判断相结合。

分箱是一种将数据排序并分组的方法，通常使用的分箱方式有等宽分箱和等频分箱。所谓等宽分箱，是用同等大小的格子来将数据范围分成 N 个间隔，箱宽为：

$$W=\frac{\max(\text{data})-\min(\text{data})}{N}$$

等宽分箱比较直观和容易操作，但是对于有尾分布的数据，等宽分箱并不是太好，可能出现许多箱中没有样本点的情况。另一种分箱方式是等频分箱，又称为等深分箱，这种分箱方法将数据分成 N 个间隔，每个间隔包含大致相同的数据样本点数，这种分箱方法具有较好的可扩展性。

将数据分箱后，可以用箱均值、箱中位数和箱边界来对数据进行平滑，平滑可以在一定程度上削弱离群点对数据的影响。下面用一个例子来说明使用分箱处理噪声数据。

例 3-2 对如下数据采用分箱方式进行平滑：4,8,9,15,21,21,24,25,26,28,29,34。

数据共有 12 个样本，故可以将其分成等频的 3 个箱，每个箱中有 4 个样本。

箱 1：4,8,9,15

箱 2：21,21,24,25

箱 3：26,28,29,34

使用箱均值进行平滑，用每个箱中数据的均值来代替每个数据。

箱 1：9，9，9，9

箱 2：22.75，22.75，22.75，22.75

箱 3：29.25，29.25，29.25，29.25

使用箱中位数进行平滑，用每个箱中数据的中位数值来代替每个数据。

箱 1：8.5，8.5，8.5，8.5

箱 2：22.5，22.5，22.5，22.5

箱 3：28.5，28.5，28.5，28.5

使用箱边界进行平滑，用每个箱边界中距离数据样本点最近的边界值来代替数据样本点。

箱 1：4，4，4，15

箱 2：21，21，25，25

箱 3：26，26，26，34

聚类分析是一种将相似数据聚在一起，将不相似的数据分开的过程。聚类通常用来发现数据中隐藏的结构，在没有标注的情况下将数据分为一些类别，通过聚类分析还可以发现数据中的离群点等信息。聚类的一个例子如图 3.13 所示，在图中数据点被分为三个数据簇(cluster)。

在数据清洗中，可以对数据进行聚类，然后使用聚类结果对数据进行处理，如舍弃离群点、对数据进行平滑等。对数据进行平滑的方法类似于分箱，可以采用中心点平滑、均值点平滑等方式来处理，这里不再赘述。

回归分析是一种确定变量依赖的定量关系的分析方法。在正确的建模下，回归分析可以揭示数据变量之间的依赖关系，通过回归分析进行的预测能够比较接近数据的真实值。图 3.14 所示是一种简单的回归分析——线性回归的一个示例。

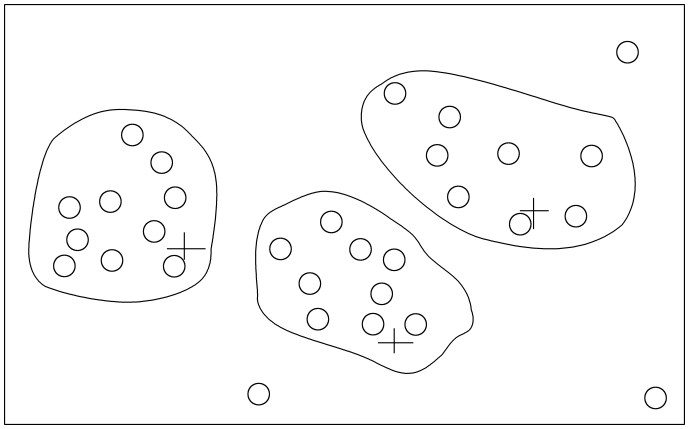


图 3.13　聚类(示意图)

对于建立好的回归分析模型，可以使用参数估计方法对模型参数进行估计。如果具有对数据的某种先验知识，使得模型符合数据的实际情况，并且参数估计是有效的，就可以使用回归分析的预测值来代替数据的样本值，以削弱数据中的噪声，并降低数据中离群点的影响。

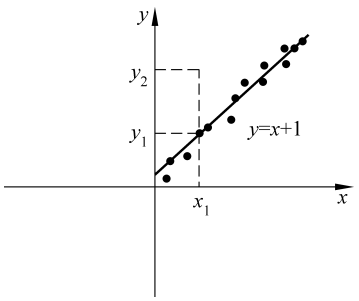


图 3.14　线性回归

数据清洗的过程通常是由两个过程的交替迭代组成的：数据异常的发现和数据的清洗。对于数据首先需要进行审查，根据先验知识如数据的取值范围、数据依赖性、数据的分布、数据的唯一性、连续性和空/非空性质等，可以发现数据中存在的异常现象。在发现数据异常后，使用数据清洗方法对数据进行转换。数据转换可以使用专门的数据迁移工具进行，通常称为ETL(Extract，Transform，Load)工具。

3.5　数据集成和转换

3.5.1　数据集成

数据集成是将不同来源的数据整合并一致地存储起来的过程。不同来源的数据可能有不同的格式、不同的元信息和不同的表示方式等。当需要对这些数据进行统一处理时，首先需要将它们变成一致的形式。通常这个过程牵涉到数据架构的集成，处理属性值冲突，处理数据冗余，对数据进行转化等处理过程。下面介绍数据集成过程中两个主要的问题：数据冗余和数据转换。

3.5.2　数据冗余

在多源数据的集成过程中经常会遇到数据冗余的问题，数据冗余可能由许多技术和业务上的原因导致，同一属性或对象在不同的数据库中的名称可能是不同的，某些属性可能是由其他属性导出的，这些原因都可能导致数据的冗余。

数据的冗余性可以通过对数据进行相关性分析发现，对发现的数据冗余进行处理，可以消除和避免冗余性，进而提高数据挖掘的效果和效率。下面介绍两种数据相关性的分析工具——皮尔森相关系数和卡方检验。

1）皮尔森相关系数

皮尔森相关系数计算两个数值向量之间的相关性，其计算方法如下：

$$r_{A,B}=\frac{\sum_{n=0}^{N-1}(A[n]-\bar{A})(B[n]-\bar{B})}{(n-1)\sigma_A\sigma_B}=\frac{\sum_{n=0}^{N-1}A[n]B[n]-n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

其中，$\bar{A}=\frac{1}{N}\sum_{n=0}^{N-1}A[n]$，$\bar{B}=\frac{1}{N}\sum_{n=0}^{N-1}B[n]$ 为样本均值，σ_A 为 A 向量的无偏标准差，σ_B 为 B 向量的无偏标准差。当相关系数大于 0 时，称两个向量正相关；当相关系数小于 0 时，称两个向量负相关；当相关系数等于 0 时，称两个向量不相关。容易得出，相关系数的取值范围是 $[-1,1]$。图 3.15 用不同相位的余弦函数展示了皮尔森相关系数的一个例子。

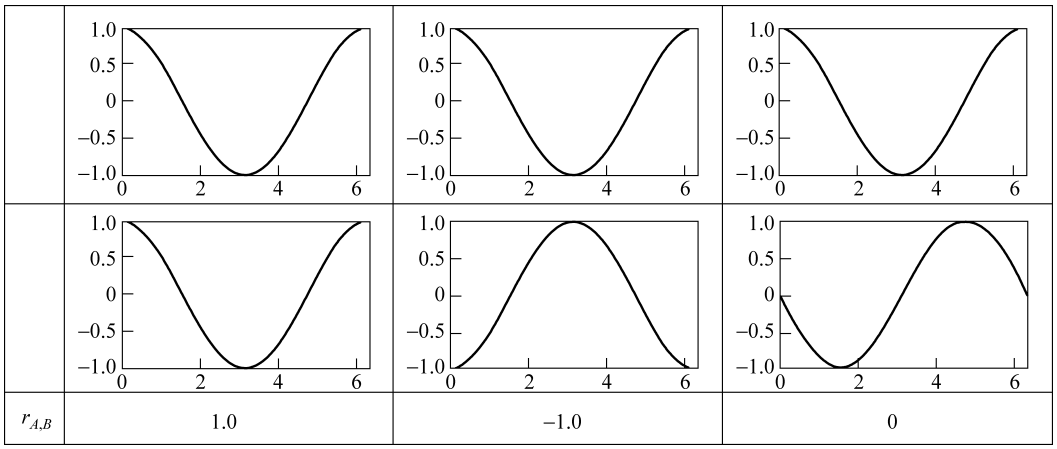


图 3.15　皮尔森相关系数

2）卡方检验

对于非数值型的变量，计算其相关性可以使用卡方检验方法进行，卡方检验的计算方式为：

$$\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$$

其中，求和是对每一种不同的变量取值情形进行的，O_i 是实际观测到的概率，而 E_i 是在变量彼此独立的假设下该情况发生概率的估计。

例 3-3　某社交网站的用户习惯调查中收集了 1000 名用户对于“是否喜欢下国际象棋”和“是否喜欢科幻小说”的回应，分析用户这两种喜好之间是否存在相关关系。

喜欢下国际象棋且喜欢科幻小说的有 300 人；

喜欢下国际象棋且不喜欢科幻小说的有 50 人；

不喜欢下国际象棋且喜欢科幻小说的有 150 人；

不喜欢下国际象棋且不喜欢科幻小说的有 500 人。

在两个变量独立的假设下：

$$E(\text{喜欢下国际象棋且喜欢科幻小说}) = P(\text{喜欢下国际象棋})P(\text{喜欢科幻小说})$$

$$P(\text{喜欢下国际象棋}) \approx \frac{300+50}{1000} = 0.35$$

$$P(\text{喜欢科幻小说}) \approx \frac{300+150}{1000} = 0.45$$

而在实际观测下：

$$O(\text{喜欢下国际象棋且喜欢科幻小说}) \approx \frac{300}{1000} = 0.3$$

类似地，可以计算出每个格子对应的 O_i 与 E_i，数据分布的卡方统计量为：

$$\chi^2 = \frac{(300-157.5)^2}{157.5} + \frac{(150-292.5)^2}{292.5} + \frac{(50-192.5)^2}{192.5} + \frac{(500-357.5)^2}{357.5} \approx 361$$

计算出统计量后，结合卡方统计量的分布，就可以通过假设检验判断变量间的相关性是否成立。在这里可以不严格地看出，这两个变量间还是存在一定的相关关系的。

在对数据进行相关性分析时，需要注意的是，相关性并不意味着因果关系。例如，一个城市居民购买节能灯泡的总数量和当地政府的节能宣传的总开支可能存在相关关系，但是并不一定构成因果关系，因为这两者可能都与城市的人口数量成正相关关系。把相关关系错当成因果关系可能会得出荒谬的结论。

3.5.3 数据转换

数据在集成过程中很多情况下需要进行转换，数据转换包括平滑、聚合、泛化、规范化、属性和特征的重构等操作。

(1) 数据平滑。数据平滑是将噪声从数据中移除的过程。数据平滑通常是对数据本身进行的，如在连续性的假设下，对时间序列进行平滑，以降低异常点的影响；数据平滑有时也指对概率的平滑，例如在自然语言处理中常用的 n 元语言模型中，对于未在训练样本中出现过的词组一般不能赋予零概率，否则会使整句话概率为 0，对这些词赋予合理的非零概率的过程也称为数据平滑。

(2) 数据聚合。数据聚合是将数据进行总结描述的过程。数据聚合的目的一般是为了对数据进行统计分析，数据立方体和在线分析处理(OLAP)都是数据聚合的形式。

(3) 数据泛化。数据泛化是将数据在概念层次上转化为较高层次的概念的过程。例如，将一个词语替换为词语的同义词的过程，将分类替换为其父分类的过程等都是数据泛化。

(4) 数据规范化。数据规范化是将数据的范围变换到一个比较小的、确定的范围的过程。数据规范化在一些机器学习方法的预处理中比较常用，可以改善分类效果和抑制过学习。常用的数据规范化方法有最小最大规范化、z-score 规范化和十进制比例规范化等。最小最大规范化是用数据的最小最大值将数据转化到某一区间的方法，如下的公式是最小最大规范化的例子，它将数据映射到 [0,1] 区间：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

z-score 规范化使用数据的均值 μ 和标准差 σ 来将数据转化到某个区间，如下的公式为

z-score 标准化的例子，规范化后的数据均值为 0，标准差为 1。

$$x' = \frac{x - \mu}{\sigma}$$

十进制比例规范化使用数据绝对值的极值进行规范化，对数据仅使用十进制放缩的方式进行规范化。如要将 564，46，－234，－19 这几个数进行规范化，其绝对值的极值为 564，要将其规范化到[－1，1]区间，对所有数据除以 1000，即大于极值的最小的 10 的整数次幂。规范化的结果为 0.564，0.046，－0.234，－0.019。十进制比例规范化只须移动小数点的位置，在一些应用中比较容易实现。在数字信号处理中，规范化经常会采用类似的二进制比例规范化，因为这种操作仅需移位运算。

特征构造是根据需要，利用数据中已经有的属性来构造新的属性的过程。

3.6 数据归约和变换

3.6.1 数据归约

在实际应用中，数据仓库可能存有海量数据，在全部数据上进行复杂的数据分析和挖掘工作所消耗的时间和空间成本巨大，这就催生了对数据进行归约的需求。数据归约是用更简化的方式来表示数据集，使得简化后的表示可以用较少的数据量来产生与挖掘全体数据类似的效果。

数据归约可以从几个方面入手：如果对数据的每个维度的物理意义很清楚，就可以舍弃某些无用的维度，并使用平均值、汇总和计数等方式来进行聚合表示，这种方式称为数据立方体聚合；如果数据只有有些维度对数据挖掘有益，就可以去除不重要的维度，保留对挖掘有帮助的维度，这种方式称为维度归约；如果数据具有潜在的相关性，那么数据实际的维度可能并不高，可以用变换的方式，用低维的数据对高维数据进行近似的表示，这种方式称为数据压缩；另外一种处理数据相关性的方式是将数据表示为不同的形式来减小数据量，如聚类、回归等，这种方式称为数据块消减。

以下分别对这些数据归约方法进行介绍。

1. 数据立方体聚合

数据立方体是一种数据表示和分析的工具，它将数据表示为多维的矩阵，可以对数据进行聚合运算如计数、求和和求平均值等操作。数据立方体将在第 3 章中详细介绍。

利用数据立方体可以对数据进行归约，从而得到能够解决问题的数据的最小表示方式。

图 3.16 展示了数据立方体聚合的一个例子，对于一张员工工资列表，如果希望挖掘不同公司中学历与工资的关系，可以使用数据立方体来将数据表示成图中右边表格的方式，约简后的数据相比源数据的数据量大大减少。

2. 特征选择

特征选择在数据预处理和迭代调整的学习中都有较多的使用，目的是对于给定数据挖掘任务，选择效果较好的较小特征集合。在预处理中，特征选择通常希望能使得在选择出的特征集合下的类别的概率分布能够尽量接近在全部特征下的类别的概率分布，这是为了权

名字	公司	学历	月工资
Amy	A公司	硕士	13 000
Alice	A公司	本科	10 000
Mike	A公司	本科	9000
Joey	B公司	硕士	5000
Tom	B公司	本科	4000
Zelda	B公司	本科	3000

归约

平均月工资		学历	
		硕士	本科
公司	A	13 000	9500
	B	5000	3500

图 3.16 数据立方体聚合

衡空间复杂度、时间复杂度和数据挖掘效果的折中。

在原始的特征有 N 维的情况下，特征子集的可能情况有 2^N 种情形，在 N 较大的情况下对这些情形一一考察其好坏是不现实的。通常使用启发式的方法进行特征选择，如前向特征选择、后向特征消减以及采用决策树归纳进行特征选择等。

前向特征选择是通过选择新的特征添加到特征集合中，使得扩充后的特征集合具有更好的特性。对于特征的衡量可以使用条件独立性，在已有特征集合的条件下，通过显著性检验来确定与类别最相关的特征。在迭代调整的学习中，一种前向特征选择的方法是随机向特征集合中添加特征，如果结果有改善则保留特征，否则就不采用新添加的特征，重新进行挑选。

后向特征消减是通过从特征集合中取出最差的特征，使得新的特征集合具有更好的特性。最佳的特征选择方法一般会结合重复地使用前向和后向特征选择方法，通过迭代调整来确定较好的特征集合。

决策树归纳方法进行特征选择是借助决策树构建来选择较小特征集合的方法。决策树是一种树状的分类模型，样例在每一个非叶子结点进行对一个属性的判断，每个指向子孙结点的路径代表一种属性值的情形，叶子结点为最终判断的类别。决策树的构建中，优先选择最好的特征来作为根结点，然后对于每种可能的情形，递归地建立子树，如果某种情形的样例集合只包含一种类别，就可以用类别标签作为叶子结点，停止子树的生长。如果建立的完整决策树不包含某些特征，就意味着不使用这些特征即可完整地描述数据的分类模型，因此可以使用决策树中非叶子结点的特征作为约简的特征集合。另外，使用剪枝方法，或限制决策树生长的层数，也可以限制决策树使用的特征数量，得到比较重要的特征作为约简的特征集合。

3. 数据压缩

数据压缩是在尽量保存原有数据中信息的基础上，用尽量少的空间表示原有的数据。数据压缩分为有损压缩和无损压缩，有损压缩后的数据信息量少于原有的数据，因而无法完全恢复成原有的数据，只能以近似的方式恢复；无损压缩没有这一限制，从压缩后的数据可以完全恢复原有数据。

无损压缩一般用于字符串的压缩，被广泛应用在文本文件的压缩中。在信息论领域，这一问题在信源编码中得到了深入研究，如哈夫曼提出的具有理论意义的 Huffman 编码，以及广泛使用于 gzip、deflate 等软件中的 LZW 算法（由 Abraham Lempel、Jacob Ziv 和 Terry Welch 提出，基于该算法的专利在 2003 年 6 月 20 日后失效）等都是无损压缩方法。

在图像和音视频压缩中通常使用有损压缩，在图像压缩中常见的离散小波变换就是一种有损压缩，仅仅保存很少一部分较强的小波分量，可以在图像质量无明显下降的情况下获得相当高的压缩率。

主成分分析(Principal Component Analysis，PCA)是一种正交线性变换，它将数据通过正交变换到新的坐标系中，其中第一个分量有最大的方差，第二个分量有第二大的方差，以此类推，数据主要的能量集中在前几个分量中。主成分分析可以帮助人们了解数据的结构，通常应用在处理维数较多的数值型数据中。

4. 其他数据归约方法

除了以上提到的方法之外，还可以对数据进行不同形式的表示，以减小数据量。一般可以将这些方法分为参数式方法和非参数式方法。参数式方法使用模型对数据进行描述，通过一定的准则(如最小错误概率准则、最小二乘准则、最大似然准则和最大后验概率准则等)来估计最佳参数，参数估计完成后，就不再使用原始数据，而是使用模型和参数来描述数据。非参数式方法不使用模型来描述数据，而是直接对数据进行转换，如采样、聚类和直方图统计等。

回归分析是一种典型的参数式方法，回归分析的一般表达式如下：

$$Y = F(X;\beta) + E$$

其中，F 为模型的表达式，X 为自变量，Y 为因变量，β 为模型的未知参数，E 为误差，X、Y、β、E 都可以是标量或矢量。回归分析的目的就是在一定条件下估计最好的参数 β。根据不同的应用问题和估计方法，通常对误差 E 有不同的假设。例如，在信号处理中经常会假设误差是高斯白噪声，各分量服从正态分布 $N(0,\sigma)$，且各分量彼此不相关，在最大似然准则下，估计 β 的问题就变成了：

$$\beta = \underset{\beta}{\operatorname{argmax}} p(Y;\beta,X)$$

$$\beta = \underset{\beta}{\operatorname{argmax}} \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \prod_i \exp\left(-\frac{1}{2}\left(\frac{y_i - f_i(X;\beta)}{\sigma}\right)^2\right)$$

$$\beta = \underset{\beta}{\operatorname{argmin}} \sum_i (y_i - f_i(X;\beta))^2$$

即最小化误差的平方之和。最后的表达式称为最小二乘方法，高斯在 1795 年就曾经使用该方法研究行星运动。最小二乘方法最大的特点在于不对误差的概率分布做假设，因此可以广泛地适用于各种回归模型。

回归分析中一类最简单的模型是线性模型，使用线性模型进行的回归分析称为线性回归。线性回归的模型如下：

$$Y = \beta^{\mathrm{T}} X + \beta_0$$

一元线性回归的模型为：

$$y = \beta_1 x + \beta_0$$

一元线性回归在平面上表现为一条直线，图 3.14 是线性回归的一个示例。

直方图是一种对数据的可视化描述方法，它将数据分箱后统计每个箱中数据的计数(总和/平均)，分箱方法有等宽分箱与等频分箱等。使用直方图来对数据进行归约后，仅仅存储每个箱中数据的计数，图 3.6 是直方图的一个示例。

聚类是根据数据相似性将数据聚成簇的方法，聚类分析在前文中已经介绍过，会在后面进行更详细的介绍。使用聚类进行数据归约后，仅仅存储聚类中心、类半径等数据。图 3.13 是聚类的一个示例。

采样方法是仅仅抽取数据的一个子集来代表数据的方法，它直接从数量上对数据量进行消减。最简单的采样方法是简单随机采样（SRS），即随机地从所有 N 个数据中抽取 M 个数据。简单随机采样分为有放回的简单随机采样（SRSWR）和无放回的简单随机采样（SRSWOR），两者的差别在于从总体数据中拿出一个数据后，是否将这个数据放回，图 3.17 是这两种简单随机采样的一个示例。有放回的简单随机采样得到一份样本的概率为：

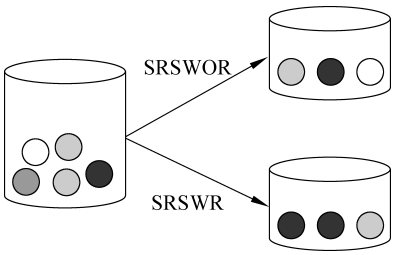


图 3.17 简单随机采样

$$P=\frac{1}{N^M}$$

而无放回的简单随机采样得到一份样本的概率则为：

$$P=\frac{1}{N}\times\frac{1}{N-1}\times\cdots\times\frac{1}{N-M+1}=\frac{(N-M)!}{N!}$$

在数据分布非常不对称的情况下，简单随机采样可能产生非常差的效果，这时可以使用更具适应性的采样方法如分层采样将数据分层，并根据每层的比例在该层中采样。

3.6.2 数据离散化

计算机存储器无法存储无限精度的值，计算机处理器也不能对无限精度的数进行处理，因此在数据预处理中需要进行数据的离散化。另外，某些数据挖掘方法需要离散值的属性，这也催生了对数据进行离散化的需要。

数据离散化是对数据的属性值进行的预处理，它将属性值划分为有限个部分，之后使用这个部分的标签来代替原来的属性值。实际上，所有采集到存储器中的数据都已经经过了离散化，这里提到的数据离散化是指显式地对数据的属性值划分部分并将属性的值替换为所属部分的标签。数据离散化的方法主要有分箱、聚类、自顶向下拆分、自底向上合并等。

使用分箱的数据离散化方法是通过先将属性值分箱，再将属性值替换为箱标签的离散化方法；使用聚类的数据离散化方法是先将属性值聚类，再使用类标签作为新的属性值的离散化方法。分箱和聚类在本章都介绍过，这里不再赘述。下面介绍三种通过拆分和合并来进行数据离散化的方法：基于信息增益的离散化、基于卡方检验的离散化和基于自然分区的离散化。

1. 基于信息增益的离散化

在进行数据离散化的过程中，如果关注点主要在于属性值的离散化能够有助于提高分类的准确性，那么可以使用信息增益来进行数据离散化。这种离散化方法是一种自顶向下的拆分方法，从属性值的整体 S 开始，使用一个边界 T 来对属性值进行划分，如果属性值划分成了 n 个分区，分别为 $S_0,\cdots,S_{n-1}$，那么 S 被 T 划分的信息增益就是：