第3章 数据预处理

数据预处理是数据挖掘中的重要一环，而且必不可少。要想更有效地挖掘出知识，就必须为它提供干净、准确、简洁的数据。然而，在实际应用系统中收集到的原始数据往往是“脏”的。

现实世界中的数据大都是不完整、不一致的脏数据，无法直接进行数据挖掘，或者挖掘结果无法令人满意。为了提高数据挖掘的质量就需要使用数据预处理技术。数据预处理有多种方法，如数据清理、数据集成、数据变换和数据归约等。这些数据预处理技术在数据挖掘之前使用可以大大提高数据挖掘模式的质量，降低实际挖掘所需要的时间。

数据清理(Data Cleaning)过程是通过填写缺失的值、光滑噪声数据、识别或删除离群点以及解决不一致性等手段来“清理”数据，主要达到如下目标：格式标准化、异常数据清除、错误纠正等。

数据集成(Data Integration)是将多个数据源中的数据结合起来并统一存储，建立数据仓库的过程。

数据变换(Data Transformation)是通过平滑聚集、数据概化、规范化等方式将数据转换成适用于数据挖掘的形式。

数据归约(Data Reduction)，数据挖掘时往往数据量非常大，进行挖掘分析需要很长的时间，数据归约技术可以得到数据集的归约表示，它比原数据集小很多，但基本可以保持原数据的完整性，对归约后的数据集进行挖掘的结果与对原数据集进行挖掘的结果相同或几乎相同。

3.1 数据预处理及任务

本节主要介绍数据预处理的必要性以及数据预处理的主要任务。

3.1.1 数据预处理的必要性

1. 原始数据存在的问题

数据挖掘使用的数据常常来源于不同的数据源，且不同数据源的用途不同。因此，数据挖掘常常不能在数据来源处控制数据质量。由于无法避免数据质量问题，因此数据挖掘对数据质量问题的控制着眼于两个方面：①数据质量的检测和纠正；②使用可以容忍低质量数据的算法。数据质量的检测和纠正，通常称为数据清理，重点关注的是测量和数据收集方面的数据质量问题，主要是测量误差和数据收集错误。

(1) 测量误差

测量误差(Measurement Error)是指测量过程中导致的问题。例如,测量记录的值与实际值不同。对于连续属性,测量值与实际值的差称为误差(Error)。测量误差的数据问题通常包括噪声、伪像、偏倚、精度和准确率。

① **噪声**(Noise):是测量误差的随机部分,收集数据的时候难以得到精确的数据。例如,收集数据的设备可能出现故障、数据输入时可能出现错误、数据传输过程中可能出现错误、存储介质可能出现损坏等,这些情况都可能导致噪声数据的出现。

处理数据时常常使用的噪声检测技术,包括基于统计的技术和基于距离的技术。即使可以检测出噪声,但要完全消除噪声也是困难的,因此许多数据挖掘工作都关注设计鲁棒算法(Robust Algorithm),即在噪声干扰下也能产生可以接受的结果。

② 伪像:数据错误可能是更确定性现象的结果如一组照片在同一个地方出现条纹。数据的这种确定性失真称为**伪像**(Artifact)。

③ 精度、偏倚和准确率:在统计学和科学实验中,测量过程和结果数据的质量用精度和偏倚度量。假定对相同的基本量进行重复测量,并且用测量值集合计算平均值来作为实际值的估计值。

精度(Precision)是对同一个量的重复测量值之间的接近程度。

偏倚(Bias)是测量值与被测量之间的系统变差。

精度通常用值集合的标准差度量,而偏倚用值集合的均值与测出的已知值之间的差度量。只有那些通过外部手段能够得到测量值的对象,其偏倚才是可确定的。

假定有 1g 标准试验重量,如果称重 5 次,得到下列值:{1.015,0.990,1.013,1.001,0.986}。这些值的均值为 1.001,因此偏倚是 0.001。用标准差度量,精度是 0.012。

准确率(Accuracy)是指被测量的测量值与实际值之间的接近度。准确率通常是更一般的表示数据测量误差程度的术语。

准确率依赖于精度和偏倚。准确率的一个重要方面是有效数字(Significant Digit)的使用。其目标是仅使用数据精度所能确定的数字位数表示的测量或计算结果。例如,对象的长度用最小刻度为毫米的米尺测量,则只能记录最接近毫米的长度数据,这种测量的精度为±0.5mm。

(2) 数据收集错误

数据收集错误(Data collection error)是指诸如遗漏数据对象或属性值,或者不当地包含了其他数据对象等错误。例如,一种特定类动物研究可能包含了相关种类的其他动物,它们只是表面上与要研究的种类相似。测量误差和数据收集错误可能是系统的,也可能是随机的。

同时涉及的测量和数据收集的数据质量问题包括:离群点、缺失值和不一致的值、重复数据。

① **离群点**(Outlier):在某种意义上,离群点是具有不同于数据集中其他大部分数据对象特征的数据对象,或者是相对于该属性的典型值来说不寻常的属性值。离群点也称为异常对象或异常值。有许多定义离群点的方法,并且统计学和数据挖掘界已经提出了很多不同的定义。此外,区别噪声和离群点这两个概念是非常重要的。离群点可以是合法的数据对象或值。因此,不像噪声,离群点本身有时是人们感兴趣的对象。例如,在欺诈和网络攻

击检测中,目标就是在大量正常对象或事件中发现不正常的对象和事件。本书第9章将详细讨论离群点检测。

② **缺失值**:一个对象缺失一个或多个属性值的情况并不少见,由于实际系统设计时可能存在的缺陷以及使用过程中人为因素所造成的影响,数据记录中可能会出现有些数据属性的值丢失或不确定的情况,还可能缺少必需的数据而造成数据不完整。例如,有的人拒绝透露年龄或体重。再如,收集数据的设备出现了故障,导致一部分数据的缺失,这就会使数据不完整。另外,实际使用的系统中可能存在大量的模糊信息,有些数据甚至还具有一定的随机性质。无论何种情况,在数据分析时都应当考虑缺失值。

③ **不一致的值**:原始数据是从各种实际应用系统(多种数据库、多种文件系统)中获取的,由于各应用系统的数据缺乏统一的标准和定义,数据结构也有较大的差异,因此各系统间的数据存在较大的不一致性,共享问题严重,往往不能直接拿来使用。例如,某数据库中两个不同的表可能都有重量这个属性,但是一个以kg为单位,一个是以g为单位,这样的数据就会有较大的杂乱性。再如,地址字段列出了邮政编码和城市名,但是有的邮政编码区域并不包含在对应的城市中。可能是人工输入该信息时录颠倒了两个数字,或许是在手写体扫描时读错了一个数字。不管导致不一致数据的原因是什么,重要的是能检测出来,并且如果可能的话还要纠正这种错误。

④ **重复数据**:数据可能包含重复或几乎重复的数据对象。许多人都收到过重复的邮件,因为它们以稍微不相同的名字多次出现在数据库中。为了检测并删除这种重复,必须处理两个主要问题。首先,两个对象实际代表同一个对象,但对应的属性值不同,必须解决这些不一致的值;其次,需要避免意外地将两个相似但并非重复的数据对象(如两个人具有相同姓名)合并在一起。去重复(Reduplication)通常用来表示处理这些问题的过程。

现实世界中收集到的原始数据存在较多的问题是:数据的不一致、噪声数据以及缺失值。

例3.1 收集的数据可能出现的问题。

假设某公司的领导想要分析某个月的销售数据。首先需要选择分析需要的属性,例如商品价格、商品ID等。如果人工录入时有输入错误,就会降低数据的准确性。再如,公司领导希望知道每种销售商品是否做过降价销售广告,但是这些信息可能是缺失的,这样就无法保证数据的完整性;存放用户具体信息的表中某用户的手机号为13110345615,但是购买记录表中的手机号被存为13110345610,这样就无法保证数据的一致性。

2. 数据质量要求

现实世界中的数据大都存在数据不一致、噪声数据以及缺失值等问题,但是数据挖掘需要的都必须是高质量的数据,即数据挖掘所处理的数据必须具有准确性(Correctness)、完整性(Completeness)、一致性(Consistency)等性质。另外,时效性(Timeliness)、可信性(Believability)和可解释性(Interpretability)也会影响数据的质量。

(1) 准确性

准确性是指数据记录的信息是否存在异常或错误。

(2) 完整性

完整性是指数据信息是否存在缺失的情况。数据缺失的情况可能是整个数据记录缺失,也可能是数据中某个字段信息的记录缺失。

(3) 一致性

一致性是指数据是否遵循了统一的规范,数据集合是否保持了统一的格式。数据质量的一致性主要体现在数据记录的规范和数据是否符合逻辑。

(4) 时效性

时效性是指某些数据是否能及时更新。更新时间越短,则时效性越强。

(5) 可信性

可信性是指用户信赖的数据的数量。用户信赖的数据越多,则可信性越好。

(6) 可解释性

可解释性是指数据自身是否易于人们理解。数据自身越容易被人们理解,则可解释性越高。

3.1.2　数据预处理的主要任务

数据预处理主要包括数据清理、数据集成、数据归约和数据变换。数据预处理的主要任务如图 3-1 所示。

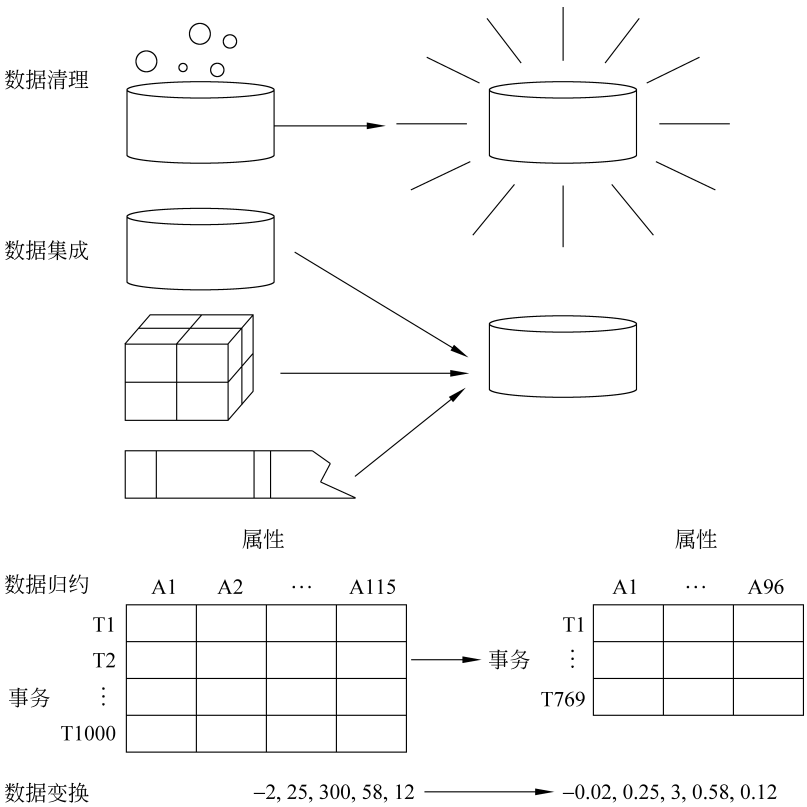


图 3-1　数据预处理的主要任务

1. 数据清理

数据清理通过填写缺失的值、光滑噪声数据、识别或删除离群点等方法去除源数据中的

噪声数据和无关数据,并且处理遗漏的数据和清洗“脏”数据,考虑时间顺序和数据变化等。数据清理主要是针对缺失值的数据处理.并完成数据类型的转换。

2. 数据集成

当需要分析挖掘的数据来自多个数据源的时候,就需要集成多个数据库、数据立方体或文件,即**数据集成**。来自多个不同数据源的数据,可能存在数据的不一致性和冗余问题:代表同一概念的属性的属性名在不同数据库中可能不同,例如在某个数据库中的商品名称的属性名为 product_name,它在另一个数据库中却是 brand_name。数据的不一致还可能出现在属性值中,例如同一个商品在第一个数据库中的商品名取值为“sofa”,在另一个数据库中值为“couch”,在第三个数据库可能还会有其他值。除此之外,还有某些属性是由其他属性导出的。

3. 数据归约

数据归约是指对数据集进行简化表示。大量的冗余数据会降低知识发现过程的性能或使之陷入混乱。因此,在数据预处理中不仅要进行数据清理,还必须采取措施避免数据集成后数据的冗余。这样既能降低数据集的规模,又可以不损害数据挖掘的结果。数据归约后,比原来小得多,但是可以得到几乎相同的分析结果。

4. 数据变换

数据变换是将数据从一种表示形式变成另一种表现形式的过程,它包括了数据的规范化、数据的离散化和概念分层,可以使数据的挖掘在多个抽象层上进行。

现实世界中的数据需要使用数据预处理提高数据的质量,这样可以提高挖掘过程的准确率和效率。因此,数据预处理是数据挖掘的重要步骤。

3.2 数据清理

现实世界中的大多数数据是不完整、有噪声和不一致的。那么就需要对“脏”数据进行数据清理。数据清理就是对数据进行重新审查和校验的过程,其目的是纠正存在的错误,并提供数据一致性。

3.2.1 缺失值、噪声和不一致数据的处理

1. 缺失值的处理

缺失值是指在现有的数据集中缺少某些信息。也就是说,某个或某些属性的值是不完全的。处理缺失值一般使用以下几种方式。

(1) 忽略元组

在数据中缺少类标号的情况下经常采用忽略元组这种方法(假定挖掘任务涉及分类)。但是,除非元组有多个属性缺失值,否则该方法就没有什么效果。当每个属性缺失值的百分比变化很大时,它的性能会特别差。

(2) 忽略属性列

如果某个属性的缺失值太多,假设超过了 80%,那么在整个数据集中就可以忽略该属性。

(3) 人工填写缺失值

一般来说,人工填写缺失值会耗费过多的人力和物力,而且如果数据集缺失了很多值或者数据集很大,该方法不方便实现。

(4) 使用属性的中心度量值填充缺失值

如果数据的分布是正常的,就可以使用均值来填充缺失值。例如,一条属于 a 类的记录在 A 属性上存在缺失值,那么可以用该属性上属于 a 类全部记录的平均值代替该缺失值。例如,对于顾客一次来超市时所消费的金额这一字段,就可以按照顾客的年龄这一字段进行分类,然后使用处于相同年龄段的顾客的平均消费金额填充缺失值。

如果数据的分布是倾斜的,则可以使用中位数来填充缺失值。

(5) 使用一个全局常量填充空缺值

使用一个全局常量填充空缺值就是对一个所有属性的所有缺失值都使用一个固定的值填补(如 Not sure 或∞)。此方法最大的优点就是简单、省事,但是也可能产生一个问题,挖掘的程序可能会误认为这是一个特殊的概念。

(6) 使用与给定元组同一类的所有样本的属性均值或中位数

该方法经常用于分类挖掘任务。例如,在对商场顾客按信用风险(credit_risk)进行分类挖掘时,可以用在同一信用风险类别下(如良好)的 income 属性的平均值,来填补所有在同一信用风险类别下属性 income 的遗漏值。

(7) 使用可能的特征值替换缺失值

以上这些简单方法的替代值都不准确,数据都有可能产生误差。为了比较准确地预测缺失值,数据挖掘者可以生成一个预测模型预测每个丢失值。例如,如果每个样本给定 3 个特征值 A、B、C,那么可以将这 3 个值作为一个训练集的样本,生成一个特征之间的关系模型。一旦有了训练好的模型,就可以提出一个包含丢失值的新样本,并产生预测值。也就是说,如果特征 A 和 B 的值已经给出,模型会生成特征 C 的值。如果丢失值与其他已知特征高度相关,这样的处理就可以为特征生成最合适的值。

当然,如果缺失值总是能够被准确地预测,就意味着这个特征在数据集中是冗余的,在进一步的数据挖掘中是不必要的。在现实世界的应用中,缺失值的特征和其他特征之间的关联应该是不完全的。所以,不是所有的自动方法都能填充出正确的缺失值。但此方法在数据挖掘中是很受欢迎的,因为它可以最大限度地使用当前数据的信息预测缺失值。

2. 噪声的处理

噪声是指被测量的变量产生的随机错误或误差。

噪声是随着随机误差出现的,包含错误点值或孤立点值。噪声数据产生的主要原因是数据输入数据库产生的纰漏及设备可能的故障。噪声检测可以降低根据大量数据做出错误决策的风险,并有助于识别、防止、去除恶意或错误行为的影响。

发现噪声数据并且从数据集中去除它们的过程可以描述为从 n 个样本中选 k 个与其余数据显著不同或例外的样本($k<<n$)。定义噪声数据的问题是非同寻常的,在多维样本中

尤其如此。常用的噪音检测的技术如下。

(1) 基于统计的技术

基于统计的噪声探测方法可以分为一元方法和多元方法,目前多数研究团体通常采用多元方法,但是这种方法并不适合高维数据集和数据分布未知的任意数据集。

多元噪声探测的统计方法常常能指出远离数据分布中心的样本。这个任务可以使用几个距离度量值完成。马哈拉诺比斯(Mahalanobis)距离(简称马氏距离)值包括内部属性之间的依赖关系,这样系统就可以比较属性组合。这个方法依赖多元分布的估计参数,给定 p 维数据集中的 n 个观察值 x_i(其中 $n>>p$),用 $\bar{\boldsymbol{x}}_n$ 表示样本平均向量,$\boldsymbol{V}_n$ 表示样本协方差矩阵,则有

$$\boldsymbol{V}_n=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}}_n)(x_i-\bar{\boldsymbol{x}}_n)^{\mathrm{T}} \tag{3-1}$$

每个多元数据点 $i(i=1,2,\cdots,n)$ 的马哈拉诺比斯距离 M_i 为

$$M_i=\Big[\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}}_n)^{\mathrm{T}}\boldsymbol{V}_n^{-1}(x_i-\bar{\boldsymbol{x}}_n)\Big]^{\frac{1}{2}} \tag{3-2}$$

于是,马氏距离很大的 n 个样本就被视为噪声数据。

(2) 基于距离的技术

基于距离的噪声检测方法与基于统计的方法最大的不同是:基于距离的噪声检测方法可以用于多维样本;而大多数的基于统计的方法仅分析一维样本,即使分析多维样本,也是单独分析每一维。这种基于距离的噪声检测方法的基本计算复杂性,在于估计 n 维数据集中所有样本间的测量距离。如果样本 S 中至少有一部分数量为 p 的样本到 s_i 的距离比 d 大,那么样本 s_i 就是数据集 S 中的一个噪声数据。也就是说,这种方法的检测标准基于参数 p 和 d,这两个参数可以根据数据的相关知识提前给出或者在迭代过程中改变,以选择最有代表性的噪声数据。

例 3.2 基于距离的噪声检测方法。

给定一组三维样本 S,$S=\{S_1,S_2,S_3,S_4,S_5,S_6\}=\{(1,2,0),(3,1,4),(2,1,5),(0,1,6),(2,4,3),(4,4,2)\}$,求在距离阈值 $d\geqslant 4$、非邻点样本的阈值部分 $p\geqslant 3$ 时的噪声数据。

解:首先,求数据集中样本的欧几里得距离,使用 $d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$,如表 3-1 所示。

表 3-1 数据集 S 的距离表

数据集	S_1	S_2	S_3	S_4	S_5	S_6
S_1	—	4.583	5.196	6.164	3.742	4.123
S_2	—	—	1.414	3.606	3.317	3.742
S_3	—	—	—	2.236	3.606	4.690
S_4	—	—	—	—	4.690	6.403
S_5	—	—	—	—	—	2.236

然后,再根据阈值距离 $d=4$ 计算出每个样本的 p 值,即距离大于或等于 d 的样本数量,计算结果如表 3-2 所示。

表 3-2　S 中每个点的距离大于或等于 d 的 p 值表

样本	p	样本	p
S_1	4	S_4	3
S_2	1	S_5	1
S_3	2	S_6	3

根据表 3-2 所示的结果，可选择 S_1、S_4、S_6 作为噪声数据（因为它们的 $p \geqslant 3$）。

3. 不一致数据的处理

数据的不一致性，是指各类数据的矛盾性和不相容性，主要是由于数据冗余、并发控制不当以及各种故障和错误造成的。由于存在很多破坏数据一致性的因素，数据库系统都会有一些相应的措施解决并保持数据库的一致性，因此可以使用数据库系统来保持数据的一致性。

但是对于某些事务中一些数据记录的不一致，可以使用其他比较权威的材料改正这些事务的数据不一致。另外，数据输入时产生的问题可以用纸上的记录改正这些数据的不一致。知识工程工具也可以用来检测违反约束条件的数据。

3.2.2　数据清理方式

噪声和缺失值都会产生“脏”数据，也就是有很多原因会使数据产生错误，在进行数据清理时，就需要对数据进行偏差检测。导致偏差的原因有很多，例如，人工输入数据时有可能误输入；数据库的字段设计自身可能产生一些问题；用户填写信息时有可能没有填写真实信息以及数据退化等。不一致的数据表示和编码的不一致使用也可能出现数据偏差，例如身高 170cm 和 1.70m，日期“2011/12/12”和“12/12/2011”。字段过载（Field Overloading）产生的原因一般是开发者将新属性的定义挤进已经定义的属性的未使用（位）的部分，例如，使用一个属性未使用的位，该属性取值已经使用了 32 位中的 31 位。

可以使用唯一性原则、连续性原则和空值原则观察数据，进行偏差检测。

（1）唯一性原则

每个值都是唯一的，一个属性的每一个值都不能和这个属性的其他值相同。

（2）连续性原则

首先要满足唯一性原则，然后每个属性的最大值和最小值之间没有缺失的值。

（3）空值原则

需要明确空白、问号、特殊符号等指示空值条件的其他串的使用，并且知道如何处理这样的值。

此外，为了统一数据格式和解决数据冲突，在数据清理时还可以使用外部源文件更正错误数据。外部源文件就是以记录的形式表示信息的文件，这些外部源文件可以从一些拥有单位或个人完整并真实的有效信息的行政部门获得，如例 3.3 所示。

例 3.3　使用外部源文件更正错误数据。

在表 3-3 所示的外部源文件中，ID 是唯一的，是关键字段。表 3-4 是一条脏记录。外部

源文件模式与脏数据的模式一致,根据外部源文件的关键字段确定脏数据中字段的格式。清理过后的结果如表3-5所示,对表中Name字段的值重新进行了调整。

表3-3 外部源文件实例

ID	Name	Address	Sex
20161009211	Zhang San	12	M
20161009212	Li Si	30	M
20161009213	Wang Wu	25	F

表3-4 一条脏记录

ID	Name	Address	Sex
20161009211	Zhang S	12	M

表3-5 清理后的记录

ID	Name	Address	Sex
20161009211	Zhang San	12	M

3.3 数据集成

数据集成主要是在数据分析任务中把不同来源、格式、特点和性质的数据合理地集中并合并起来,从而为数据挖掘提供完整的数据源(包括多个数据库、数据立方体或一般文件),然后存放在一个一致的数据存储中,这样有助于减少结果数据集的冗余和不一致,提高在这之后的挖掘过程的准确性和速度。

数据集成的过程涉及的两个问题是实体识别问题和冗余问题。

1. 实体识别问题

这个问题主要是来自多个信息源的现实世界产生的"匹配"问题。例如,一个数据库中的brand_name和另一个数据库的product_name指的是同一实体。通常,数据库和数据仓库中的元数据(关于数据的数据)可以帮助避免模式集成中的错误。

2. 冗余问题

在进行数据集成的过程中很可能会遇到冗余。某些冗余可以通过相关性分析检测出来,主要分两种情况:一种是对数值属性数据(即数值数据),使用相关系数和协方差;另一种是对标称数据,使用χ^2(卡方)检验。

(1) 数值数据的相关系数(Correlation Coefficient)

属性X和Y的相关度使用其**相关系数**$r_{X,Y}$来表示。

$$r_{X,Y}=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}(x_iy_i)-n\overline{X}\overline{Y}}{n\sigma_X\sigma_Y} \tag{3-3}$$

式(3-3) 中的 n 代表元组的个数，x_i 是元组 i 在属性 X 上的值，y_i 是元组 i 在属性 Y 上的值，$\overline{X}$ 表示 X 的均值，$\overline{Y}$ 表示 Y 的均值，σ_X 表示 X 的标准差，σ_Y 表示 Y 的标准差，$\sum_{i=1}^{n}(x_i,y_i)$ 表示每个元组中 X 的值乘 Y 的值。且 $r_{X,Y}$ 的取值范围为 $-1\leqslant r_{X,Y}\leqslant 1$。

如果 $r_{X,Y}>0$，则 X 和 Y 是正相关的，也就是说，X 值随 Y 值的变大而变大。如果 $r_{X,Y}$ 的值较大，数据可以作为冗余而被删除。

如果 $r_{X,Y}=0$，则 X 和 Y 是独立的且互不相关。

如果 $r_{X,Y}<0$，则 X 和 Y 是负相关的，也就是说，X 值随 Y 值的减小而变大，即一个字段随着另一个字段的减少而增多。

例 3.4　相关系数的计算。

已知体重与血压的 12 个样本数据如表 3-6 所示，试判断其相关性。

表 3-6　体重与血压表

指标	样本号											
	1	2	3	4	5	6	7	8	9	10	11	12
体重	68	48	56	60	83	56	62	59	77	58	75	64
血压	95	98	87	96	110	155	135	128	113	168	120	115

解：

① 由表 3-6 可计算体重 X 和血压 Y 的均值和标准差。

$$\overline{X}=\frac{68+48+56+60+83+56+62+59+77+58+75+64}{12}=63.83$$

$$\overline{Y}=\frac{95+98+87+96+110+155+135+128+113+168+120+115}{12}=118.33$$

$$\sigma_X=\sqrt[2]{\frac{1}{12}(68^2+48^2+56^2+60^2+83^2+56^2+62^2+59^2+77^2+58^2+75^2+64^2)-63.83^2}$$

$$=10.14$$

$$\sigma_Y=\sqrt[2]{\frac{1}{12}(95^2+98^2+87^2+96^2+110^2+155^2+135^2+128^2+113^2+168^2+120^2+115^2)-118.33^2}$$

$$=24.74$$

② 计算相关系数 $r_{X,Y}$。

$$r_{X,Y}=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{12}(x_i-63.83)(y_i-118.33)}{12\times 10.14\times 24.74}=-0.112$$

由于 $r_{X,Y}<0$，可知 X 和 Y 是负相关的。

但是，相关性并不代表因果关系。假设 X 和 Y 具有相关性，不能代表 X 导致 Y 或者 Y 导致 X。例如，在超市售卖货物的时候，会发现卖出的商品与货物的摆放位置是相关的，但是这并不意味着卖出的商品与商品的摆放位置是有因果关系的。

(2) 数值数据的协方差

在概率论和统计学中，**协方差**(Covariance)用于衡量两个变量的总体误差。而方差是

协方差中两个变量相同的一种特殊情况。协方差也可以评估两个变量的相互关系。

期望值就是指在一个离散性随机变量试验中每次可能结果的概率乘以其结果的总和。

设有两个属性 X 和 Y,以及有 n 次观测值的集合$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,则 X 的期望值(均值)为

$$E(X)=\overline{X}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{3-4}$$

Y 的期望值(均值)为

$$E(Y)=\overline{Y}=\frac{\sum_{i=1}^{n}y_i}{n} \tag{3-5}$$

X 和 Y 的**协方差**定义为

$$\mathrm{Cov}(X,Y)=E[(X-\overline{X})(Y-\overline{Y})]=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n} \tag{3-6}$$

将式(3-3) 与式(3-6)结合,得到

$$r_{X,Y}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \tag{3-7}$$

其中,σ_X 和 σ_Y 分别是 X 和 Y 的标准差。

还可以证明

$$\mathrm{Cov}(X,Y)=E(X\cdot Y)-\overline{X}\overline{Y} \tag{3-8}$$

当 $\mathrm{Cov}(X,Y)>0$ 时,表明 X 与 Y 正相关;当 $\mathrm{Cov}(X,Y)<0$ 时,表明 X 与 Y 负相关;当 $\mathrm{Cov}(X,Y)=0$ 时,表明 X 与 Y 不相关。

若属性 X 和 Y 是相互独立的,有

$$E(X\cdot Y)=E(X)\cdot E(Y) \tag{3-9}$$

则协方差的公式是

$$\mathrm{Cov}(X,Y)=E(X\cdot Y)-\overline{X}\,\overline{Y}=E(X)\cdot E(Y)-\overline{X}\,\overline{Y}=0$$

但是,它的逆命题是不成立的。

例 3.5 协方差的计算。

依据表 3-6 体重与血压表中的数据,求血压是否会随着体重一起变化。

解:

① 计算期望值或标准差,利用例 3.4 计算结果,如表 3-7 所示。

表 3-7 体重和血压的均值和标准差值

指标	均值	标准差
体重	63.83	10.14
血压	118.33	24.74

② 计算协方差。

$$\mathrm{Cov}(X,Y)=r_{X,Y}\cdot\sigma_X\cdot\sigma_Y=-0.112\times10.14\times24.74=-28.10$$

因为协方差为负，所以血压和体重呈负相关。

(3) 标称数据的χ^2 检验

对于标称数据，两个属性 X 和 Y 之间的相关联系可以通过χ^2(卡方)检验发现。假设 X 有 n 个不同值，分别为 $x_1, x_2, \cdots, x_n$；Y 有 r 个不同值，分别为 $y_1, y_2, \cdots, y_r$。使用列联表表示 X 和 Y 的数据，如表 3-8 所示。

表 3-8　列联表

X \ Y	y_1	y_2	…	y_j	…	y_r	sum
x_1	o_{11}	o_{12}	…	o_{1j}	…	o_{1r}	$O_{1.}$
x_2	o_{21}	o_{22}	…	o_{2j}	…	o_{2r}	$O_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	o_{i1}	o_{i2}	…	o_{ij}	…	o_{ir}	$O_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n	o_{n1}	o_{n2}	…	o_{nj}	…	o_{nr}	$O_{n.}$
sum	$O_{.1}$	$O_{.2}$	…	$O_{.j}$	…	$O_{.r}$	m

列联表是用 X 的 n 个值作为列联表的行，用 Y 的 r 个值作为列联表的列。使用(x_i, y_j)表示一个联合事件：字段 X 的值为 x_i，字段 Y 的值为 y_j，即($X=x_i$, $Y=y_j$)，每个单元 o_{ij} 都是(x_i, y_j)的联合事件。

χ^2 值又称 Pearson χ^2 统计量，其计算为

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \tag{3-10}$$

式(3-10)中的 o_{ij} 是联合事件(x_i, y_j)的观测频度(即实际计数)，而 e_{ij} 是(x_i, y_j)的期望频度。其中，e_{ij} 的计算为

$$e_{ij} = \frac{\text{count}(X=x_i)\times\text{count}(Y=y_j)}{m} = \frac{(O_{i.}\times O_{.j})}{m} \tag{3-11}$$

式(3-11)中的 m 是数据元组的个数，count($X=x_i$)是 X 上值为 x_i 的元组个数，而 count($Y=y_j$)是 Y 上值为 y_j 的元组个数。特别注意，对χ^2 值贡献最大的单元是其实际计数与期望计数极不相同的单元。

χ^2 相关检验假设的 X 和 Y 是独立的，检验基于显著水平 α，具有自由度$(r-1)\times(n-1)$。可以使用χ^2 检验两个属性是否独立。

独立性检验的步骤如下。

① 统计假设：

设 H_0：属性 X 和属性 Y 之间是独立的；则 H_1：属性 X 和属性 Y 之间是相关的。

② 计算期望频数：

$$e_{ij} = \frac{(O_{i.}\times O_{.j})}{m}$$

③ 确定自由度：

$$\text{df} = (r-1)\times(n-1)$$

④ 计算 Pearson χ^2 统计量：

$$\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

⑤ 统计推断：

χ^2>临界值(具有自由度 df 和显著水平 α)：拒绝假设 H_0。

χ^2<临界值(具有自由度 df 和显著水平 α)：接受假设 H_0。

例 3.6 使用χ^2 的标称数据的相关分析。

对从事两种工种的某一年龄段男性患某种疾病的情况进行调查，如表 3-9 所示。分析某一年龄段男性患某种疾病与从事工种是否相关。

表 3-9 患病情况调查列联表

从事工种	患病	不患病	合计
工种 1	386	895	1281
工种 2	65	322	387
合计	451	1217	1668

解：

① 统计假设：H_0 为某一年龄段男性患某种疾病与从事工种不相关。

② 计算期望频数：根据式(3-11) 计算期望频度。

$$e_{11}=1281\times451/1668=346.36$$
$$e_{12}=1281\times1217/1668=934.64$$
$$e_{21}=387\times451/1668=104.64$$
$$e_{22}=387\times1217/1668=282.36$$

③ 确定自由度 df：

$$df=(2-1)\times(2-1)=1$$

④ 计算卡方统计量：根据式(3-10)，计算卡方值。

$$\chi^2=\frac{(386-346.36)^2}{346.36}+\frac{(895-934.64)^2}{934.64}+\frac{(65-104.64)^2}{104.64}+\frac{(322-282.36)^2}{282.36}=26.80$$

⑤ 统计判断：假设取显著水平 $\alpha=0.05$，查询表 3-10 的卡方检验临界值表。

表 3-10 卡方检验临界值表(部分)

自由度	显著水平									
	0.99	0.98	0.95	0.90	0.50	0.10	**0.05**	0.02	0.01	0.005
1	0.000	0.001	0.004	0.016	0.045	2.71	**3.84**	5.41	6.46	10.83
2	0.020	0.040	0.103	0.211	1.36	4.61	5.99	7.82	9.21	13.82
3	0.115	0.185	0.352	0.584	2.366	6.25	7.82	9.84	11.34	16.27

此例中的显著水平 α 为 0.05，自由度为 1 的卡方检验临界值为 3.84，此例卡方值为 26.80，大于 3.84，因此拒绝假设 H_0，说明某一年龄段男性患某种疾病与从事工种是统计相

关的。

两个独立样本比较可以分为以下 3 种情况。

① 所有的期望频度 $e_{ij} \geqslant 5$ 并且总样本量 $m \geqslant 40$，用 Pearson 卡方进行检验。

② 如果期望频度 $e_{ij} < 5$ 但 $e_{ij} \geqslant 1$，并且 $m \geqslant 40$，用连续性校正的卡方进行检验。

$$x^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}} \tag{3-12}$$

③ 如果有期望频度 $e_{ij} < 1$ 或 $m < 40$，则用精确概率检验。

3.4　数据归约

数据归约是指在对挖掘任务和数据自身内容理解的基础上，通过删除列、删除行和减少列中值的数量，来删掉不必要的数据，以保留原始数据的特征，从而在尽可能保持数据原貌的前提下最大限度地精简数据量。

数据归约技术可以得到数据集的归约表示，虽然小，但仍大致保持原数据的完整性。在归约后的数据集上挖掘将更有效，并产生相同(或几乎相同)的分析结果。

数据归约的主要策略如下。

① 数量归约：通过直方图、聚类和数据立方体聚集等非参数方法，使用替代的、较小的数据表示形式替换原数据。

② 属性子集选择：检测并删除不相关、弱相关或冗余的属性。

③ 抽样：使用比数据小得多的随机样本表示大型的数据集。

④ 回归和对数线性模型：对数据建模，使之拟合到一条直线，主要用来近似给定的数据。

⑤ 维度归约：通过小波变换、主成分分析等特征变换方式减少特征数目。

3.4.1　直方图

直方图(Histogram)是一种常见的数据归约的形式。属性 X 的**直方图**将 X 的数据分布划分为不相交的子集或桶。通常情况下，子集或桶表示给定属性的一个连续区间。**单值桶**表示每个桶只代表单个属性值/频率对(单值桶对于存放那些高频率的离群点非常有效)。

划分桶和属性值的规则有以下两点。

① 等宽：在等宽直方图中，每个桶的宽度区间是一致的。例如，图 3-2 中的桶宽为 10。

② 等频(或等深)：在等频直方图中，每个桶的频率粗略地计为常数，即每个桶大致包含相同个数的邻近数据样本。

例 3.7　用直方图表示数据。

已知某人在不同时刻下所量的血压值为：95，98，87，96，110，155，135，128，113，168，120，115，110，155，135，128，113，158，87，96，110，98，87，94，80，93，89，95，99，101，111，123，128，113，158，128，113，168，87，96，110。

使用等宽直方图表示数据如图 3-2 所示，桶宽为 10。

如果需要继续压缩数据，可以使用桶表示某个属性的一个连续值域，如图 3-3 中的每个桶都代表不同的血压值的区间为 20。

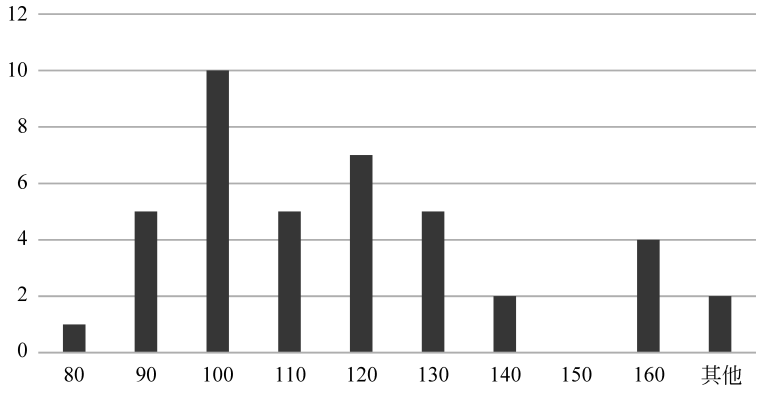


图 3-2　直方图(桶宽为 10)

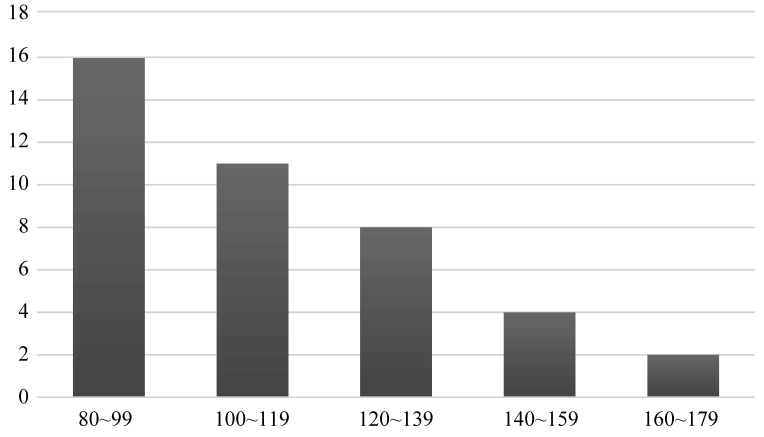


图 3-3　直方图(桶宽为 20)

3.4.2　数据立方体聚集

数据立方体是一类多维矩阵,可以使用户从多个维度探索和分析数据集,其中的数据是已经处理过的并且聚合成了立方体形式。数据立方体中的基本概念如下。

① 方体:不同层创建的数据立方体。

② 基本方体:最低抽象层创建的立方体。

③ 顶方体:最高层抽象的立方体。

④ 方体的格:每一个数据立方体。

例 3.8　某公司部分商品的销售数据立方体。

已知某公司的部分商品在不同城市前 4 个月(即 1～4 月份)每个月的销售情况,如图 3-4 所示。如果想要得到每种商品、每个地区、1～4 月份的销售总量,就可以对这些数据进行聚集。

图 3-4 是一个从商品、时间和城市 3 个维度表示的销售数据的立方体。其中,内部的每一个小立方体表示了某个城市、某个月份、销售某种商品的销售量。

右边缘上部分每个浅色的小立方体是对时间维度的汇总,如最右上边的立方体表示了

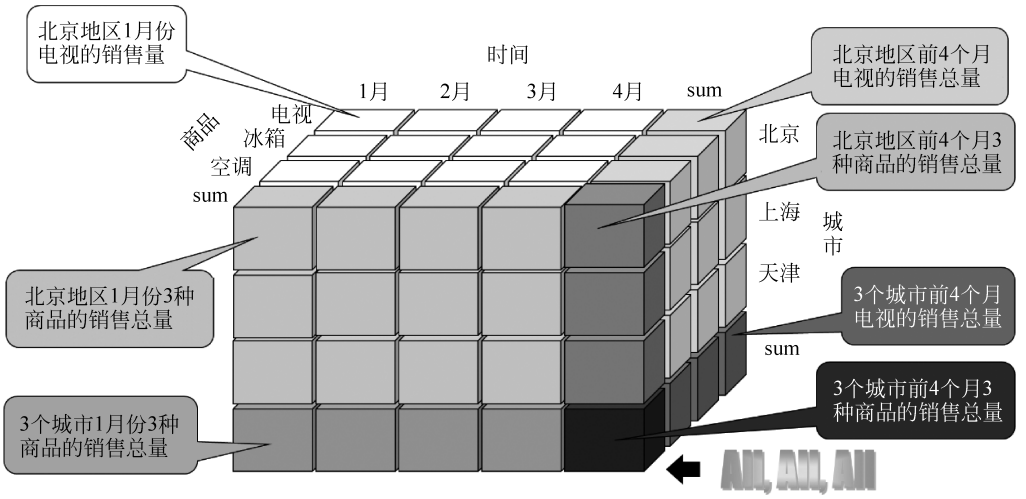


图 3-4　某公司部分商品 1～4 月份的销售数据立方体

北京地区前 4 个月电视的销售总量，它是对内部小立方体在时间维度的上一层抽象。前边缘上部分每个浅色的小立方体是对商品维度的汇总，如最左上边的立方体表示了北京地区 1 月份 3 种商品的销售总量，它是对内部小立方体在商品维度的上一层抽象。

右边缘下部分每个深色的小立方体是对时间和城市两个维度的汇总，如最右边的立方体表示了 3 个城市前 4 个月电视的销售总量，它是对上部分浅色立方体在城市维度的上一层抽象。前边缘下部分每个深色的立方体是对商品和城市两个维度的汇总，如最左边的立方体表示了 3 个城市 1 月份 3 种商品的销售总量，它是对上部分浅色立方体在城市维度的上一层抽象。

前边缘和右边缘中间的边缘上部分每个深色的立方体是对时间和商品两个维度的汇总，如最上边右侧的立方体表示了北京地区前 4 个月 3 种商品的销售总量，它是对左侧浅色立方体或右侧浅色立方体的上一层抽象。

最右下角的黑色立方体是对时间、商品和城市 3 个维度的汇总，即 3 个城市前 4 个月份 3 种商品的销售总量，它是最高层的抽象。

数据立方体的抽象层次可以用立方体的格表示，图 3-5 给出了某公司部分商品销售情况的数据抽象层次。其中的顶点表示了图 3-4 中的每个小立方体的层次概念，立方体最低层是最详细的数据，称为基本立方体。最上面一层即最高层是所有维度的汇总，称为顶立方体。中间层是对不同维度的抽象。

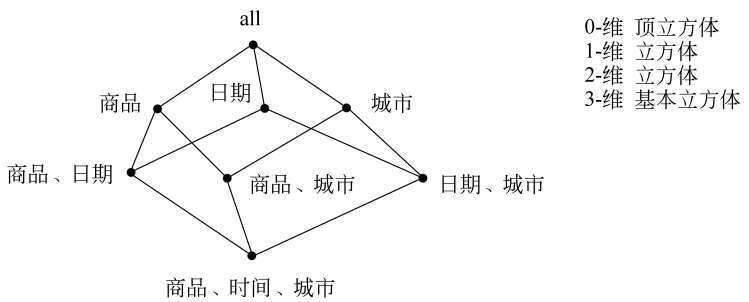


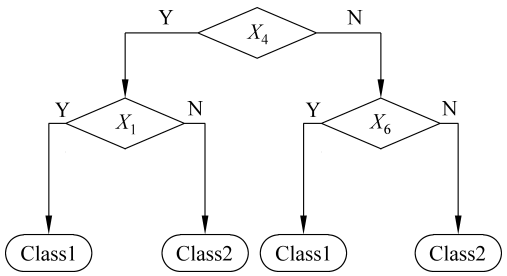
图 3-5　某公司部分商品销售情况的数据抽象层次

3.4.3 属性子集选择

属性子集选择是从一组已知属性集合中通过删除不相关或冗余的属性(或维度)来减少数据量。属性子集选择主要是为了找出最小属性集,使所选的最小属性集可以像原来的全部属性集一样能正确区分数据集中的每个数据对象。这样可以提高数据处理的效率,简化学习模型,使得模型更易于理解。

属性子集选择的基本启发式方法包括逐步向前选择、逐步向后删除以及决策树归纳,表3-11给出了属性子集选择方法。

表3-11 属性子集选择方法

向前选择	向后删除	决策树归纳
初始属性集: $\{X_1,X_2,X_3,X_4,X_5,X_6\}$ 初始化归约集: $\{\}$ $\Rightarrow\{X_1\}$ $\Rightarrow\{X_1,X_4\}$ $\Rightarrow\{X_1,X_4,X_6\}$ ⇒归约后的属性集: $\{X_1,X_4,X_6\}$	初始属性集: $\{X_1,X_2,X_3,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_2,X_3,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_3,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_4,X_6\}$ ⇒归约后的属性集: $\{X_1,X_4,X_6\}$	初始属性集: $\{X_1,X_2,X_3,X_4,X_5,X_6\}$  ⇒归约后的属性集: $\{X_1,X_4,X_6\}$

(1) 逐步向前选择

以空的属性集作为开始,首先确定原属性集中最好的属性,如表3-10所示,初始化归约集后,首先选择属性X_1,将它添加到归约后的属性集中。然后继续迭代,每次都从原属性集剩下的属性中寻找最好的属性并添加到归约后的属性集中,依次选择属性X_4和X_6,最终得到归约后的属性集$\{X_1,X_4,X_6\}$。

(2) 逐步向后删除

从原属性集开始,删除在原属性集中最差的属性,如表3-10所示,首先删除属性X_2,然后依次迭代,再依次删除属性X_3和X_5,最终得到归约后的属性集$\{X_1,X_4,X_6\}$。

(3) 决策树归纳

使用给定的数据构造决策树,假设不出现在树中的属性都是不相关的。决策树中每个非叶子结点代表一个属性上的测试,每个分支对应一个测试的结果,每个叶子结点代表一个类预测,如表3-11所示,对于属性X_1的测试,结果为“是”的对应Class1的类预测结果;结果为“否”的对应Class2的类预测结果。在每个结点上,算法选择“最好”的属性,将数据划分成类。出现在树中的属性形成归约后的属性子集。

以上这些方法的结束条件都可以是不同的,最终都通过一个度量阈值确定何时结束属性子集的选择过程。

也可以使用这些属性创造某些新属性,这就是属性构造。例如,已知属性“radius(半

径)",可以计算出"area(面积)"。这对于发现数据属性间联系的缺少信息是有用的。

3.4.4　抽样

抽样在统计中主要是在数据的事先调查和数据分析中使用。抽样是很常用的方法,用于选择数据子集,然后分析出结果。但是,抽样在统计学与数据挖掘中的使用目的是不同的。统计学使用抽样,主要是因为得到数据集太费时费力;数据挖掘使用抽样,主要是因为处理这些数据太耗费时间并且代价太大,使用抽样在某种情况下会压缩数据量。

有效抽样的理论是:假设有代表性的样本集,那么样本集和全部的数据集被使用且得到的结论是一样的。例如,假设对数据对象的均值感兴趣,并且样本的均值近似于数据集的均值,则样本是有代表性的。但是抽样是一个过程,特定的样本的代表性不是不变的,所以最好选择一个确保以很高的概率得到有代表性的样本的抽样方案。抽样的效果取决于样本的大小和抽样的方法。

假定大型数据集 D 包含 N 个元组。3 种常用的抽样方法如下。

① 无放回的简单随机抽样方法:该方法从 N 个元组中随机$\left(\text{每一数据行被选中的概率为}\frac{1}{N}\right)$抽取出 n 个元组,以构成抽样数据子集。

② 有放回的简单随机抽样方法:该方法与无放回的简单随机抽样方法类似,也是从 N 个元组中每次抽取一个元组,但是抽中的元组接着放回原来的数据集 D 中,以构成抽样数据子集。这种方法可能会产生相同的元组。

图 3-6 表示无放回和有放回的简单随机抽样方法。

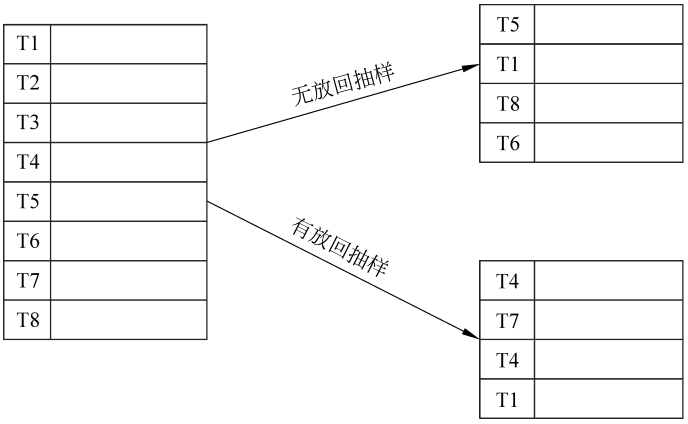


图 3-6　无放回和有放回的简单随机抽样方法示意图

③ 分层抽样:在总体由不同类型的对象组成且每种类型的对象数量差别很大时使用。分层抽样需要预先指定多个组,然后从每个组中抽取样本对象。一种方法是从每个组中抽取相同数量的对象,而不管这些组的大小是否相同。另一种方法是从每一组抽取的对象数量正比于该组的大小。

首先将大数据集 D 划分为互不相交的层,然后对每一层简单随机选样得到 D 的分层选样。例如,根据顾客的年龄组进行分层,然后再在每个年龄组中进行随机选样,从而确保

了最终获得的分层采样数据子集中的年龄分布具有代表性。

选择好了抽样技术,接下来就需要选择样本容量了。过多的样本容量会使计算变得庞杂,但是却可以使得样本更具有代表性;过少的样本容量可以使计算变得简单,但是却可能使得结果不准确。所以,确定适当的样本容量同样非常重要。

3.5 数据变换与数据离散化

3.5.1 数据变换策略及分类

数据变换是将数据转换为适合于数据挖掘的形式,数据变换策略主要包括光滑、聚集、数据泛化、规范化、属性构造和离散化。

① 光滑:去掉数据中的噪声。这类技术包括分箱、回归和聚类。

② 聚集:对数据进行汇总或聚集。例如,可以聚集某超市每一季度的销售商品数据,以获得商品年销售量。一般来说,聚集主要用来为多粒度的数据分析构造数据立方体。

③ 数据泛化:使用概念分层,用高层概念替换低层或"原始"数据。例如,可以把某超市的顾客家庭住址泛化为较高层的概念,如 city、district、street。

④ 规范化:把属性数据按比例缩放,使之落入一个特定的小区间,如 -10.0～0.0 或 0.0～10.0。

⑤ 属性构造(特征构造):通过已知的属性构建出新的属性,然后放入属性集中,有助于挖掘过程。

⑥ 离散化:数值属性(如年龄)的原始值用区间标签(如 0～10 或 11～20)或概念标签(如 youth、adult、senior)替换。这些标签可以递归地组织成更高层概念,形成数值属性的概念分层。图 3-7 就是属性年龄的离散化。

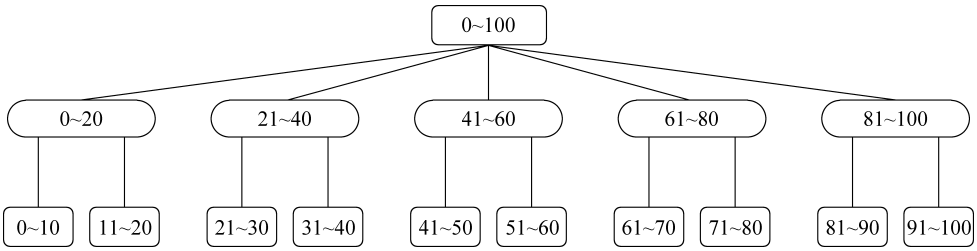


图 3-7 属性年龄的离散化

3.5.2 数据泛化

概念分层可以用来泛化数据,虽然这种方法可能会丢失某些细节,但泛化后的数据更有意义、更容易理解。

对于数值属性,概念分层可以根据数据的分布自动地构造,如用分箱、直方图分析、聚类分析、基于熵的离散化和自然划分分段等技术生成数据概念分层。

对于分类属性,有时可能具有很多个值。如果分类属性是序数属性,则可以使用类似于

处理连续属性方法的技术，以减少分类值的个数。如果分类属性是标称的或无序的，就需要使用其他方法。例如，一所大学由许多系组成，系名属性可能具有数十个值。在这种情况下，可以使用系之间的学科联系，将系合并成较大的学科；或者使用更为经验性的方法，仅当分类结果能提高分类准确率或达到某种其他数据挖掘目标时，才将值聚集到一起。

由于一个较高层概念通常包含若干从属的较低层概念，高层概念属性（如区）与低层概念属性（如街道）相比，通常包含较少数目的值。据此，可以根据给定属性集中每个属性不同值的个数自动产生概念分层。具有越多不同值的属性在分层结构中的层次就越低，属性的不同值越少，则所产生的概念在分层结构中所处的层次就越高。

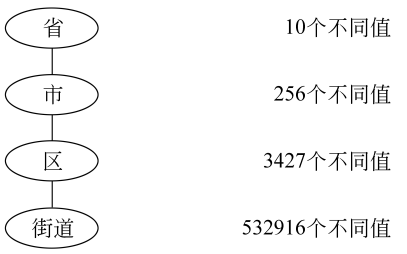


图 3-8　属性地区概念分层的自动生成

首先，根据每个属性的不同值的个数，将属性按升序排列。其次，按照排好的次序，自顶向下产生分层，第一个属性在最顶层，最后一个属性在最低层，如图 3-8 所示。

3.5.3　数据规范化

数据规范化是通过将数据压缩到一个范围内（通常是 0～1 或者－1～1），赋予所有属性相等的权重。对于神经网络的分类算法或者基于距离度量的分类和聚类，规范化是特别有用的。但是有时并不需要规范化，例如算法使用相似度函数而不是距离函数时；再如随机森林算法，它从不比较一个特征与另一个特征，因此也不需要规范化。

数据规范化的常用方法有 3 种：按小数定标规范化、最小-最大值规范化和 z-score 规范化。

1. 按小数定标规范化

通过移动属性值的小数点的位置进行规范化，通俗地说就是将属性值除以 10^j，使其值落在[－1,1]。属性 A 的值 v_i 被规范化 v_i'，其计算公式为

$$v_i'=\frac{v_i}{10^j} \tag{3-13}$$

其中，v_i 表示对象 i 的原属性值，v_i'表示规范化的属性值。j 是使 $\max(|v_i'|)<1$ 的最小整数。

例 3.9　按小数定标规范化。

设某属性的最大值为 5870，最小值为 2320，按小数定标规范化，使属性值缩小到[－1,1]的范围内。

解：题中属性的最大绝对值为 5870，显然只要将属性中的值分别除以 10000，就满足 $\max(|v_i'|)<1$，这时 $j=5870$ 规范化后为 0.587，而 2320 被规范化为 0.232。达到了将属性值缩到小的特定区间[－1,1]的目标。