

第3章 理解数据

叶允明

对数据的理解是数据挖掘过程的一个重要环节，是数据预处理及模型构建的基础。本章将首先介绍理解数据的主要任务及目标（3.1节），然后介绍理解数据的两类方法，即基于统计描述的方法（3.2节）和可视化方法（3.3节）。

3.1 数据理解的主要任务

数据理解就是通过各种统计描述方法或可视化工具对数据的概况、特性进行深入观察和分析的过程。由于每一个数据挖掘项目面对的数据通常是不同的，而理解数据的全貌对于选择哪些数据预处理方法、采用什么样的算法模型是非常重要的，因此数据理解是通用数据挖掘过程模型中的一个重要步骤。例如，通过数据理解可以发现数据中某个属性的少数几个取值与该属性的平均值差异很大，很可能是噪声数据，这就为后续选择数据预处理方法提供了依据。广义上，数据理解主要包括三个任务。

数据理解的第一个任务就是每个属性的取值分布统计描述。以结构化的二维表为例，这个任务就是对给定数据集的各个属性的取值分布情况进行统计概括。例如，对于数值型属性，我们需要计算其算术平均值，从而可以衡量该属性取值的平均状况和集中趋势，而计算其标准差可以了解该属性取值的分散程度。这些计算出来的统计指标就构成了给定数据集的总体特征描述，对后续的数据质量评估、数据预处理方法选择、乃至模型构建算法及参数的设置都具有重要的指导意义。因此这个任务是数据理解中最重要的任务。

数据理解的第二个任务是分析两个及多个属性上的数据对象取值分布情况。上述第一个任务是针对单个属性的取值分布进行统计描述，而第二个任务是针对两个或多个属性上的取值分布进行分析和理解的，其分析结果也可以用来指导数据预处理和模型构建，常采用可视化工具或属性相关性指标来度量。例如，两个数值属性下的数据对象取值分布关联性可以通过散点图来观察；而两个类别型属性相关性的计算可以用来衡量属性的冗余性（将在下一章4.5.1节中介绍），从而为属性选择提供决策依据。

数据理解的第三个任务是数据的总体质量评估。该任务的主要目标是对数据的误差、属性值缺失、噪声和不一致性等潜在的数据质量问题进行观察、分析和评价，为后续采取相应的数据预处理方法提供线索和依据。该任务通常是建立在前两个任务的基础上，并与数据预处理过程紧密结合。例如，通过对某个数值属性的均值和标准差计算，并综合利用箱线图可视化方法可以发现数据中是否存在噪声以及噪声数据对象的比例。

针对上述三个任务，数据理解一般需要借助两个工具来实现，即统计量（统计描述）和可视化工具。前者是利用统计学的方法来实现数据（属性）的总体特征描述（将在3.2节中介绍），后者是通过可视化的图形、并借助人的观察能力来实现数据特征的分析理解（将在3.3节中介绍）。

3.2 基于统计描述的数据理解方法

基于统计描述的数据理解方法力求将数据集属性所包含的所有取值用一组统计量来表示，又称为概括性度量方法，通常是针对单个属性取值分布的统计概括。该类方法主要通过三类统计指标来对属性的取值分布进行概括性描述：（1）集中趋势的度量，用于描述属性取值向“中心”集中的程度和“位置”；（2）散布程度的度量，用于描述属性取值分的分散程度；（3）取值分布形状的度量，用于描述属性取值分布的几何形态，如对称性、峰形。这些不同类型的统计量分别从不同的侧面反映了数据属性的不同特征，有助于我们对数据特征的

深入理解，也是数据预处理的重要基础。下面介绍这三类统计量的原理和计算方法。

由于本节讨论的方法都是针对单个属性的，因此我们不防假设给定的数据集为以下简化的一维数据集：

$$D = \{x_1, x_2, \dots, x_m\}$$

其中 $x_i \in D$ ($i = 1 \dots m$)，对应于一个数据对象，假设 x_i 只包含一个属性 A ，因此 x_i 也可以看作就是第 i 个数据对象在属性 A 上的取值。本章的内容中， A 可以是类别型属性，也可以是数值型属性。

3.2.1 集中趋势度量

集中趋势度量是指数据集在某属性上的大部分取值所在的范围，反映了一组数据向中心聚集的趋势，用于确定数据集的“中心”所在的具体位置。集中趋势的常用度量指标包括计算众数、中位数、均值等。对于不同类型的属性，其所采取的集中趋势度量指标是有所不同的，下面分别介绍数值型属性和类别型属性的相对度量指标及其计算方法。

(1) 数值型属性

数值型属性是数据挖掘领域的数据集最常见的属性。对于数值型属性数据，可采用均值 (mean) 或中位数 (median) 作为度量指标来反映属性数据的集中趋势。

均值即平均数，是对集中趋势最简单的度量方法，通常情况下用简单平均数就可以反映大部分数据集的集中趋势。**简单平均数**是对某属性的值取平均，即对该属性所有的值进行累加再除以数据对象的个数。

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_m}{m} = \frac{\sum x_i}{m}$$

例如针对“年龄”属性的数据集， $D(\text{"年龄"}) = \{56, 43, 55, 76, 78, 34, 49\}$ 。可以采用简单平均数的计算方法，即 $(56+43+55+76+78+34+49)/7=55.86$ 。

简单平均数容易受到“离群点”的影响，例如一组关于“月工资”属性的数据中，大部分对象的取值范围为[5000-12000]，如果有一个对象取值是“256000”（对应于公司总裁的月工资），那么这组数据的简单平均数将受到这个对象的影响而大大偏离正常取值范围[5000-12000]，从而无法反映数据的集中趋势。解决这个问题可以采用两种替代的度量指标：用“**截尾均值**”或“**中位数**”作为集中趋势的度量指标。

截尾均值的思想类似于在体育比赛中的去掉一个最高分和最低分然后取平均值的评分方法，其目的是为了过滤掉极端的数据，统计大部分数据的集中趋势。例如，在前文的信贷数据集中，属性“负债比”的取值大部分分布在 0-1.0 之间，但存在少量数据值大于 10000 的异常取值，因此在计算过程中如果采取单纯的均值将不能准确反映真正的集中趋势，这种情况下可以采取截尾均值。截尾均值的计算需要先将原始数据 D 的按属性取值的大小（一般按从小到大）进行排序，排序后的数据集 $D' = \{x_{[1]}, x_{[2]}, \dots, x_{[m]}\}$ ，这里 $x_{[i]}$ 表示排序后在第 $[i]$ 个位置的对象（其中 $[i]$ 表示对 i 取整）。截尾均值 \bar{x}_α 的计算公式如下：（公式摘自百度百科-截尾平均数）

$$\bar{x}_\alpha = \frac{x_{[m\alpha+1]} + x_{[m\alpha+2]} + \dots + x_{[m-m\alpha]}}{m - 2m\alpha}$$

其中， m 为原始数据对象的个数， α 代表截尾系数，表示被剔除数据（截尾数据）的占比。例如，取 $\alpha=0.1$ 的截尾均值，其计算是首先将数据对象按属性值排序，然后取数据排序后所占比例的范围在 (0.1, 0.9) 之间的取值。以属性“负债比”数据集为例， $D(\text{"负债比"}) = \{0.3, 0.2, 0.1, 0.4, 0.2, 0.2, 30056.7\}$ 。计算“负债比”的 $\alpha=0.1$ 截尾均值，数值排序后的结果为

$D'(\text{"负债比"}) = \{0.1, 0.2, 0.2, 0.2, 0.3, 0.4, 30056.7\}$, 排序范围在 $(m\alpha + 1, m - m\alpha)$ 即 $(1.7, 6.3)$ 之间, 取整之后取数值排序从 1 到 6 之间的数值, 30056.7 排名为 7, 因此被排除在外。“负债比” $\alpha=0.1$ 的截尾均值计算结果为 0.23, 更符合逾期。

中位数是一组数据中数值排序后处在中间位置上的数值。中位数通常可以反映数据的中心位置所在。因此可以作为一种度量数值型属性的集中趋势手段。计算中位数需要对该属性中的值进行排序, 然后取排序后中间位置的值。这里定义 $x_{[i]}$ 作为排序后在第 $[i]$ 个位置的对象 (其中 $[i]$ 表示对 i 取整), 整体排序后的数据集为 $D' = \{x_{[1]}, x_{[2]}, \dots, x_{[m]}\}$ 。当数据对象总数为奇数时中位数即 $x_{[\frac{m+1}{2}]}$, 当对象总数为偶数时其中位数是 $(x_{[\frac{m+1}{2}]} + x_{[\frac{m}{2}]})/2$ 。

例如反映“收入”的数据集, $D(\text{"收入"}) = \{5000, 4500, 7000, 6000, 4000, 4400, 5500\}$ 。

将收入排序后的结果为 $D'(\text{"收入"}) = \{4000, 4400, 4500, 5000, 5500, 6000, 7000\}$, 其中位数为 5000。

中位数是将数据集两等分的“中间”属性值, 类似的度量指标还有四分位数 (quartile) 和百分位数 (percentile): 四分位数是将数据四等分, 百分位数是将数据 100 等分。这里介绍四分位数的计算方法 (百分位数的计算方法类似)。

四分位数是指数据经过排序后处在 25% 位置和 75% 位置上的属性值, 其中处在 25% 位置的取值称为上四分位数, 而处在 75% 位置的取值称为下四分位数。上四分位数、中位数、下四分位数把数据四等分, 每个部分包含数据集中 25% 的对象 (已根据属性值排序)。上四分位数与下四分位数之间通常是大部分数据对象取值的集中区间。四分位数的计算方法与中位数的计算方法类似, 下四分位数一般可取 $x_{[\frac{m+1}{4}]}$, 而上四分位数可取 $x_{[\frac{3(m+1)}{4}]}$ 。例如, 计

算 $D'(\text{"收入"}) = \{4000, 4400, 4500, 5000, 5500, 6000, 7000\}$ 的上四分位数为 4400, 下四分位数为 6000。即数据集中大部分人的收入集中在 4400 到 6000 之间。

几何平均数是适用于比率数据的集中趋势度量指标, 即多用于计算等比或近似等比的数据, 例如计算基金产品的平均年收益率。几何平均数的计算公式如下:

$$G = \sqrt[m]{x_1 \times x_2 \times \dots \times x_m} = \sqrt[m]{\prod x_i}$$

例如, 假定某基金产品近三年的年收益率 (按复利计算): 第一年 6% (比率 1.06), 第二年 5%, 第三年 9%。则该基金产品的平均年收益率为: $\sqrt[3]{1.06 \times 1.05 \times 1.09} - 1 = 6.653\%$ 。

(2) 类别型属性

类别型属性通常不能采用均值或中位数的方法进行度量集中趋势, 因为其均值或中位数没有统计意义或实际的语义。对于类别型属性数据, 一般采取众数 (mode) 作为集中趋势度量的计算方法。众数是一组数据中出现频率最高的属性值, 一个数据集中可以有一个或多个众数, 也可以没有众数。

例如, 在一个属性“年龄”的数据集中: $D(\text{"年龄"}) = \{\text{中年}, \text{青年}, \text{老年}, \text{老年}, \text{青年}, \text{中年}, \text{中年}\}$, 其中属性值“青年”、“中年”、“老年”出现的次数分别是 2、3、2, 因此其众数为“中年”, 表示在该数据集中, 大部分对象在属性“年龄”上取值的集中趋势为“中年”。又

例如， $D(\text{"年龄"}) = \{\text{青年}, \text{青年}, \text{老年}, \text{老年}, \text{青年}, \text{中年}, \text{老年}\}$ ，其中属性值“青年”、“中年”、“老年”出现的次数分别是 3、1、3，则该数据集有两个众数：“青年”和“老年”。如果 $D(\text{"年龄"}) = \{\text{青年}, \text{青年}, \text{老年}, \text{老年}, \text{中年}, \text{中年}\}$ ，其中属性值“青年”、“中年”、“老年”出现的次数分别是 2、2、2，数据集中所有类别（属性值）出现的频次相同，则该数据集没有众数，表示该类别型属性的取值分布是均匀的，无集中趋势。

不同个数众数的数据集可以用类别频度图来直观表示，如图 3-1 所示。图中的横坐标表示各个类别型数据集中包含的类别（属性值），纵坐标为各类别（属性）出现的频率。

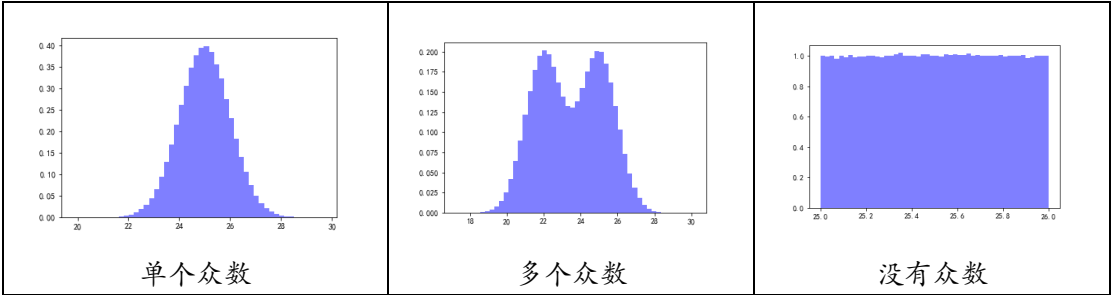


图 3-1 不同数据集的众数示意图

(3) 集中趋势度量指标的比较

总结起来，在面对实际数据时，应该针对数据的特性采取相对应的集中趋势度量指标，如图 3-2 所示：首先根据数据类型的不同，采取不同的度量方法。对于类别型数据，可采用众数作为度量指标。对于数值型数据，要根据数值型数据的不同特点，例如判断是否存在两端异常值，从而采取不同的计算方法。简单平均数是最常用、最简单的集中趋势度量指标，但容易受到异常值的干扰。截尾均值、中位数可以解决简单平均数的噪声不稳定性问题，但需要对数据进行排序，计算量更大。四分位数通常可以规避两端异常值的影响，能够反映整个数据集取值分布的集中区域。另外，在处理等比数据或近似等比数据时，一般要采用几何平均数才能得到更准确的集中趋势估计。

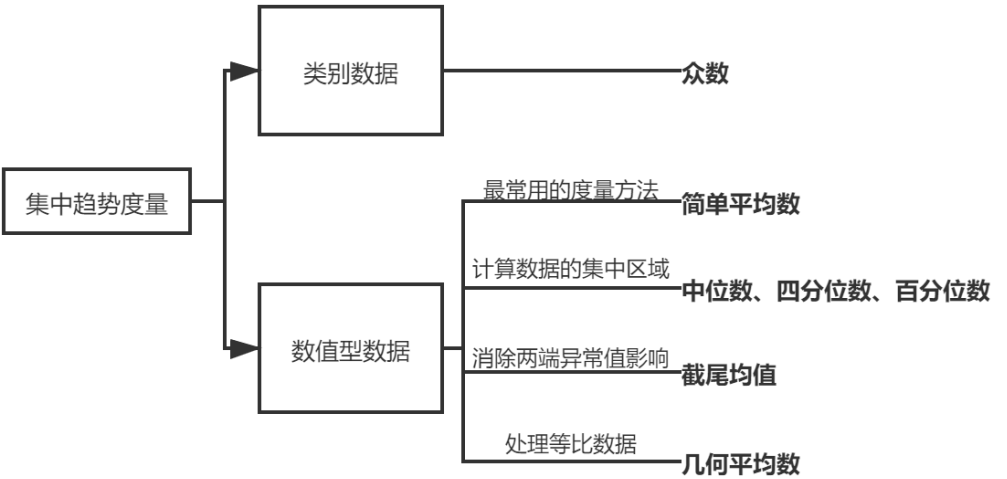


图 3-2 集中趋势度量不同方法归纳

3.2.2 散布程度度量

相对于集中趋势度量，散布程度度量表示数据向周围分散的程度。数据的散布程度越大，则集中趋势度量指标值的代表性就越差，反之则越好。不同类型数据（属性）的散布程度度量方法不同的，常见的度量指标包括异众比率、四分位距、方差和标准差等。

(1) 数值型属性

对于数值型属性数据，度量散布程度的常见指标之一是**极差**（range），即用属性值中最大值与最小值之差。极差反映了整个数据集取值的跨度，是最简单的散布度量，但同时也容易受到数据集中的极端值影响。极差 R 计算公式如下式所示：

$$R = X_{max} - X_{min}$$

其中 X_{max} 为数据集（属性值）中的最大值， X_{min} 为最小值。例如在数据集 D （“负债比”）= {0.3, 0.2, 0.1, 0.4, 0.2, 0.2, 30056.7} 中，最大值为 30056.7，最小值为 0.1，其极差为 30056.6，显然受到了极端值 30056.7 的干扰。

四分位距（interquartile range，简称为 IQR）是另一个常见的散布程度度量指标，又常称为四分位差。四分位距是上四分位数与下四分位数之差，因为其定义是基于四分位数的，因此具有不易受极端值影响的特性。例如，上述的数据集 D （“负债比”）的上四分位数为 0.35，下四分位数为 0.2，其四分位差为 $0.35 - 0.2 = 0.15$ 。四分位距反映了数据经排序后的中间 50% 属性值的分散程度，其数值越小表明中间部分的数据越集中，反之则越分散。

方差和**标准差**反映了每个数据对象相对于数据集中心位置的平均距离，是数据挖掘应用中最常用的散布程度度量指标。方差一般用 s^2 表示，标准差是方差开方得到，方差的计算公式如下：

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_m - \bar{x})^2}{m - 1} = \frac{\sum (x_i - \bar{x})^2}{m - 1}$$

式中的 \bar{x} 为上一节介绍的简单平均数（均值）。上一节我们求得数据集 D （“年龄”）= {56, 43, 55, 76, 78, 34, 49} 的均值为 55.86，通过上式可求得其方差为 $((56 - 55.86)^2 + (43 - 55.86)^2 + (55 - 55.86)^2 + (76 - 55.86)^2 + (78 - 55.86)^2 + (34 - 55.86)^2 + (49 - 55.86)^2) / 6 = 264.48$ 。

(2) 类别型属性

对于类别型属性数据，异众比率（variation ratio）是常用的散布程度度量指标。异众比率的定义是基于众数的。由上节可知众数是指数据中出现频率最高的属性值，可以在一定程度反映类别型属性数据的集中趋势。异众比率则指数据集中非众数的数据对象的数量占整个数据集的比例，其计算公式如下：

$$\rho = \frac{m - f}{m}$$

其中， ρ 表示异众比率， m 表示数据集的对象总数， f 表示数据集中取值为众数的数据对象总数。显然，异众比率越大，则数据越分散，众数对数据集的代表性越差；反之亦然。例如在数据集 D （“年龄”）= {中年, 青年, 老年, 老年, 青年, 中年, 中年} 中，众数为“中年”，异众比率为 $4/7$ ，过半数的数据并没有集中在中心区域，因此其众数反映的集中趋势代表性不强。

(3) 散布程度度量各计算方法的比较

与集中趋势度量指标的选择方法类似，数据集的散布程度度量也需要针对数据的特性采

取相对应的度量指标，如图 3-3 所示。首先，针对不同类型的属性数据应选取不同的度量方法。对于类别型属性，可选取异众比率进行估计。对于数值型属性数据，极差反映的是整个数据集的全距，易受到两端异常值的干扰；四分位差则反映了数据集的中间距，能较好规避两端异常值的影响；标准差和方差是数据挖掘项目中最常用的散布程度度量指标，但相较于极差和四分位差来说其计算过程相对复杂。

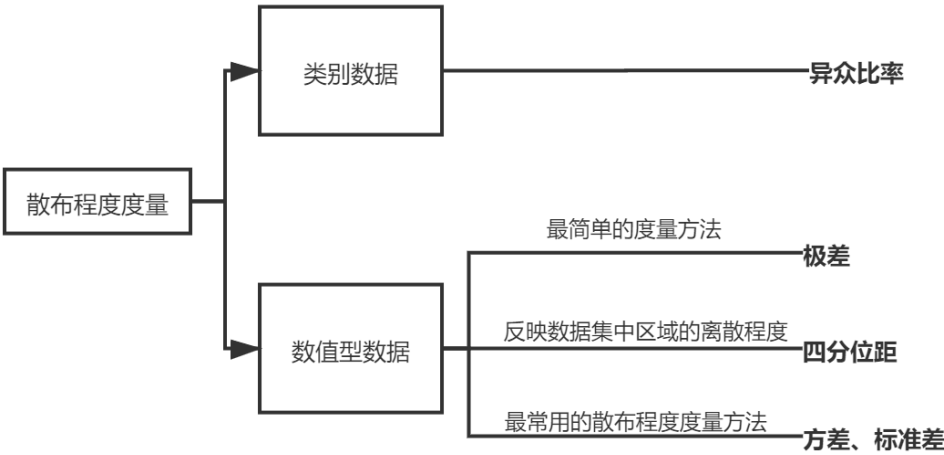


图 3-3 散布程度度量不同计算方法比较图

3.2.3 分布形态的度量

我们可以通过 3.2.1 和 3.2.2 小节介绍的方法度量数据的集中趋势与散布程度，本节将从偏态和峰度两个度量方法来进一步分析数值型数据的分布形态，从而对数据分布的几何特征有更深入的认识。

(1) 偏态度量

偏态 (skewness) 是描述数值型数据分布的对称性的度量指标。数据分布的对称性可以分为对称、左偏或右偏。偏态度量是以正态分布为基准的，即描述数据分布相对于正态分布是对称或者左偏、右偏。一般情况下，偏态类型可以通过均值、众数和中位数的大小比较来简单评估：如果数据处于正态分布（对称），则均值=众数=中位数；如果数据呈现左偏，一般均值<中位数，且均值<众数；如果数据呈现右偏，一般均值>中位数，且均值>众数。评价数据分布偏态的常用方法是采用统计量——偏态系数进行度量。偏态系数一般记为 S_k ，其计算公式如下：

$$S_k = \frac{m \sum (x_i - \bar{x})^3}{(m - 1)(m - 2)s^3}$$

其中， \bar{x} 表示数据分布的均值， s^3 表示标准差的三次方。如果偏态系数等于 0，则说明数据的分布是对称的，偏态系数越接近 0 整体数据偏斜的程度就越低，反之就越高。偏态系数大于 0 说明数据分布右偏，偏态系数小于 0 则说明数据分布左偏。例如，上一小节求得数值型数据 $D(\text{"年龄"}) = \{56, 43, 55, 76, 78, 34, 49\}$ 的方差为 264.48，标准差为 16.26，可求出其偏态 $S_k > 0$ ，因此年龄属性分布右偏。

图 3-4 是不同类型偏态（对称、左偏、右偏）的数据分布示意图，其中的横坐标表示属

性数据的取值范围，纵坐标表示各个取值的频次（或频率）。图中的符号 \bar{x} 表示数据的简单平均数， M_e 表示中位数， M_o 表示众数，从图中可以看出，不同类型偏态的数据分布其均值、众数、中位数具有不同的大小关系。

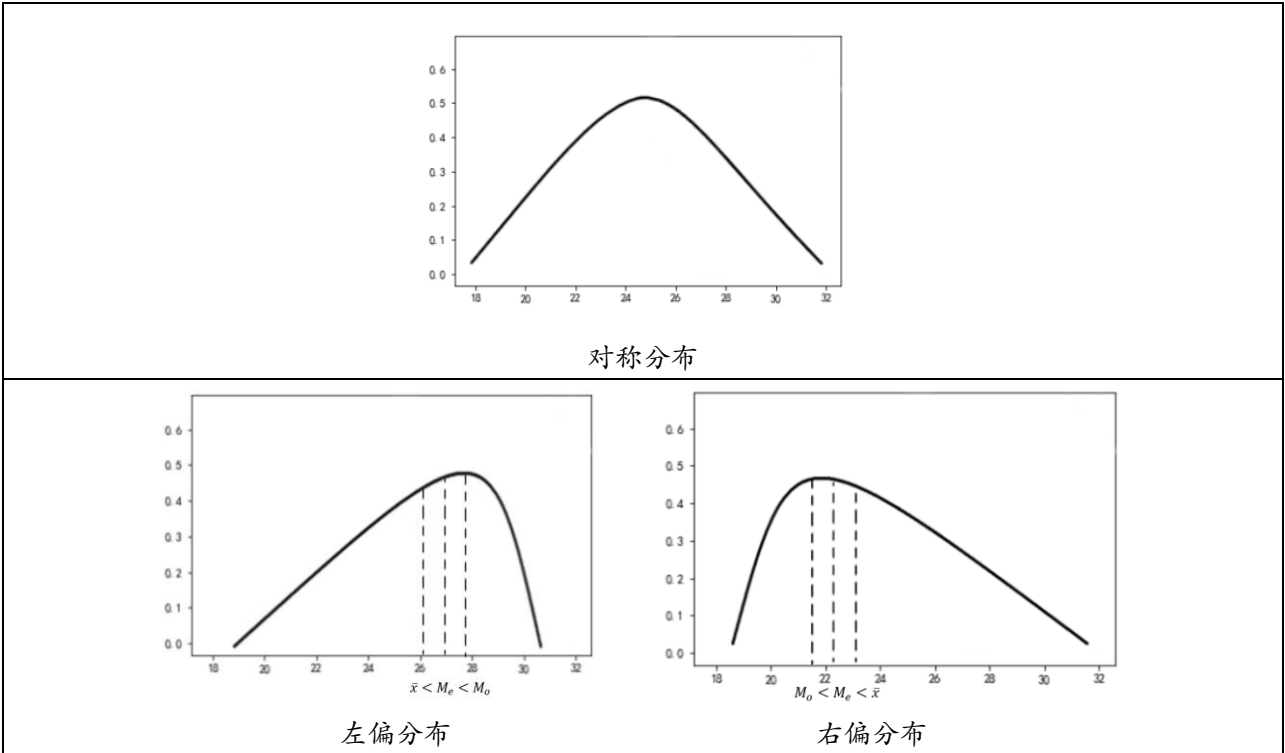


图 3-4 不同偏态的分布示意图

在数据分布的统计描述中，偏态度量是集中趋势度量和散布度量的有益补充。例如，图 3-5 表示的是某地区职工的月收入统计分布，其中的横坐标表示月收入取值，纵坐标为该收入取值下对应的统计频次。从图中可以看出，该地区的职工月收入明显存在不均衡，其月收入的分布是不对称的（呈现左偏），月收入的众数为 6500，然而绝大部分职工的月收入低于 6500；同时均值约为 5300，但仍存在大部分人收入低于均值，可见其众数和均值在描述该数据集的分布时由于数据的偏态为不对称，并不能很好反映数据的集中趋势。一般来说，数据的偏态越“偏离”对称，则均值和众数度量集中趋势的可靠性越弱。因此，在实际应用中，需要将偏态度量与均值、众数度量进行结合。

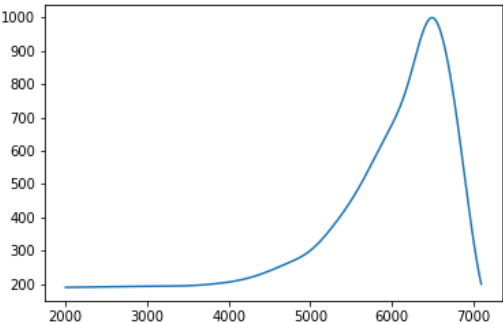


图 3-5 某地区收入统计示意图

(2) 峰度度量

数据分布的峰度是描述数值型数据分布的几何形态陡峭程度的统计量。峰度是以正态分布为基准定义的：如果数据分布与正态分布的陡峭程度相同，则峰度为 0；如果数据分布比正态分布更陡峭（分布的几何形态为尖顶峰），则峰度大于 0；如果数据分布比正态分布更平坦（分布的几何形态为平顶峰），则峰度小于 0。峰度的绝对值越大则表明数据分布的陡峭程度与正态分布的差异程度越大。峰度的计算方法如下式：

$$Kurt = \frac{m(m+1) \sum (x_i - \bar{x})^4 - 3[\sum (x_i - \bar{x})^2]^2 (m-1)}{(m-1)(m-2)(m-3)s^4}$$

如图 3-6 所示，正态分布的峰度=0，均匀分布的峰度=-1.2，峰度低于 0 一般数据越分散称之为平峰分布，峰度大于 0 数据越集中称之为尖峰分布。

在实际应用中，数据的峰度通常能够直观的反映数据的集中程度，对于如图 3-6 所示，其中正态分布、平峰分布和尖峰分布的均值和众数均相同，此时单纯从集中趋势度量的角度观察数据的集中趋势会得出三组数据集中趋势相同，因此观察数据的峰度能验证集中趋势度量指标的可靠性。峰度越高，则均值、众数等集中趋势度量的可靠性越强。

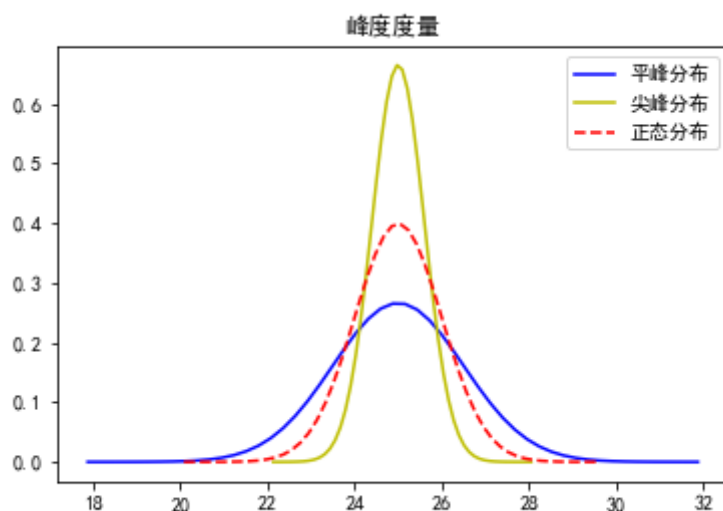


图 3-6 不同峰度的数据分布示意图

3.3 数据可视化方法

数据可视化是指通过各种图形、表格、可视化符号等工具对数据进行显示的方法，是分析和理解数据的另一类重要途径。

后者是通过可视化的图形、并借助人的观察能力来实现数据特征的分析 and 理解

上一节中我们介绍了用一个或一组统计数据来表征某个属性的全部数据，然而在很多情况下统计学的统计数据含义往往受到极端值的影响，比如均值。如果采取截尾均值我们也需要对数据的整体分布有一个先验知识，知道大概什么从位置开始会出现所谓的异常的极端值。因此这里我们引入数据可视化的方法，对数据产生一个最基本最直观的认知，通过可视化之后的数据对于数据的统计分析和后续的数据预处理工作都有很大帮助。

3.3.1 柱状图

柱状图可以把类别型数据的统计频次可视化，让使用者对该类别数据各个类别的计数统计有一个直观的认识。柱状图中，每一个数据类别的计数统计以“条柱”的形式展示，其中数据的类别的频次比较以“条柱”的高度作为评判标准。

例如对于“年龄域”的统计，其类别分类有青年、中年、老年，分别统计青年、中年、老年的频次之后将其计数统计用柱状图可视化，横坐标一般代表数据（属性取值）的类别，纵坐标为数据的频次或频率，其效果如图 3-7 所示：

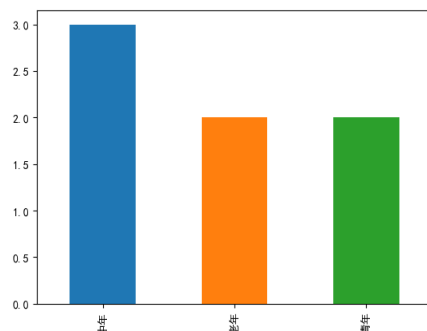


图 3-7 年龄域柱状图

3.3.2 直方图

数值型数据不同于类别型数据，其数值一般是连续的，统计数值型数据的频次可以将其归类，然后采用上一节中的柱状图进行可视化。比如“收入”的属性，可以按照划分将收入划分为 1000 以下，1000 到 10000，10000 以上等，然后依次统计，画柱状图。

对于数值型数据各数据频次的统计通常可以采用直方图。首先要保证个频次的统计是有意义的，如“收入”的属性，几乎每条数据都不相同，统计其频次是没有意义的。像“年龄”这种字段的属性，其取值在一定范围内，就可以通过直方图表示，其横坐标为数值的取值跨度，纵坐标为数值的频率或频次，如图 3-8 所示。

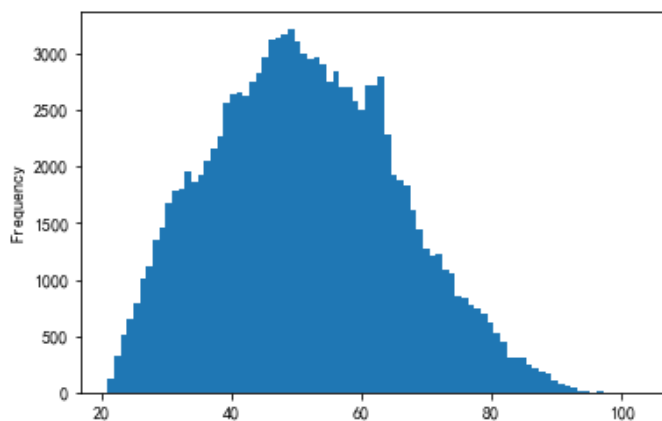


图 3-8 直方示意图

3.3.3 散点图

在数据挖掘的数据中，可能存在某两个属性直接存在一定的相关关系，而属性之间很难从数值上之间观察出他们之间的相关性，因此这里引入散点图，用于更好观察两个属性之间的数值分布。

散点图主要应用于数值型数据，其作用主要是观察两个数值型属性的数据之间的关联性，如图 3-9 所示，如果两个数值型属性存在某种线性关系，则可以比较容易利用散点图的

形式观测出来。如果属性之间存在正相关，则随着一个属性的数值增大，另一个属性的数值也具有增大的趋势；如果属性之间存在负相关，则随着一个属性的数值增大，另一个属性的数值具有减少的趋势。

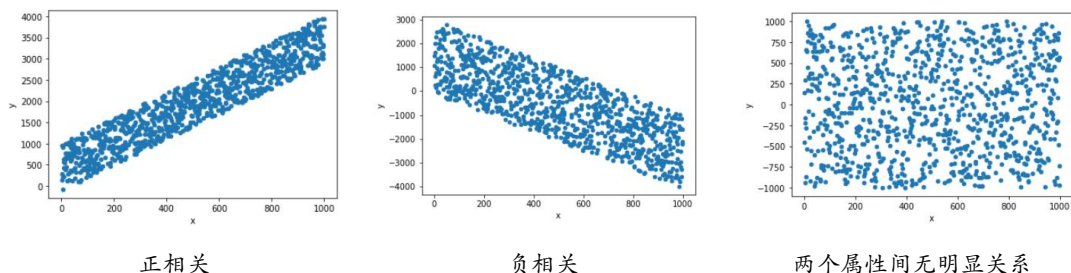


图 3-9 不同数据集的散点图相关性示意图

同时散点图可以直观的反映集中数据和离群数据的分布,还能直接反映数据集内存在的分组情况。例如图 3-10 所示,数据的集中区域有两组,还可以利用散点图同时观察三个属性的关系性,绘制三维散点图。

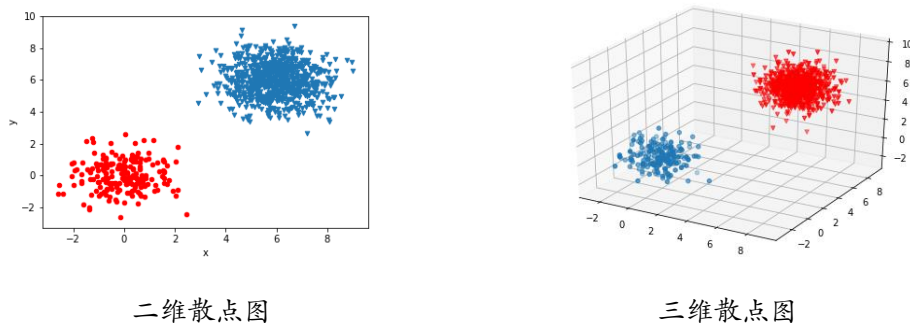


图 3-10 散点图分类显示示意图

3.3.4 箱线图

上一节中我们提到了如何计算数值型数据的中位数和上下四分位数,箱线图可以有效的将这些指标可视化。

箱线图是描述数值型数据属性的可视化方法,箱线图能够同时描述数据的中位数、上下四分位数以及上下边缘。箱线图上的上边缘取的是上四分位数加 1.5 倍四分位距与数据最大值的最大值,上边缘取的是下四分位数减 1.5 倍四分位距与数据最小值的最大值。计算方法如图 3-11 所示, $Q1$ 表示下四分位数, $Q3$ 表示上四分位数,其中 IQR 表示四分位距,即 $Q3-Q1$ 。

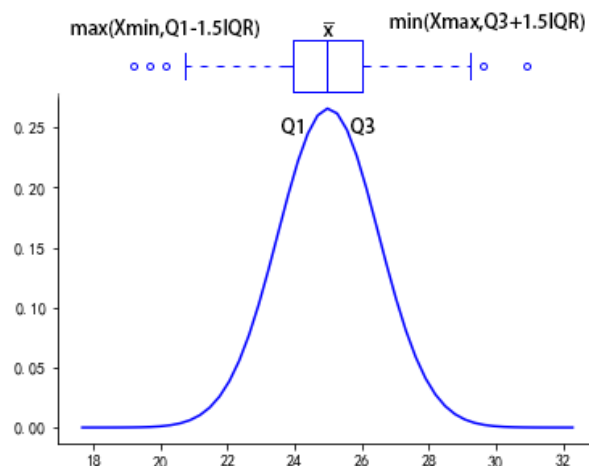


图 3-11 箱线图计算方法

超出上下边缘的数据会在箱线图中标记为异常值。如图 3-12 所示，可以看出不同属性数据的中位数、上下界，其中右侧 y 属性对应的箱线图存在低于下限的异常值。

四分位距 (interquartile range, 简称为 IQR)，又常称为四分差

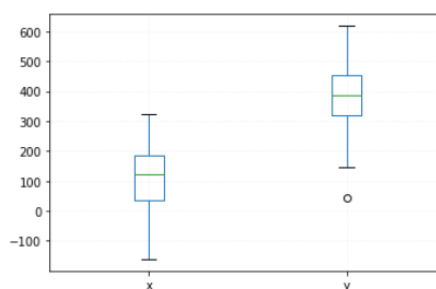


图 3-12 不同数据集的箱线示意图

3.3.5 数据矩阵图

如果想将一张完整的、数据规模不大的表可视化出来，可以采用数据矩阵图。

数据矩阵图，是描述数值型数据的可视化方法。其中，主要作用是将数值映射到不同的颜色上，通过可视化可以较为清晰看出数据的数值大小区分。

图 3-13 是一个 1949-1960 年各月的航班飞行次数累积表的矩阵可视化表示，可通过颜色的深浅较为明显看出数据集中位置。

(事例摘自 http://seaborn.pydata.org/examples/heatmap_annotation.html)

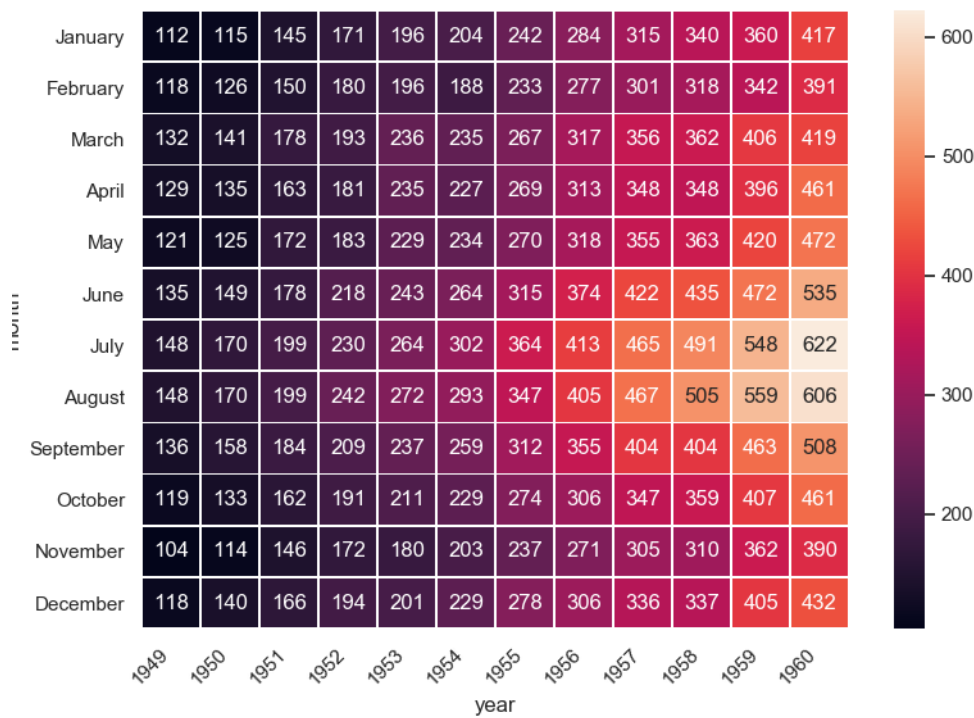


图 3-13 数据矩阵示意图

3.3.6 雷达图

以上的可视化方法大多是作用于属性列的可视化方法，如果想对每一个数据对象进行可视化，可以采用雷达图。

雷达图的主要应用于对于不同数据对象在多个数值型属性的条件下的对比，雷达图中通常包含多个维度，每个维度所代表的数值含义以及标度均有所不同。比如我们对比在信贷数据中多个数据对象在“负债比”“收入”和“年龄”三个属性维度上的对比，如图 3-14 所示。

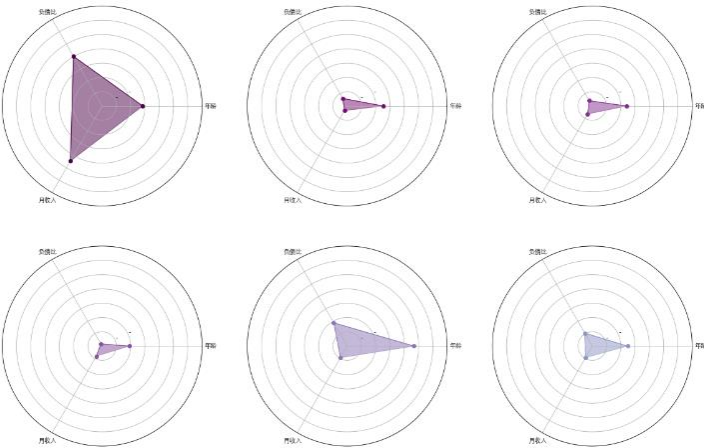


图 3-14 雷达示意图