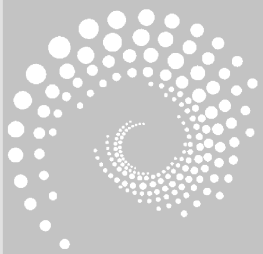
第3章

数据探索与预处理

CHAPTER 3



在互联网大数据分析和机器学习项目中，数据探索与预处理是至关重要的步骤，有助于人们理解数据的特征、分布和质量，为后续的建模和分析提供高质量的数据。数据探索是数据分析的第一步，它帮助人们了解数据的基本情况，包括数据的结构、分布、缺失值、异常值等。通过数据探索，可以发现数据中的潜在问题，为后续的数据预处理和分析提供指导。数据预处理是在数据探索之后进行的，它的目的是将原始数据转换为适合机器学习算法处理的格式。数据预处理包括数据清洗、特征工程、数据转换等步骤，这些步骤可以提高数据的质量和可用性，从而提高模型的性能。

3.1 数据属性类型

随着互联网技术和计算技术的飞速发展,每个人不仅是数据的消费者,还是生产者,数据量呈几何级数增长。数据的来源主要有交易数据、互联网数据、传感器数据、社交媒体、企业内部系统等,这些数据具有结构复杂、类型多样、更新快等特点。大数据获取是指通过各种技术手段,从各种来源和格式的数据中收集、转换、整理大量数据的过程。数据属性的类型,包括标称属性、二元属性、序数属性和数值属性,以及它们的特点和应用场景。数值属性进一步分为区间标度和比率标度,离散属性和连续属性则是根据取值方式划分。这些概念在数据预处理和机器学习中至关重要。

3.1.1 数据属性

数据集由数据对象(又称样本、实例、数据点或对象)组成,一个数据对象代表一个实体。属性是表示数据对象的一个特征的数据字段。属性向量(或特征向量)是用来描述给定对象的一组属性。属性又可以分为标称属性(Nominal Attribute,NA)、二元属性(Binary Attribute,BA)、序数属性(Ordinal Attribute,OA)、数值属性(Numerical Attribute,NA)等多种类型。

1. 标称属性

标称属性的值是一些符号或事物的名称。每个值代表某种类别、编码或状态。标称属性的值是枚举的,可以用数字表示这些符号或名称。常见的标称属性有姓名、籍贯、邮政编码或婚姻状态等。标称属性的值不仅是不同的名字,更是用于区分对象的重要信息。

2. 二元属性

二元属性作为标称属性的特殊形式,同时也是一种布尔属性,其特征为仅包含“0”和“1”两个互斥状态,分别对应不同的类别或状态。根据状态结果的重要性差异,二元属性可分为对称与非对称两类。对称二元属性的两种状态具有同等重要性,如抛硬币的“正面”与“反面”。非对称二元属性的状态重要性不同,其中一个状态(通常为“1”)代表需要重点关注的关键信息(如病毒检测的“阳性”结果),另一个状态(通常为“0”)则为默认或背景状态(如“阴性”结果)。

3. 序数属性

序数属性的取值之间具备有意义的序。例如,学生的成绩属性可分为优、良、中、差 4 个等级。序数属性适用于记录那些依赖主观判断、缺乏客观量化标准的质量评价,这类评价通常无法通过物理测量或标准化工具进行客观度量,其核心特征是属性值之间存在可排序的逻辑顺序,但相邻等级间的具体差距无法精确量化。

4. 数值属性

数值属性是可度量的量,用整数或实数值表示,有区间标度属性和比率标度两种类型。

区间标度属性用相等的单位尺度度量，区间属性的值有序。比率标度属性的度量是比率的，可以用比率来描述两个值，即一个值是另一个值的倍数。

3.1.2　数据属性的特征分析

在数据预处理阶段，数据属性的特征分析是至关重要的一步。通过对属性特征选择合适的数据处理方法，提高互联网大数据分析和机器学习的效率与准确性。

1. 属性的分布特征

属性的分布特征描述了数据在不同属性值上的分布情况。常见的分布特征包括均匀分布、正态分布、偏态分布等。属性的分布特征分析方法包括直方图、箱形图、概率密度函数和累积分布函数等。

直方图(Histogram)：通过将属性值划分为若干区间(桶)，并统计每个区间内的数据点数量，可以直观地展示数据的分布情况。直方图适用于数值属性和序数属性。

箱形图(Boxplot)：箱形图基于四分位数(Q1、Q2、Q3)和四分位距(IQR)，能够清晰地展示数据的集中趋势、离散程度以及异常值情况。箱形图特别适合用于检测数据中的异常值。

概率密度函数(Probability Density Function，PDF)和累积分布函数(Cumulative Distribution Function，CDF)：对于数值属性，可以通过计算概率密度函数和累积分布函数来更精确地描述数据的分布特征。正态分布的数据具有对称的 PDF 曲线和 S 形的 CDF 曲线。

2. 属性的相关性分析

属性的相关性分析用于评估不同属性之间的相互关系。属性之间可能存在正相关、负相关或无相关关系。属性的相关性分析方法包括散点图、相关系数和卡方检验等。

散点图(Scatter Plot)：通过在二维平面上绘制两个属性的值，可以直观地观察它们之间的关系。如果数据点呈现出明显的线性趋势，则表明两个属性之间存在相关性。

相关系数(Correlation Coefficient)：相关系数是衡量两个数值属性之间线性相关程度的统计量，包括皮尔逊相关系数(Pearson Correlation Coefficient)和斯皮尔曼秩相关系数(Spearman Rank Correlation Coefficient)。皮尔逊相关系数用于衡量线性相关性，其值介于－1 和 1 之间；斯皮尔曼秩相关系数则适用于非线性关系的评估。

卡方检验(Chi-square Test)：对于标称属性和序数属性，可以使用卡方检验来评估属性之间的独立性。如果卡方检验的 p 值小于显著性水平(如 0.05)，则可以认为两个属性之间存在显著的相关性。

3. 属性的重要性分析

属性的重要性分析用于评估每个属性对数据分析目标的贡献程度，减少数据维度，提高模型的性能。属性的重要性分析方法包括基于统计的方法、基于模型的方法、基于正则化的方法。

基于统计的方法：计算每个属性的信息增益(Information Gain)来评估其重要性，信息

增益越大,表示该属性对数据的分类或预测能力越强。

基于模型的方法:在如决策树、随机森林等机器学习模型中,可以利用模型的特征重要性评分来评估属性的重要性。这些模型会根据属性对模型预测能力的贡献程度给出相应的评分。

基于正则化的方法:例如,在 Lasso 回归中,通过引入 L1 正则化项,自动地将一些不重要的属性系数压缩为 0,实现属性选择。

3.1.3 数据属性的转换与编码

数据属性的转换与编码是数据预处理中的关键步骤之一。通过对数据属性进行适当的转换和编码,可以将数据转换为适合文本大数据分析和挖掘的格式,同时有助于提高模型的性能和解释性。

1. 数值属性转换

规范化(Normalization)是将数值属性的值缩放到一个特定的区间(如[0,1]或[-1,1]),以消除不同属性之间量纲的影响。数值属性规范化常用的方法包括最小-最大规范化、Z-score 规范化、小数定标规范化和离散化等。

最小-最大规范化:将属性值缩放到指定区间。计算方式为

$$v' = \frac{v - \min(v)}{\max(v) - \min(v)} \times (\text{new}_{\max} - \text{new}_{\min}) + \text{new\_min} \tag{3-1}$$

Z-score 规范化:将属性值转换为均值为 0、标准差为 1 的分布。计算方式为

$$v' = \frac{v - \mu}{\sigma} \tag{3-2}$$

小数定标规范化:通过移动小数点位置将属性值缩放到[-1,1]区间。计算方式为

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

式(3-3)中,j 是使得$|v'|$小于 1 的最小整数。

离散化(Discretization)是将连续的数值属性划分为离散区间的过程。等宽离散化是将数值范围按相同间隔分割为若干区间,每个区间的数值跨度一致;等频离散化则是将数据按样本数量均等原则划分为若干区间,确保每个区间包含相同数量的数据点。基于聚类的离散化方法通过 K-means 聚类算法将数值划分为若干簇,每个簇对应一个离散区间。

2. 分类数据编码

分类数据编码方法包括独热编码、标签编码和二进制编码。

独热编码(One-hot Encoding)是将标称属性的每个类别值转换为一个独立的二进制向量的过程。假设标称属性有 k 个类别值,则将其转换为 k 个二进制特征,每个特征对应一个类别值。如果某个数据点的类别值为第 i 个类别,则其对应的独热编码向量中第 i 个位置为 1,其余位置为 0。

标签编码(Label Encoding)是将标称属性的每个类别值映射到一个整数值的过程。其核心是将非数值型的类别标签转换为唯一的整数数值,以便机器学习模型能够有效识别和

处理。适用于类别值之间存在顺序关系的序数属性。假设标称属性有 k 个类别值，则将其映射到整数值区间$[0,k-1]$。需要注意的是，标签编码可能会引入类别值之间的隐含顺序关系，因此在某些情况下需要谨慎使用。

二进制编码(Binary Encoding)是一种以二进制数字(0 和 1)为基础的信息表示技术，通过特定规则将文字、符号、数值、图像、声音等各类信息转换为二进制序列。例如，对于“颜色”属性，“红色”可以编码为“00”，“绿色”可以编码为“01”，“蓝色”可以编码为“10”。

3. 时间序列数据转换

时间序列数据的转换方法包括时间戳转换、时间窗口划分和差分转换等。时间戳转换将时间序列数据的时间戳转换为日期时间格式，以便进行时间相关的分析。时间窗口划分将时间序列数据划分为固定长度的时间窗口，以便进行滑动窗口分析。差分转换用来计算时间序列数据的差分，以消除数据的趋势和季节性。

4. 时间属性的处理

时间特征提取：时间属性通常包含丰富的信息，如年份、月份等。通过提取时间特征，可将时间属性转换为多个数值属性。对于日期特征：提取其年份、月份等特征。对于时间特征：提取小时、分钟、秒等特征。对于时间周期特征：计算时间点与节假日、促销活动等某个特定时间点之间的距离。

时间序列转换：对于时间序列数据，通过差分、滑动平均等方法进行转换，以消除时间序列中的趋势和季节性成分。差分是用于计算相邻时间点之间的差值，计算方式为

$$y'_t = y_t - y_{t-1} \tag{3-4}$$

滑动平均是用于计算时间窗口内的平均值，计算方式为

$$y'_t = \frac{1}{n}\sum_{i=t-n+1}^{t} y_i \tag{3-5}$$

5. 文本数据转换

文本数据转换方法是将文本数据转换为计算机可以处理的数值形式的过程，包括词袋模型、词频-逆文档频率模型编码和词嵌入等。

词袋模型(Bag-of-Words Model)是一种基础的文本数据表示方法，其核心思路是将文本视为词语的无序集合，不考虑词语间的顺序关系及语法结构。该模型通过统计每个词语在文本中的出现频率，将文本转换为一个特征向量：向量的每个维度对应语料库中的一个独立词语，维度上的数值则表示该词语在当前文本中的出现次数。

词频-逆文档频率模型(Term Frequency-Inverse Document Frequency，TF-IDF)编码是一种广泛应用于文本处理的量化表示方法，其核心在于融合词频(TF)与逆文档频率(IDF)两个关键指标来评估词语对文档的重要程度。该模型的基本逻辑是：若某个词语在目标文档中出现的频率越高，且在整个语料库的其他文档中出现的频率越低，则该词语对当前文档的区分度和代表性就越强，计算方式为

$$\text{TF-IDF}(w,d) = \text{TF}(w,d) \cdot \text{IDF}(w) \tag{3-6}$$

其中，w 代表词语，d 代表文档。$\text{TF}(w,d)$表示词频，就是 w 在 d 中出现的次数。$\text{IDF}(w)$

表示词语 w 的逆文档频率,它用于衡量词语 w 在整个文档集合中的重要性。词嵌入作为一种将词汇映射为低维向量空间表示的技术,其核心在于通过向量形式捕捉词语间的语义关联。常见的词嵌入方法包括 Word2Vec、GloVe 和 FastText 等,将文本数据转换为低维向量表示,以便进行文本分类、情感分析等任务。

3.2 数据统计描述

数据统计描述是大数据分析与挖掘的基础,用于计算数据的统计及描述数据特征和规律,有助于了解数据分布情况。

3.2.1 数据集中趋势度量

数据集中趋势是指寻找反映事物特征的数据集合的代表值或中心值,可以很好地反映事物目前所处的位置和发展水平。通过对事物集中趋势指标的测量和比较,能刻画事物的发展和变化趋势。其主要目的是通过分析数据的统计特征,加深对数据的理解,从而利用合适的机器学习方法进行数据分析。数据描述统计包括 4 个主要部分:数据集中位置、离散程度、偏度和峰度以及单个数据变量的分布情况。

1. 算术平均值

算术平均值是一种衡量数据集中趋势的最常见方法。对于连续数据和离散数据,均值可以很好地反映数据的集中程度。此外,均值容易受到极端值的影响。如果数据中存在离群值时,均值不能很好地反映数据的中心趋势。

2. 众数

众数(Mode)是指数据集里面出现频率最高的数值。计算众数只需要统计数据中每个数值出现的次数,并找出出现次数最多的数值即可。如果数据集中存在多个数值出现次数相同且最多,则可以有多个众数。在 kNN 算法中,需要求近邻样本最多的标签,就需要用到众数。

3. 中位数

中位数(Median)是按顺序排列的一组数据中居于中间位置的数,代表一个样本、种群或概率分布中的一个数值。中位数可以将数值集合划分为相等的两部分。对于有限的数集,可以通过把所有观察值高低排序后,找出正中间的一个作为中位数。若观察值有偶数个,常取最中间的两个数值的平均数作为中位数。

4. K 百分位数

K 百分位数(K-Percentile)是指将一组数据按升序排列后,处于第 K 百分位置的数值,该数值表示有 $K\%$的数据小于或等于它。百分位数是统计学中用于描述数据分布位置的指标,它将数据分为 100 等份,每等份包含 1%的数据。通过计算 K 百分位数,可以了解

数据在某个特定位置的分布情况。

计算方法如下。

(1) 将数据从小到大排序。

(2) 计算第 K 百分位的位置：

$$P = \frac{K}{100} \times (n+1) \tag{3-7}$$

式(3-7)中，K 为百分位数(如 25、50、75 等)，n 为数据点的总数。如果 P 是整数，则第 K 百分位数为第 P 个数据点的值；如果 P 不是整数，则通过插值计算第 K 百分位数的值。常见的百分位数如图 3-1 所示。常见百分位数有第 1 四分位数(Q1)、中位数(Q2)、第 3 四分位数(Q3)。

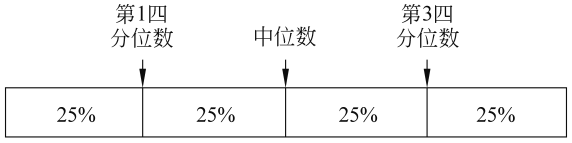


图 3-1　常见的百分位数

K 百分位数广泛用于描述数据的分布情况，特别是在分析数据的偏态和异常值时。例如，通过计算 Q2 和 Q3，可以了解数据的中间 50%的分布范围，从而评估数据的离散程度。

3.2.2　数据离散程度的度量

数据离散程度的度量是描述数据分散程度或变异程度的统计量。常见的数据离散程度度量包括极差、方差、标准差、四分位极差等。

1. 极差

极差为数据样本中的最大值与最小值的差值：

$$R = \max(x_i) - \min(x_i) \tag{3-8}$$

式(3-8)反映了数据样本的数值范围，是最基本的衡量数据离散程度的方式，是所有方式中最为简单的一种，但受极值影响较大。例如，在一次数据结构考试中，某班学生数据结构课程得分的极差为 50，反映了数据结构课程中考试分数最高的学生与得分最低的学生，得分差距为 50。

2. 四分位极差

四分位极差是第 3 四分位数(Q3)与第 1 四分位数(Q1)之间的差值，用于衡量数据中间 50%的分布范围。

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} \tag{3-9}$$

四分位极差不受极端值的影响，更能反映数据的中间分布情况。四分位极差被广泛用于数据分析中，特别是在需要识别和处理异常值时，例如，在金融数据分析中识别异常交易。

3. 五数概括和箱形图

五数概括包括数据的最小值、第 1 四分位数(Q1)、中位数(Q2)、第 3 四分位数(Q3)和

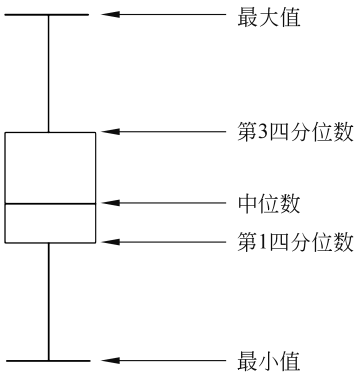


图 3-2 箱形图

最大值。箱形图是基于五数概括的一种可视化工具,用于直观展示数据的分布情况。箱形图如图 3-2 所示。

(1) 最小值:数据中的最小值。

(2) 第 1 四分位数(Q1):25%的数据小于或等于这个值。

(3) 中位数(Q2):50%的数据小于或等于这个值。

(4) 第 3 四分位数(Q3):75%的数据小于或等于这个值。

(5) 最大值:数据中的最大值。

若对称分布,则中位数位于箱子中间;若右偏分布,中位数更靠近第 1 四分位数;若左偏分布,则中位数更靠近第 3 四分位数。箱体的上下边界分别表示 Q3 和 Q1,箱体内的横线表示中位数,箱体外的"须"的长度通常是 1.5 倍的 IQR,超出"须"范围的点被认为是异常值。箱形图适用于直观展示数据的分布情况和异常值。

4. 方差和标准差

方差和标准差是衡量数据离散程度的常用指标,用于描述数据点与均值之间的偏离程度。

(1) 方差:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{3-10}$$

式(3-10)中,μ 是数据的均值,n 是数据点的总数。

(2) 标准差:

$$\sigma = \sqrt{\sigma^2} \tag{3-11}$$

方差和标准差考虑了所有数据点与均值的偏离程度,能够全面地反映数据的离散情况。标准差与原始数据具有相同的量纲,更易于解释。

5. 离散系数

离散系数,又称为差异系数,用于度量数据分散程度或波动程度,通过标准差和平均值的比值计算得出。离散系数越大,数据分散程度或波动程度越大;离散系数越小,数据分散程度或波动程度越小。离散系数可以消除数据量纲的影响,直接反映数值的波动范围。在比较不同数据集的离散程度时,使用离散系数可以更准确地评估数据的波动性。

3.2.3 数据相关性分析

数据相关性分析用于评估数据集中不同变量之间的关系,是数据探索和预处理的重要环节。通过相关性分析,可以识别出哪些变量之间存在显著的关系,从而为后续的数据分析和建模提供重要的依据。常见的相关性分析方法包括卡方检验、协方差和相关系数等。

1. 标称变量的卡方检验

卡方检验是一种统计方法，用于评估两个标称变量之间的独立性。通过计算观测频数与期望频数之间的差异，可以判断两个变量之间是否存在显著的相关性。

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3-12}$$

式(3-12)中，O_i 是观测频数，E_i 是期望频数。期望频数的计算方式为

$$E_i = \frac{\text{行总和} \times \text{列总和}}{\text{总样本数}} \tag{3-13}$$

卡方检验适用于标称变量，可以检测变量之间是否存在显著的关联。结果通常用 p 值表示，p 值越小，表示两个变量之间的相关性越显著。

2. 数值变量的协方差

协方差用于衡量两个数值变量之间的线性关系，计算方式为

$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \tag{3-14}$$

式(3-14)中，X_i 和 Y_i 分别是两个变量的第 i 个观测值，$\overline{X}$ 和 $\overline{Y}$ 分别是两个变量的均值，协方差的值可以为正、负或零，分别表示两个变量之间存在正相关、负相关或无相关。

协方差的值没有固定的范围，因此难以直接解释其大小，为正表示两个变量正相关，为负表示负相关，为零表示无相关。

3. 数值变量的相关系数

相关系数是衡量两个数值变量之间线性关系强度和方向的标准化指标。最常用的相关系数是皮尔逊相关系数(Pearson Correlation Coefficient，PCC)：

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3-15}$$

式(3-15)中，相关系数的 r 值介于 -1 和 1 之间，其中，1 表示完全正相关，-1 表示完全负相关，0 表示无相关。值越接近 1 或 -1，表示两个变量之间的线性关系越强。

想　一　想
本章学习了数据统计描述，这些方法可以在哪些场景中应用？

3.3　数据的相似性度量

在数据科学中，相似性度量是一种度量数据样本之间相互关联或紧密程度的方法，通常采用[0,1]区间的数值表示，值越大，数据样本越相似。相似性度量用于衡量两个对象之间

的相似性,这些对象可以是数字数据、图像、文本、语音或其他形式的信息。

3.3.1 基于距离的相似性度量

基于距离的相似性度量常用于衡量两个数据点之间的相似程度。该方法基于数据点间的距离来计算相似性,距离越近,相似性越高。在数据分析与机器学习领域,诸如闵可夫斯基距离在内的基于距离的相似性度量方法被广泛应用,这类方法通过量化数据点间的空间间隔来刻画样本相似程度,适用于聚类分析、分类建模、异常检测等多样化任务。在选择合适的距离度量方法时,需要考虑数据的特点和任务的需求。

1. 闵可夫斯基距离

闵可夫斯基距离是曼哈顿距离、欧几里得距离和切比雪夫距离的推广形式。对于 n 维空间中任意两个样本点 $\boldsymbol{P}=(x_1,x_2,\cdots,x_n)$ 和 $\boldsymbol{Q}=(y_1,y_2,\cdots,y_n)$,闵可夫斯基距离定义为

$$d(\boldsymbol{P},\boldsymbol{Q})=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}} \tag{3-16}$$

当 $p=1$ 时,闵可夫斯基距离变为曼哈顿距离;当 $p=2$ 时,闵可夫斯基距离变为欧几里得距离;当 $p\to\infty$ 时,闵可夫斯基距离变为切比雪夫距离。

2. 曼哈顿距离

当闵可夫斯基距离中的参数 $p=1$ 时,得到的就是曼哈顿距离。对于 n 维空间中的两个样本点 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,曼哈顿距离定义为

$$d(\boldsymbol{P},\boldsymbol{Q})=\sum_{i=1}^{n}|x_i-y_i| \tag{3-17}$$

曼哈顿距离表示在标准坐标系上的两点在水平和垂直方向上的绝对轴距总和。

3. 欧几里得距离

欧几里得距离是闵可夫斯基距离的一种特殊情况,当 $p=2$ 时,对于 n 维空间中任意两个样本点 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,欧几里得距离定义为

$$d(\boldsymbol{P},\boldsymbol{Q})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{3-18}$$

欧几里得距离是最直观的距离度量方法,适用于连续数据,在聚类分析(如 K-means 算法)和推荐系统中广泛使用。

4. 切比雪夫距离

当闵可夫斯基距离中的参数 p 趋于无穷大时,得到的就是切比雪夫距离。对于 n 维空间中的两个样本点 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,切比雪夫距离定义为

$$d(\boldsymbol{P},\boldsymbol{Q})=\max_i|x_i-y_i| \tag{3-19}$$

切比雪夫距离表示两个点在各个坐标维度上差值的最大值。它反映了在无限维空间中,两个点在最“远”的维度上的距离。

5. 汉明距离

汉明距离是一种用于度量两个等长字符串(或向量)差异程度的量化方法。具体而言，对于长度为 n 的等长字符串 S_1 和 S_2，通过逐位比较对应位置的字符，若某一位字符不同则计数加 1，最终累加的差异位数即为两者的汉明距离。其核心性质是：距离值越小，字符串的相似性越高；距离值越大，差异越显著。汉明距离的取值范围为[0，n]，其中，0 表示两个字符串完全一致，n 表示所有对应位均不相同。汉明距离在编码理论和生物信息学中广泛使用。

3.3.2　基于向量夹角的相似性度量

基于向量夹角的相似性度量常用于衡量两个向量之间的相似程度。该方法基于向量之间的夹角来计算相似性，夹角越小，相似性越高。常用的基于向量夹角的相似性度量方法包括余弦相似度、Jaccard 相似系数、皮尔逊相关系数和斯皮尔曼等级相关系数等。这些基于向量夹角的相似性度量方法可以用于各种数据分析和机器学习任务，如文本相似度计算、推荐系统、图像识别等。在选择合适的相似性度量方法时，需要考虑数据的特点和任务的需求。

1. 余弦相似度

余弦相似度(Cosine Similarity，CS)通过计算两个向量的夹角余弦值来评估它们的相似度。对于 n 维空间中的两个样本点 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$，余弦相似度定义为

$$r=\frac{\boldsymbol{P}\cdot\boldsymbol{Q}}{\|\boldsymbol{P}\|\cdot\|\boldsymbol{Q}\|}=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\times\sqrt{\sum_{i=1}^{n}y_i^2}} \tag{3-20}$$

余弦相似度的值域为[−1，1]。当两个向量的方向完全相同时，余弦相似度为 1；当两个向量的方向完全相反时，余弦相似度为−1；当两个向量正交时，余弦相似度为 0。在文本挖掘中，余弦相似度常用于衡量两个文本的相似性。例如，将文本转换为词向量，然后计算这些向量之间的余弦相似度，值越大表示文本内容越相似。

2. Jaccard 相似系数

Jaccard 相似系数(Jaccard Similarity Coefficient，JSC)用于衡量两个集合之间的相似性。对于两个集合 A 和 B，Jaccard 相似系数定义为

$$r=\frac{|A\cap B|}{|A\cup B|} \tag{3-21}$$

式(3-21)中，Jaccard 相似系数的值域为[0，1]，值越大表示集合越相似。Jaccard 相似系数在文本挖掘(如文档相似性)和生物信息学中广泛使用。

3. 皮尔逊相关系数

皮尔逊相关系数(Pearson Correlation Coefficient，PCC)用于衡量两个变量之间的线性

相关性,取值范围为[−1,1],其中,−1 表示完全负相关,1 表示完全正相关,0 表示无相关性。皮尔逊相关系数是通过计算两个变量 X 和 Y 的协方差与各自标准差的商来得到的,计算方式为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3-22}$$

式(3-22)中,x_i 和 y_i 分别是变量 X 和 Y 的数据点,$\bar{x}$ 和 $\bar{y}$ 分别是变量 X 和 Y 的平均值。

4. 斯皮尔曼等级相关系数

斯皮尔曼等级相关系数(Spearman's Rank Correlation Coefficient,SRCC)是一种非参数统计方法,用于衡量两个变量之间的单调相关性,不要求变量是连续的,取值范围同样为[−1,1]。该系数不要求数据满足正态分布,适用于各种类型的数据,包括非参数数据或等级数据,计算方式为

$$\rho=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{3-23}$$

式(3-23)中,d_i^2 是第 i 对观测值中变量 X 和 Y 的等级之差的绝对值,n 是观测值的总数。

3.3.3 基于概率和信息论的相似性度量

基于概率和信息论的相似性度量方法通过计算两个数据点之间的概率分布或信息量来评估它们的相似性。以下是几种常见的基于概率和信息论的相似性度量方法。

1. Kullback-Leibler 散度(KL 散度)

KL 散度用于衡量两个概率分布之间的差异。对于两个概率分布 P 和 Q,KL 散度定义为

$$D_{\mathrm{KL}}(P\parallel Q)=\sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{3-24}$$

式(3-24)中 KL 散度的值域为[0,+∞),值越小表示两个分布越相似。KL 散度在信息检索和机器学习中广泛使用。

2. 互信息

互信息(Mutual Information,MI)用于衡量两个随机变量之间的相互依赖性。对于两个随机变量 X 和 Y,互信息定义为

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{P(x,y)}{p(x)P(y)} \tag{3-25}$$

式(3-25)中互信息的值域为[0,+∞),值越大表示两个变量之间的依赖性越强。互信息在特征选择和信息检索中广泛使用。

想　一　想
本章学习了数据相似性度量，在使用这些技术时应该注意什么？可以用到哪些领域中？

3.4　数据清洗

数据清洗是数据预处理的重要步骤，旨在识别并纠正数据中的错误、缺失值、重复数据、不一致性和异常值等问题，以提高数据的质量和可用性。

3.4.1　缺失值处理方法

缺失值处理是数据清洗中的重要步骤，涉及识别和处理数据集中缺失的数据点。需要注意的是，在某些情况下缺失值并不是数据有错误。以下是一些常用的缺失值处理方法。

1. 删除缺失值

删除缺失值包含完全删除和按列删除。如果缺失值的数量相对较少，并且删除这些记录不会对分析结果产生重大影响，可以选择直接删除包含缺失值的记录。如果某一列的缺失值比例过高，可以考虑删除该列。

2. 填充缺失值

均值填充：对于数值型数据，可以使用该列的均值来填充缺失值。

中位数填充：对于存在异常值的数值型数据，中位数可能是一个更好的选择。

众数填充：对于分类数据，可以使用该列的众数(出现频率最高的值)来填充缺失值。

固定值填充：可以使用一个固定的值(如 0 或－1)来填充缺失值。

插值填充：根据已知数据点的值，通过插值方法(如线性插值、多项式插值等)来估计缺失值。

模型预测填充：使用机器学习模型(如回归模型、决策树模型等)来预测缺失值。

3. 多重填补

生成多个完整的数据集，每个数据集对缺失值进行不同的填充，然后对这些数据集进行分析，最后综合结果。

4. 不处理

在某些情况下，缺失值本身可能包含有价值的信息，或者缺失值的处理可能会引入偏差，此时可以选择不处理缺失值。

在实际应用中，选择哪种缺失值处理方法取决于数据的特点、缺失值的分布、分析目的以及对结果的影响等因素。通常需要尝试多种方法，并通过交叉验证等技术评估不同方法的性能，以选择最合适的处理方式。

3.4.2 噪声数据处理

噪声(Noise)是指数据中因测量误差或随机波动产生的非系统性干扰,表现为目标变量观测值与真实值之间的随机偏差。作为数据预处理的关键环节,噪声数据处理旨在通过系统性方法识别并修正数据中的错误值、异常点或不一致信息,从而提升数据的质量与可靠性。常见的噪声数据处理方法包括数据清洗、数据转换、异常值检测和处理、数据平滑和数据验证等方法。

1. 数据清洗

数据清洗方法主要包括缺失值处理、重复值处理、错误值处理等方法。

缺失值处理方法主要识别并处理数据集中的缺失值,可以使用删除、插补或预测等方法。重复值处理方法主要识别并删除数据集中的重复记录。错误值处理方法主要识别并纠正数据集的错误值,可以使用手动修正、基于规则的修正或模型预测等方法。

2. 数据转换

数据转换方法主要包括标准化、归一化和对数变换方法等。标准化方法将数据转换为均值为 0、标准差为 1 的标准正态分布,以消除数据的量纲和单位差异。归一化方法将数据缩放到一个特定的范围,如[0,1]或[−1,1],以消除数据的量纲和单位差异。对数变换方法通过数据取对数,以减少数据的偏度和峰度,使数据更接近正态分布。

3. 异常值检测和处理

异常值检测和处理主要包括基于统计的方法、基于距离的方法、基于密度的方法和基于模型的方法。基于统计的方法主要使用均值、标准差、四分位数等统计指标来检测和处理异常值。基于距离的方法主要使用距离度量(如欧氏距离、曼哈顿距离等)来检测和处理异常值。基于密度的方法主要使用数据点的密度来检测和处理异常值。基于模型的方法主要使用机器学习模型(如聚类、分类、回归等)来检测和处理异常值。

4. 数据平滑

数据平滑方法主要包括移动平均、加权移动平均和指数平滑等处理技术。数据平滑方法主要使用滑动窗口计算数据的平均值,以平滑数据的波动。加权移动平均方法主要对移动平均中的每个数据点赋予不同的权重,以更好地反映数据的趋势。指数平滑方法主要使用指数加权平均来平滑数据的波动,其中较新的数据点具有更高的权重。

5. 数据验证

数据验证方法主要验证数据的完整性、准确性和一致性。完整性主要检查数据集中的记录是否完整,是否存在缺失值、重复值或错误值。准确性主要检查数据集中的记录是否准确,是否与实际情况相符。一致性主要检查数据集中的记录是否一致,是否符合业务规则和逻辑。

在实际应用中,选择合适的噪声数据处理方法取决于数据的特点、噪声的类型和分布、

分析的目的以及对结果的影响等因素。通常需要尝试多种方法，并通过交叉验证等技术来评估不同方法的效果，以选择最合适的处理方式。

3.4.3　异常值处理

在数据集中，异常值是指那些与其他数据明显不一致的数值。它们可能是由于数据录入错误、测量误差、数据传输错误等原因产生的。异常值的检测方法如下。

1. 统计学方法

1）标准差法

原理：假设数据服从正态分布，那么大部分数据（约 99.7%）会落在均值（μ）加减 3 倍标准差（σ）的区间内。如果某个数据点超出了这个范围，就可以认为它是异常值。首先计算数据集的均值和标准差。然后确定异常值的判断区间为（$\mu-3\sigma$，$\mu+3\sigma$）。在这个区间之外的数据就是异常值。

适用场景：适用于数据分布较为接近正态分布的情况。

2）四分位数法

原理：首先将数据集按升序排列并均分为 4 等份，其中处于 25%位置的数值定义为第 1 四分位数（Q1），处于 75%位置的数值为第 3 四分位数（Q3），二者的差值即为四分位距（IQR＝Q3－Q1）。通常认为，数据的正常值分布于区间（Q1－1.5IQR，Q3＋1.5IQR）内，超出该范围的观测值会被视为异常值。该方法的核心操作流程为：先对数据排序，计算 Q1 和 Q3 以确定 IQR，再通过上述区间界定筛选出偏离整体分布的异常数据点。

适用场景：适用于各种分布的数据，尤其是当数据分布不是正态分布时。

2. 可视化方法

1）箱形图

原理：箱形图是基于四分位数法的可视化表示。它通过绘制一个箱子（箱子的上下边界分别代表 Q3 和 Q1），箱子中间的横线表示中位数（Q2）。箱子外的“胡须”长度通常是 1.5IQR，超出“胡须”范围的点就是异常值。使用 Matplotlib 等数据可视化工具绘制箱形图。

适用场景：适用于直观展示数据的分布情况和异常值。

在数据分析的初步阶段，通过箱形图可以快速发现数据中的异常情况。箱形图结构如图 3-3 所示。

2）散点图

原理：将数据点在二维坐标系中绘制出来。如果数据存在异常值，这些异常值会在散点图中明显偏离其他数据点的分布趋势。例如，对于一组包含两个变量（x 和 y）的数据，使用 Matplotlib 库绘制散点图，在生成的散点图中，观察偏离其他点的线性分布趋势，从而可以判断它是异常值。

适用场景：适用于多变量数据的异常值检测。

【例 3-1】 某年级 10 位学生的身高、体重如表 3-1 所示。

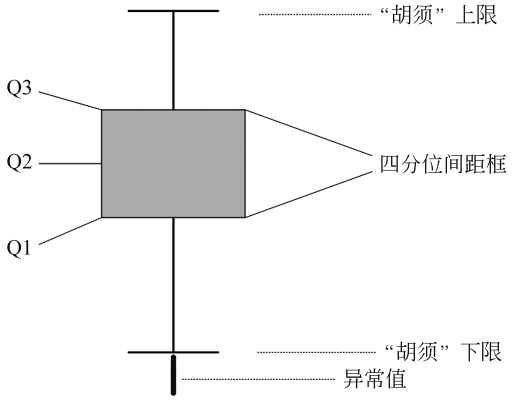


图 3-3　箱形图结构

表 3-1　某年级 10 位学生的身高、体重

学生	身高/cm	体重/kg
1	165	62
2	172	68
3	180	75
4	158	55
5	175	70
6	168	63
7	185	82
8	155	52
9	178	73
10	160	58

对应的散点图如图 3-4 所示。

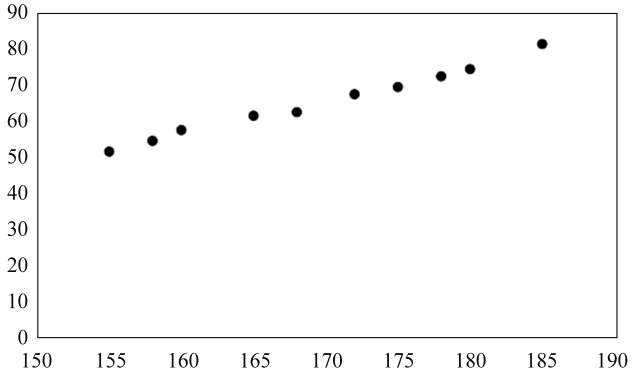


图 3-4　某年级 10 位学生的身高、体重散点图

3. 异常值的处理方法

异常值的处理方法主要包括删除异常值、修正异常值和保留异常值等。

1）删除异常值

当异常值是由于明显的错误(如数据录入错误)产生的,并且这些异常值的数量很少,对整体数据分析结果影响不大时,可以考虑删除。删除异常值可能会导致数据量减少,从而影响数据分析的全面性和准确性。如果异常值占数据集的比例较大,删除后可能会使数据失去代表性。例如,在一个医学研究数据集中,异常值可能代表了一些特殊病例,如果轻易删除,可能会遗漏重要的医学信息。

2）修正异常值

当异常值可能是由于测量误差或者数据传输过程中的小误差产生的,并且可以推测出异常值的正确范围或者可能的值时,可以进行修正。例如,在气象数据中,温度记录出现了一个异常高的值,通过对比周边气象站的数据和历史数据,可以推测出这个异常值可能是传感器误差导致的,可以将其修正为一个合理的值。

3）保留异常值

当异常值具有特殊的意义或者代表了某种重要的信息时,应该保留。

此外,可以对异常值进行标记,以便在后续数据分析中能够关注异常值。例如,可以在数据表中增加一列数据“是否异常”标记,对于异常值标记为“是”,正常值标记为“否”,进而对异常值进行单独分析或者在整体分析中给予不同的权重。

3.5　数据规范化和编码


视频讲解

数据变换作为数据预处理的关键步骤,是指对原始数据实施一系列处理与转换,通过调整数据的分布形态、组织结构或表达形式,使其更适配后续的数据分析、建模及可视化等任务。

3.5.1　数据规范化

数据规范化处理是数据挖掘领域的基础性工作,其核心目的在于解决不同评价指标因量纲和测量单位差异对数据分析产生的干扰。由于各类指标常以不同的度量标准(如长度用厘米、质量用千克)或数值范围(如百分比与绝对值)呈现,直接纳入分析会导致量纲较大的指标对结果产生不合理的主导作用。为了消除指标之间的量纲影响,需要进行数据标准化处理。数据规范化是通过比例缩放的方式将数据映射到特定数值区间(如[0,1]或[−1,1])的预处理技术,其核心是消除不同变量间量纲和取值范围的差异,使数据具备统一的比较尺度。常见的规范化方法包括最小-最大规范化、Z-score 规范化和小数定标规范化等。

(1) 最小-最大规范化：假定 m_A 和 M_A 分别为属性 A 的最小值和最大值,最小-最大规范化可通过下式计算：

$$v' = \frac{v - m_A}{M_A - m_A}(\text{new\_M}_A - \text{new\_m}_A) + \text{new\_m}_A$$

将 A 的值 v 映射到区间[new_m_A,new_M_A]中。最小-最大规范化对原始数据进行线性变换,保持原始数据值间的联系。如果数值范围在 A 的原始数据值域之外,该方法将面临“越界”错误。以下是一个使用 Python 实现最小-最大规范化的示例代码。

```
import numpy as np
def min_max_normalization(data):
    """
    对数据进行最小-最大规范化
    参数:
    data (numpy.ndarray): 需要进行规范化的数据
    返回:
    numpy.ndarray: 规范化后的数据
    """
    #计算最小值和最大值
    min_val = np.min(data)
    max_val = np.max(data)
    #进行最小-最大规范化
    normalized_data = (data - min_val) / (max_val - min_val)
    return normalized_data
#示例用法
data = np.array([1, 2, 3, 4, 5])
normalized_data = min_max_normalization(data)
print("原始数据:", data)
print("规范化后的数据:", normalized_data)
```

(2) Z-score 规范化(零均值规范化):把属性 A 的值 v 基于 A 的均值 $\overline{A}$ 和标准差规范化为 v',可由下式计算:

$$v' = \frac{(v - \overline{A})}{\sigma_A}$$

其中,$\overline{A}$ 和 σ_A 分别为属性 A 的均值和标准差。当属性 A 的真实极值(最大值与最小值)无法获取,或数据中存在的离群点导致最小-最大规范化所依赖的极值显著偏离正常分布范围时,该方法是有用的。使用 Python 实现 Z-score 规范化(零均值规范化)的示例代码如下。

```
import numpy as np
def z_score_normalization(data):
    """
    对数据进行 z-score 规范化(零均值规范化)
    参数:
    data (numpy.ndarray): 需要进行规范化的数据
    返回:
    numpy.ndarray: 规范化后的数据
    """
    #计算均值和标准差
    mean = np.mean(data)
    std = np.std(data)
```

```
    #进行 z-score 规范化
    normalized_data = (data - mean) / std
    return normalized_data
#示例用法
data = np.array([1, 2, 3, 4, 5])
normalized_data = z_score_normalization(data)
print("原始数据:", data)
print("规范化后的数据:", normalized_data)
```

(3) 小数定标规范化：通过移动属性 A 的小数点位置进行规范化。小数点的移动位数依赖于 A 的最大绝对值。A 的值 v 规范化为 v'，由下式计算。

$$v' = \frac{v}{10^j}$$

其中，j 是使得 Max($|v'|$)<1 的最小整数。

数据变换策略的选择应根据数据的特点、分析的目的及项目所使用的算法来决定。实际的项目和应用通常需要尝试多种变换策略，并使用交叉验证等方法来评估数据变换后的效果，以选择最佳的数据变换策略。此外，对数变换(Log Transformation，LT)可以将数据的分布从偏态转换为正态分布，减少数据的偏度和峰度；平方根变换(Square Root Transformation，SRT)可以用于减少数据的偏度和峰度；Box-Cox 变换可以自动寻找最佳的变换参数，进而将数据转换为正态分布。以下是一个使用 Python 进行最小-最大规范化的示例代码。

```
import pandas as pd
#生成示例数据
data = pd.DataFrame({
    '年龄': [25, 30, 35, 40, 45, 50, 55, 60, 65, 70],
    '收入': [5000, 6000, 7000, 8000, 9000, 10000, 11000, 12000, 13000, 14000]
})
#最小-最大规范化
data['年龄_规范化'] = (data['年龄'] - data['年龄'].min()) / (data['年龄'].max() -
data['年龄'].min())
data['收入_规范化'] = (data['收入'] - data['收入'].min()) / (data['收入'].max() -
data['收入'].min())
print(data)
```

3.5.2　数据编码

在数据预处理过程中，数据编码与转换是将数据从一种格式或类型转换为另一种格式或类型的重要步骤，同时是数据预处理中不可或缺的一部分，能够显著提升数据的价值和模型的性能。以下是常见的数据编码与转换方法。

1. 基于频率的编码

基于频率的编码(Frequency Encoding，FE)是一种将标称属性的每个类别值替换为其

在数据集中出现频率的方法。这种方法可以保留类别值的分布信息,对于某些机器学习算法非常有用。例如,假设“颜色”这一标称属性的类别值红色、绿色和蓝色在数据集中的频率分别为 0.4、0.3 和 0.3,那么这些类别值将分别替换为 0.4、0.3 和 0.3。基于频率的编码特别适用于处理类别值数量较多且分布不均匀的属性。

2. 目标编码

目标编码(Target Encoding,TE)是将标称属性的每个类别值替换为目标变量的统计值(如均值、中位数等)。这种方法可以将类别值与目标变量的分布联系起来,适用于分类和回归问题。例如,假设“颜色”这一标称属性的类别值红色、绿色和蓝色对应的目标变量(如价格)的平均值分别为 100、200 和 300,那么这些类别值将分别替换为 100、200 和 300。目标编码可以提高模型的性能,但需要注意防止过拟合。

3. 日期和时间编码

日期和时间编码(Date and time Encoding,DE)将日期和时间数据转换为数值型的过程。常见的转换包括提取年份、月份、星期、小时等信息。例如,假设日期属性为“2025-1-11 12:30:00”,可以提取出年份 2025、月份是一月、星期六等特征。日期和时间编码可以提取时间序列数据中的周期性和趋势信息,适用于时间序列分析和预测。

通过数据编码与转换,可以将数据转换为适合特定任务进行分析或挖掘的形式,进而提高数据的质量和模型的性能。选择合适的数据编码与转换方法取决于数据的特性和分析任务的需求。这些方法在数据预处理中具有重要的应用价值,可以帮助我们更好地准备数据,为后续的数据分析和挖掘提供支持。例如,对于分类问题,独热编码和目标编码可能是更好的选择;而对于时间序列数据,日期和时间编码则更为重要。

3.6 数据归约

借助数据归约技术可将原始数据集转换为精简的归约表示 —— 这类表示虽然数据规模显著小于原数据集,但能最大限度地保留数据的核心信息与完整性。常见的数据归约方法涵盖多个维度:通过数据立方体聚集实现数据的汇总与聚合;利用维归约(如主成分分析、特征选择)减少冗余特征;借助数据压缩技术(如无损/有损编码)实现数据体量的缩减;运用数值归约(如回归模型、直方图)近似原始数据分布;通过数据离散化与概念分层将连续数据转换为离散区间或抽象层级。这些技术通过不同策略在数据规模与信息保留间取得平衡,为高效的数据存储、处理和分析提供支持。

3.6.1 数据立方体聚集

数据立方体聚集是一种在数据仓库中常用的技术,用于对多维数据进行汇总和聚合。它通过对数据进行分组和计算,生成一个多维的汇总表,称为数据立方体。数据立方体可以帮助用户快速地获取数据的汇总信息,且不需要对原始数据进行复杂的查询和计算。数据