第二章 数据基本知识

数据分析是智能制造的核心技术之一，是提高制造业核心能力、整合产业链的核心手段，对于智能制造具有重要意义。数据是分析的基础，工业过程中数据来源广泛、类型繁多、形式各异，全面了解和认识不同数据的性质是数据分析的前提。

本章首先从数据属性角度出发，介绍数据的基本概念；其次，介绍数据的基本统计描述，主要包括数据的集中趋势度量和离散趋势度量，以把握数据的整体情况；再次，借助于图形可视化手段，清晰有效地展现数据整体分布及不同维度间的关联关系；最后，着重介绍不同属性数据间的相似性度量方法，以及通过相关性分析指标来衡量数据关联的密切程度。

2.1 数据的基本概念

数据产生于对客观事物的观察与测量，通常把被研究的对象称为实体，实体可以通过各种可观测的属性（或称为特征）来描述。例如，选矿过程中矿石有质量、颜色等属性，高炉炼铁过程中铁水有温度、颜色等属性。这些属性有时也称为变量，而这些变量的取值就是数据。

属性是一个数据字段，表示数据对象的一个特征。属性、维、特征和变量可以互换地使用，数据挖掘和数据库一般使用术语“属性”，“维”一般用在数据库中，机器学习中倾向于使用术语“特征”，而统计学中使用术语“变量”，本章统一使用属性这一描述方式。例如，描述高炉炼铁煤气利用率的属性可能包括风量、风压、富氧等。属性的类型由该属性集合的内容决定，属性的类型有标称类型、二元类型、序数类型以及数值类型等。

2.1.1 标称属性

标称属性（Nominal Attribute）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，标称属性又可以被看作是分类，与属性集合内容的排列顺序无关。

以高炉炼铁产物为例，有生铁、除尘灰、水渣、高炉煤气等产物。这里的炼铁产物就是标称属性，而生铁、除尘灰、水渣、高炉煤气就是其属性值。标称属性的值也可用数字表示，例如，对于主要产物生铁，可以指定数字 0 表示，而产物除尘灰可以用数字 1 表示，水渣用数字 2 表示。

2.1.2 二元属性

二元属性（Binary Attribute）是标称属性的一种特殊形式，只有两个类别或状态。例如，0 或 1，其中 0 通常表示该属性不出现，而 1 表示出现。依据二元属性的两种状态是否具有同等价值或相同权重，可以分为对称二元属性和非对

称二元属性。对于对称的二元属性，它的两种状态具有同等价值和相同权重，即哪个是 0 或 1 编码并无偏好。而对于非对称的二元属性，两种状态不再同等重要，会出现某一状态样本比较稀有的情况。

以高炉炼铁过程中的出铁质量检测结果为例，假设检测结果具有两种可能，值 1 表示检测结果合格，0 表示不合格；以氧气调节阀的开合为例，值 1 可以表示调节阀的打开状态，值 0 可以表示调节阀的关闭状态。

2.1.3 序数属性

序数属性（Ordinal Attribute）的值之间具有有意义的序或秩次（秩次是指对给定的一组数据按照数值从小到大的顺序进行排序，所得到的每一个数据的排序号），但是相互间的差值是未知的，序数属性通常用于等级评定中。序数属性的中心趋势可以用它的众数和中位数（有序序列的中间值）表示，但不能定义均值。

例如，高炉炼铁中的透气性指数是一个可以快速、直观、综合反映炉况的重要参数，用来表示炉子接受风量的能力。假设透气性指数有 5 个可能的值——极高、偏高、良好、偏低和极低，这些值的先后次序（对应于递增的透气性）具有意义，但是不能说“偏高”比“良好”大多少。序数属性的其他例子包括热风压力（例如，高压、过压、良好、低压和极低压等）。

这里要注意，标称、二元和序数属性都是定性描述对象的特征，不给出实际大小或数量。这种定性属性的值通常是代表类别的，如果使用整数，则代表类别编码，而不是测量值属性（例如，1 表示偏高，2 表示良好，3 表示偏小）。

2.1.4 数值属性

数值属性（Numerical Attribute）是可度量的变量，用整数或实数值表示，数值属性可以分为区间标度属性和比率标度属性。

（1）区间标度属性

区间标度属性（Interval Scaled Attribute）用等量单位尺度度量，属性的值是有序的，可以为正数、零或负数。已知温度是区间标度属性，取每一个小时内的炼铁高炉内平均温度作为一个样本，并将温度值进行排序，可以得到一定时间段内的温度值的序列。对于区间标度属性值而言，不存在真正的零点，例如区间标度属性摄氏温度和华氏温度中，0℃和 0℉都不表示“没有温度”（摄氏温度的度量单位是水在标准大气压下沸点温度与冰点温度之差的 1/100）。对于区间标度属性，差值是有意义的，例如，炼铁温度 1200℃比 600℃高出 600℃，但不能说 1200℃是 600℃的 2 倍，因为摄氏温度没有真正的零度。类似地，日期也没有绝对的零点（公元元年并不对应于时间的开始）。

（2）比率标度属性

比率标度属性（Ratio Scale Attribute）是具有零点的数值属性。如果度量是比率标度属性，则度量数据间的差和比率都是有意义的，此外，这些数据具有有序性，还可以计算值之间的均值、中位数和众数等在内的多个参数。

开氏温度属于比率标度属性，与摄氏和华氏温度不同，开氏温标（K）具有绝对零点（K=-273.15℃），且在该点，构成物质的粒子具有零动能。在炼铁过程中，其他比率标度属性的例子还有铁水质量、铁水密度和热风压力等，均可以计算差和比率等在内的多个有意义的量。

2.1.5 连续属性与离散属性

前面介绍的标称、二元、序数和数值类型属性，相互间并不互斥。按属性的取值是否为连续数值，通常把属性分成离散属性（Discrete Attribute）和连续属性（Continuous Attribute），每种类型对应不同的数据处理方法。连续属性是指在一定区间内可以任意取值，其数值是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。离散属性是指在一定区间内只能取有限个值，并且取值可以一一列举出来。

如关于对炼铁过程的描述中，连续属性有富氧率、炉温、冷风压力、热风压力以及流量等，离散属性有日出铁次数、料批数和高炉风口个数等。

2.1.6 时间序列

时间序列（Time Series）是指在不同时间收集到的数据，这类数据是按时间顺序收集到的，用以描述现象随时间变化的情况。通过对时间序列进行分析，可以发现某个现象发展变化的趋势和规律，为预测和决策提供可靠的数据支持。根据观察时间的不同，时间序列中的时间标度可以是年份、季度、月份或其他任何时间形式。对于炼铁过程，如冷风压力、热风压力、冷风流量以及热风流量等特征量，是由传感器按照一定的时间尺度（如小时、分钟、秒等）采集所得，均为时间序列。

2.2 数据的基本统计描述

制造过程中产生的数据数量庞大，在获得工业生产数据后，首先需要对数据基本统计特征进行观测，以把握数据的整体分布状态，识别数据的性质。

本节讨论两类基本的统计描述，分别为集中趋势度量和离散趋势度量。其中，集中趋势度量用来度量数据分布的中部或者中心位置，包括均值、中位数、众数和中列数等。除了估计数据集的中心趋势之外，还需要知道数据的离散状态。最常见离散趋势度量包括数据的极差、平均差、方差、标准差、协方差、四分位数以及数据的离散系数等。对于识别离群点，这些度量是有用的。此外，本节给出了统计度量的图形展示，主要包含折线图、直方图、条形图、箱线图、散点图以及平行坐标图。

2.2.1 集中趋势度量

集中趋势度量反映了整个数据样本的集中或紧密程度。本小节主要讨论评估数值数据集中或收敛的度量，这些度量包括均值、中位数、众数和中列数等。

（1）均值

均值（Mean）用于反映现象总体的一般水平，表示一组数据集中趋势的量数。令 $x_1, x_2, \cdots x_N$ 为某个数据集的 N 个观测值，则该值集合的均值 $\bar{x}$ 为：

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots x_N}{N} \tag{2.1}$$

有时，对于 $i = 1, 2, \cdots, N$，考虑到不同观测值对于现象总体的重要性或出现频率不同，因此引入权重 w_i 作为每个值 x_i 的系数。在这种情况下，可以计算加权算术平均：

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \tag{2.2}$$

尽管均值是描述数据集的最有用的单个量，但是均值对极端值（例如离群点）很敏感，并非总是度量数据中心的最佳方法。例如，一段时间内的平均冷风压力可能被几个低极端值（可能是异常值）拉低一些。为了抵消少数极端值的影响，我们可以使用截尾均值，也就是丢弃高低极端值后的均值，例如，计算比赛计算得分时，一般会去除一个最高分和一个最低分，再计算剩余得分的平均值作为真实成绩得分。此外，使用截尾均值时应该避免在两端截去太多（例如按照升序排列，两段各去掉总体数据量的 10%），因为这可能导致丢失有价值的信息。

（2）中位数

对于分布不均匀的数据，中位数（Median）可以用来较好的度量数据中心。中位数是指顺序排列的一组数据中居于中间位置的数，它将数据较高的一半与较低的一半分开。

假设给定某属性 X 的 N 个值按递增顺序排序，如果 N 是奇数，则中位数是该有序集的中间值；如果 N 是偶数，则中位数不唯一，它是最中间的两个值和它们之间的任意值，一般取作最中间两个值的平均数。

（3）众数

众数（Mode）是指在数据集中具有明显集中趋势点的数值，即是一组数据中出现次数最多的数值。此外，一组数据集中可能最高频率对应多个不同值，

导致出现了多个众数，一般地，具有一个众数的数据集合称为单峰数据集，具有两个或更多众数的数据集称为多峰数据集。在另一种极端情况下，如果每个数据值仅出现一次，则数据集没有众数。

（4）中列数

中列数（Midrange）也可以用来评估数值数据的中心趋势，中列数是数据集的最大和最小值的平均值。如集合{2, 4, 8, 9, 1, 3, 5}，它的中列数即为(1+9)/2=5。

2.2.2 离散趋势度量

离散趋势度量反映了整个数据样本远离中心值的趋势。本节主要讨论评估数值数据散布或发散的度量，这些度量包括极差、平均差、方差、标准差、协方差、四分位数和离散系数等。

（1）极差

令 $x_1, x_2, \cdots, x_N$ 为某个数值属性上的 N 个观测样本，该集合的极差（Range）是最大值 $\max(x_1, x_2, \cdots, x_N)$ 与最小值 $\min(x_1, x_2, \cdots, x_N)$ 之差。

（2）平均差

平均差（Average Deviation）指各个变量值同平均数间差值绝对值的算术平均数，平均差反应了各变量值与算术平均数之间的平均差异，平均差 A_D 计算如下：

$$A_D = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}| \tag{2.3}$$

（3）方差与标准差

方差（Variance）和标准差（Standard Deviation）是测算数据离散趋势最重要、最常用的指标。其中，方差是各变量值与其均值离差平方的平均数，标准差则为方差的算术平方根。低方差和者低标准差意味数据整体观测趋向于非常靠近均值，而高方差和高标准差表示整体数据分布较分散。方差 σ^2 计算如下：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{2.4}$$

其中，$\bar{x}$ 是观测值的均值，观测值的标准差 σ 是方差 σ^2 的平方根。

标准差是关于均值的发散，仅当选择均值作为中心度量时使用。当数据不存在发散，即所有的观测值都具有相同值时，$\sigma = 0$；否则，$\sigma > 0$。观测值一般不会远离均值超过标准差的数倍，标准差是衡量数据是否发散的良好指示器。

（4）协方差

协方差（Covariance）用于衡量两个变量（维度）的总体误差。如果两个变

量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。期望值分别为$E[\boldsymbol{x}]$与$E[\boldsymbol{y}]$的两个实随机变量$\boldsymbol{x}$与$\boldsymbol{y}$之间的协方差$\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})$定义为：

$$\begin{aligned}\mathrm{cov}(\boldsymbol{x},\boldsymbol{y}) &= E\left[\left(\boldsymbol{x}-E(\boldsymbol{x})\right)\left(\boldsymbol{y}-E(\boldsymbol{y})\right)\right] \\ &= E[\boldsymbol{xy}]-E[\boldsymbol{x}]E[\boldsymbol{y}]\end{aligned} \tag{2.5}$$

（5）四分位数

四分位数（Quartile）也称四分位点，是指把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值。将所有数据从小到大排列，其中，处在25%位置上的数值称为下四分位数Q_1，中位数则称为Q_2，处在75%位置上的数值称为上四分位数Q_3。此外，定义Q_1和Q_3间的距离为四分位极差（Interquartile Range, IQR），即IQR=Q_3-Q_1。图2.1展示了μ为10、σ^2为1的正态分布数据的四分位数图。

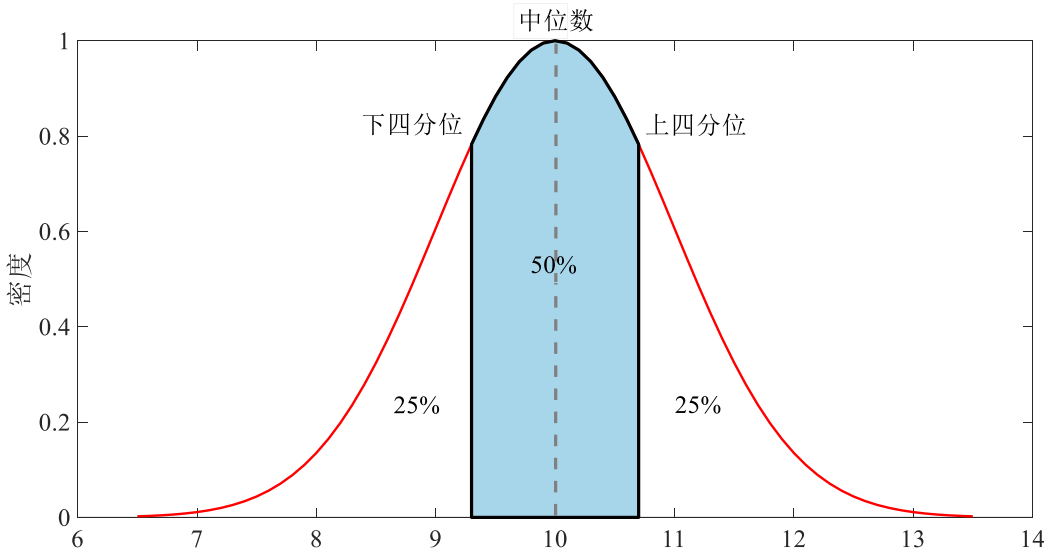


图2.1 四分位数图

（6）离散系数

离散系数（Coefficient of Variation）是度量数据离散程度的相对统计量。离散系数通常可以进行多个变量离散程度的对比，通过离散系数大小的比较可以说明不同总体平均指标（一般来说是平均数）的代表性或稳定性大小。离散系

数C_v的计算公式如下：

$$C_v = \frac{\sigma}{\bar{x}} \tag{2.6}$$

离散系数越大表示数据分布越分散，离散系数越小表示数据分布越稳定。

2.2.3 数据基本统计的图形描述

图形描述有助于直观清晰地观察出数据的变化趋势及分布状况，可以更好地服务于数据预处理，本节将讨论数据基本统计的图形描述。

（1）折线图

折线图可以显示随时间变化的连续数据，因此非常适用于显示在相等时间间隔下数据的趋势，通常横轴用来表示时间轴，纵轴用来表示实际过程中的特征参数。

图 2.2 展示了某炼铁厂 120 小时之内（采样周期为 2 个小时）的煤气利用率变化折线图，可以看出在该时间范围内，煤气利用率呈现出周期波动趋势，且波动周期约为 24 小时。

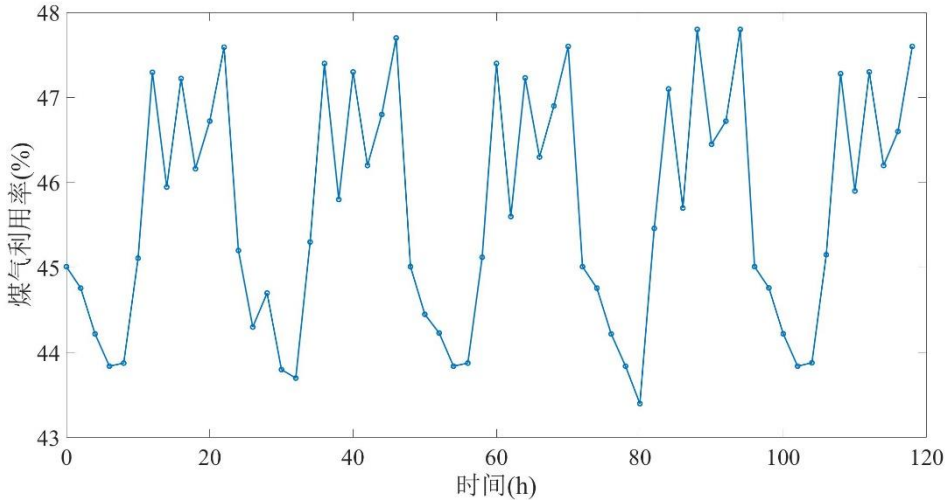
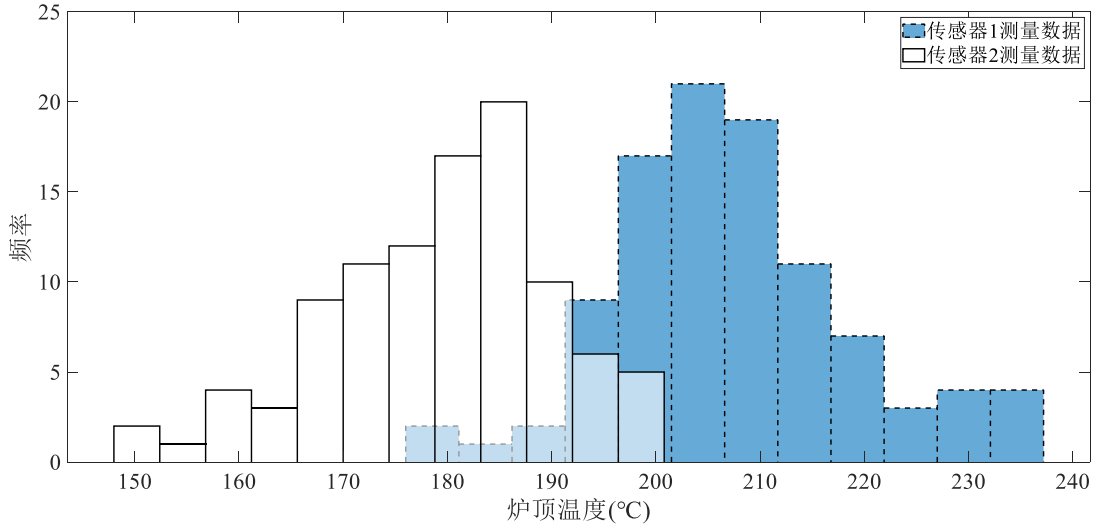


图 2.2 煤气利用率变化折线图

（2）直方图

直方图由一系列高度不等的纵向条纹来表示数据分布的情况，一般用横轴表示数据分布区间，纵轴表示对应范围内数据出现的频率。用直方图可以清晰观察出数据的频率分布情况，便于判断其总体分布情况。

图 2.3 展示了来自两个不同位置温度传感器测量得到的炉顶温度分布直方图，可以看到尽管两者测量的都是加热炉炉顶温度，但由于安装位置不同，测



量得到的温度分布还是有一定的差别。

图 2.3 不同位置传感器测量所得炉顶温度分布直方图

（3）条形图

条形图是用等宽条形的高度或长短来表示数据多少的图形，条形图可以横置或纵置，纵置时也称为柱形图。此外，条形图有简单条形图、复式条形图等形式。简单条形图可以展示一维分类数据的数据分布情况，而复式条形图可以展现两维及以上分类数据的分布情况。

图 2.4 展示了炼铁过程热风炉冷风压力与热风压力在12 h 内的条形图，在该时间段内，可以清晰明了地看出数据的数值，比较出两个变量数据的大小。

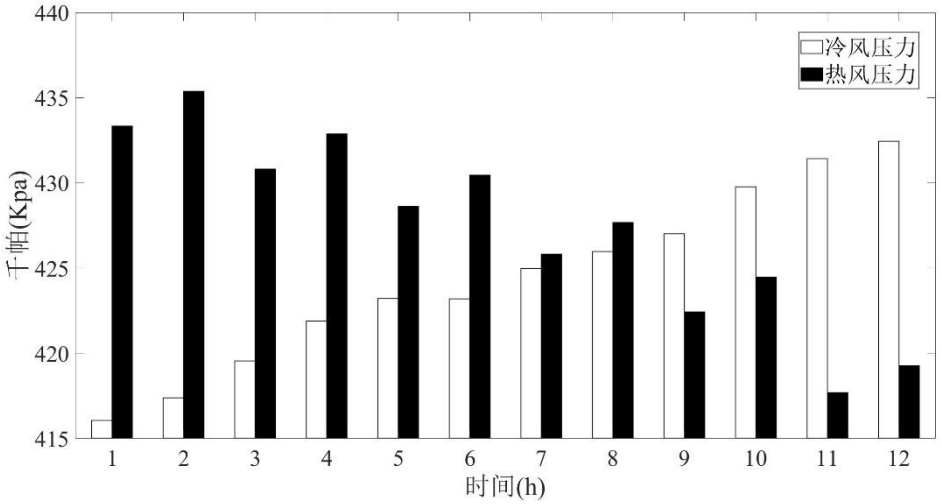


图 2.4 炼铁过程热风炉冷风压力与热风压力条形图

（4）箱线图

箱线图利用了五个统计特征对状态数据分布进行总结，用作显示一组数据的离散分布情况，这 5 个统计特征包括中位数Q_2、下四分位Q_1、上四分位Q_3、分布状态的上限和下限。其中，分布状态的上限和下限取法不唯一，比如：可以分别取作整体数据的最大及最小值，也可取作$Q_3+1.5*(Q_3-Q_1)$及$Q_1-1.5*(Q_3-Q_1)$。箱线图最大的优点就是不受异常值的影响，可以以一种相对稳定的方式描述数据的离散分布情况。

图 2.5 展示了 A、B、C、D 四组炉顶温度数据的箱线图，其中横轴表示不同的数据集或组别，纵轴代表数据取值，在该箱线图中可以清晰看出不同组数据的四分位数以及离群点的分布情况。

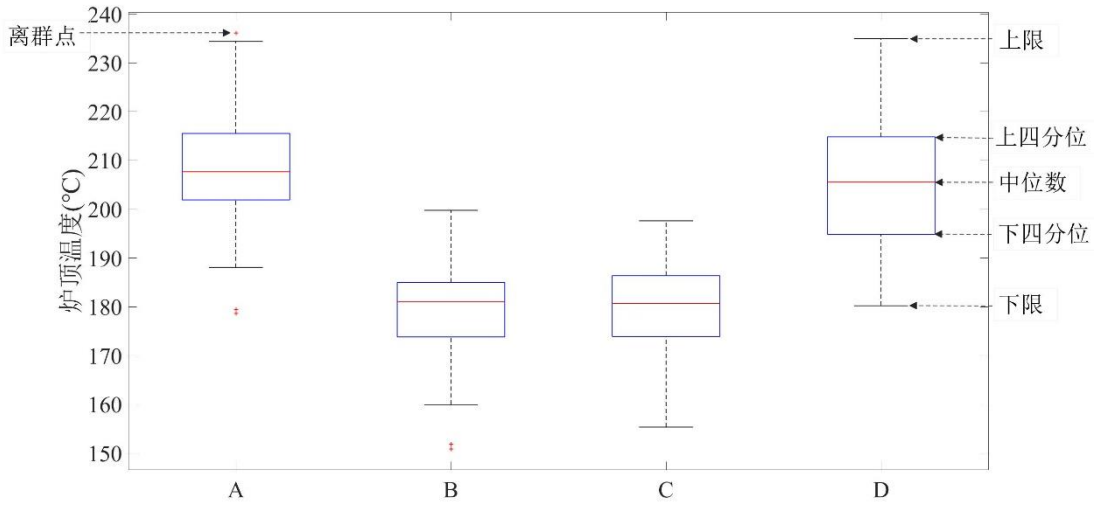


图 2.5 不同数据集的箱线图

（5）散点图

二维散点图是确定数值变量之间是否存在联系的最直观的图形描述方法之一。为了构造散点图，每个值对被视为一个代数坐标对，并作为一个点落在坐标系内。

图 2.6 展示了一组包含 50 个样本点的炼铁过程冷风压力与热风压力散点图，可以看出，冷风压力与热风压力在一定程度上呈正相关关系。此外，散点图也可以表示因变量随自变量而变化的大致趋势。

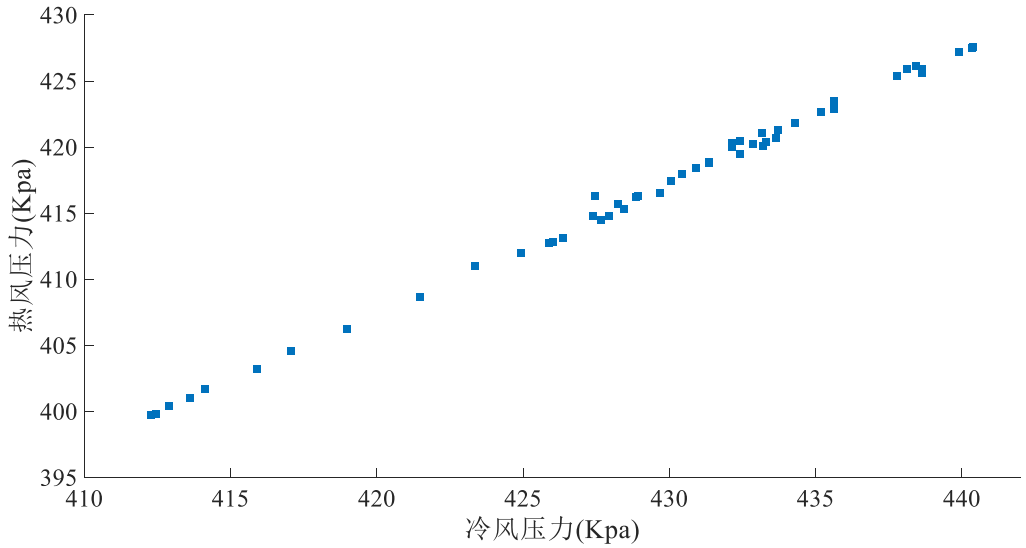


图 2.6 冷/热风压力散点图

在二维散点图中可以观察两个变量间的关系，当数据集存在多维变量，并且需要比较多维变量之间彼此两两关系时，就可以使用矩阵散点图。散点图矩阵通过将多维数据中的各个维度两两组合绘制成一系列的散点图。

图 2.7 展示了炼铁过程中在一定时间范围内采集的三个关键监测指标（数据点经过了 Z-Score 标准化，并没有改变该数据点在数据组中的位置，也没有改变这组数据的分布形状，关于数据标准化方法将在第三章进行展开介绍）间的散点图矩阵。由此散点图矩阵可以看出，冷风压力与热风压力呈现出强烈的正相关关系，这与图 2.6 得出的结论是一致的。此外，与上述散点图相比，图 2.7 同时也展示了冷风压力和热风压力分别与冷风流量呈现出一定的负相关关系。

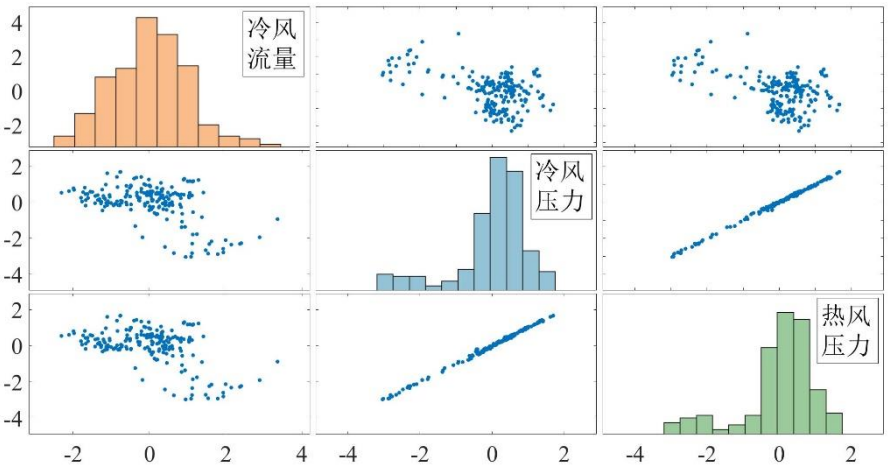


图 2.7 散点图矩阵

散点图矩阵虽然可以同时观察多个变量间的联系，但它是两两进行平面散点图的观察，实际上并不能代替高维空间的观察，有可能漏掉一些重要的信息。三维散点图就是在三个变量确定的三维空间中研究变量之间的关系，由于同时考虑了三个变量，常常可以发现两维图形中所发现不了的东西（如异常值、曲线关系等）。

图 2.8 展示了三维数据集的效果图，展示了在四个不同工况（已用椭圆标示）下，冷风压力、流量以及富氧量（Z-Score 标准化后）的数据分布情况，可以看到，在三维坐标轴下，不同工况的数据及异常值可以明显辨识。

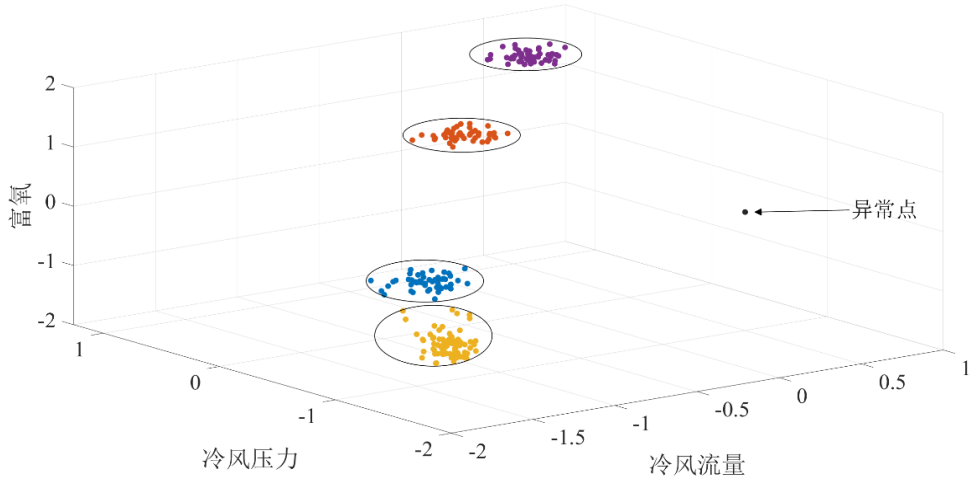


图 2.8 三维数据展示效果图

（6）平行坐标图

随着维数增加，散点图矩阵变得复杂，不利于清晰识别数据间的关系。另一种流行的技术为平行坐标图，它可以用于描述更高维度的数据。为了可视化 N 维数据点，平行坐标绘制 N 个等距离、相互平行的轴，每一个平行轴代表一个维度，纵轴表示在对应平行坐标轴维度上的取值，并用折线段将同一个数据点的不同维度取值进行连接。

图 2.9 展示了两个不同工况（加热初期和拱顶温度管理器期）下，冷风流量（CBV）、风压（BP）、2 个不同位置温度传感器测得炉顶温度 1（TT1）和炉顶温度 2（TT2）的平行坐标图。可以观测出加热初期和拱顶温度管理器期的数据呈现出明显数值差异，且拱顶温度管理器期数据在较小范围内变化。

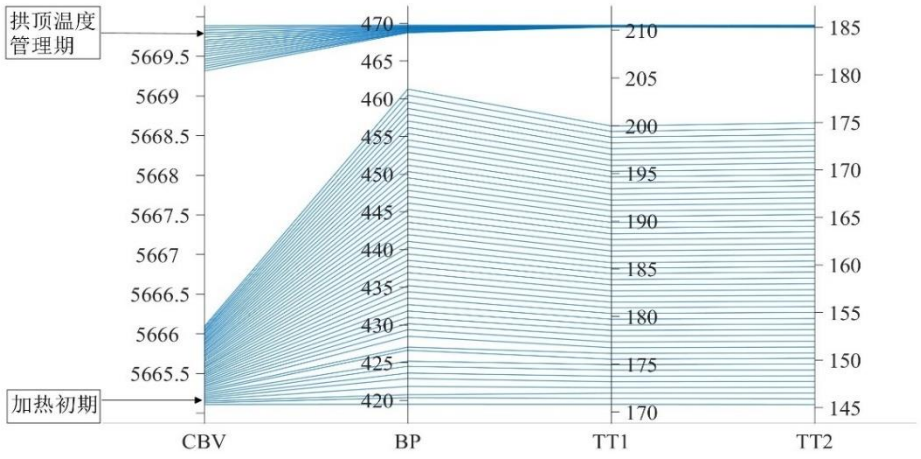


图 2.9 平行坐标图

2.3 数据的相似性度量

数据间的相似性度量可以在诸如聚类、离群点分析、因果推断中得到应用。本节将从不同属性样本的邻近性度量、衡量变量间的相关性分析指标以及因果关系推断三个方面，来讨论数据的相似性度量。

2.3.1 不同属性的邻近性度量

样本间的相似程度通过邻近性度量表示，本节主要介绍不同属性中样本数据的邻近性度量方式，主要包括标称属性、二元属性以及数值属性（数值属性的邻近性度量特称为距离度量），邻近性度量值越小，两个样本之间就越相似；度量值越大，两个样本之间就越相异。这里以 $D(\boldsymbol{x},\boldsymbol{y})$ 表示邻近性度量函数（后面对于不同邻近性度量函数进行详细定义），并介绍邻近性度量函数的通用性质：

1）非负性： $D(\boldsymbol{x},\boldsymbol{y})\geq 0$，度量值是一个非负的数值；

2）同一性： $D(\boldsymbol{x},\boldsymbol{x})=0$，一个样本与自身的度量值为 0；

3）对称性： $D(\boldsymbol{x},\boldsymbol{y})=D(\boldsymbol{y},\boldsymbol{x})$，度量函数具有对称性；

4）三角不等式： $D(\boldsymbol{x},\boldsymbol{y})\leq D(\boldsymbol{x},\boldsymbol{z})+D(\boldsymbol{z},\boldsymbol{y})$，度量函数满足三角不等式。

通常在计算多个样本间邻近性度量时，使用数据矩阵对不同样本间邻近性度量值进行存储。对于 N 个样本两两之间的邻近性度量矩阵可以表示为：

$$\begin{pmatrix} D(1,1) & D(1,2) & \cdots & D(1,N) \\ D(2,1) & D(2,2) & \cdots & D(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ D(N,1) & D(N,2) & \ldots & D(N,N) \end{pmatrix} \tag{2.8}$$

邻近性度量矩阵为对称矩阵，即 $D(\boldsymbol{x},\boldsymbol{y})=D(\boldsymbol{y},\boldsymbol{x})$，其中， $D(\boldsymbol{x},\boldsymbol{y})$ 是样本

$\boldsymbol{x}$ 和样本 $\boldsymbol{y}$ 之间的相异性或“差别”的度量。一般 $D(\boldsymbol{x},\boldsymbol{y})$ 是一个非负的数值，数值越小，样本越相似。

（1）标称属性的邻近性度量

标称属性可以取两个或多个状态，例如，炼铁矿石颜色属于标称属性，有暗红色、棕色、灰黄色等多种属性值；炼铁矿石种类属于标称属性，有赤铁矿、磁铁矿、褐铁矿等多种属性值。假设两个标称样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的属性总数为 p，取值相同的属性数目为 m，$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间的邻近性度量可根据不匹配率 $D_R(\boldsymbol{x},\boldsymbol{y})$ 来计算：

$$D_R(\boldsymbol{x},\boldsymbol{y})=\frac{p-m}{p} \tag{2.9}$$

例 2.1 有四个样本（1, 2, 3, 4）的铁矿石颜色分别为暗红色、棕色、暗红色和灰黄色，邻近性度量矩阵可表示为：

$$\begin{pmatrix} D_R(1,1) & D_R(1,2) & D_R(1,3) & D_R(1,4) \\ D_R(2,1) & D_R(2,2) & D_R(2,3) & D_R(2,4) \\ D_R(3,1) & D_R(3,2) & D_R(3,3) & D_R(3,4) \\ D_R(4,1) & D_R(4,2) & D_R(4,3) & D_R(4,4) \end{pmatrix} \tag{2.10}$$

由于上述样本只有一个标称属性（颜色属性），在式（2.10）中，令 $p=1$，使得当样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 匹配（颜色相同）时 $D_R(\boldsymbol{x},\boldsymbol{y})=0$；当样本不同时 $D_R(\boldsymbol{x},\boldsymbol{y})=1$。于是得到邻近性度量矩阵：

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{2.11}$$

可以看到，除了样本 1 和 3 外，其他样本都不相似。

对于标称数据，两个属性 A 和 B 间是否存在相关关系可以通过 χ^2（卡方）检验发现。假设 A 有 c 个不同的值，分别为 $a_1, a_2, \cdots, a_c$，B 有 r 个不同的值，分别为 $b_1, b_2, \cdots, b_r$。用 A 和 B 表示的数据可以用一个相依表展示，相依表的列数为 A 中数据的个数，行数为 B 中数据的个数，(a_i,b_j) 表示属性 A 取值 a_i、属性 B 取值 b_j 的联合事件。χ^2 值可以用式（2.12）计算：

$$\chi^2=\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \tag{2.12}$$

其中，o_{ij} 是联合事件$\left(A_i,B_j\right)$的观测频率（即实际出现的次数），而 e_{ij} 是$\left(A_i,B_j\right)$的期望频率，可以用下式计算：

$$e_{ij}=\frac{\text{count}\left(A=a_i\right)\times\text{count}\left(B=b_j\right)}{N} \tag{2.13}$$

其中，N 是样本的总个数，$\text{count}\left(A=a_i\right)$ 表示在属性 A 上具有值 a_i 的数据个数，而 $\text{count}\left(B=b_j\right)$ 表示在属性 B 上具有值 b_j 的数据个数。

注意，对 χ^2 值贡献最大的单元是观测频率与期望频率差值最大的单元，χ^2 统计检验假设两个属性 A 和 B 是彼此独立的，检验结果通过与卡方分布表中在自由度为 $(r-1)\times(c-1)$ 且人为给定的置信水平下的数值进行比较，如果可以拒绝该假设，则 A 和 B 是统计相关的。

例 2.2 假设抽调了两个工厂的 1500 台设备，每个工厂对各自设备是否属于进口的数量进行统计，并将每种可能的联合事件的观测频率（或计数）汇总在表 2.1 所显示的相依表中，其中括号中的数是期望频率。期望频率可以用式（2.13）计算得到。

表 2.1 不同工厂与选购设备是否为进口的数据相依表

	工厂 1	工厂 2	合计
进口	250（90）	200（360）	450
非进口	50（210）	1000（840）	1050
合计	300	1200	1500（N）

使用式（2.13），可以对每个单元的期望频率进行验证。例如，单元（工厂 1，进口）的期望频率是：

$$e_{11}=\frac{\text{count}\left(\text{工厂1}\right)\times\text{count}\left(\text{进口}\right)}{n}=\frac{300\times450}{1500}=90$$

注意：在任意行，期望频率的和必须等于该行总期望频率，并且任意列期望频率的和也必须等于该列的总观测频率。

使用式（2.12）对表 2.1 相依表中数据进行 χ^2 检验计算，可以得到：

$$\chi^2=\frac{\left(250\text{-}90\right)^2}{90}+\frac{\left(50\text{-}210\right)^2}{210}+\frac{\left(200\text{-}360\right)^2}{360}+\frac{\left(1000\text{-}840\right)^2}{840}=507.93$$

对于表 2.1 中数据联表，自由度计算为(2-1)(2-1)=1，在 99.9%的置信水平下，拒绝假设成立的值为 10.828（查找χ^2分布表中自自由度为 1，置信水平为 99.9%对应的置信数值）。由于计算出的值大于该值，因此可以认为不同工厂（属性）在是否选购进口设备（属性）方面是有各自偏好的，即两个属性存在相关关系。

（2）二元属性的邻近性度量

二元属性只有两种状态，可以用 0 和 1 表示，其中，0 和 1 分别表示该属性不出现和出现。如果所有属性的二元状态都被看作具有相同的权重，则可以得到一个两行两列的关联表 2.2，其中 q 是样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 都取 1 的属性数，r 是在样本 $\boldsymbol{x}$ 中取 1、在样本 $\boldsymbol{y}$ 中取 0 的属性数，s 是在样本 $\boldsymbol{x}$ 中取 0、在样本 $\boldsymbol{y}$ 中取 1 的属性数，而 t 是样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 都取 0 的属性数。属性的总数是 p，其中 $p = q + r + s +$。

表 2.2 二元属性的关联表

样本 $\boldsymbol{y}$ / 样本 $\boldsymbol{x}$	1	0
1	q	r
0	s	t

对于对称的二元属性，每个状态都同样重要（理想情况下），如果样本 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 都用对称的二元属性描述，则 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的邻近度 $D_B(\boldsymbol{x},\boldsymbol{y})$ 为：

$$D_B(\boldsymbol{x},\boldsymbol{y}) = \frac{q+t}{q+r+s+t} \tag{2.14}$$

对于两个非对称的二元属性，两个都取值 1（正匹配）比两个都取值 0（负匹配）更有意义，因此，这样的二元属性经常被认为是“一元”的（只有一种状态）。其中负匹配数 t 被认为是不重要的，计算时可忽略，则 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的邻近度 $D_J(\boldsymbol{x},\boldsymbol{y})$ 如下式所示：

$$D_J(\boldsymbol{x},\boldsymbol{y}) = \frac{q}{q+r+s} \tag{2.15}$$

式（2.15）的系数 $D_J\left(\boldsymbol{x},\boldsymbol{y}\right)$ 被称为杰卡德系数。

例 2.3 一个简易的炼铁过程二元属性例子如表 2.3 所示，其中，不同编号样本代表着取自不同时间戳的样本，并假设过程数据发生高报警为 HI，发生低报警为 LO，其中，高报警被设置为 1，低报警被设置为 0（且认为高低报警状态同等重要，是对称二元属性）。

表 2.3 炼铁过程二元属性举例

样本编号	反应釜压力	反应釜液位	反应釜温度
1	HI（1）	LO（0）	HI（1）
2	HI（1）	LO（0）	LO（0）

利用式（2.14）计算二元样本间的邻近度：

样本 1 和 2 间的邻近度 $D_B(1,2)$ 为：

$$D_B(1,2)=\frac{1+1}{1+1+0+1}\approx 0.67$$

样本 1 和 3 间的邻近度 $D_B(1,3)$ 为：

$$D_B(1,3)=\frac{0+0}{0+2+1+0}=0$$

样本 2 和 3 间的邻近度 $D_B(2,3)$ 为：

$$D_B(2,3)=\frac{0+1}{0+1+1+1}\approx 0.33$$

可以看出，在三组样本中，样本 1 和 2 所处的工况环境最为相似，样本 1 和 3 所处的工况环境差别最大。

（3）数值属性的距离度量

本节介绍几种常用的计算数值属性样本间距离度量的方法，包括欧几里德距离、曼哈顿距离、切比雪夫距离及马氏距离。

欧几里得距离（Euclidean Distance）：也称欧式距离，指在 P 维空间中两个样本（向量）间的直线距离，或者向量的自然长度，两样本 $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_P\right)$ 与 $\boldsymbol{y}=\left(y_1,y_2,\cdots,y_P\right)$ 间的欧式距离 D_E 为：

$$D_E(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{P}|x_i-y_i|^2} \tag{2.16}$$

曼哈顿距离（Manhattan Distance）：两个样本（向量）在标准坐标系的各个轴向上距离的总和。在 P 维空间中，两样本 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间的曼哈顿距离 D_M 为：

$$D_M(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{P}|x_i-y_i| \tag{2.17}$$

切比雪夫距离（Chebyshev Distance）：两个样本（向量）坐标数值差的绝对值中的最大值。在 P 维空间中，两样本 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间的切比雪夫距离 D_C 为：

$$D_C(\boldsymbol{x},\boldsymbol{y})=\max_{i=1,2,\ldots,P}(|x_i-y_i|) \tag{2.18}$$

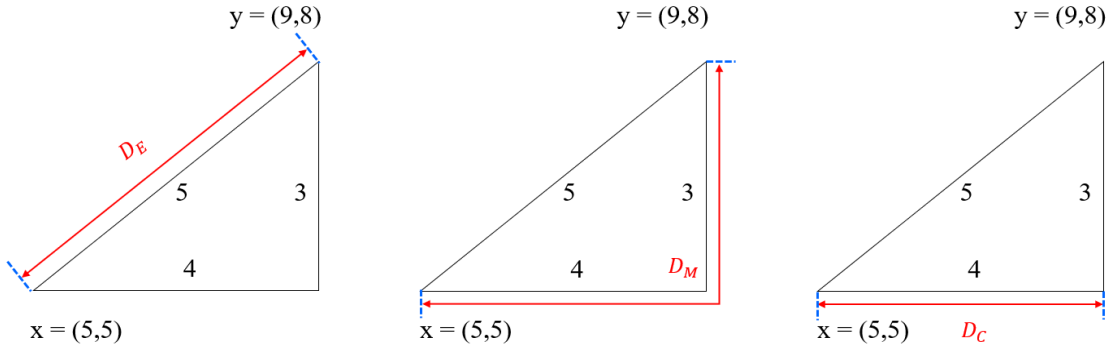
欧几里德距离、曼哈顿距离、切比雪夫距离可以统一使用闵可夫斯基距离

（Minkowski Distance）表示，定义两个样本 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 之间的闵可夫斯基距离 D_{MK} 为：

$$D_{MK}(\boldsymbol{x},\boldsymbol{y})=\left(\sum_{i=1}^{P}|x_i-y_i|^{L}\right)^{\frac{1}{L}} \tag{2.19}$$

其中，L 为变参数，依据 L 的不同，闵氏距离可以具体表示以上举例的某一种。

在二维（$d=2$）坐标系下，采用两个样本 $\boldsymbol{x}=(5,5)$ 和 $\boldsymbol{y}=(9,8)$ 进行上述三种距离的对比计算，并在图 2.10 进行展示，以便理解三种距离的定义及区别。



$D_E=\sqrt{4^2+3^2}=5$　　$D_M=4+3=7$　　$D_C=MAX\{4,3\}=4$

图 2.10 欧式距离、曼哈顿距离及切比雪夫距离对比

例 2.4 给定两个样本 $\boldsymbol{x}$=[5665.07 432.41 419.46 190.79 165.34 150.24 199.76 237.99 236.72 236.38]，$\boldsymbol{y}$=[5670.73 433.32 420.35 204.39 178.43 170.56 212.18 238.34 237.07 236.71]，计算不同 L 值下的距离。

当 $L=1$ 时，为曼哈顿距离的表达式，此时，两者间的曼哈顿距离计算如下：

$$D_M(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{10}|x_i-y_i|=67.95$$

当 $L=2$ 时，为欧式距离的表达式，2 个样本间的欧式距离计算如下：

$$D_E(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{10}|x_i-y_i|^2}=30.95$$

当 $L\to\infty$ 时，为切比雪夫距离的表达式，切比雪夫距离定义为各坐标数值差绝对值值的最大值：

$$D_C(\boldsymbol{x},\boldsymbol{y})=\max_{i=1,2,\ldots,10}(|x_i-y_i|)=20.32$$

马氏距离（Mahalanobis Distance）表示数据的协方差距离，它是一种有效的计算两个样本相异度的方法。马氏距离修正了欧式距离中各个维度尺度（单位）不一致且相关的问题，考虑到各种特征之间的联系，但独立于测量尺度。

两个长度一致的样本 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间的马氏距离 $M_D(\boldsymbol{x},\boldsymbol{y})$ 定义如下：

$$M_D(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}C^{-1}(\boldsymbol{x}-\boldsymbol{y})} \tag{2.20}$$

这里，C 代表总体样本的协方差矩阵，C^{-1} 表示求协方差矩阵的逆矩阵。

例 2.5 由图 2.7 可知，冷风压力与冷风流量数据存在相关关系，且二者的尺度不一致，符合马氏距离使用场景，这里采用冷风流量（ $\boldsymbol{x}$ ）和冷风压力（ $\boldsymbol{y}$ ）的 100 个样本进行计算。

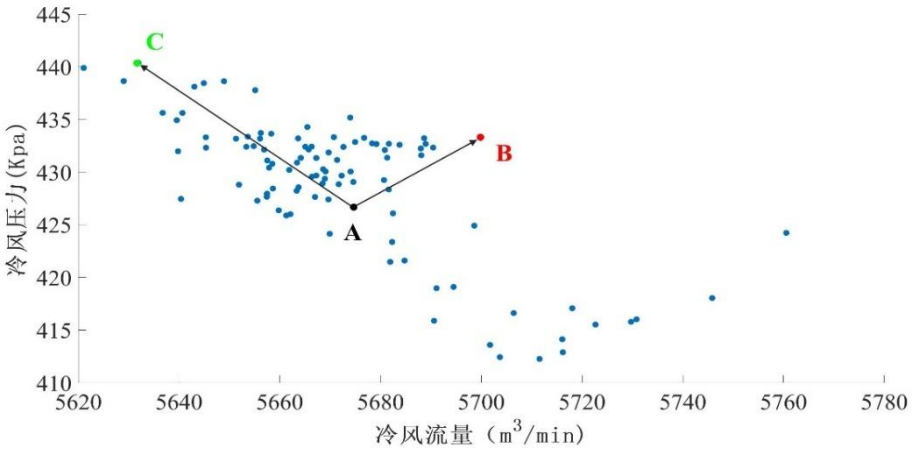


图 2.11 冷风流量冷风压力散点图

其中，A 点坐标（5674.68, 426.676），B 点坐标（5699.83, 433.312），C 点坐标（5631.67, 440.343），协方差矩阵计算为：

$$\begin{pmatrix} 613.4681 & -122.4050 \\ -122.4050 & 43.5945 \end{pmatrix}$$

直观上，B 与 A 比 C 与 A 更相近，但是从整体数据分布趋势来看，C 与整体样本分布的趋向更加一致。通过式（2.18）和（2.20）分别计算 B、C 两点到 A 点的欧式距离 D_E 和马氏距离 M_D：

$$D_{E\,A\to B}(\boldsymbol{x},\boldsymbol{y})=26.01 \qquad D_{E\,A\to C}(\boldsymbol{x},\boldsymbol{y})=45.13$$

$$M_{D\,A\to B}(\boldsymbol{x},\boldsymbol{y})=2.85 \qquad M_{D\,A\to C}(\boldsymbol{x},\boldsymbol{y})=2.09$$

对比计算结果可以知道，在欧式距离度量下，B 与 A 距离更近，在马氏距离度量下，C 与 A 间的马氏距离显然更小。

欧氏距离在各类聚类算法中应用广泛；曼哈顿距离适用于离散或二进制属性数据集，曼哈顿距离只需要做加减法，这使得计算机在大量的计算过程中代价更低；在实践中，切比雪夫距离可用于仓库物流中起重机托运货物的最短路径规划。

对于非独立同分布数据在计算欧式距离时，不同维度（变量）默认赋予相

同的“权重”，然而，量纲的不同会导致计算出的度量数值不同。由此，可以通过数据归一化或标准化消除量纲影响，然而数据归一化或标准化（通过变换将不同维度数据缩放至共同的区间）改变的是同一维度上数据的大小，却无法消除数据分布对距离度量的影响（如图 2.11）。由此，考虑到数据分布（变量相关）对距离度量的影响，引入马氏距离。马氏距离在运算过程中按照主成分进行旋转，使得维度间相互独立，然后进行标准化，让维度同分布。在选择使用距离度量时，一定要根据具体场景和问题，选择合适的度量方式。

2.3.2 相关性分析指标

样本间的相似程度通过邻近性度量表示，而变量间的相似程度可以通过相关性分析进行确定，以衡量不同变量间的相关密切程度。

（1）时间序列自相关

在时间序列分析中定义自相关系数（Auto-Correlation Function, ACF），用来衡量同一序列（变量）中不同时间间隔数据之间的相关性随时间间隔的变化情况，定义时间序列 $\boldsymbol{a}$ 的自相关系数 $C_A(k)$ 为：

$$C_A(k)=\frac{\dfrac{1}{N-k}\sum_{t=k+1}^{N}(x_t-\bar{x})(x_{t-k}-\bar{x})}{\dfrac{1}{N}\sum_{t=1}^{N}(x_t-\bar{x})^2} \tag{2.21}$$

其中，k 为时间间隔个数，N 为采样点的数目。自相关系数反映了同一事件在两个不同时期之间的相关程度。

例 2.6 已知图 2.2 中煤气利用率采样点数目 $N=60$ 个，观测变化周期约为 24h，采样周期为 2h，数据均值 $\bar{x}$=45.66%。这里设置 $k=24/2=12$，来计算此时间序列的自相关系数：

$$C_A(12)=\frac{\dfrac{1}{60-12}\sum_{t=12+1}^{60}(a_t-45.66\%)(a_{t-12}-45.66\%)}{\dfrac{1}{60}\sum_{t=1}^{60}(a_t-45.66\%)^2} \tag{2.22}$$

$$C_A(12)=\frac{\dfrac{1}{60-12}*87.60\%}{\dfrac{1}{60}*109.37\%}\approx 1.00$$

可以得出，此时间序列在间隔 $k=12$ 时，呈现出很强的自相关性。

（2）皮尔逊相关系数

在统计学中，皮尔逊（Pearson）相关系数是用于度量两个变量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 之间的线性相关关系，其值介于-1 与 1 之间。定义两个变量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 之间的皮尔逊相关

系数 $C_P(\boldsymbol{a},\boldsymbol{b})$ 为：

$$C_P(\boldsymbol{a},\boldsymbol{b})=\frac{\mathrm{cov}(\boldsymbol{a},\boldsymbol{b})}{\sigma_{\boldsymbol{a}}\sigma_{\boldsymbol{b}}}=\frac{\sum_{i=1}^{N}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i-\bar{b})^2}} \tag{2.23}$$

其中，$\mathrm{cov}(\boldsymbol{a},\boldsymbol{b})$ 是变量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 间的协方差，$\sigma_{\boldsymbol{a}}$ 和 $\sigma_{\boldsymbol{b}}$ 分别是 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的标准差，N 表示样本的数目，a_i 和 b_i 分别表示 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的第 i 个取值，$\bar{a}$ 和 $\bar{b}$ 分别是 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的均值。

例 2.7 已知图 2.7 中所展示的数据具有一定的线性相关关系，因此使用此数据段进行皮尔逊相关性分析，以定量计算数据间的线性相关强度。取其中两个变量冷风压力（$\boldsymbol{a}$）和冷风流量（$\boldsymbol{b}$）进行示例计算：

$$C_P(\boldsymbol{a},\boldsymbol{b})=\frac{\sum_{i=1}^{200}(x_i-5655.6)(y_i-429.96)}{\sqrt{\sum_{i=1}^{200}(x_i-5655.6)^2}\sqrt{\sum_{i=1}^{200}(y_i-429.96)^2}}\approx -0.48 \tag{2.24}$$

可以得出，$\boldsymbol{a}$ 与 $\boldsymbol{b}$ 之间呈现出中等的负相关关系。

（3）秩相关系数

秩相关系数，又称等级相关系数，是将两个变量的不同时刻数据按数据值的大小顺序排列位次，以样本在各个维度上的秩次来代替实际数据值而求得的一种统计量。它是反映等级相关程度的统计分析指标，常用的等级相关分析方法有斯皮尔曼（Spearman）相关系数和肯德尔（Kendall）相关系数。

Spearman 相关系数是利用两个变量的秩次做线性相关分析，用来衡量两个变量间是否单调相关，Spearman 相关系数 C_S 定义如下：

$$C_S=\frac{\sum_{i=1}^{N}(p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum_{i=1}^{N}(p_i-\bar{p})^2}\sqrt{\sum_{i=1}^{N}(q_i-\bar{q})^2}} \tag{2.25}$$

其中，p_i 和 q_i 分别为 a_i 和 b_i 的秩次，$\bar{p}$ 和 $\bar{q}$ 分别表示两个变量各自所有数据的平均秩次，若变量中出现值相等，则该值对应的秩次为这几个值对应秩次的平均值。若变量间的秩次差值可以计算，则公式（2.25）可简化为：

$$C_S=1-\frac{6\sum_{i=1}^{N}(p_i-q_i)^2}{N(N^2-1)} \tag{2.26}$$

其中，p_i 代表 a_i 的秩次，$\boldsymbol{q_i}$ 代表 b_i 的秩次，p_i-q_i 为 a_i 与 b_i 的秩次之差。若 $0<C_S\le 1$，表明两个变量呈正相关关系；若 $-1\le C_S<0$，表明呈负相关关系；

当 C_S＝ 0，则表示不相关。

例 2.8 表 2.4 展示了利用 Spearman 相关系数方法定量分析冷风流量（***a***）和冷风压力（***b***）两个变量的线性相关强度，秩次按照升序计算。

表 2.4 ***a*** 和 ***b*** 的 Spearman 相关系数计算

a	***b***	***a*** 的秩次	***b*** 的秩次	秩次差	秩次差的平方
5665.07	432.41	4	1	3	9
5670.73	433.32	5	2	3	9
5658.31	433.64	3	3.5	0.5	0.25
5636.75	433.64	2	3.5	1.5	2.25
5629.01	438.64	1	5	4	16

注：若数据存在（每维）相同的情况，将它们所在位置取算数平均数作为相同的秩次。

两个变量间的 Spearman 相关系数可计算为：

$$C_S = 1 - \frac{6\sum_{i=1}^{5}(p_i - q_i)^2}{5(5^2-1)} = -0.725$$

可以看出，该数据段的冷风流量和冷风压力呈现出较强的负相关关系。

Kendall 相关系数的应用对象是有序的类别变量，比如名次、年龄段等，它可以度量两个有序变量之间单调关系强弱。Kendall 相关系数使用了“成对”这一概念来决定相关系数的强弱。成对可以分为一致对和分歧对，一致对是指两个变量取值的相对关系一致，可以理解为 $a_2 - a_1$ 与 $b_2 - b_1$ 有相同的符号；分歧对则是指它们的相对关系不一致，$a_2 - a_1$ 与 $b_2 - b_1$ 有着相反的符号。

Kendall 相关系数通常使用两个公式进行计算，一个是 C_{K-A}，另一个是 C_{K-B}。两者的区别是 C_{K-B} 可以处理有相同值的情况，即并列，下面依次来介绍这两个公式。

对于不存在重复值的变量（样本长度均为 N ）***a*** 和 ***b***，C_{K-A} 的计算公式为：

$$C_{K-A}(\boldsymbol{a},\boldsymbol{b}) = \frac{c-d}{C_N^2} \tag{2.27}$$

其中，c 表示一致对数，d 表示分歧对数，C_N^2 表示所有样本两两组合的数量，当变量 ***a*** 和 ***b*** 中均没有重复值时，组合数量就等于 $c+d$ 。

下面以表 2.5 和 2.6 为例介绍 c 和 d 的含义（压力单位为 Kpa，温度单位为℃）：

表 2.5 一致对数示例

样本序号	冷风压力	炉顶温度	冷风压力数值顺序	炉顶温度数值顺序	一致对数

1	419	190	升序	升序	1
2	420	204			

表 2.6 分歧对数示例

样本序号	冷风压力	炉顶温度	冷风压力数值顺序	炉顶温度数值顺序	分歧对数
1	425	210	升序	降序	1
2	427	199			

当变量 $\boldsymbol{a}$ 或 $\boldsymbol{b}$ 中有重复值时，采用 C_{K-B} 计算两者间的 Kendall 相关系数。C_{K-B} 的计算公式为：

$$C_{K-B}(\boldsymbol{a},\boldsymbol{b})=\frac{c-d}{\sqrt{T-T_r}\sqrt{T-T_c}} \tag{2.28}$$

其中，$T=C_N^2$，T_r 是指变量 $\boldsymbol{a}$ 中重复值的个数：

$$T_r=\sum_{i=1}^{s}\frac{1}{2}U_i(U_i-1) \tag{2.29}$$

T_c 是指变量 $\boldsymbol{b}$ 中重复值的个数：

$$T_c=\sum_{i=1}^{t}\frac{1}{2}V_i(V_i-1) \tag{2.30}$$

这里以 T_r 计算为例，假设 $\boldsymbol{a}$ 是变量冷风压力，序数数据为[419, 420, 420, 420, 425, 425]，s 表示集合 $\boldsymbol{a}$ 中拥有的小集合（集合内数值相同）的个数，那么这里得到的 s 则为 2，因为只有 420、425 有相同元素），U_i 表示 $\boldsymbol{a}$ 中第 i 个小集合所包含的元素数。

例 2.9 利用 2.1.3 节序数属性（按照数据程度增加的方向进行排序，比如：极低序号为 1，极高序号为 5）中实例进行 C_{K-A} 计算，数据如下表所示：

表 2.7 透气性指数和热风压力属性举例

样本编号	透气性指数	透气性指数序号	热风压力	热风压力序号	c	d
1	极低	1	高压	5	0	4
2	偏低	2	低压	2	2	1
3	良好	3	极低	1	2	0
4	偏高	4	过压	4	0	1
5	极高	5	良好	3	0	0

$$C_{K-A}=\frac{4-6}{\frac{1}{2}*5(5-1)}=-0.2$$

可以看出，在该数据集中，透气性指数和热风压力的 Kendall 相关系数为-0.2，该数据段的透气性指数和热风压力相关强度较弱。

2.3.3 因果关系

因果关系（Causality）是“因”（即一个事件）和“果”（即另一个事件）之间的作用关系，其中后一事件被认为是前一事件的结果。一般来说，一个事件是很多原因综合产生的结果，原因都发生在较早时间点，而该事件又可以成为其他事件的原因。本节主要讲述格兰杰因果关系及基于滞后相关的因果推断方法。

（1）格兰杰因果推断

格兰杰因果关系推断是一种假设检定的统计方法，用以检验一个变量是否为另一个变量的原因，被广泛应用于时间序列分类和预测领域。

以平稳时间序列数据作为研究对象（平稳性即均值、方差和协方差不随时间变化，这是进行格兰杰因果推断的前提条件），两个变量$\boldsymbol{a}$和$\boldsymbol{b}$之间因果关系定义为：变量$\boldsymbol{a}$有助于预测变量$\boldsymbol{b}$的未来值，则认为变量$\boldsymbol{a}$是引致变量$\boldsymbol{b}$的格兰杰原因。可以通过建立与变量$\boldsymbol{a}$、$\boldsymbol{b}$相关的向量自回归模型，比较模型残差的方差大小判断两个变量之间是否存在因果关系。向量自回归模型如下：

$$b_{t+1}=\sum_{j=0}^{m-1}\alpha_j b_{t-1}+\varepsilon_{b,t+1} \tag{2.31}$$

$$b_{t+1}=\sum_{j=0}^{m-1}\beta_j a_{t-j}+\sum_{j=0}^{m-1}\gamma_j b_{t-j}+\varepsilon_{\boldsymbol{b}|\boldsymbol{a},t+1} \tag{2.32}$$

其中α_j、β_j和γ_j为模型的系数，m为模型的阶数，$\varepsilon_{\boldsymbol{b}}$和$\varepsilon_{\boldsymbol{b}|\boldsymbol{a}}$为模型的残差。

根据回归预测结果，通过比较向量自回归模型残差的方差大小，判断$\boldsymbol{a}\rightarrow\boldsymbol{b}$是否存在 Granger 因果关系，Granger 因果指数（Granger Causality Index, GCI）定义如下：

$$\mathrm{GCI}_{\boldsymbol{a}\rightarrow\boldsymbol{b}}=\ln\frac{\mathrm{var}\left(\varepsilon_{\boldsymbol{b}}\right)}{\mathrm{var}\left(\varepsilon_{\boldsymbol{b}|\boldsymbol{a}}\right)} \tag{2.33}$$

如果满足$\mathrm{var}\left(\varepsilon_{\boldsymbol{b}|\boldsymbol{a}}\right)<\mathrm{var}\left(\varepsilon_{\boldsymbol{b}}\right)$，即$\mathrm{GCI}_{\boldsymbol{a}\rightarrow\boldsymbol{b}}>0$则表明$\boldsymbol{a}\rightarrow\boldsymbol{b}$存在统计意义下的 Granger 因果关系。

（2）基于滞后相关的因果推断

基于滞后相关的因果推断是利用互相关函数来估计过程变量之间的时间延

迟，进而构建因果矩阵分析变量之间的因果关系。假设两个离散变量$\boldsymbol{a}(t)$和$\boldsymbol{b}(t)$之间的互相关函数$R_{\boldsymbol{ab}}(k)$为：

$$R_{\boldsymbol{ab}}(k)=E\left[\left(\boldsymbol{a}(t)-\bar{a}\right)\left(\boldsymbol{b}(t+k)-\bar{b}\right)\right] \tag{2.34}$$

其中，k为时间间隔序数（k=0, ±1, ±2,…），E表示期望值。$R_{\boldsymbol{ab}}(k)$的最大值和最小值分别表示为$\varphi^{\max}$和$\varphi^{\min}$，对应的时延常数分别表示为$k^{\max}$和$k^{\min}$，则变量$\boldsymbol{a}$和$\boldsymbol{b}$之间的时滞$\lambda_{\boldsymbol{a},\boldsymbol{b}}$为：

$$\lambda_{\boldsymbol{a},\boldsymbol{b}}=\begin{cases}k^{\max},\varphi^{\max}+\varphi^{\min}\geq 0\\ k^{\min},\varphi^{\max}+\varphi^{\min}<0\end{cases} \tag{2.35}$$

利用上式判断变量$\boldsymbol{a}$和$\boldsymbol{b}$间的影响关系，如果$\lambda_{\boldsymbol{a},\boldsymbol{b}}>0$，则表明变量$\boldsymbol{b}$受到了变量$\boldsymbol{a}$变化的影响。$\boldsymbol{a}$和$\boldsymbol{b}$之间的关联系数$\rho_{\boldsymbol{a},\boldsymbol{b}}$表示为：

$$\rho_{\boldsymbol{a},\boldsymbol{b}}=\begin{cases}\varphi^{\max},\varphi^{\max}+\varphi^{\min}\geq 0\\ \varphi^{\min},\varphi^{\max}+\varphi^{\min}<0\end{cases} \tag{2.36}$$

上述公式反映变量$\boldsymbol{a}$和$\boldsymbol{b}$之间的关联关系。基于时滞的互相关函数方法为判别时间序列因果关系提供了一种实用有效的方法，算法简单，计算复杂度低，在变量之间具有线性关系时可以准确地分析出变量间存在的因果关系。

目前因果关系分析的方法很多，但是这些方法在实际工业应用中都存在一定的局限性。例如，Granger 因果关系分析只适用于线性过程和平稳序列。这些局限性导致方法的应用受到限制，因此在原始方法的基础上，可以采用贝叶斯网络、最近邻以及传递熵等方法，提升其在非线性、非平稳等过程的适用性，以解决实际工业过程中的故障溯源等问题。

2.4 本章小结

本章为数据的基本知识，主要讲述了数据的基本概念、数据统计描述、图形展示以及数据的相似性和相异性度量方式。

基本概念：介绍了不同属性数据的基本概念、数据的基本统计描述指标、数据的相似性和相异性度量指标。此外，结合了钢铁冶金的相关数据实例，加深对数据相关概念的理解。

数据统计描述：基本统计描述为数据预处理提供了分析基础。数据的基本统计度量包括度量数据中心趋势的均值、中位数、众数以及中列数，协方差中位数和众数，度量数据散布的极差、平均差、方差、标准差、协方差、四分位数以及离散系数。

图形展示：可视化有助于清晰明了地展现数据的总体状态，本章介绍了折

线图、直方图、条形图、箱线图、散点图以及平行坐标图。

相似性度量：本章介绍了不同属性数据的邻近性度量以及相关性分析指标，以描述数据的相似程度。对象相似性用于诸如聚类、离群点分析、因果推断等应用中。杰卡德系数可以评估非对称二元属性的相似程度，欧几里得距离、曼哈顿距离、闵可夫斯基距离以及马氏距离（不同评价尺度）可以度量样本间的距离。自相关函数用来衡量时间序列自身的相关密切程度，皮尔逊相关系数用来评价数值属性变量间的线性相关程度，秩相关系数用来评价序数属性间的相关程度。格兰杰因果关系及基于滞后相关的因果推断方法，可以有效判断两个变量间是否存在因果关系。

习题

2-1 请列举生活中常见的不同属性数据。

2-2 统计数据的离散趋势度量和集中趋势度量分别有哪些？

2-3 指出皮尔逊、斯皮尔曼及肯德尔相关系数的适用条件。

2-4 请简述协方差和相关系数的区别。

2-5 讨论文中给出的不同相似性度量方法的应用场景。

2-6 除了本章给出的邻近性度量外，尝试调研并给出其他邻近性度量方法及应用场景。

2-7 简要概括标称属性对象的相异性及非对称二元属性对象的相异性。

2-8 给定两组数据 $\boldsymbol{x}=[1,3,4,8]$， $\boldsymbol{y}=[2,6,7,10]$，分别计算 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间的欧式距离、曼哈顿距离、切比雪夫距离以及马氏距离。

2-9 给出例 2.6 中所用煤气利用率中的部分数据，分别计算在 k=10, 11, 13 时的时间序列自相关系数，数据集如下：[45.01 44.76 44.22 43.84 43.88 45.11 47.30 45.95 47.22 46.16 46.72 47.59 45.20 44.30 44.71 43.82 43.70 45.30 47.40 45.79 47.3 46.21 46.79 47.72]。

2-10 分别计算两组数据 $\boldsymbol{a}$=[190.79 204.39 210.39 218.41 210.75 199.25 194.34 179.02 165.64 165.46]和 $\boldsymbol{b}$=[165.34 178.43 190.52 203.52 204.91 197.42 193.34 181.46 167.42 163.07]的皮尔逊和斯皮尔曼相关系数。

2-11 计算出两组数据 $\boldsymbol{a}=[2.1\ 0.9\ 2.2\ 3.1\ 2.0]$ 和 $\boldsymbol{b}=[1.0\ 2.1\ 3.0\ 2.0\ 1.1]$ 互相关系数最大时的 k 取值。

参考文献

[1] 贾俊平, 何晓群, 金勇进. 统计学（第四版）[M]. 中国人民大学出版社, 2009 年.

[2] 吕林根, 许子道. 解析几何（第四版）[M]. 高等教育出版社, 2006 年.

[3] Han J., Kamber M and Pei J. Data mining: concepts and techniques[M]. Morgan kaufmann, 2012.

[4] Han J., Kamber M and Pei J. 数据挖掘概念与技术（第三版）[M]. 范明, 孟小峰译. 机械工业出版社, 2012.

[5] 宋万清. 数据挖掘（第一版）[M]. 中国铁道出版社, 2019 年.

[6] 王燕. 应用时间序列分析（第四版）[M]. 中国人民大学出版社, 2015 年.

[7] Pearson K. Regression, Heredity and Panmixia[J]. *Philosophical Transactions of the Royal Society of London. Series A*, 1896, 187: 253-318.

[8] Kendall M. A new measure of rank correlation[J]. *Biometrika,*1938, 30: 81-89.

[9] Granger C. W. J. Investigating causal relations by econometric models and cross-spectral methods[J]. *Econometrica*, 1969, 37: 424-438.

[10] Spearman C. General intelligence, objectively determined and measured[J]. *American Journal of Psychology*, 1904, 15: 201-293.

[11] Bauer M., Thornhill. N. F. A practical method for identifying the propagation path of plant-wide disturbances[J]. *Journal of Process Control*, 2008, 18:707-719.

[12] Yang F, Duan P, Shah S. L. and Chen T. Capturing Connectivity and Causality in Complex Industrial Processes[M]. *Springer,* 2014.

[13] Hu W., Shah S. L. and Chen T. Framework for a smart data analytics platform towards process monitoring and alarm management[J]. *Computers & Chemical Engineering*, 2018, 114, 225-244.

[14] Runkler T. A. Data Analytics: Models and Algorithms for Intelligent Data Analysis[M]. Vieweg, 2012.