第三章数据预处理

在工业制造过程中，原始工业数据是数据挖掘的信息来源。这些数据通常含有噪声、大量的缺失值，这会影响数据挖掘的效率和结果的可信性，甚至产生一些无效归纳。对原始数据进行预处理，可为数据挖掘过程提供干净、准确、简洁的数据，为最终挖掘的结果提供保障。

针对原始数据，有大量的数据预处理技术。本章对数据预处理进行概述，首先阐述工业数据质量特性及问题，然后介绍数据清洗的基本内容，讨论数据集成需要考虑的问题。接着对数据变换策略进行概述，最后介绍数据归约技术。

3.1 工业数据质量

本节内容将讨论数据质量特性，并介绍智能制造过程中常见的数据质量问题。针对这些问题，介绍数据预处理的主要任务。

3.1.1 数据质量特性及问题

数据如果能满足其应用要求，那么它是高质量的。数据质量涉及许多因素，如图 3.1 所示，包括准确性、完整性、一致性、时效性、可信性和可解释性。

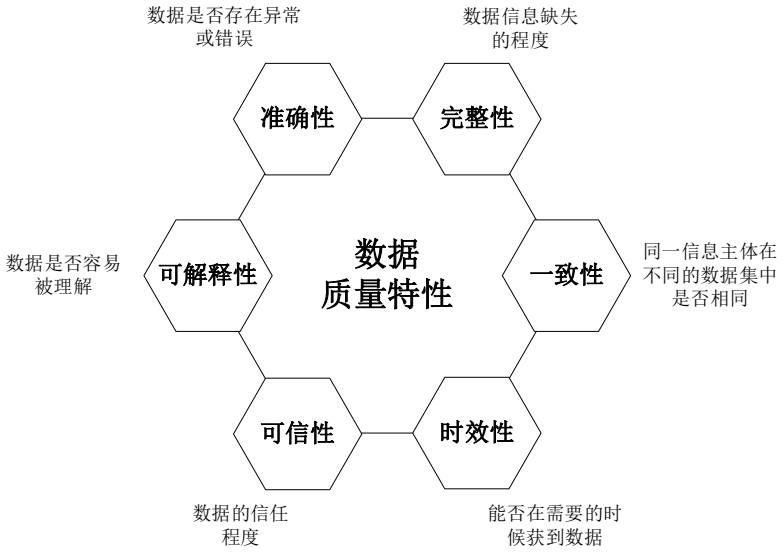

图 3.1 数据质量特性

假设你是某工厂的一名数据分析师，现需要分析高炉炼铁的部分数据，例如风温、冷风流量和冷风压力等。在研究和审查工厂的数据库时，你可能会注意到，有些属性在某些时刻或者时间段上没有值，这是典型的数据缺失现象。如果你需要分析冷风流量与温度两个属性之间的关系，数据缺失可能会导致分析结果不准确。此外，你还可能会发现数据中存在一些错误的、不寻常和不一致性的值。换

言之，实际工业数据往往存在一些普遍的质量问题，例如数据可能是不完整的（缺少某些属性值）、不正确的或含噪声的（包含错误或存在偏离期望的值），以及不一致的（不同数据库的相同属性编码存在差异）。

现在大型数据库中的数据都有着不正确、不完整和不一致的特点。不正确的数据可能由多种原因导致，例如收集数据的设备可能出现故障，人或计算机可能在数据输入时出现错误。此外，不正确的数据也可能是由命名约定或所用的数据代码不一致，或输入字段的格式不一致而导致的。

出现不完整数据可能是由于理解错误或设备故障。不一致的数据可能已经被删除，历史或修改的数据可能被忽略，从而导致数据不完整。这些存在缺失的数据需要推导出来，特别是在某些属性上存在缺失值的样本。

时效性也会影响数据的质量，假设你正在监控某高炉中炉顶温度的变化，然而，工人们未能在月末及时提交高炉中炉顶温度数据。此时数据库中的数据是不完整的，对数据质量具有负面影响。

数据质量的另外两个特征是可信性和可解释性。可信性反映有多少数据是使用者信赖的，而可解释性反映数据是否容易理解。假设在某一时刻数据库有一些错误，之后都被更正，而过去的错误已经给使用者造成了问题，因此他们不再相信该数据。有的数据可能会使用许多编码，但却不知道如何解释它们。即便该数据库现在是正确的、完整的、一致的、及时的，但是由于很差的可信性和可解释性，使用者仍然可能把它看成低质量的数据。

由于数据存在上述特征及问题，在使用数据之前，需要对数据进行预处理。

3.1.2 数据预处理的主要任务

数据预处理的主要任务包括：数据清洗、数据集成、数据变换和数据归约。如图 3.2 所示，经过这些步骤，能有效提升数据质量，保证数据挖掘的效率和结果的可信性。

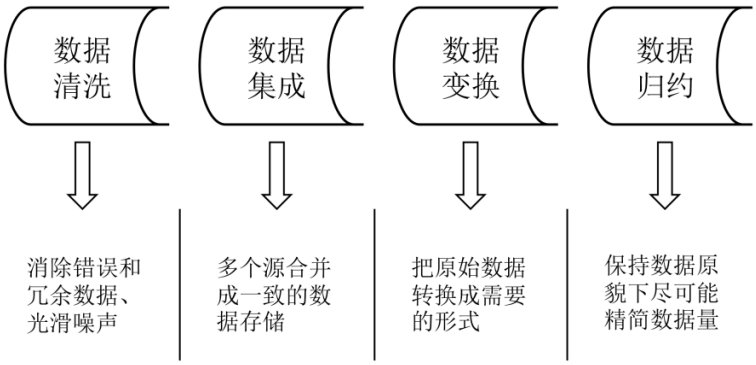

图 3.2 数据预处理的形式

数据清洗通过填写缺失的值，光滑噪声数据，识别和删除异常值，解决不一致性，来清洗数据。“脏”数据造成挖掘过程陷入混乱，导致不可靠的输出。一

个有效的预处理步骤需要通过数据清洗改善数据质量。数据清洗将在 3.2 节讨论。

数据集成用于处理来自多个数据源的数据。同一概念的属性在不同的数据库中可能具有不同的名字，这导致不一致性和冗余。例如，关于高炉炼铁过程中冷风流量在一个数据库存储中为 CBV，而在另一个数据库存储为 CBV_ID。显然，除了数据清洗之外，必须采取措施，避免数据集成时的冗余。数据集成将在 3.3 节讨论。

对于数据挖掘而言，离散化是强有力的工具，它使得数据挖掘可以在多个抽象层上进行；数据规范化可以用来把数据压缩到较小的区间；数据离散化、规范化都是某种形式的数据变换。数据变换将在 3.4 节讨论。

数据归约旨在得到数据集的简化表示，并且产生几乎同样的分析结果。数据归约策略包括维归约和数值规约：在维归约中，使用数据编码方案，以便得到原始数据的简化或“压缩”表示；在数值归约中，使用参数模型或非参数模型，用较小的表示取代数据。数据归约将在 3.5 节讨论。

总之，数据预处理技术可以改进数据的质量，有助于提高数据挖掘过程的精度和性能。检测数据异常，尽早地调整数据，并归约待分析的数据，将为决策带来高可靠性。

3.2 数据清洗

现实中工业数据一般是不完整的、有噪声的和不一致的。数据清洗主要用于填充缺失的值、光滑噪声并识别异常值、纠正数据中的不一致。数据清洗包含很多步骤和内容，每项具体任务不尽相同，并不要求都遵循相同的数据清洗步骤。

3.2.1 格式内容清洗

工业大数据往往种类繁杂、数量巨大，数据预处理对象为大量异构数据。如果数据是由系统日志而来，那么通常在格式和内容方面，会与原来的数据描述一致。如果数据是由人工收集或用户填写而来，则有很大可能在格式和内容上存在一些问题。简单来说，格式内容问题有以下几类：

（1）时间、日期、数值、全半角等显示格式不一致

这种问题通常在整合多来源数据时可能会遇到，将其处理成某种一致的格式即可。例如，某一数据库中的时间标注为 24 小时制，而在另一数据库中的时间标注为 12 小时制。

（2）内容中有不该存在的字符

某些内容可能只包括一部分字符，或者是某些数据中有不应存在的汉字、字母。这种情况下，需要以半自动校验结合半人工查验的方式来找出可能存在的问题，并去除不需要的字符。

例 3.1 当冷风压力数据中有不该存在字母、字符，对其进行格式错误清洗，结果如图 3.3 所示。

TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:00	5665.07	432.41
2015-06-01 00:01	5670.73	433.32
2015-06-01 00:02	5658.31	433.64Mpa
2015-06-01 00:03	5636.75	435.62Mpa
2015-06-01 00:04	5629.01	438.64
2015-06-01 00:05	5631.91	440.38

→

AVG（CBP） 冷风压力
432.41
433.32
433.64
435.62
438.64
440.38

图 3.3 格式错误清洗效果图

（3）内容与该字段应有内容不符

例如在高炉炼铁的数据库中，将冷风流量的数据写成了冷风压力的数据就属于这类问题。但该问题的特殊性在于并不能简单地以删除来处理，其原因可能有：人工填写错误、前端没有校验或导入数据时部分或全部列没有对齐，因此要详细识别问题类型。

3.2.2 缺失值

工业数据中样本的个别属性会出现缺失的情况，例如表 3.1 中的“空单元”或“NaN”。出现数据缺失的情况时，应该进行如下处理：

表 3.1 含缺失值的数据表

TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:06	5631.67	440.34
2015-06-01 00:07	5621.07	439.89
2015-06-01 00:08		438.63
2015-06-01 00:09	5663.64	433.20
2015-06-01 00:10	5667.24	NaN
2015-06-01 00:11	5658.63	428.45

1）忽略样本：当类标号缺少时通常这样做（假定挖掘任务涉及分类或描述）。除非样本中有多个属性缺少值，否则该方法不是很有效。当每个属性缺失值的百分比很高时，它的性能非常差。

2）人工填写缺失值：一般来说，该方法很费时。当数据集很大、缺失很多值时，该方法可能行不通。

3）使用一个全局常量填充缺失值：将缺失的属性值用同一个常量替换。

4）使用属性的中心度量（如均值或中位数）填充缺失值：正常数据用均值填充，倾斜数据则用中位数填充。

5)使用与给定样本同一类的属性均值或中位数:例如某一时刻的产量缺失，则用同一运行模态下的其他时刻的均值或者中位数来补充。如果给定数据是倾斜的，则中位数是更好的选择。

6）使用最可能的值补充缺失值：可以使用决策树归纳（第六章）、贝叶斯推理（第六章）或回归分析（第七章）确定。

例 3.2 当冷风流量某一行数据缺失，采用上述方法 4 求出其余冷风流量数据的均值，结果如图 3.4 所示。

TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:17	5698.62	424.92
2015-06-01 00:18	5718.06	417.08
2015-06-01 00:19		414.13
2015-06-01 00:20	5716.19	412.90

求三个时刻冷风流量的均值 →

AVG（CBV） 冷风流量
5698.62
5718.06
5710.96
5716.19

图 3.4 属性的平均值填补处理

例 3.3 当运行模态 1 中的数据缺失。采用上述方法 5，用同一模态下的均值来进行填补缺失值，结果如图 3.5 所示。

时间	产量	运行模态
12/26	7.6	1
12/27	9.3	2
12/28	8.1	1
12/29	8.6	1
12/30		1
12/31	10.4	2

求三个同一模态产量的均值 →

产量
7.6
9.3
8.1
8.6
8.1
10.4

图 3.5 同一模态均值法处理

3.2.3 噪声数据

噪声是测量变量的随机错误或方差，它会极大影响建立模型的鲁棒性。常用的数据光滑技术有分箱法、小波变换法、经验模态分解法等。

（1）分箱法

分箱法通过考察数据的“近邻”值（即周围的值）来光滑有序数据值，这些值通常被分到不同的“箱”中。由于分箱法考察近邻的值，因此它属于局部光滑。

常用分箱法有用箱均值光滑、用箱中位数光滑和用箱边界光滑。其中，用箱均值光滑是将箱中的每一个值用均值替换。用箱中位数光滑时，箱中的每一个值被替换为箱中的中位数。对于用箱边界光滑，给定箱中的最大和最小值，被视为箱边界，而箱中的每一个值被替换为最近的边界值。一般而言，宽度越大，光滑效果越明显。箱可以是等宽的，每个箱值的区间范围是常量。分箱也可以作为一种离散化技术使用，将在 3.4 节进一步讨论。

例 3.4 给定如下数据：4，8，15，21，21，24，25，28，34。图 3.6 中，数据首先被划分到大小为 3 的等频的箱中（即每个箱包含 3 个值），再利用上述三种方法进行数据平滑。例如，对于用箱均值光滑，箱 1 中的值 4，8 和 15 由均值 9 替换。

划分为（等频的）箱：	用箱均值光滑：	用箱中位数光滑：	用箱边界光滑：
箱1: 4, 8, 15	箱1: 9, 9, 9	箱1: 8, 8, 8	箱1: 4, 4, 15
箱2: 21, 21, 24	箱2: 22, 22, 22	箱2: 21, 21, 21	箱2: 21, 21, 24
箱3: 25, 28, 34	箱3: 29, 29, 29	箱3: 28, 28, 28	箱3: 25, 25, 34

图 3.6 数据光滑的分箱法

（2）小波变换

小波变换去噪的基本思想如下：根据噪声与信号在不同频带上的小波分解系数具有不同强度分布的特点，将各频带上噪声对应的小波系数去除，保留原始信号的小波分解系数。然后对处理后的系数进行小波重构，得到去噪后的信号。相比其它的去噪方法，小波变换在低信噪比情况下的去噪效果较好，去噪后的信号识别率较高，对时变信号和突变信号的去噪效果尤其明显。

小波变换（Wavelet Transform，WT）的本质是采用一组伸缩平移且互相正交的小波函数族对原始信号进行展开。它的主要特点是通过变换能够充分突出问题某些方面的特征，能对时间（空间）频率的局部化分析；通过伸缩平移运算对信号（函数）逐步进行多尺度细化，最终达到高频处时间细分，低频处频率细分，能自动适应时频信号分析的要求，从而可聚焦到信号的任意细节。

相比于小波变换，傅里叶变换对函数作频谱分析，反映了整个信号的时间频

谱特性，它虽较好地揭示了平稳信号的特征，但是具有一定局限性，如不具备局部化分析能力、不能分析非平稳信号等，这里不做详细介绍。小波变换继承和发展了短时傅里叶变换局部化的思想，同时又克服了窗口大小不随频率变化等缺点。它能够提供一个随频率改变的“时间-频率”窗口，是进行信号时频分析和处理的理想工具。

小波是指具有衰减性且持续时间很短的波。从数学角度上来说，当某一函数$\psi(t)$满足下列条件时，称其为小波函数或小波母函数：

1）$\psi(t)$为一平方可积函数，即是$\psi(t) \in L^2(R)$；

2）$0 < \int_{-\infty}^{+\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty$(3.1)

其中$\Psi(\omega)$为$\psi(t)$的傅里叶变换。

将小波函数$\psi(t)$进行时域上的平移和尺度上伸缩，得到小波基函数：

$$\psi(t)_{a,\tau} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \quad a > 0, \tau \in R \tag{3.2}$$

其中 a 为伸缩因子，τ 为平移因子。

把小波变换中的伸缩因子和平移因子抽样，离散结果为$a = a_0^j$，其中$j = 0, \pm 1, \pm 2, \ldots$，此时对应的小波基函数为：

$$\psi(t)_{j,\tau} = a_0^{-j/2}\psi\left(\frac{t-\tau}{a^j}\right) \tag{3.3}$$

平移因子 τ 离散化过程为在同一尺度因子下对某一初始值进行线性取值，即$\tau = \mathrm{a}_0^j \mathrm{k}\tau_0$，此时对应的小波基函数为：

$$\psi(t)_{j,k} = a_0^{-j/2}\psi\left(\frac{t-k\tau_0}{a^j}\right) k = 0, \pm 1, \pm 2, \ldots \tag{3.4}$$

通过对平移因子进行归一化处理，最后就能够得到离散化后的小波基函数为：

$$\psi(t)_{j,k} = a_0^{-j/2}\psi\left(\frac{t}{a^j} - k\right) \tag{3.5}$$

例 3.5 利用例 2.8 中某钢铁厂的煤气利用率数据，我们选取部分含噪信号，利用小波变换去噪方法，效果如图 3.7 所示。

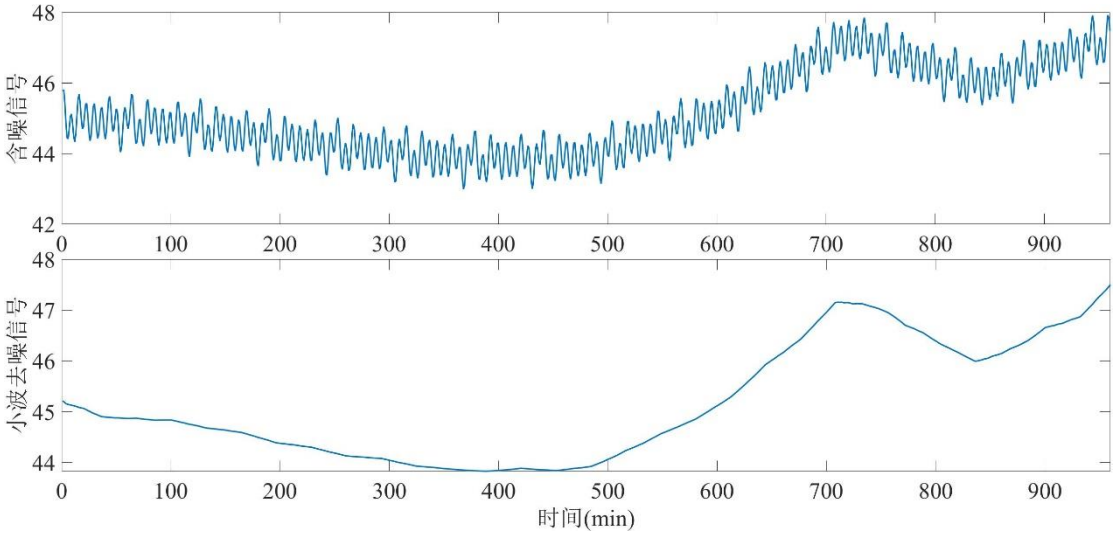

图 3.7 小波变换去噪效果图

（3）经验模态分解

经验模态分解（Empirical Mode Decomposition,EMD）是依据数据自身的时间尺度特征来进行信号分解，无须预先设定任何基函数。这一点与建立在先验性的谐波基函数和小波基函数上的傅里叶分解与小波分解方法具有本质性的差别。正是由于这样的特点，EMD 方法在理论上可以应用于任何类型信号的分解，因而在处理非平稳及非线性数据上，具有非常明显的优势。EMD 的作用是能使复杂信号分解为有限个本征模函数（Intrinsic Mode Function,IMF），所分解出来的各 IMF 分量包含了原信号不同时间尺度的局部特征信号。

EMD 方法基于以下假设条件：

1）数据至少有两个极值，一个最大值和一个最小值。

2）数据的局部时域特性是由极值点间的时间尺度唯一确定。

3）如果数据没有极值点但有拐点，则可以通过对数据微分一次或多次求得极值，然后再通过积分来获得分解结果。这种方法的本质是通过数据的特征时间尺度来获得本征波动模式，然后分解数据。

具体步骤如下：

1）找到信号$x(t)$所有的极值点。

2）用 3 次样条曲线拟合出上下极值点的包络线$e_{max}(t)$和$e_{min}(t)$，并求出上下包络线的平均值$m(t)$，在$x(t)$中减去它，求出$h(t)$：

$$h(t) = x(t) - m(t) \tag{3.6}$$

3）根据预设判据判断$h(t)$是否为 IMF。

4）如果不是，则以$h(t)$代替$x(t)$，重复以上步骤直到$h(t)$满足判据。

5）每得到一阶 IMF，就从原信号中扣除它，重复以上步骤；直到信号最后

剩余部分只是单调序列或者常值序列。这样，经过 EMD 方法分解就将原始信号$x(t)$分解成一系列 IMF 以及剩余部分的线性叠加：

$$x(t)=\sum_{i=1}^{n} f_i(t)+r_n(t) \tag{3.7}$$

式中$f_i(t)$是 EMD 分解得到的第 i 个 IMF，$r_n(t)$是分解筛除 n 个 IMF 后的信号残余分量，代表信号的直流分量或信号的趋势。

例 3.6 采用与例 3.5 相同的数据，进行经验模态分解。如图 3.8 所示，将含噪信号$x(t)$经过上述步骤的 EMD 分解，将原始信号分解成 f_1-f_5 的 5 个本征模函数与分解筛出的信号残余分量 r 的线性叠加。将 5 个本征模函数进行信号重构，去噪效果如图 3.9 所示。

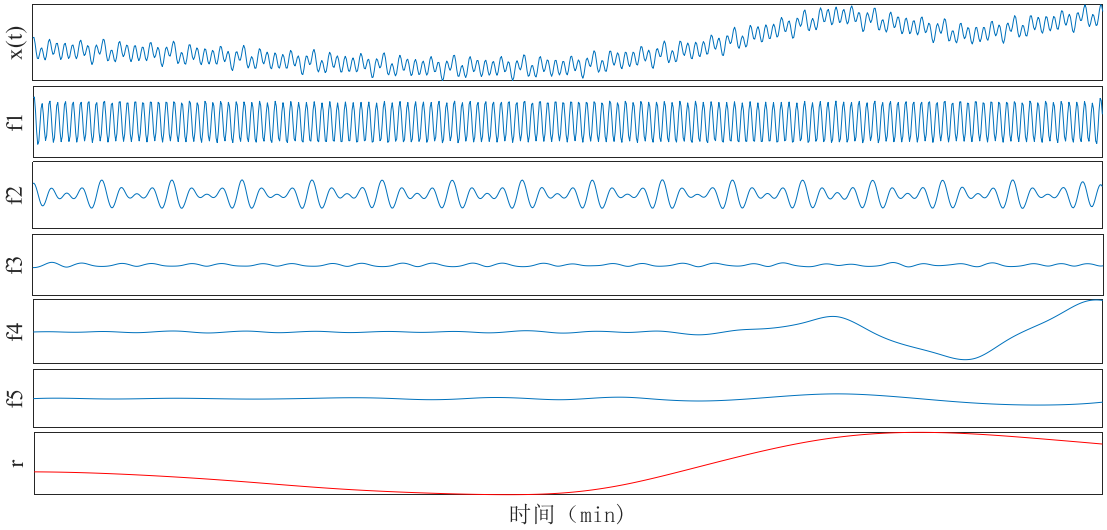

图 3.8 经验模态分解

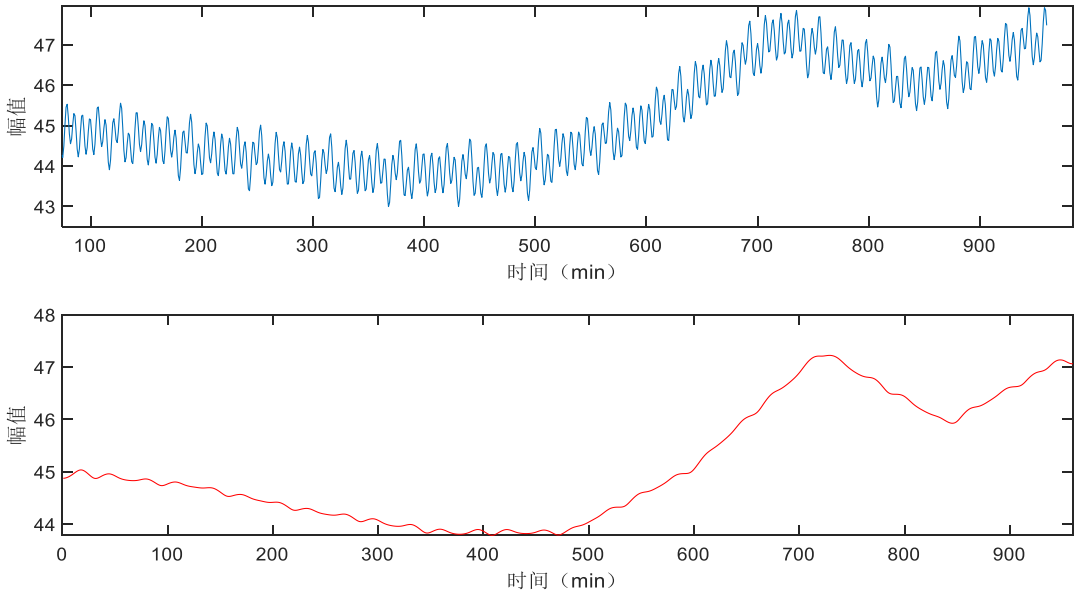

图 3.9 信号重构效果图

3.2.4 异常值清洗

检测数据有输入错误或者不合常理的数据，这些数据称为异常值。通常可以通过可视化界面来观测出异常值，如图 3.10 所示的是热风压力的一组数据样本，其中异常值一目了然。在大数据中，将全部的采集数据可视化往往是比较困难的。定义和识别异常值，往往需要采取很多其他方法，常用的方法如下：

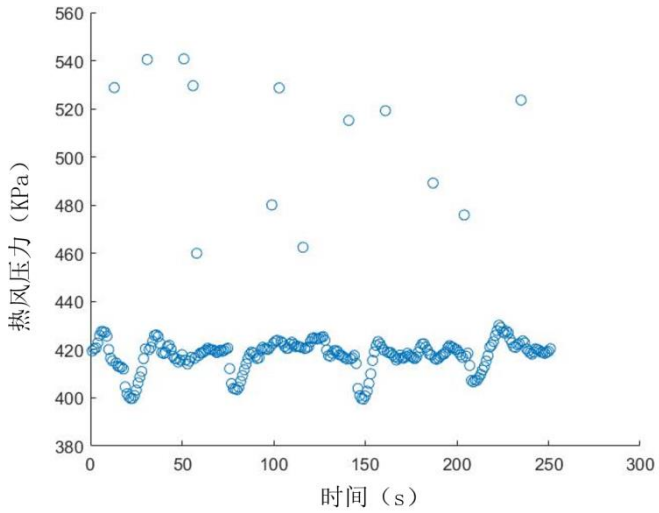

图 3.10 热风压力数据的异常值可视化

（1）简单统计分析

拿到数据后可以对数据进行一个简单的描述性统计分析，例如，结合机理知识预先设定最大、最小阈值来判断某个变量的取值是否超过了合理的范围。

例 3.7 如图 3.11 所示数据，高炉炼铁过程中的冷风流量为-5000，显然不合常理，为异常值，可通过删除该行进行异常值清洗。

TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:00	5665.07	432.41
2015-06-01 00:01	5670.73	433.32
2015-06-01 00:02	-5000	433.64
2015-06-01 00:03	5636.75	435.62
2015-06-01 00:04	5629.01	438.64
2015-06-01 00:05	5631.91	440.38

删除该行

TIME 时间	AVG（CBV） 冷风流量
2015-06-01 00:00	5665.07
2015-06-01 00:01	5670.73
2015-06-01 00:03	5636.75
2015-06-01 00:04	5629.01
2015-06-01 00:05	5631.91

图 3.11 异常值清洗效果图

（2）3σ原则

如果数据服从正态分布，在 3σ原则下，异常值为一组测定值中与平均值的偏差超过 3 倍标准差的值。距离平均值 3σ之外的值出现的概率为$P(|x-u|>3\sigma)\leq 0.003$，属于极个别的小概率事件。如果数据不服从正态分布，也可以用远距离平均值不同倍数的标准差来描述。

（3）箱线图分析

箱线图是数字数据通过其四分位数形成的图形化描述，这是一种非常简单但有效的可视化异常值的方法。箱线图提供了识别异常值的一个标准：如果一个值小于$Q_1-1.5\times IQR$或大于$Q_2+1.5\times IQR$的值，则被称为异常值。Q_1为下四分位数，表示全部观察值中有四分之一的数据取值比它小；Q_2为上四分位数，表示全部观察值中有四分之一的数据取值比它大；IQR为四分位数间距，是上四分位数Q_2与下四分位数Q_1的差值，包含了全部观察值的一半。

箱线图判断异常值的方法以四分位数和四分位距为基础，四分位数具有一定的鲁棒性：25%的数据可以变得任意远并且不会干扰四分位数，所以异常值不能影响这个标准。因此，箱线图识别异常值比较客观，在识别异常值时有一定的优越性。箱线图的更多细节可以参考 2.2.3 节。

（4）基于模型检测

首先构建一个概率分布模型，并计算对象符合该模型的概率，把具有低概率的对象视为异常点。如果模型是簇的集合，则异常是不显著属于任何簇的对象；如果模型是回归时，异常是相对远离预测值的对象。

（5）基于距离

通常可以在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象。具体可参考 2.3.2 节中的距离度量。

（6）基于密度

当一个点的局部密度显著低于它的大部分近邻时才将其分类为异常值，这种判断方法适合非均匀分布的数据。

（7）基于聚类

首先定义基于聚类的异常值，一个对象如果不强属于任何簇，则判定该对象是一个基于聚类的异常值。如果通过聚类检测异常值，则由于异常值影响聚类，存在结构是否有效的问题。处理该问题可以使用如下方法：对象聚类、删除异常值、对象再次聚类。聚类分析方法将在第 5 章进行详细介绍。

3.2.5 逻辑错误清洗

这部分的工作是去掉一些使用简单逻辑推理就可以直接发现问题的数据，防

止分析结果走偏。主要包含以下几个步骤：

（1）去重

有的数据分析把去重放在第一步，但一般情况下建议把去重放在格式内容清洗之后，因为没有经过格式清洗的数据很大可能会去重失败。

大数据中去重要特别小心，很多时候大数据的算法并不希望去重，甚至还会自动生成很多重合数据。一是人为判断的重复不一定是真正的重复，这些被认为重复而实际上不重复的数据往往表达了真实世界中的有用信息；二是某些数据的重复是必须的，比如深度学习中的对抗设计，需要去生成非真实采集的重复数据。

（2）去除不合理值

在机器收集数据的过程中，人工参与部分的偶发性错误会导致出现不合理值。例如目前时间显示为 2022 年，但是数据中填写的时间是 2202 年，这种错误往往不能靠异常值清洗去除。处理这类数据时要么删掉，要么按处理数据缺失的方法处理，如果分析得当，还可以根据数据来源进行数据重构。

逻辑错误除了以上的举例，还有很多未列举的情况，在实际操作中要酌情处理。另外，这一步骤在之后的数据分析建模过程中有可能重复，因为即使问题很简单，也并非能够一次找出所有问题。可以通过使用工具和方法，尽量减少问题出现的可能性，使分析过程更为高效。

3.3 数据集成

数据集成将多个数据源中的数据合并，存放在一个一致的数据存储中。这些数据源可能包括多个数据库、数据立方体（一种多维数据模型）或一般文件。好的数据集成方法可以减少结果数据集的冗余和不一致，这有助于提高数据挖掘的准确性和速度。在数据集成过程中，实体识别问题、冗余、样本重复问题以及数据冲突等都是需要重点考虑的问题。

3.3.1 实体识别问题

将来自多个信息源的现实世界的等价实体进行匹配，这涉及实体识别问题。例如，如何判断高炉炼铁过程中冷风流量在一个数据库存储中为 CBV，而在另一个数据库存储为 CBV_ID 是相同的属性。实际上，每个属性的数据包含名字、含义、数据类型和允许取值范围等，以及处理空白、零、NULL 值的规则（见 3.2 节数据清洗），这样可以帮助避免集成时的错误。数据集成中，当一个数据库的属性与另一个数据库的属性匹配时，必须注意其数据结构，以保证系统中函数依赖、参数约束与目标系统匹配。

3.3.2 冗余

冗余是数据集成需要考虑的另一个重要问题，如果一个属性能由另一个或几

个属性“导出”，则这个属性可能是冗余的。属性名称的不一致可能导致数据集成时产生冗余，有些冗余可以通过相关性分析检测得到。对于标称数据，可以使用χ^2（卡方）检验检测属性之间的相关性；对于数值属性，利用相关系数和协方差评估一个属性的值如何随着另一个属性值变化。标称数据的χ^2（卡方）检验和数值属性的相关系数和协方差在 2.1 节中具体介绍过

3.3.3 样本重复

除了属性之间有冗余之外，样本也可能存在冗余，也就是说一组实体数据中存在两个或多个相同的样本，可能是数据库对同一时刻的数据进行了重复记录。

3.3.4 数据冲突

数据集成时，由于不同数据源的表示方式、度量方法或编码存在区别，数据值可能存在冲突。例如，重量属性可能在一个系统中以国际单位存放，而在另一个系统中以英制单位存放。

3.4 数据变换

数据变换主要是对数据进行一定转化，变换为适当的形式，使得挖掘过程更有效，挖掘模式更易理解。

3.4.1 数据变换策略概述

在数据变换中，数据被变换或统一成适合于挖掘的形式。数据变换策略包括如下几种：

1）光滑：去掉数据中的噪音。这类技术包括分箱、小波变换和经验模态分解等。

2）属性（特征）构造：由给定的属性构造新的属性并添加到属性集。

3）聚集：对数据进行汇总和聚集。通常，这一步用来为多个抽象层的数据分析构造数据立方体。

4）规范化：把属性数据按照比例缩放，使之落入一个特定的小区间，如 -1.0-1.0 或者 0.0-1.0。

5）离散化：数值属性的原始值用区间标签（如 0-10，11-20 等）替换。

数据预处理的主要任务之间存在许多重叠，本节集中讨论后两种策略。

3.4.2 规范化

数据规范化的常用方法有三种：最小-最大（Min-Max）规范化、Z 分数（Z-Score）规范化和按小数定标规范化。假定 v 是数值属性，具有 n 个观测值 $v_1, v_2, \ldots, v_k, \ldots, v_n$，分别用三种方法进行数据规范化计算过程如下：

（1）最小-最大（Min-Max）规范化

最小-最大规范化对原始数据进行线性变换。假定$v_{\min}$和v_{max}分别为属性v的最小和最大值。最小-最大规范化通过式（3.8）

$$v' = \frac{v-v_{min}}{v_{max}-v_{min}}(v'_{max} - v'_{min}) + v'_{min} \quad (3.8)$$

把 v 的值映射到区间$[v'_{min}, v'_{max}]$中的v'。

最小-最大规范化保持原始数据值之间的相对联系。如果输入实例落在v的原数据值域之外，则该方法将面临“越界”错误。

例 3.8 假设属性热风温度的最小值与最大值分别为 1114℃和 1326℃，现在的风温值为 1200℃，我们想把热风温度映射到区间[0,1]，根据最小-最大规范化风温值 1200℃将变换为：

$$\frac{1200-1114}{1326-1114}(1-0)+0 = 0.406$$

（2）Z 分数规范化

在 Z 分数规范化中，基于v的平均值和标准差规范化。v的值被规范化为v'，由式（3.9）计算：

$$v' = \frac{v-\bar{v}}{\sigma_v} \quad (3.9)$$

当属性v的实际最大和最小值未知，或异常值导致最小-最大规范化所得结果不合理时，该方法是有用的。

例 3.9 假设属性热风温度的均值和标准差分别为 1189℃和 374℃。使用 Z 分数规范化，值 1200℃被转换为：

$$\frac{1200-1189}{374} = 0.06$$

（3）小数定标规范化

小数定标规范化通过移动属性v的小数点位置进行规范化，小数点的移动位数依赖于v的最大绝对值。v的值被规范化为v'，由式(3.10）计算：

$$v' = \frac{v}{10^j} \quad (3.10)$$

其中，j 是使得$(\max|v'|) < 1$的最小整数。

例 3.10 假设v的取值是由-986 到 917，v的最大绝对值为 986。使用小数定标规范化时，我们用每个值除以 1000（即，j=3）。这样，-986 被规范化为-0.986，而 917

被规范化为 0.917。

3.4.3 数据离散化

离散化是将一个连续属性的范围划分为区间，即使用间隔标签替换实际的数据值，通过离散化减少数据量。离散化技术可以分为有监督和无监督的离散化，也可以分为分割（自顶向下）与合并（自底向上）两种形式。离散化可以递归地对属性执行，为进一步的分析做准备。典型的数据离散化方法有：分箱、直方图、聚类、决策树等。

（1）通过分箱离散化

分箱是一种基于指定的箱个数的自上向下的分裂方式，3.2.2 节讨论了数据光滑的分箱方法，这些方法也可以用于数据归约。例如，通过使用等宽或等频分箱，然后用箱均值或中位数替换箱中的每个值；可以将属性值离散化，就像用箱的均值或箱的中位数光滑一样。

分箱并不使用类信息，因此是一种无监督的离散化技术。它对用户指定的箱个数很敏感，也容易受异常值的影响。

（2）通过直方图分析离散化

与分箱一样，直方图分析也是一种无监督离散化技术，因为它也不使用类别信息。直方图把属性的值划分成不相交的区间，称作桶或箱。

直方图可以使用各种划分规则来定义，例如在等宽直方图中，将值分成相等分区或区间。理想情况下，使用等频直方图划分数据，使得每个分区包括相同个数的数据样本。直方图分析算法可以递归地用于每个分区，可以对每一层使用最小区间长度来控制递归过程。最小区间长度设定每层每个分区的最小宽度，或每层每个分区中值的最少数目。

等宽：在等宽的直方图中，每个桶的宽度区间是一致的。

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

（3）通过聚类、决策树离散化

聚类分析是一种离散化方法。通过将属性的值划分成簇或组，来实现数据离散化。聚类考虑属性值的分布以及数据点的邻近性，因此可以产生高质量的离散化结果。聚类分析方法将在第 5 章具体介绍。

分类决策树也可以用来离散化，这类技术使用自顶向下的划分方法。决策树方法是有监督的，它使用类别标号。类分布信息用于计算和确定划分点（划分属性区间的数据值)，其主要思想是选择划分点使得一个给定的结果分区包含尽可能多的同类样本。分类分析方法将在第 6 章具体介绍。

3.5 数据归约

数据归约技术可以用来得到数据集的归约表示，规约后的数据虽小得多，但仍能接近保持原数据的完整性。数据归约策略包括维归约、数量归约和数据压缩。

维归约（Dimensionality Reduction）减少所考虑的随机变量或属性个数。常用方法包括主成分分析和小波变换，他们可以把原数据变换或投影到维度较小的空间。属性子集选择也是一种维归约方法，其中不相关、弱相关或冗余的属性被检测和删除。

数量归约（Numerosity Reduction）用替代的、较小的数据表示形式替换原数据。它可以分为参数与非参数方法：对于参数方法而言，使用模型估计数据，例如回归和对数线性模型，一般只需要存放模型参数，而不是实际数据；非参数方法包括直方图分析、聚类和抽样等。

数据压缩（Data Compression）使用变换，以便获得原数据的归约或“压缩”表示。如果原数据可以由压缩后的数据重构，而不损失信息，则数据压缩是无损的；如果只能近似重构原数据，则称为有损的。维归约和数量规约也可以视为某种形式的数据压缩。

3.5.1 维归约

维归约通过删除不相关的属性减少数据量，它包括主成分分析、小波变换、属性子集选择。

（1）主成分分析

主成分分析（Principal Component Analysis, PCA）利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法。主成分分析主要用于发现数据中的基本结构，即数据中变量之间的关系，是数据分析的有力工具。

数据集合中的样本由实数空间（正交坐标系）中的点表示，空间的一个坐标轴表示一个变量，规范化处理后得到的数据分布在原点附近。对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上。新坐标系的第一坐标轴、第二坐标轴等分别表示第一主成分、第二主成分等，数据在每一轴上的坐标值的平方表示相应变量的方差。另外，这个坐标系是在所有可能的新坐标系中，坐标轴上的方差之和最大的。

例如，数据由两个变量x_1和x_2表示，存在于二维空间中，每个点表示一个样本，如图 3.12（a）所示（对数据已做规范化处理）。可以看出，这些数据分布在以原点为中心的左下至右上倾斜的椭圆之内。很明显在这个数据中的变量x_1和x_2是线性相关的，具体地，当知道其中一个变量x_1的取值时，对另一个变量x_2的预

测不是完全随机的；反之亦然。

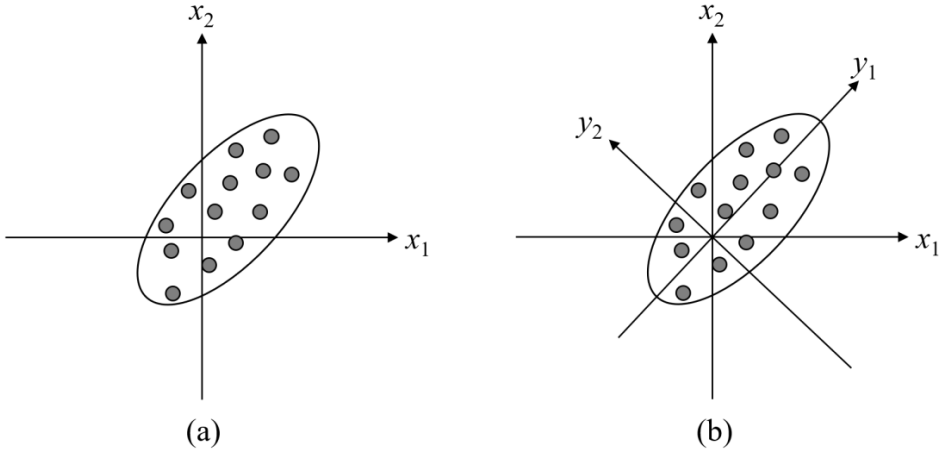

图 3.12 主成分分析的示例

主成分分析对数据进行正交变换，即对原坐标系进行旋转变换，并将数据在新坐标系表示，如图 3.12（b）所示。数据在原坐标系由变量x_1和x_2表示，通过正交变换后，在新坐标系里，由变量y_1和y_2表示。主成分分析选择方差最大的方向（第一主成分）作为新坐标系的第一坐标轴，即y_1轴，在这里意味着选择椭圆的长轴作为新坐标系的第一坐标轴；之后选择与第一坐标轴正交，且方差次之的方向（第二主成分）作为新坐标系的第二坐标轴，即y_2轴，在这里意味着选择椭圆的短轴作为新坐标系的第二坐标轴。在新坐标系里，数据中的变量y_1和y_2是线性无关的，当知道其中一个变量y_1的取值时，对另一个变量y_2的预测是完全随机的；反之亦然。如果主成分分析只取第一主成分，即新坐标系的y_1轴，那么等价于将数据投影在椭圆长轴上，用这个主轴表示数据，将二维空间的数据压缩到一维空间中。

对于正交属性空间（高维坐标系）中的样本点，如果我们需要用一个超平面（直线的高维推广，相当于降维）对样本进行恰当的表达，可以从以下两个思路入手：最近重构性，即样本点到这个超平面的距离都足够近；最大可分性，即样本点在这个超平面的投影尽可能分开。

最近重构性表示降维后忽视的坐标轴带来的信息损失尽可能最少，最大可分性表示新的坐标系尽可能代表原来样本点更多的信息。这两者本质上是一致的。基于最近重构性和最大可分性，我们可以得到主成分分析的两种等价推导：

1）最近重构性

假定数据样本进行了中心化，即$\sum_i x_i = 0$；假定投影变换后得到的新坐标系为$\{\omega_1, \omega_2 \ldots, \omega_d\}$，其中$\omega_i$是标准正交基向量，$\| \omega_i \|_2 = 1, \omega_i^T \omega_j = 0 (i \neq j)$。若丢弃新坐标系中的部分坐标，即将维度降低到$d' < d$，则样本点$x_i$在低维坐标系中的投影是$y_i = (y_{i1}; y_{i2}; \ldots; y_{id'})$，其中$y_{ij} = \omega_j^T x_i$是$x_i$在低维坐标系下第 j 维的

坐标。若基于y_i来重构x_i，则会得到$\hat{x}_i = \sum_{j=1}^{d'} y_{ij}\omega_j$。

考虑整个训练集，原样本点x_i与基于投影重构的样本点$\hat{x}_i$之间的距离为

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} y_{ij}\omega_j - x_i\right\|_2^2 = \sum_{i=1}^{m} {y_i}^T y_i - 2\sum_{i=1}^{m} {y_i}^T W^T x_i + C$$
$$\propto -\mathrm{tr}(W^T(\textstyle\sum_{i=1}^{m} x_i {x_i}^T)W) \tag{3.11}$$

其中，C为常数，$\propto$是正比符号，tr 代表矩阵的迹。

根据最近重构性，式（3.11）应被最小化，考虑到ω_j是标准正交基，$\sum_i x_i {x_i}^T$是协方差矩阵，有

$$\min_{W} -\mathrm{tr}(W^T X X^T W)$$
$$\text{s.t. } W^T W = I \tag{3.12}$$

这就是主成分分析的优化目标。

2）最大可分性

从最大可分性出发，能得到主成分分析的另一种解释。我们知道，样本点在新空间中超平面上的投影是$W^T x_i$，若所有样本点的投影能尽可能分开，则应该使投影后样本点的方差最大化。

投影后样本点的方差是$\sum_i W^T x_i {x_i}^T W$，于是优化目标可写为

$$\max_{W} tr(W^T X X^T W)$$
$$\text{s.t. } W^T W = I \tag{3.13}$$

显然，式（3.13）与（3.12）等价。

对式（3.12）或（3.13）使用拉格朗日乘子法可得

$$XX^T W = \lambda W \tag{3.14}$$

于是，只需对协方差矩阵XX^T进行特征值分解，将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$，再取前$d'$个特征值对应的特征向量构成$W = (\omega_1, \omega_2 \ldots, \omega_{d'})$，这就是主成分分析的解。

PCA 的具体算法流程如下：

设有 m 条 n 维数据。

1）将原始数据按列组成 n 行 m 列矩阵 X

2）将 X 的每一行进行零均值化，即减去这一行的均值

3）求出协方差矩阵$C=\frac{1}{m}XX^T$

4）求出协方差矩阵的特征值及对应的特征向量

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P

6）$Y=PX$即为降维到 k 维后的数据

例 3.11 现有一个二维矩阵如下所示，其中每一行均进行了零均值化，使用 PCA 方法将这组二维数据降到一维。

$$\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}$$

首先求得协方差矩阵：

$$C=\frac{1}{5}\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}\begin{pmatrix}-1 & -2\\ -1 & 0\\ 0 & 0\\ 2 & 1\\ 0 & 1\end{pmatrix}=\begin{pmatrix}\frac{6}{5} & \frac{4}{5}\\ \frac{4}{5} & \frac{6}{5}\end{pmatrix}$$

求其特征值和特征向量，求解后特征值为：$\lambda_1=2,\lambda_2=\frac{2}{5}$，其对应的特征向量经过归一化后如下：

$$\alpha_1=\begin{pmatrix}1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix},\alpha_2=\begin{pmatrix}-1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix}$$

因此矩阵 P 为$P=\begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ -1/\sqrt{2} & 1/\sqrt{2}\end{pmatrix}$

最后用 P 的第一行乘以数据矩阵，得到降维后的表示为：

$$Y=\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}=\begin{pmatrix}-\frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}$$

（2）小波变换

离散小波变换（Discrete Wavelet Transform, DWT）是一种线性信号处理技术，用于数据X时，将它变换成不同的数值小波系数X'。当这种技术用于维归约时，每个样本看作一个 n 维数据向量，即$X=(x_1,x_2,\ldots,x_n)$，描述 n 个数据库属性在样本上的 n 个测量值。

如果小波变换后的数据与原数据的长度相等，这种技术用于数据压缩的关键在于小波变换后的数据可以截短。仅存放一小部分最强的小波系数，就能保留近似的压缩数据。例如，保留大于用户设定的某个阈值的小波系数，其它系数置为0。这样，结果数据表示会非常稀疏。该技术也能用于消除噪音，而不会光滑掉数据的主要特性，也能有效地用于数据清洗（见 3.2.3 节小波变换）。

（3）属性子集选择

属性子集选择通过删除不相关或冗余的属性减少数据量。属性子集选择的目标是找出最小属性集，使得属性的概率分布尽可能地接近使用所有属性的原分布。在缩小的属性集上挖掘，减少了属性的数目，使得模式更易于理解。

属性子集选择的基本启发式方法包括以下技术：

1）逐步向前选择：该过程由空属性集开始，选择原属性集中最好的属性，并将它添加到该集合中。在其后的每一次迭代，将原属性集剩下的属性中的最好的属性添加到该集合中。

2）逐步向后删除：该过程由整个属性集开始，在每一步，删除掉尚在属性集中的最差的属性。

3）逐步向前选择和逐步向后删除的组合：可以将逐步向前选择和逐步向后删除方法结合在一起，每一步选择一个最好的属性，并在剩余属性中删除一个最差的属性。

4）决策树归纳：决策树归纳构造一个类似于流程图的结构，其每个内部（非树叶）结点表示一个属性上的测试，每个分枝对应于测试的一个结果；每个外部（树叶）结点表示一个类预测。在每个结点，算法选择“最好”的属性，将数据划分成类。当决策树归纳用于属性子集选择时，树由给定的数据构造，不出现在树中的所有属性假定是不相关的，出现在树中的属性形成归约后的属性子集。

3.5.2 数量归约

数量归约是用替代的、较小的数据表示形式替换原始数据，可以是参数和非参数的方法。对于参数方法而言，使用模型估计数据，使得一般只需要存放模型参数，而不是实际数据，回归和对数-线性模型就是例子；存放数据归约表示的非参数方法包括直方图、聚类、抽样和数据立方体聚集等。

（1）参数方法

回归和对数线性模型可以用来近似给定数据。在线性回归中，对数据建模，使之适合一条直线。例如，可以用以下公式，将随机变量 y（称作因变量）表示为另一随机变量 x（称为自变量）的线性函数。

$$y = wx + b \tag{3.15}$$

其中，假定 y 的方差是常量，x 和 y 是数据库属性，系数 w 和 b（称作回归

系数）分别为直线的斜率和 y 轴截距，系数可以用最小二乘法求解。多元回归是线性回归的扩展，允许用两个或多个自变量的线性函数对因变量 y 建模。回归方法在第 7 章进行进一步介绍。

对数线性模型（log-linear model）近似离散的多维概率分布。给定 n 个属性样本的集合，我们可以把每个样本看作 n 维空间的点。对于离散属性集，可以使用对数线性模型，基于维组合的一个较小子集，估计多维空间中每个点的概率。这使得高维数据空间可以由较低维空间构造。因此，对数线性模型也可以用于维归约（由于较低维空间的点通常比原来的数据点占据的空间要少）和数据光滑（因为与较高维空间的估计相比，较低维空间的聚集估计受抽样变化的影响较小）。

回归和对数线性模型都可以用于稀疏数据，尽管它们的应用可能是有限的。虽然两种方法都可以处理倾斜数据，但是回归的效果更好。

（2）非参数方法

下面将介绍直方图、聚类和抽样等非参数方法如何用于数量归约。

1）直方图分析：直方图使用分箱近似数据分布，是一种流行的数量归约方式。直方图曾在 3.4.3 介绍过。直方图将属性的数据分布划分为不相交的子集或桶，如果每个桶只代表单个属性值/频率对，则该桶称为单桶。通常，桶表示给定属性的一个连续区间。

对于近似稀疏和稠密数据，以及高倾斜和均匀的数据，直方图都是非常有效的。上面介绍的单属性直方图可以推广到多个属性，多维直方图可以表现属性间的依赖。

例 3.12 下面是高炉炼铁过程中部分顶炉温度的数据，对数据进行四舍五入取整并进行了排序：169，169，169，178，178，178，178，178，183，183，183，189，189，197，197，197，197，197，202，202，202，202，210，210，210，210，210，210，210，218，218，218，218，218，218，218，218，225，225，225，225，225，225。使用单桶直方图进行表示如图 3.13 所示。

图 3.13 使用单值桶显示了这些数据的直方图。为进一步压缩数据，通常让一个桶代表给定属性的一个连续值域。在图 3.14 中每个桶代表顶炉温度的一个个不同的 20 摄氏度区间。

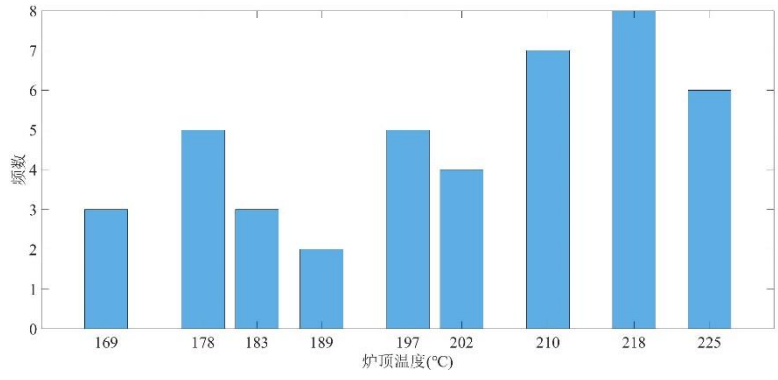

图 3.13 使用单桶的直方图

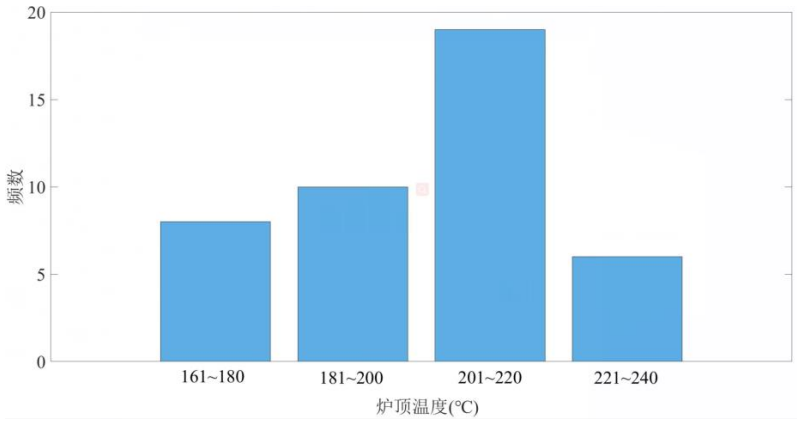

图 3.14 顶炉温度的等宽直方图

2）聚类：聚类技术将数据样本视为对象。它将对象划分为群或簇，使得在一个聚类中的对象相互“相似”，而与其他簇中的对象“相异”。通常，相似性基于距离函数，用对象在空间中的“接近”程度定义。簇的“质量”可以用“直径”表示，直径是簇中两个对象的最大距离。形心距离是簇质量的另一种度量，它定义为簇中每个对象到簇形心（表示“平均对象”，或簇空间中的平均点）的平均距离。

在数据归约时，用数据的簇代表替换实际数据。该技术的有效性依赖于数据的性质。相对于被污染的数据，对于能够组织成不同的簇的数据，该技术有效得多。

3）抽样：抽样可以作为一种数量归约技术使用，因为它允许用数据的小得多的随机样本（子集）表示大型数据集。假定大型数据集 D 包含 N 个样本。我们看看可以用于数据归约的、最常用的对 D 的抽样方法，如图 3.15 所示。

s 个样本的无放回简单随机抽样：从 D 的 N 个样本中抽取 s 个样本（$s<N$），其中 D 中任意样本被抽取的概率均为 $1/N$，即所有样本的抽取是等可能的。

s 个样本的有放回简单随机抽样：该方法类似于无放回简单随机抽样，不同之处在于当一个样本从 D 中抽取后，记录它，然后放回原处。也就是说，一个样本被抽取后，它又被放回 D，以便它可以被再次抽取。

簇抽样：如果 D 中的样本被分组，放入 M 个互不相交的“簇”，则可以得到 s 个簇的简单随机抽样，其中 $s<M$。例如，数据库中样本通常一次取一页，这样每页就可以视为一个簇。例如，可以将无放回简单随机抽样用于页，得到样本的簇样本，由此得到数据的归约表示。也可以利用其他携带更丰富语义信息的聚类标准，例如，在空间数据库中，我们可以基于不同区域位置上的邻近程度定义簇。

分层抽样：如果 D 被划分成互不相交的部分，称为“层”，则通过对每一层的 SRS 就可以得到 D 的分层抽样。特别是当数据倾斜时，这可以帮助确保样本

的代表性。

采用抽样进行数量归约的优点是，得到样本的花费正比于样本集的大小 s，而不是数据集的大小 N。其他数据归约技术至少需要完全扫描 D，对于固定的样本大小，抽样的复杂度仅随着数据的维数 n 线性地增加；而其他技术，如使用直方图，复杂度随 n 呈指数增长。

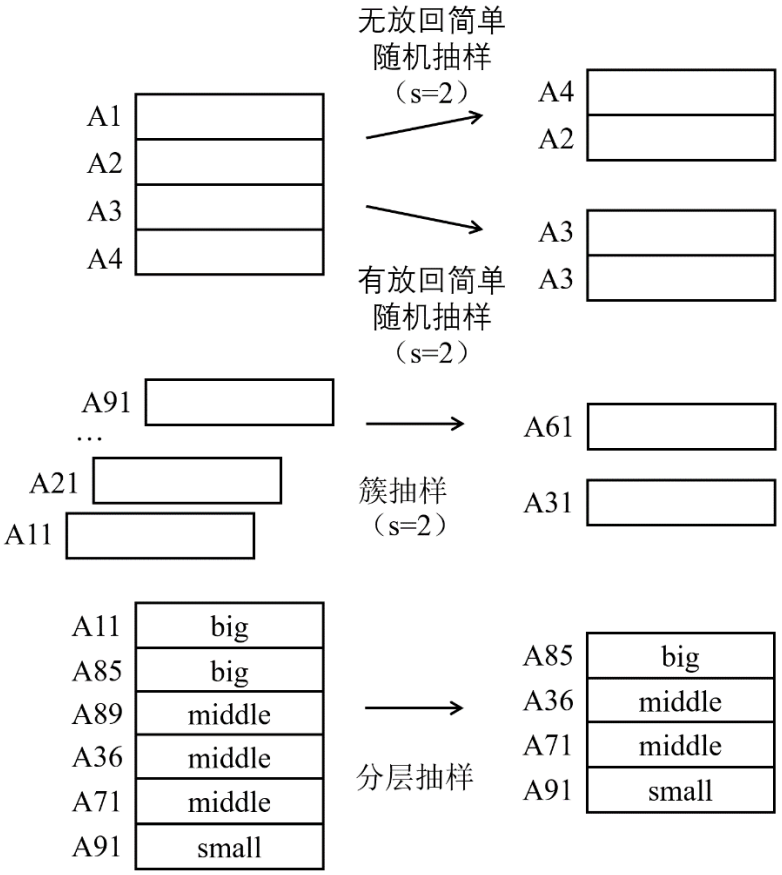

图 3.15 抽样用于数据规约

用于数据归约时，抽样最常用来估计聚集查询的回答。在指定的误差范围内，可以确定（使用中心极限定理）估计一个给定的函数所需的样本大小。样本的大小 s 相对于 N 可能非常小。对于归约数据的逐步求精，抽样是一种自然选择。通过简单地增加样本大小，这样的集合可以进一步求精。

3.6 本章小结

数据预处理为数据挖掘过程提供干净、准确、简洁的数据。本章首先对工业数据质量进行阐述，引出数据预处理的主要任务。其中，数据清洗可以用来清除数据中的噪声，纠正不一致数据。数据集成将数据由多个数据源合并成一个一致的数据存储。数据变换可以用来把数据压缩到较小的区间，或者进行数据离散化。

数据归约技术，如维归约、数量归约，可以使数据进行压缩，并尽量减小信息内容的损失。

数据质量特性：解释了准确性、完整性、一致性、时效性、可信性和可解释性等数据质量特性。

数据预处理的主要任务：现实世界中的数据多半是不完整的、有噪音的和不一致的。数据预处理包括数据清洗、数据集成、数据变换和数据归约。经过这些预处理步骤，有效提升数据质量，保证数据挖掘的效率和结果的有用性。

数据清洗：进行格式内容清洗，填补缺失的值，光滑噪声同时识别异常值，并纠正数据的异常值和逻辑错误。

数据集成：将来自多个数据源的数据整合成一致的数据存储。

数据变换：数据变换通过数据规范化或者离散化将数据变换成适于挖掘的形式。

数据归约：得到数据的归约表示，而使得信息内容的损失最小化。数据归约方法包括维归约和数量归约。维归约减少所考虑的随机变量或维的个数，方法包括主成分分析、小波变换、属性子集选择。数量归约使用参数或非参数模型，得到原数据的较小表示。参数模型包括回归和对数线性模型。非参数方法包括直方图、聚类、抽样等。数据压缩通过规约，得到原数据的压缩表示。如果原数据可以由压缩后的数据重构，而不损失信息，则数据压缩是无损的；如果只能近似重构原数据，则称为有损的。

习题

3-1 数据质量可以从多方面评估，包括准确性、完整性和一致性问题。试结合实际例子，讨论如何基于数据的应用目的评估数据质量，并提出数据质量的其他评估尺度。

3-2 如果不进行数据预处理而直接进行数据挖掘可能带来哪些问题？

3-3 在现实世界的数据中，某些属性会存在缺失值，讨论处理这一问题的方法。

3-4 如果某一属性包括如下值（以递增序）：13，15，16，16，19，20，20，21，22，22，25，25，25，25，30，33，33，35，35，35，35，36，40，45，46，52，70。

a）使用深度为 3 的箱，用箱均值光滑以上数据。说明你的步骤，讨论这种技术对给定数据的效果。

b）如何确定该数据中的异常值？

3-5 光滑数据有哪些方法？请简要阐述。

3-6 讨论数据集成需要考虑的问题。

3-7 数据变换策略有哪几种？请分别阐述。

3-8 如下规范化方法的值域是什么？

a）最小-最大规范化。

b）z 分数规范化。

c）小数定标规范化。

3-9 使用如下方法规范化如下数据组：200，300，400，600，1000

a）令 min=0，max=1，最小-最大规范化。

b）z 分数规范化。

c）小数定标规范化。

3-10 使用习题 3.4 中给出的数据，回答以下问题：

a）使用最小最大规范化将 35 变换到[0.0,1.0]区间。

b）使用 z 分数规范化，变换数据，其中标准差为 12.94。

c）使用小数定标规范化变换 35。

d）指出给定数据，你愿意使用哪种方法。陈述你的理由。

3-11 什么是数据归约？数据归约大致可分为哪几类？

3-12 讨论主成分分析的主要步骤。

参考文献

[1] 宋万清,杨寿渊,陈剑雪,高永彬.数据挖掘[M]. 北京：中国铁道出版社, 2018.12

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

[3] Little R J A, Rubin D B. Statistical analysis with missing data[M]. John Wiley & Sons, 2019.

[4] 陈封能.数据挖掘导论（完整版）[M]. 北京：人民邮电出版社, 2011.

[5] 胡广书.现代信号处理教程[M].北京：清华大学出版社, 2004.

[6] 周志华. 机器学习[M].北京：清华大学出版社,2016.

[7] 李航.统计学习方法[M].北京：清华大学出版社,2019.

[8] 赵春晖.大数据解析与应用导论[M]. 北京：化学工业出版社,2022.

[9] Redman T C. Data Quality: The Field Guide[M]. Digital Press, 2001.

[10] Walczak B, Massart D L. Noise suppression and signal compression using the wavelet packet transform[J]. Chemometrics and Intelligent Laboratory Systems, 1997, 36(2): 81-94.

[11] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J].

Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences, 1998, 454(1971): 903-995.

[12] Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 1901, 2(11): 559-572.

[13] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of educational psychology, 1933, 24(6): 417.

[14] Kosanovich K A, Piovoso M J. PCA of wavelet transformed process data for monitoring[J]. Intelligent Data Analysis, 1997, 1(1-4): 85-99.

[15] Wise B M, Ricker N L, Veltkamp D F, et al. A theoretical basis for the use of principal component models for monitoring multivariate processes[J]. Process Control and Quality, 1990, 1(1): 41-51.