数据预处理

数据清洗、数据集成、数据变换和数据规约。

数据清洗

主要是删除原始数据集中的无关数据、重复数据，平滑噪声数据，筛掉与挖掘主题无关的数据，处理缺失值、异常值等。

缺失值处理

删除记录、数据插补和不处理。

插补方法	方法描述
均值/中位数/众数插补	根据属性值的类型，用该属性取值的平均数/中位数/众数进行插补
使用固定值	将缺失的属性值用一个常量替换。如广州一个工厂普通外来务工人员的“基本工资”属性的空缺值可以用2015年广州市普通外来务工人员工资标准1895元/月，该方法就是使用固定值
最近临插补	在记录中找到与缺失样本最接近的样本的该属性值插补
回归方法	对带有缺失值的变量，根据已有数据和与其有关的其他变量（因变量）的数据建立拟合模型来预测缺失的属性值
插值法	插值法是利用已知点建立合适的插值函数$f(x)$，未知值由对应点x_i求出的函数值$f(x_i)$近似代替

如果通过简单的删除小部分记录达到既定的目标，那么删除含有缺失值的记录的方法是最有效的。然而，这种方法却有很大的局限性。它是以减少历史数据来换取数据的完备，会造成资源的大量浪费，将丢弃了大量隐藏在这些记录中的信息。尤其在数据集本来就包含很少记录的情况下，删除少量记录可能会严重影响到分析结果的客观性和正确性。一些模型可以将缺失值视作一种特殊的取值，允许直接在含有缺失值的数据上进行建模。

（1）拉格朗日插值法

根据数学知识可知，对于平面上已知的n个点（无两点在一条直线上）可以找到一个$n-1$次多项式$y=a_0+a_1x+a_2x^2+\cdots+a_{n-1}x^{n-1}$，使此多项式曲线过这$n$个点。

1）求已知的过n个点的$n-1$次多项式：

$$y=a_0+a_1x+a_2x^2+\cdots+a_{n-1}x^{n-1} \tag{4-1}$$

将n个点的坐标$(x_1,y_1),(x_2,y_2)\cdots(x_n,y_n)$代入多项式函数，得

$$\begin{aligned}
y_1&=a_0+a_1x_1+a_2x_1^2+\cdots+a_{n-1}x_1^{n-1}\\
y_2&=a_0+a_1x_1+a_2x_2^2+\cdots+a_{n-1}x_2^{n-1}\\
&\cdots\cdots\\
y_n&=a_0+a_1x_n+a_2x_n^2+\cdots+a_{n-1}x_n^{n-1}
\end{aligned}$$

解出拉格朗日插值多项式为：

$$\begin{aligned}
L(x)=&y_1\frac{(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}\\
&+y_2\frac{(x-x_1)(x-x_3)\cdots(x-x_n)}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_n)}\\
&+\cdots\cdots\\
&+y_n\frac{(x-x_1)(x-x_2)\cdots(x-x_{n-1})}{(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}
\end{aligned} \tag{4-2}$$

$$= \sum_{i=0}^{n} y_i \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

2）将缺失的函数值对应的点 x 代入插值多项式得到缺失值的近似值 $L(x)$。

拉格朗日插值公式结构紧凑，在理论分析中很方便，但是当插值节点增减时，插值多项式就会随之变化，这在实际计算中是很不方便的，为了克服这一缺点，提出了牛顿插值法。

（2）牛顿插值法

1）求已知的 n 个点对 $(x_1, y_1), (x_2, y_2)\cdots(x_n, y_n)$ 的所有阶差商公式

$$f[x_1, x] = \frac{f[x] - f[x_1]}{x - x_1} = \frac{f(x) - f(x_1)}{x - x_1} \tag{4-3}$$

$$f[x_2, x_1, x] = \frac{f[x_1, x] - f[x_2, x_1]}{x - x_2} \tag{4-4}$$

$$f[x_3, x_2, x_1, x] = \frac{f[x_2, x_1, x] - f[x_3, x_2, x_1]}{x - x_3} \tag{4-5}$$

$$\cdots\cdots$$

$$f[x_n, x_{n-1}, \cdots, x_1, x] = \frac{f[x_{n-1}, \cdots, x_1, x] - f[x_n, x_{n-1}, \cdots, x_1]}{x - x_n} \tag{4-6}$$

2）联立以上差商公式建立如下插值多项式 $f(x)$

$$\begin{aligned} f(x) &= f(x_1) + (x-x_1)f[x_2, x_1] + (x-x_1)(x-x_2)f[x_3, x_2, x_1] + \\ &\quad (x-x_1)(x-x_2)(x-x_3)f[x_4, x_3, x_2, x_1] + \cdots + \\ &\quad (x-x_1)(x-x_2)\cdots(x-x_{n-1})f[x_n, x_{n-1}, \cdots x_2, x_1] + \\ &\quad (x-x_1)(x-x_2)\cdots(x-x_n)f[x_n, x_{n-1}, \cdots x_1, x] \\ &= P(x) + R(x) \end{aligned} \tag{4-7}$$

其中：

$$\begin{aligned} P(x) &= f(x_1) + (x-x_1)f[x_2, x_1] + (x-x_1)(x-x_2)f[x_3, x_2, x_1] + \\ &\quad (x-x_1)(x-x_2)(x-x_3)f[x_4, x_3, x_2, x_1] + \cdots + \\ &\quad (x-x_1)(x-x_2)\cdots(x-x_{n-1})f[x_n, x_{n-1}, \cdots, x_2, x_1] \end{aligned} \tag{4-8}$$

$$R(x) = (x-x_1)(x-x_2)\cdots(x-x_n)f[x_n, x_{n-1}, \cdots\ x_1, x] \tag{4-9}$$

$P(x)$ 是牛顿插值逼近函数，$R(x)$ 是误差函数。

3）将缺失的函数值对应的点 x 代入插值多项式得到缺失值的近似值 $f(x)$。

牛顿插值法也是多项式插值，但采用了另一种构造插值多项式的方法，与拉格朗日插值相比，具有承袭性和易于变动节点的特点。从本质上来说，两者给出的结果是一样的（相同次数、相同系数的多项式），只不过表示的形式不同。因此，在 Python 的 Scipy 库中，只提供了拉格朗日插值法的函数（因为实现上比较容易），如果需要牛顿插值法，则需要自行编写

异常值处理

将含有异常值的记录直接删除的方法简单易行，但缺点也很明显，在观测值很少的情况下，这种删除会造成样本量不足，可能会改变变量的原有分布，从而造成分析结果的不准确。视为缺失值处理的好处是可以利用现有变量的信息，对异常值（缺失值）进行填补。

异常值处理方法	方法描述
删除含有异常值的记录	直接将含有异常值的记录删除
视为缺失值	将异常值视为缺失值，利用缺失值处理的方法进行处理
平均值修正	可用前后两个观测值的平均值修正该异常值
不处理	直接在具有异常值的数据集上进行挖掘建模

数据集成

数据挖掘需要的数据往往分布在不同的数据源中，数据集成就是将多个数据源合并存放在一个一致的数据存储（如数据仓库）中的过程。

在数据集成时，来自多个数据源的现实世界实体的表达形式是不一样的，有可能不匹配，要考虑实体识别问题和属性冗余问题，从而将源数据在最低层上加以转换、提炼和集成。

实体识别

冗余属性识别

数据集成往往导致数据冗余，例如，

1）同一属性多次出现；

2）同一属性命名不一致导致重复。

数据变换

简单函数变换

简单函数变换是对原始数据进行某些数学函数变换，常用的变换包括平方、开方、取对数、差分运算等，即：

$$x' = x^2 \tag{4-10}$$

$$x' = \sqrt{x} \tag{4-11}$$

$$x' = \log(x) \tag{4-12}$$

$$\nabla f(x_k) = f(x_{k+1}) - f(x_k) \tag{4-13}$$

简单的函数变换常用来将不具有正态分布的数据变换成具有正态分布的数据。在时间序列分析中，有时简单的对数变换或者差分运算就可以将非平稳序列转换成平稳序列。

规范化

数据规范化（归一化）处理是数据挖掘的一项基础工作。不同评价指标往往具有不同的量纲，数值间的差别可能很大，不进行处理可能会影响到数据分析的结果。为了消除指标之间的量纲和取值范围差异的影响，需要进行标准化处理，将数据按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析。

（1）最小－最大规范化

最小－最大规范化也称为离差标准化，是对原始数据的线性变换，将数值值映射到 [0,1]

之间。

转换公式如下：

$$x^* = \frac{x - min}{max - min} \tag{4-14}$$

其中，*max* 为样本数据的最大值，*min* 为样本数据的最小值。*max–min* 为极差。离差标准化保留了原来数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法。这种处理方法的缺点是若数值集中且某个数值很大，则规范化后各值会接近于 0，并且将会相差不大。若将来遇到超过目前属性 [*min*, *max*] 取值范围的时候，会引起系统出错，需要重新确定 *min* 和 *max*。

（2）零 – 均值规范化

零 – 均值规范化也称标准差标准化，经过处理的数据的均值为 0，标准差为 1。转化公式为：

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{4-15}$$

其中 $\bar{x}$ 为原始数据的均值，σ 为原始数据的标准差，是当前用得最多的数据标准化方法。

（3）小数定标规范化

通过移动属性值的小数位数，将属性值映射到 [–1,1] 之间，移动的小数位数取决于属性值绝对值的最大值。

转化公式为：

$$x^* = \frac{x}{10^k} \tag{4-16}$$

简单属性离散化

1. 离散化的过程

连续属性的离散化就是在数据的取值范围内设定若干个离散的划分点，将取值范围划分为一些离散化的区间，最后用不同的符号或整数值代表落在每个子区间中的数据值。所以，离散化涉及两个子任务：确定分类数以及如何将连续属性值映射到这些分类值。

2. 常用的离散化方法

常用的离散化方法有等宽法、等频法和（一维）聚类。

（1）等宽法

将属性的值域分成具有相同宽度的区间，区间的个数由数据本身的特点决定，或者由用户指定，类似于制作频率分布表。

（2）等频法

将相同数量的记录放进每个区间。

这两种方法简单，易于操作，但都需要人为地规定划分区间的个数。同时，等宽法的缺点在于它对离群点比较敏感，倾向于不均匀地把属性值分布到各个区间。有些区间包含许多数据，而另外一些区间的数据极少，这样会严重损坏建立的决策模型。等频法虽然避免了上述问题的产生，却可能将相同的数据值分到不同的区间以满足每个区间中固定的数据个数。

（3）基于聚类分析的方法

一维聚类的方法包括两个步骤，首先将连续属性的值用聚类算法（如 K-Means 算法）进行聚类，然后再将聚类得到的簇进行处理，合并到一个簇的连续属性值并做同一标记。聚类分析的离散化方法也需要用户指定簇的个数，从而决定产生的区间数。

属性构造

在数据挖掘的过程中，为了提取更有用的信息，挖掘更深层次的模式，提高挖掘结果的精度，我们需要利用已有的属性集构造出新的属性，并加入到现有的属性集合中。

小波分析

（1）基于小波变换的特征提取方法

基于小波变换的特征提取方法主要有：基于小波变换的多尺度空间能量分布特征提取、基于小波变换的多尺度空间的模极大值特征提取、基于小波包变换的特征提取、基于适应性小波神经网络的特征提取，详见表 4-5。

表4-5　基于小波变换的特征提取方法

基于小波变换的特征提取方法	方 法 描 述
基于小波变换的多尺度空间能量分布特征提取方法	各尺度空间内的平滑信号和细节信号能提供原始信号的时频局域信息，特别是能提供不同频段上信号的构成信息。把不同分解尺度上信号的能量求解出来，就可以将这些能量尺度顺序排列，形成特征向量供识别用
基于小波变换的多尺度空间的模极大值特征提取方法	利用小波变换的信号局域化分析能力，求解小波变换的模极大值特性来检测信号的局部奇异性，将小波变换模极大值的尺度参数 s、平移参数 t 及其幅值作为目标的特征量
基于小波包变换的特征提取方法	利用小波分解，可将时域随机信号序列映射为尺度域各子空间内的随机系数序列，按小波包分解得到的最佳子空间内随机系数序列的不确定性程度最低，将最佳子空间的熵值及最佳子空间在完整二叉树中的位置参数作为特征量，可以用于目标识别
基于适应性小波神经网络的特征提取方法	基于适应性小波神经网络的特征提取方法可以把信号通过分析小波拟合表示，进行特征提取

（2）小波基函数

小波基函数是一种具有局部支集的函数，并且平均值为 0，小波基函数满足 $\psi(0)=\int\psi(t)dt=0$。常用的小波基有 Haar 小波基、db 系列小波基等。

（3）小波变换

对小波基函数进行伸缩和平移变换：

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \tag{4-18}$$

其中，a 为伸缩因子，b 为平移因子。

任意函数 $f(t)$ 的连续小波变换（CWT）为：

$$W_f(a,b)=|a|^{-1/2}\int f(t)\psi\left(\frac{t-b}{a}\right)dt \tag{4-19}$$

可知，连续小波变换为 $f(t)\rightarrow W_f(a,b)$ 的映射，对小波基函数 $\psi(t)$ 增加约束条件 $C_\psi=\int\frac{|\widehat{\psi}(t)|^2}{t}dt<\infty$，就可以由 $W_f(a,b)$ 逆变换得到 $f(t)$。其中 $\widehat{\psi}(t)$ 为 $\psi(t)$ 的傅里叶变换。

其逆变换为：

$$f(t)=\frac{1}{C_\psi}\iint\frac{1}{a^2}W_f(a,b)\psi\left(\frac{t-b}{a}\right)da\cdot db \tag{4-20}$$

（4）基于小波变换的多尺度空间能量分布特征提取方法

应用小波分析技术可以把信号在各频率波段中的特征提取出来，基于小波变换的多尺度空间能量分布特征提取方法是对信号进行频带分析，再分别以计算所得的各个频带的能量作为特征向量。

信号 $f(t)$ 的二进小波分解可表示为：

$$f(t) = A^j + \Sigma D^j \tag{4-21}$$

其中 A 是近似信号，为低频部分；D 是细节信号，为高频部分，此时信号的频带分布如图 4-6 所示。

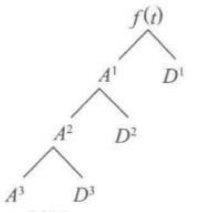
信号的总能量为：

$$E = EA_j + \Sigma ED_j \tag{4-22}$$

选择第 j 层的近似信号和各层的细节信号的能量作为特征，构造特征向量：

$$F = [EA_j, ED_1, ED_2, \cdots, ED_j] \tag{4-23}$$

利用小波变换可以对声波信号进行特征提取，提取出可以代表声波信号的向量数据，即完成从声波信号到特征向量数据的变换。本例利用小波函数对声波信号数据进行分解，得到 5 个层次的小波系数。利用这些小波系数求得各个能量值，这些能量值即可作为声波信号的特征数据。



数据规约

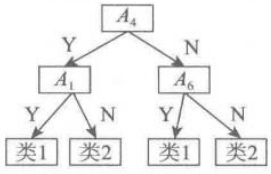
在大数据集上进行复杂的数据分析和挖掘需要很长的时间，数据规约产生更小但保持原数据完整性的新数据集。在规约后的数据集上进行分析和挖掘将更有效率。

数据规约的意义在于：

- 降低无效、错误数据对建模的影响，提高建模的准确性；
- 少量且具代表性的数据将大幅缩减数据挖掘所需的时间；
- 降低储存数据的成本。

属性规约

属性规约通过属性合并来创建新属性维数，或者直接通过删除不相关的属性（维）来减少数据维数，从而提高数据挖掘的效率、降低计算成本。属性规约的目标是寻找出最小的属性子集并确保新数据自己的概率分布尽可能地接近原来数据集的概率分布。

属性规约方法	方法描述	方法解析
合并属性	将一些旧属性合为新属性	初始属性集：$\{A_1, A_2, A_3, A_4, B_1, B_2, B_3, C\}$ $\{A_1, A_2, A_3, A_4\} \rightarrow$ A $\{B_1, B_2, B_3\} \rightarrow$ B ⇒规约后属性集：$\{A, B, C\}$
逐步向前选择	从一个空属性集开始，每次从原来属性集合中选择一个当前最优的属性添加到当前属性子集中。直到无法选择出最优属性或满足一定阈值约束为止	初始属性集：$\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\{\} \Rightarrow \{A_1\} \Rightarrow \{A_1, A_4\}$ ⇒规约后属性集：$\{A_1, A_4, A_6\}$
逐步向后删除	从一个全属性集开始，每次从当前属性子集中选择一个当前最差的属性并将其从当前属性子集中消去。直到无法选择出最差属性为止或满足一定阈值约束为止	初始属性集：$\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\} \Rightarrow \{A_1, A_4, A_5, A_6\}$ ⇒规约后属性集：$\{A_1, A_4, A_6\}$
决策树归纳	利用决策树的归纳方法对初始数据进行分类归纳学习，获得一个初始决策树，所有没有出现在这个决策树上的属性均可认为是无关属性，因此将这些属性从初始集合中删除，就可以获得一个较优的属性子集	初始属性集：$\{A_1, A_2, A_3, A_4, A_5, A_6\}$  ⇒规约后属性集：$\{A_1, A_4, A_6\}$

主成分分析	用较少的变量去解释原始数据中的大部分变量，即将许多相关性很高的变量转化成彼此相互独立或不相关的变量	详见下面计算步骤

逐步向前选择、逐步向后删除和决策树归纳是属于直接删除不相关属性（维）方法。主成分分析是一种用于连续属性的数据降维方法，它构造了原始数据的一个正交变换，新空间的基底去除了原始空间基底下数据的相关性，只需使用少数新变量就能够解释原始数据中的大部分变异。在应用中，通常是选出比原始变量个数少，能解释大部分数据中的变量的几个新变量，即所谓主成分，来代替原始变量进行建模。

1）设原始变量 $X_1, X_2, \cdots, X_p$ 的 n 次观测数据矩阵为：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(X_1, X_2, \cdots, X_P) \tag{4-24}$$

2）将数据矩阵按列进行中心标准化。为了方便，将标准化后的数据矩阵仍然记为 X。

3）求相关系数矩阵 R, $R=(r_{ij})_{p\times p}$，r_{ij} 的定义为：

$$r_{ij}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)/\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2} \tag{4-25}$$

其中，$r_{ij}=r_{ji}$, $r_{ii}=1$。

4）求 R 的特征方程 $\det(R-\lambda E)=0$ 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_p > 0$。

5）确定主成分个数 m：$\dfrac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geqslant \alpha$，$\alpha$ 根据实际问题确定，一般取 80%。

6）计算 m 个相应的单位特征向量：

$$\beta_1=\begin{bmatrix}\beta_{11}\\ \beta_{21}\\ \vdots \\ \beta_{p1}\end{bmatrix}, \beta_2=\begin{bmatrix}\beta_{12}\\ \beta_{22}\\ \vdots \\ \beta_{p2}\end{bmatrix}, \cdots, \beta_m=\begin{bmatrix}\beta_{m2}\\ \beta_{m2}\\ \vdots \\ \beta_{m2}\end{bmatrix} \tag{4-26}$$

7）计算主成分：

$$Z_i=\beta_{1i}X_1+\beta_{2i}X_2+\cdots+\beta_{pi}X_p,\ i=1, 2, \cdots, m \tag{4-27}$$

数值规约

数值规约指通过选择替代的、较小的数据来减少数据量，包括有参数方法和无参数方法两类。有参数方法是使用一个模型来评估数据，只需存放参数，而不需要存放实际数据，例如回归（线性回归和多元回归）和对数线性模型（近似离散属性集中的多维概率分布）。无参数方法就需要存放实际数据，例如直方图、聚类、抽样（采样）。

（1）直方图

直方图使用分箱来近似数据分布，是一种流行的数据规约形式。属性 A 的直方图将 A 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。

（2）聚类

聚类技术将数据元组（即记录，数据表中的一行）视为对象。它将对象划分为簇，使一个簇中的对象相互“相似”，而与其他簇中的对象“相异”。在数据规约中，用数据的簇替换实际数据。该技术的有效性依赖于簇的定义是否符合数据的分布性质。

（3）抽样

抽样也是一种数据规约技术，它用比原始数据小得多的随机样本（子集）表示原始数据集。假定原始数据集 D 包含 N 个元组，可以采用抽样方法对 D 进行抽样。下面介绍常用的

抽样方法。

s 个样本无放回简单随机抽样：从 D 的 N 个元组中抽取 s 个样本（$s < N$），其中 D 中任意元组被抽取的概率均为 $1/N$，即所有元组的抽取是等可能的。

s 个样本有放回简单随机抽样：该方法类似于无放回简单随机抽样，不同在于每次一个元组从 D 中抽取后，记录它，然后放回原处。

聚类抽样：如果 D 中的元组分组放入 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，数据库中元组通常一次检索一页，这样每页就可以视为一个簇。

分层抽样：如果 D 划分成互不相交的部分，称作层，则通过对每一层的简单随机抽样就可以得到 D 的分层样本。例如，可以得到关于顾客数据的一个分层样本，按照顾客的每个年龄组创建分层。

用于数据规约时，抽样最常用来估计聚集查询的结果。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。通常样本的大小 s 相对于 N 非常小。而通过简单地增加样本大小，这样的集合可以进一步求精。

（4）参数回归

简单线性模型和对数线性模型可以用来近似描述给定的数据。（简单）线性模型对数据建模，使之拟合一条直线。

对数线性模型一般用来近似离散的多维概率分布。在一个 n 元组的集合中，每个元组可以看作是 n 维空间中的一个点。可以使用对数线性模型基于维组合的一个较小子集，估计离散化的属性集的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规约（由于低维空间的点通常比原来的数据点占据较少的空间）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较少受抽样方差的影响）。

函 数 名	函 数 功 能	所属扩展库
interpolate	一维、高维数据插值	Scipy
unique	去除数据中的重复元素，得到单值元素列表，它是对象的方法名	Pandas/Numpy
isnull	判断是否空值	Pandas
notnull	判断是否非空值	Pandas
PCA	对指标变量矩阵进行主成分分析	Scikit-Learn
random	生成随机矩阵	Numpy