大数据技术及应用

王迪

d.wang@sjtu.edu.cn

机械与动力工程学院

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

04 数据预处理

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

1 “肮脏”的数据

- 原始业务数据（或数据仓库）是数据挖掘的信息来源，但往往存在以下问题：
  - 不完整
    - 数据收集时缺乏合适的值
    - 数据收集时和数据分析时考虑因素不同
    - 人为/硬件/软件问题
  - 有噪声
    - 数据收集工具造成
    - 数据输入时人为/计算机错误
    - 数据传输产生错误
  - 不一致
    - 数据源不同
    - 缺乏统一的分类标准和信息的编码方案

1 “肮脏”的数据

- 重复
  - 同一事物在数据库中存在两条或多条完全相同的记录
  - 相同的信息冗余地存在于多个数据源中
- 维度高
  - 在一次数据挖掘中，只需要一部分属性就可以得到期望知道的知识
- 不平衡
  - 某类样本数量明显少于其他类样本数量的数据集

- 数据预处理根据用户需求，确定挖掘任务，采用合适的方法重新组织原始数据，为数据挖掘过程提供干净、准确、简洁的数据，提高数据挖掘的效率与准确性！

2 数据预处理方法

- 数据清洗（Data Cleaning）
  - 填充空缺值
  - 识别孤立点
  - 去掉原始数据中的噪声和无关数据

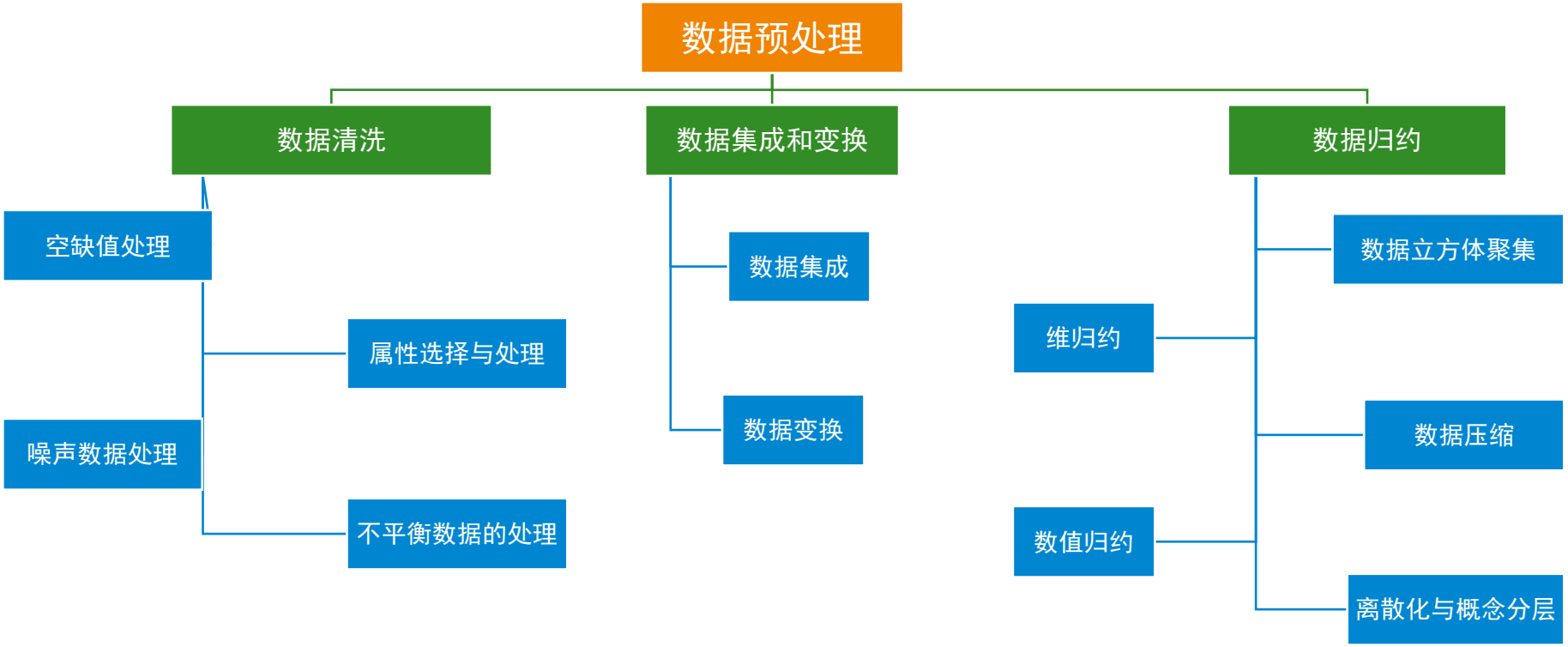
- 数据集成（Data Integration）
  - 将多个数据源中的数据结合起来存放在一个一致的数据存储中
  - 涉及多个数据源的数据匹配问题，数值冲突问题和数据的冗余问题等

2 数据预处理方法

- 数据变换（Data Transformation）
  - 把原始数据转换成为适合数据挖掘的形式
  - 对数据的汇总、聚集、概化、规范化等
  - 进行属性的构造

- 数据归约（Data reduction）
  - 产生数据的归约表示
  - 使得数据量减小，更适合于数据挖掘算法的需要，并且能够得到和原始数据相同的分析结果
  - 包括数据立方体聚集、维归约、数据压缩、数值归约、离散化和概念分层等

2 数据预处理方法


数据预处理
数据清洗
数据集成和变换
数据归约
空缺值处理
属性选择与处理
噪声数据处理
不平衡数据的处理
数据集成
数据变换
维归约
数值归约
数据立方体聚集
数据压缩
离散化与概念分层

3 数据清洗

- （1）空缺值处理
  - 忽略该记录
  - 去掉该属性
  - 写空缺值
    - 依据背景资料，手工填写
  - 使用默认值
    - Unknown
    - 挖掘算法可能认为形成了一个有用的知识
  - 使用属性平均值
  - 使用同类样本平均值
  - 预测最可能的值
    - 从现有数据的多个信息推测空缺值
    - 根据其他完整的记录数据，使用一定的预测方法，得到最可能的预测值
    - 一些数据挖掘算法在处理空值方面的能力比较强，如决策树算法、关联规则算法等，能够快速产生较为准确的知识模型！

3 数据清洗

- （2）属性选择与处理
  - 从原始数据中选取合适的属性进行数据挖掘
  - 选取原则
    - 尽可能赋予属性名和属性值明确的含义
    - 统一多数据源的属性值编码
      - 保证在各个数据源中对同一事物特征的描述是统一，如男、女，0、1，M、F等
    - 处理唯一属性
      - 原始数据中的关键属性或唯一属性对数据挖掘是无用的，如ID，姓名等
    - 去除重复属性
      - 原始数据中会出现意义相同或者可以用于表示同一信息的多个属性，如年龄和出生日期
    - 去除可忽略字段
      - 当一个属性缺失非常严重时
    - 合理选择关联字段
      - 如果属性X可以由另一个或多个属性推导或者计算出来，则认为这些字段之间的关联度高
      - 属性和它的关联属性可以选择其一，如商品的价格、数量和总价格，月薪与年薪

3 数据清洗

- （3）噪声数据处理
  - 在测量一个变量时可能产生一些误差或者错误，使得测量值相对于真实值有一定的偏差，这种偏差称之为噪声
  - 处理方法
    - 分箱
    - 聚类
    - 回归

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 通过考察相邻数据来确定最终值
  - 把待处理的数据（某列属性值）按照一定的规则放进一些箱子中，考察每一个箱子中的数据，采用某种方法分别对各个箱子中的数据进行处理
    - 数据排序
    - 确定箱子个数/每组个数（深度）
    - 采用分箱方法（统一权重，统一区间，最小熵，用户自定义区间）
    - 平滑处理（对每一个数据）
  - 箱子：按照属性值划分的子区间，如果一个属性值处于某个子区间范围内，就把该属性值放进这个子区间代表的箱子内

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 统一权重
    - 又称等深分箱法
    - 每箱有相同的记录数
    - 每箱记录数称为箱的权重
  - 统一区间
    - 也称等宽分箱法
    - 使数据集在整个属性值的区间上平均分布
    - 每个箱的区间范围是一个常量

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 最小熵
    - 使在各区间分组内的记录具有最小的熵
    - 熵是信息理论中数据无序程度的度量标准
    - 数据集的熵越低，说明数据之间的差异越小
    - 最小熵划分就是为了使每箱中的数据具有最好的相似性
    - 最小熵方法得到的各个分箱的全体，应该是各种分箱可能中，具有最小熵的分箱结果
  - 用户自定义区间
    - 按用户的需求定义某些希望观察的区间

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 数据平滑处理
    - 按平均值：对同一箱中的数据求平均值，并用这个平均值替代该箱中的所有数据
    - 按边界值：对于箱子中的每个数据，观察它与箱子两个边界值的距离，并用距离较小的那个边界值替代该数据
    - 按中值：取箱子的中值，来替代本箱中的所有数据，如果个数是偶数，用中间两个数的平均值

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 例：一个表的客户收入字段（属性），共16条记录

    5000，800，1000，2000，1800，2300，2500，3500

    4800，4500，1200，1500，2800，3000，1500，4000
  - 处理步骤：
    - 排序

      800，1000，1200，1500，1500，1800，2000，2300，

      2500，2800，3000，3500，4000，4500，4800，5000
    - 分箱个数
      - 统一权重：箱子深度为4（箱子里的数目）
      - 统一区间：箱子的数目为4
      - 自定义：箱子数目为5

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 统一权重：箱子深度为4（箱子里的数目）

    箱1：800，1000，1200，1500

    箱2：1500，1800，2000，2300，

    箱3：2500，2800，3000，3500

    箱4：4000，4500，4800，5000
  - 统一区间：箱子的数目为4
    - 数据取值范围为[800，5000]，每个箱子的宽度为（5000-800）/4，得到4个宽度相等的子区间：[800，1850]，(1850，2900]，(2900，3950]，(3950，5000]

      箱1：800，1000，1200，1500，1500，1800

      箱2：2000，2300，2500，2800

      箱3：3000，3500

      箱4：4000，4500，4800，5000

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 自定义：箱子数目为5
    - 各箱子区间为：
      - [800，1000]，(1000，2000]，(2000，3000]，(3000，4000]，(4000，5000]

箱1：800, 1000

箱2：1200，1500，1500，1800，2000

箱3：2300，2500，2800，3000

箱4：3500，4000

箱5：4500，4800，5000

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 数据平滑处理
    - 例：统一区间后的箱子

      箱1：800，1000，1200，1500，1500，1800

      箱2：2000，2300，2500，2800

      箱3：3000，3500

      箱4：4000，4500，4800，5000
    - 按平均值平滑

      箱1：1300，1300，1300，1300，1300，1300

      箱2：2400，2400, 2400, 2400

      箱3：3250，3250

      箱4：4575, 4575, 4575, 4575

3 数据清洗

- （3）噪声数据处理——分箱（binning）
  - 数据平滑处理
    - 按边界值平滑

      箱1：800，800，800，1800，1800，1800

      箱2：2000，2000，2800，2800

      箱3：3000，3500

      箱4：4000，4000，5000，5000
    - 按中值平滑

      箱1：1350，1350，1350，1350，1350，1350

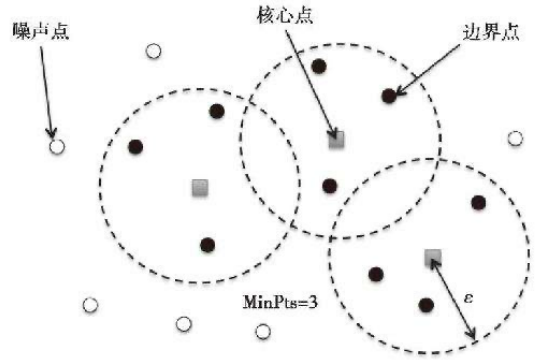
      箱2：2400，2400，2400，2400

      箱3：3250，3250

      箱4：4650，4650，4650，4650

3 数据清洗

- （3）噪声数据处理——聚类（clustering）
  - 聚类
    - 将物理的或抽象对象的集合分组为由类似的对象组成的多个类的过程
    - 聚类的结果是生成一组由数据对象组成的集合，成为簇
    - 同一簇中的对象具有相似性，并且一个对象与同簇中任何一个对象之间的相似性一定强于它与其他簇中任何一个对象之间的相似性
  - 要找到那些落在簇之外的值，称为孤立点，这些孤立点被视为噪声
  - 聚类方法不需要任何先验知识（无监督学习）



3 数据清洗

- （3）噪声数据处理——回归（regression）
  - 回归
    - 回归试图发现相关的变量之间的变化模式
    - 通过一个函数来拟合平滑数据，即通过建立数据模型来预测下一个数值
    - 回归方法分为线性回归和非线性回归
      - $Y=aX+b$　　一元线性回归
      - $Z=aX+bY+c$　　多元线性回归
      - $Y=a+bX+cX^2$　　非线性回归

3 数据清洗

- （4）不平衡数据处理
  - 各类样本数目不平衡情况下的分类学习
  - 如二分类中的正类的学习样本比负类的学习样本多得多
  - 如欺诈识别、入侵检测、医疗诊断以及文本分类等都是典型的不平衡数据问题
  - 基本思想是通过改变训练数据的分布来消除或减小数据的不平衡
  - 处理方法
    - 过抽样
      - 通过增加少数类样本来提高少数类的分类性能
      - 如复制少数类样本
      - 没有给少数类增加任何新的信息，而且可能会导致过度拟合
    - 欠抽样
      - 通过减少多数类样本来提高少数类的分类性能
      - 通过随机地去掉一些多数类样本来减少多数类的规模
      - 会丢失多数类的一些重要信息

4 数据集成与变换

- （1）数据集成
  - 把多个数据存储合并起来
  - 涉及数据的冲突问题和不一致数据的处理问题
    - 模式匹配
    - 数据冗余
    - 数据值冲突

4 数据集成与变换

- （1）数据集成
  - 模式匹配
    - 例：用户希望发现客户背景和客户购买类型、购买力的关系

表 4-1 客户基本情况表

属性名称	数据类型	说 明
id	Short int	客户标志
gender	boolean	性别
birth	date	出生日期
type	boolean	是否会员
income	Short int	月收入(元)

表 4-2 客户交易数据表

属性名称	数据类型	说 明
customer_id	int	客户标志
time	date	交易日期
goods	string	商品名称
price	real	商品价格
count	short int	商品数量
total_price	real	总价格

4 数据集成与变换

- （1）数据集成
  - 数据冗余
    - 重复：多个相同的记录
    - 冗余属性：一个属性可以由其他属性推导得出

表 4-1 客户基本情况表

属性名称	数据类型	说 明
id	Short int	客户标志
gender	boolean	性别
birth	date	出生日期
type	boolean	是否会员
income	Short int	月收入(元)

表 4-2 客户交易数据表

属性名称	数据类型	说 明
customer_id	int	客户标志
time	date	交易日期
goods	string	商品名称
price	real	商品价格
count	short int	商品数量
total_price	real	总价格

    - 相关分析法：如Pearson相关系数、秩相关等

4 数据集成与变换

- （1）数据集成
  - 数据值冲突
    - 在多个数据源中，表示同一实体的属性值可能不同
    - 如单位为元、千元；类型为0/1、Y/N等

4 数据集成与变换

- （2）数据变换
  - 为了使数据符合算法和挖掘目标的需要，如数据的取值范围、粒度等，需要对它们进行变换之后才能使用
  - 处理方法
    - 平滑
    - 聚集
    - 数据概化
    - 规范化

4 数据集成与变换

- （2）数据变换
  - 平滑
    - 去除噪声，将连续的数据离散化等
    - 分箱、聚类、回归等方法
    - 实际上是把一个区域内的值用同一个数值表示，在一定的误差允许条件下减少了属性的取值个数，进而减少挖掘算法的工作量
  - 聚集
    - 对数据进行汇总
      - 如：不使用单个客户的每次的交易明细，只需其消费总额即可
  - 数据概化
    - 将属性中的低层概念概化到高层概念
      - 如客户的出生日期，概化到年龄，再概化到年龄段，再概化到年代（80后，90后）

4 数据集成与变换

- （2）数据变换
  - 规范化
    - 将数据按比例缩放，使之落入一个特定的区域，如[0，1]，称为规范化/标准化
    - 规范化对基于距离的聚类算法和神经网络算法是非常重要的
    - 可以保证输入值在一个相对小的范围内
    - 常用方法
      - 最小-最大规范化（min-max）
      - 零-均值规范化（Z-score）
      - 小数定标规范化

4 数据集成与变换

- （2）数据变换——规范化
  - 最小-最大规范化
    - 区间映射，前提条件是属性的取值范围必须已知
    - $x' = \frac{x - old\_min}{old\_max - old\_min}(new\_max - new\_min) + new\_min$
    - 例：“客户背景数据”表中客户月收入income属性的实际值范围为[430, 4800]，需要把这个属性值规范到[0, 1]，对属性值应用上述公式得：
    - $x' = \frac{3200-430}{8000-430}(1-0) + 0 = 0.365918$
    - 特点
      - 伸缩变换
      - 改变原始数据的分布
      - 满足一些模型求解需要，提高迭代求解的精度和收敛速度
    - 缺点
      - 最大值与最小值易受异常点影响
      - 鲁棒性较差，适合传统精确小数据场景

4 数据集成与变换

- （2）数据变换——规范化
  - 零-均值规范化
    - 根据属性值的平均值和标准差进行规范化
    - 属性值范围可以未知（利用样本的全部信息构建）
    - 求样本的平均值

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

    - 求样本的标准差

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

    - 规范化

$$x' = \frac{x - \bar{X}}{\sigma_x}$$

- 特点
  - 伸缩变换
  - 不改变原始数据的分布
- 缺点
  - 在已有样本足够多的情况下比较稳定，适合现代嘈杂大数据场景

4 数据集成与变换

- （2）数据变换——规范化
  - 小数定标规范化
    - 通过移动属性值的小数点位置进行规范化
    - 需要在属性取值范围已知的条件下使用
    - 小数点移动的位数根据属性的最大绝对值确定

$$x' = \frac{x}{10^{\alpha}}$$，其中α是使$\mathrm{Max}(|x'|) < 1$的最小整数

    - 如客户收入数据，范围为800-5000

$$\mathrm{Max}\left(\left|\frac{x}{10^{\alpha}}\right|\right) = \mathrm{Max}\left(\left|\frac{5000}{10^{\alpha}}\right|\right) < 1$$

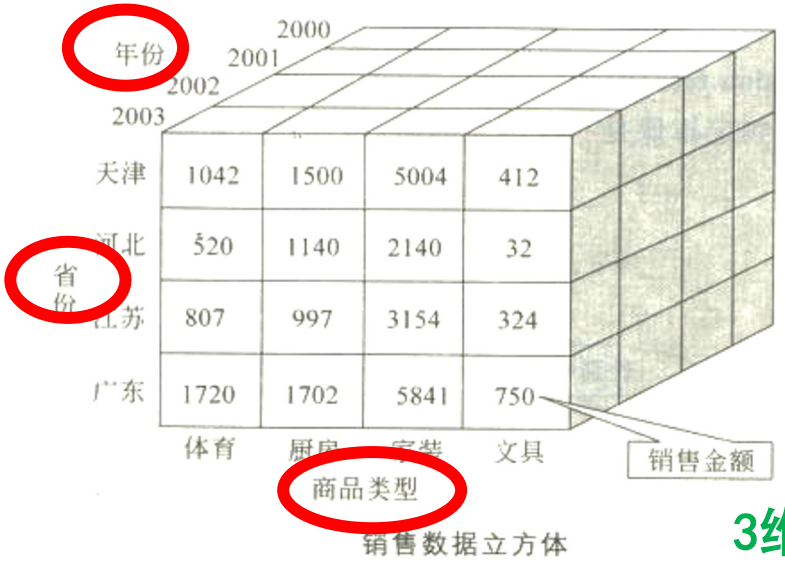
α取4，规范化后最大值5000的值为0.5

4 数据集成与变换

- （2）数据变换——属性构造
  - 小数定标规范化
    - 根据已有的属性构造新的属性添加到挖掘数据集中
      - 例如，根据客户月收入数据，构造“收入水平”属性，取值为{低、中、高}
    - 连续—〉离散
    - 数字—〉符号
    - 多-〉少

5 数据规约

- （1）数据立方体聚集
  - 一个数据立方体由维和事实组成
  - 一个数据立方体可以是n维的
  - 对数据立方体聚集就是去掉一维，变为n-1维立方体，依此类推
    - 例：如果挖掘时感兴趣的是年度的总销售量，不关心每个省份的销售量，可以进行聚集，得到2维数据立方体



销售数据立方体

3维 --> 2维

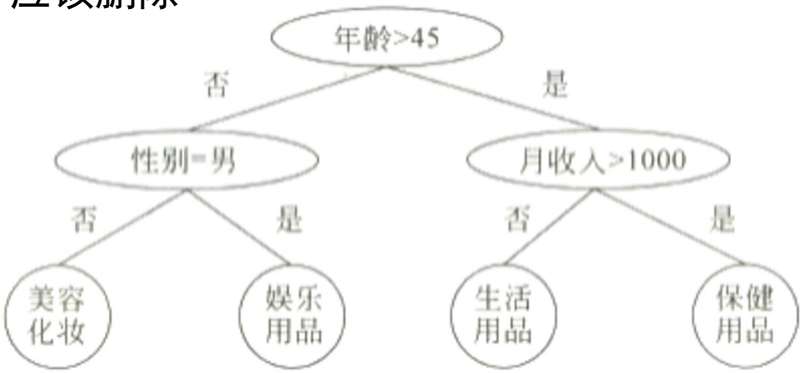
聚集后的销售数据立方体

5 数据规约

- （2）维规约
  - 去掉不相关的，即与挖掘任务无关的属性/维
  - 找到一个最小属性子集，使得这个子集能够具有和原数据集相同或近似的分布
  - 属性子集选择方法
    - 1. 逐步向前选择
      - 原属性集S和S的一个初始为空的子集S’
      - 从S中选择最好的属性（最相关的属性）a加入到S’，直到满足结束条件
    - 2. 逐步向后删除
      - 从S中选择最坏的属性（最不相关的属性）b删除，直到满足结束条件
    - 3. 向前选择和向后删除相结合
      - 每一次选择一个最好的属性，删除一个最坏的属性

5 数据规约

- （2）维规约
  - 属性子集选择方法
    - 4. 判定树归约
      - 根，全集；属性测试，子集；分支，测试结果；叶子节点，判定类
      - 判定树是一种分类算法
      - 在每一个测试点，算法从属性集中选择相关性最强的属性作为判定条件
      - 根据判定结果把数据划分成多个互斥的类
      - 算法结束时，所有内部节点代表的属性被认为是相关属性而选中，不在树中的属性被认为是不相关的，应该删除



5 数据规约

- （2）维规约
  - 属性子集选择方法
    - 5. 基于统计分析的归约
      - 用少量的特征元组去描述高维的原始知识基
      - 主成分分析、逐步回归分析、公共因素模型分析等可以直接用于维回归

5 数据规约

- （3）数据压缩
  - 用数据编码或者变换，得到原始数据的压缩表示
  - 分无损压缩和有损压缩
  - 无损
    - 基于熵的编码方法

      一段文字中的每个字母被一段不同长度的比特(Bit)所代替，n比特的信息量可以表现出2的n次方种选择
  - 有损
    - 主成分分析法：将分散在一组变量上的信息集中到某几个综合指标（主成分）上的探索性统计分析方法，创建一个由具有“最主要特征”的向量组成的集合来替换原数据，把原数据映射到一个较小的空间，实现数据压缩

5 数据规约

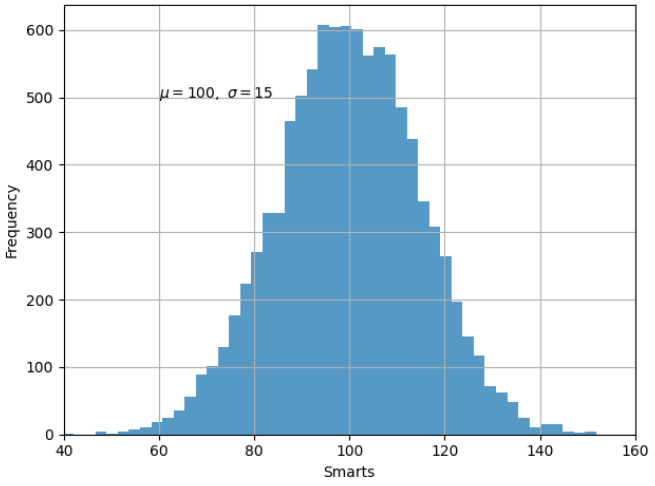
- （4）数值规约
  - 通过某种方法，选择较少的数据来替代原数据，减少数据量
  - 常用方法
    - 直方图
    - 聚类
    - 抽样
    - 线性回归
    - 非线性回归

5 数据规约

- （4）数值规约——直方图
  - 使用分箱方法对数据进行近似
  - 每个箱代表一个区域范围内的值
  - 箱的宽度代表值域范围
  - 箱的高度代表这个范围内的值的个数
  - 一维直方图：每个箱可以代表一个属性的值和频率
  - 多维直方图：每个箱可以代表两个及以上属性的值和频率
  - 单桶：每个箱只表示一个属性值

原始数据 → 箱的高度和宽度

```
mu, sigma = 100, 15
x = mu + sigma * np.random.randn(10000)
plt.hist(x, 50, alpha=0.75)  # 50代表划分为50个区间
```



5 数据规约

- （4）数值规约——聚类
  - 用数据的聚类来代表实际数据

原始数据 → 类

- （4）数值规约——线性回归、非线性回归
  - 线性回归和非线性回归用数据模型而不是记录/实际数据来近似数据
  - 只保存数据模型的参数

原始数据 → 模型参数

5 数据规约

- （4）数值规约——抽样
  - 不是对属性进行选择或者删除
  - 是对记录进行选取
  - 即用较小的数据样本集表示大的数据集
  - 样本与原数据集具有相同的数据分布
  - 抽样方法
    - 不放回简单随机抽样
    - 放回简单随机抽样
    - 聚类抽样
      - 把数据集D的数据放入M个聚类，从每个聚类中抽取样本
    - 分层抽样
      - 把数据集D划分成互不相交的部分，每一部分称为一层，从每层中抽取样本

原始数据 → 抽样数据

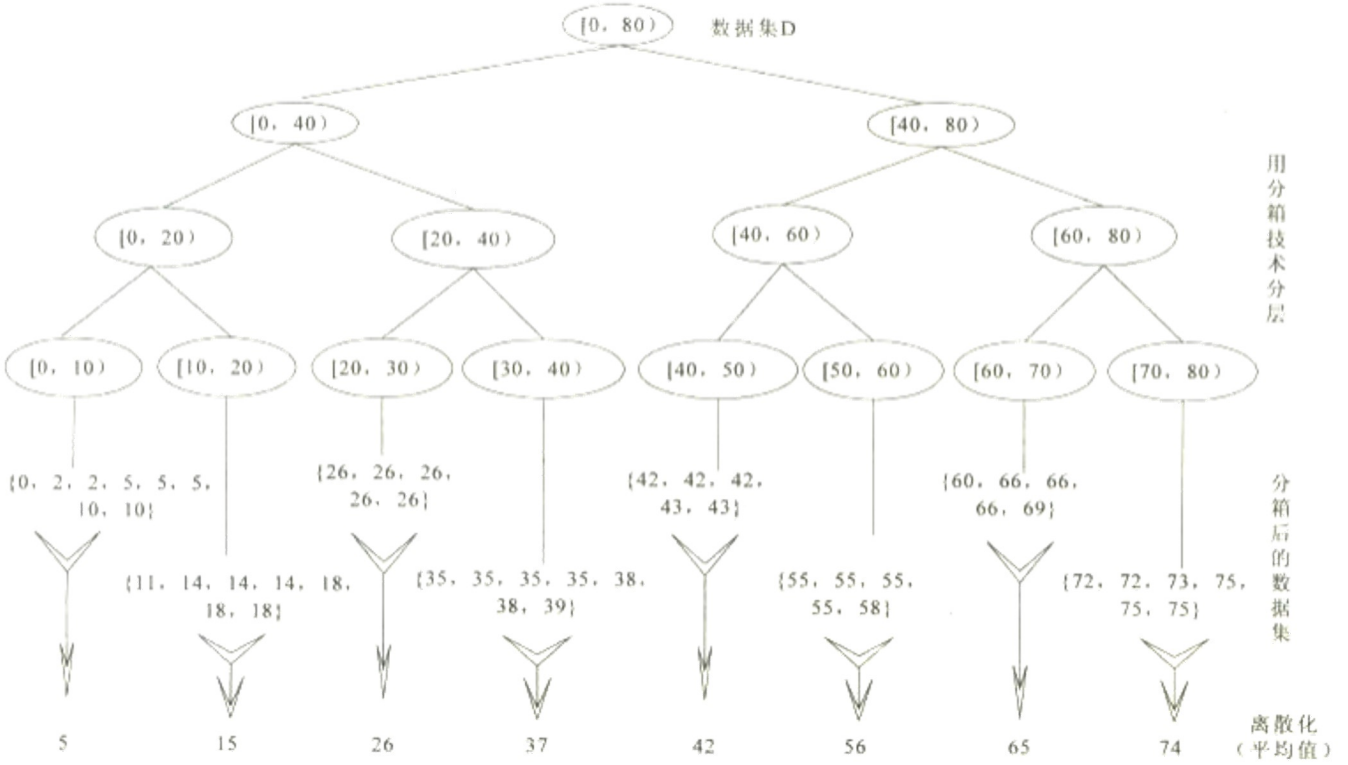
5 数据规约

- （5）离散化与概念分层
  - 将连续数据用有限数量的离散数据替代
  - 采用的方法是把数据划分区间，每个区间中的数据用一个值/符号来代替
  - 如果在数据集上递归地使用某种离散化技术，就形成了数据集的概念分层

5 数据规约

- （5）离散化与概念分层

数据集 D=(0,2,2,5,5,5,10,10,11,14,14,14,18,18,18,26,26,26,26,26,
35,35,35,35,38,38,39,42,42,42,43,43,55,55,55,55,58,60,
66,66,66,69,72,72,73,75,75,75)



5 数据规约

- （5）离散化与概念分层——数值数据
  - 数值数据的概念分层可以通过数据分析自动产生，如分箱、直方图、聚类、基于熵的离散化等
  - 缺点是划分出来的层没考虑边界值是否直观或自然
    - 如[20，30]优于[23.333，36.97]
  - 3-4-5规则
    - 自然划分分段的方法进行概念分层
    - 该规则根据最高有效位的取值范围，递归逐层地将给定的数据区域划分为3、4或5个相对等宽的区间
    - 如果一个区间在最高有效位包含3，6，7或9个不同的值，则将该区间划分成3个区间（对于3，6和9，划分成3个等宽的区间；而对于7，按2-3-2分组，划分成3个区间）
    - 如果它在最高有效位包含2，4或8个不同的值，则将区间划分成4个等宽的区间
    - 如果它在最高有效位包含1，5或10个不同的值，则将区间划分成5个等宽的区间

5 数据规约

- （5）离散化与概念分层——分类数据
  - 分类属性值所包含的数据是数值型、字符型或字符串等
  - 即具有有限个取值的属性（可枚举的）
  - 数据之间没有大小关系
  - 1）由用户或专家在模式级显式地说明数据的包含关系
  - 2）根据属性值的个数自动产生分层
    - 把具有最少不同值的属性放在最高层
    - 属性的不同值数据越多，所处的概念层越低
  - 3）根据数据语义产生分层
    - 在数据模式中加入属性的说明
    - 这些说明把属性组联系在一起
    - 当一个属性被增加进属性组时，依靠数据语义可以把所有相关的属性增加进来

谢 谢！