数据预处理方法

王继民

北京大学信息管理系

2018年11月19日

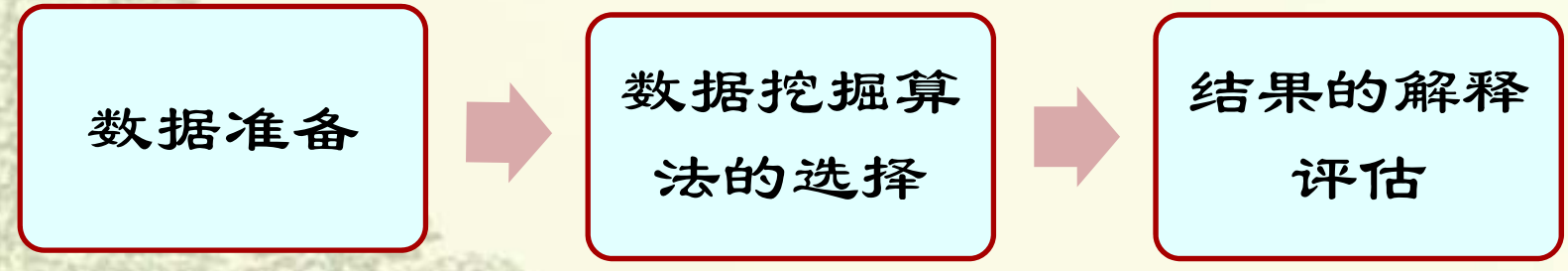
基本内容

- 引言
- 数据预处理的主要方法
  - 数据清理
  - 数据集成
  - 数据变换
  - 数据归约
  - 数据离散化
- 工具软件

引言：数据挖掘及步骤

- 数据挖掘：是指从**数据集**中识别出有效的、新颖的、潜在有用的，以及最终可理解的**模式**的非平凡**过程**

- 数据挖掘的步骤：
  - 数据准备：数据搜集与数据预处理
  - 数据挖掘算法的选择
  - 结果的解释评估



数据挖掘的主要步骤

- **数据准备：（可能要占整体工作量的60%以上）**
  - 数据搜集
  - 数据选择：目标数据
  - 数据清理：消除噪声、不一致、冗余等
  - 数据变换：连续数据离散化、数据规范
  - 数据归约：特征选择或抽取
- **数据挖掘算法的选择**
  - 首先要明确任务，如数据总结、分类、聚类、关联规则发现、序列模式发现等。
  - 考虑用户的知识需求（得到描述性的知识、预测型的知识）。
  - 根据具体的数据集合，选取有效的挖掘算法。

数据挖掘的主要步骤

- 结果的解释评估
  - 对挖掘出来的结果（模式），经用户或机器评价，剔除冗余或无关的模式。
  - 模式不满足用户需求时，返回到某一步，重新挖掘。如：重新选择数据、采用新的变换方法、设定新的数据挖掘参数，或者换一种挖掘算法（如分类方法，不同的方法对不同的数据有不同的效果）。
  - 挖掘的结果是面向用户的，对挖掘结果进行可视化或者转化为用户易于理解的形式表示。
- 评注
  - 影响挖掘结果质量的因素：采用的算法、数据本身的质量与数量
  - 数据挖掘的过程是一个不断反馈的过程
  - 可视化在数据挖掘过程的各个阶段都扮演着重要角色，如用散点图或直方图等统计可视化技术来显示有关数据，以期对数据有一个初步的了解。

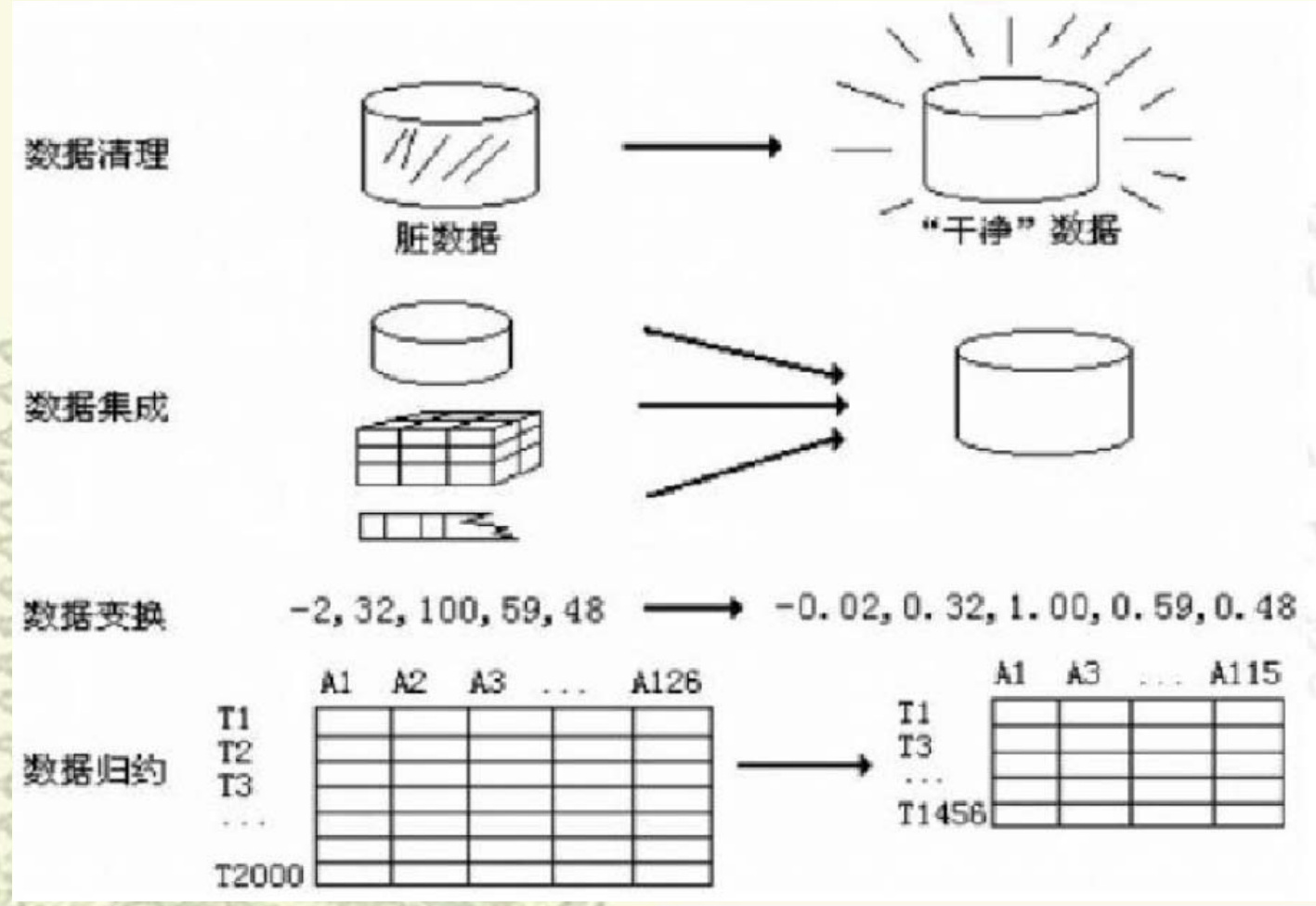
为什么要进行数据预处理？

- 现实世界的数据是“脏的”——数据多了，什么问题都会出现
  - 不完整的：有些感兴趣的属性缺少属性值，或仅包含聚集数据
  - 含噪声的：包含错误或者“孤立点”
  - 不一致的：在编码或者命名上存在差异
- 没有高质量的数据，就没有高质量的挖掘结果
  - 高质量的决策必须依赖高质量的数据
  - 数据仓库需要对高质量的数据进行一致地集成
- 数据预处理的目的：提高数据挖掘的质量(精度),降低实际挖掘所需要的时间. 即：效果+效率(性能)

数据预处理的主要方法

- **数据清理**
  - 填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性来清理数据
- **数据集成**
  - 集成多个数据库、数据立方体或文件
- **数据变换**
  - 将数据转换或统一成适合于挖掘的形式。如数据规范化
- **数据归约**
  - 可以用来得到数据集的归约（压缩）表示，它小得多，但仍保持数据的完整性。对归约后的数据集挖掘将更有效，并产生相同（或几乎相同）的分析结果
- **数据离散化**
  - 数据归约的一部分，通过数据的离散化和概念分层来规约数据

数据预处理的形式


数据清理
脏数据
“干净”数据
数据集成
数据变换
-2, 32, 100, 59, 48
-0.02, 0.32, 1.00, 0.59, 0.48
数据归约
A1 A2 A3 ... A126
T1
T2
T3
...
T2000
A1 A3 ... A115
T1
T3
...
T1456

数据预处理

- 数据清理

  主要通过填写空缺的值，平滑噪声数据，识别、删除孤立点，解决数据的不一致性问题

- 数据集成
- 数据变换
- 数据归约
- 数据离散化与概念分层

数据清理：　空缺值

- 数据并不总是完整的
  - 例如：数据库表中，很多条记录的对应字段没有相应值，比如销售表中的顾客收入
- 引起空缺值的原因
  - 设备异常
  - 与其他已有数据不一致而被删除
  - 因为误解而没有被输入的数据
  - 在输入时，有些数据应为得不到重视而没有被输入
  - 对数据的改变没有进行日志记载
- 空缺值要经过推断而补上

数据清理：如何处理空缺值

- 忽略元组：当类标号缺少时通常这么做（假定挖掘任务设计分类或描述），当每个属性缺少值的百分比变化很大时，它的效果非常差。
- 人工填写空缺值：工作量大，可行性低
- 使用一个全局变量填充空缺值：比如使用unknown或$-\infty$
- 使用属性的平均值、中位数、众数等填充空缺值
- 使用与给定元组属同一类的所有样本的平均值
- 使用*最可能的值*填充空缺值：使用像Bayesian公式或判定树这样的基于推断的方法

数据清理：　噪声数据

- 噪声：一个测量变量中的随机错误或偏差
- 引起不正确属性值的原因
  - 数据收集工具的问题
  - 数据输入错误
  - 数据传输错误
  - 技术限制
  - 命名规则的不一致
- 其它需要数据清理的数据问题
  - 重复记录
  - 不完整的数据
  - 不一致的数据

数据清理：　如何处理噪声数据

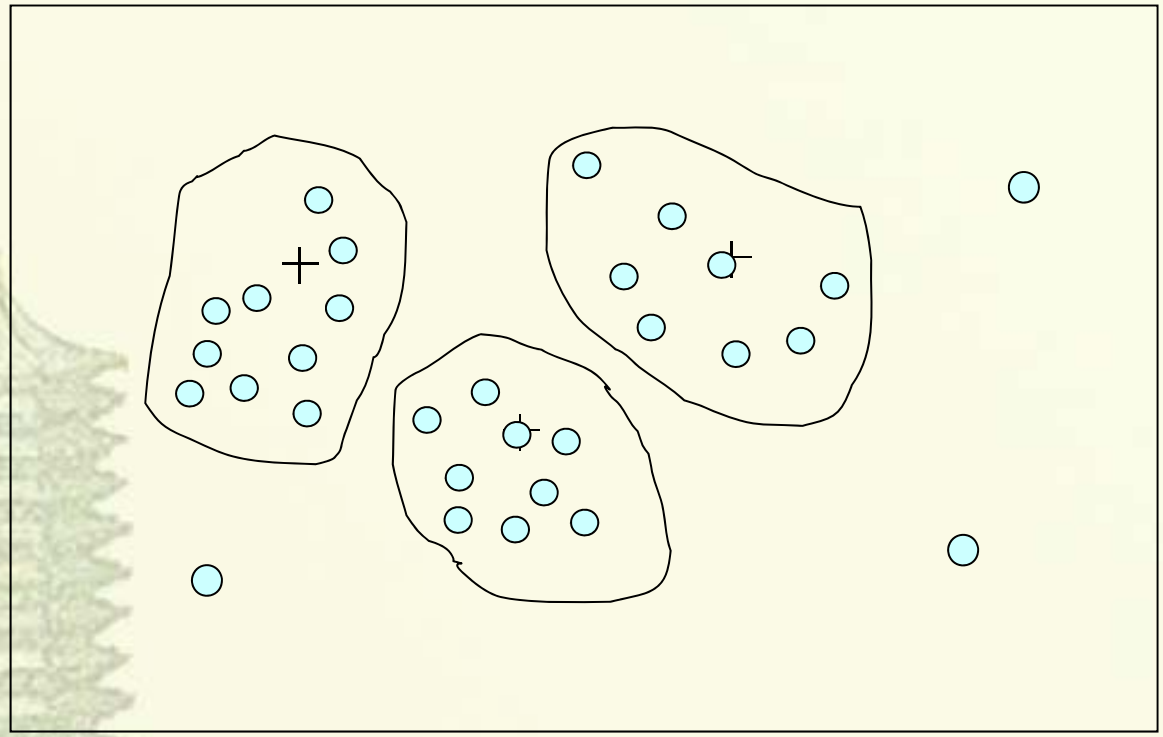
- 分箱(binning):
  - 首先排序数据，并将他们分到等深的箱中
  - 然后可以按箱的平均值平滑、按箱中值平滑、按箱的边界平滑等等
- 聚类：
  - 监测并且去除孤立点
- 计算机和人工检查结合
  - 计算机检测可疑数据，然后对它们进行人工判断
- 回归
  - 通过让数据适应回归函数来平滑数据

数据清理：数据平滑的分箱方法

- price的排序后数据（单位：美元）：4，8，15，21，21，24，25，28，34

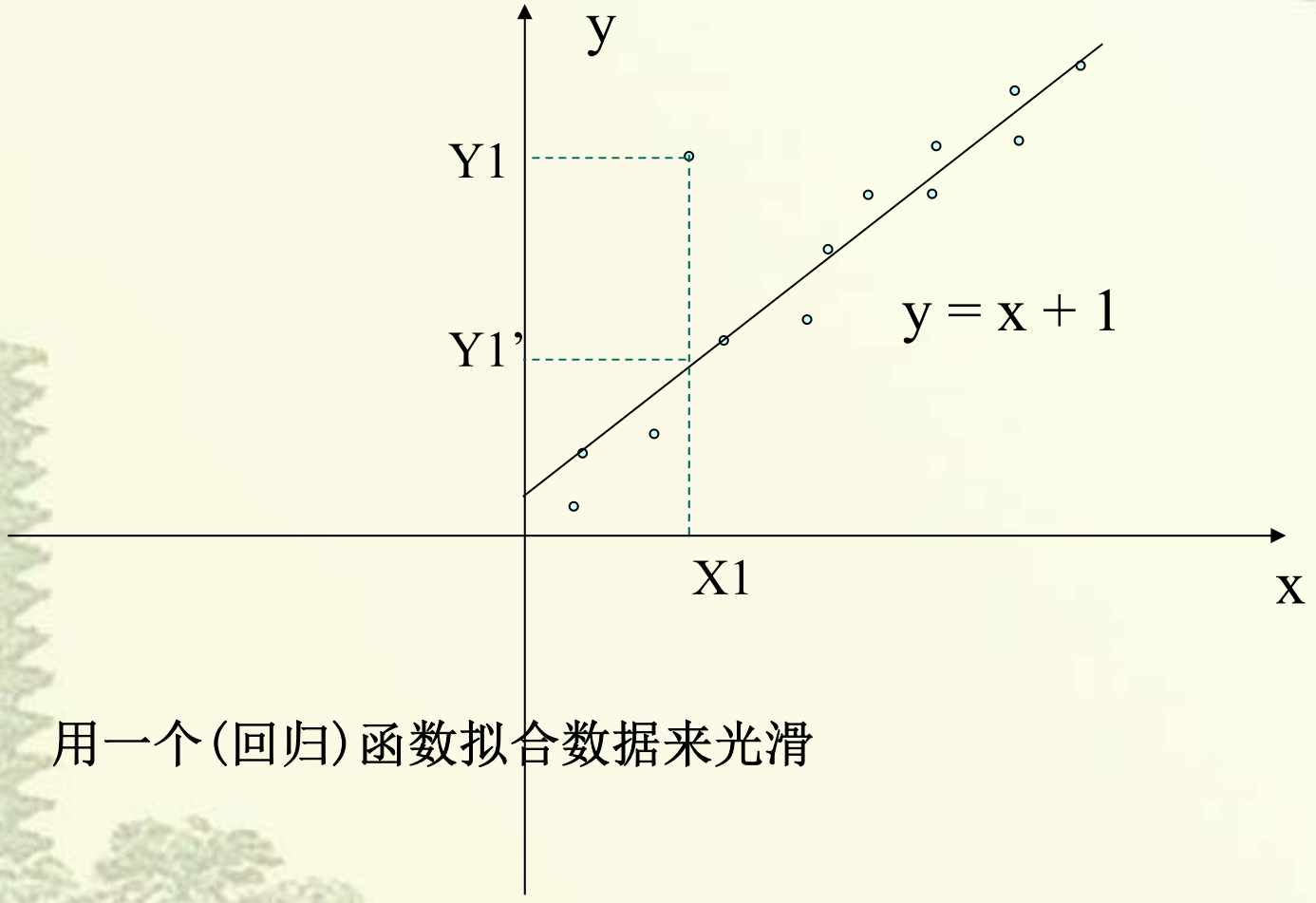
- 划分为（等深的）箱：
  - 箱1：4，8，15
  - 箱2：21，21，24
  - 箱3：25，28，34
- 用箱平均值平滑(或者：中位数)
  - 箱1：9，9，9
  - 箱2：22，22，22
  - 箱3：29，29，29
- 用箱边界平滑：
  - 箱1：4，4，15
  - 箱2：21，21，24
  - 箱3：25，25，34

数据清理：　聚类分析去除噪声数据



- 通过聚类分析查找孤立点，消除噪声

数据清理：　回归分析去除噪声数据



回归：　用一个(回归)函数拟合数据来光滑

数据预处理

- 数据清理
- 数据集成

  将多个数据源中的数据整合到一个一致的存储中
- 数据变换
- 数据归约
- 数据离散化与概念分层

数据集成

- **数据集成：**
  - 将多个数据源中的数据整合到一个一致的存储中
- **模式集成：**
  - 整合不同数据源中的元数据
  - 实体识别问题：匹配来自不同数据源的现实世界的实体，比如：A.cust-id=B.customer_no
- **检测并解决数据值的冲突**
  - 对现实世界中的同一实体，来自不同数据源的属性值可能是不同的
  - 可能的原因：不同的数据表示，不同的度量等等

处理数据集成中的冗余数据

- 集成多个数据库时，经常会出现冗余数据
  - 同一属性在不同的数据库中会有不同的字段名
  - 一个属性可以由另外一个表导出，如“年薪”,“月薪”
- 有些冗余可以被相关分析检测到

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B}$$

- 仔细将多个数据源中的数据集成起来，能够减少或避免结果数据中的冗余与不一致性，从而可以提高挖掘的速度和质量。

数据预处理

- 数据清理
- 数据集成
- 数据变换

  将数据转换或统一成适合于挖掘的形式。

  如： 数据规范化
- 数据归约
- 数据离散化与概念分层

数据变换

- 规范化：
  - 最小－最大规范化
  - z-score规范化
- 属性构造
  - 通过现有属性构造新的属性，并添加到属性集中；以增加对高维数据的结构的理解和精确度

数据变换——规范化

- 最小－最大规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score规范化

$$v' = \frac{v - mean_A}{standard\_dev_A}$$

例如：使用如下两种方法规范化数组：20，30，40，60，100。（1）min-max规范化，其中，min=0，max=1。（2）z-score规范化。

数据预处理

- 数据清理
- 数据集成
- 数据变换
- 数据归约

  可以用来得到数据集的归约（压缩）表示，它小得多，但仍保持数据的完整性。对归约后的数据集挖掘将更有效，并产生相同（或几乎相同）的分析结果。

- 数据离散化与概念分层

数据归约策略

- 数据仓库中往往存有海量数据，在其上进行复杂的数据分析与挖掘需要很长的时间
- 数据归约
  - 数据归约可以用来得到数据集的归约表示，它小得多，但可以产生相同的（或几乎相同的）分析结果
- 数据归约策略
  - 数据立方体聚集
  - 维归约
  - 数据压缩
  - 数值归约
  - 离散化和概念分层产生
- 用于数据归约的时间不应当超过或“抵消”在归约后的数据上挖掘节省的时间。

维归约

- 通过删除不相干的属性或维减少数据量
- 属性子集选择
  - 找出最小属性集，使得数据类的概率分布尽可能的接近使用所有属性的原分布
  - 减少出现在发现模式上的属性的数目，使得模式更易于理解
- 启发式的（探索性的）方法
  - 逐步向前选择
  - 逐步向后删除
  - 向前选择和向后删除相结合
  - 判定归纳树

数据压缩

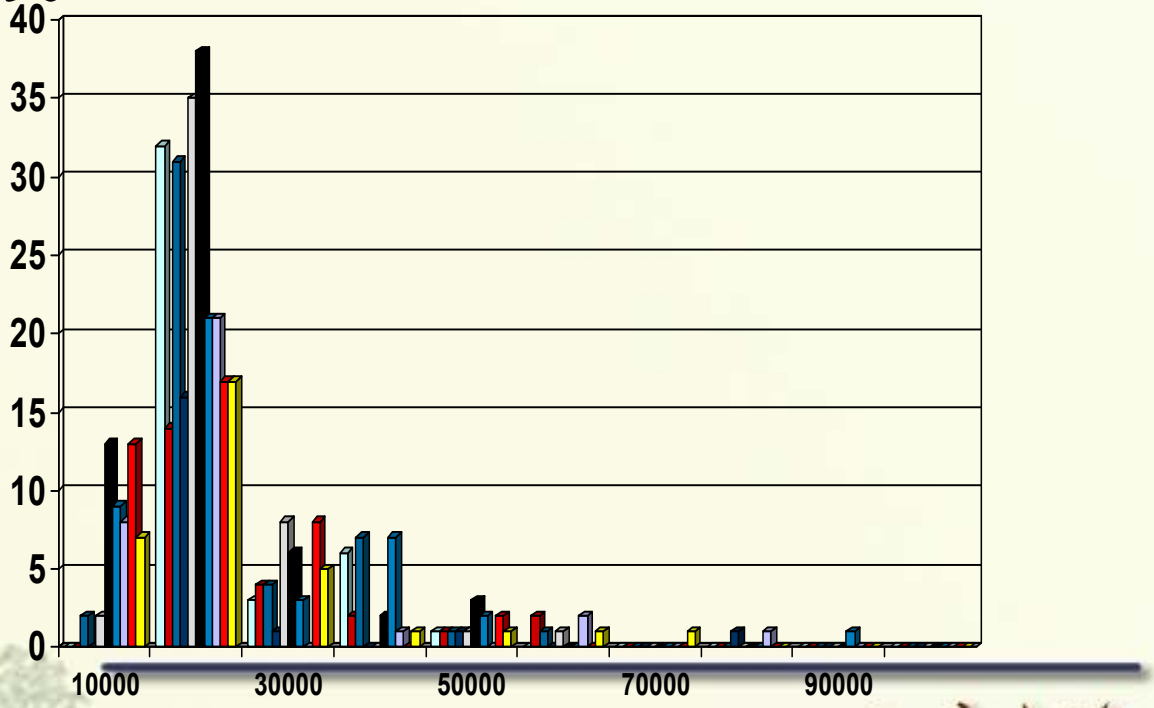
- 有损压缩 VS. 无损压缩
- 字符串压缩
  - 有广泛的理论基础和精妙的算法
  - 通常是无损压缩
  - 在解压缩前对字符串的操作非常有限
- 音频/视频压缩
  - 通常是有损压缩，压缩精度可以递进选择
  - 有时可以在不解压整体数据的情况下，重构某个片断
- 两种有损数据压缩的方法：小波变换、主成分分析

数值归约

- 通过选择替代的、较小的数据表示形式来减少数据量
- 有参方法：使用一个参数模型估计数据，最后只要存储参数即可。
  - 线性回归方法：Y=α+βX
  - 多元回归：线性回归的扩充
  - 对数线性模型：近似离散的多维数据概率分布
- 无参方法：
  - 直方图
  - 聚类
  - 选样

直方图

- 一种流行的数据归约技术
- 将某属性的数据划分为不相交的子集，或桶，桶中放置该值的出现频率
- 桶和属性值的划分规则
  - 等宽
  - 等深
  - V-最优
  - MaxDiff

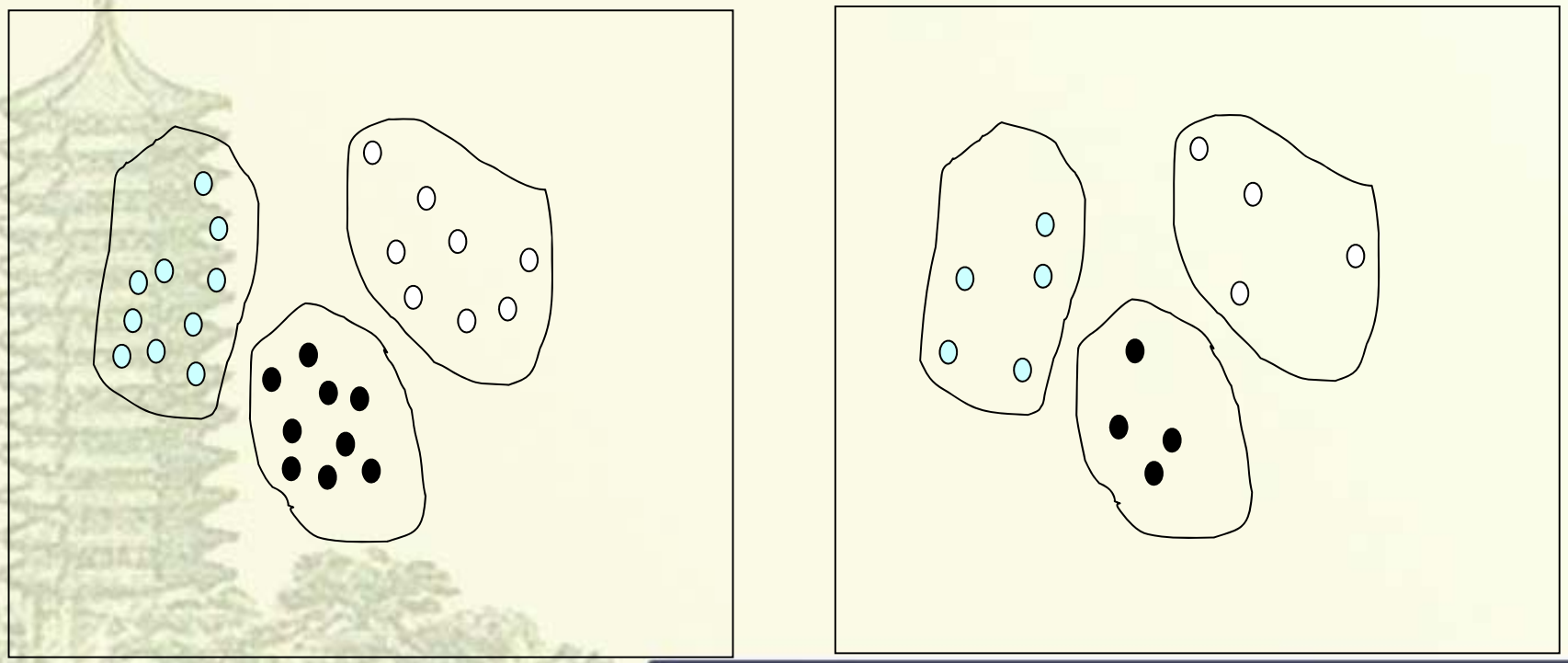


选样——SRS

选样——聚类/分层选样

原始数据

聚类/分层选样



数据预处理

- 数据清理
- 数据集成
- 数据变换
- 数据归约
- 数据离散化与概念分层
  - 数据归约的一部分，通过数据的离散化和分类数据的概念分层来规约数据。

离散化和概念分层

- 离散化
  - 通过将属性域划分为区间，减少给定连续属性值的个数。区间的标号可以代替实际的数据值。
- 概念分层
  - 通过使用高层的概念（比如：青年、中年、老年）来替代底层的属性值（比如：实际的年龄数据值）来规约数据

离散化

- 三种类型的属性值：
  - 名称型——e.g. 无序集合中的值(如颜色, 民族..)
  - 序数——e.g. 有序集合中的值 (如职称)
  - 连续值——e.g. 实数
- 离散化
  - 将连续属性的范围划分为区间
  - 有效的规约数据
    - 基于判定树的分类挖掘
  - 离散化的数值用于进一步分析

数据数值的离散化

- 分箱（binning）
  - 分箱技术递归的用于结果划分，可以产生概念分层。
- 直方图分析（histogram）
  - 直方图分析方法递归的应用于每一部分，可以自动产生多级概念分层。
- 聚类分析
  - 将数据划分成簇，每个簇形成同一个概念层上的一个节点，每个簇可再分成多个子簇，形成子节点。
- 基于信息熵的离散化
- 通过自然划分分段

通过自然划分分段

- 将数值区域划分为相对一致的、易于阅读的、看上去更直观或自然的区间。
  - 聚类分析产生概念分层可能会将一个工资区间划分为：[51263.98, 60872.34]
  - 通常数据分析人员希望看到划分的形式为[50000，60000]
- 3-4-5规则：用于将数值数据划分为相对一致和“自然的”的区间。该规则根据最重要数据的数值区域，递归地、逐层地将给定的数值区域划分为，3，4或5个等宽的区间。

自然划分的3-4-5规则

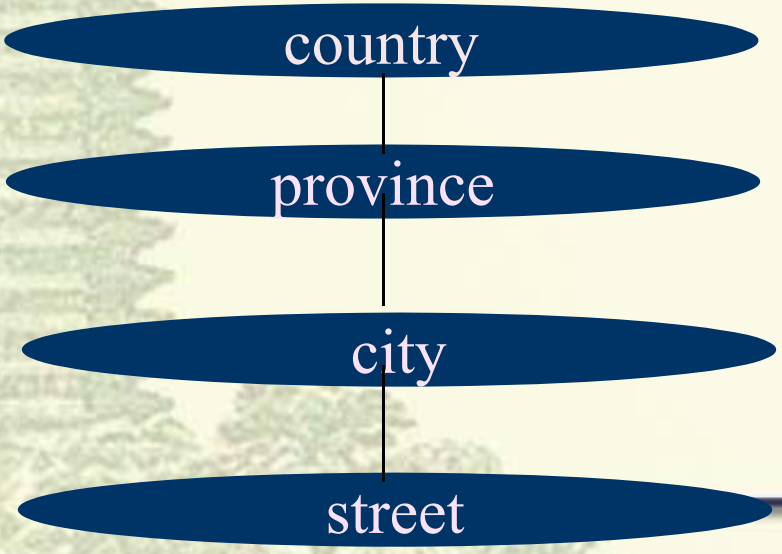
- **规则的划分步骤：**
  - 如果一个区间最高有效位上包含3，6，7或9个不同的值，就将该区间划分为3个等宽子区间；(7→2,3,2)
  - 如果一个区间最高有效位上包含2，4，或8个不同的值，就将该区间划分为4个等宽子区间；
  - 如果一个区间最高有效位上包含1，5，或10个不同的值，就将该区间划分为5个等宽子区间；
  - 将该规则递归的应用于每个子区间，产生给定数值属性的概念分层；
  - 对于数据集中出现的最大值和最小值的极端分布，为了避免上述方法出现的结果扭曲，可以在顶层分段时，选用一个大部分的概率空间。e.g. 5%-95%

分类数据的概念分层生成

- 分类数据是指无序的离散数据，它有有限个值（可能很多个）。
- 分类数据的概念分层生成方法：
  - 由用户或专家在模式级显式的说明属性的部分序。
  - 通过显示数据分组说明分层结构的一部分。
  - 说明属性集，但不说明它们的偏序，然后系统根据算法自动产生属性的序，构造有意义的概念分层。
  - 对只说明部分属性集的情况，则可根据数据库模式中的数据语义定义对属性的捆绑信息，来恢复相关的属性。

属性集的规格

- 根据在给定属性集中，每个属性所包含的不同值的个数，可以自动的生成概念分层；不同值个数最多的属性将被放在概念分层的最底层。(有例外,如日期)



5个不同值

65 个不同值

3567 个不同值

674,339 个不同值

常用的工具软件

- Python
  - sklearn的preprocessing模块
  - Pandas进行数据预处理-数据清洗
- R语言
- SPSS （ +Clementine ）
- SAS Enterprise Miner
- Matlab
- .....

参考文献

- Jiawei Han and Micheline Kamber. 数据挖掘概念与技术. 机械工业出版社. 2007.（原书第2版）
- Pang-Ning Tan, Michael Steinbach. 数据挖掘导论. 人民邮电出版社. (图灵计算机科学丛书). 2006.

基本内容

- 引言
- 数据预处理的主要方法
  - 数据清理
  - 数据集成
  - 数据变换
  - 数据归约
  - 数据离散化
- 工具软件

Thank you!