Course 2
数据预处理
Data Preprocessing

Data Mining
数据挖掘

大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

为什么要预处理数据?

- 真实世界的数据是不干净(Dirty)
  - 不完整(Incomplete):缺少属性值或某些感兴趣的属性，或仅包含聚集数据
    - 例, 职业=" "，公司人数=20 (无公司人员数据)
  - 含噪声(Noise):包含错误或存在偏离期望的离群值
    - 例, 薪资="-10"
  - 不一致(Inconsistent):在数据本身或命名上不一致
    - 例, 年龄="42" 生日="03/07/1997"
    - 例, 过去分类 "1,2,3", 现在分类 "A, B, C"
    - 重复性的纪录所造成，如: Customer_ID, Customer_Num

厦门大学智能科学系

为什么数据是不干净?

- 不完整来自
  - 输入的时候认为它是不重要的项目
  - 相关数据由于误解或设备故障没有被记录下来
  - 历史数据的记录或修改时并未仔细检查
- 噪声(不正确值)来自
  - 数据收集设备的故障
  - 数据输入时人为或计算机错误
  - 数据传输错误
- 不一致来自
  - 不同数据来源
- 重复纪录需要数据清除

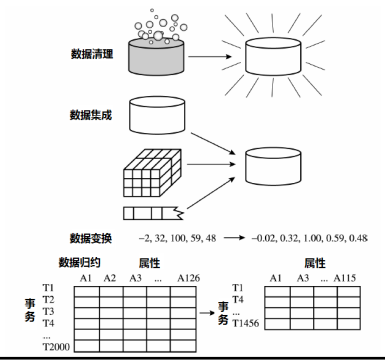
厦门大学智能科学系

数据预处理主要工作

- 数据清理 (**Data Cleaning**)
  - 填写缺失的值、光滑噪声数据、识别或删除离群点、并解决不一致
- 数据集成 (**Data Integration**)
  - 集成多个数据库、数据立方体或文件
- 数据变换 (**Data Transformation**)
  - 规范化和聚集
- 数据归约 (**Data Reduction**)
  - 降低数据集的大小并能获得相同的分析结果
- 数据离散化 (**Data Discretization**)
  - 是一种数据归约形式,对数值数据自动产生概念分层非常有用

厦门大学智能科学系

数据预处理型式



厦门大学智能科学系

为何数据预处理是重要的?

- **没有高质量的数据，就没有高质量的挖掘结果！**
  - 高质量的决策必须基于高质量的数据。
    - 重复或数据的缺失可能会使用户做出不正确的、甚至是误导性的决策。
  - 数据仓库需要将高质量的数据做一致性的集成。

厦门大学智能科学系

大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

厦门大学智能科学系

描述性数据汇总

- 动机：为了更好的理解数据
  - 获得数据的总体印像
  - 识别数据的典型性质
  - 凸显噪声或离群点
- 度量数据的中心趋势
- 度量数据的离散程度

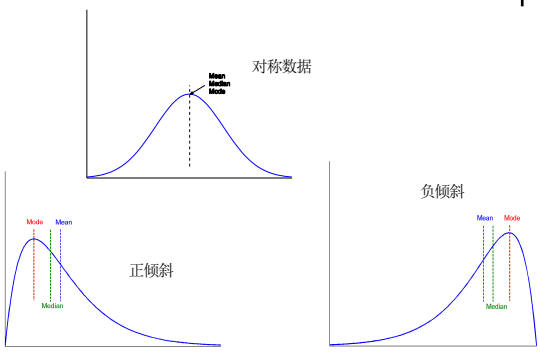
度量中心趋势 (Central Tendency)

- 平均值(Mean):
  - 加权平均:
  - 截断均值: 去掉极值后的均值
- 中位数 (Median):整体度量
  - 如果数值个数为奇数则中值为所有数字的中值, 反之则为最中间的两个数字平均
- 众数 (Mode)
  - 出现频率最高的值
  - 单峰的、双峰的和三峰的
  - 经验公式

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

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

$$mean - mode = 3 \times (mean - median)$$



度量数据的离散程度 (Dispersion)

- 方差与标准差
  - 方差:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}[\sum_{i=1}^{n} x_i^2 - \frac{1}{N}(\sum_{i=1}^{n} x_i)^2]$$

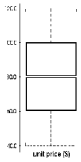
  - 标准差σ是方差σ²的平方根
- 四分位数,离群点与盒状图
  - 四分位数: Q_1(第25个百分位数), Q_3(第75个百分位数)
  - 中间四分位数极差: $IQR = Q_3 - Q_1$
  - 离群点: 高超过Q_3 /低超过Q_1 至少1.5 x IQR

数据集合的第k个百分位数是具有如下性质的值x_i：百分之k的数据项位于或低于x_i。中位数是第50个百分位数。

盒状图分析

- 五数概括:

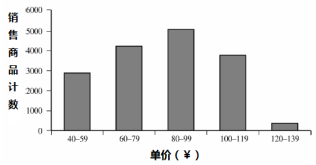
Minimum, Q1, M, Q3, Maximum



- 盒状图
  - 数据用盒状表示
  - 盒子的两边为第一与第三个四分位数，盒子的长度为中间四分位数极差
  - 盒子中的一条线代表中位数
  - 盒子外的两条线 (**whiskers,** 胡须) 延伸到最小与最大观测值

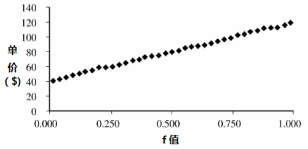
基本描述数据汇总的图形显示

- 直方图分析
  - 频率直方图
    - 是一种概括给定属性分布的图形方法
    - 数值以长方形表示，长方形的高度代表包含在这一区间的数据值的个数或相对频率



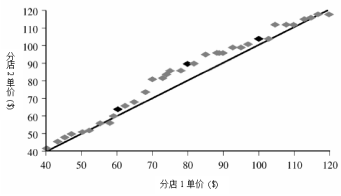
分位数图（quantile plot）

- 显示给定属性的所有数据 (允许用户评估总体情况和不寻常的出现)
  - 绘出了分位数信息
  - 将数据x_i递增排序，f_i表示大约有100%f_i的数据是低于x_i或与x_i相等



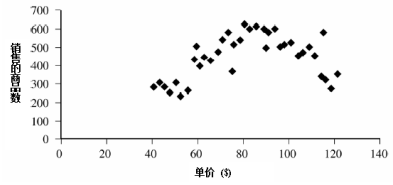
分位数-分位数图(q-q)

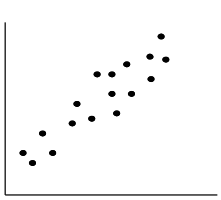
- 对照着另一个对应的分位数，绘制一个单变量分布的分位数。
- 观察数据从一个分布到另一个分布是否有移位



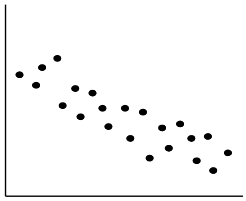
散布图（scatter plot）

- 确定两个数值属性之间看上去是否有联系、模式或趋势
- 每一对的属性值视为代数坐标对，并将这些坐标点画在平面上

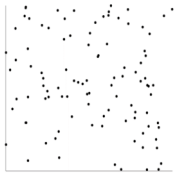




正相关



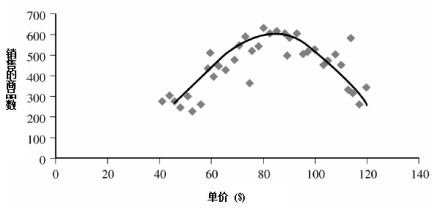
负相关



不相关

loess曲线

- 在分布图中加入一条光滑曲线，以便更好的理解依赖模式



厦门大学智能科学系

大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

厦门大学智能科学系

数据清理 (Data Cleaning)

- 重要性
  - “数据清理为数据仓库首要问题”
- 数据清理工作
  - 填充缺失值
  - 光滑噪声并识别离群点
  - 纠正数据中的不一致
  - 解决数据集成所造成重复

厦门大学智能科学系

缺失值 (Missing Data)

- 数据不是皆为available
  - 许多元组的数据有时会漏记几个字段的数值，而这些缺失值对某些特定应用非常重要。如：Customer income in sales data
- 造成缺失值的原因可能有：
  - 设备异常
  - 与其它已存在数据不一致而遭删除
  - 因为误解造成数据没有被输入
  - 在输入时，因为得不到应用的重视而没有被输入

厦门大学智能科学系

如何处理缺失值?

- 忽略元组：这个方法不是很有效，除非元组有多个属性缺少值。
- 人工填写缺失值：非常费时并不实际
- 自动填入：
  - 使用全局常量 (**global constant**) 填充缺失值：例.，“未知”，新类别!
  - 使用属性的均值来填充缺失值
  - 使用与给定元组属同一类的所有样本的属性均值：较聪明方法
  - 使用最可能的值来填充缺失值:可以用回归、使用贝叶斯的基于推理的工具或决策树归纳确定

厦门大学智能科学系

噪声数据 (Noisy Data)

- **噪声**：被测量的变量的随机误差或方差。
- 造成Noise的原因可能有：
  - 数据收集工具的故障
  - 数据输入问题
  - 数据传输问题

厦门大学智能科学系

如何处理噪声数据?

- 分箱
  - 首先将数据排序，然后分成频率相同的若干个箱子中
  - 然后用箱均值光滑，用箱中位数光滑，用箱边界光滑等
- 回归
  - 可以用一个函数拟合数据来光滑数据
- 聚类
  - 检测离群点
- 集成计算机与人检验
  - 利用人检测有疑问的值 (例.，处理可能的离群值)

简单离散化方法: 分箱

- 分箱方法通过考察数据的“近邻”来光滑有序数据的值
- 等宽(距离) 分裂
  - 切割成**N**个等宽范围: 均匀网格线
  - 如果***A***与***B***为属性的最低与最高值, 范围宽度为: ***W*= (*B*–*A*)/ *N*.**
  - 最直接但是离群值会影响结果
  - 对倾斜数据不能处理很好
- 等深(频率) 分裂
  - 切割成**N**个范围, 每个范围有大约相同样本数
  - 具有比较好的数据量度性
  - 对类别属性会较难处理

利用分箱进行数据平滑

*price*的排序后数据（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：

箱1：9, 9, 9

箱2：22, 22, 22

箱3：29, 29, 29

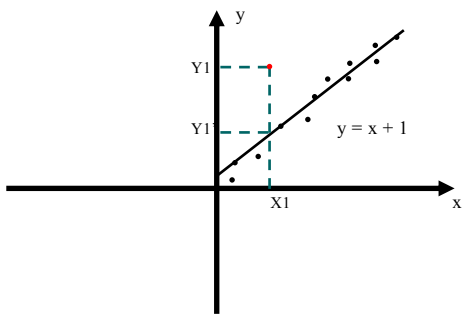
用箱边界平滑：

箱1：4, 4, 15

箱2：21, 21, 24

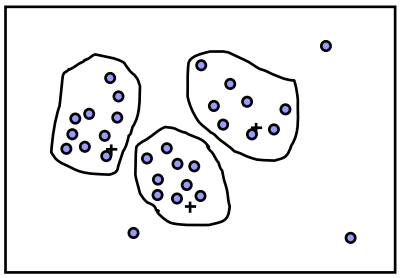
箱3：25, 25, 34

回归



厦门大学智能科学系

聚类



厦门大学智能科学系

大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

厦门大学智能科学系

数据集成 (Data Integration)

- 数据集成：
  - 合并多个数据源中的数据，存放在一个一致的数据存储（如数据仓库）中。
- 在具有多重来源的数据集成上多花心思，将有助于减少/避免数据冗余、不一致以及提高挖掘的效率与质量。
- 数据集成有几个重要的议题：
  - 模式集成与对象匹配（Schema integration）
  - 冗余（Redundancy）
  - 数据值冲突的检测与处理（Detection and resolution of data value conflicts）

- 模式集成与对象匹配：
  - 实体识别问题
  - 例., **A.cust-id≡ B.cust-#**
- 冗余数据：
  - 个体识别问题
  - 对不同数据来源的真实个体进行相关分析, 例., **Bill Clinton = William Clinton**
- 数据值冲突的检测与处理
  - 对于现实世界的同一实体，来自不同数据源的属性值可能不同
  - 可能原因: 不同表示, 比例或编码, 例., 公制与英制

在数据集成中处理重复问题

- 当集成多个数据库会导致数据重复
  - 实体识别：在不同数据库, 相同属性或个体会有不同名称
  - 冗余数据:某个属性可以从另一个数据表的属性推论得之, 例., 年盈余
- 有些冗余可以被相关分析 (**correlation analysis**) 检测到。

相关分析(数值数据)

- 相关系数(**Pearson' s product coefficient**)

$$r_{A,B}=\frac{\sum_{i=1}^{N}(a_i-\bar{A})(b_i-\bar{B})}{N\sigma_A\sigma_B}=\frac{\sum_{i=1}^{N}(a_ib_i)-N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

其中， N 是元组个数， a_i 和 b_i 分别是元组 i 中 A 和 B 的值，$\bar{A}$ 和 $\bar{B}$ 分别是 A 和 B 的均值，σ_A 和 σ_B 分别是 A 和 B 的标准差，而 $\sum(a_ib_i)$ 是 AB 叉积的和（即对于每个元组，A 的值乘以该元组 B 的值）。注意，$-1\le r_{A,B}\le+1$ 的。如果 $r_{A,B}$ 大于 0，则 A 和 B 是正相关的，意味 A 的值随 B 的值增加而增加。该值越大，每个属性蕴涵另一个的可能性越大。因此，一个较高的 $r_{A,B}$ 值表明 A（或 B）可以作为冗余而被去掉。如果结果值等于 0，则 A 和 B 是独立的，不存在相关。如果结果值小于 0，则 A 和 B 是负相关的，一个值随另一个的减少而增加。这意味每一个属性都阻止另一个出现。

厦门大学智能科学系

相关分析(类别数据)

- χ²(卡方) 检验

卡方检验假设A和B是独立的

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

- 其中

$$e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{N}$$

- χ²值愈大，变量愈有关联
- 对χ²值最大的单元是其实际计数与期望计数很不相同的单元。
- 相关并不意味因果关系
  - 医院数与汽车失窃数是相关的
  - 这两个属性都与另一个属性(人口数)有因果的连结

厦门大学智能科学系

χ^2卡方计算: 范例

	男	女	
小说	250(90)	200	
非小说	50(210)	1000(840)	1050
合计	300	1200	1500

$$e_{11} = \frac{count(男)\times count(小说)}{N} = \frac{300\times 450}{1500} = 90$$

- χ²计算 (括号的数字代表期望频率，由两个类别的数据分布计算得之)

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

- 在这一群组中，它显示**gender**与**preferred_reading**是相关

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{N}$$

厦门大学智能科学系

数据变换 (Data Transformation)

- 光滑(**Smoothing**)：移除噪声
- 聚集(**Aggregation**)：汇总或聚集，建立数据立方体
- 数据泛化(**Generalization**)：使用概念分层，用高层概念替换低层或“原始”数据。
- 规范化(**Normalization**)：将属性数据按比例缩放，使之落入一个小的特定区间。
  - **min-max** 规范化
  - **z-score**规范化
  - 小数定标规范化
- 属性构造(**Attribute Construction**)
  - 构造新的属性并添加到属性集中，以帮助挖掘过程。

厦门大学智能科学系

数据转换: 正规化

- **min-max** 规范化: 转换至 [new_min_A, new_max_A]

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

  - 例. 将收入范围 **$12,000 to $98,000** 正规化至 **[0.0, 1.0]**. 则 **$73,000** 会对应至

$$\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$$

- **Z-score**规范化(μ: 均值, σ: 标准差):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - 例. 当 μ = **54,000**, σ = **16,000**. 则

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

- 小数定标规范化

$$v' = \frac{v}{10^j}$$

*j*为当 **Max(|v'|) < 1**时最小的整数

厦门大学智能科学系

大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

厦门大学智能科学系

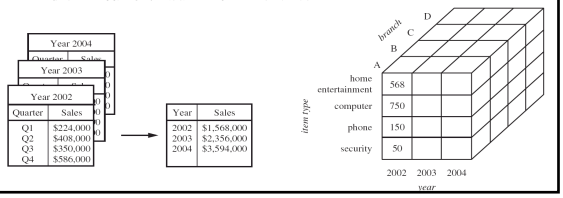
数据归约(Data Reduction Strategies)

- 为何要数据归约?
  - 数据仓库中选择用于分析的数据集极为庞大
  - 海量复杂的数据分析与挖掘会花费很长的时间，不现实或不可行
- 数据归约
  - 用来得到数据集的归约表示，它小的多，但仍接近保持原数据的完整性。
- 数据归约策略
  - 数据立方体聚集
  - 属性子集选择，例. 移除不重要属性
  - 维度归约
  - 数值归约—用替代的、较小的数据表示替换或估计数据。
  - 离散化和概念分层产生

厦门大学智能科学系

数据立方体聚集 (Data Cube Aggregation)

- 数据立方体最底层 (基本方体)
  - 对应于感兴趣的个体实体
- 数据立方体多层聚集
  - 进一步降低要处理数据的大小
- 参照适当层次
  - 使用与给定任务相关的最小可用方体

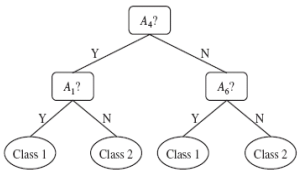


厦门大学智能科学系

属性子集选择

- 用于分析的数据集可能包含数以百计的属性，其中大部分属性与挖掘任务不相关或冗余。
- 属性子集选择：
  - 希望找到最小的属性集，使得数据类的概率分布尽可能的接近使用所有属性得到的原分布。
  - 减少了出现在发现模式的属性数目，使得模式更易于理解
- 启发式(n个属性可能有2^n数目的选择)：
  - 逐步向前选择
  - 逐步向后删除
  - 向前选择与向后删除的结合
  - 决策树归纳

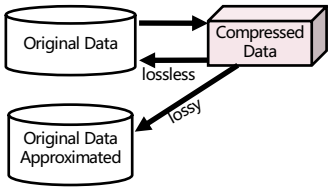
厦门大学智能科学系

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

厦门大学智能科学系

维度归约

- 维度归约使用数据编码或变换，以便得到原数据的归约或“压缩“表示。
  - 无损压缩：原数据可以由压缩数据重新构造而不丢失任何信息
  - 有损压缩：只能重新构造原数据的近似表示



厦门大学智能科学系

维度归约:小波转换

- 离散小波转换(DWT): 线性信号处理, 多分辨率分析
- 压缩式近似: 仅存放少量最强的小波系数
- 类似于离散傅里叶变换(DFT), 但DWT是一种更好的有损压缩，且对于等价的近似，DWT需要的空间更小
- 方法:
  - 长度L, 必须是2的整数幂次方(可以通过在数据向量后加入0)
  - 每个变换涉及到两个函数:光滑化, 加权差分
  - 作用于所有数据点对，进而形成两组长度为L/2的数据
  - 两个函数递归的作用于前面循环得到的数据集，直到结果数据的长度达到预设长度

厦门大学智能科学系

维度归约: 主成分分析法 (PCA)

- 搜索k个最能代表数据的n维正交向量，其中$k \leq n$
- 步骤
  - 对输入数据规范化，使得每个属性都落入相同的区间。
  - 计算k个标准正交向量, 也就是主成分
  - 输入数据为主成分的线性组合
  - 主成分按照重要性降序排序
  - 因为主成分按照重要性降序排序，所以通过去掉较弱的成分就可以达到数据归约的目的。使用最强的主成分，它应该足以近似于原始数据
- 仅适用于数值数据

数值归约

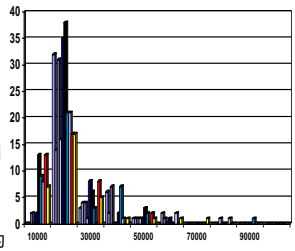
- 通过选择替代的、“较小的”数据表示形式来减少数据量
- 参数方法
  - 使用一个模型估计数据，只需要存放数据参数，而不是实际数据 (除了可能的离群值)
  - 范例: **Log-linear** 模型
- 非参数方法
  - 不假设模型
  - 主要方法: 直方图, 聚类, 抽样

数据归约方法 (1): 回归与对数线性模型

- 线性回归: $Y = w X + b$
  - 两个回归系数**w**与**b**用于设定回归线, 而这两个系数是用现有数据估计得来
  - 对已知的 $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.使用最小二乘法求解系数
- 多元线性回归: $Y = b_0 + b_1 X_1 + b_2 X_2$.
  - 是线性回归的扩充
- 对数线性模型:
  - 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。

数据归约 (2): 直方图

- 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
- 分裂规则:
  - 等宽: 每个桶范围
  - 等频率(等深)
  - **V-**最优:具有最小方差的直方图（直方图的方差是每个桶代表的原数据的加权和，其中权等于桶内值的个数。）
  - **MaxDiff:**考虑每对相邻值之间的差，桶的边界是具有β-1个最大差的对



厦门大学智能科学系

数据归约 (3): 聚类

- 根据相似度进行聚类，然后通过聚类来表示数据集 (例.,中心与直径)
- 如果数据可以组成各种不同的聚类，则该技术非常有效，反之如果数据界线模糊，则方法无效
- 数据可以分层聚类，并被存储在多层索引树中
- 聚类的定义和算法都有很多选择
- 聚类分析--第七章

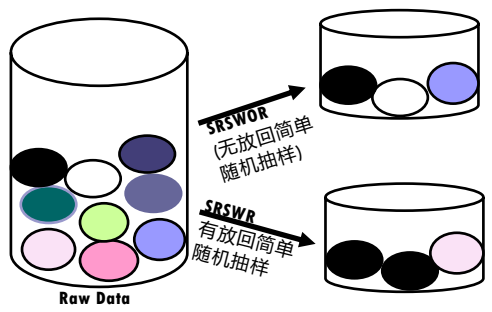
厦门大学智能科学系

数据归约方法 (4): 抽样

- 抽样:允许用数据的较小随机样本*s*（子集）表示大的数据集*N*
- 选取代表数据的子集合
  - 当数据包含倾斜时,简单随机抽样的效果会非常差
- 对数据集*D*的样本选择：
  - *s*个样本无回放简单随机抽样：由*D*的*N*个元组中抽取*s*个样本
  - *s*个样本有回放简单随机抽样：过程同上，只是元组被抽取后，将被回放，可能再次被抽取
  - 聚类抽样：*D*中元组被分入*M*个互不相交的聚类中，可在其中的*s*个聚类上进行简单随机选择（*s*<*M*）
  - 分层抽样：*D*被划分为互不相交的“层”，则可通过对每一层的简单随机抽样得到*D*的分层抽样

厦门大学智能科学系

抽样: 是否放回

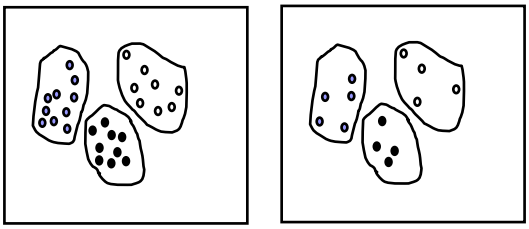


厦门大学智能科学系

抽样: 聚类或分层抽样

原始数据

聚类/分层抽样



厦门大学智能科学系

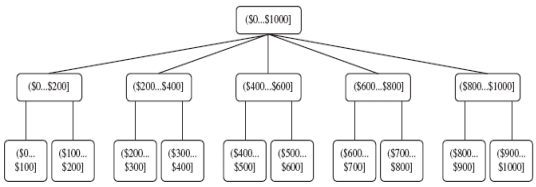
大纲

- 为什么要预处理数据?
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据归约
- 数据离散化和概念分层产生
- 总结

厦门大学智能科学系

数据离散化和概念分层产生

- 离散化
  - 通过将属性值域划分为区间，以减少给定连续属性值的个数
  - 区间的标记可以替代实际的数据值
  - 监督或无监督
  - 分裂 (由上而下) 或合并 (由下而上)
  - 离散化可以递归重复地套用在一个属性
- 概念分层形成
  - 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) 来归约数据



数值数据的离散化与概念分层的产生

- 典型方法: 所有方法可递归重复套用
  - 分箱（binning）
    - 由上而下的分裂,无监督，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）
    - 由上而下的分裂,无监督，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
    - 可以由上而下的分裂或由下而上合并,无监督，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。
  - 基于熵的离散化
    - 有监督,由上而下的分裂
  - 基于χ^2分析的区间合并:无监督,由下而上合并
  - 根据直观划分离散化：由上而下的分裂,无监督

基于熵的离散化

- 给定元组 **D**, 如果 **D**根据属性**A**和某分裂点上的划分分裂为两个区间 D_1与 D_2, 分裂后期望信息需求为

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2)$$

- 熵是根据集合中元组类分布计算得到的. 假设有***m***类别,D_1的熵为

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  p_i为类别在D_1中机率
- 在所有边界值中,使用能最小化熵函数的边界值
- 递归分裂直到停止条件满足为止
- 更有可能将区间边界（分裂点）定义在准确位置，有助于提高分类的准确性。

基于χ^2分析的区间合并

- 合并 (由下而上)
- 合并:递归的找出最佳邻近区间，然后合并它们，形成较大区间
- **ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]**
  - 初始，将数值属性A的每个不同值看作一个区间
  - 对每对相邻区间进行χ^2检验
  - 具有最小χ^2值的相邻区间进行合并
  - 合并的过程递归进行，直到满足预先定义的终止标准 (如显著程度, 最大区间, 最大不一致)

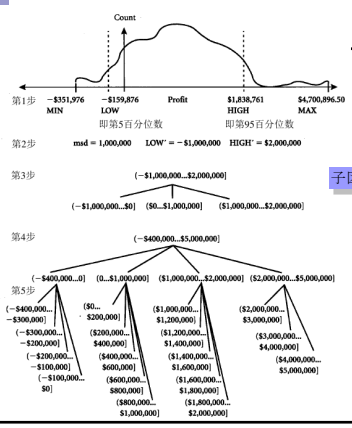
一个区间对的低卡方值表明区间是类独立的，因此可以合并

	1-10	10-30	合计
类1	250(90)	200(360)	450
类2	50(210)	1000(840)	1050
合计	300	1200	1500

根据直观划分离散化

- 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。
- 自然划分的**3-4-5**规则：
  - 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；
  - 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%**

3-4-5 规则示例

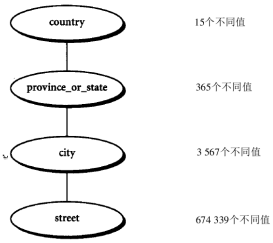


$$\text{子区间个数} = (\max - \min) \div msd$$

分类数据的概念分层产生

- 由用户或专家在模式级显式说明属性的偏序
  - 街 < 城市 < 州 < 国家
- 通过显式数据分组说明分层结构的一部分
  - **{Urbana, Champaign, Chicago} < Illinois**
- 说明属性集但不说明它们的偏序，由系统自动构建
  - 例., 仅有： 街、城市, 其它没有
- 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。
  - 例., 已知一组属性: {街，城市}，根据捆绑的信息，加入相关的属性{州, 国家}

- 概念分层可以根据属性集中属性包含的不同值来自动建立
  - 拥有最多不同值的属性会在阶层的最下层
  - 例外, 例., 星期, 月, 季, 年



15个不同值

365个不同值

3 567个不同值

674 339个不同值

总结

- 数据预处理对数据仓库与数据挖掘是一项重要的问题
- 描述性数据汇总提供数据预处理分析的基础
- 数据预处理包含
  - 数据清理与集成
  - 数据归约与属性选择
  - 离散化
- 尽管已经开发了许多数据预处理的方法，由于不一致或脏数据数量巨大以及问题本身的复杂性，数据预处理仍然是一个活跃的研究领域。