第三章 数据预处理

张静
（模式识别与智能数据研究室）
（Jingzhang@ecust.edu.cn）

主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

为什么要数据预处理?

- 真实世界的数据是“脏的”
  - 不完整: 有些感兴趣的属性缺少属性值, 或者仅仅具备聚集数据，而非具体数据
  - 不正确: 包含错误或者存在偏离期望值的离群值（噪声）
  - 不一致性: 在代码或者名称上存在差异
- 没有高质量的数据，也就没有高质量的挖掘结果!
  - 数据质量是决定数据挖掘结果的重要因素。

如何衡量数据的质量?

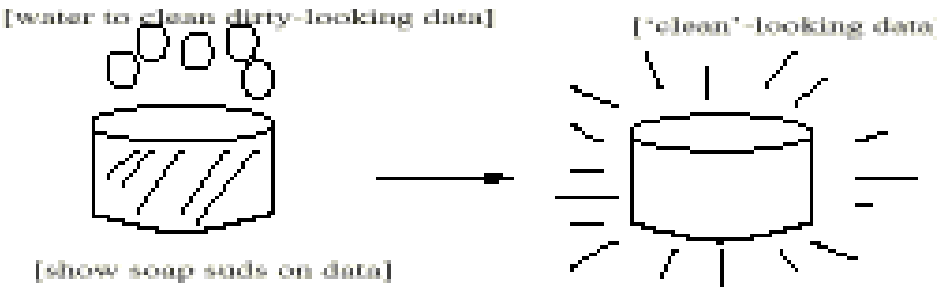
- 如何衡量数据的质量?
  - 准确性（Accuracy）
  - 完整性（Completeness）
  - 一致性（Consistency）
  - 时效性（Timeliness）
  - 可信性（Believability）
  - 可解释性（Interpretability）

数据预处理中的主要任务

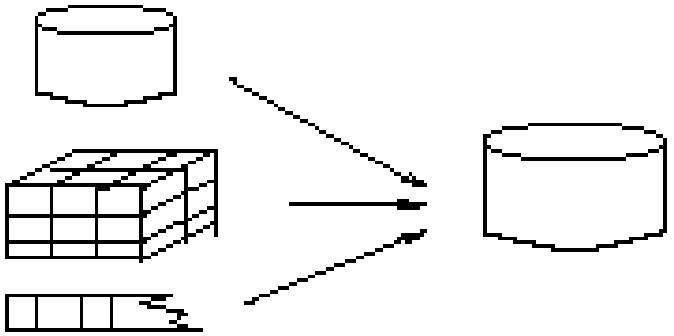
- 数据清理（Data cleaning）
  - 填充缺失值，平滑噪声数据，鉴别或者移除离群点，纠正不一致性问题
- 数据集成（Data integration）
  - 从多个数据库，数据立方体（cube）或者文件中集成
- 数据归约（Data reduction）
  - 减少数据的字段数目，但是仍然产生相同或者近似的分析结果
- 数据变换（Data transformation）
  - 规范化和聚集
- 数据离散化（Data discretization）
  - 数据归约的一部分，对于从数值数据自动产生概念分层非常有用

数据预处理的形式

Data Cleaning

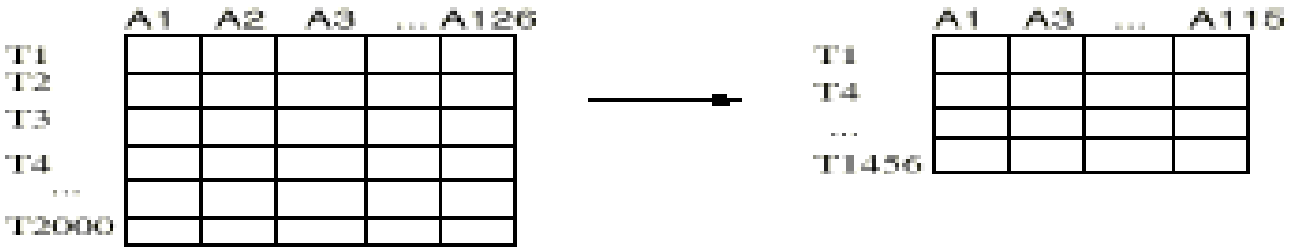


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

数据清理

- 数据清理任务
  - 填充空缺值
  - 平滑噪声数据
  - 纠正不一致的数据

空缺值

- 数据不总是可用的
  - 很多记录的许多属性难以获取，比如：在销售数据中客户的收入情况很难掌握。
- 数据缺省的原因
  - 装备的故障
  - 由于和其他数据不一致而被删除了
  - 由于理解错误而未被录入
  - 某些数据在录入的时候并不认为是重要的，因此没录入
  - 没有注册历史或者数据改变了
  - .....
- 空缺的数据可能需要被推断出来

空缺值

- 忽略该记录： 当类标号缺少时通常这样做（假定挖掘任务涉及分类或者描述）。除非元组有多个属性缺少值，否则该方法不是很有效。当每个属性缺少值的百分比变化很大时，它的性能非常差。
- 人工填写空缺值：一般地说，该方法很费时，并且当数据集很大、缺少很多值时，该方法可能行不通。
- 使用一个全局的常量填写空缺值，例如 “unknown”，但是这可能会引入一个新类型?!
- 使用属性的中心度量（如均值、中位数）填充空缺值
- 使用与给定元组属同一类的所有样本的属性均值或中位数：例如，若将顾客按credit_risk分类，使用具有相同信用度的顾客的平均收入替换income中的空缺值
- 使用最可能的值去填充空缺值：基于推导的使用贝叶斯公式或者决策树

噪声数据

- 噪声（noise）：是一个测量变量中的随机错误或偏差

- 不正确的属性值可能导致
  - 数据转换问题
  - 技术限制
  - 命名转换过程的不一致性

- 其他需要数据清理的数据问题
  - 重复记录
  - 不完整数据
  - 不一致数据

如何处理噪声数据?

- 分箱（binning）
  - 首先，把数据排序，把排序后数据分到等深的箱中
  - 接着，用按箱平均值、中心值、边界值等平滑技术平滑化数据
- 回归
  - 利用回归函数填充数据，从而平滑化数据
- 离群点分析
  - 检测和移除离群点

简单离散化方法：分箱

- 等宽划分
  - 将整个区域划分成**N**个相同大小的间隔
  - 若***A***和***B***是这个属性的最小值和最大值，则各个间隔的宽度为：$W = (B-A)/N$.
  - 等宽分箱法是最直接的分箱方法
  - 但是离群点可能会影响表示
  - 倾斜的数据并不能够很好地被处理
- 等深划分
  - 将整个区域划分为***N***个间隔，各个间隔中所包含的样本数目大致相同
  - 具有较好的数据扩展性

数据平滑的分箱方法

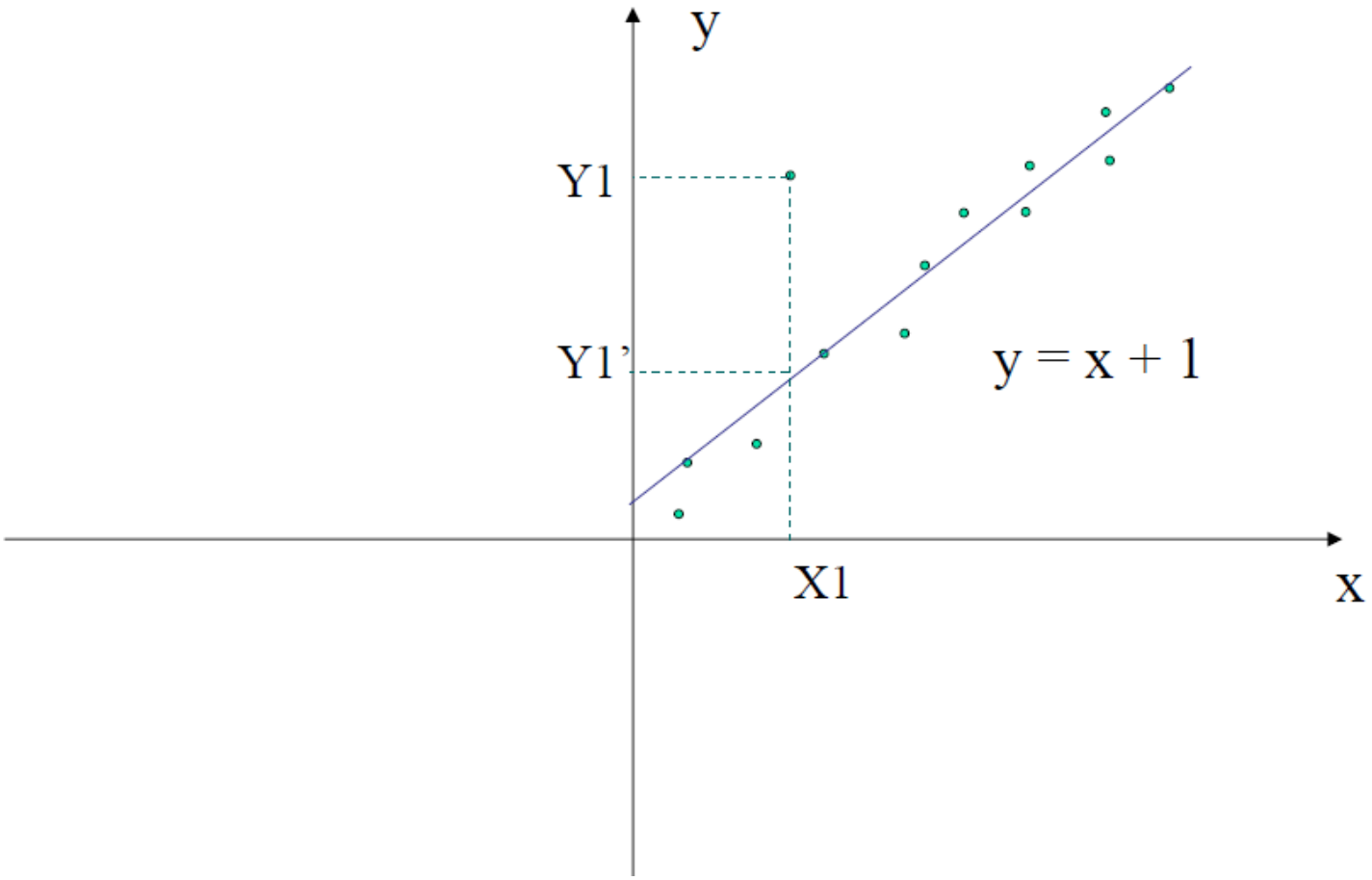
price 的排序后数据(美元): 4, 8, 15, 21, 21, 24, 25, 28, 34

- 划分为（等深的）箱:
- **-箱1: 4, 8, 15**
- **-箱2: 21, 21, 24**
- **-箱3: 25, 28, 34**

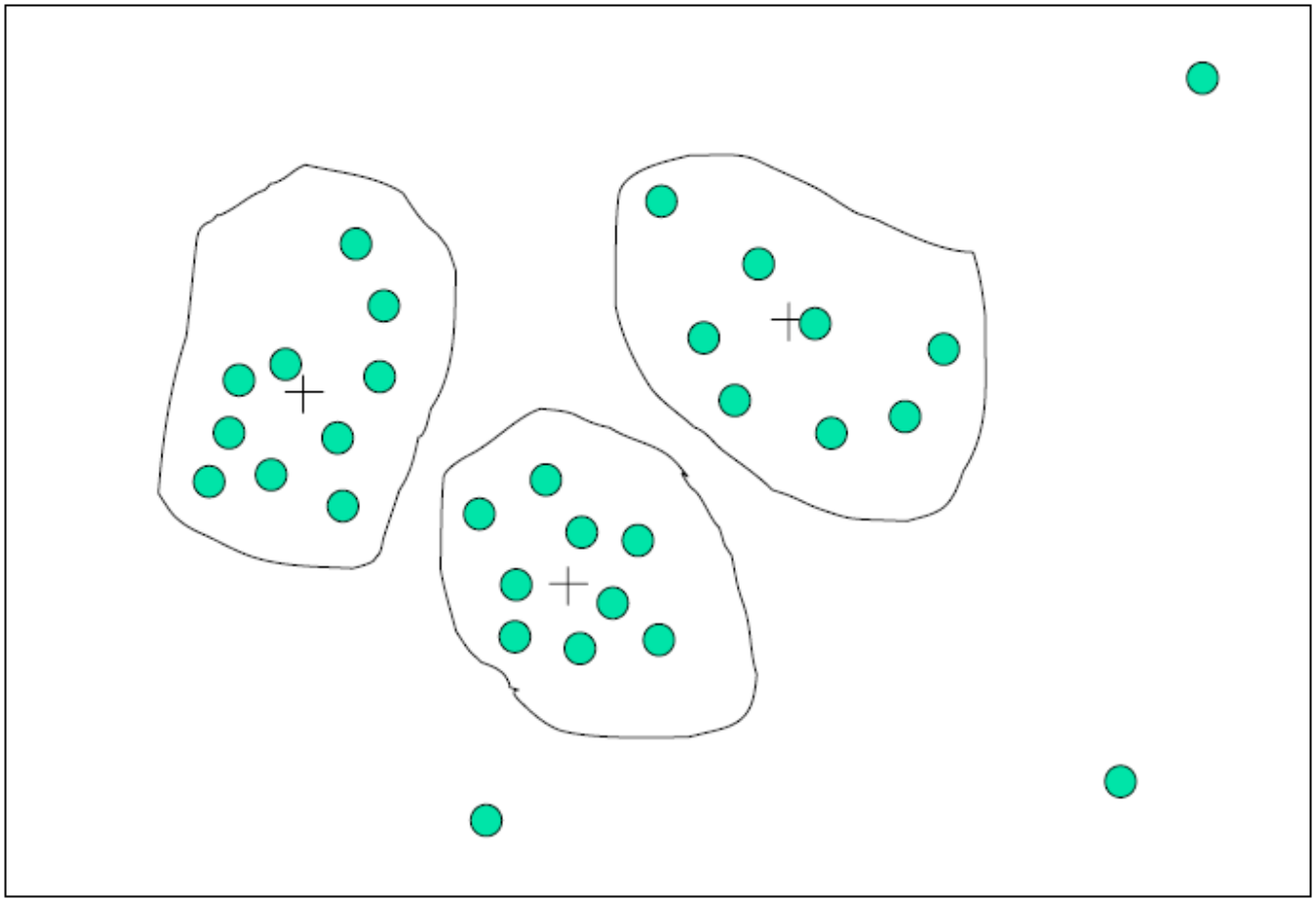
- 用箱平均值平滑:
- **-箱1: 9, 9, 9**
- **-箱2: 22, 22, 22**
- **-箱3: 29, 29, 29**

- 用箱边界值平滑:
- **-箱1: 4, 4, 15**
- **-箱2: 21, 21, 24**
- **-箱3: 25, 25, 34**

回归



离群点分析



主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

数据集成

- 数据集成
  - 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（如数据仓库）中
- 模式集成
  - 从不同的数据源集成元数据
  - 实体识别问题: 来自多个信息源的现实世界的实体如何才能“匹配”？ 例如, **A.cust-id ≡ B.cust-number**
- 数据冲突的检测与处理
  - 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: 不同的表示方式，不同的度量标准，例如公制单位和英制单位

处理数据集成中的冗余数据

- 当多个数据库的数据集成时，会产生冗余数据
  - 在不同数据库中，相同的属性可能具备不同的名称
  - 一个属性可能由另外一张表的多个字段推导出，例如：年收入
- 有些冗余可以被相关分析检测到
- 对多数据源中的数据进行仔细的数据集成，可以减少/避免冗余和矛盾，并且能提高挖掘的速度和质量

相关分析(数值数据)

- 相关系数（Correlation coefficient） (also called Pearson's product moment coefficient)

$$r_{p,q} = \frac{\sum(p-\bar{p})(q-\bar{q})}{(n-1)\sigma_p\sigma_q} = \frac{\sum(pq)-n\bar{p}\bar{q}}{(n-1)\sigma_p\sigma_q}$$

  - 其中 n 是 元组个数，$\bar{p}$ 和 $\bar{q}$分别是p 和 q 的平均值, σ_p 是 σ_q 分别是p 和 q的标准差, Σ(pq) 是pq 叉积的和（即，对于每个元组，A的值乘以该元组B的值）。
- 如果 $r_{p,q} > 0$, p 与 q 正相关，值越大，相关性越强.
- $r_{p,q} = 0$: 相互独立；$r_{p,q} < 0$: 负相关。

相关分析(离散数据)

- **X^2 (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- **X^2** 值越大, 两个变量越相关
- 实际值与期望值差别越大，则其对**X^2** 值贡献越大。
- 相关性不包含因果性
  - 例如：医院是的数量与偷车贼的数量是相关的，实际上它们都与另一个变量相关：人口数量

主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

数据归约策略

- 数据仓库可以存储数千兆字节的数据：在海量数据上进行复杂数据分析和数据挖掘需要很长时间
- 数据归约
  - 数据归约技术可以用来得到数据集的规约表示，它在规模上要小得多，但能产生同样（或几乎同样的）的分析结果
- 数据归约策略
  - 维规约：减少所考虑的随机变量或属性的个数
  - 数量规约：用替代的、较小的数据表示形式替换原数据
  - 数据压缩：使用变换以得到元数据的归约或压缩表示。

数据立方体聚集

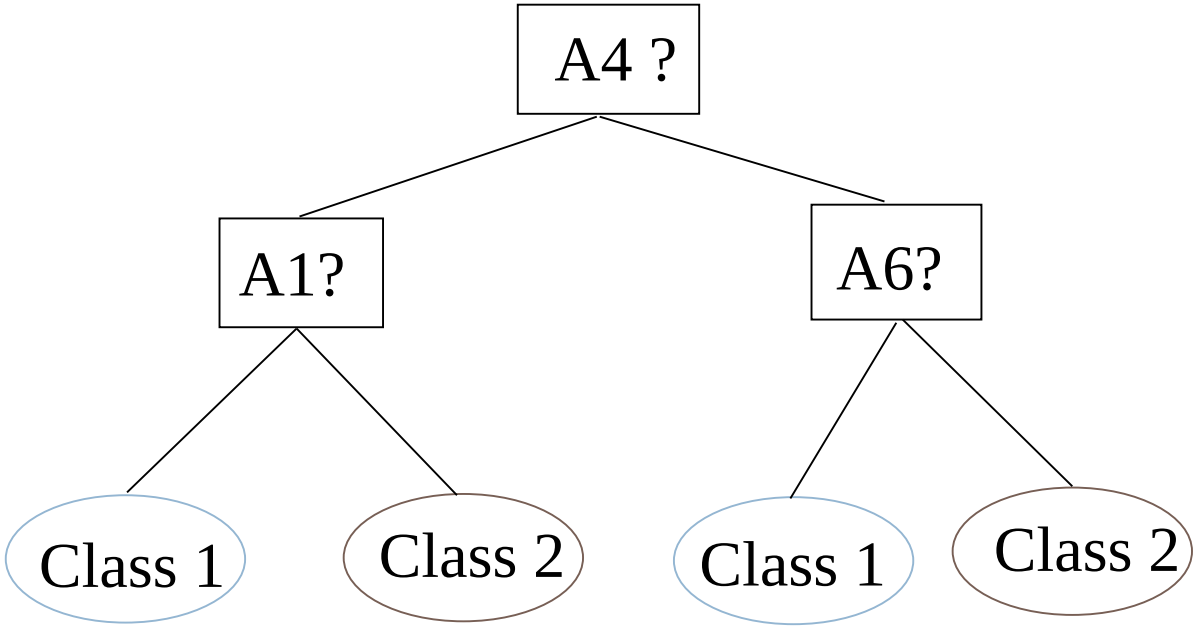
- 数据立方体的最低层为基本方体，最高层为顶点方体，中间层为方体。
  - 对应于感兴趣实体的聚集数据
- 数据立方体中聚集多层次
  - 进一步减少了要处理数据的大小
- 有关聚集信息的查询,如果可能的话,应当使用数据立方体回答
- 第四章将详细介绍

属性子集选择

- 定义：通过删除不相关或者冗余的属性（或维）减少数据集
- 目标：找出最小属性集，使得数据类的概率分布尽可能接近原数据集的概率分布
- 优点：减少了出现在发现模式上的属性的数目，使得模式更易于理解
- 启发式方法：
  - 逐步向前选择（空集开始，每次添一最优属性）
  - 逐步向后删除（满集开始，每次删一最差属性）
  - 向前选择和向后删除的结合
  - 决策树归纳

决策树归纳

初始属性集合：
{A1, A2, A3, A4, A5, A6}



-------> **归约后的属性集合：{A1, A4, A6}**

如何判断属性的重要性?

- 很多方法，例如:
  - information gain (ID3)
  - gain ratio (C4.5)
  - gini index
  - χ^2 contingency table statistics
  - uncertainty coefficient

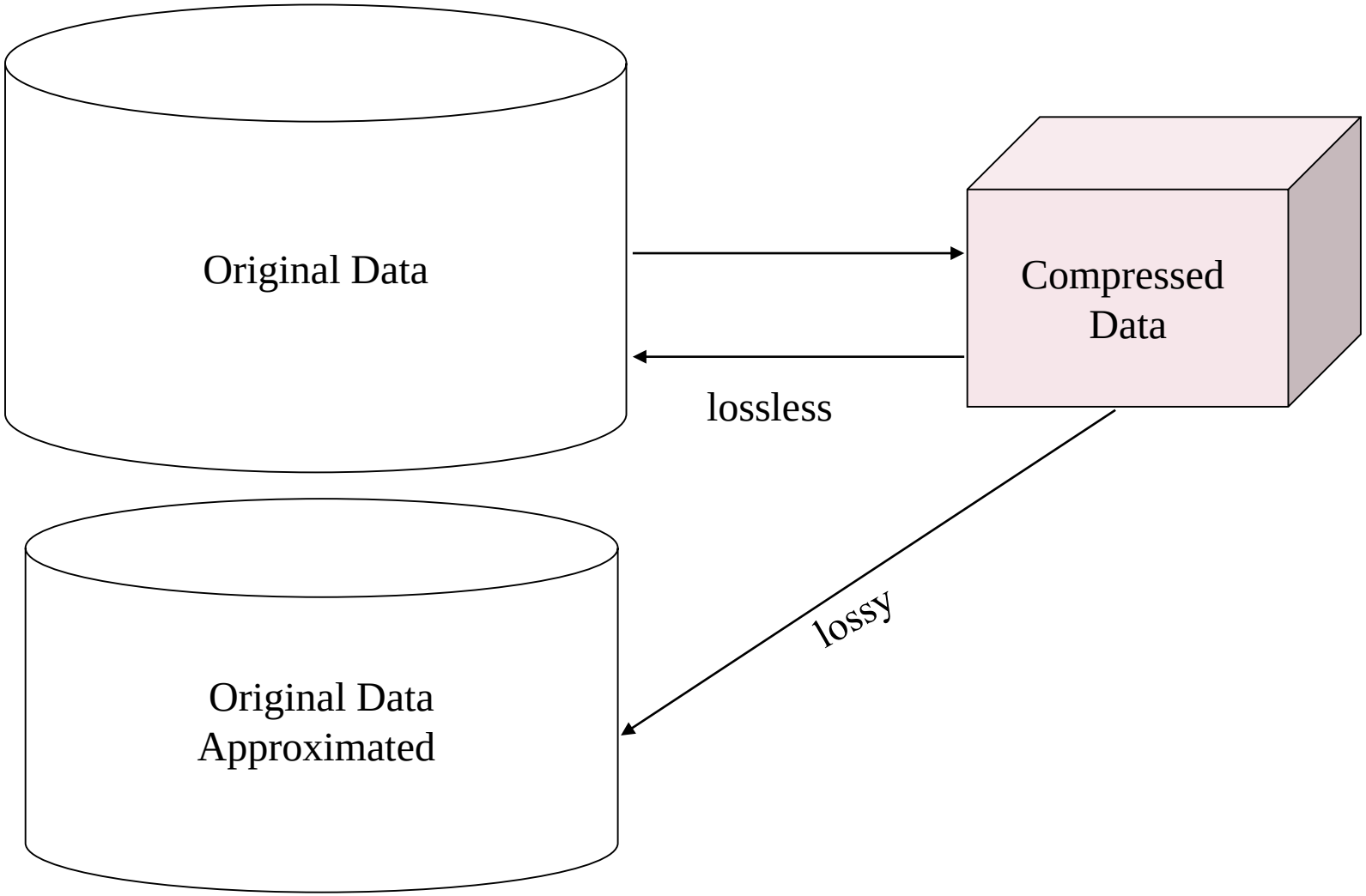
维规约

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

维归约（数据压缩）

- 数据压缩分类
  - 无损压缩
  - 有损压缩
- 串压缩
  - 典型的无损压缩
  - 已有广泛的理论和协调的算法
  - 但是只允许有限的数据操作
- 音频/图像压缩
  - 典型的有损压缩，逐步加细
  - 有时可以只重构信号的小片断，而无需重构整个信号

数据压缩


Original Data
Compressed Data
lossless
lossy
Original Data Approximated

小波变换

- 离散小波变换 (DWT): 一种线性信号处理技术
- 近似压缩技术: 仅仅保存一小部分最强的小波系数
- 类似于离散傅立叶变换 (DFT), 但是DWT是更好的有损压缩，空间局部性相当好
- 方法:
  - 输入数据向量的长度L必须是2的整数幂（必要时可在数据向量后加0）
  - 每个变换涉及两个应用函数。第一个使用数据平滑，如求和或加权平均；第二个进行加权差分，产生数据的细节特征
  - 两个函数作用于输入数据对，产生两个长度为L/2的数据集。一般地，他们分别代表输入数据平滑后的低频和高频内容
  - 两个函数递归地作用于前面循环得到的数据集，直到结果集的长度为2
  - 由以上迭代得到的数据集中选择值，指定其为数据变换的小波系数。

小波变换的优点

- 小波空间局部性好，有利于保留局部细节
- 对噪音和数据的输入顺序不敏感
- 计算复杂度为O(N)，具有较高的计算效率
- 小波变换可以用于高维数据
- 小波变换（DWT）的有损压缩比离散余弦变换（DCT）压缩效果好
  - JEPG（DCT）
  - JEPG2000（DWT）
- 小波变换有许多实际应用，包括：指纹图像压缩，计算机视觉，时间序列数据分析和数据清理。

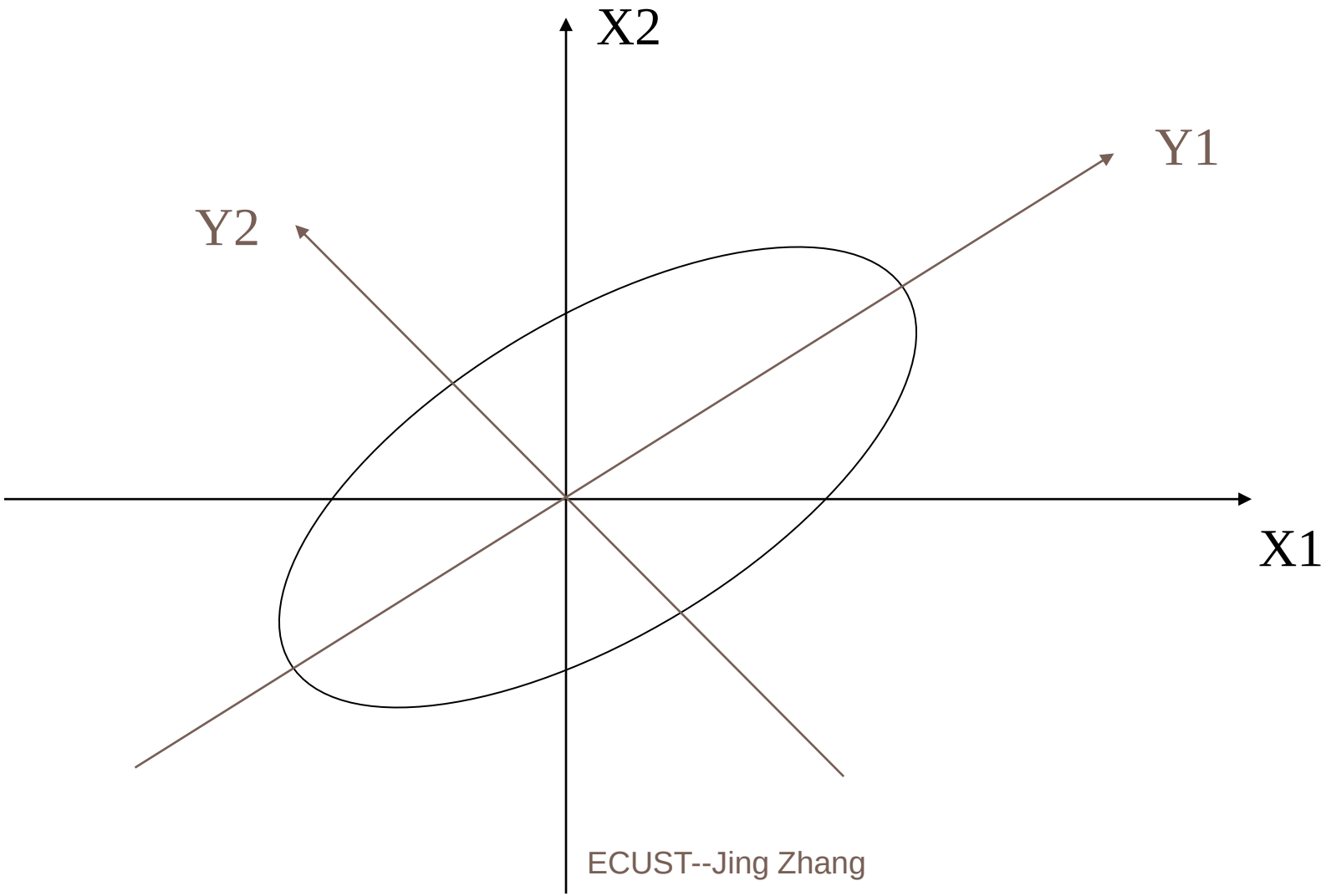
主成分分析

- 假定待压缩的数据由*N*个元组或者数据向量组成，取自*k*个维。主成分分析（PCA）搜索c（且c <= *k*） 个最能够代表数据的k-维正交向量。
  - 元数据集被归约到一个由c个主要成分上的N个数据向量构成的空间上（维归约）
  - 每一个数据矢量都是c个主要成分矢量的线性组合
  - 仅仅针对数值型数据
  - 对高维数据较为有效
- 与小波变换比，PCA能较好地处理稀疏数据，而小波变换更适合高维数据。

主成份分析

- 基本过程
  - 对输入数据规范化，使得每个属性都落入相同的区间。
  - **PCA**计算**c**个规范正交向量，作为规范化输入数据的基。这些向量被称为主成份，输入数据是主成份的线性组合。
  - 对主成分按“重要性”或强度降序排列。
  - 通过去掉较弱的成分来压缩数据。

主成分分析



数值规约

- 数值规约技术
  - 通过选择替代的、较小的数据表示形式来减少数据量。
  - 参数方法和非参数方法
- 参数方法
  - 假设数据适合一些模型，评估模型参数，使得只需存放模型参数，而不是实际数据（离群点也可能被存放）
  - 如对数线性模型：估计离散的多维概率分布。
- 非参数方法
  - 不必假设模型
  - 主要包括：直方图，聚类和选样。

回归和对数线性模型

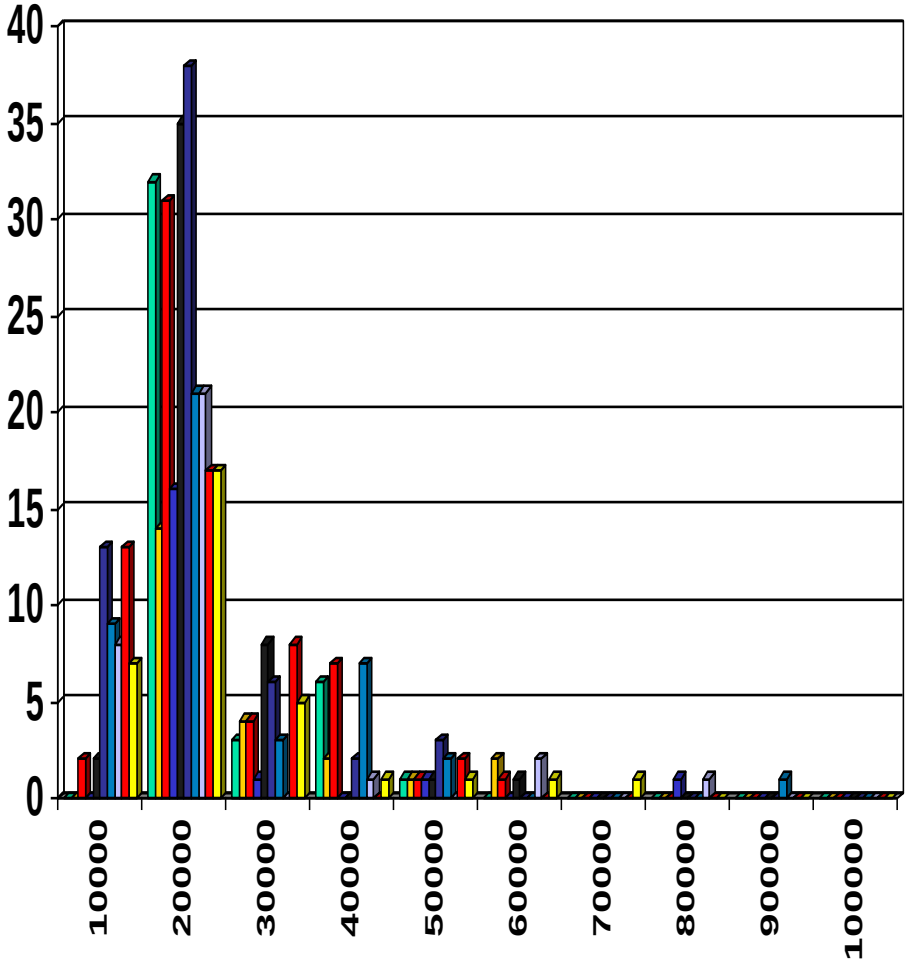
- 线性回归：对数据建模，使之适合一条直线

$$Y = \alpha + \beta X$$

  - 两个参数，α和β确定这条直线，能够利用手头的数据进行估计
  - 通常使用最小平方法来确定直线方程的系数
  - 多元回归是线性回归的扩充，相应变量是多维特征向量的线性函数。 **Y = b0 + b1 X1 + b2 X2.**
- 对数线性模型：近似离散的多维概率分布
- 回归和对数线性模型都可用于稀疏数据。对于高维数据，回归可能是计算密集的，而对数线性模型则可以表现出很好的可伸缩性。

直方图

- 一种流行的数据归约技术
- 把数据分成不同的桶，存储每个桶的平均值
- 划分规则
  - 等宽
  - 等频（等深）



聚类

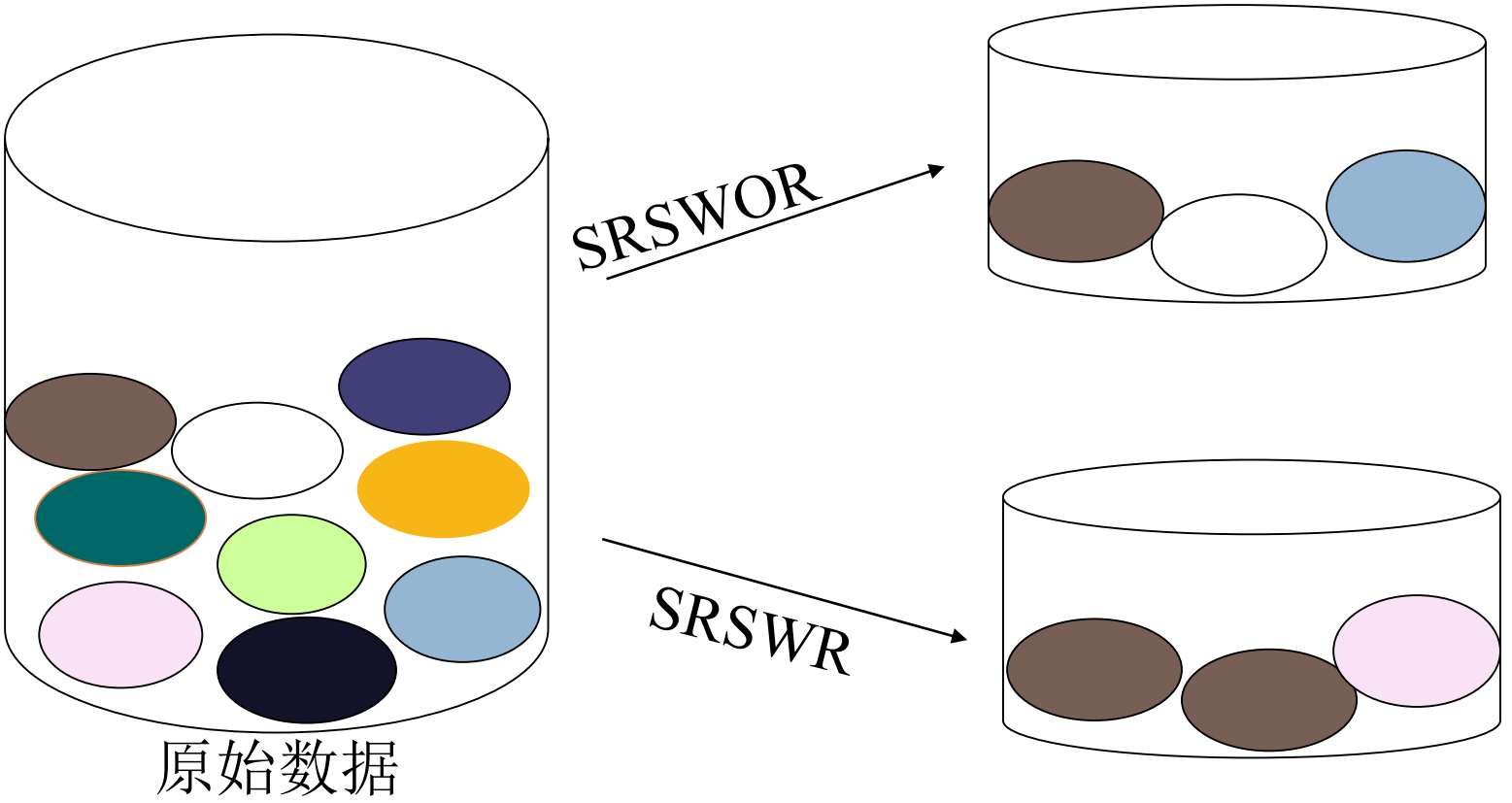
- 把数据集划分成聚类，使得类内数据相似，类间数据不相似，从而只存储聚类的表示
- 如果数据是聚集的，聚类技术将十分有效，而当数据有噪声时将失去它的有效性
- 可以层次聚类且被存储在多维索引树结构中

抽样

- 用数据的较小随机样本（子集）表示大的数据集。
- 选择数据的代表子集
  - 简单随机取样在有倾斜数据时可能会执行得比较差
- 抽样方法
  - 简单选择n个样本，不回放（SRSWOR）
  - 简单选择n个样本，回放（SRSWR）
  - 聚类抽样
  - 分层抽样
    - 把数据库D划分为互不相交的部分，称作“层”，则通过对每一层的简单随机取样就可以得到D的分层选样
      - 当数据倾斜时，可以帮助确保样本的代表性
- 抽样的复杂性子线性于数据的大小。

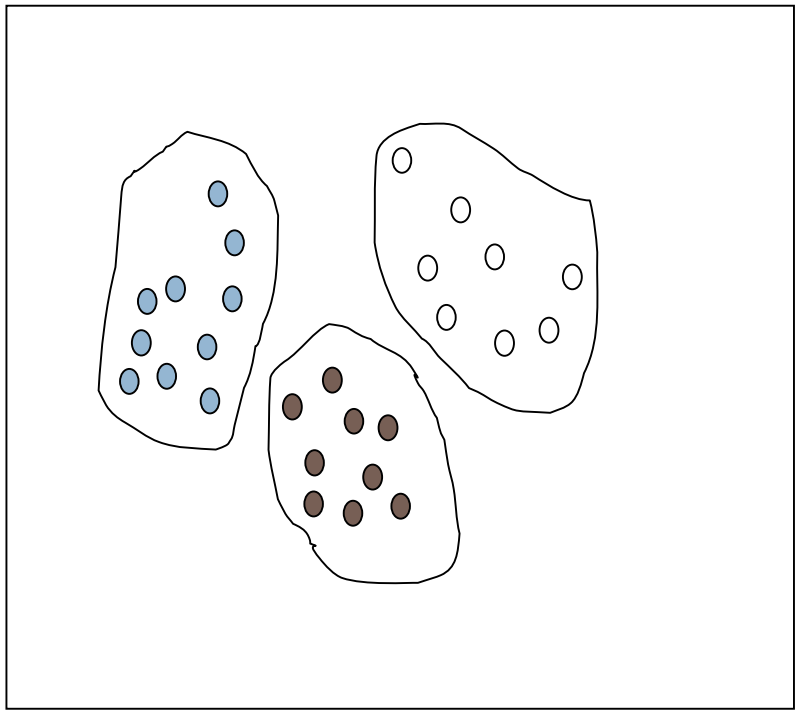
抽样（Sampling）

SRSWOR: 简单选取n个样本，不回放
SRSWR: 简单选取n个样本，回放

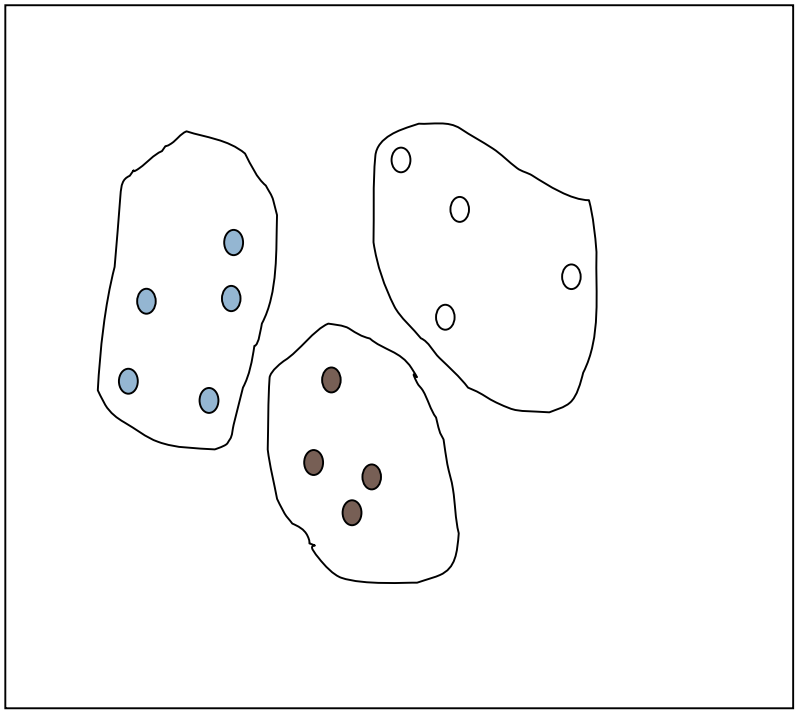


聚类抽样

原始数据



聚类抽样



主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

数据变换

- 平滑: 去掉数据中的噪声
- 属性构造：由给定的属性构造新的属性，并添加到属性集中
- 聚集: 对数据进行汇总和聚集
- 离散化: 数值属性的原始值用区间标签或概念标签替换
- 规范化: 将属性数据按比例缩放，使之落入一个小的特定区间
- 由标称数据产生概念分层：用高层次概念替换低层次“原始”数据。

数据变换：规范化

- 最小-最大规范化
  - 将A的值v映射到区间[new_min_A, new_max_A]中的v′

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

- z-score规范化
  - 属性A的值基于A的平均值和标准差规范化，A的值v被规范化为v′。

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

- 小数定标规范化
  - 通过移动属性A的小数点的位置进行规范化。小数点的移动位数依赖于A的最大绝对值。A的值v被规范化为v′。

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

其中，j 是使得Max(|v'|)<1的最小整数。

离散化

- 属性的三种类型：
  - 标称属性—来自无序集中的值
  - 序数属性—来自有序集的值
  - 连续属性—实数
- 离散化：
  - 把连续的属性值区间划分成多个区间
  - 一些分类算法只接受分类属性
  - 通过离散化压缩数据大小
  - 为进一步分析作准备

离散化和概念分层

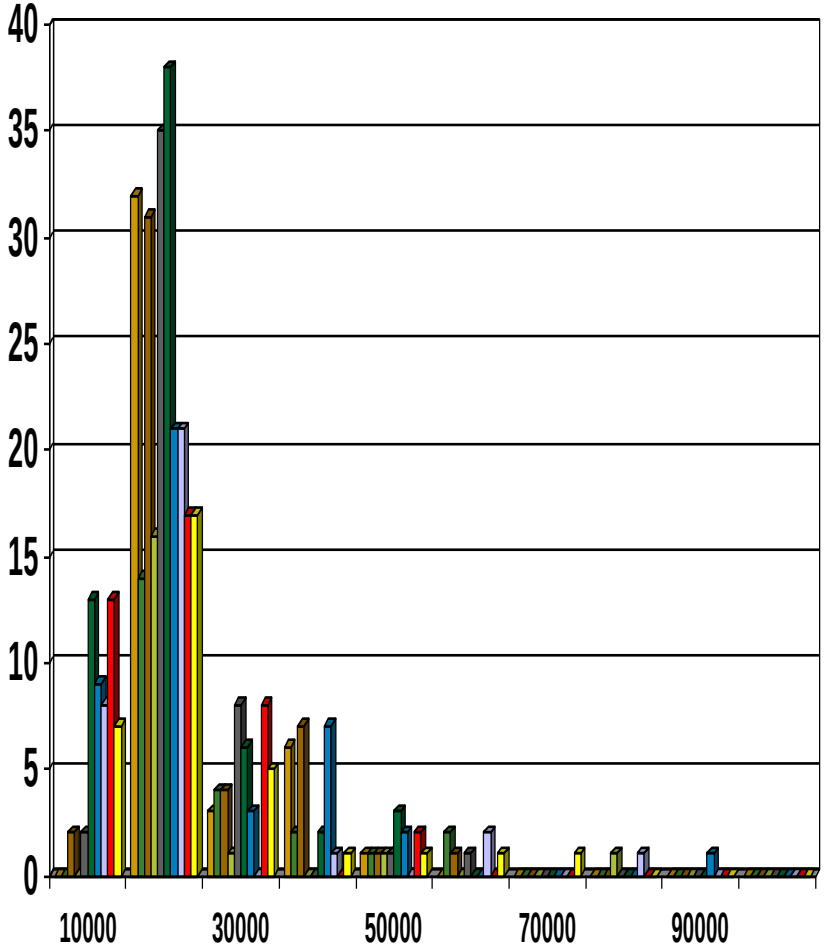
- 离散化
  - 通过将一个连续型的属性划分成少数几个间隔范围，从而降低取值的数目。间隔的标签被用于表示该字段的真实值。
- 概念分层
  - 将低级的概念(例如以数值形式表示年龄字段)转化为更高级别的概念（例如，以青年，中年，老年表示年龄字段）。

针对数值型数据的离散化和概念层次化

- 分箱 （前面已讲过）
- 直方图分析
- 基于熵的离散化
- 基于X^2 (chi-square) 分析的区间合并
- 聚类分析
- 根据直观划分离散化

直方图

- 一个常用的非监督数据规约技术
- 将数据划分成多个桶（buckets），并且以平均值（总和）表示每个桶
- 能够通过动态规划的方法优化生成



基于熵的离散化

- 利用熵的值递归地划分数值属性A的值，产生分层的离散化。
- 给定一个样本集S，基于熵对A离散化的方法如下：
  - A的每个值可以认为是一个潜在的区间边界或阈值T。
  - 给定S，所选择的阈值时这样的值，它使其后划分得到的信息增益最大。信息增益是：

$$E(S,T)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2)$$

    - 其中，s1和s2分别对应于s中满足条件A<T和A≫ T的样本。对于给定集合，它的熵函数Ent根据集合中样本的类分布来计算。例如，给定m个类，Si的熵为：（pi是类i在Si中的概率）

$$\mathrm{Ent}(S_1)=-\sum_{i=1}^{m}p_i\log_2(p_i)$$

  - 确定阈值的过程递归的用于所得到的每个划分，直到满足某个终止条件，如：

$$Ent(S)-E(T,S)>\delta$$

- 实验证明基于熵的离散化可以压缩数据量，提高分类的准确性

基于X^2 (chi-square) 分析的区间合并

- 基本思想
  - 对于精确的离散化，相对类频率在一个区间内应当相当一致。如果两个邻近的区间具有非常类似的类分布，则这两个区间可以合并。否则，它们应该保持分开
- 过程
  - 把数值属性A的每个不同值看做一个区间
  - 对每对相邻区间进行X^2检验
  - 把具有最小X^2值的相邻区间合并在一起
  - 以上各步递归进行，直到满足预先定义的终止标准

聚类

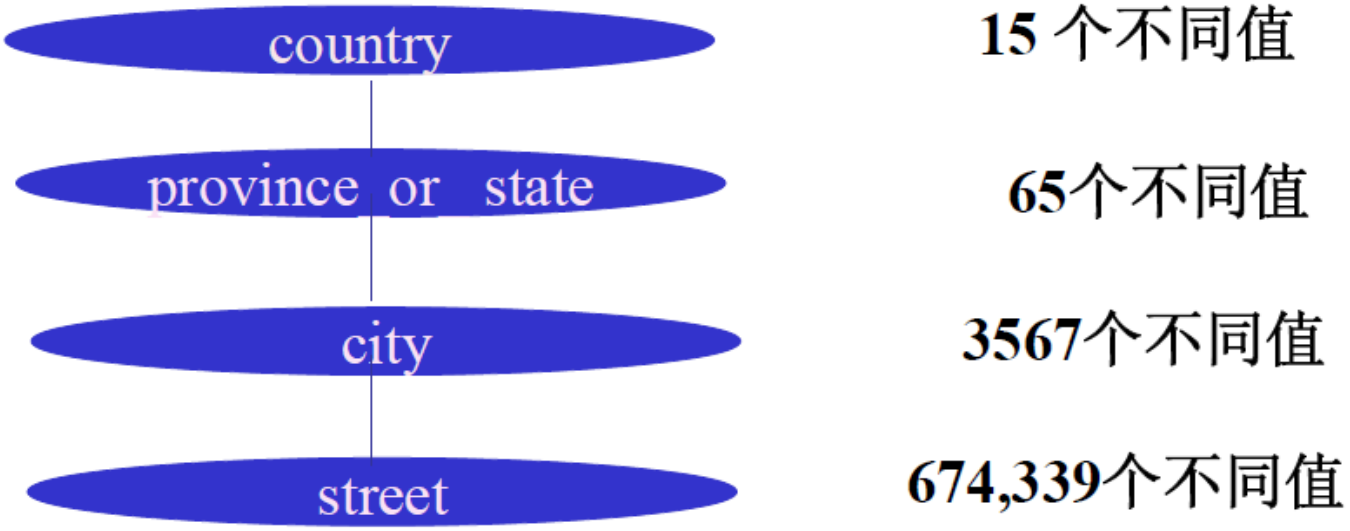
- 将数据集合划分为多个簇，然后仅仅以簇代表数据
- 如果数据本身可以分为多个簇，则较为有效
- 每一个簇可以进一步分成若干子簇，形成较低的概念层。
- 簇可以聚集在一起，以形成分层结构中较高的概念层

标称数据的概念分层生成

- 由用户或专家在模式级显示地说明属性的部分序
- 通过显式数据分组说明分层结构的一部分
- 说明属性集，但不说明他们的偏序
- 只说明部分属性集

属性集的说明

- 概念分层可以根据给定属性集中每一个属性的不同属性值的个数自动生成。具有最多不同属性值的属性放在分层中的最低层



主要内容

- 数据预处理概述
- 数据清理
- 数据集成
- 数据归约
- 数据变换与数据离散化
- 小结

小结

- 数据预处理对于数据仓库和数据挖掘都是一个重要的问题
- 数据预处理包括
  - 数据清理和数据集成
  - 数据归约和特征选择
  - 离散化和概念分层
- 尽管已经提出了一些数据预处理的方法，数据预处理仍然是一个活跃研究领域

推荐参考文献

1. R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann (in preparation)
2. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
3. U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
4. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
5. D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
6. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
7. T. M. Mitchell, Machine Learning, McGraw Hill, 1997
8. G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
9. S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
10. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2001

课后作业：P79-80

- 习题3.3
  - （a）使用深度为3的箱，分别用箱均值和箱边界值平滑以上数据。
- 习题3.6
  - （a） （b） （d）
- 习题3.9
  - （a） （b）
- 习题3.11
  - （a）

思考题

- 什么是有监督（指导）学习？什么是非监督学习？试分析下列离散化方法：分箱、直方图分析、基于熵的离散化、基于χ^2分析的区间合并、聚类分析是有监督的还是非监督的？

结束