Data Preprocessing

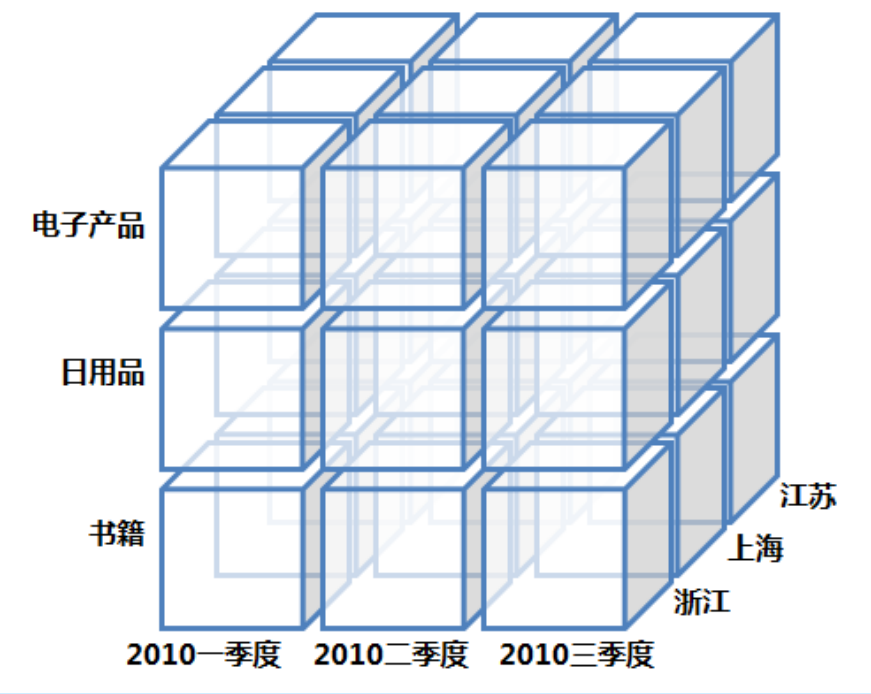
- 关于数据
- 为什么要预处理数据？
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据规约
- 小结

数据归约策略

- 一个数据仓库可能存储了兆兆字节的数据
  - ✓ 复杂的数据挖掘，在完整的数据集上运行，可能会花很长的时间
- 数据归约
  - ✓ 获得数据的简化表达方式，这些数据在容量上更小，但是会产生相同（或几乎相同）的分析结果
- 数据归约策略
  - ✓ 数据立方体聚集
  - ✓ 维度缩减—去掉不重要的属性
  - ✓ 数据压缩—小波变换
  - ✓ 数值归约—将数据拟合到模型中
  - ✓ 离散化和观念分层生成

数据立方体聚集

- 数据立方体是一类多维矩阵，让用户从多个角度探索和分析数据集
- 数据立方体的最低水平 **(base cuboid基数长方体)**
  - ✓ 为感兴趣的单个实体汇总的数据
  - ✓ 例如, 电话数据仓库中的一个客户
- 数据立方体中的多层次聚集
  - ✓ 进一步归约要处理的数据
- 参考适当的水平
  - ✓ 用最精简的表示方式就足以解决任务



维度归约

- 特征选择

  选择特征的一个最小集，这样的话，对于给定这些特征的值的不同类别的可能分布，就会与给定所有特征的值的原始分布尽可能的接近

  - ✓ **reduce # of patterns in the patterns, easier to understand**

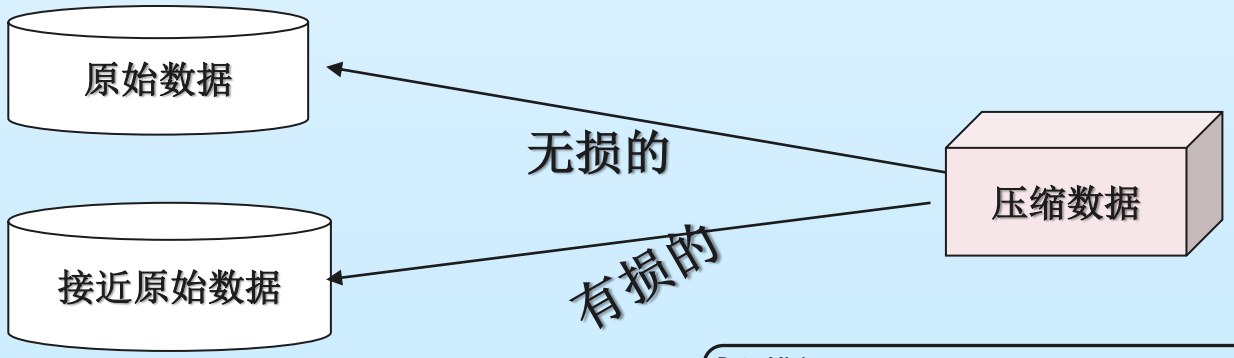
数据压缩

字符串压缩

- ✓ 有大量的理论和调整好的算法
- ✓ 通常无损的
- ✓ 但只有有限的操纵是可能的，没有扩张

音频/视频压缩

- ✓ 通常是有损压缩,逐步细化
- ✓ 有时，小片段的信号可以重建，而不用重建整个信号



数值归约

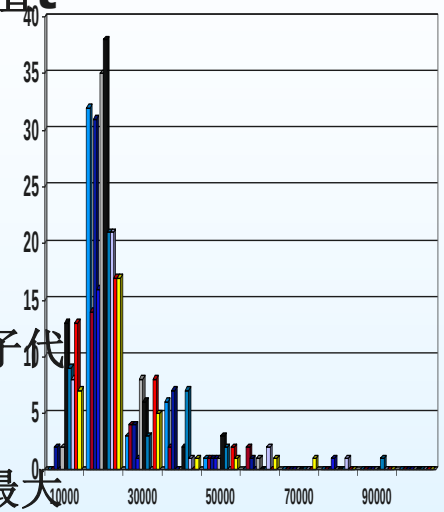
- 通过选择替代的，较小的数据表示形式来减少数据量
- 参数方法
  - ✓ 假设数据符合某个模型，估计模型的参数，仅存储参数，并丢弃数据 (除了可能的离散点)
- 非参数方法
  - ✓ 不假设模型
  - ✓ 主要成员:直方图，聚类，抽样

减少数据方法(1): 回归和对数线性模型

- **线性回归: 将数据拟合成一条直线**
  - ✓ **通常用最小二乘法拟合直线**
- **多元回归:**允许响应变量Y作为一个多维特征向量的线性函数模型
- **对数线性模型:**近似离散的多维概率分布

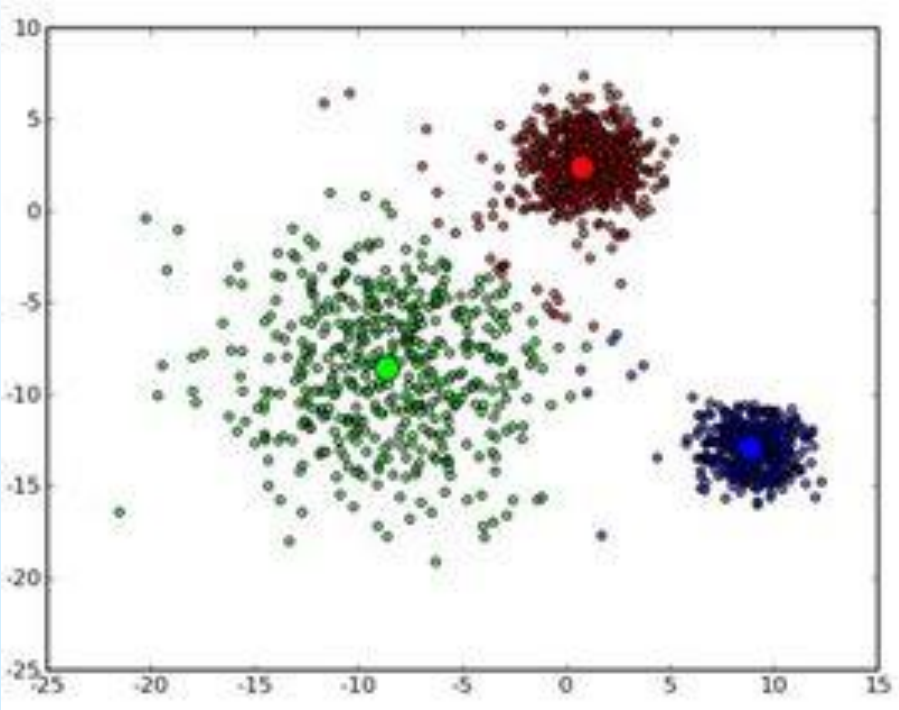
归约数据方法(2): 直方图

- 将数据分割，存储每一部分的平均值t
- 划分规则:
  - ✓ 等宽: 相等的间距
  - ✓ 等频 (或者等深)
  - ✓ V-最优: 最小直方图方差 (每个篮子代表原始值的加权和)
  - ✓ MaxDiff: 桶的边界是具有β–1个最大差的对



数据归约方法(3): 聚类

- 在相似性基础上将分区数据设置成聚类，并只储存聚类的表示方式 (例如, 质心和直径)
- 如果数据是聚类的而不是杂乱的，会非常有效

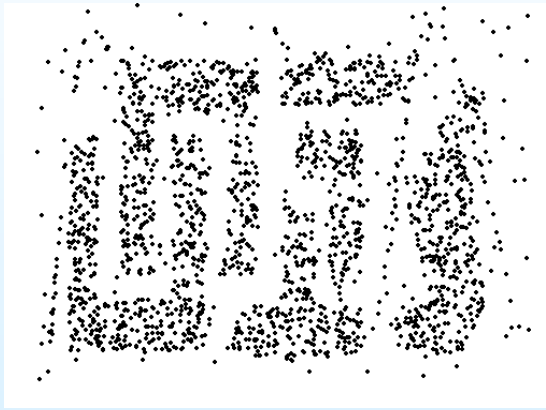


归约数据方法(4): 取样

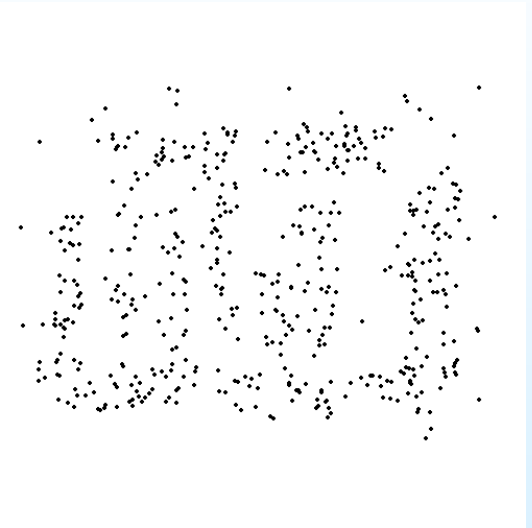
- 允许用数据的小得多的随机样本（子集）表示大型数据集
- 选取数据的一个有代表性的子集
  - ✓ 简单随机抽样，可能在存在偏差的情况下效果很差
- 自适应采样方法
  - ✓ 分层采样:
    - 整个数据库中每类的近似百分比
    - 在有偏数据的结合中使用

抽样

8000 points



2000 Points



500 Points