第三章数据预处理

2021年11月10日 13:17

1. 概述
2. 数据抽取
3. 数据清洗
4. 数据集成
5. 数据归约
6. 数据变换与数据离散化
7. 小结

1. 概述
    - 数据质量：
        - 完整性
        - 一致性
        - 有噪声
        - 准确性
        - 时效性
        - 可信性
        - 可解释性
    - 主要任务：
        - 抽取
        - 清洗
        - 集成
        - 降维
        - 转换
2. 数据抽取

3. 数据清洗
    - 潜在问题：不完整、有噪声、不一致、蓄意
    - 缺失值：
        - 原因：
            - 设备异常
            - 人为错误
            - 记录数据时，有些数据因得不到重视而没有被输入
            - 不适用
        - 处理：
            - 忽略元组
            - 人工填写空缺值：重新采集、利用领域知识等。工作量大、枯燥乏味、

可行性低

- 自动填充：
  - 用一个全局常量填充，如“unknown”或“-∞”
  - 使用属性的平均值填充
  - 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：如像Bayesian公式或判定树这样的基于推断的方法

- 噪声数据
  - 原因：
    - 数据收集工具的问题
    - 数据输入错误
    - 传输错误
    - 技术限制
  - 处理：
    - 分箱（Binning）

      a) 按递增顺序对数据进行排列

      b) 分到（等频的）箱子中

      c) 使用光滑计数（箱均值光滑、箱中位数光滑、箱边界光滑）

      Example: 4, 8, 15, 21, 21, 24, 25, 28, 34

      - 首先排序，并划分到大小为3的等频的箱中（即每个箱包含3个值）

	用箱均值光滑	用箱边界光滑
bin 1: 4, 8, 15;	9 , 9, 9	4, 4, 15
bin 2: 21, 21, 24;	22, 22,22	21,21,24
bin 3: 25, 28, 34	29, 29,29	25, 25,34

  - 离群点：
    - 聚类：将联系松散的数据当作离群点，检测并去除离群点，聚类集合之外的点即是离群点
    - 回归：让数据适应回归函数来平滑曲线
    - 盒图：通过盒图画出离群点
    - 计算机和人工检查结合：计算机检测可疑数据，人工判断

4. 数据集成
   - 概念：将多个数据源中的数据合并，存放在一个数据存储中。如放在数据仓库中
   - 作用：有助于减少结果数据集的冗余和不一致，提高挖掘过程的准确性和速度
   - 需要注意的问题：
     - 4.1实体识别问题
     - 4.2属性冗余和相关性
     - 4.3元祖重复
     - 4.4数据值冲突的检测与处理

     - 4.1实体识别问题

- 问题：分别不同数据库中的customer_id可能和cust_number是同一属性
- 解决：利用元数据。每个属性的元数据包括名字、含义、数据类型和属性的值的允许范围，以及处理空值的规则

- 4.2属性冗余和相关性
  - 问题：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的属性导出（“派生”属性），即两个属性是相关的
  - 解决：
    - 标称数据的卡方检验
    - 数值数据的相关系数
    - 数值数据的协方差
- 4.3元祖重复

  create keys=>sort=>merge
- 4.4数据值冲突的检测与处理
  - 问题：来自不同数据源的属性值可能不同。如公制单位和英制单位等

5. 数据归约
   - 维度规约：减少随机变量或属性的个数
     - 属性选择
       - 逐步向前选择：从空集开始，逐步添加
       - 逐步向后选择：从整个属性集开始，逐步删除
       - 向前和向后选择结合
       - 决策树
       - 选择标准：
         - 信息增益（IG）（看PPT计算例题18、20页）
         - 互信息（MI）
     - 主成分分析（属性重构）
     - 扩展方法（LDA、NMF）
   - 数量规约：用替代的、较小的数据表示形式替换原数据
     - 参数化数据归约：回归模型
       - 因为回归模型可以用来近似给定的数据（异常值除外）
     - 非参数化数据归约：
       - 直方图
         - 用分箱来近似数据分布
       - 聚类
         - 同一簇内的对象“相似”
       - 抽样
         - 选择数据的一个代表性子集
         - 方法：
           - 简单随机抽样、无放回随机简单抽样、有放回…
           - 簇抽样
           - 分层抽样
           - SMOTE算法：看PPT51页

6. 数据变换与数据离散化
    - 数据变换策略：
        - 平滑：去除噪声。如分箱、聚类、回归
        - 属性重构：通过现有属性构造新的属性
        - 规范化：按比例缩放，使之落入特定区间
            - 最小-最大规范化
            - z-score规范化：即数据减去均值后再除以标准差（x'表示均值）
                - v' = $\frac{v-x'}{\sigma_A}$
                - 好处：该方法在实际的最小值和最大值未知时很有用，或者离群点主导了最小-最大值的标准化
            - 小数定标规范化
                - v' = $\frac{v}{10^j}$
                  其中j是使得max(|v'|) < 1的最小整数（往往是<=1）
        - 离散化：用区间标签或概念标签替换
            - 原因：
                - 易于组织成更高层次的概念
                - 某些模型只能使用离散属性，如决策树
            - 决策树
                - 结点：属性
                - 边：属性值
                - 叶子结点：类别
            - 解释：将连续属性的范围划分为区间
                - 区间标签可以用来替换实际的数据值
                - 通过离散化减少数据取值的个数
            - 做法：
                - 分箱：
                    - 方法
                        - 等频分箱（每个箱数据个数一样）
                        - 等宽分箱（如宽度为10，则0-10内的为一个箱，11-20的为…）
                    - 并不适用类信息，是非监督的离散化技术
                    - 对箱个数很敏感，也容易受离群点的影响
                - 聚类
                    - 将数学的值划分成簇
        - 概念分层：数据属性可以泛化得到较高的概念层
            - 收集低级概念（如城市）并替换为较高级概念（如省份），可以减少数据量
7. 小结