# 第五章关联规则

2021年11月14日　　20:33

1. 概述
2. 基本概念
   - 项集：项目的集合
   - 项集的长度：项集元素的个数
   - 交易（Transaction）：每笔交易T是项集I上的一个非空子集，即T⊆ I，但通常T ⊂ $I$。每个交易有一个唯一的标识— —交易号，记为TID
   - 支持度：
     - 绝对支持度（支持度计数）：count(X⊆T)为交易集中包含X的交易数量
     - 相对支持度：项集X出现的概率，描述了X的重要性

       support(X) = $\frac{count(X \subseteq T)}{|D|}$
   - 频繁集、非频繁集
   - 频繁闭项集
   - 极大频繁项集
   - k-频繁集，记为$L_k$
   - 关联规则（Association Rule）：可以表示一个蕴含式：

     R: X=>Y，其中X⊂I，Y⊂I，并且X∩Y=Ø
     - 表示若项集X在某一交易中出现，则会导致项集Y按照某一概率也会出现在同一交易中
   - 关联规则的支持度和置信度
     - 支持度：

       support(X=>Y) = $\frac{count(X \cup Y)}{|D|}$

       联合概率：p(Y, X)
     - 置信度：

       confidence(X=>Y) = $\frac{support(X \cup Y)}{support(X)} = \frac{count(X \cup Y)}{count(X)}$

       条件概率：p(Y|X)
   - 强关联规则、弱关联规则
3. 经典算法
   a. 关联规则的任务
      i. 频繁项集产生（Frequent Itemset Generation）
      ii. 规则的产生（Rule Generation）
   b. 经典算法：
      i. 原始方法：Brute-force approach：
         1) 计算每个可能规则的支持度和置信度
         2) 代价过高，候选规则数量达指数级
         3) 产生频繁项集的时间复杂度~O(NMW)（候选集长度M、交易数N、最大

项集长度W）

ii. 优化的方法：

1) 用先验原理减少候选集数量
2) 减少比较的次数

iii. Apriori算法

1) 先验原理：
   a) 如果一个项集是频繁的，则它的所有子集一定也是频繁的
   b) 相反，如果一个项集是非频繁的，则它的所有超集也一定是非频繁的
2) 这种基于支持度度量修剪指数搜索空间的策略称为基于支持度的剪枝（support-based pruning）
   a) 这种剪枝策略依赖于支持度度量的反单调性（anti-monotone）：即一个项集的支持度决不会超过它的子集的支持度。
3) Apriori找频繁项集做法：
   a) 找出频繁1项集，以L1表示
   b) 使用Lk-1找出Lk，由两步组成：
      i) 连接步：将Lk-1与自身连接产生候选k项集的集合，标记为Ck
      ii) 剪枝步：剪除不满足最小支持度的候选项集

iv. FP-growth算法

4. 评估关联规则
   - 支持度
   - 置信度
   - 提升度

$$lift(X,Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = P(X|Y)/P(Y)$$

- 当项集X的出现独立于项集Y的出现时，P(X∪Y)=P(X)P(Y)，即lift(X,Y)=1，表明X与Y无关；
- lift(X,Y) >1表明X与Y正相关；
- lift(X,Y) <1表明X与Y负相关