数据挖掘

实验二：Apriori算法

|  |  |
| --- | --- |
|  |  |
| 教 学 班： | 计算机科学与技术一班 |
| 学 号： | 19011402 |
| 姓 名： | 南佳霖 |
| 实验地点： | 理工楼913 |
| 指导教师： | 姜珊 |

2021

# 1. 实验目的

掌握Apriori算法的原理

# 2. 实验要求

1. 理解频繁项集和关联规则
2. 熟练使用Python或其他工具实现Apriori算法

# 3. 实验器材

1. 计算机一台
2. Python或其他编程工具

# 4. 实验内容

1. **背景：**

上世纪九十年代的美国沃尔玛超市中，管理人员在分析数据时发现一个**现象**：啤酒和尿布经常出现在同一条消费记录当中，经调查，这种现象经常出现在年轻父亲身上。

经分析**原因**后得知：在美国有婴儿的家庭中，一般是母亲在家中照顾婴儿，年轻的父亲去超市购买尿布，往往会顺便为自己购买啤酒。

“啤酒和尿布”是通过**人工观察分析**得到的一种关系，是否存在一种**数学算法**，从而高效地找到这种规则？

于是，1993年Agrawal通过分析购物车中的商品集合，提出了找出商品之间的关联关系的算法，即**Apriori算法**。并根据商品之间的关系，找出客户的购买行为。

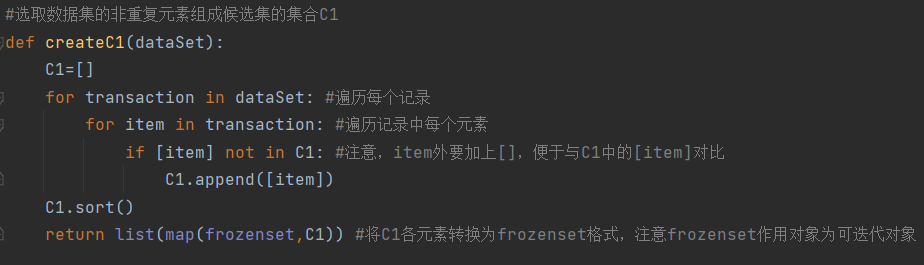
1. **关联分析相关基本概念有：**

* 关联分析：关联分析是在大规模数据集中有目的的寻找关系的任务。
* 关联分析要寻找的关系：频繁项集、关联规则。
* 支持度：数据集中包含该项集的记录所占的比例（定义为频率或概率）
* 置信度（可信度）：定义为条件概率。
* 提升度：某事件在另一事件基础上发生的概率与独立发生的概率比值
* 频繁项集：经常一起出现的项目的集合，定义为支持度大于某一阈值（最小支持度）的集合。
* 候选k项集：用来生成频繁k项集的k项集。
* 关联规则：置信度大于一定阈值（最小置信度）的关系。

1. **简述Apriori算法的原理：**
   1. 如果某个项集是频繁的，那么它的所有子集也是频繁的
   2. 如果某个项集是非频繁的，那么它的所有超集也是非频繁的
   3. 基于此，Apriori算法从单元素项集开始，通过组合满足最小支持度的项集来形成更大的集合
2. **Apriori算法的两个输入参数是数据集和最小支持度（阈值）。其流程如下：**

* 生成单个项的所有项集，遍历所有交易事件，筛选出单个项的频繁项集。
* 对于包含k个项的频繁项集，两两组合生成k+1项项集，删除非频繁项集，（可在此处进行预剪枝），获得k+1频繁项集直到算法收敛。
* 返回频繁项集。
* 将生成的频繁项集依次取出，同时计算其所有的真子集，然后以排列组合的方式形成关联规则。

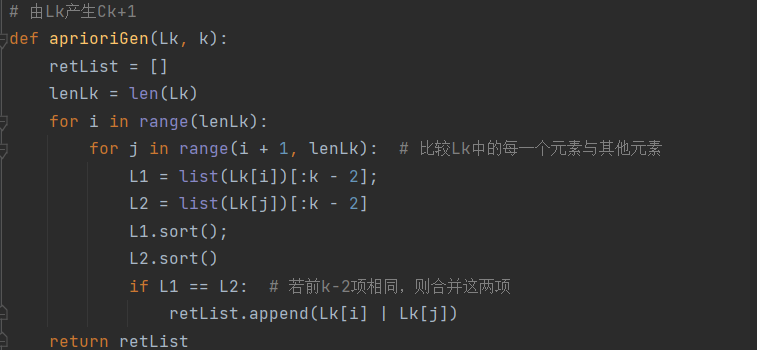
1. 生成单项候选集，**代码：**



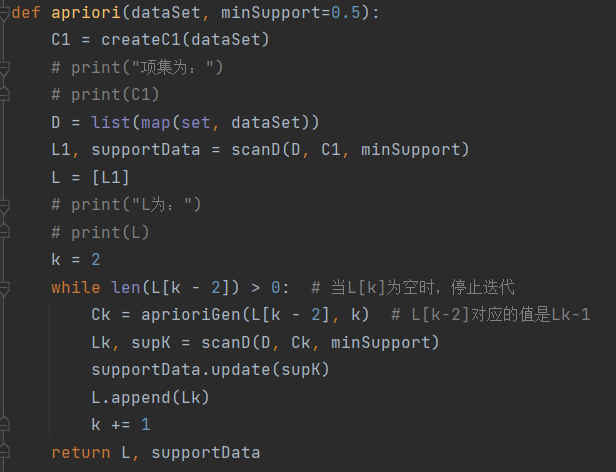
1. 计算候选集的支持度，选出k项频繁集，**代码：**



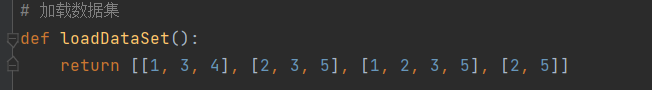
1. 由k项频繁集生成k+1项候选集，**代码：**



1. **生成频繁项集代码：**



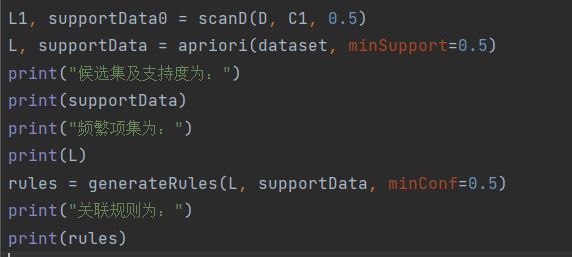
1. **输入数据集（交易事件）：**



**输入最小支持度阈值：**



**输出频繁项集结果：**



# 5. 实验心得

该算法可以很好的计算出关联规则，采用了逐层迭代的方法，易于实现。

但是时间复杂度较大，在面对大量样本的时侯，运行效率不高。

同时，该算法采用唯一支持度，没有考虑各个属性的重要程度并不相同。