实验四：关联规则及其应用

1. 案例背景

关联分析衡量的是两个商品共同被购买的可能性高低。这不同于分类或者回归，关联分析的目的不是预测某件商品是否会被购买，而是找出会被共同购买的一对商品，所以它属于无监督学习，即挖掘隐藏于数据中的模式，并以易懂的**关联规则**展示出来。

关联分析的算法广泛应用于各个领域，如零售店消费分析、推荐系统、网页在线点击分析等。其中，最流行的一类应用叫做**购物篮分析，**也就是找出那些商品会高频地与另一些商品同时被购买。为了深入了解关联规则的基本概念、表示法以及算法流程，我们将通过一个路边蔬菜摊的交易情况来进行购物篮分析的实验。具体的案例和目标如下：

假设某本地农场主建立了一个路边蔬菜摊，并提供如下销售商品：{芦笋（asparagus），豆类（beans），花椰菜（broccoli），玉米（corn），青辣椒（green peppers），南瓜（squash），西红柿（tomatoes）}。定义该项集为**I** . 客户购买的商品是**I** 的子集。假定表1 给出了在某个秋天下午路边蔬菜摊的交易情况。

1. 实验目的

本次上机实验有三个目的：

1. 了解关联规则在实际问题中的应用。
2. 掌握关联规则中Apriori 算法的核心。
3. 熟悉掌握对数据进行关联规则分析的流程和步骤。
4. 实验内容、方法与步骤：

本上机实验具体的方法与步骤如下：

1、购物篮分析的数据表示。购物篮的数据表示方法有两种：使用事物数据格式或者表数据格式。其中，采用表数据格式时，每条记录表示不同的事务，每个项采用0/1 标志表示。请将表1表示成表数据格式表2。

**表2：路边蔬菜摊的表数据格式**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 事务 | 芦笋 | 豆类 | 花椰菜 | 玉米 | 青辣椒 | 南瓜 | 西红柿 |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |  |
| 11 |  |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |  |
| 13 |  |  |  |  |  |  |  |
| 14 |  |  |  |  |  |  |  |

2、建立频繁项集。考虑表2的事务集合**D**. 这里设在D中出现次数超过4次的项是频繁项集，请给出频繁1-项集，频繁2-项集，频繁3-项集（要求**列表表示**所有频率计算过程和结果）.

3、 建立关联规则。利用频繁项集建立路边蔬菜摊中隐含的关联规则（支持度>50%，可信度>80%），并**列表表示**所有候选关联规则支持度和可信度的值。给出该案例挖掘的最终关联规则结果。

四、思考与实验总结

通过本上机实验，请简要说明关联规则中Apriori 算法的核心规则。