数据挖掘 ---- 作业二

程伟 2120160984 第10组

# 要求：

1 对数据集进行处理，转换成适合关联规则挖掘的形式；

2 找出频繁项集；

3 导出关联规则，计算其支持度和置信度

4 对规则进行评价，可使用Lift，也可以使用教材中所提及的其它指标

# 实验环境：

本实验程序基于python开发，使用的第三方库如下：

pandas、numpy-- 分析数据，并格式化存储数据

# 实验结果：

## 数据集Building\_Permits

### 1 对数据集进行处理，转换成适合关联规则挖掘的形式

考虑到数据集Building\_Permits中的属性非常多，且Apriori算法的时间开销太大，故只选择数据集中部分属性来完成本次作业。

筛选属性的依据：该属性所有可能取值的种类数大于等于2且小于等于10。

最终筛选出以下属性：['Permit Type', 'Permit Type Definition', 'Plansets', 'TIDF Compliance', 'Existing Construction Type', 'Existing Construction Type Description', 'Proposed Construction Type', 'Proposed Construction Type Description']。

为方便查看，将属性简化为由各单词首字母串联而成，则变成：['PT', 'PTD', 'P', 'TC', 'ECT', 'ECTD', 'PCT', 'PCTD']等属性。

关联数据挖掘是基于购物篮模型的，所以需要将源数据集中的各个对象的各个属性映射成唯一的“商品”。映射方式为：将原数据集中的各个元素(对应到一个对象的一个属性)取值拼接其属性名称作为该元素新的取值。但这种方式所产生的“交易记录”都是等长(意味着每个记录都包含同样数量的“商品”)的。

为解决上述问题，将源数据集中缺失数据的元素去除，从而得到不等长的“交易记录”，从而可以更好的模拟“购物篮模型”。

### 2 找出频繁项集

实现Apriori算法，可得到1到K项的频繁项集。将结果按支持度降序排列，其中前10项如下所示：

频繁项集 支持度

frequent\_item\_set support  
('PT\_8',) 0.899  
('PTD\_otc alterations permit',) 0.899  
('PTD\_otc alterations permit', 'PT\_8') 0.899  
('PCTD\_wood frame (5)',) 0.575  
('PCT\_5.0',) 0.575  
('PCT\_5.0', 'PCTD\_wood frame (5)') 0.575  
('ECT\_5.0',) 0.570  
('ECTD\_wood frame (5)',) 0.570  
('ECT\_5.0', 'ECTD\_wood frame (5)') 0.570  
('PCT\_5.0', 'ECT\_5.0') 0.563

所有的频繁项集保存在frequent\_items.txt文件中。

### 3 导出关联规则，计算其支持度和置信度

实现Apriori算法，可由第二步中得到的频繁项集导出关联规则。按置信度降序排列，其中前10项如下所示：

关联规则 支持度 置信度 lift

rule support confidence lift

('PCT\_5.0',) ==> ('PCTD\_wood frame (5)',) 0.575 1.000 1.739

('PCTD\_wood frame (5)',) ==> ('PCT\_5.0',) 0.575 1.000 1.739

('PTD\_otc alterations permit',) ==> ('PT\_8',) 0.899 1.000 1.112

('PT\_8',) ==> ('PTD\_otc alterations permit',) 0.899 1.000 1.112

('ECT\_5.0',) ==> ('ECTD\_wood frame (5)',) 0.570 1.000 1.755

('ECTD\_wood frame (5)',) ==> ('ECT\_5.0',) 0.570 1.000 1.755

('ECT\_5.0', 'PCTD\_wood frame (5)') ==> ('ECTD\_wood frame (5)',) 0.563 1.000 1.755

('ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') ==> ('ECT\_5.0',) 0.563 1.000 1.755

('P\_2.0', 'ECT\_5.0') ==> ('ECTD\_wood frame (5)',) 0.282 1.000 1.755

('P\_2.0', 'ECTD\_wood frame (5)') ==> ('ECT\_5.0',) 0.282 1.000 1.755

所有关联规则保存在association\_rules\_by\_confidence.txt文件中。

### 4 使用Lift对规则进行评价

根据Lift公式定义对每条规则打分，将结果按照Lift值降序排列，其中前10项如下所示：

关联规则 支持度 置信度 lift

rule support confidence lift

('PT\_8', 'ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') ==> ('PTD\_otc alterations permit', 'PCT\_5.0', 'ECT\_5.0') 0.508 1.000 1.969

('ECT\_5.0', 'PT\_8', 'PCTD\_wood frame (5)') ==> ('PTD\_otc alterations permit', 'PCT\_5.0', 'ECTD\_wood frame (5)') 0.508 1.000 1.969

('PCT\_5.0', 'PT\_8', 'ECTD\_wood frame (5)') ==> ('PTD\_otc alterations permit', 'ECT\_5.0', 'PCTD\_wood frame (5)') 0.508 1.000 1.969

('PCT\_5.0', 'PT\_8', 'ECT\_5.0') ==> ('PTD\_otc alterations permit', 'ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') 0.508 1.000 1.969

('PTD\_otc alterations permit', 'ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') ==> ('PCT\_5.0', 'PT\_8', 'ECT\_5.0') 0.508 1.000 1.969

('PTD\_otc alterations permit', 'ECT\_5.0', 'PCTD\_wood frame (5)') ==> ('PCT\_5.0', 'PT\_8', 'ECTD\_wood frame (5)') 0.508 1.000 1.969

('PTD\_otc alterations permit', 'PCT\_5.0', 'ECTD\_wood frame (5)') ==> ('ECT\_5.0', 'PT\_8', 'PCTD\_wood frame (5)') 0.508 1.000 1.969

('PTD\_otc alterations permit', 'PCT\_5.0', 'ECT\_5.0') ==> ('PT\_8', 'ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') 0.508 1.000 1.969

('P\_2.0', 'PT\_8', 'ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') ==> ('PTD\_otc alterations permit', 'PCT\_5.0', 'ECT\_5.0') 0.223 1.000 1.969

('ECT\_5.0', 'P\_2.0', 'PT\_8', 'PCTD\_wood frame (5)') ==> ('PTD\_otc alterations permit', 'PCT\_5.0', 'ECTD\_wood frame (5)') 0.223 1.000 1.969

所有结果保存在association\_rules\_by\_lift.txt文件中。