**数据挖掘-作业2**

王若琳 3120181040

**1 数据预处理**

Apriori算法是数据挖掘中的重要算法，它对于数据要求很严格。这里选取标称属性进行挖掘，丢弃数值属性。首先需要对数据集进行清洗，将其转换为适合挖掘的形式，重点处理缺失值。对于八个标称属性，将缺失值填补为设置好的空值。经过上述处理，数据集new\_winemag-data\_first150k.csv最终大小为150930\*8。

部分数据展示如下：

图片包含 文字, 照片

描述已自动生成

将得到的csv文件保存，用来做后续关联挖掘，保存为winemag-data\_first150k (1).csv文件。

**2 Apriori算法**

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。然后使用第1步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项，这里采用的是中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集，使用了递推的方法。

**2.1 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，其中前十项如下所示：

频繁项集 支持度

frequent\_item\_set support

('NoRegion\_2',) 0.596

('US',) 0.413

('NoDesignation',) 0.303

('California',) 0.295

('California', 'US') 0.295

('NoRegion\_1',) 0.166

('NoRegion\_1', 'NoRegion\_2') 0.166

('NoDesignation', 'NoRegion\_2') 0.160

('Italy',) 0.156

('Italy', 'NoRegion\_2') 0.156

其中所有频繁项集与关联度都保存在frequent\_items.txt文件中。

**2.2 导出关联规则，计算其支持度和置信度**

在Apriori算法中，根据上一步得到的频繁项集导出关联规则。按置信度降序排列，其中前十项如下所示：

关联规则 支持度 置信度 lift

rule support confidence lift

('Willamette Valley',) ==> ('Oregon',) 0.022 1.000 32.892

('Tuscany',) ==> ('NoRegion\_2',) 0.048 1.000 1.678

('Loire Valley',) ==> ('France',) 0.012 1.000 7.154

('Rioja',) ==> ('Spain',) 0.013 1.000 18.256

('Bordeaux',) ==> ('France',) 0.040 1.000 7.154

('Finger Lakes',) ==> ('US',) 0.010 1.000 2.419

('Marlborough',) ==> ('NoRegion\_2',) 0.011 1.000 1.678

('Tuscany',) ==> ('Italy',) 0.048 1.000 6.429

('Southwest France',) ==> ('France',) 0.011 1.000 7.154

('Napa',) ==> ('US',) 0.060 1.000 2.419

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

**2.3 对规则进行评价（使用Lift）**

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，其中前十项如下所示：

关联规则 支持度 置信度 lift

rule support confidence lift

(' Molinara"',) ==> ('"Corvina',) 0.011 1.000 89.739

('"Corvina',) ==> (' Molinara"',) 0.011 1.000 89.739

(' Molinara"',) ==> (' Rondinella', '"Corvina') 0.011 1.000 89.739

('"Corvina',) ==> (' Molinara"', ' Rondinella') 0.011 1.000 89.739

(' Molinara"', ' Rondinella') ==> ('"Corvina',) 0.011 1.000 89.739

(' Rondinella', '"Corvina') ==> (' Molinara"',) 0.011 1.000 89.739

(' Molinara"',) ==> ('"Corvina', 'Italy') 0.011 1.000 89.739

('"Corvina',) ==> (' Molinara"', 'Italy') 0.011 1.000 89.739

(' Molinara"', 'Italy') ==> ('"Corvina',) 0.011 1.000 89.739

('"Corvina', 'Italy') ==> (' Molinara"',) 0.011 1.000 89.739

其中所有关联规则把结果保存association\_rules\_by\_lift.txt文件中。