《数据挖掘》实验

班级 **计科11606**  姓名 张仁兵 成绩 \_\_\_\_\_\_\_

4. 实验四、关联规则挖掘与评测

4.0 实验内容与目的

（1）装载购物篮数据集supermarket，运行Apriori关联规则分析算法，挖掘关联规则。

（2）对生成的关联规则进行分析，从数据集中去除无关属性，再次运行Apriori算法，对生成的关联规则进行评估。

操作方法与要点如下：

4.1 装载购物篮数据集：

* 运行Weka：“开始-程序-Weka”，将启动Weka图形界面选择程序（GUI chooser）。
* 点击 “Explorer”打开主界面。
* Preprocess面板中，点击<Open file…>，选择Weka安装目录中data文件夹下的supermarket.arff文件。

4.2 查看数据集：

* 观察数据集的基本信息，回答：**该数据集有 \_\_4627\_\_\_ 条实例，\_\_\_217\_ 个属性。**
* 属性说明：除了最后一个属性total描述的是实例对应的消费水平之外，其他属性描述的都是实例的购买行为，即购买了哪些商品，或者购买了哪些部门的商品。因此，该数据集即可用于关联规则挖掘，也可用于分类学习。
* 逐个查看每个属性的取值：在Attributes版块中选择某个属性，在右侧selected attribute版块中将显示该属性的取值。回答：**在所有属性中，nominal类型的属性有\_ 217\_\_个，numeric类型的属性有\_\_0\_\_\_个。有两种取值的属性有\_\_\_1\_\_\_个，只有一种取值的属性有\_\_216\_个**。逐个查看每个属性的取值，是否存在取值为0的属性？取值为0的属性意味着数据集中所有实例在该属性上都没有取值。思考这样的属性对于后续学习有怎样的影响。

4.3 挖掘关联规则：

* 关联规则挖掘建立在项集（Item set）和频繁项集这两个概念的基础上。查看课件（P6-7），将项（Item）、项集、频繁项集的定义翻译成中文。

项是一个测试、属性值对。

项集是一组项的集合，通常指规则中发生的所有项。

频繁项集:所有满足最小支持度的项集

* 在项集基础上，关联规则挖掘围绕两个核心概念展开，即支持度（Support）和置信度（Confidence）。查看课件（P6-7），将支持度和置信度的计算公式翻译成中文。

支持度：表示同时包含A和B的事务占所有事务的比例。

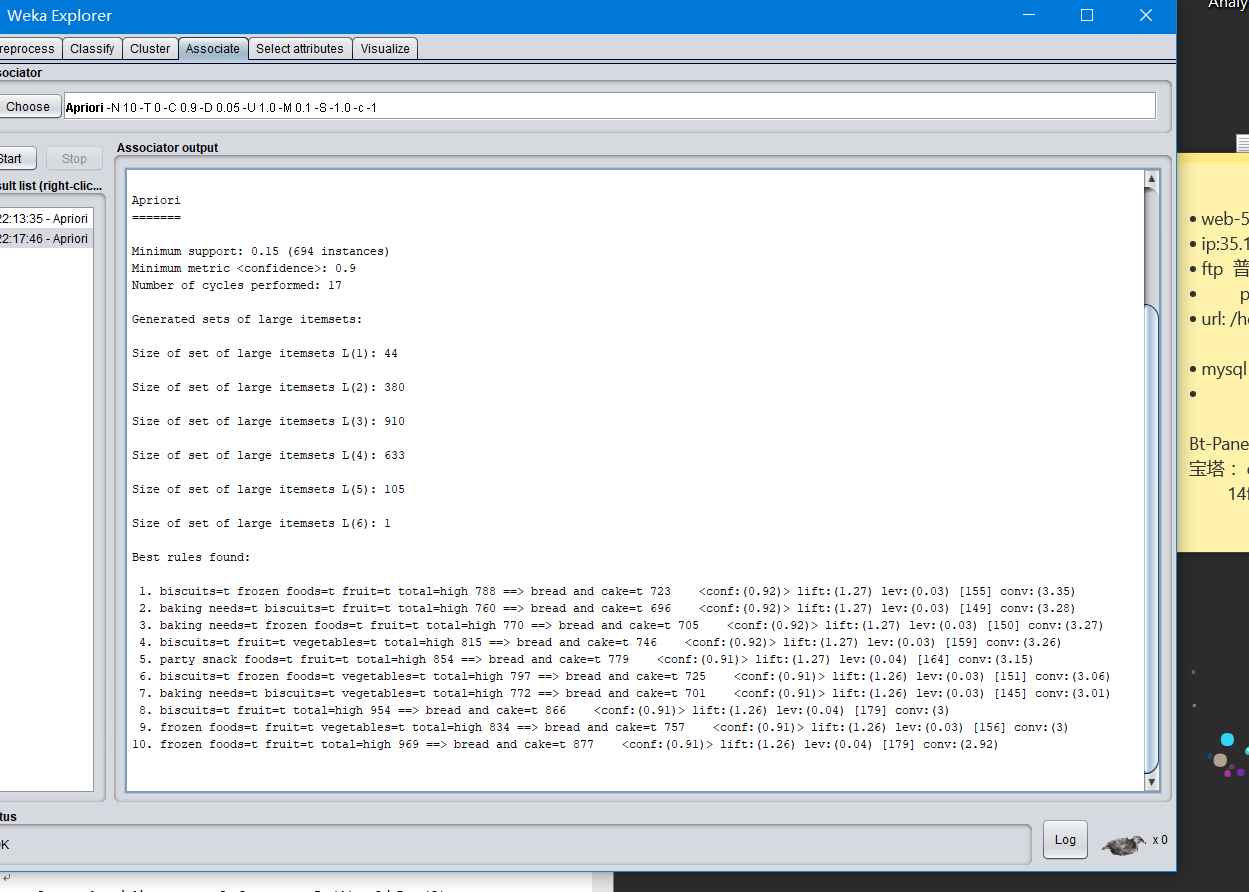
置信度：表示使用包含A的事务中同时包含B事务的比例，即同时包含A和B事务占A事务的比例 。

* 根据支持度和置信度的计算公式，计算规则potatos=》beef的支持度和置信度，列出计算公式。（提示：先用Remove按钮删除除potatoes和beef外的其他属性，再打开数据集编辑界面，点击首行属性名称，使数据按属性取值排序，再数出在这两个属性上取值不为零的实例数）

支持度：368/（1739+734）=0.1488

置信度：368/1739=0.2116

* 点击Associate面板，观察该面板与Classify和Cluster面板有何不同。
* 运行Apriori挖掘算法：直接运行Apriori挖掘算法，将以置信度（Confidence）作为关联规则的评价标准，输出排名前10的关联规则。
* **将这些关联规则截图粘贴在下面。**



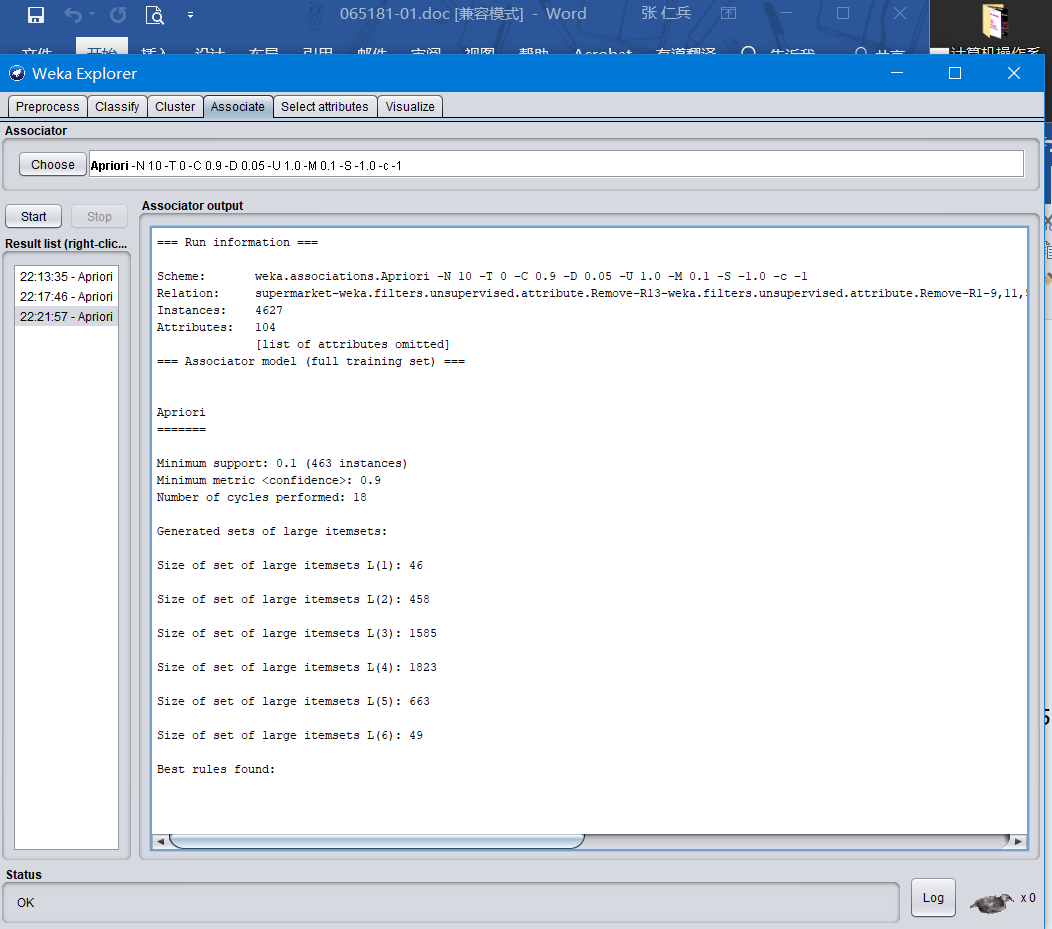
4.4 去除无关属性

* 从前面的挖掘结果中可以看到，排名前十的关联规则都跟bread and cake这个属性有关。这体现出西方的购买行为模式：基本每次购买中都会包含这个商品。然而，对于关联规则挖掘而言，这样的属性是没有太大价值的。因此，我们需要从原始数据集中去除这样的无关属性，使真正有价值的关联规则能被发现。
* 删除bread and cake属性：返回Preprocess面板，在Attributes版块中选中bread and cake，点击下方的Remove按钮，删除该属性。注意，可以用上方的Undo按钮撤销前一步操作。
* 删除其他无关属性：假设我们的目标是挖掘具体商品之间的关联关系，那么除了bread and cake属性之外，为数众多的department属性（即商场部门信息）也与当前的挖掘任务无关。如何批量删除属性呢？可以通过正则表达式匹配多个属性。点击Pattern按钮，在弹出的对话框中输入*department[0-9]+*，点击<确定>。该正则表达式将匹配名字以department开头、后接多个数字的属性。检查此时被选中的属性是否都符合该特征。点击Remove按钮，同时还可以删除total属性。

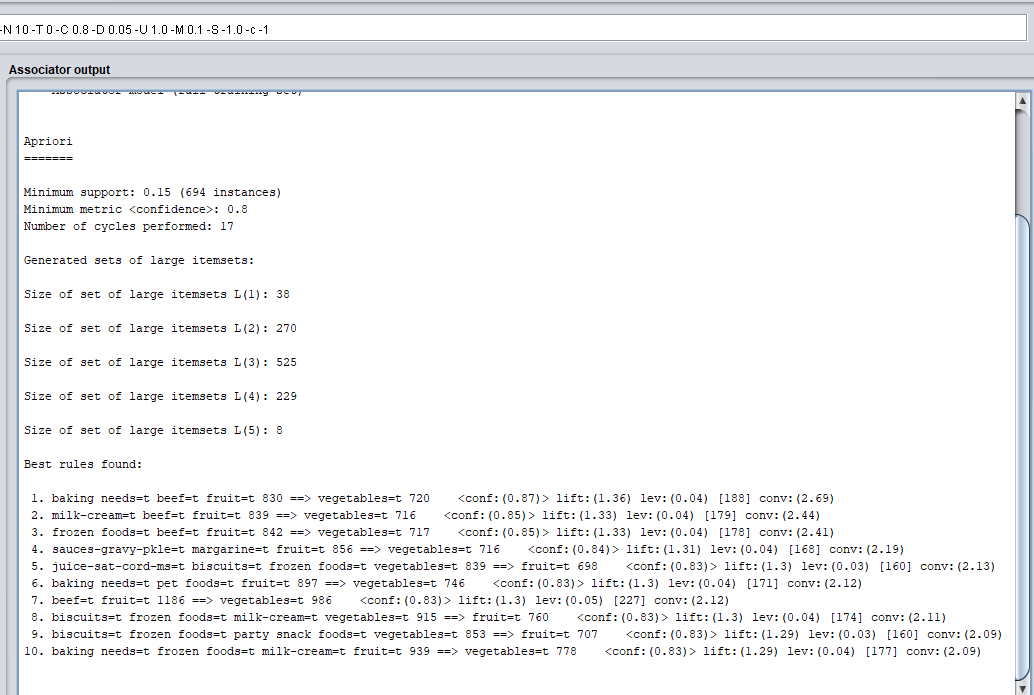
4.5 配置关联规则挖掘算法

* 点击Associate面板，运行Apriori算法，此时得到的关联规则有多少条

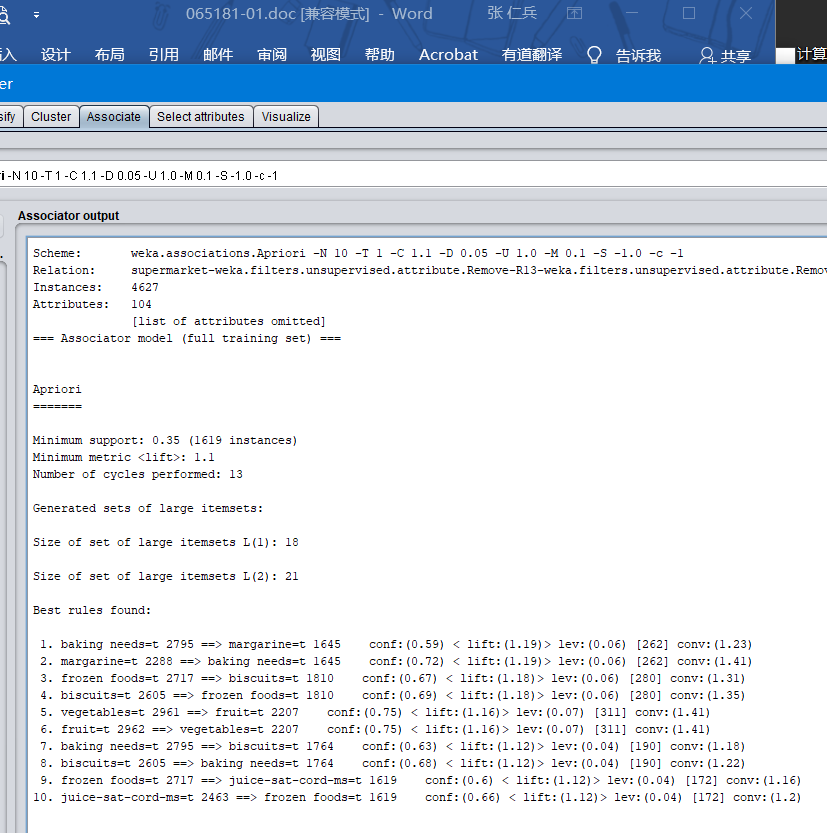
此时关联规则有6条，如下图所示



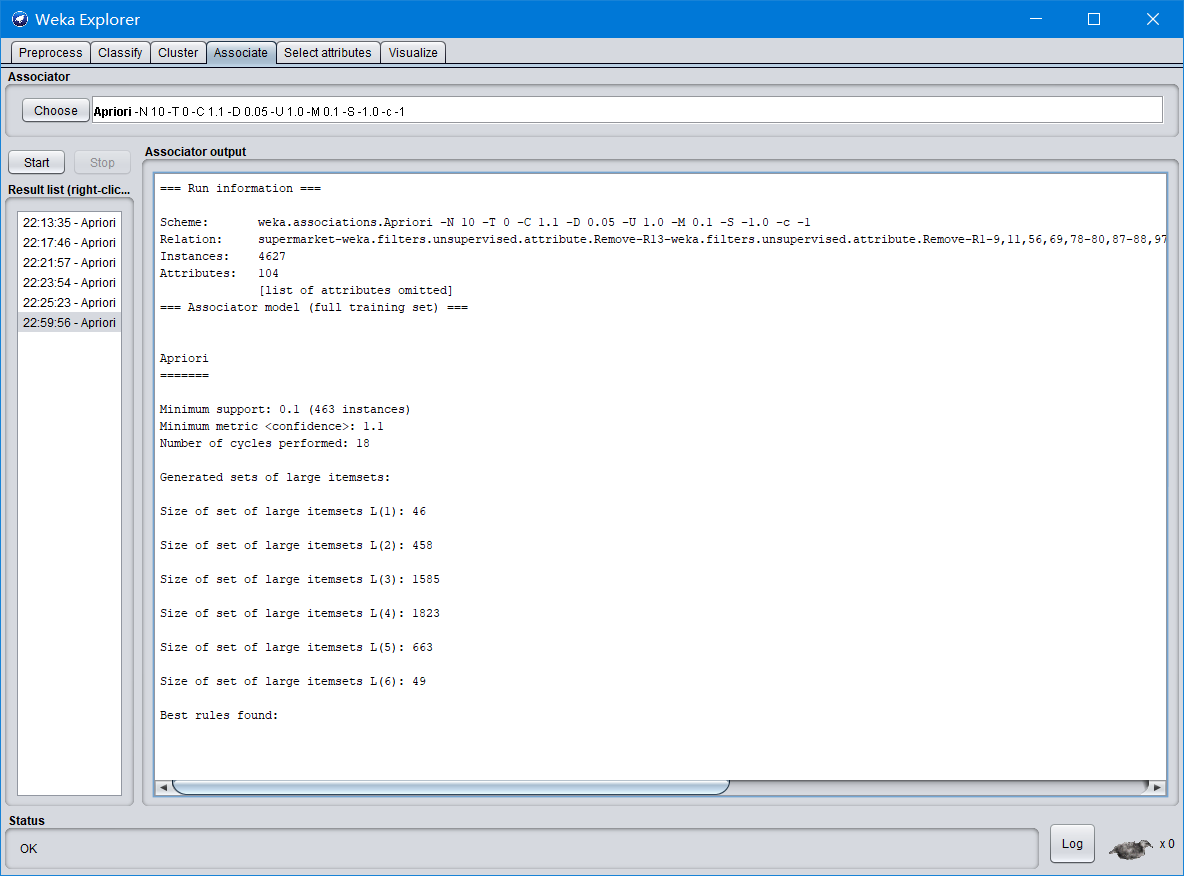
* 将置信度的最小值阈值（即minMetric）调整为0.8，再次运行Apriori算法，此时得到的关联规则是哪些？**截图并粘贴到下面。**



* 将关联规则评估方法换成提升（即将MetricType设置为Lift），再次运行Apriori算法，此时得到的关联规则是哪些？**截图并粘贴到下面。**



* 尝试不同的关联规则评估方法和阈值，**将你认为最有意义的关联规则截图粘贴到下面，并说明获取该关联规则的参数设置。**



metricType:confidence

minMteric:1.1

