实验2：基于WEKA的关联规则挖掘

**1 实验目的**

1. 掌握WEKA平台挖掘关联规则的方法。

**2 实验平台与工具**

1. Windows、Linux操作系统
2. 数据分析与挖掘系统WEKA

**3 实验内容**

1. （必做）自主设定数据挖掘的应用，采用WEKA平台完成整个挖掘过程，具体要求如下：
   1. 解释说明该挖掘应用的背景、动机和需求；
   2. 自主选择相应的数据集，数据规模不需要太大；
   3. 对于挖掘所得的关联规则具有可解释性，请解释所得的规则；
2. （选做，加分题）基于WEKA平台挖掘关联规则的挖掘性能测试与分析
   1. 随数据规模变化后的挖掘算法性能变化分析；
   2. 相同数据规模，不同支持数（或者支持度）变化后的挖掘算法性能变化分析；
   3. 相同数据规模和支持数（或者支持度），不同置信度变化后的挖掘算法性能变化分析；

**4. 规则与要求**

（1）独立完成，严禁相互抄袭（如有发现抄袭和被抄袭均判为0分），以及从网络上直接摘抄别人的观点和总结（该行为将影响报告成绩）。

（2）实验报告符合学术写作的排版要求，请参考群文件中的“报告模板.docx”和“参考文献格式.docx”的排版格式。

（3）实验报告内容详实，采用图文混合的方式叙述安装和配置过程。

Tip：Win+Shift+S 在Windows中可以快速截屏。

（4）报告文件见附件，提交报告时请以附件形式插入到超星作业中。

实验报告

报告标题：基于WEKA的关联规则挖掘

学号：21190630

姓名：黄艺杰

日期：4月13日

# 一、实验环境

1. 操作系统：Windows10

2. 软件（含版本号）：weka-3-8-6-azul-zulu-windows

# 二、实验内容及其完成情况

（1）挖掘应用的背景、动机和需求

背景：关联规则挖掘可以广泛应用于农业、医学和金融等领域，帮助我们更准确、高效地进行决策和管理。利用关联规则对大豆疾病分类特征数据挖掘分析，可以发现黄豆不同属性间的相关性，探索影响其特征的因素，进而为问题提供更好的解决方案。

动机：利用soybean.arff这个数据集来做关联规则挖掘是为了实现对豆类病害的诊断、预测和治疗，以及对豆类生长环境和特征的优化，满足农业生产管理和决策的需求。

需求：可以对关联规则进行分析，发现不同品种之间的相似性，对黄豆的品种进行分类和预测，还可通过对农田监测数据的采集和关联分析，对黄豆的生长情况和疾病抵抗力进行评估和预测，以实现精准农业管理和提高产量。此外，利用关联规则分析可以帮助农民和植物学家更好地了解某些疾病的病因和发展规律，提供更好的治疗方案。

（2）数据集介绍

本实验选用的是一个公开的、用于机器学习和数据挖掘的数据集，主要用于研究黄豆疾病的分类和识别。该数据集包含了683条记录，有35个属性和19个类别，并且经过了人工处理以保证其质量。使用该数据集可以帮助研究者更好地了解和探索黄豆领域的问题和现象，比如疾病的发生规律和根源、不同品种之间的差异等等。

1. 实验结果的解释

使用weka平台上的Apriori算法对soybean.arff数据集进行关联规则挖掘的结果。结果显示，置信度(confidence)要求为0.9，最小支持度(minimum support)为0.8，一共执行了4个循环。

在执行过程中，找到了10组最优规则。其中，1-6号规则与豆类疾病相关性较强，7-10号规则更加注重豆类生长环境和特征之间的相关性。1号规则置信度最高，为1；10号置信度最低，为0.96。

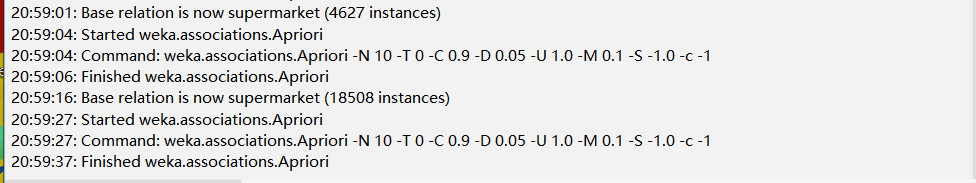
规则1和规则2说明，如果病变黄豆的颜色没有发生变化，则该黄豆不会出现硬块菌菌核。规则3说明，如果黄豆有异常叶片的情况且不存在硬块菌菌核，那么可能没有白粉病菌根系的存在。规则4-6说明，当病变黄豆缺乏菌核或者没有出现菌核时，则很少会出现菌丝体。规则8说明，如果黄豆没有发生叶片畸形，则很少会出现菌丝体。

这些规则可以用于豆类疾病的预测、诊断和治疗，同时也可以用于农业生产的管理和决策。例如，根据规则5可以判断是否需要给黄豆喷洒杀菌剂，规则9可以根据豆类的生长环境和特征来优化豆类的种植和管理决策。



1. 挖掘性能测试与分析
2. 数据规模变化:

使用同一个数据集，除了数据量不同，一个是4627条数据，一个是18508条数据，其他参数保持一致，结果数据量较小的花费2秒完成，数据量大的花费10秒完成。

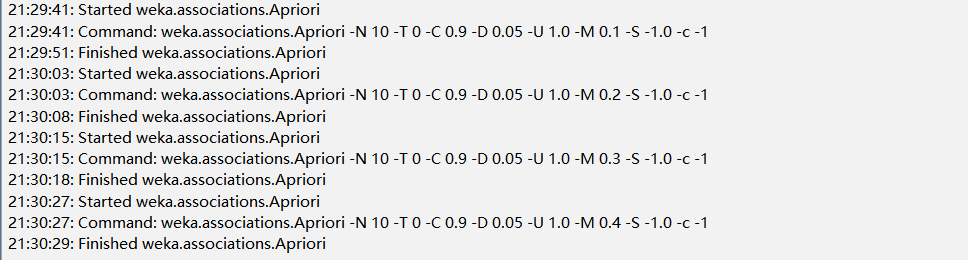


在数据量较小的情况下，算法能够很容易地找到符合要求的项集，但是在数据量增大时，算法执行时间可能会因为计算复杂度增加而变得非常慢。

1. 支持数（或者支持度）变化:

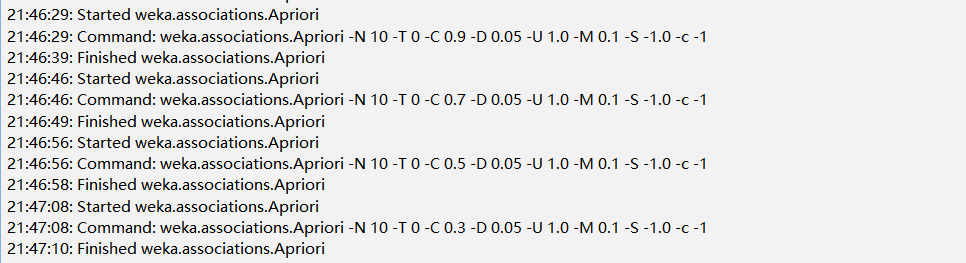
保持其他参数不变，分别将最小支持度设置为0.1-1.0，0.2-1.0，0.3-1.0，0.4-1.0，它们所花费时间分别为10秒、5秒、3秒、2秒，可以看出当最小支持度变大时，消耗时间较为显著降低。

当固定数据规模时，提高支持度设置会降低生成的频繁项集数量，从而减少了挖掘算法所需的计算量和内存需求，因此可以提高挖掘算法的运行速度。



1. 置信度变化：

控制其他参数不变，分别将最小置信度设置为0.9，0.7，0.5，0.3，结果它们花费的时间分别是10秒、3秒、2秒、2秒。



当置信度较高时，虽然产生的频繁项集和关联规则较少，但它们的质量更高，算法执行时间可能会相应增加。

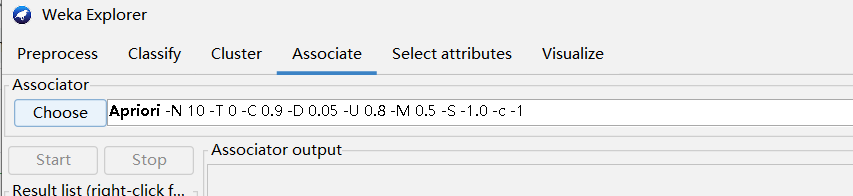
# 三、实验总结

（可以总结实验中出现的问题以及解决的思路，也可以列出没有解决的问题）

问题1：

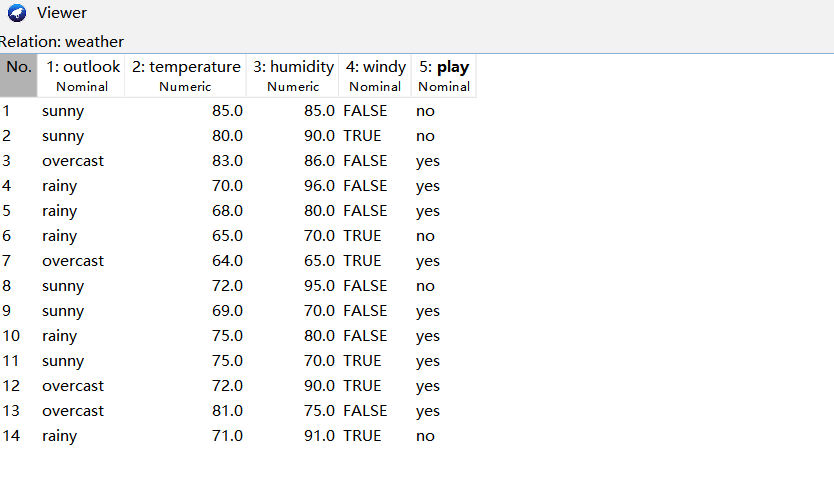
1. 问题描述

部分数据集在使用Apriori算法进行关联规则挖掘时，Associate里的Start按钮为灰色，如下图所示：



1. 问题分析（可能的原因、难点、挑战）

通过观察该数据集发现里面的数据都是具体的值，是连续的，Apriori算法主要用于处理离散数据集，不适用于连续型数据集

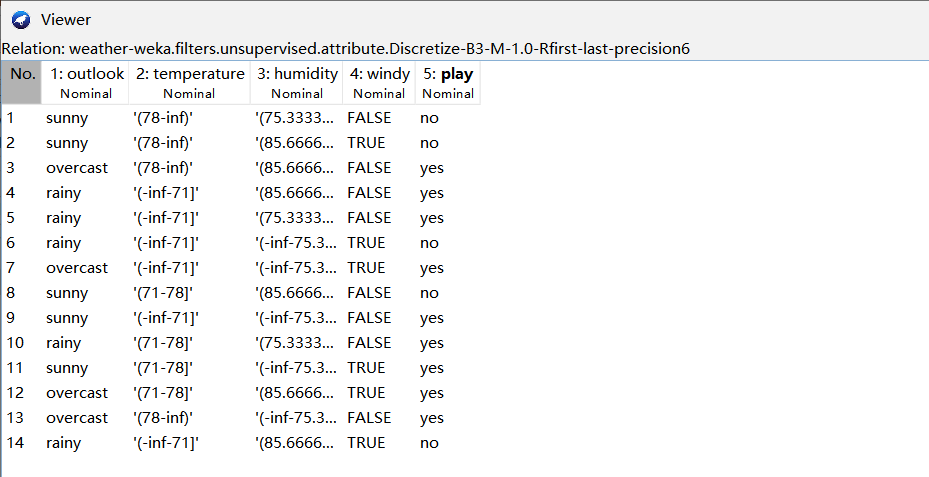


1. 解决方案

（如该问题没有被解决可以不写具体的解决方案，而给出大致的思路和方向）

可以选择其他关联规则挖掘算法或者将数据离散化处理，离散化处理过程如下：

选择要处理的文件，打开weka.filter.unsupervised.attribute.Discretize，设置参数，将bins属性设为3，最后apply，得到如下所示：



最后将相关属性区间替换为想要的新命名，得到新的数据集：

