《机器学习》读书笔记

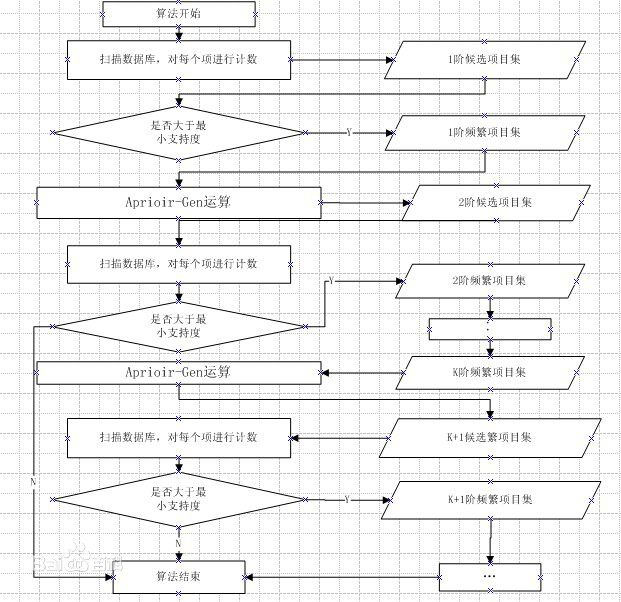
1. 算法概述（算法名称及原理）

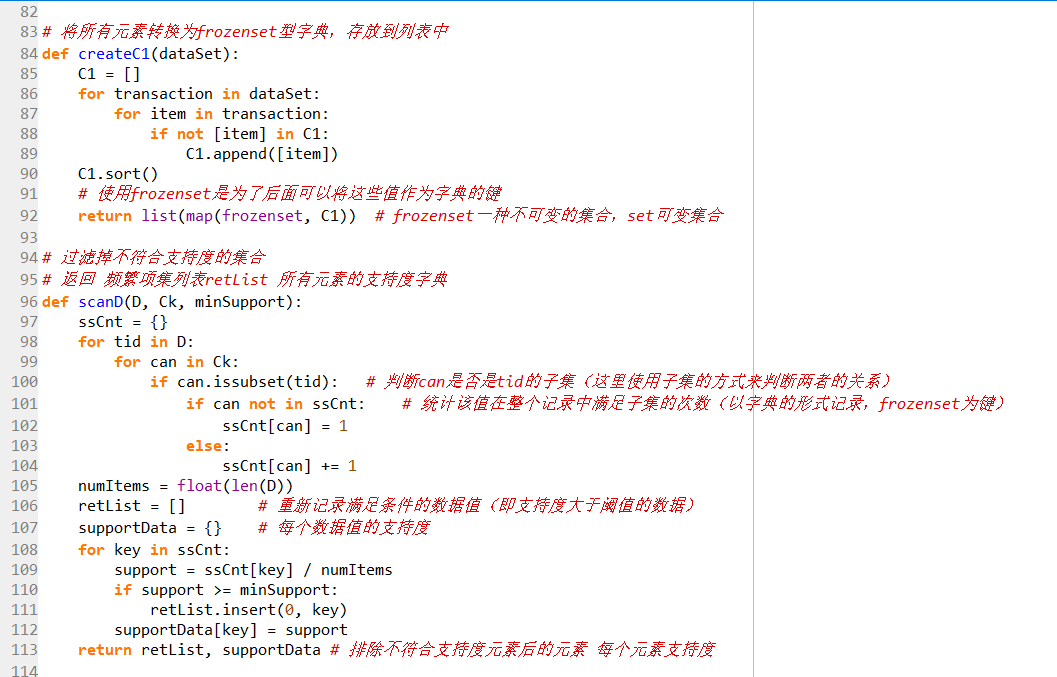
关联规则的分类规则：大于最小支持度阈值和最小置信度阈值的关联规则称为强关联规则。关联分析的最终目的是找出这些强关联规则。

Apriori算法是一种挖掘关联规则的频繁项集算法，其核心思想是通过候选集生成和情节的向下封闭检测两个阶段来挖掘频繁项集。该算法已经被广泛的应用到商业、网络安全等各个领域。

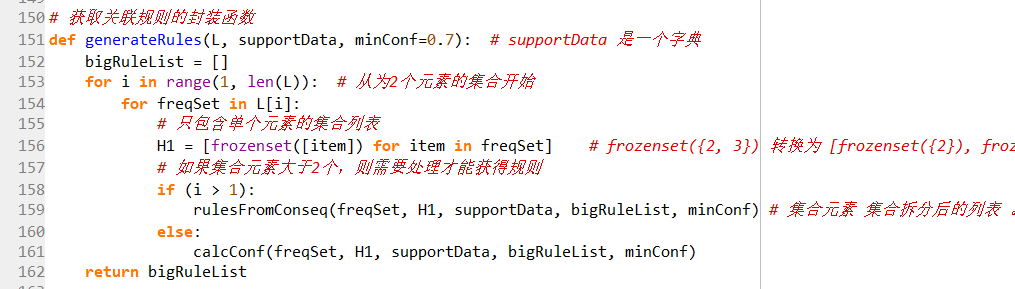
Apriori的性质：如果某个项集是频繁项集，那么它所有的子集也是频繁的。即如果 {0,1} 是频繁的，那么 {0}, {1} 也一定是频繁的。也就是说如果一个项集是非频繁的，那么它的所有超集也是非频繁的。由此可以减少候选集的数量，但是这个算法仍然无法避免对大量的候选项集进行频繁性检验，并且要重复地扫描数据集，从而使得其效率下降。

1. 算法设计（流程图及主要分段代码，附详细代码注释）



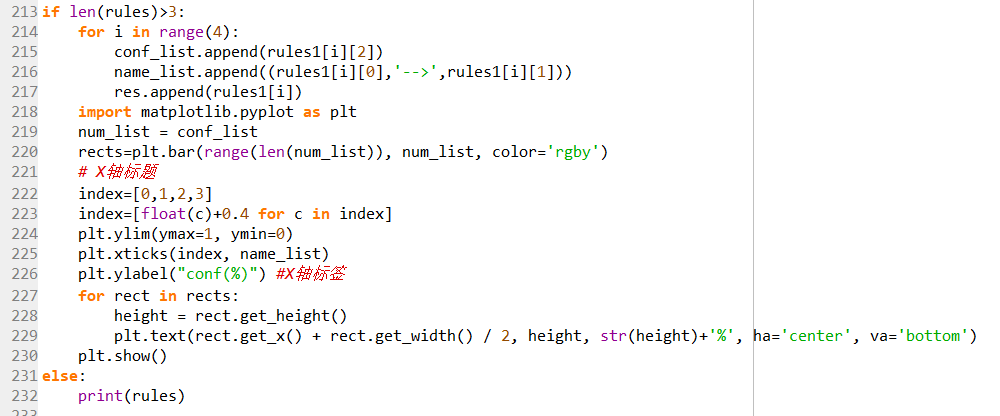












1. 选用数据（数据集描述，包括来源，行数，列数，格式等）

iris行数：150 列数：5

列属性及取值:

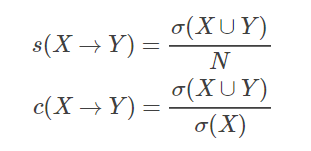
1. 萼片长度cm，数值型
2. 萼片宽度cm，数值型
3. 花瓣长度cm，数值型
4. 花瓣宽度cm数值型

类别：   
- Iris Setosa   
- Iris Versicolour   
- Iris Virginica

car，行数：1728，列数：6  
列属性及取值：  
1)  buying: vhigh, high, med, low.   
2)  maint: vhigh, high, med, low.   
3)  doors: 2, 3, 4, 5more.   
4)  persons: 2, 4, more.   
5)  lug\_boot: small, med, big.   
6)  safety: low, med, high.  
类别：  
unacc, acc, good, vgood   
  
wine，行数：178，列数：13  
属性：  
1) Alcohol   
2) Malic acid   
3) Ash   
4) Alcalinity of ash   
5) Magnesium   
6) Total phenols   
7) Flavanoids   
8) Nonflavanoid phenols   
9) Proanthocyanins   
10) Color intensity   
11) Hue   
12) OD280/OD315 of diluted wines   
13) Proline  
类别：  
Alcohol 1，2，3

1. 评价方法（说明训练集和测试集分配方法及评价指标）

支持度和置信度计算公式：

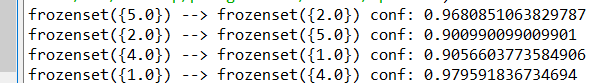


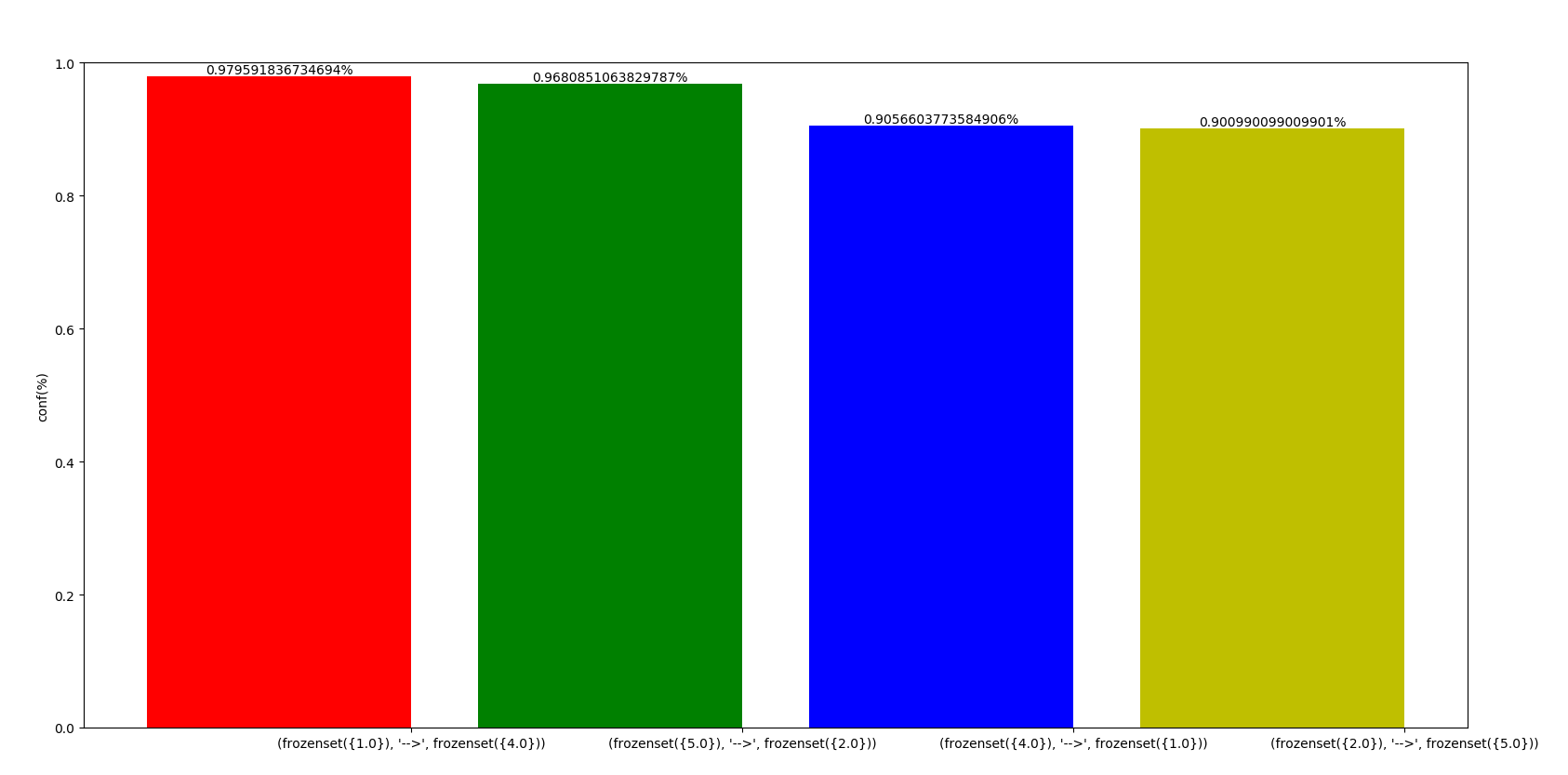
支持度反映的是关联规则得出的规则的普遍性。

置信度反映了规则的可靠性。

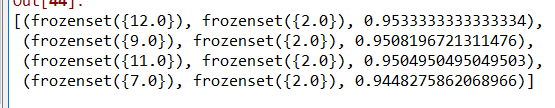
1. 实验结果截图

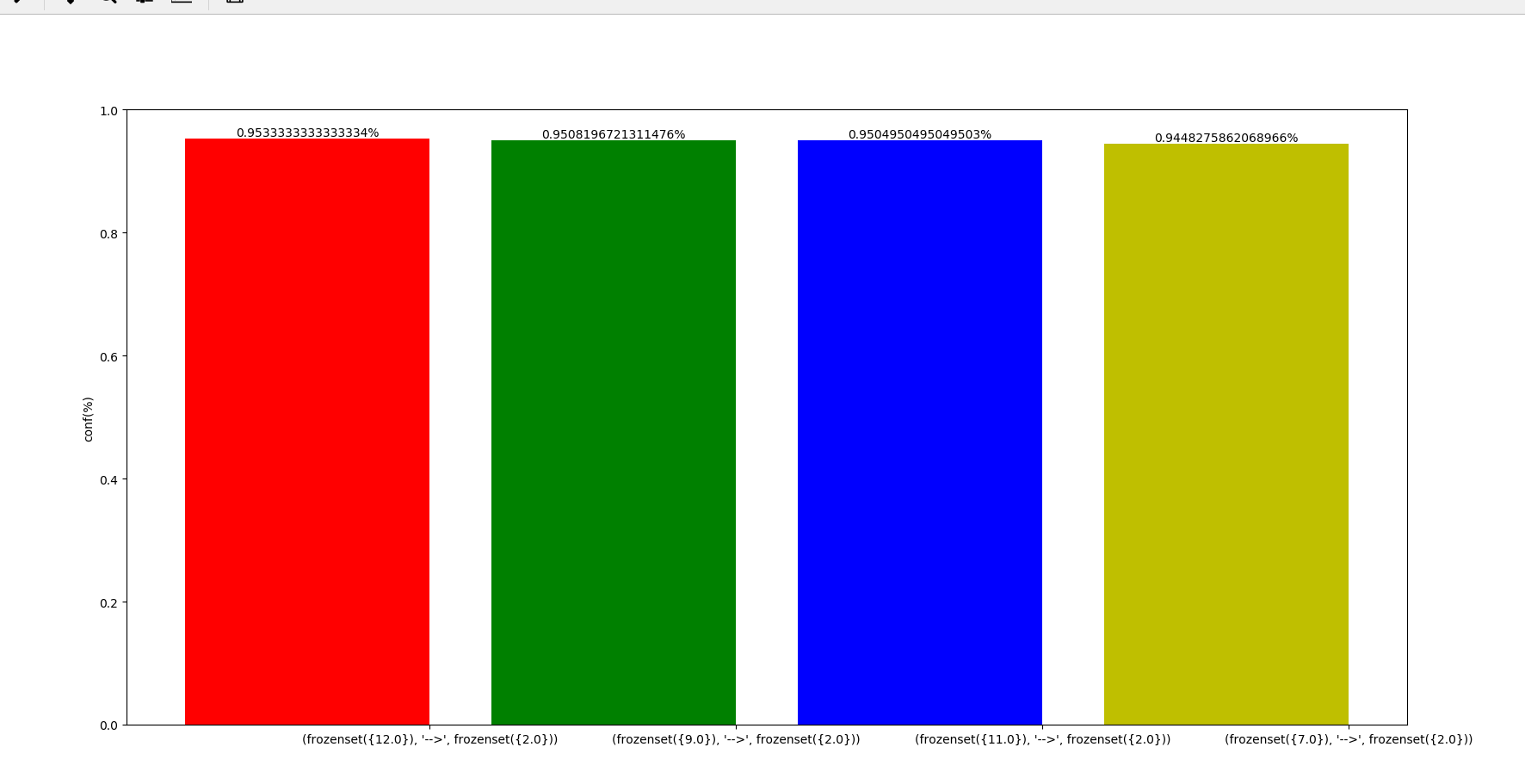
Iris：



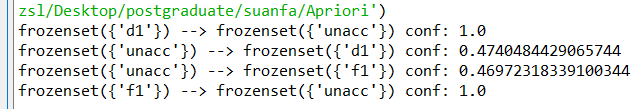


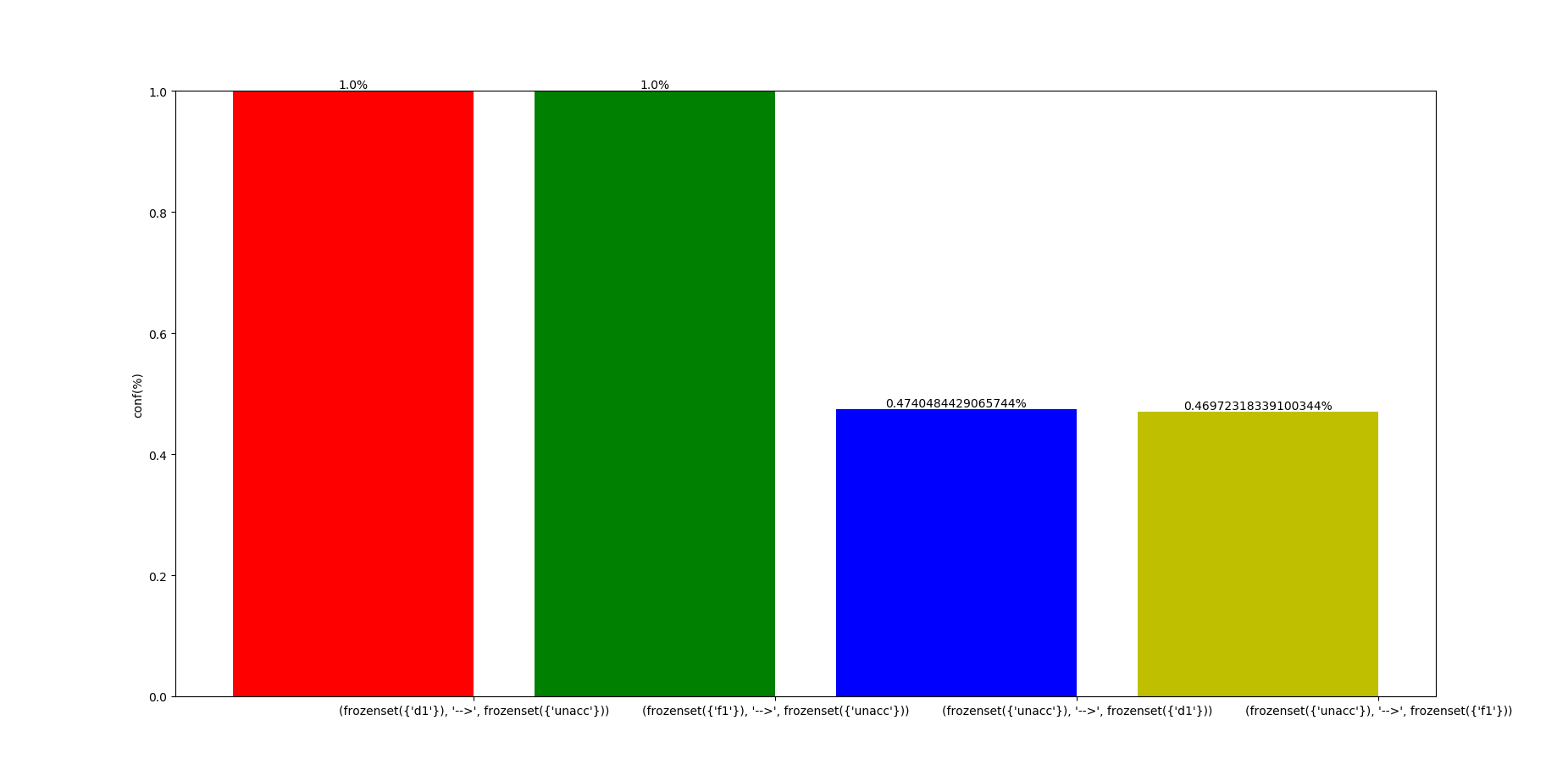
Wine：





Car：





1. 实验结果分析及比较

对于iris和wine来说，都能得到强关联规则

对于car来说，需要降低支持度和置信度的要求才能得出结果，可能本身不大适合使用这个算法，结果也不是很理想。

1. 遇到的问题及解决方法，实践心得

在做car转换的时候本身为离散型的。其他两个数据包含了连续型的，因此要先做离散化处理，再使用算法。