数据挖掘引论实验

何长鸿 2016141482154

1 数据准备

从http://archive.ics.uci.edu/ml/datasets/Wine下载数据集Wine,分别有wine.data和wine.names两个文件。根据arff格式语法，编辑整合得到wine.arff文件，部分内容如图1

```
@RELATION wine

@ATTRIBUTE class                        {1,2,3}
@ATTRIBUTE Alcohol                      REAL
@ATTRIBUTE Malic-acid                   REAL
@ATTRIBUTE Ash                          REAL
@ATTRIBUTE Alcalinity-of-ash            REAL
@ATTRIBUTE Magnesium                    REAL
@ATTRIBUTE Total-phenols                REAL
@ATTRIBUTE Flavanoids                   REAL
@ATTRIBUTE Nonflavanoid-phenols         REAL
@ATTRIBUTE Proanthocyanins              REAL
@ATTRIBUTE Color-intensity              REAL
@ATTRIBUTE Hue                          REAL
@ATTRIBUTE OD280/OD315-of-diluted-wines REAL
@ATTRIBUTE Proline                      REAL

@DATA
1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
1,14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
1,14.06,2.15,2.61,17.6,121,2.6,2.51,.31,1.25,5.05,1.06,3.58,1295
1,14.83,1.64,2.17,14,97,2.8,2.98,.29,1.98,5.2,1.08,2.85,1045
1,13.86,1.35,2.27,16,98,2.98,3.15,.22,1.85,7.22,1.01,3.55,1045
1,14.1,2.16,2.3,18,105,2.95,3.32,.22,2.38,5.75,1.25,3.17,1510
1,14.12,1.48,2.32,16.8,95,2.2,2.43,.26,1.57,5,1.17,2.82,1280
1,13.75,1.73,2.41,16,89,2.6,2.76,.29,1.81,5.6,1.15,2.9,1320
```

图 1: wine.arff数据内容

2 预处理

通过weka预处理页面观察数据集特征，可以看到红酒三个分类的数目分别为59,71,48。再观察变量与分类的相关可视化，如图2,可以看到flavanoids、Hue等多个变量大小与红酒类别有明显的分层关系。 此外，各组变量均没有缺失值。

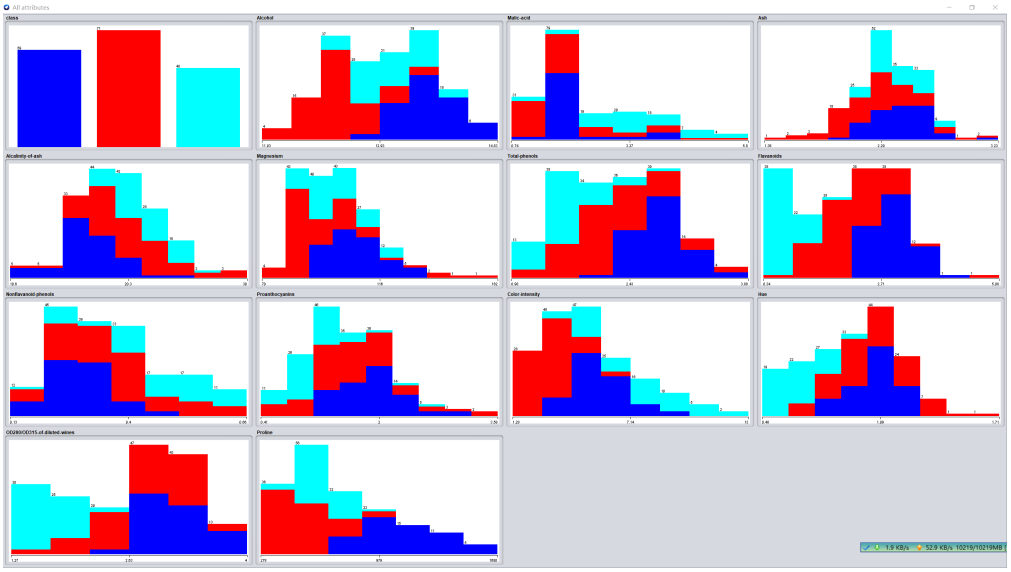

图 2: 数据分布

3 分类

先使用10-folds交叉验证测试模型，以及多种分类模型进行试验，得到多个预测结果，如图3、4分别为贝叶斯网和J48两种算法的运行结果，从混淆矩阵和各种分类评价指标来看，贝叶斯网取得了最好的分类结果。 对以上两种模型结构进行可视化，得到如下图5、6

4 数据可视化

通过wekaVisualize标签页面观察数据变量之间的相关关系，这里只选取了包括class在内的五个变量进行观察，如图7

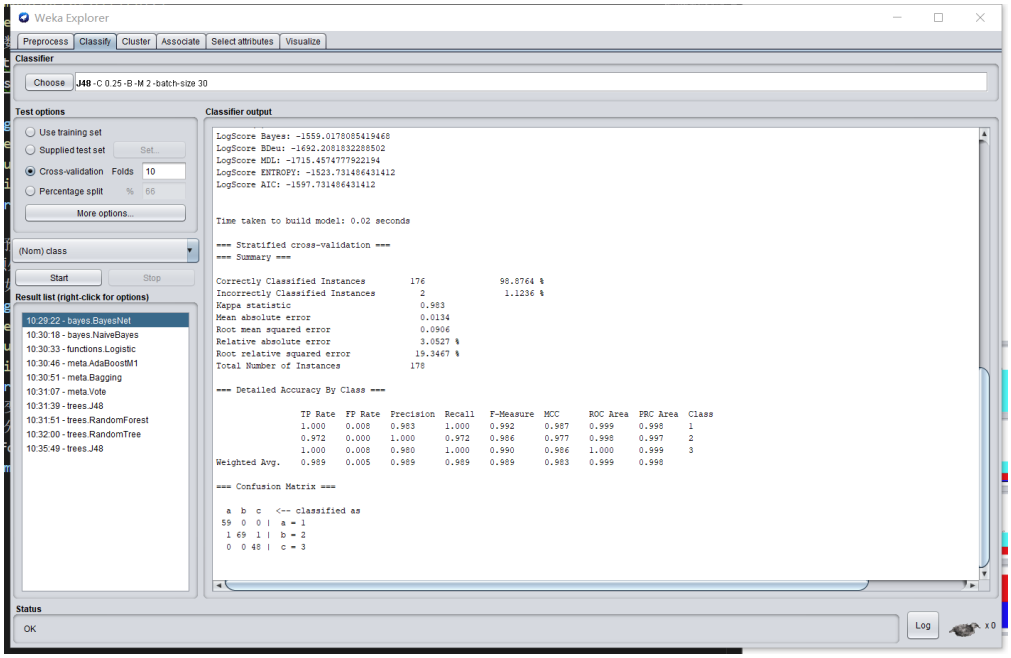

图 3: 贝叶斯网

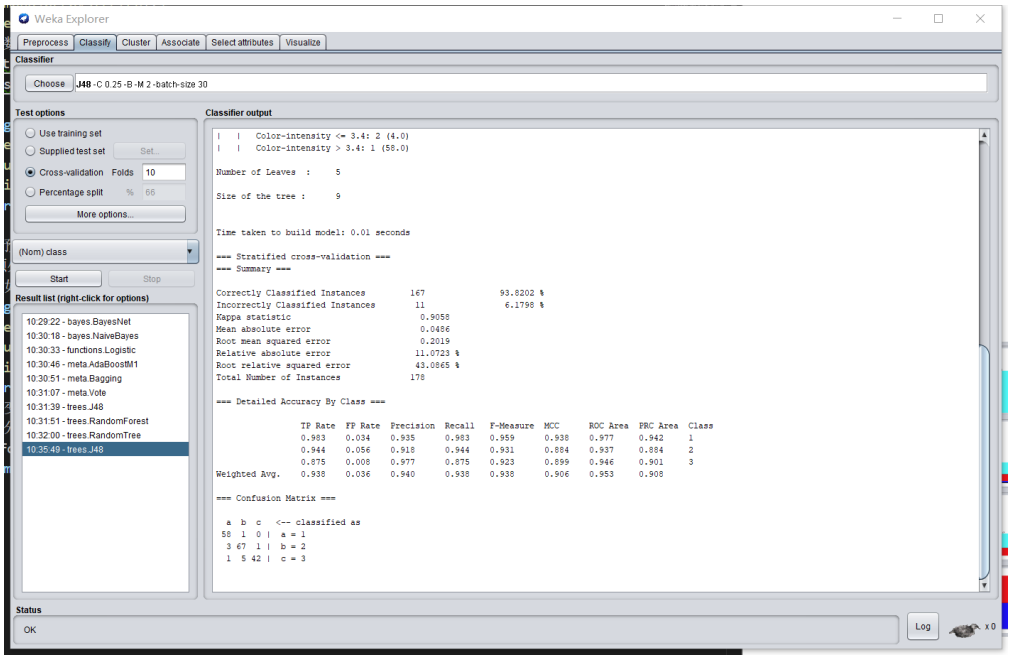

图 4: J48

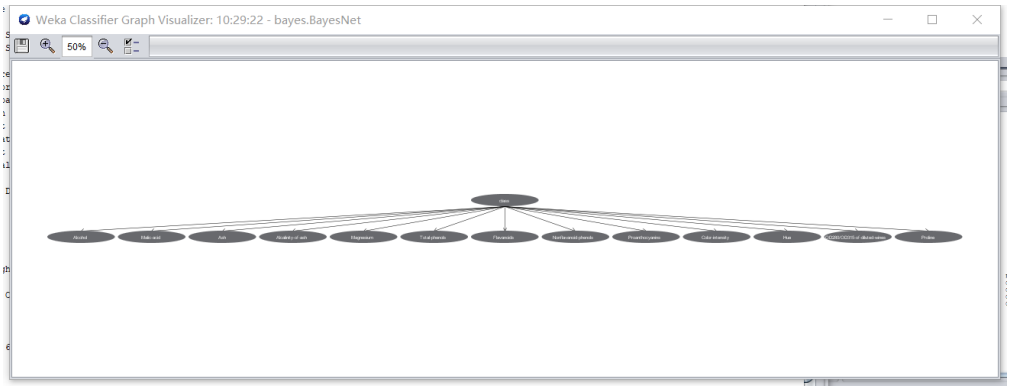

图 5: 贝叶斯网模型结构

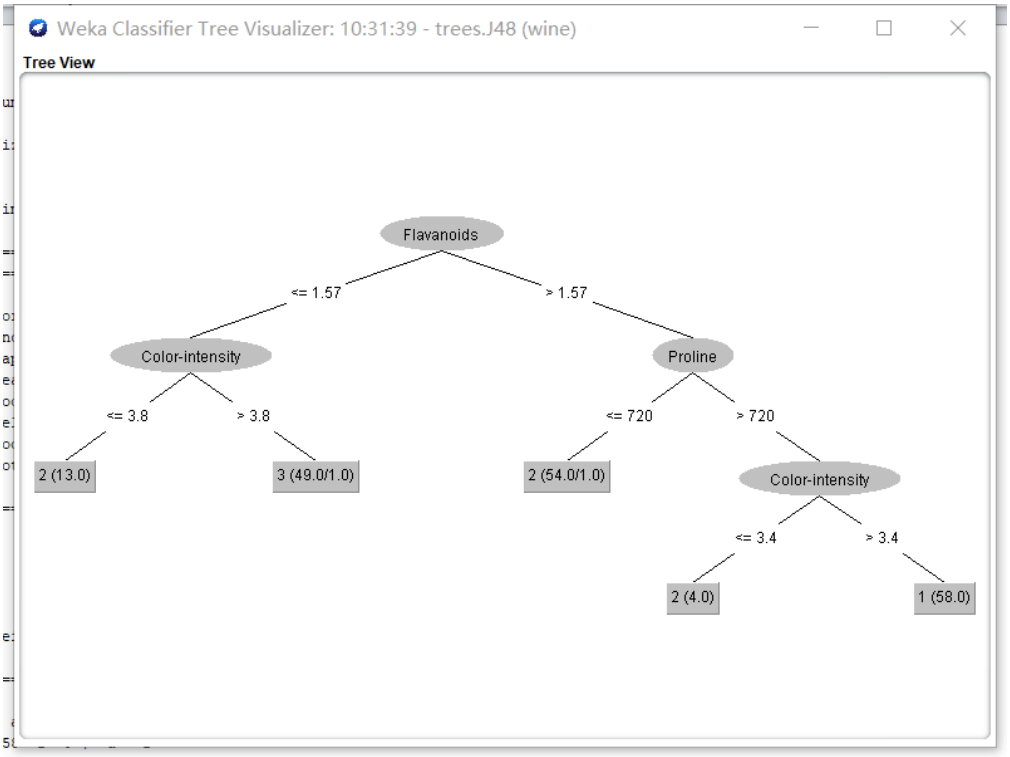

图 6: J48树结构

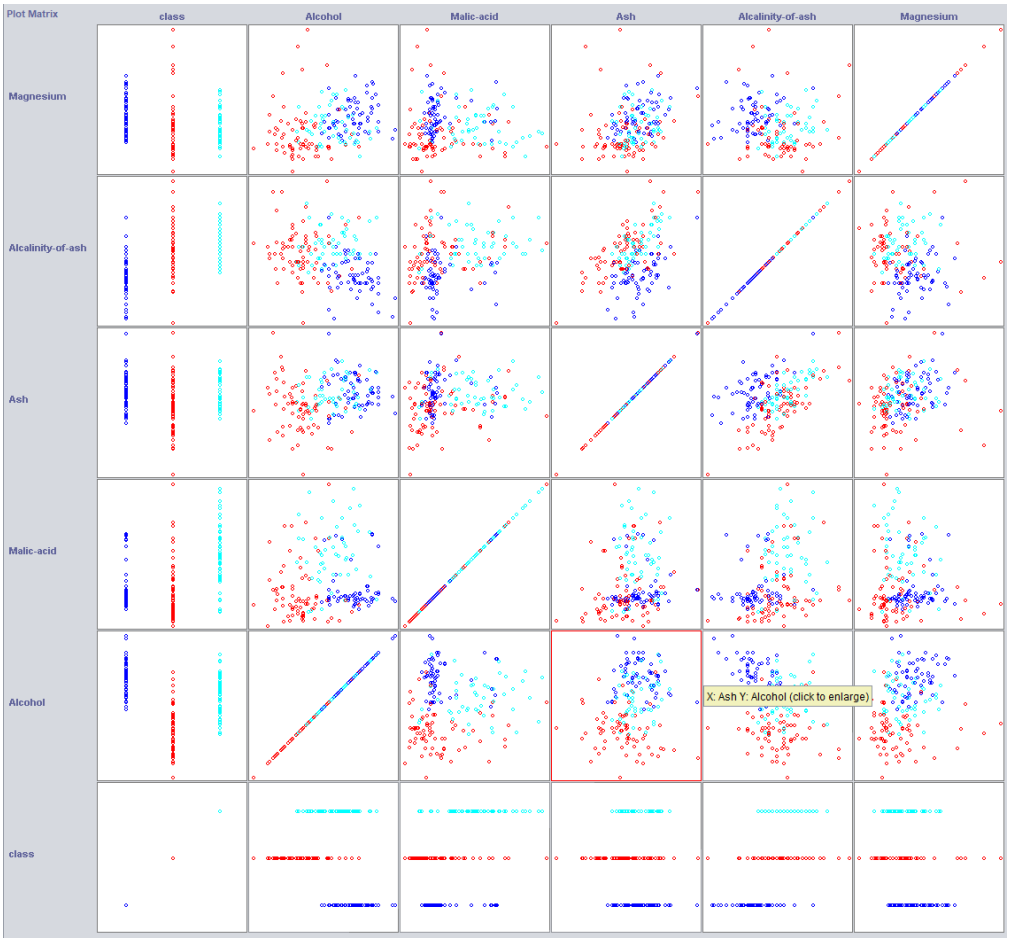
Plot Matrix
class
Alcohol
Malic-acid
Ash
Alcalinity-of-ash
Magnesium
Magnesium
Alcalinity-of-ash
Ash
Malic-acid
Alcohol
class
X: Ash Y: Alcohol (click to enlarge)

图 7: 变量关系