北京邮电大学软件学院

2019-2020 学年第 1 学期实验报告

课程名称：数据挖掘

实验名称：实验二：分类

实验完成人：

姓名：平雅霓 学号：2017211949

指导教师：牛琨

日　期：2019 年 10 月 28 日

一、 实验目的

熟悉 WEKA 软件的使用，加深对分类的理解。

二、 实验内容

(a) 实验内容如下：对给定数据集进行分类任务，并建立相应的分类器（如决策树），分析分类结果指标，比较不同的实验结果，以生成最佳模型。

(b) 解释你的模型。

三、 实验环境

Windows 环境、weka-3-8

四、 实验过程及结果

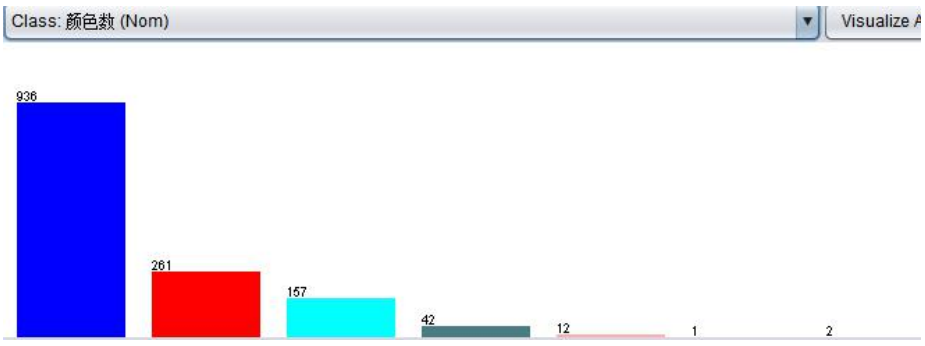
1. 分类尝试

（1） 以颜色数为 class 进行分类，前 70%为训练集，后 30%为测试集，采用 C4.5 算法。

选择颜色数作为 class 的原因：数据记录了 2003-2012 年间上市的手机，所以本次实验假设颜色和手机在市场上的受欢迎程度有关，即某一颜色的上市手机数量越多，则这个颜色越受大家欢迎。

因为颜色数的类型为 Numeric，所以在做分类操作前需要把颜色数这个属性进行离散化，使其类型变为 NomiNal 颜色数的取值范围为 1-7，所以离散化的时候将其离散为了 7 个值。

下图为颜色数离散化后的结果：



然后进行分类操作，分类的结果如下

```
Correctly Classified Instances         297               70.2128 %
Incorrectly Classified Instances       126               29.7872 %
Kappa statistic                          0
Mean absolute error                      0.1432
Root mean squared error                  0.2598
Relative absolute error                 99.503  %
Root relative squared error             99.9555 %
Total Number of Instances              423
```

可以看出，进行测试的实例有 423 个，预测的准确率为 70%左右。

下面我们来观察生成的 Confusion Matrix：

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 297   0   0   0   0   0   0 |   a = '(-inf-2]'
  71   0   0   0   0   0   0 |   b = '(2-3]'
  42   0   0   0   0   0   0 |   c = '(3-4]'
  10   0   0   0   0   0   0 |   d = '(4-4]'
   2   0   0   0   0   0   0 |   e = '(4-5]'
   1   0   0   0   0   0   0 |   f = '(5-6]'
   0   0   0   0   0   0   0 |   g = '(6-inf)'
```

由此矩阵发现，所有的测试实例都被预测成了类型 a，仔细想想这个现象其实可以解释的，因为训练集当中属于 a 类的实例占比 66%以上，所以训练出来的模型更倾向于将测试实例识别为类 a，继续进行实验发现，训练集的占比越大，正确率越接近 73%，但是所有的测试实例依然全部被识别为了类 a，由此可见，以颜色数作为 class 来训练模型是不太合理的。

（2） 以智能系统为 class 进行分类，采用 C4.5 算法。

选择智能系统作为 class 的原因：随着技术的发展，智能系统逐渐进入市场并迅速成为了占据市场极大分额的手机必备，所以此次尝试将智能系统作为了分类标签。

首先也是对智能系统的类型为进行了离散化操作，在整个 1411 条数据集中，1153 条数据信息中无智能系统，占总数据的 81.7%。

下图为总数据的前 70%为训练集的训练结果：

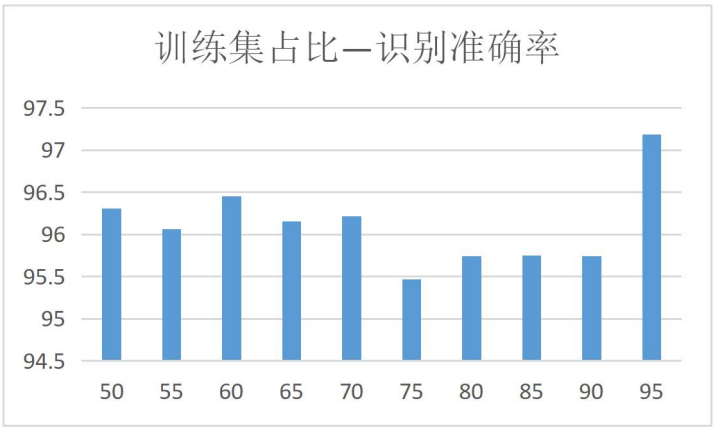
```
Correctly Classified Instances         407               96.2175 %
Incorrectly Classified Instances        16                3.7825 %
Kappa statistic                          0.8615
Mean absolute error                      0.0557
Root mean squared error                  0.1906
Relative absolute error                 19.0333 %
Root relative squared error             51.4303 %
Total Number of Instances              423
```

从图中可以看出，预测准确率约为 96.21%，这个数值是十分可观的，下面我们再来分析其 confusion matirx 来分析预测结果是否可信。

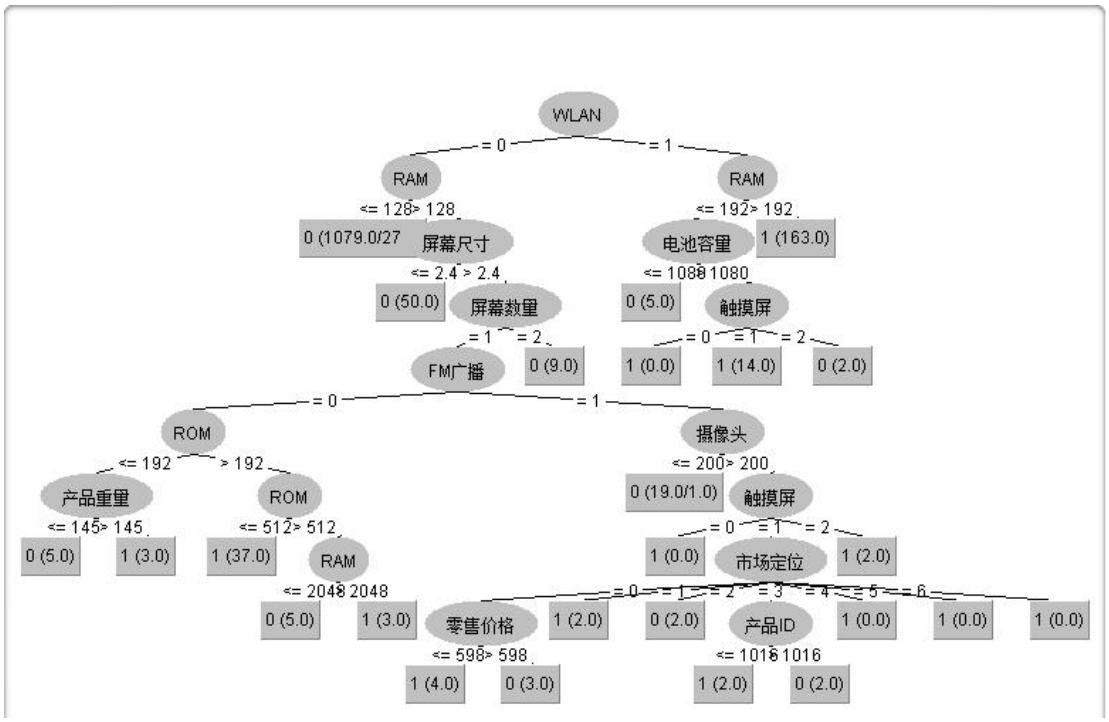
```
  a   b   <-- classified as
346   8 |   a = 0
  8  61 |   b = 1
```

在这个矩阵中，可以看出未出现上一次尝试中把所有实例都识别成一个类的情况，因此使得了本次尝试的准确率更高。

接下来，我调整了训练集占比，下图显示了训练集占比（横轴）和准确率（纵轴）的柱状图，从图中可以看出，当训练集占比为 95%的时候，准确率达到了 97%的最高值，但是实际上，训练集和测试集这样的分配比并不是相当合理，所以我认为当训练集占比为 60%的时候更为合理，此时的准确率为 96.4539%，效果也是十分可观的。



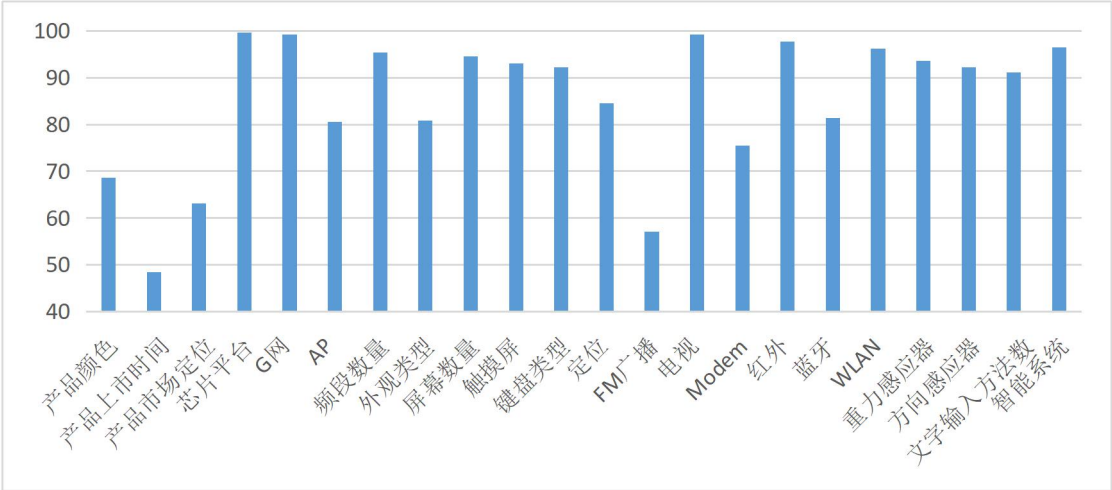
下图将展示当训练集占比为 60%时生成的决策树：



在实验一中的数据描述中可知 RAM 的平均值为 139.4232，所以决策树中 RAM 以 128 和 192 为分界是可信的，依次查看决策树中的其他节点，发现决策分支条件皆是合理的，根据此训练集训练出来的模型是可信的。在整个测试过程中，weka 根据实例的属性值沿着决策树往下走到叶节点，最终得出预测值，将预测值和原本的值进行比较，最后计算出准确率。

（3） 以其他属性为 class 进行分类，采用 C4.5 算法。

将数值个数低于 10 个的属性作为 class，采用将前 60%的数据作为训练数据，依次将产品颜色、产品上市时间、产品市场定位、芯片平台、G 网、AP、频段数量、外观类型、 屏幕数量、触摸屏、键盘类型、定位、 FM 广播、电视、Modem、红外、蓝牙、WLAN、重力感应器、方向感应器、 文字输入方法数、智能系统作为 class，进行了分类器构建，最终的准确率对比如下图所示：



从图中可以看出准确率超过 90%的有芯片平台、G 网、频段数量、屏幕数量、键盘类型、点视、红外、WLAN、重力感应器、方向感应器、文字输入方法数、智能系统这些属性，这些属性皆与硬件和新型技术有关，所以可以从中发现，这些属性都是十分重要的。

2. 算法对比：

前面的模型都是基于 C4.5 算法进行训练的，这一部分我对其他的算法也进行了比较，下面的表是其他算法对应的准确率，其中训练数据占比为 60%，使用的 class 类为智能系统。

算法	准确率	算法	准确率
J48	96.45%	SimpleLogistic	95.92%
RandomForest	95.21%	BayesNet	93.97%
RandomTree	90.96%	SMO	95.74%
REPTree	83.16%	DecisionTable	95.92%

LMT	95.92%	InputMappedClassifier	83.16%
DecisionStump	95.04%	IBk	93.26%
HoeffdingTree	93.79%		

从表中可以看出 J48（即 C4.5）算法的效果的确十分优秀，更适合于这份数据，所以在本次的实验中我最终选择的是 J48 算法来进行模型训练。

3. 结论：

经过以上分析，最终我选择的是使用 J48（C4.5）算法，将智能系统作为分类标签，训练集占比为 60%的模型，测试集准确率为 96.4539%。

此模型将智能系统作为分类标签十分具有代表性，而且经过多次实验发现，训练集占比为 60%时效果最佳，横向对比其他算法，J48 在此数据集上效果最佳。