**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**

1. **实验目的**

熟悉WEKA软件的使用，加深对分类的理解。

1. **实验内容**
2. 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。
3. 解释你的模型。
4. **实验环境**

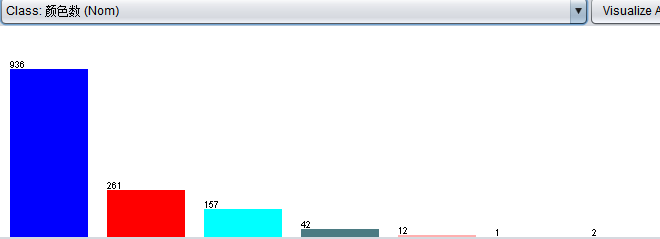
Windows环境、weka-3-8

1. **实验过程及结果**
2. **分类尝试**
3. **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。**

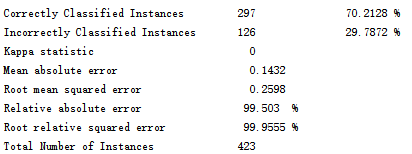
选择颜色数作为class的原因：数据记录了2003-2012年间上市的手机，所以本次实验假设颜色和手机在市场上的受欢迎程度有关，即某一颜色的上市手机数量越多，则这个颜色越受大家欢迎。

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：

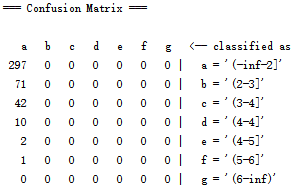


然后进行分类操作，分类的结果如下



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：



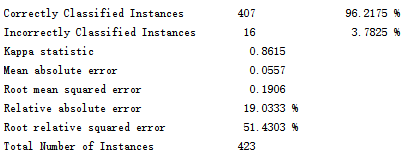
由此矩阵发现，所有的测试实例都被预测成了类型a，仔细想想这个现象其实可以解释的，因为训练集当中属于a类的实例占比66%以上，所以训练出来的模型更倾向于将测试实例识别为类a，继续进行实验发现，训练集的占比越大，正确率越接近73%，但是所有的测试实例依然全部被识别为了类a，由此可见，以颜色数作为class来训练模型是不太合理的。

1. **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：



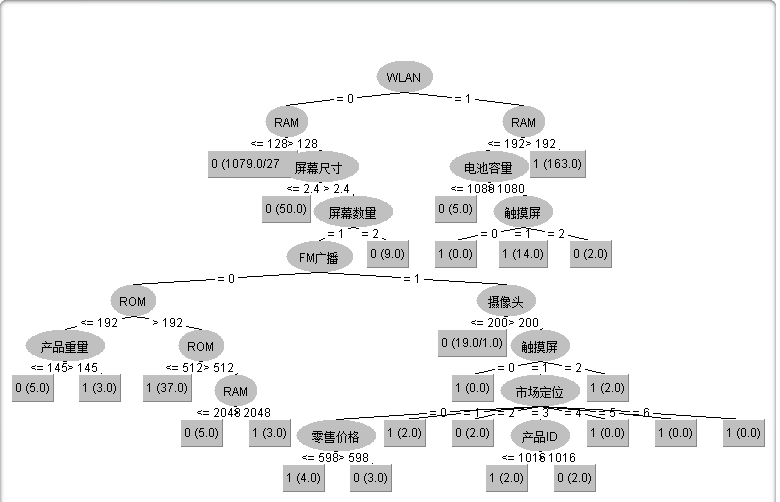
从图中可以看出，预测准确率约为96.21%，这个数值是十分可观的，下面我们再来分析其confusion matirx来分析预测结果是否可信。



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。

1. **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。

1. **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。

1. **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。