**基于WEKA的数据挖掘算法学习（二）**

# 一、数据集的选择

1. 下载UCI完备数据集：waveform-data.csv；
2. 下载UCI不完备数据集：pima-indians-diabetes-data.csv
3. 下载UCI不平衡数据集：car-data.csv

# 二、weka中classify的数据测试

## 1.J48

### （1）算法理论知识

ID3算法是一种贪心算法，用来构造决策树。在每个节点选取还尚未被用来划分的具有最高信息增益的属性作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。而C4.5（J48）没有改变ID3的算法逻辑，但在节点的划分上的划分标准做了改进。

C4.5使用信息增益率（GainRatio）来代替信息增益（Gain）进行特征的选择，即利用信息增益率来选择特征。克服了信息增益选择特征时偏向于特征值个数较多的不足。信息增益率（GainRatio）的定义如下：

### （2）weka数据测试

选择五折交叉数据验证，不断调整minNumObj参数进行数据测试，结果如下：

M=2时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| waveform | 0.737 | 0.755 | 0.755 | 0.842 | 75.48 % | 2.2s |
| pima | 0.715 | 0.712 | 0.713 | 0.729 | 71.224% | 0.03s |
| car | 0.917 | 0.916 | 0.915 | 0.972 | 91.55% | 0.06s |

M=4时：

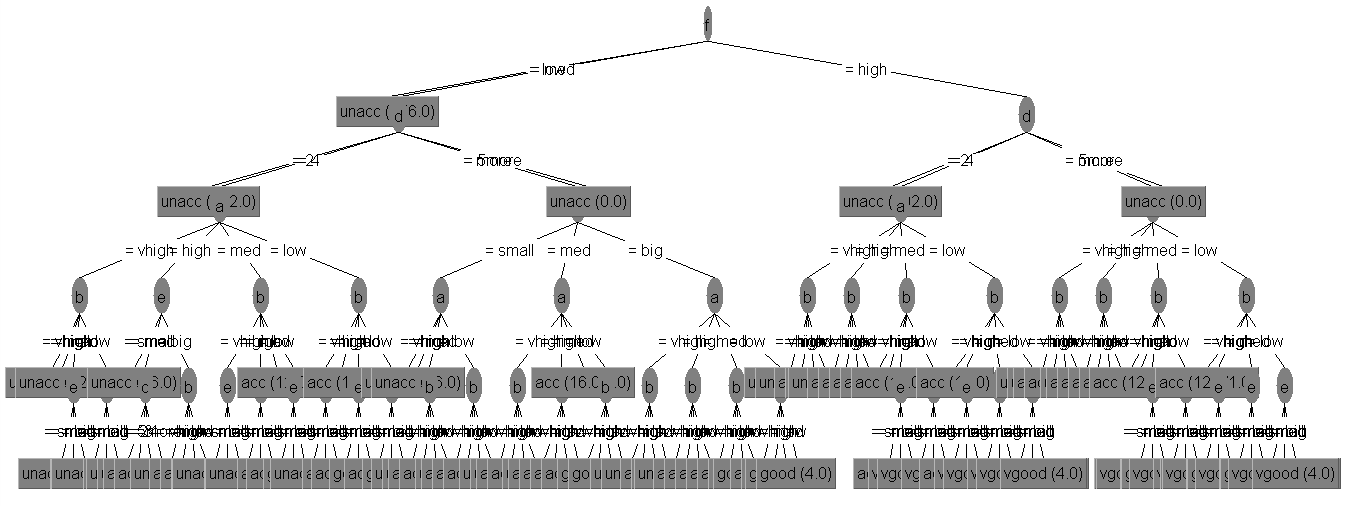
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| waveform | 0.757 | 0.758 | 0.757 | 0.868 | 75.76 % | 0.75s |
| pima | 0.731 | 0.729 | 0.730 | 0753 | 72.92% | 0.05s |
| car | 0.878 | 0.878 | 0.876 | 0.970 | 87.79% | 0.05s |

M=6时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| waveform | 0.758 | 0.759 | 0.758 | 0.888 | 75.86 % | 0.7s |
| pima | 0.736 | 0.733 | 0.734 | 0.767 | 73.71 % | 0.03s |
| car | 0.859 | 0.863 | 0.857 | 0.969 | 86.285% | 0.03s |

由实验结果可以得到，完备数据集waveform和不完备数据集pima使用J48分类器效果在M=6时效果最佳，而不平衡数据集car在M=2时效果最佳。整体来看，选择car数据集作为测试J48分类器时，M=2时的效果更好。

下面是决策树截图：



## 2.CART

### （1）算法理论知识

CART算法考虑到每个节点都有成为叶子节点的可能，对每个节点都分配类别,分配类别的方法可以用当前节点中出现最多的类别。对分类树，CART利用基尼指数（Gini index）最小化原则，进行特征选择，生成二叉树。其中基尼指数的定义如下：

分类问题中，假设k个类别，样本点属于第k类的概率为。

### （2）weka数据测试

准备使用CART算法时，找了很久很久，发现calssify里面并没有。后来通过在Tools->

Package manager->ALL->simpleCart中，





下载安装simpleCart，classify中就可以实现CART算法的数据测试。

选择五折交叉数据验证，测试数据集，结果如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| waveform | 0.766 | 0.766 | 0.766 | 0.899 | 76.62 % | 3.81s |
| pima | 0.730 | 0.736 | 0.732 | 0.763 | 73.57 % | 0.63s |
| car | 0.973 | 0.973 | 0.973 | 0.991 | 97.28% | 2.74s |

## 3.神经网络（Multilayerceptron）

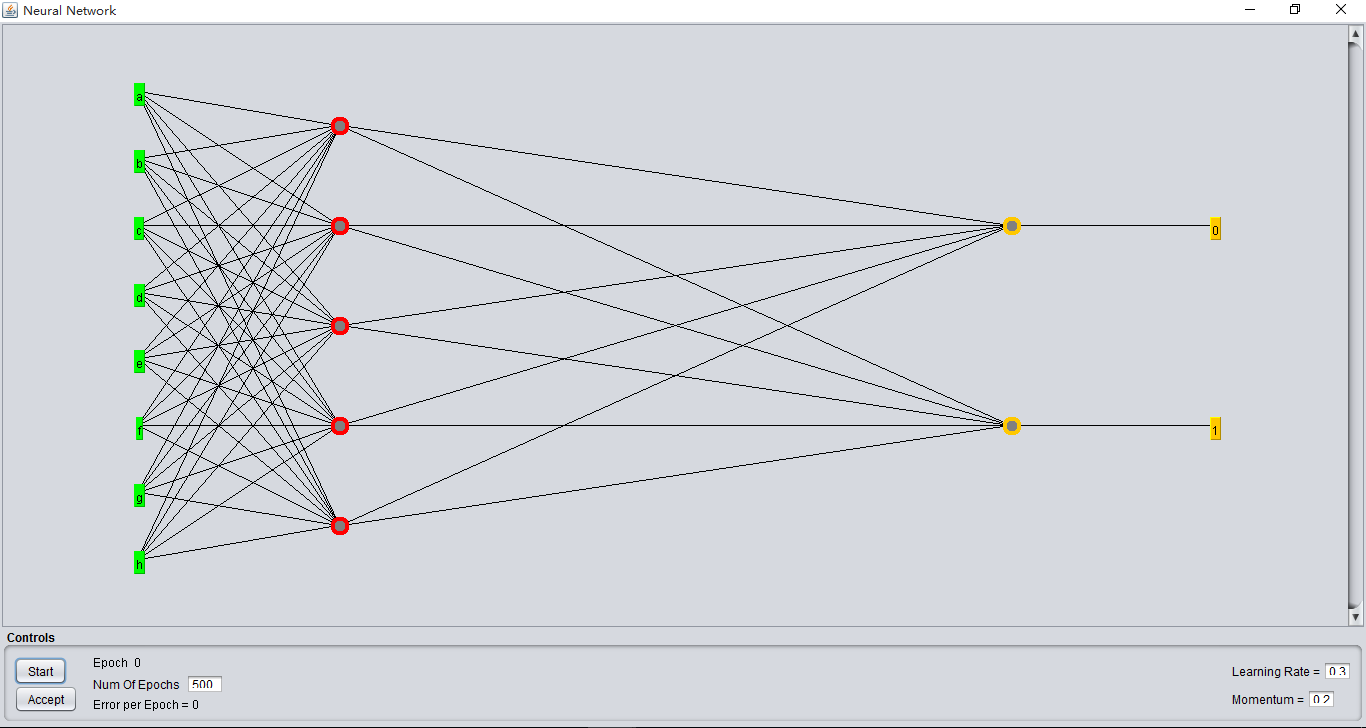
### （1）算法理论知识

Multilayerceptron多层前馈网络，由于多层前馈网络的训练经常采用误差反向传播算法，也称为BP神经网络。BP网络能学习和存贮大量的输入-输出模式映射关系，它的学习规则是使用梯度下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构：输入层（input）、隐含层(hidden layer)和输出层(output layer)。

### （2）weka数据测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| waveform | 0.836 | 0.836 | 0.835 | 0.964 | 83.6 % | 61.3s |
| pima | 0.749 | 0.751 | 0.750 | 0.809 | 75.13 % | 1.66s |
| car | 0.990 | 0.990 | 0.857 | 1 | 99.02% | 21.99s |

由结果可以看出，car数据集作为测试神经网络算法最好，且时间效率上来说明显优于waveform数据集。下面测试pima数据集时的GUI界面：



GUI界面可以允许我们在神经网络训练的过程中暂停和做一些修改。

其中标签（label）提供的原始输入（raw input）在左边，红色的节点是隐层（hidden layers），橙色的节点是输出节点（output nodes）在右边的标签展示的是输出节点表示的类别。要注意的是对于一个数值属性的类别来说，输出节点将自动的做成一个unthresholded的线性单元。

# 三、实验分析与总结

# 1.数据结果分析

### （1）完备数据集waveform

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| J48 | 0.737 | 0.755 | 0.755 | 0.842 | 75.48 % | 2.2s |
| CART | 0.766 | 0.766 | 0.766 | 0.899 | 76.62 % | 3.81s |
| 神经网络 | 0.836 | 0.836 | 0.835 | 0.964 | 83.6 % | 61.3s |

通过完备数据集waveform数据测试结果来看，神经网络算法跑出来的结果最好，但是时间性能上来说，又耗费太多的时间。

### （2）不完备数据集pima

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| J48 | 0.715 | 0.712 | 0.713 | 0.729 | 71.224% | 0.03s |
| CART | 0.730 | 0.736 | 0.732 | 0.763 | 73.57 % | 0.63s |
| 神经网络 | 0.749 | 0.751 | 0.750 | 0.809 | 75.13 % | 1.66s |

从不完备数据集pima测试结果来看，神经网络效果最佳，但时间效率最低。

### （3）不平衡数据集car

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| J48 | 0.917 | 0.916 | 0.915 | 0.972 | 91.55% | 0.06s |
| CART | 0.973 | 0.973 | 0.973 | 0.991 | 97.28% | 2.74s |
| 神经网络 | 0.990 | 0.990 | 0.857 | 1 | 99.02% | 21.99s |

根据不平衡数据集car测试三种算法结果来看，CART算法和神经网络算法结果都很不错，其中CART算法的F值（F 值即为正确率和召回率的调和平均值，用于综合反映整体的指标）大于神经网络，而神经网络的AUC值达到了1（等于1时表明效果最佳）。时间性能上来说，CART优于神经网络算法。

# 2.三种分类算法的总结

## J48

J48优点：产生的分类规则易于理解，准确率较高；其缺点是：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。从实验结果来看，J48的准确率等等评估指标是明显低于CART算法的，但是从时间效率上来看更加高效。

## CART

CART的优点是非常灵活，可以允许有部分错分成本，还可指定先验概率分布，可使用自动的成本复杂性剪枝来得到归纳性更强的树。面对存在缺失值、变量数多等问题时，CART数显得非常稳健。从实验结果来看，时间效率高效且评价指标结果都很优越。

## 神经网络

神经网络算法优点是网络实质上实现了一个从输入到输出的映射功能，而数学理论已证明它具有实现任何复杂非线性映射的功能，这使得它特别适合于求解内部机制复杂的问题，能通过学习带正确答案的实例集自动提取“合理的”求解规则，即具有自学习能力。而缺点是BP算法实质是梯度下降法，而它所要优化的目标函数又非常复杂，使得算法非常的低效或者停滞。结合我所跑的数据结果来看，确实比其他类别的分类算法低效很多，但准确率等指标却相当的优越。

# 四、数据集的选择

本周继续使用了weka中的分类算法来测试数据，了解部分了J48、CART、神经网络三类算法的理论知识和优势与劣势，但还未深入理解算法原理，还需加强巩固。