## problem2

（1）决策树生成策略

使用熵不纯度计算样本的不纯度，即

使用信息增益率最大值作为选取特征标准，即

为了防止过拟合，加入预剪枝操作，使决策树停止生成的条件如下：

1. 子树中全为同一类;
2. 可选的特征为零;
3. 最大增益率小于阈值threshold1;
4. 树的深度超过阈值threshold2;
5. 子树中的样本数小于阈值threshold3.

（2）实验结果

选取

1）停止条件为a）~c），实验结果如表1所示。

表 1 threshold1对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold1 | 0.03 | 0.05 | 0.07 | 0.1 |
| 训练集准确率（%） | 99.45 | 98.20 | 97.08 | 87.22 |
| 测试集准确率（%） | 73.33 | 73.68 | 74.13 | 70.93 |
| 决策树深度 | 208 | 203 | 202 | 192 |

由上表可以看出，随着threshold1的增大，训练集上的测试率在0.03~0.07区间上逐渐增大，但是增大幅度并不是很大；加入threshold1可以减弱不纯度下降很小时产生过拟合的情况。而当threshold1=0.1时，测试集准确率较低，说明此时决策树生成还不完全，需要继续分类。因此，在后面的试验中，均取threshold1=0.07.

2）停止条件为a）~d），实验结果如表2所示。

表 2 threshold2对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold2 | 10 | 30 | 50 | 100 |
| 训练集准确率（%） | 67.04 | 92.98 | 96.96 | 97.67 |
| 测试集准确率（%） | 62.40 | 74.90 | 74.86 | 74.44 |

由上表可以看出，在固定最大增益率，加入树的深度限制之后，在threshold2大于等于30后，测试集整体正确率大于1）中的准确率，说明减少了过拟合，另外限制树的深度既可以加快收敛速度，又可以提高测试准确率。

3）停止条件为a）~c）、e），实验结果如表3所示

表 3 threshold2对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold3 | 10 | 20 | 40 | 50 |
| 训练集准确率（%） | 67.04 | 92.98 | 96.96 | 97.67 |
| 测试集准确率（%） | 62.40 | 74.90 | 74.86 | 74.44 |

4）停止条件为a）~e），实验结果如表4所示。

（3）改进决策树