## 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. 树的深度大于阈值threshold2;
5. 子树中的样本数小于阈值threshold3.

（2）实验结果

随机选取sougou数据的80%作为测试集，20%作为训练集，进行以下实验。

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），取threshold1=0.07，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运行速度，又可以提高测试准确率。

3）停止条件为a）~c）、e），取threshold1=0.07，实验结果如表3所示

表 3 threshold3对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold3 | 10 | 20 | 40 | 50 |
| 训练集准确率（%） | 91.35 | 87.38 | 83.70 | 82.86 |
| 测试集准确率（%） | 76.29 | 75.38 | 76.66 | 77.12 |

由上表可以看出，固定最大增益率，加入子树最小个数限制后，训练集的整体准确率也高于1）中的结果，说明加入阈值threshold3也有利于减少过拟合，另外限制子树最小个数也可以加快运行孙杜，提高测试准确率。

4）停止条件为a）~e），取threshold1=0.07，threshold2=50，实验结果如表4所示。

表 4 threshold3对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold3 | 10 | 20 | 40 | 50 |
| 训练集准确率（%） | 90.63 | 87.32 | 82.68 | 83.17 |
| 测试集准确率（%） | 73.72 | 75.56 | 75.56 | 75.45 |

由上表可以看出，随着threshold3的增大，测试集准确率先增大后缩小，但是变化范围不大，相较于2）中的结果74.86%又有提升，说明在threshold1与threshold2固定时，threshold3有提高准确率的作用。

5）停止条件为a）~e），取threshold1=0.07，threshold3=20，实验结果如表5所示。

表 5 threshold2对准确率的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold2 | 10 | 30 | 50 | 100 |
| 训练集准确率（%） | 67.32 | 85.18 | 87.32 | 87.05 |
| 测试集准确率（%） | 64.20 | 74.13 | 75.56 | 76.60 |

由上表可以看出，随着threshold2的增大，测试集准确率逐渐增大，当threshold2=100时，相较于3）中的75.38%的测试集准确率有所提高，说明在threshold1与threshold3固定时，threshold2有提高准确率的作用。

（3）改进

采用多决策树，将训练集平均分为k份，训练k个决策树，用k个决策树对测试集进行分类，最后结果取k次分类的众数。

取threshold1=0.07，threshold2=100，threshold3=20，在不同的k值下，测试集准确率如表6所示。

表 6 threshold3对准确率的影响

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | 1 | 3 | 5 | 7 | 9 |
| 测试集准确率（%） | 76.60 | 76.94 | 77.24 | 77.43 | 77.66 |

由上表可以看出，多决策树与单一决策树相比，随着决策树数目的增加，测试集准确率得到了明显的提高，同时，由于训练样本缩小，改进后的算法效率也得到了提高。不过在使用多决策树时，采用信息增益来作为特征选取结果更好，原因可能是少量样本更适合使用信息增益来作为训练标准。