4.3剪枝处理

概述

目的

对付过拟合

数据集

表 4.2　西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

训练集（使用信息增益准则生成决策树）

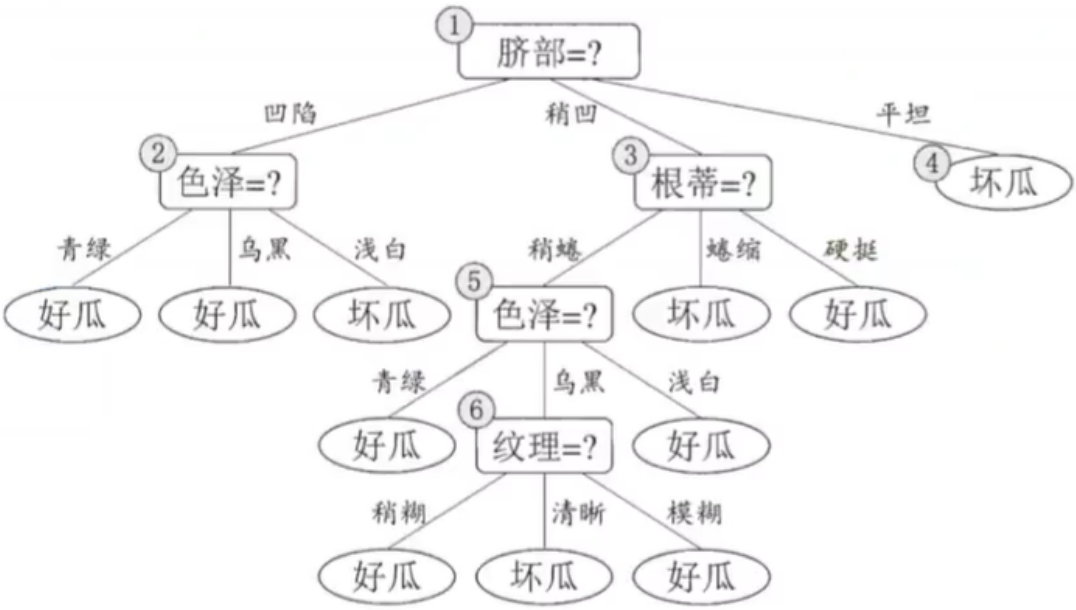
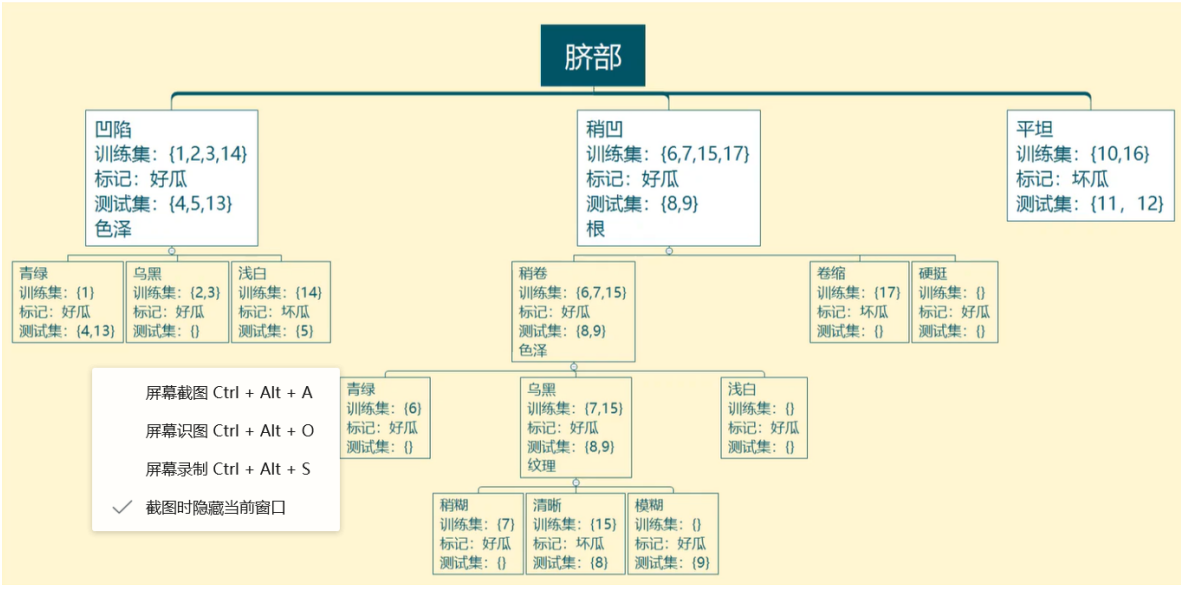
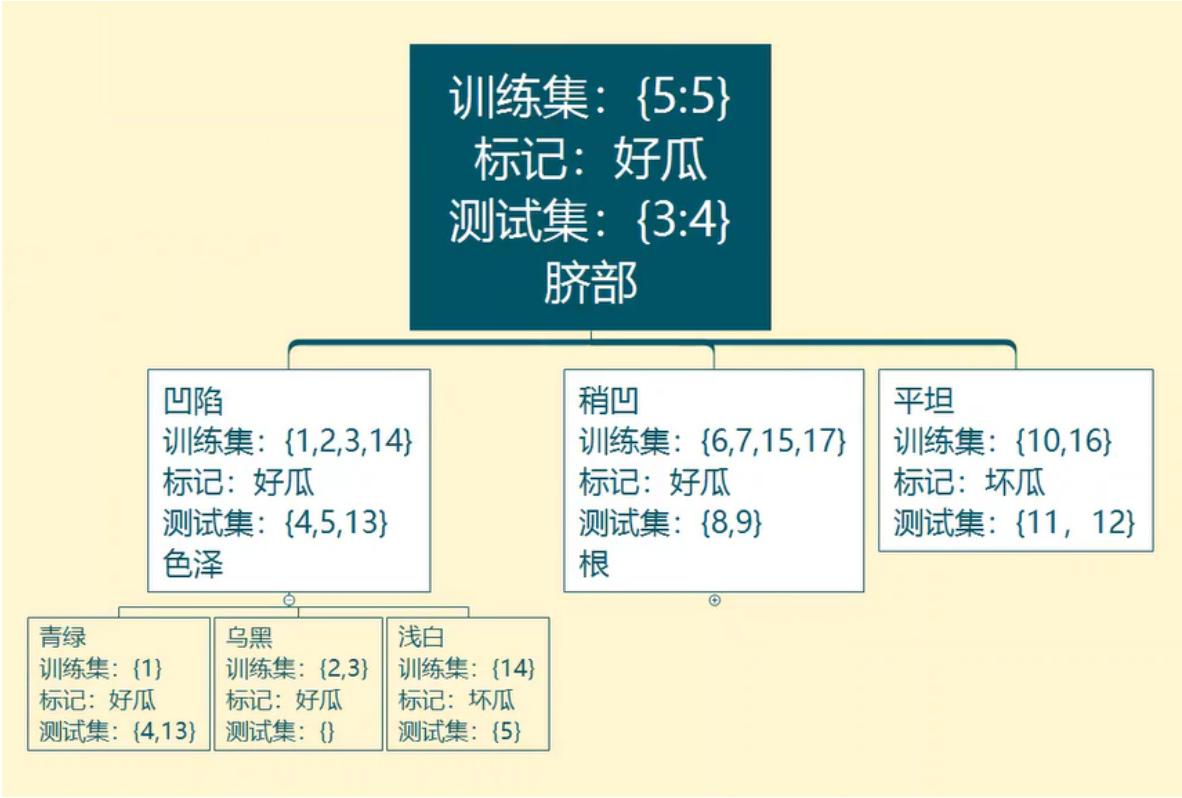


图 4.5 基于表 4.2 生成的未剪枝决策树

训练集与验证集在决策树上的情况



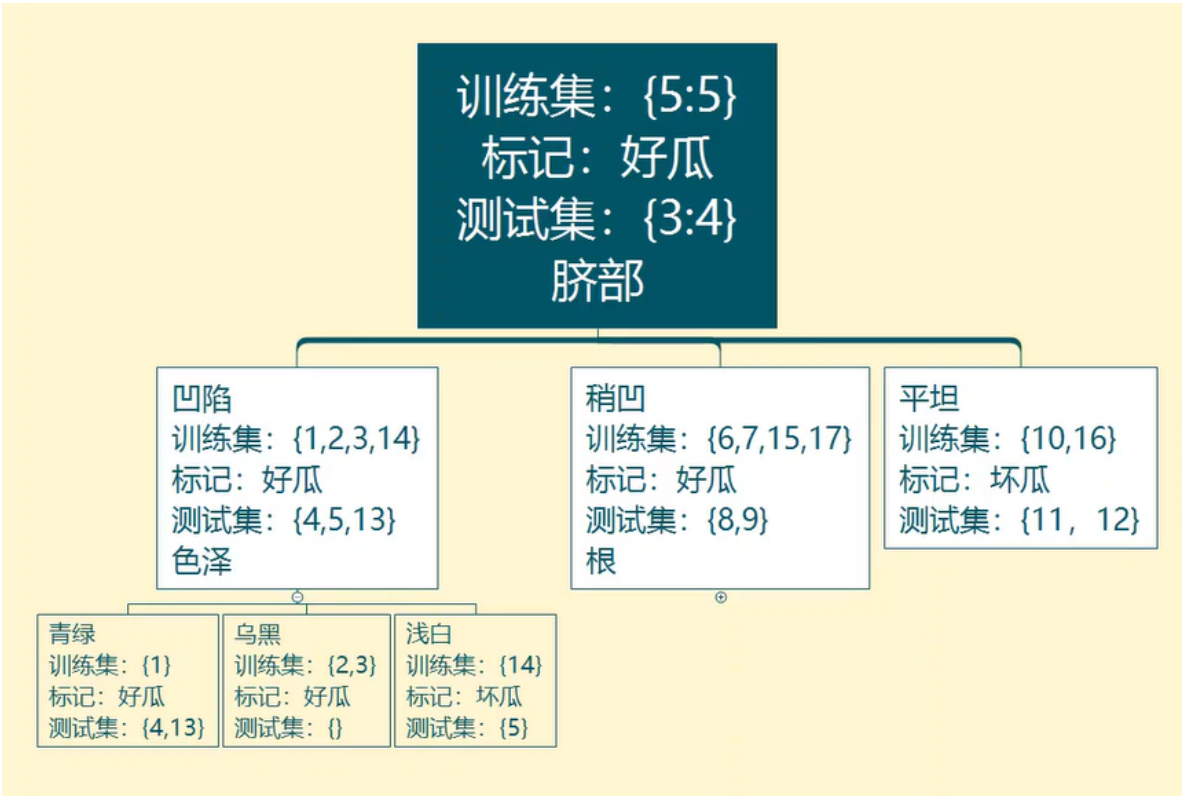
4.3.1预剪枝(prepruning)



分之前正确率为$\frac{3}{7}$

按照脐部分之后正确率为$\frac{5}{7}$

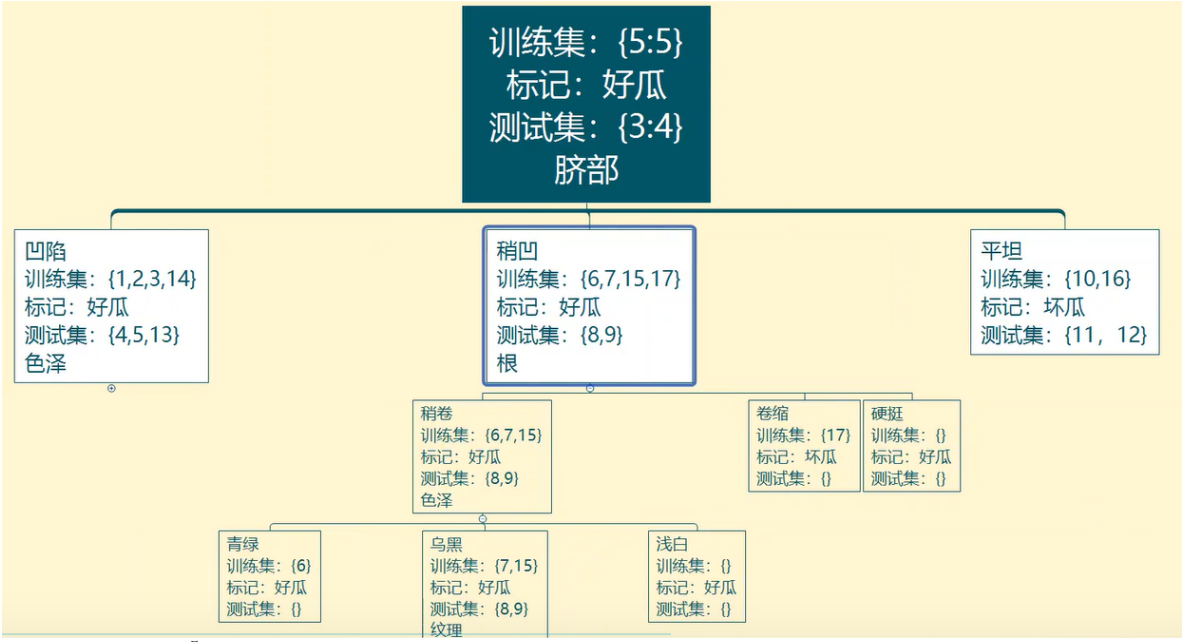
正确率上升，划分正确



分之前正确率为$\frac{5}{7}$

按照脐部分之后正确率为$\frac{4}{7}$

正确率下降，则不需要该划分，剪枝

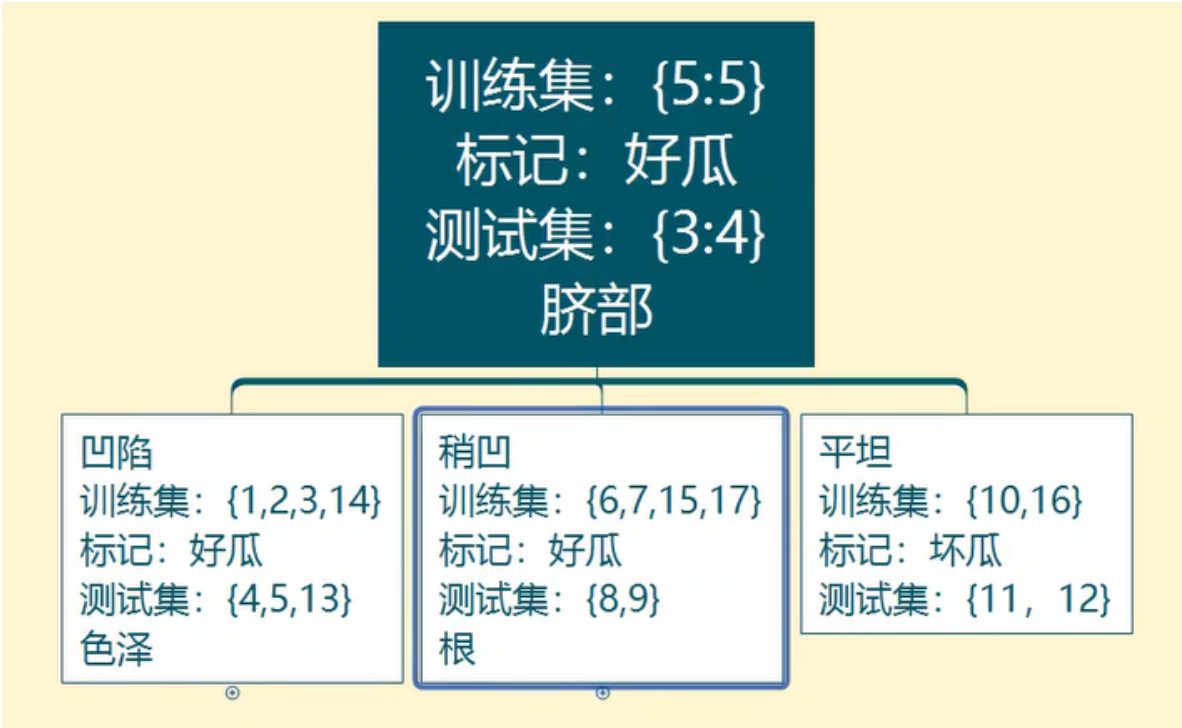


分之前正确率为$\frac{5}{7}$

按照脐部分之后正确率为$\frac{5}{7}$

正确率不变，则没必要划分，剪枝

则最终只需要对脐部进行划分：



最终决策树：

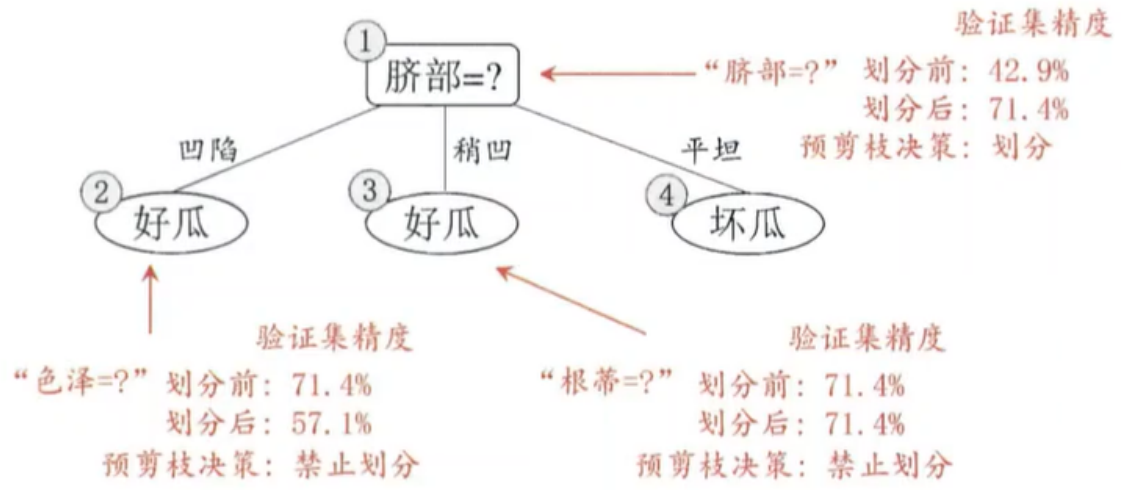
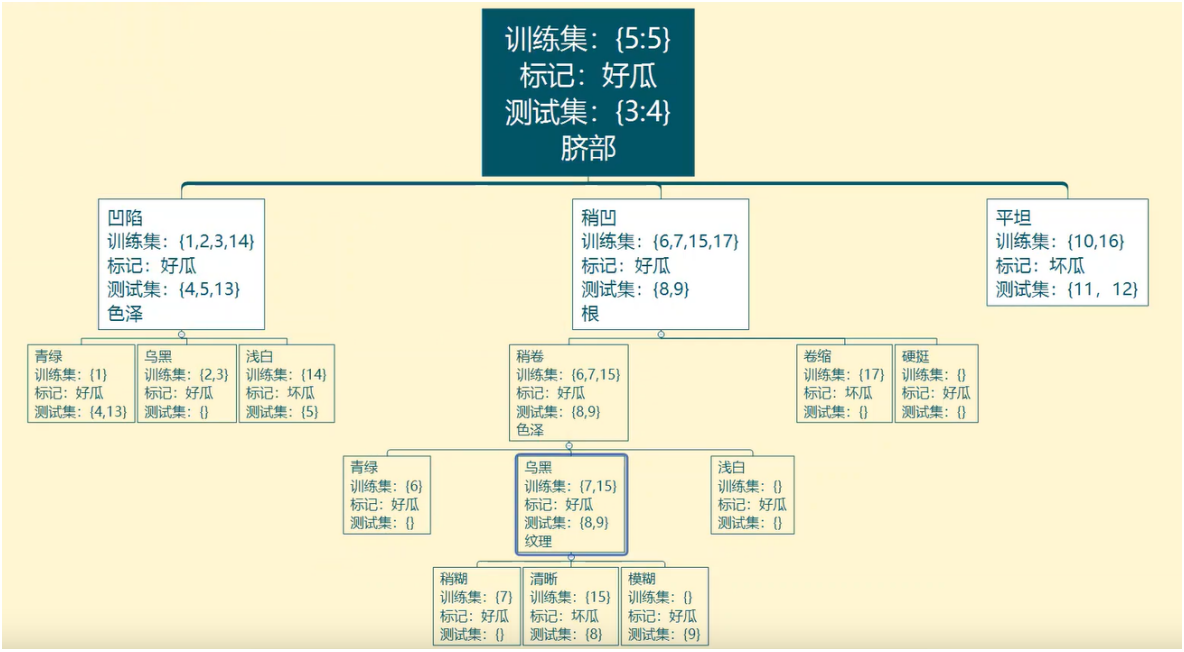


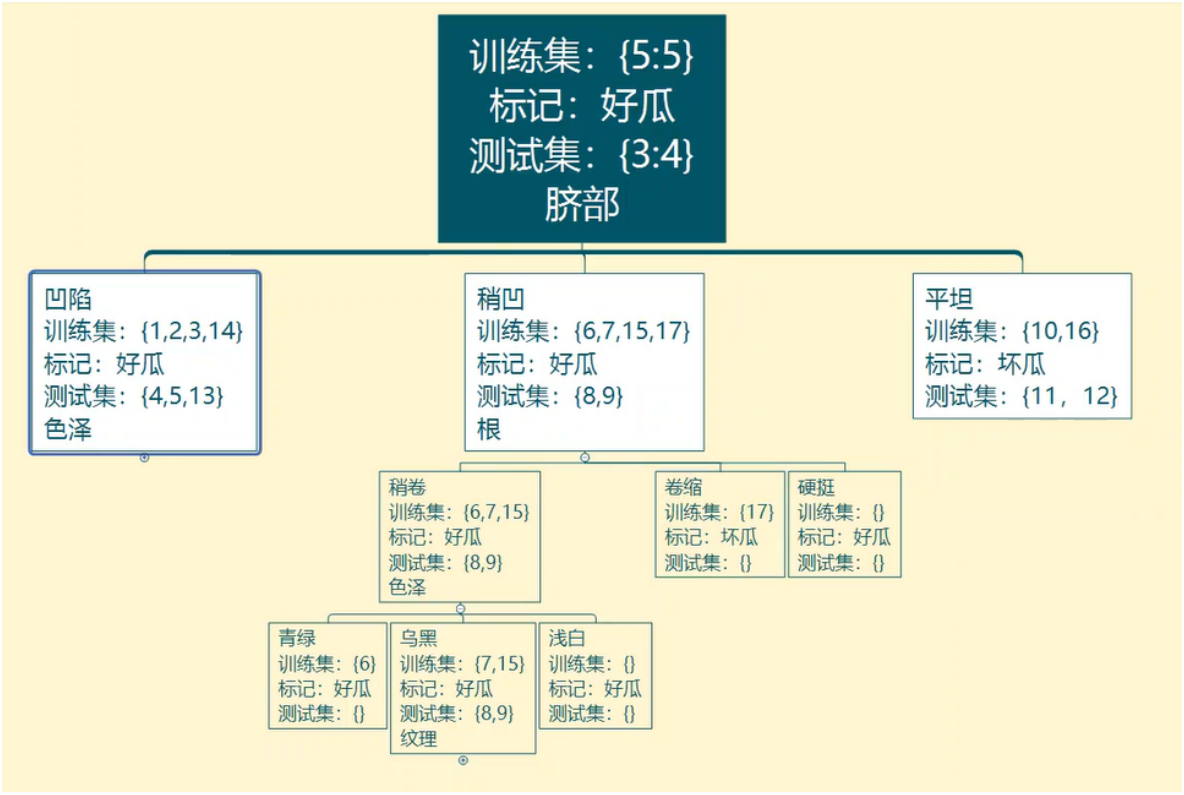
图 4.6 基于表 4.2 生成的预剪枝决策树

4.3.2后剪枝(post-pruning)

剪枝前决策树：



从叶子节点开始，自下而上计算正确率，若父节点>儿子节点正确率，则剪枝，若父节点<=儿子节点的正确率，则不剪枝。
剪掉最左边的色泽和中间最下边的纹理两个分枝：



最终决策树：

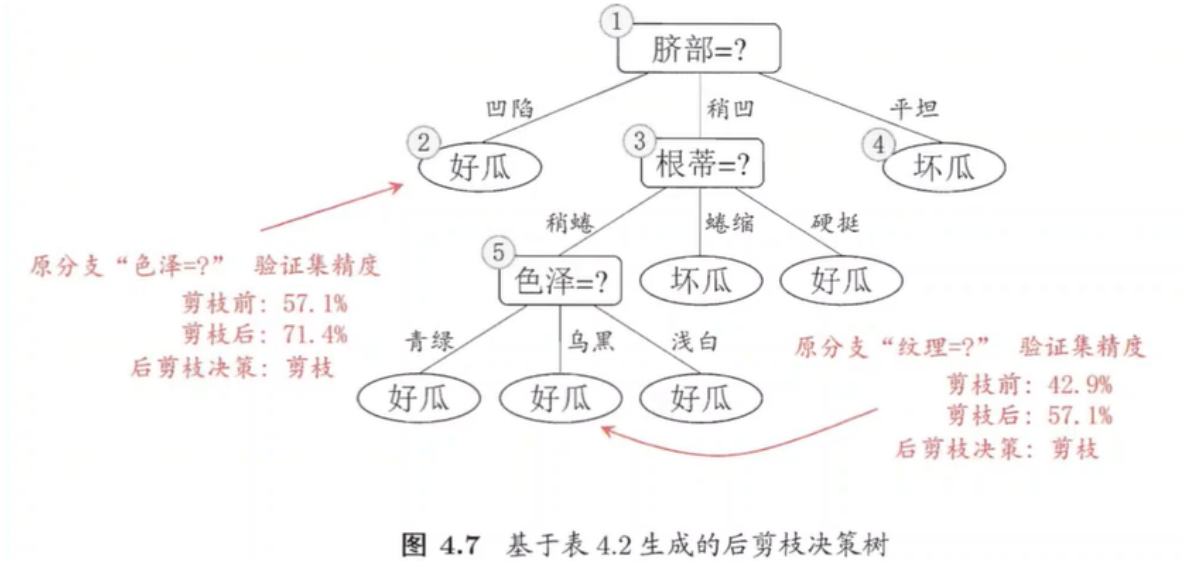


图 4.7 基于表 4.2 生成的后剪枝决策树

补充：回归剪枝的方法

详解：https://www.youtube.com/watch?v=D0efHEJsfHo