作业 7 决策树

使用软件：EXCEL, SPSS

思路过程：

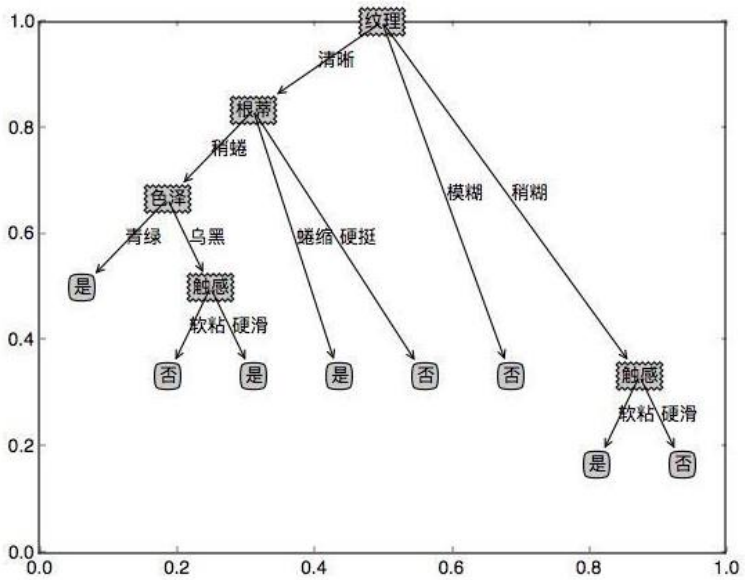
首先面临的一大困难就是数据问题

最初我试图利用国家统计局网站上的一些数据来训练，但国家统计局的数据随着年份有规律性变化，极难作为彼此独立的训练数据，训练结果非常差。

然后我试着直接搜索一些决策树训练的数据集，这次面临的是数据连续性问题，大型可用的训练数据集不少是图像型或其他连续性数据，需要进行离散化处理，通过阅读网络资源发现多为编程处理，EXCEL 虽然功能强大，但对于复杂的数据类型处理能力和便利程度还是有所欠缺。

训练数据问题困扰了我一个星期，最后我决定反向思考，既然没有合适的数据来训练决策树，那不如我根据已有的树来生成数据，再回头检验 SPSS 生成的树与我预期的树的差别，这样来观察 SPSS 的分类功能的特点反而更加清晰。

于是我在网上找到这样一棵判断瓜是否为好瓜的决策树



可以看到，这棵树是通过对瓜的纹理、根蒂、色泽和触感来判断是否为好瓜。我对这几种属性的不同值进行了编码，如下表。并使用 1 和 0 来表示是否为好瓜

	纹理	根蒂	色泽	触感
1	清晰	蜷缩	青绿	软粘
2	稍糊	稍蜷	乌黑	硬滑
3	模糊	硬挺		

接下来在 EXCEL 中利用 RANDBETWEEN 函数随机生成了 1000 个具有这四个属性的不同值的瓜，再利用 IF 函数按照上面的树给出了判断结果。由于随机数具有易失性，为便于之后对于原始数据的检查判断，将其中一次随机得到的所有数据以值的形式重新保存固定下来。将此训练集导入 SPSS，如下图所示

	纹理	根蒂	色泽	触感	是否为好瓜
1	2	3	2	1	1
2	2	3	2	2	0
3	3	1	2	2	0
4	3	3	2	1	0
5	2	1	2	2	0
6	1	1	1	2	1
7	2	1	1	2	0
8	2	2	2	2	0
9	3	2	1	2	0
10	2	1	1	1	1
11	3	1	2	2	0
12	1	2	2	2	1
13	3	2	2	2	0
14	2	3	1	1	1
15	2	2	2	1	1

数据视图 变量视图

对每个变量都添加标签如下图所示

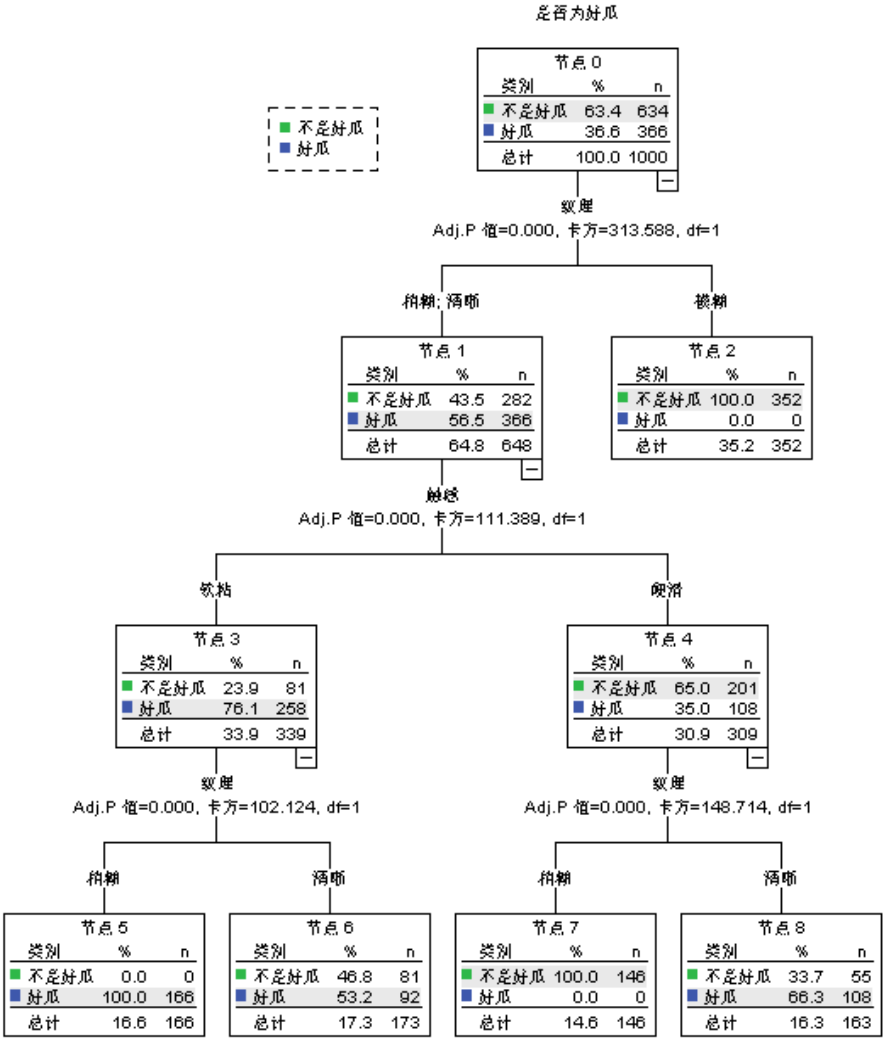
	名称	类型	宽度	小数位数	标签	值	缺失	列	对齐	测量	角色
1	纹理	数字	1	0		{1, 清晰}...	无	12	右	名义	输入
2	根蒂	数字	1	0		{1, 蜷缩}...	无	12	右	名义	输入
3	色泽	数字	1	0		{1, 青绿}...	无	12	右	名义	输入
4	触感	数字	1	0		{1, 软粘}...	无	12	右	名义	输入
5	是否为好瓜	数字	1	0		{0, 不是好瓜}...	无	12	右	名义	输入

到这里对数据的前期处理就做好了，接下来使用 SPSS 的决策树功能来生成树！

SPSS 共支持 CHAID、穷举 CHAID、CRT 和 QUEST 四种生长方法，接下来分别测试这四种方法。

1. CHAID

CHAID 生长法生成的树如下图所示，



模型摘要		
结果	包括的自变量	纹理，触感
	节点数	9
	终端节点数	5
	深度	3

由结果可知，CHAID 生长法得到的树只包含两个自变量，由于纹理有三个值，间接二分了两次。但由于少了两个自变量，预测结果不佳，由 SPSS 给出的实测可知正确率仅为 86.4%

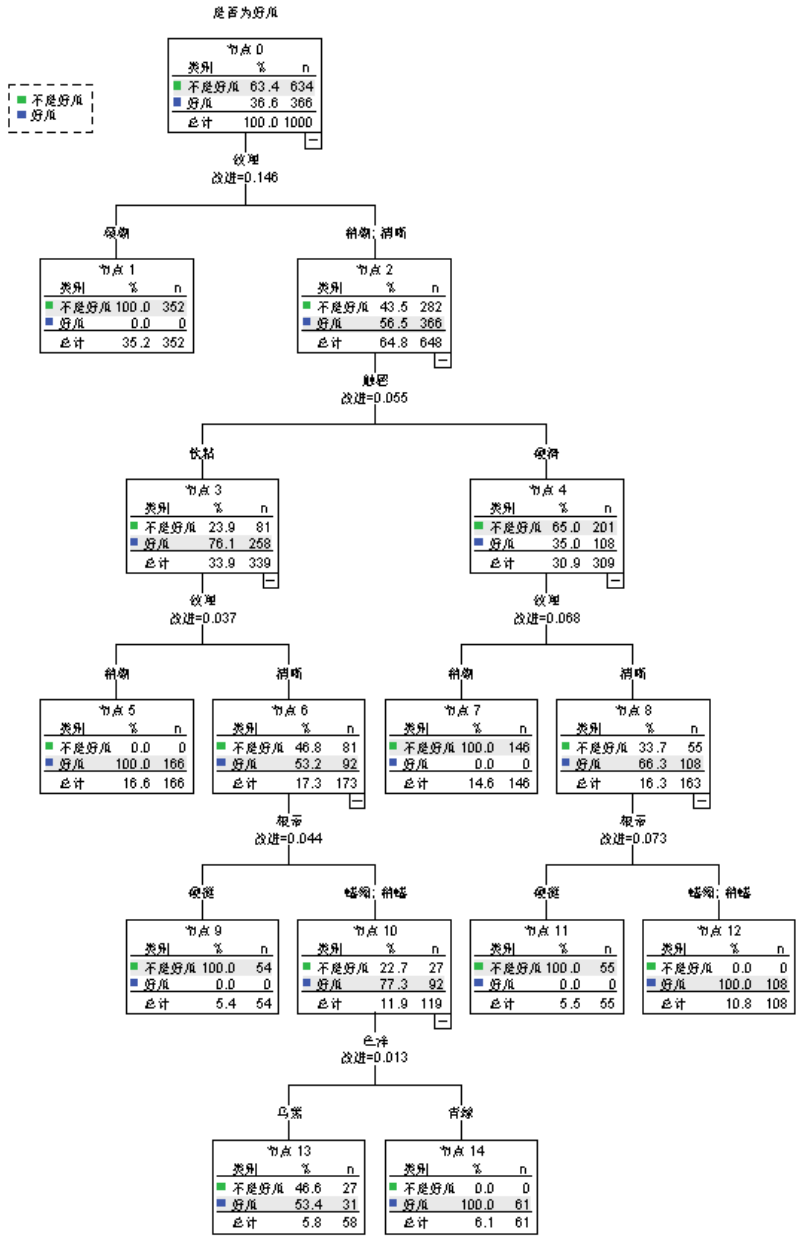
实测	预测		
	不是好瓜	好瓜	正确百分比
不是好瓜	498	136	78.5%
好瓜	0	366	100.0%
总体百分比	49.8%	50.2%	86.4%

2. 穷举 CHAID

观察生成的树发现，穷举 CHAID 与 CHAID 生成的树是完全相同的。查阅相关资料知

这两种方法的区别在于CHAID在分类过程中是通过用户设定阈值来判断是否要将两个类合并或分裂的，而穷举 CHAID 则是通过最值来做出选择的。这里 SPSS 在 CHAID 法时并没有要求输入，应该是有一定的默认参数。此处不作进一步的讨论

3. CRT



模型摘要		
结果	包括的自变量	纹理，触感，色泽，根蒂
	节点数	15
	终端节点数	8
	深度	5

CRT 生长法生成的树包含了所有的自变量，相应的正确率也提高到 97.3%，这已经是一

个比较令人满意的结果。

实测	预测		
	不是好瓜	好瓜	正确百分比
不是好瓜	607	27	95.7%
好瓜	0	366	100.0%
总体百分比	60.7%	39.3%	97.3%

观察树可以发现问题主要出在非二值的属性。对于纹理，树通过两次二分能够较好的解决，可以注意到一直到深度为 3 的树是与 CHAID 生成的树完全一样的。但 CRT 法继续向下分解，增加了树的深度，提高了预测准确率。但这里可以注意到根蒂属性也是三值，却没有经过两次二分，考虑预测结果的偏差便是由此所起。

4. QUEST

观察生成的树发现，QUEST 与 CRT 生成的树是完全相同的。查阅相关资料知这两种方法的区别在于 QUEST 运算过程比 CRT 更简单有效，对于大型 C&R 决策树可以减少分析所需的处理时间，同时减小分类树方法中常见的偏向类别较多预测变量的趋势。可见 QUEST 生长法是对 CRT 的一种优化，此处可能由于数据量的关系没有显示出来，不再作进一步的讨论。

综上，尝试了 SPSS 四种生长决策树的方法，在学习过程中试着阅读了一些使用 python 进行分类处理的代码和机器学习相关书籍的少量章节，深感水平欠缺。SPSS 属于交互良好的统计软件，提供了非常强大的功能，使用感非常不错。但若要尝试去做更加复杂的分类工作，如图像相关等，编程或许还是最优工具。希望未来能有机会学习和尝试。