

数据挖掘课程设计

——白葡萄酒质量分析

班级：信管 151 班

姓名：牟 江 秀

学号：151961

指导老师：李 向 东

2018 年 7 月 5 日

目录

1.数据描述.....	3
2.挖掘思路.....	4
3.基本目标.....	4
4.数据预处理方法和结果.....	5
5.挖掘方法和过程.....	9
6.数据挖掘结果和解释.....	16
7.感悟与体会.....	19

1.数据描述

(1) 本次课设所用的数据主要包括以下列变量：固定酸度、挥发性酸度、柠檬酸、残糖、氯化物、游离二氧化硫、总二氧化硫、密度、pH、硫酸盐、酒精、质量（分数介于 0 和 10 之间），这些变量是关于白葡萄酒的一些成分含量和等级，最后一列变量是实际评定的白葡萄酒的品级。

(2) 以下为本次课设涉及的变量已知的与白葡萄酒生产过程中的作用。

① SO₂ 是唯一在葡萄酒生产中普遍应用的添加剂。按照规定的量添加 SO₂，对于一般人是不会产生健康影响的，超标会对人体产生危害。

② 酸度源于葡萄中的酒石酸和苹果酸以及发酵产生的琥珀酸、乳酸和醋酸。酸度赋予葡萄酒清新、清脆的品尝感。酸有洁净嘴巴和让上颚焕然一新新的效果。

③ 氯化物主要来自于瓶塞污染。瓶塞污染指的是葡萄酒受到了封瓶软木塞的污染，而这种污染是由一种叫做 TCA（某种氯化物）的化学物质所致。当木塞中寄居的真菌接触到酒庄不卫生的环境或消毒残留物中的氯化物时，TCA 就形成了。因此如果酒庄使用带有 TCA 的软木塞，那么酒液也会相应受到一定程度的污染，从而产生了湿木头的味道。

④ 硫酸盐它在葡萄酒的生产和贮藏中占据不可取代的地位。几乎世界

上所有的葡萄酒都添加了亚硫酸盐，而且含量都非常低，含量从 10PPM 到 350PPM 不等（计量单位：百万分之一）。由于硫酸盐具有选择、澄清、抗氧化、增酸、溶解等作用。

2.挖掘思路

（1）判断出哪些属性对于质量变量有明显影响，这里可以使用特征选择来实现。

（2）通过自动分类次来选择最后采用的算法。

（3）最后使用建立模型进行分析，评估模型精确度，不断的改进算法的属性，最后得到精确度最高的模型。用这个模型来改进白葡萄酒在生产过程中的相关变量，以提升白葡萄酒的质量。

3.基本目标

通过一学期的数据挖掘的课程学习，本次课设运用学期到的知识，对数据进行探索和挖掘，回顾与运用知识，从而为实际应用提供参考

通过数据挖掘影响白葡萄酒品级的因素，为以后葡萄酒的生产提供参考，使得酒厂的效益提高提高葡萄酒的品级。

4.数据预处理方法和结果



	标准差 ▲	方差	偏度	tvgsdisp_tab.gif 范围	峰度	利润范围	中位数	众数	唯一	有效
0	0.003	0.000	0.978	0.035	9.794	0.070	0.994	0.992	--	4898
0	0.022	0.000	5.023	0.035	37.565	0.070	0.043	0.044	--	4898
1	0.101	0.010	1.577	0.035	5.092	0.070	0.260	0.280	--	4898
2	0.114	0.013	0.977	0.035	1.591	0.070	0.470	0.500	--	4898
2	0.121	0.015	1.282	0.035	6.175	0.070	0.320	0.300	--	4898
2	0.151	0.023	0.458	0.035	0.531	0.070	3.180	3.140	--	4898
2	0.844	0.712	0.648	0.035	2.172	0.070	6.800	6.800	--	4898
8	1.231	1.514	0.487	0.035	-0.698	0.070	10.400	9.400	--	4898

图 1 使用“数据审核”的节点的结果

由可以看出没有空缺值，但是有极值和离群值，在“数据审核”节点的“质量”选项卡处，选择“强制使极值和离群值无效”，如下图 图 2

[12 个字段] 的数据审核

文件(E) 编辑(E) 生成(G)

审核 质量 注解

完整字段(%): 100% 完整记录(%): 100%

字段	测量	离群值	极值	操作
固定酸度	连续	44	2	强制替换离群值/使极值无效
挥发性酸度	连续	72	9	强制替换离群值/使极值无效
柠檬酸	连续	77	8	强制替换离群值/使极值无效
残糖	连续	8	1	强制替换离群值/使极值无效
氯化物	连续	46	56	强制替换离群值/使极值无效
游离二氧化硫	连续	25	7	强制替换离群值/使极值无效
总二氧化硫	连续	10	2	强制替换离群值/使极值无效
密度	连续	0	3	强制替换离群值/使极值无效
ph	连续	32	0	强制替换离群值/使极值无效
硫酸盐	连续	47	1	强制替换离群值/使极值无效
酒精	连续	0	0	无
质量	有序	--	--	--

图 2 质量选项卡的内容

质量

文件(E) 生成(G) 视图(V) 预览(P)

模型 图形 摘要 设置 注解

排序方式(S): 使用 升序 降序 删除未使用模型 视图: 训练集

是否使用?	图形	模型	构建时间(分钟)	总体精确性 (%)	使用的字段编号
<input checked="" type="checkbox"/>		C5.1	< 1	87.566	9
<input checked="" type="checkbox"/>		CHAID 1	< 1	54.757	9
<input checked="" type="checkbox"/>		类神经网络...	< 1	54.594	9

确定 取消 应用(A) 重置(R)

图 3 处理前的结果

模型 图形 摘要 设置 注解					
排序方式(S): 使用 <input checked="" type="radio"/> 升序 <input type="radio"/> 降序   删除未使用模型 视图: 训练集					
是否使用?	图形	模型	构建时间 (分钟)	总体 精确性 (%)	使用的字段编号
<input checked="" type="checkbox"/>		 C5.1	< 1	84.81	9
<input checked="" type="checkbox"/>		 贝叶斯网...	< 1	56.452	9
<input checked="" type="checkbox"/>		 CHAID 1	< 1	54.349	9

图 4 处理后的结果

处理后生成了超节点，利用超节点进行后续的一系列分析，正确率不升反降，所以在后续的挖掘中，放弃对极值和离群值的处理。如图 3 与图 4 所示，处理之后由 87%降到了 84%。

重新分类字段:

quality

新字段名:

class

重新分类值:

获取

复制

清除新值

自动...

原始值	新值
6	medium
7	high
8	high
9	high






图 5 概念分层

进行概念分层，最后输出的质量总共有 3.4.5.6.7.8.9 这些数值，作为最后的类标号太多，而且 8.9.这两种质量的数据数量过少，进行概念分层，可以使输出的结果简化，而且正确率也会提高。

排名					
	序	字段	测量	重要性	值
<input checked="" type="checkbox"/>	1	酒精	连续	重要	1.0
<input checked="" type="checkbox"/>	2	挥发性酸度	连续	重要	1.0
<input checked="" type="checkbox"/>	3	总二氧化硫	连续	重要	1.0
<input checked="" type="checkbox"/>	4	氯化物	连续	重要	1.0
<input checked="" type="checkbox"/>	5	残糖	连续	重要	1.0
<input checked="" type="checkbox"/>	6	游离二氧...	连续	重要	1.0
<input checked="" type="checkbox"/>	7	固定酸度	连续	重要	1.0
<input checked="" type="checkbox"/>	8	硫酸盐	连续	重要	0.999
<input checked="" type="checkbox"/>	9	柠檬酸	连续	重要	0.997

选定字段数：9 可用字段总数：11

☒ > 0.95
 ☒ <= 0.95
 ☐ < 0.9

2 筛选的字段

	字段	测量	原因
<input type="checkbox"/>	ph	连续	变异系数低于阈值
<input type="checkbox"/>	密度	连续	变异系数低于阈值

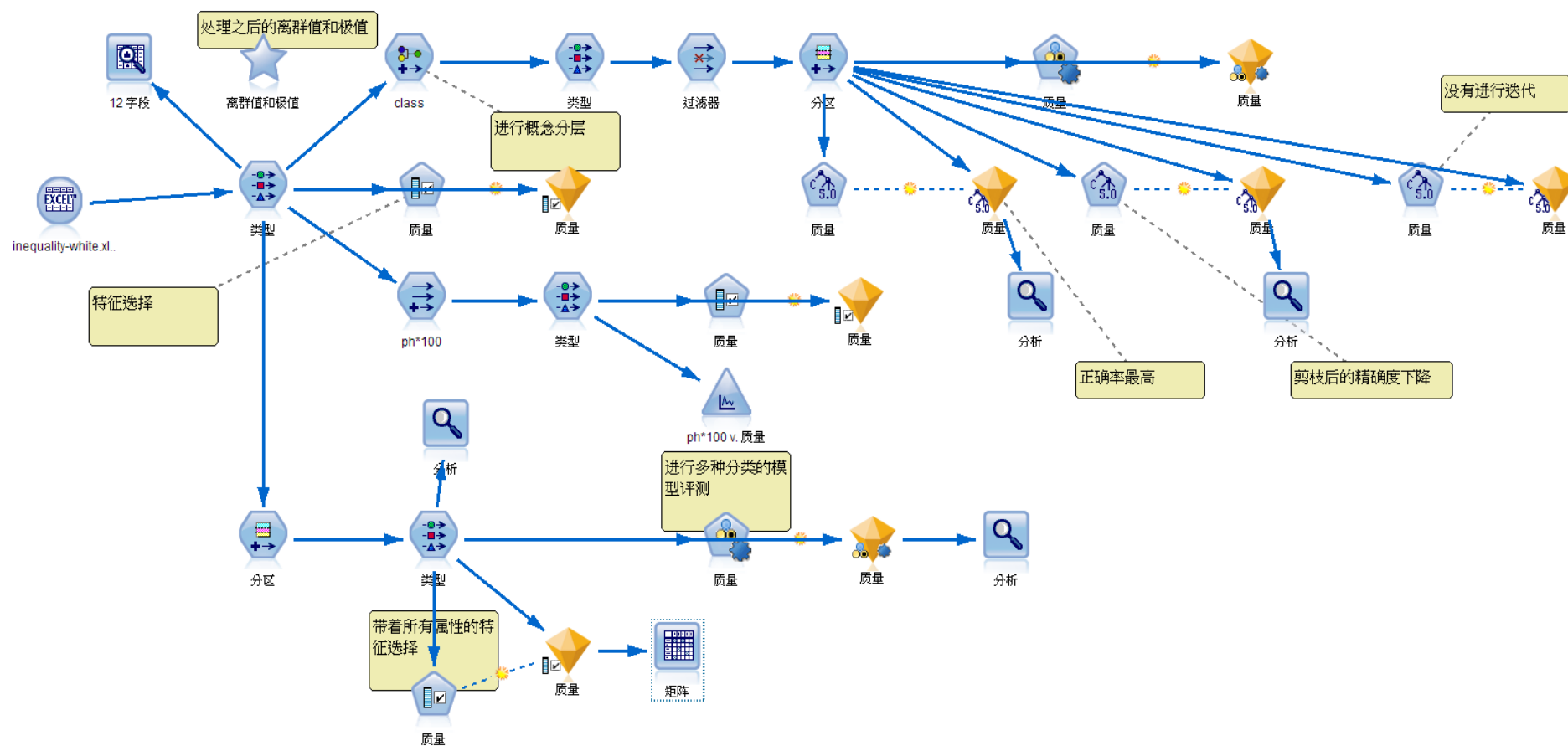
图 6 特征选择

利用“特征选择”节点，得出结果。由上图可以看出，“ph”与“密度”两个变量不会影响最后葡萄酒的质量，所以在后面的数据挖掘过程过滤“ph”与“密度”两个变量。

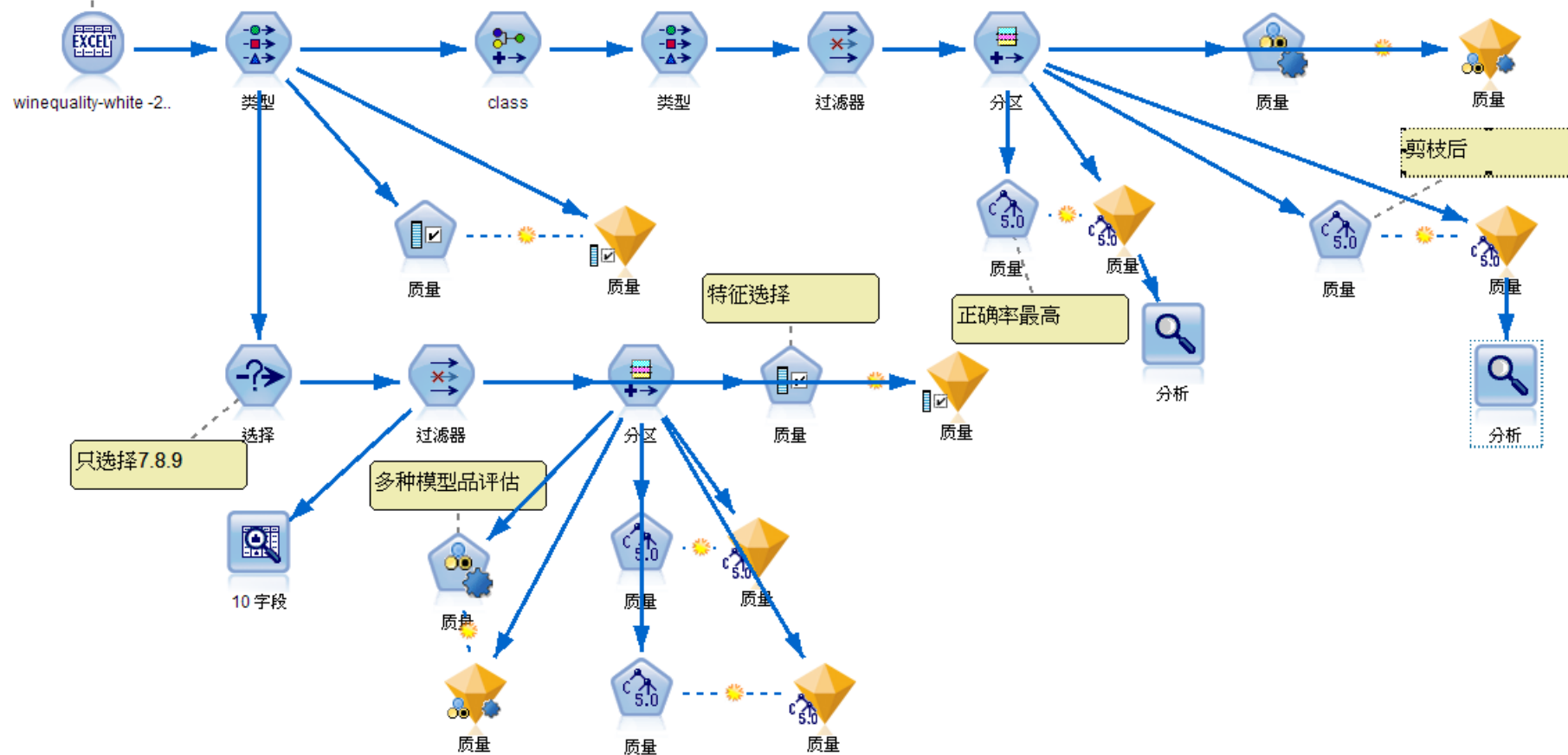
除以上工作外，还对数据进行了填充 8 和 9 的样本较少 8 只有 174 条样本 而 9 只有 5 个样本 9 原有的数据为 5 条 首先将 9 复制为 100 条 8 原有数据 174 条 将 8 复制为 700 条。

对填充后的数据进行数据挖掘，发现最后的模型的精确度比没有填充建立的模型高。

5.挖掘方法和过程



将high等级的7.8.9
填充的更多



(1) 经过预处理之后，通过“自动分类器”节点选择模型


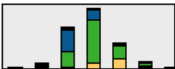
是否使用？	图形	模型	构建时间 (分钟)	总体 精确性 (%)	使用的字段编号
<input checked="" type="checkbox"/>		C5 1	< 1	86.601	9
<input checked="" type="checkbox"/>		CHAID 1	< 1	58.284	8
<input checked="" type="checkbox"/>		贝叶斯网...	< 1	50.842	9

图 7 各模型的评测结果

选取了所有的分类的算法，进行评测。得出最后的结果。根据总体精确度的排序，最后选择 C5.0 算法

(2) 使用 C5.0 来进行分析

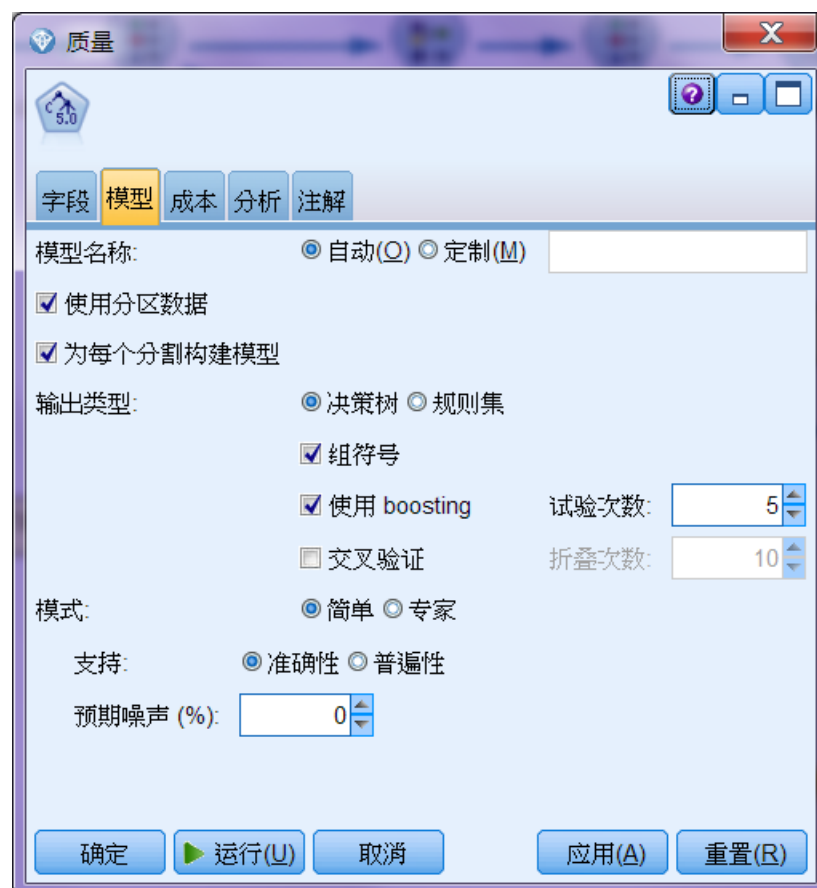


图 8，C5.0 算法中设置 boosting10 次

通过不断迭代，算法的精确度不断提高。

输出字段 质量 的结果

比较 SC-质量 与 质量

分区	1	2
正确	4,352 99.2%	775 68.22%
错误	35 0.8%	361 31.78%
总计	4,387	1,136

SC-质量 的符合矩阵（行表示实际值）

图 7 分析结果

采用分析节点得出精确度为 99.2%

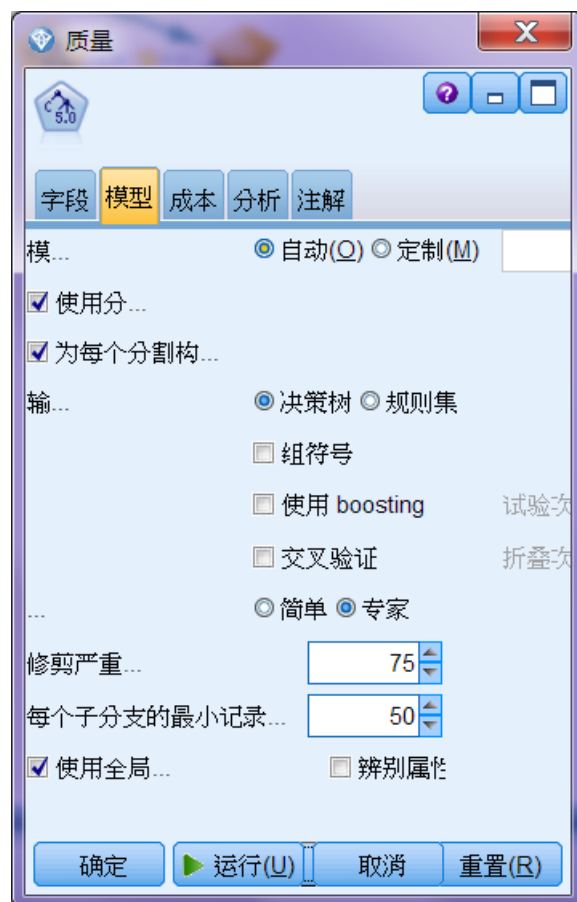


图 8 决策树进行剪枝

输出字段 质量 的结果

比较 \$C-质量 与 质量

"分区"	1		2	
正确	2,141	48.8%	564	49.65%
错误	2,246	51.2%	572	50.35%
总计	4,387		1,136	

图 9 决策树剪枝后的结果

如上图所示，决策树剪枝后，精确度有极大的下降。

(3) 增加训练集的比例

本次课设实验了训练集和测试集的 50%和 50%、60%和 40%、70%和 30%、80%和 20%。经过多次重复实验，得出的结果发现，训练集占的比例越高，精确度越高。

但是如果测试集的量过少的话，最后对模型评测的精确度本身的准确性也是不高的。所以在最终选择的模型当中，采用的 80%和 20%的训练集和测试集的比例。

(4) 增加迭代次数，迭代 20 次之后，精确度有明显的提高。

[质量] 的分析 #19

文件(E) 编辑(E)

分析 注解

全部折叠(C) 全部展开(E)

输出字段 质量 的结果

比较 \$C-质量 与 质量

"分区"	1	2
正确	3,803 97.61%	620 61.88%
错误	93 2.39%	382 38.12%
总计	3,896	1,002

\$C-质量 的符合矩阵 (行表示实际值)

"分区" = 1	3.000000	4.000000	5.000000	6.000000	7.000000	8.000000	9.000000
3.000000	13	1	0	0	0	0	0
4.000000	0	115	7	4	1	0	0
5.000000	0	1	1,162	6	1	0	0
6.000000	0	1	55	1,678	0	0	0
7.000000	0	0	1	10	698	1	0
8.000000	0	0	1	1	1	134	0
9.000000	0	0	0	1	0	0	3

\$CC-质量 的置信度值报告

"分区" = 1	
范围	0.284 - 1.0
平均正确性	0.831
平均正确性	0.564

[质量] 的分析 #18

文件(E) 编辑(E)

分析 注解

全部折叠(C) 全部展开(E)

输出字段 质量 的结果

比较 \$C-质量 与 质量

"分区"	1	2
正确	3,895 99.97%	650 64.87%
错误	1 0.03%	352 35.13%
总计	3,896	1,002

\$C-质量 的符合矩阵 (行表示实际值)

"分区" = 1	3.000000	4.000000	5.000000
3.000000	14	0	
4.000000	0	127	
5.000000	0	0	1
6.000000	0	0	
7.000000	0	0	
8.000000	0	0	
9.000000	0	0	

\$CC-质量 的置信度值报告

"分区" = 1	
范围	

(5) 对每个结果进行评估，选择最佳的模型

虽然进行概念分层会导致最后出的结果不够精细，但是本次课设的目标是提高葡萄酒的质量。当“变量”的值大于等于 7 时，白葡萄酒的质量已经够高，追求过高的质量，相对来说对生产过程的要求也极大的增强，会导致成本大幅提高，对于并不是高档的酒庄来说，这样做是不值得的。

剪枝后的决策树的精确度会下降，在精确度和决策树的复杂度二者均衡。本次课设目标的准确度是 80%，经过多次试验，每个枝最小记录数为 23，剪枝严重性为 50 时，是决策树最简洁与准确度达到目标的最佳设置。

所以最后选取的模型是填充前数据进行概念分层并经过多次迭代并且剪枝之后的模型之后的模型和填充后针对高质量白葡萄酒的单独的分类模型。

(6) 填充后的数据进行同样的处理，

■ 输出字段 质量 的结果

■ 比较 \$C-质量 与 质量

“分区”	1		2	
正确	4,352	99.2%	775	68.22%
错误	35	0.8%	361	31.78%
总计	4,387		1,136	

■ \$C-质量 的符合矩阵（行表示实际值）

图 10 填充过后的数据

可以看出填充过后的数据按照上述设置的 C5.0 算法，精确度也有同样的提高。猜测的影响是之前错误的是质量等级高的数据

(7) 选择 7.8.9 的数据进行 C5.0 模型。

因为上述的猜测是质量为 high 的数据错误率比较高，单独拿出来查看影响因素

输出字段 质量 的结果

比较 \$C-质量 与 质量

分区	1_培训		2_测试	
正确	1,328	100%	347	97.2%
错误	0	0%	10	2.8%
总计	1,328		357	

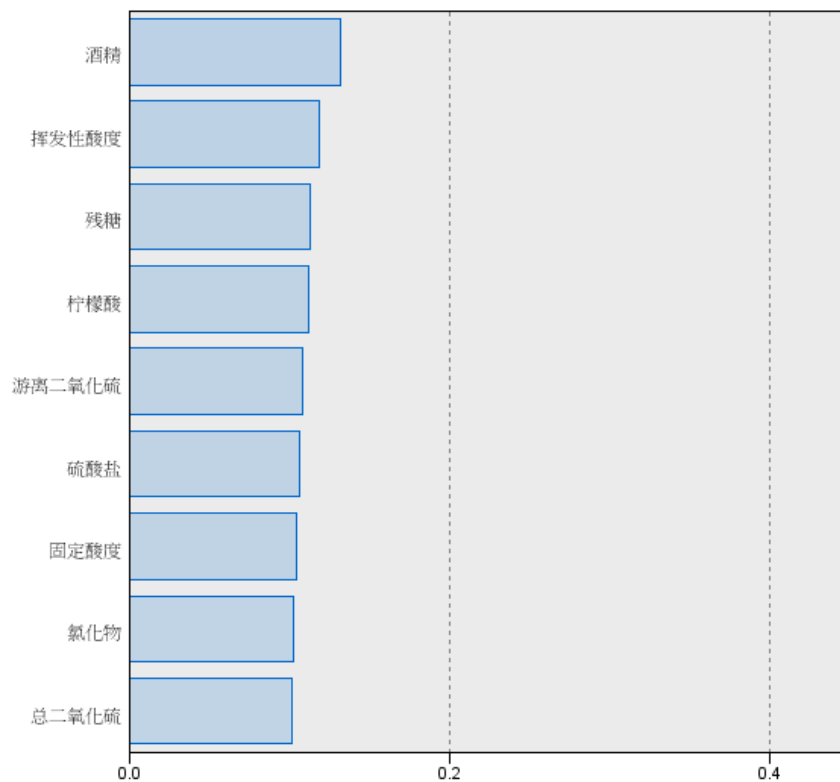
图 11 分析结果

由图看出，精确度到达 100%。可以验证上述的猜想是正确的。

6.数据挖掘结果和解释

预测变量重要性

目标：质量等级



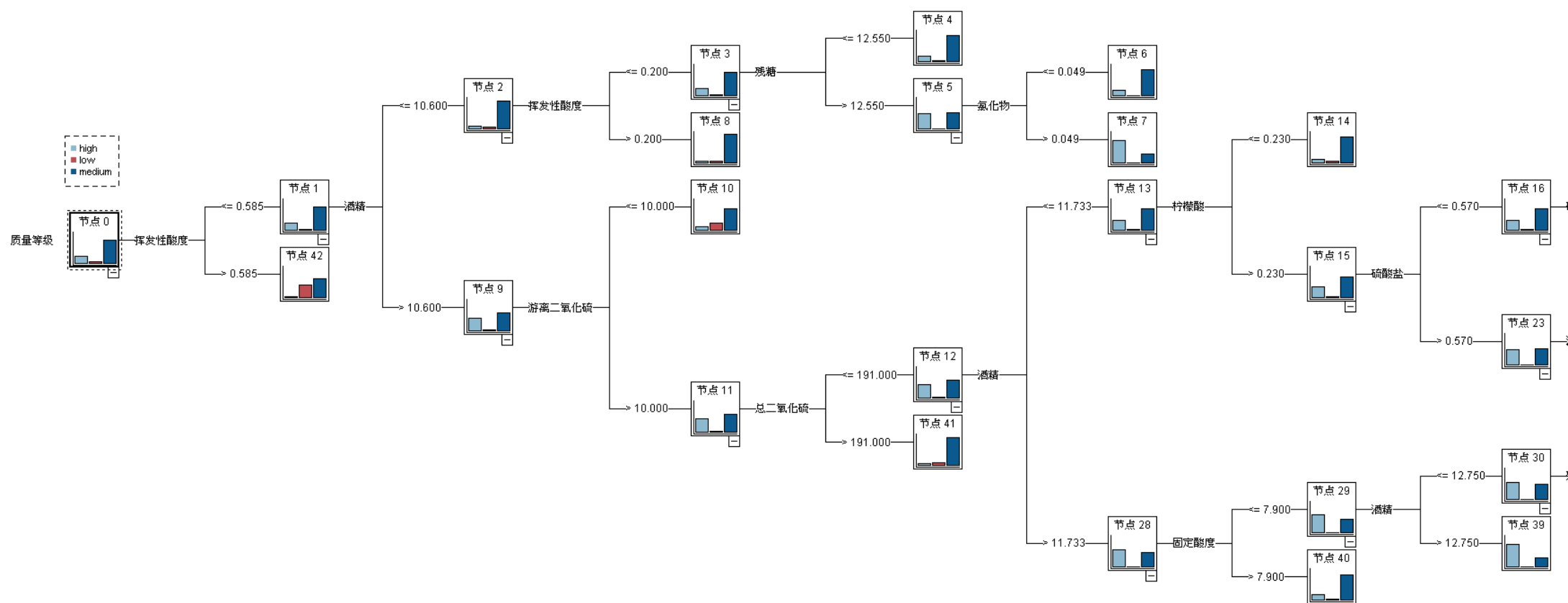


图 12 数据挖掘结果

从上面的图可以看出，当挥发性酸度含量大于等于 0.58，酒的质量为中等，挥发性酸度小于等于 0.58 时，酒精含量大于等于 10.6，游离二氧化硫小于 10 时，葡萄酒的质量为中等。依次类推，可以解读如图的模型

用于生产生活当中，可以得出，酒精含量越高来说，越高的话，葡萄酒的等级越高。游离二氧化硫的含量较高时会产生较高质量的白葡萄酒。氯化物的含量的越低，葡萄酒的质量越好。

所以在生产中，想要提高白葡萄酒的质量，可以增加葡萄的量。其次，可以加入适量的二氧化硫，尽量减少白葡萄酒制作过程中的氯化物的接触。



图 13 只选取了 7.8.9.三个值的模型结果

由图可以看出在白葡萄酒较高等级时，酒精含量小于等于 8.9，游离二氧化硫大于时输出的结果是 8；酒精含量大,12.31 且固定酸度大于 6.5 时，有输出 9.0 的节点。可以看出酒精含量更高时输出的质量更高，更容易产生高质量的葡萄酒。

7.感悟与体会

经过这次的数据挖掘的课程设计，我们小组成员真正的体验到了从数据整理，预处理，统计分析到统计建模的整个流程，对分析的步骤和工作有了更深的了解，对数据挖掘课上所讲的许多方法进行了尝试。总体来说，这次数据挖掘很好的总结了我们在本门课所学习的知识和方法，让我们很有成就感也大有收获。