葡萄酒评价问题模型

摘要

本题是一个统计分析问题，我们利用 SPSS 软件和 Excel 软件对三个附录中的大量数据进行处理，提取出有用信息。分别利用了 Mann-Whitney 模型、方差检验模型、分级模型、组合联系搜索模型、整合三阶迭代回归模型以及 BP 神经网络模型 6 个模型对问题进行求解。

针对问题一，我们首先利用置信区间法和极大-极小标准化方法，分别对数据进行标准化，并通过对比，得出用置信区间法处理后数据的准确性更高。然后我们采用 Mann-Whitney U 检验模型进行显著性的分析，得到两组品酒员的评价结果确实存在显著性差异。接下来，我们又通过比较两个评判组所得 10 个分数的方差，得出第二组结果更可信，我们的评判标准是：对同一个葡萄酒进行评价，10 个品酒员打出的 10 个得分，得分方差小的组更可信。

针对问题二，我们先对理化指标进行整理，通过主成分分析法得到 8 种主成份，然后结合问题一中的葡萄酒质量，应用系统聚类法，将红葡萄聚成 4 类，将白葡萄聚成 3 类，并分别对品质进行了排序，最终得到酿酒葡萄的分级结果。对红葡萄来说，一级葡萄有：1、2、13、25、10、5、24、16、27、19、21、4、22、20、17、14、26，二级葡萄有：9、23，三级葡萄有：3，四级葡萄有：6、7、15、18、12、11、8。对白葡萄来说，一级葡萄有：7、15、18、6、11、10、14、21、4、22、5、20、2、3、9、12、25、28、26、23，二级葡萄有：24、27，三级葡萄有：1、13、8、16、17、19。

针对问题三，我们建立了组合联系搜索模型，先进行酿酒葡萄的理化指标和葡萄酒的理化指标的相关性分析，再进行多元线性回归，并对模型所得结果中不十分理想的进行了调整，用主成分分析调整后的模型得到相关性和显著性较高的理想结果。对调整前后的回归方程筛分汇总，得到了酿酒葡萄与葡萄酒的部分相关理化指标间的函数关系，见文中式（1）、（2）。

针对问题四，由第三问我们已发现酿酒葡萄与葡萄酒的理化指标之间存在一定的线性关系。所以，此问题中我们只需要考虑葡萄酒理化指标对葡萄酒质量的影响。我们针对仅考虑理化指标的情况对葡萄酒质量的影响和考虑理化指标+芳香指标的情况对葡萄酒质量的影响进行对比，并用整合三阶迭代回归模型和 BP 神经网络模型得出的结论进行比较，确定模型的可靠性与高效性，并得出“葡萄酒和葡萄的理化指标对葡萄酒的质量有影响，但不能单纯用葡萄酒和葡萄的理化指标评价葡萄酒的质量。应用葡萄酒和葡萄的理化指标和感官指标共同评价葡萄酒的质量”的结论。

关键词：曼-惠特尼 U 检验　主成分分析　多元线性回归　整合三阶迭代　BP 神经网络

一、 问题重述

确定葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，葡萄酒和酿酒葡萄检测的理化指标会在一定程度上反映葡萄酒和葡萄的质量。附件 1 给出了某一年份一些葡萄酒的评价结果，附件 2 和附件 3 分别给出了该年份这些葡萄酒的和酿酒葡萄的成分数据。请尝试建立数学模型讨论下列问题：

1. 分析附件 1 中两组评酒员的评价结果有无显著性差异，哪一组结果更可信？
2. 根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。
3. 分析酿酒葡萄与葡萄酒的理化指标之间的联系。
4. 分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量？

二、 问题分析

本题目属于一个统计分析问题，我们需要对题中所给葡萄酒品尝评分表、指标总表和芳香物质表进行初始处理，然后从表中提取出我们所需要的信息。

葡萄酒作为一种色香味俱佳的饮品，既能满足人们的感官享受，又具有相当高的营养和保健价值。但是，它与标准产品不同，每一个葡萄酒产区都有其风格独特的葡萄酒，对其评价时没有统一的评价标准。虽然人们在如何利用现代仪器分析以确定葡萄酒的质量方面做了极大的努力，也取得了突破性的进展，但是目前感官评价还仍然是评价葡萄酒感官质量最有效的方法[1]。在葡萄酒的感官评价中，由于品酒员间存在给分区间、给分位置和品尝喜好等方面的差异，导致不同品酒员对同一酒样的评价差异很大，从而不能真实地反映不同酒样间的差异。因此，在对感官评价结果进行统计分析时，必须对品酒员的原始数据进行相应的处理，以真实反映样品间的差异。题中所给的某些数据出现了明显错误，对此我们进行修正：对附录中的错误数据进行修正：附件 1-第一组红葡萄酒品尝评分-酒样品 20-品酒员 4 号对色调的打分出现空缺，修正为此项打分为 0；附件 2-酿酒葡萄-白葡萄理化指标-葡萄样品 1 百粒质量的前两次测定值为 225.8、224.6，最后一次测定值却为 2226.1，修正为 226.1。

对于第一问来说，我们使用两种对数据进行标准化处理的方法，对于品酒员的原始数据进行处理，利用两种标准化方法得到的数据分别进行两组品酒员的显著性分析，得出更可信的结果。对于哪一组更可信，我们采用的评判标准是：对同一个葡萄酒进行评价，10 个品酒员打出的 10 个得分，得分方差小的组更可信。

对于第二问来说，对酿酒葡萄进行分级，也就是将葡萄分成几类，而我们分类的依据就是第一问得出的葡萄酒的分数以及附件 2 中给出的酿酒葡萄的理化指标。但是由于酿酒葡萄一共有 28 个理化指标，再加上葡萄酒的得分这一指标一共有 29 个，指标太多使得分级标准与理化指标间的关系过于复杂。采用科学的方法可以使存在于这些复杂关系中的问题简单化，进而更加清楚地了解它们之间的相互关系。所以我们需要先对酿酒葡萄的 28 个理化指标进行降维处理，然后再通过化简后的指标来进行分类。

对于第三问来说，分析酿酒葡萄与葡萄酒的理化指标之间的联系，也就是研究一组变量对另一组变量的相互关系，我们可以利用统计学中的回归分析方法来实现。我们将葡萄酒的理化指标作为因变量，葡萄的理化指标作为自变量，找出两组变量之间的相关程度，筛选出显著相关的自变量后再对数据进行回归分析，得到回归方程，也即得到了

酿酒葡萄与葡萄酒的理化指标之间的联系。

对于第四问来说，题目要求分析酿酒葡萄、葡萄酒的理化指标这两组因素对葡萄酒质量的影响。显然，同时研究两因素对葡萄酒质量的影响较为复杂而且其影响显著性不便得以体现，为此我们需要对影响因素分别研究。然而在原文中提出“酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系”，并且第三问中我们已发现酿酒葡萄与葡萄酒的理化指标之间存在一定的线性关系。所以，我们只需要考虑一个变量对葡萄酒质量的影响，在此我们可将本问题转化为葡萄酒的理化指标对葡萄酒质量的影响。

三、 模型假设

1．葡萄酒存放得当，不会因外界客观因素影响葡萄酒的品质。

2．对葡萄酒评价时，细则的分类是合理的，也就是说，通过每一小项评价分数的加和，最后得到的某一葡萄酒的分数可以很好地反映出葡萄酒的优劣。

3．表中所给的细则中，每一项后面写的是分数，如：外观分析 15，表示外观分析这一项的总分是 15 分。整个的评分过程采用百分制。

4．每一位评酒师是在同时同地同条件下独立品酒的，不会受到除红酒之外的其他因素影响。

5．酿酒葡萄、葡萄酒的理化指标的测定值，在误差允许范围内是准确的。

6．品质好的葡萄酒是用品质较好葡萄酿成的，对于好葡萄没有酿成好葡萄酒的情况不予考虑。

7．假设酿酒葡萄的理化指标和葡萄酒的理化指标之间呈线性关系。

8．芳香指标仅对葡萄酒的气味有影响，对别的评价项目没有影响。

9．在评价葡萄和葡萄酒的理化指标对葡萄酒质量的影响时，认为理化指标仅与品酒员对葡萄酒外观、口感的评价有关联，葡萄酒的气味不在理化指标的影响范围内。

四、 符号系统

$\overline{x_j}$：每组中 10 个品酒员对酒样品 j 所打分数的平均值

σ_j：每组中 10 个品酒员对酒样品 j 所打分数的标准差

x_{ij}：第 i 个品酒员对酒样 j 的评分

hsg：酿酒葡萄理化指标中的花色苷含量

yjs：酿酒葡萄理化指标中的有机酸含量

zf：酿酒葡萄理化指标中的总酚含量

dn：酿酒葡萄理化指标中的单宁含量

ptzht：酿酒葡萄理化指标中的葡萄总黄酮含量

hbd：酿酒葡萄理化指标中的褐变度含量

bllc：酿酒葡萄理化指标中的白藜芦醇含量

czl：酿酒葡萄理化指标中的出汁率含量

$Y=\begin{pmatrix} y_{11} & \cdots & y_{1j} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{ij} \end{pmatrix}$：主成分矩阵

$A=\begin{pmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kj} \end{pmatrix}$：系数矩阵

$X=\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} \end{pmatrix}$：理化指标矩阵

五、 模型的预备知识

5.1Mann-Whitney U检验的理论分析

1.曼-惠特尼U检验又称“曼-惠特尼秩和检验”，是由H.B.Mann和D.R.Whitney于1947年提出的。它假设两个样本分别来自除了总体均值以外完全相同的两个总体，目的是检验这两个总体的均值是否有显著的差别[4]。

曼-惠特尼秩和检验可以看作是对两均值之差的参数检验方式的T检验或相应的大样本正态检验的代用品。由于曼-惠特尼秩和检验明确地考虑了每一个样本中各测定值所排的秩，它比符号检验法使用了更多的信息。

2.曼-惠特尼U检验的步骤[5]：

第一步：将两组数据混合，并按照大小顺序编排等级。最小的数据等级为1，第二小的数据等级为2，以此类推（若有数据相等的情形，则取这几个数据排序的平均值作为其等级）。

第二步：分别求出两个样本的等级和W1、W2。

第三步：计算曼-惠特尼U检验统计量，n1为第一个样本的量，n2为第二个样本的量：

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

选择U1和U2中最小者与临界值U_α比较（U_α已经整理成为曼-惠特尼U检验的临界值表，见附录1），当U＜ U_α时，拒绝H0，接受H1。

在原假设为真的情况下，随机变量U的均值和方差分别为：

$$E(U)=\frac{n_1 n_2}{2} \qquad D(U)=\frac{n_1 n_2(n_1+n_2+1)}{12}$$

当n1和n2都不小于10时，随机变量近似服从正态分布。

第四步：作出判断：

设第一个总体的均值为μ_1，第二个总体的均值为μ_2，则有：

1)$H_0:\mu_1 \le \mu_2, H_1:\mu_1 > \mu_2$,如果$Z<-Z_\alpha$,则拒绝$H_0$;

2)$H_0:\mu_1 \ge \mu_2, H_1:\mu_1 < \mu_2$,如果$Z>Z_\alpha$,则拒绝$H_0$;

3)$H_0:\mu_1 = \mu_2, H_1:\mu_1 \ne \mu_2$,如果$Z>-\frac{Z_\alpha}{2}$,则拒绝$H_0$。

5.2 主成分分析

主成分分析是把原来多个指标变量划为少数几个综合指标的一种统计分析方法。从数学角度来看，这是一种降维处理技术。在进行降维过程中，要求这些较少的综合指标既能尽量多反映原来较多指标变量所反映的信息，同时它们之间又是彼此独立的。这种对数据进行处理的方法，用 SPSS 软件可以很容易地实现。

通过理化指标矩阵 X，计算可以得到系数矩阵 A 和主成分矩阵 Y 满足：$Y = X \cdot A$。

5.3 z-score 标准化方法

这种方法基于原始数据的均值（mean）和标准差（standard deviation）进行数据的标准化。将 A 的原始值 x 使用 z-score 标准化到 x'。

z-score 标准化方法适用于属性 A 的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况。

新数据=（原数据-均值）/标准差

标准化后的变量值围绕 0 上下波动，大于 0 说明高于平均水平，小于 0 说明低于平均水平。

六、 模型的建立与求解

6.1 关于问题一

在问题分析中我们已经提到，对于葡萄酒的感官评价，因为品酒员间存在给分区间、给分位置和品尝喜好等方面的差异，导致不同品酒员对同一酒样的评价差异很大。通过对附件 1 数据的分析可知，造成品酒员差异的主要原因有：

1．给分区间的差异。比如第一组品酒员对 28 种白葡萄酒打分的时候，3 号品酒员的分数为 75-90，给分区间为 15，而同在第一组的 2 号品酒员的分数为 40-81，给分区间竟为 41。

2．给分位置的差异。同样是第一组品酒员对 28 种白葡萄酒打分，3 号品酒员给分为 75-90，然而第一组的 4 号品酒员给分为 49-75。都是对白葡萄酒评分，但 4 号品酒员的最高分竟然和 3 号品酒员的最低分一样。

3．品尝喜好的差异。比如第二组品酒员对红葡萄酒样品 1 打分，3 号品酒员给分为 80 分，然而 4 号品酒员给分仅为 52 分。一个认为品质很好，一个认为品质很差。

通过以上的分析，可知个人的主观判断是存在很大差异的。在感官评价中，每个品酒员都可以看成一台“分析仪器”，而且它们有各自的准确度和精确度。因此，像对分析仪器进行校正一样，我们必须对品酒员的原始数据进行相应的处理，以真实反映样品间的差异。我们通过比较，选择了两种数据标准化方法，对初始数据进行了标准化处理。

6.1.1 数据预处理

1．第一种标准化方法——置信区间法[2]（推荐使用，原因后述）

为了降低品酒员的异质性，可以计算所有品酒员对同一酒样打分的平均值$(\overline{x_j})$及其标准差(σ_j)，则有品酒员 i 对酒样 j 评价的置信区间为$(\overline{x_j} \pm \sigma_j)$。

如果品酒员 i 对酒样 j 的评分(x_{ij})在置信区间内就可以直接使用；如果其评分(x_{ij})

不在置信区间范围内，则将品酒员的评价 (x_{ij}) 进行逐个调整, 使不同品酒员对同一酒样的评价值都处于 $(\overline{x_j} \pm \sigma_j)$ 范围内，即：

若 $x_{ij} < (\overline{x_j} - \sigma_j)$, 则 $x_{ij} = x_{ij} + \sigma_j$；

若 $x_{ij} > (\overline{x_j} + \sigma_j)$, 则 $x_{ij} = x_{ij} - \sigma_j$。

2．第二种标准化方法——Min-max 标准化（不推荐使用）

min-max 标准化方法是对原始数据进行线性变换。设 minA 和 maxA 分别为属性 A 的最小值和最大值，将 A 的一个原始值 x 通过 min-max 标准化映射成在区间[0,1]中的值 x'，其公式为：

新数据=（原数据-极小值）/（极大值-极小值）

6.1.2 数据处理结果

1．第一种标准化处理结果：

我们对标准化后的数据，求 10 个品酒员对同一种酒的同一样品打分的方差，得到如下两表：

表 1 红葡萄酒方差

	酒品1	酒品2	酒品3	酒品4	酒品5	酒品6	酒品7	酒品8	酒品9
第一组	37.802	20.596	12.034	29.021	14.324	6.724	18.647	9.133	14.567
第二组	21.595	6.402	10.312	11.470	2.121	5.153	27.023	31.165	10.127
	酒品10	酒品11	酒品12	酒品13	酒品14	酒品15	酒品16	酒品17	酒品18
第一组	2.597	30.411	24.204	17.678	16.000	44.044	5.297	20.804	11.699
第二组	13.954	14.379	8.379	8.260	4.575	17.944	5.122	3.803	13.526
	酒品19	酒品20	酒品21	酒品22	酒品23	酒品24	酒品25	酒品26	酒品27
第一组	14.309	7.899	36.422	22.344	12.051	33.635	16.157	14.624	17.214
第二组	9.928	40.404	10.934	11.049	8.306	2.390	18.182	14.835	21.126

表 2　白葡萄酒方差

	酒品1	酒品2	酒品3	酒品4	酒品5	酒品6	酒品7
第一组	25.906	63.556	12.660	20.471	62.086	63.204	13.542
第二组	23.929	19.498	75.567	13.229	7.975	6.210	21.243
	酒品8	酒品9	酒品10	酒品11	酒品12	酒品13	酒品14
第一组	72.934	30.837	101.142	69.462	28.559	31.771	54.105
第二组	9.003	56.416	31.973	22.347	79.363	14.363	7.009
	酒品15	酒品16	酒品17	酒品18	酒品19	酒品20	酒品21
第一组	61.423	93.033	56.686	74.556	12.528	23.091	72.580
第二组	16.894	31.992	7.811	8.236	12.553	23.436	33.034
	酒品22	酒品23	酒品24	酒品25	酒品26	酒品27	酒品28
第一组	39.423	16.757	29.327	9.742	15.920	54.206	39.745
第二组	13.972	6.609	16.611	54.084	65.175	11.235	10.867

2．第二种标准化处理结果：

我们对标准化后的数据，求 10 个品酒员对同一种酒的同一样品打分的方差，得到如下两表：

表 3 红葡萄酒方差

	酒品1	酒品2	酒品3	酒品4	酒品5	酒品6	酒品7	酒品8	酒品9
第一组	0.13742	0.09947	0.17899	0.11243	0.12812	0.11292	0.16578	0.0998	0.12869
第二组	0.10444	0.11265	0.08507	0.11437	0.11286	0.10777	0.08578	0.08931	0.10052
	酒品10	酒品11	酒品12	酒品13	酒品14	酒品15	酒品16	酒品17	酒品18
第一组	0.11875	0.07863	0.10927	0.07189	0.08163	0.12658	0.14959	0.10465	0.13078
第二组	0.29899	0.13164	0.08693	0.10617	0.09045	0.11453	0.10255	0.07576	0.08043
	酒品19	酒品20	酒品21	酒品22	酒品23	酒品24	酒品25	酒品26	酒品27
第一组	0.10743	0.08038	0.15926	0.11479	0.11242	0.15473	0.14654	0.08667	0.10284
第二组	0.13789	0.13518	0.1096	0.08397	0.17199	0.10722	0.07593	0.18469	0.06327

表 4 白葡萄酒方差

	酒品1	酒品2	酒品3	酒品4	酒品5	酒品6	酒品7
第一组	0.07978	0.10386	0.05128	0.08452	0.09236	0.10698	0.12088
第二组	0.00367	0.07258	0.08078	0.12994	0.0811	0.0033	0.09564
	酒品8	酒品9	酒品10	酒品11	酒品12	酒品13	酒品14
第一组	0.12071	0.10307	0.08858	0.10041	0.09452	0.10159	0.08813
第二组	0.10769	0.08675	0.08371	0.16602	0.08753	0.12955	0.08101
	酒品15	酒品16	酒品17	酒品18	酒品19	酒品20	酒品21
第一组	0.08652	0.10091	0.09985	0.08874	0.116	0.10304	0.11355
第二组	0.08647	0.07114	0.10653	0.10461	0.09012	0.08007	0.08834
	酒品22	酒品23	酒品24	酒品25	酒品26	酒品27	酒品28
第一组	0.07861	0.09899	0.13213	0.09384	0.11193	0.09494	0.13968
第二组	0.10132	0.116	0.11896	0.08218	0.11433	0.08062	0.07833

6.1.3 Mann-Whitney 模型（模型一）

我们利用标准化后的数据，求 10 个品酒员对同一种酒的同一样品打分的平均值，一共得到 2*4*27 个数据，“2”表示两种标准化方法，“4 表示”附件 1 的四个工作表（第一组红葡萄酒品尝评分、第一组白葡萄酒品尝评分、第二组红葡萄酒品尝评分、第二组白葡萄酒品尝评分），“27”表示酒的品种（当为白酒时是 28 个）。

将这些标准化后的平均值，输入到 SPSS19.0 软件中，利用“分析-非参数检验-旧对话框-两个独立样本”命令，对第一组打分、第二组打分进行 Mann-Whitney U 检验，得到如下结果：

表 5 Mann-Whitney U 检验结果

假设检验汇总

	原假设	测试	Sig.	决策者
1	VAR00001 的分布在 VAR00002 类别上相同。	独立样本 Mann-Whitney U 检验	.033	拒绝原假设。
2	VAR00003 的分布在 VAR00002 类别上相同。	独立样本 Mann-Whitney U 检验	.016	拒绝原假设。
3	VAR00004 的分布在 VAR00002 类别上相同。	独立样本 Mann-Whitney U 检验	.031	拒绝原假设。
4	VAR00005 的分布在 VAR00002 类别上相同。	独立样本 Mann-Whitney U 检验	.897	保留原假设。

显示渐进显著性。显著性水平是 .05。

其中：

VAR00001是用第一种标准化方法后，红葡萄酒的第一组、第二组评分的所有值列为一列

VAR00003是用第一种标准化方法后白葡萄酒的第一组、第二组评分的所有值列为一列

VAR00004是用第二种标准化方法后红葡萄酒的第一组、第二组评分的所有值列为一列

VAR00005是用第二种标准化方法后白葡萄酒的第一组、第二组评分的所有值列为一列

VAR0002 是我们自己对红葡萄酒的第一组、第二组评分的标记，如果 VAR0001 的某个数据是第一组的评分，那么 VAR0002 对应写“1”，如果是第二组的评分，则写“2”。

Sig 是显著性水平，如果 Sig<0.05，则拒绝原假设，如果 Sig>0.05，则保留原假设。

结论：

从图表结果中可以看出，第一种标准化方法得出红葡萄酒、白葡萄酒的 Sig 都小于 0.05，拒绝原假设，则说明两组品酒员评价结果有显著性差异；第二种标准化方法得出红葡萄酒的 Sig<0.05，拒绝原假设，说明红葡萄酒的第一、第二组数据有显著性差异，白葡萄酒的 Sig>0.05，无显著性差异，结论没有第一种标准化处理后的结论好。

6.1.4 方差检验模型（模型二）

对于第一组品酒员的评价结果、第二组品酒员评价结果哪个更可信这一问题，我们认为：10 个人对于同一种酒的同一样品进行打分，经过标准化后的 10 个分数方差小的组，结果更可信，也即分数越稳定的结果越可信。这是因为，葡萄酒的品质好坏是一个客观事实，如果因为不同品酒员的主观判断差异过大，导致数据的上下波动，最后所得的 10 个品酒员所打分数的平均值是不能很客观地反映葡萄酒的品质的。

我们利用 5.1.2 中的数据处理结果，对两组品尝结果中方差进行大小比较，经过统计发现如下结果：

表 6　第一种标准化方差比较

第一种标准化		
	第一组方差小的个数	第二组方差小的个数
红葡萄酒	8个	19个
白葡萄酒	8个	20个

表 7　第二种标准化方差比较

第二种标准化		
	第一组方差小的个数	第二组方差小的个数
红葡萄酒	11个	16个
白葡萄酒	9个	19个

从上述结果中跟我们看出，无论是用第一种标准化方法，还是第二种标准化方法所得的结果都是相同的，即对红葡萄酒和白葡萄酒都是第二组品酒员所打分数的方差总体比较小，也即对于葡萄酒评价的波动性小，所得结果更可信。

6.1.5 两种标准化方法结果比较

1.通过 Mann-Whitney 模型的计算，第一种标准化所得结果是红、白葡萄酒的 Sig 均小于 0.5，而第二种标准化所得结果是红葡萄酒 Sig<0.5，而白葡萄酒 Sig>0.5，第二种标准化结果没有第一种标准化结果显著。

2.通过方差检验模型的计算，第一种标准化结果得到的一、二两组的方差，比第二种标准化结果的一、二组方差区分地更开，说明第一种标准化更好。

3.根据参考文献[2]，用第一种标准化方法——置信区间法处理后的数据，品酒员方差降低，酒样间的方差明显提高，这正是我们想要的结果。因为不同样本的酒进行品尝，酒样间的方差越大说明酒样区分地越开；二对同一种酒样进行品尝，品酒员间方差越小，说明他们的意见越一致，所得葡萄酒的品质评价结果越可信。

综合分析以上三点，我们认为第一种标准化方法更能真实地反映葡萄酒的品质，所以下面的问题我们采用第一种标准化处理得到的第二组评价数据进行计算（第一种标准化所得葡萄酒品质分数的排序见附录 2）。

6.2 关于问题二——分级模型（模型三）

6.2.1 酿酒葡萄理化指标的变量整理

通过查找大量的资料，我们先对酿酒葡萄的 28 个理化指标进行整理，对其做了如下的调整：

1.将酒石酸、苹果酸、柠檬酸统一归为有机酸，有机酸含量=酒石酸+苹果酸+柠檬酸。

2.将果穗质量、百粒质量归为颗粒数指标，也即 颗粒数指标 $=\dfrac{\text{果穗质量}}{\text{百粒质量}}$。果穗质量是指一串葡萄的质量，百粒质量是指一百粒葡萄的质量，用果穗质量除百粒质量，得到一串葡萄上有几组一百粒葡萄，将此作为颗粒数指标。

3.将果皮颜色的 L*、a*、b*三个指标归为色差 $\triangle E$。附件中每个葡萄样本的 L*、a*、b*三个指标都是进行了三次试验得到的三组值，我们先取这三组数据的横向平均值 L_i、a_i、b_i（i=1，2，…27），然后再取 27 组 L、a、b 的竖向平均值得到 $\overline{L}$、$\overline{a}$、$\overline{b}$，那么 $\triangle E=\sqrt{(L_i-\overline{L})^2+(a_i-\overline{a})^2+(b_i-\overline{b})^2}$。

经过整理后，酿酒葡萄的理化指标一共有 25 个。

6.2.2 酿酒葡萄的主成分分析（PCA）

主成分分析(PCA)是将多项指标重新组合成一组新的互相无关的几个综合指标，根据实际需要从中选取尽可能少的综合指标，以达到尽可能多的反映原指标信息的分析方法[3]。

首先我们将葡萄的 25 个理化指标的数据用 SPSS19.0 软件中的“分析--描述统计--描述--将标准化得分另存为变量”命令，进行标准化处理（使用的是 z-score 标准化方法），将其化为围绕 0 上下波动的数值，大于 0 说明高于平均水平，小于 0 说明低于平均水平。

然后，我们对标准化后的数据用 SPSS19.0 软件中的“分析--降维--因子分析”命令

进行主成分分析，得到了主成分分析方差分解表和成分矩阵表。通过 SPSS 软件进行计算，将 25 个理化指标转化成为 8 种主成分 Y1-Y8。并进一步根据方差分解表和成分矩阵表计算出系数矩阵 A 和主成分矩阵 Y 的值。

6.2.3 系统聚类

八种主成分加上第一问算出的葡萄酒品质得分一共 9 个指标，将这些数据输入到 SPSS 软件中，利用“分析--分类--系统聚类”命令作系统聚类分析，得到了红葡萄和白葡萄的聚类树状图，下面我们先对红葡萄进行分析。红葡萄的聚类树状图如图 1 所示，通过观察，我们把红葡萄分成 4 类。然后将每组葡萄所酿制的葡萄酒的分数作为评价该组葡萄质量等级的标准。为了便于观察，我们将同一组的数据涂上同一种颜色，再根据问题一的葡萄酒质分数进行降序排列结果见下表 8。

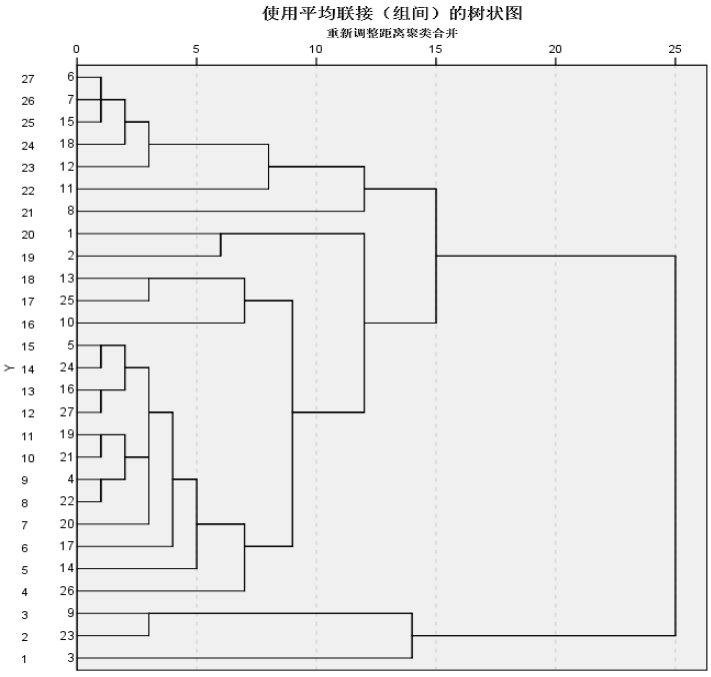


图 1　红葡萄聚类树状图

表8　红葡萄等级

红葡萄	品质分数
9	78.2
23	77.1
20	75.8
3	74.6
17	74.5
2	74
14	72.6
19	72.6
21	72.2
5	72.1
26	72
22	71.6
24	71.5
27	71.5
4	71.2
16	69.9
10	68.8
13	68.8
12	68.3
25	68.2
1	67.4
6	66.3
8	66
15	65.7
18	65.4
7	65.3
11	61.6

表 8 中，分布在列表上方的为品质最好的，分布在列表下发的为品质较差的。这样就可以确定葡萄的等级好坏。最终的分类结果及排序见下表 9。

表 9　红葡萄等级分类及排序

	排序	酿酒葡萄编号
第一类	4	6、7、15、18、12、11、8
第二类	3	1、2、13、25、10、5、24、16、27 19、21、4、22、20、17、14、26
第三类	1	9、23
第四类	2	3

从上面的分类结果可以看出，红葡萄的分级是很明显的，每种等级葡萄的品质差异较大，这点在表 8 的颜色分布上可以很直观地看出。排序号码代表着葡萄品质的优劣，其中，第三类是品质最优的葡萄，第四类品质良好，第二类品质一般，第一类品质最差。

白葡萄的聚类分析及排序进行相似处理，得到如下图 2、表 10、表 11 所示：

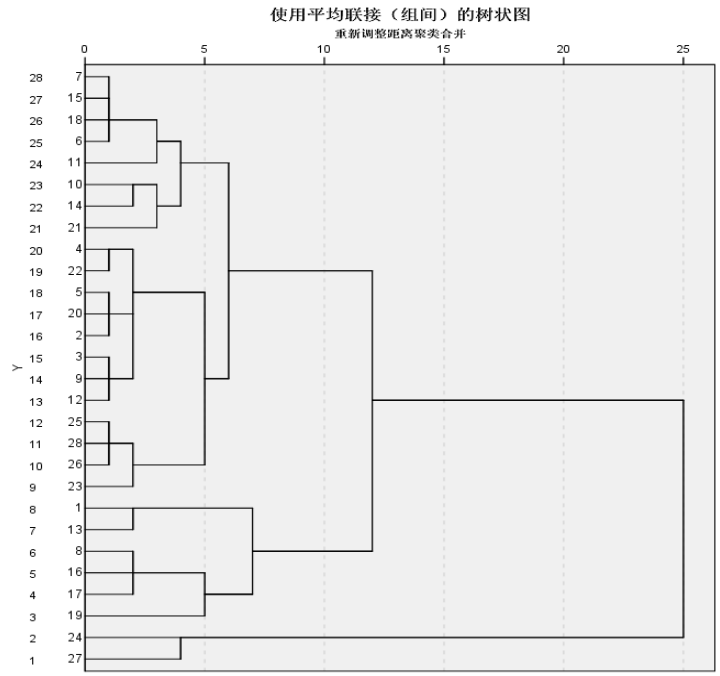


图 2　白葡萄聚类树状图

表 10　白葡萄等级

白葡萄	品质分数
5	81.5
9	80.4
17	80.3
10	79.8
28	79.6
25	79.5
22	79.4
21	79.2
15	78.4
1	77.9
23	77.4
14	77.1
27	77
4	76.9
18	76.7
20	76.6
19	76.4
24	76.1
2	75.8
3	75.6
6	75.5
26	74.3
7	74.2
13	73.9
12	72.4
8	72.3
11	71.4
16	67.3

表 11　白葡萄等级分类及排序

	排序	酿酒葡萄编号
第一类	1	7、15、18、6、11、10、14、21、4、22 5、20、2、3、9、12、25、28、26、23
第二类	3	1、13、8、16、17、19
第三类	2	24、27

关于白葡萄的等级分类，明显没有红葡萄的区分度大。白葡萄的等级相互掺杂，这一点从表 10 中可以直观看出。针对此种情况，我们将每类中葡萄的品质分数的平均值求出，利用平均值对类别进行排序。最终得到白葡萄的优劣，即第一类是品质最优的葡萄，第三类品质一般，第二类品质最差。

6.3 关于问题三

我们将葡萄酒的理化指标作为因变量，葡萄的理化指标作为自变量，找出两组变量之间的相关程度，筛选出有意义的自变量后再对数据进行回归分析，就可以得到回归方程，也即得到了酿酒葡萄与葡萄酒的理化指标之间的联系。

6.3.1 组合联系搜索模型（模型四）

一、相关性分析

将问题二中整理得到的 25 个酿酒葡萄的理化指标和葡萄酒的 7 个理化指标输入到 SPSS19.0 软件中，利用“分析-相关-偏相关”对于这两组变量进行相关性分析，得到了他们的相关性值（结果见附录 3）。

对于红葡萄，取相关性>0.6 时认为相关性显著，即该自变量与因变量确实存在某种关系，选择该自变量加入到回归方程的求解中。自变量有：花色苷、有机酸、 DPPH、总酚、单宁、葡萄总黄铜。

对于白葡萄，取相关性>0.5 时认为相关性显著，认为该自变量与因变量确实存在某种关系，选择该自变量加入到回归方程的求解中。但是白葡萄酒的不同理化指标受葡萄的理化指标影响差异较大，我们就分开描述。

二、回归方程的求解——理化指标法

用根据相关性挑选出的几个自变量，与因变量做回归线性方程拟合。用 SPSS19.0 中的“分析-回归-线性”命令，得到了各线性回归方程的系数值（结果见附录 4）。

经过整理，回归方程如下：

红葡萄酒回归方程

$$\begin{cases} \text{花色苷} = -22.248 - 8.012\text{hsg} + 1.920\text{yjs} + 0.159\text{DPPH} + 226.717\text{zf} - 7.878\text{dn} + 4.966\text{ptzht} \\ \text{单宁} = 1.241 - 0.075hsg + 0.004\text{yjs} + 0.001\text{DPPH} + 5.492\text{zf} + 0.163\text{dn} + 0.08\text{ptzht} \\ \text{总酚} = 0.716 - 0.096\text{hsg} + 0.005\text{yjs} + 5.813\text{zf} + 0.147\text{dn} + 0.055\text{ptzht} \\ \text{酒总黄酮} = -1.178 - 0.002\text{hsg} + 0.001\text{DPPH} + 0.878\text{zf} + 0.355\text{dn} + 0.007\text{ptzht} \\ \text{DPPH} = -0.051 - 0.003\text{hsg} + 0.157\text{zf} + 0.013\text{dn} + 0.002\text{ptzht} \end{cases}$$

白葡萄酒回归方程

$$\begin{cases} \text{单宁}=0.856+0.178dn \\ \text{总酚}=0.689+0.108dn+0.096ptzht \\ \text{酒总黄酮}=-4.51+0.001-0.037ptzht+0.558bllc-0.204zf \\ \text{色差}=7.477-0.29czl \end{cases}$$

三、回归方程显著性检验

我们根据 SPSS 输出的调整 R 方一值，对回归方程检验。经观察我们发现，红葡萄的调整 R 方均大于 0.5，并且大部分都大于 0.7，这样的结果已经很接近 1 了，说明回归方程很显著。但是白葡萄的调整 R 方依次为 0.526、0.590、0.681、-0.037，最后一个调整 R 方为色差的，结果很差，可以删掉。而剩下的三个数值也未超过 0.5，我们认为其显著性不是很明显，故结合其他的思路进行求解。

四、模型的调整——主成分法

我们利用第二问中得到的 8 种主成分值，将其对葡萄酒的理化指标进行相关性分析，找到与理化指标相关性大的主成分，然后与葡萄的理化指标进行线性回归方程模拟，得到了如下的白葡萄理化指标回归方程：

白葡萄酒回归方程：

$$\begin{cases} \text{单宁}=1.851+0.083y_1+0.134y_2+0.161y_3+0.2227y_6 \\ \text{总酚}=1.456+0.066y_1+0.12y_2+0.118y_3 \\ \text{酒总黄酮}=1.581+0.424y_2+0.374y_3+0.023y_6 \\ \text{色差}=5.410-0.882y_2-0.839y_8 \end{cases}$$

其中，y_i为 8 种主成分。对该回归方程进行显著性检验，发现其调整 R 方依次为 0.717、0.685、0.667、0.515，比起上一组方程来说，调整 R 方都相应有所提高，显著性水平提高。而用此种方法做出的红葡萄回归方程的显著性，没有直接用理化指标求得的显著性明显，所以就不再赘述。

其余未列出回归方程的葡萄酒的理化指标可认为与酿酒葡萄的理化指标无显著性关联。

五、酿酒葡萄与葡萄酒的理化指标的联系

红葡萄酒用理化指标法求得的方程：

$$\begin{cases} \text{花色苷}=-22.248-8.012hsg+1.920yjs+0.159DPPH+226.717zf-7.878dn+4.966ptzht \\ \text{单宁}=1.241-0.075hsg+0.004yjs+0.001DPPH+5.492zf+0.163dn+0.08ptzht \\ \text{总酚}=0.716-0.096hsg+0.005yjs+5.813zf+0.147dn+0.055ptzht \\ \text{酒总黄酮}=-1.178-0.002hsg+0.001DPPH+0.878zf+0.355dn+0.007ptzht \\ DPPH=-0.051-0.003hsg+0.157zf+0.013dn+0.002ptzht \end{cases} \quad \cdots(1)$$

而白葡萄酒用主成分法求得的方程：

$$\begin{cases} \text{单宁}=1.851+0.083y_1+0.134y_2+0.161y_3+0.2227y_6 \\ \text{总酚}=1.456+0.066y_1+0.12y_2+0.118y_3 \\ \text{酒总黄酮}=1.581+0.424y_2+0.374y_3+0.023y_6 \\ \text{色差}=5.410-0.882y_2-0.839y_8 \end{cases} \quad \cdots(2)$$

用两种方法联合求得的结果作为酿酒葡萄与葡萄酒的理化指标之间的联系，保证了回归方程显著性的满足，具有较好的合理性。

6.4 关于问题四

在问题分析中我们已经提到，因为第三问中我们已发现酿酒葡萄与葡萄酒的理化指标之间存在一定的线性关系。所以，我们只需要考虑一个变量对葡萄酒质量的影响，即我们只用求葡萄酒的理化指标对葡萄酒质量的影响。

6.4.1 仅考虑理化指标的情况

我们采取如下方式验证“能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量”：

首先，将品酒员的打分划分为芳香分数和外观口感分数，再对外观口感分数与葡萄酒的理化指标进行相关性分析。经过 SPSS 软件计算，得到与红葡萄酒外观口感分数相关性较大的理化指标为：花色苷、单宁、总酚、酒总黄铜、白藜芦醇、DPPH 半抑制体积、色差，与白葡萄酒外观口感分数相关性较大的理化指标为：单宁、总酚、酒总黄铜、白藜芦醇、DPPH 半抑制体积、色差。

找出显著相关量后，对与外观口感分数显著相关的理化指标与得分做多元线性回归，得到回归方程后用原理化指标计算新的外观口感得分，我们称之为理化得分 M。之后我们对这个得分 M 进行等比例放大：

$$M'=\frac{M}{0.6}$$

之所以选择比例系数为$\frac{1}{0.6}$，是因为在 100 分制中外观和口感总分为 60 分。

以M'作为整体分数，并与原分数比较。定义得分差值率=$\frac{M'-\mathrm{N}}{\mathrm{N}}$，N 为第一问中每种葡萄酒样品的总得分。通过对该得分差值率的分析研究，我们发现，红葡萄酒中小部分葡萄酒的得分差值率超过了 10%，白葡萄酒中所有的得分差值率均超标，因此，单纯的用葡萄酒和葡萄的理化指标来评价葡萄酒的质量是不可行的（具体得分差值率见表 12）。

6.4.2 考虑理化指标+芳香指标的情况

我们引入芳香分数，由其与外观、口感分数共同评价葡萄酒的质量，或者说用芳香物质这一感官指标与葡萄酒的理化指标共同完成对葡萄酒的评价。同样，我们先把芳香物质进行分类，将同一类别的芳香物质归为一类进行加和，作为芳香物整合指标。然后我们通过相关性分析，得出与红葡萄酒芳香分数相关性较大的芳香物整合指标为：醛、乙醚、呋喃，与白葡萄酒芳香分数相关性较大的芳香物整合指标为：烷、乙醚、联苯。

多元回归拟合得出相应方程，并用该方程求出新的芳香分数 P。按照附件一中的比例关系，求出新的葡萄酒的分数 Q：

$$Q = M + P$$

并将该得分与第一问中品酒员的评分 N 进行比较，求得得分差值率。通过观察我们发现红、白葡萄酒的新的分数与原得分基本吻合，只有极少数数据存在差异，但差异不大。

所以我们进一步得出结论，葡萄和葡萄酒的的理化指标对葡萄酒的质量有影响，但是不能仅用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。反而用芳香分数与外观后感分数共同评价葡萄酒的质量更为准确。

6.4.3 以上两种情况的比较论证

一．整合三阶迭代回归模型（模型五）

我们首先对所有的已分类芳香物质进行葡萄酒芳香得分的相关性分析，发现显著效果一般，所以我们采取逐步迭代的方法进行回归。

具体流程如下：

第一步，对得分与所有分类后的芳香物质进行多元线性回归，找出系数相对较小的，将其去除；

第二步，用剩余指标再次对得分进行多元线性回归，此时我们发现，迭代后的 R 方值相应增大。迭代次数越多，R 方值越大，显著性越好。依据这个原理，我们尽可能多次进行迭代，直到各芳香物质系数的数量级相差不大为止；

本题中做三次迭代即可。迭代方程依次为：

红葡萄：

三次迭代

得分=23.272-6.23 醛+0.168 乙醚-0.074 呋喃-0.029 烯-0.043 甘油-0.027 苯酚

得分=22.696-0.524 醛+0.146 乙醚-0.110 呋喃+0.004 烯-0.034 甘油

得分=22.790-0.505 醛+0.092 乙醚-0.999 呋喃

白葡萄：

三次迭代

得分=20.812+0.014 醇+0.180 烷+0.005 醛+0.05 烯+0.218 乙醚-0.396 联苯

得分=22.115+0.139 烷+0.344 乙醚-0.309 联苯+0.041 烯

得分=22.200+0.139 烷+0.334 乙醚-0.318 联苯

根据上述红葡萄和白葡萄的最后一次迭代方程式，算出红葡萄和白葡萄的芳香得分 P。再加上理化指标得分 M，得到总得分 Q。那么，得分差值率 $=\frac{Q-N}{N}$。通过观察我们发现红葡萄酒只有两组数据的得分差值率超过 10%，而白葡萄酒的新的分数与原得分完全吻合（具体得分差值率见表 12）。说明整合三阶迭代回归模型能够很好地使用理化指标和芳香指标共同来评价葡萄酒的质量。

二．BP 神经网络模型（模型六）

用 Matlab BP 神经网络工具箱，对得分与各整合后的芳香物质指标进行训练，得出相应的芳香分数值。

根据训练得出的芳香分数值 Q，得到得分差值率。通过观察发现，红葡萄酒只有一组数据的得分差值率超过 10%，而白葡萄酒的新的分数与原得分完全吻合（具体得分差

值率见表 12）。

注：加“*”号的是得分差值率超过 10%，不在合理性范围之内。

表 12　得分差值率

红葡萄酒		
理化指标独立得分差值率	整合三阶迭代回归得分差值率	BP神经网络得分差值率
-0.005016583	0.028695878	0.003904092
-0.042259064	0.005839629	0.058878169
-0.034584205	0.006827009	0.007265595
-0.075354942	0.000692409	0.022431679
-0.082159896	-0.005052899	0.008878489
0.02199955	0.107765933*	-0.077169156
-0.01455813	0.085539292	-0.050231086
0.011235511	0.086794066	-0.063880037
0.113349548*	-0.061588635	0.078251451
-0.035184674	0.037152462	0.005137533
0.029209334	0.100120254*	-0.092452352
-0.035433715	0.017388376	-0.023196504
-0.046764388	0.035470708	0.000789629
0.100347004*	-0.025890451	0.039718418
-0.039461594	0.062363341	-0.012490567
-0.078729949	0.008617686	0.034374574
0.097997919	-0.026494894	0.054419767
-0.013450351	0.091455638	-0.030265289
-0.03370385	0.01834006	-0.040983784
0.142542122*	-0.065516379	0.129523705*
-0.034843663	0.025075767	-0.052912128
-0.051876937	0.0095051	-0.001669833
-0.054397533	-0.011455957	0.031006741
-0.046678784	0.016431353	-0.012549313
-0.038815073	0.03895754	-0.000925351
0.106274163*	-0.027379579	0.049717071
-0.057183147	0.013181476	0.015052519

白葡萄酒		
理化指标独立得分差值率	整合三阶迭代回归得分差值率	BP神经网络得分差值率
0.587635043*	0.006527196	-0.022613681
0.543518082*	-0.000316191	0.022408254
0.943682619*	-0.007660633	-0.00956131
0.728737839*	0.012957956	0.037408961
0.657597121*	-0.023384083	-0.014677925
0.589597171*	-0.01014277	0.019506825
0.631963601*	-0.009199648	-0.005139335
0.253242149*	0.014675578	0.012252077
0.550325515*	-0.011382093	-0.013944994
0.582228579*	-0.014279893	-0.01961969
0.739710452*	0.027507041	0.03907249
1.132244257*	-0.003660481	0.005519526
0.726040269*	-0.002912822	0.00104203
0.634843295*	0.002324528	0.006402992
0.867396041*	-0.026990584	-0.006980473
0.864901361*	0.052901024	0.025984224
0.542866656*	-0.011865064	-0.027081987
0.656008077*	0.001520202	0.001607355
0.662510727*	-0.009936816	0.027499156
0.64270229*	0.031702637	0.019853231
0.437816206*	0.026637177	0.002352374
0.618687373*	-0.006840159	-0.017260257
0.517177206*	0.010077145	0.020716982
1.816946995*	0.01906962	0.058574327
0.658651912*	-0.051710771	-0.034952637
0.661660126*	0.017593259	0.038555422
1.298621464*	-0.001304256	0.014687217
0.701222352	-0.016150848	-0.012909574

表 12　得分差值率

6.4.4 结论

由表可知，理化指标独立得分的差值率对红葡萄酒有四组不符合，对白葡萄酒全都不符合；而整合三阶迭代回归差值率对红葡萄酒仅有两组不符合，对白葡萄酒则全部符合；BP 神经网络得分差值率对红葡萄酒仅有一组不符合，对白葡萄酒则全部符合。通过比较说明，后两个模型均成立且回归效果非常好。

所以我们可以认为，葡萄酒的理化指标对葡萄酒的质量有影响，但不能单纯用葡萄酒的理化指标评价葡萄酒的质量。应用葡萄酒的理化指标和感官指标共同评价葡萄酒的质量。即葡萄酒和葡萄的理化指标对葡萄酒的质量有影响，但不能单纯用葡萄酒和葡萄的理化指标评价葡萄酒的质量。应用葡萄酒和葡萄的理化指标和感官指标共同评价葡萄酒的质量。

七、 模型分析

7.1 模型一的分析

问题一中的数据标准化我们采用了两种方法，其中置信度区间法能够使同组不同品酒员的评分的方差缩小，及评分更加集中、一致，不同酒样间的方差增大，即酒样更容易区分开来。这种方法要比普通的极值标准化方法更好，推荐在类似情形下使用。

对于评价两组数据间有无显著性差异，我们采用的曼惠特尼 U 检验比 t 检验和 F 检验的效果都要好。

7.2 模型三的分析

模型三是用来对酿酒葡萄进行分级的，我们使用了系统聚类的方法，得到了较好的结果。另外，我们还使用了 K-均值聚类的方法，根据系统聚类的结果，将红葡萄和白葡萄的分类数分别设置为 4 和 3，进行了均值聚类，得到的结果与系统聚类的结果基本一致，可见，我们采用的系统聚类的方法是可靠稳定的。（K-均值聚类结果见附录 5）

7.3 模型四的分析

模型四中的回归方程方法对分析红葡萄与红葡萄酒的理化指标的关系得到了比较理想的结果，但是对白葡萄和白葡萄酒的理化指标的关系得到的结果不够好。于是我们对模型用主成分法进行了调整，使得白葡萄的回归方程的显著性水平得到了提高。这种思想可以应用在当多种指标的对目标的显著性水平不够时，可以采用主成分分析法，用主成分去对目标进行拟合。

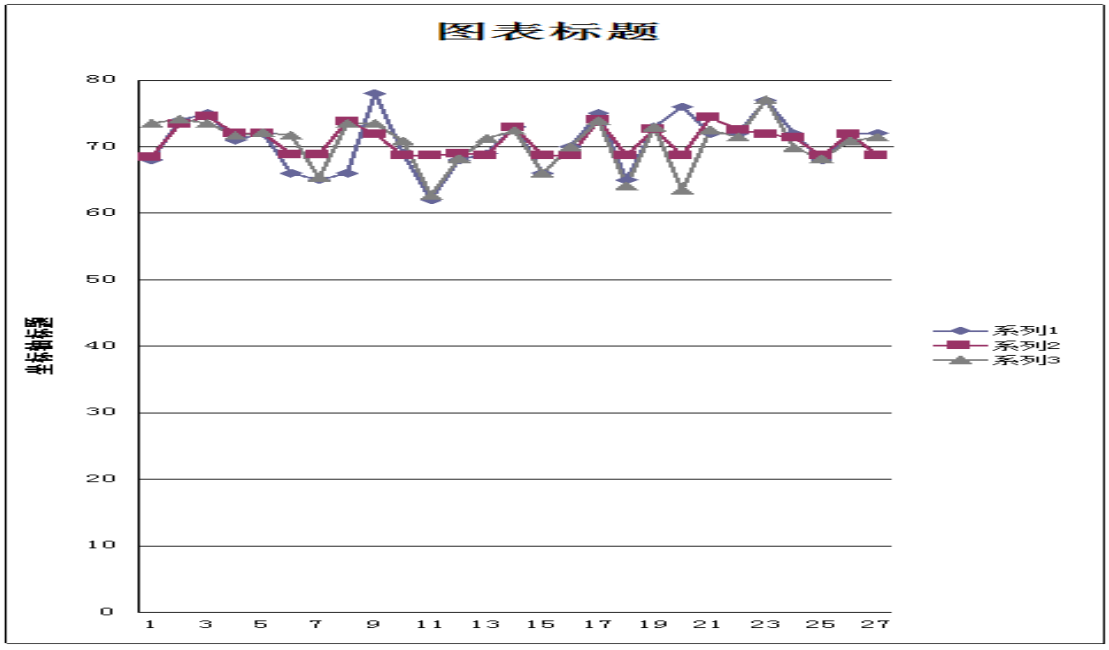
7.4 模型五、六的分析

BP 神经网络模型检验：

用该模型再对品酒师给出的红葡萄酒总酚进行验算，得出结果如下：计算值与实际值误差极小，由此可知模型准确性很高。

表 13　模型检验表

实际值	68.00	74.00	75.00	71.00	72.00	66.00	65.00	66.00	78.00
计算值	68.50	73.43	74.53	72.01	72.00	68.89	68.86	73.87	71.93
实际值	69.00	62.00	68.00	69.00	73.00	66.00	70.00	75.00	65.00
计算值	68.66	68.65	68.99	68.73	72.91	68.66	68.66	74.00	68.66
实际值	73.00	76.00	72.00	72.00	77.00	72.00	68.00	72.00	72.00
计算值	72.64	68.66	74.40	72.54	71.92	71.35	68.66	71.92	68.67



与此同时，用整合三阶迭代回归模型和 BP 神经网络模型得出的结论对比，进一步说明本问题两模型准确性很高。

八、 结论

问题一结论，我们首先利用置信区间法和极大-极小标准化方法，分别对数据进行标准化，并通过对比，得出用置信区间法处理后数据的准确性更高。然后我们采用 Mann-Whitney U 检验模型进行显著性的分析，得到两组品酒员的评价结果确实存在显著性差异。接下来，我们又通过比较两个评判组所得 10 个分数的方差，得出第二组结果更可信，我们的评判标准是：对同一个葡萄酒进行评价，10 个品酒员打出的 10 个得分，得分方差小的组更可信。

问题二结论，我们先对理化指标进行整理，通过主成分分析法得到 8 种主成份，然后结合问题一中的葡萄酒质量，应用系统聚类法，将红葡萄聚成 4 类，将白葡萄聚成 3 类，并分别对品质进行了排序，最终得到酿酒葡萄的分级结果。

	排序	酿酒葡萄编号
第一类	4	6、7、15、18、12、11、8
第二类	3	1、2、13、25、10、5、24、16、27 19、21、4、22、20、17、14、26
第三类	1	9、23
第四类	2	3

	排序	酿酒葡萄编号
第一类	1	7、15、18、6、11、10、14、21、4、22 5、20、2、3、9、12、25、28、26、23
第二类	3	1、13、8、16、17、19
第三类	2	24、27

问题三结论，我们建立了组合联系搜索模型，先进行相关性分析，再进行多元线性回归，并对模型所得结果中不十分理想的进行了调整，用主成份分析调整后的模型得到

相关性和显著性较高的理想结果。对调整前后的回归方程筛分汇总，得到了酿酒葡萄与葡萄酒的部分相关理化指标间的函数关系。

红葡萄酒用理化指标法求得的方程：

$$\begin{cases} \text{花色苷}=-22.248-8.012\text{hsg}+1.920\text{yjs}+0.159\text{DPPH}+226.717\text{zf}-7.878\text{dn}+4.966\text{ptzht} \\ \text{单宁}=1.241-0.075\text{hsg}+0.004\text{yjs}+0.001\text{DPPH}+5.492\text{zf}+0.163\text{dn}+0.08\text{ptzht} \\ \text{总酚}=0.716-0.096\text{hsg}+0.005\text{yjs}+5.813\text{zf}+0.147\text{dn}+0.055\text{ptzht} \\ \text{酒总黄酮}=-1.178-0.002\text{hsg}+0.001\text{DPPH}+0.878\text{zf}+0.355\text{dn}+0.007\text{ptzht} \\ \text{DPPH}=-0.051-0.003\text{hsg}+0.157\text{zf}+0.013\text{dn}+0.002\text{ptzht} \end{cases}$$

而白葡萄酒用主成分法求得的方程：

$$\begin{cases} \text{单宁}=1.851+0.083y_1+0.134y_2+0.161y_3+0.2227y_6 \\ \text{总酚}=1.456+0.066y_1+0.12y_2+0.118y_3 \\ \text{酒总黄酮}=1.581+0.424y_2+0.374y_3+0.023y_6 \\ \text{色差}=5.410-0.882y_2-0.839y_8 \end{cases}$$

用两种方法联合求得的结果作为酿酒葡萄与葡萄酒的理化指标之间的联系，保证了回归方程显著性的满足，具有较好的合理性。

问题四结论，由第三问我们已发现酿酒葡萄与葡萄酒的理化指标之间存在一定的线性关系。所以，此问题中我们只需要考虑一个变量对葡萄酒质量的影响。我们针对仅考虑理化指标的情况对葡萄酒质量的影响和考虑理化指标+芳香指标的情况对葡萄酒质量的影响进行对比，并用整合三阶迭代回归模型和 BP 神经网络模型得出的结论进行比较，确定模型的可靠性与高效性，并得出“葡萄酒和葡萄的理化指标对葡萄酒的质量有影响，但不能单纯用葡萄酒和葡萄的理化指标评价葡萄酒的质量。应用葡萄酒和葡萄的理化指标和感官指标共同评价葡萄酒的质量”的结论。

通过第一问中的数据处理结果、第二问中对酿酒白葡萄的分级结果以及酿酒白葡萄和白葡萄酒的理化指标之间的回复方程的分析，我们发现跟红葡萄对比起来，白葡萄的数据很不稳定，不仅在第一、第二评判组是否有显著性差异时出现了接受原假设的结果，而且在对白葡萄进行等级分类的时候，白葡萄的分类也显得相对凌乱。我们认为，这可能是白葡萄酒中的蛋白质和酒石酸不稳定对白葡萄造成的影响[6]。但是白葡萄的不稳定表现在利用了问题四中的整合三阶迭代回归模型和 BP 神经网络模型后得到了消除。

九、 参考文献

[1] 李华，葡萄酒品尝学，北京：中国青年出版社，1992 年 2 月 3 日；

[2] 王华等，葡萄酒感官评价结果的统计分析方法研究，中国食品学报，第 6 卷 2 期:126-131，2006 年 4 月；

[3] 李运等，统计分析在葡萄酒质量评价中的应用，酿酒科技，总第 178 期：79-82，2009 年；

[4] 百度百科，曼-惠特尼 U 检验，http://baike.baidu.com/view/4239142.htm，2012 年 9 月 8 日；

[5] Ken Black. Business Statistics: Contemporary Decision Making. John Wiley and Sons, 2009. ISBN:0470409010, 9780470409015；

[6] 张明霞，白葡萄酒中不稳定蛋白的研究进展，酿酒，第 33 卷第 5 期：2006 年 9 月

附录

附录 1 曼-惠特尼检验 U 的临界值表

曼-惠特尼检验U的临界值表

（仅列出单侧检验在*0.025或双侧检验在0.05处的U临界值*）

*n*2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
*n*1															
1															
2								0	0	0	0	1	1	1	1
3					0	1	1	2	2	3	3	4	4	5	5
4				0	1	2	3	4	4	5	6	7	8	9	10
5			0	1	2	3	5	6	7	8	9	11	12	13	14
6			1	2	3	5	6	8	10	11	13	14	16	17	19
7			1	3	5	6	8	10	12	14	16	18	20	22	24
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64

附录 2 第一种标准化所得葡萄酒品质分数的排序 用 Excel 软件得到

第一种标准化后的第二组排序	
红葡萄	
酒品9	78.70728
酒品23	77.59766
酒品3	75.15418
酒品17	74.5
酒品2	74
酒品20	73.9249
酒品19	73.34267
酒品21	72.79591
酒品26	72.64464
酒品14	72.6
酒品5	72.46953
酒品4	71.84256
酒品22	71.6
酒品24	71.5
酒品27	70.59446
酒品16	69.9
酒品10	69.37061
酒品1	69.00486
酒品12	68.80122
酒品13	68.8
酒品25	68.2
酒品7	66.88339
酒品6	66.75959
酒品15	66.342996
酒品8	66
酒品18	64.69101
酒品11	62.8336

第一种标准化后的第二组排序	
白葡萄	
酒品9	81.43086
酒品5	80.76376
酒品25	80.53199
酒品21	80.002496
酒品10	79.8
酒品28	79.6
酒品22	79.4
酒品17	79.28861
酒品15	78.70572
酒品23	77.4
酒品14	77.1
酒品27	77
酒品4	76.9
酒品3	76.79396
酒品18	76.7
酒品1	76.656998
酒品20	76.6
酒品24	76.42129
酒品19	76.4
酒品6	75.97668
酒品2	75.8
酒品26	75.314396
酒品7	74.2
酒品13	73.9
酒品12	73.5834
酒品11	72.33714
酒品8	72.3
酒品16	67.3

附录 3 酿酒葡萄、葡萄酒的理化指标相关性 用 SPSS19.0 软件得到

红葡萄相关性分析：

相关性									
控制变量			VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007
	VAR00010	相关性	.570	.519	.482	.458	-.004	.401	-.274
		显著性（双侧）	.003	.008	.015	.021	.984	.047	.186
		df	23	23	23	23	23	23	23

	VAR00011	相关性	-.104	-.082	-.123	-.091	-.037	-.115	-.135
		显著性（双侧）	.622	.696	.558	.665	.860	.583	.521
		df	23	23	23	23	23	23	23
	VAR00012	相关性	.910	.726	.790	.794	.489	.732	-.279
		显著性（双侧）	.000	.000	.000	.000	.013	.000	.178
		df	23	23	23	23	23	23	23
	VAR00013	相关性	.837	.698	.720	.711	.302	.646	-.307
		显著性（双侧）	.000	.000	.000	.000	.142	.000	.135
		df	23	23	23	23	23	23	23
	VAR00014	相关性	.415	.217	.232	.359	.073	.261	-.094
		显著性（双侧）	.039	.298	.264	.078	.727	.208	.654
		df	23	23	23	23	23	23	23
	VAR00015	相关性	.415	.217	.232	.359	.073	.261	-.094
		显著性（双侧）	.039	.298	.264	.078	.727	.208	.654
		df	23	23	23	23	23	23	23
	VAR00016	相关性	.771	.754	.823	.752	.478	.768	-.101
		显著性（双侧）	.000	.000	.000	.000	.016	.000	.630
		df	23	23	23	23	23	23	23

	VAR00017	相关性	.867	.848	.908	.891	.515	.882	-.237
		显著性（双侧）	.000	.000	.000	.000	.008	.000	.254
		df	23	23	23	23	23	23	23
	VAR00018	相关性	.726	.669	.697	.669	.421	.667	-.246
		显著性（双侧）	.000	.000	.000	.000	.036	.000	.236
		df	23	23	23	23	23	23	23
	VAR00019	相关性	.778	.737	.885	.847	.586	.837	-.059
		显著性（双侧）	.000	.000	.000	.000	.002	.000	.780
		df	23	23	23	23	23	23	23
	VAR00020	相关性	-.043	.056	.090	.054	.016	.081	-.107
		显著性（双侧）	.838	.789	.669	.799	.938	.700	.610
		df	23	23	23	23	23	23	23
	VAR00021	相关性	.484	.566	.371	.273	.155	.413	-.235
		显著性（双侧）	.014	.003	.068	.187	.459	.040	.259
		df	23	23	23	23	23	23	23
	VAR00022	相关性	.108	.361	.219	.203	.153	.279	-.360
		显著性（双侧）	.607	.076	.293	.330	.464	.177	.077
		df	23	23	23	23	23	23	23

	VAR00023	相关性	-.015	.143	.031	.006	-.029	.104	-.220
		显著性（双侧）	.943	.495	.881	.979	.889	.622	.290
		df	23	23	23	23	23	23	23
	VAR00024	相关性	.510	.480	.354	.438	.110	.376	-.156
		显著性（双侧）	.009	.015	.083	.028	.599	.064	.457
		df	23	23	23	23	23	23	23
	VAR00025	相关性	-.222	-.022	-.073	-.154	.093	-.028	-.057
		显著性（双侧）	.285	.918	.727	.462	.657	.894	.786
		df	23	23	23	23	23	23	23
	VAR00026	相关性	.270	.176	.172	.281	-.062	.168	-.034
		显著性（双侧）	.191	.400	.412	.174	.769	.423	.874
		df	23	23	23	23	23	23	23
	VAR00027	相关性	.374	.368	.286	.221	.326	.262	-.127
		显著性（双侧）	.065	.070	.165	.288	.111	.205	.546
		df	23	23	23	23	23	23	23
	VAR00028	相关性	.530	.413	.442	.514	.269	.447	-.216
		显著性（双侧）	.006	.040	.027	.009	.193	.025	.299
		df	23	23	23	23	23	23	23

	VAR00029	相关性	-.280	-.124	-.133	-.140	-.209	-.157	-.037
		显著性（双侧）	.175	.554	.525	.504	.315	.455	.860
		df	23	23	23	23	23	23	23
	VAR00030	相关性	.113	.386	.190	.216	.038	.284	-.303
		显著性（双侧）	.589	.057	.362	.301	.858	.169	.141
		df	23	23	23	23	23	23	23
	VAR00031	相关性	.196	.393	.260	.213	.107	.302	-.397
		显著性（双侧）	.348	.052	.210	.307	.612	.142	.049
		df	23	23	23	23	23	23	23
	VAR00032	相关性	.015	-.162	-.084	-.198	.130	-.133	.440
		显著性（双侧）	.942	.438	.690	.344	.534	.525	.028
		df	23	23	23	23	23	23	23
	VAR00033	相关性	.010	-.094	-.106	-.093	-.005	-.022	.198
		显著性（双侧）	.964	.654	.616	.659	.981	.918	.343
		df	23	23	23	23	23	23	23

白葡萄相关性分析：

相关性						
控制变量			VAR00001	VAR00002	VAR00003	VAR00004
	VAR00008	相关性	.428*	.480**	.293	-.216

	VAR00009	相关性	.362*	.419*	.605**	-.224
	VAR00010	相关性	-.165	-.100	-.155	.130
	VAR00011	相关性	-.217	-.298	-.163	.065
	VAR00012	相关性	.361*	.417*	.607**	-.224
	VAR00013	相关性	-.054	-.018	.233	-.012
	VAR00014	相关性	-.054	-.018	.233	-.012
	VAR00015	相关性	.409*	.450**	.133	.082
	VAR00016	相关性	.335*	.310	.166	.123
	VAR00017	相关性	.574**	.573**	.346*	-.002
	VAR00018	相关性	.495**	.588**	.697**	-.101
	VAR00019	相关性	-.062	.037	-.095	-.213
	VAR00020	相关性	.411*	.386*	.614**	-.095
	VAR00021	相关性	.355*	.323*	-.129	-.370*
	VAR00022	相关性	.093	.160	.084	.167
	VAR00023	相关性	.172	.117	-.174	-.045
	VAR00024	相关性	.056	.009	-.129	.053
	VAR00025	相关性	.039	.082	.138	-.159
	VAR00026	相关性	-.350*	-.434*	-.520**	.065
	VAR00027	相关性	-.295	-.293	-.048	.091
	VAR00028	相关性	-.242	-.204	-.263	.072
	VAR00029	相关性	.348*	.361*	-.060	-.158

	VAR00030	相关性	.229	.265	.114	-.015
	VAR00031	相关性	.059	.004	-.008	-.033
	VAR00032	相关性	.373*	.397*	.274	-.092

相关性

控制变量			VAR00005	VAR00006	VAR00008	VAR00009
	VAR00008	相关性	.205	-.290	1.000	.121
	VAR00009	相关性	.222	-.393*	.121	1.000
	VAR00010	相关性	.235	.110	-.300	-.225
	VAR00011	相关性	-.175	.308	-.060	-.450**
	VAR00012	相关性	.222	-.391*	.120	1.000**
	VAR00013	相关性	.047	.013	-.123	.460**
	VAR00014	相关性	.047	.013	-.123	.460**
	VAR00015	相关性	.387*	-.438**	.214	.099
	VAR00016	相关性	.306	.218	.024	.153
	VAR00017	相关性	.425*	-.377*	.272	.372*
	VAR00018	相关性	.428*	-.359*	.148	.504**
	VAR00019	相关性	-.052	.204	.274	.061
	VAR00020	相关性	.361*	-.085	.128	.370*
	VAR00021	相关性	.244	-.072	.352*	.104
	VAR00022	相关性	-.028	.032	.410*	.032
	VAR00023	相关性	.054	-.037	.186	-.181

	VAR00024	相关性	.062	-.260	.005	.186
	VAR00025	相关性	-.090	.207	.104	-.145
	VAR00026	相关性	-.050	.153	-.213	-.519**
	VAR00027	相关性	-.153	-.043	-.248	-.100
	VAR00028	相关性	-.248	1.000**	-.290	-.393*
	VAR00029	相关性	.135	-.090	.448**	-.043
	VAR00030	相关性	.048	-.024	.521**	.073
	VAR00031	相关性	.207	.349*	-.285	-.046
	VAR00032	相关性	.139	-.085	.126	.123

附录 4 各线性回归方程的系数值　用 SPSS19.0 软件得到

红葡萄回归方程：

系数 a						
模型		非标准化系数		标准系数	t	Sig.
		B	标准误差	试用版		
1	(常量)	-22.248	63.313		-.351	.729
	VAR00009	-8.012	9.714	-.068	-.825	.419
	VAR00010	1.920	.390	.748	4.924	.000
	VAR00011	.159	.076	.233	2.102	.048
	VAR00012	226.717	307.289	.110	.738	.469
	VAR00013	-7.878	6.380	-.227	-1.235	.231
	VAR00014	4.966	4.381	.143	1.133	.270
a. 因变量: 花色苷						

系数 a						
模型		非标准化系数		标准系数	t	Sig.
		B	标准误差	试用版		
1	(常量)	1.241	1.280		.969	.344
	VAR00009	-.075	.196	-.051	-.384	.705

	VAR00010	.004	.008	.123	.504	.620
	VAR00011	.001	.002	.080	.453	.656
	VAR00012	5.492	6.212	.212	.884	.387
	VAR00013	.163	.129	.372	1.264	.221
	VAR00014	.080	.089	.183	.907	.375

a. 因变量:单宁

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	.716	.880		.813	.426
	VAR00009	-.096	.135	-.074	-.710	.486
	VAR00010	.005	.005	.176	.915	.371
	VAR00011	.000	.001	.058	.413	.684
	VAR00012	5.813	4.271	.258	1.361	.189
	VAR00013	.147	.089	.386	1.657	.113
	VAR00014	.055	.061	.144	.904	.377

a. 因变量: 总酚

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	-1.178	1.125		-1.047	.308
	VAR00009	-.002	.173	-.001	-.013	.990
	VAR00010	.000	.007	.008	.039	.969
	VAR00011	.001	.001	.136	.897	.381
	VAR00012	.878	5.461	.033	.161	.874
	VAR00013	.355	.113	.788	3.129	.005
	VAR00014	.007	.078	.015	.088	.930

a. 因变量: 酒总黄酮

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	-.051	.051		-1.005	.327
	VAR00009	-.003	.008	-.048	-.402	.692
	VAR00010	-5.565E-5	.000	-.039	-.179	.860
	VAR00011	3.098E-5	.000	.082	.513	.614

	VAR00012	.157	.245	.138	.641	.529
	VAR00013	.013	.005	.680	2.569	.018
	VAR00014	.002	.003	.091	.500	.622

a. 因变量: DPPH

白葡萄回归方程：

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	.856	.280		3.056	.005
	VAR00005	.178	.078	.432	2.289	.031

a. 因变量: 单宁

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	.689	.195		3.543	.002
	VAR00005	.108	.054	.358	1.992	.057
	VAR00006	.096	.044	.392	2.180	.039

a. 因变量: 总酚

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	-.451	.679		-.665	.513
	VAR00005	-.037	.159	-.039	-.231	.819
	VAR00006	.558	.128	.728	4.358	.000
	VAR00008	.001	.001	.225	1.602	.123
	VAR00010	-.204	.241	-.119	-.846	.406

a. 因变量: 酒总黄酮

系数 a

模型		非标准化系数		标准系数		
		B	标 准 误差	试用版	t	Sig.
1	(常量)	7.477	9.453		.791	.436

	VAR00003	-.029	.132	-.043	-.219	.828

a. 因变量:色差

附录 5 K-均值聚类结果 用 SPSS19.0 软件得到

输出红葡萄类别：

聚类成员

案例号	聚类	距离
1	4	6.054
2	2	3.422
3	3	4.446
4	4	3.399
5	4	1.970
6	1	2.065
7	1	2.197
8	1	5.877
9	2	2.814
10	4	6.459
11	1	4.813
12	1	4.153
13	4	3.182
14	4	3.705
15	1	1.367
16	4	2.053
17	3	2.918
18	1	2.637
19	3	2.435
20	4	3.509
21	3	1.768
22	4	3.143
23	2	2.126
24	4	1.755
25	1	4.456
26	4	4.488
27	4	2.181

输出白葡萄类别：

聚类成员

案例号	聚类	距离
1	1	26.129
2	3	8.997
3	3	18.289
4	3	18.393

5	3	10.478
6	3	20.421
7	3	25.752
8	1	8.124
9	3	19.831
10	3	14.167
11	3	21.235
12	3	11.085
13	1	24.417
14	3	12.966
15	3	23.528
16	1	17.618
17	1	15.176
18	3	31.634
19	1	31.233
20	3	10.297
21	3	27.775
22	3	23.061
23	3	25.079
24	2	15.491
25	3	21.923
26	3	34.601
27	2	15.491
28	3	25.192