K均值聚类在葡萄酒分级中的应用

凌　佳[1]，言方荣[2,*]
（1.江苏广播电视大学，江苏南京 210036；
2.中国药科大学，江苏南京 210009）

摘　要：葡萄酒分级是葡萄酒评价中的一个重要内容，结合数据降维技术，建立葡萄酒K均值分类模型。通过实例分析证实所得结果较好，该方法在葡萄酒的评价分级中具有很好的应用价值。

关键词：葡萄酒分级，数据降维，K均值分类模型

Application of K-means clustering in wine classification

LING Jia[1], YAN Fang-rong[2,*]

(1.Jiangsu Radio and TV University, Nanjing 210036, China;
2.China Pharmaceutical University, Nanjing 210009, China)

Abstract: Wine classification is an important content in the evaluation of wine. The wine K-means classification model was established based on the data dimension reduction techniques. The result were confirmed good through the example analysis. And this method had excellent application value in wine classification.

Key words: wine classification; data dimension reduction; K-means classification model

中图分类号：TS261.7　　**文献标识码**：A　　**文 章 编 号**：1002-0306(2013)06-0104-04

葡萄酒的等级和原产地制度是反映葡萄酒质量的产区特色的重要指标，已经得到世界葡萄酒主产国政府、葡萄酒市场和消费者的认可。世界上传统葡萄酒生产国的葡萄酒工会、酒商等通常会通过品评比赛或者认证组织的评审，对葡萄酒产品按照质量高低划分等级[1-4]。葡萄酒分级是葡萄酒评价中一个非常重要的内容[5-7]。然而，如何科学的构建葡萄酒分级标准，仍然是一个值得深入研究的问题。事实上，影响葡萄酒定级的原因很多，但葡萄酒本身的品质及专家对葡萄酒的评价是两个不可忽视的指标。本文通过构建关于这两者的数学模型[8-9]，试图给葡萄酒的定级给出一种新的方案。

1　数据准备与处理

1.1　数据准备

为了构建葡萄酒分级模型[10-11]，首先收集了一批葡萄酒样品数据（数据来源cumcm 2012）。该数据收集了27个红葡萄酒样品，28个白葡萄酒样品相应的专家打分及相关葡萄酒理化指标114个。首先，我们对所收集到的数据进行标准化处理，为建模分析做准备。

1.2　数据处理

由于数据只收集了27个红葡萄酒样品，28个白葡萄酒样品，然而所需分析的指标有114个，且指标间存在明显共线性，是一个典型的“小n大p”的统计问题，即待分析指标数超过数据样本数，因此首先要考虑对数据进行“降维”处理，具体步骤如下：

a.以红、白酿酒葡萄的理化性质（包括芳香物质）为自变量，葡萄酒得分为因变量做线性回归，可以得到如下的线性方程：

$$Y_h=\beta_{h0}+\beta_{h1}x_{h1}+\cdots+\beta_{hn}x_{hn}\ (n=1,2,\cdots,114)\qquad 式(1)$$

式中，Y_h表示第h种葡萄酒的分数，对于红葡萄酒，$h=1,2,\cdots,27$；对于白葡萄酒，$h=1,2,\cdots,28$。x_{hn}为第h种酒的第n个理化指标，β为系数。

b.由于对葡萄的所有理化指标进行分析过于繁琐，所以可以考虑把其“降维”综合成一个理化指标。利用式(1)对于得到的114个系数，其中有正有负。将系数为正的项提取出来作为有利项，系数为负的作为有害项。然后将有利项与其所对应指标相乘，将有害项的绝对值与其所对应指标的绝对值相乘，计算它们和的平方根之差，分别得到红、白葡萄的综合理化指标Z。其公式如下：

$$Z_h=\sqrt{\sum_{i=1}^{m}\beta_{hi}x_{hi}}-\sqrt{\sum_{j=1}^{k}|\beta_{hj}||x_{hj}|}\qquad 式(2)$$

$$\begin{cases}h=1,2,\cdots,27 & 红葡萄\\ h=1,2,\cdots,28 & 白葡萄\end{cases}$$

式中，$\beta_{hi}>0$，β_{hi}为第h种葡萄的第i个有利项；$\beta_{hj}<0$，β_{hj}为第h种葡萄的第j个有害项。Z_h为第h种葡萄的综合理化指标。即我们用综合理化指标替代原有的114个指标，进行数据分析，避免了高维数据分析，实现了数据“降维”。

利用上述步骤，对27组红葡萄以及28组白葡萄进行数据预处理，分别得到其综合理化指标Z，其结

收稿日期：2012-10-08　*通讯联系人
作者简介：凌佳(1980-)，女，研究生，研究方向：数理统计。
基金项目：国家自然科学基金重点项目(81130068)；中央高校专项业务经费(JKQ2011032)。

果如表1所示。从表1可知，红葡萄酒综合理化指标明显不同于白葡萄酒综合理化指标，因此，在下面分析中，将分别对红葡萄酒和白葡萄酒分开讨论。

表1 红、白葡萄的综合理化指标

Table 1 Red and white grapes physical and chemical indicators

红葡萄				白葡萄			
样品	Z值	样品	Z值	样品	Z值	样品	Z值
1	3.3843	15	3.2679	1	−0.36983	15	−1.1795
2	5.2121	16	4.4078	2	−1.0192	16	−1.0602
3	4.6318	17	4.8503	3	−0.64283	17	−0.63485
4	3.4795	18	3.7113	4	−0.55278	18	−1.107
5	4.2768	19	4.5691	5	−1.247	19	−1.1921
6	4.479	20	4.4349	6	−1.4438	20	−0.68297
7	4.0619	21	4.3484	7	−0.75602	21	−0.83928
8	4.1512	22	4.151	8	−1.2701	22	−1.2156
9	4.6686	23	5.0834	9	−1.0857	23	−0.87665
10	4.5409	24	4.7349	10	−0.93767	24	−0.99733
11	3.8196	25	4.3104	11	−1.1922	25	−0.71766
12	2.1625	26	4.6787	12	−1.7366	26	−0.42062
13	4.2329	27	4.1013	13	−1.7376	27	−1.6291
14	4.3927			14	−1.145	28	−0.40195

2 数据建模

K−均值(K−means，KM)算法是最为经典的基于划分的聚类方法[12]，是十大经典数据挖掘算法之一。它的基本思想是：以空间中k个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。算法描述如下：

a.从n个数据对象中任意选择k个对象作为初始聚类中心；

b.根据每个聚类对象的means(中心对象)，计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分，将每个对象(重新)赋给最相近的类；

c.重新计算每个(有变化)聚类的means(中心对象)；

d.重复b、c，直到每个聚类不再发生变化为止。

KM算法尝试找出使平方误差函数值最小的k个聚类，其定义如下：

$$E=\sum_{i=1}^{k}\sum_{p\in c_i}|p-m_i|^2 \qquad 式(3)$$

式中，E是数据库中所有对象的平方误差总和；p是空间中的点，表示给定的数据对象；m_i是簇c_i的平均值(p和m_i都是多维的)。这个准则使生成的结果簇尽可能地紧凑和独立。

由于KM算法首先随机地选择k个对象，每个对象初始地代表了一个簇的平均或中心，对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平中心。这个过程不断重复，直到平方误差函数收敛。

另外，一个较好的分类需要满足以下两点：a.综合理化指标Z的大小应与酒的分数大小呈一定线性关系；b.基于同一类别内数据越紧凑越好，而类别间的差距越大越好的原则，这里用作为其聚类方法的评价指标，其定义为：

$$F_n=\frac{\sum_{i=1}^{n}(\max(Y_i)-\min(Y_i))}{\max(Y)-\min(Y)} \qquad 式(4)$$

式中，i表示聚类的类别数；$\max(Y_i)$为第i类评分数的上限；$\min(Y_i)$为第i类评分数的下限；$\max(Y)$表示此分类方法下评分数的最大值；$\min(Y)$表示此分类方法下评分数的最小值；F_n越大，说明类与类之间重合的部分越多；F_n越趋近于0，说明此分类方法准确性越高；而当F_n处于中间水平时对于一个未知的样本，将很难准确划分。经过尝试，发现F_n为0.75左右时，其分类最为理想。

3 结果与讨论

根据表1中得到的27种红葡萄和28种白葡萄所对应葡萄的综合理化指标以及27种红葡萄酒和28种白葡萄酒的分数(即反映其质量)，对其分别进行KM聚类。应用Matlab软件对葡萄进行KM聚类，由于KM聚类时类别数是自定的，而类别数大于6时，无实际意义，所以此处分别试了类别数从2到5的分类方法。其结果如图1和图2所示。

应用KM分类得到的结果中，葡萄的综合理化指标Z与葡萄酒的分数Y呈明显的线性关系，说明葡萄的综合理化指标Z与葡萄酒的分数有明显的相关性，即与得分高的酒分为一类的葡萄，其质量较好。

对于红、白葡萄在不同分类情况下的F值，如表2所示。

表2 红、白葡萄在不同分类情况下的F值

Table 2 The F value of different classification for red and white grapes

红葡萄		白葡萄	
类别数	F	类别数	F
2	0.96736	2	0.89855
3	0.82493	3	0.75362
4	0.96736	4	0.74396
5	1.0504	5	0.68599

由表2可知，分成3组时，其F值最小。另外，由于葡萄的综合理化指标Z与葡萄酒的分数有明显的相关性，所以对于红葡萄来说，可将其依据红葡萄酒的得分情况分为3级，即质量较好、质量中等和质量较劣的葡萄。结果见表3。

表3 红葡萄分3类时的结果

Table 3 The classification results for red grapes

类别	葡萄样品号	所对应葡萄酒得分的平均值	等级
1	12	53.90	品质较劣
2	1、4、11、15、18	64.00	品质中等
3	2、3、5、6、7、8、9、10、13、14、16、17、19、20、21、22、23、24、25、26、27	76.12	品质较优

另外由表2可知，对于白葡萄，虽然分成4、5组

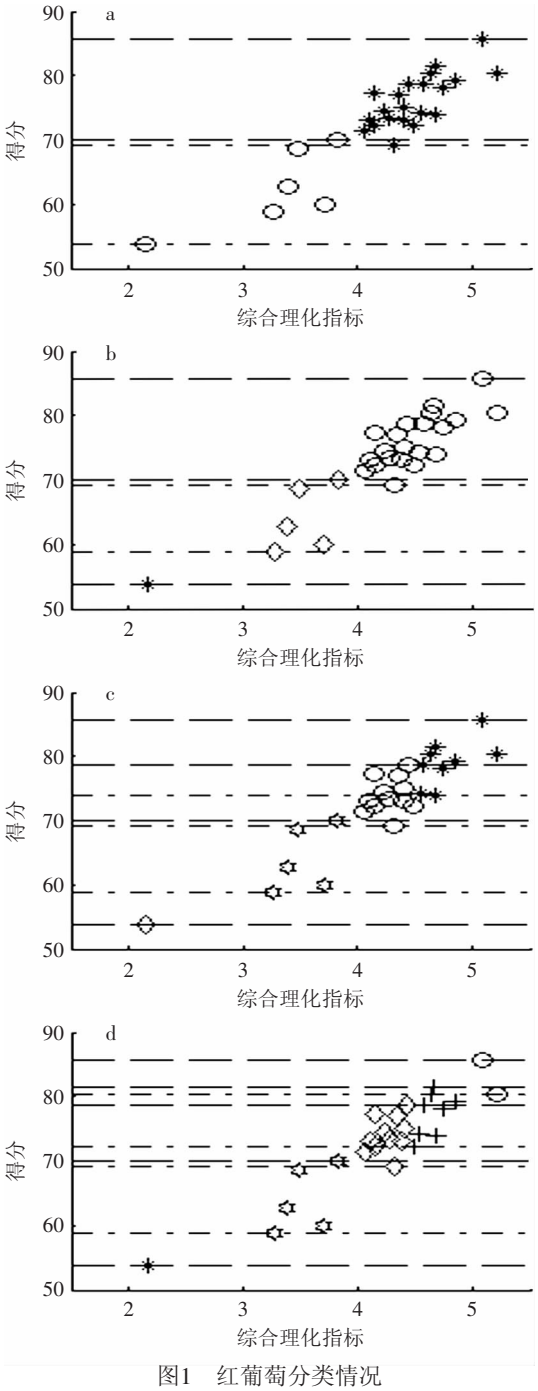

图1 红葡萄分类情况

Fig.1 Classification for red grapes

注：a：分为2类；b：分为3类；c：分为4类；d：分为5类；图2同。

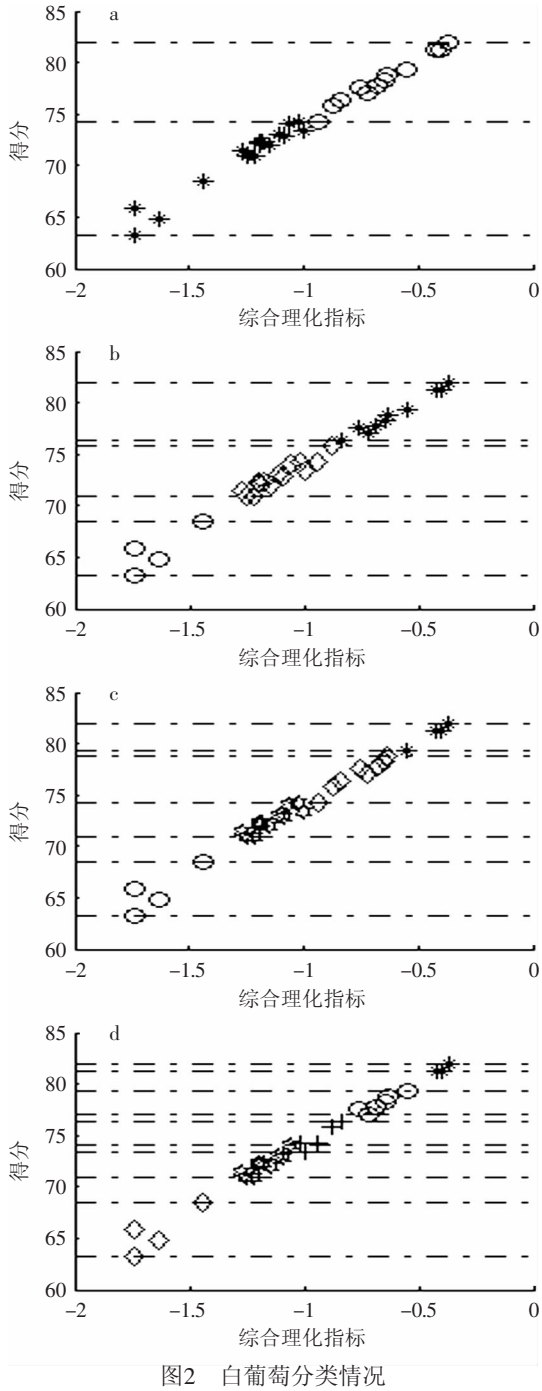

图2 白葡萄分类情况

Fig.2 Classification for white grapes

时，其F值越来越小，但是分3组时的F值最接近0.75，故认为分为3级最为合理。另外，由于葡萄的综合理化指标Z与葡萄酒的分数有明显的相关性，所以对于白葡萄来说，可也将其依据白葡萄酒的得分情况分为3级，即质量较好、质量中等和质量较劣的葡萄。结果如表4所示。

表4 白葡萄分3类时的结果

Table 4 The classification results for white grapes

类别	葡萄样品号	所对应葡萄酒得分的平均值	等级
1	6、12、13、27	65.60	品质较劣
2	2、5、8、9、10、11、14、15、16、18、19、22、23、24	72.86	品质中等
3	1、3、4、7、17、20、21、25、26、28	78.99	品质较优

4 结论

本文根据实验收集数据，探讨了利用数学模型建立葡萄酒分级方法的可行性。对于葡萄酒多指标问题，提出了数据降维技术，根据降维数据，建立了K均值聚类模型。针对K均值聚类中心的不唯一性，提出了F评价函数，确定了聚类类别数。所得结果合理，具有很好的实际应用价值。

当然，葡萄酒分级是一个复杂的问题，与很多因素有关。本文将所有理化指标归结为一个理化指标，

方便了后续数据建模处理，但同时也损失了很多有效信息。如何更好的挖掘这些信息，使得更好的为葡萄酒品质进行定级，这仍然是需要继续深入研究的问题。

参考文献

[1] 张倏翰，黄卫东. 北京地区消费者对葡萄酒等级和原产地制度的认知[J]. 食品与发酵工业，2011，37(9)：180-184.
[2] 郭其昌，郭松泉. 葡萄酒的质量等级法[J]. 中外葡萄与葡萄酒，1999(4)：64-67.
[3] 郭其昌，郭松泉，朱林. 再论葡萄酒的质量等级制[J]. 中外葡萄与葡萄酒，2000(3)：6-10.
[4] 朱加叶，乙小娟. 法国葡萄酒的等级分类[J]. 质量与生活，2008(10)：63.
[5] 王欣. 国产葡萄酒呼唤“国产”评级标准[J]. 中外葡萄与葡萄酒，2006(2)：46.
[6] 杨和财，沈忠勋，王灿辉. 我国葡萄酒质量等级制度的构建[J]. 酿酒科技，2008(3)：118-122.
[7] 杨和财，姚顺波. 国际葡萄酒质量等级制度对构建中国葡萄酒质量等级制度的启示[J]. 世界农业，2008(4)：62-65.
[8] 朱道元. 数学建模[M]. 北京：机械工业出版社，2008：237-256.
[9] 司守奎. 数学建模[M]. 青岛：海军航空工程学院出版社，2003：134-138.
[10] 王庆华，王庆斌. 应用数理统计方法评酒提高汾酒质量[J]. 酿酒，2010，37(1)：47-48.
[11] 马腾，赵丽，李军. 2008年份昌黎原产地葡萄酒理化特性的统计学分析[J]. 河北科技师范学院学报，2012，26(1)：5-11.
[12] 何迎生，段明秀. 基于改进kmeans聚类方法的RBF神网络设计[J]. 邵阳学院学报：自然科学版，2008，5(2)：48-49.

(上接第87页)

6')，121.12(C-1')，117.07(C-3'，5')，109.31(C-6)，105.05(C-10)，103.85(C-3)，93.58(C-8)，82.56(C-5'')，80.19(C-3'')，75.45(C-1'')，72.53(C-2'')，71.65(C-4'')，62.57(C-6'')。以上数据与文献[14]的报道一致，表明化合物2为异牡荆苷(isovitexin)。

确证两种化合物的结构式见图4。

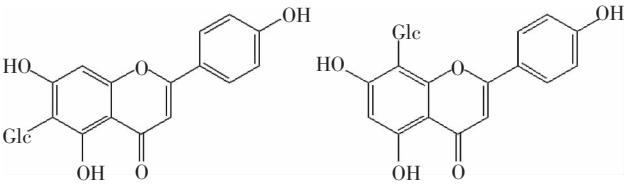

图4　2种化合物的结构式

Fig.4　The structures of compound 1 and compound 2

3　结论

将黑茶60%乙醇提取物经D101柱层析，乙醇-水梯度洗脱收集得到8个流分，以α-葡萄糖苷酶的抑制活性为评价指标，流分D-6、D-7及D-8具有显著的抑制活性且活性均较醇提取物强。综合考虑化合物组成及流分的量，确定以流分D-7为后续分离样品；以具有无需固定相载体、无化合物死吸附等优点的高速逆流色谱[15]作为后续分离的分离方法，选择合适的溶剂系统乙酸乙酯-正丁醇-甲醇-水-乙酸 (4:1:1:4:0.1，v/v)，经一步分离可获得纯度均高于95%的化合物1(牡荆苷)及化合物2(异牡荆苷)。以活性导向下追踪分离的思路应用于活性化合物的分离纯化，有效提高化合物的分离效率，并且有目的地从天然产物中分离得到具有活性的化合物。

参考文献

[1] 苗爱清，程悦，梁祈，等. 白叶单枞不同发酵茶中多酚类成分的HPLC-MS/MS分析[J]. 中国农学通报，2011，27(2)：360-366.
[2] 赖兆祥，黄国滋，赵超艺，等. 岭头单枞黑茶渥堆工艺探讨[J]. 中国茶叶加工，2008(4)：23-25.
[3] C J Greenwalt，R A Ledford，K H Steinkraus. Determination and characterization of the antimicrobial activity of the fermented tea kombucha[J]. Lebensmittel-Wisseenschaftund-Technolongie，1998，31(3)：291-296.
[4] 杨抚林，邓放明，赵玲燕，等. 黑茶微生物学研究进展[J]. 微生物学杂志，2006，26(1)：81-84.
[5] 张冬英，刘仲华，施兆鹏，等. 高通量筛选法对普洱茶降血糖血脂作用的研究[J]. 茶叶科学，2005，26(1)：49-53.
[6] 杨新河，黄建安，刘仲华，等. 普洱茶对α-葡萄糖苷酶活性影响的研究[J]. 食品工业科技，2012，33(12)：122-124.
[7] Toshiro M，Takashi T，Satomi T，*et al*. α-glucosidase inhibitory profile of catechins and theaflavins[J]. J Agric Food Chem，2007，55：99-105.
[8] Yao Y，Sang W，Zhou M J，*et al*. Antioxidant and α-glucosidase inhibitory activity of colored Granins in China[J]. J Agric Food Chem，2010，58：770-774.
[9] Miwa H，Yukihiko H. Inhibition of rat small intestinal sucrase andα-glucosidase activities by tea polyphenols[J]. Biosci Biotech Biochem，1993，57(1)：123-124.
[10] Cui H Y，Jia X Y，Zhang X，*et al*. Optimization of high-speed counter-current chromatograph for separation of polyphenols from the extract of hawthorn(*Crataegus laevigata*) with response surface methodology[J]. Sep Purif Technol，2011，77：269-274.
[11] 彭爱一，曲学伟，李慧，等. 高速逆流色谱分离纯化九里香中的黄酮类化合物[J]. 色谱，2010，28(4)：383-387.
[12] Yoo S W，Kim J S，Kang S S，*et al*. Constituents of the fruits and leaves of *Euodia daniellii*[J]. Arch Pharm Res，2002，25(6)：824-830.
[13] Kim J H，Lee B C，Kim J H，*et al*. The isolation and antioxidative effects of vitexin from *Acer palmatum*[J]. Arch Pharm Res，2005，28(2)：195-202.
[14] Takahiro H，Young S Y，Akira K. Five novel flavonoids from Wasabia japonica[J]. Tetrahedron，2005，61：7037-7044，
[15] Walter D C. Counter-current chromatography：simple process and confusing terminology[J]. J Chromatogr A，2011，1218：6015-6023.