doi:10.6041/j.issn.1000-1298.2014.01.032

基于 Pearson 系数与多元核支持向量分类的葡萄酒分析*

蒋　辉[1,2]　邓伟民[1]　陈晓青[1]

(1. 惠州学院数学系，惠州 516007；2. 中国人民大学统计学院，北京 100872)

摘要：选择 Pearson 相关系数筛选出与红葡萄酒各理化指标相关性较强的酿酒红葡萄理化指标，用逐步回归法建立回归方程确定了它们之间的数量关系。同时，采用多元核支持向量机对红葡萄酒样品进行分类，所分类别与人工口感评分所分类别基本相符，正确率达到 91.89%，结论表明酿酒红葡萄和红葡萄酒的理化指标能很好地确定葡萄酒的口感评价。

关键词：葡萄酒　理化指标　Pearson 相关系数　逐步回归　支持向量机

中图分类号：TS262.6；TS207.3　　**文献标识码**：A　　**文章编号**：1000-1298(2014)01-0203-06

引言

葡萄酒质量与其成分关系密切，是其外观、香气、口感等的综合表现。在葡萄酒的质量甄别中，其成分构成和质量评价是两个关键的因素，这两者一般通过其理化指标的检验和评酒员的主观评价来进行判别。参考国外近年的研究文献，理化指标的检验一般运用常规方法(包括密度、酒精度和 pH 值等的检验)来划分葡萄酒等级，多用于葡萄酒的发酵过程中[1]。在不同品种(或产地)的红酒评价中，理化指标检验也是常见的方法之一[2~6]。随着数据处理技术的发展，以葡萄酒的理化指标数据为样本，对葡萄酒的质量进行评价的研究越来越多[7~8]。另外，其他处理方法如逻辑斯蒂回归[9]、可视化分析[10]、独立主成分分析[11]、HPLC 技术[12]和 ICP-MS 技术[13]等也常见于近期的研究中。

国内外从葡萄酒的理化指标性质来评判其质量的研究很多。但是，研究酿酒葡萄和葡萄酒的理化指标关系的文献并不多见，且选择葡萄酒的理化指标具有一定的主观性。本文在 2012 年全国大学生数学建模竞赛提供的数据基础上，结合 Pearson(centered)相关系数，探讨红葡萄酒与酿酒红葡萄理化指标之间的联系，构建回归方程以确定它们之间的数量关系。同时，根据 Pearson(centered)系数客观地选择与葡萄酒口感评价相关性较大的理化指标作为输入特征，结合支持向量分类机对所提供的红葡萄酒样品进行分类。考虑到酿酒红葡萄各理化指标不同属性特征含有的信息量不同，采用标准支持向量机的单一尺度因子一方面可能带来信息的损失，另一方面可能纳入冗余信息。因此，对于不同的输入属性赋予不同的尺度因子，采用多元核，以模式搜索法对核参数进行调节，刻画不同属性指标对输出的不同贡献。

1　酿酒红葡萄与红葡萄酒理化指标之间的联系

1.1　Pearson 相关系数分析

酿酒红葡萄理化指标决定和影响了红葡萄酒的理化指标。在 2012 年全国大学生数学建模竞赛提供的数据中，由于酿酒红葡萄的理化指标很多，且其中有些指标与红葡萄酒中的一些指标之间的相关性很小甚至不相关，这里采用 Pearson(centered)相关系数 P_c 计算出红葡萄酒与酿酒红葡萄理化指标之间的一一对应相关程度，并遴选其中相关性较强的理化指标进行分析。

P_c 主要用来衡量两组分析对象线性关系的强弱，其取值范围为[0,1]，数学表达式为

$$P_c=\frac{N\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^2-\left(\sum x_i\right)^2}\sqrt{N\sum y_i^2-\left(\sum y_i\right)^2}} \tag{1}$$

其中，x 和 y 分别对应两组分析对象的数据序列，N 为样本个数，P_c 绝对值越大，相关性越强。

收稿日期：2013-01-30　修回日期：2013-04-06

*国家自然科学基金资助项目(61144004)、全国统计科学重点资助项目(2013LZ52)、中国博士后科学基金资助项目(2012M520495)、广东省自然科学基金资助项目(S2013010014601、S2013010013212)和惠州市科技计划项目(2012P15)

作者简介：蒋辉，副教授，中国人民大学博士后，主要从事应用统计、数据挖掘等研究，E-mail：jh@hzu.edu.cn

考虑到酿酒红葡萄的一级指标与二级指标之间有着内部性质差异,但明显二级指标是一级指标的一个分类,故只选取酿酒红葡萄所有的一级指标而不考虑二级指标,即将所有一级指标作为分析对象,再算出与红葡萄酒的理化指标间一一对应的 P_c 值。借助 SPSS 19.0 软件对数据进行处理,可得到红葡萄酒理化指标 $Y_i(i=1,2,\cdots,9)$ 和酿酒红葡萄理化指标 $X_j(j=1,2,\cdots,30)$ 之间的相关系数矩阵,找出各相关性较高的指标群,如表 1 所示,表中 $X_1 \sim X_{30}$ 分别表示氨基酸、蛋白质、维生素 C、花色苷、酒石酸、苹果酸、柠檬酸含量,多酚氧化酶活力,褐变度,DPPH 自由基、总酚、单宁、葡萄总黄酮、白藜芦醇、黄酮醇、总糖、还原糖、可溶性固形物含量,pH 值,可滴定酸含量,固酸比,干物质含量,果穗质量,百粒质量,果梗比,出汁率,果皮质量,果皮颜色 L^*、a^*、b^*。

根据表 1 中 P_c 的分析,可得出如下结论:

(1)与红葡萄酒各理化指标相关性较强的酿酒红葡萄理化指标中,出现频率 7 次(6 次在 0.01 水平下显著性相关)的指标有 DPPH 自由基含量(X_{10})、总酚含量(X_{11})和葡萄总黄酮含量(X_{13}),出现频率 6 次(均在 0.01 水平下显著性相关)的有花色苷含量(X_4)和单宁含量(X_{12}),出现频率 5 次的指标有蛋白质含量(X_2)、褐变度(X_9)和黄酮醇含量(X_{15}),它们是酿酒红葡萄的重要理化指标,与红葡萄酒理化指标关系密切。

表 1 与红葡萄酒各理化指标相关性较强的酿酒红葡萄理化指标

Tab. 1 Physicochemical indexes of grape which have strong correlation with that of wine

红葡萄酒理化指标	酿酒红葡萄理化指标
花色苷含量(Y_1)	$X_4^{\bigstar}$、$X_6^{\bigstar}$、$X_8^{\blacktriangle}$、$X_9^{\bigstar}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\blacktriangle}$、$X_{15}^{\blacktriangle}$、$X_{25}^{\bigstar}$
单宁含量(Y_2)	$X_1^{\bigstar}$、$X_2^{\blacktriangle}$、$X_4^{\bigstar}$、$X_9^{\bigstar}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\bigstar}$、$X_{15}^{\bigstar}$、$X_{18}^{\blacktriangle}$、$X_{22}^{\blacktriangle}$、$X_{25}^{\blacktriangle}$、$X_{28}^{\blacktriangle}$
总酚含量(Y_3)	$X_2^{\blacktriangle}$、$X_4^{\bigstar}$、$X_9^{\blacktriangle}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\bigstar}$、$X_{15}^{\bigstar}$、$X_{25}^{\bigstar}$、$X_{26}^{\bigstar}$、$X_{28}^{\blacktriangle}$
酒总黄酮含量(Y_4)	$X_2^{\blacktriangle}$、$X_4^{\bigstar}$、$X_9^{\blacktriangle}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\bigstar}$、$X_{26}^{\blacktriangle}$
白藜芦醇含量(Y_5)	$X_{10}^{\blacktriangle}$、$X_{11}^{\blacktriangle}$、$X_{13}^{\bigstar}$
DPPH 半抑制体积(Y_6)	$X_1^{\blacktriangle}$、$X_2^{\blacktriangle}$、$X_4^{\bigstar}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\bigstar}$、$X_{15}^{\blacktriangle}$、$X_{26}^{\blacktriangle}$、$X_{28}^{\blacktriangle}$
色泽 L^*(Y_7)	$X_2^{\blacktriangle}$、$X_4^{\bigstar}$、$X_8^{\blacktriangle}$、$X_9^{\bigstar}$、$X_{10}^{\bigstar}$、$X_{11}^{\bigstar}$、$X_{12}^{\bigstar}$、$X_{13}^{\bigstar}$、$X_{15}^{\bigstar}$、$X_{25}^{\bigstar}$、$X_{26}^{\bigstar}$、$X_{28}^{\bigstar}$、$X_{29}^{\bigstar}$
色泽 a^*(Y_8)	$X_6^{\bigstar}$、$X_{14}^{\blacktriangle}$、$X_{21}^{\blacktriangle}$、$X_{29}^{\bigstar}$、$X_{30}^{\bigstar}$
色泽 b^*(Y_9)	$X_5^{\blacktriangle}$、$X_{17}^{\bigstar}$、$X_{22}^{\blacktriangle}$

注:右上角带▲号表示在 0.05 水平显著性相关,带★号表示在 0.01 水平显著性相关。

(2)与红葡萄酒各理化指标相关性在 0.01 和 0.05 水平下均不显著的酿酒红葡萄理化指标包括维生素 C 含量(X_3)、柠檬酸含量(X_7)、总糖含量(X_{16})、pH 值(X_{19})、可滴定酸含量(X_{20})、果穗质量(X_{23})、百粒质量(X_{24})和果皮质量(X_{27}),它们与红葡萄酒各理化指标线性关系不强,影响不显著。

(3)与红葡萄酒各理化指标(花色苷含量(Y_1)、单宁含量(Y_2)、总酚含量(Y_3)、酒总黄酮含量(Y_4)、DPPH 半抑制体积(Y_6)、色泽 L^*(Y_7)、色泽 a^*(Y_8))相关性较强的酿酒红葡萄理化指标中,均包含其对应指标,且均在 0.01 水平下显著相关(如:与花色苷含量(Y_1)显著性相关的酿酒红葡萄指标中,包含花色苷含量(X_4)),由此说明红葡萄酒的这些理化指标与酿酒红葡萄的对应指标线性相关性显著。但是,对于白藜芦醇含量(Y_5)和色泽 b^*(Y_9),其分别对应的酿酒红葡萄指标白藜芦醇含量(X_{14})和果皮颜色 b^*(X_{30})的相关性不显著。由此表明在红葡萄酒的发酵与制作工艺中,酿酒红葡萄的白藜芦醇含量(X_{14})和果皮颜色 b^*(X_{30})并不直接影响红葡萄酒的白藜芦醇含量(Y_5)和色泽 b^*(Y_9)。

采用 P_c 能发现酿酒红葡萄与红葡萄酒的理化指标之间的相关方向和相关程度,但是,依然不能通过酿酒红葡萄的理化指标含量预测出红葡萄酒理化指标的大致含量,因此,可建立回归模型描述每一个红葡萄酒理化指标和多个酿酒红葡萄理化指标之间的数量关系。

1.2 理化指标关系的逐步回归分析

由表 1 可知,对于红葡萄酒的每一个理化指标,一般都有多个酿酒红葡萄理化指标与其有较强的相关性,而这些酿酒红葡萄指标之间可能存在多重共线性,特别是在各指标之间有高度依赖性时,就给回归估计的系数带来不合理的解释[14]。因此,要得到一个可靠的回归模型,就需从这众多的酿酒红葡萄理化指标中挑选出对葡萄酒指标贡献率较大者,用逐步回归方法得出最优的回归方程[15]。

逐步回归包括两个方面:引入自变量和剔除自变量。即引入对回归模型显著程度较强的变量,剔除含有多重共线性的部分变量。逐步回归中,分别以红葡萄酒理化指标观测值 $Y_i(i=1,2,\cdots,9)$ 为被解释变量,逐个引入与其相关性较强的酿酒红葡萄理化指标 X_j 作为解释变量,构成回归模型并进行模

型估计。根据拟合优度的变化决定新引入的变量是否可以用其他变量的线性组合代替。如果拟合优度变化显著，说明新引入的变量是一个独立的解释变量，在模型中应予以保留；否则，说明引入的变量可以用其他变量的线性组合代替，即该变量与其他变量之间存在共线性的关系，应予以剔除。最终，得到仅包含对红葡萄酒理化指标观测值 Y_i 影响显著的酿酒红葡萄理化指标而不包含对其影响不显著的指标。具体求解过程如下：

（1）若考虑红葡萄酒理化指标观测值 $Y_i(i=1,2,\cdots,9)$ 建立的回归方程已引入了 k 个酿酒红葡萄理化指标（变量），再引入新变量 X_j，则其对 Y_i 的方差贡献（V_{ij}）为

$$V_{ij}(X_1,X_2,\cdots,X_k)=$$
$$E_i(X_1,X_2,\cdots,X_k,X_j)-E_i(X_1,X_2,\cdots,X_k)=$$
$$\Gamma_i(X_1,X_2,\cdots,X_k)-\Gamma_i(X_1,X_2,\cdots,X_k,X_j) \quad (2)$$

式中，E 表示回归平方和，Γ 表示剩余残差平方和。引入检验变量

$$F_{ij}=V_{ij}(X_1,X_2,\cdots,X_k)/$$
$$[\Gamma_i(X_1,X_2,\cdots,X_k,X_j)/(n-k-2)] \quad (3)$$

其中，n 是样本容量。令 F_α 是置信水平 α 下 F 检验的临界值，当 $F_{ij}>F_\alpha$ 时，表明引入的自变量 X_j 有意义。

（2）接着考虑葡萄酒理化指标观测值 $Y_i(i=1,2,\cdots,9)$ 建立的回归方程中已引入了 p 个酿酒红葡萄理化指标（变量），再剔除自变量 X_r，其对 Y_i 的方差贡献（V_{ir}）和检验变量分别为

$$V_{ir}(X_1,X_2,\cdots,X_p)=E_i(X_1,X_2,\cdots,X_p)-$$
$$E_i(X_1,X_2,\cdots,X_{r-1},X_{r+1},\cdots,X_p)=$$
$$\Gamma_i(X_1,X_2,\cdots,X_{r-1},X_{r+1},\cdots,X_p)-\Gamma_i(X_1,X_2,\cdots,X_p) \quad (4)$$

$$F_{ir}=V_{ir}(X_1,X_2,\cdots,X_p)/$$
$$[\Gamma_i(X_1,X_2,\cdots,X_{r-1},X_{r+1},\cdots,X_p)/(n-p-2)] \quad (5)$$

令 F_β 是置信水平 β 下 F 检验的临界值，当 $F_{ir}\leqslant F_\beta$ 时，表明已引入的自变量 X_r 因后面变量的引入而变得不显著，应当剔除。

取显著性水平 $\alpha=\beta=0.05$，运用 Eviews 6.0 进行逐步回归可得各模型为

$$Y_1=1.936X_4+0.165X_9+0.727$$
$$(7.70)\quad(2.47)\quad(0.03)$$
$$Y_2=0.008X_4+14.350X_{10}+0.015X_{15}+$$
$$(2.34)\quad(5.18)\quad(2.43)$$
$$0.063X_{18}-12.601$$
$$(5.31)\quad(-4.48)$$
$$Y_3=0.008X_4+0.253X_{11}+1.684$$
$$(2.21)\quad(5.06)\quad(2.95)$$
$$Y_4=0.398X_{11}-0.951$$
$$(9.41)\quad(-1.40)$$
$$Y_5=-0.028X_2+0.477X_{13}+15.319$$
$$(-2.47)\quad(4.51)\quad(2.59)$$
$$Y_6=0.017X_{11}+0.024$$
$$(9.03)\quad(-0.80)$$
$$Y_7=-0.148X_4-0.136X_{15}+3.305X_{29}+55.430$$
$$(-6.78)\quad(-2.99)\quad(3.80)\quad(15.04)$$
$$Y_8=-1.303X_6-6.569X_{30}+54.708$$
$$(-2.42)\quad(-3.12)\quad(15.26)$$
$$Y_9=0.947X_5+0.115X_{17}+9.949$$
$$(2.70)\quad(3.51)\quad(-1.38)$$

模型下方括号内数据为回归系数的 t 检验值。在显著性水平为 0.05 情况下，以上方程所有解释变量系数 t 值均通过了检验，说明相应估计系数均显著异于零。为了进一步了解以上模型的合理性，表 2 中列出了模型计量分析的其他结果。

从表 2 看，以上各回归模型的 F 检验值均超过其临界值，且相应的 P 值几乎为零，说明模型整体显著；除了 Y_5、Y_8 和 Y_9 外，可决系数及其修正值都在 0.75 以上，说明大多数模型拟合优度较高；在逐步回归中，考虑了赤池信息准则（Akaike information criterion，AIC）和施瓦茨准则（Schwarze criterion，SC），这两个准则均要求仅当所增加的解释变量能减少 AIC 值或 SC 值时才在原模型中增加该解释变

表 2　回归模型的计量分析结果

Tab. 2　Econometric analysis results of the regression models

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
R^2	0.881	0.878	0.806	0.780	0.458	0.765	0.850	0.511	0.480
修正 R^2	0.872	0.855	0.789	0.771	0.413	0.756	0.831	0.470	0.437
F 检验值	89.202	39.469	49.718	88.547	11.821	81.455	43.489	12.541	11.087
P 值	0.000 1	0.000 1	0.000 1	0.000 1	0.001	0.000 1	0.000 1	0.000 1	0.000 1
AIC 值	11.767	3.202	3.237	3.622	4.534	-2.620	7.322	7.478	7.478
SC 值	11.911	3.442	3.381	3.718	4.678	-2.524	7.514	7.622	7.622

量[14]。由试验中 Eviews 6.0 的结果可知,上述回归模型 AIC 值和 SC 值均是比较理想的。

由逐步回归结果可知,表 1 中与红葡萄酒各理化指标相关性显著的酿酒红葡萄的理化指标中,指标之间关系复杂,相互影响,大多数存在多重共线性。消除多重共线性后,红葡萄酒的各理化指标仅与酿酒红葡萄的几个(甚至一个)理化指标存在确定的数量关系。但是,红葡萄酒的理化指标变化并不一定由其对应的酿酒红葡萄理化指标来决定(如回归方程 Y_2(单宁含量),其解释变量中并不包含 X_{12}(单宁含量))。另外,酿酒红葡萄中有两个指标在回归方程中比较活跃,即 X_4(花色苷含量)和 X_{11}(总酚含量)。X_4(花色苷含量)在回归方程 Y_1(花色苷含量)、Y_2(单宁含量)、Y_3(总酚含量)、Y_7(色泽 L^*)中均出现,表明该指标变化对红葡萄酒的这 4 个指标的变化均有影响,只是其影响程度不一,特别在 Y_1(花色苷含量)中影响较大;X_{11}(总酚含量)与 Y_3(总酚含量)、Y_4(酒总黄酮含量)、Y_6(DPPH 半抑制体积)存在明显的线性关系,甚至是决定 Y_4(酒总黄酮含量)和 Y_6(DPPH 半抑制体积)变化的唯一因素。

2 多元核支持向量分类机的葡萄酒口感评价分析

支持向量机(SVM)是基于 VC 维和结构风险最小化理论提出的目前比较实用的分类方法[16]。基于红葡萄酒与红葡萄理化指标数据复杂,不同指标作为输入特征含有的信息量不同。因此,对于不同的输入属性赋予不同的尺度因子,构造多元核支持向量分类机以刻画不同属性指标对输出的不同贡献。多元核支持向量分类机是标准分类机的拓展,为此,下面给出标准二分类机的一般形式。

设 $D=(\boldsymbol{x}_i,y_i)(i=1,2,\cdots,l)$ 为训练样本集,l 为训练样本的规模,对应着红葡萄(酒)样品的数量,即 $l=27$;$\boldsymbol{x}_i\in X\subset\mathbf{R}^m(i=1,2,\cdots,l)$ 为 m 维输入向量,其每一分量对应红葡萄酒或酿酒红葡萄中作为特征的某一理化指标,X 表示输入样本集,$\boldsymbol{y}_i\in Y=\{1,-1\}(i=1,2,\cdots,l)$,$Y$ 表示二分类输出样本集,引进从输入空间 $\mathbf{R}^m$ 到 Hilbert 空间 H 的变换

$$\phi:\begin{matrix}X\in{}^m\mathbf{R}^m\to H\\ x\to K(\cdot,x)\end{matrix}\tag{6}$$

则二分类支持向量机的数学形式为

$$\begin{cases}Q=\min\limits_{\omega\in H,b\in\mathbf{R},\xi\in\mathbf{R}^l}\left(C\sum\limits_{i=1}^{l}\xi_i+\dfrac{1}{2}\|\omega\|^2\right)\\ \text{s.t}\quad\begin{cases}y_i(\omega\phi(x_i)+b)\geqslant 1-\xi_i\\ \xi_i\geqslant 0\end{cases}\end{cases}\tag{7}$$

这里 ξ_i 为松弛变量,C 是惩罚参数。选取适当的核函数 $K(x,x')$ 和适当的参数 C,引入拉格朗日乘子构造并求解式(7)的对偶问题,可得

$$\begin{cases}Q=\min\limits_{\alpha}\left(\dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}y_iy_ja_ia_jK(x_i,x_j)-\sum\limits_{j=1}^{l}a_j\right)\\ \text{s.t}\quad\begin{cases}\sum\limits_{i=1}^{m}y_ia_i=0\\ 0\leqslant a_i\leqslant C\end{cases}\end{cases}\tag{8}$$

式(8)中,$K(x_i,x_j)$ 为定义在 $X\times X$ 上核函数,Q_i 为拉格朗日系数。令式(8)最优解为 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*)$,从中选取一个正分量 $0<\alpha_j^*<C$,计算阈值 $b^*=y_j-\sum\limits_{i=1}^{l}y_i\alpha_i^*K(x_i,x_j)$,最后构造决策函数 $f(x)=\operatorname{sgn}\left(\sum\limits_{i=1}^{l}\alpha_i^*y_iK(x,\boldsymbol{x}_i)+b^*\right)$,数据模式 x 的决策函数值决定了其所属的类。

为了提高二分类支持向量机的准确性,在此引入多元核(Multi-kernel)[17]。本文试验采用多元高斯核,其形式为

$$K_\sigma(x_i,x_j)=\exp\left(-\frac{1}{2}\sum_{h=1}^{m}(x_{i,h}/\sigma_h-x_{j,h}/\sigma_h)^2\right)=\exp\left(-\sum_{h=1}^{m}(x_{i,h}-x_{j,h})^2/(2\sigma_h^2)\right)\tag{9}$$

其中,K_σ 是多元核,表示不同特征的高斯核参数可取不同的最优值;$\{\sigma_h^{-1}\}_{h=1}^m$ 为多元尺度因子(核参数),控制多元核 K_σ,其最优值的获取可以采用模式搜索法[18]。引入多元核后,式(8)可变换为

$$\begin{cases}Q=\min\limits_{\alpha}\left(\dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}y_iy_ja_ia_jK_\sigma(x_i,x_j)-\sum\limits_{j=1}^{l}a_j\right)\\ \text{s.t}\quad\begin{cases}\sum\limits_{i=1}^{m}y_ia_i=0\\ 0\leqslant a_i\leqslant C\end{cases}\end{cases}\tag{10}$$

相应地,阈值计算式为

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*K_\sigma(x_i,x_j)$$

决策函数为

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i^*y_iK_\sigma(x,\boldsymbol{x}_i)+b^*\right)$$

对于本文的问题,先根据评酒员对红葡萄酒的口感评分确定各红葡萄酒样品的对应类别。分类采用如下方法:分别计算出每种红葡萄酒样品 20 位(两组)评酒员的总分,再求均值,得 27 种红葡萄酒的得分,用评分均值作为确定红葡萄酒类别的依据,把分数相近的归为同一类。本文将葡萄酒分为 4 个等级,如表 3 所示。

表 3　红葡萄酒的 4 类划分

Tab. 3　Four grade divisions of red wine

葡萄酒类别	对应样品号
1	23,9,3,20,2,17
2	19,24,21,22,26,14,5
3	16,27,13,10,4,6,8,25
4	7,11,1,18,15,12

利用 P_c，分析红葡萄酒口感评分与酿酒红葡萄和红葡萄酒理化指标之间的相关性强弱。根据 SPSS 19.0 软件计算结果，与红葡萄酒口感评分相关系数较大的理化指标有 12 项，属于酿酒红葡萄的品质指标有 6 项，分别为：花色苷含量、蛋白质含量、DPPH 自由基含量、葡萄总酚含量、pH 值、葡萄总黄酮含量；属于红葡萄酒的品质指标也有 6 项，分别为：花色苷含量、单宁含量、总酚含量、总黄酮含量、白藜芦醇含量、DPPH 半抑制体积。

以上述 12 个理化指标作为样本的输入特征，构成 12 维的输入向量 $\boldsymbol{x}_i(i=1,2,\cdots,27)$，以葡萄酒类别为输出变量 $y(y=1,2,3,4)$ 作多分类训练。一般地，一个二分类向量机只能解决两类问题。对于多类问题，可采用一对一构造方法[19]。即假定有 n 类训练样本，将每类样本与其他类样本分别构成二分类问题，共构造 $C_n^2=n(n-1)/2$ 个二分类器。测试样本经过所有的分类器进行分类，然后对所有类别进行投票，得票最多的类别即为测试样本所属的类别。本问题是一个 4 类问题，需用 $C_4^2=6$ 个二分类器。核函数选用高斯径向基（RBF）多元核，核参数 $(\{\sigma_h^{-1}\}_{h=1}^{m}, m=12)$ 采用模式搜索法以获得最优。

为检验所提方法的准确性，本节随机抽取 5 组红葡萄酒样品进行了试验，每次试验随机选取红葡萄酒测试样品后（表 4），剩余样品均作为训练样本，测试集错分样本结果如表 4。

表 4 结果表明：随机抽取葡萄酒样品作为测试集进行试验，所分类类别与实际所属类别基本相符，正确率能达到 91.89%。因此，利用多元核支持向量分类机，酿酒红葡萄和红葡萄酒的理化指标能很好地确定葡萄酒的口感评价。

表 4　多元核支持向量分类机试验结果

Tab. 4　Experiment results of multi-kernel support vector classification machine

试验序号	测试集样品号	测试集错分样本个数
1	{2、3、4、5、6、7、15}	0
2	{3、7、11、15、19、23、27}	0
3	{3、9、12、13、15、21、27}	1
4	{2、6、10、14、18、19、22}	1
5	{4、5、8、10、11、18、19、21、23}	1

3　结论

（1）与红葡萄酒各理化指标相关性较强的酿酒红葡萄理化指标中，DPPH 自由基含量、总酚含量、葡萄总黄酮含量、花色苷含量、单宁含量、蛋白质含量、褐变度和黄酮醇含量与红葡萄酒理化指标关系密切，是酿酒红葡萄的重要理化指标。

（2）酿酒红葡萄理化指标中的维生素 C 含量、柠檬酸含量、总糖含量、pH 值、可滴定酸含量、果穗质量、百粒质量和果皮质量与红葡萄酒各理化指标线性相关性不显著。

（3）红葡萄酒的部分理化指标（包括：花色苷含量、单宁含量、总酚含量、总黄酮含量、DPPH 半抑制体积、色泽 L^*、色泽 a^*）与酿酒红葡萄的对应指标均在 0.01 水平下显著相关，其线性关系明显。但是，对于红葡萄酒白藜芦醇含量和色泽 b^*，其酿酒红葡萄的对应指标白藜芦醇含量和果皮颜色（b^*）却相关性不显著。

（4）与红葡萄酒各理化指标相关性显著的酿酒红葡萄的理化指标中，指标之间关系复杂，相互影响，大多数存在多重共线性。消除多重共线性后，红葡萄酒的各理化指标仅与酿酒红葡萄的几个（甚至一个）理化指标存在确定的线性关系。

（5）利用多元核支持向量分类机，选择少数相关性强的酿酒红葡萄和红葡萄酒的理化指标就能较好地确定葡萄酒的口感评价，可为葡萄酒的质量分析提供新的思路。

参　考　文　献

1　Reddy L V A, Reddy Y H K, Reddy O V S. Wine production by guava piece immobilized yeast from Indian cultivar grapes and its volatile composition [J]. Biotechnology, 2006, 5(4): 449 ~ 454.

2　Bapat R K, Jadhav S B, Ghosh J S. Fermentation and characterization of apricot and raisin wine by *Saccharomyces cerevisiae* NCIM 3282 [J]. Research Journal of Microbiology, 2010, 5(11): 1 093 ~ 1 099.

3　Samappito S, Butkhup L. An analysis on flavonoids, phenolics and organic acids contents in brewed red wines of both non-skin contact and skin contact fermentation techniques of Mao luang ripe fruits (*Antidesma bunius*) harvested from Phupan valley in Northeast Thailand [J]. Pakistan Journal of Biological Sciences: PJBS, 2008, 11(13): 1 654 ~ 1 661.

4　Sirisantimethakom L, Laopaiboon L, Danvirutai P, et al. Olatile compounds of a traditional Thai rice wine [J]. Biotechnology, 2008, 7(3): 505 ~ 513.

5　Ogbo F C, Onuegbu J A, Achi O K. Improvement of protein content of garri by inoculation of cassava mash with biomass from

palm wine [J]. American Journal of Food Technology, 2009, 4(2): 60 ~ 65.
6 Adeleke R O, Abiodun O A. Physico-chemical properties of commercial local beverages in Osun State [J]. Pakistan Journal of Nutrition, 2010, 9(9): 853 ~ 855.
7 Cortez P, Cerdeira A, Almeida F, et al. Modeling wine preferences by data mining from physicochemical properties [J]. Decision Support Systems, 2009, 47(4): 547 ~ 553.
8 Appalasamy P, Mustapha A, Rizal N D, et al. Classification-based data mining approach for quality control in wine production [J]. Journal of Applied Sciences, 2012, 12(6): 598 ~ 601.
9 Agyemang P O. Modeling the preference of wine quality using logistic regression techniques based on physicochemical properties [D]. Youngstown: Youngstown State University, 2010: 1 ~ 84.
10 王金甲, 尹涛, 李静,等. 基于物理化学性质的葡萄酒质量的可视化评价研究[J]. 燕山大学学报, 2010, 34(2): 133 ~ 137.
Wang Jinjia, Yin Tao, Li Jing, et al. Visual evaluation of wine quality from physicochemical properties [J]. Journal of Yanshan University, 2010, 34(2): 133 ~ 137. (in Chinese)
11 吴桂芳, 蒋益虹, 王艳艳,等. 基于独立主成分和 BP 神经网络的干红葡萄酒品种的鉴别[J]. 光谱学与光谱分析, 2009, 29(5): 1 268 ~ 1 270.
Wu Guifang, Jiang Yihong, Wang Yanyan, et al. Discrimination of varieties of dry red wines based on independent component analysis and BP neural network [J]. Spectroscopy and Spectral Analysis, 2009, 29(5): 1 268 ~ 1 270. (in Chinese)
12 康俊杰, 李艳. HPLC 检测葡萄酒中有效成分的研究进展[J]. 酿酒科技, 2009(9): 112 ~ 114.
Kang Junjie, Li Yan. Research progress in the determination of effective constituents in grape wine by HPLC [J]. Liquor-making Science & Technology, 2009(9): 112 ~ 114. (in Chinese)
13 芮玉奎, 于庆泉, 金银花,等. 应用 ICP-MS 快速测定葡萄酒中 40 种元素的含量[J]. 光谱学与光谱分析, 2007, 27(5): 1 015 ~ 1 017.
Rui Yukui, Yu Qingquan, Jin Yinhua, et al. Application of ICP-MS to the detection of forty elements in wine [J]. Spectroscopy and Spectral Analysis, 2007, 27(5): 1 015 ~ 1 017. (in Chinese)
14 李子奈,潘文卿. 计量经济学[M]. 3 版. 北京:高等教育出版社, 2010.
15 赵静. 数学建模与数学实验[M]. 北京:高等教育出版社,2008: 267 ~ 269.
16 史峰, 王辉,郁磊,等. MATLAB 智能算法 30 个案例分析[M]. 北京:北京航空航天大学出版社, 2011:269 ~ 272.
17 He W, Wang Z, Jiang H. Model optimizing and feature selecting for support vector regression in time series forecasting [J]. Neurocomputing, 2008, 72(1 ~ 3): 600 ~ 611.
18 Jiang H, He W. Grey relational grade in local support vector regression for financial time series prediction [J]. Expert Systems with Applications, 2012, 39(3): 2 256 ~ 2 262.
19 许国根,贾瑛. 模式识别与智能计算的 MATLAB 实现[M]. 北京:北京航空航天大学出版社,2012:102 ~ 103.
20 陶永胜,彭传涛. 中国霞多丽干白葡萄酒香气特征与成分关联分析[J]. 农业机械学报,2012,33(3):130 ~ 139.
Tao Yongsheng, Peng Chuantao. Correlation analysis of aroma characters and volatiles in chardonnay dry white wines from five districts in China[J]. Transactions of the Chinese Society for Agricultural Machinery, 2012,33(3):130 ~ 139. (in Chinese)

Analysis of Wine Based on Pearson Coefficient and Multiple Kernel Support Vector Classification

Jiang Hui[1,2] Deng Weimin[1] Chen Xiaoqing[1]
(1. *Department of Mathematical Sciences*, *Huizhou University*, *Huizhou* 516007, *China*
2. *School of Statistics*, *Renmin University of China*, *Beijing* 100872, *China*)

Abstract: Pearson correlation coefficient was used to choose some physicochemical indexes of grape which have strong correlation with those of wine and multi factor regression equations was established to determine their quantitative relations by the stepwise regression. Each physicochemical index of wine has a specific linear relationship with several physicochemical indexes of corresponding grape or just only one. At the same time, the multi-kernel support vector machine was carried out to classify the wine samples. The results from the multi-kernel support vector machine are approximately consistent with those from the artificial with an accuracy of 91.89%. Results from this study show that the physicochemical indexes of grape and wine can determine the taste evaluation of wine well.

Key words: Wine Physicochemical index Pearson correlation coefficient Stepwise regression Support vector machine