Advances in Applied Mathematics 应用数学进展, 2015, 4(4), 376-384
Published Online November 2015 in Hans. http://www.hanspub.org/journal/aam
http://dx.doi.org/10.12677/aam.2015.44047

Hans汉斯

The Study on Evaluation System of Wine Based on Data Mining

Sizhe Wang[1], Zhigang Wang[2*], Yong He[2]

[1]Automation Professional Class 1301, School of Information Science and Engineering, Central South University, Changsha Hunan
[2]College of Information Science and Technology, Hainan University, Haikou Hainan
Email: wangsizhe@csu.edu.cn, *wzhigang@hainu.edu.cn

Received: Nov. 8th, 2015; accepted: Nov. 23rd, 2015; published: Nov. 30th, 2015

Abstract

Based on Question A of Mathematical Contest in Modeling for college students in 2012, the emphasis in this paper is mainly on the establishment of evaluation system of wine based on data mining technology. The wine quality is determined by the score of the wine tasting. We analyze the credibility of the liquor score by one-way ANOVA. We classify the wine grape by extracting common factors of some physical and chemical indicators from the wine grape, and by clustering the factor score and wine score. The stepwise regression model is established through the correlation between the physical and chemical indicators and the physical and chemical indicators of wine grapes. By the regression model between the aroma substances and the score of the wine, the key physical and chemical indicators of wine quality will be found. In the end, some shortcomings of current rating system of wine will be pointed out.

Keywords

Evaluation System of the Wine, Data Mining Technology, One-Way ANOVA, Cluster Analysis, Regression Analysis

基于数据挖掘技术的葡萄酒评价体系研究

王思哲[1]，王志刚[2*]，何　勇[2]

[1]中南大学信息科学与工程学院自动化专业1301班，湖南 长沙
[2]海南大学信息科学技术学院，海南 海口
[*]通讯作者。

文章引用: 王思哲, 王志刚, 何勇. 基于数据挖掘技术的葡萄酒评价体系研究[J]. 应用数学进展, 2015, 4(4): 376-384.
http://dx.doi.org/10.12677/aam.2015.44047

Email: wangsizhe@csu.edu.cn, *wzhigang@hainu.edu.cn

收稿日期：2015年11月8日；录用日期：2015年11月23日；发布日期：2015年11月30日

摘　要

本文以2012年高教社杯全国大学生数学建模竞赛A题为例，利用数据挖掘技术建立葡萄酒评价体系。葡萄酒质量一般是通过聘请有资质的品酒员进行品尝评分，由于品酒员主观因素导致对酒样品的评分差异悬殊，我们通过方差分析对品酒员评分进行可信性研究；通过提取酿酒葡萄多个理化指标的公共因子，对因子得分和葡萄酒评分进行聚类，将酿酒葡萄进行分级研究；通过对葡萄酒理化指标和酿酒葡萄理化的数据进行相关性分析，利用逐步回归分析模型建立它们之间的依赖关系；利用葡萄酒芳香物质与葡萄酒评分之间的回归模型，找出决定葡萄酒质量的关键理化指标，最后指出现行葡萄酒评分体系的不足。

关键词

葡萄酒评价，数据挖掘技术，方差分析，聚类分析，回归分析

1. 引言

在当今大数据时代，从数据库的挖掘出隐含的、先前未知的并有潜在价值的信息显得十分重要，多元统计方法是数据挖掘技术的关键要素。多元统计分析是处理多维同体观测数据的数学方法，是数理统计学近几十年迅速发展的一个分支，计算机技术的发展为多元统计的方法应用提供了便利的计算工具。多元统计的内容十分丰富，主要包括判别分析、聚类分析、主成分分析、因子分析、回归分析预测方法、方差分析、典型相关分析、时间序列等[1]-[11]。多元统计方法在工业、农业、医学、气象、环境以及经济管理等诸多领域中有着十分广泛的应用。

本文以 2012 年高教社杯全国大学生数学建模竞赛 A 题为例，用多元统计序列方法建立葡萄酒评价体系。确定葡萄酒质量时一般是通过聘请一批有资质的品酒员进行品评，每个品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(见竞赛试题中的附件，本文略，下同)给出了某一年份两组品酒员对两组红葡萄酒和白葡萄酒的评分结果，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，包括各种理化指标和芳香物质指标数据。

由于品酒员主观因素导致对酒样品的评分差异，我们分别构造以品酒员和酒样品为组别数据序列进行方差分析，通过比较 F 统计量值评价两组品酒员是否有显著性差异，对品酒员评分进行可信性研究；通过对酿酒葡萄的多个理化指标进行筛选，提取公共因子，并计算因子得分，然后将这些因子得分和葡萄酒评分进行聚类分析，将酿酒葡萄进行分级研究；通过对葡萄酒理化指标和酿酒葡萄理化指标相关性分析，利用逐步回归模型建立它们之间的线性关系；通过葡萄酒理化指标与葡萄酒评分之间的回归模型，建立酿酒葡萄理化指标与葡萄酒质量之间关系，给出决定葡萄酒质量的关键理化指标。

2. 葡萄酒评分的可信性研究

考虑到品酒员之间可能存在个人评酒风格等主观差异因素，导致不同品酒员对同一葡萄酒的评分悬殊，影响葡萄酒质量鉴定，因此，必须对品酒员的评分主观因素进行检验。附件 1 给出了两组红葡萄酒

品酒员对 27 个酒样品的评价得分和两组白葡萄酒品酒员对 28 组酒样品的评价得分。

对于评酒得分的偏差性检验和影响因素的数据挖掘技术，可以通过方差分析来实现。方差分析主要是检验两组品酒员评价结果有无显著性差异，进而判断出哪组评价结果更为可信。评价得分之间的差异可以分为两个部分，一部分是由于各葡萄酒样品之间的差异，称为条件误差，另一部分是各品酒员评酒风格之间的差异，称为试验误差，我们主要目的是分析得分差异是由于葡萄酒样品之间差异，还是由于品酒员主观差异造成的。通过对两组红葡萄酒和两组白葡萄酒评价得分进行正态性检验可以看出都近似服从正态分布，我们分别构造以品酒员和酒样品为组别的数据序列进行方差分析(见表 1)。

分析表 1 数据，基于品酒员和酒样品的显著性差异检验中，除第二组白葡萄酒酒样品差异不显著外，另七组的 F 统计量都大于基于显著性水平 0.01 的临界值，表明品酒员评酒风格和酒样品之间的差异都很显著。进一步比较 F 统计量数值大小，第一组红葡萄酒评分差异主要来源于酒样品之间的差异，第二组红葡萄酒评分差异主要来源于品酒员评分差异；白葡萄酒评分差异主要来源于品酒员评分差异，酒样品之间的差异不很显著。初步可以看出，对于红葡萄酒，第一组品酒员评分更为可信，两组白葡萄酒品酒员评分都不可信，品酒员间的差异过大将导致酒样质量差异的显著性被掩盖，结合实际分析，酒样评价中应尽可能缩小由于品酒员个人风格的原因而导致对同一酒样评价差异较大的情况，应尽可能将酒样之间质量的差异通过评价扩大，提高酒样的可识别度。为此，将原始数据进行处理，原始数据进行处理方法有很多，如标准化处理、聚类处理、收敛区间处理等，我们采用数据标准化处理，降低品酒员之间的主观差异性(见表 2)。

Table 1. Wine score variance analysis based on the raw data
表 1. 基于原始数据的葡萄酒评分方差分析表

	差异源	总平方和	自由度	均方差	F 统计量	F 临界值
第一组红葡萄酒	品酒员	3084.952	9	342.772	3.543	2.484
	酒样品	14,090.119	26	541.928	9.308	1.837
第二组红葡萄酒	品酒员	3228.681	9	358.742	9.999	2.484
	酒样品	4186.830	26	161.032	4.675	1.837
第一组白葡萄酒	品酒员	17,034.122	9	1892.68	26.830	2.481
	酒样品	6253.086	27	231.596	1.957	1.818
第二组白葡萄酒	品酒员	6714.442	9	746.049	19.910	2.481
	酒样品	2714.811	27	100.549	1.795	1.818

Table 2. Wine score variance analysis based on data standardization
表 2. 基于数据标准化处理的葡萄酒评分方差分析表

	差异源	总平方和	自由度	均方差	F 统计量	F 临界值
第一组红葡萄酒	品酒员	0	9	0	0	2.484
	酒样品	152.698	26	5.873	13.300	1.837
第二组红葡萄酒	品酒员	0	9	0	0	2.484
	酒样品	119.148	26	4.583	7.906	1.837
第一组白葡萄酒	品酒员	0.196	9	0.022	0.022	2.481
	酒样品	93.467	27	3.462	4.942	1.818
第二组白葡萄酒	品酒员	0.006	9	0.001	0.001	2.481
	酒样品	76.147	27	2.820	3.666	1.818

分析表 2 数据，对于四组品酒员评价数据序列，用于检验的 F 统计量值都接近于 0，远低于基于显著性水平 0.01 的 F 临界值，四组酒样品数据序列的 F 统计量都大于基于显著性水平 0.01 的 F 临界值。从数据层面上分析，对于红葡萄酒，通过数据标准化将品酒员评分差异明显降低，造成葡萄酒评分差异的主要因素是酒样品之间的差异，第一组红葡萄酒样品之间的差异比第二组红葡萄酒之间的差异更为显著，由此，第一组红葡萄酒品酒员评分更为可信，与基于原始数据的评价结果一致。对于白葡萄酒，通过标准化数据处理后品酒员之间的差异明显降低，其中第二组品酒员之间的差异的 F 统计量为 0.001，比第一组更接近 0，第一组白葡萄酒样品之间的差异比第二组白葡萄酒之间的差异更为显著，综合可以看出对于白葡萄酒，第二组品酒员评价结果更为可信。

需要指出的是，通过标准化数据处理后，品酒员之间的差异很小，仅凭酒样品的 F 统计量大小，很难精确表明哪组品酒员评价结果更为可信，综合其它方法进行处理(如置信区间法等)会收到更好的效果。

3. 因子分析法对酿酒葡萄聚类研究

酿酒葡萄的好坏与所酿葡萄酒的质量有直接关系，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，为了进一步研究酿酒葡萄的理化指标对葡萄酒质量的影响程度，需要建立它们之间的关系。附件 2 提供的酿酒葡萄的一级理化指标 30 种，经过初步筛选可知，固酸比指标值为可溶性固形物指标与可滴定酸指标之比，可以提出可溶性固形物指标和可滴定酸指标。附件 2 中存在异常值，例如酿酒葡萄理化指标中白葡萄百粒质量的第三次测量值为 2226.1 g，与前两次检测值 225.8 g 和 224.6 g 相差很大，为了避免数据异常值对数据分析带来影响，用前两次平均值来代替异常值。由于每种指标数据数值大小和波动幅度不同，为了消除这一因素的影响，对 28 种理化指标进行标准化处理。

为了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研究，将第一部分得到的葡萄酒评分作为葡萄酒质量的标准，由于酿酒葡萄理化指标较多，全部考虑会比较繁琐，且不同指标之间不是独立的，彼此之间存在一定的相关关系，可以对原始变量进行浓缩，将原始变量的信息重叠部分综合成因子，利用几个综合变量来反应这些指标之间的关系。综合多元统计因子分析思想，将 28 种指标提取公共因子，并计算因子得分，然后将这些因子得分和葡萄酒评分进行聚类分析，将酿酒葡萄进行分级研究。

利用 SPSS19.0 求出各变量之间的相关系数矩阵，可以看出，几个变量之间存在强相关性，例如，蛋白质与 DPPH 自由基(相关系数 0.748)，苹果酸与花色苷(0.633)，苹果酸与褐变度(0.644)，总酚与花色苷(0.728)，花色苷与单宁(0.688)等，考虑抽取共同因子，利用主成分分析法、考虑特征值大于或等于 1 的主成分作为初始因子，抽取 8 个共同因子。为了使抽取的因子作有效的解释，采用最大变异法对因子进行旋转，并计算因子得分作为变量储存，为下一步对酿酒葡萄进行聚类分析的作好准备。抽取的 8 个因子对方差的累计贡献率为 83.035%，且前 3 个因子对方差的贡献率分别为 18.446%，13.058%和 11.097%，能够反应原始变量的大部分信息，效果较好。

由第一部分的分析，对于红葡萄酒，第一组品酒员的评价结果更为可信，利用第一组品酒员对红葡萄酒的评价得分作为对葡萄酒的聚类分级的依据，结合抽取的 8 个因子得分，利用 SPSS19.0 中的 k-均值 Q 型聚类方法，如果要将葡萄酒分为 4 级，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的聚类结果(见表 3)。

27 种红酒酿酒葡萄样品有 13 种为优秀等级，有 7 种为良好等级，有 1 种为中等等级，有 6 种为差等级。

白葡萄酒的分析类似，将 28 种葡萄样品的 28 种理化指标进行因子分析，抽取公共因子，抽取的 9 个因子对方差的累计贡献率为 80.63%，且前 3 个因子对方差的贡献率分别为 18.770%，17.200%和 11.701%，

Table 3. The classification table of red wine grape
表 3. 红种酿酒葡萄等级分类表

样本编号	1	2	3	4	5	6	7	8	9
所属等级	差	良	良	优	差	优	优	差	良
样本编号	10	11	12	13	14	15	16	17	18
所属等级	优	中	优	优	差	差	优	良	优
样本编号	19	20	21	22	23	24	25	26	27
所属等级	优	优	良	优	良	优	优	差	优

能够反应原始变量的大部分信息，并计算因子得分，考虑第二组评分更为可信，结合第二组品酒员的评价得分，进行聚类分级，得到白种酿酒葡萄的聚类结果(见表 4)。

28 种白种酿酒葡萄样品种有 7 种为优秀等级，有 10 种为良好等级，有 10 种为中等等级，有 1 种为差等级。

需要指出的是，聚类分析的基本思想是每个样本称为一类，计算类内样本间的距离，将最近的两个类聚为一类，然后再计算新类间的距离，并将最近的两个类聚为一类，如此循环。K 均值聚类法最终的分类结果在某种程度上与初始分类有关，由于计算距离的方法很多，为了得到更为精确的分级结果，可以结合最短距离法、重心法、类平均法、可变类平均法、离差平均法等综合考虑。同时，由于酿酒葡萄理化指标较多，如果都作为分类依据计算距离会导致误差加大，抽取公共因子会损失部分信息，对最终的分级造成一定的影响。传统的 K-means 倾向于将离得近的点分在一类，但是对于子空间相交处，离得近的点并不是同一类的点，所以 K-means 这种仅仅基于距离的做法在原理上是不符合子空间聚类的，可以结合最新的研究成果，比如稀疏子空间聚类算法、谱多流形聚类算法等会收到好的效果[12] [13]。

4. 酿酒葡萄理化指标与葡萄酒理化指标之间的关系研究

酿酒葡萄的理化指标直接影响葡萄酒的理化指标，它们之间并不是相互影响而是一种因果关系，可以建立每个葡萄酒的理化指标与酿酒葡萄多个理化指标之间的联系来分析酿酒葡萄理化指标对葡萄酒理化指标的影响。考虑到酿酒葡萄理化指标(28 种一级指标)和葡萄酒理化指标(红葡萄酒 9 种，白葡萄酒 8 种)都较多，且彼此之间存在一定的相关性，样本较少，选取过多的酿酒葡萄理化指标会产生较大的误差，且很难保证模型通过检验。利用多元统计逐步回归思想，通过计算酿酒葡萄的理化指标与葡萄酒的理化指标之间的 Person 相关系数对酿酒葡萄指标进行筛选，筛选出的酿酒葡萄理化指标与葡萄酒理化指标进行回归分析。表 5 列出部分红葡萄酒理化指标和酿酒葡萄理化指标相关系数，作为筛选指标的依据(见表 5)。

利用 SPSS19.0，采用逐步筛选法建立红葡萄酒理化指标和酿酒葡萄理化指标之间的关系，每个回归方程需通过拟合优度检验和回归系数的检验

$$\text{酒花色苷指标} = 0.713 * \text{花色苷} + 0.115 * \text{苹果酸} + 0.197 * \text{褐变度} \quad (1)$$

其中模型调整后的样本决定系数 $R^2 = 0.942$，模型检验的 F 统计量值为 63.607，表 6 给出系数检验的 t 统计量分别为 7.067，1.213 和 1.926，模型检验通过，三个自变量的容忍度分别为 0.457，0.520，0.447，作出残差序列的正态检验的 P-P 图，残差序列服从正态分布。能反映酒花色苷指标与酿酒葡萄理化指标之间的关系(见表 6)。

模型的 Durbin-Watson 统计量的值为 2.853，表明残差序列存在一定程度的负相关，一些与酒花色苷相关的某些酿酒葡萄理化指标没有引进回归方程来；进一步，表 7 给出了特征根法自变量共线性性诊断表，特征根 0.303 既能刻画自变量花色苷方差较大部分比例(0.68)，又能刻画自变量褐变度方差较大部分比例(0.80)，因此，这两个自变量存在一定的多重线性性(见表 7)。

Table 4. The classification table of white wine grape
表 4. 白种酿酒葡萄等级分类表

样本编号	1	2	3	4	5	6	7	8	9	10
所属等级	优	中	中	良	中	良	良	优	优	中
样本编号	11	12	13	14	15	16	17	18	19	20
所属等级	良	中	中	良	中	优	良	良	优	优
样本编号	21	22	23	24	25	26	27	28		
所属等级	良	优	中	中	良	中	差	良		

Table 5. Correlation coefficient of the standard data of physical and chemical index between the red wine and the wine grape
表 5. 基于数据标准化红葡萄酒与酿酒葡萄理化指标之间相关系数

酒指标 / 酿酒葡萄指标	花色苷	单宁	总酚	酒总黄酮	白藜芦醇	*DPPH 半抑制体积*
花色苷	0.923	0.720	0.774	0.709		0.671
总酚	0.613	0.817	0.875	0.883	0.459	0.701
单宁	0.661	0.718	0.743	0.701	0.315	
DPPH 自由基	0.567	0.753	0.814	0.764	0.421	0.778
褐变度	0.767	0.445		0.443		
葡萄总黄酮		0.684	0.815	0.823	0.567	0.813

Table 6. Regression coefficient table of Model 1
表 6. 模型 1 回归系数表

Coefficients[a,b]							
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	花色苷	0.713	0.101	0.713	7.067	0.000	0.457	2.187
	苹果酸	0.115	0.095	0.115	1.213	0.237	0.520	1.925
	褐变度	0.197	0.102	0.197	1.926	0.066	0.447	2.240

a. Dependent Variable：花色苷，b. Linear Regression through the Origin

Table 7. The total linear diagnosis table of independent variable of Model 1 characteristic root method
表 7. 模型 1 特征根法自变量共线性性诊断表

Collinearity Diagnostics[a,b]						
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				花色苷	苹果酸(g/L)	褐变度
1	1	2.315	1.000	0.07	0.07	0.07
	2	0.381	2.464	0.25	0.92	0.14
	3	0.303	2.763	0.68	0.01	0.80

a. Dependent Variable：花色苷；b. Linear Regression through the Origin

类似地可以建立其它红葡萄酒理化指标与酿酒葡萄理化指标之间的关系。

$$\text{酒单宁指标} = 0.087*\text{花色苷} + 0.371*\text{褐变度} + 0.667*\text{DPPH自由基} - 0.205*\text{果梗比} + 0.519*\text{氨基酸总量} + 0.050*\text{总酚} - 0.101*\text{蛋白质} \quad (2)$$

其中 $R^2 = 0.844$，模型检验的 F 统计量值为 21.810。

$$\text{酒总酚} = 0.428*\text{褐变度} + 1.046*\text{DPPH自由基} - 0.313*\text{果梗比} + 0.381*\text{氨基酸总量} - 0.323*\text{蛋白质} \quad (3)$$

其中 $R^2 = 0.890$，模型检验的 F 统计量值为 44.743。

$$\text{酒总黄酮} = 0.217*\text{褐变度} + 0.147*\text{DPPH自由基} - 0.174*\text{果梗比} + 0.747*\text{总酚} \quad (4)$$

其中 $R^2 = 0.784$，模型检验的 F 统计量值为 25.462。

$$\text{酒白藜芦醇} = 0.629*\text{花色苷} - 0.800*\text{苹果酸} + 0.885*\text{DPPH自由基} - 1.271*\text{蛋白质} + 0.424*\text{酒石酸} \quad (5)$$

其中 $R^2 = 0.689$，模型检验的 F 统计量值为 12.927。

$$\text{酒DPPH半抑制体积} = 0.397*\text{褐变度} + 1.082*\text{DPPH自由基} - 0.387*\text{果梗比} + 0.458*\text{氨基酸总量} - 0.369*\text{蛋白质} \quad (6)$$

其中 $R^2 = 0.885$，模型检验的 F 统计量值为 42.593。

$$\text{酒色泽L} = -0.604*\text{花色苷} + 0.355*\text{果皮颜色}(\text{红绿}) - 0.234*\text{黄酮醇} - 0.184*\text{酒石酸} \quad (7)$$

其中 $R^2 = 0.862$，模型检验的 F 统计量值为 43.197。

$$\text{酒色泽a} = -0.369*\text{苹果酸} + 0.587*\text{还原糖} + 0.352*\text{酒石酸} \quad (8)$$

其中 $R^2 = 0.562$，模型检验的 F 统计量值为 12.534。

$$\text{酒色泽b} = -511*\text{花色苷} - 0.157*\text{苹果酸} - 0.712*\text{果皮颜色}(\text{红绿}) \quad (9)$$

其中 $R^2 = 0.608$，模型检验的 F 统计量值为 14.961。

需要指出的是，上述建立的回归方程残差序列存在一定的自相关性，回归模型存在误差，为了得到更精确的回归方程，可以将残差序列用 ARMA 模型自回归研究，提取自相关部分信息，也可以通过因子分析法抽取共同因子建立回归方程，但这时不能很好地直观地反应葡萄酒理化指标和酿酒葡萄理化指标之间的相关关系，不宜采用。

同样可以建立白葡萄酒理化指标与酿酒葡萄理化指标之间的关系式，限于篇幅，不再一一说明。

5. 酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响研究

由于葡萄酒理化指标与酿酒葡萄的理化指标之间存在高度的相关性，葡萄酒理化指标是影响葡萄酒质量的最直接因素，酿酒葡萄理化指标是通过葡萄酒理化指标间接影响葡萄酒质量的。因此，建立葡萄酒质量与葡萄酒理化指标之间的函数关系式，分析葡萄酒理化指标是否对葡萄酒质量有一定的影响。

从前面的分析我们知道，对于红葡萄酒，第一组品酒员的评价结果更为可信，我们采用第一组品酒员的得分作为红葡萄酒的质量评分，将数据标准化，利用 SPSS19.0，采用逐步筛选法，删除系数不显著变量，建立葡萄酒质量与葡萄酒理化指标之间的函数关系。

$$\text{红葡萄酒得分} = 0.504*\text{白藜芦醇} \quad (10)$$

其中模型的样本决定系数 $R^2 = 0.225$，模型检验的 F 统计量为 8.846。从模型可以看出，红葡萄酒的理化指标不能很好地反应葡萄酒的质量，由于酿酒葡萄的理化指标与葡萄酒的理化指标之间的高度相关性，可以看出酿酒葡萄的理化指标对葡萄酒质量的影响程度与葡萄酒的理化指标对葡萄酒质量的影响程度基

本一致。

通过查阅相关资料，葡萄酒的芳香物质对酿葡萄酒气味、口感等方面有很大的影响，可以通过酿酒葡萄或葡萄酒的芳香物质来评价葡萄酒的质量。在分析芳香物质对葡萄酒质量的影响时，我们将葡萄酒的评价指标分为外观分析、口感分析、香气分析和整体评价，芳香物质主要影响葡萄酒的香气和口感，根据附件的评价体系，香气分析和口感分析占比例为 74%，因此利用葡萄酒的芳香物质来评价葡萄酒质量是可行的。考虑到葡萄酒的芳香物质众多，并且有的芳香物质之间存在相关关系，通过建立芳香物质之间的相关矩阵，筛选出评价葡萄酒质量的芳香物质，并将数据进行标准化处理，建立葡萄酒质量与芳香物质之间的函数关系。

$$\begin{aligned}\text{红葡萄酒得分} = {} & 0.479*\text{乙醇} - 2.240*\text{柠檬烯} - 1.452*\text{正十三烷} - 0.182*\text{辛酸甲酯} \\ & + 0.238*\text{3,7二甲基} + 3.123*\text{反式-4-葵烯酸乙酯}\end{aligned} \tag{11}$$

模型的样本决定系数 $R^2 = 0.998$，模型检验的 F 统计量为 561.264。系数能通过检验，表明红葡萄酒评分和芳香物质之间存在一定的关系，红葡萄酒的感官指标可以在一定程度上通过酿酒葡萄和芳香物质来评价。

白葡萄酒得分与芳香物质之间回归方程的样本决定系数都很低，且系数不能通过检验，表明白葡萄酒指标和芳香物质不能拟合葡萄酒的质量。主要原因除回归方法自身缺陷外，葡萄酒的质量评分存在一定的主观性和经验性，评分的高低不一定能反应葡萄酒质量的好坏。同时表明白葡萄酒评分制度的不完善，或者白葡萄酒评分员的评分精度没有红葡萄酒高，需要建立更加完善的白葡萄酒评分体系，提高白葡萄酒品酒员的品酒水平。主要原因是在葡萄酒市场中，红葡萄酒占有份额为 95%，品酒员评分经验都占有绝对优势，随着白葡萄酒市场份额的增加和白葡萄酒品酒员素质的提高，相信在不久的将来，白葡萄酒评分会更加真实和可信，能更好地反应与葡萄酒芳香物质的关系。

6. 结论

本文的主要工作是利用数据挖掘技术建立葡萄酒评价体系。通过方差分析对品酒员评分进行可信性研究，消除因品酒员主观因素对酒质量造成的影响；通过提取酿酒葡萄多个理化指标的公共因子，对因子得分和葡萄酒评分进行聚类，将酿酒葡萄进行分级研究；通过对葡萄酒理化指标和酿酒葡萄理化指标相关性分析，利用逐步回归模型建立它们之间的依赖关系；利用葡萄酒芳香物质与葡萄酒评分之间的回归模型，找出决定葡萄酒质量的关键理化指标，同时指出现行葡萄酒评分体系的不足并提出改进方案。

基金项目

海南省自然科学基金项目(20151002)、海南省中西部高校提升综合实力工作资金项目和海南大学教育教学改革研究项目(hdjy1639)资助。

参考文献 (References)

[1] 吴孟达, 成礼智, 等. 数学建模教程[M]. 北京: 高等教育出版社, 2008.

[2] 陈光亭, 裘哲勇, 主编. 数学建模[M]. 北京: 高等教育出版社, 2010.

[3] 杨小平. 统计方法与 SPSS 应用教程[M]. 北京: 清华大学出版社, 2008.

[4] 姜启源, 谢金星, 叶俊. 数学模型[M]. 北京: 高等教育出版社, 2003.

[5] 高新波. 模糊聚类分析及其应用[M]. 西安: 西安电子科技大学出版社, 2004.

[6] 朱道元, 吴诚鸥, 秦伟良. 多元统计与软件 SAS[M]. 南京: 东南大学出版社, 2003.

[7] 汪晓银, 邹庭荣. 数学软件与数学实验[M]. 北京: 科学出版社, 2008.

[8] 王志刚. 应用随机过程[M]. 合肥: 中国科学技术大学出版社, 2009.

[9] 欧宜贵. 数学实验[M]. 合肥: 中国科学技术大学出版社, 2012.

[10] 潘伟, 王志刚. 概率论与数理统计[M]. 北京: 高等教育出版社, 2010.

[11] 符一平, 王志刚. 基于 ARCH 模型对上证综指的实证研究[J]. 应用数学进展, 2015, 4(2): 124-128.

[12] Tang, K.W., Liu, R.S., Su, Z.X. and Zhang, J. (2014) Structure-Constrained Low-Rank Representation. *IEEE Transactions on Neural Networks and Learning Systems*, **25**, 2167-2179. http://dx.doi.org/10.1109/TNNLS.2014.2306063

[13] Elhamifar, E. and Vidal, R. (2013) Sparse Subspace Clustering: Algorithm, Theory, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2765-2781. http://dx.doi.org/10.1109/TPAMI.2013.57