基于多元统计分析的葡萄酒评价与指标关联研究

丁 亮[1]* 许 文[1] 武 林[2] 刘清民[1]

(1、中国科学技术信息研究所,北京 100038 2、重庆理工大学保卫处信息科,重庆 400054)

摘 要:现行葡萄酒评价最有效的方式是通过聘请一批有资质的评酒员进行品评、给分来确定。而针对酿酒葡萄和葡萄酒的理化指标进行酿酒葡萄的分级和葡萄酒得质量评价尚需做进一步的探讨。本文通过主成份分析法、聚类分析法、典型相关分析等方法,依据酿酒葡萄和葡萄酒的理化指标建立了基于主成分分析和聚类分析的酿酒葡萄分级模型、基于典型相关分析的指标联系模型。此项研究可以通过大数据背景下的数据建模分析,减少葡萄酒评级中人工工作量,给予品酒师大数据模型决策。

关键词:多元统计分析;主成分分析;聚类分析;典型相关分析;葡萄酒评级;大数据建模

1 概述

现行葡萄酒评价最终、最有效的方式是聘请一批有资质的评酒员进行品评。酿酒葡萄的质量与所酿葡萄酒的品质有直接关系,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,本研究试建立模型解决以下问题:分析酿酒葡萄与葡萄酒的理化指标之间的联系;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。此项研究可以通过大数据背景下的数据建模分析,减少葡萄酒评级中人工工作量,给予品酒师大数据模型决策。国际上关于葡萄酒感官评价的研究很多,并且在应用中已经建立感官评价部门,但是芳香的最适宜量尚无公开数据。国内感官评价起步较晚(20 世纪 70 年代),目前国内此方面的研究主要侧重于成分、口感及其影响因素、感官分析保证体系等方面。前国际、国内尚无对酿酒葡萄等级分类的标准,对葡萄的分级研究意义在于提供生产高质量葡萄酒的酿酒葡萄的选用标准。

2 基于主成分分析和聚类分析的酿酒葡萄分级模型

因为酿酒葡萄的理化指标非常丰富繁多[1],本文在这里用主成分分析对理化指标进行信息的集中和降维,利用主成分向量和最短距离聚类法对酿酒葡萄品种进行聚类,得到酿酒葡萄的类型,再依据可信的评酒师的感官评价指标,对酿酒葡萄的类型进行分级。在降维的方法中,主成分分析法(PCA)[2]是一种比较优秀的降维方法,主成分分析法的基本思想是根据变量之间的相关性,通过正交变换用较少的变量来替代原来较多的变量,起到降维和信息集中的作用,从而便于分类或分级。

假设进行主成分分析的指标变量有 m 个,分别为 $x_1, x_2, ..., x_m$,共有 n 个评价对象,第 i 个评价对象的第 j 个指标的取值为 a_{ij}。将各个指标 a_{ij} 转化成标准化指标值 $\overline{a_{ij}}$,有

$$\overline{a_{ij}} = \frac{a_{ij} - \mu_j}{s_j}, i = 1,2,...,n; j = 1,2,...m$$

其中: $\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2}, j = 1,2,...,m,$

即 μ_j, s_j 为第 j 个指标的样本均值和样本标准差。对应地,标准化指标变量如下:$\overline{x_j} = \frac{x_j - \mu_j}{s_j}, j = 1,2,...m$

相关系数矩阵记为 $R = (r_{ij})_{m\times m}$,其中:

$$r_{ij} = \frac{\sum_{k=1}^{n} \overline{a_{ki}} \bullet \overline{a_{kj}}}{n-1}, i, j = 1,2,...,m$$

这里 $r_{ii} = 1, r_{ij} = r_j, r_{ij}$ 是第 i 个指标与第 j 个指标的相关系数。我们分别对红葡萄和白葡萄进行主成分分析,分别得到其 m 个主成分,再根据荷载矩阵,选取总方差 86%以上贡献的前几个为主成分,利用它们构建红葡萄或者白葡萄的聚类所需要的理化指标的特征向量。因为酿酒葡萄品种本身并没有进行分类,本文先利用最短距离聚类法对酿酒葡萄进行聚类,获取酿酒葡萄的类别,最短距离聚类法如下:$D(G_1, G_2) = \min_{\substack{y_j \in G_2 \\ x_i \in G_1}} \{d(x_i, y_j)\}$

通过此步分析和聚类,对葡萄酒的进行分级。以 2012 年全国数学建模竞赛葡萄酒评级题目数据为例进行试验,通过表 1 可知,对葡萄酒的理化特征提取 10 个主成分,即 m=10。主因子分析的目的之一是用尽可能少的因子来解释观测到得变量,主成分的特征根及贡献率是选择主成分的依据,从表 1 可知总方差的 86.269%的贡献来自前 10 个因子,即认为一个 10 主成分模型可以解释实验数据的 86.269%。

表 1 葡萄酒的主成分分析表

解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.606	19.016	19.016	7.606	19.016	19.016
2	6.201	15.504	34.520	6.201	15.504	34.520
3	5.038	12.595	47.115	5.038	12.595	47.115
4	4.137	10.341	57.456	4.137	10.341	57.456
5	2.892	7.229	64.686	2.892	7.229	64.686
6	2.452	6.130	70.816	2.452	6.130	70.816
7	2.030	5.074	75.890	2.030	5.074	75.890
8	1.604	4.011	79.901	1.604	4.011	79.901
9	1.394	3.486	83.387	1.394	3.486	83.387
10	1.153	2.882	86.269	1.153	2.882	86.269
11	.976	2.440	88.709			
12	.881	2.202	90.910			
13	.745	1.863	92.773			
14	.611	1.527	94.300			
15	.517	1.293	95.593			
16	.366	.916	96.509			
17	.349	.872	97.381			
18	.322	.806	98.187			
19	.269	.673	98.860			
20	.202	.506	99.365			
21	.102	.254	99.620			
22	.061	.153	99.772			
23	.050	.124	99.897			
24	.041	.103	100.000			
25	5.142E-16	1.286E-15	100.000			
26	3.897E-16	9.741E-16	100.000			
27	3.234E-16	8.086E-16	100.000			
28	2.491E-16	6.227E-16	100.000			
29	1.781E-16	4.453E-16	100.000			
30	1.619E-16	4.047E-16	100.000			

根据加入第二组评酒员对红葡萄酒的评分作为聚类指标之

作者简介:丁亮(1994-),男,硕士研究生,主要研究方向:机器翻译,自然语言处理;许文(1991-),女,硕士研究生,主要研究方向:信息分析与数据挖掘;武林(1991-),男,硕士研究生,主要研究方向:大数据技术;刘清民(1993-),男,硕士研究生,主要研究方向:自然语言处理。通讯作者:丁亮。

一和其余的 40 个指标一起对红葡萄进行最短距离法的 Q 聚类，得到红葡萄的聚类图如下：

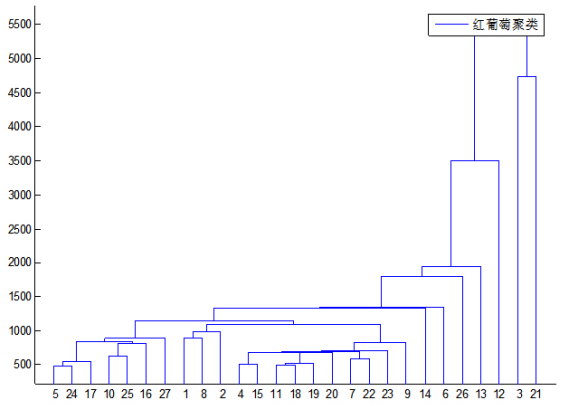

图 1 红葡萄酒聚类分析图

可以得到红葡萄酒的聚类结果如下：第 1 类有：4,7,9,11,15,18,19,20,22,23；第 2 类有：1,2,8；第 3 类有：5,10,16,17,24,25,27；第 4 类有：14；第 5 类有：6；第 6 类有：26；第 7 类有：13；第 8 类有：12；第 9 类有：3；第 10 类有：21。利用评酒员的评分进行进一步的分级，品酒员对上述 10 类的评分如下：第一类的得分为：71.2,65.3,78.2,61.6,65.7,65.4,72.6,75.8,71.6,77.1；第二类的得分为：68.1,75.8,66；第三类的得分为：72.1,68.8,69.9,74.5,71.5,68.2,71.5；第四类的得分为：72.6；第五类的得分为：66.3；第六类的得分为：72；第七类的得分为：68.8；第八类的得分为：68.3；第九类的得分为：74.6；第十类的得分为：72.2。综合聚类分析和与品酒员评分得到葡萄分级结果，见表 2 所示。

表 2 红葡萄酒分级表

红葡萄酒分级表	
第一级	14, 6, 26, 13, 12, 3, 21
第二级	5, 10, 16, 17, 24, 25, 27
第三级	2, 4, 7, 9, 19, 20, 22, 23
第四级	11, 15, 18

3 基于典型相关分析的指标联系模型

对于红葡萄酒的非芳香物质指标利用 matlab 进行典型性相关分析，得到如下的酿酒葡萄与对应葡萄酒理化指标主成分之间的相关系数表和典型相关系数表，如表 3 所示。

表 3 酿酒葡萄与对应葡萄酒理化指标主成分之间的相关系数表

	v1	v2	v3	v4	v5	v6	v7
y1	0.180009	0.134097	0.507853	0.407436	−0.71925	−0.0185	−0.08966
y2	0.158224	0.249137	0.22825	0.728004	−0.36673	−0.44228	0.026484
y3	0.013105	0.068639	0.157224	0.842227	−0.46777	−0.1775	0.10361
y4	0.293844	0.174109	−0.11741	0.763331	−0.49194	−0.0736	0.198641
y5	−0.32124	0.204405	−0.15922	0.78146	0.146426	0.31581	−0.31275
y6	0.204459	0.054552	0.011995	0.884185	−0.35723	−0.20965	−0.04158
y7	−0.05597	−0.07922	−0.27255	−0.39338	0.847609	0.190302	0.083074

可以看出，所有的七个表示红葡萄酒理化指标的变量与 u1 有大致相同的相关系数，u1 视为形容葡萄酒的指标。例如第一对典型变量的第五个成员 v5 与 y7 有较大的相关系数，说明 v5 主要代表了色泽，而 v5 与 u5 之间的相关系数为 0.99。酿酒红葡萄组的原始变量被 u1~u7 解释的比率为 41%；红葡萄酒组的原始变量被 v1~v7 解释的比率为 100%；对于红葡萄芳香类物质和白葡萄的典型性分析计算过程同上。由于篇幅限制，故只列出结果如下 酿酒红葡萄芳香物质组的原始变量被 u1~u55 解释的比率为 49.50% 红葡萄酒芳香物质组的原始变量被 v1~v55 解释的比率为 100%；酿酒白葡萄组的原始变量被 u1~u7 解释的比率为 61.32% 白葡萄酒组的原始变量被 v1~v7 解释的比率为 100% 酿酒白葡萄芳香物质组的原始变量被 u1~u55 解释的比率为 51.96%；白葡萄酒芳香物质组的原始变量被 v1~v55 解释的比率为 100%。

4 模型改进

对于葡萄酒评级问题，我们可以在已有分类条件的基础上(因为以下模型在没有分类标准时，是无法实现的，原因是进机器学习过程需要输入观测样本) 采用基于现代优化算法对 SVM 分类器修正参数的模式识别模型来对重新葡萄分类，提高分类优化精度，也可以通过训练单层前向神经网络设计神经网络分类器做同样的工作。对于葡萄酒理化指标之间的关联度，我们可以通过数据挖掘的其他算法加以对细节、隐秘关系的揭示。

5 总结与展望

对于主成分分析和聚类分析模型，主成分分析得到了方差累计度 86%以上的主成分，结果比较可信，通过主成分选出的特征指标更有代表性。Q 聚类分析举出来的类别和评分对照得出葡萄酒品级更有说服力。但是这种聚类不具有智能性，且在求解精度和算法复杂度方面欠佳。

对于典型相关分析模型，选用了所有葡萄和葡萄酒的一级理化指标，来描述其指标相关性和被解释的方差比率。这种方法以指标体系为研究对象，避免了逐个指标分析的繁冗性，但给出的解析式冗长、不直观，且在揭示指标的细节关系方面欠缺。

上面所采用的模型，利用大数据手段和统计分析方法，根据葡萄酒的理化指标对葡萄酒进行自动评级，并且针对影响葡萄酒的所有成分进行关联性计算，通过典型相关分析，得到不同指标之间的关联度和对最后口味的影响变化，总体来说本研究对于葡萄酒自动评级起到了一定的启发。

在大数据时代，我们可以考虑深度学习方案对大规模的葡萄酒样本构建多元深度神经网络模型，建立指标之间的关系，并建立更为高效准确的葡萄酒评级模型。

参考文献

[1]蒋露,薛洁,林奇,王异静.SNIF-NMR 和 IRMS 技术在葡萄酒质量评价中的初步研究[J].中国食品发酵工业研究院.
[2]徐海涛.改进的近似支持向量机在葡萄酒质量鉴定中的应用[J].安徽农业科学, 2010.
[3]刘延玲.新的 Hopfield 神经网络分类器在葡萄酒质量评价中的应用[J].天津大学系统工程研究所.
[4]李华,刘勇强,郭安鹊,梁新红,康文怀,陶永胜.运用多元统计分析确定葡萄酒感官特性的描述符[J].中国食品学报,2007,8.
[5]陈军辉,谢明勇,傅博强,杨妙峰,王小如.西洋参中无机元素的主成分分析和聚类分析[J].光谱学与光谱分析,2006,7.
[6]http://www.mathworks.com/products/neural −network/examples.html?file=/products/demos/shipping/nnet/classify_wine_demo.html
[7]中国葡萄酒国家标准 GB15037-2006
[8]李记明,李华. 食品与发酵工业[J].食品与发酵工业,1994.