2019 5th International Conference on Education Technology, Management and Humanities Science (ETMHS 2019)

Identification of red wine categories based on physicochemical properties

Xueting Bai[1], Lingbo Wang[1], Hanning Li[2]

[1]School of Statistics, Shanxi University of Finance and Economics, Taiyuan, China

[2]Faculty of International Trade, Shanxi University of Finance and Economics, Taiyuan, China

Keywords: Portrait of wine; Factor analysis; Cluster analysis; SPSS

Abstract: This paper mainly carried out the analysis of red wine categories. The red wine dataset is subjected to dimension reduction and clustering of samples through empirical analysis. First, factor analysis is performed on the 13 variables, and the complex variables are classified into five types of factors, namely the bitter trophic factor, the visual evaluation factor, the hue factor, the pH factor and the mineral element factor. Second, the samples were clustered by K-mean cluster analysis, and the samples were clustered into three different varieties. According to the cluster center, the characteristics of each variety can be summarized. Through a series of empirical analyses, a rough portrait of the red wine characteristics can be made and categories can be clustered in this data set.

1. Introduction

With the advancement of society and the continuous improvement of the life quality, the public demand for wine is gradually increasing. The wine culture is gradually becoming the civilization of all mankind. Thus people need more knowledge about wine to help them understand the wine civilization. The various physical properties of red wine (such as color, brightness and darkness) and the content of some of the ingredients can be used to quantitatively analyze the characteristics of different varieties of red wine. Different varieties have obvious differences in physical and chemical properties. Using the differences between the types reflected by these properties, even people who are unfamiliar with red wine can easily identify them. Therefore, this paper will use the various indicators of red wine to distinguish them and classify them.

2. Introduction to the method used by the model

2.1 Experimental data and variable interpretation

The data used in this experiment was obtained from the UCI database of the Italian wine data set, which contained a sample size of 178. The data contained in each variable is the result of a chemical analysis. The Italian wines shown in the sample are grown in the same area but from different varieties. The data set consists of a total of 13 numeric variables.

(1) Malic acid: It is a kind of acid with strong acidity and apple aroma. The red wine is naturally accompanied by malic acid. (2) Ash: The essence of ash is an inorganic salt, which has an effect on the overall flavor of the wine and can give the wine a fresh feeling. (3) Alkalinity of ash: It is a measure of weak alkalinity dissolved in water. (4) Magnesium: It is an essential element of the human body, which can promote energy metabolism and is weakly alkaline. (5) Total phenols: molecules containing polyphenolic substances, which have a bitter taste and affect the taste, color and taste of the wine, and belong to the nutrients in the wine. (6) Flavanoids: It is a beneficial antioxidant for the heart and anti-aging, rich in aroma and bitter. (7) Nonflavanoid phenols: It is a special aromatic gas with oxidation resistance and is weakly acidic. (8) Proanthocyanins: It is a bioflavonoid compound, which is also a natural antioxidant with a slight bitter smell. (9) Color intensity: refers to the degree of color shade. It is used to measure the style of wine to be “light” or “thick”. The color intensity is high, meanwhile the longer the wine and grape juice are in contact during the wine making process, the thicker the taste. (10) Hue: refers to the vividness of the color and the degree of warmth and coldness. It can be used to measure the variety and age of the wine.

DOI: 10.25236/etmhs.2019.309

Red wines with higher ages will have a yellow hue and increased transparency. Color intensity and hue are important indicators for evaluating the quality of a wine's appearance. (11) Proline: It is the main amino acid in red wine and an important part of the nutrition and flavor of wine. (12) OD280/OD315 of diluted wines: This is a method for determining the protein concentration, which can determine the protein content of various wines.

2.2 Model description applied to the data set

Factor and cluster analysis are applied to the known data set to determine the category to which the wine sample belongs. First, factor analysis will be performed on the 13 variables. The common factors will be extracted and the factor load matrix will be rotated, and the characteristics of the common factor will be summarized and finally the factor score formula and the comprehensive score of the sample on the common factor will be obtained. Then it will be saved as a new variable and K-mean cluster analysis will be performed on the observations, and the category characteristics of each class will be summarized according to the final cluster center on each factor.

3. Empirical analysis

3.1 Dimensionality analysis of physical and chemical variables

Tab.1 KMO and Bartlett’s Test

Sampling a sufficient degree of Kaiser-Meyer-Olkin	Measure	0.779
Bartlett's sphericity test	Approximate chi square	1317.181
	df	78
	Sig.	0

KMO test is performed on the data set from UCI, and the result is shown in Tab.1. The KMO value of the test is 0.779, which is suitable for factor analysis. When using SPSS for factor analysis, if the factor whose feature root is greater than 1, it can be retained by default. As a result, the system extracts 3 factors, and the cumulative contribution of variance is 66.53%. In order to extract more information from the variables, the model is artificially set, which means five factors are extracted, and a total variance of 80% is explained after this adjustment. The obtained composition matrix is shown in Tab.2.

Tab.2 Unrotated Load Matrix

Component Matrixa					
	1	2	3	4	5
Alcohol	.313	.764	-.249	-.017	.245
Malic acid	-.532	.355	.107	.515	-.033
Ash	-.004	.499	.753	-.205	.132
Alkalinity of ash	-.519	-.017	.736	.058	-.061
Magnesium	.308	.473	.157	-.337	-.672
Total phenols	.856	.103	.176	.190	.138
Flavanoids	.917	-.005	.181	.146	.101
Nonflavanoid phenols	-.648	.045	.205	-.195	.463
Proanthocyanins	.680	.062	.180	.383	-.126
Color intensity	-.192	.837	-.165	.063	.071
Hue	.644	-.441	.102	-.410	.160
OD280/OD315 of diluted wines	.816	-.260	.200	.177	.093
Proline	.622	.577	-.152	-.222	.146

Observing the matrix before the factor rotation is performed, there is no significant difference in the load of some common factors on the original variables. At this point, it is necessary to perform a factor rotation on the load matrix.

Tab.3 Rotating Load Matrix

Rotated Component Matrixa					
	Component				
	1	2	3	4	5
Alcohol	.148	.876	-.109	-.011	.058
Malic acid	-.146	.006	-.789	.173	-.112
Ash	.084	.243	-.032	.886	.155
Alkalinity of ash	-.182	-.424	-.294	.721	-.008
Magnesium	.098	.229	.053	.148	.907
Total phenols	.836	.268	.240	.031	.026
Flavanoids	.868	.189	.339	-.003	.051
Nonflavanoid phenols	-.565	.004	-.075	.444	-.439
Proanthocyanins	.795	.055	-.026	-.042	.155
Color intensity	-.198	.696	-.472	.118	.122
Hue	.352	-.087	.825	-.022	-.021
OD280/OD315 of diluted wines	.816	-.058	.374	-.048	-.046
Proline	.324	.775	.245	-.007	.217

The factor rotation is performed on the load matrix, and the result is shown in the load matrix after rotation in Tab.3. It is considered that the load is larger when the factor load is greater than 65%. Availability from Tab.3 is concluded below.

The first common factor carries a large load on the variables of total phenols, flavonoids, proanthocyanins and OD280/OD315 of diluted wines. It is known that chemical substances such as phenolic compounds, flavonoids, and proanthocyanins have a great influence on mouthfeel, and have a feeling of sputum and bitterness. At the same time, these chemicals have strong antioxidant properties and are, in some ways, the chemicals required by the human body. The higher absorbance ratio of OD280/OD315 indicates high protein purity. Therefore the first common factor can be named the bitter trophic factor of wine.

The second common factor has a large load on the variables of alcohol, color intensity and proline. Proline is known to be an amino acid that regulates the flavor of the wine. The color intensity refers to the degree of lightness of the color, and the greater the intensity, the darker the color. It reflects the nature of the grapes that make the wine. Therefore the second common factor can be named as the visual evaluation factor of wine.

The third common factor has a larger load on the two variables of malic acid and hue. Malic acid is known to be a natural acid that balances the sweetness of wine. Malic acid is commonly used in the production of wines for lactic acid fermentation (MLF), in which lactic acid bacteria convert the more acidic malic acid into less acidic lactic acid. In the process of MLF, the total acid decreases and the pH rises, which causes the color of the grape to change from purple to blue, thus changing the color tone of the wine. The hue refers to the vividness and warmth of the color of the wine. So the third common factor is the hue factor of wine.

The fourth common factor is heavily loaded on the two variables of ash and alkalinity of ash. It is known that ash in wine is an effective substance for neutralizing acidity, and is essentially an inorganic salt. Therefore the fourth common factor can be named the pH factor of wine.

The fifth common factor has a large load on the concentration variable of magnesium. Due to the loss of the original data set, the original mineral element variables have lost, so the magnesium element can temporarily represent the mineral element. The fifth common factor can be named as the mineral element factor of wine.

Tab.4 Factor Score Matrix

Component Score Coefficient Matrix					
	1	2	3	4	5
Alcohol	-.014	.431	-.021	-.017	-.157
Malic acid	.214	-.055	-.541	.002	-.109
Ash	.023	.089	.114	.602	.033
Alkalinity of ash	.086	-.239	-.114	.446	.061
Magnesium	-.118	-.086	.057	.049	.883
Total phenols	.287	.063	-.054	.081	-.149
Flavanoids	.276	.023	-.003	.071	-.107
Nonflavanoid phenols	-.179	.170	.203	.313	-.422
Proanthocyanins	.365	-.113	-.288	-.022	.048
Color intensity	-.041	.316	-.193	.000	-.010
Hue	-.094	.020	.512	.122	-.052
OD280/OD315 of diluted wines	.282	-.079	.003	.050	-.143
Proline	-.056	.363	.183	.032	.009

Tab.4 is a matrix of factor score coefficients. This gives the equations of the factor score, where X_i(i=1,2,3,…,13) is the value normalized by the original data.

$$Y_1 = -0.014X_1 + 0.214X_2 + 0.023X_3 + 0.086X_4 - 0.118X_5 + 0.287X_6 + 0.276X_7$$
$$-0.179X_8 + 0.365X_9 - 0.041X_{10} - 0.094X_{11} + 0.282X_{12} - 0.056X_{13}$$
$$Y_2 = 0.431X_1 - 0.055X_2 + 0.089X_3 - 0.239X_4 - 0.086X_5 + 0.063X_6 + 0.023X_7$$
$$+0.170X_8 - 0.113X_9 + 0.316X_{10} + 0.020X_{11} - 0.079X_{12} + 0.363X_{13}$$
$$Y_3 = -0.021X_1 - 0.541X_2 + 0.114X_3 - 0.114X_4 + 0.057X_5 - 0.054X_6 - 0.003X_7$$
$$+0.203X_8 - 0.288X_9 - 0.193X_{10} + 0.512X_{11} + 0.003X_{12} + 0.183X_{13}$$
$$Y_4 = -0.017X_1 + 0.002X_2 + 0.602X_3 + 0.446X_4 - 0.049X_5 + 0.081X_6 + 0.071X_7$$
$$-0.313X_8 - 0.022X_9 + 0.000X_{10} + 0.122X_{11} + 0.050X_{12} + 0.032X_{13}$$
$$Y_5 = -0.157X_1 - 0.109X_2 + 0.033X_3 + 0.061X_4 + 0.883X_5 - 0.149X_6 - 0.107X_7$$
$$-0.422X_8 + 0.048X_9 - 0.010X_{10} - 0.052X_{11} - 0.143X_{12} + 0.009X_{13}$$

$$\textit{Factor synthesis score of every Observations}$$
$$= (\frac{0.26387}{0.80162}Z_1 + \frac{0.17424}{0.80162}Z_2 + \frac{0.15435}{0.80162}Z_3 + \frac{0.12096}{0.80162}Z_4 + \frac{0.08820}{0.80162}Z_5)$$

$Z_i\left(i = 1, 2, 3, 4, 5\right)$ is the score of 178 data based on the 5 common factor.

3.2 Sample classification based on physical and chemical variables

Q-type cluster analysis was performed on the samples. The five factors obtained by factor analysis were the bitter trophic factor, the visual evaluation factor, the hue factor, the pH factor and the mineral element factor, which were saved as variables, and K-mean cluster analysis was performed on the samples according to these five factors. At this point the data has been standardized and there is no difference between the dimensions. Decide the number of clusters to be 3 categories.

Tab.5 Cluster Significance Test

	Cluster		Error			
	Mean	df	Mean	df	F	Sig.
the bitter	24.602	2	0.73	175	33.69	0
the visual	14.288	2	0.848	175	16.846	0
the hue factor	17.358	2	0.813	175	21.349	0
the pH factor	2.375	2	0.984	175	2.413	0.093
the mineral	49.665	2	0.444	175	111.9	0

One-way ANOVA was performed on the differences between the three categories of factors involved in the classification (the more significant the variance analysis has more influence on the clustering results), the test results are shown in Tab. Only the P value of the pH factor is greater than 0.05, indicating that the clustering results are generally effective.

Tab.6 Number of Clustering Cases

Number of cases in each cluster		
cluster	1	40
	2	63
	3	75
valid		178
missing		0

As shown in Tab.6, the table shows the number of cases in each cluster. It can be found that the proportions are roughly equal, and there is no large difference between the number of cases included in the class. Therefore, the clustering can be considered to be good.

Tab.7 Final Cluster Center

	Cluster		
	1	2	3
the bitter trophic factor	-0.14302	-0.60303	0.58283
the visual evaluation	0.13214	-0.52691	0.37213
the hue factor	-0.18645	0.58697	-0.39362
the pH factor	0.3028	-0.10057	-0.07701
the mineral element factor	1.38697	-0.42797	-0.38022

The final cluster center is obtained from Tab.7, and the following results can be obtained.

The first category: the bitter trophic factor score is slightly lower than the average value of 0. It can be considered that its nutrient composition is slightly lower, and the taste is sweeter. The nutritional value of this type of wine is low; the visual evaluation factor score is slightly larger than the average value. It can be considered that the variable score used for visual evaluation are moderate, while the degree of wine is moderate, the color is deep, and the flavor is easier to distinguish; the hue factor score is slightly lower than the mean. It indicates that the wine has low color saturation; the high pH factor indicates that the wine is more alkaline; the mineral element factor content is higher, indicating that the magnesium content in the wine is higher.

The second category: bitter taste nutrient content is lower and lower than the first category of wine, and the taste is sweeter. The nutritional value of this type is very low; and the visual evaluation factor has the lowest score among the three categories, with the low degree, the light color, and the Indistinguishable flavor of wine; the color saturation is the highest, so the color is bright; and the alkalinity of the wine is the lowest; the substance has the lowest magnesium content.

The third category: the bitter trophic factor has the highest score, that is, the bitter taste nutrient content is the highest. The nutritional value of this kind of wine is very high; the visual evaluation factor has the highest score. The flavor is easy to distinguish, while the color is deep and the degree of wine is higher; the liquid color saturation is the lowest; the alkalinity of wine is lower; the mineral magnesium content is lower than the average level of the three categories.

4. Conclusion

The model performs factor analysis on the variables to obtain the five types of factors, namely the bitter trophic factor, the visual evaluation factor, the hue factor, the pH factor and the mineral element factor. K-mean cluster analysis is performed on the samples using these five factors, and the samples were classified into three types. The first type of red wine is a kind of wine with a sweet taste and low nutritional value, among which the mineral element content is high and it is an alkaline wine. The second type of wine has the lowest nutritional value and belongs to sweet wine with bright colors. The third type tastes bitter, with the highest nutritional value and the deepest color.

References

[1] Generally accepted rating principles: A primer [J] .Jan Pieter Krahnen, Martin Weber. Journal of Banking and Finance .2001(1)

[2] Huang H J, Qian C B, Feng Fan, Zhou.X.Z. Based on the improved K-means algorithm, the grading method of red wine according to the physical and chemical indicators of wine grapes and wine is studied.[J].Chinese market,2017(16):196-197.

[3] Ye S P, Chen H G, Liang K H. Mathematical model for quality evaluation of wine physical and chemical indicators.[J].Anhui Agricultural Sciences, 2015, 43(12): 214-216.

[4] Liu Chan, Jiang Wei. Wine grading based on wine physical and chemical indicators [J]. Henan Science and Technology, 2014(16):30-31.

[5] Ma Jian, Yuan J H. Research on wine quality evaluation based on physical and chemical indicators [J]. Food industry technology, 2013, 34(18):137-140+143.

[6] Dong Ying, Cui R X. Quality evaluation of red wine based on factor analysis [J]. Journal of Dalian Nationalities University, 2014, 16(03):284-288.

[7] Cheng Z R, Jiu D K, Wang Y W. Graded evaluation model for wine quality [J]. China High-tech Zone, 2018(03):58.

[8] Xu S S, Tan Bing, Li Qian, He Ting. Wine quality evaluation [J]. Modern trade industry, 2018, 39(09):54-56.

[9] Yi Liu, Jiawen Peng, and Zhihao Yu. 2018. Big Data Platform Architecture under The Background of Financial Technology: In The Insurance Industry As An Example. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018). ACM, New York, NY, USA, 31-35.