Advances in Food Science and Human Nutrition (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/afshn.2023.050107
ISSN 2616-258X Vol. 5 Num. 1

Classification Technology of GC-MS Map Data of Baijiu Based on Sparse Principal Components

Zhiwen Yang*

Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin, 644000, China
**Corresponding author*

Keywords: Baijiu recognition; sparse principal component analysis; GC-MS map; elastic net penalty

Abstract: In order to achieve accurate identification of GC-MS Baijiu mapping data, the sparse principal component analysis (SPCA) of GC-MS Baijiu mapping data is achieved by introducing the elastic net penalty function and ridge regression to restrict the sparse principal components on the basis of the principal component analysis method. The sparse principal components are fed into different classifiers for classification and identification, and a Baijiu quality classification model is established. Through comparison experiments, it was demonstrated that sparse principal components better represented the information of different characteristics of liquor, and the classification recognition accuracy after classification was higher, and the recognition rates of SPCA+KNN, SPCA+DT, SPCA+SVM, and SPCA+BP reached 62%, 89%, 97% and 100%; the differences of sparse principal components of GC-MS profiles of different grades of liquor were greater than the differences of principal components, and the sparse principal components of GC-MS profiles of liquor was a nonlinear relationship. The established sparse principal component-based Baijiu quality evaluation model can effectively realize the evaluation of Baijiu grades, which provides a more effective and objective method for the control of Baijiu quality and grade identification.

1. Introduction

The acid, alcohol, ester, aldehyde, ketone and other trace components in Baijiu account for about 2% of the main components, but have a greater impact on the flavor of Baijiu. For a long time, the evaluation of Baijiu characteristics is mainly by sensory evaluation, which will inevitably be affected by human factors. Therefore, if scientific methods are used to improve the objectivity and accuracy of Baijiu characteristics evaluation, it is an urgent problem to be solved [1].

At present, the evaluation methods of Baijiu characteristics mainly obtain Baijiu map data through sensor detection technology, and use feature extraction and recognition technology to achieve the classification of Baijiu characteristics. Common methods include: Electronic nose mass spectrometry (EN-MS) [2, 3], Gas chromatography mass spectrometry (GC-MS) [4, 5], inductively coupled plasma (ICP), Fluorescence spectroscopy technology[6, 7] and techniques such as colorimetric artificial nose method. Among them, GC-MS combines the separation ability of chromatography and the qualitative ability of mass spectrometry, which can conduct qualitative analysis of multi-component mixtures in a relatively short time, thus effectively reflecting the trace components. It is the most common method in the evaluation of Baijiu characteristics. Li Xinfeng et al. [8] used gas chromatography technology to evaluate the quality grade of Luzhou flavor Baijiu, find the micro relationship between Baijiu ingredients, and provide new ideas for the micro research of Baijiu; Zhang Qi et al. [9] used GC-MS technology to obtain 32 characteristic peaks contained in several types of Luzhou flavor Baijiu. The content and proportion of flavor components represented by the characteristic peaks of various Baijiu components significantly affected the flavor and flavor of Baijiu; Xun Siying et al. [10] established a common peak map model for Baijiu by combining GC-MS determination technology with map analysis software.

The principal component analysis (PCA) is mainly used for dimension reduction of Baijiu map data based on GC-MS to achieve the main feature extraction of map data [11]. PCA uses several orthogonal principal components to represent the complete information of Baijiu map data, and achieves the purpose of data dimension reduction on the basis of maximum retention of data information [12, 13]. In recent years, with the improvement of sparse representation theory, sparse representation technology is gradually applied to the recognition of Baijiu atlas [14]. To better realize the sparse decomposition of data and reflect the weight of each component, this paper uses sparse principal component analysis (SPCA) to analyze each component of Baijiu. SPCA uses the method of aggregated data sparsity to distinguish principal components, and the unit feature vector corresponding to each principal component is filled with as many zeros as possible, so that fewer linear combinations of variables can be used to represent the original data, so as to better achieve data dimensionality reduction [15].

In this study, SPCA was applied to reduce the dimensions of Baijiu GC-MS map data, making the map data more sparse in the principal component space. The main information of GC-MS map data was screened through the cumulative contribution rate, and then K nearest neighbor (KNN), decision tree (DT), support vector machine (SVM) Various classification algorithms such as error back propagation (BP) classify the characteristic grades of Baijiu, and compare each algorithm with PCA classification algorithm.

2. Sparse Principal Component Analysis Method

Sparse principal component analysis method is an improved data statistical analysis method that adds sparsity conditions on the basis of principal component analysis method. It uses the most representative linear combination method of a few variables to represent the original data, and converts it into a regression problem with quadratic penalty, which can better simplify the load data and reduce the dimensions of the data. For the problem of solving sparse principal components, Lasso regression can be used to transform the problem into a variable selection problem, and an elastic network penalty function with a linear combination of ridge regression and Lasso penalty can be used to obtain sparse principal components.

The steps of the SPCA algorithm are as follows:

(1)The feature corresponding to the first *m* principal components of PCA is $\alpha_{\mathrm{j}}(j=1,2,3\cdots m)$.

(2)Solving the regression problem for a given $A_m=(\alpha_1,\alpha_2,\cdots\alpha_m)$:

$$\beta_j=\arg\min_{\beta}(\alpha_j-\beta)^T X^T X(\alpha_j-\beta)+\lambda\|\beta\|^2+\lambda_{1,j}\|\beta\|_1 \quad (j=1,2,\cdots m)$$

get $B_m=(\beta_1,\beta_2,\cdots\beta_m)$.

(3)For a given B_m ,perform Singular Value decomposition: $\left(X^T X\right)B_m = UDV^T$, $\overset{*}{A}_m = UV^T$.

(4)Repeat the above processes (2) and (3) until convergence occurs.

(5)Standardization $V_j = \frac{\beta_j}{\left\|\beta_j\right\|}$ $\left(j = 1,2,\cdots m\right)$.

Calculation of SPCA variance contribution rate:

Assuming $\hat{Z}$ as the extracted principal component, the total variance explained by $\hat{Z}$ can be calculated using $tr\left(\hat{Z}^{\mathrm{T}}\hat{Z}\right)$.

Let $H_{1\ldots j-1}$ be the projection on $\left(\hat{Z}_i\right)_1^{j-1}$, and $\hat{Z}_{j\cdot1,\cdots,j-1}$ be the adjusted residual:

$$\hat{Z}_{j\cdot1,\cdots j-1} = \hat{Z}_j - H_{1,\cdots,j-1}\hat{Z}_j$$

Therefore, the adjusted variance of $\hat{Z}_j$ is $\left\|\hat{Z}_{j\cdot1,\cdots j-1}\right\|^2$, and the total explanatory variance is $\sum_{j=1}^{k}\left\|\hat{Z}_{j\cdot1,\cdots,j-1}\right\|^2$.

This article adopts a sparse principal component algorithm based on elastic net penalty structure, which uses Lasso penalty with elastic net to continuously modify the regression optimization framework and obtain sparse principal components. Compared with PCA, SPCA has added sparsity conditions on the basis of PCA, which enhances the interpretation ability of PCA by sparsizing the data information payload through variance contribution rate; SPCA combines sparsity conditions to comprehensively analyze the principal component variance and load of PCA partitioning, improving the data processing ability of PCA. Finally, the data principal components generated by SPCA decomposition are not correlated. Therefore, using SPCA can better interpret and statistically analyze white wine data.

3. Experimental Methods

The sample data of Baijiu used in the experiment were selected from various distilleries in southern Sichuan, with 20 samples, which were numbered as T1, T2, T3, T4, T5, Y1, Y2, Y3, Y4, Y5, R1, R2, R3, R4, R5, S1, S2, S3, S4, S5, respectively. The GC-6800 gas chromatography-mass spectrometer produced by Jiangsu Tianrui Instrument Co., Ltd. is selected as the experimental instrument, and the DB-WAXMS chromatographic column (30 m) produced by Agilent Technologies Co., Ltd. is assembled × 0.25 mm × 0.25 μ m). The sensory evaluation of Baijiu is conducted with reference to the national standard GB/T10345-2007 Analytical Methods for Baijiu, and the sensory evaluation personnel are composed of wine tasters in southern Sichuan. The number order dark evaluation method is adopted, and the wine tasters comprehensively evaluate Baijiu, with a total score of 100 points. The average of the scores of each wine taster is the score of the Baijiu sample. Among them, scores ranging from 93.0 to 100.0 are classified as special grade, 88.0 to 92.9 are classified as first grade, 80.0 to 87.9 are classified as second grade, 70.0 to 79.9 are classified as excellent grade, and scores below 70.0 are classified as others. The sensory evaluation of the 20 samples is shown in the table 1.

Table 1: Results of sensory evaluation of Baijiu samples

ID	T1	T2	T3	T4	T5	Y1	Y2	Y3	Y4	Y5	R1	R2	R3	R4	R5	S1	S2	S3	S4	S5
Sensory score	99	98	95	94	97	89	92	91	90	91	87	83	85	87	80	76	72	78	74	71

Qualitative method of Baijiu flavor: measure 5 mL of Baijiu sample, add 100% mixed internal standard solution μ L (Amyl acetate 15.10 g/L, tert amyl alcohol 15.19 g/L, 2-ethylbutyric acid 15.09 g/L), mixed evenly and detected by GC-MS, in which the gas chromatographic condition: injection volume 1 μ L; Split ratio 20:1; Injection port temperature 250 ℃; Heating program: The initial temperature is 35 ℃ for 10 minutes, then it is raised to 120 ℃ at 2 ℃/min, then to 200 ℃ at 5 ℃/min, and finally to 245 ℃ at 10 ℃/min for 40 minutes; The carrier gas is high-purity helium (He), with a flow rate of 1 mL/min. Mass spectrometry conditions: GC-MS interface temperature 280 ℃; Ion source temperature 230 ℃; Scanning quality range 29-500 m/z; Electron impact ion source (EI); Ionization energy 70 eV. Finally, after SPCA dimensionality reduction of the obtained sample data, Baijiu samples are classified using multiple classification algorithms such as KNN, DT, SVM and BP.

4. Result and Analysis

The volatile flavor components of 20 Baijiu samples were analyzed by GC-MS, and the volatile flavor components of a Baijiu sample were randomly selected for analysis. See Table 2 for the GC-MS analysis results of volatile flavor components of a Baijiu sample.

Table 2: 46 kinds of trace components and their contents in Baijiu based liquor samples

Serial Number	chemical compound	content /(mg·L^{-1})	Serial Number	chemical compound	content /(mg·L^{-1})
1	ethyl acetate	652.19	24	butyral	69.08
2	Ethyl propionate	18.35	25	2,3-Butanediol	3.21
3	Propyl acetate	229.31	26	Ethyl Phenylpropanoic acid	31.64
4	3-Methylbutanol	25.85	27	Acetal	107.43
5	Ethyl butyrate	237.08	28	Ethyl benzene	50.63
6	Isopropyl alcohol	81.47	29	Ethyl heptanoate	173.62
7	2-Methylbutanol	44.54	30	Ethyl octanoate	167.83
8	Ethyl pentanoate	34.64	31	Phenylethyl diethyl acetal	54.53
9	N-butanol	123.24	32	Hexyl Hexanoate	131.82
10	2-butanone	11.03	33	propionic acid	0.00
11	methanol	136.62	34	Isobutyric acid	3.57
12	lactic acid	0.00	35	butyrate	112.59
13	2-Pentanone	107.84	36	Isovaleric acid	6.18
14	Ethyl Linoleic acid	203.92	37	Pentanoic acid	13.80
15	Amyl acetal	175.53	38	Caproic acid	623.92
16	P-Xylene	73.83	39	Phenethyl alcohol	1.29
17	Isoamyl alcohol	117.97	40	Heptanoic acid	5.02
18	Ethyl hexanoate	986.26	41	Octanoic acid	6.33
19	N-pentanol	19.91	42	Ethyl caproate	73.75
20	Ethyl lactate	534.19	43	Ethyl tetradecanoate	86.28
21	N-hexanol	63.80	44	Ethyl octadecanoate	557.36
22	Isopentyl hexanoate	0.00	45	Ethyl Isobutyric acid	203.81
23	Caproic acid	430.47	46	acetaldehyde	165.04

According to Table 2, 46 volatile flavor components were detected in this Baijiu sample. PCA and SPCA were used to extract features from 46 trace components in Table 2, removing the principal components with significant data deviation and irrelevant variables, while retaining the 10 principal components with significant feature values. The results are shown in Tables 3 and 4.

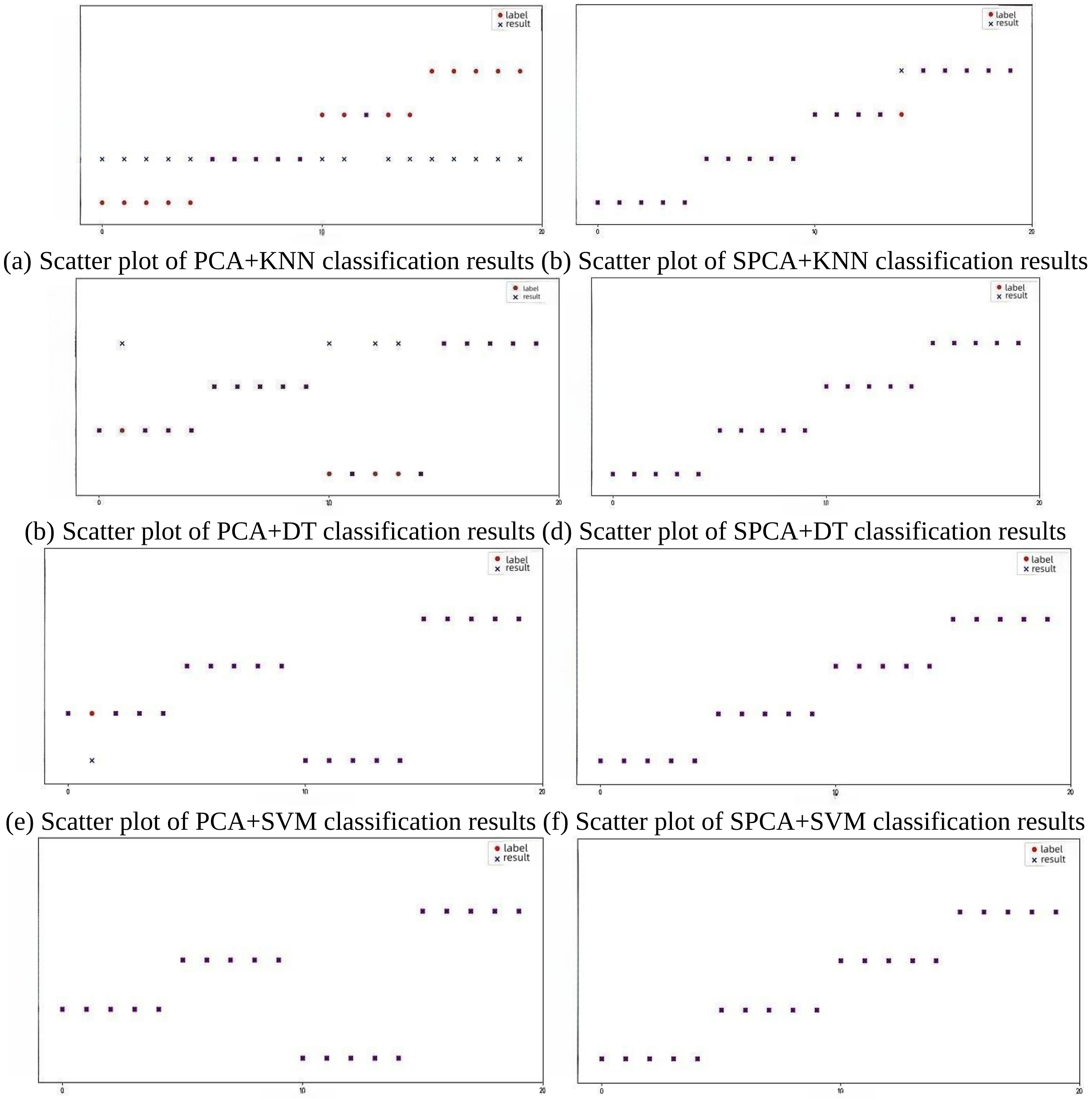
Table 3: The eigenvalues, contribution rates and cumulative variances of the 10 principal components

component	Principal Component Eigenvalues		Variance contribution rate/%	Cumulative variance contribution rate/%
1	18.960		41.218	41.218
2	7.522		16.353	57.571
3	4.714		10.249	67.820
4	3.378		7.344	75.164
5	3.022		6.570	81.734
6	2.613		5.681	87.415
7	1.383		3.006	90.421
8	1.100		2.392	92.813
9	0.905		1.968	94.781
10	0.788		1.714	96.495

Table 4: The eigenvalues, contribution rates and cumulative variances of the 10 sparse principal components

component	Principal Component Eigenvalues	Variance contribution rate/%	Cumulative variance contribution rate/%
1	12.204	40.679	40.679
2	5.829	19.432	60.111
3	3.142	10.473	70.584
4	2.206	7.352	77.936
5	2.103	7.010	84.946
6	1.515	5.048	89.994
7	1.076	3.586	93.580
8	0.726	2.420	96.000
9	0.344	1.147	97.147
10	0.268	0.893	98.040

From Tables 3 and 4, it can be seen that the cumulative variance contribution rates of the first 8 principal components and the first 7 sparse principal components extracted using PCA and SPCA methods are 92.813% and 93.580%, respectively. Therefore, in subsequent experiments, this article can select the first 8 principal components and the first 7 sparse principal components for analysis.



(a) Scatter plot of PCA+KNN classification results (b) Scatter plot of SPCA+KNN classification results

(b) Scatter plot of PCA+DT classification results (d) Scatter plot of SPCA+DT classification results

(e) Scatter plot of PCA+SVM classification results (f) Scatter plot of SPCA+SVM classification results

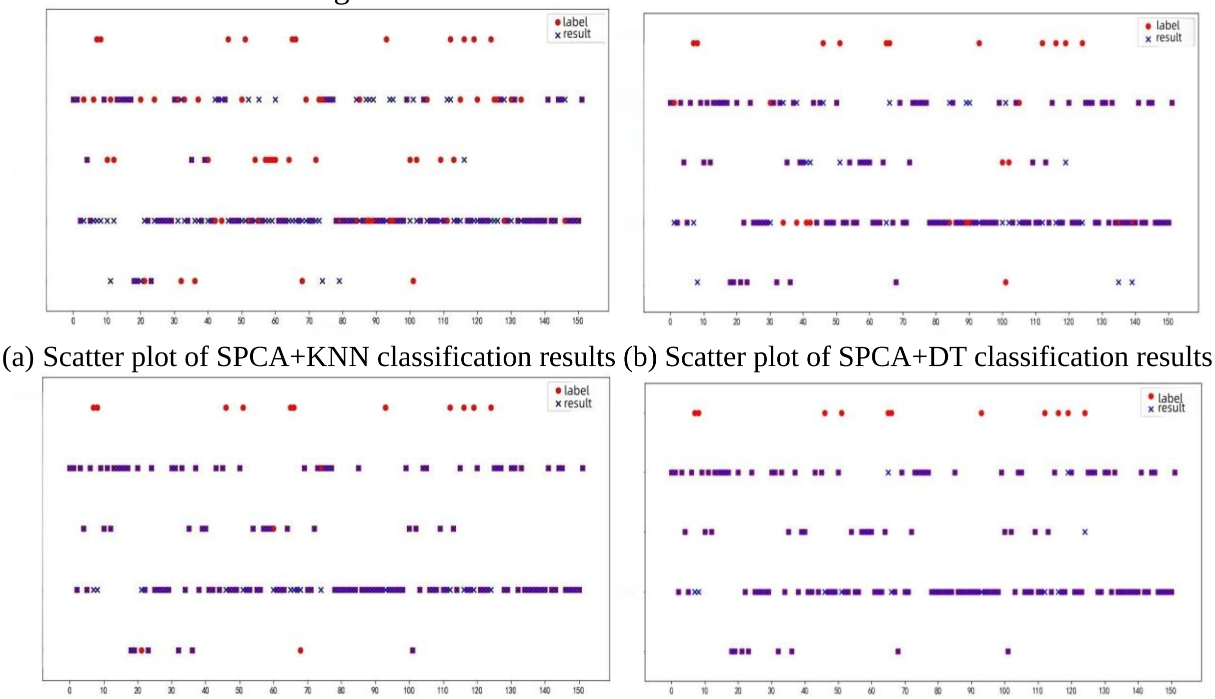
(g)Scatter plot of PCA+BP classification results (h) Scatter plot of SPCA+BP classification results

Figure 1: Classification results of PCA and SPCA combined with four classification methods

After principal component analysis (PCA) and sparse principal component analysis (SPCA) feature extraction, 20 Baijiu samples with known grades were classified according to the common classification methods of KNN, DT, SVM and BP, and compared with the grade of Baijiu evaluated by artificial Sensory analysis. From the collected Baijiu data, 170 groups of Baijiu data with different grades were selected as the training set, and another 20 groups (5 groups for each grade) of Baijiu with different grades were randomly selected as the test set. PCA and SPCA were used to process the Baijiu test set, and the extracted principal component characteristics were used as input data to establish a standard grade classification model to identify the basic attributes of the test set data. The first 8 principal components of Baijiu and the first 7 sparse principal components of Baijiu are selected as the input data of KNN, DT, SVM and BP methods, and the classification results are shown in Figure 1.

From Figure 1, it can be seen that after testing on 20 test samples, the number of false positives using PCA combined with KNN, DT, and SVM methods was 14, 4, and 1, respectively, with classification accuracy of 30%, 80%, and 95%. However, the number of false positives using SPCA combined with KNN, DT, and SVM methods was 1, 0, and 0, respectively, with classification accuracy of 95%, 100%, and 100%, respectively. It can be seen that SPCA method can more accurately extract principal components and has stronger feature extraction ability. The methods of PCA, SPCA, and BP all have no false samples, and the classification accuracy is 100%. It can be seen that the nonlinear BP method can effectively improve the recognition accuracy. The results indicate that PCA and SPCA can achieve dimensionality reduction and information interpretation in data processing to varying degrees, and SPCA has strong advantages over PCA in data dimensionality reduction, redundancy removal, preservation of original data information, and actual interpretation ability.

Using the model obtained from pre training, 150 samples of Baijiu with known grades were randomly selected from the collected Baijiu data outside the training set, and GC-MS atlas test was conducted. The test results are extracted by SPCA, and the sparse principal components extracted are used as input data, which are classified by KNN, DT, SVM and BP. The classification results of each method are shown in Figure 2.



(a) Scatter plot of SPCA+KNN classification results (b) Scatter plot of SPCA+DT classification results

(b) Scatter plot of SPCA+SVM classification results (d) Scatter plot of SPCA+BP classification results

Figure 2: Classification results of four classification methods

It can be seen from Figure 2 that 11 of the 150 Baijiu samples taken are not super Baijiu, first grade Baijiu, second grade Baijiu, and superior Baijiu; The number of false positives for SPCA combined with KNN, DT, SVM, and BP methods was 57, 17, 5, and 0, respectively, with classification accuracy rates of 62%, 89%, 97%, and 100%. The classification verification shows that the SPCA method can effectively extract the main characteristics and information of different grades of Baijiu, and combined with the nonlinear BP classification method, the accuracy of SPCA combined with BP method for different grades of Baijiu can reach 100%.

5. Conclusion

In this study, GC-MS was used to analyze the volatile flavor components of four Baijiu samples of different grades by PCA and SPCA. The results showed that, compared with the principal components, the sparse principal components of GC-MS maps of different grades of Baijiu were significantly different, which could better represent the information of Baijiu with different characteristics. Subsequently, Baijiu samples were classified based on PCA and SPCA, respectively in combination with KNN, DT, SVM and BP. The results showed that the classification accuracy of Baijiu samples based on SPCA was higher, and the accuracy of SPCA+KNN, SPCA+DT, SPCA+SVM, and SPCA+BP reached 62%, 89%, 97%, and 100%, respectively. The classification effect of SPCA combined with BP was the best, indicating that the sparse principal component coefficients of Baijiu GC-MS were non-linear. The experimental data were verified with the sample data, and the classification results and data information were completely consistent, indicating that the established Baijiu quality evaluation model based on sparse principal component analysis could effectively realize the evaluation of Baijiu grades, providing a more effective objective method for Baijiu quality control and grade identification.

Acknowledgments

This work was supported by the The Innovation Fund of Postgraduate, Sichuan University of Science & Engineering under Grant (Y2021079).

References

[1] Jian Li, Jiang Xue. Study on the identification method of Luzhou flavor pure grain Baijiu. Chinese Brewing, 2015, 34 (1): 118-121.

[2] Xiuli Chen, Hairong Gao, Zhenxu Huang, et al. Application of Electronic nose analysis method in Baijiu classification and recognition. Journal of Xinyang Normal University (Natural Science Edition), 2014,27 (3): 386-389393

[3] Ting Tian, Shuyi Qiu, Lingji Wen, et al. Differentiation and identification of different rounds of Maotai flavor Baijiu by Electronic nose technology. China brewing, 2017, 36(10):71-75.

[4] Xiaolei Zhang, Jing Rao, Chunyang Li. Application of gas chromatography and gas chromatography-mass spectrometry in quality control of Baijiu. Brewing Technology, 2016, 43(2):16-23.

[5] Zhancheng Xu, Yong Chen, Shuang Wang. Study on the Flavor Quality Characteristics of Chinese Famous Wine Jiannanchun by Using Two Dimensional Gas Chromatography and Adsorption Stirring Extraction Technology. Brewing Technology, 2012, 39(5):6-8.

[6] Ruiyu Xu, Zhuowei Zhu, Yangjun Hu, et al. Identification of several Luzhou flavor Baijiu by three-dimensional fluorescence spectroscopy combined with PCA-SVM. Spectroscopy and Spectral Analysis, 2016, 36(4): 1021-1026.

[7] Jianlei Yang, Tuo Zhu, Yan Xu, et al. Application of 3D fluorescence spectroscopy based on least squares support vector machine algorithm in the classification of Chinese Baijiu. Spectroscopy and Spectral Analysis, 2010, 30(1):243-246.

[8] Xinfeng Li, Liang Zhang, Fangfang Li, et al. Evaluation of Luzhou flavor Baijiu quality grade based on GC-QTOF MS technology. Science and Technology of Food Industry, 2019, 40(15):235-241.

[9] Qi Zhang, Yong Xu, CaiHong Shen, et al. Research progress in analysis of volatile flavor components of Luzhou flavor Baijiu. Brewing Technology, 2017(12):98-104.

[10] Siying Xun, Rui Dong, Qinrong Peng, et al. Determination of Volatile Phenols in Maotai flavor Baijiu by High-performance liquid chromatography. Food Science, 2012, 33(24):239-243.

[11] Altinel B, Ganiz M C. A new hybrid semi-supervised algorithm for text classification with class based semantics. Knowledge Based Systems, 2016, 108(C): 50-52.

[12] Xu Hu, Jinsong Li, Yongqing Tang, et al. Classification of Luzhou flavor Baijiu based on GC-MS and Chemometrics. Food and fermentation industry, 2021, 47(8):212-217.

[13] Liu Fang, Kangzhuo Yang, Jianmin Zhang, et al. Classification of Luzhou flavor Baijiu based on Electronic nose and GC-MS. Food and fermentation industry, 2020,46(2):73-78.

[14] Haiyan Wang, Hu Wang, Guoxiang Wang, et al. Classification of Baijiu flavor types based on Compressed sensing. Computer program, 2015, 41(3):172-176,181.

[15] Zhengping Hu, Junling Chen. Multi layer fusion deep local PCA subspace sparse optimization feature extraction model. Journal of Electronics, 2017, 45(10):2383-2389.