Chinese Baijiu Types Cluster Analysis using K-Means and PCA

SiyuanZhou
1024041143
Nanjing University of Posts and Telecommunications
School of Computer Science
Nanjing, China

***Abstract*—Whether by choice or obligation, most people have tried Baijiu at some point in their lives. However, regardless of the price or quality, many describe the experience with one dominant impression: a strong, spicy taste. This perception oversimplifies the complexity of Baijiu, which actually encompasses a wide range of flavors. Baijiu, one of China's most iconic spirits, is traditionally categorized into three primary flavor profiles: light aroma (Qing), strong aroma (Nong), and sauce aroma (Jiang). These broader categories are further subdivided into more specific types, including sauce aroma, soy aroma, medicinal aroma, phoenix aroma, sesame aroma, old Baigan aroma, rich aroma, special aroma, and rice aroma [9].**

The objective of classifying Baijiu types is not merely to understand its diverse flavor profiles but also to better inform consumers and producers about the characteristics that distinguish one type from another. By employing methods such as K-Means clustering and Principal Component Analysis (PCA), it becomes possible to systematically analyze the chemical and sensory attributes of various Baijiu types. Such analytical techniques help to uncover patterns in flavor composition, improve quality control, and even assist in product development. This study aims to explore how Baijiu types can be grouped and visualized based on multidimensional data, potentially contributing to a deeper appreciation of this traditional Chinese spirit and its intricate flavor diversity.

***Index Terms*—baijiu, k-means, principal component analysis.**

I. Introduction

Chinese baijiu, an integral part of traditional Chinese culture, is renowned worldwide for its unique brewing techniques, diverse aroma types, and long-standing historical significance. With the rapid development of China's economy, the baijiu industry has also shown robust growth momentum. In recent years, the market size of the baijiu industry has been continuously expanding, making it one of the key pillar industries of the Chinese economy. However, alongside this prosperity, the baijiu industry faces numerous challenges, especially the widespread prevalence of counterfeit products and inconsistent quality, which pose significant threats to consumer health and safety. These issues not only erode consumer trust in baijiu brands but also hinder the long-term and sustainable development of the industry.

A. Background

Currently, the primary methods for baijiu quality testing rely on technologies such as mass spectrometry, chromatography, and sensory analysis. While sensory analysis is a traditional and intuitive approach, its results are often subjective and susceptible to external factors such as the evaluator's physical condition and environmental changes, leading to outcomes that lack consistency and scientific rigor. On the other hand, mass spectrometry and chromatography can precisely analyze the aromatic components of baijiu, but their testing processes are complex, time-consuming, and costly [1]. Furthermore, the equipment required is expensive, bulky, and unsuitable for large-scale application. These limitations constrain the efficiency and accessibility of baijiu quality testing, making it difficult to meet the demands of the rapidly growing industry.

As a result, developing a fast, real-time, cost-effective, and efficient method for baijiu quality testing has become a critical research focus. The application of modern data analysis techniques and machine learning algorithms offers a promising solution to address the limitations of traditional testing methods. Specifically, clustering analysis can be used to classify and distinguish different types of baijiu, not only improving testing efficiency but also providing technical support for market regulation.

B. Motivation

Amid the rapid development of the baijiu industry, the market is flooded with a vast array of baijiu types and complex aroma profiles, posing significant challenges for both consumer choice and industry regulation. To better manage and regulate the baijiu market, especially in combating counterfeit or substandard products, it is essential to scientifically classify and effectively identify different types of baijiu. However, traditional testing methods face numerous practical difficulties, necessitating a more efficient and convenient approach to enable intelligent analysis of baijiu classifications.

Modern data analysis technologies provide new solutions for baijiu quality testing. Clustering analysis, as an unsupervised learning algorithm, can categorize baijiu into distinct groups based on its chemical composition or aromatic characteristics by analyzing the intrinsic distribution of data. Among clustering algorithms, the K-Means algorithm is widely used due to its simplicity, efficiency, fast computation, and ease of implementation. Additionally, Principal Component Analysis (PCA), as a dimensionality reduction technique, can effec-

tively extract the main features of data, reduce redundant information, and improve the efficiency and accuracy of data processing.

In this study, we combine the K-Means clustering algorithm with PCA to perform cluster analysis on baijiu sample data, achieving scientific classification and rapid identification of baijiu types. Compared to traditional testing methods, this approach offers the following advantages:

- Strong resistance to interference: Data analysis methods reduce the impact of human factors on testing results, ensuring objectivity and stability.
- High real-time performance: The fast computational speed of the algorithms enables quick classification analysis of baijiu samples.
- Low cost: The method does not require expensive testing equipment and relies solely on data analysis tools, making it suitable for large-scale application.

II. Related Works

The application of advanced data analysis techniques, such as clustering and dimensionality reduction, has gained increasing attention in various domains due to their ability to uncover hidden patterns in complex datasets. In this chapter, we explore how related research efforts have leveraged methods like PCA and K-Means clustering to address challenges in different fields, ranging from agriculture to finance and high-dimensional data clustering. These studies provide valuable insights into the potential of combining PCA and K-Means to improve accuracy, efficiency, and interpretability in data-driven problem-solving. By examining these approaches, we aim to identify key methodological innovations and practical applications that could inform our analysis of Chinese Baijiu classification. The following sections will delve into three representative studies, each showcasing unique advancements in utilizing PCA and K-Means, as well as their integration with other techniques. These examples will help contextualize the methods applied in this research, offering a solid foundation for further exploration of their relevance to Baijiu clustering.

A. Nutrient Detection in Crop Leaves

Research [5] on detecting nutrient content in crop leaves using spectral analysis, particularly nitrogen content, has been widely conducted. Most studies focus on annual crops like cucumber and wheat, employing near-infrared spectroscopy (NIRS) combined with modeling techniques such as Partial Least Squares Regression. However, these studies typically rely on either average spectral measurements of the entire leaf or random sampling of specific points, ignoring the uneven distribution of nitrogen within the leaf. In contrast, some research attempts to refine sampling areas through image processing techniques, such as using color-based machine vision segmentation to identify crop leaf damage or other features, as shown in Fig.1.

However, such methods are less effective for rubber tree leaves where color is not strongly correlated with nitrogen distribution. The innovation in this study lies in integrating

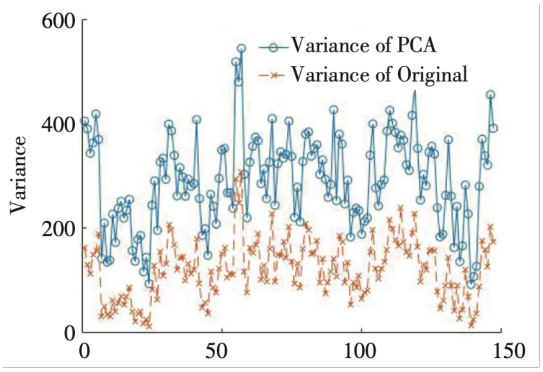

Fig. 1. Projection variance of raw hyperspectral data

PCA for spatial information extraction and K-Means clustering to group leaf pixel spectra into categories, enabling a more precise nitrogen distribution analysis and improving prediction accuracy. This method not only advances the understanding of nitrogen-sensitive areas but also provides a methodological foundation for rapid, non-destructive nutrient detection in crops.

B. High-Dimensional Data Clustering

PCA and K-Means are widely adopted for clustering high-dimensional data in various applications, such as image data and gene expression analysis. These studies [3] often leverage PCA's dimensionality reduction capability to enhance K-Means performance, improving computational efficiency and clustering accuracy. While application scenarios differ, the fundamental approach of combining PCA and K-Means remains consistent, showcasing their versatility across domains. For example, in the "Chinese Baijiu Types Cluster Analysis using K-Means and PCA" study, PCA and K-Means are utilized to classify Baijiu types [8], whereas in the rubber tree leaf research, they are applied to segment leaf pixel spectra for nitrogen content prediction. Both studies demonstrate the power of PCA and K-Means in handling high-dimensional data and uncovering meaningful patterns. These findings not only highlight the algorithmic similarities but also provide valuable insights for future research aiming to optimize data processing methods and tailor algorithms to specific applications. By building on these shared methodologies, this study offers a novel perspective on clustering analysis for Baijiu classification.

C. Quantitative Trading Strategies

In the financial domain, K-Means clustering has been widely used to classify financial assets and identify stocks with similar characteristics based on variables such as financial or technical indicators. Many studies apply clustering results to portfolio construction, risk management, or anomaly detection, while this research [4] uniquely integrates K-Means results into LSTM models for more efficient and accurate stock price predictions. LSTM and its variants are also extensively used in financial time series forecasting, such as predicting stock

prices or exchange rates. Unlike studies that apply LSTM directly to individual stocks, this research improves prediction efficiency by preprocessing data using K-Means clustering. Furthermore, combining K-Means with other machine learning or deep learning models has been a common practice in quantitative trading to leverage complementary strengths, and this study continues that tradition with a novel focus on reducing computational complexity. Additionally, strategy optimization for quantitative trading—covering aspects like trading frequency, risk control, and parameter tuning—remains a vital research area. This study contributes by optimizing cluster sizes, LSTM parameters, and trading frequencies through backtesting, ultimately enhancing profitability and risk management. The core innovation lies in the effective integration of K-Means clustering and LSTM, validated through its application to the CSI 300 index components.

The three studies reviewed illustrate diverse yet interconnected applications of PCA and K-Means clustering, showcasing their adaptability and effectiveness in solving complex problems across different fields. In crop nutrient detection, the integration of PCA and K-Means refines spectral analysis by addressing spatial variability, enabling precise and non-destructive assessments. In quantitative trading, K-Means enhances LSTM performance by preprocessing financial data, reducing computational complexity, and improving predictive accuracy for stock price forecasting. Finally, in high-dimensional data clustering, the synergy between PCA and K-Means provides a powerful framework for handling large datasets, as seen in applications ranging from gene expression analysis to Baijiu classification. These methods highlight the increasing trend of combining dimensionality reduction and clustering techniques to optimize data analysis processes, offering valuable insights for future research. By building on these advancements, this study aims to further explore the potential of PCA and K-Means in the classification of Chinese Baijiu types, contributing to the broader understanding of clustering analysis in high-dimensional data contexts.

III. PCA Dimension Reduction

The vast diversity of Chinese baijiu, coupled with the complexity of its chemical composition, necessitates dimensionality reduction techniques for effective cluster analysis. Principal Component Analysis (PCA) is a widely used dimensionality reduction method that projects high-dimensional data onto a lower-dimensional space through linear transformation. This transformation creates new variables, called principal components, which maximize the variance explained from the original data. This study employs PCA to reduce the dimensionality of a Chinese baijiu dataset and analyzes the characteristics of the reduced data.

A. PCA Principles and Procedures

The fundamental principle of PCA is to transform multiple indicators into a smaller number of composite indicators, which collectively retain as much of the original data's variation information as possible. The specific steps are as follows:

- Data Standardization: Before applying PCA, each feature in the baijiu dataset undergoes standardization to achieve a mean of 0 and a variance of 1. This is crucial because different features may have varying scales. Without standardization, features with larger scales would disproportionately influence the PCA results. The standardization formula is:

$$x' = \frac{x - \mu}{\sigma}, \tag{1}$$

  where x' is the standardized data,x is the original data, μ is the mean, and σ is the standard deviation.This research utilizes the sklearn.preprocessing.StandardScaler library function for standardization.
- Covariance Matrix Calculation: A covariance matrix is computed from the standardized data. This matrix describes the correlation between different features. The covariance matrix is a symmetric matrix; its diagonal elements represent the variance of each feature, and the off-diagonal elements represent the covariance between different features.
- Eigenvalue and Eigenvector Calculation: Eigenvalue decomposition is performed on the covariance matrix to obtain eigenvalues and eigenvectors. Eigenvalues represent the variance of the principal components, while eigenvectors represent the direction of the principal components. Larger eigenvalues indicate that the corresponding principal component explains a greater proportion of the variance.
- Principal Component Selection: Principal components are selected based on the magnitude of their eigenvalues. Typically, components whose cumulative explained variance ratio reaches a certain threshold (e.g., 85
- Data Projection: The original data is projected onto the lower-dimensional space formed by the selected principal components, resulting in the dimensionality-reduced data. The projection formula is:

$$y = XW \tag{2}$$

  where y is the reduced-dimensionality data, X is the standardized data, and W is a matrix composed of the eigenvectors corresponding to the selected principal components.

B. Selecting the Number of Principal Components

Choosing the appropriate number of principal components is critical. The selection aims to minimize information loss while reducing the number of components, simplifying the model, and improving computational efficiency. This study employs two methods for this selection:

Scree Plot: A scree plot visualizes the relationship between eigenvalues and the corresponding principal component number. The plot's elbow point—where the rate of eigenvalue decrease slows significantly—is observed. Components before the elbow point, possessing relatively larger eigenvalues, explain a substantial portion of the variance and are retained.

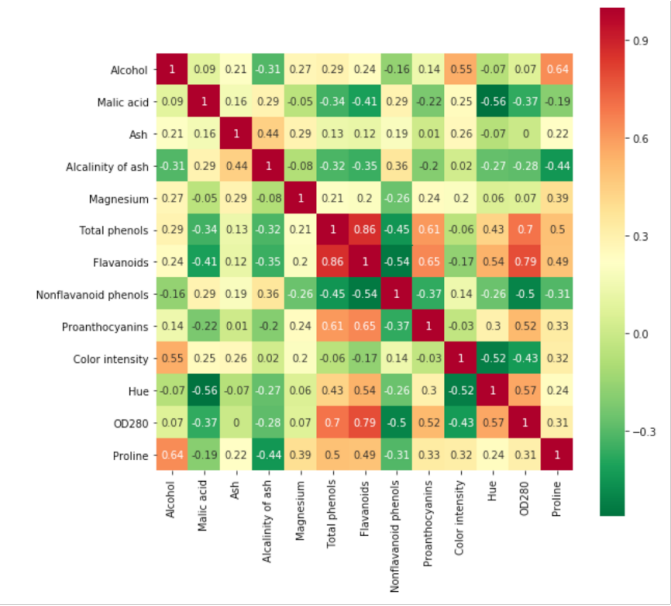

Fig. 2. Heatmap of the dataset baijiu

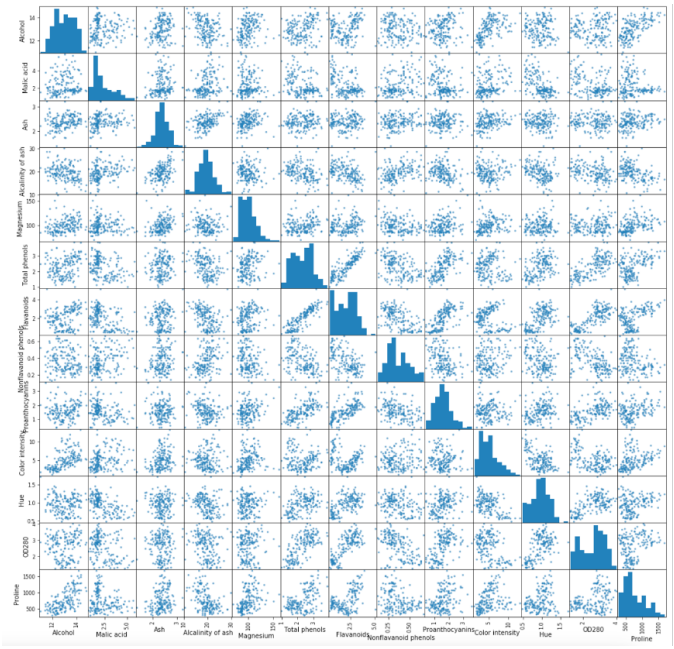

Fig. 3. the scatter plot of the dataset baijiu

Components after the elbow point, with rapidly decreasing eigenvalues, explain less variance and can be discarded.

Cumulative Explained Variance Ratio: The cumulative explained variance ratio is calculated. The number of principal components is chosen such that the cumulative explained variance ratio reaches a predefined threshold. This research combines the insights from the scree plot and the cumulative explained variance ratio to determine the optimal number of principal components.

In this researchthere are 178 instances of baijiu and 13 attributes and for each attribute, the distribution differ a lot. Then we are going to plot the scatter plot of the dataset baijiu as Fig.3.

C. PCA Dimensionality Reduction Results Analysis

After performing PCA dimensionality reduction, the results need thorough analysis, including:

- Explained Variance Proportion: The proportion of variance explained by each principal component is analyzed to understand each component's contribution to the original data. Bar charts or line graphs can visually represent this information.
- Principal Component Loadings: The loadings of each principal component are analyzed to understand the relationship between each principal component and the original features. The magnitude and sign of the loadings indicate the contribution and direction of the corresponding features to the principal component. Heatmaps [10] or loading plots effectively illustrate this information.
- Visualization of Reduced-Dimensionality Data: If the dimensionality-reduced data has low dimensionality (e.g., 2D or 3D), scatter plots can visualize the data distribution. This visualization aids in understanding the clustering structure of the reduced-dimensionality data, providing valuable insights for subsequent clustering algorithms like K-means. This allows for a better understanding of the inherent groupings within the baijiu types based on their chemical profiles. The visual representation facilitates the interpretation of the clustering results, potentially revealing relationships between specific chemical compounds and baijiu categories.

This expanded explanation provides a more comprehensive understanding of the PCA process and its application within the context of Chinese baijiu cluster analysis, aligning with the title "Chinese Baijiu Types Cluster Analysis using K-Means and PCA." The subsequent steps in the provided PDF document, namely K-Means clustering and the evaluation metrics, would then be applied to the dimensionality-reduced data generated by the PCA.

IV. K-Means Clustering

This chapter details the application of the K-Means clustering algorithm to analyze the reduced-dimensionality dataset of Chinese Baijiu types obtained through PCA in the previous chapter. K-Means, a fundamental unsupervised machine learning technique, aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean (centroid). This approach facilitates the identification of inherent groupings within the Baijiu dataset based on their chemical characteristics.

A. K-Means Algorithm Principles

The K-Means algorithm iteratively refines cluster assignments to optimize intra-cluster similarity and inter-cluster dissimilarity. The algorithm proceeds as follows:

- Initialization: The algorithm begins by randomly selecting k data points as initial centroids. The choice of k, the number of clusters, is a critical parameter discussed in Section B.
- Assignment: Each data point is assigned to the cluster whose centroid is closest. Euclidean distance is commonly used as the distance metric. The distance between a data point and a centroid is calculated as the square root of the sum of the squared differences between their corresponding feature values. Formally, given a data point x_i and a centroid c_j, the distance is:

$$d(x_i, c_j) = \sqrt{\sum_k (x_{ik} - c_{jk})^2}, \tag{3}$$

  where x_{ik} and c_{jk} are the k-th feature values of x_i and c_j , respectively.
- Update: The centroids are recalculated. The new centroid of each cluster is computed as the mean of all data points assigned to that cluster. This step involves calculating the mean for each feature across all data points within a given cluster.
- Iteration: Steps 2 and 3 are repeated until the centroids no longer change significantly or a pre-defined maximum number of iterations is reached. Convergence is typically assessed by monitoring the change in the within-cluster sum of squares (WCSS) between iterations. When the change falls below a specified threshold, the algorithm is deemed to have converged.

B. Determining the Optimal Number of Clusters (k)

Selecting the appropriate number of clusters (k) is crucial for achieving meaningful results. Improper selection can lead to over-clustering (too many clusters, obscuring underlying patterns) or under-clustering (too few clusters, masking important distinctions). This study employs two common methods:

- Elbow Method: This method [4] involves calculating the WCSS (Within-Cluster Sum of Squares) for different values of k, as is shown in Fig.4. WCSS represents the sum of squared distances of all data points to their respective cluster centroids. A plot of WCSS against k is generated. As k increases, WCSS generally decreases. The "elbow point" on the plot – the point where the rate of decrease in WCSS slows significantly – suggests an appropriate value for k. This point represents a balance between minimizing WCSS and avoiding over-clustering.
- Silhouette Score: The silhouette score is a metric that quantifies the quality of clustering by measuring how similar a data point is to its own cluster compared to other clusters. The silhouette coefficient $s(i)$, for a data point i is given by:

$$s(i) = \frac{(b(i) - a(i))}{max(a(i), b(i)}} \tag{4}$$

  where $a(i)$ is the average distance between data point i and all other points in the same cluster, and $b(i)$ is the average distance between data point i and the nearest neighboring cluster. The silhouette score ranges from -1 to 1. A score close to 1 indicates that the data point is well-clustered, while a score close to -1 suggests that the data point might be assigned to the wrong cluster. A score around 0 indicates that the data point lies between two clusters. The average silhouette score across all data points is used to select the optimal k value. Higher average silhouette scores indicate better clustering.

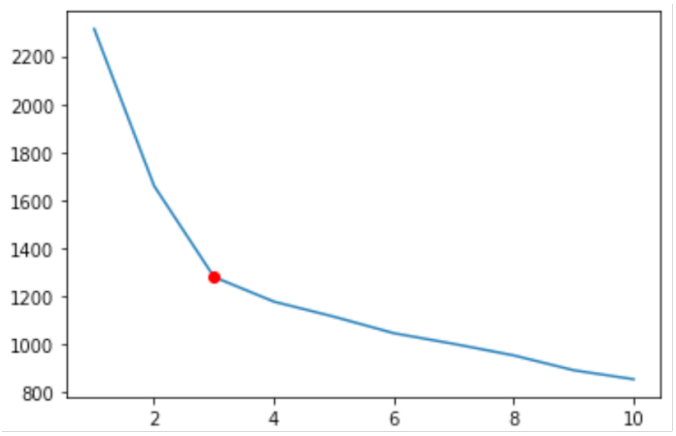

Fig. 4. Number of clusters using Elbow Method

C. K-Means Clustering Results Analysis

After performing K-Means clustering, a comprehensive analysis of the results is necessary:

- Cluster Feature Analysis: This involves examining the distribution of features within each cluster (e.g., calculating the mean, standard deviation, and other descriptive statistics for each feature in each cluster). These statistics illuminate the characteristics that distinguish different Baijiu types.
- Centroid Analysis: The location of each cluster centroid provides insights into the central tendency of each cluster in the feature space. Visualizing centroids in a reduced-dimensional space (e.g., 2D or 3D) aids in understanding the spatial relationships between clusters.
- Visualization: If the data has been reduced to a low dimensionality (ideally 2D or 3D), scatter plots can effectively visualize the clustering results. Different clusters can be represented using distinct colors or markers, making it easy to observe the spatial separation of the clusters.
- Cluster Quality Evaluation: Metrics like the silhouette score and the Adjusted Rand Index (ARI) are used to objectively assess the quality of the clustering. The ARI, in particular, is valuable because it accounts for the possibility of random clustering, providing a more robust measure of clustering performance.

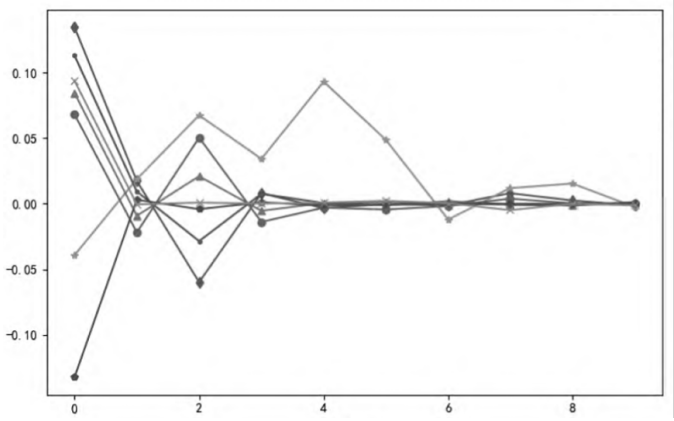

Fig. 5. Distribution of cluster centers in each cluster after dimensionality reduction

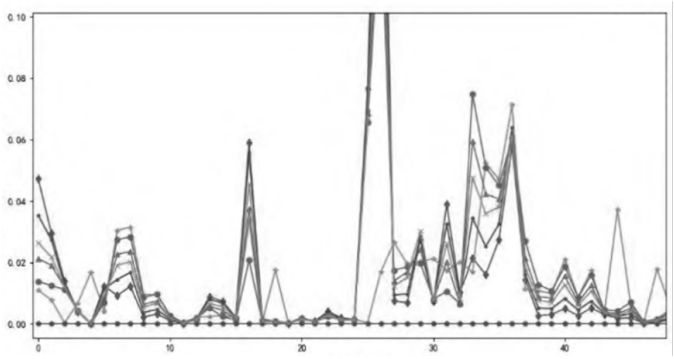

Fig. 6. The original distribution ratio of cluster centers of each cluster

D. K-Means Clustering with PCA

Combining PCA dimensionality reduction with K-Means clustering enhances efficiency and mitigates the curse of dimensionality, especially when dealing with high-dimensional data like the Baijiu dataset with numerous chemical components. PCA first extracts the principal components, capturing the most significant variance in the data. Subsequently, K-Means is applied to this reduced-dimensional representation. By carefully choosing the number of principal components retained by PCA and the optimal k for K-Means, the combined approach yields more accurate and efficient cluster assignments. Furthermore, the lower dimensionality facilitates clearer visualization and interpretation of the clustering results. The choice of the optimal number of principal components to retain is guided by the cumulative explained variance ratio, aiming to retain a sufficient amount of information while reducing dimensionality.

V. Evaluation

The application of PCA-KMeans in the cluster analysis of Chinese Baijiu data demonstrates significant effectiveness, as evidenced by the increasing stability and improved cluster performance throughout the process. PCA, employed for dimensionality reduction, effectively captures the principal components that hold the majority of the variance in the dataset, simplifying the high-dimensional data into a lower-dimensional space without substantial information loss.

Fig.5, which illustrates the distribution of cluster centers after dimensionality reduction, clearly shows that the clusters formed are more distinct and well-separated in the reduced feature space, indicating that PCA successfully enhances the separability of the clusters. This dimensionality reduction step not only improves computational efficiency but also reduces noise in the data, allowing the K-Means algorithm to focus on meaningful patterns. Fig.6, which reflects the original distribution ratio of cluster centers across each cluster, further supports the conclusion that the clustering process stabilizes over iterations.

Initially, there may be some fluctuation in cluster assignments, as seen in the sharp peaks of the distribution, but as the iterations progress, the clusters converge and exhibit more consistent distributions. This stabilization highlights the robustness of combining PCA with K-Means for clustering, as it mitigates the sensitivity of K-Means to initial cluster centroid placement and improves the algorithm's ability to identify natural groupings in the data.

Moreover, the iterative refinement of cluster centers aligns well with the physical and chemical properties of Baijiu, such as alcohol, malic acid, ash, and total phenols, which are inherently complex and interdependent. By leveraging PCA, the algorithm not only reduces redundancy in these variables but also ensures that clusters are formed based on their most defining characteristics. Overall, the PCA-KMeans approach delivers strong clustering results, with the analysis becoming increasingly stable as the clusters converge. This method is particularly well-suited for datasets with high dimensionality and complex variable interactions, such as the Baijiu dataset, and its application here demonstrates its utility in identifying meaningful patterns and groups within the data.

VI. Conclusion and Future Directions

A. Summary of Key Findings

This study has demonstrated the effectiveness of combining Principal Component Analysis (PCA) and K-Means clustering in the classification and analysis of Chinese Baijiu types. The main findings are as follows:

- Dimensionality Reduction through PCA: PCA was applied to the high-dimensional Baijiu dataset, reducing the number of features while preserving the essential variance in the data. Analysis of the explained variance ratios and principal component loadings provided insights into the underlying chemical profiles that distinguish different Baijiu types.
- Optimal Cluster Identification with K-Means: The K-Means algorithm was utilized to group the dimensionality-reduced Baijiu data into distinct clusters. The optimal number of clusters was determined using the elbow method and silhouette analysis, striking

a balance between capturing inherent groupings and avoiding over-clustering.

- Characterization of Baijiu Clusters: Examination of the cluster feature distributions and centroid locations revealed the distinguishing characteristics of each Baijiu type. Specific chemical compounds or aromatic profiles were associated with the identified clusters, enhancing the understanding of Baijiu's diverse flavor profiles.
- Visualization and Interpretation: The lower-dimensional representations obtained through PCA enabled effective visualization of the clustering results. Scatter plots and centroids plots illustrated the spatial relationships between the Baijiu clusters, facilitating the interpretation of the grouping patterns.
- Clustering Quality and Validation: Evaluation metrics, such as the silhouette score and Adjusted Rand Index, confirmed the high quality and robustness of the clustering assignments, validating the combined PCA-K-Means approach for Baijiu classification.

B. Implications and Applications

The findings of this study have several important implications for the Baijiu industry and broader research related to traditional Chinese spirits.

Baijiu Quality Assurance: The classification framework developed in this research can be leveraged for rapid, cost-effective, and objective Baijiu quality testing [1]. This can aid in combating counterfeit products and ensuring consistent quality, addressing a critical challenge facing the industry.

Product Development and Optimization: The in-depth understanding of Baijiu's chemical profiles and flavor characteristics can inform product development, allowing manufacturers to design new Baijiu varieties or optimize existing ones to cater to evolving consumer preferences.

Regulatory and Market Oversight: The systematic classification of Baijiu types can support industry regulators in monitoring the market, enforcing standards, and protecting consumers from substandard or fraudulent products.

Educational and Promotional Initiatives: The insights gained from this research can be utilized to educate consumers about the rich diversity of Baijiu, promoting a deeper appreciation for this traditional Chinese spirit and its unique flavor nuances.

C. Future Research Directions

Building upon the foundations established in this study, several avenues for future research can be explored:

Expanding the Baijiu Dataset: Incorporating a more comprehensive and diverse dataset of Baijiu samples, including a broader range of types and regions, could further enhance the robustness and generalizability of the classification framework.

Integrating Sensory Evaluation: Incorporating sensory analysis data, such as expert tasting notes and consumer perception surveys, could provide a more holistic understanding of the relationship between Baijiu's chemical composition and its organoleptic properties.

Exploring Alternative Clustering Algorithms: Investigating the application of other clustering algorithms, such as hierarchical clustering or density-based methods, could yield additional insights and potentially uncover alternative grouping structures within the Baijiu dataset.

Predictive Modeling and Quality Prediction: Building upon the classification capabilities, developing predictive models that can reliably estimate Baijiu quality attributes based on chemical profiles could aid in quality control and product optimization.

Cross-cultural Comparisons: Expanding the research to include Baijiu samples from different regions or comparing them to other traditional spirits from around the world could provide a broader perspective on the global diversity of fermented beverages.

By pursuing these future research directions, the scientific understanding of Chinese Baijiu and its complex flavor profiles can be further enhanced, ultimately contributing to the preservation, promotion, and sustainable development of this iconic traditional spirit.

REFERENCES

[1] JinZhu.Design and Research of Electronic Nose System forIdentification of Liquor Varieties[D].JiangSu University,2019.

[2] Bschaden A , Stroebele-Benschop N .Associations and patterns in lifestyle and body weight among university students over one year into the Covid-19 pandemic: A cluster analysis[J].NFS Journal[2025-01-10].DOI:10.1016/j.nfs.2024.100206.

[3] LiqiangZhao, TaoZhao, ShuixiongTang.Short-Term OD Passenger Flow Prediction of Urban Rail Transit Based on PCA-Kmeans Algorithm[J].2023,10(04):60-68.DOI:10.14103/j.issn.2095-8412.2023.08.007.

[4] HaiangFeng.Quantitative Strategy Research Based on Kmeans and LSTM[D].Guangzhou University,2023. DOI:10.27040/d.cnki.ggzdu.2023.002181.

[5] SuixiZhong, ZiboLi, RongnianTang.Near Infrared Hyperspectral Diagnostic Model for Nitrogen Content of Rubber Tree Leaves Based on PCA Kmeans Clustering Spectrum[J].HaiNan University,2020,38(03):260-269. DOI:10.15886/j.cnki.hdxbzkb.2020.0036.

[6] YunlongZhao, ZhanshuangLiu, YanLi.Evaluation of Insulator Surface Pollution State Based on PCA kmeans and Spectrum Features[J].2023,(05):193-200. DOI:10.16188/j.isa.1003-8337.2023.05.026.

[7] Xinjie Y ,Linlin Z ,Zhongjun H .A Novel Real-Time Data-Based PEMFC Performance Evaluation Model Using Improved PCA-KMEANS-XGBOOST for PEMFC Hybrid Vehicles in China[C].China Society of Automotive Engineers.Shanghai Hydrogen Propulsion Technology Co.Ltd.:,2023:1. DOI:10.26914/c.cnkihy.2023.068541.

[8] Lu S ,Yu H ,Wang X , et al.Clustering Method of Raw Meal Composition Based on PCA and Kmeans[C]//F.School of Electrical Engineering,University of Jinan;2018:4.

[9] Zhang J ,Lu J ,Yu X , et al.Characterization of aroma differences in Jiangxiangxing Baijiu with varying ethanol concentrations: Emphasis on olfactory threshold changes of aroma compounds[J].Food Chemistry,2025,469142506-142506.

[10] Zhang (R .Enhancing Clustering Stability and Efficiency: A Framework for Optimizing K-means, K-medoids, and K-shape with Intelligent Algorithms[J].Journal of Engineering Research and Reports,2024,26(12):192-206.