A Cluster Analysis in Wine Dataset

Leo Zhan zhandw.leo@gmail.com

Abstract

Data are the result of a chemical analysis of wines growing in the same region of Italy but from three different varieties. This analysis identified the number of 13 components found in each of the three wines. The purpose of our project is to conduct cluster analysis according to the type of these wines. We will first observe the distribution of the data, then standardize, then perform PCA principal component analysis, retain the main chemical composition of the wine, and then compare the results under different cluster analysis.

1 Introduce of technology

In this section, we will introduce the data analysis technologies we used in the project, including PCA dimension reduction, k-means clustering, Gaussian mixture clustering and hierarchical clustering.

1.1 Dimension reduction

In many fields of research and application, it is usually necessary to observe data containing multiple variables, collect a large amount of data, and analyze to find patterns. Multivariate large datasets undoubtedly provide rich information for research and application, but they also increase the workload of data collection to a certain extent. More importantly, in many cases, there may be correlations between many variables, which increases the complexity of problem analysis. If each indicator is analyzed separately, the analysis is often isolated and cannot fully utilize the information in the data. Therefore, blindly reducing indicators will lose a lot of useful information, leading to incorrect conclusions.

Therefore, it is necessary to find a reasonable method to reduce the number of indicators that need to be analyzed while minimizing the loss of information contained in the original indicators, in order to achieve the goal of comprehensive analysis of the collected data. Due to the existence of a certain correlation between variables, it is possible to consider transforming closely related variables into as few new variables as possible, so that these new variables are uncorrelated in pairs. Therefore, fewer comprehensive indicators can be used to represent various types of information that exist in each variable. Principal component analysis and factor analysis belong to this type of dimensionality reduction algorithm.

Dimension reduction is a preprocessing method for high-dimensional feature data. Dimension reduction is the process of preserving the most important features of high-dimensional data, removing noise and unimportant features, in order to achieve the goal of improving data processing speed. In practical production and application, dimensionality reduction within a certain range of information loss can save us a lot of time and cost. PCA dimensionality reduction is a dimensionality reduction method.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.2 Introduction of Cluster Analysis

Many biological analyses involve partitioning samples or variables into clusters on the basis of similarity or its converse, distance. For example, in a gene expression study, we might seek subsets of patients with similar expression, or take a complementary approach and identify similarly expressed genes across patients. Clustering is a type of unsupervised learning comprising many different methods. Here we will focus on three common methods: hierarchical clustering, which can use any similarity measure, gaussian mixture clustering, which can be seen as an optimization of the k-means model, and k-means clustering3, which uses Euclidean or correlation distance.

Fundamentally, all clustering methods apply the same approach. First, we calculate similarity and then use it to group objects (e.g., samples) into clusters. However, the clustering output is useful only if the clusters correspond to the data's biologically relevant features that were not used to define the grouping. To judge clusters' validity, we need external information; clusters are not known in advance. For example, our confidence in the validity of our clusters increases if patients in each cluster share a phenotype, or if genes in each cluster share a sequence motif; but confidence increases only if this information was not used to assess similarity in the first place.

2 Dataset

We obtained the data collection from the following website,

`https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering`

The data contains thirteen properties, each representing the thirteen different chemical components.
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline
Then we can view the characteristics of the data through data visualization. As well as outlier points and quantiles of data viewed through boxplot.

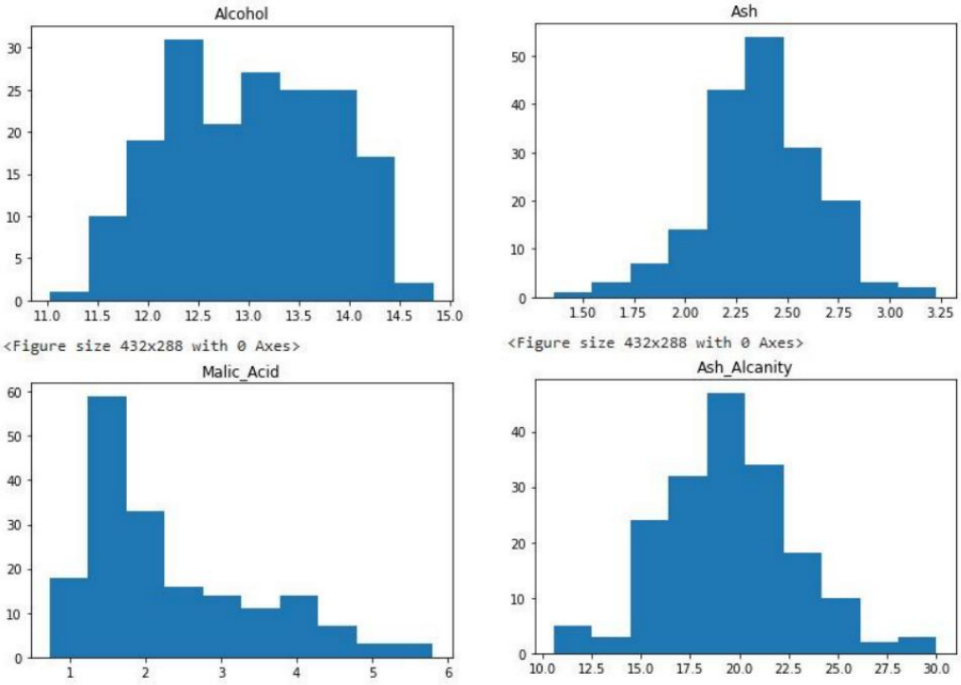

Figure 1: distribution of feature.

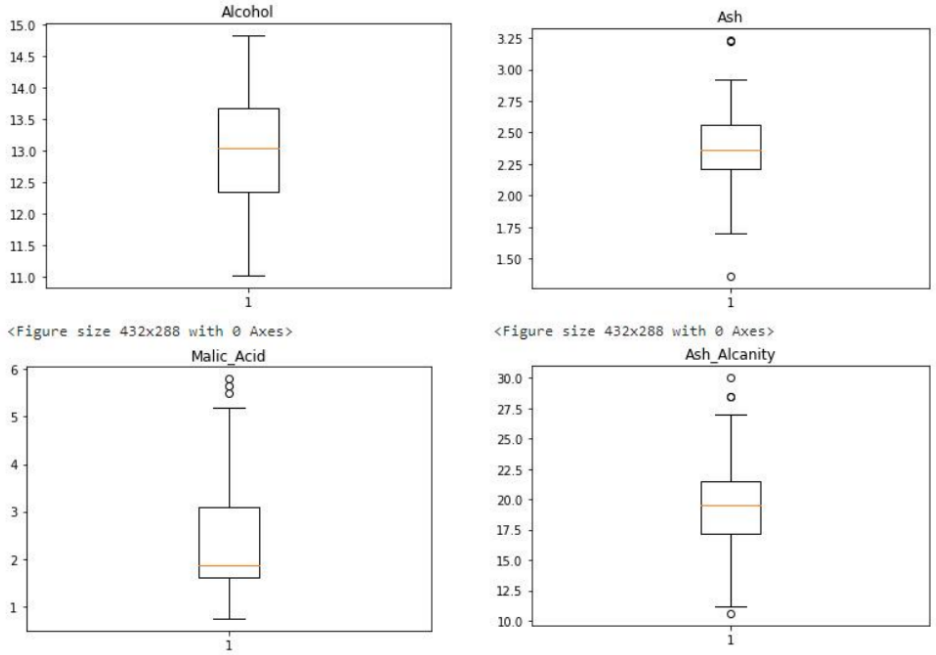

Figure 2: boxplot of feature.

There may be correlation between different data features, which is not conducive to clustering the data. Therefore, before this, we conducted data dimensionality reduction using PCA principal component analysis. Then we analyze the correlation between the features of the data, and we observe the correlation coefficient between each feature by drawing the heat map. Below are the results of the heat map.

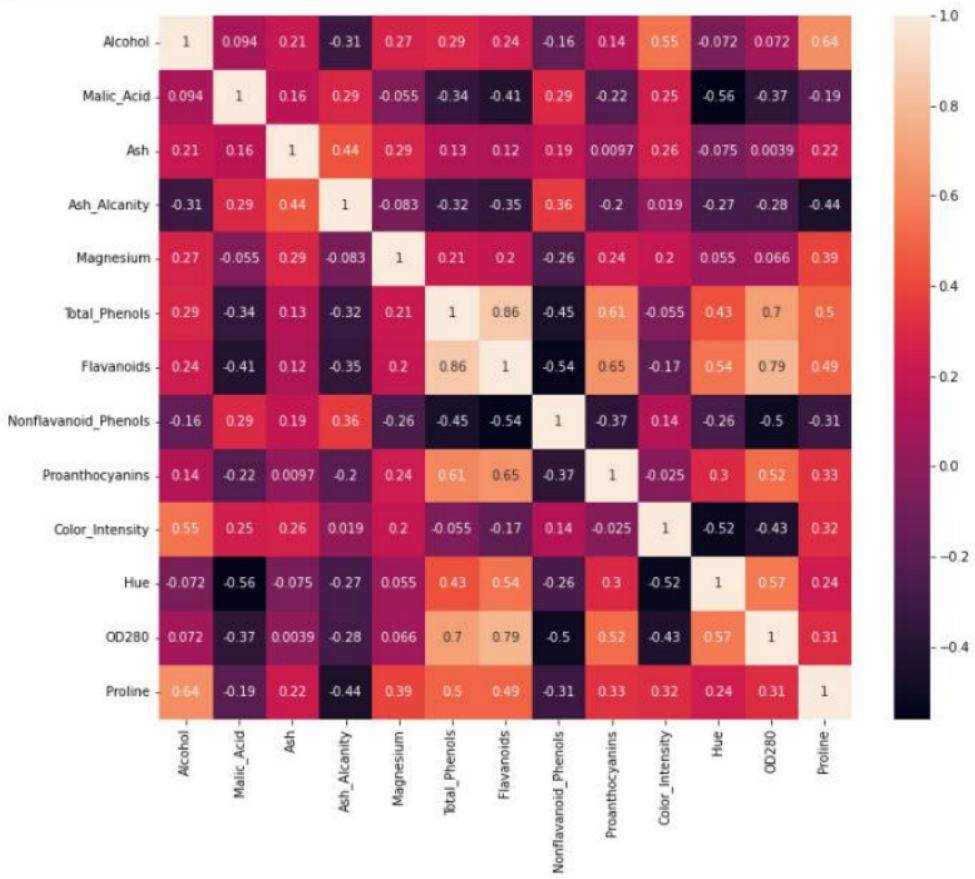

Figure 3: correlation of feature.

3 Experiment

Before data clustering, we standardize and normalized the data to normalize the data for subsequent PCA dimension reduction. Normalization and normalization of data are methods of feature scaling and a key step in data preprocessing. Different evaluation indicators often have different dimensions and dimensional units, which will affect the results of data analysis. In order to eliminate the influence of dimensions between indicators, data normalization / standardization processing is needed to solve the comparability between data indicators. After the raw data, each index is in the same order of magnitude, which is suitable for comprehensive comparative evaluation.

After PCA dimensionality reduction, we will start clustering analysis of the data. Firstly, we need to determine the k value, and we used both methods to determine that k. We calculate the square of the sum of the distance to the center of the cluster for all samples in the dataset, and find the k value corresponding to the sum of errors.

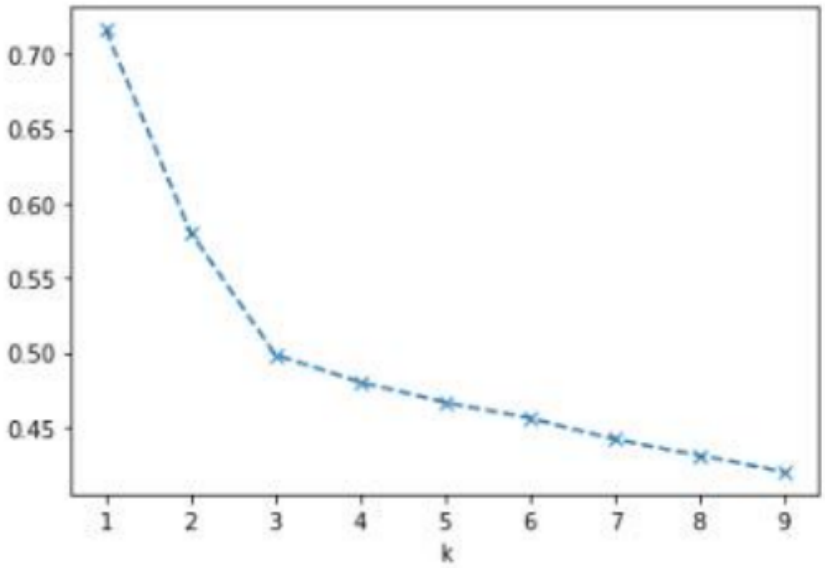

Figure 4: elbow graph.

We found that when k=3, the mutation is minimal, so select k=3. We then used the k-means for the clustering. Then We can view the k-means clustering results.

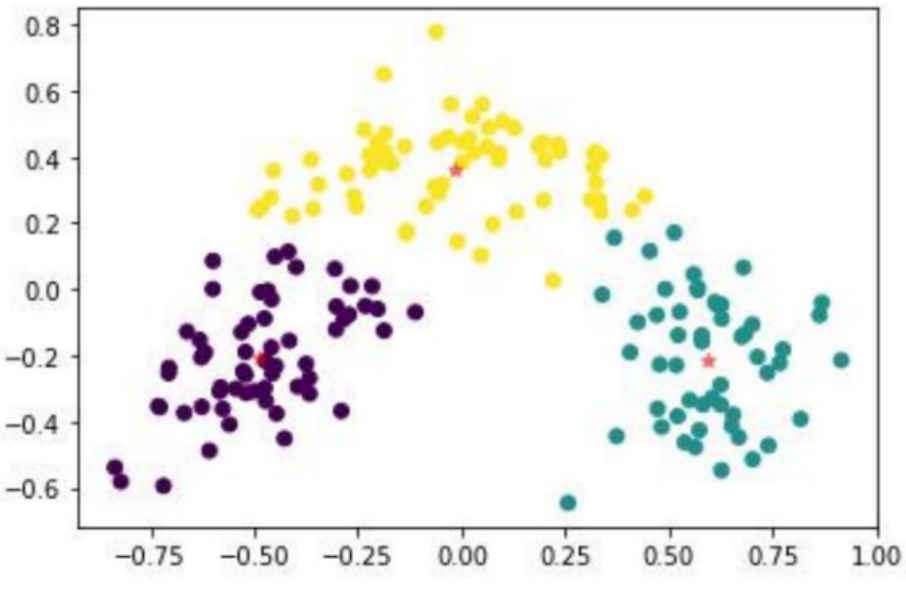

Figure 5: k-means result.

Then we used a Gaussian mixture clustering based on the probabilistic model, the data are decomposed into several models based on Gaussian probability density function formation, which results as follows.

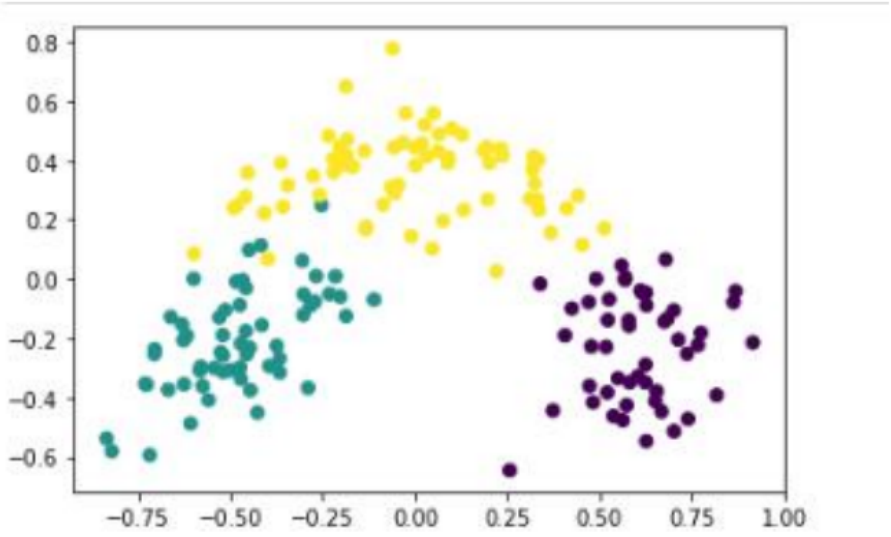

Figure 6: Gaussian mixture clustering result.

Finally, we use the approach of hierarchical clustering. Hierarchical clustering The clustering results are interpreted visually by drawing the dendrogram. The abscissa in the resulting dendrogram represents each sample, and the ordinate represents the distance. Also does not need to specify the number of clusters.

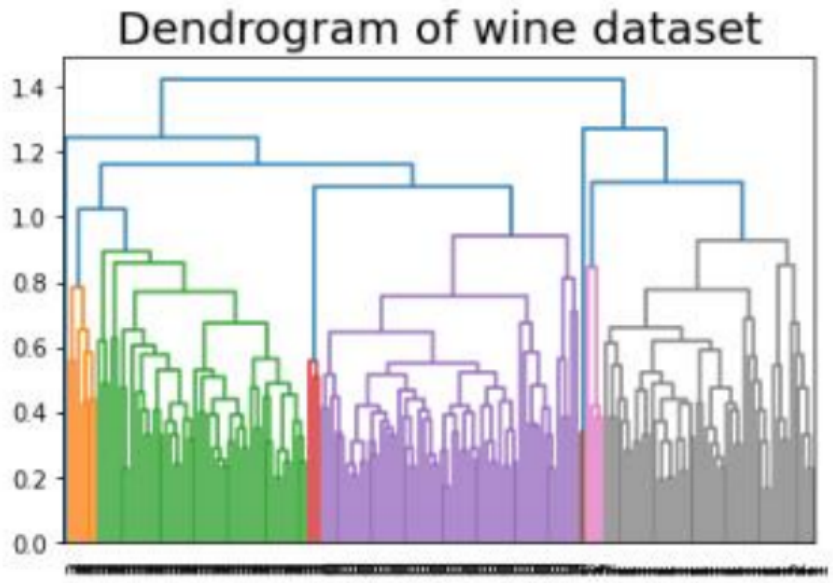

Figure 7: Hierarchical clustering.

4 Summary

The biggest feature of K-means clustering is that it can quickly process a large amount of data, but itcan only process quantitative data but not classified data. Moreover, K-means clustering needs to independently set the number of cluster categories, and cannot automatically find the optimal number

of cluster categories, which may lead to unstable quality of results. Hierarchical clustering, the basic idea is to take multiple samples as one class, calculate the distance between two samples, merge the nearest two classes into a new class, and then calculate the distance, and then merge, until there is only one class.

Hierarchical clustering can handle categorical and quantitative data, but it is relatively slow, usually subjectively judging the number of cluster categories combined with relevant results.

Gaussian mixture model clustering is a probabilistic clustering method that assumes that all data samples are generated by a mixture of mixed multivariate Gaussian distributions.. Where is the probability density function of the n-dimensional random vectors following a Gaussian distribution.

Through the experimental results, we found that after PCA dimension reduction, clustering using three algorithms can achieve good performance. The results of the experiment show that the kinds of wine can be divided into three categories according to his chemical composition.

References

[1] Ma, Y. Z. . (2013). A Tutorial on Principal Component Analysis.

[2] J. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, 1967.