Some Unsupervised Learning Clusteting Methods and its Application on a Coffee Quality Study

无监督学习聚类方法及其在咖啡品质研究中的应用

Name: **Rui Liu**

ID Number: **1718304**

Supervisor: **Dr. Mu He**

5th May 2021

Abstract

In this report, we mainly discuss four kinds of unsupervised learning methods and apply them to the study of coffee quality. Firstly, we use principal component analysis to reduce the dimension of the coffee dataset, and select the first six principal components to explain most of the information. Then, hierarchical clustering, K-means clustering and Gaussian mixture model are used to cluster the data set, and the clustering results are visualized by projection on the first six principal components. The results of the three clustering methods are different. They get 5, 2 and 3 clusters respectively, but the projection of their results on the first principal component and the second principal component can be well separated, and the performance on the projection of other principal components is not very obvious. The results of hierarchical clustering and Gaussian mixture model show that countries with similar geographical location will be divided into one cluster. The results of K-means clustering divide the two varieties of coffee in the sample into two clusters. In the future, we can further study the quality of coffee and learn more unsupervised learning methods.

Keywords: principal component analysis , hierarchical clustering, k-means clustering,Gaussian mixture model, visualization, coffee

摘要

在这篇论文中，我们主要讨论了四种无监督学习的方法，并将这四种方法应用到咖啡品质的研究中。首先我们运用了主成分分析来对咖啡数据集进行降维预处理，并选择了前 6 个主成分来解释大部分的信息。然后分别用层次聚类，k-means 聚类，高斯混合模型对数据集进行聚类分析，并将聚类结果通过投影在前 6 个主成分上进行可视化。三种聚类方法的聚类结果是不同的，分别得到了 5 个，2 个，3 个集群，但他们在第一个主成分和第二个主成分上的投影被很好的分离开来，在其他主成分的投影上表现不是很明显，都发生了一些重叠。层次聚类和高斯混合模型的聚类结果都表现出了地理位置相近的国家会被划分在一个聚类中，k-means 聚类的结果将样本中的两个品种的咖啡分为了两个聚类。未来我们还可以深入咖啡品质的研究并学习更多的无监督学习方法。

关键词：主成分分析, 层次聚类, k-means 聚类, 高斯混合模型, 可视化, 咖啡

Content

1 Introduction

Machine learning (ML) is a branch of artificial intelligence (AI) [45], before machine learning was proposed, machines operated in accordance with the rules set by human beings and the knowledge summarized. On the one hand, this approach costs too much for human, on the other hand, it can never surpass the creator. It has been suggested that if machines could learn on their own [3], all problems would be solved, then machine learning would be born. Machine learing is a method for computers to use data instead of instructions to carry out various kinds of work, so that algorithms can find patterns and features in a large amount of data, and then can get decisions and predictions through new data [36]. Machine learning has been widely used in various fields, including medicine, finance, chemistry, biology and so on, because of its powerful decision and prediction functions[14] [48] [5].

Supervised learning is to train a function model based on the given training data set, which can be used to predict the result when we bring in the new data set. For unsupervised learning, there is no known result, and the category of sample data is unknown and unmarked. The classification of sample data is based on the similarity between observation points, so as to obtain the result with the smallest gap within the cluster and the largest gap between the clusters.

In this project, we describe the literature review of coffee, hierarchical clustering, K-means clustering, density based spatial clustering of applications with noise and Gaussian mixture model in section 2, and learn the historical background, prospects and applications of these six unsupervised learning methods in various fields. In the next chapter, we apply five methods on the coffee dataset. In each section, we use pictures and tables to show the algorithm results. Each clustering method produces different results. In order to visualize the clustering results, the clustering results are projected on the first six principal components generated by principal component analysis.

2 Literature review

2.1 The historical background of coffee and its contribution to various fields

Coffee is one of the three major beverages in the world. Coffee can be traced back to more than 4000 years in the world. Arabs doing business in Ethiopia brought coffee seedlings to the Arabian Peninsula and gradually cultivated excellent "Arabian coffee" [28]. Now coffee has been widely cultivated in more than 50 countries in Asia, Africa and Latin America, and it has become a worldwide popular beverage. There are currently three main types of coffee beans, Arabica, Robusta and Libelica. The origins of Arabica coffee and Robusta coffee are Ethiopia and Africa. The production of coffee beans accounts for about 70% and 25% of the world's production respectively [39]. These two types of coffee are also the most important in the coffee market [16]. The Liberian coffee from Liberia in Africa accounts for less than 5% of the world's production. This also explains why the coffee samples to be studied in this project only include Arabica and Robusta, and Arabica coffee samples account for 79% of the total number of samples.

As a drinkable product, people are naturally concerned about food safety. The dispute about whether coffee is food, poison, or medicine has always existed. Scientists, medical scientists, and food scientists have been exploring the effects of drinking coffee on the human body in different fields. People often think that the caffeine in coffee can cause arrhythmia, but there is not a lot of scientific evidence to prove this is right. Lynn and Kissinger showed in the article that although the results of studies on different populations are different, proper caffeine intake will not affect heart rate, blood pressure and heart rhythm [33], for those who already have a heart For patients with disease, limiting caffeine will not play a positive role in disease management [38]. This does not mean that drinking coffee is completely harmless to our body. Excessive drinking of coffee will affect blood pressure and systemic vascular resistance. Arterial stiffness has an adverse effect [11]. Drinking coffee can also affect the sleep of some sensitive people. The sleep of the elderly is more susceptible than the young people. Of course, there are individual differences in the young people [30] Johnston also

mentioned in his report that the chlorogenic acid contained in coffee regulates the secretion of human gastrointestinal hormones and insulin by antagonizing the transport of glucose [23]. We cannot say that drinking coffee is necessarily harmful to our health, but drinking too much coffee is definitely not a wise choice.

Coffee is the economic center of many developing countries and the world economic center. It also plays a vital role in maintaining the sustainability of our planet and economic development. Brazil is one of the world's largest coffee producing countries [49]. Since the middle of the 19th century, it has been the world's most successful coffee economy. At the beginning of the twentieth century, Brazil controlled 70% of the world's coffee supply [10]. Indonesia, Vietnam, Colombia, Ethiopia, India, Honduras, Uganda, Mexico, Guatemala, is also the important producer of coffee, for the developing countries, a lot of coffee exports not only promote the economic growth to a certain extent, but also for these countries to provide a large number of jobs.Figures from the National Coffee Society show that the US is a huge importer of coffee and the world's largest consumer, with more than 150,000 people working full or part time [25].The actual role of coffee in the U.S. economy is even more important if you include the indirect output generated by the coffee industry, such as retail sales, production facilities, transportation, ports, warehouses, and packaging.Similarly, coffee plays an extremely important role in other countries. Coffee is slowly taking the place of tea in Japan/cite1990The. Coffee has also become the number one food item in Canada, with nearly one million people employed in the profession, almost 7% of the total employment.As for the coffee industry, the other industries it produces are also contributing to economic progress. The coffee shop needs employees, the packaging of the coffee requires bags, the drinking of the coffee requires straws, and the transportation of the coffee requires workers. The coffee industry is an extremely important industry for both producing and importing countries.

Coffee is a common drink in our lives and has been thoroughly studied by many people. In their paper, Lamparelli and colleagues used the expectation maximization to consider five clusters of coffee plants under different conditions after harvest, and used the t-test to verify similarities between clusters.The five clusters obtained are composed of different coffee crop conditions [29]. Khumaidi prepared a dataset of 170 data to predict coffee yield

using CRISP-DM and multiple linear regression algorithms [27]. In this project, I will use several different clustering methods to cluster the collected coffee samples and compare the chemical factors that affect the quality of coffee.

2.2 Principal component analysis (PCA)

Principal component analysis(PCA) is a commonly used dimensionality reduction algorithm in data mining. It is a multivariate statistical method proposed by Pearson in 1901 and later developed by Hotelling [19] in 1933. Its main use is ”dimension reduction”, in order to comprehensively analyze a problem, it tends to propose many indicators that reflect information in different degrees. Using multivariate statistical analysis of multi-indicator problems is that too many indicators will increase the complexity of the problem, and the ideal result is that less indicators reflect more information [41]. In many cases, indicators are related to each other, so the information reflected by each indicator is bound to overlap [1]. The work of principal component analysis is to establish as few indicators as possible for all the original indicators, so that the new indicators are not related to each other, and to ensure that the information reflected by the new indicators keeps the original information as much as possible[18]. The new indicator is called the principal component of the original indicator. For example, in the practical application of biology, information about hundreds or thousands of genes is usually obtained, and each of these genes can affect each other. After principal component analysis, a finite number of principal components can represent their genes[53]. This is what we call dimensionality reduction. Although the principal component analysis reduces the number of indicators in the dimension reduction process, it also eliminates the difference information of the variation degree of indicators, so the degree o f mutual influence between various indicators cannot be taken into account [44]. With the development of principal component analysis, more and more problems have been put forward. People use the method of principal component analysis to extract the features of problems and solve many problems in economy, science, development and other aspects [24]. In the field of science, principal component analysis (PCA) has important value and application prospect for decomposing spatiotemporal data, dimensionality reduction and signal extraction of satellite radar images [7]. In terms of environmental development, principal component analysis is used to analyze the pollution problem of Rybnik Reservoir [32]. The wide application of principal component analysis in various fields promotes the continuous improvement of the

method itself, showing the diversity of the method. In order to adapt to solve more complex problems, its algorithm is becoming more and more mature, and can be well combined with other methods.

2.3 Cluster analysis

As an important tool in data mining, cluster analysis can divide known data sets into meaningful or useful clusters [35]. It is an important unsupervised learning method [47], many professional fields have recognized it importance [20],such as biology, economics, and medicine [31]. Generally speaking, it is an unsupervised learning process to find a set of similar observations in the data set by dividing unmarked objects into groups. Cluster analysis is applied in many different fields, and data sets in different fields have different characteristics, so the purposes of cluster analysis on data are also different. When the data sets are different, or the purpose of use is different, the selected method of cluster analysis is also different. The following mainly introduces two clustering analysis methods, hierarchical clustering and k-means.

2.3.1 Hierarchical clustering

The interest in hierarchical clustering stems from different applications, for example, we can observe that chimpanzees are animals similar to humans, and when we think about specific categories, they are different from humans [8]. The emergence of hierarchical clustering is to better understand the relationship between unknown objects, so that our understanding of clustering is not just limited to a single granularity, the constructed hierarchical structure is composed of different granularity clustering [12]. In hierarchical clustering, each data is taken as a class separately, and a measurement method is selected to calculate the distance between classes, which can also be said to measure the level of similarity, and the most similar classes are combined together, then recalculate the distance between the new class and other classes,, and similar classes are selected for merging. This process is repeated repeatedly, reducing the number of classes with each merging until all the data becomes one class. Visualization is very important for people to explore large document collections, and hierarchical clusters solve this problem by providing views of data at different levels of granularity [56], while further narrowing the scope of data by finding a way to measure the

similarities and differences of data sets. Since hierarchical clustering generally does not need to assume data characteristics in advance, it is often more prominent than other algorithms [37].

2.3.2 K-means clustering

K-means was first published in 1955 and is one of the most popular and simplest clustering algorithms [22], many clustering algorithms have been published so far, but k-means has gained a lot of popularity. It is an iterative algorithm that divides the unlabeled data set into k different clusters, so that each data set belongs to only one group with similar attributes. This algorithm is based on the center of mass, each cluster has a center of mass associated with it [54], the purpose is to let the distance between the data points and the corresponding cluster and minimum [51]. In k-means, if there are n data observations, first select k observations as the initial clustering center, allocate the remaining data to their most similar clustering center according to the similarity with the clustering center, and then calculate the clustering center of the new class.And you keep doing this until the standard function starts to converge.K-means is still one of the most popular algorithms in data processing [6], and its higher computational efficiency may be one of the reasons [21].

2.3.3 Density-based spatial clustering of applications with noise (DBSCAN)

In 1996, Density-based spatial clustering of applications with noise(DBSCAN) was first proposed by Martin Ester, Hans-Peter Kriegel, J org Sander and Xiaowei Xu, which is the first dension-based clustering algorithm [15], aims to cluster data of arbitrary shape with noise in spatial and non-spatial high-dimensionaldatabases [43]. DBSCAN is different from the circular clustering region obtained by traditional clustering division. This method is divided according to the regional density. By using the high-density connectivity of classes, DBSCAN can quickly find classes of any shape in the spatial database with noise [17]. In DBSCAN algorithm, for each object in the class, the number of objects contained in the field with a given radius(epsilon) cannot be less than a given minimum number(MinPoints). Actually, the region with sufficient density is divided into a class, and the points of the same class are the largest set of densities connected points.Although DBSCAN is a relatively new

clustering algorithm compared with K-means and hierarchical clustering, it has been applied in various fields. DBSCAN is used to classify patients with similar diseases to support the development of the online medical Decisions [2], classifies the precise geographical location of 12 million buildings in Spain based on density [4], uses DBSCAN to classify and identify Internet traffic [52]. Therefore, this is a proven algorithm, and in 2014, it was awarded the SIGKDD Time Test Award [42].

2.3.4 Gaussian mixture model(GMM) and expectation-maximization (EM) algorithm

Gaussian mixture model has a long history, which can be traced back to 1886. It is the most popular mixture model clustering method [55]. The mixture model means that the overall distribution of observed data contains multiple sub-distributions, and the Gaussian mixture model means that these sub-distributions are Gaussian distributions. The principle of GMM is to assume that a group of observed data conforms to the Gaussian distribution, but the parameters of each Gaussian distribution are unknown. In this case, the EM algorithm is needed to estimate the parameters of GMM, and then the clustering is carried out according to the calculated parameters. EM algorithm is a method [13] proposed by Dempster, Laind and Rubin in 1977, which can estimate parameters from a non-holonomic data set. The Gaussian mixture model is widely used in various fields. In 1995, Douglas A. Reynolds and R.C. Rose proposed A paper on text independent speech recognition based on GMM, which was modeled from the perspective of the spectrum of the speaker and used to recognize the speech of multiple different speakers. Finally, he published a paper in 2000 [40]. Zivkovic's background extraction in 2004 was based on an efficient adaptive algorithm of GMM, with which parameters can be constantly updated, selecting the appropriate number of components per pixel [57]. Applications based on Gaussian mixture models do not stop there. As the fastest learning algorithm among mixture models, this is a big advantage in the increasingly important unsupervised learning environment.

3 Methodology

3.1 Principle of principal component analysis

Let's say that X is a sample set which

$$X = m * n$$

i.e.m samples and n characteristic dimensions. Write as

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

Step1: Initial data standardization

The goal of this step is to eliminate data errors so that each data makes the same contribution to the analysis. For PCA, this step is very important, because the difference in the subsequent steps will have a huge impact on the analysis results. If the range of the initial data differs greatly, the large range of data will dominate the small range of data, resulting in the deviation of analysis. We can standardize the data by using the following formula:

$$z = \frac{x - mean}{standard\ daviation}$$

Step2:Calculate the covariance matrix C

By calculating the covariance matrix, we can observe how the variables of the data set change with each other, and understand how they are related to each other. When variables are highly correlated, the information they reflect will be highly overlapping, and these correlations can be identified by calculating the covariance matrix.The following formula calculates C:

$$C = \frac{1}{m} X X^T$$

Then C is a matrix for n*n

$$C = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m}(x_1^{(i)})^2 & \frac{1}{m}\sum_{i=1}^{m}x_1^{(i)}x_2^{(i)} & \cdots & \frac{1}{m}\sum_{i=1}^{m}x_1^{(i)}x_n^{(i)} \\ \frac{1}{m}\sum_{i=1}^{m}x_2^{(i)}x_1^{(i)} & \frac{1}{m}\sum_{i=1}^{m}(x_2^{(i)})^2 & \cdots & \frac{1}{m}\sum_{i=1}^{m}x_2^{(i)}x_n^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m}\sum_{i=1}^{m}x_n^{(i)}x_1^{(i)} & \frac{1}{m}\sum_{i=1}^{m}x_n^{(i)}x_2^{(i)} & \cdots & \frac{1}{m}\sum_{i=1}^{m}(x_n^{(i)})^2 \end{bmatrix}$$

Because matrix C is a real symmetric matrix, the eigenvectors corresponding to different eigenvalues of C are orthogonal. This property will be used when identifying principal components.

Step3: Identification of principal component

The principal component is identified by calculating the eigenvalues of the covariance matrix C and the corresponding eigenvectors. In order for the dimensionless variables to be independent, in mathematics, to make the correlation between them zero, which is cov (x, y) = 0. According to the property that the covariance matrix C in 3.1.2 is a real matrix, n linearly independent non-zero eigenvectors and n non-zero eigenvalues can be easily obtained. The corresponding eigenvectors are arranged from the largest to the smallest according to the size of the eigenvalues.

The permutation matrix is expressed as:

$$V = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Then, the largest first K eigenvalues and corresponding eigenvectors are selected. Written as,

$$U = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ u_1 & u_2 & \cdots & u_k \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Finally, the original data set is projected onto the selected eigenvector, and the resulting data set is the k-dimensional data set after dimension reduction. i.e.

$$Z = XU = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ u_1 & u_2 & \cdots & u_k \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

X is the matrix of m*n, and U is the matrix of n*k, according to the calculation rules of the matrix, Z is a new data set of K dimensions, which is the result after dimension reduction.

3.2 Define a distance function to measure the similarity between observations

In the clustering algorithm, the attributes of observations are mainly represented by their relative distances in the characteristic space, so the concept of distance is significant in the clustering, which is used to reflect the anisotropy between different observations. The following are some of the most common distance calculations.Let's say we have a point i and point i'.

Euclidean Distance

It's the distance as a straight line between two observations.

$$d(x_i, x_i') = \sqrt{\sum_{i=1}^{n} (x_i - x_{i'})^2} = ||x_i - x_{i'}|| \tag{1}$$

Squared Euclidean Distance

This method is the same as the Euclidean method, except instead of taking the square root, this method is faster than the Euclidean method in terms of the size of the distance compared.

$$d(x_i, x_i') = \sum_{i=1}^{n} (x_i - x_{i'})^2 = ||x_i - x_{i'}||^2 \tag{2}$$

Cosine Distance

The idea is to measure the angle between two vectors.If the Angle is smaller, it means that the two vectors are closer in direction, and they will be grouped together in the clustering.

$$d(x_i, x_i') = \frac{\sum_{i=0}^{n-1} (x_i - x)}{\sum_{i=0}^{n-1} (x_i)^2 * \sum_{i=0}^{n-1} (x_{i'})^2} \tag{3}$$

3.3 Hierarchical clustering

3.3.1 How does hierarchical clustering work

The working principle of hierarchical clustering is introduced through the following steps:

Step 1: Suppose there are N data points and create each data point as a cluster, then we have N clusters.

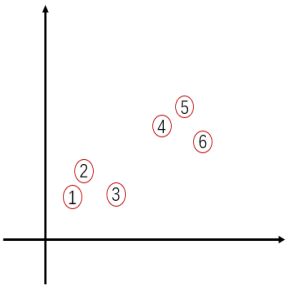


Figure 3.1: Step 1

Step 2: We need to find the shortest distance between any two data points, and once we find the shortest distance, we divide them into a group, into a cluster of multiple points, and then it becomes an $N-1$ cluster.The calculation method between two data points is described in section 3.2, and we only need to determine an appropriate method to apply. As shown in the figure below, points 4 and 5 form a cluster.Meanwhie, we can represent it as a tree-like structure called a dendrogram.

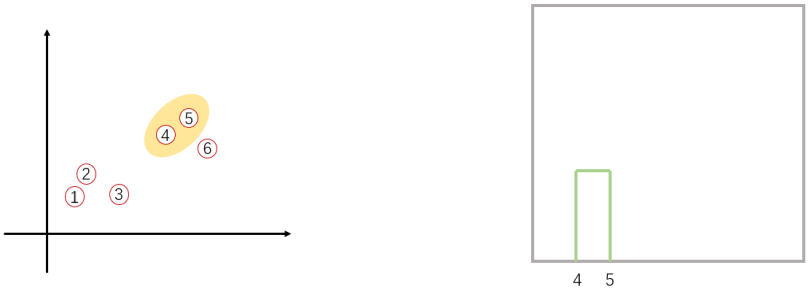


Figure 3.2: Step 2

Step 3: Repeat the work similar to Step 2, and find point 1 and point 2 to form another cluster with multiple points. At this point, there are $N-2$ clusters.The $x-y$ coordinates representation and the dendrogram are shown below:

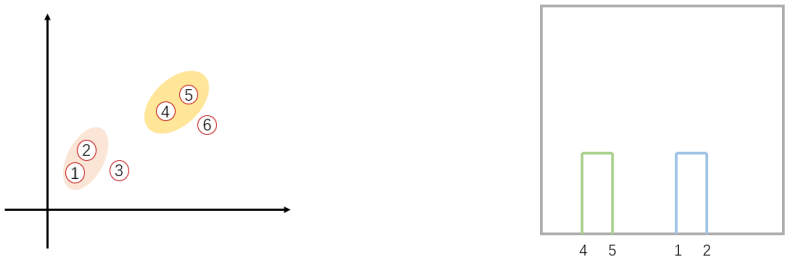


Figure 3.3: Step 3

Step 4: Repeat Step 3 until a cluster is formed.In this step, Point1, Point 2 and Point 3 form a cluster.

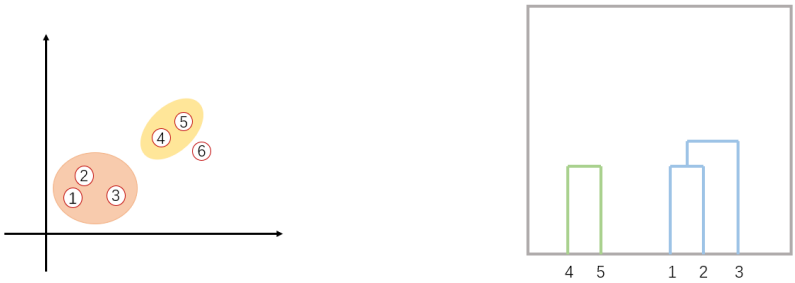
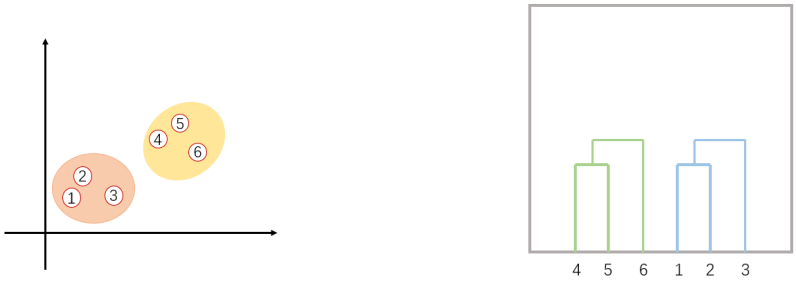
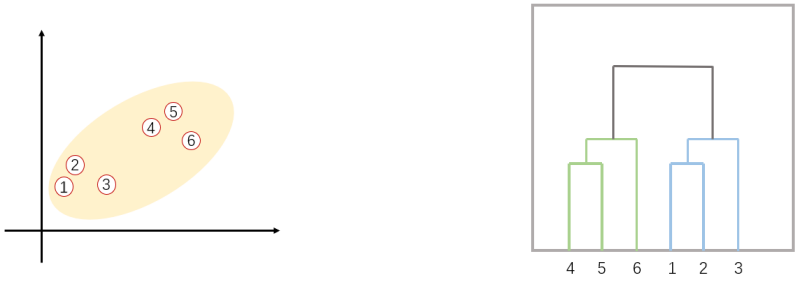


Figure 3.4: Step 4

This step shows we have two groups.



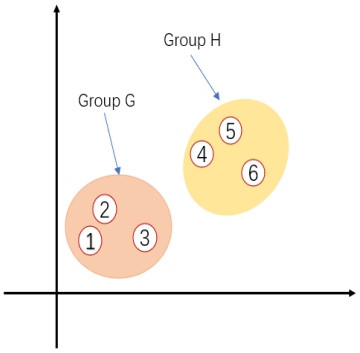
When we have only one cluster left, we are done and finally put everything together.



3.3.2 Measure for the distance between two clusters

According to the working steps of analysis hierarchy clustering in 3.2, we also need to determine the distance between clusters at multiple points, therefore, we must define a measure of dissimilarity between two clusters.

let G and H represent two such groups, and the distance between G and H is $d(G,H)$, for computing $d(G,H)$, we must observation dissimilarities $d_{ii'}$, and pair i in G, pair i' in H.

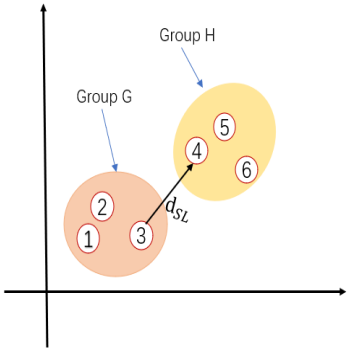


*Single linkage***(SL)**

The shortest distance between the closest points in a cluster.It could also be named the $nearest-neighbor$ technique.

$$d_{SL}(G,H) = min\ d_{ii'} \tag{4}$$

Refer to the figure below.

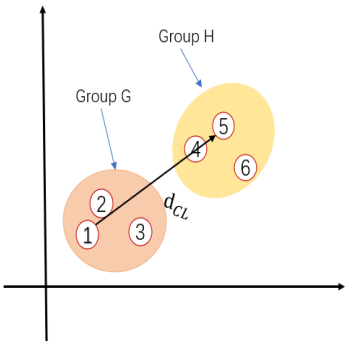


*Complete linkage***(CL)**

It's the distance between the farthest points of two different clusters.

$$d_{CL}(G,H) = max\ d_{ii'} \tag{5}$$

Refer to the figure below.



*Group average***(GA)**

It uses the average dissimilarity between the clusters

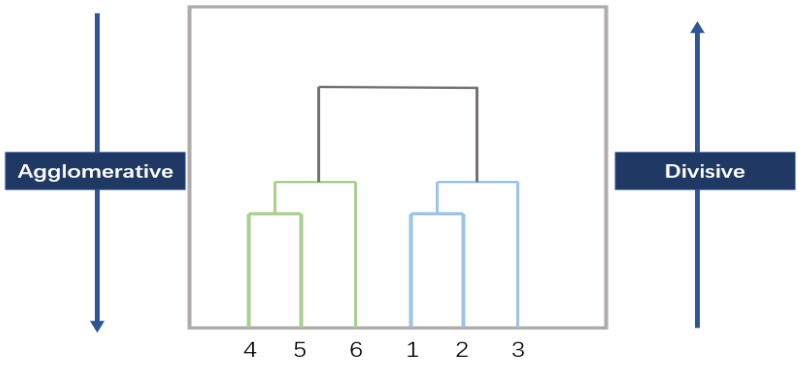
$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'} \tag{6}$$

Where N_H and N_G is the number of data point in each group.

3.3.3 The type of hierarchical clustering

Hierarchical clustering is split into agglomerative clustering and divisive clustering. Divisive clustering is top-down method, agglomerative clustering is bottom-up approach, that is, to

bring all the observations together.



3.4 The steps of k-means clustering

Step 1

Assume that there are n data points and k objects are selected from the data as the initial clustering center.

The selection of initial clustering center can refer to the result of hierarchical clustering, and select a point from each class as the initial clustering center of K-means. Or execute the algorithm for many times, whichever one has a more reasonable result.A more reasonable initial cluster center selection method can be referred to the article [26].

Step 2

The distance of each cluster object to the cluster center is calculated according to the distance calculation method in 3.2, and each data is divided into the nearest cluster center.

Step 3

Calculate the arithmetic mean of all points of each cluster in each dimension and act as the new cluster center.It is important to note that the new cluster center is not necessarily an actual data point.

Step 4

Repeat step2 and step3 until the clustering center no longer changes, that is, the iteration has converged or reached the maximum number of iterations.

3.5 Density-based spatial clustering of applications with noise

3.5.1 Basic conception

Eps(ϵ)

The neighborhood within the Eps of a given data point radius is called the Eps neighborhood of the object and is a parameter that needs to be set in advance.

Minpts

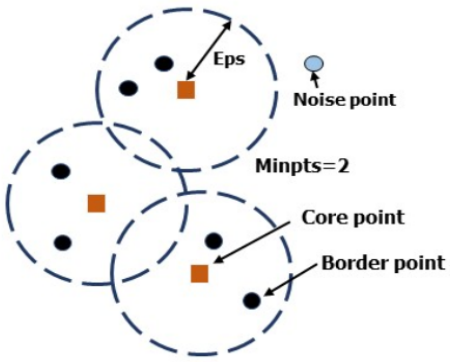
The minimum number of data points within the radius ϵ is a parameter that needs to be set in advance.

Classification of data points

Core point:If the neighborhood Eps of data point x_i contains at least Minpts samples,such that $N_\epsilon(x_i) \geq MinPts$,then data point x_i is called the core point.

Border point:If the number of samples contained in the ϵneighborhood of the data point x_i is less than Minpts, but it is in the neighborhood of other core points, the data point x_i is called the border point.

Noise point: It is neither a core point nor a border point.



directly density-rechable: If data points x_p and x_q satisfy $x_p \in N_\epsilon(x_q)$ and $|N_\epsilon(x_q)| \geq MinPts$,we say that x_p is directly debsity-reachable from x_q wth the parameter$\{Eps, MinPts\}$.
density-reachable: If you have a bunch of data points $x_1, \cdots, x_i, x_i+1, \cdots, x_n$(where $x_p = x_1$ and $x_q = x_n$),the data point $x_i + 1$ is directly density-rechable from x_i, then we say that data point x_p is density-reachable fromdata point x_q.
density-connected: If exist a data point x_0 such that x_p and x_q are density-reachable from data point x_0, then data point x_p and x_q is density-connected with the parameter $\{Eps, MinPts\}$.

3.5.2 Densitybased spatial clustering of applications with noise clustering process

Suppose there is a data set D, mark all the data points as "unvisited", select an unvisited data point x_p randomly, mark x_p as "visited", and then check whether the ϵ-neighborhood of x_p contains Minpts data points. If not, then x_p is marked as noise point. Otherwise, create a new cluster C for x_p and put all data points in the ϵ-neighborhood of x_p into a set S, then iterate over each data point in S and add data points in S that do not belong to other clusters to C. During this process, the corresponding data point x_q marked "unvisited" in S is marked as "visited", and then its ϵ-neighborhood is checked. If the ϵ-neighborhood of x_q contains at least Minpts data points, then all the data points in the ϵ-neighborhood of x_q are added to cluster C. Access x_q other neighborhood data points in turn, and so on. The cluster grows until C cannot expand, that is, the set N is empty. Then output cluster C.
In order to find the next cluster, an data point is randomly selected from the remaining "unvisited" data points and the above clustering process is repeated until all the data points are marked as "visited".

3.6 Gaussian mixture model

3.6.1 Maximum likelihood estimation

Before we can understand the EM algorithm, we need to understand what the maximum likelihood estimation is. Maximum likelihood estimation is a method used to estimate parameters of the model.
Let the population X obey the distribution $P(X;\theta)$, θ is the parameter that needs to be es-

timated, $X_1, X_2, X_3, \cdots, X_n$ is a sample from the population of X,$x_1, x_2, x_3, \cdots, x_n$ is an observation of $X_1, X_2, X_3, \cdots, X_n$， then the joint distribution of the samples is

$$L(\theta) = L(x_1, x_2, x_3, \cdots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

x_i is the known sample measurement,$L(\theta)$ is the likelihood function of the parameter $theta$ with respect to the sample set.The likelihood function can be understood as, sample set X is a known fact, and then the size of the parameter θ is adjusted to maximize the probability of the occurrence of this sample.The maximum likelihood estimator of θ is denoted as:

$$\widehat{\theta} = argmaxl(\theta)$$

What we discuss here is only the parameters of one cluster. Cluster analysis needs to discuss the mixing of multiple clusters. The EM algorithm we will discuss next is used to solve this problem.

3.6.2 The concrete process and principle of Gaussian mixture model

Firstly, the **single-Gaussian mixture model (GSM)** is understood. When the multi-dimensional variable $X = (x_1, x_2, \cdots, x_n)$ obeys the gaussian distribution, its probability density function (PDF) can be expressed as:

$$N(x; u, \sum) = \frac{1}{\sqrt{2\pi|\sum|}} exp[-\frac{1}{2}(x-u)^T \sum^{-1}(x-u)] \tag{7}$$

Where, u represents the model expectation; $\sum$ represents the covariance matrix, and describes the degree of correlation between variables of each dimension.

Then we can represent the **Gaussion Mixture Model(GMM)** as

$$p(x) = \sum_{k=1}^{K} \pi_k N(x; u_k, \sum_k) \tag{8}$$

K needs to be determined in advance, just like K in k-means. π_k is the weight factor, and $\sum_{k=1}^{K} \pi_k = 1$. Any of these Gaussian distributions $N(x; u, \sum)$ called a component of this model. And every compent is a cluster center.

The clustering algorithm of GMM is to calculate the model parameters without knowing the sample category, and then use the trained model to test the classification of the sample. Step1 is one of the K components selected at random (the probability of being selected is

π_k). step2 is to put the sample into the newly selected component. Determined if it belongs to this cluster, and return to step 1 if it does not.

Then the key problem is how to calculate the model parameters $\{\pi_k, u_k, \sum_k\}$, which requires the EM algorithm we will discuss in the following.

3.6.3 Estimate the parameters of GMM by EM algorithm

The EM algorithm is divided into two steps. The first step(E-step) is to calculate the rough values of the parameters to be estimated, and the second step(M-step) is to maximize the likelihood function using the values of the first step. Therefore, the likelihood function of GMM should be solved first. If there are n data points. We write equation (3.8) as

$$P(x; \pi, u, \sum) = \sum_{k=1}^{K} \pi_k N(x; u_k, \sum_k) \tag{9}$$

Accord to the section 3.6.1, the likelihood function of GMM is

$$L(\pi, u, \sum) = \prod_{i=1}^{n} P(x_i; \pi, u, \sum)$$

Usually, the probability of a single point is very small, and the data will be smaller after multiplication, which is easy to cause floating-point number overflowing. Therefore, the logarithm is generally taken to be log-likelihood function:

$$\sum_{i=1}^{n} logP(x_i; \pi, u, \sum) \tag{10}$$

The each sample x_i belongs to which cluster z_k is unknown.Z is the implicit variable.

E-step: Assuming that the model parameters are known, calculate the implicit variable Z, and take the expectation of $z_1, z_2, \cdots$. In GMM, it is to find the probability of data points being generated by each component.

$$\gamma(i,k) = \alpha_k P(z_k; x_i; \pi, u, \sum) \tag{11}$$

where α is the frequency of data points in the training set belonging to cluster z_k.

$$\gamma(i,k) = \frac{\pi_k N(x_i; u_k, \sum_k)}{\sum_{j=1}^{K} \pi_j N(x_j; u_j, \sum_j)} \tag{12}$$

M-step: Calculate the model parameters by maximum likelihood method. $\gamma(i,k)$ in last step is "The probability that data point x_i is generated by component K". And we can know that

$$N_k = \sum_{i=1}^{n} \gamma(i,k)$$

$$u_k = \frac{1}{N_K}\sum_{i=1}^{n}\gamma(i,k)x_i$$

$$\sum_k = \frac{1}{N_K}\sum_{i=1}^{n}\gamma(i,k)(x_i - u_k)(x_i - u_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

Check whether the parameter converges or whether the log-likelihood function converges. If not, return to E-step.

3.7 Clustering evaluation and assessment

If the clustering performance is good, it means that the samples of the same class are as close to this as possible, and the samples of different classes are as different as possible. That is to say, the clustering results have a high ”intra-cluster similarity” and a low ”inter-cluster similarity”. Clustering Evaluation and Assessment can be divided into two categories，that is external evaluation and internal evaluation. External evaluation is compare the result wit a reference model. Internal evaluation considers clustering results directly without using any reference model. The following focuses on internal evaluation.

Silhouette coefficient

For a data point x_i, $a(x_i)$ is the average distance to all other sample points in the cluster, $b(x_i)$ is the minimum of the average distance from all sample points to other clusters.
Then the silhouette coefficient of data point x_i is

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}$$

So the range of values of $s(x_i)$ is $-1 \leq s(x_i) \leq 1$.
When $a(x_i) << b(x_i)$, $s(x_i)$ is close to 1, satisfying this condition means that the cluster is appropriate;
When $a(x_i) >> b(x_i)$, $s(x_i)$ is close to -1, satisfying this condition means that x_i into adjacent clusters is more appropriate.
When $a(x_i) \approx b(x_i)$, means that the data point is at the intersection of two clusters.
The average Silhouette value is:

$$\bar{s} = \frac{1}{n}\sum_{i=q}^{n}s(x_i)$$

when $\bar{s} > 0.5$, it shows that clustering is appropriate;
when $\bar{s} < 0.2$, it shows that there is no clustering feature in the data.

4 Illustration on the coffee data

Coffee is a widely used beverage all over the world, which has been proved to be beneficial to health. Some reports have characterized coffee varieties based on the utilization of principal component analysis and cluster analysis of metal content in coffee [34]. Some reports used principal component analysis and cluster analysis to analyze the relationship between 13 variables of aroma, flavor, taste and appearance of 18 kinds of soluble coffee [9].

4.1 Description of the Dataset

The coffee data which is the chemical composition of coffee samples collected from around the world is used in this project. The data was originally reported by Streuil [46]. The dataset contains 43 coffee samples and from 29 countries. The data set contains two varieties coffees which are Arabica and Robusta. A data frame with 43 observations and 14 columns, The firs columns is variety, 1 is Arabica and 2 is Robusta. The second columns is Country. The remaining 12 columns represent the 12 chemical properties, which are Water, Bean Weight, Extract Yield, ph Value, Free Acid, Mineral Content, Fat，Caffine，Trigonelline，Chlorogenic Acid, Neochlorogenic Acid, Isochlorogenic Acid.
43 coffee samples from 29 countries were collected in the original data set. I processed the 43 coffee samples into 29 samples by taking the average value, that is, one coffee sample for each country. Then I clustered the 29 coffee samples and analyzed the clustering results.

Table 4.1 and figure 4.1 shows that caffine, isochlorogenic acid, free acid, neochlorogenic acid, mineral content and chlorogenic acid are positively correlated with each other, and these 6 chemical properties are negatively correlated with bean weight, fat and trigonelline. Bean weight have a positive correlation to trigonelline, fat and trigonelline, but it is negatively correlated the other 8 chemical properties. In addition, there is no much correlation between water and free acid, mineral content, extract yield, ph value, bean weight, trigonelline.

	Water	BeanWeight	ExtractYield	phValue	FreeAcid	MineralContent	Fat	Caffine	Trigonelline	ChlorogenicAcid	NeochlorogenicAcid	IsochlorogenicAcid
Water	1.000	0.110	-0.030	-0.070	-0.040	-0.040	0.280	-0.310	0.140	0.160	-0.230	-0.390
BeanWeight	0.110	1.000	-0.030	-0.250	-0.410	-0.220	0.440	-0.410	0.370	-0.190	-0.380	-0.490
ExtractYield	-0.030	-0.030	1.000	0.300	-0.090	0.290	0.160	-0.230	0.370	0.310	-0.100	0.070
phValue	-0.070	-0.250	0.300	1.000	-0.380	-0.120	0.010	0.190	-0.120	0.090	0.030	0.310
FreeAcid	-0.040	-0.410	-0.090	-0.380	1.000	0.550	-0.700	0.520	-0.440	0.520	0.600	0.320
MineralContent	-0.040	-0.220	0.290	-0.120	0.550	1.000	-0.340	0.240	-0.210	0.590	0.460	0.200
Fat	0.280	0.440	0.160	0.010	-0.700	-0.340	1.000	-0.840	0.710	-0.390	-0.800	-0.630
Caffine	-0.310	-0.410	-0.230	0.190	0.520	0.240	-0.840	1.000	-0.730	0.370	0.600	0.820
Trigonelline	0.140	0.370	0.370	-0.120	-0.440	-0.210	0.710	-0.730	1.000	-0.190	-0.570	-0.590
ChlorogenicAcid	0.160	-0.190	0.310	0.090	0.520	0.590	-0.390	0.370	-0.190	1.000	0.400	0.370
NeochlorogenicAcid	-0.230	-0.380	-0.100	0.030	0.600	0.460	-0.800	0.600	-0.570	0.400	1.000	0.350
IsochlorogenicAcid	-0.390	-0.490	0.070	0.310	0.320	0.200	-0.630	0.820	-0.590	0.370	0.350	1.000

Table 4.1: Correlation of all variables

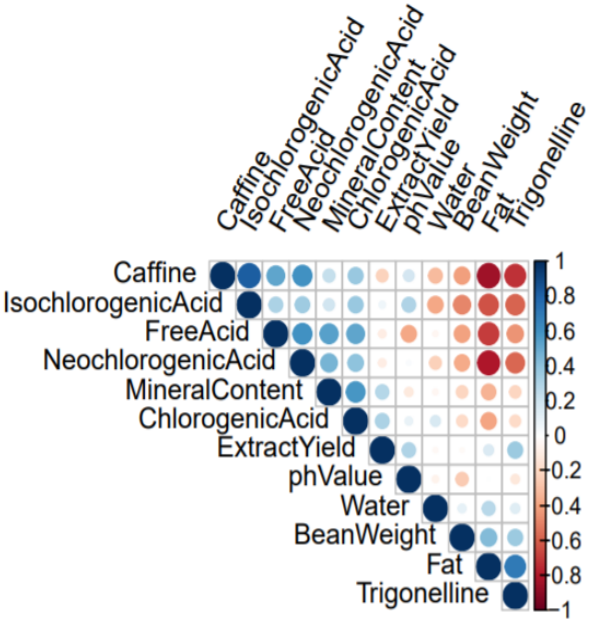


Figure 4.1: Correlation of all varianles

4.2 The coffee data was pretreated by principal component analysis for processing

Before the clustering analysis of the data set, we first performed dimensionality reduction preprocessing on the data set through PCA.This dataset contains 12 chemical properties in coffee beans, and the PCA will find a linear combination of 12 variables to account for most of the variation in the dataset.Using prcom function in R, the output results are shown in Table 4.2. The first six principal components explain about 90% of the data set changes in total, so the first six principal components are selected for further analysis.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	2.2296	1.3433	1.2938	1.0037	0.8527	0.7747	0.6477	0.5634	0.4376	0.4057	0.2711	0.2212
Proportion of Variance	0.4143	0.1504	0.1395	0.0839	0.0606	0.0500	0.0350	0.0265	0.0160	0.0137	0.0061	0.0041
Cumulative Proportion	0.4143	0.5646	0.7042	0.7881	0.8487	0.8987	0.9337	0.9601	0.9761	0.9898	0.9959	1.0000

Table 4.2: Importance of components

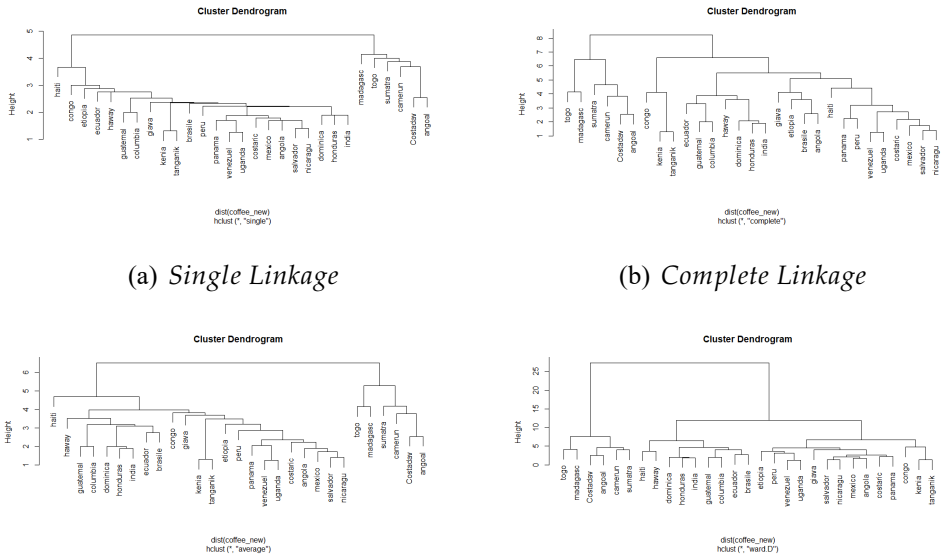
	PC1	PC2	PC3	PC4	PC5	PC6
Water	0.13	0.29	0.14	-0.82	0.04	-0.04
BeanWeight	0.26	0.06	0.19	0.21	0.86	0.07
ExtractYield	0.04	0.41	-0.54	0.24	-0.04	-0.03
phValue	-0.04	-0.20	-0.64	-0.30	0.10	0.38
FreeAcid	-0.33	0.28	0.30	0.04	-0.21	-0.19
MineralContent	-0.23	0.50	-0.01	0.17	-0.00	0.19
Fat	0.41	0.07	-0.11	-0.04	-0.07	-0.09
Caffine	-0.40	-0.22	-0.03	-0.04	0.23	-0.20
Trigonelline	0.34	0.25	-0.11	0.23	-0.16	-0.13
ChlorogenicAcid	-0.24	0.47	-0.16	-0.19	0.32	-0.20
NeochlorogenicAcid	-0.36	0.06	0.11	0.07	0.01	0.64
IsochlorogenicAcid	-0.34	-0.20	-0.29	0.04	0.11	-0.51

Figure 4.2: Weighted composition

In Figure 4.2, according to the correlation coefficient of weighted components, free acid, fat, caffine, trigonelline, chlorogenic acid and isochlorogenic acid were selected to form the principal component 1. Principal component 2 consisted of mineral content and chlorgenic acid, and principal component 3 contained extract yield and pH value. Water in principal component 4 constitutes the water content factor. Principal component 5 is the bean weight factor, and principal component 6 is neochlorogenic acid.

4.3 Hierarchical clustering

For this data set we will mainly discuss agglomerative procedures for hierarchical clustering. In section 3.3.2 we discussed several methods for calculating the distance between two clusters. For this data set, we tried using four different linkage measurements built into the "hclust" function in R(Figure 4.3).



(a) *Single Linkage* (b) *Complete Linkage*

(c) *Average Linkage* (d) ward.D

Figure 4.3: Hierarchical clustering with diferent linkage measure

As can be seen from Figure 4.3(a), it is unreasonable to use "singlelinkage" to calculate the distance between clusters, which makes no sense for dealing with practical problems. The results produced by "complete" and "ward.d" are similar. Obviously, four different methods of calculating linkage produce different results.

Here, we only show the clustering result of "ward.D" linkage measure. We could cut the tree into 5 clusters. In figure 4.4, the clustering tree already shows clearly the classification for each countries, then fucus on the chemical properties of coffee in each cluster in table 4.3. The trigonelline and extract yield content of the coffee in the first cluster1 is relatively high, and the coffee in the cluster 2 mainly contains water, bean weight. The content of the coffee in the cluster 3 is significantly different from that of other clusters in terms of free acid, mineral content, caffine, chlorogenic acid and isochlorogenic acid, while the coffee in cluster 4 mainly contains neochlorogenic acid. Cluster5 is different among other coffee in

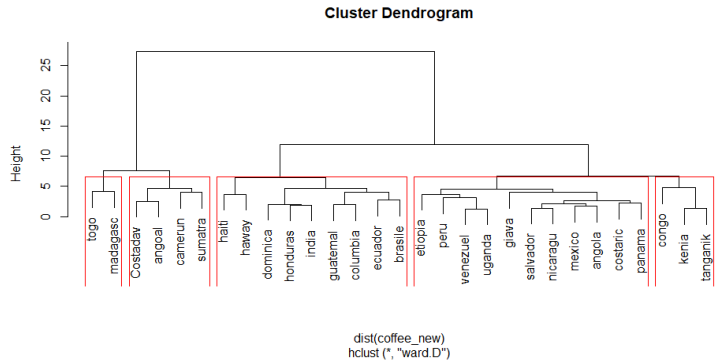


Figure 4.4: Hierarchical clustering with “ward.D” linkage measure

the fat.

	Water	BeanWeight	ExtractYield	phValue	FreeAcid	MineralContent	Fat	Caffine	Trigonelline	ChlorogenicAcid	NeochlorogenicAcid	IsochlorogenicAcid
1	-0.29	0.07	0.49	0.46	-0.71	-0.55	0.53	-0.38	0.64	-0.49	-0.51	-0.07
2	0.68	0.55	0.12	-0.77	0.24	0.50	0.28	-0.65	0.39	0.45	0.01	-0.73
3	-0.26	-1.42	0.22	0.61	1.59	1.10	-1.65	1.88	-1.23	1.51	1.15	1.88
4	-1.56	-0.30	-1.50	-0.95	1.15	-0.27	-2.08	1.44	-1.87	-0.65	1.82	0.38
5	0.43	0.20	-1.45	0.46	-0.99	-0.80	0.81	-0.14	-0.60	-1.14	-0.91	-0.30

Table 4.3: Clusters means of 5 clusters (hierarchical clustering)

As shown in Figure 4.4, all the countries in Cluster 1 belong to African countries. In Cluster 2, Sumatra comes from Asia, and all the other countries belong to Africa. However, the four countries are very similar in geographical position and altitude. In Cluster 3, all the country beans belong to America except India in Asia. Most of the countries in Cluster 4 are in South and Central America, with Etiopia, Angola and Uganda in Africa. The countries or regions in Cluster 5 are all located in eastern Africa. From the perspective of countries in each cluster, most of the countries producing coffee beans are located in America and Africa, and the countries in Cluster 3 and Cluster 4 are basically located in America. As can be seen from Table 4.3, the pH value of Cluster3 and 4 is Mineral content and chlorogenic acid.

To visualize the hierarchical clustering, we project the data onto the first four principal components. On the projection of first and second principal components(Figure4.5), we can see that cluster 3 and cluster 2 are separated in the dimention of PC1, and cluster1 and cluster2 are mixed in this projection. In addition, the dimention of PC2 separates cluster 2 and cluster 4. Cluster 4, 5 have lower values on PC2 than cluster 2, 3,and cluster 2, 3 have almost the same value. The right side of figure 4.5 show the projection of third and fourth principal components, five clusters are mixed together and not well separated.

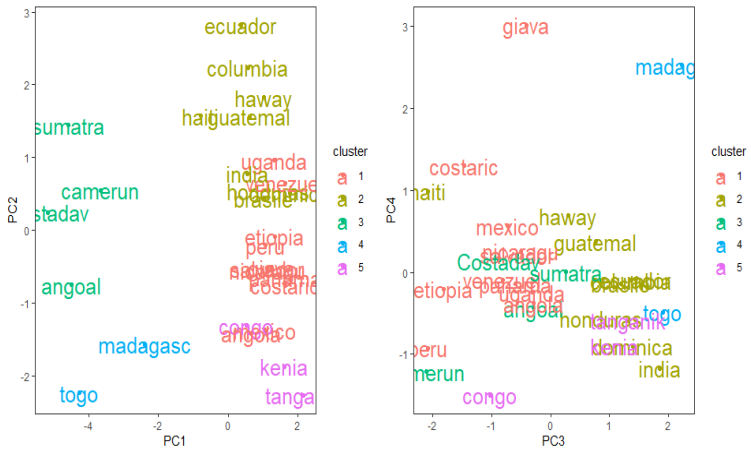


Figure 4.5: Projection the clusters onto the first four principal components

4.4 k-means clustering

Before using k-means algorithm for clustering analysis, we must first determine the number of clusters K, and the selection of K will directly affect the quality of clustering results. In Section 3.7, we mentioned a method for evaluating clustering performance - Silhouette Coefficient， we will use the package *fpc* to find the most appropriate number of clusters. In figure 4.8, the curve changes very little after k=4.In R, run the function *kmeansruns*, and the out put of average silhouette width for 1 to 10 clusters is shown in figure4.9. When k=2, average silhouette is maximized. Therefore, we pick 3 clusters.

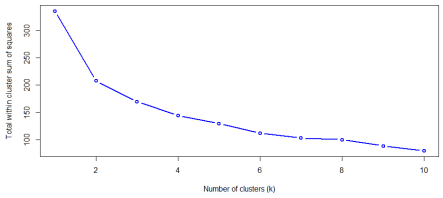


Figure 4.6: Total within cluster sum of squares

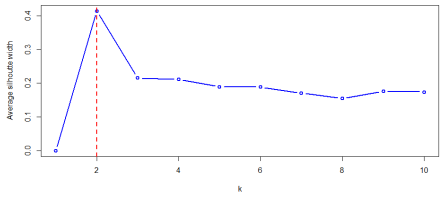


Figure 4.7: Choosing the number of cluster

The size for the 2 clusters are 6, 23. The clustering result is different from hierarchical clustering. The coffee samples from Costadav, togo, camerun, angoal, madagasc, sumatra in the cluster 1, from coffee data we know that the variety of these 6 coffee samples is Robusta. The other 23 coffee samples in cluster 2 are Arabica. We have to say that this clustering result is not so good, because we can get such a result from the original data set.

In Ttable 4.4, we can get the difference in chemical composition between Robusta and Arabic from this result， cluster 1 contains Robusta coffee with higher ph vaule, free acid, mineral content, caffine, chlorogenic acid and isochlorohgenic acid. The Arabic in cluster 2 with high water, bean weight, fat and trigonelline. Bitter taste is the biggest characteristic of coffee flavor, which is caused by caffeine. From Table 4.4, the caffeine content of coffee in cluster 1 is almost twice that of coffee in cluster 2, so we can roughly conclude that Robusta tastes more bitter than Arabica. In addition, the fat content also affects the taste of coffee, the coffee in cluster 2 contains almost 1.5 times as much fat as that in cluster 1.

clusters	Water	BeanWeight	ExtractYield	phValue	FreeAcid	MineralContent	Fat	Caffine	Trigonelline	ChlorogenicAcid	NeochlorogenicAcid	IsochlorogenicAcid
1	-0.70	-1.05	-0.35	0.09	1.44	0.65	-1.79	1.73	-1.45	0.79	1.38	1.38
2	0.18	0.27	0.09	-0.02	-0.38	-0.17	0.47	-0.45	0.38	-0.21	-0.36	-0.36

Table 4.4: Clusters means of 2 clusters (k-means clustering)

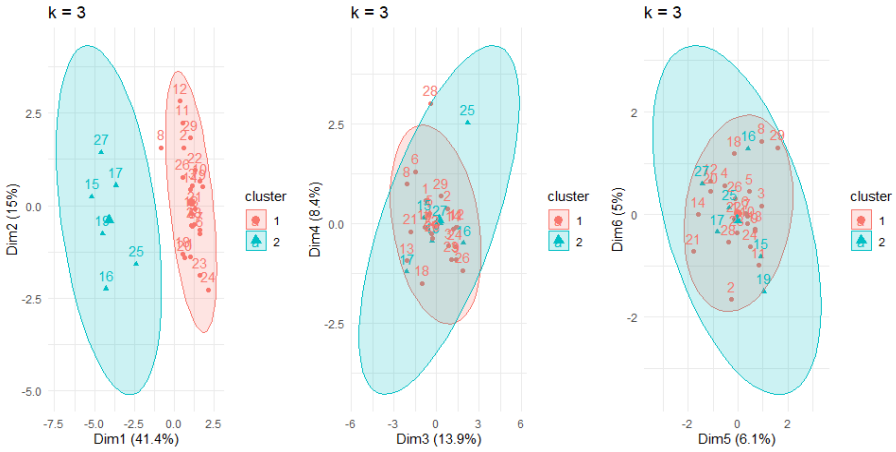


Figure 4.8: Visualization of the K-means clustering in the projections of principal components

Figure 4.8 shows a good separation of PC2 between cluster 1 and cluster 2. Then, on the projections of the third and fourth principal components, and on the projections of the fifth and sixth principal components, there is almost no separation in the clustering

4.5 Gaussian mixture model

This section we will use ”gaussian misture model” to clustering, section 3.6 has expounded the clustering principle of Gaussian mixture model, in this section we will use package

Mclust in R to implement gussian misture model. Bayesian Information Criterion(BIC) [50] will be used to select the best fitting model in package *Mclust*.
Figure 4.9 shows that BIC is maximized with VEI model with 3 components.

The size for each cluster is 13, 9, 7. The clustering result of Gaussian mixture model is

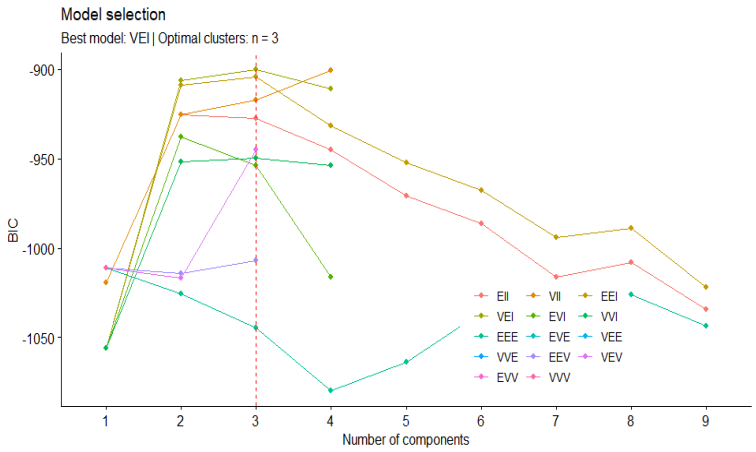


Figure 4.9: Model selection

actually very similar to the clustering result of k-means， components in the cluster 3(Table 4.5) are identical to one of the cluster from k-means. In addition, the clustering result is very similar to the hierarchical clustering. Member in cluster 1 and cluster 2(Figure 4.6) are identical to the cluster 3 in gaussian mixed model. The distribution of countries in each cluster is not uniform, so we cannot say that the clustering results are based on the geographical location of countries. It can be seen from Table 4.6 that cluster 1 has a high ph value and a high

cluster1	mexico,salvador, nicaragu, costaric, panama ,venezuel, peru, congo, angola, etiopia, kenia, tanganik, giava
cluster2	guatemal, honduras, dominica, columbia, ecuador, brasile, uganda, india, haway
cluster3	haiti, Costadav, togo, camerun, angoal, madagasc, sumatra

Table 4.5: The 3 clusters using gauission mised model

content of fat.The contents of water, bean weight and trigonelline are prominent in cluster 2. The contents of free acid, caffine, chlorogenic acid, neochlorogenic acid and isochlorogenic acid were high in cluster 3. The contents of the four main acids in coffee beans were more prominent in cluster 3. The existence of chlorogenic acid can optimize the physiological regulation of coffee spiders, promote the growth of plants and the formation of rhizomes. So

Means	cluster 1	cluster 2	cluster 3
Water	-0.19659469	0.8455212	-0.70148611
BeanWeight	0.08884548	0.5307360	-0.83898450
ExtractYield	0.02161114	0.0379069	-0.08859573
phValue	0.49399117	-0.9243169	0.24268028
FreeAcid	-0.81175071	0.3475283	1.08458604
MineralContent	-0.63562716	0.3095971	0.80181368
Fat	0.59039457	0.3840483	-1.59526723
Caffine	-0.30812925	-0.6893810	1.45176173
Trigonelline	0.35551614	0.3833041	-1.15316001
ChlorogenicAcid	-0.66691372	0.3091200	0.86118243
NeochlorogenicAcid	-0.60050205	-0.1913425	1.37021475
IsochlorogenicAcid	-0.09138601	-0.8352161	1.22934103

Table 4.6: Cluster means for the 3 clusters

coffee plants in cluster 3 are more likely to grow than coffee plants in the other two clusters. In addition, by comparing the clustering results of K-means and Gaussian mixture model, GMM divides cluster 2 into two clusters, cluster 1 and 2 (Table 4.5). The chemical composition of cluster 1 and cluster 2 in GMM is basically like this. The differences were in the contents of water, ph value, free acid, mineral content and chlorogenic.

The visualization of the cluster is shown in Figure 4.10. The coffee samples in cluster 3 are separated from each other as well as from other clusters. Cluster 1 and cluster 2 were also separated from each other, but the coffee samples in cluster 1 and cluster 2 were not well separated on the first and second principal components. In the third and fourth principal component predictions, in the fifth and sixth principal component predictions, cluster 1, 2, 3 overlapped with each other.

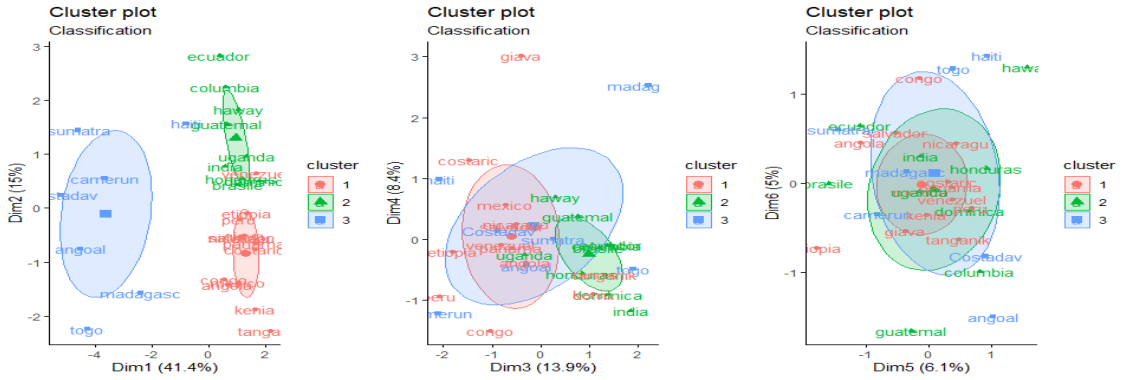


Figure 4.10: Visualization of the model-based clustering in the projections of principal components

5 Conclusion

In this report, we mainly studied several popular unsupervised learning methods, including principal component analysis (PCA), which is mainly used for data preprocessing, and four clustering methods hierarchical clustering, k-means, density based spatial clustering of application with noise (DBSCAN) and Gaussian mixture model (GMM). After understanding the development process and practical application of these methods, we learned their specific algorithm principles. Finally, we apply these methods to the data set. We select 43 coffee samples from 29 countries as the data set to find out the similarities between these coffee samples.

Through the dimensionality reduction process of principal component analysis, we select the first six principal components of the 12 principal components, which can explain more than 85% of the data set changes. The importance of components showed that PC1, PC2, PC3, PC4, PC5 and PC6 explained 41.43%, 15.04%, 13.95%, 8.39%, 6.06% and 5% of the data sets respectively.

In the next work, our goal is to use different clustering methods to cluster coffee datasets, to find out whether there are some similarities between these coffee growing countries, and to simply compare the results of different clustering methods. In hierarchical clustering, we get five clusters. The size of the cluster is not very balanced compared to the other two methods. According to the analysis of the geographical location of each cluster country, we find that

the geographical location of the countries or regions gathered together is not necessarily very similar, and the regions or countries with very similar geographical location may also be divided into different clusters. Specifically, gatemala, Columbia, Ecuador, Brasile and Peru, venezuel are located in South America, but they are divided into two clusters (cluster 3 and Cluster 4) in hierarchical clustering. The contents of free acid, fat, caffeine and trigonelline of the two clustering coffee samples are very similar, and the differences are reflected in the contents of extract yield, mineral content and chlorogenic acid. Through the projection of the first principal component and the second principal component, the separation of cluster 1, 2 and 5 is visualized. The countries or regions in the three clusters are all located in Africa. The coffee samples in cluster 1 contain more extract yield, fat and trigonelline. The difference between cluster 2 and cluster 3 lies in the content of water, bean weight, free acid, fat and caffeine.

In K-means clustering, the original data set is divided into two clusters. The coffee samples in the same cluster are from the same variety of coffee beans. 6 samples in cluster 1 were from robusta growing countries, and 23 samples in cluster 2 were from Arabica growing countries. On the projection of the first and second principal components, the separation of the two clusters is visualized. In cluster 1, Robusta coffee samples contained more free acid, mineral content, cafine, chlorogenic acid, neochlorogenic acid and isochlorogenic acid. High cafine content would lead to more bitter coffee flavor, which explains why robusta tastes more bitter than Arabica.

Gaussian mixture model separates guatemal, Honduras, Dominica, Dominica, colimbia, Ecuador, Brasile, Uganda, India, Haway from K-means clustering. The results show that the cluster size is more balanced than the other two methods. Most of the countries in cluster 1 are from America, most of the countries in cluster 2 are from Africa and South America, and most of the countries in cluster 3 are from Africa. Similar to hierarchical clustering, countries with similar geographical locations tend to cluster in the same cluster. The three clusters can be well separated on the projection of the first principal component and the second principal component.

In general, the results of k-means clustering are not very ideal, and the number of the two

clusters is not balanced. From the results of hierarchical clustering and Gaussian mixture model, we can say that most coffee samples from countries with similar geographical location will be clustered in one cluster, but some similar countries will be clustered in different clusters. There is little difference in altitude, climate and rainfall between regions or countries with similar geographical location, so the content of chemical components in coffee is relatively similar, which is divided into one category in our project. If we want to explain that coffee samples from countries with very close geographical location are divided into different clusters, we need to further understand the specific climatic conditions, even man-made conditions in these countries or regions, and analyze the specific reasons for their differences in some chemical components.

In this report, we studied the chemical factors that affect the quality of coffee beans. As consumers, we often contact the processed coffee beans. In the future, we can collect the coffee beans after grinding, baking and other operations to study whether the taste of coffee is related to the processing process. The method of unsupervised learning is also developing. It is also my future goal to continue to learn new unsupervised learning methods applied to our research.

References

[1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews Computational Statistics*, 2(4):433–459, 2010.

[2] Ahmed Al-Shammari, Rui Zhou, Mehdi Naseriparsaa, and Chengfei Liu. An effective density-based clustering and dynamic maintenance framework for evolving medical data streams. *International journal of medical informatics*, 126:176–186, 2019.

[3] Ethem Alpaydin. *Machine learning: the new AI*. MIT press, 2016.

[4] Daniel Arribas-Bel, M-À Garcia-López, and Elisabet Viladecans-Marsal. Building (s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, page 103217, 2019.

[5] Susan Athey. The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press, 2018.

[6] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *arXiv preprint arXiv:1203.6402*, 2012.

[7] D. Bueso, M. Piles, and G. Camps-Valls. Nonlinear pca for spatio-temporal analysis of earth observation data. *IEEE Transactions on Geoscience & Remote Sensing*, 2020.

[8] Ruichu Cai, Zhenjie Zhang, Anthony K.H. Tung, Chenyun Dai, and Zhifeng Hao. A general framework of hierarchical clustering and its applications. *Information Sciences*, 272:29 – 48, 2014.

[9] Amalia Mirta Calvino, MARÍA CLARA ZAMORA, and MARÍA INÉS SARCHI. Principal components and cluster analysis for descriptive sensory assessment of instant coffee. *Journal of sensory studies*, 11(3):191–210, 1996.

[10] Mbpd Camargo. The impact of climatic variability and climate change on arabic coffee crop in brazil impacto da variabilidade e da mudança climática na produção de café arábica no brasil. *Bragantia*, 2010.

[11] AD Castelnuovo. Consumption of cocoa, tea and coffee and risk of cardiovascular disease. *European Journal of Internal Medicine*, 23(1):15–25, 2012.

[12] Sanjoy Dasgupta. Performance guarantees for hierarchical clustering. In *International Conference on Computational Learning Theory*, pages 351–363. Springer, 2002.

[13] A. P. Dempster. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

[14] Thomas G Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[15] M. Ester, H. P. Kriegel, Jrg Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI Press*, 1996.

[16] Adriana Farah and Thiago Ferreira dos Santos. Chapter 1 - the coffee plant and beans: An introduction. In Victor R. Preedy, editor, *Coffee in Health and Disease Prevention*, pages 5–10. Academic Press, San Diego, 2015.

[17] Junhao Gan and Yufei Tao. Dbscan revisited: mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 519–530, 2015.

[18] Peter A Henderson and Richard MH Seaby. *A practical handbook for multivariate methods*. Pisces Conservation Lymington, England, 2008.

[19] H. Hotellings. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933.

[20] Chung-Chian Hsu, Chin-Long Chen, and Yu-Wei Su. Hierarchical clustering of mixed data based on distance hierarchy. *Information Sciences*, 177(20):4474 – 4492, 2007.

[21] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

[22] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[23] K. L. Johnston, M. N. Clifford, and L. M. Morgan. Coffee acutely modifies gastrointestinal hormone secretion and glucose tolerance in humans: glycemic effects of chlorogenic acid and caffeine. *American Journal of Clinical Nutrition*, 78(4):728–733, 2003.

[24] I. T. Jolliffe. Principal component analysis. *Journal of Marketing Research*, 87(4):513, 2002.

[25] S. U. Kang and Y. S. Na. The analysis toward consumption state, import and export in the world coffee market -the case of korea, u.s.a., japan market-. *Culinary Science & Hospitality Research*, 10, 2004.

[26] Shehroz S Khan and Amir Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern recognition letters*, 25(11):1293–1302, 2004.

[27] A. Khumaidi. Data mining for predicting the amount of coffee production using crisp-dm method. *Jurnal Techno Nusa Mandiri*, 17(1):1–8, 2020.

[28] Catherine L., Ludlow, Gareth A., Cromie, Cecilia, Garmendia-Torres, Amy, Sirr, Michelle, and Hays. Independent origins of yeast associated with coffee and cacao fermentation. *Current Biology*, 2016.

[29] Rac Lamparelli, J. A. Johann, ÉRD Santos, Jcdm Esquerdo, and J. V. Rocha. Use of data mining and spectral profiles to differentiate condition after harvest of coffee plants. *Engenharia Agrícola*, 32(1):184–196, 2012.

[30] H. P. Landolt. Coffee, caffeine, and sleep: a systematic review of epidemiological studies and randomized controlled trials. *Sleep Medicine Reviews*, 31:70–78, 2016.

[31] Richard CT Lee. Clustering analysis and its applications. In *Advances in Information Systems Science*, pages 169–292. Springer, 1981.

[32] K. Loska and Danuta Wiechu?A. Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the rybnik reservoir. *Chemosphere*, 51(8):0–733, 2003.

[33] L. A. Lynn and J. F. Kissinger. Coronary precautions: should caffeine be restricted in patients after myocardial infarction? *Heart & Lung*, 21(4):365–371, 1992.

[34] Maıa J Martın, F Pablos, and AG González. Characterization of green coffee varieties according to their metal content. *Analytica chimica acta*, 358(2):177–183, 1998.

[35] Lu Mei and Zhao Xiang-Jun. A novel pso k-modes algorithm for clustering categorical data. In *Computer, Informatics, Cybernetics and Applications*, pages 1395–1402. Springer, 2012.

[36] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning ⋯, 2006.

[37] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The computer journal*, 26(4):354–359, 1983.

[38] D. E. Newby, J M M Neilson, D. R. Jarvie, and N. A. Boon. Caffeine restriction has no role in the management of patients with symptomatic idiopathic ventricular premature beats. *Heart*, 76(4):355–7, 1996.

[39] P. Pohl, E. Stelmach, M. Welna, and A. Szymczycha-Madeja. Determination of the elemental composition of coffee using instrumental methods. *Food Analytical Methods*, 6(2):598–613, 2013.

[40] Reynolds. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1⁻3):19–41, 2000.

[41] B. Scholkopf and A. Smola. Kernel principal component analysis. In *Springer, Berlin, Heidelberg*, 1997.

[42] E. Schubert, Jrg Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems*, 42(3):1–21, 2017.

[43] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[44] J. Shlens. A tutorial on principal component analysis. *International Journal of Remote Sensing*, 51(2), 2014.

[45] Gangadhar Shobha and Shanta Rangaswamy. Chapter 8 - machine learning. In Venkat N. Gudivada and C.R. Rao, editors, *Computational Analysis and Understanding*

of Natural Languages: Principles, Methods and Applications, volume 38 of *Handbook of Statistics*, pages 197 – 228. Elsevier, 2018.

[46] H Streuli. Der heutige stand der kaffeechemie. In *ASSIC, 6e. Colloque, Bogota*, volume 61, 1973.

[47] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[48] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[49] K. Watson and M. L. Achinelli. Context and contingency: the coffee crisis for conventional small-scale coffee farmers in brazil. *Geographical Journal*, 174, 2008.

[50] WEAKLIEM and L. D. A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397, 1999.

[51] Junjie Wu. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.

[52] Caihong Yang, Fei Wang, and Benxiong Huang. Internet traffic classification using dbscan. In *2009 WASE International Conference on Information Engineering*, volume 2, pages 163–166. IEEE, 2009.

[53] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics (Oxford, England)*, 2001 年 17 卷 9 期 (9):763–74 页, 2019.

[54] Sobia Zahra, Mustansar Ali Ghazanfar, Asra Khalid, Muhammad Awais Azam, Usman Naeem, and Adam Prugel-Bennett. Novel centroid selection approaches for kmeans-clustering based recommender systems. *Information sciences*, 320:156–189, 2015.

[55] Guojun Zhang, Pascal Poupart, and George Trimponias. Comparing em with gd in mixture models of two components. In *Uncertainty in Artificial Intelligence*, pages 164–174. PMLR, 2020.

[56] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524, 2002.

[57] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004.