Task 2 report: Classification of wine data by K-means

Abstract

This report developed a clustering model using unsupervised machine learning classification method, k-means, based on the dataset Red Wine Set from UCI machine learning repository. The developed model successfully classified the data into 3 groups, which is consistent with the self-contained class identifier.

Introduction

K-means is one of the unsupervised machine learning classification methods. It can be used to cluster data by classifying them into several groups (namely, clusters), based on the common features shared by observation within the same group.

The basic idea of k-means algorithm is to: (1) Choose k points randomly as the initial clustering centers. (2) Calsculate the distance from each data point to the initial k clustering centers, and classify the data point to cluster with the nearest cluster center. (3) For each cluster, calculate the centroid $a_j = \frac{1}{|c_i|}\sum_{x \in c_i} x$, namely, the the centroid of all the data points belonging to this cluster. (4) Repeat step(2) and step(3) until reaching a certain terminal condition (for example, number of iterations, minimum change of error, etc).

The data used is Wine Data Set that has 13 attributes as well as class identifier (1,2,3). The attributes are obtained from the chemical analysis of the wine components. Hence, the model generated from wine data might have realistic significance in classifying the unlabelled wine stock and learning the different characters of wine from different classes. In this research, I will perform k-means clustering on the wine data to classify them into an appropriate number of groups (3 in this research).

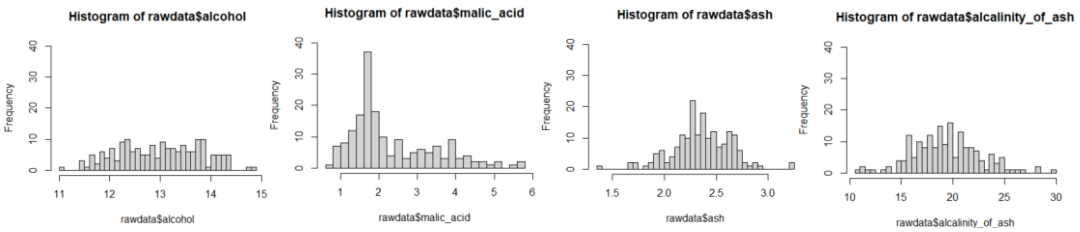
Methods and Results

(1) Data preparation

I loaded the data and named the predictors for further convenience. I deleted the first column ‘class identifier’ since an unsupervised machine learning algorithm does not need labels to train the model.

(2) Preliminary data exploring

First, I looked at the histograms of each attribute as shown in **Fig.1**. We can see these attributes show quite different distribution characters.



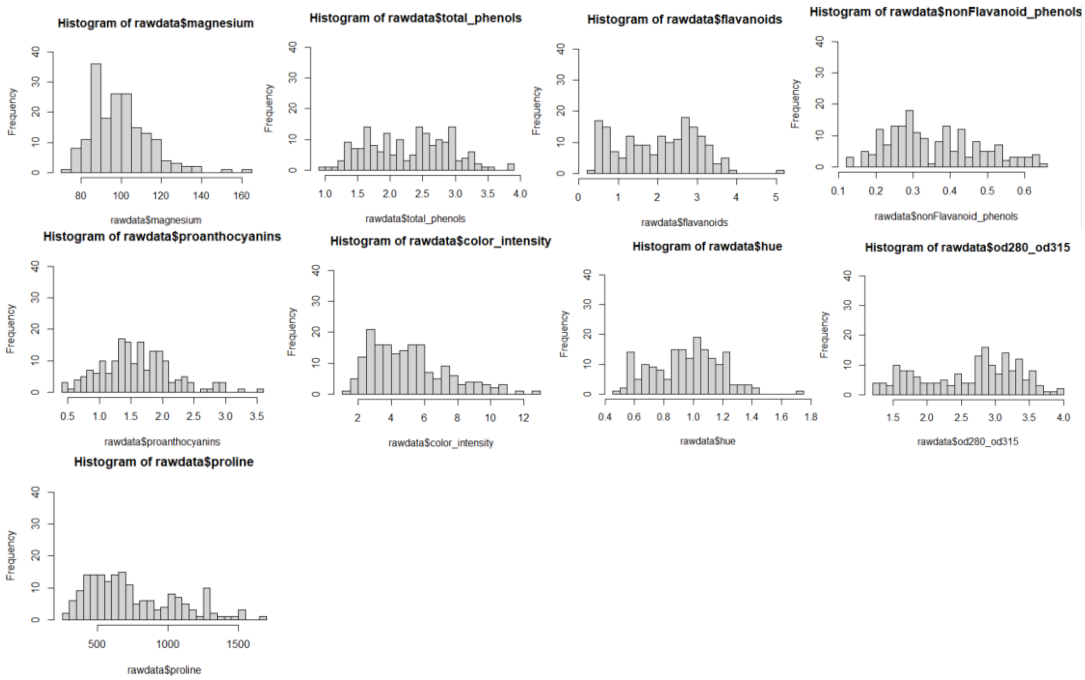


Fig.1 . Histograms of the 13 attributes.

Second, I explored the correlation between these attributes. As shown in **Fig.2**, We can see a significant linear correlation between attributes 'flavaniods' and 'total_phenols'. Hence, we fitted them into the linear regression model and confirmed the linear correlation is quite significant.

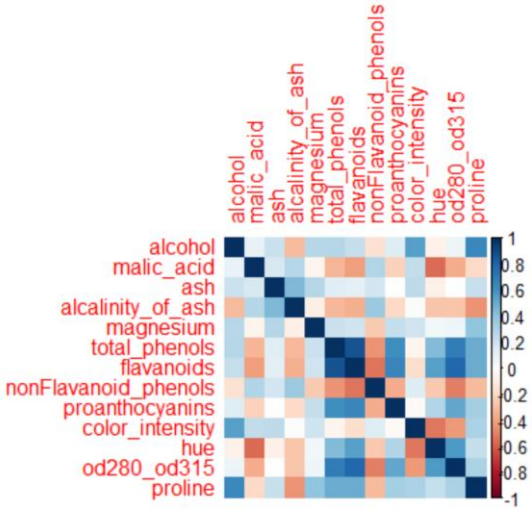


Fig.2. Heatmap representing the correlation between attributes.

(3) Data normalization

From the histograms shown in **Fig.3**, we can see great scale differences between attributes, such as 'alcohol' and 'nonFlavanoid_phenols'. Hence, we need to normalize them to make them on the same scale. As shown below, the left plot is the raw data, the right plot is normalized data.

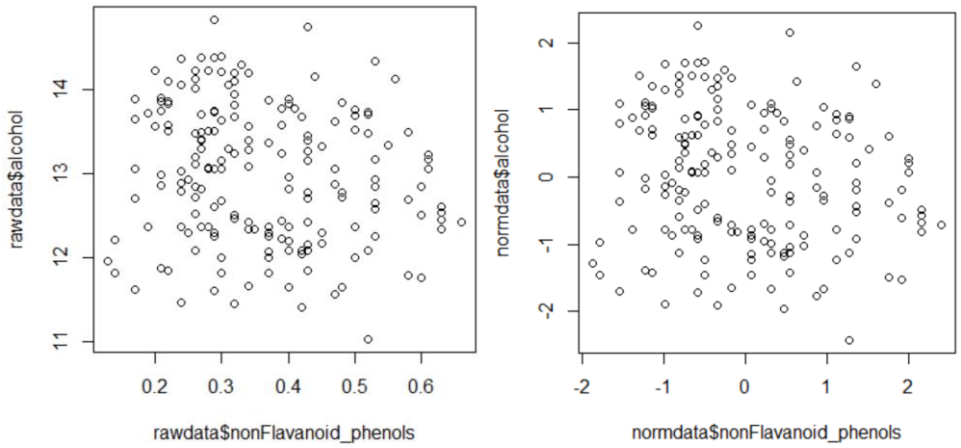


Fig.3 . Scatter plot of 'alcohol' against 'nonFlavanoid_phenols'. **Left.** Before normalization. **Right.** After normalization.

(4) Selection of k-value by Elbow Criterion Method

Here, I used Elbow Criterion Method to find the best k value. I ran the *kmeans* function in r with different k values from 1 to 10. As shown in **Fig.4**, I then used qqplot2 to plot the change of between-cluster sum of squares and total within-cluster sum of squares along with k from 1 to 10.

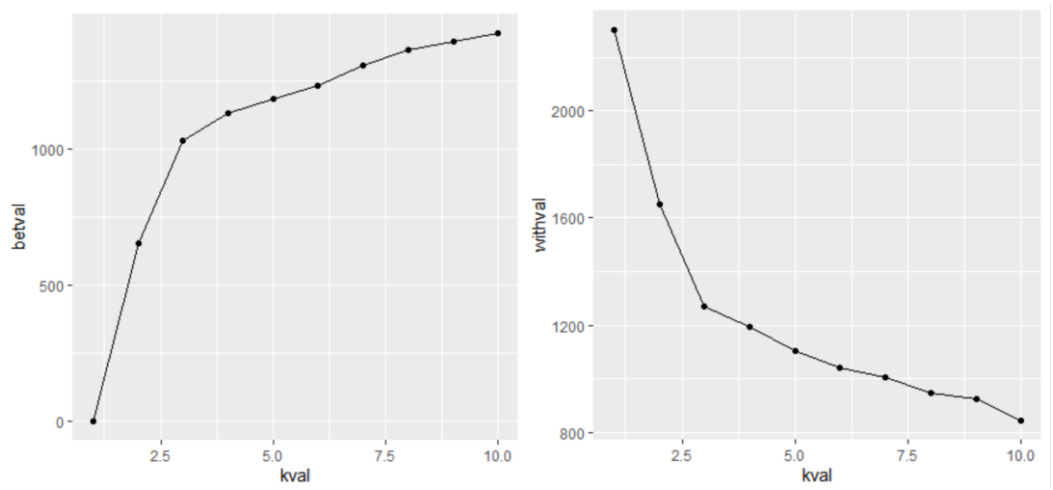


Fig.4. Curve graph showing the change of between-cluster sum of squares (left) and total within-cluster sum of squares (right) againt different k values.

We can see the elbow of the curves is when k equals to 3. After this point, both the increase of between-cluster sum of squares, and the decrease of total within-cluster sum of squares become much slower, indicate adding another cluster does not make the partition of the model much better. Hence, the optimal k value should be 3, which means 3 is the ideal cluster number of the data.

(5) Performing k-means clustering

I used k = 3 to perform k-means clustering. R output the clustering result as shown in **Fig.5.**

```
> clusterdata
K-means clustering with 3 clusters of sizes 51, 65, 62

Cluster means:
     alcohol malic_acid        ash alcalinity_of_ash   magnesium
1  0.1644436  0.8690954  0.1863726         0.5228924 -0.07526047
2 -0.9234669 -0.3929331 -0.4931257         0.1701220 -0.49032869
3  0.8328826 -0.3029551  0.3636801        -0.6084749  0.57596208
  total_phenols  flavanoids nonFlavanoid_phenols proanthocyanins
1   -0.97657548 -1.21182921           0.72402116     -0.77751312
2   -0.07576891  0.02075402          -0.03343924      0.05810161
3    0.88274724  0.97506900          -0.56050853      0.57865427
  color_intensity        hue od280_od315    proline
1       0.9388902 -1.1615122  -1.2887761 -0.4059428
2      -0.8993770  0.4605046   0.2700025 -0.7517257
3       0.1705823  0.4726504   0.7770551  1.1220202

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [35] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2
 [69] 2 2 2 2 2 3 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2
[103] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 3 2 2 2 2 2 2 2 2 1 1 1 1 1 1
[137] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[171] 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 326.3537 558.6971 385.6983
 (between_SS / total_SS =  44.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Fig.5. R output showing the clustering results

As shown in **Table.1**, I calculated the mean of each attribute in each cluster.

Table.1. Mean value of attributes in 3 clusters

Group.1	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonFlavanoid_phenols	proanthocyanins	color_intensity	hue	od280_od315	proline
1	12.25092	1.897385	2.231231	20.06308	92.73846	2.247692	2.0500000	0.3576923	1.624154	2.973077	1.0627077	2.803385	510.1692
2	13.13412	3.307255	2.417647	21.24118	98.66667	1.683922	0.8188235	0.4519608	1.145882	7.234706	0.6919608	1.696667	619.0588
3	13.67677	1.997903	2.466290	17.46290	107.96774	2.847581	3.0032258	0.2920968	1.922097	5.453548	1.0654839	3.163387	1100.2258

Then I did pairwise plot visualization of the clustering using ggpairs(), as shown in **Fig.6**.

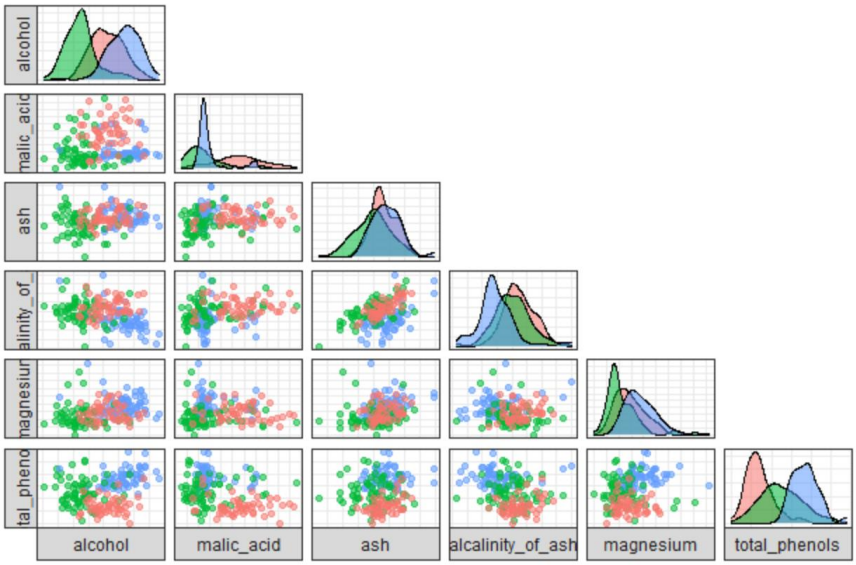


Fig.6. Pairwise plot visualization of the clustering result

I also used fviz_cluster() to visualize the clustering as shown in **Fig.7**.

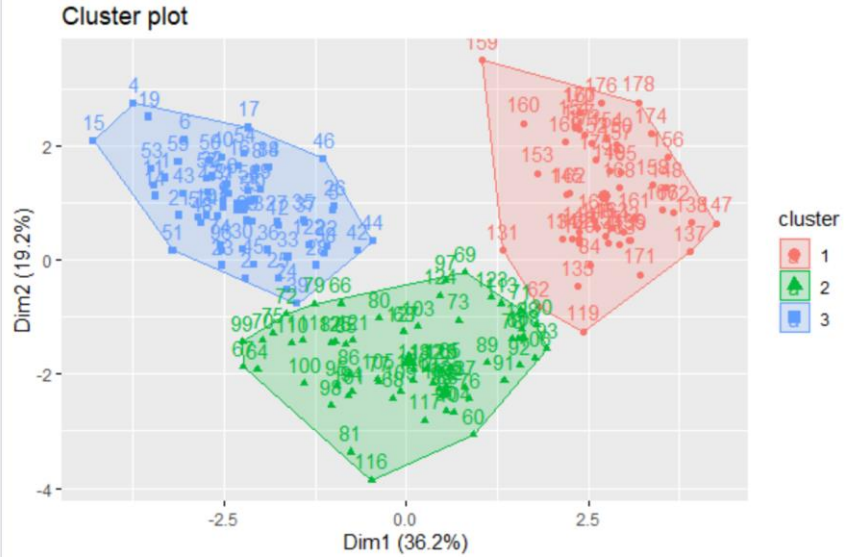


Fig.7. Clustering visualization of wine data

(6) Performance evaluation

From the class identifier (1~3) provided in raw data as shown in **Fig.8.**, and compare this classification with our clustering result. We found only 6 points are not consistent with the website-provided class identifier. Hence, we can evaluate the accuracy of the k-means clustering model as (178-6)/178 = 96.6%, which is a good performance.

```
> identifer <- rawdata[,1]
> identifer
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [35] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
 [69] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[103] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
[137] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[171] 3 3 3 3 3 3 3 3
```

Fig.8. Class identifier of wine data provided by the website.

Discussion

The reason to choose k-means to develop the classification model is: (1) Nonparametric classification methods are more flexible compared with parametric methods. K-means can ensure better scalability, especially when handling large data. (2) The implementation of K-means clustering is simple. (3) K-means methods have better performance when the distribution of the cluster is close to gaussian distribution. This can be observed in the clustering visualization results. Besides, I wish to get a clustering model that considers more about the overall differences between the multiple characters of the wine, instead of just considering the 3 cultivars.

The limitation of k-means includes (1) The clustering results are influenced by the initial K values. Along with the increase of k, the dependency also becomes higher. (2) The clustering results can be influenced by outliers since outliers can move the centroids outwards.

I explored the data by observing the distribution of each feature. The distributions are quite different between features. I also observed a strong correlation between two features, which might also influence the usefulness of this model.

Conclusion

In this report, I developed a clustering model using k-means. This model classified the wine data into 3 groups based on the differences in the features. The clustering result of the model has been proved quite consistent with the class identifier provided in the data with accuracy at 96.6%.

Appendix

Data preparation

```
> setwd("C:/Users/iefad/Desktop/course2021/bioinfosta/assignment/data-ass
ignment2")
> rawdata <- read.csv('wine_raw.csv',header = FALSE)
> rawdata <- rawdata[,-1]
> colnames(rawdata) = c('alcohol'
+                 ,'malic_acid'
+                 ,'ash'
+                 ,'alcalinity_of_ash'
+                 ,'magnesium'
+                 ,'total_phenols'
+                 ,'flavanoids'
+                 ,'nonFlavanoid_phenols'
+                 ,'proanthocyanins'
+                 ,'color_intensity'
+                 ,'hue'
+                 ,'od280_od315'
+                 ,'proline')
```

Preliminary data exploring

```
> hist(rawdata$alcohol,breaks = 30, ylim = c(0,40))
```

```
> corrplot(cor(rawdata),method = 'color')
> fit1 <- lm(rawdata$total_phenols ~ rawdata$flavanoids )
> anova(fit1)
Analysis of Variance Table

Response: rawdata$total_phenols
                   Df Sum Sq Mean Sq F value    Pr(>F)
rawdata$flavanoids   1 51.821  51.821  520.95 < 2.2e-16 ***
Residuals          176 17.508   0.099
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Normalization

```
> normdata <- scale(rawdata)
> normdata <- as.data.frame(normdata)
> plot(rawdata$alcohol ~ rawdata$nonFlavanoid_phenols)
> plot(normdata$alcohol ~ normdata$nonFlavanoid_phenols)
```

#Selection of k-value

```
> set.seed(128)
> for (i in 1:10) {
+   betweeness[i] <- kmeans(normdata, centers = i)$betweenss
+   withiess[i] <- kmeans(normdata, centers=i)$tot.withinss
+ }
>
> library(ggplot2)
> df1 <- data.frame(kval = c(1,2,3,4,5,6,7,8,9,10), betval = betweeness)
> df2 <- data.frame(kval = c(1,2,3,4,5,6,7,8,9,10), withval = withiess)
>
> ggplot(df1, aes(x= kval, y=betval, group=1)) +
+   geom_line()+
+   geom_point()
>
> ggplot(df2, aes(x= kval, y=withval, group=1)) +
+   geom_line()+
+   geom_point()
```

#perform clustering

```
> set.seed(129)
> clusterdata <- kmeans(normdata, centers = 3)
> lst <- list(clusterdata$cluster)
> aggregate(rawdata, by = lst, mean)
  Group.1  alcohol malic_acid      ash alcalinity_of_ash magnesium total_phenols flavanoids nonFlavanoid_phenols
1       1 13.13412   3.307255 2.417647          21.24118  98.66667      1.683922  0.8188235            0.4519608
2       2 12.25092   1.897385 2.231231          20.06308  92.73846      2.247692  2.0500000            0.3576923
3       3 13.67677   1.997903 2.466290          17.46290 107.96774      2.847581  3.0032258            0.2920968
  proanthocyanins color_intensity       hue od280_od315   proline
1        1.145882        7.234706 0.6919608    1.696667  619.0588
2        1.624154        2.973077 1.0627077    2.803385  510.1692
3        1.922097        5.453548 1.0654839    3.163387 1100.2258
```

```
ggpairs(cbind(rawdata, Cluster=as.factor(clusterdata$cluster)),
        columns=1:6, aes(colour=Cluster, alpha=0.5),
        upper=list(continuous="blank"),
        lower=list(continuous="points"),
        axisLabels="none", switch="both") +
  theme_bw()
```

```
> library(factoextra)
> fviz_cluster(clusterdata, data = normdata)
```