UTD 2016 Spring

CS6301.5U1 Advanced Computational Methods for Data Science

Assignment 1 – Hierarchical Clustering Mining

Name: Yu Zhang, Mingyue Sun

NetID: yxz141631, mxs151730

Part 1. K-Means Clustering

* 1. Accuracy of K-Means Clustering

According to the original data, we can easily find out that there are 3 categories of the wheat seeds. So, we set 3 as the number of centers. By comparing results with original label (V8) we can get the accuracy.





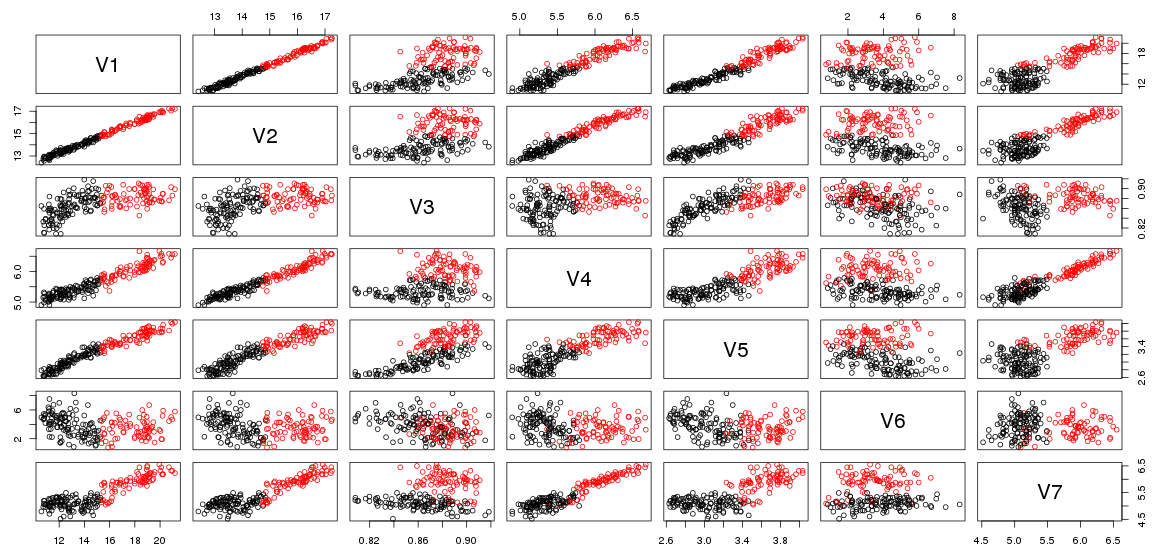


From the result we can indicates that only 19 instances are mis-clustered in this test. The total number of records is 199. Thus, the accuracy in this case is 90.45226%. One way to estimate how well the data is clustered is to estimate the sum of distances from points to centers in each clusters, i.e., the total within sum of (distance) square. In this case, we find out that the total within SS is 937.3457. We will compare this value in other cases in the following analysis.

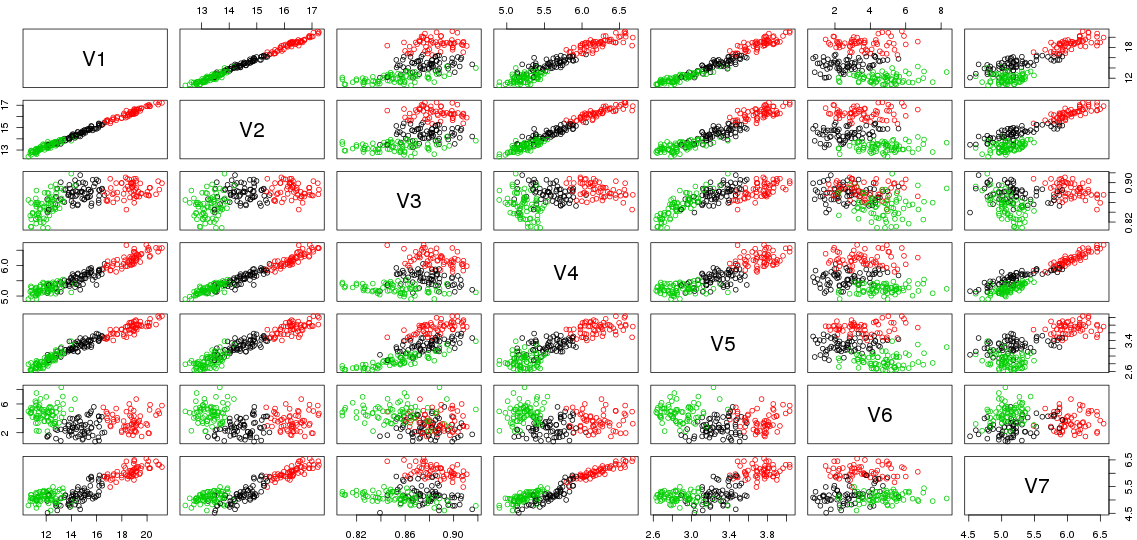
To avoid the negative influence of potential local minima issue brought by k-means algorithm. We ran k-means with 3 centers 5 times, but fortunately we got the same results (except the sequence of label names 1,2,3 or 3,2,1). Given all above, we can say that k-means clustering is relatively suitable for cluster this dataset.

* 1. Influence of Number of Centers

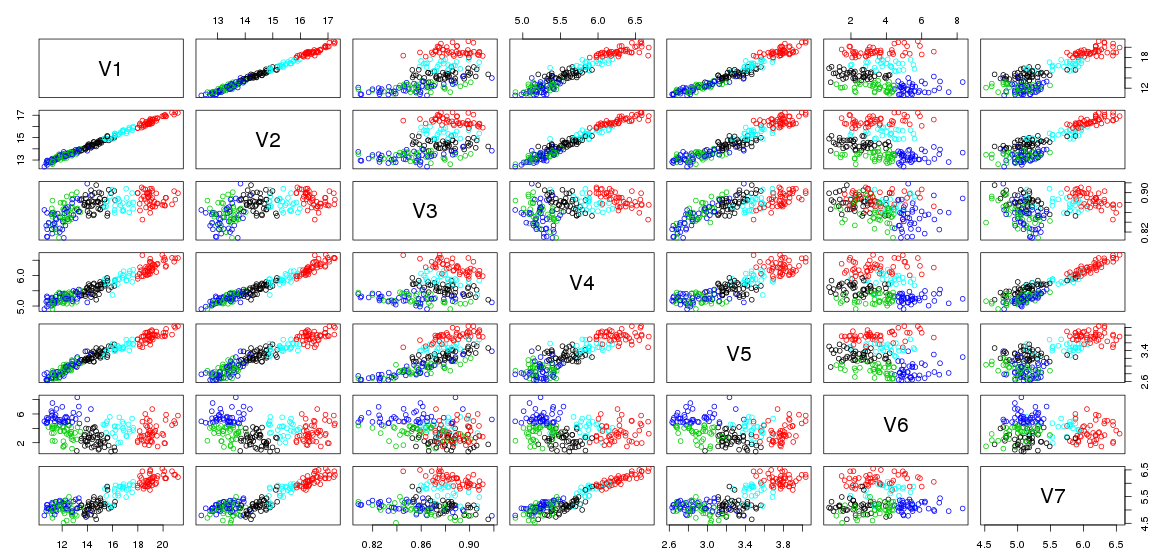
Assuming we don’t know the exact number of categories of this data. We need to try different number of clusters and compare the results with each other and reveal the influence brought by the number of centers. We will approach this analysis from two aspects: 1. The appearance of the clustering in each of the two dimensions. 2. The changes of total within sum squares (of distances from points to centers in each cluster).

* + 1. Appearance [Scheme 1: k=2]

[scheme 2: k=3]

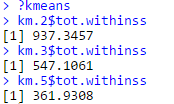


[scheme 3: k=5]



From schemes of clustering plotted in each two dimensions under different k values (number of centers), we can find out the following patterns. 1) With the increment of center number, the boundary of the clusters becomes vague. The reason behind this may be the feature of this dataset, which is the high dimension. The clustering result is influenced by all 7 attributes, each of which will influence the final clustering appearance on others. 2) The boundary is more clear in two dimensions that are linearly correlated. For example, clustering boundary in V1-V2 is more obvious than that of V3-V7.

* + 1. Total Within Sum of Squares



As we know, the k-means clustering algorithm has the potential flaw of only finding the local minima. In most cases, even with the same parameters, the results may vary in different runs (We tried several times, too. But the results remain the same. This may be caused by the mechanism that the R choose real points instead of virtual points as initial centers.). TWSS is the total square distances from points to final centers of each lusters. It from a certain aspect reflects how well the clustering performances. That is, the lower TWSS, the intensive each cluster is.

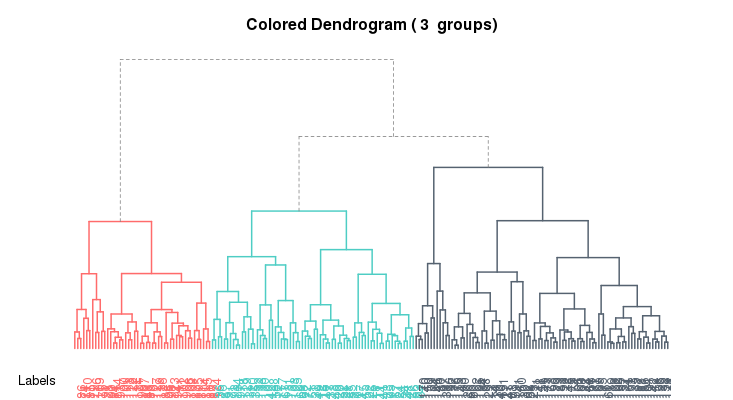
The TWSS (total within sum of squares) of 3 clusters with 2, 3, 5 centers are 937.3457, 547.1061, 361.9308. We can clearly observe the decrement of TWSS with the increasing of number of centers. However, it doesn’t mean that the more center is the more accurate the cluster will be. If k equals the number of instances (points), then the TWSS would be 0! Therefore, large size of k can reduce TWSS value, but may cause “overfitting” (over clustering) issue when we analyze dataset with k-means algorithm without knowing the exact number of clusters.

Part 2 Hierarchical Clustering on three different varieties of wheat

* 1. Overview

Hierarchical clustering, also known as hierarchical cluster analysis is a cluster analysis method which aims to construct a hierarchy of clusters and it is widely used in data mining and statistics areas. There are two types of strategies to this approach: agglomerative (also known as “bottom up”) and divisive (also known as “top down”) and normally there are three major linkage criteria to determine the distance between sets of observations: complete-linkage clustering, single-linkage clustering and average-linkage clustering. And we would like to apply those three criteria altogether to construct hierarchical clusters respectively.

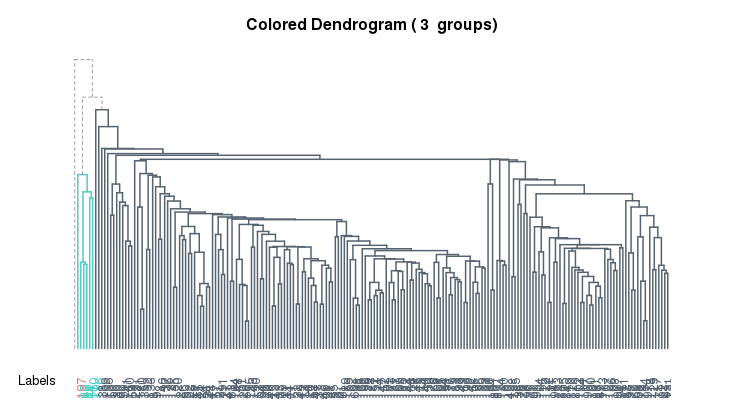
* 1. Comparison of predicated clusters and actual clusters
     1. Complete linkage



Scheme 2.2 hierarchical clusters constructed by complete-linkage clustering criteria (cut-off 3)

The colored scheme above was constructed by scripts line 13 using function A2Rplot and we can see when the cut-off point was set to 3, we have three clusters constructed by this approach (shown in red/green/black color). And since we know that in the original dataset, attribute 8 labels the categories of all those wheat kernel samples with number 1/2/3, we would like to compare these actual labels with the predicted values by our analysis. And this comparison was accomplished by line 17 to 23 which provided an accuracy of the predicted categories on the dataset comparing with the actual categories. The accuracy for this approach turned out to be 78.8944% which seems OK to us. We could observe some mislabel of original type 1 wheat to type 3 and some mislabel of original type 2 to type 1. And to our delight, the prediction for type 3 was quite good and there were rare cases where those samples were misclassified.

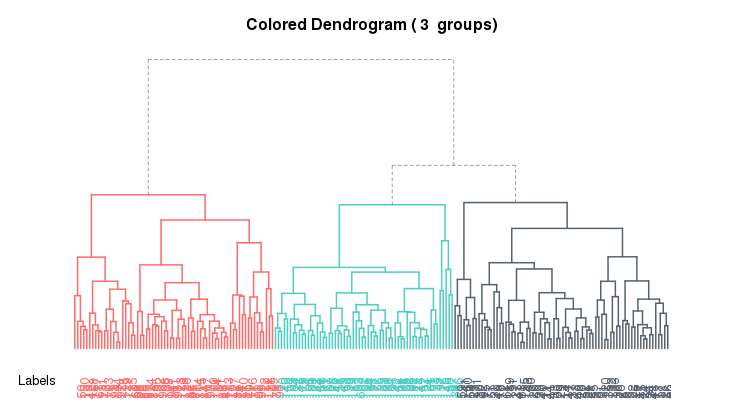
2.2.2 Single-linkage



Scheme 2.3 hierarchical clusters constructed by single-linkage clustering criteria (cut-off 3)

The colored scheme above was constructed by scripts line 14 using function A2Rplot and we can see that we also had three clusters established when the cut-off point was set to 3. We performed the same comparison analysis to calculate the accuracy of this approach and it turned to be 33.1658% which was quite disappointing. This time we could observe much more misclassification in type 1 and type 2, and even in type 3 we had much more cases of misclassification happened.

2.2.3 Average-linkage

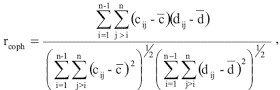


Scheme 2.4 hierarchical clusters in average-linkage clustering criteria (cut-off 3)

The colored scheme above was constructed by scripts line 15 using function A2Rplot and we can see that we also had three clusters established when the cut-off point was set to 3. The accuracy this time was only 30.1507% even worse than single-linkage approach. Needless to say, we observed even more misclassification in all types and the only conclusion or explanation we could give for this situation is that maybe complete-linkage approach is proved to be an effective one for this particular dataset while the other two apparently not

* 1. Cophenetic Correlation Analysis

Cophenetic correlation coefficient is an important measure of how a dendrogram faithfully preserves the pairwise distances between the original unmodeled data points and it is widely used in statistics and especially in biostatistics typically assessing cluster-based models of DNA sequences. Its formula is as below:

It ranges from -1 to 1 and the closer the number is to 1, the better the clustering applied in the dataset. And based on this idea, we also performed three Cophenetic correlation analyses with different approaches (complete/single/average) with scripts line 31-36 and the results were summarized below (Cor1 for complete, Cor2 for single and Cor3 for average):

Cor1: 0.742204 Cor2: 0.599655 Cor3: 0.742447

And from the result we could say complete approach was still proved to be good to this dataset while average seems also another candidate although it gave out bad accuracy in the previous part.