RDataMining

Hierarchical clustering

Dr. Liang (Leon) Chen

Instructions

- R script for hierarchical clustering is provided in the next slide, please follow it and complete the lab in R.
- You do not need to type notes (starting at #), but it's a good manner to have them in you script.
- In order to see codes and notes clearly, I show the script in RStudio.

- **The hclust function will be used in this lab. For an example about Hierarchical Clustering in R, please visit this site.**

R Script for **Hierarchical Clustering**

```
1  #define and choose the dataset
2  Lab9Data<-read.csv(file.choose(),header=T)
3  # show summary statistics of the data and check the units of each attribute
4  summary(Lab9Data)
5  # normalize the three attributes to the range 0-1
6  Lab9Data$cylinders<-(Lab9Data$cylinders-min(Lab9Data$cylinders))/(max(Lab9Data$cylinders)-min(Lab9Data$cylinders))
7  Lab9Data$cubicinches<-(Lab9Data$cubicinches-min(Lab9Data$cubicinches))/(max(Lab9Data$cubicinches)-min(Lab9Data$cubicinches))
8  Lab9Data$weightlbs<-(Lab9Data$weightlbs-min(Lab9Data$weightlbs))/(max(Lab9Data$weightlbs)-min(Lab9Data$weightlbs))
9  # show summary statistics of the normalized data
10 summary(Lab9Data)
11 # Use hclust for hierarchical clustering; hclust requires the data in the form of a distance matrix. Do this by using dist (euclidean is the default method).
12 # By default, the complete linkage method is used for hclust; use the 2nd, 3rd, and 5th columns for clustering
13 clusters <- hclust(dist(Lab9Data[, 2:3,5]))
14 # generate a cluster dendrogram
15 plot(clusters)
16 # cut off the dendrogram tree at the desired number of clusters using cutree.
17 clusterCut <- cutree(clusters, 3)
18 # generate a new column called label to save the cluster in the dataset
19 Lab9Data$label<-clusterCut
20 # use the library ggplot2
21 library(ggplot2)
22 #draw a chart to show the distribution of mpg for each cluster
23 qplot(x=label,y=mpg,data=Lab9Data)
24 #compute the average mpg for each cluster and assign a new name for the mean
25 #here I use two lines, but one line is OK: meanmpg<-setNames(aggregate(Lab9Data$mpg,list(Lab9Data$label),mean),c("cluster", "averagempg"))
26 meanmpg<-aggregate(Lab9Data$mpg,list(Lab9Data$label),mean)
27 colnames(meanmpg) <- c("cluster", "averagempg")
28 meanmpg
29 #generate a bar chart to show the mean mpg for each cluster
30 barplot(meanmpg$averagempg,main="Average MPG",names.arg=meanmpg$cluster,xlab="Cluster")
31 # conduct ANOVA test
32 summary(aov(mpg ~ factor(label), data=Lab9Data))
```

Click aggregate, dist, hclust, cutree, and aov for details about each function

Deliverable R1: take a screenshot of the dendrogram with date and time and compare it with the one generated in RM.

Deliverable R2: take a screenshot of the chart with date and time and describe it briefly.

Alternatively, you can use ggplot function: ggplot(meanmpg, aes (x=cluster,y=averagempg)) + geom_bar(stat="identity") + labs(x="Cluster",y="Average MPG")

Deliverable R3: take a screenshot of the ANOVA result with date and time and make your conclusion.

```
#save the new dataset as a csv file
write.csv(Lab9Data, file = "Lab9Data.csv")
```

Deliverable R4: save the cluster result in a csv file and then compare it with the cluster result (3-cluster model) generated at Step 4.8 in the RapidMiner lab. Are they the same? Include the screenshot of your PivotTable with date and time. Follow the same procedure we used for deliverable R4 in Week 8 R Lab.

Attention: Why do we use factor(label), instead of label in Line 32? Because label is an integer, instead of a factor. If you run this line without factor(), DF of label is 1 because label is treated as a numerical attribute.

Deliverables

- Deliverable R1: take a screenshot of the dendrogram with date and time and compare it with the one generated in RM.

- Deliverable R2: take a screenshot of the chart with date and time and describe it briefly.

- Deliverable R3: take a screenshot of the ANOVA result with date and time and make your conclusion.

- Deliverable R4: save the cluster result in a csv file and then compare it with the cluster result (3-cluster model) generated at Step 4.8 in the RapidMiner lab. Are they the same? Include the screenshot of your PivotTable with date and time. Follow the same procedure we used for deliverable R4 in Week 8 R Lab.

Q: What the y-axis Height means?

A: In hierarchical clustering, you must define a distance metric (including measure such Euclidean distance and linkage type such as complete linkage) between clusters the value of this distance metric between clusters. The y-axis is simply the value of this distance metric between clusters. For example, if you see two clusters merged at a height x, it means that the distance between those clusters was x.