クラスタリングのお勉強ノート

yoshi

2016-08-16

Table of Contents

# K-meansクラスタリング

Irvineのワインデータセットを使う。

df <- read.table("../data/wine.data", header = TRUE, sep=",")  
head(df)

## Cultivar Alcohol Malic.acid Ash lcalinity.of.ash Magnesium otal.phenols  
## 1 1 14.23 1.71 2.43 15.6 127 2.80  
## 2 1 13.20 1.78 2.14 11.2 100 2.65  
## 3 1 13.16 2.36 2.67 18.6 101 2.80  
## 4 1 14.37 1.95 2.50 16.8 113 3.85  
## 5 1 13.24 2.59 2.87 21.0 118 2.80  
## 6 1 14.20 1.76 2.45 15.2 112 3.27  
## Flavanoids Nonflavanoid.phenols Proanthocyanins olor.intensity Hue  
## 1 3.06 0.28 2.29 5.64 1.04  
## 2 2.76 0.26 1.28 4.38 1.05  
## 3 3.24 0.30 2.81 5.68 1.03  
## 4 3.49 0.24 2.18 7.80 0.86  
## 5 2.69 0.39 1.82 4.32 1.04  
## 6 3.39 0.34 1.97 6.75 1.05  
## OD280.OD315.of.diluted.wines Proline  
## 1 3.92 1065  
## 2 3.40 1050  
## 3 3.17 1185  
## 4 3.45 1480  
## 5 2.93 735  
## 6 2.85 1450

一列目は教師ラベルであるからクラスタリングをするときは取り除く。

dfNoTeacher <- df[,2:14]

３つのクラスに分類する。

wineK3 <- kmeans(x = dfNoTeacher, centers = 3)  
wineK3

## K-means clustering with 3 clusters of sizes 69, 62, 47  
##   
## Cluster means:  
## Alcohol Malic.acid Ash lcalinity.of.ash Magnesium otal.phenols  
## 1 12.51667 2.494203 2.288551 20.82319 92.34783 2.070725  
## 2 12.92984 2.504032 2.408065 19.89032 103.59677 2.111129  
## 3 13.80447 1.883404 2.426170 17.02340 105.51064 2.867234  
## Flavanoids Nonflavanoid.phenols Proanthocyanins olor.intensity Hue  
## 1 1.758406 0.3901449 1.451884 4.086957 0.9411594  
## 2 1.584032 0.3883871 1.503387 5.650323 0.8839677  
## 3 3.014255 0.2853191 1.910426 5.702553 1.0782979  
## OD280.OD315.of.diluted.wines Proline  
## 1 2.490725 458.2319  
## 2 2.365484 728.3387  
## 3 3.114043 1195.1489  
##   
## Clustering vector:  
## [1] 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 3 3 2 2 3 3 2 3 3 3 3 3 3  
## [36] 2 2 3 3 2 2 3 3 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 1 2 1 1 2 1 1 2 2  
## [71] 2 1 1 3 2 1 1 1 2 1 1 2 2 1 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 1 2 1 1 1 2  
## [106] 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 1 1 1  
## [141] 2 2 1 1 2 2 1 2 2 1 1 1 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 2 2 2 1 1 2 2 2  
## [176] 2 2 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 443166.7 566572.5 1360950.5  
## (between\_SS / total\_SS = 86.5 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

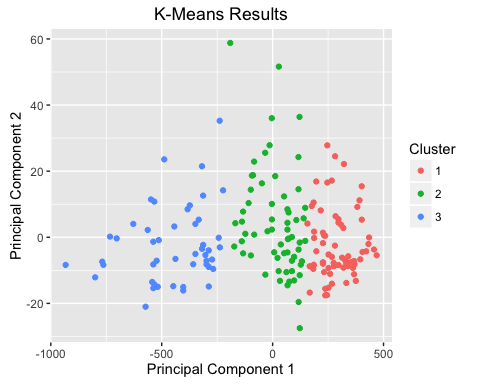
PCAして可視化する。{useful}パッケージでplotをoverrideすると簡単にkmeansの結果を可視化できる。

require(useful)

## Loading required package: useful

## Loading required package: ggplot2

plot(wineK3, data=dfNoTeacher)

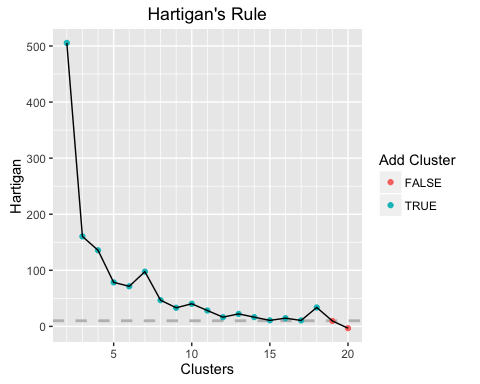


通常クラスタリングをするとき、クラスタ数は事前に分からない。クラス数の決め方としてハーティガンルールがある。{useful}パッケージにはハーティガン数を計算する関数が用意されている。

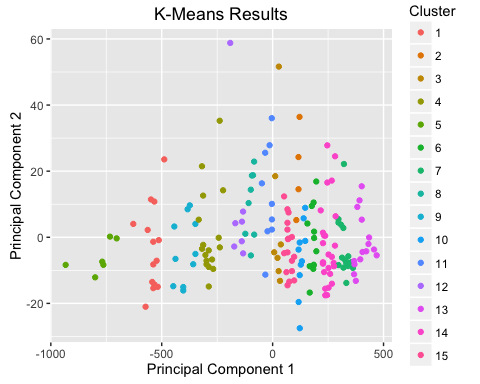
wineGrid <- FitKMeans(x=dfNoTeacher, max.clusters = 20L, nstart = 30, iter.max = 20L)  
wineGrid

## Clusters Hartigan AddCluster  
## 1 2 505.429310 TRUE  
## 2 3 160.411331 TRUE  
## 3 4 135.707228 TRUE  
## 4 5 78.445289 TRUE  
## 5 6 71.489710 TRUE  
## 6 7 97.582072 TRUE  
## 7 8 46.764658 TRUE  
## 8 9 33.197213 TRUE  
## 9 10 40.300992 TRUE  
## 10 11 28.268345 TRUE  
## 11 12 16.477020 TRUE  
## 12 13 22.263790 TRUE  
## 13 14 16.450430 TRUE  
## 14 15 10.702991 TRUE  
## 15 16 14.633768 TRUE  
## 16 17 10.609028 TRUE  
## 17 18 33.724082 TRUE  
## 18 19 9.701403 FALSE  
## 19 20 -3.229988 FALSE

PlotHartigan(wineGrid)

 ハーティガンルールによればクラス数は15にするべきである。

wineK15 <- kmeans(x = dfNoTeacher, centers = 15)  
plot(wineK15, data=dfNoTeacher)



# K-medoids法

K-meansは質的変数には使えない、外れ値に敏感であるという問題がある。

indicators <- c("BX.KLT.DINV.WD.GD.ZS", "NY.GDP.DEFL.KD.ZG",  
"NY.GDP.MKTP.CD", "NY.GDP.MKTP.KD.ZG",  
"NY.GDP.PCAP.CD", "NY.GDP.PCAP.KD.ZG",  
"TG.VAL.TOTL.GD.ZS")  
require(WDI)

## Loading required package: WDI

## Loading required package: RJSONIO

# リストの中にあるすべての国のインジケータを引っ張ってくる  
# すべての国が各指標を持ってはいない  
# いくつかの国はデータが全くない  
wbInfo <- WDI(country = "all" , indicator = indicators , start = 2011 ,  
end = 2011 , extra = TRUE)  
# 集約した情報を除く  
wbInfo <- wbInfo[wbInfo$region != "Aggregates" ,]  
# すべてのインジケータがNAの国を除く  
wbInfo <- wbInfo[which(rowSums(!is.na(wbInfo[, indicators])) > 0) ,]  
# ISOが無い行を除く  
wbInfo <- wbInfo[!is.na(wbInfo$iso2c) ,]  
  
# 国名を知ることができるようにrownamesを設定する  
rownames(wbInfo) <- wbInfo$iso2c  
# 地域を再度ファクター型に、収入や貸付は水準の変化を考慮する  
wbInfo$region <- factor(wbInfo$region)  
wbInfo$income <- factor(wbInfo$income)  
wbInfo$lending <- factor(wbInfo$lending)

各点のシルエット値は、他のクラスターの点と比べて、その点が自身のクラスター内の他の点にどれくらい相似しているかを示す尺度です。 番目の点のシルエット値 は、次のように定義されます。

$$$
S\_i = (b\_i-a\_i)/ max(a\_i,b\_i)
$$

$ ここで は 番目の点から と同じクラスターの他の点までの平均距離で、 は 番目の点から別のクラスターの点までの最小平均距離です。

# 階層型クラスタリング