Técnicas de Clusterización

Primera Parte

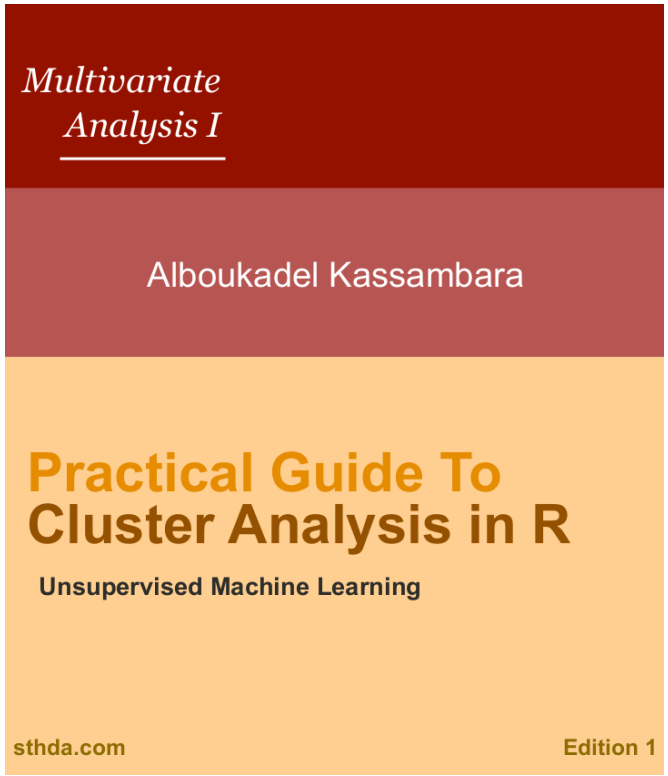
Juan C. Correa

Material de uso exclusivo para
INGENIO PANTALEON, S.A.
Diagonal 6, 10-31, Zona 10

Ciudad de Guatemala

Agenda

Clusterización



Por Clusterización se entiende al conjunto de técnicas estadísticas orientadas a la clasificación (agrupación) de observaciones con base en el cálculo de distancias o semejanzas entre cada par de observaciones.

Clusterización

Al calcular las semejanzas entre pares de observaciones, usualmente se tiene como resultado una matriz de distancias.

Algunos de los métodos para calcular las distancias o semejanzas entre pares de observaciones:

- Distancia Euclidea
- Distancia Manhattan
- Distancia por Correlación Pearson
- Distancia por Eisen Coseno
- Distancia por Correlación Spearman
- Distancia por Correlación de Kendall
- Distancias por Información Mutua

Selección del Método de Cálculo

Es muy importante saber seleccionar el método de cálculo de distancias porque este genera un impacto en la solución de clusterización. En general, el método predeterminado es la distancia Euclidea, pero no siempre necesitamos usarla.

Dependiendo de los objetivos de la investigación y los tipos de datos, pueden usarse otros métodos.

Selección del Método de Cálculo

La distancia basada en correlación considera que dos objetos son similares si sus características están altamente correlacionadas, aun cuando los valores observados puedan estar muy separados en términos de distancia Euclidia. La distancia entre dos objetos es cero cuando están perfectamente correlacionados. La correlación de Pearson es bastante sensible a los valores atípicos. Spearman es menos sensible a los valores extremos.

Selección del Método de Cálculo

Las distancias basadas en coseno son muy empleadas para tareas que buscan identificar semejanzas entre palabras, comentarios o textos. Estos casos son particularmente frecuentes en tareas de minería de textos. También es factible calcularla a través de distancias por índice de información mutua.

Ejemplo: Minería de Textos

La siguiente matriz término-documento representa la relación entre cada cliente (columna) que genera un comentario C compuesto por P palabras (filas).

	Cliente 1	Cliente 2	Cliente 3	...
buena	1	0	1	...
comida	1	0	1	...
lo	1	1	0	...
recomiendo	1	1	0	...
no	0	1	0	...
nada	0	0	1	...
. . .	...	...	...	...

La co-ocurrencia de la palabra x con la palabra y podría estimarse, por ejemplo, a través del coeficiente de información mutua, IM.

Ejemplo: Minería de Textos

$$IM(x,y) = \sum_x \sum_y P(x,y) \times log \frac{P(x,y)}{P(x)P(y)} \quad (1)$$

$IM(x,y)$ es cero cuando en un mismo comentario C una palabra está presente pero la otra no.

La distribución de frecuencias de cada palabra, entonces, suele converger a una distribución $P_n \sim 1/n^a$ y la conexión entre dos palabras distintas puede definirse si $IM \neq 0$.

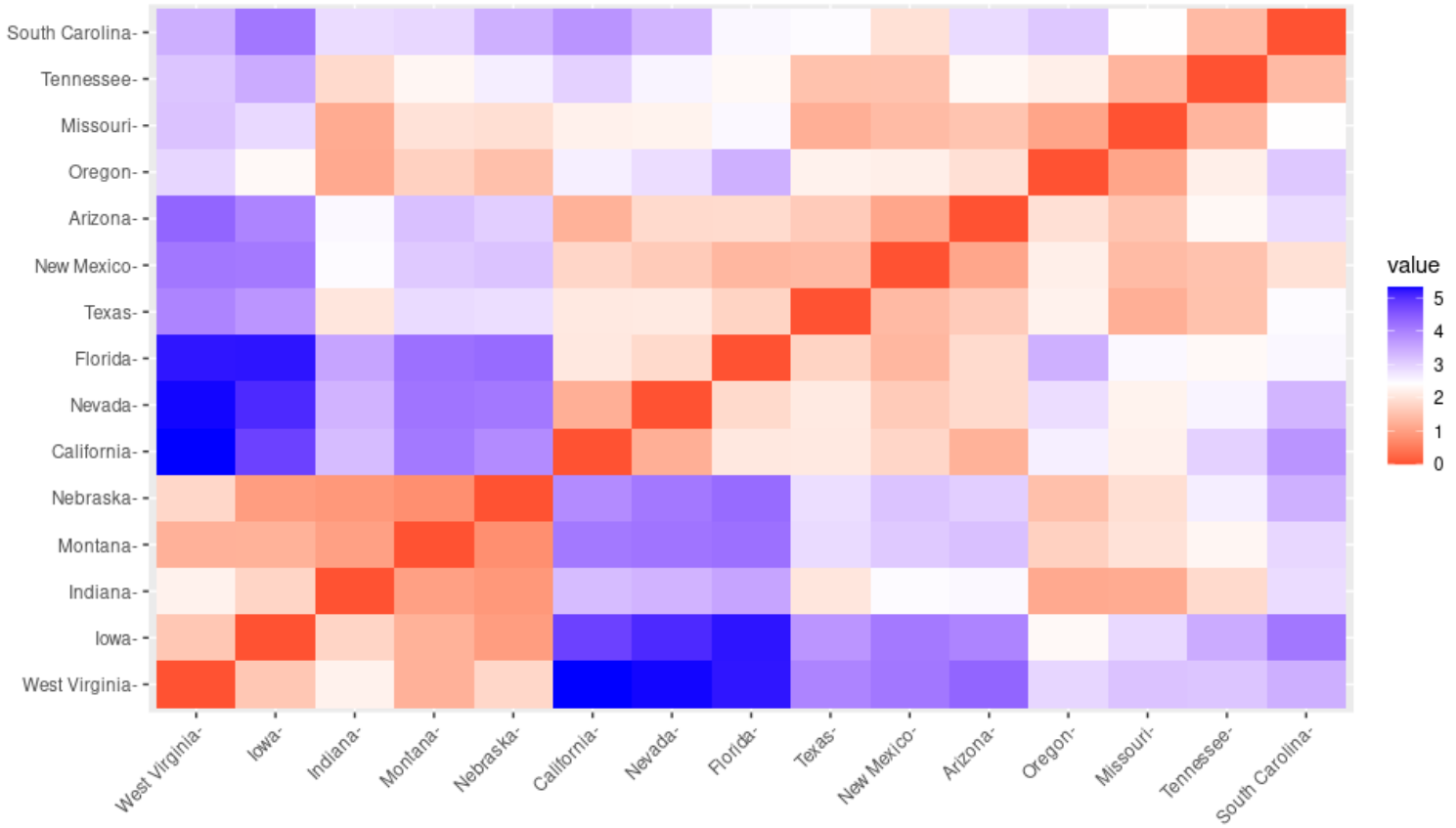
Ejemplo: Minería de Textos

El valor de las medidas de distancia está íntimamente relacionado con la escala de su medición. Por lo tanto, las variables a menudo se estandarizan antes de medir las semejanzas entre observaciones. Esto es especialmente recomendable cuando las variables se miden en diferentes escalas (por ejemplo: kilogramos, kilómetros, centímetros, ...). En la minería de textos, sin embargo, esto no es necesario porque la medición es simplemente un conteo de frecuencias.

Ejemplo: Minería de Datos

```r
data("USArrests")
df <- USArrests
# Quitemos cualquier valor perdido en los datos
df <- na.omit(df)
# Vamos a estandarizar los valores de las variables
df <- scale(df)
set.seed(123)
ss <- sample(1:50, 15)
df <- USArrests[ss, ]
df.scaled <- scale(df)
dist.eucl <- dist(df.scaled, method = "euclidean")
round(as.matrix(dist.eucl)[1:3, 1:3], 1)
library("factoextra")
dist.cor <- get_dist(df.scaled, method = "pearson")
library(factoextra)
fviz_dist(dist.eucl)
```

Ejemplo: Minería de Datos



En Rojo Alta Semejanza y en Azul Baja Semejanza.