Trabalho de Estatística Aplicada II

Matheus Machado, Leandro Dias, Maqueise de Medeiros, Caio Macedo

13/04/2021

Métodos de Agrupamentos Não-Hierárquicos

Introdução

Análise de Cluster

são técnicas estatísticas utilizada com objetivo de agrupar elementos, de forma que haja o mínimo de diferença entre os elementos de um mesmo grupo e o máximo de diferença entre elementos de grupos diferentes. O grau de diferença entre cada elemento é medido pela distância entre os mesmos, onde as noções de distância variam de situação para situação.

Quando usar?

- Quando a preocupação principal é dividir os elementos em grupos, de forma que os elementos de um mesmo grupo sejam homogêneos e os elementos em grupos diferentes sejam heterogêneos;
- Considerações teóricas, conceituais e práticas devem ser observadas ao selecionar as variáveis para a Análise de Agrupamentos;

Objetivos

Essa técnica pode ser usada para dividir e classificar os elementos entre dois ou mais grupos levando em consideração sua similiridade a partir das variáveis disponíveis. Nesse caso, o algoritmo encontra diretamente uma partição de n itens em k clusters usando dois critérios, semelhança interna e isolamento dos clusters formados. Esse tipo de técnica também pode ser usada para simplificar os dados coletados e identificar relações entre os elementos e variáveis que estão sendo analisadas.

Desvantagens

- A análise dos clusters é descritiva, não-teórica, e não inferencial;
- Divide os elementos em grupos mesmo que não haja qualquer direfença significativa entre os elementos dados;
- Não é generalizável, pois depende das variáveis usadas como base na medição de similaridade.

Métodos de agrupamento não-hierárquicos (Clustering Particional):

Existem dois tipos diferentes de técnicas de agrupamento, as hierárquicas e as não hierárquicas. Ao contrário dos agrupamentos hierárquicos, os não-hierárquicos são mutuamente exclusivos, ou seja, não há qualquer tipo de relação entre eles. Além disso, os algoritmos das técnicas não-hierárquicas possuem maior capacidade de

análise de conjunto de dados além de serem iterativos. A quantidade de número de clusters precisa ser pré-definida. Em cada etapa novos clusters podem ser formados por divisão ou junção de clusters inicialmente definidos.

Dessa forma, existem diversos métodos de agrupamentos não-hierárquicos, onde os dois principais e mais utilizados são explicados abaixo a partir da teoria e prática:

K-Médias (K-Means):

É a técnica de agrupamento não-hierárquico mais utilizada e conhecida, pois consegue uma taxa de precisão relativamente alta. Essa técnica consiste em que os objetos dentro do mesmo cluster sejam tão semelhantes quanto possível (ou seja, alta similaridade dentro de cada classe), enquanto os objetos de diferentes clusters são tão diferentes quanto possível (ou seja, baixo similaridade entre as classes). Neste método K-Médias, cada cluster é representado por seu centro (centróide) que corresponde à média dos pontos atribuídos ao cluster.

Existem diversos algoritmos disponíveis para este método, onde o de *Hartigan-Wong* é o algoritmo padrão. A ideia principal deste algoritmo é definir clusters de forma que a variação total dentro do cluster seja a menor possível. É dado da seguinte forma: $C(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$, onde x_i é um ponto de dados pertencente ao cluster C_k e μ_k é o valor médio dos pontos atribuídos ao cluster C_k.

Com isso, cada observação x_i é atribuída a um determinado cluster, de modo que a soma dos quadrados da distância da observação aos seus centros de cluster atribuídos μ_k seja minimizado.

A variação total dentro de cada cluster é dado da forma: $var.tot = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$, de forma que essa soma total do quadrado dentro de cada cluster seja a menor possível.

Por fim, as etapas deste algoritmo são dadas da forma:

1. Especificar o número de clusters K a serem criados, podendo ter o número ideal estimado a partir de alguns dos métodos abaixo: 1.1. Método do "Cotovelo" - A ideia básica por trás desse método é definir clusters de modo que a variação total dentro do cluster a menor possível através da fórmula: $minimize\left(\sum_{k=1}^{k} C(C_k) \right)$, onde C_k é o k^{th} cluster e $C(C_k)$ é a variação dentro do cluster. A soma desses elementos (wss) mede a compactação do cluster. 1.2. Método Silhoutte - Este método mede a qualidade de um agrupamento. Ou seja, determina o quão bem cada objeto se encontra em seu cluster. 1.3. Gap Statistic - Por fim, este método compara a variação total dentro dos clusters para diferentes valores de k com seus valores esperados sob uma distribuição sem agrupamento óbvio. O conjunto de dados de referência é gerado usando simulações de Monte Carlo do processo de amostragem.
2. Selecionar k objetos aleatoriamente do conjunto de dados como os centros de cluster iniciais;
3. Atribuir cada observação ao seu centróide mais próximo, com base na distância euclidiana ($d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$) entre o objeto e o centróide;
4. Para cada um dos k clusters, recalcular o centróide do cluster calculando o novo valor médio de todos os pontos de dados no cluster.
5. Minimizar iterativamente o total dentro da soma do quadrado. Repita a Etapa 3 e a Etapa 4, até que os centróides não mudem ou o número máximo de iterações seja alcançado, onde o R utiliza o valor 10 como o valor padrão para o número máximo de iterações.

Fuzzy C-Médias:

O método de agrupamento não-hierárquico Fuzzy C-Médias é dado da forma em que cada elemento tem uma determinada probabilidade de pertencer a cada cluster (agrupamento/grupo). Este método é diferente do K-Médias, onde lá cada objeto é afetado exatamente para um determinado agrupamento.

No método de Fuzzy C-Médias, o grau ao qual um elemento pertence a um determinado cluster é um valor numérico que varia de 0 a 1. Dessa forma, os objetos de dados mais próximos dos centros dos clusters tem graus de associação mais elevados do que os objetos espalhados nas bordas dos clusters.

Sendo assim, este método exige a definição inicial de k clusters. Ou seja, sendo n itens e c variáveis aleatórias, busca-se a partição que minimiza a função: $\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d(X_j, V_i)$, onde V_i é o centróide ponderado do cluster i; $m > 1$ é o parâmetro de Fuzzy; μ_{ij} é a probabilidade do item X_j pertencer ao grupo de centróides V_i; e $d(X_j, V_i)$ é a distância escolhida para o cálculo.

Estas distâncias definem como a similaridade de dois elementos x, y é calculada e como vai influenciar na formação dos clusters. As distâncias que podem ser escolhidas para este método e se encontram na função $funny()$ utilizada no exemplo são:

- Distância Euclidiana: $d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distância de Manhattan: $d_{man}(x, y) = \sum_{i=1}^{n}|(x_i - y_i)|$
- Distância Euclidiana ao quadrado.

Por fim, os clusters são modificados a cada iteração e o processo termina quando a distância entre os centroides dos dois últimos passos é $d(X_t, V_{t+1}) < \epsilon$.

Exemplo de uma aplicação para o método K-Médias

Para esse exemplo será utilizado a base de dados $mtcars$ que é um $dataframe$ nativo do R.

Sendo assim, temos os pacotes a serem utilizados e a base no código abaixo:

```
library(factoextra)
library(tidyverse)
library(cluster)
base = mtcars
```

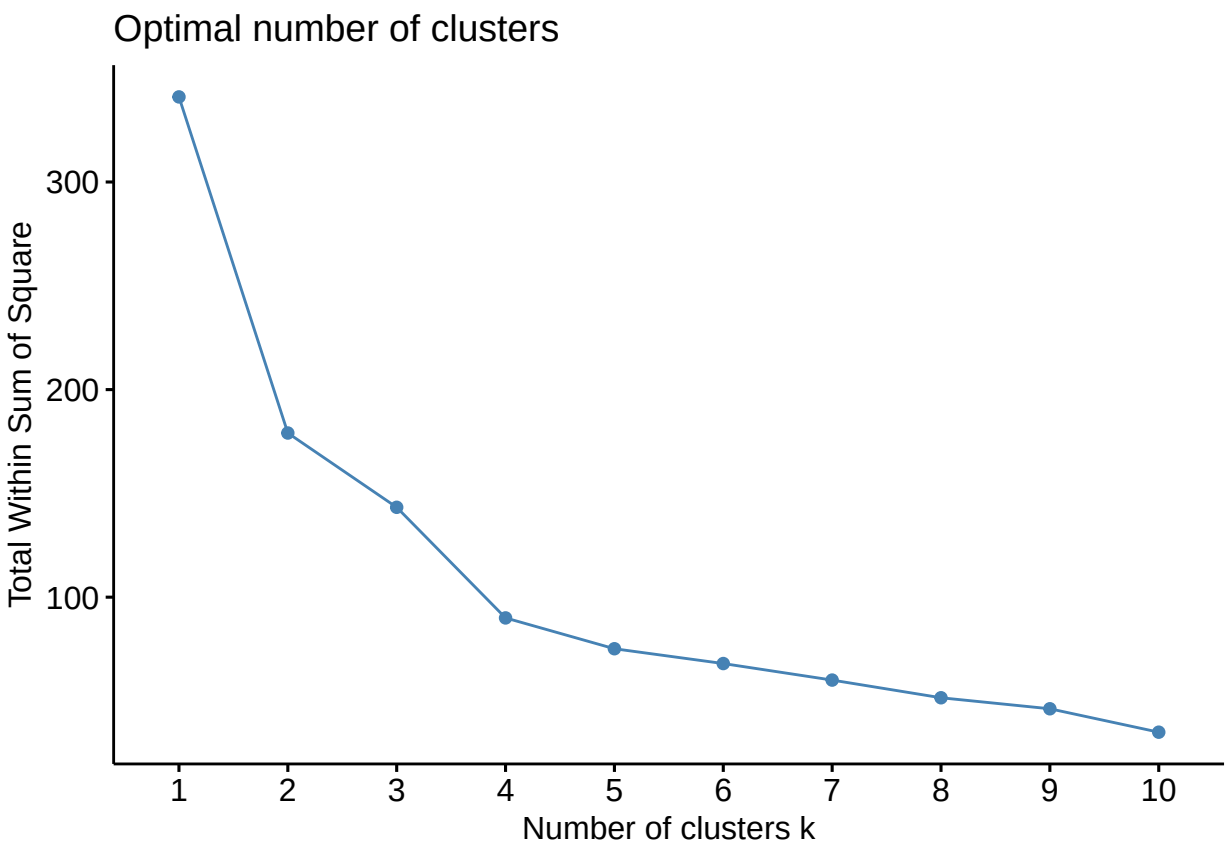
Como não é interessante que o algoritmo de agrupamento dependa de uma unidade de variável arbitrária, deve-se utilizar a função $scale()$. Esta função irá reduzir o impacto dos outliers e permitir a comparação de uma única observação com a média. Se o valor for alto podemos ter certeza de que essa observação está acima da média e um valor grande irá implicar que este ponto está longe da média em termos de desvio padrão. Ou seja, o valor sendo dois indica que o valor está a 2 desvios padrão da média.

```
df = scale(base)
head(df,n=3)
```

```
##                    mpg        cyl       disp         hp      drat         wt
## Mazda RX4     0.1508848 -0.1049878 -0.5706198 -0.5350928 0.5675137 -0.6103996
## Mazda RX4 Wag 0.1508848 -0.1049878 -0.5706198 -0.5350928 0.5675137 -0.3497853
## Datsun 710    0.4495434 -1.2248578 -0.9901821 -0.7830405 0.4739996 -0.9170046
##                    qsec         vs       am      gear       carb
## Mazda RX4     -0.7771651 -0.8680278 1.189901 0.4235542  0.7352031
## Mazda RX4 Wag -0.4637808 -0.8680278 1.189901 0.4235542  0.7352031
## Datsun 710     0.4260068  1.1160357 1.189901 0.4235542 -1.1221521
```

Estimando o número ideal de Clusters pelo método do "Cotovelo":

```
fviz_nbclust(df, kmeans, method = "wss")
```



Ao análisar o gráfico acima, nota-se que o tamanho ideal de k é 4, pois é o tamanho que aparenta ser a dobra de um cotovelo, como o próprio nome do método diz.

Sendo assim, como a abordagem sugeriu o número de clusters sendo igual a 4, foi realizada a análise final usando esse valor de k. Para o parâmetro *nstart* da função, foi utilizado sendo igual a 25, que é o que a grande parte da literatura determina, ou seja, irão ser gerados 25 configurações iniciais. Além disso, é necessário utilizar uma semente pois as configurações mudam toda vez.

```
set.seed(3)
final = kmeans(df, 4, nstart = 25)
final$cluster
```

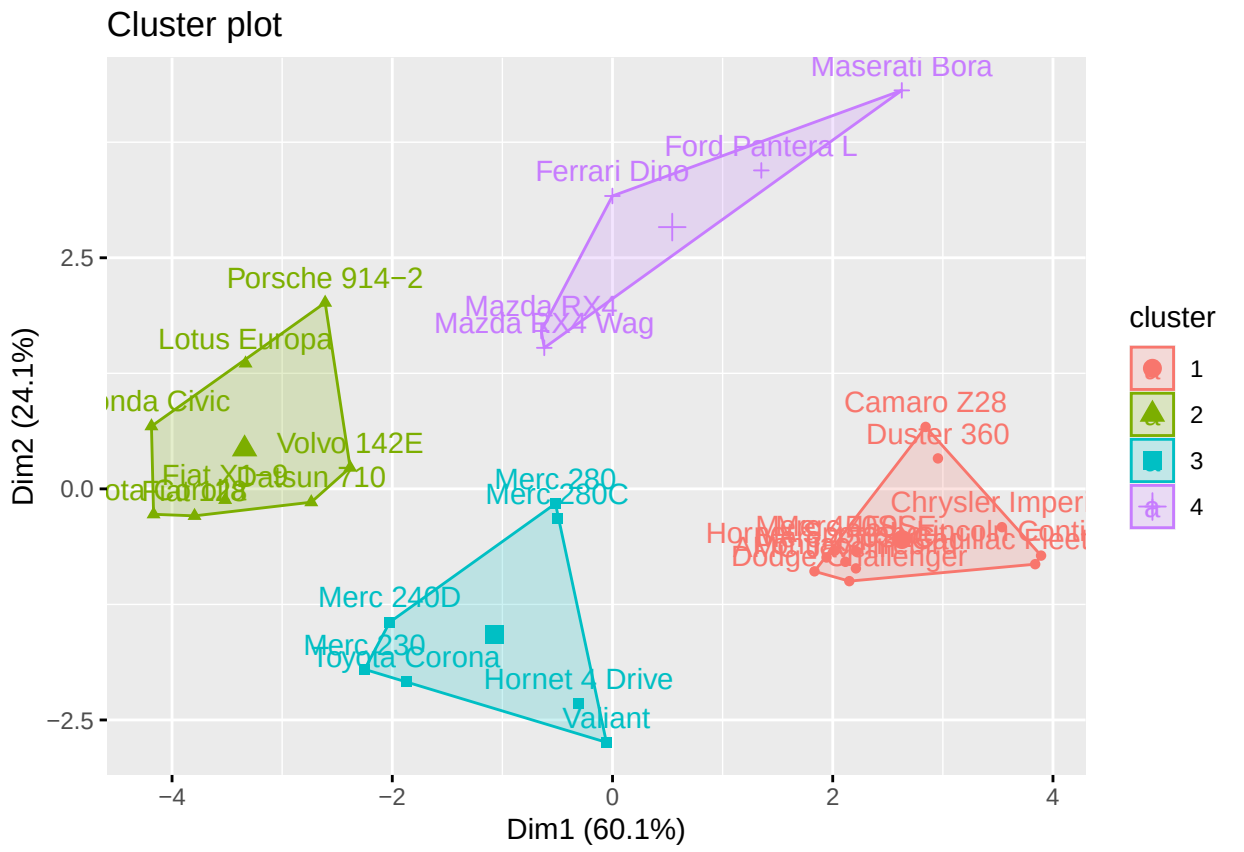
```
##          Mazda RX4      Mazda RX4 Wag         Datsun 710     Hornet 4 Drive
##                  4                  4                  2                  3
##  Hornet Sportabout            Valiant         Duster 360          Merc 240D
##                  1                  3                  1                  3
##           Merc 230           Merc 280          Merc 280C         Merc 450SE
##                  3                  3                  3                  1
##         Merc 450SL        Merc 450SLC  Cadillac Fleetwood Lincoln Continental
##                  1                  1                  1                  1
##  Chrysler Imperial           Fiat 128        Honda Civic     Toyota Corolla
##                  1                  2                  2                  2
##       Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28
```

```
##                   3                 1                 1                 1
##    Pontiac Firebird          Fiat X1-9       Porsche 914-2       Lotus Europa
##                   1                 2                 2                 2
##      Ford Pantera L       Ferrari Dino       Maserati Bora          Volvo 142E
##                   4                 4                 4                 2
```

Então, ao analisar os dados acima, temos 12 carros fazendo parte do primeiro cluster, 8 do segundo, 7 do terceiro e 5 do quarto cluster.

Por fim, para uma melhor visualização desses dados, pode-se gerar o gráfico abaixo:

```
fviz_cluster(final, data = df)
```



Exemplo de uma aplicação para o método Fuzzy C-Médias

Para esse exemplo será utilizado a base de dados *iris* que também é um *dataframe* nativo do R.

Sendo assim, os pacotes a serem utilizados são os mesmos que foram utilizados no exemplo da aplicação do método de K-Médias e a base, após uma mudança na mesma, onde foi retirada a coluna 5 e realocada no nome da linha de forma à haver uma melhor visualização da mesma está descrito no código abaixo:

```
base = iris
base = base %>% group_by(Species) %>% mutate(spec=row_number()) %>% unite('Species',
Species, spec, sep="-",remove = T) %>% column_to_rownames('Species')
```

Desse modo, como foi dito anteriormente, este método necessita de uma definição inicial para o tamanho de k, ou seja, quantos grupos desejamos. Nesse caso, a função *fanny*() abaixo, do pacote *cluster* foi utilizada onde necessitamos apenas informar a base de dados e o número de clusters iniciais. Além disso, neste método de Fuzzy C-Médias, não é necessário utilizar a função *scale*(). A função se apresenta no código abaixo, onde a segunda linha do código nos retorna os 5 primeiros valores dos clusters das bases de dados. Ou seja:

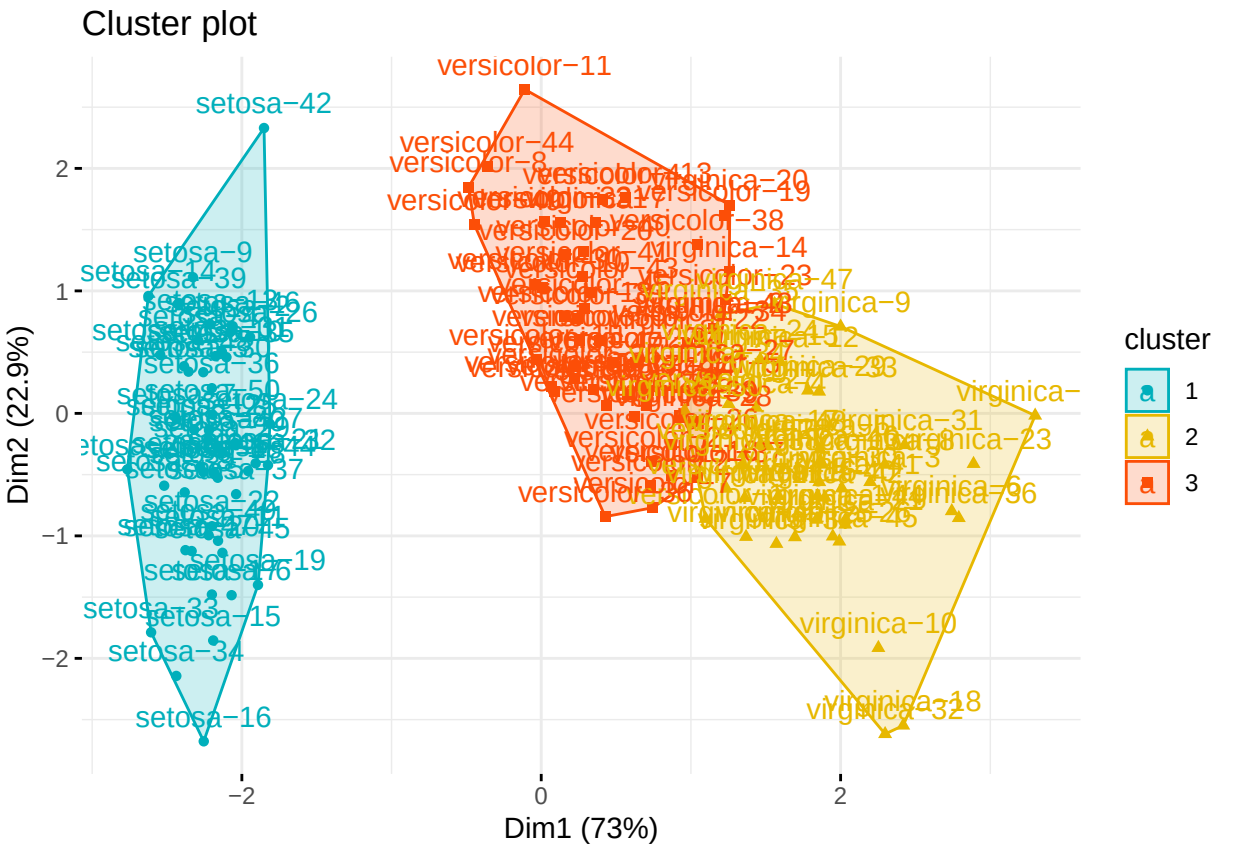
```
fuz = fanny(base,3)
head(fuz$clustering,5)
```

```
## setosa-1 setosa-2 setosa-3 setosa-4 setosa-5
##        1        1        1        1        1
```

Então, ao analisar o elemento *clustering* que é obtido após rodar a função, é verificado para quais grupos/clusters cada linha da base de dados é pertencente.

Por fim, para uma melhor visualização desses dados, pode-se gerar o gráfico abaixo:

```
fviz_cluster(fuz,ellipse.type = "convex",palette=c("#00AFBB","#E7B800","#FC4E07"),
             ggtheme = theme_minimal(), legend="right")
```



Comentários:

- Estas duas técnicas apresentadas (K-Médias e Fuzzy C-Médias) são também sensíveis às escalas e aos outliers.

- Comparando às técnicas, pode-se afirmar:
- Quando os grupos estão bem separados, qualquer técnica levará a resultados satisfatórios;
- Quando há interseção inicial entre os grupos, o método Fuzzy é melhor por gerar a probabilidade dos itens.