Agrupamento de marcas de whiskies por particionamento baseado em medóides

ANÁLISE MULTIVARIDA | MÉTODOS EM APRENDIZADO NÃO-SUPERVISIONADO DE MÁQUINAS
MI407|ME921 - Atividade prática I
Prof. Guilherme G. Ludwig

Wesley R. da Silva Satelis

20 de abril de 2021

1 Introdução

Neste trabalho foram analisadas notas dadas a diferentes marcas de whisky, por usuários do site https://www.whiskybase.com/whiskies/brands. As marcas estão dispostas por país de origem, classificação de especialistas, número de whiskies, número de votos e notas de usuários.

O objetivo desta análise foi agrupar as marcas de whisky utilizando as notas datas por usuários do site como atributo. Os grupos resultantes foram comparados com as classificações feitas por especialistas em whiskies, ordenadas de A a G, e o método de agrupamento utilizado foi o de Particionamento Baseado em Medóides ou *Partition Around Medoids (PAM).*

2 Métodos

2.1 Particionamento Baseado em Medóides

O método de particionamento baseado em medóides proposto por Kaufman e Rousseeuw (2009), é uma modificação do K-means. O método de agrupamento k-means visa agrupar atributos em um número de grupos pré-especificado. Usualmente itens são realocados de forma a minimizar o quadrado da sua distância euclidiana em relação ao centróide do grupo.

A distância euclidiana é utilizada para compor o que chamamos de matriz de dissimilariades. Seja $\mathbf{x}_i = (x_{i1}, \cdots, x_{ir})^\tau$ e $\mathbf{x}_j = \left(x_{j1}, \cdots, x_{jr}\right)^\tau$ dois pontos em R^r. Então, a matriz de dissimilaridades é definida como

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left[\left(\mathbf{x}_i - \mathbf{x}_j\right)^\tau \left(\mathbf{x}_i - \mathbf{x}_j\right)\right]^{1/2} = \left[\sum_{k=1}^{r} \left(x_{ik} - x_{jk}\right)^2\right]^{1/2}.$$

O algoritmo k-means é descrito a seguir.

1. Considera um conjunto de características X_{nxp} e a matriz de dissimilaridades.
2. Dentro de uma "caixa" que engloba os dados, escolha aleatoriamente K pontos $c_1, ..., c_k$ para $j = 1, ..., K$. Os pontos c_i são chamados centróides.
3. Conecta a observação x_i ao centróide j de menor distância. Troca a etiqueta i por j.

4. Atualiza

$$\mathbf{c}_j = \arg\min_{\mathbf{c}\in\mathbb{R}^p} \frac{1}{N_j} \sum_{i=1}^{n} \mathbf{1}\{i \text{ está no cluster } j\} d^2\left(\mathbf{x}_i, \mathbf{c}\right),$$

em que N_j é o número de observações no agrupamento j.

5. Atualiza as etiquetas e itera 3 e 4 até que as observações não mudem de grupo.

O PAM se diferencia do K-means pois procura por objetos, candidatos de centróides, entre as observações e tenta minimizar a distância com outros objetos dentro do grupo, trocando de candidato sempre que isso reduzir o valor da função objetivo. Mais detalhes podem ser encontrados em Van der Laan, Pollard, e Bryan (2003).

2.2 Gráficos de silhueta

Gráficos de silhueta são utilizados para determinar o número ideal de grupos. Suponha que C_k seja um agrupamento com K grupos.

Vamos primeiro definir as medidas de silhueta. Seja $c(i)$ o grupo que contem o i-ésimo item e a_i a similaridade média deste i-ésimo elemento em relação à todos os outros membros do mesmo agrupamento $c(i)$. Agora, considere c como algum outro grupo diferente de $c(i)$, e seja $d(i,c)$ a similaridade média entre o i-ésimo grupo com todos os membros de c. Calcule $d(i,c)$ para todos os demais grupos que não são c. Definindo $b_i = \min_{c\neq c(i)} d(i,c)$, se $b_i = d(i,c)$, então c é chamado de vizinho do i-ésimo item e é considerado o segundo melhor grupo para o i-ésimo item.

O i-ésimo valor de silhueta é dado por

$$s_i\left(C_K\right) = s_{iK} = \frac{b_i - a_i}{\max\left\{a_i, b_i\right\}},$$

de forma que $-1 \leq s_{iK} \leq 1$. Valores positivos e altos para s_{iK} indicam que o i-ésimo item está bem alocado, valores negativos e altos para s_{iK} indicam um agrupamento ruim, e $s_{iK} \approx 0$ indicam que o i-ésimo item está entre dois grupos. E ainda, se $\max_i \left\{s_{iK}\right\} < 0.25$, o procedimento de agrupamento não encontrou grupos definidos.

Em um gráfico de silhueta todos os $\{s_{iK}\}$ estão ordenados em ordem decrescente por grupo. $\bar{s}_K$ é a média de todos os $\{s_{iK}\}$ e é usada para definir o número ideal de grupos, bem como avaliar a qualidade de grupos separadamente. Mais informações podem ser encontradas em Izenman (2008).

3 Agrupamento de marcas de whiskies

Foram utilizadas as notas dadas por usuários aos whiskies do site. A fim de compor uma métrica justa, também foi incorporado o número de votos que cada whisky recebeu, gerando uma nota suavizada. Assim, as novas notas foram recalculadas por

$$\hat{u}_j^{(s)} = \frac{n_j \hat{u}_j + 1}{n_j + 2} \times 100\%,$$

em que n_j é o número de avaliadores, e u_j a nota média entre 0 e 1.

Foram consideradas somente avaliações feitas na Irlanda e marcas de whiskies com pelo menos 5 votos, resultando em um conjunto de 71 observações.

A Figura 1 mostra gráficos de agrupamentos para diferentes números de grupos prefixados. Os gráficos apenas ilustram o número de itens em cada grupo, devemos apenas nos atentar às cores, que discriminam os grupos, e ao número de observações em cada grupo. A variação ao longo do eixo y existe apenas para que que as observações não se sobreponham. Todos os agrupamentos foram feitos utilizando o método PAM, variando somente o número de grupos.

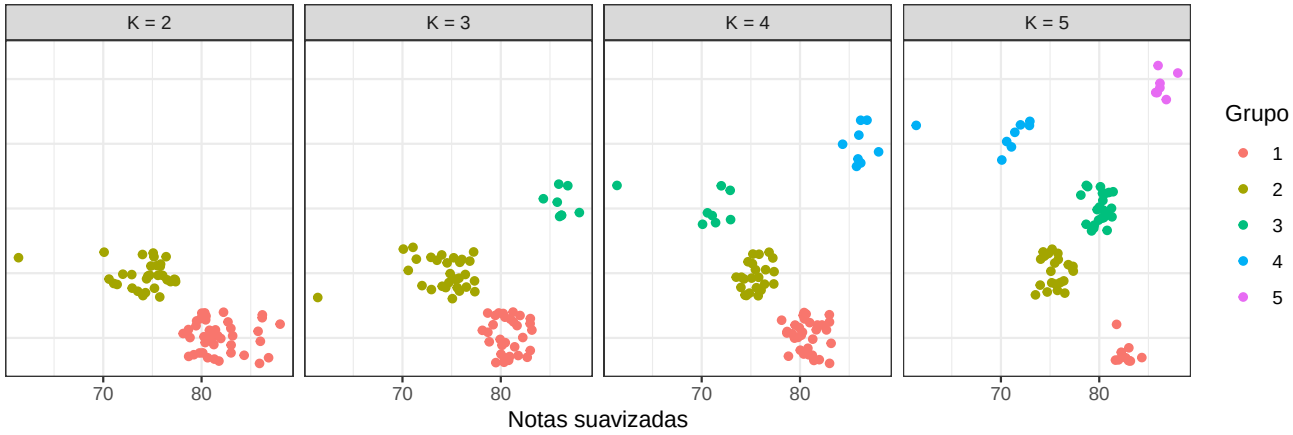

Figura 1: Gráficos de diferentes agrupamentos com K grupos. Variações ao longo do eixo y foram adicionadas apenas para que que as observações não se sobreponham e existem apenas na visualização dos dados.

Na Figura 2, temos os gráficos de silhueta para os mesmos agrupamentos expostos na Figura 1. Os agrupamentos com 4 e 5 grupos resultaram em alguns valores negativos de silhueta, apesar de não serem valores altos, é preferível evitar este tipo de resultado. As médias das silhuetas em cada agrupamento são maiores nos agrupamentos com 2 e 3 grupos, estes também não apresentam silhuetas negativas. Como o agrupamento com 2 grupos resultou em uma média de silhuetas maior que os demais, podemos dizer que este é o melhor cadidato dentre os testados. Assim, as notas de usuários foram agrupadas em dois grupos apenas.

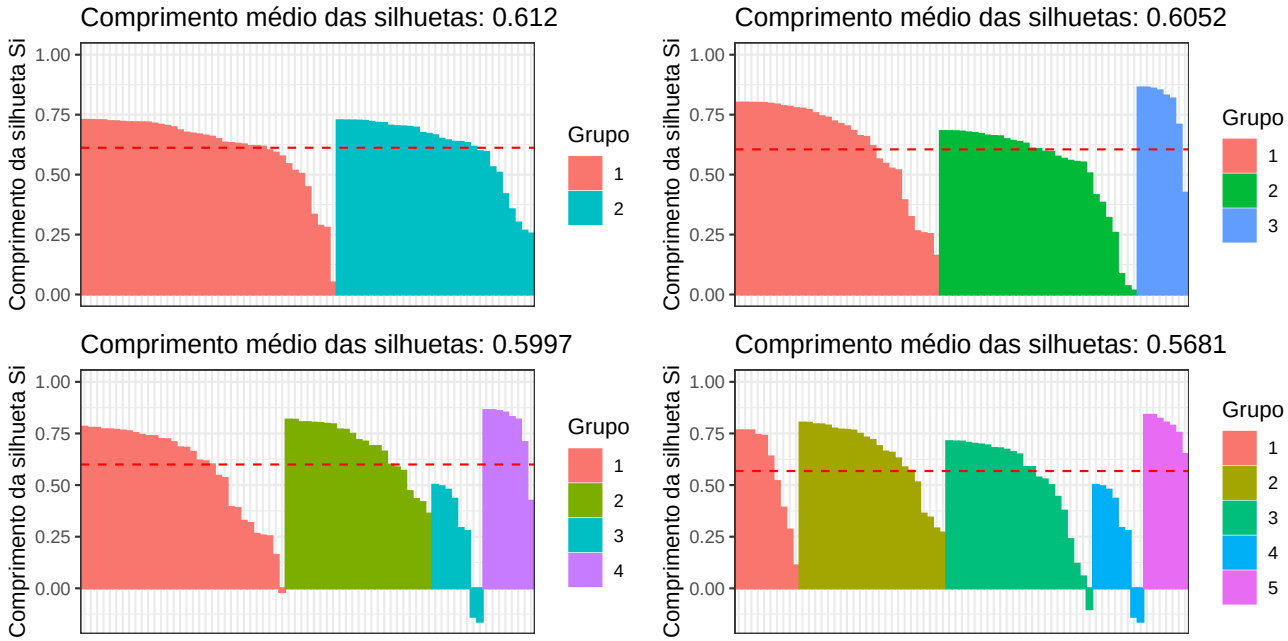

Figura 2: Gráficos de silhueta para diferentes agrupamentos com K grupos.

Com o gráfico da Figura 3, temos o objetivo de comparar as classificações dadas por especialistas com os grupos formados pelas notas dos usuários. As notas dos especialistas e dos usuários não convergem e por fim temos cinco classificações em comparação com dois grupos, respectivamente.

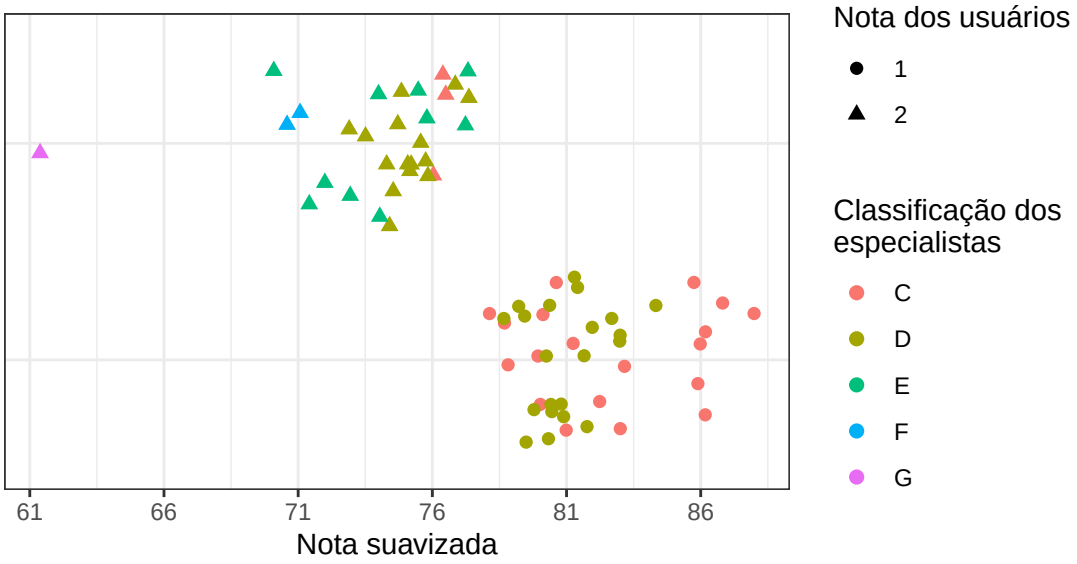

Figura 3: Comparação do agrupamento de dois grupos, usando notas de usuários como atributo, e classificações dadas por especialistas em whiskies. Variações ao longo do eixo y foram adicionadas apenas para que que as observações não se sobreponham e existem apenas na visualização dos dados.

4 Conclusões

Apesar de parecer que o agrupamento foi erroneo por não concordar com as classificações de especialistas, vale lembrar que se tratam de grupos de pessoas com habilidades diferentes.

Examinando mais de perto, nota-se que os usuários e especialistas tendem a concordar quanto às melhores marcas de whisky, marcas C. E também na maior parte das marcas D.

Faz sentido que as notas de especialistas apresentem mais grupos do que as notas de usuários, já que espera-se que estes sejam leigos no assunto e que especialistas tenham o paladar mais apurado para whiskies e sejam capazes de discernir melhor entre as categorias.

5 Referências

Izenman, Alan Julian. 2008. "Modern multivariate statistical techniques". *Regression, classification and manifold learning* 10: 978–70.

Kaufman, Leonard, e Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Van der Laan, Mark, Katherine Pollard, e Jennifer Bryan. 2003. "A new partitioning around medoids algorithm". *Journal of Statistical Computation and Simulation* 73 (8): 575–84.

6 Anexos

```r
library(tidyverse)
library(cluster)
library(data.table)

data <- read_csv("atividade1.csv") %>%
  filter(Country == "Ireland" & Votes >= 5) %>%
  select(-X7) %>%
  drop_na() %>%
  group_by(Brand) %>%
  mutate(nota_suavizada =  ((Votes * (Rating/100) + 1)/(Votes + 2))*100) %>%
  ungroup()

data_cluster <- select(data, nota_suavizada)

data <- setDT(data)
for(i in 2:5){
  cluster <- pam(data_cluster, i, keep.diss = T)$clustering
  data[, paste0("K = ", i) := cluster ]
}

data <- as_tibble(data) %>%
  gather(key="K", value="Cluster", "K = 2":"K = 5")

k_uniq <- unique(data$K)
silhuetas <- list()
for(i in 1:4){
  sil <- filter(data, K == k_uniq[i])
  silhuetas[[i]] <- silhouette(sil$Cluster, dist(sil$nota_suavizada))
}
```