Programa de Pós-Graduação em Ciência da Computação da Universidade Federal Rural do Semi-Árido (UFERSA)

Introdução a Ciência de Dados com R

K-means: Uma aplicação utilizando a ferramenta R

Leonardo Torres Marques

Rayana Souza Rocha

Mossoró – RN, 14 de novembro de

SUMÁRIO

1. Apresentação do K-means

O algoritmo *k-means* (também conhecido como: *k-médias*) é uma técnica de clusterização do grupo particionado. Esta técnica caracteriza-se por: não utilizar um supervisor para definir previamente os padrões que serão gerados; ser dependente de uma entidade externa que informe qual a quantidade *k* de clusters será formada (é daí que vem a primeira letra do nome *k-means*); tem o objetivo de criar *k* grupos (clusters, partições) iniciais, e em seguida, utilizar uma técnica de realocação iterativa fundamentada em similaridade, de forma a melhorar a posição dos centróides dentro de cada grupo.

O método k-means tem como vantagens : Ser relativamente escalável e eficiente para grandes conjuntos de dados. O método frequentemente termina num local ótimo. Entretanto, este método só pode ser aplicado quando a média (centróide) de um cluster pode ser definido. Isto pode não ser o caso em algumas aplicações, que utilizam dados com atributos categóricos (nominais) estão envolvidos.

2. Base de dados

A base de dados utilizada, foi extraída do sistema integrado de gestão de atividades (SIGAA) da UFERSA, os dados são de registros dos alunos de ciência da computação, em que 11 (onze) variáveis foram possíveis serem analisadas. As características analisados foram, sexo, raça, estado de origem, município de origem, período de ingresso, período que evadiu, ação afirmativa de ingresso, forma de ingresso, motivos da evasão, renda familiar do aluno, Índice Regular de Aproveitamento (IRA) e total de reprovações do aluno evadido. A base está disponível, no seguinte *link*: https://github.com/rayanasouza/datasciencewithR.

3. Utilizando o K-means no R

Para utilização do algoritmo de clusterização K-means, utilizou-se o pacote da biblioteca do RStudio, conforme a Figura 1.

Figura 1: Importação das bibliotecas.

```
install.packages("cluster")

library(cluster)
library(readxl)
```

Fonte: O Autor, 2018

Logo em seguida, associou-se a base de dados disponibilizada pela UFERSA à variável *"evadidos"*, como é apresenta-se na Figura 2. A variável pode ser visualizada com o comando "*view(evadidos)*".

Figura 2: Salvando a base de dados na variável "evadidos".

```
evadidos <- read_excel("datasciencewithR/clusters/evasao_evadidos (2).xlsx")
View(evadidos)
```

Fonte: O Autor, 2018

No caso da base utilizada neste estudo, devido a falta de padrão dos tipos de dados foi necessário um processo de descategorização. De modo que todos as variáveis ficaram com o mesmo tipo de dados, no caso numérico, na Figura 3 apresenta-se o comando que foi aplicado para a descategorização dos dados.

Figura 3: Descategorização dos dados.

```
# descategorização dos dados
evadidos$sexo = as.numeric(as.factor(evadidos$sexo))
evadidos$raca = as.numeric(as.factor(evadidos$raca))
evadidos$estado_origem = as.numeric(as.factor(evadidos$estado_origem))
evadidos$municipio_origem = as.numeric(as.factor(evadidos$municipio_origem))
evadidos$situacao = as.numeric(as.factor(evadidos$situacao))
evadidos$periodo_ingresso = as.numeric(as.factor(evadidos$periodo_ingresso))
evadidos$afirmativa = as.numeric(as.factor(evadidos$afirmativa))
evadidos$forma_ingresso = as.numeric(as.factor(evadidos$forma_ingresso))
evadidos$renda = as.numeric(as.factor(evadidos$renda))
evadidos$IRA = as.numeric(as.factor(evadidos$IRA))
evadidos$reprovacoes = as.numeric(as.factor(evadidos$reprovacoes))
```

Fonte: O Autor, 2018

Com os dados descategorizados, foi possível gerar os gráficos da base de acordo com a coluna escolhida. Por exemplo: *"evadidos$municipio_origem"*, *"evadidos$raca"* e *"evadidos$sexo"*, de acordo com a Figura 4.

Figura 4: Plotagem dos gráficos por variável.

```
plot(evadidos$municipio_origem, type = "p", main = "plot(x, type = \"s\")", cex = 1, col = "blue") #plotagem dos dados da base por coluna
plot(evadidos$raca) #plotagem dos dados da base por coluna
plot(evadidos$sexo, type = "p", xlab = "a", ylab = "b") #plotagem dos dados da base por coluna
```

Fonte: O Autor, 2018

Com isso, o algoritmo escolhido para este estudo foi aplicado. Na Figura 5 denota-se o código com aplicação do *K-means* na base de dados.

Figura 5: Aplicação do K-means

```
df<-scale(evadidos) # padrão centraliza e / ou dimensiona as colunas de uma matriz numérica
# kmeans - visa particionar os pontos em k grupos de forma que a soma dos quadrados dos pontos aos centros de cluster designados seja minimizada
kmf<-kmeans(df, 2) # guarda dentro de 'kmf' o resultado do que feito na chamda da função 'kmeans'
attributes(kmf) # mostra os atributos da variável 'kmf'
```

Fonte: O Autor, 2018

Uma vez aplicado o algoritmo, visualizar-se os atributos da variável "kmf", onde foi associado o resultado do comando "`kmf<-kmeans(df, 2)`", conforme a Figura 6.

Figura 6: Visualização dos atributos da variável "kmf"

```
kmf$centers

kmf$cluster
```

Fonte: O Autor, 2018

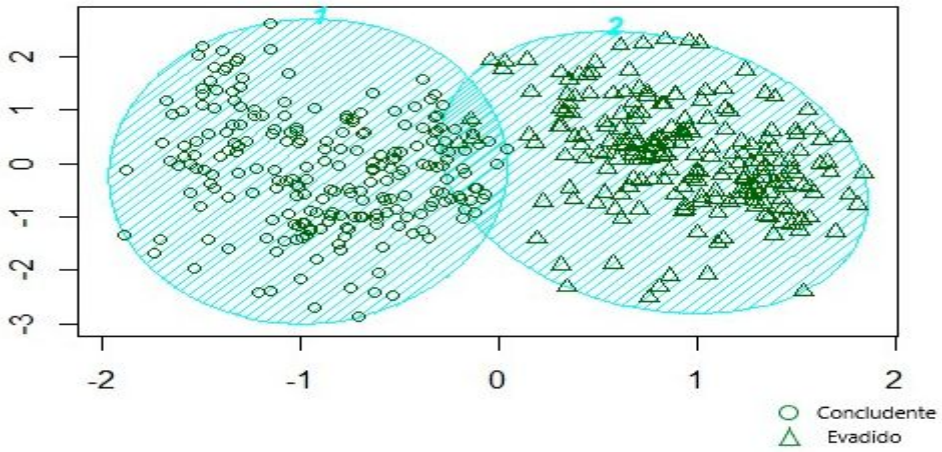
Finalmente, foi gerado o gráfico com a distribuição das observações utilizando o comando "`clusplot()`" de acordo com a Figura 7. Os *clusters* retornados, podem ser na Figura 8.

Figura 7: Plotando o gráfico.

```
clusplot(df, kmf$cluster, main = "2D representation of Cluster", shade = TRUE, labels = 5, lines = 2, stand = TRUE)
```

Fonte: O Autor, 2018

Figura 7: Clusters encontrados.



Fonte: O Autor, 2018