Uso do Algoritmo K-means na Ilustração da Técnica de Clusterização: um Estudo de Caso Utilizando R

Introdução

Mineração de dados é o método de investigação de grandes quantidades de dados com o objetivo de encontrar irregularidade, padrões e correlações para arcar a tomada de decisões e possibilitar vantagens estratégicas.

A quantidade de dados produzido está dobrando a cada dois anos. Dados não-estruturados compõem sozinhos 90% do nosso universo digital. Entretanto, mais informação não significa necessariamente mais conhecimento. A mineração de dados nos permite filtrar todo o ruído caótico e repetitivo, entender o que é relevante e, então, fazer bom uso dessa informação para avaliar os prováveis resultados

.

Função executada por técnicas de Mineração de dados

Função executada por técnicas de Mineração de dados

As técnicas de mineração de dados podem ser aplicadas a tarefas como classificação, estimativa, associação, sumarização e segmentação. Essas tarefas são descritas a seguir

Classificação

- Reconhece modelos que descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras. Equivale a construir um modelo de algum tipo que possa ser aplicado a dados não classificados visando categorizá-los em classes.

Estimativa

A estimativa é utilizada para determinar valores para alguma variável contínua desconhecida como, por exemplo, lucro, distância ou saldo na poupança . Ela lida com resultados contínuos, enquanto que a classificação lida com resultados discretos. Ela pode ser usada para executar uma tarefa de classificação, convencionando-se que diferentes intervalos de valores contínuos correspondem a diferentes classes. Como exemplos de tarefas de estimativa tem-se : estimar o número de alunos de uma escola; estimar o lucro total de uma empresa; prever a demanda de consumidores para um novo produto.

Associação

A detecção de relações entre os registros, ocorrências ligadas a um único evento. O exemplo clássico é determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado outro exemplo seria um estudo de modelos de compra em lojas de carros pode revelar que, na compra de um carro automático, 85% das pessoas, querem ele com bancada de couro.

Sumarização

A tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados.

Segmentação (ou Clustering)

Funciona de maneira semelhante a classificação quando ainda não foram definidos grupos. uma técnica em que através de métodos numéricos e a partir somente das informações das variáveis de cada caso, tem por objetivo agrupar automaticamente por aprendizado não supervisionado os n casos da base de dados em k grupos, geralmente disjuntos denominados clusters ou agrupamentos.

Um dos métodos de clusterização é o algoritmo de Análise de Agrupamento k-means um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações.

Algoritmo k-means

1º Escolha dos K elementos que formaram as sementes iniciais.Esta escolha pode ser feita de muitas formas, entre elas:

- selecionando as k primeiras observações;
- selecionando k observações aleatoriamente; e
- escolhendo k observações de modo que seus valores sejam bastante diferentes. Por exemplo, ao se agrupar uma população em três grupos de acordo com a altura dos indivíduos, poderia se escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

2ºCalcular a distância de cada elemente em relação as sementes, agrupando o elemento ao grupo que possuir a menor distância e recalculando o centróide do mesmo. . O processo é repetido até que todos os elementos façam parte de um dos clusters.

3ºApós agrupar todos os elementos, procura-se encontrar uma partição melhor do que a gerada arbitrariamente. Para isto, calcula-se o grau de homogeneidade interna dos grupos através da Soma de Quadrados Residual (SQRes), que é a medida usada para avaliar o quão boa é uma partição.

4ª Após agrupar todos os elementos, procura-se encontrar uma partição melhor do que a gerada arbitrariamente. Para isto, calcula-se o grau de homogeneidade interna dos grupos através da Soma de Quadrados Residual (SQRes), que é a medida usada para avaliar o quão boa é uma partição.

Técnicas de Mineração de dados

Não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A seguir são descritas as técnicas de mineração de dados normalmente usadas.

- **Árvores de decisão –** diagramas que permitem representar e avaliar problemas que envolvem decisões sequenciais, colocando em destaque os riscos e os resultados financeiros identificados nos diversos cursos de ação.
- **Redes neurais –** programas de computador que detectam padrões, fazem previsões e aprendem.
- **Algoritmos Genéticos** Um algoritmo genético é uma técnica de busca utilizada na ciência da computação para achar soluções aproximadas em problemas de otimização e busca

Metodologia

- Base de dados
- Linguagem R

AISP Aérea Integrada de Segurança Pública

Atuação das Delegacias, Batalhões e políticas do Setor.

Metodologia

- Variáveis de interesse:
  - **aisp** - Número da Área Integrada de Segurança Pública;
  - **mes_ano** - Mês e ano da comunicação da ocorrência;
  - **estupro** - Estupro;
  - **hom_culposo** - Homicídio culposo(trânsito);
  - **roubo_veículo** - Roubo de veículo;
  - **hom_doloso** - Homicídio doloso;
  - **tentat_hom** - Tentativa de homicídio;
  - **estelionato** - Estelionato;
  - **roubo_celular** - Roubo de celular;
  - **pessoas_desaparecidas** - Pessoas desaparecidas.

Metodologia

```
library(dplyr)

## Warning:  package 'dplyr' was built under R version 3.4.2
##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union

#Leitura do arquivo da base de dados
base <- read.csv("BaseDPEvolucaoMensalCisp.csv", sep = ';')

#Agrupamento por mês
porMes <- group_by(base, mes_ano) %>%
summarise(estupro=sum(estupro),
celular=sum(roubo_celular), tentHomicidio=sum(tentat_hom),
homDoloso=sum(hom_doloso), homCulposo = sum(hom_culposo),
rouboVeiculo=sum(roubo_veiculo), estelionato=sum(estelionato),
pessDesap=sum(pessoas_desaparecidas))

#Agrupamento por AISP
porAISP <- group_by(base, AISP) %>%
summarise(estupro=sum(estupro), celular=sum(roubo_celular),
tentHomicidio=sum(tentat_hom),homDoloso=sum(hom_doloso),
homCulposo = sum(hom_culposo), rouboVeiculo=sum(roubo_veiculo),
estelionato=sum(estelionato), pessDesap=sum(pessoas_desaparecidas))
```

Metodologia

```
#Principais variáveis estatísticas
summary(porMes)

##     mes_ano       estupro          celular       tentHomicidio
##  2003m1 :  1   Min.   :188.0   Min.   : 301.0   Min.   :243.0
##  2003m10:  1   1st Qu.:276.2   1st Qu.: 529.2   1st Qu.:325.2
##  2003m11:  1   Median :359.0   Median : 656.5   Median :365.0
##  2003m12:  1   Mean   :359.0   Mean   : 781.6   Mean   :388.6
##  2003m2 :  1   3rd Qu.:428.8   3rd Qu.: 913.0   3rd Qu.:443.2
##  2003m3 :  1   Max.   :561.0   Max.   :2548.0   Max.   :645.0
##  (Other):168
##    homDoloso       homCulposo     rouboVeiculo   estelionato
```

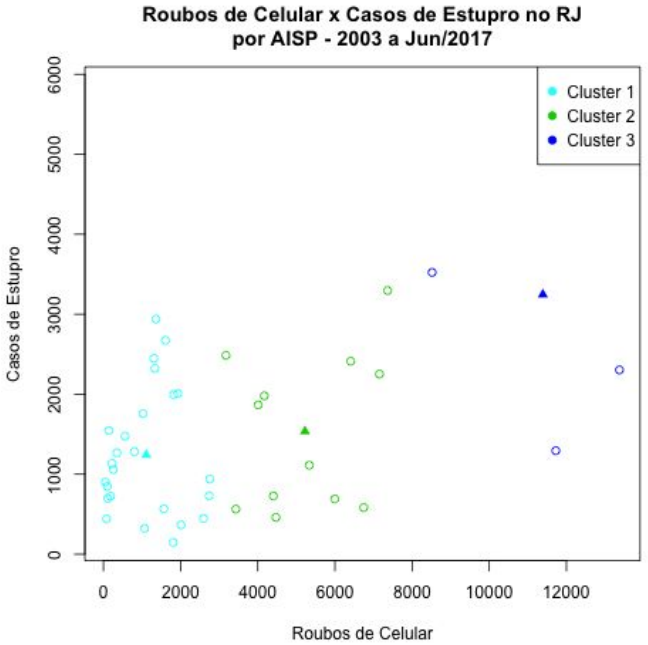
Metodologia

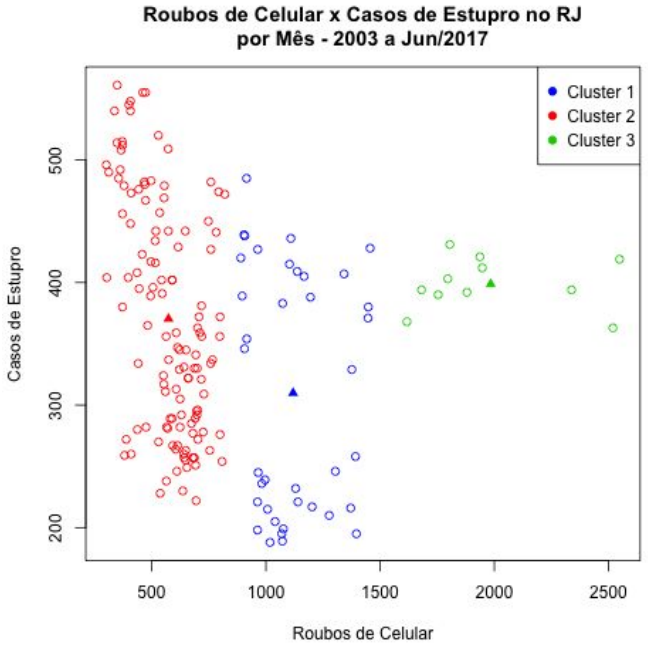
```
##  Min.   :272.0   Min.   :111.0   Min.   :1413   Min.   : 717
##  1st Qu.:377.2   1st Qu.:175.0   1st Qu.:2052   1st Qu.:1561
##  Median :446.0   Median :195.0   Median :2506   Median :2036
##  Mean   :450.7   Mean   :199.1   Mean   :2524   Mean   :2121
##  3rd Qu.:521.8   3rd Qu.:225.8   3rd Qu.:2858   3rd Qu.:2803
##  Max.   :682.0   Max.   :299.0   Max.   :5002   Max.   :3484
##
##    pessDesap
##  Min.   :236.0
##  1st Qu.:390.5
##  Median :436.5
##  Mean   :441.3
##  3rd Qu.:493.8
##  Max.   :628.0
##

summary(porAISP)

##       AISP       estupro        celular      tentHomicidio    homDoloso
##  Min.   : 1   Min.   : 148   Min.   :   52   Min.   : 178   Min.   : 108
##  1st Qu.:11   1st Qu.: 697   1st Qu.:  804   1st Qu.: 824   1st Qu.: 524
##  Median :21   Median :1268   Median : 1926   Median :1381   Median :1477
##  Mean   :21   Mean   :1524   Mean   : 3317   Mean   :1649   Mean   :1913
##  3rd Qu.:31   3rd Qu.:2252   3rd Qu.: 4471   3rd Qu.:2387   3rd Qu.:2651
##  Max.   :41   Max.   :5864   Max.   :13374   Max.   :4419   Max.   :8499
##    homCulposo      rouboVeiculo    estelionato      pessDesap
##  Min.   :  58.0   Min.   :  104   Min.   : 1310   Min.   : 282
##  1st Qu.: 433.0   1st Qu.:  852   1st Qu.: 4685   1st Qu.: 881
##  Median : 668.0   Median : 4796   Median : 6919   Median :1378
##  Mean   : 844.8   Mean   :10711   Mean   : 9003   Mean   :1873
##  3rd Qu.:1126.0   3rd Qu.:14485   3rd Qu.:13119   3rd Qu.:2596
##  Max.   :2498.0   Max.   :57974   Max.   :27322   Max.   :6644
```

Metodologia

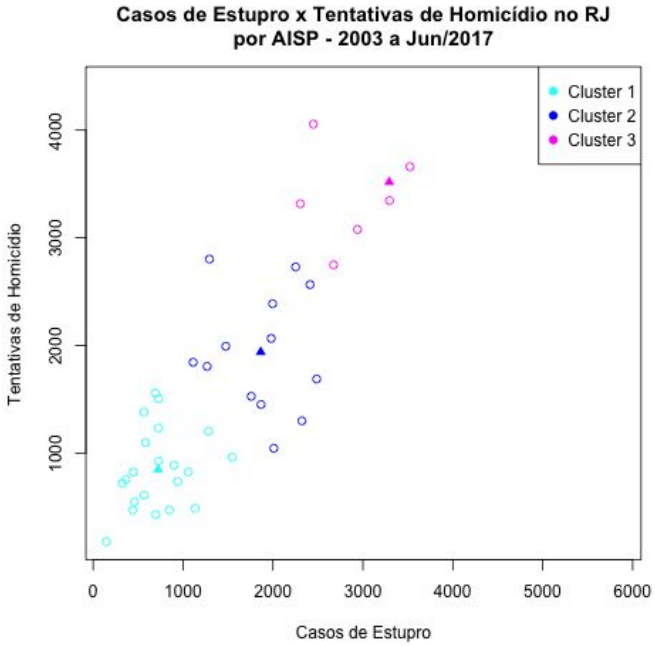
```
#Aplicação do kmeans
kCelularEstuproMes <- kmeans(data.frame(porMes$celular, porMes$estupro),
                             centers = 3)
kCelularEstuproAISP <- kmeans(data.frame(porAISP$celular, porAISP$estupro),
                              centers = 3)

kEstuproTentHomicidioMes <- kmeans(data.frame(porMes$estupro,
      porMes$tentHomicidio), centers = 3)
kEstuproTentHomicidioAISP <- kmeans(data.frame(porAISP$estupro,
              porAISP$tentHomicidio), centers = 3)

kRouboVeiculoTentHomicidioMes <- kmeans(data.frame(porMes$rouboVeiculo,
      porMes$tentHomicidio), centers = 4)
kRouboVeiculoTentHomicidioAISP <- kmeans(data.frame(porAISP$rouboVeiculo,
               porAISP$tentHomicidio), centers = 4)

porMes <- data.frame(porMes, kCelularEstuproMes$cluster,
  kEstuproTentHomicidioMes$cluster, kRouboVeiculoTentHomicidioMes$cluster)
porAISP <- data.frame(porAISP, kEstuproTentHomicidioAISP$cluster,
kEstuproTentHomicidioAISP$cluster, kRouboVeiculoTentHomicidioAISP$cluster)
```

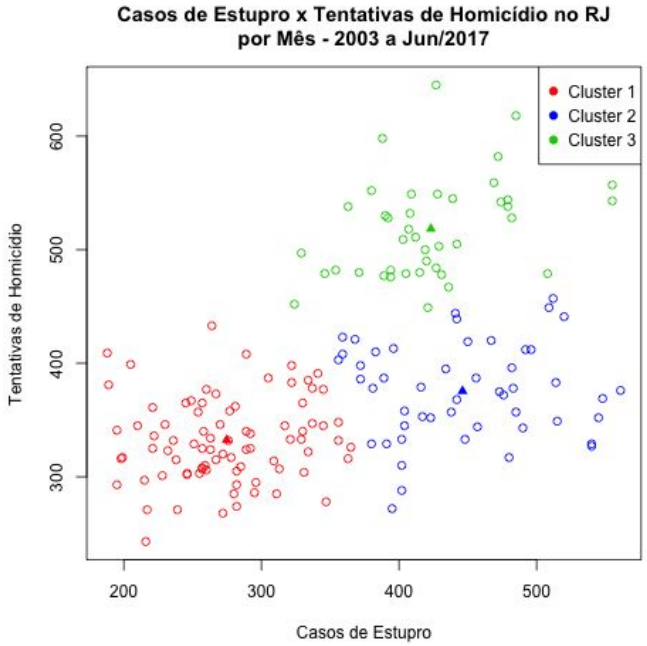
Metodologia



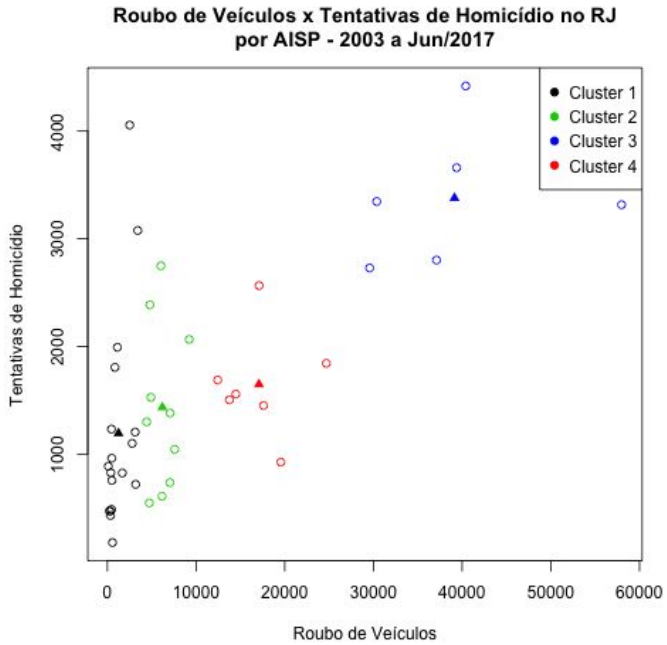


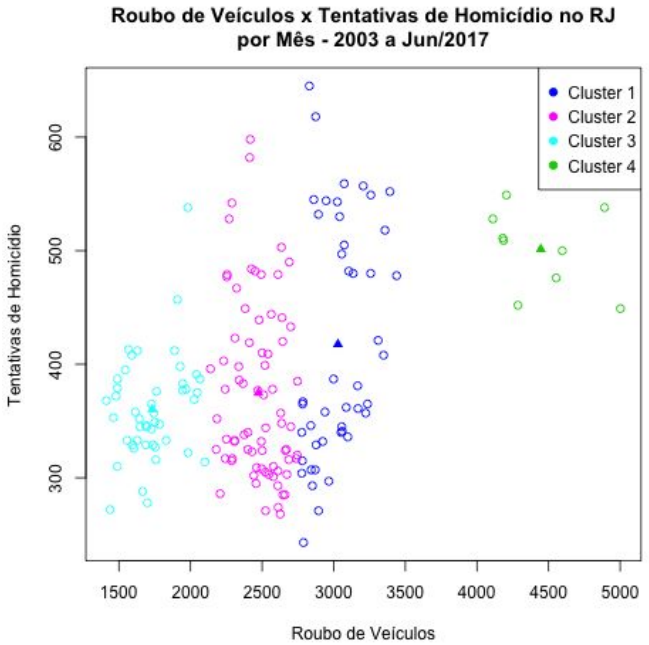
Metodolgia


Casos de Estupro x Tentativas de Homicídio no RJ
por AISP - 2003 a Jun/2017
Cluster 1
Cluster 2
Cluster 3
Tentativas de Homicídio
1000
2000
3000
4000
0
1000
2000
3000
4000
5000
6000
Casos de Estupro


Casos de Estupro x Tentativas de Homicídio no RJ
por Mês - 2003 a Jun/2017
Cluster 1
Cluster 2
Cluster 3
Tentativas de Homicídio
300
400
500
600
200
300
400
500
Casos de Estupro

Metodologia





Conclusão

A técnica de clusterização ***k-means*** é consagrada no meio estatístico como um método eficiente para agrupar dados de acordo com um determinado critério, a saber, a distância euclidiana a um determinado centróide. Utilizada com conhecimento, aliada a um poderoso software estatístico é capaz de extrair informação de onde aparentemente há apenas desordem e desinformação.