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

Eric Calasans de Barros
Fagner Freire de Oliveira

November 27, 2017

Resumo

Diante da grande quantidade de dados que são produzidos hoje em dia se faz necessário que se conheçam técnicas estatísticas e que se tenha habilidade na utilização de ferramentas computacionais para que, assim, se possa transformar essa massiva de dados em informação útil para ser utilizada para o bem da humanidade. Os autores apresentam um estudo de caso de uso da técnica de clusterização conhecida como ***K-means*** e da linguagem **R** para análise de um conjunto de dados referentes aos índices de criminalidade e violência no Estado do Rio de Janeiro.

Abstract

Facing the great amount of data made nowadays it is necessary to know statistics techniques and have expertise in handling computational tools to achieve success in change this massive quantity of data in useful information that can be use for good of humankind. The authors present a case study of application of a clustering technique named ***K-means*** and **R** language in analysis of a dataset concerning to criminal indexes and violence in Rio de Janeiro.

Palavras-chave: Clusterização, K-means, R, cluster.

1 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.

Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e muitas outras aplicações. A mineração de dados pode ser aplicada de duas formas: como um processo de verificação e como um processo de descoberta . No processo de verificação, o usuário sugere uma hipótese acerca da relação entre os dados e tenta prová-la aplicando técnicas como análises estatística e multidimensional sobre um banco de dados contendo informações passadas. No processo de descoberta não é feita nenhuma suposição antecipada. Esse processo usa técnicas, tais como, árvores de decisão, algoritmos genéticos e redes neurais.

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. Empresas de operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares: a classificação pode ajudar a descobrir as características de clientes que provavelmente virão abandoná-las e oferecer um modelo para ajudar os gerentes a prever quem são, de modo que se elabore antecipadamente campanhas especiais para reter esses clientes;
- **Estiimativa ou Regressão** - 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 Clusterização** - 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(kmeans)*** um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações. O algoritmo inicia com a 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;

- 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.

Escolhidas as sementes iniciais, é calculada a distância de cada elemento em relação às sementes, agrupando o elemento ao grupo que possuir a menor distância (mais similar)

e recalculando o centróide do mesmo. O processo é repetido até que todos os elementos façam parte de um dos clusters.

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.

Após o cálculo, move-se o primeiro objeto para os demais grupos e verifica-se se existe ganho na *SQRes*, ou seja, se ocorre uma diminuição no valor da SQRes. Existindo, o objeto é movido para o grupo que produzir o maior ganho, a *SQRes* dos grupos é recalculada e passa-se ao objeto seguinte. Depois de um certo número de iterações ou não havendo mais mudanças, o processo é interrompido.

É de conhecimento público que o Estado do Rio de Janeiro(**RJ**) tem, senão o maior, um dos maiores índices de violência do País. A mídia, por repetidas vezes, mostra não só as cenas de violência constantes no RJ mas também aponta dados estatísticos que corroboram ou complementam a informação capturada pelas câmeras. Para ajudar no combate à violência e como forma de otimizar os recursos que são destinados à Segurança Pública, se faz necessário levantamento de dados referentes e pertinentes a esta matéria, bem como uma análise detalhada e sistematizada dos mesmos, para que sejam transformados em informação útil e confiável.

Neste aspecto, o estudo da Estatística e suas técnicas de análise de dados, aliado ao uso de uma ferramenta computacional que possibilite ao analista acelerar os cálculos e produzir gráficos de qualidade configuram-se nos maiores aliados para esta tarefa.

A **linguagem R(R)** foi criada originalmente no departamento de Estatística da universidade de Auckland, Nova Zelândia, por Ross Ihaka e Robert Gentleman. Oferece uma grande variedade de funções estatísticas e de plotagem de gráficos, bem como oferece uma facilidade na hora de produzir códigos de qualidade e softwares interativos. Pode ser aplicada em várias áreas do conhecimento que requeiram manipulação de dados estatísticos: Ciências Biológicas, Sociais, Exatas, Engenharia, etc.

2 Metodologia

Os dados para análise foram obtidos diretamente do site do Instituto de Segurança Pública do Estado do Rio de Janeiro(`http://www.ispdados.rj.gov.br/Arquivos/BaseDPEvolucaoMensalCisp.csv`), constando de uma base de da-

dos(**DB**) contendo 23185 linhas, dispostas em 61 colunas. De acordo com o site, as Áreas Integradas de Segurança Pública(**AISP**) foram criadas através da **Resolução SSP N. 263 de 27 de julho de 1999**, como parte de uma política de segurança pública que tinha por objetivo estreitar a ligação entre as Polícias Civil e Militar, bem como destas com as comunidades abrangidas pelas AISP através da gestão participativa na identificação e resolução dos problemas locais de segurança pública. A figura abaixo mostra a área de jurisdição de cada AISP no RJ:

Figura 1: Distribuição das AISPs

Para efeitos de estudo foram selecionadas da DB as seguintes variáveis, cuja descrição segue abaixo:

- **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_veiculo** - Roubo de veículo;
- **hom_doloso** - Homicídio doloso;
- **tentat_hom** - Tentativa de homicídio;
- **estelionato** - Estelionato;
- **roubo_celular** - Roubo de celular;
- **pessoas_desaparecidas** - Pessoas desaparecidas.

A seguir os dados foram agrupados em dois subconjuntos de dados: por **mes_ano** e por **AISP**; desta forma a visualização da técnica de clusterização ficou mais fácil de ser perce-

bida.

Utilizando o R pode-se sumarizar os dados nas suas medidas estatísticas(média, máximo, mínimo, desvio-padrão) e se ter uma ideia da estatística descritiva destes conjuntos de dados:

Na sequência foi utilizada uma função intrínseca do R chamada ***kmeans***, que recebe como parâmetros um conjunto de dados e um número específico de centróides(maior que 2) e retorna um conjunto de dados formado por subconjuntos específicos, dos quais os de interesse para o estudo são:

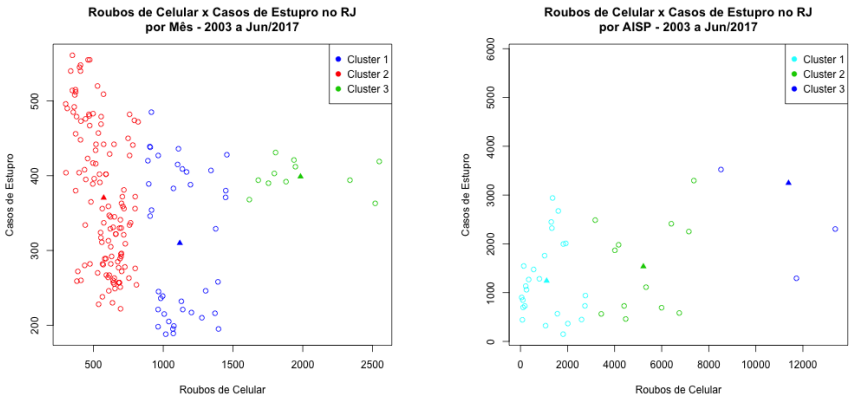
- `clusters` - número do cluster associado ao dado analisado;
- `centers` - coordenadas dos centros passados como parâmetros após a execução da função.

Para cada subconjunto de dados o *kmeans* foi aplicado nos seguintes cruzamentos de dados:

- *Roubo de celular* × *Casos de estupro*;
- *Casos de estupro* × *Tentativa de homicídio*;
- *Roubo de veículo* × *Tentativa de homicídio*.

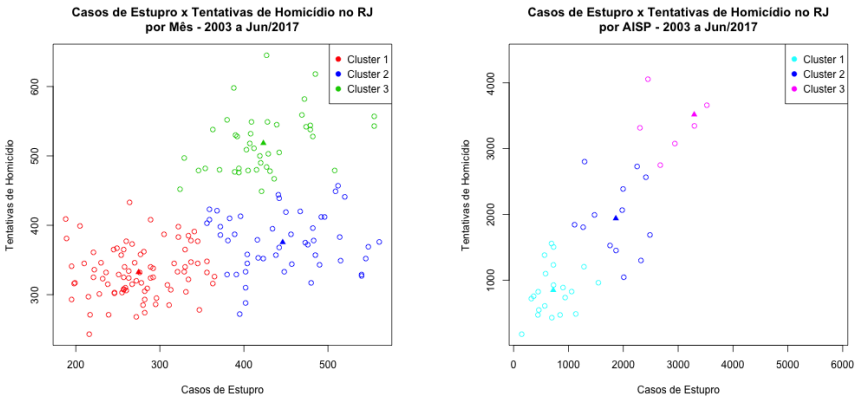
O resultado foi armazendado em variáveis que depois foram acrescentadas aos respectivos subconjuntos de dados.

Após isso foram confeccionados os gráficos onde se pode verificar a eficácia do método em agrupar os dados conforme a distância ao centróide, denotado pelo ponto marcado pelo triângulo.



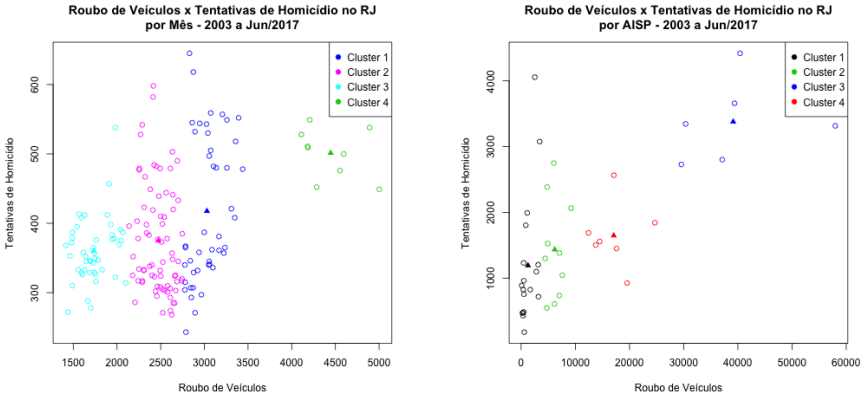
(a) Por mês (b) Por AISP

Figura 2: Roubo de celular × Casos de estupro



(a) Por mês (b) Por AISP

Figura 3: Casos de estupro × Tentativas de homicídio



(a) Por mês (b) Por AISP

Figura 4: Roubo de veículos × Tentativas de homicídio

Nos gráficos relacionados ao subconjunto agrupado por mês cada ponto representa um mês no período de estudo e nos gráficos que aludem ao agrupamento por AISP os pontos representam meses no período.

3 Conclusão

A técnica de clusterização *kmeans* 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.