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

Aguardando parte de Fagner

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 dados(**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.

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 percebida.

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: