Análise de Cluster das Especialidades de Inquéritos

Marcio Ponciano da Silva

Brasília, Brasil

Abstract

O presente trabalho aborda a análise de um conjunto de dados sobre inquéritos policiais com objetivo de verificar a aplicação de artigos de lei em diferentes áreas de especialização. A área de especialização é o contexto de natureza criminal. A proposta testa cinco áreas de especialização, a saber corrupção, defesa institucional, entorpecentes, patrimonial e previdenciário. Para cada artigo de lei, são atribuídos cinco pesos diferentes, um para cada especialização. O conjunto de dados é uma amostra de inquéritos da Polícia Federal descaracterizados para uso aberto. O resultado é a clusterização de inquéritos com base na proximidade dos pesos distribuídos nas especialidades. Foram empregados três métodos de clusterização para melhor compreensão do problema proposto. Os métodos foram K-Means, AgglomerativeClustering e DBSCAN. Foram aplicados os mesmos parâmetros para os métodos empregados. Ao final, foi observado um comportamento de concentração dos dados em razão da proximidade dos pesos, apesar do volume da amostra utilizada.

Keywords: Clusterização, Inquérito Policial, Natureza Criminal,

1. Introdução

Este trabalhou se propôs à análise de um conjunto de dados sobre registros de inquéritos policiais da Polícia Federal. O conjunto de dados abordado se referem apenas a informações que não permitem a identificação do objeto dos inquéritos, apenas a informações numéricas de classificação por pesos de acordo com a especialidade da investigação.

O trabalho apresentou o contexto da existência do inquérito policial no Brasil e sua colocação no ordenamento jurídico. Foi realizada uma abordagem acerca da informação no inquérito policial, sobre o objetivo da informação buscada no escopo do inquérito. Também foi abordado sobre o

tipo de registro encontrado no inquérito e caracterísicas como o tempo de duração e tamanho.

O inquérito policial é um instrumento de toda a polícia brasileira. Todavia, especificamente nesse trabalho foi analisado um conjunto de dados concernete à Polícia Federal, que é uma instituição pública com representação nas vinte e sete unidades federativas do país. Foi abordado ainda acerca do tipo de conteúdo que é produzido no inquérito policial, tais como documentos textuais e periciais.

A abordagem deste trabalho deteve-se ao conjunto de dados de inquéritos extraídos da base de dados da Polícia Federal com algumas de suas caracaterísticas que possibilitem uma análise de clusterização. Segundo Ingwersen [1], a clusterização é um método de técnica da Recuperação de Informação baseado em rede. Essas características do conjunto de dados examinado incluíram a informação da competência de apuração, a lei e artigos correspondentes. Somado a isso também constaram as informações dos pesos de cinco especialidades: corrupção, defesa instituciona, entorpecentes, patrimonial e previdenciário.

O problema enfrentado nesta pesquisa é a sobreposição de leis nos inquéritos conduzidos pelas áreas das cinco especialidades já mencionadas. A amostra trabalhada apresenta cerca de quinze mil registros de inquéritos com a informação da área competente que os conduziu, além de informações de peso que cada área, inclusive a que condiziu, tem sobre aquele tema.

A coleção de dados selecionada para trabalhar nesta pesquisa se refere a uma coleção de inquéritos com instauração em razão de incidência em artigos do Código Penal Brasileiro. De posse desse dataset, a pesquisa foi conduzida para encontar o número ideal de agrupamentos de inquéritos.

2. Procedimentos Metodológicos

A natureza do trabalho é a pesquisa aplicada, pelo que trata de um problema específico. Ele tem como objetivo a pesquisa exploratória, tendo como trabalho buscar conhecimento em uma coleção de inquéritos, com vistas ao agrupamento de informações.

Trata-se de estudo de caso descritivo. Irá examinar o problema detalhadamente. buscando facilitar a sua compreensão. Com relação à forma de abordagem, o método de pesquisa adotado é o quantitativo, pois neste caso os dados numéricos expressaram melhor os resultados. Eles ofereceram métricas capazes de demonstrar o resultado da amostra.

3. A informação no inquérito policial

O inquérito policial brasileiro é um procedimento administrativo utilizado como instrumento das polícias judiciárias que serve a identificar materialidade a autoria dos fatos tipificados como ilítos penais. O instrumento legal que dá legitimidade à existência do inquérito policial no Brasil é o Código de Processo Penal.

A informação no inquérito é o registro dos fatos apurados. O inquérito comunica o resultado da investigação. Ingwersen [1] aborda o conceito de informação na ciência da informação, relacionando-o com a comunicação humana e envolvendo o gerador e o receptor. Ele conclui que a recuperação de informações está relacionada aos processos envolvidos na representação, armazenamento, pesquisa e localização de informações relevantes a um usuário humano.

Presente nas vinte e sete unidades federativas, a Polícia Federal é a instituição que representa a polícia judiciária em âmbito federal. Ela tem seu ról de atribuições precipuamente delineado no art 144, § 1º, da Constituição Federal de 1988. A Polícia Federal está presente tanto nos estados brasileiros como em outros países [2].

Os inquéritos da Polícia Federal são gerenciados por sistema de informação e as informações armazenadas em banco de dados. O inquérito policial é constituído de documentos elaborados no curso da investigação. No jargão jurídico, esses documentos são chamados de peças. Segundo Luiz Flávio Gomes, o inquérito policial é um "conjunto de diligências que visa a apuração do fato punível e de sua autoria" [3].

Na anotação de Michel Misse [4] são muitas peças que compõem o inquérito policial, dentre elas surgem peças textuais e outras periciais. Algumas dessas peças periciais, tem-se o exame cadavérico, necropapiloscópico, exame de local de crime. Já algumas peças textuais são portarias, depoimentos, autos, relatórios e informação de investigação.

Quanto ao tempo de vida, o inquérito policial não tem prazo determinado para encerrar. Também não existe limite de documentos que compõem um inquérito. Na Polícia Federal, o número de páginas de um inquérito é definido em regulamento interno, por meio da Instrução Normativa nº 108/2016, que determina o desmembramento em volumes com aproximadamente duzentas páginas.

A Polícia Federal estima que são instaurados em média 70 mil inquéritos por ano [5]. A divisão de trabalho na Polícia Federal segue a organização

disposta em Regimento Interno, o qual regula as áreas de atuação nas especialidades.

4. A conjunto de inquéritos da Polícia Federal

O conjunto de dados de inquéritos que compõem o corpus desta pesquisa foi extraído do banco de dados da instituição, porém foram descaracterizados para preservar a segurança da informação. Além do que, são dados pouco descritivos, contendo colunas em sua maioria numéricas.

Há diversas especialidades que classificam os inquéritos na Polícia Federal. Essas especialidades representam naturezas de crime, as quais fazem parte do enfrentamento policial. Para este trabalho foram selecionadas cinco especialidades, contendo um conjunto de dados representativo da base de inquéritos. As especialidades escolhidas são:

1. Corrupção
2. Defesa Institucional
3. Entorpecente
4. Patrimonial
5. Previdenciário

Os inquéritos são instaurados com base em leis que descrevem as infrações penais. Cada inquérito registra uma ou mais leis que motivaram a sua instauração. Para fins de produção estatística, essas leis recebem um peso para cada especialidade. Isso significa que embora um inquérito tenha sido conduzido na delegacia da especialidade entorpecente, a lei desse inquérito pode ter classificação de peso diferente em outras especialidades.

Lei	Artigo	Peso-corrupção	Peso-Entorpecente
8666/1993	89	5	1
8666/1993	96	5	0
8666/1993	288	1	0
11343/2006	28	1	2
11343/2006	33	4	2
11343/2006	35	1	3
11343/2006	40	4	2

Tabela 1: Tabela de pesos de lei

A Tabela 1 demonstra o caso de distribuição de pesos em duas leis. Conforme observado, alguns inquéritos foram enquadrados em artigos das Leis 8.666/1993 (lei de licitações) e 11.343/2006 (lei de entorpecentes). Deve se verificar que os pesos para cada artigo são diferentes, isto é, um mesmo artigo de lei recebeu pesos diferentes nas duas especialidades, corrupção e entorpecentes.

Competência	Lei	Artigo
Corrupção	10826/03	12
Defesa Institucional	10826/03	12
Defesa Institucional	10826/03	12
Entorpecentes	10826/03	12
Entorpecentes	10826/03	12
Patrimonial	10826/03	12
Patrimonial	10826/03	12
Patrimonial	10826/03	12
Patrimonial	10826/03	12
Previdenciários	10826/03	12

Tabela 2: Tabela de pesos de lei

Embora os inquéritos policiais recebam um peso em cada especialidade, ele é conduzido apenas por uma dessas áreas de especialidade. Essa será a área de competência do inquérito.

A Tabela 2 representa um conjunto de inquéritos conduzidos por várias áreas de competências (corrupção, defesa institucional, entorpecentes patrimonial e previdenciários), mas todos eles apuram o artigo 12 da Lei nº 10.826/2003. Nesse caso, fica claro que a mesma infração penal pode ser apurada por várias áreas de competências de forma independente.

5. Problema da pesquisa

O problema proposto na pesquisa é verificar a frequência da sobreposição das leis nos inquéritos policiais, com base nas variáveis que determinam os pesos dos artigos de leis para as especialidades envolvidas. Como já foi mencionado, são analisadas cinco especialidades no corpus desta pesquisa.

A abordagem será feita com base nos algoritmos de agrupamento. Os algoritmos de clusterização agrupam um conjunto de documentos em subconjuntos e tem por objetivo criar clusters que sejam coerentes internamente, mas claramente diferentes um do outro [6].

O dataset escolhido para análise representa uma amostra de inquéritos da Polícia Federal. Como registrado, trata-se de informações descaracterizadas. Essa amostra contém pouco mais de quinze mil registros de inquéritos policiais. Esses registros informam a competência de cada inquérito, ou seja, qual a especialidade que conduziu a investigação. Eles têm também o peso que cada inquérito recebeu nas cinco especialidades observadas.

competencia	lei	artigo	peso cor	peso def	peso ent	peso pat	peso pre
Defesa Institucional	CPB	14	12	30	6	12	4
Defesa Institucional	CPB	14	12	30	6	12	4
Patrimonial	CPB	14	12	30	6	12	4
Previdenciário	CPB	14	12	30	6	12	4
Defesa Institucional	CPB	18	0	10	0	0	0
Defesa Institucional	CPB	21	0	10	0	0	0
Defesa Institucional	CPB	29	60	40	6	8	5
Entorpecentes	CPB	33	0	0	9	4	0
Entorpecentes	CPB	33	0	0	9	4	0

Tabela 3: Tabela de registros da amostra

Analisando a Tabela 3, vê-se a demonstração da amostra estudada. O dataset usado tem em particular a característica de conter apenas inquéritos policiais instaurados com base em infrações descritas no Código Penal Brasileiro (CPB).

Variável	Descrição
peso-cor	referente ao peso da especialidade corrupção
peso-def	referente ao peso da especialidade defesa institucional
peso-ent	referente ao peso da especialidade entorpecente
peso-pat	referente ao peso da especialidade patrimonial
peso-pre	referente ao peso da especialidade previdenciário

Tabela 4: Tabela de pesos de lei

Para cada artigo do CPB há um peso correspondente à cada especialidade. Conforme disposto no dataset, as variáveis do modelo são os pesos das cinco especialidades selecionadas para este trabalho. As variáveis estão descritas na forma da Tabela 4.

A proposta do trabalho é encontrar o número ideal de agrupamentos dessas especialidades com base nos pesos que cada uma delas recebeu. A clusterização revela nesse caso o quanto essas especialidades se aproximam em função dos pesos que lhe foram atribuídos. O trabalho dos algoritmos de clusterização é usar um conjunto de intâncias não rotuladas e agrupá-las [7].

6. Resultados da pesquisa

O script do código utilizado neste trabalho e o dataset correspondente estão disponibilizados no GitHub. O link para acesso é `http://github.com/mponcianos/clusterizacao`

Inicialmente, foi utilizado o Elbow Method, ou como é mais conhecido, o método do "cotovelo". Esse método foi útil para definir o número aceitável de clusters.

Foram utilizadas dez interações que verificaram o número de cluster a cada incremento. Dessa forma, é possível analisar os saltos entre as quantidades de clusters que revelará o número de K ótimo.

```
cotovelos = []

for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=0, n_init=10)
    kmeans.fit(X)
    cotovelos.append(kmeans.inertia_)
```

Figura 1: Algorítmo do cotovelo

A Figura 1 demonstra o código utilizado para gerar o incremento que criou o vetor com a interação dos clusters. O código é na linguagem Python e utilizou a biblioteca scikitlearn. O gráfico do "cotovelo"permite visualizar a quantidade relevante de K.

O atributo "inertia_"corresponde ao somatório dos erros quadrátricos das insstâncias de cada cluster. Os valores armazenados nesse atributo demonstram a partir de que número de cluster há acentuada queda.

```
...:     print(kmeans.inertia_)
8514221.990930445
3356595.784738447
1227917.963000776
767338.7486852467
594063.6167991692
423816.06938136875
332976.33217041934
251604.7543645415
192438.03867820383
156993.3418138009
```

Figura 2: Valores do atributo inertia_

Observados os valores de inertia_ na Figura 2, vê-se claramente que a partir do número de 4 clusters não há uma diminuição expressiva do valor. Esse é um método simples, mas eficiente, bastando observar o gráfico e identificar o "cotovelo", como apresentado na Figura 3.

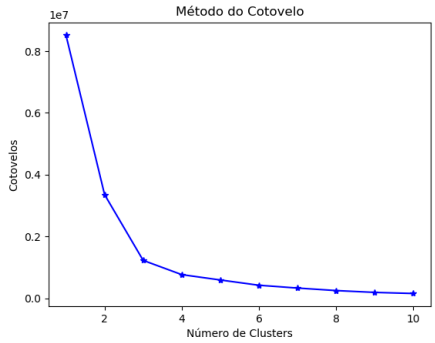

Figura 3: Método do cotovelo

Embora esteja claro o número de 4 clusters, como visto após executar o código do algoritmo do Elbow Method e observados os dados no gráfico, foi testado também outro número de cluster próximo ao "cotovelo", para uma verificação mais apurada.

Além do número de 4 clusters, foi utilizado também o número de 3 clusters. A demonstração desse experimento se baseou nas duas primeiras variáveis.

Foram testados três métodos de clusterização. O primeiro método foi o K-Means, o segundo foi AgglomerativeClustering e por fim o Density Based Spatial Clustering of Applications with Noise (DBSCAN).

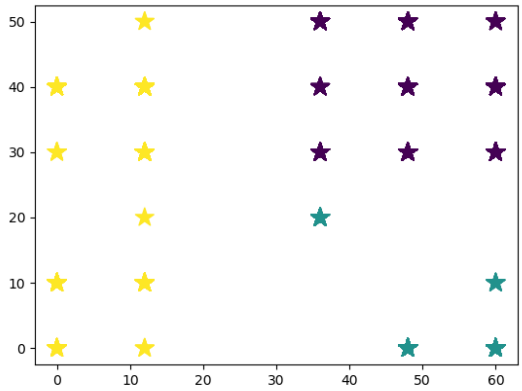

Figura 4: Número de K = 3

A Figura 4 apresenta a aplicação do método do K-Means com 3 clusters. Já a Figura 5 apresenta a aplicação do K-Means com 4 clusters. Nota-se que há uma pequena diferença no surgimento do quarto agrupamento.

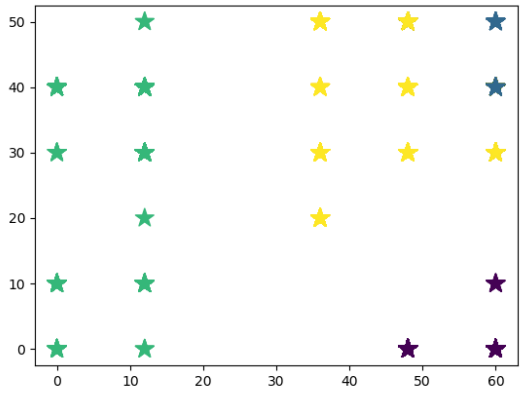

Figura 5: Número de K = 4

Também foi gerado um quadro da distribuição dos pesos da variável peso-cor por cada competência. Assim, verificou-se como essa variável foi distribuída nos clusters.

competencia	0	1	2
Corrupção	593	18	25
Defesa Institucional	1424	170	752
Entorpecentes	48	9	37
Patrimonial	226	4510	99
Previdenciário	7181	57	44

Figura 6: Distribuição com K = 3

As Figuras 6 e 7 demonstram a distribuição dos pesos. Note-se que o agrupamento foi feito na variável competencia.

competencia	0	1	2	3
Corrupção	16	160	25	435
Defesa Institucional	140	263	752	1191
Entorpecentes	6	10	37	41
Patrimonial	4486	110	99	140
Previdenciário	56	6030	44	1152

Figura 7: Distribuição com K = 4

Em seguida, foi testado o método AgglomerativeClustering. Esse método cria clusters hierárquicos. Ele utiliza parâmetro para a estratégia de mesclagem. Assim, ele tenta minimizar a variância de clusters mesclados, a média de distância entre os pares do cluster, ou a distância máxima entre pares de clusters.

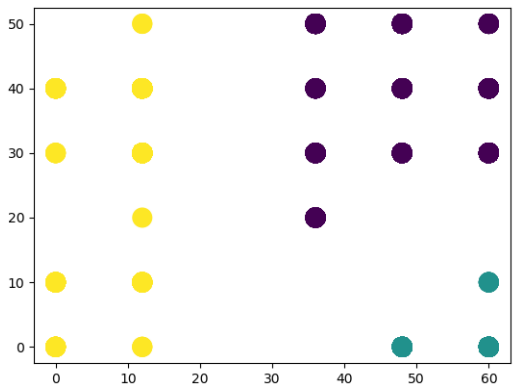

Figura 8: Número de K = 3

As Figuras 8 e 9 demonstram o método AgglomerativeClustering. A demonstração considerou o mesmo número de clusters, sendo K=3 e K=4, respectivamente.

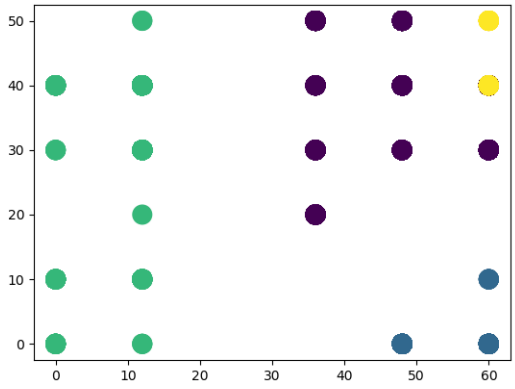

Figura 9: Número de K = 4

Com o segundo método aplicado, já é possível visualizar pequenas alterações em relação ao método do K-Means. O quadro de dispersão desse método apresenta essa diferença.

	0	1	2
competencia			
Corrupção	595	16	25
Defesa Institucional	1454	140	752
Entorpecentes	51	6	37
Patrimonial	250	4486	99
Previdenciário	7182	56	44

Figura 10: Distribuição com K = 3

As Figuras 10 e 11 demonstram a distribuição dos pesos no método AgglomerativeClustering. Note-se que, da mesma forma aplicada no método do K-Means, o agrupamento foi feito na variável competencia. Estão apresentados os agrupamentos para 3 e 4 clusters.

competencia	0	1	2	3
Corrupção	449	16	25	146
Defesa Institucional	1200	140	752	254
Entorpecentes	41	6	37	10
Patrimonial	140	4486	99	110
Previdenciário	1183	56	44	5999

Figura 11: Distribuição com K = 4

Por fim, foi testado o método DBSCAN. Esse algoritmo pode identificar clusters em grandes conjuntos de dados espaciais observando a densidade local dos elementos do banco de dados.

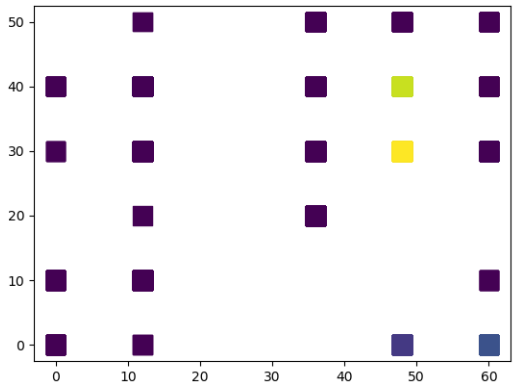

Figura 12: Densidade com 120 vizinhos

As Figuras 12 e 13 demonstram o método DBSCAN. Como se verifica, foi testado o algoritmo com dois valores para o parâmetro min_samples. Esse parâmetro especifica o número mínimo de vizinhos a formar um cluster. Os valores utilizados foram 60 e 120.

Foi experimentada uma variação de valores para o parâmetro min_samples. Todavia, as variações não foram expressivas, mantendo a demonstração com 120 e 60 vizinhos como a melhor apresentação.

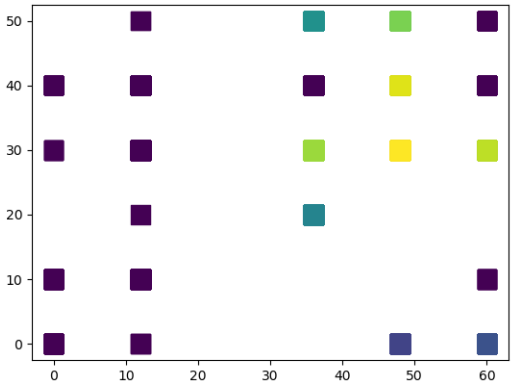

Figura 13: Densidade com 60 vizinhos

A demonstração desses três métodos apresentam como os dados se concentram em razão da proximidade dos pesos atribuídos às variáveis correspondentes às especializações. Foi garantida a aplicação dos mesmos parâmetros para permitir uma efetiva comparação entre os resultados, a fim de encontrar o agrupamento mais adequado ao caso em estudo.

O estudo indica essa concentração de dados em razão dos valores próximos. A demonstração dos três métodos apresenou poucas variações, em que pese o volume da amostra utilizada ter uma quantidade expressiva de registros.

7. Considerações finais

Esta pesquisa buscou abordar a aplicação de métodos ds clusterização para agrupar inquéritos policiais em função do pesos distribuídos em cinco variáveis que representam cinco áreas de especialidades. Para cobrir informações de negócio, foi apresentado o contexto em que o inquérito policial se insere. O dataset utilizado representa um conjunto de dados de inquéritos da Polícia Federal. Esse conjunto de dados tem pouco mais de quinze mil registros.

O problema proposto teve o objetivo de verificar a sobreposição das leis nos inquéritos policias, com base nas variáveis de pesos que foram atribuídos às áreas especializadas. Cada inquérito policial é enquadrado num determinado artigo de lei e, embora seja conduzido por uma área especializada, ele recebe peso em cada especialidade Isso ocorre por causa da sobreposição de competências nos mesmos incidentes penais.

Para responder essa questão, o dataset foi submetido a três métodos de clusterização. O primeiro método empregado foi o K-Means, utilizado para criar 3 e 4 clusters, permitindo a observação da capacidade de agrupamento. Posteriormente, foi aplicado o método AgglomerativeClustering, também com 3 e 4 clusters. Por fim, foi aplicado o método DBSCAN com agrupamento por 120 e 60 vizinhos.

Conforme se analisou a amostra, a classificação de pesos dos inquéritos policiais não são valores muito distantes. Isso permite que o agrupamento concentre diversos valores em pontos sobrepostos. Com essa concenração fica evidente a sobreposição de competências, pelo menos na amostra examinada.

[1] P. Ingwersen, Information Retrieval Interaction, Taylor Graham Publishing, 1992.

[2] P. Federal, Unidades e contatos, `http://www.pf.gov.br/institucional/unidades`, 2018. Acessado em 16 nov. 2018.

[3] L. F. Gomes, Direito Processual Penal, Revista dos Tribunais, 2005.

[4] M. Misse, O inquérito policial no Brasil: uma pesquisa empírica, NECVU/IFCS/UFRJ; BOOKLINK, 2010.

[5] P. Federal, Polícia federal lança sistema de inquérito eletrônico, `http://www.pf.gov.br/agencia/noticias/2016/10/policia-federal-lanca-sistema-de-inquerito-eletronico`, 2016. Acessado em 28 nov. 2018.

[6] C. D. Manning, P. Raghavan, H. Schutze, Introduction to information retrieval, `https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf`, 2009. Acessado em 21 nov. 2018.

[7] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in pratice, `http://ciir.cs.umass.edu/downloads/SEIRiP.pdf`, 2015. Acessado em 05 nov. 2018.