Análise de Agrupamento de Acidentes de Trânsito em Rodovias de São Paulo

Andrei Inoue Hirata
UNESP
andreihirata@unesp.br

RESUMO

Este trabalho aplica técnicas de agrupamento, uma abordagem de aprendizado não supervisionado, para identificar perfis de acidentes de trânsito nas rodovias federais de São Paulo. Utilizando dados da Polícia Rodoviária Federal (PRF) de 2023, o estudo explora os algoritmos K-Means e Agrupamento Hierárquico Aglomerativo para segmentar as ocorrências em grupos com características similares. A determinação do número ótimo de clusters é realizada através do Método do Cotovelo e do Coeficiente de Silhueta. Os modelos são avaliados e comparados utilizando tanto uma métrica de validação interna (Coeficiente de Silhueta) quanto uma externa (Índice Rand Ajustado), usando o 'tipo_acidente' como referência. Os resultados demonstram a capacidade do agrupamento em revelar padrões latentes nos dados, como a formação de clusters distintos para acidentes urbanos diurnos e acidentes noturnos em pistas simples, fornecendo insights valiosos para a segurança viária.

Palavra Chaves

Mineração de Dados, Agrupamento, K-Means, Agrupamento Hierárquico, Validação de Cluster, Cidades Inteligentes, Acidentes de Trânsito

1. INTRODUÇÃO

O agrupamento, ou clustering, é uma tarefa central da mineração de dados que visa particionar um conjunto de dados em grupos (clusters), de modo que os objetos dentro de um mesmo grupo sejam mais similares entre si do que com os de outros grupos [2]. Diferente da classificação, é uma técnica de aprendizado não supervisionado, pois não utiliza rótulos pré-definidos.

Este estudo aplica algoritmos de agrupamento no contexto de Cidades Inteligentes para analisar dados de acidentes da PRF [1]. O objetivo não é prever uma classe, mas sim descobrir se existem "perfis" ou segmentos naturais de acidentes. Essa segmentação pode auxiliar na compreensão das causas e na formulação de políticas de prevenção mais direcionadas.[1]

2. METODOLOGIA

2.1 Dataset e Pré-processamento

Utilizou-se o dataset da PRF de 2023, filtrado para o estado de São Paulo. Para preparar os dados para algoritmos baseados em distância [3], as seguintes etapas foram executadas:

- **Seleção de Features:** Foram selecionados atributos relevantes como 'dia_semana', 'uso_solo', 'hora_do_dia' (criada por discretização), 'condicao_metereologica' e 'tipo_pista'.
- **Encoding:** Variáveis categóricas foram transformadas em formato numérico usando One-Hot Encoding.
- **Escalonamento:** Todas as features foram padronizadas com 'StandardScaler' para que tivessem média 0 e desvio padrão 1, garantindo que nenhuma variável dominasse o cálculo da distância devido à sua escala.

2.2 Algoritmos e Avaliação

Foram explorados dois algoritmos de agrupamento:

- **K-Means:** Um algoritmo particional que agrupa os dados em k clusters pré-definidos. O número k foi estimado com o Método do Cotovelo e o Coeficiente de Silhueta.
- **Agrupamento Hierárquico Aglomerativo:** Um método que constrói uma hierarquia de clusters, visualizada através de um dendrograma.

A avaliação dos modelos foi realizada com duas métricas:

- **Coeficiente de Silhueta (Interna):** Mede a coesão e separação dos clusters, variando de -1 a 1. Valores mais altos indicam clusters mais bem definidos.
- **Índice Rand Ajustado (ARI - Externa):** Compara os clusters encontrados pelo algoritmo com uma classificação real (neste caso, o 'tipo_acidente'), medindo a similaridade entre as duas partições.

[1]O código-fonte, o dataset e as imagens geradas neste trabalho estão disponíveis em: `https://github.com/gitandreihirata/Unesp_Doutorado/tree/main/Mineracao%20de%20Dados/Exercicios/Ex3`

3. RESULTADOS E DISCUSSÃO

3.1 Definição do Número de Clusters (k)

As Figuras 1 e 2 mostram os resultados para a escolha do número de clusters para o K-Means. O Método do Cotovelo sugere um 'k' entre 3 e 5, enquanto o Coeficiente de Silhueta apresenta um pico em 'k=3', indicando que 3 clusters é a configuração que melhor equilibra coesão e separação. Portanto, foi adotado $k = 3$ para os experimentos.

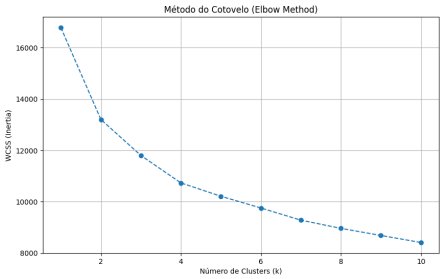

Figure 1: Método do Cotovelo para estimar o número de clusters.

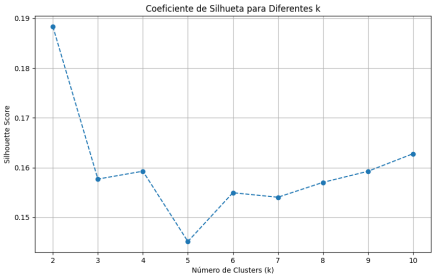

Figure 2: Coeficiente de Silhueta para diferentes valores de k.

3.2 Análise do Agrupamento Hierárquico

O dendrograma gerado (Figura 3) também sugere a formação de 3 clusters principais, corroborando a análise do K-Means.

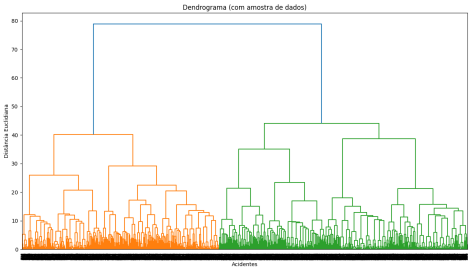

Figure 3: Dendrograma do Agrupamento Hierárquico.

3.3 Comparação dos Modelos

Os dois algoritmos foram treinados para formar 3 clusters e os resultados comparados na Tabela 1.

Table 1: Resultados das Métricas de Validação

Algoritmo	Silhueta	ARI (Externo)
K-Means	0.178	0.104
Hierárquico	0.134	0.089

O K-Means obteve um desempenho superior em ambas as métricas. O Coeficiente de Silhueta (0.178) indica uma separação moderada dos clusters. O ARI (0.104) sugere uma baixa, mas existente, concordância entre os clusters encontrados e os tipos de acidente oficiais, o que é esperado, já que o algoritmo usou muito mais informações para criar os grupos.

A Figura 4 visualiza os clusters encontrados por ambos os algoritmos, após redução de dimensionalidade com PCA.

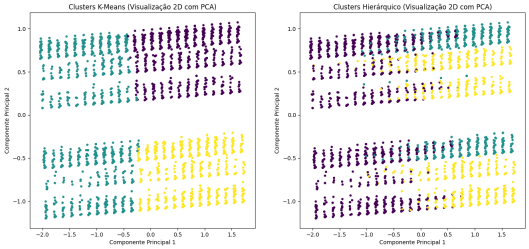

Figure 4: Visualização 2D dos clusters (K-Means à esquerda, Hierárquico à direita).

3.4 Análise dos Clusters (K-Means)

Analisando os centroides dos 3 clusters encontrados pelo K-Means, foi possível caracterizá-los:

- **Cluster 0:** Acidentes predominantemente em área urbana ('uso_solo_Urbano' alto), durante o dia ('periodo_dia_Manhã' e 'Tarde' altos) e em pista dupla.
- **Cluster 1:** Acidentes em pista simples, com maior incidência à noite e de madrugada.
- **Cluster 2:** Acidentes em área rural ('uso_solo_Rural' alto), com maior ocorrência de tempo chuvoso.

4. CONCLUSÃO

Este trabalho demonstrou com sucesso a aplicação de algoritmos de agrupamento para a segmentação de acidentes de trânsito. O K-Means, configurado com 3 clusters, se mostrou o modelo mais eficaz, gerando grupos com características distintas e interpretáveis. A avaliação interna e externa confirmou a validade da segmentação encontrada.

Conclui-se que o agrupamento é uma técnica poderosa para a descoberta de conhecimento em dados não supervisionados, capaz de revelar perfis de acidentes que não são imediatamente óbvios. Esses insights podem ser utilizados por gestores de tráfego para criar estratégias de prevenção mais focadas e eficientes, contribuindo para a segurança nas rodovias.

5. REFERENCES

[1] Polícia Rodoviária Federal. (2025). *Dados Abertos da PRF*. Acessado em 6 de outubro de 2025, de `https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf`.

[2] Carvalho, V. O. (s.d.). *Notas de Aula: Agrupamento*. UNESP.

[3] Carvalho, V. O. (s.d.). *Notas de Aula: Medidas de Dissimilaridade e Similaridade*. UNESP.