Metodologia de Agrupamento para resolução de problemas ligados ao rompimento de barragens

Carlos Henrique Tavares Brumattil[1], Mariana Albuquerque Reynaud Schaefer[1], Julio Cesar Soares dos Reis[1], Jugurta Lisboa-Filho[1]

[1] Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brasil

{carlos.h.tavares, mariana.schaefer, jreis, jugurta}@ufv.br

Abstract. *Environmental disasters, like dam failures, cause impacts that go beyond the area of occurrence. From the region of origin to its arrival at sea, the sediments can cause both environmental and economic impacts. Searching for ways to help in the recovery of these degraded areas, this work proposes the development of a methodology it utilize a Knowledge Discovery in Database Process (KDD) in order to group cities close to hydrographic basins, thus presenting the generation of knowledge groups with same characteristics, thus facilitating the allocation of resources.*

Resumo. *Desastres ambientais, como o rompimento de barragens, causam impactos que vão muito além da área de ocorrência. Da região de origem até a sua chegada ao mar, os resíduos podem causar tanto impactos ambientais quanto econômicos. Buscando formas de auxiliar na recuperação dessas áreas degradadas, neste trabalho, é proposto o desenvolvimento de uma metodologia utilizando o processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Database – KDD) a fim de agrupar cidades próximas às bacias hidrográficas, gerando grupos de cidades com características parecidas, facilitando dessa forma à alocação de recursos.*

1. Introdução

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 20151, evidenciou a íntima relação existente entre o meio biótico e o meio antrópico. Segundo [Godoy and Dias 2021], o impacto causado pelo desastre ambiental provocou sérias consequências não somente na região de ocorrência, mas em todo o percurso dos rejeitos até a sua chegada ao mar. Estas consequências não se restringem somente às questões ambientais, mas também socioeconômicas, políticas e humanas.

Diante deste cenário, soluções computacionais podem ser úteis para apoiar especialistas no processo de tomada de decisão relacionado à desastres ambientais. Uma das maneiras existentes para isso é através do processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database* - KDD), principalmente na etapa de Mineração de Dados. Segundo [Camilo and Silva 2009], tais técnicas possibilitam a resolução de tarefas como Descrição, Classificação, Estimação, Predição, Agrupamento e Associação.

O texto foi escrito conforme detalhado a seguir. A Seção 2 apresenta a fundamentação teórica. Na Seção 3 é descrita a metodologia desenvolvida. A Seção apresenta os resultados e discussões até agora alcançados. Por fim, na Seção 5 é apresentado as conclusões e trabalhos futuros a serem realizados.

2. Fundamentação Teórica

Das várias tarefas que a Mineração de Dados se propõe a resolver, destaca-se o Agrupamento. De acordo com [Jain et al. 1999], considerando um conjunto de dados, as técnicas voltadas para tal tarefa buscam gerar agrupamentos ou clusters baseados na similaridade dos elementos contidos em um mesmo grupo. Essa similaridade referida é um critério que define o quanto dois ou mais elementos são semelhantes e, consequentemente, devem pertencer a um mesmo conjunto gerado.

Utilizando o aprendizado não-supervisionado, isto é, não sendo necessário fornecer um conjunto prévio de dados para treinamento, fazendo com que o algoritmo seja executado direto sobre o conjunto de dados de interesse, considerando os algoritmos existentes, os mais tradicionais, segundo [Cassiano 2014], podem ser classificados como na Figura 1 a seguir. No primeiro nível da classificação, os algoritmos se dividem em relação a abordagem: hierárquica ou particional. A abordagem hierárquica, segundo ZHANG et al. (1996, apud CASSIANO, 2010), se caracteriza por manter o par de dados mais próximo juntos. Já a abordagem particional divide a base de dados em k grupos, sendo o número k um valor informado pelo usuário, de acordo com ESTER et al. (1996, apud CASSIANO, 2010). No segundo nível, tem-se a subdivisão da abordagem de acordo com a forma de medição da similaridade.

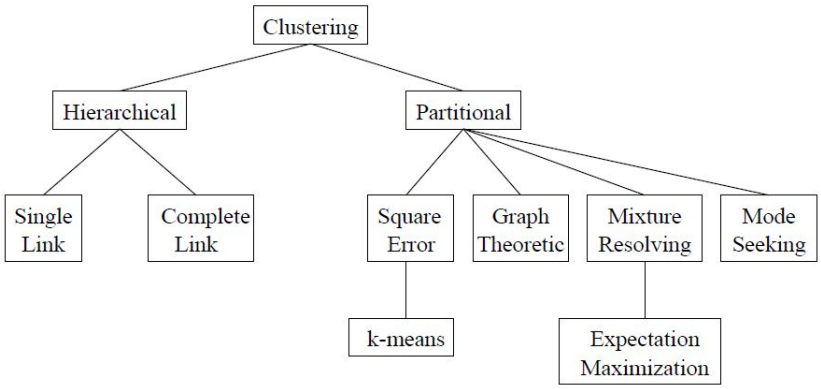

Figura 1. Classificação dos algoritmos de clusterização. Fonte: [Jain et al. 1999]

2.1. Algoritmo *K-Means*

O Algoritmo *K-Means* é um algoritmo clássico e bastante explorado na literatura . Por ser mais simples de implementar e de baixa complexidade, segundo [Lachi and da Rocha 2005], é comumente empregado para a geração de grupos. Seu pseudocódigo pode ser visto na Figura 2 a seguir. Além disso, o algoritmo foi proposto pela primeira vez em 1967, e se baseia na tentativa de minimizar o erro quadrático calculado associado a distância entre cada elemento e o centróide do seu respectivo grupo, como medida de similaridade.

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Figura 2. Pseudocódigo do algoritmo *K-Means*. Disponível em: https://realpython.com/k-means-clustering-python/. Acessado em: 01 de set. de 2022

3. Metodologia

Conforme a metodologia de trabalho proposta (Figura 3), destaca-se também a associação das etapas de trabalho com cada uma das 04 etapas do processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database* – KDD), segundo [Miller and Han 2009].

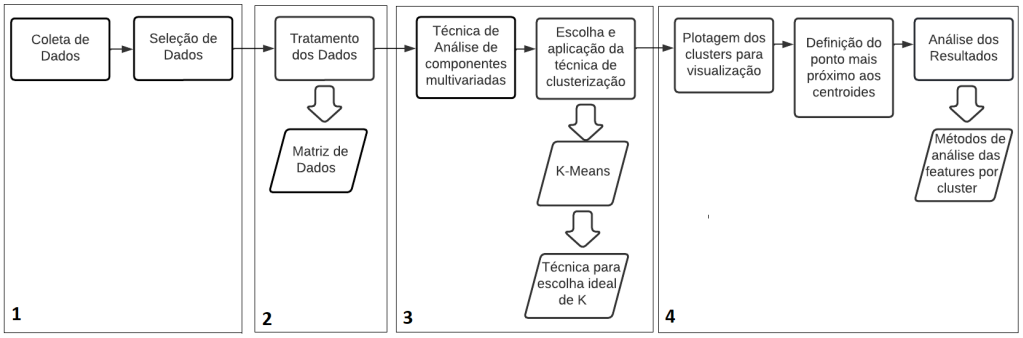

Figura 3. Visão geral da metodologia de trabalho.

Sendo assim, a etapa (1) é denominada de Seleção. Nela, ocorre a coleta e a seleção dos dados de interesse. Para esse processo, no trabalho aqui descrito, optou-se pelo uso dos dados do Cadastro Ambiental Rural (CAR) e do Instituto Brasileiro de Geografia e Estatística (IBGE). Verificou-se que os dados do CAR são pouco explorados em trabalhos desse tipo, enquanto o IBGE é uma fonte importante devido ao grande volume de informações presentes. Após a coleta dos dados, é realizado a seleção dos dados mais interessantes para o processo. Nesse momento, é interessante a presença de um usuário especialista para verificar quais dados possuem de fato impacto na regra de negócio a ser estudada.

Na próxima etapa, denominada de Pré-Processamento (2), os dados que foram selecionados com ajuda do usuário especialista passam pelo processo de tratamento. Assim, dados de diversas fontes e formatos são agora ajustados conforme a necessidade do algoritmo a ser utilizado. Por exemplo, para a metodologia proposta, os dados categóricos são convertidos em dados numéricos e os dados em ponto flutuante, dependendo do formato de entrada, precisam agora serem adequados para o formato aceitável. Ao final dessa etapa, os dados passam a serem armazenados em uma matriz de dados.

Na sequência, na etapa (3), ocorre a Redução e Projeção dos dados. Como para a metodologia aqui descrita é usado o algoritmo *K-Means* para a geração dos agrupamentos, agora é aplicado uma técnica de análise de componentes multivariadas, fazendo com que a matriz de dados, gerada na etapa anterior, seja convertida agora em uma matriz dimensional. Para isso, utilizou-se o *Principal Component Analysis* (PCA) [Bro and Smilde 2014]. Como o *K-Means* é um algoritmo de clusterização particional, o valor do número de agrupamentos k foi determinado usando o *Silhouette Score* [Rousseeuw 1987], que de maneira geral captura a consistência dentro de um determinado agrupamento de dados. É importante mencionar que, embora tenham sido aplicados algoritmos e/ou técnicas específicas baseadas em critérios pré-definidos, a metodologia é generalizável. Em outras palavras, caso seja de interesse do usuário, os algoritmos e/ou técnicas explorados em alguns componentes podem ser alterados.

Por fim, na etapa 4, logo após a geração de todos os agrupamentos e a determinação de seus centroides, ocorre a visualização dos dados. Dessa forma, os conjuntos formados podem ser vistos, assim como a distância inter-conjuntos e intra-conjunto. Nesta etapa é definido também o ponto intra-conjunto mais próximo do centroide determinado. Assim, esse ponto é o elemento mais indicado para ser um representante do agrupamento, ou seja, um elemento que possuirá as propriedades mais características do conjunto a que pertence, já que é o ponto de maior similaridade do conjunto.

4. Resultados e Discussão

Refletindo a metodologia anteriormente descrita, iniciou-se a implementação de um Sistema de Inteligência Geográfica, conforme a arquitetura da Figura 4. Nela, destaca-se o desenvolvimento de uma plataforma *web* para interação com o usuário.

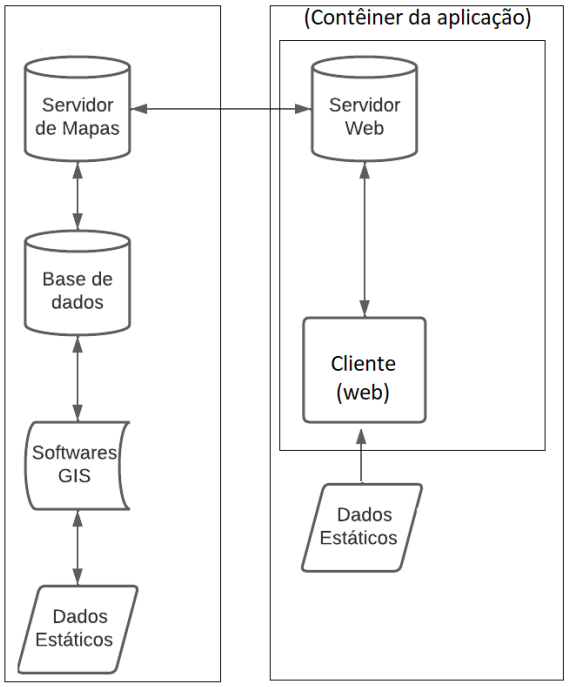

Figura 4. Arquitetura geral proposta.

Para a construção dessa plataforma *web*, utilizou-se tecnologias como *HyperText*

Markup Language (HTML), *Cascading Style Sheets* (CSS) e Javascript. Já para o *backend*, usou-se o *Python*, uma vez que os principais algoritmos para uso em Mineração de Dados já se encontram implementados. Além disso, o sistema como um todo é conectado a um servidor de mapa, provendo assim para a aplicação dados nos formatos *Web Map Service* (WMS) e *Web Feature Service* (WFS), conforme especificações da *Open Geospatial Consortium* (OGC)[1].

5. Conclusão

Este artigo apresentou uma metodologia genérica que busca auxiliar os usuários responsáveis pelo processo de tomada de decisão, seja ele pertencente a um órgão público ou empresa privada, durante a ocorrência de desastres ambientais ligados à bacias hidrográficas e que causam impacto direto nas cidades próximas. Espera-se que o trabalho proposto possa ser útil para apoiar decisões relacionadas à esses desastres ambientais provendo aos especialistas informações que possam suportar decisões neste contexto. Como trabalhos futuros, planeja-se continuar a implementação da ferramenta *opensource*, adicionando novas funcionalidades. Além disso, planeja-se também testar outros algoritmos de Agrupamento, avaliando os resultados encontrados para assim verificar qual é de fato a melhor abordagem a ser utilizada.

Referências

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9):2812–2831.

Camilo, C. O. and Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, 1(1):1–29.

Cassiano, K. M. (2014). Análise de séries temporais usando análise espectral singular (ssa) e clusterização de suas componentes baseada em densidade. *Pontifícia Universidade Católica do Rio de Janeiro*.

Godoy, S. M. and Dias, M. B. (2021). O desastre ambiental de mariana e o papel da fundação renova na reparação dos danos. *Direito e Desenvolvimento*, 12(1):37–48.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Lachi, R. L. and da Rocha, H. V. (2005). Aspectos básicos de clustering: conceitos e técnicas. *Núcleo de Informática Aplicada à Educação (Nied), UNICAMP-Instituto de Computação–Universidade Estadual de Campinas*.

Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC press.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[1] https://www.ogc.org/