Uma Metodologia para Apoio à Tomada de Decisão em Cenários de Desastres Ambientais Envolvendo Bacias Hidrográficas

Anônimo

1

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 environmental and economic impacts. Searching for ways to help in the recovery of these degraded areas, this work proposes the development of a methodology that users 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. We hope that the proposed framework can be useful to support expert decision-making in this context.*

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 semelhantes, facilitando dessa forma a alocação de recursos. Acreditamos que o arcabouço proposto possa ser útil para suportar a tomada de decisões de especialistas neste contexto.*

1. Introdução

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 2015, evidenciou a íntima relação existente entre o meio biótico e o meio antrópico. Segundo Godoy e Dias [Godoy and Dias 2021], o impacto causado pelo referido desastre provocou sérias consequências não somente na região de ocorrência, mas em todo o percurso dos rejeitos até a sua chegada ao mar. Ademais, estas consequências não se restringem somente às questões ambientais, mas também socioeconômicas, políticas e humanas.

Nesse contexto, um dos problemas identificados, por exemplo, está relacionado a alocação de recursos financeiros, de modo adequado e eficiente, para a recuperação das áreas afetadas. Segundo Barbosa et al [Barbosa et al. 2015], após o desastre, estima-se que as prefeituras das áreas envolvidas terão que gastar cerca de R$150 milhões, além de que há uma proposta da criação de um fundo de US$20 bilhões, ao longo de 10 anos, pelas empresas envolvidas.

Diante deste cenário, soluções computacionais podem ser úteis para apoiar especialistas no processo de tomada de decisão relacionado a 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 e Silva [Camilo and Silva 2009], tais técnicas possibilitam a resolução de tarefas como Descrição, Classificação, Estimação, Predição, Agrupamento e Associação. Fundamentado em KDD, neste trabalho é proposta um arcabouço que permite agrupar cidades atingidas por desastres ambientais e que são localizadas próximas à bacias hidrográficas. Acreditamos que a metodologia proposta possa ser útil para suportar a tomada de decisões de especialistas neste contexto, como por exemplo, no processo de decisão de cidades (ou áreas) atingidas por desastres ambientais e que, em detrimento de alguma característica (em comum), cuja recuperação deve ser priorizada.

O texto foi escrito conforme detalhado a seguir. A Seção 2 apresenta a fundamentação teórica. Na Seção 3 é descrita o arcabouço desenvolvido. A Seção 4 apresenta os resultados e discussões preliminares. Por fim, na Seção 5 são apresentadas as conclusões e trabalhos futuros.

2. Fundamentação Teórica

Das várias tarefas que a Mineração de Dados se propõe a resolver, destaca-se o Agrupamento. Segundo Jain et al. [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 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 [Cassiano 2014], podem ser classificados como ilustrado na Figura 1. No primeiro nível da classificação, os algoritmos se dividem em relação a abordagem: hierárquica ou particional. A abordagem hierárquica, também segundo Cassiano [Cassiano 2014], 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, conforme Cassiano [Cassiano 2014]. No segundo nível, tem-se a subdivisão da abordagem de acordo com a forma de medição da similaridade. Das várias técnicas de Agrupamento existentes, optou-se pelo *K-Means* como base para a metodologia aqui proposta, conforme detalhado a seguir.

2.1. Algoritmo *K-Means*

O Algoritmo *K-Means* é um algoritmo clássico e bastante explorado na literatura e foi proposto em 1967. Por ser mais simples de implementar e de baixa complexidade, é comumente empregado para a geração de grupos. Seu pseudocódigo pode ser visto na Figura 2. Este algoritmo 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.

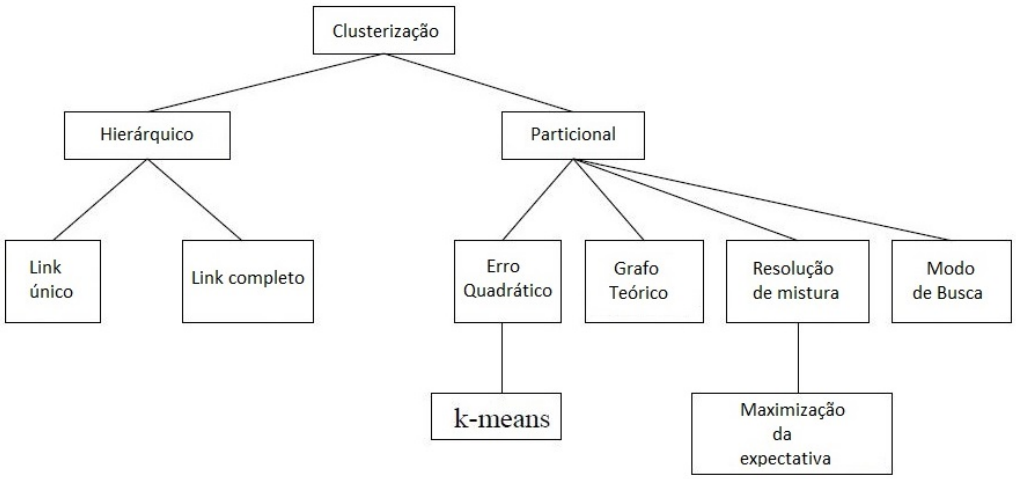

Figura 1. Classificação dos algoritmos de clusterização. Fonte: [Jain et al. 1999]

Algoritmo 1: Algoritmo KMeans

1: Especifique o número k de agrupamentos para gerar.
2: Inicialize os k centróides de forma randômica.
3: Repita
4: Atribua cada ponto ao centróide mais próxima
5: Calcule o novo valor do centróide (média) de cada agrupamento
6: Até a posição do centróide não mudar

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 Proposta

Uma visão geral da metodologia proposta neste trabalho é apresentada na Figura 3. Nela, é realizada também a associação das etapas de trabalho proposta com cada uma das 4 etapas do KDD apresentadas por Miller e Han [Miller and Han 2009].

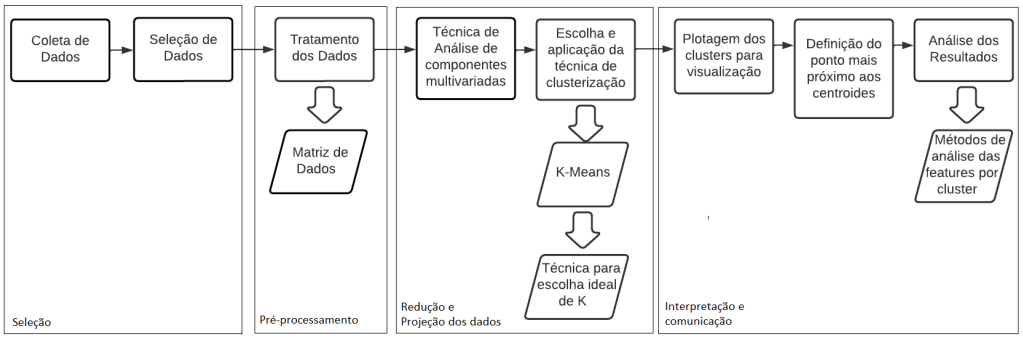

Figura 3. Visão geral da metodologia proposta

A Etapa 1, denominada de Seleção, é onde ocorre a coleta e a seleção dos dados de interesse. Para esse processo, optou-se pelo uso dos dados do Cadastro Ambiental

Rural (CAR)[1] e do Instituto Brasileiro de Geografia e Estatística (IBGE)[2]. 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 realizada 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 ter sua representação alterada, como por exemplo, trocar a vígula por ponto.

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 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 Discussões Preliminares

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

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*

[1]https://www.car.gov.br/

[2]https://www.ibge.gov.br/

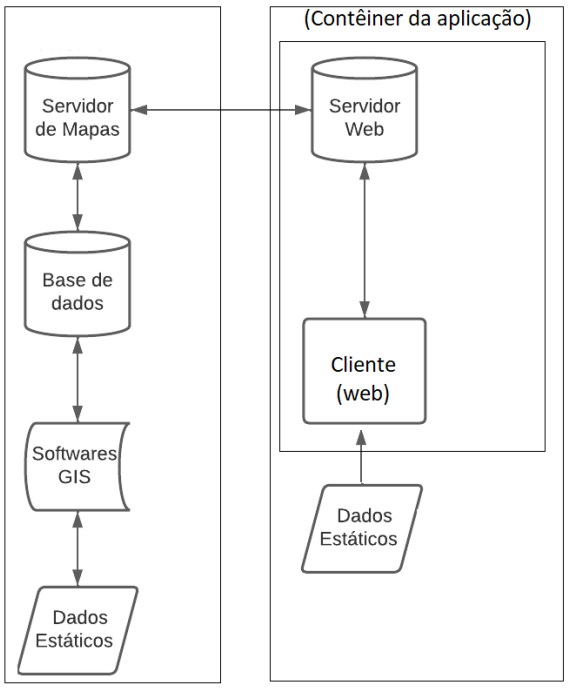

Figura 4. Arquitetura geral proposta.

Service (WMS) e *Web Feature Service* (WFS), conforme especificações da *Open Geospatial Consortium* (OGC)[3]. A Figura 5 ilustra o resultado da aplicação do algoritmo *KMeans* ao conjunto de dados já obtido para a Bacia Hidrográfica do Rio Paraopeba (BHRP). Nela, cada um dos alfinetes representa uma das cidades que compõem essa bacia e cada cor representa um dos agrupamentos distintos obtidos. Para esse exemplo, após a aplicação da metodologia aqui descrita, em especial pelo uso do algoritmo *KMeans*, foram obtidos três agrupamentos, representados pelas cores vermelho, verde e amarelo. A partir disso, o usuário pode identificar de forma visual quais as cidades que são mais semelhantes e, em seguida, buscar extrair informações estatísticas a cerca dos dados relacionados, verificando assim quais as áreas que mais necessitam dos investimentos.

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 disponibilizar a ferramenta para uso, realizando a coleta de *feedbacks* dos usuários e implementando novas funcionalidades.Além disso, planeja-se também testar outros algoritmos de Agrupamento, avaliando os resultados encontrados, para assim verificar qual é a melhor abordagem a ser utilizada.

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

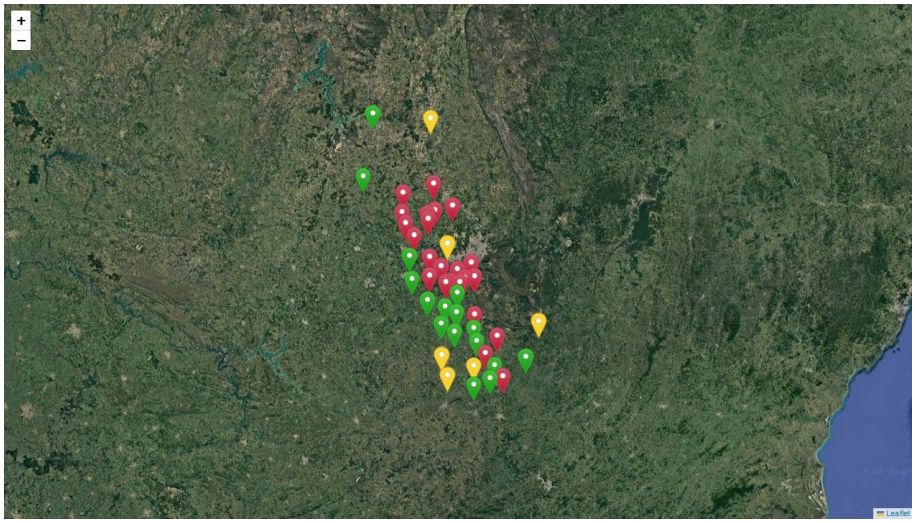

Figura 5. Plotagem dos agrupamentos obtidos como exemplo para a BHRP.

Referências

Barbosa, F. A. R., Maia-Barbosa, P. M., Nascimento, A. M. A., Rietzler, A. C., Franco, M. W., Paes, T. A., Reis, M., Moura, K. A. F., Dias, M. F., de Paula Ávila, M., et al. (2015). O desastre de mariana e suas consequências sociais, econômicas, políticas e ambientais: porque evoluir da abordagem de gestão dos recursos naturais para governança dos recursos naturais? *Arquivos do Museu de História Natural e Jardim Botânico da UFMG*, 24(1-2).

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.

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.