Algoritmos de agrupamento e de detecção de comunidades em redes: comparações e aplicações

PALAVRAS CHAVE. Agrupamento, Detecção de Comunidades, Redes.

Vitor Alves Iglesias, FCA-UNICAMP
Orientadora: Profa. Dra. Priscila Cristina Berbert Rampazzo, FCA-UNICAMP

1. Introdução

Os avanços computacionais e tecnológicos impulsionaram a geração e armazenamento de dados, criando o desafio de como extrair informações relevantes desse grande volume. No Aprendizado de Máquina, técnicas de Agrupamento (*Clustering*) visam dividir dados em subconjuntos (*clusters*), de modo que elementos de um mesmo grupo sejam mais semelhantes entre si do que com os de outros. Essa divisão depende de critérios de similaridade ou dissimilaridade, como coeficiente de correlação ou medidas de distância. Os métodos clássicos de agrupamento de dividem em quatro grupos: agrupamento hierárquico, agrupamento baseado em centróide, agrupamento baseado em grafo e agrupamento baseado em densidade [Xu e Tian, 2015], com desempenho dependente no tipo de base de dados, tipicamente dependente da dimensão do espaço vetorial no qual os pontos estão distribuídos, número de amostras e formato dos grupos.

As técnicas de agrupamento de dados baseadas em rede [da Silva et al., 2019] têm se tornado cada vez mais populares. Redes são estruturas relacionais definidas por entidades que se integram e interagem dinamicamente entre si [Teixeira et al., 2008]. Nas redes, as entidades são representadas por vértices e seus relacionamentos por arestas. Em algoritmos de agrupamento baseados em rede, o primeiro passo é transformar o conjunto de dados em uma rede. Em seguida, um algoritmo de detecção de comunidade é usado para particionar os vértices da rede em comunidades. Finalmente, as comunidades são interpretadas como grupos do conjunto de dados original. Um dos benefícios da utilização de algoritmos de detecção de comunidade é que eles podem examinar a estrutura topológica que surge da transformação dos dados em uma rede, além dos atributos dos objetos no conjunto de dados [Querobim, 2021].

Este trabalho tem o intuito de explorar algoritmos de agrupamento e de detecção de comunidades em redes com enfoque multidisciplinar. A detecção de comunidades é de grande importância em sociologia, biologia e ciência da computação [Rostami et al., 2023], disciplinas onde as relações entre sistemas são frequentemente representadas através de grafos. Este problema é muito difícil e ainda não foi resolvido de forma satisfatória, apesar do enorme esforço de uma grande comunidade interdisciplinar de cientistas que trabalharam nele nos últimos anos [Fortunato, 2010]. O objetivo geral é realizar uma análise e comparações entre algoritmos de detecção de comunidades e aplicá-los em uma situação de interesse para realização de análises e obtenção de informações.

2. Métodos e Metodologia

O estudo de diferentes algoritmos de detecção de comunidades é essencial, pois as redes podem variar em tamanho, densidade e padrões de conexão [Newman e Girvan, 2004]. Métodos distintos apresentam vantagens e limitações específicas, sendo importante compreender suas abordagens.

2.1. Louvain

O algoritmo de Louvain, proposto por Blondel et al. [2008], é heurístico e utiliza uma abordagem gulosa para otimização da modularidade. Cada vértice começa em sua própria comunidade. Iterativamente, os vértices são realocados para maximizar a modularidade. As comunidades resultantes são transformadas em "supervértices", e o processo se repete até estabilizar [Que et al., 2015].

2.2. Girvan-Newman

Este método exato, embora computacionalmente custoso, remove iterativamente as arestas com maior centralidade de intermediação para expor comunidades, baseando-se na ideia de que tais arestas conectam diferentes grupos [Girvan e Newman, 2002].

2.3. Infomap

Rosvall e Bergstrom [2008] propuseram este algoritmo heurístico baseado em Teoria da Informação. Ele usa passeios aleatórios como modelo do fluxo de informação, detectando módulos ao minimizar a descrição desses fluxos.

2.4. Label Propagation

Método heurístico em que cada vértice adota o rótulo mais frequente entre seus vizinhos. O processo se repete até a estabilização dos rótulos, sem exigir o número prévio de comunidades [Raghavan et al., 2007].

2.5. Walktrap

Baseado em passeios aleatórios, este método parte do princípio de que tais caminhadas tendem a permanecer em regiões densas da rede. Mede a similaridade entre vértices e aplica agrupamento hierárquico aglomerativo [Pons e Latapy, 2005].

2.6. Leiden

O algoritmo de Leiden é uma evolução do Louvain, com melhorias em precisão e escalabilidade [Traag et al., 2019]. O algoritmo de Leiden aplica uma técnica de otimização visando maximizar a modularidade da rede e agrupar seus nós. O algoritmo de Leiden tornou-se popular devido à sua eficácia em detectar comunidades em redes grandes e complexas, com milhões de nós e arestas.

2.7. Métrica de Avaliação

A métrica de modularidade proposta por Newman e Girvan [2004] é uma das funções mais utilizadas e conhecidas para quantificar a estrutura de comunidades em um grafo [Que et al., 2015]. A modularidade é uma medida que varia entre -1 e 1. Valores mais altos indicam comunidades mais bem definidas.

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

onde:

- A_{ij} é a matriz de adjacência;
- k_i e k_j são os graus dos nós;
- m é o número total de arestas;
- $\delta(c_i, c_j)$ é 1 se os nós estão na mesma comunidade, 0 caso contrário.

A modularidade é uma métrica que mede a qualidade da partição das comunidades comparando a densidade de arestas dentro das comunidades *versus* entre elas.

2.8. Aplicações

Overton.io[1] é a maior base de dados de políticas públicas e literatura cinzenta do mundo, permitindo analisar como a pesquisa influencia políticas públicas. Utilizamos uma base extraída da plataforma com 4.662 documentos políticos que referenciam pesquisas da Universidade Estadual de Campinas (UNICAMP), no período de 01/01/1983 a 07/07/2022. Desses documentos, foram identificadas 387 organizações distintas como fontes. Cada documento possui três palavras-chave associadas, que representam os principais temas abordados. Para análise, estruturamos um grafo onde os vértices correspondem às palavras-chave, com pesos definidos pelo número de ocorrências. As arestas representam coocorrências de palavras em um mesmo documento, ponderadas pela frequência dessa associação.

Para aplicar os métodos da Seção 2, selecionamos três organizações e seus respectivos conjuntos de documentos:

- 161 documentos do Serviço de Publicações da União Europeia (POEU).
- 486 documentos da Organização Mundial da Saúde (OMS).
- 329 documentos do Instituto de Pesquisa Econômica Aplicada (IPEA).

Consideramos as 30 palavras-chave mais frequentes de cada organização como representativas para a análise.

3. Resultados e Discussões

As estruturas foram implementadas em Python com o pacote igraph, utilizando a função DictList para construir o grafo. A escolha se deu pela flexibilidade e escalabilidade da linguagem. O algoritmo de Leiden apresentou as maiores modularidades em todas as bases (Tabela 1), sugerindo partições mais consistentes.

Algoritmo	**OMS**	**POEU**	**IPEA**
Leiden	0,4336	0,4640	0,1954
Louvain	0,4336	0,4640	0,1954
Girvan-Newman	0,3956	0,4520	0,0500
Infomap	0,4256	0,4640	0,0000
Label Propagation	0,3805	0,4640	0,0000
Walktrap	0,3181	0,4226	0,1907

Tabela 1: Melhores Resultados - Modularidade por Algoritmo e Base de Dados

A Figura 1 apresenta as comunidades detectadas para a base OMS com o algoritmo de Leiden. Neste caso, as comunidades dividem-se em: [Grupo 1] health, medicine, health sciences, clinical medicine, medical specialties, health care, public health, human activities, mental disorder, dengue, fever, diseases and disorders; [Grupo 2] childbirth, birth control, hiv/aids, maternal death, world health organization, family planning, breastfeeding, sexually transmitted infection, abortion, caesarean section; [Grupo 3] coronavirus disease 2019, cervical cancer, human papillomavirus infection, vaccine, vaccination, covid-19 pandemic, pfizer–biontech covid-19 vaccine;

[1] https://www.overton.io/

[Grupo 4] randomized controlled trial, suicide. O Grupo 1 representa pesquisas amplas relacionadas à saúde, medicina clínica, saúde pública, epidemiologia de doenças e transtornos mentais, tendo como base conceitual geral as ciências da saúde e cuidados médicos. O Grupo 2 tem foco em saúde da mulher, gestação, parto, planejamento familiar, infecções sexualmente transmissíveis e mortalidade materna; representa um subcampo especializado em saúde reprodutiva e materno-infantil, articulado com políticas públicas e recomendações de órgãos como a OMS. O Grupo 3 representa pesquisas em doenças infecciosas virais (COVID-19, HPV), vacinas e imunização, com enfoque em saúde pública e prevenção de doenças, com foco específico de epidemiologia e imunologia e influenciando práticas clínicas e políticas de saúde. O Grupo 4 apresenta pesquisas que utilizam ensaios clínicos randomizados para avaliar tratamentos ou intervenções, incluindo estudos sobre comportamentos de risco à saúde (suicídio) através de métodos de pesquisa clínica, fornecendo a base científica para validação de práticas médicas aplicadas nos 3 grupos anteriores. A estrutura conceitual dos grupos pode ser organizada como: (i) nível macro: Grupo 1 (abrange todo o sistema de saúde); (ii) áreas especializadas: Grupo 2 (reprodutiva/materno-infantil) e Grupo 3 (doenças virais e vacinação); (iii) base metodológica: Grupo 4 (ensaios clínicos, pesquisa aplicada).

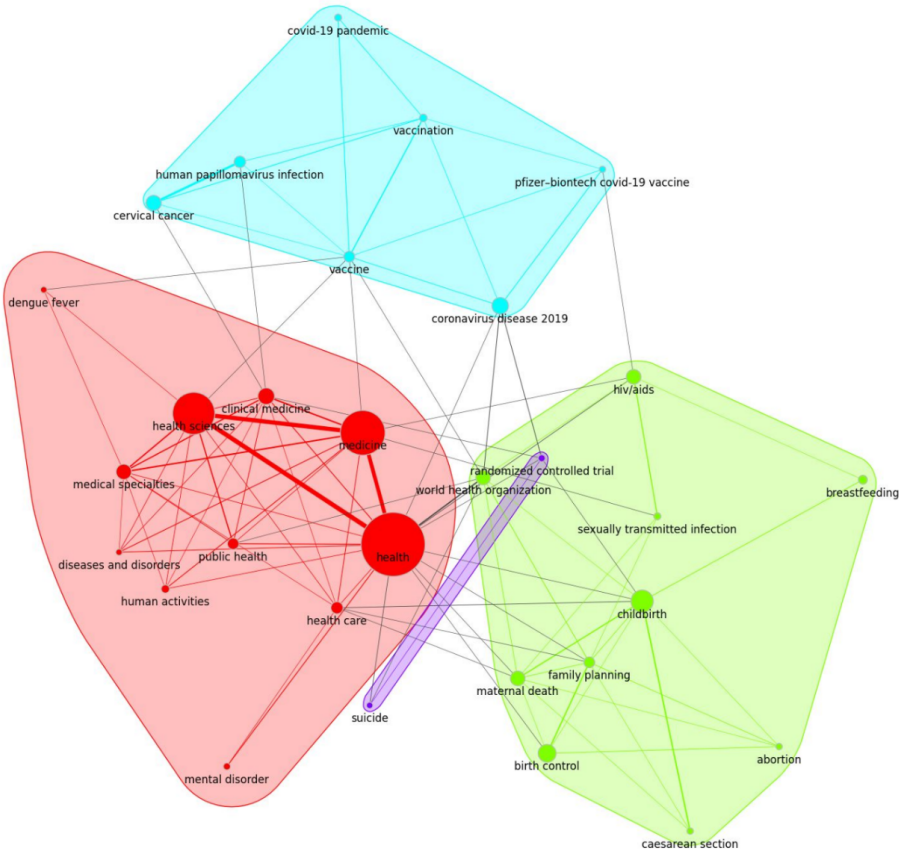

Figura 1: Base OMS, Algoritmo de Leiden.

De forma geral, o algoritmo de Leiden apresenta o melhor desempenho: ou é o melhor, ou empata com outro algoritmo. É importante destacar que esses métodos são heurísticos, ou seja, baseiam-se em estratégias aproximadas para encontrar soluções na tentativa de maximizar a conexão interna dos grupos. Como consequência, execuções distintas podem produzir partições diferentes,

mesmo com os mesmos dados de entrada.

Portanto, a escolha do algoritmo e a interpretação dos resultados devem considerar o contexto do problema, os objetivos da análise e o nível de detalhe desejado. A análise qualitativa, nesse sentido, não apenas complementa a análise quantitativa, aqui apresentada através da métrica de modularidade, como permite identificar padrões temáticos e relações semânticas relevantes que podem ser decisivos para a aplicação prática da detecção de comunidades.

Referências

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., e Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

da Silva, D. M., de Moura Brito, J. A., e Oliveira, C. S. (2019). Um estudo computacional comparativo entre algoritmos de agrupamento e de detecção de comunidades. *Simpósio de Pesquisa Operacional e Logística da Marinha*.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Girvan, M. e Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Newman, M. E. e Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Pons, P. e Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, p. 284–293. Springer.

Que, X., Checconi, F., Petrini, F., e Gunnels, J. A. (2015). Scalable community detection with the louvain algorithm. In *2015 IEEE international parallel and distributed processing symposium*, p. 28–37. IEEE.

Querobim, J. N. C. (2021). *Análise comparativa entre algoritmos de agrupamento e de detecção de comunidades em redes*. PhD thesis, Universidade Federal de São Carlos.

Raghavan, U. N., Albert, R., e Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036106.

Rostami, M., Oussalah, M., Berahmand, K., e Farrahi, V. (2023). Community detection algorithms in healthcare applications: A systematic review. *IEEE Access*, 11:30247–30272.

Rosvall, M. e Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Teixeira, L., Lima, L., e Abreu, N. (2008). Grafos que modelam redes confiáveis. *Mestrado em Engenharia de Produção, COPPE, Universidade do Rio de Janeiro, Rio de Janeiro*.

Traag, V. A., Waltman, L., e Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Xu, D. e Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.