Olavo Morais Borges Pereira - 11297792

Projeto 29 - Como usar redes complexas em aprendizado de máquina?

Abordagem Utilizada

A abordagem adotada consistiu em transformar o conjunto de dados em uma rede complexa e aplicar algoritmos de detecção de comunidades nessa rede. As comunidades identificadas foram utilizadas como clusters. Para realizar a transformação do conjunto de dados em uma rede complexa, foi empregada a estratégia de atribuir a cada entrada do conjunto de dados um nó, conectando todos os nós entre si. O peso das arestas foi determinado com base em uma métrica que mensura a proximidade entre os nós.

Métricas de Similaridade Utilizadas

As seguintes métricas foram usadas para medir distâncias entre dois nós:

- Inversa da Distância de Minkowski
  - $M_p^{-1}(x,y) = \frac{1}{M_p(x,y)}$, onde $M_p(x,y) = (\sum_{i=1}^{n}(|x_i - y_i|^p)^{\frac{1}{p}}$
  - Foi usado no código $p = 0..5$, $p = 1$(Distância de Manhattan), $p = 2$ (Distância Euclidiana) e $p = \infty$ (Distância Chebyshev)
- Exponencial da Distância de Minkowski
  - $\alpha \cdot e^{-\alpha M_p(x,y)}$
- Métrica de Fu
  - $1 - \frac{M_2(x,y)}{||x||+||y||}$
- Exponencial da Métrica de Fu
  - $\alpha \cdot e^{-\alpha \cdot \frac{1-F}{2}}$
- Exponencial de Tanimoto
  - $\alpha \cdot e^{-\alpha \cdot \frac{1-T}{2}}$ com $T = \frac{x \cdot y}{||x||^2+||y||^2-x \cdot y}$

Em todas as exponenciais, o valor de α ficou fixado em 1.

Algoritmos utilizados

A abordagem adotada consistiu em utilizar quatro algoritmos para a detecção de comunidades na rede complexa gerada a partir do conjunto de dados. Os algoritmos escolhidos foram Greedy Optimization, Label Propagation, Walktrap e Spin Glass. Esses algoritmos têm como objetivo identificar grupos de nós que possuam interconexões mais densas entre si do que com o restante da rede.

Após a detecção das comunidades, foram utilizados três algoritmos tradicionais de clusters para a comparação dos resultados obtidos. Os algoritmos selecionados foram KMeans, Mean Shift, DBSCAN, Expectation Maximization e Agglomerative Clustering. Esses algoritmos são amplamente utilizados na área de aprendizado de máquina e clustering, e

visam agrupar os dados em conjuntos semelhantes com base em diferentes critérios, como distância ou maximização da verossimilhança.

A análise comparativa dos resultados obtidos pelos algoritmos de detecção de comunidades e pelos algoritmos tradicionais de clusters permitiu avaliar a eficácia de cada método na organização e identificação de grupos dentro do conjunto de dados. Essa abordagem contribuiu para uma compreensão mais aprofundada das estruturas e relações presentes nos dados, auxiliando na obtenção de informações relevantes e na tomada de decisões embasadas.

Datasets Utilizados

A fim de validar os algoritmos empregados, foram escolhidos dois conjuntos de dados de vida real: o conjunto de dados Iris e o conjunto de dados Wine.

O conjunto de dados Iris é amplamente utilizado na área de aprendizado de máquina e consiste em medidas de quatro características diferentes de três espécies de flores Iris. Cada entrada no conjunto de dados corresponde a uma flor e é caracterizada por comprimento e largura de sépalas e pétalas. O objetivo é agrupar as flores em diferentes espécies com base em suas características.

O conjunto de dados Wine é outra referência comum em aprendizado de máquina. Ele contém informações sobre diferentes amostras de vinhos, incluindo atributos físico-químicos, como teor alcoólico, acidez e concentração de compostos. O objetivo é agrupar as amostras em diferentes categorias de vinho com base em suas características.

Para garantir uma análise adequada dos dados, foi necessário normalizá-los, levando em consideração possíveis diferenças nas escalas das características. A normalização é um processo comumente utilizado para ajustar os valores das características dentro de uma faixa específica, a fim de evitar viés e distorções nos resultados dos algoritmos de detecção de comunidades e clusters.

Ao utilizar esses dois conjuntos de dados de vida real e aplicar a normalização adequada, foi possível avaliar a eficácia dos algoritmos de detecção de comunidades e clusters na organização e agrupamento dos dados, contribuindo para a compreensão das estruturas e padrões subjacentes aos dados estudados. Foram utilizados dois datasets artificiais gerados pelas funções make_moons e make_blobs da biblioteca sklearn. A inclusão de datasets artificiais é fundamental para avaliar o desempenho dos algoritmos em cenários controlados e compreender melhor suas capacidades e limitações.

Avaliação dos Resultados

A fim de avaliar as comunidades obtidas por meio dos algoritmos de detecção de comunidades, utilizou-se a métrica de Jaccard. Essa métrica é calculada por meio da fórmula $J = \frac{a}{a+b+c}$, em que a representa o número de pares de nós que pertencem ao mesmo cluster e possuem o mesmo rótulo (label), b representa o número de pares de nós

que estão em clusters diferentes, mas possuem o mesmo rótulo, e c representa o número de pares de nós que pertencem ao mesmo cluster, mas possuem rótulos diferentes.

Essa métrica é comumente utilizada para medir a sobreposição ou similaridade entre os clusters obtidos em relação aos rótulos atribuídos aos nós. Quanto maior o valor de Jaccard, maior a similaridade entre os clusters e os rótulos. Um valor de Jaccard igual a 1 indica que todos os pares de nós pertencem ao mesmo cluster e têm o mesmo rótulo, enquanto um valor próximo de 0 indica que não há sobreposição entre os clusters e os rótulos.

Ao utilizar a métrica de Jaccard para avaliar as comunidades obtidas pelos algoritmos de detecção de comunidades, foi possível verificar a qualidade e a consistência dos agrupamentos em relação aos rótulos atribuídos aos nós. Essa avaliação contribui para a compreensão da precisão e eficácia dos algoritmos utilizados na tarefa de detecção de comunidades, auxiliando na interpretação e na validação dos resultados obtidos.

Resultados

Exibindo apenas os resultados dos algoritmos de clustering tradicionais e os 5 melhores resultados para os algoritmos de detecção de comunidade. Todos os resultados podem ser acessados no final do Notebook.

Iris Dataset

Algoritmo	Métrica	Índice de Jaccard	Número de Classes
Kmeans	-	0.695	3
Mean Shift	-	0.587	2
DBSCAN	-	0.537	3
EM (k=?)	-	0.329	1
EM (k=3)	-	0.879	3
Agglomerative Clustering	-	0.698	3
FastGreedy (k=?)	Exp Mink p=0.5	0.837	4
FastGreedy (k=3)	Exp Mink p=0.5	0.879	3
Spin Glass	Exp Mink p=1	0.702	4
Spin Glass	Exp Mink p=2	0.637	3
Walktrap	Exp Mink p=1	0.698	3

Wine Dataset

Algoritmo	Métrica	Índice de Jaccard	
KMeans	-	0.872	3
Mean Shift	-	0.330	2
DBSCAN	-	0.338	1
EM (k=?)	-	0.338	1
EM (k=3)	-	0.814	3
Agglomerative Clustering	-	0.755	3
FastGreedy (k=?)	Invers Mink p=1	0.955	3
FastGreedy (k=?)	Invers Mink p=0.5	0.932	3
FastGreedy (k=?)	Exp Mink p=2	0.809	3
LabelPropagation	Exp M p=2	0.832	3
SpinGlass	Invers Mink p=1	0.828	4

Os resultados completos podem ser conferidos no Notebook.

Comentários Finais

Ficou claro que os algoritmos de detecção de comunidades não são perfeitos e apresentam erros, mas possuem algumas vantagens. Apresentam um bom desempenho mesmo quando o número de classes é desconhecido. É visível que não existe uma métrica de distância que sempre é melhor do que outra. É necessário testar diferentes algoritmos e diferentes métricas para cada algoritmo. Outra abordagem viável para construir uma rede complexa a partir de um conjunto de dados é calcular a similaridade entre dois nós e criar uma aresta apenas quando essa similaridade excede um determinado limiar, que pode ser estabelecido através do treinamento do modelo com os conjuntos de dados disponíveis.