Relatório parcial de atividades do bolsista e do aluno voluntário da FAPESP

1 Identificação

Nome do Orientador:	Alexandre Luís Magalhães Levada
Nome do Aluno:	Antônio Cícero Amorim de Azevedo
Área do Projeto:	Ciência da Computação
Título do Projeto do Aluno:	O Algoritmo K-médias Geodésico para Agrupamento de Dados Baseado em Grafos

2 Objetivos e metas

Esse projeto de pesquisa tem o objetivo de propor e implementar um novo algoritmo para agrupamento de dados baseado em grafos, que substitui a distância euclidiana no k-médias pelo comprimento de caminhos mínimos em grafos. Com o intuito de nortear o projeto, o cronograma a seguir apresenta a divisão da pesquisa em 8 etapas a serem desenvolvidas durante sua duração:

- **1ª Etapa:** Realizar estudos sobre a fundamentação teórica do algoritmo k-médias.
- **2ª Etapa:** Realizar estudos sobre a fundamentação teórica do algoritmo de Dijkstra.
- **3ª Etapa:** Realizar estudos sobre medidas de qualidade de agrupamento.
- **4ª Etapa:** Desenvolver e implementar o algoritmo K-médias geodésico proposto em linguagem Python
- **5ª Etapa:** Realizar teste, validação e comparação do método proposto com o K-médias padrão e eventualmente outros métodos similares.
- **6ª Etapa:** Aquisição e busca de conjuntos de dados provenientes de repositórios online.
- **7ª Etapa:** Realizar testes com o método implementado em conjuntos de dados reais.
- **8ª Etapa:** Escrita do relatório técnico final do projeto de pesquisa.

3 Principais etapas executadas

Ao longo dos primeiros seis meses de projeto, todas as etapas previstas para o primeiro semestre foram executadas: 1ª, 2ª, 3ª e 4ª etapas foram incluídas; 5ª, 6ª e 7ª etapas foram desenvolvidas e estão em andamento.

Por meio da leitura da bibliografia indicada para o projeto, foi possível estabelecer uma fundamentação teórica sobre o algoritmo k-médias clássico e o algoritmo de Dijkstra. Além disso, disciplinas da grade curricular do bacharelado em Ciência da Computação serviram de complemento para a compreensão do k-médias por meio do estudo de diversos métodos de agrupamento hierárquicos e não hierárquicos, além do aprendizado em diversas estruturas de dados que compõem os algoritmos de aprendizado de máquina.

O algoritmo k-médias topológico proposto faz uso de um grafo dos k vizinhos mais próximos, em que dois vértices p e q estão conectados se a distância euclidiana entre p e q estiver entre as k-ésimas menores de p para outros vértices. A ideia é, com base nessa representação em grafo do conjunto de dados, aplicar o algoritmo de Dijkstra e utilizar o comprimento dos caminhos mais curtos como uma aproximação das distˆancias geodésicas para realização dos agrupamentos.

Entre as medidas de qualidade de agrupamento estudadas, optou-se por dez que consideram como conhecida a informação sobre os verdadeiros agrupamentos: *completeness score, fowlkes mallows score, homogeneity score, v measure score, rand score, adjusted rand score, mutual info score, adjusted mutual info score, normalized mutual info score, silhouette score*. A linguagem de programação Python, com auxílio de bibliotecas como o scikit-learn, o numpy e o scipy para o auxílio dos cálculos e da obtenção das métricas, o matplotlib para a criação de gráficos para uma melhor visualização dos dados obtidos e para o versionamento e armazenamento do projeto foram utilizados o git e o github, para mais informações sobre o código fonte acesse a partir do endereço https://github.com/azmovi/geo-k-means.

4 Resultados

Com o intuito de realizar os testes comparativos entre o k-médias topológico e sua versão clássica, alguns conjuntos de dados com alta dimensionalidade foram selecionados do repositório OpenML, acessível a partir do endereço https://www.openml.org/. A Tabela 1 descreve cada um dos conjuntos utilizados com seus nomes e números de observações, variáveis e classes. Percebe-se que os conjuntos de dados escolhidos apresentam a característica em comum de que o número de parâmetros é muito superior ao número de observações. A maldição da dimensionalidade, fenômeno que ocorre em espaços de alta dimensão, apresenta desafios para algoritmos que dependem de distâncias sensíveis a ruído, como a Euclidiana. Dessa forma, a ideia da escolha desses conjuntos de dados é verificar se o algoritmo k-médias topológico proposto consegue gerar agrupamentos melhores e mais significativos nesse cenário específico em comparação ao k-médias clássico.

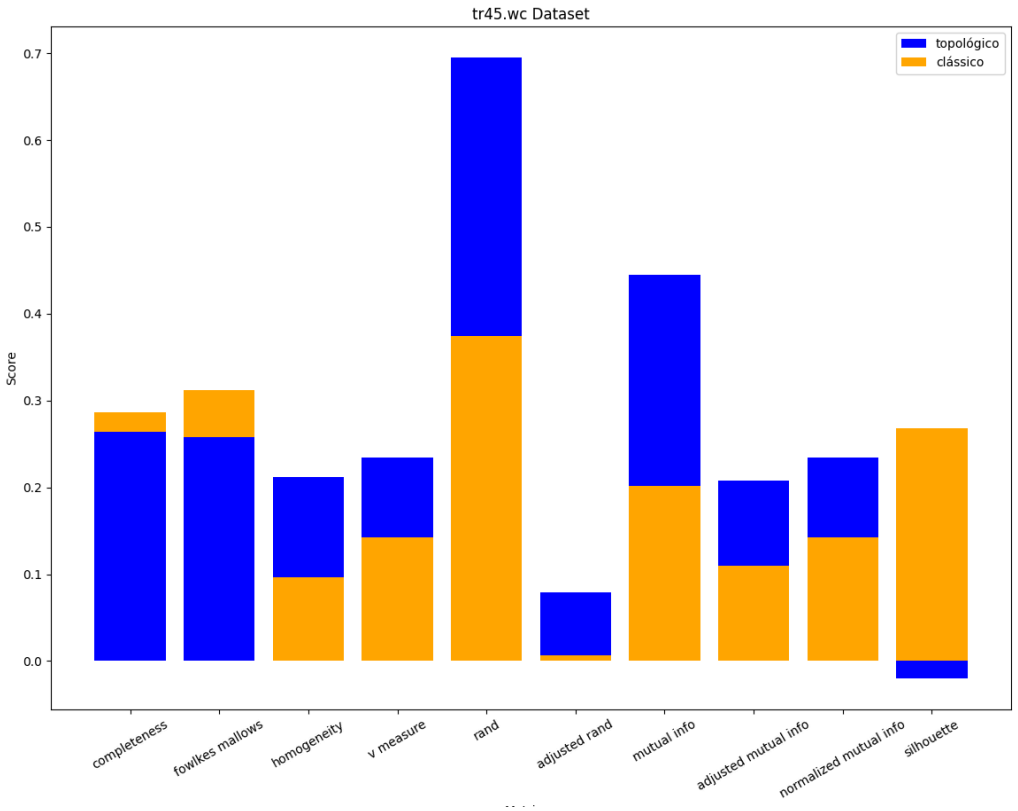
As medidas médias de qualidade de agrupamento calculadas após 30 execuções

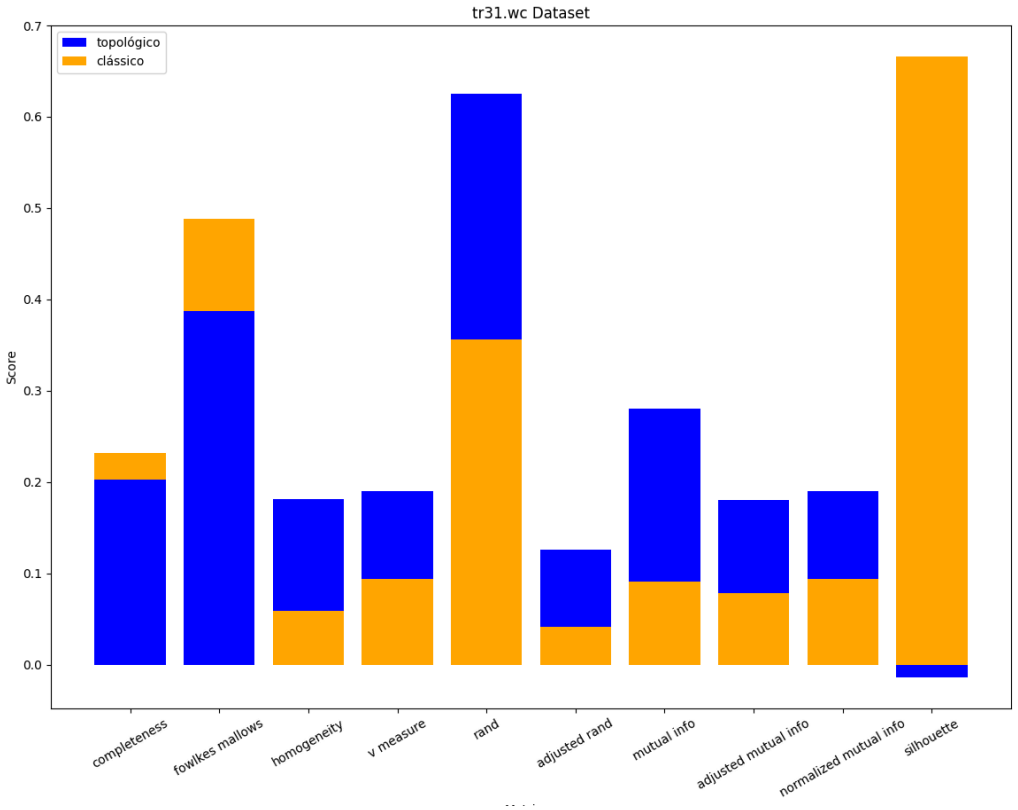
para ambos algoritmos em cada um desses conjuntos de dados estão dispostas nos gráficos a seguir que serão apresentados. Os indícios sugerem que o algoritmo k-médias topológico demonstra um desempenho superior, conforme evidenciado pelas métricas propostas, especialmente em conjuntos de dados onde o número de variáveis excede o número de observações.

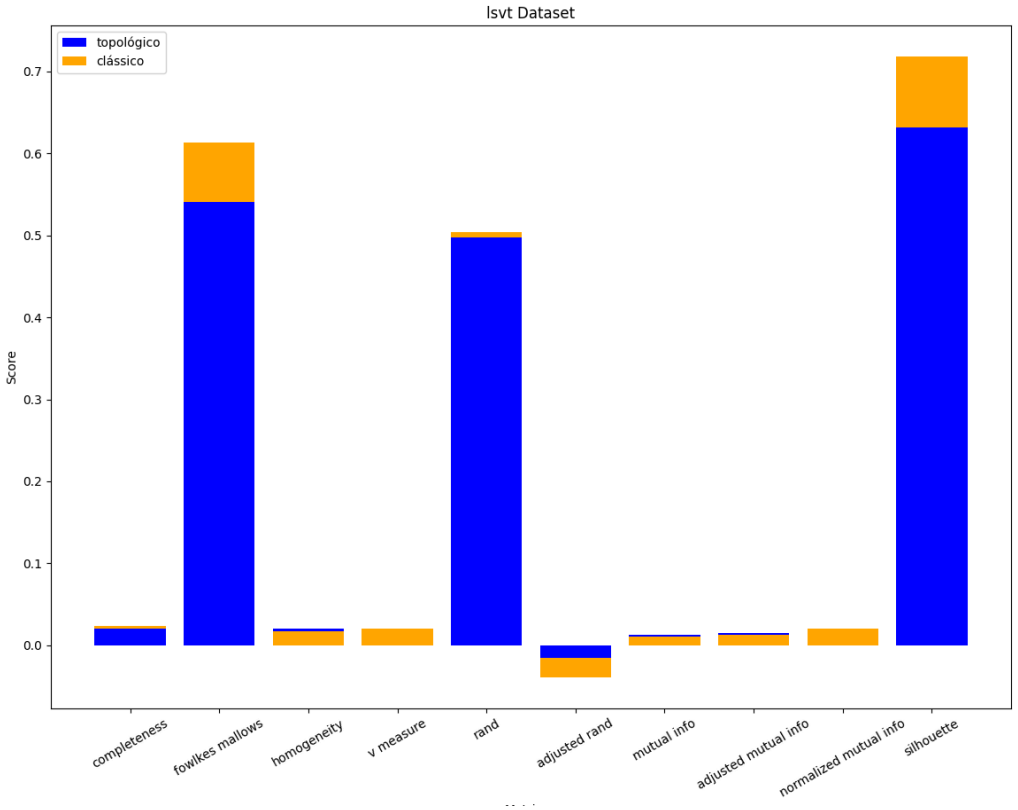
Tabela 1: Número de observações, parâmetros e grupos de cada conjunto de dados utilizado na comparação dos algoritmos.

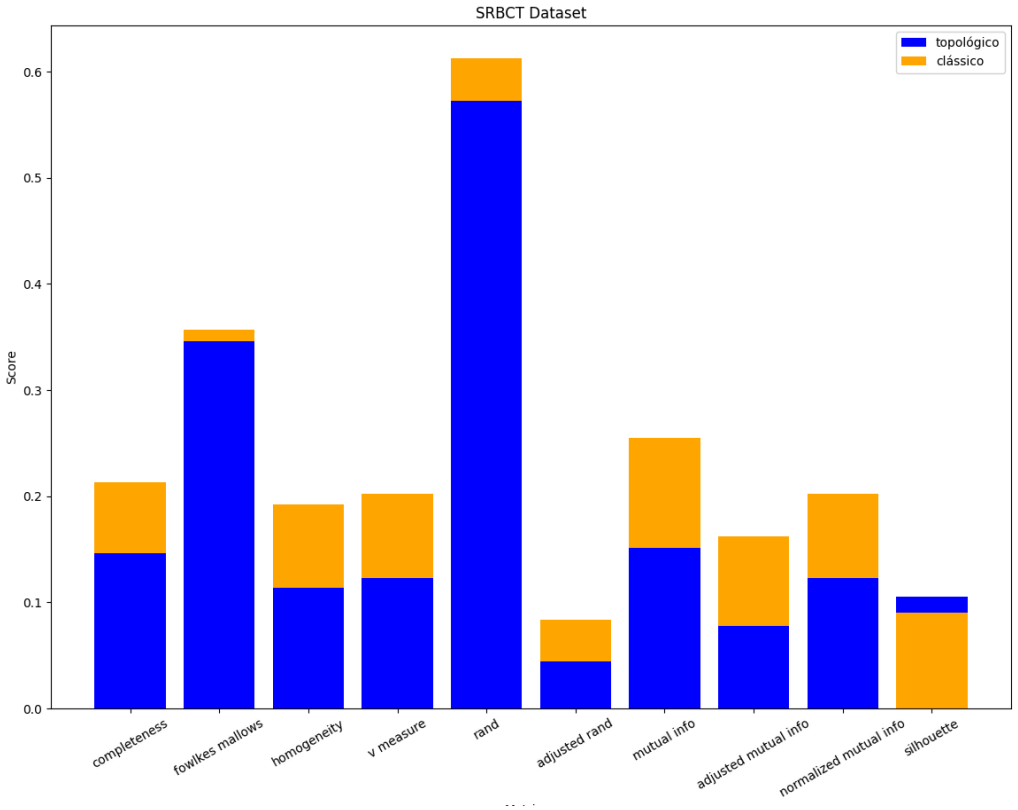
Conjunto de dados	**# observações**	**# parâmetros**	**# grupos**
AP Breast Colon	630	10935	2
AP Colon Prostate	355	10935	2
AP Lung Kidney	386	10935	2
tr31.wc	927	10128	7
tr45.wc	690	8261	10
leukemia	72	7129	2
SRBCT	83	2308	4
GCM	190	16063	14
lsvt	126	310	2

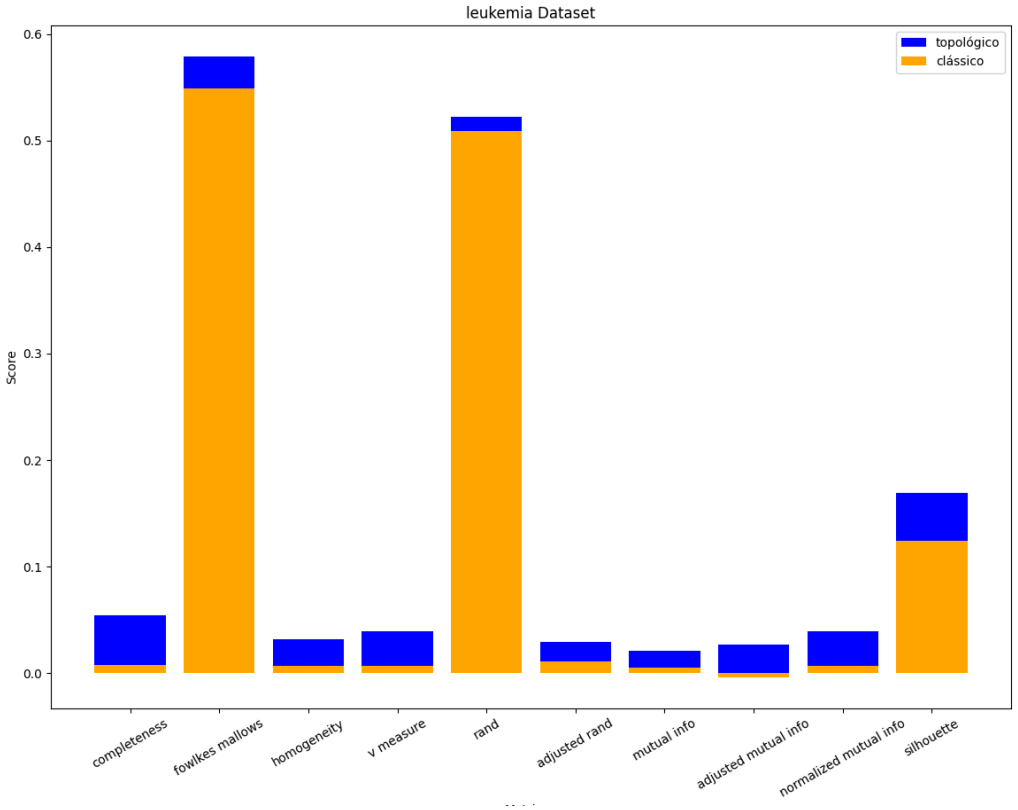
Para checar a significância das diferenças entre as medidas médias para os algoritmos, o teste não paramétrico de Wilcoxon pode ser utilizado considerando amostras pareadas. Podemos observar graficamente os valores dos datasets selecionados e observar quão das duas abordagens teve o melhor desempenho em relação às métricas utilizadas

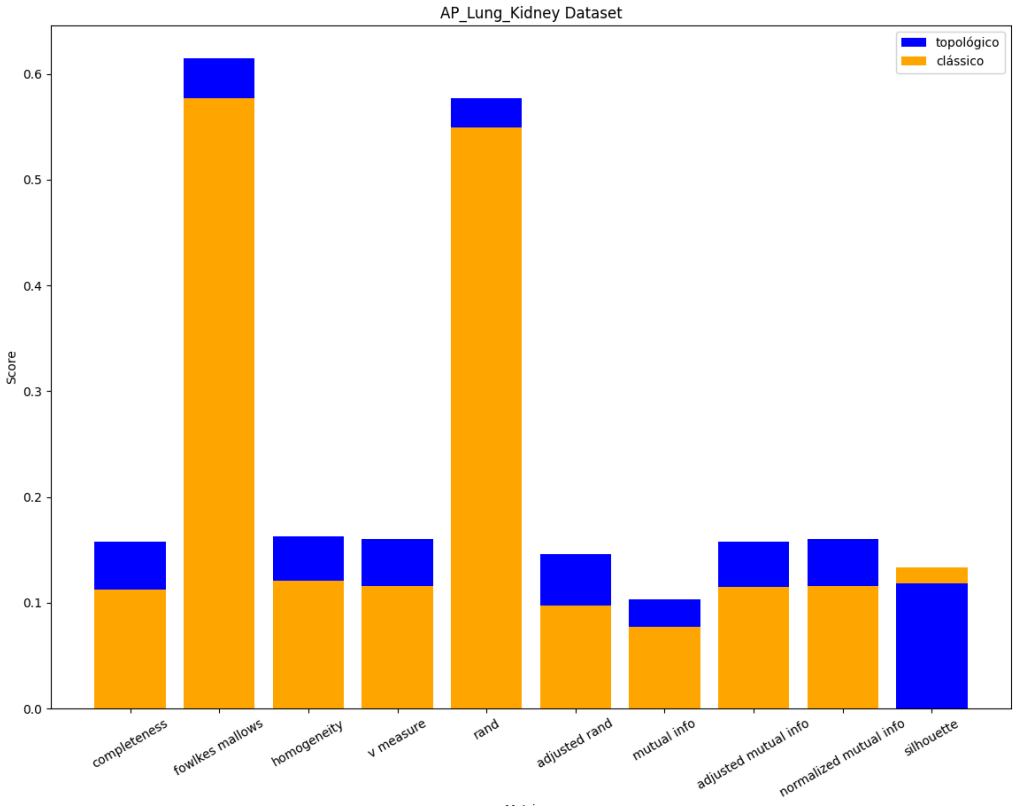


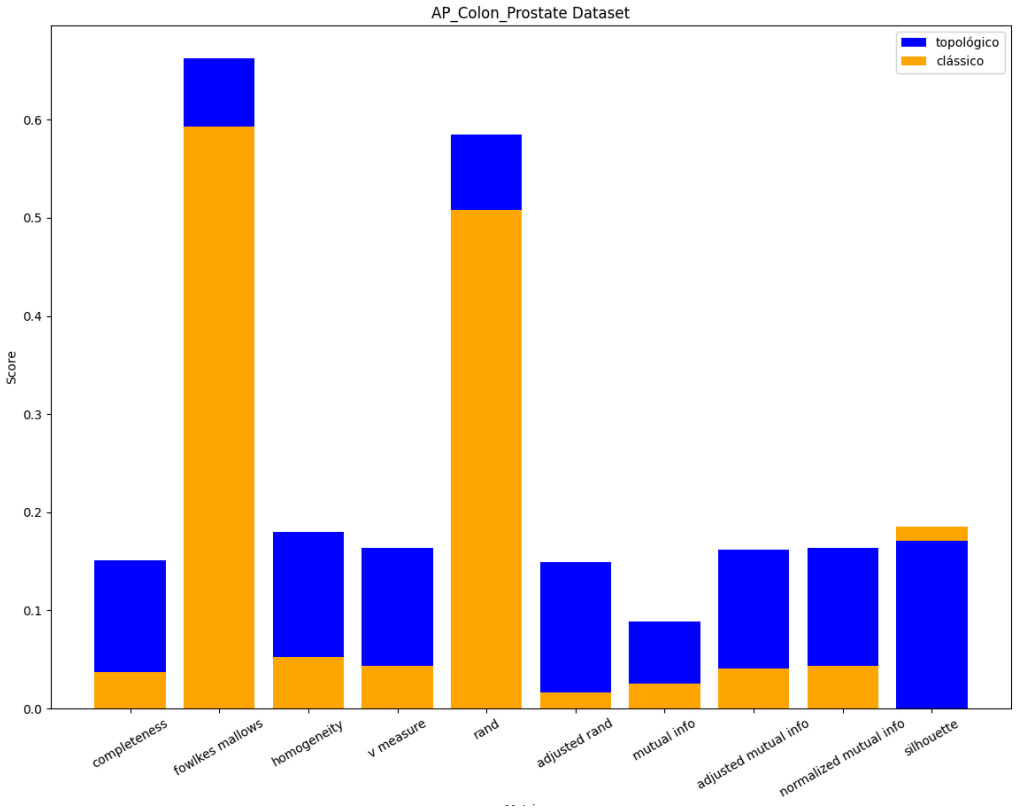

tr31.wc Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette

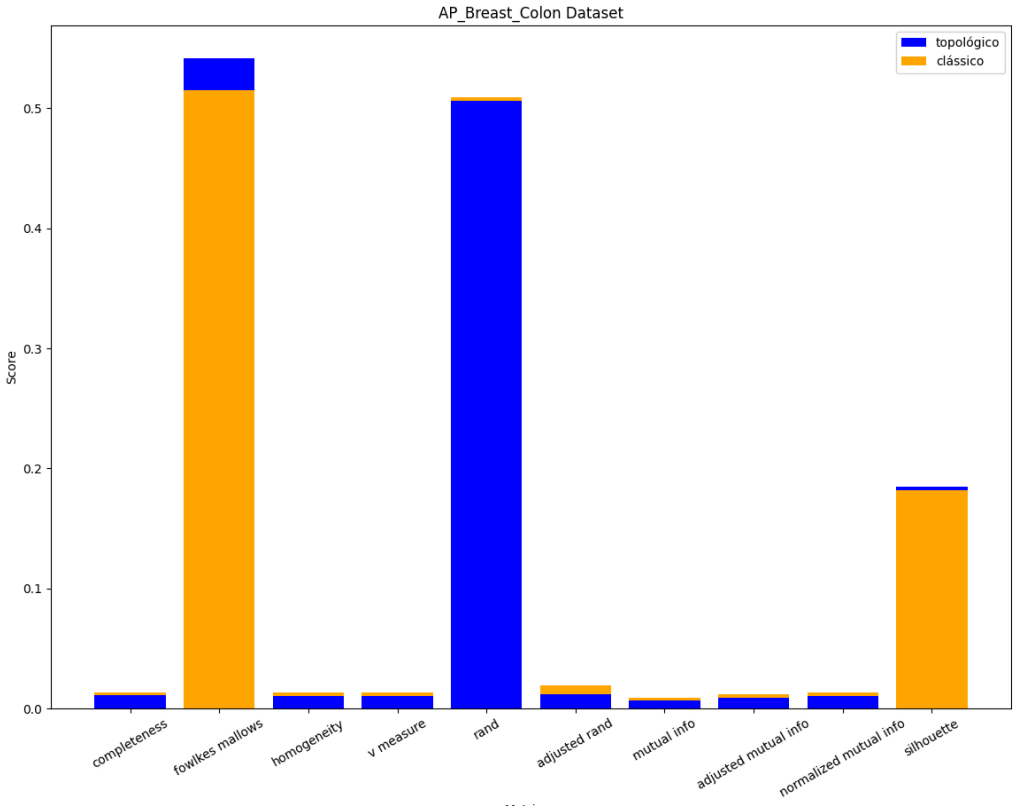

lsvt Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette

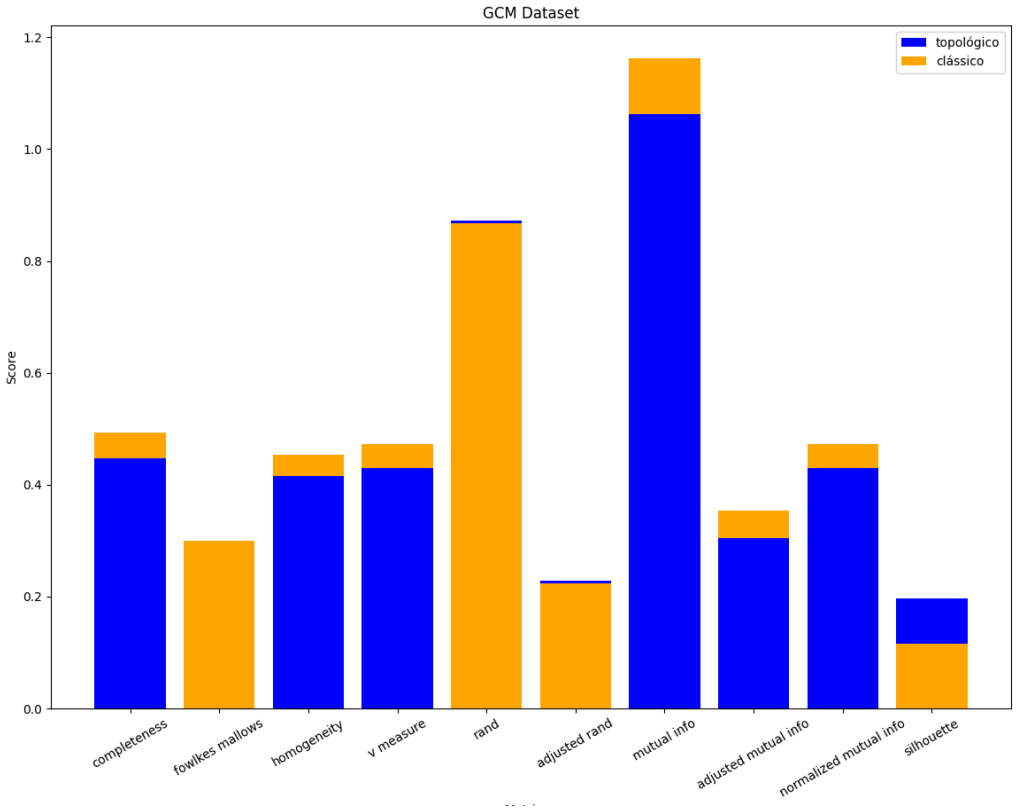

SRBCT Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette


leukemia Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette


AP_Lung_Kidney Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette


AP_Colon_Prostate Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette


AP_Breast_Colon Dataset
topológico
clássico
Score
0.0
0.1
0.2
0.3
0.4
0.5
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette


GCM Dataset
topológico
clássico
Score
0.0
0.2
0.4
0.6
0.8
1.0
1.2
completeness
fowlkes mallows
homogeneity
v measure
rand
adjusted rand
mutual info
adjusted mutual info
normalized mutual info
silhouette

Apesar dos resultados positivos do k médias topológico para essa amostra de conjuntos de dados nessas medidas de qualidade, faz-se necessário a coleta de uma maior quantidade de conjuntos com a característica desejada para que o teste não paramétrico de Wilcoxon tenha um poder estatístico maior. Além disso, outros cenários com relação aos dados podem ser testados, como dados com alta densidade ou baixa dimensionalidade, para comparação da performance de ambos algoritmos e outras medidas de qualidade de agrupamento podem ser utilizadas para complementar a inferência pelos testes estatísticos.

Em geral, há indícios de que o algoritmo proposto é capaz de produzir agrupamentos superiores ao k-médias clássico para dados de alta dimensão. Entretanto, o k-médias topológico sofre de diversas limitações presentes na sua versão clássica, como a escolha dos centroides e o custo computacional, sendo um tópico relevante na busca de possíveis melhorias

5 Cronograma das atividades

Tabela 3: Cronograma das etapas definidas para o projeto.

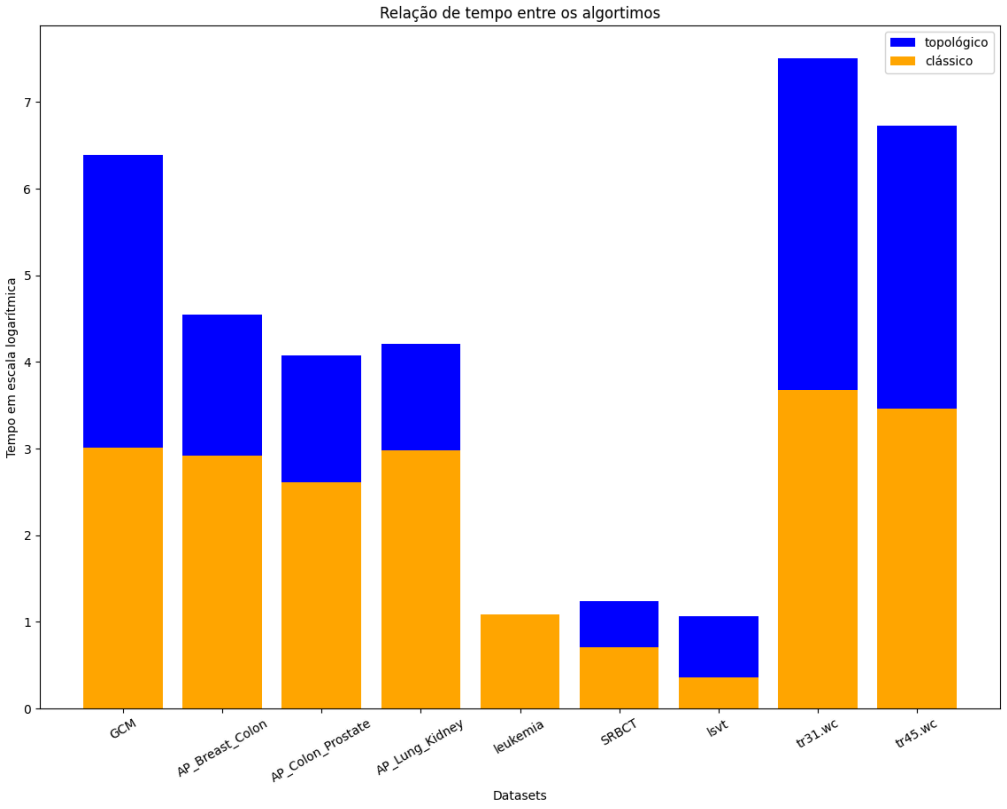
Duração do projeto												
Atividades	**Meses**											
	1	2	3	4	5	6	7	8	9	10	11	12
1ª Etapa (Prevista)	X	X	X									
1ª Etapa (Realizada)	OK	OK	OK									
2ª Etapa (Prevista)		X	X	X								
2ª Etapa (Realizada)		OK	OK	OK								
3ª Etapa (Prevista)		X	X	X	X							
3ª Etapa (Realizada)		OK	OK	OK	OK							
4ª Etapa (Prevista)			X	X	X	X	X					
4ª Etapa (Realizada)			OK	OK	OK	OK						
5ª Etapa (Prevista)				X	X	X	X	X				
5ª Etapa (Realizada)				OK	OK	OK						
6ª Etapa (Prevista)				X	X	X	X	X				
6ª Etapa (Realizada)				OK	OK	OK						
7ª Etapa (Prevista)					X	X	X	X	X	X		
7ª Etapa (Realizada)					OK	OK						
8ª Etapa (Prevista)							X	X	X	X	X	X
8ª Etapa (Realizada)												

6 Dificuldades encontradas

Apesar dos resultados positivos do k-médias topológico em dados de alta dimensão, a performance do algoritmo ainda depende fortemente da escolha dos centroides como o método clássico e, como essa etapa envolve uma aleatoriedade, não é controlável na

maioria dos casos. Dessa forma, melhores opções de inicialização dos centroides podem ser estudadas e testadas para verificar seus efeitos no algoritmo proposto.

Além disso, o tempo de execução do algoritmo k-médias topológico é levemente prejudicado pela complexidade computacional quando se está trabalhando com dados de alta dimensão. Como podemos observar no gráfico a seguir a diferença do tempo de execução de cada algoritmo em um determinado conjunto de dados:



Em: 27/04/2024

Documento assinado digitalmente
gov.br
ANTONIO CICERO AMORIM DE AZEVEDO
Data: 28/04/2024 13:55:39-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) Aluno(a)

Documento assinado digitalmente
gov.br
ALEXANDRE LUIS MAGALHAES LEVADA
Data: 28/04/2024 13:31:22-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) Orientador(a)