

Média geométrica com fator de regularização para estimação de pesos de relevância em agrupamento de dados relacionais com múltiplas visões.

Vitória Teles da Silva

Universidade Federal de Sergipe

Avenida Marcelo Deda Chagas, Rosa Elze, São Cristóvão - SE 49107-230

vitoriateles@academico.ufs.br

Renê Pereira de Gusmão

Universidade Federal de Sergipe

Avenida Marcelo Deda Chagas, Rosa Elze, São Cristóvão - SE 49107-230

renepgusmao@gmail.com

André Britto de Carvalho

Universidade Federal de Sergipe

Avenida Marcelo Deda Chagas, Rosa Elze, São Cristóvão - SE 49107-230

andre@dcomp.ufs.br

RESUMO

O agrupamento de dados com múltiplas visões é caracterizado como o desafio de agrupar dados similares e separar os dados distintos, utilizando múltiplas representações com várias fontes. A abordagem utiliza as diferentes visões que possuem dados complementares com o intuito de obter resultados mais robustos. O agrupamento de dados é uma das ferramentas de maior relevância para análise de dados e sua importância está relacionada com a fiel representação de dados reais, sendo o uso combinado de várias visões apresentando maior precisão se comparado com a visão única. Nesse contexto, a utilização de diferentes métricas para medir as similaridades entre os dados e o uso de pesos de relevância constituem uma abordagem eficiente. Este trabalho tem por objetivo investigar o uso de uma média geométrica com fator de regularização para a estimação dos pesos de relevância em agrupamento de dados relacionais com múltiplas visões.

PALAVRAS CHAVE. Agrupamento. Múltiplas visões. Pesos de relevância.

Tópicos: Análise e Ciência de Dados. Otimização discreta.

ABSTRACT

Multi-view data clustering is characterized as the challenge of grouping similar data and separating distinct data using multiple representations from multiple sources. The approach uses different views that have complementary data in order to obtain more robust results. Data clustering is one of the most relevant tools for data analysis and its importance is related to the faithful representation of real data, with the combined use of multiple views presenting greater accuracy compared to a single view. In this context, the use of different metrics to measure the similarities between data and the use of relevance weights constitute an efficient approach. This work aims to investigate the use of a geometric mean with a regularization factor for estimating relevance weights in relational data clustering with multiple views.

KEYWORDS. Clustering. Multiple views. Relevance weights.

Paper topics: Analysis and Data Science. Discrete Optimization.

1. Introdução

Com o atual crescimento acelerado de informações, os dados atingiram uma grande escala [Jiang et al., 2024]. A obtenção de informações relevantes tornou-se o principal desafio, sendo o agrupamento uma das ferramentas básicas de maior relevância na análise de dados, com aplicações em processamento de imagens, bioinformática, sistemas de recomendação, entre outras áreas de pesquisa [Zhang e Gionis, 2023].

Segundo [Gong et al., 2019], agrupamento é a técnica de agrupar objetos semelhantes, sendo seu objetivo reunir dados semelhantes no mesmo grupo com o intuito de reduzir a complexidade da análise. Porém, [Pfeifer et al., 2023] afirmam que a simples junção dos resultados de agrupamento de visualização única não é suficiente para englobar a complexidade e diversidade dos dados.

As fontes e recursos dos dados são variados, existindo a possibilidade de receber características de diferentes domínios, sendo denominados dados de múltiplas visões [Jiang et al., 2024]. Dessa forma, dados com múltiplas visões são a representação de um mesmo objeto derivado de múltiplas perspectivas. O agrupamento de dados com múltiplas visões consiste em particionar pontos de dados em conjuntos diferentes com o intuito de formar um grupo com características mais semelhantes do que os dados pertencentes aos outros grupos [Shi et al., 2021].

Porém, segundo [Liu et al., 2024], os métodos de agrupamento atuais atuam com partições rígidas ou *fuzzy* que não representam a incerteza e imprecisão em sua totalidade, intensificando o risco de erro. [Li e Zhang, 2019] asseguram que o método de agrupamento multi-visões baseado em aprendizado de peso é adequado para superar a limitação presente nos métodos convencionais. Sendo os pesos e sua relação com as diferentes visões internalizados de forma adaptativa. Uma forma explorada para integrar múltiplas visões de dados e a ponderação é o uso de matrizes de similaridade, método que relaciona os dados com pesos e distâncias correspondentes.

A escolha da métrica de distância é um ponto fundamental no agrupamento, pois é a responsável por definir a similaridade entre os pontos, influenciando diretamente na formação dos grupos [Hua e Wu, 2021]. Distâncias como euclidiana, similaridade de cosseno, baseada em grafos, Mahalanobis e Manhattan são alguns dos exemplos de abordagens mais recorrentes. Entretanto, a escolha da métrica mais adequada para cada tipo de conjunto de dados ainda é um desafio, como apontado por [Hussain et al., 2022].

Em um conjunto de dados, a atribuição de pesos para as diferentes visões é um método utilizado para otimizar o desempenho no agrupamento [Liu et al., 2024]. Já que, segundo [Zhu e Wang, 2019], as visões possuem relevâncias diferentes e não é razoável considerar todas com o mesmo grau de importância. Para estimar o impacto de cada visão na análise de dados, é necessário explorar diferentes métodos de estimação de peso. Como exposto por [Liu et al., 2024], diferentes abordagens possuem resultados distintos, consequência das características dos conjuntos de dados analisados em cada experimento.

Com a ampla variedade de métricas existentes para avaliação de desempenho, há diferentes conclusões sobre a qualidade dos resultados. Dessa forma, vários trabalhos relatam a dificuldade de identificar quais métricas são mais apropriadas para determinação de pesos de relevância. [Qian et al., 2018] refletem sobre a vulnerabilidade em relação a confiabilidade da decisão ocasionada pela distorção de dados presentes em algumas visualizações. Enquanto que [Liu et al., 2024] afirmam que as diferentes abordagens utilizadas possuem uma capacidade limitada de capturar efetivamente as correlações entre os dados de várias fontes.

Pesquisas relacionadas estudam casos específicos e exploram abordagens comparativas em algoritmos de agrupamento multi-visão, como [Lian et al., 2021], na qual a comparação é essencial para verificar a eficácia do algoritmo proposto. [Zhu e Wang, 2019] demonstram a comparação

como método eficaz para escolha do algoritmo mais adequado, enquanto que [Sharma et al., 2024] refletem sobre o uso de pesos impactando na complexidade computacional dos algoritmos e a respectiva relação com os métodos utilizados.

Nesse artigo, o objetivo é realizar experimentos para a investigação do desempenho do algoritmo utilizando a média geométrica com fator de regularização para estimação dos pesos de relevância. Para tal, serão executados testes comparativos e análise de resultados com o intuito de identificar a relação de desempenho e contribuições significativas para a eficácia geral do algoritmo.

O trabalho está estruturado em tópicos, iniciando com a apresentação do conceito e do agrupamento de dados com múltiplas visões, abordando suas principais características. Em seguida, detalha-se o algoritmo utilizado, seguido de sua versão ponderada. A seção subsequente descreve o método proposto para modificação e estimação de pesos, explicando a abordagem adotada para otimizar a eficácia do algoritmo. Por fim, a última seção é dedicada aos experimentos realizados, onde os resultados obtidos serão analisados para validar as modificações propostas.

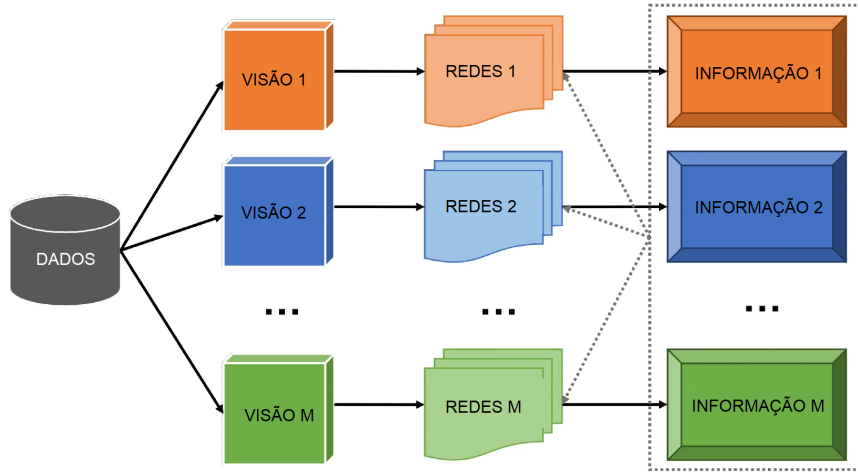
2. Agrupamento de dados com múltiplas visões

Em relação ao agrupamento tradicional de uma única visão, o agrupamento multi-visões possui uma técnica mais sofisticada, já que aproveita as informações complementares das múltiplas representações dos dados. Segundo [Khan e et al., 2022], os agrupamentos utilizam coletas do mundo real como informações externas, chamadas de informações de visualização múltipla. A representação com várias perspectivas possibilita a junção simultânea de informações com diferentes formatos sobre o mesmo objeto. Por exemplo, um ser humano pode ser visualizado por imagem de rosto, assinatura ou esboço.

Segundo [Pfeifer et al., 2023], os dados com múltiplas visões possibilitam a identificação de padrões por intermédio de suas informações, permitindo especificar grupos de assuntos ou objetos. Consolidando a ideia central do agrupamento multi-visão que é combinar as diferentes visões com o intuito de otimizar o agrupamento, formando agrupamentos mais precisos.

Com o aumento dos tipos e quantidade dos dados, a abordagem a ser utilizada é a de múltiplos recursos. Porém, os métodos de agrupamento são construídos com base no acesso de visualização única, técnicas como agrupamento espectral, fatoração de matrizes, k-means, aprendizado por subespaço e também o uso de grafos. Uma das estratégias é a concatenação de características em um espaço e a posterior aplicação de método de visualização única, tornando, segundo [Liu et al., 2024], um obstáculo para encontrar resultados satisfatórios. Procurando otimizar, muitos métodos foram derivados no intuito de adequar a abordagem para o cenário atual de dados multi-visão. Na Figura 1, está exemplificado a base de dados sendo representada por várias visões.

Figura 1: Agrupamento de dados com múltiplas visões

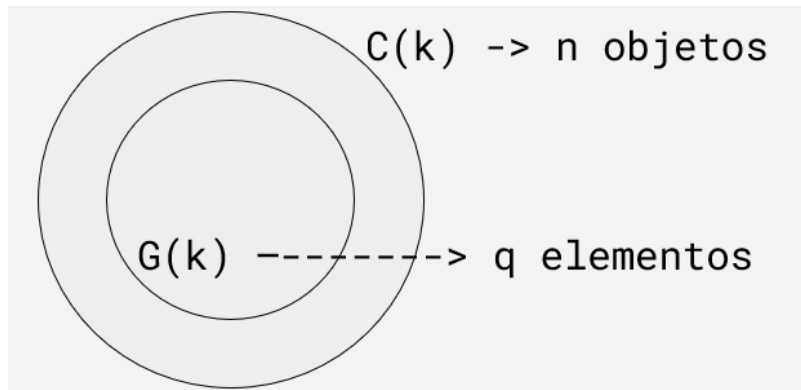


Fonte : adaptado de Fang et al. [2023]

3. Algoritmo

O algoritmo de agrupamento dinâmico baseado em múltiplas matrizes de dissimilaridade (MRDCA) desenvolvido por [Carvalho et al., 2012] é projetado para dados relacionais e utiliza as matrizes de dissimilaridade para medir as distâncias entre os dados. Sua abordagem considera um particionamento *hard*, sendo cada objeto atribuído a um único agrupamento, implicando em uma classificação mais objetiva. O método utiliza múltiplas matrizes de dissimilaridade usadas para análise das distâncias entre os dados e a prototipagem, representação de cada agrupamento $C(k)$ por um protótipo $G(k)$. O protótipo possui cardinalidade fixa dentro do conjunto total de n elementos: $E = \{e_1, \dots, e_n\}$. Como segue representado na imagem:

Figura 2: Representação da prototipagem



Fonte : o autor

O objetivo do algoritmo é encontrar a partição ótima de $P = (C_1, \dots, C_k)$ dos elementos do conjunto E em K agrupamentos. A análise de soluções são consequência da observação do comportamento da seguinte função objetivo:

$$J = \sum_{k=1}^K \sum_{e_i \in C_k} \sum_{j=1}^P D_j(e_i, G_k) = \sum_{k=1}^K \sum_{e_i \in C_k} \sum_{j=1}^P \sum_{e \in G_k} d_j(e_i, e) \quad (1)$$

Fonte: [Carvalho et al., 2012]

A função J faz uma avaliação da quantidade de agrupamentos K com relação aos objetos i em cada k , considerando a quantidade de matrizes j variando de 1 até p . Na primeira parte da equação, existe a relação dos elementos e_i do *cluster* com o representante G_k do *cluster*. Enquanto que na segunda parte, a igualdade revela a equivalência com a análise considerando um elemento e do representante G_k e medindo a distância do objeto do agrupamento para o objeto do representante. Como é definido por [Carvalho et al., 2012] no seguinte algoritmo:

Algorithm 1 Algoritmo de Clusterização Dinâmica Rígida para Dados Relacionais

```

1: Inicialização:
2: Definir o número de clusters  $K$ ;
3: Definir a cardinalidade  $1 \leq q \leq n$  dos protótipos  $G_k, k = 1, \dots, K$ ;
4: Definir  $t = 0$ ;
5: Selecionar aleatoriamente  $K$  protótipos distintos  $G_k^{(0)} \in E(q)$ ;
6: Atribuir cada objeto  $e_i$  ao protótipo mais próximo para obter a partição inicial  $P^{(0)} = \{C_1^{(0)}, \dots, C_K^{(0)}\}$ ;
7: repeat
8:    $t \leftarrow t + 1$ ;
9:   Etapa 1: Cálculo dos melhores protótipos
10:  for  $k = 1$  to  $K$  do
11:    Calcular o novo protótipo  $G_k^{(t)} = \arg \min_{G \in E(q)} \sum_{e_i \in C_k^{(t-1)}} \sum_{e \in G} d(e_i, e)$ ;
12:  end for
13:  Etapa 2: Definição da melhor partição
14:  Definir  $P^{(t)} = P^{(t-1)}$ ;
15:  Definir alterado  $\leftarrow$  falso;
16:  for  $i = 1$  to  $n$  do
17:    Determinar o cluster atual  $C_m^{(t)}$  para  $e_i$ ;
18:    Encontrar o cluster vencedor  $C_k^{(t)}$  tal que:

$$k = \arg \min_{1 \leq h \leq K} \sum_{e \in G_h^{(t)}} d(e_i, e);$$

19:    if  $k \neq m$  then
20:      Mover  $e_i$  de  $C_m^{(t)}$  para  $C_k^{(t)}$ ;
21:      alterado  $\leftarrow$  verdadeiro;
22:    end if
23:  end for
24: until alterado = falso

```

O algoritmo pode ser ainda mais específico e preciso com a atribuição de diferentes pesos de relevância para as matrizes de dissimilaridade.

3.1. Algoritmo ponderado

O algoritmo de agrupamento rígido dinâmico com peso de relevância para cada matriz de dissimilaridade estimada localmente (MRDCA-RWL) tem como seu objetivo principal encontrar uma partição $P = (C_1, C_2, \dots, C_K)$ de um conjunto de dados E em K agrupamentos, sendo

cada agrupamento representado por protótipos G_k . O algoritmo opera otimizando um critério de adequação que mede a conformidade dos dados em relação a esses protótipos. É uma variante do MRDCA, apresentada por [Carvalho et al., 2012]. Considerando diferentes pesos para as matrizes, temos o seguinte critério de adequação:

$$J = \sum_{k=1}^K \sum_{e_i \in C_k} D_{\lambda_k}(e_i, G_k) = \sum_{k=1}^K \sum_{e_i \in C_k} \sum_{j=1}^P \lambda_{kj} \sum_{e \in G_k} d_j(e_i, e) \quad (2)$$

Fonte: [Carvalho et al., 2012]

Durante seu funcionamento, o MRDCA-RWL calcula um critério de adequação J que combina as dissimilaridades dos objetos em cada cluster, ponderando-as pelos respectivos pesos de relevância. O algoritmo opera de forma iterativa, realizando ajustes nos protótipos G_k e nos pesos de relevância até que uma condição de convergência, como a estabilização dos agrupamentos ou a minimização do critério de adequação, seja alcançada.

O diferencial do algoritmo MRDCA-RWL é o ajuste automático e iterativo de pesos de relevância. O algoritmo aprende quais matrizes possuem mais relevância na formação dos agrupamentos, realizando uma adaptação dinâmica com relação às características dos dados. A variante permite a concentração em atributos mais informativos com relação aos objetivos, resultando em agrupamentos mais coesos. A principal diferença observada na função objetivo é a presença de pesos.

Os pesos de relevância λ_{kj} estão relacionados com uma matriz de dissimilaridade D_j no contexto de um algoritmo de agrupamento. O λ_{kj} associa-se com a importância da matriz D_j para a formação do agrupamento C_k . O cálculo de λ_{kj} considera a dissimilaridade entre os elementos dentro do agrupamento C_k e um protótipo G_k . Um valor maior de λ_{kj} significa que a matriz de dissimilaridade D_j é mais relevante para o agrupamento C_k ; se os elementos estão mais próximos do protótipo segundo essa matriz, isso resulta em um peso mais alto. Assim, o algoritmo dá mais importância a essa matriz ao formar e otimizar os agrupamentos. No estudo de [Carvalho et al., 2012] foram estimados os valores de λ_{kj} com base no método de multiplicadores de Lagrange. Como segue:

$$\lambda_{kj} = \frac{\left\{ \prod_{h=1}^p \left[\sum_{e_i \in C_k} D_h(e_i, G_k) \right] \right\}^{1/p}}{\sum_{e_i \in C_k} D_j(e_i, G_k)} = \frac{\left\{ \prod_{h=1}^p \left[\sum_{e_i \in C_k} \sum_{e \in G_k} d_h(e_i, e) \right] \right\}^{1/p}}{\sum_{e_i \in C_k} \sum_{e \in G_k} d_j(e_i, e)} \quad (3)$$

Fonte: [Carvalho et al., 2012]

Como definido por [Carvalho et al., 2012], o seguinte algoritmo apresenta o algoritmo ponderado:

Algorithm 2 Algoritmo de Clusterização Dinâmica Rígida com Peso de Relevância

1: **Inicialização:**

2: Definir o número de clusters K ;

3: Definir a cardinalidade $1 \leq q \leq n$ dos protótipos $G_k, k = 1, \dots, K$;

4: Definir $t = 0$;

5: Selecionar aleatoriamente K protótipos distintos $G_k^{(0)} \in E(q)$;

6: Atribuir cada objeto e_i ao protótipo mais próximo para obter a partição inicial $P^{(0)} = \{C_1^{(0)}, \dots, C_K^{(0)}\}$;

7: **repeat**

8: $t \leftarrow t + 1$;

9: **Etapa 1: Cálculo dos melhores protótipos**

10: **for** $k = 1$ to K **do**

11: Calcular o novo protótipo:

$$G_k^{(t)} = \arg \min_{G \in E(q)} \sum_{e_i \in C_k^{(t-1)}} \sum_{j=1}^p \lambda_{kj} \sum_{e \in G} d_j(e_i, e);$$

12: **end for**

13: **Etapa 2: Atualização dos pesos de relevância**

14: **for** $k = 1$ to K **do**

15: **for** $j = 1$ to p **do**

16: Calcular:

$$\lambda_{kj}^{(t)} = \frac{\left(\prod_{h=1}^p \sum_{e_i \in C_k^{(t-1)}} D_h(e_i, G_k^{(t)}) \right)^{1/p}}{\sum_{e_i \in C_k^{(t-1)}} D_j(e_i, G_k^{(t)})};$$

17: **end for**

18: **end for**

19: **Etapa 3: Definição da melhor partição**

20: Definir $P^{(t)} = P^{(t-1)}$;

21: Definir alterado $\leftarrow$ falso;

22: **for** $i = 1$ to n **do**

23: Determinar o cluster atual $C_m^{(t)}$ para e_i ;

24: Encontrar o cluster vencedor $C_k^{(t)}$ tal que:

$$k = \arg \min_{1 \leq h \leq K} \sum_{j=1}^p \lambda_{hj}^{(t)} \sum_{e \in G_h^{(t)}} d_j(e_i, e);$$

25: **if** $k \neq m$ **then**

26: Mover e_i de $C_m^{(t)}$ para $C_k^{(t)}$;

27: alterado $\leftarrow$ verdadeiro;

28: **end if**

29: **end for**

30: **until** alterado = falso

4. Adaptação proposta

A estimativa de peso compõe uma parte crucial para a análise de dados com múltiplas visões ponderados. Segundo [Liu e Lin, 2023], a distribuição de pesos adequados para as diferentes vistas melhora a precisão nos resultados e o desempenho dos algoritmos de clusterização, em contraste com os métodos tradicionais que utilizam pesos iguais ou realizam ajustes manuais seguindo parâmetros de observação subjetiva, possuindo limitações quando há ausência de conhecimento prévio dos dados.

A estimativa do peso de relevância atua diretamente na clusterização baseada em múltiplas matrizes de dissimilaridade, possuindo diferentes abordagens responsáveis pelo cálculo, impactando no desempenho do algoritmo e agrupamentos gerados. A proposta de modificação do algoritmo MRDCA-RWL considera a utilização de média geométrica acrescida de um fator de regularização para calcular pesos de relevância. Tal abordagem contribui para a proporção de cada peso em relação ao conjunto total de dados, promovendo um balanceamento mais estável e sendo útil em cenários com dados de grande variabilidade. Em contraste, o método original utiliza a média aritmética da soma de dissimilaridades dentro de cada agrupamento, considerando de maneira proporcional as dissimilaridades individuais sem considerar uma relação multiplicativa, possuindo portanto uma abordagem mais vulnerável a variações. A média geométrica proposta apresenta menor vulnerabilidade em relação a valores extremos, possuindo maior robustez numérica.

$$J = \sum_{k=1}^K \sum_{e_i \in C_k} \sum_{j=1}^p \lambda_{kj} \sum_{e \in G_k} d_j(e_i, e)$$

$$\lambda_{kj} = \frac{\left(\prod_{h=1}^p \sum_{e_i \in C_k} D_h(e_i, G_k) + \epsilon \right)^{\frac{1}{p}}}{\sum_{e_i \in C_k} D_j(e_i, G_k) + \epsilon}$$

onde:

- J é a função objetivo a ser minimizada,
- K é o número de clusters,
- C_k representa o conjunto de objetos no cluster k ,
- G_k é o protótipo do cluster k ,
- $d_j(e_i, e)$ é a dissimilaridade entre e_i e e para a métrica j ,
- λ_{kj} é o peso de relevância da métrica j no cluster k ,
- ϵ é o fator de regularização para evitar instabilidades numéricas.

Essa adaptação utiliza a média geométrica para estimar os pesos de relevância de forma mais robusta em relação a variações nos dados.

4.1. Experimentos

Nesta seção, apresentamos os experimentos realizados com o algoritmo MRDCA-RWL, bem como sua variação com um novo método de estimativa de pesos. O objetivo é examinar a eficácia dessa nova abordagem em vários conjuntos de dados e verificar seu efeito na formação de clusters. Os experimentos foram conduzidos em três conjuntos de dados populares adotados na literatura: Segmentação de Imagem, Íris e Conjunto de Dados Multi-Características.

Conjunto de dados	Número de vistas	Número de classes	Número de objetos	Número de variáveis
Imagens	16	7	2310	19
Multiple features	6	10	2000	649
Iris	1	3	150	4

Tabela 1: Tabela de Conjuntos de Dados

A principal métrica utilizada para comparação foi a minimização da função objetivo, que considerou a qualidade da estrutura interna dos agrupamentos [Han et al., 2006; Chandra et al., 2013]. Além disso, a Taxa Geral de Erro de Clusterização (OERC) e a medida F foram precisas, o que mediu a qualidade dos clusters em relação às classes de referência [Xu e Wunsch, 2005; Liu et al., 2024]. Para realização dos testes foram feitas 50 repetições do algoritmo original em comparação com a mesma quantidade aplicada no algoritmo modificado.

Tabela 2: Resultados para Image Segmentation

Métrica	Original	Proposto
OERC	0,323810	0,323810
F-Measure	0,676755	0,676755
Melhor Objetivo	5115,685899	5115,685899

Os resultados mostram que a modificação preservou a qualidade do agrupamento obtida pelo método original, mantendo métricas consistentes e garantindo estabilidade no desempenho, mesmo com a alteração na estimativa de pesos.

Tabela 3: Resultados para Iris

Métrica	Original	Proposto
OERC	0,0600	0,0600
F-Measure	0,9400	0,9400
Melhor Objetivo	935,0051	935,0051

Semelhante ao resultado anterior, Observa-se que, para o conjunto de dados Iris, o método proposto manteve desempenho equivalente ao original, evidenciando robustez e ausência de perdas na qualidade dos agrupamentos.

Tabela 4: Resultados para Multiple Features

Métrica	Original	Proposto
OERC	0,106000	0,106000
F-Measure	0,893012	0,893012
Melhor Objetivo	625050,949879	625050,949879

No conjunto Multiple Features, as métricas permaneceram estáveis em relação ao método original, confirmando que a adaptação não compromete a eficácia do algoritmo e preserva sua confiabilidade

5. Considerações Finais

Os experimentos realizados demonstram que a modificação proposta, baseada na utilização da média geométrica com fator de regularização para a estimativa de pesos de relevância, preserva integralmente o desempenho do algoritmo MRDCA-RWL original. As métricas avaliadas (OERC, F-Measure e função objetivo) mantiveram-se estáveis em todos os conjuntos de dados, evidenciando que a adaptação não compromete a qualidade dos agrupamentos.

A manutenção da eficácia indica que a abordagem proposta é robusta e estável, permitindo ajustes na forma de cálculo dos pesos sem introduzir perdas no resultado final. Essa característica amplia a flexibilidade de aplicação do método, garantindo confiabilidade mesmo diante de alterações na formulação.

Trabalhos futuros podem explorar variações adicionais do cálculo de pesos e novas estratégias de regularização, de modo a ampliar ainda mais a aplicabilidade e a adaptabilidade do MRDCA-RWL em diferentes contextos de agrupamento de dados.

Referências

- Carvalho, F. D. A. D., Lechevallier, Y., e Melo, F. M. D. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, 45(1):447–464.
- Chandra, A., Kumar, A., e Soni, P. (2013). Performance evaluation of clustering algorithms for data mining applications. *International Journal of Computer Applications*, 66(10):43–47.
- Fang, U. et al. (2023). A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge & Data Engineering*, 35(12):12350–12368. ISSN 1558-2191. URL <https://doi.ieeecomputersociety.org/10.1109/TKDE.2023.3270311>.
- Gong, X., Zhang, Y., e Liu, J. (2019). Semantic weighted multi-view clustering for web content. *IEEE Access*, 7:128097–128113.
- Han, J., Kamber, M., e Pei, J. (2006). Data mining: Concepts and techniques. *Morgan Kaufmann*.
- Hua, L. e Wu, K.-L. (2021). A novel brain mri image segmentation method using an improved multi-view fuzzy c-means clustering algorithm. *Frontiers in Neuroscience*, 15:662674.
- Hussain, S. F., Khan, K., e Jillani, R. (2022). Weighted multi-view co-clustering (wmvcc) for sparse data. *Applied Intelligence*, 52(1):398–416.
- Jiang, Q., Zhou, G., e Zhao, Q. (2024). Semi-supervised multi-view concept decomposition. *Expert Systems with Applications*, 241:122572.

- Khan, M. A. e et al. (2022). Multi-view clustering based on multiple manifold regularized non-negative sparse matrix factorization. *IEEE Access*, 10:113249–113259.
- Li, M. e Zhang, X. (2019). Robust multi-view clustering with a unified weight learning paradigm. *IEEE Access*, 7:177965–177973.
- Lian, H., Zhang, X., e Li, X. (2021). Partial multiview clustering with locality graph regularization. *International Journal of Intelligent Systems*, 36(6):2991–3010.
- Liu, Q., Wang, X., e Li, Z. (2024). Adaptive weighted multi-view subspace clustering method for recognizing urban functions from multi-source social sensing data. *Geo-spatial Information Science*, 1:1–25.
- Liu, S. S. e Lin, L. (2023). Adaptive weighted multi-view clustering. In *Conference on Health, Inference, and Learning*, PMLR, p. 19–36.
- Pfeifer, B., Bloice, M. D., e Schimek, M. G. (2023). Parea: multi-view ensemble clustering for cancer subtype discovery. *Journal of Biomedical Informatics*, 143:104406.
- Qian, P., Wei, X., Zhang, X., e et al. (2018). Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes. *IEEE Access*, 6:28594–28610.
- Sharma, S., Nayak, R., e Bhaskar, A. (2024). Multi-view feature engineering for day-to-day joint clustering of multiple traffic datasets. *Transportation Research Part C: Emerging Technologies*, 162:104607.
- Shi, S., Zhang, Z., e Liu, W. (2021). Fast multi-view clustering via prototype graph. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):443–455.
- Xu, R. e Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.
- Zhang, K. e Gionis, A. (2023). Incomplete multi-view clustering based on low-rank representation with adaptive graph regularization. *Soft Computing*, 27(11):7131–7146.
- Zhu, C. e Wang, Z. (2019). Semi-supervised soft margin consistency based multi-view maximum entropy discrimination. *Applied Computing and Informatics*, 15(2):172–181.