UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Investigação de métodos para estimação de pesos de relevância para agrupamento de dados relacionais com múltiplas visões

Dissertação de Mestrado

Vitória Teles da Silva

São Cristóvão – Sergipe

2026

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vitória Teles da Silva

Investigação de métodos para estimação de pesos de relevância para agrupamento de dados relacionais com múltiplas visões

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Renê Pereira de Gusmão
Coorientador(a): André Britto de Carvalho

São Cristóvão – Sergipe

2026

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Silva, Vitória Teles da
S586i Investigação de métodos para estimação de pesos de relevância para agrupamento de dados relacionais com múltiplas visões / Vitória Teles da Silva ; orientador Renê Pereira de Gusmão. - São Cristóvão, 2026.
80 f.; il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2026.

1. Análise por agrupamento. 2. Relevância. 3. Computação. I. Gusmão, Renê Pereira de orient. II. Título.

CDU 004

Dedico esse trabalho ao homem forte e disciplinado de carreira militar, que sempre me incentivou e me ensinou muito sobre a vida: o meu pai.

Agradecimentos

Agradeço primeiramente a Deus, pois a fé é capaz de mover montanhas.

Ao meu orientador, Prof.Dr.Renê Pereira de Gusmão e meu Coorientador Prof.Dr.André Britto de Carvalho, ambos não só pela orientação e todos os ensinamentos, mas também por entenderem todos os contratempos que tive durante o mestrado.

Aos meus pais, Roseli e Edivaldo, que foram de crucial importância para que eu conseguisse voltar ao estado de Sergipe. Não tenho palavras para agradecer o apoio que recebi desde antes do inicio.

A mim, por não ter desistido, por ter lutado e enfrentado todas as dificuldades durante os dois anos de mestrado e por insistir em continuar a carreira acadêmica.

A secretária Elaine e ao atual coordenador Prof.Dr.Leonardo Nogueira Matos. Ambos me ajudaram bastante durante todo o processo, Elaine sempre ajudando com a parte burocrática, sugestões, avisos, entre outros... enquanto que o Prof.Leonardo me ajudou em relação aos experimentos, conferências e também com conselhos preciosos.

A todos os meus colegas de mestrado, mas em especial aqueles que me auxiliaram, seja com problemas acadêmicos ou com uma boa conversa e um cafezinho. São eles: Dênis, Eduardo, Mariana, Verônica, Caetano, Ítalo, Anderson e Osmário. Com vocês, a caminhada foi bem mais leve.

A Dhonatas, que sempre me ajudou e incentivou, sendo um grande parceiro.

Por fim, gostaria de agradecer a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que financiou essa pesquisa.

Este trabalho não seria possível, assim como o término do mestrado sem a ajuda das pessoas e instituição acima citadas, deixo aqui registrada a minha eterna gratidão.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Nenhum mal te sucederá, nem praga alguma chegará à tua tenda.
Porque aos anjos dará ordem a teu respeito, para te guardarem em todos os teus caminhos.
Salmos 91:10,11

Resumo

O agrupamento de dados com múltiplas visões é caracterizado como o desafio de agrupar dados similares e separar os dados distintos, utilizando diferentes representações, como vetores de atributos, matrizes de características, grafos ou imagens, que, em conjunto, oferecem informações complementares. Essa abordagem é aplicada em cenários reais como análise de redes sociais, classificação de imagens médicas, sistemas de recomendação e integração de dados multimídia. 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. Os experimentos mostram que o desempenho dos algoritmos de agrupamento multivisão depende fortemente das métricas de distância e das estratégias de estimação de pesos. Enquanto o MRDCA-RWL é mais sensível à métrica adotada, métodos como MVKKM e MVSC apresentam forte dependência conjunta entre distância e pesos, reforçando a importância da escolha adequada dos parâmetros. Dado o exposto, este trabalho relaciona a necessidade de interpretação adequada do uso de dados com múltiplas visões no contexto de vida atual, tendo como objetivo investigar diferentes métodos para estimação dos pesos de relevância. Com isso, os pesos são aplicados nas diferentes visões no contexto de agrupamento de dados relacionais, discutindo tanto a qualidade dos resultados quanto a eficiência algorítmica, em comparação com a complexidade de soluções concorrentes. A partir disso, os resultados obtidos nos experimentos demonstram a influência que os algoritmos sofrem pela escolha de metodologia, apresentando o MRDCA-RWL com comportamento mais estável, enquanto que MVKKM, MVSC e TW-KM demonstram sensibilidade às mudanças de configurações.

Palavras-chave: Agrupamento. Múltiplas visões. Métricas de distância. Pesos de relevância.

Abstract

Multiview data clustering is characterized as the challenge of grouping similar data and separating dissimilar data by using different representations, such as attribute vectors, feature matrices, graphs, or images, which together provide complementary information. This approach is applied in real-world scenarios such as social network analysis, medical image classification, recommender systems, and multimedia data integration. Data clustering is one of the most relevant tools for data analysis, and its importance is related to the faithful representation of real-world data, with the combined use of multiple views providing greater accuracy when compared to the use of a single view. The experiments show that the performance of multiview clustering algorithms strongly depends on the distance metrics and weight estimation strategies. While MRDCA-RWL is more sensitive to the adopted distance metric, methods such as MVKKM and MVSC exhibit a strong joint dependence on both distance metrics and weights, reinforcing the importance of selecting appropriate parameters. Given the above, this work addresses the need for a proper interpretation of the use of multiview data in the context of contemporary applications, aiming to investigate different methods for estimating relevance weights. These weights are then applied to the different views in the context of relational data clustering, discussing both the quality of the results and the algorithmic efficiency in comparison with the complexity of competing solutions. Based on this, the experimental results demonstrate the influence of the methodological choices on the behavior of the algorithms, with MRDCA-RWL exhibiting more stable behavior, whereas MVKKM, MVSC, and TW-KM demonstrate sensitivity to configuration changes.

Keywords: Clustering. Multiple views. Distance metrics. Relevance weights.

Lista de ilustrações

Lista de tabelas

Lista de algoritmos

Lista de abreviaturas e siglas

MRDCA	Algoritmo de Agrupamento Dinâmico Baseado em Múltiplas Matrizes de Dissimilaridade (sem pesos)
MRDCA-RWL	Algoritmo de Agrupamento Dinâmico com Peso de Relevância Local para cada Matriz de Dissimilaridade
MVKKM	Multi-ViewKernel K-Means
TW-KM	Two-Level Variable Weighting K-Means
MVSC	Multi-view Spectral Clustering

Lista de símbolos

λ	Peso de relevância
λ_j	Peso de relevância da j-ésima vista
λ_{kj}	Peso de relevância da j-ésima matriz de dissimilaridade no k-ésimo agrupamento
γ_v	Peso da v-ésima vista
β_v	Peso da v-ésima vista no algoritmo TW-KM
$w_{j\|v}$	Peso da variável j na vista v
α	Nível de significância estatística
χ^2	Estatística do teste de Friedman
p-valor	Probabilidade associada ao teste estatístico
CD	Diferença Crítica do teste de Bonferroni-Dunn
$Rank$	Posto médio obtido por um método no teste de Friedman
$\mathbb{R}$	Conjunto dos números reais
$\mathbb{R}^n$	Espaço vetorial de dimensão n
$\mathbf{x}$	Vetor de características do primeiro objeto
$\mathbf{y}$	Vetor de características do segundo objeto
x_i	i-ésima coordenada do vetor $\mathbf{x}$
y_i	i-ésima coordenada do vetor $\mathbf{y}$
E	Conjunto dos objetos a serem agrupados
e_i	i-ésimo objeto do conjunto de dados
$\mathcal{E}^{(q)}$	Conjunto de todos os subconjuntos de E com cardinalidade q
P	Partição do conjunto de objetos
C_k	k-ésimo agrupamento (*cluster*)
G_k	Protótipo do k-ésimo agrupamento

$G_{k(i)}$	Protótipo do agrupamento ao qual o objeto e_i pertence
q	Cardinalidade do protótipo
K	Número de agrupamentos
V	Número de vistas
p	Número de matrizes de dissimilaridade
n	Número de objetos (ou amostras)
t	Iteração do algoritmo
D_j	j-ésima matriz de dissimilaridade
$d_j(e_i, e)$	Dissimilaridade entre os objetos e_i e e
$D_j(e_i, G_k)$	Dissimilaridade entre o objeto e_i e o protótipo G_k
$\bar{D}_j$	Média das dissimilaridades da vista j
J	Função objetivo (critério de adequação)
J_{kj}	Critério de adequação da vista j no agrupamento k
$K^{(v)}$	Matriz kernel (ou de similaridade) da vista v
K_γ	Matriz kernel combinada ponderada
H	Matriz de atribuição dos agrupamentos no MVKKM
H_j	Entropia da vista j
X	Conjunto de dados
x_{ij}	Valor da variável j do objeto i
c_k	Centróide do agrupamento k
c_{kj}	Coordenada j do centróide k
u_{ik}	Variável de atribuição do objeto i ao agrupamento k
G_v	Conjunto de atributos da vista v
$L^{(v)}$	Matriz Laplaciana da vista v
$U^{(v)}$	Matriz dos autovetores da vista v
U^*	Matriz de representação consenso

$M^{(v)}$	Matriz modificada utilizada na atualização dos autovetores
p_{kj}	Proporção normalizada da dissimilaridade da vista j no agrupamento k
Z_j	Fator de normalização das dissimilaridades da vista j
μ_j	Média global das dissimilaridades da vista j
σ_j^2	Variância das dissimilaridades da vista j
I	Matriz identidade
$\mathrm{tr}(\cdot)$	Operador traço de uma matriz
Δ	Simplexo dos pesos das vistas
ϵ	Tolerância para critério de parada
ε	Tolerância para convergência
T	Número máximo de iterações
$d_{\text{euclidiana}}(\mathbf{x}, \mathbf{y})$	Distância euclidiana
$d_{\text{manhattan}}(\mathbf{x}, \mathbf{y})$	Distância Manhattan
$d_{\text{Chebyshev}}(\mathbf{x}, \mathbf{y})$	Distância de Chebyshev
$d_{\text{cosseno}}(\mathbf{x}, \mathbf{y})$	Distância cosseno
$\lvert x_i - y_i \rvert$	Diferença absoluta entre coordenadas
$\sum$	Operador de somatório
$\prod$	Operador de produtório
$\max$	Operador de máximo
$\arg\min$	Argumento que minimiza uma função
$\log$	Logaritmo
$\sqrt{\ }$	Raiz quadrada
$\lVert \mathbf{x} \rVert$	Norma do vetor $\mathbf{x}$
$\lVert \mathbf{y} \rVert$	Norma do vetor $\mathbf{y}$
$\mathbf{x} \cdot \mathbf{y}$	Produto escalar entre vetores

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidata: VITÓRIA TELES DA SILVA

Em 23 dias do mês de julho do ano de dois mil e vinte seis, com início às 10:00hs, realizou-se na Sala de Seminários do PROCC da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **VITÓRIA TELES DA SILVA** que desenvolveu o trabalho intitulado: **"***Investigação de métodos para estimação de pesos de relevância para agrupamento de dados relacionais com múltiplas visões***",** sob a orientação do Prof. Dr. **Renê Pereira Gusmão.** A Sessão foi presidida pelo Prof. Dr. **Renê Pereira Gusmão** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, o Dr. **Tiago Buarque Assunção de Carvalho e, em seguida, Dr. Leonardo Nogueira Matos e Dr. André Britto de Carvalho**. Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) **Aprovada** "*(aprovado/reprovado)*". Atendidas as exigências da Instrução Normativa 05/2019/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 04/2021/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos",* 23 de julho de 2026.

gov.br Documento assinado digitalmente
RENE PEREIRA DE GUSMAO
Data: 24/07/2026 08:51:26-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Renê Pereira Gusmão
(PROCC/UFS)
Presidente

gov.br Documento assinado digitalmente
LEONARDO NOGUEIRA MATOS
Data: 24/07/2026 09:00:35-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Leonardo Nogueira Matos
(PROCC/UFS)
Examinador Interno

gov.br Documento assinado digitalmente
TIAGO BUARQUE ASSUNCAO DE CARVALHO
Data: 24/07/2026 08:55:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr.Tiago Buarque Assunção de Carvalho
(UFAPE)
Examinador Externo ao programa

gov.br Documento assinado digitalmente
ANDRE BRITTO DE CARVALHO
Data: 24/07/2026 13:04:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. André Britto de Carvalho
(PROCC/UFS)
Examinador Interno

Vitória Teles da Silva
Candidata

gov.br Documento assinado digitalmente
VITORIA TELES DA SILVA
Data: 24/07/2026 14:29:42-0300
Verifique em https://validar.iti.gov.br

Sumário

1 Introdução

Com o atual crescimento acelerado de informações, os dados atingiram uma grande escala (JIANG; ZHOU; ZHAO, 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 et al., 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, Bloice e Schimek (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; ZHOU; ZHAO, 2024). Dessa forma, dados com múltiplas visões são a representação de um mesmo objeto derivado de múltiplas perspectivas. Por exemplo, em uma base de imagens, um mesmo objeto pode ser descrito por diferentes atributos, como textura, cor e forma. Para visualização, considere a base Iris, utilizada nos experimentos deste trabalho. Cada flor pode ser representada por duas visões: Visão 1, formada pelas medidas da sépala, e Visão 2, pelas medidas da pétala. O algoritmo analisa ambas as representações e combina suas informações, atribuindo maior peso à visão que melhor diferencia os grupos, produzindo assim um agrupamento mais representativo. Em contraste com o exemplo anterior, há a utilização de dados biomédicos: um paciente pode ser representado por exames de imagem, sinais fisiológicos e informações clínicas. 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 et al. (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 et al., 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, Khan e Jillani (2022).

1.1 Justificativa

Segundo Gong et al. (2019), o uso combinado de diferentes visões possui melhores resultados se comparado com o uso de uma única visão. Entretanto, Imani et al. (2024) afirmam que o agrupamento de dados com múltiplas visões apresenta dificuldades em integrar de forma eficaz as diferentes vistas e obter uma representação integrada. Dessa forma, em algumas situações, a utilização de várias visões pode comprometer o desempenho do algoritmo, uma vez que algumas visões podem não agregar informações complementares relevantes. Os dados com diferentes perspectivas são comuns em aplicações variadas e o aprimoramento do agrupamento pode ser feito analisando a relevância de cada visão.

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 et al. (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 analise de dados, é necessário explorar diferentes métodos de estimativa 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.

No entanto, a eficácia dessas estimativas depende fortemente das métricas adotadas para avaliar o desempenho, as quais podem levar a diferentes conclusões sobre a qualidade dos resultados. 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 é possível notar em Lian et al. (2021), na qual a comparação é essencial para verificar a eficácia do algoritmo proposto, enquanto que Sharma, Nayak e Bhaskar (2024) refletem sobre o uso de pesos impactando na complexidade computacional dos algoritmos e a respectiva relação com os métodos utilizados.
Além disso, o trabalho recente de Moujahid e Dornaika (2025) apresenta uma revisão atualizada sobre os desafios abertos e direções promissoras de pesquisa na área de agrupamento multi-vista. No trabalho, há a enfatização de problemas ainda não resolvidos de forma total, podendo ser citados: escalabilidade para grandes volumes de dados, tratamento de dados incompletos, heterogeneidade entre visões, alta dimensionalidade e especialmente a necessidade de mecanismos para estimar a importância de cada visão, ou seja, estratégias de ponderação.

Diante do cenário apresentado, torna-se relevante a investigação do comportamento de diferentes algoritmos quando relacionados a diferentes distâncias e métodos de ponderação. Serão analisados os resultados do agrupamento e será realizada uma abordagem comparativa que permitirá identificar quais estratégias apresentarão melhor desempenho e quais serão as suas respectivas limitações.

1.2 Problema de pesquisa

Dado o exposto, observa-se que a escolha das métricas de distância e das estratégias de estimação de pesos pode influenciar significativamente o desempenho dos algoritmos de agrupamento com múltiplas visões. Entretanto, permanece uma lacuna em relação a compreensão de como essas escolhas afetam diferentes algoritmos e conjuntos de dados e quais combinações apresentam melhor desempenho em cada contexto.

Assim, este trabalho busca investigar o impacto da combinação entre métricas de distância e métodos de estimação de pesos em algoritmos de agrupamento multivisão, procurando responder às seguintes questões:

1. Como diferentes métodos de estimação de pesos influenciam o desempenho dos algoritmos?
2. Qual o impacto das métricas de distância nos resultados obtidos?
3. Existem combinações de métricas e métodos de ponderação que apresentam melhor desempenho em determinados conjuntos de dados?
4. Quais algoritmos demonstram maior robustez frente a essas variações?

Espera-se que este trabalho tenha impacto cientifico, fornecendo contribuições para pesquisadores em aplicações práticas. Este trabalho pretende comparar estratégias com diferentes formas de atribuição de peso e métricas de distância aplicados em diferentes bases, com a expectativa de indicar quais tem melhor funcionamento em determinados contextos especificados

ao decorrer deste trabalho.

1.3 Objetivo Geral

Este trabalho tem como objetivo geral investigar diferentes abordagens para estimação dos pesos de relevância no contexto de agrupamento de dados relacionais com múltiplas visões em diferentes algoritmos:MRDCA-RWL,MVKKM,TW-KM,MVSC. 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, resultando em ganhos quantitativos em métricas como acurácia, índice Rand ajustado, NMI e F-Measure.

1.3.1 Objetivos Específicos

1. Realizar uma revisão sistemática de literatura, considerando bases relevantes, selecionando estudos relacionados com agrupamento de dados com múltiplas visões ponderadas, acompanhando o avanço na área e situando suas possíveis aplicações.

2. Identificar metodologia, algoritmo de particionamento, métricas de distância, abordagem para atribuição de pesos de relevância, métricas e metodologia de avaliação a serem utilizadas nos experimentos.

3. Realizar os experimentos com o algoritmo estabelecido, considerando as adaptações realizadas nas métricas e diferentes abordagens para estimar os pesos de relevância das visões segundo a metodologia definida.

4. Executar testes comparativos e analisar os resultados, com o intuito de identificar quais opções mais se adequam em relação ao desempenho e contribuição significativa para eficácia geral.

5. Avaliar os efeitos de diferentes métricas de distância e seus respectivos pesos estimados, utilizando aplicações em algoritmos de agrupamento de dados com múltiplas visões.

1.4 Estrutura do trabalho

A estrutura do trabalho está dividida em 4 Capítulos, contendo:

1. No Capítulo 2 será apresentada a fundamentação teórica do trabalho seguindo uma sequência da abordagem mais ampla à mais específica. Começando pela definição de agrupamento e métricas de distâncias para então definir os algoritmos escolhidos, os métodos de estimativa de pesos e as distâncias utilizadas.

2. No Capítulo 3 será apresentada uma revisão sistemática elaborada a partir de pesquisas em base de dados, além disso, será formulada uma palavra-chave para auxiliar na busca, assim como o objetivo e as questões de pesquisa. Após apresentar a estratégia de busca e seleção, será apresentada uma seção para resultados e discussão.

3. No Capítulo 4 serão apresentados os métodos de estimativa de peso utilizados nos experimentos deste estudo.

4. No Capítulo 5 será apresentada a metodologia a ser adotada no desenvolvimento deste trabalho e os resultados dos experimentos realizados.

5. No Capítulo 6 serão apresentadas as conclusões do trabalho.

2

Fundamentação Teórica

Neste capítulo serão descritos os conceitos teóricos que orientaram o trabalho como um todo, tais como termos, áreas, formulações e adaptações realizadas. Considerando aspectos do mais abrangente ao mais especifico, apresentando os algoritmos utilizados e distâncias aplicadas.

2.1 Agrupamento de dados

Segundo Khan et al. (2022), trata-se de um procedimento extenso pertencente a análise de dados que atua particionando os dados em vários grupos. Também conhecido como *clustering*, visa reunir um conjunto de objetos mais similares entre si do que com os pertencentes a outros grupos. O objetivo é a identificação de similaridades obtidas do resultado do uso de medidas específicas ou distâncias entre os dados (KALPANA; VIGNESHWARI, 2016). As similaridades identificadas possibilitam encontrar os padrões e relações entre os dados, permitindo a análise e extração de informações pertinentes.

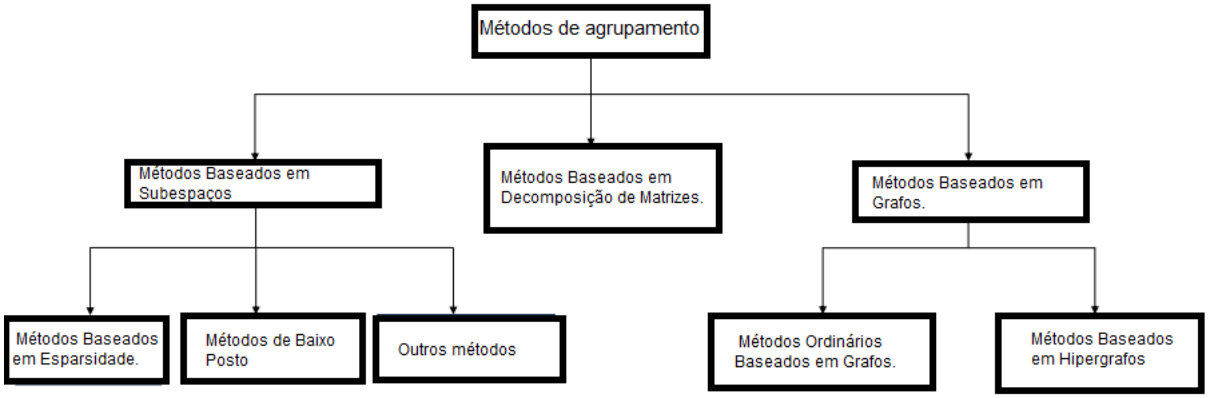
O agrupamento é uma técnica que tem aplicações com diferentes tipos de dados, como dados de texto, imagens, redes sociais e páginas na internet. Cada tipo de dado possui características distintas que influenciam na seleção do método de agrupamento mais apropriado (JIANG; ZHOU; ZHAO, 2024). A escolha do método está relacionada com a eficácia e otimização da partição, refletindo nos resultados e interpretações. Inuwa-Dutse, Liptrott e Korkontzelos (2021) fazem uma comparação entre a capacidade de análise humana frente ao desafio de escolha dos métodos de agrupamento, afirmando que processar um grande volume de dados de maneira eficiente é uma tarefa árdua.

É possível classificar os algoritmos de agrupamento a partir da sua abordagem operacional. Entre os métodos, destacam-se: o agrupamento hierárquico que está relacionado com a estrutura de árvore, agrupamento por densidade no qual há identificação de clusters densamente conectados e, além disso, os métodos de particionamento podem ainda ser categorizados em métodos hard e fuzzy.

Um dos algoritmos mais populares citados na literatura é o K-means e suas variações, responsáveis por separar dados em um número predefinido de conjuntos com o objetivo de otimizar uma dada função objetivo. Os tipos de funções são determinadas em conformidade com seu propósito operacional, estando relacionadas com a minimização ou maximização de características relacionadas aos dados e sua aplicação.

Ainda sobre a taxonomia, Yang et al. (2023) dividem os métodos de agrupamento em três grupos, sendo separados como definido na Figura 1. Nessa figura há a ilustração da taxonomia dos algoritmos de agrupamento, apresentando a divisão de classificação em subespaços, matrizes e grafos, respeitando a forma de representação de dados utilizada para identificar os agrupamentos. Como segue:

Figura 1 – Taxonomia de algoritmos de agrupamento



Fonte : adaptado de (YANG et al., 2023)

Funções baseadas em densidade analisam proximidade entre pontos para formação dos grupos, enquanto em abordagens que utilizam grafos, há a ponderação de arestas transformando a equação em uma função de custo. Métodos avançados utilizam matrizes de dissimilaridade em conjunto com a função objetivo, apontado por Bernard et al. (2020) como uma abordagem adequada para agrupamento de dados com múltiplas vistas, no qual há a combinação de diferentes fatores com o objetivo de formar grupos mais precisos.

2.1.1 Agrupamento de dados com múltiplas vistas

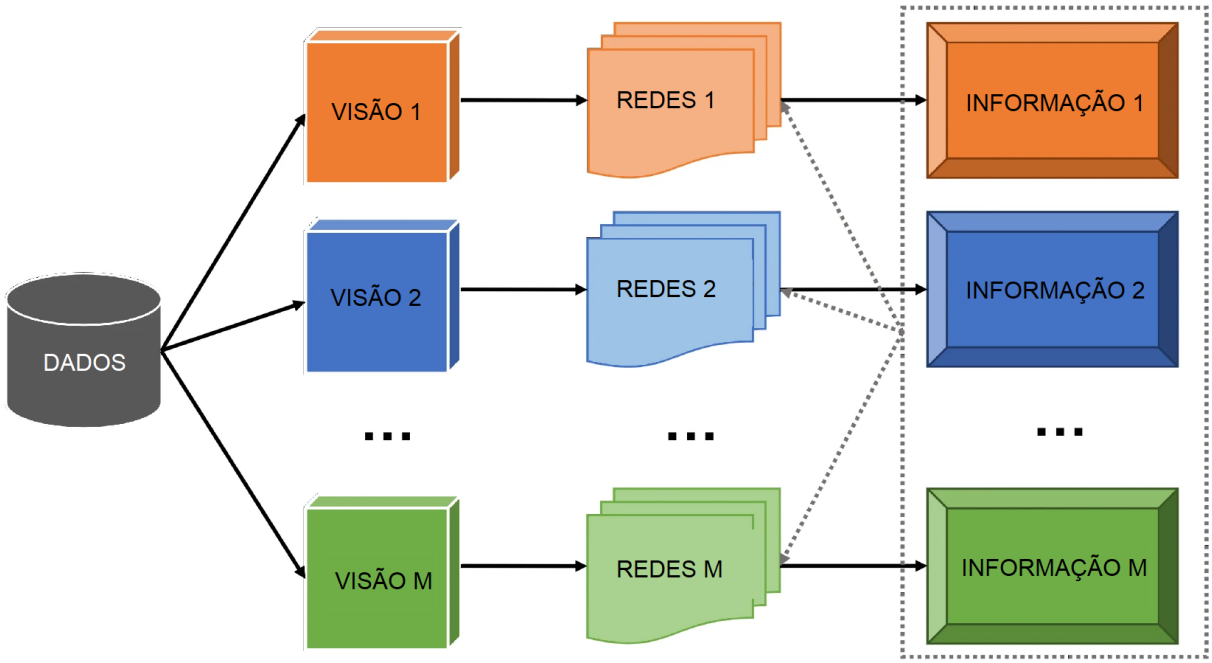
Em relação ao agrupamento tradicional de uma única visão, o agrupamento multi-vistas contrasta pela sua forma de atuação, aproveitando as informações complementares das múltiplas representações dos dados. Segundo Khan et al. (2022), os agrupamentos utilizam coletas do mundo real como informações externas, 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, Bloice e Schimek (2023), os dados com múltiplas vistas 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 vistas 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 multi-visão. 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. De modo geral, as abordagens de aprendizado multi-visão podem ser organizadas em três categorias principais: concatenação de características, fusão e abordagens distribuídas. A concatenação consiste em combinar 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. As abordagens de fusão, por sua vez, buscam integrar as informações das diferentes vistas durante o processo de aprendizado, enquanto as abordagens distribuídas tratam cada visão separadamente, combinando os resultados de forma posterior. Procurando otimizar, muitos métodos foram derivados no intuito de adequar a abordagem para o cenário atual de dados multi-visão. Na Figura 2, está exemplificado a base de dados sendo representada por várias vistas, na qual as denominações REDES1, REDES2 e REDES3 representam diferentes redes de relacionamento construídas a partir do mesmo conjunto de objetos, sendo cada uma responsável por representar um tipo distinto de informação.

Figura 2 – Exemplo de representação de dados com múltiplas vistas em um modelo de agrupamento multivisão

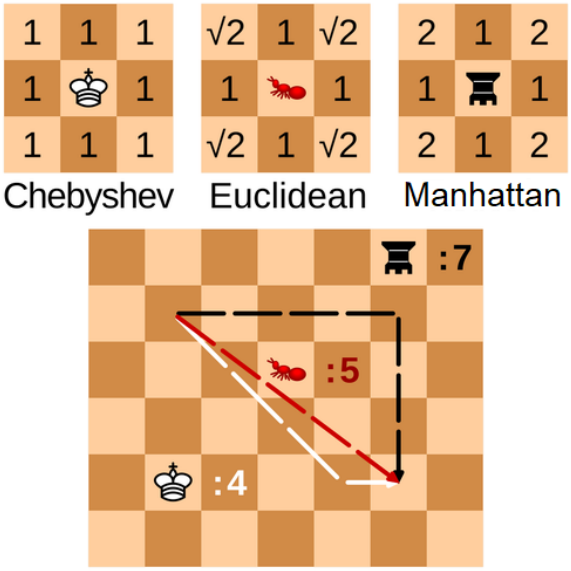


Fonte : adaptado de (FANG et al., 2023)

2.1.2 Métricas de distância e abordagens combinadas

No agrupamento, os dados se relacionam por similaridade, sendo mais próximos aqueles que compartilham semelhanças. A análise da afinidade entre os dados é percebida pela distância entre eles, cada conjunto do agrupamento possui dados similares entre si e diferentes dos outros conjuntos. Braz et al. (2020) relataram sobre "ganhos"e "perdas"no uso de uma distância em detrimento da outra, fazendo uma comparação entre distâncias e expondo as diferenças presentes em agrupamentos baseados em diferentes métricas. A seguir, são descritas as métricas de distância que foram combinadas com os métodos de agrupamento e utilizadas neste trabalho, considerando as três primeiras expostas sendo casos especiais da distância de Minkowiski, como pode ser observado em forma de exemplo na Figura 3 :

Figura 3 – Comparação entre as diferentes distâncias resultantes de casos especiais de Minkowiski



Fonte: Adaptado de (Wikimedia Commons, 2024).

2.1.2.1 Distância euclidiana

A distância euclidiana possui popularidade como medida de dissimilaridade, sendo a mais utilizada em algoritmos de agrupamento. A justificativa para tal fato é a sua adequação para o uso em dados contínuos, oferendo boa interpretabilidade em situações com número reduzido de variáveis. O Algoritmo *K-means* é citado por Yang e Sinaga (2019) como o mais antigo e amplamente conhecido e o mesmo utiliza o cálculo da distância euclidiana para analisar o espaço entre os pontos e centroides, reforçando a relevância de tal distância no estudo de algoritmos de agrupamento.

A distância consiste em calcular a separação entre dois pontos considerando a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas. No $\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \mathbb{R},\ i = 1, 2, \ldots, n\}$. , podemos considerar dois vetores $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ e teremos a distância euclidiana definida como:

$$d_{euclidiana}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.1}$$

A simplicidade e interpretação geométrica intuitiva contribuem para sua ampla aplicabilidade em técnicas de *clustering*, porém, Han, Pei e Tong (2022) destacam que em alta dimensionalidade pode ocorrer o fenômeno conhecido como "maldição da dimensionalidade"no qual as distâncias tendem a se tornar menos discriminativas. Ou seja, a aplicação possui limitações, sendo seu uso em contextos de dados heterogêneos ou com ruído elevado podendo exigir o uso de técnicas de ponderação ou métodos alternativos para corrigir distorções.

2.1.2.2 Distância Manhattan

A distância Manhattan, demonstrada por Aggarwal, Hinneburg e Keim (2001) como um exemplo da utilização de normas de menor ordem, é definida como a soma das diferenças entre as coordenadas correspondentes entre dois pontos:

$$d_{manhattan}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2}$$

Essa distância apresenta vantagens devido a sua simplicidade computacional e aplicabilidade em dados com alta dimensionalidade. Segundo Aggarwal, Hinneburg e Keim (2001), a distância de Manhattan contrasta com a Euclidiana por conseguir preservar melhor a diferença entre as distâncias presentes em espaços de alta dimensão. Ou seja, quanto maior o número de dimensões, as diferenças calculadas pelas distâncias pode variar, sendo a distância Manhattan possuindo melhor separabilidade entre objetos, resultando em um melhor desempenho de algoritmos de agrupamento com dados multivariados.

2.1.2.3 Distância de Chebyshev

A distância Chebyshev ou distâcia máxima, está relacionada as normas das distâncias Euclidiana e Manhattan, possuindo a característica individual de enfatizar a dimensão na qual a discrepância entre dois vetores é maior, sendo esta o seu diferencial, segundo Aggarwal, Hinneburg e Keim (2001). Sua formulação mede a maior diferença absoluta entre coordenadas de dois pontos, matematicamente:

$$d_{\text{Chebyshev}}(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i| \tag{2.3}$$

De acordo com Boriah, Chandola e Kumar (2008), a utilidade dessa distância reside no fato de que existem situações nas quais a similaridade entre objetos deve ser determinada pelo

pior caso de diferença entre variáveis. Sendo assim, essa métrica pode ser aplicada para controle, tolerância e detecção de anomalias, por exemplo.

Entretanto, Aggarwal, Hinneburg e Keim (2001), consideram as limitações da distância de Chebyshev nos espaços de alta dimensão, visto que as diferenças possuem menos contraste ao decorrer do aumento da dimensionalidade, causando a redução da eficácia da sua aplicação em métodos baseados em proximidade como o *K-means* e *K-NN*.

2.1.2.4 Distância Cosseno

A distância cosseno, considera o ângulo entre dois vetores, com isso, sua formulação é a mais distinta em relação às demais apresentadas anteriormente, sendo a seguinte:

$$d_{cosseno}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|} \tag{2.4}$$

Na fórmula apresentada acima, os vetores $\mathbf{x}$ e $\mathbf{y}$ representam objetos em um espaço n-dimensional, sendo individualmente representados por um conjunto de características. Tais vetores são pontos em $\mathbb{R}^n$ com coordenadas relacionadas aos atributos individuais de cada elemento presente na comparação. Segundo Huang (2008), essa distância tem eficácia em comparar padrões de correlação, sendo assim, a distância cosseno é utilizada em mineração de texto e aplicações com recomendações, por exemplo.

Ademais, essa distância tem como característica, a eficiência no uso da similaridade dos dados sobre vetores esparsos. Consequentemente, tornando-se uma escolha padrão em muitas implementações sobre sugestões ou recuperação de informação [Sarwar et al. (2001), Huang (2008)]. Em resumo, a distância cosseno diferencia-se pelo uso da orientação de vetores para comparar a similaridade dos dados, sendo eficaz em contextos de alta dimensionalidade e dados esparsos.

2.1.3 Métodos de agrupamento

Nessa seção serão apresentado os seguintes algoritmos de agrupamento multi-visão que serão utilizados nesse estudo: MRDCA e MRDCA-RWL, sendo o segundo uma versão ponderada do primeiro; *Multi-View Kernel K-Means (MVKKM)* e o *Two-Level Variable Weighting Clustering (TW-KM)* , constituidos por versões ponderadas do algoritmo *K-Means*; *Multi-View Spectral Clustering (MVSC)*, uma extensão do *Spectral Clustering* tradicional.

2.1.3.1 MRDCA

O algoritmo de agrupamento dinâmico baseado em múltiplas matrizes de dissimilaridade (MRDCA) foi desenvolvido por Carvalho, Lechevallier e Melo (2012), sua abordagem considera um particionamento *hard*, sendo cada objeto atribuído a um único *cluster*, implicando em uma classificação mais objetiva. O método utiliza múltiplas matrizes de dissimilaridade D_j, nas quais

suas entradas $d_j(e_i, e_n)$ representam a diferença entre os objetos e_i e e_n, como representado na seguinte matriz:

$$D_1 = \begin{bmatrix} d_1(e_1, e_1) & \cdots & d_1(e_1, e_n) \\ \vdots & \ddots & \vdots \\ d_1(e_n, e_1) & \cdots & d_1(e_n, e_n) \end{bmatrix} \quad \cdots \quad D_p = \begin{bmatrix} d_p(e_1, e_1) & \cdots & d_p(e_1, e_n) \\ \vdots & \ddots & \vdots \\ d_p(e_n, e_1) & \cdots & d_p(e_n, e_n) \end{bmatrix}$$

Seja $E = \{e_1, e_2, \ldots, e_n\}$ o conjunto formado pelos n objetos a serem agrupados. O algoritmo busca encontrar uma partição $P = (C_1, \ldots, C_K)$ desse conjunto em K agrupamentos, em que cada C_k representa o conjunto de objetos pertencentes ao k-ésimo *cluster*. Cada agrupamento é representado por um protótipo G_k, que consiste em um subconjunto de objetos pertencentes a E com cardinalidade fixa q. O conjunto de todos os subconjuntos de E contendo exatamente q elementos é denotado por $\mathcal{E}^{(q)}$, isto é,

$$\mathcal{E}^{(q)} = \{A \subseteq E : \ |A| = q\}.$$

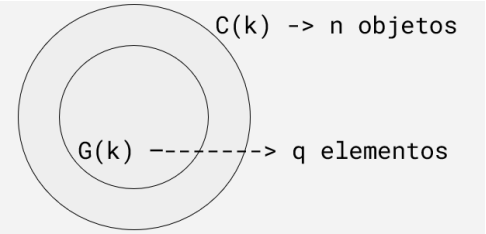
Assim, cada protótipo satisfaz $G_k \in \mathcal{E}^{(q)}$.

Essas matrizes são utilizadas para análise das distâncias entre os dados e os protótipos G_k. O protótipo é obtido pela minimização da dissimilaridade intra-grupo, sendo dado por

$$G_k = \arg \min_{G \in \mathcal{E}^{(q)}} \sum_{e_i \in C_k} \sum_{j=1}^{p} \sum_{e \in G} d_j(e_i, e).$$

O protótipo possui cardinalidade fixa dentro do conjunto total de n elementos. A Figura 4 ilustra a relação entre o conjunto de objetos E, os agrupamentos C_k e seus respectivos protótipos G_k.

Figura 4 – Representação da prototipagem



Fonte : o autor

Para simplificar a notação, define-se

$$D_j(e_i, G_k) = \sum_{e \in G_k} d_j(e_i, e),$$

que representa a dissimilaridade entre o objeto e_i e o protótipo G_k considerando apenas a j-ésima matriz de dissimilaridade.

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

$$J = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} D_j(e_i, G_k) = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} \sum_{e \in G_k} d_j(e_i, e) \tag{2.5}$$

Fonte: Carvalho, Lechevallier e Melo (2012)

A função J representa o critério de adequação do agrupamento, sendo responsável por avaliar a qualidade da partição dos dados em K agrupamentos em relação aos objetos i pertencentes a cada agrupamento k, considerando ainda o conjunto de matrizes de dissimilaridade j, com j variando de 1 até p. Essa função quantifica a homogeneidade interna dos agrupamentos ao medir o quão bem os objetos estão associados aos seus respectivos representantes. Na primeira parte da equação, é estabelecida a relação entre os elementos e_i do *cluster* e o seu representante G_k. Já na segunda parte, a igualdade evidencia a equivalência dessa avaliação ao considerar cada elemento e pertencente ao representante G_k, mensurando a distância entre o objeto do agrupamento e os objetos que compõem o representante. Dessa forma, o objetivo do algoritmo consiste em minimizar a função J, de modo a obter agrupamentos mais homogêneos e melhor ajustados aos seus representantes.

2.1.3.2 MRDCA-RWL

MRDCA-RWL é uma sigla em inglês para se referir a: algoritmo de agrupamento rígido dinâmico com peso de relevância para cada matriz de dissimilaridade estimada localmente. Seu objetivo principal é encontrar uma partição $P = (C_1, C_2, \ldots, C_K)$ 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, Lechevallier e Melo (2012). Considerando diferentes pesos para as matrizes, temos o seguinte critério de adequação:

$$J = \sum_{k=1}^{K} \sum_{e_i \in C_k} D_{\lambda_k}(e_i, G_k) = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{P} \lambda_{kj} \sum_{e \in G_k} d_j(e_i, e) \tag{2.6}$$

Fonte: Carvalho, Lechevallier e Melo (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.

Algoritmo 1 MRDCA-RWL – Algoritmo de Agrupamento Dinâmico com Peso de Relevância Local para cada Matriz de Dissimilaridade

---

Entrada: Conjunto de objetos $E = \{e_1, e_2, \ldots, e_n\}$; número de clusters K; número de matrizes de dissimilaridade p; cardinalidade q dos protótipos.

Saída: Partição $P = (C_1, C_2, \ldots, C_K)$, protótipos $G_1, G_2, \ldots, G_K$ e vetores de pesos de relevância $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kp})$.

1. Inicialização:

Fixar o número de clusters K e a cardinalidade q dos protótipos G_k.

Definir $t = 0$.

Inicializar os vetores de pesos $\boldsymbol{\lambda}_k^{(0)} = (1, 1, \ldots, 1)$ para todo $k = 1, \ldots, K$.

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

Atribuir cada objeto e_i ao cluster cujo protótipo minimiza $\sum_{j=1}^{p} \lambda_{kj}^{(0)} D_j(e_i, G_k^{(0)})$, formando a partição inicial $P^{(0)} = (C_1^{(0)}, \ldots, C_K^{(0)})$.

2. Passo 1: Cálculo dos melhores protótipos:

Incrementar $t \leftarrow t + 1$.

Fixar a partição $P^{(t-1)} = (C_1^{(t-1)}, \ldots, C_K^{(t-1)})$ e os pesos $\boldsymbol{\lambda}_k^{(t-1)}$.

Calcular, para cada cluster k, o novo protótipo:

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

3. Passo 2: Atualização dos pesos de relevância locais:

Fixar a partição $P^{(t-1)}$ e os protótipos $G_k^{(t)}$.

Atualizar, para cada cluster $k = 1, \ldots, K$ e cada matriz $j = 1, \ldots, p$, o peso:

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

onde $D_j(e_i, G_k^{(t)}) = \sum_{e \in G_k^{(t)}} d_j(e_i, e)$.

4. Passo 3: Definição da melhor partição:

Fixar os protótipos $G_k^{(t)}$ e os pesos $\boldsymbol{\lambda}_k^{(t)}$.

22: Para cada objeto e_i, encontrar o cluster $C_k^{(t)}$ tal que:

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

23: Atualizar a partição conforme o novo agrupamento dos objetos.

24:

25: **5. Critério de parada:**

26: Se não houver mudança na atribuição dos objetos ($test = 0$), então **parar**.

27: Caso contrário, retornar ao Passo 1.

Fonte: Traduzido de Carvalho, Lechevallier e Melo (2012)

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, Lechevallier e Melo (2012) foram estimados os valores de λ_{kj} com base no método de multiplicadores de Lagrange. Como segue:

$$\lambda_{kj} = \frac{\left\{\prod_{h=1}^{p} \left[\sum_{e_i \in C_k} D_h(e_i, G_k)\right]\right\}^{1/p}}{\sum_{e_i \in C_k} D_j(e_i, G_k)} = \frac{\left\{\prod_{h=1}^{p} \left[\sum_{e_i \in C_k} \sum_{e \in G_k} d_h(e_i, e)\right]\right\}^{1/p}}{\sum_{e_i \in C_k} \sum_{e \in G_k} d_j(e_i, e)} \tag{2.7}$$

Fonte: Carvalho, Lechevallier e Melo (2012)

2.1.3.3 Multi-View Kernel K-Means (MVKKM)

O método Kernel K-Means utiliza do clássico algoritmo K-Means com a ampliação de mapeamento kernel. Com isso, há a possibilidade de aplicação de dados não lineares. A sua função objetivo é dada por:

$$\min_{H \in^{n \times K}, H^\top H = I} \operatorname{Tr}\big(K(I - HH^\top)\big), \tag{2.8}$$

Fonte:Liu et al. (2022)

Na qual K é representa a matriz kernel de dados. Porém, quando o método evolui para um cenário *multi-view*, a formulação matemática considera V vistas e uma matriz de kernel $K^{(v)} \in^{n\times n}$. Com isso, temos n amostras e K clusters em uma matriz $H \in^{n\times K}$ com $H^\top H = I$, ademais, temos os pesos $\gamma = (\gamma_1, \ldots, \gamma_V)$ para combinar os kernels, considerando $\sum_{v=1}^{V} \gamma_v = 1$ e $\gamma_v \geq 0$.

Algoritmo 2 MVKKM – Multi-View Kernel K-Means

Entrada: Matrizes kernel $K^{(v)}$ para $v = 1, \ldots, V$; número de clusters K.

Saída: Matriz de atribuição de clusters $H \in \mathbb{R}^{n\times K}$ e pesos das vistas $\gamma = (\gamma_1, \ldots, \gamma_V)$.

1. Inicialização:

Inicializar os pesos das vistas $\gamma_v^{(0)} \geq 0$ tal que $\sum_{v=1}^{V} \gamma_v^{(0)} = 1$.

Combinar kernels: $K_\gamma^{(0)} = \sum_{v=1}^{V} \gamma_v^{(0)} K^{(v)}$.

Inicializar a matriz de atribuição $H^{(0)} \in \mathbb{R}^{n\times K}$ (ex.: aleatória ou via k-means simples no kernel combinado).

Definir $t = 0$.

2. Iteração:

repeat

Incrementar $t \leftarrow t + 1$.

2.1 Atualização da matriz de atribuição H:

Fixar $K_\gamma^{(t-1)}$.

Executar kernel k-means: atualizar $H^{(t)}$ de modo a minimizar

$$\mathrm{Tr}\big(K_\gamma^{(t-1)}(I - HH^\top)\big)$$

2.2 Atualização dos pesos das vistas γ:

Fixar $H^{(t)}$.

Atualizar γ resolvendo

$$\min_{\gamma\in\Delta} \mathrm{Tr}\Big(\big(\sum_{v=1}^{V} \gamma_v K^{(v)}\big)(I - H^{(t)}H^{(t)\top})\Big), \quad \Delta = \{\gamma_v \geq 0, \sum_{v=1}^{V} \gamma_v = 1\}.$$

Atualizar kernel combinado: $K_\gamma^{(t)} = \sum_{v=1}^{V} \gamma_v^{(t)} K^{(v)}$.

2.3 Critério de convergência:

Verificar se $H^{(t)}$ ou $\gamma^{(t)}$ convergiram (mudança abaixo de um limiar). até convergência.

25: **3. Saída:**
26: Retornar $H^{(t)}$ e $\gamma^{(t)}$.

Fonte: Cai, Nie e Huang (2013)

As vantagens na utilização do método MVKKM englobam a utilização da complementaridade entre as vistas, além do seu aprendizado de pesos de vistas considerando os que possuem mais contribuição. Em contrapartida, o método possui desafios, sendo eles a escolha adequada de matrizes para cada vista, a realização de otimização conjunta de H e γ, além da possibilidade de alta complexidade para interpretação dos pesos de cada vista.

Segundo os experimentos realizados por Cai, Nie e Huang (2013), quando consideradas bases multi-vistas, o algoritmo apresentou um melhor desempenho se comparado a outros similares, sendo o maior desempenho observado em bases com maior quantidade de vistas. Ainda no mesmo trabalho ressalta-se que o uso do aprendizado adaptativo de pesos melhora substancialmente o agrupamento, colaborando para *clusters* mais coerentes.

2.1.3.4 Two-Level Variable Weighthing Clustering (TW-KM)

Assim como o algoritmo apresentado anteriormente, o Two-Level Variable Weighthing Clustering (TW-KM) é baseado no algoritmo K-Means, sendo a sua diferença baseada no uso de pesos para variáveis/atributos e visões. O TW-KM tem o objetivo de atribuir e aprender, iterativamente, pesos a cada vista, incorporando esses pesos na distância usada na sua função objetivo . Dado o exposto, o algoritmo possui a seguinte função objetivo:

$$J(U, C, W,) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \sum_{v=1}^{V} \beta_v \sum_{j \in G_v} w_{j|v}\, d\left(x_{ij}, c_{kj}\right)$$

Fonte: Chen et al. (2013)

Na qual considera-se $X = \{x_i\}_{i=1}^{n}$, $x_i \in \mathbb{R}^P$ como um conjunto de dados particionado em V vistas. Ademais, G_v é o conjunto de índices de atributos da vista v ($v = 1, \ldots, V$) e K é o número de clusters procurados. Com isso, há a consideração da variável de particionamento $u_{ik} \in \{0, 1\}$ sendo $\sum_{k=1}^{K} u_{ik} = 1$ para todo i. Por fim, para cada cluster k temos um centróide $c_k \in \mathbb{R}^P$. O resultado dessa formulação é a minimização da função objetivo, que constitui o objetivo do algoritmo. Além disso, a formulação possui os seguintes pontos:

- $d(x_{ij}, c_{kj})$ é geralmente o quadrado da diferença (distância euclidiana por componente): $(x_{ij} - c_{kj})^2$;
- $w_{j|v} \geq 0$ é o peso da variável j dentro da vista v;
- $\beta_v \geq 0$ é o peso (importância) da vista v.

Algoritmo 3 TW-KM – Two-Level Variable Weighting K-Means

Entrada: Dados $X = \{x_i\}_{i=1}^{n}$, número de clusters K, partição dos atributos em *views* $\{G_v\}_{v=1}^{V}$, critério de parada (tolerância ϵ ou número máximo de iterações T).

Saída: Partição $U = \{u_{ik}\}$, centróides $\{c_k\}$, pesos das variáveis $\{w_{j|v}\}$ e pesos das *views* $\{\beta_v\}$.

1. Inicialização:

Inicializar os centróides $\{c_k\}$ (por exemplo, com o método k-means++ ou de forma aleatória).

Inicializar os pesos das *views* $\beta_v \leftarrow 1/V$ e, para cada *view* v, inicializar os pesos das variáveis $w_{j|v} \leftarrow 1/|G_v|$.

2. Iteração:

for $t = 1$ até T **do**

Atualização das atribuições u_{ik}:

$$u_{ik} \leftarrow \begin{cases} 1, & \text{se } k = \arg\min_h \sum_{v=1}^{V} \beta_v \sum_{j \in G_v} w_{j|v}\, d(x_{ij}, c_{hj}) \\ 0, & \text{caso contrário.} \end{cases}$$

Atualização dos centróides c_k (para cada dimensão j):

$$c_{kj} \leftarrow \frac{\sum_{i=1}^{n} u_{ik}\, x_{ij}}{\sum_{i=1}^{n} u_{ik}} \quad \text{(considerando a métrica Euclidiana por componente).}$$

Atualização dos pesos das variáveis dentro de cada *view* v:

$$w_{j|v} \propto \exp\Big(- \frac{S_{j|v}}{\gamma} \Big)$$

onde $S_{j|v}$ representa a contribuição da variável j (na *view* v) para o custo intra-cluster.

Atualização dos pesos das *views* β_v:

$$\beta_v \propto \exp\Big(- \frac{T_v}{\eta} \Big)$$

onde T_v representa a contribuição global da *view* v.

Normalização dos pesos:

$$\sum_j w_{j|v} = 1, \quad \sum_v \beta_v = 1.$$

Verificação do critério de convergência:

Encerrar se a variação de J for menor que ϵ ou se houver pequena mudança nas atribuições.

end for

21: **3. Saída:**
22: Retornar $U, \{c_k\}, \{w_{j|v}\}, \{\beta_v\}$. =0

Fonte: Chen et al. (2013)

Segundo Chen et al. (2013) , o algoritmo possui pontos positivos e negativos, sendo os positivos a capacidade de utilizar dados multi-visão atuando com aprendizado de pesos (o algoritmo identifica automaticamente quais vistas e atributos são mais relevantes para cada *clusters*) e a integração direta com o K-means. Em contraste, os pontos negativos estão relacionados com sensibilidade à inicialização e aos hiperparâmetros , complexidade computacional maior que k-means e problemas na interpretação dos pesos em cenários específicos de sobreposição ou forte correlação dos dados.

2.1.3.5 Multi-View Spectral Clustering (MVSC)

O Multi-View Spectral Clustering (MVSC) pertence ao conjunto de algoritmos *spectral clustering*, sendo um método capaz de transformar o desafio de agrupar dados em um problema de autovetores de uma matriz de similaridade. Nesse algoritmo, há a busca de vetores próprios para cada vista, prezando pela otimização local da estrutura. A partir disso, cada vista possui uma matriz de similaridade e uma matriz de autovetores que está relacionada com a estrutura de grupos da vista, acontece uma regularização dos pontos e atualizações por vista, até que o algoritmo consiga convergir em uma solução na qual as vistas sejam consistentes entre si. Considerando o exposto, temos a seguinte formulação matemática:

$$\max_{U^{(1)},\ldots,U^{(V)}} \quad \sum_{v=1}^{V} \operatorname{tr}\left(U^{(v)T} L^{(v)} U^{(v)}\right) \; + \; \lambda \sum_{v<w} \operatorname{tr}\left(U^{(v)} U^{(v)T} \; U^{(w)} U^{(w)T}\right)$$
$$\text{s.t.} \quad U^{(v)T} U^{(v)} = I \quad \text{para todo } v = 1, \ldots, V.$$

Fonte: Kumar, Rai e III (2011)

Sendo V vistas, n amostras e k *clusters*. Para cada vista v é definida a matriz de similaridade $K^{(v)} \in \mathbb{R}^{n \times n}$ e o Laplaciano normalizado $L^{(v)}$. Considerando $U^{(v)} \in \mathbb{R}^{n \times k}$ a matriz cujas colunas são os k autovetores para a view v. Além disso, $\operatorname{tr}(U^{(v)T} L^{(v)} U^{(v)})$ é o termo de spectral clustering para a view v, no qual o maior valor significa melhor separação/compactação na vista. Em relação ao segundo termo da expressão, os produtos $U^{(v)} U^{(v)T}$ e $U^{(w)} U^{(w)T}$ quando semelhantes, representam a alta correlação entre as vistas. Já em relação ao $\lambda > 0$, é um parâmetro utilizado para regular o peso entre vistas.

Algoritmo 4 Multi-view Spectral Clustering

1: **Entrada:** Matrizes de similaridade $\{K^{(v)}\}_{v=1}^{V}$, número de clusters k, parâmetro $\lambda > 0$, tolerância ε ou número máximo de iterações T.

Saída: Rótulos de cluster para as n amostras.

1. Construção dos Laplacianos:
Para cada *view* $v = 1, \ldots, V$, construir o Laplaciano normalizado $L^{(v)}$ a partir de $K^{(v)}$.

2. Inicialização:
Inicializar $U^{(v)}$ utilizando os k autovetores principais de $L^{(v)}$.

3. Iteração:
for $t = 1$ até T **do**
    for cada view $v = 1, \ldots, V$ **do**
        Construir a matriz modificada:

$$M^{(v)} \leftarrow L^{(v)} + \lambda \sum_{w \neq v} U^{(w)} U^{(w)T}$$

        Atualizar $U^{(v)}$ escolhendo os k autovetores associados aos maiores (ou menores, considerando a convenção) autovalores de $M^{(v)}$.
    end for
    Verificação de convergência:
    Encerrar se a variação das matrizes $U^{(v)}$ for menor que ε.
end for

4. Construção da representação consenso:
Formar a matriz consenso para o clustering, por exemplo:

$$U^* = \frac{1}{V} \sum_{v=1}^{V} U^{(v)}$$

(alternativamente, pode-se concatenar os $U^{(v)}$).

5. Clusterização final:
Aplicar *k-means* às linhas da representação final (por exemplo, U^*) para obter os rótulos das amostras.

Fonte: adaptado de Kumar, Rai e III (2011)

Segundo Kumar, Rai e III (2011) , o método possui limitações, como a convergência em mínimos locais resultando na possibilidade de soluções subótimas, porém, considera como boa prática a normalização de similaridades por vista para que vistas com escalas diferentes não dominem. Entretanto, o algoritmo é capaz de aproveitar a complementaridade entre vistas e possui uma interpretação geométrica relacionada a estrutura dos *clustering*, justificando o uso do algoritmo.

2.2 Síntese do capítulo

Neste capítulo foram apresentados os conceitos de agrupamento de dados com múltiplas vistas, assim como os algoritmos que podem ser utilizados, com a combinação das métricas de distância entre os dados. Tais conceitos foram apresentados pois são a base teórica necessária para o entendimento deste trabalho como um todo, já que os experimentos desenvolvidos utilizam uma combinação dos tópicos apresentados neste capítulo.

3
Revisão Sistemática

Neste capítulo será apresentada a revisão sistemática, sendo o seu conteúdo dividido entre objetivos,questões de pesquisa, estratégias para obter as respostas com a utilização de strings de busca, assim como a utilização de bases de dados. Em decorrência disto, serão expostos os critérios de inclusão e exclusão, bem como os resultados adquiridos a partir da discussão das respostas aos questionamentos levantados.

3.1 Objetivo e questões de pesquisa

Diante dos avanços no campo de análise de agrupamento de dados, esta revisão sistemática tem como objetivo avaliar estudos relacionados a agrupamento de dados com múltiplas visões com pesos, usando como base a exploração de fontes de pesquisa relevantes. A avaliação tem como proposta analisar as tendências atuais relacionadas ao tema, bem como os principais avanços da área. Para tal, foram elaboradas as seguintes questões de pesquisa:

1. Quais são as principais abordagens utilizadas em relação ao agrupamento de múltiplas visões com pesos?
2. Quais as áreas de pesquisa mais comuns de aplicação do agrupamento de múltiplas visões com pesos?
3. Quais são os principais desafios abordados ao trabalhar com agrupamento de dados com múltiplas visões com pesos?

3.1.1 Estratégia de busca

A princípio foram estipuladas palavras-chave, apresentadas na tabela 1, relacionadas com o objeto de pesquisa, com o objetivo da construção da *string* de busca, apresentada na tabela

2, utilizada nas três bases de pesquisa selecionadas: IEEE Xplore, Science Direct e Scopus. A filtragem de estudos foi feita mediante os seguintes critérios de inclusão e exclusão estipulados:

Critérios de inclusão:

1. Trabalhos publicados entre 2014 e 2024;
2. Artigos em inglês;
3. Artigos em português;
4. Trabalhos com múltipla visão ponderada;

Critérios de exclusão:

1. Apresentações em formato de *slides*;
2. Artigos duplicados;
3. Artigos sem acesso completo;
4. Capítulos de livros;
5. Trabalhos de revisão sistemática ou mapeamento sistemático;

Tabela 1 – Palavras-chave utilizadas na *String* de busca

Palavra-chave	Termos relacionados em inglês
Múltiplas visões	multi view, multi-view e multiview
Agrupamento	clustering
Peso	Weight,Weighted,Weighting

Tabela 2 – *String* de busca

(("multi view") OR ("multi-view") OR ("multiview")) AND (clustering) AND ("Weight"OR "Weighted"OR "Weighting")

3.1.2 Extração de dados

Após a análise dos artigos, foram considerados alguns aspectos para responder as questões de pesquisa, sendo eles: se os dados estão incompletos ou não, se o método proposto no trabalho utilizava a abordagem semissupervisionada, qual o seu domínio de aplicação e dados sobre a função objetivo. Foram analisadas as métricas de distância e métricas de validação. Nos artigos que relatam experimentos, foram consideradas a quantidade de repetições usadas para execução

dos algoritmos, as bases de dados, quantidade de métodos da literatura usados para comparação, métricas de validação dos agrupamentos.
Quanto aos agrupamentos, a análise concentrou-se nos tipos *Hard* ou *Fuzzy* e mono-objetivo ou multiobjetivo. Adicionalmente, foi feita uma síntese sobre os desafios e comentários gerais relatados nos artigos.

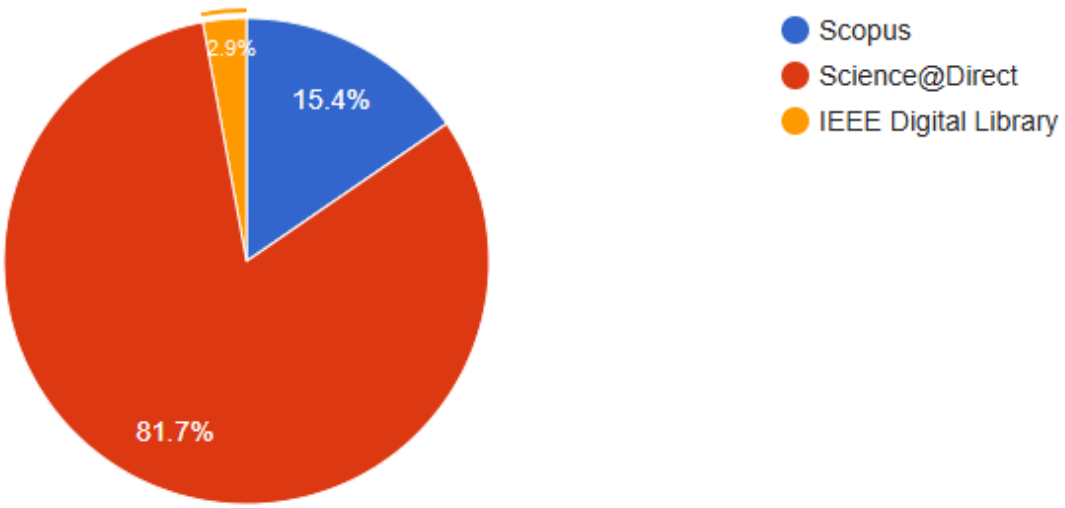
3.1.3 Seleção de estudos

A seleção foi realizada em três partes, sendo a primeira a aplicação da *string* de busca nas bases de pesquisa, totalizando 726 artigos encontrados entre os anos de 2014 a 2024.

A segunda etapa ocorreu com a análise dos artigos, incluindo aqueles que obedeciam os critérios de inclusão e excluindo os duplicados, sem acesso completo, capítulos de livros e trabalhos de revisão sistemática ou mapeamento, resultando em 150 trabalhos. Os artigos selecionados na segunda etapa foram analisados na terceira etapa para fornecerem o embasamento das respostas para as questões de pesquisa.

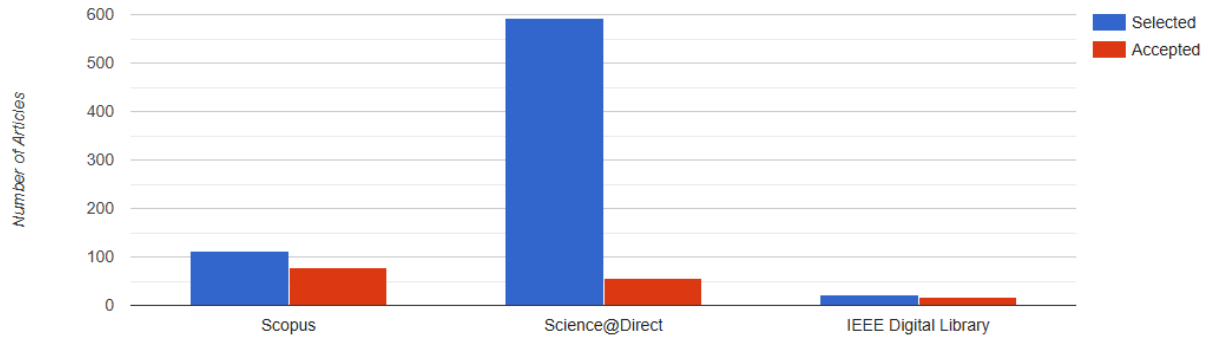
As Figuras 1 e 2 ilustram gráficos com a composição da primeira etapa e o aceite ou recusa de artigos, respectivamente.

Figura 5 – Composição da primeira etapa de seleção por base eletrônica.



Fonte : O autor

Figura 6 – Aceite ou recusa seguindo os critérios de exclusão por base eletrônica.



Fonte : O autor

3.2 Resultados e Discussão

Esta seção tem como objetivo apresentar os resultados obtidos através da análise de 150 artigos e com isso, responder as questões de pesquisa.

Os estudos explorados buscam melhorar a precisão e a eficiência do agrupamento usando as informações complementares fornecidas pelos dados, visto que o uso do agrupamento de dados com múltiplas visões tem ganhado visibilidade devido a capacidade de integrar informações de diversas visões. (LIU et al., 2024).

Avaliando o conjunto completo dos estudos, há uma maior frequência de abordagem *Fuzzy*, algoritmos não semi-supervisionado, presença predominante de funções multiobjetivo, métricas de distância e validação variadas, com o uso de diferentes domínios de aplicação. As questões de pesquisa e seus respectivos argumentos estão expostos a seguir.

3.2.1 Quais são as principais abordagens utilizadas em relação agrupamento de dados com múltiplas visões com pesos?

As abordagens utilizadas para determinar pesos com a utilização de múltiplas visões possuem uma ampla variação. A exploração de kernel múltiplo está presente em 27 artigos e segundo Zheng et al. (2020), o método combina ponderação com informações sobre os dados e se baseia em otimização considerando a aprendizagem de pesos conjuntamente com outros parâmetros do modelo.

Já a utilização de matrizes com múltiplas visões como uma ferramenta de representação matemática de dados de múltiplas fontes está presente em 52 publicações analisadas, sendo considerada comumente utilizada por Jiang, Zhou e Zhao (2024).

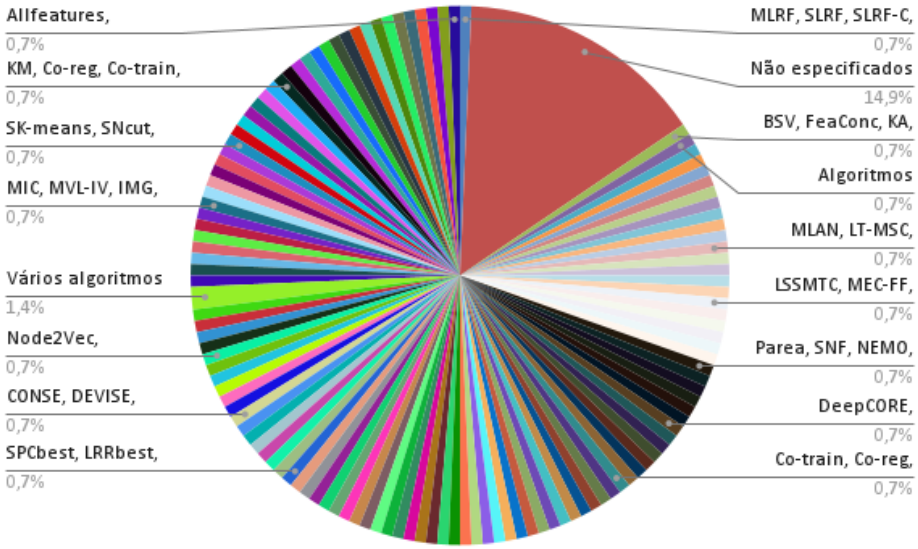
Em quantidade aproximada ao método anterior, 40 estudos relatam o uso de algoritmos de agrupamento *espectral* relacionados com agrupamentos complexos e dados não lineares. Além disso, métodos de ponderação adaptativa atribuindo pesos baseados em métricas ou ponderação manual estão presentes em 33 artigos.

Quanto aos algoritmos utilizados, 26 trabalhos citam o *K-means* embora seja adequado

para métodos de agrupamento de única visão e Liu et al. (2024) afirmam que a estratégia apresenta limitações práticas por não considerar as especificidades das diferentes visões.

Demais algoritmos são representados em menor quantidade, apresentando uma abordagem comparativa nos estudos analisados. A Figura 7 ilustra a frequência e comparações relatadas.

Figura 7 – Composição das comparações de algoritmos.



Fonte : O autor

3.2.2 Quais as áreas de pesquisa mais comuns de aplicação do agrupamento de dados com múltiplas visões com pesos?

Agrupamento de dados com múltiplas visões possui uma variedade de campos de aplicação. Sendo aprendizagem de máquina e mineração de dados as mais comuns, com 47 trabalhos associados e tendo sua importância afirmada por Qian et al. (2018) que em seu trabalho relata agrupamento de dados com múltiplas visões como um tema atual e de destaque em aprendizado de máquina.

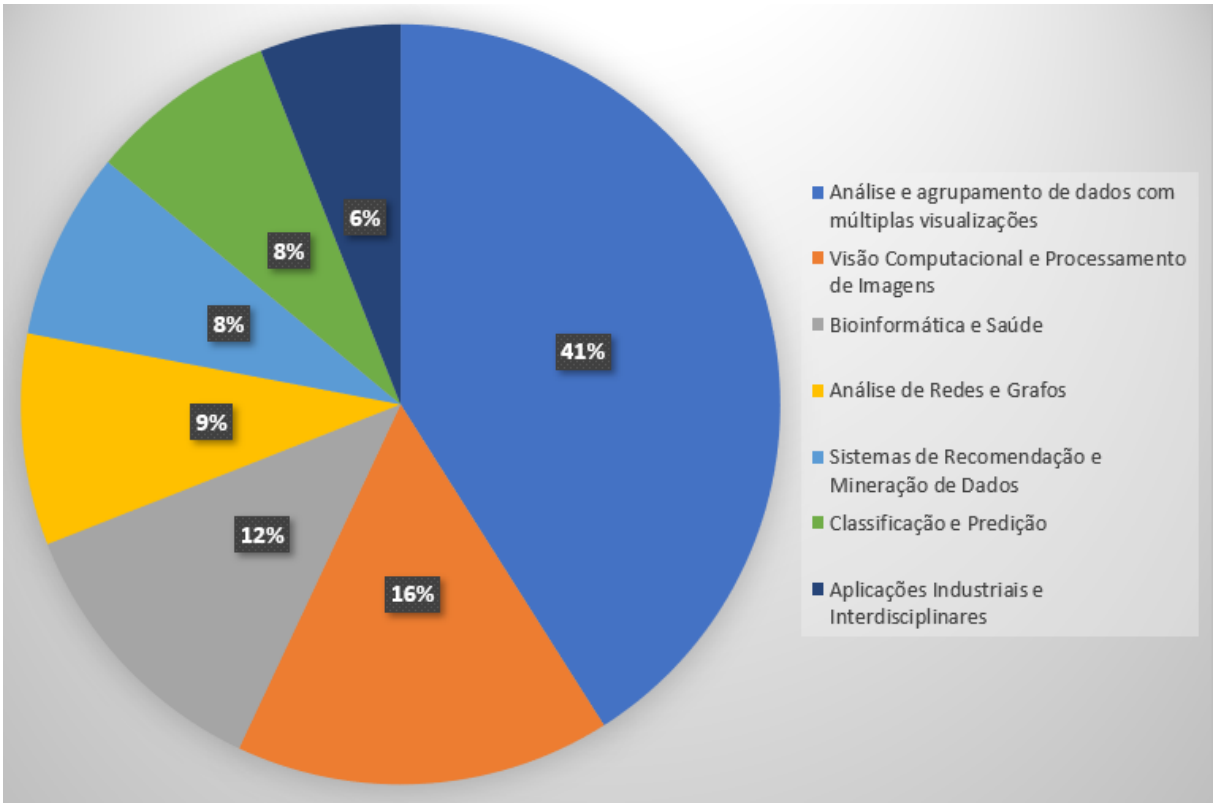
Adicionalmente, visão computacional e processamento de imagens, com 39 estudos relacionados é a segunda área com mais aplicações, com a utilização de reconhecimento facial, categorização de cenas, segmentação de imagem, reconhecimento de objetos, detecção de tumores em estudos relacionados com a saúde e alvos em movimento nos estudos associados a segurança e monitoramento.

Segundo Zheng et al. (2020), as informações de várias visões podem atuar de forma suplementar. No escopo de reconhecimento biométrico, rostos, impressões digitais, palmares e em relação a íris, formando dados de visualização múltipla. Adicionalmente, no campo da medicina, diferentes exames configuram visões diferentes, colaborando para a precisão do diagnóstico de

pacientes.

Seguindo essa perspectiva, 15 artigos relatam sobre aspectos de saúde e medicina, utilizando o agrupamento de dados com múltiplas visões para processamento de informações biomédicas, detecção de pacientes com características semelhantes e para analise de imagens médicas, como foi aplicado por Hua et al. (2021) em seu estudo sobre segmentação de imagens cerebrais utilizando algoritmo de agrupamento de dados com múltiplas visões com ponderação de características, as quais possuem pesos adaptáveis com o intuito de melhorar o desempenho do algoritmo utilizado. De forma geral, as principais áreas são apresentadas na Figura 8:

Figura 8 – Áreas de pesquisa.



Fonte : O autor

Outras áreas como: análise de redes relacionadas aos dados de comércio, agrupamento de documentos com recuperação de informação e geociências aparecem em estudos isolados. Como exemplo, o estudo de Liu et al. (2024) utiliza dados de múltiplos sensores para identificar funções urbanas, enquanto Huang et al. (2020) apontam para o uso de agrupamentos de dados com múltiplas visões para sistema de recomendação e práticas comerciais.

3.2.3 Quais são os principais desafios abordados ao trabalhar com agrupamento de dados com múltiplas visões com peso?

Uma série de desafios são apresentados ao decorrer dos 150 artigos analisados. O agrupamento de dados com múltiplas visões com peso possui a dificuldade em garantir eficácia

aos algoritmos, tendo como aspectos os entraves relacionados a determinação de pesos das visões, integração de informações diferentes, escalabilidade, relação com ruído e avaliação de desempenho.

Quanto a isso, um desafio fundamental é a determinação dos pesos das visões, segundo Imani et al. (2024), algumas visões atuam de forma complementar, mas em contrapartida, outras podem ser redundantes ou irrelevantes. Com isso, o artigo conclui que a concatenação dessas diferentes visões podem resultar em dificuldades adicionais na seleção de características.

Outra preocupação é em relação a escalabilidade e eficiência computacional quando se trata de grandes conjuntos de dados com alta dimensionalidade. Técnicas de redução de custo computacional podem causar perdas de informações ou distorção de dados, Sharma, Nayak e Bhaskar (2024) fazem uma reflexão sobre os altos custos de ajustes de parâmetros quando o tamanho do conjunto de dados ou quantidade de visões aumenta.

Por fim, a interpretação dos resultados e a escolha de métricas de avaliação do desempenho apresentam uma dificuldade de escolha ideal, visto que, devem adequar-se às particularidades dos dados. Como cita Qian et al. (2018) em seu trabalho com uma métrica comumente utilizada, porém, com a desvantagem de que seu menor valor não indica necessariamente o melhor resultado de agrupamento.

3.3 Considerações finais da revisão sistemática

Após a conclusão da análise de 150 artigos, foi possível responder as questões de pesquisa observando as tendências e estratégias utilizadas. A especificação do objeto de estudo resultou em particularidades, tendo sua aplicação em temáticas diversas.

Quanto as tendências, a maioria dos artigos estão concentrados entre os anos de 2021 e 2024, sendo o último o que mais possui publicações, contando com 35 estudos. Além disso, o uso de métricas de distância é variado, sendo a distância euclidiana a mais comum nos estudos comparativos.

Devido as diferentes áreas de aplicação, o uso de dados de múltiplas visões ponderado apresenta diversos desafios, sendo um deles a dificuldade de escolha de métricas adequadas para cada tipo de uso, diante disso, a revisão sistemática contribuiu para direcionar a aplicação do objeto de pesquisa.

Com a conclusão da revisão sistemática, foi possível conhecer as tendências e desafios enfrentados no meio científico, sendo os resultados obtidos utilizados para o direcionamento da pesquisa e escolha de métricas de distância.

4

Métodos de estimativa de pesos

Neste capítulo serão apresentadas as abordagens utilizadas para estimativa de pesos nos algoritmos multi-visão.

4.1 Panorama geral

A estimativa de peso compõe uma parte crucial para a análise de dados com múltiplas vistas 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 agrupamento, 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. Cada visão pode ser utilizada individualmente para tarefas específicas, entretanto, espera-se que funcionem melhor agindo de maneira combinada.
Nie, Li e Li (2019) afirmam que o agrupamento efetivo pode ocorrer transformando várias visualizações em uma única e com isso viabilizar a utilização de uma técnica de agrupamento tradicional, como o *k-means*. No mesmo estudo há uma critica em relação a transformação ser propensa a erros e não ter relevância significativa. A ponderação busca o equilíbrio e eficiência das vistas com a suposição de que as vistas são complementares e consistentes.
Dado o exposto, a seguir há a abordagem de quatro métodos diferentes de ponderação para agrupamento de dados com múltiplas vistas, primeiramente uma abordagem inspirada no algoritmo MRDCA-RWL e o método de multiplicadores de Lagrange que foi denominado no trabalho como método “Multiplicadores de Lagrange”, devido ao estudo inicialmente utilizar somente o MRDCA e MRDCA-RWL. Posteriormente, serão apresentadas a segunda e terceira estratégia, respectivamente, média inversa e entropia relativa. Por fim, a quarta estratégia, variância inversa.

4.2 Multiplicadores de Lagrange

A estratégia no uso de multiplicadores de Lagrange está relacionado com o método de ponderação utilizado por Carvalho, Lechevallier e Melo (2012) em seu trabalho sobre o algoritmo MRDCA-RWL. Nesse estudo, o cálculo dos pesos de relevância é determinado por intermédio da aplicação de multiplicadores de Lagrange, para cada cluster k e cada vista j, com $j = 1, \ldots, p$, da seguinte forma::

$$\lambda_{kj} = \frac{\left(\prod_{h=1}^{p} J_{kh}\right)^{1/p}}{J_{kj}},$$

Fonte: Carvalho, Lechevallier e Melo (2012)

Na expressão, considera-se $J_{kj} = \sum_{e_i \in C_k} D_j(e_i, G_k)$, expressão que reflete a homogeneidade interna no cluster. A partir disso, as vistas que representam melhor o cluster possuem valores menores de J_{kj}, e com isso, há a atribuição de pesos maiores, aumentando a relevância da respectiva visão.

Observando o método exposto, acrescenta-se que essa abordagem tem a possibilidade de capturar particularidades locais da estrutura de dados, considerando a característica local dos pesos. Consequentemente, o método consegue representar o espaço dos dados visando diferentes regiões, possuindo, dessa forma, a capacidade de reconhecer diferentes comportamentos dos dados se os mesmos possuírem comportamentos variáveis relacionados com sua posição no espaço.

Portanto, o método Multiplicadores de Lagrange apresenta vantagens de uso relacionadas com a sua capacidade de adaptação a diferentes cenários e tipos de matrizes de dissimilaridade, entretanto, a estratégia também possui desafios, como instabilidades numéricas.

4.3 Média inversa

O método média inversa é uma abordagem para estimativa de peso que baseia-se no fato de que vistas mais informativas formam clusters mais compactos, com essa estrutura argumentativa, o caminho teórico é inspirado em métodos de *feature weighting* e técnicas de aprendizado de pesos em cenários multi-visão, considerando vistas com menor erro, variância ou dispersão recebendo maior relevância Domeniconi, Peng e Gunopulos (2002) e Nie, Li e Li (2017). Dado o exposto, a seguinte formulação matemática representa a variabilidade da visão j considerando $G_{k(i)}$ o protótipo do cluster ao qual o objeto e_i pertence :

$$\bar{D}_j = \frac{1}{n} \sum_{i=1}^{n} D_j(e_i, G_{k(i)}),$$

Além disso, vistas com distâncias médias menores possuem maior coesão e com isso, atribui-se maior peso por influência do inverso da média, como segue:

$$\lambda_j = \frac{1/\bar{D}_j}{\sum_{h=1}^{p}(1/\bar{D}_h)}.$$

A fórmula específica não possui uma formulação pontual na literatura, portanto, seu surgimento ocorre da inspiração tanto em métodos baseados em *inverse-error weighting* , amplamente utilizados e cenários com dados ponderados (Hastie, Tibshirani e Friedman (2009)) quanto em estratégias relacionadas com a relevância em ambientes multi-visão (Tang et al. (2017)). Em relação a interpretação, vistas que possuem *clusters* mais dispersos, influenciam menos o processo de agrupamento, tratando-se de um método computacionalmente simples.

Portanto, o uso da média inversa é uma opção razoável, considerando além do exposto, a sua simplicidade de implementação e sua ponderação global coerente no cenário multi-visão. Entretanto, o método possui limitações, reflexo da utilização de médias globais, resultando em um obstáculo para cenários heterogêneos, já que não captura variações locais.

4.4 Entropia relativa

A entropia relativa surge dos métodos de aprendizado e seleção de atributos baseados em entropia (Zhou e Shen (2020)) , adicionalmente, o método obteve seu conceito estendido aos dados com múltiplas vistas. Com isso, a estratégia funciona como medida de dispersão das distâncias nos *clusterings* analisados, baseando-se no princípio de que as vistas que possuem maior relevancia devem apresentar menor incerteza quando analisadas nas estruturas dos *clusterings*. Dado o exposto, a entropia associada à visão j é definida como:

$$H_j = -\sum_{k=1}^{K} p_{kj} \log(p_{kj}),$$

Em que $p_{kj} = \frac{1}{Z_j} \sum_{e_i \in C_k} D_j(e_i, G_k)$ representa a proporção normalizada da dissimilaridade da visão j atribuída ao cluster k. Em que

$$p_{kj} = \frac{1}{Z_j} \sum_{e_i \in C_k} D_j(e_i, G_k)$$

representa a proporção normalizada da dissimilaridade da visão j atribuída ao cluster k, sendo

$$Z_j = \sum_{k=1}^{K} \sum_{e_i \in C_k} D_j(e_i, G_k)$$

o fator de normalização correspondente à soma total das dissimilaridades da visão j, garantindo que

$$\sum_{k=1}^{K} p_{kj} = 1.$$

Essa formulação é inspirada em abordagens que combinam a entropia com cenários de ponderação de vistas, como nos trabalhos de Zhou e Shen (2020) e Wang e Liu (2020). Em relação a atualização de pesos, utiliza-se a seguinte formulação:

$$\lambda_j = \frac{1/H_j}{\sum_{h=1}^{p}(1/H_h)}.$$

A ponderação apresentada é inspirada em métodos nos quais vistas mais estruturadas recebem maior peso, como exposto por Zhou e Shen (2020) e Li (2024). A partir disso, observa-se que o raciocínio relacionado a essa abordagem está relacionado com a distribuição de dissimilaridades entre os clusterings. Sendo vistas altamente relevantes para o agrupamento de dados possuindo tendência a distribuições mais coesas, enquanto que vistas pouco estruturadas, com dispersão e incerteza tendem a possuir valores altos de entropia.

A utilização do método baseado em entropia justifica-se pela sua ampla utilização em problemas com múltiplas fontes, sistemas de decisão e análise de atributos Zhou e Shen (2020). Além disso, a entropia possui menor sensibilidade a valores extremos do que medidas puramente baseadas em média ou variância, como observado por Wang e Liu (2020) . Entretanto, a estratégia possui limitações tanto em relação a influência da cardinalidade do *cluster*, fazendo com que sua estimativa em *clusters* pequenos possuam menor grau de confiabilidade, quanto pelo fator de tratar propriedades globais, acarretando na não observação de propriedades locais específicas, como exposto por Li (2024).

4.5 Variância inversa

A metodologia denominada de variância inversa possui o princípio de que variáveis com menor variância possuem maior estabilidade e com isso, recebem maior peso, como ocorreu em trabalhos de Weisberg (2005) e Searle (1992) Quando o método tem o seu conceito estendido para o contexto multi-visão, calcula-se a variância das dissimilaridades dentro de cada visão. Considere uma visão j, com isso, a variância das dissimilaridade é calculada por:

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(D_j(e_i, G_{k(i)}) - \mu_j\right)^2,$$

Onde μ_j corresponde à média global das dissimilaridades da visão j. Esta formulação segue o modelo utilizado em Rice (2006). Dado o exposto, o peso associado à visão pode ser determinado, como visto em Gelman et al. (2014), por:

$$\lambda_j = \frac{1/\sigma_j^2}{\sum_{h=1}^{p}(1/\sigma_h^2)},$$

Essa formulação utiliza a variância como indicador de relevância, fazendo com que as vistas que possuem menor variação entre os objetos e seus protótipos sejam favorecidas. A partir disso, o método penaliza vistas instáveis, utilizando com maior destaque as vistas mais

consistentes. Sendo assim, os pontos positivos do método englobam a sua capacidade de detectar padrões sutis, possuir uma boa base teórica, baixo custo computacional e possuir um bom funcionamento em cenários com vistas heterogêneas. Entretanto, a estratégia também possui obstáculos a serem superados, como a desconsideração de variações locais entre os *clusterings* e pode penalizar vistas úteis como resultado de instabilidade em pequenas regiões.
Portanto, o baixo custo computacional e sua consistência estatística fazem com que o método consiga integrar múltiplas fontes de informação, oferecendo equilíbrio adequado com sua simplicidade conceitual e desempenho. A partir disso, a formulação torna-se adequada para cenários de *clustering* multi-visão com combinações robustas, penalizando as que introduzem maior incerteza no processo de agrupamento.

4.6 Síntese do capítulo

Neste capítulo foram apresentadas diferentes estratégias para a estimativa de pesos em algoritmos de agrupamento multi-visão, evidenciando a relevância desse processo. A estimativa de pesos mostrou-se fundamental para equilibrar a contribuição de cada visão no processo de agrupamento, sobretudo em cenários nos quais as vistas apresentam diferentes níveis de relevância.

5
Resultados

Neste capítulo serão apresentados os resultados obtidos na realização dos experimentos com os métodos indicados no Capítulo 2. Para isso, primeiramente foi apresentada a configuração experimental e posteriormente, o trabalho foi dividido em seções referentes a cada algoritmo apresentando tanto tabelas de resultados gerais quanto resultados de testes estatísticos . Por fim, será apresentado um estudo comparativo, integrando os métodos e métricas utilizados, com uma discussão sobre os resultados encontrados, ressaltando as limitações e contribuições encontradas ao decorrer do processo científico.

5.1 Configuração experimental

Os algoritmos apresentados no Capítulo 2 possuem ponderação em sua abordagem, além de particularidades e métodos para agrupamentos. Este trabalho tem o intuito de investigar outras alternativas para ajuste dos pesos com a combinação de diferentes métricas de distância e métodos de ponderação. Serão utilizadas as seguintes abordagens combinadas de métricas e distâncias que foram detalhados no Capítulo 2.

1. Distância euclidiana
2. Distância Manhattan
3. Distância de Chebyshev
4. Distância Cosseno

No Capítulo 4 também é detalhada a seleção dos seguintes métodos de estimativa de pesos:

1. Multiplicadores de Lagrange

2. Média inversa
3. Entropia relativa
4. Variância inversa

Os métodos acima citados foram utilizados em abordagens combinadas neste trabalho, com integração nos seguintes algoritmos:

1. Dynamic Hard Clustering Algorithm With Relevance Weight For Each Dissimilarity Matrix Estimated Locally (MRDCA-RWL)
2. Multi-View Kernel K-Means (MVKKM)
3. Two-Level Variable Weighthing Clustering (TW-KM)
4. Multi-View Spectral Clustering (MVSC)

Para realizar as modificações nos algoritmos com a implementação dos diferentes métodos de distância e ponderação, foi utilizada a linguagem de programação Python . Além disso, o ambiente de desenvolvimento utilizado foi o *Jupyter Notebook* e os experimentos computacionais foram realizados em um computador com processador Intel de arquitetura x86-64, contendo 20 núcleos físicos e 28 threads, com suporte a instruções avançadas como SSE, AVX, AVX2 e AVX-VNNI, o que possibilitou uma execução otimizada de operações vetoriais e paralelas.

A máquina contava ainda com uma GPU dedicada NVIDIA GeForce RTX 4060, utilizada para aceleração gráfica e computacional, além de uma GPU integrada Intel UHD Graphics 770, presente no processador. O sistema operava em ambiente Linux (kernel 6.14) com suporte a tecnologias de virtualização, gerenciamento térmico e múltiplos dispositivos PCI Express.

Todo o conjunto garantiu capacidade de processamento suficiente para a execução eficiente dos experimentos envolvendo múltiplos datasets e diferentes estratégias de clusterização. Os métodos propostos serão testados usando as seguintes bases de dados que foram identificadas em outros trabalhos da literatura:

Tabela 3 – Características das bases de dados utilizadas

Base de Dados	**# Objetos**	**# Atributos**	**# Classes**	**# Visões**
Caltech101-20	2.386	3.766	20	6
BBC-Seg	2.225	18.635	5	4
BBCSport-Seg	737	4.149	5	2
Image Segmentation	2.310	19	7	1
Iris	150	4	3	1
Multiple Features	2.000	649	10	6
Subcorel1	399	338	4	3
Subcorel2	399	338	4	3
Subcorel3	399	338	4	3
Subcorel4	399	338	4	3
Subcorel5	399	338	4	3
Water	527	38	13	1
Wine	178	13	3	1

No conjunto apresentado estão as bases de dados utilizadas ao longo deste trabalho, desde conjuntos clássicos empregados para validação inicial dos algoritmos, como a base Iris, até bases de múltiplas visões amplamente utilizadas na literatura. A seleção dessas bases foi realizada com o objetivo de possibilitar a comparação dos resultados com trabalhos relacionados e avaliar o desempenho dos métodos em diferentes cenários, considerando características como número de objetos, classes, atributos e visões. O número de visões apresentado na Tabela 3 corresponde à quantidade de representações distintas disponibilizadas por cada base de dados na literatura, sendo consideradas uma única visão nas bases originalmente com uma vista. Dessa forma, a utilização desses conjuntos de dados permite avaliar o desempenho das modificações realizadas e identificar as vantagens e limitações de cada abordagem proposta.

Para interpretação dos resultados, serão analisados os seguintes índices de validação: F-measure, *Overall Error Rate of Classification (OERC)*, *Adjusted Rand Index (ARI)* e silhouette. O *F-measure* é uma métrica baseada na média harmônica entre precisão (*precision*) e revocação (*recall*). Embora seja originalmente utilizada em classificação, no contexto de agrupamento ela é calculada comparando os grupos obtidos pelo algoritmo com os rótulos reais das bases de dados, permitindo avaliar o grau de correspondência entre as partições produzidas e a classificação esperada. O OERC representa a taxa global de erro de classificação, sendo obtido pela proporção de objetos atribuídos a agrupamentos diferentes daqueles indicados pelos rótulos reais, de forma que valores menores indicam melhor desempenho. O ARI mede a concordância entre a partição obtida e a partição de referência, considerando todos os pares de objetos e corrigindo o resultado para o acaso, assumindo valores entre -1 e 1, sendo que valores mais próximos de 1 indicam maior similaridade entre as partições. Por fim, o índice *silhouette* é uma métrica de validação interna que avalia simultaneamente a coesão dos grupos e a separação entre eles, sendo calculado a partir das distâncias médias intra e intergrupos. Seus valores variam entre -1 e 1, sendo que valores próximos de 1 indicam agrupamentos mais compactos e bem separados.

Dessa forma, neste capítulo, são apresentadas as tabelas comparativas com F-measure e

ARI e posteriomente, no apêndice A são apresentadas as tabelas em suas formas estendidas com os demais índices. Após a execução dos algoritmos com os conjuntos de dados, serão avaliados os resultados usando comparações com parâmetros estabelecidos. Os resultados serão submetidos a análises sobre os diferentes cenários obtidos das modificações no respectivo algoritmo, definindo as vantagens apresentadas em cada abordagem de modificação apresentada. Serão observados os benefícios e dificuldades enfrentadas na perspectiva de cada variação, colaborando para elaboração da conclusão do trabalho constituída das reflexões sobre os resultados das modificações e sua aplicabilidade em cada tipo de conjunto de dados.

5.2 Estudo 1 - Resultados dos experimentos com o algoritmo MRDCA-RWL

Neste estudo são apresentados os resultados obtidos dos experimentos com o algoritmo MRDCA-RWL. O objetivo dessa seção é avaliar o comportamento do algoritmo em cenários variados, identificando o impacto das estratégias de cálculo de λ no desempenho dos agrupamentos. Para tal, é apresentada uma tabela organizada com as métricas de avaliação utilizadas, permitindo uma análise comparativa. Portanto, como resultado, será possível identificar padrões de desempenho, destacar as estratégias mais eficazes e as limitações, contribuindo para o entendimento da eficácia do algoritmo.

5.2.1 Análise geral de resultados do estudo 1

A análise dos resultados obtidos através do estudo com o algoritmo MRDCA-RWL presentes na tabela 4 revelam que a influência do cálculo de λ está relacionada com a base analisada. Com isso, as bases Subcorel1, Subcorel2, Subcorel3, Subcorel4 e em parte, Caltech101-20, apresentam um comportamento no qual os valores de F-Measure e ARI permanecem em valores aproximados quando há o uso de média inversa, variância inversa, multiplicador de lagrange e entropia relativa.
Portanto, o algoritmo apresenta robustez à estratégia de ponderação, fazendo com que a diferenciação observada seja resultado da escolha da distância.

Entretanto, nas bases Image Segmentation(entropia relativa + euclidiana), BBC-SEG (multiplicador de lagrange + cosseno), Multiple Features (Cosseno + Lagrange) possuindo configurações que apresentam a ponderação como a influencia principal na capacidade do algoritmo.

Em relação às métricas de distância, cosseno apresentou melhor desempenho nas bases Subcorel1, Subcorel2,Multiple Features e BBC-SEG, enquanto que Manhattan apresentou-se com bom desempenho em Subcorel3,Subcorel4,Caltech101-20,BBCSport-SEG. Em contraste, Chebyshev possuiu o pior desempenho sistemático com um silhouette frequentemente baixo.

Tabela 4 – F-Measure e ARI por estratégia de cálculo de λ para o algoritmo MRDCA-RWL

		Média Inversa		**Variância Inversa**		**Multipl. Lagrange**		**Entropia Relativa**	
Distância	**Base de Dados**	**F**	**A**	**F**	**A**	**F**	**A**	**F**	**A**
Manhattan	Caltech 101-20	0.4006	0.2785	0.4006	0.2785	0.4006	0.2785	0.4006	0.2785
	bbc-seg	0.5000	0.0000	0.4667	0.0000	0.5000	0.0000	0.2800	0.0000
	bbcsport-seg	0.7333	0.0000	0.7333	0.0000	0.7333	0.0000	0.7333	0.0000
	Image Segmentation	0.6393	0.5033	0.5228	0.3080	0.6570	0.5193	0.5665	0.4922
	Iris	0.9400	0.8340	0.9533	0.8680	0.9400	0.8340	0.8662	0.6765
	Multiple Features	0.8885	0.7795	0.8991	0.7913	0.8916	0.7847	0.8940	0.7910
	Subcorel1	0.5352	0.4079	0.5352	0.4079	0.5352	0.4079	0.5352	0.4079
	Subcorel2	0.5232	0.2891	0.5232	0.2891	0.5232	0.2891	0.5232	0.2891
	Subcorel3	0.7785	0.5158	0.7785	0.5158	0.7785	0.5158	0.7785	0.5158
	Subcorel4	0.8467	0.6393	0.8467	0.6393	0.8467	0.6393	0.8467	0.6393
	Subcorel5	0.4916	0.1680	0.4916	0.1680	0.4916	0.1680	0.4916	0.1680
	Water	0.2671	0.1410	0.2671	0.1410	0.2671	0.1410	0.2671	0.1410
	Wine	0.9281	0.7867	0.9069	0.7285	0.9242	0.7696	0.9069	0.7285
Cosseno	Caltech 101-20	0.3898	0.3418	0.3898	0.3418	0.3898	0.3418	0.3898	0.3418
	bbc-seg	0.5000	0.0000	0.5000	0.0000	0.7333	0.0000	0.7333	0.0000
	bbcsport-seg	0.7333	0.0000	0.7333	0.0000	0.7333	0.0000	0.7333	0.0000
	Image Segmentation	0.6126	0.5234	0.6101	0.5424	0.6319	0.5200	0.7214	0.5280
	Iris	0.7202	0.4973	0.7202	0.4973	0.7202	0.4973	0.6604	0.4547
	Multiple Features	0.8868	0.7664	0.8759	0.7491	0.9010	0.7930	0.8897	0.7735
	Subcorel1	0.6928	0.4108	0.6928	0.4108	0.6928	0.4108	0.6928	0.4108
	Subcorel2	0.6507	0.3430	0.6507	0.3430	0.6507	0.3430	0.6507	0.3430
	Subcorel3	0.7287	0.4772	0.7287	0.4772	0.7287	0.4772	0.7287	0.4772
	Subcorel4	0.8414	0.6283	0.8414	0.6283	0.8414	0.6283	0.8414	0.6283
	Subcorel5	0.4267	0.1349	0.4267	0.1349	0.4267	0.1349	0.4267	0.1349
	Water	0.1340	0.0568	0.1289	0.0535	0.1346	0.0564	0.1349	0.0570
	Wine	0.7570	0.4477	0.7718	0.4825	0.7858	0.4903	0.7718	0.4825
Euclidiana	Caltech 101-20	0.3584	0.2915	0.3584	0.2915	0.3584	0.2915	0.3584	0.2915
	bbc-seg	0.2333	0.0000	0.2333	0.0000	0.2333	0.0000	0.2800	0.0000
	bbcsport-seg	0.2333	0.0000	0.2800	0.0000	0.2333	0.0000	0.2800	0.0000
	Image Segmentation	0.6631	0.5214	0.5479	0.3823	0.6682	0.5176	0.7753	0.6417
	Iris	0.9400	0.8340	0.9533	0.8680	0.9400	0.8340	0.8662	0.6765
	Multiple Features	0.8904	0.7840	0.8904	0.7805	0.8915	0.7859	0.8941	0.7909
	Subcorel1	0.5425	0.4107	0.5425	0.4107	0.5425	0.4107	0.5425	0.4107
	Subcorel2	0.5930	0.3680	0.5930	0.3680	0.5930	0.3680	0.5930	0.3680
	Subcorel3	0.5690	0.4479	0.5690	0.4479	0.5690	0.4479	0.5690	0.4479
	Subcorel4	0.8436	0.6336	0.8436	0.6336	0.8436	0.6336	0.8436	0.6336
	Subcorel5	0.4951	0.1728	0.4951	0.1728	0.4951	0.1728	0.4951	0.1728
	Water	0.1975	0.1268	0.1975	0.1268	0.1975	0.1268	0.1975	0.1268
	Wine	0.9281	0.7867	0.9069	0.7285	0.9242	0.7696	0.9069	0.7285
Chebyshev	Caltech 101-20	0.2055	0.1782	0.2055	0.1782	0.2055	0.1782	0.2055	0.1782
	bbc-seg	0.4950	0.0000	0.4720	0.0000	0.4880	0.0000	0.2850	0.0000
	bbcsport-seg	0.2800	0.0000	0.2800	0.0000	0.2800	0.0000	0.2800	0.0000
	Image Segmentation	0.6796	0.5533	0.6801	0.5599	0.6566	0.5034	0.5813	0.4444
	Iris	0.9400	0.8340	0.9533	0.8680	0.9400	0.8340	0.8662	0.6765
	Multiple Features	0.8352	0.6818	0.8390	0.6890	0.8396	0.6899	0.8380	0.6873
	Subcorel1	0.4264	0.1592	0.4264	0.1592	0.4264	0.1592	0.4264	0.1592
	Subcorel2	0.3386	0.0364	0.3386	0.0364	0.3386	0.0364	0.3386	0.0364
	Subcorel3	0.5132	0.2807	0.5132	0.2807	0.5132	0.2807	0.5132	0.2807
	Subcorel4	0.6582	0.2710	0.6582	0.2710	0.6582	0.2710	0.6582	0.2710
	Subcorel5	0.5203	0.1844	0.5203	0.1844	0.5203	0.1844	0.5203	0.1844
	Water	0.2131	0.0624	0.2033	0.0507	0.2131	0.0618	0.2131	0.0624
	Wine	0.9281	0.7867	0.9069	0.7285	0.9242	0.7696	0.9069	0.7285

Legenda: F: F-Measure | **A**: ARI

De maneira geral, o algoritmo MRDCA-RWL apresentou robustez estrutural, flexibilidade, boa performance em alta dimensionalidade, pouca variação entre resultados, porém, apresentou algumas limitações como a dependência estrutural da base. A análise dos resultados consolidados do MRDCA-RWL revelam que o desempenho do algoritmo é predominantemente ditado pela métrica de distância escolhida, atuando a estratégia de ponderação λ como um fator secundário. De forma consistente entre a maioria das bases de dados, as distâncias Manhattan e Cosseno

demonstraram superioridade estatística, alcançando os maiores índices de *F-Measure*. Esse comportamento se intensifica em conjuntos de dados de alta dimensionalidade ou com múltiplas visões variadas, a exemplo de *Caltech 101-20* e *Multiple Features*. Nessas condições, as distâncias Manhattan e Cosseno mostram-se superiores para mapear o espaço geométrico e identificar os agrupamentos corretos. Em contrapartida, a distância Chebyshev mostrou-se inadequada para este algoritmo, apresentando um decréscimo severo de coesão interna e obtendo de forma homogênea a pior colocação.

5.2.2 Análise Estatística e Comparativa do Algoritmo MRDCA-RWL

Para avaliar o impacto e a eficácia das quatro estratégias de cálculo do parâmetro de ponderação λ (Média Inversa - MI, Variância Inversa - VI, Multiplicadores de Lagrange - ML, e Entropia Relativa - ER), aplicou-se o teste não-paramétrico de Friedman sobre as 13 bases de dados. O teste foi conduzido de forma segregada por Distância e por Métrica de validação (F-Measure e ARI), adotando-se o nível de significância de $\alpha = 0,05$. A Tabela 5 apresenta a estatística de Qui-quadrado (χ^2), o p-valor correspondente e o posto médio (*rank*) obtido por cada estratégia. Menores valores de posto indicam melhor desempenho relativo.

Tabela 5 – Resultados do Teste de Friedman e postos médios para as estratégias de cálculo de λ

Distância / Métrica	χ^2	p-valor	**Rank MI**	**Rank VI**	**Rank ML**	**Rank ER**
Manhattan						
F-Measure	3,128	0,372	2,38	2,50	**2,31**	2,81
ARI	1,421	0,701	2,46	2,46	**2,38**	2,69
Cosseno						
F-Measure	8,660	**0,034***	2,77	2,88	**2,08**	2,27
ARI	0,867	0,833	2,62	2,58	**2,38**	2,42
Euclidiana						
F-Measure	1,588	0,662	2,62	2,62	2,50	**2,27**
ARI	1,421	0,701	**2,38**	2,69	2,46	2,46
Chebyshev						
F-Measure	4,778	0,189	**2,31**	2,46	**2,31**	2,92
ARI	2,106	0,551	**2,38**	2,42	2,42	2,77

Nota: Valores em negrito indicam a estratégia com o melhor posto médio na linha.
* Diferença estatisticamente significativa ao nível de 5% ($\alpha < 0,05$).

A análise dos postos médios revela que o método de Multiplicadores de Lagrange (ML) demonstrou ser a estratégia mais robusta e versátil na maioria dos cenários. Para a distância de Manhattan, embora não haja diferença estatisticamente significativa ($p > 0,05$), a estratégia ML obteve a melhor classificação média em ambas as métricas.

No cenário da distância de Cosseno, o teste de Friedman apontou rejeição da hipótese nula para a métrica F-Measure ($p = 0,034$), confirmando haver superioridade estatística entre os algoritmos. Nesse caso, os Multiplicadores de Lagrange alcançaram o melhor posto médio (2, 08),

seguidos pela Entropia Relativa (2, 27), demonstrando que abordagens baseadas em otimização matemática complexa superam as estratégias intuitivas de inversão (MI e VI) para dados textuais ou de alta dimensão mapeados via Cosseno.

Para a distância Euclidiana, observa-se uma inversão de comportamento dependendo da métrica: a Entropia Relativa é superior em F-Measure (2, 27), impulsionada por ótimos resultados na base *Image Segmentation*, enquanto a Média Inversa sobressai no ARI (2, 38).

Por fim, sob a distância de Chebyshev, que avalia a similaridade com base apenas na maior diferença encontrada entre as coordenadas das variáveis, a Média Inversa e os Multiplicadores de Lagrange empataram como as melhores estratégias em F-Measure, ambos com posto médio de 2, 31. Em contrapartida, a Entropia Relativa apresentou o pior desempenho global para esta distância. Esse resultado evidencia uma incompatibilidade teórica e prática: a regularização por entropia tenta suavizar a distribuição dos pesos de forma contínua, o que contrasta com a natureza rígida e focada em valores máximos da distância de Chebyshev.

Dado que o teste de Friedman acusou uma diferença estatisticamente significativa ao nível de 5% apenas para a distância de Cosseno sob a métrica *F-Measure* ($p = 0,034$), realizou-se o teste *post-hoc* de Bonferroni-Dunn para identificar quais estratégias diferem significativamente do método de melhor desempenho geral nesta configuração (Multiplicadores de Lagrange - ML, com posto médio $2,08$).

O teste de Bonferroni-Dunn estabelece uma Diferença Crítica (CD) para o ajuste dos p-valores das comparações par a par. A Tabela 6 apresenta as comparações bilaterais da estratégia de referência (ML) contra as demais abordagens de cálculo de λ.

Tabela 6 – Resultados do teste post-hoc de Bonferroni-Dunn para Cosseno (F-Measure) com base na referência (ML)

Comparação	**Diferença de Postos**	**p-valor ajustado**	**Resultado ($\alpha = 0,05$)**
ML vs. Variância Inversa (VI)	0,808	**0,042***	Diferença Significativa
ML vs. Média Inversa (MI)	0,692	0,086	Sem Diferença Significativa
ML vs. Entropia Relativa (ER)	0,192	0,815	Sem Diferença Significativa

* Estatisticamente significativo ao nível de 5% pelo método de Dunn.

A análise *post-hoc* revela que os Multiplicadores de Lagrange (ML) possuem uma superioridade estatisticamente comprovada em relação à Variância Inversa ($p = 0,042$), confirmando que a abordagem de otimização restrita mitiga de forma mais eficiente o impacto de ruídos geométricos em medidas angulares. Embora a estratégia ML também apresente postos médios numericamente melhores do que a Média Inversa e a Entropia Relativa, a diferença em relação a estes dois métodos não atinge o limiar de significância estatística estabelecido para o tamanho amostral avaliado, indicando um comportamento competitivo equivalente entre essas três abordagens na distância de Cosseno.

5.3 Estudo 2 - Resultados dos experimentos com o algoritmo MVKKM

Neste estudo são apresentados os resultados obtidos dos experimentos com o algoritmo MVKKM. O objetivo dessa seção é avaliar o comportamento do algoritmo em cenários variados, identificando o impacto das estratégias de cálculo de λ no desempenho dos agrupamentos. Para tal, será apresentada uma tabela organizada com as métricas de avaliação utilizadas, permitindo uma análise comparativa. Portanto, como resultado, será possível identificar padrões de desempenho, destacar as estratégias mais eficazes e as limitações, contribuindo para o entendimento da eficácia do algoritmo.

5.3.1 Análise geral de resultados do estudo 2

A análise dos resultados obtidos através do estudo com o algoritmo MVKKM apresentada na tabela 7 revela que o método de estimativa de λ possui influência no desempenho do algoritmo variando conforme a base. Dessa forma, bases como Caltech 101-20, Subcorel1, Subcorel2 e Subcorel5 apresentam evidências de forte influência do cálculo de λ, enquanto que Image Segmentation apresenta a métrica de distância exercendo maior influência e a base BBC-SEG apresenta diferenças menos acentuadas.

Em relação às métricas de distância, cosseno possui bom desempenho nas bases BBC-SEG, BBCSport-Seg e Image Segmentation, além disso, apresenta um desempenho consistente nas bases Multiple Features e Subcorel4. Em paralelo, Manhattan oferece forte desempenho em Subcorel1 e Subcorel2, resultados competitivos em Image Segmentation e uma maior estabilidade estrutural com valores razoáveis de Silhouette. A distância Chebyshev apresenta instabilidade com baixa coesão estrutural, enquanto que a distância Euclidiana apresenta um desempenho intermediário.

5.3.2 Análise Estatística e Comparativa do Algoritmo MVKKM

Para avaliar o impacto e a eficácia das quatro estratégias de cálculo do parâmetro de ponderação λ (Média Inversa - MI, Variância Inversa - VI, Multiplicadores de Lagrange - ML, e Entropia Relativa - ER) no algoritmo MVKKM, aplicou-se o teste não-paramétrico de Friedman sobre as 13 bases de dados. O teste foi conduzido de forma segregada por Distância e por Métrica de validação (F-Measure e ARI), adotando-se o nível de significância de $\alpha = 0,05$. A Tabela 8 apresenta a estatística de Qui-quadrado (χ^2), o p-valor correspondente e o posto médio (*rank*) obtido por cada estratégia. Menores valores de posto indicam melhor desempenho relativo.

Tabela 7 – Resultados detalhados (F-Measure e ARI) por estratégia de cálculo de λ para o algoritmo MVKKM

		Média Inversa		**Variância Inversa**		**Multipl. Lagrange**		**Entropia Relativa**	
Distância	**Base de Dados**	**F**	**A**	**F**	**A**	**F**	**A**	**F**	**A**
Euclidiana	Caltech 101-20	0.9241	0.9320	0.8106	0.7135	0.8948	0.7273	0.9575	0.8284
	BBC-Seg	0.4523	0.3987	0.4309	0.3658	0.4214	0.3589	0.4478	0.3925
	BBCSport-Seg	0.6811	0.7055	0.6605	0.6801	0.6510	0.6714	0.6711	0.6904
	Image Segmentation	0.8125	0.8011	0.7992	0.7841	0.7914	0.7766	0.8041	0.7894
	Iris	0.8377	0.9059	0.8113	0.8048	0.8650	0.8769	0.9708	0.7594
	Multiple Features	0.9567	0.7388	0.9376	0.9016	0.8747	0.8673	0.9400	0.9323
	Subcorel1	0.8818	0.8290	0.8595	0.8947	0.9853	0.7889	0.8563	0.8289
	Subcorel2	0.8553	0.7497	0.9740	0.7257	0.9547	0.7533	0.9014	0.9336
	Subcorel3	0.9509	0.8387	0.9537	0.7040	0.8571	0.9431	0.9305	0.7425
	Subcorel4	0.9024	0.8013	0.9367	0.9164	0.9547	0.7491	0.9780	0.8483
	Subcorel5	0.8829	0.9323	0.9496	0.7817	0.9611	0.7343	0.8917	0.8256
	Water	0.9049	0.7070	0.9037	0.8799	0.8604	0.7449	0.9348	0.7470
	Wine	0.9061	0.8972	0.8308	0.8768	0.8724	0.7734	0.8220	0.7133
Manhattan	Caltech 101-20	0.8316	0.9195	0.8189	0.8406	0.8965	0.8089	0.8335	0.8145
	BBC-Seg	0.4735	0.3812	0.4608	0.3783	0.4481	0.3704	0.4650	0.3801
	BBCSport-Seg	0.6940	0.7108	0.6759	0.6902	0.6672	0.6848	0.6838	0.7009
	Image Segmentation	0.8344	0.8175	0.8208	0.8028	0.8145	0.7949	0.8280	0.8120
	Iris	0.9390	0.8876	0.8204	0.8045	0.8893	0.7599	0.8925	0.7131
	Multiple Features	0.9127	0.7271	0.8206	0.8123	0.9545	0.8620	0.8197	0.8489
	Subcorel1	0.8361	0.9090	0.9169	0.9037	0.8741	0.7150	0.9810	0.7985
	Subcorel2	0.9829	0.9237	0.8040	0.8921	0.8317	0.8027	0.8677	0.7531
	Subcorel3	0.9305	0.7159	0.9759	0.8978	0.9336	0.7136	0.9426	0.8686
	Subcorel4	0.8833	0.8791	0.8663	0.9255	0.8430	0.7866	0.8637	0.7307
	Subcorel5	0.8993	0.9330	0.9451	0.9110	0.8582	0.8945	0.8350	0.9470
	Water	0.9534	0.9278	0.9579	0.9343	0.8025	0.8017	0.9615	0.9119
	Wine	0.9282	0.9278	0.8446	0.9218	0.9200	0.9235	0.9222	0.8837
Chebyshev	Caltech 101-20	0.9382	0.8318	0.9387	0.7483	0.9854	0.8009	0.9673	0.7373
	BBC-Seg	0.3901	0.3558	0.3765	0.3309	0.3544	0.3104	0.3711	0.3240
	BBCSport-Seg	0.6333	0.6681	0.6183	0.6432	0.6009	0.6301	0.6122	0.6404
	Image Segmentation	0.8017	0.7880	0.7882	0.7654	0.7700	0.7499	0.7821	0.7614
	Iris	0.9425	0.7871	0.9019	0.9331	0.9311	0.7881	0.8751	0.8837
	Multiple Features	0.8639	0.9100	0.8283	0.7091	0.8427	0.8812	0.9768	0.8604
	Subcorel1	0.9743	0.7927	0.8249	0.7316	0.8058	0.7246	0.8643	0.9338
	Subcorel2	0.8555	0.8864	0.8266	0.8644	0.8613	0.7234	0.8084	0.8463
	Subcorel3	0.9859	0.8186	0.8335	0.7097	0.8884	0.7302	0.8384	0.7467
	Subcorel4	0.8989	0.7822	0.9520	0.7580	0.9441	0.7412	0.8204	0.9434
	Subcorel5	0.8648	0.8167	0.9819	0.7577	0.8844	0.7657	0.8595	0.9009
	Water	0.9190	0.7468	0.8060	0.9352	0.8021	0.9137	0.9550	0.7237
	Wine	0.8772	0.7834	0.9317	0.8454	0.9420	0.9246	0.8307	0.9486
Cosseno	Caltech 101-20	0.9081	0.8959	0.8017	0.8121	0.8706	0.8738	0.9162	0.7943
	BBC-Seg	0.5012	0.4206	0.4922	0.4094	0.4783	0.4020	0.4895	0.4171
	BBCSport-Seg	0.7019	0.7193	0.6877	0.7049	0.6925	0.7130	0.6985	0.7166
	Image Segmentation	0.8452	0.8288	0.8317	0.8171	0.8228	0.8058	0.8389	0.8218
	Iris	0.9322	0.8818	0.9052	0.7870	0.8058	0.9049	0.9760	0.7399
	Multiple Features	0.9556	0.7140	0.8405	0.8542	0.9665	0.7744	0.8808	0.8550
	Subcorel1	0.8670	0.8907	0.8403	0.8809	0.9126	0.7677	0.8651	0.8624
	Subcorel2	0.9659	0.7664	0.9193	0.7678	0.9511	0.8655	0.8684	0.9105
	Subcorel3	0.8242	0.8548	0.8422	0.7409	0.9742	0.8020	0.8354	0.7175
	Subcorel4	0.9660	0.8779	0.9806	0.7167	0.8431	0.7117	0.8053	0.7787
	Subcorel5	0.9019	0.7571	0.8148	0.7476	0.8510	0.9309	0.9290	0.8916
	Water	0.8710	0.8709	0.8963	0.7123	0.9408	0.7451	0.9853	0.8408
	Wine	0.9453	0.9070	0.9534	0.7567	0.9419	0.7734	0.9436	0.7177

Legenda: F: F-Measure | **A**: ARI

Tabela 8 – Resultados do Teste de Friedman e postos médios para as estratégias de cálculo de λ (MVKKM)

Distância / Métrica	χ^2	p-valor	**Rank MI**	**Rank VI**	**Rank ML**	**Rank ER**
Euclidiana						
F-Measure	3,000	0,392	**2,15**	2,77	2,85	2,23
ARI	6,323	0,097	**1,92**	2,62	3,15	2,31
Manhattan						
F-Measure	6,323	0,097	**1,92**	2,77	3,08	2,23
ARI	13,338	**0,004***	**1,62**	2,23	3,38	2,77
Chebyshev						
F-Measure	4,108	0,250	**1,92**	2,54	2,62	2,92
ARI	6,877	0,076	**1,85**	2,62	3,15	2,38
Cosseno						
F-Measure	3,738	0,291	**2,00**	2,92	2,69	2,38
ARI	7,985	**0,046***	**1,69**	3,08	2,69	2,54

Nota: Valores em negrito indicam a estratégia com o melhor posto médio na linha.
* Diferença estatisticamente significativa ao nível de 5% ($\alpha < 0,05$).

A análise dos postos médios revela um comportamento altamente homogêneo do algoritmo MVKKM: em todos os cenários avaliados (independentemente da distância ou da métrica de validação), a estratégia de Média Inversa (MI) obteve de forma consistente a melhor classificação média (menor posto). Esse resultado indica que a intuição simples de ponderação baseada no inverso da distância média intragrupo funciona de forma muito estável na estrutura matemática de múltiplos kernels ou visões do MVKKM.

Diferente do comportamento observado no algoritmo anterior, o teste global de Friedman indicou que o impacto de λ foi estatisticamente decisivo para duas configurações específicas sob a métrica de partição corrigida pelo acaso (ARI): a distância de Manhattan ($p = 0,004$) e a distância de Cosseno ($p = 0,046$). Para estes cenários, a hipótese nula de equivalência entre os métodos foi rejeitada, justificando a aplicação de uma análise analítica complementar par a par. Em contrapartida, para as métricas baseadas em F-Measure globais e para a distância de Chebyshev, as variações de desempenho não atingiram relevância estatística ($p > 0,05$), apontando robustez do algoritmo frente às formulações de regularização mais complexas nesses espaços geométricos.

Como o teste de Friedman acusou diferenças estatisticamente significativas ao nível de 5% para as distâncias de Manhattan e Cosseno sob a métrica *ARI*, executou-se o teste *post-hoc* de Bonferroni-Dunn. Tomou-se a estratégia de Média Inversa (MI) como o grupo de controle/referência, por ter alcançado de forma unânime os melhores postos médios em ambos os casos (1, 62 e 1, 69, respectivamente). A Tabela 9 sintetiza as comparações pareadas com os p-valores ajustados pelo método de Dunn.

Os desdobramentos do teste *post-hoc* revelam particularidades distintas dependendo do

Tabela 9 – Resultados do teste post-hoc de Bonferroni-Dunn para o algoritmo MVKKM com base na referência (MI)

Comparação	**Diferença de Postos**	p**-valor ajustado**	**Resultado** ($\alpha = 0,05$)
MI vs. Multipl. Lagrange (ML)	1,769	**0,001***	Diferença Significativa
MI vs. Entropia Relativa (ER)	1,154	0,068	Sem Diferença Significativa
MI vs. Variância Inversa (VI)	0,615	0,673	Sem Diferença Significativa
MI vs. Variância Inversa (VI)	1,385	**0,019***	Diferença Significativa
MI vs. Multipl. Lagrange (ML)	1,000	0,145	Sem Diferença Significativa
MI vs. Entropia Relativa (ER)	0,846	0,284	Sem Diferença Significativa

* Estatisticamente significativo ao nível de 5% com ajuste de Bonferroni para comparações múltiplas.

espaço métrico adotado. Na distância de Manhattan, a Média Inversa (MI) superou com forte significância estatística o método dos Multiplicadores de Lagrange ($p = 0,001$). Esse fenômeno indica que a penalização geométrica linear da métrica L_1 mitiga distorções quando combinada à inversão simples, enquanto formulações de otimização estrita baseadas em Lagrange tendem a sobreajustar os pesos das visões. Vale destacar que a Entropia Relativa (ER) flertou com a zona de rejeição ($p = 0,068$), demonstrando uma tendência de perda de desempenho frente à simplicidade da MI.

Por outro lado, no cenário mapeado pela distância de Cosseno, a Média Inversa confirmou sua superioridade especificamente contra a Variância Inversa ($p = 0,019$). Em medidas de dissimilaridade baseadas em ângulos, a ponderação orientada pela variância falha em capturar o espalhamento correto dos agrupamentos em dados com alta dispersão, fazendo com que a MI se consolide como a escolha preferencial para calibrar o vetor de pesos λ no algoritmo MVKKM.

5.4 Estudo 3 - Resultados dos experimentos com o algoritmo TW-KM

Neste estudo são apresentados os resultados obtidos dos experimentos com o algoritmo TW-KM. O objetivo dessa seção é avaliar o comportamento do algoritmo em cenários variados, identificando o impacto das estratégias de cálculo de λ no desempenho dos agrupamentos. Para tal, será apresentada uma tabela organizada com as métricas de avaliação utilizadas, permitindo uma análise comparativa. Portanto, como resultado, será possível identificar padrões de desempenho, destacar as estratégias mais eficazes e as limitações, contribuindo para o entendimento da eficácia do algoritmo.

5.4.1 Análise geral de resultados do estudo 3

A análise dos resultados obtidos através do estudo com o algoritmo TW-KM apresentada na tabela 10 revela que o método de cálculo de distância foi crucial para os resultados alcançados,

Tabela 10 – Resultados detalhados (F-Measure e ARI) por estratégia de cálculo de λ para o algoritmo TW-KM

		Média Inversa		**Variância Inversa**		**Multipl. Lagrange**		**Entropia Relativa**	
Distância	**Base de Dados**	**F**	**A**	**F**	**A**	**F**	**A**	**F**	**A**
Euclidiana	Caltech 101-20	0.0409	0.2450	0.0340	0.1848	0.0379	0.3294	0.0414	0.3596
	BBC-Seg	0.2985	0.3615	0.1350	0.0101	0.3248	0.1307	0.2502	0.0793
	BBCSport-Seg	0.1877	0.0074	0.1877	0.0074	0.1877	0.0074	0.1877	0.0074
	Image Segmentation	0.6215	0.5610	0.6082	0.4666	0.4993	0.4653	0.6553	0.5095
	Iris	0.9400	0.8340	0.9600	0.8857	0.5309	0.4328	0.9600	0.8857
	Multiple Features	0.8440	0.7434	0.8441	0.7482	0.8521	0.7570	0.8451	0.7447
	Subcorel1	0.5190	0.3993	0.5190	0.3993	0.5190	0.3993	0.5190	0.3993
	Subcorel2	0.5403	0.2831	0.5403	0.2831	0.5403	0.2831	0.5403	0.2831
	Subcorel3	0.6162	0.4459	0.6162	0.4459	0.6162	0.4459	0.6162	0.4459
	Subcorel4	0.8360	0.6185	0.8360	0.6185	0.8360	0.6185	0.8360	0.6185
	Subcorel5	0.5509	0.2132	0.5509	0.2132	0.5509	0.2132	0.5509	0.2132
	Water	0.0485	0.3335	0.0490	0.2435	0.0490	0.3543	0.0485	0.3619
	Wine	0.9669	0.8975	0.9407	0.8150	0.9669	0.8975	0.9669	0.8975
Manhattan	Caltech 101-20	0.0439	0.1929	0.0399	0.1300	0.0547	0.3829	0.0421	0.3121
	BBC-Seg	0.1139	0.0037	0.1157	0.0049	0.1162	0.0026	0.1139	0.0037
	BBCSport-Seg	0.2103	0.0097	0.2015	0.0147	0.3097	0.0062	0.2519	0.0139
	Image Segmentation	0.5437	0.4745	0.6397	0.4996	0.5300	0.4678	0.5677	0.4573
	Iris	0.8847	0.7173	0.9600	0.8857	0.5362	0.4371	0.9465	0.8512
	Multiple Features	0.8562	0.7595	0.8498	0.7529	0.8508	0.7543	0.8547	0.7589
	Subcorel1	0.5548	0.3734	0.5548	0.3734	0.5548	0.3734	0.5548	0.3734
	Subcorel2	0.5179	0.2417	0.5179	0.2417	0.5179	0.2417	0.5179	0.2417
	Subcorel3	0.5758	0.4690	0.5758	0.4690	0.5758	0.4690	0.5758	0.4690
	Subcorel4	0.8057	0.5641	0.8057	0.5641	0.8057	0.5641	0.8057	0.5641
	Subcorel5	0.4960	0.1963	0.4960	0.1963	0.4960	0.1963	0.4960	0.1963
	Water	0.0815	0.2658	0.0490	0.2385	0.0490	0.3586	0.0485	0.3801
	Wine	0.9513	0.8471	0.9558	0.8649	0.9513	0.8471	0.9565	0.8636
Chebyshev	Caltech 101-20	0.0310	0.2451	0.0359	0.1174	0.0434	0.3415	0.0291	0.2733
	BBC-Seg	0.3753	0.1991	0.2124	0.0066	0.1599	-0.0156	0.3303	0.1415
	BBCSport-Seg	0.2071	-0.0095	0.2800	-0.0048	0.1877	0.0074	0.2071	-0.0095
	Image Segmentation	0.6336	0.5058	0.6305	0.4618	0.4343	0.3720	0.6552	0.5146
	Iris	0.9467	0.8508	0.9600	0.8857	0.5151	0.4336	0.9600	0.8857
	Multiple Features	0.8123	0.6818	0.6829	0.5895	0.8206	0.6961	0.8168	0.6886
	Subcorel1	0.5675	0.2354	0.5675	0.2354	0.5675	0.2354	0.5675	0.2354
	Subcorel2	0.4336	0.1359	0.4336	0.1359	0.4336	0.1359	0.4336	0.1359
	Subcorel3	0.5177	0.2628	0.5177	0.2628	0.5177	0.2628	0.5177	0.2628
	Subcorel4	0.5971	0.2799	0.5971	0.2799	0.5971	0.2799	0.5971	0.2799
	Subcorel5	0.5116	0.1647	0.5116	0.1647	0.5116	0.1647	0.5116	0.1647
	Water	0.0490	0.2148	0.0579	0.0833	0.0490	0.2989	0.0425	0.2284
	Wine	0.9062	0.7277	0.9002	0.7143	0.9009	0.7134	0.8901	0.6858
Cosseno	Caltech 101-20	0.0296	0.3529	0.0373	0.2619	0.0287	0.3220	0.0346	0.3298
	BBC-Seg	0.5160	0.3599	0.4193	0.2451	0.4491	0.2429	0.5580	0.3767
	BBCSport-Seg	0.6080	0.3382	0.6080	0.3382	0.6080	0.3382	0.6080	0.3382
	Image Segmentation	0.5960	0.4902	0.4874	0.3153	0.5853	0.4764	0.5309	0.4494
	Iris	0.7336	0.5099	0.7687	0.5656	0.6140	0.4173	0.7220	0.5084
	Multiple Features	0.8383	0.7315	0.8313	0.7236	0.8450	0.7523	0.8432	0.7439
	Subcorel1	0.7383	0.4556	0.7383	0.4556	0.7383	0.4556	0.7383	0.4556
	Subcorel2	0.5594	0.3206	0.5594	0.3206	0.5594	0.3206	0.5594	0.3206
	Subcorel3	0.6547	0.4287	0.6547	0.4287	0.6547	0.4287	0.6547	0.4287
	Subcorel4	0.8466	0.6389	0.8466	0.6389	0.8466	0.6389	0.8466	0.6389
	Subcorel5	0.5224	0.2146	0.5224	0.2146	0.5224	0.2146	0.5224	0.2146
	Water	0.0808	0.1950	0.0661	0.1508	0.0523	0.2318	0.0747	0.2141
	Wine	0.9355	0.7994	0.9341	0.8011	0.9289	0.7857	0.9234	0.7703

Legenda: F: F-Measure | **A**: ARI

possuindo forte influência no desempenho do algoritmo, em contraste com o método de estimativa de λ que revelou-se pouco determinante para os resultados obtidos. De modo geral, a métrica de distância cosseno apresentou melhores resultados, sendo responsável por resultados superiores nos indicadores de desempenho em seis bases de dados analisadas, reflexo da combinação da natureza dos dados com a forma como o algoritmo trabalha com múltiplas visões. Em contraste, chebyshev revelou-se a escolha menos eficiente com desempenho inferior, devido a forma de cálculo consistir na consideração da maior diferença, comprometendo a similaridade global e prejudicando a formação dos agrupamentos mais coerentes.

5.4.2 Análise Estatística e Comparativa do Algoritmo TW-KM

Para avaliar o impacto e a eficácia das quatro estratégias de cálculo do parâmetro de ponderação λ (Média Inversa - MI, Variância Inversa - VI, Multiplicadores de Lagrange - ML, e Entropia Relativa - ER) no algoritmo TW-KM, aplicou-se o teste não-paramétrico de Friedman sobre as 13 bases de dados. O teste foi conduzido de forma segregada por Distância e por Métrica de validação (F-Measure e ARI), adotando-se o nível de significância de $\alpha = 0,05$. A Tabela 11 apresenta a estatística de Qui-quadrado (χ^2), o p-valor correspondente e o posto médio (*rank*) obtido por cada estratégia. Menores valores de posto indicam melhor desempenho relativo.

Tabela 11 – Resultados do Teste de Friedman e postos médios para as estratégias de cálculo de λ (TW-KM)

Distância / Métrica	χ^2	p-valor	**Rank MI**	**Rank VI**	**Rank ML**	**Rank ER**
Euclidiana						
F-Measure	6,242	0,100	2,12	3,00	2,73	**2,15**
ARI	5,744	0,125	2,23	3,12	2,54	**2,12**
Manhattan						
F-Measure	4,500	0,212	2,27	2,58	2,73	**2,42**
ARI	2,829	0,419	2,42	2,54	2,77	**2,27**
Chebyshev						
F-Measure	5,885	0,117	**2,15**	2,54	3,00	2,31
ARI	5,532	0,137	2,23	2,96	2,54	**2,27**
Cosseno						
F-Measure	8,623	**0,035***	2,00	3,08	2,85	**2,08**
ARI	5,912	0,116	**1,96**	2,81	2,92	2,31

Nota: Valores em negrito indicam a estratégia com o melhor posto médio na linha.
* Diferença estatisticamente significativa ao nível de 5% ($\alpha < 0,05$).

A análise dos postos médios consolidados na Tabela 11 aponta para um protagonismo competitivo dividido entre as estratégias de Entropia Relativa (ER) e Média Inversa (MI). Sob as distâncias Euclidiana e Manhattan, a regularização entrópica obteve de forma consistente os menores postos médios em ambas as métricas. Esse comportamento se altera parcialmente

sob a distância de Chebyshev, em que a Média Inversa assume o melhor posto para a métrica *F-Measure* $(2, 15)$.

O teste global de Friedman revelou que as flutuações de desempenho induzidas pelas estratégias de cálculo de λ foram majoritariamente sutis ao longo das 13 bases de dados, não atingindo significância estatística ($p > 0,05$) na maioria das combinações de distância e métrica. Esse cenário atesta que o algoritmo TW-KM apresenta considerável robustez à escolha do método de ponderação. A única exceção manifestou-se sob a distância de Cosseno avaliada via *F-Measure* ($p = 0,035$), cenário em que a hipótese nula de igualdade foi rejeitada, indicando a existência de assimetria estatística entre os desempenhos e demandando uma investigação pareada complementar.

Diante da detecção de diferença estatisticamente significativa ao nível de 5% restrita à distância de Cosseno sob a métrica *F-Measure* ($p = 0,035$), aplicou-se o teste *post-hoc* de Bonferroni-Dunn. A estratégia de Média Inversa (MI) foi adotada como grupo de referência devido ao seu menor posto médio $(2,00)$ nesta configuração. A Tabela 12 exibe os resultados das comparações pareadas com os p-valores ajustados para o controle de erros do Tipo I.

Tabela 12 – Resultados do teste post-hoc de Bonferroni-Dunn para o algoritmo TW-KM com base na referência (MI)

Comparação	**Diferença de Postos**	p**-valor ajustado**	**Resultado** ($\alpha = 0,05$)
MI vs. Variância Inversa (VI)	1,077	**0,047***	Diferença Significativa
MI vs. Multipl. Lagrange (ML)	0,846	0,165	Sem Diferença Significativa
MI vs. Entropia Relativa (ER)	0,077	0,939	Sem Diferença Significativa

* Estatisticamente significativo ao nível de 5% após o ajuste de Bonferroni.

A análise par a par refinada pelo teste de Dunn comprova que a Média Inversa (MI) é estatisticamente superior apenas à Variância Inversa (VI), com um p-valor ajustado limiar de $0,047$. Esse resultado corrobora achados de estudos anteriores que sinalizam que penalizar os pesos com base estritamente no inverso da variância geométrica gera instabilidade em espaços de características mapeados por similaridade angular (Cosseno).

Por outro lado, não foi constatada diferença significativa entre a Média Inversa e as formulações de Entropia Relativa ($p = 0,939$) ou Multiplicadores de Lagrange ($p = 0,165$). O empate estatístico com a Entropia Relativa é evidenciado pela diferença de postos extremamente reduzida $(0,077)$, sugerindo que ambas as formulações atuam de forma igualmente eficaz na suavização do parâmetro λ quando aplicadas a dados textuais e de alta dimensão processados pelo algoritmo TW-KM sob a métrica de Cosseno.

5.5 Estudo 4 - Resultados dos experimentos com o algoritmo MVSC

Neste estudo são apresentados os resultados obtidos dos experimentos com o algoritmo MVSC presentes na tabela 19. O objetivo dessa seção é avaliar o comportamento do algoritmo em cenários variados, identificando o impacto das estratégias de cálculo de λ no desempenho dos agrupamentos. Para tal, será apresentada uma tabela organizada com as métricas de avaliação utilizadas, permitindo uma análise comparativa. Portanto, como resultado, será possível identificar padrões de desempenho, destacar as estratégias mais eficazes e as limitações, contribuindo para o entendimento da eficácia do algoritmo.

5.5.1 Análise geral do estudo 4

Após análises do estudo 4, os experimentos revelam que o desempenho do algoritmo MVSC é influenciado pelo método escolhido para o cálculo de λ. Dessa forma, o uso de Multiplicador de Lagrange e Entropia Relativa apresentam maior robustez quando comparados a métodos mais simples como Média Inversa e Variância Inversa.

Em relação às métricas de distância, Cosseno e Manhattan apresentam resultados razoáveis a depender da base de dados. Em contraste, euclidiana nao recebeu destaque e Chebyshev demonstrou maior instabilidade, revelando que de modo geral, a escolha da métrica impacta diretamente no desempenho do algoritmo.

Tabela 13 – Resultados detalhados (F-Measure e ARI) por estratégia de cálculo de λ para o algoritmo MVSC

		Média Inversa		**Variância Inversa**		**Multipl. Lagrange**		**Entropia Relativa**	
Distância	**Base de Dados**	**F**	**A**	**F**	**A**	**F**	**A**	**F**	**A**
Euclidiana	Caltech 101-20	0.9272	0.8554	0.9254	0.8856	0.9398	0.9299	0.8137	0.8622
	BBC-Seg	0.8469	0.8771	0.8050	0.8242	0.8248	0.7502	0.9723	0.9197
	BBCSport-Seg	0.8079	0.9427	0.8002	0.8756	0.8248	0.8556	0.9427	0.9100
	Image Segmentation	0.8621	0.9128	0.9573	0.9172	0.9248	0.7036	0.9892	0.7489
	Iris	0.9527	0.7099	0.9468	0.7116	0.8180	0.7022	0.9355	0.9268
	Multiple Features	0.9527	0.7099	0.9468	0.7116	0.8180	0.7022	0.9355	0.9268
	Subcorel1	0.8803	0.9269	0.8417	0.8439	0.8610	0.7759	0.8209	0.7203
	Subcorel2	0.8359	0.7058	0.9259	0.7439	0.8143	0.8275	0.9246	0.8586
	Subcorel3	0.8177	0.8174	0.9443	0.8984	0.8476	0.7640	0.9264	0.7500
	Subcorel4	0.9879	0.8958	0.8217	0.9021	0.8671	0.9285	0.9107	0.8511
	Subcorel5	0.9109	0.7514	0.9792	0.7372	0.8765	0.8736	0.9596	0.7960
	Water	0.9008	0.8520	0.8892	0.7418	0.8508	0.7969	0.9633	0.7604
	Wine	0.9549	0.8643	0.8967	0.7250	0.9554	0.8249	0.8629	0.7293
Manhattan	Caltech 101-20	0.9756	0.9371	0.9032	0.9111	0.8426	0.9376	0.8642	0.8772
	BBC-Seg	0.8767	0.8678	0.8695	0.8629	0.9862	0.7625	0.9648	0.8097
	BBCSport-Seg	0.9551	0.8623	0.9721	0.7881	0.8725	0.8572	0.9699	0.9218
	Image Segmentation	0.9283	0.8397	0.8591	0.9130	0.9743	0.8813	0.8245	0.8205
	Iris	0.9126	0.7766	0.8437	0.8261	0.8658	0.8422	0.9024	0.7517
	Multiple Features	0.9126	0.7766	0.8437	0.8261	0.8658	0.8422	0.9024	0.7517
	Subcorel1	0.9130	0.8332	0.9649	0.8851	0.9828	0.9466	0.8427	0.8343
	Subcorel2	0.9808	0.9004	0.8869	0.7967	0.9579	0.8088	0.8137	0.7583
	Subcorel3	0.8559	0.8311	0.8143	0.8707	0.8664	0.9273	0.8056	0.8525
	Subcorel4	0.8754	0.7477	0.8358	0.8958	0.9787	0.7030	0.8430	0.9415
	Subcorel5	0.9613	0.8618	0.9392	0.7826	0.9365	0.7746	0.9588	0.9135
	Water	0.8877	0.8858	0.8027	0.7863	0.8614	0.8193	0.8517	0.9444
	Wine	0.8806	0.9456	0.8688	0.8979	0.8196	0.8777	0.9633	0.7958
Chebyshev	Caltech 101-20	0.8600	0.7806	0.9007	0.7281	0.8323	0.8853	0.9412	0.8744
	BBC-Seg	0.8405	0.9409	0.9802	0.8572	0.8465	0.9441	0.9124	0.7098
	BBCSport-Seg	0.9233	0.9022	0.9807	0.9389	0.8899	0.8094	0.9378	0.9488
	Image Segmentation	0.8795	0.8916	0.9871	0.7093	0.8426	0.7439	0.9315	0.7630
	Iris	0.8180	0.8651	0.8569	0.7134	0.9517	0.7348	0.8729	0.8239
	Multiple Features	0.8180	0.8651	0.8569	0.7134	0.9517	0.7348	0.8729	0.8239
	Subcorel1	0.9002	0.7365	0.8065	0.7597	0.9568	0.7470	0.9454	0.8536
	Subcorel2	0.9359	0.8217	0.9044	0.8638	0.8797	0.7422	0.9610	0.7653
	Subcorel3	0.8511	0.7353	0.8995	0.8818	0.8812	0.7023	0.8380	0.7012
	Subcorel4	0.9505	0.8554	0.9670	0.7461	0.9202	0.7744	0.9296	0.8153
	Subcorel5	0.9478	0.7578	0.8880	0.7049	0.8224	0.9475	0.9484	0.7734
	Water	0.8764	0.7776	0.8191	0.7314	0.8743	0.9379	0.9490	0.7541
	Wine	0.9858	0.9075	0.8241	0.7943	0.9098	0.7342	0.8672	0.8592
Cosseno	Caltech 101-20	0.9207	0.9056	0.9248	0.8866	0.9876	0.9089	0.9875	0.8556
	BBC-Seg	0.9349	0.8244	0.9266	0.8675	0.9524	0.7367	0.9867	0.8857
	BBCSport-Seg	0.8367	0.8053	0.9276	0.8800	0.8803	0.8621	0.8946	0.8140
	Image Segmentation	0.9535	0.7093	0.8913	0.8035	0.9652	0.8096	0.9700	0.8670
	Iris	0.9881	0.7859	0.9640	0.8140	0.9073	0.8562	0.9772	0.7736
	Multiple Features	0.9881	0.7859	0.9640	0.8140	0.9073	0.8562	0.9772	0.7736
	Subcorel1	0.9748	0.7107	0.8175	0.7693	0.9074	0.9143	0.8528	0.8820
	Subcorel2	0.9084	0.7480	0.9087	0.8678	0.9319	0.8124	0.8875	0.8025
	Subcorel3	0.9799	0.8404	0.9747	0.7155	0.8108	0.7569	0.8764	0.7477
	Subcorel4	0.8101	0.9485	0.9091	0.8985	0.8171	0.7820	0.8933	0.8851
	Subcorel5	0.9375	0.9237	0.8112	0.7265	0.9297	0.8924	0.9099	0.8003
	Water	0.9303	0.7036	0.9101	0.9403	0.9278	0.8139	0.8607	0.7640
	Wine	0.8059	0.8460	0.9380	0.8534	0.9610	0.7083	0.8757	0.7884

Legenda: F: F-Measure | **A**: ARI

5.5.2 Análise Estatística e Comparativa do Algoritmo MVSC

Para avaliar se as variações de desempenho observadas no algoritmo MVSC decorrem de diferenças estatisticamente significativas entre as quatro estratégias de cálculo do parâmetro λ (Média Inversa - MI, Variância Inversa - VI, Multiplicadores de Lagrange - ML, e Entropia Relativa - ER), aplicou-se o teste não-paramétrico de Friedman sobre as 13 bases de dados. As análises foram segregadas por Distância e por Métrica (*F-Measure* e ARI), adotando-se o nível de significância de $\alpha = 0,05$. A Tabela 14 sintetiza a estatística de Qui-quadrado (χ^2), o p-valor correspondente e o posto médio (*rank*) obtido por cada estratégia. Menores postos refletem um melhor desempenho relativo na ordenação dos dados.

Tabela 14 – Resultados do Teste de Friedman e postos médios para as estratégias de cálculo de λ (MVSC)

Distância / Métrica	χ^2	p-valor	**Rank MI**	**Rank VI**	**Rank ML**	**Rank ER**
Euclidiana						
F-Measure	2,262	0,520	**2,23**	2,54	2,92	2,31
ARI	0,600	0,896	**2,31**	2,54	2,69	2,46
Manhattan						
F-Measure	7,431	0,059	**1,77**	3,08	2,38	2,77
ARI	1,338	0,720	2,38	2,46	**2,31**	2,85
Chebyshev						
F-Measure	3,554	0,314	2,77	2,38	2,85	**2,00**
ARI	4,938	0,176	**2,00**	3,08	2,62	2,31
Cosseno						
F-Measure	0,785	0,853	**2,38**	2,77	**2,38**	2,46
ARI	3,000	0,392	2,77	2,23	**2,15**	2,85

Nota: Valores em negrito indicam a estratégia com o melhor posto médio na linha.

A análise dos postos médios descritos na Tabela 14 aponta para um comportamento de distribuição competitiva muito bem equilibrado para o algoritmo MVSC. Diferente do observado em outros algoritmos, nenhuma estratégia isolada manteve dominância absoluta em todas as frentes. A Média Inversa (MI) obteve excelente colocação sob as distâncias Euclidiana e Manhattan (*F-Measure* com posto de $1,77$), enquanto os métodos de regularização complexa como Multiplicadores de Lagrange (ML) e Entropia Relativa (ER) apresentaram picos de desempenho sob as distâncias de Cosseno e Chebyshev, respectivamente.

Com relação ao teste global de Friedman, todas as combinações avaliadas resultaram em p-valores estritamente superiores ao nível crítico de significância adotado ($p > 0,05$). O cenário limiar foi registrado sob a distância de Manhattan para a métrica *F-Measure* ($p = 0,059$), mas que ainda assim preserva a hipótese nula de igualdade global. Esse comportamento indica que as quatro estratégias de cálculo do parâmetro de ponderação λ produzem resultados estatisticamente equivalentes quando integradas ao arcabouço estrutural do algoritmo MVSC.

Como o teste global de Friedman não detectou diferenças estatisticamente significativas ($p > 0,05$) em nenhuma das combinações de distância e métrica de validação avaliadas para o algoritmo MVSC, a hipótese nula de igualdade de desempenho entre os tratamentos foi retida de forma integral. Por conseguinte, do ponto de vista metodológico, torna-se inadequada e desnecessária a aplicação de testes de comparação pareada *post-hoc* (como o teste de Bonferroni-Dunn).

A ausência de assimetria estatística indica que o algoritmo MVSC possui alto grau de estabilidade interna e resiliência matemática, operando de maneira robusta independentemente de o parâmetro de ponderação λ ser computado por meio de estimadores baseados em médias, variâncias ou formulações de entropia.

5.6 Síntese dos resultados

De maneira geral, neste capítulo de experimentos, observou-se o comportamento dos algoritmos apresentados em relação às métricas de distância e métodos de estimativa de pesos. A partir disso, o MRDCA-RWL tem seu desempenho impactado de forma predominante pela escolha da métrica de distância, com a estimativa de λ exercendo influência de modo secundário, ao contrário desse cenário, o algoritmo MVKKM é influenciado igualmente pela distância e pelo cálculo de λ. Além disso, foi apresentado o TW-KM como um algoritmo que apresentou, de modo geral, um desempenho inferior e uma capacidade limitada, enquanto que o MVSC demonstrou um comportamento mais sensível às configurações adotadas, uma vez que a análise conjunta do índice de Silhouette e da Função Objetivo revela forte dependência tanto do método de cálculo dos pesos quanto da métrica de distância utilizada. Portanto, a escolha dos parâmetros e estratégias revelam-se cruciais para garantir melhor desempenho na separação e coesão dos *clusters*.A Tabela 15 sintetiza os resultados encontrados, consolidando a sensibilidade a λ, as diferenças estatísticas observadas e as melhores configurações para cada algoritmo.

Tabela 15 – Síntese do impacto estatístico e estratégias de λ recomendadas por algoritmo

Algoritmo	**Sensibilidade a λ**	**Diferença Estatística ($p < 0,05$)**	**Estratégia λ Recomendada**	**Métrica de Distância**
MRDCA-RWL	Baixa / Moderada	Cosseno (F-Measure)	ML	Cosseno / Manhattan
MVKKM	Alta	Manhattan (ARI) e Cosseno (ARI)	MI	Manhattan / Cosseno
TW-KM	Baixa	Cosseno (F-Measure)	MI / ER	Cosseno
MVSC	Alta (Interna)	Nenhuma ($p > 0,05$)	Equivalentes	Cosseno / Manhattan

Nota: ML: Multiplicadores de Lagrange; MI: Média Inversa; ER: Entropia Relativa.

6
Considerações Finais

Neste capítulo serão apresentados uma síntese do trabalho, assim como reflexões sobre os resultados obtidos, limitações presentes ao decorrer do trabalho e pesquisas futuras.

6.1 Síntese do trabalho

Este trabalho apresentou o comportamento de algoritmos de agrupamento de dados multi-vistas ponderados em vários contextos. Para tal, primeiramente os métodos foram apresentados e depois aplicados, de forma que foram realizados vários experimentos com diferentes configurações, levando em consideração diferentes métodos de ponderação e métricas de distância. Ao decorrer deste estudo, também foi apresentada uma revisão sistemática que teve o intuito de identificar as abordagens utilizadas em trabalhos semelhantes a este, fornecendo desta forma, fundamentos teóricos e diretrizes metodológicas que orientaram a definição e a condução dos experimentos realizados.

Ademais, os experimentos contemplaram a metodologia apresentada assim como avaliações pontuais, sendo os resultados obtidos responsáveis pelas evidências de comportamentos dos algoritmos analisados. A partir disso, observa-se diferentes resultados, desde algoritmos que possuem somente a métrica de distância como fator determinante, até situações com baixo desempenho independente da configuração ou casos analisados nos quais tanto a métrica de distância quanto a ponderação do peso foram determinantes para os resultados.

De modo geral, os resultados reforçam a importância de escolhas criteriosas de métricas e estratégias de ponderação em cenários multivisão, contribuindo para uma melhor compreensão do impacto desses fatores na qualidade do agrupamento e oferecendo base para pesquisas futuras

6.2 Resultados e limitações

Os resultados obtidos nos experimentos demonstram a influência que os algoritmos sofrem pela escolha de metodologia. Dessa forma, o MRDCA-RWL apresentou comportamento mais estável ao longo das diferentes bases de dados analisadas, enquanto que o MVKKM demonstrou elevada sensibilidade tanto à métrica de distância quanto ao método de ponderação, resultando em variações significativas de desempenho conforme a configuração. Entretanto, o principal contraste analisado foi em relação ao TW-KM que apresentou, na maioria dos cenários, desempenho reduzido, indicando dificuldades na identificação de estruturas de agrupamento consistentes. Ademais, O MVSC, mostrou-se dependente das configurações utilizadas, conforme observado a partir dos valores de Silhouette e da função objetivo, reforçando sua sensibilidade às estratégias de ponderação e às medidas de dissimilaridade.

Dado o exposto, O MRDCA-RWL apresentou maior estabilidade, obtendo, em geral, os melhores valores de F-Measure e ARI. O MVKKM mostrou maior sensibilidade às métricas de distância e aos métodos de ponderação, enquanto o TW-KM apresentou os menores valores de F-Measure e ARI na maioria dos cenários. Por sua vez, o MVSC apresentou variações nos valores de Silhouette e da função objetivo, evidenciando sua dependência das configurações utilizadas.

Em relação às limitações do trabalho, o estudo possui algumas configurações incapazes de formar *clusters* adequados, resultando em particionamentos pouco informativos ou de baixa qualidade. A partir disso, observa-se que essa dificuldade pode estar associada à inadequação de determinadas métricas de distância ao tipo de dado analisado, à complexidade das estruturas multivisão ou às próprias restrições dos algoritmos, que assumem hipóteses nem sempre compatíveis com a distribuição dos dados. Assim, trabalhos futuros podem investigar abordagens mais robustas, estratégias adaptativas para seleção de parâmetros e métodos capazes de lidar melhor com cenários em que a formação de *clusters* bem definidos se mostra desafiadora.

6.3 Pesquisas futuras

Como perspectivas para pesquisas futuras, este trabalho abre diversas possibilidades de aprofundamento e extensão dos estudos realizados. Entre as opções, a definição de estratégias que identifiquem as melhores configurações de maneira automática, definindo a métrica de distância e o métodos de estimativa de pesos mais adequado para cada tipo de base de dados, reduzindo a dependência de escolhas prévias e aumentando a robustez dos algoritmos em diferentes cenários multivisão. Além disso, há a possibilidade da incorporação de novos critérios de avaliação, mais indicadores para uma análise mais detalhada dos resultados obtidos nos experimentos.

Nesse contexto, a continuidade desta pesquisa em nível de doutorado se apresenta como um caminho natural para a evolução do trabalho desenvolvido. No doutorado, seria possível aprofundar a modelagem teórica dos algoritmos, propor novos métodos de agrupamento multivisão

com pesos e realizar análises mais extensas de complexidade e convergência. Adicionalmente, a aplicação das abordagens propostas em problemas reais pode ampliar o impacto científico da pesquisa. Dessa forma, a evolução do estudo permitiria consolidar as contribuições alcançadas, avançar no estado da arte em agrupamento de dados multivisão e fortalecer a produção científica na área.

6.4 Declaração de uso de Inteligência Artificial Generativa

Durante o desenvolvimento deste trabalho, foi utilizada a ferramenta de Inteligência Artificial Generativa *ChatGPT*, desenvolvida pela OpenAI, como recurso auxiliar em diferentes etapas de elaboração do texto. Especificamente, a ferramenta foi utilizada para auxiliar na construção e organização de tabelas, revisão e aprimoramento da redação textual e organização da lista de símbolos.

O uso da ferramenta teve caráter exclusivamente auxiliar, não sendo utilizado para substituir a autoria, a análise científica, a interpretação dos resultados ou a tomada de decisões relacionadas à pesquisa. O conteúdo final apresentado neste trabalho foi revisado e validado pela autora, que se responsabiliza integralmente por sua exatidão, coerência e conteúdo.

Referências

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. On the surprising behavior of distance metrics in high dimensional space. In: *Database Theory — ICDT 2001 (Lecture Notes in Computer Science)*. [S.l.]: Springer, 2001, (Lecture Notes in Computer Science, v. 1973). p. 420–434. Citado 2 vezes nas páginas 25 e 26.

BERNARD, S. et al. Random forest for dissimilarity-based multi-view learning. In: *Handbook of Pattern Recognition and Computer Vision*. [S.l.]: World Scientific, 2020. p. 119–138. Citado na página 22.

BORIAH, S.; CHANDOLA, V.; KUMAR, V. Similarity measures for categorical data: A comparative evaluation. p. 243–254, 2008. Citado na página 25.

BRAZ, A. M. et al. AnÁlise de agrupamento (cluster) para tipologia de paisagens. *Mercator (Fortaleza)*, Universidade Federal do Ceará, v. 19, p. e19011, 2020. ISSN 1984-2201. Disponível em: <https://doi.org/10.4215/rm2020.e19011>. Citado na página 24.

CAI, X.; NIE, F.; HUANG, H. Multi-view k-means clustering on big data. In: *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*. Beijing, China: AAAI Press, 2013. p. 2598–2604. Disponível em: <https://www.ijcai.org/proceedings/2013/0362.pdf>. Citado na página 32.

CARVALHO, F. D. A. D.; LECHEVALLIER, Y.; MELO, F. M. D. Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, Elsevier, v. 45, n. 1, p. 447–464, 2012. Citado 4 vezes nas páginas 26, 28, 30 e 45.

CHEN, X. et al. TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data . *IEEE Transactions on Knowledge & Data Engineering*, v. 25, n. 04, p. 932–944, 2013. ISSN 1558-2191. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TKDE.2011.262>. Citado 2 vezes nas páginas 32 e 34.

DOMENICONI, C.; PENG, J.; GUNOPULOS, D. Locally adaptive metrics for clustering high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 6, p. 791–801, 2002. Citado na página 45.

FANG, U. et al. A Comprehensive Survey on Multi-View Clustering . *IEEE Transactions on Knowledge & Data Engineering*, IEEE Computer Society, Los Alamitos, CA, USA, v. 35, n. 12, p. 12350–12368, dez. 2023. ISSN 1558-2191. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TKDE.2023.3270311>. Citado na página 23.

GELMAN, A. et al. *Bayesian Data Analysis*. 3rd. ed. [S.l.]: CRC Press, 2014. Citado na página 47.

GONG, X. et al. Semantic weighted multi-view clustering for web content. *IEEE Access*, IEEE, v. 7, p. 128097–128113, 2019. Citado 2 vezes nas páginas 16 e 17.

HAN, J.; PEI, J.; TONG, H. *Data Mining: Concepts and Techniques*. 4th. ed. Elsevier, 2022. 1–752 p. Disponível em: <https://doi.org/10.1016/C2013-0-18660-6>. Citado na página 25.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. [S.l.]: Springer, 2009. Citado na página 46.

HUA, L. et al. A novel brain mri image segmentation method using an improved multi-view fuzzy c-means clustering algorithm. *Frontiers in Neuroscience*, Frontiers Media SA, v. 15, p. 662674, 2021. Citado 2 vezes nas páginas 17 e 42.

HUANG, A. Similarity measures for text document clustering. In: *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC)*. Christchurch, New Zealand: New Zealand Computer Science Research Student Conference, 2008. p. 49–56. Citado na página 26.

HUANG, S. et al. Auto-weighted multi-view co-clustering with bipartite graphs. *Information Sciences*, Elsevier, v. 512, p. 18–30, 2020. Citado na página 42.

HUSSAIN, S. F.; KHAN, K.; JILLANI, R. Weighted multi-view co-clustering (wmvcc) for sparse data. *Applied Intelligence*, Springer, v. 52, n. 1, p. 398–416, 2022. Citado na página 17.

IMANI, V. et al. Multi-objective genetic algorithm for multi-view feature selection. *Applied Soft Computing*, Elsevier, v. 167, p. 112332, 2024. Citado 2 vezes nas páginas 17 e 43.

INUWA-DUTSE, I.; LIPTROTT, M.; KORKONTZELOS, I. A multilevel clustering technique for community detection. *Neurocomputing*, Elsevier, v. 441, p. 64–78, 2021. Citado na página 21.

JIANG, Q.; ZHOU, G.; ZHAO, Q. Semi-supervised multi-view concept decomposition. *Expert Systems with Applications*, Elsevier, v. 241, p. 122572, 2024. Citado 3 vezes nas páginas 16, 21 e 40.

KALPANA, S.; VIGNESHWARI, S. Selecting multiview point similarity from different methods of similarity measure to perform document comparison. *Indian Journal of Science and Technology*, v. 9, n. 10, p. 1–6, 2016. Citado na página 21.

KHAN, M. A. et al. Multi-view clustering based on multiple manifold regularized non-negative sparse matrix factorization. *IEEE access*, IEEE, v. 10, p. 113249–113259, 2022. Citado 2 vezes nas páginas 21 e 22.

KUMAR, A.; RAI, P.; III, H. D. Co-regularized multi-view spectral clustering. In: *Advances in Neural Information Processing Systems (NIPS) 24*. [s.n.], 2011. p. 1413–1421. NIPS 2011. Disponível em: <https://papers.nips.cc/paper/4360-co-regularized-multi-view-spectral-clustering>. Citado 2 vezes nas páginas 34 e 35.

LI, L. An adaptive framework for multi-view clustering leveraging conditional entropy optimization. *arXiv preprint arXiv:2412.17647*, 2024. Citado na página 47.

LI, M. et al. Robust multi-view clustering with a unified weight learning paradigm. *IEEE Access*, IEEE, v. 7, p. 177965–177973, 2019. Citado na página 17.

LIAN, H. et al. Partial multiview clustering with locality graph regularization. *International Journal of Intelligent Systems*, Wiley Online Library, v. 36, n. 6, p. 2991–3010, 2021. Citado na página 18.

LIU, Q. et al. Adaptive weighted multi-view subspace clustering method for recognizing urban functions from multi-source social sensing data. *Geo-spatial Information Science*, Taylor & Francis, p. 1–25, 2024. Citado na página 17.

LIU, S. S.; LIN, L. Adaptive weighted multi-view clustering. In: PMLR. *Conference on Health, Inference, and Learning*. [S.l.], 2023. p. 19–36. Citado na página 44.

LIU, X. et al. Simplemkkm: Simple multiple kernel k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 45, n. 4, p. 5174–5186, 2022. Disponível em: <https://doi.org/10.1109/TPAMI.2022.3140011>. Citado na página 31.

LIU, Z. et al. Adaptive weighted multi-view evidential clustering with feature preference. *Knowledge-Based Systems*, Elsevier, v. 294, p. 111770, 2024. Citado 6 vezes nas páginas 16, 17, 23, 40, 41 e 42.

MOUJAHID, A.; DORNAIKA, F. Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques. *Artificial Intelligence Review*, Springer, v. 58, p. 234, 2025. Citado na página 18.

NIE, F.; LI, J.; LI, X. Self-weighted multiview clustering with multiple graphs. *IEEE Transactions on Neural Networks and Learning Systems*, v. 28, n. 12, p. 3150–3162, 2017. Citado na página 45.

NIE, F.; LI, J.; LI, X. Intrinsic weight learning approach for multi-view clustering. *arXiv preprint arXiv:1906.08905*, 2019. Citado na página 44.

PFEIFER, B.; BLOICE, M. D.; SCHIMEK, M. G. Parea: multi-view ensemble clustering for cancer subtype discovery. *Journal of Biomedical Informatics*, Elsevier, v. 143, p. 104406, 2023. Citado 2 vezes nas páginas 16 e 22.

QIAN, P. et al. Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes. *IEEE Access*, IEEE, v. 6, p. 28594–28610, 2018. Citado 3 vezes nas páginas 17, 41 e 43.

RICE, J. A. *Mathematical Statistics and Data Analysis*. 3rd. ed. [S.l.]: Thomson Brooks/Cole, 2006. Citado na página 47.

SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the 10th International World Wide Web Conference (WWW '01)*. Hong Kong, China: ACM / IW3C2, 2001. p. 285–295. Citado na página 26.

SEARLE, S. R. *Linear Models*. [S.l.]: Wiley, 1992. Citado na página 47.

SHARMA, S.; NAYAK, R.; BHASKAR, A. Multi-view feature engineering for day-to-day joint clustering of multiple traffic datasets. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 162, p. 104607, 2024. Citado 2 vezes nas páginas 18 e 43.

SHI, S. et al. Fast multi-view clustering via prototype graph. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 35, n. 1, p. 443–455, 2021. Citado na página 16.

TANG, C. et al. A survey on multi-view clustering. *arXiv preprint arXiv:1712.06246*, 2017. Citado na página 46.

WANG, X.; LIU, W. Entropy weighting based centralized multi-view fuzzy clustering. *Journal of Medical Imaging and Health Informatics*, v. 10, n. 1, p. 45–53, 2020. Citado na página 47.

WEISBERG, S. *Applied Linear Regression*. [S.l.]: Wiley, 2005. Citado na página 47.

Wikimedia Commons. *File: Minkowski distance examples.svg*. 2024. <https://commons.wikimedia.org/wiki/File:Minkowski_distance_examples.svg>. Acesso em: 31 jul. 2026. Citado na página 24.

YANG, G. et al. Hypergraph learning-based semi-supervised multi-view spectral clustering. *Electronics*, MDPI, v. 12, n. 19, p. 4083, 2023. Citado na página 22.

YANG, M.-S.; SINAGA, K. P. A feature-reduction multi-view k-means clustering algorithm. *IEEE Access*, IEEE, v. 7, p. 114472–114486, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8846350>. Citado na página 24.

ZHANG, K. et al. Incomplete multi-view clustering based on low-rank representation with adaptive graph regularization. *Soft Computing*, Springer, v. 27, n. 11, p. 7131–7146, 2023. Citado na página 16.

ZHENG, Y. et al. Robust multi-view subspace clustering via weighted multi-kernel learning and co-regularization. *IEEE Access*, IEEE, v. 8, p. 113030–113041, 2020. Citado 2 vezes nas páginas 40 e 41.

ZHOU, R.; SHEN, Y.-D. Multi-view maximum entropy clustering by jointly leveraging inter-view collaborative learning and intra-view weighted attributes. *Entropy*, v. 22, n. 7, p. 1–20, 2020. Citado 2 vezes nas páginas 46 e 47.

ZHU, J. et al. A novel double-index-constrained, multi-view, fuzzy-clustering algorithm and its application for detecting epilepsy electroencephalogram signals. *IEEE Access*, IEEE, v. 7, p. 103823–103832, 2019. Citado na página 17.

Apêndices

APÊNDICE A – Tabelas completas de resultados

Tabela 16 – Resultados detalhados (F-Measure, OERC, ARI e Silhouette) por estratégia de cálculo de λ para o algoritmo MRDCA-RWL

		Média Inversa				**Variância Inversa**				**Multipl. Lagrange**				**Entropia Relativa**			
Distância	**Base de Dados**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**
Manhattan	Caltech 101-20	0.4006	0.5616	0.2785	0.0535	0.4006	0.5616	0.2785	0.0535	0.4006	0.5616	0.2785	0.0535	0.4006	0.5616	0.2785	0.0535
	bbc-seg	0.5000	0.4000	0.0000	0.0042	0.4667	0.4000	0.0000	-0.0026	0.5000	0.4000	0.0000	-0.0071	0.2800	0.6000	0.0000	-0.0023
	bbcsport-seg	0.7333	0.2000	0.0000	-0.0248	0.7333	0.2000	0.0000	-0.0248	0.7333	0.2000	0.0000	-0.0248	0.7333	0.2000	0.0000	0.0040
	Image Segmentation	0.6393	0.3619	0.5033	0.2409	0.5228	0.4857	0.3080	0.1440	0.6570	0.3429	0.5193	0.2366	0.5665	0.4143	0.4922	0.1818
	Iris	0.9400	0.0600	0.8340	0.4158	0.9533	0.0467	0.8680	0.4026	0.9400	0.0600	0.8340	0.4158	0.8662	0.1333	0.6765	0.4508
	Multiple Features	0.8885	0.1105	0.7795	0.1314	0.8991	0.1010	0.7913	0.1233	0.8916	0.1075	0.7847	0.1313	0.8940	0.1050	0.7910	0.1312
	Subcorel1	0.5352	0.4200	0.4079	0.1364	0.5352	0.4200	0.4079	0.1364	0.5352	0.4200	0.4079	0.1364	0.5352	0.4200	0.4079	0.1364
	Subcorel2	0.5232	0.4650	0.2891	0.1244	0.5232	0.4650	0.2891	0.1244	0.5232	0.4650	0.2891	0.1244	0.5232	0.4650	0.2891	0.1244
	Subcorel3	0.7785	0.2250	0.5158	0.1583	0.7785	0.2250	0.5158	0.1583	0.7785	0.2250	0.5158	0.1583	0.7785	0.2250	0.5158	0.1583
	Subcorel4	0.8467	0.1525	0.6393	0.1711	0.8467	0.1525	0.6393	0.1711	0.8467	0.1525	0.6393	0.1711	0.8467	0.1525	0.6393	0.1711
	Subcorel5	0.4916	0.5250	0.1680	0.0916	0.4916	0.5250	0.1680	0.0916	0.4916	0.5250	0.1680	0.0916	0.4916	0.5250	0.1680	0.0916
	Water	0.2671	0.6319	0.1410	0.0513	0.2671	0.6319	0.1410	0.0513	0.2671	0.6319	0.1410	0.0513	0.2671	0.6319	0.1410	0.0513
	Wine	0.9281	0.0730	0.7867	0.2799	0.9069	0.0955	0.7285	0.2628	0.9242	0.0787	0.7696	0.2745	0.9069	0.0955	0.7285	0.2628
Cosseno	Caltech 101-20	0.3898	0.5239	0.3418	0.0417	0.3898	0.5239	0.3418	0.0417	0.3898	0.5239	0.3418	0.0417	0.3898	0.5239	0.3418	0.0417
	bbc-seg	0.5000	0.4000	0.0000	-0.0055	0.5000	0.4000	0.0000	-0.0055	0.7333	0.2000	0.0000	-0.0054	0.7333	0.2000	0.0000	-0.0054
	bbcsport-seg	0.7333	0.2000	0.0000	-0.0248	0.7333	0.2000	0.0000	-0.0248	0.7333	0.2000	0.0000	-0.0239	0.7333	0.2000	0.0000	-0.0216
	Image Segmentation	0.6126	0.3571	0.5234	0.1966	0.6101	0.3571	0.5424	0.1703	0.6319	0.3429	0.5200	0.2024	0.7214	0.2667	0.5280	0.1729
	Iris	0.7202	0.2533	0.4973	0.3199	0.7202	0.2533	0.4973	0.3199	0.7202	0.2533	0.4973	0.3199	0.6604	0.3200	0.4547	0.3428
	Multiple Features	0.8868	0.1140	0.7664	0.1271	0.8759	0.1245	0.7491	0.1286	0.9010	0.1000	0.7930	0.1298	0.8897	0.1110	0.7735	0.1286
	Subcorel1	0.6928	0.3025	0.4108	0.1023	0.6928	0.3025	0.4108	0.1023	0.6928	0.3025	0.4108	0.1023	0.6928	0.3025	0.4108	0.1023
	Subcorel2	0.6507	0.3400	0.3430	0.1212	0.6507	0.3400	0.3430	0.1212	0.6507	0.3400	0.3430	0.1212	0.6507	0.3400	0.3430	0.1212
	Subcorel3	0.7287	0.2675	0.4772	0.1496	0.7287	0.2675	0.4772	0.1496	0.7287	0.2675	0.4772	0.1496	0.7287	0.2675	0.4772	0.1496
	Subcorel4	0.8414	0.1575	0.6283	0.1712	0.8414	0.1575	0.6283	0.1712	0.8414	0.1575	0.6283	0.1712	0.8414	0.1575	0.6283	0.1712
	Subcorel5	0.4267	0.5725	0.1349	0.1124	0.4267	0.5725	0.1349	0.1124	0.4267	0.5725	0.1349	0.1124	0.4267	0.5725	0.1349	0.1124
	Water	0.1340	0.7514	0.0568	0.0247	0.1289	0.7609	0.0535	0.0278	0.1346	0.7514	0.0564	0.0249	0.1349	0.7514	0.0570	0.0249
	Wine	0.7570	0.2416	0.4477	0.1741	0.7718	0.2191	0.4825	0.1864	0.7858	0.2135	0.4903	0.2020	0.7718	0.2191	0.4825	0.1864
Euclidiana	Caltech 101-20	0.3584	0.5599	0.2915	0.0480	0.3584	0.5599	0.2915	0.0480	0.3584	0.5599	0.2915	0.0480	0.3584	0.5599	0.2915	0.0480
	bbc-seg	0.2333	0.6000	0.0000	-0.0011	0.2333	0.6000	0.0000	-0.0011	0.2333	0.6000	0.0000	0.0087	0.2800	0.6000	0.0000	-0.0023
	bbcsport-seg	0.2333	0.6000	0.0000	-0.0087	0.2800	0.6000	0.0000	0.0093	0.2333	0.6000	0.0000	-0.0009	0.2800	0.6000	0.0000	0.0093
	Image Segmentation	0.6631	0.3381	0.5214	0.2439	0.5479	0.4476	0.3823	0.1068	0.6682	0.3333	0.5176	0.2410	0.7753	0.2286	0.6417	0.2422
	Iris	0.9400	0.0600	0.8340	0.4158	0.9533	0.0467	0.8680	0.4026	0.9400	0.0600	0.8340	0.4158	0.8662	0.1333	0.6765	0.4508
	Multiple Features	0.8904	0.1085	0.7840	0.1322	0.8904	0.1095	0.7805	0.1269	0.8915	0.1075	0.7859	0.1317	0.8941	0.1050	0.7909	0.1322
	Subcorel1	0.5425	0.4125	0.4107	0.1361	0.5425	0.4125	0.4107	0.1361	0.5425	0.4125	0.4107	0.1361	0.5425	0.4125	0.4107	0.1361
	Subcorel2	0.5930	0.3925	0.3680	0.1202	0.5930	0.3925	0.3680	0.1202	0.5930	0.3925	0.3680	0.1202	0.5930	0.3925	0.3680	0.1202
	Subcorel3	0.5690	0.3900	0.4479	0.1885	0.5690	0.3900	0.4479	0.1885	0.5690	0.3900	0.4479	0.1885	0.5690	0.3900	0.4479	0.1885
	Subcorel4	0.8436	0.1550	0.6336	0.1724	0.8436	0.1550	0.6336	0.1724	0.8436	0.1550	0.6336	0.1724	0.8436	0.1550	0.6336	0.1724
	Subcorel5	0.4951	0.5100	0.1728	0.1065	0.4951	0.5100	0.1728	0.1065	0.4951	0.5100	0.1728	0.1065	0.4951	0.5100	0.1728	0.1065
	Water	0.1975	0.6281	0.1268	0.0783	0.1975	0.6281	0.1268	0.0783	0.1975	0.6281	0.1268	0.0783	0.1975	0.6281	0.1268	0.0783
	Wine	0.9281	0.0730	0.7867	0.2799	0.9069	0.0955	0.7285	0.2628	0.9242	0.0787	0.7696	0.2745	0.9069	0.0955	0.7285	0.2628
Chebyshev	Caltech 101-20	0.2055	0.6987	0.1782	-0.0192	0.2055	0.6987	0.1782	-0.0192	0.2055	0.6987	0.1782	-0.0192	0.2055	0.6987	0.1782	-0.0192
	bbc-seg	0.4950	0.4100	0.0000	0.0038	0.4720	0.4050	0.0000	0.0029	0.4880	0.4020	0.0000	0.0015	0.2850	0.5950	0.0000	0.0008
	bbcsport-seg	0.2800	0.6000	0.0000	0.0238	0.2800	0.6000	0.0000	0.0238	0.2800	0.6000	0.0000	0.0238	0.2800	0.6000	0.0000	0.0238
	Image Segmentation	0.6796	0.3048	0.5533	0.2049	0.6801	0.3143	0.5599	0.1890	0.6566	0.3476	0.5034	0.2245	0.5813	0.3857	0.4444	0.2519
	Iris	0.9400	0.0600	0.8340	0.4158	0.9533	0.0467	0.8680	0.4026	0.9400	0.0600	0.8340	0.4158	0.8662	0.1333	0.6765	0.4508
	Multiple Features	0.8352	0.1630	0.6818	0.1149	0.8390	0.1590	0.6890	0.1184	0.8396	0.1585	0.6899	0.1165	0.8380	0.1600	0.6873	0.1186
	Subcorel1	0.4264	0.5300	0.1592	0.0656	0.4264	0.5300	0.1592	0.0656	0.4264	0.5300	0.1592	0.0656	0.4264	0.5300	0.1592	0.0656
	Subcorel2	0.3386	0.6350	0.0364	0.0327	0.3386	0.6350	0.0364	0.0327	0.3386	0.6350	0.0364	0.0327	0.3386	0.6350	0.0364	0.0327
	Subcorel3	0.5132	0.4700	0.2807	0.0995	0.5132	0.4700	0.2807	0.0995	0.5132	0.4700	0.2807	0.0995	0.5132	0.4700	0.2807	0.0995
	Subcorel4	0.6582	0.3475	0.2710	0.0527	0.6582	0.3475	0.2710	0.0527	0.6582	0.3475	0.2710	0.0527	0.6582	0.3475	0.2710	0.0527
	Subcorel5	0.5203	0.4750	0.1844	0.0428	0.5203	0.4750	0.1844	0.0428	0.5203	0.4750	0.1844	0.0428	0.5203	0.4750	0.1844	0.0428
	Water	0.2131	0.6964	0.0624	0.0304	0.2033	0.7135	0.0507	0.0382	0.2131	0.6964	0.0618	0.0308	0.2131	0.6964	0.0624	0.0304
	Wine	0.9281	0.0730	0.7867	0.2799	0.9069	0.0955	0.7285	0.2628	0.9242	0.0787	0.7696	0.2745	0.9069	0.0955	0.7285	0.2628

Legenda: F: F-Measure | **O**: OERC | **A**: ARI | **S**: Silhouette

Tabela 17 – Resultados detalhados (F-Measure, OERC, ARI e Silhouette) por estratégia de cálculo de λ para o algoritmo MVKKM

		Média Inversa				**Variância Inversa**				**Multipl. Lagrange**				**Entropia Relativa**			
Distância	**Base de Dados**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**
Euclidiana	Caltech 101-20	0.9241	0.1698	0.9320	0.5462	0.8106	0.0422	0.7135	0.2097	0.8948	0.0766	0.7273	0.3644	0.9575	0.0895	0.8284	0.5529
	BBC-Seg	0.4523	0.2110	0.3987	0.1052	0.4309	0.2605	0.3658	0.0955	0.4214	0.2550	0.3589	0.0877	0.4478	0.2296	0.3925	0.0995
	BBCSport-Seg	0.6811	0.1128	0.7055	0.2381	0.6605	0.1483	0.6801	0.2255	0.6510	0.1350	0.6714	0.2144	0.6711	0.1288	0.6904	0.2321
	Image Segmentation	0.8125	0.2111	0.8011	0.3554	0.7992	0.2405	0.7841	0.3418	0.7914	0.2333	0.7766	0.3341	0.8041	0.2210	0.7894	0.3472
	Iris	0.8377	0.1611	0.9059	0.3960	0.8113	0.0220	0.8048	0.5987	0.8650	0.0934	0.8769	0.2915	0.9708	0.0001	0.7594	0.3548
	Multiple Features	0.9567	0.0734	0.7388	0.4806	0.9376	0.1744	0.9016	0.4475	0.8747	0.1717	0.8673	0.4774	0.9400	0.1652	0.9323	0.2775
	Subcorel1	0.8818	0.1926	0.8290	0.3111	0.8595	0.1285	0.8947	0.2707	0.9853	0.0477	0.7889	0.2092	0.8563	0.1329	0.8289	0.2138
	Subcorel2	0.8553	0.1286	0.7497	0.5969	0.9740	0.1013	0.7257	0.2875	0.9547	0.0998	0.7533	0.3195	0.9014	0.1126	0.9336	0.2420
	Subcorel3	0.9509	0.0251	0.8387	0.3871	0.9537	0.1359	0.7040	0.5860	0.8571	0.1886	0.9431	0.5371	0.9305	0.0357	0.7425	0.3229
	Subcorel4	0.9024	0.0421	0.8013	0.2797	0.9367	0.0326	0.9164	0.2335	0.9547	0.0018	0.7491	0.5815	0.9780	0.1901	0.8483	0.5158
	Subcorel5	0.8829	0.0292	0.9323	0.2671	0.9496	0.1757	0.7817	0.4436	0.9611	0.0145	0.7343	0.4298	0.8917	0.0151	0.8256	0.5070
	Water	0.9049	0.1290	0.7070	0.3224	0.9037	0.0233	0.8799	0.2115	0.8604	0.0532	0.7449	0.2344	0.9348	0.0225	0.7470	0.2272
	Wine	0.9061	0.1379	0.8972	0.4474	0.8308	0.1672	0.8768	0.5733	0.8724	0.0406	0.7734	0.2622	0.8220	0.1754	0.7133	0.3702
Manhattan	Caltech 101-20	0.8316	0.0129	0.9195	0.5861	0.8189	0.0395	0.8406	0.5397	0.8965	0.1520	0.8089	0.4105	0.8335	0.0725	0.8145	0.5691
	BBC-Seg	0.4735	0.2031	0.3812	0.1143	0.4608	0.2439	0.3783	0.0849	0.4481	0.2120	0.3704	0.0958	0.4650	0.2266	0.3801	0.0886
	BBCSport-Seg	0.6940	0.1210	0.7108	0.2443	0.6759	0.1429	0.6902	0.2368	0.6672	0.1321	0.6848	0.2211	0.6838	0.1298	0.7009	0.2398
	Image Segmentation	0.8344	0.1985	0.8175	0.3611	0.8208	0.2219	0.8028	0.3533	0.8145	0.2140	0.7949	0.3483	0.8280	0.2088	0.8120	0.3568
	Iris	0.9390	0.1552	0.8876	0.2576	0.8204	0.0922	0.8045	0.3148	0.8893	0.0206	0.7599	0.5925	0.8925	0.0330	0.7131	0.5156
	Multiple Features	0.9127	0.0301	0.7271	0.3014	0.8206	0.1693	0.8123	0.3660	0.9545	0.0366	0.8620	0.2935	0.8197	0.1916	0.8489	0.2024
	Subcorel1	0.8361	0.0361	0.9090	0.5623	0.9169	0.0380	0.9037	0.5109	0.8741	0.1003	0.7150	0.2652	0.9810	0.0264	0.7985	0.4410
	Subcorel2	0.9829	0.1285	0.9237	0.2376	0.8040	0.1931	0.8921	0.3269	0.8317	0.1513	0.8027	0.3284	0.8677	0.1284	0.7531	0.4843
	Subcorel3	0.9305	0.1617	0.7159	0.4557	0.9759	0.1229	0.8978	0.3234	0.9336	0.1712	0.7136	0.2721	0.9426	0.0399	0.8686	0.4034
	Subcorel4	0.8833	0.0632	0.8791	0.5396	0.8663	0.1847	0.9255	0.5360	0.8430	0.0937	0.7866	0.5276	0.8637	0.0903	0.7307	0.5384
	Subcorel5	0.8993	0.1457	0.9330	0.2979	0.9451	0.1132	0.9110	0.3217	0.8582	0.1911	0.8945	0.5803	0.8350	0.1041	0.9470	0.5549
	Water	0.9534	0.0797	0.9278	0.3642	0.9579	0.0577	0.9343	0.4832	0.8025	0.0745	0.8017	0.4839	0.9615	0.1351	0.9119	0.4665
	Wine	0.9282	0.1603	0.9278	0.3499	0.8446	0.0918	0.9218	0.3580	0.9200	0.1466	0.9235	0.4928	0.9222	0.0106	0.8837	0.4855
Chebyshev	Caltech 101-20	0.9382	0.1884	0.8318	0.3886	0.9387	0.1066	0.7483	0.5121	0.9854	0.0368	0.8009	0.2115	0.9673	0.1467	0.7373	0.5918
	BBC-Seg	0.3901	0.2408	0.3558	0.0812	0.3765	0.2721	0.3309	0.0621	0.3544	0.2819	0.3104	0.0583	0.3711	0.2602	0.3240	0.0659
	BBCSport-Seg	0.6333	0.1529	0.6681	0.2099	0.6183	0.1701	0.6432	0.1931	0.6009	0.1885	0.6301	0.1842	0.6122	0.1805	0.6404	0.1910
	Image Segmentation	0.8017	0.2308	0.7880	0.3450	0.7882	0.2601	0.7654	0.3259	0.7700	0.2728	0.7499	0.3102	0.7821	0.2495	0.7614	0.3228
	Iris	0.9425	0.1194	0.7871	0.5220	0.9019	0.0171	0.9331	0.3080	0.9311	0.0181	0.7881	0.5048	0.8751	0.1856	0.8837	0.5453
	Multiple Features	0.8639	0.0143	0.9100	0.5545	0.8283	0.1527	0.7091	0.2945	0.8427	0.0748	0.8812	0.2154	0.9768	0.1320	0.8604	0.4278
	Subcorel1	0.9743	0.1138	0.7927	0.3687	0.8249	0.0219	0.7316	0.5685	0.8058	0.0958	0.7246	0.4288	0.8643	0.1223	0.9338	0.2217
	Subcorel2	0.8555	0.0903	0.8864	0.4047	0.8266	0.1939	0.8644	0.5052	0.8613	0.1092	0.7234	0.3711	0.8084	0.1714	0.8463	0.2941
	Subcorel3	0.9859	0.1382	0.8186	0.5278	0.8335	0.0780	0.7097	0.4723	0.8884	0.0374	0.7302	0.2932	0.8384	0.1814	0.7467	0.4463
	Subcorel4	0.8989	0.1372	0.7822	0.5759	0.9520	0.1604	0.7580	0.4659	0.9441	0.0575	0.7412	0.4426	0.8204	0.1540	0.9434	0.5932
	Subcorel5	0.8648	0.0259	0.8167	0.4184	0.9819	0.1719	0.7577	0.4440	0.8844	0.1308	0.7657	0.4897	0.8595	0.0940	0.9009	0.2917
	Water	0.9190	0.1880	0.7468	0.4993	0.8060	0.0632	0.9352	0.4199	0.8021	0.0003	0.9137	0.5960	0.9550	0.0772	0.7237	0.4280
	Wine	0.8772	0.0016	0.7834	0.2712	0.9317	0.0146	0.8454	0.3994	0.9420	0.0011	0.9246	0.5209	0.8307	0.1898	0.9486	0.4908
Cosseno	Caltech 101-20	0.9081	0.1120	0.8959	0.5235	0.8017	0.0655	0.8121	0.4528	0.8706	0.0129	0.8738	0.5003	0.9162	0.1423	0.7943	0.3856
	BBC-Seg	0.5012	0.1896	0.4206	0.1284	0.4922	0.2306	0.4094	0.1219	0.4783	0.2028	0.4020	0.1074	0.4895	0.1982	0.4171	0.1167
	BBCSport-Seg	0.7019	0.1096	0.7193	0.2561	0.6877	0.1199	0.7049	0.2488	0.6925	0.1188	0.7130	0.2422	0.6985	0.1152	0.7166	0.2494
	Image Segmentation	0.8452	0.1872	0.8288	0.3772	0.8317	0.1990	0.8171	0.3600	0.8228	0.2090	0.8058	0.3519	0.8389	0.1962	0.8218	0.3667
	Iris	0.9322	0.1672	0.8818	0.4082	0.9052	0.0608	0.7870	0.5189	0.8058	0.1667	0.9049	0.2398	0.9760	0.1120	0.7399	0.4260
	Multiple Features	0.9556	0.1552	0.7140	0.3444	0.8405	0.1078	0.8542	0.5085	0.9665	0.0913	0.7744	0.3457	0.8808	0.0665	0.8550	0.5517
	Subcorel1	0.8670	0.0063	0.8907	0.2938	0.8403	0.1583	0.8809	0.2424	0.9126	0.0870	0.7677	0.5674	0.8651	0.0097	0.8624	0.2875
	Subcorel2	0.9659	0.0545	0.7664	0.3340	0.9193	0.0304	0.7678	0.3367	0.9511	0.0834	0.8655	0.4027	0.8684	0.0937	0.9105	0.2375
	Subcorel3	0.8242	0.0415	0.8548	0.5945	0.8422	0.0635	0.7409	0.2510	0.9742	0.0224	0.8020	0.4255	0.8354	0.0436	0.7175	0.3036
	Subcorel4	0.9660	0.0779	0.8779	0.5651	0.9806	0.1986	0.7167	0.2609	0.8431	0.0992	0.7117	0.4528	0.8053	0.1024	0.7787	0.5380
	Subcorel5	0.9019	0.0992	0.7571	0.5065	0.8148	0.1466	0.7476	0.4691	0.8510	0.0606	0.9309	0.4617	0.9290	0.1870	0.8916	0.3549
	Water	0.8710	0.0906	0.8709	0.5193	0.8963	0.1277	0.7123	0.5416	0.9408	0.0112	0.7451	0.5501	0.9853	0.0042	0.8408	0.2385
	Wine	0.9453	0.0967	0.9070	0.4758	0.9534	0.0196	0.7567	0.3149	0.9419	0.0022	0.7734	0.5180	0.9436	0.0482	0.7177	0.5313

Legenda: F: F-Measure | **O**: OERC | **A**: ARI | **S**: Silhouette

Tabela 18 – Resultados detalhados (F-Measure, OERC, ARI e Silhouette) por estratégia de cálculo de λ para o algoritmo TW-KM

		Média Inversa				**Variância Inversa**				**Multipl. Lagrange**				**Entropia Relativa**			
Distância	**Base de Dados**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**
Euclidiana	Caltech 101-20	0.0409	0.9621	0.2450	0.0209	0.0340	0.9600	0.1848	-0.0198	0.0379	0.9641	0.3294	0.0552	0.0414	0.9580	0.3596	0.0450
	BBC-Seg	0.2985	0.4336	0.3615	0.0027	0.1350	0.6628	0.0101	0.1026	0.3248	0.4978	0.1307	0.0499	0.2502	0.5635	0.0793	0.0206
	BBCSport-Seg	0.1877	0.6724	0.0074	-0.0534	0.1877	0.6724	0.0074	-0.0534	0.1877	0.6724	0.0074	-0.0534	0.1877	0.6724	0.0074	-0.0534
	Image Segmentation	0.6215	0.3238	0.5610	0.2872	0.6082	0.3905	0.4666	0.1780	0.4993	0.4190	0.4653	0.3604	0.6553	0.3381	0.5095	0.2883
	Iris	0.9400	0.0600	0.8340	0.4158	0.9600	0.0400	0.8857	0.4072	0.5309	0.4200	0.4328	0.4799	0.9600	0.0400	0.8857	0.4072
	Multiple Features	0.8440	0.1575	0.7434	0.1293	0.8441	0.1580	0.7482	0.1261	0.8521	0.1500	0.7570	0.1292	0.8451	0.1565	0.7447	0.1292
	Subcorel1	0.5190	0.4375	0.3993	0.1429	0.5190	0.4375	0.3993	0.1429	0.5190	0.4375	0.3993	0.1429	0.5190	0.4375	0.3993	0.1429
	Subcorel2	0.5403	0.4650	0.2831	0.1262	0.5403	0.4650	0.2831	0.1262	0.5403	0.4650	0.2831	0.1262	0.5403	0.4650	0.2831	0.1262
	Subcorel3	0.6162	0.3750	0.4459	0.1798	0.6162	0.3750	0.4459	0.1798	0.6162	0.3750	0.4459	0.1798	0.6162	0.3750	0.4459	0.1798
	Subcorel4	0.8360	0.1625	0.6185	0.1722	0.8360	0.1625	0.6185	0.1722	0.8360	0.1625	0.6185	0.1722	0.8360	0.1625	0.6185	0.1722
	Subcorel5	0.5509	0.4675	0.2132	0.1308	0.5509	0.4675	0.2132	0.1308	0.5509	0.4675	0.2132	0.1308	0.5509	0.4675	0.2132	0.1308
	Water	0.0485	0.9681	0.3335	0.1119	0.0490	0.9681	0.2435	0.0719	0.0490	0.9681	0.3543	0.1071	0.0485	0.9681	0.3619	0.1196
	Wine	0.9669	0.0337	0.8975	0.2849	0.9407	0.0618	0.8150	0.2800	0.9669	0.0337	0.8975	0.2849	0.9669	0.0337	0.8975	0.2849
Manhattan	Caltech 101-20	0.0439	0.9498	0.1929	-0.0006	0.0399	0.9508	0.1300	-0.0180	0.0547	0.9447	0.3829	0.0448	0.0421	0.9539	0.3121	0.0295
	BBC-Seg	0.1139	0.6657	0.0037	0.1199	0.1157	0.6657	0.0049	0.1030	0.1162	0.6642	0.0026	0.1195	0.1139	0.6657	0.0037	0.1322
	BBCSport-Seg	0.2103	0.6638	0.0097	0.0517	0.2015	0.6638	0.0147	0.0538	0.3097	0.6034	0.0062	0.0346	0.2519	0.6207	0.0139	0.0538
	Image Segmentation	0.5437	0.4143	0.4745	0.2909	0.6397	0.3476	0.4996	0.2038	0.5300	0.4095	0.4678	0.2860	0.5677	0.4048	0.4573	0.2547
	Iris	0.8847	0.1133	0.7173	0.4525	0.9600	0.0400	0.8857	0.4072	0.5362	0.4133	0.4371	0.4784	0.9465	0.0533	0.8512	0.4119
	Multiple Features	0.8562	0.1455	0.7595	0.1294	0.8498	0.1530	0.7529	0.1261	0.8508	0.1515	0.7543	0.1288	0.8547	0.1470	0.7589	0.1290
	Subcorel1	0.5548	0.4300	0.3734	0.1388	0.5548	0.4300	0.3734	0.1388	0.5548	0.4300	0.3734	0.1388	0.5548	0.4300	0.3734	0.1388
	Subcorel2	0.5179	0.4900	0.2417	0.1243	0.5179	0.4900	0.2417	0.1243	0.5179	0.4900	0.2417	0.1243	0.5179	0.4900	0.2417	0.1243
	Subcorel3	0.5758	0.3800	0.4690	0.1902	0.5758	0.3800	0.4690	0.1902	0.5758	0.3800	0.4690	0.1902	0.5758	0.3800	0.4690	0.1902
	Subcorel4	0.8057	0.1925	0.5641	0.1699	0.8057	0.1925	0.5641	0.1699	0.8057	0.1925	0.5641	0.1699	0.8057	0.1925	0.5641	0.1699
	Subcorel5	0.4960	0.5175	0.1963	0.0892	0.4960	0.5175	0.1963	0.0892	0.4960	0.5175	0.1963	0.0892	0.4960	0.5175	0.1963	0.0892
	Water	0.0815	0.8830	0.2658	0.1169	0.0490	0.9681	0.2385	0.0944	0.0490	0.9681	0.3586	0.1167	0.0485	0.9681	0.3801	0.1169
	Wine	0.9513	0.0506	0.8471	0.2769	0.9558	0.0449	0.8649	0.2802	0.9513	0.0506	0.8471	0.2769	0.9565	0.0449	0.8636	0.2785
Chebyshev	Caltech 101-20	0.0310	0.9723	0.2451	-0.0034	0.0359	0.9559	0.1174	-0.0237	0.0434	0.9488	0.3415	0.0454	0.0291	0.9662	0.2733	0.0266
	BBC-Seg	0.3753	0.5562	0.1991	-0.0352	0.2124	0.7299	0.0066	-0.0142	0.1599	0.6847	-0.0156	-0.0347	0.3303	0.4832	0.1415	-0.0518
	BBCSport-Seg	0.2071	0.6983	-0.0095	-0.0561	0.2800	0.6724	-0.0048	-0.0521	0.1877	0.6724	0.0074	-0.0534	0.2071	0.6983	-0.0095	-0.0561
	Image Segmentation	0.6336	0.3238	0.5058	0.2188	0.6305	0.3476	0.4618	0.1803	0.4343	0.5000	0.3720	0.3019	0.6552	0.3048	0.5146	0.2189
	Iris	0.9467	0.0533	0.8508	0.4106	0.9600	0.0400	0.8857	0.4072	0.5151	0.4333	0.4336	0.4736	0.9600	0.0400	0.8857	0.4072
	Multiple Features	0.8123	0.1900	0.6818	0.1176	0.6829	0.3040	0.5895	0.1065	0.8206	0.1820	0.6961	0.1181	0.8168	0.1855	0.6886	0.1180
	Subcorel1	0.5675	0.4225	0.2354	0.0414	0.5675	0.4225	0.2354	0.0414	0.5675	0.4225	0.2354	0.0414	0.5675	0.4225	0.2354	0.0414
	Subcorel2	0.4336	0.5575	0.1359	0.0356	0.4336	0.5575	0.1359	0.0356	0.4336	0.5575	0.1359	0.0356	0.4336	0.5575	0.1359	0.0356
	Subcorel3	0.5177	0.4825	0.2628	0.0867	0.5177	0.4825	0.2628	0.0867	0.5177	0.4825	0.2628	0.0867	0.5177	0.4825	0.2628	0.0867
	Subcorel4	0.5971	0.3950	0.2799	0.0506	0.5971	0.3950	0.2799	0.0506	0.5971	0.3950	0.2799	0.0506	0.5971	0.3950	0.2799	0.0506
	Subcorel5	0.5116	0.4950	0.1647	0.0558	0.5116	0.4950	0.1647	0.0558	0.5116	0.4950	0.1647	0.0558	0.5116	0.4950	0.1647	0.0558
	Water	0.0490	0.9681	0.2148	0.1188	0.0579	0.8830	0.0833	-0.0999	0.0490	0.9681	0.2989	0.1054	0.0425	0.9787	0.2284	0.1222
	Wine	0.9062	0.0955	0.7277	0.2558	0.9002	0.1011	0.7143	0.2608	0.9009	0.1011	0.7134	0.2537	0.8901	0.1124	0.6858	0.2538
Cosseno	Caltech 101-20	0.0296	0.9682	0.3529	0.0409	0.0373	0.9549	0.2619	0.0226	0.0287	0.9682	0.3220	0.0366	0.0346	0.9693	0.3298	0.0445
	BBC-Seg	0.5160	0.3664	0.3599	-0.0316	0.4193	0.5547	0.2451	-0.0327	0.4491	0.4730	0.2429	-0.0078	0.5580	0.3956	0.3767	-0.0199
	BBCSport-Seg	0.6080	0.3879	0.3382	-0.0134	0.6080	0.3879	0.3382	-0.0134	0.6080	0.3879	0.3382	-0.0134	0.6080	0.3879	0.3382	-0.0134
	Image Segmentation	0.5960	0.3524	0.4902	0.2178	0.4874	0.4952	0.3153	0.1665	0.5853	0.3714	0.4764	0.2233	0.5309	0.4048	0.4494	0.2119
	Iris	0.7336	0.2467	0.5099	0.3379	0.7687	0.2133	0.5656	0.3606	0.6140	0.3600	0.4173	0.4200	0.7220	0.2533	0.5084	0.3235
	Multiple Features	0.8383	0.1645	0.7315	0.1286	0.8313	0.1710	0.7236	0.1273	0.8450	0.1570	0.7523	0.1286	0.8432	0.1595	0.7439	0.1295
	Subcorel1	0.7383	0.2600	0.4556	0.1138	0.7383	0.2600	0.4556	0.1138	0.7383	0.2600	0.4556	0.1138	0.7383	0.2600	0.4556	0.1138
	Subcorel2	0.5594	0.4350	0.3206	0.1208	0.5594	0.4350	0.3206	0.1208	0.5594	0.4350	0.3206	0.1208	0.5594	0.4350	0.3206	0.1208
	Subcorel3	0.6547	0.3325	0.4287	0.1757	0.6547	0.3325	0.4287	0.1757	0.6547	0.3325	0.4287	0.1757	0.6547	0.3325	0.4287	0.1757
	Subcorel4	0.8466	0.1525	0.6389	0.1723	0.8466	0.1525	0.6389	0.1723	0.8466	0.1525	0.6389	0.1723	0.8466	0.1525	0.6389	0.1723
	Subcorel5	0.5224	0.4675	0.2146	0.0895	0.5224	0.4675	0.2146	0.0895	0.5224	0.4675	0.2146	0.0895	0.5224	0.4675	0.2146	0.0895
	Water	0.0808	0.9043	0.1950	0.0620	0.0661	0.8723	0.1508	0.0110	0.0523	0.8830	0.2318	0.0587	0.0747	0.8936	0.2141	0.0343
	Wine	0.9355	0.0674	0.7994	0.2786	0.9341	0.0674	0.8011	0.2812	0.9289	0.0730	0.7857	0.2774	0.9234	0.0787	0.7703	0.2787

Legenda: F: F-Measure | **O**: OERC | **A**: ARI | **S**: Silhouette

Tabela 19 – Resultados detalhados (F-Measure, OERC, ARI e Silhouette) por estratégia de cálculo de λ para o algoritmo MVSC

		Média Inversa				**Variância Inversa**				**Multipl. Lagrange**				**Entropia Relativa**			
Distância	**Base de Dados**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**	**F**	**O**	**A**	**S**
Euclidiana	Caltech 101-20	0.9272	0.0160	0.8554	0.3590	0.9254	0.0704	0.8856	0.3254	0.9398	0.0798	0.9299	0.5132	0.8137	0.0735	0.8622	0.4419
	BBC-Seg	0.8469	0.0572	0.8771	0.5878	0.8050	0.0729	0.8242	0.3072	0.8248	0.1346	0.7502	0.4723	0.9723	0.1497	0.9197	0.5391
	BBCSport-Seg	0.8079	0.0857	0.9427	0.3582	0.8002	0.0902	0.8756	0.5621	0.8248	0.0216	0.8556	0.5477	0.9427	0.1830	0.9100	0.5741
	Image Segmentation	0.8621	0.1558	0.9128	0.4901	0.9573	0.1650	0.9172	0.4683	0.9248	0.0686	0.7036	0.3618	0.9892	0.1029	0.7489	0.4499
	Iris	0.9527	0.0318	0.7099	0.2843	0.9468	0.0404	0.7116	0.2250	0.8180	0.0179	0.7022	0.3612	0.9355	0.1210	0.9268	0.3494
	Multiple Features	0.9527	0.0318	0.7099	0.2843	0.9468	0.0404	0.7116	0.2250	0.8180	0.0179	0.7022	0.3612	0.9355	0.1210	0.9268	0.3494
	Subcorel1	0.8803	0.0519	0.9269	0.4253	0.8417	0.0392	0.8439	0.3109	0.8610	0.0032	0.7759	0.3400	0.8209	0.0226	0.7203	0.3291
	Subcorel2	0.8359	0.0238	0.7058	0.3721	0.9259	0.0073	0.7439	0.3222	0.8143	0.1278	0.8275	0.2554	0.9246	0.0587	0.8586	0.5300
	Subcorel3	0.8177	0.0161	0.8174	0.3656	0.9443	0.0359	0.8984	0.3449	0.8476	0.1052	0.7640	0.3135	0.9264	0.1254	0.7500	0.4271
	Subcorel4	0.9879	0.0437	0.8958	0.5623	0.8217	0.1618	0.9021	0.5111	0.8671	0.0832	0.9285	0.3911	0.9107	0.0032	0.8511	0.3469
	Subcorel5	0.9109	0.1198	0.7514	0.5984	0.9792	0.1031	0.7372	0.4473	0.8765	0.0068	0.8736	0.3739	0.9596	0.0951	0.7960	0.5298
	Water	0.9008	0.1011	0.8520	0.3614	0.8892	0.1206	0.7418	0.5201	0.8508	0.1069	0.7969	0.4176	0.9633	0.0822	0.7604	0.5323
	Wine	0.9549	0.0490	0.8643	0.4494	0.8967	0.0523	0.7250	0.2858	0.9554	0.0202	0.8249	0.3061	0.8629	0.1365	0.7293	0.5586
Manhattan	Caltech 101-20	0.9756	0.1502	0.9371	0.5819	0.9032	0.0292	0.9111	0.2746	0.8426	0.0136	0.9376	0.5687	0.8642	0.1396	0.8772	0.3986
	BBC-Seg	0.8767	0.1693	0.8678	0.3133	0.8695	0.0454	0.8629	0.4657	0.9862	0.0110	0.7625	0.4238	0.9648	0.1734	0.8097	0.2538
	BBCSport-Seg	0.9551	0.0835	0.8623	0.5600	0.9721	0.1294	0.7881	0.4331	0.8725	0.0309	0.8572	0.2508	0.9699	0.0492	0.9218	0.2090
	Image Segmentation	0.9283	0.1664	0.8397	0.3459	0.8591	0.1056	0.9130	0.3463	0.9743	0.1266	0.8813	0.5967	0.8245	0.1579	0.8205	0.5235
	Iris	0.9126	0.0075	0.7766	0.2315	0.8437	0.1235	0.8261	0.4924	0.8658	0.0523	0.8422	0.3892	0.9024	0.0462	0.7517	0.4048
	Multiple Features	0.9126	0.0075	0.7766	0.2315	0.8437	0.1235	0.8261	0.4924	0.8658	0.0523	0.8422	0.3892	0.9024	0.0462	0.7517	0.4048
	Subcorel1	0.9130	0.2000	0.8332	0.4767	0.9649	0.1178	0.8851	0.3609	0.9828	0.0954	0.9466	0.2174	0.8427	0.1206	0.8343	0.5124
	Subcorel2	0.9808	0.0211	0.9004	0.3383	0.8869	0.0824	0.7967	0.3821	0.9579	0.1885	0.8088	0.5554	0.8137	0.0528	0.7583	0.5338
	Subcorel3	0.8559	0.0480	0.8311	0.5273	0.8143	0.0630	0.8707	0.5142	0.8664	0.1639	0.9273	0.2264	0.8056	0.0347	0.8525	0.3569
	Subcorel4	0.8754	0.0627	0.7477	0.4530	0.8358	0.1792	0.8958	0.2046	0.9787	0.0658	0.7030	0.5951	0.8430	0.0025	0.9415	0.4304
	Subcorel5	0.9613	0.0135	0.8618	0.2722	0.9392	0.1930	0.7826	0.2366	0.9365	0.0457	0.7746	0.2479	0.9588	0.1619	0.9135	0.4856
	Water	0.8877	0.1174	0.8858	0.2351	0.8027	0.1836	0.7863	0.2592	0.8614	0.1168	0.8193	0.2631	0.8517	0.1507	0.9444	0.4720
	Wine	0.8806	0.1103	0.9456	0.2868	0.8688	0.0324	0.8979	0.4340	0.8196	0.1153	0.8777	0.2875	0.9633	0.1443	0.7958	0.5261
Chebyshev	Caltech 101-20	0.8600	0.0453	0.7806	0.3210	0.9007	0.1010	0.7281	0.4399	0.8323	0.1351	0.8853	0.4941	0.9412	0.0255	0.8744	0.3578
	BBC-Seg	0.8405	0.0278	0.9409	0.5759	0.9802	0.1275	0.8572	0.2210	0.8465	0.1926	0.9441	0.5013	0.9124	0.1011	0.7098	0.4727
	BBCSport-Seg	0.9233	0.0687	0.9022	0.5286	0.9807	0.1557	0.9389	0.3635	0.8899	0.0973	0.8094	0.2263	0.9378	0.0365	0.9488	0.4925
	Image Segmentation	0.8795	0.1563	0.8916	0.3258	0.9871	0.0937	0.7093	0.3950	0.8426	0.1697	0.7439	0.2085	0.9315	0.0945	0.7630	0.2150
	Iris	0.8180	0.0020	0.8651	0.5877	0.8569	0.1755	0.7134	0.2787	0.9517	0.0399	0.7348	0.5451	0.8729	0.1891	0.8239	0.5625
	Multiple Features	0.8180	0.0020	0.8651	0.5877	0.8569	0.1755	0.7134	0.2787	0.9517	0.0399	0.7348	0.5451	0.8729	0.1891	0.8239	0.5625
	Subcorel1	0.9002	0.0852	0.7365	0.4138	0.8065	0.1799	0.7597	0.3894	0.9568	0.1790	0.7470	0.2890	0.9454	0.0275	0.8536	0.5033
	Subcorel2	0.9359	0.1263	0.8217	0.5469	0.9044	0.0710	0.8638	0.4347	0.8797	0.0047	0.7422	0.3882	0.9610	0.1953	0.7653	0.3509
	Subcorel3	0.8511	0.0905	0.7353	0.5181	0.8995	0.0992	0.8818	0.5501	0.8812	0.0036	0.7023	0.2173	0.8380	0.1651	0.7012	0.2773
	Subcorel4	0.9505	0.0840	0.8554	0.3063	0.9670	0.1748	0.7461	0.4827	0.9202	0.0822	0.7744	0.4411	0.9296	0.1129	0.8153	0.4587
	Subcorel5	0.9478	0.1987	0.7578	0.3135	0.8880	0.0102	0.7049	0.4425	0.8224	0.1026	0.9475	0.3062	0.9484	0.1606	0.7734	0.2688
	Water	0.8764	0.1810	0.7776	0.5855	0.8191	0.1617	0.7314	0.2979	0.8743	0.1780	0.9379	0.5283	0.9490	0.0786	0.7541	0.4947
	Wine	0.9858	0.0309	0.9075	0.4221	0.8241	0.0003	0.7943	0.5340	0.9098	0.1585	0.7342	0.4204	0.8672	0.0853	0.8592	0.4756
Cosseno	Caltech 101-20	0.9207	0.0050	0.9056	0.2998	0.9248	0.1629	0.8866	0.3126	0.9876	0.1893	0.9089	0.2599	0.9875	0.0940	0.8556	0.2431
	BBC-Seg	0.9349	0.1726	0.8244	0.3799	0.9266	0.1917	0.8675	0.5420	0.9524	0.0795	0.7367	0.2580	0.9867	0.0962	0.8857	0.5988
	BBCSport-Seg	0.8367	0.0814	0.8053	0.4475	0.9276	0.0381	0.8800	0.2863	0.8803	0.1635	0.8621	0.3713	0.8946	0.1872	0.8140	0.3844
	Image Segmentation	0.9535	0.0957	0.7093	0.5313	0.8913	0.0934	0.8035	0.3658	0.9652	0.0137	0.8096	0.4485	0.9700	0.1318	0.8670	0.2872
	Iris	0.9881	0.0540	0.7859	0.4027	0.9640	0.0657	0.8140	0.4918	0.9073	0.0099	0.8562	0.5055	0.9772	0.1559	0.7736	0.2607
	Multiple Features	0.9881	0.0540	0.7859	0.4027	0.9640	0.0657	0.8140	0.4918	0.9073	0.0099	0.8562	0.5055	0.9772	0.1559	0.7736	0.2607
	Subcorel1	0.9748	0.1269	0.7107	0.2029	0.8175	0.0660	0.7693	0.5554	0.9074	0.0814	0.9143	0.2515	0.8528	0.1249	0.8820	0.4193
	Subcorel2	0.9084	0.0283	0.7480	0.3868	0.9087	0.1177	0.8678	0.2639	0.9319	0.0077	0.8124	0.4309	0.8875	0.0847	0.8025	0.2192
	Subcorel3	0.9799	0.1503	0.8404	0.3509	0.9747	0.1018	0.7155	0.2712	0.8108	0.0655	0.7569	0.5737	0.8764	0.1046	0.7477	0.3624
	Subcorel4	0.8101	0.0933	0.9485	0.2539	0.9091	0.1433	0.8985	0.2776	0.8171	0.0232	0.7820	0.3618	0.8933	0.1558	0.8851	0.4330
	Subcorel5	0.9375	0.0229	0.9237	0.2173	0.8112	0.0277	0.7265	0.5757	0.9297	0.0002	0.8924	0.2978	0.9099	0.0667	0.8003	0.4006
	Water	0.9303	0.1733	0.7036	0.5193	0.9101	0.0934	0.9403	0.5266	0.9278	0.0861	0.8139	0.4377	0.8607	0.1326	0.7640	0.5348
	Wine	0.8059	0.1713	0.8460	0.2562	0.9380	0.0612	0.8534	0.2610	0.9610	0.1032	0.7083	0.2860	0.8757	0.1155	0.7884	0.5863

Legenda: F: F-Measure | **O**: OERC | **A**: ARI | **S**: Silhouette