Pós-Graduação em Ciência da Computação

Renê Pereira de Gusmão

Métodos Híbridos para Agrupamento de Dados Relacionais com Múltiplas Visões

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Renê Pereira de Gusmão

Métodos Híbridos para Agrupamento de Dados Relacionais com Múltiplas Visões

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional
Orientador: Prof. Francisco de Assis Tenório de Carvalho

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

G982m Gusmão, Renê Pereira de
Métodos híbridos para agrupamento de dados relacionais com múltiplas visões / Renê Pereira de Gusmão. – 2019.
160 f.: il., fig., tab.

Orientador: Francisco de Assis Tenório de Carvalho.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Otimização. 3. Análise de cluster. I. Carvalho, Francisco de Assis Tenório (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2019-073

Renê Pereira de Gusmão

"Métodos Híbridos para Agrupamento de Dados Relacionais com Múltiplas Visões"

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 26/04/2019.

Orientador: Prof. Francisco de Assis Tenório de Carvalho

BANCA EXAMINADORA

Prof. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática/UFPE

Prof. Adriano Lorena Inácio de Oliveira
Centro de Informática/UFPE

Prof. Cleber Zanchettin
Centro de Informática/UFPE

Prof. Carmelo Jose Albanez Bastos Filho
Escola Politécnica de Pernambuco/UPE

Prof. Lucidio dos Anjos Formiga Cabral
Centro de Informática/UFPB

Dedico este trabalho a minha esposa, Cleonides, que foi meu porto seguro perante as dificuldades durante todo o percurso.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela existência e pela saúde.

Agradeço ao meu orientador, Prof. Francisco de Assis, por me dar a oportunidade de fazer parte do seu grupo de pesquisa. Agradeço por seu contínuo suporte em meus estudos e pesquisa, e também por sua motivação, disponibilidade, paciência e conhecimento.

Agradeço ao Departamento de Computação da Universidade Federal de Sergipe por ter aprovado o afastamento necessário para a conclusão deste trabalho.

Agradeço ao Laboratório de Computação de Alto Desempenho da Universidade Federal de Sergipe e ao Laboratório responsável pelo *Cluster* computacional do CIn-UFPE pelo uso de recursos computacionais para execução de parte dos experimentos.

Agradeço a Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) pelo suporte financeiro recebido.

Agradeço a toda minha família pelo apoio e torcida. Em especial, agradeço a minha amada esposa por todo suporte e compreensão.

Agradeço a todos os professores da banca que pacientemente avaliaram e contribuiram para o aprimoramento da minha pesquisa.

Por fim, agradeço a todos os professores que tive pelo caminho e que tanto contribuíram para minha formação.

RESUMO

O agrupamento de dados com múltiplas visões é um problema emergente e que vem sendo pesquisado nos últimos anos. Modelos para agrupamento de dados relacionais presentes na literatura apresentam rápida convergência e, consequentemente, o espaço de soluções não é explorado de forma adequada. Esta tese de doutorado teve como objetivo a investigação e desenvolvimento de métodos híbridos baseados em otimização por nuvem de partículas para resolver o problema do agrupamento de dados relacionais com múltiplas visões. Os métodos híbridos desenvolvidos combinam otimização por nuvem de partículas com métodos de agrupamento baseados em matrizes de dissimilaridades com o objetivo de se beneficiar das características de cada abordagem e explorar de melhor forma o espaço de soluções. A abordagem desenvolvida foi avaliada para agrupamento rígido e nebuloso de dados relacionais. Além disso, devido a importância da escolha de uma função de aptidão apropriada, diversos índices para validação de agrupamentos foram investigados e adaptados para considerar dissimilaridades fornecidas por várias matrizes bem como pesos de relevância para cada matriz. Seis estudos foram realizados para validação dos modelos híbridos desenvolvidos. No primeiro estudo, o modelo híbrido para agrupamento rígido de dados relacionais com única visão foi comparado a outros métodos da literatura e obteve resultados competitivos. No segundo estudo, os agrupamentos rígidos gerados por onze funções de aptidão para diversas bases de dados reais foram avaliados em termos dos índices externos medida F e índice ajustado de Rand tanto para o modelo que considera dados com única visão quanto para os modelos para dados com múltiplas visões. As funções de aptidão que se destacaram dentre as demais foram: índice da silhueta, índice de Xu e homogeneidade intra-cluster. Os resultados obtidos pelo índice da silhueta e pela homogeneidade foram selecionados para comparação com os resultados obtidos por outros métodos da literatura no terceiro estudo. Verificou-se que a abordagem proposta apresentou melhores resultados para a maioria dos casos analisados. Três estudos também foram realizados para validação dos modelos híbridos para agrupamento nebuloso de dados com única visão e com várias visões. As funções de aptidão para agrupamento nebuloso que se destacaram dentre as demais foram: silhueta simplificada e coeficiente da partição. A análise dos resultados mostrou que a abordagem proposta para agrupamento nebuloso também obteve desempenho competitivo e melhor em alguns casos em comparação a outros métodos da literatura. Os resultados demonstraram que o problema do agrupamento de dados relacionais com múltiplas visões pode ser melhorado de forma significativa através de métodos híbridos baseados em otimização de enxame. Portanto, tais achados reforçam a importância da aplicação de técnicas tais como algoritmos baseados em otimização de enxame no campo da mineração de dados.

Palavras-chaves: Otimização por nuvem de partículas. Análise de cluster. Dados relacionais. Agrupamento de dados com múltiplas visões.

ABSTRACT

Clustering of multi-view data is an emerging problem that has been researched in recent years. Existing models for relational data clustering in the literature present fast convergence and, consequently, the solution space is not adequately explored. This thesis aimed at the investigation and development of hybrid methods based on particle swarm optimization to solve the problem of clustering multi-view relational data. The hybrid methods developed combine particle swarm optimization with clustering methods based on dissimilarity matrices in order to benefit from the characteristics of each approach and to better exploit the solution space. The approach developed was evaluated for hard and fuzzy clustering of relational data. In addition, due to the importance of choosing an appropriate fitness function, several clustering validation indices have been investigated and adapted to consider dissimilarities provided by several matrices as well as relevance weights for each matrix. Six studies were carried out to validate the hybrid models developed. In the first study, the hybrid model for rigid clustering of single-view relational data was compared to other methods in the literature and obtained competitive results. In the second study, the hard clusterings generated by eleven fitness functions for various real data sets were evaluated in terms of the external indexes F-measure and Adjusted Rand index for both the model that considers single-view data and the models for multi-view data. The fitness functions that stood out among the others were: silhouette index, Xu index and intra-cluster homogeneity. The results obtained by the silhouette index and homogeneity were selected for comparison with the results obtained by other methods of the literature in the third study. It was found that the proposed approach presented better results for most of the cases analyzed. Three studies were also carried out to validate hybrid models for fuzzy clustering of single-view data and multi-view data. The fitness functions for fuzzy clustering that stood out among others were: simplified silhouette and partition coefficient. The analysis of the results showed that the approach proposed for fuzzy clustering also obtained competitive performance and better in some cases in comparison with other methods of the literature. The results demonstrated that the problem of clustering multi-view relational data can be significantly improved through hybrid methods based on swarm optimization. Therefore, the results reinforce the importance of the application of techniques such as algorithms based on swarm optimization in the field of data mining.

Keywords: Particle swarm optimization. Cluster analysis. Relational data. Clustering of multi-view data.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Algoritmos de agrupamento particionam um conjunto de objetos em grupos ou *clusters*, de tal forma que os objetos que estão no mesmo grupo possuem elevado grau de similaridade e os objetos que estão em grupos diferentes possuem alto grau de dissimilaridade entre eles. Esses algoritmos são amplamente utilizados em muitas áreas, incluindo: mineração de dados, estatística, biologia, aprendizado de máquina e reconhecimento de padrões (JAIN; MURTY; FLYNN, 1999; HAN; KAMBER; PEI, 2012; LONG; ZHANG; YU, 2010; Chao; Sun; Bi, 2017). Um exemplo de agrupamento de dados é ilustrado na Figura 1. Na Figura 1, após o agrupamento através de algum algoritmo, quatro grupos bem separados e compactos foram encontrados.

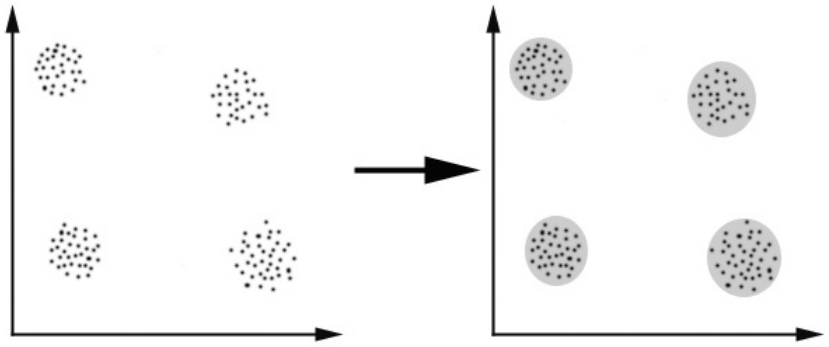


Figura 1 – Exemplo de agrupamento de dados
Fonte: (BHARNE; GULHANE; YEWALE, 2011)

Dado um conjunto E contendo n objetos, algoritmos de particionamento constroem K grupos disjuntos dos dados (Equações (1.1) e (1.3)), em que cada grupo representa um *cluster* C_i $(i = 1, .., K)$. Além disso, cada partição deve ser não-vazia (Eq. (1.2)).

$$\cup_{i=1}^{K} \quad C_i = E, \tag{1.1}$$

$$C_i \neq \emptyset, \quad (1 \leq i \leq K), \tag{1.2}$$

$$C_i \cap C_j = \emptyset, \quad (1 \leq i, j \leq K; i \neq j). \tag{1.3}$$

Algoritmos de particionamento podem ser classificados em duas categorias: rígido (*hard*) e difuso/nebuloso (*fuzzy*). No primeiro caso, cada objeto deve ser atribuído a apenas um *cluster* (Eq. (1.3)) enquanto que, métodos nebulosos atribuem um grau de pertinência de *cluster* a cada objeto (DAS; ABRAHAM; KONAR, 2009; HRUSCHKA et al., 2009). Uma matriz $U_{n \times K}$ pode ser utilizada para representar a pertinência de cada objeto e_i do conjunto E

ao *cluster* C_k. Dessa forma, o elemento u_{ik} da matriz U informa o grau de pertinência do objeto e_i ao *cluster* C_k.

$$U_{n\times K} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1K} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nK} \end{bmatrix}$$

No agrupamento rígido, os elementos u_{ik} da matriz podem ter valor zero (caso o objeto não esteja atribuído ao *cluster* C_k) ou um (caso o objeto pertença ao *cluster* C_k). Portanto, como cada objeto e_i deve pertencer a apenas um *cluster*, a Eq. (1.4) deve ser satisfeita para todos os objetos em E. No agrupamento nebuloso, além da Eq. (1.4), considera-se que os objetos podem pertencer parcialmente aos *clusters* de forma que se tenha graus de pertinência e isso é expresso pela Eq. (1.5).

$$\sum u_{ik} = 1 \quad \forall k, \tag{1.4}$$

$$0 \leq u_{ik} \leq 1 \quad \forall i, k. \tag{1.5}$$

As duas formas de representação de objetos mais comuns que o agrupamento de dados pode se basear são: dados vetoriais e dados relacionais. No caso dos dados vetoriais, cada objeto é descrito por um vetor de valores quantitativos ou qualitativos. Neste caso, o conjunto de dados é representado por uma matriz $n \times p$ de dados, como a matriz abaixo, assumindo que os n objetos possuem p atributos.

$$X = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix}$$

Por outro lado, quando cada par de objetos é representado por uma relação (como uma distância, por exemplo), então os dados são ditos relacionais. Dessa forma, o conjunto de objetos pode ser representado por uma matriz $n \times n$ contendo as descrições relacionais de todos os pares de objetos. Segundo Frigui, Hwang e Rhee (2007), o agrupamento relacional é mais geral no sentido de que é aplicável a situações em que os objetos a serem agrupados não podem ser descritos por valores numéricos. Ainda segundo Frigui, Hwang e Rhee (2007), o agrupamento relacional também é mais mais prático para situações em que a complexidade computacional da distância é alta, quando a medida de distância não possui uma solução de forma fechada ou quando grupos de objetos similares não podem ser representados eficientemente por um único prototipo (CARVALHO; LECHEVALLIER; MELO, 2012; JAIN; DUBES, 1988). Outro aspecto importante na utilização de dados relacionais é que, atributos considerados confidenciais não são disponibilizados, apenas as relações

entre os objetos. Dessa forma, restrições de segurança podem ser preservadas no processo de agrupamento.

$$E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \Leftrightarrow D = \begin{bmatrix} d(e_1, e_1) & \cdots & d(e_1, e_n) \\ \vdots & \ddots & \vdots \\ d(e_n, e_1) & \cdots & d(e_n, e_n) \end{bmatrix}$$

O caso em que os dados são representados por uma única matriz de dissimilaridade é um caso particular e, neste caso, os dados possuem uma única visão relacional (do inglês, *single-view relational data*). Nesta tese, assume-se que a matriz de dissimilaridades satisfaz as restrições apresentadas nas equações (1.6), (1.7) e (1.8).

$$d(e_i, e_j) \geq 0; \quad \forall 1 \leq i, j \leq n, \tag{1.6}$$

$$d(e_i, e_i) = 0; \quad 1 \leq i \leq n, \tag{1.7}$$

$$d(e_i, e_j) = d(e_j, e_i); \quad \forall 1 \leq i, j \leq n. \tag{1.8}$$

Em diversas aplicações práticas, os dados possuem múltiplas representações ou fontes, as quais usualmente contém informações complementares e compatíveis. Estas informações podem ser úteis no agrupamento de dados (WANG et al., 2016; ZHANG et al., 2016; ZHANG et al., 2014; TZORTZIS; LIKAS, 2009; KUMAR; RAI; DAUMÉ III, 2011; YANG; WANG, 2018). No caso em que cada par de objetos pode ser representado por múltiplas relações, existem p matrizes de dissimilaridade e, portanto, os dados possuem múltiplas visualizações relacionais (do inglês, *multi-view relational data*).

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

Cada tipo de visão pode representar diferentes perspectivas dos dados (JIANG; QIU; WANG, 2016). Jiang et al. (2016b) exemplificam que em dados de cursos online abertos massivos (do inglês, *massive open online courses* - MOOCs), os dados de registro de cursos dos alunos descrevem informações demográficas e notas do histórico academico (visão 1), e os dados de comportamento *online* registram os comportamentos de interação durante a aprendizagem (visão 2), tais como postagens em fóruns, leitura das tarefas, leitura em fóruns e submissão de *quizzies*, estes comportamentos refletem o nível de engajamento da aprendizagem. Portanto, é de suma importância integrar essas visões heterogêneas para gerar resultados de agrupamento mais robustos e precisos, ao invés de considerar apenas uma única visão dos dados (LI et al., 2015).

Outro exemplo sobre o uso de dados com múltiplas visões pode ser visto na ciência médica, em que diversas medidas obtidas em uma série de exames médicos são documentadas para cada sujeito, incluindo medidas clínicas, imunológicas, sorológicas e cognitivas, estas medidas sendo obtidas a partir de múltiplas fontes. É desejável combinar todas informações de forma efetiva para se obter um diagnóstico mais preciso de doenças (CAO et al., 2014).

Cai, Nie e Huang (2013) afirmaram que na última década, uma grande quantidade de dados foram e continuam sendo coletados por múltiplas fontes ou representados por múltiplas visões, em que diferentes visões descrevem perspectivas diferentes dos dados. Segundo Xu et al. (2017), Long, Yu e Zhang (2008), o agrupamento eficiente de dados com múltiplas visões é um desafio. Xu, Wang e Lai (2016) aponta que combinar as múltiplas visões de um mesmo conjunto de objetos para obter melhores desempenhos no agrupamento de dados é um problema de pesquisa significativo. Jiang et al. (2016b), Yin et al. (2015) acrescentam que, além do principal desafio ser o de encontrar uma forma de simultaneamente explorar as informações complementares fornecidas por todas as visões, estas podem fornecer conflitos visto que são obtidas por diferentes medidas. Apesar de que cada visão poderia ser usada individualmente para encontrar padrões pelo agrupamento, o desempenho pode ser mais preciso ao explorar a rica quantidade de informação contida nas múltiplas visões (LIU et al., 2013; CHIKHI, 2016; WANG; CHEN, 2017).

Segundo Cleuziou et al. (2009) e (ZHU; ZHU; ZHENG, 2018), as abordagens para agrupamento de dados com múltiplas visões podem ser divididas em três grupos dependendo se a combinação das visões é feita antes, durante ou depois do processo de agrupamento. As três abordagens são: (i) estratégia de concatenação ou fusão, (ii) estratégia distribuída ou (iii) estratégia centralizada. Na primeira, a combinação das visões ocorre antes do processo de agrupamento e todas as visões são concatenadas em uma única visão. No segundo caso, cada visão é utilizada de forma independente para agrupar os dados e depois uma partição final é gerada com base nos vários agrupamentos obtidos. A terceira estratégia utiliza todas as visões simultaneamente no processo de agrupamento e representa um importante desafio. Os métodos desenvolvidos neste trabalho se encaixam na estratégia centralizada.

1.1 COMPLEXIDADE DO PROBLEMA

Conforme Jain e Dubes (1988), o número de partições NP de n objetos em K *clusters* é definido na Eq. (1.9). Esse número cresce consideravelmente quando o tamanho do conjunto de dados e o número de *clusters* aumentam.

$$NP(n, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^n. \tag{1.9}$$

Para ilustrar a complexidade do problema, ao particionar um pequeno conjunto de dados contendo 19 objetos em 4 *clusters*, o número de partições possíveis é 11,259,666,950

(MURTY; DEVI, 2015). Portanto, devido ao elevado número de possíveis partições à medida que o conjunto de dados e o número de *clusters* crescem, a enumeração exaustiva de todas as possíveis partições para encontrar a partição ótima global é um problema NP-Completo (MAIMON; ROKACH, 2005; MURTY; DEVI, 2015; MURTY; DEVI, 2011; JAIN; DUBES, 1988).

1.2 JUSTIFICATIVA

Ao invés de enumerar exaustivamente todas as possíveis partições para alcançar a partição ótima global, métodos heurísticos podem ser usados para obter soluções de boa qualidade em tempo computacionalmente aceitável. Em particular, metaheurísticas são métodos que coordenam soluções no espaço de busca tentando evitar soluções que representam ótimos locais para resolver problemas de otimização genéricos (GENDREAU; POTVIN, 2010). Uma vez que o problema do agrupamento de dados pode ser modelado matematicamente como um problema de otimização combinatória, metaheurísticas se tornam uma opção interessante e importante para gerar partições sub-ótimas do conjunto de dados.

A Otimização por Nuvem de Partículas (do inglês, *Particle Swarm Optimization* - PSO) é uma metaheurística baseada em inteligência de enxame importante devido a sua simplicidade e versatilidade (KENNEDY; EBERHART, 1995; CLERC; KENNEDY, 2002; TRELEA, 2003). Isto motivou diversos pesquisadores a proporem algoritmos baseados em PSO para agrupamento rígido e nebuloso de dados vetoriais. Métodos de agrupamento baseados em PSO mostraram-se importantes ferramentas e apresentaram resultados melhorados quando comparados aos algoritmos de agrupamento tradicionais como *K-means*, *K-medoids* e *Fuzzy c-means* (RANA; JASOLA; KUMAR, 2011).

Como apontado por Ding et al. (2011), otimização por inteligência de enxame é uma técnica de inteligência artificial inovadora para otimização. Algoritmos de inteligência de enxame são modelos matemáticos inspirados no comportamento de diversos enxames de animais e insetos sociais que compartilham informações. PSO foi introduzido há mais de duas décadas e tem sido aplicada para solucionar problemas de otimização em diversas áreas. Ainda segundo Ding et al. (2011) e Bharne, Gulhane e Yewale (2011), algoritmos de inteligência de enxame são métodos de busca eficientes, adaptativos e robustos que conseguem produzir soluções próximas da solução ótima e possuem grande quantidade de paralelismo implícito. Por outro lado, o agrupamento de dados pode ser bem formulado como um problema de otimização global difícil e, portanto, torna a aplicação de algoritmos de inteligência de enxame uma opção promissora e apropriada.

Conforme Frigui, Hwang e Rhee (2007), o agrupamento de dados vetoriais tem sido um campo de pesquisa muito estudado e diversos métodos foram propostos para solucionar o problema. Por outro lado, o agrupamento de dados relacionais recebeu pouca atenção, mesmo quando diversas aplicações podem se beneficiar de algoritmos de agrupamento relacional. Por exemplo, na recuperação de imagens baseada em conteúdo ou mineração de dados na web, foi mostrado que as medidas de dissimilaridade mais efetivas não possuem

uma forma fechada. Portanto, essas medidas não poderiam ser usadas no agrupamento de dados vetoriais.

Segundo Frigui e Hwang (2008), outra questão importante no agrupamento de dados complexos é que a relação entre objetos pode ser descrita por múltiplas matrizes de dissimilaridade (do inglês, *multi-view data*). Um exemplo do uso de múltiplas dissimilaridades é na categorização de bancos de dados de imagens, em que é possível ter uma matriz de dissimilaridade para informações de cor, outra matriz para informações sobre textura e outra matriz para estrutura da informação. De acordo com Long, Yu e Zhang (2008), dados com múltiplas visões fazem emergir um novo problema natural e ainda não completamente padronizado: como aprender um padrão consensual a partir de múltiplas representações, tal que seja mais preciso e robusto do que os padrões baseados em uma única visão? Devido ao enorme impacto dos dados com múltiplas visões em muitas aplicações, o aprendizado com múltiplas visões está atraindo mais e mais atenção.

Ainda de acordo com Frigui, Hwang e Rhee (2007) e Frigui e Hwang (2008), a maior parte dos algoritmos existentes para agrupamento de dados relacionais pode operar considerando apenas uma única matriz de dissimilaridade. Dessa forma, para particionar dados com múltiplas visões, seria necessário particionar os objetos considerando cada matriz separadamente ou particionar os objetos usando uma única matriz que combinasse uniformemente todas as outras. Contudo, a influência de cada matriz pode não ser igualmente importante para a definição dos grupos aos quais os objetos pertencem. Portanto, para obter *clusters* significativos considerando todas as matrizes, é necessário estimar os pesos de relevância para cada matriz de dissimilaridade.

Portanto, a justificativa em trabalhar com o problema do agrupamento de dados relacionais com múltiplas visões nesta tese se deve ao fato de ser um problema emergente, relativamente novo e pouco explorado, desafiador e com grande impacto em diversas aplicações do mundo real. Consequentemente, novos métodos de agrupamento de dados relacionais com múltiplas visões são necessários, visto que essa área recebeu muito pouca atenção e a modelagem de métodos híbridos de agrupamento relacional baseados em inteligência de enxame, como PSO, se torna uma opção importante a ser explorada e promissora devido às suas características para problemas de otimização e seu desempenho já avaliado em diversos trabalhos aplicados a problemas de otimização de forma geral e também no agrupamento de dados vetoriais. Além disso, o PSO possui características interessantes como: simplicidade, facilidade de adaptação para processamento paralelo e convergência global. Estas características também motivaram o uso dos conceitos desse modelo para composição dos métodos híbridos desenvolvidos nesta tese.

1.3 DEFINIÇÃO DO PROBLEMA

Baseado neste contexto, a principal questão de pesquisa investigada nesta tese é:

Problema da pesquisa *O agrupamento rígido e nebuloso de dados relacionais com múltiplas visões pode ser melhorado de forma significativa usando métodos híbridos baseados na otimização de enxame?*

Com o objetivo de responder essa pergunta, é necessário entender as abordagens atuais na literatura, escolher os modelos mais adequados que podem servir de base para o desenvolvimento dos modelos híbridos e avaliar de forma coerente os resultados encontrados pelos modelos propostos. Dessa forma, o objetivo do trabalho descrito nesta tese pode ser formulado como:

Objetivo da pesquisa *Investigar e desenvolver modelos híbridos baseados em otimização por nuvem de partículas para agrupamento rígido e nebuloso de dados relacionais com múltiplas visões.*

Objetivo específicos

Para atingir o objetivo geral deste trabalho, os seguintes objetivos específicos foram definidos:

1. *Revisar a literatura sobre agrupamento rígido e nebuloso de dados relacionais com visão única e com várias visões*;
2. *Desenvolver modelo híbrido de agrupamento rígido de dados relacionais com única visão usando otimização por nuvem de partículas* ;
3. *Desenvolver modelo híbrido de agrupamento rígido de dados relacionais com múltiplas visões usando otimização por nuvem de partículas* ;
4. *Realizar estudo sobre quais funções objetivo podem ser mais adequadas para agrupamento rígido de dados relacionais com múltiplas visões*;
5. *Desenvolver modelo híbrido de agrupamento nebuloso de dados relacionais com única visão usando otimização por nuvem de partículas* ;
6. *Desenvolver modelo híbrido de agrupamento nebuloso de dados relacionais com múltiplas visões usando otimização por nuvem de partículas* ;
7. *Realizar estudo sobre quais funções objetivo podem ser mais adequadas para agrupamento nebuloso de dados relacionais com múltiplas visões*;
8. *Comparar desempenho dos métodos desenvolvidos com algoritmos da literatura usando índices externos para validação dos agrupamentos gerados.*

1.4 CONTRIBUIÇÕES

As principais contribuições desta tese são:

1. Uma revisão sobre agupamento de dados relacionais com múltiplas visões, com ênfase na utilização de múltiplas matrizes;

2. Modelagem de método híbrido baseado em otimização por nuvem de partículas para agrupamento rígido de dados relacionais com única visão. Publicado na conferência *IEEE International Conference on Systems, Man and Cybernetics* (GUSMãO; CARVALHO, 2016);

3. Modelagem de métodos híbridos baseados na otimização por nuvem de partículas para agrupamento rígido de dados relacionais com múltiplas visões. Publicado na revista *Expert Systems with Applications* (GUSMãO; CARVALHO, 2019);

4. Modelagem de métodos híbridos baseados na otimização por nuvem de partículas para agrupamento nebuloso de dados relacionais com múltiplas visões. Submetido na revista *International Journal of Pattern Recognition and Artificial Intelligence*;

5. Análise, adaptação e estudo comparativo de diversos índices para validação de agrupamento rígido e nebuloso de dados relacionais com múltiplas visões; e

6. Validação experimental das abordagens desenvolvidas através de seis estudos empíricos usando bases de dados reais.

1.5 ORGANIZAÇÃO DO TRABALHO

O capítulo 2 apresenta trabalhos relacionados envolvendo agrupamento de dados vetoriais e relacionais com visão única bem como com várias visões. O capítulo 3 apresenta uma fundamentação teórica contendo os principais conceitos e algoritmos que serão utilizados nesta tese. O capítulo 4 apresenta alguns modelos para agrupamento de dados baseados em PSO. O capítulo 5 apresenta os métodos desenvolvidos. O capítulo 6 apresenta os índices para validação de agrupamentos que foram utilizados nesta tese. O capítulo 7 apresenta os resultados referentes aos algoritmos de agrupamento rígidos e o capítulo 8 apresenta os resultados referentes aos algoritmos de agrupamento nebulosos. Por fim, o capítulo 9 apresenta as considerações finais desta tese.

2 MÉTODOS DE AGRUPAMENTO DE DADOS RELACIONAIS

Neste capítulo serão apresentados os principais conceitos e alguns algoritmos de agrupamento rígido e nebuloso de dados relacionais com apenas uma visão (*single-view*) ou múltiplas visões (*multi-view*) necessários ao entendimento desta tese. Na seção 2.1, alguns métodos para agrupamento rígido são apresentados. Na seção 2.2, alguns métodos para agrupamento nebuloso são apresentados.

2.1 AGRUPAMENTO RÍGIDO

Nesta seção serão descritos alguns algoritmos para agrupamento rígido de dados relacionais. Os algoritmos descritos a seguir são: Hard c-Medoids (KRISHNAPURAM; FREG, 1992), Algoritmo rígido baseado em múltiplas matrizes de dissimilaridades, Algoritmo rígido baseado em múltiplas matrizes de dissimilaridades com pesos de relevância estimados localmente e Algoritmo rígido baseado em múltiplas matrizes de dissimilaridades com pesos de relevância estimados globalmente (CARVALHO; LECHEVALLIER; MELO, 2012).

2.1.1 Hard c-Medoids

Seja $E = \{e_1, ..., e_n\}$ um conjunto de n objetos e seja D uma matriz de dissimilaridades $D = [d(e_i, e_o)]$, onde $d(e_i, e_o)$ representa a dissimilaridade entre os objetos e_i e e_o $(i, o = 1, ..., n)$.

$$E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \Leftrightarrow D = \begin{bmatrix} d(e_1, e_1) & \cdots & d(e_1, e_n) \\ \vdots & \ddots & \vdots \\ d(e_n, e_1) & \cdots & d(e_n, e_n) \end{bmatrix}$$

Este algoritmo é a versão rígida do algoritmo nebuloso proposto por (KRISHNAPURAM; JOSHI; YI, 1999). Este método assume que o representativo de um *cluster* C_k é um objeto do conjunto de objetos E, ou seja, $e_j{}' \in E(j = 1, .., K)$. O algoritmo procura por uma partição $\mathcal{P} = (C_1, ..., C_K)$ de E em K *clusters* e pelos representativos $(e_1^{'}, ..., e_K^{'})$ correspondentes de cada *cluster* em $\mathcal{P}$ de forma que otimize a função objetivo, a qual mede a adequação entre os objetos em cada grupo e os representativos.

Neste algoritmo, J é o critério de adequação (também chamado de função objetivo) dos objetos aos *clusters*, e é definido como

$$J = \sum_{k=1}^{K} \sum_{e_i \in C_k} d(e_i, e_k^{'}). \tag{2.1}$$

Dessa forma, o objetivo é encontrar uma partição com grupos homogêneos, isto é, em que os objetos estejam o mais próximo possível dos representativos. Os passos do *Hard*

c-medoids são definidos conforme o Algoritmo 1. Na fase de inicialização, os representativos são escolhidos aleatoriamente para representar os *clusters* $(C_1, .., C_K)$ e os objetos são atribuídos aos *clusters* conforme a menor distância entre eles e os representativos. Dessa forma, a partição inicial é gerada.

Após a inicialização, no Passo 1, ocorre a atualização dos representativos de cada *cluster*. Dessa forma, o novo representativo do *cluster* C_1, por exemplo, será o objeto que estiver mais próximo de todos os objetos atribuídos ao *cluster*. No Passo 2, uma nova partição é gerada de acordo com as novas atribuições dos objetos aos *clusters* conforme a menor distância entre cada objeto e os novos representativos. Ao término da definição da nova partição, se algum objeto tiver sido atribuído a um *cluster* diferente do que estava no passo anterior, o algoritmo retorna ao passo 1.

Consequentemente, o algoritmo é finalizado quando não é mais possível movimentar os objetos, pois a solução representa um ponto de mínimo local.

Algoritmo 1: Hard c-Medoids

Inicialização
  Escolher K objetos aleatórios distintos para o conjunto inicial de representativos.
  Atribuir e_i ao *cluster* $C_k^{(0)}(k = 1, .., K)$ se $e_k^{'}$ for o representativo mais próximo.
  Dessa forma, a partição inicial $\mathcal{P}^{(0)} = (C_1^{(0)}, ..., C_K^{(0)})$ é obtida.

Passo 1: Definição dos melhores representativos
  $t \leftarrow t + 1$;
  A partição $\mathcal{P}^{(t-1)} = (C_1^{(t-1)}, ..., C_K^{(t-1)})$ é fixada;
  Computar o novo representativo $e^* \in E$ do *cluster* $C_k^{(t-1)}(k = 1, .., K)$:
    $argmin_{e_j \in E} \sum_{e_i \in C_k} d(e_i, e_j)$

Passo 2: Definição da melhor partição
  $test \leftarrow 0$;
  $\mathcal{P}^{(t)} \leftarrow \mathcal{P}^{(t-1)}$
  Para $i = 1$ até n **faça**
    Encontrar o *cluster* $C_m^{(t)}$ que e_i pertence
    Encontrar o *cluster* $C_k^{(t)}$ de forma que
      $k = argmin_{1 \leq h \leq K} d(e_i, e_h)$.
    Se $k \neq m$ **então**
      $test \leftarrow 1$;
      $C_k^{(t)} = C_k^{(t)} \cup \{e_i\}$;
      $C_m^{(t)} = C_m^{(t)} \setminus \{e_i\}$;

  Se $test = 0$, **então** PARE;
  Senão vá para o Passo 1.

2.1.2 Agrupamento rígido baseado em múltiplas matrizes

Seja $E = \{e_1, ..., e_n\}$ um conjunto de n objetos e sejam p matrizes de dissimilaridades $D_j = [d_j(e_i, e_o)]$, onde $d_j(e_i, e_o)$ representa a dissimilaridade entre os objetos e_i e e_o $(i, o = 1, ..., n)$ da matriz D_j.

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

O método de agrupamento rígido baseado em múltiplas matrizes, denominado MRDCA, assume que o representativo de um *cluster* C_k é um subconjunto de cardinalidade fixa $1 \leq q << n$ do conjunto de objetos E, isto é, $G_k \in E^{(q)} = \{A \subset E : |A| = q\}$ (CARVALHO; LECHEVALLIER; MELO, 2012). O algoritmo HCMdd é um caso particular tendo apenas uma matriz de dissimilaridade e $q = 1$, ou seja, apenas um objeto como representante por *cluster*. O algoritmo procura por uma partição $\mathcal{P} = (C_1, ..., C_K)$ de E em K *clusters* e os representativos correspondentes $(G_1, ..., G_K)$ representando os clusters em $\mathcal{P}$ de forma que otimize a função objetivo. Para esse algoritmo, a função objetivo também mede a adequação entre os representativos de cada *cluster* e os objetos.

Considerando p matrizes, o MRDCA leva em consideração simultaneamente essas p matrizes de dissimilaridades. Para tanto, o critério de adequação é modificado para

$$J = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} D_j(e_i, G_k) = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} \sum_{e \in G_k} d_j(e_i, e) \tag{2.2}$$

A função objetivo utilizada nesse algoritmo considera que todas matrizes de dissimilaridades possuem o mesmo peso no processo de agrupamento de dados. Além disso, o algoritmo MRDCA irá diferir do algoritmo HCMdd da seguinte forma:

- Na inicialização, em que cada objeto e_i será atribuído ao *cluster* cujo representativo G_k for mais próximo ao objeto. Portanto, o cálculo deve considerar as p matrizes e os elementos de G_k, ou seja, $\sum_{j=1}^{p} \sum_{e \in G_k} d_j(e_i, e)$

- No Passo 1, o cálculo do representativo passará a ser feito como definido na Eq. (2.3). Nesse caso, serão selecionados os q objetos que estiverem mais próximos aos objetos de C_k considerando as dissimilaridades de todas p as matrizes.

$$argmin_{G \in E^{(q)}} \sum_{e_i \in C_k} \sum_{j=1}^{p} \sum_{e \in G} d_j(e_i, e). \tag{2.3}$$

- No Passo 2, para encontrar o *cluster* vencedor, o cálculo passará a ser definido de acordo com a Eq. (2.4). Nesse caso, o grupo vencedor será aquele que possuir os representativos mais próximos a cada objeto.

$$k = argmin_{1 \leq h \leq K} \sum_{j=1}^{p} \sum_{e \in G_h} d_j(e_i, e). \tag{2.4}$$

2.1.3 Agrupamento rígido baseado em múltiplas matrizes com pesos de relevância

A abordagem do algoritmo MRDCA pode não ser adequada, pois a influência de cada matriz de dissimilaridade nem sempre é igualmente importante para definição de cada *cluster*. Uma estratégia alternativa para tratar esse problema é atribuir um peso de relevância para cada matriz de dissimilaridades. Essa estratégia tem inspiração na abordagem usada para calcular o peso de relevância de cada variável em cada *cluster* no algoritmo de agrupamento dinâmico baseado em distâncias adaptativas. Essa abordagem é considerada no método MRDCA-RWL.

Além da partição e representativos, este algoritmo também tem como saída os pesos de relevância estimados localmente, ou seja, os pesos de relevância são diferentes para cada matriz de dissimilaridades e também diferem para cada *cluster*. O critério de adequação passa a ser definido como

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

Os pesos de relevância de cada matriz de dissimilaridades estimados localmente foram propostos por Carvalho, Lechevallier e Melo (2012), e são calculados usando o método multiplicador de Lagrange sob as restrições $\lambda_{kj} > 0$ e $\prod_{j=1}^{p} \lambda_{kj} = 1$. Os pesos são definidos na Eq. (2.6). Os pesos foram definidos dessa forma para minimizar o valor da função objetivo tendo a partição e representativos fixados.

$$\lambda_{kj} = \frac{\{\prod_{h=1}^{p} [\sum_{e_i \in C_k} D_h(e_i, G_k)]\}^{\frac{1}{p}}}{[\sum_{e_i \in C_k} D_j(e_i, G_k)]} = \frac{\{\prod_{h=1}^{p} [\sum_{e_i \in C_k} \sum_{e \in G_k} d_h(e_i, e)]\}^{\frac{1}{p}}}{[\sum_{e_i \in C_k} \sum_{e \in G_k} d_j(e_i, e)]} \tag{2.6}$$

Os passos desse método são descritos no Algoritmo 2. Note que, em relação ao algoritmo MRDCA, o passo para o cálculo do vetor de pesos de relevância foi adicionado. Além disso, os passos de inicialização, definição dos melhores representativos e definição da melhor partição também foram alterados. Na inicialização, o vetor de pesos de relevância é inicializado com todos os elementos iguais a 1, ou seja, as matrizes de dissimilaridades possuem mesma relevância na geração da partição inicial. Os Passos 1 e 3 são alterados para considerar os pesos de relevância no cálculo dos representativos e no momento de atribuir cada objeto a um *cluster*.

O algoritmo MRDCA-RWL apresenta problemas com instabilidades númericas no cálculo do peso de relevância de cada matriz de dissimilaridade quando, por exemplo, o

Algoritmo 2: MRDCA-RWL e MRDCA-RWG

Inicialização
  $t \leftarrow 0$
  $MRDCA_{RWL}$: Inicializar $\lambda_k^{(0)} = (\lambda_{k1}^{(0)}, ..., \lambda_{kp}^{(0)}) = (1, ..., 1)(k = 1, ..., K)$
  $MRDCA_{RWG}$: Inicializar $\lambda^{(0)} = (\lambda_1^{(0)}, ..., \lambda_p^{(0)}) = (1, ..., 1)$
  Selecionar aleatoriamente K prototipos distintos $G_k^{(0)} \in E^{(q)}$
  Atribuir cada objeto e_i ao prototipo mais próximo para obter a partição inicial $\mathcal{P}^{(0)} = (C_1^{(0)}, ..., C_K^{(0)})$
Passo 1: definição dos melhores representativos
  $t \leftarrow t + 1$;
  A partição $\mathcal{P}^{(t-1)} = (C_1^{(t-1)}, ..., C_K^{(t-1)})$ e os pesos de relevância são fixados;
  $MRDCA_{RWL}$: computar $G_k^{(t)} = G^* \in E^{(q)}$ of *cluster* $C_k^{(t-1)}(k = 1, .., K)$: $argmin_{G \in E^{(q)}} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_{kj}^{(t-1)} \sum_{e \in G} d_j(e_i, e)$
  $MRDCA_{RWG}$: computar $G_k^{(t)} = G^* \in E^{(q)}$ of *cluster* $C_k^{(t-1)}(k = 1, .., K)$: $argmin_{G \in E^{(q)}} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_j^{(t-1)} \sum_{e \in G} d_j(e_i, e)$
    $argmin_{G \in E^{(q)}} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_{kj}^{(t-1)} \sum_{e \in G} d_j(e_i, e)$
Passo 2: cálculo do vetor de pesos de relevância
  Os pesos são calculados de acordo com a Eq. 2.6 se os pesos forem estimados localmente; ou de acordo com a Eq. 2.8 se forem estimados globalmente
Passo 3: definição da melhor partição
  Os representativos e os vetores de peso de relevância são fixados.
  $test \leftarrow 0$;
  $\mathcal{P}^{(t)} \leftarrow \mathcal{P}^{(t-1)}$
  Para $i = 1$ até n **faça**
    Encontrar o *cluster* $C_m^{(t)}$ que e_i pertence
    $MRDCA_{RWL}$: encontrar o *cluster* $C_k^{(t)}$ da seguinte forma:
      $k = argmin_{1 \leq h \leq K} \sum_{j=1}^{p} \lambda_{hj}^{(t)} \sum_{e \in G_h^{(t)}} d_j(e_i, e)$.
    $MRDCA_{RWG}$: encontrar o *cluster* $C_k^{(t)}$ da seguinte forma:
      $k = argmin_{1 \leq h \leq K} \sum_{j=1}^{p} \lambda_j^{(t)} \sum_{e \in G_h^{(t)}} d_j(e_i, e)$.
    Se $k \neq m$ **então**
      $test \leftarrow 1$;
      $C_k^{(t)} = C_k^{(t)} \cup \{e_i\}$;
      $C_m^{(t)} = C_m^{(t)} \setminus \{e_i\}$;
  Se $test = 0$, então PARE;
  Senão vá para o Passo 1

algoritmo produz um único *cluster* ou *clusters* com alguns objetos que possuem dissimilaridade zero entre eles (CARVALHO; LECHEVALLIER; MELO, 2012). Para evitar esse problema, Carvalho, Lechevallier e Melo (2012) propuseram o MRDCA-RWG, em que o peso de relevância para cada matriz muda nas iterações do algoritmo, mas o peso de relevância é o mesmo para todos os *clusters*, ou seja, o cálculo dos pesos de relevância é feito globalmente e que difere do algoritmo MRDCA-RWL, pois este último realiza o cálculo localmente para cada grupo.

O critério de adequação passa a ser definido como na Eq. (2.7) e os pesos estimados globalmente passam a ser calculados de acordo com a Eq. (2.8).

$$J = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_j D_j(e_i, G_k) = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_j \sum_{e \in G_k} d_j(e_i, e) \tag{2.7}$$

$$\lambda_j = \frac{\{\prod_{h=1}^{p} (\sum_{k=1}^{K} [\sum_{e_i \in C_k} D_h(e_i, G_k)])\}^{\frac{1}{p}}}{\sum_{k=1}^{K} [\sum_{e_i \in C_k} D_j(e_i, G_k)]} = \frac{\{\prod_{h=1}^{p} (\sum_{k=1}^{K} [\sum_{e_i \in C_k} \sum_{e \in G_k} d_h(e_i, e)])\}^{\frac{1}{p}}}{\sum_{k=1}^{K} [\sum_{e_i \in C_k} \sum_{e \in G_k} d_j(e_i, e)]}$$

(2.8)

A partir da partição inicial, este algoritmo de agrupamento alterna entre três passos e encerra quando não ocorre mais movimentação dos objetos entre os grupos. Os passos desse método também são descritos no Algoritmo 2. A mudança em relação ao MRDCA-RWL se dá essencialmente no vetor de pesos e no cálculo dos mesmos, mas que tem impacto no agrupamento dos dados.

2.2 AGRUPAMENTO NEBULOSO

Nesta seção serão descritos alguns algoritmos para agrupamento nebuloso de dados relacionais. Os algoritmos descritos a seguir são: Fuzzy c-Medoids (SFCMdd), Algoritmo nebuloso baseado em múltiplas matrizes de dissimilaridades, Algoritmo nebuloso baseado em múltiplas matrizes de dissimilaridades com pesos de relevância.

2.2.1 Agrupamento nebuloso baseado em uma matriz de dissimilaridade

De acordo com (CARVALHO; LECHEVALLIER; MELO, 2013), o critério de adequação usado no SFCMdd mede a homogeneidade da partição nebulosa como a soma das homogeneidades em cada *cluster* nebuloso. Este algoritmo nebuloso para dados relacionais procura por uma partição $P = (C_1, ..., C_K)$ do conjunto E em K *clusters* nebuloso representados por $U = (u_1, ..., u_n)$, em que $u_i = (u_{i1}, ..., u_{iK})$ $(i = 1, .., n)$, e os representativos correspondentes $G = (G_1, ..., G_K)$ de tal forma que o critério de adequação J é otimizado.

Os passos do SFCMdd são definidos no Alg. 3. O algoritmo define uma partição nebulosa inicial e alterna entre os Passos 1 e 2 até a convergência. O critério de adequação passa a ser definido como na Eq. (2.9).

$$J = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D(e_i, G_k) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \sum_{e\in G_k} d(e_i, e) \tag{2.9}$$

Algoritmo 3: SFCMdd

Inicialização
  $t \leftarrow 0$
  Selecionar aleatoriamente K prototipos distintos $G_k^{(0)} \in E^{(q)}$
  Para cada objeto e_i, computar os seus graus de pertinência $u_{ik}^{(0)}$ $(k = 1, .., K)$:
  $u_{ik}^{(0)} = \left[\sum_{h=1}^{K}\left(\frac{D(e_i,G_k^{(0)})}{D(e_i,G_h^{(0)})}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{h=1}^{K}\left(\frac{\sum_{e\in G_k^{(0)}} d(e_i,e)}{\sum_{e\in G_h^{(0)}} d(e_i,e)}\right)^{\frac{1}{m-1}}\right]^{-1}$
  Computar:
  $J^{(0)} = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik}^{(0)})^m D(e_i, G_k^{(0)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik}^{(0)})^m \sum_{e\in G_k^{(0)}} d(e_i, e)$
Passo 1: definição dos melhores representativos
  $t \leftarrow t + 1$;
  A partição nebuloso representada por $U^{(t-1)} = (u_1^{(t-1)}, ..., u_n^{(t-1)})$ é fixada;
  Computar $G_k^{(t)}$ do *cluster* nebuloso $C_k^{(t-1)}$ $(k = 1, .., K)$ de acordo com:
  $G_k \leftarrow \emptyset$
  Faça
    Encontrar $e_l \in E$, $e_l \notin G^*$ de forma que
    $l = argmin_{1\leq h\leq n}\sum_{i=1}^{n}(u_{ik})^m d(e_j, e_h)$
    $G_k \leftarrow G_k \cup e_l$
  Enquanto $|G_k| < q$
Passo 2: definição da melhor partição nebuloso
  Os representativos $G^{(t)} = (G_1^{(t)}, .., G_K^{(t)})$ são fixados.
  Computar o grau de pertinência $u_{ik}^{(t)}$ do objeto e_i $(i = 1, .., n)$ no *cluster* nebuloso C_k $(k = 1, .., K)$ de acordo com:
  $u_{ik}^{(t)} = \left[\sum_{h=1}^{K}\left(\frac{D(e_i,G_k^{(t)})}{D(e_i,G_h^{(t)})}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{h=1}^{K}\left(\frac{\sum_{e\in G_k^{(t)}} d(e_i,e)}{\sum_{e\in G_h^{(t)}} d(e_i,e)}\right)^{\frac{1}{m-1}}\right]^{-1}$
Critério de parada
  Computar:
  $J^{(t)} = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik}^{(t)})^m D(e_i, G_k^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik}^{(t)})^m \sum_{e\in G_k^{(t)}} d(e_i, e)$
  Se $|J^{(t)} - J^{(t-1)}| \leq \epsilon$ ou $t > T$, **então** PARE;
  Senão vá para o Passo 1

2.2.2 Agrupamento nebuloso baseado em mútiplas matrizes de dissimilaridades

Considerando p matrizes, o algoritmo denotado por MFCMdd leva em consideração simultaneamente p matrizes de dissimilaridades. Para tanto, o critério de adequação é

modificado conforme definido na Eq. (2.10).

$$J = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D(e_i, G_k) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \sum_{e\in G_k}\sum_{j=1}^{p} d_j(e_i, e) \tag{2.10}$$

O algoritmo MFCMdd irá diferir do algoritmo SFCMdd da seguinte maneira:

- No Passo 1, o cálculo do representativo passará a considerar as dissimilaridades de todas as matrizes simultaneamente, isto é,

$$l = argmin_{1\leq h\leq n} \sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}^{p} d_j(e_i, e_h).$$

- No Passo 2, o grau de pertinência do objeto e_i ao *cluster* nebuloso também é modificado para considerar as dissimilaridades contidas em todas as matrizes, ou seja,

$$u_{ik}^{(t)} = \left[\sum_{h=1}^{K}\left(\frac{D(e_i, G_k^{(t)})}{D(e_i, G_h^{(t)})}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{h=1}^{K}\left(\frac{\sum_{j=1}^{p}\sum_{e\in G_k^{(t)}} d_j(e_i, e)}{\sum_{j=1}^{p}\sum_{e\in G_h^{(t)}} d_j(e_i, e)}\right)^{\frac{1}{m-1}}\right]^{-1}.$$

De acordo com Carvalho, Lechevallier e Melo (2013), esta abordagem é similar àquela que consiste na agrupamento do conjunto de objetos baseado em uma matriz de dissimilaridades global, a qual considera que as p matrizes possem igual importância no processo de agrupamento. Como apontado por Frigui, Hwang e Rhee (2007), considerar múltiplas matrizes com igual relevância pode não ser uma estratégia efetiva, uma vez que a influência de cada matriz de dissimilaridade pode não ser igual na definição dos *clusters*.

2.3 ALGORITMO DE AGRUPAMENTO NEBULOSO BASEADO EM MÚTIPLAS MATRIZES COM PESOS DE RELEVÂNCIA

Segundo Carvalho, Lechevallier e Melo (2013), o método denotado por MFCMdd-RWL foi desenvolvido para fornecer a partição nebuloso e um prototipo para cada *cluster* nebuloso. Além disso, também foi projetado para aprender um peso de relevância para cada matriz de dissimilaridade, estes pesos mudam ao longo das iterações e são diferentes para cada *cluster* nebuloso.

O critério de adequação passa a ser definido como

$$J = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D_{(\lambda_k, s)}(e_i, G_k). \tag{2.11}$$

Ainda segundo Carvalho, Lechevallier e Melo (2013), duas funções de correspondência com pesos de relevância para cada matriz de dissimilaridades estimados locamente são considerados a depender se a soma dos pesos é igual a um (inspirado da computação dos graus de pertinência de um objeto a um *cluster* nebuloso (BEZDEK, 1981)) ou se o produto dos pesos é igual a um (inspirado da computação de um peso de relevância para cada

variável em um *cluster* do *framework* do algoritmo de agrupamento dinamico baseado em distâncias adaptativas (DIGAY; GOVAERT, 1977)).

A função de correspondência, definida na Eq. (2.12), é parametrizada pelo vetor de pesos de relevância $\lambda_k = (\lambda_{k1}, ..., \lambda_{kp})$, em que $\lambda_{kj} > 0$ e $\prod_{j=1}^{p} \lambda_{kj} = 1$, e associado com o *cluster* C_k $(k = 1, ..., K)$

$$D_{\lambda_k}(e_i, G_k) = \sum_{j=1}^{p} \lambda_{kj} D_j(e_i, G_k) = \sum_{j=1}^{p} \lambda_{kj} \sum_{e \in G_k} d_j(e_i, e) \tag{2.12}$$

Dessa forma, o peso de relevância estimado localmente é definido de acordo com a Eq. (2.13).

$$\lambda_{kj} = \frac{\{\prod_{h=1}^{p}[\sum_{l=1}^{n}(u_{lk})^m D_h(e_l, G_k)]\}^{\frac{1}{p}}}{[\sum_{l=1}^{n}(u_{lk})^m D_j(e_l, G_k)]} = \frac{\{\prod_{h=1}^{p}[\sum_{l=1}^{n}(u_{lk})^m \sum_{e \in G_k} d_h(e_l, e)]\}^{\frac{1}{p}}}{[\sum_{l=1}^{n}(u_{lk})^m \sum_{e \in G_k} d_j(e_l, e)]} \tag{2.13}$$

Em Carvalho, Lechevallier e Melo (2012), a função de correspondência também é definida sendo parametrizada por um parâmetro $1 < s < \infty$, além de considerar $\lambda_{kj} \in [0, 1]$ e $\sum_{j=1}^{p} \lambda_{kj} = 1$. Contudo, essa definição não foi considerada neste trabalho.

O grau de pertinencia de cada objeto passa a ser atualizado de acordo com a Eq. (2.14).

$$u_{ik}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{D_{\lambda_k^{(t)}}(e_i, G_k^{(t)})}{D_{\lambda_k^{(t)}}(e_i, G_h^{(t)})}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \lambda_{kj}^{(t)} \sum_{e \in G_k^{(t)}} \sum_{j=1}^{p} d_j(e_i, e)}{\sum_{j=1}^{p} \lambda_{kj}^{(t)} \sum_{e \in G_h^{(t)}} \sum_{j=1}^{p} d_j(e_i, e)}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{2.14}$$

2.3.1 Agrupamento nebuloso baseado em mútiplas matrizes com pesos de relevância calculados globalmente

Como apontado por (CARVALHO; LECHEVALLIER; MELO, 2013), o algoritmo de MFCMdd-RWL apresentado na subseção anterior apresenta instabilidades numéricas (como divisão por zero) no cálculo do peso de relevância de cada matriz de dissimilaridade em cada *cluster* nebuloso quando o algoritmo produz *clusters* nebuloso de tal forma que $\sum_{i=1}^{n}(u_{ik})^m D_j(e_i, G_k) \to 0$. Para diminuir significativamente a chance disso ocorrer, (CARVALHO; LECHEVALLIER; MELO, 2013) propuseram o MFCMdd-RWG, este aprende um peso de relevância para cada matriz de dissimilaridades e que muda ao longo das iterações, mas é o mesmo peso para todos os *clusters* nebulosos.

O critério de adequação passa a ser definido como

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m D_\lambda(e_i, G_k) \tag{2.15}$$

Assim como no algoritmo MFCMdd-RWL, apresentado anteriormente, duas funções de correspondência são consideradas a depender se a soma dos pesos é igual a um ou se o produto dos pesos é igual a um. A função de correspondência é:

Algoritmo 4: Agrupamento nebuloso baseado em múltiplas matrizes com pesos de relevância

Inicialização
  Inicializar $t = 0$ e $\lambda_k^{(0)} = (\lambda_{k1}^{(0)}, ..., \lambda_{kp}^{(0)}) = (1, ..., 1)$ $(k = 1, .., K)$
  Selecionar aleatoriamente K prototipos distintos $G_k^{(0)} \in E^{(q)}$
  Computar o grau de pertinência $u_{ik}^{(t)}$ do objeto e_i $(i = 1, .., n)$ no *cluster* nebuloso C_k $(k = 1, .., K)$ de acordo com (2.14) para MFCMdd-RWL ou (2.18) para MFCMdd-RWG
  Computar J de acordo com (2.13) para MFCMdd-RWL ou de acordo com (2.17) para MFCMdd-RWG
Passo 1: definição dos melhores representativos
  Atualizar $t = t + 1$;
  A partição nebuloso representada por $U^{(t-1)} = (u_1^{(t-1)}, ..., u_n^{(t-1)})$ é fixada;
  Computar o representativo $G_k^{(t)} = G^* \in E^{(q)}$ do *cluster* nebuloso $C_k^{(t-1)}$ $(k = 1, .., K)$ de acordo com:
  $G_k \leftarrow \emptyset$
  Faça
    Encontrar $e_l \in E$, $e_l \notin G^*$ de forma que
    $l = argmin_{1 \leq h \leq n} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (\lambda_{kj})^s d_j(e_i, e_h)$
    $G_k \leftarrow G_k \cup e_l$
  Enquanto $|G_k| < q$
Passo 2: definição dos pesos de relevância
  Computar de acordo com (2.13) para MFCMdd-RWL ou (2.17) para MFCMdd-RWG
Passo 3: definição da melhor partição nebulosa
  Os representativos $G^{(t)} = (G_1^{(t)}, .., G_K^{(t)})$ são fixados.
  Atualizar partiao nebulosa de acordo com (2.14) para MFCMdd-RWL ou (2.18) para MFCMdd-RWG
Critério de parada
  Computar J de acordo com (2.13) para MFCMdd-RWL ou de acordo com (2.17) para MFCMdd-RWG
  Se $|J^{(t)} - J^{(t-1)}| \leq \epsilon$ ou $t > T$, **então** PARE;
  Senão vá para o Passo 1

Função parametrizada pelo vetor de pesos de relevância $\lambda = (\lambda_1, ..., \lambda_p)$, em que $\lambda_{kj} > 0$ e $\prod_{j=1}^{p} \lambda_j = 1$, e associado com o *cluster* C_k $(k = 1, ..., K)$

$$D_\lambda(e_i, G_k) = \sum_{j=1}^{p} \lambda_j D_j(e_i, G_k) = \sum_{j=1}^{p} \lambda_j \sum_{e \in G_k} d_j(e_i, e) \tag{2.16}$$

$$\lambda_j = \frac{\{\prod_{h=1}^{p} (\sum_{k=1}^{K} [\sum_{l=1}^{n} (u_{lk})^m D_h(e_l, G_k)])\}^{\frac{1}{p}}}{\sum_{k=1}^{K} [\sum_{l=1}^{n} (u_{lk})^m D_j(e_l, G_k)]} = \frac{\{\prod_{h=1}^{p} (\sum_{k=1}^{K} [\sum_{l=1}^{n} (u_{lk})^m \sum_{e \in G_k} d_h(e_l, e)])\}^{\frac{1}{p}}}{\sum_{k=1}^{K} [\sum_{l=1}^{n} (u_{lk})^m \sum_{e \in G_k} d_j(e_l, e)]} \tag{2.17}$$

O grau de pertinência de cada objeto passa a ser atualizado de acordo com a Eq. (2.18).

$$u_{ik}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \lambda_k^{(t)} \sum_{e \in G_k^{(t)}} \sum_{j=1}^{p} d_j(e_i, e)}{\sum_{j=1}^{p} \lambda_k^{(t)} \sum_{e \in G_h^{(t)}} \sum_{j=1}^{p} d_j(e_i, e)} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{2.18}$$

2.4 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados alguns algoritmos para particionamento rígido e nebuloso de dados relacionais com múltiplas visões e com pesos de relevância para as matrizes de dissimilaridades. Tais modelos foram apresentados porque servirão de base para entendimento dos algoritmos desenvolvidos e que serão descritos na metodologia desta tese.

3 REVISÃO DE LITERATURA

Neste capítulo serão apresentados alguns trabalhos relacionados que foram considerados relevantes. Em relação ao particionamento de dados, existem diversos trabalhos abordando o dados vetoriais utilizando métodos tradicionais e usando inteligência de enxame para resolver o problema.

3.1 AGRUPAMENTO DE DADOS VETORIAIS

Segundo Jain (2010), o algoritmo *K-means* é um dos mais populares algoritmos de agrupamento de dados vetoriais devido as seguintes características: fácil implementação, simplicidade, eficiência e sucesso empírico. O algoritmo encontra uma partição tal que o erro quadrado entre o centróide de um *cluster* e os pontos do *cluster* é minimizado. O objetivo do *K-means* é minimizar a soma dos erros quadrados de todos os *clusters*. Minimizar essa função objetivo é um problema NP-Difícil (MEILa, 2006). O algoritmo começa com uma partição inicial com K *clusters* e atribui os pontos aos *clusters* de forma a reduzir o erro quadrado. Outro algoritmo para agrupamento de dados vetoriais é o *K-medoids* (KAUFMAN; ROUSSEEUW, 1990), este é uma variação do algoritmo *K-means* e usa o objeto mais próximo ao centróide, chamado *medóide*, como representante ao invés da média. O principal problema desses algoritmos é que eles são muito sensíveis a escolha inicial dos representantes de cada *cluster*.

No que diz respeito ao agrupamento de dados vetoriais com múltiplas visões, Bickel e Scheffer (2004) trataram o problema assumindo que os dados foram gerados por um modelo mistura (do inglês, *mixture model*). Dessa forma, dados que foram gerados por um mesmo componente mistura seriam atribuídos a um mesmo grupo e objetos gerados por componentes diferentes seriam atribuídos a grupos diferentes. Outra suposição feita no trabalho é que os atributos disponíveis no conjunto de dados poderiam ser divididos em dois subconjuntos independentes e que cada subconjunto seria suficiente para aprendizado. Nesse trabalho, métodos foram estudados e desenvolvidos para agrupamento de dados textuais com múltiplas visões. Resultados empíricos mostraram que a versão do algoritmo *k-means* para dados com múltiplas visões foi melhor do que a sua versão para os dados com apenas uma visão. Nesse estudo, seis bases de dados foram consideradas, em que não foi possível encontrar uma divisão natural dos atributos e, dessa forma, divisões aleatórias dos atributos foram consideradas. O desempenho médio foi avaliado para 10 possíveis divisões dos atributos.

Outro trabalho para agrupamento de dados com múltiplas visões foi realizado por Kumar e III (2011). Os autores propuseram um algoritmo de agrupamento espectral para dados com duas visões, em que cada visão é usada independentemente para agrupamento

dos dados. Neste trabalho, os autores consideram a hipótese de que no agrupamento ideal os dados seriam atribuídos aos mesmos *clusters* independentemente da visão utilizada. Dessa forma, a abordagem proposta usa uma restrição para buscar agrupamentos que são os mesmos entre as visões. A abordagem proposta utiliza quatro passos. No primeiro passo, um agrupamento espectral é realizado para obter os autovetores em cada visão. No segundo passo, os dados são agrupados de acordo com os autovetores encontrados referentes a primeira visão, e após isso, o agrupamento produzido é usado para modificar a estrutura do grafo na visão 2. No terceiro passo, os autovetores da segunda visão são usados para agrupar os dados e o agrupamento é usado para modificar o grafo da primeira visão. No último passo, ocorre uma verificação relativa ao critério de parada. Uma vantagem dessa abordagem é que ela não possui parâmetros para definir. Para avaliar o método proposto, conjuntos de dados reais e sintéticos foram usados. Os índices externos usados para avaliar a qualidade das partições geradas foram: *f-measure*, precisão, *recall*, entropia, informação mútua normalizada e o índice de rand ajustado.

Em Tzortzis e Likas (2012), uma abordagem de agrupamento com múltiplas visões com pesos baseada em *kernel* foi apresentada. Nesse trabalho, dois algoritmos foram usados: *multi-view kernel k-means* e *multi-view spectral clustering*. Cada visão foi expressada como uma matriz *kernel*. O peso de cada visão bem como o agrupamento final são aprendidos minimizando as divergências entre as diferentes visões. O índice externo que calcula a informação mútua normalizada foi usado para avaliar os agrupamentos resultantes e comparar a abordagem proposta com outras da literatura.

Na pesquisa realizada por Cai, Nie e Huang (2013), os autores criaram um método robusto de agrupamento *multi-view* que integra múltiplas representações de grandes quantidades de dados (*big data*) vetoriais. O método proposto foi avaliado com o uso de seis conjuntos de dados *benchmark* e comparado com diversas abordagens de agrupamento usadas comumente. Jiang et al. (2016b) propuseram a formulação de otimização multiobjetivo para agrupamento de dados *multi-view*, em que cada visão é tratada separadamente em um dos objetivos. Nesse trabalho, foram utilizados cinco algoritmos evolucionários multiobjetivos: NSGA-II, SPEA2,MOEA/D, SMS-EMOA e NSGA-III. Esses algoritmos foram aplicados a seis conjuntos de dados do mundo real para avaliação e comparação dos resultados. Na comparação feita entre os algoritmos mencionados, o SPEA2 foi melhor do que os outros em termos de acurácia, índice de *rand* e tempo de execução. Dentre as populações de soluções fornecidas pelos algoritmos, a abordagem semi-supervisionada foi utilizada. Dessa forma, dentre as soluções do conjunto de Pareto fornecido, a solução com maior acurácia foi selecionada. Na comparação feita com outros algoritmos para dados com apenas um objetivo, o SPEA2 demonstrou ter obtido resultados competitivos.

Chen et al. (2013) propuseram o algoritmo *TW-k-means*, um algoritmo de agrupamento com pesos em dois níveis para dados *multi-view*, em que é possível computar simultaneamente pesos para as diferentes visões e variáveis individuais. Além disso, dois novos passos

foram adicionados ao processo de agrupamento do *k-means* tradicionais para computar as visões em dois níveis. A função objetivo utilizada considera o somatório das dispersões em cada *cluster* e dois termos referentes a pesos de entropia. O algoritmo resolve iterativamente quatro problemas de minimização: atualização da partição, atualização dos centróides, atualização dos pesos das variáveis e atualização dos pesos das visões. Para investigar as propriedades de dois tipos de pesos para o algoritmo, dois conjuntos de dados reais foram usados. O TW-k-Means foi comparado a outros cinco algoritmos de agrupamento utilizando três bases de dados. Os resultados mostraram que o TW-k-Means foi melhor significativamente do que os outros métodos usando quatro índices para validação de agrupamentos, a saber: precisão, *recall*, *f-measure* e acurácia.

Recentemente, Jiang, Qiu e Wang (2016) desenvolveram um novo *framework* com base no algoritmo *k-means* para agrupamento de dados vetoriais com múltiplas visões que simultaneamente atribui pesos para as diferentes variáveis como também pesos para as diferentes visões, isto é, existe a atribuição de pesos em dois níveis. Segundo os autores, a vantagem principal dessa estratégia é que pode simultaneamente discriminar as visões e variáveis e que, portanto, os métodos propostos conseguem ser mais flexíveis e robustos. A Figura 2 ilustra diferentes métodos de atribuição de pesos em dados com múltiplas visões. Um conjunto de dados possuindo três visões com cinco objetos e dez variáveis é usado no exemplo. As variáveis com mesma cor pertencem a mesma visão. No exemplo (a), cada variável x_i possui um peso w_i. No exemplo (b), ocorre a atribuição de um mesmo peso para todas as variáveis em uma mesma visão. Por fim, Jiang, Qiu e Wang (2016) propuseram a atribuição de pesos em dois níveis ilustrados no exemplo (c). Dois índices externos foram usados para avaliação dos particionamentos gerados: acurácia e o índice de Rand.

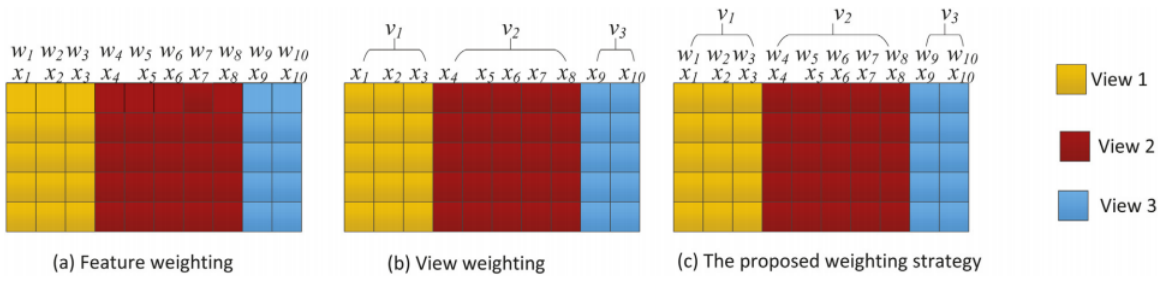


Figura 2 – Ilustração de diferentes métodos de atribuição de pesos em dados com múltiplas visões

Fonte: Extraída de (JIANG; QIU; WANG, 2016)

Em Xu, Wang e Lai (2016), os autores desenvolveram um algoritmo, nomeado wmcfs, que pode simultaneamente desempenhar o agrupamento de dados com múltiplas visões e seleção de variáveis. Uma função objetivo global é usada que leva em consideração o aprendizado com múltiplas visões e a seleção de variáveis. Para resolver a função objetivo, os autores usaram iterações do tipo *Expectation Maximization*, as quais conseguem convergir para resultados de agrupamento aceitáveis. Dois índices externos são usados para avaliar

os particionamentos gerados: a taxa de classificação (do inglês, *classification rate* - CR) e a informação mútua normalizada (do inglês, *normalized mutual information*).

Chikhi (2016) propuseram a abordagem *Multi-View Normalized Cuts*-MVNC, esta é um algoritmo de dois passos para agrupamento de dados com múltiplas visões. No primeiro passo, um particionamento inicial é gerado usando um algoritmo de agrupamento espectral. No segundo passo, um procedimento de busca local é usado para refinar a partição inicial gerada. O método proposto foi avaliado e comparado a outras abordagens da literatura usando três conjuntos de dados *multi-view*. Os índices externos utilizados para avaliar os particionamentos gerados foram: pureza, medida F e NMI.

Wang e Chen (2017) apresentaram uma nova abordagem de agrupamento nebuloso de dados com múltiplas visões, chamada MinimaxFCM, que usa otimização minimax baseada no conhecido algoritmo *fuzzy c-means*. Nesta abordagem, o agrupamento consensual é gerado baseado na otimização minimax em que a divergência máxima entre as visões ponderadas é minimizada. O peso de cada visão é aprendido automaticamente no processo de agrupamento. Índices externos são usados para comparar os métodos utilizados, sendo eles: acurácia, informação mútua normalizada e *f-measure*.

3.2 ONP E AGRUPAMENTO DE DADOS VETORIAIS

Uma revisão detalhada sobre metaheurísticas inspiradas na natureza foi fornecida por Nanda e Panda (2014) e Hasan e Ramakrishnan (2011). Métodos de agrupamento baseados em ONP foram revisados recentemente por Alam et al. (2014) e por Rana, Jasola e Kumar (2011). Durante o levantamento de trabalhos relacionados, nenhum trabalho foi encontrado que abordasse o partionamento de dados relacionais com múltiplas visões por nenhuma metaheurística inspirada em inteligência de enxame.

Por outro lado, diversos pesquisadores estiveram trabalhando com agrupamento rígido e metaheurísticas. Merwe e Engelbrecht (2003) propuseram um algoritmo de agrupamento baseado em ONP para dados vetoriais, o qual utiliza o algoritmo tradicional *k-means*, além de usar centróides como representativos de cada *cluster*. No modelo proposto, cada partícula é representada por k centróides. Outras abordagens baseadas em ONP para dados vetoriais foram apresentadas em Das, Abraham e Konar (2009). Yang, Sun e Zhang (2009a) desenvolveram o método híbrido PSO-KHM que é baseado em otimização por partículas e no algoritmo *K-Harmonic Means*. O PSO-KHM foi projetado com o objetivo de ajudar o KHM a escapar de soluções ótimas locais e também para melhorar a velocidade de convergência do PSO. O PHO-KHM foi comparado ao PSO e ao KHM, e obteve melhores resultados. Filho et al. (2015) apresentaram métodos híbridos para agrupamento nebuloso de dados vetoriais. Os métodos híbridos apresentados se baseiam em uma versão melhorada chamada IDPSO, em que os fatores de aceleração e peso de inércia mudam dinamicamente e, dessa forma, permitem melhor balanço entre exploração local e global.

No trabalho de Jiang et al. (2016a), os autores propuseram um método de agrupamento considerando múltiplas visões com atribuição de pesos em dois níveis com ênfase em pesos nebulosos para as visões e variáveis. Além disso, foi utilizada uma estratégia de busca global que combina otimização por nuvem de partículas e otimização de gradiente descendente para encontrar os melhores centróides para os *clusters* e melhores vetores de pesos. O desempenho do algoritmo proposto foi avaliado em comparação com outros algoritmos da literatura para três bases de dados usando dois índices externos: acurácia e índice de Rand.

Alswaitti, Albughdadi e Isa (2018) desenvolveram uma abordagem para agrupamento de dados baseado em nuvem de partículas combinado com estimação de densidade baseada em *kernel*. A abordagem foi desenvolvida para lidar com a prematura convergência e com o desafio de estimação dos coeficientes de aprendizado. A abordagem proposta foi comparada aos métodos K-means, PSO, PSO-KM e FCM. Os resultados obtidos no estudo mostraram que o método proposto obteve resultados competitivos e superiores em alguns casos.

3.3 AGRUPAMENTO DE DADOS RELACIONAIS

Em relação aos algoritmos de agrupamento rígido usando descrições relacionais dadas por uma única matriz de dissimilaridade, existem trabalhos que foram propostos por alguns autores. Kaufman e Rousseeuw (1990) propuseram o algoritmo *Partitioning Around Medoids*, que pode usar dados de características ou dados relacionais (matriz de dissimilaridade). O algoritmo PAM usa diferentes representantes chamados *medóides*, de modo que a soma das dissimilaridades pode ser minimizada em cada *cluster*. *Clustering Large Applications*, proposto em Kaufman e Rousseeuw (1990), é uma versão modificada do algoritmo PAM que consegue lidar com grandes bases de dados.

O *Sequential Agglomerative Hierarchical Non-overlapping - SAHN* é um dos métodos mais conhecidos e foi proposto por Sneath e Sokal (1973). O SAHN é outro algoritmo que usa as descrições relacionais de objetos para agrupá-los os objetos até chegar a um único conjunto contendo todos os objetos. O algoritmo SAHN utiliza uma abordagem *bottom-up* que gera os *clusters* ao aglomerar sequencialmente *clusters* menores similares.

O *Relational Hard c-Means - RHCM* é outro algoritmo para agrupamento relacional proposto por Hathaway, Davenport e Bezdek (1989). Nesse trabalho, o algoritmo faz uso da matriz de pertinência U, esta é uma matriz $k \times N$ contendo a pertinência dos objetos para cada *cluster*. O algoritmo RHCM é a versão relacional do algoritmo *Hard c-Means*.

Krishnapuram, Joshi e Yi (1999) propuseram um outro algoritmo para o agrupamento de dados relacionais utilizando matrizes de dissimilaridade denominado algoritmo rígido *c-medoids*, também denotado como como HCMdd. Este algoritmo também usa *medoids* como representantes e inicialmente seleciona k objetos distintos para representar os *clusters*. Em cada etapa, um novo objeto é selecionado para representar cada *cluster* de acordo

com a distância a todos os objetos em cada *cluster* de tal forma que a soma de distância dentro de cada *cluster* pode ser minimizada.

Frigui, Hwang e Rhee (2007) propuseram um algoritmo para agrupamento e agregação de dados relacionais nomeado (do inglês, *Clustering and Aggregating Relational Data*). Nesse trabalho, supõe-se que os dados estão disponíveis na forma relacional, isto é, cada par de objetos está representado por um grau indicando a relação entre eles. Além disso, assume-se que a informação relacional está representada por múltiplas matrizes de dissimilaridades. O algoritmo CARD foi projetado para particionar os objetos com base nas múltiplas matrizes (*multi-view data*) e aprender pesos de relevância das matrizes para cada *cluster*. O algoritmo proposto é baseado nos algoritmos relacionais FCM e FANNY.

Horta e Campello (2010) desenvolveram duas versões de um algoritmo evolucionário para dados relacionais chamadas F-EARC-BKM e F-EARC-RHCM. A principal diferença entre as duas versões se dá em relação ao uso algoritmos diferentes na etapa de busca local. A primeira versão utiliza o algoritmo *Basic-k-medoids* e a segunda versão utiliza o algoritmo RHCM. (HORTA; ANDRADE; CAMPELLO, 2011) apresentaram a versão nebulosa do método F-EARC-BKM, chamada F-EARFC. O método utiliza o algoritmo *Fuzzy c-Medoids* como busca local aplicada a todos os indivíduos da população.

Mei e Chen (2011) desenvolveram um método para agrupamento nebuloso de dados relacionais com única visão chamado *fuzzy clustering with multi-medoids (FMMdd)*. Este metódo foi desenvolvido para utilizar múltiplos *medóides* para representação dos grupos. Carvalho, Lechevallier e Melo (2012) apresentaram algoritmos de agrupamento rígido que levam em consideração p matrizes de dissimilaridade simultaneamente, chamados MRDCA, MRDCA-RWL e MRDCA-RWG, os dois últimos calculam pesos de relevância para cada matriz de dissimilaridades de forma que as matrizes possuem diferentes influências no processo de agrupamento. Levando-se em consideração o agrupamento nebuloso, Carvalho, Lechevallier e Melo (2013) desenvolveram os algoritmos de agrupamento nebuloso MFCMdd, MFCMdd-RWL e MFCMdd-RWG, os quais também baseiam-se em p matrizes de dissimilarides, sendo os dois últimos capazes de estimar pesos de relevância para as matrizes de dissimilaridades. No capítulo 3, esses algoritmos de agrupamento rígido e nebuloso serão explicados com detalhes, pois esses métodos são usados na abordagem híbrida desenvolvida nesta tese.

Pio et al. (2018) desenvolveram o método HENPC, o qual trabalha com redes heterogêneas com estrutura arbitrária. Este método identifica o número ótimo de grupos automaticamente baseando-se na distribuição dos dados. Este método tem quatro etapas principais: (i) identificação da força das relações entre nós; (ii) identificação de um conjunto inicial de grupos na forma de cliques multitipo; (iii) construção de uma hierarquia de grupos heterogêneos; e (iv) identificação de funções de classifição baseado nos grupos encontrados. Os experimentos do trabalho demonstraram que a abordagem foi superior a outros métodos em termos de qualidade e acurácia.

3.4 ONP E AGRUPAMENTO DE DADOS RELACIONAIS

Gusmão e Carvalho (2016) propuseram um algoritmo de agrupamento rígido baseado na otimização por nuvem de partículas aplicado a dados relacionais com única visão descritos considerando uma única matriz contendo as dissimilaridades entre todos os pares de objetos. Nesse trabalho, o algoritmo proposto foi comparado com o RHCM, HCMdd e um algoritmo espectral, desenvolvidos para particionamento de dados relacionais. A qualidade das partições resultantes foi comparada utilizando cinco índices externos. As matrizes de dissimilaridade para cada base de dados foram calculadas considerando a matriz de dados normalizada. Na maioria dos casos, o algoritmo proposto obteve particionamentos das bases de dados melhores, em termos de índices externos, do que os outros três.

3.5 ÍNDICES PARA VALIDAÇÃO DE AGRUPAMENTOS

Índices para validação de agrupamentos tem um papel muito importante na maioria dos algoritmos de agrupamento. Devido a esse fato, diversos trabalhos foram feitos com o intuito de comparar diversos índices e avaliar quais deles poderiam fornecer melhores particionamentos dos dados. Nesta subseção, alguns desses trabalhos são apresentados.

Maulik e Bandyopadhyay (2002) avaliaram o desempenho de três algoritmos, *K-means*, *Single Linkage* e uma técnica baseada em *Simulated Annealing*, em que quatro índices para validação de agrupamentos foram usados, a saber: Davies-Bouldin, Dunn, Calinsky-Harabasz e o índice I. Os índices foram usados para estimar o número apropriado de grupos. Liu et al. (2010) apresentaram um estudo sobre 11 índices internos para validação de agrupamentos rígidos. Os resultados experimentais demonstraram que o índice S_Dbw foi o único índice que obteve bom desempenho considerando todos os cinco aspectos convencionais de agrupamentos.

Xu, Xu e Wunsch (2012) realizaram um estudo comparativo de índices para métodos de agrupamentos baseados em inteligência de enxame. O método utilizado no trabalho foi o algoritmo híbrido chamado DEPSO (*Differential Evolution Particle Swarm Optimization*) Nesse estudo, oito índices foram avaliados: *Calinsky-Harabasz*, CS, Davies-Bouldin, o índice de Dunn com duas versões generalizadas, índice I e índice da Silhueta. De acordo com os resultados do estudo, o índice da Silhueta se destacou dentre os demais para as bases de dados consideradas.

Raitoharju et al. (2017) propuseram uma nova estratégia para calcular a aptidão das partículas usando agrupamento baseado em enxame de partículas. Na estratégia proposta, a aptidão das partículas passa a ser calculada com os centróides computacionais dos grupos formados ao invés dos centróides que representam as partículas. Nesse estudo, 31 funções de aptidão foram consideradas, estas sendo baseadas em 17 índices para validação de agrupamento. Três critérios para avaliação de qualidade de agrupamentos foram usados

e as funções que se destacaram no estudo foram: o índice de XU, o índice WB e uma variante do índice de Dunn.

3.6 SÍNTESE DO CAPÍTULO

Nesta capítulo foram apresentados diversos trabalhos relacionados o agrupamento de dados vetoriais e relacionais usando uma única visão dos dados ou múltiplas visões. Além disso, também foram apresentados alguns trabalhos sobre a modelagem da otimização de enxame para resolver o problema da agrupamento de dados *single-view* e *multi-view*.

4 OTIMIZAÇÃO POR NUVEM DE PARTÍCULAS E AGRUPAMENTO DE DADOS

Neste capítulo serão apresentados os principais conceitos relacionados a PSO bem como alguns modelos baseados em PSO para agrupamento de dados. Estes modelos auxiliarão no entendimento dos modelos desenvolvidos nesta tese.

4.1 OTIMIZAÇÃO POR NUVEM DE PARTÍCULAS

Otimização por nuvem de partículas (do inglês, *Particle Swarm Optimization - PSO*) é um processo de busca estocástico baseado em população, modelado a partir do comportamento social de pássaros proposto por um psicólogo social e por um engenheiro eletricista (KENNEDY; EBERHART, 1995). O algoritmo utiliza uma população de partículas, em que cada partícula representa uma solução potencial para o problema de otimização em questão e possui um valor numérico associado chamado de aptidão (do inglês, *fitness*) que mede a qualidade da solução.

O conjunto de partículas é chamado de enxame. O objetivo é encontrar a posição que resulta na melhor avaliação de uma dada função objetivo. No modelo proposto por Kennedy e Eberhart (1995), cada partícula é representada por uma posição no espaço dimensional R^p. A atualização da posição de cada partícula é feita usando a velocidade, a qual é calculada com base em três componentes: a posição atual da partícula, a melhor posição encontrada pelo enxame e pela melhor posição encontrada pela partícula. A melhor posição encontrada por cada partícula é denotada como *pbest* e a melhor posição global encontrada pelo enxame é denotada como *gbest*. Cada partícula mantém as seguintes informações:

$x^{(t)}$: posição atual;

$y^{(t)}$: melhor posição da partícula;

$v^{(t)}$: velocidade atual;

A posição de cada partícula é ajustada de acordo com a Eq. (4.1).

$$x^{(t)} = x^{(t-1)} + v^{(t)} \tag{4.1}$$

onde

$$v^{(t)} = w \times v^{(t-1)} + c_1 \times r_1 \times (y^{(t)} - x^{(t)}) + c_2 \times r_2 \times (\bar{y}^{(t)} - x^{(t)}) \tag{4.2}$$

Na Eq. (4.2), r_1 e r_2 são números aleatórios dentro do intervalo [0,1]. As constantes c_1 e c_2, chamadas fatores de aceleração, são dados de entrada. A constante w é chamada fator de inércia e também é um parâmetro de entrada. $\bar{y}^{(t)}$ representa a melhor posição global encontrada pelo enxame na iteração t. A posição atual das partículas é atualizada

de acordo com a Eq. (4.1). Em (KENNEDY; EBERHART, 1995), os pesos da fórmula da velocidade foram fixados: $w = 1$, $c_1 = c_2 = 2$. O termo $y^{(t)} - x^{(t)}$ é calculado como a diferença entre dois vetores. Os passos gerais do PSO são descritos no Alg. 5.

Algoritmo 5: Otimização por nuvem de partículas

Passo 1: Initialização do enxame
    Para $i = 1$ até n_p **faça**
        Inicialize p_i
        $y_i^{(t)} \leftarrow$ posição atual de p_i
    Definir *gbest* como a posição com melhor aptidão

Passo 2: Atualização da velocidade e da posição
    $t = 0$
    Repita
        Para cada particula p_i no enxame S
            Atualizar velocidade de acordo com (4.2)
            Atualizar posição de acordo com (4.1)
            Se (*melhoria local*) **entao**
                Atualizar $y_i^{(t)}$

        Se (*melhoria global*) **entao**
            Atualizar $\bar{y}^{(t)}$
        $t = t + 1$
    Até $t = T_{max}$

Em abordagens mais recentes, o valor de w decresce linearmente durante a execução do algoritmo, sendo atualizado de acordo com a Eq. (4.3).

$$w^{(t)} = w_{max} - ((w_{max} - w_{min})/T_{max}) \times t \tag{4.3}$$

Na Eq. (4.3), w_{max} e w_{min} são os valores máximo e mínimo que o peso de inércia pode ter, T_{max} é o número máximo de iterações e t é a iteração corrente.

4.2 AGRUPAMENTO DE DADOS USANDO OTIMIZAÇÃO POR NUVEM DE PARTÍCULAS

No trabalho desenvolvido em Merwe e Engelbrecht (2003), o agrupamento de dados foi abordado com Otimização por nuvem de partículas. Na modelagem apresentada nesse trabalho, cada partícula p_i foi representada como k vetores de centróides, isto é,

$$x_i = (m_{i1}, ..., m_{ij}, ..., m_{ik}) \tag{4.4}$$

em que m_{ij} representa o vetor referente ao centróide do j-ésimo grupo da partícula p_i. Dessa forma, o enxame representa um grupo de agrupamentos em potencial. A aptidão

(*fitness*) das partículas foi mensurada de acordo com a Eq. (4.5).

$$J = \frac{\sum_{j=1}^{K}[\sum_{\forall o_l \in C_{ik}} d(o_l, m_{ij})]/|C_{ij}|]}{K} \tag{4.5}$$

Algoritmo 6: Agrupamento de dados baseado em PSO

```
1: Inicializar partículas com K centróides aleatórios
2: Para t = 1 até T_max faça
3:     Para cada partícula p_i no enxame S
4:         Para cada vetor de dados x_l
5:             Calcular a distância Euclidiana d(o_l, m_ij) para todos os centróides
6:             Atribuir o_l ao cluster C_ij de forma que
7:                 d(o_l, m_ij) = min_{∀c=1,...,K} d(o_l, m_ic)
8:             Calcular a aptidao usando a Eq. (4.5)
9:         Atualizar pbest
10:
11:    Se (melhoria global) entao
12:        Atualizar os centroides do gbest
```

Diferentemente do algoritmo *k-means*, o agrupamento de dados baseado em PSO tem menor sensibilidade a inicialização, pois o método parte de vários pontos diferentes em paralelo. É importante também salientar que existem outras abordagens na literatura em que a topologia das partículas é diferente. No modelo *gbest* apresentado, todas as partículas são vizinhas entre si. Outro modelo utilizado é o *lbest - local best*, em que a topologia usada representa um anel e cada partícula possui apenas dois vizinhos. Embora exista discussão sobre qual topologia apresenta melhor desempenho para determinados problemas, Engelbrecht (2013) realizou diversos experimentos comparando as duas topologias e verificou que nenhuma das duas abordagens pode ser considerada melhor que a outra.

4.3 AGRUPAMENTO BASEADO EM MÉTODO HÍBRIDO PSO-K-MEANS

No trabalho desenvolvido em Ahmadyfard e Modares (2008), os autores apresentaram uma hibridização entre a PSO e o método K-means. Este método foi desenvolvido para se beneficiar da característica de exploração global do PSO juntamente com a característica de rápida convergência do K-means. A aptidão das partículas foi avaliada de acordo com a Eq. (4.6).

$$Aptidão(p_i) = \frac{\sum_{j=1}^{K} \sum_{x_l \in C_{ij}} d(x_l, m_{ij})}{n} \tag{4.6}$$

Nesse trabalho, após a atualização da posição, verifica-se também se a posição recebe valor fora do intervalo $[X_{min}, X_{max}]$ e, caso ocorra, o valor é definido como X_{min} ou X_{max}. Similarmente, a velocidade também obedece ao intervalo $[V_{min}, V_{max}]$ e deve ser definida como um dos limites caso ocorra atualização do valor para fora do intervalo. Os passos do método são descritos no Alg. 7.

Algoritmo 7: Agrupamento de dados baseado em método híbrido PSO-KM

1: Inicializar partículas a velocidade e posição aleatoriamente
2: Avaliar a aptidão das partículas de acordo com a Eq. (4.6)
3: Se o número de iterações exceder o limite máximo, vá para o Passo 7, senão continue para o Passo 4
4: A posição da melhor partícula é armazenada. A posição das partículas é atualizada de acordo com as Eq. (4.1) e (4.2)
5: Reduzir o peso de inércia
6: Se o *gbest* não melhora para um número definido de iterações, vá para o próximo passo; senão volte para o Passo 3.
7: Usar o K-means para finalizar a tarefa de agrupamento.

Como é possível perceber nos passos do PSO-KM, inicialmente, o PSO é utilizado para realizar uma busca inicial partindo de vários pontos em paralelo. Caso o PSO não consiga melhorar o *gbest* dentro de um número determinado de iterações ou caso o limite de iterações seja atingido, o *k-means* é acionado, tendo como os centróides apresentandos pelo *gbest* como iniciais, com o intuito de refinar a solução e finalizar a tarefa.

4.4 AGRUPAMENTO NEBULOSO BASEADO EM OTMIZAÇÃO POR NUVEM DE PARTÍCULAS

Izakian e Abraham (2011) apresentaram um modelo para agrupamento nebuloso de dados baseado em PSO. Nesse modelo, a posição de cada partícula é representada pela matriz de pertinência U, em que cada elemento da matriz deve respeitar as restrições (4.7) e (4.8). A velocidade de cada partícula também é representada por uma matriz $n \times K$, em que os elementos da matriz pertencem ao intervalo $[-1, 1]$.

$$U_{n\times K} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1K} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nK} \end{bmatrix}$$

$$\sum u_{lk} = 1 \quad \forall k, \tag{4.7}$$

$$0 \leq u_{lk} \leq 1 \quad \forall l, k. \tag{4.8}$$

Após a atualização da posição, as restrições (4.7) e (4.8) podem ser violadas. Dessa forma, é necessário normalizar a matriz da posição. Primeiro, elementos negativos são zerados. Se todos os elementos de uma mesma linha estiverem com valor zero, eles recebem valores aleatórios dentro do intervalo $[0, 1]$ e a matriz é atualizada da seguinte forma:

$$U_{normalizada} = \begin{bmatrix} u_{11}/\sum_{j=1}^{K} u_{j1} & \cdots & u_{1K}/\sum_{j=1}^{K} u_{j1} \\ \vdots & \ddots & \vdots \\ u_{n1}/\sum_{j=1}^{K} u_{jn} & \cdots & u_{nK}/\sum_{j=1}^{K} u_{jn} \end{bmatrix}$$

A aptidão de cada partícula foi quantificada de acordo com a Eq. (4.9). J_m representa a função objetivo do algoritmo *Fuzzy C-Means*. A medida que J_m decresce, melhor se torna o agrupamento e a aptidão da partícula se torna maior. Os passos desse método são descritos em 8.

$$Aptidão(p_i) = \frac{K}{J_m} \tag{4.9}$$

$$J_m = \sum_{j=1}^{K} \sum_{l=1}^{n} u_{lj} dist(o_l, z_j) \tag{4.10}$$

O centróide z_{ij} do j-ésimo grupo da partícula p_i é obtido de acordo com a Eq. (4.11).

$$z_{ij} = \frac{\sum_{l=1}^{n} u_{lj}^m o_l}{\sum_{l=1}^{n} u_{lj}^m} \tag{4.11}$$

Algoritmo 8: Agrupamento nebuloso de dados baseado em PSO

1: Inicializar os parâmetros incluindo tamanho da população, c1, c2, w, e o número máximo de iterações.
2: Criar um enxame com P partículas (X, pbest, gbest e V são representados por matrizes $n \times K$).
3: Inicializar X, V, pbest para cada partícula.
4: Calcular os centróides de cada cluster para cada partícula usando a Eq. (4.11).
5: Calcular a aptidão de cada partícula usando a Eq. (4.9).
6: Calcular pbest para cada partícula.
7: Calcular gbest para o enxame.
8: Atualizar a matrix de velocidade para cada partícula usando Eq. (4.2).
9: Atualizar a matrix de posição para partícula usando Eq. (4.1).
10: Se a condição de término não satisfeita, voltar para o Passo 4.

4.5 OUTROS TRABALHOS

Diversos outros trabalhos foram propostos como discutido em Rana, Jasola e Kumar (2011). Mudanças relativas ao uso de diferentes funções de aptidão, representação das partículas, variações de k-means, hibridizações com outros métodos e variações do PSO foram investigadas. Por exemplo, em Toreini e Mehrnejad (2011), os autores consideraram a função objetivo utilizada no FCM ao invés do erro de quantização utilizada em trabalhos anteriores e essa mudança mostrou-se promissora segundo os resultados apresentados no trabalho. Devido a importância da escolha do critério para validação de agrupamento,

os autores em Xu, Xu e Wunsch (2012) e Liu et al. (2012), dentre outros trabalhos, investigaram diversos índices como funções de aptidão para agrupamento de dados baseado em otimização de enxame.

4.6 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados os conceitos basicos sobre PSO e agrupamento de dados baseado em PSO. O próximo capítulo abordará os modelos híbridos para agrupamento de dados relacionais com múltiplas visões desenvolvidos neste trabalho.

5 MÉTODOS HÍBRIDOS

Nesta seção será apresentada a metodologia utilizada nesta tese. As seções seguintes contém detalhes sobre como os conceitos da otimização por enxame foram utilizados para agrupamento de dados relacionais com múltiplas visões, mais especificamente detalhes sobre: representação das partículas, inicialização das partículas, atualização da velocidade e método para atualização da posição das partículas. Por último, a complexidade de tempo dos métodos desenvolvidos é apresentada.

5.1 OTIMIZAÇÃO POR NUVEM DE PARTÍCULAS PARA AGRUPAMENTO DE DADOS RELACIONAIS

A Figura 3 ilustra o fluxograma geral da abordagem desenvolvida neste trabalho. Nesta tese, a abordagem desenvolvida combina PSO com algoritmos para agrupamento de dados baseados em múltiplas matrizes de dissimilaridades com o objetivo de melhorar o balanço entre busca local e global. A escolha do PSO como modelo de otimização de exame para composição dos métodos híbridos se deve a ampla utilização em análise de agrupamentos bem como às suas características interessantes para problemas de otimização.

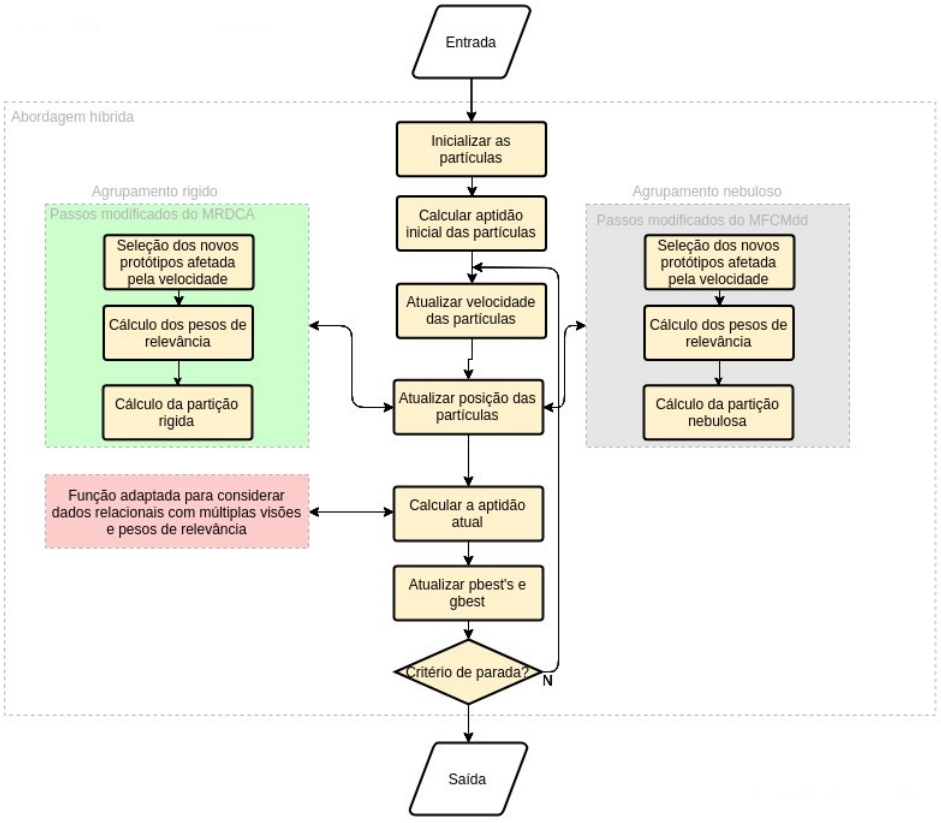


Figura 3 – Fluxograma geral da abordagem desenvolvida para dados com múltiplas visões

O PSO possui características como: simplicidade, facilidade de adaptação para processamento paralelo e possibilidade de obter convergência global. Estas características também motivaram o uso desse modelo para composição dos métodos híbridos desenvolvidos nesta tese. Nesta tese, seis modelos híbridos foram desenvolvidos:

1. PSO-SV: modelo para agrupamento rígido de dados relacionais com única visão. Neste modelo o PSO é combinado ao algoritmo HCMdd na fase de atualização da posição;

2. PSO-RWL: modelo para agrupamento rígido de dados relacionais com múltiplas visões. Neste modelo o PSO é combinado ao algoritmo MRDCA-RWL na fase de atualização da posição;

3. PSO-RWG: modelo para agrupamento rígido de dados relacionais com múltiplas visões. Neste modelo o PSO é combinado ao algoritmo MRDCA-RWG na fase de atualização da posição. Este modelo difere do anterior devido a estimação dos pesos de relevância que ocorre globalmente;

4. FPSO-SV: modelo para agrupamento nebuloso dados relacionais com única visão. Neste modelo o PSO é combinado ao algoritmo SFCMdd na fase de atualização da posição;

5. FPSO-RWL: modelo para agrupamento nebuloso de dados relacionais com múltiplas visẽs. Neste modelo o PSO é combinado ao algoritmo MFCMdd-RWL na fase de atualização da posição;

6. FPSO-RWG: modelo para agrupamento nebuloso de dados relacionais com múltiplas visẽs. Neste modelo o PSO é combinado ao algoritmo MFCMdd-RWG na fase de atualização da posição. Este modelo difere do anterior devido a estimação dos pesos de relevância que ocorre globalmente.

5.1.1 Representação da partícula

Em um dos primeiros trabalhos propostos usando otimização por nuvem de partículas para agrupamento de dados (MERWE; ENGELBRECHT, 2003), cada partícula do enxame foi representada por k centróides, onde cada centróide é representado como sendo um ponto no espaço R^p. Além disso, a velocidade das partículas foi representada por um vetor $k \times p$, onde cada componente do vetor é atualizado de acordo com a Eq. definida no Cap.2. Dessa forma, para atualizar a posição das partículas, seria necessário atualizar a matriz velocidade e, posteriormente, usar a matriz velocidade para atualizar a matriz que representa os centróides.

No entanto, ao agrupar dados relacionais, o conceito de centróide não pode ser usado diretamente, pois os atributos ou características dos objetos não estão disponíveis. Uma

abordagem alternativa é realizar o agrupamento de dados relacionais usando *medóides*, pois parece ser mais adequado usar os próprios objetos da base de dados para representar os grupos. A utilização de medóides para representação de grupos tem duas principais vantagens. A primeira delas é que, como os protótipos são formados por objetos da própria base de dados, o cálculo da dissimilaridade entre os objetos e os protótipos se torna trivial, pois as dissimilaridades já são fornecidas nas matrizes de entrada. A segunda vantagem é que algoritmos baseados em medóides podem lidar naturalmente com matrizes relacionais de qualquer tipo, incluindo as não euclidianas (HORTA, 2013). Contudo, a utilização de um único *medóide* como representativo pode ser insuficiente para caracterizar os grupos. Dessa forma, a utilização de múltiplos *medóides* pode servir para representar cada grupo de forma mais precisa (MEI; CHEN, 2010; WANG; CHEN; MEI, 2014).

Nesta tese, cada partícula p_i do enxame foi definida como um vetor de representativos $(G_{i,1}, ..., G_{i,K})$ em que o representativo $G_{i,k}$, associado ao *cluster* $C_{i,k}$, é subconjunto de cardinalidade fixa q do conjunto de dados E, isto é, $G_{i,k}$ é um subconjunto de E de cardinal q. De forma simplificada, cada grupo é representado por múltiplos *medóides*.

$$p_i = \begin{bmatrix} G_{i,1} \\ \vdots \\ G_{i,K} \end{bmatrix}$$

Cada partícula p_i do enxame procura por uma partição $P_i = (C_{i,1}, ..., C_{i,K})$ de E em K *clusters* e nos representativos correspondentes $(G_{i,1}, ..., G_{i,K})$ que vão representar os *clusters* na partição P_i de forma que a função de aptidão seja otimizada, isto é, minimizada ou maximizada a depender da função escolhida.

A melhor posição da partícula p_i, denotada como $pbest_i$, representa o melhor conjunto de representativos $(G^*_{i,1}, .., G^*_{i,K})$ encontrado baseado no valor da aptidão. A melhor posição do enxame inteiro denotada como *gbest* corresponde a melhor posição $(G_{g,1}, .., G_{g,K})$ encontrada por todo o enxame. Cada partícula p_i tem uma velocidade $v_i^{(t)} = (v_{i,1}^{(t)}, .., v_{i,n}^{(t)})$ na iteração t, esta velocidade será usada para atualizar a posição atual representada pelo conjunto de representativos.

5.1.2 Entradas e saídas

A Tabela 1 apresenta as entradas e saídas dos métodos desenvolvidos. Os dados de entrada são: matrizes de dissimilaridades, número de grupos, número de partículas, número de iterações, fatores de aceleração, pesos de inércia e a função de aptidão a ser otimizada. Os métodos produzem as seguintes saídas: melhores representativos encontrados para cada grupo, partição final e os pesos de relevância estimados para cada matriz.

As subseções seguintes descrevem a inicialização das partículas, o cálculo para atualização da velocidade e o método de busca local para atualização da posição das partículas.

Tabela 1 – Entradas e saídas

Tipo	Descrição	Notação
entrada	matrizes de dissimilaridades	D_j, $1 \leq j \leq p$
entrada	número de grupos	K
entrada	número de partículas	n_p
entrada	número de iterações	T_{max}
entrada	fatores de aceleração	c_1, c_2
entrada	pesos de inércia w	w_{min}, w_{max}
entrada	função de aptidão a ser otimizada	ff
saída	melhores representativos	$(G_{g,1}, .., G_{g,K})$
saída	partição final	$P = \{C_1, .., C_K\}$
saída	vetor de pesos de relevância estimados globalmente	λ
saída	vetor de pesos de relevância estimados localmente	λ_k

Após a inicialização das partículas, a posição com melhor aptidão será usada para representar o *gbest*. A cada iteração, a velocidade é atualizada e é usada para influenciar na atualização dos representativos e, consequentemente, modificar a partição atual. A aptidão das partículas é calculado novamente, o *pbest* de cada partícula e o *gbest* são atualizados se uma posição melhor for encontrada. Este processo continua até que o critério de parada seja satisfeito.

5.1.3 Inicialização das partículas

A inicialização é descrita no Algoritmo 9 para os métodos rígidos. A inicialização de cada partícula define a posição inicial no espaço de busca, isto é, define os múltiplos *medóides* iniciais que irão representar os grupos e, consequentemente, a partição inicial no espaço de busca das soluções a ser explorado pelas partículas. Primeiramente, os representativos da partícula p_i, $G_{i,k}$ de cada *cluster* $C_{i,k}$ são selecionados de forma aleatória. Para os métodos PSO-RWL e PSO-RWG, os vetores de pesos de relevância também são inicializados, linhas 5 e 6, com todas as visões tendo pesos iguais. Em seguida, cada objeto é atribuído ao *cluster* que possui o representativo mais próximo. No caso do método que considera apenas uma matriz, representado na linha 6 do algoritmo, os pesos de relevância não são considerados. Os casos em que existem várias matrizes sem pesos de relevância, várias matrizes com pesos estimados localmente e vários pesos estimados globalmente são apresentados nas linhas 7, 8 e 9 do Algoritmo, respectivamente. Por fim, a aptidão inicial é computada.

Para agrupamento nebuloso, a inicialização é descrita no Algoritmo 10. Neste caso, diferentemente do que ocorre no agrupamento rígido, um grau de pertinência de cada objeto para cada grupo é definido. A linha 5 do algoritmo é utilizada no método que considera apenas uma matriz. Os casos em que existem várias matrizes sem pesos de relevância, várias matrizes com pesos estimados localmente e vários pesos estimados globalmente são apresentados nas linhas 6, 7 e 8 do Algoritmo, respectivamente. Por fim, a aptidão inicial é computada.

Algoritmo 9: Inicialização de cada partícula p_i para agrupamento rígido

1: **RWL**: Inicializar $\lambda_k^{(0)} = (\lambda_{k1}^{(0)}, ..., \lambda_{kp}^{(0)}) = (1, ..., 1)(k = 1, ..., K)$
2: **RWG**: Inicializar $\lambda^{(0)} = (\lambda_1^{(0)}, ..., \lambda_k^{(0)}) = (1, ..., 1)(k = 1, ..., K)$
3:
4: Aleatoriamente selecionar K protótipos distintos $G_{i,k}^{(0)} \in E^{(q)}(k = 1, .., K)$ para obter o vetor de representativos $(G_{i,1}^{(0)}, .., G_{i,k}^{(0)})$;
5: Atribuir cada objeto e_i ao cluster com representativo mais próximo
6:
7: A partição $P_i^{(0)} = (C_{i,1}^{(0)}, ..., C_{i,K}^{(0)})$ é obtida de acordo com:
8: **SV**: $C_{i,k}^{(0)} = \{e_i \in E : D(e_j, G_{i,k}^{(0)}) \leq D(e_j, G_{i,h}^{(0)}), (h = 1, .., K))\}$
9: **MV**: $C_{i,k}^{(0)} = \{e_i \in E : \sum_{j=1}^{p} (0) D_j(e_j, G_{i,k}^{(0)}) \leq \sum_{j=1}^{p} D_j(e_j, G_{i,h}^{(0)})\}$
10: **RWL**: $C_{i,k}^{(0)} = \{e_i \in E : \sum_{j=1}^{p} \lambda_{kj}^{(0)} D_j(e_j, G_{i,k}^{(0)}) \leq \sum_{j=1}^{p} \lambda_{hj}^{(0)} D_j(e_j, G_{i,h}^{(0)})\}$
11: **RWG**: $C_{i,k}^{(0)} = \{e_i \in E : \sum_{j=1}^{p} \lambda_k^{(0)} D_j(e_j, G_{i,k}^{(0)}) \leq \sum_{j=1}^{p} \lambda_h^{(0)} D_j(e_j, G_{i,h}^{(0)})\}$
12:
13: Computar a aptidão inicial

Algoritmo 10: Inicialização de cada partícula p_i para agrupamento nebuloso

1: **RWL**: Inicializar $\lambda_k^{(0)} = (\lambda_{k1}^{(0)}, ..., \lambda_{kp}^{(0)}) = (1, ..., 1)(k = 1, ..., K)$
2: **RWG**: Inicializar $\lambda^{(0)} = (\lambda_1^{(0)}, ..., \lambda_p^{(0)}) = (1, ..., 1)$
3:
4: Aleatoriamente selecionar K protótipos distintos $G_{i,k}^{(0)} \in E^{(q)}(k = 1, .., K)$ para obter o vetor de representativos $(G_{i,1}^{(0)}, .., G_{i,k}^{(0)})$;
5:
6: Para cara objeto e_l, computar o grau de pertinência $u_{lk}^{(0)}$ $(k = 1, .., K)$:
7: **SV**: $u_{lk}^{(0)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{e \in G_k^{(0)}} d(e_l, e)}{\sum_{e \in G_h^{(0)}} d(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$
8: **MV**: $u_{lk}^{(0)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \sum_{e \in G_k^{(0)}} d_j(e_l, e)}{\sum_{j=1}^{p} \sum_{e \in G_h^{(0)}} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$
9: **RWL**: $u_{lk}^{(0)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \sum_{e \in G_k^{(0)}} \lambda_{kj} d_j(e_l, e)}{\sum_{j=1}^{p} \sum_{e \in G_h^{(0)}} \lambda_{kj} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$
10: **RWL**: $u_{lk}^{(0)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \sum_{e \in G_k^{(0)}} \lambda_k d_j(e_l, e)}{\sum_{j=1}^{p} \sum_{e \in G_h^{(0)}} \lambda_j d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$
11:
12: Computar a aptidão inicial

5.1.4 Atualização da posição

A atualização da posição é descrita nos algoritmos 12 e 13. O primeiro descreve os passos para atualização no caso de agrupamento rígido e o segundo descreve os passos para agrupamento nebuloso.

Algoritmo 11: Cálculo dos representativos para a partícula p_i e *cluster* C_k

Seja $T = \{< e_1, c_1 >, \ldots, < e_l, c_l >, \ldots, < e_n, c_n >\}$

Para cada tupla $< e_l, c_l > \in T$ **faça**

- **SV** : $c_l \leftarrow \sum_{e_m \in C_{i,k}^{(t)}} d(e_m, e_l) + v_{i,k,l}^{(t)}$
- **MV** : $c_l \leftarrow \sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} d_j(e_m, e_l) + v_{i,k,l}^{(t)}$
- **RWL** : $c_l \leftarrow \sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{kj}^{(t)} d_j(e_m, e_l) + v_{i,k,l}^{(t)}$
- **RWG** : $c_l \leftarrow \sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{j}^{(t)} d_j(e_m, e_l) + v_{i,k,l}^{(t)}$

Ordenar T de acordo com o valor do segundo componente das tuplas

$G_{i,k}^{(t)} \leftarrow \emptyset$

Para $l = 1$ até q **faça**

- Selecionar a l^{th} tupla $< e_m, c_m >$ of T
- $G_{i,k}^{(t)} = G_{i,k}^{(t)} \cup \{e_m\}$

Em todas as abordagens, o procedimento descrito no Alg. 11 modifica a posição atual de cada partícula, ou seja, os representativos $G_{i,k}(k = 1, .., K)$ de forma que a partição atual mude e seja possível explorar o espaço de busca na tentativa de encontrar soluções melhores. Isto é feito utilizando a velocidade da partícula corrente, esta velocidade irá afetar a escolha dos próximos representativos de cada *cluster*, isto é, os próximos representantes vão ser escolhidos de acordo com a minimização das equações (5.1), (5.2), (5.3) e (5.4).

$$\sum_{e_m \in C_{i,k}^{(t)}} d(e_m, e_l) + v_{i,k,l}^{(t)}. \tag{5.1}$$

$$\sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} d_j(e_m, e_l) + v_{i,k,l}^{(t)}. \tag{5.2}$$

$$\sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{kj}^{(t)}\, d_j(e_m, e_l) + v_{i,k,l}^{(t)}. \tag{5.3}$$

$$\sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{j}^{(t)}\, d_j(e_m, e_l) + v_{i,k,l}^{(t)}. \tag{5.4}$$

Em modelagens anteriores que utilizaram PSO para agrupamento de dados vetoriais, as partículas foram representados por k centróides e a aplicação da velocidade para alterar

a posição ocorria com uma soma de matrizes, em que a matriz que representa a velocidade é somada a matriz posição. No entanto, para agrupamento de dados relacionais isso não é possível. Dessa forma, pensou-se na estratégia de aplicação da velocidade para afetar a escolha dos representativos.

Para cada objeto será calculado a média das distâncias para todos os objetos dentro de um grupo somado a velocidade. Portanto, o próximo representativo será composto pelos objetos que terão os menores valores para essa soma. Dessa forma, a velocidade afeta a escolha dos representativos, o que diversifica e torna possível uma diferente exploração no espaço de busca. Diferentemente do que foi realizado em (CARVALHO; LECHEVALLIER; MELO, 2012), em que os representativos são escolhidos considerando a minimização de (5.5) ou (5.6) para os pesos de relevância estimados localmente e globalmente, respectivamente. A estratégia teve como objetivo prevenir que as partículas tenham uma convergencia rápida e fiquem presas em pontos que representam mínimos locais. Essa estratégia foi usada em todos os cenários avaliados neste trabalho.

$$\sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{kj}^{(t)} \, d_j(e_m, e_l) \tag{5.5}$$

$$\sum_{e_m \in C_{i,k}^{(t)}} \sum_{j=1}^{p} \lambda_{j}^{(t)} \, d_j(e_m, e_l) \tag{5.6}$$

5.1.5 Atualização da velocidade

A velocidade de cada partícula é usada para atualizar a posição atual. Neste caso, a velocidade é um vetor n-dimensional $v_i^{(t)} = (v_{i,1}^{(t)}, .., v_{i,n}^{(t)})$, este vetor será usado para afetar a seleção dos novos representativos da fase de busca local e fornecer mais diversidade ao processo de escolha dos próximos representativos.

A velocidade é atualizada de acordo com a seguinte equação:

$$v_{i,k,j}^{(t)} = w^{(t)} \times v_{i,k,j}^{(t-1)} + c_1 \times r_1 \times \Delta_c(G_{i,k}^{*(t)}, e_j) + c_2 \times r_2 \times \Delta_s(G_{g,k}^{(t)}, e_j). \tag{5.7}$$

em que w, c_1, c_2, r_1 e r_2 são

- w representa o peso de inércia na atualização da velocidade;
- c_1 e c_2 são fatores de aceleração e influenciam na importância dos componentes cognitivo e social para atualização da velocidade;
- r_1 e r_2 são números aleatórios entre 0 e 1.

Na versão proposta por (KENNEDY; EBERHART, 1995) do algoritmo otimização por nuvem de partículas, a velocidade é calculada baseando-se em três componentes principais: (1) velocidade anterior, (2) componente cognitivo e (3) componente social.

Algoritmo 12: Atualização da posição de p_i para agrupamento rígido

Passo 1: Busca pelos melhores representativos

O representativo $G_{i,k} = G^* \in E^{(q)}$ do cluster $C_{i,k}(k = 1, .., K)$ é computado de acordo com o Alg. 11

Passo 2: Atualização da partição

Os representativos $(G_{i,1}, .., G_{i,K})$ são fixados para atualizar a partição;

$P_i^{(t)} = P_i^{(t-1)}$;

Definir $test = 0$

Para $j = 1$ até n **faça**

Encontrar índice m:$e_j \in P_{i,m}^{(t)}$

Determinar o *cluster* vencedor $C_{i,k}^{(t)}$ tal que:

SV: $k = argmin_{1 \leq h \leq K} D(e_j, G_{i,k})$

MV: $k = argmin_{1 \leq h \leq K} D(e_j, G_{i,k})$

RWL: $k = argmin_{1 \leq h \leq K} D_{\lambda_k}(e_j, G_{i,k})$

RWL: $k = argmin_{1 \leq h \leq K} D_{\lambda}(e_j, G_{i,k})$

Se $k \neq m$ **então**

$test = 1$;

$C_{i,k}^{(t)} = C_{i,k}^{(t)} \cup \{e_j\}$;

$C_{i,m}^{(t)} = C_{i,m}^{(t)} / e_j$;

Computar a aptidão atual

A dissimilaridade entre um objeto e_j e o $pbest_i$ é medida pela maior relação entre o objeto e_j e o representativo do *cluster* k do *pbest* de forma que $e_j \in C_{i,k}^*$. Esta dissimilaridade é usada no componente cognitivo e é denotada por $\Delta(G_{i,k}^*, e_j)$. Nos casos em que há apenas uma visão ou múltiplas visões com pesos iguais, $\Delta_s(G_{i,k}^*, e_j)$ é definido em (5.8) e (5.9). Ao considerar os pesos estimados localmente, $\Delta(G_{i,k}^*, e_j)$ é definido em (5.10) e, considerar pesos globais, é definida em (5.11).

$$\Delta_1(G_{i,k}^*, e_j) = \max_{e \in G_{i,k}} d(e_j, e). \tag{5.8}$$

$$\Delta_2(G_{i,k}^*, e_j) = \max_{e \in G_{i,k}} \sum_{l=1}^{p} d_l(e_j, e). \tag{5.9}$$

$$\Delta_3(G_{i,k}^*, e_j) = \max_{e \in G_{i,k}} \sum_{l=1}^{p} \lambda_{kl}\, d_l(e_j, e). \tag{5.10}$$

$$\Delta_4(G_{i,k}^*, e_j) = \max_{e \in G_{i,k}} \sum_{l=1}^{p} \lambda_l\, d_l(e_j, e). \tag{5.11}$$

Algoritmo 13: Atualização da posição de p_i para agrupamento nebuloso

Passo 1: busca pelos representativos

Os representativos $G_{i,k}^{(t)}$ $(k = 1, .., K)$ são calculados de acordo com o Alg. 11

Passo 2: cálculo dos pesos de relevância

Calcular os pesos de acordo com Eq. (2.6) para $FPSO_{RWL}$ ou de acordo com Eq. (2.8) para $FPSO_{RWG}$

Passo 3: cálculo da partição nebulosa

Para $l = 1$ até n **faça**

SV: $u_{lk}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{e \in G_k^{(t)}} d(e_l, e)}{\sum_{e \in G_h^{(t)}} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$

MV: $u_{lk}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \sum_{e \in G_k^{(t)}} d_j(e_l, e)}{\sum_{j=1}^{p} \sum_{e \in G_h^{(t)}} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$

RWL: $u_{lk}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \lambda_{kj} \sum_{e \in G_k^{(t)}} d_j(e_l, e)}{\sum_{j=1}^{p} \lambda_{kj} \sum_{e \in G_h^{(t)}} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$

RWG: $u_{lk}^{(t)} = \left[\sum_{h=1}^{K} \left(\frac{\sum_{j=1}^{p} \lambda_j \sum_{e \in G_k^{(t)}} d_j(e_l, e)}{\sum_{j=1}^{p} \lambda_j \sum_{e \in G_h^{(t)}} d_j(e_l, e)} \right)^{\frac{1}{m-1}} \right]^{-1}$

Computar a aptidão atual

A dissimilaridade entre um objeto e_j e o *gbest* é mensurada pela maior relação entre o objeto e_j e o representativo do *cluster* k associado ao *gbest* de forma que $e_j \in C_{g,k}$. Esta dissimilaridade é usada no componente social e é denotada por $\Delta_s(G_{g,k}, e_j)$.

$$\Delta_{s1}(G_{g,k}, e_j) = \max_{e \in G_{g,k}} d(e_j, e). \tag{5.12}$$

$$\Delta_{s2}(G_{g,k}, e_j) = \max_{e \in G_{g,k}} \sum_{l=1}^{p} d_l(e_j, e). \tag{5.13}$$

$$\Delta_{s3}(G_{g,k}, e_j) = \max_{e \in G_{g,k}} \sum_{l=1}^{p} \lambda_{kl}\, d_l(e_j, e). \tag{5.14}$$

$$\Delta_{s4}(G_{g,k}, e_j) = \max_{e \in G_{g,k}} \sum_{l=1}^{p} \lambda_l\, d_l(e_j, e). \tag{5.15}$$

É importante ressaltar que as dissimilaridades $\Delta(G_{i,k}^{*}, e_j)$ e $\Delta(G_{g,k}, e_j)$ poderiam ser definidas de outras formas como, por exemplo, a dissimilaridade mínima ou, ainda, a dissimilaridade média entre e_j e todos os objetos de $G_{i,k}^{*}$ ou $G_{g,k}$. Estas definições não foram consideradas neste trabalho. Contudo, vale a pena ressaltar que as dissimilaridades mínimas e médias entre os objetos e os representantes de cada grupo poderiam tornar a convergência das partículas mais lenta e, com isso, o enxame possivelmente exploraria mais o espaço de busca.

5.2 COMPLEXIDADE COMPUTACIONAL

Nesta seção são apresentadas as complexidades computacionais de tempo associadas aos métodos para agrupamento rígido e nebuloso de dados relacionais com múltiplas visões.

5.2.1 Agrupamento rígido

A complexidade de tempo do PSO_{RWL} pode ser analisada considerando a complexidade de cada etapa. Seja n o número de objetos, $K << n$ o número de *clusters*, $q << n$ a cardinalidade de cada representativo e p o número de matrizes de dissimilaridades, os passos do algoritmo podem ser analisados da seguinte forma:

- Inicialização das partículas: A inicialização do vetor de pesos de relevância tem complexidade $O(K \times p)$. A seleção aleatória dos K protótipos distintos ($K \times q$ objetos distintos) é feita em $O(K \times q)$. A atribuição de cada objeto ao *cluster* com protótipo mais próximo tem complexidade $O(n \times K \times q \times p)$. Dessa forma, como existem n_p partículas no enxame, o processo de inicialização é feito em $O(n_p \times n \times K \times q \times p)$.

- Atualização da velocidade: cada partícula tem sua velocidade atualizada com complexidade de $O(K \times n \times q \times p)$. Para todo o enxame, a complexidade é $O(n_p \times K \times n \times q \times p)$.

- Atualização da posição:

  1. Cálculo dos protótipos: para cada objeto, um custo é associado que mede a distância entre o objeto e todos os objetos em cada *cluster*. Esse processo tem complexidade $O(n^2 \times K \times p)$. A seleção de q objetos para compor os representativos de cada *cluster* precisa ordenar os objetos com base no custo calculado anteriormente, a ordenação é feita em $O(K \times n \times log(n))$. Logo, esse passo tem complexidade $O(n^2 \times K \times p)$.
  2. Cálculo do vetor de pesos de relevância: $O(n \times K \times p \times q)$.
  3. Definição da melhor partição: nesse passo, é necessário computar a distância entre cada objeto e os protótipos para decidir em qual *cluster* os objetos serão atrbuídos. Este passo tem complexidade $O(n \times q \times K \times p)$.

- Dessa forma, a complexidade para atualizar a posição de todas as partículas é $O(n_p \times n^2 \times K \times p)$.

- Finalmente, o PSO_{RWL} executa com um número máximo de iterações T_{max}. Portanto, a complexidade do método é $O(n_p \times n^2 \times K \times p \times T_{max})$.

Um raciocínio similar pode ser feito para o método PSO_{RWG}. Dessa forma, é possível concluir que o método tem a mesma complexidade.

5.2.2 Agrupamento nebuloso

A complexidade de tempo do método $FPSO_{RWL}$ pode ser analisada da seguinte forma:

- Inicialização: A inicialização do vetor de pesos custa $O(K \times p)$. A escolha aleatória dos K representativos distintos ($K \times q$ distinct objects) pode ser efetuada em $O(K \times q)$. O cálculo do grau de pertinência de cada objeto tem complexidade $O(n \times K^2 \times q \times p)$. Como existe n_p partículas no enxame, a inicialização leva $O(n_p \times n \times K^2 \times q \times p)$.
- Atualização da velocidade: cada partícula atualiza sua velocidade em $O(K \times n \times q \times p)$. Consequentemente, para o enxame inteiro, a atualização custa $O(n_p \times K \times n \times q \times p)$.
- Atualização da posição:
  1. Cálculo dos prototipos: para cada objeto, um custo medindo a dissimilaridade entre o objeto e todos os elementos atribuídos a cada grupo considerando todas as matrizes é calculado. Esse processo custa $O(n^2 \times K \times p)$. A seleção de q objetos precisa ordenar os objetos com base no custo computado anteriormente, a ordenação pode ser efetuada em $O(K \times n \times log(n))$. O cálculo dos prototipos custa $O(n^2 \times K \times p)$.
  2. Cálculo do vetor de pesos de relevância: $O(n \times K \times p \times q)$.
  3. Definição da partição nebulosa: neste passo, é necessário computar os graus de pertinência. Esse passo custa $O(n \times K^2 \times q \times p)$.
- Dessa forma, a complexidade para atualizar a posição de todo o enxame é $O(n_p \times n^2 \times K \times p)$.
- Finalmente, o método $FPSO - RWL$ tem T_{max} iterações. Por conseguinte, a complexidade de tempo do método é $O(n_p \times n^2 \times K \times p \times T_{max})$.

Um raciocínio similar pode ser feito para o método $FPSO_{RWG}$. Dessa forma, é possível concluir que o método tem a mesma complexidade.

5.3 SÍNTESE DA SEÇÃO

Neste capítulo, a metodologia desenvolvida nesta tese foi apresentada. Métodos híbridos para agrupamento rígido e nebuloso de dados relacionais com múltiplas visões foram desenvolvidos. Os métodos híbridos são inspirados na otimização por nuvem de partículas para convergência global e em algoritmos baseados em múltiplas matrizes de dissimilaridade com estimação de pesos de relevância para exploração local. Detalhes sobre representação das partículas, inicialização das partículas, atualização da posição, atualização da velocidade e sobre a complexidade foram detalhados. No próximo capítulo será apresentada a validação experimental realizada para os métodos para agrupamento rígido.

6 ÍNDICES PARA VALIDAÇÃO DE AGRUPAMENTOS

Neste capítulo serão apresentados os índices para validação de agrupamentos que foram utilizados nesta tese. No caso dos métodos híbridos para agrupamento de dados relacionais com várias visões, alguns desses índices foram adaptados. A adaptação envolveu três mudanças em relação a definição tradicional dos índices para dados vetoriais: (i) os índices considerados para agrupamento de dados relacionais com múltiplas visões consideram múltiplos *medóides* como representativos dos grupos diferentemente dos centróides quando os dados são representados por vetores de atributos; (ii) os índices passam a considerar dissimilaridades fornecidas por várias matrizes simultaneamente; e (iii) as dissimilaridades fornecidas pelas matrizes são ponderadas através do uso de pesos de relevância estimados localmente ou globalmente.

Na seção 6.1.1, serão apresentados os critérios internos que foram usados como funções de aptidão para os métodos híbridos desenvolvidos para agrupamento rígido. Os critérios internos considerados para agrupamento nebuloso são apresentados na seção 6.1.2. Por fim, os critérios externos de validação de agrupamentos que foram utilizados são definidos na seção 6.2.

6.1 CRITÉRIOS INTERNOS

Nas subseções 6.1.1 e 6.1.2, cinco definições serão apresentadas para a maioria dos índices. A primeira definição é a que considera o conjunto de dados representados na forma de uma matriz $X = \{x_1, ..., x_l, ..., x_n\}$, em que os *clusters* são representados por um vetor de centróides $(c_1, ..., c_K)$.

$$X_{n\times r} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nr} \end{bmatrix}$$

As demais definições de cada índice são referentes aos casos em que os dados são representados por: (i) apenas uma matriz, (ii) p matrizes com pesos de relevância iguais no processo de agrupamento, (iii) p matrizes com pesos de relevância estimados localmente para cada grupo e (iv) p matrizes com pesos de relevância estimados globalmente. Além disso, os grupos passam a ser representados por múltiplos *medóides*, denotados por $(G_1, ..., G_K)$, em que $G_k \in E^{(q)} = \{A \subset E : |A| = q\}$ $(k = 1, .., K)$.

6.1.1 Agrupamento rígido

Os índices para agrupamento rígido que foram considerados nesta tese são sumarizados na Tabela 2.

Tabela 2 – Sumário de índices internos para agrupamento rígido

Notação	Critério	Objetivo	Fórmula
HM	Homogeneidade	Minimização	6.1
QE	Erro de quantização	Minimização	6.57
SIL	Silhueta	Maximização	6.11
SS	Silhueta simplificada	Maximização	6.11
DB	Davies & Bouldin	Minimização	6.35
DN	Dunn	Maximização	6.46
HA	Hartigan	Maximização	6.62
CS	CS	Maximização	6.68
CH	Calinsky-Harabasz	Maximização	6.76
XU	Xu	Minimização	6.80
WB	WB	Minimização	6.85

6.1.1.1 Homogeneidade

Este índice mede a homogeneidade da partição $\mathcal{P}$ como a soma das homogeneidades em cada grupo. Na equação (6.1) e em equações que serão apresentadas posteriormente , *dist* representa a distância euclideana. Na Eq. (6.1), por exemplo, $dist(x_j, c_k)$ representa a distância euclideana entre o objeto x_j e o centróide c_k.

$$HM = \sum_{k=1}^{K} \sum_{x_l \in C_k} dist(x_l, c_k) \tag{6.1}$$

Neste trabalho, o índice é definido conforme (6.2), (6.3), (6.4) e (6.5) para os casos em que os dados possuem apenas uma visão, quando os dados possuem várias visões com mesmo peso, quando possuem várias visões com pesos locais e com várias visões com pesos globais, respectivamente..

$$HM = \sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{e \in G_k} d(e_l, e) \tag{6.2}$$

$$HM = \sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{j=1}^{p} \sum_{e \in G_k} d_j(e_l, e) \tag{6.3}$$

$$HM = \sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{j=1}^{p} \lambda_{kj} \sum_{e \in G_k} d_j(e_l, e) \tag{6.4}$$

$$HM = \sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{j=1}^{p} \lambda_{j} \sum_{e \in G_k} d_j(e_l, e) \tag{6.5}$$

6.1.1.2 Silhueta

Outro índice comumente usado para agrupamento de dados é o índice da silhueta proposto por Rousseeuw (1987). O índice da silhueta considera a homogeneidade em cada grupo bem como a separação entre eles. Este índice é definido como a média da silhueta de todos os objetos (6.6).

$$SIL = \frac{1}{n} \sum_{x_l \in X} s(x_l) \tag{6.6}$$

onde

$$s(x_l) = \begin{cases} \frac{b(x_l)-a(x_l)}{max(b(x_l),a(x_l))} & if \quad a(x_l) \neq b(x_l) \\ 0 & if \quad a(x_l) = b(x_l) \end{cases} \tag{6.7}$$

A silhueta de um objeto x_l é definida na equação (6.7). Considerando todos os grupos C_t tal que $x_l \notin C_t$, calcula-se $d(x_l, C_m)$ como a distância média do objeto x_l para todos os objetos do grupo C_m. Além disso, calcula-se $a(x_l)$ como sendo a distância média entre x_l e todos os outros objetos pertencentes ao mesmo grupo. O componente $b(e_l)$ é a distância média mínima de x_l para outros grupos (Eq. 6.21).

$$a(x_l) = \frac{\sum_{x \in C_k} dist(x_l, x)}{|C_k|} \tag{6.8}$$

$$d(x_l, C_m) = \frac{\sum_{x \in C_m} dist(x_l, x)}{|C_m|} \tag{6.9}$$

$$b(x_l) = \min_{C_m \neq C_k} d(x_l, C_m) \tag{6.10}$$

A silhueta de um objeto e_l é definida na Eq. (6.12). Considerando todos os grupos C_t tal que $e_l \notin C_t$, calcula-se $d(e_l, C_t)$ como a dissimilaridade média do objeto e_l para todos os objetos do grupo C_t. A Figura 4 ilustra o cálculo dos componentes $a(e_i)$ e $d(e_i, C_t)$. No exemplo, e_l pertence a C_1. Dessa forma, calcula-se $a(e_l)$ como sendo a dissimilaridade média entre e_l e todos os outros objetos pertencentes a C_1. O componente $b(e_l)$ é a dissimilaridade média mínima de e_l para outros *clusters* (Eq. 6.21).

$$SIL = \frac{1}{n} \sum_{e_l \in E} s(e_l) \tag{6.11}$$

$$s(e_l) = \begin{cases} \frac{b(e_l)-a(e_l)}{max(b(e_l),a(e_l))} & se \quad a(e_l) \neq b(e_l) \\ 0 & se \quad a(e_i) = b(e_i) \end{cases} \tag{6.12}$$

Quando os dados são representados através de matrizes de dissimilaridades, $a(e_l)$ passa a ser definido como definido em (6.13), (6.14), (6.15) e (6.16) para os casos em que os dados possuem apenas uma visão, quando os dados possuem várias visões com mesmo

peso, quando possuem várias visões com pesos locais e com várias visões com pesos globais, respectivamente.

$$a(e_l) = \frac{\sum_{e \in C_k} D(e_l, e)}{|C_k|} = \frac{\sum_{e \in C_k} d(e_l, e)}{|C_k|} \tag{6.13}$$

$$a(e_l) = \frac{\sum_{e \in C_k} D(e_l, e)}{|C_k|} = \frac{\sum_{e \in C_k} \sum_{j=1}^{p} d_j(e_l, e)}{|C_k|} \tag{6.14}$$

$$a(e_l) = \frac{\sum_{e \in C_k} D_{\lambda_k}(e_l, e)}{|C_k|} = \frac{\sum_{e \in C_k} \sum_{j=1}^{p} \lambda_{kj} d_j(e_l, e)}{|C_k|} \tag{6.15}$$

$$a(e_l) = \frac{\sum_{e \in C_k} D_{\lambda}(e_l, e)}{|C_k|} = \frac{\sum_{e \in C_k} \sum_{j=1}^{p} \lambda_j d_j(e_l, e)}{|C_k|} \tag{6.16}$$

Similarmente, $d(e_l, C_t)$ passa a ser definido como definido em (6.17), (6.18), (6.19) e (6.20) para os casos em que os dados possuem apenas uma visão, quando os dados possuem várias visões com mesmo peso, quando possuem várias visões com pesos locais e com várias visões com pesos globais, respectivamente.

$$d(e_l, C_t) = \frac{\sum_{e \in C_t} D(e_l, e)}{|C_t|} = \frac{\sum_{e \in C_t} d(e_l, e)}{|C_t|} \tag{6.17}$$

$$d(e_l, C_t) = \frac{\sum_{e \in C_t} D(e_l, e)}{|C_t|} = \frac{\sum_{e \in C_t} \sum_{j=1}^{p} d_j(e_l, e)}{|C_t|} \tag{6.18}$$

$$d(e_l, C_t) = \frac{\sum_{e \in C_t} D_{\lambda_t}(e_l, e)}{|C_t|} = \frac{\sum_{e \in C_t} \sum_{j=1}^{p} \lambda_{tj} d_j(e_l, e)}{|C_t|} \tag{6.19}$$

$$d(e_l, C_t) = \frac{\sum_{e \in C_t} D_{\lambda_t}(e_l, e)}{|C_t|} = \frac{\sum_{e \in C_t} \sum_{j=1}^{p} \lambda_t d_j(e_l, e)}{|C_t|} \tag{6.20}$$

$$b(e_l) = \min_{C_j \neq C_k} d(e_l, C_j) \tag{6.21}$$

O valor da silhueta de um objeto está no intervalo [-1,1]. Valores próximos a 1 indicam que o objeto está bem agrupado; valores próximos a 0 indicam que o objeto está entre *clusters*, e valores negativos indicam que o objeto está no grupo errado. Dessa forma, o índice da silhueta do agrupamento deve ser maximizado.

6.1.1.3 Silhueta Simplificada

Devido a complexidade computacional do índice da silhueta, a silhueta simplificada foi proposta por Naldi et al. (2011). Neste caso, os termos $a(x_l)$ e $d(x_l, C_m)$ são modificados e passam a ser calculados considerando apenas a distância dos objetos aos representativos de cada grupo.

$$SS = \frac{1}{n} \sum_{x_l \in X} ss(x_l) \tag{6.22}$$

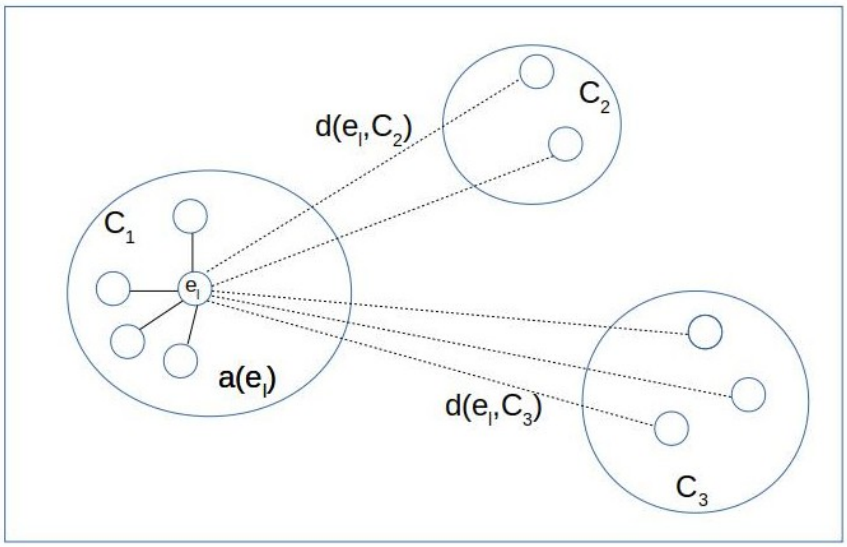


Figura 4 – Exemplo sobre cálculo de $a(e_l)$ e $d(e_l, C_t)$

onde

$$ss(x_l) = \begin{cases} \frac{b'(x_l)-a'(x_l)}{max(b'(x_l),a'(x_l))} & se \quad a'(x_l) \neq b'(x_l) \\ 0 & se \quad a'(x_l) = b'(x_l) \end{cases} \tag{6.23}$$

$$a'(x_l) = dist(x_l, c_k) \tag{6.24}$$

$$d'(x_l, C_m) = dist(x_l, c_m) \tag{6.25}$$

$$b'(x_l) = \min_{C_m \neq C_k} d'(x_l, C_m) \tag{6.26}$$

Quando os dados são representados através de matrizes de dissimilaridades, a' e d' passam a ser definidos como em (6.27),(6.28), (6.29), (6.30) e (6.31), (6.32), (6.33), (6.34), respectivamente.

$$a'(e_l) = \frac{D(e_l, G_k)}{|G_k|} = \frac{\sum_{e \in G_k} d(e_l, e)}{|G_k|} \tag{6.27}$$

$$a'(e_l) = \frac{D(e_l, G_k)}{|G_k|} = \frac{\sum_{j=1}^{p} \sum_{e \in G_k} d_j(e_l, e)}{|G_k|} \tag{6.28}$$

$$a'(e_l) = \frac{D_{\lambda_k}(e_l, G_k)}{|G_k|} = \frac{\sum_{j=1}^{p} \lambda_{kj} \sum_{e \in G_k} d_j(e_l, e)}{|G_k|} \tag{6.29}$$

$$a'(e_l) = \frac{D_{\lambda}(e_l, G_k)}{|G_k|} = \frac{\sum_{j=1}^{p} \lambda_{j} \sum_{e \in G_k} d_j(e_l, e)}{|G_k|} \tag{6.30}$$

$$d'(e_l, C_t) = \frac{D_{\lambda_t}(e_l, G_t)}{|G_t|} = \frac{\sum_{e_t \in G_t} d(e_l, e_t)}{|G_t|} \tag{6.31}$$

$$d'(e_l, C_t) = \frac{D_{\lambda_t}(e_l, G_t)}{|G_t|} = \frac{\sum_{j=1}^{p} \sum_{e_t \in G_t} d_j(e_l, e_t)}{|G_t|} \tag{6.32}$$

$$d'(e_l, C_t) = \frac{D_{\lambda_t}(e_l, G_t)}{|G_t|} = \frac{\sum_{j=1}^{p} \lambda_{tj} \sum_{e_t \in G_t} d_j(e_l, e_t)}{|G_t|} \tag{6.33}$$

$$d'(e_l, C_t) = \frac{D_{\lambda}(e_l, G_t)}{|G_t|} = \frac{\sum_{j=1}^{p} \lambda_{j} \sum_{e_t \in G_t} d_j(e_l, e_t)}{|G_t|} \tag{6.34}$$

6.1.1.4 Davies-Bouldin

O índice de Davies-Bouldin (DB) também foi proposto para minimizar a distância dentro dos grupos e maximizar a distância entre os grupos (DAVIES; BOULDIN, 1979). Para definição do índice DB é necessário calcular a dispersão e a similaridade entre os grupos. O termo o_k representa a dispersão de C_k e M_{km} representa a dissimilaridade entre C_k e C_m. Este índice é definido na Eq. (6.35).

$$DB = \frac{1}{K} \sum_{k=1}^{K} max_{m \in \{1,..,K\}} \left(\frac{o_k + o_m}{M_{km}} \right) \tag{6.35}$$

$$o_k = \left(\frac{1}{|C_k|} \sum_{x_l \in C_k} dist(x_l, c_k)) \right)^{\frac{1}{q}} \tag{6.36}$$

$$M_{km} = \max_{1 \leq i,k \leq K} dist(c_k, c_m) \tag{6.37}$$

Neste trabalho, para a representação de dados relacionais, o_k e M_{km} são calculados de acordo com (6.38),(6.40),(6.42), (6.44) e a Eq. (6.39),(6.41), (6.43) e (6.45), respectivamente.

$$o_k = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} D(e_l, G_k)) \right)^{\frac{1}{q}} = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} \sum_{e \in G_k} d(e_l, e)) \right)^{\frac{1}{q}} \tag{6.38}$$

$$M_{km} = \max_{e_l \in G_k, e_t \in G_m} d(e_l, e_t) \tag{6.39}$$

$$o_k = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} D(e_l, G_k)) \right)^{\frac{1}{q}} = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} d_j(e_l, e)) \right)^{\frac{1}{q}} \tag{6.40}$$

$$M_{km} = \max_{e_l \in G_k, e_t \in G_m} \sum_{j=1}^{p} d_j(e_l, e_t) \tag{6.41}$$

$$o_k = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} D_{\lambda_k}(e_l, G_k)) \right)^{\frac{1}{q}} = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_{kj} d_j(e_l, e)) \right)^{\frac{1}{q}} \tag{6.42}$$

$$M_{km} = \max_{e_l \in G_k, e_t \in G_m} \sum_{j=1}^{p} \lambda_{kj} d_j(e_l, e_t) \tag{6.43}$$

$$o_k = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} D_\lambda(e_l, G_k)) \right)^{\frac{1}{q}} = \left(\frac{1}{|C_k|} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_j d_j(e_l, e)) \right)^{\frac{1}{q}} \tag{6.44}$$

$$M_{km} = \max_{e_l \in G_k, e_t \in G_m} \sum_{j=1}^{p} \lambda_j d_j(e_l, e_t) \tag{6.45}$$

Este índice deve ser minimizado.

6.1.1.5 Dunn

Este índice foi proposto por Dunn (1973) e define a razão entre a mínima distância inter-cluster e o máximo diametro dos *clusters*. Os termos $\delta(C_i, C_k)$ e $\Delta(C_k)$ definem a distância entre os *clusters* (distância entre os representantes) e o diâmetro de C_k, respectivamente. Este índice é propenso a *outliers* uma vez que apenas considera distâncias máximas e mínimas. O índice é definido na Eq. (6.46) e deve ser maximizado.

$$DN = \frac{min_{k \in \{1,..,K\}, k \neq i} \delta(C_i, C_k)}{max_{k \in \{1,..,K\}} \Delta(C_k)} \tag{6.46}$$

$$\delta(C_i, C_k) = min_{i \neq k} dist(c_i, c_k) \tag{6.47}$$

$$\Delta(C_k) = \max_{x_m, x_t \in C_k} dist(x_m, x_t) \tag{6.48}$$

Neste trabalho, a dissimilaridade $\delta(C_i, C_k)$ e o diâmetro $\Delta(C_k)$ são definidos de acordo com as Eq. (6.49),(6.50), (6.51), (6.52) e com (6.53),(6.54),(6.55), (6.56), respectivamente.

$$\delta(C_i, C_k) = \min_{e_m \in G_i, e_t \in G_k} d(e_m, e_t) \tag{6.49}$$

$$\delta(C_i, C_k) = \min_{e_m \in G_i, e_t \in G_k} \sum_{j=1}^{p} d_j(e_m, e_t) \tag{6.50}$$

$$\delta(C_i, C_k) = \min_{e_m \in G_i, e_t \in G_k} \sum_{j=1}^{p} \lambda_{ij} d_j(e_m, e_t) \tag{6.51}$$

$$\delta(C_i, C_k) = \min_{e_m \in G_i, e_t \in G_k} \sum_{j=1}^{p} \lambda_j d_j(e_m, e_t) \tag{6.52}$$

$$\Delta(C_k) = \max_{e_m, e_t \in C_k} d(e_m, e_t) \tag{6.53}$$

$$\Delta(C_k) = \max_{e_m, e_t \in C_k} \sum_{j=1}^{p} d_j(e_m, e_t) \tag{6.54}$$

$$\Delta(C_k) = \max_{e_m, e_t \in C_k} \sum_{j=1}^{p} \lambda_{kj} d_j(e_m, e_t) \tag{6.55}$$

$$\Delta(C_k) = \max_{e_m, e_t \in C_k} \sum_{j=1}^{p} \lambda_{j} d_j(e_m, e_t) \tag{6.56}$$

6.1.1.6 Erro de quantização

O erro de quantização (MERWE; ENGELBRECHT, 2003) é outro índice comumente usado para calcular a aptidão de partículas. Este índice é definido na Eq. (6.57). Neste trabalho, este índice foi definido de acordo com as Eq. (6.58),(6.59), (6.60) e (6.61).

$$QE = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{x_l \in C_k} dist(x_l, c_k)}{|C_k|} \tag{6.57}$$

$$QE = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{e_l \in C_k} \sum_{e \in G_k} d(e_l, e)}{|C_k|} \tag{6.58}$$

$$QE = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} d_j(e_l, e)}{|C_k|} \tag{6.59}$$

$$QE = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_{kj} \times d_j(e_l, e)}{|C_k|} \tag{6.60}$$

$$QE = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_{j} \times d_j(e_l, e)}{|C_k|} \tag{6.61}$$

6.1.1.7 Hartigan

O índice de Hartigan proposto por Milligan e Cooper (1985) também foi projetado com o objetivo de minimizar a distância dentro de cada grupo e maximizar a distância entre eles. Este índice é definido na Eq. (6.62) para dados vetoriais.

$$HA = log\left(\frac{\sum_{i=1}^{K} |C_k| dist(\overline{x}, c_k)}{\sum_{k=1}^{K} \sum_{x_j \in C_k} dist(x_j, c_k)}\right) \tag{6.62}$$

Neste trabalho, o índice de Hartigan é definido de acordo com as Eq. (6.63), (6.64), (6.65) e (6.66).

$$HA = log\left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} d(g, e)}{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} d(e_j, e)}\right) \tag{6.63}$$

$$HA = log\left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \sum_{l=1}^{p} d_l(g, e)}{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \sum_{l=1}^{p} d_l(e_j, e)}\right) \tag{6.64}$$

$$HA = log\left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{kl} \times d_l(g, e)}{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{kl} \times d_l(e_j, e)}\right) \tag{6.65}$$

$$HA = log\left(\frac{\sum_{k=1}^{K}|C_k|\sum_{e\in G_k}\sum_{l=1}^{p}\lambda_l \times d_l(g,e)}{\sum_{k=1}^{K}\sum_{e_j\in C_k}\sum_{e\in G_k}\sum_{l=1}^{p}\lambda_l \times d_l(e_j,e)}\right) \tag{6.66}$$

O índice de Hartigan deve ser maximizado.

6.1.1.8 CS

O índice CS, proposto por Chou, Su e Lai (2004), é uma combinação entre diametros dos *clusters* e as distâncias entre os centros de cada grupo. Este índice foi desenvolvido para lidar com *clusters* com diferentes densidades e tamanhos. Este índice é definido na Eq. (6.68) e deve ser minimizado.

$$CS = \frac{\sum_{k=1}^{K}(\frac{1}{|C_k|}\sum_{x_j\in C_k} max_{x_l\in C_k} dist(x_j,x_l))}{\sum_{i=1}^{K}(min_{j\in\{1,..,K\},j\neq i} dist(c_i,c_j))} \tag{6.67}$$

Neste trabalho, o índice CS é definido de acordo com a Eq. (6.68). O termo $D(G_i,G_k)$ representa a dissimilaridade entre os representantes e é definido de acordo com as Eq. (6.69), (6.70), (6.71) e (6.72).

$$CS = \frac{\sum_{k=1}^{K}(\frac{1}{|C_k|}\sum_{e_j\in C_k} max_{e_l\in C_k} d(e_j,e_l))}{\sum_{i=1}^{K}(min_{j\in\{1,..,K\},j\neq i} D(G_i,G_j))} \tag{6.68}$$

$$D(G_i,G_k) = \min_{e_m\in G_i, e_t\in G_k} d(e_m,e_t) \tag{6.69}$$

$$D(G_i,G_k) = \min_{e_m\in G_i, e_t\in G_k} \sum_{j=1}^{p} d_j(e_m,e_t) \tag{6.70}$$

$$D(G_i,G_k) = \min_{e_m\in G_i, e_t\in G_k} \sum_{j=1}^{p} \lambda_{ij} d_j(e_m,e_t) \tag{6.71}$$

$$D(G_i,G_k) = \min_{e_m\in G_i, e_t\in G_k} \sum_{j=1}^{p} \lambda_j d_j(e_m,e_t) \tag{6.72}$$

6.1.1.9 Calinsky-Harabasz

O índice de Calinsky-Harabasz também foi desenvolvido para minimizar a distância intra-cluster e maximizar a separação entre os *clusters* (MILLIGAN; COOPER, 1985). Este índice é definido nas Eq. (6.74),(6.75),(6.74) e (6.77), e deve ser maximizado.

$$CH = \frac{\sum_{k=1}^{K}|C_k|dist(\overline{x},c_k)}{\sum_{k=1}^{K}\sum_{x_j\in C_k} dist(x_j,c_k)} \times \frac{n-K}{K-1} \tag{6.73}$$

$$CH = \left(\frac{\sum_{k=1}^{K}|C_k|\sum_{e\in G_k} d(g,e)}{\sum_{k=1}^{K}\sum_{e_l\in C_k}\sum_{e\in G_k} d(e_l,e)}\right)\left(\frac{n-K}{K-1}\right) \tag{6.74}$$

$$CH = \left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} d(g,e)}{\sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} d_j(e_l,e)} \right) \left(\frac{n-K}{K-1} \right) \tag{6.75}$$

$$CH = \left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} d(g,e)}{\sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_{kj} d_j(e_l,e)} \right) \left(\frac{n-K}{K-1} \right) \tag{6.76}$$

$$CH = \left(\frac{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} d(g,e)}{\sum_{k=1}^{K} \sum_{e_l \in C_k} \sum_{e \in G_k} \sum_{j=1}^{p} \lambda_j d_j(e_l,e)} \right) \left(\frac{n-K}{K-1} \right) \tag{6.77}$$

O representativo geral $g = e_l \in E$ é definido de acordo com a Eq. (6.78) para representação com única matriz e de acordo com a Eq. (6.79) para múltiplas matrizes.

$$l = argmin_{1 \leq h \leq n} \sum_{l=1}^{n} d(e_l, e_h) \tag{6.78}$$

$$l = argmin_{1 \leq h \leq n} \sum_{l=1}^{n} \sum_{j=1}^{p} d_j(e_l, e_h) \tag{6.79}$$

6.1.1.10 Xu

Este índice foi proposto por Xu (1997) e considera apenas a homogeneidade intra-cluster. O índice original é definido na Eq. (6.80). O índice é definido neste trabalho de acordo com as Eq. (6.81), (6.82), (6.83) e (6.84).

$$XU = log \left(\sqrt{\frac{\sum_{k=1}^{K} \sum_{x_j \in C_k} dist(x_j, c_k)}{n^2}} \right) + logK \tag{6.80}$$

$$XU = log \left(\sqrt{\frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} d(e_j, e)}{n^2}} \right) + logK \tag{6.81}$$

$$XU = log \left(\sqrt{\frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{l=1}^{p} \sum_{e \in G_k} d_l(e_j, e)}{n^2}} \right) + logK \tag{6.82}$$

$$XU = log \left(\sqrt{\frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{l=1}^{p} \lambda_{kl} \sum_{e \in G_k} d_l(e_j, e)}{n^2}} \right) + logK \tag{6.83}$$

$$XU = log \left(\sqrt{\frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{l=1}^{p} \lambda_j \sum_{e \in G_k} d_l(e_j, e)}{n^2}} \right) + logK \tag{6.84}$$

6.1.1.11 WB

O índice WB proposto por Zhao, Xu e Fränti (2009) é definido na Eq. (6.85). Este índice deve ser minimizado.

$$WB = K \times \frac{\sum_{k=1}^{K} \sum_{x_j \in C_k} dist(x_j, c_k)}{\sum_{k=1}^{K} |C_k| dist(\overline{x}, c_k)} \tag{6.85}$$

$$WB = K \times \frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \times d(e_j, e)}{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \times d(g, e)} \tag{6.86}$$

$$WB = K \times \frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \sum_{l=1}^{p} \times d_l(e_j, e)}{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \sum_{l=1}^{p} \times d_l(g, e)} \tag{6.87}$$

$$WB = K \times \frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{kl} \times d_l(e_j, e)}{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{kl} \times d_l(g, e)} \tag{6.88}$$

$$WB = K \times \frac{\sum_{k=1}^{K} \sum_{e_j \in C_k} \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{j} \times d_l(e_j, e)}{\sum_{k=1}^{K} |C_k| \sum_{e \in G_k} \sum_{l=1}^{p} \lambda_{j} \times d_l(g, e)} \tag{6.89}$$

6.1.2 Agrupamento nebuloso

Os índices para agrupamento nebuloso que foram estudados e utilizados nesta tese são sumarizados na Tabela 3.

Tabela 3 – Sumário de índices internos fuzzy

Notação	Critério	Objetivo	Fórmula
AWCD	Distância intra-cluster média	Minimização	6.93
HM	Homogeneidade	Minimização	6.93
FCS	Silhueta nebulosa	Maximização	6.95
FSS	Silhueta nebulosa simplificada	Maximização	6.96
FS	Fukuyama & Sugeno	Minimização	6.109
PC	Coeficiente de partição	Maximização	6.107
PE	Entropia da partição	Maximização	6.108
XB	Xie-Beni	Maximização	6.97

6.1.2.1 Homogeneidade

Este índice é a versão nebulosa da homogeneidade definida na seção anterior. Dessa forma, as informações contidas na partição nebulosoa U são consideradas. Assim como sua versão para agrupamento rígido, este índice quantifica a homogeneidade da partição e deve ser minimizado. O parametro $m \in (1, \infty)$ controla o grau de nebulosidade de cada objeto e_i.

$$HM = \sum_{k=1}^{K} \sum_{l=1}^{n} (u_{lk})^m dist(x_l, c_k) \tag{6.90}$$

$$HM = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m D(e_i, G_k) = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{e_l \in G_k} d(e_i, e_l) \tag{6.91}$$

$$HM = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m D(e_i, G_k) = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \sum_{e_l \in G_k} d_j(e_i, e_l) \tag{6.92}$$

$$HM = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D_{(\lambda_k,s)}(e_i, G_k) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_{kj}\sum_{e_l\in G_k} d_j(e_i, e_l) \tag{6.93}$$

$$HM = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D_{(\lambda,s)}(e_i, G_k) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_{j}\sum_{e_l\in G_k} d_j(e_i, e_l) \tag{6.94}$$

6.1.2.2 Silhueta nebulosa

A versão nebulosa da silhueta foi proposta por Campello e Hruschka (2006), pois a silhueta tradicional não é capaz de discriminar grupos sobrepostos uma vez que não faz uso da informação contida na partição nebulosa. A silhueta nebulosa é definida na Eq. 6.95, em que $s(e_j)$ é a silhueta do objeto e_j definida anteriormente de acordo com a Eq. (6.12), u_{pj} e u_{qj} são o maior e o segundo maior graus de pertinência do j-esimo objeto, respectivamente, e $\alpha \geq 0$ é um coeficiente de peso.

$$FSil = \frac{\sum_{i=1}^{n}(u_{pj} - u_{qj})^\alpha s(e_j)}{\sum_{i=1}^{n}(u_{pj} - u_{qj})^\alpha} \tag{6.95}$$

6.1.2.3 Silhueta nebulosa simplificada

Assim como a versão rígida, a silhueta nebulosa também possui uma versão simplificada. Dessa forma, a silhueta simplificada nebulosa faz uso da silhueta simplificada definida anteriormente. A silhueta nebulosa simplificada é definida na Eq. (6.96).

$$FSS = \frac{\sum_{i=1}^{n}(u_{pj} - u_{qj})^\alpha ss(e_j)}{\sum_{i=1}^{n}(u_{pj} - u_{qj})^\alpha} \tag{6.96}$$

6.1.2.4 Xie-Beni

O índice proposto por Xie e Beni (1991) é um índice de validação de agrupamento nebuloso que quantifica a média geral de compacidade e separação de uma partição nebuloso. A função objetivo proposta no clássico *Fuzzy C-Means* busca por *clusters* compactos. A mínima distância entre prototipos no denominador é chamada de separação. A razão entre a compacidade e a separação da partição nebulosa define o índice Xie-Beni. Este índice é definido na Eq. (6.97). Portanto, quando o número de *clusters* é fixado, quanto menor o valor do índice, melhor a partição gerada.

$$XB = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m dist(x_l, c_k)}{\min_{l\neq m} dist(c_m, c_l)} \tag{6.97}$$

$$XB = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}(u_{ik})^m \sum_{e_l\in G_k} d(e_i, e_l)}{n \min_{l\neq m} D(G_m, G_l)} \tag{6.98}$$

$$XB = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}\sum_{e_l\in G_k} d_j(e_i,e_l)}{n\min_{l\neq m} D(G_m,G_l)} \tag{6.99}$$

$$XB = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}\sum_{e_l\in G_k} \lambda_{kj} d_j(e_i,e_l)}{n\min_{l\neq m} D(G_m,G_l)} \tag{6.100}$$

$$XB = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}\sum_{e_l\in G_k} \lambda_{j} d_j(e_i,e_l)}{n\min_{l\neq m} D(G_m,G_l)} \tag{6.101}$$

6.1.2.5 Distância Intra-Cluster Média

A distância intra-cluster média (do inglês, *Average Within-Cluster Distance*) é definida de acordo com a Eq. (6.102). Este índice é o valor médio das distâncias intra-cluster computado para todos os clusters (HRUSCHKA; CAMPELLO; CASTRO, 2006).

Para dados relacionais, o índice passa a ser definido como definido em (6.103), (6.104), (6.105) e (6.106) para os casos em que os dados possuem apenas uma visão, quando os dados possuem várias visões com mesmo peso, quando possuem várias visões com pesos locais e com várias visões com pesos globais, respectivamente.

$$AWCD = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m dist(x_l,c_k)}{\sum_{l=1}^{n}(u_{lk})^m} \tag{6.102}$$

$$AWCD = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m \sum_{e\in G_k} d(e_l,e)}{\sum_{l=1}^{n}(u_{lk})^m} \tag{6.103}$$

$$AWCD = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_k} d_j(e_l,e)}{\sum_{l=1}^{n}(u_{lk})^m} \tag{6.104}$$

$$AWCD = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_k} \lambda_{kj} d_j(e_l,e)}{\sum_{l=1}^{n}(u_{lk})^m} \tag{6.105}$$

$$AWCD = \frac{1}{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_k} \lambda_{j} d_j(e_l,e)}{\sum_{l=1}^{n}(u_{lk})^m} \tag{6.106}$$

6.1.2.6 Coeficiente de partição

O coefiente de partição (do inglês, *Partition Coefficient - PC*) proposto por Bezdek (1974) é o primeiro índice de validação associado ao algoritmo nebuloso c-*means*. Este índice é definido na Eq. (6.107) e também depende apenas da partição nebuloso. O índice PC de um agrupamento tem valor dentro do intervalo $[\frac{1}{K}, 1]$, onde K é o número de *clusters*. Quanto mais próximo de 1, mais rígido o agrupamento se torna. No caso de todos os graus de pertinência de uma partição nebulosa serem iguais, isto é, $u_{ik} = 1$, o índice PC obtém seu valor mais baixo. Dessa forma, valores próximos a $1/K$ indicam que não há tendência de agrupamento no conjunto de dados considerado ou que o algoritmo de agrupamento

não conseguiu encontrá-lo (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001). Este índice deve ser maximizado.

$$PC = \frac{1}{n}\sum_{i=1}^{K}\sum_{l=1}^{n} u_{il}^2 \tag{6.107}$$

6.1.2.7 Entropia da partição

A entropia de partição (do inglês, *Partition Entropy - PE*) proposta por Bezdek (1981) é outro índice de validação para agrupamento nebuloso. Este índice é definido na Eq. (6.108) e também depende apenas da partição nebulosa. O índice PE tem valor dentro do intervalo $[0, log(K)]$. Quanto mais próximo o valor de PE estiver de 0, mais rígido o agrupamento se torna. Como no caso do coeficiente da partição, valores deste índice próximos a seu limite superior indicam a ausência de estrutura de agrupamento ou que o algoritmo não conseguiu revelá-la. Tanto o índice PC quanto o índice PE possuem a desvantagem de não usar informações dos dados propriamente e, por isso, estes índicem não possuem conexão com a geometria dos dados em suas definições (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

$$PE = -\frac{1}{n}\sum_{i=1}^{K}\sum_{l=1}^{n} u_{il}\log(u_{il}) \tag{6.108}$$

6.1.2.8 Fukuyama and Sugeno

Fukuyama e Sugeno (1989) propuseram outro índice em que *clusters* bem compactos e separados são procurados. Agrupamentos com *clusters* compactos e bem separados produzem valores pequenos para este índice. Portanto, este índice deve ser minimizado e é definido na Eq. (6.109). O primeiro termo dentro do colchetes mede a compacidade e o segundo termo mede a distância dos prototipos. $\overline{x}$ representa o vetor médio considerando todos os objetos da matriz de dados X (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

$$FS = \sum_{k=1}^{K}\sum_{l=1}^{n}(u_{lk})^m[dist(x_l, c_k) - dist(c_k, \overline{x})] \tag{6.109}$$

Para dados relacionais, o índice passa a ser definido como definido em (6.110), (6.111), (6.112) e (6.113) para os casos em que os dados possuem apenas uma visão, quando os dados possuem várias visões com mesmo peso, quando possuem várias visões com pesos locais e com várias visões com pesos globais, respectivamente.

$$FS = \sum_{k=1}^{K}\sum_{l=1}^{n}(u_{lk})^m \sum_{e\in G_i}[d(e_l, e) - d(e, \overline{g})] \tag{6.110}$$

$$FS = \sum_{k=1}^{K}\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_i}[d_j(e_l, e) - d_j(e, \overline{g})] \tag{6.111}$$

$$FS = \sum_{k=1}^{K}\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_i}\lambda_{kj}[d_j(e_l,e) - d_j(e,\overline{g})] \tag{6.112}$$

$$FS = \sum_{k=1}^{K}\sum_{l=1}^{n}(u_{lk})^m \sum_{j=1}^{p}\sum_{e\in G_i}\lambda_{j}[d_j(e_l,e) - d_j(e,\overline{g})] \tag{6.113}$$

O representativo geral do conjunto de dados $\overline{g}$ é definido em (6.114) quando apenas uma matriz é considerada e definido em (6.115) quando várias matrizes são consideradas.

$$\overline{g} = \{e_t \in E : \sum_{e\in E} d_j(e_t,e) = \min_{e_h\in E}\sum_{e\in E} d(e_h,e)\} \tag{6.114}$$

$$\overline{g} = \{e_t \in E : \sum_{j=1}^{p}\sum_{e\in E} d_j(e_t,e) = \min_{e_h\in E}\sum_{j=1}^{p}\sum_{e\in E} d_j(e_h,e)\} \tag{6.115}$$

6.2 CRITÉRIOS EXTERNOS

A qualidade das partições fornecidas pelos algoritmos podem ser comparadas usando índices externos quando há informação sobre as partições *a priori*. Dessa forma, é possível avaliar o desempenho dos algoritmos de agrupamento. O índice de rand ajustado (do inglês, *adjusted rand index* - ARI) (HUBERT; ARABIE, 1985) e a medida F (do inglês, *F-measure*) (RIJSBERGEN, 1979) são dois índices externos que são utilizados em diversos trabalhos para avaliar a qualidade das partições geradas. Como diversos trabalhos relacionados encontrados na literatura utilizaram esses índices externos para avaliação das partições geradas pelos algoritmos de agrupamento, optou-se por também utilizá-los para verificar a qualidade dos agrupamentos obtidos. Esses índices externos são definidos a seguir.

6.2.1 Índice de Rand Ajustado

Seja $P = \{P_1, ..., P_m\}$ a partição a priori de E em m classes e $Q = \{Q_1, ..., Q_K\}$ a partição rígido em K *clusters* dado por um algoritmo de clusterização. O índice de rand ajustado é:

$$ARI = \frac{\sum_{i=1}^{m}\sum_{j=1}^{K}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{K}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{m}\binom{n_i}{2} + \sum_{j=1}^{K}\binom{n_j}{2}] - \binom{n}{2}^{-1}\sum_{i-1}^{m}\binom{n_i}{2}\sum_{j=1}^{K}\binom{n_j}{2}} \tag{6.116}$$

em que $\binom{n}{2} = \frac{n(n-1)}{2}$ e n_{ij} representa o número de objetos que estão na classe P_i e no grupo Q_j; $n_{i.}$ indica o número de objetos na classe P_i; $n_{.j}$ indica o número de objetos no grupo Q_j; e n é o número total de objetos na base de dados.

O ARI avalia o grau de conformidade (similaridade) entre uma partição a priori e uma partição fornecida por um algoritmo de clusterização. Além disso, o ARI não é sensível ao número de classes nas partições ou a distribuição dos itens nos *clusters*. O ARI tem valor dentro do intervalo [-1,1], em que 1 indica concordância perfeita entre as partições comparadas.

6.2.2 Medida *F*

Dado um conjunto de objetos, a matriz de confusão exibe o número de classificações preditas pelo algoritmo de agrupamento contra as classificações reais de cada objeto. Dessa forma, é possível determinar índices de qualidade como precisão, cobertura e medida F (*F-measure*). Contudo, a matriz de confusão só pode ser utilizada quando existe a informação da classe dos objetos a serem classificados.

Tabela 4 – Matriz de confusão

	Clusters					
Classes	Q_1	$\cdots$	Q_j	$\cdots$	Q_K	$\sum$
P_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1K}	$n_{1\cdot} = \sum_{j=1}^{K} n_{1j}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
P_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iK}	$n_{i\cdot} = \sum_{j=1}^{K} n_{ij}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
P_m	n_{m1}	$\cdots$	n_{mj}	$\cdots$	n_{mK}	$n_{m\cdot} = \sum_{j=1}^{K} n_{mj}$
$\sum$	$n_{\cdot 1} = \sum_{i=1}^{m} n_{i1}$	$\cdots$	$n_{\cdot 2} = \sum_{i=1}^{m} n_{i2}$	$\cdots$	$n_{\cdot 3} = \sum_{i=1}^{m} n_{i3}$	$n = \sum_{i=1}^{m} \sum_{j=1}^{K} n_{ij}$

A precisão entre a classe $P_i (i = 1, ..., m)$ e o grupo $Q_i (i = 1, ..., K)$ é definida como a razão entre o número de objetos que estão simultaneamente na classe P_i e no grupo Q_j pelo número de objetos no grupo Q_j:

$$Precisão(P_i, Q_j) = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}} \tag{6.117}$$

A revocação (*Recall*) entre a classe $P_i (i = 1, ..., m)$ e o grupo $Q_i (i = 1, ..., K)$ é definido como a razão entre o número de objetos que estão simultaneamente na classe P_i e no grupo Q_j pelo número de objetos no grupo P_i:

$$Revocação(P_i, Q_j) = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{ij}}{\sum_{i=1}^{K} n_{ij}} \tag{6.118}$$

O índice *F-measure* entre uma classe $P_i (i = 1, ..., m)$ e um grupo $Q_i (i = 1, ..., K)$ é a média harmonica entre a precisão e *recall*:

$$F - measure(P_i, Q_j) = 2 \times \frac{Precisão(P_i, Q_j) \times Revocação(P_i, Q_j)}{Precisão(P_i, Q_j) + Revocação(P_i, Q_j)} \tag{6.119}$$

O índice *F-measure* entre uma partição a priori $P = \{P_1, .., P_m\}$ e a partição rígida $Q = \{Q_1, ..., Q_K\}$ fornecida por um algoritmo de agrupamento é definida como:

$$F - measure(P, Q) = \frac{1}{n} \sum_{i=1}^{m} n_{i\cdot} \max_{i \leq j \leq K} F - measure(P_i, Q_j) \tag{6.120}$$

O índice F-measure tem valor entre o intervalo $[0, 1]$, em que 1 indica perfeita concordância entre partições.

6.2.3 Taxa Global de Erro de Classificação

A taxa global de erro de classificação (do inglês, *Overall Error Rater of Classification - OERC*) tem como objetivo medir a habilidade de um algoritmo de agrupamento de encontrar classes a priori presentes em uma base de dados. Este índice é computado de acordo com a Eq. (6.121). O índice OERC tem valor entre o intervalo [0,1], em que 0 indica perfeita concordância entre partições.

$$OERC = 1 - \frac{\sum_{j=1}^{K} \max\limits_{1 \leq i \leq m} n_{ij}}{n} \tag{6.121}$$

6.2.4 Exemplo de agrupamento da base Íris

Para ilustrar o uso dos índices externos mencionados anteriormente, a base de dados Íris[1] será utilizada. Esta base contém 150 elementos e 3 classes. A Figura 4.1 mostra as partições $P = \{P_1, P_2, P_3\}$ e $Q = \{Q_1, Q_2, Q_3\}$, em que a partição P representa a partição a priori da base de dados Íris e a partição Q é fornecida como resultado de uma execução do algoritmo MRDCA. Para avaliar a qualidade da partição gerada, usam-se os índices externos.

Figura 5 – Exemplo de agrupamento de dados para base Íris

[1] https://archive.ics.uci.edu/ml/datasets/iris

A matriz de confusão mostrada na Tabela 5 fornece a quantidade de elementos que estão simultaneamente em uma classe e um grupo. Neste exemplo, a classe P_1 e o grupo Q_2 possuem os mesmos 50 elementos. Como mostrado a seguir, na Tabela 6, o grau de correspondencia de acordo com o índice *F-Measure* é igual a 1, ou seja, perfeita correspondencia.

Tabela 5 – Matriz de confusão do exemplo

	Grupos			
Classes	Q_1	Q_2	Q_3	$\sum$
P_1	0	50	0	50
P_2	46	0	4	50
P_3	14	0	36	50
$\sum$	60	50	40	150

Tabela 6 – Precisão, Revocação e F-measure entre as classes e grupos

	Precisão			Revocação			F-Measure		
	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
P_1	0	1	0	0	1	0	0	1	0
P_2	0,76	0	0,10	0,92	0	0,08	0,85	0	0,09
P_3	0,23	0	0,90	0,28	0	0,72	0,25	0	0,80

Após o cálculo da precisão, revocação e *F-measure* entre as classes e grupos, é possível calcular o índice *F-measure* entre as partições P e Q. Dessa forma, para o agrupamento fornecido no exemplo, o valor do índice *F-Measure* é igual a 0,88.

$F - measure(P,Q) = \frac{1}{n}\sum_{i=1}^{m} n_{i\cdot} \max_{i\leq j\leq K} F - measure(P_i, Q_j) = \frac{1}{150} \times (50 \times 1 + 50 \times 0,85 + 50 \times 0,8) = 0,88$.

O ARI para esse agrupamento da base de dados Íris é:

$ARI = \frac{\binom{n_{12}}{2}+\binom{n_{21}}{2}+\binom{n_{23}}{2}+\binom{n_{31}}{2}+\binom{n_{33}}{2}-\binom{n}{2}^{-1}[\binom{n_{1\cdot}}{2}+\binom{n_{2\cdot}}{2}+\binom{n_{3\cdot}}{2}][\binom{n_{\cdot 1}}{2}+\binom{n_{\cdot 2}}{2}+\binom{n_{\cdot 3}}{2}]}{\frac{1}{2}[\binom{n_{1\cdot}}{2}+\binom{n_{2\cdot}}{2}+\binom{n_{3\cdot}}{2}+\binom{n_{\cdot 1}}{2}+\binom{n_{\cdot 2}}{2}+\binom{n_{\cdot 3}}{2}]-\binom{n}{2}^{-1}[\binom{n_{1\cdot}}{2}+\binom{n_{2\cdot}}{2}+\binom{n_{3\cdot}}{2}][\binom{n_{\cdot 1}}{2}+\binom{n_{\cdot 2}}{2}+\binom{n_{\cdot 3}}{2}]} = 0,702.$

Por fim, a taxa global de erro é:

$$OERC = 1 - \frac{\sum_{j=1}^{K} \max\limits_{1\leq i\leq m} n_{ij}}{n} = 1 - \frac{50+46+36}{150} = 1 - 0,88 = 0,12.$$

6.3 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados os índices internos para validação de agrupamentos que serão usados como funções de aptidão para os métodos desenvolvidos. Alguns desses índices foram adaptados para considerar várias matrizes de dissimilaridades simultaneamente bem como considerar pesos de relevância para cada matriz. Além dos índices internos, também foram apresentados os índices de validação de agrupamentos externos que serão utilizados nesta tese para avaliar os agrupamentos encontrados pelos algoritmos. Estes agrupamentos encontrados serão avaliados em relação aos agrupamentos *a priori* conhecidos das bases de dados.

7 RESULTADOS PARA AGRUPAMENTO RÍGIDO

Neste capítulo serão apresentados os resultados encontrados pelos métodos híbridos desenvolvidos. O primeiro estudo realizado nesta tese é apresentado na subseção 7.1. Nesse primeiro estudo, o desempenho do modelo rígido desenvolvido para dados relacionais com única visão é avaliado juntamente com o desempenho de outros métodos presentes na literatura. O segundo estudo presente neste capítulo é apresentado na seção 7.2. As bases de dados com múltiplas visões usadas nos experimentos do segundo estudo são descritas na seção 7.2.1. Um estudo comparativo entre as funções de aptidão usadas é realizado e apresentado na subseção 7.2.3. Outro estudo comparativo entre os métodos desenvolvidos para dados relacionais com várias visões e outros métodos da literatura é realizado e apresentado na subseção 7.3. Por fim, uma discussão sobre os resultados encontrados é feita.

7.1 ESTUDO 1 - AGRUPAMENTO DE DADOS RELACIONAIS COM ÚNICA VISÃO

Este estudo teve como objetivo avaliar o desempenho do método híbrido para dados relacionais com única visão, chamado $PSO - SV$, em comparação com outros algoritmos para agrupamento de dados relacionais considerando-se a representação dos dados através de uma única matriz de dissimilaridades. Os métodos comparados foram: RHCM (HATHAWAY; DAVENPORT; BEZDEK, 1989), SSC (ZELNIK-MANOR; PERONA, 2004) e HCMdd (KRISHNAPURAM; FREG, 1992).

Neste estudo, dez conjuntos de dados obtidos do Repositório de Aprendizado de Máquina UCI [1] foram considerados. Os conjuntos utilizados foram: Abalone, Australian, Bupa, Image, Iris plants, Mammographic mass, Pima indians Diabetes, Satellite, Thyroid gland e Wine. A Tabela 10 sumariza as bases usadas com informações como número de atributos, número de objetos e número de classes. Primeiramente, os atributos foram normalizados de acordo com a abordagem de normalização min-max (7.1).

$$v' = \frac{v - min_a}{max_a - min_a}(nmax_a - nmin_a) + nmin_a \tag{7.1}$$

Por exemplo, uma variável Y tem seus valores variando de 5 a 100 e se quer normalizá-los entre $[0, 1]$, então o valor 25 seria mapeado como $\frac{25-5}{100-5}(1-0)+0 = 0.21$. Depois, a matriz de dissimilaridades foi computada usando a distância euclidiana entre os objetos de cada base considerando as variáveis normalizadas.

Todos os algoritmos foram executados 30 vezes. O número máximo de iterações do PSO-SV foi fixado em 50 para cada execução. Os parâmetros usados para execução do PSO-SV foram: $w_{min} = 0.4$, $w_{max} = 0.9$, $c_1 = c_2 = 2$. O tamanho do enxame utilizado foi

[1] https://archive.ics.uci.edu/ml/index.php

Tabela 7 – Sumário das bases de dados

Base	Nº de variáveis	Nº de classes	Nº de objetos
Abalone	8	3	4177
Australian	14	2	690
Bupa	7	2	345
Image	16	7	2310
Iris	4	3	150
Mammography	5	2	961
Pima	8	2	768
Satellite	36	7	6435
Thyroid	5	3	215
Wine	13	3	178

20 partículas. Depois de todas as execuções, a solução com melhor resultado de acordo com a função objetivo foi selecionada. A Tabela 8 apresenta o desempenho dos algoritmos RHCM, SSC, HCMdd e PSO-SV de acordo com os índices: medida F, ARI e OERC. O teste de Friedman foi utilizado para avaliar o desempenho dos algoritmos para as bases usadas. Para tanto, o *rank* médio de cada algoritmo foi calculado.

A Tabela 9 apresenta o *rank* médio dos algoritmos para cada índice. (DEMsAR, 2006) afirma que o *rank* médio por si só já fornece uma comparação justa entre os métodos. Considerando-se apenas os *ranks* médios, para todos os índices, o PSO-SV ficou em primeiro lugar, o HMCdd ficou em segundo, o RHCM ficou em terceiro e o SSC em quarto.

O teste de Friedman (DEMsAR, 2006) rejeitou a hipótese nula que considera que todos os algoritmos possuem desempenho equivalente para os índices. A aplicação do pós-teste Bonferroni-Dunn (DEMsAR, 2006) mostrou que o PSO-SV foi significativamente melhor do que os algoritmos RHCM e SSC (com $\alpha = 0.05$ e $\alpha = 0.10$) para o índice ARI. Considerando os índices F-Measure e OERC, o PSO-SV mostrou-se melhor de forma signifiativa do que o SSC (com $\alpha = 0.05$ e $\alpha = 0.10$). Com base nos *ranks* médios, o PSO-SV foi melhor do que HCMdd, mas não foi estatisticamente significativo. Com base nesse primeiro estudo, a abordagem híbrida para agrupamento de dados relacionais com apenas uma visão mostrou-se promissora considerando-se as bases utilizadas.

Tabela 8 – Desempenho dos algoritmos

	F-measure			
Data sets	HCMdd	RHCM	SSC	PSO-R
Abalone	0.533 (2)	0.514 (3)	0.394 (4)	0.539 (1)
Austral.	0.794 (1,5)	0.578 (4)	0.669 (3)	0.794 (1,5)
Bupa	0.530 (4)	0.652 (1)	0.632 (2)	0.547 (3)
Image	0.679 (2)	0.514 (3)	0.355 (4)	0.697 (1)
Iris	0.893 (2)	0.637 (3)	0.537 (4)	0.959 (1)
Mamm.	0.790 (1,5)	0.652 (4)	0.668 (3)	0.790 (1,5)
Pima	0.659 (3)	0.678 (1)	0.515 (4)	0.660 (2)
Satel.	0.670 (2)	0.660 (3)	0.620 (4)	0.672 (1)
Thyroid	0.467 (4)	0.776 (1)	0.509 (3)	0.618 (2)
Wine	0.925 (2)	0.648 (3)	0.604 (4)	0.954 (1)
	ARI			
Data sets	HCMdd	RHCM	SSC	PSO-R
Abalone	0.144 (2)	0.102 (3)	0.000 (4)	0.155 (1)
Austral.	0.345 (1,5)	0.000 (3.5)	0.000 (3.5)	0.345 (1,5)
Bupa	0.000 (4)	0.002 (3)	0.046 (1)	0.005 (2)
Image	0.494 (2)	0.332 (3)	0.113 (4)	0.511 (1)
Iris	0.727 (2)	0.216 (3)	0.131 (4)	0.885 (1)
Mamm.	0.337 (1,5)	0.036 (3)	0.000 (4)	0.337 (1,5)
Pima	0.091 (1,5)	0.045 (3)	-0.003 (4)	0.091 (1,5)
Satel.	0.468 (2)	0.423 (3)	0.403 (4)	0.493 (1)
Thyroid	0.060 (4)	0.523 (1)	-0.003 (3)	0.143 (2)
Wine	0.784 (2)	0.295 (3)	0.134 (4)	0.863 (1)
	OERC			
Data sets	HCMdd	RHCM	SSC	PSO-R
Abalone	48.24% (2)	52.16% (3)	63.41% (4)	47.68% (1)
Austral.	20.57% (1,5)	44.49% (3,5)	44.49% (3,5)	20.57% (1,5)
Bupa	42.02% (3)	42.02% (3)	37.97% (1)	42.02% (3)
Image	32.20% (2)	48.44% (3)	65.23% (4)	30.21% (1)
Iris	10.66% (2)	43.33% (3)	50.00% (4)	4.00% (1)
Mamm.	20.91% (1,5)	39.95% (3)	46.20% (4)	20.91% (1,5)
Pima	34.63% (2)	33.59% (1)	34.89% (4)	34.76% (3)
Satel.	25.36% (2)	32.10% (3)	35.95% (4)	24.94% (1)
Thyroid	30.23% (3)	15.34% (1)	30.23% (3)	30.23% (3)
Wine	7.30% (2)	39.32% (3)	40.44% (4)	4.49% (1)

Tabela 9 – Ranks médios.

	Médias		
Algoritmos	F-measure	ARI	OERC
HCMdd	2.400	2.150	2.200
RHCM	2.600	2.850	2.800
SSC	3.500	3.650	3.600
PSO-SV	1.500	1.350	1.400

7.2 AGRUPAMENTO RÍGIDO DE DADOS RELACIONAIS COM MÚLTIPLAS VISÕES

Nesta seção, os resultados encontrados pelos métodos híbridos para agrupamento rígido de dados relacionais com múltiplas visões são apresentados.

7.2.1 Bases de dados

As bases de dados utilizadas são sumarizadas na Tabela 10. O número de visões, número de classes a priori, número de objetos e número de variáveis de cada base de dados são informados na Tabela. Além disso, cada base de dados é brevemente apresentada abaixo.

Tabela 10 – Sumário das bases de dados

Bases	Nº de visões	Nº de classes	Nº de objetos	Nº de variáveis
Animals-1	6	5	1500	10940
Animals-2	6	5	2000	10940
Corel-1	7	4	400	338
Corel-2	7	4	400	338
Flowers	4	17	1360	-
Image	2	7	2310	19
Internet	4	2	2359	1492
Multiple features	6	10	2000	649
Phoneme	3	5	2000	150
3-Sources	3	6	169	10259

7.2.1.1 Animals with Attributes

Esta base de dados[2] consiste de 30475 imagens categorizadas em 50 classes de animais com seis visões. Dois subconjuntos, Animals-1 e Animals-2, foram extraídos do conjunto de dados. O primeiro subconjunto contém 1500 animais (300 por classe) e o segundo contém 2000 animais (400 por classe) obtidos de forma aleatória do conjunto de dados inteiro. As

[2] https://cvml.ist.ac.at/AwA/

categorias de cada subconjunto são descritas na Tabela 11. Os atributos são divididos em seis visões e estão descritos na Tabela 12.

Tabela 11 – Categorias

Subconjunto	Categorias				
Animals-1	antelope	chihuahua	collie	polar bear	siamese cat
Animals-2	antelope	chihuahua	collie	polar bear	siamese cat

Tabela 12 – Visões

Visão	Descrição	Variáveis
CQ	Color Histogram features	2688
SURF	SURF features	2000
LSS	Local Self-Similarity features	2000
RGSIFT	ColorSIFT features	2001
PHOG	PyramidHOG (PHOG) features	252
SIFT	SIFT features	2000

7.2.1.2 Corel Images

Esta base de dados[3] foi extraída do banco de dados COREL. Cinco subconjuntos foram extraídos contendo quatro classes, cada uma possuindo 100 imagens. As categorias dos cinco subconjuntos estão descritas na Tabela 38. Os atributos são divididos em sete visões e descritos na Tabela 13.

Tabela 13 – Visões do conjunto Corel

Visão	Descrição	Variáveis
ColorHsvHistogram64	Color histogram	64
ColorLuvMoment123	Color moment	9
ColorHsvCoherence64	Color coherence	128
CoarsnessVector	Coarsness - tamura texture	10
Directionality	Directionality - tamura texture	8
WaveletTwtTexture	Wavelet texture	104
MRSAR	MASAR texture	15

[3] http://www.cs.virginia.edu/CB9Cxj3a/research/CBIR/Download.htm

Tabela 14 – Subconjuntos extraídos da base de dados Corel

Subconjunto	Categorias			
Corel-1	buses	leopards	trains	ships
Corel-2	buses	leopards	cars	deer

7.2.1.3 Flowers

Esta base de dados [4] possui 1370 imagens de flores divididas em 17 classes. A base é constituída de imagens que foram coletadas na Web e de imagens de flores que foram fotografadas. As flores contidas nessa base de dados são flores comumente encontradas no Reino Unido. Existem 80 imagens por classe e quatro matrizes de dissimilaridades são fornecidas.

Figura 6 – Exemplos de imagens de três classes da base Flowers

[4] http://www.robots.ox.ac.uk/ vgg/data/flowers/17/index.html

7.2.1.4 Image Segmentation

Esta base também foi obtida do Repositório da UCI e contém 2310 objetos divididos em sete classes, em que cada classe possui 330 objetos. Os objetos desta classe são descritos por 19 variáveis divididas em duas visões como apresentado na Tabela 15.

Tabela 15 – Image Segmentation Views

Visão	Descrição	Variáveis
shape	Informação de forma	9
rgb	Valores RGB	10

7.2.1.5 Internet Advertisement

Esta base de dados pode ser encontrada no repositório da UCI e representa um conjunto de possíveis propagandas em páginas da Internet. Esta base contém 2359 objetos e um total de 1558 variáveis, as quais podem ser divididas em seis visões descritas na Tabela 16.

Tabela 16 – Visões da base Internet

Visão	Descrição	Variáveis
geometry	altura, largura e razão de imagem	76
base url	palavras contidas na url	216
image url	palavras sobre imagem	64
target url	palavras sobre destinatario	240

7.2.1.6 Multiple Features

Esta base de dados foi obtida no repositório da UCI. Esta base contém 2000 objetos classificados em 10 classes (‘0’–‘9’), em que cada classe possui 200 objetos e cada objeto representa um numeral escrito a mão. Cada objeto da base é descrito por 649 variáveis que são divididas em seis diferentes visões apresentadas na Tabela 17.

Tabela 17 – Visões da base Multiple Features

Visão	Descrição	Variáveis
mfeat-fou	Coeficientes de Fourier	76
mfeat-fac	Correlações de perfil	216
mfeat-kar	Coeficientes de Karhunen-Love	64
mfeat-pix	Média de pixel em 2x3	240
mfeat-zer	Momentos Zernike	47
mfeat-mor	Características morfológicas	6

7.2.1.7 Phoneme

Esta base de dados[5] contém 2000 objetos contendo 5 classes de fonemas. Cada objeto é descrito como (x_i, y_i) $(i = 1..., n)$, em que y_i fornece a pertinência de classe (fonema) e $x_i = (x_i(t_1), ..., x_i(t_{150}))$ é o vetor de atributos. Como foi realizado em (CARVALHO; LECHEVALLIER; MELO, 2012), a partir dessa base, duas bases adicionais correspondendo a velocidade e aceleração foram obtidas. Dessa forma, três visões são utilizadas para representar a base de dados.

Tabela 18 – Visões da base Phoneme

Visão	Descrição	Variáveis
npfda-position	Posição	150
npfda-velocity	Velocidade	149
npfda-acceleration	Aceleração	148

7.2.1.8 Water Treatment Plant

Esta base de dados também foi obtida do repositório UCI e descreve medidas diárias de sensores de uma estação de tratamento de desperdício de água urbano. Esta base contém 527 instâncias e 38 variáveis. As variáveis podem ser divididas em quatro visões como descrito na Tabela 19.

Tabela 19 – Visões da base Water

Visão	Descrição	Variáveis
input	Condições de entrada	22
output	Demandas de saída	7
performance	Demandas de desempenho de entrada	5
global	Demandas de entrada de desempenho geral	4

7.2.1.9 3-Sources

Esta base de dados de texto set[6] foi coletada a partir de três fontes de notícias *online*: BBC, *Reuters* e *The Guardian*. Esta base contém 169 histórias que foram reportadas em todas as três fontes de notícias. Além disso, esta base tem seis classes: *business*, *entertainment*, *health*, *politics*, *sport* e *technology*.

[5] https://statweb.stanford.edu/ tibs/ElemStatLearn/
[6] http://mlg.ucd.ie/datasets/3sources.html

7.2.2 Configurações dos experimentos

Todos os métodos foram implementados neste trabalho, executados 50 vezes de forma independente e tiveram o número máximo de iterações fixado em 100. Os índices externos foram computados para cada partição final fornecida em cada execução. Os métodos foram executados com enxame de 50 partículas. Os outros parâmetros usados para execução foram: $w_{min} = 0.4$, $w_{max} = 0.9$, $c_1 = c_2 = 2$. O peso de inércia, $w(t)$, foi decrescido linearmente de 0.9 até 0.4. Os componentes de velocidade foram inicializados com valores aleatórios entre [0,1].

Para cada visão, uma matriz de dissimilaridade foi computada considerando todos os atributos da visão. Nesta tese, a dissimilaridade entre cada par de objetos foi computada de acordo com a distância Euclidiana (L_2). Para os algoritmos que consideram apenas uma visão dos dados, uma única matriz de dissimilaridade foi computada considerando todas as variáveis de todas as visões. Para todas as matrizes, a dissimilaridade entre quaisquer dois objetos $e_l, e_{l'}$ foi normalizada, conforme realizado em (CARVALHO; LECHEVALLIER; MELO, 2012), como $\frac{d(e_l,e_{l1})}{T}$ onde T, definido em (7.2), é a dispersão geral e g é o representante global.

$$T = \sum_{k=1}^{n} d(e_k, g). \tag{7.2}$$

O representante global é definido como $g = e_l \in E$, em que

$$l = argmin_{1 \leq h \leq n} \sum_{k=1}^{n} d(e_k, e_h). \tag{7.3}$$

7.2.2.1 Influência do parâmetro $|G_k|$

Com o objetivo de avaliar a influência do parâmetro $|G_k|$ no desempenho dos algoritmos, três bases de dados foram selecionadas. O parâmetro $|G_k|$ variou dentro do intervalo $[2, 10]$. Os métodos foram executados 30 vezes e 20 partículas foram usadas com número máximo de iterações definido em 50. Os outros parâmetros usados para execução foram: $w_{min} = 0.4$, $w_{max} = 0.9$, $c_1 = c_2 = 2$. As Figuras 7 - 9 apresentam os valores médios obtidos, em termos de medida F, OERC e índice da silhueta, para cada valor de $|G_k|$.

Observa-se que o parâmetro tem influência no desempenho, pois o número de objetos que é usado para representar cada grupo pode gerar representativos melhores ou piores, dependendo da base de dados considerada. Por exemplo, para a base 3-Sources ilustrada na Figura 7, percebe-se uma melhora na medida F e na taxa de erro até $|G_k| = 6$, após esse valor, os índices começam a piorar. O índice da silhueta, por outro lado, apresentou crescimento até o valor máximo testado do parâmetro. Este último resultado sugere que, nem sempre, a otimização do índice da silhueta, o qual busca grupos bem separados e coesos, irá necessariamente coincidir com os grupos a priori conhecidos de uma base de dados.

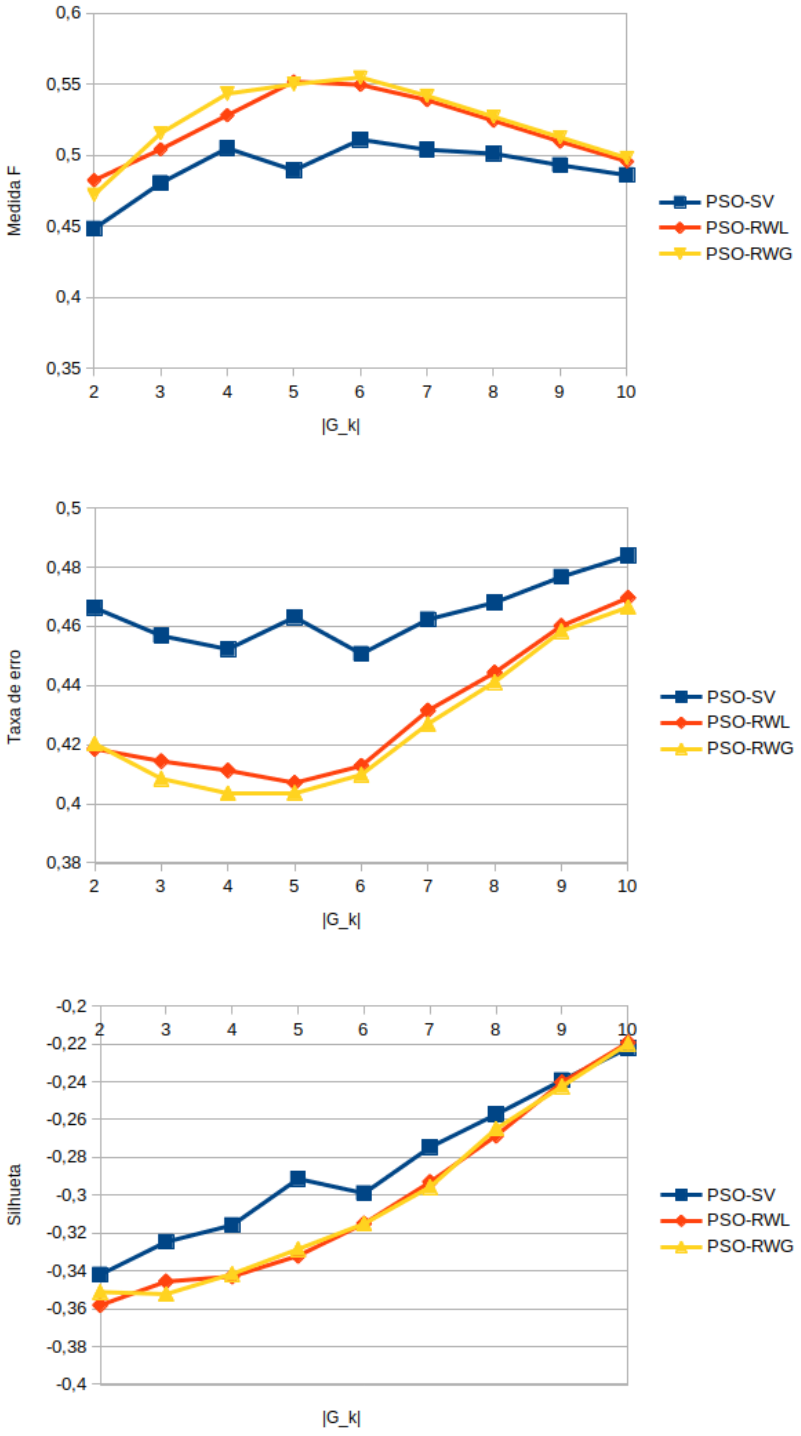


Figura 7 – Influência de $|G_k|$ sobre a base 3-Sources

Ao analisar a influência sobre o desempenho dos métodos para as bases *Image Segmentation* e *Multiple Features*, observa-se pouca variação para os três índices. Apenas o método PSO-SV apresentou maior variação nos valores do índice da silhueta para $|G_k| = \{8, 9, 10\}$

No entanto, é importante ressaltar que, considerando-se os índices externos, esse tipo de análise só é possível quando se tem a informação sobre as classes a priori e não pode ser usada em bases de dados para as quais não existe essa informação. Portanto, para os

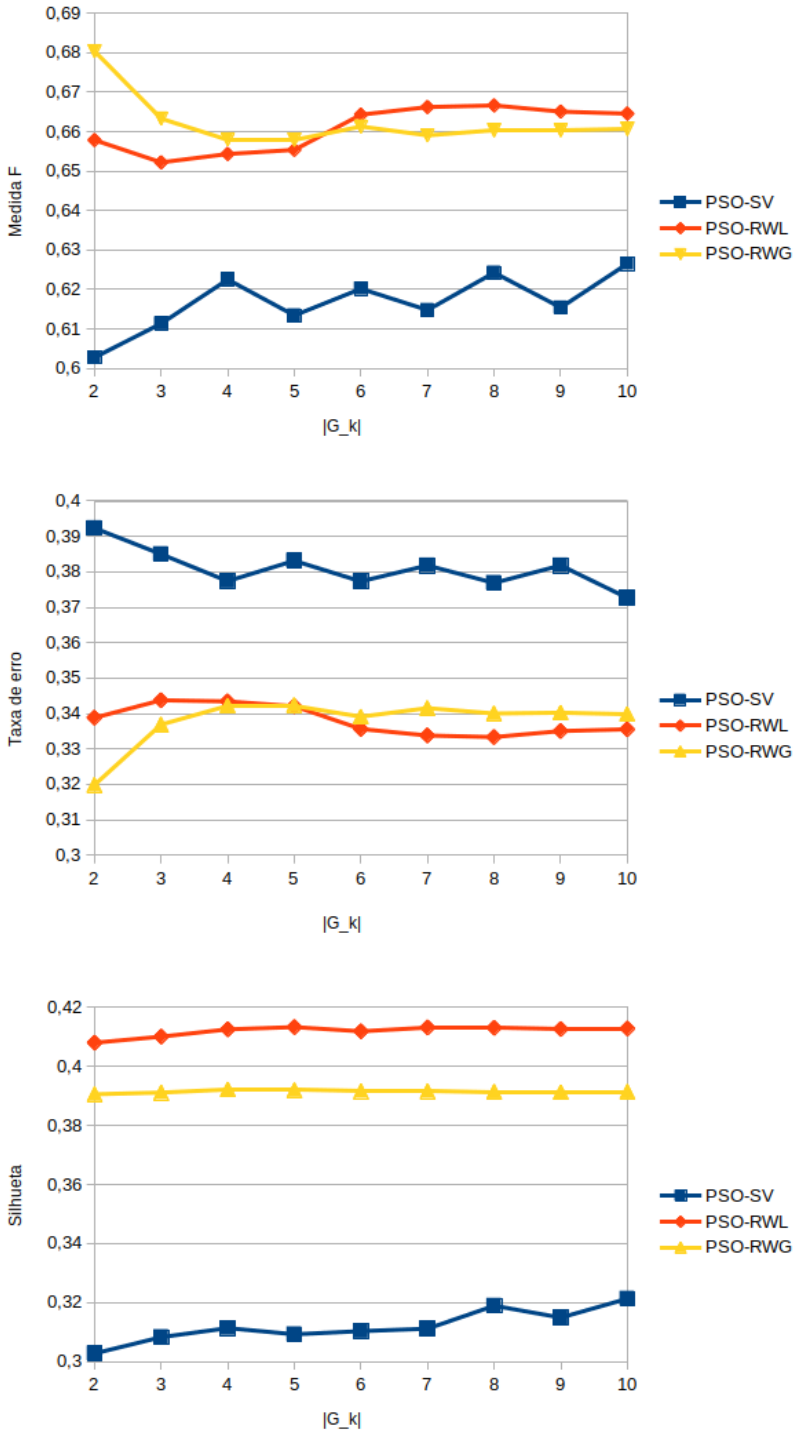


Figura 8 – Influência de $|G_k|$ sobre a base Image

próximos estudos empíricos, o número de *medóides* em cada representativo foi definido como $|G_k| = 5$ para todos os métodos híbridos desenvolvidos bem como para os outros métodos que utilizam múltiplos *medóides* para representar os grupos.

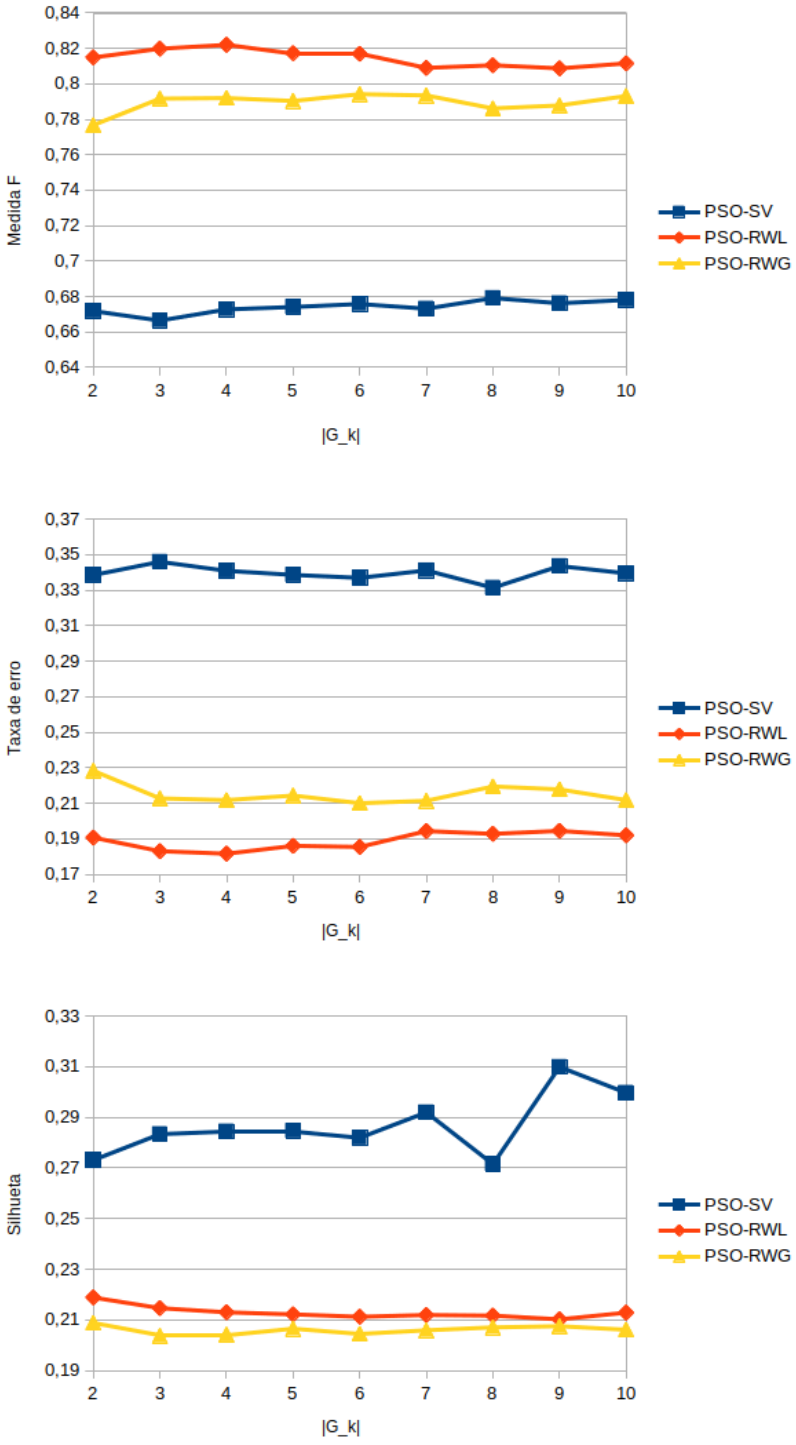


Figura 9 – Influência de $|G_k|$ sobre a base Multiple features

7.2.3 Estudo 2

Em experimentos preliminares, apenas a homogeneidade intra-cluster tinha sido considerada como função de aptidão. No entanto, devido a existência de outros índices para validação de agrupamentos na literatura, verificou-se a necessidade de investigar outros índices. Neste estudo, as funções de aptidão avaliadas foram: Calinsky-Harabasz, índice CS, Davies-Bouldin, Dunn, Hartigan, Homogeneidade, Erro de quantização, índice da Silhueta, Silhueta

simplificada, índice WB e o índice de Xu. Estas funções de aptidão foram avaliados com o objetivo de verificar se alguma se destacaria dentre as demais e seria capaz de gerar melhores agrupamentos dos conjuntos de dados.

Este estudo foi realizado com o objetivo de investigar e responder as seguintes perguntas de pesquisa:

1. *Considerando os métodos híbridos desenvolvidos, explorar múltiplas visões de dados relacionais produz melhores agrupamentos?*

2. *Dentre as funções de aptidão consideradas neste trabalho, é possível afirmar que alguma delas produz melhores agrupamentos?*

3. *O uso de pesos de relevância estimados localmente nos métodos desenvolvidos é capaz de gerar melhores agrupamentos do que o uso de pesos de relevância estimados globalmente?*

As Tabelas 20 - 30 apresentam os resultados dos métodos $PSO-SV$, $PSO-RWL$ e $PSO-RWG$ obtidos para cada função de aptidão separadamente para os índices medida F e ARI. A média e o desvio padrão são apresentados. As maiores médias encontradas para cada base de dados estão destacadas em negrito.

Nessas Tabelas, é possível observar qual dos três métodos obteve melhor desempenho médio e maior medida de dispersão para cada índice e para cada base de dados. Consequentemente, é possível verificar para quais bases de dados, a hipótese de que explorar bases de dados com várias visões produz melhores agrupamentos dos dados pode ser confirmada ou não. Além disso, também é possível comparar o método que utiliza pesos de relevância estimados localmente com o método que usa os pesos de relevância globais. Por fim, o grupo de funções também será analisado para cada base e cada índice.

Tabela 20 – Resultados relativos a função CH

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,01(.02)	**0,21(.00)**	0,41(.02)	0,28(.04)	0,09(.01)	0,40(.00)	0,31(.08)	0,48(.02)	0,60(.09)	0,01(.00)	0,32(.05)
	F-measure	0,34(.02)	**0,49(.01)**	0,66(.03)	0,56(.04)	0,28(.02)	0,58(.00)	0,83(.03)	0,67(.03)	0,76(.07)	0,21(.01)	0,54(.03)
PSO-RWL	ARI	**0.20(.06)**	0.18(.03)	**0.49(.03)**	**0.36(.06)**	0.28(.01)	**0.43(.03)**	**0.54(.02)**	0.53(.05)	0.59(.14)	**0.03(.00)**	0.33(.07)
	F-measure	**0.50(.04)**	**0.49(.02)**	**0.73(.02)**	**0.62(.06)**	**0.46(.02)**	**0.60(.02)**	**0.90(.00)**	**0.70(.04)**	0.75(.10)	**0.26(.01)**	0.55(.05)
PSO-RWG	ARI	0.04(.04)	0.00(.00)	**0.49(.04)**	0.33(.05)	**0.29(.01)**	0.42(.01)	0.45(.12)	0.46(.05)	**0.63(.08)**	**0.03(.00)**	**0.35(.05)**
	F-measure	0.35(.04)	0.32(.00)	0.72(.03)	0.60(.05)	**0.49(.01)**	0.59(.01)	0.87(.04)	0.62(.05)	**0.78(.06)**	0.24(.01)	**0.56(.03)**

Tabela 21 – Resultados relativos a função CS

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,00(.00)	0,10(0,11)	0,36(.02)	0,26(.02)	0,09(.01)	0,39(.05)	**0,50(.00)**	0,45(.04)	**0,56(0,11)**	0,01(.00)	0,29(.05)
	F-measure	0,33(.00)	0,41(.08)	0,62(.01)	0,53(.02)	0,28(.01)	0,58(.04)	**0,89(.00)**	0,62(.04)	**0,72(.08)**	0,21(.01)	0,53(.03)
PSO-RWL	ARI	**0.18(.06)**	**0.36(.15)**	**0.46(.01)**	**0.36(.07)**	0.27(.02)	0.37(.08)	0.10(.24)	**0.51(.04)**	0.53(.12)	**0.03(.00)**	0.16(.07)
	F-measure	**0.49(.06)**	**0.59(.10)**	**0.70(.01)**	**0.63(.06)**	0.47(.02)	0.56(.07)	0.82(.05)	**0.69(.04)**	0.70(.08)	**0.26(.01)**	0.46(.05)
PSO-RWG	ARI	0.00(.00)	0.01(.02)	0.45(.01)	0.31(.04)	**0.29(.02)**	**0.42(.02)**	0.08(.04)	0.48(.05)	0.53(.10)	**0.03(.00)**	0.18(.07)
	F-measure	0.33(.01)	0.34(.02)	0.69(.00)	0.59(.04)	**0.49(.02)**	**0.61(.02)**	0.83(.00)	0.66(.05)	0.71(.07)	0.25(.02)	0.47(.05)

Tabela 22 – Resultados relativos a função DB

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,01(.02)	**0,10(0,11)**	0,39(.05)	0,27(.05)	0,06(.01)	0,32(.04)	**0,50(.01)**	0,26(.05)	0,45(0,12)	0,02(.01)	0,29(.06)
	F-measure	0,34(.03)	0,41(.08)	0,67(.03)	0,54(.04)	0,21(.01)	0,54(.03)	**0,89(.00)**	0,43(.04)	0,65(.07)	0,34(.03)	0,53(.04)
PSO-RWL	ARI	**0.04(.04)**	0.08(.12)	**0.47(.04)**	**0.37(.07)**	0.27(.03)	**0.48(.01)**	0.22(.28)	**0.54(.06)**	0.29(.10)	**0.04(.02)**	**0.26(.09)**
	F-measure	**0.36(.04)**	0.37(.09)	**0.71(.03)**	**0.65(.06)**	**0.46(.04)**	**0.65(.02)**	0.85(.05)	**0.71(.05)**	0.54(.07)	**0.38(.05)**	0.50(.06)
PSO-RWG	ARI	0.01(.01)	0.01(.03)	0.45(.04)	0.33(.06)	**0.28(.03)**	0 **.48(.03)**	-0.02(.08)	0.52(.06)	**0.47(.08)**	0.03(.01)	**0.26(.09)**
	F-measure	0.33(.01)	0.33(.02)	0.70(.03)	0.61(.06)	**0.46(.03)**	**0.65(.02)**	0.80(.01)	0.68(.06)	**0.67(.04)**	0.29(.03)	**0.51(.05)**

Tabela 23 – Resultados relativos a função DN

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,00(.00)	0.16(0.09)	0,40(.02)	0,37(.03)	0.09(0.05)	0.40(0.03)	**0,50(.01)**	0.43(.04)	0,68(.06)	0,01(.00)	0,15(.05)
	F-measure	0,33(.00)	0.46(0.07)	0,66(.02)	0,64(.02)	0.27(0.09)	0.58(0.03)	**0,89(.00)**	0.58(.04)	0,82(.05)	0,22(.01)	0,44(.05)
PSO-RWL	ARI	**0.20(.02)**	**0.37(.14)**	**0.49(.08)**	**0.46(.04)**	0.29(.02)	**0.43(.05)**	0.19(.27)	0.56(.05)	0.67(.08)	**0.03(.00)**	0.22(.11)
	F-measure	**0.51(.03)**	**0.61(.10)**	**0.72(.05)**	**0.72(.03)**	0.47(.03)	**0.61(.04)**	0.84(.05)	0.71(.05)	**0.83(.05)**	**0.26(.02)**	0.49(.07)
PSO-RWG	ARI	0.00(.00)	0.05(.05)	0.37(.13)	**0.46(.05)**	**0.31(.02)**	0.42(.03)	0.09(.02)	0.57(.06)	0.64(.11)	**0.03(.00)**	**0.25(.11)**
	F-measure	0.33(.00)	0.37(.05)	0.64(.10)	0.71(.03)	**0.49(.02)**	0.60(.03)	0.83(.00)	0.72(.05)	0.81(.07)	**0.26(.02)**	**0.50(.07)**

Tabela 24 – Resultados relativos a função HA

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,01(.02)	**0,21(.01)**	0,41(.02)	0,30(.04)	0,09(.01)	0,40(.00)	0,32(.09)	0,49(.02)	0,60(.09)	0,01(.00)	0,30(.06)
	F-measure	0,34(.02)	**0,49(.00)**	0,65(.02)	0,57(.05)	0,28(.01)	0,58(.00)	0,83(.03)	0,67(.03)	0,76(.07)	0,21(.01)	0,53(.04)
PSO-RWL	ARI	**0.22(.05)**	0.19(.03)	**0.49(.03)**	**0.35(.05)**	0.28(.01)	**0.42(.02)**	**0.54(.01)**	**0.53(.05)**	0.58(.12)	**0.03(.00)**	0.33(.07)
	F-measure	**0.51(.04)**	**0.49(.02)**	**0.73(.02)**	**0.61(.06)**	0.47(.02)	**0.60(.02)**	**0.90(.00)**	**0.71(.04)**	0.75(.08)	**0.27(.01)**	0.55(.04)
PSO-RWG	ARI	0.04(.05)	0.00(.00)	**0.49(.04)**	0.34(.05)	**0.29(.01)**	**0.42(.01)**	0.46(.11)	0.46(.05)	**0.66(.07)**	0.03(.00)	**0.36(.04)**
	F-measure	0.35(.04)	0.32(.00)	**0.73(.03)**	**0.61(.06)**	**0.48(.01)**	0.59(.01)	0.88(.05)	0.62(.05)	**0.80(.05)**	0.24(.01)	**0.57(.02)**

Tabela 25 – Resultados relativos a função HM

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,06(.00)	0,02(.01)	0,40(.00)	0,27(.00)	0,10(.01)	0,40(.00)	0,50(.00)	0,50(.00)	0,71(.00)	0,01(.00)	0,34(.03)
	F-measure	0,37(.00)	0,33(.00)	0,64(.00)	0,54(.00)	0,29(.01)	0,58(.00)	0,89(.00)	0,68(.00)	0,85(.00)	0,21(.01)	0,56(.03)
PSO-RWL	ARI	**0.18(.02)**	**0.44(.02)**	**0.46(.00)**	**0.41(.00)**	0.27(.01)	**0.49(.00)**	**0.54(.00)**	**0.67(.00)**	**0.72(.00)**	**0.03(.00)**	0.32(.03)
	F-measure	**0.49(.01)**	**0.66(.01)**	**0.69(.00)**	**0.69(.00)**	0.46(.01)	**0.66(.00)**	**0.90(.00)**	**0.83(.00)**	**0.85(.00)**	0.26(.01)	0.55(.02)
PSO-RWG	ARI	0.01(.00)	0.01(.00)	0.44(.00)	0.39(.00)	**0.29(.01)**	**0.49(.00)**	0.50(.00)	0.62(.00)	**0.72(.00)**	**0.03(.00)**	**0.36(.02)**
	F-measure	0.33(.00)	0.32(.00)	**0.69(.00)**	0.67(.00)	**0.49(.01)**	**0.66(.00)**	0.89(.00)	0.80(.00)	**0.85(.00)**	**0.27(.01)**	**0.57(.02)**

Tabela 26 – Resultados relativos a função QE

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,05(.01)	0,02(.01)	0,38(.03)	0,26(.03)	0,09(.01)	0,47(.05)	**0,50(.00)**	0,50(.02)	**0,70(.05)**	0,01(.00)	0,31(.04)
	F-measure	0,37(.02)	0,37(.01)	0,64(.02)	0,54(.03)	0,27(.02)	0,63(.04)	**0,89(.00)**	0,70(.02)	**0,82(.04)**	**0,26(.02)**	**0,54(.03)**
PSO-RWL	ARI	**0.15(.07)**	**0.40(.13)**	**0.47(.01)**	**0.36(.04)**	0.26(.01)	0.47(.02)	0.19(.28)	0.55(.05)	0.67(.10)	**0.02(.00)**	**0.32(.03)**
	F-measure	**0.48(.05)**	**0.62(.09)**	**0.70(.01)**	**0.66(.04)**	0.44(.02)	0.65(.01)	0.84(.05)	0.72(.04)	0.80(.07)	0.25(.01)	**0.54(.02)**
PSO-RWG	ARI	0.01(.01)	0.01(.02)	**0.47(.01)**	**0.36(.04)**	**0.28(.02)**	**0.48(.02)**	-0.00(.13)	**0.57(.05)**	**0.70(.06)**	**0.02(.01)**	**0.32(.03)**
	F-measure	0.34(.01)	0.33(.02)	**0.70(.00)**	0.65(.05)	**0.47(.02)**	**0.66(.01)**	0.80(.02)	**0.74(.04)**	**0.82(.05)**	**0.26(.01)**	**0.54(.02)**

Tabela 27 – Resultados relativos a função SIL

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,05(.05)	0,21(.01)	0,41(.01)	0,33(.06)	0,09(.01)	0,46(.01)	**0,51(.01)**	0,37(.03)	0,57(0,10)	0,01(.00)	**0,13(.05)**
	F-measure	0,37(.04)	**0,49(.01)**	0,65(.00)	0,60(.06)	0,28(.01)	0,65(.01)	**0,89(.00)**	0,52(.03)	0,73(.07)	0,22(.01)	**0,42(.04)**
PSO-RWL	ARI	**0.13(.06)**	**0.24(.18)**	**0.53(.03)**	**0.45(.01)**	0.27(.01)	**0.50(.00)**	0.12(.16)	**0.66(.03)**	**0.72(.01)**	**0.03(.00)**	0.12(.05)
	F-measure	**0.42(.05)**	**0.49(.15)**	**0.76(.03)**	**0.70(.01)**	0.46(.01)	**0.67(.00)**	0.70(.08)	**0.82(.03)**	**0.85(.01)**	**0.25(.01)**	0.41(.04)
PSO-RWG	ARI	0.04(.02)	0.10(.06)	0.52(.02)	**0.45(.02)**	**0.30(.01)**	**0.50(.00)**	0.09(.01)	0.61(.04)	**0.72(.02)**	**0.03(.00)**	**0.13(.05)**
	F-measure	0.34(.02)	0.39(.05)	0.75(.03)	0.69(.01)	**0.49(.01)**	**0.67(.01)**	0.70(.01)	0.78(.03)	**0.85(.01)**	**0.25(.01)**	**0.42(.04)**

Ao analisar as Tabelas 20 - 30, é possível perceber que os métodos $PSO-RWL$ e $PSO-RWG$ obtiveram melhores resultados em comparação ao método $PSO-SV$ para a

Tabela 28 – Resultados relativos a função SS

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,00(.00)	0,17(.08)	**0,41(.01)**	0,28(.05)	0,09(.01)	0,45(.03)	0,50(.01)	0,48(.04)	**0,50(.05)**	0,01(.00)	0,32(.02)
	F-measure	0,33(.00)	0,00(.00)	0,65(.01)	0,56(.05)	0,27(.01)	0,62(.03)	0,89(.00)	0,65(.05)	**0,69(.03)**	0,23(.01)	0,55(.01)
PSO-RWL	ARI	**0.03(.07)**	0.29(.12)	**0.41(.10)**	**0.38(.06)**	0.30(.01)	**0.47(.00)**	0.50(.16)	**0.54(.07)**	0.51(.06)	**0.03(.00)**	0.33(.03)
	F-measure	**0.36(.06)**	**0.57(.08)**	**0.66(.07)**	**0.65(.06)**	0.49(.01)	**0.63(.00)**	0.89(.05)	**0.69(.07)**	0.68(.05)	0.26(.01)	0.55(.02)
PSO-RWG	ARI	0.00(.00)	0.00(.00)	0.40(.09)	0.34(.06)	**0.31(.01)**	0.43(.03)	**0.52(.01)**	**0.54(.06)**	0.49(.03)	**0.03(.00)**	**0.34(.03)**
	F-measure	0.33(.00)	0.33(.00)	**0.66(.07)**	0.62(.06)	**0.50(.02)**	0.61(.02)	**0.90(.00)**	**0.69(.06)**	0.67(.02)	**0.28(.01)**	**0.56(.02)**

Tabela 29 – Resultados relativos a função WB

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,01(.01)	**0,21(.02)**	0,37(.05)	0,27(.04)	0,06(.01)	0,40(.03)	0,50(.00)	0,46(.03)	0.44(0.14)	0,01(.00)	0,27(.06)
	F-measure	0,34(.02)	**0,49(.01)**	0,64(.04)	0,56(.04)	0,22(.02)	0,58(.02)	0,89(.00)	0,62(.04)	0.65(0.08)	0,21(.01)	0,52(.04)
PSO-RWL	ARI	**0.21(.08)**	0.19(.03)	**0.49(.03)**	**0.35(.05)**	0.28(.01)	**0.43(.03)**	**0.53(.05)**	**0.67(.00)**	0.62(.09)	**0.03(.00)**	0.33(.05)
	F-measure	**0.50(.06)**	**0.49(.02)**	**0.73(.02)**	**0.62(.06)**	0.47(.02)	**0.60(.02)**	**0.90(.01)**	**0.83(.00)**	0.77(.07)	**0.26(.01)**	0.55(.03)
PSO-RWG	ARI	0.03(.04)	0.00(.00)	0.48(.05)	0.34(.05)	**0.29(.01)**	0.42(.01)	0.47(.10)	0.45(.04)	**0.65(.06)**	**0.03(.00)**	**0.34(.07)**
	F-measure	0.35(.04)	0.32(.00)	0.72(.04)	0.60(.06)	**0.48(.02)**	0.59(.01)	0.88(.05)	0.61(.05)	**0.79(.05)**	0.24(.01)	**0.56(.05)**

Tabela 30 – Resultados relativos a função XU

Método	Índice	Base de dados										
		AwA1	AwA2	C1	C2	FLO	IMG	INT	MFT	PHO	WAT	3-S
PSO-SV	ARI	0,05(.00)	0,02(.00)	0,40(.00)	0,30(.04)	0,10(.01)	0,40(.00)	0,50(.00)	0,50(.01)	0.71(0.00)	0,01(.00)	0,31(.04)
	F-measure	0,37(.01)	0,33(.00)	0,64(.00)	0,57(.04)	0,29(.01)	0,58(.00)	0,89(.00)	0,68(.01)	0.85(0.00)	0,21(.01)	0,54(.03)
PSO-RWL	ARI	**0.19(.02)**	**0.44(.02)**	**0.46(.00)**	**0.41(.00)**	0.27(.01)	**0.49(.00)**	**0.54(.00)**	**0.67(.00)**	**0.72(.00)**	**0.03(.00)**	0.33(.03)
	F-measure	**0.49(.02)**	**0.67(.01)**	**0.69(.00)**	**0.69(.00)**	0.46(.01)	**0.66(.00)**	**0.90(.00)**	**0.83(.00)**	**0.85(.00)**	**0.27(.01)**	0.55(.02)
PSO-RWG	ARI	0.01(.00)	0.01(.00)	0.44(.00)	0.39(.00)	**0.29(.01)**	**0.49(.00)**	0.50(.00)	0.62(.00)	**0.72(.00)**	**0.03(.00)**	**0.35(.02)**
	F-measure	0.33(.00)	0.32(.00)	**0.69(.00)**	0.67(.00)	**0.49(.01)**	**0.66(.00)**	0.89(.00)	0.80(.00)	**0.85(.00)**	**0.27(.01)**	**0.57(.02)**

maioria dos casos. Esses resultados sugerem que, ao considerar múltiplas visões dos dados, torna-se possível agrupá-los com maior eficiência. Esses resultados também servem para responder a primeira pergunta de pesquisa considerada nesse estudo. Contudo, também é importante destacar que o modelo para dados com única visão obteve melhores médias para algumas bases e considerando algumas funções de aptidão. Por exemplo, para a base Internet, o método $PSO - SV$ conseguiu melhores médias com cinco funções. Para a base Animals-2, o $PSO - SV$ também apresentou maiores médias para quatro funções. Ainda, o método $PSO - SV$ apresentou menor dispersão em relação aos outros dois para a maioria dos casos. Todos os métodos apresentaram menor dispersão para os índices de Xu e homogeneidade.

As Tabelas 31 e 32 sumarizam os resultados encontrados, relativos as índices ARI e medida F, respectivamente, pelos métodos híbridos para agrupamento rígido considerando todas as funções de aptidão. A média e desvio padrão (em parênteses) são apresentados para cada índice. Alternativamente, os gráficos *box-plots* desses dados também foram gerados e estão apresentados no Apêndice A. As melhores médias para cada índice e base de dados são apresentadas em negrito. A Análise de Variância simples (do inglês, One way ANOVA) (MILLER, 1997) foi realizada com o intuito de analisar estatísticamente os resultados encontrados pelos modelos que utilizam múltiplas matrizes para cada base de dados, índice e grupo de funções, e verificar se existia diferença estatísticamente significativa entre os resultados encontrados. É possível utilizar a Análise de Variância devido ao tamanho da

amostra e também pelo fato das amostras serem independentes. A hipótese nula H_0, que considera que as médias apresentadas são iguais, foi rejeitada em todas as 44 análises realizadas. O pós-teste Holm-Bonferroni (HOLM, 1979) foi aplicado para verificar quais pares de funções apresentaram diferenças estatisticamente significativas.

A Tabela 53 sumariza os resultados obtidos pela aplicação do teste Holm-Bonferroni. A parte triangular superior da tabela apresenta os resultados das comparações para o ARI e a parte triangular inferior contém os resultados das comparações para a medida F. Para cada par de função e algoritmo, um par de valores é apresentado, onde o primeiro valor representa o número total de vezes que a função linha foi significativamente melhor considerando todas as bases de dados e o segundo valor representa o número total de vezes que a função coluna foi significativamente melhor considerando todas as bases de dados,

A Figura 14 mostra o número total de vezes que cada função foi melhor considerando todas as comparações significativas em todas as bases e ambos os índices. As três funções que se destacaram para os métodos $PSO - RWL$ e $PSO - RWG$ foram: o índice da silhueta, a homogeneidade e o índice de Xu. O índice da silhueta se destacou entre todas as funções para o método PSO_{RWL}, e a homogeneidade foi a função que se destacou para o método PSO_{RWG}. Dessa forma, os resultados dessas duas funções de aptidão foram selecionados para serem comparados aos resultados obtidos pelos outros métodos relacionais na próxima subseção. Esses resultados respondem a segunda pergunta de pesquisada considerada neste estudo.

Outra observação importante que deve ser feita é que, apesar do índice da silhueta ter obtido ótima colocação em comparação com as outras funções de aptidão, o cálculo desse índice é um dos mais complexos e que, por isso, a medida que a base de dados crescer, o tempo computacional exigido para computar esse índice também aumentará bastante. Consequentemente, se a base de dados for muito grande, o tempo computacional da silhueta não será interessante.

Além disso, as Figuras 10 e 11 apresentam os valores dos índices apresentados pelas melhores soluções encontradas pela otimização de cada função de aptidão.

Por último, Testes t (nível de significância $\alpha = 0.05$) foram realizados para comparar os resultados dos índices medida F e ARI obtidos pelo PSO_{RWL} e PSO_{RWG} tendo o índice da silhueta e a homogeneidade como funções de aptidão, isto é, os pares comparados foram PSO^{SIL}_{RWL}-PSO^{SIL}_{RWG} e PSO^{HM}_{RWL}-PSO^{HM}_{RWG}. Em 22 testes realizados, considerando o índice da homogeneidade intra-cluster, as diferenças foram significativas em dezoito comparações, onde o PSO_{RWL} apresentou melhores médias em nove testes e o PSO_{RWG} apresentou maiores médias nos outros nove. Por outro lado, considerando o índice da silhueta em 22 testes, as diferenças foram significativas em catorze comparações, onde o PSO_{RWL} apresentou melhores médias em nove testes e o PSO_{RWG} apresentou maiores médias nos outros cinco. Portanto, o PSO_{RWL} apresentou melhores agrupamentos de dados relacionais com múltiplas visões na maior parte dos casos em que houve diferença

Tabela 31 – Resultados relativos a ARI com diferentes funções

	Index	CH	CS	DB	DN	HA	HM	QE	SIL	SS	WB	XU
Animals-1	PSO-SV	0.01(.02)	0.00(.00)	0.01(.02)	0.00(.00)	0.01(.02)	0.06(.00)	0.05(.01)	0.05(.05)	0.00(.00)	0.01(.01)	0.05(.00)
	PSO-RWL	0.03(.07)	0.08(.09)	0.18(.07)	0.21(.01)	0.03(.06)	0.20(.02)	0.17(.08)	**0.30**(.03)	0.17(.08)	0.04(.07)	0.21(.02)
	PSO-RWG	0.03(.04)	0.00(.00)	0.01(.00)	0.00(.00)	**0.04**(.04)	0.01(.00)	0.01(.01)	**0.04**(.02)	0.00(.00)	0.03(.03)	0.01(.00)
Animals-2	PSO-SV	0.21(.00)	0.10(.11)	0.10(.11)	0.16(.09)	0.21(.01)	0.02(.01)	0.02(.01)	0.21(.01)	0.17(.08)	0.21(.02)	0.02(.00)
	PSO-RWL	0.21(.03)	0.21(.10)	0.32(.14)	0.24(.12)	0.21(.03)	**0.54**(.05)	0.49(.17)	0.47(.14)	0.40(.09)	0.22(.03)	0.53(.05)
	PSO-RWG	0.19(.04)	0.16(.09)	0.26(.08)	0.12(.07)	0.19(.04)	0.29(.00)	**0.40**(.06)	0.29(.07)	0.22(.08)	0.19(.04)	0.29(.00)
Corel-1	PSO-RWL	0.41(.02)	0.36(.02)	0.39(.05)	0.40(.02)	0.41(.02)	0.40(.00)	0.38(.03)	0.41(.01)	0.41(.01)	0.37(.05)	0.40(.00)
	PSO-RWL	0.49(.03)	0.46(.01)	0.47(.04)	0.49(.08)	0.49(.03)	0.46(.00)	0.46(.01)	**0.53**(.03)	0.41(.10)	0.49(.03)	0.46(.00)
	PSO-RWG	0.49(.04)	0.44(.01)	0.45(.04)	0.37(.13)	0.49(.04)	0.44(.00)	0.47(.00)	**0.52**(.02)	0.40(.09)	0.48(.05)	0.44(.00)
Corel-2	PSO-SV	0.28(.04)	0.26(.02)	0.27(.05)	0.37(.03)	0.30(.04)	0.27(.00)	0.26(.03)	0.33(.06)	0.28(.05)	0.27(.04)	0.30(.04)
	PSO-RWL	0.36(.05)	0.36(.07)	0.37(.07)	**0.46**(.04)	0.35(.05)	0.41(.00)	0.36(.04)	0.45(.01)	0.38(.06)	0.35(.05)	0.41(.00)
	PSO-RWG	0.33(.04)	0.31(.04)	0.33(.06)	**0.46**(.05)	0.33(.05)	0.39(.00)	0.36(.04)	0.45(.02)	0.34(.06)	0.33(.05)	0.39(.00)
Flowers	PSO-SV	0.09(.01)	0.09(.01)	0.06(.01)	0.07(.01)	0.09(.01)	0.10(.01)	0.09(.01)	0.09(.01)	0.09(.01)	0.06(.01)	0.10(.01)
	PSO-RWL	0.28(.01)	0.27(.02)	0.27(.03)	0.28(02)	0.28(.01)	0.27(.01)	0.26(.01)	0.27(.01)	**0.30**(.01)	0.28(.01)	0.27(.01)
	PSO-RWG	0.29(.01)	0.29(.02)	0.28(.03)	0.30(.02)	0.29(.01)	0.28(.02)	0.30(.01)	0.30(.02)	**0.31**(.01)	0.29(.01)	0.29(.01)
Image	PSO-SV	0.40(.00)	0.39(.05)	0.32(.04)	0.40(.03)	0.40(.00)	0.40(.00)	0.47(.05)	0.46(.01)	0.45(.03)	0.40(.03)	0.40(.00)
	PSO-RWL	0.43(.03)	0.37(.08)	0.48(.01)	0.43(.05)	0.42(.02)	0 .49(.00)	0.47(.02)	**0.50**(.00)	0.47(.00)	0.43(.03)	0.49(.00)
	PSO-RWG	0.42(.01)	0.42(.02)	0.48(.03)	0.42(.03)	0.42(.01)	0.49(.00)	0.48(.02)	**0.50**(.00)	0.43(.03)	0.42(.01)	0.49(.00)
Internet	PSO-SV	0.31(.08)	0.50(.00)	0.50(.01)	0.50(.01)	0.32(.09)	0.50(.00)	0.50(.00)	0.51(.01)	0.50(.01)	0.50(.00)	0.50(.00)
	PSO-RWL	**0.54**(.02)	0.09(.24)	0.22(.28)	0.18(.27)	**0.54**(.00)	0.53(.00)	0.19(.28)	0.53(.00)	0.50(.16)	0.53(.05)	0.53(.00)
	PSO-RWG	0.45(.12)	0.00(.12)	-0.02(.08)	-0.02(.08)	0.46(.10)	0.50(.00)	0.00(.13)	0.09(.01)	**0.52**(.01)	0.46(.10)	0.50(.00)
Mfeat	PSO-SV	0.48(.02)	0.45(.04)	0.26(.05)	0.43(.04)	0.49(.02)	0.50(.00)	0.50(.02)	0.37(.03)	0.48(.04)	0.46(.03)	0.50(.01)
	PSO-RWL	0.52(.05)	0.51(.04)	0.54(.05)	0.56(.05)	0.53(.05)	**0.67**(.00)	0.55(.05)	0.66(.03)	0.53(.07)	0.53(.04)	**0.67**(.00)
	PSO-RWG	0.45(.04)	0.48(.04)	0.52(.06)	0.57(.06)	0.46(.04)	**0.62**(.00)	0.57(.05)	0.62(.04)	0.52(.07)	0.45(.04)	**0.62**(.00)
Phoneme	PSO-SV	0.60(.09)	0.56(.11)	0.45(.12)	0.68(.06)	0.60(.09)	0.71(.00)	0.70(.05)	0.57(.10)	0.50(.05)	0.65(.08)	0.85(.00)
	PSO-RWL	0.59(.14)	0.52(.12)	0.29(.10)	0.66(.08)	0.58(.12)	0.72(.00)	0.67(.10)	**0.72**(.01)	0.51(.05)	0.62(.10)	**0.72**(.00)
	PSO-RWG	0.63(.07)	0.53(.10)	0.47(.08)	0.64(.10)	0.66(.07)	**0.72**(.00)	0.70(.06)	**0.72**(.01)	0.50(.04)	0.65(.06)	**0.72**(.00)
Water	PSO-SV	0.01(.00)	0.01(.00)	0.02(.01)	0.01(.00)	0.01(.00)	0.01(.00)	0.01(.00)	0.01(.00)	0.01(.00)	0.01(.00)	0.01(.00)
	PSO-RWL	0.03(.00)	0.03(.00)	**0.04**(.02)	0.02(.00)	0.03(.00)	0.03(.00)	0.02(.00)	0.03(.00)	0.03(.00)	0.03(.00)	0.03(.00)
	PSO-RWG	**0.03**(.00)	**0.03**(.00)	**0.03**(.01)	**0.03**(.00)	**0.03**(.00)	**0.03**(.00)	0.02(.00)	**0.03**(.00)	**0.03**(.00)	**0.03**(.00)	**0.03**(.00)
3-Sources	PSO-SV	0.20(.07)	0.19(.07)	0.19(.05)	0.15(.05)	0.22(.06)	0.23(.03)	0.21(.04)	0.07(.03)	0.24(.04)	0.17(.05)	0.20(.04)
	PSO-RWL	**0.33**(.07)	0.16(.06)	0.26(.09)	0.22(.10)	**0.33**(.07)	0.32(.02)	0.32(.03)	0.11(.05)	**0.33**(.03)	**0.33**(.05)	**0.33**(.03)
	PSO-RWG	0.35(.05)	0.18(.07)	0.26(.09)	0.25(.10)	**0.36**(.04)	**0.36**(.02)	0.32(.03)	0.12(.05)	0.34(.03)	0.34(.07)	0.35(.02)

significativa para as bases e índices considerados. Os resultados desses testes servem como base para responder a terceira pergunta de pesquisa considerada neste estudo. Portanto, a depender da função de aptidão considerada, a estimação dos pesos tem influência. Para a silhueta, por exemplo, o uso dos pesos localmente estimados produziu melhores resultados.

7.2.3.1 Análise de robustez

A robustez das funções foi analisada com o critério utilizado em (LIU et al., 2010). O desempenho relativo da função m sobre uma base de dados é representada pela razão b_m entre o valor médio do índice avaliado e o maior valor médio do índice encontrado por todas as funções comparadas. Por exemplo, para a medida F, b_m é calculado de acordo com a Eq. (7.4), onde k representa o número de funções.

$$b_m = \frac{F_m}{max_k F_k} \tag{7.4}$$

A partir da Eq. (7.4). a melhor função m^* sobre a base de dados possui $b_m = 1$ e todas as outras funções tem $b_m \leq 1$. Quanto maior o valor de b_m, melhor é o desempenho da função em relação ao melhor desempenho naquela base de dados. Portanto, a soma de b_m

Tabela 32 – Resultados relativos a medida F com diferentes funções

	Index	CH	CS	DB	DN	HA	HM	QE	SIL	SS	WB	XU
Animals-1	PSO-SV	0.34(.02)	0.33(.00)	0.34(.03)	0.00(.00)	0.34(.02)	0.37(.00)	0.37(.02)	0.37(.04)	0.33(.00)	0.34(.02)	0.37(.01)
	PSO-RWL	0.36(.06)	0.40(.08)	0.49(.06)	0.51(.00)	0.36(.06)	0.50(.02)	0.47(.07)	**0.54**(.03)	0.48(.07)	0.37(.07)	0.50(.01)
	PSO-RWG	**0.35**(.04)	0.33(.00)	0.33(.01)	0.33(.00)	**0.35**(.04)	0.33(.00)	0.34(.01)	0.34(.02)	0.33(.00)	**0.35**(.04)	0.33(.00)
Animals-2	PSO-SV	0.49(.01)	0.41(.08)	0.41(.08)	0.46(.07)	0.49(.00)	0.33(.00)	0.37(.01)	0.49(.01)	0.47(.08)	0.49(.01)	0.33(.00)
	PSO-RWL	0.51(.03)	0.50(.07)	0.58(.09)	0.53(.09)	0.50(.03)	**0.71**(.02)	0.70(.10)	0.66(.10)	0.64(.07)	0.51(.02)	**0.71**(.02)
	PSO-RWG	0.48(.04)	0.48(.07)	0.54(.05)	0.45(.05)	0.48(.04)	0.53(.01)	**0.62**(.04)	0.54(.05)	0.52(.05)	0.48(.04)	0.54(.01)
Corel-1	PSO-SV	0.66(.03)	0.62(.01)	0.67(.03)	0.66(.02)	0.65(.02)	0.64(.00)	0.64(.02)	0.65(.00)	0.65(.01)	0.64(.04)	0.64(.00)
	PSO-RWL	0.73(.02)	0.70(.00)	0.71(.03)	0.72(.05)	0.73(.02)	0.68(.00)	0.70(.01)	**0.75**(.03)	0.66(.07)	0.73(.02)	0.68(.00)
	PSO-RWG	0.72(.03)	0.69(.00)	0.70(.03)	0.64(.10)	0.73(.03)	0.69(.00)	0.70(.00)	**0.74**(.03)	0.66(.07)	0.72(.04)	0.69(.00)
Corel-2	PSO-SV	0.56(.04)	0.53(.02)	0.54(.04)	0.64(.02)	0.57(.05)	0.54(.00)	0.54(.03)	0.60(.06)	0.56(.05)	0.56(.04)	0.57(.04)
	PSO-RWL	0.62(.06)	0.63(.06)	0.65(.06)	**0.72**(.03)	0.61(.06)	0.69(.00)	0.66(.04)	0.69(.01)	0.65(.06)	0.62(.05)	0.69(.00)
	PSO-RWG	0.60(.05)	0.58(.04)	0.61(.06)	**0.71**(.03)	0.60(.06)	0.67(.00)	0.65(.04)	0.69(.01)	0.61(.06)	0.59(.06)	0.67(.00)
Flowers	PSO-SV	0.28(.02)	0.28(.01)	0.21(.01)	0.24(.02)	0.28(.01)	0.29(.01)	0.27(.02)	0.28(.01)	0.27(.01)	0.22(.02)	0.29(.01)
	PSO-RWL	0.46(.02)	0.47(.02)	0.45(.03)	0.47(.03)	0.47(.02)	0.46(.01)	0.44(.01)	0.46(.01)	**0.49**(.01)	0.46(.02)	0.46(.01)
	PSO-RWG	0.48(.01)	**0.49**(.02)	0.45(.03)	**0.49**(.02)	**0.49**(.01)	0.47(.02)	**0.49**(.01)	**0.49**(.02)	**0.49**(.02)	0.48(.01)	**0.49**(.01)
Image	PSO-SV	0.58(.00)	0.58(.04)	0.54(.03)	0.58(.03)	0.58(.00)	0.58(.00)	0.63(.04)	0.65(.01)	0.62(.03)	0.58(.02)	0.58(.00)
	PSO-RWL	0.60(.02)	0.56(.06)	0.65(.02)	0.61(.04)	0.59(.02)	0.65(.00)	0.65(.01)	**0.67**(.00)	0.63(.00)	0.60(.02)	0.65(00)
	PSO-RWG	0.59(.01)	0.61(.02)	0.65(.02)	0.60(.03)	0.59(.01)	0.66(.00)	0.65(.01)	**0.67**(.01)	0.60(.02)	0.59(.01)	0.66(.00)
Internet	PSO-SV	0.83(.03)	0.89(.00)	0.89(.00)	0.89(.00)	0.83(.03)	0.89(.00)	0.89(.00)	0.89(.00)	0.89(.00)	0.89(.00)	0.89(.00)
	PSO-RWL	**0.90**(.00)	0.82(.04)	0.85(.05)	0.84(.05)	**0.90**(.00)	**0.90**(.00)	0.84(.05)	**0.90**(.00)	0.89(.04)	**0.90**(.01)	**0.90**(.00)
	PSO-RWG	0.87(.04)	0.80(.01)	0.80(.01)	0.80(.01)	0.87(.05)	**0.89**(.00)	0.80(.02)	0.69(.01)	**0.89**(.00)	0.88(.05)	**0.89**(.00)
Mfeat	PSO-SV	0.67(.03)	0.62(.04)	0.43(.04)	0.58(.04)	0.67(.03)	0.68(.00)	0.70(.02)	0.52(.03)	0.65(.05)	0.62(.04)	0.68(.01)
	PSO-RWL	0.70(.04)	0,69(.04)	0.71(.05)	0.71(.05)	0.71(.04)	**0.83**(.00)	0.72(.04)	0.82(.03)	0.68(.06)	0.71(.04)	**0.83**(.00)
	PSO-RWG	0.62(.05)	0.66(.05)	0.68(.06)	0.71(.05)	0.62(.05)	**0.80**(.00)	0.74(.04)	0.79(.03)	0.68(.07)	0.61(.05)	**0.80**(.00)
Phoneme	PSO-SV	0.76(.07)	0.72(.08)	0.65(.07)	0.82(.05)	0.76(.07)	0.85(.00)	0.82(.04)	0.73(.07)	0.69(.03)	0.44(.13)	0.72(.00)
	PSO-RWL	0.75(.10)	0.70(.08)	0.54(.07)	0.82(.05)	0.75(.08)	**0.85** (.00)	0.80(.07)	**0.85**(.00)	0.69(.04)	0.77(.07)	**0.85**(.00)
	PSO-RWG	0.78(.06)	0.70(.07)	0.67(.04)	0.81(.07)	0.80(.05)	**0.85**(.00)	0.82(.04)	**0.85**(.01)	0.68(.03)	0.79(.05)	**0.85**(.00)
Water	PSO-SV	0.21(.01)	0.21(.01)	0.34(.03)	0.22(.01)	0.21(.01)	0.21(.01)	0.26(.02)	0.22(.01)	0.23(.01)	0.21(.01)	0.21(.01)
	PSO-RWL	**0.33**(.07)	0.16(.06)	0.26(.09)	0.22(.10)	**0.33**(.07)	0.32(.02)	0.32(.03)	0.11(.05)	**0.33**(.03)	**0.33**(.05)	**0.33**(.03)
	PSO-RWG	0.35(.05)	0.18(.07)	0.26(.09)	0.25(.10)	**0.36**(.04)	**0.36**(.02)	0.32(.03)	0.12(.05)	0.34(.03)	0.34(.07)	0.35(.02)
3-Sources	PSO-SV	0.47(.05)	0.47(.05)	0.48(.03)	0.44(.05)	0.48(.05)	0.51(.02)	0.49(.03)	0.37(.03)	0.51(.02)	0.46(.04)	0.49(.03)
	PSO-RWL	0.26(.01)	0.26(.01)	**0.38**(.05)	0.26(.01)	0.27(.01)	0.26(.02)	0.25(.01)	0.25(.00)	0.27(.01)	0.26(.01)	0.26(.01)
	PSO-RWG	0.56(.03)	0.48(.05)	0.51(.05)	0.50(.06)	0.56(.02)	**0.57**(.02)	0.54(.02)	0.42(.04)	0.56(.02)	0.55(.05)	**0.57**(.02)

para todas as bases de dados fornece uma medida de robustez da função m. Valores altos para essa soma indicam boa robustez (LIU et al., 2010).

As Figuras 15 e 16 apresentam as distribuições de b_m de cada função. Verifica-se que, para o método $PSO - RWL$ e ambos os índices externos, os índices de Xu e a homogeneidade intra-cluster foram mais robustos dentre as funções comparadas. Para o método $PSO - RWG$ e ambos os índices, os índices de Xu e a homogeneidade intra-cluster também foram mais robustos.

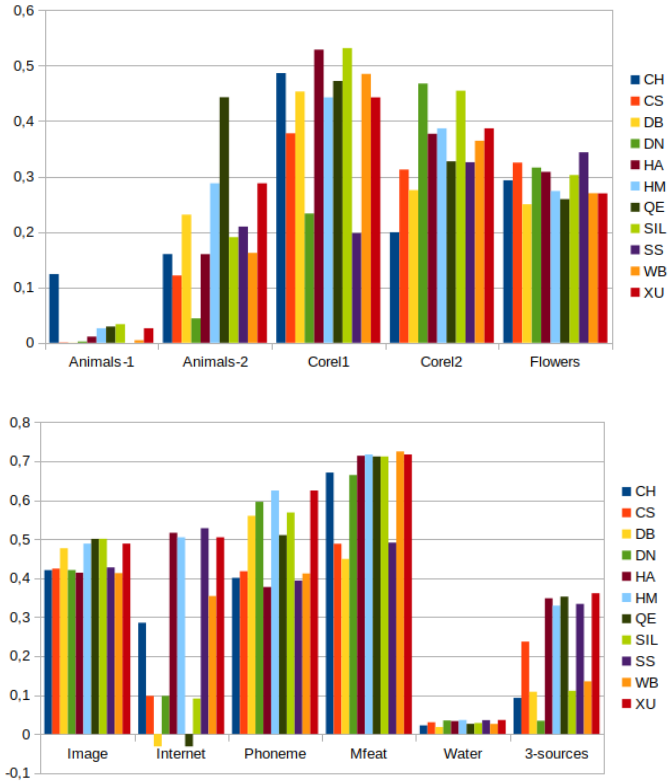


Figura 10 – Melhor solução encontrada pelo PSO-RWL com cada função em termos de F-measure

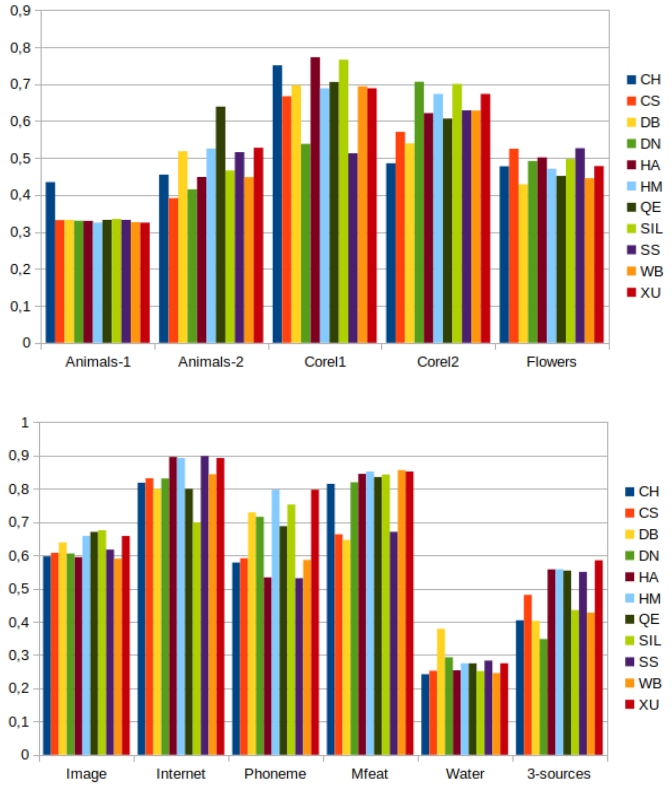


Figura 11 – Melhor solução encontrada pelo PSO-RWL com cada função em termos de ARI

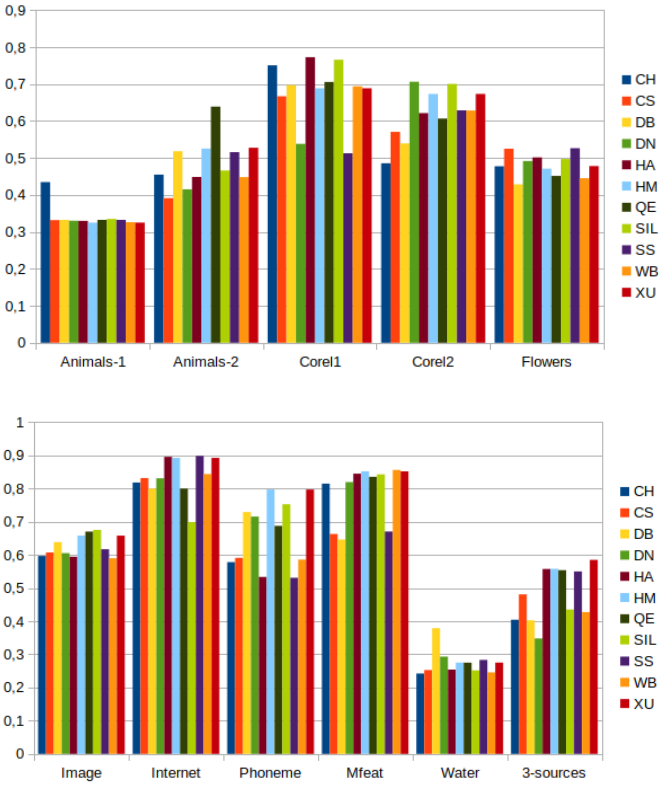


Figura 12 – Melhor solução encontrada pelo PSO-RWG com cada função em termos de F-measure

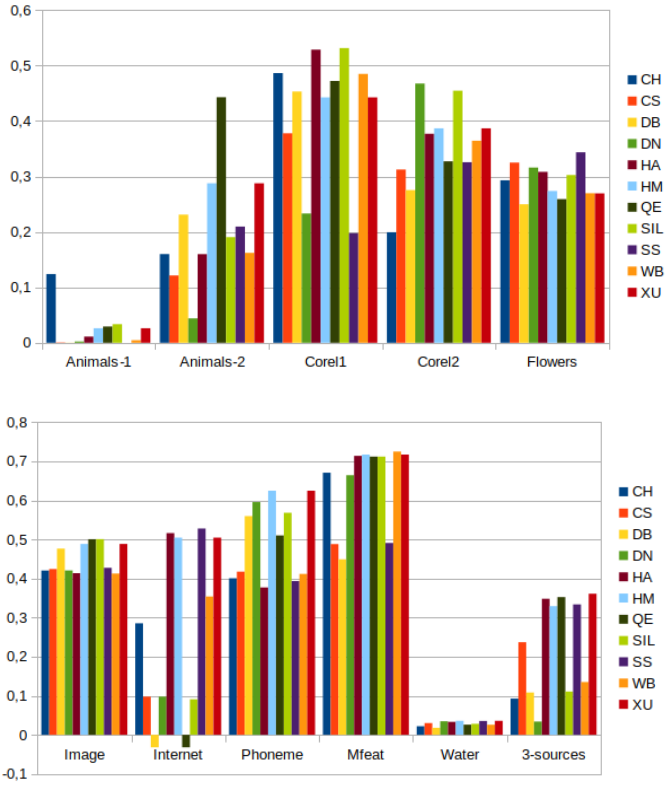


Figura 13 – Melhor solução encontrada pelo PSO-RWGL com cada função em termos de ARI

Tabela 33 – Sumário dos resultados encontrados pela aplicação do teste Holm-Bonferroni para todos os pares

	Algoritmo	CH	CS	DB	DN	HA	HM	QE	SIL	SS	WB	XU
CH	PSO-RWL	-	4-0	3-3	2-4	0-0	2-6	3-4	3-4	2-6	0-0	2-6
	PSO-RWG	-	4-0	5-3	4-5	0-1	1-8	4-6	2-7	2-4	0-1	1-8
CS	PSO-RWL	0-5	-	2-3	0-6	1-3	0-8	1-6	1-4	1-6	0-5	0-8
	PSO-RWG	1-4	-	3-4	2-5	0-4	0-9	1-8	1-6	1-7	1-4	0-9
DB	PSO-RWL	3-5	4-2	-	2-2	4-3	0-7	1-3	1-4	2-5	3-3	0-7
	PSO-RWG	2-7	2-4	-	3-6	4-5	1-5	1-5	2-7	3-3	5-4	1-5
DN	PSO-RWL	3-2	5-0	2-3	-	3-2	2-6	3-3	2-6	4-5	2-2	2-6
	PSO-RWG	3-3	4-1	5-4	-	5-4	1-8	2-5	2-6	2-4	4-5	2-8
HA	PSO-RWL	0-0	4-1	3-6	3-2	-	3-6	4-3	3-7	2-5	0-0	3-5
	PSO-RWG	1-0	4-1	5-4	4-2	-	1-7	4-6	2-7	2-4	0-0	1-7
HM	PSO-RWL	6-1	8-1	7-2	6-3	6-2	-	6-0	2-5	4-2	5-2	0-1
	PSO-RWG	7-2	8-2	6-2	8-2	6-2	-	7-3	3-4	7-2	7-1	0-0
QE	PSO-RWL	4-4	7-2	3-1	3-4	5-4	2-6	-	1-6	3-2	3-3	0-5
	PSO-RWG	6-3	8-3	5-2	4-4	5-5	3-7	-	3-6	5-4	6-2	2-8
SIL	PSO-RWL	7-3	7-3	7-3	6-3	7-3	4-4	7-2	-	5-3	7-3	5-2
	PSO-RWG	6-2	6-2	6-3	5-4	6-2	4-4	6-4	-	7-4	8-2	4-3
SS	PSO-RWL	5-2	6-1	5-3	5-4	5-3	1-5	3-5	4-5	-	3-3	1-6
	PSO-RWG	6-2	5-2	3-2	4-2	7-2	2-5	5-6	3-5	-	3-5	2-6
WB	PSO-RWL	0-0	5-0	4-4	2-3	1-2	1-6	3-5	2-7	3-5	-	2-6
	PSO-RWG	0-1	4-1	5-4	3-3	0-0	2-6	4-6	2-7	1-5	-	1-7
XU	PSO-RWL	6-1	8-1	7-2	6-3	5-2	0-1	7-1	4-4	5-1	7-1	-
	PSO-RWG	7-2	8-2	6-2	8-2	7-2	0-1	8-1	4-4	5-2	6-2	-

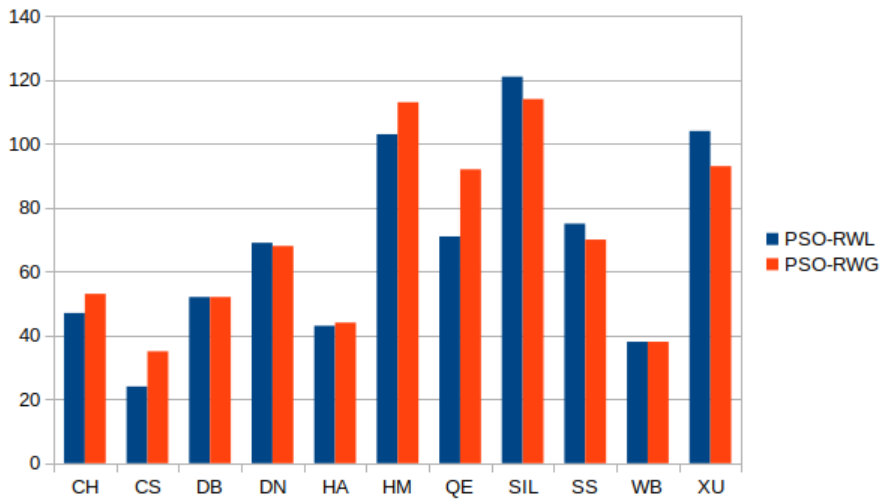
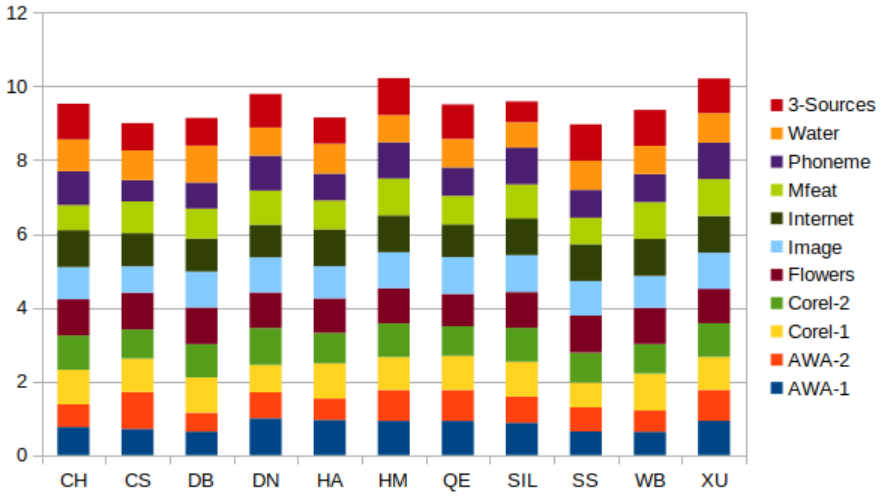
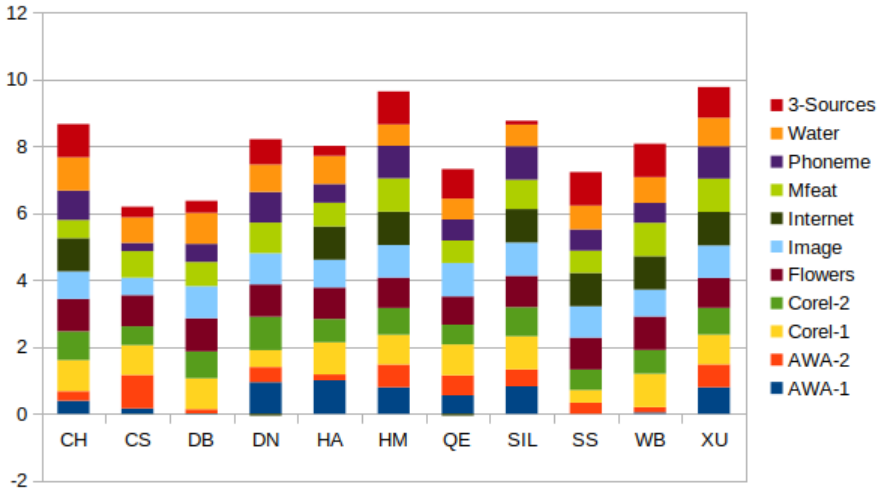


Figura 14 – Sumário dos resultados Bonferroni

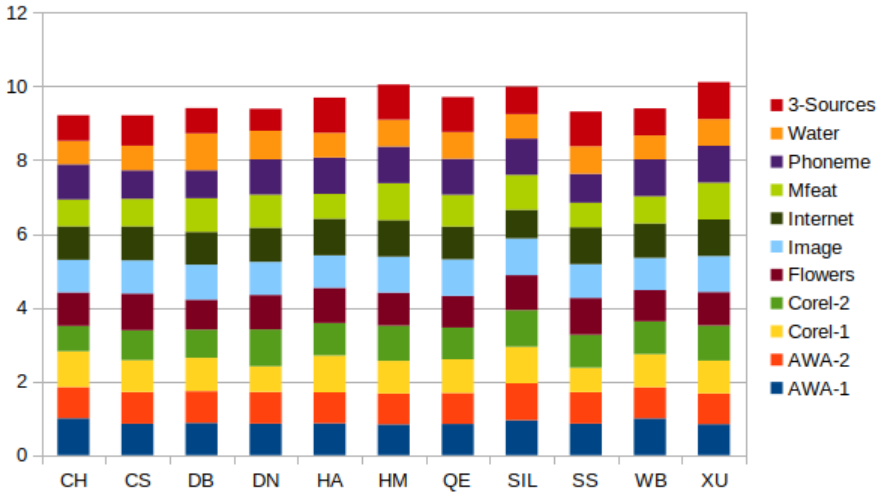


(a) Medida F

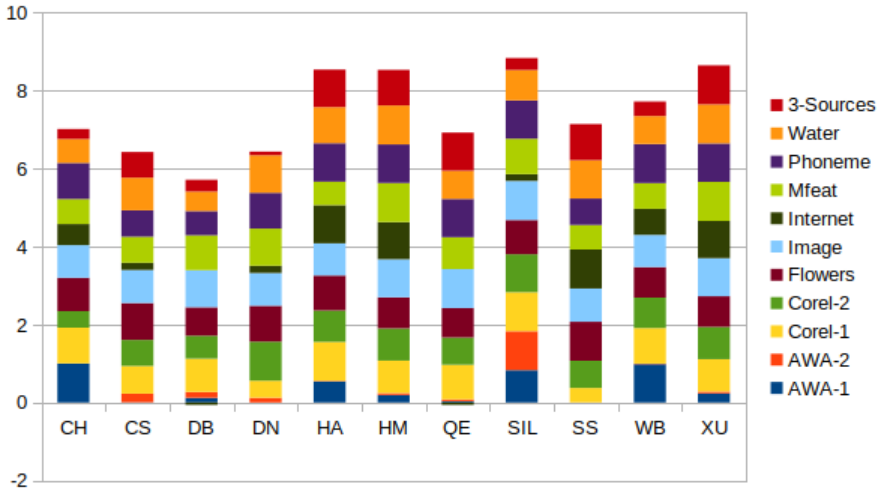


(b) ARI

Figura 15 – Robustez das funções de aptidão para o PSO-RWL



(a) Medida F



(b) ARI

Figura 16 – Robustez das funções de aptidão para o PSO-RWG

7.3 ESTUDO 3 - COMPARAÇÃO COM OUTROS MÉTODOS

Este estudo teve como objetivo comparar o desempenho obtido pelos métodos híbridos desenvolvidos com outros métodos presentes na literatura. Os métodos comparados foram: Hard C-Medoids (HCMdd) (KRISHNAPURAM; JOSHI; YI, 1999), FANNY (KAUFMAN; ROUSSEEUW, 1990), NERF (HATHAWAY; BEZDEK, 1994), PSO_SV (GUSMãO; CARVALHO, 2016), $CARD_R$ (FRIGUI; HWANG; RHEE, 2007), $MRDCA_{RWL}$ e $MRDCA_{RWG}$ (CARVALHO; LECHEVALLIER; MELO, 2012). Além desses métodos, os algoritmos TW-k-Means e EW-k-Means (CHEN et al., 2013), adequados para agrupamento de dados vetoriais com múltiplas visões, também foram selecionados para comparação.

Todos os algoritmos foram executados 50 vezes com número máximo de iterações definido em 100 para cada execução. Os índices externos foram computados para cada partição final encontrada em cada execução. O parâmetro $|G_k|$ dos algoritmos $HCMdd$, PSO_R, $MRDCA_{RWL}$ e $MRDCA_{RWG}$ foi definido com valor igual a 5. Os parâmetros m, T e ϵ dos métodos $FANNY$, $NERF$ e $CARD_R$ foram definidos, respectivamente, com valores iguais a 2, 100 ed 10^{-9}. Com relação aos algoritmos TW-k-Means e EW-k-Means, uma análise paramétrica, similar a que foi realizada em Chen et al. (2013), foi feita para selecionar os melhores parâmetros para cada base de dados.

A Tabela 34 sumariza os resultados encontrados pelos métodos. A média e desvio padrão são apresentados para cada índice externo. O melhor valor para cada índice e cada base de dados é apresentado em negrito. Os resultados dos algoritmos TW-k-Means e EW-k-Means para a base Flowers não são mostrados, pois os atributos não foram disponibilizados, apenas as matrizes de dissimilaridades. Portanto, apenas métodos adequados para lidar com dados relacionais podem ser avaliados para essa base de dados.

Uma Análise de Variância simples (do inglês, One way ANOVA) (MILLER, 1997) foi realizada para cada base de dados, índice e grupo de funções para verificar se existia diferença estatísticamente significante entre os resultados encontrados pelo grupo. A hipótese nula foi rejeitada para todos os casos. Após a constatação de que havia diferenças significativas nos resultados pelos grupos, o pós-teste Holm-Bonferroni (HOLM, 1979) foi aplicado para comparar os métodos propostos com os outros algoritmos.

No total, 396 comparações foram realizadas. O símbolo "*"indica que a diferença entre as médias, considerando o método PSO-RWL, foi significante estatisticamente. O símbolo "+"indica que a diferença entre as médias, considerando o método PSO-RWG, foi significante estatisticamente. Se o símbolo estiver vermelho, então a média apresentada pelo algoritmo comparado foi maior do que o método proposto.

Levando em consideração as comparações feitas com o PSO-RWL, as diferenças foram significantes estatisticamente em 95 testes para a medida F e em 92 testes para o ARI, onde o PSO-RWL foi melhor em 72 testes para a medida F e em 74 testes para o ARI. Considerando as comparações feitas com o PSO-RWG, as diferenças foram significantes estatisticamente em 94 testes para a medida F e em 92 testes para o ARI, onde o PSO-RWG

foi melhor em 76 testes para a medida F e em 77 testes para o ARI.

Estes resultados demonstram a efetividade dos métodos híbridos desenvolvidos para agrupamento rígido de dados relacionais com múltiplas visões. A abordagem desenvolvida apresentou resultados competitivos quando comparada a outros métodos presentes na literatura e ainda melhores resultados na maioria dos casos.

Tabela 34 – Sumário dos resultados encontrados pelos outros algoritmos

Base de dados	Indice	HCMdd	NERF	$FANNY$	PSO_R	$CARD_R$	RWL	RWG	TW-k-Means	EW-k-Means
Animals-1	ARI	0.04(.01)$^{*+}$	0.10(.01)$^{*+}$	0.21(.05)$^{*+}$	0.06(.00)$^{*+}$	0.18(.06)$^{*+}$	**0.22**(.04)$^{*+}$	0.03(.02)*	0.00(.00)$^{*+}$	0.00(.00)$^{*+}$
	F-measure	0.33(.01)$^{*+}$	0.41(.01)$^{*+}$	0.50(.05)$^{*+}$	0.39(.00)$^{*+}$	0.48(.05)$^{*+}$	**0.50**(.03)$^{*+}$	0.32(.02)*	0.32(.01)*	0.33(.00)$^{*+}$
Animals-2	ARI	0.20(.00)$^{*+}$	0.01(.00)$^{*+}$	0.16(.09)$^{*+}$	0.23(.02)$^{*+}$	0.16(.00)$^{*+}$	**0.40**(.10)$^{+}$	0.03(.02)$^{*+}$	0.01(.02)$^{*+}$	0.01(.04)$^{*+}$
	F-measure	0.43(.01)$^{*+}$	0.32(.01)$^{*+}$	0.45(.07)$^{*+}$	0.48(.02)$^{*+}$	0.46(.00)$^{*+}$	**0.63**(.07)$^{*+}$	0.33(.02)$^{*+}$	0.34(.02)$^{*+}$	0.34(.03)$^{*+}$
Corel-1	ARI	0.39(.06)$^{*+}$	0.26(.03)$^{*+}$	0.36(.14)$^{*+}$	0.40(.00)$^{*+}$	0.29(.05)$^{*+}$	**0.48**(.09)$^{*+}$	0.47(.05)$^{*+}$	0.47(.07)$^{*+}$	0.36(.08)$^{*+}$
	F-measure	0.66(.05)$^{*+}$	0.56(.02)$^{*+}$	0.63(.09)$^{*+}$	0.64(.00)$^{*+}$	0.58(.04)$^{*+}$	**0.71**(.07)*	0.70(.04)*	**0.71**(.04)$^{*+}$	0.63(.08)$^{*+}$
Corel-2	ARI	0.28(.06)$^{*+}$	0.23(.00)$^{*+}$	0.33(.10)$^{*+}$	0.27(.00)$^{*+}$	0.23(.02)$^{*+}$	0.38(.05)*	0.38(.06)*	**0.52**(.05)$^{*+}$	0.28(.05)$^{*+}$
	F-measure	0.57(.05)$^{*+}$	0.52(.00)$^{*+}$	0.61(.06)$^{*+}$	0.54(.00)$^{*+}$	0.51(.01)$^{*+}$	0.65(.05)$^{*+}$	0.65(.06)$^{*+}$	**0.76**(.04)$^{*+}$	0.56(.05)$^{*+}$
Flowers	ARI	0.25(.03)$^{*+}$	0.14(.02)$^{*+}$	0.20(.05)$^{*+}$	**0.35**(.01)$^{*+}$	0.10(.03)$^{*+}$	0.27(.02)$^{+}$	0.28(.03)*	-	-
	F-measure	0.43(.03)$^{*+}$	0.28(.02)$^{*+}$	0.35(.04)$^{*+}$	**0.53**(.02)$^{*+}$	0.24(.03)$^{*+}$	0.46(.02)$^{+}$	0.47(.03)$^{+}$	-	-
Image	ARI	0.33(.05)$^{*+}$	0.35(.05)$^{*+}$	0.34(.05)$^{*+}$	**0.41**(.01)$^{*+}$	0.40(.01)$^{*+}$	0.40(.09)$^{*+}$	0.38(.09)$^{*+}$	0.36(.04)$^{*+}$	0.33(.07)$^{*+}$
	F-measure	0.56(.05)$^{*+}$	0.57(.05)$^{*+}$	0.56(.06)$^{*+}$	**0.63**(.01)$^{*+}$	0.59(.02)$^{*+}$	0.59(.06)$^{*+}$	0.58(.07)$^{*+}$	0.57(.03)$^{*+}$	0.57(.06)$^{*+}$
Internet	ARI	0.40(.13)$^{*+}$	0.29(.00)$^{*+}$	0.29(.00)$^{*+}$	**0.50**(.00)$^{*+}$	0.20(.04)$^{*+}$	0.27(.20)$^{+}$	0.29(.21)$^{+}$	0.12(.25)$^{+}$	0.20(.25)$^{*+}$
	F-measure	0.85(.05)$^{*+}$	0.82(.00)$^{*+}$	0.82(.00)$^{*+}$	**0.89**(.00)$^{*+}$	0.77(.03)$^{*+}$	0.80(.08)$^{*+}$	0.82(.06)$^{+}$	0.76(.09)$^{*+}$	0.76(.11)$^{*+}$
Mfeat	ARI	0.39(.04)$^{*+}$	0.32(.00)$^{*+}$	0.33(.01)$^{*+}$	0.50(.00)$^{*+}$	0.22(.05)$^{*+}$	**0.58**(.06)$^{*+}$	0.56(.06)$^{*+}$	0.35(.06)$^{*+}$	0.43(.03)$^{*+}$
	F-measure	0.58(.05)$^{*+}$	0.48(.00)$^{*+}$	0.50(.02)$^{*+}$	0.68(.00)$^{*+}$	0.39(.05)$^{*+}$	**0.74**(.05)$^{*+}$	0.72(.06)$^{*+}$	0.53(.06)$^{*+}$	0.60(.04)$^{*+}$
Phoneme	ARI	0.48(.08)$^{*+}$	0.20(.01)$^{*+}$	0.39(.09)$^{*+}$	**0.71**(.00)$^{*+}$	0.21(.04)$^{*+}$	0.62(.10)$^{*+}$	0.60(.11)$^{*+}$	**0.71**(.05)	0.41(.28)$^{*+}$
	F-measure	0.68(.07)$^{*+}$	0.47(.01)$^{*+}$	0.63(.07)$^{*+}$	**0.85**(.00)$^{*+}$	0.51(.03)$^{*+}$	0.77(.08)$^{*+}$	0.75(.09)$^{*+}$	0.84(.04)	0.62(.20)$^{*+}$
Water	ARI	0.02(.00)$^{*+}$	0.02(.00)$^{*+}$	0.01(.00)$^{*+}$	0.01(.00)$^{*+}$	−0.02(.00)$^{*+}$	0.03(.02)	0.03(.00)*	**0.04**(.01)$^{*+}$	0.02(.00)$^{*+}$
	F-measure	0.29(.01)$^{*+}$	0.21(.01)$^{*+}$	0.21(.02)$^{*+}$	0.20(.00)$^{*+}$	**0.33**(.01)$^{*+}$	0.28(.02)$^{*+}$	0.28(.02)$^{*+}$	0.31(.02)$^{*+}$	0.23(.01)$^{*+}$
3-sources	ARI	0.18(.03)$^{*+}$	**0.39**(.03)$^{*+}$	0.07(.02)$^{*+}$	0.32(.02)$^{*+}$	**0.39**(.03)$^{*+}$	0.23(.06)$^{*+}$	0.25(.06)$^{*+}$	0.01(.01)$^{*+}$	0.01(.1)$^{*+}$
	F-measure	0.46(.02)$^{*+}$	0.58(.00)$^{*+}$	0.38(.02)$^{*+}$	0.54(.01)$^{*+}$	**0.59**(.00)$^{*+}$	0.49(.04)$^{*+}$	0.51(.05)$^{*+}$	0.37(.01)$^{*+}$	0.37(.00)$^{*+}$

Para fins de ilustração, a Figura 17 apresenta o tempo de execução dos métodos para agrupamento de dados relacionais com várias visões para as bases Multiple features, Image e Phoneme. Comparado aos métodos $MRDCA_{RWL}$ e $MRDCA_{RWG}$, o método PSO_{RWL} é mais lento uma vez que integra passos modificados do $MRDCA-RWL$ dentro da otimização de enxame e foi projetado para realizar uma melhor busca no espaço de soluções. Contudo, a qualidade das partições encontradas pelo método PSO_{RWL} foi melhor de forma significativa para essas bases de dados e tendo essas duas funções como referência. Também é possível observar que o índice da silhueta é mais custoso em termos de tempo computacional, como já era esperado, quando comparado a homogeneidade como função de aptidão, pois o cálculo do índice da silhueta é mais complexo. O método $CARD-R$, por sua vez, também apresentou tempos altos em comparação aos outros métodos, isso se deve também ao fato de que ele possui uma etapa adicional para calcular os graus de pertinência dos objetos aos grupos, pois é um método de agrupamento nebuloso.

7.3.1 Aplicação: Image Segmentation

A Tabela 35 apresenta a matriz de confusão fornecida pela melhor solução encontrada pelo PSO_{RWL} com a homogeneidade como função objetivo para a base de dados *Image Segmentation*. Os clusters 1 e 6 correspondem as classes a priori 2 e 7, respectivamente.

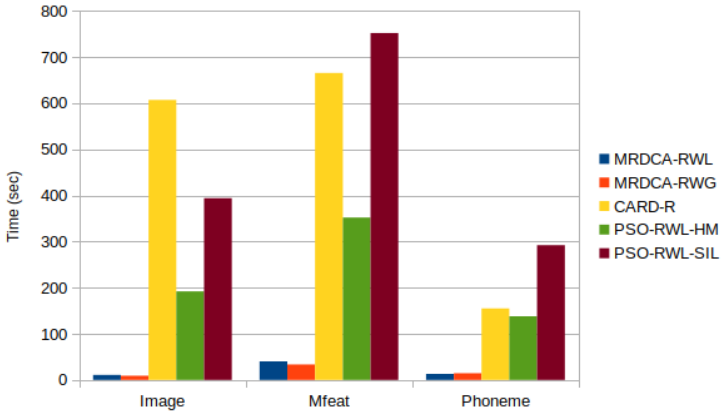


Figura 17 – Tempo computacional dos métodos para agrupamentos de dados relacionais com múltiplas visões

Observa-se também que o algoritmo falhou em descobrir a maior parte das outras classes a priori.

Tabela 35 – Matriz de confusão da partição apresentada pelo PSO_{RWL} com a homogeneidade como função de aptidão

	Clusters						
Classes	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
P_1	0	68	184	78	0	0	0
P_2	330	0	0	0	0	0	0
P_3	0	63	52	193	6	2	14
P_4	0	8	54	5	116	0	147
P_5	0	194	73	60	0	0	3
P_6	0	0	35	0	137	0	158
P_7	0	3	0	5	0	322	0

A Tabela 36 apresenta os pesos de relevância das visões encontrados pela melhor solução obtida por cada método para dados relacionais com múltiplas visões considerando a base *Image Segmentation*. Os pesos de relevância mostram a importância de cada visão para definição de cada grupo. Observa-se que o peso de relevância da matriz de dissimilaridades "RGB"foi mais importante para definição de todos os grupos para todos os algoritmos. Em relação aos pesos encontrados pela homogeneidade como função de aptidão, os grupos 5, 6 e 7 obtiveram pesos mais altos para a visão 2 (RGB). Portanto, esses grupos foram mais homogêneos considerando essa matriz de dissimilaridades.

As Figuras 18 - 28 mostram os gráficos box-plots dos índices externos computados para cada partição encontrada por cada algoritmo em todas as execuções. Com esses gráficos, é possível constatar quais métodos apresentaram maior variação nos valores dos índices e quais deles obtiveram maiores medianas, isto é, quais deles conseguiram encontrar melhores partições nas execuções. Os métodos híbridos desenvolvidos apresentaram, de

Tabela 36 – Image: vetores de pesos de relevância

	$PSO_{RWL} - SIL$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$PSO_{RWG} - SIL$
1 - shape	0,539	0,722	0,537	0,431	0,662	0,524	0,371	0,547
2 - rgb	1,853	1,384	1,860	2,318	1,510	1,907	2,688	1,827
	$PSO_{RWL} - HM$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$PSO_{RWG} - HM$
1 - shape	0,743	0,628	0,557	0,552	0,382	0,434	0,443	0,539
2 - rgb	1,344	1,591	1,793	1,808	2,612	2,299	2,255	1,854
	$MRDCA_{RWL}$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$MRDCA_{RWG}$
1 - shape	0,743	0,382	0,442	0,557	0,553	0,617	0,435	0,539
2 - rgb	1,344	2,612	2,261	1,795	1,806	1,619	2,299	1,853
	$CARD - R$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
1 - shape	0,384	0,277	0,467	0,408	0,349	0,210	0,377	
2 - rgb	0,616	0,723	0,533	0,592	0,651	0,790	0,623	

forma geral, pouca variação e obtiveram medianas mais altas em comparação com outros os métodos na maior parte dos casos.

7.4 DISCUSSÃO

O Estudo 1 foi um estudo piloto e teve como objetivo observar se o método híbrido baseado em PSO seria capaz de apresentar bons resultados, quando comparado a outros métodos, para o agrupamento de dados relacionais com apenas uma visão. A partir do Estudo 1, observou-se que o método desenvolvido apresentou resultados interessantes e promissores, o que motivou a investigação para o cenário com várias visões. Inicialmente, apenas uma função de aptidão havia sido considerada. No entanto, observou-se a necessidade de considerar outras funções, visto que existem várias na literatura e cada uma avalia aspectos diferentes dos agrupamentos. Tendo isso em mente, formulou-se o Estudo 2.

A partir do Estudo 2, foi possível investigar respostas para três perguntas de pesquisa. Os resultados obtidos no estudo sugerem que agrupar dados relacionais com múltiplas visões pode trazer melhores resultados em relação ao agrupamento de dados com visão única. Além disso, as três funções de aptidão que se destacaram, com base nas análises efetuadas, foram: o índice da Silhueta, o índice de Xu e a Homogeneidade intra-grupo. É interessante salientar que apenas o índice da silhueta avalia simultaneamente a separação dos grupos quanto a coesão interna deles. Os outros dois índices buscam por partições que possuem grupos mais homogeneos.

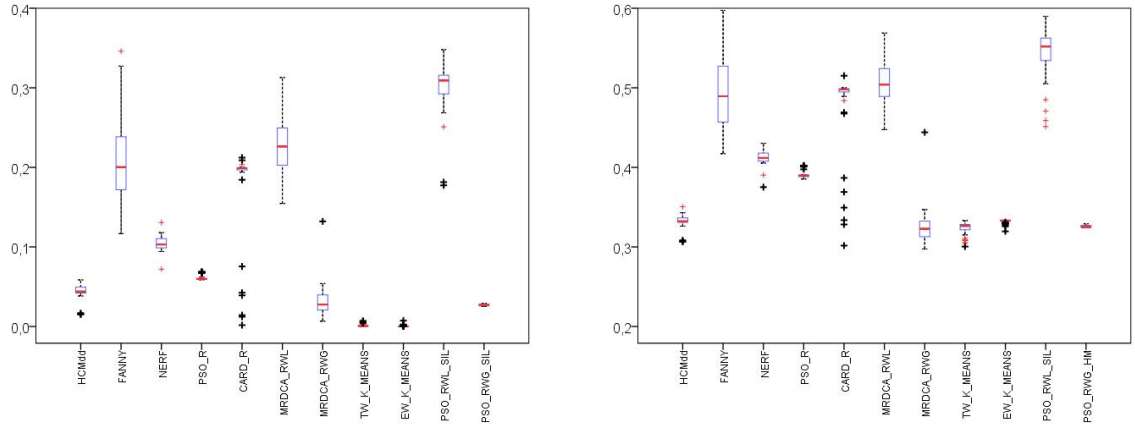
Outros pesquisas obtiveram resultados similares relacionados a avaliação de índices internos para validação de agrupamentos. Por exemplo, o índice de Xu obteve segundo lugar dentre os índices avaliados por Dimitriadou, Dolničar e Weingessel (2002), e esteve entre os três melhores índices na pesquisa feita por Raitoharju et al. (2017). Em outro trabalho da literatura, os autores Xu, Xu e Wunsch (2012) observaram que o índice da

silhueta se destacou dentre os oito índices analisados.

Em relação a comparação entre o desempenho obtido pelo método para dados com unicao visão com o desempenho obtido pelos métodos para dados com múltiplas visões, a exploração das múltiplas visões trouxe ganhos significativos para a maioria dos casos apresentados. Este fato confirma que, a depender da base de dados considerada, as visões possuem importância diferente para o agrupamento de dados e que a estimação dos pesos de relevância se torna muito importanto para o processo. Uma limitação deste trabalho foi não ter avaliado outras formas de calcular os pesos de relevância, pois isso teria provável impacto no desempenho dos métodos em dois momentos: na definição dos grupos e no avaliação da aptidão das soluções. Consequentemnte, o cálculo dos pesos influencia na busca do espaço de soluções. Outra limitação deste trabalho foi quanto a complexidade dos métodos desenvolvidos. Outras formas de atualização da posição não foram estudadas e, por isso, os métodos desenvolvidos tem complexidade $O(n^2)$. Apesar disso, outros trabalhos da literatura para dados relacionais com múltiplas visões como Frigui, Hwang e Rhee (2007) e Carvalho, Lechevallier e Melo (2012) também apresetaram métodos com a mesma complexidade.

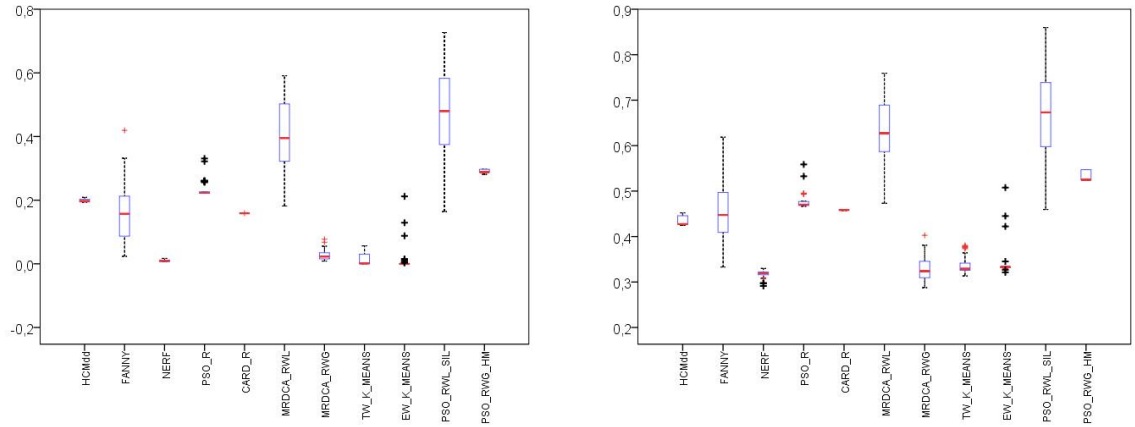
Quanto a estimação de pesos localmente ou globalmente, de forma geral, não houve grandes diferenças entre as soluções encontradas pelos métodos. Dessa forma, é possível concluir que mesmo com essa diferença, os métodos convergiram para as mesmas soluções na maior parte dos casos. Na maioria dos casos, o *gbest* dos métodos desenvolvidos convergiu dentro das 50 primeiras iterações.

Na comparação feita com outros algoritmos presentes na literatura, a abordagem desenvolvida neste trabalho obteve resultados competitivos e melhores significativamente para a maioria dos casos. Por consequência disso, essas evidências empíricas suportam, ao menos parcialmente, a tese deste trabalho que afirma que a utilização de métodos híbridos baseados em PSO melhora de forma significativa o agrupamento rígido de dados relacionais com múltiplas visões.



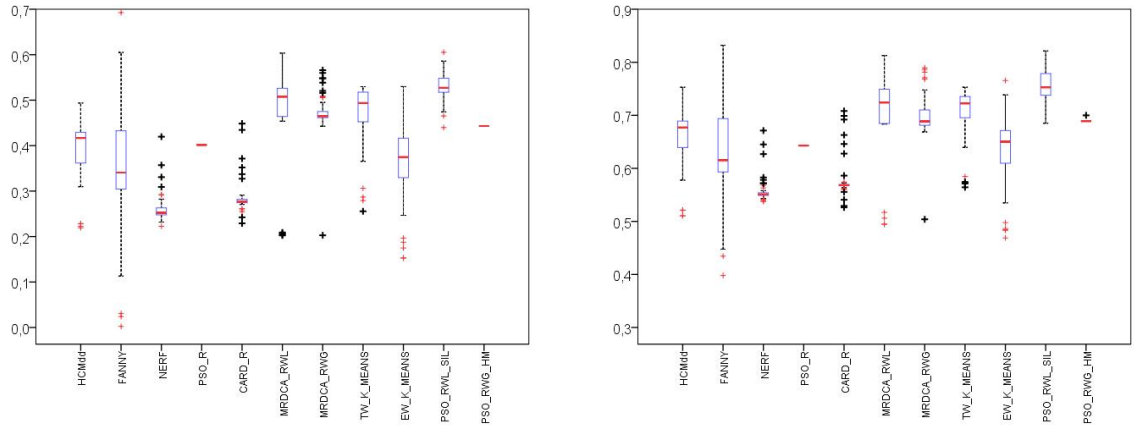
(a) ARI (b) F-measure

Figura 18 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Animals-1



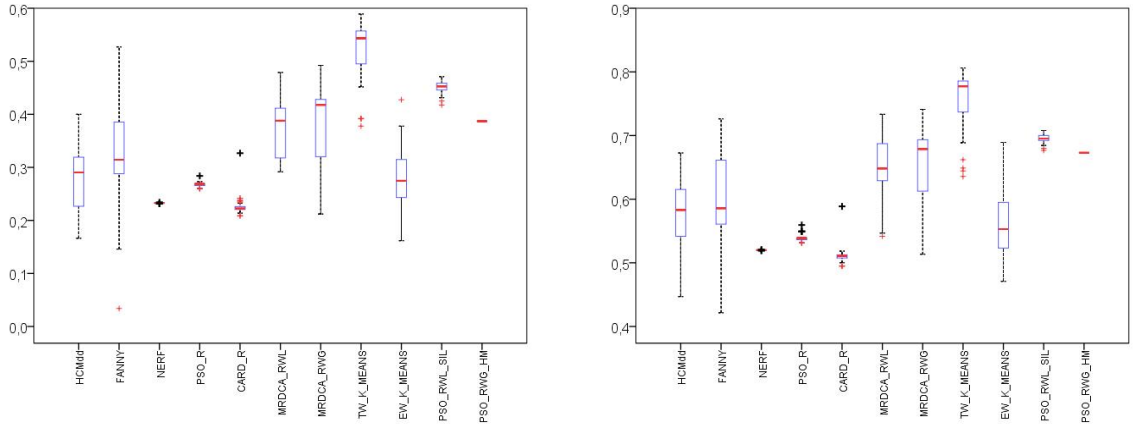
(a) ARI (b) F-measure

Figura 19 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Animals-2



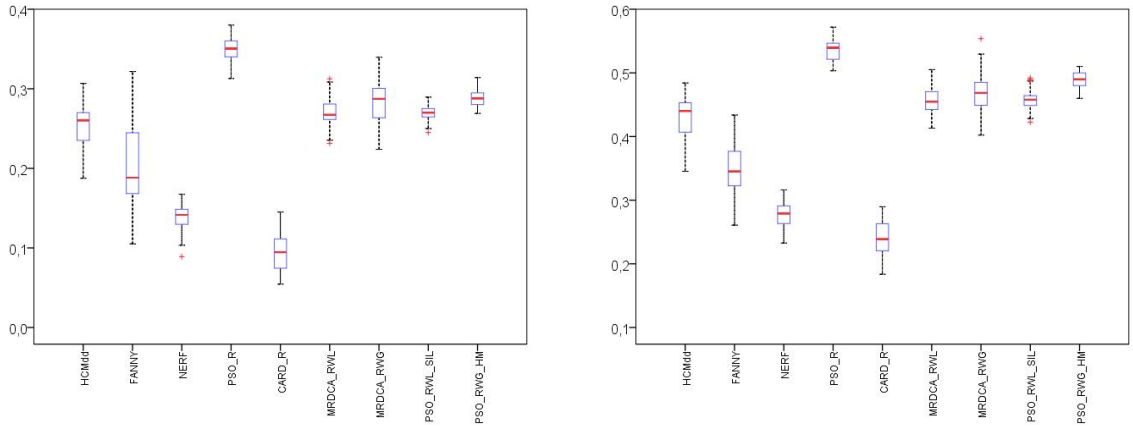
(a) ARI (b) F-measure

Figura 20 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-1



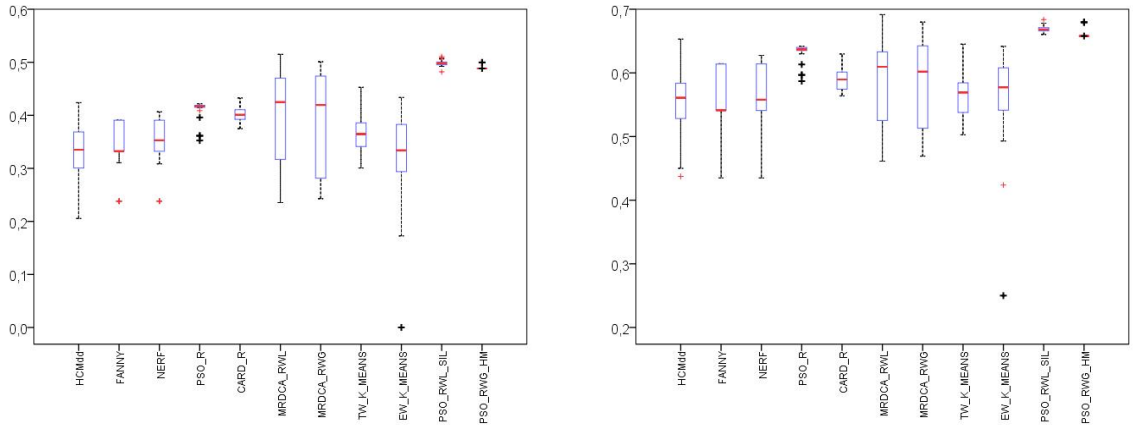
(a) ARI (b) F-measure

Figura 21 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-2



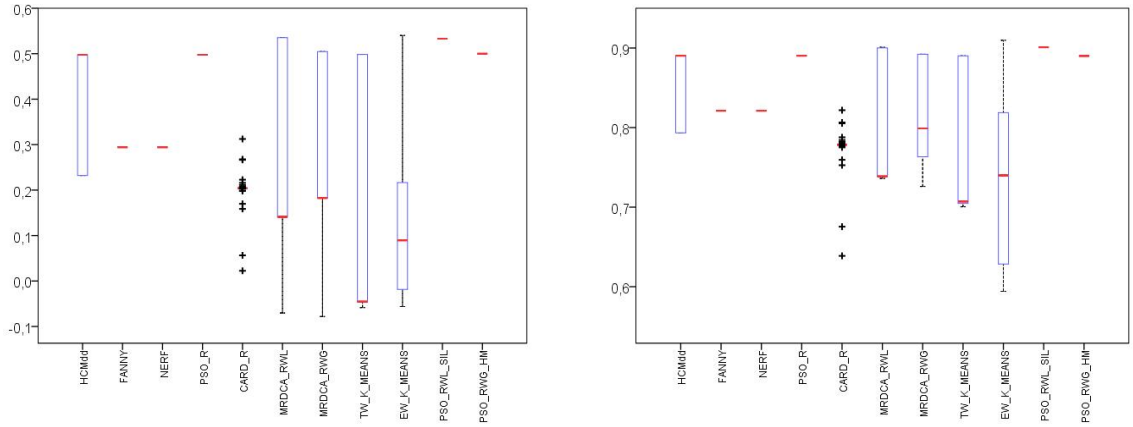
(a) ARI (b) F-measure

Figura 22 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Flowers



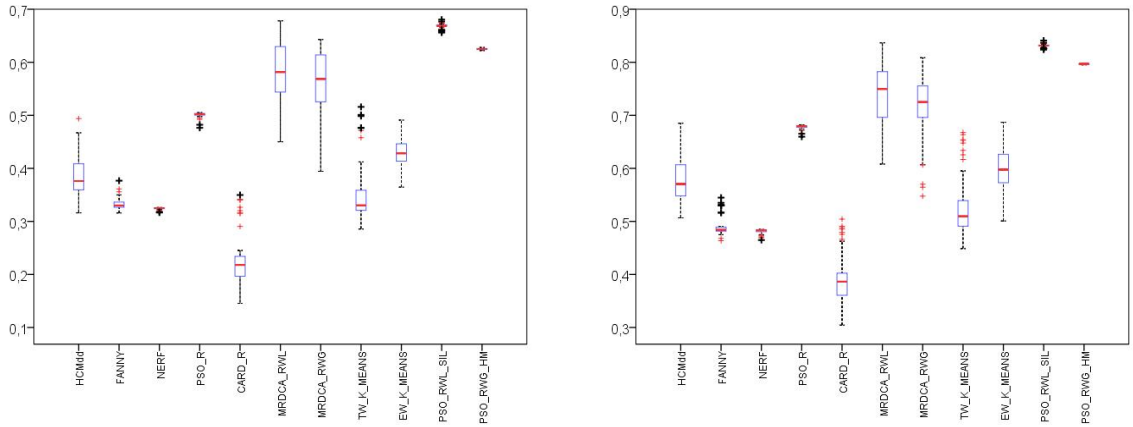
(a) ARI (b) F-measure

Figura 23 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Image



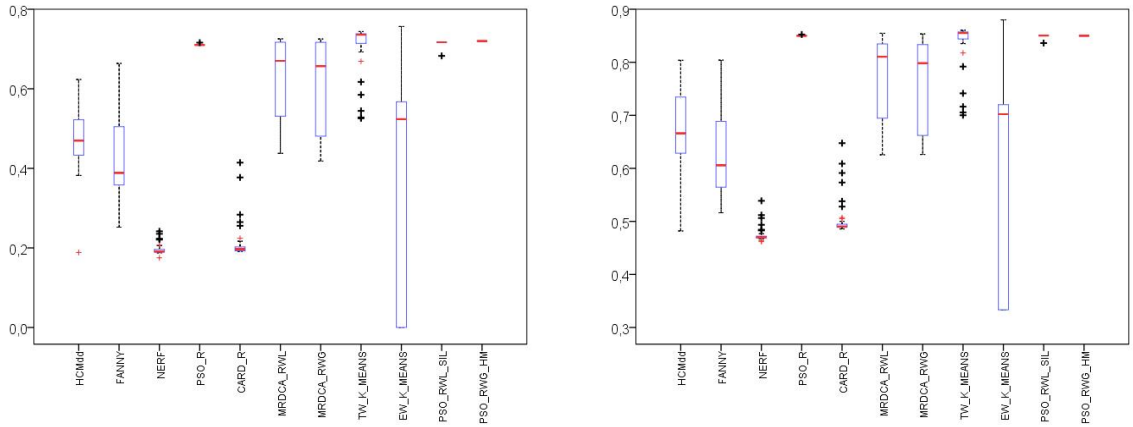
(a) ARI (b) F-measure

Figura 24 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Internet



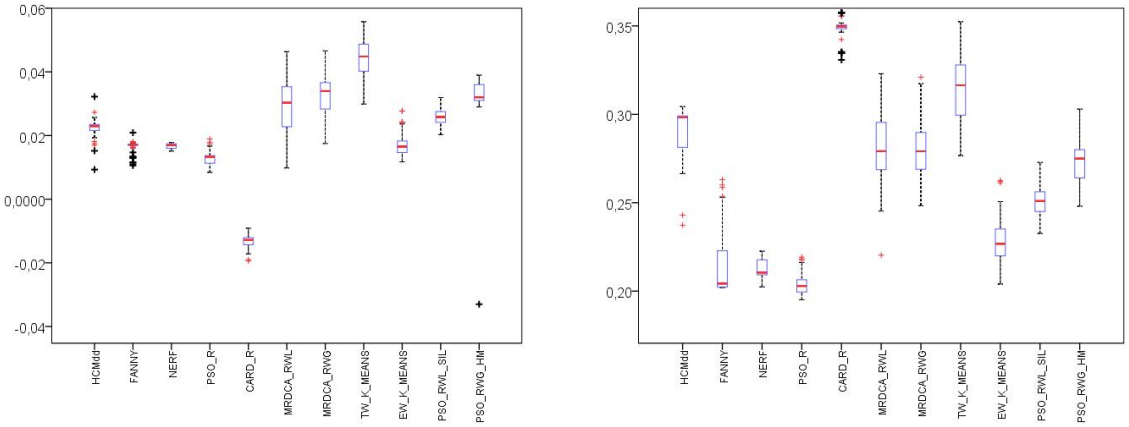
(a) ARI (b) F-measure

Figura 25 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Multiple features



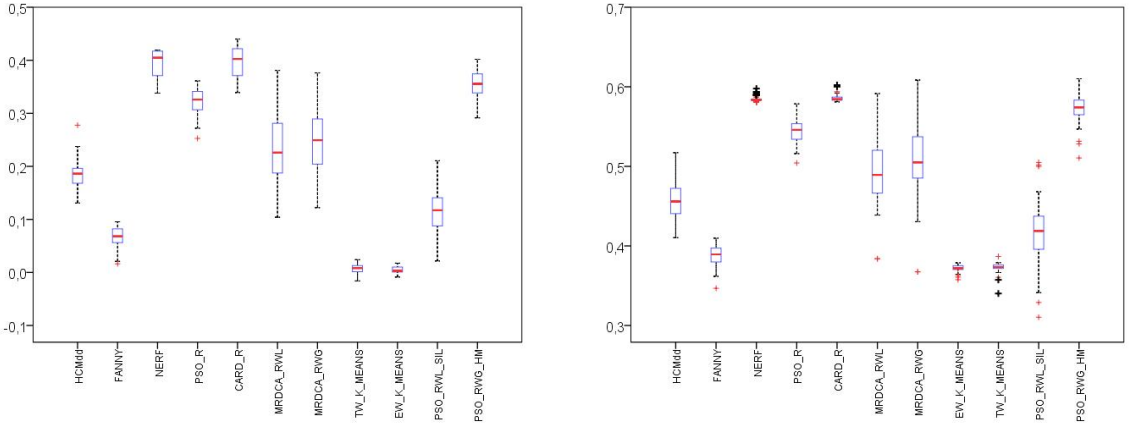
(a) ARI (b) F-measure

Figura 26 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Phoneme



(a) ARI (b) F-measure

Figura 27 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Water



(a) ARI (b) F-measure

Figura 28 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base 3-Sources

7.5 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados os resultados encontrados pelos métodos desenvolvidos para agrupamento rígido de dados relacionais. Três estudos foram realizados para avaliar esses resultados. No primeiro, o modelo para dados com visão única foi avaliado e comparado com outros três algoritmos. No segundo estudo, diversas funções de aptidão foram avaliadas e comparadas. No terceiro estudo, o desempenho dos métodos desenvolvidos foi comparado com o desempenho de outros métodos para dados com várias visões. No próximo capítulo será apresentada a validação experimental realizada para os métodos para agrupamento nebuloso.

8 RESULTADOS PARA AGRUPAMENTO NEBULOSO

Neste capítulo serão apresentados os resultados empíricos encontrados pelos métodos propostos. Primeiramente, as bases de dados usadas nos experimentos são apresentadas. Segundo, os parâmetros usados em todos os experimentos são descritos. Quarto, um estudo comparativo entre as funções de aptidão usadas é feito. Quinto, um estudo comparando os métodos propostos aos algoritmos da literatura é realizado. Por fim, uma discussão sobre os resultados coletados é feita.

8.1 BASES DE DADOS

As bases de dados apresentadas abaixo também foram usadas nos experimentos deste capítulo além das bases de dados descritas no capítulo anterior. Contudo, após a adição das novas bases e devido a quantidade de simulações, houve a necessidade de retirar duas das bases de dados descritas no capítulo anterior. Portanto, as bases Animals-1 e Water não foram consideradas nos experimentos deste capítulo.

8.1.1 Caltech

Esta base de dados[1] contém 8677 imagens agrupadas em 101 classes de animais. Um subconjunto, Caltech101-7, foi extraído do conjunto inteiro. Este subconjunto tem 1474 imagens pertencentes as categorias: faces, motorbikes, garfield, snoopy, stop-sign, windsor-hair and dolla-bill. Os atributos estão divididos em seis visões descritas na Tabela 37.

Tabela 37 – Visões do conjunto Caltech101-7

Visão	Descrição	Variáveis
Centrist	Centrist features	254
GIST	GIST features	512
LBP	LBP features	928
GABOR	GABOR features	48
HOG	HOG features	1984
WM	WM features	40

8.1.2 Corel Images

Outros três subconjuntos foram extraídos contendo quatro classes, cada uma possuindo 100 imagens. As categorias dos três subconjuntos estão descritas na Tabela 38. Os atributos são os mesmos descritos na Tabela 13 que está no capítulo anterior.

[1] http://www.vision.caltech.edu/ImageDatasets/Caltech101/

Tabela 38 – Subconjuntos extraídos da base de dados Corel

Subconjunto	Categorias			
Corel-3	owls	tigers	trains	ships
Corel-4	owls	tigers	flowers	eagle
Corel-5	eagles	elephants	cars	deer

8.2 CONFIGURAÇÕES DOS EXPERIMENTOS

Esses algoritmos de agrupamento nebuloso de dados foram aplicados as bases de dados para obter uma partição nebulosa representada por $U = (u_1, ..., u_n)$, com $u_l = (u_{l1}, ..., u_{lK})$ $(l = 1, ..., n)$. Depois, uma partição rígida $Q = (Q_1, ..., Q_K)$ é obtida a partir da partição nebulosa, onde a partição rígida $Q_k (k = 1, ..., K)$ é definida como $Q_k = \{e_l : u_{ik} \geq u_{im} \forall m \in \{1, ..., K\}\}$. As partições rígidas obtidas desses algoritmos são comparadas com as partições a priori conhecidas. Todos os algoritmos foram implementados na linguagem C e foram executados em um cluster computacional contendo 22 nós (OS: Red Hat Linux 6.4, Memória: 64 GB, Processador: Intel Xeon Ten-Core E5-2660v2 2.2 GHz).

Os métodos foram executados 30 vezes para cada função de aptidão e o número máximo de iterações foi fixado em 100 para cada execução. Os índices externos foram computados para cada partição final fornecida em cada execução. O parâmetro $|G_k|$ foi definido com valor 5. O número de partículas usado foi 30. O peso de inércia, $w(t)$, foi decrescido linearmente de 0.9 até 0.4. Os pesos de aceleração c_1 e c_2 foram definidos com valor 2 como sugerido em (KENNEDY; EBERHART, 1995; SHI; EBERHART, 1998) e amplamente usados desde então (JORDEHI; JASNI, 2013). Os componentes de velocidade foram inicializados com valores aleatórios entre 0 e 1. O parâmetro m foi definido com valor igual a 2. A mudança quanto ao número de execuções e número de partículas foi devido a limitações de recursos computacionais bem como pelo fato dos métodos nebulosos serem mais lentos devido ao passo adicional para computar os graus de pertinência.

8.3 RESULTADOS DOS ALGORITMOS DE AGRUPAMENTO NEBULOSO

Nesta seção, três estudos são apresentados. No primeiro estudo, o método $FPSO - SV$ é comparado com outros três algoritmos para agrupamento nebuloso de dados. No segundo estudo, diversas funções de aptidão para agrupamento nebuloso são consideradas e avaliadas. No terceiro, os resultados de uma função selecionada são comparados com os resultados obtidos por outros métodos para agrupamento de dados com múltiplas visões.

8.3.1 Estudo 4 - Abordagem para dados com uma visão

O objetivo desse estudo foi avaliar o desempenho da abordagem híbrida para o agrupamento nebuloso de dados relacionais com uma visão em comparação com outros métodos

da literatura. Os métodos utilizados na comparação foram: SFCMdd (CARVALHO; LECHEVALLIER; MELO, 2013), NERF (HATHAWAY; BEZDEK, 1994) e FANNY (KAUFMAN; ROUSSEEUW, 1990). A Tabela 39 apresenta os resultados obtidos pelos algoritmos. A média e o desvio padrão (parenteses) são apresentados. O desempenho médio de cada método recebeu uma classificação de acordo com o valor obtido em cada índice.

Tabela 39 – Resultados encontrados pelos métodos para agrupamento de dados com única visão

Data set	Index	$SFCMdd$	$NERF$	$FANNY$	$FPSO-SV$
Animals	ARI	0.00(.00) (4)	0.17(.07) (2,5)	0.17(.07) (2,5)	0.21(.02) (1)
	F-measure	0.32(.00) (4)	0.46(.05) (1,5)	0.46(.05) (1,5)	0.42(.01) (3)
Caltech101-7	ARI	0.28(.02) (3)	0.66(.09) (1)	−0.07(.03) (4)	0.389(.04) (2)
	F-measure	0.63(.01) (3)	0.78(.04) (1)	0.46(.02) (4)	0.69(.02) (2)
Corel-1	ARI	0.26(.08) (3,5)	0.26(.01) (3,5)	0.42(.12) (1)	0.335(.05) (2)
	F-measure	0.55(.07) (3,5)	0.55(.01) (3,5)	0.68(.08) (1)	0.62(.04) (2)
Corel-2	ARI	0.29(.09) (3)	0.23(.00) (4)	0.33(.08) (2)	0.406(.02) (1)
	F-measure	0.55(.07) (3)	0.52(.00) (4)	0.60(.05) (2)	0.65(.02) (1)
Corel-3	ARI	0.29(.05) (3)	0.23(.00) (4)	0.33(.00) (2)	0.466(.03) (1)
	F-measure	0.55(.05) (3)	0.52(.00) (4)	0.60(.00) (2)	0.72(.03) (1)
Corel-4	ARI	0.29(.04 (3)	0.23(.00) (4)	0.33(.02) (1)	0.31(.01) (2)
	F-measure	0.55(.04 (2)	0.52(.00) (4)	0.60(.02) (1)	0.57(.01) (2)
Corel-5	ARI	0.29(.03) (2)	0.23(.01) (3)	0.33(.02) (1)	0.119(.03) (4)
	F-measure	0.55(.05) (2)	0.52(.01) (3)	0.60(.01) (1)	0.49(.03) (4)
Flowers	ARI	0.10(.01) (3)	0.14(.02) (2)	0.21(.05) (1)	0.065(.01) (4)
	F-measure	0.27(.02) (2)	0.28(.02) (3)	0.36(.04) (1)	0.22(.01) (4)
Image	ARI	0.21(.04) (3,5)	0.35(.05) (1,5)	0.35(.03) (1,5)	0.210(.01) (3,5)
	F-measure	0.43(.04) (3)	0.56(.05) (2)	0.57(.04) (1)	0.42(.01) (4)
Internet	ARI	0.12(.00) (4)	0.29(.00) (1,5)	0.29(.00) (1,5)	0.131(.00) (3)
	F-measure	0.71(.00) (4)	0.82(.00) (1,5)	0.82(.00) (1,5)	0.73(.00) (3)
Mfeat	ARI	0.37(.03) (1)	0.32(.00) (3)	0.34(.01) (2)	0.305(.03)(4)
	F-measure	0.53(.03) (1)	0.48(.01) (4)	0.50(.02) (2)	0.49(.03) (3)
Phoneme	ARI	0.63(.06) (1)	0.20(.05) (4)	0.44(.09) (3)	0.563(.04)(2)
	F-measure	0.78(.04) (1)	0.48(.03) (4)	0.63(.07) (3)	0.74(.03) (2)
3-Sources	ARI	0.28(.06) (2)	0.39(.03) (1)	0.06(.02) (4)	0.099(.04) (3)
	F-measure	0.51(.04) (2)	0.58(.00) (1)	0.38(.02) (4)	0.47(.03)(3)

A Tabela 40 apresenta a classificação média obtida por cada método para cada índice. O FPSO-SV ficou em segundo lugar para os dois índices. Em relação ao que foi apresentado no Estudo 1 pelo modelo para agrupamento rígido, o FPSO-SV mostrou desempenho um pouco inferior levando-se em consideração os dois índices externos avaliados. Além disso, diferentes parametrizações poderiam melhorar o desempenho do método.

Tabela 40 – Ranks médios.

	Médias	
Algoritmos	F-measure	ARI
SFCMdd	2,65 (3)	2,77 (4)
NERF	2,81 (4)	2,68 (3)
FANNY	1,92 (1)	2,04 (1)
FPSO-SV	2,62 (2)	2,50 (2)

8.3.2 Estudo 5 - Avaliação das funções de aptidão

Este estudo teve como objetivo avaliar e comparar o desempenho das várias funções de aptidão consideradas nesta tese para agrupamento nebuloso de dados relacionais. Neste estudo, as funções de aptidão avaliadas foram: *Average Within-Cluster Distance*, Homogeneidade, índice da Silhueta, Silhueta simplificada, índice da silhueta nebulosa, silhueta nebulosa simplificada, o índice de Xie-Beni, o índice de Fukuyama-Sugeno, Entropia da partição e Coeficiente da partição. Além disso, também teve como objetivo comparar os três métodos híbridos para cada função de aptidão. As Tabelas 41 - 50 apresentam os resultados obtidos por cada método híbrido para cada função de aptidão separadamente. A partir desses resultados, é possível investigar respostas para as três perguntas de pesquisa abaixo:

1. *Considerando os métodos híbridos desenvolvidos para agrupamento nebuloso, explorar múltiplas visões de dados relacionais produz melhores agrupamentos?*

2. *Dentre as diversas funções de aptidão para agrupamento nebuloso consideradas neste trabalho, é possível afirmar que alguma delas produz melhores agrupamentos?*

3. *O uso de pesos de relevância estimados localmente nos métodos desenvolvidos é capaz de gerar melhores agrupamentos do que o uso de pesos de relevância estimados globalmente?*

Tabela 41 – Resultados relativos a função AWCD

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,17(.02)	**0,40(.03)**	0,07(.01)	**0,40(.04)**	**0,37(.05)**	**0,47(.02)**	0,29(.04)	0,12(.04)	0,23(.02)	**0,13(.00)**	**0,36(.03)**	0,55(.04)	0,08(.03)
	F-measure	0,39(.01)	**0,69(.02)**	0,23(.01)	**0,66(.03)**	**0,63(.04)**	**0,73(.01)**	0,60(.03)	**0,48(.03)**	0,44(.02)	**0,73(.00)**	**0,53(.03)**	0,72(.04)	0,38(.04)
FPSO-RWL	ARI	**0,22(.00)**	0,25(.04)	**0,11(.01)**	0,28(.05)	0,36(.04)	0,43(.05)	**0,48(.02)**	**0,15(.02)**	0,26(.04)	0,08(.00)	0,34(.03)	**0,64(.00)**	0,16(.06)
	F-measure	**0,47(.00)**	0,59(.04)	**0,27(.01)**	0,59(.05)	**0,63(.02)**	0,66(.06)	**0,72(.02)**	0,46(.02)	0,46(.04)	0,69(.00)	0,51(.03)	**0,80(.00)**	0,44(.05)
FPSO-RWG	ARI	**0,22(.00)**	0,27(.06)	**0,11(.01)**	0,28(.04)	0,32(.04)	0,44(.05)	**0,48(.01)**	0,13(.02)	**0,27(.03)**	0,07(.01)	0,34(.03)	0,33(.08)	**0,17(.04)**
	F-measure	**0,47(.00)**	0,60(.06)	**0,27(.02)**	0,58(.03)	0,60(.03)	0,68(.05)	**0,72(.01)**	0,45(.02)	**0,47(.02)**	0,68(.01)	0,51(.02)	0,58(.06)	**0,45(.04)**

Ao analisar as Tabelas, é possível constatar que os modelos para dados com várias visões apresentaram melhores médias dos índices considerando todas as funções ou quase todas para seis bases: Animals-2, Flowers, Corel-5, Image, Phoneme e 3-Sources. Por outro lado, o modelo para dados com única visão apresentou melhores médias para as

Tabela 42 – Resultados relativos a função CS

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,23(.01)	**0,22(.06)**	0,07(.01)	**0,26(.05)**	0,25(.06)	**0,37(.03)**	0,25(.02)	**0,16(.05)**	**0,30(.04)**	**0,17(.00)**	**0,36(.03)**	0,40(.10)	0,08(.03)
	F-measure	0,42(.01)	0,53(.06)	0,23(.01)	**0,57(.04)**	0,54(.05)	**0,65(.02)**	0,55(.03)	**0,51(.05)**	**0,51(.04)**	**0,75(.00)**	**0,51(.03)**	0,63(.07)	0,37(.03)
FPSO-RWL	ARI	**0,26(.02)**	0,19(.08)	**0,11(.01)**	0,25(.06)	0,41(.07)	0,34(.08)	**0,36(.06)**	0,09(.03)	0,28(.08)	0,08(.01)	0,34(.03)	0,60(.02)	**0,13(.03)**
	F-measure	0,47(.03)	0,48(.06)	**0,28(.02)**	0,56(.05)	**0,67(.05)**	0,63(.06)	**0,66(.06)**	0,42(.03)	0,50(.06)	0,69(.01)	**0,51(.03)**	0,78(.01)	**0,41(.04)**
FPSO-RWG	ARI	0,24(.01)	**0,22(.06)**	**0,11(.01)**	0,23(.05)	**0,43(.09)**	0,35(.08)	0,33(.07)	0,10(.04)	0,26(.04)	0,08(.00)	0,29(.03)	**0,61(.02)**	0,11(.04)
	F-measure	**0,49(.01)**	**0,54(.06)**	**0,28(.02)**	0,54(.06)	**0,67(.06)**	0,64(.07)	0,63(.07)	0,44(.04)	0,47(.03)	0,69(.00)	0,48(.03)	**0,79(.01)**	**0,41(.04)**

Tabela 43 – Resultados relativos a função FCS

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,22(.01)	0,23(.05)	0,07(.01)	0,24(.07)	0,23(.06)	**0,36(.02)**	0,25(.03)	**0,17(.04)**	0,22(.05)	**0,12(.04)**	0,35(.03)	0,40(.10)	0,08(.03)
	F-measure	0,42(.01)	0,55(.06)	0,23(.01)	0,53(.06)	0,53(.05)	0,64(.02)	0,55(.04)	**0,53(.03)**	0,44(.04)	**0,72(.03)**	0,49(.04)	0,62(.08)	0,37(.03)
FPSO-RWL	ARI	0,23(.02)	**0,34(.09)**	**0,11(.02)**	**0,26(.07)**	0,36(.07)	0,35(.06)	**0,35(.07)**	0,09(.04)	0,25(.05)	0,06(.01)	0,34(.03)	0,58(.06)	**0,12(.03)**
	F-measure	0,48(.02)	**0,63(.06)**	**0,28(.02)**	0,56(.07)	0,62(.05)	0,63(.04)	**0,66(.06)**	0,44(.05)	**0,48(.05)**	0,67(.01)	0,51(.03)	0,76(.04)	0,41(.04)
FPSO-RWG	ARI	**0,24(.01)**	0,21(.07)	0,10(.01)	0,25(.08)	**0,41(.09)**	**0,36(.06)**	0,33(.08)	0,10(.05)	**0,26(.06)**	0,06(.01)	0,28(.02)	**0,61(.04)**	**0,12(.03)**
	F-measure	**0,49(.01)**	0,53(.06)	0,27(.02)	**0,57(.07)**	**0,65(.06)**	**0,65(.05)**	0,63(.06)	0,44(.05)	0,47(.05)	0,67(.01)	0,46(.03)	**0,78(.02)**	**0,42(.03)**

Tabela 44 – Resultados relativos a função FS

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,19(.03)	**0,39(.06)**	0,07(.01)	**0,27(.08)**	**0,41(.03)**	**0,47(.03)**	0,25(.04)	**0,15(.05)**	0,23(.02)	**0,27(.05)**	**0,38(.04)**	0,38(.04)	0,09(.03)
	F-measure	0,41(.01)	**0,68(.03)**	0,23(.01)	**0,57(.06)**	**0,65(.03)**	**0,73(.03)**	0,56(.04)	**0,50(.03)**	0,44(.02)	**0,81(.02)**	**0,53(.04)**	0,60(.03)	0,40(.03)
FPSO-RWL	ARI	0,21(.00)	0,31(.09)	**0,11(.01)**	0,22(.04)	0,34(.05)	0,41(.03)	**0,47(.04)**	0,11(.04)	**0,27(.04)**	0,07(.01)	0,33(.03)	**0,59(.08)**	0,21(.05)
	F-measure	**0,47(.01)**	0,63(.06)	**0,27(.02)**	0,53(.04)	0,62(.04)	0,63(.04)	**0,71(.03)**	0,44(.03)	**0,46(.04)**	0,68(.01)	0,50(.02)	**0,77(.05)**	0,48(.04)
FPSO-RWG	ARI	**0,22(.00)**	0,28(.05)	0,10(.01)	0,26(.05)	0,36(.04)	0,40(.04)	**0,47(.02)**	0,11(.04)	0,25(.04)	0,07(.01)	0,33(.03)	0,31(.10)	**0,23(.06)**
	F-measure	**0,47(.00)**	0,61(.04)	**0,27(.02)**	0,56(.05)	0,62(.04)	0,62(.04)	**0,71(.02)**	0,44(.04)	0,45(.04)	0,68(.01)	0,49(.03)	0,57(.07)	**0,50(.04)**

Tabela 45 – Resultados relativos a função FSS

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	**0,22(.01)**	0,33(.06)	0,07(.01)	**0,34(.06)**	0,29(.03)	0,28(.03)	0,26(.06)	**0,18(.04)**	0,24(.06)	**0,44(.01)**	0,34(.02)	0,31(.08)	0,11(.04)
	F-measure	0,45(.01)	0,64(.05)	0,23(.01)	**0,61(.04)**	0,54(.04)	0,56(.02)	0,57(.05)	**0,52(.03)**	0,45(.05)	**0,87(.00)**	0,47(.02)	0,56(.06)	0,40(.03)
FPSO-RWL	ARI	0,15(.09)	**0,43(.06)**	**0,11(.02)**	0,27(.07)	**0,35(.05)**	0,36(.05)	0,37(.09)	0,15(.03)	0,23(.03)	0,08(.03)	**0,34(.03)**	**0,51(.09)**	**0,31(.03)**
	F-measure	0,45(.07)	**0,70(.04)**	**0,27(.02)**	0,57(.06)	**0,61(.04)**	**0,63(.04)**	**0,66(.07)**	0,48(.03)	0,44(.03)	0,69(.01)	**0,51(.03)**	**0,70(.07)**	**0,54(.02)**
FPSO-RWG	ARI	**0,22(.01)**	0,33(.05)	**0,11(.02)**	0,28(.09)	0,29(.05)	**0,37(.05)**	**0,38(.09)**	0,16(.02)	**0,27(.03)**	0,09(.01)	0,35(.04)	0,41(.11)	0,31(.05)
	F-measure	**0,49(.02)**	0,67(.03)	0,26(.02)	0,57(.07)	0,57(.04)	0,62(.04)	**0,66(.06)**	0,49(.02)	**0,47(.04)**	0,70(.01)	0,49(.04)	0,63(.08)	0,53(.03)

Tabela 46 – Resultados relativos a função HM

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,21(.02)	**0,39(.04)**	0,06(.01)	**0,34(.05)**	**0,41(.02)**	**0,47(.03)**	0,24(.01)	**0,12(.03)**	0,21(.01)	**0,13(.00)**	**0,36(.03)**	0,56(.02)	0,10(.04)
	F-measure	0,42(.01)	**0,69(.02)**	0,22(.02)	**0,62(.04)**	**0,65(.02)**	**0,72(.03)**	0,54(.01)	**0,49(.03)**	0,43(.02)	**0,73(.00)**	**0,52(.03)**	0,74(.01)	0,40(.04)
FPSO-RWL	ARI	**0,22(.00)**	0,26(.06)	**0,11(.01)**	0,25(.05)	0,34(.05)	0,40(.01)	0,45(.04)	0,10(.02)	**0,29(.05)**	0,06(.01)	0,34(.03)	**0,62(.01)**	**0,20(.04)**
	F-measure	**0,48(.00)**	0,61(.05)	**0,27(.01)**	0,55(.06)	0,62(.04)	0,61(.00)	0,69(.04)	0,44(.01)	**0,50(.04)**	0,68(.01)	0,51(.03)	**0,79(.01)**	**0,47(.03)**
FPSO-RWG	ARI	0,21(.00)	0,28(.06)	**0,11(.01)**	0,28(.04)	0,37(.04)	0,40(.01)	**0,48(.01)**	0,12(.02)	0,26(.03)	0,08(.01)	0,31(.03)	0,31(.08)	**0,20(.04)**
	F-measure	0,47(.00)	0,61(.05)	**0,27(.02)**	0,59(.04)	0,64(.04)	0,61(.00)	**0,72(.01)**	0,45(.01)	0,46(.03)	0,69(.01)	0,48(.04)	0,57(.06)	**0,47(.03)**

Tabela 47 – Resultados relativos a função PC

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,22(.02)	**0,36(.07)**	0,07(.01)	**0,38(.05)**	**0,34(.04)**	**0,46(.03)**	0,28(.04)	0,11(.03)	0,23(.02)	**0,43(.03)**	**0,39(.03)**	**0,54(.07)**	0,11(.04)
	F-measure	0,43(.02)	0,66(.05)	0,23(.02)	**0,64(.04)**	**0,62(.03)**	**0,72(.02)**	0,60(.03)	0,48(.02)	0,45(.02)	**0,87(.01)**	**0,55(.03)**	0,71(.05)	0,41(.04)
FPSO-RWL	ARI	**0,21(.00)**	0,29(.04)	**0,12(.02)**	0,32(.06)	0,31(.05)	0,43(.06)	0,43(.07)	0,16(.04)	0,27(.04)	0,08(.01)	0,32(.05)	0,52(.03)	0,20(.04)
	F-measure	**0,47(.00)**	0,64(.03)	**0,29(.02)**	0,62(.05)	0,58(.03)	0,68(.05)	0,69(.04)	0,48(.04)	0,47(.04)	0,69(.01)	0,49(.05)	0,71(.02)	0,47(.03)
FPSO-RWG	ARI	**0,21(.00)**	**0,36(.04)**	**0,12(.01)**	0,29(.07)	0,33(.06)	0,42(.05)	**0,48(.02)**	**0,18(.04)**	**0,27(.04)**	0,08(.01)	0,33(.05)	**0,54(.03)**	**0,21(.04)**
	F-measure	**0,47(.00)**	**0,67(.02)**	0,28(.02)	0,59(.06)	0,60(.04)	0,66(.05)	**0,72(.01)**	**0,50(.04)**	**0,48(.04)**	0,69(.01)	0,50(.04)	**0,74(.02)**	**0,48(.04)**

Tabela 48 – Resultados relativos a função PE

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,21(.02)	0,35(.06)	0,07(.01)	**0,36(.08)**	**0,35(.05)**	**0,47(.02)**	0,28(.03)	0,11(.02)	0,23(.03)	**0,44(.00)**	**0,36(.03)**	0,53(.04)	0,08(.03)
	F-measure	0,42(.02)	0,65(.05)	0,22(.01)	**0,63(.06)**	**0,62(.03)**	**0,72(.01)**	0,60(.03)	0,47(.02)	0,44(.03)	**0,88(.00)**	**0,52(.04)**	**0,70(.04)**	0,39(.04)
FPSO-RWL	ARI	**0,22(.00)**	0,28(.06)	**0,12(.02)**	0,30(.08)	0,30(.04)	0,43(.05)	**0,48(.02)**	**0,17(.04)**	**0,27(.04)**	0,08(.01)	0,32(.04)	**0,59(.03)**	**0,20(.04)**
	F-measure	**0,48(.00)**	0,62(.06)	**0,29(.02)**	0,60(.07)	0,58(.03)	0,67(.06)	**0,72(.02)**	**0,49(.04)**	**0,48(.04)**	0,69(.01)	0,49(.04)	**0,77(.02)**	**0,48(.03)**
FPSO-RWG	ARI	0,21(.00)	**0,39(.03)**	**0,12(.02)**	0,31(.07)	0,33(.04)	0,40(.05)	0,47(.05)	**0,17(.04)**	**0,27(.05)**	0,08(.01)	0,32(.04)	0,55(.03)	0,19(.04)
	F-measure	0,47(.00)	**0,69(.01)**	0,28(.02)	0,61(.06)	0,60(.03)	0,65(.04)	0,71(.03)	**0,49(.04)**	0,47(.05)	0,69(.01)	0,49(.04)	0,74(.02)	0,47(.04)

bases Corel-1, Corel-2 e Internet. Para as bases Corel-3 e Corel-4, o método FPSO-SV apresentrou melhores médias em seis funções contra três funções dos outros modelos. Para

Tabela 49 – Resultados relativos a função SS

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,22(.00)	0,36(.04)	0,07(.01)	0,40(.04)	**0,37(.05)**	0,41(.05)	0,25(.02)	**0,18(.02)**	0,22(.02)	**0,30(.08)**	**0,39(.03)**	0,48(.09)	0,14(.05)
	F-measure	0,45(.01)	0,68(.02)	0,23(.01)	0,65(.02)	0,62(.05)	0,66(.06)	0,57(.03)	**0,52(.02)**	0,43(.02)	**0,82(.03)**	**0,53(.03)**	0,68(.07)	0,43(.04)
FPSO-RWL	ARI	0,19(.05)	**0,45(.03)**	**0,11(.01)**	**0,33(.04)**	**0,37(.10)**	**0,44(.05)**	**0,48(.01)**	0,17(.04)	**0,28(.04)**	0,08(.01)	0,34(.03)	**0,59(.04)**	**0,35(.06)**
	F-measure	0,48(.04)	**0,72(.01)**	**0,27(.02)**	**0,63(.03)**	**0,63(.07)**	**0,67(.05)**	**0,72(.01)**	0,50(.03)	**0,48(.04)**	0,69(.01)	0,51(.03)	0,76(.03)	**0,56(.03)**
FPSO-RWG	ARI	**0,25(.00)**	0,33(.05)	**0,11(.01)**	**0,33(.05)**	0,33(.06)	**0,44(.05)**	**0,48(.02)**	0,17(.04)	**0,28(.05)**	0,08(.01)	0,38(.03)	**0,59(.03)**	0,33(.06)
	F-measure	**0,52(.00)**	0,66(.03)	0,26(.01)	0,62(.03)	0,60(.05)	0,66(.05)	**0,72(.02)**	0,50(.04)	**0,48(.05)**	0,69(.01)	**0,53(.03)**	**0,77(.01)**	0,55(.04)

Tabela 50 – Resultados relativos a função XB

	Index	AwA2	CAL7	FLOW	C1	C2	C3	C4	C5	IMG	INT	MFT	PHO	3-S
FPSO-SV	ARI	0,20(.02)	0,19(.06)	0,07(.01)	**0,30(.07)**	**0,36(.10)**	0,33(.04)	0,26(.05)	0,10(.02)	0,24(.03)	**0,25(.01)**	**0,39(.03)**	0,37(.03)	0,08(.03)
	F-measure	0,45(.01)	0,55(.04)	0,22(.01)	**0,59(.05)**	**0,61(.07)**	0,58(.02)	0,58(.05)	**0,46(.02)**	**0,46(.03)**	**0,80(.01)**	0,52(.04)	0,60(.02)	0,39(.03)
FPSO-RWL	ARI	0,21(.00)	0,32(.12)	**0,12(.02)**	0,27(.06)	0,30(.07)	**0,41(.04)**	**0,43(.07)**	0,10(.03)	0,19(.07)	0,13(.16)	0,34(.03)	0,34(.11)	0,18(.07)
	F-measure	**0,49(.00)**	0,61(.07)	**0,32(.02)**	0,58(.05)	0,57(.06)	**0,63(.04)**	**0,70(.04)**	0,43(.03)	0,42(.06)	0,75(.08)	0,51(.03)	0,59(.07)	0,46(.06)
FPSO-RWG	ARI	**0,22(.02)**	**0,32(.07)**	0,11(.02)	0,26(.04)	0,37(.10)	0,39(.04)	**0,43(.06)**	**0,11(.03)**	**0,25(.02)**	0,19(.20)	0,38(.03)	**0,46(.14)**	**0,23(.07)**
	F-measure	**0,49(.02)**	**0,65(.05)**	0,28(.02)	0,57(.04)	0,61(.07)	0,62(.03)	**0,70(.03)**	0,45(.03)	0,45(.02)	0,77(.09)	**0,53(.03)**	**0,68(.09)**	**0,50(.05)**

Tabela 51 – Resultados do ARI pelos métodos híbridos com diferentes funções de aptidão

Base de dados	Algoritmo	AWCD	CS	FCS	FS	FSS	HM	PC	PE	SS	XB
Animals	FPSO-SV	0.169(.02)	**0.225**(.01)	**0.225**(.01)	0.190(.03)	0.220(.01)	0.210(.02)	0.216(.02)	0.213(.02)	0.224(.00)	0.204(.01)
	FPSO-RWL	0.225(.00)	**0.256**(.01)	0.235(.02)	0.211(.00)	0.153(.00)	0.217(.00)	0.213(.00)	0.215(.00)	0.194(.00)	0.207(.00)
	FPSO-RWG	0.215(.01)	0.259(.03)	**0.260**(.04)	0.225(.01)	0.180(.07)	0.216(.00)	0.201(.05)	0.174(.09)	0.188(.06)	0.220(.03)
Caltech101-7	FPSO-SV	**0.402**(.03)	0.216(.06)	0.225(.05)	0.385(.06)	0.328(.06)	0.389(.04)	0.357(.07)	0.356(.06)	0.364(.04)	0.193(.06)
	FPSO-RWL	0.253(.04)	0.340(.07)	0.340(.09)	0.306(.09)	0.427(.06)	0.265(.06)	0.293(.04)	0.282(.06)	**0.447**(.03)	0.321(.12)
	FPSO-RWG	0.268(.01)	0.221(.08)	0.210(.07)	0.276(.08)	0.334(.08)	0.284(.02)	0.359(.03)	**0.385**(.04)	0.326(.05)	0.323(.08)
Corel-1	FPSO-SV	**0.402**(.04)	0.264(.05)	0.237(.07)	0.273(.08)	0.337(.05)	0.335(.05)	0.383(.05)	0.365(.08)	0.400(.03)	0.296(.07)
	FPSO-RWL	0.284(.08)	0.236(.06)	0.257(.08)	0.224(.04)	0.270(.05)	0.248(.04)	0.318(.04)	0.296(.08)	**0.329**(.06)	0.270(.05)
	FPSO-RWG	0.276(.04)	0.226(.05)	0.256(.08)	0.260(.05)	0.276(.08)	0.277(.03)	0.289(.07)	0.309(.07)	**0.329**(.05)	0.257(.03)
Corel-2	FPSO-SV	0.374(.05)	0.246(.06)	0.227(.06)	**0.410**(.03)	0.290(.03)	0.406(.02)	0.344(.04)	0.349(.04)	0.371(.05)	0.361(.10)
	FPSO-RWL	0.366(.03)	0.324(.07)	**0.430(.07)**	0.355(.05)	0.414(.05)	0.289(.05)	0.307(.05)	0.299(.04)	0.373(.10)	0.302(.07)
	FPSO-RWG	0.358(.04)	**0.413**(.09)	0.345(.08)	0.356(.04)	0.354(.05)	0.340(.04)	0.328(.05)	0.331(.04)	0.328(.06)	0.368(.10)
Corel-3	FPSO-SV	**0.473**(.02)	0.371(.03)	0.362(.02)	0.466(.03)	0.280(.03)	0.466(.03)	0.461(.03)	0.468(.02)	0.413(.05)	0.335(.04)
	FPSO-RWL	0.426(.05)	0.341(.08)	0.352(.06)	0.413(.03)	0.356(.05)	0.397(.01)	0.430(.06)	0.428(.05)	**0.441**(.05)	0.408(.04)
	FPSO-RWG	0.**437**(.05)	0.354(.08)	0.363(.06)	0.399(.04)	0.368(.05)	0.396(.01)	0.418(.05)	0.401(.04)	0.435(.05)	0.389(.04)
Corel-4	FPSO-SV	**0.291**(.04)	0.248(.02)	0.245(.03)	0.254(.04)	0.256(.06)	0.241(.01)	0.282(.04)	0.279(.03)	0.253(.02)	0.263(.05)
	FPSO-RWL	0.478(.02)	0.364(.06)	0.352(.07)	0.468(.04)	0.367(.09)	0.451(.04)	0.430(.07)	0.478(.02)	**0.481**(.01)	0.434(.06)
	FPSO-RWG	0.480(.01)	0.333(.07)	0.329(.07)	0.470(.02)	0.385(.09)	0.476(.01)	0.480(.02)	0.465(.04)	**0.483**(.02)	0.433(.06)
Corel-5	FPSO-SV	0.125(.04)	0.160(.05)	0.167(.04)	0.146(.04)	**0.181**(.03)	0.119(.03)	0.115(.03)	0.107(.02)	0.180(.02)	0.105(.02)
	FPSO-RWL	0.147(.02)	0.088(.03)	0.094(.04)	0.106(.04)	0.151(.03)	0.104(.02)	0.161(.04)	**0.169**(.04)	**0.169**(.04)	0.098(.03)
	FPSO-RWG	0.135(.02)	0.102(.04)	0.102(.05)	0.106(.04)	0.156(.02)	0.117(.02)	**0.178**(.04)	0.165(.03)	0.174(.04)	0.111(.03)
Flowers	FPSO-SV	0.068(.01)	0.070(.01)	0.069(.01)	0.067(.01)	**0.071**(.01)	0.065(.01)	0.070(.01)	0.069(.01)	**0.071**(.01)	0.065(.01)
	FPSO-RWL	0.113(.01)	0.109(.02)	0.110(.01)	0.107(.01)	0.114(.02)	0.108(.01)	**0.123**(.02)	**0.123**(.02)	0.112(.01)	0.118(.01)
	FPSO-RWG	0.107(.01)	0.100(.02)	0.100(.02)	0.101(.01)	0.107(.02)	0.107(.01)	0.116(.02)	**0.120**(.02)	0.108(.02)	0.111(.03)
Image	FPSO-SV	0.229(.02)	**0.296**(.04)	0.222(.05)	0.229(.02)	0.237(.05)	0.210(.01)	0.230(.02)	0.227(.02)	0.223(.02)	0.240(.03)
	FPSO-RWL	0.255(.03)	0.282(.05)	0.255(.05)	0.266(.04)	0.226(.04)	**0.289**(.04)	0.271(.04)	0.273(.04)	0.275(.05)	0.189(.07)
	FPSO-RWG	0.289(.02)	0.290(.05)	0.259(.04)	0.294(.04)	0.231(.03)	0.281(.01)	0.297(.03)	**0.302**(.02)	0.279(.03)	0.253(.04)
Internet	FPSO-SV	0.131(.00)	0.167(.00)	0.117(.03)	0.269(.05)	0.443(.01)	0.131(.00)	0.433(.03)	**0.444**(.00)	0.296(.08)	0.254(.01)
	FPSO-RWL	0.076(.02)	0.077(.01)	-0.004(.02)	0.074(.00)	0.237(.14)	0.077(.00)	**0.248**(.09)	0.235(.07)	0.244(.09)	0.077(.08)
	FPSO-RWG	0.071(.02)	0.079(.00)	-0.005(.03)	0.075(.00)	0.231(.12)	0.077(.00)	0.249(.10)	**0.269**(.09)	0.221(.04)	0.109(.10)
Mfeat	FPSO-SV	0.291(.03)	0.320(.03)	0.311(.03)	0.285(.04)	0.311(.02)	0.305(.03)	0.312(.03)	0.301(.03)	0.325(.02)	**0.331**(.03)
	FPSO-RWL	0.336(.03)	0.368(.05)	0.356(.05)	0.327(.03)	0.351(.05)	0.342(.03)	0.320(.05)	0.319(.04)	**0.401**(.03)	0.384(.04)
	FPSO-RWG	0.379(.04)	0.364(.06)	0.354(.05)	0.327(.04)	0.359(.04)	0.340(.03)	0.386(.05)	0.378(.05)	**0.395**(.03)	0.386(.03)
Phoneme	FPSO-SV	0.554(.06)	0.401(.11)	0.403(.08)	0.383(.01)	0.313(.09)	**0.563**(.04)	0.544(.04)	0.531(.08)	0.481(.11)	0.373(.02)
	FPSO-RWL	**0.638**(.00)	0.599(.02)	0.583(.06)	0.592(.06)	0.508(.02)	0.618(.03)	0.519(.03)	0.591(.03)	0.583(.04)	0.343(.11)
	FPSO-RWG	0.328(.05)	**0.615**(.09)	0.608(.12)	0.310(.07)	0.410(.09)	0.313(.03)	0.545(.10)	0.546(.11)	0.595(.10)	0.458(.09)
3-sources	FPSO-SV	0.077(.03)	0.078(.03)	0.077(.03)	0.091(.03)	0.109(.03)	0.099(.04)	0.107(.04)	0.081(.03)	**0.137**(.05)	0.082(.03)
	FPSO-RWL	0.167(.05)	0.104(.05)	0.127(.05)	0.230(.05)	0.307(.07)	0.212(.06)	0.222(.06)	0.194(.05)	**0.347**(.05)	0.227(.07)
	FPSO-RWG	0.166(.04)	0.112(.04)	0.116(.02)	0.234(.06)	0.305(.05)	0.200(.04)	0.210(.04)	0.186(.04)	**0.333**(.06)	0.230(.07)

a base Caltech, os modelos para múltiplas visões obtiveram melhores médias para cinco funções contra três do FPSO-SV. Tendo isso em vista, pode-se afirmar que o agrupamento nebuloso de dados relacionais com múltiplas visões conseguiu obter desempenho médio superior em comparação ao agrupamento de dados com única visão para a maioria dos casos considerados e para a parametrização usada. Esse resultado também demonstra

que nem sempre é possível obter melhores agrupamentos ao explorar múltiplas visões dos conjuntos de dados, pois as visões podem conter informações conflitantes ou que não são complementares.

As Tabelas 51 e 52 sumarizam os resultados das abordagens híbridas considerando a medida F e o ARI. Os valores mostrados nas tabelas são as médias e desvios dos índices considerando-se todas as execuções dos métodos. Os gráficos *box-plots* desses dados também foram gerados e estão apresentados no Apêndice A. Nessas Tabelas é possível ter uma visão mais global dos resultados e observar, por exemplo, qual função obteve melhor desempenho médio para cada método e para cada base de dados. Uma Análise de Variância simples (MILLER, 1997) foi realizada para cada base de dados, índice e grupo de funções para verificar se existia diferença estatísticamente significativa entre os resultados encontrados. A hipótese nula não foi rejeitada para os seguintes casos: Corel-1 (FPSO-RWL para medida F), Flowers (FPSO-SV para os dois índices externos), Image (FSPO-RWG para o ARI) e Mfeat (FPSO-RWL para os dois índices). Para os casos em que a hipótese nula foi rejeitada, o pós-teste Holm-Bonferroni (HOLM, 1979) foi aplicado para cada combinação de função para realizar as comparações entre todos os pares de funções.

A Tabela 53 sumariza os resultados obtidos pela aplicação do teste Holm-Bonferroni. A parte triangular superior da tabela apresenta os resultados das comparações para o ARI e a parte triangular inferior contém os resultados das comparações para a medida F. Para cada par de função e algoritmo, um par de valores é apresentado, onde o primeiro valor representa o número de vezes que a função linha foi significativamente melhor considerando todas as bases de dados e o segundo valor representa o número de vezes que a função coluna foi significativamente melhor considerando todas as bases de dados.

As Figuras 29 e 30 fornecem outra forma de visualizar os resultados contidos na Tabela 33 e apresentam o número total de vezes que cada função foi melhor considerando todas as comparações significativas para todas as bases e ambos os índices. A silhueta simplificada apresentou o maior número e se destacou dentre todas as funções consideradas neste estudo. Dessa forma, os resultados obtidos pela silhueta simplificada serão utilizados para comparação com os resultados obtidos por outros métodos para agrupamento nebuloso de dados relacionais.

Por último, Testes t (nível de significância $\alpha = 0.05$) foram realizados para comparar os resultados dos índices medida F e ARI obtidos pelo $FPSO_{RWL}$ e $FPSO_{RWG}$ tendo o índice da silhueta simplificada e a homogeneidade como funções de aptidão, isto é, os pares comparados foram $FPSO^{SS}_{RWL}$-$FPSO^{SS}_{RWG}$ e $FPSO^{PC}_{RWL}$-$FPSO^{PC}_{RWG}$. Em 26 testes realizados, considerando o índice da silhueta simplificada, as diferenças foram significativas em apenas oito comparações, onde o $FPSO_{RWL}$ apresentou melhores médias em seis testes e o $FPSO_{RWG}$ apresentou maiores médias nos outros dois. Por outro lado, considerando o coeficiente da partição em 26 testes, as diferenças foram significativas em onze comparações, onde o $FPSO_{RWL}$ apresentou melhores médias em quatro testes e o $FPSO_{RWG}$ apresentou

Tabela 52 – Resultados da medida F pelos métodos híbridos com diferentes funções de aptidão

Base de dados	Algoritmo	AWCD	CS	FCS	FS	FSS	HM	PC	PE	SS	XB
Animals	FPSO-SV	0.392(.01)	0.420(.01)	0.421(.01)	0.409(.01)	**0.449**(.01)	0.416(.01)	0.434(.02)	0.424(.02)	**0.449**(.01)	0.448(.01)
	FPSO-RWL	0.467(.00)	0.468(.01)	0.476(.02)	0.475(.00)	0.446(.00)	0.475(.00)	0.470(.00)	0.475(.00)	0.476(.00)	**0.490**(.00)
	FPSO-RWG	0.474(.01)	0.469(.04)	0.474(.04)	**0.490**(.02)	0.465(.05)	0.473(.00)	0.472(.04)	0.452(.06)	0.471(.05)	0.477(.03)
Caltech101-7	FPSO-SV	**0.692**(.02)	0.528(.06)	0.546(.06)	0.681(.03)	0.636(.05)	0.686(.02)	0.661(.05)	0.646(.05)	0.678(.02)	0.548(.04)
	FPSO-RWL	0.594(.04)	0.631(.06)	0.630(.08)	0.628(.06)	0.702(.04)	0.608(.05)	0.638(.03)	0.619(.06)	**0.717**(.01)	0.614(.07)
	FPSO-RWG	0.600(.00)	0.544(.07)	0.532(.06)	0.612(.07)	0.665(.06)	0.613(.02)	0.674(.02)	**0.683**(.02)	0.658(.04)	0.654(.05)
Corel-1	FPSO-SV	0.657(.03)	0.566(.04)	0.535(.06)	0.569(.06)	0.612(.04)	0.621(.04)	0.640(.04)	0.632(.06)	**0.654**(.02)	0.586(.04)
	FPSO-RWL	0.590(.07)	0.555(.05)	0.563(.06)	0.535(.04)	0.573(.04)	0.554(.04)	0.617(.03)	0.597(.06)	**0.626**(.05)	0.580(.04)
	FPSO-RWG	0.584(.03)	0.543(.06)	0.566(.06)	0.561(.05)	0.574(.07)	0.585(.03)	0.589(.06)	0.606(.06)	**0.621**(.03)	0.566(.03)
Corel-2	FPSO-SV	0.630(.04)	0.542(.05)	0.527(.06)	**0.653**(.02)	0.541(.04)	0.650(.02)	0.616(.03)	0.616(.03)	0.623(.05)	0.614(.07)
	FPSO-RWL	0.641(.02)	0.597(.05)	**0.672**(.05)	0.625(.04)	0.652(.04)	0.571(.04)	0.583(.03)	0.581(.03)	0.625(.06)	0.566(.06)
	FPSO-RWG	0.633(.03)	**0.670**(.06)	0.621(.06)	0.617(.04)	0.611(.04)	0.621(.04)	0.599(.04)	0.601(.03)	0.601(.05)	0.614(.07)
Corel-3	FPSO-SV	**0.730**(.01)	0.651(.02)	0.645(.02)	0.726(.03)	0.557(.02)	0.722(.03)	0.717(.02)	0.724(.01)	0.659(.05)	0.577(.02)
	FPSO-RWL	0.661(.06)	0.629(.06)	0.630(.04)	0.627(.04)	0.627(.04)	0.612(.00)	**0.675**(.05)	0.666(.06)	0.669(.05)	0.628(.03)
	FPSO-RWG	**0.675**(.05)	0.644(.06)	0.648(.05)	0.622(.04)	0.622(.04)	0.610(.00)	0.660(.05)	0.646(.04)	0.662(.05)	0.616(.03)
Corel-4	FPSO-SV	**0.603**(.03)	0.550(.03)	0.554(.03)	0.559(.04)	0.567(.05)	0.542(.01)	0.600(.03)	0.596(.02)	0.567(.03)	0.582(.05)
	FPSO-RWL	0.718(.02)	0.662(.06)	0.656(.06)	0.708(.03)	0.655(.07)	0.694(.04)	0.689(.04)	0.717(.02)	**0.721**(.01)	0.697(.04)
	FPSO-RWG	**0.720**(.01)	0.634(.05)	0.632(.06)	0.709(.02)	0.663(.06)	0.715(.01)	**0.720**(.01)	0.710(.03)	**0.720**(.01)	0.695(.03)
Corel-5	FPSO-SV	0.485(.03)	0.514(.04)	**0.532**(.03)	0.499(.03)	0.521(.02)	0.486(.03)	0.478(.02)	0.469(.02)	0.520(.02)	0.462(.02)
	FPSO-RWL	0.461(.02)	0.422(.03)	0.437(.05)	0.443(.03)	0.481(.03)	0.441(.01)	0.481(.04)	0.488(.04)	**0.495**(.03)	0.432(.03)
	FPSO-RWG	0.450(.02)	0.440(.04)	0.442(.05)	0.442(.04)	0.494(.02)	0.447(.01)	0.496(.04)	0.488(.04)	**0.499**(.04)	0.449(.03)
Flowers	FPSO-SV	0.227(.01)	0.231(.01)	0.227(.01)	0.226(.01)	0.230(.01)	0.225(.01)	0.228(.02)	0.225(.01)	**0.232**(.01)	0.223(.01)
	FPSO-RWL	0.269(.01)	0.276(.02)	0.283(.02)	0.268(.02)	0.269(.02)	0.271(.00)	0.291(.02)	0.287(.02)	0.268(.02)	**0.317**(.02)
	FPSO-RWG	0.271(.01)	0.266(.02)	0.266(.02)	0.266(.02)	0.261(.02)	0.271(.01)	0.280(.02)	**0.285**(.02)	0.262(.03)	0.275(.03)
Image	FPSO-SV	0.444(.02)	**0.505**(.04)	0.437(.04)	0.449(.02)	0.453(.05)	0.424(.01)	0.448(.02)	0.442(.03)	0.435(.02)	0.460(.03)
	FPSO-RWL	0.462(.04)	**0.501**(.04)	0.476(.04)	0.464(.04)	0.438(.04)	0.496(.04)	0.472(.04)	0.476(.04)	0.479(.05)	0.424(.06)
	FPSO-RWG	0.506(.03)	0.513(.05)	0.481(.04)	0.501(.03)	0.448(.03)	0.494(.02)	0.512(.02)	**0.520**(.02)	0.485(.04)	0.449(.05)
Internet	FPSO-SV	0.730(.00)	0.754(.00)	0.717(.03)	0.809(.05)	0.875(.01)	0.730(.00)	0.871(.03)	**0.875**(.00)	0.819(.08)	0.803(.01)
	FPSO-RWL	0.685(.02)	0.687(.01)	0.615(.02)	0.684(.00)	0.787(.06)	0.687(.00)	**0.795**(.04)	0.791(.03)	0.793(.04)	0.701(.05)
	FPSO-RWG	0.681(.02)	0.689(.00)	0.627(.02)	0.685(.00)	0.787(.05)	0.686(.00)	0.794(.04)	**0.806**(.04)	0.785(.02)	0.721(.06)
Mfeat	FPSO-SV	0.485(.03)	**0.515**(.03)	0.511(.04)	0.472(.04)	0.476(.02)	0.491(.03)	0.465(.03)	0.471(.04)	0.513(.03)	0.512(.04)
	FPSO-RWL	0.514(.03)	**0.552**(.05)	0.541(.05)	0.499(.02)	0.494(.04)	0.511(.03)	0.489(.04)	0.493(.04)	0.546(.02)	0.542(.03)
	FPSO-RWG	0.543(.03)	**0.556**(.06)	0.547(.05)	0.496(.04)	0.506(.04)	0.510(.03)	0.544(.05)	0.536(.04)	0.545(.03)	0.544(.03)
Phoneme	FPSO-SV	0.718(.05)	0.629(.07)	0.617(.06)	0.602(.01)	0.560(.06)	**0.739**(.03)	0.770(.03)	0.700(.05)	0.681(.08)	0.604(.03)
	FPSO-RWL	**0.799**(.00)	0.777(.01)	0.762(.04)	0.769(.04)	0.705(.01)	0.788(.03)	0.709(.02)	0.767(.02)	0.762(.03)	0.592(.07)
	FPSO-RWG	0.582(.04)	**0.788**(.06)	0.782(.09)	0.574(.05)	0.631(.08)	0.570(.02)	0.738(.07)	0.739(.08)	0.771(.07)	0.682(.08)
3-Sources	FPSO-SV	0.382(.04)	0.369(.03)	0.372(.03)	0.399(.03)	0.404(.03)	0.403(.04)	0.410(.04)	0.389(.03)	**0.431**(.04)	0.386(.03)
	FPSO-RWL	0.443(.04)	0.414(.04)	0.413(.04)	0.482(.04)	0.537(.04)	0.473(.04)	0.474(.04)	0.479(.04)	**0.559**(.03)	0.491(.05)
	FPSO-RWG	0.452(.04)	0.407(.04)	0.415(.03)	0.498(.04)	0.529(.03)	0.471(.03)	0.477(.03)	0.466(.04)	**0.548**(.04)	0.496(.05)

maiores médias nos outros sete testes.

As Figuras 31 - 34 apresentam índices calculados para os melhores agrupamentos encontrados para cada função. Para ambos os índices e ambos os métodos, observa-se que a silhueta simplificada e o coeficiente da partição estiveram entre as funções que apresentaram melhores agrupamentos para a maioria das bases. Para o ARI,

8.3.2.1 Análise de robustez

A robustez das funções foi analisada com o mesmo método usado no capítulo anterior. As Figuras 35 e 36 apresentam as distribuições de b_m de cada função. Quanto maior o valor apresentado nas figuras, maior a robustez. Verifica-se que, para o método $FPSO - RWL$ e ambos os índices externos, a silhueta simplificada foi mais robusta dentre as funções comparadas. Para o método $FPSO-RWG$ e medida F, o coeficiente da partição apresentou maior robustez, enquanto que, para o ARI, foram mais robustos juntamente com a silhueta o coeficiente da partição juntamente com a silhueta simplificada aprensetaram maior robustez.

Tabela 53 – Sumário dos resultados encontrados pela aplicação do teste Holm-Bonferroni para todos os pares

	Algorithm	AWCD	CS	FCS	FS	FSS	HM	PC	PE	SS	XB
AWCD	FPSO-SV	-	6-4	6-2	2-2	6-4	2-3	0-2	1-2	4-5	4-3
	FPSO-RWL	-	5-5	5-5	3-1	3-4	6-1	3-2	3-1	4-5	2-0
	FPSO-RWG	-	6-3	8-1	1-1	2-5	2-1	1-4	1-3	0-8	5-5
CS	FPSO-SV	3-6	-	2-0	2-5	3-5	4-5	2-5	2-7	1-8	3-4
	FPSO-RWL	1-4	-	3-1	2-2	5-3	6-3	6-5	4-4	2-4	4-0
	FPSO-RWG	3-5	-	2-0	3-4	3-5	3-3	3-8	3-7	2-8	2-5
FSC	FPSO-SV	2-7	1-3	-	1-5	2-5	3-5	1-8	1-7	0-8	3-3
	FPSO-RWL	1-3	0-2	-	3-2	4-2	4-3	5-6	4-5	1-5	3-1
	FPSO-RWG	2-5	0-2	-	2-4	5-2	2-5	2-8	2-7	1-8	2-5
FS	FPSO-SV	2-4	5-4	6-3	-	5-4	1-2	2-4	2-3	2-4	4-0
	FPSO-RWL	2-3	3-1	3-1	-	3-6	2-1	4-3	3-4	1-4	2-1
	FPSO-RWG	1-1	3-4	4-2	-	1-5	0-1	2-5	2-5	0-10	0-3
FSS	FPSO-SV	3-7	5-3	5-2	4-5	-	3-4	1-5	1-3	1-5	4-3
	FPSO-RWL	2-3	5-2	3-2	4-2	-	6-3	5-2	4-3	0-3	3-1
	FPSO-RWG	4-2	4-3	5-2	4-1	-	6-2	2-3	3-3	0-3	2-1
HM	FPSO-SV	2-3	6-3	5-1	3-1	6-3	-	1-2	3-2	3-6	3-3
	FPSO-RWL	3-5	2-1	2-2	0-2	3-4	-	2-3	2-3	2-7	2-2
	FPSO-RWG	0-5	4-4	4-3	0-0	2-5	-	1-3	1-3	0-6	1-3
PC	FPSO-SV	2-0	9-3	9-3	4-4	5-2	2-3	-	1-0	4-1	5-0
	FPSO-RWL	4-3	2-1	5-2	4-3	3-3	4-2	-	0-1	2-6	1-1
	FPSO-RWG	3-2	5-3	6-2	5-1	4-2	4-1	-	0-0	0-4	3-3
PE	FPSO-SV	2-2	6-3	6-2	4-4	4-2	2-3	1-1	-	3-4	5-1
	FPSO-RWL	3-2	3-0	3-1	3-1	3-3	4-1	3-0	-	2-5	2-0
	FPSO-RWG	3-2	5-3	7-2	5-2	4-3	5-1	0-0	-	1-4	5-4
SS	FPSO-SV	4-2	6-1	7-0	4-3	4-1	5-2	2-3	3-3	-	8-0
	FPSO-RWL	3-2	4-0	5-1	5-0	5-0	7-1	5-1	3-1	-	8-0
	FPSO-RWG	8-0	7-2	8-0	10-0	4-1	8-0	3-1	4-2	-	7-1
XB	FPSO-SV	2-5	4-3	5-2	1-3	3-4	4-4	1-3	1-3	1-6	-
	FPSO-RWL	3-2	3-3	4-2	3-1	3-3	3-2	3-4	3-3	2-7	-
	FPSO-RWG	5-4	5-4	5-1	4-0	3-1	6-0	2-4	3-3	1-7	-

8.3.3 Estudo 6 - Comparação com outros métodos

Os resultados obtidos pela otimização da silhueta simplificada como função de aptidão para os métodos FPSO-RWL e FPSO-RWG foram selecionados para comparação com os resultados obtidos por outros métodos adequados para agrupamento nebuloso de dados relacionais. Os algoritmos comparados também foram executados 30 vezes e tiveram o número máximo de iterações igual a 100. A Tabela 54 apresenta os resultados dos métodos

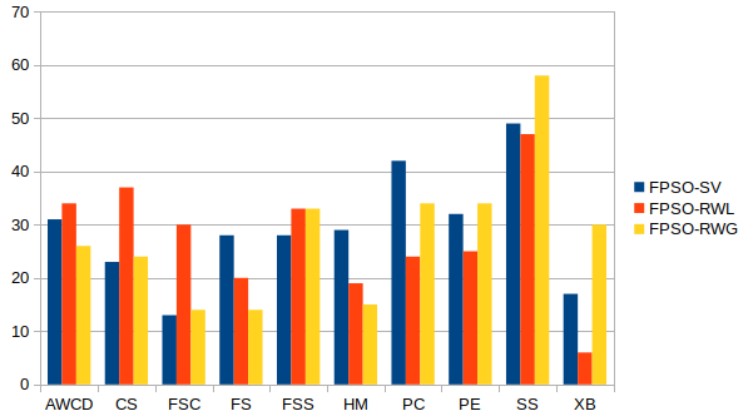


Figura 29 – Sumário dos resultados Bonferroni para ARI

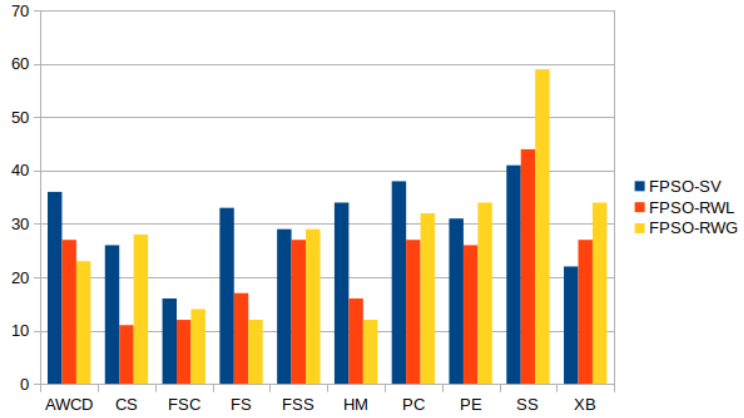


Figura 30 – Sumário dos resultados Bonferroni para a medida F

comparados. A Análise de Variância Simples foi utilizada para verificar se existia diferença estatisticamente significativa. A hipótese nula foi rejeitada em todos os casos. O pós-teste Holm-Bonferroni (HOLM, 1979) foi aplicado para comparar os métodos híbridos com os outros algoritmos par a par.

No total, 364 comparações foram realizadas. O símbolo "*"indica que a diferença entre as médias, considerando o método FPSO-RWL, foi significante estatisticamente. O símbolo "+"indica que a diferença entre as médias, considerando o método FPSO-RWG, foi significante estatisticamente. Se o símbolo estiver vermelho, então a média apresentada pelo algoritmo comparado foi maior do que o método proposto comparado.

Levando em consideração as comparações feitas com o FPSO-RWL, as diferenças foram significantes estatisticamente em 70 testes para a medida F e em 71 testes para o ARI, onde o FPSO-RWL foi melhor em 45 testes para a medida F e em 43 testes para o ARI. Considerando as comparações feitas com o FPSO-RWG, as diferenças foram significantes estatisticamente em 72 testes para a medida F e em 75 testes para o ARI, onde o FPSO-RWG foi melhor em 46 testes para a medida F e em 48 testes para o ARI. Portanto, estes resultados demonstram a efetividade da abordagem proposta para o agrupamento nebuloso de dados relacionais com múltiplas visões. A abordagem híbrida

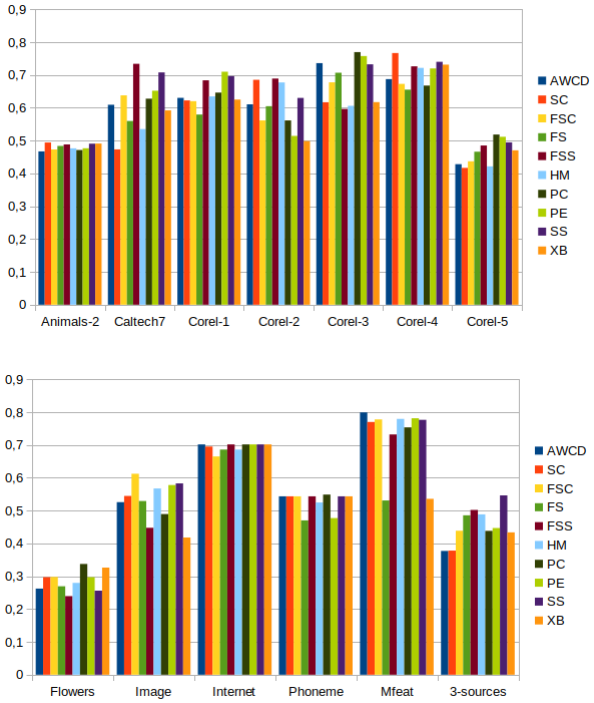


Figura 31 – Melhor solução encontrada pelo FPSO-RWL com cada função em termos de F-measure

proposta apresentou resultados competitivos e melhores resultados em alguns casos.

Tabela 54 – Sumário dos resultados apresentados pelos métodos comparados

Data set	Index	$SFCMdd$	$NERF$	$FANNY$	$CARD_R$	$CARD_F$	$MFCMdd_{RWL}$	$MFCMdd_{RWG}$
Animals	ARI	0.00(.00)$^{*+}$	0.17(.07)$^{+}$	0.17(.07)$^{+}$	0.16(.00)$^{*+}$	0.24(.06)*	**0.25**(.06)*	0.01(.00)$^{*+}$
	F-measure	0.32(.00)$^{*+}$	0.46(.05)$^{+}$	0.46(.05)$^{+}$	0.46(.00)$^{+}$	**0.51**(.04)*	**0.51**(.03)*	0.32(.00)$^{*+}$
Caltech101-7	ARI	0.28(.02)$^{*+}$	**0.66**(.09)$^{*+}$	−0.07(.03)$^{*+}$	0.44(.12)$^{+}$	0.29(.10)*	0.24(.02)$^{*+}$	0.30(.02)$^{*+}$
	F-measure	0.63(.01)$^{*+}$	**0.78**(.04)$^{*+}$	0.46(.02)$^{*+}$	0.68(.08)*	0.60(.06)$^{*+}$	0.57(.03)$^{*+}$	0.65(.01)*
Corel-1	ARI	0.26(.08)$^{*+}$	0.26(.01)$^{*+}$	**0.42**(.12)$^{*+}$	0.31(.07)	0.30(.05)	0.18(.04)$^{*+}$	0.21(.03)$^{*+}$
	F-measure	0.55(.07)$^{*+}$	0.55(.01)$^{*+}$	**0.68**(.08)$^{*+}$	0.59(.05)$^{*+}$	0.58(.04)$^{*+}$	0.51(.04)$^{*+}$	0.53(.04)$^{*+}$
Corel-2	ARI	0.29(.09)*	0.23(.00)$^{*+}$	0.33(.08)	0.22(.01)$^{*+}$	0.22(.01)$^{*+}$	0.25(.09)$^{*+}$	**0.34**(.06)
	F-measure	0.55(.07)$^{*+}$	0.52(.00)$^{*+}$	0.60(.05)	0.51(.01)$^{*+}$	0.50(.01)$^{*+}$	0.55(.06)$^{*+}$	**0.61**(.05)
Corel-3	ARI	0.29(.05)$^{*+}$	0.23(.00)$^{*+}$	0.33(.00)$^{*+}$	0.22(.00)$^{*+}$	0.22(.00)$^{*+}$	0.25(.03)$^{*+}$	**0.34**(.02)$^{*+}$
	F-measure	0.55(.05)	0.52(.00)$^{*+}$	0.60(.00)$^{*+}$	0.51(.00)$^{*+}$	0.50(.00)$^{*+}$	0.55(.03)$^{*+}$	**0.61**(.02)$^{*+}$
Corel-4	ARI	0.29(.04^{*+}	0.23(.00)$^{*+}$	0.33(.02)$^{*+}$	0.22(.00)$^{*+}$	0.22(.00)$^{*+}$	0.25(.08)$^{*+}$	**0.34**(.09)$^{+}$
	F-measure	0.55(.04^{*+}	0.52(.00)$^{*+}$	0.60(.02)$^{*+}$	0.51(.00)$^{*+}$	0.50(.00)$^{*+}$	0.55(.07)$^{*+}$	**0.61**(.07)$^{*+}$
Corel-5	ARI	0.29(.03)$^{*+}$	0.23(.01)$^{*+}$	0.33(.02)$^{*+}$	0.22(.00)$^{*+}$	0.22(.00)$^{*+}$	0.25(.06)$^{*+}$	**0.34**(.05)$^{*+}$
	F-measure	0.55(.05)	0.52(.01)$^{+}$	0.60(.01)$^{+}$	0.51(.00)$^{*+}$	0.50(.00)$^{*+}$	0.55(.05)$^{*+}$	**0.61**(.05)$^{*+}$
Flowers	ARI	0.10(.01)	0.14(.02)$^{*+}$	**0.21**(.05)$^{*+}$	0.10(.02)	0.12(.02)	0.11(.01)	0.11(.01)
	F-measure	0.27(.02)	0.28(.02)$^{+}$	**0.36**(.04)$^{*+}$	0.24(.02)$^{*+}$	0.27(.02)	0.27(.01)	0.26(.01)
Image	ARI	0.21(.04)$^{*+}$	0.35(.05)$^{*+}$	0.35(.03)$^{*+}$	**0.40**(.01)$^{*+}$	0.19(.03)$^{*+}$	0.20(.03)$^{*+}$	0.25(.06)
	F-measure	0.43(.04)$^{*+}$	0.56(.05)$^{*+}$	0.57(.04)$^{*+}$	**0.58**(.01)$^{*+}$	0.43(.02)$^{*+}$	0.43(.03)$^{*+}$	0.46(.06)
Internet	ARI	0.12(.00)$^{*+}$	**0.29**(.00)$^{*+}$	**0.29**(.00)$^{*+}$	0.20(.05)$^{*+}$	0.08(.00)$^{*+}$	0.03(.03)$^{*+}$	0.05(.00)$^{*+}$
	F-measure	0.71(.00)$^{*+}$	**0.82**(.00)$^{*+}$	**0.82**(.00)$^{*+}$	0.77(.04)$^{*+}$	0.69(.00)	0.79(.05)$^{*+}$	0.72(.07)
Mfeat	ARI	**0.37**(.03)*	0.32(.00)$^{+}$	0.34(.01)$^{+}$	0.22(.04)$^{*+}$	0.26(.04)$^{*+}$	0.32(.02)$^{+}$	0.34(.01)$^{+}$
	F-measure	**0.53**(.03)	0.48(.01)$^{*+}$	0.50(.02)$^{*+}$	0.38(.04)$^{*+}$	0.43(.04)$^{*+}$	0.48(.02)$^{*+}$	0.49(.02)$^{*+}$
Phoneme	ARI	**0.63**(.06)$^{*+}$	0.20(.05)$^{*+}$	0.44(.09)$^{*+}$	0.21(.03)$^{*+}$	0.19(.01)$^{*+}$	0.58(.10)	0.56(.07)$^{*+}$
	F-measure	**0.78**(.04)	0.48(.03)$^{*+}$	0.63(.07)$^{*+}$	0.50(.03)$^{*+}$	0.48(.01)$^{*+}$	0.74(.07)	0.74(.08)
3-Sources	ARI	0.28(.06)$^{*+}$	**0.39**(.03)$^{*+}$	0.06(.02)$^{*+}$	0.37(.04)$^{*+}$	0.02(.02)$^{*+}$	0.28(.05)$^{*+}$	0.29(.05)$^{*+}$
	F-measure	0.51(.04)$^{*+}$	0.58(.00)$^{*+}$	0.38(.02)$^{*+}$	**0.59**(.00)$^{*+}$	0.35(.02)$^{*+}$	0.53(.03)$^{*+}$	0.52(.03)$^{*+}$

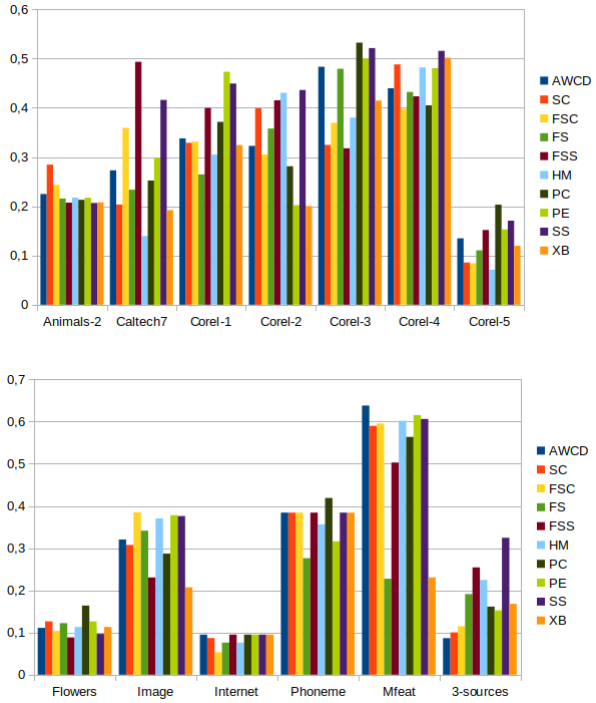


Figura 32 – Melhor solução encontrada pelo FPSO-RWL com cada função em termos de ARI

8.3.4 Aplicação: Image Segmentation

A Tabela 55 apresenta os pesos de relevância das visões encontrados pela melhor solução obtida por cada método de agrupamento nebuloso para dados relacionais com múltiplas visões considerando a base *Image Segmentation*. Como dito anteriormente, os pesos de relevância expressam a importância de cada visão para definição dos grupos. Portanto, observa-se que o peso de relevância da matriz de dissimilaridades "RGB"foi mais importante para definição de todos os grupos para todos os algoritmos, exceto para o CARD-F e para o MFCMdd-RWL, em que a visão "shape"foi mais importante na definição dos grupos C_1 e C_7, por exemplo.

Além disso, as Figuras 37 - 47 apresentam os gráficos *box-plots* para as bases de dados considerando os dois índices externos calculados para cada partição fornecida por cada método em cada execução. É possível observar quais algoritmos apresentaram maiores variações e quais apresentaram medianas mais altas, isto é, quais deles encontraram melhores partições de acordo com os valores dos índices externos. De forma geral, percebe-se que os métodos desenvolvidos apresentaram baixa dispersão e conseguiram apresentar medianas mais altas em comparação aos outros métodos na maioria dos casos.

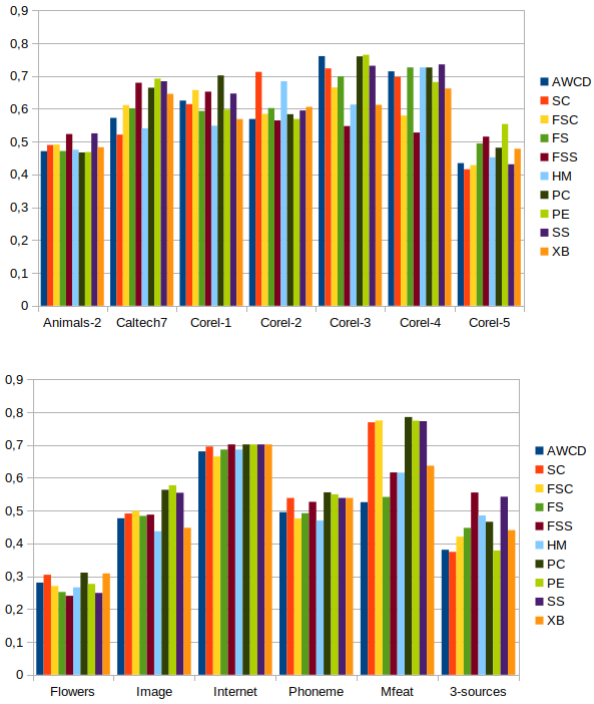


Figura 33 – Melhor solução encontrada pelo FPSO-RWG com cada função em termos de F-measure

Tabela 55 – Image: vetores de pesos de relevancia

	$FPSO_{RWL} - SS$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$FPSO_{RWG} - SS$
1 - shape	0.975	0.895	0.909	0.984	0.953	0.938	0.970	0.929
2 - rgb	1.024	1.116	1.099	1.016	1.048	1.065	1.030	1.076
	$MFCMdd_{RWL}$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$MFCDMdd_{RWG}$
1 - shape	1.052	0.998	1.036	0.848	0.788	0.802	1.087	0.899
2 - rgb	0.950	1.001	0.965	1.178	1.267	1.246	0.919	1.111
	$CARD - R$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
1 - shape	0.466	0.349	0.377	0.210	0.417	0.277	0.375	
2 - rgb	0.533	0.650	0.622	0.789	0.582	0.722	0.624	
	$CARD - F$							
View	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
1 - shape	0.624	0.382	0.475	0.612	0.327	0.423	0.577	
2 - rgb	0.375	0.617	0.524	0.387	0.672	0.576	0.422	

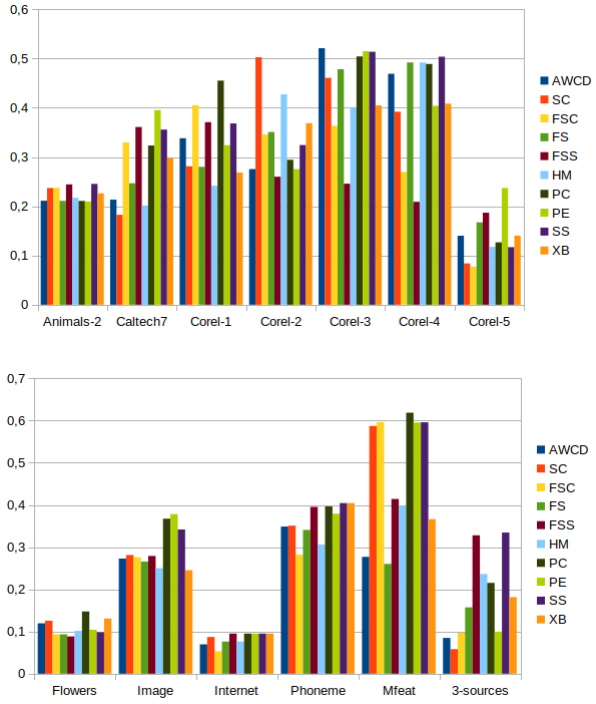
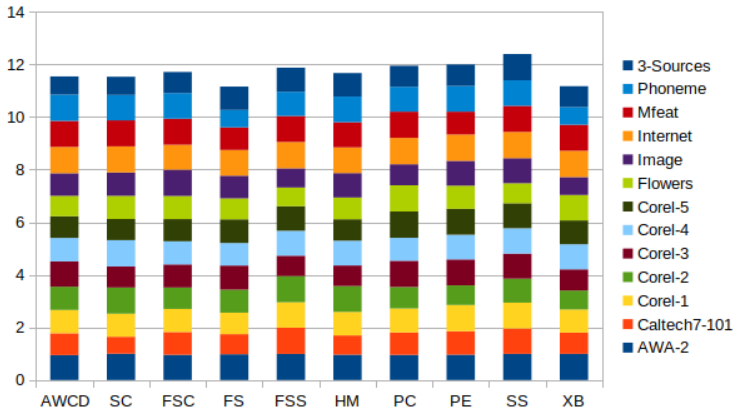
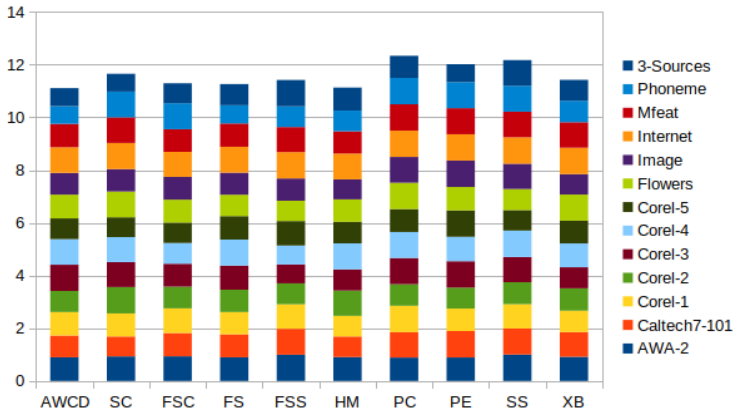


Figura 34 – Melhor solução encontrada pelo FPSO-RWG com cada função em termos de ARI

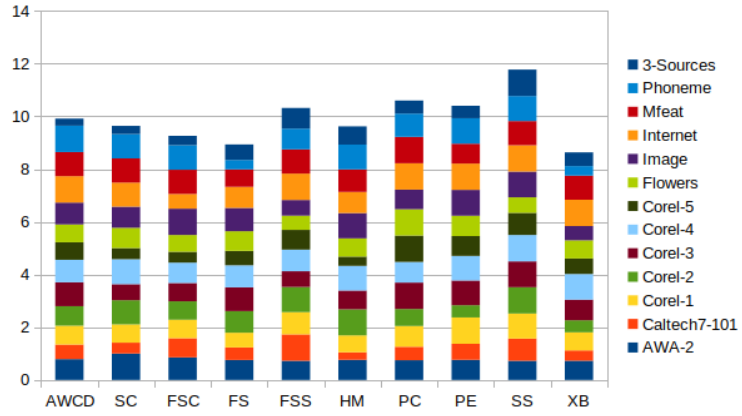


(a) FPSO-RWL

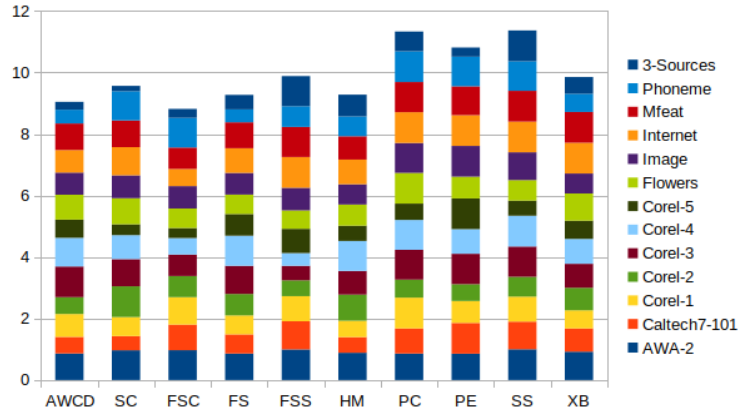


(b) FPSO-RWG

Figura 35 – Robustez das funções de aptidão considerando o F-measure



(a) FPSO-RWL



(b) FPSO-RWG

Figura 36 – Robustez das funções de aptidão considerando o ARI

8.4 DISCUSSÃO

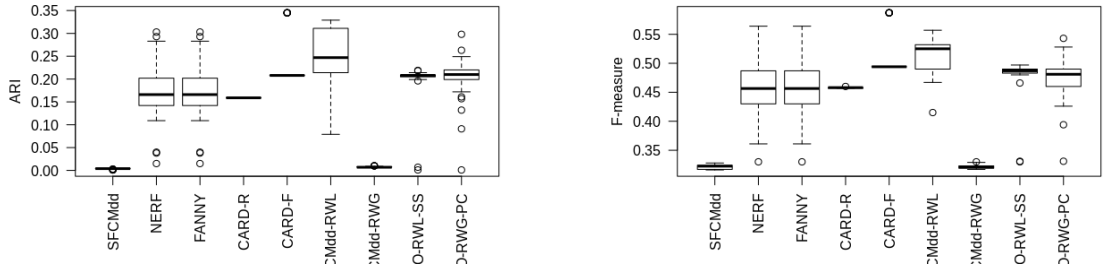
O Estudo 4 teve como objetivo observar se o método FPSO-SV seria capaz de apresentar bons resultados, quando comparado a outros métodos, para o agrupamento nebuloso de dados relacionais com apenas uma visão. A partir do Estudo 4, observou-se que o método desenvolvido apresentou resultados competitivos. Nesse estudo, a função utilizada foi a homogeneidade. Como existem várias funções na literatura, o desempenho do método poderia ter sido melhor considerando-se outra função. Além disso, outras parametrizações também poderiam ter melhorado o desempenho do método.

A partir do Estudo 5, foi possível investigar respostas para três perguntas de pesquisa. Os resultados obtidos no estudo sugerem que agrupar dados relacionais com múltiplas visões pode trazer melhores resultados em relação ao agrupamento nebuloso de dados com visão única a depender da base de dados e da função de aptidão utilizada. Além disso, as duas funções de aptidão que se destacaram, com base nas análises efetuadas, foram: silhueta simplificada e o coeficiente da partição. É interessante observar que um deles utiliza apenas a partição nebulosa para avaliar um agrupamento e que o outro índice utiliza o agrupamento, mas não considera a partição nebulosa. Além disso, a silhueta simplificada também avalia simultaneamente a separação dos grupos quanto a coesão interna deles.

Em relação a comparação entre o desempenho obtido pelo método para dados com únicao visão com o desempenho obtido pelos métodos para dados com múltiplas visões, a exploração das múltiplas visoes trouxe ganhos para a maioria dos casos. Este fato confirma que, a depender da base de dados considerada, as visões possuem importância diferente para o agrupamento nebuloso de dados assim como foi observado para o agrupamento rígido.

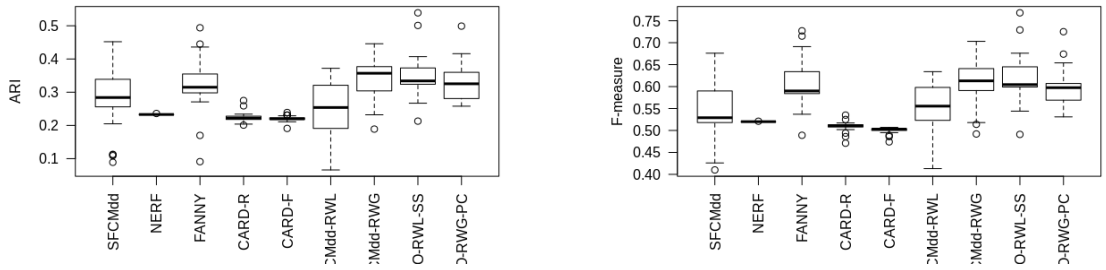
A estimação de pesos localmente apresentou melhores resultados para a silhueta simplificada e o coeficiente da partição apresentou melhores resultados com a estimação de pesos globais para os índices e bases de dados considerados. Os resultados desses testes servem como base para responder a terceira pergunta de pesquisa considerada neste estudo. Portanto, a depender da função de aptidão considerada, a estimação dos pesos tem influência. Nos experimentos realizados nesse estudo, as duas formas de estimação apresentaram resultados similares que divergiram de forma significativa em menos da metade dos casos avaliados.

Na comparação feita com outros algoritmos de agrupamento nebulosos, a abordagem desenvolvida neste trabalho obteve resultados competitivos e melhores significativamente para alguns casos. Por consequência disso, essas evidências empíricas suportam, ao menos parcialmente, a tese deste trabalho que afirma que a utilização de métodos híbridos baseados em PSO melhora de forma significativa o agrupamento de dados relacionais com múltiplas visões.



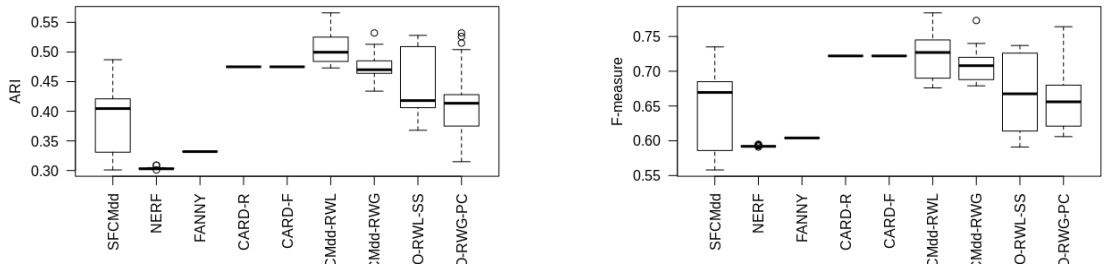
(a) ARI (b) F-measure

Figura 37 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Animals



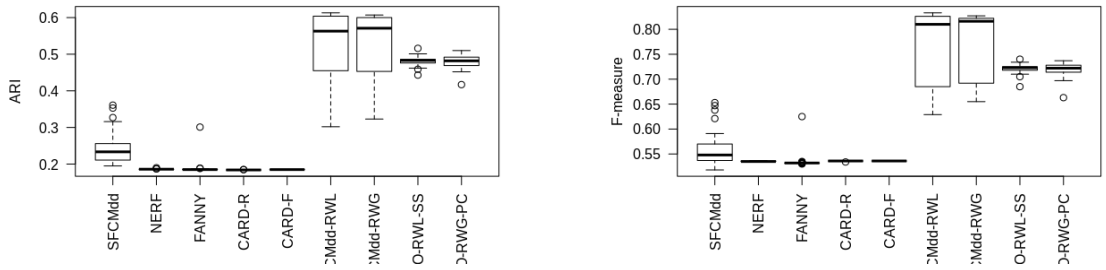
(a) ARI (b) F-measure

Figura 38 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-2



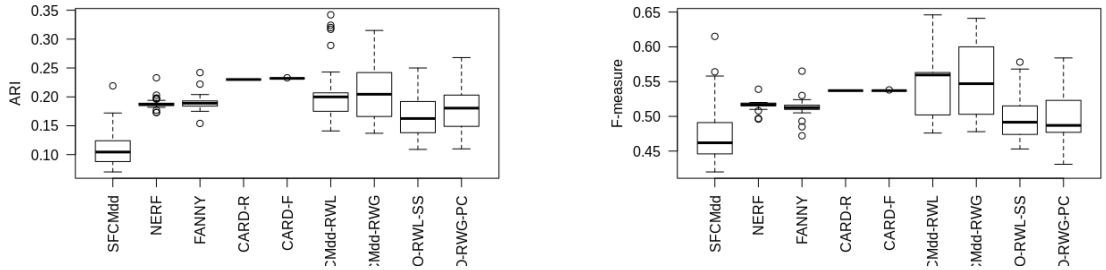
(a) ARI (b) F-measure

Figura 39 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-3



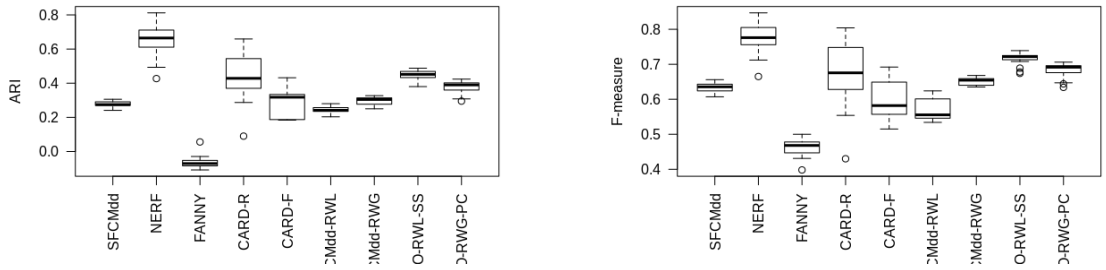
(a) ARI (b) F-measure

Figura 40 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-4



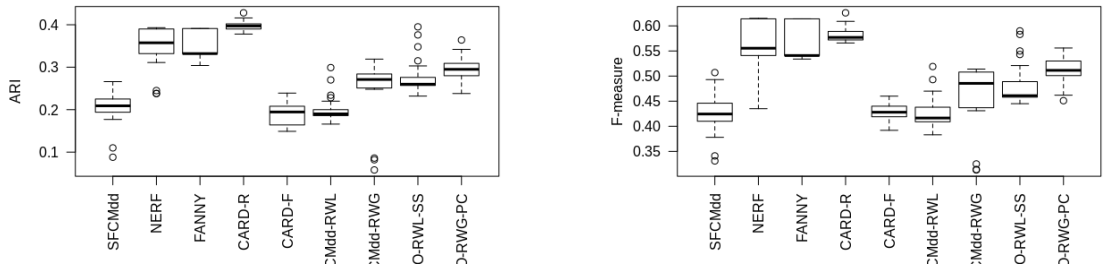
(a) ARI (b) F-measure

Figura 41 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Corel-5



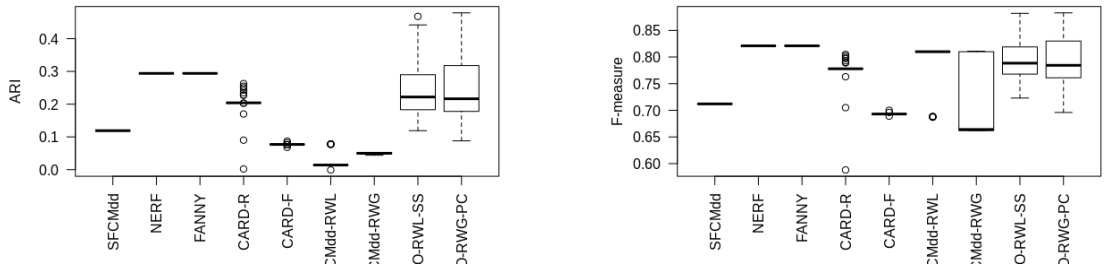
(a) ARI (b) F-measure

Figura 42 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Caltech101-7



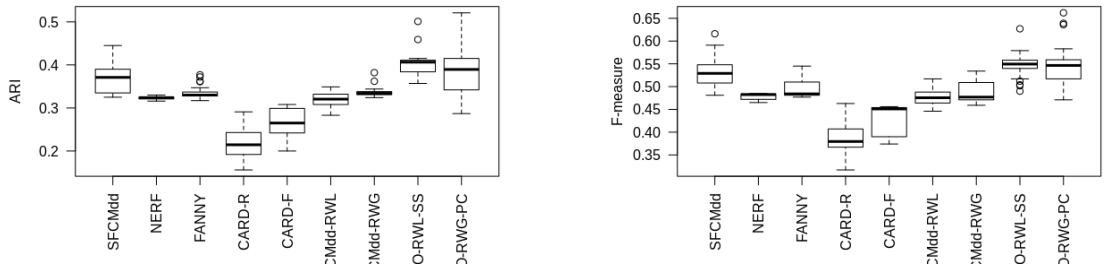
(a) ARI (b) F-measure

Figura 43 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Image



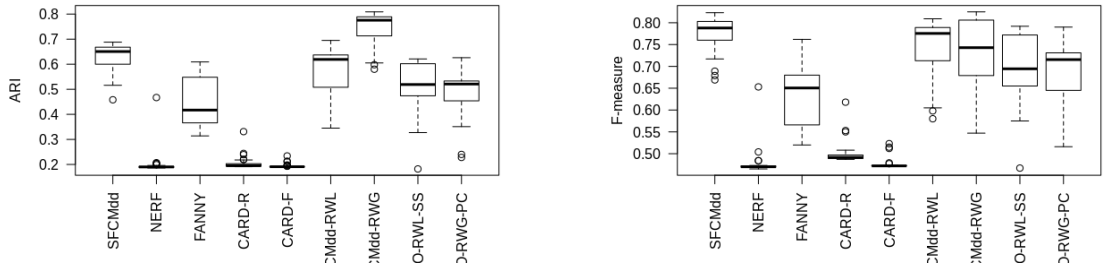
(a) ARI (b) F-measure

Figura 44 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Internet



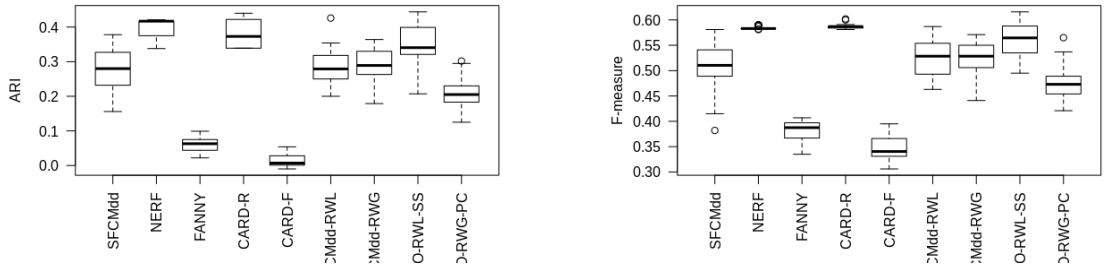
(a) ARI (b) F-measure

Figura 45 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Multiple features



(a) ARI (b) F-measure

Figura 46 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base Phoneme



(a) ARI (b) F-measure

Figura 47 – Gráfico Box-Plot com os índices externos obtidos pelos métodos para a base 3-Sources

8.5 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados os resultados encontrados pelas abordagens híbridas para dados relacionais. O desempenho de diversas funções de aptidão foi avaliado. A abordagem híbrida foi comparada com outros métodos de agrupamento nebuloso adequados para dados relacionais. Os resultados encontrados demonstraram que a abordagem híbrida foi competitiva em relação aos outros métodos e conseguiu apresentar melhores resultados para alguns casos.

9 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo a investigação e desenvolvimento de métodos híbridos baseados em otimização por nuvem de partículas para o agrupamento rígido e nebuloso de dados relacionais com múltiplas visões. Os métodos híbridos combinam PSO com métodos de agrupamento baseados em matrizes de dissimilaridades com o objetivo de se beneficiar das vantagens de cada abordagem e, consequentemente, explorar de melhor forma o espaço de busca. A modelagem dos métodos híbridos baseados em PSO foi uma das principais contribuições deste trabalho.

Na modelagem desenvolvida, a representação das partículas foi definida com base em múltiplos *medóides*, pois com a representação de dados relacionais não é possível ter acesso aos atributos dos objetos e calcular os centróides diretamente. Na definição da velocidade, os componentes sociais e cognitivos foram propostos baseando-se nas dissimilaridades entre os objetos e os múltiplos *medóides* que representam os grupos. A aplicação da velocidade também foi definida de forma diferente do que foi realizado em trabalhos anteriores existentes na literatura. A atualização da posição foi realizada com base em algoritmos de agrupamento baseados em várias matrizes que foram modificados. Por fim, a aplicação da velocidade foi realizada para afetar a escolha dos representativos de cada grupo em cada iteração dos métodos.

Além disso, diversas funções de aptidão foram investigadas tanto para o agrupamento rígido quanto para o agrupamento nebuloso. A adaptação dos índices foi outra contribuição deste trabalho. A adaptação envolveu três mudanças em relação a definição tradicional dos índices para dados vetoriais: (i) os índices considerados para agrupamento de dados relacionais com múltiplas visões consideram múltiplos *medóides* como representativos dos grupos diferentemente dos centróides quando os dados são representados por vetores de atributos; (ii) os índices passam a considerar dissimilaridades fornecidas por várias matrizes simultaneamente; e (iii) as dissimilaridades fornecidas pelas matrizes consideram os pesos estimados localmente ou globalmente. Os métodos desenvolvidos foram validados em vários estudos empíricos. Portanto, este trabalho cumpriu todos os objetivos estabelecidos.

9.1 SUMÁRIO DE RESULTADOS

Diversos experimentos foram realizados com o objetivo de avaliar o desempenho dos métodos desenvolvidos. Primeiro, devido a importância da escolha do índice de validação de agrupamento como função de aptidão, um estudo comparativo foi realizado com onze funções adaptadas a partir de índices tradicionais da literatura. Testes estatísticos foram realizados e confirmaram que as funções abordadas apresentaram diferenças significativas para os índices externos avaliados. Dentre as funções consideradas, o índice da silhueta, o

índice de Xu e a homogeneidade intra-grupo se destacaram. A partir disso, os resultados obtidos pelo índice da silhueta e pela homogeneidade foram selecionados para comparação com os resultados obtidos por outros métodos da literatura. Os métodos desenvolvidos para agrupamento rígido obtiveram desempenho superior de forma significativa em relação a outros métodos úteis na maioria dos casos. Esses resultados suportam a tese de que os métodos híbridos baseados em PSO conseguem melhorar de forma significativa o agrupamento de dados relacionais com múltiplas visões. Além disso, os resultados também mostram que os métodos híbridos tem potencial para ser uma alternativa promissora para lidar com o problema. Outros pesquisadores também obtiveram resultados interessantes ao propor métodos híbridos para agrupamento rígido e nebuloso de dados (FILHO et al., 2015; YANG; SUN; ZHANG, 2009b).

A mesma metodologia foi utilizada para avaliar o desempenho dos métodos desenvolvidos para agrupamento nebuloso. Dez funções de aptidão baseadas em índices para validação de agrupamento nebuloso foram utilizadas. Dentre as funções consideradas, a silhueta simplificada e o coeficiente de partição se destacaram. Os resultados obtidos pelo índice da silhueta simplificada foram utilizados para comparação com outros métodos de agrupamento nebuloso. A abordagem proposta nesta teste apresentou desempenho competitivo e superior de forma significativa para alguns casos.

9.2 CONTRIBUIÇÕES

As principais contribuições desta tese são:

1. Uma revisão sobre agupamento de dados relacionais com múltiplas visões, com ênfase na utilização de múltiplas matrizes;

2. Modelagem de método híbrido baseado em otimização por nuvem de partículas para agrupamento rígido de dados relacionais com única visão.

3. Modelagem de métodos híbridos baseados na otimização por nuvem de partículas para agrupamento rígido de dados relacionais com múltiplas visões.

4. Modelagem de métodos híbridos baseados na otimização por nuvem de partículas para agrupamento nebuloso de dados relacionais com múltiplas visões.

5. Análise, adaptação e estudo comparativo de diversos índices para validação de agrupamento rígido e nebuloso de dados relacionais com múltiplas visões; e

6. Avaliação das abordagens propostas desenvolvidas em comparação com outros algoritmos da literatura usando índices externos para avaliar a qualidade das partições geradas.

9.3 LIMITAÇÕES DO TRABALHO

Apesar das contribuições, este trabalho possui várias limitações. Por exemplo, este trabalho não considerou outros índices para validação de agrupamento presentes na litetura como alguns presentes em (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002). Outro exemplo de limitação está no fato dos métodos não serem capazes de encontrar o número adequados de grupos para as bases de dados automaticamente, isto é, o número de grupos deve ser informado como dado de entrada. Isso pode ser realizado com o uso dos próprios índices de validação como, por exemplo, o índice da silhueta que já foi utilizado para esse fim em outros trabalhos, mas que isso não foi explorado nesta tese. Outro aspecto não explorado neste trabalho foi relacionado a seleção de visões, pois é assumido que o uso de todas as visões simultaneamente pode resultar em melhores agrupamentos. No entanto, algumas visões também podem apresentar informações conflitantes e, consequentemente, degradar o desempenho dos métodos. Dessa forma, a seleção de visões também não foi considerada neste trabalho e isso também é uma limitação.

Dessa forma, é necessária a realização de mais pesquisas que envolvam o tema, tendo em vista que o objetivo não foi o esgotamento sobre esse assunto. Primeiro devido a amplitude e possibilidades diversas de direções de pesquisa. Segundo, devido ao escopo e tempo limitados. Por fim, não foi possível realizar outros experimentos com diferentes parametrizações devido a limitações dos recursos computacionais disponíveis para execução dos experimentos.

9.4 TRABALHOS FUTUROS

Existem várias pesquisas que podem ser realizadas como trabalhos futuros. Primeiramente, a investigação de formas de reduzir a complexidade dos métodos desenvolvidos é muito importante e deve ser feita, pois possibilitará uma melhor exploração do espaço de soluções em menos tempo computacional. Um segundo estudo importante que poderá ser realizado é a investigação de outros métodos para a fase de atualização de posição das partículas bem como outras formas de inicializar as partículas. Outro trabalho futuro interessante pode ser a análise de variações da otimização por nuvem de partículas para verificar o impacto sobre a convergência dos métodos e também sobre a qualidade dos agrupamentos gerados. Outro caminho de pesquisa interessante é o desenvolvimento dos métodos híbridos para otimização multiobjetivo, pois a otimização de um único objetivo resulta em agrupamentos com estruturas particulares. Com a abordagem multiobjetivo é possível gerar vários agrupamentos com estruturas diferentes e, a partir de algum critério, escolher o mais adequado.

Além das possibilidades citadas anteriormente, estudos sobre a aplicação de outros métodos baseados em inteligência de enxame como, por exemplo, otimização por colônia de formigas e colônia de abelhas artificiais também podem ser investigados, pois tais

métodos possuem diferentes propriedades e, consequentemente, o espaço de soluções pode ser explorado de forma diferente. Finalmente, os métodos desenvolvidos são capazes de estimar um peso de relevância para cada visão com o objetivo de levar em consideração a sua importância para a tarefa de agrupamento. Esta estratégia é interessante porque não exige que seja feita uma seleção de visões em uma fase de preprocessamento dos métodos. Contudo, podem existir casos em que a seleção de visões seria benéfica e resultar em melhores agrupamentos, além de reduzir a complexidade da tarefa. Portanto, outra pesquisa futura deverá ser a de investigar formas para realizar a seleção de visões para agrupamento de dados relacionais com múltiplas visões.

9.5 ARTIGOS PRODUZIDOS

9.5.1 Publicados

1. R. P. de Gusmão and F. d. A. T. de Carvalho, "Particle Swarm Optimization applied to relational data clustering,"2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 001690-001695. doi: 10.1109/SMC.2016.7844480

2. R. P. de Gusmão, F.A.T. de Carvalho, "Clustering of multi-view relational data based on particle swarm optimization", Expert Systems with Applications, Volume 123, 2019, Pages 34-53, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.12.053.

9.5.2 Submetidos

1. R. P. de Gusmão, F.A.T. de Carvalho, "An approach based on PSO for fuzzy clustering of multi-view relational data". International Journal of Pattern Recognition and Artificial Intelligence. Submetido em 2018.

REFERÊNCIAS

AHMADYFARD, A.; MODARES, H. Combining pso and k-means to enhance data clustering. In: *2008 International Symposium on Telecommunications*. [S.l.: s.n.], 2008. p. 688–691.

ALAM, S.; DOBBIE, G.; KOH, Y. S.; RIDDLE, P.; REHMAN, S. U. Research on particle swarm optimization based clustering: A systematic review of literature and techniques. *Swarm and Evolutionary Computation*, v. 17, p. 1–13, 2014.

ALSWAITTI, M.; ALBUGHDADI, M.; ISA, N. A. M. Density-based particle swarm optimization algorithm for data clustering. *Expert Systems with Applications*, v. 91, p. 170 – 186, 2018. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417417305912>.

BEZDEK, J. Numerical taxonomy with fuzzy sets. v. 1, p. 57–71, 05 1974.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BHARNE, P. K.; GULHANE, V. S.; YEWALE, S. K. Data clustering algorithms based on swarm intelligence. In: *2011 3rd International Conference on Electronics Computer Technology*. [S.l.: s.n.], 2011. v. 4, p. 407–411.

BICKEL, S.; SCHEFFER, T. Multi-view clustering. In: *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*. [S.l.: s.n.], 2004. p. 19–26.

CAI, X.; NIE, F.; HUANG, H. Multi-view k-means clustering on big data. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2013. (IJCAI '13), p. 2598–2604. ISBN 978-1-57735-633-2.

CAMPELLO, R.; HRUSCHKA, E. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, v. 157, n. 21, p. 2858 – 2875, 2006. ISSN 0165-0114.

CAO, B.; HE, L.; KONG, X.; YU, P. S.; HAO, Z.; RAGIN, A. B. Tensor-based multi-view feature selection with applications to brain diseases. In: *Proceedings of the 2014 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2014. (ICDM '14), p. 40–49. ISBN 978-1-4799-4302-9.

CARVALHO, F. A. T. D.; LECHEVALLIER, Y.; MELO, F. M. D. Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, Elsevier Science Inc., New York, NY, USA, v. 45, n. 1, p. 447–464, jan. 2012. ISSN 0031-3203.

CARVALHO, F. D. A. T. D.; LECHEVALLIER, Y.; MELO, F. M. D. Relational partitioning fuzzy clustering algorithms based on multiple dissimilarity matrices. *Fuzzy Sets Syst.*, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 215, p. 1–28, mar. 2013. ISSN 0165-0114.

Chao, G.; Sun, S.; Bi, J. A Survey on Multi-View Clustering. *ArXiv e-prints*, dez. 2017.

CHEN, X.; XU, X.; HUANG, J. Z.; YE, Y. Tw-k-means: Automated two-level variable weighting clustering algorithm for multiview data. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 4, p. 932–944, April 2013. ISSN 1041-4347.

CHIKHI, N. F. Multi-view clustering via spectral partitioning and local refinement. *Information Processing & Management*, v. 52, n. 4, p. 618 – 627, 2016. ISSN 0306-4573.

CHOU, C.-H.; SU, M.-C.; LAI, E. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, v. 7, n. 2, p. 205–220, Jul 2004. ISSN 1433-755X.

CLERC, M.; KENNEDY, J. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *Trans. Evol. Comp*, IEEE Press, Piscataway, NJ, USA, v. 6, n. 1, p. 58–73, fev. 2002. ISSN 1089-778X.

Cleuziou, G.; Exbrayat, M.; Martin, L.; Sublemontier, J. Cofkm: A centralized method for multiple-view clustering. In: *2009 Ninth IEEE International Conference on Data Mining*. [S.l.: s.n.], 2009. p. 752–757. ISSN 1550-4786.

DAS, S.; ABRAHAM, A.; KONAR, A. *Metaheuristic Clustering*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2009. ISBN 3540921729, 9783540921721.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, April 1979. ISSN 0162-8828.

DEMsAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1248547.1248548>.

DIGAY, E.; GOVAERT, G. Classification Automatique avec Distances Adaptatives. In: *Informatique Computer Science 11 (4)*. [S.l.]: R.A.I.R.O., 1977. p. 329–349.

DIMITRIADOU, E.; DOLNIČAR, S.; WEINGESSEL, A. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, v. 67, n. 1, p. 137–159, Mar 2002. ISSN 1860-0980.

DING, J.; SHAO, J.; HUANG, Y.; SHENG, L.; FU, W.; LI, Y. Swarm intelligence based algorithms for data clustering. In: *Proceedings of 2011 International Conference on Computer Science and Network Technology*. [S.l.: s.n.], 2011. v. 1, p. 577–581.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, Taylor Francis, v. 3, n. 3, p. 32–57, 1973.

Engelbrecht, A. P. Particle swarm optimization: Global best or local best? In: *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*. [S.l.: s.n.], 2013. p. 124–135. ISSN 2377-0589.

FILHO, T. M. S.; PIMENTEL, B. A.; SOUZA, R. M.; OLIVEIRA, A. L. Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization. *Expert Systems with Applications*, v. 42, n. 17, p. 6315 – 6328, 2015. ISSN 0957-4174.

FRIGUI, H.; HWANG, C. Semi-supervised clustering and aggregation of relational data. In: *2008 IEEE Symposium on Computers and Communications*. [S.l.: s.n.], 2008. p. 590–595. ISSN 1530-1346.

FRIGUI, H.; HWANG, C.; RHEE, F. C.-H. Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recogn.*, Elsevier Science Inc., New York, NY, USA, v. 40, n. 11, p. 3053–3068, nov. 2007. ISSN 0031-3203.

FUKUYAMA, Y.; SUGENO, M. A new method of choosing the number of clusters for the fuzzy c-means method. In: *Proceeding of fifth Fuzzy Syst. Symp.* [S.l.: s.n.], 1989. p. 247–250.

GENDREAU, M.; POTVIN, J.-Y. *Handbook of Metaheuristics*. 2nd. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 1441916636, 9781441916631.

GUSMãO, R. P.; CARVALHO, F. A. T. de. Particle swarm optimization applied to relational data clustering. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.: s.n.], 2016. p. 001690–001695.

GUSMãO, R. P. de; CARVALHO, F. de A.T. de. Clustering of multi-view relational data based on particle swarm optimization. *Expert Systems with Applications*, v. 123, p. 34 – 53, 2019. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417418308248>.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. *Journal of Intelligent Information Systems*, v. 17, n. 2, p. 107–145, Dec 2001. ISSN 1573-7675.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity checking methods: Part ii. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 31, n. 3, p. 19–27, set. 2002. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/601858.601862>.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining Concepts and Techniques*. 3rd. ed. [S.l.]: Morgan Kaufmann Publishers, 2012.

HASAN, M. J. A.; RAMAKRISHNAN, S. A survey: hybrid evolutionary algorithms for cluster analysis. *Artif. Intell. Rev.*, v. 36, n. 3, p. 179–204, 2011.

HATHAWAY, R. J.; BEZDEK, J. C. Nerf c-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, v. 27, n. 3, p. 429 – 437, 1994. ISSN 0031-3203.

HATHAWAY, R. J.; DAVENPORT, J. W.; BEZDEK, J. C. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, v. 22, n. 2, p. 205–212, 1989.

HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, v. 6, p. 65–70, 1979.

HORTA, D. *Algoritmos e técnicas de validação em agrupamento de dados multi-representados, agrupamento possibilístico e bi-agrupamento*. Tese (PhD dissertation) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013.

HORTA, D.; ANDRADE, I. C. de; CAMPELLO, R. J. Evolutionary fuzzy clustering of relational data. *Theoretical Computer Science*, v. 412, n. 42, p. 5854 – 5870, 2011. ISSN 0304-3975. Rough Sets and Fuzzy Sets in Natural Computing.

HORTA, D.; CAMPELLO, R. J. G. B. Evolutionary clustering of relational data. *Int. J. Hybrid Intell. Syst.*, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 7, n. 4, p. 261–281, dez. 2010. ISSN 1448-5869. Disponível em: <http://dl.acm.org/citation.cfm?id=1923391.1923395>.

HRUSCHKA, E. R.; CAMPELLO, R. J.; CASTRO, L. N. de. Evolving clusters in gene-expression data. *Information Sciences*, v. 176, n. 13, p. 1898 – 1927, 2006. ISSN 0020-0255.

HRUSCHKA, E. R.; CAMPELLO, R. J. G. B.; FREITAS, A. A.; CARVALHO, A. C. P. L. F. D. A survey of evolutionary algorithms for clustering. *Trans. Sys. Man Cyber Part C*, IEEE Press, Piscataway, NJ, USA, v. 39, n. 2, p. 133–155, mar. 2009. ISSN 1094-6977.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985. ISSN 1432-1343.

IZAKIAN, H.; ABRAHAM, A. Fuzzy c-means and fuzzy swarm for fuzzy clustering problem. *Expert Systems with Applications*, v. 38, n. 3, p. 1835 – 1838, 2011. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417410007402>.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651 – 666, 2010. ISSN 0167-8655. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)19th International Conference in Pattern Recognition (ICPR).

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 31, n. 3, p. 264–323, set. 1999. ISSN 0360-0300.

JIANG, B.; QIU, F.; WANG, L. Multi-view clustering via simultaneous weighting on views and features. *Applied Soft Computing*, v. 47, p. 304 – 315, 2016. ISSN 1568-4946.

JIANG, B.; QIU, F.; WANG, L.; ZHANG, Z. Bi-level weighted multi-view clustering via hybrid particle swarm optimization. *Information Processing & Management*, v. 52, n. 3, p. 387 – 398, 2016. ISSN 0306-4573.

JIANG, B.; QIU, F.; YANG, S.; WANG, L. Evolutionary multi-objective optimization for multi-view clustering. In: *2016 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2016. p. 3308–3315.

JORDEHI, A. R.; JASNI, J. Parameter selection in particle swarm optimisation: a survey. *Journal of Experimental & Theoretical Artificial Intelligence*, Taylor Francis, v. 25, n. 4, p. 527–542, 2013.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data : an introduction to cluster analysis*. New York: Wiley, 1990. (Wiley series in probability and mathematical statistics). A Wiley-Interscience publication. ISBN 0-471-87876-6.

KENNEDY, J.; EBERHART, R. C. Particle swarm optimization. In: *Proceedings of the IEEE International Conference on Neural Networks*. [S.l.: s.n.], 1995. p. 1942–1948.

KRISHNAPURAM, R.; FREG, C.-P. Fitting an unknown number of lines and planes to image data through compatible cluster merging. *Pattern Recognition*, v. 25, n. 4, p. 385 – 400, 1992. ISSN 0031-3203.

KRISHNAPURAM, R.; JOSHI, A.; YI, L. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In: *Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99. 1999 IEEE International*. [S.l.: s.n.], 1999. v. 3, p. 1281–1286 vol.3. ISSN 1098-7584.

KUMAR, A.; III, H. D. A co-training approach for multi-view spectral clustering. In: GETOOR, L.; SCHEFFER, T. (Ed.). *ICML*. [S.l.]: Omnipress, 2011. p. 393–400.

KUMAR, A.; RAI, P.; DAUMÉ III, H. Co-regularized multi-view spectral clustering. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. USA: Curran Associates Inc., 2011. (NIPS'11), p. 1413–1421. ISBN 978-1-61839-599-3.

LI, Y.; NIE, F.; HUANG, H.; HUANG, J. Large-scale multi-view spectral clustering via bipartite graph. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2015. (AAAI'15), p. 2750–2756. ISBN 0-262-51129-0.

LIU, J.; WANG, C.; GAO, J.; HAN, J. Multi-View Clustering via Joint Nonnegative Matrix Factorization. In: *Proc. of 2013 SIAM Data Mining Conf. (SDM'13)*. [S.l.: s.n.], 2013.

LIU, R.; SUN, X.; JIAO, L.; LI, Y. A comparative study of different cluster validity indexes. *Transactions of the Institute of Measurement and Control*, v. 34, n. 7, p. 876–890, 2012.

LIU, Y.; LI, Z.; XIONG, H.; GAO, X.; WU, J. Understanding of internal clustering validation measures. In: *2010 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2010. p. 911–916. ISSN 1550-4786.

LONG, B.; YU, P. S.; ZHANG, Z. M. A general model for multiple view unsupervised learning. In: *SDM*. [S.l.]: SIAM, 2008. p. 822–833. ISBN 978-1-61197-278-8.

LONG, B.; ZHANG, Z.; YU, P. S. *Relational Data Clustering: Models, Algorithms, and Applications*. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2010. ISBN 1420072617, 9781420072617.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005. ISBN 0387244352, 9780387244358.

MAULIK, U.; BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 12, p. 1650–1654, Dec 2002. ISSN 0162-8828.

MEI, J.-P.; CHEN, L. Fuzzy clustering with weighted medoids for relational data. *Pattern Recognition*, v. 43, n. 5, p. 1964 – 1974, 2010. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S003132030900452X>.

MEI, J.-P.; CHEN, L. Fuzzy relational clustering around medoids: A unified view. *Fuzzy Sets and Systems*, v. 183, n. 1, p. 44 – 56, 2011. ISSN 0165-0114. Theme : Information processing.

MEILa, M. The uniqueness of a good optimum for k-means. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM, 2006. (ICML '06), p. 625–632. ISBN 1-59593-383-2.

MERWE, D. W. van der; ENGELBRECHT, A. P. Data clustering using particle swarm optimization. In: *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*. [S.l.: s.n.], 2003. v. 1, p. 215–220 Vol.1.

MILLER, R. G. *Beyond ANOVA: Basics of Applied Statistics (Texts in Statistical Science Series)*. Chapman & Hall/CRC, 1997. Hardcover. ISBN 0412070111. Disponível em: <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0412070111>.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, v. 50, n. 2, p. 159–179, Jun 1985. ISSN 1860-0980.

MURTY, M.; DEVI, V. *Introduction to Pattern Recognition and Machine Learning*. [S.l.]: World Scientific, 2015. (IISc lecture notes series). ISBN 9789814335454.

MURTY, N.; DEVI, V. S. *Pattern recognition : an algorithmic approach*. New York: Springer, 2011. (Undergraduate topics in computer science). ISBN 978-0-85729-494-4.

NALDI, M. C.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R.; CARVALHO, A. C. P. L. F. Efficiency issues of evolutionary k-means. *Applied Soft Computing*, v. 11, n. 2, p. 1938–1952, 2011.

NANDA, S. J.; PANDA, G. A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary Computation*, v. 16, p. 1–18, 2014.

PIO, G.; SERAFINO, F.; MALERBA, D.; CECI, M. Multi-type clustering and classification from heterogeneous networks. *Information Sciences*, v. 425, p. 107 – 126, 2018. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025516321570>.

RAITOHARJU, J.; SAMIEE, K.; KIRANYAZ, S.; GABBOUJ, M. Particle swarm clustering fitness evaluation with computational centroids. *Swarm and Evolutionary Computation*, Elsevier, v. 34, p. 103–118, 2 2017. ISSN 2210-6502.

RANA, S.; JASOLA, S.; KUMAR, R. A review on particle swarm optimization algorithms and their applications to data clustering. *Artif. Intell. Rev.*, Kluwer Academic Publishers, Norwell, MA, USA, v. 35, n. 3, p. 211–222, mar. 2011. ISSN 0269-2821.

RIJSBERGEN, C. J. V. *Information Retrieval*. 2nd. ed. Newton, MA, USA: Butterworth-Heinemann, 1979. ISBN 0408709294.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. Supplement C, p. 53 – 65, 1987. ISSN 0377-0427.

SHI, Y.; EBERHART, R. C. Parameter selection in particle swarm optimization. In: _____. *Evolutionary Programming VII: 7th International Conference, EP98 San Diego, California, USA, March 25–27, 1998 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 591–600. ISBN 978-3-540-68515-9.

SNEATH, P. H. A.; SOKAL, R. R. *Numerical taxonomy : the principles and practice of numerical classification*. San Francisco: W. H. Freeman, 1973. (A Series of books in biology). ISBN 0-7167-0697-0.

Toreini, E.; Mehrnejad, M. Clustering data with particle swarm optimization using a new fitness. In: *2011 3rd Conference on Data Mining and Optimization (DMO)*. [S.l.: s.n.], 2011. p. 266–270. ISSN 2155-6946.

TRELEA, I. C. The particle swarm optimization algorithm: Convergence analysis and parameter selection. *Inf. Process. Lett.*, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 85, n. 6, p. 317–325, mar. 2003. ISSN 0020-0190.

TZORTZIS, G.; LIKAS, A. Convex mixture models for multi-view clustering. In: _____. *Artificial Neural Networks – ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 205–214. ISBN 978-3-642-04277-5.

TZORTZIS, G.; LIKAS, A. Kernel-based weighted multi-view clustering. In: *2012 IEEE 12th International Conference on Data Mining*. [S.l.: s.n.], 2012. p. 675–684. ISSN 1550-4786.

WANG, Q.; DOU, Y.; LIU, X.; LV, Q.; LI, S. Multi-view clustering with extreme learning machine. *Neurocomputing*, v. 214, p. 483 – 494, 2016. ISSN 0925-2312.

WANG, Y.; CHEN, L. Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources. *Expert Systems with Applications*, v. 72, p. 457 – 466, 2017. ISSN 0957-4174.

WANG, Y.; CHEN, L.; MEI, J. Incremental fuzzy clustering with multiple medoids for large data. *IEEE Transactions on Fuzzy Systems*, v. 22, n. 6, p. 1557–1568, Dec 2014. ISSN 1063-6706.

XIE, X. L.; BENI, G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 13, n. 8, p. 841–847, Aug 1991. ISSN 0162-8828.

XU, J.; HAN, J.; NIE, F.; LI, X. Re-weighted discriminatively embedded k-means for multi-view clustering. *IEEE Transactions on Image Processing*, PP, n. 99, p. 1–1, 2017. ISSN 1057-7149.

XU, L. Bayesian ying–yang machine, clustering and number of clusters. *Pattern Recognition Letters*, v. 18, n. 11, p. 1167 – 1178, 1997. ISSN 0167-8655.

XU, R.; XU, J.; WUNSCH, D. C. A comparison study of validity indices on swarm-intelligence-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, v. 42, n. 4, p. 1243–1256, Aug 2012. ISSN 1083-4419.

XU, Y.-M.; WANG, C.-D.; LAI, J.-H. Weighted multi-view clustering with feature selection. *Pattern Recognition*, v. 53, p. 25 – 35, 2016. ISSN 0031-3203.

YANG, F.; SUN, T.; ZHANG, C. An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. *Expert Systems with Applications*, v. 36, n. 6, p. 9847 – 9852, 2009. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417409001407>.

YANG, F.; SUN, T.; ZHANG, C. An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. *Expert Systems with Applications*, v. 36, n. 6, p. 9847 – 9852, 2009. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417409001407>.

YANG, Y.; WANG, H. Multi-view clustering: A survey. *Big Data Mining and Analytics*, v. 1, n. 2, p. 83–107, June 2018. ISSN 2096-0654.

YIN, Q.; WU, S.; HE, R.; WANG, L. Multi-view clustering via pairwise sparse subspace representation. *Neurocomputing*, v. 156, p. 12 – 21, 2015. ISSN 0925-2312.

ZELNIK-MANOR, L.; PERONA, P. Self-tuning spectral clustering. In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2004. (NIPS'04), p. 1601–1608. Disponível em: <http://dl.acm.org/citation.cfm?id=2976040.2976241>.

ZHANG, X.; WANG, Z.; ZONG, L.; YU, H. Multi-view clustering via graph regularized symmetric nonnegative matrix factorization. In: *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. [S.l.: s.n.], 2016. p. 109–114.

ZHANG, X.; ZHAO, L.; ZONG, L.; LIU, X.; YU, H. Multi-view clustering via multi-manifold regularized nonnegative matrix factorization. In: *2014 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2014. p. 1103–1108. ISSN 1550-4786.

ZHAO, Q.; XU, M.; FRÄNTI, P. Sum-of-squares based cluster validity index and significance analysis. In: _____. *Adaptive and Natural Computing Algorithms: 9th International Conference, ICANNGA 2009, Kuopio, Finland, April 23-25, 2009, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 313–322. ISBN 978-3-642-04921-7.

ZHU, Y.; ZHU, X.; ZHENG, W. Robust multi-view learning via half-quadratic minimization. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 2018. p. 3278–3284. Disponível em: <https://doi.org/10.24963/ijcai.2018/455>.

APÊNDICE A – GRÁFICOS BOX-PLOTS DAS FUNÇÕES DE APTIDÃO

Nesta seção são apresentados os gráficos box-plots gerados para os resultados obtidos na comparação das funções de aptidão. As caixas em vermelho são relativas ao modelo para dados com única visão, as caixas em azul são relativas aos modelos com pesos locais e as caixas em laranja são referentes aos modelos que usam pesos globais.

A.0.1 Agrupamento rígido

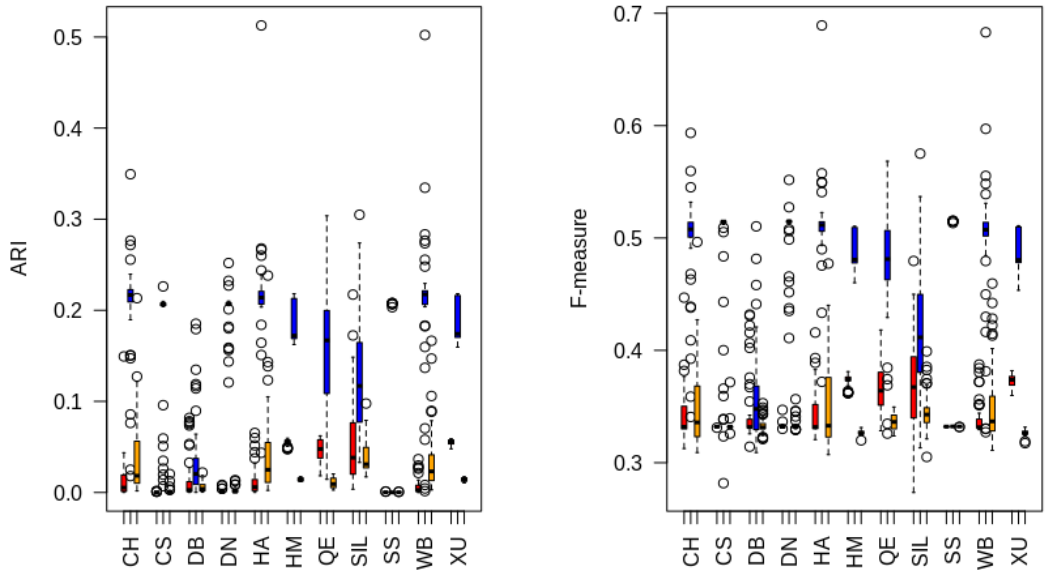


Figura 48 – Gráfico Box-plot das funções de aptidão para base Animals-1

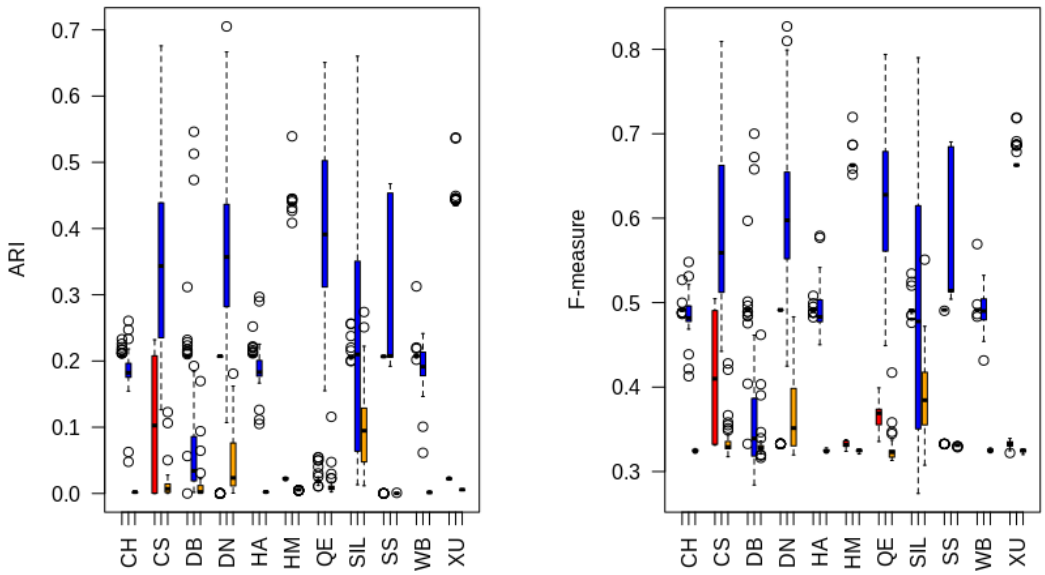


Figura 49 – Gráfico Box-plot das funções de aptidão para base Animals-2

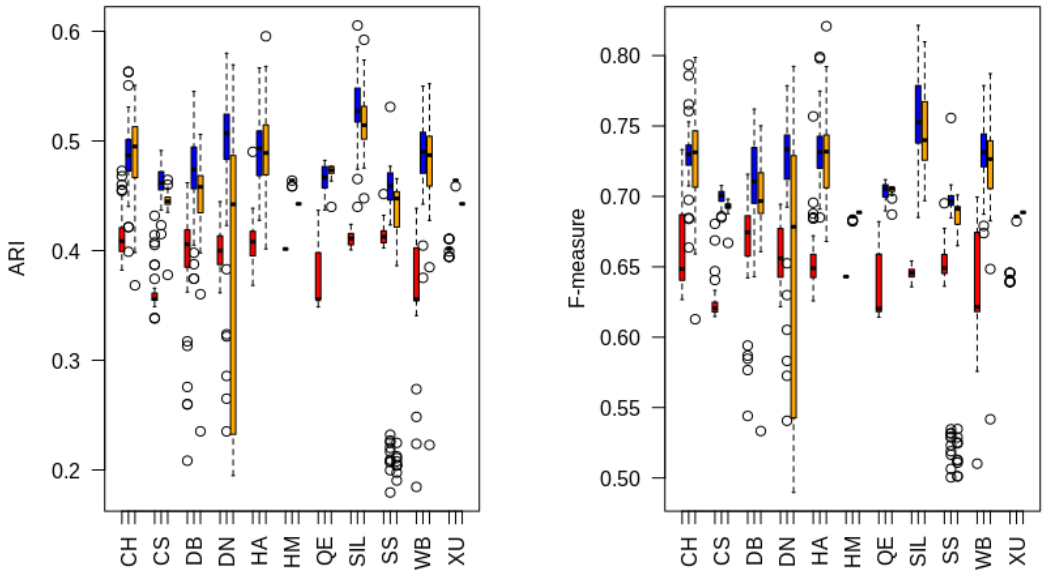


Figura 50 – Gráfico Box-plot das funções de aptidão para base Corel-1

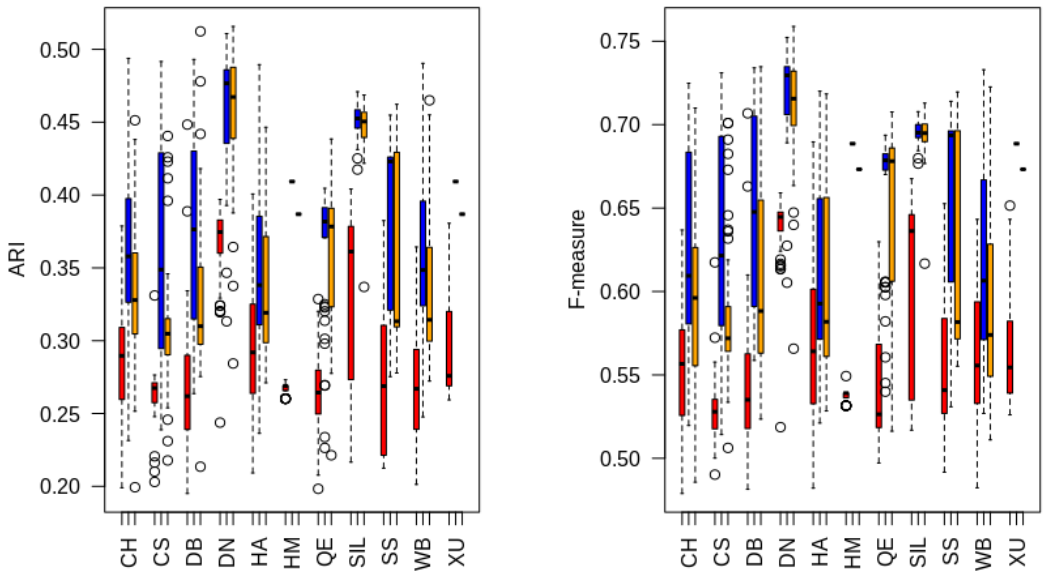


Figura 51 – Gráfico Box-plot das funções de aptidão para base Corel-2

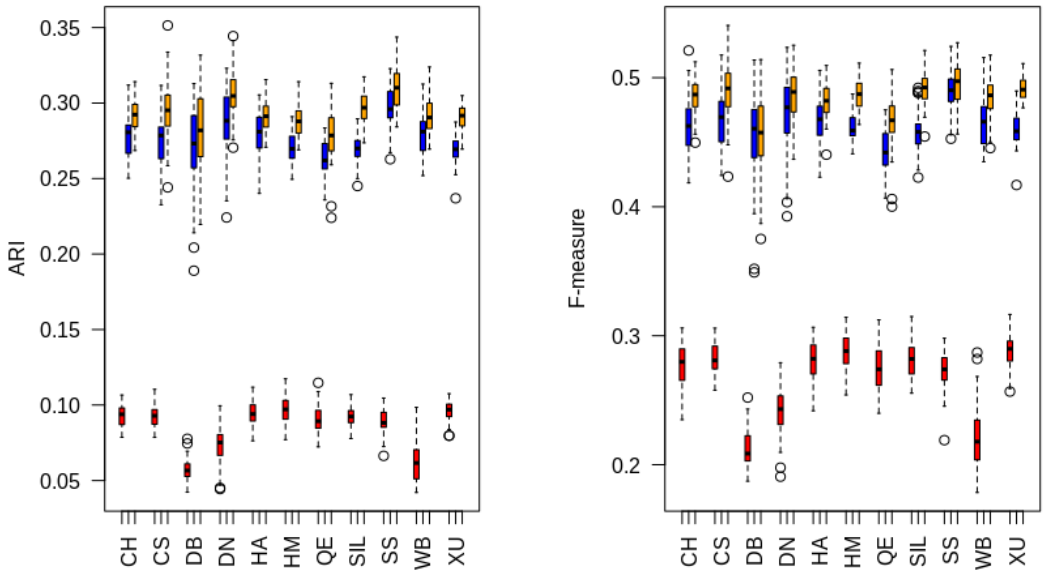


Figura 52 – Gráfico Box-plot das funções de aptidão para base Flowers

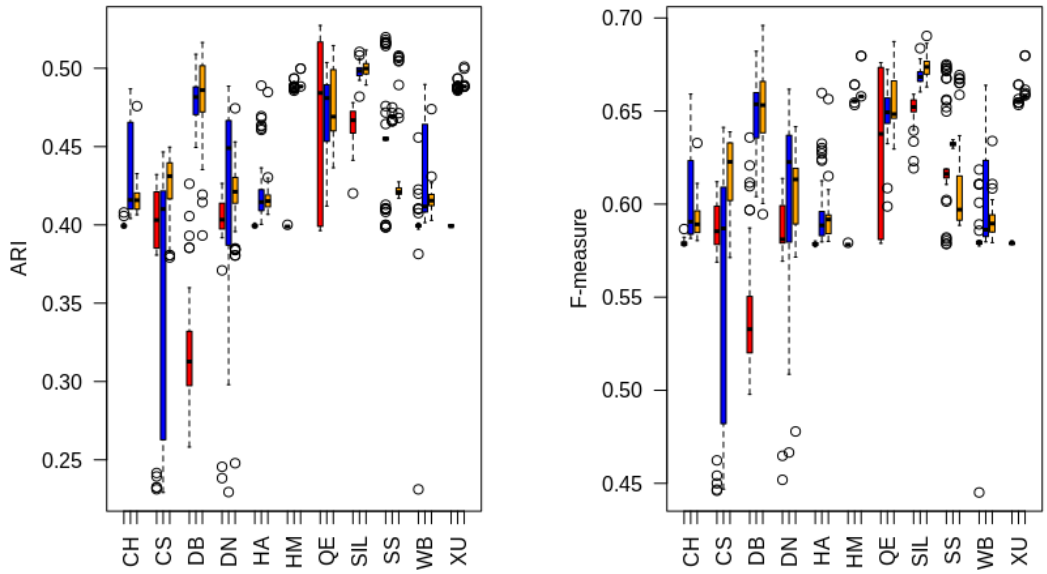


Figura 53 – Gráfico Box-plot das funções de aptidão para base Image

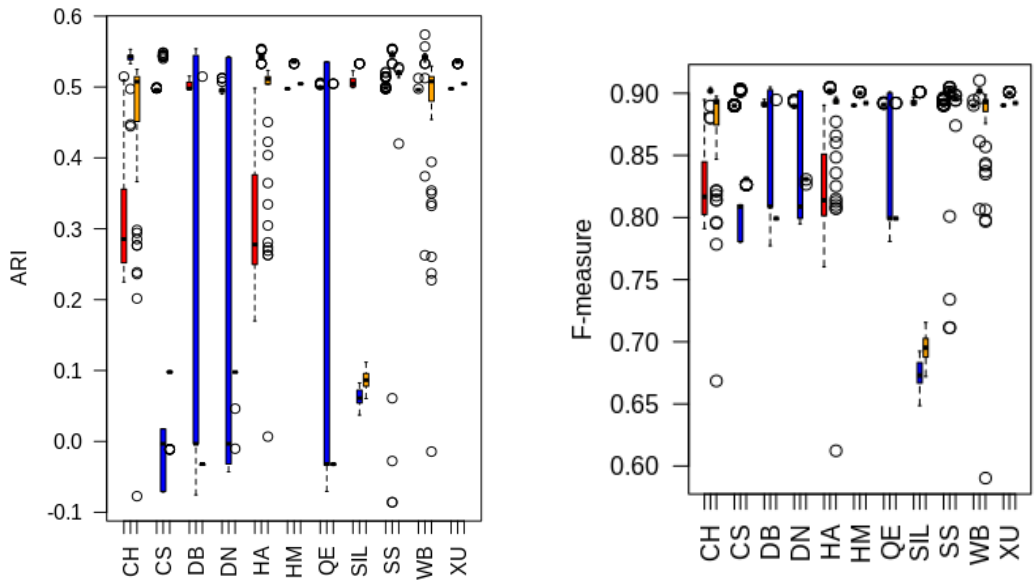


Figura 54 – Gráfico Box-plot das funções de aptidão para base Internet

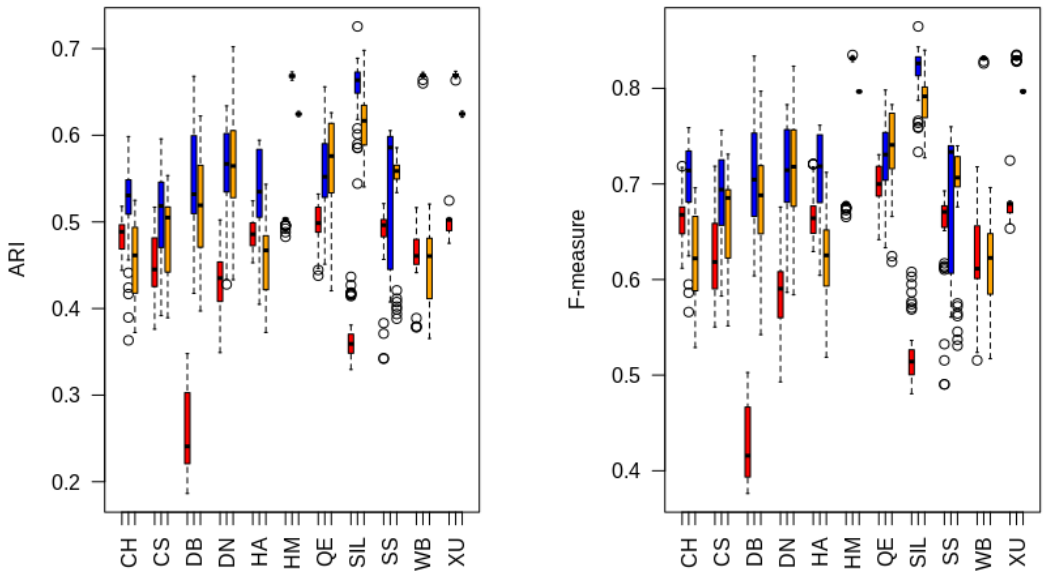


Figura 55 – Gráfico Box-plot das funções de aptidão para base Mfeat

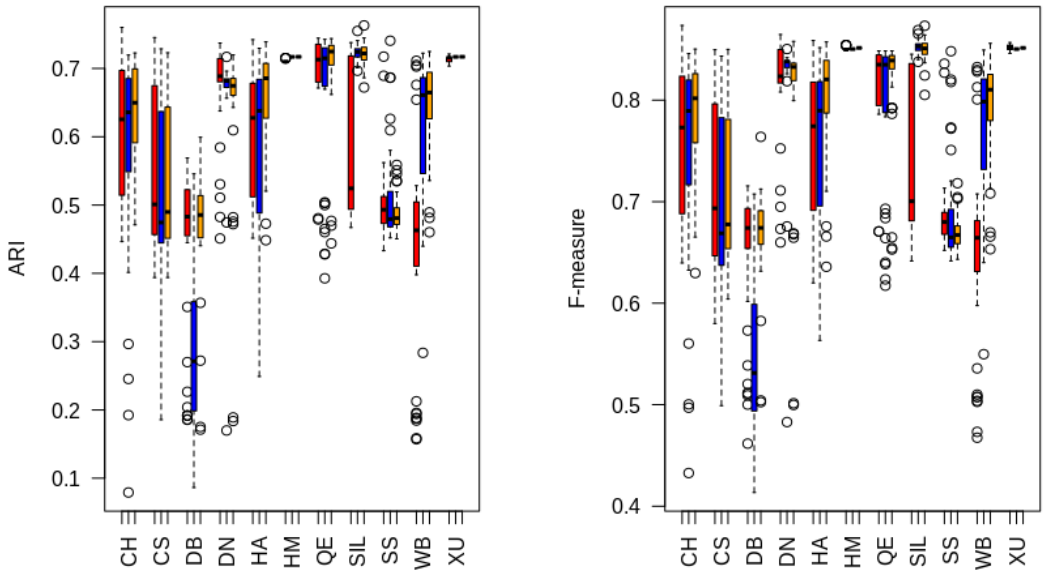


Figura 56 – Gráfico Box-plot das funções de aptidão para base Phoneme

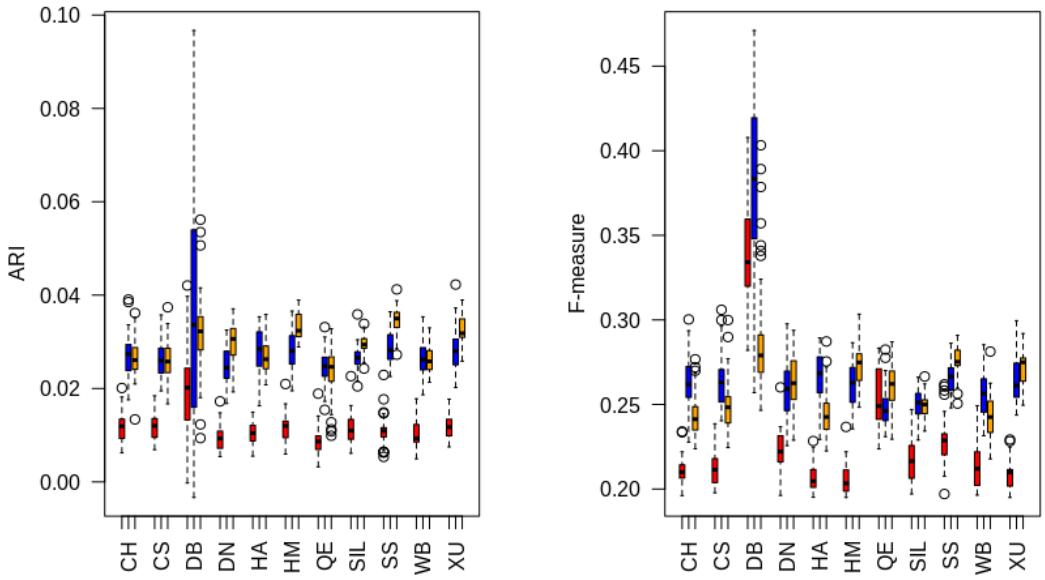


Figura 57 – Gráfico Box-plot das funções de aptidão para base Water

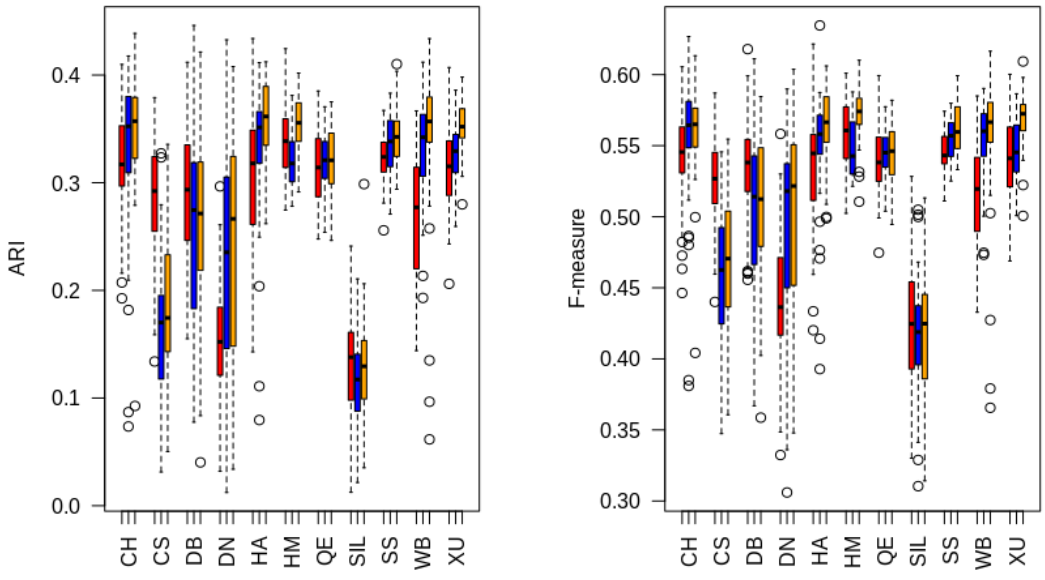


Figura 58 – Gráfico Box-plot das funções de aptidão para base 3-Sources

A.1 AGRUPAMENTO NEBULOSO

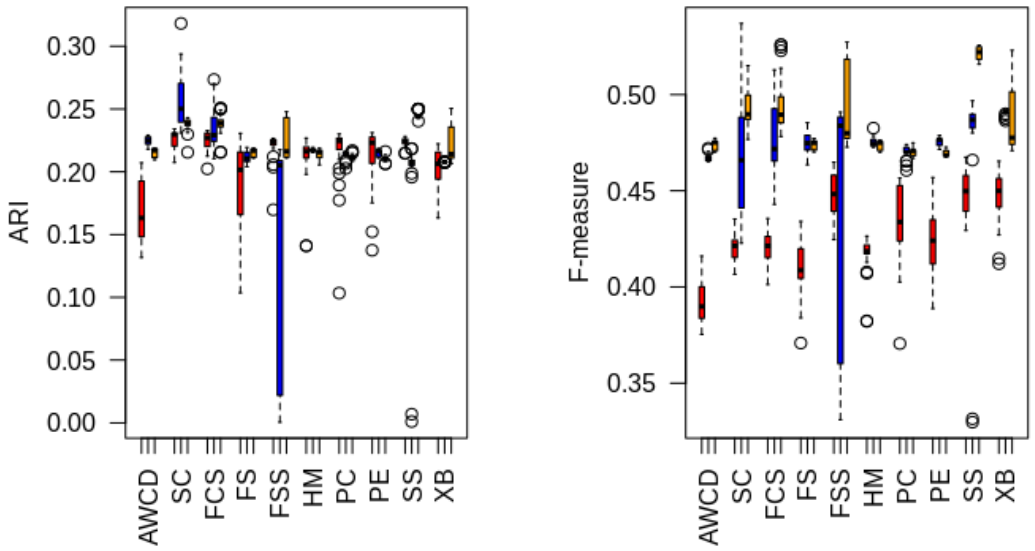


Figura 59 – Gráfico Box-plot das funções de aptidão para base Animals-2

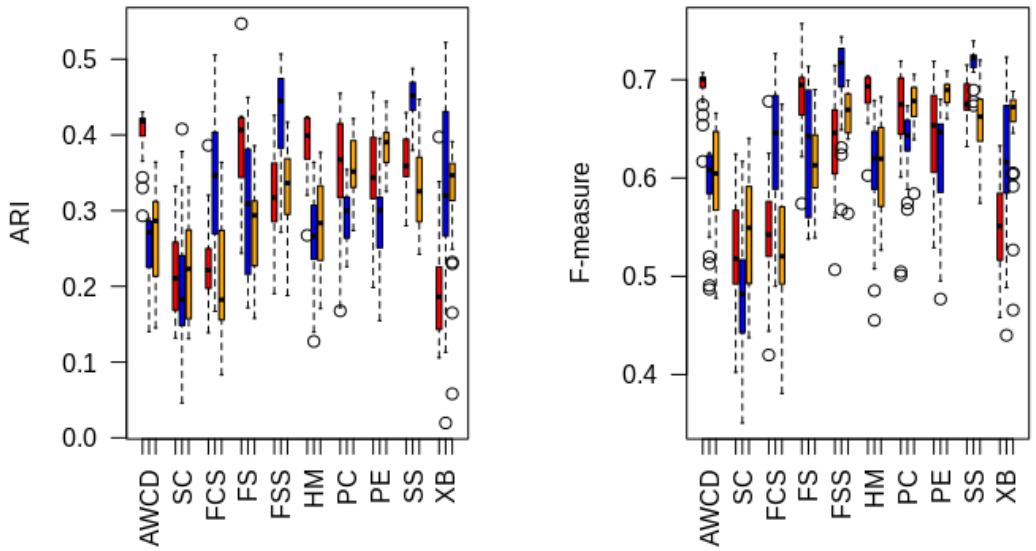


Figura 60 – Gráfico Box-plot das funções de aptidão para base Caltech101-7

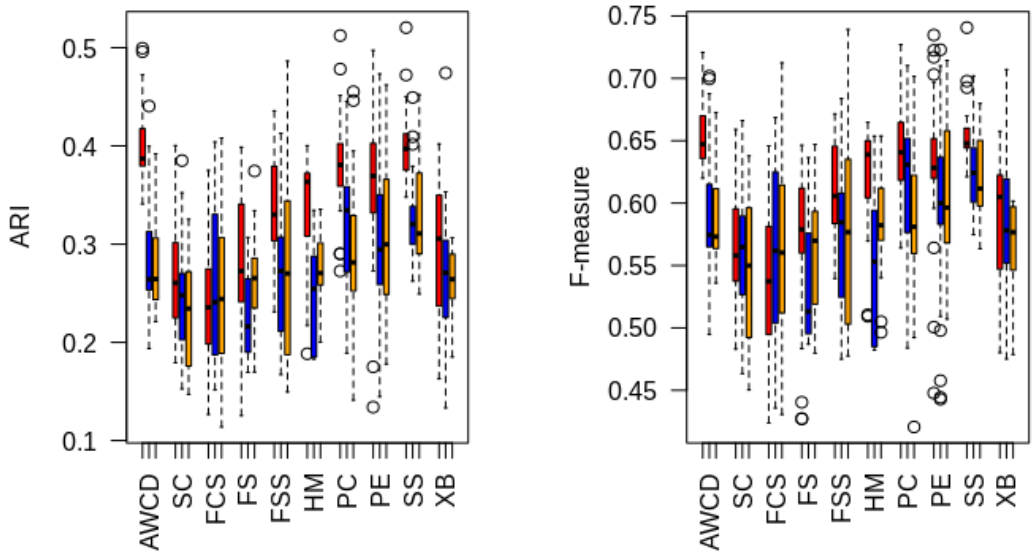


Figura 61 – Gráfico Box-plot das funções de aptidão para base Corel-1

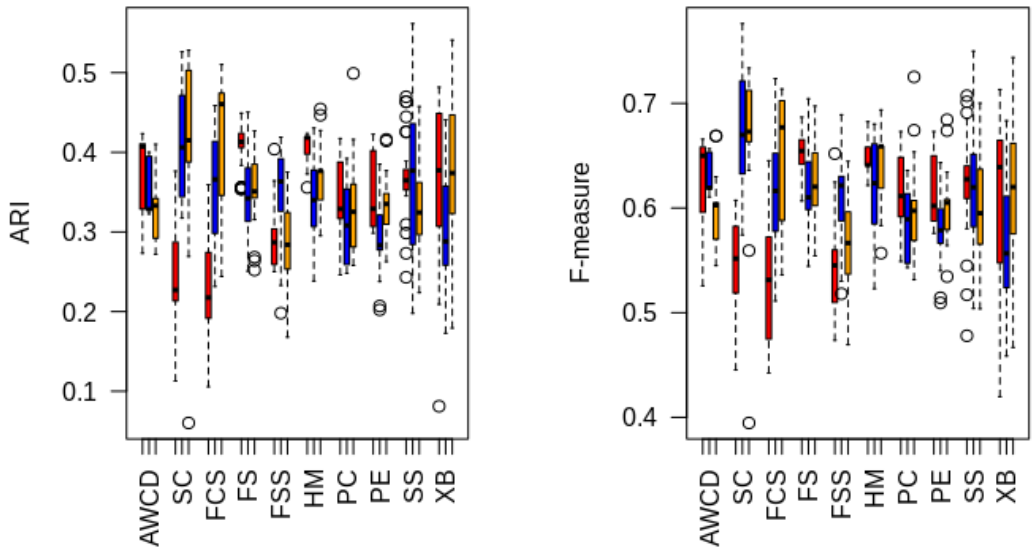


Figura 62 – Gráfico Box-plot das funções de aptidão para base Corel-2

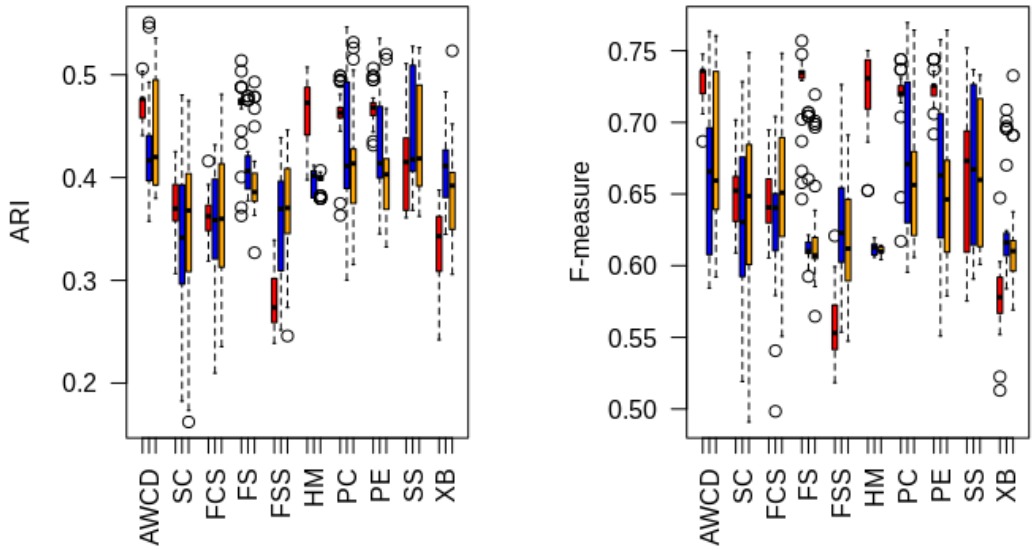


Figura 63 – Gráfico Box-plot das funções de aptidão para base Corel-3

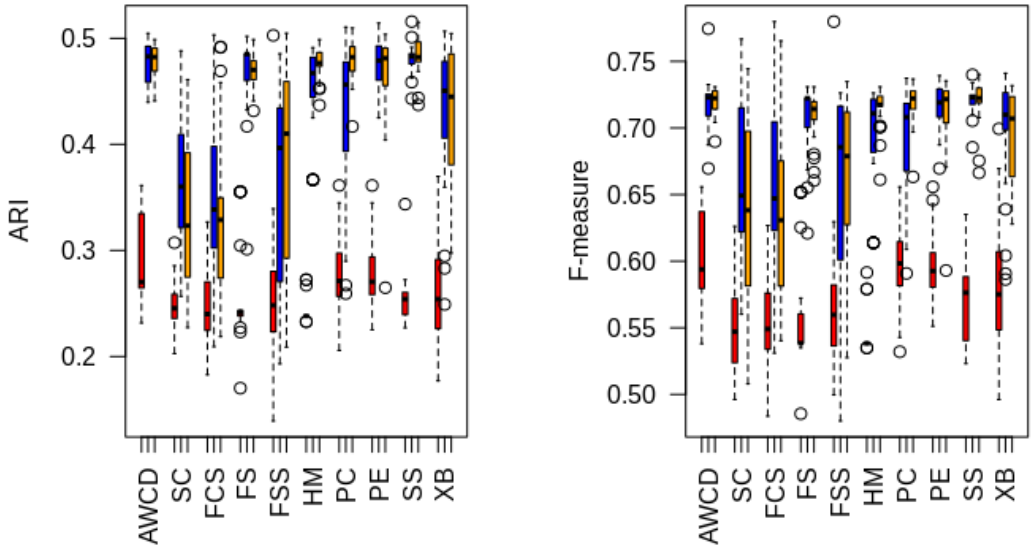


Figura 64 – Gráfico Box-plot das funções de aptidão para base Corel-4

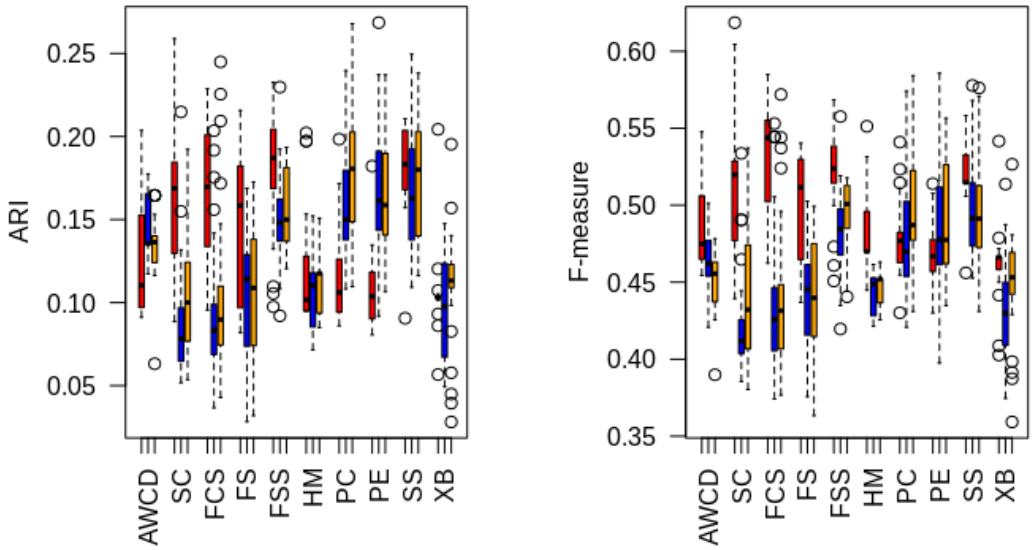


Figura 65 – Gráfico Box-plot das funções de aptidão para base Corel-5

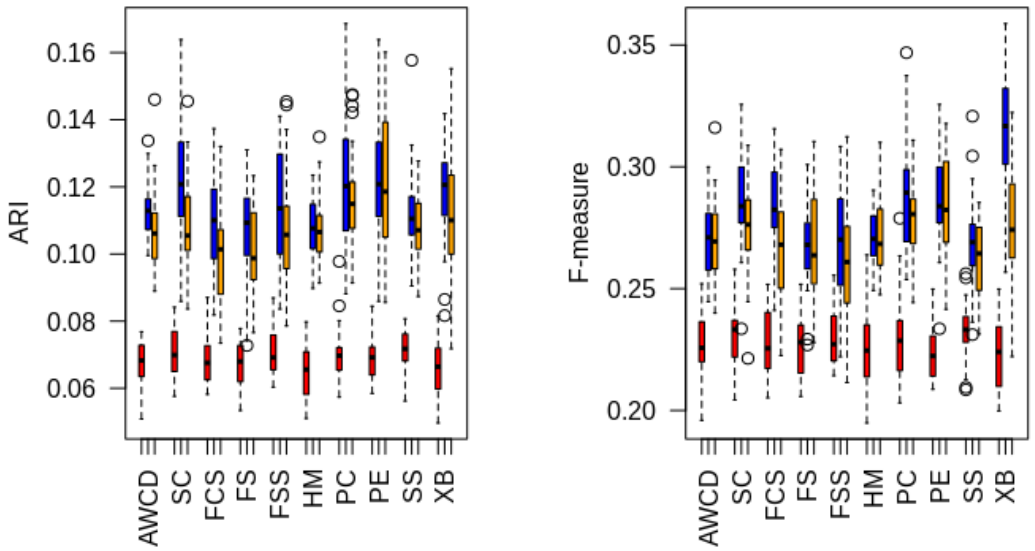


Figura 66 – Gráfico Box-plot das funções de aptidão para base Flowers

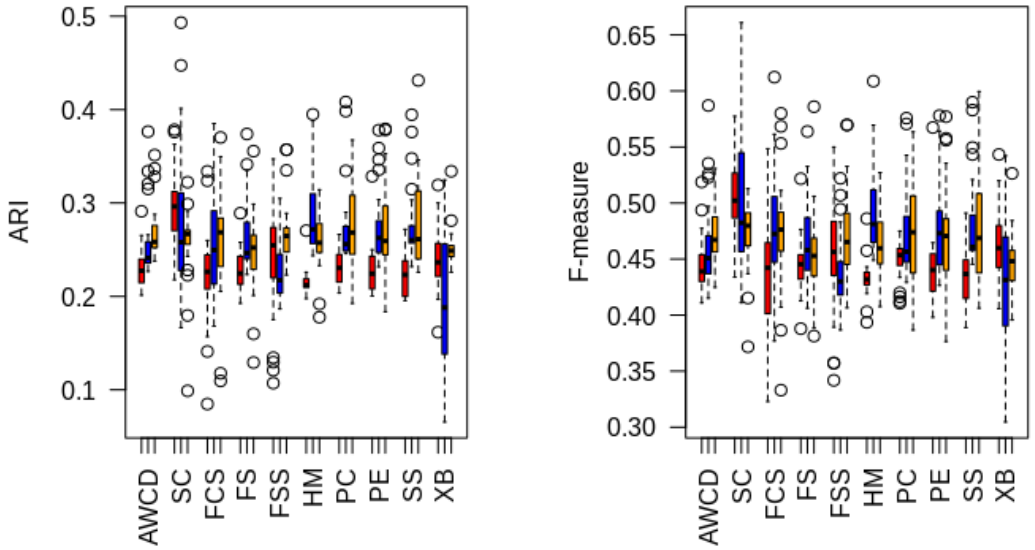


Figura 67 – Gráfico Box-plot das funções de aptidão para base Image

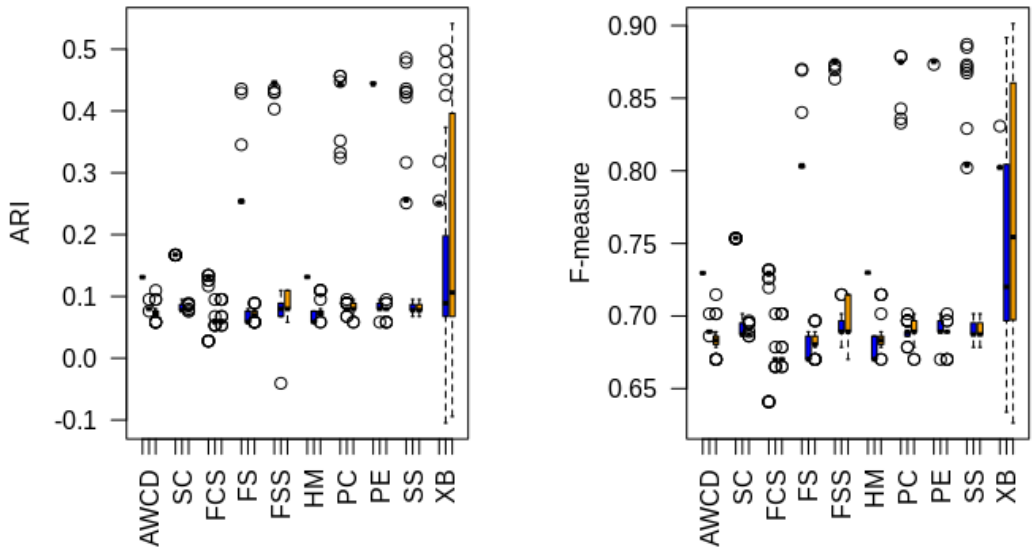


Figura 68 – Gráfico Box-plot das funções de aptidão para base Internet

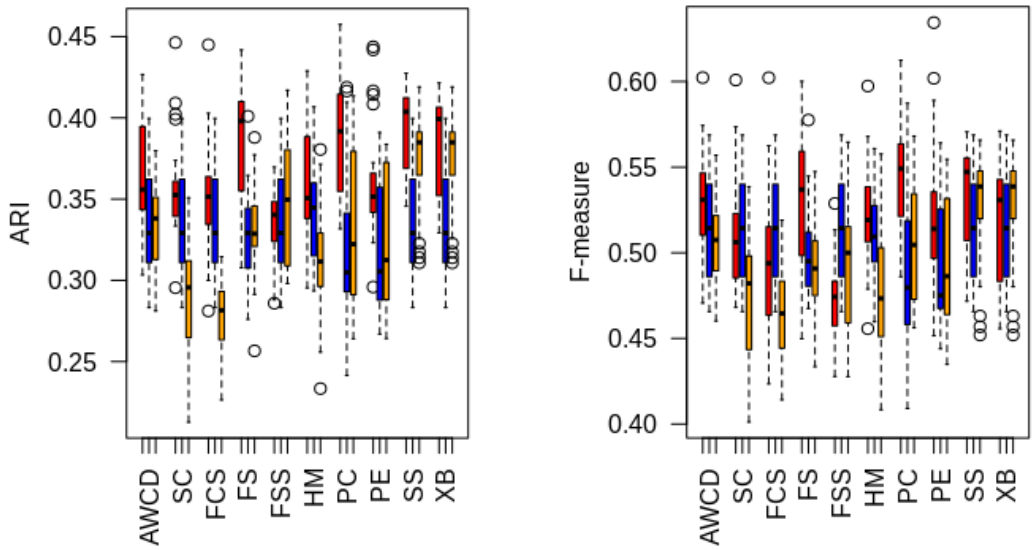


Figura 69 – Gráfico Box-plot das funções de aptidão para base Mfeat

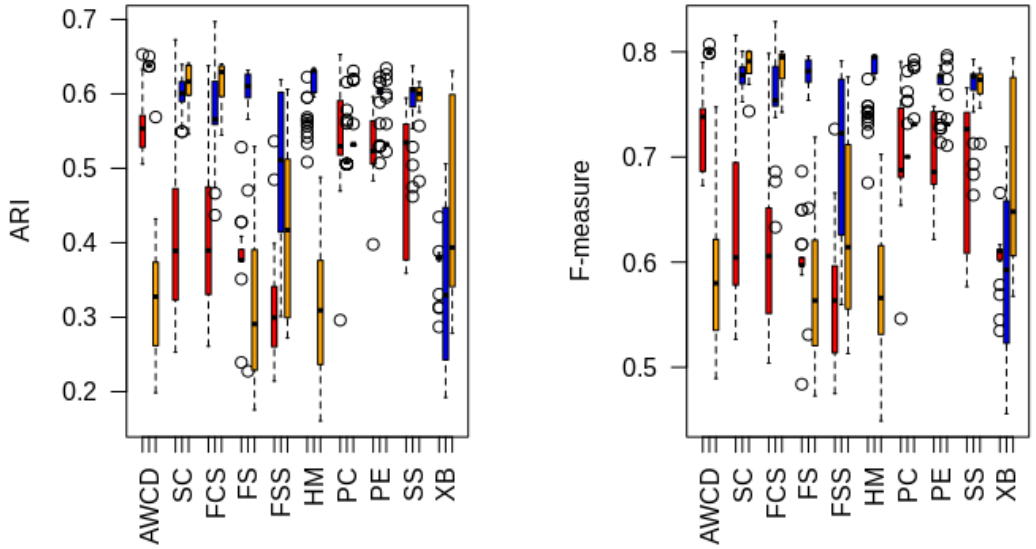


Figura 70 – Gráfico Box-plot das funções de aptidão para base Phoneme

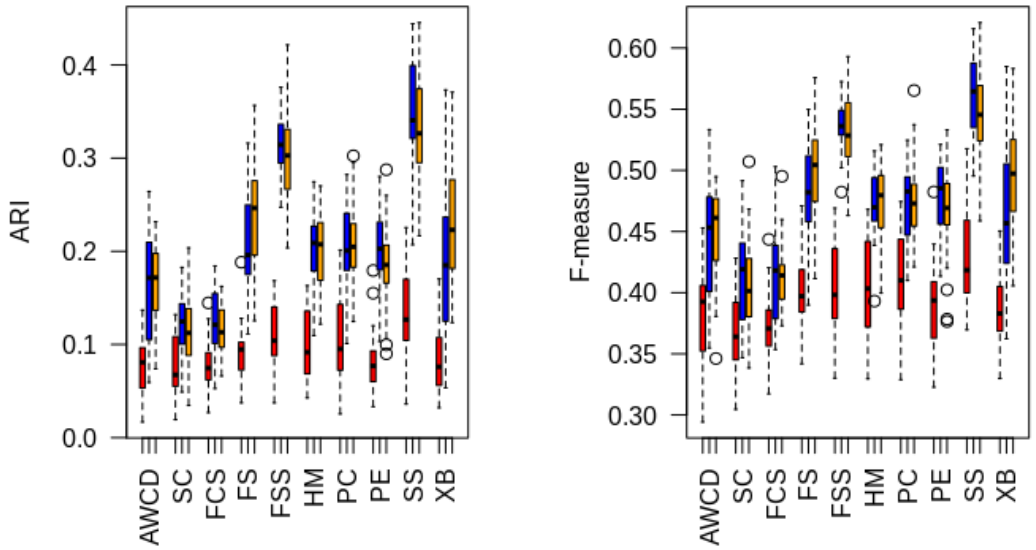


Figura 71 – Gráfico Box-plot das funções de aptidão para base 3-Sources

APÊNDICE B – ANÁLISE PARAMÉTRICA

Com o objetivo de avaliar o impacto causado pelo número de partículas no desempenho, o método PSO-RWL e duas bases, 3-Sources e Water, foram selecionados. Além disso, de todas as funções de aptidão consideradas, quatro foram escolhidas para realizar este estudo. O número de iterações foi definido com valor igual a 100. Os fatores de aceleração definidos foram $c_1 = c_2 = 2$. Os pesos mínimos de inércia foram $w_{min} = 0.4$ e $w_{max} = 0.9$. As Tabelas 56 e 57 apresentam os resultados obtidos em termos do valor da função de aptidão, medida F e índice ajustado de rand. O método foi executado 50 vezes. O melhor valor da função de aptidão encontrado é apresentado e os índices externos computados para a solução com melhor aptidão são apresentados. Além disso, as médias e desvios dos índices também são apresentados.

Como esperado, percebe-se que, com o aumento do número de partículas, o valor médio e o melhor valor das funções de aptidão, considerando todas as execuções, melhoram. Contudo, isso não é refletido necessariamente nos valores ds índices externos, pois a otimização de uma função de aptidão pode não levar a agrupamentos iguais ao agrupamento *a priori* de uma base. Outra observação que pode ser feita é que, em termos de melhor solução encontrada, para alguns casos, o enxame tendo 40 partículas consegue encontrar soluções próximas ou até melhores do que quando $n_p = 50$. Por exemplo, para as funções CS e HM considerando a base 3-Sources.

Tabela 56 – Resultados para a base 3-Sources

n_p	CS		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	1,4535	1,6873(0,1281)	0,5436	0,5352(0,0212)	0,3272	0,3050(0,0368)
20	1,3589	1,6377(0,1178)	0,5420	0,5394(0,0202)	0,3173	0,3175(0,0377)
30	1,4055	1,6162(0,0931)	0,5448	0,5376(0,0155)	0,3071	0,3095(0,0286)
40	1,3356	1,5906(0,0837)	0,5417	0,5365(0,0183)	0,3526	0,3135(0,0318)
50	1,3520	1,5583(0,0949)	0,5231	0,5346(0,0153)	0,2953	0,3092(0,0277)
n_p	DB		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	8,4884	10,1060(0,5988)	0,5588	0,5225(0,0406)	0,3216	0,2774(0,0650)
20	8,4214	9,8364(0,6531)	0,5742	0,5197(0,0407)	0,3754	0,2720(0,0640)
30	7,4370	9,5460(0,7248)	0,5781	0,5267(0,0529)	0,3434	0,2832(0,0755)
40	7,7202	9,4588(0,7251)	0,5600	0,5187(0,0551)	0,3581	0,2787(0,0851)
50	7,7268	9,2158(0,6049)	0,5388	0,5269(0,0428)	0,2923	0,2898(0,0565)
n_p	HM		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	16,3565	16,5357(0,0829)	0,5359	0,5507(0,0228)	0,3210	0,3284(0,0333)
20	16,3775	16,5123(0,0629)	0,5434	0,5472(0,0248)	0,3083	0,3259(0,0345)
30	16,3595	16,4680(0,0506)	0,5706	0,5430(0,0229)	0,3283	0,3211(0,0333)
40	16,3344	16,4626(0,0631)	0,5183	0,5490(0,0210)	0,2938	0,3258(0,0276)
50	16,3358	16,4578(0,0621)	0,5611	0,5495(0,0197)	0,3428	0,3251(0,0304)
n_p	SS		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	0,2505	0,2200(0,0127)	0,5519	0,5542(0,0207)	0,3256	0,3341(0,0317)
20	0,2572	0,2265(0,0101)	0,5489	0,5484(0,0179)	0,3221	0,3246(0,0248)
30	0,2510	0,2292(0,0099)	0,5457	0,5507(0,0156)	0,3163	0,3305(0,0252)
40	0,2577	0,2321(0,0120)	0,5533	0,5472(0,0158)	0,3264	0,3222(0,0258)
50	0,2629	0,2343(0,0096)	0,5479	0,5469(0,0162)	0,3272	0,3201(0,0278)

Tabela 57 – Resultados para a base Water

n_p	CS		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	0,0770	0,0825(0,0025)	0,2728	0,2570(0,0154)	0,0236	0,0241(0,0044)
20	0,0771	0,0818(0,0022)	0,2298	0,2537(0,0121)	0,0213	0,0245(0,0044)
30	0,0732	0,0809(0,0025)	0,2651	0,2561(0,0125)	0,0235	0,0232(0,0036)
40	0,0765	0,0807(0,0019)	0,2859	0,2557(0,0132)	0,0241	0,0227(0,0039)
50	0,0769	0,0803(0,0015)	0,2843	0,2548(0,0149)	0,0258	0,0233(0,0048)
n_p	DB		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	36,5592	39,9010(1,2724)	0,3603	0,2995(0,0311)	0,0434	0,0253(0,0094)
20	36,6490	38,9585(1,0030)	0,3231	0,3047(0,0359)	0,0303	0,0289(0,0113)
30	36,5098	38,9693(0,8981)	0,2664	0,2998(0,0311)	0,0113	0,0264(0,0110)
40	36,3040	38,7089(0,9107)	0,3029	0,3052(0,0333)	0,0214	0,0279(0,0114)
50	36,6941	38,5098(0,7903)	0,2666	0,3036(0,0365)	0,0231	0,0255(0,0105)
n_p	HM		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	11,5755	11,8183(0,1105)	0,2536	0,2551(0,0137)	0,0204	0,0258(0,0033)
20	11,6572	11,7934(0,0519)	0,2653	0,2553(0,0119)	0,0265	0,0265(0,0036)
30	11,5134	11,7502(0,0694)	0,2593	0,2531(0,0126)	0,0257	0,0254(0,0035)
40	11,5570	11,7568(0,0481)	0,2499	0,2542(0,0105)	0,0197	0,0251(0,0031)
50	11,4372	11,7532(0,0578)	0,2841	0,2540(0,0114)	0,0222	0,0263(0,0030)
n_p	SS		$F-measure$		ARI	
	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$	Melhor	$\mu(\sigma)$
10	0,3165	0,3032(0,0073)	0,2559	0,2536(0,0115)	0,0249	0,0246(0,0036)
20	0,3184	0,3084(0,0051)	0,2562	0,2580(0,0108)	0,0215	0,0258(0,0035)
30	0,3188	0,3097(0,0047)	0,2774	0,2596(0,0125)	0,0316	0,0260(0,0032)
40	0,3195	0,3108(0,0047)	0,2551	0,2560(0,0113)	0,0248	0,0250(0,0033)
50	0,3196	0,3116(0,0044)	0,2501	0,2546(0,0120)	0,0265	0,0246(0,0031)