# UM ESTUDO DA APLICAÇÃO DE TÉCNICAS DE COMBINAÇÃO DE AGRUPAMENTOS

**Augusto César Fadel**
Escola Nacional de Ciências Estatísticas – ENCE
augustofadel@gmail.com

**Gustavo da Silva Semaan**
Instituto Noroeste Fluminense de Educação Superior da Universidade Federal Fluminense – INFES/UFF
gustavosemaan@id.uff.br

**José André de Moura Brito**
Escola Nacional de Ciências Estatísticas – ENCE/IBGE
jose.m.brito@ibge.gov.br

**Resumo**

A combinação de agrupamentos é um método altamente promissor para melhorar a qualidade de soluções de agrupamento, combinando diferentes partições (soluções) em uma única partição consenso. Nesse sentido, o presente estudo considera a aplicação desse método para dados do Censo Demográfico 2010, Produto Interno Bruto dos municípios brasileiros em 2010 e Índice de Desenvolvimento Humano dos municípios brasileiros em 2010. Além desses, foram consideradas algumas instâncias tradicionais da literatura de agrupamentos, cujos grupos foram construídos a partir da aplicação dos algoritmos k-means, PAM e DBSCAN. As soluções obtidas foram validadas mediante a aplicação do índice silhueta e submetidas a um comitê de agrupamento, com função consenso baseada em coassociação. A partir da aplicação dessa função, foi possível produzir uma solução consenso, cuja similaridade com as partições base foi quantificada através da informação mútua normalizada, o índice mais utilizado na literatura para esse fim.

**Palavras-Chaves:** Agrupamento; Combinação de agrupamento; Algoritmo; Particional; Densidade; K-means; PAM; DBSCAN;

**Abstract**

The clustering ensembles is a very promising method to improve the quality of clustering solutions, combining different partitions in a single consensus partition. In this sense, this study consider the appliance of this method to data from the 2010 Brazilian census, 2010 Brazilian cities Gross Product and 2013 Brazilian cities Human Development Index. Besides these, some instances from the clustering literature were considered, whose clusters were constructed by the appliance of the algorithms k-means, PAM and DBSCAN. The obtained solutions were validated by a silhouette index criteria and were submitted to a cluster ensemble method with consensus function based on coassociation. Applying this function, it was possible to produce a consensus solution, whose similarity with the partitions was measured by the Normalized Mutual Information, the most commonly used index for this purpose in the literature.

**Keywords:** Cluster; Cluster Ensemble; Algorithm; Partitioning; Density-based; K-means; PAM; DBSCAN;

# 1. INTRODUÇÃO

A evolução tecnológica observada nas últimas décadas tem permitido a captação e o acúmulo de uma vasta quantidade de informações, produzindo enormes conjuntos de dados. Nesse sentido, a aplicação de abordagens tradicionais no tratamento e na interpretação dessas bases tem se mostrado ineficiente ou até mesmo inviável, no que diz respeito à compreensão dos fenômenos a elas associados. Esse cenário motivou o desenvolvimento de técnicas computacionais que combinam os métodos tradicionais de análise de dados com algoritmos sofisticados, capazes de descobrir padrões úteis, outrora ignorados, em grandes bases de dados. Essas técnicas estão associadas à mineração de dados.

A análise de agrupamentos é uma técnica que emprega algoritmos computacionais para agrupar objetos semelhantes (considerando os seus atributos) contidos em uma base de dados, ainda que não haja um conhecimento prévio sobre classes ou valores associados a esses dados. Cada algoritmo de agrupamento baseia-se em critérios específicos para obter as soluções de agrupamento (Naldi, 2011). Dessa forma, alguns algoritmos são mais aptos para identificar certas estruturas do que outros. Em geral, não é possível obter um único agrupamento ideal, mas sim um conjunto de soluções traduzem diferentes características do conjunto de dados. A combinação de agrupamentos é uma técnica capaz de, a partir desse conjunto de soluções, produzir uma nova solução, que pode ser inclusive de qualidade superior no que diz respeito à adequada alocação dos objetos.

O presente trabalho tem o objetivo de apresentar um estudo da aplicação de técnicas de combinação de agrupamentos e foi organizado em cinco seções. A seção dois descreve o problema de agrupamento e apresenta a metodologia proposta. Mais especificamente, são descritos, sucintamente, os algoritmos de agrupamento utilizados e as técnicas empregadas para definir seus parâmetros de entrada. Nessa seção encontra-se, também, a descrição dos critérios baseados no índice silhueta, sendo tais critérios utilizados para validar as soluções produzidas pelos algoritmos supracitados. Por fim, é apresentada a técnica de combinação de agrupamentos. A seção quatro relaciona os 24 conjuntos de dados (instâncias) utilizados no trabalho. Na seção cinco são apresentados os resultados dos experimentos computacionais realizados. A seção seis traz a conclusão do estudo e possíveis temas para trabalhos futuros.

# 2. METODOLOGIA

## 2.1. O PROBLEMA DE AGRUPAMENTO

O problema clássico de agrupamento pode ser definido como: dado um conjunto formado por $n$ objetos $X=\{ x_1,x_2,...,x_i,...,x_n \}$ em que com cada objeto $x_i \in X$ possui $p$ atributos, deve-se construir $k$ grupos $C_j$ ( $j=1,...,k$ ) a partir de $X$, de forma a garantir que os objetos de cada grupo sejam homogêneos segundo alguma medida de similaridade. Uma solução (ou partição) pode ser representada como $\pi = \{ C_1,C_2,...,C_k \}$ (Han et al., 2012).

Uma ampla variedade de algoritmos para o Problema de Agrupamento é encontrada na literatura. Alguns, devido à possibilidade de variação de seus parâmetros, produzem soluções diversificadas. Neste sentido, surge então a combinação de agrupamentos (ou comitê de agrupamento, do inglês "*cluster ensemble*") que pode ser definida como: dado um conjunto de soluções do problema de agrupamento de tamanho $q$, $\prod = \{\pi_1,\pi_2,...,\pi_q\}$, também definido como conjunto de partições base, deve-se encontrar uma única solução consenso. Esse conjunto é constituído por soluções resultantes da aplicação de um algoritmo de agrupamento várias vezes (considerando variação de seus parâmetros) ou da aplicação de alguns algoritmos de agrupamento em determinado conjunto de dados $X$ (Naldi, 2011).

## 2.2. PROCEDIMENTO PROPOSTO

A metodologia adotada no presente trabalho foi desenvolvida em duas etapas, quais

sejam: (i) obtenção do conjunto de partições base; (ii) obtenção da partição consenso.

Na primeira etapa foram construídas as soluções de agrupamento, chamadas de partições base, através da aplicação dos conhecidos algoritmos da literatura: *k*-means, PAM e DBSCAN (Han et al., 2012). Cada algoritmo de agrupamento produziu diversas partições para cada instância, considerando diferentes configurações dos parâmetros de entrada ou da variação do número de grupos desejado, submetido como parâmetro de entrada quando possível. As partições base foram então avaliadas mediante a aplicação de um critério de validação baseado no índice silhueta e, aquelas validadas, formaram os conjuntos de partições base.

Na segunda etapa, um algoritmo de combinação de agrupamentos, orientado por uma função consenso baseada em coassociação (Naldi, 2011), foi aplicado aos conjuntos de partições base, gerando uma nova solução de agrupamento para cada conjunto, denominada partição consenso. A qualidade das partições consenso foi avaliada através da informação mútua normalizada, critério capaz de mensurar a similaridade entre uma partição consenso e o conjunto de partições base que lhe deu origem. Quanto maior essa similaridade, maior o sucesso da combinação. Os algoritmos e métodos mencionados acima são descritos nas seções subsequentes.

### 2.3. ALGORITMOS DE AGRUPAMENTO

Foram adotados três algoritmos particionais e exclusivos (*hard clustering*). Tais algoritmos atribuem cada objeto do conjunto de dados a um e somente um grupo da partição gerada. O algoritmo *k*-means básico é completo (todos os objetos devem pertencer a um grupo) e sua complexidade computacional é $O(nk)$. É baseado em protótipos definidos em termos de centróides, sendo o centróide de um grupo correspondente à média de cada um dos atributos dos objetos nele contidos. A sua eficácia pode ser prejudicada na presença de *outliers*. O algoritmo PAM (*Partitioning Around Medoids*) é uma realização particular do método de agrupamento *k*-medoids (Han et al., 2012) e possui complexidade computacional $O(k(n-k)^2)$. Ele também é baseado em protótipos, definidos em termos de medóides, sendo o medóide de um grupo correspondente ao objeto mais representativo nele contido. Por ser um algoritmo determinístico, múltiplas execuções desse algoritmo em uma mesma instância resultam sempre na mesma solução. A fim de ampliar a variedade de soluções, o método CLARA (*Clustering LARge Applications*) também foi aplicado. Tal método, proposto por Kaufman e Rousseeuw (1986), é uma variação do método PAM para aplicação em grandes conjuntos de dados, visando reduzir os requisitos de memória e de capacidade de processamento. Sua complexidade computacional é dada por $O\left(k^2 + k(n-k)\right)$.

O algoritmo DBSCAN (*Density-Based Spatial Clustering of Application with Noise*) é parcial, ou seja, objetos podem ser rotulados como ruído e não serem incluídos em nenhum grupo. Ele é baseado no conceito de densidade e tem complexidade computacional $O(n^2)$.

### 2.4. DEFINIÇÃO DOS PARÂMETROS DE CONFIGURAÇÃO DOS ALGORITMOS

Os algoritmos *k*-means e PAM necessitam que o usuário especifique o número de grupos, $k$, a serem formados. Este parâmetro foi estimado individualmente para cada instância a partir da execução destes dois algoritmos com $k \in [2, \sqrt{n}]$, intervalo considerado na maioria dos trabalhos publicados. As partições associadas aos maiores índice silhueta, em cada um dos algoritmos, foram armazenadas. As execuções finais adotaram os números de grupos estimados em ambos os algoritmos e também um grupo a menos e um grupo a mais, ou seja: $v = \{k_{k-means} - 1, k_{k-means}, k_{k-means} + 1\} \cup \{k_{PAM} - 1, k_{PAM}, k_{PAM} + 1\}$. Dessa forma, o vetor $v$ terá no mínimo três diferentes valores, caso $k_{k-means} = k_{PAM}$, e no máximo seis diferentes valores. No entanto, caso ocorra $k_{k-means} = 2$ ou $k_{PAM} = 2$, o primeiro elemento do vetor $v$ é desconsiderado. O algoritmo DBSCAN, entretanto, não permite estabelecer a priori o número de grupos desejado. Nesse caso, os parâmetros necessários são o raio da circunferência de abrangência e o número de objetos a partir do qual uma região deve ser

considerada densa. A metodologia utilizada para definí-los foi proposta por Semaan et al. (2012) em que 28 pares de parâmetros para configuração são considerados.

### 2.5. VALIDAÇÃO DAS PARTIÇÕES BASE

O índice de validação adotado foi o índice silhueta, proposto por Rousseeuw (1987). Esse índice define a qualidade dos agrupamentos com base na proximidade entre os objetos de determinado grupo e na distância desses objetos ao grupo mais próximo. O índice silhueta original é calculado para cada objeto de um grupo, indicando quais objetos estão bem situados no mesmo e quais seriam situados melhor em outro grupo. Pode ser calculado com qualquer medida de dissimilaridade ou similaridade e retorna valores no intervalo [-1,1], onde valores positivos próximos de 1 indicam que o objeto está bem situado em seu grupo e valores negativos próximos de -1 indicam que o objeto está mais próximo de outro grupo.

Como a silhueta depende apenas do agrupamento resultante, e não do algoritmo de agrupamento empregado, a média das silhuetas dos objetos (ASWC, do inglês *Average Silhouette Width Criterion*), $\bar{s}$, pode ser usada para executar uma análise dos grupos obtidos, para comparar as partições geradas por diferentes algoritmos, ou diferentes execuções do mesmo algoritmo, em uma mesma instância. Esse índice é obtido a partir da Equação 1 em que $a_i$ é a dissimilaridade média do objeto $x_i$ para os demais objetos contidos no grupo $C_l$, do qual $x_i$ faz parte; Além disso, $b_i$ é a dissimilaridade média do objeto $x_i$ para os objetos do grupo vizinho mais próximo $C_s$, do qual $x_j$ faz parte.

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max(a_i, b_i)} \qquad (1)$$

$$a_i = \frac{1}{(n_{C_1} - 1)}\sum_{j \in C_l, j \neq i} d(x_i \qquad b_i = \min_{C_s \neq C_l} d(i \qquad d(i, C_s) = \frac{1}{n_{C_s}}\sum_{j \in C_s} d(x_i, x$$

As partições base podem ser classificadas, segundo sua silhueta média, conforme apresentado na Tabela 1 (Retzer e Shan, 2007).

**Tabela 1 - Critério de classificação segundo silhueta média**

| Silhueta média | Classificação |
|---|---|
| 0,71 a 1,00 | Uma estrutura consistente foi encontrada |
| 0,51 a 0,70 | Uma razoável estrutura foi encontrada |
| 0,26 a 0,50 | Uma estrutura fraca e possivelmente artificial foi encontrada |
| ≤ 0,25 | Nenhuma estrutura substancial foi encontrada |

Com base nesse critério, as instâncias que obtiveram ao menos uma solução com silhueta média superior a 0,70, deram origem a três conjuntos de partições base, são eles: $\Pi_{70}$ (soluções com $\bar{s}$ > 0,70); $\Pi_{50}$ (soluções com $\bar{s}$ > 0,50); $\Pi_{25}$ (soluções com $\bar{s}$ > 0,25).

Nas instâncias em que nenhuma solução apresentou silhueta média superior a 0,70, foram formados três conjuntos de partições base, conforme as respectivas distribuições do índice de validação, são eles: $\Pi_{q_3}$ (soluções com $\bar{s}$ acima do terceiro quartil); $\Pi_{q_2}$(soluções com $\bar{s}$ acima do segundo quartil); $\Pi_{25}$ (soluções com $\bar{s}$ > 0,25).

A silhueta média foi o único critério adotado na construção dos conjuntos de partições base, não sendo aplicado nenhum filtro no que diz respeito ao número de grupos. Assim, todas as soluções validadas foram contempladas em ao menos um conjunto, independente do algoritmo utilizado e do número de grupos resultante.

### 2.6. COMBINAÇÃO DE AGRUPAMENTOS

Não é trivial determinar a melhor solução de agrupamento ou o algoritmo mais apropriado para um determinado problema. Em geral, é obtido um conjunto de soluções no qual muitas se equivalem em qualidade, em termos do critério estabelecido. A técnica de combinação de agrupamentos permite, a partir de um conjunto suficientemente diversificado no que concerne às soluções, obter uma solução consenso, que pode ser inclusive de qualidade superior no que diz respeito à adequada alocação dos objetos.

Muitos algoritmos podem gerar diferentes resultados para um mesmo conjunto de dados. Dentre os motivos para tal, podem ser considerados o uso de funções objetivo distintas, inicializações distintas, como, por exemplo, diferentes conjuntos de centróides iniciais no algoritmo *k*-means, ou ainda diferentes valores para os parâmetros livres de configuração, como o raio de abrangência e o critério de densidade do algoritmo DBSCAN (Hartigan *apud* Naldi, 2011). É necessário considerar que, apesar de um pequeno número de objetos agrupados em poucos grupos pode resultar em muitas diferentes soluções, a maioria delas não apresenta grande diversidade entre si, no que se refere à composição dos grupos, uma vez que muitos objetos acabam sendo alocados sempre no mesmo grupo. Combinações com maior diversidade tendem a produzir partições de melhor qualidade (Kuncheva e Hadjitodorov *apud* Naldi, 2011). Além disso, partições semelhantes produzidas por metodologias distintas indicam a presença de uma forte estrutura de agrupamento (Naldi, 2011).

### 2.7. FUNÇÃO CONSENSO

No presente trabalho, foi adotada função consenso baseada em coassociação. Em tais funções, a similaridade entre dois objetos é determinada pelo número de grupos compartilhados entre eles em todas as partições base. Essa similaridade representa a força de coassociação entre os objetos e é organizada em uma matriz de coassociação.

Para construção da partição consenso, foi aplicado modelo proposto por Gordon e Vichi (2001), que adota uma a abordagem de otimização, utilizando um algoritmo SUMT (*Sequencial Unconstrained Minimization Technique*) para minimizar a função objetivo apresentada na Equação 2, em que $M$ representa o número de partições base; $\pi_m$ é a *m*-ésima partição do conjunto de partições base; $d$ é a dissimilaridade dada pela matriz de coassociação; $w_m$ é o peso atribuído à *m*-ésima partição.

$$L(\pi) = \sum_{m=1}^{M} w_m \, d(\pi, \pi_m)^2 \tag{2}$$

### 2.8. VALIDAÇÃO DAS PARTIÇÕES CONSENSO

Como foi apresentado anteriormente, a partição consenso é construída a partir de um conjunto de partições base que, por sua vez, foram previamente submetidas a um critério de validação (vide seção 2.5). Em seguida, é necessário avaliar a qualidade da partição consenso.

> O sucesso do processo de combinação pode ser medido por meio da similaridade entre a partição consenso, $\pi_{\mathbb{C}}$, e o conjunto de partições base, $\Pi$. Quanto maior a similaridade entre a partição consenso e o conjunto de partições base, melhor o resultado da combinação representa as partições que lhe deram origem. (Naldi, 2011)

A maioria dos índices de validação externos, como os índices de Jaccard e de Rand, pode ser utilizada para medir a similaridade entre a partição consenso e cada uma das partições do conjunto de partições base. Contudo, o método mais utilizado na literatura é a informação mútua normalizada (NMI, do inglês *Normalized Mutual Information*), que consiste em uma medida estatística capaz de qualificar a informação comum entre duas distribuições (Naldi, 2011). Esta medida é dada pela Equação 3, em que $k^a$ e $k^b$ são os números de grupos das partições $\pi^a$ e $\pi^b$, respectivamente; $C_h^a$ é o h-ésimo grupo da a-ésima partição ($\pi^a$); $n_o$ é o número de objetos do conjunto de dados.

$$\phi^{(NMI)}(\pi_a, \pi_b) = \frac{\sum_{h=1}^{k^a}\sum_{l=1}^{k^b} |C_h^a \cap C_l^b| \log\left(\frac{n_o |C_h^a \cap C_l^b|}{|C_h^a||C_l^b|}\right)}{\sqrt{\left[\sum_{h=1}^{k^a} |C_h^a| \log\left(\frac{|C_h^a|}{n_o}\right)\right]\left[\sum_{l=1}^{k^b} |C_l^b| \log\left(\frac{|C_l^b|}{n_o}\right)\right]}} \tag{3}$$

A NMI é invariante no que diz respeito ao número de grupos das partições avaliadas e atinge o valor máximo de 1 se, e somente se, houver uma correspondência direta, objeto a

objeto, entre as duas partições $\pi_a$ e $\pi_b$. A similaridade entre a partição consenso, $\pi_C$, e todo o conjunto de partições base, $\Pi$, foi medida através da informação mútua normalizada média (*ANMI*, do inglês *Average Normalized Mutual Information*), dada pela Equação 4 em que $n_\pi$ é correspondente ao número de partições contidas no conjunto de partições base.

$$\phi^{(ANMI)}(\pi_C, \Pi) = \frac{1}{n_\pi} \sum_{j=1}^{n_\pi} \phi^{(NMI)}(\pi_C, \pi_j) \quad (4)$$

## 3. BASES DE DADOS

Os algoritmos de agrupamento foram aplicados em quatro conjuntos de instâncias, três deles compostos por dados reais e um composto por dados amplamente utilizados na literatura de análise de agrupamentos. Cada conjunto foi objeto de uma análise distinta.

Uma vez que todas as instâncias têm variáveis (atributos) apenas quantitativos, foi efetuada uma padronização, a fim de que nenhuma sobressaísse em importância às demais variáveis de sua instância, na execução dos algoritmos. Nesse sentido, cada observação $x_{ij}$ foi transformada no valor padronizado $z_{ij}$ correspondente como apresenta a Equação 5 em que $x_{ij}$ é a $i$-ésima observação do $j$-ésimo atributo; $\mu_j$ é a média do $j$-ésimo atributo; $\sigma_j$ é o desvio padrão do $j$-ésimo atributo; $z_{ij}$ é o valor padronizado da observação $x_{ij}$.

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \begin{cases} i = 1, \ldots, n \\ j = 1, \ldots, p \end{cases} \quad (5)$$

No que diz respeito a valores atípicos, nenhum tratamento foi aplicado nas instâncias da literatura, pois a utilização de tais conjuntos de dados é uma prática recomendada para possibilitar a correspondência com outros estudos e análises, o que ficaria seriamente comprometido com a remoção de qualquer observação. Nas instâncias de dados reais, foram excluídas todas as observações posicionadas três vezes a distância interquartílica além do quartil superior ou três vezes a distância interquartílica aquém do quartil inferior.

### 3.1. DADOS REAIS

Os conjuntos de instâncias reais foram compostos por três tipos de instâncias consolidadas a partir de dados disponibilizados no site do Instituto Brasileiro de Geografia e Estatística (IBGE). Mais especificamente, dados do censo demográfico, do PIB (Produto Interno Bruto) e do IDH (Índice de Desenvolvimento Humano). Essas instâncias possuem número de objetos compreendidos entre 71 a 1780 e número de atributos entre 3 e 5, como relacionado na Tabela 2. Em todas as instâncias reais os objetos correspondem a municípios brasileiros e os atributos estão associados, de alguma forma, ao respectivo grau de desenvolvimento desses.

**Tabela 2 - Conjunto de instâncias de dados reais**

| Instância | | Número de objetos | Número de atributos |
|---|---|---|---|
| Censo Demográfico 2010 | Pará | 128 | 5 |
| | Maranhão | 190 | 5 |
| | Mato Grosso do Sul | 73 | 5 |
| | Rio de Janeiro | 88 | 5 |
| | Santa Catarina | 256 | 5 |
| PIB Municipal *per capita* 2010 | Tocantins | 117 | 4 |
| | Ceará | 164 | 4 |
| | Mato Grosso | 122 | 4 |
| | Espírito Santo | 71 | 4 |
| | Paraná | 369 | 4 |
| IDHM 2010 | Norte | 449 | 3 |
| | Nordeste | 1780 | 3 |
| | Centro-Oeste | 465 | 3 |
| | Sudeste | 1668 | 3 |
| | Sul | 1188 | 3 |

As instâncias do Censo Demográfico 2010 possuem 5 atributos, quais sejam: (1) Valor do rendimento nominal médio mensal das pessoas de 10 anos ou mais de idade (com rendimento); (2) Proporção de domicílios particulares permanentes com banheiro de uso exclusivo dos moradores ou sanitário e esgotamento sanitário via rede geral de esgoto ou pluvial; (3) Proporção de domicílios particulares permanentes sem banheiro de uso exclusivo dos moradores; (4) Proporção de domicílios particulares permanentes com lixo coletado; (5) Proporção de domicílios particulares permanentes com energia elétrica de companhia distribuidora.

Nas instâncias do PIB municipal 2010, os atributos são os valores adicionados brutos no ano de 2010, *per capita*, em quatro setores, quais sejam: agropecuária, indústria, serviços (excluídos administração, saúde, educação pública e seguridade social) e administração. Em relação às instâncias do IDH Municipal 2010, os atributos correspondem às três dimensões do índice, quais sejam: longevidade, educação e renda.

### 3.2. DADOS DA LITERATURA

O conjunto de dados da literatura foi composto por nove instâncias bem conhecidas (Semaan et al., 2012). Essas instâncias possuem número de objetos variando entre 75 e 1484 e número de dimensões entre 2 e 60, como relacionado na Tabela 3.

**Tabela 3 - Conjunto de instâncias da literatura**

| Instância | Número de objetos | Número de atributos |
|---|---|---|
| 200DATA | 200 | 2 |
| chart | 600 | 60 |
| gauss9 | 900 | 2 |
| iris | 150 | 4 |
| maronna | 200 | 2 |
| ruspini | 75 | 2 |
| spherical_4d3c | 400 | 3 |
| wine | 178 | 13 |
| yeast | 1484 | 7 |

## 4. EXPERIMENTOS COMPUTACIONAIS

Os resultados computacionais foram divididos em duas seções. A primeira seção traz as soluções geradas pelas execuções dos algoritmos de agrupamento e os respectivos valores do índice de validação, o que corresponde à primeira etapa da implementação da metodologia proposta. A segunda seção mostra os resultados da combinação dos agrupamentos validados na primeira seção, as partições consenso, segunda etapa da metodologia. Todos os experimentos computacionais foram desenvolvidos em linguagem R e realizados em um computador dotado de um processador i7 de 2.6 GHz e 4 *cores* (núcleos), 8GB de RAM e sistema operacional MAC OS X, versão 10.8.4.

### 4.1. PARTIÇÕES BASE

A Figura 1 apresenta o quantitativo de soluções válidas por algoritmo de agrupamento. O número muito superior de soluções do DBSCAN deve-se ao fato desse algoritmo permitir a configuração de dois parâmetros, contra apenas um nos demais algoritmos. Esses parâmetros, combinados, permitiram executar o algoritmo com 28 configurações diferentes para cada instância, como descrito na seção 2.4.

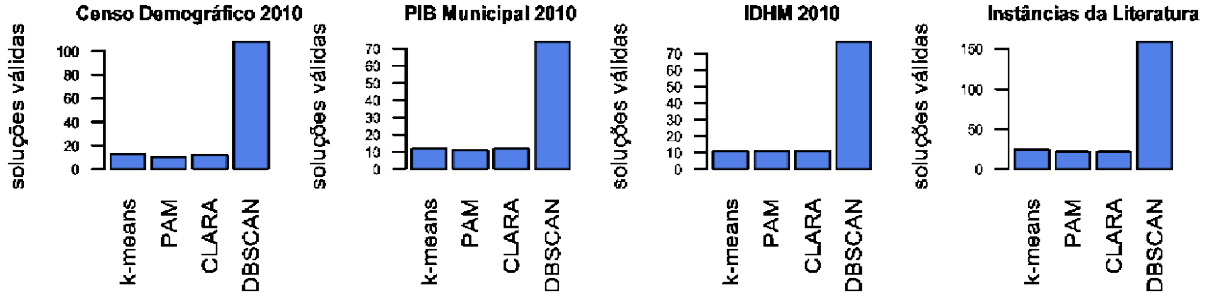

**Figura 1: Quantitativo de soluções válidas por algoritmo de agrupamento**

A Tabela 4 apresenta um resumo do número de partições válidas obtidas por instância. A segunda coluna apresenta o total de partições construídas pelos três algoritmos, a terceira coluna traz o número soluções que obtiveram mais de um grupo, pois o algoritmo DBSCAN, conforme os valores submetidos como parâmetros de entrada, pode produzir soluções com um único grupo, o que é indesejável. A última coluna apresenta o número de partições aptas a compor os conjuntos de partições base.

**Tabela 4 - Resumo do número de partições base obtidas nas instâncias do Censo Demográfico 2010**

| Instância | | Total de partições | Partições com $k > 1$ | Partições válidas $\bar{s} > 0{,}25$ |
|---|---|---|---|---|
| Censo Demográfico 2010 | Pará | 34 | 21 | 12 |
| | Maranhão | 34 | 27 | 23 |
| | Mato Grosso do Sul | 34 | 26 | 17 |
| | Rio de Janeiro | 43 | 36 | 25 |
| | Santa Catarina | 34 | 27 | 20 |
| PIB Municipal *per capita* 2010 | Tocantins | 34 | 23 | 22 |
| | Ceará | 37 | 30 | 27 |
| | Mato Grosso | 37 | 21 | 18 |
| | Espírito Santo | 34 | 21 | 17 |
| | Paraná | 34 | 24 | 20 |
| IDHM 2010 | Norte | 34 | 20 | 17 |
| | Nordeste | 34 | 19 | 12 |
| | Centro-Oeste | 34 | 16 | 12 |
| | Sudeste | 34 | 14 | 8 |
| | Sul | 34 | 22 | 16 |
| Instâncias da Literatura | 200DATA | 37 | 33 | 31 |
| | chart | 37 | 17 | 9 |
| | gauss9 | 37 | 30 | 16 |
| | íris | 34 | 26 | 25 |
| | maronna | 37 | 28 | 24 |
| | ruspini | 37 | 28 | 27 |
| | spherical 4d3c | 37 | 29 | 20 |
| | wine | 37 | 25 | 7 |
| | yeast | 40 | 37 | 20 |

### 4.2. CONJUNTOS DE PARTIÇÕES BASE

A fim de possibilitar a composição de conjuntos com partições de boa qualidade e preservar a diversidade, as soluções obtidas pelos algoritmos de agrupamento foram organizadas em três subconjuntos, conforme descrito na seção 2.5. Cada subconjunto respeita um critério de silhueta média, quais sejam:

- ☐ Critério 1: partições com silhueta média superior a 0,25 ($\Pi_{25}$);
- ☐ Critério 2: partições com silhueta média superior a 0,50 ou acima do segundo quartil, para aqueles conjuntos que obtiveram silhueta média máxima inferior a 0,70, ($\Pi_{50}$ ou $\Pi_{q_2}$);
- ☐ Critério 3: partições com silhueta média superior a 0,70 ou acima do terceiro quartil, para aqueles conjuntos que obtiveram silhueta média máxima inferior a 0,70, ($\Pi_{70}$ ou $\Pi_{q_3}$).

As distribuições das silhuetas médias das soluções base obtidas para cada instância, com exceção da instância Wine, apresentam mediana acima do critério mínimo de Retzer e Shan, 0,25 (ver Tabela 1), indicando que mais de 50% das partições base obtidas são válidas para aplicação da combinação de agrupamentos.

Os algoritmos obtiveram soluções de agrupamento de elevada qualidade. Segundo o critério de Retzer e Shan, em mais de 50% das soluções das instâncias Ruspini e spherical_4d3c foi encontrada uma estrutura razoável e em mais de 50% das soluções da instância 200DATA uma estrutura consistente foi encontrada. Era esperado bom desempenho no agrupamento dessas instâncias, uma vez que sua estrutura é conveniente à aplicação dos algoritmos utilizados nesse estudo. Para as instâncias de dados reais, em mais de 70% das

instâncias foram obtidas soluções onde uma estrutura razoável foi encontrada, ainda segundo o critério de Retzer e Shan.

### 4.3. PARTIÇÕES CONSENSO

Os resultados obtidos na aplicação da metodologia proposta foram avaliados com base na ANMI. Para tal, em cada uma das 24 instâncias, e em cada um dos critérios de composição dos conjuntos de partições base descritos na seção 2.5, foi calculada a ANMI de cada partição base em relação ao seu conjunto de partições base, bem como a ANMI da partição consenso em relação ao conjunto de partições base que lhe deu origem.

Os diagramas a seguir (Figuras 2 a 5) apresentam os resultados obtidos para algumas instâncias. Nesses diagramas, cada *boxplot* representa a distribuição das ANMIs das partições base de determinado conjunto, cujo limite inferior da silhueta média encontra-se indicado no eixo horizontal, e o ponto azul representa a ANMI da respectiva partição consenso obtida.

Nas instâncias do Censo Demográfico 2010, em 14 dos 15 conjuntos de partições base formados, o valor da ANMI da partição consenso foi igual ou superior à maior ANMI obtida pelas partições base. A única exceção foi observada no conjunto $\Pi_{q_2}$ da instância composta por dados do Estado do Mato Grosso do Sul, onde $\phi^{(ANMI)}(\pi_C, \Pi_{q_2}) = 0{,}34$ e $\max\left(\phi^{(ANMI)}(\pi_i, \Pi_{q_2})\right) = 0{,}39$.

Outro resultado notável foi a robustez da função consenso baseada em coassociação às partições base pouco representativas. O conjunto $\Pi_{q_2}$ da instância do Estado de Santa Catarina contém uma partição base com ANMI inferior a 0,01 e ANMI máximo de 0,48, ainda assim a partição consenso obteve ANMI de aproximadamente 0,57, conforme indicado na Figura 2. O valor da ANMI próximo de zero, indica que a partição é substancialmente diferente das demais partições do conjunto de partições base do qual ela faz parte. No entanto, ela faz parte do conjunto de partições com silhueta média mais elevada, logo é uma solução de boa qualidade e contribuiu para o aprimoramento da partição consenso.

Nas instâncias do PIB Municipal 2010, em 13 dos 15 conjuntos de partições base formados, o valor da ANMI da partição consenso foi igual ou superior à maior ANMI obtida pelas partições base. As duas exceções foram observadas no conjunto $\Pi_{q_2}$ da instância composta por dados do Estado do Espírito Santo, onde $\phi^{(ANMI)}(\pi_C, \Pi_{q_2}) = 0{,}48$ e $\max\left(\phi^{(ANMI)}(\pi_i, \Pi_{q_2})\right) = 0{,}49$, e no conjunto $\Pi_{25}$ da instância composta por dados do Estado do Mato Grosso, onde $\phi^{(ANMI)}(\pi_C, \Pi_{q_2}) = 0{,}31$ e $\max\left(\phi^{(ANMI)}(\pi_i, \Pi_{q_2})\right) = 0{,}39$.

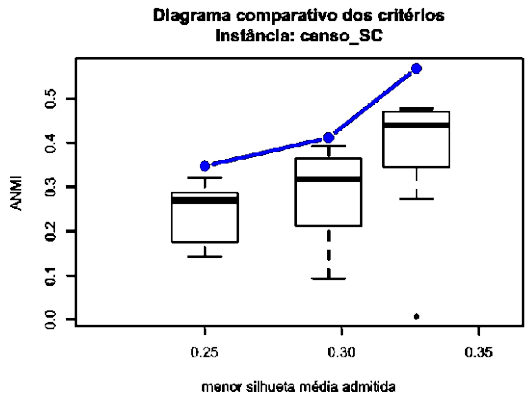


**Figura 2: Diagrama comparativo de instância do CD 2010 – Estado de Santa Catarina**

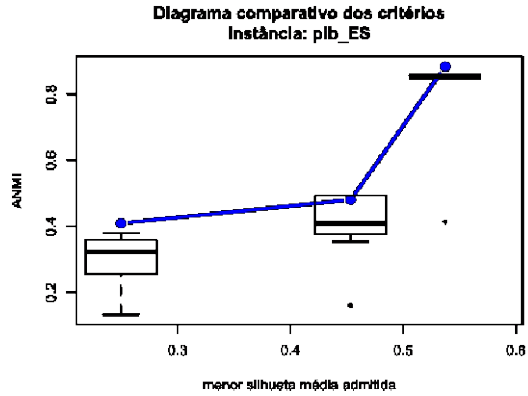


**Figura 3: Diagrama comparativo de instância do PIB municipal 2010 – Estado do Espírito Santo**

O conjunto $\Pi_{q_2}$ da instância do Estado do Espírito Santo contém uma partição base com ANMI de 0,41 e ANMI máxima de 0,85, conforme indicado na Figura 3. Ainda assim, a partição consenso obteve ANMI de aproximadamente 0,88, indicando que não houve degradação da qualidade da partição consenso mesmo na presença de solução base pouco representativa. Comportamento semelhante pôde ser visto na instância de dados do Paraná.

Nas instâncias do IDHM 2010, em 14 dos 15 conjuntos de partições base formados, o valor da ANMI da partição consenso foi igual ou superior à maior ANMI obtida pelas partições base. A única exceção foi observada no conjunto $\Pi_{25}$ da instância composta por dados da região Norte, onde $\phi^{(ANMI)}(\pi_C, \Pi_{q_{25}}) = 0{,}25$ e $\max\left(\phi^{(ANMI)}(\pi_i, \Pi_{q_{25}})\right) = 0{,}26$, conforme representado na Figura 4.

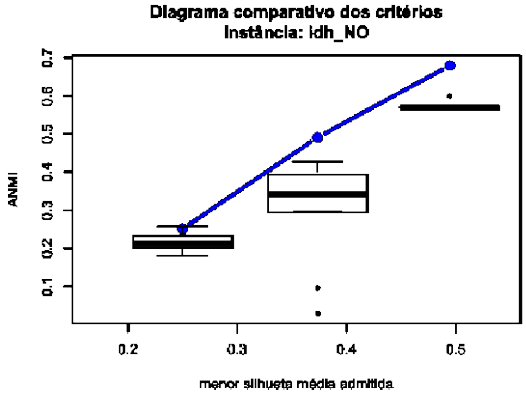


**Figura 4: Diagrama comparativo de instância do IDHM 2010 – Região Norte**

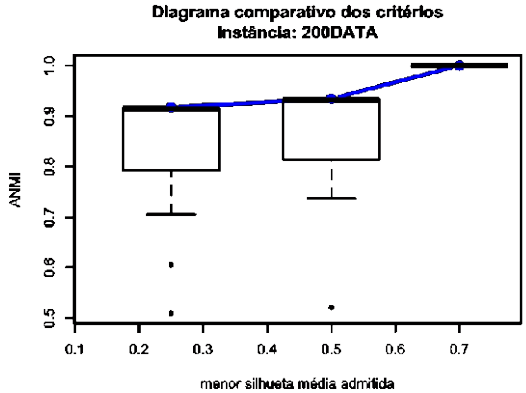


**Figura 5: Diagrama comparativo de instâncias da literatura – 200DATA**

O melhor resultado dentre as instâncias do IDHM 2010 foi obtido para os dados das regiões Centro-Oeste e Nordeste, onde a ANMI da partição consenso foi de aproximadamente 1,00. A partição consenso que apresentou a menor ANMI, de aproximadamente 0,25, foi aquela construída a partir das partições base do conjunto $\Pi_{25}$ da instância da região Norte.

As instâncias da literatura 200DATA, Maronna, Ruspini e spherical_4d3c, em função de suas estruturas naturais de agrupamento, são especialmente adequadas à aplicação da metodologia proposta, como já mencionado. Não por acaso, foram obtidos excelentes resultados a partir dessas instâncias.

Em 200DATA (Figura 5), Iris, Ruspini, spherical_4d3c e Yeast, as partições consenso obtiveram ANMI igual a ANMI máxima obtida pelas partições base. Esse resultado indica que a combinação de agrupamentos convergiu na direção das melhores partições base obtidas diretamente pelos algoritmos, sugerindo que essas são soluções de elevada consistência.

As instâncias Maronna e Gauss9 têm grupos de forma esférica, porém as fronteiras dos grupos não são muito bem definidas. Isso permitiu a construção de muitas soluções de boa qualidade, porém sutilmente diferentes, em função das características dos algoritmos e das configurações de parâmetros adotadas. Para soluções com um mesmo número de grupos, os objetos nos núcleos dos grupos são basicamente os mesmos, entretanto os objetos nas fronteiras não. Por essa razão, a combinação de agrupamentos construiu uma nova solução de qualidade ligeiramente superior àquelas obtidas diretamente pelos algoritmos. Em ambas as instâncias, Gauss9 e Maronna, a ANMI máxima obtida pelas soluções base foi de 0,94, enquanto que a ANMI das partições consenso foi de 0,95.

Dentre as instâncias da literatura, o resultado mais peculiar foi obtido na instância Wine, onde, ao contrário do observado em todas as demais instâncias analisadas, o conjunto de partições base formado pelas soluções de silhuetas médias mais elevadas não foi o que produziu partição consenso de maior ANMI.

Esse conjunto é de fato particular, pois foram obtidas apenas 7 soluções válidas, com silhuetas médias entre 0,26 e 0,32. Foi o pior resultado em termos do número de soluções válidas e o segundo pior no que diz respeito à silhueta média. O número restrito de partições válidas fez com que o conjunto $\Pi_{q_{25}}$ contasse com apenas duas partições base, ambas geradas pelo mesmo algoritmo, oferecendo pouca diversidade na construção da partição consenso.

Ainda assim, a combinação produziu partição com ANMI de 0,75, contra ANMI máxima de 0,50 das partições base, mostrando robustez da técnica de combinação de agrupamentos mesmo em condições adversas.

Comportamento similar foi observado nos resultados da instância Chart, com características semelhantes aos resultados obtidos para instância Wine. Tais dados obtiveram apenas 9 soluções de agrupamento válidas, segundo pior resultado, com silhuetas médias entre 0,26 e 0,31, pior resultado em termos de silhueta média.

Seu diagrama comparativo dos critérios mostra que a partição consenso de ANMI mais baixa foi obtida pelo conjunto $\Pi_{q_2}$ e não pelo $\Pi_{q_3}$, como na instância Wine, nem pelo $\Pi_{25}$, como nas demais instâncias.

Diferente do $\Pi_{q_3}$ da instância Wine, o $\Pi_{q_3}$ da instância Chart conta com quatro soluções, com três números de grupos ($k$) diferentes, produzidas por dois algoritmos distintos. Pode-se concluir então que, nesse caso, o critério mais restrito na seleção de soluções base de $\Pi_{q_2}$ para $\Pi_{q_3}$ elevou a qualidade do conjunto sem prejudicar a diversidade em demasia. Já o valor inferior da ANMI da partição consenso do conjunto $\Pi_{q_2}$ em relação à do conjunto $\Pi_{25}$ pode ser atribuído ao fato de que a transição de um conjunto para o outro reduziu a diversidade sem elevar a qualidade de forma substancial. De qualquer forma, as partições consenso produzidas a partir dos três conjuntos base apresentaram ANMI superior à ANMI máxima das partições base.

## 5. CONCLUSÕES E TRABALHOS FUTUROS

Os resultados apresentados na seção 4.1 mostram que os algoritmos k-means, PAM e DBSCAN formam um conjunto bastante versátil na obtenção de partições base, capaz de produzir soluções de razoável qualidade, medida pelo índice silhueta, e grande diversidade, medida pela informação mútua normalizada, nos mais diferentes conjuntos de dados.

No que diz respeito às instâncias de dados reais, o algoritmo DBSCAN praticamente dominou as melhores soluções. Um vez que foi aplicado tratamento de valores extremos, tal resultado sugere a provável presença de grupos de formas arbitrárias, inibindo o bom desempenho individual dos demais algoritmos, apropriados para grupos de formas esféricas.

O desempenho superior do algoritmo DBSCAN também pode ser creditado à metodologia abrangente e eficaz de configuração de parâmetros proposta por Semaan et. al. (2012).

A qualidade das partições base obtidas para as instâncias de dados reais poderia ainda ser elevada através da execução de um procedimento mais criterioso para seleção de atributos, o que não foi escopo desse trabalho. Todavia deve-se ressaltar que, em todas as instâncias de dados reais, mais de 50% do total de soluções obtidas pelos quatro algoritmos adotados foi considerada válida segundo o critério de Ratzer e Shan.

Os resultados apresentados na seção 4.3 mostram as excelentes soluções produzidas pela combinação de agrupamentos com função consenso baseada em coassociação. A qualidade da partição consenso foi avaliada através da sua similaridade com o conjunto e partições base que lhe deu origem e quantificada pela informação mútua normalizada média.

Para as 24 instâncias estudadas foi construído um total de 72 conjuntos de partições base e cada um deu origem a uma partição consenso. Desse universo, em menos de 6% dos casos a partição consenso obteve ANMI inferior à ANMI máxima obtida pelo conjunto de partições base e em quase 67% dos casos a ANMI da partição consenso superou a ANMI máxima do conjunto base. Nos cerca de 27% restantes, a partição consenso apresentou ANMI igual a ANMI máxima do conjunto de partições base.

O método de composição dos conjuntos de partições base, descrito na seção 4.2, mostrou-se efetivo, pois de maneira geral os conjuntos apresentaram diversidade de soluções satisfatória e em 23 das 24 instâncias analisadas a ANMI da partição consenso obtida a partir do conjunto mais restritivo ($\Pi_{70}$ ou $\Pi_{q_3}$) superou as ANMIs das partições consenso obtidas a partir dos demais conjuntos de partições base. A exceção ficou por conta da instância wine e tal resultado pode ser atribuído ao número reduzido de soluções válidas obtidas nesse caso, uma vez que o conjunto $\Pi_{q_3}$ contou com apenas duas partições base, o que é insuficiente para

executar a combinação de agrupamentos.

Em resumo, o conjunto de algoritmos adotado mostrou-se adequado a diferentes tipos de dados e a combinação de agrupamentos com função consenso baseada em coassociação se mostrou eficaz na obtenção de uma única solução final a partir de várias soluções de grande diversidade. Portanto, os resultados apresentados permitem concluir que a metodologia proposta no presente trabalho é eficaz na obtenção não supervisionada de soluções de qualidade para problemas de agrupamento.

A combinação de agrupamentos é um tema extremamente amplo. Uma grande variedade de algoritmos de agrupamento, índices de validação e funções consenso está disponível. Além disso, a definição, supervisionada ou não, dos parâmetros e critérios adotados em cada uma das etapas de uma metodologia de combinação de agrupamentos tem impacto decisivo no seu desempenho. Apesar de a metodologia proposta no presente trabalho ter se mostrado eficaz na obtenção não supervisionada de soluções de qualidade para problemas de agrupamento, não foi conduzida nenhuma avaliação no que diz respeito ao custo computacional. O número de soluções base construídas pelo algoritmo DBSCAN, mais custoso entre os algoritmos adotados, pode implicar em demanda elevada de recursos computacionais e tempo de processamento, especialmente para instâncias de grandes dimensões.

Embora limitar as soluções desse algoritmo não pareça uma boa estratégia, uma vez que ele foi responsável por muitas soluções de qualidade, um futuro trabalho poderia procurar identificar o impacto dessa medida no custo computacional e na qualidade da partição consenso. Tal análise produziria, para diferentes instâncias, diagramas (tempo para obtenção da partição consenso X qualidade da partição consenso) indicando até que momento (variedade/quantidade de parâmetros do DBSCAN) o custo reflete em efetiva melhora do resultado final.

Ainda no que diz respeito ao custo computacional, a linguagem R oferece suporte à programação *multithreaded*. Tal recurso permite usufruir dos benefícios dos processadores de múltiplos núcleos, presentes em computadores pessoais desde meados da década passada, e poderia reduzir drasticamente o tempo de execução da metodologia proposta. Uma breve análise realizada por Hee (2013) mostra a redução de cerca de 80% no tempo decorrido de execução de uma rotina na linguagem R utilizando seis núcleos em vez de apenas um.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS


[1] Gordon, A.; Vichi, M. Fuzzy Partition Models for Fittind a Set of Partitions. **Psychometrika**, vol. 66, 2001.

[2] Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques. 3. ed. Morgan Kaufmann Publishers, 2012. ISBN 978-0-12-381479-1.

[3] Hee, Y. E. Speeding up performance with multicore programming [Internet]. **National University of Singapore, Computer Centre**; [citado em 11 de abr. de 2013]. Disponível em: <http://comcen.nus.edu.sg/technus/hpc/speeding-up-performance-with-multicore-programming/>. Acesso em: 13 de nov. 2013.

[4] Kaufman, L.; Rousseeuw, P. J. Clustering Large Data Sets. E. S. Gelsema and L. N. Kanal (Eds.), **Pattern Recognition in Pratice 2**, 1986.

[5] Naldi, M. C. Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados. Tese (Doutorado). São Carlos: USP, 2011.

[6] Retzer, J.; Shan, M. Cluster Ensemble Analysis and Graphical Depiction of Cluster Partitions. **Proceedings of the 2007 Sawtooth Software Conference**. Sequim WA, 2007.

[7] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, vol. 20, 1987.

[8] Semaan, G. et al. Proposta de um método de classificação baseado em densidade para a determinação do número ideal de grupos em problemas de clusterização. **Learning &**

**Nonlinear Models (L&NLM)**, vol. 10, n. 4, 2012.