PROPOSTA DE MÉTODO PARA SOLUÇÃO DO PROBLEMA DE AGRUPAMENTO AUTOMÁTICO

Augusto Fadel[1]
Instituto Brasileiro de Geografia e Estatística – IBGE
augustofadel@gmail.com

Gustavo Silva Semaan
Instituto do Noroeste Fluminense de Educação Superior – INFES/UFF
gsemaan@ic.uff.br

José André de Moura Brito
Escola Nacional de Ciências Estatísticas – ENCE/IBGE
jambrito@gmail.com

RESUMO

O problema clássico de agrupamento consiste, basicamente, em identificar a estrutura de grupos que porventura existe em um conjunto de dados. Entretanto**,** dado que existe uma grande gama de algoritmos de agrupamentos que incorporam diferentes conceitos e técnicas**,** identificar qual algoritmo é mais adequado para determinada instância (base de dados) não é uma tarefa trivial. Neste sentido, o presente trabalho apresenta um estudo de caso do problema de agrupamento automático. Mais especificamente, partições base são obtidas através dos algoritmos K-Means, CLARA, BRKGA-Medoids, DBSCAN e BIRCH, cujos parâmetros são configurados automaticamente e validadas segundo os índices silhueta, S_Dbw e silhueta baseada em densidade. Em seguida é aplicada função consenso baseada em coassociação e as partições finais obtidas são avaliadas segundo a informação mútua normalizada média. Os resultados apresentados neste estudo indicam que a combinação de agrupamentos foi capaz de construir uma solução que representa melhor o conjunto de partições base do que qualquer solução nele contida.

PALAVRAS CHAVE: Análise de agrupamentos. Agrupamento automático. Combinação de agrupamento.

ABSTRACT

The clustering problem basically consists in identifying the structure of groups that may exist on a dataset. However there are several clustering algorithms that address different concepts and techniques and identify which algorithm is best suited for particular instance is not a trivial task. This paper present a case study of the automatic clustering problem. More specifically, clustering solutions were obtained using the K-Means, CLARA, BRKGA-Medoids, DBSCAN and BIRCH algorithms, whose input parameters were automatically setted, and validated according to the silhouette, S_Dbw and density-based silhouette indexes. Next was applied consensus function based on co-membership and obtained final partitions were evaluated according to average normalized mutual information. The results indicate that the ensemble was able to build a solution that best represents the set of partitions that any solution contained therein.

KEYWORDS: Cluster analysis. Automatic clustering. Cluster ensambles.

[1] O IBGE está isento de qualquer responsabilidade pelas opiniões, informações, dados e conceitos emitidos neste artigo, que são de exclusiva responsabilidade do autor.

1 Introdução

A técnica de análise de agrupamentos é uma das mais úteis para definir grupos e identificar distribuições e padrões em conjuntos de dados (Halkidi et al., 2001). Têm por objetivo particionar o conjunto de dados em um número finito de grupos tomando, como base, as similaridades entre seus objetos (Naldi et al., 2013).

São quatro os passos básicos para desenvolver um processo de agrupamento (Fayyad et al. *apud* Halkidi et al., 2001), quais sejam: (i) a seleção de atributos, ou seja, das dimensões disponíveis no conjunto de dados analisado, a escolha daquelas que são representativas no que diz respeito à definição de grupos e ou identificação de padrões; (ii) a escolha do algoritmo de agrupamento, o que compreende a escolha da medida que será utilizada para identificar a similaridade entre os objetos do conjunto e a função objetivo, dentre outros aspectos; (iii) o critério de validação dos resultados, ou seja, um critério quantitativo objetivo capaz de oferecer parâmetro de comparação entre diferentes soluções obtidas; (iv) a interpretação dos resultados obtidos através da avaliação de um especialista, validando que os grupos formados ou padrões identificados pelo algoritmo apresentam coerência e de fato emergiram do conjunto de dados estudado.

Todavia, tal processo, ainda que robusto, pode ser excessivamente custoso do ponto de vista computacional, tornando-se inviável para muitas aplicações reais. A seleção de atributos e, principalmente, a interpretação dos resultados são etapas que, de maneira geral, exigem intensa dedicação de um ou mais profissionais com grande expertise no que concerne ao tema estudado. A etapa de interpretação dos resultados nunca será omitida, entretanto, é possível minimizar seu custo através de processos mais sofisticados de validação dos resultados, permitindo que os esforços dos especialistas sejam dedicados mais à interpretação dos resultados e menos à validação das soluções obtidas.

Quanto às etapas de definição do algoritmo de agrupamento e do critério de validação, em função da variedade disponível na literatura e da velocidade com que surgem versões aprimoradas e novas propostas, uma análise exaustiva dos recursos disponíveis não é producente.

No que diz respeito aos algoritmos de agrupamento, é fato conhecido que alguns são mais aptos para identificar certas estruturas do que outros. Entretanto, caso exista uma estrutura de grupos no conjunto de objetos analisado, essa será desconhecida a priori, o que implica a impossibilidade de escolher, de maneira consistente, um único algoritmo de agrupamento, ou mesmo uma família de algoritmos de agrupamento. Além disso, ainda que tal escolha seja feita, em geral, não é possível obter um único agrupamento ideal, mas sim um conjunto de soluções, que revelam diferentes características do conjunto de dados.

Nesse contexto surge a aplicação da técnica de combinação de agrupamentos, capaz de combinar soluções produzidas por diferentes algoritmos de agrupamento, com diferentes configurações de parâmetros, em uma única solução consenso. Considerando que, como mencionado, diferentes algoritmos são capazes de captar diferentes estruturas, essa técnica tende a tornar os processos de agrupamento mais robustos e adequados às diversas aplicações, incluindo o tratamento de ruídos e *outliers*, reaproveitamento de conhecimento e computação distribuída (Naldi et al., 2013). Pode, portanto, produzir soluções de qualidade superior àquelas produzidas pelos algoritmos de agrupamento, no que diz respeito à alocação adequada dos objetos aos grupos.

O presente trabalho apresenta um estudo de caso da aplicação de um comitê de agrupamento na obtenção de uma solução final de boa qualidade para o problema de agrupamento e está organizado da seguinte forma: a seção dois descreve o problema de agrupamento; a seção três apresenta a metodologia proposta, incluindo os algoritmos de agrupamento e os índices de validação adotados; a seção 4 descreve os experimentos computacionais realizados e os resultados obtidos; a seção 5, por fim, traz as conclusões obtidas.

2 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_t ($t=1,...,k$) 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 no que diz respeito ao número de grupos e à alocação dos objetos, ou seja, a estrutura de agrupamento final. 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 q soluções produzidas para o problema de agrupamento, definidas por $\prod = \{\pi_1,\pi_2,...,\pi_q\}$, 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).

3 Metodologia Proposta

O processo proposto foi desenvolvido em duas etapas, a saber: (i) obtenção do conjunto de partições base; (ii) obtenção da partição consenso.

A primeira etapa consistiu na construção das partições base a partir de execuções de cinco algoritmos de agrupamento bem conhecidos da literatura, quais sejam, K-Means, CLARA, BRGKA-Medoids, MRDBSCAN e BIRCH, descritos resumidamente na seção 3.1, e na validação dessas partições de acordo com quatro estatísticas utilizadas como índices de validação, quais sejam, Silhueta, Silhueta Densidade e S_Dbw, descritos resumidamente na seção 3.3.

A segunda etapa abrangeu a formação dos conjuntos de partições base, conforme descrito na seção 3.3, a construção das partições consenso utilizando função consenso baseada em coassociação, descrita na seção 3.4, e, por fim sua validação, onde foi adotada a Informação Mútua Normalizada Média (ANMI, do inglês *Average Normalized Mutual Information*), também descrita na seção 3.4.

3.1 Algoritmos de Agrupamento

Os algoritmos de agrupamento escolhidos para compor o processo proposto apresentam, em sua concepção, características distintas no que diz respeito à estratégia de formação de grupos, possibilitando a obtenção de conjuntos de soluções de grande diversidade.

Os cinco algoritmos adotados são exclusivos, ou seja, atribuem cada objeto do conjunto de dados a somente um grupo da partição gerada. O algoritmo K-Means básico, além de exclusivo, é completo, ou seja, todos os objetos são alocados a um e somente 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. Muitas versões deste algoritmo estão disponíveis na literatura, sendo adotada nesse trabalho a versão proposta por Hartigan e Wong (1979).

O método CLARA (*Clustering LARge Applications*), proposto por Kaufman e Rousseeuw (1986), é uma variação do algoritmo PAM (*Partitioning Around Medoids*) para aplicação em grandes conjuntos de dados, visando reduzir os requisitos de memória e de capacidade de processamento. Também é completo e baseado em protótipos, porém estes são definidos em termos de medóides, onde o medóide de um grupo correspondente ao objeto mais representativo nele contido. Sua complexidade computacional é dada por $O\left(k^2 + k(n-k)\right)$.

O BRKGA-Medoids (ou BRKGA-RC) foi proposto por Brito et al. (2014) e consiste, basicamente, na aplicação do BRKGA (*Biased Random-Key Genetic Algorithm*) ao problema dos k-medoids. Algoritmos Genéticos são capazes de resolver problemas complexos de otimização combinatória, sendo especialmente eficazes para busca de soluções ótimas (Reeves, 2010), pois não impõem muitas das limitações encontradas nos métodos de busca tradicionais. O método adota ainda o procedimento de reconexão por caminhos (*Path-Relinking*) na intensificação das soluções obtidas na fase de busca local, a fim de produzir soluções de qualidade superior.

O algoritmo DBSCAN (*Density-Based Spatial Clustering of Application with Noise*), proposto por Ester et al. (1996), é parcial, ou seja, objetos podem ser rotulados como ruído e não serem incluídos em nenhum grupo, oferecendo recurso para tratamento de valores extremos. É baseado no conceito de densidade, o que permite a identificação de grupos de formas arbitrárias, e tem complexidade computacional $O(n^2)$. O método MRDBSCAN (*Multiple Runs of DBSCAN*), é uma variante proposta por Semaan et al. (2012), na qual, a partir de parâmetros iniciais, diferentes valores de entrada são determinados segundo a técnica de calibração de parâmetros *DistK* apresentada por Tan et al. (2009).

O algoritmo BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), proposto por Zhang (1997), é um método hierárquico baseado no conceito de CF-tree (*clustering feature tree*) e é especialmente adequado à aplicação em grandes conjuntos de dados.

3.2 Definição dos Parâmetros

Os algoritmos K-Means e CLARA não determinam o número de grupos, k, automaticamente. Destarte, o intervalo $[2; \sqrt{n}]$, sendo n o número de objetos da instância, foi estabelecido para esse parâmetro e todos os valores nele contidos foram considerados. Além disso, a fim de evitar que a ocorrência de um mínimo local na função objetivo resulte em soluções de qualidade inferior, cada algoritmo foi executado 10 vezes para cada valor de k. Para cada uma das $10 \times (\sqrt{n} - 1)$ execuções, o K-Means foi configurado ainda para realizar 100 inicializações aleatórias de centroides e o CLARA foi configurado para selecionar 10 amostras de tamanho $\frac{n}{5}$, como sugerido por Kaufman e Rousseeuw (1986).

O BRKGA-Medoids, por sua vez, foi configurado de acordo com resultados apresentados por Brito et al. (2014). Portanto, os parâmetros de entrada receberam os seguintes valores: $p = 100$ (tamanho da população inicial de vetores de chaves aleatórias), $p_e = 0{,}20p$ (tamanho do conjunto de soluções elite de cada geração), $p_m = 0{,}15p$ (número de soluções mutantes introduzidas a cada geração), $\rho_e = 0{,}7$ (probabilidade de um cromossomo filho herdar um gene de um cromossomo pai oriundo do conjunto elite) e $m = 500$ (total de gerações).

Os parâmetros de entrada do algoritmo DBSCAN 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 defini-los foi a proposta por Semaan et al. (2012), adotando o vetor de k-ésimo vizinho mais próximo $k^* = \{3, 4, 5, 10, 15, 20, 50\}$, o que deu origem a 28 pares de parâmetros para cada instância.

O algoritmo BIRCH gera soluções de agrupamento através da construção da estrutura denominada CF-tree. Para tal, considera os parâmetros fator de ramificação (*branching factor*), o qual define quantos subgrupos um nó (*node*) pode ter, e limite T (*threshold T*), distância a partir da qual dois pontos são alocados em diferentes nós. Dada a similaridade de conceitos entre esses parâmetros e os utilizados no algoritmo DBSCAN, aliada ao bom desempenho da metodologia proposta por Semaan et al. (2012) em calibra-los, um experimento preliminar foi conduzido e os resultados sugeriram a adequação dessa técnica também para o algoritmo BIRCH. Portanto, o mesmo recurso foi empregado para definir seus parâmetros de entrada.

3.3 Partições Base

Os algoritmos de agrupamento produzem uma solução mesmo que não exista uma estrutura de grupos natural na instância analisada. Portanto, é imprescindível validar cada partição obtida. Os índices de validação relativos oferecem critérios para comparação de soluções obtidas através de diferentes métodos, permitindo não apenas validá-las, mas ordená-las. Uma grande variedade de índices de validação relativos está disponível na literatura. A fim de conferir diversidade ao conjunto de partições base, foram escolhidos índices capazes de destacar diferentes características, são eles o índice silhueta, proposto por Rousseeuw (1987), o índice silhueta baseado em densidade, proposto por Menardi (2010) e o S_Dbw, proposto por Halkidi e Vazirgiannis (2001).

O índice silhueta, proposto por Rousseeuw (1987), 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. A qualidade de uma solução de agrupamento pode então ser medida pela média das silhuetas dos objetos (ASWC, do inglês *Average Silhouette Width Criterion*), $\bar{s}$, sendo obtida 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.

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

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

Segundo Naldi (2011), esse índice é mais apropriado para avaliar estruturas de agrupamento com grupos de formas aproximadamente esféricas (ou hiperesféricas) e, nessa condição, apresenta desempenho superior aos índices VCR, DB (Davies-Bouldin) e Dunn.

O índice silhueta baseado em densidade (dbs, do inglês *density-based silhouette*), segundo índice de validação considerado, foi proposto por Menardi (2010). Corresponde a uma variante do índice silhueta original adaptada para avaliar especialmente partições obtidas a partir de algoritmos baseados em densidade. Cada objeto é avaliado individualmente, recebendo valores iguais ou inferiores a 1, porém, de maneira geral, não muito menores do que 0. Quanto maior é o valor atribuído a um objeto, mais forte é a evidência de que sua alocação foi correta, ao passo que valores pequenos sugerem baixa confiança na alocação realizada. Valores negativos indicam que o objeto tem elevada probabilidade a posteriori de pertencer a outro grupo, representando, portanto, uma forte evidência de alocação incorreta.

Uma vez que a distribuição real dos grupos não é conhecida, a dbs do i-ésimo objeto de X é estimada através da Equação 2, onde, dado que $x_i \in X$ é proveniente de uma função densidade de probabilidade f, $\hat{\tau}_{C_t}(x_i)$ é o estimador da probabilidade a posteriori de que x_i pertença ao grupo C_t, considerando que π_{C_t} é a probabilidade a priori de C_t e $\hat{f}_{C_t}(x_i)$ é a densidade estimada em x_i obtida a partir dos objetos alocados no grupo C_t. Além disso, C_l é o grupo no qual x_i foi alocado e C_s é o grupo vizinho mais próximo.

A qualidade de uma partição pode então ser avaliada a partir de medidas resumo dos valores de dbs dos objetos nela contidos. No presente trabalho foram utilizadas a dbs média, $\overline{dbs}$, obtida através da Equação 3, e a dbs mediana, dbs_{med}, obtida através da Equação 4.

$$\widehat{dbs}_i = \frac{\log\left(\frac{\hat{\tau}_{C_l}(x_i)}{\hat{\tau}_{C_s}(x_i)}\right)}{\max_{j=1,\ldots,n}\left|\log\left(\frac{\hat{\tau}_{C_l}(x_j)}{\hat{\tau}_{C_s}(x_j)}\right)\right|} \quad (2) \qquad \hat{\tau}_{C_t}(x_i) = \frac{\pi_{C_t}\hat{f}_{C_t}(x_i)}{\sum_{t=1}^{k}\pi_{C_t}\hat{f}_{C_t}(x_i)}$$

$$\overline{\widehat{dbs}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{dbs}_i \quad (3) \qquad dbs_{med} = \widehat{dbs}_{\left(\left\lfloor\frac{n+1}{2}\right\rfloor\right)} \quad (4)$$

O terceiro índice de validação adotado, S_Dbw, considera duas grandezas para avaliar a qualidade de uma partição, quais sejam: (i) a compacidade do grupo, em termos de variância interna (*intra-cluster variance*); (ii) a densidade entre grupos (*inter-cluster density*).

O índice S_Dbw de uma partição de k grupos, $S\_Dbw(k)$, pode ser obtido através da Equação 5. O espalhamento (do inglês, *scattering*) médio de uma partição de k grupos, $Scat(k)$, é obtida a partir da razão entre as normas dos vetores de variância interna de cada grupo, $\sigma^2(C_l)$, e a de variância de todos os objetos da instância, $\sigma^2(X)$, em seus p atributos (dimensões), como representado na Equação 6. A densidade entre grupos**,** média de uma partição com k grupos, $Dens\_bw(k)$, é obtida a partir da razão entre a densidade, medida em termos de número de objetos, na região entre grupos e a maior densidade na região em torno do centro dos grupos, como representado na Equação 7, onde v_l e v_s são os centros dos grupos C_l e C_s, respectivamente, e u_{ls} é o ponto médio do segmento de reta traçado entre os dois centros. A região (ou vizinhança) de um ponto é demarcada por uma hiperesfera de raio σ descrita ao redor deste ponto, em seus p atributos, onde σ representa o desvio padrão médio dos k grupos. A Figura 1 traz uma representação esquemática do cálculo da densidade em uma instância bidimensional hipotética.

$$S\_Dbw(k) = Scat(k) + Dens\_bw(k) \quad (5)$$

$$Scat\ (k) = \frac{1}{k}\sum_{l=1}^{k}\frac{\|\sigma^2(C_l)\|}{\|\sigma^2(X)\|} \quad (6) \qquad \|\sigma^2(C_l)\| = \sqrt{\sum_{r=1}^{p}\left(\frac{1}{n_{C_l}}\sum_{i\in C_l}\left(x_i^r - \bar{x}_{C_l}^r\right)^2\right)}$$

$$Dens\_bw(k) = \frac{1}{k(k-1)}\sum_{s=1}^{k}\left(\sum_{\substack{l=1\\ l\neq s}}^{k}\frac{\text{densidade}(u_{ls})}{\max\left(\text{densidade}(v_l),\text{densidade}(v_s)\right)}\right) \quad (7)$$

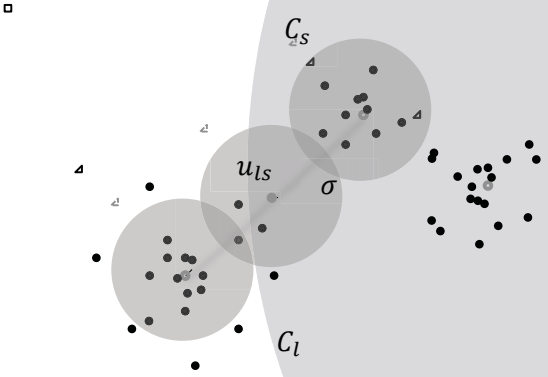

Figura 1: Esquema para cálculo da densidade entre grupos

Segundo Halkidi e Vazirgiannis (2001) este índice apresenta desempenho superior ao de muitos índices propostos na literatura, no que diz respeito à validação de soluções obtidas por algoritmos particionais, hierárquicos ou baseados em densidade. Entretanto, não oferece bons resultados na presença de grupos de formas arbitrárias (não convexas).

Com base no método BRP (do inglês *Best Ranking Position*), proposto por Naldi et al. (2013), três conjuntos de partições base foram então construídos para cada instância, de acordo com os seguintes critérios: conjunto Π_A – melhores soluções de cada algoritmo, segundo cada medida de validação adotada; conjunto Π_B – melhores soluções dos algoritmos K-Means, CLARA e BRKGA-Medoids, segundo cada medida de validação adotada; conjunto Π_C – melhores soluções dos algoritmos MRDBSCAN e BIRCH, segundo cada medida de validação adotada.

Assim, o número de soluções em cada conjunto foi limitado aos intervalos [5;20], [3;12] e [2;8], respectivamente, dado que uma solução foi selecionada de acordo com cada medida de validação. Porém, se a mesma solução (produzida pelo mesmo algoritmo com igual configuração de parâmetros) foi selecionada segundo mais de uma medida de validação, ela não foi novamente incluída no conjunto de partições base.

O critério adotado para estabelecer os grupos Π_B e Π_C foi o método utilizado pelo algoritmo, separando os que realizam agrupamento particional dos demais.

3.4 Combinação de Agrupamentos

Na obtenção das partições consenso foi adotada função consenso baseada em coassociação, ou seja, a similaridade entre objetos foi determinada pelo número de grupos compartilhados entre eles em todas as partições contidas no conjunto de partições base. Tal similaridade representa a força de coassociação entre os objetos e é organizada em uma matriz de coassociação.

Foi aplicado o modelo proposto por Gordon e Vichi (2001), que adota uma abordagem de otimização, utilizando um algoritmo SUMT (*Sequencial Unconstrained Minimization Technique*) para minimizar a função objetivo apresentada na Equação 8, em que M representa o número de partições base; π_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.

$$L(\pi) = \sum_{m=1}^{M} w_m \, d(\pi, \pi_m)^2 \tag{8}$$

Uma vez obtida a partição consenso, assim como as soluções geradas pelos algoritmos de agrupamento, esta precisa ser validada. Segundo Naldi (2011), uma partição consenso é de boa qualidade se apresenta elevada similaridade com as partições base que lhe deram origem.

No presente trabalho, a similaridade entre duas partições foi avaliada através da informação mútua normalizada (NMI, do inglês *Normalized Mutual Information*), proposta por Strehl et al. (2000), que tem a propriedade de ser invariante ao número de grupos e gera valores compreendidos no intervalo $[0; 1]$, porém atinge o valor máximo de 1 se, e somente se, houver uma correspondência direta, objeto a objeto, entre as duas partições π_a e π_b. É obtida através da Equação 9, onde k_a e k_b são os números de grupos das partições π_a e π_b, respectivamente; C_l^a é o l-ésimo grupo da a-ésima partição (π_a); n é o número de objetos da instância X.

$$\phi^{(NMI)}(\pi_a, \pi_b) = \frac{\sum_{l=1}^{k^a}\sum_{l=1}^{k^b} |C_l^a \cap C_s^b| \log\left(\frac{n_o |C_l^a \cap C_s^b|}{|C_l^a||C_s^b|}\right)}{\sqrt{\left[\sum_{l=1}^{k^a} |C_l^a| \log\left(\frac{|C_l^a|}{n}\right)\right]\left[\sum_{s=1}^{k^b} |C_s^b| \log\left(\frac{|C_s^b|}{n}\right)\right]}} \tag{9}$$

A informação mútua normalizada média (*ANMI*, do inglês *Average Normalized Mutual Information*), proposta por Strehl e Ghosh (2003), dada pela Equação 10, permite medir a similaridade entre a partição consenso π_c e um conjunto de partições Π, em que n_π corresponde ao número de partições contidas no conjunto de partições base.

$$\phi^{(ANMI)}(\pi_C, \Pi) = \frac{1}{n_\pi} \sum_{j=1}^{n_\pi} \phi^{(NMI)}(\pi_C, \pi_j) \tag{10}$$

4 Experimentos Computacionais

Os resultados computacionais foram divididos em duas subseções, a primeira traz as soluções geradas pela execução dos algoritmos de agrupamento e construção dos conjuntos de partições base, o que corresponde à primeira etapa da implementação da metodologia proposta. A segunda subseção mostra a avaliação das partições consenso obtidas na 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 10.9.5.

4.1 Bases de Dados

Os algoritmos de agrupamento foram aplicados em 25 instâncias, as quais podem ser divididas em três grupos, quais sejam: (1) compostas a partir de dados do Censo Demográfico 2010 e do Produto Interno Bruto e do Índice de Desenvolvimento Humano dos municípios brasileiros em 2010, disponibilizados nos sites do Instituto Brasileiro de Geografia e Estatística (IBGE – *www.ibge.org.br*) e do Programa das Nações Unidas para o Desenvolvimento (PNUD – *www.pnud.org.br*); (2) instâncias tradicionais da literatura de análise de agrupamento; (3) sintéticas, geradas no software R. O primeiro grupo permite avaliar a aplicabilidade do procedimento proposto em situações reais, onde a estrutura de grupos, se existir, é absolutamente desconhecida. O segundo grupo oferece parâmetro de comparação com outras metodologias e o terceiro permite conhecer o desempenho do procedimento proposto quando a distribuição dos dados é conhecida. A Tabela 1 apresenta as características das 25 instâncias, onde a coluna n indica o número de objetos, a coluna p indica o número de atributos e a coluna f indica o grupo de origem.

Tabela 1: Características das instâncias utilizadas

Instâncias	n	p	f
censo_Pará	128	5	1
censo_Maranhão	190	5	1
censo_Mato Grosso do Sul	73	5	1
censo_Rio de Janeiro	88	5	1
pib_Tocantins	117	4	1
pib_Ceará	164	4	1
pib_Mato Grosso	122	4	1
pib_Espírito Santo	71	4	1
pib_Paraná	369	4	1
idh_norte	449	3	1
idh_nordeste	1780	3	1
idh_centro-oeste	465	3	1
idh_sudeste	1668	3	1
idh_sul	1188	3	1

Instâncias	n	p	f
DS1-200DATA	200	2	2
DS1-gauss9	900	2	2
DS1-iris	150	4	2
DS1-maronna	200	2	2
DS1-ruspini	75	2	2
DS1-spherical_4d3c	400	3	2
DS1-yeast	1484	7	2
DS4-Gamma400	500	3	3
DS4-Normal300	300	2	3
DS4-Uniform400	400	2	3
DS4-Uniform700	700	3	3

Nas instâncias do Censo Demográfico 2010, identificadas pelo prefixo “censo”, os cinco atributos correspondem ao valor do rendimento nominal médio mensal das pessoas de 10 anos ou mais de idade (com rendimento), proporção de domicílios particulares permanentes com banheiro de uso exclusivo dos moradores ou sanitário e esgotamento sanitário via rede geral de esgoto ou pluvial, proporção de domicílios particulares permanentes sem banheiro de uso exclusivo dos moradores, proporção de domicílios particulares permanentes com lixo coletado e proporção de domicílios particulares permanentes com energia elétrica de companhia distribuidora. Nas instâncias do PIB municipal 2010, identificadas pelo prefixo “pib”, 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, identificadas pelo prefixo “idh”, os atributos correspondem às três dimensões do índice, quais sejam: longevidade, educação e renda (Fadel et al., 2014).

Uma vez que todas as instâncias têm variáveis (atributos) apenas quantitativos, foi efetuada uma padronização, a fim de que, na execução dos algoritmos, nenhuma sobressaísse em importância às demais, impactando no cálculo das dissimilaridades. Nesse sentido, cada observação x_{ir} foi transformada no valor padronizado z_{ir} correspondente como apresenta a Equação 11 em que x_{ir} é a i-ésima observação do r-ésimo atributo; μ_r é a média do r-ésimo atributo; σ_r é o desvio padrão do r-ésimo atributo; z_{ir} é o valor padronizado da observação x_{ir}.

$$z_{ir} = \frac{x_{ir} - \mu_r}{\sigma_r} \quad , \qquad \begin{cases} i = 1, \dots, n \\ r = 1, \dots, p \end{cases} \tag{11}$$

4.2 Conjuntos de Partições Base

Os cinco algoritmos de agrupamento produziram, para as 25 instâncias, um total de 12.206 soluções, das quais 319 foram identificadas como melhores soluções e organizadas em conjuntos de partições base, conforme descrito na seção 3.3. A distribuição das melhores soluções por algoritmo de origem é apresentada na Figura 2.

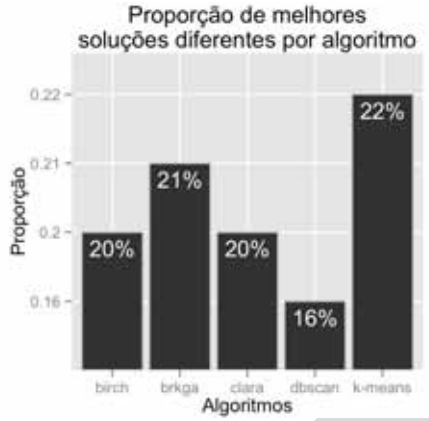

Figura 2: Proporção de melhores soluções diferentes por algoritmo

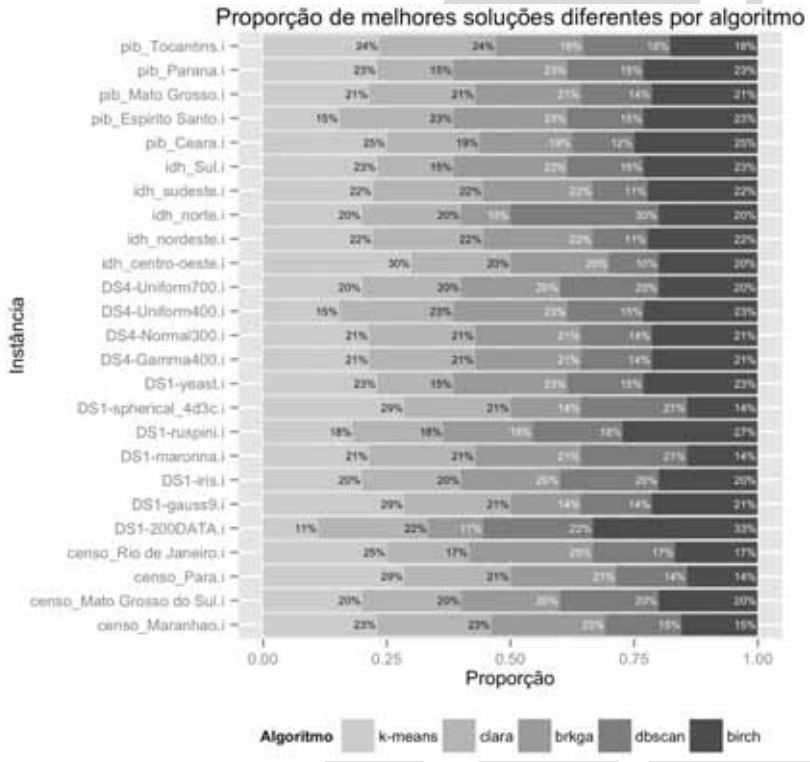

Figura 3: Proporção de melhores soluções por algoritmo, para cada instância

A Figura 3 mostra a contribuição dos algoritmos no provimento de soluções para os conjuntos de partições base de cada instância. Observa-se que apenas a instância "DS1-200DATA" não apresentou predomínio de melhores soluções diferentes obtidas através dos algoritmos particionais, embora esses sejam superiores em número. Tal fato indica que, para esta instância, os algoritmos hierárquico e baseado em densidade foram superiores em diversidade de soluções. A instância "DS1-200DATA" possui apenas 2 atributos, o que permite uma análise visual da distribuição dos seus objetos. Essa é uma instância tradicional da literatura de agrupamento, é especialmente adequada à aplicação de algoritmos particionais, pois seus objetos distribuem-se em grupos bem separados, aproximadamente esféricos e razoavelmente bem definidos, sem a presença de objetos muito dispersos. Devido a tais características, os algoritmos particionais contribuíram em qualidade**,** mas não em diversidade para o conjunto de partições base. Portanto, o predomínio de soluções provenientes dos algoritmos não particionais no conjunto de partições base dessa instância revela a capacidade do procedimento proposto em garantir diversidade de soluções, propriedade extremamente desejável na combinação de agrupamento.

4.3 Partições Consenso

A combinação de agrupamento foi executada de maneira independente para cada um dos conjuntos de partições base construídos e os resultados foram analisados em termos de ANMI, conforme descrito na seção 3.4. A Tabela 2 apresenta um resumo dos resultados obtidos para cada um dos três conjuntos de partições base de cada uma das 25 instâncias. A coluna n_π indica o número de partições base contidas no conjunto, a coluna min indica a ANMI mínima do conjunto, a coluna max indica a ANMI máxima do conjunto e a coluna π_c indica a ANMI da partição consenso obtida.

Tabela 2: ANMI dos conjuntos de partição base

Instâncias	Conjunto Π_A				Conjunto Π_B				Conjunto Π_C			
	n_π	ANMI			n_π	ANMI			n_π	ANMI		
		min	max	π_c		min	max	π_c		min	max	π_c
censo_Pará	14	0,091	0,546	0,518	10	0,457	0,597	**0,638**	4	0,130	0,480	**0,520**
censo_Maranhão	13	0,063	0,563	0,559	9	0,551	0,636	**0,673**	4	0,115	0,323	**0,365**
censo_Mato Grosso do Sul	15	0,135	0,533	0,521	9	0,433	0,636	**0,702**	6	0,264	0,434	**0,467**
censo_Rio de Janeiro	12	0,122	0,546	0,504	8	0,247	0,533	**0,574**	4	0,164	0,463	**0,498**
pib_Tocantins	17	0,134	0,489	**0,491**	11	0,393	0,604	**0,639**	6	0,289	0,383	**0,505**
pib_Ceará	16	0,103	0,599	**0,604**	10	0,458	0,674	**0,703**	6	0,143	0,580	0,561
pib_Mato Grosso	14	0,129	0,564	**0,599**	9	0,543	0,666	**0,691**	5	0,158	0,462	**0,513**
pib_Espírito Santo	13	0,309	0,599	0,580	8	0,445	0,659	**0,694**	5	0,441	0,538	**0,631**
pib_Paraná	13	0,097	0,505	0,462	8	0,482	0,567	**0,582**	5	0,190	0,466	0,395
idh_norte	10	0,075	0,380	**0,459**	5	0,475	0,640	**0,712**	5	0,139	0,329	**0,463**
idh_nordeste	9	0,049	0,476	0,466	6	0,494	0,536	**0,614**	3	0,094	0,329	**0,506**
idh_centro-oeste	10	0,061	0,498	**0,504**	7	0,431	0,557	**0,589**	3	0,089	0,284	**0,497**
idh_sudeste	9	0,028	0,480	**0,543**	6	0,520	0,592	**0,661**	3	0,047	0,351	0,297
idh_sul	13	0,040	0,454	**0,474**	8	0,496	0,556	**0,646**	5	0,081	0,396	0,270
DS1-200DATA	9	0,547	0,859	**0,875**	4	0,620	0,873	**0,905**	5	0,549	0,813	**0,851**
DS1-gauss9	14	0,056	0,650	0,647	9	0,369	0,749	**0,765**	5	0,128	0,416	0,319
DS1-iris	10	0,601	0,740	0,707	6	0,650	0,736	**0,757**	4	0,596	0,735	**0,801**
DS1-maronna	14	0,602	0,782	**0,797**	9	0,588	0,774	**0,799**	5	0,573	0,754	**0,797**
DS1-ruspini	11	0,812	0,943	**0,948**	6	0,811	0,944	**0,953**	5	0,818	0,927	**0,941**
DS1-spherical_4d3c	14	0,527	0,803	**0,817**	9	0,585	0,791	**0,814**	5	0,574	0,778	**0,795**
DS1-yeast	13	0,143	0,489	0,458	8	0,360	0,497	**0,550**	5	0,204	0,532	0,428
DS4-Gamma400	14	0,091	0,490	**0,510**	9	0,361	0,523	**0,585**	5	0,216	0,418	0,340
DS4-Normal300	14	0,078	0,487	0,451	9	0,428	0,532	**0,616**	5	0,175	0,462	0,365
DS4-Uniform400	13	0,050	0,556	**0,587**	8	0,389	0,629	**0,676**	5	0,090	0,639	0,543
DS4-Uniform700	15	0,222	0,614	0,613	9	0,525	0,665	**0,705**	6	0,283	0,629	0,593

As partições consenso cujas ANMI foram destacadas em negrito superaram a ANMI máxima do respectivo conjunto de partições base, indicando que a combinação de agrupamento foi capaz de construir uma solução que representa melhor o conjunto de partições base do que qualquer solução nele contida. Em todas as instâncias, tal condição foi observada em ao menos um dos conjuntos de partições base, sugerindo que o procedimento proposto foi eficaz na obtenção de uma partição consenso de qualidade, no que diz respeito à adequada alocação dos objetos da instância. Mesmo nos casos onde essa condição não foi observada, a ANMI da partição consenso apresentou valor próximo da ANMI máxima do conjunto de partições base.

Além disso, pode-se verificar que, em todas as instâncias, os conjuntos Π_B, compostos apenas pelas soluções obtidas a partir dos algoritmos particionais, deram origem a partições consenso com ANMI superior a ANMI máxima do conjunto de partições base.

No que diz respeito à diversidade, a amplitude entre as ANMI mínima e máxima mostra que, de maneira geral, esta propriedade foi atendida. Nesse aspecto, a instância "DS1-200DATA", sofreu uma análise pormenorizada. A Tabela 3 mostra o número de grupos, k, das soluções selecionadas por cada medida de validação para essa instância. Verifica-se que a silhueta baseada em densidade selecionou soluções pouco apropriadas para os algoritmos CLARA e BIRCH. Todavia, houve grande predominância de soluções de boa qualidade, o que, devido à função consenso adotada, resultou em partições consenso com a correta alocação dos objetos.

Tabela 3: Melhores soluções para a instância "DS1-200DATA"

Algoritmo	$\bar{s}$	S_Dbw	$\overline{dbs}$	dbs_{med}
K-Means	k=3	k=3	k=3	k=3
CLARA	k=3	k=3	**k=14**	**k=14**
BRKGA-Medoids	k=3	k=3	k=3	k=3
DBSCAN	k=3 (r=0.48 d=50)	k=3 (r=0.48 d=50)	k=2 (r=1.62 d=50)	k=2 (r=1.62 d=50)
BIRCH	k=3 (t=2.71 b=20)	k=4 (t=2.30 b=20)	**k=12** (t=0.17 b=10)	**k=12** (t=0.17 b=10)

5 Conclusões

No que diz respeito à composição dos conjuntos de partições base, o conjunto Π_B reuniu os algoritmos que adotam método de agrupamento particional, que, de maneira geral, não são aptos a tratar ruídos e *outliers* ou identificar grupos de formas arbitrárias (Halkidi et al., 2001), implicando sérias limitações às aplicações reais. Porém, a execução da combinação de agrupamentos a partir das partições contidas nesse conjunto apresentou excelentes resultados. Tal desempenho pode ser atribuído à presença das soluções obtidas pelo algoritmo BRKGA-Medoids e à abordagem nele adotada para tratar o problema dos k-medoids, que, segundo Brito et al. (2014), apresenta bom desempenho em encontrar soluções ótimas globais.

As partições consenso obtidas a partir do conjunto Π_A, que reuniu soluções obtidas a partir de três algoritmos particionais, um algoritmo hierárquico e um algoritmo baseado em densidade, quando não superaram o valor da ANMI máxima das partições base, aproximaram-se muito deste, mostrando-se capazes de representar as soluções do conjunto. Este resultado, aliado à diversidade muito superior de soluções do conjunto Π_A em relação aos demais conjuntos, sugere que a composição deste conjunto, dentre as opções estudadas no presente trabalho, seja a mais adequada na obtenção da solução final.

No que diz respeito ao desempenho dos algoritmos, dos bons resultados obtidos pelo algoritmo BIRCH emerge a conclusão de que a metodologia de calibração de parâmetros proposta inicialmente para os parâmetros do algoritmo DBSCAN, por Semaan et al. (2012), é adequada também para este algoritmo.

Em resumo, o procedimento proposto apresentou resultados satisfatórios no que diz respeito à obtenção de uma solução final de boa qualidade, em termos da correta alocação dos objetos, para o problema de agrupamento. O sucesso de sua aplicação a instâncias com características amplamente diferentes sugere grande versatilidade e robustez do método.

6 Referências Bibliográficas

[1] Brito J, Semaan G & Brito R. 2014. Resolução do Problema dos K-Medoids via Algoritmo Genético de Chaves Aleatórias Viciadas. In: *Anais do XVII Simpósio de Pesquisa Operacional e Logística da Marinha*, 1(1):50-61.

[2] Ester M, Kriegel H, Jörg S & Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* [edited by E. Simoudis, J. Han and U. Fayyad], AAAI Press, 226:231.

[3] Fadel A, Semaan G & Brito J. 2014. Um Estudo da Aplicação de Técnicas de Combinação de Agrupamentos. In: *Anais do XVII Simpósio de Pesquisa Operacional e Logística da Marinha*, 1(1):188-200.

[4] Gordon A & Vichi M. 2001. Fuzzy Partition Models for Fittind a Set of Partitions. *Psychometrika*, 66(2):229-248.

[5] Halkidi M & Vazirgiannis M. 2001. Clustering Validity Assessment: Finding the optimal partitioning of a data set. *ICDM 2001, Proceedings IEEE International Conference on Data Mining*, IEEE, 187-194.

[6] Halkidi M, Batistakis Y & Vazirgiannis M. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2/3):107-145.

[7] Han J, Kamber M & Pei J. 2012. Data Mining: Concepts and Techniques. 3. ed. Waltham: Morgan Kaufmann Publishers.

[8] Hartigan J & Wong M. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 28(1):100-108.

[9] Kaufman L & Rousseeuw P. 1986. Clustering Large Data Sets. In: *Pattern Recognition in Practice 2* [edited by E. Gelsema and L. Kanal], Elsevier, 425-437.

[10] Menardi G. 2011. Density-based Silhouette diagnostics for clustering methods. *Statistics and Computing*, 21(3):295-308.

[11] Naldi M, Carvalho A & Campello R. 2013. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, 27(2):259-289.

[12] Naldi M. 2011. Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados. Ph.D thesis, USP, São Carlos, SP, Brazil.

[13] Reeves C. 2010. Genetic Algorithms. In: *Handbook of Metaheuristics*, International Series in Operations Research & Management Science 146 [edited by M. Gendreau and J. Potvin], Springer Science Business Media, 109-140.

[14] Rousseeuw P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53-65.

[15] Semaan G, Vasconcelos R, Brito J & Ochi L. 2012. 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)*, 10(4):242-262.

[16] Strehl A & Ghosh J. 2003. Cluster Ensembles: A Knowledge Reuse Framework for Combining Partitionings. *The Journal of Machine Learning Research*, 3:583-617.

[17] Strehl A, Ghosh J & Mooney R. 2000. Impact of Similarity Measures on Web-page Clustering. AAAI Technical Report WS-00-01.

[18] Tan P, Steinbach M & Kumar V. 2009. Introduction to Data Mining. Pearson Addison Wesley.

[19] Zhang T, Ramakrishnan R & Livny M. 1997. BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, 1:141-182.