Aplicação da heurística híbrida geração de colunas com *path-reliking* para problemas de agrupamentos

Rudinei Martins de Oliveira
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, SP
rudmart@gmail.com

Antonio Augusto Chaves
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, SP
antoniochaves@gmail.com

Luiz Antonio Nogueira Lorena
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos, SP
lorena@lac.inpe.br

Geraldo Regis Mauri
Universidade Federal do Espírito Santo - UFES
Alegre, ES
geraldo.mauri@ufes.br

RESUMO

Este artigo apresenta uma heurística híbrida para a resolução do problema de agrupamentos. Um problema de agrupamentos pode ser definido como o processo de separação de um conjunto de dados em grupos de tal forma que os membros de cada grupo são semelhantes entre si. O método proposto baseia-se na aplicação de uma técnica de geração de colunas com *path-relinking*. O processo de geração de colunas gera as soluções do problema de agrupamentos resolvendo o problema de *p*-medianas. Essas soluções são melhoradas pelo método *path relinking*. As soluções finais são ainda validadas pelo índice Crand.

PALAVRAS CHAVE. Geração de Colunas, Problema *p*-medianas, Problema de Agrupamentos.

Área Principal: OC - Otimização Combinatória

ABSTRACT

This paper presents a hybrid heuristic for solving clustering problems. The clustering problem can be defined as the process of separating a set of objects into groups such that members of a group are similar to each other. The method is based on the application of a column generation technical with path-relinking. The column generation process generates solutions for the clustering problem solving the problem of *p*-medians. The column generation process generates solutions that are further improved by the path relinking method. The finale values are further validated by Crand index.

KEYWORDS. Column Generation, *p*-median problems, clustering problems.

Main Area: CO - Combinatorial Optimization

1. Introdução

Problemas de agrupamentos envolvem a análise de grandes volumes de dados podendo ter várias aplicações no dia a dia, tais como, identificar pacientes com quadros clínicos semelhantes, agrupar problemas de saúde pública, encontrar padrões no comportamento dos clientes que fazem compras de produtos semelhantes, verificar o comportamento dos usuários em determinada página da *web*, agrupar ações com as mesmas flutuações de preço entre outros (Babaki et al. 2014).

Assim, dado um conjunto de dados que apresenta atributos e representa algo do mundo real, o problema de agrupamentos consiste no processo de separar o conjunto de dados em grupos, tal que os membros de cada grupo sejam similares entre si. A dificuldade ocorre quando alguns atributos não são definidos de forma clara ou geram interpretações equivocadas. A questão a ser resolvida é a maneira adequada de agrupar esses dados.

Para que os dados sejam agrupados é necessário identicar quão próximos ou distantes eles estão. Um dos caminhos seria a criação de uma matriz de distâncias e em seguida determinar a similaridade usando uma métrica. Se os dados estão próximos ou não, depende da escolha da métrica utilizada. Métricas como a distancia Euclidiana, *City Block*, Pearson e Cosseno usadas neste artigo auxiliam a definir se um objeto é similar a outro.

Para solução do problema de agrupamentos é necessário definir os grupos que armazenarão as seleções dos objetos similares. Os grupos são obtidos pela técnica de geração de colunas para solução do problema de *p*-medianas. Soluções de *p*-medianas minimizam a soma das distâncias entre objetos e sua mediana mais próxima.

Nesse contexto, o objetivo deste artigo é resolver o problema de agrupamentos, utilizando a heurística híbrida geração de colunas com *path-relinking*. A validação dos resultados foi realizada através do cálculo do valor do índice de CRand, usado para comparar a eficiência dos métodos.

O artigo está organizado da seguinte forma: a seção 2 apresenta uma breve revisão bibliográfica sobre o problemas de agrupamentos, a seção 3 apresenta a descrição geral da técnica de Geração de Colunas para o problema de *p*-medianas. A seção 4 descreve o método proposto para resolução do problema de agrupamentos. Os resultados computacionais obtidos são apresentados na seção 5.

2. Revisão da literatura para agrupamentos

O problema de agrupamentos tem sido muito estudado e pesquisadores utilizam-se de uma grande diversidade de métodos buscando resolvê-lo. Como exemplo, o trabalho de Rand (1971) propõe critérios que isolam aspectos do desempenho de um método, tais como, retorno, sensibilidade e estabilidade. Esses critérios dependem de uma medida de similaridade entre dois agrupamentos diferentes do mesmo conjunto de dados. A medida considera essencialmente o modo como cada par de pontos de dados é atribuído em cada conjunto.

Handl et al. (2005) mostram a grande quantidade de técnicas disponíveis para a validação dos resultados obtidos para o problema de agrupamento. O foco principal do trabalho é a análise de dados em pós-genômica. Os autores usam dados biológicos sintéticos e reais para demonstrar os benefícios, e também alguns riscos da validação.

O trabalho de Al-Sultan (1995) apresenta uma busca tabu para resolver o problema de agrupamentos. A função objetivo minimiza a distância entre os pontos pertencentes a um mesmo agrupamento e o seu centro. Mitra e Banka (2006) introduziram um método *multi objetivo biclustering* evolutivo com estratégias de busca local.

Chang et al. (2009) propõem um algoritmo de separação baseado em um algoritmo genético com rearranjamento de genes para o problema *k-means clustering*, o qual busca remover degenerações para o propósito de uma busca mais eficiente. Um operador de *crossover* que explora similaridades entre os cromossomos em uma população também é apresentado.

Nascimento et al. (2010) apresentaram uma formulação matemática e um algoritmo *greedy randomized adaptive search procedure* (GRASP) para resolver o problema de agrupamentos usando dados biológicos. Os resultados computacionais foram comparados com a aplicação do

CPLEX, *k-means*, *k-medians* e *partitioning around medoids* (PAM). O índice CRand é usado para comparar os métodos.

3. Geração de colunas para o problema de *p*-medianas

A heurística híbrida para resolver o problema de agrupamentos começa a partir de um conjunto de dados e sem qualquer informação dos padrões e constroem grupos que contenham características semelhantes entre seus objetos. Os grupos são obtidos pela técnica de Geração de Colunas (GC) para solução do problema de *p*-medianas. Assim, uma possível solução viável para o problema de agrupamentos resultaria da separação de um grupo em p subgrupos distintos, na qual cada grupo contém uma facilidade e os nós alocados a ela.

Dada a matriz de distâncias $[d_{ij}]_{n\times n}$, de acordo com Senne et al. (2007), o problema de p-medianas pode ser modelado como o seguinte problema de particionamento de conjuntos:

$$\textbf{Minimizar:} \quad \sum_{k=1}^{m} c_k y_k \tag{1}$$

$$\textbf{Sujeito a:} \quad \sum_{k=1}^{m} A_k y_k = 1 \tag{2}$$

$$\sum_{k=1}^{m} y_k = p \tag{3}$$

$$y_k \in \{0,1\} \tag{4}$$

O conjunto $S = \{S_1, S_2, ..., S_m\}$, é formado por subconjuntos dos objetos $N = \{1, ..., n\}$; $M = \{1, 2, ..., m\}$ é o conjunto dos índices correspondentes as colunas; o subconjunto S_k corresponde a uma coluna A_k do conjunto de restrições 2; $A_k = [a_i]_{n\times 1}$, para $k \in M$; com $a_i = 1$ se $i \in S_k$, e $a_i = 0$ caso contrário; $c_k = Min_{i\in S_k}\left(\sum_{j\in S_k} d_{ij}\right)$, para $k \in M$; y_k são as variáveis de decisão, com $y_k = 1$ se o subconjunto S_k é escolhido e $y_k = 0$ caso contrário. Para cada conjunto S_k, a escolha da mediana é realizada pelo cálculo do custo (c_k).

Como o número de colunas pode ser muito grande, o problema a ser resolvido é uma relaxação de programação linear de (1) - (4) conhecido como Problema Mestre (PM) e definido da seguinte forma:

$$\textbf{Minimizar:} \quad \sum_{k=1}^{m} c_k y_k \tag{5}$$

$$\textbf{Sujeito a:} \quad \sum_{k=1}^{m} A_k y_k \geq 1 \tag{6}$$

$$\sum_{k=1}^{m} y_k = p \tag{7}$$

$$y_k \in [0,1] \tag{8}$$

Depois de definir um conjunto de colunas iniciais, o PM é resolvido e os seus custos duais finais (μ_i, para $i = 1, ..., n$) e ρ são usados para gerar novas colunas ($\beta j = [\beta_{ij}]_{n\times 1}$), resolvendo o seguinte subproblema:

$$Min_{j\in N}\left[Min_{\beta_{ij}\in\{0,1\}} \sum_{i=1}^{n} (d_{ij} - \mu_i)\beta_{ij}\right] \tag{9}$$

O problema (9) é resolvido considerando cada $j \in N$ como uma mediana, e $\beta_{ij} = 1$, se $(d_{ij} - \mu_i \leq 0)$ e $\beta_{ij} = 0$, se $(d_{ij} - \mu_i) > 0$. Para o novo conjunto S_j é definido o vértice i tal que $\beta_{ij} = 1$ para o subproblema (9). Então a coluna $\left[\frac{\beta_j}{1}\right]$ é adicionada ao PM se o valor da solução do subproblema (9) é menor que ρ. Todas as colunas encontradas satisfazendo a desigualdade (10) para $j = 1, ..., n$, podem ser adicionadas ao conjunto de colunas, acelerando o processo de GC.

$$\left[Min_{\beta_{ij} \in \{0,1\}} \sum_{i=1}^{n} (d_{ij} - \mu_i)\beta_{ij}\right] < \rho \tag{10}$$

O algoritmo de GC está resumido no fluxograma da Figura 1. Após a definição de um conjunto inicial de colunas, o *software* CPLEX (ILOG, 2009) é usado para resolver o PM e obter os valores duais μ_i, $i = 1, \ldots, n$ e ρ. Esses valores são utilizados para gerar novas colunas ($\left[\frac{\beta_j}{1}\right]$) por meio da solução do subproblema (9). Todas as colunas ($\left[\frac{\beta_j}{1}\right]$) que satisfazem a desigualdade (10) (para $i = 1, \ldots, n$) são colunas de entrada para o PM. Algumas colunas com alto custo reduzido são removidas para manutenção das boas soluções obtidas pelo PM. O processo iterativo termina se não forem encontradas novas colunas.

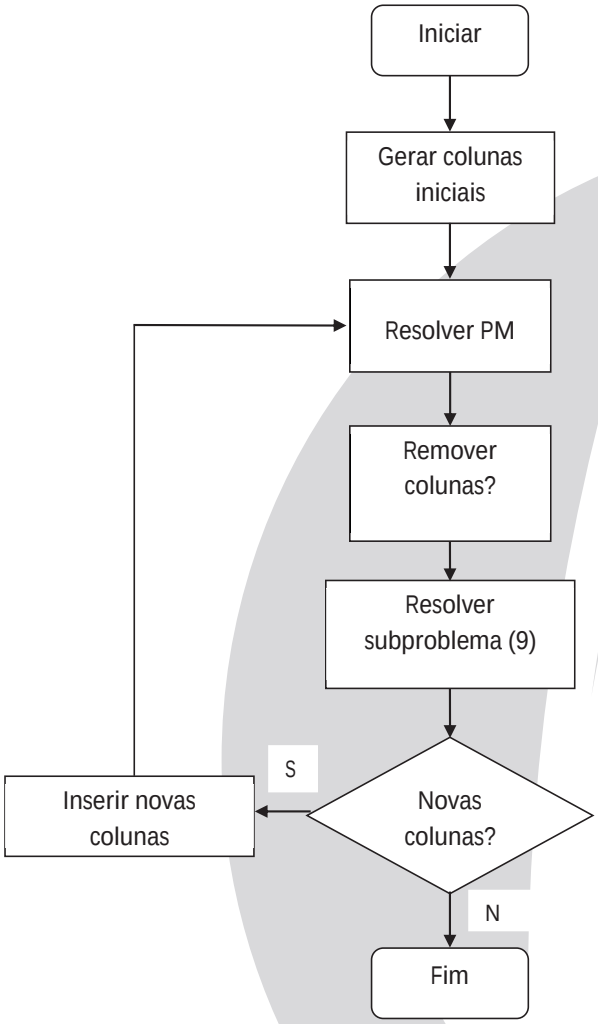

Figura 1: Algoritmo GC.

4. Heurística híbrida para o problema de agrupamentos (HGC-PR)

A heurística híbrida examinada neste artigo pode ser classificada como uma combinação de meta-heurística e GC. Duas possibilidades básicas são exploradas na literatura: aplicar meta-

heurística no subproblema gerador de colunas ou diretamente no PM, para gerar colunas iniciais ou gerar colunas candidatas a participar do GC (MAURI; LORENA, 2007; PIRKWIESER; RAIDL, 2010; FILHO; LORENA, 2000; MASSEN et al., 2013). A heurística híbrida proposta neste trabalho pode ser vista como uma terceira opção na qual o processo de GC gera soluções do problema de agrupamentos que são melhoradas a seguir pelo método *path relinking* - PR (Glover e Martí, 2000).

A heurística proposta gera soluções viáveis de p-medianas (agrupamentos) obtidas no processo de GC. O valor da função custo (c_k) do subconjunto S_k no modelo (5) - (8), será o valor obtido com a soma das distâncias de todas as arestas para os vértices nos agrupamentos (uma clique), além do calculado em Senne et al. (2007), que seria a soma das distâncias dos pontos para a mediana mais próxima (p-medianas).

O valor de CRand varia entre $[-1, 1]$ e quanto mais próximos de 1 as partições serão mais similares. Assim, sejam U e V duas partições, tal que N é o número de objetos de um conjunto de dados, n_i é o número de objetos no grupo i de U, n_j o número de objetos no grupo j de V, os índices i e j variam de acordo com o número de grupos das partições. Assim, o índice CRand proposto por Hubert e Arabie (1985) é definido da seguinte forma:

$$CRand = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}/\binom{N}{2}} \tag{11}$$

O algoritmo PR é usado para intensificar e diversificar a busca em um grupo de soluções. Ele realiza movimentos exploratórios em uma vizinhança a partir de uma solução inicial buscando atingir uma solução guia. O objetivo é encontrar as melhores soluções no caminho que conectam estas soluções. Os movimentos gradualmente introduzem informações dos atributos da solução guia para a solução inicial (RESENDE e RIBEIRO, 2005).

A Figura 2 apresenta um exemplo do PR usado nesse artigo. Supõe-se que a solução inicial (5, 10, 15) e a solução guia (6, 11, 16) são as medianas de cada solução. O PR calcula a diferença (Δ) entre as soluções (medianas), ou seja, o número de posições diferentes entre elas. A solução inicial e a solução guia diferem em três posições. A partir das próximas três soluções possíveis, tal que $\Delta = 3$, o PR escolhe a melhor solução de p-medianas e assim por diante até atingir $\Delta = 0$, na qual o processo é finalizado com a melhor solução no caminho.

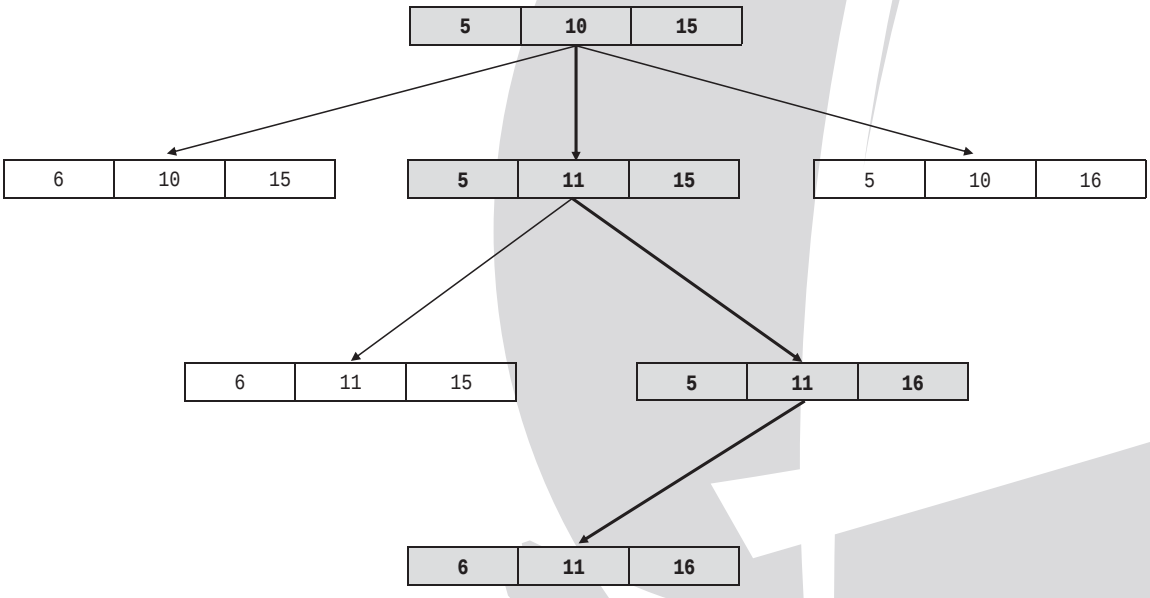

Figura 2: Aplicação do PR.

A HGC-PR utiliza o PR em soluções intermediárias y_j (colunas $\left[\frac{y_j}{1}\right]$) do PM. As soluções do PM são consideradas as soluções iniciais e guia a cada trinta iterações do processo de GC (Seção 3). Ao final de execução do algoritmo todas as soluções de agrupamentos geradas (S_{sol}) são avaliadas pelo índice CRand.

As colunas de inicialização do algoritmo HGC-PR são obtidas utilizando as colunas encontradas pela heurística de solução inicial (HSI) (Figura 3). A HSI começa com a leitura dos dados e cálculo da matriz de distâncias entre os nós, construindo dessa forma, um conjunto inicial de colunas para o PM. Os grupos individuais são construídos (S_j) ($j = 1, ..., p$), compostos pela mediana j, escolhida de forma aleatória, e os vértices mais próximos à mediana. Os grupos criados são convertidos em colunas de uns (se $i \in S_j$) e zeros (se $i \notin S_j$), e as colunas são adicionadas ao conjunto inicial de colunas de PM. Todo o processo é repetido até que o número de colunas (Num_col) atinja o número máximo de colunas geradas (Num_Max_Col).

Solução Inicial:
1 Ler dados();
2 Cálculo da matriz de distâncias;
3 Dado Num_Max_Col o número máximo de GC;
4 $Num\_col \leftarrow 0$;
5 Repita
6 Dado $P = (n_1, ..., n_p)$ um conjunto de vértices escolhidos de forma aleatória
7 Para (j=1, ..., p) faça
8 $S_j \leftarrow \{n_j\} \cup \left\{ q \in N - P | d_{qn_j} = \min_{t \in P} \{d_{qt}\} \right\}$
9 $c_j \leftarrow \min_{t \in S_j} \left\{ \sum_{i \in S_j} \{d_{it}\} \right\}$
10 Para $(i = 1, ..., n)$ faça
11 Se $(i \in S_j)$, faça $a_{ij} \leftarrow 1$;
12 Se $i \notin S_j$, faça $a_{ij} \leftarrow 0$;
13 Adicione a coluna $\left[\frac{A_j}{1}\right]$ ao conjunto inicial de colunas;
14 $Num\_col \leftarrow Num\_col + 1$;
15 Enquanto $(Num\_col < Num\_Max\_Col)$;
16 Fim

Figura 3: Algoritmo HSI
Adaptada de Pereira et al. (2007).

5. Resultados

Os dados utilizados são no total de 5 instâncias, sendo eles: Íris, BreastA, BreastB, DLBCLA e DLBCLB. O dado de Íris foi obtido no repositório UCI (ABBASI; YOUNIS, 2007). Os dados de BreastA, BreastB, DLBCLA e DLBCLB são do repositório de dados do programa de câncer (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi).

Os experimentos computacionais foram executados em um PC AMD Athlon 64 bits, processador dual core com 2.5 GHz e 4 GB de memória RAM. Os parâmetros utilizados foram definidos de forma empírica. A seguir são apresentados os que obtiveram os melhores resultados.

- número de colunas iniciais geradas foram 500,
- número máximo de iterações = 500 (ou até que o algoritmo atinja um máximo de 10.000 colunas geradas),
- o algoritmo GC foi executado 100 vezes para cada conjunto de dados,

- o número de medianas variou de 2 a 30. Durante os experimentos verificou-se que os resultados pioraram com valores acima de 10 medianas. Assim, os resultados finais foram realizados para até 10 medianas, e os melhores resultados são apresentados nas tabelas seguintes.

A heurística propostas considerou os dados representados em forma de grafos completos ponderados. Os pesos das arestas (distâncias ou dissimilaridades) são calculados por quatro métricas: A distância euclidiana, a distância City Block, a correlação de Pearson e a correlação cosseno.

Definindo x_{ik} como o k-ésimo atributo do objeto i e n_a como o número de atributos de um conjunto de dados, as métricas utilizadas são definidas da seguinte forma:

- **Euclidiana**: é a distância geométrica entre os objetos.

$$d_{ij} = \sqrt{\sum_{k=1}^{n_a} (x_{ik} - x_{jk})^2} \tag{12}$$

- **City block ou Manhattan**: é a soma das diferenças absolutas entre os atributos de dois objetos.

$$d_{ij} = \sum_{k=1}^{n_a} |x_{ik} - x_{jk}| \tag{13}$$

- **Correlação de Pearson**: mede o grau de correlação entre dois objetos produzindo um valor entre $[-1, 1]$ (direção negativa ou positiva) no caso de valor zero não existe dependência linear entre os objetos.

$$c_{ij} = \frac{n_a \sum x_{ik}x_{jk} - \sum x_{ik} \sum x_{jk}}{\sqrt{n_a \sum x_{ik}^2 - (\sum x_{ik})^2}\sqrt{n_a \sum x_{jk}^2 - (\sum x_{jk})^2}} \tag{14}$$

  A dissimilaridade entre dois objetos i e j pela correlação de Pearson é da forma: $d_{ij} = 1 - |c_{ij}|$.

- **Correlação cosseno**: é a correlação geométrica definida pelo ângulo entre dois objetos.

$$D_{ij} = \frac{\sum_{k=1}^{n_a} x_{ik}x_{jk}}{\sum_{k=1}^{n_a} x_{ik}^2 \sum_{k=1}^{n_a} x_{jk}^2} \tag{15}$$

  O valor D_{ij} está entre $[-1, 1]$. Se $D_{ij} = 1$ significa que o ângulo entre dois objetos é $0°$. Se $D_{ij} = -1$ significa que o ângulo entre os objetos é $180°$. Assim, a dissimilaridade entre dois objetos i e j pela correlação cosseno é: $d_{ij} = 1 - |D_{ij}|$.

As Tabelas 1 a 4 apresentam a comparação dos valores de CRand para os métodos propostos neste artigo para as duas formas de cálculo da função custo na solução do problema de agrupamentos: a somatória entre todos os pontos dos grupos (clique) e a somatória de todos os pontos à mediana mais próxima (p-mediana tradicional). Além disso, são mostradas as diferenças entre eles. Os valores apresentados nas tabelas são os melhores valores de CRand para todas as instâncias e mostram que em geral, o método proposto encontrou bons índices de CRand para todas as amostras analisadas.

Analisando os resultados, pode-se observar que, para a maioria das soluções, o resultado foi melhor quando se usou a somatória da clique no cálculo da fução objetivo. Para a distância

city block usando clique a HGC-PR apresentou os melhores resultados em quatro instâncias (Íris, BreastA, BreastB e DLBCLB). Para a distância euclidiana usando clique a HGC-PR apresentou os melhores resultados em cinco instâncias (Íris, BreastA, BreastB, DLBCLA e DLBCLB). Para a correlação de Pearson a usando clique a HGC-PR foi melhor em quatro instâncias (Íris, BreastA, BreastB e DLBCLA). Para a correlação Cosseno a usando clique a HGC-PR foi melhor em três instâncias (Íris, BreastB e DLBCLA).

Tabela 1: Diferenças para a distância *City Block*.

Dados	GC-PR		
	Clique	*p*-medianas	Melhorias
	Crand	CRand	(%)
Íris	0,904	0,759	19,10
BreastA	0,722	0,660	9,39
BreastB	0,590	0,347	70,03
DLBCLA	0,674	0,740	-8,92
DLBCLB	0,702	0,608	15,46

Tabela 2: Diferenças para a distância Euclidiana.

Dados	GC-PR		
	Clique	*p*-medianas	Melhorias
	Crand	CRand	%
Íris	0,868	0,760	14,21
BreastA	0,610	0,567	7,58
BreastB	0,590	0,238	147,90
DLBCLA	0,508	0,395	28,61
DLBCLB	0,689	0,545	26,42

Tabela 3: Diferenças para a correlação de Pearson.

Dados	GC-PR		
	Clique	*p*-medianas	Melhorias
	Crand	CRand	%
Íris	0,922	0,886	4,06
BreastA	0,587	0,496	18,35
BreastB	0,661	0,403	64,02
DLBCLA	0,685	0,642	6,70
DLBCLB	0,117	0,159	-26,42

Tabela 4: Diferenças para a correlação Cosseno.

Dados	GC-PR		
	Clique	*p*-medianas	Melhorias
	Crand	CRand	%
Íris	0,885	0,670	32,09
BreastA	0,638	0,759	-15,94
BreastB	0,444	0,195	127,69
DLBCLA	0,619	0,055	1.025,45
DLBCLB	0,111	0,149	-25,50

6. Conclusões

Para resolver o problema de agrupamentos na literatura são usados métodos como a busca tabu (Al-Sultan 1995), algoritmo multi objetivo evolutivo com estratégias de busca local (Mitra e Banka 2006), *k-means* (Chang et al. 2009), algoritmo *greedy randomized adaptive search procedure* (GRASP) (Nascimento et al. 2010), e geração de colunas (Oliveira et al. 2014).

A proposta deste artigo foi desenvolver um método combinando geração de colunas com *path-relinking* para resolver o problema de agrupamentos de forma satisfatória. A solução esperada é a separação do conjunto de dados em grupos, tal que os objetos pertencentes ao mesmo grupo sejam similares. A configuração dos grupos foi obtido pela solução do problema *p*-medianas pela técnica geração de colunas. Para este problema de *p*-medianas, além de usar o cálculo do custo na função objetivo (c_k), analisou-se também quando c_k era definido a partir de uma clique. Assim, como é possível observar nas Tabelas de 1 a 4, quando se usou a clique na função custo (c_k) os resultados foram mais favoráveis. De modo geral, os resultados obtidos demonstraram que o método proposto pode ser usado como uma nova ferramenta para solução do problema de agrupamentos abordado nesse artigo.

Apesar dos bons resultados apresentados nesse artigo, o método utilizado pode ser melhorado. Assim, para um trabalho futuro pretende-se estudar o problema de agrupamentos abordando o caso de agrupamentos com *outliers* e agrupamentos com restrições (*must-link* ou *cannot-link*). O problema de agrupamentos com *outliers* consiste em estudar a maneira de selecionar os *outliers*. No problema agrupamentos com restrições antes de agrupar os dados deve-se definir os objetos a serem ligados os *must-link* (estar juntos no mesmo grupo) ou não ligados os *cannot-link* (separados em grupos distintos). *Must-link* indica que, se um objeto estiver associado a um grupo, o outro deve estar também, já se for não ligado, indica que se um objeto não estiver associado ao grupo o outro não deve estar também.

Neste contexto, a proposta para melhora do método é baseada na aplicação da técnica *Clustering Search* (CS), utilizando a heurística de geração de colunas como geradora de soluções. O algoritmo geração de colunas poderá criar as soluções iniciais. Soluções estas que seriam encaminhadas ao processo de agrupamento do *clustering search*. Dessa forma, espera-se que haja uma melhoria na criação dos grupos facilitando assim, a atuação do algoritmo de busca local que compõe o CS.

7. Agradecimentos

Os autores agradecem ao CNPq (processos 303052/2013-9, 301836/2014-0 e 454569/2014-9) e à CAPES pelo suporte financeiro.

Referências

Abbasi, A. A.; Younis, M. (2007), A survey on clustering algorithms for wireless sensor networks, *Computer Communications*, 30, (14-15), 2826-2841.

Al-Sultan, K. S. (1995), A tabu search approach to the clustering problem, *Pattern Recognition*, 28 (9), 1443-1451.

Babaki, B. e Guns, T. e Nijssen, S. (2014), Constrained Clustering Using Column Generation *Springer International Publishing*, 8451, 438-454.
Chang, Dong-Xia, e Zhang, Xian-Da e Zheng, Chang-Wen (2009), A genetic algorithm with gene rearrangement for K-means clustering *Pattern Recognition*, 42 (7), 1210-1222.
Filho, G. R. e Lorena, L. A. N. (1985), Constructive Genetic Algorithm and Column Generation: an Application to Graph Coloring.
Glover, F, e MartÃ, R. (2000), Fundamentals of scatter search and path relinking, *Control and Cybernetics*, 39, 653-684.
Handl, J., Knowles, J. e Kell, D. B. (2005), Computational cluster validation in post-genomic data analysis *Bioinformatics*, 21 (12), 3201-3212.
Hubert, L. e Arabie, P. (1985), Comparing partitions *Journal of Classification*, 2, 218–218.
ILOG (2009), ILOG CPLEX 12.1 *user's manual*.
Nascimento, M. C. V., Toledo, F. M. B. e Carvalho, A. C. P. L. F. (2010), Investigation of a new GRASP-based clustering algorithm applied to biological data, *Computers & Operations Research*, 37 (8), 1381-1388.
Massen, F. e López, I. M. e Stutzle, T. e Deville, Y. (2013), Experimental Analysis of Pheromone-Based Heuristic Column Generation Using irace, *Lecture Notes in Computer Science*, 7919, 92-106.
Mauri G. R. e Lorena L. A. N. (2007), A new hybrid heuristic for driver scheduling, *International Journal of Hybrid Intelligent Systems*, 4, 39-47.
Mitra, S. e Banka, H. (2006), Multi-objective evolutionary biclustering of gene expression data, *Pattern Recogn*, 39 (12), 2464-2477.
Pirkwieser, S. e Raidl, G. R. (2010), Multilevel Variable Neighborhood Search for Periodic Routing Problems, *Lecture Notes in Computer Science*, 6022, 226-238.
Rand, W. M. (1971), Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, 336 (66), 846-850.
Resende, M. G. C. e Ribeiro, C. (2005), GRASP with path-relinking: recent advances and applications, *Springer*, 29–63.
Oliveira, R. M., Lorena, L. A. N., Chaves, A. A., and Mauri, G. R. (2014), Hybrid heuristics based on column generation with path-relinking for clustering problems, *Expert Systems with Applications*,41 (11), 5277 - 5284.
Pereira, M. A. e Lorena, L. A. N. e Senne, E. L. F. (2007), A column generation approach for the maximal covering location problem, *International Transactions in Operational Research*,14, 4, 349–364.
Senne, E. L. F., Lorena, L. A. N e Pereira, M. A. (2007), A simple stabilizing method for column generation heuristics: an application to p-median location problems, *International Journal of Operations Research*, 4, 1-9.