Multiplicação lagrangena com divisão em clusters e o problema da diversidade máxima

Carlos Renato S. de Almeida Jr[1] e Luiz Antonio Nogueira Lorena[2].

Instituto Nacional de Pesquisas Espaciais - INPE,
Laboratório Associado de Computação e Matemática Aplicada – CAP/LAC,
São José dos Campos – SP – Brasil

cr_seabra@yahoo.com.br, Lorena@lac.inpe.br

Abstract. *The Maximum Diversity Problem consists in select the 'm' most distinguished elements, according to studied parameters, between the 'n' elements of a given set. Many techniques have already been studied to solve this problem, such as GRASP heuristics, Tabu Search, Genetic Algorithm and others. This work's objective is to contribute with the study of applications of graph partitioning and lagrangean relaxation.*

Resumo. *O problema da diversidade máxima consiste em selecionar os 'm' elementos mais distintos, de acordo com os parâmetros estudados, dentre os 'n' elementos de um conjunto determinado. Muitas técnicas já foram estudadas para resolver esse problema, como utilização de heurísticas GRASP, Tabu Search, Algoritmo Genético e outras. Este trabalho tem o objetivo de contribuir com o estudo de aplicações de técnicas de particionamento de grafos e relaxação lagrangeana.*

1. Introdução

O Problema da Diversidade Máxima (PDM) consiste em selecionar um determinado número de elementos pertencentes a um dado conjunto, de modo que o subconjunto constituído pelos itens selecionados seja o mais diversificado possível. Isto é, seja N um conjunto com 'n' elementos e dada a matriz de distâncias que representa a diferença entre os elementos de N, deseja-se selecionar um subconjunto $M \subset N$, com 'm' elementos, tal que 'm' < 'n', de modo que M possua a maior diversidade possível.

Uma abordagem interessante, concebida por Glover et al. (1995), tem o objetivo de otimizar a aplicação de recursos na preservação de espécies garantindo a maior diversidade biológica possível respeitando limites de recursos. Mais aplicações já citadas na literatura são o gerenciamento de recursos humanos e mineração de dados em Aringhieri e Cordone (2006), VLSI design e tratamento médico em Aringhieri et al. (2006) dentre outras.

O PDM foi introduzido por Glover et al. (1977), quando o autor apresenta uma formulação matemática inteira para o problema. Os mesmos autores demonstraram que o PDM pertence à classe de problemas NP-Hard através da redução do problema K-Cliques (Ghosh 1996). Em 1993, Glover et al. (1993) sugerem a linearização da formulação quadrática do PDM.

Uma das técnicas mais estudadas para aplicações em PDM's é a utilização de heurísticas GRASP (Greedy Randomized Adptative Search Procedure) para o cálculo de soluções (Silva et al. (2003); Silva et al. (2004a); Aringhieri et al. 2005; Duarte et al. (2007b) e Silva et al. (2007)).

Em 2006, Aringhieri e Cordone 2006 atingem bons resultados propondo um algoritmo TS com um mecanismo de memória e duas metaheurísticas denominadas Variable Neighbohood Search (VNS) e Scatter Search (SS) e Aringhieri et. al (2006) apresentam um estudo de cálculo de limitantes semidefinidos para PDM's.

Duarte et al. (2007a) apresentam uma solução exata para o PDM eficiente para casos com até 100 elementos podem ser resolvidos utilizando-se métodos exatos.

3. Modelos

Dada a matriz de diversidade $D = \lfloor d_{ij} \rfloor_{nxn}$ simétrica, onde d_{ij} corresponde à diversidade entre os elementos i e j, a formulação quadrática do PDM é:

$$v(PDM) = \max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_i . x_j . d_{ij}$$

sujeito a

$$\sum_{i=1}^{n} x_i = m$$

$$x_i \in \{0,1\}, \ \forall i = 1 \ldots n$$

No processo de linearização a seguir, as variáveis $x_i x_j$ são substituídas pelos termos y_{ij} e restrições que garantem que $y_{ij} = x_i x_j$:

$$v(PDM) = \max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} . y_{ij}$$

sujeito a

$$\sum_{i=1}^{n} x_i = m$$

$$x_i + x_j - y_{ij} \leq 1, \ \forall i, j \in Q$$

$$y_{ij} - x_i \leq 0, \ \forall i, j \in Q$$

$$y_{ij} - x_j \leq 0, \ \forall i, j \in Q$$

$$y_{ij} \in \{0,1\}, \ \forall i, j \in Q$$

$$x_i \in \{0,1\}, \ \forall i = 1 \ldots n$$

$$Q = \{(i, j) \mid i, j \in N, i < j\}$$

3.1. LagClus – Relaxação Lagrangeana com divisão em Clusters

Embora recente e sem registro de aplicação em PDM's, a *LagClus* tem sido utilizada em problemas semelhantes e apresentado bons resultados (Lorena e Ribeiro, 2005; Lorena e Ribeiro, 2007; Lorena e Ribeiro, 2008). A proposta deste trabalho aborda a utilização da LagClus, tendo em vista que até o momento o cálculo de limitantes do PDM não foi abordado com a relaxação lagrangeana ou teve sua representação em grafo explorada.

A idéia do método se encontra no particionamento do grafo PDM em p clusters, que dividem o problema principal em p subproblemas, cuja formulação inclui a relaxação lagrangeana das arestas cortadas que conectam os elementos do próprio cluster aos demais fora deste.

Assim, dada a matriz de distâncias $D=[d_{ij}]_{nxn}$, cria-se o grafo $G=(V,A)$, sendo $V=N$ e $A=D$ (matriz de adjacencias com pesos). Do particionamento de G em p clusters obtém-se:

a) p clusters G_i, $i=1,..p$;

b) $V = V_1 \cup V_2 \cup \ldots \cup V_p$, onde V_i é o conjunto de vértices do cluster G_i;

c) $X_i = V - V_i$;

d) p subproblemas $LagPDM_k$, k=1...p.

Para cada $LagPDM_k$ identifica-se as seguintes restrições de ligação:

$$\begin{aligned}
&\text{r1) } x_i + x_j - y_{ij} \le 1, \ (i,j) \in R \\
&\text{r2) } y_{ij} - x_i \le 0, \ (i,j) \in R \\
&\text{r3) } y_{ij} - x_j \le 0, \ (i,j) \in R \\
&\text{onde } R = \{(i,j) \mid i \in V_k, j \in X_k\}
\end{aligned}$$

A restrição principal, $\sum_{i=1}^{n} x_i = m$, continua fazendo parte dos subproblemas, porém com uma alteração: $\sum_{i \in V_k} x_i = \lceil m/p \rceil$

Para cada subproblema $LagPDM_k$, os multiplicadores de lagrange serào trabalhados da seguinte forma:

a) $\alpha_{ij} \ge 0$, para as restrições do tipo r1;

b) $\beta_{ij} \ge 0$, para as restrições do tipo r2;

c) $\lambda_{ij} \ge 0$, para as restrições do tipo r3.

Assim, cada $LagPDM_k$ será escrito de modo a seguir:

$$v(LagPDM_k) = \max$$

$$\sum_{(i,j)\in R} \alpha_{ij} x_i + \sum_{(i,j)\in R} \beta_{ij} x_j - \sum_{(i,j)\in R} \lambda_{ij} x_i - \sum_{(i,j)\in R} \lambda_{ij} x_j + \sum_{(i,j)\in V_k} d_{ij} y_{ij} + \sum_{(i,j)\in R} y_{ij}(d_{ij} - \alpha_{ij} - \beta_{ij}) + \sum_{(i,j)\in R} y_{ij}(d_{ij} + \lambda_{ij})$$

$$\begin{aligned}
&\text{sujeito a :} \\
&x_i + x_j - y_{ij} \le 1, \ \forall i, j \in V_k \\
&y_{ij} - x_i \le 0, \ \forall i, j \in V_k \\
&y_{ij} - x_j \le 0, \ \forall i, j \in V_k \\
&y_{ij} \in \{0,1\}, \ \forall i \in V_k, \ j \in X_k \\
&x_i \in \{0,1\}, \ \forall i \in V_k
\end{aligned}$$

Após a otimização de cada $LagPDM_k$, a solução do problema será:

$$v(Lag^p PDM) = \sum_{k=1}^{p} LagPDM_k + \sum_{(i,j)\in Q, cl(i)\ne cl(j)} \lambda_{ij}$$

Onde $cl(i)$=cluster do elemento i, E seu dual lagrangeano:

$$v(DLCPDM^p) = \min\{Lag^p PDM\}$$

4. Testes Computacionais

A implementação foi desenvolvida sobre a linguagem C++ e o solver XPRESS (Dash Optimization, 2006) foi utilizado para resolver de forma exata os subproblemas da *LagClus* e a relaxação linear das instâncias.

Para a clusterização, foi utilizada a biblioteca **METIS** (Karypis e Kumar, 1998a; Karypis e Kumar, 1998b; Karypis e Kumar, 1998c) desenvolvida especificamente para o particionamento otimizado de grafos. Como a natureza do PDM é a dispersão, o particionamento foi feito com o objetivo de maximizar a soma dos pesos das arestas cortadas. A otimização do dual lagrangeano do modelo e obtenção dos limitantes foi calculada utilizando-se o algoritmo de subgradientes (Wolsey 1998; Narciso e Lorena 1999).

Até o momento as pesquisas têm explorado a obtenção de soluções viáveis de qualidade. Esses trabalhos apresentam resultados cada vez melhores, porém o estudo de técnicas para análise dessas soluções com o cálculo de limitantes é uma novidade para os PDM's.

Foram realizados diversos experimentos em ambiente Windows Vista 32 bits, Processador Intel Core 2, 1.66 GB e 2 GB RAM. Os resultados foram feitos com variações no número de partições, valores iniciais dos multiplicadores de lagrange e valor do peso do passo do método dos subgradientes (Narciso e Lorena 1999). Os testes foram divididos nas seguintes categorias:

a) **AA-** Análise dos melhores resultados apresentados em Aringhieri et al. (2005) para instâncias geradas por Andrade, com n = 100, 150, 200 e 250;
b) **AS-** Análise dos melhores resultados apresentados em Aringhieri et al. (2005) para instâncias geradas por Silva, com n = 300, 400 e 500;
c) **SS-** Análise dos melhores resultados apresentados em Silva et al. (2004b) para instâncias geradas por Silva, com n = 10, 20, 30, 40 e 50. A maior parte desses resultados é ótima e foi obtida através de métodos exatos.

As tabelas 4.1, 4.2 e 4.3 estão estruturadas da seguinte forma: o campo **Instância** indica o nome, sem extensão, do arquivo que contém a instância; **n** é o valor de n (número de elementos de N); **m** , o de de m (número de elementos a serem selecionados); **p**, o número de clusters da *LagClus* que apresentou o melhor resultado; $\overline{\mathbf{T}}$ é o tempo médio de processamento, em segundos; **GAP (%)**, o campo do valor do gap, de acordo com a seguinte equação: $GAP(\%) = \frac{v(sol) - z}{z} \times 100$. Onde z é o valor da melhor solução conhecida para a instância utilizada e $v(sol)$ é o valor do limitante obtido com a relaxação linear convencional no caso **v(PDML)**, ou o valor obtido com a *LagClus* no caso **v(DLCPDM^{p})**. O melhor limitante de cada instância está identificado em negrito.

A Tabela 4.1 apresenta os resultados obtidos na análise da categoria AA de experimentos. Os GAP's obtidos com a *LagClus* são semelhantes, porém ainda melhores do que os obtidos com a relaxação linear. Durante a execução dos testes, observou-se que o número de partições exerce uma forte influência tanto no tempo de processamento quanto na qualidade do limitante.

Tabela 4.1 – Análise de soluções da categoria AA

Instância	n	m	p	v(PDML) GAP (%)	v(DLCPDM^{p}) GAP (%)	$\overline{T}$
09a250m50	250	50	25	312,81	**306,62**	186
18b200m80	200	80	20	124,98	**123,09**	932
19b250m50	250	50	10	318,27	**302,53**	2460
20b250m100	250	100	20	127,51	**125,66**	650
23c100m20	100	20	5	186,14	**158,97**	858
36d150m60	150	60	15	111,76	**62,27**	559
38d200m80	200	80	20	110,86	**108,66**	922
40d250m100	250	100	20	110,00	**107,84**	819

As análises do tipo AS estão consolidadas na Tabela 4.2, onde pode ser verificado que os limitantes da *LagClus* são ligeiramente melhores do que os da relaxação linear. Contudo, devido ao tamanho das instâncias, há uma diferença importante na estragégia dos subgradientes para esta categoria. O peso do passo de atualização (Narciso e Lorena, 1999) decresce a cada **primeira** não atualização do limitante enquanto seu valor for maior ou igual a 0,1. Essa modificação foi feita na tentativa de conseguir uma convergência incial mais rápida, melhorando o tempo de processamento sem perder significativamente a qualidade do limitante.

Tabela 4.2 – Análise de soluções da categoria AS

Instância	n	m	p	v(PDML) GAP (%)	v(DLCPDM^P) GAP (%)	$\overline{T}$
matrizn300m60	300	60	15	315,97	**299,23**	2293
matrizn300m90	300	90	15	191,62	**181,83**	3370
matrizn300m120	300	120	15	125,68	**119,66**	4292
matrizn400m120	400	120	20	196,94	**190,24**	4409
matrizn400m160	400	160	32	129,92	**126,25**	526
matrizn500m150	500	150	25	187,76	**183,62**	789
matrizn500m200	500	200	25	130,54	**127,97**	1775

Na Tabela 4.3 são apresentadas as análises do tipo SS. A diferença entre os limtantes é considerável, sendo sempre os da *LagClus* significativamente melhores. A facilidade de trabalar com poucos clusters (2 a 5 partições) em pequenas instâncias é o motivo aparente para essa diferença na qualidade dos limitantes nesses casos. Os "*" marcam as instâncias cujo valor de z utilizado foi o valor **ótimo** obtido através de método exato em Silva et al. (2004b). Para as instâncias com o nome sublinhado, o valor do peso passo do algoritmo dos subgradientes (Narciso e Lorena, 1999) foi atualizado dinamicamente, conforme descrito anteriormente.

Tabela 4.3 – Análise de soluções da categoria SS

Instância	n	m	p	v(PDML) GAP (%)	v(DLCPDM^P) GAP (%)	$\overline{T}$
matrizn10m4*	10	2	2	377,78	**179,05**	10
matrizn10m4*	10	3	3	158,00	**77,83**	9
matrizn10m4*	10	4	2	82,98	**21,90**	9
matrizn20m4*	20	2	2	833,33	**393,91**	19
matrizn20m4*	20	4	2	236,00	**108,56**	41
matrizn20m4*	20	6	2	131,19	**57,82**	61
matrizn20m4*	20	8	2	85,64	**34,72**	39
matrizn30m3*	30	3	3	609,63	**412,39**	10
matrizn30m3*	30	6	2	216,69	**102,96**	150
matrizn30m3*	30	9	3	126,30	**85,56**	48
matrizn30m3*	30	12	2	86,02	**37,91**	254
matrizn40m4*	40	4	2	570,00	**301,40**	137
matrizn40m4*	40	8	2	258,69	**134,27**	366
matrizn40m4	40	12	2	158,57	**81,19**	537
matrizn40m4	40	16	2	106,45	**54,91**	172
matrizn50m5	50	5	2	532,33	**427,32**	25
matrizn50m5	50	10	5	249,71	**163,88**	140
matrizn50m5	50	15	3	150,21	**102,09**	147
matrizn50m5	50	20	4	100,11	**74,41**	51

5. Considerações Finais

Este trabalho apresenta uma nova abordagem para o cálculo de limitantes de PDM's, a *LagClus*. Foram realizados diversos experimentos e dependendo do número de clusters trabalhados e dos parâmetros do algoritmo dos subgradientes a *LagClus* pode apresentar limitantes melhores do que a relaxação linear das variáveis de decisão da formulação linearizada do problema.

Contudo o PDM possui características que dificultam a aplicação da *LagClus*:

d) A matriz de adjacências possui uma densidade de 100%;
e) A restrição principal não é trabalhada com a mesma facilidade que as demais nos subproblemas;
f) Os subproblemas, se muito grandes, ainda são difíceis de serem resolvidos por algum método exato. Nesses casos foi imposto um limite de tempo de 20 segundos após a primeira solução inteira encontrada e a solução considerada foi a melhor solução inteira dentro desse limite de tempo. Por esse motivo, a qualidade dos limitantes apresentou um comportamento inverso ao esperado em grandes instâncias, melhorando o resultado para um número maior de partições, uma vez que subproblemas pequenos foram resolvidos até sua otimlidade, enquanto os grandes podem não atingir o ponto ótimo.

Os resultados sugerem o investimento em esforços para o aprimoramento de métodos baseados na divisão em clusters, a fim de encontrar limitantes melhores e em tempos menores.

6. Referências

Aringhieri, R.; Cordone, R. & Melzani, Y., "Tabu Search vs. GRASP for the Maximum Diversity Problem". DTI - University of Milano, Note del Polo 89, 2005.

Aringhieri, R.; Bruglieri, M. & Cordone, R., "Semidefinite Bounds for the Maximum Diversity Problem". DTI - University of Milano, Note del Polo 95, 2006.

Aringhieri, R. & Cordone, R., "Better and Faster Solutions for the Maximum Diversity Problem". DTI - University of Milano, Note del Polo 93, 2006.

Dash Optimization, XPRESS-MP 2006B, "Xpress-Optimizer Reference Manual", 371 p, 2006.

Duarte A.; Gallego M. & Martí R., "An exact method for the maximum diversity problem". Universitat de València, 2007a, Disponível em: http://www.uv.es/sestio/TechRep/technicals.html , acesso em 28 nov. 2007.

Duarte A.; Gallego, M.; Laguna, M. & Martí R., "Hybrid Heuristics for the Maximum Diversity Problem". Computational Optimization and Applications, 2007b, disponível em: http://leeds-faculty.colorado.edu/Laguna/publications.htm , acesso em 26 nov. 2007.

Ghosh, J. B., "Computational aspects of the maximum diversity problem". Operations Research Letters, Vol. 19, p. 175–181, 1996.

Glover, F.; Hersh G. & Mcmilian, C., "Selecting subset of maximum diversity". MS/IS 77-9, University of Colorado at Boulder, 1977.

Glover, F.; Kuo C.C. & Dhir, K.S., "Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming". Decision Sciences, Vol. 24, Issue 6, p. 1171-1185, 1993.

Glover, F.; Kuo, C.C. & Dhir, K.S., "A discrete optimization problem model for preserving biological diversity". Applied Mathematical Modeling, Vol. 19, n. 11, p. 696-701, 1995.

Karypis, G. & Kumar, V., “A fast and high quality multilevel scheme for partitioning irregular graphs”. SIAM Journal on Scientific Computing, Vol. 20, p. 359–392, 1998a.

Karypis, G. & Kumar, V., “Multilevel algorithms for multi-constraint graph partitiong”. Army HPC Research Center, Dep. of Computer Science: University of Minnesota, Technical Report 98-019, 1998b.

Karypis, G. & Kumar, V., “Multilevel k-way partitioning scheme for irregular graphs”. Journal of Parallel and Distributed Computing, Vol. 48, n. 1, p. 96–129, 1998c.

Lorena, L.A.N. & Ribeiro, G.M., “Relaxação lagrangeana com divisão em clusters para alguns problemas de otimização modelados em grafos de conflitos”. Proposta de teste de doutorado do curso de pós graduação em computação aplicada – INPE, 2005.

Lorena, L.A.N. & Ribeiro, G.M., “Lagrangean relaxation with clusters and column generation for the manufacturer's pallet loading problem”. Computers & Operations Research, Vol. 34(9), p. 2695-2708, 2007.

Lorena, L.A.N. & Ribeiro, G.M., “Lagrangean relaxation with clusters for point-feature cartographic label placement problems”. Computers & Operations Research, Vol. 35, p. 2129-2140, 2008.

Narciso, M.G. & Lorena, L.A.N., “Lagrangean/surrogate relaxation for assignment problems”. European Journal of Operational Research, Vol. 114, p. 165-177, 1999.

Silva, G.C.; Ochi, L.S. & Martins, S.L. “O Problema da Diversidade Máxima: Proposta e Análise de Metaheurística GRASP”. Tendências em Matemática Aplicada e Computacional da SBMAC, submetido em 2003. Disponível em: http://www.ic.uff.br/~gsilva/publicacoes.html. Acesso em 30 nov. 2007.

Silva, G.C.; Andrade, M. R.; Martins, S.L. & Ochi, L.S., “Experimental comparison of greedy randomized adaptive search procedures for the Maximum diversity problem”. Lecture Notes in Computer Science, issue 3059, p. 498-512, 2004a.

Silva, G.C.; Martins, S.L. & Ochi, L.S., “Análise de Heurísticas GRASP para o Problema da Diversidade Máxima”. Dissertação de Mestrado em Ciências da Computação – UFF – RJ, 2004b.

Silva, G.C.; Andrade, M.R.; Martins, S.L.; Ochi, L.S. & Plastino, A., “New heuristics for the maximum diversity problem”. Journal of Heuristics, ago. 2007, Vol.13, issue 4, p. 315-336, 2007.

Wolsey, L.A., “Integer Programming, Editora” John Wiley & Sons, 1998, p. 288, ISBN: 0-471-28366-5, 1998.