Formulação de fluxo em arcos para problemas de agrupamento capacitado

Vítor G. Chagas[1], **Manuel Iori**[2]

[1] Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Cidade Universitária Zeferino Vaz - Barão Geraldo, Campinas, SP

[2]DISMI – University of Modena and Reggio Emilia (UNIMORE)
Via Amendola 2, 42122 Reggio Emilia, Itália

Abstract. *In this work, we present a formulation based on arc-flow models, originally designed for bin packing problems, and now applied to clustering problems. We consider the Correlation Clustering and Graph Partitioning problems with node weights and a cost associated to each cluster. We discuss the advantages of this formulation in comparison with other strategies presented in the literature.*

Resumo. *Neste trabalho, apresentamos uma formulação baseada em modelos de fluxo em arcos, originalmente projetados para problemas de empacotamento, e agora aplicada a problemas de agrupamento. Consideramos os problemas de Correlation Clustering e Graph Partitioning com pesos nos vértices e custo associado a cada cluster utilizado. Discutimos as vantagens de tal formulação comparada a outras estratégias presentes na literatura.*

1. Introdução

O problema de agrupar um conjunto de dados que possuem informações similares é um problema com bastante interesse teórico e prático, possuindo aplicações em várias áreas, como mineração de dados, biologia computacional, compressão de dados, aprendizado de máquina, reconhecimento de padrões e visão computacional.

Um tipo de problema de agrupamento (*clustering*) é o *Correlation Clustering* (CC), onde é dado um grafo $G = (V, E)$ tal que cada aresta $e \in E$ possui um sinal $s(e) \in \{+, -\}$. A versão de minimização de discordâncias visa encontrar um particionamento dos vértices em clusters tal que o número de arestas $-$ dentro de clusters mais o número de arestas $+$ entre clusters seja minimizado. A versão de maximização de concordâncias é definida de forma similar. Um *survey* sobre o problema e suas variações pode ser encontrado em [Becker 2005].

Outro problema de *clustering* bastante estudado é o problema de Particionamento em Grafos (*Graph Partitioning* – GP). Nesta variante é dado um grafo $G = (V, E)$ e uma função de pesos $d : E \to \mathbb{R}$ que representa a distância ou dissimilaridade entre os vértices. Assim como no CC, também se deseja encontrar um particionamento de V em clusters, porém o objetivo é minimizar a soma dos pesos das arestas que estão em um mesmo cluster. Um *survey* sobre o problema pode ser encontrado em [Buluç et al. 2016].

Neste trabalho, investigamos a variante do CC e GP com pesos nos vértices, e com um custo associado a cada cluster utilizado. Tal cenário surge no contexto de

logística, em que se deseja realizar o carregamento de uma grande variedade de produtos em contêineres, e se queira minimizar a quantidade de contêineres utilizados, porém considerando que não se deseja carregar produtos muito diferentes juntos, como remédios e veneno, ou eletrônicos e alimentos. Outra aplicação é relacionada com anúncios, em que há um conjunto de propagandas que devem ser dispostas em *banners* de uma página web ou uma aplicação móvel. Deseja-se encontrar uma disposição que utilize uma quantidade pequena de *banners*, considerando que propagandas de produtos bem relacionados estejam dispostas no mesmo *banner*. Nessas situações, é relevante não só a escolha dos elementos que estarão agrupados, como também a quantidade de grupos utilizados.

O restante do texto é organizado da seguinte forma: na Seção 2 descrevemos os problemas que serão abordados formalmente. Na Seção 3 apresentamos formulações matemáticas para tais problemas, e na Seção 4 concluímos com algumas considerações.

2. Definição dos problemas

Seja $N = \{1, \ldots, n\}$ um conjunto de n itens, $w : N \to \mathbb{R}_+$ uma função que representa o peso de cada item, $d : N\text{x}N \to \mathbb{R}$ uma função que representa a distância entre pares de itens, com $d_{ii} = 0$, $d_{ij} = d_{ji}\ \forall i, j \in N$, e $W, c \in \mathbb{R}_+$ a capacidade e o custo de uso de um cluster, respectivamente. Sem perda de generalidade, consideramos que $w_i \leq W$ e $w_i \geq w_j$ para $i < j$. No problema de GP capacitado (CGP), queremos encontrar uma partição de N dada por $\mathcal{C} = \{C_1, \ldots, C_k\}$ tal que $ck + \sum_{C \in \mathcal{C}} \sum_{i,j \in C} d_{ij}$ é mínimo e para todo $C \in \mathcal{C}$, $\sum_{i \in C} w_i \leq W$.

Para o problema de CC capacitado (CCC), consideramos que cada aresta $e \in E$ possui um rótulo $+$ ou $-$, definido por $s(e)$. Para $S, T \subseteq V$, sendo $E_+[S]$ (resp. $E_-[S]$) o conjunto das arestas induzidas por S com rótulo $+$ (resp. $-$), e sendo $\delta_+(S,T) = \{(u,v) : u \in S, v \in T, s(u,v) = +\}$, o objetivo do CCC é encontrar uma partição de N dada por $\mathcal{C} = \{C_1, \ldots, C_k\}$ que minimiza $ck + \sum_{i=1}^{k}(\sum_{e \in E_-[C_i]} d_e) + \sum_{i \neq j}(\sum_{e \in \delta_+(C_i, C_j)} d_e)$, também satisfazendo $\sum_{i \in C} w_i \leq W$ para todo $C \in \mathcal{C}$.

3. Formulações

Inicialmente, apresentamos uma formulação mais tradicional para o CGP. Para $1 \leq i \leq j \leq n$ e $1 \leq k \leq n$, seja x_{ij}^k uma variável de decisão binária que indica se os itens i e j são atribuídos ao cluster k (consequentemente, x_{ii}^k indica se o item i pertence ao cluster k) e y_k uma variável que indica se o cluster k está sendo utilizado. Uma formulação para o problema é fornecida a seguir:

$$
\begin{aligned}
(F_T)\ \min \quad & \textstyle\sum_{k \in N} \sum_{i \in N} \sum_{j > i} d_{ij} x_{ij}^k + c \sum_{k \in N} y_k && && (1)\\
\text{s. a} \quad & x_{ii}^k \leq y_k && \forall i, k \in N && (2)\\
& \textstyle\sum_{k \in N} x_{ii}^k = 1 && \forall i \in N && (3)\\
& \textstyle\sum_{i \in N} w_i x_{ii}^k \leq W && \forall k \in N && (4)\\
& x_{ij}^k \geq x_{ii}^k + x_{jj}^k - 1 && \forall i, j, k \in N, i < j && (5)\\
& x_{ij}^k \leq x_{ii}^k && \forall i, j, k \in N, i < j && (6)\\
& x_{ij}^k \leq x_{jj}^k && \forall i, j, k \in N, i < j && (7)\\
& \text{variáveis } x \text{ e } y \text{ binárias} && && (8)
\end{aligned}
$$

As restrições (2) impõem que um item é atribuído apenas a clusters utilizados. As restrições (3) garantem que cada item é atribuído a exatamente um cluster. As restrições (4) representam as restrições de capacidade de cada cluster. As restrições (5) indicam que $x_{ij}^k = 1$ sempre que os itens i e j forem atribuídos ao cluster k, enquanto as restrições (6) e (7) fazem com que $x_{ij}^k = 0$ se o item i ou item j não foram atribuídos ao cluster k. Por fim, as restrições (8) se referem às restrições de integralidade das variáveis.

Apresentamos outra formulação para o CGP, baseada nas formulações de *arc-flow* para problemas de empacotamento (veja [Delorme e Iori 2020]). Podemos interpretar a atribuição de itens a um cluster com capacidade W como um empacotamento dos itens em um recipiente (*bin*) de mesma capacidade. Além disso, a atribuição de itens em um bin pode ser modelada como um problema de caminhos em um digrafo. Seja digrafo $G = (V, A)$, onde $V = \{0, \ldots, W\}$ e $A = A_i \cup A_l$, com $A_i = \{(u, v) : u, v \in V, \exists j \in N : v - u = w_j\}$ e $A_l = \{(u, u + 1) : u \in \{1, \ldots, n - 1\}\}$. O conjunto de vértices V representa as possíveis posições em que um item pode ser empacotado. Um arco $(u, v) \in A_i$ representa um item de tamanho $v - u$ empacotado na posição u, e um arco $(u, u + 1) \in A_l$ representa uma unidade de espaço vazio no bin. Dessa forma, um empacotamento é representado por um caminho de 0 a W em G. Com isso, podemos derivar uma formulação pseudopolinomial para o CGP. Considere os itens artificiais 0 e $n + 1$, com $w_0 = w_{n+1} = 0$, que são utilizados para representar o início e o fim de um caminho que representa um empacotamento, e seja $N' = N \cup \{0, n + 1\}$. Seja $x_{i,p,j,p+w_j}$ uma variável binária que vale 1 se o item j é empacotado imediatamente após o item i com início na posição p e término na posição $p + w_j$, e 0 caso contrário. Note que o quarto índice é redundante e é usado apenas para melhorar o entendimento, não interferindo na quantidade de variáveis. Para reduzir o espaço de possibilidades, os valores p em que um item pode ser atribuído é limitado pelos padrões normais, definido como $\mathcal{N} = \{x \ : \ x \leq W, x = \sum_{i \in N} w_i \epsilon_i, \ \epsilon_i \in \{0, 1\} \ \forall i \in N\}$, e representa o conjunto de possíveis coordenadas em que um item pode ser atribuído se todos os itens forem deslocados o máximo possível para a esquerda. Seja também y_{ij} uma variável binária que indica se o item j é empacotado imediatamente após o item i. Tal variável é redundante com as variáveis x, como será visto na formulação a seguir, e é utilizada apenas para simplificar o modelo. Por fim, sejam as variáveis binárias z_{ij}, que indica se os itens i e j estão no mesmo *bin*, e b_i^k, que indica se o item i foi atribuído ao *bin* k. Consideramos apenas as variáveis x, y e z com $i < j$. A formulação, denotada por F_{CGP}, é dada em (9) – (19).

A função objetivo (9) minimiza a soma dos pesos das arestas que estão em um mesmo cluster, mais o custo de uso dos clusters, pois como todo empacotamento é feito a partir da posição 0, cada caminho que parte do vértice 0 equivale a um cluster na solução. As restrições (10) relacionam as variáveis y com as variáveis x. As restrições (11) e (12) impõem que todo item é precedido e sucedido por exatamente um outro item (que pode ser um dos itens 0 ou $n + 1$), respectivamente. As restrições (13) se referem às restrições de conservação de fluxo. As restrições (14) garantem que todo item pertença a exatamente um bin. As restrições (15) fazem com que o índice do primeiro item que é alocado em um bin seja usado como índice do bin ao qual ele pertence, a fim de evitar simetrias. As restrições (16) indicam que se o item i precede j em algum bin, então i e j estão no mesmo bin, e as restrições (17) triangulam as variáveis z. As restrições (18) fazem com que b_j^k seja 1 se os itens i e j estão no mesmo bin e o item i está no bin k. Além disso, esse conjunto de restrições é responsável por fixar z_{ij} em 0 se os itens i e j estão em bins

diferentes. Tal fixação é importante quando $d_{ij} < 0$. Se todas as arestas possuem peso não-negativo, podemos remover as variáveis b e as restrições (14), (15) e (18). Por fim, as restrições (19) se referem às restrições de integralidade das variáveis.

(F_{CGP})

$$\min \quad \sum_{i\in N}\sum_{j>i} d_{ij} z_{ij} + c\sum_{i\in N} x_{0,0,i,w_i} \tag{9}$$

$$\text{s. a} \quad y_{ij} = \sum_{\substack{p\in\mathcal{N} \\ w_i \le p \le W - w_j}} x_{i,p,j,p+w_j} \qquad \forall i,j \in N', i<j \tag{10}$$

$$\sum_{i<j} y_{ij} = 1 \qquad \forall j \in N \tag{11}$$

$$\sum_{j>i} y_{ij} = 1 \qquad \forall i \in N \tag{12}$$

$$\sum_{\substack{i<j \\ p-w_j \ge w_i}} x_{i,p-w_j,j,p} - \sum_{\substack{k>j \\ p+w_k \le W}} x_{j,p,k,p+w_k} = 0 \qquad \forall j \in N, p \in \mathcal{N} \tag{13}$$

$$\sum_{k\in N} b_i^k = 1 \qquad \forall i \in N \tag{14}$$

$$b_i^i = y_{0i} \qquad \forall i \in N \tag{15}$$

$$z_{ij} \ge y_{ij} \qquad \forall i,j \in N, i<j \tag{16}$$

$$z_{ik} \ge z_{ij} + z_{jk} - 1 \qquad \forall i,j,k \in N, i<j<k \tag{17}$$

$$b_j^k \ge b_i^k + z_{ij} - 1 \qquad \forall i,j,k \in N, i<j \tag{18}$$

$$\text{variáveis } x, y, z, b \text{ binárias} \tag{19}$$

Note que com as variáveis z, sabemos exatamente o conjunto de arestas cujos extremos estão dentro de um mesmo cluster. Dessa forma, torna-se simples adaptar F_{CGP} para resolver o CCC, bastando alterar a função objetivo da forma apresentada em F_{CCC}.

$$(F_{\mathrm{CCC}}) \ \min \quad \sum_{(i,j)\in E_-} d_{ij} z_{ij} + \sum_{(i,j)\in E_+} d_{ij}(1 - z_{ij}) + c\sum_{i\in N} x_{0,0,i,w_i} \tag{20}$$

$$\text{s. a} \quad (10) - (19)$$

4. Considerações finais

Apresentamos duas formulações para o CGP, F_T e F_{CGP}. Experimentos computacionais indicam que a formulação F_{CGP} geralmente encontra soluções ótimas mais rapidamente do que F_T em cenários com grafos esparsos e com custos de cluster altos comparados com os valores de dissimilaridade entre os itens. Outra característica relevante em relação à F_{CGP} é a facilidade de adaptação da formulação para outros problemas de *clustering*, como mostrado em F_{CCC}. Ademais, ela pode ser adaptada para situações em que há necessidade de se colocar um separador entre itens adjacentes, isto é, se os itens i e j são empacotados lado a lado, então um separador de tamanho c_{ij} deve ser colocado entre eles. Para tal, podemos adaptar as variáveis x de $x_{i,p,j,p+w_j}$ para $x_{i,p,j,p+w_j+c_{ij}}$.

Agradecimentos. O presente trabalho foi realizado com apoio do CNPq (Proc. 425340/2016-3), da FAPESP (Proc. 2015/11937-9, 2016/01860-1 e 2019/12728-5) e do Quinto Andar.

Referências

Becker, H. (2005). A survey of correlation clustering. *Advanced Topics in Computational Learning Theory*, pages 1–10.

Buluç, A., Meyerhenke, H., Safro, I., Sanders, P., e Schulz, C. (2016). *Recent Advances in Graph Partitioning*, pages 117–158. Springer International Publishing, Cham.

Delorme, M. e Iori, M. (2020). Enhanced pseudo-polynomial formulations for bin packing and cutting stock problems. *INFORMS Journal on Computing*, 32(1):101–119.